arXiv:1506.07216v3 [cs.LG] 10 May 2016

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

Mark Braverman[1], Ankit Garg[1], Tengyu Ma[1], Huy L. Nguyen[2], and David P. Woodruff[3]

[1]Princeton University
[2]Toyota Technological Institute at Chicago
[3]IBM Research Almaden

May 11, 2016

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 θ which is promised to be k-sparse. The machines communicate by message passing and aim to estimate the mean θ. 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 results improve upon [Sha14, SD15] by allowing multi-round iterative 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.

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].

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 θ. 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[1]. 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 cost per machine needs to be $\Omega(\min\{n, d\})$ where d is the ambient dimension and n is the dimension of the observation 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 Theorem C.1).

[1] the dependency on the ambient dimension is typically logarithmic.

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 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 $X_1, \ldots, X_m$ from μ_v and the j-th machine receives one sample X_j, for $j = 1, \ldots, m$. We use $\Pi \in \{0, 1\}^*$ to denote the sequences of messages that are communicated by the machines. We will refer to Π 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 $\hat{v} : \{0, 1\}^* \to \{0, 1\}$ that takes the transcript Π 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 Π carries $\Omega(1)$ information about the random variable V. Therefore, the mutual information $\mathrm{I}(V; \Pi)$ 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 Π form a simple Markov chain $V \to X \to \Pi$. The channel $V \to X$ is defined as $X \sim \mu_v$, conditioned on $V = v$. The strong data processing inequality (SDPI) captures the relative ratio between $\mathrm{I}(V; \Pi)$ and $\mathrm{I}(X; \Pi)$.

Definition 1 (Special case of SDPI)**.** Let $V \sim B_{1/2}$ and the channel $V \to X$ be defined as above. Then there exists a constant $\beta \leq 1$ that depends on μ_0 and μ_1, such that for any Π that depends only on X (that is, $V \to X \to \Pi$ forms a Markov Chain), we have

$$\mathrm{I}(V; \Pi) \leq \beta \cdot \mathrm{I}(X; \Pi). \tag{1}$$

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

Observe that the LHS of (1) measures how much information Π carries about V, which is closely related to the accuracy of the protocol. The RHS of (1) is a lower bound on the expected

[2] Inequality (1) is always true for a Markov chain $V \to X \to \Pi$ with $\beta = 1$ and this is called the data processing inequality.

length of Π, that is, the expected communication cost. Therefore the inequality relates two quantities that we are interested in - the statistical quality of the protocol and the communication cost of the protocol. Concretely, when $q = 1/2$, in order to recover V from Π, we need that $\mathrm{I}(V;\Pi) \geq \Omega(1)$, and therefore inequality (1) gives that $\mathrm{I}(X;\Pi) \geq \Omega(\beta^{-1})$. Then it follows from Shannon's source coding theory that the expected length of Π (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.[3]

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 β is such that $(1-\sqrt{\beta})\mu_1 \leq \mu_0 \leq (1+\sqrt{\beta})\mu_1$, then

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

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[4]. Moreover, the relation between β, μ_0 and μ_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 β 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 β, and $\mu_0 \leq O(1)\mu_1$, then the following prior-free multi-machine SDPI is true with the same constant β (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,*

$$\mathrm{h}^2(\Pi|_{V=0}, \Pi|_{V=1}) \leq Kc\beta(\mu_0,\mu_1) \cdot \min\{\mathrm{I}(X_1 \dots X_m;\Pi \mid V=0), \mathrm{I}(X_1 \dots X_m;\Pi \mid V=1)\} \quad (2)$$

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

Moreover, for any μ_0 and μ_1 which satisfy the condition of the theorem, there exists a protocol that produces transcript Π 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.

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 Π as input) can recover v with probability $3/4$. Then*

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

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 β. 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.

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

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

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 $\mathrm{I}(X_1 \dots X_m; \Pi \mid V)$ 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 $\mathrm{I}(\Pi; V)$, 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 Π. 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 $\mathrm{I}(V; \Pi)$ 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 $\mathrm{I}(V; \Pi)$ (see Lemma 10, and the lower bound side was proved by [BYJKS04]),

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

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 $\mathrm{h}^2(\Pi|_{V=0}, \Pi|_{V=1}) = \Omega(1)$. In fact, a protocol for which inequality (2) is tight is one in which only a single machine sends a message Π which maximizes $\mathrm{I}(\Pi; V)/\mathrm{I}(\Pi; X)$.

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 and in Section A we prove Theorem 4.3 and Corollary 4.8. In Section 6 we provide tools for proving single machine strong data processing inequality and prove Theorem 4.1. In Section B we present our matching upper bound in the simultaneous communication model. In section C we give a simple proof of distributed gap majority problems using our machinery.

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 θ using m machines. Machine j receives n i.i.d samples $X_j^{(1)}, \dots, X_j^{(n)}$ from distribution μ_θ. 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 μ^n denotes the product of n copies of μ. 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 θ 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 Π. We will refer to Π as the transcript and note that $\Pi \in \{0,1\}^*$ is written in bits and the communication cost is defined as the length of Π, denoted by $|\Pi|$. We will call the algorithm that the machines follow to produce Π a protocol. With a slight abuse of notation, we ue Π to denote both the protocol and the transcript produced by the protocol.

One of the machines needs to estimate the value of θ using an estimator $\hat{\theta} : \{0,1\}^* \to \mathbb{R}^d$ which takes Π as input. The accuracy of the estimator on θ 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 θ,

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

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

$$\mathrm{IC}(\Pi) = \sup_{\theta \in \Omega} \mathrm{I}_\theta(\Pi; X \mid R_{\mathrm{pub}}) \tag{4}$$

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

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

$$\text{min-IC}(\Pi) = \min_{v \in \{0,1\}} \mathrm{I}_v(\Pi; X \mid R_{\mathrm{pub}}) \tag{5}$$

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 $\mathrm{IC}(\Pi) \le I$, $\mathrm{CC}(\Pi) \le C$ and $R(\Pi, \hat{\theta}) \le R$.

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,

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

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 $\mathrm{IC}(\Pi) \le I$, $R_{det}(\Pi, \hat{v}) \le 1/4$.

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 $\mathrm{GD}(n, m, \delta, \sigma^2)$.

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 $\mathrm{GME}(n, m, d, \sigma^2)$. When $\Omega = \{\theta \in \mathbb{R}^d : |\theta|_0 \le k\}$, the corresponding problem is referred to as distributed sparse Gaussian mean estimation, abbreviated $\mathrm{SGME}(n, m, d, k, \sigma^2)$.

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 $\Omega = \{\theta \in$

$\mathbb{R}^d : |\theta|_0 \le k\}$. 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 $j = 1, \ldots, mn$, and each machine receives n data points among them[5]. Formally, let μ_θ denote the joint distribution of (A_j, y_j) here, and let $\mathcal{P} = \{\mu_\theta : \theta \in \Omega\}$. We use $\mathrm{SLR}(n, m, d, k, \sigma^2)$ 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 $f, g : \Omega \to \mathbb{R}$. The square of the Hellinger distance between f and g is defined as

$$\mathrm{h}^2(f, g) := \frac{1}{2} \cdot \int_\Omega \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx$$

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

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

where $p_{i,\pi}(\cdot)$ is a function that only depends on i and the entire transcript π . To see this, one could simply write the density of π as a products of density of each messages 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 π).

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

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

$$\Pr_{X \sim \mu_{\boldsymbol{b}}}[\Pi = \pi] = q_{1,\pi}(b_1) \ldots q_{m,\pi}(b_m), \tag{7}$$

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

Proposition 2.1 (Cut-paste property of a protocol)**.** *For any $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}, \boldsymbol{d}$ with $\{a_i, b_i\} = \{c_i, d_i\}$ (in a multi-set sense) for every $i \in [m]$,*

$$\Pi_{\boldsymbol{a}}(\pi) \cdot \Pi_{\boldsymbol{b}}(\pi) = \Pi_{\boldsymbol{c}}(\pi) \cdot \Pi_{\boldsymbol{d}}(\pi) \tag{8}$$

and therefore,

$$\mathrm{h}^2(\Pi_{\boldsymbol{a}}, \Pi_{\boldsymbol{b}}) = \mathrm{h}^2(\Pi_{\boldsymbol{c}}, \Pi_{\boldsymbol{d}}) \tag{9}$$

[5]We note that here for convenience, we use subscripts for samples, which is different from the notation convention used for previous problems.

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 $\mu_1 \le c \cdot \mu_0$, and $\beta(\mu_0, \mu_1) = \beta$, we have*

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

where K is an absolute constant.

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

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

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: $X'_j \sim \mu_W$, and for any $j \neq i$, $X'_j \sim \mu_0$. We apply the protocol on the input X', and view the resulting transcript Π' as communication between the i-th machine and the remaining machines. Then we are in the situation of a single machine case, that is, $W \to X'_i \to \Pi'$ forms a Markov Chain. Applying the data processing inequality (1), we obtain that

$$\mathrm{I}(W; \Pi') \le \beta \mathrm{I}(X'_i; \Pi')\cdot \tag{12}$$

Using Lemma 10, we can lower bound the LHS of (12) by the Hellinger distance and obtain

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

Let $\boldsymbol{e_i} = (0, 0, \ldots, 1, \ldots, 0)$ be the unit vector that only takes 1 in the ith 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 $\Pi_{\boldsymbol{e}_i}$. Then we can rewrite the equation above as

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\boldsymbol{e}_i}) \le \beta \cdot \mathrm{I}(X'_i; \Pi') \tag{13}$$

Observe that the RHS of (13) is close to the first entry of the LHS of (11) since the joint distribution of (X'_1, Π') is not very far from $X, \Pi \mid V = 0$. (The only difference is that X'_1 is drawn from a mixture of μ_0 and μ_1, and note that μ_0 is not too far from μ_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 $\mu_1 \le c \cdot \mu_0$, and $\beta(\mu_0, \mu_1) = \beta$, then*

$$\mathrm{h}^2(\Pi_{\boldsymbol{e_i}}, \Pi_{\mathbf{0}}) \le \frac{(c+1)\beta}{2} \cdot \mathrm{I}(X_i; \Pi \mid V = 0) \tag{14}$$

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

$$\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \le O(1) \cdot \sum_{i=1}^{m} \mathrm{h}^2(\Pi_{\boldsymbol{e_i}}, \Pi_{\mathbf{0}}) \tag{15}$$

Using Lemma 1 and Lemma 2, we obtain Theorem 3.1 straightforwardly by combining inequalities (11), (14) and (15)[6].

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

Proof of Lemma 1. Let W be uniform Bernoulli random variable and define X' and Π' as follows: Conditioned on $W = 0$, $X' \sim \mu_{\mathbf{0}}$ and conditioned on $W = 1$, $X' \sim \mu_{\boldsymbol{e_i}}$. We run protocol on X' and get transcript Π'.

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

$$\beta \cdot \mathrm{I}(X'_i; \Pi') \geq \mathrm{I}(V; \Pi'). \tag{16}$$

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 10, to lower bound the right hand side. Note that Z in Lemma 10 corresponds to V here and ϕ_{z_1}, ϕ_{z_2} corresponds to $\Pi_{\boldsymbol{e_i}}$ and $\Pi_{\mathbf{0}}$. Therefore,

$$\mathrm{I}(V; \Pi') \geq \mathrm{h}^2(\Pi_{\boldsymbol{e_i}}, \Pi_{\mathbf{0}}). \tag{17}$$

It remains to relate $\mathrm{I}(X'_i; \Pi')$ to $\mathrm{I}(X_i; \Pi \mid V = 0)$. Note that the difference between joint distributions of (X'_i, Π') 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 11) that since $\mu_0 \geq \frac{2}{c+1}(\frac{\mu_0+\mu_1}{2})$, we have

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

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 μ_0 or μ_1. We will denote the product of n i.i.d copies of μ_v by μ_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 μ_0 and μ_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 μ'_0, μ'_1, the truncation of μ_0 and μ_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 β. [Rag14] proves that Gaussian distributions μ_0 and μ_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 μ'_0 and μ'_1. Our proof essentially uses the fact that $(\mu'_0 + \mu'_1)/2$ is a log-concacve 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 Section 6.3.

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

Theorem 4.1. *Let μ_0' and μ_1' be the distributions obtained by truncating μ_0 and μ_1 on support $[-\tau, \tau]$ for some $\tau > 0$. If $\delta \le \sigma$, we have $\beta(\mu_0', \mu_1') \le \delta^2/\sigma^2$.*

As a corollary, the SDPI constant between n copies of μ_0' and μ_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 μ_0^n and μ_1^n outside the ball $\mathcal{B} = \{x \in \mathbb{R}^n : |x_1 + \cdots + x_n| \le \tau\}$. Then when $\sqrt{n}\delta \le \sigma$, we have*

$$\beta(\tilde{\mu}_0, \tilde{\mu}_1) \le n\delta^2/\sigma^2$$

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, $\Omega\left(\frac{\sigma^2}{n\delta^2}\right)$ 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 $\mathrm{GD}(n, m, \delta, \sigma^2)$ with $\delta \le \sigma/\sqrt{n}$ requires communication cost and minimum information cost at least,*

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

Remark 3. The condition $\delta \le \sigma/\sqrt{n}$ captures the interesting regime. When $\delta \gg \sigma/\sqrt{n}$, a single machine can even distinguish μ_0 and μ_1 by its local n samples.

Proof of Theorem 4.3. Let $\Pi_{\mathbf{0}}$ and $\Pi_{\mathbf{1}}$ be the distribution of $\Pi|V = 0$ and $\Pi|V = 1$ as defined in Section 2.2. Since $\hat{v}$ solves the detection problem, we have that $\|\Pi_{\mathbf{0}} - \Pi_{\mathbf{1}}\|_{\mathrm{TV}} \ge 1/4$. It follows from Lemma 9 that $\mathrm{h}(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \ge \Omega(1)$.

We pick a threshold $\tau = 20\sigma$, and let $\mathcal{B} = \{z \in \mathbb{R}^n : |z_1 + \cdots + z_n| \le \sqrt{n}\tau\}$. Let $F = 1$ denote the event that $X = (X_1, \ldots, X_n) \in \mathcal{B}$, and otherwise $F = 0$. Note that $\Pr[F = 1] \ge 0.95$ 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

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

We run our whole argument conditioning on the event $F = 1$. First note that for any Markov chain $V \to X \to \Pi$, and any random variable F that only depends on X, the chain $V|_{F=1} \to X|_{F=1} \to \Pi|_{F=1}$ 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 $\beta(\tilde{\mu}_0, \tilde{\mu}_1) \le n\delta^2/\sigma^2$. Also note that by the choice of τ and the fact that $\delta \le O(\sigma/\sqrt{n})$, we have that for any $z \in \mathcal{B}$, $\tilde{\mu}_0(z) \le O(1) \cdot \tilde{\mu}_1(z)$.

Therefore we are ready to apply Theorem 3.1 and conclude that

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

Note that Π is independent with F conditioned on X and $V = 0$. Therefore we have that

$$\mathrm{I}(X; \Pi \mid V = 0) \ge \mathrm{I}(X; \Pi \mid F, V = 0) \ge \mathrm{I}(X; \Pi | F = 1, V = 0) \Pr[F = 1 \mid V = 0] = \Omega(\frac{\sigma^2}{n\delta^2}).$$

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

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

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 Π' for d-dimensional k-sparse estimation problem with information cost I and loss R, we can construct a protocol Π' for the detection problem with information cost roughly I/d and loss R/k. The protocol Π' 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 Π 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 Π' has small information cost, but only able to show that Π' has a small minimum information cost [7]. 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 $\Omega_{d,k,\delta} = \{\theta : \theta \in \{0,\delta\}^d, |\theta|_0 \le k\}$ 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 $d \ge 2k$, 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 $R \le \frac{1}{16}k\delta^2$, 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 $d \ge 2k$. Any protocol estimator pair $(\Pi, \hat{v})$ that solves the k-sparse Gaussian mean problem* SGME(n, m, d, k, σ^2) *with mean-squared loss R and information cost I and communication cost C satisfy that*

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

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.

Our theorem gives a tight tradeoff between C and R up to logarithmic factor, since it is known [GMN14] that for any communication budget C, there exists protocol which uses C bits and has error $R \le O\left(\min\left\{\frac{\sigma^2 k}{n}, \max\left\{\frac{\sigma^2 dk}{nC}, \frac{\sigma^2 k}{nm}\right\}\right\} \cdot \log d\right)$.

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 $R \le \frac{1}{16}\frac{k\sigma^2}{n}$ then we are done. Otherwise, let $\delta := \sqrt{16R/k} \le \sigma/\sqrt{n}$. 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 sparse Gaussian mean estimation problem SGME(n, m, d, k, σ^2), and $T(n, m, \mathcal{P})$ is the distributed Gaussian mean detection problem

[7] This might be inevitable because protocol Π might reveal a lot information for the nonzero coordinate of θ but since there are very few non-zeros, the total information revealed is still not too much.

$\mathrm{GD}(n, m, \delta, \sigma^2)$. Therefore, by Theorem 4.4, there exists $(\Pi', \hat{v}')$ that solves $\mathrm{GD}(n, m, \delta, \sigma^2)$ 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. □

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 $\mathrm{GME}(n, m, d, \sigma^2)$ 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 B, though we mention a few aspects here. We first give a protocol under the assumption that $|\theta|_\infty \leq \frac{\sigma}{\sqrt{n}}$. 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. In Section B.1 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 η 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 $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with signal strength η 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 $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with signal strength η 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 $\mathrm{SGME}(n, m, d, k, \sigma^2)$ with signal strength η, mean-squared loss R and information cost I (where $\eta^2 \geq 16R/k$), we can find a protocol Π' that solves the

Gaussian mean detection problem 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). Π' would be exactly the same as Protocol 1 but with μ_0 replaced by $\mathcal{N}(0, \sigma^2)$, μ_1 replaced by $\mathcal{N}(\eta, \sigma^2)$ and δ replaced by η. We leave the details to the reader. □

4.3 Lower bound for Sparse Linear Regression

In this section we consider the sparse linear regression problem SLR(n, m, d, k, σ^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 λ 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 $R \cdot C \geq \Omega(\frac{\sigma^2 kd}{\lambda^2 n})$.*

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 $R = O(\frac{k\sigma^2}{nm})$, 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 $d/n \to c$ 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 Section A.

5 Direct-sum Theorem for Sparse Parameters

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

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

1. All machines publicly sample k independent coordinates $I_1, \ldots, I_k \subset [d]$ (without replacement).
2. Each machine j locally prepares data $\widetilde{X}_j = \left(\widetilde{X}_{j,1}, \ldots, \widetilde{X}_{j,d}\right)$ as follows: The I_1-th coordinate is embedded with the true data, $\widetilde{X}_{j,I_1} = X_j$. For $r = 2, \ldots, k$, j-th the machine draws $\widetilde{X}_{j,I_r}$ privately from distribution μ_1^n. For any coordinate $i \in [d]\backslash\{I_1, \ldots, I_k\}$, the j-th machine draws privately $\widetilde{X}_{j,i}$ from the distribution μ_0^n.
3. The machines run protocol Π with input data $\widetilde{X}$.
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 Π' 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 $r = 2, \ldots, k$, and $\theta_i = 0$ for $i \in [d] \backslash \{I_1, \ldots, I_k\}$. We can see that by our construction, the distribution of $\widetilde{X}_j$ is the same as μ_θ^n, and all X_j's are independent. Also note that θ is k-sparse. Therefore when Π' invokes Π on data $\widetilde{X}$, Π will have loss R and information cost I with respect to $\widetilde{X}$.

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

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

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

and when $v = 0$, we have

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

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

Proof. We know that Π has mean-squared loss R, that is,

$$\begin{aligned} R((\Pi, \hat{\theta}), \theta) &= \mathbb{E}\left[||\hat{\theta}(\Pi) - \theta||_2^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{d} |\hat{\theta}(\Pi)_i - \theta_i|^2\right] \end{aligned}$$

Here the expectation is over the randomness of the protocol Π and randomness of the samples $\widetilde{X}_1, \ldots, \widetilde{X}_m$. We first prove equation (22), that is

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

Here the expectation is over $I_1, \ldots, I_k$ in addition to being over the randomness of Π and the samples $\widetilde{X}_1, \ldots, \widetilde{X}_m$. We will in fact prove this claim for any fixing of $I_2, \ldots, I_k$ to some $i_2, \ldots, i_k$. Then I_1 is a random coordinate in $[d] \backslash \{i_2, \ldots, i_k\}$. Then

$$\begin{aligned} \mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2 \mid I_r = i_r, r \geq 2\right] &= \frac{1}{d-k+1} \sum_{i \in [d] \backslash \{i_2, \ldots, i_k\}} \mathbb{E}\left[|\hat{\theta}(\Pi)_i|^2 \mid I_r = i_r, r \geq 2\right] \\ &\leq \frac{1}{d-k+1} \left(\sum_{i \in [d] \backslash \{i_2, \ldots, i_k\}} \mathbb{E}\left[|\hat{\theta}(\Pi)_i|^2 \mid I_r = i_r, r \geq 2\right] \right. \\ &\quad \left. + \sum_{i \in \{i_2, \ldots, i_k\}} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid I_r = i_r, r \geq 2\right] \right) \end{aligned}$$

Taking expectation over $I_2, \ldots, I_r$ we obtain

$$\begin{aligned} \mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1}|^2\right] \leq \frac{1}{d-k+1} \sum_{i=1}^{d} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \theta|^2\right] &= \frac{1}{d-k+1} R((\Pi, \hat{\theta}), \theta) \\ &\leq \frac{R}{d-k+1} \end{aligned}$$

In order to prove equation (21), we prove the statement for every fixing of $\{I_1, \ldots, I_k\}$ to some $S \subset [d]$.

$$\begin{aligned}
&\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1} - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right] \\
&= \frac{1}{k}\sum_{i \in S} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right] \\
&\leq \frac{1}{k}\left(\sum_{i \in S} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right] + \sum_{i \notin S} \mathbb{E}\left[|\hat{\theta}(\Pi)_i|^2 \mid \{I_1, \ldots, I_k\} = S\right]\right) \\
&= \frac{1}{k}\sum_{i=1}^{d} \mathbb{E}\left[|\hat{\theta}(\Pi)_i - \delta|^2 \mid \{I_1, \ldots, I_k\} = S\right]
\end{aligned}$$

Taking expectation over $I_1, \ldots, I_k$ we obtain,

$$\mathbb{E}\left[\mathbb{E}\left[|\hat{\theta}(\Pi)_{I_1} - \delta|^2\right] \mid \{I_1, \ldots, I_k\} = S\right] = \frac{1}{k} R((\Pi, \hat{\theta}), \theta) \leq \frac{R}{k}$$

The last statement of proposition follows easily from Markov's inequality and the assumption that $R \leq k\delta^2/16$. □

Now we prove the information cost of the protocol Π' under the case $v = 0$ is small.

Proposition 5.2. *Under the assumption of Theorem 4.4, we have*

$$\text{min-IC}(\Pi') \leq \mathrm{I}_0(\Pi'; X_1, \ldots, X_m \mid R'_{\text{pub}}) \leq \frac{I}{d - k + 1}$$

where $X_j \sim \mu_0^n$ and R'_{pub} is the public coin used by Π'.

Proof. Let us denote $\left(\widetilde{X}_{j,i}^{(1)}, \ldots, \widetilde{X}_{j,i}^{(n)}\right)$ by $\widetilde{X}_{j,i}$, that is, $\widetilde{X}_{j,i}$ is the collection of i-th coordinates of the samples on machine j. Let R_{pub} be the public coins used by protocol Π. Note that R'_{pub} are just $I_1, \ldots, I_k$ and R_{pub}, therefore, the information cost of Π' is

$$\begin{aligned}
\mathrm{I}_0(\Pi'; X_1, \ldots, X_m \mid R'_{\text{pub}}) &= \mathrm{I}(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, \ldots, I_k, R_{\text{pub}}) \\
&= \mathop{\mathbb{E}}_{i_2, \ldots, i_k}\left[I(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}})\right] \quad (23)
\end{aligned}$$

For each $i_2, \ldots, i_k$, we will prove that $\mathrm{I}(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \leq I/(d - k + 1)$. Note that conditioned on $I_r = i_r$ for $r \geq 2$, I_1 is uniform over $[d] \backslash \{i_2, \ldots, i_k\}$

$$\begin{aligned}
&\mathrm{I}(\Pi; \widetilde{X}_{1,I_1}, \ldots, \widetilde{X}_{m,I_1} | I_1, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \quad (24) \\
&= \frac{1}{d - k + 1} \sum_{i \in [d] \backslash \{i_2, \ldots, i_k\}} \mathrm{I}(\Pi; \widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i} | I_1 = i, I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \\
&= \frac{1}{d - k + 1} \sum_{i \in [d] \backslash \{i_2, \ldots, i_k\}} \mathrm{I}(\Pi; \widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i} | I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \\
&\leq \frac{1}{d - k + 1} \mathrm{I}\left(\Pi; \left(\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i}\right)_{i \in [d] \backslash \{i_2, \ldots, i_k\}} | I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}\right) \\
&\leq \frac{1}{d - k + 1} \mathrm{I}(\Pi; \widetilde{X}_1, \ldots, \widetilde{X}_m | I_2 = i_2, \ldots, I_k = i_k, R_{\text{pub}}) \quad (25)
\end{aligned}$$

The second equality follows from the fact that the distribution of $\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i}$ for $i \in [d] \backslash \{i_2, \ldots, i_k\}$ does not depend on i and the protocol Π is also oblivious of I_1 and hence we can remove the conditioning on $I_1 = i$. First inequality follows from lemma 12 and the fact that $\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{m,i}$ are independent across i. The second inequality follows from the fact that $I(A; B) \leq I(A; B, C)$.

Finally, note that Π performs the task $T(n, m, \mathcal{Q})$ with information cost $I = \sup_\theta \mathrm{I}_\theta(\Pi; \tilde{X} \mid R_{\mathrm{pub}})$. Note that conditioned on $I_r = i_r$ and $I_1 = i$, $\tilde{X}$ are drawn from some valid μ_θ with a k-sparse θ. Therefore by the definition of information cost, we have that

$$\mathrm{I}(\Pi; \widetilde{X}_1, \ldots, \widetilde{X}_m | I_1 = i, I_2 = i_2, \ldots, I_k = i_k, R_{\mathrm{pub}}) \leq I \tag{26}$$

Hence it follows from equations (23) and (25) and (26), we have that

$$\mathrm{I}_0(\Pi'; X_1, \ldots, X_m \mid R'_{\mathrm{pub}}) \leq \frac{I}{d - k + 1} \tag{27}$$

ant it follows by definition that min-IC$(\Pi') \leq \frac{I}{d-k+1}$. □

6 Data Processing Inequality for Truncated Gaussian

In this section, we prove Theorem 4.1, the SDPI for truncated guassian distributions. We first survey the connection between SDPI and transportation inequalities established by Raginsky [Rag14] in Section 6.1. Then we prove in Section 6.2 that when a distribution has log-concave density function on a finite interval, it satisfies the transportation inequalities. These preparations imply straightforwardly Theorem 4.1, which is proved in Section 6.3.

6.1 SDPI Constant and Transportation Inequality

Usually in literature, the inequality (1) is referred to SDPI for mutual information. Here we introduce the more common version of strong data processing inequality, which turns out to be generally equivalent to SDPI for mutual information.

Lemma 3. *Consider the joint distribution of (V, X) where $V \sim B_{1/2}$ and conditioned on $V = v$, we have $X \sim \mu_v$. Note that X is distributed according to the distribution $\mu = (\mu_0 + \mu_1)/2$. By Bayes' rule, we can define the reverse channel $K : X \to V$ with transition probabilities $\{K(v|x) : v \in \{0, 1\}, x \in \mathbb{R}\}$ the same as the conditional probabilities $P_{V|X}$ of the above joint distribution. For any distribution ν over $\mathbb{R}$, let νK denote the distribution of the output v of K if the input x is distributed according to ν. Then*

$$\beta(\mu_0, \mu_1) = \sup_{\nu \neq \mu} \frac{\mathrm{D}_{\mathrm{kl}}(\nu K \| \mu K)}{\mathrm{D}_{\mathrm{kl}}(\nu \| \mu)} \tag{28}$$

Thus, it suffices to bound from above the RHS of (28). We use the technique developed in Theorem 3.7 of [Rag14], which relates the strong data processing inequality with the concentration of measure and specifically the transportation inequality.

To state the transportation inequality, we define the Wasserstein distance $w_1(\cdot, \cdot)$ between two probability measures,

Definition 5. The w_1 distance between two probability measure μ, ν over $\mathbb{R}$ is defined as

$$w_1(\nu, \mu) = \sup_{f : f \text{ is 1-Lipschitz}} \left| \int f d\nu - \int f d\mu \right| \tag{29}$$

We will prove a simple transportation inequality relates the cost of transporting ν to μ in Wasserstein distance w_1 with the KL-divergence between ν and μ,

$$w_1(\nu,\mu)^2 \le \alpha \mathrm{D}_{\mathrm{kl}}(\nu\|\mu). \tag{30}$$

for a certain value of α in section 6.2. For a complete survey of transportation inequalities with other cost functions, please see the survey of Gozlan and Léonard [GL10]. However, before proving the transportation inequality, we show how to use it to derive a bound on $\beta(\mu_0,\mu_1)$.

Lemma 4 (A special case of Theorem 3.7 [Rag14])**.** *Suppose for any $v \in \{0,1\}$, $f_v(x) = \Pr[V = v \mid X = x]$ is L-Lipschitz, and transportation inequality* (30) *is true for $\mu = (\mu_0+\mu_1)/2$ and any measure ν, then*

$$\beta(\mu_0,\mu_1) = \sup_{\nu \ne \mu} \frac{\mathrm{D}_{\mathrm{kl}}(\nu K\|\mu K)}{\mathrm{D}_{\mathrm{kl}}(\nu\|\mu)} \le \alpha L^2 \tag{31}$$

Proof of Lemma 4. We basically follow the proof of Theorem 3.7 of [Rag14] with some simplifications and modifications. Note μK is the unbiased Bernoulli distribution and by the fact that KL divergence is not greater than χ^2 distance, we have

$$\begin{aligned}\mathrm{D}_{\mathrm{kl}}(\nu K\|\mu K) \le \chi^2(\nu K\|\mu K) &= \sum_{v\in\{0,1\}} \frac{(\mu K(v)-\nu K(v))^2}{\mu K(v)} \\ &= 2\sum_{v\in\{0,1\}} (\mu K(v)-\nu K(v))^2 \end{aligned} \tag{32}$$

Fixing any $v \in \{0,1\}$, we have that

$$\begin{aligned}|\mu K(v)-\nu K(v)| &= \left|\int \Pr[V=v\mid X=x]d\mu - \int \Pr[V=v\mid X=x]d\nu\right| \\ &= \left|\int f_v(x)d\mu - \int f_v(x)d\nu\right| \\ &\le L w_1(\nu,\mu) \end{aligned} \tag{33}$$

where the last inequality is by the definition of Wasserstein distance and the fact that $f_v(x)$ is L-Lipschitz.

It follows from (33) and (32) that

$$\mathrm{D}_{\mathrm{kl}}(\nu K\|\mu K) \le L^2 w_1^2(\nu,\mu).$$

Then by transportation inequality (30) we have that

$$\mathrm{D}_{\mathrm{kl}}(\nu K\|\mu K) \le L^2 w_1^2(\nu,\mu) \le \alpha L^2 D(\nu\|\mu).$$

□

6.2 Proving transportation inequality via concentration of measure

In this subsection, we show that if μ is log-concave then it satisfies transportation inequality (30). To obtain the following theorem, we use a series of tools from the theory of concentration of measures in a straightforward way, albeit that in our setting, μ has only support on a finite interval and therefore we need to take some additional care.

Theorem 6.1. *Suppose μ is a measure defined on $[a,b]$ with $d\mu = \exp(-U(x))dx$, and $\nabla^2 u(x) \geq c$, then for any measure ν we have*

$$w_1(\nu,\mu)^2 \leq \frac{2}{c} \cdot \mathrm{D}_{\mathrm{kl}}(\nu\|\mu). \tag{34}$$

In addition, it can be proved by direct calculation that if both μ_0 and μ_1 are log-concave and μ_0 and μ_1 are not too far away in some sense, then $\mu = (\mu_0+\mu_1)/2$ is also log-concave with similar parameters.

Lemma 5. *Suppose distribution μ_0 and μ_1 has supports on $[a,b]$ with $d\mu_0 = \exp(-u_0(x))dx$ and $d\mu_1 = \exp(-u_1(x))dx$. Suppose $\nabla^2 u_0(x) \geq c$, and $\nabla^2 u_0(x) \geq c$, and $|\nabla u_0(x) - \nabla u_1(x)| \leq \sqrt{2c}$ then then $\mu = \frac{1}{2}(\mu_0+\mu_1)$ satisfies that $d\mu = \exp(-u(x))dx$ with $\nabla^2 u(x) \geq \frac{c}{2}$.*

To prove Theorem 6.1, we exploit the well-established connections between transportation inequality, concentration of measure and log-Sobolev inequalities. First of all, transportation inequality (34) with Wasserstein w_1 and KL-divergence ties closely to the concentration of probability measure μ. The theorem of Bobkov-Gotze established the exact connection:

Theorem 6.2 (Bobkov-Gotze [BG99] Theorem 3.1)**.** *Let $\mu \in \mathbb{P}_1$ be a probability measure on a metric space $(\mathbb{X}, d)$. Then the following two are equivalent for $X \sim \mu$.*

1. *$w_1(\nu,\mu) \leq \sqrt{2\sigma^2 \mathrm{D}_{\mathrm{kl}}(\nu\|\mu)}$ for all ν.*
2. *$f(X)$ is σ^2-subgaussian for every 1-Lipschitz function f.*

Using Theorem 6.2, in order to prove Theorem 6.1, it suffices to prove the concentration of measure for $f(X)$ when $X \sim \mu$, and f is 1-Lipschitz. Although one might prove $f(X)$ is subgaussian directly by definition, we use the log-Sobolev inequality to get around the tedious calculation. We begin by defining the entropy of a nonnegative random variable.

Definition 6. The entropy of the a nonnegative random variable Z is defined as

$$\mathrm{Ent}[Z] := \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z] \tag{35}$$

Entropy is very useful for proving concentration of measure. As illustrated in the following lemma, to prove X is subgaussian we only need to bound $\mathrm{Ent}[e^{\lambda X}]$ by $\mathbb{E}[e^{\lambda X}]$.

Lemma 6 (Herbst, c.f. [Led01])**.** *Suppose that for some random variable X, we have*

$$\mathrm{Ent}[e^{\lambda X}] \leq \frac{\lambda^2\sigma^2}{2}\mathbb{E}[e^{\lambda X}], \qquad \text{for all } \lambda \geq 0 \tag{36}$$

Then

$$\psi(\lambda) := \log\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{\lambda^2\sigma^2}{2}, \qquad \text{for all } \lambda \geq 0$$

and as an immediate consequences, X is a σ^2-subgaussian random variable.

Therefore by Theorem 6.2 and Lemma 6, in order to prove transportation inequality, it suffices to to upper bound $\mathrm{Ent}_\mu[e^{\lambda f}]$ by $\mathbb{E}[e^{\lambda f}]$. It turns out that as long as the measure μ is log-concave, we get the concentration inequality for $f(X)$ with 1-Lipschitz function f.

Theorem 6.3 (Theorem 5.2 of [Led01])**.** *Let $d\mu = e^{-U}dx$ where for some $c > 0$, $\nabla^2 U(x) \geq c$ for all $x \in \mathbb{R}$. Then for all smooth function f on $\mathbb{R}$,*

$$\mathrm{Ent}_\mu(f^2) \leq \frac{2}{c}\int |\nabla f|^2 d\mu$$

As a direct corollary, we obtain inequality (36) that we are interested in.

Corollary 6.4. *Let $d\mu = e^{-U}dx$ where for some $c > 0$, $\nabla^2 U(x) \geq c$ for all $x \in \mathbb{R}$. Then for all 1-Lipschitz and smooth function f on $\mathbb{R}$, and any $\lambda \geq 0$, we have*

$$\mathrm{Ent}_\mu(e^{\lambda f}) \leq \frac{\lambda^2}{2c}\,\mathbb{E}[e^{\lambda f}]$$

Proof of Corollary 6.4. Applying directly Theorem 6.3 on $e^{\lambda f/2}$ we obtain,

$$\mathrm{Ent}_\mu[e^{\lambda f}] \leq \frac{2}{c}\int |\nabla e^{\lambda f/2}|^2 d\mu = \frac{2}{c}\int |e^{\lambda f/2}\cdot \lambda \nabla f/2|^2 d\mu$$

Note that if f is 1-Lipschitz, we have $|\nabla e^{\lambda f/2}| \leq |\frac{1}{2}\lambda e^{\lambda f/2}|$, and therefore

$$\mathrm{Ent}_\mu[e^{\lambda f}] \leq \frac{\lambda^2}{2c}\int e^{\lambda f} d\mu = \frac{\lambda^2}{2c}\mathbb{E}_\mu[e^{\lambda f}]$$

□

The distributions that we are interested has continuous density function on a finite support and 0 elsewhere. Therefore we need to use a non-continuous version of the Corollary above to be rigorous.

Corollary 6.5. *Let $S = [a, b]$ be a finite interval in $\mathbb{R}$. Let $d\mu = e^{-U}dx$ for $x \in S$ and $d\mu = 0$ for $x \notin S$. Suppose for some $c > 0$, we have $\nabla^2 U(x) \geq c$ for all $x \in S$.Then the conclusion of Corollary 6.4 is still true.*

Proof of Corollary 6.5. We first extend Theorem 6.3 to the finite support case. Let g be an extension of f to $\mathbb{R}$, such that g is nonnegative and bounded above by some constant C, and ∇g is also bounded by C. Let U_n be a series of extensions of U to $\mathbb{R}$ such that the following happens: a) U_n is twice-differentiable b) $\nabla^2 U_n(x) \geq c$ for all $x \in \mathbb{R}$ c) $\mu_n = e^{-U_n}dx$ approaches to μ in TV norm as n tends to infinity. (The following choice will work for example, $U_n(x) = U(x) + \mathbf{1}_{x>b}\cdot\big(\nabla U(b)(x-b) + \nabla^2 U(b)(x-b)^2 + \exp(n(x-b)^4)\big) + \mathbf{1}_{x<a}\cdot\big(\nabla U(b)(x-a) + \nabla^2 U(b)(x-a)^2 + \exp(n(x-a)^4)\big).$)

Since g and ∇g are bounded, we have that $|\,\mathbb{E}_{\mu_n}(g^2) - \mathbb{E}_\mu(g^2)| = \int g^2(d\mu_n - d\mu) \leq C^2\|\mu_n - \mu\|_{\mathrm{TV}} \to 0$ as n tends to infinity. Similarly we have that $\mathrm{Ent}_{\mu_n}(g^2) \to \mathrm{Ent}_\mu(g^2)$ and $\mathbb{E}_{\mu_n}[|\nabla g|^2] \to \mathbb{E}_\mu[|\nabla g|^2]$. Note that under μ, g agrees with f and therefore we have that $\mathrm{Ent}_{\mu_n}(g^2) \to \mathrm{Ent}_\mu(f^2)$ and $\mathbb{E}_{\mu_n}[|\nabla g|^2] \to \mathbb{E}_\mu[|\nabla f|^2]$.

Also note that μ_n satisfies the condition of Theorem 6.3, therefore

$$\mathrm{Ent}_{\mu_n}(g^2) \leq \frac{2}{c}\int |\nabla g|^2 d\mu_n$$

and the desired result follows by taking n to infinity. □

Finally we provide the proof of Lemma 5, which is obtained by direct calculation of the second derivatives of $u(x)$.

6.3 SDPI for truncated Gaussian

We first check that the Lipschitz constants for $f_v(x) = \Pr[V = 0 \mid X = x]$ as defined in Lemma 4. The proof of the following lemma is deferred to Section D.3.

Lemma 7. *When X is generated by $X \sim \mu_v$ conditioned on $V = v$, let $f_v(x) = \Pr[V = 0 \mid X = x]$, we have that $f_v(x)$ is $\mu/4\sigma^2$-Lipschitz for any $v \in \{0, 1\}$.*

We first prove Theorem 4.1 using Lemma 5, Theorem 6.1 and Lemma 4.

Proof of Theorem 4.1. Note that by definition on support $[-\tau, \tau]$, $d\mu_0' = \gamma_0 \exp(-u_0(x))dx$, and $\mathrm{m}_0' = \gamma_1 \exp(-u_0(x))dx$ with $u_0(x) = -\frac{x^2}{2\sigma^2}$ and $u_1(x) = -\frac{(x-\delta)^2}{2\sigma^2}$. By Lemma 5, we have that $\mu = (\mu_0' + \mu_1')/2$ is $1/\sigma^2$-log concave, and therefore by Theorem 6.1, we have

$$w_1(\nu, \mu)^2 \le 2\sigma^2 \cdot \mathrm{D}_{\mathrm{kl}}(\nu \| \mu).$$

By Lemma 7, we have that f_v's are $\delta/4\sigma^2$-Lipschitz and therefore by Lemma 4, we have that

$$\beta(\mu_0, \mu_1) \le \delta^2/\sigma^2$$

□

Then we present the proof of Corollary 4.2, which relies on the following observation, whose proof is given in Section D.2.

Lemma 8. *Suppose $V \to (X_1, \ldots, X_n) \to \Pi$ forms a Markov Chain, where conditioned on $V = v$, $(X_1, \ldots, X_n)$ are distributed according to $\tilde{\mu}_v$. Then $V \to X_1 + \cdots + X_n \to (X_1, \ldots, X_n) \to \Pi$ also forms a Markov Chain.*

Now we are ready to prove Corollary 4.2.

Proof. (Of corollary 4.2) Let us restate what we want to prove. Suppose $V \sim B_{1/2}$, $(X_1, \ldots, X_n)|V = 0 \sim \tilde{\mu}_0$ and $(X_1, \ldots, X_n)|V = 1 \sim \tilde{\mu}_1$ and $V \to (X_1, \ldots, X_n) \to \Pi$ be a Markov chain. Then

$$I(\Pi; V) \le \frac{n\delta^2}{\sigma^2} I(\Pi; X_1, \ldots, X_n)$$

By lemma 8, $V \to X_1 + \cdots + X_n \to (X_1, \ldots, X_n) \to \Pi$ also forms a Markov chain. Then

$$I(\Pi; V) \le \frac{n\delta^2}{\sigma^2} I(\Pi; X_1 + \cdots + X_n) \le \frac{n\delta^2}{\sigma^2} I(\Pi; X_1, \ldots, X_n)$$

where the first inequality follows from Theorem 4.1 and the fact that the distribution of $X_1 + \cdots + X_n|V = 0$ is the Gaussian $\mathcal{N}(0, n\sigma^2)$ truncated to $[-\tau, \tau]$ and the distribution of $X_1 + \cdots + X_n|V = 1$ is the Gaussian $\mathcal{N}(n\delta, n\sigma^2)$ truncated to $[-\tau, \tau]$. The second inequality follows from data processing. □

Acknowledgments. We thank Yuchen Zhang for suggesting to us the version of 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.

References

[AG76] R. Ahlswede and P. Gacs. Spreading of sets in product spaces and hypercontraction of the markov operator. *Annals of Probability*, 4:925–939, 1976.

[BG99] Sergej G Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[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.

[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.

[CR11] Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. *STOC*, 2011.

[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.

[GL10] Nathael Gozlan and Christian Léonard. Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*, 2010.

[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.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[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 $Φ$-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.

[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.

A Proofs of Results in Section 4

In this section, we prove Theorem 4.3 and Corollary 4.8.

Proof of Corollary 4.8. Suppose there exists such a protocol with mean-squared loss R and communication cost C for sparse linear regression problem $\mathrm{SLR}(n, m, k, d, \sigma^2)$. We are going to use it to solve the sparse linear regression problem $\mathrm{SGME}(m, 1, d, k, \sigma_0)$ as follows. Suppose the i^{th} machine has data $X_i \sim \mathcal{N}(\theta, \sigma_0^2 I_{d\times d})$ with $\sigma_0 = \frac{\sigma}{\lambda\sqrt{n}}$. Then the machines can prepare

$$y_{S_i} = A_{S_i} X_i + b_i$$

where $b_i \sim \mathcal{N}(0, \sigma^2 I - \sigma_0^2 A_{S_i} A_{S_i}^T)$. Note that by the bound $\|A_{S_i}\| \leq \lambda/\sqrt{n}$, we have that $\sigma^2 I - \sigma_0^2 A_{S_i} A_{S_i}^T$ is positive semidefinite. Note that then y_{S_i} can written in the form

$$y_{S_i} = A_{S_i}\theta + \xi_i$$

where ξ_i's are independent distributed according to $\mathcal{N}(0, \sigma^2 I_{n\times n})$

Then the machines call the protocol for the sparse linear regression problem with data (y_{S_i}, A_{S_i}). Therefore we obtain a protocol that solves $\mathrm{SGME}(m, 1, d, k, \sigma_0)$ with communication R and C. Then by Theorem 4.5, we know that

$$R \cdot C \geq \Omega(\sigma_0^2 kd) = \Omega(\frac{\sigma^2 kd}{\lambda^2 n})$$

□

B 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 $\mathrm{GME}(n, m, d, \sigma^2)$ with the assumption that $|\theta|_\infty \le \frac{\sigma}{\sqrt{n}}$.

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 αm machines, and let the other machines be idle.

Unknown parameter $\theta \in [-\sigma/\sqrt{n}, \sigma/\sqrt{n}]$
Inputs: Machine i gets n samples $(X_i^{(1)}, \ldots, X_i^{(n)})$ 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

$$B_i = \begin{cases} 1 & \text{if } X_i \ge 0 \\ -1 & \text{otherwise} \end{cases}$$

- Machine 1 computes

$$T = \sqrt{2} \cdot \mathrm{erf}^{-1}\left(\frac{1}{m}\sum_{i=1}^{m} B_i\right)$$

  where 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 B.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 θ, 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,

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

For any $i \in [m]$, $\Pr[0 \le X_i] = \Pr[-X_i \le 0] = \Phi_{-\bar{\theta},1}(0)$, where Φ_{μ,σ^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

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

Hence,

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

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

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

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

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

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

Hence we have that

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

□

B.1 Extension to general θ

Now we do not assume that $\theta_\ell \in [-\sigma/\sqrt{n}, \sigma/\sqrt{n}]$ for each dimension $\ell \in [d]$, and still show how to achieve a 1-round protocol with $O(md)$ bits of communication, up to low order terms. We will make the simplifying and standard assumptions though, that $|\theta_\ell| \le U = \operatorname{poly}(md)$ for each $\ell \in [d]$, as well as $\log(mdn/\sigma) = o(m)$ and $mdn/\sigma \ge (mdn)^c$ for a constant $c > 0$.

The protocol. As before, it suffices to consider a one-dimensional problem. Protocol 3 solves the one-dimensional Gaussian mean estimation problem using $O(m + \log^2(mdn/\sigma))$ bits of communication. To solve the d-dimensional problem, we run the protocol independently on each coordinate. The total communication will be $O(md + d\log^2(mdn/\sigma))$ bits. We fix $\ell \in [d]$ and let $\theta = \theta_\ell$. Let $\bar{\theta} = \theta\sqrt{n}/\sigma$, where now we no longer assume $\bar{\theta} \le 1$. We will show the output $\hat{\theta}$ satisfies:

$$\mathbb{E}[|\hat{\theta} - \bar{\theta}|^2] = O\left(\frac{1}{m}\right),$$

Unknown parameter θ

Inputs: Machine i gets n samples $(X_i^{(1)}, \ldots, X_i^{(n)})$ 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. If $i \leq r = O(\log(mdn/\sigma))$, machine i sends its first $O(\log(mdn/\sigma))$ bits of X_i to the coordinator (Machine 1)
  3. Else if $i > r$, machine i
     (a) Computes $R_i = X_i - \lfloor X_i \rfloor$, $R_i' = X_i + 1/5 - \lfloor X_i + 1/5 \rfloor$
     (b) Sends B_i and B_i'

$$B_i = \begin{cases} 1 & \text{with probability } R_i \\ 0 & \text{with probability } 1 - R_i \end{cases}$$

$$B_i' = \begin{cases} 1 & \text{with probability } R_i' \\ 0 & \text{with probability } 1 - R_i' \end{cases}$$

- Machine 1
  1. Computes an estimate $\gamma = \frac{\sqrt{n}}{\sigma}$ times the median of X_i's sent by the first r machines.
  2. Computes

$$T = \frac{1}{m-r} \sum_{i=r+1}^{m} B_i, T' = \frac{1}{m-r} \sum_{i=r+1}^{m} B_i'$$

  3. Returns $\frac{\sigma}{\sqrt{n}}\hat{\theta}$ where $\hat{\theta}$ is a multiple of $1/\sqrt{m-r}$ satisfying $|\gamma - \hat{\theta}| < 1/100$ and certain agreement conditions with T, T' described in the text.

Protocol 3: A simultaneous algorithm for estimating the mean of a normal distribution in the distributed setting without assuming $|\theta| \leq \sigma/\sqrt{n}$.

from which it follows that

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

We now describe the one-dimensional problem for a given unknown mean $\bar{\theta}$. The first $r = O(\log(mdn/\sigma))$ machines i send the first $O(\log(mdn/\sigma))$ bits of their (averaged) input Gaussians $X_i = \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} X_i^{(j)}$ to the coordinator. Note that the random variables X_i are distributed according to $\mathcal{N}(\bar{\theta}, 1)$.

Since $O(\log(mdn/\sigma))$ bits of each X_i are communicated to the coordinator, since $\bar{\theta} \leq \text{poly}(md) \cdot \sqrt{n}/\sigma$ (here we use our assumption that $|\theta_\ell| \leq \text{poly}(md)$ for each $\ell \in [d]$), and since each X_i has variance 1, it follows by standard Chernoff bounds that the median γ of $X_1, \ldots, X_r$ is within an additive $\frac{1}{100}$ of $\bar{\theta}$ with probability $1 - \frac{1}{(mdn/\sigma)^\alpha}$ for an arbitrarily large constant $\alpha > 0$ depending on the value $r = O(\log(mdn/\sigma))$. We call this event $\mathcal{E}$, so $\Pr[\mathcal{E}] \geq 1 - \frac{1}{(mdn/\sigma)^\alpha}$.

In parallel, machines $r+1, r+2, \ldots, m$ do the following. Let $R_i \in [0, 1)$ be such that $R_i = X_i - \lfloor X_i \rfloor$. Similarly, let $R_i' \in [0, 1)$ be such that $R_i' = X_i + 1/5 - \lfloor X_i + 1/5 \rfloor$.

For $i = r+1, \ldots, m$, the i-th machine sends a bit $B_i \in \{0, 1\}$, where

$$\Pr[B_i = 1] = R_i,$$

and the i-th matchine also sends a bit $B_i' \in \{0,1\}$ where

$$\Pr[B_i' = 1] = R_i'.$$

We describe the output of the coordinator in the proof of correctness below. Observe that the overall communication is $O(m + \log^2(mdn/\sigma))$, as desired.

Correctness. Consider the "sawtooth" wave $f(x)$, which for a parameter L, satisfies $f(x) = x/(2L)$ for $x \in [0, 2L)$, and is periodic with period $2L$. Its Fourier series[8] is given by

$$f(x) = \frac{1}{2} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\sin\left(\frac{k\pi x}{L}\right).$$

We set $L = 1/2$ and note that $f(X_i) = R_i$. Then, for $X \sim N(\bar{\theta}, 1)$, using a standard transformation of the Gaussian distribution,

$$\mathbf{E}[\sin(tX)] = e^{-t^2/2}\sin(t\bar{\theta}),$$

we have

$$\begin{aligned}
\mathbf{E}[B_i] &= \mathbf{E}[R_i] \\
&= \mathbf{E}[f(X_i)] \\
&= \frac{1}{2} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}e^{-(k\pi/L)^2/2}\sin(k\pi\bar{\theta}/L) \\
&= \frac{1}{2} - \frac{1}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}e^{-2k^2\pi^2}\sin(2k\pi\bar{\theta}).
\end{aligned}$$

Let $B = \frac{1}{m}\sum_{i=r+1}^{m} B_i$, so that $\mathbf{E}[B] = \mathbf{E}[B_i]$. Since the B_i are Bernoulli random variables,

$$\mathbf{E}[|B - \mathbf{E}[B]|^2] \leq \frac{1}{m-r} \leq \frac{2}{m}, \tag{37}$$

where the second inequality uses that $r = O(\log(mdn/\sigma))$ is at most $m/2$ under our assumption that $\log(mdn/\sigma) = o(m)$. In an analogous fashion the coordinator computes a B' using the B_i'.

If event $\mathcal{E}$ occurs, then the coordinator knows γ satisfying $|\gamma - \bar{\theta}| < \frac{1}{100}$, and using γ together with B, will output its estimate to $\bar{\theta}$ as follows. Let $\{x\} = x - \lfloor x \rfloor$. The coordinator checks which of the two conditions γ satisfies:

1. $1/50 < \{\gamma\} < 49/50$ and $|\{\gamma\} - 1/4| \geq 3/100$ and $|\{\gamma\} - 3/4| \geq 3/100$
2. $1/50 < \{\gamma + 1/5\} < 49/50$ and $|\{\gamma + 1/5\} - 1/4| \geq 3/100$ and $|\{\gamma + 1/5\} - 3/4| \geq 3/100$.

We note that one of these two conditions must be satisfied. To see this, suppose the first condition is not satisfied. If it is not satisfied because $\{\gamma\} < 1/50$, then $\{\gamma + 1/5\} \in [1/5, 1/5 + 1/50]$, which satisfies the second of the two conditions. If it is not satisfied because $\{\gamma\} > 49/50$, then $\{\gamma + 1/5\} \in [1/5 - 1/50, 1/5]$, which satisfies the second of the two conditions. If the first condition is not satisfied because $\{\gamma\} \in [1/4 - 1/50, 1/4 + 1/50]$, then $\{\gamma + 1/5\} \in [9/20 - 1/50, 9/20 + 1/50]$ and the second condition is satisfied. If the first condition is not satisfied because $\{\gamma\} \in [3/4 - 1/50, 3/4 + 1/50]$, then $\{\gamma + 1/5\} \in [19/20 - 1/50, 19/20 + 1/50]$, which satisfies the second condition.

If the first condition holds, the coordinator will use B and estimate $\bar{\theta}$ below, otherwise it will use B' and estimate $\bar{\theta} + 1/5$ below. We will analyze the first case; the second case is analogous.

[8] See, e.g., `http://mathworld.wolfram.com/FourierSeriesSawtoothWave.html`

Note that since $\{\gamma\} > 1/50$, and $|\gamma - \bar{\theta}| < \frac{1}{100}$, the coordinator learns $Z = \lfloor \bar{\theta} \rfloor$. Its estimate $\hat{\theta}$ for $\bar{\theta}$ is then $Z + g(B)$, for a function $g(B)$ to be specified (in the other case the coordinator would have learned $\{\bar{\theta} + 1/5\}$ and $\hat{\theta}$ would have been $\{\bar{\theta} + 1/5\} + g(B') - 1/5$).

To define $g(B)$, we need the following claim. Note that in the first case $|\{\gamma\} - 1/4| \geq 3/100$ and so by the triangle inequality $|\{\bar{\theta}\} - 1/4| \geq 3/100 - \gamma = 1/50$. Similarly, $|\{\bar{\theta}\} - 3/4| \geq 1/50$, so the conditions of the following claim hold for $\{\bar{\theta}\}$.

Claim 1. *Define $h(x) = \sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi x)$. There exists a constant $C > 0$ with the following guarantee. If $|\{\bar{\theta}\} - 1/4| \geq 1/50$ and $|\{\bar{\theta}\} - 3/4| \geq 1/50$ then for any number $x \in [\{\bar{\theta}\} - 1/100, \{\bar{\theta}\} + 1/100]$,*

$$C \leq h'(x) \leq 1.$$

Before proving the claim, we conclude the correctness proof. The coordinator guesses $\frac{i}{\sqrt{m}}$ for each integer i for which $|Z + \frac{i}{\sqrt{m}} - \gamma| < \frac{1}{100}$. For each guess $\frac{i}{\sqrt{m}}$, the coordinator checks if

$$|\sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi \frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - B)| \leq \frac{1}{\sqrt{m}} \tag{38}$$

Note that, since the above Fourier series is periodic between succesive integers, we need not add Z to $\frac{i}{\sqrt{m}}$ in (38). Let $g(B)$ be the first guess which passes the check. The coordinator outputs $\hat{\theta} = Z + g(B)$ as its estimate to $\bar{\theta}$ (the second case is analogous, in which Z corresponds to $\lfloor \bar{\theta} + 1/5 \rfloor$ and $g(B')$ is defined in the same way). If there is no such $g(B)$ the coordinator just outputs γ. Note also that if its output ever exceeds our assumed upper bound $U = \text{poly}(mnd/\sigma)$ on the magnitude of $\bar{\theta}$, then we instead output U.

Then

$$\begin{aligned}
\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2] &= \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}] \Pr[\mathcal{E}] + \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \neg\mathcal{E}] \Pr[\neg\mathcal{E}] \\
&= \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}](1 - \frac{1}{(mdn/\sigma)^{\alpha}}) + 4U^2 \cdot \frac{1}{(nmd/\sigma)^{\alpha}} \\
&\leq \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}](1 - \frac{1}{(mdn)^{c\alpha}}) + 4U^2 \cdot \frac{1}{(mdn)^{c\alpha}} \\
&\leq \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}] + \frac{1}{m},
\end{aligned} \tag{39}$$

where the first inequality uses our assumption that $(mdn/\sigma) \geq (mdn)^c$ for a constant $c > 0$, and the second inequality holds for a sufficiently large constant $\alpha > 0$.

Conditioned on $\mathcal{E}$, we have $\hat{\theta} - \bar{\theta} = g(B) - \{\theta\}$. If (38) holds for a given $\frac{i}{\sqrt{m}}$, then

$$|\sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi \frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - B)| \leq \frac{1}{\sqrt{m}}.$$

Let $\mathcal{F}$ be the event that the coordinator finds such an $\frac{i}{\sqrt{m}}$ for which (38) holds. We use the

shorthand $h(z)$ to denote $\sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi z)$.

$$\begin{aligned}
\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E} \wedge \mathcal{F}] &= \mathbf{E}[|\frac{i}{\sqrt{m}} - \{\bar{\theta}\}|^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[|h(\frac{i}{\sqrt{m}}) - h(\{\bar{\theta}\})|^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[(|h(\frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - B)| + |\pi(\frac{1}{2} - B) - h(\{\bar{\theta}\})|)^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[(\frac{1}{\sqrt{m}} + |\pi(\frac{1}{2} - B) - \pi(\frac{1}{2} - \mathbf{E}[B])|)^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \mathbf{E}[(\frac{1}{\sqrt{m}} + \pi|B - \mathbf{E}[B]|)^2 \mid \mathcal{E} \wedge \mathcal{F}] \\
&\leq \frac{2}{m} + 2\pi^2 \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \mathcal{F}]
\end{aligned}$$

where the first equality follows from $\hat{\theta} - \bar{\theta} = g(B) - \{\theta\}$, the first inequality uses the fact that the algorithm ensures $|\frac{i}{\sqrt{m}} - \{\bar{\theta}\}| \leq \frac{1}{100}$ given that $\mathcal{E}$ occurs and therefore one can apply Claim 1 with $x = \frac{i}{\sqrt{m}}$ to conclude that $|h(\frac{i}{\sqrt{m}}) - h(\{\bar{\theta}\})| \leq |\frac{i}{\sqrt{m}} - \{\bar{\theta}\}|$, the second inequality is the triangle inequality, the third inequality uses the guarantee on the value $\frac{i}{\sqrt{m}}$ chosen by the coordinator and the definition of $\mathbf{E}[B]$, the fourth inequality rearranges terms, and the fifth inequality uses $(a+b)^2 \leq 2a^2 + 2b^2$.

If there is no value $\frac{i}{\sqrt{m}}$ for which (38) holds, then since $\mathcal{E}$ occurs it means there is no integer multiple of $\frac{1}{\sqrt{m}}$, call it x, with $|x - \{\bar{\theta}\}| \leq \frac{1}{100}$ for which $|h(x) - \pi(\frac{1}{2} - B)| \leq \frac{1}{\sqrt{m}}$. If it were the case that $|\mathbf{E}[B] - B| < \frac{C}{100\pi}$, where $C > 0$ is the constant of Claim 1, then $|\frac{1}{2} - \frac{1}{\pi}h(\bar{\theta}) - B| < \frac{C}{100\pi}$, or equivalently, $|\pi(\frac{1}{2} - B) - h(\bar{\theta})| < \frac{C}{100}$. By Claim 1, though, we can find an x which is an integer multiple of $\frac{1}{\sqrt{m}}$ which is within $\frac{1}{\sqrt{m}}$ of y, where $h(y) = \pi(\frac{1}{2} - B)$. This follows since the derivative on $[\{\bar{\theta}\} - 1/100, \{\bar{\theta}\} + 1/100]$ is at least C. But then $|h(x) - h(y)| \leq |x - y| \leq \frac{1}{\sqrt{m}}$, contradicting that (38) did not hold. It follows that in this case $|\mathbf{E}[B] - B| \geq \frac{C}{100\pi}$. Now in this case, we obtain an additive $\frac{1}{100}$ approximation, and so $|\hat{\theta} - \bar{\theta}|^2 \leq \frac{\pi^2}{C^2}|B - \mathbf{E}[B]|^2$. Hence,

$$\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E} \wedge \neg\mathcal{F}] \leq O(1) \cdot \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \neg\mathcal{F}],$$

and so

$$\begin{aligned}
\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}] &\leq \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}, \mathcal{F}] \Pr[\mathcal{F}] + \mathbf{E}[|\hat{\theta} - \bar{\theta}|^2 \mid \mathcal{E}, \neg\mathcal{F}] \Pr[\neg\mathcal{F}] \\
&\leq \frac{2}{m} + 2\pi^2 \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \mathcal{F}] \Pr[\mathcal{F}] + O(1) \cdot \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E} \wedge \neg\mathcal{F}] \Pr[\neg\mathcal{F}] \\
&\leq O\left(\frac{1}{m}\right) + O(1) \cdot \mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E}] \\
&\leq O\left(\frac{1}{m}\right),
\end{aligned}$$

where the final inequality uses $\mathbf{E}[|B - \mathbf{E}[B]|^2 \mid \mathcal{E}] \leq \frac{\mathbf{E}[|B - \mathbf{E}[B]|^2]}{\Pr[\mathcal{E}]} \leq 2\mathbf{E}[|B - \mathbf{E}[B]|^2]$, and (37).

Combining this with (39) completes the proof that $\mathbf{E}[|\hat{\theta} - \bar{\theta}|^2] = O(1/m)$.

Proof of Claim. We need to understand the derivative, with respect to x, of the function

$$h(x) = \sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi x),$$

which is equal to

$$h'(x) = \sum_{k=1}^{\infty} 2\pi e^{-2k^2\pi^2} \cos(2k\pi x).$$

Note that the function is periodic in x with period 1, so we can restrict to $x \in [0,1)$. Consider $z = 2\pi x$. Suppose first that $|z - \pi/2| > \epsilon$ and $|z - 3\pi/2| > \epsilon$ for a constant $\epsilon > 0$ to be determined. Then,

$$|\cos(2\pi z)| \geq \cos(\pi/2 - \epsilon) = \sin(\epsilon) \geq 2\epsilon/\pi,$$

using that $\cos(\pi/2-\epsilon) = \sin(\epsilon)$ and that $\sin(x)/x \geq 2/\pi$ for $0 < x < \pi/2$. In this case, it follows that

$$|h'(x)| \quad \geq \quad (2\pi)e^{-2\pi^2}2\epsilon/\pi - \sum_{k>1} 2\pi e^{-2k^2\pi^2} \geq 4e^{-2\pi^2}\epsilon - 4\pi e^{-8\pi^2},$$

using that the summation is dominated by a geometric series. Note that this expression is at least $4e^{-2\pi^2}(\epsilon - \pi e^{-6\pi^2})$, and so setting $\epsilon = 2\pi e^{-6\pi^2}$ shows that $|h'(x)| = \Omega(1)$. Notice that x satisfies $|2\pi x - \pi/2| > \epsilon$ provided $|x - 1/4| \geq 1/100 > \epsilon/(2\pi)$ and that x satisfies $|2\pi x - 3\pi/2| > \epsilon$ provided that $|x - 3/4| \geq 1/100 > \epsilon/(2\pi)$. As $|\{\bar{\theta}\} - 1/4| \geq 1/50$ and $|\{\bar{\theta}\} - 3/4| \geq 1/50$, it follows that $x \in [\{\bar{\theta}\} - 1/100, \{\bar{\theta}\} + 1/100]$. Hence, $|h'(x)| = \Omega(1)$ for such x, as desired.

On the other hand, it is clear that $h'(x) \leq 1$, by upper bounding $\cos(2k\pi x)$ by 1 and using a geometric series to bound $h'(x)$. □

C Distributed Gap Majority

Our techniques can also be used to obtain a cleaner proof of the lower bound on the information complexity of distributed gap majority due to Woodruff and Zhang [WZ12]. In this problem, there are k parties/machines and the i^{th} machine receives a bit z_i. The machines communicate via a shared blackboard and their goal is to decide whether $\sum_{i=1}^k z_i \leq k/2 - \sqrt{k}$ or $\sum_{i=1}^k z_i \geq k/2 + \sqrt{k}$. In [WZ12], it was proven that the information complexity of this problem is $\Omega(k)$. We give a different proof using strong data processing inequalities.

The distribution we will consider is the following: let $B \sim B_{1/2}$. Denote $B_{1/2+10/\sqrt{k}}$ by μ_1 and $B_{1/2-10/\sqrt{k}}$ by μ_0. If $B = 1$, sample $Z_1, \ldots, Z_k$ according to μ_1^k. If $B = 0$, sample $Z_1, \ldots, Z_k$ according to μ_0^k.

Theorem C.1. *Suppose π is a k-party protocol (with inputs $Z_1, \ldots, Z_k$) and π solves the gap majority problem (up to some error). Then $I(\Pi; Z_1, \ldots, Z_k | B = 0) \geq \Omega(k)$.*

Π is the random variable for the transcript of the protocol π. The intuition for the proof is pretty simple. It is not hard to verify that since π solves the gap majority problem, it should be able to estimate B as well i.e. $I(\Pi; B) \geq \Omega(1)$. However since each Z_i has only $\Theta(1/k)$ information about B, the protocol needs to gather information about $\Omega(k)$ of the Z_i's. It is satisfying that this intuition can indeed be formalized! Perhaps worth noting that similar intuition can be drawn for the two-party gap hamming distance problem but there we don't have a completely information theoretic proof of the linear lower bound [CR11]. We will be using the strong data processing inequality for the binary symmetric channel first proven by [AG76]. it studies how information decays on a binary symmetric channel. Suppose X be a bit distributed according to $B_{1/2}$. Y be another bit obtained from X by passing it through a binary symmetric channel with error $1/2 - \epsilon$ (i.e. Y remains X w.p. $1/2 + \epsilon$ and gets flipped w.p. $1/2 - \epsilon$). Then for any random variable U s.t. $U - X - Y$ is a Markov chain, $I(U; Y) \leq 4\epsilon^2 I(U; X)$.

Proof. We will denote by $\Pi_{b_1,\ldots,b_k}$ the transcript of the protocol π when the inputs to π are sampled according to $\mu_{b_1} \otimes \mu_{b_2} \otimes \cdots \otimes \mu_{b_k}$. Since $I(\Pi; B) \geq \Omega(1)$, we know that $h^2(\Pi_{0^k}, \Pi_{1^k}) \geq$

$\Omega(1)$. Now

$$I(\Pi; Z_1, \ldots, Z_k | B = 0) \geq \sum_{i=1}^{k} I(\Pi; Z_i | B = 0)$$

Lets denote our distribution of $\Pi, Z_1, \ldots, Z_k, B$ by ρ. We will tweak this distribution a little bit. Take an independent $B' \sim B_{1/2}$. All the variables are distributed the same as ρ except Z_i which is taken to be independently distributed as $\mu_{B'}$. Denote the new distribution as ρ'. It is easy to verify that

$$I(\Pi; Z_i | B = 0)_{\rho} \geq I(\Pi; Z_i | B = 0)_{\rho'}/2$$

This is true since in ρ, conditioned on $B = 0$, Z_i has the distribution $B_{1/2-10/sqrtk}$ and in ρ' it is $B_{1/2}$ (and hence use Lemma 11). We can also see that

$$\begin{aligned} I(\Pi; Z_i | B = 0)_{\rho'} &\geq \Omega\left(k \cdot I(\Pi; B' | B = 0)_{\rho'}\right) \\ &\geq \Omega\left(k \cdot h^2(\Pi_{e_i}, \Pi_{0^k})\right) \end{aligned}$$

The first inequality is by strong data processing inequality for the binary symmetric channel and the second by Lemma 10. Now

$$\begin{aligned} I(\Pi; Z_1, \ldots, Z_k | B = 0) &\geq \sum_{i=1}^{k} I(\Pi; Z_i | B = 0) \\ &\geq \sum_{i=1}^{k} \Omega\left(k \cdot h^2(\Pi_{e_i}, \Pi_{0^k})\right) \\ &\geq \Omega\left(k \cdot h^2(\Pi_{0^k}, \Pi_{1^k})\right) \\ &\geq \Omega(k) \end{aligned}$$

The third inequality is by noting that $\Pi_{b_1,\ldots,b_k}$ satisfies a cut-and-paste property because π is a k-party protocol and hence Theorem E.1 applies. □

D Missing Proofs in Section 6

D.1 Proof of Lemma 5

Proof of Lemma 5. Let $u(x)$ be such that $d\mu = \exp(-u(x))dx$, that is, $u(x) = -\ln(\frac{1}{2}(\exp(-u_0(x)) + \exp(-u_1(x))))$. We calculate $u''(x)$ as follows:

We can simply calculate the derivatives of u. For simplicity of notation, let $h = \exp(-u_0(x)) + \exp(-u_1(x))$. We have that

$$h' = -u_0' \exp(-u_0) - u_1 u_1' \exp(-u_1),$$

and

$$h'' = (u_0'^2 - u_0'') \exp(-u_0) + (u_1'^2 - u_1'') \exp(-u_1).$$

Therefore we have

$$\begin{aligned} u'' &= \frac{-hh'' + h'^2}{h^2} \\ &= \frac{u_0'' \exp(-2u_0) + u_1'' \exp(-2u_1) + (u_0'' + u_1'' - (u_0' - u_1')^2) \exp(-u_1 - u_2)}{((u_0'^2 - u_0'') \exp(-u_0) + (u_1'^2 - u_1'') \exp(-u_1))^2} \end{aligned}$$

With some simple algebraic manipulations we have that $h'' \geq t$ (for $t \leq \min\{\mu_0'', \mu_1''\}$) is equivalent to

$$\left(\sqrt{\mu_0'' - t}\exp(-u_0) - \sqrt{\mu_1'' - t}\exp(-u_1)\right)^2 + \left(\left(\sqrt{\mu_0'' - t} + \sqrt{\mu_1'' - t}\right)^2 - (u_0' + u_1')^2\right)\exp(-u_0 - u_1) \geq 0$$

Therefore, taking $t = \frac{1}{2c}$ and under our assumptions that $|\mu_0'(x) - \mu_1'(x)| \leq \sqrt{2c}$ for any $x \in [a, b]$, we have that $u'' \geq \frac{c}{2}$ as desired. □

D.2 Proof of Lemma 8

Let us look at the density of $(X_1, \ldots, X_n)$ conditioned on $X_1 + \cdots + X_n = l \leq \tau$ and $V = v$. Suppose $x_1, \cdots, x_n$ be such that $\sum_i x_i = l$, then for some normalizing constant C

$$\begin{aligned} p(x_1, \cdots, x_n | l, v) &= C\frac{e^{-(x_1 - v\delta)^2/2\sigma^2} \cdots e^{-(x_n - v\delta)^2/2\sigma^2}}{e^{-(l - nv\delta)^2/2n\sigma^2}} \\ &= Ce^{(l-nv\delta)^2/2n\sigma^2 - \sum_i (x_i - v\delta)^2/2\sigma^2} \\ &= Ce^{\frac{(l-nv\delta)^2 - n\sum_i (x_i - v\delta)^2}{2n\sigma^2}} \\ &= Ce^{\frac{l^2 - n\sum_i x_i^2}{2n\sigma^2}} \end{aligned}$$

which is independent of v and that proves the lemma. Note that we used the fact that $\sum_i x_i = l$ to simplify the expression.

D.3 Proof of Lemma 7

The proof is by direct calculation. Note that by definition on support $[-\tau, \tau]$, $d\mu_0' = \gamma_0 \exp(-u_0(x))dx$, and $\mu_0' = \gamma_1 \exp(-u_0(x))dx$ with $u_0(x) = -\frac{x^2}{2\sigma^2}$ and $u_1(x) = -\frac{(x-\delta)^2}{2\sigma^2}$, where γ_0 and γ_1 are scaling constants. Note that by the definition of the reverse channel K,

$$f_0(x) = \Pr[V = 0 \mid X = x] = \frac{\gamma_0 e^{-\frac{x^2}{2\sigma^2}}}{\gamma_0 e^{-\frac{x^2}{2\sigma^2}} + \gamma_1 e^{-\frac{(x-\delta)^2}{2\sigma^2}}}$$

Therefore

$$f_0'(x) = \left(\gamma_0 + \gamma_1 \exp(\frac{2x\delta - \delta^2}{2\sigma^2})\right)^{-2} \cdot \gamma_0\gamma_1 \frac{\delta}{\sigma^2} \exp(\frac{2x\delta - \delta^2}{2\sigma^2})$$

By AM-GM inequality we have

$$f_0'(x) \leq \left(4\gamma_0\gamma_1 \exp(\frac{2x\delta - \delta^2}{2\sigma^2})\right)^{-1} \cdot \gamma_0\gamma_1 \frac{u}{\sigma^2} \exp(\frac{2x\delta - \delta^2}{2\sigma^2}) = \frac{4\delta}{\sigma^2}$$

Similarly for $f_1(v)$ we have

$$f_1(x) = \frac{\gamma_1 e^{-\frac{(x-\delta)^2}{2\sigma^2}}}{\gamma_0 e^{-\frac{x^2}{2\sigma^2}} + \gamma_1 e^{-\frac{(x-\delta)^2}{2\sigma^2}}}$$

and

$$f_1'(x) = \left(\gamma_1 + \gamma_0 \exp(\frac{-2x\delta + \delta^2}{2\sigma^2})\right)^{-2} \cdot \gamma_0\gamma_1 \frac{-\delta}{\sigma^2} \exp(\frac{-2x\delta + \delta^2}{2\sigma^2}) \geq \frac{-\delta}{4\sigma^2}$$

Also note that $f_0' \geq 0$ and $f_1' \leq 0$. Therefore for any v, f_v' is $\frac{\delta}{4\sigma^2}$-Lipschitz

E Toolbox

Lemma 9 (Folklore, Hellinger v.s. total variation)**.** *For any two distribution P, Q, we have*

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

Lemma 10. *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,*

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

Proof. The lower bound of the mutual information follows from Lemma 6.2 of [BJKS04]. For the upper bound, we assume that for simplicity ϕ has discrete support $\mathcal{X}$, though the proof extends continuous random variable directly. We have

$$\begin{aligned}
\mathrm{I}(Z;\phi(Z)) &= \frac{1}{2}\mathrm{D}_{\mathrm{kl}}(\phi_1 \| (\phi_1+\phi_2)/2) + \frac{1}{2}\mathrm{D}_{\mathrm{kl}}(\phi_2 \| (\phi_1+\phi_2)/2) \\
&\le \frac{1}{2}\chi^2(\phi_1 \| (\phi_1+\phi_2)/2) + \frac{1}{2}\chi^2(\phi_2 \| (\phi_1+\phi_2)/2) \\
&= \frac{1}{4}\sum_{x\in\mathcal{X}} \frac{(\phi_1(x)-\phi_2(x))^2}{\phi_1(x)+\phi_2(x)} + \frac{1}{4}\sum_{x\in\mathcal{X}} \frac{(\phi_1(x)-\phi_2(x))^2}{\phi_1(x)+\phi_2(x)} \\
&\le \sum_{x\in\mathcal{X}} \frac{(\phi_1(x)-\phi_2(x))^2}{(\sqrt{\phi_1(x)}+\sqrt{\phi_2(x)})^2} \\
&= 2\mathrm{h}^2(\phi_1,\phi_2)
\end{aligned}$$

where the first inequality uses that KL-divergence is less than χ^2 distance and the second one uses the inequality $a^2+b^2 \ge \frac{(a+b)^2}{2}$. □

Theorem E.1 (Corollary of Theorem 7 of [Jay09])**.** *Suppose a family of distribution $\{P_{\boldsymbol{b}} : \boldsymbol{b} \in \{0,1\}^m\}$ satisfies the cut-paste property: for any for any $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}, \boldsymbol{d}$ with $\{a_i, b_i\} = \{c_i, d_i\}$ (in a multi-set sense) for every $i \in [m]$, $\mathrm{h}^2(\Pi_{\boldsymbol{a}}, \Pi_{\boldsymbol{b}}) = \mathrm{h}^2(\Pi_{\boldsymbol{c}}, \Pi_{\boldsymbol{d}})$. Then we have*

$$\sum_{i=1}^{m} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_i}}) \ge \Omega(1) \cdot \mathrm{h}^2(P_{\mathbf{0}}, P_{\mathbf{1}}) \tag{40}$$

where $\mathbf{0}$ and $\mathbf{1}$ are all 0's and all 1's vectors respectively, and $\boldsymbol{e_i}$ is the unit vector that only takes 1 in the ith entry.

Proof. Theorem 7 of [Jay09] already proves a stronger version of this theorem for the $m = 2^t$ case. Suppose on the other hand $m = 2^t + \ell$ for $\ell < 2^t$. We divide $[m] = \{1, \ldots, m\}$ into a collection of 2^t subsets $A_1, \ldots, A_{2^t}$, each of which contains at most 2 elements. Let $\boldsymbol{f_i}$ be the indicator vector of the subset A_i. For example, if $A_i = \{p, q\}$, then $\boldsymbol{f_i} = \boldsymbol{e_p} + \boldsymbol{e_q}$. We claim that $\sum_{j\in A_i} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_j}}) \ge \Omega(1)\mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{f_i}})$. This is trivial when $|A_i| = 1$ and when $A_i = \{p, q\}$, we have that by CauchySchwarz inequality and the cut-paste property

$$\mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_p}}) + \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_q}}) \ge \frac{1}{2}\mathrm{h}^2(P_{\boldsymbol{e_p}}, P_{\boldsymbol{e_q}}) = \frac{1}{2}\mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_q}+\boldsymbol{e_q}}).$$

Therefore, we can lowerbound LHS as

$$\sum_{i=1}^{m} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{e_i}}) \ge \frac{1}{2}\sum_{i=1}^{2^t} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{f_i}}).$$

Then applying Theorem 7 of [Jay09] on the RHS of the inequality above we have

$$\frac{1}{2}\sum_{i=1}^{2^t} \mathrm{h}^2(P_{\mathbf{0}}, P_{\boldsymbol{f_i}}) \geq \Omega(1) \cdot \mathrm{h}^2(P_{\mathbf{0}}, P_{\mathbf{1}}),$$

and the theorem follows. □

Lemma 11. *Suppose two distribution μ, μ' 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*

$$\mathrm{I}(X; \Pi(X)) \geq c \cdot \mathrm{I}(X'; \Pi(X')) \tag{41}$$

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

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

Then note that

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

It follows that

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

□

Lemma 12 (Folklore)**.** *When X is drawn from a product distribution, then*

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