The Role of Interactivity in Structured Estimation

Jayadev Acharya ACHARYA@CORNELL.EDU
Cornell University

Clément L. Canonne CLEMENT.CANONNE@SYDNEY.EDU.AU
University of Sydney

Ziteng Sun ZITENGSUN@GOOGLE.COM
Google Research

Himanshu Tyagi HTYAGI@IISC.AC.IN
Indian Institute of Science

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study high-dimensional sparse estimation under three natural constraints: communication constraints, local privacy constraints, and linear measurements (compressive sensing). Without sparsity assumptions, it has been established that interactivity cannot improve the minimax rates of estimation under these information constraints. The question of whether interactivity helps with natural inference tasks has been a topic of active research. We settle this question in the affirmative for the prototypical problems of high-dimensional sparse mean estimation and compressive sensing, by demonstrating a gap between interactive and noninteractive protocols. We further establish that the gap increases when we have more structured sparsity: for *block sparsity* this gap can be as large as *polynomial* in the dimensionality. Thus, the more structured the sparsity is, the greater is the advantage of interaction. Proving the lower bounds requires a careful breaking of a sum of correlated random variables into independent components using Baranyai's theorem on decomposition of hypergraphs, which might be of independent interest.

Keywords: Mean estimation, parameter estimation, communication constraints, compressive sensing, adaptive sensing, local privacy interactivity, adaptivity, distributed inference

1. Introduction

Estimating high-dimensional parameters is a central task arising in various scientific disciplines and data-driven applications. Modern applications often involve data from distributed or online sources which restrict the mechanism via which we have access to the data; for instance, limitations may be placed due to ease of implementation, or due to stringent communication constraints (bandwidth), or legal constraints (privacy).

Understanding the interplay between these restrictions and the task at hand is the key to designing better and more efficient algorithms for these tasks. In this paper, we make progress on that front, considering the fundamental question of *sparse parameter estimation*.

Sparse mean estimation: Upon observing independent samples $X_1, \ldots, X_n$ from a high-dimensional product distribution over $\{\pm 1\}^d$ with mean vector $\mu \in [-1, 1]^d$, the goal is to output an estimate $\hat{\mu}$ such that

$$\Pr[\|\hat{\mu} - \mu\|_2 > \varepsilon] \leq 1/10, \tag{1}$$

i.e., to achieve good accuracy of estimation under ℓ_2 loss. In addition, we are promised that the unknown parameter μ is s-sparse; namely, it has at most s nonzero coordinates and $\|\mu\|_0 \le s$.

However, the observations are subject to an ℓ-bit *communication constraint*, where $1 \le \ell \le d$. Namely, each X_t must be compressed to an ℓ-bit message Y_t, and the estimate $\hat{\mu}$ is then computed from the n messages $Y_1, \ldots, Y_n \in \{0,1\}^\ell$ only. Our results also apply to *local differential privacy* (LDP) constraints (Dwork et al., 2006; Kasiviswanathan et al., 2011), where each message is required not to reveal too much about the observation; we relegate the details to the appendix.

Block-sparse mean estimation: The task is very similar, but the sparsity structure imposed on μ is now more restrictive. Specifically, we are promised that the (at most) s nonzero coordinates of the unknown parameter μ are contiguous:

$$\exists 1 \le i \le d - s : \forall j \notin \{i, i+1, \ldots, i+s\},\ \mu(j) = 0. \tag{2}$$

Compressive sensing: There is an unknown s-sparse vector $x \in \mathbb{R}^d$, which can only be observed through noisy linear measurements given by

$$Y_t := A_t x + Z_t, \qquad 1 \le t \le n, \tag{3}$$

where $Z_1, \ldots, Z_n$ are i.i.d. $\mathcal{N}(0, I_m)$ random variables (noise), and $A_1, \ldots, A_n \in \mathbb{R}^{m \times d}$ are measurement unitary matrices[1] chosen (possibly adaptively) by the protocol. The goal is to estimate x to ℓ_2 loss ε using observations $Y_1, \ldots, Y_n \in \mathbb{R}^m$, minimizing the number $m \cdot n$ of overall measurements. When the matrices A_t are chosen interactively, this is known as *adaptive sensing* (Arias-Castro et al., 2012); specifically, adaptive sensing considers the case $m = 1$ and allows each measurement to be of the form $\langle \mathbf{a}_t, x \rangle + z_t$ for a vector $\mathbf{a}_t$ that is adaptively chosen dependent on $Y_1, \ldots, Y_{t-1}$.

All these tasks have received significant attention in recent years. But the role of *interactivity* in communication protocols is not completely understood. Interactivity allows clients to choose the messaging scheme based on clients' outputs from previous communication rounds. Formally, for (sequentially) *interactive* protocols, the messaging scheme from X_t to Y_t is allowed to be chosen based on previous messages $Y_1, \ldots, Y_{t-1}$ while for *noninteractive* protocols, the mapping from X_t to Y_t is chosen independently without observing others' messages. Although interactivity brings flexibility in the protocol design, it often comes with extra cost. For example, interaction may lead to time delays since each client needs messages from previous clients, which can be prohibitive for large-scale distributed learning systems such as those used for Federated Learning (Kairouz et al., 2021). Despite these overheads, it is not fully understood whether *interactivity* can lead to significant savings.

We make progress in this direction and show that for the three examples above interactivity *does* enable more data-efficient solutions. At a high level, our results can be interpreted as follows:

> Interactivity allows one to leverage the *structure* (sparsity) of the three tasks considered to obtain provably more data-efficient estimation algorithms (in a minimax sense).

1. More generally, this can be relaxed to requiring only that each row vector has bounded ℓ_2 norm.

This is to be put in contrast to two related tasks. First, it has recently been shown that for unstructured estimation tasks, allowing for interactivity does not yield any speedup over noninteractive protocols, or, indeed, even over *private-coin* protocols (where the users do not have access to any common random seed, but instead are fully independent) (Braverman et al., 2016; Han et al., 2018; Acharya et al., 2020a). That is, for unstructured estimation, neither public randomness nor interactivity are useful. Second, for communication (or local privacy) constraints such as the ones considered in this work, a sequence of papers (Acharya et al., 2020b,c, 2022, 2020e) showed that *goodness-of-fit testing* (not estimation) could be more data-efficient when allowing for some coordination between users; however, this gain in efficiency was enabled by the use of a common random seed (i.e., public-coin protocols), a weaker setting than interactivity. Moreover, once this common random seed was available, letting the users interact would not lead to any additional saving: put differently, under those constraints public randomness helps for testing, but interactivity does not.

We state the formal statements of our results in Section 1.1, and then put them in context and discuss prior work in Section 1.2. We discuss details about sparse estimation and block-sparse estimation in Section 2 and Section 3 respectively. We present extensions to adaptive sensing and estimation under local privacy constraints in the appendix.

1.1. Our results and contributions

Our first result concerns the lower bound of noninteractive protocols for sparse mean estimation, which provides a lower that establishes a strict separation between the performance of interactive and noninteractive protocols:

Theorem 1 *For any $s \geq 4 \log d$, any ℓ-bit noninteractive protocol for mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ must have sample complexity $\Omega\Big(\frac{sd}{\varepsilon^2 \ell} \log \frac{ed}{s}\Big)$.*

Combined with previously known results Braverman et al. (2016); Acharya et al. (2020a) for sparse mean estimation (detailed in Section 2), this lower bound immediately implies the following:

Corollary 2 *For any $s \geq 4 \log d$, the* noninteractive *sample complexity of mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $\Theta\Big(\frac{sd}{\varepsilon^2 \ell} \log \frac{ed}{s}\Big)$, while the* interactive *sample complexity is $\Theta\Big(\frac{sd}{\varepsilon^2 \ell} + \frac{s}{\varepsilon^2} \log \frac{ed}{s}\Big)$.*

This shows that interactive protocols outperform the noninteractive ones by a factor up to $\Omega(\log d/s)$. We emphasize that prior to our work this gap was only known for $\varepsilon \ll \sqrt{\ell/d}$, from (Han et al., 2018), even for the case $s = 1$. [2]

Our second set of results focuses on *block sparsity*.

Theorem 3 *For any $s \geq 1$, the* noninteractive *sample complexity of mean estimation of s-block sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $\Theta\Big(\frac{sd + d \log d}{\varepsilon^2 \ell}\Big)$, while the* interactive *sample complexity is $\tilde{O}\Big(\frac{s^2 + d}{\varepsilon^2 \ell} + \frac{s}{\varepsilon^2}\Big)$ and $\Omega\Big(\frac{s^2 + d}{\varepsilon^2 \ell} + \frac{s}{\varepsilon^2}\Big)$.*

2. The result in (Han et al., 2018) holds under the setting where the data comes from Gaussian distributions with identity covariance matrices, which implies the lower bound for product distributions over $\{\pm 1\}^d$. However, the restriction on the lower bound that $\varepsilon \ll \sqrt{\ell/d}$ still holds even if we modify the technique in (Han et al., 2018) to the setting considered in this paper.

Only a restricted version of the upper bound in the interactive case was previously known;[3] our results, by complementing them with the required lower bounds (as well as the noninteractive upper bound) establish that interactivity leads to significant savings under this more structured sparsity constraint. As an example, for $s \approx \sqrt{d}$, the sample complexity for interactive protocols is $\tilde{\Theta}(d/(\varepsilon^2\ell))$ whereas that of noninteractive protocols is $\tilde{\Theta}(d^{3/2}/(\varepsilon^2\ell))$. Interestingly, establishing the lower bound in the noninteractive case (Lemma 14) requires circumventing many technical hurdles, and in particular handling high-order correlations between random variables when trying to bound the expectation of a multivariate polynomial with the information bound of Lemma 7. To achieve this, we carefully decompose the dependency (hyper)graph of the resulting monomials into sums of independent terms, taking recourse to a result of Baranyai (1974) on factorization of hypergraphs (Lemma 17). We believe this strategy to be of independent interest, with applications to other statistical lower bounds in distributed settings.

Finally, our third set of results departs from communication constraints, and instead focuses on the well-studied question of *compressive sensing*. Specifically, as discussed earlier, we consider the problem of estimating (under the ℓ_2 loss) an s-sparse signal, when the only measurements allowed are m-dimensional noisy linear measurements (as defined in Eq. (3)).

Theorem 4 *For any $s \geq 4\log d$, there exists an interactive protocol for compressive sensing for s-sparse vectors using m-dimensional noisy linear measurements with sample complexity $O\Big(\frac{sd}{\varepsilon^2 m} + \frac{s}{\varepsilon^2}\log\frac{ed}{s}\Big)$.*

Combined with known results Raskutti et al. (2011); Arias-Castro et al. (2012); Wu (2020) on compressive sensing (detailed in Appendix C), our upper bound readily implies the following:

Corollary 5 *For any $s \geq 4\log d$, the* noninteractive *sample complexity of compressive sensing for s-sparse vectors using m-dimensional random measurements is $\Theta\Big(\frac{sd}{\varepsilon^2 m}\log\frac{ed}{s}\Big)$, while the* interactive *sample complexity is $\Theta\Big(\frac{sd}{\varepsilon^2 m} + \frac{s}{\varepsilon^2}\log\frac{ed}{s}\Big)$.*

Taken together, our three sets of results show that, across various tasks, interactivity *does* help for estimation under constraints, as soon as sparsity enters the picture. Further, it is not too hard to show that the analogues of Corollary 2 and Theorem 3 hold for *local privacy* constraints as well, replacing ℓ by square of the privacy parameter, which demonstrates corresponding separations under LDP.

Theorem 6 (Local privacy (LDP)) *All the bounds from Corollary 2 and Theorem 3 hold when considering ϱ-LDP constraints instead of ℓ-bit communication constraints, replacing ℓ by ϱ^2 in the corresponding expressions for any value of the privacy parameter $\varrho \in (0, 1]$.*

We provide the necessary definitions and the proof of this theorem in Appendix A.

1.2. Prior and related work

In the recent years, there has been a significant work on both distribution mean estimation and signal estimation under various constraints. We highlight below the most relevant to our work.

Distributed mean estimation under both communication and privacy constraints has been extensively considered (Shamir, 2014; Erlingsson et al., 2014; Duchi et al., 2018; Han et al., 2018; Barnes

3. That is, the existing upper bound worked under an additional promise on the block sparsity, which was that all biased coordinates had the same magnitude (Acharya et al., 2021).

et al., 2020; Braverman et al., 2016; Ye and Barg, 2018; Acharya et al., 2020b, 2019). Most of these results pertain to noninteractive protocols, namely schemes where the measurements/messaging schemes are decided simultaneously, not allowing for dependence on the outcomes from prior symbols. There are some notable exceptions. Braverman et al. (2016); Duchi and Rogers (2019) establish interactive lower bounds for estimating high dimensional distributions under communication and local privacy constraints. Their strong results establish that the minimax rates of interactive and noninteractive schemes are the same. However, these minimax lower bounds are tight only for *dense* distributions. Braverman et al. (2016) considered sparse high-dimensional mean estimation under communication and establish lower bounds for interactive schemes and upper bounds for noninteractive schemes; still, their result leave open the existence of a gap between the two for sparse mean estimation. Similarly, Duchi and Rogers (2019) consider sparse mean estimation under local privacy: their work also leaves unanswered the existence of a gap between the interactive lower bounds and their noninteractive upper bounds. Shamir (2014) consider 1-sparse mean estimation for d-dimensional product distributions, and their bounds also have a similar gap.

Block-sparse signals are common in several applications such as DNA microarrays, sensor networks and MIMO communication systems (Elhamifar and Vidal, 2013; Stojnic et al., 2009; Barbotin et al., 2012; Baron et al., 2009; Gogineni and Nehorai, 2011; Shoukry and Tabuada, 2015; Vorobyov et al., 2004; Baraniuk et al., 2010). Estimating distributions with block-sparse means was considered in Acharya et al. (2021). They study the constraint where one has access to a few coordinates of each sample and showed that for this constraint there is a separation between interactive and noninteractive protocols. This is in the context of first-order optimization, where they used a reduction to this mean estimation problem in order to show that adaptive processing of gradients can lead to faster convergence rates for distributed optimization.

Compressive sensing has been immensely popular since the pioneering works of Candès et al. (2006); Donoho (2006). Adaptive sensing, *i.e.*, choosing the measurements adaptively, was studied in Arias-Castro et al. (2012) for the case $m = 1$. Their results leave open a logarithmic (in the dimension) gap between upper and lower bounds on the number of measurements (sample complexity). For a slightly different problem of exact support recovery, Malloy and Nowak (2014) shows that adaptive sensing can help reduce the minimum required signal level in the support. However, it does not show a separation in terms of sample complexity.

The question of whether interactivity helps under local privacy constraints has been extensively studied, starting with the influential work of Kasiviswanathan et al. (2008), who designed a problem for which there show a separation between interactive and noninteractive schemes. Daniely and Feldman (2019) designed a class of Boolean functions for which learning under interactive LDP protocols is exponentially more expensive than noninteractive schemes. Dagan and Feldman (2020) showed that exponentially more samples are needed to learn linear models with convex loss without interaction than that with, under both privacy and communication constraints. Joseph et al. (2019) went a step further and showed that allowing for fully interactive schemes can provide an advantage over sequentially interactive schemes. Ullman (2018) proves a lower bound for locally private hypothesis selection for noninteractive protocols, which can be viewed as a 1-sparse mean estimation problem. The role of interactivity for the problem of locally private hypothesis selection was discussed in Gopi et al. (2020). But the result doesn't imply strict separation between interactive and non-interactive protocols in the LDP setting.

Another line of work (Agarwal et al., 2017; Jin et al., 2019; Thananjeyan et al., 2021) shows that interactivity brings advantage for the task of best arm identification in multi-armed bandits. The

feedback model in multi-armed bandit can be simulated by a 1-bit communication protocol, hence our result would imply the same separation.

1.3. Notation and Preliminaries.

We use $\log$ and $\ln$ for logarithm in base 2 and natural logarithm respectively. Throughout the paper, we use standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$, with asymptotics to be taken as $d, s \gg 1$ and small ε. Our lower bounds will routinely involve both Kullback–Leibler (KL) and chi-squared (χ^2) divergences between probability distributions, defined as

$$\mathrm{KL}(\mathbf{p} \,||\, \mathbf{q}) := \sum_{x \in \mathcal{X}} \mathbf{p}(x) \ln \frac{\mathbf{p}(x)}{\mathbf{q}(x)}, \qquad \chi^2(\mathbf{p} \,||\, \mathbf{q}) := \sum_{x \in \mathcal{X}} \frac{(\mathbf{p}(x) - \mathbf{q}(x))^2}{\mathbf{q}(x)}$$

for any two distributions $\mathbf{p}, \mathbf{q}$ over a (discrete) domain $\mathcal{X}$, with the convention that $0 \ln 0 = 0$. These divergences satisfy $\mathrm{KL}(\mathbf{p} \,||\, \mathbf{q}) \le \chi^2(\mathbf{p} \,||\, \mathbf{q})$. We will also require the notion of (Shannon) entropy $H(X) = -\sum_{x \in \mathcal{X}} \mathbf{p}_X(x) \log \mathbf{p}_X(x)$ of a random variable X with distribution $\mathbf{p}_X$, as well as that of the mutual information $I(X \wedge Y)$ between two random variables X, Y, defined as

$$I(X \wedge Y) := \mathrm{KL}(\mathbf{p}_{XY} \,||\, \mathbf{p}_X \otimes \mathbf{p}_Y),$$

where $\mathbf{p}_{XY}, \mathbf{p}_X, \mathbf{p}_Y$ are the joint distribution of (X, Y) and the marginal distributions of X and Y, respectively, and $\mathbf{p} \otimes \mathbf{q}$ is the product distribution with marginals $\mathbf{p}, \mathbf{q}$. We will also use the conditional mutual information $I(X \wedge Y \mid Z)$, defined as $I(X \wedge Y \mid Z) := \mathbb{E}_Z\Big[\mathrm{KL}\Big(\mathbf{p}_{XY|Z} \,||\, \mathbf{p}_{X|Z} \otimes \mathbf{p}_{Y|Z}\Big)\Big]$ (where $\mathbf{p}_{XY|Z}, \mathbf{p}_{X|Z}, \mathbf{p}_{Y|Z}$ are now the analogous distributions, conditioned on Z). For more on these notions and their properties, we refer the reader to the textbook by Cover and Thomas (2006).

Throughout the paper, we often use the term *channel* to refer to the probabilistic mapping from the user's observation to messages. Formally, the tth user selects a channel $W_t \colon \mathcal{X} \to \mathcal{Y}$, where, for all input $x \in \mathcal{X}$ and possible output $y \in \mathcal{Y}$,

$$W_t(y \mid x) = \Pr[Y_t = y \mid X_t = x].$$

For instance, by restricting the output space $\mathcal{Y}$ to satisfy $|\mathcal{Y}| \le 2^\ell$, the formulation captures ℓ-bit communication constraints. In *noninteractive* protocols, users must select their channels independently without observing each other's message. In contrast, for (sequentially) *interactive*[4] protocols, the tth user can select their channel based on previous users' messages $Y_1, Y_2, \ldots, Y_{t-1}$. For both interactive and non-interactive protocols considered in this paper, we assume all users and the server have access to a public random seed U, which is independent of the samples.[5]

Our proof will rely on the following information bound adapted from Acharya et al. (2020d).

Lemma 7 *Consider a random variable X taking values in $\mathcal{X}$. Let $\Phi \colon \mathcal{X} \to \mathbb{R}^d$ be such that the random vector $\Phi(X)$ has independent coordinates and is σ^2-subgaussian. Let $W : \mathcal{X} \to \mathcal{Y}$ be an ℓ-bit channel. Then, we have*

$$\sum_{y \in \mathcal{Y}} \frac{\|\mathbb{E}[\Phi(X) W(y \mid X)]\|_2^2}{\mathbb{E}[W(y \mid X)]} \le 2(\ln 2)\sigma^2 \ell.$$

4. The lower bounds in the paper also holds for fully interactive protocols (the so-called *blackboard model*) while the provided upper bounds only require sequential interactivity. We focus on sequentially interactive protocols in this paper for clarity of presentation.
5. For a formal definition, see Acharya et al. (2022).

2. Sparse mean estimation under communication constraints

We first establish Theorem 1, thus establishing the claimed gap between interactive and noninteractive communication-constrained sparse mean estimation.

Of the four ingredients required to prove Theorem 1 (two upper bounds, and two lower bounds), three follow from the literature; we restate them below for completeness. The following statements establish the sample complexity for interactive sparse mean estimation.

Lemma 8 ((Acharya et al., 2020a, Proposition 2)) *For any $s \geq 1$, the* interactive *sample complexity of mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $O\Big(\frac{sd}{\varepsilon^2 \ell} + \frac{s}{\varepsilon^2} \log \frac{ed}{s}\Big)$.*

Lemma 9 (Braverman et al. (2016); Acharya et al. (2020a)) *For any $s \geq 1$, the* interactive *sample complexity of mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $\Omega\Big(\frac{sd}{\varepsilon^2 \ell} + \frac{s}{\varepsilon^2} \log \frac{ed}{s}\Big)$.*

The algorithm achieving Lemma 8 is based on successive elimination and requires interaction between clients. Turning to noninteractive estimation, similar bounds can be obtained, but with an extra logarithmic factor.

Lemma 10 *For any $s \geq 1$, the* noninteractive *sample complexity of mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $O\Big(\frac{sd \log(ed/s)}{\varepsilon^2 \ell}\Big)$.*

Proof The key observation is that, by using the "simulate-and-infer" idea of Acharya et al. (2020c) (specifically used in the context of product distributions over $\{\pm 1\}^d$ in Acharya et al. (2020d)), it suffices to show an $O\Big(\frac{s \log(ed/s)}{\varepsilon^2}\Big)$ upper bound in the *unconstrained* setting (where all the observations are fully available), as any such algorithm can be simulated by a private-coin protocol under ℓ-bit communication constraints at the cost of a factor d/ℓ in the sample complexity. The idea is to partition d coordinates into $\lceil d/\ell \rceil$ blocks of size at most ℓ. Then $\lceil d/\ell \rceil$ users can send their observation within each block using ℓ bits. By independence of the coordinates, we get a valid sample from the source distribution by combining the messages. With samples from the original distribution, the $O\Big(\frac{s \log(ed/s)}{\varepsilon^2}\Big)$ sample complexity upper bound, in turn, is well-known, and is attained by *e.g.*, the maximum likelihood estimator. See, for instance, Wu (2020, Section 20.2). ■

The final component needed to show the additional logarithmic factor is necessary is the noninteractive sample complexity lower bound. As discussed earlier, the required lower bound is shown in Han et al. (2018, Theorem 3), but under the restriction that $n \geq \frac{sd^2 \log(ed/s)}{\ell^2}$, making the lower bound vacuous unless $\varepsilon \ll \sqrt{\ell/d}$. We provide a proof of this lower bound, which removes this restriction on n. The crux in removing this regularity condition is to handle the dependent terms in the obtained information bound (Eq. (5)) directly through careful conditioning; while previous techniques consider linearization of the information vector, which results in loose bounds.

Lemma 11 *For any $s \geq 4 \log d$, the* noninteractive *sample complexity of mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $\Omega\Big(\frac{sd \log(ed/s)}{\varepsilon^2 \ell}\Big)$.*

Proof Consider the following set of s-sparse product distributions, which we will use as the "hard instances" for our lower bound. Setting $\gamma := \frac{\varepsilon}{\sqrt{s}}$, for any $z \in \{-1, 0, +1\}^d$ we define $\theta_z \in \mathbb{R}^d$ by $\theta_{z,i} = \gamma z(i)$, $i \in [d]$. Let Z be a random variable on $\{-1, 0, +1\}^d$ satisfying

$$\Pr[Z(i) = +1] = \frac{s}{4d}, \quad \Pr[Z(i) = -1] = \frac{s}{4d}, \quad \Pr[Z(i) = 0] = 1 - \frac{s}{2d}.$$

Note that θ_Z is then $(s/2)$-sparse in expectation, and further $\mathbb{E}[Z(i)] = 0$, $\sigma^2 := \mathbb{E}\big[Z(i)^2\big] = \frac{s}{2d}$ for all i. By a Chernoff bound, we also get that θ_Z is s-sparse with high probability: if $s \geq 4 \log d$, $\Pr[\|\theta_Z\|_0 \leq s] \geq 1 - \frac{s}{4d}$. This will be enough for our purposes, and allows us to consider the random prior of hard instances above (product distributions over $\{-1, 0, +1\}^d$, with mean θ_Z for random Z with independent coordinates) instead of enforcing s-sparsity with probability one (details follow).

Consider the following generative process. First pick Z at random from $\{-1, 0, +1\}^d$ as above. Then, each of the n users observes one sample X_t from the product distribution $\mathbf{p}_Z$ with mean vector θ_Z and sends its samples through a channel $W_t \colon \{\pm 1\}^d \to \{0, 1\}^\ell$ to compress it to a message Y_t.

The next claim states that any sufficiently accurate estimation protocol must provide enough information about each $Z(i)$ from the tuple of messages Y^n.

Claim 1 (Assouad-type Bound) *For any protocol that estimates s-sparse product distributions to ℓ_2 accuracy $\varepsilon/4$, we must have $\sum_{i=1}^d I(Z(i) \wedge Y^n) = \Omega\big(s \log \frac{ed}{s}\big)$. In particular, by independence of the coordinates of Z, this implies $I(Z \wedge Y^n) = \Omega\big(s \log \frac{ed}{s}\big)$.*

Proof [Proof of Claim 1] Fix any such protocol, and consider the corresponding estimator $\hat{\theta} = \hat{\theta}(Y^n)$. From there, define an estimator $\hat{Z}$ for Z by choosing

$$\hat{Z} = \arg \min_{z \in \{-1,0,+1\}^d} \|\theta_z - \hat{\theta}\|_2 .$$

In particular, $\|\theta_{\hat{Z}} - \theta_Z\|_2 \leq 2\|\hat{\theta} - \theta_Z\|_2$ with probability 1, and

$$\mathbb{E}\Big[\|\theta_{\hat{Z}} - \theta_Z\|_2^2\Big] \leq \mathbb{E}\Big[\|\theta_{\hat{Z}} - \theta_Z\|_2^2 \mathbb{1}_{\{\|\theta_Z\|_0 \leq s\}}\Big] + \frac{s}{4d} \cdot \max_{z,z'} \|\theta_z - \theta_{z'}\|_2^2 \leq 2 \cdot \frac{\varepsilon^2}{16} + \frac{s}{4d} \cdot \frac{\varepsilon^2}{s} \cdot d = \frac{3\varepsilon^2}{8},$$

where we used the fact that $\hat{\theta}$ has the guarantees of a good estimator (to ℓ_2 loss $\varepsilon/4$) whenever θ_Z is s-sparse, our bound on the probability that Z is not s-sparse, and the fact that the maximum distance between any two of the mean vectors $\theta_z, \theta_{z'}$ from our construction is $\varepsilon/\sqrt{d}$. Since $\|\theta_{\hat{Z}} - \theta_Z\|_2^2 = \frac{\varepsilon^2}{s} \sum_{i=1}^d \mathbb{1}_{\{Z(i) \neq \hat{Z}(i)\}}$, this implies

$$\sum_{i=1}^d \Pr\Big[Z(i) \neq \hat{Z}(i)\Big] \leq \frac{3s}{8} .$$

By the data processing inequality, considering the Markov chain $Z(i) - Y^n - \hat{Z}(i)$, we have

$$\sum_{i=1}^d I\Big(Z(i) \wedge \hat{Z}(i)\Big) \leq \sum_{i=1}^d I(Z(i) \wedge Y^n) .$$

Thus, it is enough to show that $\sum_{i=1}^d I\Big(Z(i) \wedge \hat{Z}(i)\Big) = \Omega\Big(s \log \frac{ed}{s}\Big)$. Towards that, we have by Fano's inequality that for all i $I\Big(Z(i) \wedge \hat{Z}(i)\Big) = H(Z(i)) - H(Z(i) \mid \hat{Z}(i)) \geq h(\frac{s}{2d}) -$

$h(\Pr\big[Z(i) \neq \hat{Z}(i)\big])$, where $h(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy. It follows that

$$\begin{aligned}\sum_{i=1}^{d} I\Big(Z(i) \wedge \hat{Z}(i)\Big) &\geq d\left(h\left(\frac{s}{2d}\right) - \frac{1}{d}\sum_{i=1}^{d} h\Big(\Pr\big[Z(i) \neq \hat{Z}(i)\big]\Big)\right) \\ &\geq d\left(\left(\frac{s}{2d}\right) - h\left(\frac{1}{d}\sum_{i=1}^{d}\Pr\big[Z(i) \neq \hat{Z}(i)\big]\right)\right) \\ &\geq d\left(h\left(\frac{s}{2d}\right) - h\left(\frac{3s}{8d}\right)\right) \geq \frac{3}{100} s\log\frac{es}{d}\,,\end{aligned}$$

where the second inequality by concavity and monotonicity (on $[0, 1/2]$) of h, respectively, and the last by observing that

$$\inf_{x\in[0,1]} \frac{h(x/2) - h(3x/8)}{x\log(e/x)} > 0.03\,.$$

This concludes the proof. ■

The next (key) claim below states that, under communication constraints, the mutual information scales as the total number of bits communicated from the users.

Claim 2 *For any noninteractive protocol with ℓ bits from each of the n users, we must have $I(Z \wedge Y^n) = O\Big(\frac{n\varepsilon^2\ell}{d}\Big)$.*

Proof First, we note that while the noninteractive protocol might allow for public randomness U shared between users (public-coin protocols), it is enough to establish the bound for private-coin protocols. This is because we can condition on a particular realization u of the public randomness U: by obtaining a uniform upper bound on $I(Z \wedge Y^n \mid U = u)$ for all u, the same applies to the conditional mutual information $I(Z \wedge Y^n \mid U) = I(Z \wedge Y^n, U)$ which is the quantity of interest.

With that in mind, note that for private-coin protocols the messages $Y_1, Y_2, \ldots, Y_n$ are mutually independent conditioned on Z. This implies that

$$I(Z \wedge Y^n) \leq \sum_{t=1}^{n} I(Z \wedge Y_t)\,,$$

and thus it is enough to bound each term of the sum as $I(Z \wedge Y_t) = O\big(\varepsilon^2\ell/d\big)$. To do so, fix any $1 \leq t \leq n$, and denote $\mathbf{u}$ the uniform distribution over $\{\pm 1\}^d$. For the channel $W_t\colon \{\pm1\}^d \to \{0,1\}^\ell$ used by user t, let $W_t^{\mathbf{p}}$ be the distribution on $\mathcal{Y} := \{0,1\}^\ell$ induced by an input X drawn from $\mathbf{p}$:

$$W_t^{\mathbf{p}}(y) = \mathbb{E}_{X\sim\mathbf{p}}[W_t(y \mid X)], \qquad y \in \mathcal{Y}\,. \tag{4}$$

We can rewrite and bound the mutual information as

$$I(Z \wedge Y_t) = \mathbb{E}_Z\big[\mathrm{KL}(W_t^{\mathbf{p}_Z} \mid\mid W_t^{\mathbf{u}})\big] \leq \mathbb{E}_Z\Big[\chi^2(W_t^{\mathbf{p}_Z} \mid\mid W_t^{\mathbf{u}})\Big]\,.$$

We bound the mutual information for each user t and drop the subscript t from W_t when it is clear from context. Expanding out the chi-square divergence, we obtain the following bound on the mutual information:

$$I(Z \wedge Y_t) \leq \sum_{y\in\mathcal{Y}}\Big(\sigma^2\gamma^2 \sum_{i\in[d]} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} + \sum_{r=2}^{d} \sigma^{2r}\gamma^{2r} \sum_{\substack{B\subseteq[d]\\|B|=r}} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)\prod_{i\in B} X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}\Big), \tag{5}$$

where $\gamma = \varepsilon/\sqrt{s}$ and $\sigma^2 = \frac{s}{2d}$.

We defer the proof of Eq. (5) to Appendix D, and proceed to bound the right-hand-side. For the first term, since X is 1-subgaussian, we can invoke Lemma 7 to get

$$\sigma^2 \sum_{y\in\mathcal{Y}} \gamma^2 \sum_{i\in[d]} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} = \frac{s}{2d}\gamma^2 \sum_{y\in\mathcal{Y}} \frac{\|\mathbb{E}_{\mathbf{u}}[XW(y \mid X)]\|_2^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} \le (\ln 2)\frac{\varepsilon^2\ell}{d}. \tag{6}$$

Next we handle the second-order terms, i.e.,

$$\sigma^4\gamma^4 \sum_{y\in\mathcal{Y}} \sum_{i=1}^{d} \sum_{j\neq i} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)X(i)X(j)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]},$$

For all $i \in [d]$, we have

$$\begin{aligned}
\sum_{j\neq i} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)X(i)X(j)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} &\le \sum_{j\neq i} \frac{\frac{1}{2}\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)X(j)]^2 + \frac{1}{2}\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)X(j)]^2}{\frac{1}{2}\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)] + \frac{1}{2}\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)]} \\
&\le \sum_{j\neq i} \left(\frac{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)]} + \frac{\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)]} \right),
\end{aligned}$$

and so

$$\begin{aligned}
&\sigma^4\gamma^4 \sum_{y\in\mathcal{Y}} \sum_{i=1}^{d} \sum_{j\neq i} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)X(i)X(j)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} \\
&\quad\le \sigma^4\gamma^4 \sum_{y\in\mathcal{Y}} \sum_{i=1}^{d} \sum_{j\neq i} \left(\frac{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)]} + \frac{\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)]} \right) \\
&\quad= \sigma^4\gamma^4 \sum_{i=1}^{d} \sum_{y\in\mathcal{Y}} \left(\frac{\sum_{j\neq i}\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)]} + \frac{\sum_{j\neq i}\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)]} \right) \\
&\quad= \sigma^4\gamma^4 \sum_{i=1}^{d} \sum_{y\in\mathcal{Y}} \left(\frac{\|\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)X_{-i}]\|_2^2}{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)]} + \frac{\|\mathbb{E}_{\mathbf{u}|X(i)=-1}[W(y \mid X)X_{-i}]\|_2^2}{\mathbb{E}_{\mathbf{u}|X(i)=1}[W(y \mid X)]} \right) \\
&\quad\le 2(\ln 2)\sigma^4\gamma^4 \cdot (2d\ell) = (\ln 2)\frac{\ell\varepsilon^2}{d} \cdot \varepsilon^2.
\end{aligned} \tag{7}$$

Similarly, we can bound the jth-order terms as

$$\sigma^{2r}\gamma^{2r} \sum_{y\in\mathcal{Y}} \sum_{\substack{B\subseteq[d] \\ |B|=r}} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} \le \frac{\ell\varepsilon^2}{d} \cdot (\ln 2)(\varepsilon^2)^{j-1}. \tag{8}$$

We defer the details to Appendix D. And thus, summing over all terms, we get

$$I(Z \wedge Y_t) \le \frac{(\ln 2)\varepsilon^2\ell}{d} \sum_{j=1}^{\infty} \varepsilon^{2(j-1)} \le \frac{2(\ln 2)\ell\varepsilon^2}{d}.$$

Summing over $1 \le t \le n$, we get the desired result. ■

Putting together Claims 1 and 2 then completes the proof of Lemma 11. ■

Remark 12 *As a byproduct, the proof of Lemma 11 above (for the noninteractive case) has an interesting corollary: the lower bound framework of Acharya et al. (2020a) for the interactive case, which proceeds by bounding a quantity termed* average discrepancy, *could not possibly go through in the sparse case with* $\sum_{i=1}^{d} I(Z(i) \wedge Y^n)$ *instead of average discrepancy. Indeed, if the bound of Acharya et al. (2020a) applied to* $\sum_{i=1}^{d} I(Z(i) \wedge Y^n)$ *as well, we would get the same lower bound above for interactive protocols, which in turn will contradict the upper bound of Lemma 8 for interactive protocols.*

Combining Lemmas 10, 8, 9, and 11 establishes Theorem 1. Finally, we mention that while our noninteractive lower bound (Lemma 11) requires $s = \Omega(\log d)$, we are able to establish separately the case $s = 1$ via a simple, different proof (see Theorem 24). We provide this result in Appendix B, as we believe it to be of independent interest and will also be requiring it in the proof of Lemma 14.

3. Block-sparse mean estimation under communication constraints

In this section, we establish Theorem 3, our result for s-block-sparse mean estimation under ℓ-bit communication constraints. In order to establish the result, we need an upper and a lower bound on the sample complexity of both noninteractive and interactive protocols.

Of these four bounds, only a restricted version of the interactive upper bound was known, which assumed that all coordinates of the block-sparse mean had the same magnitude and that $\ell = 1$ (Acharya et al., 2021, Theorem 13). While the algorithm can easily be made to extend to $\ell > 1$, it crucially relies on the former assumption on the structure of the block-sparse mean, and thus does not translate to our setting.

We proceed to prove separately the four bounds, starting with the noninteractive upper bound.

Lemma 13 *For any* $s \geq 1$, *the* noninteractive *sample complexity of mean estimation of* s*-block sparse product distributions over* $\{\pm 1\}^d$ *under* ℓ*-bit communication constraints is* $O\Big(\frac{sd + d\log d}{\varepsilon^2 \ell}\Big)$.

Proof As in the proof of Theorem 10, by using the "simulate-and-infer" idea of Acharya et al. (2020c) it suffices to show an $O\Big(\frac{s+\log d}{\varepsilon^2}\Big)$ upper bound in the *unconstrained* setting (where all the observations are fully available). This $O\Big(\frac{s+\log d}{\varepsilon^2}\Big)$ sample complexity upper bound then can be obtained by the following simple estimator: partition the d coordinates in $\lceil d/s \rceil$ consecutive blocks of (at most) s coordinates, and, using the same samples, separately estimate the $\lceil d/s \rceil$ mean subvectors to ℓ_2 loss $\varepsilon^2/3$, with probability of success $\delta := s/(10d)$. This can be done with

$$O\Big(\frac{s + \log(1/\delta)}{\varepsilon^2}\Big) = O\Big(\frac{s + \log \frac{d}{s}}{\varepsilon^2}\Big) = O\Big(\frac{s + \log d}{\varepsilon^2}\Big)$$

samples, by (sub) Gaussian concentration of measure. By a union bound, all of the $\lceil d/s \rceil$ estimates are simultaneously accurate, with probability at least $9/10$. Since the "true" block overlaps at most 2 consecutive blocks of the $\lceil d/s \rceil$ considered, it then suffices to output the vector $\hat{\mu}$ consisting of only the two estimated subvectors with largest magnitude (and all other coordinates set to zero). ■

Next, we establish a matching lower bound for noninteractive protocols.

Lemma 14 *For any $s \geq 1$, the* noninteractive *sample complexity of mean estimation of s-block sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is $\Omega\Big(\frac{sd+d\log d}{\varepsilon^2\ell}\Big)$.*

Proof The $\Omega\Big(\frac{d\log d}{\varepsilon^2\ell}\Big)$ term follows from the 1-sparse estimation lower bound established in Theorem 24, since any 1-sparse product distribution is s-block-sparse for every s. We thus focus on the main term, and establish the $\Omega\Big(\frac{sd}{\varepsilon^2\ell}\Big)$ lower bound.

To do so, consider the following set of s-block sparse distributions. Partition $[d]$ into $b := d/s$ consecutive nonoverlapping blocks, $B_1, B_2, \ldots, B_b$, each of size at most s. For all $z \in \{\pm 1\}^d$ and $j \in [b]$, define $\mathbf{p}_{z,j}$ as a product distribution over $\{\pm 1\}^d$ with mean $\theta_{z,j}$ given by

$$\theta_{z,j}(i) = \begin{cases} \frac{\varepsilon}{\sqrt{s}} z(i), & \text{if } i \in B_j, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Consider the following generative process. First independently pick Z uniformly at random from $\{\pm 1\}^d$ and J uniformly from $[b]$. Then, each of the n users observes one sample X_t from the product distribution $\mathbf{p}_{Z,J}$ with mean vector $\theta_{Z,J}$ and sends its samples through a channel $W_t \colon \{\pm 1\}^d \to \{0,1\}^\ell$ to compress it to a message Y_t.

The next result states that any sufficiently accurate estimation protocol must provide enough information about each $Z(i)$ from the tuple of messages Y^n, even if J is known.

Lemma 15 (Assouad-type Bound) *For any protocol that estimates s-block-sparse product distributions to ℓ_2 accuracy ε, we must have $\sum_{i=1}^d I(Z(i) \wedge Y^n \mid J) = \Omega(s)$.*

Proof Let $\hat{Z}$ be an estimator of Z based on Y^n. By the data processing inequality, it is be enough to prove $\sum_{i=1}^d I\Big(Z(i) \wedge \hat{Z}(i) \mid J\Big) = \Omega(s)$. By definition,

$$\sum_{i=1}^d I\Big(Z(i) \wedge \hat{Z}(i) \mid J\Big) = \frac{s}{d}\sum_{j=1}^b \sum_{i=1}^d I\Big(Z(i) \wedge \hat{Z}(i) \mid J = j\Big) = \frac{s}{d}\sum_{j=1}^b \sum_{i \in B_j} I\Big(Z(i) \wedge \hat{Z}(i) \mid J = j\Big). \tag{10}$$

Now,

$$\begin{aligned} \sum_{j=1}^b \sum_{i \in B_j} I\Big(Z(i) \wedge \hat{Z}(i) \mid J = j\Big) &= \sum_{j=1}^b \sum_{i \in B_j} \Big(H(Z(i) \mid J = j) - H\Big(Z(i) \mid \hat{Z}(i), J = j\Big)\Big) \\ &\geq \sum_{j=1}^b \sum_{i \in B_j} \Big(H(Z(i) \mid J = j) - h\Big(\Pr\Big[\hat{Z}(i) \neq Z(i) \mid J = j\Big]\Big)\Big) \\ &= d - \sum_{j=1}^b \sum_{i \in B_j} h\Big(\Pr\Big[\hat{Z}(i) \neq Z(i) \mid J = j\Big]\Big) \\ &\geq d - d \cdot h\left(\frac{1}{d}\sum_{j=1}^b \sum_{i \in B_j} \Pr\Big[\hat{Z}(i) \neq Z(i) \mid J = j\Big]\right), \end{aligned} \tag{11}$$

where $h \colon [0,1] \to \mathbb{R}$ is the binary entropy function. By construction, for any valid protocol, we must have

$$\frac{\varepsilon^2}{10} \geq \mathbb{E}\left[\sum_{i=1}^d \frac{\varepsilon^2}{s} \mathbb{1}_{\{\hat{Z}(i) \neq Z(i)\}}\right] = \frac{\varepsilon^2}{s}\sum_{i=1}^d \Pr\Big[\hat{Z}(i) \neq Z(i)\Big]$$

$$= \frac{\varepsilon^2}{s}\frac{s}{d}\sum_{j=1}^{b}\sum_{i=1}^{d}\Pr\Big[\hat{Z}(i) \neq Z(i) \mid J = j\Big]$$
$$\geq \frac{\varepsilon^2}{d}\sum_{j=1}^{b}\sum_{i\in B_j}\Pr\Big[\hat{Z}(i) \neq Z(i) \mid J = j\Big],$$

which implies

$$\frac{1}{d}\sum_{j=1}^{b}\sum_{i\in B_j}\Pr\Big[\hat{Z}(i) \neq Z(i) \mid J = j\Big] \leq \frac{1}{10}.$$

Combining this with Eqs. (10) and (11) completes the proof of the lemma. ■

Using independence of $Z(i)$'s, by additivity of mutual information this claim then implies that

$$I(Z \wedge Y^n \mid J) \geq \sum_{i=1}^{d} I(Z(i) \wedge Y^n \mid J) = \Omega(s). \tag{12}$$

Having obtained a lower bound on the mutual information, we now provide an upper bound for it; combining the two will yield our lower bound for sample complexity.

Lemma 16 *For any noninteractive protocol using ℓ-bit communication constraints, we must have*

$$I(Z \wedge Y^n \mid J) = O\left(\frac{n\varepsilon^2\ell}{d}\right).$$

Proof Note that, since $I(Z \wedge Y^n \mid J) = \frac{1}{b}\sum_{j\in[b]} I(Z \wedge Y^n \mid J = j)$, it is enough to prove that $\sum_{j\in[b]} I(Z \wedge Y^n \mid J = j) = O\Big(\frac{n\varepsilon^2\ell}{s}\Big)$. Similar to Eq. (5), the first step of the proof is to bound the mutual information at each time step. Let $\gamma := \varepsilon/\sqrt{s}$, at each user t, the following inequality holds.

$$\sum_{j\in[b]} I(Z \wedge Y_t \mid J = j)$$
$$\leq \sum_{j\in[b]}\sum_{y\in\mathcal{Y}}\Big(\gamma^2\sum_{i\in B_j}\frac{\mathbb{E}_X[W(y \mid X)X(i)]^2}{\mathbb{E}_X[W(y \mid X)]} + \sum_{r=2}^{s}\gamma^{2r}\sum_{\substack{B\subseteq B_j\\|B|=r}}\frac{\mathbb{E}_X[W(y \mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_X[W(y \mid X)]}\Big). \tag{13}$$

We defer the proof of Eq. (13) to Appendix D, and proceed to bound the RHS. For all $r \in [s]$, let

$$\zeta_r := \gamma^{2r}\sum_{j\in[b]}\sum_{y\in\mathcal{Y}}\sum_{B\subseteq B_j, |B|=r}\frac{\mathbb{E}_X[W(y \mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_X[W(y \mid X)]},$$

whereby we can rewrite the earlier bound as

$$\sum_{j\in[b]} I(Z \wedge Y_t \mid J = j) \leq \sum_{r\in[s]}\zeta_r. \tag{14}$$

We will bound each ζ_r separately. To bound ζ_1, since X is 1-subgaussian, from Lemma 7 we have

$$\sum_{j\in[b]}\sum_{y\in\mathcal{Y}}\frac{\varepsilon^2}{s}\sum_{i\in B_j}\frac{\mathbb{E}_X[W(y \mid X)X(i)]^2}{\mathbb{E}_X[W(y \mid X)]} = \frac{\varepsilon^2}{s}\sum_{y\in\mathcal{Y}}\frac{\|\mathbb{E}_X[XW(y \mid X)]\|_2^2}{\mathbb{E}_X[W(y \mid X)]} \leq 2(\ln 2)\frac{\varepsilon^2\ell}{s},$$

Next we bound

$$\zeta_2 = \gamma^4 \sum_{y\in\mathcal{Y}} \sum_{j\in[b]} \sum_{i<i'\in B_j} \frac{\mathbb{E}_X[W(y \mid X)X(i)X(i')]^2}{\mathbb{E}_X[W(y \mid X)]}.$$

Note that each term in the summation above is a product of two entries of X, which are not independent: hence, we cannot use Lemma 7 directly. To resolve this, we use the following lemma, which is a consequence of Baranyai's Theorem (Baranyai, 1974).

Lemma 17 *Let $[s] = \{1, 2, \ldots, s\}$. $\{B \subset [s] : |B| = r\}$, the set of all size-$r$ subsets of $[s]$, can be partitioned into $m \leq 2\binom{s-1}{r-1}$ sets $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ such that all subsets within each $\mathcal{M}_i$ are disjoint.*

Without loss of generality, assume that, for all $j \in [b]$, $B_j = \{(j-1)s+1, \ldots, js\}$. Using Lemma 17 with $r = 2$, we can rewrite ζ_2 as

$$\begin{aligned}
\zeta_2 &= \gamma^4 \sum_{y\in\mathcal{Y}} \sum_{j\in[b]} \sum_{i<i'\in B_j} \frac{\mathbb{E}_X[W(y \mid X)X(i)X(i')]^2}{\mathbb{E}_X[W(y \mid X)]} \\
&= \gamma^4 \sum_{y\in\mathcal{Y}} \sum_{j\in[b]} \sum_{k\in[s-1]} \sum_{(i,i')\in\mathcal{M}_k} \frac{\mathbb{E}_X[W(y \mid X)X((j-1)s+i)X((j-1)s+i')]^2}{\mathbb{E}_X[W(y \mid X)]} \\
&= \gamma^4 \sum_{k\in[s-1]} \sum_{y\in\mathcal{Y}} \sum_{j\in[b]} \sum_{(i,i')\in\mathcal{M}_k} \frac{\mathbb{E}_X[W(y \mid X)X((j-1)s+i)X((j-1)s+i')]^2}{\mathbb{E}_X[W(y \mid X)]}.
\end{aligned}$$

Note that in the summation above, for each subset $\mathcal{M}_k$, the pairwise products have disjoint entries and hence independent. Moreover, $X(i)X(i')$ is 1-subgaussian as well since it is supported on $\{\pm 1\}$ with mean zero. For $k \in [s-1]$, let $\Phi_k(X)$ be the vector whose coordinates are $X((j-1)s+i)X((j-1)s+i')$ for $j \in [b]$ and $(i,i') \in \mathcal{M}_k$. Then we have $\Phi_k(X)$ have independent coordinates and is 1-subgaussian. Proceeding from above, applying Lemma 7, the equation can be bounded by

$$\begin{aligned}
\zeta_2 &= \gamma^4 \sum_{k\in[s-1]} \sum_{y\in\mathcal{Y}} \sum_{j\in[b]} \sum_{(i,i')\in\mathcal{M}_k} \frac{\mathbb{E}_X[W(y \mid X)X((j-1)s+i)X((j-1)s+i')]^2}{\mathbb{E}_X[W(y \mid X)]} \\
&= \gamma^4 \sum_{k\in[s-1]} \sum_{y\in\mathcal{Y}} \frac{\|\mathbb{E}_X[W(y \mid X)\Phi_k(X)]\|_2^2}{\mathbb{E}_X[W(y \mid X)]} \\
&\leq 2(\ln 2)\frac{\varepsilon^2\ell}{s} \cdot \varepsilon^2.
\end{aligned}$$

By similar grouping techniques on the higher order terms, using Lemma 17, we can prove that for all $r \geq 3$ $\zeta_r \leq 2(\ln 2)\frac{\varepsilon^2\ell}{s} \cdot (\varepsilon^2)^{r-1}$. Combining these and Eq. (14), we get

$$\sum_{j\in[b]} I(Z \wedge Y_t \mid J = j) \leq 2(\ln 2)\frac{\varepsilon^2\ell}{s} \sum_{r\in[s]} (\varepsilon^2)^{r-1} \leq 4(\ln 2)\frac{\varepsilon^2\ell}{s}.$$

The claim follows from the observation above since, conditioned on Z, $Y_1, Y_2, \ldots, Y_n$ are independent,[6] we have

$$\frac{1}{b}\sum_{j\in[b]} I(Z \wedge Y^n \mid J=j) \le \frac{1}{b}\sum_{t=1}^{n}\sum_{j\in[b]} I(Z \wedge Y_t \mid J=j) \le \frac{1}{b}\cdot 4(\ln 2)\frac{n\varepsilon^2\ell}{s} = 4(\ln 2)\frac{n\varepsilon^2\ell}{d},$$

showing the result. ■

Combining the two claims concludes the proof, as this implies that one must have $\frac{n\varepsilon^2\ell}{d} = \Omega(s)$. ■

We now turn to the upper bound for interactive protocols. The algorithm has a two-stage procedure. In the first stage, users first detects the "active" block with size $\Theta(s)$. Then in the second stage, the users will focus on learning coordinates within the detected block, which needs less samples.

Lemma 18 *For any $s \ge 1$, the* interactive *sample complexity of mean estimation of s-block sparse product distributions over $\{\pm 1\}^d$ under ℓ-bit communication constraints is*

$$O\left(\frac{s^2 + d\log(d/s)\log(s/\varepsilon)}{\varepsilon^2\ell} + \frac{s\log(d/s)\log(s/\varepsilon)}{\varepsilon^2}\right).$$

Proof The algorithm works in two stages: *detection* and *estimation*. We start by partitioning the d coordinates into $T := \lceil d/s \rceil$ consecutive blocks of (at most) s coordinates, $B_1, \ldots, B_T$. Let μ_{B_j} be the mean vector restricted on block B_j. Since the actual support of the mean vector overlaps at most 2 such blocks, if $\|\mu\|_2^2 > \varepsilon^2$ there exists some $j \in [T]$ such that $\|\mu_{B_j}\|_2^2 > \varepsilon^2/2$. On the other hand, if $\|\mu\|_2^2 \le \varepsilon^2$, then no such j may exist, but our task in that case will be simpler.

The algorithm proceeds in the following two stages:

1. **Detection:** Identify, with probability at least $19/20$, a block B_j such that $\|\mu_{B_j}\|_2^2 > \varepsilon^2/2$, if there exists one, using $O\Big(\frac{d\log(d/s)\log(s/\varepsilon)}{\ell\varepsilon^2} + \frac{s\log(d/s)\log(s/\varepsilon)}{\varepsilon^2}\Big)$ samples. This detection step is the most involved, and will constitute most of the proof below.

2. **Estimation:** If no such block was identified, output the zero vector (which is a good estimate); otherwise, consider the union of the 3 blocks $B_{j-1} \cup B_j \cup B_{j+1}$, which has at most $3s$ coordinates and contains the support of the unknown s-sparse vector μ. Use the noninteractive estimation algorithm (with "$d = 3s$") to learn, with probability $19/20$, the corresponding mean with $O\Big(\frac{s^2}{\min(s,\ell)\varepsilon^2}\Big) = O\Big(\frac{s^2}{\varepsilon^2\ell} + \frac{s}{\varepsilon^2}\Big)$ new samples.

The overall algorithm has the claimed sample complexity and, by a union bound, is successful overall with probability $9/10$. Details for the first stage follow.

Our algorithm will use public randomness as follows. All users jointly draw a Rademacher vector $\xi = (\xi(i))_{i\in[d]}$ uniformly at random. Let $\Delta := 5\sqrt{s\log(s/\varepsilon)}$. Any given user computes the T bits $M(1), \ldots, M(T)$ based on ξ as follows. For every $j \in [T]$, upon observing X (and knowing ξ) each user computes $M(j)$ based on $\bar{X}_j := \sum_{i\in B_j} X(i)\xi(i)$ using a stochastic rounding algorithm:

$$M(j) = \begin{cases} +1, & \text{with probability } \frac{\Delta + \mathrm{Clip}_\Delta(\bar{X}_j)}{2\Delta}, \\ -1, & \text{with probability } \frac{\Delta - \mathrm{Clip}_\Delta(\bar{X}_j)}{2\Delta}. \end{cases}$$

6. We can here ignore public randomness, as we can bound the quantity under each fixed realization of the public coins.

where $\mathrm{Clip}_\Delta(x) := \max\{\min\{x, \Delta\}, -\Delta\}$ denotes the clipping function on the interval $[-\Delta, \Delta]$. It can be seen that $(M(1), \dots, M(T))$ follows a product distribution over $\{\pm 1\}^T$. Next we analyze the mean on each coordinate, conditioned on ξ.

$$\mathbb{E}[M(j) \mid \xi] = 2 \cdot \mathbb{E}\left[\frac{\Delta + \mathrm{Clip}_\Delta\left(\bar{X}_j\right)}{2\Delta} \;\middle|\; \xi\right] - 1 = \frac{\mathbb{E}\left[\mathrm{Clip}_\Delta\left(\bar{X}_j\right) \;\middle|\; \xi\right]}{\Delta}. \tag{15}$$

Let $\bar{\mu}(B_j, \xi) := \mathbb{E}\left[\bar{X}_j \mid \xi\right] = \sum_{i \in B_j} \xi(i)\mu(i)$. Note that when a block j does not intersect with the support of μ, then $\bar{\mu}(B_j, \xi) = 0$. Further, since each $X(i)$ in B_j is then symmetric, the clipping does not change the mean: thus, for any $j \in [T]$ such that B_j does not intersect the support of μ,

$$\mathbb{E}[M(j) \mid \xi] = 0.$$

That is, we then have $\Pr[M(j) = 1 \mid \xi] = 1/2$ regardless of the realization of the shared random variable ξ.

Suppose now that B_j *does* intersect the support of the mean vector μ, and specifically that $\|\mu_{B_j}\|_2^2 > \varepsilon^2/2$. We then show the following:

Claim 3 *If $\|\mu_{B_j}\|_2^2 > \varepsilon^2/2$, then with probability at least $1/8$ over the choice of ξ, we have*

$$|\mathbb{E}[M(j) \mid \xi]| \geq \frac{\varepsilon}{40\sqrt{s \log(s/\varepsilon)}}.$$

Proof We first show that before performing stochastic rounding, with probability at least $1/4$ over the randomness of ξ it is the case that

$$|\bar{\mu}(B_j, \xi)| \geq \frac{\varepsilon}{4}. \tag{16}$$

To see this, notice that the second moment of $\bar{\mu}(B_j, \xi)$ is large:

$$\mathbb{E}_\xi\left[\bar{\mu}(B_j, \xi)^2\right] = \mathbb{E}_\xi\left[\left(\sum_{i \in B_j} \xi_i \mu_i\right)^2\right] = \sum_{i \in B_j} \mu_i^2 \geq \frac{\varepsilon^2}{2}.$$

We also can control the fourth moment of $\bar{\mu}(B_j, \xi)$ as follows:

$$\mathbb{E}_\xi\left[\bar{\mu}(B_j, \xi)^4\right] = \mathbb{E}_\xi\left[\left(\sum_{i \in B_j} \xi_i \mu_i\right)^4\right] = \sum_{i \in B_j} \mu_i^4 + 6 \sum_{i < i' \in B_j} \mu_i^2 \mu_{i'}^2 \leq 3\left(\sum_{i \in B_j} \mu_i^2\right)^2.$$

Hence, by the Paley–Zygmund inequality, we have

$$\Pr\left[|\bar{\mu}(B_j, \xi)| > \frac{\varepsilon}{4}\right] \geq \Pr\left[\bar{\mu}(B_j, \xi)^2 > \frac{1}{8}\mathbb{E}\left[\bar{\mu}(B_j, \xi)^2\right]\right] \geq \frac{3}{4}\frac{\mathbb{E}\left[\bar{\mu}(B_j, \xi)^2\right]^2}{\mathbb{E}[\bar{\mu}(B_j, \xi)^4]} \geq \frac{1}{4},$$

which proves that, as stated, (16) holds with probability at least $1/4$. Next we prove that the clipping does not affect this too much; namely, that with probability least $8/9$ over the randomness of ξ,

$$\left|\mathbb{E}\left[\mathrm{Clip}_\Delta\left(\bar{X}_j\right) \mid \xi\right] - \bar{\mu}(B_j, \xi)\right| \leq \frac{\varepsilon}{8}. \tag{17}$$

Before proving the above statement, we note that by a union bound, both (16) and (17) simultaneously happen with probability at least $1/4 - 1/9 > 1/8$. Combining this with (15) and the value of Δ then establishes the claim. Thus, to conclude it only remains to prove (17).

Since $\mathbb{E}_\xi[\bar{\mu}(B_j,\xi)^2] = \sum_{i\in B_j}\mu_i^2 \le s$, by Markov's inequality, with probability at least $8/9$,

$$|\bar{\mu}(B_j,\xi)| \le 3\sqrt{s}.$$

Call this event $\mathcal{E}$. Conditioning on $\mathcal{E}$, we bound the probability that the sum gets clipped. By Hoeffding's inequality, we have

$$\Pr\Big[\bar{X}_j \notin [-\Delta,\Delta] \mid \xi,\mathcal{E}\Big] \le \Pr\left[\left|\bar{X}_j - \bar{\mu}(B_j,\xi)\right| \ge 2\sqrt{s\log(s/\varepsilon)} \;\middle|\; \xi,\mathcal{E}\right] \le 2\left(\frac{\varepsilon}{s}\right)^2.$$

Hence assuming $\varepsilon \le 1/16$, we can upper bound the clipping error by

$$\left|\mathbb{E}\Big[\mathrm{Clip}_\Delta\Big(\bar{X}_j\Big) \mid \xi\Big] - \bar{\mu}(B_j,\xi)\right| \le s\cdot\Pr\Big[\bar{X}_j \notin [-\Delta,\Delta]\Big] \le \frac{2\varepsilon^2}{s} \le \frac{\varepsilon}{8},$$

concluding the proof. ■

With this claim in hand, we can analyze the detection step as follows. We have, after the above transformation and conditioned on ξ, each user obtains $(M(1),M(2),\dots,M(T))$ from a product distribution over $\{\pm1\}^T$. By the "simulate-and-infer" trick (Acharya et al., 2020c), the mean vector of the product distribution can be learned to ℓ_∞ distance $\frac{\varepsilon}{20\sqrt{s\log(s/\varepsilon)}}$ with

$$O\left(\frac{T\log T}{\min(T,\ell)(\varepsilon/\sqrt{s\log(s/\varepsilon)})^2}\right) = O\left(\frac{d\log(d/s)\log(s/\varepsilon)}{\min(d/s,\ell)\varepsilon^2}\right)$$

samples, allowing us to detect (with probability at least $99/100$) the at most 2 biased coordinates. Of course, overall, we may only detect them when the choice of ξ was good (so that the coordinates corresponding to the (at most two) biased blocks ended up indeed $\Omega(\varepsilon/\sqrt{s})$-biased); but since this happens with constant probability, one can pay a constant factor in the sample complexity and amplify this, to get a $99/100$ success probability overall. This concludes the proof. ■

Finally, we prove the matching lower bound (up to logarithmic factors).

Lemma 19 *For any $s \ge 1$, the* interactive *sample complexity of mean estimation of s-block sparse product distributions over $\{\pm1\}^d$ under ℓ-bit communication constraints is $\Omega\Big(\frac{s^2+d}{\varepsilon^2\ell} + \frac{s}{\varepsilon^2}\Big)$.*

Proof The $\Omega\Big(\frac{s}{\varepsilon^2}\Big)$ term is simply the (unconstrained) "oracle bound," as $\Omega(\frac{s}{\varepsilon^2})$ samples are required even without communication constraints and knowing which block of coordinates corresponds to the support of the mean vector.

The $\Omega(\frac{d}{\varepsilon^2\ell})$ term follows from the case of 1-sparse estimation (Shamir, 2014, Theorem 2) (since, again, any 1-sparse product distribution is s-block-sparse for any $s \ge 1$). Finally, the last term of the lower bound, $\Omega(\frac{s^2}{\varepsilon^2\ell})$, follows from the lower bound on mean estimation under communication constraints (see, *e.g.*, Acharya et al. (2020a, Theorem 3)) in the *nonsparse* case with $d = s$, since even knowing the location of the block we still have a mean estimation task under information constraints, with dimensionality s. ■

References

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/acharya19a.html.

Jayadev Acharya, Clément L. Canonne, Ziteng Sun, and Himanshu Tyagi. Unified lower bounds for interactive high-dimensional estimation under information constraints. *CoRR*, abs/2010.06562, 2020a.

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: Lower bounds from chi-square contraction. *IEEE Trans. Inform. Theory*, 66(12):7835–7855, 2020b. ISSN 0018-9448. doi: 10.1109/TIT.2020.3028440. URL https://doi.org/10.1109/TIT.2020.3028440. Preprint available at arXiv:abs/1812.11476.

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints II: Communication constraints and shared randomness. *IEEE Trans. Inform. Theory*, 66(12):7856–7877, 2020c. ISSN 0018-9448. doi: 10.1109/TIT.2020.3028439. URL https://doi.org/10.1109/TIT.2020.3028439.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Distributed signal detection under communication constraints. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 41–63. PMLR, 09–12 Jul 2020d. URL https://proceedings.mlr.press/v125/acharya20b.html.

Jayadev Acharya, Clément L. Canonne, Yanjun Han, Ziteng Sun, and Himanshu Tyagi. Domain compression and its application to randomness-optimal distributed goodness-of-fit. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3–40. PMLR, 09–12 Jul 2020e. URL http://proceedings.mlr.press/v125/acharya20a.html.

Jayadev Acharya, Clement Canonne, Prathamesh Mayekar, and Himanshu Tyagi. Information-constrained optimization: can adaptive processing of gradients help? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7126–7138. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/398475c83b47075e8897a083e97eb9f0-Paper.pdf.

Jayadev Acharya, Clément L. Canonne, Yuhan Liu, Ziteng Sun, and Himanshu Tyagi. Interactive inference under information constraints. *IEEE Transactions on Information Theory*, 68(1):502–516, 2022.

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*,

volume 65 of *Proceedings of Machine Learning Research*, pages 39–75. PMLR, 07–10 Jul 2017. URL https://proceedings.mlr.press/v65/agarwal17c.html.

Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on information theory*, 56(4):1982–2001, 2010.

Zsolt Baranyai. On the factorization of the complete uniform hypergraphs. *Infinite and finite sets*, 1974.

Yann Barbotin, Ali Hormati, Sundeep Rangan, and Martin Vetterli. Estimation of sparse mimo channels with common support. *IEEE Transactions on Communications*, 60(12):3705–3716, 2012.

Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE J. Sel. Areas Inf. Theory*, 1(3):645–659, 2020.

Dror Baron, Marco F Duarte, Michael B Wakin, Shriram Sarvotham, and Richard G Baraniuk. Distributed compressive sensing. *arXiv preprint arXiv:0901.3403*, 2009.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Symposium on Theory of Computing Conference, STOC'16*, pages 1011–1020. ACM, 2016.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

Yuval Dagan and Vitaly Feldman. Interaction is necessary for distributed learning with privacy or communication constraints. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 450–462, 2020.

Amit Daniely and Vitaly Feldman. Locally private learning without interaction requires separation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/d01c25576ff1c53de58e0e6970a2d510-Paper.pdf.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1161–1191, Phoenix, USA, June 2019. PMLR.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.*, 113(521):182–201, 2018. ISSN 0162-1459.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, volume 3876 of *Lecture Notes in Comput. Sci.*, pages 265–284. Springer, Berlin, 2006.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2765–2781, 2013.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

Sandeep Gogineni and Arye Nehorai. Target estimation using sparse modeling for distributed mimo radar. *IEEE Transactions on Signal Processing*, 59(11):5315–5325, 2011.

Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally private hypothesis selection. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1785–1816. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/gopi20a.html.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Proceedings of the 31st Conference on Learning Theory, COLT 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 3163–3188. PMLR, 2018.

Tianyuan Jin, Jieming SHI, Xiaokui Xiao, and Enhong Chen. Efficient pure exploration in adaptive round model. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Matthew Joseph, Jieming Mao, Seth Neel, and Aaron Roth. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 94–105. IEEE, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL http://dx.doi.org/10.1561/2200000083.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008*, pages 531–540. IEEE, October 2008.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011. ISSN 0097-5397.

Matthew L. Malloy and Robert D. Nowak. Near-optimal adaptive compressed sensing. *IEEE Transactions on Information Theory*, 60(7):4001–4012, 2014. doi: 10.1109/TIT.2014.2321552.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994, 2011.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems 27*, pages 163–171, 2014.

Yasser Shoukry and Paulo Tabuada. Event-triggered state observers for sparse sensor noise/attacks. *IEEE Transactions on Automatic Control*, 61(8):2079–2091, 2015.

Mihailo Stojnic, Farzad Parvaresh, and Babak Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Transactions on Signal Processing*, 57(8): 3075–3085, 2009.

Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael I. Jordan, Ken Goldberg, and Joseph E. Gonzalez. Pac best arm identification under a deadline. *arXiv preprint arXiv:2106.03221*, 2021.

Jonathan Ullman. Tight lower bounds for locally differentially private selection. *arXiv preprint arXiv:1802.02638*, 2018.

Sergiy A Vorobyov, Alex B Gershman, and Kon Max Wong. Maximum likelihood direction-of-arrival estimation in unknown noise fields using sparse sensor arrays. *IEEE Transactions on Signal Processing*, 53(1):34–43, 2004.

Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics, 2020. URL http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018. ISSN 0018-9448. doi: 10.1109/TIT.2018.2809790. URL https://doi.org/10.1109/TIT.2018.2809790.

Appendix A. Results for the local privacy setting

In this section, we discuss how the results can be extended to the local privacy setting (LDP). In particular, we will focus on estimating the mean of sparse product distributions over $\{\pm 1\}^d$. The results on the block-sparse case will follow similarly. Under LDP constraints, each observation X_t must be privatized using an ϱ-LDP channel to get Y_t, which the estimate is based on.

Definition 20 *For $\varrho > 0$, a channel $W : \mathcal{X} \to \mathcal{Y}$ is said to be* ϱ-LDP *if, for all $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$,*

$$\frac{W(y \mid x)}{W(y \mid x')} \leq e^{\varrho}.$$

We focus on the high privacy regime, *i.e.*, when $\rho = O(1)$, and state the results below. Note that, in this regime, $(e^{\varrho} - 1)^2 = O(\varrho^2)$.

Theorem 21 *For any $s \geq 4 \log d$, any ρ-LDP noninteractive protocol for mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ must have sample complexity $\Omega\Big(\frac{sd}{\varepsilon^2 \varrho^2} \log \frac{ed}{s}\Big)$.*

Combined with previously known results for sparse mean estimation, this lower bound immediately implies the following:

Corollary 22 *For any $s \geq 4 \log d$, the* noninteractive *sample complexity of mean estimation of s-sparse product distributions over $\{\pm 1\}^d$ under ϱ-LDP constraints is $\Theta\Big(\frac{sd}{\varepsilon^2 \varrho^2} \log \frac{ed}{s}\Big)$, while the* interactive *sample complexity is $\Theta\Big(\frac{sd}{\varepsilon^2 \varrho^2}\Big)$.*

Of these bounds, the interactive upper and lower bounds are shown in Acharya et al. (2020a) and Duchi and Rogers (2019). The noninteractive upper bound was established in Duchi et al. (2018). The proof of Theorem 21, the noninteractive lower bound, follows similar steps as the proof of Theorem 1[7]. We now discuss how to modify the argument for estimation under LDP constraints.

We first follow the same steps as in the proof of Theorem 1 until Eq. (5), which we write below.

$$I(Z \wedge Y_t) \leq \sum_{y \in \mathcal{Y}} \Big(\frac{s\gamma^2}{2d} \sum_{i \in [d]} \frac{\mathbb{E}_X[W(y \mid X) X_i]^2}{\mathbb{E}_X[W(y \mid X)]} + \sum_{r=2}^{d} \Big(\frac{s\gamma^2}{2d} \Big)^r \sum_{\substack{B \subseteq [d] \\ |B| = r}} \frac{\mathbb{E}_X[W(y \mid X) \prod_{i \in B} X_i]^2}{\mathbb{E}_X[W(y \mid X)]} \Big).$$

To bound each term, we need the lemma below, proved in Acharya et al. (2020a), which follows from direct application of Bessel's inequality.

Lemma 23 *Let $\phi_i \colon \mathcal{X} \to \mathbb{R}$, for $i \leq 1$, be a family of functions. If the functions satisfy, for all i, j,*

$$\mathbb{E}_X[\phi_i(X) \phi_j(X)] = \mathbb{1}_{\{i=j\}},$$

then, for any ϱ-LDP channel W, we have

$$\sum_i \mathbb{E}_X[\phi_i(X) W(y \mid X)]^2 \leq \mathrm{Var}_X\left[W(y \mid X)\right].$$

Note that for the first term,

$$\begin{aligned} \sum_{y \in \mathcal{Y}} \frac{s}{2d} \gamma^2 \sum_{i \in [d]} \frac{\mathbb{E}_{X \sim \mathbf{u}}[W(y \mid X) X_i]^2}{\mathbb{E}_X[W(y \mid X)]} &\leq \frac{s}{2d} \gamma^2 \sum_{y \in \mathcal{Y}} \frac{\mathrm{Var}_X\left[W(y \mid X)\right]}{\mathbb{E}_X[W(y \mid X)]} \\ &\leq \frac{s}{2d} \gamma^2 \sum_{y \in \mathcal{Y}} \frac{(e^{\varrho} - 1)^2 \mathbb{E}_X[W(y \mid X)]^2}{\mathbb{E}_X[W(y \mid X)]} \\ &= \frac{\varepsilon^2}{2d} (e^{\varrho} - 1)^2. \end{aligned}$$

7. For the case of $s = 1$, a lower bound of $\Omega\Big(\frac{d}{\varepsilon^2 \varrho^2}\Big)$ is shown in Ullman (2018).

As in the proof of Theorem 1, we can use Lemma 23 to bound the jth order term by $(e^{\varrho}-1)^2\frac{\varepsilon^2}{2d}(\varepsilon^2/2)^{j-1}$. And thus, summing over all terms, we get

$$I(Z \wedge Y_t) \le (e^{\varrho}-1)^2\frac{\varepsilon^2}{2d}\sum_{j=1}^{\infty}(\varepsilon/2)^{2(j-1)} \le (e^{\varrho}-1)^2\frac{\varepsilon^2}{d}.$$

Since $I(Z \wedge Y^n) \le \sum_{t=1}^n I(Z \wedge Y_t)$, we conclude the proof using Theorem 15, thus establishing Theorem 21.

Upper and lower bounds for the block-sparse case. For the lower bound part, similar to the above derivation, we can change the proof of Theorem 14 by applying Theorem 23 whenever Theorem 7 is applied and result in a bound which replaces $2(\ln 2)\ell$ by $(e^{\rho}-1)^2$. To get the corresponding upper bound, we notice that we just need to replace the ℓ_∞ mean estimation step (for the **Detection** phase) and the ℓ_2 mean estimation step (for the **Estimation** phase) with an LDP protocol, which already exists in the literature (e.g., see Acharya et al. (2020a)). We ignore the details here since it will mostly resemble the proof for the communication constrained case.

Appendix B. One-sparse noninteractive lower bound

In this section, we prove the following result for 1-sparse estimation under communication constraints.

Theorem 24 *Any ℓ-bit noninteractive protocol for mean estimation of 1-sparse product distributions over $\{\pm 1\}^d$ must have sample complexity $\Omega\Big(\frac{d\log d}{\varepsilon^2\ell}\Big)$.*

Proof Consider the following family of distributions. For $i \in [d]$, $\mathbf{p}_i$ is a product distribution over $\{\pm 1\}^d$ with mean $\theta_j = 2\varepsilon\mathbb{1}_{\{i=j\}}$ for $1 \le j \le d$. Consider the generative process where we first sample J uniformly from $[d]$ and then each user observes one sample from $\mathbf{p}_J$ and follows the protocol, thus obtaining a tuple Y^n of messages.
By Fano's inequality, we have that for any 1-sparse estimation protocol the following must hold:

$$I(J \wedge Y^n) = \Omega(\log d).$$

It remains to provide an upper bound on $I(J \wedge Y^n)$. Since $(Y_1, Y_2, \ldots, Y_n)$ are independent conditioned on J, we have

$$I(J \wedge Y^n) \le \sum_{t=1}^n I(J \wedge Y_t)\,,$$

and therefore it suffices to show that $I(J \wedge Y_t) = O\Big(\frac{\varepsilon^2\ell}{d}\Big)$ for every $t \in [n]$. Using the same notation as in the proof of Lemma 11 we have

$$I(J \wedge Y_t) \le \mathbb{E}_J\big[\mathrm{KL}(W_t^{\mathbf{p}_J} \,||\, W_t^{\mathbf{u}})\big] \le \mathbb{E}_J\Big[\chi^2(W_t^{\mathbf{p}_J} \,||\, W_t^{\mathbf{u}})\Big].$$

Now, we can expand

$$\mathbb{E}_J\Big[\chi^2(W_t^{\mathbf{p}_J} \,||\, W_t^{\mathbf{u}})\Big] = \mathbb{E}_J\left[\sum_{y\in\mathcal{Y}}\frac{\big(\sum_x W(y\mid x)(\mathbf{p}_J(x)-\mathbf{u}(x))\big)^2}{\sum_x W(y\mid x)\mathbf{u}(x)}\right]$$

$$= \mathbb{E}_J\left[\sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)2\varepsilon X(J)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}\right]$$
$$= \frac{4\varepsilon^2}{d}\sum_{y\in\mathcal{Y}}\sum_{j\in[d]} \frac{\mathbb{E}_{\mathbf{u}}[W(y \mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}$$
$$= \frac{4\varepsilon^2}{d}\sum_{y\in\mathcal{Y}} \frac{\|\mathbb{E}_{\mathbf{u}}[W(y \mid X)X]\|_2^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}\,.$$

Note that the uniformly random vector X is 1-subgaussian. Hence using Lemma 7, we get $I(J \wedge Y_t) \le \frac{8(\ln 2)\varepsilon^2\ell}{d}$, which lets us conclude the proof as we get that n must satisfy $\Omega(\log d) = I(J \wedge Y^n) \le \sum_{t=1}^n I(J \wedge Y_t) \le n \cdot \frac{8(\ln 2)\varepsilon^2\ell}{d}$. ■

Appendix C. Adaptive sensing from m-dimensional measurements

In this section, we prove Theorem 4. The theorem states that there is an algorithm which estimates a sparse signal up to ℓ_2 accuracy ε with $m \cdot n = O\left(\frac{sd}{\varepsilon^2} + \frac{s}{\varepsilon^2}\log\frac{ed}{s}\right)$ noisy linear measurements, which is optimal as shown by the adaptive sensing lower bound from Arias-Castro et al. (2012) and the sparse mean estimation lower bound from the centralized case (see, *e.g.*, Wu (2020, Section 19)). Moreover, as shown in Raskutti et al. (2011), $m \cdot n = \Omega\left(\frac{sd}{\varepsilon^2}\log\frac{ed}{s}\right)$ measurement are required for a noninteractive protocol. All together, this demonstrates a separation between noninteractive compressed sensing and adaptive sensing.

Proof [Proof of Theorem 4.] We establish the result by a reduction to estimating the mean of a sparse product distribution over $\{\pm1\}^d$, which we have considered in previous sections.

Let e_i be the ith standard base vector in $\mathbb{R}^d$. Consider the family of selection matrices containing, for every $S \subseteq [d]$ of size $|S| = m$, the matrix $A_S := [e_i]_{i\in S}$. Then by Eq. (3), for any $S \subseteq [d]$, $Y \sim \mathcal{N}(x_S, I_m)$, where $x_S \in \mathbb{R}^m$ denotes the subvector of x restricted to coordinates indexed by S. Let $Y' = (\mathrm{sign}(Y(i)))_{i\in S}$. Then Y' has a product distribution such that, for every $i \in S$, $Y_i' \in \{\pm1\}$ has mean

$$\mu(i) := \mathbb{E}[Y(i)'] = 2\Pr[Y(i) > 0] - 1 = \mathrm{Erf}\left(\frac{x(i)}{\sqrt{2}}\right),$$

where Erf is the Gaussian error function. For $x \in [-1,+1]^d$, $\mu \in [-\mathrm{Erf}(1/\sqrt{2}), \mathrm{Erf}(1/\sqrt{2})] \subset [-1,+1]^d$. We will rely on the following lemma from Acharya et al. (2020a), which states that a good estimate for μ is also a good estimate for x.

Lemma 25 (Acharya et al. (2020a, Lemma 7)) *For $\widehat{\mu} \in [-\eta,\eta]^d$, define $\widehat{x} \in [-1,1]^d$ by $\widehat{x}(i) := \sqrt{2}\,\mathrm{Erf}^{-1}(\widehat{\mu}(i))$, for all $i \in [d]$. Then $\|\widehat{x} - x\|_2 \le \sqrt{\frac{e\pi}{2}} \cdot \|\mu - \widehat{\mu}\|_2$.*

It only remains to establish an upper bound on estimating the mean of a product distribution over $\{\pm1\}^d$ by observing a subset of m coordinates from each sample (in particular, this is a more restricted constraint than ℓ-bit communication, where the message is not restricted to consist of bits of the original sample). Nonetheless, in the protocol in Acharya et al. (2020a) which achieves Theorem 8, each user does actually send ℓ coordinates of the observed sample, meaning that it can be directly applied here by setting $\ell = m$. Plugging m for ℓ in Theorem 8, we get the desired bound. ■

Appendix D. Missing proofs in Sections 2 and 3

We now provide the proofs of the two inequalities used in Sections 2 and 3, respectively.

Proof [Proof of Eq. (5)] We can rewrite and bound the mutual information as

$$I(Z \wedge Y_t) = \mathbb{E}_Z\big[\mathrm{KL}(W_t^{\mathbf{p}_Z} \mid\mid W_t^{\mathbf{u}})\big] \le \mathbb{E}_Z\Big[\chi^2(W_t^{\mathbf{p}_Z} \mid\mid W_t^{\mathbf{u}})\Big]\,.$$

We drop the subscript t from W_t when it is clear from context. Using the definition of chi-square divergence and Eq. (4), for X, X' generated i.i.d. from $\mathbf{u}$, we have

$$\begin{aligned}
&\mathbb{E}_Z\Big[\chi^2(W_t^{\mathbf{p}_Z} \mid\mid W_t^{\mathbf{u}})\Big] \\
&\quad= \mathbb{E}_Z\left[\sum_{y\in\mathcal{Y}} \frac{\left(\sum_x W(y \mid x)(\mathbf{p}_Z(x) - \mathbf{u}(x))\right)^2}{\sum_x W(y \mid x)\mathbf{u}(x)}\right] \\
&\quad= \sum_{y\in\mathcal{Y}} \mathbb{E}_Z\left[\frac{\mathbb{E}_{\mathbf{u}}\Big[W(y \mid X)\Big(\prod_{i=1}^d (1+\gamma Z(i)X(i)) - 1\Big)\Big]^2}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}\right] \\
&\quad= \sum_{y\in\mathcal{Y}} \mathbb{E}_Z\left[\frac{\mathbb{E}_{X,X'\sim\mathbf{u}}\Big[W(y \mid X)W(y \mid X')\Big(\prod_{i=1}^d (1+\gamma Z(i)X(i)) - 1\Big)\Big(\prod_{i=1}^d (1+\gamma Z(i)X(i)') - 1\Big)\Big]}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}\right],
\end{aligned}$$

where we recall that $\gamma = \varepsilon/\sqrt{s}$. Note that since $\mathbb{E}_Z[Z(i)] = 0$ and $\mathbb{E}_Z[Z(i)^2] = \frac{s}{2d} = \sigma^2$ for all $i \in [d]$ and the $Z(i)$'s are independent, we further obtain that

$$\begin{aligned}
&\mathbb{E}_Z\left[\left(\prod_{i=1}^d (1+\gamma X(i)Z(i)) - 1\right)\left(\prod_{i=1}^d (1+\gamma X(i)'Z(i)) - 1\right)\right] \\
&= \mathbb{E}_Z\left[\prod_{i=1}^d (1+\gamma Z(i)X(i))(1+\gamma Z(i)X(i)')\right] - 2\mathbb{E}_Z\left[\prod_{i=1}^d (1+\gamma Z(i)X(i))\right] + 1 \\
&= \prod_{i=1}^d (1+\sigma^2\gamma^2 X(i)X(i)') - 1\,.
\end{aligned}$$

Plugging this into the previous expression, we get

$$\begin{aligned}
\mathbb{E}_Z\Big[\chi^2(W_t^{\mathbf{p}_Z} \mid\mid W_t^{\mathbf{u}})\Big] &= \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X,X'\sim\mathbf{u}}\Big[W(y \mid X)W(y \mid X')\Big(\prod_{i=1}^d (1+\sigma^2\gamma^2 X(i)X(i)') - 1\Big)\Big]}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} \\
&= \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X,X'\sim\mathbf{u}}\Big[W(y \mid X)W(y \mid X')\Big(\sum_{r=1}^d \sum_{B\subseteq[d],|B|=r} \sigma^{2r}\gamma^{2r} \prod_{i\in B} X(i)X(i)'\Big)\Big]}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} \\
&= \sum_{y\in\mathcal{Y}} \Big(\sigma^2\gamma^2 \sum_{i\in[d]} \frac{\mathbb{E}_{X,X'\sim\mathbf{u}}[W(y \mid X)W(y \mid X')X(i)X(i)']}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]} \\
&\qquad + \sum_{r=2}^d \sum_{\substack{B\subseteq[d] \\ |B|=r}} \sigma^{2r}\gamma^{2r} \frac{\mathbb{E}_{X,X'\sim\mathbf{u}}[W(y \mid X)W(y \mid X') \prod_{i\in B} X(i)X(i)']}{\mathbb{E}_{\mathbf{u}}[W(y \mid X)]}\Big)
\end{aligned}$$

$$= \sum_{y\in\mathcal{Y}}\Big(\sigma^2\gamma^2 \sum_{i\in[d]} \frac{\mathbb{E}_{\mathbf{u}}[W(y\mid X)X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y\mid X)]} + \sum_{r=2}^{d}\sigma^{2r}\gamma^{2r}\sum_{\substack{B\subseteq[d]\\|B|=r}} \frac{\mathbb{E}_{\mathbf{u}}[W(y\mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y\mid X)]}\Big),$$

which is the inequality we wanted to establish. ■

Proof [Proof of Eq. (8)] For the rth order term, we have:

$$\begin{aligned}
\sum_{\substack{B\subseteq[d]\\|B|=r}} \frac{\mathbb{E}_{\mathbf{u}}[W(y\mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y\mid X)]} &= \sum_{\substack{B'\subseteq[d]\\|B'|=r-1}}\sum_{j\notin B'} \frac{\mathbb{E}_{\mathbf{u}}[W(y\mid X)X(j)\prod_{i\in B'}X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y\mid X)]}\\
&\le \sum_{\substack{B'\subseteq[d]\\|B'|=r-1}}\sum_{j\notin B'} \frac{\sum_{x\in\{\pm1\}^{r-1}}\frac{1}{2^{r-1}}\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)X(j)\prod_{i\in B'}X(i)]^2}{\sum_{x\in\{\pm1\}^{r-1}}\frac{1}{2^{r-1}}\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)]}\\
&\le \sum_{\substack{B'\subseteq[d]\\|B'|=r-1}}\sum_{x\in\{\pm1\}^{r-1}}\sum_{j\notin B'} \frac{\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)X(j)]^2}{\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)]}\\
&= \sum_{\substack{B'\subseteq[d]\\|B'|=r-1}}\sum_{x\in\{\pm1\}^{r-1}} \frac{\|\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)X_{-B'}]\|_2^2}{\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)]},
\end{aligned}$$

where $X_{B'}$ denotes the sub-vector of X indexed by elements in B' and $X_{-B'}$ denotes the sub-vector without elements in B'. Hence the rth order term can be bounded by

$$\sigma^{2r}\gamma^{2r}\sum_{y\in\mathcal{Y}}\sum_{\substack{B\subseteq[d]\\|B|=r}} \frac{\mathbb{E}_{\mathbf{u}}[W(y\mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_{\mathbf{u}}[W(y\mid X)]} \le \left(\frac{\varepsilon^2}{2d}\right)^r \sum_{\substack{B'\subseteq[d]\\|B'|=r-1}}\sum_{x\in\{\pm1\}^{r-1}}\sum_{y\in\mathcal{Y}} \frac{\|\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)X_{-B'}]\|_2^2}{\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)]}$$

Invoking Theorem 7, we get for all x, B',

$$\sum_{y\in\mathcal{Y}} \frac{\|\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)X_{-B'}]\|_2^2}{\mathbb{E}_{\mathbf{u}|X_{B'}=x}[W(y\mid X)]} \le 2(\ln 2)\ell.$$

Hence we can further bound the rth order term by

$$\left(\frac{\varepsilon^2}{2d}\right)^r \binom{d}{r-1}\cdot 2^{r-1}\cdot 2(\ln 2)\cdot\ell \le \frac{\ell\varepsilon^2}{d}\cdot(\ln 2)(\varepsilon^2)^{r-1},$$

where we use $\binom{d}{r-1} \le d^{r-1}$. ■

Proof [Proof of Eq. (13)] Let $\mathbf{u}$ denote the uniform distribution over $\{\pm1\}^d$, and $\gamma := \varepsilon/\sqrt{s}$. For all $j\in[b]$, we have

$$I(Z\wedge Y_t \mid J=j)$$

$$
\begin{aligned}
&\le \mathbb{E}_Z[\mathrm{KL}(W^{\mathbf{p}_{Z,j}} \,||\, W^{\mathbf{u}})] \\
&\le \mathbb{E}_Z\left[\chi^2(W^{\mathbf{p}_{Z,j}} \,||\, W^{\mathbf{u}})\right] \\
&= \mathbb{E}_Z\left[\sum_{y\in\mathcal{Y}} \frac{\left(\sum_x W(y \mid X)(\mathbf{p}_{Z,j}(x) - \mathbf{u}(x))\right)^2}{\sum_x W(y \mid X)\mathbf{u}(x)}\right] \\
&= \sum_{y\in\mathcal{Y}} \mathbb{E}_Z\left[\frac{\mathbb{E}_X\left[W(y \mid X)\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i)) - 1\right)\right]^2}{\mathbb{E}_X[W(y \mid X)]}\right] \\
&= \sum_{y\in\mathcal{Y}} \mathbb{E}_Z\left[\frac{\mathbb{E}_{X,X'}\left[W(y \mid X)W(y \mid X')\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i)) - 1\right)\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i') - 1\right)\right]}{\mathbb{E}_X[W(y \mid X)]}\right] \\
&= \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X,X'}\left[W(y \mid X)W(y \mid X')\mathbb{E}_Z\left[\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i)) - 1\right)\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i') - 1\right)\right]\right]}{\mathbb{E}_X[W(y \mid X)]},
\end{aligned}
\tag{18}
$$

where $X, X' \sim \mathbf{u}$ are independent. Note that since $\mathbb{E}_Z[Z(i)] = 0$ and $\mathbb{E}_Z\left[Z(i)^2\right] = 1$ for all $i \in [d]$ and the $Z(i)$'s are independent, we have

$$
\begin{aligned}
&\mathbb{E}_Z\left[\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i)) - 1\right)\left(\prod_{i\in B_j}(1+\gamma Z(i)X(i') - 1\right)\right] \\
&\qquad = \prod_{i\in B_j}\mathbb{E}_Z[(1+\gamma Z(i)X(i))(1+\gamma Z(i)X(i'))] - 1 = \prod_{i\in B_j}(1+\gamma^2 X(i)X(i')) - 1\,.
\end{aligned}
$$

Plugging this into Eq. (18), we obtain

$$
\begin{aligned}
I(Z \wedge Y_t \mid J = j) &\le \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X,X'}\left[W(y \mid X)W(y \mid X')\left(\prod_{i\in B_j}(1+\gamma^2 X(i)X(i')) - 1\right)\right]}{\mathbb{E}_X[W(y \mid X)]} \\
&= \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X,X'}\left[W(y \mid X)W(y \mid X')\left(\sum_{r=1}^{s}\sum_{B\subseteq B_j,|B|=r}\gamma^{2r}\prod_{i\in B}X(i)X(i')\right)\right]}{\mathbb{E}_X[W(y \mid X)]} \\
&= \sum_{y\in\mathcal{Y}}\left(\gamma^2\sum_{i\in B_j}\frac{\mathbb{E}_{X,X'}[W(y \mid X)W(y \mid X')X(i)X(i')]}{\mathbb{E}_X[W(y \mid X)]}\right. \\
&\qquad\qquad \left. + \sum_{r=2}^{s}\sum_{\substack{B\subseteq B_j\\|B|=r}}\gamma^{2r}\frac{\mathbb{E}_{X,X'}[W(y \mid X)W(y \mid X')\prod_{i\in B}X(i)X(i')]}{\mathbb{E}_X[W(y \mid X)]}\right) \\
&= \sum_{y\in\mathcal{Y}}\left(\gamma^2\sum_{i\in B_j}\frac{\mathbb{E}_X[W(y \mid X)X(i)]^2}{\mathbb{E}_X[W(y \mid X)]} + \sum_{r=2}^{s}\gamma^{2r}\sum_{\substack{B\subseteq B_j\\|B|=r}}\frac{\mathbb{E}_X[W(y \mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_X[W(y \mid X)]}\right).
\end{aligned}
$$

Summing over all the blocks, we get

$$\sum_{j\in[b]} I(Z \wedge Y_t \mid J=j) \leq \sum_{j\in[b]}\sum_{y\in\mathcal{Y}} \Big(\gamma^2 \sum_{i\in B_j} \frac{\mathbb{E}_X[W(y\mid X)X(i)]^2}{\mathbb{E}_X[W(y\mid X)]} + \sum_{r=2}^{s}\gamma^{2r}\sum_{\substack{B\subseteq B_j\\|B|=r}} \frac{\mathbb{E}_X[W(y\mid X)\prod_{i\in B}X(i)]^2}{\mathbb{E}_X[W(y\mid X)]}\Big).$$

which is what we set out to prove. ■