Information-theoretic lower bounds for distributed statistical estimation with communication constraints

Yuchen Zhang[1] **John C. Duchi**[1] **Michael I. Jordan**[1,2] **Martin J. Wainwright**[1,2]
[1]Department of Electrical Engineering and Computer Science and [2]Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
{yuczhang,jduchi,jordan,wainwrig}@eecs.berkeley.edu

Abstract

We establish lower bounds on minimax risks for distributed statistical estimation under a communication budget. Such lower bounds reveal the minimum amount of communication required by any procedure to achieve the centralized minimax-optimal rates for statistical estimation. We study two classes of protocols: one in which machines send messages independently, and a second allowing for interactive communication. We establish lower bounds for several problems, including various types of location models, as well as for parameter estimation in regression models.

1 Introduction

Rapid growth in the size and scale of datasets has fueled increasing interest in statistical estimation in distributed settings [see, e.g., 5, 23, 7, 9, 17, 2]. Modern data sets are often too large to be stored on a single machine, so that it is natural to consider methods that involve multiple machines, each assigned a smaller subset of the full dataset. An essential design parameter in such methods is the amount of communication required between machines or chips. Bandwidth limitations on network and inter-chip communication often impose significant bottlenecks on algorithmic efficiency.

The focus of the current paper is the communication complexity of various classes of statistical estimation problems. More formally, suppose that we are interested in estimating the parameter θ of some unknown distribution P, based on a dataset of N i.i.d. samples. In the classical setting, one considers *centralized estimators* that have access to all N samples, and for a given estimation problem, the optimal performance over all centralized schemes can be characterized by the minimax rate. By way of contrast, in the distributed setting, one is given m different machines, and each machine is assigned a subset of samples of size $n = \lfloor \frac{N}{m} \rfloor$. Each machine is allowed to perform arbitrary operations on its own subset of data, and then communicate results of these intermediate computations to the other processors, or to a central fusion node. In this paper, we try to answer the following question: what is the minimal number of bits that must be exchanged in order to achieve the optimal estimation error achievable by centralized schemes?

There is a substantial literature on communication complexity in many settings, including function computation in theoretical computer science (e.g., [21, 1, 13]), decentralized detection and estimation (e.g., [18, 16, 15]) and information theory [11]. For instance, Luo [15] considers architectures in which machines may send only a single bit to a centralized processor; for certain problems, he shows that if each machine receives a single one-dimensional sample, it is possible to achieve the optimal centralized rate up to constant factors. Among other contributions, Balcan et al. [2] provide lower bounds for Probably Approximately Correct (PAC) learning in the distributed setting; however, their stated lower bounds do not involve the number of machines. In contrast, our work focuses on scaling issues, both in terms of the number of machines as well as the dimensionality of the underlying data, and formalizes the problem in terms of statistical minimax theory.

More precisely, we study the following problem: given a budget B of the total number of bits that may be communicated from the m distributed datasets, what is the minimax risk of any estimator based on the communicated messages? While there is a rich literature connecting information-theoretic techniques with the risk of statistical estimators (e.g. [12, 22, 20, 19]), little of it characterizes the effects of limiting communication. In this paper, we present some minimax lower bounds for distributed statistical estimation. By comparing our lower bounds with results in statistical estimation, we can identify the minimal communication cost that a distributed estimator must pay to have performance comparable to classical centralized estimators. Moreover, we show how to leverage recent work [23] so as to achieve these fundamental limits.

2 Problem setting and notation

We begin with a formal description of the statistical estimation problems considered here. Let $\mathcal{P}$ denote a family of distributions and let $\theta : \mathcal{P} \to \Theta \subseteq \mathbb{R}^d$ denote a function defined on $\mathcal{P}$. A canonical example throughout the paper is the problem of mean estimation, in which $\theta(P) = \mathbb{E}_P[X]$. Suppose that, for some fixed but unknown member P of $\mathcal{P}$, there are m sets of data stored on individual machines, where each subset $X^{(i)}$ is an i.i.d. sample of size n from the unknown distribution P.[1] Given this distributed collection of local data sets, our goal is to estimate $\theta(P)$ based on the m samples $X^{(1)}, \ldots, X^{(m)}$, but using limited communication.

We consider a class of distributed protocols Π, in which at each round $t = 1, 2, \ldots$, machine i sends a message $Y_{t,i}$ that is a measurable function of the local data $X^{(i)}$, and potentially of past messages. It is convenient to model this message as being sent to a central fusion center. Let $\overline{Y}_t = \{Y_{t,i}\}_{i \in [m]}$ denote the collection of all messages sent at round t. Given a total of T rounds, the protocol Π collects the sequence $(\overline{Y}_1, \ldots, \overline{Y}_T)$, and constructs an estimator $\widehat{\theta} := \widehat{\theta}(\overline{Y}_1, \ldots, \overline{Y}_T)$. The length $L_{t,i}$ of message $Y_{t,i}$ is the minimal number of bits required to encode it, and the total $L = \sum_{t=1}^{T} \sum_{i=1}^{m} L_{t,i}$ of all messages sent corresponds to the *total communication cost* of the protocol. Note that the communication cost is a random variable, since the length of the messages may depend on the data, and the protocol may introduce auxiliary randomness.

It is useful to distinguish two different classes, namely *independent* versus *interactive* protocols. An independent protocol Π is based on a single round ($T = 1$) of communication, in which machine i sends message $Y_{1,i}$ to the fusion center. Since there are no past messages, the message $Y_{1,i}$ can depend only on the local sample $X^{(i)}$. Given a family $\mathcal{P}$, the class of independent protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\mathrm{ind}}(B, \mathcal{P}) = \left\{ \text{independent protocols } \Pi \text{ such that} \quad \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\sum_{i=1}^{m} L_i \right] \leq B \right\}. \tag{1}$$

(For simplicity, we use Y_i to indicate the message sent from processor i and L_i to denote its length in the independent case.) It can be useful in some situations to have more granular control on the amount of communication, in particular by enforcing budgets on a per-machine basis. In such cases, we introduce the shorthand $B_{1:m} = (B_1, \ldots, B_m)$ and define

$$\mathcal{A}_{\mathrm{ind}}(B_{1:m}, \mathcal{P}) = \left\{ \text{independent protocols } \Pi \text{ such that} \quad \sup_{P \in \mathcal{P}} \mathbb{E}_P[L_i] \leq B_i \text{ for } i \in [m] \right\}. \tag{2}$$

In contrast to independent protocols, the class of interactive protocols allows for interaction at different stages of the message passing process. In particular, suppose that machine i sends message $Y_{t,i}$ to the fusion center at time t, who then posts it on a "public blackboard," where all machines can read $Y_{t,i}$. We think of this as a global broadcast system, which may be natural in settings in which processors have limited power or upstream capacity, but the centralized fusion center can send messages without limit. In the interactive setting, the message $Y_{t,i}$ should be viewed as a measurable function of the local data $X^{(i)}$, and the past messages $\overline{Y}_{1:t-1}$. The family of interactive protocols with budget $B \geq 0$ is given by

$$\mathcal{A}_{\mathrm{inter}}(B, \mathcal{P}) = \left\{ \text{interactive protocols } \Pi \text{ such that} \quad \sup_{P \in \mathcal{P}} \mathbb{E}_P[L] \leq B \right\}. \tag{3}$$

[1] Although we assume in this paper that every machine has the same amount of data, our technique generalizes easily to prove tight lower bounds for distinct data sizes on different machines.

We conclude this section by defining the minimax framework used throughout this paper. We wish to characterize the best achievable performance of estimators $\widehat{\theta}$ that are functions of only the messages $(\overline{Y}_1, \ldots, \overline{Y}_T)$. We measure the quality of a protocol and estimator $\widehat{\theta}$ by the mean-squared error

$$\mathbb{E}_{P,\Pi}\left[\|\widehat{\theta}(\overline{Y}_1, \ldots, \overline{Y}_T) - \theta(P)\|_2^2\right],$$

where the expectation is taken with respect to the protocol Π and the m i.i.d. samples $X^{(i)}$ of size n from distribution P. Given a class of distributions $\mathcal{P}$, parameter $\theta : \mathcal{P} \to \Theta$, and communication budget B, the *minimax risk for independent protocols* is

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B) := \inf_{\Pi \in \mathcal{A}_{\text{ind}}(B,\mathcal{P})} \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{P,\Pi}\left[\left\|\widehat{\theta}(Y_1, \ldots, Y_m) - \theta(P)\right\|_2^2\right]. \tag{4}$$

Here, the infimum is taken jointly over all independent procotols Π that satisfy the budget constraint B, and over all estimators $\widehat{\theta}$ that are measurable functions of the messages in the protocol. This minimax risk should also be understood to depend on both the number of machines m and the individual sample size n. The *minimax risk for interactive protocols*, denoted by $\mathfrak{M}^{\text{inter}}$, is defined analogously, where the infimum is instead taken over the class of interactive protocols. These communication-dependent minimax risks are the central objects in this paper: they provide a sharp characterization of the optimal rate of statistical estimation as a function of the communication budget B.

3 Main results

With our setup in place, we now turn to the statement of our main results, along with some discussion of their consequences.

3.1 Lower bound based on metric entropy

We begin with a general but relatively naive lower bound that depends only on the geometric structure of the parameter space, as captured by its metric entropy. In particular, given a subset $\Theta \subset \mathbb{R}^d$, we say $\{\theta^1, \ldots, \theta^K\}$ are δ-separated if $\left\|\theta^i - \theta^j\right\|_2 \geq \delta$ for $i \neq j$. We then define the *packing entropy* of Θ as

$$\log M_\Theta(\delta) := \log_2\left[\max\left\{K \in \mathbb{N} \mid \{\theta_1, \ldots, \theta^K\} \subset \Theta \text{ are } \delta\text{-separated}\right\}\right]. \tag{5}$$

The function $\theta \mapsto \log M_\Theta(\delta)$ is left-continuous and non-increasing in δ, so we may define the inverse function $\log M_\Theta^{-1}(B) := \sup\{\delta \mid \log M_\Theta(\delta) \geq B\}$.

Proposition 1 *For any family of distributions $\mathcal{P}$ and parameter set $\Theta = \theta(\mathcal{P})$, the interactive minimax risk is lower bounded as*

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \geq \left(\frac{1}{4}\log M_\Theta^{-1}(2B+2)\right)^2. \tag{6}$$

Of course, the same lower bound also holds for $\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B)$, since any independent protocol is a special case of an interactive protocol. Although Proposition 1 is a relatively generic statement, not exploiting any particular structure of the problem, it is in general unimprovable by more than constant factors, as the following example illustrates.

Example: Bounded mean estimation. Suppose that our goal is to estimate the mean $\theta = \theta(P)$ of a class of distributions $\mathcal{P}$ supported on the interval $[0, 1]$, so that $\Theta = \theta(\mathcal{P}) = [0, 1]$. Suppose that a single machine ($m = 1$) receives n i.i.d. observations X_i according to P. Since the packing entropy is lower bounded as $\log M_\Theta(\delta) \geq \log(1/\delta)$, the lower bound (6) implies

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B) \geq \mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \geq \frac{e^{-2}}{4}\, e^{-2B}.$$

Thus, setting $B = \frac{1}{2}\log n$ yields the lower bound $\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}([0,1]), B) \geq \frac{e^{-2}}{4n}$. This lower bound is sharp up to the constant pre-factor, since it can be achieved by a simple method. Given its n observations, the single machine can compute the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Since the sample mean lies in the interval $[0, 1]$, it can be quantized to accuracy $1/n$ using $\log(n)$ bits, and this quantized version $\widehat{\theta}$ can be transmitted. A straightforward calculation shows that $\mathbb{E}[(\widehat{\theta} - \theta)^2] \leq \frac{2}{n}$, so Proposition 1 yields an order-optimal bound in this case.

3.2 Multi-machine settings

We now turn to the more interesting multi-machine setting ($m > 1$). Let us study how the *budget* B—meaning the of bits required to achieve the minimax rate—scales with the number of machines m. We begin by considering the uniform location family $\mathcal{U} = \{P_\theta,\ \theta \in [-1, 1]\}$, where P_θ is the uniform distribution on the interval $[\theta - 1,\ \theta + 1]$. For this problem, a direct application of Proposition 1 gives a nearly sharp result.

Corollary 1 *Consider the uniform location family $\mathcal{U}$ with n i.i.d. observations per machine:*

(a) Whenever the communication budget is upper bounded as $B \leq \log(mn)$, there is a universal constant c such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{U}, B) \geq \frac{c}{(mn)^2}.$$

(b) Conversely, given a budget of $B = \big[2 + 2\ln m\big] \log(mn)$ bits, there is a universal constant c' such that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{U}, B) \leq \frac{c'}{(mn)^2}.$$

If each of m machines receives n observations, we have a total sample size of mn, so the minimax rate over all centralized procedures scales as $1/(mn)^2$ (for instance, see [14]). Consequently, Corollary 1(b) shows that the number of bits required to achieve the centralized rate has only *logarithmic* dependence on the number m of machines. Part (a) shows that this logarithmic dependence on m is unavoidable.

It is natural to wonder whether such logarithmic dependence holds more generally. The following result shows that it does not: for some problems, the dependence on m must be (nearly) linear. In particular, we consider estimation in a normal location family model, where each machine receives an i.i.d. sample of size n from a normal distribution $\mathsf{N}(\theta, \sigma^2)$ with unknown mean θ.

Theorem 1 *For the univariate normal family $\mathcal{N} = \{\mathsf{N}(\theta, \sigma^2) \mid \theta \in [-1, 1]\}$, there is a universal constant c such that*

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{N}, B) \geq c\, \frac{\sigma^2}{mn}\ \min\left\{\frac{mn}{\sigma^2},\ \frac{m}{\log m},\ \frac{m}{B\ \log m}\right\}. \tag{7}$$

The centralized minimax rate for estimating a univariate normal mean based on mn observations is $\frac{\sigma^2}{mn}$; consequently, the lower bound (7) shows that at least $B = \Omega\big(\frac{m}{\log m}\big)$ bits are required for a decentralized procedure to match the centralized rate in this case. This type of scaling is dramatically different than the logarithmic scaling for the uniform family, showing that establishing sharp communication-based lower bounds requires careful study of the underlying family of distributions.

3.3 Independent protocols in multi-machine settings

Departing from the interactive setting, in this section we focus on independent protocols, providing somewhat more general results than those for interactive protocols. We first provide lower bounds for the problem of mean estimation in the parameter for a *d-dimensional normal location family*

$$\mathcal{N}_d = \{\mathsf{N}(\theta, \sigma^2 I_{d\times d}) \mid \theta \in \Theta = [-1, 1]^d\}, \tag{8}$$

Theorem 2 *For $i = 1, \ldots, m$, assume that each machine has communication budget B_i, and receives an i.i.d. sample of size n from a distribution $P \in \mathcal{N}_d$. There exists a universal (numerical) constant c such that*

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{N}_d, B_{1:m}) \geq c\frac{\sigma^2 d}{mn}\ \min\left\{\frac{mn}{\sigma^2},\ \frac{m}{\log m},\ \frac{m}{\big(\sum_{i=1}^m \min\{1, \frac{B_i}{d}\}\big)\ \log m}\right\}. \tag{9}$$

Given centralized access to the full mn-sized sample, a reasonable procedure would be to compute the sample mean, leading to an estimate with mean-squared error $\frac{\sigma^2 d}{mn}$, which is minimax optimal.

Consequently, Theorem 2 shows that to achieve an order-optimal mean-squared error, the total number of bits communicated must (nearly) scale with the product of the dimension d and number of machines m, that is, as $dm/\log m$. If we ignore logarithmic factors, this lower bound is achievable by a simple procedure: each machine computes the sample mean of its local data and quantizes each coordinate to precision σ^2/n using $\mathcal{O}(d\,\log(n/\sigma^2))$ bits. These quantized sample averages are communicated to the fusion center using $B = \mathcal{O}(dm\log(n/\sigma^2))$ total bits. The fusion center averages them, obtaining an estimate with mean-squared error of optimal order $\sigma^2 d/(mn)$ as required.

We finish this section by presenting a result that is sharp up to numerical constant prefactors. It is a minimax lower bound for mean estimation over the family $\mathcal{P}_d = \{P \text{ supported on } [-1,1]^d\}$.

Proposition 2 *Assume that each of m machines receives a single sample ($n = 1$) from a distribution in $\mathcal{P}_d$. There exists a universal (numerical) constant c such that*

$$\mathfrak{M}^{\mathrm{ind}}(\theta, \mathcal{P}^d, B_{1:m}) \geq c\,\frac{d}{m}\;\min\left\{m,\;\frac{m}{\sum_{i=1}^m \min\{1, \frac{B_i}{d}\}}\right\}, \tag{10}$$

where B_i is the budget for machine i.

The standard minimax rate for d-dimensional mean estimation scales as d/m. The lower bound (10) shows that in order to achieve this scaling, we must have $\sum_{i=1}^m \min\{1, \frac{B_i}{d}\} \gtrsim m$, showing that each machine must send $B_i \gtrsim d$ bits.

Moreover, this lower bound is achievable by a simple scheme. Suppose that machine i receives a d-dimensional vector $X_i \in [-1,1]^d$. Based on X_i, it generates a Bernoulli random vector $Z_i = (Z_{i1}, \ldots, Z_{id})$ with $Z_{ij} \in \{0,1\}$ taking the value 1 with probability $(1 + X_{ij})/2$, independently across coordinates. Machine i uses d bits to send the vector $Z_i \in \{0,1\}^d$ to the fusion center. The fusion center then computes the average $\widehat{\theta} = \frac{1}{m}\sum_{i=1}^m (2Z_i - 1)$. This average is unbiased, and its expected squared error is bounded by d/m.

4 Consequences for regression

In this section, we turn to identifying the minimax rates for a pair of important estimation problems: linear regression and probit regression.

4.1 Linear regression

We consider a distributed instantiation of linear regression with fixed design matrices. Concretely, suppose that each of m machines has stored a fixed design matrix $A^{(i)} \in \mathbb{R}^{n\times d}$ and then observes a response vector $b^{(i)} \in \mathbb{R}^d$ from the standard linear regression model

$$b^{(i)} = A^{(i)}\theta + \varepsilon^{(i)}, \tag{11}$$

where $\varepsilon^{(i)} \sim \mathsf{N}(0, \sigma^2 I_{n\times n})$ is a noise vector. Our goal is to estimate unknown regression vector $\theta \in \Theta = [-1,1]^d$, shared across all machines, in a distributed manner, To state our result, we assume uniform upper and lower bounds on the eigenvalues of the rescaled design matrices, namely

$$0 < \lambda_{\min} \leq \min_{i\in\{1,\ldots,m\}} \frac{\eta_{\min}(A^{(i)})}{\sqrt{n}} \quad \text{and} \quad \max_{i\in\{1,\ldots,m\}} \frac{\eta_{\max}(A^{(i)})}{\sqrt{n}} \leq \lambda_{\max}. \tag{12}$$

Corollary 2 *Consider an instance of the linear regression model* (11) *under condition* (12)*.*

(a) Then there is a universal positive constant c such that

$$\mathfrak{M}^{\mathrm{ind}}(\theta, \mathcal{P}, B_{1:m}) \geq c\,\frac{\sigma^2 d}{mn}\;\min\left\{\frac{mn}{\sigma^2},\;\frac{m}{\lambda_{\max}^2 \log m},\;\frac{m}{\lambda_{\max}^2\big(\sum_{i=1}^m \min\{1, \frac{B_i}{d}\}\big)\log m}\right\}.$$

(b) Conversely, given budgets $B_i \geq d\log(mn)$ for $i = 1, \ldots, m$, there is a universal constant c' such that

$$\mathfrak{M}^{\mathrm{ind}}(\theta, \mathcal{P}, B_{1:m}) \leq \frac{c'}{\lambda_{\min}^2}\,\frac{\sigma^2 d}{mn}.$$

It is a classical fact (e.g. [14]) that the minimax rate for d-dimensional linear regression scales as $d\sigma^2/(nm)$. Part (a) of Corollary 2 shows this optimal rate is attainable only if the budget B_i at each machine is of the order $d/\log(m)$, meaning that the total budget $B = \sum_{i=1}^m B_i$ must grow as $\frac{dm}{\log m}$. Part (b) of the corollary shows that the minimax rate is achievable with budgets that match the lower bound up to logarithmic factors.

Proof: The proof of part (b) follows from techniques of Zhang et al. [23], who show that solving each regression problem separately and then performing a form of approximate averaging, in which each machine uses $B_i = d\log(mn)$ bits, achieves the minimax rate up to constant prefactors.

To prove part (a), we show that solving an arbitrary Gaussian mean estimation problem can be reduced to solving a specially constructed linear regression problem. This reduction allows us to apply the lower bound from Theorem 2. Given $\theta \in \Theta$, consider the Gaussian mean model

$$X^{(i)} = \theta + w^{(i)}, \;\; \text{where} \;\; w^{(i)} \sim \mathsf{N}\left(0, \frac{\sigma^2}{\lambda_{\max}^2 n} I_{d\times d}\right).$$

Each machine i has its own design matrix $A^{(i)}$, and we use it to construct a response vector $b^{(i)} \in \mathbb{R}^n$. Since $\eta_{\max}(A^{(i)}/\sqrt{n}) \leq \lambda_{\max}$, the matrix $\Sigma^{(i)} := \sigma^2 I_{n\times n} - \frac{\sigma^2}{\lambda_{\max}^2 n} A^{(i)}(A^{(i)})^\top$ is positive semidefinite. Consequently, we may form a response vector via

$$b^{(i)} = A^{(i)} X^{(i)} + z^{(i)}, \quad z^{(i)} \sim \mathsf{N}\left(0, \Sigma^{(i)}\right) \text{ is drawn independently of } w^{(i)}. \tag{13}$$

The independence of $w^{(i)}$ and $z^{(i)}$ guarantees that $b^{(i)} \sim \mathsf{N}(A^{(i)}\theta, \sigma^2 I_{n\times n})$, so that the pair $(b^{(i)}, A^{(i)})$ is faithful to the regression model (11).

Now consider any protocol $\Pi \in \mathcal{A}_{\text{ind}}(B, \mathcal{P})$ that can solve any regression problem to within accuracy δ, so that $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \leq \delta^2$. By the previously described reduction, the protocol Π can also solve the mean estimation problem to accuracy δ, in particular via the pair $(A^{(i)}, b^{(i)})$ constructed via expression (13). Combined with this reduction, the corollary thus follows from Theorem 2. ■

4.2 Probit regression

We now turn to the problem of binary classification, in particular considering the probit regression model. As in the previous section, each of m machines has a fixed design matrix $A^{(i)} \in \mathbb{R}^{n\times d}$, where $A^{(i,k)}$ denotes the kth row of $A^{(i)}$. Machine i receives n binary responses $Z^{(i)} = (Z^{(i,1)}, \ldots, Z^{(i,n)})$, drawn from the conditional distribution

$$\mathbb{P}(Z^{(i,k)} = 1 \mid A^{(i,k)}, \theta) = \Phi(A^{(i,k)}\theta) \quad \text{for some fixed } \theta \in \Theta = [-1, 1]^d, \tag{14}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. The log-likelihood of the probit model (14) is concave [4, Exercise 3.54]. Under condition (12) on the design matrices, we have:

Corollary 3 *Consider the probit model* (14) *under condition* (12)*. Then*

(a) There is a universal constant c such that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \geq c\frac{d}{mn} \; \min\left\{mn, \; \frac{m}{\lambda_{\max}^2 \log m}, \; \frac{m}{\lambda_{\max}^2 \left(\sum_{i=1}^m \min\{1, \frac{B_i}{d}\}\right) \log m}\right\}.$$

(b) Conversely, given budgets $B_i \geq d\log(mn)$ for $i = 1, \ldots, m$, there is a universal constant c' such that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \leq \frac{c'}{\lambda_{\min}^2} \frac{d}{mn}.$$

Proof: As in the previous case with linear regression, Zhang et al.'s study of distributed convex optimization [23] gives part (b): each machine solves the local probit regression separately, after which each machine sends $B_i = d\log(mn)$ bits to average its local solution.

To prove part (a), we show that linear regression problems can be solved via estimation in a specially constructed probit model. Consider an arbitrary $\theta \in \Theta$; assume we have a regression problem of the

form (11) with noise variance $\sigma^2 = 1$. We construct the binary responses for our probit regression $(Z^{(i,1)}, \ldots, Z^{(i,n)})$ by

$$Z^{(i,k)} = \begin{cases} 1 & \text{if } b^{(i,k)} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

By construction, we have $\mathbb{P}(Z^{(i,k)} = 1 \mid A^{(i)}, \theta) = \Phi(A^{(i,k)}\theta)$ as desired for our model (14). By inspection, any protocol $\Pi \in \mathcal{A}_{\text{ind}}(B, \mathcal{P})$ solving the probit regression problem provides an estimator with the same error (risk) as the original linear regression problem via the construction (15). Corollary 2 provides the desired lower bound. ■

5 Proof sketches for main results

We now give an outline of the proof of each of our main results (Theorems 1 and 2), providing a more detailed proof sketch for Proposition 2, since it displays techniques common to our arguments.

5.1 Broad outline

Most of our lower bounds follow the same basic strategy of reducing an estimation problem to a testing problem. Following this reduction, we then develop inequalities relating the probability of error in the test to the number of bits contained in the messages Y_i sent from each machine. Establishing these links is the most technically challenging aspect.

Our reduction from estimation to testing is somewhat more general than the classical reductions (e.g., [22, 20]), since we do not map the original estimation problem to a strict test, but rather a test that allows some errors. Let $\mathcal{V}$ denote an index set of finite cardinality, where $\nu \in \mathcal{V}$ indexes a family of probability distributions $\{P(\cdot \mid \nu)\}_{\nu \in \mathcal{V}}$. For each member of this family, associate with a parameter $\theta_\nu := \theta(P(\cdot \mid \nu)) \in \Theta$, where Θ denotes the parameter space. In our proofs applicable to d-dimensional problems, we set $\mathcal{V} = \{-1, 1\}^d$, and we index vectors θ_ν by $\nu \in \mathcal{V}$. Now, we sample V uniformly at random from $\mathcal{V}$. Conditional on $V = \nu$, we then sample X from a distribution $P_X(\cdot \mid V = \nu)$ satisfying $\theta_\nu := \theta(P_X(\cdot \mid \nu)) = \delta\nu$, where $\delta > 0$ is a fixed quantity that we control. We define $d_{\text{ham}}(\nu, \nu')$ to be the Hamming distance between $\nu, \nu' \in \mathcal{V}$. This construction gives

$$\|\theta_\nu - \theta_{\nu'}\|_2 = 2\delta\sqrt{d_{\text{ham}}(\nu, \nu')}.$$

Fixing $t \in \mathbb{R}$, the following lemma reduces the problem of estimating θ to finding a point $\nu \in \mathcal{V}$ within distance t of the random variable V. The result extends a result of Duchi and Wainwright [8]; for completeness we provide a proof in Appendix H.

Lemma 1 *Let V be uniformly sampled from $\mathcal{V}$. For any estimator $\widehat{\theta}$ and any $t \in \mathbb{R}$, we have*

$$\sup_{P \in \mathcal{P}} \mathbb{E}[\|\widehat{\theta} - \theta(P)\|_2^2] \geq \delta^2(\lfloor t \rfloor + 1) \inf_{\widehat{\nu}} \mathbb{P}\left(d_{\text{ham}}(\widehat{\nu}, V) > t\right),$$

where the infimum ranges over all testing functions.

Lemma 1 shows that minimax lower lower bound can be derived by showing that, for some $t > 0$ to be chosen, it is difficult to identify V within a radius of t. The following extension of Fano's inequality [8] can be used to control this type of error probability:

Lemma 2 *Let $V \to X \to \widehat{V}$ be a Markov chain, where V is uniform on $\mathcal{V}$. For any $t \in \mathbb{R}$, we have*

$$\mathbb{P}(d_{\text{ham}}(\widehat{V}, V) > t) \geq 1 - \frac{I(V; X) + \log 2}{\log \frac{|\mathcal{V}|}{N_t}},$$

where $N_t := \max_{\nu \in \mathcal{V}} |\{\nu' \in \mathcal{V} : d_{\text{ham}}(\nu, \nu') \leq t\}|$ is the size of the largest t-neighborhood in $\mathcal{V}$.

Lemma 2 allows flexibility in the application of the minimax bounds from Lemma 1. If there is a large set $\mathcal{V}$ for which it is easy to control $I(V; X)$, whereas neighborhoods in $\mathcal{V}$ are relatively small (i.e., N_t is small), then we can obtain sharp lower bounds.

In a distributed protocol, we have a Markov chain $V \to X \to Y$, where Y denotes the messages the different machines send. Based on the messages Y, we consider an arbitrary estimator $\widehat{\theta}(Y)$. For $0 \le t \le \lceil d/3 \rceil$, we have $N_t = \sum_{\tau=0}^{t} \binom{d}{\tau} \le 2\binom{d}{t}$. Since $\binom{d}{t} \le (de/t)^t$, for $t \le d/6$ we have

$$\log \frac{|\mathcal{V}|}{N_t} \ge d \log 2 - \log 2 \binom{d}{t} \ge d \log 2 - \frac{d}{6} \log(6e) - \log 2 = d \log \frac{2}{2^{1/d} \sqrt[6]{6e}} > \frac{d}{6}$$

for $d \ge 12$ (the case $d < 12$ can be checked directly). Thus, combining Lemma 1 and Lemma 2 (using the Markov chain $V \to X \to Y \to \widehat{\theta}$), we find that for $t = \lfloor d/6 \rfloor$,

$$\sup_{P \in \mathcal{P}} \mathbb{E}\left[\|\widehat{\theta}(Y) - \theta(P)\|_2^2 \right] \ge \delta^2 (\lfloor d/6 \rfloor + 1) \left(1 - \frac{I(Y;V) + \log 2}{d/6}\right). \tag{16}$$

With inequality (16) in hand, it then remains to upper bound the mutual information $I(Y;V)$, which is the main technical content of each of our results.

5.2 Proof sketch of Proposition 2

Following the general outline of the previous section, let V be uniform on $\mathcal{V} = \{-1, 1\}^d$. Letting $0 < \delta \le 1$ be a positive number, for $i \in [m]$ we independently sample $X^{(i)} \in \mathbb{R}^d$ according to

$$P(X_j^{(i)} = \nu_j \mid V = \nu) = \frac{1+\delta}{2} \quad \text{and} \quad P(X_j^{(i)} = -\nu_j \mid V = \nu) = \frac{1-\delta}{2}. \tag{17}$$

Under this distribution, we can give a sharp characterization of the mutual information $I(V; Y_i)$. In particular, we show in Appendix B that under the sampling distribution (17), there exists a numerical constant c such that

$$I(V; Y_i) \le c\delta^2 I(X^{(i)}; Y_i). \tag{18}$$

Since the random variable X takes discrete values, we have

$$I(X^{(i)}; Y_i) \le \min\{H(X^{(i)}), H(Y_i)\} \le \min\{d, H(Y_i)\}.$$

Since the expected length of message Y_i is bounded by B_i, Shannon's source coding theorem [6] implies that $H(Y_i) \le B_i$. In particular, inequality (18) establishes a link between the initial distribution (17) and the number of bits used to transmit information, that is,

$$I(V; Y_i) \le c\delta^2 \min\{d, B_i\}. \tag{19}$$

We can now apply the quantitative data processing inequality (19) in the bound (16). By the independence of the communication scheme, $I(V; Y_{1:m}) \le \sum_{i=1}^m I(V; Y_i)$, and thus inequality (16) simplifies to

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \ge \delta^2 (\lfloor d/6 \rfloor + 1) \left(1 - \frac{c\delta^2 \sum_{i=1}^m \min\{d, B_i\} + \log 2}{d/6}\right).$$

Assuming $d \ge 9$, so $1 - 6 \log 2/d > 1/2$, we see that choosing $\delta^2 = \min\{1, \frac{d}{24c \sum_{i=1}^m \min\{B_i, d\}}\}$ implies

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \ge \frac{\delta^2(\lfloor d/6 \rfloor + 1)}{4} = \frac{\lfloor d/6 \rfloor + 1}{4} \min\left\{1, \frac{d}{24c \sum_{i=1}^m \min\{B_i, d\}}\right\}.$$

Rearranging slightly gives the statement of the proposition.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback and comments. JCD was supported by a Facebook Graduate Fellowship. Our work was supported in part by the U.S. Army Research Laboratory, U.S. Army Research Office under grant number W911NF-11-1-0391, and Office of Naval Research MURI grant N00014-11-1-0688.

References

[1] H. Abelson. Lower bounds on information transfer in distributed computations. *Journal of the ACM*, 27(2):384–392, 1980.

[2] M.-F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity and privacy. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012. URL `http://arxiv.org/abs/1204.3514`.

[3] K. Ball. An elementary introduction to modern convex geometry. In S. Levy, editor, *Flavors of Geometry*, pages 1–58. MSRI Publications, 1997.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 2011.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition*. Wiley, 2006.

[7] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

[8] J. C. Duchi and M. J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv [cs.IT]*, to appear, 2013.

[9] J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3): 592–606, 2012.

[10] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. *arXiv:1302.3203 [math.ST]*, 2013. URL `http://arXiv.org/abs/1302.3203`.

[11] S. Han and S. Amari. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, 1998.

[12] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, 1981.

[13] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[14] E. L. Lehmann and G. Casella. *Theory of Point Estimation, Second Edition*. Springer, 1998.

[15] Z.-Q. Luo. Universal decentralized estimation in a bandwidth constrained sensor network. *IEEE Transactions on Information Theory*, 51(6):2210–2219, 2005.

[16] Z.-Q. Luo and J. N. Tsitsiklis. Data fusion with minimal communication. *IEEE Transactions on Information Theory*, 40(5):1551–1563, 1994.

[17] R. McDonald, K. Hall, and G. Mann. Distributed training strategies for the structured perceptron. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2010.

[18] J. N. Tsitsiklis. Decentralized detection. In *Advances in Signal Processing, Vol. 2*, pages 297–344. JAI Press, 1993.

[19] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[20] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[21] A. C.-C. Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on the Theory of Computing*, pages 209–213. ACM, 1979.

[22] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

[23] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems 26*, 2012.

Appendices

A Notation and proof setup

In these appendices, we provide the proofs of all our major results. Note that we prove the theorems out of the order in which they are presented: many of the theorems build on one another, so we present them in (rough) order of most basic to most complex. Before proceeding to the proofs proper, we give notation.

Notation in proofs

The distributed machines are indexed by $i \in \{1, \ldots, m\}$. For machine i, it receives local dataset D_i. If D_i contains multiple examples, we may denote the k-th example by $X^{(i,k)}$. If each example has more than one coordinate, then the j-th coordinate is represented by $X_j^{(i,k)}$.

For a random variable X, we let P_X denote the probability measure on X, so that $P_X(S) = P(X \in S)$, and we abuse notation by writing p_X for the probability mass function or density of X, depending on the situation, so that $p_X(x) = P(X = x)$ in the discrete case and denotes the density of X at x when p_X is a density. For discrete random variable X, we let $H(X) = -\sum_x p_X(x) \log p_X(x)$ denote the (Shannon) entropy, and for probability distributions P, Q on a set $\mathcal{X}$, with densities p, q with respect to a base measure μ, we write the KL-divergence as

$$D_{\rm kl}\left(P \| Q\right) := \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\mu(x).$$

The mutual information $I(X; Y)$ between random variables X and Y where Y has distribution P_Y is defined as

$$I(X; Y) := \mathbb{E}_{P_X}\left[D_{\rm kl}\left(P_Y(\cdot \mid X) \| P_Y(\cdot)\right)\right] = \int D_{\rm kl}\left(P_Y(\cdot \mid X = x) \| P_Y(\cdot)\right) dP_X(x).$$

Le Cam's method

In low-dimensional settings, it is sometimes difficult to apply our incarnation of Fano's inequality as outlined in Section 5.1. In these settings, we use a minimax lower bound based on a two-point family. In this setting, we let $\mathcal{V} = \{-1, 1\}$, and define $\theta_\nu = \theta(P_\nu)$ as usual. Then Le Cam's inequality (e.g. [22] or [19, Theorem 2.2]) guarantees that for V chosen uniformly as $V = 1$ or $V = -1$ we have

$$\inf_{\widehat{\nu}} \mathbb{P}(\widehat{\nu} \neq V) \geq \frac{1}{2} - \frac{1}{2} \left\|P_1 - P_{-1}\right\|_{\rm TV}.$$

As a consequence, if by construction $\theta_\nu = \delta\nu$, then Lemma 1 implies that

$$\inf_{\widehat{\theta}} \max_{P \in \{P_1, P_{-1}\}} \mathbb{E}[\|\widehat{\theta} - \theta(P)\|_2^2] \geq \delta^2 \left(\frac{1}{2} - \frac{1}{2} \left\|P_1 - P_{-1}\right\|_{\rm TV}\right). \tag{20}$$

We use arguments based on Le Cam's method (20) when the dimension d is small.

In addition, it will be useful to have a few simple upper bounds on the distance $\|P_1 - P_{-1}\|_{\rm TV}$. We claim that if we have the Markov chain $V \to Y$, for any random variable Y, then for V chosen uniformly in a set $\mathcal{V} = \{\nu, \nu'\}$,

$$\left\|P_Y(\cdot \mid V = \nu) - P_Y(\cdot \mid V = \nu')\right\|_{\rm TV}^2 \leq 2I(Y, V). \tag{21}$$

To see inequality (21), let P_ν be shorthand for $P_Y(\cdot \mid V = \nu)$. The triangle inequality implies that

$$\left\|P_\nu - P_{\nu'}\right\|_{\rm TV} \leq \left\|P_\nu - (1/2)(P_\nu + P_{\nu'})\right\|_{\rm TV} + \frac{1}{2} \left\|P_\nu - P_{\nu'}\right\|_{\rm TV},$$

and similarly swapping the roles of ν' and ν, whence

$$\left\|P_\nu - P_{\nu'}\right\|_{\rm TV} \leq 2 \min\{\left\|P_\nu - (1/2)(P_{\nu'} + P_\nu)\right\|_{\rm TV}, \left\|P_{\nu'} - (1/2)(P_{\nu'} + P_\nu)\right\|_{\rm TV}\}.$$

By Pinsker's inequality, we thus have the upper bound

$$\begin{aligned}\left\|P_\nu - P_{\nu'}\right\|_{\rm TV}^2 &\leq 2 \min\{D_{\rm kl}\left(P_\nu \| (1/2)(P_\nu + P_{\nu'})\right), D_{\rm kl}\left(P_{\nu'} \| (1/2)(P_\nu + P_{\nu'})\right)\} \\ &\leq D_{\rm kl}\left(P_\nu \| (1/2)(P_\nu + P_{\nu'})\right) + D_{\rm kl}\left(P_{\nu'} \| (1/2)(P_\nu + P_{\nu'})\right) = 2I(Y; V)\end{aligned}$$

by the definition of mutual information.

Tensorization of information

We also require a type of tensorization inequality in each of our proofs for independent protocols. When Y_i is constructed based only on $X^{(i)}$, we have

$$\begin{aligned} I(V; Y_{1:m}) = \sum_{i=1}^m I(V; Y_i \mid Y_{1:i-1}) &= \sum_{i=1}^m H(Y_i \mid Y_{1:i-1}) - H(Y_i \mid V, Y_{1:i-1}) \\ &\le \sum_{i=1}^m H(Y_i) - H(Y_i \mid V, Y_{1:i-1}) \\ &= \sum_{i=1}^m H(Y_i) - H(Y_i \mid V) = \sum_{i=1}^m I(V; Y_i) \end{aligned} \tag{22}$$

where we have used that conditioning reduces entropy and Y_i is conditionally independent of $Y_{1:i-1}$ given V.

B Proof of Proposition 2

The proof of this proposition follows the basic outline described in Section 5.

We first describe the distribution of the step $V \to X$. Given $\nu \in \mathcal{V}$, we assume that each machine i receives a d-dimensional sample $X^{(i)}$ with coordinates independently sampled according to

$$P(X_j = \nu_j \mid \nu) = \frac{1 + \delta\nu_j}{2} \quad \text{and} \quad P(X_j = -\nu_j \mid \nu) = \frac{1 - \delta\nu_j}{2}.$$

Let $\delta \le \frac{1}{4}$. Then $\theta_\nu = \mathbb{E}_\nu[X]$, and moreover we have the likelihood ratio bound

$$\frac{P(X_j \in S \mid \nu)}{P(X_j \in S \mid \nu')} \le \frac{1+\delta}{1-\delta} \le \exp\left(\frac{17}{8}\delta\right), \quad \text{and} \quad \exp\left(\frac{17}{4}\delta\right) \le 1 + 8\delta.$$

We now present a lemma that relates this ratio bound via a type of quantitative data processing inequality. The lemma is actually somewhat more general than what we require, and we prove it in Section B.1. The result is similar to recent results of Duchi et al. [10, Theorems 1 and 2], who show similar strong data processing inequalities in the context of privacy-preserving data analysis. Our proof, however, is different, as we have the Markov chain $V \to X \to Y$, and instead of a likelihood ratio bound on the channel $X \to Y$, we place a likelihood ratio bound on $V \to X$.

Lemma 3 *Let V be sampled uniformly at random from $\{-1, 1\}^d$. For any (i, j), assume that $X_j^{(i)}$ is independent of $\{X_{j'}^{(i)} : j' \neq j\} \cup \{V_{j'} : j' \neq j\}$ given V_j. Let $\mathbb{P}_{X_j}$ be the probability measure of $X_j^{(i)}$ and assume in addition that*

$$\sup_{S \in \sigma(X_j)} \frac{\mathbb{P}_{X_j}(S \mid V = \nu)}{\mathbb{P}_{X_j}(S \mid V = \nu')} \le \exp(\alpha).$$

Then

$$I(V; Y_i) \le 2(e^{2\alpha} - 1)^2 I(X^{(i)}; Y_i).$$

Lemma 3 provides a quantitative data processing inequality relating the mutual information in the channel $X^{(i)} \to Y_i$ to that in $V \to Y_i$. In particular, we find that

$$I(V; Y_i) \le 2\left(e^{(17/4)\delta} - 1\right)^2 I(X^{(i)}; Y_i) \le 128\delta^2 I(X^{(i)}; Y_i).$$

This is the claimed strong data processing inequality (18), which almost completes our proof. To complete the proof, note that $\theta(P(\cdot \mid V = \nu)) = \mathbb{E}[Y \mid \nu] = \delta\nu$. Recalling the tensorization inequality (22), we also have

$$I(V; Y_{1:m}) \le \sum_{i=1}^m I(V; Y_i) \le 128\delta^2 \sum_{i=1}^m I(Y_i; X^{(i)}).$$

The remainder of the proof we break into two cases: when $d \geq 9$ and when $d < 9$. For the case $d \geq 9$, our proof sketch in Section 5.2, beginning from inequality (18) with $c = 128$, completes the proof. When $d < 9$, we use a slightly different argument. By a reduction to smaller dimensions, we may assume without loss of generality that $d = 1$, and we set $\mathcal{V} = \{-1, 1\}$. In this case, Le Cam's method (20) coupled with the subsequent information inequality (21) implies that

$$\mathfrak{M}^{\rm ind}(\theta, \mathcal{P}, B_{1:m}) \geq \delta^2 \left(\frac{1}{2} - \frac{1}{2}\sqrt{2I(V; Y_{1:m})}\right). \tag{23}$$

Applying our previous bound $I(V; Y_{1:m}) \leq 128\delta^2 \sum_{i=1}^m I(Y_i; X^{(i)})$, and noting that $I(X^{(i)}; Y_i) \leq \min\{H(X^{(i)}), H(Y_i)\} \leq \min\{1, H(Y_i)\}$ since $X^{(i)} \in \{-1, 1\}$, we obtain

$$\mathfrak{M}^{\rm ind}(\theta, \mathcal{P}, B_{1:m}) \geq \delta^2 \left(\frac{1}{2} - 8\left(\delta^2 \sum_{i=1}^m \min\{1, H(Y_i)\}\right)^{\frac{1}{2}}\right).$$

Since $H(Y_i) \leq B_i$ by Shannon's source coding theorem [6], setting

$$\delta^2 = \min\left\{1, \frac{1}{400 \sum_{i=1}^m \min\{1, B_i\}}\right\}$$

completes the proof.

B.1 Proof of Lemma 3

Let $Y = Y_i$; we suppress the dependence on the index i (and similarly let $X = X^{(i)}$ denote a single fixed sample). We begin with the simple observation that, by the chain rule for mutual information,

$$I(V; Y) = \sum_{j=1}^d I(V_j; Y \mid V_{1:j-1}).$$

Using the definition of mutual information and non-negativity of the KL-divergence, we have

$$\begin{aligned} I(V_j; Y \mid V_{1:j-1}) &= \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\rm kl}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right) \mid V_{1:j-1}\right]\right] \\ &\leq \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\rm kl}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right)\right.\right. \\ &\quad \left.\left. + D_{\rm kl}\left(P_{V_j}(\cdot \mid V_{1:j-1}) \| P_{V_j}(\cdot \mid Y, V_{1:j-1})\right) \mid V_{1:j-1}\right]\right]. \end{aligned}$$

Now, we require an argument that builds off of a technical lemma we present in Appendix G, Lemma 8. We claim that Lemma 8 implies that

$$\begin{aligned} &|P(V_j = \nu_j \mid V_{1:j-1}, Y) - P(V_j = \nu_j \mid V_{1:j-1})| \\ &\quad \leq 2(e^{2\alpha} - 1) \min\left\{P(V_j = \nu_j \mid V_{1:j-1}, Y), P(V_j = \nu_j \mid V_{1:j-1})\right\} \\ &\quad\quad \times \left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\rm TV}. \end{aligned} \tag{24}$$

Indeed, making the identification

$$V_j \leftrightarrow A, \quad X_j \leftrightarrow B, \quad V_{1:j-1} \leftrightarrow C, \quad Y \leftrightarrow D$$

gives inequality (24), by our independence assumptions. Expanding our KL divergence bound, we have

$$\begin{aligned} &D_{\rm kl}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right) \\ &\quad \leq \sum_{\nu_j} \left(P_{V_j}(\nu_j \mid Y, V_{1:j-1}) - P_{V_j}(\nu_j \mid V_{1:j-1})\right) \log \frac{P_{V_j}(\nu_j \mid Y, V_{1:j-1})}{P_{V_j}(\nu_j \mid V_{1:j-1})}. \end{aligned}$$

Now, using the elementary inequality for $a, b \geq 0$ that

$$\left|\log \frac{a}{b}\right| \leq \frac{|a - b|}{\min\{a, b\}},$$

we have

$$\begin{aligned}
&\left(P_{V_j}(\nu_j \mid Y, V_{1:j-1}) - P_{V_j}(\nu_j \mid V_{1:j-1})\right) \log \frac{P_{V_j}(\nu_j \mid Y, V_{1:j-1})}{P_{V_j}(\nu_j \mid V_{1:j-1})} \\
&\quad \le \frac{(P_{V_j}(\nu_j \mid Y, V_{1:j-1}) - P_{V_j}(\nu_j \mid V_{1:j-1}))^2}{\min\{P_{V_j}(\nu_j \mid Y, V_{1:j-1}), P_{V_j}(\nu_j \mid V_{1:j-1})\}} \\
&\quad \le 4(e^{2\alpha}-1)^2 \min\left\{P_{V_j}(\nu_j \mid Y, V_{1:j-1}), P_{V_j}(\nu_j \mid V_{1:j-1})\right\} \\
&\qquad \times \left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}^2
\end{aligned}$$

by inequality (24).

Substituting this into our bound on KL-divergence, we obtain

$$\begin{aligned}
&I(V_j; Y \mid V_{1:j-1}) \\
&\quad = \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\mathrm{kl}}\left(P_{V_j}(\cdot \mid Y, V_{1:j-1}) \| P_{V_j}(\cdot \mid V_{1:j-1})\right) \mid V_{1:j-1}\right]\right] \\
&\quad \le 4(e^{2\alpha}-1)^2 \mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[\left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}^2 \mid V_{1:j-1}\right]\right].
\end{aligned}$$

Using Pinsker's inequality, we then find that

$$\begin{aligned}
&\mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[\left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y) - P_{X_j}(\cdot \mid V_{1:j-1})\right\|_{\mathrm{TV}}^2 \mid V_{1:j-1}\right]\right] \\
&\le \frac{1}{2}\mathbb{E}_{V_{1:j-1}}\left[\mathbb{E}_Y\left[D_{\mathrm{kl}}\left(P_{X_j}(\cdot \mid Y, V_{1:j-1}) \| P_{X_j}(\cdot \mid V_{1:j-1})\right) \mid V_{1:j-1}\right]\right] = \frac{1}{2} I(X_j; Y \mid V_{1:j-1}).
\end{aligned}$$

In particular, we have

$$I(V_j; Y \mid V_{1:j-1}) \le 2\left(e^{2\alpha}-1\right)^2 I(X_j; Y \mid V_{1:j-1}) \tag{25}$$

Lastly, we argue that $I(X_j; Y \mid V_{1:j-1}) \le I(X_j; Y \mid X_{1:j-1})$. Indeed, we have by definition[2] that

$$\begin{aligned}
I(X_j; Y \mid V_{1:j-1}) &\stackrel{(i)}{=} H(X_j) - H(X_j \mid Y, V_{1:j-1}) \\
&\stackrel{(ii)}{\le} H(X_j) - H(X_j \mid Y, V_{1:j-1}, X_{1:j-1}) \\
&\stackrel{(iii)}{=} H(X_j \mid X_{1:j-1}) - H(X_j \mid Y, X_{1:j-1}) = I(X_j; Y \mid X_{1:j-1}).
\end{aligned}$$

Here, equality (i) follows since X_j is independent of $V_{1:j-1}$, inequality (ii) because conditioning reduces entropy, and equality (iii) because X_j is independent of $X_{1:j-1}$. Thus

$$I(V; Y) = \sum_{j=1}^d I(V_j; Y \mid V_{1:j-1}) \le 2(e^{2\alpha}-1)^2 \sum_{j=1}^d I(X_j; Y \mid X_{1:j-1}) = 2(e^{2\alpha}-1)^2 I(X_{1:d}; Y),$$

which completes the proof.

C Proof of Theorem 2

In this section, we represent the ith sample by an n_i sample matrix $X^{(i)} \in \mathbb{R}^{d \times n_i}$, where the kth column of $X^{(i)}$ is $X^{(i,k)}$ and jth row of $X^{(i)}$ is $X_j^{(i)}$. As usual, we assume the testing Markov chain $V \to X^{(i)} \to Y_i$, as in the setup for our proofs. We assume that $m \ge 4$, since otherwise the interactive lower bound (Proposition 1) provides a stronger result.

We have the following lemma, which is an analogue of Lemma 3.

Lemma 4 *Let V be sampled uniformly at random from $\{-1, 1\}^d$. For any (i, j), assume that $X_j^{(i)}$ is independent of $\{X_{j'}^{(i)} : j' \ne j\} \cup \{V_{j'} : j' \ne j\}$ given V_j. Let P_{X_j} be the probability measure of $X_j^{(i)}$ and assume in addition that*

$$\sup_{S \in \sigma(B_j)} \frac{P_{X_j}(S \mid V = \nu)}{P_{X_j}(S \mid V = \nu')} \le \exp(\alpha).$$

[2] We assume that X is discrete or has a density with respect to Lebesgue measure.

Define the random variable $E_j = 1$ if $X_j^{(i)} \in B_j$ and 0 otherwise. Then

$$I(V; Y_i) \le 2\left(e^{4\alpha} - 1\right)^2 I(X^{(i)}; Y_i) + \sum_{j=1}^{d} H(E_j) + \sum_{j=1}^{d} P(E_j = 0).$$

For the next lemma, we assume that as usual $\mathcal{V} = \{-1, 1\}^d$, and the parameter θ_ν has coordinates given by $(\theta_\nu)_j = \nu_j \delta$. Moreover, we assume that each machine i has n_i independent samples from a $\mathsf{N}(\nu\delta, \sigma^2 I)$ distribution, so $\mathbb{E}_\nu[X] = \theta_\nu$. For conciseness we define the shorthand

$$b_i = \min\left\{128\frac{a^2}{\sigma^2}H(Y_i), d\right\}.$$

Lemma 5 *Let $a > 0$ and $\delta > 0$ be chosen such that $\frac{\sqrt{n_i}a\delta}{\sigma^2} \le \frac{1.2564}{4}$ for any $i \in \{1, \ldots, m\}$, and let $h(p) = -p\log(p) - (1-p)\log(1-p)$ be binary entropy. Then*

$$\begin{aligned} I(V; Y_i) \le \frac{n_i\delta^2}{\sigma^2} \min\left\{128\frac{a^2}{\sigma^2}H(Y_i), d\right\} + dh\left(2\exp\left(-\frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2}\right)\right) \\ + 2d\exp\left(-\frac{(a - \sqrt{n_i}\delta)^2}{2\sigma^2}\right). \end{aligned} \tag{26}$$

With the bound (26) on the mutual information $I(Y_i; V)$, we may now divide our proof into two cases: when $d \ge 9$ and $d < 9$. Let us being with $d \ge 9$. Recalling our earlier minimax bound (16), we have—since $\theta(P_\nu) = \delta\nu$—that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B_{1:m}) \ge \delta^2(\lfloor d/6 \rfloor + 1)\left(1 - \frac{I(Y_{1:m}; V) + \log 2}{d/6}\right).$$

If we can choose appropriate δ so that $I(Y_{1:m}; V) < 3/10$, then (since $d \ge 9$), we will obtain that the minimax error is lower bounded by $\delta^2(\lfloor d/6 \rfloor + 1)/2$, which will complete the proof.

Now, we consider each of the terms in the bound in Lemma 5 in turn, finding settings of δ and a so that each is small. Specifically, recalling the assumption that $m \ge 2$, we will find settings of δ and a so that the sum is bounded by $3/10$. We begin with the third term in the bound, where we note that if

$$\delta_3^2 \le \frac{\sigma^2}{25 \cdot 16\log(m)\max_i n_i} \quad \text{and} \quad a = 5\sigma\sqrt{\log m}, \tag{27a}$$

then the condition $\frac{\sqrt{n_i}a\delta}{\sigma^2} \le \frac{1.2564}{4}$ in Lemma 5 is satisfied. In addition, we have $(a - \sqrt{n_i}\delta_3)^2 \ge (5 - 1/20)^2\sigma^2\log m \ge 24\sigma^2\log m$, so

$$\sum_{i=1}^{m} 4\exp\left(-\frac{(a - \sqrt{n_i}\delta_3)^2}{2\sigma^2}\right) \le 4m\exp\left(-12\log m\right) = \frac{4}{m^{11}} < 10^{-6}.$$

For the first term in the bound from Lemma 5, we note that with the identical choice of $a = 5\sigma\sqrt{n\log m}$, by taking

$$\delta_1^2 \le \frac{d\sigma^2}{10\sum_{i=1}^m b_i n_i}, \tag{27b}$$

we have that $\sum_{i=1}^m 2b_i n_i \delta_1^2/(d\sigma^2) \le 1/5$. Lastly, we have $h(q) \le (6/5)\sqrt{q}$ for $q \ge 0$. As a consequence, we see that for δ_2^2 chosen identically to the choice (27a) for δ_3, we have

$$\sum_{i=1}^{m} 2h\left(2\exp\left(-\frac{(a - \sqrt{n_i}\delta_2)^2}{2\sigma^2}\right)\right) \le \frac{12m}{5}\sqrt{2}\exp\left(-\frac{24}{4}\log m\right) \le \frac{1}{300}.$$

In particular, combining bounds (27a) and (27b), we see that if we choose

$$\delta^2 = \min\left\{1, \frac{\sigma^2}{400\log(m)\max_i n_i}, \frac{d\sigma^2}{10\sum_{i=1}^m b_i n_i}\right\} \quad \text{and} \quad a = 5\sigma\sqrt{\log m},$$

then

$$\sum_{i=1}^{m} \frac{2b_i n_i \delta^2}{d\sigma^2} + 2h\left(2\exp\left(-\frac{(a-\sqrt{n_i}\delta)^2}{2\sigma^2}\right)\right) + 4\exp\left(-\frac{(a-\sqrt{n_i}\delta)^2}{2\sigma^2}\right) < \frac{3}{10}.$$

This completes the proof for the case that $d \geq 9$, since

$$b_i \leq \min\left\{128\frac{a^2}{\sigma^2}H(Y_i), d\right\} = \min\left\{25 \cdot 128 H(Y_i)\log m, d\right\} \leq \min\left\{25 \cdot 128 B_i \log m, d\right\}$$

by Shannon's source coding theorem.

When $d < 9$, an appeal to Le Cam's method (20), as in the proof of Proposition 2, and an identical series of steps to bound the mutual information using inequality (26) (i.e., again applying inequalities (27a)–(27b)) completes the proof.

C.1 Proof of Lemma 4

The proof is substantially similar to that of Lemma 3, but we exhibit some care since we must condition on the event that $X_j^{(i)} \in B_j$. For notational simplicity, we again suppress all dependence of X and Y on the machine index i.

We begin by noting that given E_j, the variable V_j is independent of $V_{1:j-1}$, $X_{1:j-1}$, $V_{j+1:d}$, and $X_{j+1:d}$. Moreover, by the assumption in the lemma we have for any $S \in \sigma(B_j)$ that

$$\frac{P_{X_j}(S \mid V = \nu, E_j = 1)}{P_{X_j}(S \mid V = \nu', E_j = 1)} = \frac{P_{X_j}(S \mid V = \nu)}{P_{X_j}(X_j \in B_j \mid V = \nu)} \frac{P_{X_j}(X_j \in B_j \mid V = \nu')}{P_{X_j}(X_j \in S \mid V = \nu')} \leq \exp(2\alpha),$$

so we have the analogue of the bound (24) that

$$\begin{aligned} &P(V_j = \nu_j \mid V_{1:j-1}, Y, E_j = 1) - P(V_j = \nu_j \mid V_{1:j-1}, E_j = 1) \\ &\leq 2\left(e^{4\alpha} - 1\right)\left\|P_{X_j}(\cdot \mid V_{1:j-1}, Y, E_j = 1) - P_{X_j}(\cdot \mid V_{1:j-1}, E_j = 1)\right\|_{\mathrm{TV}} \cdot \ldots \\ &\qquad \min\left\{P(V_j = \nu_j \mid V_{1:j-1}, Y, E_j = 1), P(V_j = \nu_j \mid V_{1:j-1}, E_j = 1)\right\}. \end{aligned} \tag{28}$$

Thus, proceeding as in the proof of Lemma 3 (specifically the argument preceding inequality (25)), the expression (28) implies

$$I(V_j; Y \mid V_{1:j-1}, E_j = 1) \leq 2\left(e^{4\alpha} - 1\right)^2 I(X_j; Y \mid V_{1:j-1}, E_j = 1). \tag{29}$$

The bound (29) as stated conditions on E_j, which makes it somewhat unwieldy. We turn to removing this conditioning. By the definition of (conditional) mutual information, we have

$$\begin{aligned} &P(E_j = 1) I(V_j; Y \mid V_{1:j-1}, E_j = 1) \\ &\quad = I(V_j; Y \mid V_{1:j-1}, E_j) - I(V_j; Y \mid V_{1:j-1}, E_j = 0) P(E_j = 0) \\ &\quad = I(V_j; E_j, Y \mid V_{1:j-1}) - I(V_j; E_j \mid V_{1:j-1}) - I(V_j; Y \mid V_{1:j-1}, E_j = 0) P(E_j = 0) \end{aligned}$$

Since conditioning reduces entropy,

$$\begin{aligned} I(V_j; E_j, Y \mid V_{1:j-1}) &= H(V_j \mid V_{1:j-1}) - H(V_j \mid E_j, Y, V_{1:j-1}) \\ &\geq H(V_j \mid V_{1:j-1}) - H(V_j \mid Y, V_{1:j-1}) = I(V_j; Y \mid V_{1:j-1}), \end{aligned}$$

and noting that $I(V_j; Y \mid V_{1:j-1}, E_j = 0) \leq H(V_j) \leq 1$ and $I(V_j; E_j \mid V_{1:j-1}) \leq H(E_j)$ gives

$$P(E_j = 1) I(V_j; Y \mid V_{1:j-1}, E_j = 1) \geq I(V_j; Y \mid V_{1:j-1}) - H(E_j) - P(E_j = 0). \tag{30}$$

We now combine inequalities (30) and (29) to complete the proof of the lemma. By the definition of conditional mutual information,

$$I(X_j; Y \mid V_{1:j-1}, E_j = 1) \leq \frac{I(X_j; Y \mid V_{1:j-1}, E_j)}{P(E_j = 1)} \leq \frac{I(X_j; Y \mid V_{1:j-1})}{P(E_j = 1)}.$$

Combining this with inequalities (30) and (29) yields

$$I(V_j; Y \mid V_{1:j-1}) \leq H(E_j) + P(E_j = 0) + 2\left(e^{4\alpha} - 1\right)^2 I(X_j; Y \mid V_{1:j-1}).$$

Up to the additive terms, this is equivalent to the earlier bound (25) in the proof of Lemma 3; proceeding *mutatis mudandis* we complete the proof.

C.2 Proof of Lemma 5

Inequality (26) is the consequence of two intermediate upper bounds, which we prove separately:

$$I(V;Y_i) \le \frac{dn_i\delta^2}{\sigma^2}, \tag{31}$$

$$I(V;Y_i) \le 128\frac{\delta^2 a^2}{\sigma^4} n_i H(Y_i) + dh\left(2\exp\left(-\frac{(a-\sqrt{n_i}\delta)^2}{2\sigma^2}\right)\right) + 2d\exp\left(-\frac{(a-\sqrt{n_i}\delta)^2}{2\sigma^2}\right). \tag{32}$$

To prove inequality (31), we note that $V \to X^{(i)} \to Y_i$ forms a Markov chain. Thus, the data-processing inequality [6] implies that

$$I(V;Y_i) \le I(V;X^{(i)}) \le \sum_{j=1}^{n_i} I(V;X^{(i,j)}) = n_i I(V;X^{(i,1)})$$

where the last inequality comes from the independence of the $X^{(i,j)}$. Let P_ν denote the conditional distribution of $X^{(i,j)}$ given $V=\nu$. Then the convexity of the KL-divergence implies

$$I(V;X^{(i,j)}) \le \frac{1}{|\mathcal{V}|^2}\sum_{\nu,\nu'\in\mathcal{V}} D_{\mathrm{kl}}\left(P_\nu \| P_{\nu'}\right) = \frac{\delta^2}{2\sigma^2}\frac{1}{|\mathcal{V}|^2}\sum_{\nu,\nu'\in\mathcal{V}}\|\nu-\nu'\|_2^2 = \frac{d\delta^2}{\sigma^2}.$$

This establishes inequality (31).

To prove inequality (32), we apply Lemma 4. First, we note that by taking a ratio of the densities of two normals with n_i independent samples, one with mean δ and the other with mean $-\delta$, both with variance σ^2, we have

$$\frac{\exp(-\frac{1}{2\sigma^2}\sum_{l=1}^{n_i}(x_l-\delta)^2)}{\exp(-\frac{1}{2\sigma^2}\sum_{l=1}^{n_i}(x_l+\delta)^2)} = \exp\left(\frac{2\delta}{2\sigma^2}\sum_{l=1}^{n_i}x_l\right) \le \exp\left(\frac{\sqrt{n_i}\delta a}{\sigma^2}\right)$$

whenever $|\sum_l x_l| \le \sqrt{n_i}a$. As a consequence, we see that by taking the sets

$$B_j = \left\{x \in \mathbb{R}^{n_i} : \left|\sum_{l=1}^{n_i} x_l\right| \le \sqrt{n_i}a\right\},$$

we satisfy the conditions of Lemma 4 with $\alpha = \sqrt{n_i}\delta a/\sigma^2$. In addition, when $\alpha \le 1.2564$, we have $\exp(\alpha)-1 \le 2\alpha$, so under the conditions of the lemma, $\exp(4\alpha)-1 = \exp(4\sqrt{n_i}\delta a/\sigma^2)-1 \le 8\sqrt{n_i}\delta a/\sigma^2$. Recalling the definition of the event $E_j = \{X_j^{(i)} \in B_j\}$ from Lemma 4, we obtain

$$I(V;Y_i) \le 128\frac{\delta^2 a^2}{\sigma^4} n_i I(X^{(i)};Y_i) + \sum_{j=1}^{d} H(E_j) + \sum_{j=1}^{d} P(E_j=0).$$

Comparing this inequality with inequality (32), we see that we must bound the probability of the event $E_j = 0$.

Bounding $P(E_j=0)$ is not challenging, however. From standard Gaussian tail bounds, we have for Z_i distributed i.i.d. according to $\mathsf{N}(\delta,\sigma^2)$ that

$$\begin{aligned} P(E_j=0) &= P\left(\left|\sum_{l=1}^{n_i} Z_l\right| \ge \sqrt{n_i}a\right) \\ &= P\left(\sum_{l=1}^{n_i}(Z_l-\delta) \ge \sqrt{n_i}a - n\delta\right) + P\left(\sum_{l=1}^{n_i}(Z_l-\delta) \le \sqrt{n_i}a - n\delta\right) \\ &\le 2\exp\left(-\frac{(a-\sqrt{n_i}\delta)^2}{2\sigma^2}\right). \end{aligned}$$

D Proof of Proposition 1

We prove the lower bound via a standard information-theoretic argument. Fix $\delta > 0$, and let $\mathcal{V} = [2^{\log M_\Theta(2\delta)}]$ index a maximal 2δ-packing of Θ, which we identify by $\{\theta_\nu\}_{\nu\in\mathcal{V}} \subset \Theta$. Fix an (arbitrary) protocol Π for communication.

Following the standard reduction from (worst-case) estimation to testing [20, 22, 19], let V be sampled uniformly from $\mathcal{V}$. For messages $Y = (Y_1, \ldots, Y_T)$ sent by the protocol Π, let $\widehat{\theta}(Y)$ denote the estimator of θ based on Y and define $\widehat{V} = \operatorname{argmin}_{\nu\in\mathcal{V}} \|\widehat{\theta}(Y) - \theta_\nu\|_2$. Then $\|\widehat{\theta}(Y) - \theta_\nu\|_2 \geq \delta$ if $\widehat{V} \neq V$, and we have

$$\begin{aligned}\max_{\nu\in\mathcal{V}} \mathbb{E}\left[\|\widehat{\theta}(Y) - \theta_\nu\|_2^2\right] &\geq \sum_{\nu\in\mathcal{V}} \mathbb{P}(V = \nu)\mathbb{E}\left[\|\widehat{\theta}(Y) - \theta_V\|_2^2 \mid V = \nu\right] \\ &\geq \sum_{\nu\in\mathcal{V}} \delta^2 \mathbb{P}(V = \nu)\mathbb{P}(\widehat{V} \neq V \mid V = \nu) = \delta^2 \mathbb{P}(\widehat{V} \neq V). \qquad (33)\end{aligned}$$

By Fano's inequality [6], the testing error (33) is lower bounded by

$$\mathbb{P}(\widehat{V} \neq V) \geq 1 - \frac{I(V;Y) + 1}{\log M_\Theta(2\delta)} \geq 1 - \frac{H(Y) + 1}{\log M_\Theta(2\delta)},$$

since $H(Y) \geq I(V;Y)$. Shannon's source coding theorem [6, Chapter 5] guarantees the lower bound $B \geq H(Y)$. Since the protocol Π was arbitrary, we have as an immediate consequence of inequality (33) that

$$\mathfrak{M}^{\text{inter}}(\theta, \mathcal{P}, B) \geq \delta^2 \left(1 - \frac{B+1}{\log M_\Theta(2\delta)}\right) \quad \text{for any } \delta \geq 0. \qquad (34)$$

Using inequality (34), the remainder of the proof is straightforward. Indeed, we have

$$1 - \frac{B+1}{\log M_\Theta(2\delta)} \geq \frac{1}{2} \text{ iff } \frac{\log M_\Theta(2\delta)}{B+1} \geq 2 \text{ iff } 2\delta \geq \log M_\Theta^{-1}(2B+2).$$

Setting $\delta = \frac{1}{2}\log M_\Theta^{-1}(2B+2)$ thus gives the result of the theorem.

E Proof of Theorem 1

We follow a standard hypothesis testing setup (recall Section 5.1) to choose a variable $V \in \{-1, 1\}$ uniformly at random and then sample $X^{(i)}$ w.r.t. $\mathsf{N}(\delta V, \sigma^2)$ independently on each of the m machines. However, in this situation, while the local samples are independent, the messages are not: the sequence of random variables $Y = (Y_1, \ldots, Y_T)$ is generated such that the distribution of Y_t is a measurable function of $(X^{(i_t)}, Y_{1:t-1})$ where $i_t \in \{1, \ldots, m\}$ is the index the existing sample upon which Y_t is based. We assume without loss of generality that the sequence $\{i_1, i_2, \ldots, \}$ is fixed in advance—if the choice of index i_t is based on $Y_{1:t-1}$ and X, then we simply say there exists a default value (say $Y_t = \perp$) that indicates "nothing."

Lemma 6 *Assume that $|\mathcal{V}| = 2$. Also assume that there is a set B such that for any $\nu, \nu' \in \mathcal{V}$ we have*

$$\sup\left\{\frac{P_{X^{(i)}}(S \mid \nu)}{P_{X^{(i)}}(S \mid \nu')} \middle| S \in \sigma(B),\ \nu, \nu' \in \mathcal{V}\right\} \leq e^\alpha. \qquad (35)$$

Let the random variable $\mathcal{E} = 1$ if $X^{(i)} \in B$ for all i and $\mathcal{E} = 0$ otherwise. Then

$$I(V;Y) \leq 2\left(e^{4\alpha} - 1\right)^2 I(X;Y) + H(\mathcal{E}) + P(\mathcal{E} = 0).$$

Consider the following scheme. Given $\nu \in \{-1, 1\}$, we assume that each machine i receives n sample $X^{(i,k)}$ $(k = 1, \ldots, n)$ independently sampled according to

$$X \sim \mathsf{N}(\delta\nu, \sigma^2)$$

Following the low dimension case of Proposition 2, inequality (23) implies that

$$\text{if } I(V;Y) \le \frac{3}{10} \text{ then } \sup_{\theta\in\Theta} \mathbb{E}[(\widehat{\theta}-\theta)^2] > \frac{\delta^2}{10}. \tag{36}$$

We focus on showing the conditions for the implication (36) hold. By defining $B = \{x \in \mathbb{R}^n : |\sum_{i=1}^n x_i \le \sqrt{n}a\}$ and the condition of Lemma 6 is satisfied with $\alpha = \sqrt{n}\delta a/\sigma^2$. If we assume that $\alpha \le 1.2564$ (which is satisfied by the assignment described below), then $\exp(\alpha) - 1 \le 2\alpha$ and hence $\exp(4\alpha) - 1 = \exp(4\sqrt{n}\delta a/\sigma^2) - 1 \le 8\sqrt{n}\delta a/\sigma^2$. We obtain

$$I(V;Y) \le 128\frac{\delta^2 n a^2}{\sigma^4} H(Y) + H(\mathcal{E}) + P(\mathcal{E}=0).$$

Let E_i be the random variable such that $E_i = 1$ if $X^{(i)} \in B$ and $E_i = 0$ otherwise. Since $\mathcal{E} = \prod_{i=1}^m E_i$, we have $P(\mathcal{E}=0) \le \sum_{i=1}^m P(\mathbb{E}_i = 0)$. We apply the last inequality in the proof of Lemma 5 to upper bounds $P(E_i = 0)$, which yields that

$$P(\mathcal{E}=0) \le \sum_{i=1}^m P(E_i = 0) \le 2m \exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right).$$

Consequently,

$$I(V;Y) \le 128\frac{\delta^2 n a^2}{\sigma^4} H(Y) + mh\left(2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right)\right) + 2m\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right), \tag{37}$$

where $h(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function. We also used the convexity of h in $[0, 1/2]$, so that $h(p) \le mh(p/m)$ for $0 \le p \le 1/2$.

Given upper bound (37), we follow the proof of Theorem 2 to see that by choosing

$$\delta^2 = \min\left\{1, \frac{\sigma^2}{400\log(m)n}, \frac{\sigma^2}{10\cdot 128\cdot 36\log(m)nH(Y)}\right\} \quad \text{and} \quad a = 5\sigma\sqrt{\log m},$$

we obtain $I(V;Y) \le \frac{3}{10}$. Thus, there is a universal constant c such that

$$\max_{\nu\in\mathcal{V}} \mathbb{E}[(\widehat{\theta}-\theta)^2] > c\min\left\{1, \frac{\sigma^2}{\log(m)n}, \frac{\sigma^2}{\log(m)nH(Y)}\right\}.$$

Applying the source coding theorem to bound $H(Y) \le B$ completes the proof.

E.1 Proof of Lemma 6

Lemma 7 *Consider the hypothesis testing problem described in the second paragraph of Appendix E, but assume that $X^{(i)}$ is sampled from another probability measure $Q(\cdot \mid \nu)$ satisfying*

$$\sup\left\{\frac{Q(S \mid \nu)}{Q(S \mid \nu')} \mid S \in \sigma(\mathcal{X}),\ \nu, \nu' \in \mathcal{V}\right\} \le e^{\alpha}. \tag{38}$$

Then we have

$$I(V;Y) \le 2\left(e^{2\alpha}-1\right)^2 I(X;Y).$$

With Lemma 7 established, the proof of Lemma 6 follows, *mutatis mutandis*, as in the proof of Lemma 4 from Lemma 3. Thus, it only remains to prove Lemma 7.

Proof of Lemma 7 By the chain-rule for mutual information, we have that

$$I(V;Y) = \sum_{t=1}^T I(V; Y_t \mid Y_{1:t-1}).$$

Let $P_{Y_t}(\cdot \mid Y_{1:t-1})$ denote the (marginal) distribution of Y_t given $Y_{1:t-1}$ and define $P_V(\cdot \mid Y_{1:t})$ to be the distribution of V conditional on $Y_{1:t}$. Then we have by marginalization that

$$P_V(\cdot \mid Y_{1:t-1}) = \int P_V(\cdot \mid Y_{1:t-1}, y_t) dP_{Y_t}(y_t \mid Y_{1:t-1})$$

and thus

$$I(V; Y_t \mid Y_{1:t-1}) = \mathbb{E}_{Y_{1:t-1}}\left[\mathbb{E}_{Y_t}\left[D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) \mid Y_{1:t-1}\right]\right]. \tag{39}$$

We now bound the above KL divergence using the assumptions in the lemma.

By the nonnegativity of the KL divergence, we have

$$\begin{aligned} &D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) \\ &\quad \le D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) + D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t-1}) \| P_V(\cdot \mid Y_{1:t})\right) \\ &\quad = \sum_{\nu \in \mathcal{V}} \left(p_V(\nu \mid Y_{1:t-1}) - p_V(\nu \mid Y_{1:t})\right) \log \frac{p_V(\nu \mid Y_{1:t-1})}{p_V(\nu \mid Y_{1:t})} \end{aligned}$$

where p_V denotes the p.m.f. of V. We claim that Lemma 8 implies that

$$\begin{aligned} &|p_V(\nu \mid Y_{1:t-1}) - p_V(\nu \mid Y_{1:t})| \\ &\le 2\left(e^{2n\alpha} - 1\right) \min\left\{p_V(\nu \mid Y_{1:t-1}), p_V(\nu \mid Y_{1:t})\right\} \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}. \end{aligned} \tag{40}$$

Deferring the proof of inequality (40) to the end of this section, we give the remainder of the proof. First, by a first-order convexity argument, we have that for any $a, b > 0$

$$\log \frac{a}{b} \le \frac{|a-b|}{\min\{a, b\}}.$$

As a consequence, we find

$$\begin{aligned} &\left(p_V(\nu \mid Y_{1:t-1}) - p_V(\nu \mid Y_{1:t})\right) \log \frac{p_V(\nu \mid Y_{1:t-1})}{p_V(\nu \mid Y_{1:t})} \le \frac{\left(p_V(\nu \mid Y_{1:t-1}) - p_V(\nu \mid Y_{1:t})\right)^2}{\min\{p_V(\nu \mid Y_{1:t-1}), p_V(\nu \mid Y_{1:t})\}} \\ &\quad \le 4\left(e^{2n\alpha} - 1\right)^2 \min\left\{p_V(\nu \mid Y_{1:t-1}), p_V(\nu \mid Y_{1:t})\right\} \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2 \end{aligned}$$

by using inequality (40). Using the fact that p_V is a p.m.f., we thus have

$$\begin{aligned} &D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t}) \| P_V(\cdot \mid Y_{1:t-1})\right) + D_{\mathrm{kl}}\left(P_V(\cdot \mid Y_{1:t-1}) \| P_V(\cdot \mid Y_{1:t})\right) \\ &\le 4\left(e^{2n\alpha} - 1\right)^2 \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2 \sum_{\nu \in \mathcal{V}} \min\left\{p_V(\nu \mid Y_{1:t-1}), p_V(\nu \mid Y_{1:t})\right\} \\ &\le 4\left(e^{2n\alpha} - 1\right)^2 \left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2. \end{aligned}$$

Using Pinsker's inequality, we then find that

$$\begin{aligned} &\mathbb{E}_{Y_{1:t-1}}\left[\mathbb{E}_{Y_t}\left[\left\|P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) - P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right\|_{\mathrm{TV}}^2 \mid Y_{1:t-1}\right]\right] \\ &\le \frac{1}{2}\mathbb{E}_{Y_{1:t-1}}\left[\mathbb{E}_{Y_t}\left[D_{\mathrm{kl}}\left(P_{X^{(i_t)}}(\cdot \mid Y_{1:t}) \| P_{X^{(i_t)}}(\cdot \mid Y_{1:t-1})\right) \mid Y_{1:t-1}\right]\right] = \frac{1}{2} I(X^{(i_t)}; Y_t \mid Y_{1:t-1}). \end{aligned}$$

Since conditioning reduces entropy and Y is discrete, we have

$$\begin{aligned} I(X^{(i_t)}; Y_t \mid Y_{1:t-1}) &= H(Y_t \mid Y_{1:t-1}) - H(Y_t \mid X^{(i_t)}, Y_{1:t-1}) \\ &\le H(Y_t \mid Y_{1:t-1}) - H(Y_t \mid X, Y_{1:t-1}) = I(X; Y_t \mid Y_{1:t-1}). \end{aligned}$$

This completes the proof of the lemma, since $\sum_{t=1}^{T} I(X; Y_t \mid Y_{1:t-1}) = I(X; Y)$ by the chain rule for information.

Proof of inequality (40) To establish the inequality, we give a one-to-one correspondence between the variables in Lemma 8 and the variables in Lemma 7. We make the following identifications:

$$V \leftrightarrow A \qquad X^{(i_t)} \leftrightarrow B \qquad Y_{1:t-1} \leftrightarrow C \qquad Y_t \leftrightarrow D.$$

For Lemma 8 to hold, we must verify conditions (43), (44), and (45). For condition (43) to hold, Y_t must be independent of V given $\{Y_{1:t-1}, X^{(i_t)}\}$. Since the distribution of $P_{Y_t}(\cdot \mid Y_{1:t-1}, X^{(i_t)})$ is measurable-$\{Y_{1:t-1}, X^{(i_t)}\}$, Condition (45) is satisfied by the assumption in the lemma.

Finally, for condition (44) to hold, we must be able to factor the conditional probability of $Y_{1:t-1}$ given $\{V, X^{(i_t)}\}$ as

$$P(Y_{1:t-1} = y_{1:t-1} \mid V, X^{(i_t)}) = \Psi_1(V, y_{1:t-1})\Psi_2(X^{(i_t)}, y_{1:t-1}). \tag{41}$$

To prove this decomposition, notice that

$$P(Y_{1:t-1} = y_{1:t-1} \mid V, X^{(i_t)}) = \prod_{k=1}^{t-1} P(Y_k = y_k \mid Y_{1:k-1}, V, X^{(i_t)}).$$

For any $k \in \{1, \ldots, t-1\}$, if $i_k = i_t$—that is, the message Y_k is generated based on sample $X^{(i_t)} = X^{(i_k)}$—then Y_k is independent of V given $\{X^{(i_t)}, Y_{1:k-1}\}$. Thus, $P_{Y_k}(\cdot \mid Y_{1:k-1}, V, X^{(i_t)})$ is measurable-$\{X^{(i_t)}, Y_{1:k-1}\}$. If the kth index $i_k \neq i_t$, then Y_k is independent of $X^{(i_t)}$ given $\{Y_{1:k-1}, V\}$ by construction, which means $P_{Y_k}(\cdot \mid Y_{1:k-1}, V, X^{(i_t)}) = P_{Y_k}(\cdot \mid Y_{1:k-1}, V)$. The decomposition (41) thus holds, and we have verified that each of the conditions of Lemma 8 holds. We thus establish inequality (40).

F Proof of Corollary 1

We prove Corollary 1 in two parts: the upper bound (for part (a)) and lower bound (for part (b)). We prove the upper bound by exhibiting an interactive protocol Π^* and prove the lower bound by applying Proposition 1.

Upper bound on the minimax risk We consider the following communication protocol $\Pi^* \in \mathcal{A}_{\text{inter}}(B, \mathcal{P})$:

1. Machine $i \in \{1, \ldots, m\}$ computes its local minimum $a^{(i)} = \min\{X^{(i,k)} : k = 1, \ldots, n\}$.
2. Machine 1 broadcasts $a^{(1)}$ using $2\log(mn)$ bits. Upon receiving the broadcast, all machines initialize global minimum variables $s \leftarrow a^{(1)}$.
3. In the order $i = 2, 3, \ldots, m$, machine i performs the following operations:
   (i) Check if $a^{(i)} < s$. If so, machine i performs the update $s \leftarrow a^{(i)}$ and broadcasts s, otherwise it does nothing.
   (ii) All other machines update their local s after receiving machine i's update. All real numbers in the message are rounded down to $2\log(mn)$-bit discrete values.
4. One machine outputs $\widehat{\theta} = s + 1$.

According to the protocol described above, Π^* computes a global minima

$$s = \min\left\{X^{(i,k)} : i = 1, \ldots, m;\ k = 1, \ldots, n\right\}$$

to accuracy of $\mathcal{O}(1/(mn)^2)$ since because real numbers are encoded with $2\log(mn)$ bits. Then classical convergence analysis [14] yields estimator $\widehat{\theta} = s + 1$ achieves minimax optimal convergence rate $\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \lesssim 1/(mn)^2$.

To analyze the communication complexity of the protocol Π^*, we study Steps 2–3. In Step 2, machine 1 sends $2\log(mn)$ bits as message Y_1. In Step 3, machine i sends $2\log(mn)$ bits only if $a^{(i)} < \min\{a^{(1)}, \cdots, a^{(i-1)}\}$. By inspection, this event happens with probability bounded by $1/i$, so we find that the expected length of message Y_i is

$$\mathbb{E}[L_i] \leq \frac{2\log(mn)}{i}.$$

Putting all pieces together, we obtain that

$$\mathbb{E}[L] = \sum_{i=1}^{m} \mathbb{E}[L_i] \leq 2\log(mn) + \sum_{i=2}^{m} \frac{2\log(dmn)}{i} \leq 2\log(mn) + 2\ln(m)\log(mn).$$

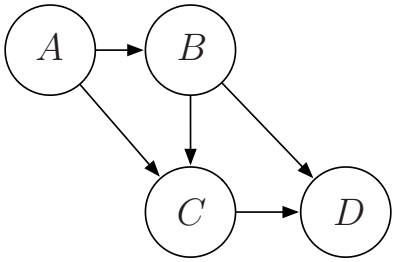

Figure 1: Graphical model for Lemma 8

Lower bound on the minimax risk To prove the lower bound, we simply evaluate packing entropies by using a volume argument [3]. Since $\Theta = [-1, 1]$, the size of a maximal 2δ-packing can be lower bounded by

$$2^{\log M_\Theta(2\delta)} \geq \frac{\text{Volume}(\Theta)}{\text{Volume}(\{x \in \mathbb{R} : \|x\|_2 \leq 2\delta\})} \geq \frac{1}{2\delta}. \tag{42}$$

Taking logarithms and inverting $B = \log M_\Theta(\delta) = \log M_\Theta(1/(mn))$ yields the lower bound.

G Total variation contraction

In this section, we prove a technical lemma that is essential to the proof of our results.

Consider four random variables A, B, C, D, of which we assume that A, C, and D have discrete distributions. We denote the conditional distribution of A given B by $P_{A|B}$ and their full joint distribution by $P_{A,B,C,D}$. We assume that the random variables have conditional indpendence structure specified by the graphical model in Figure 1, that is, that we can write the joint distribution as the product

$$P_{A,B,C,D} = P_A P_{B|A} P_{C|A,B} P_{D|B,C}. \tag{43}$$

We denote the domain of a random variable by the identical calligraphic letter, so $A \in \mathcal{A}$, $B \in \mathcal{B}$, and so on. We write $\sigma(\mathcal{A})$ for the sigma-field on $\mathcal{A}$ with respect to which our measures are defined. Sometimes we write $P_A(\cdot \mid B)$ for the conditional distribution of A given B. In addition to the conditional independence assumption (43), we assume that the conditional distribution of C given A, B factorizes in the following specific form. There exist functions $\Psi_1 : \mathcal{A} \times \sigma(\mathcal{C}) \to \mathbb{R}_+$ and $\Psi_2 : \mathcal{B} \times \sigma(\mathcal{C}) \to \mathbb{R}_+$ such that for any (measureable) set S in the range $\mathcal{C}$ of C, we have

$$P_C(S \mid A, B) = \Psi_1(A, S)\Psi_2(B, S). \tag{44}$$

Since C is assumed discrete, we abuse notation and write $P(C = c \mid A, B) = \Psi_1(A, c)\Psi_2(B, c)$. Lastly, we assume that for any $a, a' \in \mathcal{A}$, we have the following likelihood ratio bound:

$$\sup_{S \in \sigma(\mathcal{B})} \frac{P_B(S \mid A = a)}{P_B(S \mid A = a')} \leq \exp(\alpha). \tag{45}$$

Lemma 8 *Under assumptions* (43), (44), *and* (45), *the following inequality holds:*

$$\begin{aligned} &|P(A = a \mid C, D) - P(A = a \mid C)| \\ &\quad \leq 2\left(e^{2\alpha} - 1\right) \min\left\{P(A = a \mid C), P(A = a \mid C, D)\right\} \left\|P_B(\cdot \mid C, D) - P_B(\cdot \mid C)\right\|_{\text{TV}}. \end{aligned}$$

Proof: By assumption, A is independent of D given $\{B, C\}$. Thus we may write

$$P(A = a \mid C, D) - P(A = a \mid C) = \int P(A = a \mid B = b, C) \left(dP_B(b \mid C, D) - dP_B(b \mid C)\right)$$

Combining this equation with the inequality

$$\int P(A = a \mid C) \left(dP_B(b \mid C, D) - dP_B(b \mid C)\right) = 0$$

we find that

$$\begin{aligned}&P(A=a\mid C,D)-P(A=a\mid C)\\&=\int\left(P(A=a\mid B=b,C)-P(A=a\mid C)\right)\left(dP_B(b\mid C,D)-dP_B(b\mid C)\right).\end{aligned}$$

Using the fact that $|\int f(b)d\mu(b)|\leq \sup_b\{|f(b)|\}\int|d\mu(b)|$ for any signed measure μ on $\mathcal{B}$, we conclude from the previous equality that for *any* version $P_A(\cdot\mid B,C)$ of the conditional probability of A given $\{B,C\}$ that since $\int|d\mu|=\|\mu\|_{\mathrm{TV}}$,

$$\begin{aligned}&|P(A=a\mid C,D)-P(A=a\mid C)|\\&\quad\leq 2\sup_{b\in\mathcal{B}}\{|P(A=a\mid B=b,C)-P(A=a\mid C)|\}\,\|P_B(\cdot\mid C,D)-P_B(\cdot\mid C)\|_{\mathrm{TV}}.\end{aligned}$$

Thus, to prove the lemma, it is sufficient to show (for some version of the conditional distribution[3] $P_A(\cdot\mid B,C)$) that for any $b\in\mathcal{B}$

$$|P(A=a\mid B=b,C)-P(A=a\mid C)|\leq(e^{2\alpha}-1)\min\{P(A=a\mid C),P(A=a\mid C,D)\}. \tag{46}$$

To prove this upper bound, we consider the joint distribution (43) and likelihood ratio bound (46). The distributions $\{P_B(\cdot\mid A=a)\}_{a\in\mathcal{A}}$ are all absolutely continuous with respect to one another by assumption (46), so it is no loss of generality to assume that there exists a density $p_B(\cdot\mid A=a)$ for which $P(B\in S\mid A=a)=\int p_B(b\mid A=a)d\mu(b)$, for some fixed measure μ, and for which the ratio $p_B(b\mid A=a)/p_B(b\mid A=a')\in[e^{-\alpha},e^{\alpha}]$ for all b. By elementary conditioning we have for any $S_b\in\sigma(\mathcal{B})$ and $c\in\mathcal{C}$

$$\begin{aligned}&P(A=a\mid B\in S_b,C=c)\\&\quad=\frac{P(A=a,B\in S_b,C=c)}{P(B\in S_b,C=c)}\\&\quad=\frac{P(B\in S_b,C=c\mid A=a)P(A=a)}{\sum_{a'\in\mathcal{A}}P(A=a')P(B\in S_b,C=c\mid A=a)}\\&\quad=\frac{P(A=a)\int_{S_b}P(C=c\mid B=b,A=a)p_B(b\mid A=a)d\mu(b)}{\sum_{a'\in\mathcal{A}}P(A=a')\int_{S_b}P(C=c\mid B=b,A=a')p_B(b\mid A=a')d\mu(b)},\end{aligned}$$

where for the last equality we used the conditional independence assumptions (43). But now we recall the decomposition formula (44), and we can express the likelihood functions by

$$P(A=a\mid B\in S_b,C=c)=\frac{P(A=a)\int_{S_b}\Psi_1(a,c)\Psi_2(b,c)p_B(b\mid A=a)d\mu(b)}{\sum_{a'}P(A=a')\int_{S_b}\Psi_1(a',c)\Psi_2(b,c)p_B(b\mid A=a')d\mu(b)}.$$

As a consequence, there is a version of the conditional distribution of A given B and C such that

$$P(A=a\mid B=b,C=c)=\frac{P(A=a)\Psi_1(a,c)p_B(b\mid A=a)}{\sum_{a'}P(A=a')\Psi_1(a',c)p_B(b\mid A=a')}. \tag{47}$$

Define the shorthand

$$\beta=\frac{P(A=a)\Psi_1(a,c)}{\sum_{a'\in\mathcal{A}}P(A=a')\Psi_1(a',c)}.$$

We claim that

$$e^{-\alpha}\beta\leq P(A=a\mid B=b,C=c)\leq e^{\alpha}\beta. \tag{48}$$

Assuming the correctness of bound (48), we establish inequality (46). Indeed, since $P(A=a\mid C=c)$ is a weighted average of $P(A=a\mid B=b,C=c)$, we also have the same upper and lower bound for $P(A=a\mid C)$, that is

$$e^{-\alpha}\beta\leq P(A=a\mid C)\leq e^{\alpha}\beta,$$

[3]If $P(A=a\mid C)$ is undefined, we simply set it to have value 1 and assign $P(A=a\mid B,C)=1$ as well.

while the conditional independence assumption that A is independent of D given B, C (recall Figure 1 and the product (43)) implies

$$\begin{aligned} P(A = a \mid C = c, D = d) &= \int_{\mathcal{B}} P(A = a \mid B = b, C = c, D = d) dP_B(b \mid C = c, D = d) \\ &= \int_{\mathcal{B}} P(A = a \mid B = b, C = c) dP_B(b \mid C = c, D = d), \end{aligned}$$

and the final integrand belongs to $\beta[e^{-\alpha}, e^{\alpha}]$. Combining the preceding three displayed expressions, we find that

$$\begin{aligned} |P(A = a \mid B = b, C) - \mathbb{P}(A = a \mid C)| &\leq \left(e^{\alpha} - e^{-\alpha}\right) \beta \\ &\leq \left(e^{\alpha} - e^{-\alpha}\right) e^{\alpha} \min\left\{P(A = a \mid C), P(A = a \mid C, D)\right\}. \end{aligned}$$

This completes the proof of the upper bound (46).

It remains to prove inequality (48). We observe from expression (47) that

$$P(A = a \mid B = b, C) = \frac{P(A = a)\Psi_1(a, C)}{\sum_{a' \in \mathcal{A}} P(A = a')\Psi_1(a', C) \frac{p_B(b \mid A = a')}{p_B(b \mid A = a)}}.$$

By the likelihood ratio bound (45), we have $p_B(b \mid A = a')/p_B(b \mid A = a) \in [e^{-\alpha}, e^{\alpha}]$, and combining this inequality with the above equation yields inequality (48). ■

H Proof of Lemma 1

For any $\Delta > 0$ and any estimator $\widehat{\theta}$, if V is a random variable uniformly chosen from $\mathcal{V}$, then we have

$$\max_{\nu \in \mathcal{V}} \mathbb{E}\left[\|\widehat{\theta} - \theta_\nu\|_2^2\right] \geq \mathbb{E}\left[\|\widehat{\theta} - \theta_V\|_2^2\right] \geq \mathbb{E}\left[\Delta^2 1_{\left(\|\widehat{\theta} - \theta_V\|_2 \geq \Delta\right)}\right] = \Delta^2 \mathbb{P}(\|\widehat{\theta} - \theta_V\|_2 \geq \Delta). \quad (49)$$

We now lower bound $\mathbb{P}(\|\widehat{\theta} - \theta_V\|_2 \geq \Delta)$ by a testing-like probability claimed in the lemma. Define the testing function

$$\widehat{\nu} := \operatorname*{argmin}_{\nu \in \mathcal{V}} \|\theta_\nu - \widehat{\theta}\|_2.$$

The triangle inequality implies that

$$\|\theta_{\widehat{\nu}} - \theta_V\|_2 \leq \|\theta_{\widehat{\nu}} - \widehat{\theta}\|_2 + \|\widehat{\theta} - \theta_V\|_2 \leq 2\|\widehat{\theta} - \theta_V\|_2 \quad (50)$$

Recall that $\theta_\nu = \delta\nu$ where $\nu \in \{-1, 1\}^d$, we have $\|\theta_{\widehat{\nu}} - \theta_V\|_2 = 2\delta\sqrt{d_{\text{ham}}(\widehat{\nu}, V)}$. Combining this equation with inequality (50) implies that

$$\text{if } \; d_{\text{ham}}(\widehat{\nu}, V) > t \; \text{ then } \; \|\widehat{\theta} - \theta_V\|_2^2 \geq \delta^2(\lfloor t \rfloor + 1).$$

Consequently,

$$P\left(\|\widehat{\theta} - \theta_V\|_2^2 \geq \delta^2(\lfloor t \rfloor + 1)\right) \geq P(d_{\text{ham}}(\widehat{\nu}, V) > t). \quad (51)$$

Combining inequality (49) and (51) with $\Delta^2 = \delta^2(\lfloor t \rfloor + 1)$, we have

$$\max_{\nu \in \mathcal{V}} \mathbb{E}\left[\|\widehat{\theta} - \theta_\nu\|_2^2\right] \geq \delta^2(\lfloor t \rfloor + 1) P(d_{\text{ham}}(\widehat{\nu}, V) > t).$$

On the righthand side of the above inequality, taking infinium over all testing functions establishes the result.