S. Ruiz

On Distributed Computations Under Communication Constraints

Bachelor thesis

July 21, 2017

Thesis supervisor: dr. B. Szabo

Leiden University
Mathematical Institute

Contents

Introduction

There is an ever increasing amount of information available because data scientists have developed computer systems that are constantly collecting data all over the world. An example of this is the internet, where Google is collecting search queries and information about how users interact with web pages. Machines cannot keep up with the information that needs to be processed. Therefore data scientists are developing new scalable learning and statistical methods to handle big data problems.

A scalable approach is to use distributed methods where data is split among many machines such that computations can be done in parallel and each machine analyses only a subset of the complete dataset. The local machines send their results to a global machine which aggregates the local results to give an overall interpretation of the data. Optimally, the heavy computations are done on the local machines.

Distributed methods also have other applications; they can can offer security in terms of privacy. The data that needs to be processed may be sensitive, for example patient hospital records, where collecting data in one central location could be dangerous for patient confidentiality. Instead communicating as little data as possible from multiple locations to obtain a similar result would be advantageous. We will investigate how to approach this, with emphasis on communication constraints between the machines; in other words how to communicate as little information between machines as possible while still being able to determine accurate results. This has practical reasons in the sense that communicating over some networks can be expensive, for example communicating with satellites.

We consider some idealised models where the data is sampled from well known distributions and where we investigate the theoretical limitations of communication constraints through simulations and theoretical results. From the theoretical results we prove some hypotheses determined from the simulations.

Some interesting methods that we look at are: A method with data from a normal distribution, where on each local machine we describe the data with a one bit value and rely on many local machines such that the aggregated result is still a good estimate for the property of the distribution we are evaluating. In another method with data from a uniform distribution, local machines have multiple communication rounds in which they can communicate with the global machine and where the local machines can view the communications of all machines from earlier communication rounds. Both of these methods aim to cut down on the communication required between machines while still giving good enough results.

We investigate the work of [6] where they explore theoretic lower bounds for communication constraints. They find a lower bound for the minimax risk of any estimator based on the communicated data, given that the data is normally distributed. We calculate the constants exactly for the proof they give, and further extend these results to Laplace distributed data. We find that the Laplace lower bound for the minimax risk is the same as for normal, up to constant factors. This brings us to ask questions about how other distributions would perform.

Distributed computing is used for all sorts of systems and projects in the real world. We highlight below a few projects that everyone has access to.

- **Folding@home [3]:** The Folding@home project by Stanford university focuses on disease research that simulates protein folding. The problems require solving many computer calculations. The project uses idle processing resources of thousands of computers to perform these computations.

- **Pooled bitcoin mining [5]:** Bitcoins are a virtual currency that are rewarded by mining a block. A block is mined by doing many computer calculations. The more blocks generated, the harder is becomes to mine. In pooled bitcoin mining, multiple clients contribute to the generation of a block, and then split the block reward according to the contributed processing power.

- **Great Internet Mersenne Prime Search (GIMPS) [4]:** Computers taking part in the project receive tasks from the central computer. Each computer searches for Mersenne prime numbers.

1 Problem Setting

In this section we introduce various communication methods in order to formally outline the problem. We will immediately use these definitions for our idealised examples in the next section.

Definition 1.1 (Global and local machines)**.** *The global machine receives messages communicated by the local machines. Each local machine has a subset of the complete dataset and can communicate with the global machine.*

In a common scenario the local machines compute some estimator from the data they have and communicate the result to the global machine. The global machine aggregates the results to determine some estimator from the complete dataset.

Definition 1.2 (The dataset)**.** *Let $m \in \mathbb{N}$ be the number of machines and $n \in \mathbb{N}$ the size of the data per machine. Let $\mathcal{P}$ be the set of all probability distributions. Consider a fixed probability distribution $P \in \mathcal{P}$. Let $X := \{X^{(1)}, ..., X^{(m)}\}$ be the complete dataset or sample, where machine i has data $X^{(i)} := \{X_1^{(i)}, ..., X_n^{(i)}\}$, where $X_j^{(i)} \overset{iid}{\sim} P$ for all i, j.*

The entire dataset has size mn.

Definition 1.3 (Communication rounds)**.** *Let $T \in \mathbb{N}$ denote the number of communication rounds. For each communication round all m machines may communicate once with the global machine and the local machines may be able to read data published by the global machine about earlier communication rounds.*

We call a communication method between global and local machines a *protocol*.

Definition 1.4 (Protocols)**.** *A protocol $\Pi(T)$ defines at each round $t \in \{1, 2, ..., T\}$, the communication of machine i to the global machine, given by $Y_{t,i}$ a measurable function of the data $X^{(i)}$, and potentially of past communication between the global and all local machines from earlier communication rounds.*

Let $\mathbf{Y_t} := \{Y_{t,1}, ..., Y_{t,m}\}$ be the set of all messages sent at round t. The sequence $\mathbf{Y} := (\mathbf{Y_1}, ..., \mathbf{Y_T})$ gives all communication between global and local machines. The protocol $\Pi(T)$ then defines an estimator $\hat{\theta} := \hat{\theta}(\mathbf{Y_1}, ..., \mathbf{Y_T})$, where $\hat{\theta}$ is intended to estimate some function θ on P.

A protocol can be represented by a Markov chain (Definition 5.7) $X \to \mathbf{Y} \to \hat{\theta}$, showing how a protocol uses a dataset X to give some $\mathbf{Y}$, the communication from X between the local and global machines, which in-turn defines an estimator $\hat{\theta}$.

For each protocol, we want to examine how much data is communicated, to do this we look at communication costs and we name different types of protocols.

Definition 1.5 (Total communication cost)**.** *Let $L_{t,i}$ be the minimal number of bits required to encode $Y_{t,i}$. Then the total communication cost is given by $L = \sum_{t=1}^{T} \sum_{i=1}^{m} L_{t,i}$.*

Definition 1.6 (Non-distributive, independent and interactive protocols)**.** *Non-distributive protocols have 0 rounds of communication, $T = 0$. Independent protocols have 1 round of communication, $T = 1$. Interactive protocols have multiple rounds of communication, $T > 1$.*

For the non-distributive protocol we assume that there are no local machines and that the global machine has access to all the information.

In the examples in Section 2 we consider protocols that are non-distributive, independent and interactive. In Section 6 and Section 7 where we investigate the minimax risk over different models, we consider only independent protocols.

In the following sections we will focus on models that take samples from the Uniform, normal and Laplace distributions.

Definition 1.7 (Uniform, normal and Laplace means models)**.** *We define a means model to take a set of probability distributions $\mathcal{P}$ and a set of protocols $\mathbf{\Pi}$ that estimate the mean $\theta(P)$ for $P \in \mathcal{P}$.*

- *Uniform means model is a means model with $\mathcal{P}$ a family of uniform distributions.*
- *Normal means model is a means model with $\mathcal{P}$ a family of normal distributions.*
- *Laplace means model is a means model with $\mathcal{P}$ a family of Laplace distributions.*

2 Example Protocols

We consider two different models in this section, the normal means model and the uniform means model. In the normal means model we consider normal distributed data and protocols that aim to estimate the true mean of the data, and similarly for the uniform means model.

We consider some example protocols under the normal and uniform means models. To determine how good an estimator is, we consider a risk function.

Definition 2.1 (Loss and risk functions of an estimator)**.** *Let $\hat{\theta}$ be an estimator for $\theta \in \Theta$. The loss function L_2 is given by $L_2(\theta, \hat{\theta}) := ||\theta - \hat{\theta}||_2^2$. The risk function R_2 of an estimator $\hat{\theta}$ is given by $R_2(\theta, \hat{\theta}) := \mathbb{E}_\theta(L_2(\theta, \hat{\theta}))$ the average L_2 loss between the estimator and the parameter of interest.*

Using computer simulations in Section 3 we will compare the protocols by considering the total communication cost and the L_2 loss of the estimator.

We summarise the protocols we have chosen, where the protocol types are given in Definition 1.6. For the uniform means model we consider the following protocols investigated in Section 2.1 in details, similarly below in the case of the other model.

1. A non-distributive protocol, where the data is not distributed and the mean of the uniform distribution is estimated on the global machine.
2. An independent protocol, where we take the minimum of the data on each machine.
3. An interactive protocol, where local machines can read the communication of earlier communication rounds from all the local machines.

For the normal means model we consider the following protocols from Section 2.2 to estimate the mean of the sample.

1. A non-distributive protocol, where the data is not distributed and the mean is estimated on the global machine given all the data.
2. An independent protocol, where we take the mean of the local data on each machine.
3. An independent protocol, where we communicate a one bit random variable from each machine. The random variable is Bernoulli distributed with parameter dependent on the local data.

The above protocols are defined rigorously in the following sections.

2.1 Protocols for the Uniform Means Model

Let $\theta \in \mathbb{R}$, and $\delta \in \mathbb{R}_{>0}$. Let $X_j^{(i)} \overset{\text{iid}}{\sim} U(\theta - \delta, \theta + \delta)$ for all i, j, where δ is the offset. In the simulations we take $\delta = 10$.

We want to estimate θ with some estimator $\hat{\theta}$ using 3 different protocols, a non-distributive, an independent and an interactive protocol.

Define the rounding function $\text{round}(x, k)$ as x rounded in decimal representation to k significant figures.

Example 2.1 (Protocol 1, A non-distributive protocol)**.** *Let Π_1^{unif} be the following non-distributive protocol ($T = 0$). On the global machine let $\hat{m} := \min X$ of the entire dataset and let $\hat{\theta}^{non\text{-}dist} := \hat{m} + \delta$.*

In the non-distributive protocol, no distributed computing takes place and we use as a benchmark protocol to compare the following protocols to.

Example 2.2 (Protocol 2, An independent protocol)**.** *Let Π_2^{unif} be the following independent protocol ($T = 1$). Each local machine communicates the local minimum rounded to 3 significant figures, $Y_i := round(\min X^{(i)}, 3)$ to the global machine. On the global machine let $\hat{\theta}^{ind} := \min\{Y_1, ..., Y_m\} + \delta$.*

Example 2.3 (Protocol 3, An interactive protocol)**.** *Let Π_3^{unif} be the following interactive protocol with $T = m$ communication rounds. For all t, i let $Y_{t,i} := round(\min X^{(i)}, 3)$ if $\min X^{(i)} < \min \mathbf{Y}_j$ for all $j \in \{1, ..., i-1\}$ and otherwise do not communicate $Y_{t,i}$. On the global machine let $\hat{\theta}^{int} := \min\{\mathbf{Y}_1, ..., \mathbf{Y}_m\} + \delta$.*

The idea behind Example 2.3 is that in each communication round, all the machines can view the communication made by the local machines in all the earlier communication rounds. Each machine communicates the computed minimum of its local data if the minimum is less than all the minima communicated by the local machines in all previous communication rounds. The global machine takes the minimum of all the communicated minima to determine an estimator for the absolute minimum of the data.

We note that protocols Π_2^{unif} and Π_3^{unif} must return the same result for the L_2 loss by definition. They do differ however in how much data is communicated. We will investigate this in our simulations in Section 3.

2.2 Protocols for the Normal Means Model

Let $\theta \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{\geq 0}$. Let $X_j^{(i)} \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ for all i, j, where for simplicity we take $\sigma^2 = 1$.

Example 2.4 (Protocol 1, A non-distributive protocol)**.** *Let Π_1^{norm} be the following non-distributive protocol ($T = 0$). On the global machine let $\hat{\theta}^{non\text{-}dist} := \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} X_j^{(i)}$. No distributed computing takes place and we use this as a benchmark protocol.*

Example 2.5 (Protocol 2, An independent protocol)**.** *Let Π_2^{norm} be the following independent protocol ($T = 1$). Each local machine communicates the local sample average rounded to 3 significant figures $Y_i := round(\frac{1}{n} \sum_{j=1}^{n} X_j^{(i)}, 3)$, to the global machine. On the global machine let $\hat{\theta}^{ind} := \frac{1}{m} \sum_{i=1}^{m} Y_i$.*

Example 2.6 (Protocol 3, An independent protocol)**.** *Let Π_3^{norm} be the following independent protocol ($T = 1$). Let us assume that $\theta \in [-a, a]$ for some $a \in \mathbb{R}_{>0}$. On each local machine compute the sample average $\overline{X}^{(i)} := \frac{1}{n} \sum_{j=1}^{n} X_j^{(i)}$ and then take*

$\overline{X}_*^{(i)} := (\overline{X}^{(i)} \vee -2a) \wedge 2a$. *Let* Z_i *be a Bernoulli random variable with parameter* $p_i := \frac{\overline{X}_*^{(i)}+2a}{4a}$. *The global machine computes the estimator* $\hat{\theta}^{ind} := \frac{1}{m}\sum_{i=1}^{m}(Z_i 4a - 2a)$.

The idea behind Example 2.6 is that on each machine we encode the sample average as a Bernoulli random variable with the parameter given by the scaled local sample average. Thus each machine communicates only 1 bit. The global machine takes the sample average of all these random variables to give a value between 0 and 1 which we can scale to give an estimator for the true mean.

We have that Π_3^{norm} communicates only 1 bit per machine, which is much less than the communication in Π_2^{norm}. We are interested in comparing the performance of protocols Π_2^{norm} and Π_3^{norm} from the above examples. We will investigate this in our simulations in Section 3.

3 Simulations

In this section we will simulate some of the protocols covered in the pervious examples.

We compare different protocols given in the examples. In figures Figure 1, Figure 2 and Figure 3, we vary the number of machines m and the size of the dataset per machine n, such that $mn = 10^4$ is taken to be fixed. We hereby fix the global sample size to study how the number of machines effects the number of bits communicated and the sample risk.

In each figure below we have 5 bar charts and a boxplot. Each bar chart represents a protocol with some m, n given in the title of the chart.

In each bar chart we make 100 trials. For each trial we generate new sample data from the applicable distribution. We carry out the protocol with associated m and n, which gives us an estimator $\hat{\theta}$ for the true mean θ, allowing us to compute the loss $L_2(\hat{\theta}, \theta)$. Let L_{outcomes} be the set of 100 outcomes of $L_2(\hat{\theta}, \theta)$ with the estimator $\hat{\theta}$ computed in each trial. We discretise the 100 outcomes of L_{outcomes} into 50 equally sized bins, where each bin is labelled with the L_2 average that it represents. Each bar chart is composed of 50 bars. The height of each bar depends on how many outcomes fall into its associated bin and we call this the frequency. The risk can be approximated by a sample risk $\frac{1}{100}\sum_i L_2(\hat{\theta}_i, \theta)$, the sample mean of the L_2 loss outcomes.

In the box plot we display the number of bits communicated with the global machine by each of the protocols over the 100 trials. We note that the scale of the bits communicated is logarithmic.

3.1 Simulations of Protocols for the Uniform Means Model

We compare protocols Π_1^{unif}, Π_2^{unif} and Π_3^{unif} from Example 2.1, Example 2.2 and Example 2.3 in Figure 1.

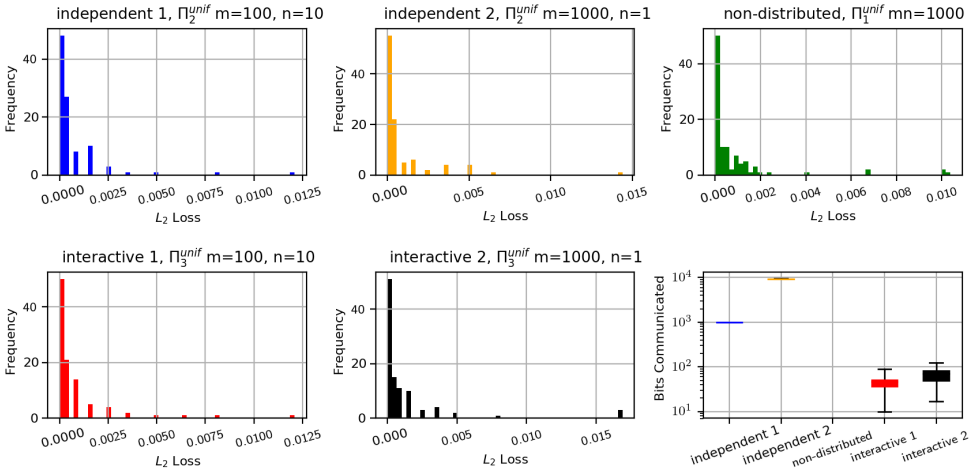

Figure 1: Comparison of protocols Π_1^{unif}, Π_2^{unif} and Π_3^{unif}.

We see that the interactive protocol communicates fewer bits than the independent protocol for the same m (and associated n), and that the protocols perform equally well in terms of the value of the L_2 loss. We can see from how the protocols Π_2^{unif} and Π_3^{unif} are defined in Example 2.2 and Example 2.3 that they must return the same values for the L_2 loss.

3.2 Simulations of Protocols for the Normal Means Model

We compare protocols Π_1^{norm}, Π_2^{norm} and Π_3^{norm} from Example 2.4, Example 2.5 and Example 2.6 in Figure 2 and Figure 3. In the following figures we show the number of machines m for $m = 10, 10^2, 10^3$ and 10^4 with the associated number of data points per machine n, such that $mn = 10^4$.

Not surprisingly we observe that Π_1^{norm}, the non-distributive protocol gives us the best results. We focus on comparing the independent protocols Π_2^{norm} and Π_3^{norm} against each other. This is of interest because from the definition of Π_3^{norm} it is not clear how well it will perform.

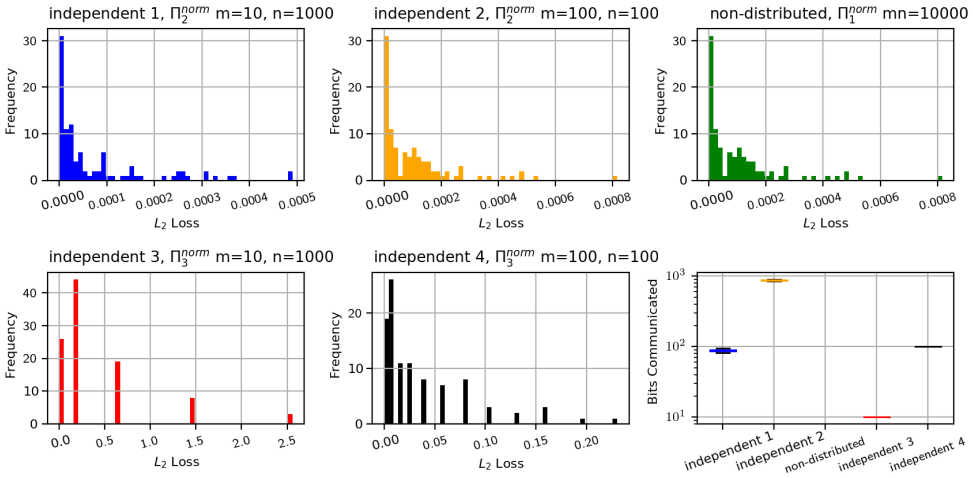

Figure 2: Comparison of protocols Π_1^{norm}, Π_2^{norm} and Π_3^{norm}.

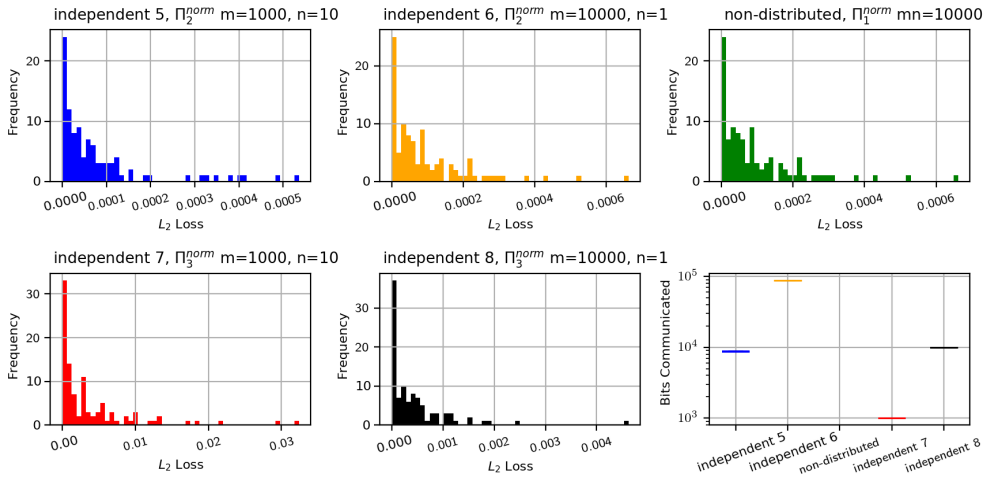

Figure 3: Comparison of protocols Π_1^{norm}, Π_2^{norm} and Π_3^{norm}.

We examine Figure 2, Figure 3 and the protocol Π_2^{norm}. We find for all values of m (and associated n) that Π_2^{norm} performs equally well when considering only the L_2 loss. Increasing m directly increases the number of bits communicated since each machine communicates a set number of bits independent of n.

We investigate protocol Π_3^{norm}. We find that the smaller the value of m, the larger the L_2 loss is. For the same values of $m = 10, 10^2$ and 10^3, we find that Π_2^{norm} gives a much smaller L_2 loss than Π_3^{norm}.

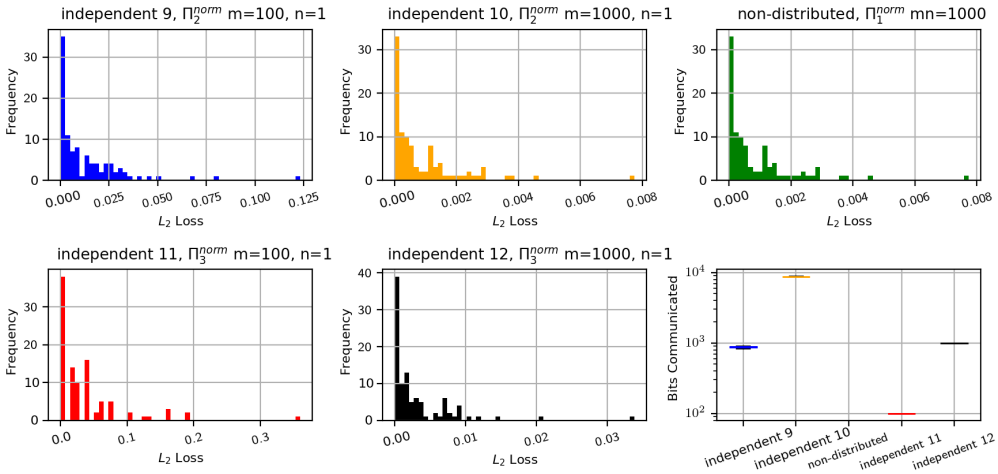

Figure 4: Comparison of protocols Π_1^{norm}, Π_2^{norm} and Π_3^{norm}.

In Figure 4 we keep $n = 1$ and vary only m. We see that Π_2^{norm} and Π_3^{norm} for the same $m = 10^2, 10^3$ and 10^4, give approximately the same values for the L_2 loss.

Our hypothesis is that Π_2^{norm} performs similarly well to Π_3^{norm} in terms of L_2 loss, for

$n = 1$.

We note that for each value of m with $n = 1$ that Π_3^{norm} communicates a factor of 10 less bits than Π_2^{norm}. This makes protocol Π_3^{norm} better in all senses than Π_2^{norm} given that $n = 1$.

4 Theoretical Results for the Example Protocols

Let $X_j^{(i)} \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ for all i, j, where $\sigma^2 = 1$.

Proposition 4.1. *Consider the protocol* Π_2^{norm} *from Example 2.5. Then*

$$R(\theta, \hat{\theta}^{ind}) = \frac{\sigma^2}{mn} + O\left(\frac{1}{(mn)^p}\right).$$

The proof is given in Section A.1.

Proposition 4.2. *Consider the protocol* Π_3^{norm} *from Example 2.6. Then*

$$R(\theta, \hat{\theta}^{ind}) = \frac{1}{m}\left(4a^2 - \theta^2\right) + \mathcal{O}\left(\sqrt{n}e^{-\frac{1}{2}na^2}\right).$$

The proof is given in Section A.2.

We find that the risk function for Π_2^{norm} depends on both m and n while the R_2 risk for Π_3^{norm} depends only on m. Hence for large n protocol Π_2^{norm} performs much better than Π_3^{norm}.

From Proposition 4.1 we see that protocol Π_2^{norm} achieves the minimax lower bound given in Theorem 6.8, ignoring logarithmic factors, as discussed in the end of that section. An important question that we do not cover in this paper is if it is possible to give a protocol that achieves the lower bound given in Theorem 6.8 including logarithmic factors.

5 Review of Information Theory Concepts

In order to understand the proofs given in the upcoming sections we need to review some concepts from information theory. The definitions are summarised from [1] where they can be found for a more comprehensive explanation. Of importance is Definition 5.5 which is key in determining the theoretical lower bounds for the minimax risk in Theorem 6.8 and Theorem 7.4.

Definition 5.1 (Entropy)**.** *Entropy provides an absolute limit on the best possible average length of lossless encoding or compression of an information source.*

The entropy $H(X)$ of a discrete random variable X with sample space $\mathcal{X}$ is defined by

$$H(X) := -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x).$$

The differential entropy $h(V)$ of a continuous random variable V with density $f(v)$ and support S of the random variable is defined as

$$h(V) := \int_S f(v)\log_2 f(v)\, dv.$$

Definition 5.2 (Joint Entropy)**.** *If X and Y are independent, then their joint entropy is the sum of their individual entropies. The joint entropy $H(X,Y)$ of discrete random variables X and Y with $p(x,y)$ and sample space $\mathcal{X}$ and $\mathcal{Y}$, given by*

$$H(X,Y) := -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2(p(x,y)),$$

which can also be written as

$$H(X,Y) = -\mathbb{E}(\log_2 p(X,Y))$$

where $\mathbb{E}$ is the expectation and $p(x,y)$ is the joint probability of x and y occurring together and by convention $0\cdot\log_2(0) := 0$.

The differential entropy of a set $V_1, V_2, ..., V_n$ of continuous random variables with density $f(\mathbf{v})$ where $\mathbf{v} := (v_1, ..., v_n)$, is defined as

$$h(V_1, V_2, ..., V_n) = -\int f(\mathbf{v})\log_2 f(\mathbf{v})\, d\mathbf{v}.$$

Definition 5.3 (Conditional Entropy)**.** *The conditional entropy $H(X|Y)$ is defined as*

$$H(Y|X) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log_2(p(y|x)).$$

If continuous random variables X, Y have a joint density function $f(x,y)$, we can define the conditional differential entropy $h(X|Y)$ as

$$h(X|Y) = -\int f(x,y)\log_2 f(x|y)\, dx\, du.$$

We give some useful properties of entropy H.

1. $H(X) \geq 0$
2. $H_b(X) = (\log_b a) H_a(X)$ where $H_a(X) := -\sum_{x \in \mathcal{X}} p(x) \log_a p(x)$ and $H := H_2$.
3. For any two random variables, X, Y, we have
$$H(X|Y) \leq H(X)$$
with equality if and only if X and Y are independent.
4. $H(X_1, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$, with equality if and only if the X_i are independent.
5. $H(X) \leq \log_2 |\mathcal{X}|$ with equality if and only if X is distributed uniformly over the sample space $\mathcal{X}$.
6. $H(p)$ is concave in p.

Definition 5.4 (Relative Entropy)**.** *The relative entropy, or Kullback-Leibler divergence between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) := \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} = \mathbb{E}_p \log_2 \frac{p(X)}{q(X)}.$$

It is the expectation of the logarithmic difference between the probabilities p and q, where the expectation is taken using probability mass function p. We use the convention that $0 \log_2 \frac{0}{0} := 0$, $0 \log_2 \frac{0}{q} := 0$ and $p \log_2 \frac{p}{0} := \infty$.

The relative entropy $D(f||g)$ between two densities f and g is defined by

$$D(f||g) = \int f(v) \log_2 \frac{f(v)}{g(v)} dv.$$

Relative entropy compares the entropy of two distributions over the same random variable. We note that it is not a metric since it does not satisfy sub-additivity nor symmetry. A relative entropy of 0 indicates that we can expect similar behaviour of the two distributions and a large relative entropy indicates that the two distributions have very much different behaviours.

Definition 5.5 (Mutual Information)**.** *The mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y, given by*

$$\begin{aligned} I(X;Y) : &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)||p(x)p(y)) \\ &= H(X) - H(X|Y). \end{aligned} \tag{1}$$

The mutual information $I(V;W)$ between two continuous random variables with joint density $f(v,w)$ is defined by

$$I(V;W) := \int f(v,w) \log_2 \frac{f(v,w)}{f(v)f(w)} \, dv \, dw.$$

By symmetry we have that

$$I(X;Y) = H(Y) - H(Y|X).$$

Intuitively, the mutual information of $I(X;Y)$ quantifies the number of bits obtained about X, through Y; it measures the information that X and Y share.

Definition 5.6 (Conditional Mutual Information)**.** *The conditional mutual information of random variables X and Y given Z is defined by*

$$I(X;Y|Z) := H(X|Z) - H(X|Y,Z) = \mathbb{E}_{p(x,y,z)} \log_2 \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z))}. \tag{2}$$

Theorem 5.1 (Chain Rules)**.** *We have that H satisfies the chain rules*

$$H(X,Y) = H(X) + H(Y|X),$$

and

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z).$$

Let $X_1, ..., X_n$ be drawn according to $p(x_1, ..., x_n)$, then

$$H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_1, ..., X_{i-1}).$$

The chain rule for mutual information, is given by

$$I(X_1, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, ..., X_{i-1}). \tag{3}$$

Definition 5.7 (Markov Chain)**.** *Random variables X, Y and Z form a Markov chain $X \to Y \to Z$ if the condition distribution of Z depends only on Y and is conditionally independent of X.*

More formally X, Y and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as $p(x, y, z) = p(x)p(y|x)p(z|y)$.

Theorem 5.2 (Data Processing Inequality)**.** *If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.*

6 Minimax Lower Bound for the Normal Means Model

In this section we investigate the work in [6] to find a theoretical lower bound for the minimax risk under the normal means model. For completeness we recall their results and in doing so we remove some of the inaccuracies found in their proofs. We use this as a basis for the next section in which we extend this work to give a lower bound for the minimax risk under the Laplace means model.

We look at only independent protocols under the normal means model in this section. We define the set of independent protocols $\mathcal{A}_{\text{ind}}(B, \mathcal{P})$ satisfying certain budget constraints.

Consider arbitrary number of machines $m \in \mathbb{N}$ where each machine has $n \in \mathbb{N}$ samples from some probability distribution $P \in \mathcal{P}$.

Definition 6.1. *Let each machine have communication budget of $B_i \in \mathbb{N}$ number of bits and let $B = \{B_1, ..., B_m\}$. For some protocol and dataset, let the minimum number of bits communicated be given by L_i as in Definition 1.5.*

Given a family of distributions $\mathcal{P}$, the class of independent protocols is given by

$$\mathcal{A}_{ind}(B, \mathcal{P}) := \left\{ \text{independent protocols } \Pi \ : \ \sup_{P \in \mathcal{P}} \mathbb{E}_P \left(L_i \right) \leq B_i, \ \text{for all } i \in \{1, ..., m\} \right\}.$$

We use the minimax principal to minimise the possible loss for a maximum loss scenario for the R_2 risk of an estimator given by a protocol.

Definition 6.2. *Let each machine have communication budget of $B_i \in \mathbb{N}$ number of bits and let $B = \{B_1, ..., B_m\}$. Let θ be some function of P. A protocol $\Pi \in \mathcal{A}_{ind}(B, \mathcal{P})$ gives an estimator $\hat{\theta}_\Pi := \hat{\theta}(\mathbf{Y})$. Define the minimax risk for the independent protocol as*

$$\mathfrak{M}^{ind}(\theta, \mathcal{P}, B) := \inf_{\Pi \in \mathcal{A}_{ind}(B, \mathcal{P})} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left(||\hat{\theta}_\Pi(\mathbf{Y}) - \theta(P)||_2^2 \right). \tag{4}$$

We will build up to Theorem 6.8, where we will find a lower bound for the above minimax risk given in Definition 6.2, where for $\mathcal{P}$ we take a d-dimensional normal distribution family and where we let $\theta(P)$ be the mean of probability distribution P for $P \in \mathcal{P}$. The lower bound is of interest because it will give us the minimum number of bits required such that the minimax risk gives an order optimal result. The proof to find the lower bound is not constructive and thus it does not give a protocol that satisfies this lower bound.

For the moment we will consider the general case where $\mathcal{P}$ is a set of probability distributions and where $\theta(P)$ is some function on P for $P \in \mathcal{P}$.

Let $\mathcal{V} := \{-1, 1\}^d$ where $\mathcal{V}$ indexes a family of probability distributions $\{P_\nu\}_{\nu \in \mathcal{V}} \subset \mathcal{P}$, such that $\theta_\nu := \theta(P_\nu) = \delta\nu \in \Theta$ where $\delta > 0$ fixed. Sample V uniformly at random from $\mathcal{V}$. Sample X from distribution $P_{V=\nu}$. Then from Equation (4)

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B) \geq \inf_{\Pi \in \mathcal{A}_{\text{ind}}(B, \{P_\nu\}_{\nu \in \mathcal{V}})} \sup_{P \in \{P_\nu\}_{\nu \in \mathcal{V}}} \mathbb{E}_P \left(||\hat{\theta}_\Pi(\mathbf{Y}) - \theta(P)||_2^2 \right). \tag{5}$$

This lower bound is useful since we now take the supremum over a countable set.

We consider the data to be d-dimensional. Let machine $i \in \{1, ..., m\}$ have data sample $X^{(i)} \in \mathbb{R}^{d \times n}$. The k^{th} column of $X^{(i)}$ is $X^{(i,k)}$ and the j^{th} row of $X^{(i)}$ is $X_j^{(i)}$.

We have from above that V gives us the probability distribution where we sample X from. Hence the data $X^{(i)}$ in the i^{th} machine depends on V. We transmit the information Y_i based only on $X^{(i)}$ to the global machine. Thus we have the Markov chain $V \to X^{(i)} \to Y_i$.

Definition 6.3. *The Hamming distance between $\nu, \nu' \in \mathcal{V}$ is the number of positions at which the corresponding values in the vector are different, given by $d_{ham}(\nu, \nu')$.*

We use Lemma 6.1 and Lemma 6.2 to give a lower bound for the minimax risk (5) in terms of mutual information between V and $\mathbf{Y}$, where the result is given in Lemma 6.3.

Lemma 6.1 (Lemma 1 of [6])**.** *Let V be uniformly sampled from $\mathcal{V}$. For any estimator $\hat{\theta}$ and any $t \in \mathbb{R}$ where $t \geq \frac{1}{4}$ we have*

$$\sup_{P \in \{P_\nu\}_{\nu \in \mathcal{V}}} \mathbb{E}(||\hat{\theta} - \theta(P)||_2^2) \geq \delta^2(\lfloor t \rfloor + 1) \inf_{\hat{\nu} \in \mathcal{V}} \mathbb{P}(d_{ham}(\hat{\nu}, V) > t).$$

The proof is given in Section A.3.

Lemma 6.2 (Corollary 1 of [2])**.** *Let $V \to X \to \hat{\nu}$ be a Markov chain, where V is uniformly distributed on $\mathcal{V}$. For $t < \frac{1}{2}(d-1)$, $d \geq 2$ and $\hat{\nu} \in \mathcal{V}$ we have*

$$\mathbb{P}(d_{ham}(\hat{\nu}, V) > t) \geq 1 - \frac{I(V;X) + \log_2 2}{\log_2 \frac{|\mathcal{V}|}{N_t^{max}}}$$

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

The proof is given in Section A.4.

Remark 1. *We have that $N_t^{max} \leq 2\binom{d}{t}$ for $0 \leq t \leq \frac{d+1}{3}$.*

The proof is given in Section A.5.

Remark 2. *We have that $\binom{d}{t} \leq \left(\frac{de}{t}\right)^t$.*

The proof is given in Section A.6. We use Remark 1 and Remark 2 to help prove the following Lemma 6.3.

Lemma 6.3 (Lemma 2 of [6])**.** *For $d \geq 3$ we have from Lemma 6.1 and Lemma 6.2 that*

$$\sup_{P \in \{P_\nu\}_{\nu \in \mathcal{V}}} \mathbb{E}(||\hat{\theta}(\mathbf{Y}) - \theta(P)||_2^2) \geq \delta^2(\lfloor d/6 \rfloor + 1)\left(1 - \frac{I(V;\mathbf{Y}) + 1}{d/6}\right).$$

The proof is given in Section A.7. We use this to prove Theorem 6.8 for the case that the dimension $d \geq 9$, by giving an upper bound for the mutual information $I(V;\mathbf{Y})$.

For small d we need to consider a different lower bound for the minimax risk.

Lemma 6.4 (Appendix A, Le Cam's Method of [6])**.** *We consider the d-dimensional case for $d \leq 8$. By taking the supremum over a smaller set, we have*

$$\begin{aligned}\mathfrak{M}^{ind}(\theta, \mathcal{P}, B) &\overset{(5)}{\geq} \inf_{\Pi \in \mathcal{A}_{ind}(B, \{P_\nu\}_{\nu \in \mathcal{V}})} \sup_{P \in \{P_\nu\}_{\nu \in \mathcal{V}}} \mathbb{E}_P\left(||\hat{\theta}_\Pi(\mathbf{Y}) - \theta(P)||_2^2\right) \\ &\geq \delta^2 \frac{d}{16}\left(1 - \sqrt{2I(V;\mathbf{Y})}\right).\end{aligned}$$

The proof is given in Section A.8.

We use Lemma 6.4 to prove Theorem 6.8 for $d < 9$ by finding an upper bound for $I(V; \mathbf{Y})$.

Proposition 6.1 (Appendix A , Tensorisation of information [6]). *Consider a protocol* Π *such that* Y_i *is constructed based only on* $X^{(i)}$. *Then from we have*

$$I(V; \mathbf{Y}) \leq \sum_{i=1}^{m} I(V; Y_i).$$

The proof is given in Section A.9

Using Proposition 6.1 we need to find an upper bound for $I(V; Y_i)$.

Lemma 6.5 (Lemma 4 of [6]). *Let* V *be sampled uniformly at random from* $\{-1, 1\}^d$. *For any pair* (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.

Define $S_0 := \{x \in \mathbb{R}^n : |\sum_{l=1}^{n} x_l| \leq \sqrt{n}a\}$. *Assume that*

$$\sup_{S \in \sigma(S_0)} \frac{P_\nu(S)}{P_{\nu'}(S)} \leq \exp(\alpha),$$

where $\sigma(S_0)$ *the collection of measurable subsets of the set* S_0. *Define the random variable* $E_j := \mathbb{1}_{X_j^{(i)} \in S_0}$. *Then*

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

The proof is given in [6, Lemma 4].

We note that Lemma 6.4, Lemma 6.3 and Lemma 6.5 have not depended on any distribution family for $\mathcal{P}$. This is useful because it means we can use the same theory in order to find lower bounds for the minimax risk for different distribution families using the same principals.

Consider the d-dimensional normal family, with $\sigma^2 \in \mathbb{R}_{>0}$ as

$$\mathcal{N}_d := \left\{N(\theta, \sigma^2 I_{d\times d}) : \theta \in [-1, 1]^d\right\} \subset \mathcal{P}.$$

Then Lemma 6.7 will give us an upper bound for $I(V; Y_i)$ assuming that the data is sampled from P for $P \in \mathcal{N}_d$ and that $\theta(P)$ gives the mean of normal probability distribution P. For the proof of Lemma 6.7 we require Kullback-Leibler divergence for the multivariate normal distribution.

Lemma 6.6 (Kullback-Leibler divergence for the Multivariate Normal distribution). *Let* $p(\mathbf{x})$ *and* $q(\mathbf{x})$ *be probability density functions with* $p(\mathbf{x})$ *mean* μ *and variance* $\sigma^2 I_{d\times d}$, *and* $q(\mathbf{x})$ *with mean* m *and variance* $\sigma^2 I_{d\times d}$. *Then*

$$D(p||q) = \frac{1}{2\sigma^2} \sum_{i=1}^{d} (\mu_i - m_i)^2.$$

The proof is given in Section A.11.

Lemma 6.7 (Lemma 5 of [6]). *Let $a > 0$ and $\delta > 0$ be chosen such that $\frac{\sqrt{n}a\delta}{\sigma^2} \leq \frac{1.2564}{4}$ for any $i \in \{1, ..., m\}$, and let $h(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ be the binary entropy. Let*

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

Then using Lemma 6.5 we have

$$I(V;Y_i) \leq \frac{n\delta^2}{\sigma^2}b_i + dh\left(2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right)\right) + 2d\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right). \tag{6}$$

The proof is given in [6, Lemma 5].

We split the proof for Theorem 6.8 into two cases, for $d \geq 9$ and for $d < 9$. For $d \geq 9$ we use Lemma 6.3 and for $d < 9$ we use Lemma 6.4, both giving a lower bound for the minimax risk from (5) in terms of the mutual information $I(V;\mathbf{Y})$. Using Lemma 6.7 we give an upper bound for $I(V;\mathbf{Y})$. With this in hand we can give upper bounds for the three terms on the right hand side of (6) and together to give Theorem 6.8.

Theorem 6.8 (Theorem 2 of [6]). *Let $B = \{B_1, ..., B_m\}$ where each machine has communication budget $B_i \in \mathbb{N}$ and receives an i.i.d. sample of size n from a distribution $P \in \mathcal{N}_d$. Then*

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

where $c = 4.6875 \cdot 10^{-8}$.

The proof is given in Section A.10.

To achieve an order optimal result for the lower bound in Theorem 6.8, the total number of bits communicated per machine must scale with d. Then the lower bound is of order $\frac{\sigma^2 d}{mn\log m}$. We can construct a protocol that almost achieves this order optimal result up to a factor of $\frac{1}{\log m}$. We gave such a protocol earlier in Proposition 4.1 using protocol Π_2^{norm} for the 1-dimensional case. A protocol that achieves the same order but for the d-dimensional case follows similarly to that example.

We wonder how the lower bound for the minimax risk depends on the probability distribution family. In the next section we will consider the Laplace means model and show that for a d-dimensional Laplace family we obtain the same lower bound for the minimax risk up to constant factors. This is surprising since the normal distribution has much lighter tails than the Laplace distribution. We go on to ask ourselves how the set of distribution families that attain the same lower bound up to constant factors would be described, however this goes beyond the bounds of this paper.

7 Minimax Lower Bound for the Laplace Means Model

In this section we will derive a similar proof to Theorem 6.8 but for the d-dimensional Laplace family. To do this we need to formulate a similar proof to Lemma 6.7 but for Laplace, this is given in Lemma 7.3.

Define the d-dimensional Laplace family, with $b \in \mathbb{R}_{>0}$ as

$$\mathcal{L}_d := \left\{ \text{Laplace}(\theta, bI_{d\times d}) : \theta \in [-1,1]^d \right\}.$$

Similarly to Lemma 6.7 for the normal means model, the proof of Lemma 7.3 requires the Kullback-Leiber divergence for the multivariate Laplace distribution.

Lemma 7.1 (Kullback-Leibler divergence for Univariate Laplace distribution)**.** *Let $p(x)$ and $q(x)$ be Laplace probability density functions with $p(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$ and $q(x) = \frac{1}{2b}\exp\left(-\frac{|x-m|}{b}\right)$, where $\mu, m \in \mathbb{R}$ are the locations and $b \in \mathbb{R}_{>0}$ is the scale. Then*

$$D(p||q) = e^{-\frac{|\mu-m|}{b}} + \frac{|\mu-m|}{b} - 1.$$

The proof is given in Section A.12.

Lemma 7.2 (Kullback-Leibler divergence for Multivariate Laplace distribution)**.** *Let $p(\mathbf{x})$ and $q(\mathbf{x})$ be Laplace probability density functions where $p(\mathbf{x})$ has location μ and covariance $bI_{d\times d}$ and $q(\mathbf{x})$ has location $\mathbf{m}$ and covariance $bI_{d\times d}$. Then*

$$\begin{aligned} D(p||q) &= e^{-\frac{1}{b}\sum_{i=1}^d |\mu_i - m_i|} + \frac{1}{b}\sum_{i=1}^d |\mu_i - m_i| - 1 \\ &= e^{-\frac{1}{b}||\mu-m||_1} + \frac{1}{b}||\mu - m||_1 - 1. \end{aligned}$$

Proof. This follows similarly to Lemma 7.1. □

We split the proof of Theorem 7.4 into two cases, for $d \geq 9$ and for $d < 9$. For both cases we can use the same lower bounds used for Theorem 6.8, namely for $d \geq 9$ we use Lemma 6.3 and for $d < 9$ we use Lemma 6.4. Both give us a lower bound in terms of $I(V;\mathbf{Y})$ that is not dependent on the distribution family. In order to give an upper bound for $I(V;\mathbf{Y})$ we consider Lemma 7.3 and use this in conjunction with Proposition 6.1.

Lemma 7.3. *Let $a > 0$ and $\delta > 0$ be chosen such that $\frac{\delta\sqrt{n_i}a}{b} \leq \frac{1.2564}{4}$ for any $i \in \{1,...,m\}$, and let $h(p) = -p\log(p) - (1-p)\log(1-p)$ be the binary entropy. Then*

$$I(V;Y_i) \leq \frac{n_i\delta^2}{b^2}\min\left\{128a^2H(Y_i), d\right\} + dh\left(2\exp\left(-\frac{a^2}{32}\right)\right) + 2d\exp\left(-\frac{a^2}{32}\right). \quad (7)$$

The proof is given in Section A.13.

The upper bound from Lemma 7.3 depends on the Laplace distribution family $\mathcal{L}_d$. With this in hand, we can give upper bounds for the three terms on the right hand side of Equation (7) to give an upper bound on $I(V;\mathbf{Y})$. We plug this result back into Lemma 6.3 for $d \geq 9$ and Lemma 6.4 for $d < 9$ to give the result in Theorem 7.4.

Theorem 7.4. *Let $B = \{B_1, ..., B_m\}$ where each machine has communication budget $B_i \in \mathbb{N}$ and receives an i.i.d. sample of size n from a distribution $P \in \mathcal{L}_d$. Then*

$$\mathfrak{M}^{ind}(\theta, \mathcal{L}_d, B) \geq c\frac{\sigma^2 d}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log m}, \frac{m}{\log m \left(\sum_{i=1}^{m} \min\left\{1, \frac{B_i}{d}\right\}\right)}\right\}.$$

where $c = 2.44 \cdot 10^{-9}$.

The proof is given in Section A.14.

We find that the lower bound given here is the same as the lower bound found in Theorem 6.8, ignoring constant factors.

A Appendix

We give the proofs referred to in the paper.

A.1 Proof of Proposition 4.1

Proof. Let $\overline{X}^{(i)} := \frac{1}{n}\sum_{j=1}^{n} X_j^{(i)}$ calculated by machine i.

Remark 3 (Integers and Binary)**.** *A positive integer $n \in \mathbb{N}$ can be written with b bits in binary notation when $2^{b-1} \leq n \leq 2^b - 1$. Then $b = \lfloor \log_2(n) \rfloor + 1$.*

Let $\overline{X}^{(i)} = y_i + r_i$ where y_i is the first $b := \lceil p \log_2(mn) \rceil$ bits of $\overline{X}^{(i)}$ we transmit for $p \in \mathbb{N}$, and $r_i = \overline{X}^{(i)} - y_i$ the error term in the transmitted data.

We take the binary representation $\overline{X}^{(i)} = \pm(1.a_1a_2...a_b...)2^k$ for $a_1, a_2, ... \in \{0, 1\}$ digits and $k \in \mathbb{Z}$. Then $y_i = \pm(1.a_1a_2...a_b)2^k$ and $r_i = \pm(0.0...0a_{b+1}...)2^k$. We have $|\overline{X}^{(i)}| \geq 1 \cdot 2^k$ and $|r_i| \leq (0.0...01)2^k$ where the 1 is on the place of a_b. Then

$$\left| \frac{r_i}{\overline{X}^{(i)}} \right| \leq \frac{(0.0...01)2^k}{(1)2^k} \leq \frac{2^{-b}}{1} \leq \frac{1}{(mn)^p}.$$

Then

$$y_i = \overline{X}^{(i)} - \overline{X}^{(i)} \left(\frac{r_i}{\overline{X}^{(i)}} \right) = \overline{X}^{(i)} \left(1 + O\left(\frac{1}{(mn)^p} \right) \right),$$

and

$$\sum_{i=1}^{m} y_i = \left(1 + O\left(\frac{1}{(mn)^p} \right) \right) \sum_{i=1}^{m} \overline{X}^{(i)}.$$

In the global machine we receive $y_1, ..., y_m$ and calculate $\bar{y} := \frac{1}{m}\sum_{i=1}^{m} y_i$. Then $\bar{y} = \frac{1}{m}\left(1 + O\left(\frac{1}{(mn)^p}\right)\right)\sum_{i=1}^{m} \overline{X}^{(i)}$. We want to determine $\mathbb{E}((\bar{y} - \theta)^2)$.

We have that

$$\mathbb{E}((\bar{y} - \theta)^2) = \mathbb{E}(\bar{y}^2) + \theta^2 - 2\theta\mathbb{E}(\bar{y}). \tag{8}$$

We determine the terms on the right hand side of the above equation. We have

$$\begin{aligned}
\mathbb{E}(\bar{y}) &= \frac{1}{m}\left(1 + O\left(\frac{1}{(mn)^p}\right)\right)\mathbb{E}\left(\sum_{i=1}^{m} \overline{X}^{(i)}\right) \\
&= \left(1 + O\left(\frac{1}{(mn)^p}\right)\right)\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(\overline{X}^{(i)}\right) \\
&= \left(1 + O\left(\frac{1}{(mn)^p}\right)\right)\theta.
\end{aligned}$$

By squaring $\bar{y}^2$ we have

$$\mathbb{E}(\bar{y}^2) = \mathbb{E}\left(\frac{1}{m^2}\left(1+O\left(\frac{1}{(mn)^p}\right)\right)^2\left(\sum_{i=1}^{m}\overline{X}^{(i)}\right)^2\right) \tag{9}$$

$$= \frac{1}{m^2}\left(1+O\left(\frac{1}{(mn)^p}\right)\right)^2\mathbb{E}\left(\left(\sum_{i=1}^{m}\overline{X}^{(i)}\right)^2\right). \tag{10}$$

Since $\overline{X}^{(i)}$ are independent random variables, $\mathbb{V}\text{ar}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ and $\mathbb{V}\text{ar}(aX) = a^2\mathbb{V}\text{ar}(X)$ we have

$$\begin{aligned}
\mathbb{E}\left(\left(\sum_{i=1}^{m}\overline{X}^{(i)}\right)^2\right) &= \mathbb{V}\text{ar}\left(\sum_{i=1}^{m}\overline{X}^{(i)}\right) + \mathbb{E}^2\left(\sum_{i=1}^{m}\overline{X}^{(i)}\right) \\
&= \sum_{i=1}^{m}\mathbb{V}\text{ar}(\overline{X}^{(i)}) + \left(\sum_{i=1}^{m}\mathbb{E}(\overline{X}^{(i)})\right)^2 \\
&= \sum_{i=1}^{m}\mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{j=1}^{n}X_j^{(i)}\right) + \left(\sum_{i=1}^{m}\theta\right)^2 \\
&= \sum_{i=1}^{m}\frac{1}{n^2}\mathbb{V}\text{ar}\left(\sum_{j=1}^{n}X_j^{(i)}\right) + (m\theta)^2 \\
&= \sum_{i=1}^{m}\frac{\sigma^2}{n} + (m\theta)^2 \\
&= \frac{m\sigma^2}{n} + m^2\theta^2.
\end{aligned}$$

Hence by substituting the above into Equation (9) we arrive at

$$\begin{aligned}
\mathbb{E}(\bar{y}^2) &= \frac{1}{m^2}\left(1+O\left(\frac{1}{(mn)^p}\right)\right)^2\left(\frac{m\sigma^2}{n}+m^2\theta^2\right) \\
&= \left(1+O\left(\frac{1}{(mn)^p}\right)\right)\left(\frac{\sigma^2}{mn}+\theta^2\right).
\end{aligned}$$

Then by substituting into Equation (8) we find

$$\begin{aligned}
\mathbb{E}((\bar{y}-\theta)^2) &= \left(1+O\left(\frac{1}{(mn)^p}\right)\right)\left(\frac{\sigma^2}{mn}+\theta^2\right) + \theta^2 - 2\theta^2\left(1+O\left(\frac{1}{(mn)^p}\right)\right) \\
&= \frac{\sigma^2}{mn} + \theta^2 + \left(\frac{\sigma^2}{mn}+\theta^2\right)O\left(\frac{1}{(mn)^p}\right) + \theta^2 - 2\theta^2 - 2\theta^2 O\left(\frac{1}{(mn)^p}\right) \\
&= \frac{\sigma^2}{mn} - \left(\frac{\sigma^2}{mn}-\theta^2\right)O\left(\frac{1}{(mn)^p}\right) \\
&= \frac{\sigma^2}{mn} + O\left(\frac{1}{(mn)^p}\right).
\end{aligned}$$

□

A.2 Proof of Proposition 4.2

Proof. Let us assume that $\theta \in [-a, a]$ for some $a \in \mathbb{R}_{>0}$ and $\overline{X}^{(i)} \sim N\left(\theta, \frac{1}{n}\right)$ as in protocol Π_3^{norm} given in Example 2.6. We introduce the definition for the truncated normal distribution, since in the above protocol we compute the sample average $\overline{X}^{(i)}$ and then take $\overline{X}_*^{(i)} := (\overline{X}^{(i)} \vee -2a) \wedge 2a$, the truncation of $\overline{X}^{(i)}$ to the interval $[-2a, 2a]$.

Definition A.1 (Truncated normal distribution)**.** *Suppose $X \sim N(\mu, \sigma^2)$ has a normal distribution and lies within the interval $X \in (a, b)$ with $a < b$ and $a, b \in \mathbb{R}$. Then X conditional on $a < X < b$ has a truncated normal distribution $N_{trunc}(\mu, \sigma^2)$ with probability density function f for $a \leq x \leq b$ given by*

$$f(x) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma\left(\Phi\left(\beta\right) - \Phi\left(\alpha\right)\right)} \tag{11}$$

and by $f = 0$ otherwise, where $\alpha = \frac{a-\mu}{\sigma}$ and $\beta = \frac{b-\mu}{\sigma}$.

Here ϕ is the probability density function of the standard normal distribution, given by $\phi(\zeta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\zeta^2\right)$ and Φ the cumulative distribution function.

The mean is given by

$$\mathbb{E}(X|a < X < b) = \mu + \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\sigma.$$

The variance is given by

$$\mathbb{V}ar(X|a < X < b) = \sigma^2\left(1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right)^2\right).$$

By Definition A.1 we have that $\overline{X}_*^{(i)}$ is truncated normal distributed $\overline{X}_*^{(i)} \sim N_{\text{trunc}}(\mu, \sigma^2)$ with $\mu := \theta, \sigma^2 := \frac{1}{n}$. Take

$$\alpha := \frac{-2a - \mu}{\sigma} = \sqrt{n}(-2a - \theta) \quad \text{and} \quad \beta := \frac{2a - \mu}{\sigma} = \sqrt{n}(2a - \theta).$$

Then by definition

$$\mathbb{E}(\overline{X}_*^{(i)}) = \theta + \frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\frac{1}{\sqrt{n}}.$$

We have that

$$\alpha \leq -\sqrt{n}a \quad \text{and} \quad \beta \geq \sqrt{n}a.$$

Note that Φ is non-decreasing, $\Phi(x) = 1 - \Phi(-x)$ and $\Phi(0) = 1/2$. We have

$$\begin{aligned}
\Phi(\beta) - \Phi(\alpha) &\geq \Phi(\sqrt{n}a) - \Phi(-\sqrt{n}a) \\
&= \Phi(\sqrt{n}a) - \left(1 - \Phi(\sqrt{n}a)\right) \\
&= 2\Phi(\sqrt{n}a) - 1 \\
&= 2\left(\Phi(\sqrt{n}a) - \Phi(0)\right).
\end{aligned} \tag{12}$$

Then for any a and large enough n such that $\sqrt{n}a > 1$, we have

$$\begin{aligned}\Phi(\beta) - \Phi(\alpha) &\geq 2\left(\Phi(1) - \Phi(0)\right) \\ &> 2(0.84 - 1/2) = 0.17 > 0.\end{aligned}$$

We have

$$\alpha^2 = n(-2a - \theta)^2 \geq na^2 \quad \text{and} \quad \beta^2 = n(2a - \theta)^2 \geq na^2.$$

Then

$$\phi(\alpha) \vee \phi(\beta) \leq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}na^2}.$$

Since $\phi(\zeta) > 0$ we have

$$|\phi(\alpha) - \phi(\beta)| \leq \max\left\{\phi(\alpha), \phi(\beta)\right\} \leq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}na^2}. \tag{13}$$

Thus by Equation (12) and Equation (13) we get

$$\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} = \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right).$$

Hence

$$\mathbb{E}(\overline{X}_*^{(i)}) = \theta + \frac{1}{\sqrt{n}} \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right). \tag{14}$$

By definition, the variance of $\overline{X}_*^{(i)}$ is given by

$$\mathbb{V}\mathrm{ar}(\overline{X}_*^{(i)}) = \sigma^2 \left(1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right)^2\right).$$

Applying again Equation (12) and Equation (13) we get

$$\left(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right)^2 = \mathcal{O}\left(e^{-na^2}\right).$$

Similarly as above we have

$$\begin{aligned}|\alpha\phi(\alpha) - \beta\phi(\beta)| &\leq \max\left\{|\alpha|\phi(\alpha), \beta\phi(\beta)\right\} \\ &\leq \max\left\{\sqrt{n}|-2a-\theta|\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}na^2}, \sqrt{n}(2a-\theta)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}na^2}\right\} \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{1}{2}na^2} \max\left\{|-2a-\theta|, (2a-\theta)\right\} \\ &\leq 3a\frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{1}{2}na^2}.\end{aligned}$$

Then $\left|\frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right| = \mathcal{O}\left(\sqrt{n}e^{-\frac{1}{2}na^2}\right)$. Thus

$$\mathbb{V}\mathrm{ar}(\overline{X}_*^{(i)}) = \sigma^2 + \mathcal{O}\left(\sqrt{n}e^{-\frac{1}{2}na^2}\right). \tag{15}$$

Consider $Z_1, ..., Z_m$ Bernoulli random variables where the probability of success for Z_i is

$$p_i := \frac{\overline{X}_*^{(i)} + 2a}{4a}.$$

Consider an estimator $\hat{\theta}$ for the mean θ given by

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\left(Z_i 4a - 2a\right).$$

We have that $\hat{\theta}$ is an asymptotically unbiased estimator since, using the law of total expectation

$$\begin{aligned}
\mathbb{E}(\hat{\theta}) = \mathbb{E}(\mathbb{E}(\hat{\theta}|X)) &= \mathbb{E}\left(\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\left(Z_i 4a - 2a\right)|X\right)\right) \\
&= \mathbb{E}\left(\frac{1}{m}\left(\sum_{i=1}^{m}\mathbb{E}(Z_i|X)4a - 2a\right)\right) = \mathbb{E}\left(\frac{1}{m}\left(\sum_{i=1}^{m}\frac{\overline{X}_*^{(i)} + 2a}{4a}4a - 2a\right)\right) \\
&= \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\overline{X}_*^{(i)}\right) = \frac{1}{m}\sum_{i=1}^{m}E(\overline{X}_*^{(i)}) = \theta + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right). \qquad (16)
\end{aligned}$$

We now want to determine an upperbound for the expectation of the L_2 distance of $\hat{\theta}$ from θ, in other words of

$$\begin{aligned}
\mathbb{E}((\hat{\theta} - \theta)^2) &= \mathbb{E}(\hat{\theta}^2) + \theta^2 - 2\theta\mathbb{E}(\hat{\theta}) \\
&\overset{(16)}{=} \mathbb{E}(\hat{\theta}^2) + \theta^2 - 2\theta^2 + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right) \\
&= \mathbb{E}\left(\left[\frac{1}{m}\sum_{i=1}^{m}\left(Z_i 4a - 2a\right)\right]^2\right) - \theta^2 + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right) \\
&= \mathbb{E}\left(\left[-2a + \frac{4a}{m}\sum_{i=1}^{m}Z_i\right]^2\right) - \theta^2 + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right) \\
&= \mathbb{E}\left(4a^2 + \left(\frac{4a}{m}\sum_{i=1}^{m}Z_i\right)^2 - \frac{16a^2}{m}\sum_{i=1}^{m}Z_i\right) - \theta^2 + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right) \\
&= 4a^2 + \frac{16a^2}{m^2}\mathbb{E}\left(\left(\sum_{i=1}^{m}Z_i\right)^2\right) - \frac{16a^2}{m}\mathbb{E}\left(\sum_{i=1}^{m}Z_i\right) - \theta^2 + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right). \qquad (17)
\end{aligned}$$

We now want to find the unknown terms on the right hand side of Equation (17). Again using the law of total expectation

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{m}Z_i\right) &= \sum_{i=1}^{m}\mathbb{E}(\mathbb{E}(Z_i|X)) = \sum_{i=1}^{m}\mathbb{E}\left(\frac{\overline{X}_*^{(i)} + 2a}{4a}\right) \\
&= \sum_{i=1}^{m}\frac{\mathbb{E}(\overline{X}_*^{(i)}) + 2a}{4a} = \frac{m(\theta + 2a)}{4a} + \mathcal{O}\left(me^{-\frac{1}{2}na^2}\right).
\end{aligned}$$

As a consequence we have

$$\mathbb{E}^2\left(\sum_{i=1}^{m}Z_i\right) = \frac{m^2(\theta + 2a)^2}{16a^2} + \mathcal{O}\left(m^2e^{-\frac{1}{2}na^2}\right), \qquad (18)$$

and by the law of total variance

$$\mathbb{E}\left(\left(\sum_{i=1}^{m} Z_i\right)^2\right) = \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{m} Z_i\right) + \mathbb{E}^2\left(\sum_{i=1}^{m} Z_i\right)$$
$$= \mathbb{E}\left(\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{m} Z_i \middle| X\right)\right) + \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left(\sum_{i=1}^{m} Z_i \middle| X\right)\right) + \mathbb{E}^2\left(\sum_{i=1}^{m} Z_i\right). \tag{19}$$

Since the variance of the Bernoulli distribution with parameter p is given by $p(1-p)$, the first term on the right hand side of Equation (19) can be reformulated as

$$\begin{aligned}
\mathbb{E}\left(\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{m} Z_i \middle| X\right)\right) &= \sum_{i=1}^{m} \mathbb{E}\left(\frac{\overline{X}_*^{(i)} + 2a}{4a}\left(1 - \frac{\overline{X}_*^{(i)} + 2a}{4a}\right)\right) \\
&= \frac{1}{4a}\sum_{i=1}^{m} \mathbb{E}\left(\overline{X}_*^{(i)} + 2a - \frac{1}{4a}\left((\overline{X}_*^{(i)})^2 + (-2a)^2 + 4a\overline{X}_*^{(i)}\right)\right) \\
&= \frac{1}{4a}\sum_{i=1}^{m} \left[\mathbb{E}\left(\overline{X}_*^{(i)}\right) + 2a - \frac{1}{4a}\left(\mathbb{E}\left((\overline{X}_*^{(i)})^2\right) + 4a^2 + 4a\mathbb{E}\left(\overline{X}_*^{(i)}\right)\right)\right] \\
&\overset{(14)}{=} \frac{1}{4a}\sum_{i=1}^{m} \left[\theta + 2a - \frac{1}{4a}\left(\mathbb{E}\left((\overline{X}_*^{(i)})^2\right) + 4a^2 + 4a\theta\right)\right] + \mathcal{O}\left(me^{-\frac{1}{2}na^2}\right) \\
&= \frac{1}{4a}\sum_{i=1}^{m} \left[a - \frac{1}{4a}\left(\mathbb{V}\mathrm{ar}\left(\overline{X}_*^{(i)}\right) + \mathbb{E}^2\left(\overline{X}_*^{(i)}\right)\right)\right] + \mathcal{O}\left(me^{-\frac{1}{2}na^2}\right) \\
&\overset{(14)(15)}{=} \frac{m}{4a}\left[a - \frac{1}{4a}\left(\frac{1}{n} + \theta^2\right)\right] + \mathcal{O}\left(m\sqrt{n}e^{-\frac{1}{2}na^2}\right).
\end{aligned}$$

Using the independence of $\overline{X}_*^{(i)}$ and the properties of the variance, the second term on the right hand side of Equation (19) can be written as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\mathbb{E}\left(\sum_{i=1}^{m} Z_i \middle| X\right)\right) &= \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{m} \frac{\overline{X}_*^{(i)} + 2a}{4a}\right) = \frac{1}{16a^2}\sum_{i=1}^{m} \mathbb{V}\mathrm{ar}\left(\overline{X}_*^{(i)} + 2a\right) \\
&= \frac{1}{16a^2}\sum_{i=1}^{m} \mathbb{V}\mathrm{ar}\left(\overline{X}_*^{(i)}\right) \overset{(15)}{=} \frac{m}{16na^2} + \mathcal{O}\left(m\sqrt{n}e^{-\frac{1}{2}na^2}\right).
\end{aligned}$$

Then by substitution of the above into Equation (19) gives

$$\begin{aligned}
\mathbb{E}\left(\left(\sum_{i=1}^{m} Z_i\right)^2\right) &= \frac{m}{4a}\left(a - \frac{1}{4a}\left(\frac{1}{n} + \theta^2\right)\right) + \frac{m}{16na^2} + \frac{m^2(\theta+2a)^2}{16a^2} + \mathcal{O}\left(m^2\sqrt{n}e^{-\frac{1}{2}na^2}\right) \\
&= \frac{m}{4a}\left(a - \frac{\theta^2}{4a}\right) + \frac{m^2(\theta+2a)^2}{16a^2} + \mathcal{O}\left(m^2\sqrt{n}e^{-\frac{1}{2}na^2}\right).
\end{aligned} \tag{20}$$

Finally substituting Equation (18) and Equation (20) into Equation (17) gives

$$\begin{aligned}\mathbb{E}((\hat{\theta}-\theta)^2) =& 4a^2 + \frac{16a^2}{m^2}\left[\frac{m}{4a}\left(a-\frac{\theta^2}{4a}\right) + \frac{m^2(\theta+2a)^2}{16a^2} + \mathcal{O}\left(m^2\sqrt{n}e^{-\frac{1}{2}na^2}\right)\right] \\ & - \frac{16a^2}{m}\left[\frac{m(\theta+2a)}{4a} + \mathcal{O}\left(me^{-\frac{1}{2}na^2}\right)\right] - \theta^2 + \mathcal{O}\left(e^{-\frac{1}{2}na^2}\right) \\ =& 4a^2 + \frac{4a}{m}\left(a-\frac{\theta^2}{4a}\right) + (\theta+2a)^2 - 4a(\theta+2a) - \theta^2 + \mathcal{O}\left(\sqrt{n}e^{-\frac{1}{2}na^2}\right) \\ =& \frac{1}{m}\left(4a^2-\theta^2\right) + \mathcal{O}\left(\sqrt{n}e^{-\frac{1}{2}na^2}\right).\end{aligned}$$

This finishes the proof. □

A.3 Proof of Lemma 6.1

Proof (Lemma 1 of [6]). Consider arbitrary $\triangle > 0$ and arbitrary estimator $\hat{\theta}$. If V is a random variable uniformly chosen from $\mathcal{V}$, then we have

$$\begin{aligned}\sup_{P\in\mathcal{P}} \mathbb{E}\left(||\hat{\theta}-\theta(P)||_2^2\right) &\geq \max_{\nu\in\mathcal{V}} \mathbb{E}\left(||\hat{\theta}-\theta_\nu||_2^2\right) \\ &\geq \mathbb{E}\left(||\hat{\theta}-\theta_V||_2^2\right) \\ &\geq \mathbb{E}\left(\triangle^2 1_{\{||\hat{\theta}-\theta_V||_2\geq\triangle\}}\right) \\ &= \triangle^2 \mathbb{P}\left(||\hat{\theta}-\theta_V||_2 \geq \triangle\right), \end{aligned} \tag{21}$$

where the third inequality follows from

$$\begin{aligned}\mathbb{E}(||\hat{\theta}-\theta_V||_2^2) &= \int ||\hat{\theta}-\theta_V||_2^2 \, dP \\ &= \int_{\{||\hat{\theta}-\theta_V||_2^2\geq\triangle^2\}} ||\hat{\theta}-\theta_V||_2^2 \, dP + \int_{\{||\hat{\theta}-\theta_V||_2^2<\triangle^2\}} ||\hat{\theta}-\theta_V||_2^2 \, dP \\ &\geq \int_{\{||\hat{\theta}-\theta_V||_2^2\geq\triangle^2\}} ||\hat{\theta}-\theta_V||_2^2 \, dP \\ &\geq \triangle^2 \int_{\{||\hat{\theta}-\theta_V||_2\geq\triangle\}} 1 \, dP \\ &= \triangle^2 \mathbb{E}\left(1_{\{||\hat{\theta}-\theta_V||_2\geq\triangle\}}\right).\end{aligned}$$

We now lower bound $\mathbb{P}(||\hat{\theta}-\theta_V||_2 \geq \triangle)$ from Equation (21) by considering

$$\hat{\nu} := \text{argmin}_{\nu\in\mathcal{V}} ||\hat{\theta}-\theta_\nu||_2.$$

Then $||\theta_{\hat{\nu}}-\hat{\theta}||_2 \leq ||\hat{\theta}-\theta_V||_2$. The triangle inequality implies that

$$||\theta_{\hat{\nu}}-\theta_V||_2 \leq ||\theta_{\hat{\nu}}-\hat{\theta}||_2 + ||\hat{\theta}-\theta_V||_2 \leq 2||\hat{\theta}-\theta_V||_2.$$

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

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

Consequently,

$$\mathbb{P}\left(||\hat{\theta}-\theta_V||_2^2 \geq \delta^2(\lfloor t \rfloor+1)\right) \geq \mathbb{P}\left(d_{\text{ham}}(\hat{\nu},V) > t\right). \tag{22}$$

Combining the inequalities in Equation (21) and Equation (22) with $\Delta^2 = \delta^2(\lfloor t \rfloor+1)$, gives

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

On the right hand side of the above inequality, taking infimum over all $\hat{\nu} \in \mathcal{V}$ establishes the result. □

A.4 Proof of Lemma 6.2

Proof (Corollary 1 of [2]). Let $V \to X \to \hat{\nu}$ be a Markov chain, where V is uniform on $\mathcal{V}$. Let

$$N_t^{\max} := \max_{\nu\in\mathcal{V}} |\{\nu' \in \mathcal{V} : d_{\text{ham}}(\nu,\nu') \leq t\}| \quad \text{and} \quad N_t^{\min} := \min_{\nu\in\mathcal{V}} |\{\nu' \in \mathcal{V} : d_{\text{ham}}(\nu,\nu') \leq t\}|.$$

We note that $N_t^{\max} = N_t^{\min}$. We have $|\mathcal{V}| = 2^d$ and $N_t^{\max} = \sum_{i=0}^{t}\binom{d}{i}$.

Let $\rho : V \times V \to \mathbb{R}$ be a symmetric function defined on $V \times V$. From [2, Corollary 1] if V is uniform on $\mathcal{V}$, $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$ and $\hat{\nu} \in \mathcal{V}$, then

$$\mathbb{P}(\rho(\hat{\nu},V) > t) \geq 1 - \frac{I(V;X)+1}{\log_2 \frac{|\mathcal{V}|}{N_t^{\max}}}. \tag{23}$$

We want to find t such that $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$, so that we can use Equation (23). Hence we want to find the largest t such that

$$\sum_{i=0}^{t}\binom{d}{i} < \frac{1}{2}2^d. \tag{24}$$

From the binomial theorem we have that $2^d = \sum_{i=0}^{d}\binom{d}{i}$. For d even we have $2^d = \binom{d}{\frac{1}{2}d} + 2\sum_{i=0}^{\frac{1}{2}d-1}\binom{d}{i}$. Then $\frac{1}{2}2^d > \sum_{i=0}^{\frac{1}{2}d-1}\binom{d}{i}$. So by choosing $t \leq \frac{1}{2}d-1$, (or equivalently $t < \frac{1}{2}d$) we get Equation (24).

For d odd, we have that $2^d = 2\cdot\sum_{i=0}^{\frac{1}{2}(d-1)}\binom{d}{i}$. Then $\frac{1}{2}2^d = \sum_{i=0}^{\frac{1}{2}(d-1)}\binom{d}{i}$. By choosing $t < \frac{1}{2}(d-1)$ we get Equation (24).

By taking $t < \frac{1}{2}(d-1)$ and $t \geq \frac{1}{4}$ we must take $d \geq 2$. Then $(|\mathcal{V}| - N_t^{\min}) > N_t^{\max}$.

Now by Equation (23) we let $\rho := d_{\text{ham}}$ giving

$$\mathbb{P}(d_{\text{ham}}(\hat{\nu},V) > t) \geq 1 - \frac{I(V;X)+1}{\log_2 \frac{|\mathcal{V}|}{N_t^{\max}}}$$

for $t < \frac{1}{2}(d-1)$ and $d \geq 2$. □

A.5 Proof of Remark 1

Proof. Consider $d \in \mathbb{N}$ and fix $0 \leq t \leq \frac{d+1}{3}$.

We have that $\binom{d}{i-1} = \binom{d}{i}\frac{i}{d+1-i} \leq \frac{1}{2}\binom{d}{i}$ for $0 \leq i \leq \frac{d+1}{3}$, since

$$\frac{i}{d+1-i} \leq \frac{\frac{d+1}{3}}{d+1-\frac{d+1}{3}} = \frac{d+1}{3} \cdot \frac{3}{2d+2} = \frac{1}{2}.$$

Then

$$\begin{aligned} N_t^{\max} = \sum_{i=0}^{t} \binom{d}{i} &\leq \frac{1}{2^t}\binom{d}{t} + \frac{1}{2^{t-1}}\binom{d}{t} + \ldots + \frac{1}{2}\binom{d}{t} + \binom{d}{t} \\ &\leq \binom{d}{t} \sum_{i=0}^{\infty} \frac{1}{2^i} \\ &= 2\binom{d}{t}. \end{aligned}$$

□

A.6 Proof of Remark 2

Proof. Since

$$e^t = \sum_{i=0}^{\infty} \frac{t^i}{i!}$$

we have $e^t > \frac{t^t}{t!}$. Hence

$$\binom{d}{t} = \frac{d!}{(d-t)!t!} = \frac{d \cdot (d-1) \cdots (d-(t-1))}{t!} \leq \frac{d^t}{t!} \leq \left(\frac{ed}{t}\right)^t.$$

□

A.7 Proof of Lemma 6.3

Proof (Lemma 2 of [6]). Let $d \geq 9$ and $t \leq \frac{d}{6} \leq \frac{d+1}{3}$. We note that

$$\frac{d}{dt}\left(\frac{de}{t}\right)^t = \left(\frac{de}{t}\right)^t \ln\left(\frac{d}{t}\right) > \left(\frac{de}{t}\right)^t \ln(6) > 0, \tag{25}$$

so $\left(\frac{de}{t}\right)^t$ is increasing in t. By using Remark 1 and Remark 2 we have

$$\begin{aligned}
\log_2 \frac{|\mathcal{V}|}{N_t^{\max}} &\geq \log_2 \left(\frac{2^d}{2\binom{d}{t}} \right) \\
&= d - \log_2 \left(2\binom{d}{t} \right) \geq d - \log_2 \left(2 \left(\frac{de}{t} \right)^t \right) \\
&\overset{(25)}{\geq} d - \log_2 \left(2 \left(\frac{de}{d/6} \right)^{\frac{d}{6}} \right) \\
&= d - \frac{d}{6} \log_2 (6e) - 1 \\
&= d \log_2 \left(\frac{2}{(6e)^{\frac{1}{6}} \cdot 2^{\frac{1}{d}}} \right) \\
&> \frac{d}{6}.
\end{aligned}$$

It can be checked that $\log_2 \frac{|\mathcal{V}|}{N_t^{\max}} > \frac{d}{6}$ holds for $d < 8$, however for better readability we do not include it here.

Thus combining Lemma 6.1 and Lemma 6.2 using the Markov chain $V \to X \to Y \to \hat{\theta}$, we find that for $t = \frac{d}{6}$

$$\begin{aligned}
\sup_{P \in \mathcal{P}} \mathbb{E}(||\hat{\theta} - \theta(P)||_2^2) &\geq \delta^2(\lfloor t \rfloor + 1) \inf_{\hat{\nu} \in \mathcal{V}} \mathbb{P}(d_{\text{ham}}(\hat{\nu}, V) > t) \\
&\geq \delta^2(\lfloor t \rfloor + 1) \left(1 - \frac{I(V;X) + 1}{\log_2 \frac{|\mathcal{V}|}{N_t^{\max}}} \right) \\
&\geq \delta^2(\lfloor d/6 \rfloor + 1) \left(1 - \frac{I(V;X) + 1}{d/6} \right).
\end{aligned}$$

Since from Lemma 6.2 we have that $t < \frac{1}{2}(d-1)$, let us assume that $d \geq 3$. □

A.8 Proof of Lemma 6.4

Proof (Appendix A, Le Cam's Method of [6]). We consider the d-dimensional case for $d \leq 8$. We note that $\mathcal{V} := \{-1, 1\}^d$. Define $\nu_1 := (1, ..., 1)$ and $\nu_{-1} := (-1, ..., -1)$ with vector length d. By taking the supremum over a smaller (or equal for $d = 1$) set $\{P_{\nu_1}, P_{\nu_{-1}}\} \subset \{P_\nu\}_{\nu \in \mathcal{V}}$ we have

$$\begin{aligned}
\mathfrak{M}^{\text{ind}}(\theta, \mathcal{P}, B) &\overset{(5)}{\geq} \inf_{\Pi \in \mathcal{A}_{\text{ind}}(B, \{P_\nu\}_{\nu \in \mathcal{V}})} \sup_{P \in \{P_\nu\}_{\nu \in \mathcal{V}}} \mathbb{E}_P \left(||\hat{\theta}_\Pi(\mathbf{Y}) - \theta(P)||_2^2 \right) \\
&\geq \inf_{\Pi \in \mathcal{A}_{\text{ind}}(B, \{P_{\nu_1}, P_{\nu_{-1}}\})} \sup_{P \in \{P_{\nu_1}, P_{\nu_{-1}}\}} \mathbb{E}_P \left(||\hat{\theta}_\Pi(\mathbf{Y}) - \theta(P)||_2^2 \right) \\
&\overset{(*)}{\geq} \delta^2 \frac{1}{2} \left(1 - \sqrt{2I(V;\mathbf{Y})} \right) \\
&\geq \delta^2 \frac{d}{16} \left(1 - \sqrt{2I(V;\mathbf{Y})} \right).
\end{aligned}$$

where $(*)$ follow from Le Cam's Method of [6]. □

A.9 Proof of Proposition 6.1

Proof (Appendix A, Tensorisation of information [6]). Consider a protocol Π such that Y_i is constructed based only on $X^{(i)}$. Then

$$\begin{aligned} I(V;\mathbf{Y}) &\overset{(3)}{=} \sum_{i=1}^{m} I(V;Y_i|Y_1,Y_2,...,Y_{i-1}) \\ &\overset{(2)}{=} \sum_{i=1}^{m} H(Y_i|Y_1,...,Y_{i-1}) - H(Y_i|V,Y_1,...,Y_{i-1}) \\ &\leq \sum_{i=1}^{m} H(Y_i) - H(Y_i|V,Y_1,...,Y_{i-1}) \\ &= \sum_{i=1}^{m} H(Y_i) - H(Y_i|V) \overset{(1)}{=} \sum_{i=1}^{m} I(V;Y_i), \end{aligned}$$

where the inequality follows since conditioning reduces entropy, and in the last line we use that the Y_is are conditionally independent. □

A.10 Proof of Theorem 6.8

Proof (Theorem 2 of [6]). We want to prove the lower bound for $\mathfrak{M}^{\text{ind}}(\theta,\mathcal{N}_d,B)$ given in Theorem 6.8. To do this we will consider two cases, where the dimension $d \geq 9$ and $d < 9$. For $d \geq 9$ we will use Lemma 6.3 to give the lower bound

$$\begin{aligned} \mathfrak{M}^{\text{ind}}(\theta,\mathcal{N}_d,B) &\geq \sup_{P\in\{P_\nu\}_{\nu\in\mathcal{V}}} \mathbb{E}(||\hat{\theta}(\mathbf{Y}) - \theta(P)||_2^2) \\ &\geq \delta^2(\lfloor d/6 \rfloor + 1)\left(1 - \frac{I(V;\mathbf{Y})+1}{d/6}\right). \end{aligned} \tag{26}$$

For $d < 9$ we will use Lemma 6.4 to give the lower bound

$$\mathfrak{M}^{\text{ind}}(\theta,\mathcal{N}_d,B) \geq \delta^2 \frac{d}{16}\left(1 - \sqrt{2I(V;\mathbf{Y})}\right). \tag{27}$$

For both cases we need to find an upper bound for $I(V;\mathbf{Y})$. From Proposition 6.1 and Lemma 6.7 we have

$$\begin{aligned} \frac{2}{d}I(V;\mathbf{Y}) &\leq \frac{2}{d}\sum_{i=1}^{m} I(V;Y_i) \\ &\leq \frac{2}{d}\sum_{i=1}^{m} \frac{n\delta^2}{\sigma^2}b_i + dh\left(2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right)\right) + 2d\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right) \\ &= \sum_{i=1}^{m}\left[\frac{2n\delta^2}{d\sigma^2}b_i + 2h\left(2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right)\right) + 4\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right)\right], \end{aligned} \tag{28}$$

where $\mathbf{Y} = (Y_1,...,Y_m)$. We will upper bound all three terms in the summation on the right of Equation (28). Choose $a = 5\sigma\sqrt{\log_2(m)}$. For the first term in the summation

choose $\delta_1^2 \leq \frac{d\sigma^2}{20\sum_{i=1}^m b_i n}$. Then the first term is lower bounded by

$$\sum_{i=1}^m \frac{2n\delta_1^2}{d\sigma^2} b_i \leq \sum_{i=1}^m \frac{2b_i n}{20\sum_{j=1}^m b_j n} = \frac{1}{10}. \tag{29}$$

For the second and third term choose $\delta_2^2 \leq \frac{\sigma^2}{400\log_2(m)n}$. Note for $\log_2(m) \geq 1$ we have $\frac{\sqrt{n}}{\sqrt{\log_2(m)n}} \leq 1$, and

$$\begin{aligned}
(a - \sqrt{n}\delta_2)^2 &= \left(5\sigma\sqrt{\log_2(m)} - \sqrt{n}\frac{\sigma}{20\sqrt{\log_2(m)n}}\right)^2 \\
&= \left(5 - \frac{1}{20\log_2(m)}\right)^2 \sigma^2 \log_2(m) \\
&\geq \left(5 - \frac{1}{20}\right)^2 \sigma^2 \log_2(m) \\
&\geq 24\sigma^2 \log_2(m).
\end{aligned}$$

For all $-1 < x < 0$ we have $\frac{2x}{1+x} < \log_2(1+x)$. For $m \geq 2$ we have $-1 < -2m^{-12} < 0$, hence $\frac{-4m^{-12}}{1-2m^{-12}} < \log_2(1-2m^{-12})$. Note that $-\log_2(1-2m^{-12}) > 0$. Thus

$$\begin{aligned}
h(2m^{-12}) &= -\log_2(1-2m^{-12})(1-2m^{-12}) - 2m^{-12}\log_2(2m^{-12}) \\
&\leq \frac{4m^{-12}}{1-2m^{-12}}(1-2m^{-12}) - 2m^{-12}\log_2(2m^{-12}) \\
&= 4m^{-12} - 2m^{-12}\log_2(2m^{-12}) \\
&= 2m^{-12}(2 - \log_2(2m^{-12})).
\end{aligned}$$

Then for the upper bound on the second term on the right hand side of Equation (28) we have

$$\begin{aligned}
\sum_{i=1}^m 2h\left(2\exp\left(-\frac{(a-\sqrt{n}\delta)^2}{2\sigma^2}\right)\right) &\leq \sum_{i=1}^m 2h\left(2\exp\left(-\frac{24\sigma^2\log_2(m)}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^m 2h\left(2\exp\left(-12\log_2(m)\right)\right) \\
&\leq 4m^{-11}(2 - \log_2(2m^{-12})).
\end{aligned} \tag{30}$$

For the third term on the right hand side of Equation (28) we have

$$\sum_{i=1}^m 4\exp\left(-\frac{(a-\sqrt{n}\delta_2)^2}{2\sigma^2}\right) \leq 4m\exp\left(-12\log_2(m)\right) = 4m\cdot m^{-12} = 4m^{-11}. \tag{31}$$

Choose $\delta^2 := \min\left\{1, \frac{\sigma^2}{400\log_2(m)n}, \frac{d\sigma^2}{20\sum_{i=1}^m b_i n}\right\}$. Then the conditions of Lemma 6.7 are satisfied since $\delta^2 \leq \delta_2^2$ and

$$\frac{\sqrt{n}a\delta_2}{\sigma^2} \leq \frac{\sqrt{n}}{\sigma^2}5\sigma\sqrt{\log_2(m)}\frac{\sigma}{20\sqrt{\log_2(m)n}} \leq \frac{5\sqrt{n}}{20\sqrt{n}} \leq \frac{5}{20} < \frac{1.2564}{4}.$$

Let $g(m)$ be the sum of the lower bounds found in (29), (30) and (31) giving

$$\begin{aligned} g(m) &= \frac{1}{10} + 4m^{-11}(2 - \log_2(2m^{-12})) + 4m^{-11} \\ &= \frac{1}{10} + 4m^{-11}(3 - \log_2(2m^{-12})). \end{aligned}$$

Then from Equation (28) we have

$$I(V; \mathbf{Y}) \leq \sum_{i=1}^{m} I(V; Y_i) \leq g(m)\frac{d}{2}.$$

We note that $g(3) < 0.1005$ and that $g(m)$ is a decreasing function for $m \geq 3$, hence $g(m) < 0.1005$ for $m \geq 3$. Then $1 - \frac{6}{d}\left(g(m)\frac{d}{2} + 1\right) > 0$ holds for $m \geq 3$. Thus

$$I(V; \mathbf{Y}) < 0.1005 \cdot \frac{d}{2}. \tag{32}$$

For the case $d \geq 9$ we have that

$$1 - \frac{6}{d}\left(I(V; \mathbf{Y}) + 1\right) > 1 - \frac{6}{d}\left(\left(0.1005 \cdot \frac{d}{2}\right) + 1\right) > 0, \tag{33}$$

where middle term is increasing in d.

We have that $\sup_{P \in \mathcal{P}} \mathbb{E}_P(L_i) \leq B_i$ where L_i is the number of bits required to encode Y_i. Then by Shannon's coding theorem [1], we have $H(Y_i) \leq B_i$. Thus

$$\begin{aligned} b_i &= \min\left\{128\frac{a^2}{\sigma^2}H(Y_i), d\right\} \\ &= \min\left\{25 \cdot 128 H(Y_i)\log(m), d\right\} \leq \min\left\{25 \cdot 128 B_i \log_2(m), d\right\}. \end{aligned} \tag{34}$$

We note that by algebraic manipulation we get

$$\begin{aligned} \delta^2 &= \min\left\{1, \frac{\sigma^2}{400\log_2(m)n}, \frac{d\sigma^2}{20\sum_{i=1}^{m} b_i n}\right\} \\ &\geq \min\left\{1, \frac{\sigma^2}{400\log_2(m)n}, \frac{\sigma^2}{20\sum_{i=1}^{m}\min\left\{25 \cdot 128\log_2(m)\frac{B_i}{d}, 1\right\} n}\right\} \\ &= \frac{\sigma^2}{mn}\min\left\{\frac{mn}{\sigma^2}, \frac{mn}{400\log_2(m)n}, \frac{mn}{20\sum_{i=1}^{m}\min\left\{25 \cdot 128\log_2(m)\frac{B_i}{d}, 1\right\} n}\right\} \\ &\geq \frac{\sigma^2}{mn}\min\left\{\frac{mn}{\sigma^2}, \frac{m}{400\log_2(m)}, \frac{m}{20 \cdot 25 \cdot 128\sum_{i=1}^{m}\min\left\{\log_2(m)\frac{B_i}{d}, 1\right\}}\right\} \\ &\geq \frac{1}{20 \cdot 25 \cdot 128}\frac{\sigma^2}{mn}\min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m)\sum_{i=1}^{m}\min\left\{\frac{B_i}{d}, 1\right\}}\right\}. \end{aligned} \tag{35}$$

Then for $d \geq 9$ we have

$$\begin{aligned}
\mathfrak{M}^{\text{ind}}(\theta, \mathcal{N}_d, B) &\overset{(26)}{\geq} \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1 \right) \left(1 - \frac{6}{d} I(V; \mathbf{Y}) - \frac{6}{d} \right) \\
&\overset{(33)}{\geq} \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1 \right) \left(1 - 6 \left(0.1005 \cdot \frac{1}{2} \right) - \frac{6}{9} \right) \\
&\geq \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1 \right) 0.0235 \\
&\geq \delta^2 d \frac{0.0235}{6} \\
&\overset{(35)}{\geq} c_1 \frac{\sigma^2 d}{mn} \min \left\{ \frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^m \min\left\{ \frac{B_i}{d}, 1 \right\}} \right\},
\end{aligned}$$

where $c_1 = \frac{0.0235}{6} \cdot \frac{1}{20 \cdot 28 \cdot 128}$.

We now consider the case $d < 9$. We use from Equation (32) that $I(V; \mathbf{Y}) < 0.1005 \cdot \frac{d}{2}$.

We note that $\left(1 - \sqrt{0.1005 d}\right) > 0$ for $d < 9$ and it is decreasing in d. Then

$$\left(1 - \sqrt{0.1005 \cdot d}\right) > \left(1 - \sqrt{0.1005 \cdot 9}\right) > 0.048.$$

From Lemma 6.4,

$$\begin{aligned}
\mathfrak{M}^{\text{ind}}(\theta, \mathcal{N}_d, B) &\overset{(27)}{\geq} \delta^2 \frac{d}{16} \left(1 - \sqrt{2 I(V; \mathbf{Y})}\right) \\
&\overset{(32)}{\geq} \delta^2 \frac{d}{16} \left(1 - \sqrt{2 \cdot 0.1005 \cdot \frac{d}{2}}\right) \\
&= \delta^2 \frac{d}{16} \left(1 - \sqrt{0.1005 \cdot d}\right) \\
&\overset{(35)}{\geq} c_2 \frac{\sigma^2 d}{mn} \min \left\{ \frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^m \min\left\{ \frac{B_i}{d}, 1 \right\}} \right\},
\end{aligned} \tag{36}$$

where $c_2 = 0.048 \cdot \frac{1}{16} \cdot \frac{1}{20 \cdot 25 \cdot 128}$.

For general $d \in \mathbb{N}$ we find that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{N}_d, B) \geq c \frac{\sigma^2 d}{mn} \min \left\{ \frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^m \min\left\{ \frac{B_i}{d}, 1 \right\}} \right\}$$

where $c = c_2 < c_1$. We can write c as $c = 4.6875 \cdot 10^{-8}$. □

A.11 Proof of Lemma 6.6

Proof. Let $p(\mathbf{x})$ and $q(\mathbf{x})$ be normal probability density functions where $p(\mathbf{x})$ has mean μ and variance $\sigma^2 I_{d \times d}$, and $q(\mathbf{x})$ has mean m and variance $\sigma^2 I_{d \times d}$. With a little abuse of notation we let $p(x_i)$ be a normal probability density function with mean μ and variance σ^2, and $q(x_i)$ be a normal probability density function with mean m and variance σ^2. We

have that $\int p(x_i)\,dx_i = 1$, $\int (x_i-\mu)^2 p(x_i)\,dx_i = \sigma^2$ and $\int (x_i-\mu)p(x_i)\,dx_i = 0$. We have

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{\sqrt{|2\pi\sigma^2 I_{d\times d}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T(\sigma^2 I_{d\times d})^{-1}(\mathbf{x}-\mu)\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^{2d}}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^{2d}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{d}(x_i-\mu_i)^2\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^{2d}}} \prod_{i=1}^{d} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_i)^2\right).
\end{aligned}$$

Similarly

$$q(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^{2d}}} \prod_{i=1}^{d} \exp\left(-\frac{1}{2\sigma^2}(x_i-m_i)^2\right).$$

Then

$$\begin{aligned}
D(p||q) &= \int p(\mathbf{x}) \log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})}\, d\mathbf{x} \\
&= \int p(\mathbf{x}) \log_2 \left(\frac{\frac{1}{\sqrt{2\pi\sigma^{2d}}} \prod_{i=1}^{d} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_i)^2\right)}{\frac{1}{\sqrt{2\pi\sigma^{2d}}} \prod_{i=1}^{d} \exp\left(-\frac{1}{2\sigma^2}(x_i-m_i)^2\right)} \right) d\mathbf{x} \\
&= \int p(\mathbf{x}) \sum_{i=1}^{d} \log_2 \left(\frac{\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_i)^2\right)}{\exp\left(-\frac{1}{2\sigma^2}(x_i-m_i)^2\right)} \right) d\mathbf{x} \\
&= \int p(\mathbf{x}) \sum_{i=1}^{d} \left(-\frac{1}{2\sigma^2}(x_i-\mu_i)^2 + \frac{1}{2\sigma^2}(x_i-m_i)^2 \right) d\mathbf{x} \\
&= \sum_{i=1}^{d} \left(-\frac{1}{2\sigma^2} \int p(\mathbf{x})(x_i-\mu_i)^2\, d\mathbf{x} + \frac{1}{2\sigma^2} \int p(\mathbf{x})(x_i-m_i)^2\, d\mathbf{x} \right) \\
&= \sum_{i=1}^{d} \left(-\frac{\sigma^2}{2\sigma^2} + \frac{1}{2\sigma^2} \int p(\mathbf{x})(x_i-\mu_i+\mu_i-m_i)^2\, d\mathbf{x} \right) \\
&= \sum_{i=1}^{d} \Big(-\frac{1}{2} + \frac{1}{2\sigma^2} \Big(\int p(x_i)(x_i-\mu_i)^2\, dx_i + \int p(x_i)(\mu_i-m_i)^2\, dx_i \\
&\qquad\qquad - 2(\mu_i-m_i) \int p(x_i)(x_i-\mu_i)\, dx_i \Big) \Big) \\
&= \frac{1}{2} \sum_{i=1}^{d} \left(-1 + \frac{1}{\sigma^2} \left(\sigma^2 + (\mu_i-m_i)^2 \right) \right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{d} (\mu_i-m_i)^2.
\end{aligned}$$

□

A.12 Proof of Lemma 7.1

Proof. Let $p(x)$ and $q(x)$ be Laplace probability density functions with $p(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$ and $q(x) = \frac{1}{2b}\exp\left(-\frac{|x-m|}{b}\right)$, where $\mu, m \in \mathbb{R}$ are the locations and $b \in \mathbb{R}_{>0}$ is the scale.

Define $t := x - \mu$, then $dx = dt$. We have

$$\begin{aligned}
\int p(x)|x-\mu|\,dx &= \frac{1}{2b}\int |x-\mu|\exp\left(-\frac{|x-\mu|}{b}\right)\,dx \\
&= \frac{1}{2b}\int |t|\exp\left(-\frac{|t|}{b}\right)\,dt \\
&= \frac{1}{2b}\left(-\int_{-\infty}^{0} t\exp\left(\frac{t}{b}\right)\,dt + \int_{0}^{\infty} t\exp\left(-\frac{t}{b}\right)\,dt\right) \\
&= \frac{1}{2b}\left(\int_{0}^{\infty} t\exp\left(-\frac{t}{b}\right)\,dt + \int_{0}^{\infty} t\exp\left(-\frac{t}{b}\right)\,dt\right) \\
&= \frac{1}{b}\int_{0}^{\infty} t\exp\left(-\frac{t}{b}\right)\,dt \\
&= \frac{1}{b}b^2 = b.
\end{aligned}$$

Define $\mu - m := k$. We consider two separate cases, firstly where $k \geq 0$ and secondly where $k < 0$. Suppose $k \geq 0$. By splitting the integral into intervals to remove the absolute signs and applying integration by parts we get

$$\begin{aligned}
\int p(x)|x-m|dx &= \frac{1}{2b}\int |x-m|\exp\left(-\frac{|x-\mu|}{b}\right)dx \\
&= \frac{1}{2b}\int |t+k|\exp\left(-\frac{|t|}{b}\right)\,dt \\
&= \frac{1}{2b}\left(\int_{-\infty}^{0} |t+k|\exp\left(\frac{t}{b}\right)\,dt + \int_{0}^{\infty} |t+k|\exp\left(\frac{-t}{b}\right)\,dt\right) \\
&= \frac{1}{2b}\Bigg(-\int_{-\infty}^{-k}(t+k)\exp\left(\frac{t}{b}\right)\,dt + \int_{-k}^{0}(t+k)\exp\left(\frac{t}{b}\right)\,dt \\
&\qquad + \int_{0}^{\infty}(t+k)\exp\left(-\frac{t}{b}\right)\,dt\Bigg) \\
&= \frac{1}{2b}\left(\left(b^2e^{-\frac{k}{b}}\right) + \left(b^2(e^{-\frac{k}{b}}-1)+kb\right) + \left(b^2+bk\right)\right) \\
&= \frac{1}{2b}\left(2b^2e^{-\frac{k}{b}} + 2kb\right) = be^{-\frac{k}{b}} + k = \mu - m + be^{\frac{m-\mu}{b}}.
\end{aligned}$$

Then

$$\begin{aligned}
D(p||q) &= \int p(x) \log_2 \frac{p(x)}{q(x)}\, dx \\
&= \int p(x) \log_2 \left(\frac{\frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)}{\frac{1}{2b} \exp\left(-\frac{|x-m|}{b}\right)} \right) \\
&= \int p(x) \left(-\frac{|x-\mu|}{b} + \frac{|x-m|}{b} \right) dx \\
&= -\frac{1}{b} \int p(x)|x-\mu|\, dx + \frac{1}{b} \int p(x)|x-m|\, dx \\
&= -\frac{1}{b}(b) + \frac{1}{b}\left(\mu - m + be^{\frac{m-\mu}{b}}\right) \\
&= e^{\frac{m-\mu}{b}} + \frac{\mu - m}{b} - 1.
\end{aligned}$$

Suppose that $k > 0$, then using the same method as before we get

$$\begin{aligned}
\int p(x)|x-m|dx &= \frac{1}{2b} \int |x-m| e^{-\frac{|x-\mu|}{b}} dx \\
&= \frac{1}{2b} \int |t+\mu-m| e^{-\frac{|t|}{b}}\, dt \\
&= \frac{1}{2b} \left(\int_{-\infty}^{0} |t-k| e^{\frac{t}{b}}\, dt + \int_{0}^{\infty} |t-k| e^{-\frac{t}{b}}\, dt \right) \\
&= \frac{1}{2b} \left(-\int_{-\infty}^{0} (t-k) e^{\frac{t}{b}}\, dt - \int_{0}^{k} (t-k) e^{-\frac{t}{b}}\, dt + \int_{k}^{\infty} (t-k) e^{-\frac{t}{b}}\, dt \right) \\
&= \frac{1}{2b} \left(\left(b^2 + bk\right) + \left(b^2 e^{-\frac{k}{b}} - b^2 + bk\right) + \left(b^2 e^{-\frac{k}{b}}\right) \right) \\
&= k + be^{-\frac{k}{b}} \\
&= m - \mu + be^{\frac{\mu-m}{b}}.
\end{aligned}$$

Then similarly

$$\begin{aligned}
D(p||q) &= -\frac{1}{b} \int p(x)|x-\mu|\, dx + \frac{1}{b} \int p(x)|x-m|\, dx \\
&= -\frac{1}{b}(b) + \frac{1}{b}\left(m - \mu + be^{\frac{\mu-m}{b}}\right) \\
&= e^{\frac{\mu-m}{b}} + \frac{m-\mu}{b} - 1.
\end{aligned}$$

It follows for all k that

$$D(p||q) = e^{-\frac{|\mu-m|}{b}} + \frac{|\mu-m|}{b} - 1.$$

□

A.13 Proof of Lemma 7.3

Proof. Inequality (7) is the consequence of two intermediate upper bounds, which we prove separately:

$$I(V;Y_i) \leq \frac{\delta^2 dn}{b^2} \tag{37}$$

$$I(V;Y_i) \leq 128\frac{n\delta^2 a^2}{b^2} I(X^{(i)};Y_i) + dh\left(2\exp\left(-\frac{a^2}{32}\right)\right) + 2d\exp\left(-\frac{a^2}{32}\right). \tag{38}$$

We note that $V \to X^{(i)} \to Y_i$ forms a Markov chain. We have that for arbitrary $k \in \{1,...,n\}$

$$I(V;Y_i) \leq I(V;X^{(i)}) \leq \sum_{j=1}^{n} I(V;X^{(i,j)}) = nI(V;X^{(i,k)})$$

where Theorem 5.2 gives the first inequality, Proposition 6.1 the second, and the final equality is since the $X^{(i,j)}$'s are identical for $j \in \{1,...,n\}$.

We note that by the Taylor expansion

$$e^{-\frac{1}{b}||\delta\nu-\delta\nu'||_1} = 1 - \frac{1}{b}||\delta\nu-\delta\nu'||_1 + \frac{1}{2b^2}||\delta\nu-\delta\nu'||_1^2 - \frac{1}{6b^3}||\delta\nu-\delta\nu'||_1^3 e^{\zeta},$$

where $\zeta \in (-\frac{1}{b}||\delta\nu-\delta\nu'||_1, 0)$. Then $\zeta < 0$ so $0 < e^{\zeta} < 1$. Thus

$$-\frac{1}{6b^3}||\delta\nu-\delta\nu'||_1^3 e^{\zeta} < 0.$$

Let P_ν denote the conditional distribution of $X^{(i,j)}$ given $V = \nu$. By the above we have

$$\begin{aligned} D(P_\nu||P_{\nu'}) &= e^{-\frac{1}{b}||\delta\nu-\delta\nu'||_1} + \frac{1}{b}||\delta\nu-\delta\nu'||_1 - 1 \\ &\leq 1 - \frac{1}{b}||\delta\nu-\delta\nu'||_1 + \frac{1}{2b^2}||\delta\nu-\delta\nu'||_1^2 + \frac{1}{b}||\delta\nu-\delta\nu'||_1 - 1 \\ &= \frac{\delta^2}{2b^2}||\nu-\nu'||_1^2. \end{aligned}$$

For $\nu \in \mathcal{V}$ we have

$$\sum_{\nu' \in \mathcal{V}} ||\nu-\nu'||_1^2 = \frac{1}{2}\left(\sum_{\nu' \in \mathcal{V}} ||\nu-\nu'||_1^2 + \sum_{\nu' \in \mathcal{V}} ||\nu+\nu'||_1^2\right) = \frac{1}{2}\sum_{\nu' \in \mathcal{V}} 4d = 2d|\mathcal{V}|.$$

Thus

$$\sum_{\nu,\nu' \in \mathcal{V}} ||\nu-\nu'||_1^2 = 2d|\mathcal{V}|^2.$$

Remark 4. *Consider two random variables X and Y with joint probability density $p(x,y)$. Let $p(x|y) = \frac{p(x,y)}{p(y)}$. Then*

$$\begin{aligned} I(X;Y) &= \sum_y p(y) \sum_x p(x|y) \log_2 \frac{p(x|y)}{p(x)} \\ &= \sum_y p(y)\ D(p(x|y)\|p(x)) \\ &= \mathbb{E}_Y\left(D(p(x|y)\|p(x))\right). \end{aligned}$$

Let P denote the distribution of $X^{(i,j)}$. From Remark 4 we have

$$\begin{aligned} I(V;X^{(i,j)}) &= \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} D(P_\nu\|P) \\ &= \frac{1}{|\mathcal{V}|}\sum_{\nu\in\mathcal{V}} D\left(P_\nu\Big\|\frac{1}{|\mathcal{V}|}\sum_{\nu'\in\mathcal{V}} P_{\nu'}\right) \\ &\le \frac{1}{|\mathcal{V}|^2}\sum_{\nu,\nu'\in\mathcal{V}} D\left(P_\nu\|P_{\nu'}\right). \end{aligned} \tag{39}$$

Then using the Kullback-Leiber divergence for the multivariate Laplace given in Lemma 7.2, we have

$$\begin{aligned} I(V;X^{(i,j)}) &\overset{(39)}{\le} \frac{1}{|\mathcal{V}|^2}\sum_{\nu,\nu'\in\mathcal{V}} D(P_\nu||P_{\nu'}) \\ &\le \frac{1}{|\mathcal{V}|^2}\sum_{\nu,\nu'\in\mathcal{V}} \frac{\delta^2}{2b^2}||\nu-\nu'||_1^2 \\ &\le \frac{\delta^2 d}{b^2}. \end{aligned}$$

This establishes the inequality in Equation (37). To prove the inequality in Equation (38), we apply Lemma 6.5.

Take a ratio of the densities of two Laplace distributions with n independent samples, one with mean δ and the other with mean $-\delta$, both with location parameter b. We have

$$\frac{\exp\left(-\frac{1}{b}\sum_{l=1}^n |x_l-\delta|\right)}{\exp\left(-\frac{1}{b}\sum_{l=1}^n |x_l+\delta|\right)} = \exp\left(\frac{1}{b}\sum_{l=1}^n |x_l+\delta|-|x_l-\delta|\right) \le \exp\left(\frac{\delta\sqrt{n}a}{b}\right),$$

whenever $\sum_{l=1}^n |x_l+\delta|-|x_l-\delta| \le \delta\sqrt{n}a$ for some $a\in\mathbb{R}$. Taking the sets

$$S_0 := \left\{x\in\mathbb{R}^n : \sum_{l=1}^n |x_l+\delta|-|x_l-\delta| \le \delta\sqrt{n}a\right\},$$

satisfies the conditions of Lemma 6.5 with $\alpha = \delta\sqrt{n}a/b$. When $\alpha \le 1.2564$ we have $\exp(\alpha)-1 \le 2\alpha$. Then $\exp(4\alpha)-1 = \exp(4\delta\sqrt{n}a/b)-1 \le 8\delta\sqrt{n}a/b$. Hence $2(e^{4\alpha}-1)^2 \le 2(8\delta\sqrt{n}a/b)^2 = 128n\delta^2a^2/b^2$. Let us take $E_j = \mathbb{1}_{X_j^{(i)}\in S_0}$ for some i. Then from Lemma 6.5 we have

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

We have

$$\begin{aligned} \sum_{l=1}^n |x_l+\delta|-|x_l-\delta| &= \sum_{l=1}^n 2\delta\cdot\mathbb{1}_{x_l>\delta} - 2\delta\cdot\mathbb{1}_{x_l<-\delta} + 2x_l\cdot\mathbb{1}_{-\delta<x_l<\delta} \\ &= \delta\sum_{l=1}^n 2\cdot\mathbb{1}_{x_l>\delta} - 2\cdot\mathbb{1}_{x_l<-\delta} + \frac{2x_l}{\delta}\cdot\mathbb{1}_{-\delta<x_l<\delta} \\ &= \delta\sum_{l=1}^n (C_l+D_l), \end{aligned}$$

where $C_l := 2 \cdot \mathbb{1}_{x_l > \delta} - 2 \cdot \mathbb{1}_{x_l < -\delta}$ and $D_l := \frac{2x_l}{\delta} \cdot \mathbb{1}_{-\delta < x_l < \delta}$. Note that $\mathbb{E}(C_l) = 0$ and $\mathbb{E}(D_l) = 0$ from symmetry since x_l's have mean 0. The variance of the random variables C_l and D_l can be bounded from above as

$$\begin{aligned}\mathbb{V}\mathrm{ar}(C_l) &= \mathbb{E}\left((C_l - \mathbb{E}(C_l))^2\right) = \mathbb{E}\left(C_l^2\right) = 4\mathbb{E}\left(\left(\mathbb{1}_{x_l > \delta} - \mathbb{1}_{x_l < -\delta}\right)^2\right) \\ &= 4\mathbb{E}\left(\mathbb{1}_{x_l > \delta} + \mathbb{1}_{x_l < -\delta}\right) = 4\left(\mathbb{P}\left(\mathbb{1}_{x_l > \delta}\right) + \mathbb{P}\left(\mathbb{1}_{x_l < -\delta}\right)\right) \\ &= 4c_l,\end{aligned}$$

for some constant $0 < c_l < 1$ close to 1. Furthermore

$$\mathbb{V}\mathrm{ar}(D_l) = \mathbb{E}\left((D_l - \mathbb{E}(D_l))^2\right) = \mathbb{E}\left(D_l^2\right) = \frac{4}{\delta^2}\mathbb{E}\left(\left(x_l \cdot \mathbb{1}_{-\delta < x_l < \delta}\right)^2\right) \leq 4.$$

Since the C_l's and D_l's are i.i.d. random variables, from the central limit theorem

$$\frac{\sqrt{n}}{n}\sum_{l=1}^{n} C_l \xrightarrow{d} N(0, \sigma_1^2), \quad \frac{\sqrt{n}}{n}\sum_{l=1}^{n} D_l \xrightarrow{d} N(0, \sigma_2^2),$$

where $\sigma_1^2, \sigma_2^2 \leq 4$. Let $Z_1, Z_2 \sim N(0,1)$. Then

$$\begin{aligned}\mathbb{P}(E_j = 0) = \mathbb{P}(X_j^{(i)} \notin B_i) &= \mathbb{P}\left(\delta\sum_{l=1}^{n}(C_l + D_l) > \delta\sqrt{n}a\right) \\ &= \mathbb{P}\left(\frac{\sqrt{n}}{n}\sum_{l=1}^{n}(C_l + D_l) > a\right) \\ &\approx \mathbb{P}(\sigma_1 Z_1 + \sigma_2 Z_2 > a) \\ &\leq \mathbb{P}\left(Z_1 > \frac{a}{2\sigma_1}\right) + \mathbb{P}\left(Z_2 > \frac{a}{2\sigma^2}\right) \\ &\leq \exp\left(-\frac{1}{2}\frac{a^2}{4\sigma_1^2}\right) + \exp\left(-\frac{1}{2}\frac{a^2}{4\sigma_2^2}\right) \\ &= 2\exp\left(-\frac{a^2}{32}\right).\end{aligned}$$

Then

$$\begin{aligned}I(V; Y_i) &\leq 128\frac{n\delta^2 a^2}{b^2} I(X^{(i)}; Y_i) + \sum_{j=1}^{d} H(E_j) + \sum_{j=1}^{d}\mathbb{P}(E_j = 0) \\ &\leq 128\frac{n\delta^2 a^2}{b^2} I(X^{(i)}; Y_i) + dh\left(2\exp\left(-\frac{a^2}{32}\right)\right) + 2d\exp\left(-\frac{a^2}{32}\right).\end{aligned}$$

Note that $I(X^{(i)}; Y_i) = H(Y_i) - H(Y_i|X^{(i)})$,and therefore $I(X^{(i)}; Y_i) \leq H(Y_i)$. From Equation (37) we get

$$I(V; Y_i) \leq \frac{n\delta^2}{b^2}\min\left\{128a^2 H(Y_i), d\right\} + dh\left(2\exp\left(-\frac{a^2}{32}\right)\right) + 2d\exp\left(-\frac{a^2}{32}\right).$$

□

A.14 Proof of Theorem 7.4

Proof. We want to prove the lower bound for $\mathfrak{M}^{\mathrm{ind}}(\theta, \mathcal{L}_d, B)$ given in Theorem 7.4. This proof follows the same structure as the proof for Theorem 6.8 given in Section A.10 where

we split the proof into 2 cases, for $d \geq 9$ and for $d < 9$. However, now instead of applying Lemma 6.7 for the normal means model, we apply Lemma 7.3 for the Laplace means model.

For both cases we need to find an upper bound for $I(V; \mathbf{Y})$. From Proposition 6.1 and Lemma 7.3 we have

$$\begin{aligned}\frac{2}{d}\sum_{i=1}^{m} I(V;\mathbf{Y}) &\leq \frac{2}{d}\sum_{i=1}^{m} I(V;Y_i) \\ &\leq \sum_{i=1}^{m}\left[\frac{dn\delta^2}{2b^2}b_i + 2h\left(2\exp\left(-\frac{a^2}{32}\right)\right) + 4\exp\left(-\frac{a^2}{32}\right)\right], \end{aligned} \tag{40}$$

where $b_i = \min\left\{128a^2H(Y_i), d\right\}$. We will upper bound all three terms in the summation on the right of Equation (40). Choose $a = \sqrt{320 \cdot \log_2(m)}$. For the first term choose $\delta_1^2 \leq \frac{db^2}{20\sum_{i=1}^{m} b_i n}$. Then we have

$$\sum_{i=1}^{m} \frac{2n\delta_1^2}{db^2}b_i \leq \sum_{i=1}^{m} \frac{2b_i n}{20\sum_{j=1}^{m} b_j n_j} = \frac{1}{10}. \tag{41}$$

Similarly to Section A.10 for $m \geq 2$ we have

$$h(2m^{-10}) \leq 2m^{-10}(2 - \log_2(2m^{-10})). \tag{42}$$

Then for the upper bound on the second term on the right hand side of Equation (40) we have

$$\begin{aligned}\sum_{i=1}^{m} 2h\left(2\exp\left(-\frac{a^2}{32}\right)\right) &= 2mh\left(2\exp\left(\frac{-320\log_2(m)}{32}\right)\right) \\ &= 2mh\left(2m^{-10}\right) \\ &\leq 4m^{-9}(2 - \log_2(2m^{-10})). \end{aligned} \tag{43}$$

For the third term on the right hand side of Equation (40) we have

$$\sum_{i=1}^{m} 4\exp\left(-\frac{a^2}{32}\right) = \sum_{i=1}^{m} 4\exp\left(-\frac{320\log_2(m)}{32}\right) = \sum_{i=1}^{m} 4m^{-10} = 4m^{-9}. \tag{44}$$

Choose $\delta_2^2 := \frac{1}{3245\log_2(m)n}$ and $\delta^2 := \min\left\{1, \delta_1^2, \delta_2^2\right\}$. Then the conditions of Lemma 7.3 are satisfied since $\delta^2 \leq \delta_2^2$ and

$$\frac{\sqrt{n}a\delta_2}{b} \leq \frac{\sqrt{n}}{b}\sqrt{320 \cdot \log_2(m)}\frac{1}{\sqrt{3245 \cdot \log_2(m)n}} \leq \sqrt{\frac{320}{3245}} < \frac{1.2564}{4}.$$

Let $g(m)$ be the sum of the lower bounds found in (41), (43) and (44) giving

$$\begin{aligned} g(m) &= \frac{1}{10} + 4m^{-9}(2 - \log_2(2m^{-10})) + 4m^{-9} \\ &= \frac{1}{10} + 4m^{-9}(3 - \log_2(2m^{-10})). \end{aligned}$$

Then from Equation (40) we have

$$I(V;\mathbf{Y}) \leq \sum_{i=1}^{m} I(V;Y_i) \leq g(m)\frac{d}{2}.$$

Similarly to Equation (32), we have $g(m) < 0.104$ for $m \geq 3$. Thus

$$I(V; \mathbf{Y}) < 0.104 \cdot \frac{d}{2}. \tag{45}$$

Similarly to Equation (34) we have

$$b_i = \min\left\{128a^2 H(Y_i), d\right\} \leq \min\left\{128 \cdot 320 \cdot \log_2(m) B_i, d\right\}.$$

From Lemma 6.3 we have

$$\sup_{P \in \{P_\nu\}_{\nu \in \mathcal{V}}} \mathbb{E}\left[||\hat{\theta}(Y) - \theta(P)||_2^2\right] \geq \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1\right)\left(1 - \frac{6}{d} I(V; \mathbf{Y}) - \frac{6}{d}\right). \tag{46}$$

Similar to in Equation (35) we have

$$\begin{aligned}
\delta^2 &= \min\left\{1, \frac{1}{3245 \log_2(m) n}, \frac{db^2}{20 \sum_{i=1}^m b_i n}\right\} \\
&\geq \frac{1}{20 \cdot 128 \cdot 320} \frac{\sigma^2}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^m \min\left\{\frac{B_i}{d}, 1\right\}}\right\}.
\end{aligned} \tag{47}$$

Then for $d \geq 9$

$$\begin{aligned}
\mathfrak{M}^{\text{ind}}(\theta, \mathcal{L}_d, B) &\overset{(26)}{\geq} \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1\right)\left(1 - \frac{6}{d} I(V; \mathbf{Y}) - \frac{6}{d}\right) \\
&\overset{(45)}{\geq} \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1\right)\left(1 - 6\left(0.104 \cdot \frac{1}{2}\right) - \frac{6}{9}\right) \\
&\geq \delta^2 \left(\left\lfloor \frac{d}{6} \right\rfloor + 1\right) 0.021 \\
&\geq \delta^2 d \frac{0.021}{6} \\
&\overset{(47)}{\geq} c_1 \frac{\sigma^2 d}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^m \min\left\{\frac{B_i}{d}, 1\right\}}\right\},
\end{aligned}$$

where $c_1 = \frac{0.021}{6} \cdot \frac{1}{20 \cdot 128 \cdot 320}$.

We now consider the case $d < 9$. We use from above that $I(V; \mathbf{Y}) < 0.104 \cdot \frac{d}{2}$.

We note that $\left(1 - \sqrt{0.104d}\right) > 0$ for $d < 9$ and it is decreasing in d. Then

$$\left(1 - \sqrt{0.104 \cdot d}\right) > \left(1 - \sqrt{0.104 \cdot 9}\right) > 0.032.$$

From Lemma 6.4,

$$\begin{aligned}
\mathfrak{M}^{\text{ind}}(\theta, \mathcal{L}_d, B) &\overset{(27)}{\geq} \delta^2 \frac{d}{16}\left(1 - \sqrt{2I(V; \mathbf{Y})}\right) \\
&\overset{(45)}{\geq} \delta^2 \frac{d}{16}\left(1 - \sqrt{2 \cdot 0.104 \cdot \frac{d}{2}}\right) \\
&= \delta^2 \frac{d}{16}\left(1 - \sqrt{0.104 \cdot d}\right) \\
&\overset{(47)}{\geq} c_2 \frac{\sigma^2 d}{mn} \min\left\{\frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^m \min\left\{\frac{B_i}{d}, 1\right\}}\right\},
\end{aligned}$$

where $c_2 = 0.032 \cdot \frac{1}{16} \cdot \frac{1}{20 \cdot 128 \cdot 320}$.

For general $d \in \mathbb{N}$ we find that

$$\mathfrak{M}^{\text{ind}}(\theta, \mathcal{L}_d, B) \geq c \frac{\sigma^2 d}{mn} \min \left\{ \frac{mn}{\sigma^2}, \frac{m}{\log_2(m)}, \frac{m}{\log_2(m) \sum_{i=1}^{m} \min\left\{\frac{B_i}{d}, 1\right\}} \right\}$$

where $c := 2.44 \cdot 10^{-9} < c_2 < c_1$, where c is equal to c_2 rounded down to 3 significant figures. □

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2005, pp. 1–748. ISBN: 9780471241959. DOI: 10.1002/047174882X. arXiv: ISBN0-471-06259-6.

[2] John C Duchi and Martin J Wainwright. "Distance-based and continuum Fano inequalities with applications to statistical estimation". In: *ArXiv preprint* (2013), p. 16. arXiv: 1311.2669. URL: http://arxiv.org/abs/1311.2669.

[3] *Folding@home*. 2017. URL: http://folding.stanford.edu/ (visited on 06/06/2017).

[4] *Great Internet Mersenne Prime Search*. 2016. URL: https://www.mersenne.org/ (visited on).

[5] *Pooled mining*. 2017. URL: https://en.bitcoin.it/wiki/Pooled%7B%5C_%7Dmining (visited on).

[6] Yuchen Zhang et al. "Information-theoretic lower bounds for distributed statistical estimation with communication constraints". In: *Advances in Neural Information Processing Systems 26*. Ed. by C J C Burges et al. Curran Associates, Inc., 2013, pp. 2328–2336.