Dartmouth College Computer Science Technical Report TR2015-777

COMMUNICATION COMPLEXITY OF DISTRIBUTED STATISTICAL ALGORITHMS

By James Brofos

Dartmouth College

This paper constructs bounds on the minimax risk under loss functions when statistical estimation is performed in a distributed environment and with communication constraints. We treat this problem using techniques from information theory and communication complexity. In many cases our bounds rely crucially on metric entropy conditions and the classical reduction from estimation to testing. A number of examples exhibit how bounds on the minimax risk play out in practice. We also study distributed statistical estimation problems in the context of PAC-learnability and derive explicit algorithms for solving classical problems. We study the communication complexity of these algorithms.

1. Introduction. This paper is concerned with the theory of the minimax risk and PAC-learning and their connections with distributed computing. We introduce definitions of parallel protocols emerging from communication complexity and then describe a natural extension of the minimax risk for distributed computing following [4]. We also discuss PAC-learnability in distributed environments for the problems of half-space and disjunction learning. The goal of this work is to begin to carefully characterize the amount of communication that is necessary in order to estimate parameters of statistical models.

With the explosion of big data in recent years, it has become increasingly important for researchers to investigate and develop algorithmic approaches to handling massive data sets. Since these data sets, and indeed the algorithms used for interpreting them, may sometimes be too memory-intensive for a single computer, it makes sense to consider distributing the task of learning from data across multiple computers. This work focuses on understanding the amount of communication that is necessary for solving particular statistical estimation problems.

Until recently, statistics dealt primarily with the case where n observations are drawn i.i.d. from a probability distribution $f(\cdot;\theta)$ parametrized by θ. The task at hand is to compute an estimator of θ, commonly denoted $\hat{\theta}$, from the available data. A natural question that arises from statistical estimation theory is whether or it is possible to quantify, in some sense, the desirability of a particular estimator. This question has given rise to the

theory of minimax estimators [13]. In such instances where the data and, in fact, parameters of the algorithm, are distributed across multiple computers, minimax theory of estimation is remarkably lacking in substance. The only prior work, as far as the author knows, in this area can be found in [4].

In this work we expand on the prior literature by carefully analyzing performance of estimators under the ℓ_1 risk measure and develop an explicit information-theoretic inequality for this purpose. We also introduce a helpful relationship between Lipschitz parametrizations and show how this can be informative in deriving distributed measures of complexity for statistical algorithms. We also expand on the prior literature by analyzing PAC-learnable quantities following [2], but focus on developing explicit algorithms with complexities measured by bits. This includes the development and analysis of an algorithm for PAC-learning disjunctive normal forms and half-spaces.

2. Notation. Given a parameter space $\Theta \subset \mathbb{R}^d$ with $|\Theta| \geq 2$ (if $|\Theta| < 2$ then there is nothing to estimate), a collection of K points (conveniently thought of as parameters) in Θ, denoted $\{\theta_1, \ldots, \theta_K\}$, is called δ-separated if for $i \neq j$ the ℓ_1-distance of θ_i and θ_j is lower bounded by δ. A maximal δ-packing has size,

$$\max\left\{K \in \mathbb{N} : \{\theta_1, \ldots, \theta_K\} \subset \Theta \text{ is } \delta\text{-separated}\right\}. \tag{2.1}$$

The notion of a maximal δ-packing will be crucial to the information-theoretic results that are formulated later in the work.

Furthermore, we say that every $\theta \in \Theta$ induces a probability distribution from which samples are drawn. In particular, we write that $f_\theta = f(\cdot; \theta)$ is a probability distribution parametrized by θ. We denote $\mathbf{F} = \{f_\theta : \theta \in \Theta\}$ to be the class of probability distributions parametrized by points of Θ. We use the notation $[N]$ to refer to the set of natural numbers up to N, namely $\{1, 2, \ldots, N\}$.

For some of our analyses it will be important for us to define a notion of *quantization*. This refers to the process of approximating a real number by a finite bit-string. Indeed, if x is a real number in the range $[-a, +a]$, then one can construct a quantization of x, denoted $\tilde{x}$, using $\log_2 m$-bits for some fixed $m \in \mathbb{N}$. Each configuration of the $\log_2 m$ bits (of which there are m) refers to equally-spaced points in the range of $[-a, +a]$. Hence it can be readily verified that $|x - \tilde{x}| \leq \frac{2a}{m}$ so that the quantization becomes more accurate as more bits are used in the representation.

2.1. *Protocols.* Let Γ, a natural number, denote the number of computers and suppose that each is provided a distinct data set $\mathbf{X}_i$ for $i \in [\Gamma]$, where we

assume that all of the $\mathbf{X}_i \in \mathbb{R}^{n \times p}$. Therefore, the total number of data points present in the statistical estimation procedure is $n \times \Gamma$. Given an estimator $\hat{\theta}$ of θ, this setup seeks to construct $\hat{\theta}$ via local operations on each machine and a limited amount of communication within the system. Thereby, the protocol attempts to recover the original $\theta \in \Theta$ on the basis of data.

We formally define this framework in the language of communication complexity [9]. We focus on multi-computer protocols Π such that the message broadcast at each round of communication is a measurable function of the available data $\mathbf{X}_i$ and (conceivably) of prior broadcasts. Let M denote the set of all messages sent in all rounds. We say that $\hat{\theta}$ is mapping of M into the parametrization space Θ.

If Π permits m rounds of communication, then denote by $M_{i,j}$ the message sent by the i^{th} computer at the j^{th} iteration. Then the quantity,

$$C = \sum_{i=1}^{\Gamma} \sum_{j=1}^{m} \text{Length}\left(M_{i,j}\right), \tag{2.2}$$

is the total communication cost of the protocol on the input. The function $\text{Length}\left(\cdot\right)$ gives the length, in bits, of the shortest encoding of the message argument.

Throughout our analysis we will impose restrictions on the size of C. In particular, if Π consists of only a single round, then we impose an upper bound on the message length for the i^{th} machine $\text{Length}\left(M_{i,1}\right) \leq B_i$. In each round, the computers write their message on a "blackboard," which the other computers may read from at no additional cost.

Remark 2.1. We assume that computations performed by computers locally consume no cost and that, additionally, the computation of $\hat{\theta}\left(M\right)$ consumes no cost beyond that inherent in broadcasting the set of messages.

In this work we focus primarily on one-round protocols. Although the preceding discussion attempts to offer an intuitive understanding of a protocol, some may find it useful to have a rigorous definition of a protocol. For the purposes of distributed statistical estimation, the following definition will suffice (similar to the more general definition provided in [9]).

Definition 2.1 (Protocols). For every $i \in [\Gamma]$, let $\mathbf{X}_i$ be the data on the i^{th} machine. Let $\mathcal{X}$ be the set of all possible datasets consisting of n points drawn from distributions in $\mathbf{F}$. A protocol Π with with domain $\mathcal{X} \times \ldots \times \mathcal{X}$ (Γ-times) and range Θ is a binary tree where internal nodes v are labeled

by functions $f_v^{(i)} : \mathcal{X} \to \{0,1\}$ for a single $i \in [\Gamma]$ and leaf nodes are labeled by points of Θ.

The behavior of the protocol Π on input $(\mathbf{X}_1, \ldots, \mathbf{X}_\Gamma)$ is given by walking on the tree until a leaf node is encountered. At each internal node v, Π moves left if $f_v^{(i)}(\mathbf{X}_i) = 0$ and otherwise moves right. The cost of the protocol Π is the height of the tree.

2.2. *Minimax Risk Theory.* The theory of minimax risk plays a central role in the analysis of statistical estimators [13]. Minimax risk seeks to quantify the worst-case performance (with respect to the underlying distribution) of the most effective estimator. In this context, "most effective" refers to the expected value of a statistical loss function. A common loss function is the squared-error, though in this paper we also analyze the absolute loss.

We quantitatively capture the quality of an estimator $\hat{\theta}$ by its expected absolute deviation from truth in the ℓ_1 metric,

$$\mathcal{R}\left(\hat{\theta}, \theta\right) = \mathbb{E}\left[\mathcal{D}\left(\hat{\theta}, \theta\right)\right], \tag{2.3}$$

where $\mathcal{D}(\cdot,\cdot)$ is a metric on Θ. From here, we define the minimax risk for protocols as,

$$\mathfrak{M}(\Theta, B) = \inf_{\Pi} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathcal{R}\left(\hat{\theta}, \theta\right). \tag{2.4}$$

This can be intuitively thought of as the worst-case quality (with respect to the class distributions) of the best estimator with the best implementation under the specified communication budget B; this is to say, Π must obey the communicatin budget B when constructing $\hat{\theta}$. Note that the expectation is taken over randomness in the underlying data (and therefore randomness in the messages). We do not consider randomized protocols, although this represents an interesting direction for continuing research.

2.3. *PAC-Learnability.* The notion of the probably approximately correct (PAC) estimator is an important one in machine learning [1]. The intuition behind the PAC estimator is as follows: that the estimator may be made to have arbitrarily small deviation from its true value if the number of samples used to compute the estimator is allowed to grow in an unbounded fashion. For our purposes, it is equivalent to think of as estimator as being PAC if it is consistent with its true value.

Example 2.1. As an example of PAC-learnability, consider flipping a coin with probability of success p. Suppose that n samples are drawn from

the underlying Bernoulli probability mass function and denote the samples $\{X_1, \ldots, X_n\}$. Then calculating the empirical average $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, we apply Hoeffding's inequality,

$$\mathbb{P}\left[\left|\bar{X} - p\right| > \epsilon\right] \leq 2e^{-2n\epsilon^2}. \tag{2.5}$$

This example shows that estimating the success probability of a Bernoulli random variable is possible with the probability of ϵ-large deviations exponentially vanishing with n.

2.4. *Information Theory and Fano's Inequality.* A staple of our analysis is Fano's Inequality [3], which characterizes, in terms of the mutual information, the error probability of an estimator of a quantity. The usual form of Fano's Inequality is as follows.

THEOREM 2.1 (Fano's Inequality). *Let X be a random variable drawn uniformly at random from one of $k \geq 2$ probability distributions parametrized on $\Theta \subset \mathbb{R}^d$, denoted $\{f_{\theta_1}, \ldots, f_{\theta_k}\}$. Let $\Xi \in [k]$ be the index of that $\theta_\Xi \in \Theta$ giving rise to X. Then an estimator of Ξ, denoted $\hat{\Xi}$, has error probability,*

$$\mathbb{P}\left[\hat{\Xi} \neq \Xi\right] \geq 1 - \frac{\beta + 1}{\log_2 k}, \tag{2.6}$$

where β is an upper bound on the Kullback-Leibler divergence for any two (ordered) pairs of probability distributions f_i and f_j for $i \neq j$ [11].

COROLLARY 2.1. *Fano's Inequality may be, in a sense, generalized to characterize expected error according to some metric [14]. Let $\mathcal{D}(\cdot, \cdot)$ be a metric on Θ. Then if for $i \neq j$ we have $\mathcal{D}(\theta_i, \theta_j) \geq \alpha$ then,*

$$\max_{\Xi \in [k]} \mathbb{E}\left[\mathcal{D}\left(\theta_\Xi, \theta_{\hat{\Xi}}\right)\right] \geq \frac{\alpha}{2}\left(1 - \frac{\beta + 1}{\log_2 k}\right). \tag{2.7}$$

REMARK 2.2. Theorem 2.1 and Corollary 2.1 are standard results in information theory. For proofs of these results, one may refer to the classic text on information theory by Cover [3] and for a treatment of the generalization in the corollary refer to [14].

3. Main Results of Information Theory.

THEOREM 3.1. *This proposition is a modified version of the one found in [4] to be simultaneously tighter and easier to construct. Indeed, for any*

class of distributions $\mathbf{F}$ *and for any protocol* Π *with communication budget* B *and parameter space* Θ*,*

$$\mathfrak{M}(\Theta, B) \;\geq\; \delta\left(1 - \frac{B+1}{\log_2 K_\delta}\right), \tag{3.1}$$

for any $\delta > 0$*. Here,* K_δ *denotes the maximal* 2δ *packing number of* Θ *in the* ℓ_1 *distance measure.*

Proof. Fix a distance parameter $\delta > 0$ and form a collection of point parameters $\{\theta_1, \ldots, \theta_{K_\delta}\}$ that form a maximal 2δ-packing of the parameter space Θ.

Suppose that we generate an index Ξ uniformly at random from $[K_\delta]$. We then generate a set of n points from the distribution f_{θ_Ξ} parametrized by θ_Ξ, denoted $\mathbf{X}^n$. We use the classical approach of reducing the estimation problem to a testing problem. Denote by $M = \{M_1, \ldots, M_m\}$ the set of messages communicated in a m-round protocol Π. Denote by $\hat{\theta}$ an arbitrary estimator of θ based on M and define the testing function,

$$\hat{\Xi} = \underset{k \in [K_\delta]}{\arg\min} \left|\left|\hat{\theta}(M) - \theta_k\right|\right|_1. \tag{3.2}$$

Since the point parameter set forms a maximal 2δ-packing of Θ, we are guaranteed that $||\hat{\theta}(M) - \theta_k||_1 \geq 2\delta$ whenever $\hat{\Xi} \neq \Xi$. Leveraging Fano's Inequality,

$$\max_{k \in [K_\delta]} \mathbb{E}\left[\left|\left|\hat{\theta}(M) - \theta_k\right|\right|_1\right] \;\geq\; \delta \cdot \mathbb{P}\left[\hat{\Xi} \neq \Xi\right] \tag{3.3}$$

$$\geq\; \delta\left(1 - \frac{\mathcal{I}(\Xi : M) + 1}{\log_2 K_\delta}\right). \tag{3.4}$$

Obviously, $\mathcal{I}(\Xi : M) \leq H(M) \leq B$ by Shannon's source coding theorem. The result follows immediately. □

Example 3.1. Let Θ denote the unit $||\cdot||_1$-ball in $\mathbb{R}^d$. It is a classical result that $K_\delta \geq \left(\frac{1}{\delta}\right)^d$. Suppose that each machine i receives n observations $\mathbf{X}_i$ generated from f, where f belongs to a class of distributions $\mathbf{F}$ parametrized by Θ. By applying Theorem 3.1 we obtain,

$$\mathfrak{M}(\Theta, B) \geq \delta\left(1 + \frac{B+1}{d \log_2 \delta}\right). \tag{3.5}$$

Because δ is not constrained (aside from being positive), we choose to arbitrarily fix $\delta = \frac{1}{10}$. Then if,

$$B = \frac{30dn \log_2 10 - d \log_2 10 - 30n}{30n}, \tag{3.6}$$

we obtain,

$$\mathfrak{M}(\Theta, B) = \Omega\left(n^{-1}\right), \tag{3.7}$$

by elementary algebra.

EXAMPLE 3.2. Consider a class of distributions $\mathbf{F}$ whose support is the unit interval. If $X_1, \ldots, X_n \sim f \in \mathbf{F}$ are generated i.i.d. from a particular distribution in the class, we may suppose that we are interested in computing the mean value of f based on our data, denoted θ. Denoting the space of mean parameters $\Theta = [0, 1]$ and setting $B = \log_2 n$ (and assuming $\mathbb{E}[X_i] = \theta$), we will illustrate that for a single-computer algorithm $\mathfrak{M}(\Theta, B) = \mathcal{O}\left(n^{-1/2}\right)$.

Clearly a natural estimator of the distribution mean is the sample average. We define the estimator $\hat{\theta} = n^{-1}\sum_{i=1}^{n} X_i$. It is apparent that $\hat{\theta} \in [0, 1]$ since the data points are individually within that range. A straightforward calculation of ℓ_2 mean squared error is,

$$\mathbb{E}\left[\left|\left|\hat{\theta} - \theta\right|\right|_2^2\right] = \mathbb{V}\left[\hat{\theta}\right] = \mathbb{V}\left[n^{-1}\sum_{i=1}^{n} X_i\right] \tag{3.8}$$

$$= n^{-2}\sum_{i=1}^{n}\mathbb{V}[X_i] = n^{-1}\mathbb{V}[X_1] \tag{3.9}$$

$$\leq \frac{1}{n}. \tag{3.10}$$

We now define a quantized version of $\hat{\theta}$ denoted $\tilde{\theta}$. In particular, let $\tilde{\theta}$ be $\hat{\theta}$ quantized to $\log_2 n$ bits. Because $\hat{\theta} \in [0, 1]$, we are guaranteed that $\left|\hat{\theta} - \tilde{\theta}\right| \leq n^{-1}$. As such, we may equivalently write $\tilde{\theta} = \hat{\theta} + \epsilon$ for $|\epsilon| \leq n^{-1}$. It can be verified (refer to Appendix B) that the worst-case performance in ϵ is achieved when $\epsilon = n^{-1}$. We will now analyze the mean squared error of $\tilde{\theta}$ by first noting the common identity,

$$\mathbb{E}\left[\left(\tilde{\theta} - \theta\right)^2\right] = \text{Bias}\left(\tilde{\theta}\right)^2 + \mathbb{V}\left[\tilde{\theta}\right]. \tag{3.11}$$

Then by direct calculation we see,

$$\text{Bias}\left(\tilde{\theta}\right) = \mathbb{E}\left[\tilde{\theta} - \theta\right] \leq \mathbb{E}\left[\hat{\theta} + n^{-1} - \theta\right] = n^{-1}. \tag{3.12}$$

Analysis of the variance of $\tilde{\theta}$ is similarly straightforward.

$$\mathbb{V}\left[\tilde{\theta}\right] = \mathbb{E}\left[\left(\tilde{\theta} - \mathbb{E}\left[\tilde{\theta}\right]\right)^2\right] \tag{3.13}$$

$$\leq \mathbb{E}\left[\left(\hat{\theta} + n^{-1} - \mathbb{E}\left[\hat{\theta} + n^{-1}\right]\right)^2\right] \tag{3.14}$$

$$= \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \tag{3.15}$$

Combining our earlier inequality on the final quantity, these results along with the conclusion in Lemma A.1 yield the result,

$$\mathbb{E}\left[\left|\tilde{\theta} - \theta\right|\right] \leq \sqrt{\mathbb{E}\left[\left(\tilde{\theta} - \theta\right)^2\right]} \tag{3.16}$$

$$= \sqrt{\frac{1}{n^2} + \frac{2}{n}} \tag{3.17}$$

$$\leq \sqrt{\frac{4}{n}}, \tag{3.18}$$

where the last inequality follows when $n \geq 1$, which must be true for any reasonable estimation task.

Definition 3.1 (L-Lipschitz). Let $\mathbf{F}$ be a class of functions parametrized by $\Theta \subset \mathbb{R}^d$. Then if $||\cdot||_{\mathbf{F}}$ is a norm on $\mathbf{F}$ and if $||\cdot||_{\Theta}$ is a norm on Θ, then the mapping $\theta \mapsto f(\cdot;\theta)$ is L-Lipschitz if,

$$\left|\left|f(\cdot;\theta) - f\left(\cdot;\theta'\right)\right|\right|_{\mathbf{F}} \leq L\left|\left|\theta - \theta'\right|\right|_{\Theta}. \tag{3.19}$$

Theorem 3.2. *Let $\mathbf{F}$ be a class of parametrized functions; that is, $\mathbf{F} = \{f(\cdot;\theta) : \theta \in \Theta, f : [0,1]^s \to [0,1]\}$ for an arbitrary space $\Theta \subset \mathbb{R}^d$, where both $d, s \in \mathbb{N}$. Let the $||\cdot||_1$ be a norm on Θ and let $||f||_{\infty} = \sup_{x\in[0,1]} f(x;\theta)$ be a norm on $\mathbf{F}$. If the mapping $\theta \mapsto f(\cdot;\theta)$ is L-Lipschitz, and if $L \geq 2\delta$, then the minimax risk is lower bounded like,*

$$\mathfrak{M}(\Theta, B) \geq \delta\left(1 - \frac{B+1}{\left(\frac{1}{(2\delta L)^s}\right)}\right), \tag{3.20}$$

for any $\delta > 0$.

Proof. The L-Lipschitz permits the lower bound on the covering number in $\mathbf{F}$-space to be expanded to a lower bound on the maximal packing number

in Θ-space. This idea is captured precisely in Lemma A.2. A straightforward analysis reveals that,

$$\log_2 K_\delta \geq \frac{1}{2\delta L}. \tag{3.21}$$

A careful treatment of this lower bound can be found in Appendix C. The desired inequality follows immediately from Theorem 3.1. □

EXAMPLE 3.3. Suppose $\Theta = [0,1]$. Then we may define the class,

$$\mathbf{F} = \left\{ f : f(x;\theta) = \theta^x (1-\theta)^{1-x} \text{ and } x \in \{0,1\} \right\}. \tag{3.22}$$

In other words, we permit the class of functions $\mathbf{F}$ to represent the class of Bernoulli densities with success probability in the unit interval. Notice that Θ and $\mathbf{F}$ are Lipschitz with $L = 1$. This can easily be seen as follows, allowing $x \in \{0,1\}$,

$$\sup_x \left| f(x;\theta) - f\left(x;\theta'\right) \right| = \sup_x \left| \theta^x (1-\theta)^{1-x} - \theta'^x \left(1-\theta'\right)^{1-x} \right| \tag{3.23}$$

$$= \left| \theta - \theta' \right| \tag{3.24}$$

$$= \left|\left| \theta - \theta' \right|\right|_1. \tag{3.25}$$

A Lipschitz parametrization permits the covering number of the function space to be converted into a lower bound for the packing number of the parameter space. Therefore, by Theorem 3.2 it can be seen that,

$$\mathfrak{M}([0,1], B) \geq \delta \left(1 - \frac{B+1}{\left(\frac{1}{(2\delta)}\right)} \right), \tag{3.26}$$

for any $\delta > 0$. This lower bound is positive whenever $B > 0$ and $0 < \delta < \frac{1}{2(B+1)}$.

4. Main Results of PAC-Learnability. We demonstrate two results relating to two classical problems in theoretical machine learning. In particular, we treat the distributed learning of 3-term disjunctive normal forms and of multi-dimensional half-spaces. We find that these two learning problems fit naturally within the distributed setting, wherein communication may be restricted to a particular number of bits. We begin with a well-known result for conjunctions that will prove to be useful later on.

EXAMPLE 4.1. This example is similar to the one in [2]. Denote by $X_i \in \{0,1\}^{2p}$ a boolean vector and suppose that we form a dataset by $\mathbf{X}^n = \{X_i\}_{i=1}^n$. Then let $\mathcal{C}$ be the set of conjunctions of the measured boolean random variables on p boolean random variables and their negations. For instance, with $p = 5$, the concept $c(X) = X_1 \wedge X_3 \wedge X_5$ would represent a valid conjunction. Let $\mathbf{X}^n$ be labeled according to a particular $c \in \mathcal{C}$ so that for every $i \in [n]$ we obtain $y_i = c\,(X_i)$.

Now form a candidate hypothesis by taking $h(X) = X_1 \wedge \bar{X}_1 \wedge \ldots \wedge X_p \wedge \bar{X}_p$, where $\bar{X}$ is the logical negation of X. It is known that it is possible to PAC-learn the out of sample error rate of h (which should be zero) by taking those X_i with the property that $y_i = 1$ and deleting from h all those logical literals that are inconsistent with the positive examples of the data. It is important to note that when h is constructed in this way, logical literals that are present in c are never removed from the hypothesis so that only errors of the misdetection type are possible.

One equivalent way to think about the construction of h is simply to take all of the positive examples of $\mathbf{X}^n$ and take their bitwise AND down the features. Those boolean variables whose entries were positive in all of the positive training examples constitute the learned conjunction. In a distributed computing environment, this h can be obtained by having each computer take the bitwise AND of its positive examples and transmit that $2p$-vector to the blackboard. This computation takes exactly $2p\Gamma$-bits of communication.

DEFINITION 4.1. A 3-term conjunctive normal form (CNF) is a conjunction with at most three logical literals per logical clause [12]. In particular,

$$\bigwedge_i (x_i \vee y_i \vee z_i)\,, \tag{4.1}$$

is an example of a conjunctive normal form.

DEFINITION 4.2. A 3-term disjunctive normal form (DNF) is a disjunction of at most three logical clauses, each of which may consist may consist of an arbitrary number of logical literals [12]. In particular, if S_1, S_2, and S_3 are three logical clauses, then their DNF is,

$$S_1 \vee S_2 \vee S_3. \tag{4.2}$$

PROPOSITION 4.1. *Every* 3*-term DNF is representable by a* 3*-term CNF, though the converse is not true. Indeed, we obtain,*

$$S_1 \vee S_2 \vee S_3 = \bigwedge_{x \in S_1} \bigwedge_{y \in S_2} \bigwedge_{z \in S_3} (x \vee y \vee z)\,. \tag{4.3}$$

PROPOSITION 4.2. *Let there be k computers. Let $\mathcal{C}$ be the class of 3-DNFs. Then if $c \in \mathcal{C}$ and $\mathbf{X}^n \in \{0,1\}^{n \times p}$ is labeled according to c, and if the examples of $\mathbf{X}^n$ are equally partitioned across the k computers, then c is PAC-learnable with,*

$$\min\left\{k \cdot n \cdot d, k \cdot \binom{d}{3} \cdot 2^3\right\} \tag{4.4}$$

bits of communication.

PROOF. First observe by Lemma A.5 that the number of 3-CNFs obtainable from $2p$ boolean random variables (a number which includes the negations of p of those random variables) is $\binom{d}{3} \cdot 8$. One approach to PAC-learning the 3-DNF is to "explode" the $2p$ boolean random variables into the $\binom{p}{3} \cdot 8$ boolean random variables obtained by forming new variables by taking 3-term disjunctions. This transforms the problem of estimating the 3-DNF directly into an equivalent problem of estimating a large 3-CNF. This is known to be possible by taking the bitwise logical AND of the data.

On the other hand, if $n \cdot p$ is less than the size of the representation obtained by exploding the boolean random variables into their equivalent 3-CNF form, each computer may simply transmit their entire dataset, thereby rendering the problem trivial to solve.

This completes the proof, and shows that n must be $\mathcal{O}\left(p^3\right)$ in order to make distributed learning more interesting. □

We turn our attention now to the problem of learning multi-dimensional half-spaces [8]. To explain the concept more clearly, we consider the 1-dimensional setting: Consider $x \in \mathbb{R}$ and let $c(x) = \mathbf{1}\{x > \theta\}$ for some $\theta \in \mathbb{R}$. Learning the half-space amounts to constructing an estimator $\hat{\theta}$ of the threshold parameter which labels x on the basis of data collected from the system.

Assuming that the collected data, denoted $\mathbf{X}^n \in \mathbb{R}^n$, where each $X_i \sim \mathcal{U}(-a, a)$, has both positively and negatively labeled examples, it is natural to take an estimator of θ of the form,

$$\hat{\theta} = \frac{\max\{X_i : c(X_i) = 0\} + \min\{X_i : c(X_i) = 1\}}{2}. \tag{4.5}$$

Unfortunately, this estimator can be shown to be generally biased (although the bias goes to zero as the number of samples goes to infinity). This will not do for our analysis, so we instead develop a new estimator.

LEMMA 4.1. *Supposing that one samples from the 1-dimensional system n times. Suppose further that n_1 samples are labeled as belonging to the negative class and n_2 samples are labeled positively such that $n_1 + n_2 = n$. We further impose the restriction that we have at least one example from both the negative and positive classes. Denote by Z_1 the maximum of the negatively labeled examples and Z_2 the minimum of positive examples. Furthermore, define Z_{max} as the maximum of the positive examples and Z_{min} as the minimum of the negative examples. Then the estimator,*

$$\hat{\theta} = \frac{Z_1 + Z_2}{2} + \frac{Z_{max} + Z_{min}}{2}, \tag{4.6}$$

is unbiased [7].

PROOF. The proof of this is easy with the contents of Appendix D in hand, relating to the extrema of uniform random variables. By direct computation,

$$\mathbb{E}\left[\hat{\theta}\right] = \frac{1}{2}\mathbb{E}\left[Z_1\right] + \frac{1}{2}\mathbb{E}\left[Z_2\right] + \frac{1}{2}\mathbb{E}\left[Z_{max}\right] + \frac{1}{2}\mathbb{E}\left[Z_{min}\right] \tag{4.7}$$

$$= \theta + \frac{-\theta - a}{2\left(n_1 + 1\right)} + \frac{a - \theta}{2\left(n_2 + 1\right)} + \frac{\theta - n_1 a}{2\left(n_1 + 1\right)} + \frac{n_2 a + \theta}{2\left(n_2 + 1\right)} \tag{4.8}$$

$$= \theta. \tag{4.9}$$

This completes the proof. □

Now, we know that the interval $(-a, a)$ may be quantized to accuracy $\frac{2a}{m}$ using $\log_2 m$-bits. From this, we obtain the following proposition.

PROPOSITION 4.3. *Half-space learning in Γ dimensions with a data set split across Γ computers can be learned to accuracy $\frac{2a\Gamma}{m}$ using $\Gamma \log_2 m$-bits of communication. Since the estimator, by inspection, has vanishing variance with n, for appropriately large m, the quantized estimator can have arbitrarily low probability of ϵ-large deviations from the truth. Hence, the estimator is probably approximately correct.*

5. Conclusion. In this work we have demonstrated some results in an area at the intersection of statistics and computation. As data becomes more ubiquitous, and in ever-increasing quantities, it is important to develop a theory of statistical estimation and machine learning that pays heed to the computational costs involved. Indeed, with the availability of cluster computing, it is especially important to obtain a theoretical sense of what is achievable under communication budgets.

This work in particular has dealt with two realms of analysis with respect to statistical learning under communication budgets. We have developed some general theory, which makes heavy use of Fano's Inequality, and which demonstrates in a number of cases a non-trivial lower bound on the number of bits required to estimate statistical quantities from data. We also consider the concept of PAC-learnability, and give a more explicit treatment of certain classical problems from the field of theoretical machine learning. This analysis yields specific distributed algorithms for solving learning problems under communication constraints.

6. Acknowledgments. To begin with, I would like to thank Professor Peter Winkler for agreeing to supervise this research officially. I would like to thank Professor Peter Doyle for agreeing to supervise this research unofficially (and thereby making the whole thing possible). Professor Amit Chakrabarti was also invaluable as a source of discussion about elements of the work. Professor Eugene Demidenko also provided helpful feedback regarding the statistical components of the work. Professor Craig Sutton and James Drain '17 were also helpful for discussion bounds on the metric entropy.

StackOverflow was also a useful resource when I was having trouble with concepts, and I am indebted to the many helpful members there.

Additionally, I would like to dedicate this thesis to Rui Shu '15 for his endless support, optimism, and enthusiasm over the past few weeks.

APPENDIX A: TECHNICAL LEMMAS

LEMMA A.1. *Let $\hat{\theta}$ be an arbitrary estimator of a scalar parameter of interest θ. We can relate the ℓ_2 mean squared error to that of the ℓ_1 by the concavity of the square root function and an application of Jensen's inequality,*

$$\mathbb{E}\left[\left|\left|\hat{\theta}-\theta\right|\right|_1\right] = \mathbb{E}\left[\left|\hat{\theta}-\theta\right|\right] \tag{A.1}$$

$$= \mathbb{E}\left[\sqrt{\left(\hat{\theta}-\theta\right)^2}\right] \tag{A.2}$$

$$\leq \sqrt{\mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right]}, \tag{A.3}$$

where the final inequality follows from the result that $\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}$.

LEMMA A.2. *Denote the covering number of a set by $N(\cdot,\cdot,\cdot)$ and the maximal packing number $M(\cdot,\cdot,\cdot)$. Then covering number and maximal packing number of a set Θ under an arbitrary pseudo-metric $||\cdot||$ satisfy the relations,*

$$M(2\epsilon,\Theta,||\cdot||) \leq N(\epsilon,\Theta,||\cdot||) \leq M(\epsilon,\Theta,||\cdot||), \tag{A.4}$$

for all $\epsilon > 0$.

LEMMA A.3. *Denote by Y_i the message sent by the i^{th} machine. Then if Y_i is a $\mathbf{X}_i$-measurable function, we have by the data processing inequality,*

$$\mathcal{I}\left(V : \{Y_i\}_{i=1}^{\Gamma}\right) \leq \sum_{i=1}^{\Gamma} \mathcal{I}(V : Y_i) \leq \sum_{i=1}^{\Gamma} \mathcal{I}(\mathbf{X}_i : Y_i). \tag{A.5}$$

where the data $\mathbf{X}_i$ is drawn from the distribution $\mathbb{P}[\cdot;\theta_V]$ for V drawn uniformly at random from the d-hypercube.

LEMMA A.4. *For any $\Theta \subset \mathbb{R}^d$ if $\{\theta_1,\ldots,\theta_{K_\delta}\}$ is a maximal packing of Θ for a norm $||\cdot||$, then,*

$$K_\delta \geq \frac{Volume(\Theta)}{Volume(\{x \in \mathbb{R}^d : ||x|| \leq \delta\})}. \tag{A.6}$$

PROOF. Clearly,

$$\Theta \subset \bigcup_{i=1}^{K_\delta} \left\{x \in \mathbb{R}^d : ||x-\theta_i|| \leq \delta\right\}, \tag{A.7}$$

since otherwise $\exists\, \theta_{K_\delta+1}$ such that $\{\theta_1,\ldots,\theta_{K_\delta+1}\}$ is a packing, contradicting the assumption that K_δ is the maximal packing number. Thus,

$$\text{Volume}(\Theta) \leq K_\delta \cdot \text{Volume}\left(\left\{x \in \mathbb{R}^d : ||x|| \leq \delta\right\}\right). \tag{A.8}$$

The result follows immediately from simple rearrangement of terms. □

LEMMA A.5. *Let $X_1, \bar{X}_1 \ldots, X_d, \bar{X}_d$ be a set of boolean random variables and their negations. Then the number of 3-CNFs that can be formed from this set logical literals is,*

$$|3\text{-}CNF| = \binom{d}{3} \cdot 2^3. \tag{A.9}$$

This can be verified through a simple counting argument.

APPENDIX B: LARGE DEVIATIONS ARE THE WORST CASE

Denote by $\tilde{\theta}_{1/n}$ the quantization of $\hat{\theta}$ for which $\epsilon = n^{-1}$. This represents the worst-case corruption of $\hat{\theta}$. We will illustrate that,

$$\mathbb{E}\left[\left(\tilde{\theta}-\theta\right)^2\right] \leq \mathbb{E}\left[\left(\tilde{\theta}_{1/n}-\theta\right)^2\right]. \tag{B.1}$$

First begin by recalling that the mean squared error of $\tilde{\theta}_{1/n}$ is $\frac{2}{n}+\frac{1}{n^2}$. Denote by $\bar{\epsilon}$ the mean value of ϵ. Then we obtain,

$$\begin{aligned}
\mathbb{E}\left[\left(\tilde{\theta}-\theta\right)^2\right] &= \mathbb{E}\left[\left(\hat{\theta}+\epsilon-\theta-\bar{\epsilon}\right)^2\right] && \text{(B.2)}\\
&= \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2+(\epsilon-\bar{\epsilon})^2+2\left(\hat{\theta}-\theta\right)(\epsilon-\bar{\epsilon})\right] && \text{(B.3)}\\
&= \frac{2}{n}+\mathbb{V}[\epsilon]+\sigma\left(\hat{\theta},\epsilon\right) && \text{(B.4)}\\
&= \frac{2}{n}+\mathbb{V}[\epsilon]. && \text{(B.5)}
\end{aligned}$$

Since $\mathbb{V}[\epsilon] = \mathbb{E}\left[\epsilon^2\right]-\bar{\epsilon}^2 \leq \mathbb{E}\left[\epsilon^2\right]$ it is easy to see that $\mathbb{V}[\epsilon] \leq n^{-2}$. This yields the claim.

APPENDIX C: COVERING NUMBER OF L-LIPSCHITZ FUNCTIONS

Consider the set of L-Lipschitz functions mapping $[0,1] \to [0,1]$. Equipped the supremum norm, we consider the subset of functions that are piecewise linear on $[0,1]$. In particular, we decompose the domain into subintervals, each of length $\frac{\epsilon}{L}$. We also partition the codomain into subintervals of length ϵ.

Now consider the set of intervals on the domain,

$$S = \left\{\left[\frac{i\epsilon}{L}, \frac{(i+1)\,\epsilon}{L}\right) : 0 \leq i \leq \left\lceil\frac{L}{\epsilon}\right\rceil - 1\right\}. \tag{C.1}$$

Assume that $L \geq 2\epsilon$. For each $T \subset S$, define,

$$\phi_T(x) = \begin{cases} L & \text{if } x \in \bigcup_{t\in T} t \\ 0 & \text{otherwise} \end{cases}. \tag{C.2}$$

For every $T_1 \neq T_2$ then we have,

$$||\phi_{T_1} - \phi_{T_2}||_\infty \geq L \geq 2\epsilon. \tag{C.3}$$

Therefore, we have that $\{\phi_T : T \subset S\}$ is a collection of $2^{|S|} = 2^{\lceil L/\epsilon \rceil}$ functions mutually L-apart. Therefore, they are also 2ϵ-apart. Plainly, we must then have that the ϵ-covering number of the space is at least $2^{\lceil L/\epsilon \rceil}$.

The proof of an analogous lower bound for Lipschitz functions mapping $[0,1]^s \to [0,1]$ for $s \in \mathbb{N}$ may be obtained in a similar manner. In particular, rather than considering 1-dimensional intervals of the domain, one considers s-dimensional "panels." There are $(\lceil L/\epsilon \rceil)^s$ such panels and the result follows.

APPENDIX D: EXPECTED VALUE OF THE EXTREMES OF UNIFORM RANDOM VARIABLES

Let $X_1, \ldots, X_n$ be n i.i.d. random variables distributed uniformly at random on (a, b). Define the random variable, $Y = \min_{i \in [n]} \{X_i\}$. We will compute the expectation of Y. We begin by computing the cumulative density function of Y. By inspection,

$$\mathbb{P}[Y \leq y] = 1 - \mathbb{P}[Y > y] \tag{D.1}$$

$$= 1 - \mathbb{P}\left[\min_{i \in [n]} \{X_i\} > y\right] \tag{D.2}$$

$$= 1 - \prod_{i=1}^{n} \mathbb{P}[X_i > y] \tag{D.3}$$

$$= 1 - \mathbb{P}[X_1 > y]^n \tag{D.4}$$

$$= 1 - \left(\frac{b-y}{b-a}\right)^n \tag{D.5}$$

From this cumulative density function we obtain the probability density,

$$\mathbb{P}[Y = y] = \left(\frac{n}{b-a}\right)\left(\frac{b-y}{b-a}\right)^n, \tag{D.6}$$

defined for $y \in (a, b)$. Hence the expectation of Y can be computed through an elementary integral. Indeed,

$$\mathbb{E}[Y] = \int_a^b \left(\left(\frac{n}{b-a}\right)\left(\frac{b-y}{b-a}\right)^n\right) \cdot y dy \tag{D.7}$$

$$= \frac{b - na}{n+1}. \tag{D.8}$$

If on the other hand we had defined $Z = \max_{i \in [n]} \{X_i\}$, we would have obtained that $\mathbb{E}[Z] = \frac{nb+a}{n+1}$, as expected [5].

REFERENCES

[1] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*, First Edition. AMLBook, 2012.

[2] M.-F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed Learning, Communication Complexity and Privacy. In *COLT*, 2012.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*, Second Edition. Wiley, 2006.

[4] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Yuchen Zhang. Optimality guarantees for distributed statistical estimation. *arXiv:1405.0782*, 2014.

[5] "Expected Value of Maximum of Two Random Variables from Uniform Distribution." *Mathematics – StackExchange*. StackExchange, 8 May 2014. Web. 08 May 2015.

[6] Peng He, Changshui Zhang: Exploiting the Limits of Structure Learning via Inherent Symmetry. AISTATS 2014: 328-337.

[7] "Interval Learning Unbiased Estimator." *Mathematics – StackExchange*. StackExchange, 5 Apr. 2015. Web. 08 May 2015.

[8] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.

[9] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*, First Edition. Cambridge University Press, 2006.

[10] Rowland, Todd and Weisstein, Eric W. "Lipschitz Function." From MathWorld–A Wolfram Web Resource.

[11] N. Santhanam and M. J. Wainwright (2012). *Information-theoretic limits of selecting binary graphical models in high dimensions*. IEEE Transactions on Information Theory, 58(7): 4117–4134, July 2012.

[12] Weisstein, Eric W. "Disjunctive Normal Form." From MathWorld–A Wolfram Web Resource.

[13] Yang, Yuhong and Barron, Andrew. Information-theoretic determination of minimax rates of convergence. Ann. Statist. 27 (1999), no. 5, 1564–1599. doi:10.1214/aos/1017939142.

[14] Yu, Bin. *Festschrift for Lucien Le Cam Research Papers in Probability and Statistics*. New York: Springer Verlag, 2013. Print.

E-mail: james.a.brofos.15@dartmouth.edu
May 28, 2015