Optimal Distributed Learning with Two Rounds of Communication

Anonymous Authors[1]

Abstract

We present the OWA algorithm for distributed empirical risk minimization. OWA requires only two rounds of communication, and the total number of bits transfered is comparable to the naive averaging distributed estimator. Given m machines, each with n data points, the mean squared error of OWA shrinks at the rate of $O(1/mn)$, which matches the optimal error rate of the single machine oracle. The analysis relies on simple properties of sub-Gaussian random vectors, and does not require the loss to be convex.

1. INTRODUCTION

Many modern datasets are too large to fit in the memory of a single machine, so they must be partitioned onto many machines. To analyze these datasets, we need distributed algorithms. Existing distributed algorithms can be classified as either interactive or non-interactive depending on their communication complexity. In this paper we propose an algorithm that exhibits the benefits of both types.

Interactive algorithms require many rounds of communication between machines. These algorithms often resemble standard iterative algorithms where each iteration is followed by a communication step. The appeal of interactive algorithms is that they enjoy the same statistical regret bounds as standard sequential algorithms. But, there are two downsides. First, these algorithms can be too slow in practice because communication is the main bottleneck in modern distributed architectures. Second, these algorithms require special implementations and do not work with off-the-shelf statistics libraries provided by (for example) Python, R, and Matlab.

Non-interactive algorithms require only a single round of communication. They are significantly faster than interactive algorithms and easily implemented with standard libraries. The downside is worse regret bounds. Recent work (discussed in Section 3.2) has shown that no non-interactive algorithm can achieve regret bounds comparable to an interactive one.

In this paper, we propose a *semi-interactive* distributed algorithm called *optimal weighted averaging* (OWA). Our algorithm performs two rounds of communication, so it is not subject to the existing regret bounds of non-interactive algorithms. The algorithm has two tunable parameters that let the user trade better statistical performance for worse communication complexity. These parameters do not require cross-validation to set correctly; increasing the amount of communication will always improve the statistical performance. Therefore, these parameters are determined by the communication constraints and not the underlying data. The OWA algorithm is easily implemented in a MapReduce architecture with standard packages.

In the next section, we formally describe the OWA algorithm. In Section 3, we compare OWA to existing distributed algorithms. We highlight how the analysis of existing algorithms requires more limiting assumptions than our own, and show in detail why existing non-interactive regret bounds do not apply to OWA. Section 4 shows that OWA's regret bounds interpolate between the averaging estimator's regret and the optimal regret. As part of the analysis, we provide novel, more general regret bounds for the averaging estimator. Section 5 shows experimentally that our algorithm performs well. We emphasize that our algorithm is robust to the strength of regularization, which is one of the reasons it performs well in practice.

2. THE OWA ALGORITHM

This section first formally introduces the problem of communication efficient distributed estimation, then describes our proposed OWA distributed estimator.

2.1. Problem Setting

Let $\mathcal{Y} \subseteq \mathbb{R}$ be the space of response variables, $X \subseteq \mathbb{R}^d$ be the space of covariates, and $\Theta \subseteq \mathbb{R}^d$ be the parameter space. We assume a linear model where the loss of data point $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ given the parameter $\theta^* \in \Theta$ is denoted by $\ell(y, \mathbf{x}^\mathsf{T}\theta^*)$. We do not require that the model be correctly specified, nor do we require

that ℓ be convex with respect to θ. Let $Z \subset \mathcal{X} \times \mathcal{Y}$ be a dataset of mn i.i.d. observations. Finally, let $r : \Theta \to \mathbb{R}$ be a regularization function (typically the L1 or L2 norm) and $\lambda \in \mathbb{R}$ be the regularization strength. Then the regularized empirical risk minimizer (ERM) is

$$\hat{\theta}^{erm} = \arg\max_{\theta} \sum_{(\mathbf{x},y)\in Z} \ell(y, \mathbf{x}^\mathsf{T}\theta) + \lambda r(\theta). \quad (1)$$

In the remainder of this paper, it should be understood that all ERMs are regularized.

Assume that the dataset Z has been partitioned onto m machines so that each machine i has dataset Z_i of size n, and all the Z_i are disjoint. Then each machine calculates the local ERM

$$\hat{\theta}_i^{erm} = \arg\max_{\theta} \sum_{(\mathbf{x},y)\in Z_i} \ell(y, \mathbf{x}^\mathsf{T}\theta) + \lambda r(\theta). \quad (2)$$

Solving for $\hat{\theta}_i^{erm}$ requires no communication with other machines. Our goal is to merge the $\hat{\theta}_i^{erm}$ into a single improved estimate. A baseline merging procedure is the averaging estimator

$$\hat{\theta}^{ave} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^{erm}. \quad (3)$$

This estimator is well studied, and in Section 3 we compare this previous work to our own.

2.2. Proposed Solution

We propose a modification to the averaging estimator called the *optimal weighted average* (OWA). OWA uses a second round of optimization to calculate the optimal linear combination of the $\hat{\theta}_i^{erm}$s. This second optimization occurs over a small fraction of the dataset, so its computational and communication cost is negligible.

To motivate our estimator, we first present an estimator that uses the entire dataset for the second round of optimization. Define the matrix $\hat{W} : \mathbb{R}^{d\times m}$ to have ith column equal to $\hat{\theta}_i^{erm}$. Now consider the estimator

$$\hat{\theta}^{owa,full} = \hat{W}\hat{\alpha}^{full}, \quad (4)$$

where

$$\hat{\alpha}^{full} = \arg\max_{\alpha} \sum_{(\mathbf{x},y)\in Z} \ell\left(y, \mathbf{x}^\mathsf{T}\hat{W}\alpha\right) + \lambda r(\hat{W}\alpha). \quad (5)$$

Notice that $\hat{\theta}^{owa,full}$ is the empirical risk minimizer when the parameter space Θ is restricted to $\hat{\Theta}^{owa} = \text{span}\{\hat{\theta}_i^{erm}\}_{i=1}^m$. In other words, the $\hat{\alpha}^{full}$ vector contains the optimal weights to apply to each $\hat{\theta}_i^{erm}$ when averaging. Figure 1 shows graphically that no other estimator in $\hat{\Theta}^{owa}$ can have lower regularized empirical loss than $\hat{\theta}^{owa,full}$.

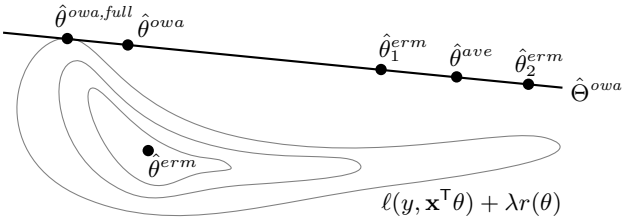

Figure 1. Our method performs a second round of optimization to find the best parameter vector in $\hat{\Theta}^{owa}$. Since $\hat{\Theta}^{owa}$ has low dimension, we can use relatively few data points in the second round of optimization to ensure that with high probability $\hat{\theta}^{owa}$ has lower empirical loss than $\hat{\theta}^{ave}$.

Calculating the weights $\hat{\alpha}^{full}$ directly is infeasible because it requires access to the full dataset. Fortunately, we do not need to consider all the data points for an accurate estimator. The parameter space $\hat{\Theta}^{owa}$ is m-dimensional. So intuitively, we only need $O(m)$ data points to solve the second optimization to our desired accuracy. This intuition motivates the OWA estimator. Let $Z_i^{owa} \subset Z_i$ be a set of n^{owa} data points uniformly sampled from Z_i without replacement, and let Z^{owa} be the union of the Z_i^{owa}s. Then the OWA estimator is defined as

$$\hat{\theta}^{owa} = \hat{W}\hat{\alpha}, \quad (6)$$

where

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{(\mathbf{x},y)\in Z^{owa}} \ell\left(y, \mathbf{x}^\mathsf{T}\hat{W}\alpha\right) + \lambda r(\hat{W}\alpha). \quad (7)$$

Algorithm 1 shows the steps for efficiently computing $\hat{\theta}^{owa}$. In the first round, each machine calculates $\hat{\theta}_i^{erm}$ independently and broadcasts the result to every other machine. A total of $O(dm^2)$ bits are transmitted in this round. (The parameter vector has d dimensions, there are m machines, and each machine transmits to each other machine.) In the second round, each machine projects its local dataset Z_i^{owa} onto the space $\hat{\Theta}^{owa}$. These projected data points are then transmitted to a predesignated master machine. A total of $O(m^2 n^{owa})$ bits are transmitted. (The projected data points each have dimension m, each of the m machines makes a single communication, and there are n^{owa} total data points per machine.) In general, n^{owa} will be set to $O(m) < O(d)$. So the total data transmitted in both rounds is $O(dm^2)$. The averaging estimator transmits $O(dm)$ bits, so whenever $d > m^2$, OWA has the same communication complexity as averaging.

Algorithm 1 Distributed calculation of $\hat{\theta}^{owa}$

Round 1, each machine i independently:
  loads dataset Z_i
  calculates $\hat{\theta}_i^{erm}$
  broadcasts $\hat{\theta}_i^{erm}$ to all other machines
Round 2, each machine i independently:
  constructs $\hat{W}$
  randomly selects a dataset $Z_i^{owa} \subset Z_i$
  calculates $Z_i^{proj} = \{(\mathbf{x}^\mathsf{T}\hat{W}, y) : (\mathbf{x}, y) \in Z^{owa}\}$
  broadcasts Z_i^{proj} to a master machine
The master calculates $\hat{\theta}^{owa}$

FIXME: argue that regularization isn't really even needed at all.

Equations 5 and 7 cannot be solved directly using off the shelf optimizers because existing optimizers do not support the non-standard regularization term $r(\hat{W}\alpha)$. In practice, it is sufficient to approximate this regularization by L2 regularization directly on the α vector:

$$\lambda r(\hat{W}\alpha) \approx \lambda_2 \|\alpha\|. \tag{8}$$

Intuitively, this is because even when we want the parameter vector θ to be sparse (and so are regularizing by r equal to the L1 norm), we have no reason to believe that the α vector should be sparse. The desired sparsity is induced by the regularization when solving for $\hat{\theta}_i^{erm}$s and maintained in any linear combination of the $\hat{\theta}_i^{erm}$s. The new λ_2 regularization parameter should be set by cross validation. This will be a fast procedure, however, because there are few data points to optimize over, and the L2 regularized problem is much easier to solve than the L1 problem. With this minor modification, our distributed estimator can be implemented using any existing optimizer.

3. RELATED WORK

Related work can be divided into two categories: alternative estimators, and bounds on the communication complexity of distributed learning.

3.1. Alternative estimators

The simplest and most popular communication-efficient estimator is the averaging estimator

$$\hat{\theta}^{ave} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^{erm}. \tag{9}$$

Previous analysis of $\hat{\theta}^{ave}$ makes a number of limiting assumptions. McDonald et al. (2009) analyze $\hat{\theta}^{ave}$ in the special case of L2 regularized maximum entropy models. They provide tail bounds on $\|\theta^* - \hat{\theta}^{ave}\|$, showing that the deviation $\|\mathbb{E}\hat{\theta}^{ave} - \hat{\theta}^{ave}\|$ reduces as $O((nm)^{-1/2})$, but they do not show a reduction in bias. Their analysis uses a martingale technique that requires the radius of the dataset be independent of the size of the dataset. This is a particularly limiting assumption as even the simple case of normally-distributed data does not satisfy it. Zhang et al. (2012) provide a more general analysis showing that the mean squared error (MSE) $\mathbb{E}\|\theta^* - \hat{\theta}^{ave}\|^2$ decays as $O((nm)^{-1} + n^{-2})$. This matches the optimal MSE of $\hat{\theta}^{erm}$ whenever $m < n$. Their analysis also requires a number of technical assumptions. For example, they assume the parameter space Θ is bounded. This assumption does not hold under the standard Bayesian interpretation of L2 regularization as a Gaussian prior of the parameter space. They further make strong convexity and $8th$ order smoothness assumptions which guarantee that $\hat{\theta}_i^{erm}$ is a "nearly unbiased estimator" of θ^*. Most recently, Rosenblatt and Nadler (2016) analyze $\hat{\theta}^{ave}$ in the asymptotic regime as the number of data points $n \to \infty$. This analysis is more general than previous analyses, but it does not hold in the finite sample regime. In Section 4, we provide a simple analysis of $\hat{\theta}^{ave}$ that relaxes all previous assumptions of boundedness or convexity and holds in the finite sample regime.

Other research has focused on modifications to the $\hat{\theta}^{ave}$ estimator to reduce bias. Zinkevich et al. (2010) show that if the training sets partially overlap each other (instead of being disjoint), then the resulting estimator will have lower bias. Lee et al. (2015) and Battey et al. (2015) independently developed techniques for debiasing L1 regularized problems. Zhang et al. (2012) provides a debiasing technique that works for any estimator. It works as follows. Let $r \in (0, 1)$, and Z_i^r be a bootstrap sample of Z_i of size rn. Then the bootstrap average estimator is

$$\hat{\theta}^{boot} = \frac{\hat{\theta}^{ave} - r\hat{\theta}^{ave,r}}{1 - r}, \tag{10}$$

where

$$\begin{aligned} \hat{\theta}^{ave,r} &= \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^{erm,r}, \\ \hat{\theta}_i^{erm,r} &= \arg\max_{\theta} \sum_{(\mathbf{x},y)\in Z_i^r} \ell(y, \mathbf{x}^\mathsf{T}\theta) + \lambda r(\theta). \end{aligned} \tag{11}$$

The intuition behind this estimator is to use the bootstrap sample to directly estimate and correct for the bias. This estimator enjoys a MSE that decays as $O((nm)^{-1} + n^{-3})$ under similar assumptions as their analysis of $\hat{\theta}^{ave}$. There are two main limitations to

$\hat{\theta}^{boot}$. First, the optimal value of r is not obvious and setting the parameter requires cross validation on the entire data set. Our proposed $\hat{\theta}^{owa}$ estimator has a similar parameter λ_2 that needs tuning, but this tuning happens on a small fraction of the data and always with the L2 regularizer. So properly tuning λ_2 is more efficient than r. Second, performing a bootstrap on an unbiased estimator increases the variance. This means that $\hat{\theta}^{boot}$ could perform worse than $\hat{\theta}^{ave}$ on unbiased estimators. Our $\hat{\theta}^{owa}$ estimator, in contrast, will perform at least as well as $\hat{\theta}^{ave}$ with high probability as seen in Figure 1. In Section 5, we show that our estimator has better empirical performance.

Liu and Ihler (2014) propose a more Bayesian approach inspired by Merugu and Ghosh (2003). Instead of averaging the model's parameters, they directly "average the models" with the following KL-average estimator:

$$\hat{\theta}^{kl} = \arg\min_{\theta} \sum_{i=1}^{m} \mathrm{KL}\left(p(\cdot; \hat{\theta}_i^{erm}) \,\middle\|\, p(\cdot; \theta) \right). \qquad (12)$$

The minimization is performed via a bootstrap sample from the smaller models. This method has three main advantages. First, it is robust to reparameterizations of the model. Second, it is statistically optimal for the class of non-interactive optimization methods. (We show in the next section that this optimality bound does not apply to our $\hat{\theta}^{owa}$ estimator due to our semi-interactive setting.) Third, this method is general enough to work for any model, whereas our proposed OWA method works only for linear models. The main downside of the KL-average is that the minimization has a prohibitively high computational cost. Let n^{kl} be the size of the bootstrap sample. Then the original implementation's MSE shrinks as $O((nm)^{-1} + (nn^{kl})^{-1})$. This implies that the bootstrap procedure requires as many samples as the original problem to get a MSE that shrinks at the same rate as the averaging estimator. Han and Liu (2016) provide a method to reduce this rate to $O((nm)^{-1} + (n^2 n^{kl})^{-1})$ using control variates, but the procedure remains prohibitively expensive. Their experiments show the procedure scaling only to datasets of size $nm \approx 10^4$, whereas our experiments involve a dataset of size $nm \approx 10^8$.

Surprisingly, Zhang et al. (2013b) show that in the special case of kernel ridge regression, a reduction in bias is not needed to have the MSE of $\hat{\theta}^{ave}$ decay at the optimal sequential rate. By a careful choice of regularization parameter, they cause $\hat{\theta}_i^{erm}$ to have lower bias but higher variance, so that the final estimate of $\hat{\theta}^{ave}$ has both reduced bias and variance. This suggests that a merging procedure that reduces bias is not crucial to good performance if we set the regularization parameter correctly. Typically there is a narrow range of good regularization parameters, and finding a λ in this range is expensive computationally. We show experimentally in Section 5 that our method has significantly reduced sensitivity to λ. Therefore, it is computationally cheaper to find a good λ for our method than for the other methods discussed in this section.

3.2. Performance bounds

Performance bounds come in two flavors: statistical and information theoretic. On the statistical side, Liu and Ihler (2014) show that for any non-interactive distributed estimator $\hat{\theta}$, the quantity $\|\hat{\theta} - \hat{\theta}^{erm}\|^2$ decays as $\Omega(\gamma_{\theta^*}^2 \mathcal{I}_{\theta^*}^{-1} / n^2)$. Here γ_{θ^*} is the statistical curvature of the model and $\mathcal{I}_{\theta^*}$ is the Fisher information. Furthermore, they show that their KL-averaging estimator $\hat{\theta}^{kl}$ matches this bound. This bound is not relevant for our $\hat{\theta}^{owa}$ estimator because of our semi-interactive setting. A crucial assumption of Liu and Ihler's analysis is that the merge function not depend on the data.

Shamir (2014), Zhang et al. (2013a), and Garg et al. (2014) all provide information theoretic lower bounds on the sample complexity of non-interactive learning problems. As above, however, their results are not applicable in our semi-interactive setting. There is one information theoretic lower bound that does apply to us. Let the true parameter vector θ^* be k-sparse. That is, $\|\theta^*\|_0 \leq k$. Surprisingly, Braverman et al. (2015) show that the minimax optimal error rate for least squares regression requires $\Omega(m \cdot \min\{n, d\})$ bits of communication (independent of k) even in the fully interactive setting. This is important because sparsity does not reduce the amount of communication required, and this bound does apply in our setting.

4. ANALYSIS

In this section, we analyze the statistical performance of our $\hat{\theta}^{owa}$ estimator.

4.1. The Sub-Gaussian Tail Condition

The main condition of our analysis is that the estimation error has sub-Gaussian tails. As we will see, this is a mild condition that holds in most situations.

Definition 1. *We say that a random vector* $\mathbf{x}$ *is* sub-Gaussian *with variance proxy* σ^2 *if it obeys the following concentration bound:*

$$\Pr\left[\|\mathbf{x}\| < t\right] \geq 1 - \exp(-\sigma^2 t^2 / 2). \qquad (13)$$

Note in particular that if $\mathbf{x}$ is a Gaussian random vector with mean μ and covariance Σ, then $\Sigma^{1/2}(\mathbf{x}-\mu)$ is sub-Gaussian with $\sigma^2 = 1$. Sub-Gaussian random vectors have recently become an important tool in the analysis of high dimensional statistics. Vershynin (2012) provides an accessible tutorial of these results. We are now ready to state our condition.

The Sub-Gaussian Tail (SGT) Condition. *Let $\hat{\theta}$ be an estimator trained on n data points. Then the random vector*

$$\Delta_{\hat{\theta}} = \sqrt{n}\mathcal{I}^{1/2}\big(\theta^* - \hat{\theta}\big) \tag{14}$$

is sub-Gaussian for some σ^2. Above, $\mathcal{I}$ is the positive definite Fisher information matrix at the parameter vector's true value θ^*.

The SGT condition is mild and known to hold in many situations of interest. In the asymptotic regime when $n \to \infty$, very strong results are known. Theorem 7.5.2 of Lehmann (1999) is an elementary example that shows $\Delta_{\hat{\theta}}$ is an isotropic centered Gaussian (and hence sub-Gaussian). Lehman's theorem requires only that ℓ be three times differentiable and that the data points be i.i.d.. More sophisticated analyses show that the SGT Condition holds very generally in the asymptotic regime. For example Spokoiny (2012) shows that $\Delta_{\hat{\theta}}$ is normally distributed even when the data points are correlated.

Similar results hold in the non-asymptotic case $n < \infty$. The simplest results require that the data points also be sub-Gaussian. For example, Negahban et al. (2009) considers the case when the data points are sub-Gaussian, the likelihood satisfies a "restricted strong convexity condition," and the regularizer is decomposable. More recently, Sivakumar et al. (2015) showed the SGT Condition holds when the data are only sub-exponential. The strongest non-asymptotic results on the estimation error known to the authors are due to Spokoiny (2012). Spokoiny does not place any distributional assumption on the data, but shows that the SGT Condition is satisfied only up to $t < O(n)$.

We emphasize that the SGT condition is strictly more general than the conditions in previous work. For example, Zhang et al. (2012) requires that the parameter space Θ be bounded (in addition to other moment conditions). A bounded parameter space automatically implies that $\hat{\theta}_i^{erm}$ satisfies the SGT Condition because every bounded random variable is sub-Gaussian by definition.

4.2. Analyzing the ERM estimator

As a warm up, we show how the SGT assumption gives us a useful concentration bound for the error of $\hat{\theta}^{erm}$ (the ERM oracle estimator trained with all data on a single machine). In subsequent sections, we will measure the quality of the distributed estimators by comparing their error to the error of $\hat{\theta}^{erm}$.

Theorem 1. *Assume that $\hat{\theta}^{erm}$ satisfies the SGT condition. Let $t > 0$. Then with probability at least $1 - \exp(-t)$,*

$$\|\theta^* - \hat{\theta}^{erm}\| \leq \sigma\sqrt{\frac{v_t}{mn}}. \tag{15}$$

where

$$v_t = \operatorname{tr}\mathcal{I}^{-1} + 2\sqrt{\operatorname{tr}\left(\mathcal{I}^{-2}\right)t} + 2\|\mathcal{I}^{-1}\|t \tag{16}$$

Proof. By the definition of $\Delta_{\hat{\theta}^{erm}}$, we can rewrite

$$\|\theta^* - \hat{\theta}^{erm}\| = (mn)^{-1/2}\|\mathcal{I}^{-1/2}\Delta_{\hat{\theta}^{erm}}\|. \tag{17}$$

The result is then an immediate consequence of the following theorem due to Hsu et al. (2012).

Theorem 2. *Let $\mathbf{x}$ be a sub-Gaussian random variable with variance proxy σ^2 and A be a positive semidefinite matrix. Let $t > 0$. Then with probability at least $\exp(-t)$,*

$$\|A^{1/2}\mathbf{x}\|^2 \geq \sigma^2\left(\operatorname{tr}(A) + 2\sqrt{\operatorname{tr}A^2 t} + 2\|A\|t\right) \tag{18}$$

□

This theorem will be our main tool as we prove bounds on the distributed estimators.

4.3. Analyzing of the averaging estimator

We now provide a simple bound that shows that averaging improves the variance, but not the bias of an estimator. Similar bounds are well known (see Section 3.1), but our analysis has the following advantages: It requires fewer assumptions, has a simpler proof, and has an easy to interpret constant factor.

Theorem 3. *Assume that the $\hat{\theta}_i^{erm}$s satisfy the SGT Condition. Let $t > 0$. Then with probability at least $1 - \exp(-t)$,*

$$\|\theta^* - \hat{\theta}^{ave}\| \leq \|\theta^* - \mathbb{E}\hat{\theta}_i^{erm}\| + \sigma\sqrt{\frac{v_t}{mn}} \tag{19}$$

where v_t is defined as in Equation 16.

Proof. We have by the triangle inequality that

$$\|\theta^* - \hat{\theta}^{ave}\| \leq \|\theta^* - \mathbb{E}\hat{\theta}^{ave}\| + \|\mathbb{E}\hat{\theta}^{ave} - \hat{\theta}^{ave}\|. \quad (20)$$

The leftmost term in (20) is the estimator's bias. By the linearity of expectation, we have that

$$\mathbb{E}\hat{\theta}^{ave} = \mathbb{E}\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i^{erm} = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\hat{\theta}_i^{erm} = \mathbb{E}\hat{\theta}_i^{erm}, \quad (21)$$

and so $\|\theta^* - \mathbb{E}\hat{\theta}^{ave}\| = \|\theta^* - \mathbb{E}\hat{\theta}_i^{erm}\|$.

For the deviation term, we have that

$$\left\|\hat{\theta}^{ave} - \mathbb{E}\hat{\theta}^{ave}\right\| = \left\|\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i^{erm} - \mathbb{E}\hat{\theta}^{ave}\right\| \quad (22)$$

$$= \frac{1}{m}\left\|\sum_{i=1}^{m}\left(\hat{\theta}_i^{erm} - \mathbb{E}\hat{\theta}_i^{erm}\right)\right\| \quad (23)$$

$$= \frac{1}{m\sqrt{n}}\left\|\mathcal{I}^{-1/2}\sum_{i=1}^{m}\Delta_{\hat{\theta}_i^{erm}}\right\|. \quad (24)$$

Notice that each of the $\Delta_{\hat{\theta}_i^{erm}}$ are i.i.d. sub-Gaussian random vectors with variance proxy σ^2. The sum of m sub-Gaussians is also sub-Gaussian with variance proxy $m\sigma^2$. (See the Appendix for a proof.) Therefore, applying Theorem 2 gives the stated result. □

4.4. Analyzing the OWA variants

In this section, we provide concentration inequalities on the error of $\hat{\theta}^{owa}$. Our main result shows that the mean error $\mathbb{E}\|\hat{\theta}^{owa} - \theta^*\|$ decays at the optimal rate of $O(1/\sqrt{mn})$. To prove this result, we prove concentration inequalities similar to the previous sections. The concentration of $\|\hat{\theta}^{owa} - \theta^*\|$ is significantly looser than the single machine estimate.

Proof. We have that with probability at most $\exp(-tm)$,

$$\|\theta^* - \hat{\theta}^{owa}\| > \sqrt{\frac{v_t}{mn^{owa}}} + \sqrt{\left(\frac{q_{\text{hi}}}{q_{\text{lo}}}\right)\left(\frac{v_t}{n}\right)}. \quad (25)$$

A change of variables substitution gives us

$$\Pr\left[\|\theta^* - \hat{\theta}^{owa}\| > s\right] \leq \exp(-(\frac{s}{\sigma})^2\frac{mn}{\|\mathcal{I}^{-1}\|}) \quad (26)$$

One of the properties of expectations of non-negative random variables is that

$$\mathbb{E}\|\theta^* - \hat{\theta}^{erm}\| = \int_0^\infty \Pr\left[\|\theta^* - \hat{\theta}^{erm}\| > s\right] ds. \quad (27)$$

We will do a change of variables with

$$s = \sqrt{\frac{v_t}{mn^{owa}}} + \sqrt{\left(\frac{q_{\text{hi}}}{q_{\text{lo}}}\right)\left(\frac{v_t}{n}\right)} \quad (28)$$

□

Lemma 1. *Assume the $\hat{\theta}_i^{erm}$s satisfy the SGT condition. Let $t > 0$. Then with probability at least $1 - \exp(-t)$,*

$$\min_{\theta\in\hat{\Theta}^{owa}}\|\theta - \theta^*\| \leq \sigma\sqrt{\frac{v_{t/m}}{n}}. \quad (29)$$

where

$$v_{t/m} = \operatorname{tr}\mathcal{I}^{-1} + 2\sqrt{\operatorname{tr}\left(\mathcal{I}^{-2}\right)\frac{t}{m}} + 2\|\mathcal{I}^{-1}\|\frac{t}{m}. \quad (30)$$

Proof of Lemma 1. We have that

$$\Pr\left[\min_{\theta\in\hat{\Theta}^{owa}}\|\theta - \theta^*\| \geq \sqrt{\frac{v_t}{n}}\right] \quad (31)$$

$$\leq \Pr\left[\min_{i=[m]}\|\hat{\theta}_i^{erm} - \theta^*\| \geq \sqrt{\frac{v_t}{n}}\right] \quad (32)$$

$$= \Pr\left[\forall_{i\in[m]}\|\hat{\theta}_i^{erm} - \theta^*\| \geq \sqrt{\frac{v_t}{n}}\right] \quad (33)$$

$$= \left(\Pr\left[\|\hat{\theta}_1^{erm} - \theta^*\| \geq \sqrt{\frac{v_t}{n}}\right]\right)^m \quad (34)$$

$$= \left(\Pr\left[\|(n\mathcal{I})^{-1/2}\Delta_{\hat{\theta}_1^{erm}}\| \geq \sqrt{\frac{v_t}{n}}\right]\right)^m \quad (35)$$

$$= \left(\Pr\left[\|\mathcal{I}^{-1/2}\Delta_{\hat{\theta}_1^{erm}}\| \geq \sqrt{v_t}\right]\right)^m \quad (36)$$

$$\leq \exp(-tm). \quad (37)$$

□

The Hessian Condition. *For any vector θ satisfying $\|\theta\| \leq \|\theta^* - \hat{\theta}^{owa,full}\|$, we have that*

$$q_{lo}\|\theta\|^2 \leq \mathcal{L}(\theta^* + \theta) - \mathcal{L}(\theta) \leq q_{hi}\|\theta\|^2 \quad (38)$$

where $\mathcal{L}(\theta) = \sum_{(\mathbf{x},y)\in Z}\ell(y;\mathbf{x}^\mathsf{T}\theta) + \lambda r(\theta)$.

This condition is somewhat easier to understand in the asymptotic regime as $n \to \infty$. In this regime, we have that $\|\theta^* - \hat{\theta}^{owa,full}\| \to 0$, so (40) is equivalent to requiring that the condition number of the Hessian $\nabla^2\mathcal{L}(\theta^*)$ be $q_{\text{hi}}/q_{\text{lo}}$.

Theorem 4. *Assume the Hessian Condition and that the $\hat{\theta}_i^{erm}$s satisfy the SGT condition. Then we have that with probability at least $1 - \exp(-tm)$,*

$$\|\hat{\theta}^{owa,full} - \theta^*\| \leq \sqrt{\left(\frac{q_{hi}}{q_{lo}}\right)\left(\frac{v_t}{n}\right)} \quad (39)$$

Proof. By the Hessian condition, we have that

$$q_{\text{lo}}\|\hat{\theta}^{owa,full} - \theta^*\|^2 \leq \mathcal{L}(\hat{\theta}^{owa,full}) - \mathcal{L}(\theta^*) \quad (40)$$
$$\leq \mathcal{L}(\pi_{\hat{\Theta}^{owa}}\theta^*) - \mathcal{L}(\theta^*) \quad (41)$$
$$\leq q_{\text{hi}}\|\pi_{\hat{\Theta}^{owa}}\theta^*\|^2. \quad (42)$$

And so

$$\|\hat{\theta}^{owa,full} - \theta^*\| \leq \sqrt{\frac{q_{\text{hi}}}{q_{\text{lo}}}}\|\pi_{\hat{\Theta}^{owa}}\theta^*\|. \quad (43)$$

The result follows by Lemma 1. □

Corollary 1. *Assume the Hessian Condition and that the $\hat{\theta}_i^{erm}$s satisfy the SGT condition. Then we have that*

$$\mathbb{E}\|\hat{\theta}^{owa,full} - \theta^*\| \leq O\left(\sqrt{\left(\frac{q_{hi}}{q_{lo}}\right)\left(\frac{\operatorname{tr}\mathcal{I}^{-1}}{mn}\right)}\right) \quad (44)$$

Proof. We have that

$$\mathbb{E}\|\hat{\theta}^{owa} - \theta^*\| = \int_0^\infty \Pr\left[\|\hat{\theta}^{owa} - \theta^*\| > s\right] ds \quad (45)$$
$$\leq \int_0^\infty \exp(-tm)ds, \quad (46)$$

where the relationship between s and t is given by

$$s = \sqrt{\left(\frac{q_{\text{hi}}}{q_{\text{lo}}}\right)\left(\frac{v_t}{n}\right)}. \quad (47)$$

Substituting for v_t and rearranging gives us

$$s^2 mn\left(\frac{q_{\text{lo}}}{q_{\text{hi}}}\right) = \operatorname{tr}\mathcal{I}^{-1} + \sqrt{\operatorname{tr}\mathcal{I}^{-2}t} + \|\mathcal{I}^{-1}\|t \quad (48)$$
$$\leq (1+t)(\operatorname{tr}\mathcal{I}^{-1} + \sqrt{\operatorname{tr}\mathcal{I}^{-2}} + \|\mathcal{I}^{-1}\|) \quad (49)$$
$$\leq 3(1+t)\operatorname{tr}\mathcal{I}^{-1}. \quad (50)$$

And so,

$$t \geq \left(\frac{s^2 mn}{3\operatorname{tr}\mathcal{I}^{-1}}\right)\left(\frac{q_{\text{lo}}}{q_{\text{hi}}}\right) - 1 \quad (51)$$

Substituting (53) into (48) and solving the integral gives the result. □

Theorem 5. *Assume the Hessian Condition and that both the $\hat{\theta}_i^{erm}$s and α satisfy the SGT condition. Then, for all $t > 0$, with probability at least $1 - \exp(-t)$,*

$$\|\theta^* - \hat{\theta}^{owa}\| \leq \sqrt{\frac{v_t}{mn^{owa}}} + \sqrt{\left(\frac{q_{hi}}{q_{lo}}\right)\left(\frac{v_{t/m}}{n}\right)} \quad (52)$$

Proof of Theorem 4. We have by the triangle inequality that

$$\|\hat{\theta}^{owa} - \theta^*\| \leq \|\hat{\theta}^{owa} - \hat{\theta}^{owa,full}\| + \|\hat{\theta}^{owa,full} - \theta^*\|. \quad (53)$$

Theorem 5 bounds the left term, and the SGT condition combined with Theorem 2 bounds the right term. □

5. EXPERIMENTS

We evaluate OWA on two logistic regression tasks. The first task uses synthetic data. The second task uses real world ad-click data from the Tencent search engine. In each experiment, we compare our $\hat{\theta}^{owa}$ estimator with four baseline estimators: the naive estimator using the data from only a single machine $\hat{\theta}_i^{erm}$; the averaging estimator $\hat{\theta}^{ave}$; the bootstrap estimator $\hat{\theta}^{boot}$; and the oracle estimator of all data trained on a single machine $\hat{\theta}^{erm}$. The $\hat{\theta}^{boot}$ estimator has a parameter r that needs to be tuned. In all experiments we evaluate $\hat{\theta}^{boot}$ with $r \in \{0.005, 0.01, 0.02, 0.04, 0.1, 0.2\}$, which is a set recommended in the original paper (Zhang et al., 2012), and then report only the value of r with highest true likelihood. Thus we are reporting an overly optimistic estimate of the performance of $\hat{\theta}^{boot}$, and as we shall see our estimator $\hat{\theta}^{owa}$ still tends to perform better.

5.1. Synthetic Data

We generate the data according to the following sparse logistic regression model. Each component of the true parameter vector θ^* is sampled i.i.d. from a spike and slab distribution. With probability 0.9, the component is 0; with probability 0.1, the component is sampled from a standard normal distribution. The data points are then sampled as

$$\mathbf{x}_i \sim \mathcal{N}(0, I), \quad y_i = \left(1 + \exp(-\mathbf{x}_i^{\mathsf{T}}\theta^*)\right)^{-1}. \quad (54)$$

In all experiments, we use the L1 regularizer to induce sparsity in our estimates of θ^*. Our estimators will be biased because the model is misspecified. The true model for L1 regularization has a Laplace prior on θ^*.

Our first experiment shows the sensitivity of the estimators to the strength of regularization λ. In this experiment, λ was allowed to vary from 10^{-4} to 10^4. For each value of λ, we randomly generated 50 datasets (with $n = 1000$) and then calculated the corresponding estimators. Our $\hat{\theta}^{owa}$ estimator was trained with $n^{owa} = 128$. Figure 2 shows the average of the results for three choices of m and d. Our $\hat{\theta}^{owa}$ estimator is significantly less sensitive to the choice of λ than the other distributed estimators. Surprisingly, $\hat{\theta}^{owa}$ even

Figure 2. OWA is robust to the regularization strength. Surprisingly, additional regularization introduced by OWA lets it outperform the oracle estimator $\hat{\theta}^{erm}$ in some cases. Our theory states that as $m \to d$, $\hat{\theta}^{owa} \to \hat{\theta}^{erm}$. This is confirmed in the middle experiment. In the leftmost experiment, $m < d$, but $\hat{\theta}^{owa}$ still behaves similarly to $\hat{\theta}^{erm}$. In the rightmost experiment, $\hat{\theta}^{owa}$ has similar performance as $\hat{\theta}^{ave}$ and $\hat{\theta}^{boot}$ but is less sensitive to λ.

Figure 3. $\hat{\theta}^{owa}$ scales well with the number of machines. Surprisingly, it outperforms the oracle estimator trained on all of the data $\hat{\theta}^{erm}$ in some situations. This is likely due to the additional regularization introduced by the OWA algorithm, as seen in Figure 2.

outperforms the oracle $\hat{\theta}^{erm}$ in some regimes. This is likely due to additional regularization induced by the approximation in Equation 8.

Our second experiment shows how the parallel algorithms scale as the number of machines m increases. We fix $n = 1000$ data points per machine, so the size of the dataset mn grows as we add more machines. This simulates the typical "big data" regime where data is abundant, but processing resources are scarce. Each machine independently uses cross validation to select the λ that best fits the data locally. There are three advantages to this model selection procedure. First, there is no additional communication because model selection is a completely local task. Second, existing optimizers have built-in model selection routines which make the process easy to implement. We used the default model selection procedure from Python's SciKit-Learn (Pedregosa et al., 2011). Third, the data may be best fit using different regularization strengths for each machine. The results are shown in Figure 3. The performance of $\hat{\theta}^{owa}$ scales much better than $\hat{\theta}^{ave}$ and $\hat{\theta}^{boot}$. Because of the regularization induced by Equation 8 observed in the previous experiment, $\hat{\theta}^{owa}$ even scales better than $\hat{\theta}^{erm}$ in some regimes.

5.2. Real World Advertising Data

We now evaluate our estimator on real world data from the KDD 2012 Cup (Niu et al., 2012). The goal is to predict whether a user will click on an ad from the Tencent internet search engine. This dataset was previously used to evaluate the performance of the bootstrap average estimator (Zhang et al., 2012). This dataset is too large to fit on a single machine, so we must use distributed estimators, and we do not provide results of the oracle estimator $\hat{\theta}^{erm}$ in our figures. There are 235,582,879 distinct data points, each of dimension 741,725. The data points are sparse, so we use the L1 norm to encourage sparsity in our final solution. The regularization strength was set using cross validation in the same manner as for the synthetic data. For each test, we split the data into 80 percent training data and 20 percent test data. The training data is further subdivided into 128 partitions, one for each of the machines used. It took about 1 day to train the local model on each machine.

Our first experiment tests the sensitivity of the n^{owa} parameter on large datasets. We fix $m = 128$, and allow n^{owa} to vary from 2^0 to 2^{20}, which is approximately the size of the full dataset. We repeated the experiment 50 times, each time using a different randomly selected set Z^{owa} for the second optimization. Figure 4 shows the results. Our $\hat{\theta}^{owa}$ estimator has lower loss than the $\hat{\theta}^{ave}$ using only 16 data points per machine (approximately 4×10^{-8} percent of the full training set) and $\hat{\theta}^{owa}$ has converged to its final loss value with only 1024 data points per machine (approximately 2.7×10^{-6} percent of the full training set). This justifies our claim that only a small number of data points are needed for the second round of optimization, and so the communication complexity of $\hat{\theta}^{owa}$ is essentially the same as $\hat{\theta}^{ave}$. The computation was also very fast. Even when $n^{owa} = 2^{20}$ computing the merged model took only several minutes, which is negligible compared to the approximately 1 day it took to train the models on the individual machines.

Figure 4. Relatively few data points are needed in the second round of optimization for $\hat{\theta}^{owa}$ to converge.

Figure 5. Performance of the parallel estimators on advertising data as the number of machines m increases.

Our last experiment shows the performance as we scale the number of machines m. The results are shown in Figure 5. Here, our $\hat{\theta}^{owa}$ performs especially well in the low m setting. For large m, $\hat{\theta}^{owa}$ continues to slightly outperform $\hat{\theta}^{boot}$ without the need for an expensive model selection procedure to determine the r parameter.

6. CONCLUSION

We introduced a new distributed estimation algorithm called OWA. OWA has the speed advantages of non-interactive distributed estimators, but has better accuracy due to a (cheap) second round of optimization. Unlike other algorithms, OWA does not require expensive hyperparameter tuning. Furthermore, our analysis is more general than the analysis of similar algo-

rithms. As part of our analysis, we also provided a more general analysis of the averaging estimator.

References

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*, 2015.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. *arXiv preprint arXiv:1506.07216*, 2015.

Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

Jun Han and Qiang Liu. Bootstrap model aggregation for distributed statistical learning. *arXiv preprint arXiv:1607.01036*, 2016.

Daniel Hsu, Sham M Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52): 1–6, 2012.

Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E Taylor. Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*, 2015.

Erich Leo Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 1999.

Qiang Liu and Alexander T Ihler. Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems*, pages 1098–1106, 2014.

Ryan McDonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems*, pages 1231–1239, 2009.

Srujana Merugu and Joydeep Ghosh. Privacy-preserving distributed clustering using generative models. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 211–218. IEEE, 2003.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

Yanzhi Niu, Yi Wang, Gordon Sun, Aden Yue, Brian Dalessandro, Claudia Perlich, and Ben Hamner. The tencent dataset and kdd-cup'12. 2012.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Jonathan D Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference*, 5(4):379–404, 2016.

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

Vidyashankar Sivakumar, Arindam Banerjee, and Pradeep K Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. In *Advances in Neural Information Processing Systems*, pages 2206–2214, 2015.

Vladimir Spokoiny. Parametric estimation. finite sample theory. *The Annals of Statistics*, 40(6):2877–2909, 2012.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5. 2012.

Yuchen Zhang, Martin J Wainwright, and John C Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013a.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Divide and conquer kernel ridge regression. In *COLT*, 2013b.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

Appendix A: Proof of Theorem 1

Appendix

Lemma 2. *Let $\mathbf{x}_1, ..., \mathbf{x}_m$ be a sequence of i.i.d. subgaussian random vectors with variance proxy σ^2. Then the random vector $\sum_{i=1}^m \mathbf{x}_i$ is subgaussian with variance proxy $m\sigma^2$.*

Proof. The lemma follows from the following straightforward calculation of the moment generating function.

$$\mathbb{E}\exp(\alpha^\mathsf{T} \sum_{i=1}^m \mathbf{x}_i) = \mathbb{E}\prod_{i=1}^m \exp(\alpha^\mathsf{T} \mathbf{x}_i) \tag{55}$$

$$= \prod_{i=1}^m \mathbb{E}\exp(\alpha^\mathsf{T} \mathbf{x}_i) \tag{56}$$

$$\leq \prod_{i=1}^m \exp(\|\alpha\|^2 \sigma^2/2) \tag{57}$$

$$= \exp(\|\alpha\|^2 m\sigma^2/2) \tag{58}$$

□