Journal of Machine Learning Research 14 (2013) 3321-3363 Submitted 3/13; Published 11/13

Communication-Efficient Algorithms for Statistical Optimization

Yuchen Zhang YUCZHANG@EECS.BERKELEY.EDU
John C. Duchi JDUCHI@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720 USA

Martin J. Wainwright WAINWRIG@STAT.BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720 USA

Editor: Francis Bach

Abstract

We analyze two communication-efficient algorithms for distributed optimization in statistical settings involving large-scale data sets. The first algorithm is a standard averaging method that distributes the N data samples evenly to m machines, performs separate minimization on each subset, and then averages the estimates. We provide a sharp analysis of this average mixture algorithm, showing that under a reasonable set of conditions, the combined parameter achieves mean-squared error (MSE) that decays as $O(N^{-1}+(N/m)^{-2})$. Whenever $m \leq \sqrt{N}$, this guarantee matches the best possible rate achievable by a centralized algorithm having access to all N samples. The second algorithm is a novel method, based on an appropriate form of bootstrap subsampling. Requiring only a single round of communication, it has mean-squared error that decays as $O(N^{-1}+(N/m)^{-3})$, and so is more robust to the amount of parallelization. In addition, we show that a stochastic gradient-based method attains mean-squared error decaying as $O(N^{-1}+(N/m)^{-3/2})$, easing computation at the expense of a potentially slower MSE rate. We also provide an experimental evaluation of our methods, investigating their performance both on simulated data and on a large-scale regression problem from the internet search domain. In particular, we show that our methods can be used to efficiently solve an advertisement prediction problem from the Chinese SoSo Search Engine, which involves logistic regression with $N \approx 2.4 \times 10^8$ samples and $d \approx 740{,}000$ covariates.

Keywords: distributed learning, stochastic optimization, averaging, subsampling

1. Introduction

Many procedures for statistical estimation are based on a form of (regularized) empirical risk minimization, meaning that a parameter of interest is estimated by minimizing an objective function defined by the average of a loss function over the data. Given the current explosion in the size and amount of data available in statistical studies, a central challenge is to design efficient algorithms for solving large-scale problem instances. In a centralized setting, there are many procedures for solving empirical risk minimization problems, among them standard convex programming approaches (e.g., Boyd and Vandenberghe, 2004) as well as stochastic approximation and optimization algorithms (Robbins and Monro, 1951; Hazan et al., 2006; Nemirovski et al., 2009). When the size of the data set becomes extremely large, however, it may be infeasible to store all of the data on a

single computer, or at least to keep the data in memory. Accordingly, the focus of this paper is the study of some distributed and communication-efficient procedures for empirical risk minimization.

Recent years have witnessed a flurry of research on distributed approaches to solving very large-scale statistical optimization problems. Although we cannot survey the literature adequately—the papers Nedić and Ozdaglar (2009), Ram et al. (2010), Johansson et al. (2009), Duchi et al. (2012a), Dekel et al. (2012), Agarwal and Duchi (2011), Recht et al. (2011), Duchi et al. (2012b) and references therein contain a sample of relevant work—we touch on a few important themes here. It can be difficult within a purely optimization-theoretic setting to show explicit benefits arising from distributed computation. In statistical settings, however, distributed computation can lead to gains in computational efficiency, as shown by a number of authors (Agarwal and Duchi, 2011; Dekel et al., 2012; Recht et al., 2011; Duchi et al., 2012b). Within the family of distributed algorithms, there can be significant differences in communication complexity: different computers must be synchronized, and when the dimensionality of the data is high, communication can be prohibitively expensive. It is thus interesting to study distributed estimation algorithms that require fairly limited synchronization and communication while still enjoying the greater statistical accuracy that is usually associated with a larger data set.

With this context, perhaps the simplest algorithm for distributed statistical estimation is what we term the *average mixture* (AVGM) algorithm. It is an appealingly simple method: given m different machines and a data set of size N, first assign to each machine a (distinct) data set of size $n = N/m$, then have each machine i compute the empirical minimizer θ_i on its fraction of the data, and finally average all the parameter estimates θ_i across the machines. This approach has been studied for some classification and estimation problems by Mann et al. (2009) and McDonald et al. (2010), as well as for certain stochastic approximation methods by Zinkevich et al. (2010). Given an empirical risk minimization algorithm that works on one machine, the procedure is straightforward to implement and is extremely communication efficient, requiring only a single round of communication. It is also relatively robust to possible failures in a subset of machines and/or differences in speeds, since there is no repeated synchronization. When the local estimators are all unbiased, it is clear that the the AVGM procedure will yield an estimate that is essentially as good as that of an estimator based on all N samples. However, many estimators used in practice are biased, and so it is natural to ask whether the method has any guarantees in a more general setting. To the best of our knowledge, however, no work has shown rigorously that the AVGM procedure generally has greater efficiency than the naive approach of using $n = N/m$ samples on a single machine.

This paper makes three main contributions. First, in Section 3, we provide a sharp analysis of the AVGM algorithm, showing that under a reasonable set of conditions on the population risk, it can indeed achieve substantially better rates than the naive approach. More concretely, we provide bounds on the mean-squared error (MSE) that decay as $O((nm)^{-1} + n^{-2})$. Whenever the number of machines m is less than the number of samples n per machine, this guarantee matches the best possible rate achievable by a centralized algorithm having access to all $N = nm$ samples. In the special case of optimizing log likelihoods, the pre-factor in our bound involves the trace of the Fisher information, a quantity well-known to control the fundamental limits of statistical estimation. We also show how the result extends to stochastic programming approaches, exhibiting a stochastic gradient-descent based procedure that also attains convergence rates scaling as $O((nm)^{-1})$, but with slightly worse dependence on different problem-specific parameters.

Our second contribution is to develop a novel extension of simple averaging. It is based on an appropriate form of resampling (Efron and Tibshirani, 1993; Hall, 1992; Politis et al., 1999), which

we refer to as the *subsampled average mixture* (SAVGM) approach. At a high level, the SAVGM algorithm distributes samples evenly among m processors or computers as before, but instead of simply returning the empirical minimizer, each processor further subsamples its own data set in order to estimate the bias of its own estimate, and returns a subsample-corrected estimate. We establish that the SAVGM algorithm has mean-squared error decaying as $O(m^{-1}n^{-1}+n^{-3})$. As long as $m < n^2$, the subsampled method again matches the centralized gold standard in the first-order term, and has a second-order term smaller than the standard averaging approach.

Our third contribution is to perform a detailed empirical evaluation of both the AVGM and SAVGM procedures, which we present in Sections 4 and 5. Using simulated data from normal and non-normal regression models, we explore the conditions under which the SAVGM algorithm yields better performance than the AVGM algorithm; in addition, we study the performance of both methods relative to an oracle baseline that uses all N samples. We also study the sensitivity of the algorithms to the number of splits m of the data, and in the SAVGM case, we investigate the sensitivity of the method to the amount of resampling. These simulations show that both AVGM and SAVGM have favourable performance, even when compared to the unattainable "gold standard" procedure that has access to all N samples. In Section 5, we complement our simulation experiments with a large logistic regression experiment that arises from the problem of predicting whether a user of a search engine will click on an advertisement. This experiment is large enough—involving $N \approx 2.4 \times 10^8$ samples in $d \approx 740,000$ dimensions with a storage size of approximately 55 gigabytes—that it is difficult to solve efficiently on one machine. Consequently, a distributed approach is essential to take full advantage of this data set. Our experiments on this problem show that SAVGM—with the resampling and correction it provides—gives substantial performance benefits over naive solutions as well as the averaging algorithm AVGM.

2. Background and Problem Set-up

We begin by setting up our decision-theoretic framework for empirical risk minimization, after which we describe our algorithms and the assumptions we require for our main theoretical results.

2.1 Empirical Risk Minimization

Let $\{f(\cdot;x),\, x \in \mathcal{X}\}$ be a collection of real-valued and convex loss functions, each defined on a set containing the convex set $\Theta \subseteq \mathbb{R}^d$. Let P be a probability distribution over the sample space $\mathcal{X}$. Assuming that each function $x \mapsto f(\theta;x)$ is P-integrable, the *population risk* $F_0 : \Theta \to \mathbb{R}$ is given by

$$F_0(\theta) := \mathbb{E}_P[f(\theta;X)] = \int_{\mathcal{X}} f(\theta;x)dP(x).$$

Our goal is to estimate the parameter vector minimizing the population risk, namely the quantity

$$\theta^* := \operatorname*{argmin}_{\theta\in\Theta} F_0(\theta) = \operatorname*{argmin}_{\theta\in\Theta} \int_{\mathcal{X}} f(\theta;x)dP(x),$$

which we assume to be unique. In practice, the population distribution P is unknown to us, but we have access to a collection S of samples from the distribution P. Empirical risk minimization is based on estimating θ^* by solving the optimization problem

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta} \Big\{ \frac{1}{|S|} \sum_{x\in S} f(\theta;x) \Big\}.$$

Throughout the paper, we impose some regularity conditions on the parameter space, the risk function F_0, and the instantaneous loss functions $f(\cdot;x) : \Theta \to \mathbb{R}$. These conditions are standard in classical statistical analysis of M-estimators (e.g., Lehmann and Casella, 1998; Keener, 2010). Our first assumption deals with the relationship of the parameter space to the optimal parameter θ^*.

Assumption 1 (Parameters) *The parameter space $\Theta \subset \mathbb{R}^d$ is a compact convex set, with $\theta^* \in \operatorname{int}\Theta$ and ℓ_2-radius $R = \max_{\theta\in\Theta} \|\theta - \theta^*\|_2$.*

In addition, the risk function is required to have some amount of curvature. We formalize this notion in terms of the Hessian of F_0:

Assumption 2 (Local strong convexity) *The population risk is twice differentiable, and there exists a parameter $\lambda > 0$ such that $\nabla^2 F_0(\theta^*) \succeq \lambda I_{d\times d}$.*

Here $\nabla^2 F_0(\theta)$ denotes the $d \times d$ Hessian matrix of the population objective F_0 evaluated at θ, and we use $\succeq$ to denote the positive semidefinite ordering (i.e., $A \succeq B$ means that $A - B$ is positive semidefinite.) This local condition is milder than a global strong convexity condition and is required to hold only for the population risk F_0 evaluated at θ^*. It is worth observing that some type of curvature of the risk is required for any method to consistently estimate the parameters θ^*.

2.2 Averaging Methods

Consider a data set consisting of $N = mn$ samples, drawn i.i.d. according to the distribution P. In the distributed setting, we divide this N-sample data set evenly and uniformly at random among a total of m processors. (For simplicity, we have assumed the total number of samples is a multiple of m.) For $i = 1, \ldots, m$, we let $S_{1,i}$ denote the data set assigned to processor i; by construction, it is a collection of n samples drawn i.i.d. according to P, and the samples in subsets $S_{1,i}$ and $S_{1,j}$ are independent for $i \neq j$. In addition, for each processor i we define the (local) empirical distribution $P_{1,i}$ and empirical objective $F_{1,i}$ via

$$P_{1,i} := \frac{1}{|S_1|} \sum_{x\in S_{1,i}} \delta_x, \quad \text{and} \quad F_{1,i}(\theta) := \frac{1}{|S_{1,i}|} \sum_{x\in S_{1,i}} f(\theta; x).$$

With this notation, the AVGM algorithm is very simple to describe.

2.2.1 AVERAGE MIXTURE ALGORITHM

(1) For each $i \in \{1, \ldots, m\}$, processor i uses its local data set $S_{1,i}$ to compute the local empirical minimizer

$$\theta_{1,i} \in \operatorname*{argmin}_{\theta\in\Theta} \Big\{ \underbrace{\frac{1}{|S_{1,i}|} \sum_{x\in S_{1,i}} f(\theta; x)}_{F_{1,i}(\theta)} \Big\}. \tag{1}$$

(2) These m local estimates are then averaged together—that is, we compute

$$\bar{\theta}_1 = \frac{1}{m} \sum_{i=1}^{m} \theta_{1,i}. \tag{2}$$

The subsampled average mixture (SAVGM) algorithm is based on an additional level of sampling on top of the first, involving a fixed subsampling rate $r \in [0,1]$. It consists of the following additional steps:

2.2.2 SUBSAMPLED AVERAGE MIXTURE ALGORITHM

(1) Each processor i draws a subset $S_{2,i}$ of size $\lceil rn \rceil$ by sampling uniformly at random without replacement from its local data set $S_{1,i}$.

(2) Each processor i computes both the local empirical minimizers $\theta_{1,i}$ from Equation (1) and the empirical minimizer

$$\theta_{2,i} \in \operatorname*{argmin}_{\theta \in \Theta} \Big\{ \underbrace{\frac{1}{|S_{2,i}|} \sum_{x \in S_{2,i}} f(\theta; x)}_{F_{2,i}(\theta)} \Big\}.$$

(3) In addition to the previous average (2), the SAVGM algorithm computes the bootstrap average $\overline{\theta}_2 := \frac{1}{m} \sum_{i=1}^m \theta_{2,i}$, and then returns the weighted combination

$$\overline{\theta}_{\text{SAVGM}} := \frac{\overline{\theta}_1 - r\overline{\theta}_2}{1-r}. \tag{3}$$

The intuition for the weighted estimator (3) is similar to that for standard bias correction procedures using the bootstrap or subsampling (Efron and Tibshirani, 1993; Hall, 1992; Politis et al., 1999). Roughly speaking, if $b_0 = \theta^* - \theta_1$ is the bias of the first estimator, then we may approximate b_0 by the subsampled estimate of bias $b_1 = \theta^* - \theta_2$. Then, we use the fact that $b_1 \approx b_0/r$ to argue that $\theta^* \approx (\theta_1 - r\theta_2)/(1-r)$. The re-normalization enforces that the relative "weights" of $\overline{\theta}_1$ and $\overline{\theta}_2$ sum to 1.

The goal of this paper is to understand under what conditions—and in what sense—the estimators (2) and (3) approach the *oracle performance*, by which we mean the error of a centralized risk minimization procedure that is given access to all $N = nm$ samples.

2.2.3 NOTATION

Before continuing, we define the remainder of our notation. We use ℓ_2 to denote the usual Euclidean norm $\|\theta\|_2 = (\sum_{j=1}^d \theta_j^2)^{\frac{1}{2}}$. The ℓ_2-operator norm of a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ is its maximum singular value, defined by

$$|||A|||_2 := \sup_{v \in \mathbb{R}^{d_2}, \|v\|_2 \leq 1} \|Av\|_2.$$

A convex function F is λ-strongly convex on a set $U \subseteq \mathbb{R}^d$ if for arbitrary $u, v \in U$ we have

$$F(u) \geq F(v) + \langle \nabla F(v), u - v \rangle + \frac{\lambda}{2} \|u - v\|_2^2.$$

(If F is not differentiable, we may replace ∇F with any subgradient of F.) We let $\otimes$ denote the Kronecker product, and for a pair of vectors u, v, we define the outer product $u \otimes v = uv^\top$. For a

three-times differentiable function F, we denote the third derivative tensor by $\nabla^3 F$, so that for each $u \in \operatorname{dom} F$ the operator $\nabla^3 F(u) : \mathbb{R}^{d \times d} \to \mathbb{R}^d$ is linear and satisfies the relation

$$\left[\nabla^3 F(u)(v \otimes v)\right]_i = \sum_{j,k=1}^{d} \left(\frac{\partial^3}{\partial u_i \partial u_j \partial u_k} F(u) \right) v_j v_k.$$

We denote the indicator function of an event $\mathcal{E}$ by $1_{(\mathcal{E})}$, which is 1 if $\mathcal{E}$ is true and 0 otherwise.

3. Theoretical Results

Having described the AVGM and SAVGM algorithms, we now turn to statements of our main theorems on their statistical properties, along with some consequences and comparison to past work.

3.1 Smoothness Conditions

In addition to our previously stated assumptions on the population risk, we require regularity conditions on the empirical risk functions. It is simplest to state these in terms of the functions $\theta \mapsto f(\theta; x)$, and we note that, as with Assumption 2, we require these to hold only locally around the optimal point θ^*, in particular within some Euclidean ball $U = \{\theta \in \mathbb{R}^d \mid \|\theta^* - \theta\|_2 \le \rho\} \subseteq \Theta$ of radius $\rho > 0$.

Assumption 3 (Smoothness) *There are finite constants G, H such that the first and the second partial derivatives of f exist and satisfy the bounds*

$$\mathbb{E}[\|\nabla f(\theta; X)\|_2^8] \le G^8 \quad \text{and} \quad \mathbb{E}[\left|\left|\left|\nabla^2 f(\theta; X) - \nabla^2 F_0(\theta)\right|\right|\right|_2^8] \le H^8 \quad \text{for all } \theta \in U.$$

In addition, for any $x \in \mathcal{X}$, the Hessian matrix $\nabla^2 f(\theta; x)$ is $L(x)$-Lipschitz continuous, meaning that

$$\left|\left|\left|\nabla^2 f(\theta'; x) - \nabla^2 f(\theta; x)\right|\right|\right|_2 \le L(x) \left\|\theta' - \theta\right\|_2 \quad \text{for all } \theta, \theta' \in U. \tag{4}$$

We require that $\mathbb{E}[L(X)^8] \le L^8$ and $\mathbb{E}[(L(X) - \mathbb{E}[L(X)])^8] \le L^8$ for some finite constant L.

It is an important insight of our analysis that some type of smoothness condition on the Hessian matrix, as in the Lipschitz condition (4), is *essential* in order for simple averaging methods to work. This necessity is illustrated by the following example:

Example 1 (Necessity of Hessian conditions) *Let X be a Bernoulli variable with parameter $\frac{1}{2}$, and consider the loss function*

$$f(\theta; x) = \begin{cases} \theta^2 - \theta & \text{if } x = 0 \\ \theta^2 1_{(\theta \le 0)} + \theta & \text{if } x = 1, \end{cases} \tag{5}$$

where $1_{(\theta \le 0)}$ is the indicator of the event $\{\theta \le 0\}$. The associated population risk is $F_0(\theta) = \frac{1}{2}(\theta^2 + \theta^2 1_{(\theta \le 0)})$. Since $|F_0'(w) - F_0'(v)| \le 2|w - v|$, the population risk is strongly convex and smooth, but it has discontinuous second derivative. The unique minimizer of the population risk is $\theta^* = 0$, and by an asymptotic expansion given in Appendix A, it can be shown that $\mathbb{E}[\theta_{1,i}] = \Omega(n^{-\frac{1}{2}})$. Consequently, the bias of $\overline{\theta}_1$ is $\Omega(n^{-\frac{1}{2}})$, and the AVGM algorithm using $N = mn$ observations must suffer mean squared error $\mathbb{E}[(\overline{\theta}_1 - \theta^*)^2] = \Omega(n^{-1})$.

The previous example establishes the necessity of a smoothness condition. However, in a certain sense, it is a pathological case: both the smoothness condition given in Assumption 3 and the local strong convexity condition given in Assumption 2 are relatively innocuous for practical problems. For instance, both conditions will hold for standard forms of regression, such as linear and logistic, as long as the *population* data covariance matrix is not rank deficient and the data has suitable moments. Moreover, in the linear regression case, one has $L = 0$.

3.2 Bounds for Simple Averaging

We now turn to our first theorem that provides guarantees on the statistical error associated with the AVGM procedure. We recall that θ^* denotes the minimizer of the population objective function F_0, and that for each $i \in \{1, \ldots, m\}$, we use S_i to denote a data set of n independent samples. For each i, we use $\theta_i \in \operatorname{argmin}_{\theta \in \Theta}\{\frac{1}{n}\sum_{x \in S_i} f(\theta; x)\}$ to denote a minimizer of the empirical risk for the data set S_i, and we define the averaged vector $\overline{\theta} = \frac{1}{m}\sum_{i=1}^m \theta_i$. The following result bounds the mean-squared error between this averaged estimate and the minimizer θ^* of the population risk.

Theorem 1 *Under Assumptions 1 through 3, the mean-squared error is upper bounded as*

$$\begin{aligned} \mathbb{E}\left[\left\|\overline{\theta} - \theta^*\right\|_2^2\right] &\leq \frac{2}{nm}\mathbb{E}\left[\left\|\nabla^2 F_0(\theta^*)^{-1}\nabla f(\theta^*; X)\right\|_2^2\right] \\ &\quad + \frac{c}{\lambda^2 n^2}\left(H^2 \log d + \frac{L^2 G^2}{\lambda^2}\right)\mathbb{E}\left[\left\|\nabla^2 F_0(\theta^*)^{-1}\nabla f(\theta^*; X)\right\|_2^2\right] \\ &\quad + O(m^{-1}n^{-2}) + O(n^{-3}), \end{aligned} \tag{6}$$

where c is a numerical constant.

A slightly weaker corollary of Theorem 1 makes it easier to parse. In particular, note that

$$\left\|\nabla^2 F_0(\theta^*)^{-1}\nabla f(\theta^*; x)\right\|_2 \overset{(i)}{\leq} \left|\left|\left|\nabla^2 F_0(\theta^*)^{-1}\right|\right|\right|_2 \|\nabla f(\theta^*; x)\|_2 \overset{(ii)}{\leq} \frac{1}{\lambda}\|\nabla f(\theta^*; x)\|_2, \tag{7}$$

where step (i) follows from the inequality $|||Ax|||_2 \leq |||A|||\,\|x\|_2$, valid for any matrix A and vector x; and step (ii) follows from Assumption 2. In addition, Assumption 3 implies $\mathbb{E}[\|\nabla f(\theta^*; X)\|_2^2] \leq G^2$, and putting together the pieces, we have established the following.

Corollary 2 *Under the same conditions as Theorem 1,*

$$\mathbb{E}\left[\left\|\overline{\theta} - \theta^*\right\|_2^2\right] \leq \frac{2G^2}{\lambda^2 nm} + \frac{cG^2}{\lambda^4 n^2}\left(H^2 \log d + \frac{L^2 G^2}{\lambda^2}\right) + O(m^{-1}n^{-2}) + O(n^{-3}). \tag{8}$$

This upper bound shows that the leading term decays proportionally to $(nm)^{-1}$, with the pre-factor depending inversely on the strong convexity constant λ and growing proportionally with the bound G on the loss gradient. Although easily interpretable, the upper bound (8) can be loose, since it is based on the relatively weak series of bounds (7).

The leading term in our original upper bound (6) involves the product of the gradient $\nabla f(\theta^*; X)$ with the inverse Hessian. In many statistical settings, including the problem of linear regression, the effect of this matrix-vector multiplication is to perform some type of standardization. When the

loss $f(\cdot;x) : \Theta \to \mathbb{R}$ is actually the negative log-likelihood $\ell(x \mid \theta)$ for a parametric family of models $\{P_\theta\}$, we can make this intuition precise. In particular, under suitable regularity conditions (e.g., Lehmann and Casella, 1998, Chapter 6), we can define the Fisher information matrix

$$I(\theta^*) := \mathbb{E}\left[\nabla\ell(X \mid \theta^*)\nabla\ell(X \mid \theta^*)^\top\right] = \mathbb{E}[\nabla^2\ell(X \mid \theta^*)].$$

Recalling that $N = mn$ is the total number of samples available, let us define the neighbourhood $B_2(\theta, t) := \{\theta' \in \mathbb{R}^d : \|\theta' - \theta\|_2 \le t\}$. Then under our assumptions, the Hájek-Le Cam minimax theorem (van der Vaart, 1998, Theorem 8.11) guarantees for *any estimator* $\widehat{\theta}_N$ based on N samples that

$$\lim_{c\to\infty} \liminf_{N\to\infty} \sup_{\theta \in B_2(\theta^*, c/\sqrt{N})} N\,\mathbb{E}_\theta\left[\left\|\widehat{\theta}_N - \theta\right\|_2^2\right] \ge \operatorname{tr}(I(\theta^*)^{-1}).$$

In connection with Theorem 1, we obtain:

Corollary 3 *In addition to the conditions of Theorem 1, suppose that the loss functions $f(\cdot;x)$ are the negative log-likelihood $\ell(x \mid \theta)$ for a parametric family $\{P_\theta, \theta \in \Theta\}$. Then the mean-squared error is upper bounded as*

$$\mathbb{E}\left[\left\|\overline{\theta}_1 - \theta^*\right\|_2^2\right] \le \frac{2}{N}\operatorname{tr}(I(\theta^*)^{-1}) + \frac{cm^2\operatorname{tr}(I(\theta^*)^{-1})}{\lambda^2 N^2}\left(H^2\log d + \frac{L^2G^2}{\lambda^2}\right) + O(mN^{-2}),$$

where c is a numerical constant.

Proof Rewriting the log-likelihood in the notation of Theorem 1, we have $\nabla\ell(x \mid \theta^*) = \nabla f(\theta^*;x)$ and all we need to note is that

$$\begin{aligned} I(\theta^*)^{-1} &= \mathbb{E}\left[I(\theta^*)^{-1}\nabla\ell(X \mid \theta^*)\nabla\ell(X \mid \theta^*)^\top I(\theta^*)^{-1}\right] \\ &= \mathbb{E}\left[\left(\nabla^2 F_0(\theta^*)^{-1}\nabla f(\theta^*;X)\right)\left(\nabla^2 F_0(\theta^*)^{-1}\nabla f(\theta^*;X)\right)^\top\right]. \end{aligned}$$

Now apply the linearity of the trace and use the fact that $\operatorname{tr}(uu^\top) = \|u\|_2^2$. ■

Except for the factor of two in the bound, Corollary 3 shows that Theorem 1 essentially achieves the best possible result. The important aspect of our bound, however, is that we obtain this convergence rate without calculating an estimate on all $N = mn$ samples: instead, we calculate m independent estimators, and then average them to attain the convergence guarantee. We remark that an inspection of our proof shows that, at the expense of worse constants on higher order terms, we can reduce the factor of $2/mn$ on the leading term in Theorem 1 to $(1+c)/mn$ for any constant $c > 0$; as made clear by Corollary 3, this is unimprovable, even by constant factors.

As noted in the introduction, our bounds are certainly to be expected for unbiased estimators, since in such cases averaging m independent solutions reduces the variance by $1/m$. In this sense, our results are similar to classical distributional convergence results in M-estimation: for smooth enough problems, M-estimators behave asymptotically like averages (van der Vaart, 1998; Lehmann and Casella, 1998), and averaging multiple independent realizations reduces their variance. However, it is often desirable to use biased estimators, and such bias introduces difficulty in the analysis, which we explore more in the next section. We also note that in contrast to classical asymptotic results, our results are applicable to finite samples and give explicit upper bounds on the mean-squared error. Lastly, our results are not tied to a specific model, which allows for fairly general sampling distributions.

3.3 Bounds for Subsampled Mixture Averaging

When the number of machines m is relatively small, Theorem 1 and Corollary 2 show that the convergence rate of the AVGM algorithm is mainly determined by the first term in the bound (6), which is at most $\frac{G^2}{\lambda^2 mn}$. In contrast, when the number of processors m grows, the second term in the bound (6), in spite of being $O(n^{-2})$, may have non-negligible effect. This issue is exacerbated when the local strong convexity parameter λ of the risk F_0 is close to zero or the Lipschitz continuity constant H of ∇f is large. This concern motivated our development of the subsampled average mixture (SAVGM) algorithm, to which we now return.

Due to the additional randomness introduced by the subsampling in SAVGM, its analysis requires an additional smoothness condition. In particular, recalling the Euclidean ρ-neighbourhood U of the optimum θ^*, we require that the loss function f is (locally) smooth through its third derivatives.

Assumption 4 (Strong smoothness) *For each $x \in \mathcal{X}$, the third derivatives of f are $M(x)$-Lipschitz continuous, meaning that*

$$\left\| \left(\nabla^3 f(\theta; x) - \nabla^3 f(\theta'; x) \right) (u \otimes u) \right\|_2 \leq M(x) \left\| \theta - \theta' \right\|_2 \|u\|_2^2 \quad \text{for all } \theta, \theta' \in U, \text{ and } u \in \mathbb{R}^d,$$

where $\mathbb{E}[M^8(X)] \leq M^8$ for some constant $M < \infty$.

It is easy to verify that Assumption 4 holds for least-squares regression with $M = 0$. It also holds for various types of non-linear regression problems (e.g., logistic, multinomial etc.) as long as the covariates have finite eighth moments.

With this set-up, our second theorem establishes that bootstrap sampling yields improved performance:

Theorem 4 *Under Assumptions 1 through 4, the output $\overline{\theta}_{\text{SAVGM}} = (\overline{\theta}_1 - r\overline{\theta}_2)/(1-r)$ of the bootstrap* SAVGM *algorithm has mean-squared error bounded as*

$$\mathbb{E}\left[\left\| \overline{\theta}_{\text{SAVGM}} - \theta^* \right\|_2^2 \right] \leq \frac{2+3r}{(1-r)^2} \cdot \frac{1}{nm} \mathbb{E}\left[\left\| \nabla^2 F_0(\theta^*)^{-1} \nabla f(\theta^*; X) \right\|_2^2 \right] + c\left(\frac{M^2 G^6}{\lambda^6} + \frac{G^4 L^2 d \log d}{\lambda^4} \right) \left(\frac{1}{r(1-r)^2} \right) n^{-3} + O\left(\frac{1}{(1-r)^2} m^{-1} n^{-2} \right) \tag{9}$$

for a numerical constant c.

Comparing the conclusions of Theorem 4 to those of Theorem 1, we see that the the $O(n^{-2})$ term in the bound (6) has been eliminated. The reason for this elimination is that subsampling at a rate r reduces the bias of the SAVGM algorithm to $O(n^{-3})$, whereas in contrast, the bias of the AVGM algorithm induces terms of order n^{-2}. Theorem 4 suggests that the performance of the SAVGM algorithm is affected by the subsampling rate r; in order to minimize the upper bound (9) in the regime $m < N^{2/3}$, the optimal choice is of the form $r \propto C\sqrt{m}/n = Cm^{3/2}/N$ where $C \approx (G^2/\lambda^2) \max\{MG/\lambda, L\sqrt{d \log d}\}$. Roughly, as the number of machines m becomes larger, we may increase r, since we enjoy averaging affects from the SAVGM algorithm.

Let us consider the relative effects of having larger numbers of machines m for both the AVGM and SAVGM algorithms, which provides some guidance to selecting m in practice. We define $\sigma^2 =$

$\mathbb{E}[\left\|\nabla^2 F_0(\theta^*)^{-1}\nabla f(\theta^*;X)\right\|_2^2]$ to be the asymptotic variance. Then to obtain the optimal convergence rate of σ^2/N, we must have

$$\frac{1}{\lambda^2}\max\left\{H^2\log d, L^2G^2\right\}\frac{m^2}{N^2}\sigma^2 \leq \frac{\sigma^2}{N} \quad \text{or} \quad m \leq N^{\frac{1}{2}}\sqrt{\frac{\lambda^2}{\max\{H^2\log d, L^2G^2/\lambda^2\}}} \tag{10}$$

in Theorem 1. Applying the bound of Theorem 4, we find that to obtain the same rate we require

$$\max\left\{\frac{M^2G^2}{\lambda^6}, \frac{L^2 d\log d}{\lambda^4}\right\}\frac{G^4m^3}{rN^3} \leq \frac{(1+r)\sigma^2}{N} \text{ or } m \leq N^{\frac{2}{3}}\left(\frac{\lambda^4 r(1+r)\sigma^2}{\max\{M^2G^6/\lambda^2, G^4L^2 d\log d\}}\right)^{\frac{1}{3}}.$$

Now suppose that we replace r with $Cm^{3/2}/N$ as in the previous paragraph. Under the conditions $\sigma^2 \approx G^2$ and $r = o(1)$, we then find that

$$m \leq N^{\frac{2}{3}}\left(\frac{\lambda^2\sigma^2 m^{3/2}}{G^2\max\left\{MG/\lambda, L\sqrt{d\log d}\right\}N}\right)^{\frac{1}{3}} \text{ or } m \leq N^{\frac{2}{3}}\left(\frac{\lambda^2}{\max\left\{MG/\lambda, L\sqrt{d\log d}\right\}}\right)^{\frac{2}{3}}. \tag{11}$$

Comparing inequalities (10) and (11), we see that in both cases m may grow polynomially with the global sample size N while still guaranteeing optimal convergence rates. On one hand, this asymptotic growth is faster in the subsampled case (11); on the other hand, the dependence on the dimension d of the problem is more stringent than the standard averaging case (10). As the local strong convexity constant λ of the *population risk* shrinks, both methods allow less splitting of the data, meaning that the sample size per machine must be larger. This limitation is intuitive, since lower curvature for the population risk means that the local empirical risks associated with each machine will inherit lower curvature as well, and this effect will be exacerbated with a small local sample size per machine. Averaging methods are, of course, not a panacea: the allowed number of partitions m does not grow linearly in either case, so blindly increasing the number of machines proportionally to the total sample size N will not lead to a useful estimate.

In practice, an optimal choice of r may not be apparent, which may necessitate cross validation or another type of model evaluation. We leave as intriguing open questions whether computing multiple subsamples at each machine can yield improved performance or reduce the variance of the SAVGM procedure, and whether using estimates based on resampling the data with replacement, as opposed to without replacement as considered here, can yield improved performance.

3.4 Time Complexity

In practice, the exact empirical minimizers assumed in Theorems 1 and 4 may be unavailable. Instead, we need to use a finite number of iterations of some optimization algorithm in order to obtain reasonable approximations to the exact minimizers. In this section, we sketch an argument that shows that both the AVGM algorithm and the SAVGM algorithm can use such approximate empirical minimizers, and as long as the optimization error is sufficiently small, the resulting averaged estimate achieves the same order-optimal statistical error. Here we provide the arguments only for the AVGM algorithm; the arguments for the SAVGM algorithm are analogous.

More precisely, suppose that each processor runs a finite number of iterations of some optimization algorithm, thereby obtaining the vector θ_i' as an approximate minimizer of the objective function $F_{1,i}$. Thus, the vector θ_i' can be viewed as an approximate form of θ_i, and we let $\overline{\theta}' = \frac{1}{m}\sum_{i=1}^m \theta_i'$ denote

the average of these approximate minimizers, which corresponds to the output of the approximate AVGM algorithm. With this notation, we have

$$\mathbb{E}\left[\left\|\overline{\theta}' - \theta^*\right\|_2^2\right] \overset{(i)}{\leq} 2\mathbb{E}[\left\|\overline{\theta} - \theta^*\right\|_2^2] + 2\mathbb{E}\left[\left\|\overline{\theta}' - \overline{\theta}\right\|_2^2\right] \overset{(ii)}{\leq} 2\mathbb{E}[\left\|\overline{\theta} - \theta^*\right\|_2^2] + 2\mathbb{E}[\left\|\theta_1' - \theta_1\right\|_2^2], \quad (12)$$

where step (i) follows by triangle inequality and the elementary bound $(a+b)^2 \leq 2a^2 + 2b^2$; step (ii) follows by Jensen's inequality. Consequently, suppose that processor i runs enough iterations to obtain an approximate minimizer θ_1' such that

$$\mathbb{E}[\left\|\theta_i' - \theta_i\right\|_2^2] = O((mn)^{-2}). \quad (13)$$

When this condition holds, the bound (12) shows that the average $\overline{\theta}'$ of the approximate minimizers shares the same convergence rates provided by Theorem 1.

But how long does it take to compute an approximate minimizer θ_i' satisfying condition (13)? Assuming processing one sample requires one unit of time, we claim that this computation can be performed in time $O(n\log(mn))$. In particular, the following two-stage strategy, involving a combination of stochastic gradient descent (see the following subsection for more details) and standard gradient descent, has this complexity:

(1) As shown in the proof of Theorem 1, with high probability, the empirical risk F_1 is strongly convex in a ball $B_\rho(\theta_1)$ of constant radius $\rho > 0$ around θ_1. Consequently, performing stochastic gradient descent on F_1 for $O(\log^2(mn)/\rho^2)$ iterations yields an approximate minimizer that falls within $B_\rho(\theta_1)$ with high probability (e.g., Nemirovski et al., 2009, Proposition 2.1). Note that the radius ρ for local strong convexity is a property of the population risk F_0 we use as a prior knowledge.

(2) This initial estimate can be further improved by a few iterations of standard gradient descent. Under local strong convexity of the objective function, gradient descent is known to converge at a geometric rate (see, e.g., Nocedal and Wright, 2006; Boyd and Vandenberghe, 2004), so $O(\log(1/\varepsilon))$ iterations will reduce the error to order ε. In our case, we have $\varepsilon = (mn)^{-2}$, and since each iteration of standard gradient descent requires $O(n)$ units of time, a total of $O(n\log(mn))$ time units are sufficient to obtain a final estimate θ_1' satisfying condition (13).

Overall, we conclude that the speed-up of the AVGM relative to the naive approach of processing all $N = mn$ samples on one processor, is at least of order $m/\log(N)$.

3.5 Stochastic Gradient Descent with Averaging

The previous strategy involved a combination of stochastic gradient descent and standard gradient descent. In many settings, it may be appealing to use only a stochastic gradient algorithm, due to their ease of their implementation and limited computational requirements. In this section, we describe an extension of Theorem 1 to the case in which each machine computes an approximate minimizer using only stochastic gradient descent.

Stochastic gradient algorithms have a lengthy history in statistics, optimization, and machine learning (Robbins and Monro, 1951; Polyak and Juditsky, 1992; Nemirovski et al., 2009; Rakhlin et al., 2012). Let us begin by briefly reviewing the basic form of stochastic gradient descent (SGD). Stochastic gradient descent algorithms iteratively update a parameter vector θ^t over time based on

randomly sampled gradient information. Specifically, at iteration t, a sample X_t is drawn at random from the distribution P (or, in the case of a finite set of data $\{X_1, \ldots, X_n\}$, a sample X_t is chosen from the data set). The method then performs the following two steps:

$$\theta^{t+\frac{1}{2}} = \theta^t - \eta_t \nabla f(\theta^t; X_t) \quad \text{and} \quad \theta^{t+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \left\| \theta - \theta^{t+\frac{1}{2}} \right\|_2^2 \right\}. \tag{14}$$

Here $\eta_t > 0$ is a stepsize, and the first update in (14) is a gradient descent step with respect to the random gradient $\nabla f(\theta^t; X_t)$. The method then projects the intermediate point $\theta^{t+\frac{1}{2}}$ back onto the constraint set Θ (if there is a constraint set). The convergence of SGD methods of the form (14) has been well-studied, and we refer the reader to the papers by Polyak and Juditsky (1992), Nemirovski et al. (2009), and Rakhlin et al. (2012) for deeper investigations.

To prove convergence of our stochastic gradient-based averaging algorithms, we require the following smoothness and strong convexity condition, which is an alternative to the Assumptions 2 and 3 used previously.

Assumption 5 (Smoothness and Strong Convexity II) *There exists a function $L : \mathcal{X} \to \mathbb{R}_+$ such that*

$$\left\| \left\| \nabla^2 f(\theta; x) - \nabla^2 f(\theta^*; x) \right\| \right\|_2 \leq L(x) \left\| \theta - \theta^* \right\|_2 \quad \textit{for all } x \in \mathcal{X},$$

and $\mathbb{E}[L^2(X)] \leq L^2 < \infty$. There are finite constants G and H such that

$$\mathbb{E}[\|\nabla f(\theta; X)\|_2^4] \leq G^4, \quad \textit{and} \quad \mathbb{E}[\left\| \left\| \nabla^2 f(\theta^*; X) \right\| \right\|_2^4] \leq H^4 \quad \textit{for each fixed } \theta \in \Theta.$$

In addition, the population function F_0 is λ-strongly convex over the space Θ, meaning that

$$\nabla^2 F_0(\theta) \succeq \lambda I_{d \times d} \quad \textit{for all } \theta \in \Theta.$$

Assumption 5 does not require as many moments as does Assumption 3, but it does require each moment bound to hold globally, that is, over the entire space Θ, rather than only in a neighbourhood of the optimal point θ^*. Similarly, the necessary curvature—in the form of the lower bound on the Hessian matrix $\nabla^2 F_0$—is also required to hold globally, rather than only locally. Nonetheless, Assumption 5 holds for many common problems; for instance, it holds for any linear regression problem in which the covariates have finite fourth moments and the domain Θ is compact.

The averaged stochastic gradient algorithm (SGDAVGM) is based on the following two steps:

(1) Given some constant $c > 1$, each machine performs n iterations of stochastic gradient descent (14) on its local data set of n samples using the stepsize $\eta_t = \frac{c}{\lambda t}$, then outputs the resulting local parameter θ_i'.

(2) The algorithm computes the average $\overline{\theta}^n = \frac{1}{m} \sum_{i=1}^m \theta_i'$.

The following result characterizes the mean-squared error of this procedure in terms of the constants

$$\alpha := 4c^2 \quad \text{and} \quad \beta := \max \left\{ \left\lceil \frac{cH}{\lambda} \right\rceil, \frac{c\alpha^{3/4} G^{3/2}}{(c-1)\lambda^{5/2}} \left(\frac{\alpha^{1/4} L G^{1/2}}{\lambda^{1/2}} + \frac{4G + HR}{\rho^{3/2}} \right) \right\}.$$

Theorem 5 *Under Assumptions 1 and 5, the output $\bar{\theta}^n$ of the* SAVGM *algorithm has mean-squared error upper bounded as*

$$\mathbb{E}\left[\left\|\bar{\theta}^n - \theta^*\right\|_2^2\right] \leq \frac{\alpha G^2}{\lambda^2 mn} + \frac{\beta^2}{n^{3/2}}. \tag{15}$$

Theorem 5 shows that the averaged stochastic gradient descent procedure attains the optimal convergence rate $O(N^{-1})$ as a function of the total number of observations $N = mn$. The constant and problem-dependent factors are somewhat worse than those in the earlier results we presented in Theorems 1 and 4, but the practical implementability of such a procedure may in some circumstances outweigh those differences. We also note that the second term of order $O(n^{-3/2})$ may be reduced to $O(n^{(2-2k)/k})$ for any $k \geq 4$ by assuming the existence of kth moments in Assumption 5; we show this in passing after our proof of the theorem in Appendix D. It is not clear whether a bootstrap correction is possible for the stochastic-gradient based estimator; such a correction could be significant, because the term $\beta^2/n^{3/2}$ arising from the bias in the stochastic gradient estimator may be non-trivial. We leave this question to future work.

4. Performance on Synthetic Data

In this section, we report the results of simulation studies comparing the AVGM, SAVGM, and SGDAVGM methods, as well as a trivial method using only a fraction of the data available on a single machine. For each of our simulated experiments, we use a fixed total number of samples $N = 100{,}000$, but we vary the number of parallel splits m of the data (and consequently, the local data set sizes $n = N/m$) and the dimensionality d of the problem solved.

For our experiments, we simulate data from one of three regression models:

$$y = \langle u, x\rangle + \varepsilon, \tag{16}$$

$$y = \langle u, x\rangle + \sum_{j=1}^{d} v_j x_j^3 + \varepsilon, \quad \text{or} \tag{17}$$

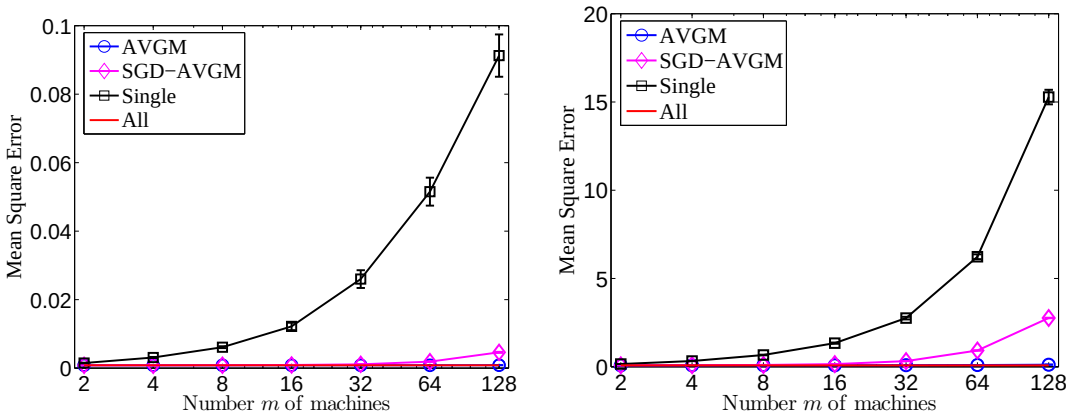
$$y = \langle u, x\rangle + h(x)|\varepsilon|, \tag{18}$$

where $\varepsilon \sim N(0,1)$, and h is a function to be specified. Specifically, the data generation procedure is as follows. For each individual simulation, we choose fixed vector $u \in \mathbb{R}^d$ with entries u_i distributed uniformly in $[0,1]$ (and similarly for v), and we set $h(x) = \sum_{j=1}^{d}(x_j/2)^3$. The models (16) through (18) provide points on a curve from correctly-specified to grossly mis-specified models, so models (17) and (18) help us understand the effects of subsampling in the SAVGM algorithm. (In contrast, the standard least-squares estimator is unbiased for model (16).) The noise variable ε is always chosen as a standard Gaussian variate $N(0,1)$, independent from sample to sample.

In our simulation experiments we use the least-squares loss

$$f(\theta; (x,y)) := \frac{1}{2}(\langle \theta, x\rangle - y)^2.$$

The goal in each experiment is to estimate the vector θ^* minimizing $F_0(\theta) := \mathbb{E}[f(\theta; (X,Y))]$. For each simulation, we generate N samples according to either the model (16) or (18). For each $m \in \{2, 4, 8, 16, 32, 64, 128\}$, we estimate $\theta^* = \arg\min_\theta F_0(\theta)$ using a parallel method with data split into m independent sets of size $n = N/m$, specifically



(a) $d = 20$ (b) $d = 200$

Figure 1: The error $\|\widehat{\theta} - \theta^*\|_2^2$ versus number of machines, with standard errors across twenty simulations, for solving least squares with data generated according to the normal model (16). The oracle least-squares estimate using all N samples is given by the line "All," while the line "Single" gives the performance of the naive estimator using only $n = N/m$ samples.

(i) The AVGM method

(ii) The SAVGM method with several settings of the subsampling ratio r

(iii) The SGDAVGM method with stepsize $\eta_t = d/(10(d+t))$, which gave good performance.

In addition to (i)–(iii), we also estimate θ^* with

(iv) The empirical minimizer of a single split of the data of size $n = N/m$

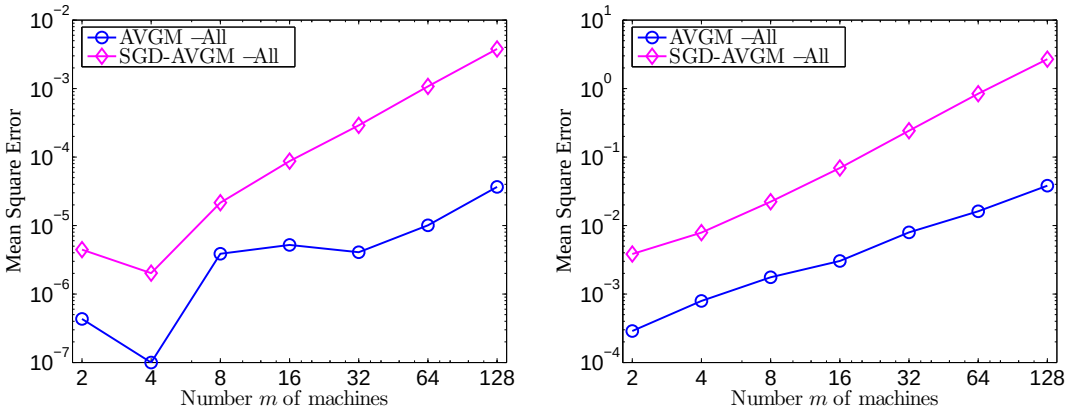
(v) The empirical minimizer on the full data set (the oracle solution).

4.1 Averaging Methods

For our first set of experiments, we study the performance of the averaging methods (AVGM and SAVGM), showing their scaling as the number of splits of data—the number of machines m—grows for fixed N and dimensions $d = 20$ and $d = 200$. We use the standard regression model (16) to generate the data, and throughout we let $\widehat{\theta}$ denote the estimate returned by the method under consideration (so in the AVGM case, for example, this is the vector $\widehat{\theta} := \overline{\theta}_1$). The data samples consist of pairs (x, y), where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ is the target value. To sample each x vector, we choose five distinct indices in $\{1, \ldots, d\}$ uniformly at random, and the entries of x at those indices are distributed as $N(0,1)$. For the model (16), the population optimal vector θ^* is u.

In Figure 1, we plot the error $\|\widehat{\theta} - \theta^*\|_2^2$ of the inferred parameter vector $\widehat{\theta}$ for the true parameters θ^* versus the number of splits m, or equivalently, the number of separate machines available for use. We also plot standard errors (across twenty experiments) for each curve. As a baseline in each plot, we plot as a red line the squared error $\|\widehat{\theta}_N - \theta^*\|_2^2$ of the centralized "gold standard," obtained by applying a batch method to all N samples.

From the plots in Figure 1, we can make a few observations. The AVGM algorithm enjoys excellent performance, as predicted by our theoretical results, especially compared to the naive



(a) $d = 20$ (b) $d = 200$

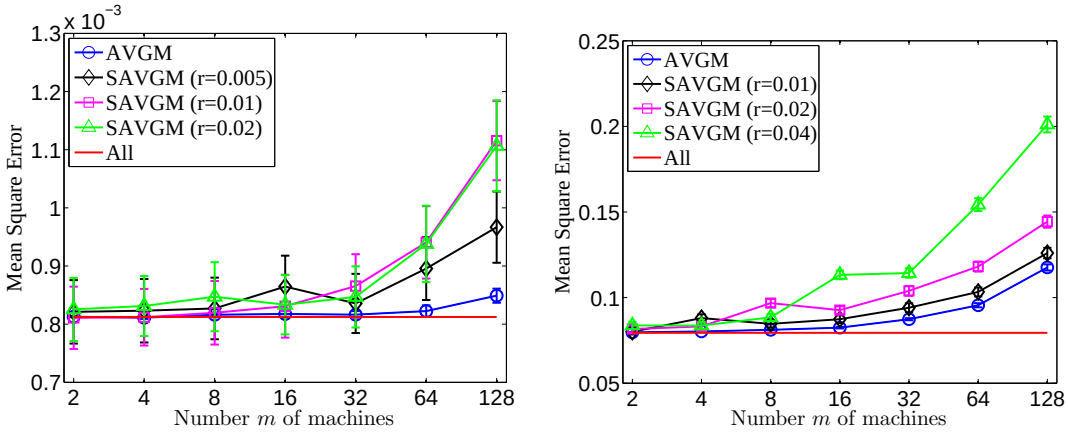
Figure 2: Comparison of AVGM and SGDAVGM methods as in Figure 1 plotted on logarithmic scale. The plot shows $\|\widehat{\theta} - \theta^*\|_2^2 - \|\theta_N - \theta^*\|_2^2$, where θ_N is the oracle least-squares estimator using all N data samples.

solution using only a fraction $1/m$ of the data. In particular, if $\widehat{\theta}$ is obtained by the batch method, then AVGM is almost as good as the full-batch baseline even for m as large as 128, though there is some evident degradation in solution quality. The SGDAVGM (stochastic-gradient with averaging) solution also yields much higher accuracy than the naive solution, but its performance degrades more quickly than the AVGM method's as m grows. In higher dimensions, both the AVGM and SGDAVGM procedures have somewhat worse performance; again, this is not unexpected since in high dimensions the strong convexity condition is satisfied with lower probability in local data sets.

We present a comparison between the AVGM method and the SGDAVGM method with somewhat more distinguishing power in Figure 2. For these plots, we compute the gap between the AVGM mean-squared-error and the unparallel baseline MSE, which is the accuracy lost due to parallelization or distributing the inference procedure across multiple machines. Figure 2 shows that the mean-squared error grows polynomially with the number of machines m, which is consistent with our theoretical results. From Corollary 3, we expect the AVGM method to suffer (lower-order) penalties proportional to m^2 as m grows, while Theorem 5 suggests the somewhat faster growth we see for the SGDAVGM method in Figure 2. Thus, we see that the improved run-time performance of the SGDAVGM method—requiring only a single pass through the data on each machine, touching each datum only once—comes at the expense of some loss of accuracy, as measured by mean-squared error.

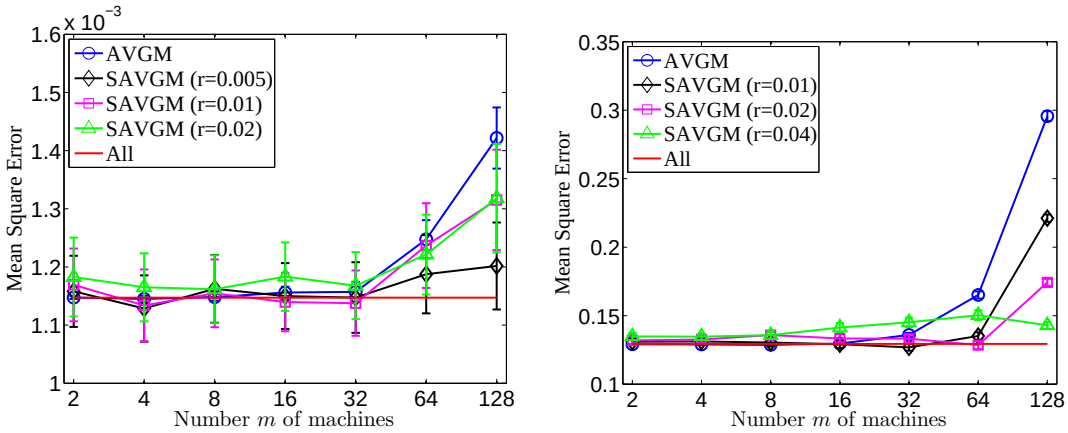
4.2 Subsampling Correction

We now turn to developing an understanding of the SAVGM algorithm in comparison to the standard average mixture algorithm, developing intuition for the benefits and drawbacks of the method. Before describing the results, we remark that for the standard regression model (16), the least-squares solution is unbiased for θ^*, so we expect subsampled averaging to yield little (if any) improvement. The SAVGM method is essentially aimed at correcting the bias of the estimator $\overline{\theta}_1$, and de-biasing an unbiased estimator only increases its variance. However, for the mis-specified models (17) and (18)



(a) $d = 20$ (b) $d = 200$

Figure 3: The error $\|\widehat{\theta} - \theta^*\|_2^2$ plotted against the number of machines m for the AVGM and SAVGM methods, with standard errors across twenty simulations, using the normal regression model (16). The oracle estimator is denoted by the line "All."

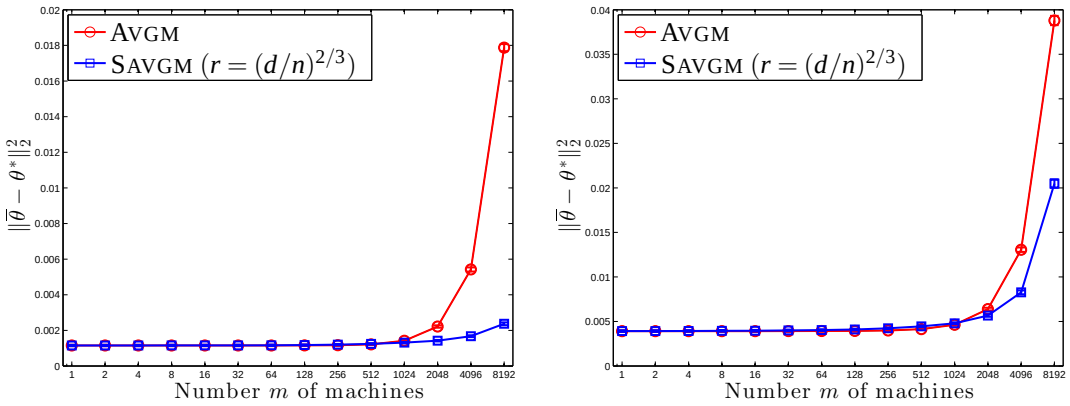


(a) $d = 20$ (b) $d = 200$

Figure 4: The error $\|\widehat{\theta} - \theta^*\|_2^2$ plotted against the number of machines m for the AVGM and SAVGM methods, with standard errors across twenty simulations, using the non-normal regression model (18). The oracle estimator is denoted by the line "All."

we expect to see some performance gains. In our experiments, we use multiple sub-sampling rates to study their effects, choosing $r \in \{0.005, 0.01, 0.02, 0.04\}$, where we recall that the output of the SAVGM algorithm is the vector $\widehat{\theta} := (\overline{\theta}_1 - r\overline{\theta}_2)/(1-r)$.

We begin with experiments in which the data is generated as in the previous section. That is, to generate a feature vector $x \in^d$, choose five distinct indices in $\{1, \ldots, d\}$ uniformly at random, and the entries of x at those indices are distributed as $N(0,1)$. In Figure 3, we plot the results of simulations comparing AVGM and SAVGM with data generated from the normal regression model (16). Both algorithms have have low error rates, but the AVGM method is slightly better than the SAVGM method for both values of the dimension d and all and sub-sampling rates r. As expected, in this



(a) $d = 20$ (b) $d = 40$

Figure 5: The error $\|\widehat{\theta} - \theta^*\|_2^2$ plotted against the number of machines m for the AVGM and SAVGM methods using regression model (17).

case the SAVGM method does not offer improvement over AVGM, since the estimators are unbiased. (In Figure 3(a), we note that the standard error is in fact very small, since the mean-squared error is only of order 10^{-3}.)

To understand settings in which subsampling for bias correction helps, in Figure 4, we plot mean-square error curves for the least-squares regression problem when the vector y is sampled according to the non-normal regression model (18). In this case, the least-squares estimator is biased for θ^* (which, as before, we estimate by solving a larger regression problem using $10N$ data samples). Figure 4 shows that both the AVGM and SAVGM method still enjoy good performance; in some cases, the SAVGM method even beats the oracle least-squares estimator for θ^* that uses all N samples. Since the AVGM estimate is biased in this case, its error curve increases roughly quadratically with m, which agrees with our theoretical predictions in Theorem 1. In contrast, we see that the SAVGM algorithm enjoys somewhat more stable performance, with increasing benefit as the number of machines m increases. For example, in case of $d = 200$, if we choose $r = 0.01$ for $m \leq 32$, choose $r = 0.02$ for $m = 64$ and $r = 0.04$ for $m = 128$, then SAVGM has performance comparable with the oracle method that uses all N samples. Moreover, we see that all the values of r—at least for the reasonably small values we use in the experiment—provide performance improvements over a non-subsampled distributed estimator.

For our final simulation, we plot results comparing SAVGM with AVGM in model (17), which is mis-specified but still a normal model. We use a simpler data generating mechanism, specifically, we draw $x \sim N(0, I_{d\times d})$ from a standard d-dimensional normal, and v is chosen uniformly in $[0,1]$; in this case, the population minimizer has the closed form $\theta^* = u + 3v$. Figure 5 shows the results for dimensions $d = 20$ and $d = 40$ performed over 100 experiments (the standard errors are too small to see). Since the model (17) is not that badly mis-specified, the performance of the SAVGM method improves upon that of the AVGM method only for relatively large values of m, however, the performance of the SAVGM is always at least as good as that of AVGM.

Feature Name	Dimension	Description
Query	20000	Word tokens appearing in the query.
Gender	3	Gender of the user
Keyword	20000	Word tokens appearing in the purchase keywords.
Title	20000	Word tokens appearing in the ad title.
Advertiser	39191	Advertiser's ID
AdID	641707	Advertisement's ID.
Age	6	Age of the user
UserFreq	25	Number of appearances of the same user.
Position	3	Position of advertisement on search page.
Depth	3	Number of ads in the session.
QueryFreq	25	Number of occurrences of the same query.
AdFreq	25	Number of occurrences of the same ad.
QueryLength	20	Number of words in the query.
TitleLength	30	Number of words in the ad title.
DespLength	50	Number of words in the ad description.
QueryCtr	150	Average click-through-rate for query.
UserCtr	150	Average click-through-rate for user.
AdvrCtr	150	Average click-through-rate for advertiser.
WordCtr	150	Average click-through-rate for keyword advertised.
UserAdFreq	20	Number of times this user sees an ad.
UserQueryFreq	20	Number of times this user performs a search.

Table 1: Features used in online advertisement prediction problem.

5. Experiments with Advertising Data

Predicting whether a user of a search engine will click on an advertisement presented to him or her is of central importance to the business of several internet companies, and in this section, we present experiments studying the performance of the AVGM and SAVGM methods for this task. We use a large data set from the Tencent search engine, `soso.com` (Sun, 2012), which contains 641,707 distinct advertisement items with $N = 235{,}582{,}879$ data samples.

Each sample consists of a so-called *impression*, which in the terminology of the information retrieval literature (e.g., see the book by Manning et al., 2008), is a list containing a user-issued search, the advertisement presented to the user in response to the search, and a label $y \in \{+1, -1\}$ indicating whether the user clicked on the advertisement. The ads in our data set were presented to 23,669,283 distinct users.

Transforming an impression into a useable set of regressors x is non-trivial, but the Tencent data set provides a standard encoding. We list the features present in the data in Table 1, along with some description of their meaning. Each text-based feature—that is, those made up of words, which are Query, Keyword, and Title—is given a "bag-of-words" encoding (Manning et al., 2008). This encoding assigns each of 20,000 possible words an index, and if the word appears in the query (or Keyword or Title feature), the corresponding index in the vector x is set to 1. Words that do not appear are encoded with a zero. Real-valued features, corresponding to the bottom fifteen features in Table 1 beginning with "Age", are binned into a fixed number of intervals $[-\infty, a_1], (a_1, a_2], \ldots, (a_k, \infty]$,

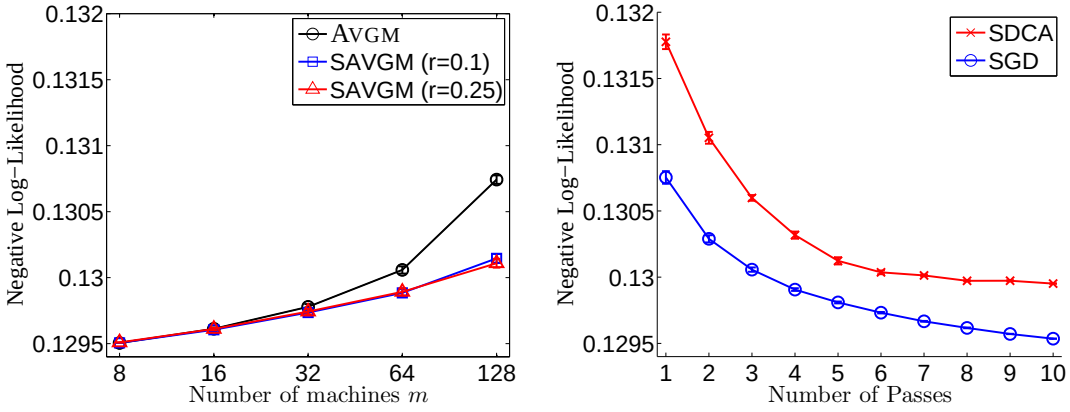


Figure 6: The negative log-likelihood of the output of the AVGM, SAVGM, and stochastic methods on the held-out data set for the click-through prediction task. (a) Performance of the AVGM and SAVGM methods versus the number of splits m of the data. (b) Performance of SDCA and SGD baselines as a function of number of passes through the entire data set.

each of which is assigned an index in x. (Note that the intervals and number thereof vary per feature, and the dimension of the features listed in Table 1 corresponds to the number of intervals). When a feature falls into a particular bin, the corresponding entry of x is assigned a 1, and otherwise the entries of x corresponding to the feature are 0. Each feature has one additional value for "unknown." The remaining categorical features—gender, advertiser, and advertisement ID (AdID)—are also given $\{0,1\}$ encodings, where only one index of x corresponding to the feature may be non-zero (which indicates the particular gender, advertiser, or AdID). This combination of encodings yields a binary-valued covariate vector $x \in \{0,1\}^d$ with $d = 741{,}725$ dimensions. Note also that the features incorporate information about the user, advertisement, and query issued, encoding information about their interactions into the model.

Our goal is to predict the probability of a user clicking a given advertisement as a function of the covariates in Table 1. To do so, we use a logistic regression model to estimate the probability of a click response

$$P(y = 1 \mid x; \theta) := \frac{1}{1 + \exp(-\langle \theta, x \rangle)},$$

where $\theta \in \mathbb{R}^d$ is the unknown regression vector. We use the negative logarithm of P as the loss, incorporating a ridge regularization penalty. This combination yields instantaneous loss

$$f(\theta; (x,y)) = \log\left(1 + \exp(-y \langle \theta, x \rangle)\right) + \frac{\lambda}{2} \|\theta\|_2^2.$$

In all our experiments, we assume that the population negative log-likelihood risk has local strong convexity as suggested by Assumption 2. In practice, we use a small regularization parameter $\lambda = 10^{-6}$ to ensure fast convergence for the local sub-problems.

For this problem, we cannot evaluate the mean-squared error $\|\widehat{\theta} - \theta^*\|_2^2$, as we do not know the true optimal parameter θ^*. Consequently, we evaluate the performance of an estimate $\widehat{\theta}$ using

log-loss on a held-out data set. Specifically, we perform a five-fold validation experiment, where we shuffle the data and partition it into five equal-sized subsets. For each of our five experiments, we hold out one partition to use as the test set, using the remaining data as the training set for inference. When studying the AVGM or SAVGM method, we compute the local estimate θ_i via a trust-region Newton-based method (Nocedal and Wright, 2006) implemented by LIBSVM (Chang and Lin, 2011).

The data set is too large to fit in the memory of most computers: in total, four splits of the data require 55 gigabytes. Consequently, it is difficult to provide an oracle training comparison using the full N samples. Instead, for each experiment, we perform 10 passes of stochastic dual coordinate ascent (SDCA) (Shalev-Shwartz and Zhang, 2012) and 10 passes of stochastic gradient descent (SGD) through the data set to get two rough baselines of the performance attained by the empirical minimizer for the entire training data set. Figure 6(b) shows the hold-out set log-loss after each of the sequential passes through the training data finishes. Note that although the SDCA enjoys faster convergence rate on the regularized empirical risk (Shalev-Shwartz and Zhang, 2012), the plot shows that the SGD has better generalization performance.

In Figure 6(a), we show the average hold-out set log-loss (with standard errors) of the estimator $\overline{\theta}_1$ provided by the AVGM method versus number of splits of the data m, and we also plot the log-loss of the SAVGM method using subsampling ratios of $r \in \{.1, .25\}$. The plot shows that for small m, both AVGM and SAVGM enjoy good performance, comparable to or better than (our proxy for) the oracle solution using all N samples. As the number of machines m grows, however, the de-biasing provided by the subsampled bootstrap method yields substantial improvements over the standard AVGM method. In addition, even with $m = 128$ splits of the data set, the SAVGM method gives better hold-out set performance than performing two passes of stochastic gradient on the entire data set of m samples; with $m = 64$, SAVGM enjoys performance as strong as looping through the data four times with stochastic gradient descent. This is striking, since doing even one pass through the data with stochastic gradient descent gives minimax optimal convergence rates (Polyak and Juditsky, 1992; Agarwal et al., 2012). In ranking applications, rather than measuring negative log-likelihood, one may wish to use a direct measure of prediction error; to that end, Figure 7 shows plots of the area-under-the-curve (AUC) measure for the AVGM and SAVGM methods; AUC is a well-known measure of prediction error for bipartite ranking (Manning et al., 2008). Broadly, this plot shows a similar story to that in Figure 6.

It is instructive and important to understand the sensitivity of the SAVGM method to the value of the resampling parameter r. We explore this question in Figure 8 using $m = 128$ splits, where we plot the log-loss of the SAVGM estimator on the held-out data set versus the subsampling ratio r. We choose $m = 128$ because more data splits provide more variable performance in r. For the `soso.com` ad prediction data set, the choice $r = .25$ achieves the best performance, but Figure 8 suggests that mis-specifying the ratio is not terribly detrimental. Indeed, while the performance of SAVGM degrades to that of the AVGM method, a wide range of settings of r give improved performance, and there does not appear to be a phase transition to poor performance.

6. Discussion

Large scale statistical inference problems are challenging, and the difficulty of solving them will only grow as data becomes more abundant: the amount of data we collect is growing much faster than the speed or storage capabilities of our computers. Our AVGM, SAVGM, and SGDAVGM meth-

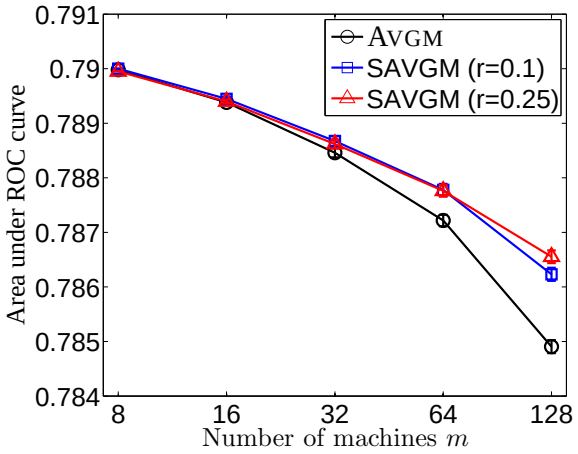


Figure 7: The area-under-the-curve (AUC) measure of ranking error for the output of the AVGM and SAVGM methods for the click-through prediction task.

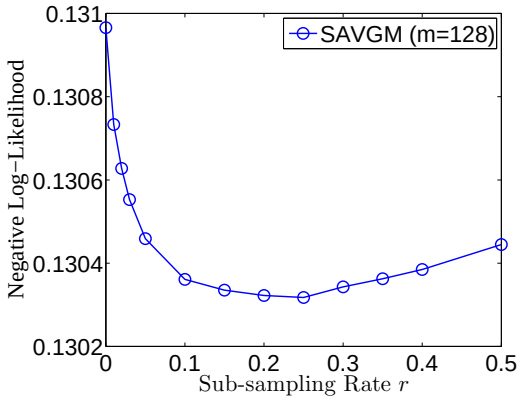


Figure 8: The log-loss on held-out data for the SAVGM method applied with $m = 128$ parallel splits of the data, plotted versus the sub-sampling rate r.

ods provide strategies for efficiently solving such large-scale risk minimization problems, enjoying performance comparable to an oracle method that is able to access the entire large data set. We believe there are several interesting questions that remain open after this work. First, nonparametric estimation problems, which often suffer superlinear scaling in the size of the data, may provide an interesting avenue for further study of decomposition-based methods. Our own recent work has addressed aspects of this challenge in the context of kernel methods for non-parametric regression (Zhang et al., 2013). More generally, an understanding of the interplay between statistical efficiency and communication could provide an avenue for further research, and it may also be interesting to study the effects of subsampled or bootstrap-based estimators in other distributed environments.

Acknowledgments

We thank Joel Tropp for some informative discussions on and references for matrix concentration and moment inequalities. We also thank Ohad Shamir for pointing out a mistake in the statements of results related to Theorem 1, and the editor and reviewers for their helpful comments and feedback. JCD was supported by the Department of Defence under the NDSEG Fellowship Program and by a Facebook PhD fellowship. This work was partially funded by Office of Naval Research MURI grant N00014-11-1-0688 to MJW.

Appendix A. The Necessity of Smoothness

Here we show that some version of the smoothness conditions presented in Assumption 3 are necessary for averaging methods to attain better mean-squared error than using only the n samples on a single processor. Given the loss function (5), let $n_0 = \sum_{i=1}^n 1_{(X_i=0)}$ to be the count of 0 samples. Using θ_1 as shorthand for $\theta_{1,i}$, we see by inspection that the empirical minimizer θ_1 is

$$\theta_1 = \begin{cases} \frac{n_0}{n} - \frac{1}{2} & \text{when } n_0 \le n/2 \\ 1 - \frac{n}{2n_0} & \text{otherwise.} \end{cases}$$

For simplicity, we may assume that n is odd. In this case, we obtain that

$$\begin{aligned} \mathbb{E}[\theta_1] &= \frac{1}{4} + \mathbb{E}\left[\frac{n_0}{n} 1_{(n_0 < n/2)}\right] - \mathbb{E}\left[\frac{n}{2n_0} 1_{(n_0 > n/2)}\right] \\ &= \frac{1}{4} + \frac{1}{2^n} \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{i} \frac{i}{n} - \frac{1}{2^n} \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} \frac{n}{2i} = \frac{1}{4} + \frac{1}{2^n} \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{i} \left[\frac{i}{n} - \frac{n}{2(n-i)}\right] \end{aligned}$$

by the symmetry of the binomial. Adding and subtracting $\frac{1}{2}$ from the term within the braces, noting that $P(n_0 < n/2) = 1/2$, we have the equality

$$\mathbb{E}[\theta_1] = \frac{1}{2^n} \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{i} \left[\frac{i}{n} - \frac{n}{2(n-i)} + \frac{1}{2}\right] = \frac{1}{2^n} \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{i} \frac{i(n-2i)}{2n(n-i)}.$$

If Z is distributed normally with mean $1/2$ and variance $1/(4n)$, then an asymptotic expansion of the binomial distribution yields

$$\begin{aligned} \left(\frac{1}{2}\right)^n \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n}{i} \frac{i(n-2i)}{2n(n-i)} &= \mathbb{E}\left[\frac{Z(1-2Z)}{2-2Z} \mid 0 \le Z \le \frac{1}{2}\right] + o(n^{-1/2}) \\ &\ge \frac{1}{2}\mathbb{E}\left[Z - 2Z^2 \mid 0 \le Z \le \frac{1}{2}\right] + o(n^{-1/2}) = \Omega(n^{-\frac{1}{2}}), \end{aligned}$$

the final equality following from standard calculations, since $\mathbb{E}[|Z|] = \Omega(n^{-1/2})$.

Appendix B. Proof of Theorem 1

Although Theorem 1 is in terms of bounds on 8^{th} order moments, we prove a somewhat more general result in terms of a set of (k_0, k_1, k_2) moment conditions given by

$$\mathbb{E}[\|\nabla f(\theta; X)\|_2^{k_0}] \le G^{k_0}, \qquad \mathbb{E}[\left|\left|\left|\nabla^2 f(\theta; X) - \nabla^2 F_0(\theta)\right|\right|\right|_2^{k_1}] \le H^{k_1},$$
$$\mathbb{E}[L(X)^{k_2}] \le L^{k_2} \quad \text{and} \quad \mathbb{E}[(L(X) - \mathbb{E}[L(X)])^{k_2}] \le L^{k_2}$$

for $\theta \in U$. (Recall the definition of U prior to Assumption 3). Doing so allows sharper control if higher moment bounds are available. The reader should recall throughout our arguments that we have assumed $\min\{k_0, k_1, k_2\} \ge 8$. Throughout the proof, we use F_1 and θ_1 to indicate the local empirical objective and empirical minimizer of machine 1 (which have the same distribution as those of the other processors), and we recall the notation $1_{(\mathcal{E})}$ for the indicator function of the event $\mathcal{E}$.

Before beginning the proof of Theorem 1 proper, we begin with a simple inequality that relates the error term $\overline{\theta} - \theta^*$ to an average of the errors $\theta_i - \theta^*$, each of which we can bound in turn. Specifically, a bit of algebra gives us that

$$\begin{aligned}
\mathbb{E}[\left\|\overline{\theta} - \theta^*\right\|_2^2] &= \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \theta_i - \theta^*\right\|_2^2\right] \\
&= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}[\|\theta_i - \theta^*\|_2^2] + \frac{1}{m^2}\sum_{i \ne j} \mathbb{E}[\left\langle \theta_i - \theta^*, \theta_j - \theta^* \right\rangle] \\
&\le \frac{1}{m}\mathbb{E}[\|\theta_1 - \theta^*\|_2^2] + \frac{m(m-1)}{m^2}\left\|\mathbb{E}[\theta_1 - \theta^*]\right\|_2^2 \\
&\le \frac{1}{m}\mathbb{E}[\|\theta_1 - \theta^*\|_2^2] + \left\|\mathbb{E}[\theta_1 - \theta^*]\right\|_2^2. \qquad (19)
\end{aligned}$$

Here we used the definition of the averaged vector $\overline{\theta}$ and the fact that for $i \ne j$, the vectors θ_i and θ_j are statistically independent, they are functions of independent samples. The upper bound (19) illuminates the path for the remainder of our proof: we bound each of $\mathbb{E}[\|\theta_i - \theta^*\|_2^2]$ and $\|\mathbb{E}[\theta_i - \theta^*]\|_2^2$. Intuitively, since our objective is locally strongly convex by Assumption 2, the empirical minimizing vector θ_1 is a nearly unbiased estimator for θ^*, which allows us to prove the convergence rates in the theorem.

We begin by defining three events—which we (later) show hold with high probability—that guarantee the closeness of θ_1 and θ^*. In rough terms, when these events hold, the function F_1 behaves similarly to the population risk F_0 around the point θ^*; since F_0 is locally strongly convex, the minimizer θ_1 of F_1 will be close to θ^*. Recall that Assumption 3 guarantees the existence of a ball $U_\rho = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2 < \rho\}$ of radius $\rho \in (0,1)$ such that

$$\left|\left|\left|\nabla^2 f(\theta; x) - \nabla^2 f(\theta'; x)\right|\right|\right|_2 \le L(x)\left\|\theta - \theta'\right\|_2$$

for all $\theta, \theta' \in U_\rho$ and any x, where $\mathbb{E}[L(X)^{k_2}] \le L^{k_2}$. In addition, Assumption 2 guarantees that $\nabla^2 F_0(\theta^*) \succeq \lambda I$. Now, choosing the potentially smaller radius $\delta_\rho = \min\{\rho, \rho\lambda/4L\}$, we can define

the three "good" events

$$\mathcal{E}_0 := \left\{ \frac{1}{n} \sum_{i=1}^{n} L(X_i) \leq 2L \right\},$$
$$\mathcal{E}_1 := \left\{ \left| \left| \left| \nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*) \right| \right| \right|_2 \leq \frac{\rho\lambda}{2} \right\}, \quad \text{and} \tag{20}$$
$$\mathcal{E}_2 := \left\{ \|\nabla F_1(\theta^*)\|_2 \leq \frac{(1-\rho)\lambda\delta_\rho}{2} \right\}.$$

We then have the following lemma:

Lemma 6 *Under the events $\mathcal{E}_0$, $\mathcal{E}_1$, and $\mathcal{E}_2$ previously defined* (20)*, we have*

$$\|\theta_1 - \theta^*\|_2 \leq \frac{2\|\nabla F_1(\theta^*)\|_2}{(1-\rho)\lambda}, \quad \textit{and} \quad \nabla^2 F_1(\theta) \succeq (1-\rho)\lambda I_{d\times d}.$$

The proof of Lemma 6 relies on some standard optimization guarantees relating gradients to minimizers of functions (e.g., Boyd and Vandenberghe, 2004, Chapter 9), although some care is required since smoothness and strong convexity hold only locally in our problem. As the argument is somewhat technical, we defer it to Appendix E.

Our approach from here is to give bounds on $\mathbb{E}[\|\theta_1 - \theta^*\|_2^2]$ and $\|\mathbb{E}[\theta_1 - \theta^*]\|_2^2$ by careful Taylor expansions, which allows us to bound $\mathbb{E}[\|\overline{\theta}_1 - \theta^*\|_2^2]$ via our initial expansion (19). We begin by noting that whenever the events $\mathcal{E}_0$, $\mathcal{E}_1$, and $\mathcal{E}_2$ hold, then $\nabla F_1(\theta_1) = 0$, and moreover, by a Taylor series expansion of ∇F_1 between θ^* and θ_1, we have

$$0 = \nabla F_1(\theta_1) = \nabla F_1(\theta^*) + \nabla^2 F_1(\theta')(\theta_1 - \theta^*)$$

where $\theta' = \kappa\theta^* + (1-\kappa)\theta_1$ for some $\kappa \in [0,1]$. By adding and subtracting terms, we have

$$0 = \nabla F_1(\theta^*) + (\nabla^2 F_1(\theta') - \nabla^2 F_1(\theta^*))(\theta_1 - \theta^*) + (\nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*))(\theta_1 - \theta^*) + \nabla^2 F_0(\theta^*)(\theta_1 - \theta^*). \tag{21}$$

Since $\nabla^2 F_0(\theta^*) \succeq \lambda I$, we can define the inverse Hessian matrix $\Sigma^{-1} := [\nabla^2 F_0(\theta^*)]^{-1}$, and setting $\Delta := \theta_1 - \theta^*$, we multiply both sides of the Taylor expansion (21) by Σ^{-1} to obtain the relation

$$\Delta = -\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(\nabla^2 F_1(\theta^*) - \nabla^2 F_1(\theta'))\Delta + \Sigma^{-1}(\nabla^2 F_0(\theta^*) - \nabla^2 F_1(\theta^*))\Delta. \tag{22}$$

Thus, if we define the matrices $P = \nabla^2 F_0(\theta^*) - \nabla^2 F_1(\theta^*)$ and $Q = \nabla^2 F_1(\theta^*) - \nabla^2 F_1(\theta')$, equality (22) can be re-written as

$$\theta_1 - \theta^* = -\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)(\theta_1 - \theta^*). \tag{23}$$

Note that Equation (23) holds when the conditions of Lemma 6 hold, and otherwise we may simply assert only that $\|\theta_1 - \theta^*\|_2 \leq R$. Roughly, we expect the final two terms in the error expansion (23) to be of smaller order than the first term, since we hope that $\theta_1 - \theta^* \to 0$ and additionally that the Hessian differences decrease to zero at a sufficiently fast rate. We now formalize this intuition.

Inspecting the Taylor expansion (23), we see that there are several terms of a form similar to $(\nabla^2 F_0(\theta^*) - \nabla^2 F_1(\theta^*))(\theta_1 - \theta^*)$; using the smoothness Assumption 3, we can convert these terms into higher order terms involving only $\theta_1 - \theta^*$. Thus, to effectively control the expansions (22) and (23), we must show that higher order terms of the form $\mathbb{E}[\|\theta_1 - \theta^*\|_2^k]$, for $k \geq 2$, decrease quickly enough in n.

B.0.1 CONTROL OF $\mathbb{E}[\|\theta_1 - \theta^*\|_2^k]$

Recalling the events (20), we define $\mathcal{E} := \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$ and then observe that

$$\begin{aligned}\mathbb{E}[\|\theta_1 - \theta^*\|_2^k] &= \mathbb{E}[1_{(\mathcal{E})} \|\theta_1 - \theta^*\|_2^k] + \mathbb{E}[1_{(\mathcal{E}^c)} \|\theta_1 - \theta^*\|_2^k] \\ &\leq \frac{2^k \mathbb{E}[1_{(\mathcal{E})} \|\nabla F_1(\theta^*)\|_2^k]}{(1-\rho)^k \lambda^k} + \mathbb{P}(\mathcal{E}^c) R^k \\ &\leq \frac{2^k \mathbb{E}[\|\nabla F_1(\theta^*)\|_2^k]}{(1-\rho)^k \lambda^k} + \mathbb{P}(\mathcal{E}^c) R^k,\end{aligned}$$

where we have used the bound $\|\theta - \theta^*\|_2 \leq R$ for all $\theta \in \Theta$, from Assumption 1. Our goal is to prove that $\mathbb{E}[\|\nabla F_1(\theta^*)\|_2^k] = O(n^{-k/2})$ and that $\mathbb{P}(\mathcal{E}^c) = O(n^{-k/2})$. We move forward with a two lemmas that lay the groundwork for proving these two facts:

Lemma 7 *Under Assumption 3, there exist constants C and C' (dependent only on the moments k_0 and k_1 respectively) such that*

$$\mathbb{E}[\|\nabla F_1(\theta^*)\|_2^{k_0}] \leq C \frac{G^{k_0}}{n^{k_0/2}}, \quad \text{and} \tag{24}$$

$$\mathbb{E}[\left|\left|\left|\nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*)\right|\right|\right|_2^{k_1}] \leq C' \frac{\log^{k_1/2}(2d) H^{k_1}}{n^{k_1/2}}. \tag{25}$$

See Appendix F for the proof of this claim.

As an immediate consequence of Lemma 7, we see that the events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined by (20) occur with reasonably high probability. Indeed, recalling that $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$, Boole's law and the union bound imply

$$\begin{aligned}&\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(\mathcal{E}_0^c \cup \mathcal{E}_1^c \cup \mathcal{E}_2^c) \\ &\leq \mathbb{P}(\mathcal{E}_0^c) + \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) \\ &\leq \frac{\mathbb{E}[|\frac{1}{n}\sum_{i=1}^n L(X_i) - \mathbb{E}[L(X)]|^{k_2}]}{L^{k_2}} + \frac{2^{k_1}\mathbb{E}[\left|\left|\left|\nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*)\right|\right|\right|_2^{k_1}]}{\rho^{k_1}\lambda^{k_1}} + \frac{2^{k_0}\mathbb{E}[\|\nabla F_1(\theta^*)\|_2^{k_0}]}{(1-\rho)^{k_0}\lambda^{k_0}\delta_\rho^{k_0}} \\ &\leq C_2 \frac{1}{n^{k_2/2}} + C_1 \frac{\log^{k_1/2}(2d) H^{k_1}}{n^{k_1/2}} + C_0 \frac{G^{k_0}}{n^{k_0/2}}\end{aligned} \tag{26}$$

for some universal constants C_0, C_1, C_2, where in the second-to-last line we have invoked the moment bound in Assumption 3. Consequently, we find that

$$\mathbb{P}(\mathcal{E}^c) R^k = O(R^k(n^{-k_1/2} + n^{-k_2/2} + n^{-k_0/2}) \quad \text{for any } k \in \mathbb{N}.$$

In summary, we have proved the following lemma:

Lemma 8 *Let Assumptions 2 and 3 hold. For any $k \in \mathbb{N}$ with $k \leq \min\{k_0, k_1, k_2\}$, we have*

$$\mathbb{E}[\|\theta_1 - \theta^*\|_2^k] = O\left(n^{-k/2} \cdot \frac{G^k}{(1-\rho)^k \lambda^k} + n^{-k_0/2} + n^{-k_1/2} + n^{-k_2/2}\right) = O\left(n^{-k/2}\right),$$

where the order statements hold as $n \to +\infty$.

Now recall the matrix $Q = \nabla^2 F_1(\theta^*) - \nabla^2 F_1(\theta')$ defined following Equation (22). The following result controls the moments of its operator norm:

Lemma 9 *For $k \leq \min\{k_2, k_1, k_0\}/2$, we have $\mathbb{E}[|||Q|||_2^k] = O(n^{-k/2})$.*

Proof We begin by using Jensen's inequality and Assumption 3 to see that

$$|||Q|||^k \leq \frac{1}{n}\sum_{i=1}^n \left|\left|\left|\nabla^2 f(\theta'; X_i) - \nabla^2 f(\theta^*; X_i)\right|\right|\right|^k \leq \frac{1}{n}\sum_{i=1}^n L(X_i)^k \left\|\theta' - \theta^*\right\|_2^k.$$

Now we apply the Cauchy-Schwarz inequality and Lemma 8, thereby obtaining

$$\mathbb{E}[|||Q|||_2^k] \leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n L(X_i)^k\right)^2\right]^{\frac{1}{2}} \mathbb{E}\left[\|\theta_1 - \theta^*\|_2^{2k}\right]^{\frac{1}{2}} = O\left(L^k \frac{G^k}{(1-\rho)^k \lambda^k} n^{-k/2}\right),$$

where we have used Assumption 3 again. ■

Lemma 8 allows us to control the first term from our initial bound (19) almost immediately. Indeed, using our last Taylor expansion (23) and the definition of the event $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\begin{aligned}\mathbb{E}[\|\theta_1 - \theta^*\|_2^2] &= \mathbb{E}\left[1_{(\mathcal{E})}\left\|-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)(\theta_1 - \theta^*)\right\|_2^2\right] + \mathbb{E}[1_{(\mathcal{E}^c)}\|\theta_1 - \theta^*\|_2^2] \\ &\leq 2\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\Sigma^{-1}(P+Q)(\theta_1 - \theta^*)\right\|_2^2\right] + \mathbb{P}(\mathcal{E}^c)R^2,\end{aligned}$$

where we have applied the inequality $(a+b)^2 \leq 2a^2 + 2b^2$. Again using this same inequality, then applying Cauchy-Schwarz and Lemmas 8 and 9, we see that

$$\begin{aligned}\mathbb{E}\left[\left\|\Sigma^{-1}(P+Q)(\theta_1 - \theta^*)\right\|_2^2\right] &\leq 2\left|\left|\left|\Sigma^{-1}\right|\right|\right|_2^2 \left(\mathbb{E}[|||P|||_2^2 \|\theta_1 - \theta^*\|_2^2] + \mathbb{E}[|||Q|||_2^2 \|\theta_1 - \theta^*\|_2^2]\right) \\ &\leq 2\left|\left|\left|\Sigma^{-1}\right|\right|\right|_2^2 \left(\sqrt{\mathbb{E}[|||P|||_2^4]\mathbb{E}[\|\theta_1 - \theta^*\|_2^4]} + \sqrt{\mathbb{E}[|||Q|||_2^4]\mathbb{E}[\|\theta_1 - \theta^*\|_2^4]}\right) \\ &= O(n^{-2}),\end{aligned}$$

where we have used the fact that $\min\{k_0, k_1, k_2\} \geq 8$ to apply Lemma 9. Combining these results, we obtain the upper bound

$$\mathbb{E}[\|\theta_1 - \theta^*\|_2^2] \leq 2\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right] + O(n^{-2}), \tag{27}$$

which completes the first part of our proof of Theorem 1.

B.0.2 CONTROL OF $\|\mathbb{E}[\theta_1 - \theta^*]\|_2^2$

It remains to consider the $\|\mathbb{E}[\theta_1 - \theta^*]\|_2^2$ term from our initial error inequality (19). When the events (20) occur, we know that all derivatives exist, so we may recursively apply our expansion (23) of $\theta_1 - \theta^*$ to find that

$$\begin{aligned}\theta_1 - \theta^* &= -\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)(\theta_1 - \theta^*) \\ &= \underbrace{-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)\left[-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)(\theta_1 - \theta^*)\right]}_{=:v}\end{aligned} \tag{28}$$

where we have introduced v as shorthand for the vector on the right hand side. Thus, with a bit of algebraic manipulation we obtain the relation

$$\theta_1 - \theta^* = 1_{(\mathcal{E})} v + 1_{(\mathcal{E}^c)}(\theta_1 - \theta^*) = v + 1_{(\mathcal{E}^c)}(\theta_1 - \theta^*) - 1_{(\mathcal{E}^c)} v = v + 1_{(\mathcal{E}^c)}(\theta_1 - \theta^* - v). \quad (29)$$

Now note that $\mathbb{E}[\nabla F_1(\theta^*)] = 0$ thus

$$\begin{aligned}\mathbb{E}[v] &= \mathbb{E}\left[-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)[-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)(\theta_1 - \theta^*)]\right] \\ &= \mathbb{E}\left[\Sigma^{-1}(P+Q)\Sigma^{-1}\left[(P+Q)(\theta_1 - \theta^*) - \nabla F_1(\theta^*)\right]\right].\end{aligned}$$

Thus, by re-substituting the appropriate quantities in (29) and applying the triangle inequality, we have

$$\begin{aligned}&\|\mathbb{E}[\theta_1 - \theta^*]\|_2 \\ &\leq \left\|\mathbb{E}[\Sigma^{-1}(P+Q)\Sigma^{-1}\left((P+Q)(\theta_1 - \theta^*) - \nabla F_1(\theta^*)\right)]\right\|_2 + \left\|\mathbb{E}[1_{(\mathcal{E}^c)}(\theta_1 - \theta^* - v)]\right\|_2 \\ &\leq \left\|\mathbb{E}[\Sigma^{-1}(P+Q)\Sigma^{-1}\left((P+Q)(\theta_1 - \theta^*) - \nabla F_1(\theta^*)\right)]\right\|_2 + \mathbb{E}[1_{(\mathcal{E}^c)}\|\theta_1 - \theta^*\|_2] \\ &\quad + \mathbb{E}\left[1_{(\mathcal{E}^c)}\left\|-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)\Sigma^{-1}\left[-\nabla F_1(\theta^*) + (P+Q)(\theta_1 - \theta^*)\right]\right\|_2\right]. \quad (30)\end{aligned}$$

Since $\|\theta_1 - \theta^*\|_2 \leq R$ by assumption, we have

$$\mathbb{E}[1_{(\mathcal{E}^c)}\|\theta_1 - \theta^*\|_2] \leq \mathbb{P}(\mathcal{E}^c)R \overset{(i)}{=} O(Rn^{-k/2})$$

for any $k \leq \min\{k_2, k_1, k_0\}$, where step (i) follows from the inequality (26). Hölder's inequality also yields that

$$\begin{aligned}\mathbb{E}\left[1_{(\mathcal{E}^c)}\left\|\Sigma^{-1}(P+Q)\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2\right] &\leq \mathbb{E}\left[1_{(\mathcal{E}^c)}\left|\left|\left|\Sigma^{-1}(P+Q)\right|\right|\right|_2\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2\right] \\ &\leq \sqrt{\mathbb{P}(\mathcal{E}^c)}\mathbb{E}\left[\left|\left|\left|\Sigma^{-1}(P+Q)\right|\right|\right|_2^4\right]^{1/4}\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^4\right]^{1/4}.\end{aligned}$$

Recalling Lemmas 7 and 9, we have $\mathbb{E}[\left|\left|\left|\Sigma^{-1}(P+Q)\right|\right|\right|_2^4] = O(\log^2(d)n^{-2})$, and we similarly have $\mathbb{E}[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^4] = O(n^{-2})$. Lastly, we have $\mathbb{P}(\mathcal{E}^c) = O(n^{-k/2})$ for $k \leq \min\{k_0, k_1, k_2\}$, whence we find that for any such k,

$$\mathbb{E}\left[1_{(\mathcal{E}^c)}\left\|\Sigma^{-1}(P+Q)\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2\right] = O\left(\sqrt{\log(d)}n^{-k/4-1}\right).$$

We can similarly apply Lemma 8 to the last remaining term in the inequality (30) to obtain that for any $k \leq \min\{k_2, k_1, k_0\}$,

$$\begin{aligned}&\mathbb{E}\left[1_{(\mathcal{E}^c)}\left\|-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)\left[-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(P+Q)(\theta_1 - \theta^*)\right]\right\|_2\right] \\ &\qquad = O(n^{-k/2} + n^{-k/4-1}).\end{aligned}$$

Applying these two bounds, we find that

$$\|\mathbb{E}[\theta_1 - \theta^*]\|_2 \leq \left\|\mathbb{E}\left[\Sigma^{-1}(P+Q)\Sigma^{-1}\left((P+Q)(\theta_1 - \theta^*) - \nabla F_1(\theta^*)\right)\right]\right\|_2 + O(n^{-k}) \quad (31)$$

for any k such that $k \leq \min\{k_0, k_1, k_2\}/2$ and $k \leq \min\{k_0, k_1, k_2\}/4 + 1$.

In the remainder of the proof, we show that part of the bound (31) still consists only of higher-order terms, leaving us with an expression not involving $\theta_1 - \theta^*$. To that end, note that

$$\mathbb{E}\left[\left\|\Sigma^{-1}(P+Q)\Sigma^{-1}(P+Q)(\theta_1-\theta^*)\right\|_2^2\right] = O(n^{-3})$$

by three applications of Hölder's inequality, the fact that $\|Ax\|_2 \leq |||A|||_2 \|x\|_2$, and Lemmas 7, 8 and 9. Coupled with our bound (31), we use the fact that $(a+b)^2 \leq 2a^2+2b^2$ to obtain

$$\|\mathbb{E}[\theta_1-\theta^*]\|_2^2 \leq 2\left\|\mathbb{E}[\Sigma^{-1}(P+Q)\Sigma^{-1}\nabla F_1(\theta^*)]\right\|_2^2 + O(n^{-3}). \tag{32}$$

We focus on bounding the remaining expectation. We have the following series of inequalities:

$$\begin{aligned}
\left\|\mathbb{E}[\Sigma^{-1}(P+Q)\Sigma^{-1}\nabla F_1(\theta^*)]\right\|_2 &\overset{(i)}{\leq} \mathbb{E}\left[|||\Sigma^{-1}(P+Q)|||_2 \left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2\right] \\
&\overset{(ii)}{\leq} \left(\mathbb{E}\left[|||\Sigma^{-1}(P+Q)|||_2^2\right]\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right]\right)^{\frac{1}{2}} \\
&\overset{(iii)}{\leq} \left(2\mathbb{E}\left[|||\Sigma^{-1}P|||_2^2 + |||\Sigma^{-1}Q|||_2^2\right]\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right]\right)^{\frac{1}{2}}.
\end{aligned}$$

Here step (i) follows from Jensen's inequality and the fact that $\|Ax\|_2 \leq |||A|||_2 \|x\|_2$; step (ii) uses the Cauchy-Schwarz inequality; and step (iii) follows from the fact that $(a+b)^2 \leq 2a^2+2b^2$. We have already bounded the first two terms in the product in our proofs; in particular, Lemma 7 guarantees that $\mathbb{E}[|||P|||_2^2] \leq CH\log d/n$, while

$$\mathbb{E}[|||Q|||_2^2] \leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n L(X_i)^4\right]^{\frac{1}{2}} \mathbb{E}[\|\theta_1-\theta^*\|_2^4]^{\frac{1}{2}} \leq C\frac{L^2G^2}{(1-\rho)^2\lambda^2}\cdot n^{-1}$$

for some numerical constant C (recall Lemma 9). Summarizing our bounds on $|||P|||_2$ and $|||Q|||_2$, we have

$$\begin{aligned}
&\left\|\mathbb{E}\left[\Sigma^{-1}(P+Q)\Sigma^{-1}\nabla F_1(\theta^*)\right]\right\|_2^2 \\
&\leq 2\left|||\Sigma^{-1}|||_2^2\left(\frac{2H^2(\log d+1)}{n} + 2C\frac{L^2G^2}{(1-\rho)^2\lambda^2 n} + O(n^{-2})\right)\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right].
\end{aligned} \tag{33}$$

From Assumption 3 we know that $\mathbb{E}[\|\nabla F_1(\theta^*)\|_2^2] \leq G^2/n$ and $\left|||\Sigma^{-1}|||_2 \leq 1/\lambda$, and hence we can further simplify the bound (33) to obtain

$$\begin{aligned}
\|\mathbb{E}[\theta_1-\theta^*]\|_2^2 &\leq \frac{C}{\lambda^2}\left(\frac{H^2\log d + L^2G^2/\lambda^2(1-\rho)^2}{n}\right)\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right] + O(n^{-3}) \\
&= \frac{C}{\lambda^2}\left(\frac{H^2\log d + L^2G^2/\lambda^2(1-\rho)^2}{n^2}\right)\mathbb{E}\left[\left\|\Sigma^{-1}\nabla f(\theta^*;X)\right\|_2^2\right] + O(n^{-3})
\end{aligned}$$

for some numerical constant C, where we have applied our earlier inequality (32). Noting that we may (without loss of generality) take $\rho < \frac{1}{2}$, then applying this inequality with the bound (27) on $\mathbb{E}[\|\theta_1-\theta^*\|_2^2]$ we previously proved to our decomposition (19) completes the proof.

Appendix C. Proof of Theorem 4

Our proof of Theorem 4 begins with a simple inequality that mimics our first inequality (19) in the proof of Theorem 1. Recall the definitions of the averaged vector $\bar{\theta}_1$ and subsampled averaged vector $\bar{\theta}_2$. Let θ_1 denote the minimizer of the (an arbitrary) empirical risk F_1, and θ_2 denote the minimizer of the resampled empirical risk F_2 (from the same samples as θ_1). Then we have

$$\mathbb{E}\left[\left\|\frac{\bar{\theta}_1 - r\bar{\theta}_2}{1-r} - \theta^*\right\|_2^2\right] \le \left\|\mathbb{E}\left[\frac{\theta_1 - r\theta_2}{1-r} - \theta^*\right]\right\|_2^2 + \frac{1}{m}\mathbb{E}\left[\left\|\frac{\theta_1 - r\theta_2}{1-r} - \theta^*\right\|_2^2\right]. \tag{34}$$

Thus, parallel to our proof of Theorem 1, it suffices to bound the two terms in the decomposition (34) separately. Specifically, we prove the following two lemmas.

Lemma 10 *Under the conditions of Theorem 4,*

$$\left\|\mathbb{E}\left[\frac{\theta_1 - r\theta_2}{1-r} - \theta^*\right]\right\|_2^2 \le O(1)\frac{1}{r(1-r)^2}\left(\frac{M^2G^6}{\lambda^6} + \frac{G^4L^2}{\lambda^4}d\log d\right)\frac{1}{n^3}. \tag{35}$$

Lemma 11 *Under the conditions of Theorem 4,*

$$\mathbb{E}\left[\|\theta_1 - \theta^* - r(\theta_2 - \theta^*)\|_2^2\right] \le (2+3r)\mathbb{E}\left[\left\|\nabla^2 F_0(\theta^*)^{-1}\nabla F_1(\theta^*)\right\|_2^2\right] + O(n^{-2}) \tag{36}$$

In conjunction, Lemmas 10 and 11 coupled with the decomposition (34) yield the desired claim. Indeed, applying each of the lemmas to the decomposition (34), we see that

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{\bar{\theta}_1 - r\bar{\theta}_2}{1-r} - \theta^*\right\|_2^2\right] &\le \frac{2+3r}{(1-r)^2 m}\mathbb{E}\left[\left\|\nabla^2 F_0(\theta^*)^{-1}\nabla F_1(\theta^*)\right\|_2^2\right] \\ &\quad + O\left(\frac{1}{(1-r)^2}m^{-1}n^{-2}\right) + O\left(\frac{1}{r(1-r)^2}n^{-3}\right),\end{aligned}$$

which is the statement of Theorem 4.

The remainder of our argument is devoted to establishing Lemmas 10 and 11. Before providing their proofs (in Appendices C.3 and C.4 respectively), we require some further set-up and auxiliary results. Throughout the rest of the proof, we use the notation

$$Y = Y' + \mathcal{R}_k$$

for some random variables Y and Y' to mean that there exists a random variable Z such that $Y = Y' + Z$ and $\mathbb{E}[\|Z\|_2^2] = O(n^{-k})$.[1] The symbol $\mathcal{R}_k$ may indicate different random variables throughout a proof and is notational shorthand for a moment-based big-O notation. We also remark that if we have $\mathbb{E}[\|Z\|_2^2] = O(a^k n^{-k})$, we have $Z = a^{k/2}\mathcal{R}_k$, since $(a^{k/2})^2 = a^k$. For shorthand, we also say that $\mathbb{E}[Z] = O(h(n))$ if $\|\mathbb{E}[Z]\|_2 = O(h(n))$, which implies that if $Z = \mathcal{R}_k$ then $\mathbb{E}[Z] = O(n^{-k/2})$, since

$$\|\mathbb{E}[Z]\|_2 \le \sqrt{\mathbb{E}[\|Z\|_2^2]} = O(n^{-k/2}).$$

1. Formally, in our proof this will mean that there exist random vectors Y, Y', and Z that are measurable with respect to the σ-field $\sigma(X_1, \ldots, X_n)$, where $Y = Y' + Z$ and $\mathbb{E}[\|Z\|_2^2] = O(n^{-k})$.

C.1 Optimization Error Expansion

In this section, we derive a sharper asymptotic expansion of the optimization errors $\theta_1 - \theta^*$. Recall our definition of the Kronecker product $\otimes$, where for vectors u, v we have $u \otimes v = uv^\top$. With this notation, we have the following expansion of $\theta_1 - \theta^*$. In these lemmas, $\mathcal{R}_3$ denotes a vector Z for which $\mathbb{E}[\|Z\|_2^2] \le cn^{-3}$ for a numerical constant c.

Lemma 12 *Under the conditions of Theorem 4, we have*

$$\begin{aligned}\theta_1 - \theta^* = &-\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(\nabla^2 F_1(\theta^*) - \Sigma)\Sigma^{-1}\nabla F_1(\theta^*) \\ &- \Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*)) \otimes (\Sigma^{-1}\nabla F_1(\theta^*))\right) \\ &+ \left(M^2G^6/\lambda^6 + G^4L^2 d\log(d)/\lambda^4\right)\mathcal{R}_3.\end{aligned} \tag{37}$$

We prove Lemma 12 in Appendix G. The lemma requires careful moment control over the expansion $\theta_1 - \theta^*$, leading to some technical difficulty, but is similar in spirit to the results leading to Theorem 1.

An immediately analogous result to Lemma 12 follows for our sub-sampled estimators. Since we use $\lceil rn \rceil$ samples to compute θ_2, the second level estimator, we find

Lemma 13 *Under the conditions of Theorem 4, we have*

$$\begin{aligned}\theta_2 - \theta^* = &-\Sigma^{-1}\nabla F_2(\theta^*) + \Sigma^{-1}(\nabla^2 F_2(\theta^*) - \Sigma)\Sigma^{-1}\nabla F_2(\theta^*) \\ &- \Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_2(\theta^*)) \otimes (\Sigma^{-1}\nabla F_2(\theta^*))\right) \\ &+ r^{-\frac{3}{2}}\left(M^2G^6/\lambda^6 + G^4L^2 d\log(d)/\lambda^4\right)\mathcal{R}_3.\end{aligned}$$

C.2 Bias Correction

Now that we have given Taylor expansions that describe the behaviour of $\theta_1 - \theta^*$ and $\theta_2 - \theta^*$, we can prove Lemmas 10 and 11 (though, as noted earlier, we defer the proof of Lemma 11 to Appendix C.4). The key insight is that expectations of terms involving $\nabla F_2(\theta^*)$ are nearly the same as expectations of terms involving $\nabla F_1(\theta^*)$, except that some corrections for the sampling ratio r are necessary.

We begin by noting that

$$\frac{\theta_1 - r\theta_2}{1-r} - \theta^* = \frac{\theta_1 - \theta^*}{1-r} - r\frac{\theta_2 - \theta^*}{1-r}. \tag{38}$$

In Lemmas 12 and 13, we derived expansions for each of the right hand side terms, and since

$$\mathbb{E}[\Sigma^{-1}\nabla F_1(\theta^*)] = 0 \quad \text{and} \quad \mathbb{E}[\Sigma^{-1}\nabla F_2(\theta^*)] = 0,$$

Lemmas 12 and 13 coupled with the rewritten correction (38) yield

$$\begin{aligned}\mathbb{E}[\theta_1 - \theta^* - r(\theta_2 - \theta^*)] = &-r\mathbb{E}[\Sigma^{-1}(\nabla^2 F_2(\theta^*) - \Sigma)\Sigma^{-1}\nabla F_2(\theta^*)] \\ &+ \mathbb{E}[\Sigma^{-1}(\nabla^2 F_1(\theta^*) - \Sigma)\Sigma^{-1}\nabla F_1(\theta^*)] \\ &+ r\mathbb{E}[\Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_2(\theta^*)) \otimes (\Sigma^{-1}\nabla F_2(\theta^*))\right)] \\ &- \mathbb{E}[\Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*)) \otimes (\Sigma^{-1}\nabla F_1(\theta^*))\right)] \\ &+ O(1)r^{-1/2}\left(M^2G^6/\lambda^6 + G^4L^2 d\log(d)/\lambda^4\right)n^{-3/2}.\end{aligned} \tag{39}$$

Here the remainder terms follow because of the $r^{-3/2}\mathcal{R}_3$ term on $\theta_2 - \theta^*$.

C.3 Proof of Lemma 10

To prove the claim in the lemma, it suffices to show that

$$r\mathbb{E}[\Sigma^{-1}(\nabla^2 F_2(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_2(\theta^*)] = \mathbb{E}[\Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_1(\theta^*)] \tag{40}$$

and

$$\begin{aligned} &r\mathbb{E}[\Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_2(\theta^*))\otimes(\Sigma^{-1}\nabla F_2(\theta^*))\right)] \\ &\quad = \mathbb{E}[\Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*))\right)] \end{aligned} \tag{41}$$

Indeed, these two claims combined with the expansion (39) yield the bound (35) in Lemma 10 immediately.

We first consider the difference (40). To make things notationally simpler, we define functions $A : \mathcal{X} \to \mathbb{R}^{d\times d}$ and $B : \mathcal{X} \to \mathbb{R}^d$ via $A(x) := \Sigma^{-1}(\nabla^2 f(\theta^*;x)-\Sigma)$ and $B(x) := \Sigma^{-1}\nabla f(\theta^*;x)$. If we let $S_1 = \{X_1,\ldots,X_n\}$ be the original samples and $S_2 = \{Y_1,\ldots,Y_{rn}\}$ be the subsampled data set, we must show

$$r\mathbb{E}\left[\frac{1}{(rn)^2}\sum_{i,j}^{rn} A(Y_i)B(Y_j)\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}^{n} A(X_i)B(X_j)\right].$$

Since the Y_i are sampled without replacement (i.e., from P directly), and $\mathbb{E}[A(X_i)] = 0$ and $\mathbb{E}[B(X_i)] = 0$, we find that $\mathbb{E}[A(Y_i)B(Y_j)] = 0$ for $i \neq j$, and thus

$$\sum_{i,j}^{rn}\mathbb{E}[A(Y_i)B(Y_j)] = \sum_{i=1}^{rn}\mathbb{E}[A(Y_i)B(Y_i)] = rn\mathbb{E}[A(Y_1)B(Y_1)].$$

In particular, we see that the equality (40) holds:

$$\begin{aligned} \frac{r}{(rn)^2}\sum_{i,j}^{rn}\mathbb{E}[A(Y_i)B(Y_j)] = \frac{r}{rn}\mathbb{E}[A(Y_1)B(Y_1)] &= \frac{1}{n}\mathbb{E}[A(X_1)B(X_1)] \\ &= \frac{1}{n^2}\sum_{i,j}^{n}\mathbb{E}[A(X_i)B(X_j)]. \end{aligned}$$

The statement (41) follows from analogous arguments.

C.4 Proof of Lemma 11

The proof of Lemma 11 follows from that of Lemmas 12 and 13. We first claim that

$$\theta_1 - \theta^* = -\Sigma^{-1}\nabla F_1(\theta^*) + \mathcal{R}_2 \quad \text{and} \quad \theta_2 - \theta^* = -\Sigma^{-1}\nabla F_2(\theta^*) + r^{-1}\mathcal{R}_2. \tag{42}$$

The proofs of both claims similar, so we focus on proving the second statement. Using the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and Lemma 13, we see that

$$\begin{aligned} \mathbb{E}\left[\left\|\theta_2 - \theta^* + \Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^2\right] &\leq 3\mathbb{E}\left[\left\|\Sigma^{-1}(\nabla^2 F_2(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^2\right] \\ &\quad + 3\mathbb{E}\left[\left\|\Sigma^{-1}\nabla^3 F_0(\theta^*)\left((\Sigma^{-1}\nabla F_2(\theta^*))\otimes(\Sigma^{-1}\nabla F_2(\theta^*))\right)\right\|_2^2\right] \\ &\quad + 3r^{-3}O(n^{-3}). \end{aligned} \tag{43}$$

We now bound the first two terms in inequality (43). Applying the Cauchy-Schwarz inequality and Lemma 7, the first term can be upper bounded as

$$\begin{aligned}
&\mathbb{E}\left[\left\|\Sigma^{-1}(\nabla^2 F_2(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^2\right] \\
&\leq \left(\mathbb{E}\left[\left|\left|\left|\Sigma^{-1}(\nabla^2 F_2(\theta^*)-\Sigma)\right|\right|\right|_2^4\right]\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^4\right]\right)^{1/2} \\
&= \left(r^{-2}\right)O(\log^2(d)n^{-2})\cdot r^{-2}O(n^{-2})\big)^{1/2} \;=\; r^{-2}O(n^{-2}),
\end{aligned}$$

where the order notation subsumes the logarithmic factor in the dimension. Since $\nabla^3 F_0(\theta^*):\mathbb{R}^{d^2}\to\mathbb{R}^d$ is linear, the second term in the inequality (43) may be bounded completely analogously as it involves the outer product $\Sigma^{-1}\nabla F_2(\theta^*)\otimes\Sigma^{-1}\nabla F_2(\theta^*)$. Recalling the bound (43), we have thus shown that

$$\mathbb{E}\left[\left\|\theta_2-\theta^*+\Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^2\right]=r^{-2}O(n^{-2}),$$

or $\theta_2-\theta^*=-\Sigma^{-1}\nabla F_2(\theta^*)+r^{-1}\mathcal{R}_2$. The proof of the first equality in Equation (42) is entirely analogous.

We now apply the equalities (42) to obtain the result of the lemma. We have

$$\mathbb{E}\left[\|\theta_1-\theta^*-r(\theta_2-\theta^*)\|_2^2\right]=\mathbb{E}\left[\left\|-\Sigma^{-1}\nabla F_1(\theta^*)+r\Sigma^{-1}\nabla F_2(\theta^*)+\mathcal{R}_2\right\|_2^2\right].$$

Using the inequality $(a+b)^2\leq(1+\eta)a^2+(1+1/\eta)b^2$ for any $\eta\geq 0$, we have

$$\begin{aligned}
(a+b+c)^2 &\leq (1+\eta)a^2+(1+1/\eta)(b+c)^2 \\
&\leq (1+\eta)a^2+(1+1/\eta)(1+\alpha)b^2+(1+1/\eta)(1+1/\alpha)c^2
\end{aligned}$$

for any $\eta,\alpha\geq 0$. Taking $\eta=1$ and $\alpha=1/2$, we obtain $(a+b+c)^2\leq 2a^2+3b^2+6c^2$, so applying the triangle inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\|\theta_1-\theta^*-r(\theta_2-\theta^*)\|_2^2\right] &= \mathbb{E}\left[\left\|-\Sigma^{-1}\nabla F_1(\theta^*)+r\Sigma^{-1}\nabla F_2(\theta^*)+\mathcal{R}_2\right\|_2^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right]+3r^2\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^2\right]+O(n^{-2}).
\end{aligned} \tag{44}$$

Since F_2 is a sub-sampled version of F_1, algebraic manipulations yield

$$\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_2(\theta^*)\right\|_2^2\right]=\frac{n}{rn}\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right]=\frac{1}{r}\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^2\right]. \tag{45}$$

Combining equations (44) and (45), we obtain the desired bound (36).

Appendix D. Proof of Theorem 5

We begin by recalling that if θ^n denotes the output of performing stochastic gradient on one machine, then from the inequality (19) we have the upper bound

$$\mathbb{E}[\left\|\overline{\theta}^n-\theta^*\right\|_2^2]\leq\frac{1}{m}\mathbb{E}[\|\theta^n-\theta^*\|_2^2]+\|\mathbb{E}[\theta^n-\theta^*]\|_2^2.$$

To prove the error bound (15), it thus suffices to prove the inequalities

$$\mathbb{E}[\|\theta^n - \theta^*\|_2^2] \le \frac{\alpha G^2}{\lambda^2 n}, \quad \text{and} \tag{46}$$

$$\|\mathbb{E}[\theta^n - \theta^*]\|_2^2 \le \frac{\beta^2}{n^{3/2}}. \tag{47}$$

Before proving the theorem, we introduce some notation and a few preliminary results. Let $g_t = \nabla f(\theta^t; X_t)$ be the gradient of the t^{th} sample in stochastic gradient descent, where we consider running SGD on a single machine. We also let

$$\Pi(v) := \operatorname*{argmin}_{\theta \in \Theta} \left\{ \|\theta - v\|_2^2 \right\}$$

denote the projection of the point v onto the domain Θ.

We now state a known result, which gives sharp rates on the convergence of the iterates $\{\theta^t\}$ in stochastic gradient descent.

Lemma 14 (Rakhlin et al., 2012) *Assume that $\mathbb{E}[\|g_t\|_2^2] \le G^2$ for all t. Choosing $\eta_t = \frac{c}{\lambda t}$ for some $c \ge 1$, for any $t \in \mathbb{N}$ we have*

$$\mathbb{E}\left[\left\|\theta^t - \theta^*\right\|_2^2\right] \le \frac{\alpha G^2}{\lambda^2 t} \quad \textit{where} \quad \alpha = 4c^2.$$

With these ingredients, we can now turn to the proof of Theorem 5. Lemma 14 gives the inequality (46), so it remains to prove that $\overline{\theta}^n$ has the smaller bound (47) on its bias. To that end, recall the neighborhood $U_\rho \subset \Theta$ in Assumption 5, and note that

$$\begin{aligned}
\theta^{t+1} - \theta^* &= \Pi(\theta^t - \eta_t g_t - \theta^*) \\
&= \theta^t - \eta_t g_t - \theta^* + 1_{\left(\theta^{t+1} \notin U_\rho\right)} \left(\Pi(\theta^t - \eta_t g_t) - (\theta^t - \eta_t g_t)\right)
\end{aligned}$$

since when $\theta \in U_\rho$, we have $\Pi(\theta) = \theta$. Consequently, an application of the triangle inequality gives

$$\left\|\mathbb{E}[\theta^{t+1} - \theta^*]\right\|_2 \le \left\|\mathbb{E}[\theta^t - \eta_t g_t - \theta^*]\right\|_2 + \mathbb{E}[\left\|(\Pi(\theta^t - \eta_t g_t) - (\theta^t - \eta_t g_t))1(\theta^{t+1} \notin U_\rho)\right\|_2].$$

By the definition of the projection and the fact that $\theta^t \in \Theta$, we additionally have

$$\left\|\Pi(\theta^t - \eta_t g_t) - (\theta^t - \eta_t g_t)\right\|_2 \le \left\|\theta^t - (\theta^t - \eta_t g_t))\right\|_2 \le \eta_t \left\|g_t\right\|_2.$$

Thus, by applying Hölder's inequality (with the conjugate choices $(p, q) = (4, \frac{4}{3})$) and Assumption 5, we have

$$\begin{aligned}
\left\|\mathbb{E}[\theta^{t+1} - \theta^*]\right\|_2 &\le \left\|\mathbb{E}[\theta^t - \eta_t g_t - \theta^*]\right\|_2 + \eta_t \mathbb{E}[\|g_t\|_2 1_{\left(\theta^{t+1} \notin U_\rho\right)}] \\
&\le \left\|\mathbb{E}[\theta^t - \eta_t g_t - \theta^*]\right\|_2 + \eta_t \sqrt[4]{\mathbb{E}[\|g_t\|_2^4]} \left(\mathbb{E}[1^{4/3}_{\left(\theta^t \notin U_\rho\right)}]\right)^{3/4} \\
&\le \left\|\mathbb{E}[\theta^t - \eta_t g_t - \theta^*]\right\|_2 + \eta_t G \left(\mathbb{P}(\theta^t \notin U_\rho)\right)^{3/4} \\
&\le \left\|\mathbb{E}[\theta^t - \eta_t g_t - \theta^*]\right\|_2 + \eta_t G \left(\frac{\mathbb{E}\left\|\theta^{t+1} - \theta^*\right\|_2^2}{\rho^2}\right)^{3/4},
\end{aligned} \tag{48}$$

the inequality (48) following from an application of Markov's inequality. By applying Lemma 14, we finally obtain

$$\begin{aligned}\left\|\mathbb{E}[\theta^{t+1}-\theta^*]\right\|_2 &\leq \left\|\mathbb{E}[\theta^t-\eta_t g_t-\theta^*]\right\|_2+\eta_t G\left(\frac{\alpha G^2}{\lambda^2\rho^2 t}\right)^{3/4}\\ &= \left\|\mathbb{E}[\theta^t-\eta_t g_t-\theta^*]\right\|_2+\frac{c\alpha^{3/4}G^{5/2}}{\lambda^{5/2}\rho^{3/2}}\cdot\frac{1}{t^{7/4}}. \end{aligned} \tag{49}$$

Now we turn to controlling the rate at which $\theta^t-\eta_t g_t$ goes to zero. Let $f_t(\cdot)=f(\cdot;X_t)$ be shorthand for the loss evaluated on the t^{th} data point. By defining

$$r_t = g_t-\nabla f_t(\theta^*)-\nabla^2 f_t(\theta^*)(\theta^t-\theta^*),$$

a bit of algebra yields

$$g_t=\nabla f_t(\theta^*)+\nabla^2 f_t(\theta^*)(\theta^t-\theta^*)+r_t.$$

Since θ^t belongs to the σ-field of $X_1,\ldots,X_{t-1}$, the Hessian $\nabla^2 f_t(\theta^*)$ is (conditionally) independent of θ^t and

$$\mathbb{E}[g_t]=\nabla^2 F_0(\theta^*)\mathbb{E}[\theta^t-\theta^*]+\mathbb{E}[r_t 1_{\left(\theta^t\in U_\rho\right)}]+\mathbb{E}[r_t 1_{\left(\theta^t\notin U_\rho\right)}]. \tag{50}$$

If $\theta^t\in U_\rho$, then Taylor's theorem implies that r_t is the Lagrange remainder

$$r_t=(\nabla^2 f_t(\theta')-\nabla^2 f_t(\theta^*))(\theta'-\theta^*),$$

where $\theta'=\kappa\theta^t+(1-\kappa)\theta^*$ for some $\kappa\in[0,1]$. Applying Assumption 5 and Hölder's inequality, we find that since θ^t is conditionally independent of X_t,

$$\begin{aligned}\mathbb{E}\left[\left\|r_t 1_{\left(\theta^t\in U_\rho\right)}\right\|_2\right] &\leq \mathbb{E}\left[\left|\left|\left|\nabla^2 f(\theta';X_t)-\nabla^2 f(\theta^*;X_t)\right|\right|\right|\left\|\theta^t-\theta^*\right\|_2 1_{\left(\theta^t\in U_\rho\right)}\right]\\ &\leq \mathbb{E}\left[L(X_t)\left\|\theta^t-\theta^*\right\|_2^2\right]=\mathbb{E}[L(X_t)]\mathbb{E}[\left\|\theta^t-\theta^*\right\|_2^2]\\ &\leq L\mathbb{E}\left[\left\|\theta^t-\theta^*\right\|_2^2\right]\leq\frac{\alpha L G^2}{\lambda^2 t}.\end{aligned}$$

On the other hand, when $\theta^t\notin U_\rho$, we have the following sequence of inequalities:

$$\begin{aligned}\mathbb{E}\left[\left\|r_t 1_{\left(\theta^t\notin U_\rho\right)}\right\|_2\right] &\overset{(i)}{\leq} \sqrt[4]{\mathbb{E}[\|r_t\|_2^4]}\left(\mathbb{P}(\theta^t\notin U_\rho)\right)^{3/4}\\ &\overset{(ii)}{\leq} \sqrt[4]{3^3\left(\mathbb{E}[\|g_t\|_2^4]+\mathbb{E}[\|\nabla f_t(\theta^*)\|_2^4]+\mathbb{E}[\|\nabla^2 f_t(\theta^*)(\theta^t-\theta^*)\|_2^4]\right)}\left(\mathbb{P}(\theta^t\notin U_\rho)\right)^{3/4}\\ &\leq 3^{3/4}\sqrt[4]{G^4+G^4+H^4R^4}\left(\mathbb{P}(\theta^t\notin U_\rho)\right)^{3/4}\\ &\overset{(iii)}{\leq} 3(G+HR)\left(\frac{\alpha G^2}{\lambda^2\rho^2 t}\right)^{3/4}.\end{aligned}$$

Here step (i) follows from Hölder's inequality (again applied with the conjugates $(p,q)=(4,\frac{4}{3})$); step (ii) follows from Jensen's inequality, since $(a+b+c)^4\leq 3^3(a^4+b^4+c^4)$; and step (iii) follows

from Markov's inequality, as in the bounds (48) and (49). Combining our two bounds on r_t, we find that

$$\mathbb{E}[\|r_t\|_2] \leq \frac{\alpha L G^2}{\lambda^2 t} + \frac{3\alpha^{3/4} G^{3/2}(G+HR)}{\lambda^{3/2}\rho^{3/2}} \cdot \frac{1}{t^{3/4}}. \tag{51}$$

By combining the expansion (50) with the bound (51), we find that

$$\begin{aligned}\left\|\mathbb{E}[\theta^t - \eta_t g_t - \theta^*]\right\|_2 &= \left\|\mathbb{E}[(I - \eta_t \nabla^2 F_0(\theta^*))(\theta^t - \theta^*) + \eta_t r_t]\right\|_2 \\ &\leq \left\|\mathbb{E}[(I - \eta_t \nabla^2 F_0(\theta^*))(\theta^t - \theta^*)]\right\|_2 + \frac{c\alpha L G^2}{\lambda^3 t^2} + \frac{3c\alpha^{3/4} G^{3/2}(G+HR)}{\lambda^{5/2}\rho^{3/2}} \cdot \frac{1}{t^{7/4}}.\end{aligned}$$

Using the earlier bound (49), this inequality then yields

$$\left\|\mathbb{E}[\theta^{t+1} - \theta^*]\right\|_2 \leq \left|\left|\left|I - \eta_t \nabla^t F_0(\theta^*)\right|\right|\right| \left\|\mathbb{E}[\theta^t - \theta^*]\right\|_2 + \frac{c\alpha^{3/4} G^{3/2}}{\lambda^{5/2} t^{7/4}} \left(\frac{\alpha^{1/4} L G^{1/2}}{\lambda^{1/2} t^{1/4}} + \frac{4G + HR}{\rho^{3/2}}\right).$$

We now complete the proof via an inductive argument using our immediately preceding bounds. Our reasoning follows a similar induction given by Rakhlin et al. (2012). First, note that by strong convexity and our condition that $\left|\left|\left|\nabla^2 F_0(\theta^*)\right|\right|\right| \leq H$, we have

$$\left|\left|\left|I - \eta_t \nabla^2 F_0(\theta^*)\right|\right|\right| = 1 - \eta_t \lambda_{\min}(\nabla^2 F_0(\theta^*) \leq 1 - \eta_t \lambda$$

whenever $1 - \eta_t H \geq 0$. Define $\tau_0 = \lceil cH/\lambda \rceil$; then for $t \geq t_0$ we obtain

$$\left\|\mathbb{E}[\theta^{t+1} - \theta^*]\right\|_2 \leq (1 - c/t) \left\|\mathbb{E}[\theta^t - \theta^*]\right\|_2 + \frac{1}{t^{7/4}} \cdot \frac{c\alpha^{3/4} G^{3/2}}{\lambda^{5/2}} \left(\frac{\alpha^{1/4} L G^{1/2}}{\lambda^{1/2} t^{1/4}} + \frac{4G + HR}{\rho^{3/2}}\right). \tag{52}$$

For shorthand, we define two intermediate variables

$$a_t = \left\|\mathbb{E}(\theta^t - \theta^*)\right\|_2 \quad \text{and} \quad b = \frac{c\alpha^{3/4} G^{3/2}}{\lambda^{5/2}} \left(\frac{\alpha^{1/4} L G^{1/2}}{\lambda^{1/2}} + \frac{4G + HR}{\rho^{3/2}}\right).$$

Inequality (52) then implies the inductive relation $a_{t+1} \leq (1 - c/t) a_t + b/t^{7/4}$. Now we show that by defining $\beta = \max\{\tau_0 R, b/(c-1)\}$, we have $a_t \leq \beta/t^{3/4}$. Indeed, it is clear that $a_1 \leq \tau_0 R$. Using the inductive hypothesis, we then have

$$a_{t+1} \leq \frac{(1 - c/t)\beta}{t^{3/4}} + \frac{b}{t^{7/4}} = \frac{\beta(t-1)}{t^{7/4}} - \frac{\beta(c-1) - b}{t^2} \leq \frac{\beta(t-1)}{t^{7/4}} \leq \frac{\beta}{(t+1)^{3/4}}.$$

This completes the proof of the inequality (47). ■

D.0.1 REMARK

If we assume kth moment bounds instead of 4th, that is, $\mathbb{E}[\left|\left|\left|\nabla^2 f(\theta^*; X)\right|\right|\right|_2^k] \leq H^k$ and $\mathbb{E}[\|g_t\|_2^k] \leq G^k$, we find the following analogue of the bound (52):

$$\begin{aligned}\left\|\mathbb{E}[\theta^{t+1} - \theta^*]\right\|_2 \leq (1 - c/t) &\left\|\mathbb{E}[\theta^t - \theta^*]\right\|_2 \\ &+ \frac{1}{t^{\frac{2k-1}{k}}} \cdot \frac{c\alpha^{\frac{k-1}{k}} G^{\frac{2k-2}{k}}}{\lambda^{\frac{3k-2}{k}}} \left[\frac{\left(54^{1/k} + 1\right) G + 54^{1/k} HR}{\rho^{\frac{2k-2}{k}}} + \frac{\alpha^{1/k} L G^{2/k}}{\lambda^{2/k} t^{1/k}}\right].\end{aligned}$$

In this case, if we define

$$b = \frac{c\alpha^{\frac{k-1}{k}} G^{\frac{2k-2}{k}}}{\lambda^{\frac{3k-2}{k}}} \left[\frac{\left(54^{1/k}+1\right) G + 54^{1/k} HR}{\rho^{\frac{2k-2}{k}}} + \frac{\alpha^{1/k} L G^{2/k}}{\lambda^{2/k}} \right] \quad \text{and} \quad \beta = \max\left\{ \tau_0 R, \frac{b}{c-1} \right\},$$

we have the same result except we obtain the bound $\|\mathbb{E}[\theta^n - \theta^*]\|_2^2 \leq \beta^2 / n^{\frac{2k-2}{k}}$.

Appendix E. Proof of Lemma 6

We first prove that under the conditions given in the lemma statement, the function F_1 is $(1-\rho)\lambda$-strongly convex over the ball $U := \left\{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_2 < \delta_\rho\right\}$ around θ^*. Indeed, fix $\gamma \in U$, then use the triangle inequality to conclude that

$$\begin{aligned} \left|\left|\left|\nabla^2 F_1(\gamma) - \nabla^2 F_0(\theta^*)\right|\right|\right|_2 &\leq \left|\left|\left|\nabla^2 F_1(\gamma) - \nabla^2 F_1(\theta^*)\right|\right|\right|_2 + \left|\left|\left|\nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*)\right|\right|\right|_2 \\ &\leq L \|\gamma - \theta^*\|_2 + \frac{\rho\lambda}{2}. \end{aligned}$$

Here we used Assumption 3 on the first term and the fact that the event $\mathcal{E}_1$ holds on the second. By our choice of $\delta_\rho \leq \rho\lambda/4L$, this final term is bounded by $\lambda\rho$. In particular, we have

$$\nabla^2 F_0(\theta^*) \succeq \lambda I \quad \text{so} \quad \nabla^2 F_1(\gamma) \succeq \lambda I - \rho\lambda I = (1-\rho)\lambda I,$$

which proves that F_1 is $(1-\rho)\lambda$-strongly convex on the ball U.

In order to prove the conclusion of the lemma, we argue that since F_1 is (locally) strongly convex, if the function F_1 has small gradient at the point θ^*, it must be the case that the minimizer θ_1 of F_1 is near θ^*. Then we can employ reasoning similar to standard analyses of optimality for globally strongly convex functions (e.g., Boyd and Vandenberghe, 2004, Chapter 9). By definition of (the local) strong convexity on the set U, for any $\theta' \in \Theta$, we have

$$F_1(\theta') \geq F_1(\theta^*) + \left\langle \nabla F_1(\theta^*), \theta' - \theta^* \right\rangle + \frac{(1-\rho)\lambda}{2} \min\left\{ \left\|\theta^* - \theta'\right\|_2^2, \delta_\rho^2 \right\}.$$

Rewriting this inequality, we find that

$$\begin{aligned} \min\left\{ \left\|\theta^* - \theta'\right\|_2^2, \delta_\rho^2 \right\} &\leq \frac{2}{(1-\rho)\lambda} \left[F_1(\theta') - F_1(\theta^*) + \left\langle \nabla F_1(\theta^*), \theta' - \theta^* \right\rangle \right] \\ &\leq \frac{2}{(1-\rho)\lambda} \left[F_1(\theta') - F_1(\theta^*) + \|\nabla F_1(\theta^*)\|_2 \left\|\theta' - \theta^*\right\|_2 \right]. \end{aligned}$$

Dividing each side by $\|\theta' - \theta^*\|_2$, then noting that we may set $\theta' = \kappa\theta_1 + (1-\kappa)\theta^*$ for any $\kappa \in [0,1]$, we have

$$\min\left\{ \kappa \|\theta_1 - \theta^*\|_2, \frac{\delta_\rho^2}{\kappa \|\theta_1 - \theta^*\|_2} \right\} \leq \frac{2\left[F_1(\kappa\theta_1 + (1-\kappa)\theta^*) - F_1(\theta^*)\right]}{\kappa(1-\rho)\lambda \|\theta_1 - \theta^*\|_2} + \frac{2\|\nabla F_1(\theta^*)\|_2}{(1-\rho)\lambda}.$$

Of course, $F_1(\theta_1) < F_1(\theta^*)$ by assumption, so we find that for any $\kappa \in (0,1)$ we have the strict inequality

$$\min\left\{ \kappa \|\theta_1 - \theta^*\|_2, \frac{\delta_\rho^2}{\kappa \|\theta_1 - \theta^*\|_2} \right\} < \frac{2\|\nabla F_1(\theta^*)\|_2}{(1-\rho)\lambda} \leq \delta_\rho,$$

the last inequality following from the definition of $\mathcal{E}_2$. Since this holds for any $\kappa \in (0,1)$, if $\|\theta_1 - \theta^*\|_2 > \delta_\rho$, we may set $\kappa = \delta_\rho / \|\theta_1 - \theta^*\|_2$, which would yield a contradiction. Thus, we have $\|\theta_1 - \theta^*\|_2 \leq \delta_\rho$, and by our earlier inequalities,

$$\|\theta_1 - \theta^*\|_2^2 \leq \frac{2}{(1-\rho)\lambda}\left[F_1(\theta_1) - F_1(\theta^*) + \|\nabla F_1(\theta^*)\|_2 \|\theta_1 - \theta^*\|_2\right] \leq \frac{2\|\nabla F_1(\theta^*)\|_2}{(1-\rho)\lambda}\|\theta_1 - \theta^*\|_2.$$

Dividing by $\|\theta_1 - \theta^*\|_2$ completes the proof. ■

Appendix F. Moment Bounds

In this appendix, we state two useful moment bounds, showing how they combine to provide a proof of Lemma 7. The two lemmas are a vector and a non-commutative matrix variant of the classical Rosenthal inequalities. We begin with the case of independent random vectors:

Lemma 15 (de Acosta, 1981, Theorem 2.1) *Let $k \geq 2$ and X_i be a sequence of independent random vectors in a separable Banach space with norm $\|\cdot\|$ and $\mathbb{E}[\|X_i\|^k] < \infty$. There exists a finite constant C_k such that*

$$\mathbb{E}\left[\left|\left\|\sum_{i=1}^n X_i\right\| - \mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|\right]\right|^k\right] \leq C_k\left[\left(\sum_{i=1}^n \mathbb{E}[\|X_i\|^2]\right)^{k/2} + \sum_{i=1}^n \mathbb{E}[\|X_i\|^k]\right].$$

We say that a random matrix X is symmetrically distributed if X and $-X$ have the same distribution. For such matrices, we have:

Lemma 16 (Chen et al., 2012, Theorem A.1(2)) *Let $X_i \in \mathbb{R}^{d\times d}$ be independent and symmetrically distributed Hermitian matrices. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|^k\right]^{1/k} \leq \sqrt{2e\log d}\left\|\left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\right)^{1/2}\right\| + 2e\log d\left(\mathbb{E}[\max_i \|X_i\|^k]\right)^{1/k}.$$

Equipped with these two auxiliary results, we turn to our proof Lemma 7. To prove the first bound (24), let $2 \leq k \leq k_0$ and note that by Jensen's inequality, we have

$$\mathbb{E}[\|\nabla F_1(\theta^*)\|_2^k] \leq 2^{k-1}\mathbb{E}\left[\left|\,\|\nabla F_1(\theta^*)\|_2 - \mathbb{E}[\|\nabla F_1(\theta^*)\|_2]\right|^k\right] + 2^{k-1}\mathbb{E}\left[\|\nabla F_1(\theta^*)\|_2\right]^k.$$

Again applying Jensen's inequality, $\mathbb{E}[\|\nabla f(\theta^*;X)\|_2^2] \leq G^2$. Thus by recalling the definition $\nabla F_1(\theta^*) = \frac{1}{n}\sum_{i=1}^n \nabla f(\theta^*;X_i)$ and applying the inequality

$$\mathbb{E}[\|\nabla F_1(\theta^*)\|_2] \leq \mathbb{E}[\|\nabla F_1(\theta^*)\|_2^2]^{1/2} \leq n^{-1/2}G,$$

we see that Lemma 15 implies $\mathbb{E}\left[\|\nabla F_1(\theta^*)\|_2^k\right]$ is upper bounded by

$$2^{k-1}C_k\left[\left(\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\theta;X_i)\|_2^2]\right)^{k/2} + \frac{1}{n^k}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\theta^*;X_i)\|_2^k]\right] + 2^{k-1}\mathbb{E}[\|\nabla F_1(\theta^*)\|_2]^k$$
$$\leq 2^{k-1}\frac{C_k}{n^{k/2}}\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\theta^*;X_i)\|_2^2]\right)^{k/2} + \frac{1}{n^{k/2}}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\theta^*;X_i)\|_2^k]\right] + \frac{2^{k-1}G^k}{n^{k/2}}.$$

Applying Jensen's inequality yields

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f(\theta^*;X_i)\|_2^2]\right)^{k/2} \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\nabla f(\theta^*;X_i)\|_2^2]^{k/2} \leq G^k,$$

completes the proof of the inequality (24).

The proof of the bound (25) requires a very slightly more delicate argument involving symmetrization step. Define matrices $Z_i = \frac{1}{n}\left(\nabla^2 f(\theta^*;X_i) - \nabla^2 F_0(\theta^*)\right)$. If $\varepsilon_i \in \{\pm 1\}$ are i.i.d. Rademacher variables independent of Z_i, then for any integer k in the interval $[2,k_2]$, a standard symmetrization argument (e.g., Ledoux and Talagrand, 1991, Lemma 6.3) implies that

$$\mathbb{E}\left[\left|\left|\left|\sum_{i=1}^{n} Z_i\right|\right|\right|^k\right]^{1/k} \leq 2\mathbb{E}\left[\left|\left|\left|\sum_{i=1}^{n}\varepsilon_i Z_i\right|\right|\right|^k\right]^{1/k}.$$

Now we may apply Lemma 16, since the matrices $\varepsilon_i Z_i$ are Hermitian and symmetrically distributed; by expanding the definition of the Z_i, we find that

$$\mathbb{E}\left[\left|\left|\left|\nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*)\right|\right|\right|^k\right]^{1/k} \leq 5\sqrt{\log d}\left|\left|\left|\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[(\nabla^2 f(\theta;X_i) - \nabla^2 F_0(\theta^*))^2]\right)^{1/2}\right|\right|\right| + 4e\log d\left(n^{-k}\mathbb{E}[\max_i \left|\left|\left|\nabla^2 f(\theta^*;X_i) - \nabla^2 F_0(\theta^*)\right|\right|\right|^k]\right)^{1/k}.$$

Since the X_i are i.i.d., we have

$$\left|\left|\left|\left(\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[(\nabla^2 f(\theta;X_i) - \nabla^2 F_0(\theta^*))^2]\right)^{1/2}\right|\right|\right| = \left|\left|\left|n^{-1/2}\mathbb{E}\left[\left(\nabla^2 f(\theta^*;X) - \nabla^2 F_0(\theta^*)\right)^2\right]^{1/2}\right|\right|\right| \leq n^{-1/2}\mathbb{E}\left[\left|\left|\left|\nabla^2 f(\theta^*;X) - \nabla^2 F_0(\theta^*)\right|\right|\right|^2\right]^{1/2}$$

by Jensen's inequality, since $\left|\left|\left|A^{1/2}\right|\right|\right| = |||A|||^{1/2}$ for semidefinite A. Finally, noting that

$$\frac{1}{n^k}\mathbb{E}\left[\max_i \left|\left|\left|\nabla^2 f(\theta^*;X_i) - \nabla^2 F_0(\theta^*)\right|\right|\right|^k\right] \leq \frac{n}{n^k}\mathbb{E}\left[\left|\left|\left|\nabla^2 f(\theta^*;X) - \nabla^2 F_0(\theta^*)\right|\right|\right|^k\right] \leq n^{1-k}H^k$$

completes the proof of the second bound (25).

Appendix G. Proof of Lemma 12

The proof follows from a slightly more careful application of the Taylor expansion (21). The starting point in our proof is to recall the success events (20) and the joint event $\mathcal{E} := \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$. We begin by arguing that we may focus on the case where $\mathcal{E}$ holds. Let C denote the right hand side of the equality (37) except for the remainder $\mathcal{R}_3$ term. By Assumption 3, we follow the bound (26) (with $\min\{k_0,k_1,k_2\} \geq 8$) to find that

$$\mathbb{E}\left[1_{(\mathcal{E}^c)}\|\theta_1 - \theta^*\|_2^2\right] = O\left(R^2 n^{-4}\right),$$

so we can focus on the case where the joint event $\mathcal{E} = \mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$ does occur.

Defining $\Delta = \theta_1 - \theta^*$ for notational convenience, on $\mathcal{E}$ we have that for some $\kappa \in [0,1]$, with $\theta' = (1-\kappa)\theta_1 + \kappa\theta^*$,

$$\begin{aligned} 0 &= \nabla F_1(\theta^*) + \nabla^2 F_1(\theta^*)\Delta + \nabla^3 F_1(\theta')(\Delta \otimes \Delta) \\ &= \nabla F_1(\theta^*) + \nabla^2 F_0(\theta^*)\Delta + \nabla^3 F_0(\theta^*)(\Delta \otimes \Delta) \\ &\quad + (\nabla^2 F_1(\theta^*) - \nabla^2 F_0(\theta^*))\Delta + (\nabla^3 F_1(\theta') - \nabla^3 F_0(\theta^*))(\Delta \otimes \Delta). \end{aligned}$$

Now, we recall the definition $\Sigma = \nabla^2 F_0(\theta^*)$, the Hessian of the risk at the optimal point, and solve for the error Δ to see that

$$\begin{aligned} \Delta = &-\Sigma^{-1}\nabla F_1(\theta^*) - \Sigma^{-1}(\nabla^2 F_1(\theta^*) - \Sigma)\Delta - \Sigma^{-1}\nabla^3 F_1(\theta^*)(\Delta \otimes \Delta) \\ &+ \Sigma^{-1}(\nabla^3 F_0(\theta^*) - \nabla^3 F_1(\theta'))(\Delta \otimes \Delta) \end{aligned} \tag{53}$$

on the event $\mathcal{E}$. As we did in the proof of Theorem 1, specifically in deriving the recursive equality (28), we may apply the expansion (23) of $\Delta = \theta_1 - \theta^*$ to obtain a clean asymptotic expansion of Δ using (53). Recall the definition $P = \nabla^2 F_0(\theta^*) - \nabla^2 F_1(\theta^*)$ for shorthand here (as in the expansion (23), though we no longer require Q).

First, we claim that

$$1_{(\mathcal{E})}(\nabla^3 F_0(\theta^*) - \nabla^3 F_1(\theta'))(\Delta \otimes \Delta) = \left(M^2 G^6/\lambda^6 + G^4 L^2 d \log(d)/\lambda^4\right) \mathcal{R}_3. \tag{54}$$

To prove the above expression, we add and subtract $\nabla^3 F_1(\theta^*)$ (and drop $1_{(\mathcal{E})}$ for simplicity). We must control

$$(\nabla^3 F_0(\theta^*) - \nabla^3 F_1(\theta^*))(\Delta \otimes \Delta) + (\nabla^3 F_1(\theta^*) - \nabla^3 F_1(\theta'))(\Delta \otimes \Delta).$$

To begin, recall that $|||u \otimes v|||_2 = \left|\left|\left|uv^\top\right|\right|\right|_2 = \|u\|_2 \|v\|_2$. By Assumption 4, on the event $\mathcal{E}$ we have that $\nabla^3 F_1$ is $(1/n)\sum_{i=1}^n M(X_i)$-Lipschitz, so defining $M_n = (1/n)\sum_{i=1}^n M(X_i)$, we have

$$\begin{aligned} \mathbb{E}\left[1_{(\mathcal{E})} \left\|\left(\nabla^3 F_1(\theta^*) - \nabla^3 F_1(\theta')\right)(\Delta \otimes \Delta)\right\|_2^2\right] &\le \mathbb{E}\left[M_n^2 \left\|\theta^* - \theta'\right\|_2^2 \|\Delta\|_2^4\right] \\ &\le \mathbb{E}\left[M_n^8\right]^{1/4} \mathbb{E}\left[\|\theta_1 - \theta^*\|_2^8\right]^{3/4} \le O(1) M^2 \frac{G^6}{\lambda^6 n^3} \end{aligned}$$

by Hölder's inequality and Lemma 8. The remaining term we must control is the derivative difference $\mathbb{E}[\|(\nabla^3 F_1(\theta^*) - \nabla^3 F_0(\theta^*))(\Delta \otimes \Delta)\|_2^2]$. Define the random vector-valued function $\mathsf{G} = \nabla(F_1 - F_0)$, and let G_j denote its jth coordinate. Then by definition we have

$$(\nabla^3 F_1(\theta^*) - \nabla^3 F_0(\theta^*))(\Delta \otimes \Delta) = \left[\Delta^\top(\nabla^2 \mathsf{G}_1(\theta^*))\Delta \;\; \cdots \;\; \Delta^\top(\nabla^2 \mathsf{G}_d(\theta^*))\Delta\right]^\top \in \mathbb{R}^d.$$

Therefore, by the Cauchy-Schwarz inequality and the fact that $x^\top A x \le |||A|||_2 \|x\|_2^2$,

$$\begin{aligned} \mathbb{E}\left[\left\|(\nabla^3 F_1(\theta^*) - \nabla^3 F_0(\theta^*))(\Delta \otimes \Delta)\right\|_2^2\right] &= \sum_{j=1}^d \mathbb{E}\left[\left(\Delta^\top(\nabla^2 \mathsf{G}_j(\theta^*))\Delta\right)^2\right] \\ &\le \sum_{j=1}^d \left(\mathbb{E}\left[\|\Delta\|_2^8\right] \mathbb{E}\left[\left|\left|\left|\nabla^2 \mathsf{G}_j(\theta^*)\right|\right|\right|_2^4\right]\right)^{1/2}. \end{aligned}$$

Applying Lemma 8 yields that $\mathbb{E}[\|\Delta\|_2^8] = O(G^8/(\lambda^2 n)^4)$. Introducing the shorthand notation $g(\cdot;x) := \nabla f(\cdot;x) - \nabla F_0(\cdot)$, we can write

$$\nabla^2 G_j(\theta^*) = \frac{1}{n}\sum_{i=1}^n \nabla^2 g_j(\theta^*;X_i)$$

For every coordinate j, the random matrices $\nabla^2 g_j(\theta^*;X_i)$ $(i = 1,\ldots,n)$ are i.i.d. and mean zero. By Assumption 3, we have $\left|\left|\left|\nabla^2 g_j(\theta^*;X_i)\right|\right|\right|_2 \le 2L(X_i)$, whence we have $\mathbb{E}[\left|\left|\left|\nabla^2 g_j(\theta^*;X_i)\right|\right|\right|_2^8] \le 2^8 L^8$. Applying Lemma 16, we obtain

$$\mathbb{E}\left[\left|\left|\left|\nabla^2 G_j(\theta^*)\right|\right|\right|_2^4\right] \le O(1)L^4 n^{-2}\log^2(d),$$

and hence

$$\mathbb{E}\left[\left\|(\nabla^3 F_1(\theta^*) - \nabla^3 F_0(\theta^*))(\Delta\otimes\Delta)\right\|_2^2\right] \le O(1)\frac{G^4L^2}{\lambda^4} d\log(d) n^{-3},$$

which implies the desired result (54). From now on, terms of the form $\mathcal{R}_3$ will have no larger constants than those in the equality (54), so we ignore them.

Now we claim that

$$1_{(\mathcal{E})}\nabla^3 F_1(\theta^*)(\Delta\otimes\Delta) = \nabla^3 F_1(\theta^*)((\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*))) + \mathcal{R}_3. \tag{55}$$

Indeed, applying the expansion (23) to the difference $\Delta = \theta_1 - \theta^*$, we have on $\mathcal{E}$ that

$$\begin{aligned}\Delta\otimes\Delta = {} & (\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*)) + (\Sigma^{-1}P\Delta)\otimes(\Sigma^{-1}P\Delta) \\ & - (\Sigma^{-1}P\Delta)\otimes(\Sigma^{-1}\nabla F_1(\theta^*)) - (\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}P\Delta).\end{aligned}$$

We can bound each of the second three outer products in the equality above similarly; we focus on the last for simplicity. Applying the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\left|\left|\left|(\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}P\Delta)\right|\right|\right|_2^2\right] \le \left(\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^4\right]\mathbb{E}\left[\left\|\Sigma^{-1}P(\theta_1-\theta^*)\right\|_2^4\right]\right)^{\frac{1}{2}}.$$

From Lemmas 8 and 9, we obtain that

$$\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^4\right] = O(n^{-2}) \quad \text{and} \quad \mathbb{E}\left[\left\|\Sigma^{-1}P(\theta_1-\theta^*)\right\|_2^4\right] = O(n^{-4})$$

after an additional application of Cauchy-Schwarz for the second expectation. This shows that

$$(\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}P\Delta) = \mathcal{R}_3,$$

and a similar proof applies to the other three terms in the outer product $\Delta\otimes\Delta$. Using the linearity of $\nabla^3 F_1(\theta^*)$, we see that to prove the equality (55), all that is required is that

$$1_{(\mathcal{E}^c)}\nabla^3 F_1(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*))\right) = \mathcal{R}_3. \tag{56}$$

For this, we apply Hölder's inequality several times. Indeed, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|1_{(\mathcal{E}^c)}\nabla^3 F_1(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*))\right)\right\|_2^2\right] \\
&\leq \mathbb{E}[1_{(\mathcal{E}^c)}]^{1/4}\mathbb{E}\left[\left\|\nabla^3 F_1(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*))\right)\right\|_2^{8/3}\right]^{3/4} \\
&\leq \mathbb{E}[1_{(\mathcal{E}^c)}]^{1/4}\mathbb{E}\left[\left|\left|\left|\nabla^3 F_1(\theta^*)\right|\right|\right|^{8/3}\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^{16/3}\right]^{3/4} \\
&\leq \mathbb{E}[1_{(\mathcal{E}^c)}]^{1/4}\mathbb{E}\left[\left|\left|\left|\nabla^3 F_1(\theta^*)\right|\right|\right|^{8}\right]^{1/4}\mathbb{E}\left[\left\|\Sigma^{-1}\nabla F_1(\theta^*)\right\|_2^{8}\right]^{2/4} = O(n^{-1}\cdot L^2\cdot n^{-2}).
\end{aligned}$$

For the final asymptotic bound, we used Equation (26) to bound $\mathbb{E}[1_{(\mathcal{E}^c)}]$, used the fact (from Assumption 3) that $\mathbb{E}[L(X)^8] \leq L^8$ to bound the term involving $\nabla^3 F_1(\theta^*)$, and applied Lemma 7 to control $\mathbb{E}[\|\Sigma^{-1}\nabla F_1(\theta^*)\|_2^8]$. Thus the equality (56) holds, and this completes the proof of the equality (55).

For the final step in the lemma, we claim that

$$-1_{(\mathcal{E})}\Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Delta = \Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_1(\theta^*) + \mathcal{R}_3. \tag{57}$$

To prove (57) requires an argument completely parallel to that for our claim (55). As before, we use the expansion (23) of the difference Δ to obtain that on $\mathcal{E}$,

$$\begin{aligned}
&-\Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Delta \\
&= \Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_1(\theta^*) - \Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Sigma^{-1}P\Delta.
\end{aligned}$$

Now apply Lemmas 8 and 9 to the final term after a few applications of Hölder's inequality. To finish the equality (57), we argue that $1_{(\mathcal{E}^c)}\Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_1(\theta^*) = \mathcal{R}_3$, which follows exactly the line of reasoning used to prove the remainder (56).

Applying equalities (54), (55), and (57) to our earlier expansion (53) yields that

$$\begin{aligned}
\Delta &= 1_{(\mathcal{E})}\big[-\Sigma^{-1}\nabla F_1(\theta^*) - \Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Delta - \Sigma^{-1}\nabla^3 F_1(\theta^*)(\Delta\otimes\Delta) \\
&\qquad + \Sigma^{-1}(\nabla^3 F_0(\theta^*) - \nabla^3 F_1(\theta'))(\Delta\otimes\Delta)\big] + 1_{(\mathcal{E}^c)}\Delta \\
&= -\Sigma^{-1}\nabla F_1(\theta^*) + \Sigma^{-1}(\nabla^2 F_1(\theta^*)-\Sigma)\Sigma^{-1}\nabla F_1(\theta^*) \\
&\qquad - \Sigma^{-1}\nabla^3 F_1(\theta^*)\left((\Sigma^{-1}\nabla F_1(\theta^*))\otimes(\Sigma^{-1}\nabla F_1(\theta^*))\right) + \mathcal{R}_3 + 1_{(\mathcal{E}^c)}\Delta.
\end{aligned}$$

Finally, the bound (26) implies that $\mathbb{E}[1_{(\mathcal{E}^c)}\|\Delta\|_2^2] \leq \mathbb{P}(\mathcal{E}^c)R^2 = O(n^{-4})$, which yields the claim.

References

A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems 24*, 2011.

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58 (5):3235–3249, May 2012.

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

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

R. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference*, to appear, 2012.

A. de Acosta. Inequalities for b-valued random vectors with applications to the strong law of large numbers. *The Annals of Probability*, 9:157–161, 1981.

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

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

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012b.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer, 1992.

E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2009.

R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

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

G. Mann, R. McDonald, M. Mohri, N. Silberman, and D. Walker. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems 22*, pages 1231–1239, 2009.

C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

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

A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54:48–61, 2009.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

S. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

S. S. Ram, A. Nedić, and V. V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147(3):516–545, 2010.

B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: a lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, 2011.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *arXiv preprint arXiv:1209.1873*, 2012.

G. Sun. KDD cup track 2 `soso.com` ads prediction challenge, 2012. URL `http://www.kddcup2012.org/c/kddcup2012-track2`. Accessed August 1, 2012.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Divide and conquer kernel ridge regression. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, Princeton, NJ, July 2013.

M. A. Zinkevich, A. Smola, M. Weimer, and L. Li. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*, 2010.