Journal of Machine Learning Research 18 (2017) 1-30 Submitted 1/16; Revised 2/16; Published 1/17

Communication-efficient Sparse Regression

Jason D. Lee JasonLee@marshall.usc.edu
Marshall School of Business
University of Southern California
Los Angeles, CA 90089

Qiang Liu qliu@cs.dartmouth.edu
Department of Computer Science
Dartmouth University
Hanover, NH 02714

Yuekai Sun yuekai@umich.edu
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Jonathan E. Taylor jonathan.taylor@stanford.edu
Department of Statistics
Stanford University
Stanford, CA 94305

Editor: Zhihua Zhang

Abstract

We devise a communication-efficient approach to distributed sparse regression in the high-dimensional setting. The key idea is to average "debiased" or "desparsified" lasso estimators. We show the approach converges at the same rate as the lasso as long as the dataset is not split across too many machines, and consistently estimates the support under weaker conditions than the lasso. On the computational side, we propose a new parallel and computationally-efficient algorithm to compute the approximate inverse covariance required in the debiasing approach, when the dataset is split across samples. We further extend the approach to generalized linear models.

Keywords: Distributed Sparse Regression, Averaging, Debiasing, lasso, high-dimensional statistics

1. Introduction

Explosive growth in the size of modern datasets has fueled interest in distributed statistical learning. For examples, we refer to Boyd et al. (2011); Dekel et al. (2012); Duchi et al. (2012); Zhang et al. (2013) and the references therein. The problem arises, for example, when working with datasets that are too large to fit on a single machine and must be distributed across multiple machines. The main bottleneck in the distributed setting is usually communication between machines/processors, so the overarching goal of algorithm design is to minimize communication costs.

In distributed statistical learning, the simplest and most popular approach is *averaging*: each machine forms a local estimator $\hat{\theta}_k$ with the portion of the data stored locally, and a "master" averages the local estimators to produce an aggregate estimator: $\bar{\theta} = \frac{1}{m}\sum_{k=1}^{m} \hat{\theta}_k$. Averaging was first studied by Mcdonald et al. (2009) for multinomial regression. They derive non-asymptotic error bounds on the estimation error that show averaging reduces the variance of the local estimators, but has no effect on the bias (from the centralized solution). In follow-up work, Zinkevich et al. (2010) studied a variant of averaging where each machine computes a local estimator with stochastic gradient descent (SGD) on a random subset of the dataset. They show, among other things, that their estimator converges to the centralized estimator.

More recently, Zhang et al. (2013) studied averaged empirical risk minimization (ERM). They show that the mean squared error (MSE) of the averaged ERM decays like $O\big(N^{-\frac{1}{2}} + \frac{m}{N}\big)$, where m is the number of machines and N is the total number of samples. Thus, so long as $m \lesssim \sqrt{N}$, the averaged ERM matches the $N^{-\frac{1}{2}}$ convergence rate of the centralized ERM. Even more recently, Rosenblatt and Nadler (2014) studied the optimality of averaged ERM in two asymptotic settings: $N \to \infty$, m, p fixed and $p, n \to \infty$, $\frac{p}{n} \to \mu_l \in (0, 1)$, where $n = \frac{N}{m}$ is the number of samples per machine. They show that in the $n \to \infty$, p fixed setting, the averaged ERM is first-order equivalent to the centralized ERM. However, when $p, n \to \infty$, the averaged ERM is suboptimal (versus the centralized ERM).

We develop a divide and conquer approach to statistical learning. In the high-dimensional setting, regularization is essential. The key idea is to average *debiased* or *desparsified* regularized M-estimators. Under suitable conditions, it is possible to show that the local debiased estimators are asymptotically normal. thus the averaged estimator delivers the same statistical performance as the computationally infeasible centralized M-estimator.

Formally, we show that the error of the averaged estimator decomposes into a $\widetilde{O}_P\big(\frac{1}{\sqrt{N}}\big)$ asymptotically normal term and a remainder term. As long as $m \lesssim \frac{\sqrt{N}}{s \log p}$, where s is the sparsity of the unknown regression coefficients, the reminder term is asymptotically negligible. Thus the averaged estimator converges at the same rate as a centralized estimator. Further, the averaged estimator is model selection consistent under a weak minimum signal strength condition. In the following section, we review the theoretical properties of the lasso and debiased lasso and describe our contributions more formally.

2. A divide-and-conquer approach to sparse regression

To keep things simple, we focus on sparse linear regression. Consider the sparse linear model

$$y = X\beta^* + \epsilon,$$

where the rows of $X \in \mathbf{R}^{n \times p}$ are predictors, and the components of $y \in \mathbf{R}^n$ are the responses. To keep things simple, we assume

(A1) the predictors $x \in \mathbf{R}^p$ are independent σ_x-subgaussian random vectors with whose covariance Σ has smallest eigenvalue $\sigma_p(\Sigma) > \lambda_{\min}$ and largest eigenvalue $\sigma_1(\Sigma) < \lambda_{\max}$;

(A2) the regression coefficients $\beta^* \in \mathbf{R}^p$ are s-sparse, i.e. all but s components of β^* are zero;

(A3) the components of the noise ϵ are independent, mean zero σ_y-subgaussian random variables.

Given the predictors and responses, the lasso estimates β^* by

$$\hat{\beta} := \arg\min_{\beta \in \mathbf{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

There is a well-developed theory of the lasso that says, under suitable assumptions on X, the lasso estimator $\hat{\beta}$ is nearly minimax optimal(e.g. see Hastie et al. (2015), Chapter 11 for an overview). More precisely, under some conditions on $\frac{1}{n}X^T X$, the MSE of the lasso estimator is roughly $\frac{s \log p}{n}$, which is the minimax rate.

2.1 Background on the lasso and debiasing

However, the lasso estimator is also biased[1]. Since averaging only reduces variance, not bias, we gain (almost) nothing by averaging the biased lasso estimators. That is, it is possible to show if we naively averaged local lasso estimators, the MSE of the averaged estimator is of the same order as that of the local estimators. The key to overcoming the bias of the averaged lasso estimator is to "debias" the lasso estimators before averaging.

The *debiased lasso estimator* by Javanmard and Montanari (2013a) is

$$\hat{\beta}^d := \hat{\beta} + \frac{1}{n}\hat{\Theta} X^T (y - X\hat{\beta}), \tag{1}$$

where $\hat{\beta}$ is the lasso estimator and $\hat{\Theta} \in \mathbf{R}^{p \times p}$ is an approximate inverse to $\hat{\Sigma} = \frac{1}{n}X^T X$. Intuitively, the debiased lasso estimator trades bias for variance. The trade-off is obvious when $\hat{\Sigma}$ is non-singular: setting $\hat{\Theta} = \hat{\Sigma}^{-1}$ gives the ordinary least squares (OLS) estimator $(X^T X)^{-1} X^T y$.

Another way to interpret the debiased lasso estimator is a corrected estimator that compensates for the bias incurred by shrinkage. By the optimality conditions of the lasso, the correction term $\frac{1}{n}X^T(y - X\hat{\beta})$ is a subgradient of $\lambda \|\cdot\|_1$ at $\hat{\beta}$. By adding a term proportional to the subgradient of the regularizer, the debiased lasso estimator compensates for the bias incurred by regularization. The debiased lasso estimator has previously been used to perform inference on the regression coefficients in high-dimensional regression models. We refer to the papers by Javanmard and Montanari (2013a); van de Geer et al. (2013); Zhang and Zhang (2014); Belloni et al. (2011) for details.

The choice of $\hat{\Theta}$ in the correction term is crucial to the performance of the debiased estimator. Javanmard and Montanari (2013a) suggest forming $\hat{\Theta}$ row by row: the j-th row of $\hat{\Theta}$ is the optimum of

$$\begin{aligned} \underset{\theta \in \mathbf{R}^p}{\text{minimize}} \quad & \theta^T \hat{\Sigma} \theta \\ \text{subject to} \quad & \|\hat{\Sigma}\theta - e_j\|_\infty \leq \delta. \end{aligned} \tag{2}$$

1. We refer to Section 2.2 in Javanmard and Montanari (2013a) for a more formal discussion of the bias of the lasso estimator.

The parameter δ should large enough to keep the problem feasible, but as small as possible to keep the bias (of the debiased lasso estimator) small. As we shall see, when the rows of X are subgaussian, setting $\delta \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$ is usually large enough to keep (2) feasible.

Definition 1 (Generalized coherence) *Given $X \in \mathbf{R}^{n\times p}$, let $\hat{\Sigma} = \frac{1}{n}X^TX$. The* generalized coherence *between $\hat{\Sigma}$ and $\Theta \in \mathbf{R}^{p\times p}$ is*

$$\mathrm{GC}(\hat{\Sigma}, \Theta) = \max_{j\in[p]} \|\hat{\Sigma}\Theta_j^T - e_j\|_\infty.$$

The preceding definition is a generalization of the usual notion of coherence as it appears in the compressed sensing literature. Assume the columns of X are normalized so that $\|x_j\|_2 = 1$, and $\Theta = I$. The diagonal entries of $\hat{\Sigma}\Theta - I$ vanish. Thus $\mathrm{GC}(\hat{\Sigma}, \Theta)$ is the largest off diagonal entry of $\hat{\Sigma}$, which is the largest inner product between columns of X: $\frac{1}{n}\max_{i\neq j} \left|e_i^T X^T X e_j\right|$. We recognize the preceding quantity as the coherence of X.

Lemma 2 (Javanmard and Montanari (2013a)) *Under (A1), when $16\kappa\sigma_x^4 n > \log p$, the event*

$$\mathcal{E}_{\mathrm{GC}}(\hat{\Sigma}) := \left\{\mathrm{GC}(\hat{\Sigma}, \Sigma^{-1}) \leq \frac{8}{\sqrt{c_1}}\sqrt{\kappa}\sigma_x^2\left(\frac{\log p}{n}\right)^{\frac{1}{2}}\right\}$$

occurs with probability at least $1 - 2p^{-2}$ for some $c_1 > 0$, where $\kappa := \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ is the condition number of Σ.

As we shall see, the bias of the debiased lasso estimate is of higher order than its variance under suitable conditions on $\hat{\Sigma}$. In particular, we require $\hat{\Sigma}$ to satisfy the *restricted eigenvalue (RE) condition.*

Definition 3 (RE condition) *For any $\mathcal{S} \subset [p]$, let*

$$\mathcal{C}(\mathcal{S}) := \{\Delta \in \mathbf{R}^p \mid \|\Delta_{\mathcal{S}^c}\|_1 \leq 3\|\Delta_{\mathcal{S}}\|_1\}.$$

We say $\hat{\Sigma}$ satisfies the RE condition on the cone $\mathcal{C}(\mathcal{S})$ when

$$\Delta^T\hat{\Sigma}\Delta \geq \mu_l\|\Delta_{\mathcal{S}}\|_2^2$$

for some $\mu_l > 0$ and any $\Delta \in \mathcal{C}(\mathcal{S})$.

The RE condition requires $\hat{\Sigma}$ to be positive definite on $\mathcal{C}(\mathcal{S})$. When the rows of $X \in \mathbf{R}^{n\times p}$ are *i.i.d.* Gaussian random vectors, Raskutti et al. (2010) show there are constants $\mu_1, \mu_2 > 0$ such that

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \mu_1\|\Delta\|_2^2 - \mu_2\frac{\log p}{n}\|\Delta\|_1^2 \text{ for any } \Delta \in \mathbf{R}^p$$

with probability at least $1 - c_2\exp(-c_2 n)$. Their result implies the RE condition holds on $\mathcal{C}(\mathcal{S})$ (for any $\mathcal{S} \subset [p]$) as long as $n \gtrsim |S|\log p$, even when there are dependencies among the predictors. Their result was extended to subgaussian designs by Rudelson and Zhou (2013), also allowing for dependencies among the covariates. We state their result verbatim.

Lemma 4 (Rudelson and Zhou (2013)) *Under (A1), if* $n > 4000\tilde{s}\sigma_x^2 \log\left(\frac{Cp}{\tilde{s}}\right)$ *and* $p > \tilde{s}$, *where* $\tilde{s} := s + 25920\kappa s = C's$, *the event*

$$\mathcal{E}_{\mathrm{RE}}(X) = \left\{\Delta^T\hat{\Sigma}\Delta \geq \frac{1}{2}\lambda_{\min}(\Sigma)\|\Delta_S\|_2^2 \text{ for any } \Delta \in \mathcal{C}(S)\right\}$$

occurs with probability at least $1 - 2e^{-\frac{n}{4000\sigma_x^4}}$, *where* C *and* C' *are universal constants.*

Proof The lemma is a consequence of Rudelson and Zhou (2013), Theorem 6. In their notation, we set $\delta = \frac{1}{\sqrt{2}}$, $k_0 = 3$ and bound $\max_{j\in[p]} \|Ae_j\|_2^2$ and $K(s_0, k_0, \Sigma^{\frac{1}{2}})$ by $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)^{-\frac{1}{2}}$. ■

When the RE condition holds, the lasso and debiased lasso estimators are consistent for a suitable choice of the regularization parameter λ. The parameter λ should be large enough to dominate the "empirical process" part of the problem: $\frac{1}{n}\left\|X^T y\right\|_\infty$, but as small as possible to reduce the bias incurred by regularization. As we shall see, setting $\lambda \sim \sigma_y\left(\frac{\log p}{n}\right)^{\frac{1}{2}}$ is a good choice.

Lemma 5 *Under (A3),*

$$\frac{1}{n}\|X^T\epsilon\|_\infty \leq \max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}}\sigma_y\left(\frac{3\log p}{c_2 n}\right)^{\frac{1}{2}}$$

with probability at least $1 - ep^{-2}$ *for any (non-random)* $X \in \mathbf{R}^{n\times p}$.

When $\hat{\Sigma}$ satisfies the RE condition and λ is large enough, Negahban et al. (2012) show that the lasso is consistent.

Lemma 6 (Negahban et al. (2012)) *If in addition to (A2) and (A3),*

1. $\hat{\Sigma}$ *satisfies the RE condition on* $\mathcal{C}(\mathrm{supp}(\beta^*))$ *with constant* μ_l
2. $\frac{1}{n}\|X^T\epsilon\|_\infty \leq \lambda$,

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{3}{\mu_l}s\lambda \text{ and } \|\hat{\beta} - \beta^*\|_2 \leq \frac{3}{\mu_l}\sqrt{s}\lambda.$$

When the lasso estimator is consistent, the debiased lasso estimator is also consistent. Further, it is possible to show that the bias of the debiased estimator is of higher order than its variance. Similar results by Javanmard and Montanari (2013a); van de Geer et al. (2013); Zhang and Zhang (2014); Belloni et al. (2011) are the key step in showing the asymptotic normality of the (components of) the debiased lasso estimator. The result we state is essentially Javanmard and Montanari (2013a), Theorem 2.3.

Lemma 7 *If in addition to (A2) and (A3),*

1. $\hat{\Sigma}$ *satisfies the RE condition on* $\mathcal{C}(\mathrm{supp}(\beta^*))$ *with constant* μ_l
2. $\frac{1}{n}\|X^T\epsilon\|_\infty \leq \lambda$,

3. $(\hat{\Sigma}, \hat{\Theta})$ *has generalized incoherence* δ,

the debiased lasso estimator has the form

$$\hat{\beta}^d = \beta^* + \frac{1}{n}\hat{\Theta}X^T\epsilon + \hat{\Delta},$$

where $\|\hat{\Delta}\|_\infty \leq \frac{3\delta}{\mu_l} s\lambda$.

Lemma 7, together with Lemmas 5 and 2, shows that the bias of the debiased lasso estimator is of higher order than its variance. In particular, setting λ and δ to be the order of the upper bounds on $\inf_{\Theta \in \mathbf{R}^{p\times p}} \mathrm{GC}(\hat{\Sigma}, \Theta)$ and $\frac{1}{n}\|X^T\epsilon\|_\infty$ given by Lemmas 5 and 2 gives a bias term $\|\hat{\Delta}\|_\infty$ that is $O_P\big(\frac{s\log p}{n}\big)$. By comparison, the variance term $\frac{1}{n}\|\hat{\Theta}X^T\epsilon\|_\infty$ is the maximum of p subgaussian random variables with mean zero and variances of $O(1)$, which is $O\big(\big(\frac{\log p}{n}\big)^{\frac{1}{2}}\big)$. Thus the bias term is of higher order than the variance term as long as $n \gtrsim s^2 \log p$.

Corollary 8 *If in addition to the conditions of Lemma 6,*

1. $(\hat{\Sigma}, \hat{\Theta})$ *has generalized incoherence* $\delta'\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$,
2. $\lambda = \max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}}\sigma_y\big(\frac{3\log p}{c_2 n}\big)^{\frac{1}{2}}$,

$$\|\hat{\Delta}\|_\infty \leq \frac{3\sqrt{3}}{\sqrt{c_2}} \frac{\delta' \max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}}}{\mu_l}\sigma_y \frac{s\log p}{n}.$$

The rest of the paper is organized as follows. In the subsequent section, we describe a divide-and-conquer approach to sparse regression and derive its theoretical properties. We show that

1. the averaged estimator converges at the same rate as the communication-intensive centralized lasso estimator. In particular, a thresholded version of the averaged estimator attains the same estimation rate as the centralized lasso estimator $\|\bar{\beta}^{ht} - \beta^*\|_2 \lesssim \big(\frac{s\log p}{N}\big)^{\frac{1}{2}}$, and only requires one round of communication.
2. a thresholded version of the averaged estimator is model selection consistent as long as the minimum signal strength is at least $\big(\frac{\log p}{N}\big)^{\frac{1}{2}}$. We remark that the model selection consistency result does not require X to obey an irrepresentability condition, which the centralized lasso does require.

Although the divide-and-conquer approach is communication efficient, it is costly in terms of floating point operations. The parallel runtime of debiasing is roughly equivalent to the cost of evaluating p lasso estimators, due to computation of $\hat{\Theta}$. In the rest of this section, we describe a more sophisticated approach to debiasing: each machine debiases $\frac{p}{m}$, instead of all p, regression coefficients. Thus the parallel runtime of the more sophisticated approach is roughly m times smaller than that of a naive approach.

In Section 4, we further refine the approach to reduce the sample complexity from $n \gtrsim ms^2 \log p$ to $n \gtrsim ms \log p$. In Section 6, we show via simulation experiments that

the averaged debiased estimator outperforms averaging local lasso estimates, and performs as well as the centralized lasso. Section 5 generalizes our approach from least-squares to generalized linear models such as logistic regression.

Finally in Section 7, we show the optimality of our estimator in terms of the amount of communication, and rounds of communication using recent work on communication lower bounds. We also provide a comparison of the average debiased estimator and the centralized lasso estimator. The parallel runtime of the averaging debiased estimator is only larger than the centralized lasso by a constant multiplicative factor.

2.2 Averaging debiased lassos

Recall the problem setup: we are given N samples of the form (x_i, y_i) distributed across m machines:

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

The k-th machine has local predictors $X_k \in \mathbf{R}^{n_k \times p}$ and responses $y_k \in \mathbf{R}^{n_k}$. To keep things simple, we assume the data is evenly distributed, i.e. $n_1 = \cdots = n_k = n = \frac{N}{m}$. The *averaged debiased lasso* estimator (for lack of a better name) is

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k^d = \frac{1}{m}\sum_{k=1}^{m} \big(\hat{\beta}_k + \hat{\Theta}_k X_k^T (y_k - X_k \hat{\beta}_k)\big), \tag{3}$$

We begin by studying the error of the $\bar{\beta}$ in the ℓ_∞ norm.

Lemma 9 *Suppose the local sparse regression problem on each machine satisfies the conditions of Corollary 8, that is when $m \le p$,*

1. *$\{\hat{\Sigma}_k\}_{k\in[m]}$ satisfy the RE condition on $\mathcal{C}(\mathrm{supp}(\beta^*))$ with constant μ_l,*
2. *$\{(\hat{\Sigma}_k, \hat{\Theta}_k)\}_{k\in[m]}$ have generalized incoherence $c_{\mathrm{GC}}\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$,*
3. *$\lambda_1 = \cdots = \lambda_m = c_\Sigma \sigma_y \big(\frac{3\log p}{c_2 n}\big)^{\frac{1}{2}}$.*

Then

$$\|\bar{\beta} - \beta^*\|_\infty \le c\sigma_y \Big(\Big(\frac{c_\Omega \log p}{N}\Big)^{\frac{1}{2}} + \frac{c_{\mathrm{GC}} c_\Sigma}{\mu_l} \sigma_y \frac{s \log p}{n} \Big)$$

with probability at least $1 - ep^{-1}$, where $c > 0$ is a universal constant, $c_\Omega := \max_{j\in[p],\, k\in[m]}((\hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T)_{j,j})$ and $c_\Sigma := \max_{j\in[p],k\in[m]}((\hat{\Sigma}_k)_{j,j})^{\frac{1}{2}}$.

Lemma 10 hints at the performance of the averaged debiased lasso. In particular, we note the first term is $O\big(\big(\frac{\log p}{N}\big)^{\frac{1}{2}}\big)$, which matches the convergence rate of the centralized estimator. When n is large enough, $\frac{s\log p}{n}$ is negligible compared to $\big(\frac{\log p}{N}\big)^{\frac{1}{2}}$, and the error is $O\big(\big(\frac{\log p}{N}\big)^{\frac{1}{2}}\big)$.

Next, we show the conditions of Lemma 10 occur with high probability when the rows of X are independent subgaussian random vectors.

Lemma 10 *Under (A1), (A2), and (A3), when $m < p$, $p > \tilde{s}$,*

1. $n > \max\{4000\tilde{s}\sigma_x^2 \log(\frac{Cp}{\tilde{s}}),\ 8000\sigma_x^4 \log p,\ \frac{3}{c_1}\max\{\sigma_x^2, \sigma_x\}\log p\}$,
2. $\lambda_1 = \cdots = \lambda_m = \max_{j\in[p],k\in[m]}((\hat{\Sigma}_k)_{j,j})^{\frac{1}{2}}\sigma_y\big(\frac{3\log p}{c_2 n}\big)^{\frac{1}{2}}$,
3. $\delta_1 = \cdots = \delta_m = \frac{8}{\sqrt{c_1}}\sqrt{\kappa}\sigma_x^2\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ *and form* $\{\hat{\Theta}_k\}_{k\in[m]}$ *by* (2),

$$\|\bar{\beta} - \beta^*\|_\infty \le c\Big(\sigma_y\Big(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}\Big)$$

with probability at least $1-(8+e)p^{-1}$ for some universal constant $c > 0$.

The averaged debiased lasso has one serious drawback versus the lasso: $\bar{\beta}$ is usually dense. The density of $\bar{\beta}$ detracts from the intrepretability of the coefficients and makes the estimation error large in the ℓ_2 and ℓ_1 norms. To remedy both problems, we threshold the averaged debiased lasso:

$$\begin{aligned}\mathrm{HT}_t(\bar{\beta}) &\leftarrow \bar{\beta}_j \cdot \mathbf{1}_{\{|\bar{\beta}_j|\ge t\}},\\ \mathrm{ST}_t(\bar{\beta}) &\leftarrow \mathrm{sign}(\bar{\beta}_j)\cdot\max\{|\bar{\beta}_j| - t,\, 0\}.\end{aligned}$$

As we shall see, both hard and soft-thresholding give sparse aggregates that are close to β^* in ℓ_2 norm.

Lemma 11 *As long as $t > \|\bar{\beta} - \beta^*\|_\infty$, $\bar{\beta}^{ht} := \mathrm{HT}_t(\bar{\beta})$ satisfies*

1. $\|\bar{\beta}^{ht} - \beta^*\|_\infty \le 2t$,
2. $\|\bar{\beta}^{ht} - \beta^*\|_2 \le 2\sqrt{2s}t$,
3. $\|\bar{\beta}^{ht} - \beta^*\|_1 \le 2\sqrt{2}st$.

The analogous result also holds for $\bar{\beta}^{st} := \mathrm{ST}_t(\bar{\beta})$.

Proof By the triangle inequality,

$$\begin{aligned}\|\bar{\beta}^{ht} - \beta^*\|_\infty &\le \|\bar{\beta}^{ht} - \bar{\beta}\|_\infty + \|\bar{\beta} - \beta^*\|_\infty\\ &\le t + \big\|\bar{\beta} - \beta^*\big\|_\infty\\ &\le 2t.\end{aligned}$$

Since $t > \big\|\bar{\beta} - \beta^*\big\|_\infty$, $\bar{\beta}^{ht}_j = 0$ whenever $\beta^*_j = 0$. Thus $\bar{\beta}^{ht}$ is s-sparse and $\bar{\beta}^{ht} - \beta^*$ is $2s$-sparse. By the equivalence between the ℓ_∞ and ℓ_2, ℓ_1 norms,

$$\begin{aligned}\|\bar{\beta}^{ht} - \beta^*\|_2 &\le 2\sqrt{2s}t,\\ \|\bar{\beta}^{ht} - \beta^*\|_1 &\le 2\sqrt{2}st.\end{aligned}$$

The argument for $\bar{\beta}^{st}$ is similar. ■

By combining Lemma 11 with Lemma 10, we show that $\bar{\beta}^{ht}$ converges at the same rates as the centralized lasso.

Theorem 12 *Under the conditions of Lemma 10, hard-thresholding $\bar{\beta}$ at $\sigma_y\Big(\frac{4\max_{j\in[p]}\Sigma^{-1}_{j,j}\log p}{c_2 N}\Big)^{\frac{1}{2}} + \frac{48\sqrt{6}}{\sqrt{c_1 c_2}}\frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}$ gives*

1. $\|\bar{\beta}^{ht}-\beta^*\|_\infty \lesssim_P \sigma_y\Big(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}$,
2. $\|\bar{\beta}^{ht}-\beta^*\|_2 \lesssim_P \sigma_y\Big(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}s\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s^{\frac{3}{2}}\log p}{n}$,
3. $\|\bar{\beta}^{ht}-\beta^*\|_1 \lesssim_P \sigma_y\Big(\frac{\max_{j\in[p]}\Sigma^{-1}_{j,j}s^2\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s^2\log p}{n}$.

Remark 13 *By Theorem 12, when $m \lesssim \frac{n}{s^2\log p}$, the variance term is dominant and the convergence rates given by the theorem simplify:*

1. $\|\bar{\beta}^{ht}-\beta^*\|_\infty \lesssim_P \big(\frac{\log p}{N}\big)^{\frac{1}{2}}$,
2. $\|\bar{\beta}^{ht}-\beta^*\|_2 \lesssim_P \big(\frac{s\log p}{N}\big)^{\frac{1}{2}}$,
3. $\|\bar{\beta}^{ht}-\beta^*\|_1 \lesssim_P \big(\frac{s^2\log p}{N}\big)^{\frac{1}{2}}$.

The convergence rates for the centralized lasso estimator $\hat{\beta}$ are identical (modulo constants):

1. $\|\hat{\beta}-\beta^*\|_\infty \lesssim_P \big(\frac{\log p}{N}\big)^{\frac{1}{2}}$,
2. $\|\hat{\beta}-\beta^*\|_2 \lesssim_P \big(\frac{s\log p}{N}\big)^{\frac{1}{2}}$,
3. $\|\hat{\beta}-\beta^*\|_1 \lesssim_P \big(\frac{s^2\log p}{N}\big)^{\frac{1}{2}}$.

The estimator $\bar{\beta}^{ht}$ matches the convergence rates of the centralized lasso in ℓ_1, ℓ_2, and ℓ_∞ norms. Furthermore, $\bar{\beta}^{ht}$ can be evaluated in a communication-efficient manner by a one-shot averaging approach.

Corollary 14 *Under the conditions of Lemma 10, further assume*

1. $m \lesssim \frac{n}{s^2\log p}$,
2. *β-min:* $|\beta^*_j| \gtrsim \big(\frac{\log p}{N}\big)^{\frac{1}{2}}$ *for any* $j \in \mathrm{supp}(\beta^*)$.

Then $\mathrm{supp}(\bar{\beta}^{ht}) = \mathrm{supp}(\beta^*)$.

Proof As long as we threshold at $t > \|\bar{\beta}^{ht}-\beta^*\|_\infty$, $\mathrm{supp}(\bar{\beta}^{ht}) \subset \mathrm{supp}(\beta^*)$. That is, all the zero components of β^* are correctly estimated. Further, as long as the non-zero components of β^* have magnitude at least $2t$, they are not set to zero by thresholding at t. By Theorem 12, there is such a $t \sim \big(\frac{\log p}{N}\big)^{\frac{1}{2}}$. ■

3. A distributed approach to debiasing

The averaged estimator we studied has the form

$$\bar{\beta} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k + \hat{\Theta}_k X_k^T (y - X_k \hat{\beta}_k).$$

The estimator requires each machine to form $\hat{\Theta}_k$ by the solution of (2). Since the dual of (2) is an ℓ_1-regularized quadratic program:

$$\underset{\gamma \in \mathbf{R}^p}{\text{minimize}} \frac{1}{2}\gamma^T \hat{\Sigma}_k \gamma - \hat{\Sigma}_k \gamma + \delta \left\|\gamma\right\|_1, \tag{4}$$

forming $\hat{\Theta}_k$ is (roughly speaking) p times as expensive as solving the local lasso problem, making it the most expensive step (in terms of floating point operations) of evaluating the averaged estimator. To trim the cost of the debiasing step, we consider an estimator that forms only a single $\hat{\Theta}$:

$$\tilde{\beta} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k + \frac{1}{N}\hat{\Theta}\sum_{k=1}^{m} X_k^T (y - X_k \hat{\beta}_k). \tag{5}$$

To evaluate (5),

1. each machine sends $\hat{\beta}_k$ and $\frac{1}{n}X_k^T(y - X_k\hat{\beta}_k)$ to a central server,
2. the central server forms $\frac{1}{m}\sum_{k=1}^m \hat{\beta}_k$ and $\frac{1}{N}\sum_{k=1}^m X_k^T(y - X_k\hat{\beta}_k)$ and sends the averages to all the machines,
3. each machine, given the averages, forms $\frac{p}{m}$ rows of $\hat{\Theta}$ and debiases $\frac{p}{m}$ coefficients:

$$\tilde{\beta}_j = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_j + \hat{\Theta}_{j,\cdot}\Big(\frac{1}{N}\sum_{k=1}^{m} X_k^T (y - X_k \hat{\beta}_k)\Big),$$

   where $\hat{\Theta}_{j,\cdot} \in \mathbf{R}^p$ is a row vector.

As we shall see, each machine can perform debiasing with only the data stored locally. Thus, forming the estimator (5) requires two rounds of communication.

The question that remains is how to form $\hat{\Theta}_{j,\cdot}$. We consider an estimator proposed by van de Geer et al. (2013): nodewise regression on the predictors. For some $j \in [p]$ that machine k is debiasing, the machine solves

$$\hat{\gamma}_j := \underset{\gamma \in \mathbf{R}^{p-1}}{\arg\min} \frac{1}{2n}\|X_{k,j} - X_{k,-j}\gamma\|_2^2 + \lambda_j \|\gamma\|_1,\ j \in [p],$$

where $X_{k,-j} \in \mathbf{R}^{n \times (p-1)}$ is X_k less its j-th column $X_{k,j}$. Implicitly, we are forming

$$\hat{C} := \begin{bmatrix} 1 & -\hat{\gamma}_{1,2} & \dots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \dots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \dots & -\hat{\gamma}_{p,p} \end{bmatrix},$$

where the components of $\hat{\gamma}_j$ are indexed by $k \in \{1, \ldots, j-1, j+1, \ldots, p\}$. We scale the rows of $\hat{C}$ by $\mathbf{diag}\Big(\big[\hat{\tau}_1, \ldots, \hat{\tau}_p\big]\Big)$, where

$$\hat{\tau}_j = \big(\frac{1}{n}\|X_j - X_{-j}\hat{\gamma}_j\|_2^2 + \lambda_j\|\hat{\gamma}_j\|_1\big)^{\frac{1}{2}},$$

to form $\hat{\Theta} = \hat{T}^{-2}\hat{C}$. Each row of $\hat{\Theta}$ is given by

$$\hat{\Theta}_{j,\cdot} = -\frac{1}{\hat{\tau}_j^2}\begin{bmatrix}\hat{\gamma}_{j,1} & \cdots & \hat{\gamma}_{j,j-1} & 1 & \hat{\gamma}_{j,j+1} & \cdots & \hat{\gamma}_{j,p}\end{bmatrix}. \tag{6}$$

Since $\hat{\gamma}_j$ and $\hat{\tau}_j$ only depend on X_k, they can be formed without any communication.

Before we justify the choice of $\hat{\Theta}$ theoretically, we mention that it is an approximate "inverse" of $\hat{\Sigma}$ (in a component-wise sense). By the optimality conditions of nodewise regression,

$$\begin{aligned}\hat{\tau}_j^2 &= \frac{1}{n}\|X_j - X_{-j}\hat{\gamma}_j\|_2^2 + \lambda_j\|\hat{\gamma}_j\|_1 \\ &= \frac{1}{n}\|X_j - X_{-j}\hat{\gamma}_j\|_2^2 + \frac{1}{n}(X_j - X_{-j}\hat{\gamma}_j)^T X_{-j}^T\hat{\gamma}_j \\ &= \frac{1}{n}X_j(X_j - X_{-j}\hat{\gamma}).\end{aligned}$$

Recalling the defintition of $\hat{\Theta}$, we have

$$\begin{aligned}\frac{1}{n}\hat{\Theta}_{j,\cdot}X^T X_j = \frac{1}{\hat{\tau}_j^2}\frac{1}{n}(X_j - \hat{\gamma}_j^T X_{-j})^T X_j = \lambda_j \text{ and} \\ \frac{1}{n}\|\hat{\Theta}_{j,\cdot}X^T X_{-j}\|_\infty = \frac{1}{\hat{\tau}_j^2}\Big\|\frac{1}{n}(X_j - \hat{\gamma}_j^T X_{-j})^T X_{-j}\Big\|_\infty \le \frac{\lambda_j}{\hat{\tau}_j^2}\end{aligned}$$

for any $j \in [\,p\,]$. Thus

$$\max_{j\in[\,p\,]}\|\hat{\Theta}_{j,\cdot}\hat{\Sigma} - e_j\|_\infty \le \frac{\lambda_j}{\hat{\tau}_j^2}. \tag{7}$$

van de Geer et al. (2013) show that when the rows of X are *i.i.d.* subgaussian random vectors and the precision matrix Σ^{-1} is sparse, $\hat{\Theta}_{j,\cdot}$ converges to Σ_j^{-1} at the usual convergence rate of the lasso. For completeness, we restate their result.

We consider a sequence of regression problems indexed by the sample size N, dimension p, sparsity s_0 that satisfies (A1), (A2), and (A3). As N grows to infinity, both $p = p(N)$ and $s = s(N)$ may also grow as a function of N. To keep notation manageable, we drop the index N. We further assume

(A4) the covariance of the predictors (rows of X) has smallest eigenvalue $\lambda_{\min}(\Sigma) \sim \Omega(1)$ and largest diagonal entry $\max_{j\in[p]}\Sigma_{j,j} \sim O(1)$,

(A5) the rows of Σ^{-1} are sparse: $\max_{j\in[p]}\frac{s_j^2\log p}{n} \sim o(1)$, where s_j is the sparsity of Σ_j^{-1}.

Lemma 15 (van de Geer et al. (2013), Theorem 2.4) *Under (A1)–(A5),* (6) *with suitable parameters* $\lambda_j \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$ *satisfies*

$$\|\hat{\Theta}_{j,\cdot} - \Sigma_j^{-1}\|_1 \lesssim_P \left(\frac{s_j^2 \log p}{n}\right)^{\frac{1}{2}} \text{ for any } j \in [p].$$

We show that the averaged estimator (5) matches the convergence rate of the centralized lasso.

Theorem 16 *Under (A1)–(A5),* (5)*, where* $\hat{\Theta}$ *is given by* (6)*, with suitable parameters* $\lambda_j, \lambda_k \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$*,* $j \in [p]$*,* $k \in [m]$ *satisfies*

$$\|\bar{\beta} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac{1}{2}} + \frac{s_{\max} \log p}{n},$$

where $s_{\max} := \max\{s_0, s_1, \ldots, s_p\}$.

Proof See the appendix. ■

By combining the Lemma 11 with Theorem 16, we can show that $\tilde{\beta}^{ht} := \mathrm{HT}(\tilde{\beta}, t)$ for an appropriate threshold t converges to β^* at the same rates as the centralized lasso.

Theorem 17 *Under the conditions of Theorem 16, hard-thresholding* $\tilde{\beta}$ *at* $t \sim \left(\frac{\log p}{N}\right)^{\frac{1}{2}} + \frac{s_{\max} \log p}{n}$ *gives*

1. $\|\tilde{\beta}^{ht} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac{1}{2}} + \frac{s_{\max} \log p}{n}$,
2. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \left(\frac{s_0 \log p}{N}\right)^{\frac{1}{2}} + \frac{\sqrt{s_0} s_{\max} \log p}{n}$,
3. $\|\tilde{\beta}^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s_0^2 \log p}{N}\right)^{\frac{1}{2}} + \frac{s_0 s_{\max} \log p}{n}$.

Theorem 17 shows that for $m \lesssim \frac{n}{s_{\max}^2 \log p}$, the variance term is dominant, so the convergence rates simplify:

1. $\|\tilde{\beta}^{ht} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac{1}{2}}$,
2. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \left(\frac{s_{\max} \log p}{N}\right)^{\frac{1}{2}}$,
3. $\|\tilde{\beta}^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s_{\max}^2 \log p}{N}\right)^{\frac{1}{2}}$.

Thus, estimator $\tilde{\beta}^{ht}$ shares the advantages of $\bar{\beta}^{ht}$ over the centralized lasso (cf. Remark 13). It also achieves computational gains over $\bar{\beta}^{ht}$ by amortizing the cost of debiasing across m machines.

4. A sharper estimation result

It is possible to obtain a sharper estimation result by forgoing the ℓ_∞ norm convergence rate. By sharper, we mean the sample complexity of the averaged estimator from $m \lesssim \frac{n}{s_0^2 \log p}$ to $m \lesssim \frac{n}{s_0 \log p}$.

The sharper estimation result depends on a result by Javanmard and Montanari (2013b), which we combine with Lemma 15 and restate for completeness. Before stating the results, we define the (∞, l) norm of a point $x \in \mathbf{R}^p$ as

$$\|x\|_{(\infty,l)} := \max_{\mathcal{A} \subset [p], |\mathcal{A}| \geq l} \frac{\|x_{\mathcal{A}}\|_2}{\sqrt{l}}.$$

When $l = 1$, the (∞, l) norm of x is its ℓ_∞ norm. When $l = p$, the (∞, l) norm is the ℓ_2 norm (rescaled by $\frac{1}{\sqrt{p}}$). Thus the (∞, l) norm interpolates between the ℓ_2 and ℓ_∞ norms. Javanmard and Montanari (2013b), Theorem 2.3 shows that the bias of the debiased lasso is of order $\frac{\sqrt{s_0} \log p}{n}$.

Lemma 18 *Under the conditions of Theorem 16,*

$$\|\hat{\Delta}_k\|_{(\infty, c' s_0)} \lesssim_P \frac{c\sqrt{s_0} \log p}{n} \text{ for any } k \in [m] \text{ for any } c' > 0,$$

where c is a constant that depends only on c' and Σ.

By Lemma 18, the estimator (5) is consistent in the (∞, s_0) norm. The argument is similar to the proof of Theorem 16.

Theorem 19 *Under the conditions of Theorem 16,*

$$\|\bar{\beta} - \beta^*\|_{(\infty, c' s_0)} \sim O_P\Big(\Big(\frac{\log p}{N}\Big)^{\frac{1}{2}} + \frac{\sqrt{s_0} \log p}{n}\Big).$$

Theorem 20 *Under the conditions of Theorem 16, hard-thresholding $\tilde{\beta}$ at $t = |\tilde{\beta}|_{(\hat{s}_0)}$ for some $\hat{s}_0 \sim s_0$, i.e. setting all but the largest $\hat{s}_0$ debiased coefficients to zero, gives*

1. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \left(\frac{s_0 \log p}{N}\right)^{\frac{1}{2}} + \frac{s_0 \log p}{n}$,
2. $\|\bar{\beta}^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s_0^2 \log p}{N}\right)^{\frac{1}{2}} + \frac{s_0^{3/2} \log p}{n}$.

By Theorem 20, when $m \lesssim \frac{N}{s_0 \log p}$, the variance term is dominant and the convergence rates given by the theorem simplify to the convergence rates of the (centralized) lasso estimator:

1. $\|\bar{\beta}^{ht} - \beta^*\|_2 \lesssim_P \left(\frac{s_0 \log p}{N}\right)^{\frac{1}{2}}$,
2. $\|\bar{\beta}^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s_0^2 \log p}{N}\right)^{\frac{1}{2}}$.

Thus, by forgoing estimation error in the ℓ_∞ norm, it is possible to reduce the sample complexity of the averaged estimator to $m \lesssim \frac{s_0 \log p}{N}$. When $m = 1$, we recover the sample complexity of the centralized lasso estimator.

5. Averaging debiased ℓ_1 regularized M-estimators

The distributed approach to debiasing extends readily to ℓ_1 regularized M-estimators. As before, we are given N pairs (x_i, y_i) stored on m machines. Let $\rho(y_i, a)$ be a loss function function, which is convex in a, and $\dot{\rho}$, $\ddot{\rho}$ be its derivatives with respect to a. That is

$$\dot{\rho}(y,a) = \frac{d}{da}\rho(y,a), \quad \ddot{\rho}(y,a) = \frac{d^2}{da^2}\rho(y,a).$$

We define $\ell_k(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho(y_i, x_i^T\beta)$, where the sum is only over the pairs on machine k. The averaged estimator is

$$\bar{\beta} := \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k + \hat{\Theta}\Big(\frac{1}{m}\sum_{k=1}^{m} \nabla \ell_k(\hat{\beta}_k)\Big), \tag{8}$$

where $\hat{\beta}_k$ is the local ℓ_1 regularized M-estimator: $\hat{\beta}_k := \arg\min_{\beta\in\mathbf{R}^p} \ell_k(\beta) + \lambda_k\|\beta\|_1$. As before, we form $\hat{\Theta}$ by nodewise regression on the weighted design matrix $X_{\hat{\beta}_k} := W_{\hat{\beta}_k} X_k$, where $W_{\hat{\beta}_k}$ is diagonal and its diagonal entries are

$$\big(W_{\hat{\beta}_k}\big)_{i,i} := \ddot{\rho}(y_i, x_i^T\hat{\beta}_k)^{\frac{1}{2}}.$$

That is, for some $j \in [p]$ that machine k is debiasing, the machine solves

$$\hat{\gamma}_j := \underset{\gamma\in\mathbf{R}^{p-1}}{\arg\min} \frac{1}{2n}\|X_{\hat{\beta}_k,j} - X_{\hat{\beta}_k,-j}\gamma\|_2^2 + \lambda_j\|\gamma\|_1,\ j \in [p],$$

and forms

$$\hat{\Theta}_{j,\cdot} = -\frac{1}{\hat{\tau}_j^2}\begin{bmatrix}\hat{\gamma}_{j,1} & \dots & \hat{\gamma}_{j,j-1} & 1 & \hat{\gamma}_{j,j+1} & \dots & \hat{\gamma}_{j,p}\end{bmatrix},$$

where

$$\hat{\tau}_j = \big(\frac{1}{n}\|X_{\hat{\beta}_k,j} - X_{\hat{\beta}_k,-j}\hat{\gamma}_j\|_2^2 + \lambda_j\|\hat{\gamma}_j\|_1\big)^{\frac{1}{2}}.$$

We assume

(B1) the pairs $\{(x_i, y_i)\}_{i\in[N]}$ are *i.i.d.*; the predictors are bounded:

$$\max_{i\in[N]} \|x_i\|_\infty \lesssim 1;$$

the projection of $X_{\beta^*,j}$ on $\mathcal{R}(X_{\beta^*,-j})$ in the $\mathbf{E}\left[\nabla^2\ell_k(\beta^*)\right]$ inner product is bounded: $\|X_{\beta^*,-j}\gamma_{\beta^*,j}\|_\infty \lesssim 1$ for any $j \in [p]$, where

$$\gamma_{\beta^*,j} := \underset{\gamma\in\mathbf{R}^{p-1}}{\arg\min}\, \mathbf{E}\left[\|X_{\beta^*,j} - X_{\beta^*,-j}\gamma\|_2^2\right].$$

(B2) the rows of $\mathbf{E}\left[\nabla^2\ell_k(\beta^*)\right]^{-1}$ are sparse: $\max_{j\in[p]} \frac{s_j^2\log p}{n} \sim o(1)$, where s_j is the sparsity of $\left(\mathbf{E}\left[\nabla^2\ell_k(\beta^*)\right]^{-1}\right)_{j,\cdot}$.

(B3) the smallest eigenvalue of $\mathbf{E}\left[\nabla^2\ell_k(\beta^*)\right]$ is bounded away from zero and its entries are bounded.

(B4) for any β such that $\|\beta - \beta^*\|_1 \leq \delta$ for some $\delta > 0$, the diagonal entries of W_β stays away from zero, and

$$|\ddot{\rho}(y, x^T\beta) - \ddot{\rho}(y, x^T\beta^*)| \leq |x^T(\beta - \beta^*)|.$$

(B5) we have $\frac{1}{n}\|X_k(\hat{\beta}_k - \beta^*)\|_2^2 \lesssim_P s_0\lambda_k^2$ and $\|\hat{\beta}_k - \beta^*\|_1 \lesssim_P s_0\lambda_k$.

(B6) the derivatives $\dot{\rho}(y, a)$, $\ddot{\rho}(y, a)$ is locally Lipschitz:

$$\max_{i\in[N]} \sup_{|a,a'-x_i^T\beta^*|\leq\delta} \sup_y \frac{|\ddot{\rho}(y,a)-\ddot{\rho}(y,a')|}{|a-a'|} \leq K \text{ for some } \delta > 0.$$

Further,

$$\max_{i\in[N]} \sup_y |\dot{\rho}(y, x_i^T\beta)| \sim O(1),$$
$$\max_{i\in[N]} \sup_{|a-x_i^T\beta^*|\leq\delta} \sup_y |\ddot{\rho}(y, a)| \sim O(1).$$

(B7) the diagonal entries of

$$\mathbf{E}\big[\nabla^2\ell_k(\beta^*)\big]^{-1}\,\mathbf{E}\big[\nabla\ell_k(\beta^*)\nabla\ell_k(\beta^*)^T\big]\,\mathbf{E}\big[\nabla^2\ell_k(\beta^*)\big]^{-1}$$

are bounded.

The preceding assumptions deserve elaboration. Assumptions (B1), (B4), (B6), and (B7) are standard in the literature on high-dimensional regression. They ensure the various intermediate quantities, such as $\rho(y, x^T\beta)$ and its derivaties, remain bounded. Assumption (B2) is perhaps the most restrictive. The assumption serves to ensure that the debiasing step is effective in reducing the bias of the regularized estimator. It may be relaxed (at the cost of additional technicalities) to the rows of $\mathbf{E}\left[\nabla^2\ell_k(\beta^*)\right]^{-1}$ admit a s_j-sparse approximation. We refer to Bühlmann and Van De Geer (2011) for the details. Assumption (B3) is a quantitative version of the usual rank condition in regression. It ensures the regression coefficients are identifiable in the limit. Assumption (B5) is not necessary; it is implied by the other assumptions. We refer to Bühlmann and Van De Geer (2011), Chapter 6 for the details. Here we state it as an assumption to simplify the exposition.

We are ready to state our main results concerning the averaged estimator(8). It shows the averaged estimator achieves the convergence rate of the centralized ℓ_1-regularized M-estimator.

Theorem 21 *Under (B1)–(B7),* (8) *with suitable parameters* $\lambda_j, \lambda_k \sim \left(\frac{\log p}{n}\right)^{\frac{1}{2}}$, $j \in [p]$, $k \in [m]$ *satisfies*

$$\|\bar{\beta} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac{1}{2}} + \frac{s_{\max}\log p}{n}, \tag{9}$$

where $s_{\max} := \max\{s_0, s_1, \ldots, s_p\}$.

Proof The averaged estimator is given by

$$\bar{\beta} - \beta^* = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k - \hat{\Theta}\nabla \ell_k(\hat{\beta}_k)(\hat{\beta}_k - \beta^*) - \beta^*.$$

By the smoothness of ρ,

$$\dot{\rho}(y_i, x_i^T\hat{\beta}_k) = \dot{\rho}(y_i, x_i^T\beta^*) + \ddot{\rho}(y_i, \tilde{a}_i)x_i^T(\hat{\beta}_k - \beta^*),$$

where $\tilde{a}_i$ is a point between $x_i^T\hat{\beta}_k$ and $x_i^T\beta^*$. Thus

$$\begin{aligned}\bar{\beta} - \beta^* &= \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_k - \hat{\Theta}(\nabla \ell_k(\beta^*) + Q_k(\hat{\beta}_k - \beta^*)) - \beta^* \\ &= -\hat{\Theta}\big(\tfrac{1}{m}\textstyle\sum_{k=1}^m \nabla \ell_k(\beta^*)\big) + \dfrac{1}{m}\displaystyle\sum_{k=1}^{m}\big(I - \hat{\Theta}Q_k\big)(\hat{\beta}_k - \beta^*).\end{aligned}$$

where $Q_k = \frac{1}{n}\sum_{i=1}^n \ddot{\rho}(y_i, \tilde{a}_i)x_i x_i^T$, where the sum is over the data points on machine k. Taking norms, we obtain

$$\|\bar{\beta} - \beta^*\|_\infty \le \big\|\hat{\Theta}\big(\tfrac{1}{m}\textstyle\sum_{k=1}^m \nabla \ell_k(\beta^*)\big)\big\|_\infty + \dfrac{1}{m}\displaystyle\sum_{k=1}^{m}\big\|\big(I - \hat{\Theta}Q_k\big)(\hat{\beta}_k - \beta^*)\big\|_\infty.$$

It is possible to show that $\big\|\hat{\Theta}\big(\frac{1}{m}\sum_{k=1}^m \nabla \ell_k(\beta^*)\big)\big\|_\infty \lesssim_P \big(\frac{\log p}{N}\big)^{\frac{1}{2}}$, which corresponds to the first term in (9). We refer to Bühlmann and Van De Geer (2011), Chapter 6 for the details.

We turn our attention to the second term. By the triangle inequality,

$$\begin{aligned}&\|(I - \hat{\Theta}Q_k)(\hat{\beta}_k - \beta^*)\|_\infty \\ &\quad\le \big\|\big(I - \hat{\Theta}\nabla^2\ell_k(\hat{\beta}_k)\big)(\hat{\beta}_k - \beta^*)\big\|_\infty + \big\|\hat{\Theta}(\nabla^2\ell_k(\hat{\beta}_k) - Q_k)(\hat{\beta}_k - \beta^*)\big\|_\infty \\ &\quad\le \max_{j\in[p]}\big\|e_j^T - \hat{\Theta}_{j,\cdot}\nabla^2\ell_k(\hat{\beta}_k)\big\|_\infty \|\hat{\beta}_k - \beta^*\|_1 \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n}\|\hat{\Theta}x_i\|_\infty\big|\ddot{\rho}(y_i, x_i^T\hat{\beta}_k) - \ddot{\rho}(y_i, \tilde{a}_i)x_i^T(\hat{\beta}_k - \beta^*)\big|.\end{aligned}$$

We proceed term by term. By (7),

$$\max_{j\in[p]}\big\|e_j^T - \hat{\Theta}_{j,\cdot}\nabla^2\ell_k(\hat{\beta}_k)\big\|_\infty \le \frac{\lambda_j}{\hat{\tau}_j^2} \lesssim \frac{1}{\hat{\tau}_j^2}\Big(\frac{\log p}{n}\Big)^{\frac{1}{2}}.$$

By van de Geer et al. (2013), Theorem 3.2,

$$|\hat{\tau}_j^2 - \tau_j^2| \lesssim_P \Big(\frac{\max\{s_0, s_j\}\log p}{n}\Big)^{\frac{1}{2}}$$

Thus $\max_{j\in[p]}\big\|e_j^T - \hat{\Theta}_{j,\cdot}\nabla^2\ell_k(\hat{\beta}_k)\big\|_\infty \lesssim_P \big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ and, by (B5),

$$\max_{j\in[p]}\big\|e_j^T - \hat{\Theta}_{j,\cdot}\nabla^2\ell_k(\hat{\beta}_k)\big\|_\infty \|\hat{\beta}_k - \beta^*\|_1 \lesssim_P \frac{s_{\max}\log p}{n}.$$

We turn our attention to the second term. We have $\|\hat{\Theta}x_i\|_\infty \lesssim_P 1$ because

$$\begin{aligned}\|\hat{\Theta}x_i\|_\infty &\leq \max_{j\in[p]} \|\hat{\Theta}_{j,\cdot}X_k^T\|_\infty \lesssim \max_{j\in[p]} \|\hat{\Theta}_{j,\cdot}X_{k,\beta^*}^T\|_\infty \\ &\leq \max_{j\in[p]} \frac{1}{\hat{\tau}_j^2}\|(X_{k,\beta^*})_j - (X_{k,\beta^*})_{-j}\hat{\gamma}_j\|_\infty.\end{aligned}$$

Again, by van de Geer et al. (2013), Theorem 3.2,

$$\begin{aligned}&\lesssim_P \max_{j\in[p]} \frac{1}{\tau_j^2}\|(X_{k,\beta^*})_j - (X_{k,\beta^*})_{-j}\hat{\gamma}_j\|_\infty \\ &\lesssim_P \max_{j\in[p]} \frac{1}{\tau_j^2}\|(X_{k,\beta^*})_j - (X_{k,\beta^*})_{-j}\gamma_j\|_\infty \\ &\quad + \frac{1}{\tau_j^2}\|(X_{k,\beta^*})_j\|_\infty\|(\hat{\gamma}_j - \gamma_j)\|_1.\end{aligned}$$

which, by (B1) and van de Geer et al. (2013), Theorem 3.2,

$$\lesssim_P 1 + \frac{s_j \log p}{n}.$$

Thus

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \|\hat{\Theta}x_i\|_\infty \big|\ddot{\rho}(y_i, x_i^T\hat{\beta}_k) - \ddot{\rho}(y_i, \tilde{a}_i)x_i^T(\hat{\beta}_k - \beta^*)\big| \\ &\quad \lesssim_P \frac{1}{n}\sum_{i=1}^n \big|\ddot{\rho}(y_i, x_i^T\hat{\beta}_k) - \ddot{\rho}(y_i, \tilde{a}_i)x_i^T(\hat{\beta}_k - \beta^*)\big|,\end{aligned}$$

which, by (B5) and (B6), is at most

$$\lesssim \frac{1}{n}\|X_k(\hat{\beta}_k - \beta^*)\|_2^2 \lesssim_P \frac{s_0 \log p}{n}.$$

We put the pieces together to deduce $\frac{1}{m}\sum_{k=1}^m \big\|\big(I - \hat{\Theta}Q_k\big)(\hat{\beta}_k - \beta^*)\big\|_\infty \lesssim_P \frac{s_{\max}\log p}{n}$. ■

By combining the Lemma 11 with Theorem 16, we can show that $\tilde{\beta}^{ht} := \mathrm{HT}(\tilde{\beta}, t)$ for an appropriate threshold t converges to β^* at the same rates as the centralized ℓ_1-regularized M-estimator.

Theorem 22 *Under the conditions of Theorem 21, hard-thresholding $\tilde{\beta}$ at $t \sim \big(\frac{\log p}{N}\big)^{\frac{1}{2}} + \frac{\max_{j\in[p]} s_j \log p}{n}$ gives*

1. $\|\tilde{\beta}^{ht} - \beta^*\|_\infty \lesssim_P \big(\frac{\log p}{N}\big)^{\frac{1}{2}} + \frac{s_{\max}\log p}{n}$,
2. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \big(\frac{s_0 \log p}{N}\big)^{\frac{1}{2}} + \frac{\sqrt{s_0}s_{\max}\log p}{n}$,

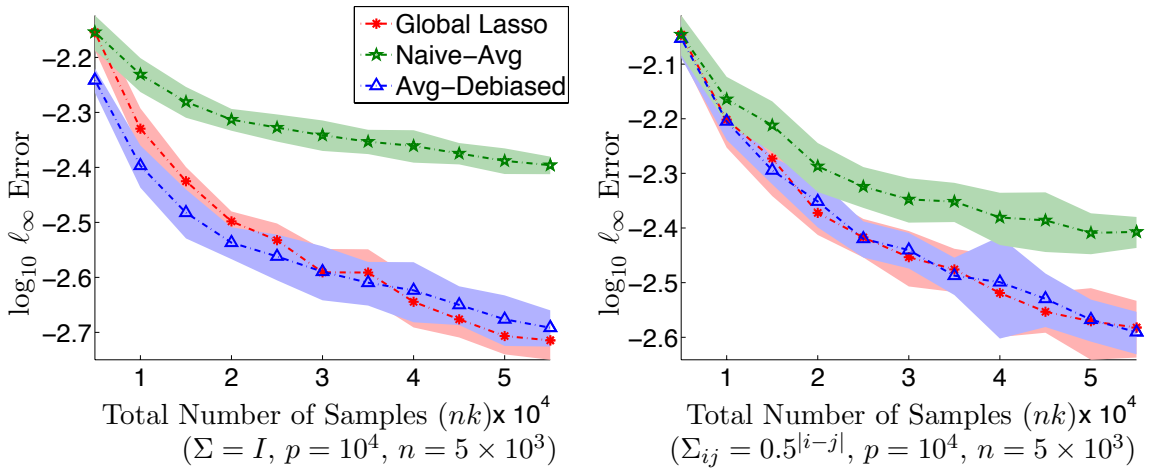


Figure 1: The estimation error (in ℓ_∞ norm) of the averaged debiased lasso estimator versus that of the centralized lasso when the predictors are Gaussian. In both settings, the estimation error of the averaged debiased estimator is comparable to that of the centralized lasso, while that of the naive averaged lasso is much worse.

3. $\|\tilde{\beta}^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s_0^2 \log p}{N}\right)^{\frac{1}{2}} + \frac{s_0 \max_{j\in[p]} s_j \log p}{n}$.

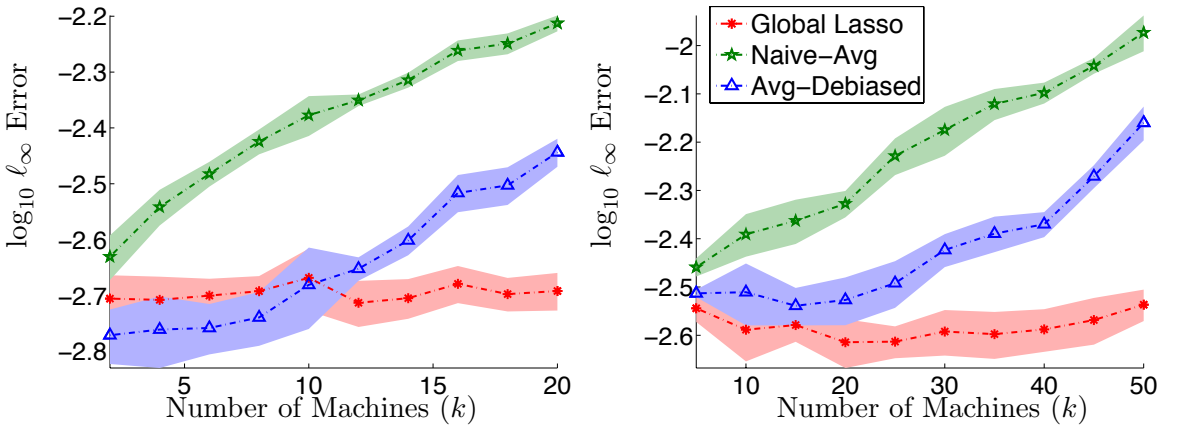
Assuming $s_0 \sim s_{\max}$, Theorem 22 shows when $m \lesssim \frac{n}{s_0^2 \log p}$, the variance term is dominant, so the convergence rates simplify to

1. $\|\tilde{\beta}^{ht} - \beta^*\|_\infty \lesssim_P \left(\frac{\log p}{N}\right)^{\frac{1}{2}}$,
2. $\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim_P \left(\frac{s_0 \log p}{N}\right)^{\frac{1}{2}}$,
3. $\|\tilde{\beta}^{ht} - \beta^*\|_1 \lesssim_P \left(\frac{s_0^2 \log p}{N}\right)^{\frac{1}{2}}$.

6. Simulations

We validate our theoretical results with simulations. First, we study the estimation error of the averaged debiased lasso in ℓ_∞ norm. To focus on the effect of averaging, we grow the number of machines m linearly with the (total) sample size N. In other words, we fix the sample size per machine n and grow the total sample size N by adding machines. The tuning parameters were set to their oracle values stated in the Theorem 12. Figure 1 compares the estimation error (in ℓ_∞ norm) of the averaged debiased lasso estimator with that of the centralized lasso. We see the estimation error of the averaged debiased lasso estimator is comparable to that of the centralized lasso, while that of the naive averaged lasso is much worse.

We conduct a second set of simulations to study the effect of the number of machines on the estimation effort of the averaged estimator. To focus on the effect of the number of



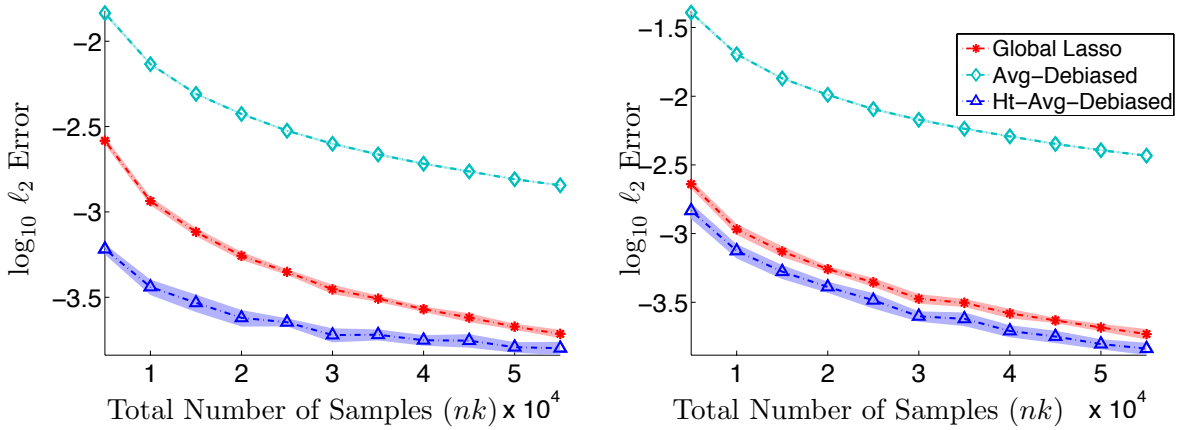
$(\Sigma = I,\ p = 10^4,\ nk = 2 \times 10^5)$ $(\Sigma_{ij} = 0.5^{|i-j|},\ p = 10^4,\ nk = 2 \times 10^5)$

Figure 2: The estimation error (in ℓ_∞ norm) of the averaged estimator as the number of machines k vary. When the number of machines is small, the error is comparable to that of the centralized lasso. However, when the number of machines exceeds a certain threshold, the bias term (which grows linearly in k) is dominant, and the performance of the averaged estimator degrades.

machines k, we fix the (total) sample size N and vary the number of machines the samples are distributed across. The tuning parameters were again set to the oracle values stated in the Theorem 12. Figure 2 shows how the estimation error (in ℓ_∞ norm) of the averaged estimator grows as the number of machines grows. When the number of machines is small, the estimation error of the averaged estimator is comparable to that of the centralized lasso. However, when the number of machines exceeds a certain threshold, the estimation error grows with the number of machines. This transition occurs when $\frac{s \log p}{n} \gtrsim \left(\frac{\log p}{N}\right)^{\frac{1}{2}}$, or equivalently, when $k \gtrsim \left(\frac{N}{s^2 \log p}\right)^{\frac{1}{2}}$. The preceding observation is consistent with the prediction of Lemma 10: when the number of machines exceeds a certain threshold, the bias term of order $\frac{s \log p}{n}$ becomes dominant. Since $\frac{s \log p}{n} \propto k$, we expect the error to grow linearly with k, which agrees with the trends in Figure 2.

We conduct a third set of simulations to study the effect of thresholding on the estimation error in ℓ_2 norm. The tuning parameters were set to the oracle values stated in the Theorem 12. Figure 3 compares the estimation error incurred by the averaged estimator with and without thresholding versus that of the centralized lasso. Since the averaged estimator is usually dense, its estimation error (in ℓ_2 norm) is large compared to that of the centralized lasso. However, after thresholding, the averaged estimator performs comparably versus the centralized lasso. This demonstrates the importance of the thresholding step to achieve low ℓ_2 error.

In practice, it is possible to set the tuning parameter δ by the bisection method in the accompanying code to Javanmard and Montanari (2013a); via bisection, they search for the smallest δ such that the optimization program in (2) is feasible. The lasso tuning parameter λ is set by first estimating the noise variance using the residuals and then using the formula



$(\Sigma = I,\ p = 10^4,\ n = 5 \times 10^3)$ $(\Sigma_{ij} = 0.5^{|i-j|},\ p = 10^4,\ n = 5 \times 10^3)$

Figure 3: The estimation error (in ℓ_2 norm) of the averaged estimator with and without thresholding versus that of the centralized lasso when the predictors are Gaussian. In both settings, thresholding reduces the estimation error by order(s) of magnitude. Although the estimation error of the averaged estimator is large compared to that of the centralized lasso, the thresholded averaged estimator performs comparably, or even better than, the centralized lasso.

$\lambda = \sigma\sqrt{2\log p}$. The parameter λ can be chosen independently of σ, if we replace the lasso with the sqrt-lasso Belloni et al. (2011); all of the same theoretical guarantees still apply, since the sqrt-lasso has the same consistency guarantees. For generalized linear models, the oracle choice of λ depends on known quantities Negahban et al. (2012).

7. Summary and discussion

We devised a communication-efficient approach to distributed sparse regression in the high-dimensional setting. The key idea is first "debiasing" local lasso estimators, and then averaging the debiased estimators. We show that as long as the data is not split across too many machines, the averaged estimator achieves the convergence rate of the centralized lasso estimator. In the appendix, we show that by foregoing consistency in the ℓ_∞ norm, it is possible to further reduce the sample complexity of the averaged estimator to that of the centralized lasso estimator. Further, the distributed approach to debiasing extends readily to other ℓ_1 regularized M-estimators. In concurrent work, the approach of averaging debiased M-estimators was proposed by Battey et al. (2015) for high-dimensional inference.

7.1 Communication and Computational complexity

In recent years, there has a been a flurry of work on establishing communication lower bounds for mean estimation in the Gaussian distribution. In other words, they establish the minimum communication C needed to obtain ℓ_2^2 risk R , where $\|\hat{\beta} - \beta^*\|_2^2 \leq R$ (Duchi et al., 2014; Garg et al., 2014). These results are not directly applicable to sparse linear

regression, since they do not impose sparsity on the mean. In Braverman et al. (2015), the authors established that to obtain risk $R \leq \frac{s \log p}{N}$ at least $\Omega\big(\frac{m \min(n,p)}{\log p}\big)$ bits of communication is required. Our approach communicates $\tilde{O}(mp)$ bits to achieve risk of $\frac{s \log p}{N}$, so is communication-optimal when $p \lesssim n$. To our knowledge, lowest known communication complexity for solving the lasso is at least $mp\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)} \log \frac{1}{\epsilon}$, for any desired accuracy $\epsilon \gtrsim \sqrt{\frac{s \log p}{n}}$ (Agarwal et al., 2012). This communication cost is larger than our algorithm by a multiplicative factor of $\frac{\sigma_{\max}(\Sigma)}{\sigma_{\min}(\Sigma)} \log \frac{1}{\epsilon}$, which is substantially larger when the problem is poorly conditioned.

In light of the fact that our approach is essentially optimal in terms of communication cost, we turn to the computational complexity of our method. The most intensive step of our approach is the computation of $\hat{\Theta}$. To compute one row of $\hat{\Theta}$ requires solving an optimization problem whose cost is equivalent to a lasso in dimension p with n samples. For the purpose of comparison, let us assume that the lasso solver performs T iterations. Thus the complexity of solving a lasso in dimension p with n samples is $O(npT)$. In the simple approach of Section 2 where each machine computes its own $\hat{\Theta}$, the parallel runtime is $O(np^2T)$. However using the approach of Section 3, each machine is only computing p/m rows of $\hat{\Theta}$. This brings down the parallel runtime to $O(\frac{np^2T}{m})$. In comparison, the cost of solving the lasso using a state-of-art optimization method such as Agarwal et al. (2012) has parallel runtime $O(npT)$, so our computational cost is larger by a factor of $O(\frac{p}{m})$. It is reasonable to think of $\frac{p}{m}$ as a constant, since as the number of variables p increases the dataset sizes increases and we will be forced to use a larger cluster size m due to memory constraints on a single machine. Although our computational complexity is larger in the distributed setting, the dominant factor is often bottlenecked by the communication and latency limitations, rather than local computation.

Appendix A. Proofs of Lemmas

Proof [Proof of Lemma 2] Let $z_i = \Sigma^{-\frac{1}{2}} x_i$. The generalized coherence between X and Σ^{-1} is given by

$$|||\Sigma^{-1}\hat{\Sigma} - I|||_\infty = |||\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)(\Sigma^{\frac{1}{2}} z_i)^T - I|||_\infty,$$

where $|||X|||_\infty$ is the maximum entry of a X. Each entry of $\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}} z_i)(\Sigma^{\frac{1}{2}} z_i)^T - I$ is a sum of independent subexponential random variables. Their subexponential norms are bounded by

$$\|(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k - \mathbf{1}\{j=k\}\|_{\psi_1} \leq 2\|(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k\|_{\psi_1},$$

where $\|X\|_{\psi_1}$ and $\|Y\|_{\psi_2}$ are the sub-exponential and sub-Gaussian norms of the random variables X and Y. Recall for any two subgaussian random variables X, Y, we have

$$\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

Thus

$$\|(\Sigma^{-\frac{1}{2}} z_i)_j(\Sigma^{\frac{1}{2}} z_i)_k - \delta_{j,k}\|_{\psi_1} \leq 4\|(\Sigma^{-\frac{1}{2}} z_i)_j\|_{\psi_2}\|(\Sigma^{\frac{1}{2}} z_i)_k\|_{\psi_2} \leq 4\sqrt{\kappa}\sigma_x^2,$$

where $\sigma_x = \|z_i\|_{\psi_2}$. By a Bernstein-type inequality,

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}}z_i)_j(\Sigma^{\frac{1}{2}}z_i)_k - \delta_{j,k} \geq t\Big) \leq 2e^{-c_1 \min\{\frac{nt^2}{\tilde{\sigma}_x^4}, \frac{nt}{\tilde{\sigma}_x^2}\}},$$

where $c_1 > 0$ is a universal constant and $\tilde{\sigma}_x^2 := 4\sqrt{\kappa}\sigma_x^2$. Since $\tilde{\sigma}_x^4 n > \log p$, we set $t = \frac{2\tilde{\sigma}_x^2}{\sqrt{c_1}}\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$ to obtain

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}}z_i)_j(\Sigma^{\frac{1}{2}}z_i)_k - \delta_{j,k} \geq \frac{2\tilde{\sigma}_x^2}{\sqrt{c_1}}\Big(\frac{\log p}{n}\Big)^{\frac{1}{2}}\Big) \leq 2p^{-4}.$$

We obtain the stated result by taking a union bound over the p^2 entries of $\frac{1}{n}\sum_{i=1}^{n}(\Sigma^{-\frac{1}{2}}z_i)(\Sigma^{\frac{1}{2}}z_i)^T - I$. ■

Proof [Proof of Lemma 5] By Vershynin (2010), Proposition 5.10,

$$\mathbf{Pr}\Big(\frac{1}{n}|x_j^T\epsilon| > t\Big) \leq e\exp\Big(-\frac{c_2n^2t^2}{\sigma_y^2\|x_j^T\|_2^2}\Big) \leq e\exp\Big(-\frac{c_2n^2t^2}{\sigma_y^2\max_{j\in[p]}\hat{\Sigma}_{j,j}}\Big).$$

We take a union bound over the p components of $\frac{1}{n}X^T\epsilon$ to obtain

$$\mathbf{Pr}\Big(\frac{1}{n}\|X^T\epsilon\|_\infty > t\Big) \leq e\exp\Big(-\frac{c_2n^2t^2}{\sigma_y^2\max_{j\in[p]}\hat{\Sigma}_{j,j}} + \log p\Big).$$

We set $t = \max_{j\in[p]}\hat{\Sigma}_{j,j}^{\frac{1}{2}}\sigma_y\big(\frac{3\log p}{c_2n}\big)^{\frac{1}{2}}$ to obtain the desired conclusion. ■

Proof [Proof of Lemma 7] We start by substituting the linear model into (1):

$$\hat{\beta}^d = \hat{\beta} + \frac{1}{n}\hat{\Theta}X^T(y - X\hat{\beta}) = \beta^* + \hat{\Theta}\hat{\Sigma}(\beta^* - \hat{\beta}) + \frac{1}{n}\hat{\Theta}X^T\epsilon.$$

By adding and subtracting $\hat{\Delta} = \beta^* - \hat{\beta}$, we obtain

$$\hat{\beta}^d = \beta^* + \frac{1}{n}\hat{\Theta}X^T(y - X\hat{\beta}) = \beta^* + (\hat{\Theta}\hat{\Sigma} - I)(\beta^* - \hat{\beta}) + \frac{1}{n}\hat{\Theta}X^T\epsilon.$$

We obtain the expression of $\hat{\beta}^d$ by setting $\hat{\Delta}$.

To show $\|\hat{\Delta}\|_\infty \leq \frac{3\delta}{\mu}s\lambda$, we apply Hölder's inequality to each component of $\hat{\Delta}$ to obtain

$$|(\hat{\Theta}\hat{\Sigma} - I)(\beta^* - \hat{\beta})| \leq \max_j \|\hat{\Sigma}m_j^T - e_j\|_\infty\|\hat{\beta} - \beta^*\|_1 \leq \delta\|\hat{\beta} - \beta^*\|_1, \tag{10}$$

where δ is the generalized incoherence between X and $\hat{\Theta}$. By Lemma 6, $\|\hat{\beta} - \beta^*\|_1 \leq \frac{3}{\mu}s\lambda$. We combine the bound on $\|\hat{\beta} - \beta^*\|_1$ with (10) to obtain the stated bound on $\|\hat{\Delta}\|_\infty$. ■

Proof [Proof of Lemma 10] By Lemma 7,

$$\bar{\beta} - \beta^* = \frac{1}{N}\sum_{k=1}^{m} \hat{\Theta}_k X_k^T \epsilon_k + \frac{1}{m}\sum_{k=1}^{m} \hat{\Delta}_k.$$

We take norms to obtain

$$\|\bar{\beta} - \beta^*\|_\infty \leq \Big\| \frac{1}{N}\sum_{k=1}^{m} \hat{\Theta}_k X_k^T \epsilon_k \Big\|_\infty + \frac{1}{m}\sum_{k=1}^{m} \|\hat{\Delta}_k\|_\infty.$$

We focus on bounding the first term. Let $a_j^T := e_j^T \begin{bmatrix} \hat{\Theta}_1 X_1^T & \dots & \hat{\Theta}_m X_m^T \end{bmatrix}$. By Vershynin (2010), Proposition 5.10,

$$\mathbf{Pr}\Big(\big| \frac{1}{N} a_j^T \epsilon \big| > t \Big) \leq e \exp\Big(- \frac{c_2 N^2 t^2}{\|a_j\|_2^2 \sigma_y^2} \Big)$$

for some universal constant $c_2 > 0$. Further,

$$\|a_j\|_2^2 = \sum_{k=1}^{m} \|X_k \hat{\Theta}_k^T e_j\|_2^2 = n \sum_{k=1}^{m} \big(\hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T \big)_{j,j} \leq c_\Omega N,$$

where $c_\Omega := \max_{j\in[p], k\in[m]} \big(\hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T \big)_{j,j}$. By a union bound over $j \in [p]$,

$$\mathbf{Pr}\Big(\max_{j\in[p]} \big| \frac{1}{N} a_j^T \epsilon \big| > t \Big) \leq e \exp\Big(- \frac{c_2 N t^2}{c_\Omega \sigma_y^2} + \log p \Big).$$

We set $t = \sigma_y \big(\frac{2 c_\Omega \log p}{c_2 N} \big)^{\frac{1}{2}}$ to deduce

$$\mathbf{Pr}\Big(\max_{j\in[p]} \big| \tfrac{1}{N} a_j^T \epsilon \big| \geq \sigma_y \big(\tfrac{2 c_\Omega \log p}{c_2 N} \big)^{\frac{1}{2}} \Big) \leq e p^{-1}.$$

We turn our attention to bounding the second term. By Lemma 5 and a union bound over $j \in [p]$, when we set

$$\lambda_1 = \dots = \lambda_m = \lambda := \max_{j\in[p], k\in[m]} \big((\hat{\Sigma}_k)_{j,j} \big)^{\frac{1}{2}} \sigma_y \Big(\frac{3 \log p}{c_2 n} \Big)^{\frac{1}{2}},$$

we have $\frac{1}{n} \|X_k^T \epsilon\|_\infty \leq \lambda$ for any $k \in [m]$ with probability at least $1 - \frac{em}{p^2} \geq 1 - e p^{-1}$. By Lemma 7, when

1. $\{\hat{\Sigma}_k\}_{k\in[m]}$ satisfy the RE condition on $\mathcal{C}(\mathrm{supp}(\beta^*))$ with constant μ_l,
2. $\{(\hat{\Sigma}_k, \hat{\Theta}_k)\}_{k\in[m]}$ have generalized incoherence $c_{\mathrm{GC}} \big(\frac{\log p}{n} \big)^{\frac{1}{2}}$,

the second term is at most $\frac{3\sqrt{3}}{\sqrt{c_2}} \frac{c_{\mathrm{GC}} c_\Sigma}{\mu_l} \sigma_y \frac{s \log p}{n}$. We put the pieces together to obtain

$$\|\bar{\beta} - \beta^*\|_\infty \leq \sigma_y \Big(\frac{2 c_\Omega \log p}{c_2 N} \Big)^{\frac{1}{2}} + \frac{3\sqrt{3}}{\sqrt{c_2}} \frac{c_{\mathrm{GC}} c_\Sigma}{\mu_l} \sigma_y \frac{s \log p}{n},$$

■

Proof [Proof of Lemma 10] We start with the conclusion of Lemma 10:

$$\|\bar{\beta} - \beta^*\|_\infty \leq \sigma_y \Big(\frac{2c_\Omega \log p}{c_2 N}\Big)^{\frac{1}{2}} + \frac{3\sqrt{3}}{\sqrt{c_2}} \frac{c_{\mathrm{GC}} c_\Sigma}{\mu_l} \sigma_y \frac{s \log p}{n}.$$

First, we show that the two constants $c_\Omega = \max_{j\in[p],\, k\in[m]} (\hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T)_{j,j}$ and $c_\Sigma := \max_{j\in[p],k\in[m]} ((\hat{\Sigma}_k)_{j,j})^{\frac{1}{2}}$ are bounded with high probability.

Lemma 23 *Under (A1),*

$$\mathbf{Pr}\big(\max_{j\in[p]} \Sigma_j^{-1} \hat{\Sigma} \Sigma_j^{-1} > 2 \max_{j\in[p]} \Sigma_{j,j}^{-1}\big) \leq 2pe^{-c_1 \min\{\frac{n}{\sigma_x^2}, \frac{n}{\sigma_x}\}}$$

for some universal constant $c_1 > 0$.

Proof [Proof of Lemma 23] We express

$$\Sigma_{j,\cdot}^{-1} \hat{\Sigma} \Sigma_{j,\cdot}^{-1} = \Sigma_{j,\cdot}^{-1} \hat{\Sigma} \Sigma_{j,\cdot}^{-1} - \Sigma_{j,j}^{-1} + \Sigma_{j,j}^{-1} = \frac{1}{n} \sum_{i=1}^n (x_i^T \Sigma_{\cdot,j}^{-1})^2 - \Sigma_{j,j}^{-1} + \Sigma_{j,j}^{-1}.$$

Since the subgaussian norm of $z_i = \Sigma^{-\frac{1}{2}} x_i$ is σ_x, $x_i^T \Sigma_{\cdot,j}^{-1}$ is also subgaussian with subgaussian norm bounded by

$$\|x_i^T \Sigma_{\cdot,j}^{-1}\|_{\psi_2} \leq \|z_i\|_{\psi_2} \|\Sigma_{\cdot,j}^{-\frac{1}{2}}\|_2 \leq \sigma_x (\Sigma_{j,j}^{-1})^{\frac{1}{2}}.$$

We recognize $\frac{1}{n}\sum_{i=1}^n (x_i^T \Sigma_{\cdot,j}^{-1})^2 - \Sigma_{j,j}^{-1}$ as a sum of *i.i.d.* subexponential random variables with subexponential norm bounded by

$$\|(x_i^T \Sigma_{\cdot,j}^{-1})^2 - \Sigma_{j,j}^{-1}\|_{\psi_1} \leq 2\|(x_i^T \Sigma_{\cdot,j}^{-1})^2\|_{\psi_1} \leq 4\|x_i^T \Sigma_{\cdot,j}^{-1}\|_{\psi_2}^2 \leq 4\sigma_x^2 \Sigma_{j,j}^{-1}.$$

By Vershynin (2010), Proposition 5.16, we have

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^n (x_i^T \Sigma_{\cdot,j}^{-1})^2 - \Sigma_{j,j}^{-1} > t\Big) \leq 2e^{-c_1 \min\{\frac{nt^2}{16\sigma_x^2 (\Sigma_{j,j}^{-1})^2}, \frac{nt}{4\sigma_x \Sigma_{j,j}^{-1}}\}}$$

for some absolute constant $c_1 > 0$. For $t = \Sigma_{j,j}^{-1}$, the bound simplifies to

$$\mathbf{Pr}\Big(\frac{1}{n}\sum_{i=1}^n (x_i^T \Sigma_{\cdot,j}^{-1})^2 - \Sigma_{j,j}^{-1} > \Sigma_{j,j}^{-1}\Big) \leq 2e^{-c_1 \min\{\frac{n}{16\sigma_x^2}, \frac{n}{4\sigma_x}\}}.$$

We take a union bound over $j \in [p]$ to obtain the stated result. ■

Since we form $\{\hat{\Theta}_k\}_{k\in[m]}$ by (2),

$$(\hat{\Theta}_k \hat{\Sigma}_k \hat{\Theta}_k^T)_{j,j} \leq \max_{j\in[p]} (\Sigma^{-1} \hat{\Sigma}_k \Sigma^{-1}))_{j,j}.$$

Lemma 23 implies

$$\max_{j\in[p]} (\Sigma^{-1} \hat{\Sigma}_k \Sigma^{-1}))_{j,j} \leq 2 \max_{j\in[p]} \Sigma_{j,j}^{-1} \text{ for each } k \in [m]$$

with probability at least $1 - 2pe^{-c_1 \min\{\frac{n}{\sigma_x^2}, \frac{n}{\sigma_x}\}}$.

Lemma 24 *Under (A1),*

$$\mathbf{Pr}(\max_{j\in[p]}(\hat{\Sigma}_{j,j})^{\frac{1}{2}} > \sqrt{2}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}) \le 2pe^{-c_1\min\{\frac{n}{16\sigma_x^2},\frac{n}{4\sigma_x}\}}$$

for some universal constant $c_1 > 0$.

We put the pieces together to obtain the stated result:

1. By Lemma 23 (and a union bound over $k \in [m]$),

$$\mathbf{Pr}(c_\Omega \ge 2\max_j \Sigma^{-1}_{j,j}) \le 2mpe^{-c_1\min\{\frac{n}{\sigma_x^2},\frac{n}{\sigma_x}\}}.$$

   Since $m \le p$, when $n > \frac{3}{c_1}\max\{\sigma_x^2,\sigma_x\}\log p$,

$$\mathbf{Pr}\big(c_\Omega < 2\max_j \Sigma^{-1}_{j,j}\big) \ge 1 - 2p^{-1}.$$

2. By Lemma 24 (and a union bound over $k \in [m]$),

$$\mathbf{Pr}(c_\Sigma < \sqrt{2}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}) \ge 1 - 2mpe^{-c_1\min\{\frac{n}{16\sigma_x^2},\frac{n}{4\sigma_x}\}}.$$

   When $n > \frac{3}{c_1}\max\{\sigma_x^2,\sigma_x\}\log p$, the right side is again at most $2p^{-1}$.

3. By Lemma 4, as long as

$$n > \max\{4000\tilde{s}\sigma_x^2\log(\tfrac{Cp}{\tilde{s}}),\, 8000\sigma_x^4\log p\},$$

   $\hat{\Sigma}_1,\ldots,\hat{\Sigma}_m$ all satisfy the RE condition with probability at least

$$1 - 2me^{-\frac{n}{4000\sigma_x^4}} \ge 1 - 2p^{-1}.$$

4. By Lemma 2,

$$\mathbf{Pr}\big(\cap_{k\in[m]}\mathcal{E}_{\mathrm{GC}}(\hat{\Sigma}_k)\big) \ge 1 - 2p^{-2}.$$

   Since $m < p$, the probability is at least $1 - 2p^{-1}$.

We apply the bounds $c_\Omega \le 2\max_{j\in[p]}\Sigma^{-1}_{j,j}$, $c_\Sigma \le \sqrt{2}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}$, $c_{\mathrm{GC}} = \frac{8}{\sqrt{c_1}}\sqrt{\kappa}\sigma_x^2$, and $\mu_l = \frac{1}{2}\lambda_{\min}(\Sigma)$ to obtain

$$\|\bar{\beta} - \beta^*\|_\infty \le \sigma_y\left(\frac{4\max_{j\in[p]}\Sigma^{-1}_{j,j}\log p}{c_2 N}\right)^{\frac{1}{2}} + \frac{48\sqrt{6}}{\sqrt{c_1 c_2}}\frac{\sqrt{\kappa}\max_{j\in[p]}(\Sigma_{j,j})^{\frac{1}{2}}}{\lambda_{\min}(\Sigma)}\sigma_x^2\sigma_y\frac{s\log p}{n}.$$

■

Proof [Proof of Lemma 24] We follow a similar argument as the proof of Lemma 23:

$$\hat{\Sigma}_{k;j,j} = \hat{\Sigma}_{j,j} = \hat{\Sigma}_{j,j} - \Sigma_{j,j} + \Sigma_{j,j} = \frac{1}{n}\sum_{i=1}^n x_{i,j}^2 - \Sigma_{j,j} + \Sigma_{j,j}.$$

Since the $z_i = \Sigma^{-\frac{1}{2}} x_i$ is subgaussian with subgaussian norm σ_x, $x_{i,j}$ is also subgaussian with subgaussian norm bounded by

$$\|x_{i,j}\|_{\psi_2} \leq \|\Sigma_{j,\cdot}^{\frac{1}{2}} z_i\|_{\psi_2} \leq \sigma_x (\Sigma_{j,j})^{\frac{1}{2}}.$$

We recognize $\hat{\Sigma}_{j,j} - \Sigma_{j,j} = \frac{1}{n}\sum_{i=1}^n x_{i,j}^2 - \Sigma_{j,j}$ as a sum of *i.i.d.* subexponential random variables with subexponential norm bounded by

$$\|\hat{\Sigma}_{j,j} - \Sigma_{j,j}\|_{\psi_1} \leq 2\|x_{i,j}^2\|_{\psi_1} \leq 4\|x_{i,j}\|_{\psi_2}^2 \leq 4\sigma_x^2 \Sigma_{j,j}.$$

By Vershynin (2010), Proposition 5.16, we have

$$\mathbf{Pr}(|\hat{\Sigma}_{j,j} - \Sigma_{j,j}| > t) \leq 2e^{-c_1 \min\{\frac{nt^2}{16\sigma_x^2 \Sigma_{j,j}^2}, \frac{nt}{\sigma_x \Sigma_{j,j}}\}}$$

for some absolute constant $c_1 > 0$. For $t = \Sigma_{j,j}$, the bound simplifies to

$$\mathbf{Pr}(|\hat{\Sigma}_{j,j} - \Sigma_{j,j}| > \Sigma_{j,j}) \leq 2e^{-c_1 \min\{\frac{n}{16\sigma_x^2}, \frac{n}{4\sigma_x}\}}.$$

We take a union bound over $j \in [p]$ to obtain the stated result. ■

Proof [Proof of Theorem 16] We start by substituting the linear model into (5):

$$\begin{aligned}
\tilde{\beta} &= \frac{1}{m}\sum_{k=1}^m \hat{\beta}_k - \hat{\Theta}\hat{\Sigma}_k(\hat{\beta}_k - \beta^*) + \frac{1}{n}\hat{\Theta} X_k^T \epsilon_k \\
&= \frac{1}{m}\sum_{k=1}^m \hat{\beta}_k - \hat{\Theta}\hat{\Sigma}_k(\hat{\beta}_k - \beta^*) + \frac{1}{N}\hat{\Theta} X^T \epsilon.
\end{aligned}$$

Subtracting β^* and taking norms, we obtain

$$\|\tilde{\beta} - \beta^*\|_\infty \leq \frac{1}{m}\sum_{k=1}^m \|(I - \hat{\Theta}\hat{\Sigma}_k)(\hat{\beta}_k - \beta^*)\|_\infty + \big\|\frac{1}{N}\hat{\Theta} X^T \epsilon\big\|_\infty. \tag{11}$$

By Vershynin (2010), Proposition 5.16, and Lemma (23), it is possible to show that

$$\big\|\frac{1}{N}\hat{\Theta} X^T \epsilon\big\|_\infty \lesssim_P \Big(\frac{\log p}{N}\Big)^{\frac{1}{2}}.$$

We turn our attention to the first term in (11). It's straightforward to see each term in the sum is bounded by

$$\begin{aligned}
&\|(I - \hat{\Theta}\hat{\Sigma}_k)(\hat{\beta}_k - \beta^*)\|_\infty \\
&\quad \leq \|(I - \Sigma^{-1}\hat{\Sigma}_k)(\hat{\beta}_k - \beta^*)\|_\infty + \|(\Sigma^{-1} - \hat{\Theta})\hat{\Sigma}_k(\hat{\beta}_k - \beta^*)\|_\infty \\
&\quad \leq \max_{j\in[p]} \|e_j^T - \Sigma_j^{-1}\hat{\Sigma}_k\|_\infty \|\hat{\beta}_k - \beta^*\|_1 + \|\Sigma_j^{-1} - \hat{\Theta}_{j,\cdot}\|_1 \|\hat{\Sigma}_k(\hat{\beta}_k - \beta^*)\|_\infty.
\end{aligned}$$

We put the pieces together to deduce each term is $O\big(\frac{s_{\max}\log p}{n}\big)$:

1. By Lemmas 4, 6, 24, $\|\hat{\beta}_k - \beta^*\|_1 \lesssim_P \sqrt{s_0}\lambda_k$.
2. By Lemma 15, $\|\Sigma_j^{-1} - \hat{\Theta}_{j,\cdot}\|_1 \lesssim_P s_j\big(\frac{\log p}{n}\big)^{\frac{1}{2}}$.
3. By the triangle inequality,

$$\|\hat{\Sigma}_k(\hat{\beta}_k - \beta^*)\|_\infty \leq \Big\|\frac{1}{n}X_k^T(y_k - X_k\hat{\beta}_k)\Big\|_\infty + \Big\|\frac{1}{n}X_k^T\epsilon_k\Big\|_\infty.$$

   By the optimality conditions of the (local) lasso estimators, the first term is λ_k, and it is possible to show, by Lemma 23 and Vershynin (2010), Proposition 5.16, that the second term is $O_P\big(\big(\frac{\log p}{n}\big)^{\frac{1}{2}}\big)$.

Since $\lambda_k \asymp \big(\frac{\log p}{n}\big)^{\frac{1}{2}}$, by a union bound over $k \in [m]$, we obtain

$$\|\bar{\beta} - \beta^*\|_\infty \sim O_P\Big(\Big(\frac{\log p}{N}\Big)^{\frac{1}{2}} + \frac{s_{\max}\log p}{n}\Big),$$

where $s_{\max} := \max\{s_0, s_1, \ldots, s_p\}$. ■

Proof [Proof of Lemma 18] The result is essentially Javanmard and Montanari (2013b), Theorem 2.3 with $\hat{\Omega} = \hat{\Theta}$ given by (6). Lemma 15 shows that

$$\max_{j\in[p]} \|\hat{\Theta}_{j,\cdot} - \Sigma_j^{-1}\|_1 \lesssim_P s_j\big(\tfrac{\log p}{n}\big)^{\frac{1}{2}},$$

Since $\frac{\max_{j\in[p]} s_j^2 \log p}{n} \sim o(1)$, $\hat{\Theta}$ satisfies the conditions of Javanmard and Montanari (2013b), Theorem 2.3:

$$\|\hat{\Delta}_k\|_{(\infty,c's_0)} \lesssim_P \frac{c\sqrt{s_0}\log p}{n} \text{ for any } k \in [m],$$

The bound is uniform in $k \in [m]$ by a union bound for suitable parameters $\lambda_k \sim \big(\frac{\log p}{n}\big)^{\frac{1}{2}}$. ■

Proof [Proof of Theorem 19] We start by substituting the linear model into (5):

$$\tilde{\beta} = \frac{1}{m}\sum_{k=1}^m \hat{\Delta}_k + \frac{1}{N}\hat{\Theta}X^T\epsilon.$$

Subtracting β^* and taking norms, we obtain

$$\|\tilde{\beta} - \beta^*\|_{(\infty,c's_0)} \leq \frac{1}{m}\sum_{k=1}^m \|\hat{\Delta}_k\|_{(\infty,c's_0)} + \big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\big\|_{(\infty,c's_0)}. \tag{12}$$

By Lemma 18, the first (bias) term is of order $\frac{c\sqrt{s_0}\log p}{n}$. We focus on showing the second (variance) term is of order $\big(\frac{\log p}{N}\big)^{\frac{1}{2}}$. Since the (∞, l) norm is non-increasing in l,

$$\big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\big\|_{(\infty,c's_0)} \leq \big\|\frac{1}{N}\hat{\Theta}X^T\epsilon\big\|_\infty.$$

By Vershynin (2010), Proposition 5.16 and Lemma 23, it is possible to show that

$$\|\frac{1}{N}\hat{\Theta}X^T\epsilon\|_\infty \sim O_P\Big(\Big(\frac{\log p}{N}\Big)^{\frac{1}{2}}\Big).$$

Thus the second term in (12) is of order $\big(\frac{\log p}{N}\big)^{\frac{1}{2}}$. We put all the pieces together to obtain the stated conclusion. ■

Proof [Proof of Theorem 20] Since $\tilde{\beta}^{ht} - \beta^*$ is $2s_0$-sparse,

$$\|\tilde{\beta}^{ht} - \beta^*\|_2^2 \lesssim s_0\|\tilde{\beta}^{ht} - \beta^*\|_{(\infty,c's_0)}^2$$

or, equivalently,

$$\|\tilde{\beta}^{ht} - \beta^*\|_2 \lesssim \sqrt{s_0}\|\tilde{\beta}^{ht} - \beta^*\|_{(\infty,c's_0)}.$$

By the triangle inequality,

$$\begin{aligned}\|\tilde{\beta}^{ht} - \beta^*\|_{(\infty,c's_0)} &\le \|\tilde{\beta}^{ht} - \tilde{\beta}\|_{(\infty,c's_0)} + \|\tilde{\beta} - \beta^*\|_{(\infty,c's_0)} \\ &\le 2\|\tilde{\beta} - \beta^*\|_{(\infty,c's_0)},\end{aligned}$$

where the second inequality is by the fact that thresholding at $t = |\tilde{\beta}|_{(c's_0)}$ minimizes $\|\beta - \beta^*\|_{(\infty,c's_0)}$ over $c's_0$-sparse points β. Thus

$$\|\tilde{\beta}^{ht} - \beta^*\|_2 = O_P\Big(\Big(\frac{s_0\log p}{N}\Big)^{\frac{1}{2}} + \frac{s_0\log p}{n}\Big).$$

To complete the proof, we observe that the estimation error bound of $\tilde{\beta}^{ht}$ in the ℓ_1 norm follows by the fact that $\tilde{\beta}^{ht} - \beta^*$ is $2s_0$-sparse. ■

References

Alekh Agarwal, Sahand Negahban, Martin J Wainwright, et al. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40 (5):2452–2482, 2012.

Heather Battey, Jianqing Fan, Han Liu, and Junwei Lu. Splitotic analysis for distributed estimation and hypothesis testing. *preprint (personal communication)*, 2015.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

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

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *The Journal of Machine Learning Research*, 13(1):165–202, 2012.

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

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

Ankit Garg, Tengyu Ma, and Huy L Nguyen. Lower bound for high-dimensional statistical learning problem via direct-sum theorem. *arXiv preprint arXiv:1405.1665*, 2014.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and its generalizations*. CRC Press, 2015.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013a.

Adel Javanmard and Andrea Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1311.0274*, 2013b.

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

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, 2010.

Jonathan Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *arXiv preprint arXiv:1407.2724*, 2014.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. *Information Theory, IEEE Transactions on*, 59(6):3434–3447, 2013.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603, 2010.