A. Illustrative Examples of General Sparse Learning Problems

In this section we discuss additional examples of high-dimensional statistical learning problems for which Theorem 6 is applicable.

A.1. Sparse Logistic Regression

For logistic model, performing maximum likelihood estimation (MLE) on (12) leads to the logistic loss function $\ell(y_{ji}, \langle \boldsymbol{\beta}, \mathbf{x}_{ji} \rangle) = \log(1 + \exp(-y_{ji}\langle \boldsymbol{\beta}, \mathbf{x}_{ji} \rangle))$. For high-dimensional problems, when we add a ℓ_1 regularization, we obtain the ℓ_1 regularized logistic regression model (Zhu & Hastie, 2004, Wu et al., 2009):

$$\widehat{\boldsymbol{\beta}}_{\text{centralize}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{mn} \sum_{j\in[m]} \sum_{i\in[n]} \log(1 + \exp(-y_{ji}\langle \boldsymbol{\beta}, \mathbf{x}_{ji} \rangle)) + \lambda ||\boldsymbol{\beta}||_1.$$

The logistic loss is $\frac{1}{4}$-smooth, and we also know $M = \frac{1}{4}$ because of self-concordance (Zhang & Xiao, 2015). Let $\mathcal{L}_j(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i\in[n]} \log(1 + \exp(-\mathbf{y}_{ji}\langle \boldsymbol{\beta}, \mathbf{x}_{ji} \rangle))$, (Negahban et al., 2012) showed that if $\mathbf{x}_{ji}$ are drawn from mean zero distribution with sub-Gaussian tails, then $\mathcal{L}_1(\boldsymbol{\beta})$ satisfies the restricted strong condition (5). Moreover, we have the following control on the quantity $\left|\left|\frac{1}{m}\sum_{j\in[m]} \nabla \mathcal{L}_j(\boldsymbol{\beta}^*)\right|\right|_\infty$.

Lemma 10. *Then we have the following upper bound holds in probability at least* $1 - \delta$*:*

$$\left|\left|\frac{1}{m} \sum_{j\in[m]} \nabla \mathcal{L}_j(\boldsymbol{\beta}^*)\right|\right|_\infty \lesssim ||\mathbf{x}_{ji}||_\infty \sqrt{\frac{2\log(p/\delta)}{mn}}.$$

The following ℓ_1 error bound states the estimation error for logistic regression with ℓ_1 regularization, which was established, for example, in (van de Geer, 2008, Negahban et al., 2012).

Lemma 11. *Under the model* (12)*, when* $n \geqslant (64/\kappa)s\log p$*, we have the following estimation error bound for* $\widehat{\boldsymbol{\beta}}_0$ *holds with probability at least* $1 - \delta$*:*

$$||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*||_1 \lesssim \frac{s\sigma_X}{\kappa}\sqrt{\frac{2\log(np/\delta)}{n}}.$$

With above analysis for sparse logistic regression model with random design, we are ready to present the results for the estimation error bound which established local exponential convergence.

Corollary 12. *Under sparse logistic regression model with random design, and set* λ_{t+1} *as* (9)*. If the following condition holds for some* $T \geqslant 0$*:*

$$||\widehat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*||_1 \leqslant 4\sqrt{\frac{\log(2p/\delta)}{n}}. \tag{13}$$

Then with probability at least $1 - 2\delta$*, we have the following estimation error bound for all* $t \geqslant T$*:*

$$||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 \leqslant \frac{1 - a_n^{t-T+1}}{1 - a_n} \frac{96s\sigma\sigma_X}{\kappa}\sqrt{\frac{\log(p/\delta)}{mn}} + 4a_n^{t-T+1}\sqrt{\frac{\log(2p/\delta)}{n}}, \tag{14}$$

$$||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2 \leqslant \frac{1 - a_n^{t-T+1}}{1 - a_n} \frac{4\sqrt{s}\sigma\sigma_X}{\kappa}\sqrt{\frac{\log(p/\delta)}{mn}} + 4a_n^{t-T}b_n\sqrt{\frac{\log(2p/\delta)}{n}}, \tag{15}$$

where

$$a_n = \frac{24s\sigma\sigma_X}{\kappa}\sqrt{\frac{\log(2p/\delta)}{n}} \quad \textit{and} \quad b_n = \frac{\sqrt{s}\sigma\sigma_X}{\kappa}\sqrt{\frac{\log(np/\delta)}{n}}.$$

A.2. High-dimensional Generalized Linear Models

The results are readily extendable to other high-dimensional generalized linear models (McCullagh & Nelder, 1989, van de Geer, 2008), where the response variable $y_{ji} \in \mathcal{Y}$ is drawn from the distribution

$$\mathbb{P}(y_{ji}|\mathbf{x}_{ji}) \propto \exp\left(\frac{y_{ji}\langle \mathbf{x}_{ji}, \boldsymbol{\beta}^* \rangle - \Phi(\langle \mathbf{x}_{ji}, \boldsymbol{\beta}^* \rangle)}{A(\sigma)}\right),$$

where $\Phi(\cdot)$ is a link function and $A(\sigma)$ is a scale parameter. Under the random subgaussian design, as long as the loss function has Lipschitz gradient, then the algorithm and corresponding estimation error bound and be applied.

A.3. High-dimensional Graphical Models

The results can also be used for the distributed unsupervised learning setting where the task is to learn a sparse graphical structure that represents the conditional independence between variables. Widely studied graphical models are Gaussian graphical models (Meinshausen & Bühlmann, 2006, Yuan & Lin, 2007) for continuous data and Ising graphical models (Ravikumar et al., 2010) for binary observations. As shown in (Meinshausen & Bühlmann, 2006, Ravikumar et al., 2010), these model selection problems can be reduced to solving parallel ℓ_1 regularized linear regression and logistic regression problems, respectively. Thus the approach presented in this paper can be readily applicable for these tasks.

B. Proofs

The section contains proofs of some theorems and lemmas stated in the main paper.

B.1. Proof of Lemma 8

Proof. Recall the definition of $\widetilde{\mathcal{L}}_1$ from (11). We have

$$\begin{aligned}\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t) &= \nabla\mathcal{L}_1(\boldsymbol{\beta}^*) + \frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t)\\ &= \frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\boldsymbol{\beta}^*) + \nabla\mathcal{L}_1(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t) - \left(\frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\boldsymbol{\beta}^*) - \frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t)\right).\end{aligned}$$

Using the triangle inequality

$$\begin{aligned}&\left\|\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t)\right\|_\infty\\ &\leqslant \left\|\frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\boldsymbol{\beta}^*)\right\|_\infty + \left\|\nabla\mathcal{L}_1(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t) - \left(\frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\boldsymbol{\beta}^*) - \frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t)\right)\right\|_\infty.\end{aligned}$$

We focus on bounding the second term in the right-hand-side inequality above. Let $\tau_{ji} = \ell'(y_{ji}, \langle\boldsymbol{\beta}^*, \mathbf{x}_{ji}\rangle)$ and define $\mathbf{v}_{ji}(\widehat{\boldsymbol{\beta}}_t) \in \mathbb{R}^p$:

$$\begin{aligned}\mathbf{v}_{ji}(\widehat{\boldsymbol{\beta}}_t) &= \mathbf{x}_{ji}(\ell'(y_{ji},\langle\boldsymbol{\beta}^*,\mathbf{x}_{ji}\rangle) - \ell'(y_{ji},\langle\widehat{\boldsymbol{\beta}}_t,\mathbf{x}_{ji}\rangle))\\ &= \tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T\left(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*\right) + \mathbf{x}_{ji}\frac{\ell'''(y_{ji},\mathbf{u}_{ji})}{2}(\langle\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*, \mathbf{x}_{ji}\rangle)^2\end{aligned}$$

where $\mathbf{u}_{ji}$ is a number between $\langle\widehat{\boldsymbol{\beta}}_t, \mathbf{x}_{ji}\rangle$ and $\langle\boldsymbol{\beta}^*, \mathbf{x}_{ji}\rangle$. With this notation

$$\begin{aligned}&\left\|\nabla\mathcal{L}_1(\boldsymbol{\beta}^*) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t) - \left(\frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\boldsymbol{\beta}^*) - \frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t)\right)\right\|_\infty\\ &\leqslant \left\|\frac{1}{n}\sum_{i\in[n]}\mathbf{v}_{1i}(\widehat{\boldsymbol{\beta}}_t) - \frac{1}{mn}\sum_j\sum_i\mathbf{v}_{ji}(\widehat{\boldsymbol{\beta}}_t)\right\|_\infty\\ &\leqslant \left\|\frac{1}{n}\sum_i\tau_{1i}\mathbf{x}_{1i}\mathbf{x}_{1i}^T(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*) - \frac{1}{mn}\sum_j\sum_i\tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*)\right\|_\infty + M\cdot\left(\max_{j,i}||\mathbf{x}_{ji}||_\infty^3\right)\cdot||\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*||_1^2.\end{aligned}$$

The first term above can be further upper bounded by

$$\begin{aligned}
&\left|\left|\frac{1}{n}\sum_j \tau_{1i}\mathbf{x}_{1i}\mathbf{x}_{1i}^T(\widehat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}^*)-\frac{1}{mn}\sum_j\sum_i \tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T(\widehat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}^*)\right|\right|_\infty \\
&\quad\leqslant \left|\left|\frac{1}{n}\sum_j \tau_{1i}\mathbf{x}_{1i}\mathbf{x}_{1i}^T-\frac{1}{mn}\sum_j\sum_i \tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T\right|\right|_\infty \cdot ||\widehat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}^*||_1 \,. \\
&\quad\leqslant \left(\left|\left|\frac{1}{n}\sum_{i\in[n]} \tau_{1i}\mathbf{x}_{1i}\mathbf{x}_{1i}^T-\mathbb{E}\left[\tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T\right]\right|\right|_\infty+\left|\left|\frac{1}{mn}\sum_j\sum_i \tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T-\mathbb{E}\left[\tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T\right]\right|\right|_\infty\right)\cdot ||\widehat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}^*||_1 \,.
\end{aligned}$$

Using Hoeffding's inequality together with a union bound, we have with probability at least $1-\delta$,

$$\left|\left|\frac{1}{n}\sum_{i\in[n]} \tau_{1i}\mathbf{x}_{1i}\mathbf{x}_{1i}^T-\mathbb{E}\left[\tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T\right]\right|\right|_\infty \leqslant L\left(\max_{j,i}||\mathbf{x}_{ji}||_\infty^2\right)\sqrt{\frac{2\log(2p/\delta)}{n}},$$

and

$$\left|\left|\frac{1}{mn}\sum_j\sum_i \tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T-\mathbb{E}\left[\tau_{ji}\mathbf{x}_{ji}\mathbf{x}_{ji}^T\right]\right|\right|_\infty \leqslant L\left(\max_{j,i}||\mathbf{x}_{ji}||_\infty^2\right)\sqrt{\frac{2\log(2p/\delta)}{mn}}.$$

Combining the bounds, the proof of the lemma is complete. □

B.2. Proof of Lemma 9

Proof. The proof uses ideas presented in (Negahban et al., 2012). By triangle inequality we have

$$\begin{aligned}
||\widehat{\boldsymbol{\beta}}_{t+1}||_1-||\boldsymbol{\beta}^*||_1 =&||\boldsymbol{\beta}^*+(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}+(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1-||\boldsymbol{\beta}^*||_1 \\
\geqslant&||\boldsymbol{\beta}^*+(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1-||\boldsymbol{\beta}^*||_1 \\
=&||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1.
\end{aligned}$$

By the optimality of $\widehat{\boldsymbol{\beta}}_{t+1}$ for (4), we have

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{\beta}}_{t+1},\widehat{\boldsymbol{\beta}}_t)+\lambda_{t+1}||\widehat{\boldsymbol{\beta}}_{t+1}||_1-\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t)-\lambda_{t+1}||\boldsymbol{\beta}^*||_1\leqslant 0.$$

Thus

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{\beta}}_{t+1},\widehat{\boldsymbol{\beta}}_t)-\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t)+\lambda_{t+1}(||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1)\leqslant 0.$$

By the convexity of $\widetilde{\mathcal{L}}_1(\cdot,\widehat{\boldsymbol{\beta}}_t)$, we further have

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{\beta}}_{t+1},\widehat{\boldsymbol{\beta}}_t)-\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t)\geqslant\langle\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t),\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*\rangle.$$

Thus by Hölder's inequality

$$\begin{aligned}
0\geqslant&\langle\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t),\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*\rangle+\lambda_{t+1}(||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1) \\
\geqslant&-||\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*,\widehat{\boldsymbol{\beta}}_t)||_\infty||\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*||_1+\lambda_{t+1}(||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1).
\end{aligned}$$

Under the assumption on λ_{t+1} we further have

$$\begin{aligned}
0\geqslant&-\frac{\lambda_{t+1}}{2}||\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*||_1+\lambda_{t+1}(||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1) \\
=&\frac{\lambda_{t+1}}{2}||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_{S^c}||_1-\frac{3\lambda_{t+1}}{2}||(\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*)_S||_1,
\end{aligned}$$

which completes the proof. □

B.3. Proof of Theorem 6

Proof. For the term $\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{\beta}}_{t+1}, \widehat{\boldsymbol{\beta}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}_t)$ we have

$$\begin{aligned}
\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{\beta}}_{t+1}, \widehat{\boldsymbol{\beta}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}_t) =& \mathcal{L}_1(\widehat{\boldsymbol{\beta}}_{t+1}) + \left\langle \frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t), \widehat{\boldsymbol{\beta}}_{t+1} \right\rangle \\
& - \mathcal{L}_1(\boldsymbol{\beta}^*) - \left\langle \frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t), \boldsymbol{\beta}^* \right\rangle \\
\geqslant & \langle \nabla\mathcal{L}_1(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^* \rangle + \kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2 \\
& + \left\langle \frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t), \widehat{\boldsymbol{\beta}}_{t+1} \right\rangle \\
& - \left\langle \frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t), \boldsymbol{\beta}^* \right\rangle \\
= & \left\langle \nabla\mathcal{L}_1(\boldsymbol{\beta}^*) + \frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\widehat{\boldsymbol{\beta}}_t) - \nabla\mathcal{L}_1(\widehat{\boldsymbol{\beta}}_t), \widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^* \right\rangle \\
& + \kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2 \\
= & \langle \nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}_t), \widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^* \rangle + \kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2,
\end{aligned}$$

where the first inequality we use the restricted strong convexity condition (5). Also by the optimality of $\widehat{\boldsymbol{\beta}}_{t+1}$ for (4), we have

$$\widetilde{\mathcal{L}}_1(\widehat{\boldsymbol{\beta}}_{t+1}, \widehat{\boldsymbol{\beta}}_t) - \widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}_t) + \lambda_{t+1}||\widehat{\boldsymbol{\beta}}_{t+1}||_1 - \lambda_{t+1}||\boldsymbol{\beta}^*||_1 \leqslant 0.$$

Combining above two inequalities we obtain with probability at least $1-\delta$:

$$\begin{aligned}
\lambda_{t+1}||\boldsymbol{\beta}^*||_1 - \lambda_{t+1}||\widehat{\boldsymbol{\beta}}_{t+1}||_1 \geqslant & \langle \nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}_t), \widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^* \rangle + \kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2 \\
\geqslant & -||\nabla\widetilde{\mathcal{L}}_1(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}_t)||_\infty||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 + \kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2 \\
\geqslant & -\frac{\lambda_{t+1}}{2}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 + \kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2.
\end{aligned}$$

By triangle inequality that $\lambda_{t+1}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 \geqslant \lambda_{t+1}||\boldsymbol{\beta}^*||_1 - \lambda_{t+1}||\widehat{\boldsymbol{\beta}}_{t+1}||_1$, we have

$$\begin{aligned}
\kappa||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2^2 \leqslant & \frac{3\lambda_{t+1}}{2}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 \\
= & \frac{3\lambda_{t+1}}{2}(||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_1 + ||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_{S^c}||_1) \\
\leqslant & \frac{3\lambda_{t+1}}{2}(||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_1 + 3||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_1) \\
= & 6\lambda_{t+1}||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_1 \\
\leqslant & 6\sqrt{s}\lambda_{t+1}||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_2 \\
\leqslant & 6\sqrt{s}\lambda_{t+1}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2.
\end{aligned}$$

We get

$$||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2 \leqslant \frac{6\sqrt{s}\lambda_{t+1}}{\kappa}.$$

Substitute λ_{t+1} in (9) concludes the proof for ℓ_2 estimation error bound. For $||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1$, we know

$$\begin{aligned}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 \leqslant& ||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_1 + ||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_{S^c}||_1 \\ \leqslant& 4||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_1 \leqslant 4\sqrt{s}||(\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*)_S||_2 \\ \leqslant& 4\sqrt{s}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2 \leqslant \frac{24s\lambda_{t+1}}{\kappa},\end{aligned}$$

which obtains the desired bound. □

B.4. Proof of Theorem 3

Proof. Theorem 3 follows from Theorem 6 after we verify some conditions. First, it is easy to see that the quadratic loss $L = 1, M = 0$. Under conditions of Theorem, with probability $1 - \delta$,

$$\left|\left|\frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\boldsymbol{\beta}^*)\right|\right|_\infty \lesssim \sigma\sigma_X\sqrt{\frac{\log(p/\delta)}{mn}}.$$

This follows from Corollary 5.17 of Vershynin (2012). Furthermore, with probability at least $1 - \delta$, we have

$$\max_{j\in[m],i\in[n]} ||\mathbf{x}_{ji}||_\infty \lesssim \sigma_X\sqrt{\log(mnp/\delta)}.$$

Finally,

$$||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*||_1 \lesssim \frac{s\sigma\sigma_X}{\kappa}\sqrt{\frac{\log(np/\delta)}{n}},$$

with probability at least $1 - \delta$ (Wainwright, 2009, Meinshausen & Yu, 2009, Bickel et al., 2009). Plugging these bounds into Theorem 6 completes the proof.

□

B.5. Proof of Corollary 7

Proof. The proof proceeds by recursively applying Theorem 6 and sum a geometric sequence. For notation simplicity let

$$\begin{aligned}a =& \frac{48s}{\kappa}\left|\left|\frac{1}{m}\sum_{j\in[m]} \nabla\mathcal{L}_j(\boldsymbol{\beta}^*)\right|\right|_\infty, \\ b =& \left(\frac{48sL}{\kappa}\left(\max_{j,i}||\mathbf{x}_{ji}||_\infty^2\right)\sqrt{\frac{4\log(2p/\delta)}{n}}\right), \\ c =& \frac{48sM}{\kappa}\left(\max_{j,i}||\mathbf{x}_{ji}||_\infty^3\right).\end{aligned}$$

By Theorem 6 we have

$$\begin{aligned}||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_1 \leqslant& a + b||\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*||_1 + c||\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*||_1^2 \\ \leqslant& a + 2b||\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*||_1 \\ \leqslant& a + 2b(a + 2b||\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^*||_1) \leqslant \ldots \\ \leqslant& a\sum_{k=0}^{t}(2b)^k + (2b)^{t+1}||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*||_1. \\ =& \frac{a(1-(2b)^{t+1})}{1-2b} + (2b)^{t+1}||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*||_1, \end{aligned} \tag{16}$$

which completes the ℓ_1 estimation error bound. For $||\widehat{\boldsymbol{\beta}}_{t+1} - \boldsymbol{\beta}^*||_2$, we first use (16) to obtain

$$||\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^*||_1 \leqslant \frac{a(1-(2b)^t)}{1-(2b)} + (2b)^t||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*||_1.$$

Then apply Theorem 6 to obtain that

$$\begin{aligned}||\widehat{\boldsymbol{\beta}}_{t+1}-\boldsymbol{\beta}^*||_2 \leqslant & \frac{a}{4\sqrt{s}}+\frac{(2b)}{4\sqrt{s}}||\widehat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}^*||_1 \leqslant \frac{a}{4\sqrt{s}}+\frac{b}{4\sqrt{s}}\left(\frac{a(1-(2b)^t)}{1-(2b)}+(2b)^t||\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}^*||_1\right)\\ =&\frac{1}{4\sqrt{s}}\left(a+\frac{a((2b)-(2b)^{t+1})}{1-(2b)}\right)+\frac{(2b)^{t+1}||\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}^*||_1}{4\sqrt{s}}\\ =&\frac{a(1-(2b)^{t+1})}{4\sqrt{s}(1-(2b))}+\frac{(2b)^{t+1}||\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}^*||_1}{4\sqrt{s}},\end{aligned}$$

which concludes the proof. □

B.6. Proof of Lemma 10

Proof. By the definition of $\mathcal{L}_j(\boldsymbol{\beta})$, we have

$$\frac{1}{m}\sum_{j\in[m]}\nabla\mathcal{L}_j(\boldsymbol{\beta}^*)=\frac{1}{mn}\sum_{j\in[m]}\sum_{i\in[n]}\mathbf{x}_{ji}\left(y_{ji}-\frac{y_{ji}}{1+\exp(-y_{ji}\langle\boldsymbol{\beta},\mathbf{x}_{ji}\rangle)}\right).$$

It is easy to check that

$$\mathbb{E}\left[y_{ji}-\frac{y_{ji}}{1+\exp(-y_{ji}\langle\boldsymbol{\beta},\mathbf{x}_{ji}\rangle)}\right]=0,\quad\text{and}\quad\left|y_{ji}-\frac{y_{ji}}{1+\exp(--y_{ji}\langle\boldsymbol{\beta},\mathbf{x}_{ji}\rangle)}\right|\leqslant 1$$

and thus

$$\begin{aligned}&\mathbb{E}\left[\mathbf{x}_{ji}\left(y_{ji}-\frac{y_{ji}}{1+\exp(-y_{ji}\langle\boldsymbol{\beta},\mathbf{x}_{ji}\rangle)}\right)\right]=0,\\ &\left|\left|\mathbf{x}_{ji}\left(y_{ji}-\frac{y_{ji}}{1+\exp(--y_{ji}\langle\boldsymbol{\beta},\mathbf{x}_{ji}\rangle)}\right)\right|\right|_\infty\leqslant\max_{ji}\left(||\mathbf{x}_{ji}||_\infty\right).\end{aligned}$$

Appling Azuma-Hoeffding inequality (Hoeffding, 1963) and the union bound over $[p]$ leads to the desired bound. □

C. Full Experimental Results

We run the algorithms for both distributed regression and classification problems. The algorithms to be compared are:

- Local: the first machine just solves a related ℓ_1 regularized problem (lasso or ℓ_1 regularized logistic regression) with the optimal λ, and outputs the solution. Obviously this approach is communication free.
- Centralize: the master gathers all data from different machines together, and solves a centralized ℓ_1 regularized loss minimization problem with the optimal λ, and outputs the solution. This approach is communication expensive as all data needs to be communicated, but it usually gives us the best estimation and prediction performance.
- Prox GD: the distributed proximal gradient descent is ran on the ℓ_1 regularized objective, where we initialized the starting point with the first machine's solution.
- Avg-Debias: the method proposed in Lee et al. (2015b), with fine tuned regularization and hard thresholding parameters. This approach only requires one round of communication, where each machine sends a p-dimensional vector. However, Avg-Debias is computationally prohibitive because of the debiasing operation.
- EDSL: the proposed efficient distributed sparse learning approach, where the regularization level at each iteration is fine tuned on a held out test data set.

C.1. Simulations

The full experimental results plotted in Figure 3 and Figure 4, with various settings of (n, p, m, s), and condition numbers $1/\kappa$. We have the following observations:

Table 2. List of real-world datasets used in the experiments.

Name	#Instances	#Features	Task
a9a	48,842	123	Classification
connect-4	67,557	127	Regression
dna	2,000	181	Regression
mitface	6,977	362	Classification
mnist 1 vs 2	14,867	785	Classification
mnist	60,000	785	Regression
mushrooms	8,124	113	Classification
protein	17,766	358	Regression
spambase	4,601	57	Classification
usps	7,291	257	Regression
w8a	64,700	301	Classification
year	51,630	91	Regression

- The Avg-Debias approach obtained much better estimation error compared to Local after one round of communication and sometimes performed quite close to Centralize. However, in most cases, there is still a gap compared with Centralize, especially when the problem is not well-conditioned or the number of machines m is large.
- When the problem is well conditioned ($\Sigma_{ij} = 0.5^{|i-j|}$ case), Prox GD converges reasonably fast. However, it becomes very slow when the condition number becomes bad ($\Sigma_{ij} = 0.5^{|i-j|/5}$ case). We expect to observe a similar phenomenon for other first-order distributed optimization algorithms, such as accelerated proximal gradient or ADMM.
- As theory suggests, EDSL obtained a solution that is competitive with Avg-Debias after one round of communication. The estimation error decreases to match performance of Centralize within few rounds of communications; typically less than 5, even though the theory suggests EDSL will match the performance of centralize within $\mathcal{O}(\log m)$ rounds of communication.

C.2. Real-world Data Evaluation

In real world data evaluation presented in Section 5.2, the datasets are publicly available from the LIBSVM website[7] and UCI Machine Learning Repository[8]. The statistics of these datasets are summarized in Table 2, where some of the multi-class classification datasets are adopted under the regression setting with squared losses. The results are plotted in Figure 5 where for some datasets the performance of Avg-Debias is significantly worse than others (mostly because the debiasing step fails), thus we omit these plots. The plots are shown in Figure 5 We have the following observations

- Since there is no well-specified model on these datasets, the curves behave quite differently on different data sets. However, a large gap between the local and centralized procedure is consistent as the later uses 10 times more data.
- Avg-Debias often fails on these real datasets and performs much worse than in simulations. The main reason might be that the assumptions, such as well-specified model or generalized coherence condition, fail, then Avg-Debias can totally fail and produce solution even much worse than the local.
- Prox GD approach still converges slowly in most of the cases.
- The proposed EDSL is quite robust on real world data sets, and can output a solution which is highly competitive with the centralized model within a few rounds of communications.
- There exits a slight "zig-zag" behavior for EDSL approach on some data sets. For example, on the mushrooms data set, the predictive performance of EDSL is not stable.

[7] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
[8] http://archive.ics.uci.edu/ml/

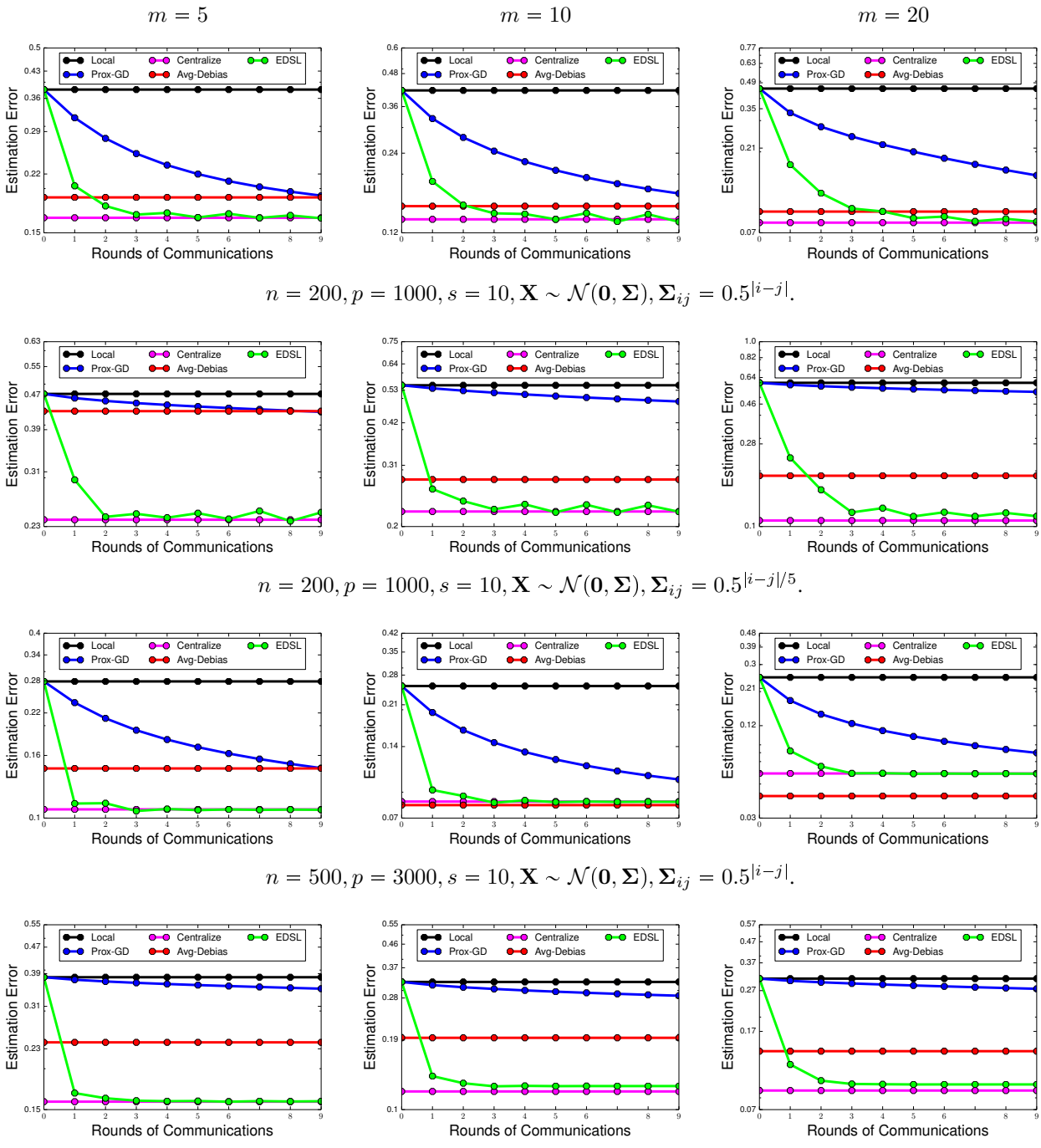
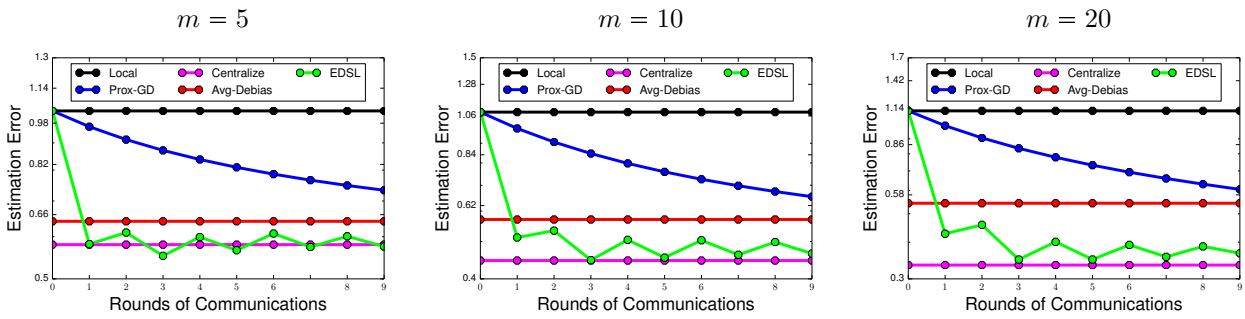
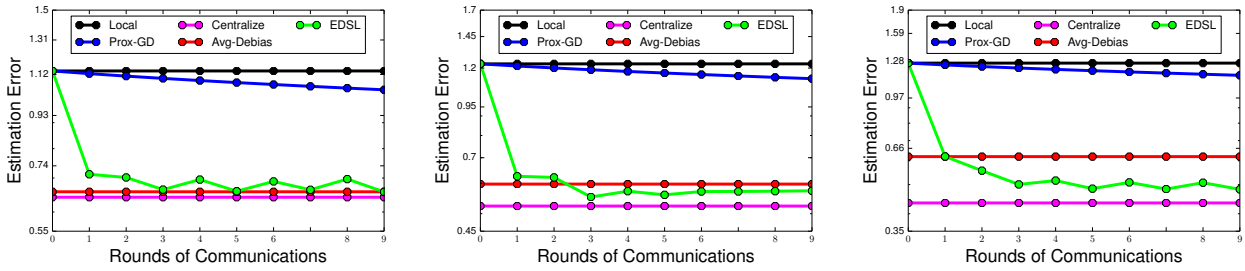


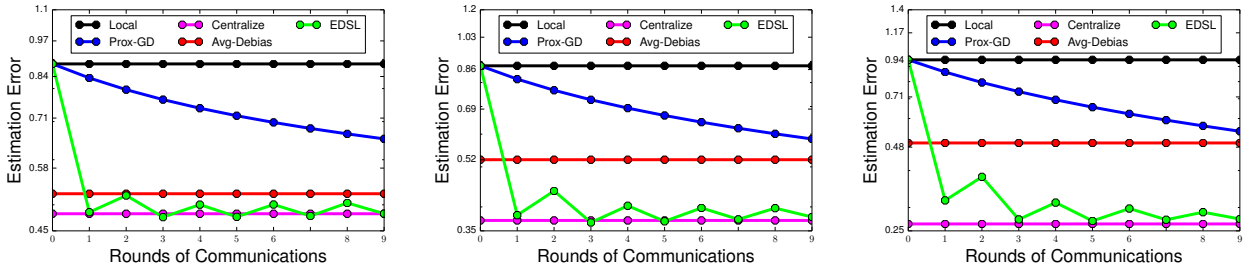
Figure 3. Comparison of various algorithms for distributed sparse regression, 1st and 3rd row: well-conditioned cases, 2nd and 4th row: ill-conditioned cases.



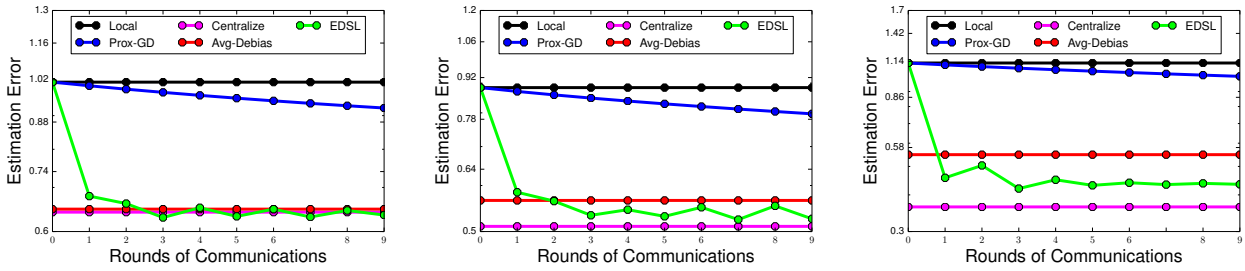
$n = 500, p = 1000, s = 10, \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \mathbf{\Sigma}_{ij} = 0.5^{|i-j|}$.



$n = 500, p = 1000, s = 10, \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \mathbf{\Sigma}_{ij} = 0.5^{|i-j|/5}$.



$n = 1000, p = 3000, s = 10, \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \mathbf{\Sigma}_{ij} = 0.5^{|i-j|}$.



$n = 1000, p = 3000, s = 10, \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \mathbf{\Sigma}_{ij} = 0.5^{|i-j|/5}$.

Figure 4. Comparison of various algorithms for distributed sparse classification (logistic regression), 1st and 3rd row: well-conditioned cases, 2nd and 4th row: ill-conditioned cases.

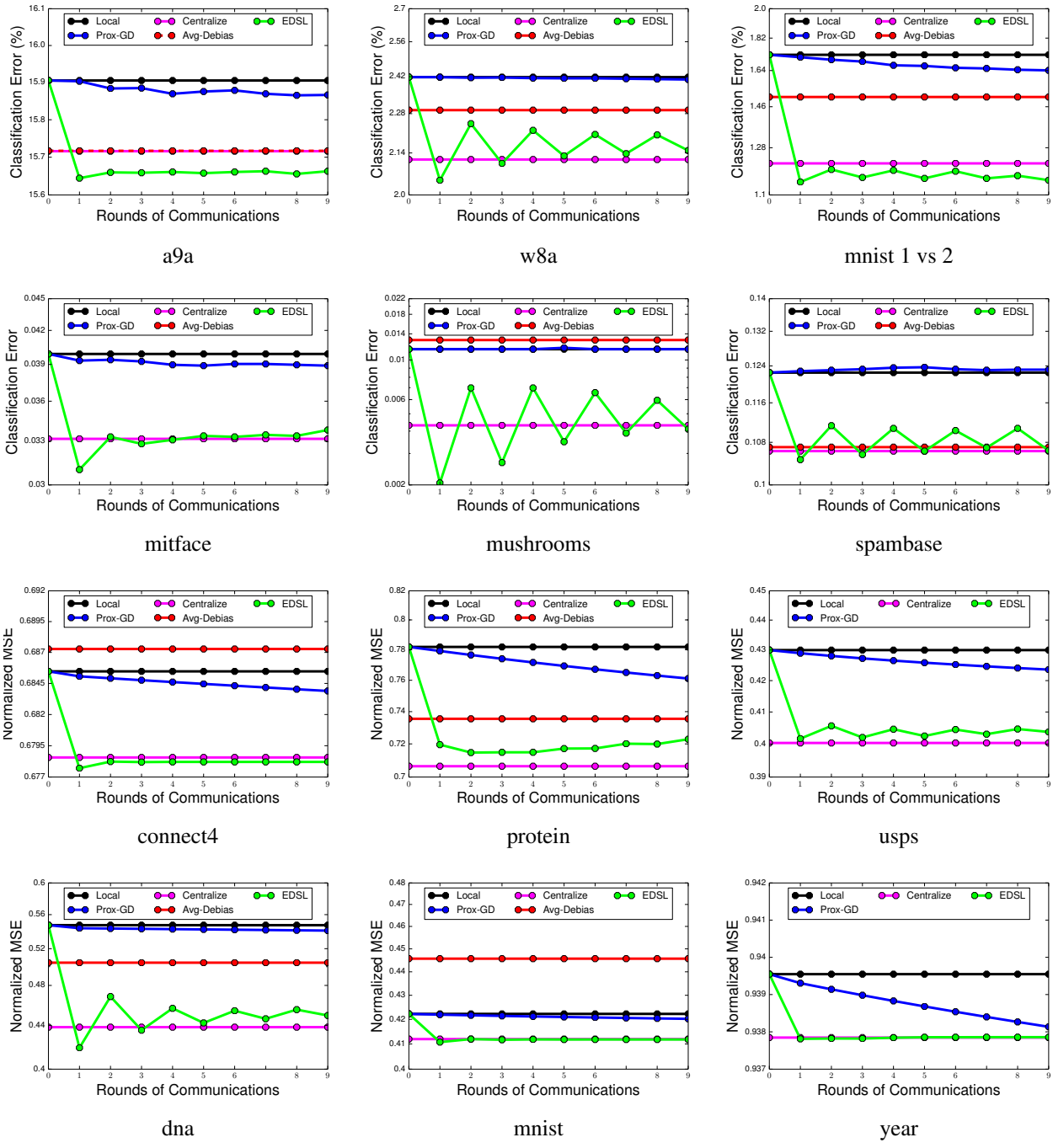


Figure 5. Comparison of various approaches for distributed sparse regression and classification on real world datasets. (Avg-Debias is omitted when it is significantly worse than others.)