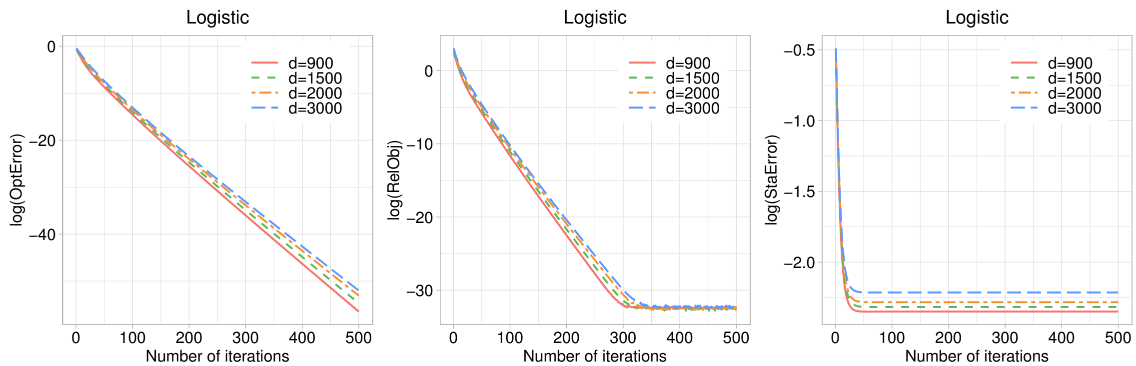

Fig. 9: Performances for logistic regression under different dimensions d.

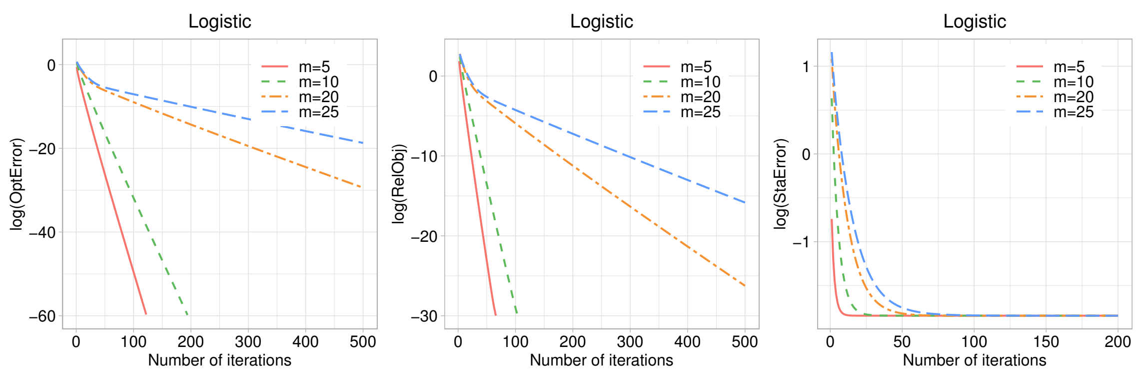

Fig. 10: Performances for logistic regression under different numbers of node m with fixed $N = 5000$.

Appendix
Proof of Some Technical Results.

Proof of Proposition 2. Using restricted smoothness of f,

$$\begin{aligned}
& f(\boldsymbol{\theta}_i^{t+1}) + \lambda\|\boldsymbol{\theta}_i^{t+1}\|_1 \\
\leq\ & f(\mathbf{z}_i^t) + \langle \nabla f(\mathbf{z}_i^t), \boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t \rangle + \frac{L}{2}\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|^2 \\
& + \tau_L \|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|_1^2 + \lambda\|\boldsymbol{\theta}_i^{t+1}\|_1 \\
=\ & f(\mathbf{z}_i^t) + \langle \mathbf{s}_i^t, \boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t \rangle + \frac{1}{2\eta}\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|^2 \\
& + \lambda\|\boldsymbol{\theta}_i^{t+1}\|_1 + \langle \nabla f(\mathbf{z}_i^t) - \mathbf{s}_i^t, , \boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t \rangle \\
& + (\frac{L}{2} - \frac{1}{2\eta})\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|^2 + \tau_L \|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|_1^2.
\end{aligned}$$

Let $\boldsymbol{\theta}_\alpha = \alpha\widehat{\boldsymbol{\theta}} + (1-\alpha)\mathbf{z}_i^t$, $\alpha \in [0,1]$. Define $\mathfrak{m}(\boldsymbol{\theta}) := f(\mathbf{z}_i^t) + \langle \mathbf{s}_i^t, \boldsymbol{\theta} - \mathbf{z}_i^t \rangle + \frac{1}{2\eta}\|\boldsymbol{\theta} - \mathbf{z}_i^t\|^2 + \lambda\|\boldsymbol{\theta}\|_1$. Using that $\boldsymbol{\theta}_i^{t+1}$ is the minimizer of $\mathfrak{m}(\boldsymbol{\theta})$, we have $\mathfrak{m}(\boldsymbol{\theta}_i^{t+1}) \leq \mathfrak{m}(\boldsymbol{\theta}_\alpha)$ and thus

$$\begin{aligned}
& f(\boldsymbol{\theta}_i^{t+1}) + \lambda\|\boldsymbol{\theta}_i^{t+1}\|_1 \\
\leq\ & f(\mathbf{z}_i^t) + \langle \mathbf{s}_i^t, \boldsymbol{\theta}_\alpha - \mathbf{z}_i^t \rangle + \frac{1}{2\eta}\|\boldsymbol{\theta}_\alpha - \mathbf{z}_i^t\|^2 + \lambda\|\boldsymbol{\theta}_\alpha\|_1 \\
& + \langle \nabla f(\mathbf{z}_i^t) - \mathbf{s}_i^t, , \boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t \rangle + (\frac{L}{2} - \frac{1}{2\eta})\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|^2 \\
& + \tau_L \|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|_1^2 \\
=\ & f(\mathbf{z}_i^t) + \langle \nabla f(\mathbf{z}_i^t), \boldsymbol{\theta}_\alpha - \mathbf{z}_i^t \rangle + \frac{1}{2\eta}\|\boldsymbol{\theta}_\alpha - \mathbf{z}_i^t\|^2 + \lambda\|\boldsymbol{\theta}_\alpha\|_1 \\
& + \langle \nabla f(\mathbf{z}_i^t) - \mathbf{s}_i^t, \boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t \rangle + \langle \nabla f(\mathbf{z}_i^t) - \mathbf{s}_i^t, \mathbf{z}_i^t - \boldsymbol{\theta}_\alpha \rangle \\
& + (\frac{L}{2} - \frac{1}{2\eta})\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|^2 + \tau_L \|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|_1^2.
\end{aligned}$$

Using $\langle \nabla f(\mathbf{z}_i^t), \boldsymbol{\theta}_\alpha - \mathbf{z}_i^t \rangle = \alpha\langle \nabla f(\mathbf{z}_i^t), \widehat{\boldsymbol{\theta}} - \mathbf{z}_i^t \rangle \leq -\frac{\alpha\mu}{2}\|\widehat{\boldsymbol{\theta}} - \mathbf{z}_i^t\|^2 + \alpha\tau_\mu\|\widehat{\boldsymbol{\theta}} - \mathbf{z}_i^t\|_1^2 + \alpha f(\widehat{\boldsymbol{\theta}}) - \alpha f(\mathbf{z}_i^t)$ (RSC) and $\lambda\|\boldsymbol{\theta}_\alpha\|_1 \leq \lambda\alpha\|\widehat{\boldsymbol{\theta}}\|_1 + \lambda(1-\alpha)\|\mathbf{z}_i^t\|_1$ (using the convexity of penalty), we then get

$$\begin{aligned}
& f(\boldsymbol{\theta}_i^{t+1}) + \lambda\|\boldsymbol{\theta}_i^{t+1}\|_1 \\
\leq\ & (1-\alpha)(f(\mathbf{z}_i^t) + \lambda\|\mathbf{z}_i^t\|_1) + \alpha(f(\widehat{\boldsymbol{\theta}}) + \lambda\|\widehat{\boldsymbol{\theta}}\|_1) \\
& - \frac{\alpha\mu}{2}\|\mathbf{z}_i^t - \widehat{\boldsymbol{\theta}}\|^2 + \alpha\tau_\mu\|\mathbf{z}_i^t - \widehat{\boldsymbol{\theta}}\|_1^2 + \frac{\alpha^2}{2\eta}\|\mathbf{z}_i^t - \widehat{\boldsymbol{\theta}}\|^2 \\
& + (\frac{L}{2} - \frac{1}{2\eta})\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|^2 + \tau_L \|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|_1^2 \\
& + \langle \nabla f(\mathbf{z}_i^t) - \mathbf{s}_i^t, \boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t \rangle + \alpha\langle \nabla f(\mathbf{z}_i^t) - \mathbf{s}_i^t, \mathbf{z}_i^t - \widehat{\boldsymbol{\theta}} \rangle.
\end{aligned}$$

Using $\|\boldsymbol{\theta}_i^{t+1} - \mathbf{z}_i^t\|_1^2 \leq 64s(\|\boldsymbol{\theta}_i^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{z}_i^t - \widehat{\boldsymbol{\theta}}\|^2) + 8\nu^2$, $\|\mathbf{z}_i^t - \widehat{\boldsymbol{\theta}}\|_1^2 \leq 32s\|\mathbf{z}_i^t - \widehat{\boldsymbol{\theta}}\|^2 + 2\nu^2$, and summing over i, we have

$$\begin{aligned}
& \sum_i (\varphi(\boldsymbol{\theta}_i^{t+1}) - \varphi(\widehat{\boldsymbol{\theta}})) \\
\leq\ & (1-\alpha)\sum_i (\varphi(\mathbf{z}_i^t) - \varphi(\widehat{\boldsymbol{\theta}})) - \frac{\alpha\mu}{2}\|\mathbf{Z}^t - \widehat{\boldsymbol{\Theta}}\|^2 \\
& + \alpha\tau_\mu(32s\|\mathbf{Z}^t - \widehat{\boldsymbol{\Theta}}\|^2 + 2m\nu^2) + \frac{\alpha^2}{2\eta}\|\mathbf{Z}^t - \widehat{\boldsymbol{\Theta}}\|^2
\end{aligned}$$

$$+(\frac{L}{2}-\frac{1}{2\eta})\|\boldsymbol{\Theta}^{t+1}-\mathbf{Z}^t\|^2+\tau_L(64s\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2 +64s\|\mathbf{Z}^t-\widehat{\boldsymbol{\Theta}}\|^2+8m\nu^2)+\sum_i\langle\nabla f(\mathbf{z}_i^t)-\mathbf{s}_i^t,\boldsymbol{\theta}_i^{t+1}-\mathbf{z}_i^t\rangle +\alpha\sum_i\langle\nabla f(\mathbf{z}_i^t)-\mathbf{s}_i^t,\mathbf{z}_i^t-\widehat{\boldsymbol{\theta}}\rangle,$$

which is the main iterative relationship connecting iteration t and iteration $t+1$. $\square$

Proof of Proposition 3. The proofs for the two bounds are essentially the same, and we only show the second one. For simplicity of notation, without loss of generality, we assume $T=0$ in the proof. We write

$$\begin{aligned}
&\langle\nabla f(\mathbf{z}_i^t)-\mathbf{s}_i^t,\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle\\
=\;&\langle\nabla f(\mathbf{z}_i^t)-\nabla f(\bar{\mathbf{z}}^t),\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle\\
&+\langle\nabla f(\bar{\mathbf{z}}^t)-\frac{1}{m}\sum_j\nabla f_j(\mathbf{z}_j^t),\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle\\
&+\langle\frac{1}{m}\sum_j\nabla f_j(\mathbf{z}_j^t)-\mathbf{s}_i^t,\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle. \qquad (15)
\end{aligned}$$

We note that $\mathbf{Z}^t=\mathbf{W}\boldsymbol{\Theta}^t$ and thus

$$\begin{aligned}
\mathbf{Z}^t-\bar{\mathbf{Z}}^t &= (\mathbf{W}-\mathbf{J})\boldsymbol{\Theta}^t\\
&= (\mathbf{W}-\mathbf{J})\mathbf{Z}^{t-1}+(\mathbf{W}-\mathbf{J})(\boldsymbol{\Theta}^t-\mathbf{Z}^{t-1})\\
&= (\mathbf{W}-\mathbf{J})(\mathbf{Z}^{t-1}-\bar{\mathbf{Z}}^{t-1})\\
&\quad+(\mathbf{W}-\mathbf{J})(\boldsymbol{\Theta}^t-\mathbf{Z}^{t-1})\\
&= \cdots\\
&= (\mathbf{W}-\mathbf{J})^t(\mathbf{Z}^0-\bar{\mathbf{Z}}^0)\\
&\quad+\sum_{s=0}^{t-1}(\mathbf{W}-\mathbf{J})^{t-s}(\boldsymbol{\Theta}^{s+1}-\mathbf{Z}^s),
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{Z}^t-\bar{\mathbf{Z}}^t\| &\le \rho^t\|\mathbf{Z}^0-\bar{\mathbf{Z}}^0\| \qquad (16)\\
&\quad+\sum_{s=0}^{t-1}\rho^{t-s}\|\boldsymbol{\Theta}^{s+1}-\mathbf{Z}^s\|.
\end{aligned}$$

Using that f satisfies $RSM'(L,\tau_L)$, we have

$$\begin{aligned}
&\sum_i\left|\langle\nabla f(\mathbf{z}_i^t)-\nabla f(\bar{\mathbf{z}}^t),\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle\right|\\
\le\;&\sum_i\left(L\|\mathbf{z}_i^t-\bar{\mathbf{z}}^t\|^2+\tau_L\|\mathbf{z}_i^t-\bar{\mathbf{z}}^t\|_1^2\right)^{1/2}\\
&\quad\cdot\left(L\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|^2+\tau_L\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|_1^2\right)^{1/2}\\
\le\;&\left(L\sum_i\|\mathbf{z}_i^t-\bar{\mathbf{z}}^t\|^2+\tau_L\sum_i\|\mathbf{z}_i^t-\bar{\mathbf{z}}^t\|_1^2\right)^{1/2}\\
&\quad\cdot\left(L\sum_i\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|^2+\tau_L\sum_i\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|_1^2\right)^{1/2}\\
\le\;&\left\{L\left(\sum_{s=0}^{t-1}\rho^{t-s}\|\boldsymbol{\Theta}^{s+1}-\mathbf{Z}^s\|+\rho^t\|\mathbf{Z}^0-\bar{\mathbf{Z}}^0\|^2\right)^2\right.\\
&\quad\left.+\tau_L(128s\|\mathbf{Z}^t-\widehat{\boldsymbol{\Theta}}\|^2+8m\nu^2)\right\}^{1/2}\\
&\quad\cdot\left\{L\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2\right.\\
&\quad\left.+\tau_L\left(32s\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2+2m\nu^2\right)\right\}^{1/2}\\
\le\;&\left(\sum_{s=0}^{t-1}\rho^{t-s}\sqrt{a_s}+\rho^t\sqrt{L}\|\mathbf{Z}^0-\bar{\mathbf{Z}}^0\|+\sqrt{b_t}\right)\sqrt{c_t}\\
\le\;&\sum_{s=0}^{t-1}\rho^{t-s}\left(\frac{a_s}{\gamma}+\frac{\gamma c_t}{4}\right)+\rho^t\frac{L\|\mathbf{Z}^0-\bar{\mathbf{Z}}^0\|^2}{\gamma}\\
&+\frac{\gamma c_t\rho^t}{4}+\frac{b_t}{\gamma}+\frac{\gamma c_t}{4}\\
=\;&\frac{L}{\gamma}\sum_{s=0}^{t-1}\rho^{t-s}\|\boldsymbol{\Theta}^{s+1}-\mathbf{Z}^s\|^2+\frac{L}{\gamma}\rho^t\|\mathbf{Z}^0-\bar{\mathbf{Z}}^0\|^2\\
&+\frac{8\tau_L}{\gamma}(16s\|\mathbf{Z}^t-\widehat{\boldsymbol{\Theta}}\|^2+m\nu^2)+\frac{\gamma L}{2(1-\rho)}\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2\\
&+\frac{\gamma\tau_L}{1-\rho}(16s\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2+m\nu^2), \qquad (17)
\end{aligned}$$

where we defined $a_s=L\|\boldsymbol{\Theta}^{s+1}-\mathbf{Z}^s\|^2$, $b_t=\tau_L(128s\|\mathbf{Z}^t-\widehat{\boldsymbol{\Theta}}\|^2+8m\nu^2)$ and $c_t=L\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2+\tau_L(32s\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2+2m\nu^2)$ for ease of notation, the first inequality is based on Assumption 2, the second step used the Cauchy-Schwarz inequality, the third inequality used (16), the fourth inequality used the elementary inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, and the fifth inequality is due to the Cauchy-Schwarz inequality again.

Similarly,

$$\begin{aligned}
&\sum_i\left|\langle\nabla f(\bar{\mathbf{z}}^t)-\frac{1}{m}\sum_j\nabla f_j(\mathbf{z}_j^t),\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle\right|\\
\le\;&\frac{1}{m}\sum_i\sum_j\left|\langle\nabla f_j(\bar{\mathbf{z}}^t)-\nabla f_j(\mathbf{z}_j^t),\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\rangle\right|\\
\le\;&\frac{1}{m}\sum_j\left(L\|\mathbf{z}_j^t-\bar{\mathbf{z}}^t\|^2+\tau_\ell\|\mathbf{z}_j^t-\bar{\mathbf{z}}^t\|_1^2\right)^{1/2}\\
&\quad\cdot\sum_i\left(L\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|^2+\tau_\ell\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|_1^2\right)^{1/2}\\
\le\;&\left(L\sum_i\|\mathbf{z}_i^t-\bar{\mathbf{z}}^t\|^2+\tau_\ell\sum_i\|\mathbf{z}_i^t-\bar{\mathbf{z}}^t\|_1^2\right)^{1/2}\\
&\quad\cdot\left(L\sum_i\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|^2+\tau_\ell\sum_i\|\boldsymbol{\theta}_i^{t+1}-\widehat{\boldsymbol{\theta}}\|_1^2\right)^{1/2}\\
\le\;&\frac{L}{\gamma}\sum_{s=0}^{t-1}\rho^{t-s}\|\boldsymbol{\Theta}^{s+1}-\mathbf{Z}^s\|^2+\frac{8\tau_\ell}{\gamma}(16s\|\mathbf{Z}^t-\widehat{\boldsymbol{\Theta}}\|^2+m\nu^2) \qquad (18)\\
&+\frac{\gamma L}{4(1-\rho)}\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2+\frac{\gamma\tau_\ell}{2(1-\rho)}(16s\|\boldsymbol{\Theta}^{t+1}-\widehat{\boldsymbol{\Theta}}\|^2+m\nu^2).
\end{aligned}$$

Furthermore, by the relationship $\mathbf{s}_i^{t+1}=\sum_j w_{ij}(\mathbf{s}_j^t+\nabla f_j(\mathbf{z}_j^{t+1})-\nabla f_j(\mathbf{z}_j^t))$, we have

$$\mathbf{S}^{t+1}-\bar{\mathbf{S}}^{t+1} = (\mathbf{W}-\mathbf{J})(\mathbf{S}^t-\bar{\mathbf{S}}^t+\nabla f(\mathbf{Z}^{t+1})-\nabla f(\mathbf{Z}^t)),$$

and thus

$$\begin{aligned}
&\langle \mathbf{S}^t - \bar{\mathbf{S}}^t, \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\rangle \quad (19)\\
=\ & \Big\langle (\mathbf{W}-\mathbf{J})^t(\mathbf{S}^0 - \bar{\mathbf{S}}^0) \\
&+ \sum_{s=0}^{t-1} (\mathbf{W}-\mathbf{J})^{t-s}(\nabla f(\mathbf{Z}^{s+1}) - \nabla f(\mathbf{Z}^s)), \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}} \Big\rangle .
\end{aligned}$$

Using Lemma 4 and

$$\begin{aligned}
&|\langle \nabla f_j(\mathbf{z}_j^{s+1}) - \nabla f_j(\mathbf{z}_j^s), \boldsymbol{\theta}_i^{t+1} - \widehat{\boldsymbol{\theta}}\rangle| \\
\leq\ & \left(L\|\mathbf{z}_j^{s+1} - \mathbf{z}_j^s\|^2 + \tau_\ell \|\mathbf{z}_j^{s+1} - \mathbf{z}_j^s\|_1^2\right)^{1/2} \\
&\cdot \left(L\|\boldsymbol{\theta}_i^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + \tau_\ell \|\boldsymbol{\theta}_i^{t+1} - \widehat{\boldsymbol{\theta}}\|_1^2\right)^{1/2}, \quad (20)
\end{aligned}$$

we get

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\langle \mathbf{S}^t - \bar{\mathbf{S}}^t, \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\rangle \\
\leq\ & \frac{1}{\sqrt{m}} \left\langle (\mathbf{W}-\mathbf{J})^t(\mathbf{S}^0 - \bar{\mathbf{S}}^0), \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}} \right\rangle \\
&+ \sum_{s=0}^{t-1} \rho^{t-s} \left(L \sum_i \|\mathbf{z}_i^{s+1} - \mathbf{z}_i^s\|^2 + \tau_\ell \sum_i \|\mathbf{z}_i^{s+1} - \mathbf{z}_i^s\|_1^2\right)^{1/2} \\
&\cdot \left(L \sum_i \|\boldsymbol{\theta}_i^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + \tau_\ell \sum_i \|\boldsymbol{\theta}_i^{t+1} - \widehat{\boldsymbol{\theta}}\|_1^2\right)^{1/2} \\
\leq\ & \frac{1}{\sqrt{m}} \left\langle (\mathbf{W}-\mathbf{J})^t(\mathbf{S}^0 - \bar{\mathbf{S}}^0), \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}} \right\rangle \\
&+ \sum_{s=0}^{t-1} \rho^{t-s} \Big\{L\|\mathbf{Z}^{s+1} - \mathbf{Z}^s\|^2 + \tau_\ell \Big(64s\|\mathbf{Z}^{s+1} - \widehat{\mathbf{\Theta}}\|^2 \\
&\qquad +64s\|\mathbf{Z}^s - \widehat{\mathbf{\Theta}}\|^2 + 8m\nu^2\Big)\Big\}^{1/2} \\
&\cdot \Big\{L\|\mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\|^2 + \tau_\ell \Big(32s\|\mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\|^2 \\
&\qquad +2m\nu^2\Big)\Big\}^{1/2} \\
\leq\ & \frac{1}{\sqrt{m}} \left\langle (\mathbf{W}-\mathbf{J})^t(\mathbf{S}^0 - \bar{\mathbf{S}}^0), \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}} \right\rangle \\
&+ \sum_{s=0}^{t-1} \frac{\rho^{t-s}}{2} \Big(L\|\mathbf{Z}^{s+1} - \mathbf{Z}^s\|^2 + \tau_\ell(64s\|\mathbf{Z}^{s+1} - \widehat{\mathbf{\Theta}}\|^2 \\
&\qquad +64s\|\mathbf{Z}^s - \widehat{\mathbf{\Theta}}\|^2 + 8m\nu^2)\Big) \\
&+ \sum_{s=0}^{t-1} \frac{\rho^{t-s}}{2} \Big(L\|\mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\|^2 + \tau_\ell(32s\|\mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\|^2 \\
&\qquad +2m\nu^2)\Big),
\end{aligned}$$

where the 1st inequality used (19) and (20), the 2nd used (6), and the 3rd used the Cauchy-Schwarz inequality.

Then, using $\mathbf{Z}^{s+1} - \mathbf{Z}^s = \mathbf{W}\mathbf{\Theta}^{s+1} - \mathbf{Z}^s = (\mathbf{W}-\mathbf{I})(\mathbf{Z}^s - \bar{\mathbf{Z}}^s) + \mathbf{W}(\mathbf{\Theta}^{s+1} - \mathbf{Z}^s)$,

$$\begin{aligned}
&\|\mathbf{Z}^{s+1} - \mathbf{Z}^s\|^2 \\
\leq\ & 8\|\mathbf{Z}^s - \bar{\mathbf{Z}}^s\|^2 + 2\|\mathbf{\Theta}^{s+1} - \mathbf{Z}^s\|^2 \\
\leq\ & 8\Big(\sum_{r=0}^{s-1} \rho^{s-r}\|\mathbf{\Theta}^{r+1} - \mathbf{Z}^r\| + \rho^s\|\mathbf{Z}^0 - \bar{\mathbf{Z}}^0\|\Big)^2 \\
&+2\|\mathbf{\Theta}^{s+1} - \mathbf{Z}^s\|^2 \\
\leq\ & \frac{16\rho}{1-\rho} \sum_{r=0}^{s-1} \rho^{s-r}\|\mathbf{\Theta}^{r+1} - \mathbf{Z}^r\|^2 \\
&+16\rho^s\|\mathbf{Z}^0 - \bar{\mathbf{Z}}^0\|^2 + 2\|\mathbf{\Theta}^{s+1} - \mathbf{Z}^s\|^2.
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\langle \mathbf{S}^t - \bar{\mathbf{S}}^t, \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\rangle \\
\leq\ & \frac{1}{\sqrt{m}} \left\langle (\mathbf{W}-\mathbf{J})^t(\mathbf{S}^0 - \bar{\mathbf{S}}^0), \mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}} \right\rangle \\
&+ \sum_{s=0}^{t-1} \frac{\rho^{t-s}}{2} \Big(\frac{16\rho L}{1-\rho} \sum_{r=0}^{s-1} \rho^{s-r}\|\mathbf{\Theta}^{r+1} - \mathbf{Z}^r\|^2 \\
&\qquad +16\rho^s\|\mathbf{Z}^0 - \bar{\mathbf{Z}}^0\|^2 + 2L\|\mathbf{\Theta}^{s+1} - \mathbf{Z}^s\|^2 \\
&\qquad +\tau_\ell(64s\|\mathbf{Z}^{s+1} - \widehat{\mathbf{\Theta}}\|^2 + 64s\|\mathbf{Z}^s - \widehat{\mathbf{\Theta}}\|^2 \\
&\qquad +8m\nu^2)\Big) \\
&+ \sum_{s=0}^{t-1} \frac{\rho^{t-s}}{2} \Big(L\|\mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\|^2 + \tau_\ell(32s\|\mathbf{\Theta}^{t+1} - \widehat{\mathbf{\Theta}}\|^2 \\
&\qquad +2m\nu^2)\Big). \quad (21)
\end{aligned}$$

Combining (15),(17),(18) and (21) completes the proof. □

Proof of Proposition 4. Using Proposition 3 in (7), we get that for all $t \geq 0$,

$$\begin{aligned}
&\sum_i (\varphi(\boldsymbol{\theta}_i^{T+t+1}) - \varphi(\widehat{\boldsymbol{\theta}})) \\
\leq\ & (1-\alpha)\sum_i (\varphi(\mathbf{z}_i^{T+t}) - \varphi(\widehat{\boldsymbol{\theta}})) \\
&+ \left(64s\tau_L + \frac{\gamma(L + 16s(\tau_L + \tau_\ell))}{1-\rho} + \frac{\rho\sqrt{m}}{1-\rho}(L + 32s\tau_\ell)\right) \\
&\qquad \cdot\|\mathbf{\Theta}^{T+t+1} - \widehat{\mathbf{\Theta}}\|^2 \\
&+ \Big(\frac{\alpha^2}{2\eta} - \frac{\alpha\mu}{2} + 32\alpha s\tau_\mu + \frac{384s(\tau_L + \tau_\ell)}{\gamma} \\
&\qquad + \frac{2\gamma(L + 16s(\tau_L + \tau_\ell))}{1-\rho} + \frac{\rho\sqrt{m}}{1-\rho}(L + 32s\tau_\ell) + 64s\tau_L\Big) \\
&\qquad \cdot\|\mathbf{Z}^{T+t} - \widehat{\mathbf{\Theta}}\|^2 \\
&+192\tau_\ell\rho\sqrt{m}\sum_{s=0}^{t} \rho^{t-s}\|\mathbf{Z}^{T+s} - \widehat{\mathbf{\Theta}}\|^2 \\
&+(\frac{L}{2} - \frac{1}{2\eta})\|\mathbf{\Theta}^{T+t+1} - \mathbf{Z}^{T+t}\|^2 \\
&+(\frac{6L}{\gamma} + 3L\sqrt{m})\sum_{s=0}^{t-1} \rho^{t-s}\|\mathbf{\Theta}^{T+s+1} - \mathbf{Z}^{T+s}\|^2 \\
&+\frac{24\sqrt{m}\rho L}{1-\rho}\sum_{r=0}^{t-2}(t-r-1)\rho^{t-r}\|\mathbf{\Theta}^{T+r+1} - \mathbf{Z}^{T+r}\|^2 \\
&+(24t\sqrt{m} + \frac{6L}{\gamma})\rho^t\|\mathbf{Z}^T - \bar{\mathbf{Z}}^T\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\left(2\alpha\tau_\mu + 8\tau_L + \frac{24(\tau_L+\tau_\ell)}{\gamma} + \frac{3\gamma(\tau_L+\tau_\ell)}{1-\rho}\right.\\
&\qquad\left.+15\tau_\ell\frac{\sqrt{m}\rho}{1-\rho}\right)m\nu^2\\
&+2\left|\left\langle(\mathbf{W}-\mathbf{J})^t(\mathbf{S}^T-\bar{\mathbf{S}}^T),\mathbf{Z}^{T+t}-\widehat{\boldsymbol{\Theta}}\right\rangle\right|\\
&+\left|\left\langle(\mathbf{W}-\mathbf{J})^t(\mathbf{S}^T-\bar{\mathbf{S}}^T),\boldsymbol{\Theta}^{T+t+1}-\widehat{\boldsymbol{\Theta}}\right\rangle\right|,
\end{aligned}
$$

for any $\gamma > 0$ (γ will be set to be sufficiently small below). We note that in the above, the terms containing the sum $\sum_{s=0}^{t-1}$, $\sum_{r=0}^{t-2}$ are interpreted as 0 when $t < 1$ and $t < 2$, respectively.

Using Lemma 3 and that $\|\mathbf{Z}^T - \bar{\mathbf{Z}}^T\| \le \|\boldsymbol{\Theta}^T - \widehat{\boldsymbol{\Theta}}\|$ (due to the convexity of $\|.\|$), we then get

$$
\begin{aligned}
&\sum_i(\varphi(\boldsymbol{\theta}_i^{T+t+1})-\varphi(\widehat{\boldsymbol{\theta}}))\\
\le\;& (1-\alpha)\sum_i(\varphi(\mathbf{z}_i^{T+t})-\varphi(\widehat{\boldsymbol{\theta}}))\\
&+\left(64s\tau_L+\frac{\gamma(L+16s(\tau_L+\tau_\ell))}{1-\rho}+\frac{\rho\sqrt{m}}{1-\rho}(L+32s\tau_\ell)\right)\\
&\qquad\cdot\|\boldsymbol{\Theta}^{T+t+1}-\widehat{\boldsymbol{\Theta}}\|^2\\
&+\left(\frac{\alpha^2}{2\eta}-\frac{\alpha\mu}{2}+\alpha 32s\tau_\mu+\frac{384s(\tau_L+\tau_\ell)}{\gamma}\right.\\
&\qquad\left.+\frac{2\gamma(L+16s(\tau_L+\tau_\ell))}{1-\rho}+\frac{\rho\sqrt{m}}{1-\rho}(L+32s\tau_\ell)+64s\tau_L\right)\\
&\qquad\cdot\|\mathbf{Z}^{T+t}-\widehat{\boldsymbol{\Theta}}\|^2\\
&+192\tau_\ell\rho\sqrt{m}\sum_{s=0}^{t}\rho^{t-s}\|\mathbf{Z}^{T+s}-\widehat{\boldsymbol{\Theta}}\|^2\\
&+(\frac{L}{2}-\frac{1}{2\eta})\|\boldsymbol{\Theta}^{T+t+1}-\mathbf{Z}^{T+t}\|^2\\
&+(\frac{6L}{\gamma}+3L\sqrt{m})\sum_{s=0}^{t-1}\rho^{t-s}\|\boldsymbol{\Theta}^{T+s+1}-\mathbf{Z}^{T+s}\|^2\\
&+\frac{24\sqrt{m}\rho L}{1-\rho}\sum_{r=0}^{t-2}(t-r-1)\rho^{t-r}\|\boldsymbol{\Theta}^{T+r+1}-\mathbf{Z}^{T+r}\|^2\\
&+\left(24t\sqrt{m}+\frac{6L}{\gamma}\right)\rho^t\|\boldsymbol{\Theta}^T-\widehat{\boldsymbol{\Theta}}\|^2+\rho^t\omega\\
&+\left(2\alpha\tau_\mu+8\tau_L+\frac{24(\tau_L+\tau_\ell)}{\gamma}+\frac{3\gamma(\tau_L+\tau_\ell)}{1-\rho}\right.\\
&\qquad\left.+15\tau_\ell\frac{\sqrt{m}\rho}{1-\rho}\right)m\nu^2.
\end{aligned}
$$

Using Lemma 2 and $\sum_i \varphi(\mathbf{z}_i^t) \le \sum_i \varphi(\boldsymbol{\theta}_i^t)$ (due to convexity of φ and that $\mathbf{z}_i^t = \sum_j w_{ij}\boldsymbol{\theta}_j^t$), with the quantities defined before the statement of the proposition, we get that

$$
\begin{aligned}
&(1-\frac{A_1}{\mu/2-32s\tau_\mu})\sum_i(\varphi(\boldsymbol{\theta}_i^{T+t+1})-\varphi(\widehat{\boldsymbol{\theta}}))\\
\le\;&(1-\alpha)\sum_i(\varphi(\boldsymbol{\theta}_i^{T+t})-\varphi(\widehat{\boldsymbol{\theta}}))+A_2\|\mathbf{Z}^{T+t}-\widehat{\boldsymbol{\Theta}}\|^2\\
&+192\tau_\ell\rho\sqrt{m}\sum_{s=0}^{t}\rho^{t-s}\|\mathbf{Z}^{T+s}-\widehat{\boldsymbol{\Theta}}\|^2+(\frac{L}{2}-\frac{1}{2\eta})\|\boldsymbol{\Theta}^{t+1}-\mathbf{Z}^t\|^t\\
&+(\frac{6L}{\gamma}+3L\sqrt{m})\sum_{s=0}^{t-1}\rho^{t-s}\|\boldsymbol{\Theta}^{T+s+1}-\mathbf{Z}^{T+s}\|^2\\
&+\frac{24\sqrt{m}\rho L}{1-\rho}\sum_{r=0}^{t-2}(t-r-1)\rho^{t-r}\|\boldsymbol{\Theta}^{T+r+1}-\mathbf{Z}^{T+r}\|^2\\
&+\left(24t\sqrt{m}+\frac{6L}{\gamma}\right)\rho^t\|\boldsymbol{\Theta}^T-\widehat{\boldsymbol{\Theta}}\|^2+\rho^t\omega+A_5m\nu^2
\end{aligned}
\tag{22, 23}
$$

Dividing both sides by σ^{t+1}, and summing over $t = 0,\ldots,T'-1$, by defining $a^{T'}(\sigma) = \sum_{t=0}^{T'}\sum_i(\varphi(\mathbf{z}_i^{T+t}) - \varphi(\widehat{\boldsymbol{\theta}}))/\sigma^t$, $b^{T'}(\sigma) = \sum_{t=0}^{T'}\|\mathbf{Z}^{T+t} - \widehat{\boldsymbol{\Theta}}\|^2/\sigma^t$, $c^{T'}(\sigma) = \sum_{t=1}^{T'+1}\|\boldsymbol{\Theta}^{T+t} - \mathbf{Z}^{T+t-1}\|^2/\sigma^t$, recalling A_6, we get

$$
\begin{aligned}
&(1-\frac{A_1}{\mu/2-32s\tau_\mu})\left(a^{T'}(\sigma)-\sum_i(\varphi(\boldsymbol{\theta}_i^T)-\varphi(\widehat{\boldsymbol{\theta}}))\right)\\
\le\;&\frac{1-\alpha}{\sigma}a^{T'-1}(\sigma)+A_3b^{T'-1}(\sigma)+A_6c^{T'-1}(\sigma)\\
&+\left(\frac{24\sqrt{m}\rho}{(\sigma-\rho)^2}+\frac{6L}{\gamma(\sigma-\rho)}\right)\|\boldsymbol{\Theta}^T-\widehat{\boldsymbol{\Theta}}\|^2+\frac{\omega}{\sigma-\rho}\\
&+\frac{A_5}{\sigma^{T'}(1-\sigma)}m\nu^2,
\end{aligned}
$$

if $\sum_i(\varphi(\boldsymbol{\theta}_i^t) - \varphi(\boldsymbol{\theta}_0)) \le \xi$ for all $t \ge T$, using that $\sum_{s=0}^{t}(\frac{\rho}{\sigma})^{t-s} \le \frac{\rho}{\sigma-\rho}$ and $\sum_{r=0}^{t-2}(t-r-1)(\frac{\rho}{\sigma})^{t-r} \le \frac{\rho^2}{(\sigma-\rho)^2}$. Since we assume $A_3 \le 0$, $A_4 > 0$, $A_6 \le 0$, we get

$$
\begin{aligned}
&\sum_i(\varphi(\boldsymbol{\theta}_i^{T+t})-\varphi(\widehat{\boldsymbol{\theta}}))\\
\le\;&\frac{\sigma^t}{A_4}(1-\frac{A_1}{\mu/2-32s\tau_\mu})\sum_i(\varphi(\boldsymbol{\theta}_i^T)-\varphi(\widehat{\boldsymbol{\theta}}))\\
&+\frac{\sigma^t}{A_4}\left(\frac{24\sqrt{m}\rho}{(\sigma-\rho)^2}+\frac{6L}{\gamma(\sigma-\rho)}\right)\|\boldsymbol{\Theta}^T-\widehat{\boldsymbol{\Theta}}\|^2\\
&+\frac{\sigma^t}{A_4(\sigma-\rho)}\omega+\frac{A_5}{A_4(1-\sigma)}m\nu^2\\
\le\;&\sigma^t\eta_1+\eta_2 m(\varepsilon_{stat}+\varepsilon)^2,
\end{aligned}
$$

where $\varepsilon = 2\min\{\xi/\lambda, R\}$. □

The following Lemmas are used in various parts of the main proofs above.

Lemma 2: When $\sum_i(\varphi(\boldsymbol{\theta}_i^t) - \varphi(\widehat{\boldsymbol{\theta}})) \le \xi$,

$$(\frac{\mu}{2} - 32s\tau_\mu)\|\boldsymbol{\Theta}^t - \widehat{\boldsymbol{\Theta}}\|^2 \le \sum_i(\varphi(\boldsymbol{\theta}_i^t) - \varphi(\widehat{\boldsymbol{\theta}})) + 2\tau_\mu m\nu^2.$$

We note that ν implicitly depends on ξ.

Proof of Lemma 2. This directly follows from Lemma 1 and the strong convexity of f. □

Lemma 3: For any $t_1, t_2 \ge 0$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_m)^\top$ with $\|\boldsymbol{\theta}_i\|_1 \le R, i = 1,\ldots,m$ we have that

$$\left|\left\langle(\mathbf{W}-\mathbf{J})^{t_1}(\mathbf{S}^{t_2}-\bar{\mathbf{S}}^{t_2}),\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\right\rangle\right| \le \rho^{t_1}\omega,$$

where $\omega = m^{3/2}(8(L+\tau_\ell)R^2 + 4MR + \frac{8(L+\tau_\ell)R^2}{1-\rho})$.

Proof of Lemma 3. Using (19), we have

$$\begin{aligned}&\left\langle(\mathbf{W}-\mathbf{J})^{t_1}(\mathbf{S}^{t_2}-\bar{\mathbf{S}}^{t_2}),\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\right\rangle\\ =&\Bigg\langle(\mathbf{W}-\mathbf{J})^{t_1+t_2}(\mathbf{S}^0-\bar{\mathbf{S}}^0)+\sum_{s=0}^{t_2-1}(\mathbf{W}-\mathbf{J})^{t_1+t_2-s}\\ &\cdot(\nabla f(\mathbf{Z}^{s+1})-\nabla f(\mathbf{Z}^s)),\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\Bigg\rangle.\end{aligned}$$

Note that $\mathbf{S}^0=(\nabla f_1(\boldsymbol{\theta}_1^0),\ldots,\nabla f_m(\boldsymbol{\theta}_m^0))^\top$, and we have

$$\begin{aligned}&\langle\nabla f_i(\boldsymbol{\theta}_i^0)-\frac{1}{m}\sum_k\nabla f_k(\boldsymbol{\theta}_k^0),\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle\\ =&\langle\nabla f_i(\boldsymbol{\theta}_0),\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle+\langle\nabla f_i(\boldsymbol{\theta}_i^0)-\nabla f_i(\boldsymbol{\theta}_0),\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle\\ &-\langle\nabla f(\boldsymbol{\theta}_0),\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle-\frac{1}{m}\sum_k\langle\nabla f_k(\boldsymbol{\theta}_k^0)-\nabla f_k(\boldsymbol{\theta}_0),\\ &\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle.\end{aligned}$$

The first term is bounded by $M\|\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\|_1\leq 2RM$, the second term is bounded by $\sqrt{L\|\boldsymbol{\theta}_i^0-\boldsymbol{\theta}_0\|^2+\tau_\ell\|\boldsymbol{\theta}_i^0-\boldsymbol{\theta}_0\|_1^2}\sqrt{L\|\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\|^2+\tau_\ell\|\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\|_1^2}\leq 4(L+\tau_\ell)R^2$, and the third and fourth terms can be bounded similarly, which yields

$$\begin{aligned}&\left|\langle\nabla f_i(\boldsymbol{\theta}_i^0)-\frac{1}{m}\sum_k\nabla f_k(\boldsymbol{\theta}_k^0),\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle\right|\\ \leq&\ 8(L+\tau_\ell)R^2+4MR.\end{aligned}$$

Similarly,

$$\left|\langle\nabla f_i(\mathbf{z}_i^{s+1})-\nabla f_i(\mathbf{z}_i^s),\boldsymbol{\theta}_j-\widehat{\boldsymbol{\theta}}\rangle\right|\leq 4(L+\tau_\ell)R^2.$$

Thus, by Lemma 4,

$$\begin{aligned}&\Big\langle(\mathbf{W}-\mathbf{J})^{t_1+t_2}(\mathbf{S}^0-\bar{\mathbf{S}}^0)\\ &+\sum_{s=0}^{t_2-1}(\mathbf{W}-\mathbf{J})^{t_1+t_2-s}(\nabla f(\mathbf{Z}^{s+1})-\nabla f(\mathbf{Z}^s)),\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}\Big\rangle\\ \leq&\ \sqrt{m}\rho^{t_1+t_2}m(8(L+\tau_\ell)R^2+4MR)\\ &+\sum_{s=0}^{t_2-1}\sqrt{m}\rho^{t_1+t_2-s}\cdot 4m(L+\tau_\ell)R^2\\ \leq&\ m^{3/2}\rho^{t_1}(8(L+\tau_\ell)R^2+4MR+\frac{8(L+\tau_\ell)R^2}{1-\rho}).\end{aligned}$$

□

Lemma 4: For any matrices $\mathbf{A},\mathbf{B}\in\mathbb{R}^{m\times d}$ with rows $\mathbf{a}_i,\mathbf{b}_i$, $i=1,\ldots,m$, if $|\langle\mathbf{a}_i,\mathbf{b}_j\rangle|\leq c_id_j$, $\forall i,j\in\{1,\ldots,m\}$, for some non-negative vectors $\mathbf{c}=(c_1,\ldots,c_m)^\top$, $\mathbf{d}=(d_1,\ldots,d_m)^\top$, then

$$\langle(\mathbf{W}-\mathbf{J})^s\mathbf{A},\mathbf{B}\rangle\leq\sqrt{m}\rho^s\|\mathbf{c}\|\|\mathbf{d}\|.$$

Proof of Lemma 4. For a matrix $\mathbf{A}$ with entries a_{ij}, $\|\mathbf{A}\|_{op}$ denotes its operator norm and $\|\mathbf{A}\|_\infty=\max_i\sum_j|a_{ij}|$ (maximum row sum) is the matrix norm induced by vector ℓ_∞ norm. Let $\mathbf{U}=(\mathbf{W}-\mathbf{J})^s=\{u_{ij}\}_{i,j=1}^m$ and $|\mathbf{U}|$ indicates the matrix $\{|u_{ij}|\}_{i,j=1}^m$. First we show $\||\mathbf{U}|\|_{op}\leq\sqrt{m}\|\mathbf{U}\|_{op}$. In fact,

$$\||\mathbf{U}|\|_{op}\leq\||\mathbf{U}|\|_\infty=\|\mathbf{U}\|_\infty\leq\sqrt{m}\|\mathbf{U}\|_{op},$$

where the first inequality is due to that operator norm is the smallest among all matrix norms (Theorem 5.6.9 of [25]) and the second inequality is due to $\|\mathbf{U}\|_\infty^2=\max_{\|\mathbf{c}\|_\infty\leq 1}\max_i(\sum_j u_{ij}c_j)^2\leq\max_{\|\mathbf{c}\|\leq\sqrt{m}}\sum_i(\sum_j u_{ij}c_j)^2=m\|\mathbf{U}\|_{op}^2$. Then we have

$$\begin{aligned}|\langle(\mathbf{W}-\mathbf{J})^s\mathbf{A},\mathbf{B}\rangle|&=\left|\sum_{i,j}u_{ij}\langle\mathbf{a}_j,\mathbf{b}_i\rangle\right|\leq\sum_{i,j}|u_{ij}|c_jd_i\\ &=\mathbf{d}^\top|\mathbf{U}|\mathbf{c}\leq\||\mathbf{U}|\|_{op}\|\mathbf{c}\|\|\mathbf{d}\|\\ &\leq\sqrt{m}\|\mathbf{U}\|_{op}\|\mathbf{c}\|\|\mathbf{d}\|,\end{aligned}$$

which established the lemma. □

Acknowledgment

The research of Heng Lian is partially supported by NSFC 12371297 at CityU Shenzhen Research Institute, NSF of Jiangxi Province under Grant 20223BCJ25017, and by Hong Kong RGC general research fund 11300519, 11300721 and 11311822, and by CityU internal grant 7005514 and 9680239.

References

[1] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.
[2] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.
[3] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *Journal of Machine Learning Research*, vol. 14, pp. 3321–3363, 2013.
[4] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.
[5] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 4051–4064, 2015.
[6] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, pp. 1835–1854, 2016.
[7] A. Mokhtari and A. Ribeiro, "DSA: Decentralized double stochastic averaging gradient algorithm," *Journal of Machine Learning Research*, vol. 17, pp. 1–35, 2016.
[8] J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor, "Communication-efficient sparse regression," *Journal of Machine Learning Research*, vol. 18, pp. 1–30, 2017.
[9] A. Nedic and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.
[10] X. Lian, C. Zhang, H. Zhang, C. J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2017.
[11] Y. Nesterov and I. U. E. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Netherlands, 2004.
[12] Y. Nesterov, *Lectures on convex optimization*. New York: Springer, 2018.
[13] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.
[14] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.
[15] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2018.

[16] P. McCullagh and J. A. Nelder, *Generalized linear models*, 2nd ed. London New York: Chapman and Hall, 1989.
[17] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B-Methodological*, vol. 58, no. 1, pp. 267–288, 1996.
[18] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, pp. 101–148, 2010.
[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.
[20] C. Geoffrey, L. Guillaume, and L. Matthieu, "Robust high dimensional learning for Lipschitz and convex losses," *Journal of Machine Learning Research*, vol. 21, pp. 1–47, 2020.
[21] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *Annals of Statistics*, vol. 40, pp. 2452–2482, 2012.
[22] Z. Wang, H. Liu, and T. Zhang, "Optimal computational and statistical rates of convergence for sparse nonconvex learning problems," *Annals of Statistics*, vol. 42, pp. 2164–2201, 2014.
[23] Y. Ji, G. Scutari, Y. Sun, and H. Honnappa, "Distributed sparse regression via penalization," *arXiv preprint arXiv:2111.06530*, 2021.
[24] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statistical Science*, vol. 27, pp. 538–557, 2012.
[25] R. Horn and C. Johnson, *Matrix Analysis*, 2nd ed. New York, NY: Cambridge University Press, 2012.
[26] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 2008.
[27] A. Belloni and V. Chernozhukov, "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, vol. 19, no. 2, pp. 521–547, 2013.
[28] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic control*, vol. 57, no. 3, pp. 592–606, 2011.
[29] H. Robbins and S. Monro, "A Stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

Heng Lian is currently a Professor in the Department of Mathematics, City University of Hong Kong. He previously worked as an Assistant Professor at Nanyang Technological University, Singapore, and later as a Senior Lecturer at the University of New South Wales, Australia. His research interests include mathematical statistics, machine learning, and pattern recognition.

Junzhuo Gao