Supplemental Materials

Proof to Lemma 1

**Proof.** Denote $\text{supp}(P_{\Omega(t,s)}(w))$ and $\text{supp}(P_{\Omega(t,s)}(w))$ by $A$ and $B$ respectively for short. We first prove $B \subseteq A$.

Suppose $B \not\subseteq A$, then we can find an element $b \in B$ but $b \not\in A$. Without the loss of generality, we assume that $b$ is in a certain group $g$. Since $A \cap g$ contains the indices of the $t_g$ largest (magnitude) elements of group $g$, there exists at least one element $a \in A \cap g$ and $a \not\in B \cap g$ (otherwise $|B \cap g| \geq t_g + 1$). Replacing $b$ by $a$ in $B$, the constraints are still satisfied, but we can get a better solution since $|w_a| > |w_b|$. This contradicts $B = \text{supp}(P_{\Omega(t,s)}(w))$.

Because we already know $B \subseteq A$, we can construct $B$ by selecting the $A$’s elements corresponding to the largest $s$ (magnitude) elements. Therefore, $\text{supp}(P_{\Omega(s,t)}(w)) = \text{supp}(P_{\Omega(s,\infty)}(P_{\Omega(t,s)}(w)))$, which proves Lemma 1. \qed 

**Lemma 5.** ∀$\text{supp}(w - \bar{w}) \subseteq S, S \in \Omega(s, t)$, if $2\eta - \eta^2 \rho_+(s, t)) > 0$, then

$$
\|w - \bar{w} - \eta[\nabla f(w) - \nabla f(\bar{w})]_S \| \leq (1 - 2\eta \rho_-(s, t) + \eta^2 \rho_+(s, t))\|w - \bar{w}\|^2.
$$

\[ (6) \]

**Proof.**

$$
\|w - \bar{w} - \eta[\nabla f(w) - \nabla f(\bar{w})]_S \|^2 \\
= \|w - \bar{w}\|^2 + \eta^2 \|\nabla f(w) - \nabla f(\bar{w})\|^2 - 2\eta(w - \bar{w}, [\nabla f(w) - \nabla f(\bar{w})]_S) \\
\leq \|w - \bar{w}\|^2 + \eta^2 \rho_+(s, t) - 2\eta(w - \bar{w}, [\nabla f(w) - \nabla f(\bar{w})]_S) \\
\leq \|w - \bar{w}\|^2 - (2\eta - \eta^2 \rho_+(s, t))\rho_-(s, t)\|w - \bar{w}\|^2 \\
= (1 - 2\eta \rho_-(s, t) + \eta^2 \rho_+(s, t))\|w - \bar{w}\|^2.
$$

It completes the proof. \qed 

Proof to Theorem 2

**Proof.** Let us prove the first claim.

$$
\|w^{k+1} - (w^k - \eta \nabla f(w^k))\|^2 \\
= \|w^{k+1} - w^k + \|w^k - (w^k - \eta \nabla f(w^k))\|^2 + 2\|w^{k+1} - w^k - \eta \nabla f(w^k))\|^2 \\
\leq 2\|w^{k+1} - w^k, [w^k - \eta \nabla f(w^k)]_{\Omega_{k+1}}\) \\
\leq 2\|w^{k+1} - w^k\|\|w^k - \eta \nabla f(w^k) - \bar{w}\|_{\Omega_{k+1}}\|.
$$

Define $\Omega = \text{supp}(\bar{w}), \Omega_{k+1} = \text{supp}(w^{k+1})$, and $\Omega_{k+1} = \Omega \cup \Omega_{k+1}$. From $\|w^{k+1} - (w^k - \eta \nabla f(w^k))\|^2 \leq \|w - (w^k - \eta \nabla f(w^k))\|^2$, we have

$$
\|w^{k+1} - \bar{w}\|^2 \leq 2\|w^{k+1} - \bar{w}, [w^k - \eta \nabla f(w^k)]_{\Omega_{k+1}}\) \\
\leq 2\|w^{k+1} - \bar{w}\|\|w^k - \eta \nabla f(w^k) - \bar{w}\|_{\Omega_{k+1}}\|
$$

It follows

$$
\|w^{k+1} - w\| \leq 2\|w^k - \eta \nabla f(w^k) - \bar{w}\|_{\Omega_{k+1}}\| \\
= 2\|w^k - \eta \nabla f(w^k) - \bar{w} + \eta \nabla f(w) - \eta \nabla f(w)\|_{\Omega_{k+1}}\| \\
\leq 2\|w^k - \eta \nabla f(w^k) - \bar{w} + \eta \nabla f(w)\|_{\Omega_{k+1}}\| + 2\eta\|\nabla f(w)\|_{\Omega_{k+1}}\| \\
\leq 2\|w^k - \eta \nabla f(w^k) - \bar{w} + \eta \nabla f(w)\|_{\Omega_{k+1}}\| + 2\eta\|\nabla f(w)\|_{\Omega_{k+1}}\| \\
= 2\|w^k - \bar{w} - \eta \nabla f(w^k) - \nabla f(w)\|_{\Omega_{k+1}}\| + 2\eta\|\nabla f(w)\|_{\Omega_{k+1}}\|.
$$
From the inequality of Lemma 5, we have
\[
\|w^{k+1} - w\| \leq \alpha \|w^k - w\| + 2\eta \|\nabla f(w)\|_{\Omega_{k+1}}
\leq \alpha \|w^k - w\| + 2\eta \max_j \|\nabla f(w)\|_{\Omega_{j+1}}
\leq \alpha \|w^k - w\| + 2\eta \Delta.
\] (7)

Since \(\Delta\) is constant, using the recursive relation of (7), we have
\[
\|w^k - w\| \leq \alpha^k \|w^0 - w\| + 2\eta \Delta \sum_{i=0}^{k-1} \alpha^i
= \alpha^k \|w^0 - w\| + 2\eta \Delta \frac{1 - \alpha^k}{1 - \alpha}
\leq \alpha^k \|w^0 - w\| + 2\eta \Delta \frac{1}{1 - \alpha}.
\] (8)

Then we move to (2), when \(k \geq \lceil \log \frac{2\Delta}{(1 - \alpha)p_+ (3s, 3t)} \|w^0 - w\| / \log \alpha \rceil\), from the conclusion of (1), we have
\[
\|w^k - w\| \leq \frac{4\Delta}{(1 - \alpha)p_+ (3s, 3t)}.
\] (9)

For any \(j \in \tilde{\Omega}\),
\[
\|w^k - w\|_{\infty} \geq \|w^k - \bar{w}\|_{j}
\geq -\|w^k\|_{j} + \|\bar{w}\|_{j}.
\]

So
\[
\|w^k\|_{j} \geq \|\bar{w}\|_{j} - \|w^k - w\|_{\infty}
\geq \|\bar{w}\|_{j} - \frac{4\Delta}{(1 - \alpha)p_+ (3s, 3t)}.
\]

Therefore, \(w^k\) is non-zero if \(\|\bar{w}\|_{j} > \frac{4\Delta}{(1 - \alpha)p_+ (3s, 3t)}\), and (2) is proved. \(\square\)

**Lemma 6.** The value of \(\Delta\) is bounded by
\[
\Delta \leq \min \left( O\left( \sqrt{n} \log p + \log \frac{1}{\eta'} \right), \ O\left( \sqrt{\max_{g \in G} \log |g| \sum_{g \in G} t_g + \log \frac{1}{\eta'}} \right) \right),
\] (10)

with high probability \(1 - \eta'\).

**Proof.** We introduce the following notation for matrix and it is different from the vector notation. For a matrix \(X\in \mathbb{R}^{n \times p}\), \(X_h\) will be a \(\mathbb{R}^{n \times |h|}\) matrix that only keep the columns corresponding to the index set \(h\). Here we restrict \(h\) by \(w_h \in \Omega(s, t)\) for any \(w \in \mathbb{R}^p\). We denote \(\Sigma_h = X_h^T X_h\). For the theorem, we can first show that \(\|X_h^T \epsilon\|_{\infty} \leq \sqrt{n} \left( \sqrt{|h|} + \sqrt{2p_+ (2s, 2t) \log \frac{1}{\eta}} \right)\) with probability \(1 - \eta\). To this end, we have to point out that our columns of \(X\) are normalized to \(\sqrt{n}\) and hence \(X_h^T \epsilon\) will be a \(\mathbb{R}^n\)-variate Gaussian random variable with \(n\) on the diagonal of covariance matrix. We further use \(\lambda_1\) as the eigenvalues of \(\Sigma_h\) with decreasing order, i.e., \(\lambda_1\) being the largest, or equivalently, \(\lambda_1 = \|\Sigma_h\|_{\text{spec}}\).
Also, using the trick that $tr(\Sigma^T_h) = \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{|h|}^2$ and Proposition 1.1 from [16], we have

$$e^{-t} \geq \Pr \left( \|X_h^T \epsilon\|^2 > \sum_{i=1}^{|h|} \lambda_i + 2 \sqrt{\sum_{i=1}^{|h|} \lambda_i^2 t + 2 \lambda_1 t} \right)$$

$$\geq \Pr \left( \|X_h^T \epsilon\|^2 > \sum_{i=1}^{|h|} \lambda_i + 2 \sqrt{2 \sum_{i=1}^{|h|} \lambda_i \lambda_1 t + 2 \lambda_1 t} \right)$$

$$\geq \Pr \left( \|X_h^T \epsilon\| > \sqrt{\sum_{i=1}^{|h|} \lambda_i + 2 \lambda_1 t} \right).$$

Substitute $t$ with $\log(\frac{1}{\eta})$ and the facts that $\sum_{i=1}^{|h|} \lambda_i = |h|n$ and $\lambda_1 = \|\Sigma\|_{\text{spec}} \leq n \rho_+(2s, 2t)$, we have

$$\|X_h^T \epsilon\| \leq \sqrt{n} \left( \sqrt{|h|} + \sqrt{2 \rho_+(2s, 2t) \log(1/\eta)} \right)$$

with probability $1 - \eta$.

For the least square loss, we have $\nabla f(\bar{w}) = \frac{1}{n} X^T (X \bar{w} - y) = \frac{1}{n} X^T \epsilon$. To estimate the upper bound of $\|P_{\Omega(2s, 2t)}(\nabla f(\bar{w}))\|$, we use the following fact

$$\|P_{\Omega(2s, 2t)}(\nabla f(\bar{w}))\| = \|P_{\Omega(2s, 2t)}(X^T \epsilon)\| \leq \min \left( \|P_{\Omega(2s, \infty)}(X^T \epsilon)\|, \|P_{\Omega(\infty, 2t)}(X^T \epsilon)\| \right).$$

We consider the upper bounds of $\|P_{\Omega(2s, \infty)}(X^T \epsilon)\|$ and $\|P_{\Omega(\infty, 2t)}(X^T \epsilon)\|$ respectively:

$$\Pr \left( \|P_{\Omega(2s, \infty)}(X^T \epsilon)\| \geq n^{-1/2} \left( \sqrt{2s} + \sqrt{2 \rho_+(2s, 2t) \log(1/\eta)} \right) \right)$$

$$= \Pr \left( \max_{|h|=2s} \|X_h^T \epsilon\| \geq n^{-1/2} \left( \sqrt{2s} + \sqrt{2 \rho_+(2s, 2t) \log(1/\eta)} \right) \right)$$

$$\leq \sum_{|h|=2s} \Pr \left( \|X_h^T \epsilon\| \geq n^{-1/2} \left( \sqrt{2s} + \sqrt{2 \rho_+(2s, 2t) \log(1/\eta)} \right) \right)$$

$$\leq \left( \frac{p}{2s} \right) \eta.$$ 

By taking $\eta' = \eta \left( \frac{p}{2s} \right)$, we obtain

$$\eta' \geq \Pr \left( \|P_{\Omega(2s, \infty)}(X^T \epsilon)\| \geq n^{-1/2} \left( \sqrt{2s} + \sqrt{2 \rho_+(2s, 2t) \log \left( \left( \frac{p}{2s} \right) / \eta' \right)} \right) \right)$$

$$\geq \Pr \left( \|P_{\Omega(2s, \infty)}(X^T \epsilon)\| \geq O \left( \frac{s \log(p) + \log 1/\eta'}{n} \right) \right),$$

where the last inequality uses the fact that $\rho_+(2s, 2t)$ is bounded by a constant with high probability.
Next we consider the upper bound of $\|P_{\Omega(2t)}(X^T \epsilon)\|$. Similarly, we have

$$\Pr \left( \|P_{\Omega(2t)}(X^T \epsilon)\| \geq n^{-1/2} \left( \frac{2 \sum_{g \in \mathcal{G}} t_g + \sqrt{2 \rho_+ (2s, 2t) \log(1/\eta)}}{2} \right) \right)$$

Thus, by taking $\eta' = \eta \prod_{g \in \mathcal{G}} \left( \frac{|g|}{2t_g} \right)$, we have

$$\eta' \geq \Pr \left( \|P_{\Omega(2t)}(X^T \epsilon)\| \geq n^{-1/2} \left( \frac{2 \sum_{g \in \mathcal{G}} t_g + \sqrt{2 \rho_+ (2s, 2t) \log \left( \prod_{g \in \mathcal{G}} \left( \frac{|g|}{2t_g} \right) / \eta' \right)}}{2} \right) \right)$$

$$\geq \Pr \left( \|P_{\Omega(2t)}(X^T \epsilon)\| \geq n^{-1/2} \left( \frac{2 \sum_{g \in \mathcal{G}} t_g + \sqrt{4 \rho_+ (2s, 2t) \sum_{g \in \mathcal{G}} t_g \log|g| + 2 \rho_+ (1) \log(1/\eta')}}{2} \right) \right)$$

$$\geq \Pr \left( \|P_{\Omega(2t)}(X^T \epsilon)\| \geq n^{-1/2} \left( \frac{2 \sum_{g \in \mathcal{G}} t_g + \sqrt{4 \rho_+ (2s, 2t) \max \sum_{g \in \mathcal{G}} t_g + 2 \rho_+ (1) \log(1/\eta')}}{2} \right) \right)$$

$$\geq \Pr \left( \|P_{\Omega(2t)}(X^T \epsilon)\| \geq O \left( \frac{\max_{g \in \mathcal{G}} \log|g| \sum_{g \in \mathcal{G}} t_g + \log 1/\eta'}{n} \right) \right).$$

Summarizing two upper bounds, we have with high probability $(1 - 2\eta')$

$$\|P_{\Omega(2t)}(\nabla f(\mathbf{w}))\| \leq \min \left( O \left( \frac{s \log p + \log 1/\eta'}{n} \right), O \left( \frac{\max_{g \in \mathcal{G}} \log|g| \sum_{g \in \mathcal{G}} t_g + \log 1/\eta'}{n} \right) \right).$$

Lemma 7. For the least square loss, assume that matrix $X$ to be sub-Gaussian with zero mean and has independent rows or columns. If the number of samples $n$ is more than

$$O \left( \min \left\{ s \log p, \log(\max_{g \in \mathcal{G}} |g|) \sum_{g \in \mathcal{G}} t_g \right\} \right),$$

then with high probability, we have with high probability

$$\rho_+ (3s, 3t) \leq \frac{3}{2} \quad (11)$$

$$\rho_- (3s, 3t) \geq \frac{1}{2}. \quad (12)$$

Thus, $\alpha$ defined in (3) is less than 1 by appropriately choosing $\eta$ (for example, $\eta = 1/\rho_+ (3s, 3t)$).
Proof. For the linear regression loss, we have
\[
\rho_+^{1/2}(3s, 3t) \leq \frac{1}{\sqrt{n}} \max_{w \in \Omega(3s, 3t)} \|Xw\| = \max_{|h| \leq 3s, |h\cap q| \leq t_q} \|X_h\|
\]
\[
\rho_-^{1/2}(3s, 3t) \geq \frac{1}{\sqrt{n}} \min_{w \in \Omega(3s, 3t)} \|Xw\| = \min_{1 \leq |h| \leq 3s, |h\cap q| \leq t_q} \|X_h\|
\]
From the random matrix theory [35, Theorem 5.39], we have
\[
\Pr \left( \|X_h\| \geq \sqrt{n} + O(\sqrt{3s}) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right) \leq O(\eta)
\]
Then we have
\[
\Pr \left( \sqrt{n} \rho_+^{1/2}(3s, 3t) \geq \sqrt{n} + O(\sqrt{s}) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right)
\]
\[
\leq \Pr \left( \max_{|h| \leq 3s, |h\cap q| \leq t_q} \|X_h\| \geq \sqrt{n} + O(\sqrt{s}) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right)
\]
\[
\leq \Pr \left( \|X_h\| \geq \sqrt{n} + O(\sqrt{s}) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right) = \left( \frac{p}{3s} \right) \Pr \left( \|X_h\| \geq \sqrt{n} + O(\sqrt{s}) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right) \leq O\left( \left( \frac{p}{3s} \right) \eta \right)
\]
which implies (by taking \( \eta' = \left( \frac{p}{3s} \right) \eta \)):
\[
\Pr \left( \sqrt{n} \rho_+^{1/2}(3s, 3t) \geq \sqrt{n} + O\left(\sqrt{s \log p}\right) + O\left(\sqrt{\log \frac{1}{\eta'}}\right) \right) \leq \eta'
\]
Taking \( n = O(s \log p) \), we have \( \rho_+^{1/2}(3s, 3t) \leq \sqrt{\frac{3}{2}} \) with high probability. Next, we consider it from a different perspective.
\[
\Pr \left( \sqrt{n} \rho_+^{1/2}(3s, 3t) \geq \sqrt{n} + O\left(\sum_{g \in \mathcal{V}} |t_g|\right) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right)
\]
\[
\leq \Pr \left( \sqrt{n} \rho_+^{1/2}(+\infty, 3t) \geq \sqrt{n} + O\left(\sum_{g \in \mathcal{V}} |t_g|\right) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right)
\]
\[
= \Pr \left( \max_{|h\cap q| \leq t_q} \|X_h\| \geq \sqrt{n} + O\left(\sum_{g \in \mathcal{V}} |t_g|\right) + O\left(\sqrt{\log \frac{1}{\eta}}\right) \right)
\]
\[
\leq \prod_{g \in \mathcal{V}} \Pr \left( |g| \leq \eta \log \max_{g \in \mathcal{V}} |g| \sum_{g \in \mathcal{V}} t_g \right)
\]
\[
\leq \eta \sum_{g \in \mathcal{V}} |g| t_g \leq \eta \log \max_{g \in \mathcal{V}} |g| \sum_{g \in \mathcal{V}} t_g
\]
\[
\Rightarrow \Pr \left( \sqrt{n} \rho_+^{1/2}(3s, 3t) \geq \sqrt{n} + O\left(\sum_{g \in \mathcal{V}} |t_g| \log \max_{g \in \mathcal{V}} |g|\right) + O\left(\log \frac{1}{\eta'}\right) \right) \leq \eta'$
It indicates that if \( n \geq O(\sum_{g \in G} t_g \max_{g \in G} |g|) \), then we have \( \rho_{\pm}^{1/2}(3s, 3t) \leq \sqrt{\frac{3}{2}} \) with high probability as well. Similarly, we can prove \( \rho_{\pm}^{1/2}(3s, 3t) \leq \sqrt{\frac{1}{2}} \) with high probability.

\[ \text{Proof to Theorem 3} \]

\[ \text{Proof.} \] Since \( n \) is large enough as shown in (4), from Lemma 7, we have \( \alpha < 1 \) and are allowed to apply Theorem 2. Since \( \Delta = 0 \) for the noiseless case, we prove the theorem by letting \( \bar{w} \) be \( w^* \).

\[ \text{Proof to Theorem 4} \]

\[ \text{Proof.} \] Since \( n \) is large enough as shown in (4), from Lemma 7, we have \( \alpha < 1 \) and are allowed to apply Theorem 2. From Lemma 6, we obtain the upper bound for \( \Delta \). When the number of iterations \( k \) is large enough such that \( \alpha^k \|w^0 - \bar{w}\| \) reduces the magnitude of \( \Delta \), we can easily prove the error bound of \( w^k \) letting \( \bar{w} \) be \( w^* \). The second claim can be similarly proven by applying the second claim in Theorem 2.