Dual Iterative Hard Thresholding: From Non-convex Sparse Minimization to Non-smooth Concave Maximization
Supplementary File

A. Technical Proofs

A.1. Proof of Theorem 1

Proof. "$\Leftarrow$": If the pair $(\bar{w}, \bar{\alpha})$ is a sparse saddle point for L, then from the definition of conjugate convexity and inequality (3) we have

$$P(\bar{w}) = \max_{\alpha \in \mathcal{F}^N} L(\bar{w}, \alpha) \leq L(\bar{w}, \bar{\alpha}) \leq \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}).$$

On the other hand, we know that for any $\|w\|_0 \leq k$ and $\alpha \in \mathcal{F}^N$

$$L(w, \alpha) \leq \max_{\alpha' \in \mathcal{F}^N} L(w, \alpha') = P(w).$$

By combining the preceding two inequalities we obtain

$$P(\bar{w}) \leq \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}) \leq \min_{\|w\|_0 \leq k} P(w) \leq P(\bar{w}).$$

Therefore $P(\bar{w}) = \min_{\|w\|_0 \leq k} P(w)$, i.e., $\bar{w}$ solves the problem in (1), which proves the necessary condition (a). Moreover, the above arguments lead to

$$P(\bar{w}) = \max_{\alpha \in \mathcal{F}^N} L(\bar{w}, \alpha) = L(\bar{w}, \bar{\alpha}).$$

Then from the maximizing argument property of convex conjugate we know that $\bar{\alpha}_i \in \partial l_i(\bar{w}^\top x_i)$. Thus the necessary condition (b) holds. Note that

$$L(w, \bar{\alpha}) = \frac{\lambda}{2} \left\| w + \frac{1}{N\lambda} \sum_{i=1}^N \bar{\alpha}_i x_i \right\|^2 - \frac{1}{N} \sum_{i=1}^N l_i^*(\bar{\alpha}_i) + C, \tag{11}$$

where C is a quantity not dependent on w. Let $\bar{F} = \mathrm{supp}(\bar{w})$. Since the above analysis implies $L(\bar{w}, \bar{\alpha}) = \min_{\|w\|_0 \leq k} L(w, \bar{\alpha})$, it must hold that

$$\bar{w} = \mathrm{H}_{\bar{F}} \left(-\frac{1}{N\lambda} \sum_{i=1}^N \bar{\alpha}_i x_i \right) = \mathrm{H}_k \left(-\frac{1}{N\lambda} \sum_{i=1}^N \bar{\alpha}_i x_i \right).$$

This validates the necessary condition (c).

"$\Rightarrow$": Conversely, let us assume that $\bar{w}$ is a k-sparse solution to the problem (1) (i.e., conditio(a)) and let $\bar{\alpha}_i \in \partial l_i(\bar{w}^\top x_i)$ (i.e., condition (b)). Again from the maximizing argument property of convex conjugate we know that $l_i(\bar{w}^\top x_i) = \bar{\alpha}_i \bar{w}^\top x_i - l_i^*(\bar{\alpha}_i)$. This leads to

$$L(\bar{w}, \alpha) \leq P(\bar{w}) = \max_{\alpha \in \mathcal{F}^N} L(\bar{w}, \alpha) = L(\bar{w}, \bar{\alpha}). \tag{12}$$

The sufficient condition (c) guarantees that $\bar{F}$ contains the top k (in absolute value) entries of $-\frac{1}{N\lambda} \sum_{i=1}^N \bar{\alpha}_i x_i$. Then based on the expression in (11) we can see that the following holds for any k-sparse vector w

$$L(\bar{w}, \bar{\alpha}) \leq L(w, \bar{\alpha}). \tag{13}$$

By combining the inequalities (12) and (13) we get that for any $\|w\|_0 \leq k$ and $\alpha \in \mathcal{F}^N$,

$$L(\bar{w}, \alpha) \leq L(\bar{w}, \bar{\alpha}) \leq L(w, \bar{\alpha}).$$

This shows that $(\bar{w}, \bar{\alpha})$ is a sparse saddle point of the Lagrangian L. $\square$

A.2. Proof of Theorem 2

Proof. "$\Rightarrow$": Let $(\bar{w}, \bar{\alpha})$ be a saddle point for L. On one hand, note that the following holds for any k-sparse w' and $\alpha' \in \mathcal{F}^N$

$$\min_{\|w\|_0 \le k} L(w, \alpha') \le L(w', \alpha') \le \max_{\alpha \in \mathcal{F}^N} L(w', \alpha),$$

which implies

$$\max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \le k} L(w, \alpha) \le \min_{\|w\|_0 \le k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha). \tag{14}$$

On the other hand, since $(\bar{w}, \bar{\alpha})$ is a saddle point for L, the following is true:

$$\begin{aligned} \min_{\|w\|_0 \le k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha) &\le \max_{\alpha \in \mathcal{F}^N} L(\bar{w}, \alpha) \\ &\le L(\bar{w}, \bar{\alpha}) \le \min_{\|w\|_0 \le k} L(w, \bar{\alpha}) \\ &\le \max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \le k} L(w, \alpha). \end{aligned} \tag{15}$$

By combining (14) and (15) we prove the equality in (4).

"$\Leftarrow$": Assume that the equality in (4) holds. Let us define $\bar{w}$ and $\bar{\alpha}$ such that

$$\begin{aligned} \max_{\alpha \in \mathcal{F}^N} L(\bar{w}, \alpha) &= \min_{\|w\|_0 \le k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha) \\ \min_{\|w\|_0 \le k} L(w, \bar{\alpha}) &= \max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \le k} L(w, \alpha) \end{aligned}.$$

Then we can see that for any $\alpha \in \mathcal{F}^N$,

$$L(\bar{w}, \bar{\alpha}) \ge \min_{\|w\|_0 \le k} L(w, \bar{\alpha}) = \max_{\alpha' \in \mathcal{F}^N} L(\bar{w}, \alpha') \ge L(\bar{w}, \alpha),$$

where the "$=$" is due to (4). In the meantime, for any $\|w\|_0 \le k$,

$$L(\bar{w}, \bar{\alpha}) \le \max_{\alpha \in \mathcal{F}^N} L(\bar{w}, \alpha) = \min_{\|w'\|_0 \le k} L(w', \bar{\alpha}) \le L(w, \bar{\alpha}).$$

This shows that $(\bar{w}, \bar{\alpha})$ is a sparse saddle point for L. ☐

A.3. Proof of Lemma 1

Proof. For any fixed $\alpha \in \mathcal{F}^N$, then it is easy to verify that the k-sparse minimum of $L(w, \alpha)$ with respect to w is attained at the following point:

$$w(\alpha) = \arg\min_{\|w\|_0 \le k} L(w, \alpha) = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N} \alpha_i x_i\right).$$

Thus we have

$$\begin{aligned} D(\alpha) &= \min_{\|w\|_0 \le k} L(w, \alpha) = L(w(\alpha), \alpha) \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(\alpha_i w(\alpha)^\top x_i - l_i^*(\alpha_i)\right) + \frac{\lambda}{2}\|w(\alpha)\|^2 \\ &\overset{\zeta_1}{=} \frac{1}{N}\sum_{i=1}^{N} -l_i^*(\alpha_i) - \frac{\lambda}{2}\|w(\alpha)\|^2, \end{aligned}$$

where "ζ_1" follows from the above definition of $w(\alpha)$.

Now let us consider two arbitrary dual variables $\alpha', \alpha'' \in \mathcal{F}^N$ and any $g(\alpha'') \in \frac{1}{N}[w(\alpha'')^\top x_1 - \partial l_1^*(\alpha_1''), ..., w(\alpha'')^\top x_N - \partial l_N^*(\alpha_N'')]$. From the definition of $D(\alpha)$ and the fact that $L(w, \alpha)$ is concave with respect to α at any fixed w we can derive that

$$\begin{aligned} D(\alpha') &= L(w(\alpha'), \alpha') \\ &\le L(w(\alpha''), \alpha') \\ &\le L(w(\alpha''), \alpha'') + \langle g(\alpha''), \alpha' - \alpha'' \rangle. \end{aligned}$$

This shows that $D(\alpha)$ is a concave function and its super-differential is as given in the theorem.

If we further assume that $w(\alpha)$ is unique and $\{l_i^*\}_{i=1,...,N}$ are differentiable at any α, then $\partial D(\alpha) = \frac{1}{N}[w(\alpha)^\top x_1 - \partial l_1^*(\alpha_1), ..., w(\alpha)^\top x_N - \partial l_N^*(\alpha_N)]$ becomes unique, which implies that $\partial D(\alpha)$ is the unique super-gradient of $D(\alpha)$. □

A.4. Proof of Theorem 3

Proof. "⇒": Given the conditions in the theorem, it can be known from Theorem 1 that the pair $(\bar{w}, \bar{\alpha})$ forms a sparse saddle point of L. Thus based on the definitions of sparse saddle point and dual function $D(\alpha)$ we can show that

$$D(\bar{\alpha}) = \min_{\|w\|_0 \leq k} L(w, \bar{\alpha}) \geq L(\bar{w}, \bar{\alpha}) \geq L(\bar{w}, \alpha) \geq D(\alpha).$$

This implies that $\bar{\alpha}$ solves the dual problem in (5). Furthermore, Theorem 2 guarantees the following

$$D(\bar{\alpha}) = \max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \leq k} L(w, \alpha) = \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha) = P(\bar{w}).$$

This indicates that the primal and dual optimal values are equal to each other.

"⇐": Assume that $\bar{\alpha}$ solves the dual problem in (5) and $D(\bar{\alpha}) = P(\bar{w})$. Since $D(\bar{\alpha}) \leq P(w)$ holds for any $\|w\|_0 \leq k$, $\bar{w}$ must be the sparse minimizer of $P(w)$. It follows that

$$\max_{\alpha \in \mathcal{F}^N} \min_{\|w\|_0 \leq k} L(w, \alpha) = D(\bar{\alpha}) = P(\bar{w}) = \min_{\|w\|_0 \leq k} \max_{\alpha \in \mathcal{F}^N} L(w, \alpha).$$

From the "⇐" argument in the proof of Theorem 2 and Corollary 1 we get that the conditions (a)∼(c) in Theorem 1 should be satisfied for $(\bar{w}, \bar{\alpha})$. □

A.5. Proof of Theorem 4

We need a series of technical lemmas to prove this theorem. The following lemmas shows that under proper conditions, $w(\alpha)$ is locally smooth around $\bar{w} = w(\bar{\alpha})$.

Lemma 2. *Let $X = [x_1, ..., x_N] \in \mathbb{R}^{d \times N}$ be the data matrix. Assume that $\{l_i\}_{i=1,...,N}$ are differentiable and*

$$\bar{\epsilon} := \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0.$$

If $\|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\sigma_{\max}(X)}$, then supp$(w(\alpha)) =$ supp$(\bar{w})$ and

$$\|w(\alpha) - \bar{w}\| \leq \frac{\sigma_{\max}(X, k)}{N\lambda}\|\alpha - \bar{\alpha}\|.$$

Proof. For any $\alpha \in \mathcal{F}^N$, let us define

$$\tilde{w}(\alpha) = -\frac{1}{N\lambda}\sum_{i=1}^{N} \alpha_i x_i.$$

Consider $\bar{F} = \text{supp}(\bar{w})$. Given $\bar{\epsilon} > 0$, it is known from Theorem 3 that $\bar{w} = \mathrm{H}_{\bar{F}}(\tilde{w}(\bar{\alpha}))$ and $\frac{P'(\bar{w})}{\lambda} = \mathrm{H}_{\bar{F}^c}(-\tilde{w}(\bar{\alpha}))$. Then $\bar{\epsilon} > 0$ implies $\bar{F}$ is unique, i.e., the top k entries of $\tilde{w}(\bar{\alpha})$ is unique. Given that $\|\alpha - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\sigma_{\max}(X)}$, it can be shown that

$$\|\tilde{w}(\alpha) - \tilde{w}(\bar{\alpha})\| = \frac{1}{N\lambda}\|X(\alpha - \bar{\alpha})\| \leq \frac{\sigma_{\max}(X)}{N\lambda}\|\alpha - \bar{\alpha}\| \leq \frac{\bar{\epsilon}}{2}.$$

This indicates that $\bar{F}$ still contains the (unique) top k entries of $\tilde{w}(\alpha)$. Therefore,

$$\text{supp}(w(\alpha)) = \bar{F} = \text{supp}(\bar{w}).$$

Then it must hold that

$$\begin{aligned}
\|w(\alpha) - w(\bar{\alpha})\| &= \|\mathrm{H}_{\bar{F}}(\tilde{w}(\alpha)) - \mathrm{H}_{\bar{F}}(\tilde{w}(\bar{\alpha}))\| \\
&= \frac{1}{N\lambda}\|X_{\bar{F}}(\alpha - \bar{\alpha})\| \\
&\leq \frac{\sigma_{\max}(X, k)}{N\lambda}\|\alpha - \bar{\alpha}\|.
\end{aligned}$$

This proves the desired bound. □

The following lemma bounds the estimation error $\|\alpha - \bar{\alpha}\| = O(\sqrt{\langle D'(\alpha), \bar{\alpha} - \alpha\rangle})$ when the primal loss $\{l_i\}_{i=1}^N$ are smooth.

Lemma 3. *Assume that the primal loss functions $\{l_i(\cdot)\}_{i=1}^N$ are $1/\mu$-smooth. Then the following inequality holds for any $\alpha, \alpha'' \in \mathcal{F}$ and $g(\alpha'') \in \partial D(\alpha'')$:*

$$D(\alpha') \leq D(\alpha'') + \langle g(\alpha''), \alpha' - \alpha''\rangle - \frac{\lambda N\mu + \sigma_{\min}^2(X,k)}{2\lambda N^2}\|\alpha' - \alpha''\|^2.$$

Moreover, $\forall \alpha \in \mathcal{F}$ and $g(\alpha) \in \partial D(\alpha)$,

$$\|\alpha - \bar{\alpha}\| \leq \sqrt{\frac{2\lambda N^2 \langle g(\alpha), \bar{\alpha} - \alpha\rangle}{\lambda N\mu + \sigma_{\min}^2(X,k)}}.$$

Proof. Recall that

$$D(\alpha) = \frac{1}{N}\sum_{i=1}^N -l_i^*(\alpha_i) - \frac{\lambda}{2}\|w(\alpha)\|^2,$$

Now let us consider two arbitrary dual variables $\alpha', \alpha'' \in \mathcal{F}$. The assumption of l_i being $1/\mu$-smooth implies that its convex conjugate function l_i^* is μ-strongly-convex. Let $F'' = \mathrm{supp}(w(\alpha''))$. Then

$$\begin{aligned}
D(\alpha') =& \frac{1}{N}\sum_{i=1}^N -l_i^*(\alpha_i') - \frac{\lambda}{2}\|w(\alpha')\|^2 \\
=& \frac{1}{N}\sum_{i=1}^N -l_i^*(\alpha_i') - \frac{\lambda}{2}\left\|\mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^N \alpha_i' x_i\right)\right\|^2 \\
\leq& \frac{1}{N}\sum_{i=1}^N \left(-l_i^*(\alpha_i'') - l_i^{*'}(\alpha_i'')(\alpha_i' - \alpha_i'') - \frac{\mu}{2}(\alpha_i' - \alpha_i'')^2\right) - \frac{\lambda}{2}\left\|\mathrm{H}_{F''}\left(-\frac{1}{N\lambda}\sum_{i=1}^N \alpha_i' x_i\right)\right\|^2 \\
\leq& \frac{1}{N}\sum_{i=1}^N \left(-l_i^*(\alpha_i'') - l_i^{*'}(\alpha_i'')(\alpha_i' - \alpha_i'') - \frac{\mu}{2}(\alpha_i' - \alpha_i'')^2\right) - \frac{\lambda}{2}\|w(\alpha'')\|^2 + \frac{1}{N}\sum_{i=1}^N x_i^\top w(\alpha'')(\alpha_i' - \alpha_i'') \\
&- \frac{1}{2\lambda N^2}(\alpha' - \alpha'')^\top X_{F''}^\top X_{F''}(\alpha' - \alpha'') \\
\leq& D(\alpha'') + \langle g(\alpha''), \alpha' - \alpha''\rangle - \frac{\lambda N\mu + \sigma_{\min}^2(X,k)}{2\lambda N^2}\|\alpha' - \alpha''\|^2.
\end{aligned}$$

This proves the first desirable inequality in the lemma. By invoking the above inequality and using the fact $D(\alpha) \leq D(\bar{\alpha})$ we get that

$$\begin{aligned}
D(\bar{\alpha}) \leq& D(\alpha) + \langle g(\alpha), \bar{\alpha} - \alpha\rangle - \frac{\lambda N\mu + \sigma_{\min}^2(X,k)}{2\lambda N^2}\|\alpha - \bar{\alpha}\|^2 \\
\leq& D(\bar{\alpha}) + \langle g(\alpha), \bar{\alpha} - \alpha\rangle - \frac{\lambda N\mu + \sigma_{\min}^2(X,k)}{2\lambda N^2}\|\alpha - \bar{\alpha}\|^2,
\end{aligned}$$

which leads to the second desired bound. □

The following lemma gives a simple expression of the gap for properly related primal-dual pairs.

Lemma 4. *Given a dual variable $\alpha \in \mathcal{F}^N$ and the related primal variable*

$$w = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^N \alpha_i x_i\right).$$

The primal-dual gap $\epsilon_{PD}(w, \alpha)$ can be expressed as:

$$\epsilon_{PD}(w,\alpha) = \frac{1}{N}\sum_{i=1}^N \left(l_i(w^\top x_i) + l_i^*(\alpha_i) - \alpha_i w^\top x_i\right).$$

Proof. It is directly to know from the definitions of $P(w)$ and $D(\alpha)$ that

$$
\begin{aligned}
& P(w) - D(\alpha) \\
=& \frac{1}{N}\sum_{i=1}^{N} l_i(w^\top x_i) + \frac{\lambda}{2}\|w\|^2 - \left(\frac{1}{N}\sum_{i=1}^{N}\left(\alpha_i w^\top x_i - l_i^*(\alpha_i)\right) + \frac{\lambda}{2}\|w\|^2\right) \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(l_i(w^\top x_i) + l_i^*(\alpha_i) - \alpha_i w^\top x_i\right).
\end{aligned}
$$

This shows the desired expression. □

Based on Lemma 4, we can derive the following lemma which establishes a bound on the primal-dual gap.

Lemma 5. *Consider a primal-dual pair (w, α) satisfying*

$$
w = \mathrm{H}_k\left(-\frac{1}{N\lambda}\sum_{i=1}^{N}\alpha_i x_i\right).
$$

Then the following inequality holds for any $g(\alpha) \in \partial D(\alpha)$ and $\beta \in [\partial l_1(w^\top x_1), ..., \partial l_N(w^\top x_N)]$:

$$
P(w) - D(\alpha) \le \langle g(\alpha), \beta - \alpha\rangle.
$$

Proof. For any $i \in [1, ..., N]$, from the maximizing argument property of convex conjugate we have

$$
l_i(w^\top x_i) = w^\top x_i l_i'(w^\top x_i) - l_i^*(l_i'(w^\top x_i)),
$$

and

$$
l_i^*(\alpha_i) = \alpha_i l_i^{*'}(\alpha_i) - l_i(l_i^{*'}(\alpha_i)).
$$

By summing both sides of above two equalities we get

$$
\begin{aligned}
& l_i(w^\top x_i) + l_i^*(\alpha_i) \\
=& w^\top x_i l_i'(w^\top x_i) + \alpha_i l_i^{*'}(\alpha_i) - (l_i(l_i^{*'}(\alpha_i)) + l_i^*(l_i'(w^\top x_i))) \\
\overset{\zeta_1}{\le}& w^\top x_i l_i'(w^\top x_i) + \alpha_i l_i^{*'}(\alpha_i) - l_i^{*'}(\alpha_i) l_i'(w^\top x_i),
\end{aligned}
\tag{16}
$$

where "ζ_1" follows from Fenchel-Young inequality. Therefore

$$
\begin{aligned}
& \langle g(\alpha), \beta - \alpha\rangle \\
=& \frac{1}{N}\sum_{i=1}^{N}(w^\top x_i - l_i^{*'}(\alpha_i))(l_i'(w^\top x_i) - \alpha_i) \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(w^\top x_i l_i'(w^\top x_i) - l_i^{*'}(\alpha_i) l_i'(w^\top x_i) - \alpha_i w^\top x_i + \alpha_i l_i^{*'}(\alpha_i)\right) \\
\overset{\zeta_2}{\ge}& \frac{1}{N}\sum_{i=1}^{N}(l_i(w^\top x_i) + \alpha_i l_i^*(\alpha_i) - w^\top x_i) \\
\overset{\zeta_3}{=}& P(w) - D(\alpha),
\end{aligned}
$$

where "ζ_2" follows from (16) and "ζ_3" follows from Lemma 4. This proves the desired bound. □

The following simple result is also needed in our iteration complexity analysis.

Lemma 6. *For any $\epsilon > 0$,*

$$
\frac{1}{t} + \frac{\ln t}{t} \le \epsilon
$$

holds when $t \ge \max\left\{\frac{3}{\epsilon}\ln\frac{3}{\epsilon}, 1\right\}$.

Proof. Obviously, the inequality $\frac{1}{t} + \frac{\ln t}{t} \le \epsilon$ holds for $\epsilon \ge 1$. When $\epsilon < 1$, it holds that $\ln(\frac{3}{\epsilon}) \ge 1$. Then the condition on t implies that $\frac{1}{t} \le \frac{\epsilon}{3}$. Also, we have

$$\frac{\ln t}{t} \le \frac{\ln(\frac{3}{\epsilon}\ln\frac{3}{\epsilon})}{\frac{3}{\epsilon}\ln\frac{3}{\epsilon}} \le \frac{\ln(\frac{3}{\epsilon})^2}{\frac{3}{\epsilon}\ln\frac{3}{\epsilon}} = \frac{2\epsilon}{3},$$

where the first "$\le$" follows the fact that $\ln t/t$ is decreasing when $t \ge 1$ while the second "$\le$" follows $\ln x < x$ for all $x > 0$. Therefore we have $\frac{1}{t} + \frac{\ln t}{t} \le \epsilon$. $\square$

We are now in the position to prove the main theorem.

of Theorem 4. **Part(a):** Let us consider $g^{(t)} \in \partial D(\alpha^{(t)})$ with $g_i^{(t)} = \frac{1}{N}(x_i^\top w^{(t)} - l_i^{*\prime}(\alpha_i^{(t)}))$. From the expression of $w^{(t)}$ we can verify that $\|w^{(t)}\| \le r/\lambda$. Therefore we have

$$\|g^{(t)}\| \le c_0 = \frac{r + \lambda\rho}{\lambda\sqrt{N}}.$$

Let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)}, \bar{\alpha} - \alpha^{(t)}\rangle$. The concavity of D implies $v^{(t)} \ge 0$. From Lemma 3 we know that $h^{(t)} \le \sqrt{2\lambda N^2 v^{(t)}/(\lambda N\mu + \sigma_{\min}(X,k))}$. Then

$$\begin{aligned}
(h^{(t)})^2 =& \|\mathrm{P}_{\mathcal{F}^N}\left(\alpha^{(t-1)} + \eta^{(t-1)} g^{(t-1)}\right) - \bar{\alpha}\|^2 \\
\le& \|\alpha^{(t-1)} + \eta^{(t-1)} g^{(t-1)} - \bar{\alpha}\|^2 \\
=& (h^{(t-1)})^2 - 2\eta^{(t-1)} v^{(t-1)} + (\eta^{(t-1)})^2 \|g^{(t-1)}\|^2 \\
\le& (h^{(t-1)})^2 - \frac{\eta^{(t-1)}(\lambda N\mu + \sigma_{\min}(X,k))}{\lambda N^2}(h^{(t-1)})^2 + (\eta^{(t-1)})^2 c_0^2.
\end{aligned}$$

Let $\eta^{(t)} = \frac{\lambda N^2}{(\lambda N\mu + \sigma_{\min}(X,k))(t+1)}$. Then we obtain

$$(h^{(t)})^2 \le \left(1 - \frac{1}{t}\right)(h^{(t-1)})^2 + \frac{\lambda^2 N^4 c_0^2}{(\lambda N\mu + \sigma_{\min}(X,k))^2 t^2}.$$

By recursively applying the above inequality we get

$$(h^{(t)})^2 \le \frac{\lambda^2 N^4 c_0^2}{(\lambda N\mu + \sigma_{\min}(X,k))^2}\left(\frac{1}{t} + \frac{\ln t}{t}\right) = c_1\left(\frac{1}{t} + \frac{\ln t}{t}\right).$$

This proves the desired bound in part(a).

Part(b): Let us consider $\epsilon = \frac{\lambda N\bar{\epsilon}}{2\sigma_{\max}(X)}$. From part(a) and Lemma 6 we obtain

$$\|\alpha^{(t)} - \bar{\alpha}\| \le \epsilon$$

after $t \ge t_0 = \frac{3c_1}{\epsilon^2}\ln\frac{3c_1}{\epsilon^2}$. It follows from Lemma 2 that $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$.

Let $\beta^{(t)} := [l_1'((w^{(t)})^\top x_1), ..., l_N'((w^{(t)})^\top x_N)]$. According to Lemma 5 we have

$$\begin{aligned}
\epsilon_{PD}^{(t)} &= P(w^{(t)}) - D(\alpha^{(t)}) \\
&\le \langle g^{(t)}, \beta^{(t)} - \alpha^{(t)}\rangle \\
&\le \|g^{(t)}\|(\|\beta^{(t)} - \bar{\alpha}\| + \|\bar{\alpha} - \alpha^{(t)}\|).
\end{aligned}$$

Since $\bar{\epsilon} = \bar{w}_{\min} - \frac{1}{\lambda}\|P'(\bar{w})\|_\infty > 0$, it follows from Theorem 2 that $\bar{\alpha} = [l_1'(\bar{w}^\top x_1), ..., l_N'(\bar{w}^\top x_N)]$. Given that $t \ge t_0$, from the smoothness of l_i and Lemma 2 we get

$$\|\beta^{(t)} - \bar{\alpha}\| \le \frac{1}{\mu}\|w^{(t)} - \bar{w}\| \le \frac{\sigma_{\max}(X,k)}{\mu\lambda N}\|\alpha^{(t)} - \bar{\alpha}\|, .$$

where in the first "$\leq$" we have used $\|x_i\| \leq 1$. Therefore, the following is valid when $t \geq t_0$:

$$\begin{aligned}\epsilon_{PD}^{(t)} &\leq \|g^{(t)}\|(\|\beta^{(t)} - \bar{\alpha}\| + \|\bar{\alpha} - \alpha^{(t)}\|) \\ &\leq c_0 \left(1 + \frac{\sigma_{\max}(X,k)}{\mu\lambda N}\right) \|\alpha^{(t)} - \bar{\alpha}\|.\end{aligned}$$

Since $t \geq t_1$, from part(a) and Lemma 6 we get $\|\alpha^{(t)} - \bar{\alpha}\| \leq \frac{\epsilon}{c_0\left(1+\frac{\sigma_{\max}(X,k)}{\mu\lambda N}\right)}$, which according to the above inequality implies $\epsilon_{PD}^{(t)} \leq \epsilon$. This proves the desired bound. □

A.6. Proof of Theorem 5

Proof. **Part(a):** Let us consider $g^{(t)}$ with $g_j^{(t)} = \frac{1}{N}(x_j^\top w^{(t)} - l_j^{*\prime}(\alpha_i^{(t)}))$. Let $h^{(t)} = \|\alpha^{(t)} - \bar{\alpha}\|$ and $v^{(t)} = \langle g^{(t)}, \bar{\alpha} - \alpha^{(t)}\rangle$. The concavity of D implies $v^{(t)} \geq 0$. From Lemma 3 we know that $h^{(t)} \leq \sqrt{2\lambda N^2 v^{(t)}/(\lambda N\mu + \sigma_{\min}(X,k))}$. Let $g_{B_i}^{(t)} := \mathrm{H}_{B_i^{(t)}}(g^{(t)})$ and $v_{B_i}^{(t)} := \langle g_{B_i}^{(t)}, \bar{\alpha} - \alpha^{(t)}\rangle$ Then

$$\begin{aligned}(h^{(t)})^2 =& \|\mathrm{P}_{\mathcal{F}^\mathcal{N}}\left(\alpha^{(t-1)} + \eta^{(t-1)} g_{B_i}^{(t-1)}\right) - \bar{\alpha}\|^2 \\ \leq& \|\alpha^{(t-1)} + \eta^{(t-1)} g_{B_i}^{(t-1)} - \bar{\alpha}\|^2 \\ =& (h^{(t-1)})^2 - 2\eta^{(t-1)} v_{B_i}^{(t-1)} + (\eta^{(t-1)})^2 \|g_{B_i}^{(t-1)}\|^2.\end{aligned}$$

By taking conditional expectation (with respect to uniform random block selection, conditioned on $\alpha^{(t-1)}$) on both sides of the above inequality we get

$$\begin{aligned}&\mathbb{E}[(h^{(t)})^2 \mid \alpha^{(t-1)}] \\ \leq& (h^{(t-1)})^2 - \frac{1}{m}\sum_{i=1}^m 2\eta^{(t-1)} v_{B_i}^{(t-1)} + \frac{1}{m}\sum_{i=1}^m (\eta^{(t-1)})^2 \|g_{B_i}^{(t-1)}\|^2 \\ =& (h^{(t-1)})^2 - \frac{2\eta^{(t-1)}}{m} v^{(t-1)} + \frac{(\eta^{(t-1)})^2}{m}\|g^{(t-1)}\|^2 \\ \leq& (h^{(t-1)})^2 - \frac{\eta^{(t-1)}(\lambda N\mu + \sigma_{\min}(X,k))}{\lambda m N^2}(h^{(t-1)})^2 + \frac{(\eta^{(t-1)})^2}{m} c_0^2..\end{aligned}$$

Let $\eta^{(t)} = \frac{\lambda m N^2}{(\lambda N\mu + \sigma_{\min}(X,k))(t+1)}$. Then we obtain

$$\mathbb{E}[(h^{(t)})^2 \mid \alpha^{(t-1)}] \leq \left(1 - \frac{1}{t}\right)(h^{(t-1)})^2 + \frac{\lambda^2 m N^4 c_0^2}{(\lambda N\mu + \sigma_{\min}(X,k))^2 t^2}.$$

By taking expectation on both sides of the above over $\alpha^{(t-1)}$, we further get

$$\mathbb{E}[(h^{(t)})^2] \leq \left(1 - \frac{1}{t}\right)\mathbb{E}[(h^{(t-1)})^2] + \frac{\lambda^2 m N^4 c_0^2}{(\lambda N\mu + \sigma_{\min}(X,k))^2 t^2}.$$

This recursive inequality leads to

$$\mathbb{E}[(h^{(t)})^2] \leq \frac{\lambda^2 m N^4 c_0^2}{(\lambda N\mu + \sigma_{\min}(X,k))^2}\left(\frac{1}{t} + \frac{\ln t}{t}\right) = c_2\left(\frac{1}{t} + \frac{\ln t}{t}\right).$$

This proves the desired bound in part(a).

Part(b): Let us consider $\epsilon = \frac{\lambda N\bar{\epsilon}}{2\sigma_{\max}(X)}$. From part(a) and Lemma 6 we obtain

$$\mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|] \leq \delta\epsilon$$

after $t \geq t_2 = \frac{3c_2}{\delta^2\epsilon^2}\ln\frac{3c_2}{\delta^2\epsilon^2}$. Then from Markov inequality we know that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \mathbb{E}[\|\alpha^{(t)} - \bar{\alpha}\|]/\delta \leq \epsilon$ holds with probability at least $1-\delta$. Lemma 2 shows that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \epsilon$ implies $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$. Therefore when $t \geq t_2$, the event $\mathrm{supp}(w^{(t)}) = \mathrm{supp}(\bar{w})$ occurs with probability at least $1-\delta$.

Similar to the proof arguments of Theorem 4(b) we can further show that when $t \geq 4t_2$, with probability at least $1 - \delta/2$

$$\|\alpha^{(t)} - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\sigma_{\max}(X)},$$

which then leads to

$$\epsilon_{PD}^{(t)} \leq c_0 \left(1 + \frac{\sigma_{\max}(X,k)}{\mu \lambda N}\right) \|\alpha^{(t)} - \bar{\alpha}\|.$$

Since $t \geq t_3$, from the arguments in part(a) and Lemma 6 we get that $\|\alpha^{(t)} - \bar{\alpha}\| \leq \frac{\epsilon}{c_0\left(1+\frac{\sigma_{\max}(X,k)}{\mu\lambda N}\right)}$ holds with probability at least $1 - \delta/2$. Let us consider the following events:

- $\mathcal{A}$: the event of $\epsilon_{PD}^{(t)} \leq \epsilon$;
- $\mathcal{B}$: the event of $\|\alpha^{(t)} - \bar{\alpha}\| \leq \frac{\lambda N \bar{\epsilon}}{2\sigma_{\max}(X)}$;
- $\mathcal{C}$: the event of $\|\alpha^{(t)} - \bar{\alpha}\| \leq \frac{\epsilon}{c_0\left(1+\frac{\sigma_{\max}(X,k)}{\mu\lambda N}\right)}$.

When $t \geq \max\{4t_2, t_3\}$, we have the following holds:

$$\mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{A} \mid \mathcal{B})\mathbb{P}(\mathcal{B}) \geq \mathbb{P}(\mathcal{C} \mid \mathcal{B})\mathbb{P}(\mathcal{B}) \geq (1 - \delta/2)^2 \geq 1 - \delta.$$

This proves the desired bound. $\square$