A Projection Region

In Section 2, we compare the projection regions of different algorithms in Figure 2, and conclude that the orthant step of OBProx-SG enjoys a much larger projection to map a trial iterate to zero so that it is a more aggressive sparsity promotion mechanism than the others. Hence the solutions computed by OBProx-SG tends to be more sparse. In this Appendix, we present the deduction of the projection region for Orthant Step in 1-dimensional example. The 1-dimensional result can be easily extended to higher-dimensional space.

Proposition 1. *If $k \in \mathcal{S}_{\mathcal{O}}$, $0 < x_k \in \mathbb{R}^1$, the Orthant Step of OBProx-SG yields next iterate x_{k+1} based on the trial iterate $\hat{x}_{k+1} = x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)$ as follows*

$$x_{k+1} = \begin{cases} \hat{x}_{k+1} - \alpha_k \lambda & \textit{if } \hat{x}_{k+1} > \alpha_k \lambda, \\ 0 & \textit{otherwise}. \end{cases} \tag{29}$$

Therefore, the projection region of Orthant Step is $(-\infty, \alpha_k \lambda]$ to map $\hat{x}_{k+1}$ to zero if $x_k > 0$. Similarly, the projection region as $[-\alpha_k \lambda, \infty)$ is attained if $x_k < 0$.

Proof. It follows the definition of $\tilde{F}$ as (9) and $x_k > 0$ that

$$\tilde{F}_{\mathcal{B}_k}(x) = f_{\mathcal{B}_k}(x) + \lambda x, \tag{30}$$

$$\nabla \tilde{F}_{\mathcal{B}_k}(x) = \nabla f_{\mathcal{B}_k}(x) + \lambda \tag{31}$$

By the update mechanism of Orthant Step in Algorithm 3, the next iterate x_{k+1} is computed by the following

$$x_{k+1} = \begin{cases} x_k - \alpha_k \nabla \tilde{F}_{\mathcal{B}_k}(x_k) & \text{if } x_k - \alpha_k \nabla \tilde{F}_{\mathcal{B}_k}(x_k) > 0, \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

Combining with (31) and $\hat{x}_{k+1} = x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)$, (32) is equivalent to

$$x_{k+1} = \begin{cases} \hat{x}_{k+1} - \alpha_k \lambda & \text{if } \hat{x}_{k+1} > \alpha_k \lambda, \\ 0 & \text{otherwise,} \end{cases}$$

which completes the proof.

Finally, we remark here that the projection region of Orthant Step in OBProx-SG is a superset of that of Prox-SG and Prox-SVRG, where the trial iterate of Prox-SVRG is computed under SVRG [8]. RDA possesses a different projection region as $[-\lambda, \lambda]$ to produce zero elements if the dual averaging inhabits [7].

B Convergence Analysis Proofs

In Appendix-B, we present the proofs of the theorems stated in Section 3. We first describe the sufficient decrease properties of Prox-SG Step and Orthant Step in Section B.1. We then derive the main convergence results for convex settings in Section B.2. We establish an non-asymptotic upper bound of $N_{\mathcal{P}}$ for OBProx-SG+ in Section B.3. Finally, we generalize our conclusions in non-convex scenario in Section B.4.

B.1 Sufficient decrease by Prox-SG Step and Orthant Step

The lemma below is well known for proximal operator under our notations. We include this proof for completeness.

Lemma 1. *Suppose $k \in \mathcal{S}_{\mathcal{P}}$, line 3 of Algorithm 2 yields that $x_{k+1} = x_k - \alpha_k \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)$, where*

$$\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) \in \nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \|x_{k+1}\|_1. \tag{33}$$

And the objective value $F_{\mathcal{B}_k}$ satisfies

$$F_{\mathcal{B}_k}(x_{k+1}) \leq F_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \left\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_2^2. \tag{34}$$

Proof. It follows from the line (3) in Algorithm 2 and the definitions of proximal operator that

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \frac{1}{2\alpha_k} \left\|x - (x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k))\right\|_2^2 + \lambda \left\|x\right\|_1 \\ &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \nabla f_{\mathcal{B}_k}(x_k)^T (x - x_k) + \lambda \left\|x\right\|_1 + \frac{1}{2\alpha_k} \left\|x - x_k\right\|_2^2 \end{aligned} \tag{35}$$

By the optimal condition, we have

$$0 \in \frac{1}{\alpha_k}(x_{k+1} - x_k) + \nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \left\|x_{k+1}\right\|_1. \tag{36}$$

Since $x_{k+1} = x_k - \alpha_k \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)$, we have

$$0 \in -\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) + \nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \left\|x_{k+1}\right\|_1, \tag{37}$$

which implies that

$$\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) \in \nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \left\|x_{k+1}\right\|_1. \tag{38}$$

And thus there exists some $v \in \partial \left\|x_{k+1}\right\|_1$ such that

$$\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) = \nabla f_{\mathcal{B}_k}(x_k) + \lambda v. \tag{39}$$

By Lipschitz continuity of $\nabla f_{\mathcal{B}_k}$ and convexity of $\left\|\cdot\right\|_1$, we have

$$\begin{aligned} f_{\mathcal{B}_k}(x_{k+1}) &= f_{\mathcal{B}_k}(x_k - \alpha_k \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)) \\ &\leq f_{\mathcal{B}_k}(x_k) - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)^T \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2} \left\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_2^2 \end{aligned} \tag{40}$$

and

$$\begin{aligned} \lambda \left\|x_{k+1}\right\|_1 &= \lambda \left\|x_k - \alpha_k \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_1 \\ &\leq \lambda \left\|x_k\right\|_1 + \lambda v^T (x_k - \alpha_k \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) - x_k) \\ &= \lambda \left\|x_k\right\|_1 - \alpha_k \lambda v^T \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k). \end{aligned} \tag{41}$$

Hence, by (39), (40) and (41), the objective $F_{\mathcal{B}_k}(x_{k+1})$ satisfies

$$
\begin{aligned}
&F_{\mathcal{B}_k}(x_{k+1}) = f_{\mathcal{B}_k}(x_{k+1}) + \lambda \left\|x_{k+1}\right\|_1 \\
\leq &f_{\mathcal{B}_k}(x_k) - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)^T \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2} \left\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_2^2 + \lambda \left\|x_k\right\|_1 - \alpha_k \lambda v^T \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) \\
= &F_{\mathcal{B}_k}(x_k) - \alpha_k (\nabla f_{\mathcal{B}_k}(x_k) + \lambda v)^T \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2} \left\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_2^2 \\
= &F_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \left\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_2^2,
\end{aligned}
$$

which completes the proof.

We then establish a useful lemma for Orthant Step.

Lemma 2. *Suppose $k \in \mathcal{S}_{\mathcal{O}}$, line 3 of Algorithm 3 yields that $x_{k+1} = x_k + \alpha_k d_k$, where*

$$d_k \in -\left(\nabla f_{\mathcal{B}_k}(x_k) + \mathcal{N}_{\mathcal{O}_k}(x_{k+1})\right), \quad \textit{and} \tag{42}$$

$$\mathcal{N}_{\mathcal{O}_k}(x_{k+1}) := \left\{v : v^T(x_{k+1} - x) \geq 0, \forall x \in \mathcal{O}_k\right\} \tag{43}$$

is the normal cone of the orthant face $\mathcal{O}_k$ at x_{k+1}. Moreover, the objective value $F_{\mathcal{B}_k}$ satisfies

$$F_{\mathcal{B}_k}(x_{k+1}) \leq F_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \left\|d_k\right\|_2^2. \tag{44}$$

Proof. Using the fact that Euclidean projection on a set $\mathcal{O}_k$ is a proximal mapping of indicator function $I_{\mathcal{O}_k}(x)$, we have

$$
\begin{aligned}
x_{k+1} &= \mathrm{Proj}_{\mathcal{O}_k}(x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)) \\
&= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2\alpha_k} \left\|x - (x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k))\right\|_2^2 + I_{\mathcal{O}_k}(x) \\
&= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \ \nabla f_{\mathcal{B}_k}(x_k)^T (x - x_k) + I_{\mathcal{O}_k}(x) + \frac{1}{2\alpha_k} \left\|x - x_k\right\|_2^2.
\end{aligned} \tag{45}
$$

It follows $\mathcal{O}_k$ is convex that $I_{\mathcal{O}_k}(x)$ is convex. Combining with optimal condition, we have

$$0 \in \frac{1}{\alpha_k}(x_{k+1} - x_k) + \nabla f_{\mathcal{B}_k}(x_k) + \partial I_{\mathcal{O}_k}(y_{t+1}). \tag{46}$$

Let $x_{k+1} = x_k + \alpha_k d_k$, and utilizing the fact that the subdifferential of the indicator function $I_{\mathcal{O}_k}(x)$ at x_{k+1} is the normal cone $\mathcal{N}_{\mathcal{O}_k}(x_{k+1})$ [1, Example 5.4.1], we obtain

$$0 \in d_k + \nabla f_{\mathcal{B}_k}(x_k) + \mathcal{N}_{\mathcal{O}_k}(x_{k+1}), \tag{47}$$

which implies that

$$d_k \in -\left(\nabla f_{\mathcal{B}_k}(x_k) + \mathcal{N}_{\mathcal{O}_k}(x_{k+1})\right). \tag{48}$$

And thus there exists some $v \in \mathcal{N}_{\mathcal{O}_k}(x_{k+1})$ such that

$$d_k = -(\nabla f_{\mathcal{B}_k}(x_k) + v). \tag{49}$$

Let us define an auxiliary function $\Phi(x) : \mathbb{R}^n \to \mathbb{R}$ as

$$\Phi(x) := \tilde{F}_{\mathcal{B}_k}(x) + I_{\mathcal{O}_k}(x). \tag{50}$$

Note that $\tilde{F}_{\mathcal{B}_k}$ is differentiable by the definition. It follows the line 3 of Algorithm 3 that $x_k, x_{k+1} \in \mathcal{O}_k$. Combining with definition of indicator function $I_{\mathcal{O}_k}$, we have

$$\begin{aligned} \Phi(x_k) &= \tilde{F}_{\mathcal{B}_k}(x_k) + I_{\mathcal{O}_k}(x_k) = \tilde{F}_{\mathcal{B}_t}(x_k) \\ \Phi(x_{k+1}) &= \tilde{F}_{\mathcal{B}_k}(x_{k+1}) + I_{\mathcal{O}_k}(x_{k+1}) = \tilde{F}_{\mathcal{B}_k}(x_{k+1}). \end{aligned} \tag{51}$$

Similar to the proof of Lemma 1, we have

$$\begin{aligned} &\Phi(x_{k+1}) = \Phi(x_k + \alpha_k d_k) \\ =&\tilde{F}_{\mathcal{B}_k}(x_k + \alpha_k d_k) + I_{\mathcal{O}_k}(x_k + \alpha_k d_k) \\ \leq&\tilde{F}_{\mathcal{B}_k}(x_k) + \alpha_k \nabla \tilde{F}_{\mathcal{B}_k}(x_k)^T d_k + \frac{\alpha_k^2 L}{2} \|d_k\|_2^2 + I_{\mathcal{O}_k}(x_k) + \alpha_k v^T d_k \\ =&\Phi(x_k) + \frac{\alpha_k^2 L}{2} \|d_k\|_2^2 + \alpha_k \left(\nabla \tilde{F}_{\mathcal{B}_t}(x_k) + v\right)^T d_k \\ =&\Phi(x_k) + \frac{\alpha_k^2 L}{2} \|d_k\|_2^2 - \alpha_k \|d_k\|_2^2 \end{aligned} \tag{52}$$

where the last equality follows from (49). Therefore, we obtain

$$\Phi(x_{k+1}) \leq \Phi(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \|d_k\|_2^2. \tag{53}$$

Finally, it follows (9), (51) and (53) that

$$\begin{aligned} F_{\mathcal{B}_k}(x_{k+1}) = \tilde{F}_{\mathcal{B}_k}(x_{k+1}) = \Phi(x_{k+1}) \leq \Phi(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \|d_k\|_2^2 \\ = \tilde{F}_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \|d_k\|_2^2 = F_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \|d_k\|_2^2 \end{aligned}$$

which completes the proof.

According to Lemma 1 and Lemma 2, the objective value on a mini-batch tends to achieve a sufficient decrease in both Prox-SG Step and Orthant Step given α_k is small enough. By taking the expectation on both sides, we obtain the following result characterizing the sufficient decrease from $F(x_k)$ to $\mathbb{E}\left[F(x_{k+1})\right]$.

Corollary 2. *For iteration k, we have*

(i) if $k \in \mathcal{S}_{\mathcal{P}}$, then

$$\mathbb{E}\left[F(x_{k+1})\right] \leq F(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left[\|\mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k)\|_2^2\right]. \tag{54}$$

(ii) if $k \in \mathcal{S}_{\mathcal{O}}$, then

$$\mathbb{E}\left[F(x_{k+1})\right] \leq F(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left[\|d_k\|_2^2\right]. \tag{55}$$

Corollary 2 shows that the bound of F depends on step size α_k and norm of search direction. It further indicates that both Orthant Step and Prox-SG Step can make some progress to optimality with proper selection of α_k.

B.2 Proof of Theorem 1 for convex settings

In order to establish our main convergence results for convex settings, we require the following two lemmas. The first one shows the continuity of the subdifferential for convex functions from [6, Theorem 3].

Lemma 3. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be convex and let $\{x_k\}$ converges to some $x^* \in \mathbb{R}^n$. Let $s_k \in \partial h(x_k)$ for all k. Then, the sequence $\{s_k\}$ is bounded and every of its limit points is a subgradient of h at x^*.*

The second lemma shows that the vectors $\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)$ and d_k correspond to a valid optimality measure for target problem.

Lemma 4. *Let $\mathcal{S}$ be an infinite set of positive integers such that $\{x_k\}_{k\in\mathcal{S}} \to z^*$. If one of the following cases satisfies:*

(i) $\{\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\}_{k\in\mathcal{S}_\mathcal{P}\bigcap\mathcal{S}} \to 0$. In other words, Prox-SG Step performs infinitely many times, and the proximal mapping converges to zero.
*(ii) $\{d_k\}_{k\in\mathcal{S}_\mathcal{O}\bigcap\mathcal{S}} \to 0$ and the optimal solution x^* lies in $\mathcal{O}_k$ for all $k \in \mathcal{S}_\mathcal{O}\bigcap\mathcal{S}$.
In other words, Prox-SG Step has explored orthant face inhabited by the optimal solution. Then Orthant Step runs infinitely many times on $\{\mathcal{O}_k\}_{k\in\mathcal{S}_\mathcal{O}\bigcap\mathcal{S}}$, and the projected mapping converges to zero.*

then the z^* is an optimal solution to problem (1).

Proof. Suppose case (i) holds. Then by Lemma 1, we have that for $k \in \mathcal{S}_\mathcal{P}\bigcap\mathcal{S}$

$$\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) = \nabla f(x_k) + \lambda v_{k+1},$$

where $v_{k+1} \in \partial\left\|x_{k+1}\right\|_1$. It follows the continuity of ∇f , $\{\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\}_{k\in\mathcal{S}_\mathcal{P}\bigcap\mathcal{S}} \to 0$ and $\{x_k\}_{k\in\mathcal{S}} \to z^*$ that there exists $v^* \in \mathbb{R}^n$ such that v^* is the unique limit point of $\{v_{k+1}\}_{k\in\mathcal{S}_\mathcal{P}\bigcap\mathcal{S}}$, namely

$$\{v_{k+1}\}_{k\in\mathcal{S}_\mathcal{P}\bigcap\mathcal{S}} \to v^*.$$

Combining with the convexity of $\left\|\cdot\right\|_1$, by Lemma 3, v^* belongs to sudifferential of $\left\|z^*\right\|_1$. Overall, we obtain

$$\nabla f(z^*) + \lambda v^* = 0$$

which means z^* is an optimal solution to problem (1).

Suppose case (ii) holds, then problem (10) shares the same solution with problem (1). Then using the similar analysis for case (i), we obtain that z^* is the optimal solution of (10) and (1).

Now we prove the first case of Theorem 1 in Section 3.1.

Proof of Theorem 1(i): We know that Algorithm 1 performs an infinite sequence of iterations. It follows Corollary 2 that for any $\ell \in \mathcal{S}_\mathcal{P}\bigcup\mathcal{S}_\mathcal{O}$,

$$\begin{aligned}
&\mathbb{E}F(x_0) - \mathbb{E}F(x_{\ell+1}) = \sum_{k=0}^{\ell}\mathbb{E}F(x_k) - \mathbb{E}F(x_{k+1}) \\
&\geq \sum_{\substack{k\in\mathcal{S}_\mathcal{P}\\ k\leq\ell}}\left(\alpha_k - \frac{\alpha_k^2 L}{2}\right)\mathbb{E}\left\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\right\|_2^2 + \sum_{\substack{k\in\mathcal{S}_\mathcal{O}\\ k\leq\ell}}\left(\alpha_k - \frac{\alpha_k^2 L}{2}\right)\mathbb{E}\left\|d_k\right\|_2^2
\end{aligned} \tag{56}$$

Combining the assumption that F is bounded below on the level set $\mathcal{L} := \{x \in \mathbb{R}^n : F(x) \leq F(x_0)\}$, $|\mathcal{S}_{\mathcal{P}}| = \infty$ and letting $\ell \to \infty$, we obtain

$$\sum_{k \in \mathcal{S}_{\mathcal{P}}} \left(\alpha_k - \frac{\alpha_k^2 L}{2} \right) \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 < \infty \tag{57}$$

Suppose case (16) holds, i.e. $0 < \alpha_k \equiv \alpha \leq \frac{1}{L}$, then

$$\frac{1}{2L} \sum_{k \in \mathcal{S}_{\mathcal{P}}} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 < \infty.$$

Consequently, we have

$$\lim_{k \in \mathcal{S}_{\mathcal{P}}} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 = 0 \ , \ \lim_{k \in \mathcal{S}_{\mathcal{P}}} \mathbb{E}[\mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k)] = 0 \tag{58}$$

By the compactness of level set $\mathcal{L}$, the infinite sequence $\{x_k\}$ has a subsequence that converges to a point in $\mathcal{L}$ in expectation. Given this fact, it follows from Lemma 4 and (58) that the limit point is one optimal solution x^* of (1). Now following the continuity of F, the monotonically decrease of F in expectation, we have that

$$\lim_{k \to \infty} \mathbb{E} \left[F(x_k) \right] = F(x^*). \tag{59}$$

If the uniqueness of x^* is given, we then have that

$$\lim_{k \to \infty} \mathbb{E} \left[x_k \right] = x^*. \tag{60}$$

Suppose case (17) holds, rewrite (79) as

$$\sum_{k \in \mathcal{S}_{\mathcal{P}}} \alpha_k \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 - \sum_{k \in \mathcal{S}_{\mathcal{P}}} \frac{\alpha_k^2 L}{2} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 < \infty \tag{61}$$

It follows Assumption 1, (17) and Lemma 3 that

$$\sum_{k \in \mathcal{S}_{\mathcal{P}}} \frac{\alpha_k^2 L}{2} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 < \infty, \tag{62}$$

which implies that

$$\sum_{k \in \mathcal{S}_{\mathcal{P}}} \alpha_k \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 < \infty. \tag{63}$$

combining with $\alpha_k > 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, we obtain

$$\liminf_{k \in \mathcal{S}_{\mathcal{P}}} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 = 0. \tag{64}$$

(64) indicates that there exists a subsequence $\mathcal{S}'$ in $\mathcal{S}_{\mathcal{P}}$ such that

$$\lim_{k \in \mathcal{S}'} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 = 0. \tag{65}$$

Similar to the proof of constant size case, we have that $\lim_{k\in\infty}\mathbb{E}[F(x_k)] = F(x_*)$, and $\lim_{k\in\infty}\mathbb{E}[x_k] = x_*$ if optimal solution is unique as claimed.

Next, we start to consider the case (ii) of Theorem 1. Toward that end, we at first show if x_k is sufficiently close to x^*, the optimal solution inhabits the orthant face $\mathcal{O}_k$ constructed by x_k.

Lemma 5. *If $\|x_k - x^*\|_2 \leq \delta$, then for each $i \in \mathcal{I}^{\neq 0}(x^*)$,*

$$\operatorname{sign}([x_k]_i) = \operatorname{sign}([x^*]_i). \tag{66}$$

Consequently $\mathcal{I}^0(x_k) \subseteq \mathcal{I}^0(x^*)$, and the optimal solution x^* lies in the Orthant face of x_k defined as (7)*, i.e., $x^* \in \mathcal{O}_k$.*

Proof. To obtain the conclusion, observe that since $\|x_k - x^*\|_2 \leq \delta$ by assumption, it follows from the definition of δ as (15) that

$$\begin{aligned}\text{for any } i \in \mathcal{I}^-(x^*) : [x_k]_i &= [x_k]_i - [x^*]_i + [x^*]_i \leq |[x_k - x^*]_i| - 2\delta\\ &\leq \|x_k - x^*\|_2 - 2\delta \leq \delta - 2\delta = -\delta < 0.\\ \text{for any } i \in \mathcal{I}^+(x^*) : [x_k]_i &= [x_k]_i - [x^*]_i + [x^*]_i \geq -|[x_k - x^*]_i| + 2\delta\\ &\geq -\|x_k - x^*\|_2 + 2\delta \leq -\delta + 2\delta = \delta > 0.\end{aligned} \tag{67}$$

Hence, we have that for each $i \in \mathcal{I}^{\neq 0}(x^*)$, $\operatorname{sign}([x_k]_i) = \operatorname{sign}([x^*]_i)$. By the definition $\mathcal{O}_k$, the x^* inhabits $\mathcal{O}_k$, namely $x^* \in \mathcal{O}_k$.

Once x_k is close enough to x^*, if the step size α_k is properly selected, then the yielded zero elements by employing one Orthant Step belongs to $\mathcal{I}^0(x^*)$ as stated in Lemma 6.

Lemma 6. *If $\|x_k - x^*\|_2 \leq \delta$, $k \in \mathcal{S}_{\mathcal{O}}$, and $\alpha_k \in (0, 2\delta/M)$, then we have that*

$$\mathcal{I}^0(x_{k+1}) \subseteq \mathcal{I}^0(x^*). \tag{68}$$

Proof. To prove it by contradiction, suppose there exists some $i \in \mathcal{I}^0(x_{k+1})$ such that $i \notin \mathcal{I}^0(x^*)$. Since $i \in \mathcal{I}^0(x_{k+1})$, $i \notin \mathcal{I}^0(x^*)$, and $\mathcal{I}^0(x_k) \subseteq \mathcal{I}^0(x^*)$ by Lemma 5, then $i \notin \mathcal{I}^0(x_k)$, consequently $\operatorname{sign}([x_{k+1}]_i) \neq \operatorname{sign}([x_{k1}]_i)$,

$$[x_{k+1}]_i[x_k]_i = [x_k - \alpha_k \nabla\tilde{F}_{\mathcal{B}_k}(x_k)]_i[x_k]_i \leq 0 \tag{69}$$

On the other hand, combining (69) with (15) and the assumption $\alpha_k \in (0, 2\delta/M)$, we have that

$$\begin{aligned}[x_{k+1}]_i[x_k]_i &= [x_k - \alpha_k \nabla\tilde{F}_{\mathcal{B}_k}(x_k)]_i[x_k]_i\\ &= \|[x_k]_i\|^2 - \alpha_k \nabla[\tilde{F}_{\mathcal{B}_k}(x_k)]_i[x_k]_i\\ &\geq 4\delta^2 - \alpha_k|[\nabla\tilde{F}_{\mathcal{B}_k}(x_k)]_i| \cdot |[x_k]_i|\\ &\geq 4\delta^2 - \alpha_k \left\|\nabla\tilde{F}_{\mathcal{B}_k}(x_k)\right\|_2 |[x_k]_i|\\ &\geq 4\delta^2 - \alpha_k M 2\delta > 0,\end{aligned} \tag{70}$$

contradicting (69) which completes the proof.

We then reveal that the Euclidean distance between next iterate x_{k+1} computed by Orthant Step and the optimal solution x^* does not increase under sufficiently small step size α_k in Lemma 7.

Lemma 7. *If* $\|x_k - x^*\|_2 \leq \delta$, $k \in \mathcal{S}_{\mathcal{O}}$, *and* $\alpha_k \in (0, \min\{2/L, 2\delta/M\})$, *then*

$$\|x_{k+1} - x^*\|_2 \leq \delta. \tag{71}$$

Proof. For simplicity, let $\mathcal{I}^0_{k+1}$ as $\mathcal{I}^0(x_{k+1})$ and $\mathcal{I}^{\neq 0}_{k+1}$ as $\mathcal{I}^{\neq 0}(x_{k+1})$.

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \left\|\left[x_k - \alpha_k \nabla \tilde{F}_{\mathcal{B}_k}(x_k) - x^*\right]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2 + \left\|[x^*]_{\mathcal{I}^0_{k+1}}\right\|_2^2 \\ &= \left\|[x_k - x^*]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2 - 2\alpha_k \left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k)\right]^T_{\mathcal{I}^{\neq 0}_{k+1}} [x_k - x^*]_{\mathcal{I}^{\neq 0}_{k+1}} \\ &\quad + \alpha_k^2 \left\|\left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k)\right]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2 + \left\|[x^*]_{\mathcal{I}^0_{k+1}}\right\|_2^2\end{aligned} \tag{72}$$

It follows the convexity of $F_{\mathcal{B}_k}(x)$ and its Lipschitz continuous gradient that

$$\left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k) - \nabla \tilde{F}_{\mathcal{B}_k}(x^*)\right]^T_{\mathcal{I}^{\neq 0}_{k+1}} [x_k - x^*]_{\mathcal{I}^{\neq 0}_{k+1}} \geq \frac{1}{L} \left\|\left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k) - \nabla \tilde{F}_{\mathcal{B}_k}(x^*)\right]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2. \tag{73}$$

Combining with the optimal condition and $x^* \in \mathcal{O}_k$, (73) can be rewritten as

$$\left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k)\right]^T_{\mathcal{I}^{\neq 0}_{k+1}} [x_k - x^*]_{\mathcal{I}^{\neq 0}_{k+1}} \geq \frac{1}{L} \left\|\left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k)\right]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2. \tag{74}$$

Additionally, it follows the assumption of this lemma and Lemma 6 that

$$\left\|[x^*]_{\mathcal{I}^0_{k+1}}\right\|_2^2 = 0. \tag{75}$$

By the above (74) and (75), (72) can be further simplified as

$$\|x_{k+1} - x^*\|_2^2 \leq \left\|[x_k - x^*]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2 - \left(\frac{2\alpha_k}{L} - \alpha_k^2\right) \left\|\left[\nabla \tilde{F}_{\mathcal{B}_k}(x_k)\right]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2. \tag{76}$$

Now it follows $0 < \alpha_k < 2/L$ that

$$\|x_{k+1} - x^*\|_2^2 \leq \left\|[x_k - x^*]_{\mathcal{I}^{\neq 0}_{k+1}}\right\|_2^2 \leq \|x_k - x^*\|_2^2 = \delta^2 \tag{77}$$

which completes the proof.

The Lemma 8 below shows if current iterate x_k locates closely enough to x^* and step size α_k is properly selected, then x^* inhabits all subsequently Orthant faces.

Lemma 8. *If* $\|x_K - x^*\|_2 \leq \delta$, $\{k : k \geq K, k \in \mathbb{Z}^+\} \subseteq \mathcal{S}_{\mathcal{O}}$ *and* $\alpha_k \in (0, \min\{2/L, 2\delta/M\})$, *then* $x^* \in \mathcal{O}_k$ *for any* $k \geq K$.

Proof. It follows Lemma 5 and the assumption of this lemma that $x^* \in \mathcal{O}_K$. Combining with $\alpha_k \in (0, \min\{2/L, 2\delta/M\})$ and Lemma 7, that the assumption of Lemma 5 still holds for $K+1$, hence $x^* \in \mathcal{O}_{K+1}$. Therefore, we can iteratively employing Lemma 5 and 7 to show that $x^* \in \mathcal{O}_k$ holds for any $k \in \{k : k \geq K, k \in \mathbb{Z}^+\} \subseteq \mathcal{S}_{\mathcal{O}}$.

We now establish the proof for the second case of Theorem 1.

Proof of Theorem 1(ii): We know that Algorithm 1 performs an infinite sequence of iterations. It follows Corollary 2 that for any $\ell \in \mathcal{S}_{\mathcal{P}} \bigcup \mathcal{S}_{\mathcal{O}}$,

$$\begin{aligned} &\mathbb{E}F(x_0) - \mathbb{E}F(x_{\ell+1}) = \sum_{k=0}^{\ell} \mathbb{E}F(x_k) - \mathbb{E}F(x_{k+1}) \\ \geq &\sum_{\substack{k \in \mathcal{S}_{\mathcal{P}} \\ k \leq \ell}} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left\|\mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k)\right\|_2^2 + \sum_{\substack{k \in \mathcal{S}_{\mathcal{O}} \\ k \leq \ell}} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left\|d_k\right\|_2^2 \end{aligned} \tag{78}$$

Combining the assumption that F is bounded below on the level set $\mathcal{L} := \{x \in \mathbb{R}^n : F(x) \leq F(x_0)\}$, $|\mathcal{S}_{\mathcal{O}}| = \infty$ and letting $\ell \to \infty$, we obtain

$$\sum_{k \in \mathcal{S}_{\mathcal{O}}} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left\|d_k\right\|_2^2 < \infty \tag{79}$$

Similarly to the proof of Theorem 13 that under the step size setting (16) and (17), there exists a subsequence of $\mathcal{S} \in \mathcal{S}_{\mathcal{O}}$ such that

$$\lim_{k \in \mathcal{S}} \mathbb{E}\left\|d_k\right\|_2^2 = 0 \tag{80}$$

It follows $\alpha_k \in (0, \min\{2/L, 2\delta/M\})$, Lemma 8 and Lemma 4(ii) that $\lim_{k\to\infty} \mathbb{E}[F(x_k)] = F^*$ and $\lim_{k\to\infty} \mathbb{E}[x_k] = x^*$ if optimal solution is unique.

B.3 Proof of Theorem 2

We first show that the general PL condition (21) implies a different Proximal PL condition in [4], i.e., there exists a $\mu > 0$ such that

$$\mathcal{D}_{\lambda\|\cdot\|_1}(x, \eta) \geq 2\mu(F(x) - F^*) \tag{81}$$

where

$$\mathcal{D}_{\lambda\|\cdot\|_1}(x, \eta) = -2\eta \min_{y \in \mathbb{R}^n} \left\{\nabla f(x)^T(y - x) + \frac{\eta}{2}\left\|y - x\right\|_2^2 + \lambda\left\|y\right\|_1 - \lambda\left\|x\right\|_1\right\}. \tag{82}$$

Lemma 9. *If there exists a $\mu > 0$ such that for all $x \in \mathbb{R}^n$*

$$\|\mathcal{G}_\alpha(x)\|_2^2 \geq 2\mu(F(x) - F^*), \tag{83}$$

then for all $x \in \mathbb{R}^n$, the $\mathcal{D}_{\lambda\|\cdot\|_1}(x, 1/\alpha)$ satisfies

$$\mathcal{D}_{\lambda\|\cdot\|_1}(x, 1/\alpha) \geq 2\mu(F(x) - F^*). \tag{84}$$

Proof. Let $\hat{y} = \operatorname{argmin}_y \left\{\nabla f(x)^T(y - x) + \frac{1}{2\alpha}\|y - x\|_2^2 + \lambda\|y\|_1 - \lambda\|x\|_1\right\}$, then

$$\begin{aligned} 0 &\in \nabla f(x) + \frac{1}{\alpha}(\hat{y} - x) + \lambda\partial\|\hat{y}\|_1, \\ \hat{y} - x &\in -\alpha(\nabla f(x) + \lambda\partial\|\hat{y}\|_1). \end{aligned} \tag{85}$$

It follows the definition of $\mathcal{D}_{\lambda\|\cdot\|_1}(x, 1/\alpha)$ that

$$\begin{aligned} &\mathcal{D}_{\lambda\|\cdot\|_1}(x, 1/\alpha) \\ =&\frac{-2}{\alpha}\left\{\nabla f(x)^T(\hat{y} - x) + \frac{1}{2\alpha}\|\hat{y} - x\|_2^2 + \lambda\|\hat{y}\|_1 - \lambda\|x\|_1\right\} \\ \in&\frac{-2}{\alpha}\left\{-\alpha\nabla f(x)^T(\nabla f(x) + \lambda\partial\|\hat{y}\|_1) + \frac{1}{2\alpha}\alpha^2\|\nabla f(x) + \lambda\partial\|\hat{y}\|_1\|^2 + \lambda\|\hat{y}\|_1 - \lambda\|x\|_1\right\} \\ =&2\nabla f(x)^T(\nabla f(x) + \lambda\partial\|\hat{y}\|_1) + 2/\alpha(\lambda\|\hat{x}\|_1 - \lambda\|y\|_1) - \|\nabla f(x) + \lambda\partial\|\hat{y}\|_1\|^2 \\ \geq&2\nabla f(x)^T(\nabla f(x) + \lambda\partial\|\hat{y}\|_1) + 2/\alpha\lambda\partial\|\hat{y}\|_1(x - \hat{y}) - \|\nabla f(x) + \lambda\partial\|\hat{y}\|_1\|^2 \\ =&2\nabla f(x)^T(\nabla f(x) + \lambda\partial\|\hat{y}\|_1) + \lambda\partial\|\hat{y}\|_1(\nabla f(x) + \lambda\partial\|\hat{y}\|_1) - \|\nabla f(x) + \lambda\partial\|\hat{y}\|_1\|^2 \\ =&2\|(\nabla f(x) + \lambda\partial\|\hat{y}\|_1)\|^2 - \|\nabla f(x) + \lambda\partial\|\hat{y}\|_1\|^2 \\ =&\|\nabla f(x) + \lambda\partial\|\hat{y}\|_1\|^2 \end{aligned} \tag{86}$$

On the other hand, the gradient mapping $\mathcal{G}_\alpha(x)$ exactly belongs to $\nabla f(x) + \lambda\partial\|\hat{y}\|_1$. Consequently, the following inequality holds

$$\mathcal{D}_{\lambda\|\cdot\|_1}(x, 1/\alpha) \geq \|\mathcal{G}_\alpha(x)\|_2^2 \geq 2\mu(F(x) - F^*) \tag{87}$$

for any $x \in \mathbb{R}^n$ by the assumption of this lemma, which completes the proof.

To distinguish these two different PL conditions, we refer the PL condition in (21) as G-PL condition and the one in (81) as D-PL condition.

The highlight idea of Theorem 2 is now presented as follows: if $f(x)$ is convex and satisfies PL condition like (21), when the step size α is sufficiently small, and the size of mini-batch is sufficiently large, there exists an upper bound $N_{\mathcal{P}}$ such that $\|x - x^*\|_2 \leq \delta$ can be achieved by employing $N_{\mathcal{P}}$ Prox-SG Steps with high probability.

Proof of Theorem 2: At first, since $F(x)$ satisfies the G-PL condition (21), it also satisfies D-PL condition due to Lemma 9. It then follows [4, Appendix G], specifically D-PL condition implies the Proximal Error Bound that there exists some $\frac{1}{2L} > \gamma > 0$ such that

$$\|x - x^*\|_2 \leq \gamma \|\mathcal{G}_\eta(x)\|_2 \tag{88}$$

holds for any $x \in \mathbb{R}^n$ and any $\eta > 0$.

For any $k \leq N_\mathcal{P}$, $k \in \mathcal{S}_\mathcal{P}$, at k-th iteration, let $\mathcal{G}_{\alpha_k}(x)$ be the full gradient mapping at point x, let $\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x)$ be the mini-batch gradient mapping at point x with $\mathbb{E}_{\mathcal{B}_k}[\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x)|x] = \mathcal{G}_{\alpha_k}(x)$, and let $e_k(x)$ be the difference between full gradient mapping and mini-batch gradient mapping such that

$$\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x) = \mathcal{G}_{\alpha_k}(x) + e_k, \tag{89}$$

with $\mathbb{E}_{\mathcal{B}_k}[e_k|x_k] = 0$ where x_k denotes the starting point at k-th iteration. Notice that condition on x_k, $\mathcal{G}_{\alpha_k}(x_k)$ is independent with e_k.

Based on Lemma 1, given x_k and a random sampled mini-batch $\mathcal{B}_k$, the expected Euclidean distance square between next iterate x_{k+1} and the solution x^* given x_k can be computed as follows

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_k}[\|x_{k+1} - x^*\|_2^2 \,|x_k]\\
=&\mathbb{E}_{\mathcal{B}_k}[\|x_k - \alpha_k \mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k) - x^*\|_2^2 \,|x_k]\\
=&\mathbb{E}_{\mathcal{B}_k}[\|x_k - x^*\|_2^2 \,|x_k] - 2\alpha_k (x_k - x^*)^T \mathbb{E}_{\mathcal{B}_k}[\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)|x_k] + \alpha_k^2 \mathbb{E}_{\mathcal{B}_k}[\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 \,|x_k]\\
=&\|x_k - x^*\|_2^2 - 2\alpha_k (x_k - x^*)^T \mathcal{G}_{\alpha_k}(x_k) + \alpha_k^2 \{\|\mathbb{E}_{\mathcal{B}_k}[\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)|x_k]\|^2 + \mathbb{E}_{\mathcal{B}_k}[\|e_k(x_k)\|^2 \,|x_k]\}\\
=&\|x_k - x^*\|^2 - 2\alpha_k (x_k - x^*)^T \mathcal{G}_{\alpha_k}(x_k) + \alpha_k^2 \{\|\mathcal{G}_{\alpha_k}(x_k)\|^2 + \mathbb{E}_{\mathcal{B}_k}[\|e_k(x_k)\|^2 \,|x_k]\\
=&\|x_k - \alpha_k \mathcal{G}_{\alpha_k}(x_k) - x^*\|_2^2 + \alpha_k^2 \mathbb{E}_{\mathcal{B}_k}[\|e_k(x_k)\|^2 \,|x_k]
\end{aligned} \tag{90}$$

where the first term $\|x_k - \alpha_k \mathcal{G}_{\alpha_k}(x_k) - x^*\|_2^2$ is the distance square obtained via starting at x_k followed by doing a proximal *full gradient descent* step, and the second term $\alpha_k^2 \mathbb{E}_{\mathcal{B}_k}[\|e_k(x_k)\|^2 \,|x_k]$ is the random noise generated from the kth mini-batch stochastic gradient descent step combining with step size α_k.

To upper bound the first term, notice that for a proximal full gradient descent, it follows Proximal Error Bound (88), $\alpha_k \in (0, 1/L]$ and [3, Theorem 3.2] that

$$\|x_k - \alpha_k \mathcal{G}_{\alpha_k}(x_k) - x^*\|_2^2 \leq \left(1 - \frac{1}{2L\gamma}\right) \hat{C}(F(x_k) - F^*) \tag{91}$$

where $\hat{C}$ is a constant as $\frac{2}{L(1-\sqrt{1-(2L\gamma)^{-1}})^2}$. Based on [4, Theorem 4], since F has L-Lipschitz continuous gradient and satisfies G-PL condition, if we use a constant step size $\alpha_k \equiv \alpha < \frac{1}{2\mu}$, then we obtain a linear convergence rate up to a solution level that is proportional to α,

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq (1 - 2\mu\alpha)^k \left(F(x_0) - F^*\right) + \frac{LD^2\alpha}{4\mu}, \tag{92}$$

where D is the bound of norm of gradient mapping estimation which is well defined by Assumption 1 and (1).

To upper bound the second term, since the norm of gradient mapping is bounded, let $\mathcal{B}_i$ be a one-point mini-batch, then for any x, there exists a constant $\sigma > 0$ such that

$$\sigma^2 \geq \mathbb{E}_{\mathcal{B}_i \sim \text{Unif}[n]} \left[\left(\mathcal{G}_{\alpha_k, \mathcal{B}_k}(x) - \mathcal{G}_{\alpha_k}(x) \right)^2 \right] \tag{93}$$

i.e., σ^2 is an upper bound for the noise generated from a one-point mini-batch stochastic proximal gradient descent step. By computational, we have

$$\mathbb{E}_{\mathcal{B}_k} \left[\|e_k\|^2 \,\middle|\, y_k \right] \leq \frac{\sigma^2}{|\mathcal{B}_k|}, \tag{94}$$

which gives an upper bound propotion to $\frac{1}{|\mathcal{B}_k|}$.

Therefore, combining (90), (92), and (94)

$$\begin{aligned}
&\mathbb{E}[\|x_{k+1} - x^*\|_2^2] \\
=&\mathbb{E}[\|x_k - \alpha_k \mathcal{G}_{\alpha_k}(x_k) - x^*\|_2^2] + \alpha_k^2 \mathbb{E}_{\mathcal{B}_k}[\|e_k(x_k)\|^2 \,|x_k] \\
\leq& \left(1 - \frac{1}{2L\gamma}\right) \hat{C} \left[(1 - 2\mu\alpha)^k \left(F(x_0) - F^*\right) + \frac{LD^2\alpha}{4\mu} \right] + \frac{\sigma^2}{|\mathcal{B}_k|}.
\end{aligned} \tag{95}$$

Now for any $1 > \tau > 0$, if the step size α is sufficient small and satisfies

$$\alpha < \frac{8\gamma\mu\tau\delta^2}{(2L\gamma - 1)\hat{C}D^2}, \tag{96}$$

then

$$8\gamma\mu\tau\delta^2 - (2L\gamma - 1)\hat{C}D^2\alpha > 0 \tag{97}$$

Moreover, if mini-batch size is sufficiently large and satisfies

$$|\mathcal{B}_k| > \frac{8\gamma\mu\sigma^2}{8\gamma\mu\tau\delta^2 - (2L\gamma - 1)\hat{C}D^2\alpha} \tag{98}$$

then

$$\tau\delta^2 - \frac{\sigma^2}{|\mathcal{B}_k|} - \left(1 - \frac{1}{2L\gamma}\right) \hat{C} \frac{LD^2\alpha}{4\mu} > 0. \tag{99}$$

Thus, there exist some well-defined $k \geq 0$ such that

$$\left(1 - \frac{1}{2L\gamma}\right) \hat{C} (1 - 2\mu\alpha)^k \left(F(x_0) - F^*\right) \leq \tau\delta^2 - \frac{\sigma^2}{|\mathcal{B}_k|} - \left(1 - \frac{1}{2L\gamma}\right) \hat{C} \frac{LD^2\alpha}{4\mu} \tag{100}$$

Notice that the right hand side of (100) is a polynomial of $\tau\delta^2, 1/|\mathcal{B}_k|$ and α, and $\left(1 - \frac{1}{2L\gamma}\right) \hat{C}$ on the left hand side of (100) is a constant given F. Thus to let (100) hold, k should satisfy

$$k \geq K := \left\lceil \frac{\log\left(\text{poly}(\tau\delta^2, 1/|\mathcal{B}_k|, \alpha)/(F(x_0) - F^*)\right)}{\log\left(1 - 2\mu\alpha\right)} \right\rceil \tag{101}$$

where poly$(\tau\delta^2, 1/|\mathcal{B}_k|, \alpha)$ represents a polynomial of $\tau\delta^2, 1/|\mathcal{B}_k|$ and α.

Now, it follows (95) that if (96), (98) and (101) hold, then

$$\mathbb{E}[\|x_{k+1} - x^*\|_2^2] \leq \tau\delta^2, \tag{102}$$

now combine with Markov inequality that

$$\mathbb{P}\left(\|x_{k+1} - x^*\|_2^2 \geq \delta^2\right) \leq \frac{\mathbb{E}[\|x_{k+1} - x^*\|_2^2]}{\delta^2} \leq \tau. \tag{103}$$

which indicates the event $\|x_{k+1} - x^*\|_2^2 \leq \delta^2$ holds with probability at least $1 - \tau$ for any $k \geq K$.

B.4 Proof of Theorem 3 for nonconvex settings

In this Appendix, we present the proofs of the convergence theorem for nonconvex settings.

Proof of Theorem 3(i): Similar to proof of Theorem 1(i), we have that

$$\sum_{k\in\mathcal{S}_\mathcal{P}} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 < \infty \tag{104}$$

Suppose case (16) holds, i.e. $0 < \alpha_k \equiv \alpha \leq \frac{1}{L}$, then

$$\frac{1}{2L} \sum_{k\in\mathcal{S}_\mathcal{P}} \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 < \infty.$$

Consequently, we have

$$\lim_{k\in\mathcal{S}_\mathcal{P}} \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 = 0\ ,\ \lim_{k\in\mathcal{S}_\mathcal{P}} \mathbb{E}[\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)] = 0 \tag{105}$$

Suppose case (17) holds, rewrite (79) as

$$\sum_{k\in\mathcal{S}_\mathcal{P}} \alpha_k \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 - \sum_{k\in\mathcal{S}_\mathcal{P}} \frac{\alpha_k^2 L}{2} \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 < \infty \tag{106}$$

It follows Assumption 1, (17) and Lemma 3 that

$$\sum_{k\in\mathcal{S}_\mathcal{P}} \frac{\alpha_k^2 L}{2} \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 < \infty, \tag{107}$$

which implies that

$$\sum_{k\in\mathcal{S}_\mathcal{P}} \alpha_k \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 < \infty. \tag{108}$$

combining with $\alpha_k > 0$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, we obtain

$$\liminf_{k\in\mathcal{S}_\mathcal{P}} \mathbb{E}\,\|\mathcal{G}_{\alpha_k,\mathcal{B}_k}(x_k)\|_2^2 = 0, \tag{109}$$

which completes the proof.

Now we establish the proof of Theorem 3(ii) for OBProx-SG+. Remark here that although $f(x)$ is nonconvex $\mathbb{R}^n$, we assume $f(x)$ is convex on a subset $\mathcal{X} := \{x : \|x - \hat{x}\|_2 \leq \delta\}$ for some stationary point $\hat{x}$.

Proof of Theorem 3(ii): Similarly to proof of Theorem 1(ii), we have that under the step size setting (16) and (17), there exists a subsequence of $\mathcal{S} \in \mathcal{S}_{\mathcal{O}}$ such that

$$\lim_{k \in \mathcal{S}} \mathbb{E} \left\| \nabla \tilde{F}_{\mathcal{B}_k}(x_k) \right\|_2^2 = 0 \tag{110}$$

At first, it follows $f(x)$ is convex on $\mathcal{X} := \{x : \|x - \hat{x}\|_2 \leq \delta\}$ that Lemma, 5, 6, 7, 8 are still applicable for $\hat{x}$. Hence, combining with $\alpha_k \leq \min\{2/L, 2\delta/M\}$, the stationary point $\hat{x}$ of problem (1) is also one stationary point of subproblem (8) for any $k \in \mathcal{S}_{\mathcal{O}}$. Therefore, by proof on contradiction, (110) indicates that for the subsequence of $\mathcal{S} \in \mathcal{S}_{\mathcal{O}}$

$$\lim_{k \in \mathcal{S}} \mathbb{E} \left\| \mathcal{G}_{\alpha_k, \mathcal{B}_k}(x_k) \right\|_2^2 = 0. \tag{111}$$

C Switching Mechanism Comparison

In this section, we dive into the performance of OBProx-SG under different switching mechanisms to numerically demonstrate the superiority of the control mechanism under $N_{\mathcal{P}}$ and $N_{\mathcal{O}}$ presented in the main body of this paper.

As a competitor, we design another switching mechanism stated as Algorithm 5, by making use of the optimality measure inspired by the multi-routine deterministic optimization algorithms [2,5]. Particularly, at kth iteration, we at first compute a minimum-norm subgradient $g(x)$ defined as follows

$$[g(x)]_i = \begin{cases} [\nabla f(x)]_i + \lambda & \text{if } [x]_i > 0 \text{ or } ([x]_i = 0 \text{ and } [\nabla f(x)]_i + \lambda < 0) \\ [\nabla f(x)]_i - \lambda & \text{if } [x]_i < 0 \text{ or } ([x]_i = 0 \text{ and } [\nabla f(x)]_i - \lambda > 0) \\ 0 & \text{otherwise} \end{cases} \tag{112}$$

on x_k, or its estimator on a subset of full data points $\hat{\mathcal{B}}$, see line 2 in Algorithm 5. Then we compute the norm of subvector in $g_{\hat{\mathcal{B}}}(x_k)$ corresponding to the indices of zero entries on x_k, and the norm of subvector in $g_{\hat{\mathcal{B}}}(x_k)$ for the non-zero entries. If $\left\| [g_{\hat{\mathcal{B}}}(x_k)]_{\mathcal{I}^0(x_k)} \right\|_2 \geq \left\| [g_{\hat{\mathcal{B}}}(x_k)]_{\mathcal{I}^{\neq 0}(x_k)} \right\|_2$, then the progress by freeing zero variables on x_k to non-zero may produce more progress to the optimality. Since the Prox-SG Step mainly serves as predicting the supports (non-zero entries) of the solution, then employing Prox-SG Step at current iteration is a reasonable choice. Otherwise, we select Orthant Step to promote the sparsity.

Next, we test OBProx-SG under the switching mechanism as Algorithm 5 on the convex experiments in Section 4.1, where at each iteration $\hat{\mathcal{B}}$ is constructed by uniformly sampling 5% data points. The numerical results are provided in Table 7 for

Algorithm 5 Switching Mechanism by Optimality Measure.

```
Input: $k, x_k, \hat{\mathcal{B}}$.
Compute the minimum-norm subgradient (112) on $\hat{\mathcal{B}}$, denoted as $g_{\hat{\mathcal{B}}}(x_k)$.
if $\left\|[g_{\hat{\mathcal{B}}}(x_k)]_{\mathcal{I}^0(x_k)}\right\|_2 \geq \left\|[g_{\hat{\mathcal{B}}}(x_k)]_{\mathcal{I}^{\neq 0}(x_k)}\right\|_2$ then
    Return Prox-SG step is selected.
else
    Return Orthant step is selected.
```

final objective function values (F/f) and Table 8 for density of solutions. We observe that OBProx-SG under different switching mechanisms can achieve quite competitive objective function values F/f on these convex problems. However, it is apparent that OBProx-SG under switching mechanism by optimality measure computes solutions with obviously lower sparsity (higer density) comparing with OBProx-SG under switching mechanism by $N_{\mathcal{P}}$ and $N_{\mathcal{O}}$. It is because the randomness of $\hat{\mathcal{B}}$ may not guarantee the OBProx-SG ends with Orthant Step but Prox-SG Step which is highly likely to deteriorate the progress of sparsity exploration. Therefore due to additionally computational cost of Algorithm 5 and the unreliability of sparsity promotion, we recommend to use Algorithm 4 as the default switch.

Table 7: Objective function values F/f on convex problems.

Problems	Switch		
	Algorithm 4	Algorithm 4	Algorithm 5
	$N_{\mathcal{P}} = 5, N_{\mathcal{O}} = 5$	$N_{\mathcal{P}} = 15, N_{\mathcal{O}} = \infty$	
a9a	**0.327 / 0.326**	0.329 / 0.328	0.331 / 0.329
higgs	**0.326 / 0.326**	**0.326 / 0.326**	**0.326 / 0.326**
ijcnn1	**0.198 / 0.197**	**0.198 / 0.197**	0.199 / 0.198
kdda	**0.102 / 0.102**	**0.102 / 0.102**	**0.102 / 0.102**
news20	**0.413 / 0.355**	**0.413 / 0.355**	**0.413 / 0.355**
real-sim	**0.164 / 0.125**	**0.164 / 0.125**	0.165 / 0.126
rcv1	**0.242 / 0.179**	**0.242 / 0.179**	**0.242 / 0.179**
susy	**0.376 / 0.376**	**0.376 / 0.376**	**0.376 / 0.376**
url_combined	0.050 / 0.049	**0.047 / 0.046**	0.090 / 0.090
w8a	**0.052 / 0.048**	**0.052 / 0.048**	**0.052 / 0.048**
best	**9**	**9**	6
second best	**1**	**1**	0
sum	**10**	**10**	6

Table 8: Density of solutions on convex problems.

Problems	Switch Algorithm 4 $N_{\mathcal{P}} = 5, N_{\mathcal{O}} = 5$	Switch Algorithm 4 $N_{\mathcal{P}} = 15, N_{\mathcal{O}} = \infty$	Algorithm 5
a9a	62.10	**59.68**	**59.68**
higgs	**70.69**	**70.69**	89.66
ijcnn1	**56.52**	**56.52**	**56.52**
kdda	0.08	**0.06**	0.34
news20	0.20	**0.19**	2.22
real-sim	22.44	22.15	**21.52**
rcv1	4.36	**4.33**	10.42
susy	**73.68**	**73.68**	94.74
url_combined	3.26	**3.00**	4.91
w8a	78.03	74.75	**71.10**
best	3	**8**	4
second best	**4**	2	0
sum	7	**10**	4

References

1. Bertsekas, D.P.: Convex optimization theory. Athena Scientific Belmont (2009)
2. Chen, T., Curtis, F.E., Robinson, D.P.: A reduced-space algorithm for minimizing ℓ_1-regularized convex functions. SIAM Journal on Optimization **27**(3), 1583–1610 (2017)
3. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Mathematics of Operations Research (2018)
4. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer (2016)
5. Keskar, N.S., Nocedal, J., Oztoprak, F., Waechter, A.: A second-order method for convex ℓ_1-regularized optimization with active set prediction. arXiv preprint arXiv:1505.04315 (2015)
6. Nedich, A.: Ie 598 an, lecture 18: Subdifferential properties (Fall 2008)
7. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. Journal of Machine Learning Research **11**(Oct), 2543–2596 (2010)
8. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization **24**(4), 2057–2075 (2014)