SUPPLEMENTARY MATERIAL: A HALF-SPACE STOCHASTIC PROJECTED GRADIENT METHOD FOR GROUP SPARSITY REGULARIZATION

Anonymous authors
Paper under double-blind review

As a summary, we organize the supplementary material as follows. We first restate basic notations and definitions in Section 1, then move on to restate the HSPG method in Section 2. Next we restate convergence analysis theorems with corresponding proofs in Appendix C. Finally, we present additional experimental setting and numerical results in Appendix D. For ease of reference, we highlight the referenced points appeared in the main body of paper as below:

- The sufficient decrease of Half-Space Step (Lemma 1 in the main body) is restated as Lemma 1 with detailed proof in Appendix C.1.
- The projection region of Half-Space Step is presented in Appendix A.
- The Non-Lipschitz continuity of Ψ around the origin point is revealed in Appendix B.
- Proofs of Theorem 1 and 2 are shown in Appendix C.2 and C.3 respectively.
- Proofs of (Proposition 1 in the main body) is reordered as Proposition 2 provided in Appendix C.4.
- Additional linear and logistic regression experiments are reported in Appendix D.1 and D.2.
- The procedure of fine tuning ϵ and final f comparison in non-convex experiments are described in Appendix D.3.

1 BASIC NOTATIONS AND DEFINITIONS

Consider the mixed ℓ_1/ℓ_2-regularized optimization problem (Group Lasso) in the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \left\{\Psi(x) \overset{\text{def}}{=} f(x)+\lambda\Omega(x)=\frac{1}{N}\sum_{i=1}^{N} f_i(x)+\lambda\sum_{g\in\mathcal{G}}\|[x]_g\|\right\}, \tag{1}$$

where $\lambda > 0$ is a weighting factor, $\|\cdot\|$ denotes ℓ_2-norm, $f(x)$ is the average of numerous N continuously differentiable instance functions $f_i : \mathbb{R}^n \to \mathbb{R}$.

Since that fundamental to Half-Space Step is the manner in which we handle the zero and non-zero group of variables, we define the following index sets for any $x \in \mathbb{R}^n$:

$$\mathcal{I}^0(x) := \{g : g\in\mathcal{G}, [x]_g = 0\} \text{ and } \mathcal{I}^{\neq 0}(x) := \{g : g\in\mathcal{G}, [x]_g \neq 0\}, \tag{2}$$

where $\mathcal{I}^0(x)$ represents the indices of groups of zero variables at x, and $\mathcal{I}^{\neq 0}(x)$ indexes the groups of nonzero variables at x.

To proceed, we further define an artificial set that x lies in:

$$\mathcal{S}(x) := \left\{z\in\mathbb{R}^n : [z]_g^\top [x]_g \geq \epsilon\,\|[x]_g\|^2 \text{ if } g\in\mathcal{I}^{\neq 0}(x), \text{ and } [z]_g = 0 \text{ if } g\in\mathcal{I}^0(x)\right\}\bigcup\{0\}, \tag{3}$$

which consists of half-spaces and the origin. Hence, x inhabits $\mathcal{S}(x_k)$, *i.e.*, $x\in\mathcal{S}(x_k)$, only if: (i) $[x]_g$ lies in the upper half-space for all $g\in\mathcal{I}^{\neq 0}(x_k)$ for some prescribed $\epsilon\in[0,1)$; and (ii) $[x]_g$ equals to zero for all $g\in\mathcal{I}^0(x_k)$.

In HSPG, Half-Space Step involves minimizing $\Psi(x)$ over $\mathcal{S}_k$ as follows:

$$x_{k+1} = \underset{x\in\mathcal{S}_k}{\arg\min}\ \Psi(x) = f(x)+\lambda\Omega(x). \tag{4}$$

A novel group projection operator is used to solve (4) and effectively promote group sparsity, so we need the following definition of the novel half-space group projection operator which projects a group of variables to zero if it falls outside of $\mathcal{S}_k$:

$$\left[\text{Proj}_{\mathcal{S}_k}(z)\right]_g := \begin{cases} [z]_g & \text{if } [z]_g^\top [x_k]_g > \epsilon \left\|[x_k]_g\right\|^2, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The above projector of form (5) is not the standard Euclidean projection operator in most cases, but still satisfies the following two advantages: (i) the progress to the optimum is made via sufficient decrease property; and (ii) effectively project groups of variables to zero simultaneously. In contrast, the Euclidean projection operator is far away effective to promote group sparsity, as shown in Figure 1 in the main body of this paper.

2 THE HSPG METHOD

In this section, we restate our main algorithm HSPG (Algorithm 1), and the subroutines to proceed a Prox-SG Step (Algorithm 2) and a Half-Space Step (Algorithm 3).

Algorithm 1 Outline of HSPG for solving (1).

```
Input: $x_0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$, $\epsilon \in [0,1)$, and $N_{\mathcal{P}} \in \mathbb{Z}^+$.
for $k = 0, 1, 2, \ldots$ do
    if $k < N_{\mathcal{P}}$ then
        Compute $x_{k+1} \leftarrow$ Prox-SG($x_k, \alpha_k$) by Algorithm 2.
    else
        Compute $x_{k+1} \leftarrow$ Half-Space($x_k, \alpha_k, \epsilon$) by Algorithm 3.
    Update $\alpha_{k+1}$.
```

Algorithm 2 Prox-SG Step.

```
Input: Current iterate $x_k$, and step size $\alpha_k$.
Compute the stochastic gradient of $f$ on mini-batch $\mathcal{B}_k$
    $$\nabla f_{\mathcal{B}_k}(x_k) \leftarrow \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla f_i(x_k). \tag{6}$$
Return $x_{k+1} \leftarrow \text{Prox}_{\alpha_k \lambda \Omega(\cdot)}\left(x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)\right)$.
```

Algorithm 3 Half-Space Step

```
Input: Current iterate $x_k$, step size $\alpha_k$, and $\epsilon$.
Compute the stochastic gradient of $\Psi$ on $\mathcal{I}^{\neq 0}(x_k)$ by mini-batch $\mathcal{B}_k$
    $$[\nabla \Psi_{\mathcal{B}_k}(x_k)]_{\mathcal{I}^{\neq 0}(x_k)} \leftarrow \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} [\nabla \Psi_i(x_k)]_{\mathcal{I}^{\neq 0}(x_k)} \tag{7}$$
Compute $[\tilde{x}_{k+1}]_{\mathcal{I}^{\neq 0}(x_k)} \leftarrow [x_k - \alpha_k \nabla \Psi_{\mathcal{B}_k}(x_k)]_{\mathcal{I}^{\neq 0}(x_k)}$ and $[\tilde{x}_{k+1}]_{\mathcal{I}^{0}(x_k)} \leftarrow 0$.
for each group $g$ in $\mathcal{I}^{\neq 0}(x_k)$ do
    if $[\tilde{x}_{k+1}]_g^\top [x_k]_g \leq \epsilon \left\|[x_k]_g\right\|^2$ then
        $[\tilde{x}_{k+1}]_g \leftarrow 0$.
Return $x_{k+1} \leftarrow \tilde{x}_{k+1}$.
```

A PROJECTION REGION

In this Appendix, we derive the projection region of HSPG, and reveal that is a superset of those of Prox-SG, Prox-SVRG and Prox-Spider under the same α_k and λ.

Proposition 1. *The Half-Space Step of HSPG 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 for each $g \in \mathcal{I}^{\neq 0}(x_k)$*

$$[x_{k+1}]_g = \begin{cases} [\hat{x}_{k+1}]_g - \alpha_k \lambda \frac{[x_k]_g}{\|[x_k]_g\|} & \textit{if } [\hat{x}_{k+1}]_g^\top [x_k]_g > (\alpha_k\lambda + \epsilon)\left\|[x_k]_g\right\| \\ 0 & \textit{otherwise.} \end{cases} \tag{8}$$

Consequently, if $\|[\hat{x}_{k+1}]_g\| \leq \alpha_k\lambda$, then $[x_{k+1}]_g = 0$ for any $\epsilon \geq 0$.

Proof. For $g \in \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^{\neq 0}(x_{k+1})$, by Algorithm 3, it is equivalent to

$$\begin{aligned} \left[x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k) - \alpha_k \lambda \frac{[x_k]_g}{\|[x_k]_g\|}\right]_g^\top [x_k]_g > \epsilon \left\|[x_k]_g\right\|^2, \\ [\hat{x}_{k+1}]_g^\top [x_k]_g - \alpha_k\lambda \left\|[x_k]_g\right\| > \epsilon \left\|[x_k]_g\right\|^2, \\ [\hat{x}_{k+1}]_g^\top [x_k]_g > (\alpha_k\lambda + \epsilon\left\|[x_k]_g\right\|)\left\|[x_k]_g\right\|. \end{aligned} \tag{9}$$

Similarly, $g \in \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^{0}(x_{k+1})$ is equivalent to

$$\begin{aligned} \left[x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k) - \alpha_k \lambda \frac{[x_k]_g}{\|[x_k]_g\|}\right]_g^\top [x_k]_g \leq \epsilon \left\|[x_k]_g\right\|^2, \\ [\hat{x}_{k+1}]_g^\top [x_k]_g - \alpha_k\lambda \left\|[x_k]_g\right\| \leq \epsilon \left\|[x_k]_g\right\|^2, \\ [\hat{x}_{k+1}]_g^\top [x_k]_g \leq (\alpha_k\lambda + \epsilon\left\|[x_k]_g\right\|)\left\|[x_k]_g\right\|. \end{aligned} \tag{10}$$

If $\|[\hat{x}_{k+1}]_g\| \leq \alpha_k\lambda$, then

$$[\hat{x}_{k+1}]_g^\top [x_k]_g \leq \left\|[\hat{x}_{k+1}]_g\right\| \left\|[x_k]_g\right\| \leq \alpha_k\lambda \left\|[x_k]_g\right\|. \tag{11}$$

Hence $[x_{k+1}]_g = 0$ holds for any $\epsilon \geq 0$ by (10), which implies that the projection region of Prox-SG and its variance reduction variants, *e.g.*, Prox-SVRG, Prox-Spider and SAGA are the subsets of HSPG's. □

B NON-LIPSCHITZ CONTINUITY OF $\nabla\Psi(x)$ ON $\mathbb{R}^n$

The first-derivative of $\Psi(x)$ at $x \neq 0$ can be written as

$$\nabla\Psi(x) = \nabla f(x) + \lambda \sum_{g\in\mathcal{G}} \frac{[x]_g}{\|[x]_g\|} \tag{12}$$

We next show $\frac{[x]_g}{\|[x]_g\|}$ is not Lipschitz continuous on $\mathbb{R}^n$ if $|g| \geq 2$. Take a example for $[x]_g = (x_1, x_2)^\top \in \mathbb{R}^2$, and select $x_1 = (t, a_1 t), x_2 = (t, a_2 t), a_1 \neq a_2$ and $t \in \mathbb{R}$. Then suppose there exists a positive constant $L < \infty$ such that Lipschitz continuity holds as follows

$$\begin{aligned} \left\| \frac{x_1}{\|x_1\|} - \frac{x_2}{\|x_2\|} \right\| &\leq L \left\| x_1 - x_2 \right\| \\ \left\| \frac{(1, a_1)}{\sqrt{1+a_1^2}} - \frac{(1, a_2)}{\sqrt{1+a_2^2}} \right\| &\leq L|a_1 - a_2| \cdot |t| \end{aligned} \tag{13}$$

holds for any $t \in \mathbb{R}$, and note the left hand side is a positive constant. However, letting $t \to 0$, we have that $L \to \infty$ which contradicts the $L < \infty$. Therefore, $\frac{[x]_g}{\|[x]_g\|}$ is not Lipschitz continuous on $\mathbb{R}^2$, specifically the region surrounding the origin point.

Although $[\nabla\Psi(x)]_{\mathcal{I}^{\neq 0}(x)}$ is not Lipscthiz continuous on $\mathbb{R}^n$, the Lipschitz continuity still holds on by excluding a fixed size ℓ_2-ball centered at the origin for the group of non-zero variables $\mathcal{I}^{\neq 0}(x)$ from $\mathbb{R}^n$. For our paper, we define the region where Lipscthiz continuity of $[\nabla\Psi(x)]_{\mathcal{I}^{\neq 0}(x)}$ still holds as

$$\mathcal{X} = \{x : \|[x]_g\| \geq \delta_1 \text{ for each } g \in \mathcal{I}^{\neq 0}(x), \text{ and } [x]_g = 0 \text{ for each } g \in \mathcal{I}^0(x)\}. \tag{14}$$

C CONVERGENCE ANALYSIS PROOF

Our analysis uses the following assumption that is assumed to hold throughout this section.

Assumption 1. *Each $f_i : \mathbb{R}^n \to \mathbb{R}$, for $i = 1, 2, \cdots, N$, is differentiable and bounded below. Their gradients $\nabla f_i(x)$ are Lipschitz continuous, and let L be the shared Lipschitz constant.*

Denote the sets of groups which are projected or not onto zero as

$$\hat{\mathcal{G}}_k := \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^0(x_{k+1}), \text{ and} \tag{15}$$

$$\tilde{\mathcal{G}}_k := \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^{\neq 0}(x_{k+1}). \tag{16}$$

Denote $\mathcal{X} := \{x : \|[x]_g\| \geq \delta_1 \text{ for each } g \in \mathcal{G}\}$ where the Lipschitz continuity of $\nabla \Psi_{\mathcal{B}}(x)$ still holds by excluding a ℓ_2-ball centered at the origin with radius δ_1 from $\mathbb{R}^n$. Let M denote one upper bound of $\|\partial \Psi\|$ and $\|\xi\|$.

Additionally, establishing some convergence results require the below constants to measure the least and largest magnitude of non-zero group variables in x^*,

$$0 < \delta_1 := \frac{1}{2} \min_{g \in \mathcal{I}^{\neq 0}(x^*)} \|[x^*]_g\|, \text{ and} \tag{17}$$

$$0 < \delta_2 := \frac{1}{2} \max_{g \in \mathcal{I}^{\neq 0}(x^*)} \|[x^*]_g\|. \tag{18}$$

and a subsequent results of strict complementary assumption on any $\mathcal{B}$ uniformly,

$$0 < \delta_3 := \frac{1}{2} \min_{g \in \mathcal{I}^0(x^*)} (\lambda - \|[\nabla f_{\mathcal{B}}(x^*)]_g\|) \tag{19}$$

And denote the following frequently used constant R describing the size of neighbor around x^*.

$$R := \min \left\{ \frac{-(\delta_1 + 2\epsilon\delta_2) + \sqrt{(\delta_1 + 2\epsilon\delta_2)^2 - 4\epsilon^2\delta_2 + 4\epsilon\delta_1^2}}{\epsilon}, \delta_1 \right\} > 0. \tag{20}$$

Remark: (20) is well defined as $0 < \epsilon < \frac{\delta_1^2}{\delta_2}$, and degenerated to δ_1 as $\epsilon = 0$.

C.1 SUFFICIENT DECREASE OF PROX-SG STEP AND HALF-SPACE STEP

Our convergence analysis relies on the following sufficient decrease properties of Half-Space Step and Prox-SG Step.

Sufficient Decrease of Half-Space Step: We restate Lemma 1 presented in the paper main body formally as the below Lemma 1 with corresponding proof as follows.

Lemma 1. *Suppose $x_k \in \mathcal{X}$ as (14). Algorithm 3 yields the next iterate x_{k+1} as $Proj_{\mathcal{S}_k}(x_k - \alpha_k \partial \Psi_{\mathcal{B}_k}(x_k))$ and the search direction $d_k := (x_{k+1} - x_k)/\alpha_k$, then*

- *(i) d_k is a descent direction for $\Psi_{\mathcal{B}_k}(x_k)$, i.e., $d_k^\top \partial \Psi_{\mathcal{B}_k}(x_k) < 0$; and*
- *(ii) the objective function value $\Psi_{\mathcal{B}_k}(x_{k+1})$ satisfies*

$$\Psi_{\mathcal{B}_k}(x_{k+1}) \leq \Psi_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \|[\partial \Psi_{\mathcal{B}_k}(x_k)]_g\|^2 - \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[x_k]_g\|^2. \tag{21}$$

Proof. It follows Algorithm 3 and the definition of $\tilde{\mathcal{G}}_k$ and $\hat{\mathcal{G}}_k$ as (16) and (15) that $x_{k+1} = x_k + \alpha_k d_k$ where d_k is

$$[d_k]_g = \begin{cases} -[\partial \Psi_{\mathcal{B}_k}(x_k)]_g & \text{if } g \in \tilde{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^{\neq 0}(x_{k+1}), \\ -[x_k]_g/\alpha_k & \text{if } g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^0(x_{k+1}), \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

We also notice that for any $g \in \hat{\mathcal{G}}_k$, the following holds

$$\begin{aligned}[x_k - \alpha_k \partial \Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g < \epsilon \left\|[x_k]_g\right\|^2, \\ (1-\epsilon)\left\|[x_k]_g\right\|^2 < \alpha_k [\partial \Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g.\end{aligned} \tag{23}$$

For simplicity, let $\mathcal{I}_k^{\neq 0} := \mathcal{I}^{\neq 0}(x_k)$. Since $[d_k]_g = 0$ for any $g \in \mathcal{I}^0(x_k)$, then by (22) and (23), we have

$$\begin{aligned} d_k^\top \partial \Psi_{\mathcal{B}_k}(x_k) &= [d_k]_{\mathcal{I}_k^{\neq 0}}^\top [\partial \Psi_{\mathcal{B}_k}(x_k)]_{\mathcal{I}_k^{\neq 0}} \\ &= -\sum_{g \in \tilde{\mathcal{G}}_k} \left\|[\partial \Psi_{\mathcal{B}_k}(x_k)]_g\right\|^2 - \sum_{g \in \hat{\mathcal{G}}_k} \frac{1}{\alpha_k} [x_k]_g^\top [\partial \Psi_{\mathcal{B}_k}(x_k)]_g \\ &\leq -\sum_{g \in \tilde{\mathcal{G}}_k} \left\|[\partial \Psi_{\mathcal{B}_k}(x_k)]_g\right\|^2 - \sum_{g \in \hat{\mathcal{G}}_k} \frac{1}{\alpha_k^2}(1-\epsilon)\left\|[x_k]_g\right\|^2 < 0, \end{aligned} \tag{24}$$

holds for any $\epsilon \in [0, 1)$, which implies that d_k is a descent direction for $\Psi_{\mathcal{B}_k}(x_k)$.

Now, we start to prove the suffcient decrease of Half-Space Step. By the descent lemma, $x_k \in \mathcal{X}$ and the Lipschitz continuity of $[\partial \Psi_{\mathcal{B}_k}]_{\mathcal{I}_k^{\neq 0}}$ on $\mathcal{X}$, we have that

$$\Psi_{\mathcal{B}_k}(x_k + \alpha_k d_k) \leq \Psi_{\mathcal{B}_k}(x_k) + \alpha_k [\partial \Psi_{\mathcal{B}_k}(x_k)]_{\mathcal{I}_k^{\neq 0}}^\top [d_k]_{\mathcal{I}_k^{\neq 0}} + \frac{L}{2}\alpha_k^2 \left\|[d_k]_{\mathcal{I}_k^{\neq 0}}\right\|^2. \tag{25}$$

Then it follows (22) that (25) can be rewritten as follows

$$\begin{aligned} &\Psi_{\mathcal{B}_k}(x_k + \alpha_k d_k) \\ \leq &\Psi_{\mathcal{B}_k}(x_k) + \alpha_k [\partial \Psi_{\mathcal{B}_k}(x_k)]_{\mathcal{I}_k^{\neq 0}}^\top [d_k]_{\mathcal{I}_k^{\neq 0}} + \frac{L}{2}\alpha_k^2 \left\|[d_k]_{\mathcal{I}_k^{\neq 0}}\right\|^2 \\ = &\Psi_{\mathcal{B}_k}(x_k) - \sum_{g \in \tilde{\mathcal{G}}_k} \left\|[\partial \Psi_{\mathcal{B}_k}(x_k)]_g\right\|^2 \left(\alpha_k - \frac{L}{2}\alpha_k^2\right) - \sum_{g \in \hat{\mathcal{G}}_k} \left\{[\partial \Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g - \frac{L}{2}\left\|[x_k]_g\right\|^2\right\} \end{aligned} \tag{26}$$

Consequently, combining with $\epsilon \in [0, 1)$ and (23), (26) can be further shown as

$$\Psi_{\mathcal{B}_k}(x_{k+1}) \leq \Psi_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \left\|[\partial \Psi_{\mathcal{B}_k}(x_k)]_g\right\|^2 - \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \left\|[x_k]_g\right\|^2, \tag{27}$$

which completes the proof.

□

Sufficient Decrease of Prox-SG Step: The second lemma is well known for proximal operator under our notations. We include this proof for completeness.

Lemma 2. *Line 3 of Algorithm 2 yields that* $x_{k+1} = x_k - \alpha_k \xi_{\alpha_k, \mathcal{B}_k}(x_k)$, *where*

$$\xi_{\alpha_k, \mathcal{B}_k}(x_k) \in -\left(\nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \Omega(x_{k+1})\right). \tag{28}$$

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

$$\Psi_{\mathcal{B}_k}(x_{k+1}) \leq \Psi_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \left\|\xi_{\alpha_k, \mathcal{B}_k}(x_k)\right\|^2. \tag{29}$$

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

$$\begin{aligned} x_{k+1} &= \underset{x \in \mathbb{R}^n}{\arg\min}\ \frac{1}{2\alpha_k}\left\|x - (x_k - \alpha_k \nabla f_{\mathcal{B}_k}(x_k))\right\|^2 + \lambda \Omega(x) \\ &= \underset{x \in \mathbb{R}^n}{\arg\min}\ \nabla f_{\mathcal{B}_k}(x_k)^\top (x - x_k) + \lambda \Omega(x) + \frac{1}{2\alpha_k}\left\|x - x_k\right\|^2 \end{aligned} \tag{30}$$

By the optimal condition, we have

$$0 \in \frac{1}{\alpha_k}(x_{k+1} - x_k) + \nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \Omega(x_{k+1}). \tag{31}$$

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

$$0 \in -\xi_{\alpha_k, \mathcal{B}_k}(x_k) + \nabla f_{\mathcal{B}_k}(x_k) + \lambda \partial \Omega(x_{k+1}), \tag{32}$$

which implies that

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

And thus there exists some $v \in \partial \Omega(x_{k+1})$ such that

$$\xi_{\alpha_k, \mathcal{B}_k}(x_k) = \nabla f_{\mathcal{B}_k}(x_k) + \lambda v. \tag{34}$$

By Lipschitz continuity of $\nabla f_{\mathcal{B}_k}$ and convexity of $\Omega(\cdot)$, we have

$$\begin{aligned} f_{\mathcal{B}_k}(x_{k+1}) &= f_{\mathcal{B}_k}(x_k - \alpha_k \xi_{\alpha_k, \mathcal{B}_k}(x_k)) \\ &\leq f_{\mathcal{B}_k}(x_k) - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)^\top \xi_{\alpha_k, \mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2} \left\| \xi_{\alpha_k, \mathcal{B}_k}(x_k) \right\|^2 \end{aligned} \tag{35}$$

and

$$\begin{aligned} \lambda \Omega(x_{k+1}) &= \lambda \Omega(x_k - \alpha_k \xi_{\alpha_k, \mathcal{B}_k}(x_k)) \\ &\leq \lambda \Omega(x_k) + \lambda v^\top (x_k - \alpha_k \xi_{\alpha_k, \mathcal{B}_k}(x_k) - x_k) \\ &= \lambda \Omega(x_k) - \alpha_k \lambda v^\top \xi_{\alpha_k, \mathcal{B}_k}(x_k). \end{aligned} \tag{36}$$

Hence, by (34), (35) and (36), the objective $\Psi_{\mathcal{B}_k}(x_{k+1})$ satisfies

$$\begin{aligned} &\Psi_{\mathcal{B}_k}(x_{k+1}) = f_{\mathcal{B}_k}(x_{k+1}) + \lambda \Omega(x_{k+1}) \\ \leq & f_{\mathcal{B}_k}(x_k) - \alpha_k \nabla f_{\mathcal{B}_k}(x_k)^\top \xi_{\alpha_k, \mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2} \left\| \xi_{\alpha_k, \mathcal{B}_k}(x_k) \right\|^2 + \lambda \Omega(x_k) - \alpha_k \lambda v^\top \xi_{\alpha_k, \mathcal{B}_k}(x_k) \\ = & \Psi_{\mathcal{B}_k}(x_k) - \alpha_k (\nabla f_{\mathcal{B}_k}(x_k) + \lambda v)^\top \xi_{\alpha_k, \mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2} \left\| \xi_{\alpha_k, \mathcal{B}_k}(x_k) \right\|^2 \\ = & \Psi_{\mathcal{B}_k}(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2} \right) \left\| \xi_{\alpha_k, \mathcal{B}_k}(x_k) \right\|^2, \end{aligned}$$

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 Half-Space Step given α_k is small enough. By taking the expectation on both sides, we obtain the following result characterizing the sufficient decrease from $\Psi(x_k)$ to $\mathbb{E}\left[\Psi(x_{k+1})\right]$.

Corollary 1. *For iteration k, we have*

(i) if kth iteration conducts Prox-SG Step, then

$$\mathbb{E}\left[\Psi(x_{k+1})\right] \leq \Psi(x_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2} \right) \mathbb{E}\left[\left\| \xi_{\alpha_k, \mathcal{B}_k}(x_k) \right\|^2 \right]. \tag{37}$$

(ii) if kth iteration conducts Half-Space Step, $x_k \in \mathcal{X}$, then

$$\mathbb{E}\left[\Psi(x_{k+1})\right] \leq \Psi(x_k) - \sum_{g \in \tilde{\mathcal{G}}_k} \left(\alpha_k - \frac{\alpha_k^2 L}{2} \right) \mathbb{E}\left[\left\| \partial \Psi_{\mathcal{B}_k}(x_k) \right\|^2 \right] - \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2} \right) \sum_{g \in \hat{\mathcal{G}}_k} \left\| [x_k]_g \right\|^2. \tag{38}$$

Corollary 1 shows that the bound of Ψ depends on step size α_k and norm of search direction. It further indicates that both Half-Space Step and Prox-SG Step can make some progress to optimality with proper selection of α_k.

C.2 PROOF OF THEOREM 1

Toward that end, we first show that if the optimal distance from x_k to the local minimizer x^* is sufficiently small, then HSPG already covers the supports of x^*, *i.e.*, $\mathcal{I}^{\neq 0}(x^*) \subseteq \mathcal{I}^{\neq 0}(x_k)$.

Lemma 3. *If $\|x_k - x^*\| \leq R$, then $\mathcal{I}^{\neq 0}(x^*) \subseteq \mathcal{I}^{\neq 0}(x_k)$.*

Proof. For any $g \in I^{\neq 0}(x^*)$, by the assumption of this lemma and the definition of R as (20) and δ_1 as (17), we have that

$$\begin{aligned} \|[x^*]_g\| - \|[x_k]_g\| \leq \|[x_k - x^*]_g\| \leq \|x_k - x^*\| \leq R \leq \delta_1 \\ \|[x_k]_g\| \geq \|[x^*]_g\| - \delta_1 \geq 2\delta_1 - \delta_1 = \delta_1 > 0 \end{aligned} \tag{39}$$

Hence $\|[x_k]_g\| \neq 0$, *i.e.*, $g \in \mathcal{I}^{\neq 0}(x_k)$. Therefore, $\mathcal{I}^{\neq 0}(x^*) \subseteq \mathcal{I}^{\neq 0}(x_k)$. $\square$

The next lemma shows that if the distance between current iterate x_k and x^*, *i.e.*, $\|x_k - x^*\|$ is sufficiently small, then x^* inhabits the reduced space $\mathcal{S}_k := \mathcal{S}(x_k)$.

Lemma 4. *Under Assumption 1, if $0 \leq \epsilon < \frac{\delta_1^2}{\delta_2}$, $\|x_k - x^*\| \leq R$, then for each $g \in \mathcal{I}^{\neq 0}(x^*)$,*

$$[x_k]_g^\top [x^*]_g \geq \epsilon \left\|[x_k]_g\right\|^2 \tag{40}$$

Consequently, it implies $x^* \in \mathcal{S}_k$ by the definition as (3).

Proof. It follows the assumption of this lemma and the definition of R in (20), δ_1 and δ_2 in (20), (17) and (18) that for any $g \in \mathcal{I}^{\neq 0}(x^*)$,

$$\|[x_k]_g\| \leq \|[x^*]_g\| + R \leq 2\delta_2 + R, \tag{41}$$

and the $\left[-(\delta_1 + 2\epsilon\delta_2) + \sqrt{(\delta_1 + 2\epsilon\delta_2)^2 - 4\epsilon^2\delta_2 + 4\epsilon\delta_1^2}\right]/\epsilon$ in (20) is actually the solution of $\epsilon z^2 + (4\epsilon\delta_2 + 2\delta_1)z + 4\epsilon\delta_2^2 - 4\delta_1^2 = 0$ regarding $z \in \mathbb{R}^+$. Then we have that

$$\begin{aligned} [x_k]_g^\top [x^*]_g =& [x_k - x^* + x^*]_g^\top [x^*]_g \\ =& [x_k - x^*]_g^\top [x^*]_g + \left\|[x^*]_g\right\|^2 \\ \geq& \left\|[x^*]_g\right\|^2 - \|[x_k - x^*]_g\| \, \|[x^*]_g\| \\ =& \|[x^*]_g\| \left(\|[x^*]_g\| - \|[x_k - x^*]_g\|\right) \\ \geq& 2\delta_1(2\delta_1 - R) \geq \epsilon(2\delta_2 + R)^2 \\ \geq& \epsilon \left\|[x_k]_g\right\|^2 \end{aligned} \tag{42}$$

holds for any $g \in \mathcal{I}^{\neq 0}(x^*)$, where the second last inequality holds because that $2\delta_1(2\delta_1 - R) = \epsilon(2\delta_2 + R)^2$ as $R = \left[-(\delta_1 + 2\epsilon\delta_2) + \sqrt{(\delta_1 + 2\epsilon\delta_2)^2 - 4\epsilon^2\delta_2 + 4\epsilon\delta_1^2}\right]/\epsilon$. Now combing with the definition of $\mathcal{S}_k$ as (3), we have x^* inhabits $\mathcal{S}_k$, which completes the proof. $\square$

Furthermore, if $\|x_k - x^*\|$ is small enough and the step size is selected properly, every recovery of group sparsity by Half-Space Step can be guaranteed as successful as stated in the following lemma.

Lemma 5. *Suppose $k \geq N_{\mathcal{P}}$, $\|x_k - x^*\| \leq R$, $0 \leq \epsilon < \frac{2\delta_1 - R}{2\delta_2 + R}$ and $0 < \alpha_k \leq \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}$, then for any $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^0(x_{k+1})$, g must be in $\mathcal{I}^0(x^*)$, i.e., $g \in \mathcal{I}^0(x^*)$.*

Proof. To prove it by contradiction, suppose there exists some $g \in \hat{\mathcal{G}}_k$ such that $g \in \mathcal{I}^{\neq 0}(x^*)$. Since $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^0(x_{k+1})$, then the group projection (5) is trigerred at g such that

$$\begin{aligned} [\tilde{x}_{k+1}]_g^\top [x_k]_g &= [x_k - \alpha \nabla\Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g \\ &= \left\|[x_k]_g\right\|^2 - \alpha_k [\nabla\Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g < \epsilon \left\|[x_k]_g\right\|^2 . \end{aligned} \tag{43}$$

On the other hand, it follows the assumption of this lemma and $g \in \mathcal{I}^{\neq 0}(x^*)$ that

$$\|[x_k - x^*]_g\| \leq \|x_k - x^*\| \leq R \tag{44}$$

Combining the definition of δ_1 as (17) and δ_2 as (18), we have that

$$\begin{aligned} \|[x_k]_g\| &\geq \|[x^*]_g\| - R \geq 2\delta_1 - R \\ \|[x_k]_g\| &\leq \|[x^*]_g\| + R \leq 2\delta_2 + R \end{aligned} \tag{45}$$

It then follows $0 < \alpha_k \leq \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}$, where note $2\delta_1 - R - \epsilon(2\delta_2 + R) > 0$ as $R \leq \delta_1$ and $\epsilon < \frac{2\delta_1 - R}{2\delta_2 + R}$, that

$$\begin{aligned} [\tilde{x}_{k+1}]_g^\top [x_k]_g &= \|[x_k]_g\|^2 - \alpha_k [\nabla \Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g \\ &\geq \|[x_k]_g\|^2 - \alpha_k \|[\nabla \Psi_{\mathcal{B}_k}(x_k)]_g\| \, \|[x_k]_g\| \\ &= \|[x_k]_g\| \left(\|[x_k]_g\| - \alpha_k \|[\nabla \Psi_{\mathcal{B}_k}(x_k)]_g\|\right) \\ &\geq \|[x_k]_g\| \left(\|[x_k]_g\| - \alpha_k M\right) \\ &\geq \|[x_k]_g\| \left[(2\delta_1 - R) - \alpha_k M\right] \\ &\geq \|[x_k]_g\| \left[(2\delta_1 - R) - \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M} M\right] \\ &\geq \|[x_k]_g\| \left[(2\delta_1 - R) - 2\delta_1 + R + \epsilon(2\delta_2 + R)\right] \\ &\geq \epsilon \|[x_k]_g\| (2\delta_2 + R) \\ &\geq \epsilon \|[x_k]_g\|^2 \end{aligned} \tag{46}$$

which contradicts with (43). Hence, we conclude that any g of variables projected to zero, *i.e.*, $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(x_k) \bigcap \mathcal{I}^0(x_{k+1})$ are exactly also the zeros on the optimal solution x^*, *i.e.*, $g \in \mathcal{I}^0(x^*)$. □

We next present that if the iterate of Half-Space Step is close enough to the optimal solution x^*, then x^* inhabits all reduced spaces constructed by the subsequent iterates of Half-Space Step with high probability. To establish this results, we require the below two lemmas. The first bounds the accumulated error because of random sampling.

Lemma 6. *Given any $\theta > 1$, $K \geq N_{\mathcal{P}}$, let $k := K + t$, $t \in \mathbb{Z}^+ \bigcup \{0\}$, then there exists $\alpha_k = \mathcal{O}(1/t)$ and $|\mathcal{B}_k| = \mathcal{O}(t)$, such that for any $y_t \in \mathbb{R}^n$,*

$$\max_{\{y_t\}_{t=0}^\infty \in \mathcal{X}^\infty} \sum_{t=0}^{\infty} \alpha_k \|e_{\mathcal{B}_k}(y_t)\|_2 \leq \frac{3R^2}{8(4R+1)}$$

holds with probability at least $1 - \frac{1}{\theta^2}$.

Proof. Define random variable $Y_t := \alpha_{K+t} \|e_{\mathcal{B}_{K+t}}(y_t)\|_2$ for all $t \geq 0$. Since $\{y_t\}_{t=0}^\infty$ are arbitrarily chosen, then the random variables $\{Y_t\}_{t=0}^\infty$ are independent. Let $Y := \sum_{t=0}^\infty Y_t$. Using Chebshev's inequality, we obtain

$$\mathbb{P}\left(Y \geq \mathbb{E}[Y] + \theta\sqrt{\mathrm{Var}[Y]}\right) \leq \mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq \theta\sqrt{\mathrm{Var}[Y]}\right) \leq \frac{1}{\theta^2}. \tag{47}$$

And based on the Assumption 1, there exists an upper bound $\sigma^2 > 0$ for the variance of random noise $e(x)$ generated from the one-point mini-batch, *i.e.*, $\mathcal{B} = \{i\}, i = 1, \ldots, N$. Consequently, for each $t \geq 0$, we have $\mathbb{E}[Y_t] \leq \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}}$ and $\mathrm{Var}[Y_t] \leq \frac{\alpha_{K+t}^2 \sigma^2}{|\mathcal{B}_{K+t}|}$, then combining with (47), we have

$$\begin{aligned} Y &\leq \mathbb{E}[Y] + \theta\sqrt{\mathrm{Var}[Y]} && (48) \\ &\leq \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{k+t}|}} + \theta \cdot \sum_{t=0}^{\infty} \frac{\alpha_{K+t}^2 \sigma^2}{|\mathcal{B}_{K+t}|} && (49) \\ &\leq \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{k+t}|}} + \theta \cdot \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} = (1+\theta) \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} && (50) \end{aligned}$$

holds with probability at least $1 - \frac{1}{\theta^2}$. Here, for the second inequality, we use the property that the equality $\mathbb{E}[\sum_{t=0}^{\infty} Y_i] = \sum_{t=0}^{\infty} \mathbb{E}[Y_i]$ holds whenever $\sum_{t=0}^{\infty} \mathbb{E}[|Y_i|]$ convergences, see Section 2.1 in Mitzenmacher (2005); and for the third inequality, we use $\frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} \leq 1$ without loss of generality as the common setting of large mini-batch size and small step size.

Given any $\theta > 1$, there exists some $\alpha_k = \mathcal{O}(1/t)$ and $|\mathcal{B}_k| = \mathcal{O}(t)$, the above series converges and satisfies that

$$(1+\theta)\sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} \leq \frac{3R^2}{8(4R+1)}$$

holds. Notice that the above proof holds for any given sequence $\{y_t\}_{t=0}^{\infty} \in \mathcal{X}^{\infty}$, thus

$$\max_{\{y_t\}_{t=0}^{\infty} \in \mathcal{X}^{\infty}} \sum_{t=0}^{\infty} \alpha_k \|e_{\mathcal{B}_k}(y_t)\|_2 \leq \frac{3R^2}{8(4R+1)}$$

holds with probability at least $1 - \frac{1}{\theta^2}$. □

The second lemma draws if previous iterate of Half-Space Step falls into the neighbor of x^*, then under appropriate step size and mini-batch setting, the current iterate also inhabits the neighbor with high probability.

Lemma 7. *Under the assumptions of Lemma 6, suppose* $\|x_K - x^*\| \leq R/2$*; for any* ℓ *satisfying* $K \leq \ell < K + t$, $0 < \alpha_\ell \leq \min\{\frac{1}{L}, \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}\}$, $|B_\ell| \geq N - \frac{N}{2M}$ *and* $\|x_\ell - x^*\| \leq R$ *holds, then*

$$\|x_{K+t} - x^*\| \leq R. \tag{51}$$

holds with probability at least $1 - \frac{1}{\theta^2}$.

Proof. It follows the assumptions of this lemma, Lemma 5, (15) and (16) that for any ℓ satisfying $K \leq \ell < K + t$

$$\|[x^*]_g\| = 0, \text{ for any } g \in \hat{\mathcal{G}}_\ell. \tag{52}$$

Hence we have that for $K \leq \ell < K + t$,

$$\begin{aligned}
&\|x_{\ell+1} - x^*\|^2 \\
=& \sum_{g\in\tilde{\mathcal{G}}_\ell} \|[x_\ell - x^* - \alpha_\ell \nabla\Psi(x_\ell) - \alpha_\ell e_{\mathcal{B}_\ell}(x_\ell)]_g\|^2 + \sum_{g\in\hat{\mathcal{G}}_k} \|[x_\ell - x^* - x_\ell]_g\|^2 \\
=& \sum_{g\in\tilde{\mathcal{G}}_\ell} \left\{\|[x_\ell - x^*]_g\|^2 - 2\alpha_\ell [x_\ell - x^*]_g^\top [\nabla\Psi(x_\ell) + e_{\mathcal{B}_\ell}(x_\ell)]_g + \alpha_\ell^2 \|[\nabla\Psi(x_\ell) + e_{\mathcal{B}_\ell}(x_\ell)]_g\|^2\right\} + \sum_{g\in\hat{\mathcal{G}}_\ell} \|[x^*]_g\|^2 \\
=& \sum_{g\in\tilde{\mathcal{G}}_\ell} \left\{\|[x_\ell - x^*]_g\|^2 - 2\alpha_\ell [x_\ell - x^*]_g^\top [\nabla\Psi(x_\ell)]_g - 2\alpha_\ell [x_\ell - x^*]_g^\top [e_{\mathcal{B}_\ell}(x_\ell)]_g + \alpha_\ell^2 \|[\nabla\Psi(x_\ell) + e_{\mathcal{B}_\ell}(x_\ell)]_g\|^2\right\} \\
\leq& \sum_{g\in\tilde{\mathcal{G}}_\ell} \|[x_\ell - x^*]_g\|^2 - \|[\nabla\Psi(x_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) - 2\alpha_\ell [x_\ell - x^*]_g^\top [e_{\mathcal{B}_\ell}(x_\ell)]_g + \alpha_\ell^2 \|[e_{\mathcal{B}_\ell}(x_\ell)]_g\|^2 \\
&+ 2\alpha_\ell^2 [\nabla\Psi(x_\ell)]_g^\top [e_{\mathcal{B}_\ell}(x_\ell)]_g \\
\leq& \sum_{g\in\tilde{\mathcal{G}}_\ell} \|[x_\ell - x^*]_g\|^2 - \|[\nabla\Psi(x_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) + 2\alpha_\ell \|[x_\ell - x^*]_g\| \|[e_{\mathcal{B}_\ell}(x_\ell)]_g\| + \alpha_\ell^2 \|[e_{\mathcal{B}_\ell}(x_\ell)]_g\|^2 \\
&+ 2\alpha_\ell^2 \|[\nabla\Psi(x_\ell)]_g\| \|[e_{\mathcal{B}_\ell}(x_\ell)]_g\| \\
\leq& \sum_{g\in\tilde{\mathcal{G}}_\ell} \|[x_\ell - x^*]_g\|^2 - \|[\nabla\Psi(x_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) + (2\alpha_\ell + 2\alpha_\ell^2 L) \|[x_k - x^*]_g\| \|[e_{\mathcal{B}_\ell}(x_\ell)]_g\| + \alpha_\ell^2 \|[e_{\mathcal{B}_\ell}(x_\ell)]_g\|^2 \\
\leq& \sum_{g\in\tilde{\mathcal{G}}_\ell} \left\{\|[x_\ell - x^*]_g\|^2 - \|[\nabla\Psi(x_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right)\right\} + (2\alpha_\ell + 2\alpha_\ell^2 L) \|x_\ell - x^*\| \|e_{\mathcal{B}_\ell}(x_\ell)\| + \alpha_\ell^2 \|e_{\mathcal{B}_\ell}(x_\ell)\|^2
\end{aligned} \tag{53}$$

On the other hand, by the definition of $e_{\mathcal{B}}(x)$, we have that

$$\begin{aligned}
e_{\mathcal{B}}(x) =&[\nabla\Psi_{\mathcal{B}}(x) - \nabla\Psi(x)]_{\mathcal{I}^{\neq 0}(x)} = [\nabla f_{\mathcal{B}}(x) - \nabla f(x)]_{\mathcal{I}^{\neq 0}(x)} \\
=&\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}[\nabla f_j(x)]_{\mathcal{I}^{\neq 0}(x)} - \frac{1}{N}\sum_{i=1}^{N}[\nabla f_i(x)]_{\mathcal{I}^{\neq 0}(x)} \\
=&\frac{1}{N}\sum_{j\in\mathcal{B}}\left[\frac{N}{|\mathcal{B}|}[\nabla f_j(x)]_{\mathcal{I}^{\neq 0}(x)} - [\nabla f_j(x)]_{\mathcal{I}^{\neq 0}(x)}\right] - \frac{1}{N}\sum_{\substack{i=1\\ i\notin\mathcal{B}}}^{N}[\nabla f_i(x)]_{\mathcal{I}^{\neq 0}(x)} \\
=&\frac{1}{N}\sum_{j\in\mathcal{B}}\left[\frac{N-|\mathcal{B}|}{|\mathcal{B}|}[\nabla f_j(x)]_{\mathcal{I}^{\neq 0}(x)}\right] - \frac{1}{N}\sum_{\substack{i=1\\ i\notin\mathcal{B}}}^{N}[\nabla f_i(x)]_{\mathcal{I}^{\neq 0}(x)}
\end{aligned}\tag{54}$$

Thus taking the norm on both side of (54) and using triangle inequality results in the following:

$$\begin{aligned}
\|e_{\mathcal{B}}(x)\| \leq& \frac{1}{N}\sum_{j\in\mathcal{B}}\left[\frac{N-|\mathcal{B}|}{|\mathcal{B}|}\left\|[\nabla f_j(x)]_{\mathcal{I}^{\neq 0}(x)}\right\|\right] + \frac{1}{N}\sum_{\substack{i=1\\ i\notin\mathcal{B}}}^{N}\left\|[\nabla f_i(x)]_{\mathcal{I}^{\neq 0}(x)}\right\| \\
\leq& \frac{1}{N}\frac{N-|\mathcal{B}|}{|\mathcal{B}|}|\mathcal{B}_k|M + \frac{1}{N}(N-|\mathcal{B}|)M \leq \frac{2(N-|\mathcal{B}|)M}{N}.
\end{aligned}\tag{55}$$

Since $\alpha_\ell \leq 1$, and $|B_\ell| \geq N - \frac{N}{2M}$ hence $\alpha_\ell\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| \leq 1$. Then combining with $\alpha_\ell \leq 1/L$, (53) can be further simplified as

$$\begin{aligned}
&\left\|x_{\ell+1} - x^*\right\|^2 \\
\leq& \sum_{g\in\tilde{\mathcal{G}}_\ell}\left\{\left\|[x_\ell - x^*]_g\right\|^2 - \left\|[\nabla\Psi(x_\ell)]_g\right\|^2\left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right)\right\} + (2\alpha_\ell + 2\alpha_\ell^2 L)\left\|x_\ell - x^*\right\|\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| + \alpha_\ell^2\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\|^2 \\
\leq& \sum_{g\in\tilde{\mathcal{G}}_\ell}\left\{\left\|[x_\ell - x^*]_g\right\|^2 - \frac{1}{L^2}\left\|[\nabla\Psi(x_\ell)]_g\right\|^2\right\} + 4\alpha_\ell\left\|x_\ell - x^*\right\|\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| + \alpha_\ell^2\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\|^2 \\
\leq& \left\|x_\ell - x^*\right\|^2 + 4\alpha_\ell\left\|x_\ell - x^*\right\|\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| + \alpha_\ell\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\|
\end{aligned}\tag{56}$$

Following from the assumption that $\|x_\ell - x^*\| \leq R$, then (56) can be further simplified as

$$\begin{aligned}
\left\|x_{\ell+1} - x^*\right\|^2 \leq& \left\|x_\ell - x^*\right\|^2 + 4\alpha_\ell R\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| + \alpha_k\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| \\
\leq& \left\|x_\ell - x^*\right\|^2 + (4R+1)\alpha_\ell\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\|
\end{aligned}\tag{57}$$

Summing the the both side of (57) from $\ell = K$ to $\ell = K + t - 1$ results in

$$\left\|x_{K+t} - x^*\right\|^2 \leq \left\|x_K - x^*\right\|^2 + (4R+1)\sum_{\ell=K}^{K+t-1}\alpha_\ell\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| \tag{58}$$

It follows Lemma 6 that the followng holds with probability at least $1 - \frac{1}{\theta^2}$,

$$\sum_{\ell=K}^{\infty}\alpha_\ell\|e_{\mathcal{B}_\ell}(x_\ell)\| \leq \frac{3R^2}{4(4R+1)}. \tag{59}$$

Thus we have that

$$\begin{aligned}
\left\|x_{K+t} - x^*\right\|^2 \leq& \left\|x_K - x^*\right\|^2 + (4R+1)\sum_{\ell=K}^{K+t-1}\alpha_\ell\left\|e_{\mathcal{B}_\ell}(x_\ell)\right\| \\
\leq& \left\|x_K - x^*\right\|^2 + (4R+1)\sum_{\ell=K}^{\infty}\alpha_\ell\|e_{\mathcal{B}_\ell}(x_\ell)\| \\
\leq& \frac{R^2}{4} + (4R+1)\frac{3R^2}{4(4R+1)} \leq \frac{R^2}{4} + \frac{3R^2}{4} \leq R^2,
\end{aligned}\tag{60}$$

holds with probability at least $1 - \frac{1}{\theta^2}$, which completes the proof.

$\square$

Based on the above lemmas, the Lemma 8 below shows if initial iterate of Half-Space Step locates closely enough to x^*, step size α_k polynomially decreases, and mini-batch size $\mathcal{B}_k$ polynomially increases, then x^* inhabits all subsequent reduced space $\{\mathcal{S}_k\}_{k=K}^{\infty}$ constructed in Half-Space Step with high probability.

Lemma 8. *Suppose* $\|x_K - x^*\| \leq \frac{R}{2}$, $K \geq N_{\mathcal{P}}$, $k = K + t$, $t \in \mathbb{Z}^+$, $0 < \alpha_k = \mathcal{O}(1/(\sqrt{N}t)) \leq \min\{\frac{2(1-\epsilon)}{L}, \frac{1}{L}, \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}\}$ *and* $|\mathcal{B}_k| = \mathcal{O}(t) \geq N - \frac{N}{2M}$. *Then for any constant* $\tau \in (0, 1)$, $\|x_k - x^*\| \leq R$ *with probability at least* $1 - \tau$ *for any* $k \geq K$.

Proof. It follows Lemma 4 and the assumption of this lemma that $x^* \in \mathcal{S}_K$. Moreover, it follows the assumptions of this lemma, Lemma 6 and 7, the definition of finite-sum $f(x)$ in (1), and the bound of error as (55) that

$$\mathbb{P}(\{x_k\}_{k=K}^{\infty} \in \{x : \|x - x^*\| \leq R\}^{\infty}) \geq \left(1 - \frac{1}{\theta^2}\right)^{\mathcal{O}(N-K)} \geq 1 - \tau, \tag{61}$$

where the last two inequalities comes from that the error vanishing to zero as $|\mathcal{B}_k|$ reaches the upper bound N, and θ is sufficiently large depending on τ and $\mathcal{O}(N - K)$. $\square$

Corollary 2. *Lemma 8 further implies* x^* *inhabits all subsequent* $\mathcal{S}_k$, *i.e.,* $x^* \in \mathcal{S}_k$ *for any* $k \geq K$.

Next, we establish that after finitely number of iterations, HSPG generates sequences that inhabits in the feasible domain $\mathcal{X}$ where Lipschitz continuity of Ψ holds.

Lemma 9. *Suppose the assumptions of Lemma 8 hold, then after finite number of iterations, all subsequent iterates* $x_k \in \mathcal{X}$ *with high probability.*

Proof. It follows Lemma 8 that all subsequent x_k satisfying $\|x_k - x^*\| \leq R$ with high probability. Combining with Lemma 3, we have that $\mathcal{I}^{\neq 0}(x^*) \subseteq \mathcal{I}^{\neq 0}(x_k)$ for all $k \geq K$ with high probability. Then for any $g \in \mathcal{I}^{\neq 0}(x_k)$, there are two possbilities, either $g \in \mathcal{I}^{\neq 0}(x^*)$ or $g \in \mathcal{I}^{0}(x^*)$. For the first case $g \in \mathcal{I}^{\neq 0}(x^*) \bigcap \mathcal{I}^{\neq 0}(x_k)$, it follows the definitions of R as (20) and δ_1 as (17) that

$$\begin{aligned} \|[x_k - x^*]_g\| \leq \|x_k - x^*\| \leq R \leq \delta_1 \\ \|[x^*]_g\| - \|[x_k]_g\| \leq \delta_1 \\ \|[x_k]_g\| \geq \|[x^*]_g\| - \delta_1 \geq 2\delta_1 - \delta_1 = \delta_1 \end{aligned} \tag{62}$$

For any $g \in \mathcal{I}^{0}(x^*) \bigcap \mathcal{I}^{\neq 0}(x_k)$, by Algorithm 3, its norm is bounded below by

$$\delta_1 \geq \|[x_k - x^*]_g\| = \|[x_k]_g\| \geq \epsilon^t \|[x_K]_g\|, \tag{63}$$

where by the Theorem 2 will shown in Appendix C.3, if $\|[x_k]_g\| \leq \frac{2\alpha_k \delta_3}{1 - \epsilon + \alpha_k L}$, then $[x_{k+1}]_g$ equals to zero and will be fixed as zero since Algorithm 3 operates on $\mathcal{S}_k$ as (3). Note $\alpha_k = \mathcal{O}(1/t)$, following (Karimi et al., 2016, Theorem 4) and (Drusvyatskiy & Lewis, 2018, Theorem 3.2), $\mathbb{E}[\|[x_k]_g\|^2] = \mathcal{O}(1/t)$. If $\epsilon > 0$, then after finite number of iterations $\mathcal{O}(1/\epsilon^2)$, $g \in \mathcal{I}^{0}(x^*) \bigcap \mathcal{I}^{\neq 0}(x_k)$ becomes zero. If $\epsilon = 0$, note $\mathcal{B}_k = \mathcal{O}(t)$ and f is finite-sum, then similar result holds by (Gower, 2018, Theorem 2.3, Theorem 3.2) (f needs further strongly convexity on $\tilde{\mathcal{X}}$). Hence with high probability, after finite number of iterations, denoted by T, all subsequent x_k, $k \geq K + T$ inhabits $\mathcal{X}$. Regarding $[x_k]_{g \in \mathcal{I}^{0}(x^*) \bigcap \mathcal{I}^{\neq 0}(x_k)}$ for $K \leq k \leq K + T$, note $\epsilon^t \|[x_K]_g\|$ is also bounded below by constant $\epsilon^T \|[x_K]_g\| > 0$ given x_K, for similicity, denote the Lipschitz constant of $[\nabla \Psi(x_k)]_g$ as L as well. $\square$

We now prove the first main theorem of HSPG, *i.e.*, Theorem 1.

Theorem 1. *Suppose f is convex on $\widetilde{\mathcal{X}}$, $\epsilon \in \left[0, \min\{\frac{\delta_1^2}{\delta_2}, \frac{2\delta_1 - R}{2\delta_2 + R}\}\right)$, $\|x_K - x^*\| < \frac{R}{2}$ for some $K \geq N_{\mathcal{P}}$. Set $k := K + t$, $(t \in \mathbb{Z}^+)$, step size $\alpha_k = \mathcal{O}(\frac{1}{\sqrt{N}t}) \in (0, \min\{\frac{2(1-\epsilon)}{L}, \frac{1}{L}, \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}\})$, and mini-batch size $|\mathcal{B}_k| = \mathcal{O}(t) \leq N - \frac{N}{2M}$. Then for any $\tau \in (0,1)$, we have $\{x_k\}$ converges to some stationary point in expectation with probability at least $1 - \tau$, i.e., $\mathbb{P}(\lim_{k\to\infty} \mathbb{E}\left[\|\xi_{\alpha_k, \mathcal{B}_k}(x_k)\|\right] = 0) \geq 1 - \tau$.*

Proof. We know that Algorithm 1 performs an infinite sequence of iterations. It follows Corollary 1 that for any $\ell \in \mathbb{Z}^+$,

$$\begin{aligned} &\mathbb{E}[\Psi(x_K)] - \mathbb{E}[\Psi(x_{\ell+1})] = \sum_{k=K}^{\ell} \{\mathbb{E}[\Psi(x_k)] - \mathbb{E}[\Psi(x_{k+1})]\} \\ \geq &\sum_{K \leq k \leq \ell} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] + \sum_{K \leq k \leq \ell} \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[x_k]_g\|^2 . \end{aligned} \tag{64}$$

Combining the assumption that Ψ is bounded below and letting $\ell \to \infty$, we obtain

$$\sum_{k \geq K} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] + \sum_{k \geq K} \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[x_k]_g\|^2 < \infty \tag{65}$$

By Algorithm 3, variables on $\mathcal{I}^0(x_k)$ are fixed during kth Half-Space Step and n is finite, then the group projection appears finitely many times, consequently,

$$\sum_{k \geq K} \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[x_k]_g\|^2 < \infty. \tag{66}$$

Thus (65) implies that

$$\sum_{k \geq K} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] \tag{67}$$

$$= \sum_{k \geq K} \alpha_k \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] - \sum_{k \geq K} \frac{\alpha_k^2}{L} \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] < \infty \tag{68}$$

Since $\alpha_k = \mathcal{O}(1/(\sqrt{N}t))$, then $\sum_{k \geq K} \alpha_k = \infty$ and $\sum_{k \geq K} \alpha_k^2 \leq \infty$. Combining with (67) and the boundness of $\partial \Psi$, it implies

$$\sum_{k \geq K} \alpha_k \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] < \infty. \tag{69}$$

By $\sum_{k \geq K} \alpha_k = \infty$ and (69), we have that

$$\liminf_{k \geq K} \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] = 0 \tag{70}$$

then there exists a subsequence $\mathcal{K}$ such that

$$\lim_{k \in \mathcal{K}} \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\|[\nabla \Psi(x_k)]_g\|^2\right] = 0 \tag{71}$$

It follows from the assumptions of this theorem and Lemma 3 to 8 and Corollay 2 that with high probability at least $1 - \tau$, for each $k \geq K$, x^* inhabits $\mathcal{S}_k$. Note as $|\mathcal{B}_k| = \mathcal{O}(t)$ linearly increases, the error of gradient estimate vanishes. Hence, (71) naturally implies that the sequence $\{x_k\}_{k \in \mathcal{K}}$ converges to some stationary point with high probability. And we can extend $\mathcal{K}$ to $\{k : k \geq K\}$ due to the non-decreasing distance to optimal solution as shown in the Lemma 8. By the above, we conclude that

$$\mathbb{P}(\lim_{k \to \infty} \mathbb{E}\left[\|\xi_{\alpha_k, \mathcal{B}_k}(x_k)\|\right] = 0) \geq 1 - \tau. \tag{72}$$

□

C.3 PROOF OF THEOREM 2

In this Appendix, we compare the group sparsity identification property of HSPG and Prox-SG. We first show the generic sparsity identification property of Prox-SG for any mixed ℓ_1/ℓ_p regularization for $p \geq 1$.

Lemma 10. *If $\|x_k - x^*\|_{p'} \leq \min\{\delta_3/L, \alpha_k\delta_3\}$, where $1/p + 1/p' = 1$ ($p' = \infty$ if $p = 1$), then the Prox-SG yields that for each $g \in \mathcal{I}^0(x^*)$, $[x_{k+1}]_g = 0$ holds, i.e., $\mathcal{I}^0(x^*) \subseteq \mathcal{I}^0(x_{k+1})$.*

Proof. It follows from the reverse triangle inequality, basic norm inequalities, Lipschitz continuity of $\nabla f(x)$ and the assumption of this lemma that for any $g \in \mathcal{G}$,

$$\begin{aligned}\|[\nabla f_{\mathcal{B}_k}(x_k)]_g\|_{p'} - \|[\nabla f_{\mathcal{B}_k}(x^*)]_g\|_{p'} &\leq \|[\nabla f_{\mathcal{B}_k}(x_k) - \nabla f_{\mathcal{B}_k}(x^*)]_g\|_{p'}\\ &\leq \|\nabla f_{\mathcal{B}_k}(x_k) - \nabla f_{\mathcal{B}_k}(x^*)\|_{p'}\\ &\leq L\,\|x_k - x^*\|_{p'} \leq L \cdot \frac{\delta_3}{L} = \delta_3.\end{aligned} \tag{73}$$

By (73), we have that for any $g \in \mathcal{I}^0(x^*)$,

$$\begin{aligned}\|[\nabla f_{\mathcal{B}_k}(x_k)]_g\|_{p'} &\leq \|[\nabla f_{\mathcal{B}_k}(x^*)]_g\|_{p'} + \delta_3\\ &\leq \lambda - 2\delta_3 + \delta_3 = \lambda - \delta_3\end{aligned} \tag{74}$$

Combining (74) and the assumption of this lemma, the following holds for any $\alpha_k > 0$ that

$$\begin{aligned}\|[x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g\|_{p'} &\leq \|[x_k]_g\|_{p'} + \|[\alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g\|_{p'}\\ &\leq \alpha_k\delta_3 + \alpha_k(\lambda - \delta_3) = \alpha_k\lambda\end{aligned} \tag{75}$$

which further implies that the Ecludiean projection yields that

$$\mathrm{Proj}^E_{\mathcal{B}(\|\cdot\|_{p'}, \alpha_k\lambda)}([x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g) = [x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g. \tag{76}$$

Combining with (76), the fact that proximal operator is the residual of identity operator subtracted by Euclidean project operator onto the dual norm ball and $[x_k]_g = 0$ for any $g \in \mathcal{I}^0(x^*)$ (Chen, 2018), we have that

$$\begin{aligned}[x_{k+1}]_g &= \mathrm{Prox}_{\alpha_k\lambda\|\cdot\|_p}([x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g)\\ &= \left[I - \mathrm{Proj}^E_{\mathcal{B}(\|\cdot\|_{p'}, \alpha_k\lambda)}\right][x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g\\ &= [x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g - [x_k - \alpha_k\nabla f_{\mathcal{B}_k}(x_k)]_g = 0,\end{aligned} \tag{77}$$

consequently $\mathcal{I}^0(x^*) \subseteq \mathcal{I}^0(x_{k+1})$, which completes the proof. ☐

Now we establish the group-sparsity identification of HSPG as the restated Theorem 2.

Theorem 2. *If $k \geq N_{\mathcal{P}}$ and $\|x_k - x^*\| \leq \frac{2\alpha_k\delta_3}{1-\epsilon+\alpha_k L}$, then HSPG yields next iterate x_{k+1} such that $\mathcal{I}^0(x^*) \subseteq \mathcal{I}^0(x_{k+1})$.*

Proof. Suppose $\|x_k - x^*\| \leq \frac{2\alpha_k\delta_3}{1-\epsilon+\alpha_k L}$. There is nothing to prove if $g \in \mathcal{I}^0(x^*) \bigcap \mathcal{I}^0(x_k)$. For $g \in \mathcal{I}^0(x^*) \bigcap \mathcal{I}^{\neq 0}(x_k)$, we compute that

$$\begin{aligned}&[x_k - \alpha_k\nabla\Psi_{\mathcal{B}_k}(x_k)]_g^\top[x_k]_g - \epsilon\,\|[x_k]_g\|^2\\ =&\|[x_k]_g\|^2 - \alpha_k[\nabla\Psi_{\mathcal{B}_k}(x_k)]_g^\top[x_k]_g - \epsilon\,\|[x_k]_g\|^2\\ =&(1-\epsilon)\,\|[x_k]_g\|^2 - \alpha_k\left([\nabla f_{\mathcal{B}_k}(x_k)]_g + \lambda\frac{[x_k]_g}{\|[x_k]_g\|}\right)^\top[x_k]_g\\ =&(1-\epsilon)\,\|[x_k]_g\|^2 - \alpha_k[\nabla f_{\mathcal{B}_k}(x_k)]_g^\top[x_k]_g - \alpha_k\lambda\,\|[x_k]_g\|\\ \leq&(1-\epsilon)\,\|[x_k]_g\|^2 + \alpha_k\,\|[\nabla f_{\mathcal{B}_k}(x_k)]_g\|\,\|[x_k]_g\| - \alpha_k\lambda\,\|[x_k]_g\|\\ =&\|[x_k]_g\|\,\{(1-\epsilon)\,\|[x_k]_g\| + \alpha_k\,\|[\nabla f_{\mathcal{B}_k}(x_k)]_g\| - \alpha_k\lambda\}\end{aligned} \tag{78}$$

By the Lipschitz continuity of ∇f, we have that for each $g \in \mathcal{I}^0(x^*) \bigcap \mathcal{I}^{\neq 0}(x_k)$,

$$\begin{aligned} \|[\nabla f_{\mathcal{B}_k}(x_k) - \nabla f_{\mathcal{B}_k}(x^*)]_g\| &\leq L \|[x_k - x^*]_g\| = L \|[x_k]_g\| \\ \|[\nabla f_{\mathcal{B}_k}(x_k)]_g\| &\leq L \|[x_k]_g\| + \|[\nabla f_{\mathcal{B}_k}(x^*)]_g\| \end{aligned} \tag{79}$$

Combining with the definition of δ_3, which implies that $\|[\nabla f_{\mathcal{B}_k}(x^*)]_g\| \leq \lambda - 2\delta_3$ that

$$\|[\nabla f_{\mathcal{B}_k}(x_k)]_g\| \leq L \|[x_k]_g\| + \lambda - 2\delta_3 \tag{80}$$

Hence combining with $\|[x_k]_g\| \leq \frac{2\alpha_k\delta_3 + \epsilon}{1+\alpha_k L}$, (78) can be further written as

$$\begin{aligned} & [x_k - \alpha_k \nabla \Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g - \epsilon \|[x_k]_g\|^2 \\ \leq & \|[x_k]_g\| \left\{(1-\epsilon)\|[x_k]_g\| + \alpha_k \|[\nabla f_{\mathcal{B}_k}(x_k)]_g\| - \alpha_k \lambda\right\} \\ \leq & \|[x_k]_g\| \left\{(1-\epsilon)\|[x_k]_g\| + \alpha_k L \|[x_k]_g\| + \alpha_k\lambda - 2\alpha_k\delta_3 - \alpha_k\lambda\right\} \\ = & \|[x_k]_g\| \left\{(1-\epsilon+\alpha_k L)\|[x_k]_g\| - 2\alpha_k\delta_3\right\} \\ \leq & \|[x_k]_g\| \left\{(1-\epsilon+\alpha_k L)\frac{2\alpha_k\delta_3}{1-\epsilon+\alpha_k L} - 2\alpha_k\delta_3\right\} \\ = & \|[x_k]_g\| (2\alpha_k\delta_3 - 2\alpha_k\delta_3) = 0. \end{aligned} \tag{81}$$

which shows that $[x_k - \alpha_k \nabla \Psi_{\mathcal{B}_k}(x_k)]_g^\top [x_k]_g \leq \epsilon \|[x_k]_g\|^2$. Hence the group projection operator is trigerred on g to map the variables to zero, then $g \in \mathcal{I}^0(x_{k+1})$, *i.e.*, $[x_{k+1}]_g = 0$. Therefore, the group sparsity of x^* can be successfully identified by Half-Space Step, *i.e.*, $\mathcal{I}^0(x^*) \subseteq \mathcal{I}^0(x_{k+1})$.

□

In the end, if further assumptions hold, we can further show its group-support recovery.

Corollary 3. *Under the assumption of Theorem 2, moreover, if* $\|x_k - x^*\| \leq R$, $x^* \in \mathcal{S}_k$, $0 \leq \epsilon < \min\left\{\frac{\delta_1^2}{\delta_2}, \frac{2\delta_1 - R}{2\delta_2 + R}\right\}$ *and* $\alpha_k \leq \frac{2\delta_1 - R - \epsilon(2\delta_2+R)}{M}$, *then* $\mathcal{I}^0(x^*) = \mathcal{I}^0(x_{k+1})$ *and* $\mathcal{I}^{\neq 0}(x_{k+1}) = \mathcal{I}^{\neq 0}(x^*)$.

Proof. Moreover, besides $\|x_k - x^*\| \leq \frac{2\alpha_k\delta_3}{1-\epsilon+\alpha_k L}$, suppose $\|x_k - x^*\| \leq R$, $x^* \in \mathcal{S}_k$, $0 \leq \epsilon < \min\left\{\frac{\delta_1^2}{\delta_2}, \frac{2\delta_1 - R}{2\delta_2 + R}\right\}$ and $\alpha_k \leq \frac{2\delta_1 - R - \epsilon(2\delta_2+R)}{M}$. Then $x^* \in \mathcal{S}_k$ indicates that $\mathcal{I}^{\neq 0}(x^*) \subseteq \mathcal{I}^{\neq 0}(x_k)$ by the definition of $\mathcal{S}_k$. It still holds for x_{k+1} by Lemma 5, *i.e.*, $\mathcal{I}^{\neq 0}(x^*) \subseteq \mathcal{I}^{\neq 0}(x_{k+1})$. Combining with $\mathcal{I}^0(x^*) \subseteq \mathcal{I}^0(x_k)$, we have that both group-supports and group sparsity of x^* are identified by HSPG, *i.e.*, $\mathcal{I}^{\neq 0}(x^*) = \mathcal{I}^{\neq 0}(x_{k+1})$ and $\mathcal{I}^0(x^*) = \mathcal{I}^0(x_{k+1})$. □

C.4 PROOF OF PROPOSITION 2

This result is established under a popular Polyak-Lojasiewicz (PL) condition for non-smooth problem Li & Li (2018), *i.e.*, there exists a $\mu > 0$ such that for any $x \in \mathbb{R}^n$ and $\eta > 0$,

$$\|\xi_\eta(x)\|^2 \geq 2\mu(\Psi(x) - \Psi^*). \tag{82}$$

We first show that the general PL condition (82) implies a different Proximal PL condition in Karimi et al. (2016), i.e., there exists a $\mu > 0$ such that

$$\mathcal{D}_{\lambda\Omega(\cdot)}(x, \eta) \geq 2\mu(\Psi(x) - \Psi^*) \tag{83}$$

where

$$\mathcal{D}_{\lambda\Omega(\cdot)}(x, \eta) = -2\eta \min_{y \in \mathbb{R}^n} \left\{\nabla f(x)^\top (y - x) + \frac{\eta}{2}\|y - x\|^2 + \lambda\Omega(y) - \lambda\Omega(x)\right\}. \tag{84}$$

Lemma 11. *If there exists a* $\mu > 0$ *such that for all* $x \in \mathbb{R}^n$

$$\|\xi_\alpha(x)\|^2 \geq 2\mu(\Psi(x) - \Psi^*), \tag{85}$$

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

$$\mathcal{D}_{\lambda\Omega(\cdot)}(x, 1/\alpha) \geq 2\mu(\Psi(x) - \Psi^*). \tag{86}$$

Proof. Let $\hat{y} = \arg\min_y \left\{ \nabla f(x)^\top (y-x) + \frac{1}{2\alpha}\left\|y-x\right\|^2 + \lambda\Omega(y) - \lambda\Omega(x) \right\}$, then

$$\begin{aligned} 0 &\in \nabla f(x) + \frac{1}{\alpha}(\hat{y}-x) + \lambda\partial\Omega(\hat{y}), \\ \hat{y}-x &\in -\alpha(\nabla f(x) + \lambda\partial\Omega(\hat{y})). \end{aligned} \tag{87}$$

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

$$\begin{aligned} &\mathcal{D}_{\lambda\Omega(\cdot)}(x, 1/\alpha) \\ =&\frac{-2}{\alpha}\left\{ \nabla f(x)^\top(\hat{y}-x) + \frac{1}{2\alpha}\left\|\hat{y}-x\right\|^2 + \lambda\Omega(\hat{y}) - \lambda\Omega(x) \right\} \\ \in&\frac{-2}{\alpha}\left\{ -\alpha\nabla f(x)^\top(\nabla f(x) + \lambda\partial\Omega(\hat{y})) + \frac{1}{2\alpha}\alpha^2\left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 + \lambda\Omega(\hat{y}) - \lambda\Omega(x) \right\} \\ =&2\nabla f(x)^\top(\nabla f(x) + \lambda\partial\Omega(\hat{y})) + 2/\alpha(\lambda\Omega(\hat{x}) - \lambda\Omega(y)) - \left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 \\ \geq&2\nabla f(x)^\top(\nabla f(x) + \lambda\partial\Omega(\hat{y})) + 2/\alpha\lambda\partial\Omega(\hat{y})^\top(x-\hat{y}) - \left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 \\ =&2\nabla f(x)^\top(\nabla f(x) + \lambda\partial\Omega(\hat{y})) + \lambda\partial\Omega(\hat{y})^\top(\nabla f(x) + \lambda\partial\Omega(\hat{y})) - \left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 \\ =&2\left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 - \left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 \\ =&\left\|\nabla f(x) + \lambda\partial\Omega(\hat{y})\right\|^2 \end{aligned} \tag{88}$$

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

$$\mathcal{D}_{\lambda\Omega(\cdot)}(x, 1/\alpha) \geq \left\|\xi_\alpha(x)\right\|^2 \geq 2\mu(\Psi(x) - \Psi^*) \tag{89}$$

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 (82) as G-PL condition and the one in (83) as D-PL condition.

We now establish the linear convergence rate of Prox-SG Step under G-PL condition by extending (Karimi et al., 2016, Theorem 4) from SGD to Prox-SG.

Lemma 12. *Suppose Ψ satisfies the G-PL condition (82), we use a constant $\alpha_k \equiv \alpha < \frac{1}{2\mu}$, then we obtain a linear convergence rate up to a solution level that is proportional to α,*

$$\mathbb{E}[\Psi(x_k) - \Psi^*] \leq (1-2\mu\alpha)^k[\Psi(x_0) - \Psi^*] + \frac{LD^2\alpha}{4\mu} \tag{90}$$

where D is the bound of norm of gradient mapping estimation.

Proof. By using the update rule of Prox-SG shown in the proof of Lemma 2, we have

$$\begin{aligned} &\Psi(x_{k+1}) = f(x_{k+1}) + \lambda\Omega(x_{k+1}) \\ \leq&f(x_k) - \alpha_k\nabla f(x_k)^\top\xi_{\alpha_k,\mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2}\left\|\xi_{\alpha_k,\mathcal{B}_k}(x_k)\right\|^2 + \lambda\Omega(x_k) - \alpha_k\lambda v^\top\xi_{\alpha_k,\mathcal{B}_k}(x_k) \\ =&\Psi(x_k) - \alpha_k(\nabla f(x_k) + \lambda v)^\top\xi_{\alpha_k,\mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2}\left\|\xi_{\alpha_k,\mathcal{B}_k}(x_k)\right\|^2 \end{aligned} \tag{91}$$

holds for any $v \in \partial\Omega(x_{k+1})$. Select the v to make $\nabla f(x_k) + \lambda v$ as $\xi_{\alpha_k}(x_k)$. (91) can be simplified as

$$\Psi(x_{k+1}) \leq \Psi(x_k) - \alpha_k\xi_{\alpha_k}(x_k)^\top\xi_{\alpha_k,\mathcal{B}_k}(x_k) + \frac{\alpha_k^2 L}{2}\left\|\xi_{\alpha_k,\mathcal{B}_k}(x_k)\right\|^2 \tag{92}$$

Taking the expectation of both sides with respect to $\mathcal{B}_k$ and combining the assumption of this lemma, we have

$$\begin{aligned} \mathbb{E}[\Psi(x_{k+1})] &\leq \Psi(x_k) - \alpha_k\xi_{\alpha_k}(x_k)^\top\mathbb{E}[\xi_{\alpha_k,\mathcal{B}_k}(x_k)] + \frac{\alpha_k^2 L}{2}\mathbb{E}[\left\|\xi_{\alpha_k,\mathcal{B}_k}(x_k)\right\|^2] \\ &\leq \Psi(x_k) - \alpha_k\left\|\xi_{\alpha_k}(x_k)\right\|^2 + \frac{\alpha_k^2 L}{2}\mathbb{E}[\left\|\xi_{\alpha_k,\mathcal{B}_k}(x_k)\right\|^2] \\ &\leq \Psi(x_k) - 2\alpha_k\mu(\Psi(x_k) - \Psi^*) + \frac{\alpha_k^2 LM^2}{2} \end{aligned} \tag{93}$$

where M is the bound of norm of gradient mapping estimation which is well defined by Assumption. Subtracting Ψ^* from both sides yields:

$$\mathbb{E}[\Psi(x_{k+1}) - \Psi^*] \leq (1 - 2\mu\alpha_k)(\Psi(x_k) - \Psi^*) + \frac{\alpha_k^2 L D^2}{2} \tag{94}$$

Choosing $\alpha_k \equiv \alpha$ for any $\alpha < \frac{1}{2\mu}$ and applying (94) recursively yields

$$\begin{aligned}
\mathbb{E}[\Psi(x_{k+1}) - \Psi^*] &\leq (1 - 2\mu\alpha)^k(\Psi(x_0) - \Psi^*) + \frac{LD^2\alpha^2}{2}\sum_{i=0}^{k}(1 - 2\mu\alpha)^i \\
&\leq (1 - 2\mu\alpha)^k(\Psi(x_0) - \Psi^*) + \frac{LD^2\alpha^2}{2}\sum_{i=0}^{\infty}(1 - 2\mu\alpha)^i \\
&= (1 - 2\mu\alpha)^k(\Psi(x_0) - \Psi^*) + \frac{LD^2\alpha}{4\mu}
\end{aligned} \tag{95}$$

where the last line uses that $\alpha < \frac{1}{2\mu}$ and the limit of the geometric series which completes the proof. $\square$

The highlight idea of Theorem 2 is now presented as follows: if $f(x)$ is convex and satisfies PL condition like (82), 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 $\|x - x^*\| \leq R/2$ can be achieved by employing $N_{\mathcal{P}}$ Prox-SG Steps with high probability.

Proposition 2. *Suppose f is convex and Ψ satisfies the PL condition* (82)*. There exist some constants $C > 0, \gamma \in (0, 1/2L)$, for any fixed $\tau \in (0, 1)$, if $\alpha_k \equiv \alpha < \min\left\{\frac{2\gamma\mu\tau R^2}{(2L\gamma - 1)C}, \frac{1}{2\mu}, \frac{1}{L}\right\}$, and mini-batch size $|\mathcal{B}_k| \equiv |\mathcal{B}| > \frac{32\gamma\mu M^2}{2\gamma\mu\tau R^2 - (2L\gamma-1)C\alpha}$ for any $k < N_{\mathcal{P}}$, then $\|x_{N_{\mathcal{P}}} - x^*\| \leq R/2$ holds with probability at least $1 - \tau$, i.e., $\mathbb{P}(\|x_{N_{\mathcal{P}}} - x^*\| \leq R/2) \geq 1 - \tau$ for any $N_{\mathcal{P}} \geq K$ with $K := \left\lceil \frac{\log\left(poly(\tau R^2, 1/|\mathcal{B}|, \alpha)/(\Psi(x_0) - \Psi^*)\right)}{\log(1 - 2\mu\alpha)} \right\rceil$, where $poly(\cdot)$ is some polynomial of assembled variables.*

Proof of Proposition 2: At first, since $\Psi(x)$ satisfies the G-PL condition (82), it also satisfies D-PL condition due to Lemma 11. It then follows (Karimi et al., 2016, Appendix G), specifically D-PL condition implies the Proximal Error Bound that there exists some $\frac{1}{2L} > \gamma > 0$ such that

$$\|x - x^*\| \leq \gamma \|\xi_\eta(x)\| \tag{96}$$

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

For any $k \leq N_{\mathcal{P}}$, based on Lemma 2, 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

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_k}[\|x_{k+1} - x^*\|^2 \,|x_k] \\
=&\mathbb{E}_{\mathcal{B}_k}[\|x_k - \alpha_k\xi_{\alpha_k,\mathcal{B}_k}(x_k) - x^*\|^2 \,|x_k] \\
=&\mathbb{E}_{\mathcal{B}_k}[\|x_k - x^*\|^2 \,|x_k] - 2\alpha_k(x_k - x^*)^\top\mathbb{E}_{\mathcal{B}_k}[\xi_{\alpha_k,\mathcal{B}_k}(x_k)|x_k] + \alpha_k^2\mathbb{E}_{\mathcal{B}_k}[\|\xi_{\alpha_k,\mathcal{B}_k}(x_k)\|^2 \,|x_k] \\
=&\|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)^\top\xi_{\alpha_k}(x_k) + \alpha_k^2\{\|\mathbb{E}_{\mathcal{B}_k}[\xi_{\alpha_k,\mathcal{B}_k}(x_k)|x_k]\|^2 + \mathbb{E}_{\mathcal{B}_k}[\|e(x_k)\|^2 \,|x_k]\} \\
=&\|x_k - x^*\|^2 - 2\alpha_k(x_k - x^*)^\top\xi_{\alpha_k}(x_k) + \alpha_k^2\{\|\xi_{\alpha_k}(x_k)\|^2 + \mathbb{E}_{\mathcal{B}_k}[\|e(x_k)\|^2 \,|x_k]\} \\
=&\|x_k - \alpha_k\xi_{\alpha_k}(x_k) - x^*\|^2 + \alpha_k^2\mathbb{E}_{\mathcal{B}_k}[\|e(x_k)\|^2 \,|x_k]
\end{aligned} \tag{97}$$

where the first term $\|x_k - \alpha_k\xi_{\alpha_k}(x_k) - x^*\|^2$ 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 (96), $\alpha_k \in (0, 1/L]$ and (Drusvyatskiy & Lewis, 2018, Theorem 3.2) that

$$\|x_k - \alpha_k \xi_{\alpha_k}(x_k) - x^*\|^2 \leq \left(1 - \frac{1}{2L\gamma}\right) \hat{C}(\Psi(x_k) - \Psi^*) \tag{98}$$

where $\hat{C}$ is a constant as $\frac{2}{L(1-\sqrt{1-(2L\gamma)^{-1}})^2}$. It follows Lemma 12 that if we use a constant step size $\alpha_k \equiv \alpha < \frac{1}{2\mu}$, we obtain a linear convergence rate up to a solution level that is proportional to α,

$$\mathbb{E}\left[\Psi(x_k) - \Psi^*\right] \leq (1 - 2\mu\alpha)^k \left(\Psi(x_0) - \Psi^*\right) + \frac{LM^2\alpha}{4\mu}, \tag{99}$$

where M is the bound of norm of gradient mapping estimation.

To upper bound the second term, since the norm of gradient mapping is bounded, let $\mathcal{B}_i$ be a one-point mini-batch, by the definition M, then for any x

$$4M^2 \geq \mathbb{E}_{\mathcal{B}_i \sim \text{Unif}[n]} \left[\|\xi_{\alpha_k, \mathcal{B}_k}(x) - \xi_{\alpha_k}(x)\|^2 \right] \tag{100}$$

By computation, we have

$$\mathbb{E}_{\mathcal{B}_k} \left[\|\boldsymbol{e}(x_k)\|^2 \,\middle|\, \boldsymbol{x}_k \right] \leq \frac{4M^2}{|\mathcal{B}_k|}, \tag{101}$$

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

Therefore, combining (97), (99), (101) and $\alpha_k \in (0, 1]$,

$$\begin{aligned} &\mathbb{E}[\|x_{k+1} - x^*\|^2] \\ =&\mathbb{E}[\|x_k - \alpha_k \xi_{\alpha_k}(x_k) - x^*\|^2] + \alpha_k^2 \mathbb{E}_{\mathcal{B}_k}[\|e(x_k)\|^2 \,|x_k] \\ \leq& \left(1 - \frac{1}{2L\gamma}\right) \hat{C} \left[(1 - 2\mu\alpha)^k \left(\Psi(x_0) - \Psi^*\right) + \frac{LD^2\alpha}{4\mu} \right] + \frac{4M^2}{|\mathcal{B}_k|}. \end{aligned} \tag{102}$$

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

$$\alpha < \frac{2\gamma\mu\tau R^2}{(2L\gamma - 1)\hat{C}D^2}, \tag{103}$$

then

$$2\gamma\mu\tau R^2 - (2L\gamma - 1)\hat{C}D^2\alpha > 0 \tag{104}$$

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

$$|\mathcal{B}_k| > \frac{32\gamma\mu M^2}{2\gamma\mu\tau R^2 - (2L\gamma - 1)\hat{C}D^2\alpha} \tag{105}$$

then

$$\frac{\tau R^2}{4} - \frac{4M^2}{|\mathcal{B}_k|} - \left(1 - \frac{1}{2L\gamma}\right) \hat{C} \frac{LD^2\alpha}{4\mu} > 0. \tag{106}$$

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

$$\left(1 - \frac{1}{2L\gamma}\right) \hat{C} (1 - 2\mu\alpha)^k \left(\Psi(x_0) - \Psi^*\right) \leq \frac{\tau R^2}{4} - \frac{4M^2}{|\mathcal{B}_k|} - \left(1 - \frac{1}{2L\gamma}\right) \hat{C} \frac{LD^2\alpha}{4\mu} \tag{107}$$

Notice that the right hand side of (107) is a polynomial of $\tau R^2, 1/|\mathcal{B}_k|$ and α, and $\left(1 - \frac{1}{2L\gamma}\right)\hat{C}$ on the left hand side of (107) is a constant given Ψ. Thus to let (107) hold, k should satisfy

$$k \geq K := \left\lceil \frac{\log\left(\text{poly}(\tau R^2, 1/|\mathcal{B}_k|, \alpha)/(\Psi(x_0) - \Psi^*)\right)}{\log\left(1 - 2\mu\alpha\right)} \right\rceil \tag{108}$$

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

Now, it follows (102) that if (103), (105) and (108) hold, then

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le \frac{\tau R^2}{4}, \tag{109}$$

now combine with Markov inequality that

$$\mathbb{P}\left(\|x_{k+1} - x^*\|^2 \ge \frac{R^2}{4}\right) \le \frac{\mathbb{E}[\|x_{k+1} - x^*\|^2]}{R^2/4} \le \tau. \tag{110}$$

which indicates the event $\|x_{k+1} - x^*\| \le \frac{R}{2}$ holds with probability at least $1 - \tau$ for any $k \ge K$.

D ADDITIONAL NUMERICAL EXPERIMENTS

In this section, we provide additional numerical experiments to *(i)* demonstrate the validness of group sparsity identification of HSPG; *(ii)* provide comprehensive comparison to Prox-SG, RDA and Prox-SVRG on benchmark convex problems; and *(iii)* describe more details regarding our non-convex deep learning experiments shown in the main body.

D.1 LINEAR REGRESSION ON SYNTHETIC DATA

We first numerically validate the proposed HSPG on group sparsity identification by linear regression problems with ℓ_1/ℓ_2 regularizations using synthetic data. Consider a data matrix $A \in \mathbb{R}^{N\times n}$ consisting of N instances and the target variable $y \in \mathbb{R}^N$, we are interested in the following problem:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2N}\|Ax-y\|^2+\lambda\sum_{g\in\mathcal{G}}\|[x]_g\|. \tag{111}$$

Our goal is to empirically show that HSPG is able to identify the ground truth zero groups with synthetic data. We conduct the experiments as follows: *(i)* generate the data matrix A whose elements are uniformly distributed among $[-1,1]$; *(ii)* generate a vector x^* working as the ground truth solution, where the elements are uniformly distributed among $[-1,1]$ and the coordinates are equally divided into 10 groups ($|\mathcal{G}|=10$); *(iii)* randomly set a number of groups of x^* to be 0 according to a pre-specified group sparsity ratio; *(iv)* compute the target variable $y=Ax^*$; (v) solve the above problem (111) for x with A and y only, and then evaluate the Intersection over Union (IoU) with respect to the identities of the zero groups between the computed solution estimate $\hat{x}$ by HSPG and the ground truth x^*.

We test HSPG on (111) under different problem settings. For a slim matrix A where $N\geq n$, we test with various group sparsity ratios among $\{0.1,0.3,0.5,0.7,0.9\}$, and for a fat matrix A where $N<n$, we only test with a certain group sparsity value since a recovery of x^* requires that the number of non-zero elements in x^* is bounded by N. Throughout the experiments, we set λ to be $100/N$, the mini-batch size $|\mathcal{B}|$ to be 64, step size α_k to be 0.1 (constant), and fine-tune ϵ per problem. Based on a similar statistical test on objective function stationarity (Zhang et al., 2020), we switch to Half-Space Step roughly after 30 epoches. Table 1 shows that under each setting, the proposed HSPG correctly identifies the groups of zeros as indicated by IoU$(\hat{x},x^*)=1.0$, which is a strong evidence to show the correctness of group sparsity idenfitication of HSPG.

Table 1: Linear regression problem settings and IoU of the recovered solutions by HSPG.

	N	n	Group sparsity ratio of x^*	IoU$(\hat{x},x^*)$
Slim A	10000	1000	{0.1, 0.3, 0.5, 0.7, 0.9}	1.0
	10000	2000	{0.1, 0.3, 0.5, 0.7, 0.9}	1.0
	10000	3000	{0.1, 0.3, 0.5, 0.7, 0.9}	1.0
	10000	4000	{0.1, 0.3, 0.5, 0.7, 0.9}	1.0
Fat A	200	1000	0.9	1.0
	300	1000	0.8	1.0
	400	1000	0.7	1.0
	500	1000	0.6	1.0

D.2 LOGISTIC REGRESSION

We then focus on the benchmark convex logistic regression problem with the mixed ℓ_1/ℓ_2-regularization given N examples $(d_1,l_1),\cdots,(d_N,l_N)$ where $d_i\in\mathbb{R}^n$ and $l_i\in\{-1,1\}$ with the form

$$\underset{(x;b)\in\mathbb{R}^{n+1}}{\text{minimize}}\ \frac{1}{N}\sum_{i=1}^{N}\log(1+e^{-l_i(x^Td_i+b)})+\lambda\sum_{g\in\mathcal{G}}\|[x]_g\|, \tag{112}$$

for binary classification with a bias $b\in\mathbb{R}$. We set the regularization parameter λ as $100/N$ throughout the experiments since it yields high sparse solutions and low object value f's, equally decompose the variables into 10 groups to form $\mathcal{G}$, and test problem (112) on 8 standard publicly available

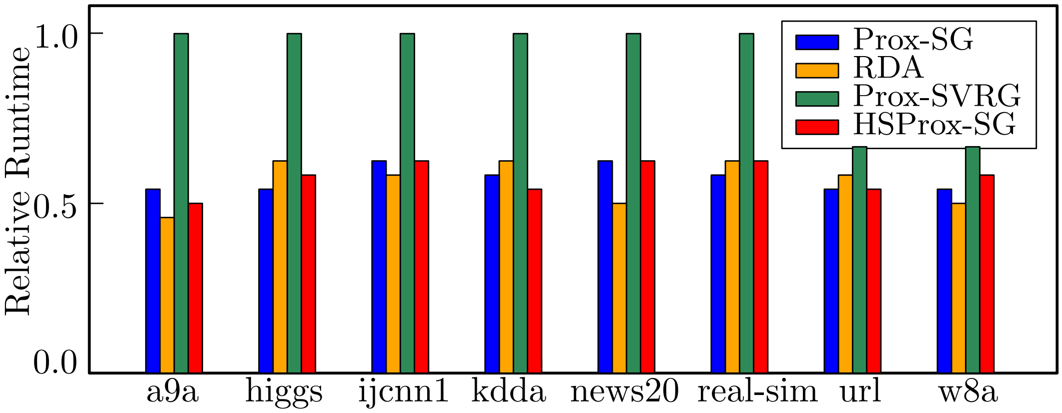

Figure 1: Relative runtime.

large-scale datasets from LIBSVM repository (Chang & Lin, 2011) as summarized in Table 2. All convex experiments are conducted on a 64-bit operating system with an Intel(R) Core(TM) i7-7700K CPU @ 4.20 GHz and 32 GB random-access memory.

We run the solvers with a maximum number of epochs as 60. The mini-batch size $|\mathcal{B}|$ is set to be $\min\{256, \lceil 0.01N \rceil\}$ similarly to (Yang et al., 2019). The step size α_k setting follows [Section 4](Xiao & Zhang, 2014). Particularly, we first compute a Lipschitz constant L as $\max_i \|d_i\|^2/4$, then fine tune and select constant $\alpha_k \equiv \alpha = 1/L$ to Prox-SG and Prox-SVRG since it exhibits the best results. For RDA, the step size parameter γ is fined tuned as the one with the best performance among all powers of 10. For HSPG, we set α_k as the same as Prox-SG and Prox-SVRG in practice. We set $N_{\mathcal{P}}$ as $30N/|\mathcal{B}|$ such that Half-Space Step is triggered after employing Prox-SG Step 30 epochs similarly to Appendix D.1, and the control parameter ϵ in (5) as 0.05. We select two ϵ's as 0 and 0.05. The final objective value Ψ and f, and group sparsity in the solutions are reported in Table 3-5, where we mark the best values as bold to facilitate the comparison. Furthermore, Figure 1 plots the relative runtime of these solvers for each dataset, scaled by the runtime of the most time-consuming solver.

Table 5 shows that our HSPG is definitely the best solver on exploring the group sparsity of the solutions. In fact, HSPG under $\epsilon = 0.05$ performs all the best except *ijcnn1*. Prox-SVRG is the second best solver on group sparsity exploration, which demonstrates that the variance reduction techniques works well in convex setting to promote sparsity, but not in non-convex settings. HSPG under $\epsilon = 0$ performs much better than Prox-SG which matches the better sparsity recovery property of HSPG as stated in Theorem 2 even under ϵ as 0. Moreover, as shown in Table 3 and 4, we observe that all solvers perform quite competitively in terms of final objective values (round up to 3 decimals) except RDA, which demonstrates that HSPG reaches comparable convergence as Prox-SG and Prox-SVRG in practice. Finally, Figure 1 indicates that Prox-SG, RDA and HSPG have similar computational cost to proceed, except Prox-SVRG due to its periodical full gradient computation.

Table 2: Summary of datasets.

Dataset	N	n	Attribute	Dataset	N	n	Attribute
a9a	32561	123	binary $\{0, 1\}$	news20	19996	1355191	unit-length
higgs	11000000	28	real $[-3, 41]$	real-sim	72309	20958	real [0, 1]
ijcnn1	49990	22	real [-1, 1]	url_combined	2396130	3231961	real $[-4, 9]$
kdda	8407752	20216830	real $[-1, 4]$	w8a	49749	300	binary $\{0, 1\}$

D.3 DEEP LEARNING EXPERIMENTS

We conduct all deep learning experiments on one GeForce GTX 1080 Ti GPU, and describe how to fine-tune the control parameter ϵ in (5) in details. According to Theorem 2, a larger ϵ results in a faster group sparsity identification, while by Lemma 1 on the other hand too large ϵ may cause a significant regression on the target objective Ψ value, *i.e.*, the Ψ value increases a lot. Hence, in our experiments, from the point of view of optimization, we search a proper ϵ in the following

Table 3: Final objective values Ψ for tested algorithms on convex problems.

Dataset	Prox-SG	RDA	Prox-SVRG	HSPG	
				ϵ as 0	ϵ as 0.05
a9a	**0.355**	0.359	**0.355**	**0.355**	**0.355**
higgs	**0.357**	0.360	0.365	0.358	0.358
ijcnn1	**0.248**	0.278	**0.248**	**0.248**	**0.248**
kdda	**0.103**	0.124	**0.103**	**0.103**	**0.103**
news20	**0.538**	0.693	**0.538**	**0.538**	**0.538**
real-sim	**0.242**	0.666	0.244	**0.242**	**0.242**
url_combined	0.397	0.579	**0.391**	0.405	0.405
w8a	**0.110**	0.111	0.112	**0.110**	**0.110**

Table 4: Final objective values f for tested algorithms on convex problems.

Dataset	Prox-SG	RDA	Prox-SVRG	HSPG	
				ϵ as 0	ϵ as 0.05
a9a	**0.329**	0.338	**0.329**	**0.329**	**0.329**
higgs	**0.357**	0.360	0.365	0.358	0.358
ijcnn1	**0.213**	0.270	**0.213**	**0.213**	0.214
kdda	**0.103**	0.124	**0.103**	**0.103**	**0.103**
news20	0.373	0.693	0.381	**0.372**	**0.372**
real-sim	**0.148**	0.665	0.159	**0.148**	**0.148**
url_combined	0.397	0.579	**0.391**	0.405	0.405
w8a	**0.089**	0.098	0.091	**0.089**	**0.089**

Table 5: Group sparsity for tested algorithms on convex problems.

Dataset	Prox-SG	RDA	Prox-SVRG	HSPG	
				ϵ as 0	ϵ as 0.05
a9a	20%	**30%**	**30%**	**30%**	**30%**
higgs	0%	10%	0%	0%	**30%**
ijcnn1	50%	**70%**	60%	60%	60%
kdda	0%	0%	0%	0%	**80%**
news20	20%	80%	**90%**	80%	**90%**
real-sim	0%	0%	**80%**	0%	**80%**
url_combined	0%	0%	0%	0%	**90%**
w8a	**0%**	**0%**	**0%**	**0%**	**0%**

ways: start from $\epsilon = 0.0$ and the models trained by employing $N_{\mathcal{P}}$ Prox-SG Steps, incrementally increase ϵ by 0.01 and check if the Ψ on the first Half-Space Step has an obvious increase, then accept the largest ϵ without regression on Ψ as our fine tuned ϵ shown in the main body of the paper. Particularly, the fine tuned ϵ's equal to 0.03, 0.05, 0.02 and 0.02 for VGG16 with CIFAR10, VGG16 with Fashion-MNIST, ResNet18 with CIFAR10 and ResNet18 with Fashion-MNIST respectively. Note from the perspective of different applications, there are different criterions to fine tune ϵ, *i.e.*, for model compression, we may accept ϵ based on the validation accuracy regression to reach higher group sparsity.

Additionally, we also report the final f comparison in Table 6 and its evolution on ResNet18 with CIFAR10 in Figure 2, where we can see that all tested algorithms can achieve competitive f values as they do in convex settings. And the evolution of f is similar to that of Ψ, *i.e.*, the raw objective f generally monotonically decreases for small $\epsilon = 0$ to 0.02, and experiences a mild pulse after switch to Half-Space Step for larger ϵ, *e.g.*, 0.05, which matches Lemma 1.

Table 6: Final objective values f for tested algorithms on non-convex problems.

Backbone	Dataset	Prox-SG	Prox-SVRG	HSPG	
				ϵ as 0	fine tuned ϵ
VGG16	CIFAR10	0.010	0.036	0.010	**0.009**
	Fashion-MNIST	0.181	**0.165**	0.181	0.182
ResNet18	CIFAR10	**0.001**	0.002	**0.001**	0.004
	Fashion-MNIST	0.006	0.008	**0.005**	0.010
MobileNetV1	CIFAR10	**0.021**	0.031	**0.021**	0.031
	Fashion-MNIST	0.074	**0.057**	0.074	0.088

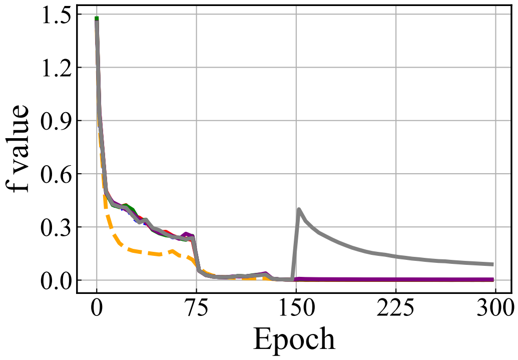

Figure 2: Evolution of f value on ResNet18 with CIFAR10.

REFERENCES

Chih-Chung Chang and Chih-Jen Lin. Libsvm: Data repository. 2011. URL `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Tianyi Chen. *A Fast Reduced-Space Algorithmic Framework for Sparse Optimization*. PhD thesis, Johns Hopkins University, 2018.

Dmitriy Drusvyatskiy and Adrian S Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Robert M. Gower. Convergence theorems for gradient descent. *University of Illinois, Urbana Champaign*, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. 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.

Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. In *Advances in neural information processing systems*, 2018.

Michael Mitzenmacher. Probability and computing-randomized algorithms and probabilistic analysis. *JOURNAL-OPERATIONAL RESEARCH SOCIETY*, 56(12):1454, 2005.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Minghan Yang, Andre Milzarek, Zaiwen Wen, and Tong Zhang. A stochastic extra-step quasi-newton method for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1910.09373*, 2019.

Pengchuan Zhang, Hunter Lang, Qiang Liu, and Lin Xiao. Statistical adaptive stochastic gradient methods. *arXiv preprint arXiv:2002.10597*, 2020.