Only Train Once: A One-Shot Neural Network Training And Pruning Framework

Tianyi Chen*
Microsoft
tiachen@microsoft.com

Bo Ji
National University of Singapore
jibo@comp.nus.edu.sg

Tianyu Ding
Johns Hopkins University
tding1@jhu.edu

Biyi Fang
Microsoft
bif@microsoft.com

Guanyi Wang
Georgia Institute of Technology
gwang93@gatech.edu

Zhihui Zhu
University of Denver
zhihui.zhu@du.edu

Luming Liang
Microsoft
lulian@microsoft.com

Yixin Shi
Microsoft
yixshi@microsoft.com

Sheng Yi
Microsoft
shengyi@microsoft.com

Xiao Tu
Microsoft
xiaotu@microsoft.com

Abstract

Structured pruning is a commonly used technique in deploying deep neural networks (DNNs) onto resource-constrained devices. However, the existing pruning methods are usually heuristic, task-specified, and require an extra fine-tuning procedure. To overcome these limitations, we propose a framework that compresses DNNs into slimmer architectures with competitive performances and significant FLOPs reductions by Only-Train-Once (OTO). OTO contains two keys: (i) we partition the parameters of DNNs into zero-invariant groups, enabling us to prune zero groups without affecting the output; and (ii) to promote zero groups, we then formulate a structured-sparsity optimization problem and propose a novel optimization algorithm, Half-Space Stochastic Projected Gradient (HSPG), to solve it, which outperforms the standard proximal methods on group sparsity exploration and maintains comparable convergence. To demonstrate the effectiveness of OTO, we train and compress full models simultaneously from scratch without fine-tuning for inference speedup and parameter reduction, and achieve state-of-the-art results on VGG16 for CIFAR10, ResNet50 for CIFAR10 and Bert for SQuAD and competitive result on ResNet50 for ImageNet. The source code is available at https://github.com/tianyic/only_train_once.

1 Introduction

Deep neural networks (DNNs) have been shown to be effective in various real applications (51; 28). It is widely acknowledged that large-scale DNN models not only learn faster but also outperform their slimmer counterparts. However, such heavy models pose a great challenge to the deployment stage due to their resource-consuming nature. In addressing this issue, many model compression

*Corresponding author.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

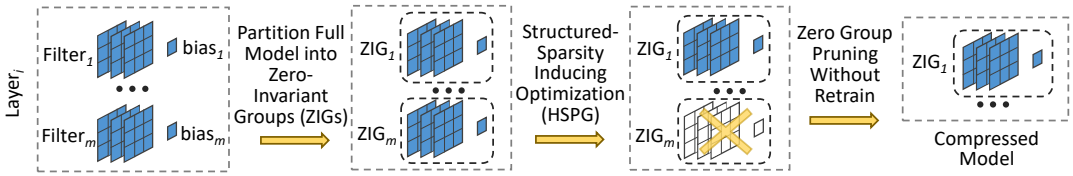


Figure 1: Overview of OTO. Without loss of generality, we illustrate OTO on a model with only vanilla convolutional layers, and for simplicity we only show Layer$_i$ with m 3D filters and their biases. The key to its success is twofold: (i) identify and partition the parameters of the model into zero-invariant groups (ZIGs); and (ii) solve the structured-sparsity regularization problem using HSPG. Finally, we obtain the compressed model by directly pruning the zero groups, *i.e.*, ZIG$_m$.

techniques (5; 11) are proposed in the past decade that aim at compressing those large and complex models into slimmer and simpler ones while suffering negligible loss in performance.

Pruning methods as one of the main categories of model compression, focus on identifying and pruning redundant structures via various mechanisms to achieve a slimmer architecture, and thus improve the interpretability of a DNN model (26; 11; 65). For example, (32; 33) adopt fine-grained pruning via ℓ_1 or ℓ_2 regularization, which prune the small-weight connections based on some hard threshold. (36; 57; 60) measure the importance of filters to accelerate the networks by removing insignificant feature maps. (37; 7) utilize reinforcement learning agent to predict compression action.

Nevertheless, many of the existing pruning methods (i) often rely on criteria based on heuristics or empirical cues, *e.g.*, magnitude of a connection weight and importance score of a filter, to identify redundant parameters, which may cause divergence during optimization; (ii) thus require complex multi-stage pipelines that involve either a retraining or fine-tuning procedure to regain the accuracy during constructing a slimmer model, which is time-consuming; and (iii) are specific to certain architectures or applications, and are consequently less applicable to various downstream scenarios. Recently, there have been a few efforts (14; 58; 8) to directly train the network with sparsity inducing regularizers, which provide generality and convergence guarantee. However, these approaches focus on either merely the individual sparsity of the parameters or the group sparsity of the filters, and thus cannot directly remove those zero components (still require subsequent fine-tuning) since the zeros are entangled with other commonly used components, *e.g.*, bias, batch normalization or skip connection. Furthermore, the optimization algorithms used in (14; 58) lack sufficient capability to explore (group) sparsity in DNNs effectively and require a post-processing step to yield exact zeros.

In this paper, we overcome the above limitations of existing pruning methods by proposing a *one-shot* neural network pruning framework, with which we are able to train a full heavy model from scratch only once, and obtain a slim architecture without fine-tuning while maintain high performance. As shown in Figure 1, the key to its success is twofold: (i) we identify and partition the parameters of DNNs into zero-invariant groups (ZIGs), enabling us to prune redundant structures according to zero groups without affecting the output of the network; and (ii) to promote zero groups, we formulate the pruning task as a structured-sparsity optimization problem and propose a novel optimization method, Half-Space Stochastic Projected Gradient (HSPG), to solve it, which outperforms the standard proximal methods on sparsity exploration and maintains comparable convergence. We highlight that both zero-invariant group partition and the novel optimization algorithm in promoting zero group lead to achieve one-shot neural network training and pruning regardless of its architecture.

Our main contributions are summarized as follows.

- **One-Shot Training and Pruning.** We propose OTO, a training and pruning framework that compresses a full neural network with competitive performance by Only-Train-Once, thereby one-shot. OTO dramatically simplifies the complex multi-stage training pipelines of the existing pruning approaches, fits various architectures and applications, and hence is generic and efficient.
- **Zero-Invariant Group.** We define zero-invariant groups for neural networks. If a network is partitioned into ZIGs, it allows us to prune the zero groups without affecting the output, which results in one-shot pruning. Such property is applicable to various popular structures from plain fully connected layers to sophisticated ones such as residual blocks and multi-head attention.
- **Novel Structured-Sparsity Optimization Algorithm.** We propose Half-Space Stochastic Projected Gradient (HSPG), a method that solves structured-sparsity inducing regularization problem. We show and analyze the superiority of HSPG in promoting zero groups of networks than the

standard proximal methods and the competitive objective convergence in practice. The fact that ZIG and HSPG are designed agnostic to networks makes OTO generic to various applications.

- **Experimental Results.** We train and compress full models simultaneously from scratch without fine-tuning for inference speedup and parameter reduction, and achieve state-of-the-art results on compression benchmark VGG for CIFAR10, ResNet50 for CIFAR10/ImageNet, Bert for SQuAD.

2 Related Work

Structured pruning focuses on identifying and pruning the redundant structures in a full model to achieve slimmer architectures for efficient model inference and storage (26; 32), where there have been numerous efforts dedicated. For CNN compression, the general procedure can be largely summarized as: (*i*) train a full model; (*ii*) identify and prune the redundant structures to build a slimmer model based on various criteria, including (structured) sparsity (58; 85; 14; 56; 102; 27; 102; 62; 91), Bayesian pruning (101; 65; 59; 81), ranking importance (54; 60; 41; 36; 57; 100), reinforcement learning (37; 7), adversarial robustness (76), scientific control (79), lottery ticket (23; 24; 72), joint quantization learning (80; 90), etc.; (*iii*) retrain or iteratively fine-tune the slimmer model to regain the accuracy regression during pruning. These methods cannot avoid the extra and usually time-consuming fine-tuning step because the identified redundant structures, even parametrized with zeros, actually contribute to the model output, thereby additional fine-tuning step is an absolute necessity.

For pruning Bert (82), knowledge distillation (40) and LayerDropout (21) shorten Bert by reducing the number of layers directly. Other methods (29; 75; 30) build slimmer Berts in the manner of individual sparsity, but require specially designed data structure for storage and computing library to take advantage of sparse data (31; 10), and typically cannot achieve inference speedup against the highly optimized library (16) for dense model due to the discontiguous memory allocation (9).

The structured sparsity for weight pruning is the most relevant to the algorithm described in this paper. The existing structure learning works (58; 85; 14; 56; 102) have the respective disadvantages: *(i)* multiple trainings during the whole procedure since their group partition cannot isolate the impact of pruned structures to the model output; and *(ii)* heuristic post-processing to generate zero groups as the standard proximal methods (19; 87; 88; 12) and ADMM (100; 58; 4) defective on the sparsity exploration for deep learning (8), which may deteriorate the performance of the model significantly.

Avoiding fine-tuning step during the whole pruning procedure is receiving more and more attentions because of its efficiency. In particular, SNIP (52) and GraSP (83) identify redundancy via salience scores at the initialization stage to construct pruned structures, then train the pruned models by the standard optimizers. SCP (48) isolates the impact of batch normalization, while lacks the consideration of more general DNN architectures.

3 OTO

In essence, OTO frames the network training and pruning as a structure learning problem. Given a full model $\mathcal{M}$, OTO trains and compresses it simultaneously from scratch *without* fine-tuning, and achieves significant reduction in both FLOPs and parameters. Particularly, as stated in Algorithm 1, the trainable parameters of $\mathcal{M}$ are firstly partitioned into a ZIG set $\mathcal{G}$ (Section 3.1). We then construct and solve a structured-sparsity inducing optimization problem (Section 3.2) by proposed stochastic optimizer (HSPG) to seek a highly group-sparse solution $\boldsymbol{x}^*_{\text{HSPG}}$ (Section 3.3). Lastly, we obtain a compressed model $\mathcal{M}^*$ by directly pruning these zero groups (Section 3.4).

Algorithm 1 Outline of OTO.

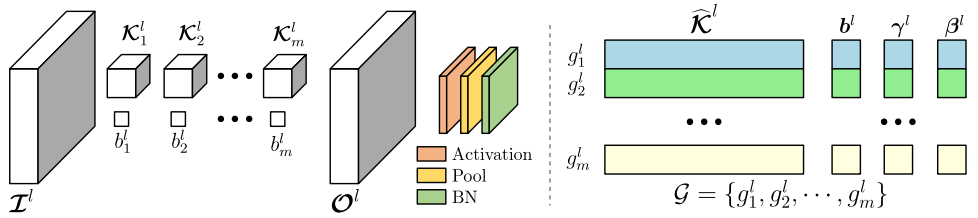
1: **Input:** Full model $\mathcal{M}$ (no need to be pretrained).
2: **Construct $\mathcal{G}$:** Partition the trainable parameters of $\mathcal{M}$ into a ZIG set $\mathcal{G}$.
3: **Train:** Train the model $\mathcal{M}$ using HSPG (Algorithm. 2) to obtain a group-sparse solution $\boldsymbol{x}^*_{\text{HSPG}}$.
4: **Prune:** Construct a slimmer model architecture $\mathcal{M}^*$ by directly pruning zero groups of $\boldsymbol{x}^*_{\text{HSPG}}$.
5: **Output:** Compressed model $\mathcal{M}^*$.

3.1 Zero-Invariant Group

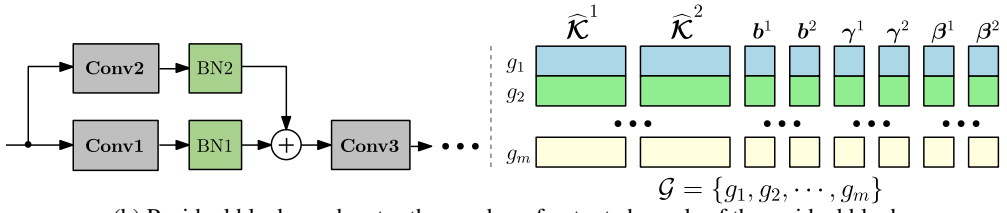
The root cause of the existing methods having multi-stage training pipeline is that despite the pruned structure (*e.g.*, 3D filter) being zeros, its associated structures (*e.g.*, non-zero bias) still contribute to its corresponding output to the next layer (*e.g.*, feature map). As a result, the model accuracy regresses, hence an extra step of fine-tuning is necessary. OTO avoids the necessity by partitioning the parameters of DNNs into a set of so-called zero-invariant groups (ZIGs) $\mathcal{G}$ defined as follows.

Definition 1 (Zero-Invariant Groups (ZIGs))**.** *For a layer-wise DNN, we partition its entire trainable parameters into disjoint groups $\mathcal{G} = \{g\}$. Then we call $\mathcal{G}$ zero-invariant groups (ZIGs) if each group $g \in \mathcal{G}$ is zero-invariant in the sense that all of the parameters in g being zeros results in its corresponding output to the next layer to be zeros as well.*

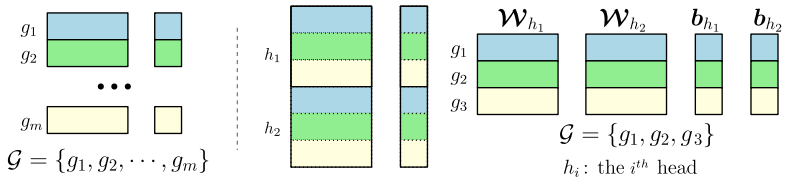
In effect, if and only if a DNN model is partitioned into a ZIG set $\mathcal{G}$ and one or more of its element g are parameterized by zeros, the entire corresponding structures contribute none to the model outputs and hence can be pruned directly. Such partition is applicable to various structures of DNN models. Without loss of generality, we define and describe ZIG partition for three most popular structures: *(i)* Conv-BN, *(ii)* Residual Block, and *(iii)* Fully Connected and Multi-Head Attention Layer.



(a) Conv-BN. m denotes the number of channels in $\mathcal{O}^l$.



(b) Residual block. m denotes the number of output channels of the residual block.



(c) Fully connected layer (Left). Multi-head attention layer (Right). m denotes the length of output vector.

Figure 2: Zero-invariant group partition for three popular structures.

ZIG of Conv-BN. Convolutional layer (Conv) followed by batch-normalization layer (BN) is extensively used in DNN models. Figure 2a shows the ZIG partition for Conv-BN. The 4D filter tensor $\boldsymbol{\kappa}^l$ is flattened into a filter matrix $\hat{\boldsymbol{\kappa}}^l$. During the forward pass, the input tensor $\boldsymbol{\mathcal{I}}^l$ is transformed into the output tensor $\boldsymbol{\mathcal{O}}^l$ of Conv and then into the input tensor of the $(l+1)^{th}$ layer $\boldsymbol{\mathcal{I}}^{l+1}$ by

$$\boldsymbol{\mathcal{O}}^l \leftarrow \boldsymbol{\mathcal{I}}^l \otimes \hat{\boldsymbol{\mathcal{K}}}^l + \boldsymbol{b}^l,\ \boldsymbol{\mathcal{I}}^{l+1} \leftarrow \frac{a(\boldsymbol{\mathcal{O}}^l) - \boldsymbol{\mu}^l}{\boldsymbol{\sigma}^l} \odot \boldsymbol{\gamma}^l + \boldsymbol{\beta}^l, \tag{1}$$

where denoted by $\otimes$ the convolutional operation, $\odot$ the element-wise multiplication and $a(\cdot)$ the activation function. BN is parameterized by mean $\boldsymbol{\mu}^l$, standard deviation $\boldsymbol{\sigma}^l$, weight $\boldsymbol{\gamma}^l$ and bias $\boldsymbol{\beta}^l$ respectively. The activation function needs to be zero-invariant, *i.e.*, $a(\mathbf{0}) = \mathbf{0}$, where most instances satisfy, *e.g.*, ReLU (25), PReLU (34), GELU (39) and LeakyReLU (89). Hence, each row of the flattened filter matrix $\hat{\boldsymbol{\kappa}}^l$ and its bias $\boldsymbol{b}^l$ belong to one ZIG because they being zeros results in their corresponding channel of $\boldsymbol{\mathcal{O}}^l$ (*i.e.*, feature map) to be zeros as well. Subsequently, $\boldsymbol{\gamma}^l$ and $\boldsymbol{\beta}^l$ of this corresponding channel in BN are also included into this ZIG to avoid the value shift (zero to

non-zero) during normalization. Note that grouping these four sets of parameters channel-wisely makes Conv-BN zero-invariant regardless of the value of $\boldsymbol{\mu}^l$ and $\boldsymbol{\sigma}^l$, and hence they are excluded from the ZIG. For illustration, each ZIG is highlighted in the same color (*e.g.*, g_1^l in blue).

ZIG of Residual Block. The residual block adds another layer of challenge because its output tensor is the summation of the outputs of two Conv-BNs. Figure 2b shows the ZIG partition for the residual block. As illustrated, before propagated to Conv3, the outputs of Conv1-BN1 and Conv2-BN2 are summarized and hence share the same dimension. As such, to make residual block zero-invariant, we group the four sets of parameters channel-wisely of both Conv1-BN1 and Conv2-BN2 into ZIGs, *i.e.*, each row of $\hat{\boldsymbol{\kappa}}^1$, $\boldsymbol{b}^1$, $\boldsymbol{\gamma}^1$, $\boldsymbol{\beta}^1$ of Conv1-BN1 and each row of $\hat{\boldsymbol{\kappa}}^2$, $\boldsymbol{b}^2$, $\boldsymbol{\gamma}^2$, $\boldsymbol{\beta}^2$ of Conv2-BN2. In Appendix A.1, we describe the zero-invariant group partition of ResNet50 in greater detail.

ZIG of Fully Connected and Multi-Head Attention Layer. Figure 2c shows the ZIG partition for fully connected and multi-head attention layer. Particularly, we partition each row of weight matrix and its associated bias into a ZIG, and therefore any input element is turned to zero if that ZIG is parameterized with zeros, making the fully connected layer zero-invariant. Multi-head attention layer is the key building block of the transformer architectures (82). Its trainable parameters contain a weight matrix and bias vector, consisting of the sub-matrix and sub-vector of each head (we use two heads as an example). We form ZIG by grouping each row of every sub-matrix and sub-vector, *i.e.*, each row of $\boldsymbol{\mathcal{W}}_{\boldsymbol{h}_1}$, $\boldsymbol{b}_{\boldsymbol{h}_1}$, $\boldsymbol{\mathcal{W}}_{\boldsymbol{h}_2}$ and $\boldsymbol{b}_{\boldsymbol{h}_2}$ of h_1 and h_2, respectively.

Automatic ZIG Partition. Based on the above illustrating examples, we provide prescribed ZIG partition for the tested DNNs in Section 4. Furthermore, given an arbitrary DNN architecture, the procedure of partitioning variables into ZIGs could be automatically proceeded, wherein the key would be identifying the connections among various layers. We will leave the automatic ZIG partition for arbitrary DNNs as future work.

3.2 Structured-Sparsity Regularization

We now formulate a structured-sparsity regularization problem over the ZIG set $\mathcal{G}$ for the trainable parameters of the full model $\mathcal{M}$ as follows

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ \psi(\boldsymbol{x}) := f(\boldsymbol{x}) + \lambda r(\boldsymbol{x}),\ r(\boldsymbol{x}) := \sum_{g\in\mathcal{G}} \|[\boldsymbol{x}]_g\|\,, \tag{2}$$

where $\lambda > 0$ is a weighting coefficient, $f(\boldsymbol{x})$ is a task-specific loss function, and $r(\boldsymbol{x})$ is an augmented structured-sparsity inducing regularization term encoding the topological structure of $\mathcal{M}$ over $\mathcal{G}$. A larger λ typically results in a higher group sparsity while sacrifices more on the bias of model estimation. We aim at computing a local optimum to achieve both low loss and high group sparsity.

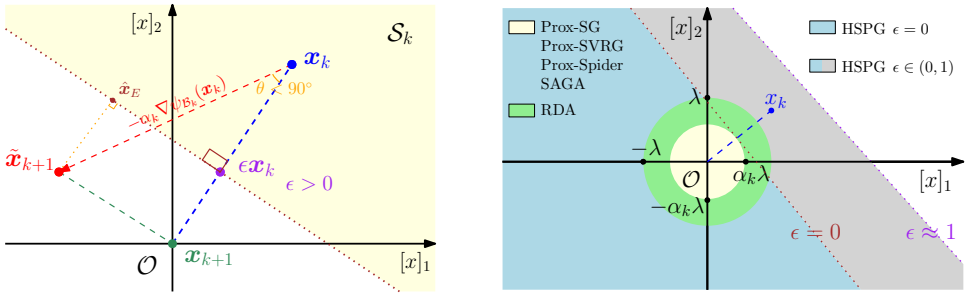
To induce group sparsity onto the solution of (2), there exist several candidates for $r(\boldsymbol{x})$, including mixed ℓ_1/ℓ_p norm ($p > 1$) (1; 20) and group Minmax Concave Penalty (MCP) (96). Among these candidates, the mixed ℓ_1/ℓ_2 norm as defined in (2) is arguably the most popular choice in classical machine learning applications (1; 92), where $\|\cdot\|$ is the ℓ_2-norm, and each component $g \in \mathcal{G}$ indexes a group of variables. In this paper, we will demonstrate the effectiveness of OTO by selecting $r(\boldsymbol{x})$ as the mixed ℓ_1/ℓ_2 norm. We highlight OTO is applicable for other group sparsity regularizers as well.

3.3 Half-Space Stochastic Projected Gradient (HSPG)

To solve the non-smooth regularization problem as (2) in deep learning applications, the standard proximal method and the ADMM lack capability to effectively identify group sparsity; see the discussions later in this Section. Therefore, we propose a novel stochastic optimization algorithm so-called Half-Space Stochastic Projected Gradient (HSPG) to enhance the group sparsity exploration more effectively than the classical methods while maintain a similar convergence property.

Outline. We state the outline of HSPG in Algorithm 2. It contains two stages: Initialization Stage and Group-Sparsity Stage. The first Initialization Stage employs Stochastic Gradient Descent (SGD) step to search for a good but usually non-sparse solution estimate. Then the second stage proceeds Half-Space step started with the non-sparse iterate to effectively exploit the group sparsity within a sequence of reduced spaces and converges to the group-sparse solutions. Half-Space step performs SGD update along with a novel projection operator so-called Half-Space Projection, which significantly outperforms the standard proximal operators on sparsity exploration.

Initialization Stage. The Initialization Stage performs the vanilla SGD to find a good initial point for the subsequent Group-Sparsity Stage. At k^{th} iteration, a stochastic gradient of f, *e.g.*, based on a mini-batch, is generated denoted as $\nabla \tilde{f}$. Since the group sparsity inducing regularizer $r(\boldsymbol{x})$ in the form as (2) is non-smooth, we select a subgradient $\zeta(\boldsymbol{x}_k)$ from its subdifferential $\partial r(\boldsymbol{x}_k)$ to form a stochastic subgradient of $\psi(\boldsymbol{x}_k)$ as $\nu(\boldsymbol{x}_k) := \nabla \tilde{f}(\boldsymbol{x}_k) + \lambda \zeta(\boldsymbol{x}_k)$. We then compute the next iterate as $\boldsymbol{x}_{k+1} := \boldsymbol{x}_k - \alpha_k \nu(\boldsymbol{x}_k)$ by subgradient descent update.



(a) Half-Space Projection

(b) Projection Region For Mixed ℓ_1/ℓ_2 Regularization

Figure 3: Illustration of Half-Space Step with projection in (6), where $\mathcal{G} = \{\{1, 2\}\}$.

Group-Sparsity Stage. The Group-Sparsity Stage is designed to effectively determine the groups of zero variables and capitalize convergence characteristic, which is in sharp contrast to other heuristic aggressive weight pruning methods that typically lack theoretical guarantees (55; 60). The intuition of Half-Space Step is to project $[\boldsymbol{x}_k]_g$ to zero only if $-[\boldsymbol{x}_k]_g$ serves as a descent step to $\psi(\boldsymbol{x}_k)$, *i.e.*, $-[\boldsymbol{x}_k]_g^\top [\nabla\psi(\boldsymbol{x}_k))]_g < 0$, hence updating $[\boldsymbol{x}_{k+1}]_g \leftarrow [\boldsymbol{x}_k]_g - [\boldsymbol{x}_k]_g = 0$ still results in some progress to the optimality. In particular, we first define the following index sets for any $\boldsymbol{x} \in \mathbb{R}^n$:

$$\mathcal{I}^0(\boldsymbol{x}) := \{g : g \in \mathcal{G}, [\boldsymbol{x}]_g = 0\} \text{ and } \mathcal{I}^{\neq 0}(\boldsymbol{x}) := \{g : g \in \mathcal{G}, [\boldsymbol{x}]_g \neq 0\}, \tag{3}$$

where $\mathcal{I}^0(\boldsymbol{x})$ represents the indices of groups of zero variables at $\boldsymbol{x}$, and $\mathcal{I}^{\neq 0}(\boldsymbol{x})$ indexes the groups of nonzero variables at $\boldsymbol{x}$. To proceed, we further define an artificial set that $\boldsymbol{x}$ lies in:

$$\mathcal{S}(\boldsymbol{x}) := \{\boldsymbol{0}\} \bigcup \{\boldsymbol{z} \in \mathbb{R}^n : [\boldsymbol{z}]_g = \boldsymbol{0} \text{ if } g \in \mathcal{I}^0(\boldsymbol{x}), \text{and } [\boldsymbol{z}]_g^\top [\boldsymbol{x}]_g \geq \epsilon \, \|[\boldsymbol{x}]_g\|^2 \text{ if } g \in \mathcal{I}^{\neq 0}(\boldsymbol{x})\}, \tag{4}$$

which consists of half-spaces and the origin. Here the parameter $\epsilon \geq 0$ controls how aggressively we promote group sparsity, and is typically fixed as zero in practice. Hence, $\boldsymbol{x} \in \mathcal{S}_k := \mathcal{S}(\boldsymbol{x}_k)$ only if: *(i)* $[\boldsymbol{x}]_g$ lies in the upper half-space for all $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$ for some prescribed $\epsilon \in [0, 1)$ as shown in Figure 3a; and *(ii)* $[\boldsymbol{x}]_g$ equals to zero for all $g \in \mathcal{I}^0(\boldsymbol{x}_k)$. Intuitively, $\mathcal{S}_k$ establishes the region where important structures inhabit, thereby redundant structures vanish if falling outside.

Ideally, the Initialization Stage has produced reasonably well but typically non-sparse iterate $\boldsymbol{x}_k$ nearby a group-sparse solution $\boldsymbol{x}^*$ of problem (2), , *i.e.*, the optimal distance $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ is sufficiently small. As seen in Appendix B, it further indicates that the group-sparse optimal solution $\boldsymbol{x}^*$ inhabits $\mathcal{S}_k$, and $\mathcal{S}_k$ has already covered the group-support of $\boldsymbol{x}^*$, *i.e.*, $\mathcal{I}^{\neq 0}(\boldsymbol{x}^*) \subseteq \mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$. Our goal now becomes minimizing $\psi(\boldsymbol{x})$ over $\mathcal{S}_k$ to identify the remaining zero groups, *i.e.*, $\mathcal{I}^0(\boldsymbol{x}^*)/\mathcal{I}^0(\boldsymbol{x}_k)$, which is formulated as the following problem:

$$\underset{\boldsymbol{x} \in \mathcal{S}_k}{\text{minimize}} \; \psi(\boldsymbol{x}) = f(\boldsymbol{x}) + \lambda r(\boldsymbol{x}). \tag{5}$$

The next iterate $\boldsymbol{x}_{k+1}$ is computed as an solution estimate of problem (5).

Algorithm 2 Outline of HSPG for solving (2).

Input: $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\alpha_0 > 0, \epsilon \in [0, 1)$, and $N \in \mathbb{Z}^+$.
Output: a group-sparse solution $\boldsymbol{x}^*_{\text{HSPG}}$ from $\{\boldsymbol{x}_k\}$.
for $k = 0, 1, 2, \ldots$ **do**
  Compute a stochastic subgradient $\nu(\boldsymbol{x}_k)$ of $\psi(\boldsymbol{x}_k)$.
  if $k < N$ **then**
    Subgradient Descent Update:
    Set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k - \alpha_k \nu(\boldsymbol{x}_k)$.
  else
    Half-Space Update:
    Set a trial iterate $\tilde{\boldsymbol{x}}_{k+1}$ as
    $[\tilde{\boldsymbol{x}}_{k+1}]_{\mathcal{I}^{\neq 0}(\boldsymbol{x}_k)} \leftarrow [\boldsymbol{x}_k - \alpha_k \nu(\boldsymbol{x}_k)]_{\mathcal{I}^{\neq 0}(\boldsymbol{x}_k)}$
    $[\tilde{\boldsymbol{x}}_{k+1}]_{\mathcal{I}^0(\boldsymbol{x}_k)} \leftarrow \boldsymbol{0}$.
    for each group g in $\mathcal{G}$ **do**
      $[\boldsymbol{x}_{k+1}]_g \leftarrow [\text{Proj}^{HS}_{\mathcal{S}_k}(\tilde{\boldsymbol{x}}_{k+1})]_g$.
  Update α_{k+1}.

Particularly, in Algorithm 2, $[\boldsymbol{x}_{k+1}]_{\mathcal{I}^0(\boldsymbol{x}_k)} \equiv \mathbf{0}$ will not be updated, and only the entries in $\mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$ are free to move. Hence $\psi(\boldsymbol{x})$ is smooth on $\mathcal{S}_k$, and (5) is a reduced space optimization problem. A standard way to solve problem (5) would be the stochastic gradient descent equipped with Euclidean projection (68). However, such a projected method rarely produces zero (group) variables, as the dense Euclidean projected point $\hat{\boldsymbol{x}}_E \neq \mathbf{0}$ illustrated in Figure 3a. To address, we introduce a novel half-space projection operator to effectively project an entire group of variables to zeros.

As line 4 and 9-12 in Algorithm 2, we first approximate the (sub)gradient of ψ on the free variables by $[\nu(\boldsymbol{x}_k)]_{\mathcal{I}^{\neq 0}(\boldsymbol{x}_k)}$, then employ gradient descent over $\mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$ to compute a trial point $\widetilde{\boldsymbol{x}}_{k+1}$ which is passed into a fresh half-space projection operator $\mathrm{Proj}_{\mathcal{S}_k}^{HS}(\cdot)$ defined as

$$\left[\mathrm{Proj}_{\mathcal{S}_k}^{HS}(\boldsymbol{z})\right]_g := \begin{cases} 0 & \text{if } [\boldsymbol{z}]_g^\top [\boldsymbol{x}_k]_g < \epsilon \left\|[\boldsymbol{x}_k]_g\right\|^2, \\ [\boldsymbol{z}]_g & \text{otherwise.} \end{cases} \tag{6}$$

The above projector of form (6) is not the standard one in Euclidean sense[2], and it has two advantages: *(i)* the actual search direction $\boldsymbol{d}_k := (\mathrm{Proj}_{\mathcal{S}_k}^{HS}(\tilde{\boldsymbol{x}}_{k+1}) - \boldsymbol{x}_k)/\alpha_k$ performs as a descent direction to $\psi(\boldsymbol{x}_k)$, *i.e.*, $[\boldsymbol{d}_k]_g^\top [\nu(\boldsymbol{x}_k))]_g < 0$ as $\theta < 90^\circ$ in Figure 3a, hence the progress to the optimum is made via the sufficient decrease property drawn as Lemma 1 in Appendix B; then *(ii)* it effectively projects entire groups of variables to zero if the inner product of corresponding entries is sufficiently small. In contrast, the Euclidean projection operator is far away effective to promote group sparsity.

Superiority of HSPG on Group Sparsity Identification. We now intuitively illustrate the strength of HSPG on group sparsity exploration. In fact, the half-space projection (6) is a more effective sparsity promotion mechanism compared to the standard proximal methods. Particularly, it benefits from a much larger projection region to map a reference point $\hat{\boldsymbol{x}}_{k+1} := \boldsymbol{x}_k - \alpha_k \nabla \tilde{f}(\boldsymbol{x}_k)$ or its variants to zero. As the 2D case described in Figure 3b, the projection regions of the state-of-the-art Prox-SG (19), Prox-SVRG (88), Prox-Spider (97) and SAGA (12) for (2) are ℓ_2-balls with radius as $\alpha_k \lambda$. In deep learning applications, the step size α_k is usually selected around 10^{-3} to 10^{-4} or even smaller for convergence. Together with the common setting of $\lambda \ll 1$, their projection regions would vanish rapidly, resulting in the difficulties to produce group sparsity. As a sharp contrast, even though $\alpha_k \lambda$ is near zero, the projection region of HSPG $\{\boldsymbol{x} : \boldsymbol{x}_k^\top \boldsymbol{x} < (\alpha_k \lambda + \epsilon \left\|\boldsymbol{x}_k\right\|) \left\|\boldsymbol{x}_k\right\|\}$ (seen in Appendix B) is still an open half-space which contains those ℓ_2 balls as well as RDA (87)'s if ϵ is large enough. Conversely, vanilla ADMM alone lacks the mechanism to project a group of variables to zero, unless equips with extra post-processing step (100; 58). In Appendix B, we further reveal that HSPG still maintains the convergence to the optimality as drawn in Theorem 1. Moreover, we numerically demonstrate the superiority of HSPG in the sense of optimization in Appendix C.

3.4 Pruning Without Fine-Tuning

The group-sparse solution $\boldsymbol{x}^*_{\text{HSPG}}$ over ZIGs to the full model $\mathcal{M}$ is leveraged to construct the slimmer model $\mathcal{M}^*$. Particularly, we prune the redundant structures identified as zero groups $\mathcal{I}^0$ and retain non-zero groups $\mathcal{I}^{\neq 0}$ in $\boldsymbol{x}^*_{\text{HSPG}}$. Because the parameters of full model are partitioned into ZIGs, the pruned structures contribute none to the model output. Therefore, given the same input, the slimmer model $\mathcal{M}^*$ computes the identical output as the full model $\mathcal{M}$ parameterized with $\boldsymbol{x}^*_{\text{HSPG}}$.

4 Experiment

In this section, we numerically demonstrate the effectiveness of OTO by one-shot training and pruning without fine-tuning on several benchmark compression tasks for CNNs, *i.e.*, VGG16 (77) for CIFAR10 (49) and ResNet50 (35) for CIFAR10 (49) and ImagetNet (ILSVRC2012) (15). We also verify the scalibility of OTO onto Bert (82) evaluated on SQuAD (69). All datasets are free to academic usage and do not contain personally identifiable information or offensive content. CIFAR10 is under the MIT license, consisting of 50,000 training and 10,000 test images from 10 classes. ImagetNet is a large-scale dataset without license and contains about 1.2 million and 50,000 images in training and validation sets from 1,000 classes. SQuAD is under the CC BY-SA 4.0 license with about 100,000 question/answer pairs splitted into train/dev/test sets as (80/10/10%). We conduct all experiments on a Nvidia RTX8000 GPU and provide implementation details in Appendix A.

[2]Note that when $r(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ where each $g \in \mathcal{G}$ is singleton, then $\mathcal{S}_k$ becomes an orthant face (8).

Table 1: VGG16 and VGG16-BN for CIFAR10. Convolutional layers are in bold.

Method	BN	Architecture	FLOPs	# of Params	Top-1 Acc.
Baseline	✗	**64-64-128-128-256-256-256-512-512-512-512-512-512**-512-512	100%	100%	91.6%
SBP (65)	✗	**47-50-91-115-227-160-50-72-51-12-34-39-20**-20-272	31.1%	5.9%	**91.0%**
BC (59)	✗	**51-62-125-128-228-129-38-13-9-6-5-6-6**-6-20	38.5%	5.4%	**91.0%**
RBC (101)	✗	**43-62-120-120-182-113-40-12-20-11-6-9-10**-10-22	32.3%	3.9%	90.5%
RBP (101)	✗	**50-63-123-108-104-57-23-14-9-8-6-7-11**-11-12	28.6%	2.6%	**91.0%**
OTO	✗	**21-45-82-110-109-68-37-13-9-7-3-5-8**-170-344	**16.3%**	**2.5%**	**91.0%**
Baseline	✓	**64-64-128-128-256-256-256-512-512-512-512-512-512**-512-512	100%	100%	93.2%
EC (55)	✓	**32-64-128-128-256-256-256-256-256-256-256-256-256**-512-512	65.8%	37.0%	93.1%
Hinge (56)	✓	–	60.9%	20.0%	93.6%
SCP (48)	✓	–	33.8%	7.0%	**93.8%**
OTO	✓	**22-56-93-123-182-125-95-45-27-21-10-13-19**-244-392	**26.8%**	**5.5%**	93.3%

4.1 Deep Convolutional Neural Network

The results on CNN experiments are summarized in Table 1, 2 and 4. In particular, we compare OTO to its state-of-the-art counterparts by Top-1/5 accuracy, remaining FLOPs and parameters against the corresponding baseline (full model). We report the numbers of other methods based on the corresponding literature and leave as '-' if not reported. The best pruning results are marked as bold.

VGG16 for CIFAR10. We consider the standard VGG16 and the version with batch normalization layer after each convolutional layer, referred to as VGG16-BN. OTO partitions the parameters into ZIGs following Section 3.1, then trains and prunes the model via HSPG, and finally constructs the slimmer model without fine-tuning. For VGG16, as shown in Table 1, the pruned architecture of OTO indicates that OTO identifies similar redundancy of the intermediate and late convolutional layers compared to other methods, but significantly more of the early convolutional layers. As a result, OTO achieves 83.7% $(1 - 16.3\%)$ FLOPs reduction and 97.5% $(1 - 2.5\%)$ parameter reduction with the best Top-1 accuracy, which outperforms other state-of-the-arts significantly. For VGG16-BN, among all, OTO reduces FLOPs and parameters to the lowest 26.8% and 5.5%, respectively. EC (55) and Hinge (56) achieve the same level of Top-1 accuracy as OTO, but are substantially outperformed when it comes to FLOPs and parameter reduction. In Table 8 of Appendix A.4, we further present the FLOPs reductions per layer of OTO.

ResNet50 for CIFAR10. Since OTO is able to automatically learn a slimmer model of high performance, we compare it with two state-of-the-art automatic neural network compression frameworks, *i.e.*, AMC (37) and ANNC (90). AMC trains a reinforcement learning agent to predict compression action for each layer environment. ANNC jointly proceeds pruning and quantization within energy constraint. We conduct OTO on their shared experiment, *i.e.*, ResNet50 on CIFAR10. ResNet50 includes both the standard convolutional layers and the layers with residual connections, which are partitioned into ZIGs following Section 3.1. We report the results in Table 2 along with other competitors from (61; 78). Based on the results, all methods achieve competitive validation accuracies, where most of them are even higher than the baseline reported in (37). OTO outperforms AMC, ANNC without quantization, PruneTrain and N2NSkip by using only 12.8% FLOPs and 8.8% parameters. Note that no FLOPs reduction is reported in (37) and (90). Finally, we highlight that OTO is flexible to incorporate quantization as the two techniques are complementary and will leave to future work.

Table 2: ResNet50 for CIFAR10.

Method	FLOPs	# of Params	Top-1 Acc.
Baseline	100%	100%	93.5%
AMC (37)	–	60.0%	93.6%
ANNC (90)	–	50.0%	**95.0%**
PruneTrain (61)	30.0%	–	93.1%
N2NSkip (78)	–	10.0%	94.4%
OTO	**12.8%**	**8.8%**	94.4%

Ablation Study of Switching Parameter N. We provide ablation study regarding the impact the switch (parameterized as N) between initialization stage and projection stage in Algorithm 1. In theory, as shown in Theorem 1 of Appendix B.4, the projection stage should start when the iterate falls nearby a group sparse local minimizer. In practice, we relaxed it to start projection stage once the iterate falling into some stationary status regarding the validation accuracy. As described in Appendix A.2, throughout all experiments, we periodically decay the learning rate per fixed number of epochs parameterized as T. At the end of each T epochs,

Table 3: OTO Under Different Switchings ($N = T, 2T.3T$) for VGG16, VGG16-BN and ResNet50 on CIFAR10

Backend	FLOPs	# of Params	Top-1 Acc.
VGG16	$15.8\% \pm 2.4\%$	$2.5\% \pm 0.3\%$	$90.8\% \pm 0.2\%$
VGG16-BN	$25.3\% \pm 1.9\%$	$4.8\% \pm 1.1\%$	$93.1\% \pm 0.5\%$
ResNet50	$11.7\% \pm 3.5\%$	$8.1\% \pm 1.2\%$	$93.8\% \pm 0.6\%$

we then proceed a statistical test similar to (98) but on the validation accuracy and found that the validation accuracy fell into stationarity on the late epochs of each period. Therefore, in our pruning experiments, we switch to the group-sparsity stage from initialization stage right after the first T epochs. Table 3 describes the performance of OTO under varying switching parameters, from which we observe that OTO is not largely sensitive to the switching parameter if the group sparsity stage starts after some stationary condition has been numerically satisfied.

ResNet50 for ImageNet. We now evaluate OTO on ResNet50 for ImageNet. As shown in Table 4, OTO prunes $64.5\%(1 - 35.5\%)$ parameters to achieve $65.5\%(1-34.5\%)$ FLOPs reduction with only $1.4\%/0.8\%$ Top-1/5 accuracy regression compared to the baseline. OTO consistently outperforms the majority of counterparts especially on the FLOPs reduction and the parameter reduction. We note that Hinge (56) prunes CNNs via structured-sparsity optimization by employing standard stochastic proximal gradient method. It requires several trainings including fine-tuning the pruned model, because it partitions the parameters into non-ZIGs and relies on an empirical truncation mechanism to generate zero groups due to the weakness of proximal operator in deep learning applications (8). In contrast, OTO only trains and prunes the full model from scratch once and obtains better pruning results. The comparison between OTO and Hinge stand as evidence of the superiority of OTO due to ZIGs and HSPG. Furthermore, if conducting more training efforts, OTO reaches higher Top-1/5 accuracy, marked as * in Table 4 and becomes more competitive to other competitors, such as GBN, Group-HS and ResRep.

Table 4: ResNet50 for ImageNet.

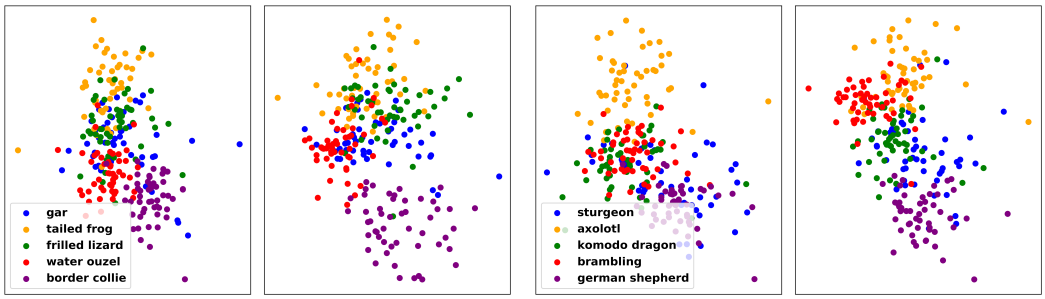
Method	FLOPs	# of Params	Top-1 Acc.	Top-5 Acc.
Baseline	100%	100%	76.1%	92.9%
DDS-26 (43)	57.0%	61.2%	71.8%	91.9%
CP (38)	66.7%	–	72.3%	90.8%
ThiNet-50 (45)	44.2%	48.3%	71.0%	90.0%
RBP (101)	43.5%	48.0%	71.1%	90.0%
RRBP (101)	45.4%	–	73.0%	91.0%
SFP (36)	41.8%	–	74.6%	92.1%
Hinge (56)	46.6%	–	74.7%	–
GBN-50 (94)	44.9%	46.6%	75.2%	92.4%
GBN-60 (94)	59.5%	68.2%	76.2%	92.8%
Group-HS (2e-5) (91)	32.4%	-	75.2%	92.5%
Group-HS (1e-5) (91)	52.9%	-	**76.4%**	**93.1%**
ResRep (18)	45.5%	-	76.2%	92.9%
SCP (48)	45.7%	-	74.2%	92.0%
OTO	**34.5%**	**35.5%**	74.7%	92.1%
OTO*	**34.5%**	**35.5%**	75.1%	92.5%

Representation of Deep Features of ImageNet. It is widely acknowledged that deep neural architectures could be treated as non-linear feature representation extractors. Therefore, we further study the feature representation extracted by OTO to demonstrate its generalizability to other visual applications besides image classification. Figure 4 shows the clustering results of ImageNet validation images using the deep feature extracted by both the baseline ResNet50 and the pruned ResNet50 by OTO. Specifically, we extract the deep features over the validation samples in ImageNet, *i.e.*, the tensors fed into the fully connected layer, and project them onto a 2-dimensional space via PCA (47). For illustration, following the hierarchy of ImageNet (3), two sets of five classes are randomly selected[3]. We observe that the deep features of the pruned ResNet50 by OTO remain structured in the sense that distinct classes are well separated from each other. Over all 1000-class ImageNet validation images, OTO achieves 48.2% clustering accuracy compared to 42.5% of the baseline ResNet50 using k-means. Both observations indicate that with only 35.5% parameters and 34.5% FLOPs, the pruned ResNet50 is still able to extract highly discriminative deep features. We argue that during model compression, OTO not only achieves parameter and FLOPs reduction, but also preserves the ability of capturing perceptual properties (99). This is especially important in training and compressing models for many vision tasks, *e.g.*, object detection (70; 71), frame interpolation (2; 17; 67) and video synthesis (84; 50). We leave the application of OTO to broader tasks to future work.

4.2 Large-Scale Transformer

We show the scalability of OTO by pruning the large-scale transformer Bert (82), evaluated on SQuAD, a question-answering benchmark (69). Bert mainly includes embedding layers, fully connected layers and multi-head attention layers. The fully connected layers and the multi-head attention layers are partitioned into ZIGs following Section 3.1. For fair comparisons, we follow the prior Bert compression works (14; 75) and do not prune the embedding layers.

[3]Each selected class belongs to a disjoint upper category.



(a) (b)

Figure 4: Clustering results of ImageNet validation images using deep features extracted by full ResNet50 (left of a and b) and pruned ResetNet50 by OTO (right of a and b). The points are visualized by projecting deep features onto a two-dimensional space via PCA.

Table 5: Pruning Bert on SQuAD

Method	# of Params	Exact	F1-score	SpeedUp
Baseline	100%	81.0%	88.3%	1×
MaP (75)	**10.0%**	67.7%	78.5%	1×*
MvP (75)	**10.0%**	71.9%	81.7%	1×*
ProxSSI (14)	83.4%†	72.3%	82.0%	1×
OTO	91.0%	**75.0%**	**84.1%**	1.1×
OTO	76.2%	72.3%	82.1%	1.2×
OTO	66.7%	71.9%	82.0%	1.3×
OTO	53.3%	71.4%	81.5%	1.5×
OTO	40.0%	70.9%	81.1%	**1.8×**

* Based on the statement in the official git repository of (75).
† Approximate value based on the group sparsity reported in (14).

To the best of our knowledge, OTO is the first work that compresses Bert by exploring group sparsity on individual layers and achieves significant parameter reduction and inference speedup[4]. In contrast, the existing works (29; 75; 30) prune individual parameters instead, *i.e.*, the generated sparsity is not structured. Hence, the computed models typically do not have inference speedup (75), unless are executed by specialized hardware and sparse computing library (31; 10). As shown in Table 5, under different group sparsity upper bound constraints, OTO reduces 9% to 60% parameters and achieves up to 1.8× inference speedup based on the average model execution time [5]. In comparison, despite that the pruned model contains 10% parameters, MaP and MvP (75) do not have any inference speedup. On the other hand, the structured sparsity on Bert is studied in (14) (referred to as ProxSSI), where an adaptive proximal method is proposed to yield group-sparse solution. Nonetheless, ProxSSI optimizes over non-ZIGs and relies on proximal operator to identify group sparsity. Therefore, the groups even parameterized with zeros have to be retained in the model rather than pruned. As a consequence, ProxSSI is not competitive to OTO on parameter reduction, and there is no reported inference speedup. Note that all the pruning methods achieve comparable exact match rate and F1-score.

5 Conclusion And Future Work

We propose OTO, a one-shot deep neural networks (DNNs) training and pruning framework, that compresses full DNNs into slimmer architectures with competitive performances and significant FLOPs and parameter reduction without fine-tuning. OTO contains two fundamentals: *(i)* partitions the trainable parameters of DNNs into zero-invariant groups (ZIGs), thereby pruning zero groups does not affect the model output, and *(ii)* trains by a novel optimizer, Half-Space Stochastic Projected Gradient (HSPG), which outperforms proximal methods on group sparsity exploration and maintains comparable convergence. We numerically demonstrate OTO on benchmark experiments, *i.e.*, VGG16 for CIFAR10, ResNet50 for CIFAR10/ImageNet and Bert for SQuAD, and achieve state-of-the-art pruning results. We leave automatically generating ZIGs for arbitrary DNNs, incorporating quantization and applying OTO to other tasks to future work.

[4] Knowledge distillation (40) and LayerDropout (21) compresses Bert by pruning layers in their entirety.
[5] Run by OnnxRuntime (16)

References

[1] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019.

[3] Mike Bostock. Imagenet hierarchy. https://observablehq.com/@mbostock/imagenet-hierarchy.

[4] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[5] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[7] Jianda Chen, Shangyu Chen, and Sinno Jialin Pan. Storage efficient and dynamic flexible runtime channel pruning via deep reinforcement learning. 2019.

[8] Tianyi Chen, Tianyu Ding, Bo Ji, Guanyi Wang, Yixin Shi, Sheng Yi, Xiao Tu, and Zhihui Zhu. Orthant based proximal stochastic gradient method for ell_1-regularized optimization. *arXiv preprint arXiv:2004.03639*, 2020.

[9] Tianyi Chen, Yixin Shi, and Sheng Yi. Spatially sparse convolutional neural networks for inking applications, Sept. 17 2020. US Patent App. 16/355,702.

[10] Xuhao Chen. Escoin: Efficient sparse convolutional neural network inference on gpus. *arXiv preprint arXiv:1802.10280*, 2018.

[11] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

[12] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[13] Aaron Defazio and Léon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. In *Advances in Neural Information Processing Systems*, 2019.

[14] Tristan Deleu and Yoshua Bengio. Structured sparsity inducing adaptive optimizers for deep learning. *arXiv preprint arXiv:2102.03869*, 2021.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] ONNX Runtime developers. Onnx runtime. https://onnxruntime.ai/, 2021.

[17] Tianyu Ding, Luming Liang, Zhihui Zhu, and Ilya Zharkov. Cdfi: Compression-driven network design for frame interpolation. *arXiv preprint arXiv:2103.10559*, 2021.

[18] Xiaohan Ding, Tianxiang Hao, Jianchao Tan, Ji Liu, Jungong Han, Yuchen Guo, and Guiguang Ding. Lossless cnn channel pruning via decoupling remembering and forgetting. *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[19] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[20] Marwa El Halabi, Francis Bach, and Volkan Cevher. Combinatorial penalties: Which structures are preserved by convex relaxations? In *International Conference on Artificial Intelligence and Statistics*, pages 1551–1560. PMLR, 2018.

[21] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019.

[22] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[23] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[24] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*, 2019.

[25] Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[26] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

[27] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In *European Conference on Computer Vision*, pages 702–721. Springer, 2020.

[28] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[29] Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. Compressing bert: Studying the effects of weight pruning on transfer learning. *arXiv preprint arXiv:2002.08307*, 2020.

[30] Fu-Ming Guo, Sijia Liu, Finlay S Mungall, Xue Lin, and Yanzhi Wang. Reweighted proximal pruning for large-scale language representation. *arXiv preprint arXiv:1909.12486*, 2019.

[31] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A Horowitz, and William J Dally. Eie: Efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News*, 44(3):243–254, 2016.

[32] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[33] Song Han, Jeff Pool, John Tran, and William J Dally. Learning both weights and connections for efficient neural networks. *arXiv preprint arXiv:1506.02626*, 2015.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[36] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*, 2018.

[37] Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018.

[38] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[39] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[40] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[41] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

[42] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12(Nov):3371–3412, 2011.

[43] Zehao Huang and Naiyan Wang. Data-driven sparse structure selection for deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 304–320, 2018.

[44] Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373, 2010.

[45] Jianxin Wu Jian-Hao Luo and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *ICCV*, pages 5058–5066, 2017.

[46] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[47] Ian Jolliffe. *Principal Component Analysis*, pages 1094–1096. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[48] Minsoo Kang and Bohyung Han. Operation-aware soft channel pruning using differentiable masks. In *International Conference on Machine Learning*, pages 5122–5131. PMLR, 2020.

[49] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[50] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: Image and video synthesis using graph cuts. *Acm transactions on graphics (tog)*, 22(3):277–286, 2003.

[51] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[52] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. Snip: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340*, 2018.
[53] Sangkyun Lee and Stephen J Wright. Manifold identification in dual averaging for regularized stochastic online learning. *The Journal of Machine Learning Research*, 13(1):1705–1744, 2012.
[54] Bailin Li, Bowen Wu, Jiang Su, and Guangrun Wang. Eagleeye: Fast sub-net evaluation for efficient neural network pruning. In *European Conference on Computer Vision*, pages 639–654. Springer, 2020.
[55] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.
[56] Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group sparsity: The hinge between filter pruning and decomposition for network compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8018–8027, 2020.
[57] Yuchao Li, Shaohui Lin, Baochang Zhang, Jianzhuang Liu, David Doermann, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Exploiting kernel sparsity and entropy for interpretable cnn compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2019.
[58] Shaohui Lin, Rongrong Ji, Yuchao Li, Cheng Deng, and Xuelong Li. Toward compact convnets via structure-sparsity regularized filter pruning. *IEEE transactions on neural networks and learning systems*, 31(2):574–588, 2019.
[59] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *Advances in neural information processing systems*, pages 3288–3298, 2017.
[60] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pages 5058–5066, 2017.
[61] Sangkug Lym, Esha Choukse, Siavash Zangeneh, Wei Wen, Sujay Sanghavi, and Mattan Erez. Prunetrain: fast neural network training by dynamic sparse model reconfiguration. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–13, 2019.
[62] Fanxu Meng, Hao Cheng, Ke Li, Huixiang Luo, Xiaowei Guo, Guangming Lu, and Xing Sun. Pruning filter in filter. *arXiv preprint arXiv:2009.14410*, 2020.
[63] Michael Mitzenmacher. Probability and computing-randomized algorithms and probabilistic analysis. *JOURNAL-OPERATIONAL RESEARCH SOCIETY*, 56(12):1454, 2005.
[64] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *The Journal of Machine Learning Research*, 18(1):4671–4703, 2017.
[65] Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, and Dmitry P Vetrov. Structured bayesian pruning via log-normal multiplicative noise. In *Advances in Neural Information Processing Systems*, pages 6775–6784, 2017.
[66] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 2009.
[67] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017.
[68] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
[69] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
[70] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
[71] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
[72] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. *arXiv preprint arXiv:2003.02389*, 2020.
[73] Volker Roth and Bernd Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855, 2008.

[74] Nicolas L Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in neural information processing systems*, pages 2663–2671, 2012.
[75] Victor Sanh, Thomas Wolf, and Alexander M Rush. Movement pruning: Adaptive sparsity by fine-tuning. *arXiv preprint arXiv:2005.07683*, 2020.
[76] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 7, 2020.
[77] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[78] Arvind Subramaniam and Avinash Sharma. N2nskip: Learning highly sparse networks using neuron-to-neuron skip connections. In *BMVC*, 2020.
[79] Yehui Tang, Yunhe Wang, Yixing Xu, Dacheng Tao, Chunjing Xu, Chao Xu, and Chang Xu. Scop: Scientific control for reliable neural network pruning. *arXiv preprint arXiv:2010.10732*, 2020.
[80] Frederick Tung and Greg Mori. Clip-q: Deep network compression learning by in-parallel pruning-quantization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
[81] Mart van Baalen, Christos Louizos, Markus Nagel, Rana Ali Amjad, Ying Wang, Tijmen Blankevoort, and Max Welling. Bayesian bits: Unifying quantization and pruning. *arXiv preprint arXiv:2005.07093*, 2020.
[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
[83] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376*, 2020.
[84] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.
[85] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *arXiv preprint arXiv:1608.03665*, 2016.
[86] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.
[87] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.
[88] Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.
[89] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.
[90] Haichuan Yang, Shupeng Gui, Yuhao Zhu, and Ji Liu. Automatic neural network compression by sparsity-quantization joint learning: A constrained optimization-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2178–2188, 2020.
[91] Huanrui Yang, Wei Wen, and Hai Li. Deephoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures. *arXiv preprint arXiv:1908.09979*, 2019.
[92] Haiqin Yang, Zenglin Xu, Irwin King, and Michael R Lyu. Online learning for group lasso. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1191–1198, 2010.
[93] 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.
[94] Zhonghui You, Kun Yan, Jinmian Ye, Meng Ma, and Ping Wang. Gate decorator: Global filter pruning method for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1909.08174*, 2019.
[95] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
[96] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.
[97] Junyu Zhang and Lin Xiao. Multi-level composite stochastic optimization via nested variance reduction. *arXiv preprint arXiv:1908.11468*, 2019.

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

[99] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[100] Tianyun Zhang, Shaokai Ye, Kaiqi Zhang, Jian Tang, Wujie Wen, Makan Fardad, and Yanzhi Wang. A systematic dnn weight pruning framework using alternating direction method of multipliers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–199, 2018.

[101] Yuefu Zhou, Ya Zhang, Yanfeng Wang, and Qi Tian. Accelerate cnn via recursive bayesian pruning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3306–3315, 2019.

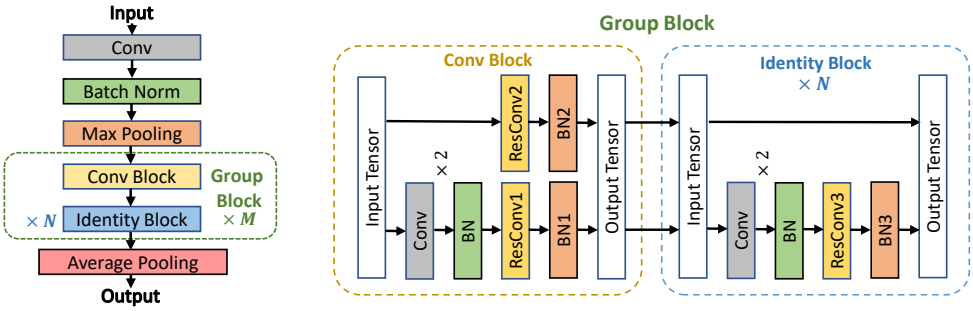
[102] Tao Zhuang, Zhixuan Zhang, Yuheng Huang, Xiaoyi Zeng, Kai Shuang, and Xiang Li. Neuron-level structured pruning using polarization regularizer. *Advances in Neural Information Processing Systems*, 33, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In the abstract and introduction, we highlighted the scope of this work and summarized the main contributions: (i) OTO, (ii) ZIG, (iii) HSPG, and (iv) CNN and Bert experiments.
   (b) Did you describe the limitations of your work? [Yes] See Section 5. OTO does not proceed quantization jointly yet.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work focuses on a general neural network training and pruning framework. We did not expect any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] As the supplementary materials to the proposed HSPG, we described and discussed the main theorems with the full set of assumptions in Appendix B.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix B, we included the complete proofs to the main theorems.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provided our source code, detailed data information, and instructions in supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We included the training details in Appendix A.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In Appendix A.3, we reported the error bars for all the main results shown in Section 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [Yes] See Section 4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] For the extensible experiments shown in Appendix B, we used several datasets from LIBSVM and the Fashion-MNIST with proper citation and description.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We have mentioned that all used datasets are publicly available to academic usage in Section 4.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We described the used datasets in Section 4 and concluded that they do not include personally identifiable information and offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Implementation Details of OTO

A.1 ZIG for ResNet50



(a) ResNet50.

(b) Group block.

Figure 5: ResNet50 Architecture.

Without loss of generality, we illustrate ZIGs for the general ResNet class with ResNet50. As shown in Figure 5a, ResNet50 begins with a Conv-BN, a pooling layer, and extracts features by M Group Blocks that each contains one Conv Block and N Identity Block. The extracted features are ultimately fed into an average pooling layer for different downstream computations. There exist two types of convolution structures inside ResNet50: *(i)* the regular Conv-BN (see Section 3.1), marked as gray and green blocks in Figure 5b, and *(ii)* the ResConv-BN of which the output shares the same dimension with another ResConv-BN, marked as yellow and brown in Figure 5b.

For ResNet50, we partition regular Conv-BN following Section 3.1. For ResConv-BN, within each Group Block, the intermediate input/output tensors in Conv/Indentity Blocks share the same dimension, and hence all the ResConv-BNs in one Group block share the same number of 3D filters. Consequently, their flattened filter matrices has the same number of rows. Figure 5b breaks down the architecture of a Group Block. The output tensors of ResConv-BN1 and ResConv-BN2 in Conv Block, denoted as $\mathcal{O}^1$ and $\mathcal{O}^2$, are computed by (7) and (8) respectively. They are then summed up as the input tensor of the subsequent identify block $\mathcal{I}^{I_1}$, indicating that $\mathcal{O}^1$ and $\mathcal{O}^2$ have the same shape and their flattened filter matrices $\hat{\mathcal{K}}^1$ and $\hat{\mathcal{K}}^2$ has the same number of rows. As (11), $\mathcal{I}^{I_1}$ later sums the output tensor of ResConv-BN3 $\mathcal{O}^3$ to yield the input tensor to the next Identity Block $\mathcal{I}^{I_2}$, implying the filter matrix of ResConv-BN3 $\hat{\mathcal{K}}^3$ has the same number of rows as $\hat{\mathcal{K}}^1$ and $\hat{\mathcal{K}}^2$.

$$\mathcal{O}^1 \leftarrow \frac{a(\mathcal{I}^1 \otimes \hat{\mathcal{K}}^1 + \boldsymbol{b}^1) - \boldsymbol{\mu}^1}{\boldsymbol{\sigma}^1} \odot \boldsymbol{\gamma}^1 + \boldsymbol{\beta}^1 \tag{7}$$

$$\mathcal{O}^2 \leftarrow \frac{a(\mathcal{I}^2 \otimes \hat{\mathcal{K}}^2 + \boldsymbol{b}^2) - \boldsymbol{\mu}^2}{\boldsymbol{\sigma}^2} \odot \boldsymbol{\gamma}^2 + \boldsymbol{\beta}^2 \tag{8}$$

$$\mathcal{I}^{I_1} \leftarrow \mathcal{O}^1 + \mathcal{O}^2 \tag{9}$$

$$\mathcal{O}^3 \leftarrow \frac{a(\mathcal{I}^3 \otimes \hat{\mathcal{K}}^3 + \boldsymbol{b}^3) - \boldsymbol{\mu}^3}{\boldsymbol{\sigma}^3} \odot \boldsymbol{\gamma}^3 + \boldsymbol{\beta}^3 \tag{10}$$

$$\mathcal{I}^{I_2} \leftarrow \mathcal{I}^{I_1} + \mathcal{O}^3 \tag{11}$$

Therefore, based on (7) to (11), to make the entire Group Block zero-invariant, we group each ith row of the filter matrix for all the Res-Conv-BNs of same group block. In doing so, any one row of parameters being zeros results in the output, *i.e.*, the corresponding channel of feature map, being zeros. Figure 6 shows ZIG for the three ResConv-BN of a Group Block. Regardless of the input, the ith channel-wise matrix of $\mathcal{I}^{I_1}$ are zeros if and only if both i^{th} channel-wise matrices of $\mathcal{O}^1$ and $\mathcal{O}^2$ are equal to zero. This is equivalent to both i^{th} rows of $\hat{\mathcal{K}}^1$ and $\hat{\mathcal{K}}^2$ being zeros. Similarly, i^{th} channel-wise matrix of $\mathcal{I}^{I_2}$ being zeros regardless of the input further requires the ith row of $\hat{\mathcal{K}}^3$ to be grouped in the ZIG.

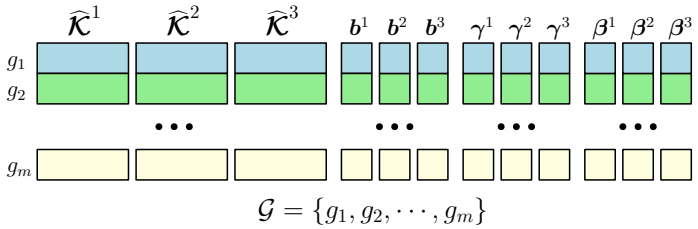


Figure 6: Zero Invariant Groups for the three ResConv-BN of a Group Block.

A.2 Training Details

We implement OTO in PyTorch. The key ingredient HSPG is packaged as an optimizer class which is flexible to various applications. During the experiment, the trainable parameters of the full model $\mathcal{M}$ are firstly partitioned into a ZIG set $\mathcal{G}$ wherein each group is tagged as its corresponding atomic layer category, *e.g.*, fully-connected layer or convolutional layer. The ZIG set $\mathcal{G}$ is treated as an argument to the HSPG constructor. Then $\mathcal{M}$ is trained from scratch by HSPG where the group-wise variables are updated based on their tagged layer category. In our repository, we provide the prescribed ZIG partitions for the DNNs used in this paper, *i.e.*, VGG16, VGG16-BN, ResNet50 and Bert. For other models, one can easily follow Section 3.1 and Appendix A.1 to construct a ZIG partition and feed it as an argument to the HSPG optimizer. After training, a full group-sparse model with high performance is achieved. Finally, a slimmer pruned model $\mathcal{M}^*$ is constructed following Section 3.4 without fine-tuning and has the identical performance as the full group-sparse model. We provide the implementation in `https://github.com/tianyic/only_train_once` and the pruned models associated with the corresponding full group sparse models in `https://tinyurl.com/nips20212271`.

Parameter Settings. We conduct all experiments on an Nvidia RTX8000 graphics card with 48 GB memory. For all experiments, the step size (learning rate) α_k is initialized as 10^{-1}, and decayed by a factor 0.1 periodically T epochs till the minimum value 10^{-4}. The selection of T depends on the max number of epochs K. We follow various benchmark online resources to select K. Particularly, for all CIFAR10 experiments, we follow the model pre-training settings in (56) and set K as 300. Note that by using the same number of epochs, OTO achieves both slimmer model and competitive performance simultaneously. For the ImageNet experiment, following (35), we set T as 30 and K as 120. For all Bert experiments, we set K as 8 and T as 2.

We set the mini-batch size as the commonly used 64 for CIFAR10, 256 for ImageNet and 32 for SQuAD experiments. For all experiments, we initialize the regularization coefficient λ as 10^{-3} to balance between model performance and group sparsity. In particular, λ as 10^{-3} is the maximum value from the candidate set $\{10^{-2}, 10^{-3}, 10^{-4}\}$ which returns competitive evaluation results to the majority of the tested models trained without regularization. In addition, to favor more on the model performance, if group sparsity becomes stationary, we decay λ by a factor 0.1 periodically after stepping into Group-Sparsity Stage. The control parameter $\epsilon \in [0, 1)$ in the half-space projection (6) controls the aggressiveness level of group sparsity promotion, which is typically fixed as 0 since for most of experiments, $\epsilon \equiv 0$ has resulted in sufficiently good experiment results. In case if group sparsity is not sufficiently yielded, we provide an adaptive mechanism to increase ϵ by 0.1 till the upper bound 0.999. For the setting of $N_{\mathcal{P}}$ which controls when switching to the Group-Sparsity Stage, we proceed a test on objective value stationarity similarly to (98, Section 2.1) and empirically set $N_{\mathcal{P}} \equiv T$ since the objective function values become stationary at the late epochs till T. Hence, the Group-Sparsity Stage starts after T epochs and is accompanied with the α_k decaying.

Additional Remarks. For the ZIG partition of ResNet50 on CIFAR10, we include all trainable variables of ResNet50 and apply the ZIG partition described in Appendix A.1 for ResConv-BN and the ZIG partition described in Section 3.1 for standard Conv-BN. For the ZIG partition of ResNet50 on ImageNet, we construct ZIGs for standard Conv-BN only. This is because we observe that ZIG partition for ResConv-BN lead to accuracy regression in spite of more FLOPs reduction, (15% FLOPs with up to 71% Top-1 Accuracy). The cause is that it decreases the number of features maps generated by the entire Group Block. Additionally, for Bert experiments, to study the accuracy evolution against different compression rates, we set extra constraints to bound the maximum group sparsity ratio, *e.g.*,

30%, 50%, 70%, and do not yield new zero groups if the upper bound has been exceeded. Note that without any constraint, OTO reaches about 95% group sparsity ratio with 80% F1-score.

A.3 Error Bar Analysis

In this section, we report the overall statistics of the experiments and analyze the error bar. We note that for fair comparison with others, in the main body of paper, we report the best results in terms of remaining FLOPs/parameters and Top-1/5 accuracy. We conduct all experiments three times with different random seeds.

Table 6: OTO for CNN Experiments (mean $\pm$ std)

Model	Dataset	FLOPs	# of Params	Top-1 Acc.
VGG16	CIFAR10	$16.9\% \pm 1.5\%$	$2.7\% \pm 0.2\%$	$90.7\% \pm 0.3\%$
VGG16-BN	CIFAR10	$26.9\% \pm 0.1\%$	$5.5\% \pm 0.1\%$	$93.2\% \pm 0.1\%$
ResNet50	CIFAR10	$11.9\% \pm 1.7\%$	$8.8\% \pm 0.4\%$	$93.9\% \pm 0.5\%$
ResNet50	ImageNet	$34.8\% \pm 1.8\%$	$35.9\% \pm 1.7\%$	$73.3\% \pm 1.4\%$

Training neural networks is equivalent to solving a non-convex optimization problem which has numerous local minimizers, thereby training from scratch like OTO may generate solutions close to stationary points with different attributes. As shown in Table 6, we can see that for the CNN experiments, OTO performs reliably to achieve significant FLOPs and parameters reduction and competitive Top-1 accuracy with small fluctuations. As shown in Table 7, we observe that for the Bert on SQuAD experiments, OTO achieves reliable exact match rates and f1-scores under different upper bounds of group sparsity ratios.

Table 7: OTO for Bert on SQuAD (mean $\pm$ std)

# of Params	Exact	F1-score
91.0%	$74.6\% \pm 0.4\%$	$83.5\% \pm 0.5\%$
76.2%	$71.9\% \pm 0.4\%$	$81.8\% \pm 0.3\%$
66.7%	$71.7\% \pm 0.2\%$	$81.7\% \pm 0.3\%$
53.3%	$71.1\% \pm 0.3\%$	$81.4\% \pm 0.1\%$
40.0%	$70.8\% \pm 0.1\%$	$80.9\% \pm 0.2\%$

A.4 FLOPs Reduction Breakdown

We provide the layer-wise FLOPs reduction for VGG16 on CIFAR10. As shown in Table 8, the majority of the FLOPs reduction via OTO comes from a few middle ConvLayers (over 10% to the overall FLOPs reductions) instead of the first ConvLayer (0.45% to the overall FLOPs reduction). In general, the distribution of FLOPs reduction per Layer of OTO is similar to other pruning baselines.

Table 8: FLOPs Reduction Breakdown for the ConvLayers of VGG16 on CIFAR10

ConvLayer Index	# of Output Channels		FLOPs Reduction	
	Original	Pruned	Quantity (Million)	Percentage (%)
1	64	21	1.19M	0.45%
2	64	45	29.04M	11.07%
3	128	82	10.47M	3.99%
4	128	110	17.22M	6.57%
5	256	109	11.97M	4.56%
6	256	68	33.48M	12.77%
7	256	37	36.30M	13.84%
8	512	13	18.81M	7.17%
9	512	9	37.73M	14.38%
10	512	7	37.74M	14.39%
11	512	3	9.44M	3.60%
12	512	5	9.44M	3.60%
13	512	8	9.44M	3.60%

B Convergence Analysis of HSPG

In this section, we provide theoretical analysis of HSPG. We focus on the most popular setting of optimization problem (2) as follows

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ \psi(\boldsymbol{x}) := f(\boldsymbol{x}) + \lambda r(\boldsymbol{x}),\ f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x}), \tag{12}$$

Here $f(\boldsymbol{x})$ is defined as the average of N task-specific loss functions $f_i : \mathbb{R}^n \mapsto \mathbb{R},\ \forall\, i = 1, \dots, N$. The stochastic gradient $\nabla \tilde{f}$ proposed in Section 3.3 can be obtained via a uniformly chosen mini-batch $\mathcal{B} \subseteq [N]$ as follows: for any $\boldsymbol{x} \in \mathbb{R}^n$, given $\mathcal{B}$, we have

$$\nabla \tilde{f}(\boldsymbol{x}) = \nabla \underbrace{\left(\frac{1}{|\mathcal{B}|} \sum_{i\in\mathcal{B}} f_i(\boldsymbol{x}) \right)}_{=: f_{\mathcal{B}}(\boldsymbol{x})}, \tag{13}$$

in short, we denote above term as $\nabla f_{\mathcal{B}}(\boldsymbol{x})$ where $f_{\mathcal{B}}(\boldsymbol{x})$ is the average of loss functions with repsect to mini-batch $\mathcal{B}$. Similarly, let $\psi_{\mathcal{B}}(\boldsymbol{x}) := f_{\mathcal{B}}(\boldsymbol{x}) + \lambda r(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

Organization. The Section B is organized as follows: From Section B.1 to Section **??**, we present the convergence result and the sparse recovery guarantee for Half-Space Step. More specifically,

- In Section B.1, we first presented the existing related work of solving the problem (12).
- In Section B.2, we show the sufficient decrease of Half-Space Step under Assumption 1.
- In Section B.3, we derive the projection region of Half-Space Step and compare this projection region with existing methods.
- In Section B.4, we give the convergence result of Half-Space Step as stated in Theorem 1 under the Assumption (2, 3).

To complete the story, in Section B.5, we show that the "close enough" condition required in Theorem **??** can be achieved by the *Sub-gradient Descent Step* under the Assumption 5. Moreover, we further point out that: (1) the *Sub-gradient Descent Step* we used to achieve a "close enough" solution can be replaced by other methods, and (2) the Assumption 4 is only a sufficient condition that we could use to show the "close enough" condition.

B.1 Related Work

Problem (12) has been well studied in deterministic optimization with various algorithms that are capable of returning solutions with both low objective value and high group sparsity under proper λ (95; 73; 42; 64). Proximal methods are classical approaches to solve the structured non-smooth optimization (12), including the popular proximal gradient method (Prox-FG) which only uses the first-order derivative information. When N is huge, stochastic methods become ubiquitous to operate on a small subset to avoid the costly evaluation over all instances in deterministic methods for large-scale problems. Proximal stochastic gradient method (Prox-SG) (19) is the natural stochastic extension of Prox-FG. Regularized dual-averaging method (RDA) (87; 92) is proposed by extending the dual averaging scheme in (66). To improve the convergence rate, there exists a set of incremental gradient methods inspired by SAG (74) to utilizes the average of accumulated past gradients. For example, proximal stochastic variance-reduced gradient method (Prox-SVRG) (88) and proximal spider (Prox-Spider) (97) are developed to adopt multi-stage schemes based on the well-known variance reduction technique SVRG proposed in (46) and Spider developed in (22) respectively. SAGA (12) stands as the midpoint between SAG and Prox-SVRG.

Compared to deterministic methods, the studies of structured sparsity regularization (12) in stochastic field become somewhat rare and limited. Prox-SG, RDA, Prox-SVRG, Prox-Spider and SAGA are valuable state-of-the-art stochastic algorithms for solving problem (12) but with apparent weakness. Particularly, these existing stochastic algorithms typically meet difficulties to achieve both decent convergence and effective group sparsity identification simultaneously (*e.g.*, small function values but merely dense solutions), because of the randomness and the limited sparsity-promotion mechanisms. In depth, Prox-SG, RDA, Prox-SVRG, Prox-Spider and SAGA derive from proximal gradient method

to utilize the proximal operator to produce group of zero variables. Such operator is generic to extensive non-smooth problems, consequently perhaps not sufficiently insightful if the target problems possess certain properties, *e.g.*, the group sparsity structure as problem (12). In fact, in convex setting, the proximal operator suffers from variance of gradient estimate; and in non-convex setting, especially deep learning, the discreet step size (learning rate) further deteriorates its effectiveness on the group sparsity promotion, as shown in Section 3.3 of the main body that the projection region vanishes rapidly except RDA. RDA has superiority on finding manifold structure to others (53), but inferiority on the objective convergence. Besides, the variance reduction techniques are typically required to measure over a huge mini-batch data points in both theory and practice which is probably prohibitive for large-scale problems, and have been observed as sometimes noneffective for deep learning applications (13). On the other hand, to introduce sparsity, there exist heuristic weight pruning methods (55; 60), whereas they commonly do not equip with theoretical guarantee, so that easily diverge and hurt generalization accuracy.

B.2 Sufficient Decrease of Half-Space Step

Before we present the convergence result of Half-Space Step to the global group-sparsity solution, in this part, we first show that the sufficient decrease property holds for Half-Space Step under the following Assumption 1.

Assumption 1. *Assume the following assumptions hold.*

- ***(A1-1).*** $f : \mathbb{R}^n \mapsto \mathbb{R}$ *is differentiable and* L *smooth.*
- ***(A1-2).*** $r : \mathbb{R}^n \mapsto \mathbb{R}$ *is sub-differentiable and convex.*
- ***(A1-3).*** $\psi = f + \lambda r : \mathbb{R}^n \mapsto \mathbb{R}$ *is sub-differentiable over all points* $\boldsymbol{x} \in \mathbb{R}^n$.

For any $k > N_{\mathcal{P}}$ (in Half-Space Step of Algorithm 2), recall the next iterate $\boldsymbol{x}_{k+1}$ and the search direction

$$\boldsymbol{d}_k := \frac{\boldsymbol{x}_{k+1} - \boldsymbol{x}_k}{\alpha_k} = \frac{\mathrm{Proj}_{\mathcal{S}_k}^{\mathrm{HS}}(\boldsymbol{x}_k - \alpha_k \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)) - \boldsymbol{x}_k}{\alpha_k}. \tag{14}$$

Define

$$\hat{\mathcal{G}}_k := \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \cap \mathcal{I}^{0}(\boldsymbol{x}_{k+1}) \tag{15}$$

$$\tilde{\mathcal{G}}_k := \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \cap \mathcal{I}^{\neq 0}(\boldsymbol{x}_{k+1}) \tag{16}$$

be the sets of groups which projects or not onto zero. We claim that the following Lemma 1 holds.

Lemma 1. *Under Assumption 1, the search direction* $\boldsymbol{d}_k$ *is a descent direction for* $\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)$*, i.e.,* $\boldsymbol{d}_k^\top \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) < 0$*. Moreover, we have the following sufficient decrease property holds,*

$$\psi_{\mathcal{B}_k}(\boldsymbol{x}_{k+1}) \leq \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2 - \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[\boldsymbol{x}_k]_g\|^2 . \tag{17}$$

Proof. **Proof of Descent Direction.** It follows the Half-Space Step in Algorithm 2 and the definition of $\tilde{\mathcal{G}}_k$ and $\hat{\mathcal{G}}_k$ as (16) and (15) that $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ where $\boldsymbol{d}_k$ is

$$[\boldsymbol{d}_k]_g = \begin{cases} -[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g & \text{if } g \in \tilde{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^{\neq 0}(\boldsymbol{x}_{k+1}), \\ -[\boldsymbol{x}_k]_g / \alpha_k & \text{if } g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^{0}(\boldsymbol{x}_{k+1}), \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

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

$$\begin{aligned} &[\boldsymbol{x}_k - \alpha_k \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g^\top [\boldsymbol{x}_k]_g < \epsilon \|[\boldsymbol{x}_k]_g\|^2 , \\ &(1-\epsilon) \|[\boldsymbol{x}_k]_g\|^2 < \alpha_k [\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g^\top [\boldsymbol{x}_k]_g . \end{aligned} \tag{19}$$

For simplicity, let $\mathcal{I}_k^{\neq 0} := \mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$. Since $[\boldsymbol{d}_k]_g = \mathbf{0}$ for any $g \in \mathcal{I}^0(\boldsymbol{x}_k)$, then by (18) and (19), we have

$$\begin{aligned}
\boldsymbol{d}_k^\top \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) &= [\boldsymbol{d}_k]_{\mathcal{I}_k^{\neq 0}}^\top [\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_{\mathcal{I}_k^{\neq 0}} \\
&= -\sum_{g \in \tilde{\mathcal{G}}_k} \|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2 - \sum_{g \in \hat{\mathcal{G}}_k} \frac{1}{\alpha_k} [\boldsymbol{x}_k]_g^\top [\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g \\
&\leq -\sum_{g \in \tilde{\mathcal{G}}_k} \|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2 - \sum_{g \in \hat{\mathcal{G}}_k} \frac{1}{\alpha_k^2} (1-\epsilon) \|[\boldsymbol{x}_k]_g\|^2 < 0,
\end{aligned} \tag{20}$$

holds for any $\epsilon \in [0, 1)$, which implies that $\boldsymbol{d}_k$ is a descent direction for $\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)$.

Proof of Sufficient Decrease. Now, we start to prove the suffcient decrease of Half-Space Step. By assumption, $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L smooth and $r : \mathbb{R}^n \mapsto \mathbb{R}$ is convex. Therefore

$$\begin{aligned}
& \psi_{\mathcal{B}_k}(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) && (21) \\
&= f_{\mathcal{B}_k}(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) + \lambda r(\boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k) && (22) \\
&\leq f_{\mathcal{B}_k}(\boldsymbol{x}_k) + \alpha_k \boldsymbol{d}_k^\top \nabla f_{\mathcal{B}}(\boldsymbol{x}_k) + \frac{\alpha_k^2 L}{2} \|\boldsymbol{d}_k\|^2 \qquad \text{by Assumption 1} && (23) \\
&\quad + \lambda r(\boldsymbol{x}_k) + \alpha_k \lambda \boldsymbol{d}_k^\top \zeta(\boldsymbol{x}_k) && (24) \\
&= \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) + \alpha_k \boldsymbol{d}_k^\top \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) + \frac{\alpha_k^2 L}{2} \|\boldsymbol{d}_k\|^2 && (25) \\
&\leq \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \sum_{g \in \tilde{\mathcal{G}}_k} \|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2 \qquad \text{by inequality (20) \& } \boldsymbol{d}_k \text{ definition} && (26) \\
&\quad - \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[\boldsymbol{x}_k]_g\|^2, && (27)
\end{aligned}$$

which completes the proof. □

According to Lemma 1, the objective value $\psi_{\mathcal{B}}(\boldsymbol{x})$ with $\mathbb{E}[\psi_{\mathcal{B}}(\boldsymbol{x})|\boldsymbol{x}] = \psi(\boldsymbol{x})$ achieves a sufficient decrease in Half-Space Step given α_k is small enough. Taking the expectation over mini-batch $\mathcal{B}$ on both sides, it is straight-forward to obtain the expectation version of the sufficient decrease property.

Corollary 1. *Similarly, under Assumption 1, for all $k > N_{\mathcal{P}}$, we have*

$$\psi(\boldsymbol{x}_{k+1}) \leq \psi(\boldsymbol{x}_k) - \sum_{g \in \tilde{\mathcal{G}}_k} \left(\alpha_k - \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left[\|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right] - \left(\frac{1-\epsilon}{\alpha_k} - \frac{L}{2}\right) \sum_{g \in \hat{\mathcal{G}}_k} \|[\boldsymbol{x}_k]_g\|^2. \tag{28}$$

B.3 Projection Region of Half-Space Step

In this part, we derive the projection region of Half-Space Step, and reveal that is a superset of the projection region of existing methods, e.g. Prox-SG, Prox-SVRG and Prox-Spider, under the same α_k and λ.

Proposition 1. *For any $k > N_{\mathcal{P}}$, given $\boldsymbol{x}_k$, the next iterate $\boldsymbol{x}_{k+1}$ obtained by the Half-Space Step satisfies that: for any group $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$,*

$$[\boldsymbol{x}_{k+1}]_g = \begin{cases} [\hat{\boldsymbol{x}}_{k+1}]_g - \alpha_k \lambda \frac{[\boldsymbol{x}_k]_g}{\|[\boldsymbol{x}_k]_g\|} & \text{if } [\hat{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g > (\alpha_k \lambda + \epsilon) \|[\boldsymbol{x}_k]_g\|, \\ 0 & \text{otherwise}, \end{cases} \tag{29}$$

where $\hat{\boldsymbol{x}}_{k+1} := \boldsymbol{x}_k - \alpha_k \nabla f_{\mathcal{B}_k}(\boldsymbol{x}_k)$. Moreover, we claim that if $\|[\hat{\boldsymbol{x}}_{k+1}]_g\| \leq \alpha_k \lambda$, then $[\boldsymbol{x}_{k+1}]_g = 0$ for any $\epsilon \geq 0$.

Proof. For $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^{\neq 0}(\boldsymbol{x}_{k+1})$, by line 11-12 in Algorithm 2, it is equivalent to

$$\begin{aligned} \left[\boldsymbol{x}_k - \alpha_k \nabla f_{\mathcal{B}_k}(\boldsymbol{x}_k) - \alpha_k \lambda \frac{[\boldsymbol{x}_k]_g}{\|[\boldsymbol{x}_k]_g\|}\right]_g^\top [\boldsymbol{x}_k]_g &> \epsilon \|[\boldsymbol{x}_k]_g\|^2 , \\ [\hat{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g - \alpha_k \lambda \|[\boldsymbol{x}_k]_g\| &> \epsilon \|[\boldsymbol{x}_k]_g\|^2 , \\ [\hat{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g &> (\alpha_k \lambda + \epsilon \|[\boldsymbol{x}_k]_g\|) \|[\boldsymbol{x}_k]_g\| . \end{aligned} \tag{30}$$

Similarly, $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^{0}(\boldsymbol{x}_{k+1})$ is equivalent to

$$\begin{aligned} \left[\boldsymbol{x}_k - \alpha_k \nabla f_{\mathcal{B}_k}(\boldsymbol{x}_k) - \alpha_k \lambda \frac{[\boldsymbol{x}_k]_g}{\|[\boldsymbol{x}_k]_g\|}\right]_g^\top [\boldsymbol{x}_k]_g &\leq \epsilon \|[\boldsymbol{x}_k]_g\|^2 , \\ [\hat{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g - \alpha_k \lambda \|[\boldsymbol{x}_k]_g\| &\leq \epsilon \|[\boldsymbol{x}_k]_g\|^2 , \\ [\hat{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g &\leq (\alpha_k \lambda + \epsilon \|[\boldsymbol{x}_k]_g\|) \|[\boldsymbol{x}_k]_g\| . \end{aligned} \tag{31}$$

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

$$[\hat{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g \leq \|[\hat{\boldsymbol{x}}_{k+1}]_g\| \|[\boldsymbol{x}_k]_g\| \leq \alpha_k \lambda \|[\boldsymbol{x}_k]_g\| . \tag{32}$$

Hence $[\boldsymbol{x}_{k+1}]_g = 0$ holds for any $\epsilon \geq 0$ by (31), 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.4 Convergence Analysis of Half-Space Step

In this section, we give the convergence result of Half-Space Step under the following Assumptions for the properties of the objective function and the global optimal solution $\boldsymbol{x}^*$ of (2).

Assumption 2. *Assume the following assumptions hold.*

- ***(A2-1).*** *For $i = 1, 2, \cdots, N$, each $f_i : \mathbb{R}^n \to \mathbb{R}$ is differentiable and bounded below.*
- ***(A2-2).*** *For $i = 1, 2, \cdots, N$, each $f_i : \mathbb{R}^n \to \mathbb{R}$ is L_i smooth.*
- ***(A2-3).*** *$\psi_{\mathcal{B}} = f_{\mathcal{B}} + \lambda r : \mathbb{R}^n \mapsto \mathbb{R}$ has bounded sub-gradient (i.e., $\mathbb{E}[\|\nabla \psi_{\mathcal{B}}(\boldsymbol{x})\|^2] \leq M^2$ for some universal constant M) over all points $\boldsymbol{x} \in \mathbb{R}^n$ with respect to any mini-batch $\mathcal{B} \subseteq [N]$.*
- ***(A2-4).*** *The stochastic gradient $\nabla f_{\mathcal{B}}(\boldsymbol{x})$ satisfies $\mathbb{E}_{\mathcal{B}}[\nabla f_{\mathcal{B}}(\boldsymbol{x})|\boldsymbol{x}] = \nabla f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*
- ***(A2-5).*** *The stochastic gradient $\nabla f_{\mathcal{B}}(\boldsymbol{x})$ satisfies $Var_{\mathcal{B}}[\nabla f_{\mathcal{B}}(\boldsymbol{x})|\boldsymbol{x}] \leq \sigma^2$ for all $\boldsymbol{x} \in \mathbb{R}^n$, where $\sigma^2 > 0$ is a universal constant.*

Notice that this Assumption 2 is a variant of the Assumption 1, to be concise, we set L proposed in Assumption 1 as $L := \max_{i=1}^N \{L_i\}$.

Assumption 3. *Assume the following assumptions hold.*

- ***(A3-1).*** $\sum_{k \geq N_{\mathcal{P}}} \alpha_k = \infty$.
- ***(A3-2).*** $\sum_{k \geq N_{\mathcal{P}}} \alpha_k^2 < \infty$.

Assumption 4. *Assume the following assumptions hold.*

- ***(A4-1).*** *The least and the largest ℓ_2-norm of non-zero groups in $\boldsymbol{x}^*$ are lower and upper bounded by some constants,*

$$0 < 2\delta_1 := \min_{g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}^*)} \|[\boldsymbol{x}^*]_g\| \leq \max_{g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}^*)} \|[\boldsymbol{x}^*]_g\| =: 2\delta_2. \tag{33}$$

- ***(A4-2).*** *The regularizer parameter λ satisfies that: for any mini-batch $\mathcal{B} \subseteq [N]$, there exists a constant δ_3 such that*

$$0 < 2\delta_3 := \min_{g \in \mathcal{I}^{0}(\boldsymbol{x}^*)} (\lambda - \|[\nabla f_{\mathcal{B}}(\boldsymbol{x}^*)]_g\|) . \tag{34}$$

Theorem 1. *Under Assumptions (1, 2, 3, 4), set*

$$R \in \left(0,\ \min\left\{\frac{1}{\epsilon}\cdot\left[-(\delta_1+2\epsilon\delta_2)+\sqrt{(\delta_1+2\epsilon\delta_2)^2-4\epsilon^2\delta_2+4\epsilon\delta_1^2}\right],\delta_1\right\}\right), \tag{35}$$

$$\epsilon \in \left[0,\ \min\left\{\frac{\delta_1^2}{\delta_2},\frac{2\delta_1-R}{2\delta_2+R}\right\}\right), \tag{36}$$

$$\alpha_k \in \left(0,\ \min\left\{\frac{2(1-\epsilon)}{L},\frac{1}{L},\frac{2\delta_1-R-\epsilon(2\delta_2+R)}{M}\right\}\right),\quad \forall k\geq N_{\mathcal{P}}. \tag{37}$$

If there exists a $K \geq N_{\mathcal{P}}$ such that

$$\|\boldsymbol{x}_K-\boldsymbol{x}^*\|\leq\frac{R}{2}. \tag{38}$$

Given any $\tau\in(0,1)$, there exists some $\alpha_k=\mathcal{O}(1/(1+\sqrt{\tau})(k-K))$ and $|\mathcal{B}_k|=\mathcal{O}(k-K)$ for all $k\geq K$ such that the sequence $\{\boldsymbol{x}_k\}_{k\geq K}$ obtained from the Algorithm 2 converges to some stationary point with probability at least $1-\tau$, i.e.,

$$\liminf_k \mathbb{E}\left[\|\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|\right]=0\quad \textit{with probability}\quad 1-\tau. \tag{39}$$

Proof. **Proof Sketch.** We split the proof of showing the convergence to some stationary points into two parts. In the first part, we show the convergence holds for all groups in $\tilde{\mathcal{G}}_k$; and in the second part, we show the convergence also holds in $\hat{\mathcal{G}}_k$.

Convergence in $\tilde{\mathcal{G}}_k$ part. For any $t\in\mathbb{N}_+$, applying Corollary 1 yields

$$\psi(\boldsymbol{x}_{N_{\mathcal{P}}})-\psi(\boldsymbol{x}_{N_{\mathcal{P}}+t}) \tag{40}$$

$$=\sum_{k=N_{\mathcal{P}}}^{N_{\mathcal{P}}+t-1}\psi(\boldsymbol{x}_k)-\psi(\boldsymbol{x}_{k+1}) \tag{41}$$

$$\geq\sum_{k=N_{\mathcal{P}}}^{N_{\mathcal{P}}+t-1}\sum_{g\in\tilde{\mathcal{G}}_k}\left(\alpha_k-\frac{\alpha_k^2L}{2}\right)\mathbb{E}\left[\|[\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right]+\sum_{k=N_{\mathcal{P}}}^{N_{\mathcal{P}}+t-1}\left(\frac{1-\epsilon}{\alpha_k}-\frac{L}{2}\right)\sum_{g\in\hat{\mathcal{G}}_k}\|[\boldsymbol{x}_k]_g\|^2. \tag{42}$$

Combining the assumption that ψ is bounded below and letting $t\to\infty$ yield

$$\underbrace{\sum_{k=N_{\mathcal{P}}}^{\infty}\sum_{g\in\tilde{\mathcal{G}}_k}\left(\alpha_k-\frac{\alpha_k^2L}{2}\right)\mathbb{E}\left[\|[\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right]}_{=:T_1}+\underbrace{\sum_{k=N_{\mathcal{P}}}^{\infty}\left(\frac{1-\epsilon}{\alpha_k}-\frac{L}{2}\right)\sum_{g\in\hat{\mathcal{G}}_k}\|[\boldsymbol{x}_k]_g\|^2}_{=:T_2}<\infty. \tag{43}$$

Given $\alpha_k\in(0,2(1-\epsilon)/L)$, we have $T_1>0,T_2>0$, combining with $T_1+T_2<\infty$ implies

$$\sum_{k=N_{\mathcal{P}}}^{\infty}\sum_{g\in\tilde{\mathcal{G}}_k}\left(\alpha_k-\frac{\alpha_k^2L}{2}\right)\mathbb{E}\left[\|[\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right] \tag{44}$$

$$=\sum_{k=N_{\mathcal{P}}}^{\infty}\sum_{g\in\tilde{\mathcal{G}}_k}\alpha_k\mathbb{E}\left[\|[\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right]-\sum_{k=N_{\mathcal{P}}}^{\infty}\sum_{g\in\tilde{\mathcal{G}}_k}\frac{\alpha_k^2L}{2}\mathbb{E}\left[\|[\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right]. \tag{45}$$

Based on the boundness of sub-gradient in Assumptions 2 and the choice of stepsize in 3, we have

$$\sum_{k=N_{\mathcal{P}}}^{\infty}\sum_{g\in\tilde{\mathcal{G}}_k}\frac{\alpha_k^2L}{2}\mathbb{E}\left[\|[\nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|^2\right]<\infty, \tag{46}$$

which yields

$$\sum_{k=N_{\mathcal{P}}}^{\infty} \sum_{g \in \tilde{\mathcal{G}}_k} \alpha_k \mathbb{E}\left[\left\|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\right\|^2\right] < \infty \tag{47}$$

$$\Rightarrow \quad \liminf_{k \geq N_{\mathcal{P}}} \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\left\|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\right\|^2\right] = 0 \tag{48}$$

$$\Rightarrow \quad \lim_{k \geq \mathcal{K}} \sum_{g \in \tilde{\mathcal{G}}_k} \mathbb{E}\left[\left\|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\right\|^2\right] = 0, \quad \exists\, \mathcal{K} \subseteq \{N_{\mathcal{P}}, \ldots\} \tag{49}$$

Convergence in $\hat{\mathcal{G}}_k$ part. Under Assumption 4, Lemma (2, 3, 4) show that if there exists a $K \geq N_{\mathcal{P}}$ such that

$$\|\boldsymbol{x}_K - \boldsymbol{x}^*\| \leq R, \tag{50}$$

then we have the following results hold

$$\mathcal{I}^{\neq 0}(\boldsymbol{x}^*) \subseteq \mathcal{I}^{\neq 0}(\boldsymbol{x}_K), \qquad \text{non-zero group coverage}, \tag{51}$$

$$\boldsymbol{x}^* \in \mathcal{S}_K, \qquad \text{correct optimal inclusion } \mathcal{S}_K, \tag{52}$$

$$\mathcal{I}^{\neq 0}(\boldsymbol{x}_K) \cap \mathcal{I}^{=0}(\boldsymbol{x}_{K+1}) \subseteq \mathcal{I}^{=0}(\boldsymbol{x}^*), \qquad \text{correct zero group projection}. \tag{53}$$

Under Assumption (2, 3, 4), Lemma (5, 6, 7) and Corollary 2 show that: given any $\tau \in (0, 1)$, with probability at least $1 - \tau$, for any $k \geq K$, $\boldsymbol{x}^*$ inhabits $\mathcal{S}_k$. Therefore, for any $k \geq K$, any group $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \cap \mathcal{I}^{=0}(\boldsymbol{x}_{k+1})$ will be projected to zero group correctly with probability at least $1 - \tau$.

Convergence over the whole space. Based on the discussion in $\hat{\mathcal{G}}_k$ part, it is sufficient to focus on the subspace of $\tilde{\mathcal{G}}_k$. Hence, (49) naturally implies that the sequence $\{\boldsymbol{x}_k\}_{k \in \mathcal{K}}$ converges to some stationary point with high probability. By the above, we conclude that

$$\mathbb{P}\left(\liminf_{k} \mathbb{E}\left[\|\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|\right] = 0\right) \geq 1 - \tau. \tag{54}$$

□

B.4.1 Support Lemma in the Proof of Theorem 1

The Lemma 2 shows that if the optimal distance from the current iterate $\boldsymbol{x}_k$ to any local minimizer $\boldsymbol{x}^*$ is sufficiently small, then HSPG already covers the supports of $\boldsymbol{x}^*$, *i.e.*, $\mathcal{I}^{\neq 0}(\boldsymbol{x}^*) \subseteq \mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$.

Lemma 2. *Under Assumption 4, given any $R \leq \delta_1$, for any $k \geq N_{\mathcal{P}}$, if $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq R$, then we have $\mathcal{I}^{\neq 0}(\boldsymbol{x}^*) \subseteq \mathcal{I}^{\neq 0}(\boldsymbol{x}_k)$.*

Proof. For any $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}^*)$, we have that

$$\begin{aligned} \|[\boldsymbol{x}^*]_g\| - \|[\boldsymbol{x}_k]_g\| \leq \|[\boldsymbol{x}_k - \boldsymbol{x}^*]_g\| \leq \|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq R \leq \delta_1 \\ \|[\boldsymbol{x}_k]_g\| \geq \|[\boldsymbol{x}^*]_g\| - \delta_1 \geq 2\delta_1 - \delta_1 = \delta_1 > 0 \end{aligned} \tag{55}$$

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

The Lemma 3 shows that if the distance between the current iterate $\boldsymbol{x}_k$ and $\boldsymbol{x}^*$, *i.e.*, $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ is sufficiently small, then $\boldsymbol{x}^*$ inhabits the reduced space $\mathcal{S}_k := \mathcal{S}(\boldsymbol{x}_k)$.

Lemma 3. *Under Assumption 4, for any $k \geq N_{\mathcal{P}}$, given $\epsilon \in [0, \delta_1^2/\delta_2)$ and*

$$R \leq R^* := \frac{1}{\epsilon} \cdot \left[-(\delta_1 + 2\epsilon\delta_2) + \sqrt{(\delta_1 + 2\epsilon\delta_2)^2 - 4\epsilon^2\delta_2 + 4\epsilon\delta_1^2}\right], \tag{56}$$

if $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq R$, we have

$$[\boldsymbol{x}_k]_g^\top [\boldsymbol{x}^*]_g \geq \epsilon \left\|[\boldsymbol{x}_k]_g\right\|^2, \quad g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}^*). \tag{57}$$

Consequently, it implies $\boldsymbol{x}^* \in \mathcal{S}_k$ by the definition as (4).

Proof. For any $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}^*)$,

$$\|[\boldsymbol{x}_k]_g\| \leq \|[\boldsymbol{x}^*]_g\| + R \leq 2\delta_2 + R, \tag{58}$$

and the R^* defined in (56) is one of the roots of the quadratic $\epsilon z^2 + (4\epsilon\delta_2 + 2\delta_1)z + 4\epsilon\delta_2^2 - 4\delta_1^2 = 0$ regarding $z \in \mathbb{R}$. Thus

$$\begin{aligned}
[\boldsymbol{x}_k]_g^\top [\boldsymbol{x}^*]_g =& [\boldsymbol{x}_k - \boldsymbol{x}^* + \boldsymbol{x}^*]_g^\top [\boldsymbol{x}^*]_g \\
=& [\boldsymbol{x}_k - \boldsymbol{x}^*]_g^\top [\boldsymbol{x}^*]_g + \|[\boldsymbol{x}^*]_g\|^2 \\
\geq& \|[\boldsymbol{x}^*]_g\|^2 - \|[\boldsymbol{x}_k - \boldsymbol{x}^*]_g\| \, \|[\boldsymbol{x}^*]_g\| \\
=& \|[\boldsymbol{x}^*]_g\| \left(\|[\boldsymbol{x}^*]_g\| - \|[\boldsymbol{x}_k - \boldsymbol{x}^*]_g\|\right) \\
\geq& 2\delta_1(2\delta_1 - R) \geq \epsilon(2\delta_2 + R)^2 \\
\geq& \epsilon \, \|[\boldsymbol{x}_k]_g\|^2
\end{aligned} \tag{59}$$

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

The Lemma 4 shows that if $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$ 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 4. *Under Assumption 4, for any $k \geq N_{\mathcal{P}}$, given $\epsilon \in \left[0, \frac{2\delta_1 - R}{2\delta_2 + R}\right)$, $\alpha_k \in \left(0, \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}\right)$ and $R \in (0, \min\{R^*, \delta_1\})$, if $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq R$, then for any $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^0(\boldsymbol{x}_{k+1})$, we have $g \in \mathcal{I}^0(\boldsymbol{x}^*)$.*

Proof. To prove it by contradiction, suppose there exists some $g \in \hat{\mathcal{G}}_k$ such that $g \in \mathcal{I}^{\neq 0}(\boldsymbol{x}^*)$. Since $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^0(\boldsymbol{x}_{k+1})$, then the group projection (6) is trigerred at g such that

$$\begin{aligned}
[\tilde{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g &= [\boldsymbol{x}_k - \alpha \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g^\top [\boldsymbol{x}_k]_g \\
&= \|[\boldsymbol{x}_k]_g\|^2 - \alpha_k [\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g^\top [\boldsymbol{x}_k]_g < \epsilon \, \|[\boldsymbol{x}_k]_g\|^2 .
\end{aligned} \tag{60}$$

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

$$\|[\boldsymbol{x}_k - \boldsymbol{x}^*]_g\| \leq \|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq R \tag{61}$$

Combining the definition of δ_1 and δ_2 in Assumption 4, we have that

$$\begin{aligned}
\|[\boldsymbol{x}_k]_g\| &\geq \|[\boldsymbol{x}^*]_g\| - R \geq 2\delta_1 - R \\
\|[\boldsymbol{x}_k]_g\| &\leq \|[\boldsymbol{x}^*]_g\| + R \leq 2\delta_2 + R
\end{aligned} \tag{62}$$

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

$$\begin{aligned}
[\tilde{\boldsymbol{x}}_{k+1}]_g^\top [\boldsymbol{x}_k]_g &= \|[\boldsymbol{x}_k]_g\|^2 - \alpha_k [\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g^\top [\boldsymbol{x}_k]_g \\
&\geq \|[\boldsymbol{x}_k]_g\|^2 - \alpha_k \, \|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\| \, \|[\boldsymbol{x}_k]_g\| \\
&= \|[\boldsymbol{x}_k]_g\| \left(\|[\boldsymbol{x}_k]_g\| - \alpha_k \, \|[\nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)]_g\|\right) \\
&\geq \|[\boldsymbol{x}_k]_g\| \left(\|[\boldsymbol{x}_k]_g\| - \alpha_k M\right) \\
&\geq \|[\boldsymbol{x}_k]_g\| \left[(2\delta_1 - R) - \alpha_k M\right] \\
&\geq \|[\boldsymbol{x}_k]_g\| \left[(2\delta_1 - R) - \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M} M\right] \\
&\geq \|[\boldsymbol{x}_k]_g\| \left[(2\delta_1 - R) - 2\delta_1 + R + \epsilon(2\delta_2 + R)\right] \\
&\geq \epsilon \, \|[\boldsymbol{x}_k]_g\| \, (2\delta_2 + R) \\
&\geq \epsilon \, \|[\boldsymbol{x}_k]_g\|^2
\end{aligned} \tag{63}$$

which contradicts with (60). Hence, we conclude that any g of variables projected to zero, *i.e.*, $g \in \hat{\mathcal{G}}_k = \mathcal{I}^{\neq 0}(\boldsymbol{x}_k) \bigcap \mathcal{I}^0(\boldsymbol{x}_{k+1})$ are exactly also the zeros on the optimal solution $\boldsymbol{x}^*$, *i.e.*, $g \in \mathcal{I}^0(\boldsymbol{x}^*)$. □

We next present that if the iterate of Half-Space Step is close enough to the optimal solution $\boldsymbol{x}^*$, then $\boldsymbol{x}^*$ inhabits all reduced spaces constructed by the subsequent iterates of Half-Space Step with high probability.

To establish this results, we require the following two lemmas (Lemma 5 and Lemma 6). The Lemma 5 bounds the accumulated error because of random sampling. Here we introduce the error of gradient estimator on $\mathcal{I}^{\neq 0}(\boldsymbol{x})$ for ψ on mini-batch $\mathcal{B}$ as

$$\boldsymbol{e}_{\mathcal{B}}(\boldsymbol{x}) := [\nabla\psi_{\mathcal{B}}(\boldsymbol{x}) - \nabla\psi(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})}, \tag{64}$$

where by the definition of r in problem (12), we have $\boldsymbol{e}_{\mathcal{B}}(\boldsymbol{x})$ also equals to the error of estimation for ∇f, i.e., $\boldsymbol{e}_{\mathcal{B}}(\boldsymbol{x}) = [\nabla f_{\mathcal{B}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})}$.

Lemma 5. *Under Assumption 2, given any $\theta > 1$, $K \geq N_{\mathcal{P}}$, let $k := K + t$ with $t \in \mathbb{Z}_{\geq 0}$, then there exists a sequence of stepsize $\alpha_k = \mathcal{O}(1/(1+\theta)t)$ and corresponding size of mini-batch $|\mathcal{B}_k| = \mathcal{O}(t)$, such that for any $y_t \in \mathbb{R}^n$,*

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

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

Proof. Define random variable $Y_t := \alpha_{K+t}\|e_{\mathcal{B}_{K+t}}(\boldsymbol{y}_t)\|_2$ for all $t \geq 0$. Since $\{\boldsymbol{y}_t\}_{t=0}^{\infty}$ 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

$$\mathbb{P}\left(Y \geq \mathbb{E}[Y] + \theta\sqrt{\mathrm{Var}[Y]}\right) \leq \mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq \theta\sqrt{\mathrm{Var}[Y]}\right) \leq \frac{1}{\theta^2}. \tag{65}$$

And based on the Assumption 2, there exists an upper bound $\sigma^2 > 0$ for the variance of random noise $e_{\mathcal{B}}(\boldsymbol{x})$ 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 (65), we have

$$\begin{aligned}
Y &\leq \mathbb{E}[Y] + \theta\sqrt{\mathrm{Var}[Y]} && (66)\\
&\leq \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{k+t}|}} + \theta \cdot \sum_{t=0}^{\infty} \frac{\alpha_{K+t}^2\sigma^2}{|\mathcal{B}_{K+t}|} && (67)\\
&\leq \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{k+t}|}} + \theta \cdot \sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} = (1+\theta)\sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} && (68)
\end{aligned}$$

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 (63); 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 $\alpha_k = \mathcal{O}(1/(1+\theta)t)$ and $|\mathcal{B}_k| = \mathcal{O}(t)$, the above series converges and satisfies that

$$(1+\theta)\sum_{t=0}^{\infty} \frac{\alpha_{K+t}\sigma}{\sqrt{|\mathcal{B}_{K+t}|}} \leq \frac{3R^2}{8(4R+1)} \tag{69}$$

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

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

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

The Lemma 6 draws if previous iterate of Half-Space Step falls into the neighbor of $\boldsymbol{x}^*$, then under appropriate step size and mini-batch setting, the current iterate also inhabits the neighbor with high probability.

Lemma 6. *Under the assumptions of Lemma 5, suppose $\|\boldsymbol{x}_K - \boldsymbol{x}^*\| \le R/2$; for any ℓ satisfying $K \le \ell < K+t$, $0 < \alpha_\ell \le \min\{\frac{1}{L}, \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}\}$, $|B_\ell| \ge N - \frac{N}{2M}$ and $\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\| \le R$ holds, then*

$$\|\boldsymbol{x}_{K+t} - \boldsymbol{x}^*\| \le R. \tag{70}$$

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

Proof. It follows the assumptions of this lemma, Lemma 4, (15) and (16) that for any ℓ satisfying $K \le \ell < K+t$

$$\|[\boldsymbol{x}^*]_g\| = 0, \text{ for any } g \in \hat{\mathcal{G}}_\ell. \tag{71}$$

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

$$\begin{aligned}
&\|\boldsymbol{x}_{\ell+1} - \boldsymbol{x}^*\|^2 \\
=& \sum_{g \in \tilde{\mathcal{G}}_\ell} \|[\boldsymbol{x}_\ell - \boldsymbol{x}^* - \alpha_\ell \nabla\Psi(\boldsymbol{x}_\ell) - \alpha_\ell \boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\|^2 + \sum_{g \in \hat{\mathcal{G}}_k} \|[\boldsymbol{x}_\ell - \boldsymbol{x}^* - \boldsymbol{x}_\ell]_g\|^2 \\
=& \sum_{g \in \tilde{\mathcal{G}}_\ell} \left\{ \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\|^2 - 2\alpha_\ell [\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g^\top [\nabla\Psi(\boldsymbol{x}_\ell) + \boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g + \alpha_\ell^2 \|[\nabla\Psi(\boldsymbol{x}_\ell) + \boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\|^2 \right\} + \sum_{g \in \hat{\mathcal{G}}_\ell} \|[\boldsymbol{x}^*]_g\|^2 \\
=& \sum_{g \in \tilde{\mathcal{G}}_\ell} \left\{ \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\|^2 - 2\alpha_\ell [\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g^\top [\nabla\Psi(\boldsymbol{x}_\ell)]_g - 2\alpha_\ell [\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g^\top [\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g + \alpha_\ell^2 \|[\nabla\Psi(\boldsymbol{x}_\ell) + \boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\|^2 \right\} \\
\le& \sum_{g \in \tilde{\mathcal{G}}_\ell} \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\|^2 - \|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) - 2\alpha_\ell [\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g^\top [\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g + \alpha_\ell^2 \|[\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\|^2 \\
&+ 2\alpha_\ell^2 [\nabla\Psi(\boldsymbol{x}_\ell)]_g^\top [\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g \\
\le& \sum_{g \in \tilde{\mathcal{G}}_\ell} \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\|^2 - \|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) + 2\alpha_\ell \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\| \|[\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\| + \alpha_\ell^2 \|[\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\|^2 \\
&+ 2\alpha_\ell^2 \|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\| \|[\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\| \\
\le& \sum_{g \in \tilde{\mathcal{G}}_\ell} \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\|^2 - \|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) + (2\alpha_\ell + 2\alpha_\ell^2 L) \|[\boldsymbol{x}_k - \boldsymbol{x}^*]_g\| \|[\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\| + \alpha_\ell^2 \|[\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)]_g\|^2 \\
\le& \sum_{g \in \tilde{\mathcal{G}}_\ell} \left\{ \|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\|^2 - \|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) \right\} + (2\alpha_\ell + 2\alpha_\ell^2 L) \|\boldsymbol{x}_\ell - \boldsymbol{x}^*\| \|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\| + \alpha_\ell^2 \|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\|^2
\end{aligned} \tag{72}$$

On the other hand, by the definition of $\boldsymbol{e}_{\mathcal{B}}(\boldsymbol{x})$ as (64), we have that

$$\begin{aligned}
\boldsymbol{e}_{\mathcal{B}}(\boldsymbol{x}) =& [\nabla\Psi_{\mathcal{B}}(\boldsymbol{x}) - \nabla\Psi(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} = [\nabla f_{\mathcal{B}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \\
=& \frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} [\nabla f_j(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} - \frac{1}{N} \sum_{i=1}^{N} [\nabla f_i(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \\
=& \frac{1}{N} \sum_{j \in \mathcal{B}} \left[\frac{N}{|\mathcal{B}|} [\nabla f_j(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} - [\nabla f_j(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \right] - \frac{1}{N} \sum_{\substack{i=1 \\ i \notin \mathcal{B}}}^{N} [\nabla f_i(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \\
=& \frac{1}{N} \sum_{j \in \mathcal{B}} \left[\frac{N - |\mathcal{B}|}{|\mathcal{B}|} [\nabla f_j(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \right] - \frac{1}{N} \sum_{\substack{i=1 \\ i \notin \mathcal{B}}}^{N} [\nabla f_i(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})}
\end{aligned} \tag{73}$$

Thus taking the norm on both side of (73) and using triangle inequality results in the following:

$$\begin{aligned}
\|\boldsymbol{e}_{\mathcal{B}}(\boldsymbol{x})\| \le& \frac{1}{N} \sum_{j \in \mathcal{B}} \left[\frac{N - |\mathcal{B}|}{|\mathcal{B}|} \left\| [\nabla f_j(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \right\| \right] + \frac{1}{N} \sum_{\substack{i=1 \\ i \notin \mathcal{B}}}^{N} \left\| [\nabla f_i(\boldsymbol{x})]_{\mathcal{I}^{\neq 0}(\boldsymbol{x})} \right\| \\
\le& \frac{1}{N} \frac{N - |\mathcal{B}|}{|\mathcal{B}|} |\mathcal{B}_k| M + \frac{1}{N} (N - |\mathcal{B}|) M \le \frac{2(N - |\mathcal{B}|)M}{N}.
\end{aligned} \tag{74}$$

Since $\alpha_\ell \leq 1$, and $|B_\ell| \geq N - \frac{N}{2M}$ hence $\alpha_\ell \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| \leq 1$. Then combining with $\alpha_\ell \leq 1/L$, (72) can be further simplified as

$$
\begin{aligned}
&\left\|\boldsymbol{x}_{\ell+1} - \boldsymbol{x}^*\right\|^2 \\
\leq &\sum_{g \in \tilde{\mathcal{G}}_\ell} \left\{ \left\|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\right\|^2 - \left\|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\right\|^2 \left(2\frac{\alpha_\ell}{L} - \alpha_\ell^2\right) \right\} + (2\alpha_\ell + 2\alpha_\ell^2 L) \left\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\right\| \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| + \alpha_\ell^2 \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\|^2 \\
\leq &\sum_{g \in \tilde{\mathcal{G}}_\ell} \left\{ \left\|[\boldsymbol{x}_\ell - \boldsymbol{x}^*]_g\right\|^2 - \frac{1}{L^2} \left\|[\nabla\Psi(\boldsymbol{x}_\ell)]_g\right\|^2 \right\} + 4\alpha_\ell \left\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\right\| \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| + \alpha_\ell^2 \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\|^2 \\
\leq &\left\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\right\|^2 + 4\alpha_\ell \left\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\right\| \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| + \alpha_\ell \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\|
\end{aligned}
\tag{75}
$$

Following from the assumption that $\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\| \leq R$, then (75) can be further simplified as

$$
\begin{aligned}
\left\|\boldsymbol{x}_{\ell+1} - \boldsymbol{x}^*\right\|^2 &\leq \left\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\right\|^2 + 4\alpha_\ell R \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| + \alpha_k \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| \\
&\leq \left\|\boldsymbol{x}_\ell - \boldsymbol{x}^*\right\|^2 + (4R+1)\alpha_\ell \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\|
\end{aligned}
\tag{76}
$$

Summing the the both side of (76) from $\ell = K$ to $\ell = K + t - 1$ results in

$$
\left\|\boldsymbol{x}_{K+t} - \boldsymbol{x}^*\right\|^2 \leq \left\|\boldsymbol{x}_K - \boldsymbol{x}^*\right\|^2 + (4R+1) \sum_{\ell=K}^{K+t-1} \alpha_\ell \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| \tag{77}
$$

It follows Lemma 5 that the followng holds with probability at least $1 - \frac{1}{\theta^2}$,

$$
\sum_{\ell=K}^{\infty} \alpha_\ell \|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\| \leq \frac{3R^2}{4(4R+1)}. \tag{78}
$$

Thus we have that

$$
\begin{aligned}
\left\|\boldsymbol{x}_{K+t} - \boldsymbol{x}^*\right\|^2 &\leq \left\|\boldsymbol{x}_K - \boldsymbol{x}^*\right\|^2 + (4R+1) \sum_{\ell=K}^{K+t-1} \alpha_\ell \left\|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\right\| \\
&\leq \left\|\boldsymbol{x}_K - \boldsymbol{x}^*\right\|^2 + (4R+1) \sum_{\ell=K}^{\infty} \alpha_\ell \|\boldsymbol{e}_{\mathcal{B}_\ell}(\boldsymbol{x}_\ell)\| \\
&\leq \frac{R^2}{4} + (4R+1)\frac{3R^2}{4(4R+1)} \leq \frac{R^2}{4} + \frac{3R^2}{4} \leq R^2,
\end{aligned}
\tag{79}
$$

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

Based on the above lemmas, the Lemma 7 shows if initial iterate of Half-Space Step locates closely enough to $\boldsymbol{x}^*$, step size α_k polynomially decreases, and mini-batch size $\mathcal{B}_k$ polynomially increases, then $\boldsymbol{x}^*$ inhabits all subsequent reduced space $\{\mathcal{S}_k\}_{k=K}^{\infty}$ constructed in Half-Space Step with high probability.

Lemma 7. *If $\|\boldsymbol{x}_K - \boldsymbol{x}^*\| \leq \frac{R}{2}$, $K \geq N_{\mathcal{P}}$, $k = K + t$, $t \in \mathbb{Z}^+$, $0 < \alpha_k = \mathcal{O}(1/(\sqrt{N}t)) \leq \min\{\frac{2(1-\epsilon)}{L}, \frac{1}{L}, \frac{2\delta_1 - R - \epsilon(2\delta_2 + R)}{M}\}$ and $|\mathcal{B}_k| = \mathcal{O}(t) \geq N - \frac{N}{2M}$. Then for any constant $\tau \in (0,1)$, $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq R$ with probability at least $1 - \tau$ for any $k \geq K$.*

Proof. It follows Lemma 3 and the assumption of this lemma that $\boldsymbol{x}^* \in \mathcal{S}_K$. Moreover, it follows the assumptions of Lemma (5, 6, 7), the definition of finite-sum $f(\boldsymbol{x})$ in (12), and the bound of error as (74) that

$$
\mathbb{P}(\{\boldsymbol{x}_k\}_{k=K}^{\infty} \in \{x : \|\boldsymbol{x} - \boldsymbol{x}^*\| \leq R\}^{\infty}) \geq \left(1 - \frac{1}{\theta^2}\right)^{\mathcal{O}(N-K)} \geq 1 - \tau, \tag{80}
$$

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 $\mathcal{O}(N - K)$. □

Corollary 2. *Lemma 7 further implies $\boldsymbol{x}^*$ inhabits all subsequent $\mathcal{S}_k$, i.e., $\boldsymbol{x}^* \in \mathcal{S}_k$ for any $k \geq K$.*

B.5 The Initialization Stage

In previous parts, we show that the Half-Space Step guarantees to converge to the optimal solution, and ensures to recover the no-zero groups of the optimal solution under some assumptions with a "close-enough" initialization point $\boldsymbol{x}_{N_{\mathcal{P}}}$. To complete the story, in this part, we show that the iterate obtained from the *Subgradient Descent Update* in Algorithm 2 satisfies the "close-enough" condition with high probability. Remark here that the proximal methods, such as Prox-SG, Prox-SVRG and SAGA, may also serve in the initialization stage. However, for the general regularization $r(\boldsymbol{x})$, they may not have closed-form solution for the corresponding inherent subproblems, implying non-explicit update mechanism to the next iterate. Hence, people may have to inconveniently approximate the solutions of proximal operator by other techniques, whereas the sub-gradient method does not have these drawbacks. Therefore, for the generality of HSPG, we select the sub-gradient method in the Initialization Stage by default.

B.5.1 Convergence Analysis of Initialization Stage

In this part, we show that the "close enough" condition

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \frac{2\alpha_k \delta_3}{1 - \epsilon + \alpha_k L} \tag{81}$$

proposed in Theorem **??** can be achieved via the Initialization Stage (*Subgradient Descent Update*) in Algorithm 2 under the Assumption 5.

Assumption 5. *Assume the following assumptions hold.*

- ***(A5-1).*** *$f : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable and μ-strongly convex. $r : \mathbb{R}^n \mapsto \mathbb{R}$ is convex.*
- ***(A5-2).*** *There exists an universal constant M such that the stochastic gradient $\nabla f_{\mathcal{B}}(\boldsymbol{x})$ satisfies $\|\nabla f_{\mathcal{B}}(\boldsymbol{x})\|_2 \leq M$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and mini-batch $\mathcal{B}$.*
- ***(A5-3).*** *The stochastic gradient $\nabla f_{\mathcal{B}}(\boldsymbol{x})$ satisfies $\mathbb{E}_{\mathcal{B}}[\nabla f_{\mathcal{B}}(\boldsymbol{x})|\boldsymbol{x}] = \nabla f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Proposition 2. *Under Assumption 5, for any $R > 0$, any $\tau \in (0, 1)$, set*

$$N_{\mathcal{P}} = \left\lceil \log\left(\frac{\tau R}{2\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}\right) \Big/ \log\left(1 - \frac{\tau R}{2M}\right) \right\rceil, \tag{82}$$

$$\alpha_0 = \alpha_1 = \ldots = \alpha_{N_{\mathcal{P}}-1} = \frac{\tau \mu R}{2M^2}, \tag{83}$$

where

$$R = \left(\frac{2\alpha_k \delta_3}{1 - \epsilon + \alpha_k L}\right)^2 \tag{84}$$

*based on the setting of Theorem **??**. We have the Algorithm 1 (Subgradient Descent Update) returns a solution $\boldsymbol{x}_{N_{\mathcal{P}}}$ that satisfies $\|\boldsymbol{x}_{N_{\mathcal{P}}} - \boldsymbol{x}^*\|_2^2 \leq R$ with probability $1 - \tau$.*

Proof. Let $\boldsymbol{x}^*$ be the global optimal solution of (2). Let $\nabla\psi(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \lambda\zeta(\boldsymbol{x})$ and $\nabla\psi_{\mathcal{B}}(\boldsymbol{x}) = \nabla f_{\mathcal{B}}(\boldsymbol{x}) + \lambda\zeta(\boldsymbol{x})$ given any point $\boldsymbol{x} \in \mathbb{R}^n$ and mini-batch $\mathcal{B}$. Consider

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x}_k - \alpha_k \nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \boldsymbol{x}^*\|_2^2 \tag{85}$$

$$= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - 2\alpha_k \langle \nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^*\rangle + \|\alpha_k \nabla\psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|_2^2. \tag{86}$$

Due to (A1) in Assumption 5, the μ-strongly convexity of f and the convexity of r yields

$$\psi(\boldsymbol{x}^*) \geq \psi(\boldsymbol{x}_k) + \langle \nabla\psi(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k\rangle + \frac{\mu}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2. \tag{87}$$

Thus

$$
\begin{aligned}
&\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2 && (88)\\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - 2\alpha_k \langle \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* \rangle + \|\alpha_k \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|_2^2 && (89)\\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + 2\alpha_k \langle \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle + \|\alpha_k \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|_2^2 && (90)\\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + 2\alpha_k \langle \nabla \psi(\boldsymbol{x}_k) - \nabla \psi(\boldsymbol{x}_k) + \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle + \|\alpha_k \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|_2^2 && (91)\\
&\leq \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + 2\alpha_k \left(\psi(\boldsymbol{x}^*) - \psi(\boldsymbol{x}_k) - \frac{\mu}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \right) && (92)\\
&\quad + 2\alpha_k \langle \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \nabla \psi(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle + \|\alpha_k \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k)\|_2^2 && (93)\\
&\leq (1 - \alpha_k \mu) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - 2\alpha_k (\psi(\boldsymbol{x}_k) - \psi(\boldsymbol{x}^*)) + \alpha_k^2 \|\nabla \psi(\boldsymbol{x}_k)\|_2^2 && (94)\\
&\quad + 2\alpha_k \langle \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \nabla \psi(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle && (95)\\
&\leq (1 - \alpha_k \mu) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + \alpha_k^2 M^2 + 2\alpha_k \langle \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \nabla \psi(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle. && (96)
\end{aligned}
$$

Given $\boldsymbol{x}_k$, due to (A5-2) in Assumption 5, taking expectation over $\mathcal{B}_k$ yields

$$
\mathbb{E}_{\mathcal{B}_k}[\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2 | \boldsymbol{x}_k] \leq (1 - \alpha_k \mu) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + \alpha_k^2 M^2, \tag{97}
$$

where the above inequality holds by (A5-3) in Assumption 5

$$
\mathbb{E}_{\mathcal{B}_k}[\langle \nabla \psi_{\mathcal{B}_k}(\boldsymbol{x}_k) - \nabla \psi(\boldsymbol{x}_k), \boldsymbol{x}^* - \boldsymbol{x}_k \rangle | \boldsymbol{x}_k] = 0. \tag{98}
$$

For any $k \in \mathbb{N}_+$, any constant $c > 0$, and initial point $\boldsymbol{x}_0$, setting $\alpha_k = \frac{\mu}{cM^2}$, apply above inequality recursively yields

$$
\mathbb{E}_{\mathcal{H}} \left[\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \right] \leq \left(1 - \frac{1}{cM^2} \right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 + \frac{1}{c}, \tag{99}
$$

where $\mathcal{H} = \{\mathcal{B}_0, \ldots, \mathcal{B}_{k-1}\}$ denotes the whole history until step k.

Non-asymptotic bounds. Combine above together, given any $R > 0$, for any $\tau \in (0, 1)$, set

$$
N_{\mathcal{P}} = \left\lceil \log \left(\frac{\tau R}{2\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2} \right) \Big/ \log \left(1 - \frac{\tau R}{2M} \right) \right\rceil \tag{100}
$$

$$
\alpha_0 = \alpha_1 = \ldots = \alpha_{N_{\mathcal{P}}-1} = \frac{\tau \mu R}{2M^2}, \tag{101}
$$

by Markov inequality, we have

$$
\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \leq R \tag{102}
$$

holds with probability $1 - \tau$. □

Remark & Discussion. The Assumption 4 required by the Proposition 2 is a sufficient condition to show the "close enough" condition. Actually, for any general function f which satisfies Assumption 1 or Assumption 2, when $\boldsymbol{x}_k$ is "close enough" to any group-sparse *local optimal solution* $\boldsymbol{x}^*$ which satisfies the Assumption 4, the Theorem **??** identifies the zero-group with respect to the local optimal solution $\boldsymbol{x}^*$.

C Extensive Numerical Experiments

In this Appendix, we include extensive numerical experiments in the view of optimization to demonstrate the superiority of HSPG to other classical proximal methods on the sparsity exploration and the competitiveness on objective convergence in both convex and nonconvex settings. Particularly, in Appendix C.1, we provide convex experiments to *(i)* demonstrate the validness of group sparsity identification of HSPG; *(ii)* present comprehensive comparison to Prox-SG, RDA and Prox-SVRG on benchmark convex problems. In Appendix C.2, we show additional nonconvex experiments to reveal the superiority of HSPG to competitors on group sparsity exploration.

C.1 Convex Experiments

Linear Regression on Synthetic Data We 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 $\boldsymbol{y} \in \mathbb{R}^N$, we are interested in the following problem:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{2N}\|A\boldsymbol{x}-\boldsymbol{y}\|^2 + \lambda \sum_{g\in\mathcal{G}} \|[\boldsymbol{x}]_g\|\,. \tag{103}$$

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 $\boldsymbol{x}^*$ 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 $\boldsymbol{x}^*$ to be 0 according to a pre-specified group sparsity ratio; *(iv)* compute the target variable $\boldsymbol{y} = A\boldsymbol{x}^*$; (v) solve the above problem (103) for $\boldsymbol{x}$ with A and $\boldsymbol{y}$ only, and then evaluate the Intersection over Union (IoU) with respect to the identities of the zero groups between the computed solution estimate $\hat{\boldsymbol{x}}$ by HSPG and the ground truth $\boldsymbol{x}^*$.

We test HSPG on (103) 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 $\boldsymbol{x}^*$ requires that the number of non-zero elements in $\boldsymbol{x}^*$ 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 (98), we switch to Half-Space Step roughly after 30 epoches. Table 9 shows that under each setting, the proposed HSPG correctly identifies the groups of zeros as indicated by $\text{IoU}(\hat{\boldsymbol{x}}, \boldsymbol{x}^*) = 1.0$, which is a strong evidence to show the correctness of group sparsity identification of HSPG.

Table 9: Linear regression problem settings and IoU of the recovered solutions by HSPG.

	N	n	Group sparsity ratio of $\boldsymbol{x}^*$	$\text{IoU}(\hat{x}, x^*)$
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

Logistic Regression We then focus on the benchmark convex logistic regression problem with the mixed ℓ_1/ℓ_2-regularization given N examples $(\boldsymbol{d}_1, l_1), \cdots, (\boldsymbol{d}_N, l_N)$ where $\boldsymbol{d}_i \in \mathbb{R}^n$ and $l_i \in \{-1, 1\}$ with the form

$$\underset{(\boldsymbol{x};b)\in\mathbb{R}^{n+1}}{\text{minimize}}\ \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-l_i(\boldsymbol{x}^T\boldsymbol{d}_i+b)}) + \lambda \sum_{g\in\mathcal{G}} \|[\boldsymbol{x}]_g\|\,, \tag{104}$$

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 (104) on 8 standard publicly available large-scale datasets from LIBSVM repository (6) as summarized in Table 10. 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 following (8). The mini-batch size $|\mathcal{B}|$ is set to be $\min\{256, \lceil 0.01N \rceil\}$ similarly to (93). The step size α_k setting follows [Section 4](88). Particularly, we first compute a Lipschitz constant L as $\max_i \|\boldsymbol{d}_i\|^2/4$, 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.

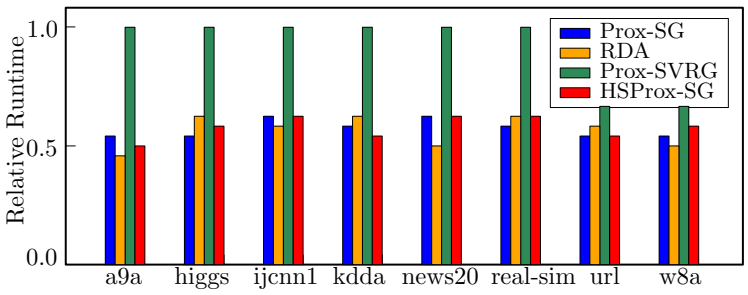


Figure 7: Relative runtime.

For HSPG, we set α_k as the same as Prox-SG and Prox-SVRG in practice. We select two ϵ's as 0 and 0.8. The final objective value ψ and group sparsity in the solutions are reported in Table 11-12, where we mark the best values as bold to facilitate the comparison. Furthermore, Figure 7 plots the relative runtime of these solvers for each dataset, scaled by the runtime of the most time-consuming solver.

Table 12 shows that our HSPG is definitely the best solver on exploring the group sparsity of the solutions. In fact, HSPG under $\epsilon = 0.8$ 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 **??** even under ϵ as 0. Moreover, as shown in Table 11, 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 7 indicates that Prox-SG, RDA and HSPG have similar computational cost to proceed, except Prox-SVRG due to its periodical full gradient computation.

Table 10: 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 $\{0, 1\}$

Table 11: Final objective values ψ for tested algorithms on convex problems.

Dataset	Prox-SG	RDA	Prox-SVRG	HSPG	
				ϵ as 0	ϵ as 0.8
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 12: Group sparsity for tested algorithms on convex problems.

Dataset	Prox-SG	RDA	Prox-SVRG	HSPG	
				ϵ as 0	ϵ as 0.8
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%**

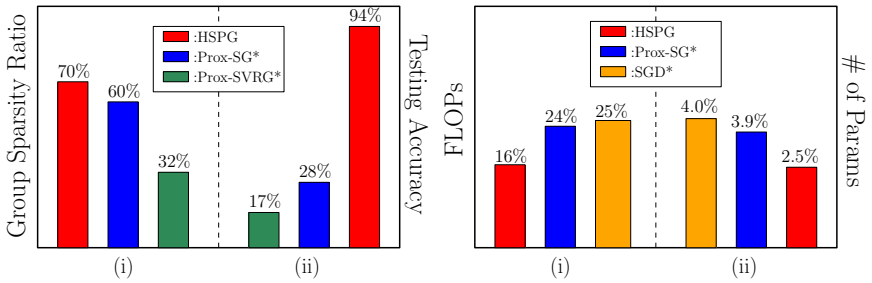
C.2 Nonconvex Experiments

To illustrate, among the state-of-the-art proximal stochastic optimizers, we exclude RDA because of no acceptable results attained during our following tests with the step size parameter γ setting throughout all powers of 10 from 10^{-3} to 10^{3}, and skip Prox-Spider and SAGA since Prox-SVRG has been a superb representative to the proximal incremental gradient methods. We consider the popular image classification tasks, with popular architectures, *i.e.*, VGG16 and ResNet18 on benchmark datasets CIFAR10 and Fashion-MNIST (86), where the group partition $\mathcal{G}$ is defined as 3D kernel following (14; 58), which are not ZIGs.

Table 13: Final ψ/group sparsity ratio/testing accuracy on non-convex problems over non-ZIGs.

Backbone	Dataset	Prox-SG	Prox-SVRG	HSPG
VGG16	CIFAR10	0.59 / 52.58% / 90.50%	0.85 / 14.13% / 89.16%	**0.58** / **76.47%** / **91.93%**
	Fashion-MNIST	0.52 / 12.31% / **92.83%**	2.66 / 0.38% / 92.72%	**0.52** / **47.82%** / 92.87%
ResNet18	CIFAR10	**0.31** / 20.27% / 94.36%	0.37 / 4.60% / 94.11%	**0.31** / **69.98%** / **94.40%**
	Fashion-MNIST	0.14 / 0.00% / **94.94%**	0.18 / 0.00% / 94.70%	**0.13** / **77.08%** / 94.61%
MobileNetV1	CIFAR10	**0.40** / 58.05% / 91.54%	0.65 / 29.20% / 89.68%	**0.40** / **71.36%** / 92.04%
	Fashion-MNIST	**0.22** / 62.62% / 94.22%	0.40 / 41.99% / 94.19%	0.26 / **84.26%** / **94.52%**

Table 13 demonstrates the effectiveness and superiority of HSPG, where we mark the best values as bold, and the group sparsity ratio is defined as the percentage of zero groups. In particular, *(i)* HSPG computes remarkably higher group sparsity than other methods on all tests, of which the solutions are typically multiple times sparser in the manner of group than those of Prox-SG, while Prox-SVRG performs not comparably since the variance reduction techniques may not work as desired for deep learning applications (13); *(ii)* HSPG performs competitively with respect to the final objective values ψ. In addition, all the methods reach a comparable generalization performance on unseen test data. On the other hand, sparse regularization methods may yield solutions with entries that are not exactly zero but are very small. Sometimes all entries below certain threshold ($\mathcal{T}$) are set to zero (44; 20). However, such simple truncation mechanism is heuristic-rule based, hence may hurt convergence and accuracy. To illustrate this, we set the groups of the solutions of Prox-SG and Prox-SVRG to zero if the magnitudes of the group variables are less than some $\mathcal{T}$, and denote the corresponding solutions as Prox-SG* and Prox-SVRG*.



(a) HSPG VS Truncation over non-ZIGs. (b) HSPG VS Truncation over ZIGs.

Figure 8: HSPG versus simple truncation. (a) On ResNet18 with CIFAR10 over non-ZIGs. (b) On VGG16 with CIFAR10 over ZIGs.

As shown in Figure 8a(i), under the $\mathcal{T}$ with no accuracy regression, Prox-SG* and Prox-SVRG* reach higher group sparsity ratio as 60% and 32% compared to Table 13, but still significantly lower than the 70% of HSPG without simple truncation. Under the $\mathcal{T}$ to reach the same group sparsity ratio as HSPG, the testing accuracy of Prox-SG* and Prox-SVRG* regresses drastically to 28% and 17% in Figure 8a(ii) respectively. Remark here that although further refitting the models from Prox-SG* and Prox-SVRG* on active (non-zero) groups of weights may recover the accuracy regression, it requires additional engineering efforts and training cost, which is less attractive and convenient than HSPG (with no need to refit). Similarly, as shown in Figure 8b, under the ZIG partition and the $\mathcal{T}$ without accuracy regression, the FLOPs and number of parameters reductions achieved by SGD* (subgradient descent with simple truncation) and Prox-SG* are not comparable with those achieve by HSPG, *i.e.*, HSPG achieves about $1.5\times$ fewer FLOPs and number of parameters.