Appendices

A Proofs

This part contains the proofs of Lemma 3.1 and Theorem 3.2. For notational convenience, we use $f(\cdot)$ to substitute the $R_s(\cdot)$, and use $\boldsymbol{x}$ to substitute the $\boldsymbol{\gamma}$, in Equations (2) and (3) of Section 3.2. We also restate Lemma 3.1 and Theorem 3.2 using the new notations here, so that this part can be self-contained. Under the new notations here, Equation (2) in Section 3.2 becomes:

$$f(\boldsymbol{x}) = t\|\boldsymbol{x}\|_1 - \|\boldsymbol{x} - \bar{x}\mathbf{1}_n\|_1 = \sum_{i=1}^{n} t|x_i| - |x_i - \bar{x}|, (t \in \mathbb{R}, x_i \in [0, a]) \tag{5}$$

And Lemma 3.1 becomes:

Lemma A.1. *$f(\boldsymbol{x})$ defined in (5) is concave.*

Proof. For any $0 < \theta < 1$ and $x_i, y_i \in [0, a]$, we have:

$$\begin{aligned} &|\theta x_i + (1-\theta)y_i - [\theta\bar{x} + (1-\theta)\bar{y}]| \\ =&|\theta(x_i - \bar{x}) + (1-\theta)(y_i - \bar{y})| \\ \leq&\theta|x_i - \bar{x}| + (1-\theta)|y_i - \bar{y}| \end{aligned} \tag{6}$$

Using $t(\theta x_i + (1-\theta)y_i)$ to subtract both sides of inequality (6), we have:

$$\begin{aligned} &t(\theta x_i + (1-\theta)y_i) - |\theta x_i + (1-\theta)y_i - [\theta\bar{x} + (1-\theta)\bar{y}]| \\ \geq&\theta\left(tx_i - |x_i - \bar{x}|\right) + (1-\theta)\left(ty_i - |y_i - \bar{y}|\right) \end{aligned}$$

Therefore,

$$\begin{aligned} &\sum_{i=1}^{n} t(\theta x_i + (1-\theta)y_i) - |\theta x_i + (1-\theta)y_i - [\theta\bar{x} + (1-\theta)\bar{y}]| \\ \geq&\theta\sum_{i=1}^{n} tx_i - |x_i - \bar{x}| + (1-\theta)\sum_{i=1}^{n} ty_i - |y_i - \bar{y}| \end{aligned}$$

Note that when $x_i, y_i \in [0, a]$,

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{n} tx_i - |x_i - \bar{x}| \\ f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) &= \sum_{i=1}^{n} t(\theta x_i + (1-\theta)y_i) - |\theta x_i + (1-\theta)y_i - [\theta\bar{x} + (1-\theta)\bar{y}]| \end{aligned}$$

So we have

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \geq \theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y})$$

which shows that $f(\boldsymbol{x})$ is concave on $\boldsymbol{x} \in [0, a]^n$.

□

Under the new notations here, Equation (3) in Subection 3.2 becomes:

$$\min_{\boldsymbol{x}\in[0,a]^n} f(\boldsymbol{x}) = \sum_{i=1}^{n} t|x_i| - |x_i - \bar{x}|, (t \in \mathbb{R}) \tag{7}$$

Given k indices: $1 \leq i_1, \cdots, i_k \leq n$, and k values: $v_1, \cdots, v_k \in \{0, a\}$, we define the notation:

$$\mathbb{F}_{i_{1,\cdots,k}, v_{1,\cdots,k}} = \{\boldsymbol{x} \in [0, a]^n | x_{i_1} = v_1, \cdots, x_{i_k} = v_k\} \tag{8}$$

Then we prove the following lemma:

Lemma A.2. *The optimal value of (7) is attained on the vertices of* $[0,a]^n$.

Proof. $[0,a]^n$ is an n-dimensional cube, and its boundary consists of $2n$ facets:

$$\begin{aligned}\partial[0,a]^n &= \bigcup_{i=1}^{n} (\{\boldsymbol{x} \in [0,a]^n | x_i = 0\} \cup \{\boldsymbol{x} \in [0,a]^n | x_i = a\}) \\ &= \bigcup_{i=1}^{n} (\mathbb{F}_{i,0} \cup \mathbb{F}_{i,a})\end{aligned}$$

where the notation of (8) is used. Obviously, $\mathbb{F}_{i,v} (v = 0 \text{ or } a)$ is a facet of $[0,a]^n$. Note that $\mathbb{F}_{i,v} = \{\boldsymbol{x} | x_i = v, \forall j \neq i, x_j \in [0,a]\}$. So $\mathbb{F}_{i,v}$ is a $n-1$ dimensional cube on the dimensions $1 \leq j \leq n, j \neq i$. Actually, $\mathbb{F}_{i_{1,\cdots,k}, v_{1,\cdots,k}}$ is a $n-k$ dimensional cube on dimensions $1 \leq j \leq n, j \notin \{i_{1,\cdots,k}\}$.

First step, we prove that the optimal value of (7) is attained on one facet of $[0,a]^n$. Assume $\boldsymbol{x}^* \in [0,a]^n$ is an optimal solution of (7). Because $[0,a]^n$ is a bounded and closed convex set on $\mathbb{R}^n$, then if $\boldsymbol{x}^* \notin \partial[0,a]^n$, there exists a line segment in $[0,a]^n$ passing through $\boldsymbol{x}^*$ and intersects with the boundary of $[0,a]^n$ on two points: $\boldsymbol{x}_1^{(n)}, \boldsymbol{x}_2^{(n)}$, i.e. $\exists \boldsymbol{x}_1^{(n)}, \boldsymbol{x}_2^{(n)} \in \partial[0,a]^n$, and $\exists \theta \in (0,1)$, such that $\boldsymbol{x}^* = \theta \boldsymbol{x}_1^{(n)} + (1-\theta)\boldsymbol{x}_2^{(n)}$. According to Lemma A.1, $f(\boldsymbol{x})$ is concave on $[0,a]^n$. Therefore $f(\boldsymbol{x}^*) \geq \theta f(\boldsymbol{x}_1^{(n)}) + (1-\theta) f(\boldsymbol{x}_2^{(n)}) \geq \min\{f(\boldsymbol{x}_1^{(n)}), f(\boldsymbol{x}_2^{(n)})\}$. Without loss of generality, suppose $f(\boldsymbol{x}_1^{(n)}) \leq f(\boldsymbol{x}_2^{(n)})$, then $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}_1^{(n)})$, which means $\boldsymbol{x}_1^{(n)}$ is also an optimal solution. Because $\boldsymbol{x}_1^{(n)} \in \partial[0,a]^n$, $\boldsymbol{x}_1^{(n)}$ must lie on one facet of $[0,a]^n$. Let $\mathbb{F}_{i_1,v_1}$ denote the facet that $\boldsymbol{x}_1^{(n)}$ lies on.

Second step, note that $\mathbb{F}_{i_1,v_1}$ is an $n-1$ dimensional cube, $\boldsymbol{x}_1^{(n)} \in \mathbb{F}_{i_1,v_1}$ is an optimal solution of (7), and $f(\boldsymbol{x})$ is concave on $\mathbb{F}_{i_1,v_1}$. For the same reasons as in the first step, the optimal value of (7) is attained on one facet of $\mathbb{F}_{i_1,v_1}$. Without loss of generality, suppose $\mathbb{F}_{i_{1,2},v_{1,2}}$ is the facet that the optimal solution lies on. Obviously, $\mathbb{F}_{i_{1,2},v_{1,2}}$ is an $n-2$ dimensional cube.

We can iteratively apply the same reasoning as the 2 steps above until we reach the n-th step, in which case the optimal solution $\boldsymbol{y}^* = (y_1^*, y_2^*, \cdots, y_n^*)$ lies on a 1-dimensional cube $\mathbb{F}_{i_{1,\cdots,n-1}, v_{1,\cdots,n-1}}$. Let $\{j\} = \{1,\cdots,n\} \setminus \{i_{1,\cdots,n-1}\}$, then $y_j^* \in [0,a]$, and $\forall i \neq j (1 \leq i \leq n), y_i^* \in \{0,a\}$. Suppose $y_j^* = \xi a, (0 \leq \xi \leq 1)$, and let $\boldsymbol{y}_1 = (y_1^*, \cdots, y_{j-1}^*, a, y_{j+1}^*, \cdots, y_n^*)$ and $\boldsymbol{y}_2 = (y_1^*, \cdots, y_{j-1}^*, 0, y_{j+1}^*, \cdots, y_n^*)$, then both $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are vertices of $[0,a]^n$, and $\boldsymbol{y}^* = \xi \boldsymbol{y}_1 + (1-\xi)\boldsymbol{y}_2$. Because $f(\boldsymbol{x})$ is concave, we have $f(\boldsymbol{y}^*) \geq \xi f(\boldsymbol{y}_1) + (1-\xi) f(\boldsymbol{y}_2) \geq \min\{f(\boldsymbol{y}_1), f(\boldsymbol{y}_2)\}$, i.e. $f(\boldsymbol{y}^*) \geq f(\boldsymbol{y}_1)$ or $f(\boldsymbol{y}^*) \geq f(\boldsymbol{y}_2)$. Because both $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are vertices of $[0,a]^n$, this shows that the optimal value is attained on vertices of $[0,a]^n$. □

Under the new notations here, Theorem 3.2 in Subection 3.2 becomes:

Theorem A.3. *A class of optimal solutions of (7) is that either* $\lfloor n\rho \rfloor$ *or* $\lfloor n\rho \rfloor + 1$ *number of* $x_i, (1 \leq i \leq n)$ *are a, and the rest are 0, where:*

$$\rho = \begin{cases} -t/4 + 1/2, & -2 \leq t \leq 2 \\ 0, & t > 2 \\ 1, & t < -2 \end{cases} \tag{9}$$

Proof. By Lemma A.2, the optimal solution of (7) can be found on the vertices of $[0,a]^n$: $\{\boldsymbol{x} | x_i \in \{0,a\}, 1 \leq i \leq n\}$. So (7) is solved by:

$$\min_{x_i \in \{0,a\}} f(\boldsymbol{x}) = \sum_{i=1}^{n} t|x_i| - |x_i - \bar{x}| \tag{10}$$

Suppose k variables of $x_i, (1 \leq i \leq n)$ equal a, and the rest $n-k$ variables equal 0. Then

$$\begin{aligned} f(\boldsymbol{x}) &= a\left(tk - \left[k|1 - \frac{k}{n}| + (n-k)|0 - \frac{k}{n}|\right]\right) \\ &= a\left(k(t - 1 + \frac{k}{n}) - (n-k)\frac{k}{n}\right) \end{aligned}$$

This shows that k determines the value of $f(\boldsymbol{x})$, which is permutation invariant. And because $a > 0$, solving (10) is equivalent to solving the following problem:

$$\min_{1\leq k\leq n} g(k) = k(t-1+\frac{k}{n}) - (n-k)\frac{k}{n} \tag{11}$$

Let $\rho = k/n$, and define

$$\begin{aligned} h(\rho) = \frac{g(k)}{n} &= \rho(t-1+\rho) - (1-\rho)\rho \\ &= 2\rho^2 + (t-2)\rho \end{aligned}$$

Because k can only be integers, the values of ρ are discrete. We first relax this discreteness constraint and assume ρ is a continuous variable in $[0, 1]$. Then (11) is simplified to:

$$\min_{0\leq\rho\leq 1} h(\rho) = 2\rho^2 + (t-2)\rho \tag{12}$$

Solving (12), we get the optimal solution:

$$\rho^* = \begin{cases} -t/4 + 1/2, & -2 \leq t \leq 2 \\ 0, & t > 2 \\ 1, & t < -2 \end{cases}$$

Now we consider the discreteness constraint. When $t > 2$, the optimal solution of (11) is 0. When $t < -2$, the optimal solution of (11) is n. When $t \in [-2, 2]$, $\rho^* \in [0, 1]$. If $\rho^* = m/n$, where $m \in \{0, 1, \cdots, n\}$, then the optimal solution of (11) is m. Otherwise, ρ^* must lie in some interval $(m/n, (m+1)/n)$, where $m \in \{0, 1, \cdots, n-1\}$. Because $\rho = \rho^*$ is the symmetric axis of the quadratic function (12), the minimal value of (12) on the discrete points $\{0, 1/n, 2/n, \cdots, 1\}$ must be attained either on m/n or on $(m+1)/n$. So the optimal solution of (11) is either m or $m+1$. □

B Detailed Parameters in Training

The detailed parameters in the training and pruning stages of our method are listed in Table 4. We use the most commonly used batch size, training epochs and learning rate decay scheme as in [Zagoruyko and Komodakis, 2016, Liu et al., 2017]. For example, in the training stage on ImageNet, the learning rate decay scheme "0.1,0.01,0.001@30,60" means that the learning rate in the first 30 epochs is 0.1, in epochs 31 to 60 is 0.01, and in epochs after 60 is 0.001. And "0.05 cosine lr decay" means using cosine learning rate decay strategy with the initial learning rate 0.05.

Table 4: Training hyper-parameters for sparsity training and fine-tuning. Learning rate denotes the learning rates and corresponding decay milestones.

Dataset	Model	Batch size	Training		Fine-tuning	
			Epoch	Learning rate	Epoch	Learning rate
ImageNet	ResNet	512	120	$\{0.1, 0.01, 0.001\}$ @{40,80}	128	0.05 cosine lr decay [Loshchilov and Hutter, 2017]
ImageNet	MobileNet v2	1024	256	0.4 cosine lr decay with warmup	256	0.05 cosine lr decay
CIFAR	ResNet & VGG	128	200	$\{0.01, 0.1, 0.02, 0.004, 0.0008\}$ @{1,60,120,160}	200	$\{0.001, 0.0005, 0.00025, 0.0001\}$ @{20,60,150}

C Details on Pruning ResNet-50

The main building block of ResNet-50 is the bottleneck block [He et al., 2016], as shown in Figure 5. The polarization regularizer is applied to the scale factors of the first two BN layers in each bottleneck block, i.e. the "`bn1`" and "`bn2`" in Figure 5. In the baseline ResNet-50 model, many scaling factors in the "`bn3`" of a bottleneck block are already 0 or very close to 0. So we do not apply any extra regularization on "`bn3`", either in our polarization pruning method or in our implementation of the NS method [Liu et al., 2017]. We visualize the layer-wise distribution of scaling factors in Figure 6. Figure 6 compares the layer-wise distributions of scaling factors between the baseline ResNet-50 model and the model trained with our polarization regularizer on ImageNet dataset. Due to space limit, we only visualize the distributions in the "conv2" and "conv3" layers as defined in [He et al., 2016], where "conv2" and "conv3" contains 3 and 4 bottleneck blocks respectively. Figure 6 shows that our polarization regularizer also makes the scaling factors more separable in layer-level.

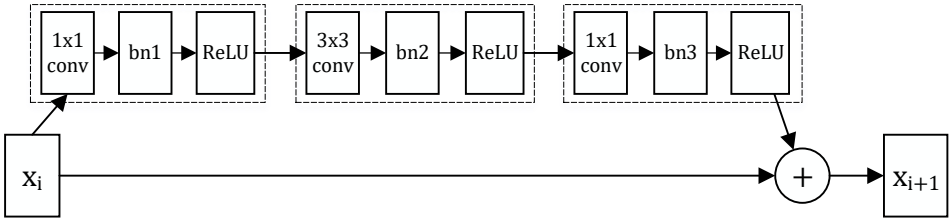

Figure 5: A "bottleneck" block in ResNet-50.

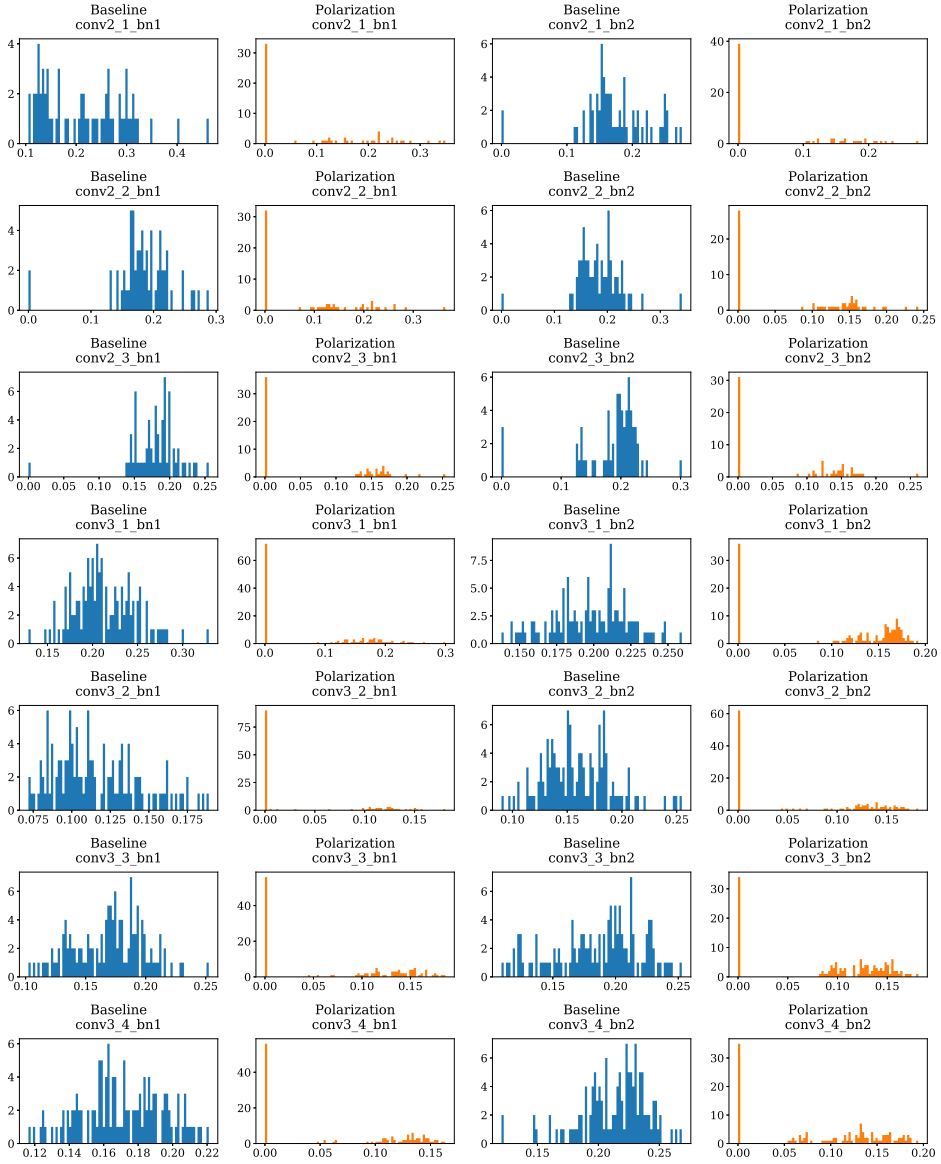

Figure 6: Comparison of the layer-wise scaling factor distributions between baseline ResNet-50 model and the model trained with our polarization regularizer on ImageNet dataset. The left two columns show the distributions of scaling factors in "bn1" of the bottleneck block. The right two columns show the distributions of scaling factors in "bn2" of the bottleneck block. The top 3 rows show scaling factor distributions in each bottleneck block of the "conv2" layer. The bottom 4 rows show scaling factor distributions in each bottleneck block of the "conv3" layer.