Appendix

A Limitations and Future Work

Theory We leverage our proposed PQ Index and derive a PQI-bound to indicate the number of retained parameters r. Our result is a lower bound for r, which only suggests the maximum number of model parameters we should prune. One potential future work is to develop an upper for r so that we can better understand the relationship between sparsity and pruning. Furthermore, our result introduces an additional term η_r, which is unavailable before determining the pruning ratio. We treat it as a hyper-parameter in our algorithm and experiments. However, it is desirable to develop a tighter bound without such approximation. Additionally, how p and q together impact the sparsity measure and model pruning has not been thoroughly analyzed. Finally, the theoretical justification of the proposed hypothesis is lacking.

Method Our corroborated hypothesis indicates that a dynamic relationship exists between the model's sparsity and compressibility. However, our proposed SAP algorithm determines the number of pruned parameters based on a static PQI-bound at each iteration. Thus, one potential future work is to further develop the SAP algorithm by considering the dynamics of sparsity. For example, stopping pruning when the PQI starts to increase or the pruning ratio is below some threshold. In this work, we demonstrate the relationship between the performance of iterative pruning and the dynamics of PQI. It is interesting to analyze the critical factors that may determine such dynamics, such as initialization and model architecture. Recent works also introduce gradual magnitude pruning which can outperform iterative pruning algorithms Gale et al. (2019); Renda et al. (2020). Therefore, it is also interesting to study how PQI are related to various pruning methods Evci et al. (2020); Hoefler et al. (2021). We demonstrate that SAP with proper choice of hyper-parameters can outperform LT. It is interesting to explore when SAP can outperform LT with respect to the accuracy-compression trade-off.

Application We apply the proposed PQ Index for model compression because model compression is one of the most important topics related to sparsity. However, many other interesting topics could leverage the PQ Index. For example, one may consider directly optimizing the objective function and PQ Index together for regularization. Furthermore, fairness that advocates similar performances of various groups may also benefit from our PQ Index. Finally, other fields using Gini Index, such as sociology and economics, may also find the proposed alternative index interesting.

B Theoretical Analysis

Proof of Theorem 1:
Note that for any $0 < p < q$, Hölder's inequality gives that

$$\|w\|_q \leq \|w\|_p \leq d^{\frac{1}{p}-\frac{1}{q}}\|w\|_q.$$

We immediately obtain $\mathrm{I}(w) \in [0, 1 - d^{\frac{1}{q}-\frac{1}{p}}]$ from the inequality above.

Next, we prove that I satisfies six properties, which implies that $\mathrm{I}(w)$ is larger if w is sparser. Hurley & Rickard (2009) prove that (P1) and (P2) are automatically satisfied as long as (D1)-(D4) are met. Furthermore, we note that $f(x) = 1 - 1/x$ is monotonous for $x > 0$. Therefore, we only have to prove (D1)-(D4) hold for $S(w) = d^{\frac{1}{p}-\frac{1}{q}}\|w\|_q/\|w\|_p$.

(D2) Scaling is automatically satisfied for $S(w)$ since ℓ_q-norm is homogeneous for any $q > 0$, and (D4) Cloning is clear from the definition of $S(w)$.

For (D1) Robin Hood, we only need to prove that for any $w_i > w_j > 0$, the derivative of $f(t)$ at $t = 0$ is negative, where

$$f(t) = \frac{\{(w_i - t)^q + (w_j + t)^q\}^{1/q}}{\{(w_i - t)^p + (w_j + t)^p\}^{1/p}}.$$

Let $a = w_i - t$ and $b = w_j + t$. Take the derivative with respect to t, we have

$$\begin{aligned} f'(t) &= \frac{(a^q + b^q)^{\frac{1}{q}-1}(-a^{q-1} + b^{q-1})(a^p + b^p)^{1/p}}{(a^p + b^p)^{2/p}} \\ &\quad - \frac{(a^q + b^q)^{1/q}(a^p + b^p)^{\frac{1}{p}-1}(-a^{p-1} + b^{p-1})}{(a^p + b^p)^{2/p}} \\ &= f(t)\left[\frac{-a^{q-1} + b^{q-1}}{a^q + b^q} - \frac{-a^{p-1} + b^{p-1}}{a^p + b^p}\right] \\ &= f(t)\frac{(a^{p-q} - b^{p-q})(a+b)a^{q-1}b^{q-1}}{(a^q + b^q)(a^p + b^p)}. \end{aligned}$$

Note that when $t = 0$, $f(0) > 0$ and $a > b > 0$. Thus, $f'(t) < 0$.

Similarly, for (D3) Rising tide, we need to show that $f'(t)$ is negative, where

$$f(t) = \frac{(\sum_{i=1}^d (w_i + t)^q)^{1/q}}{(\sum_{i=1}^d (w_i + t)^p)^{1/p}}.$$

We can verify that

$$f'(0) = f(0)\left[\frac{\sum_{i=1}^d w_i^{q-1}}{\sum_{i=1}^d w_i^q} - \frac{\sum_{i=1}^d w_i^{p-1}}{\sum_{i=1}^d w_i^p}\right].$$

Thus, we conclude the proof by showing that $h(t) = (\sum_{i=1}^d w_i^{t-1})/(\sum_{i=1}^d w_i^t)$ is a monotonously decreasing function for $t > 0$. This is done by showing $h'(t) < 0$ for all $t > 0$. Actually, since $w_i \geq 0$ and w_i's are not all the same, we know

$$\begin{aligned} h'(t) &= \frac{(\sum_{i=1}^d w_i^{t-1}\ln(w_i))(\sum_{i=1}^d w_i^t) - (\sum_{i=1}^d w_i^{t-1})(\sum_{i=1}^d w_i^t \ln(w_i))}{(\sum_{i=1}^d w_i^t)^2} \\ &= \frac{\sum_{1\leq i<j\leq d}(w_i - w_j)(\ln(w_i) - \ln(w_j))w_i^{t-1}w_j^{t-1}}{(\sum_{i=1}^d w_i^t)^2} < 0. \end{aligned}$$

Proof of Theorem 2:

Recall that M_r is the largest r components of w, and η_r is a constant such that $\sum_{i\notin M_r} |w_i|^p \leq \eta_r \sum_{i\in M_r} |w_i|^p$. Therefore,

$$\begin{aligned} \|w\|_p &= \left(\sum_{1\leq i\leq d} |w_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i\in M_r} |w_i|^p + \sum_{i\notin M_r} |w_i|^p\right)^{\frac{1}{p}} \\ &\leq \left(\sum_{i\in M_r} |w_i|^p + \eta_r \sum_{i\in M_r} |w_i|^p\right)^{\frac{1}{p}} = \left(\sum_{i\in M_r} |w_i|^p\right)^{\frac{1}{p}} (1+\eta_r)^{\frac{1}{p}} \\ &\leq \left(\sum_{i\in M_r} |w_i|^q\right)^{\frac{1}{q}} r^{\frac{1}{p}-\frac{1}{q}} (1+\eta_r)^{\frac{1}{p}} \leq \|w\|_q r^{\frac{1}{p}-\frac{1}{q}} (1+\eta_r)^{\frac{1}{p}}. \end{aligned}$$

Rearranging the above inequality gives

$$r \geq d(1+\eta_r)^{-q/(q-p)}[1 - \mathrm{I}(w)]^{\frac{qp}{q-p}}.$$

C EXPERIMENTAL SETUP

Table 1 and 2 summarizes the model architecture of MLP and CNN used in our experiments. Table 3 shows the statistics of model architecture and hyper-parameters used in our experiments.

Table 1: The model architecture of Multi-Layer Perceptron (MLP) used in our experiments. The n_c, H, W represent the shape of images, namely the number of image channels, height, and width, respectively. K is the number of classes in the classification task. The ReLU layers follow Linear(input channel size, output channel size) layers, apart from the last one.

Image $x \in \mathbb{R}^{n_c \times H \times W}$
Linear($n_c \times H \times W$, 128)
Linear(128, 256)
Linear(256, K)

Table 2: The model architecture of Convolutional Neural Networks (CNN) used in our experiments. The n_c, H, W represent the shape of images, namely the number of image channels, height, and width, respectively. K is the number of classes in the classification task. The BatchNorm and ReLU layers follow Conv2d(input channel size, output channel size, kernel size, stride, padding) layers. The MaxPool2d(output channel size, kernel size) layer reduces the height and width by half.

Image $x \in \mathbb{R}^{n_c \times H \times W}$
Conv2d(n_c, 64, 3, 1, 1)
MaxPool2d(64, 2)
Conv2d(64, 128, 3, 1, 1)
MaxPool2d(128, 2)
Conv2d(128, 256, 3, 1, 1)
MaxPool2d(256, 2)
Conv2d(256, 512, 3, 1, 1)
MaxPool2d(512, 2)
Global Average Pooling
Linear(512, K)

Table 3: Statistics of the models and hyper-parameters used in our experiments for training and pruning.

Dataset		FashionMNIST				CIFAR10			
Model Architecture		Linear	MLP	CNN	ResNet18	Linear	MLP	CNN	ResNet18
Model Size		7.9 K	136.1 K	1.6 M	11.2 M	30.7 K	428.9 K	1.6 M	11.2 M
FLOPS		3.9 M	67.8 M	20.1 G	114.4 G	15.4 M	214.2 M	29.4 G	139.4 G
Train	Epoch E	200							
	Batch size	250							
	Optimizer	SGD							
	Learning rate	1E-01							
	Momentum	0.9							
	Weight decay	5E-04							
	Nesterov	✓							
	Scheduler	Cosine Annealing (Loshchilov & Hutter, 2016)							
Prune	T	30		15		30		15	
	P	0.2							

D EXPERIMENTAL RESULTS

D.1 PQ INDEX

We visualize the PQ Index of pruned models at the global scale with various combinations of p and q. We use zero to indicate the numerical overflow may happen when $p = 0.1$. Note that we use $p = 0.5$ and $q = 1.0$ to compute PQ Index for other figures. The results show that various combinations of p and q also corroborate our hypothesis in different scales, e.g (d) One Shot in Figure 7 and (b) SAP ($p = 1.0$, $q = 2.0$) of Figure 8.

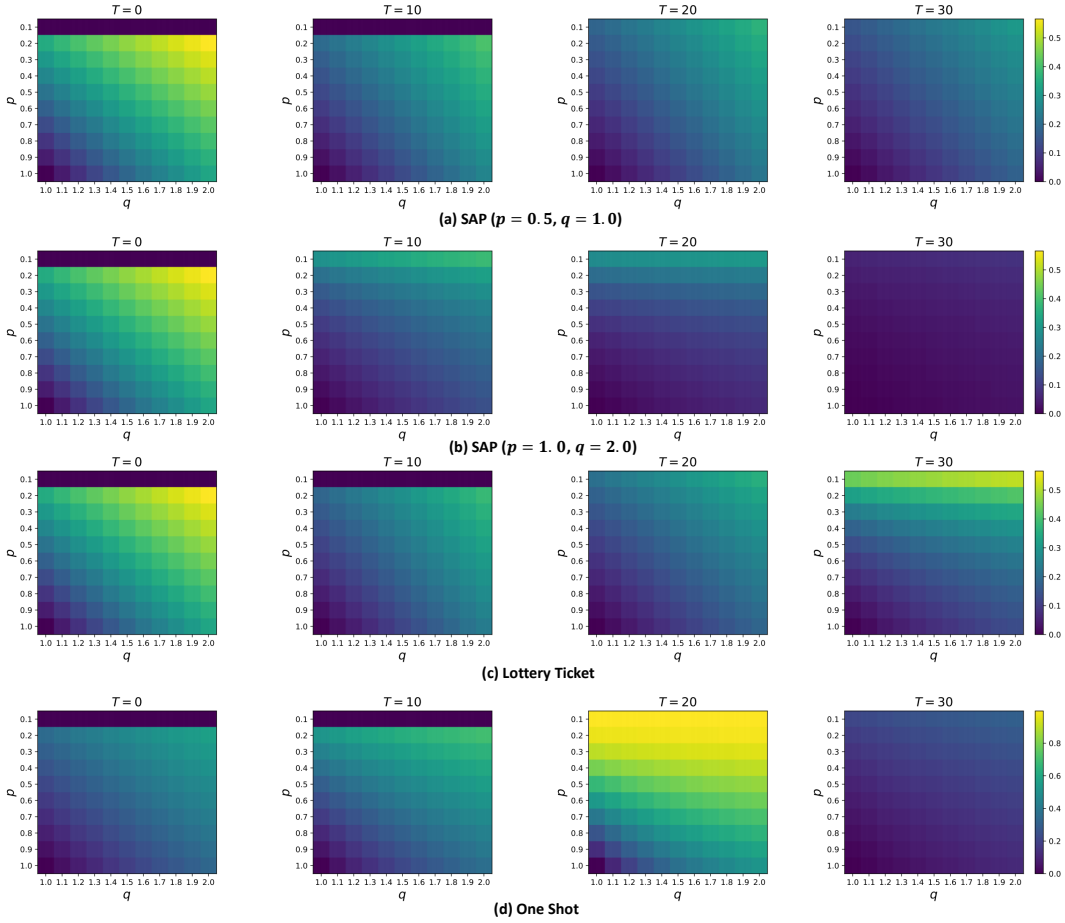

Figure 7: Results of PQ index visualized with various combinations of p and q for FashionMNIST and MLP.

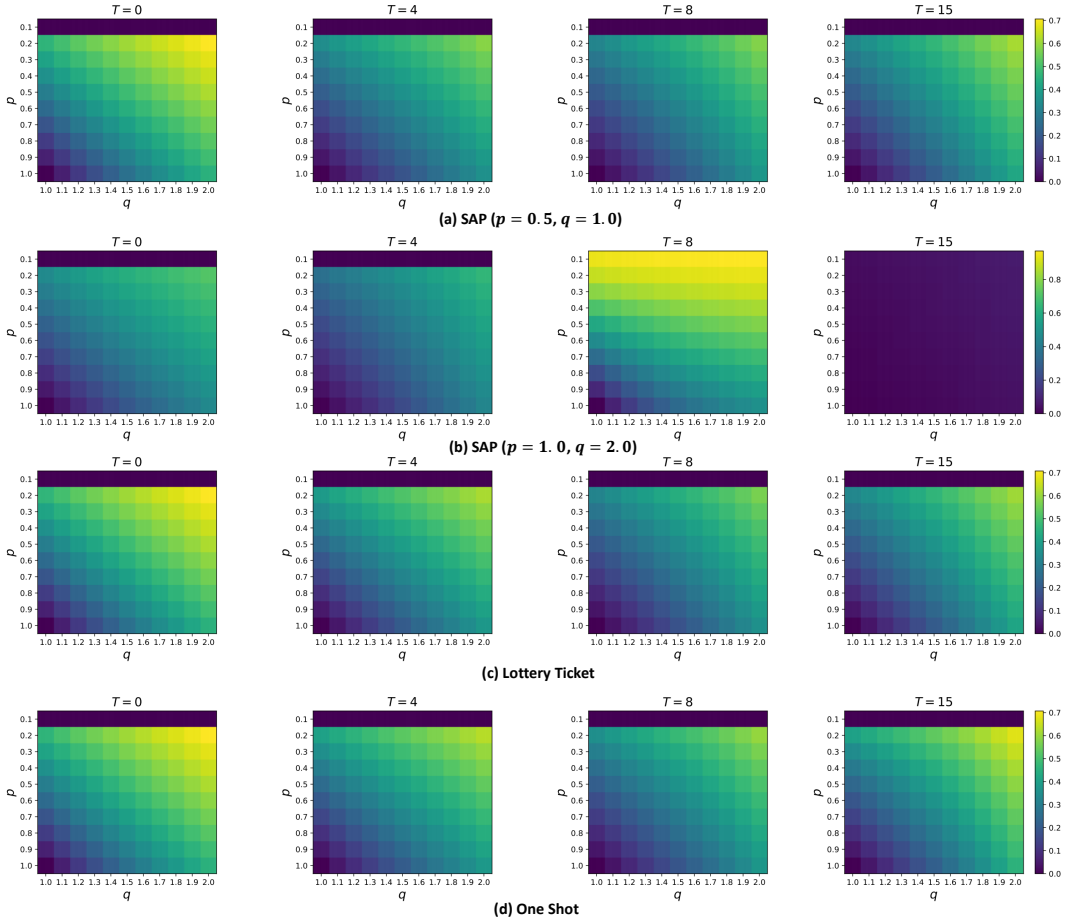

Figure 8: Results of PQ index visualized with various combinations of p and q for CIFAR10 and ResNet18.

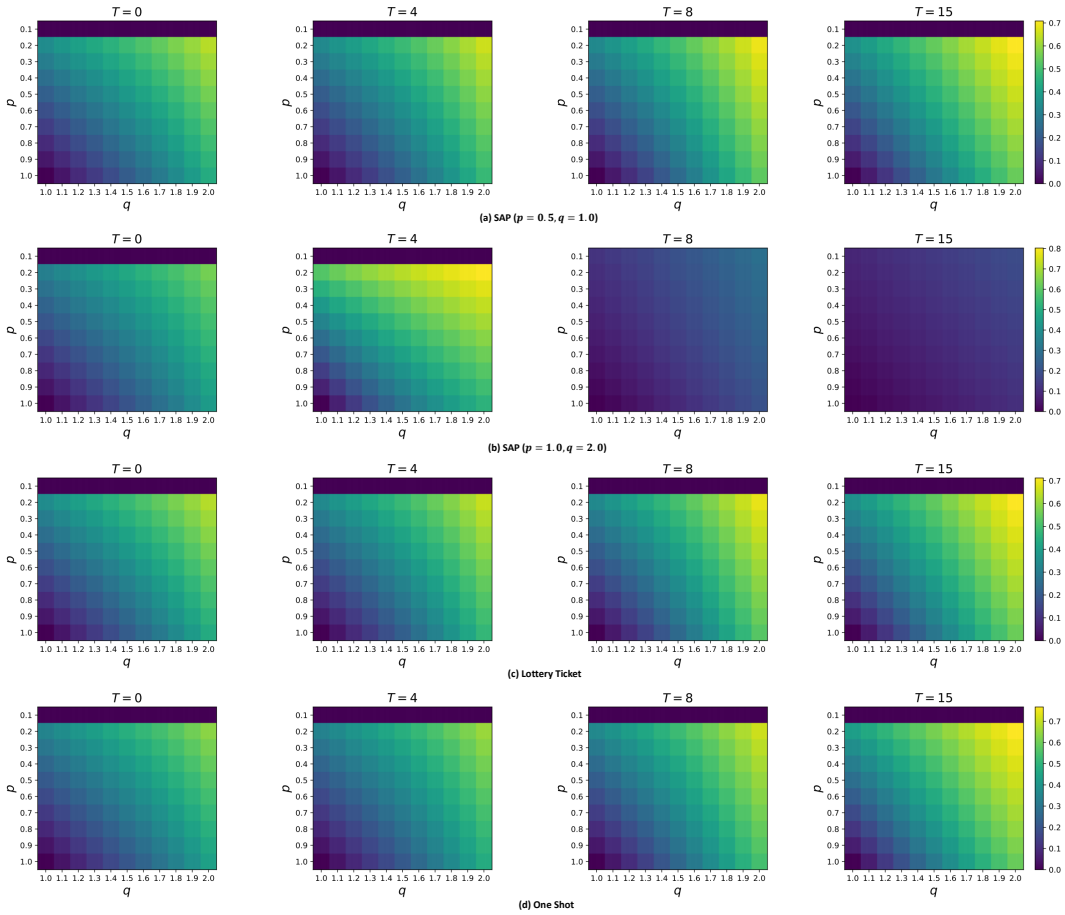

Figure 9: Results of PQ index visualized with various combinations of p and q for CIFAR100 and WResNet28x8.

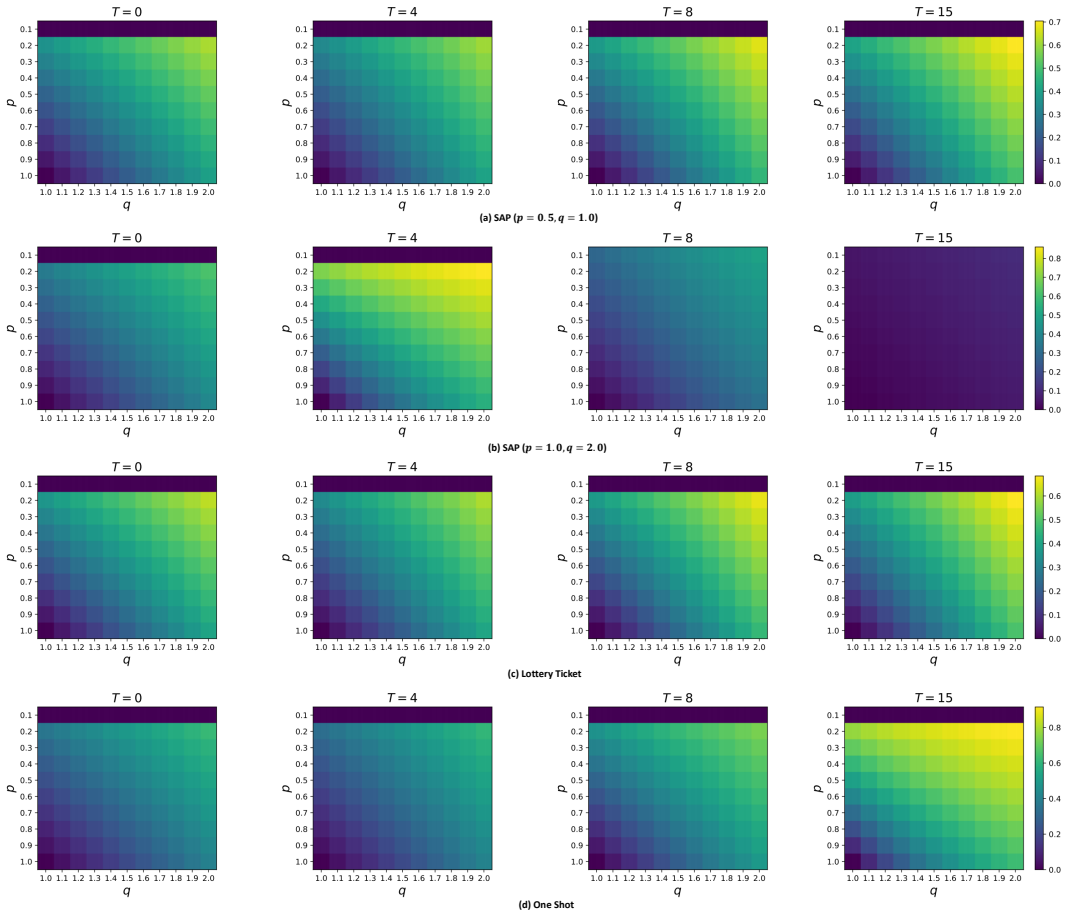

Figure 10: Results of PQ index visualized with various combinations of p and q for TinyImageNet and ResNet50.

D.2 RETRAINED AND PRUNED MODELS

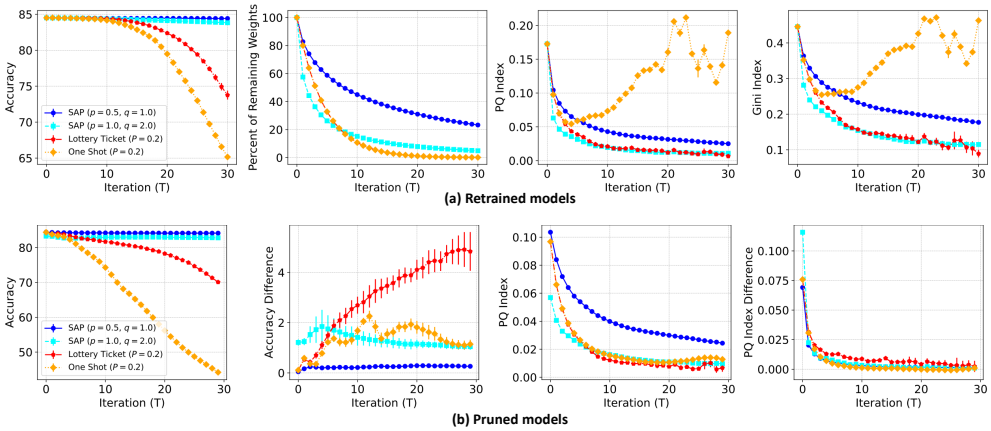

Figure 11: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with FashionMNIST and Linear.

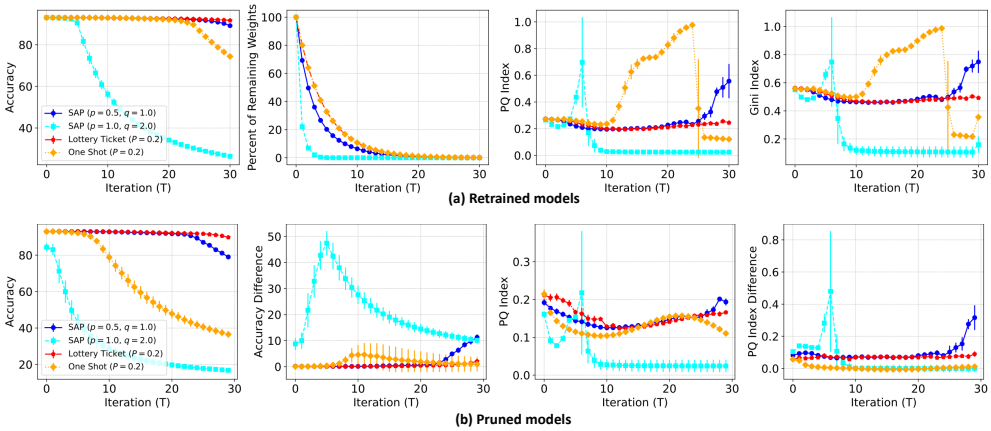

Figure 12: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with FashionMNIST and CNN.

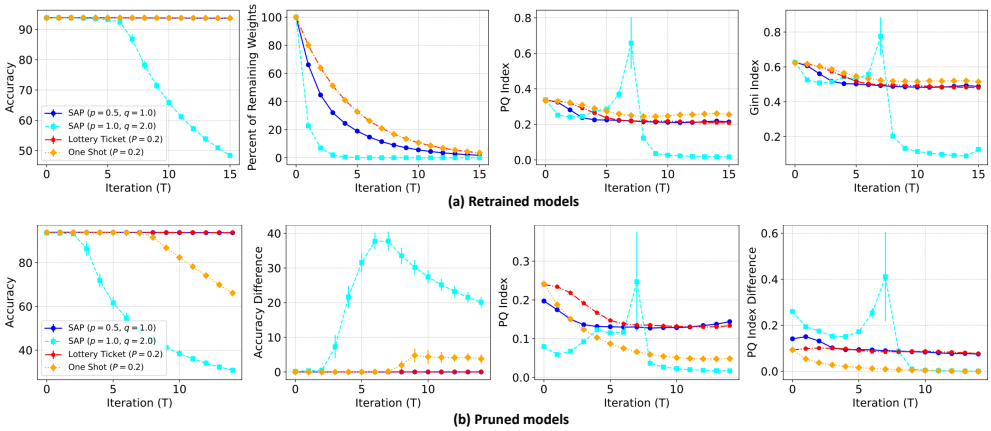

Figure 13: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with FashionMNIST and ResNet18.

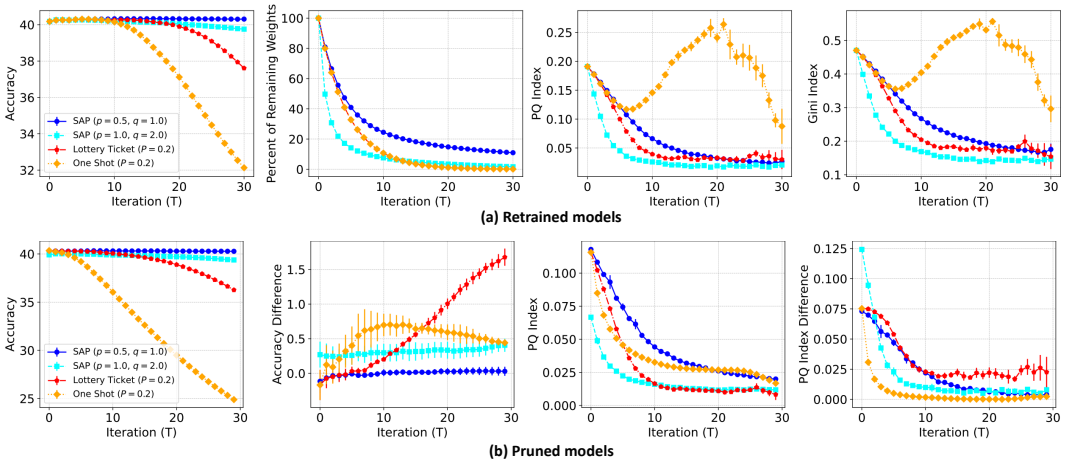

Figure 14: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with CIFAR10 and Linear.

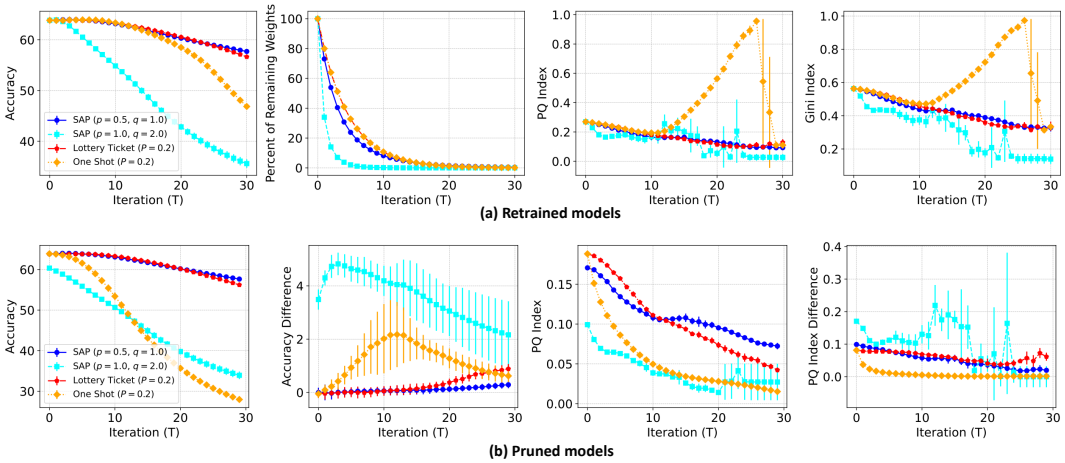

Figure 15: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with CIFAR10 and MLP.

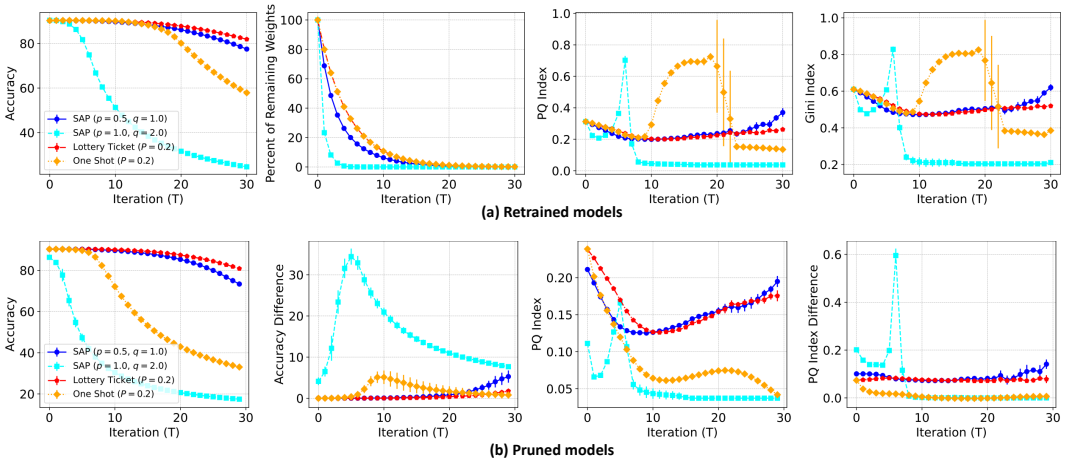

Figure 16: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with CIFAR10 and CNN.

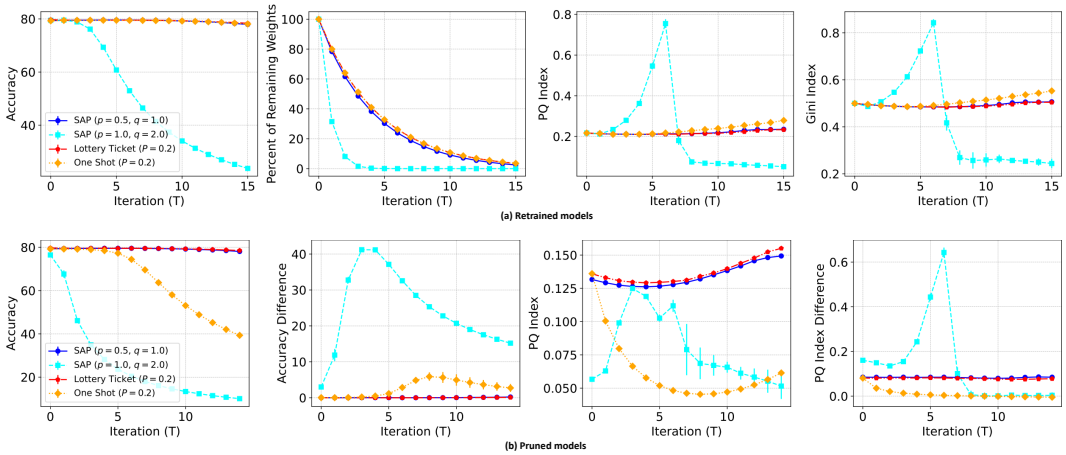

Figure 17: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with CIFAR100 and WResNet28x8.

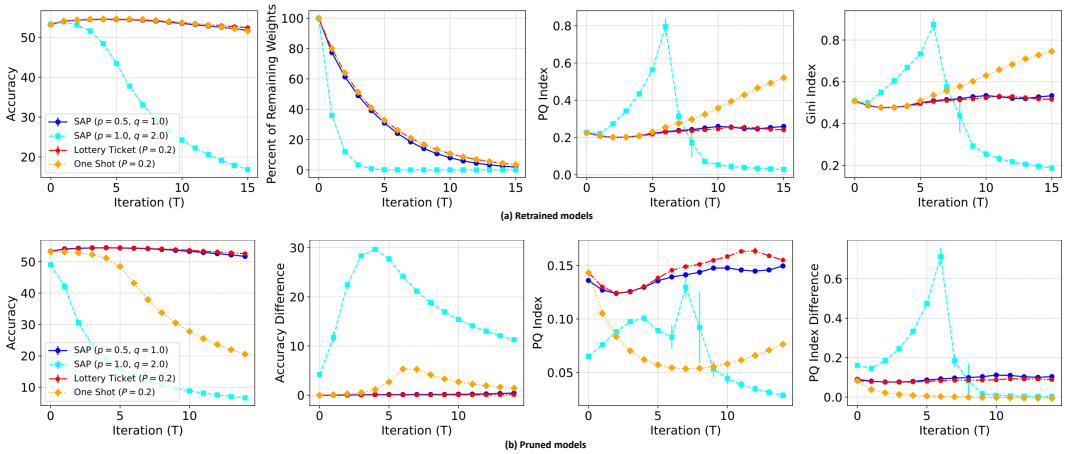

Figure 18: Results of (a) retrained and (b) pruned models at each pruning iteration for 'Global Pruning' with TinyImageNet and ResNet50.

D.3 PRUNING SCOPES

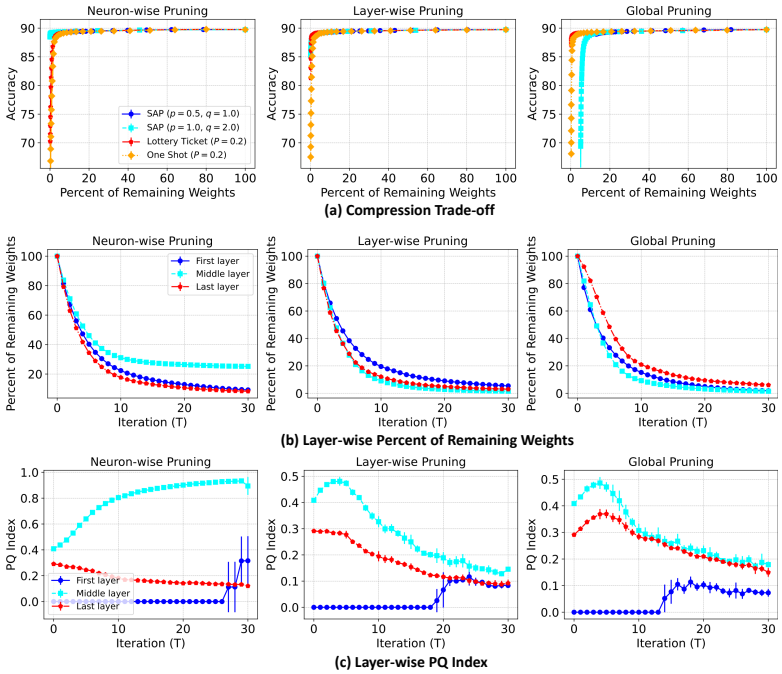

Figure 19: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for FashionMNIST and MLP. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

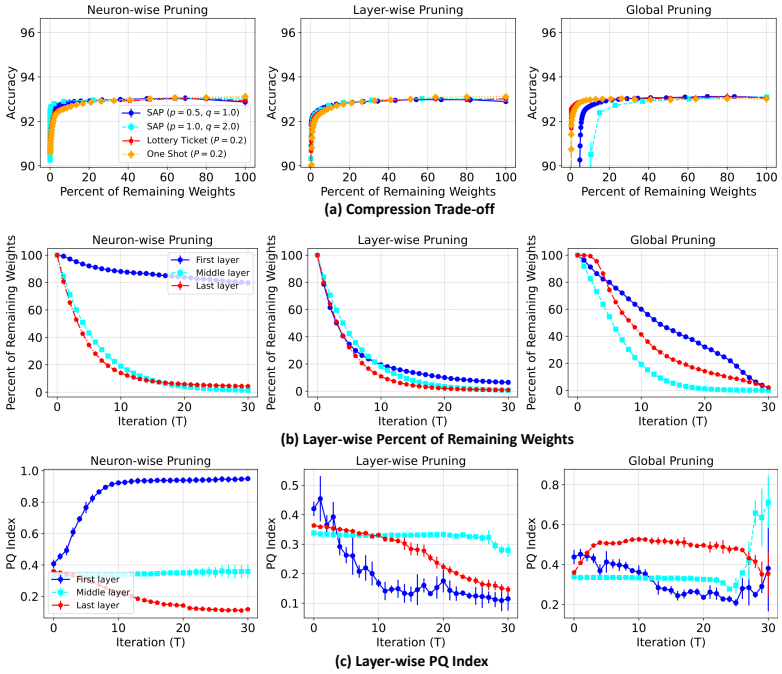

Figure 20: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for FashionMNIST and CNN. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

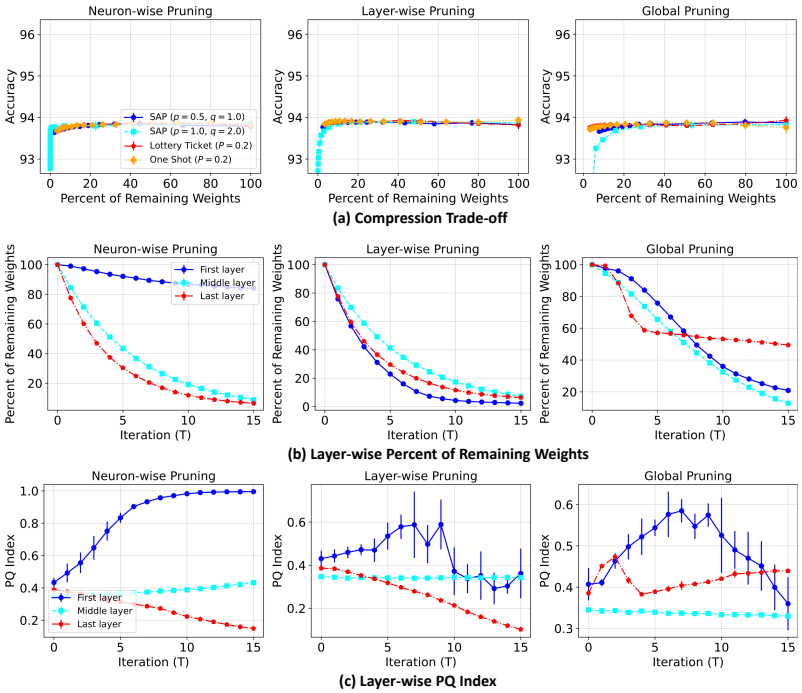

Figure 21: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for FashionMNIST and ResNet18. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

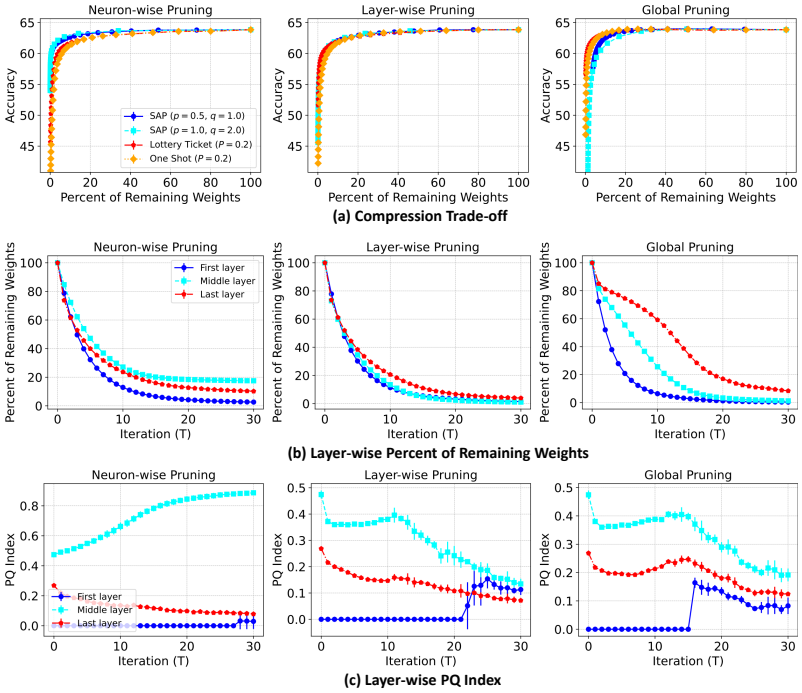

Figure 22: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for CIFAR10 and MLP. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

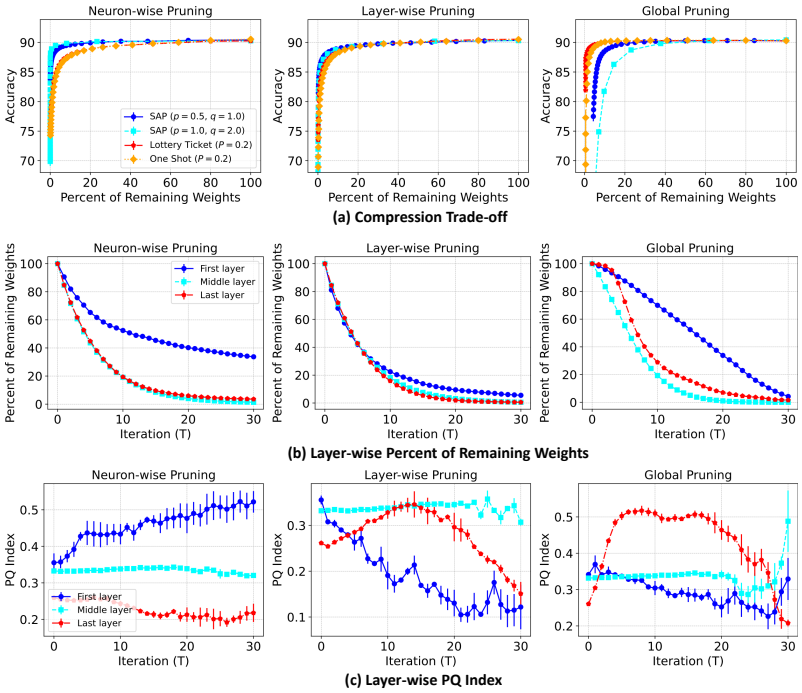

Figure 23: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for CIFAR10 and CNN. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

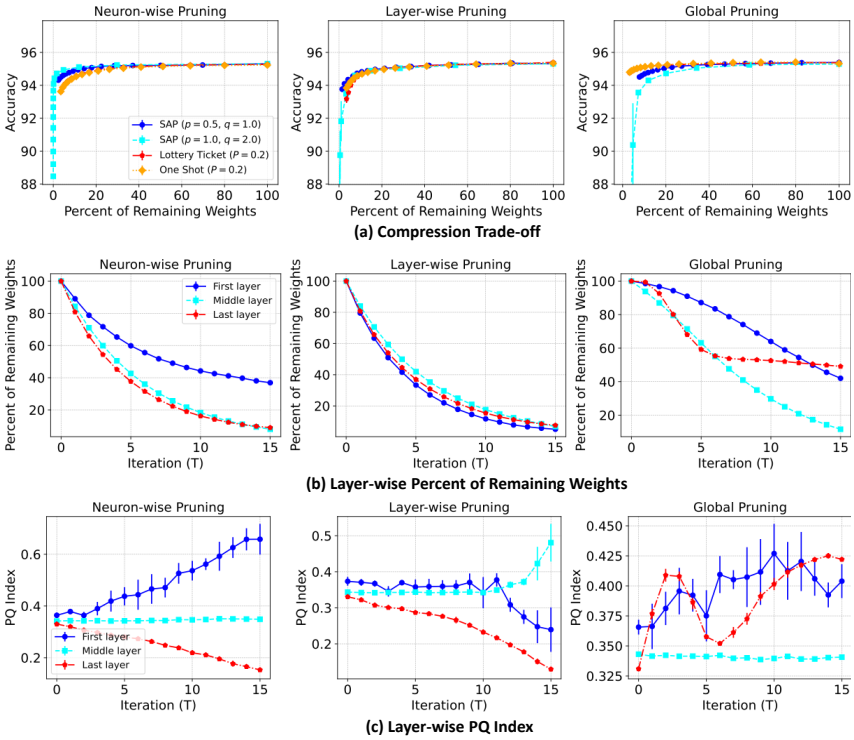

Figure 24: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for CIFAR10 and ResNet18. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

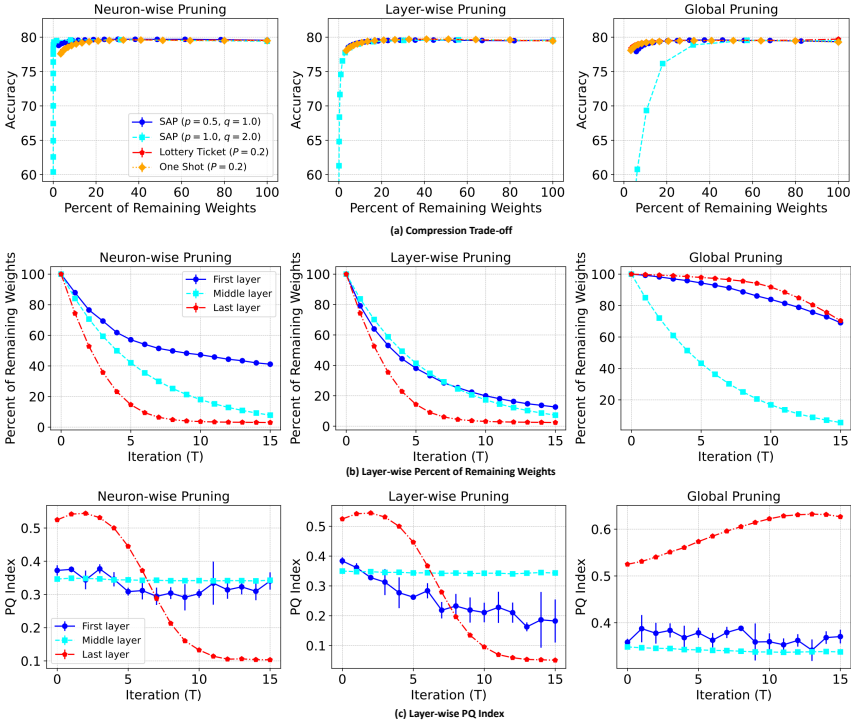

Figure 25: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for CIFAR100 and WResNet28x8. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

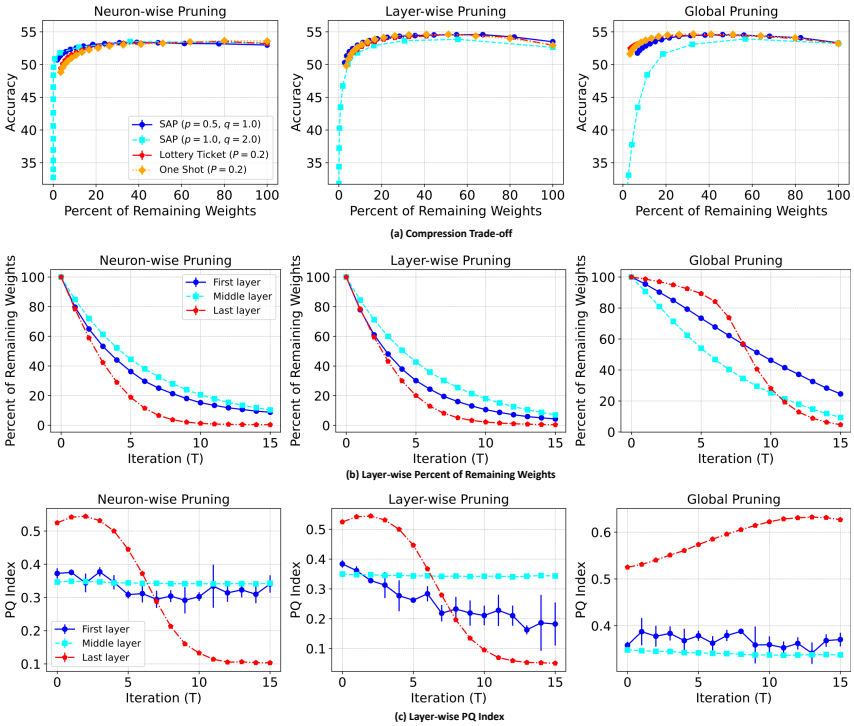

Figure 26: Results of various pruning scopes regarding (a) compression trade-off, (b) layer-wise percent of remaining weights, and (c) layer-wise PQ Index for TinyImageNet and ResNet50. (b, c) are performed with SAP ($p = 0.5$, $q = 1.0$).

D.4 EFFECTS OF p AND q

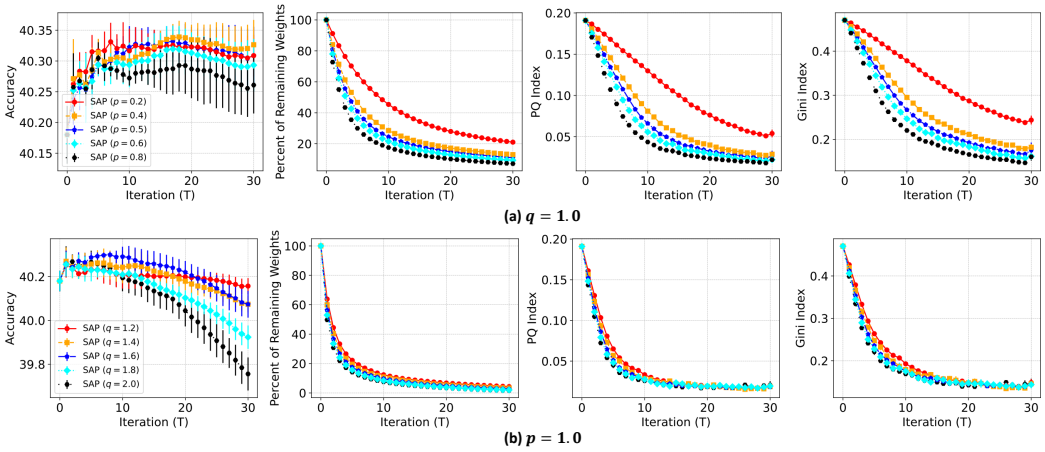

Figure 27: Ablation studies of p and q for global pruning with CIFAR10 and Linear.

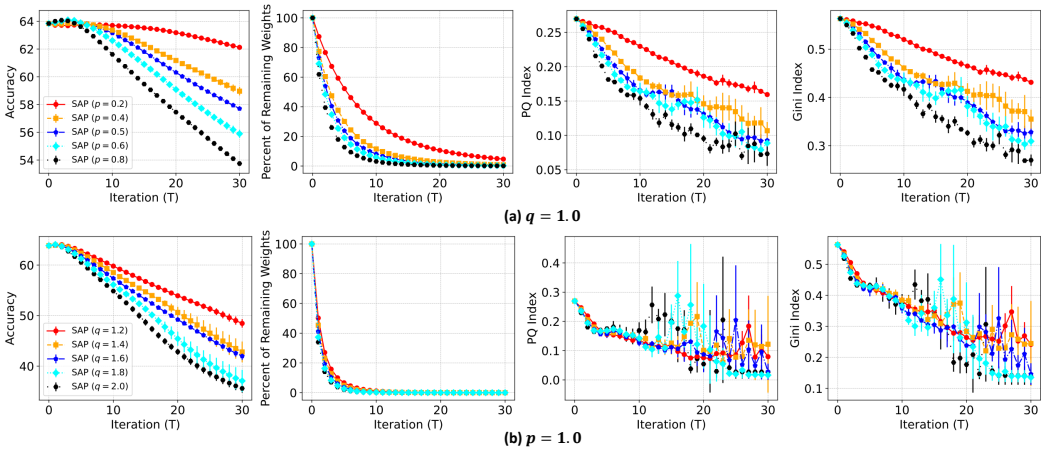

Figure 28: Ablation studies of p and q for global pruning with CIFAR10 and MLP.

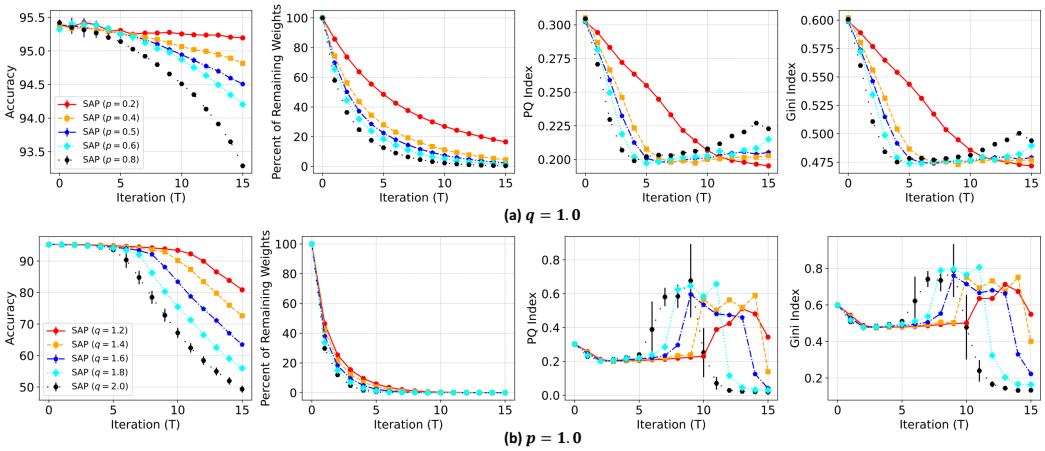

Figure 29: Ablation studies of p and q for global pruning with CIFAR10 and ResNet18.

D.5 EFFECTS OF η_r AND γ

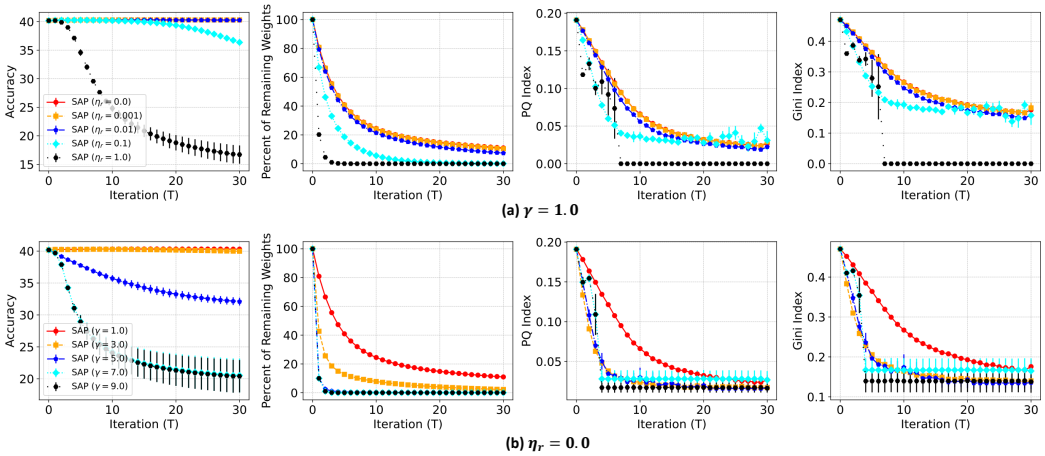

Figure 30: Ablation studies of η_r and γ for global pruning with CIFAR10 and Linear.

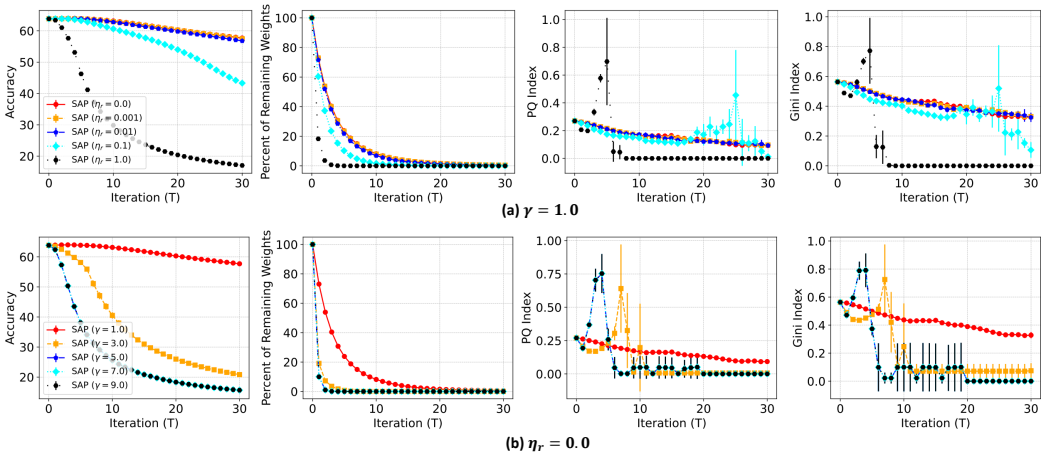

Figure 31: Ablation studies of η_r and γ for global pruning with CIFAR10 and MLP.

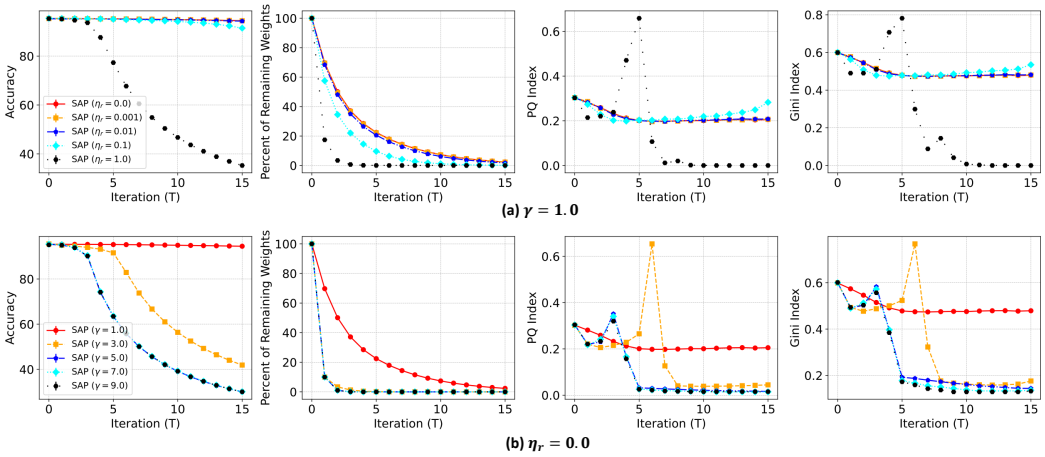

Figure 32: Ablation studies of η_r and γ for global pruning with CIFAR10 and ResNet18.