DIRECTIONAL PRUNING OF DEEP NEURAL NETWORKS

Shih-Kang Chao[1], Zhanyu Wang[2], Yue Xing[2], and Guang Cheng[2]

[1]Department of Statistics, University of Missouri-Columbia
[2]Department of Statistics, Purdue University

Mizzou University of Missouri

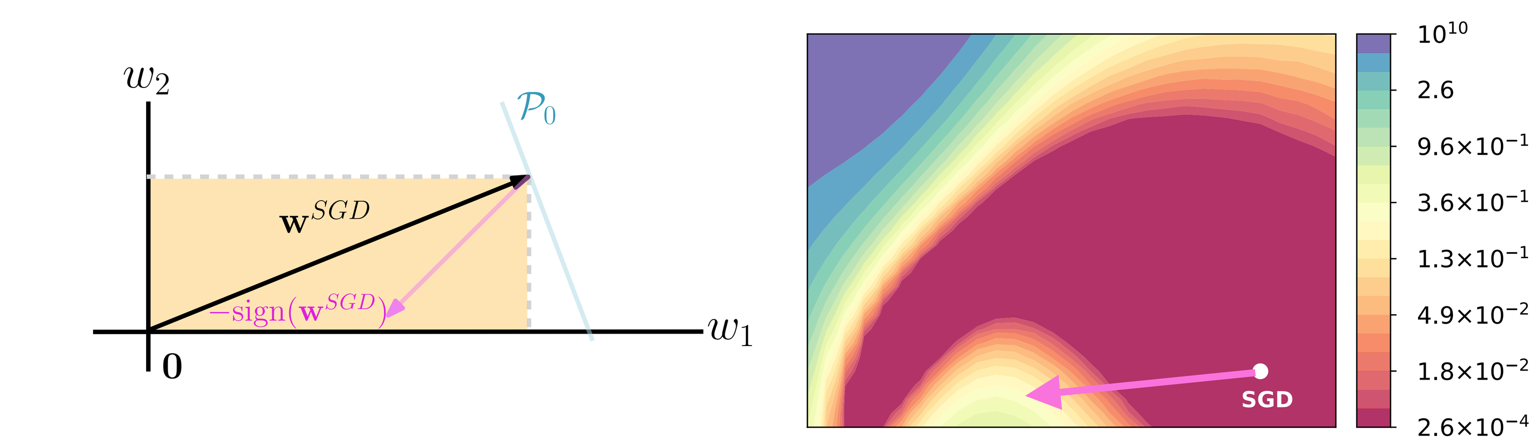
PURDUE UNIVERSITY

Introduction

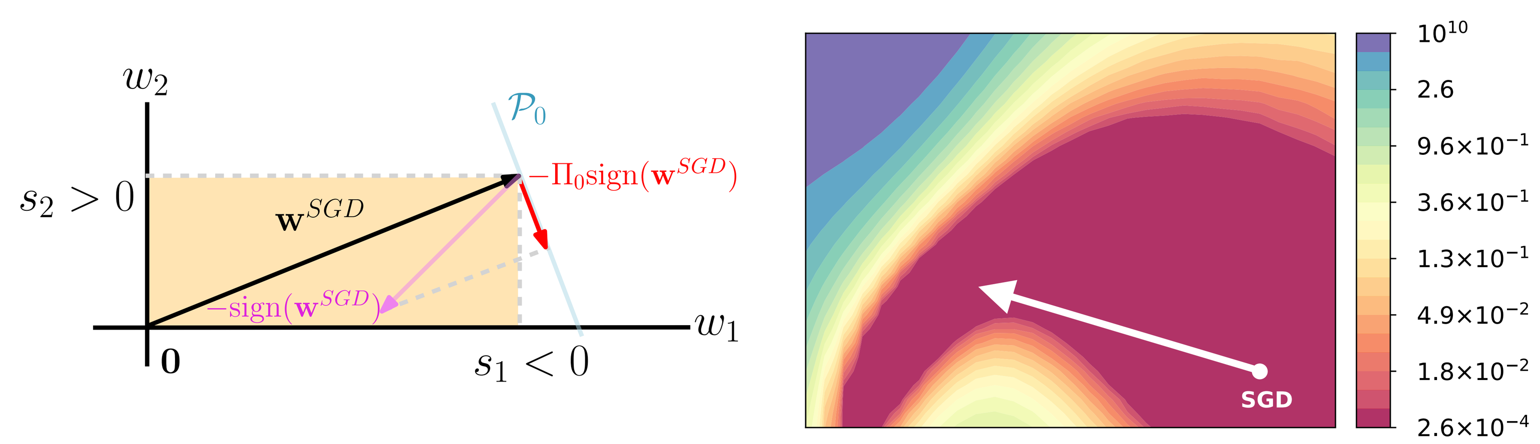
- For post-training pruning methods, re-training and fine-tuning the model bring additional computational costs.
- Most existing studies on pruning do not provide theoretical study.
- **Contributions**: We provide directional pruning (DP), a new pruning strategy that preserves training loss while maximizing the sparsity and doesn't require retraining, and a theoretically provable ℓ_1 **proximal gradient algorithm** to achieve DP.

Directional Pruning

1. SGD reaches a flat valley $\mathcal{P}_0 \subset \mathbb{R}^d$ (subspace of flat directions, or the eigenspace associated with zero eigenvalues of the Hessian). Magnitude pruning needs retraining: perturbing $\mathbf{w}^{SGD}$ to yellow area increases the training error since it is no longer in $\mathcal{P}_0$)

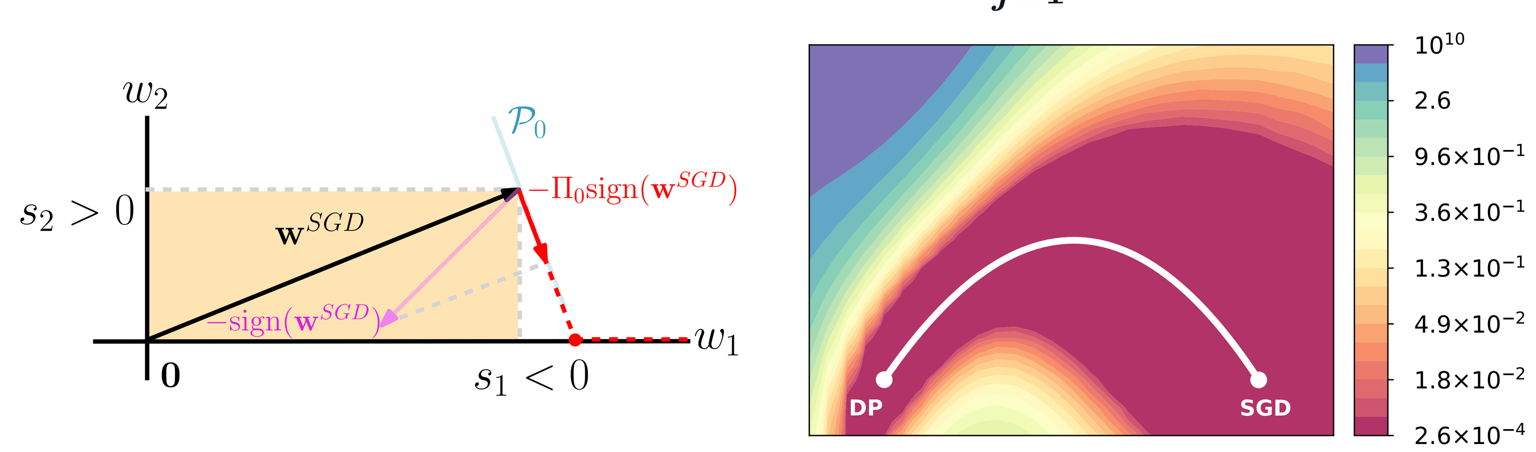


2. Pruning in a flat direction: $-\Pi_0 \text{sign}(\mathbf{w}^{SGD}) \in \mathcal{P}_0$ maximizes sparsity; score $s_j := \text{sign}(w_j^{SGD}) \cdot (\Pi_0\{\text{sign}(\mathbf{w}^{SGD})\})_j > 0$ iff. pruning w_j doesn't increase the loss



3. Directional Pruning:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{argmin}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}^{SGD}\|_2^2 + \lambda \sum_{j=1}^d s_j |w_j|$$



Algorithm and Theory

- Challenge: $\mathcal{P}_0$ is associated with the zero eigenspace of Hessian, which can't be estimated.
- Generalized regularized dual averaging (gRDA):

$$\mathbf{w}_{n+1} = \mathcal{S}_{g(n,\gamma)}\left(\mathbf{w}_0 - \gamma \sum_{k=0}^{n} \nabla f(\mathbf{w}_k; Z_{i_{k+1}})\right)$$

where $\mathcal{S}_{g(n,\gamma)}$ is soft-thresholding with $g(n,\gamma) = c\gamma^{1/2}(n\gamma)^{\mu}$, γ is learning rate, $Z_{i_{k+1}}$ is one data batch, $c, \mu > 0$ are hyperparameters.

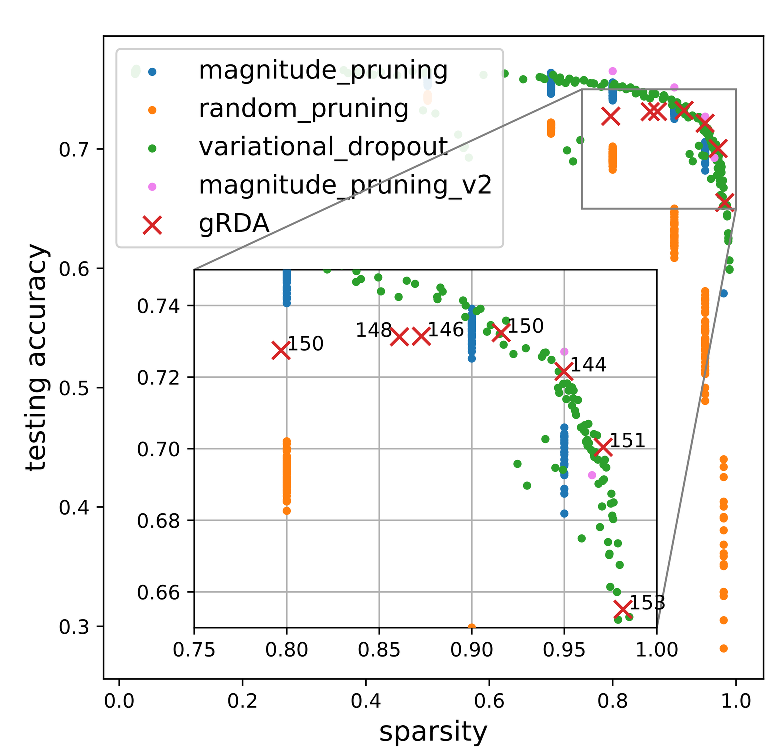
Theorem: gRDA asymptotically solves DP

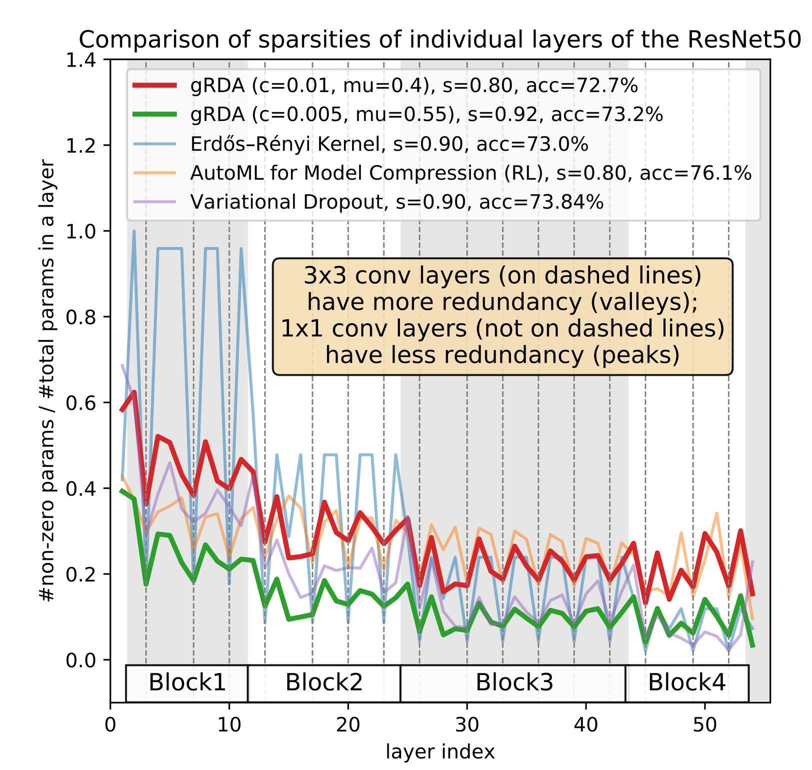
Under regularity conditions (details in paper), assume $\mu \in (0.5, 1)$ and $c > 0$ in $g(n,\gamma)$ of gRDA. Then, as $\gamma \to 0$, gRDA asymptotically performs directional pruning based on $\mathbf{w}^{SGD}(t)$; particularly,

$$\mathbf{w}_\gamma(t) \overset{d}{\approx} \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\left\{\frac{1}{2}\|\mathbf{w}^{SGD}(t) - \mathbf{w}\|_2^2 + \lambda_{\gamma,t}\sum_{j=1}^d \bar{s}_j |w_j|\right\}, \ \forall t > \bar{T},$$

where $\overset{d}{\approx}$ means "asymptotic in distribution" under the empirical probability measure of the gradients, $\lambda_{\gamma,t} = c\sqrt{\gamma}t^{\mu}$ and the $\bar{s}_j$ satisfies $\lim_{t\to\infty}|\bar{s}_j - s_j| = 0$ for all j.

ResNet50 on ImageNet





- gRDA achieves promising performance among many others for sparsity > 90%. We gratefully acknowledge Gale et al. who share the data [2].
- gRDA generates a layerwise sparsity pattern similar to other pruning algorithms [3, 1, 4].

Connectivity

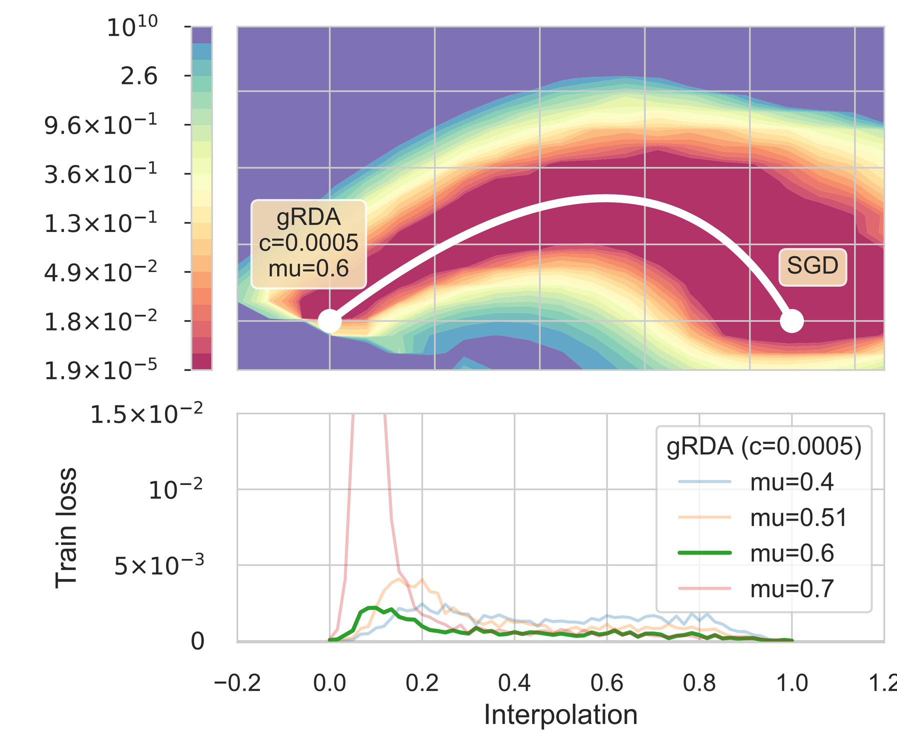


Figure 1: VGG16/CIFAR-10/Train loss

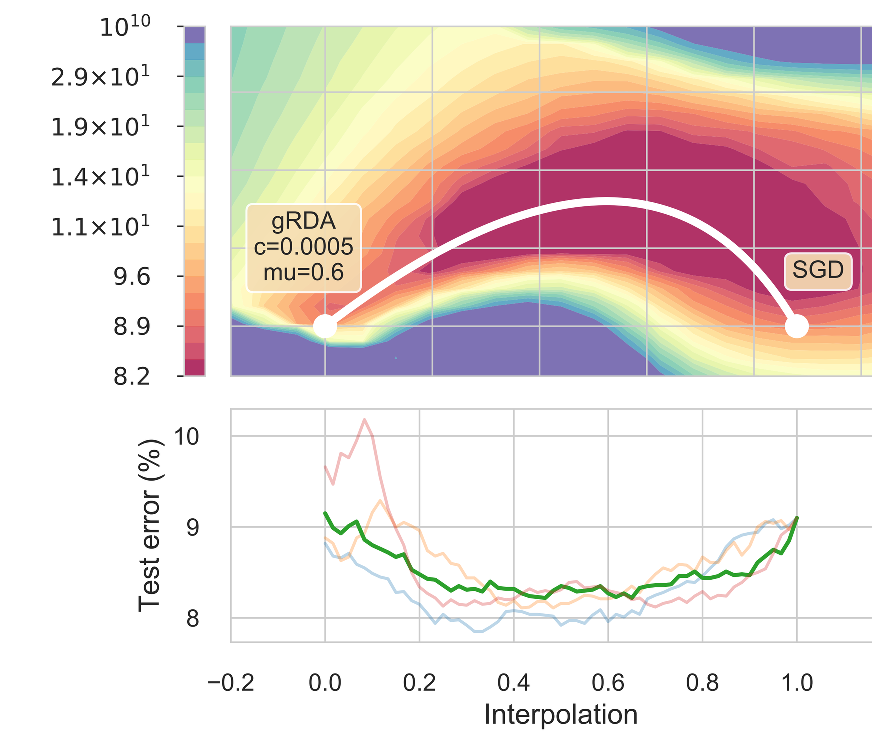


Figure 2: VGG16/CIFAR-10/Test error

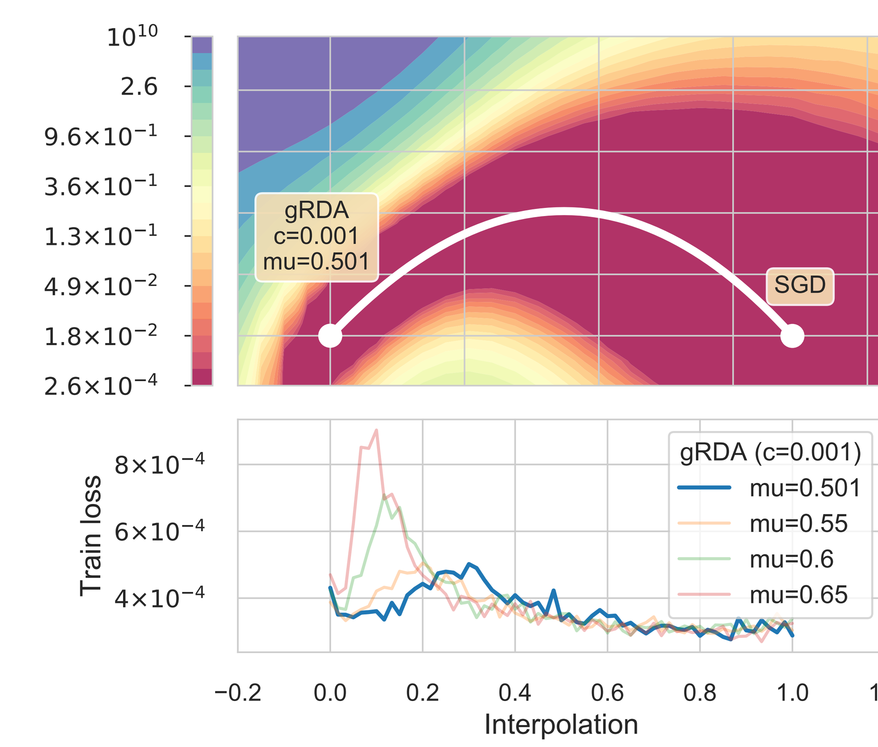


Figure 3: WRN28x10/CIFAR-100/Train loss

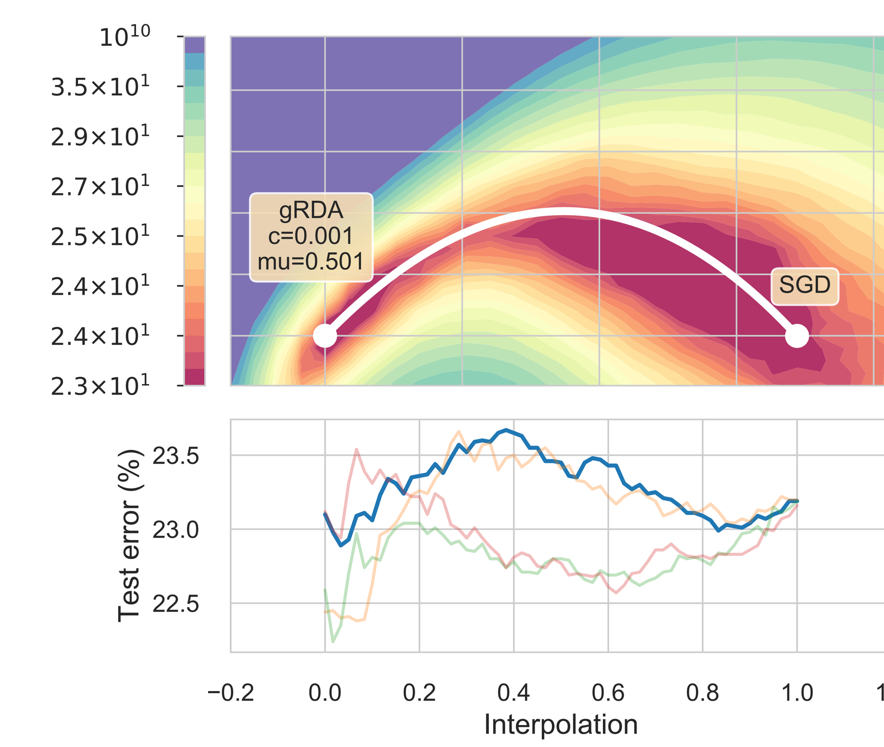


Figure 4: WRN28x10/CIFAR-100/Test error

The white curve (Bézier) traces the minimal training loss that interpolates the minimizers found by the SGD and the gRDA.

References

[1] Utku Evci et al. "Rigging the Lottery: Making All Tickets Winners". In: *arXiv preprint arXiv:1911.11134* (2019).

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

[3] Yihui He et al. "AMC: Automl for model compression and acceleration on mobile devices". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 784–800.

[4] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. "Variational dropout sparsifies deep neural networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 2498–2507.