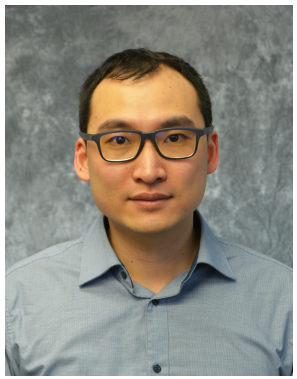
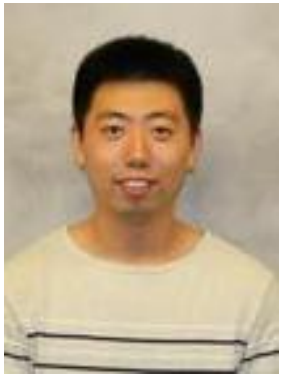
Directional Pruning of Deep Neural Networks

To appear in *NeurIPS 2020*

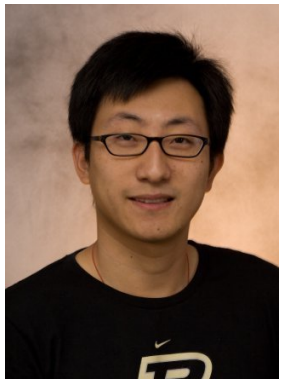


Shih-Kang Chao
University of Missouri



Zhanyu Wang
Purdue University

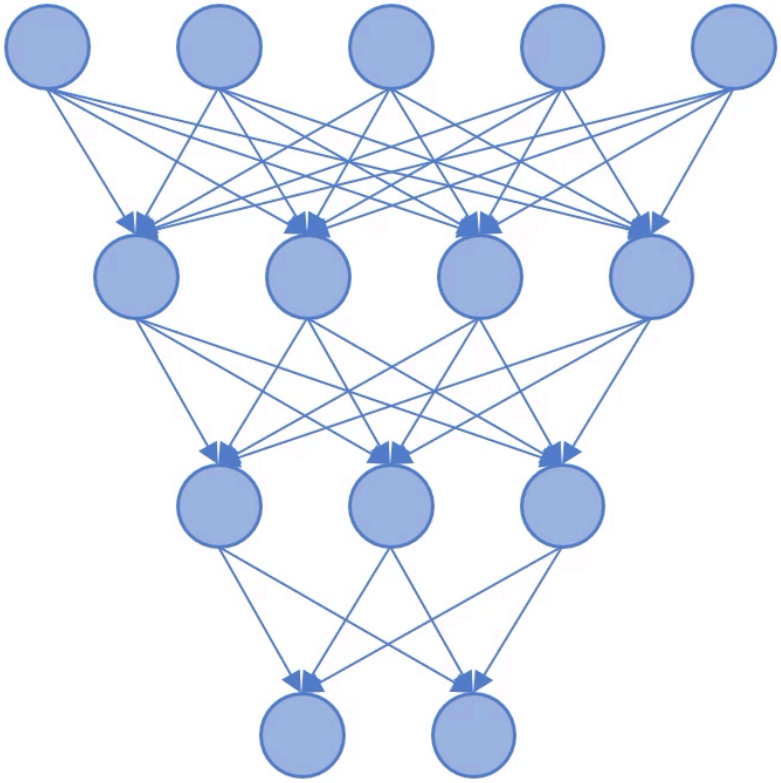
Yue Xing
Purdue University



Guang Cheng
Purdue University

GitHub: gRDA-optimizer

Pruning



- Goal: achieve **the same accuracy** as a dense networks with only **a few** nonzero weights
- Pruning is a very active research area. More than 85 papers https://github.com/he-y/Awesome-Pruning

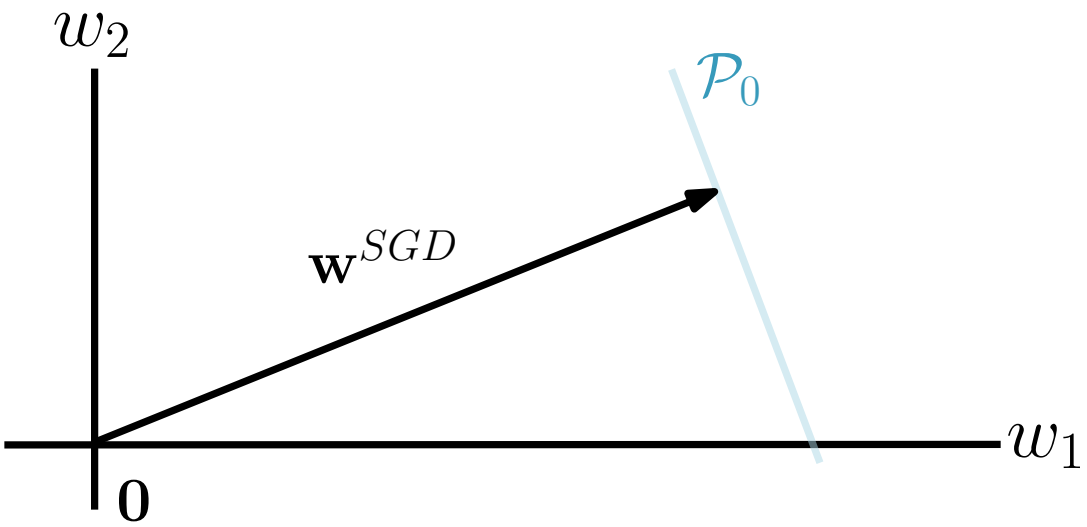
Our Contributions

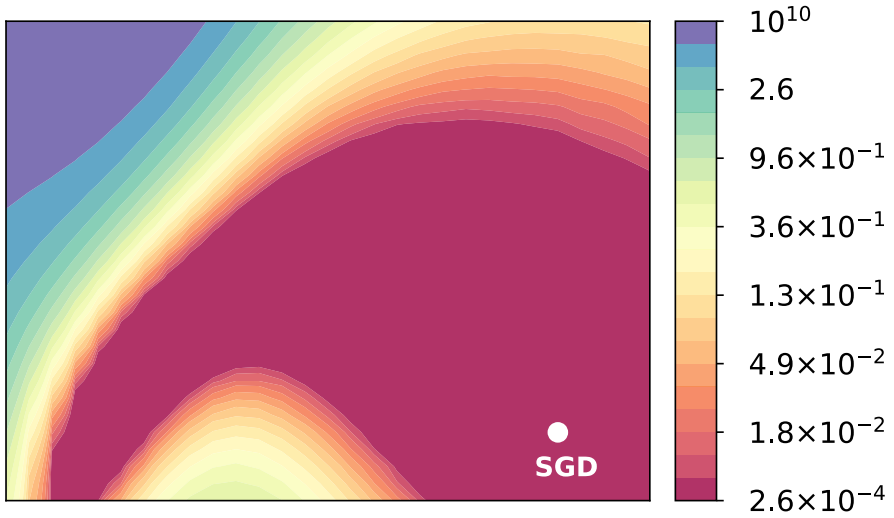
We provide **directional pruning (DP),** a new pruning strategy that

- preserves training loss while maximizing the sparsity
- doesn't require retraining

and **a theoretically provable ℓ_1 proximal gradient algorithm** to achieve DP

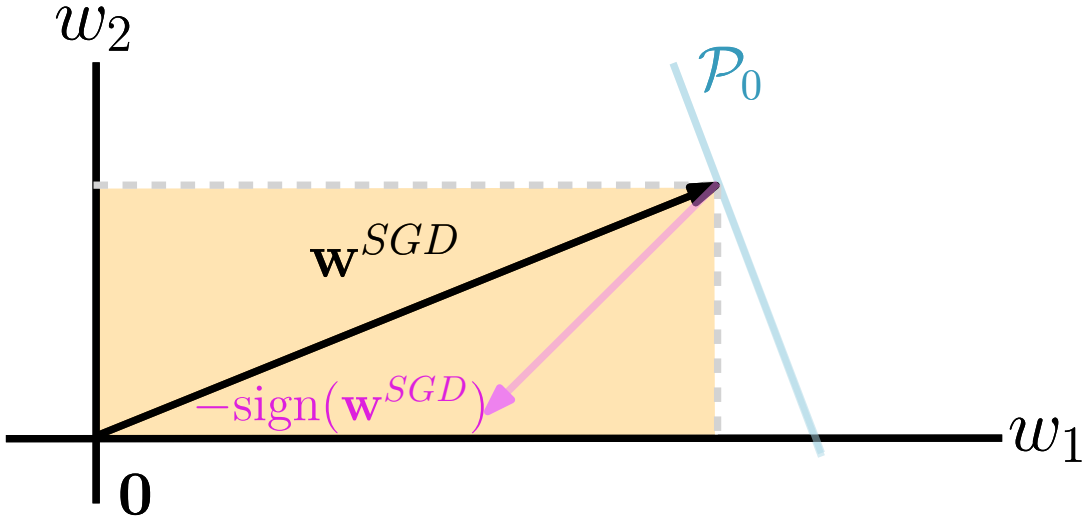
SGD reaches a flat valley

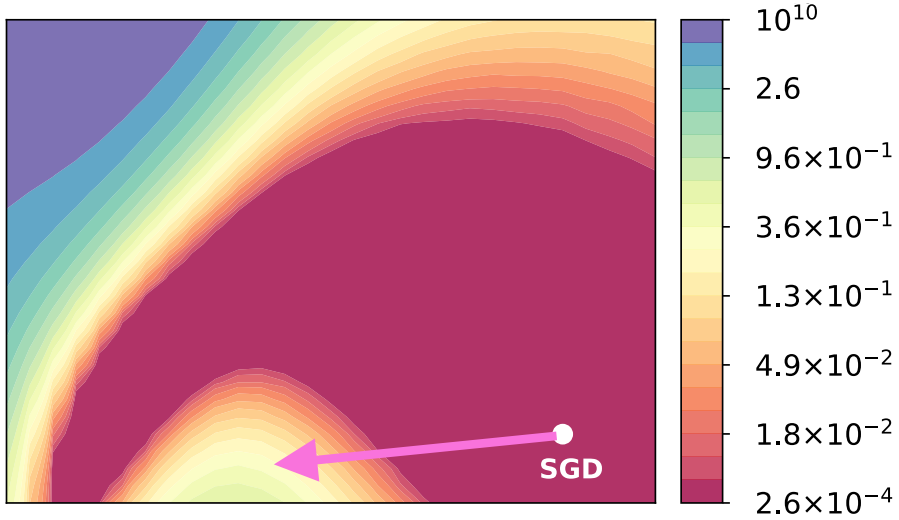




$\mathscr{P}_0 \subset \mathbb{R}^d$: subspace of flat directions, or the eigenspace associated with zero eigenvalues of the Hessian matrix

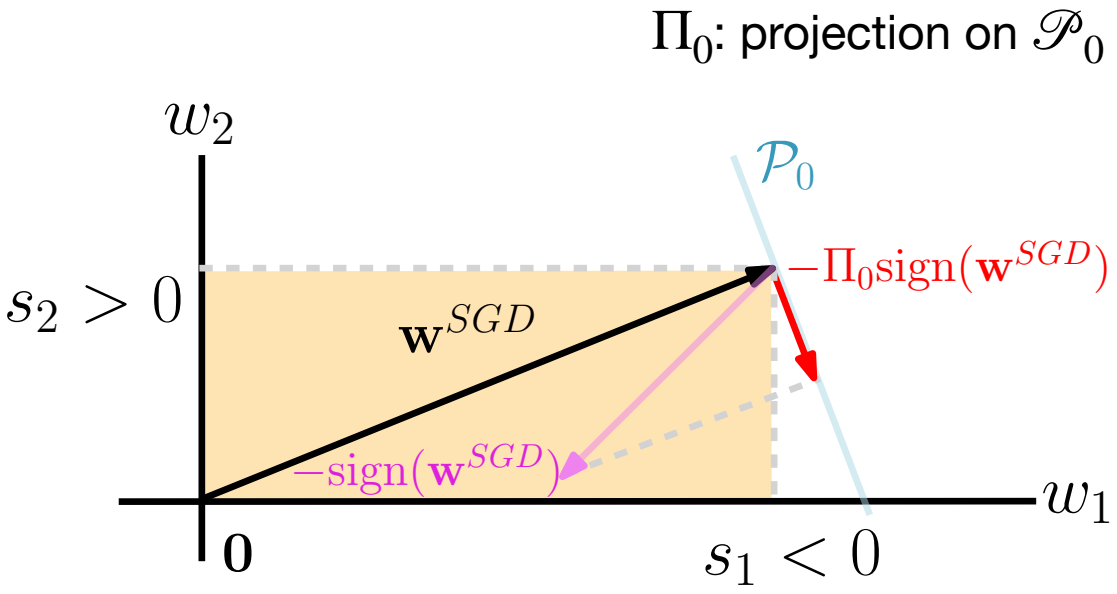
Magnitude pruning needs retraining

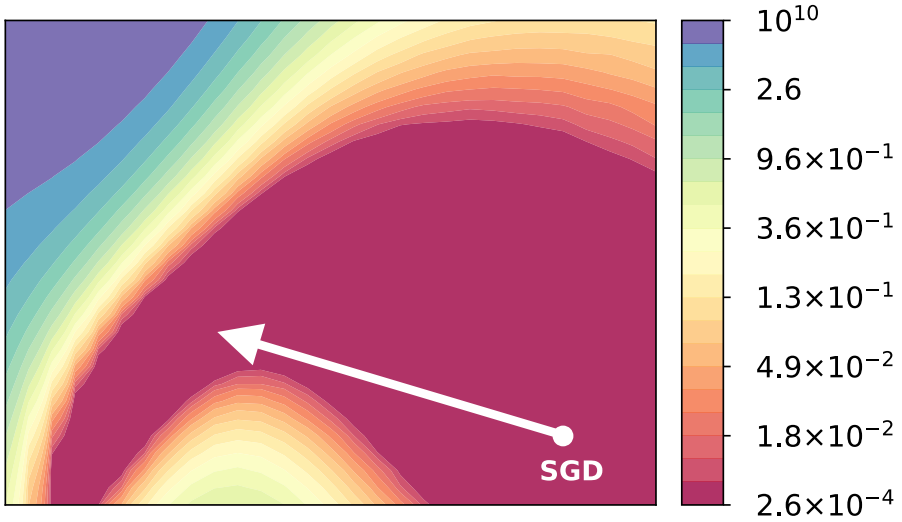




- Magnitude pruning: perturbing $\mathbf{w}^{SGD}$ to yellow area
- As the yellow area does not overlap $\mathcal{P}_0$ (documented in prior research), training error increases

Pruning in a flat direction

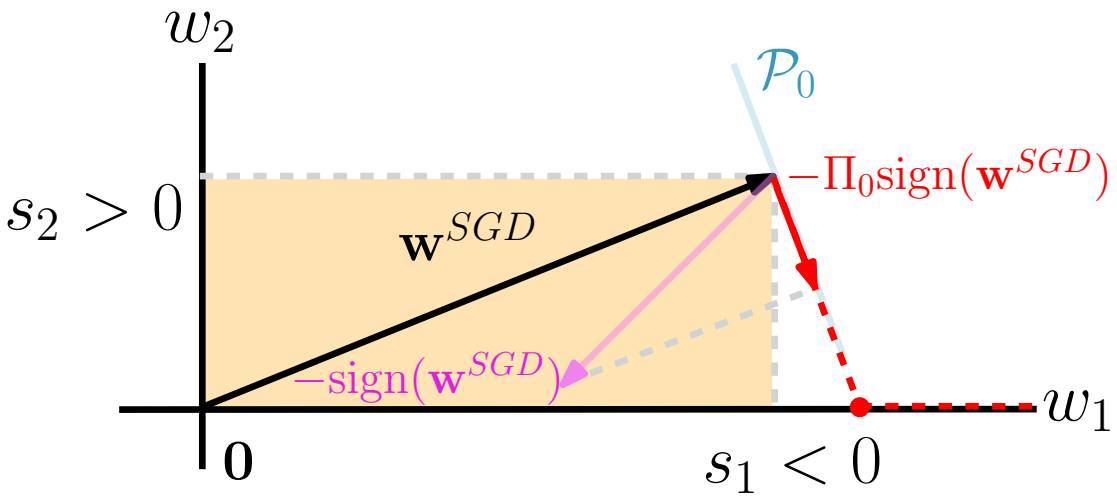


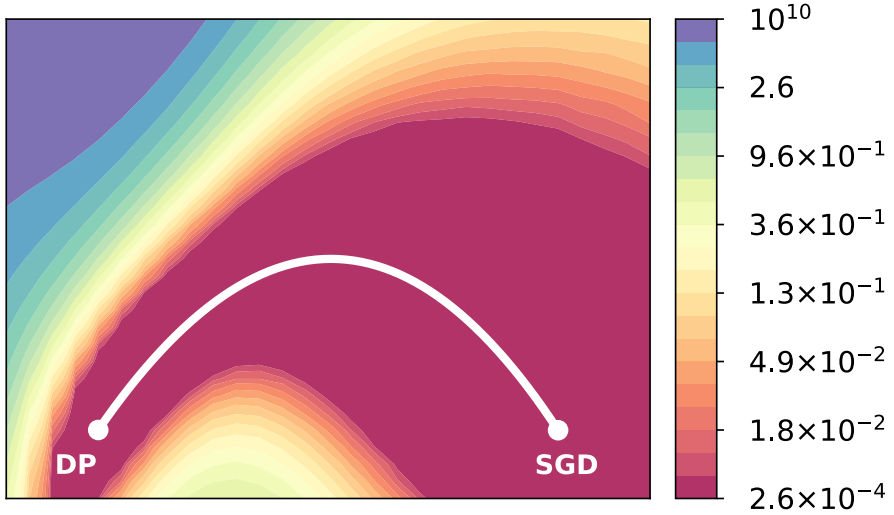


- $-\Pi_0\text{sign}(\mathbf{w}^{SGD})$: the direction in $\mathscr{P}_0$ that maximizes sparsity
- Score s_j >0 iff. pruning w_j doesn't increase the loss

  $\mathbf{s} = \text{sign}(w^{SGD}) \odot \Pi_0\text{sign}(\mathbf{w}^{SGD})$ elementwise product

Directional Pruning (DP)





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

with a large constant $\lambda > 0$

Implementation

- Challenge: $\mathscr{P}_0$ **is associated with the zero eigenspace of Hessian, which can't be estimated**

- **Theorem**: **gRDA asymptotically solves DP when** γ **small and** n **to infinity.**

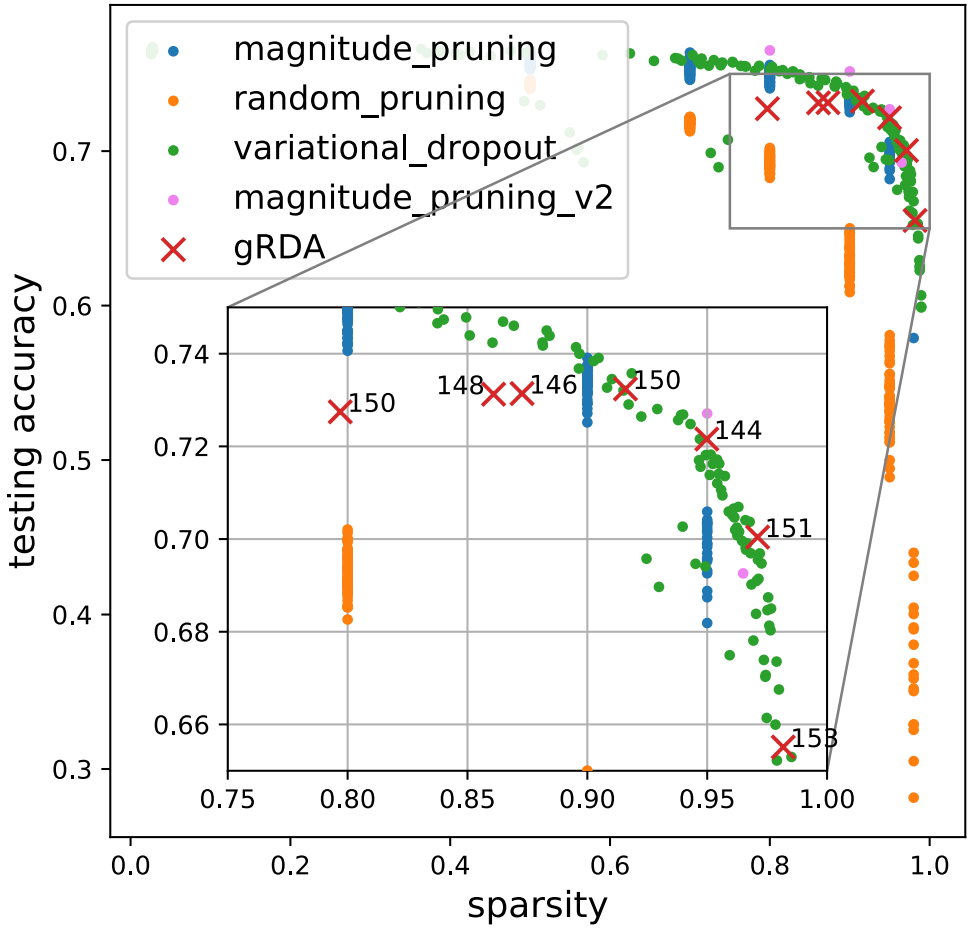
- Generalized regularized dual averaging (gRDA):

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

$\mathcal{S}_{g(n,\gamma)}$ is soft-thresholding with $g(n,\gamma) = c\gamma^{1/2}(n\gamma)^{\mu}$

γ: learning rate; c, μ>0: hyperparameters

ResNet50 on ImageNet



gRDA achieves promising performance among many others for sparsity > 90%

We gratefully acknowledge Gale, Elsen and Hooker (2019) who share the data with us

Recap

- **Propose:** directional pruning for deep neural networks
- **Implementation:** gRDA algorithm, low computational cost
- **Prove:** gRDA asymptotically performs DP
- **Demonstrate:** ResNet50 on ImageNet and more (see paper)

Thank you