# Training Structured Neural Networks Through Manifold Identification and Variance Reduction

HUANG Zih-Syuan, LEE Ching-pei

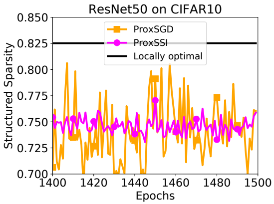
The 10th International Conference on Learning Representations

**Poster #6536**

# Structured Neural Networks

- In many scenarios, it is desirable to obtain neural networks with a certain structure
- Achieved by adding a regularizer to training/optimization objective
- Examples (regularizer in the bracket):
  - Structured or unstructured sparsity ($\ell_1$-norm or group-LASSO norm)
  - Binary/discrete neural networks (indicator function of the feasible set, or penalty for violating constraints)

# State of the Art in Deep Learning




- Only convergence guarantees to stationary points, but no guarantee for the structure of their output model
- Proximal stochastic gradient methods: ProxSGD (Yang et al., 2019, ICLR'19), ProxSSI (Deleu and Bengio, 2021), Yun et al. (2021, NeurIPS'21): artificial structure from the proximal operator
  - The output structure can be far from that of the point of convergence, due to the variance of the stochastic gradients
  - Known to output unstable and highly suboptimal structure in the convex setting (Sun et al., 2019; Poon et al., 2018)

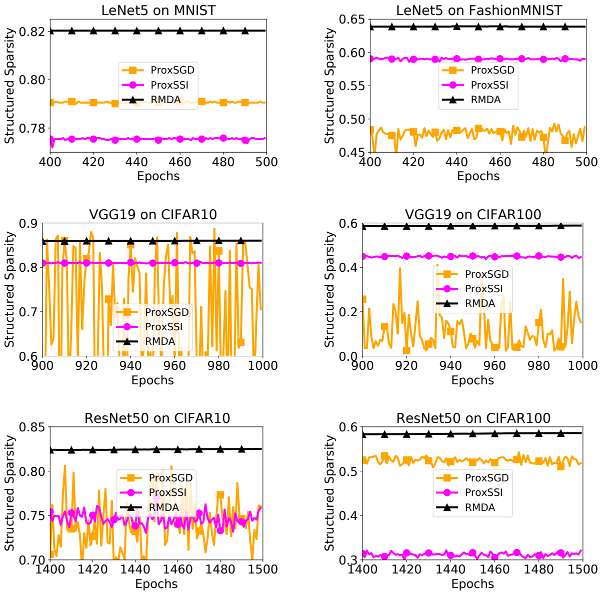
# Our Work

- We propose RMDA: Regularized Dual Averaging + Momentum
- Variance reduction beyond finite-sum (for data augmentation) with low cost
- Guaranteed optimal structure identification in finite steps, by tools from partly smooth regularizers and manifold identification
- Superior empirical performance over state of the art for
  1. Structured sparsity
  2. Pruning

# Results: Group Sparsity

- Group-LASSO norm:
  - Each channel in convolutional layers as one group
  - All outputs from one neuron in fully-connected layers as one group
- Compare with the following state of the art methods for this task
  - RMDA: Our method
  - ProxSGD (Yang et al., 2019, ICLR'19): A simple proxMSGD algorithm.
  - ProxSSI (Deleu and Bengio, 2021): This is a special case of the adaptive proximal SGD framework of Yun et al. (2021, NeurIPS'21)
  - Dense: a dense baseline by SGD with momentum

# Structured Sparsity v.s. Epochs



# Final Structured Sparsity and Validation Accuracy

| Algorithm | Validation acc. | Group sparsity | Validation acc. | Group sparsity |
|---|---|---|---|---|
| | LeNet5/MNIST | | LeNet5/FashionMNIST | |
| Dense | $99.4 \pm 0.1\%$ | - | $92.0 \pm 0.0\%$ | - |
| ProxSGD | $99.1 \pm 0.0\%$ | $76.6 \pm 2.3\%$ | $91.0 \pm 0.2\%$ | $50.5 \pm 2.7\%$ |
| ProxSSI | $99.1 \pm 0.0\%$ | $77.8 \pm 1.6\%$ | $90.9 \pm 0.0\%$ | $60.5 \pm 1.1\%$ |
| RMDA | $99.1 \pm 0.1\%$ | $79.8 \pm 1.6\%$ | $91.4 \pm 0.1\%$ | $66.2 \pm 1.7\%$ |
| | VGG19/CIFAR10 | | VGG19/CIFAR100 | |
| Dense | $94.0 \pm 0.1\%$ | - | $74.6 \pm 0.2\%$ | - |
| ProxSGD | $92.4 \pm 0.3\%$ | $72.6 \pm 6.0\%$ | $71.9 \pm 0.1\%$ | $08.6 \pm 4.9\%$ |
| ProxSSI | $92.5 \pm 0.0\%$ | $81.1 \pm 0.2\%$ | $66.2 \pm 0.4\%$ | $46.4 \pm 1.4\%$ |
| RMDA | $93.6 \pm 0.2\%$ | $86.4 \pm 0.3\%$ | $72.2 \pm 0.2\%$ | $58.9 \pm 0.4\%$ |
| | ResNet50/CIFAR10 | | ResNet50/CIFAR100 | |
| Dense | $95.7 \pm 0.0\%$ | - | $79.1 \pm 0.2\%$ | - |
| ProxSGD | $92.4 \pm 0.1\%$ | $76.8 \pm 4.1\%$ | $75.5 \pm 0.5\%$ | $51.8 \pm 0.3\%$ |
| ProxSSI | $94.1 \pm 0.1\%$ | $74.8 \pm 1.3\%$ | $74.5 \pm 0.3\%$ | $32.8 \pm 2.5\%$ |
| RMDA | $94.3 \pm 0.0\%$ | $83.0 \pm 0.5\%$ | $76.1 \pm 0.5\%$ | $57.7 \pm 3.8\%$ |

# Comparison with Pruning

- $\ell_1$ norm for pruning/unstructured sparsity
- Compare RMDA with a state-of-the-art pruning method: RigL (Evci et al., 2020, ICML'20) by Google Brain/DeepMind
- $1,000$ epochs for both RMDA and RigL

| | ResNet50 with CIFAR10 | | ResNet50 with CIFAR100 | |
|---|---|---|---|---|
| Algorithm | Sparsity | Accuracy | Sparsity | Accuracy |
| Dense baseline | | 94.81% | | 74.61% |
| RMDA | 98.36% | 93.78% | 98.32% | 74.32% |
| RigL | 98.00% | 93.41% | 98.00% | 70.88% |

Code at https://www.github.com/zihsyuan1214/rmda

Full paper at https://openreview.net/pdf?id=mdUYT5QV0O

See you at poster #6536