Regularized Adaptive Momentum Dual Averaging with an Efficient Inexact Subproblem Solver for Training Structured Neural Network

HUANG Zih-Syuan, LEE Ching-pei
r11922210@ntu.edu.tw, chingpei@ism.ac.jp

The Institute of Statistical Mathematics

Structured Neural Networks

- We aim to train neural network models with a certain structure
- Achieved by adding a regularizer to training/optimization objective
- Examples (regularizer in the bracket):
  - Structured or unstructured sparsity (ℓ_1-norm or group-LASSO norm)
  - Binary/discrete neural networks (indicator function of the feasible set, or penalty for violating constraints)
  - Low-rank structure at each layer (nuclear norm)

Our Method

- We propose RAMDA: Adaptiveness + Momentum + Regularized Dual Averaging
- Adaptiveness: Better generalization ability for various modern models including transformers
- Dual averaging: Asymptotic variance reduction with low cost
- Guaranteed to find a locally optimal structure
- Superior empirical performance over state of the art for structured sparsity with competitive running time

Inexact Subproblem Solver

- Adaptiveness + regularizer: subproblem may not have a closed-form solution
- Proposal: using the subgradient of the subproblem objective as a measurable stopping condition
- Solve the subproblem approximately using proximal gradient
- Efficiet computation and rapid convergence for the subproblems
- Retains the guarantees for structure identification and convergence of the whole algorithm

Results: Group Sparsity – Vision

Weighted group sparsity and validation accuracy on ImageNet/ResNet50.

Algorithm	Accuracy	Sparsity
RAMDA	74.53 ± 0.10%	29.19 ± 0.94%
RMDA (ICLR'22)	74.47 ± 0.08%	25.20 ± 1.69%
ProxSGD (ICLR'20)	73.50 ± 0.20%	17.54 ± 1.26%
ProxGen (NeurIPS'21)	74.17 ± 0.08%	20.29 ± 0.22%

Results: Group Sparsity – Language Modeling

Weighted group sparsity and validation perplexity on Transformer-XL with WikiText-103.

Alg.	Perplexity	Sparsity	Time/epoch
RAMDA	26.97 ± 0.10	36.2 ± 0.3%	6954 ± 30s
RMDA (ICLR'22)	27.10 ± 0.08	36.0 ± 2.7%	6184 ± 20s
ProxSGD (ICLR'20)	27.42 ± 0.02	33.1 ± 1.5%	6167 ± 12s
ProxGen (NeurIPS'21)	27.49 ± 0.19	30.5 ± 0.6%	6652 ± 21s

Results: Group Sparsity – Speech Synthesis

Weighted group sparsity and validation loss on Tacotron2 with LJSpeech.

Alg.	Loss	Sparsity	Time/epoch
RAMDA	0.44 ± 0.01	52.9 ± 1.6%	443 ± 1s
RMDA (ICLR'22)	0.46 ± 0.01	45.9 ± 1.7%	431 ± 2s
ProxSGD (ICLR'20)	0.50 ± 0.00	34.3 ± 1.6%	431 ± 0s
ProxGen (NeurIPS'21)	0.45 ± 0.01	45.6 ± 0.9%	438 ± 2s