A Single-Step, Sharpness-Aware Minimization is All You Need to Achieve Efficient and Accurate Sparse Training

Jie Ji, Gen Li, Jingjing Fu, Fatemeh Afghah, Linke Guo, Xiaoyong Yuan, Xiaolong Ma

Clemson University

jji@g.Clemson.edu

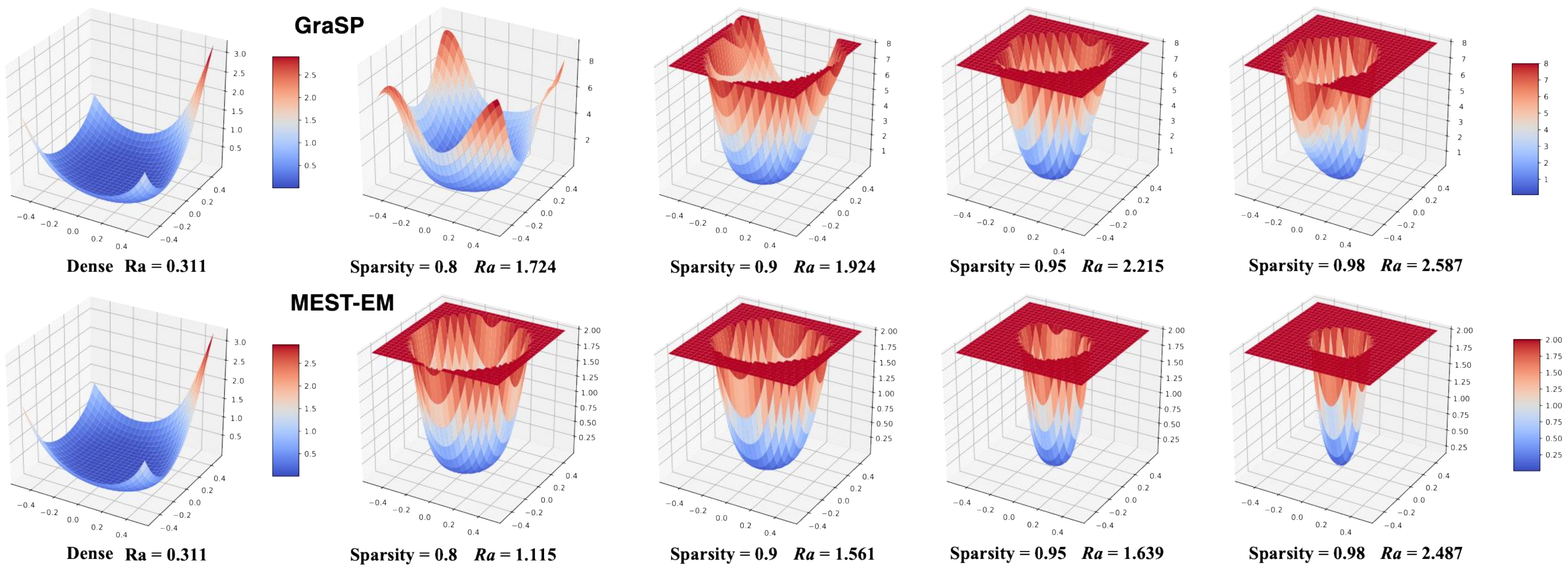
NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- **Context**: Growing size and complexity of DNNs demand efficient training methods.
- **Challenge**: Sparse training helps but struggles with generalization due to chaotic loss surfaces.
- **Objective**: Introduce S2-SAM to enhance sparse training with no extra computational cost.

Motivation & Problem Statement

- **Insight**: Sparse networks suffer from steep, chaotic loss surfaces as shown in visualization (Fig. 1).



Motivation & Problem Statement

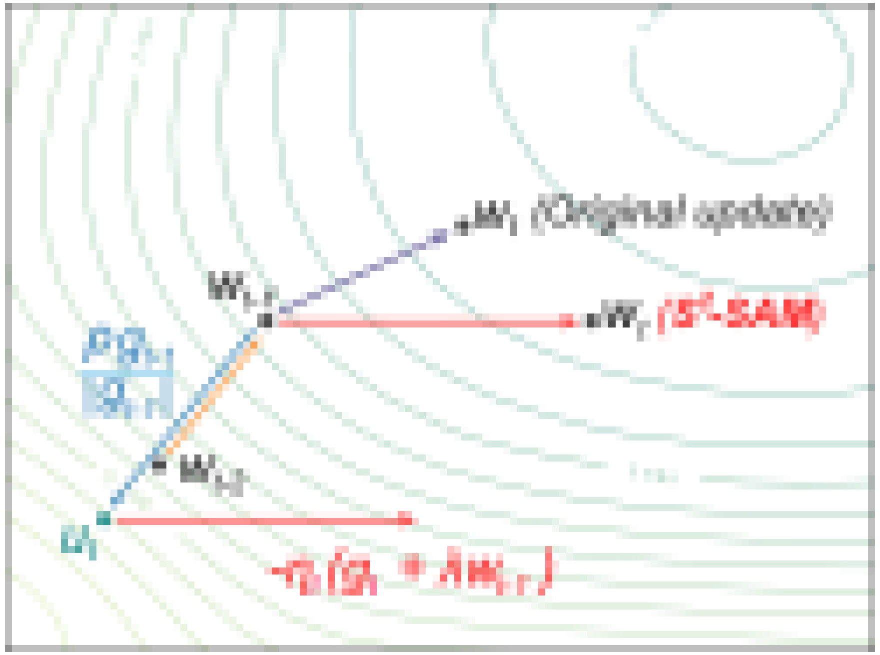
- **Insight**: Sparse networks suffer from steep, chaotic loss surfaces as shown in visualization (Fig. 1).
- **Problem**: Current sparse training methods often fail to achieve optimal generalization.
- **Key Question**: Can we improve generalization without sacrificing sparsity or efficiency?

Proposed Method - S2-SAM

- **Concept**: Single-step Sharpness-Aware Minimization.
- **Innovation**: Leverages prior gradient information to approximate perturbation in a single step.
- **Benefit**: Zero additional computational cost compared to traditional SAM.

Technical Overview

- **SAM Recap**: Two-step approach to find flatter minima.
- **S2-SAM Approach**: Uses prior gradients to compute perturbation (Equation reference from the paper).
- **Diagram**: Visualize the gradient flow (Fig. 2).



Theoretical Analysis

- **Convergence Proof**: Overview of theoretical backing and conditions.
- **Assumptions**: Unbiased gradient, smooth loss function, and bounded variance.
- **Conclusion**: S2-SAM guarantees convergence with minimal modifications.

Theorem 1. *Under Assumption 1, assume that $\ell(\mathbf{y}, f(\mathbf{w}, \mathbf{x}))$ is L-smooth and B-Lipschitz, suppose $\widehat{F}_{\mathcal{S}}(\mathbf{w})$ satisfies Assumption 2 and $\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}_{\mathcal{S}}(\mathbf{w}) \leq F_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}^*) + \frac{\lambda}{2}\|\mathbf{w}_t\|^2$ with $\lambda = 2L$, where $\mathbf{w}_t$ is the intermediate solution of $\mathcal{A}$, then*

$$\mathrm{E}_{R,\mathcal{A},\mathcal{S}}\left[F\left(\mathbf{w}_R\right)\right] - \mathrm{E}_{\mathcal{S}}\left[F\left(\mathbf{w}_*\right)\right] \leq \frac{\widehat{F}_{\mathcal{S}}\left(\mathbf{w}_0\right)}{\mu\eta T} + \frac{\eta(L+\lambda)\sigma^2}{2\mu} + \frac{6B+1}{n}$$

where $\mathcal{A}$ is SGD.

The $\mathbf{w}_{\mathcal{S}}^* \in \min_{\mathbf{w}\in\mathcal{W}} F_{\mathcal{S}}(\mathbf{w})$ is an optimal solution, and we show that the generalization error is bounded. Based on Lemma 1, the proof of Theorem 1 is derived in Appendix C.

Experimental Results Overview

- **Datasets & Models**: CIFAR-10, CIFAR-100 with ResNet-32 and VGG-19; ImageNet-1K with ResNet-50.
- **Key Metrics**: Accuracy improvement and training throughput.
- **Summary**: Consistent improvement across various sparsity levels and methods.

Detailed Experimental Results

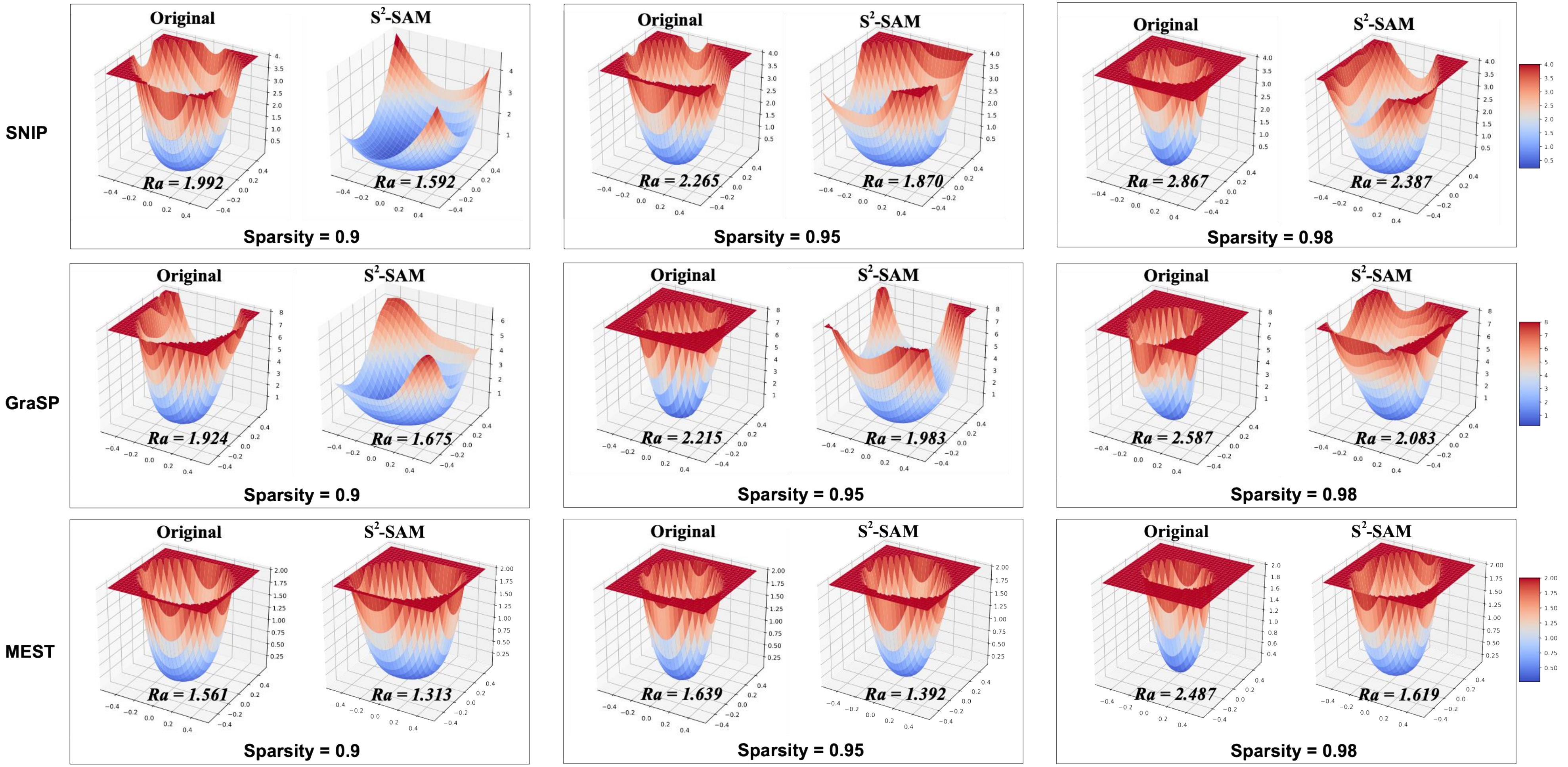
- **Table 1**: Show test accuracy improvements with S2-SAM on CIFAR-10/100.
- **Visualization**: Loss surface comparison (Fig. 3) illustrating smoother loss with S2-SAM.
- **Key Takeaway**: Higher sparsity benefits more from S2-SAM.

Detailed Experimental Results

Table 1: Test accuracy (%) of pruned ResNet-32 on CIFAR-10/100.

Datasets	CIFAR-10			CIFAR-100		
Pruning ratio	90%	95%	98%	90%	95%	98%
ResNet-32	94.58 (Dense)			74.89 (Dense)		
LT [20]	92.31	91.06	88.78	68.99	65.02	57.37
LT+ S^2-SAM (ours)	**92.58±0.07** (0.27↑)	**91.47±0.10** (0.41↑)	**89.35±0.11** (0.57↑)	**69.34±0.09** (0.35↑)	**65.45±0.11** (0.43↑)	**57.76±0.13** (0.39↑)
SNIP [21]	92.59±0.10	91.01±0.21	87.51±0.31	68.89±0.45	65.02±0.69	57.37±1.43
SNIP+ S^2-SAM (ours)	**93.17±0.16** (0.58↑)	**91.59±0.22** (0.58↑)	**88.08±0.29** (0.57↑)	**69.33±0.28** (0.44↑)	**65.66±0.49** (0.64↑)	**58.25±0.77** (0.88↑)
GraSP [32]	92.38±0.21	91.39±0.25	88.81±0.14	69.24±0.24	66.50±0.11	58.43±0.43
GraSP+ S^2-SAM (ours)	**92.87±0.14** (0.49↑)	**91.98±0.22** (0.59↑)	**89.66±0.29** (0.85↑)	**69.98±0.22** (0.74↑)	**67.12±0.18** (0.62↑)	**59.45±0.19** (1.02↑)
SET [2]	92.30	90.76	88.29	69.66	67.41	62.25
SET+ S^2-SAM (ours)	**92.92±0.23** (0.62↑)	**91.50±0.19** (0.74↑)	**88.78±0.20** (0.49↑)	**70.23±0.20** (0.57↑)	**68.28±0.15** (0.87↑)	**63.56±0.19** (1.31↑)
DSR [33]	92.97	91.61	88.46	69.63	68.20	61.24
DSR+ S^2-SAM (ours)	**93.49±0.21** (0.52↑)	**92.08±0.22** (0.47↑)	**89.11±0.17** (0.65↑)	**70.11±0.16** (0.48↑)	**68.87±0.16** (0.67↑)	**62.00±0.17** (0.76↑)
RigL [23]	93.07	91.83	89.00	70.34	68.22	64.07
RigL+ S^2-SAM (ours)	**93.55±0.14** (0.48↑)	**92.11±0.21** (0.28↑)	**90.40±0.17** (1.40↑)	**72.38±0.11** (2.04↑)	**70.29±0.14** (2.07↑)	**64.98±0.06** (0.91↑)
RigL (ERK) [23]	93.55	92.39	90.22	70.62	68.47	64.14
RigL (ERK)+ S^2-SAM (ours)	**93.75±0.19** (0.20↑)	**92.81±0.08** (0.42↑)	**91.16±0.11** (0.94↑)	**72.56±0.07** (1.94↑)	**70.33±0.10** (1.86↑)	**65.15±0.12** (1.01↑)
MEST (EM) [3]	92.56±0.07	91.15±0.29	89.22±0.11	70.44±0.26	68.43±0.32	64.59±0.27
MEST (EM) + S^2-SAM (ours)	**93.43±0.12** (0.87↑)	**91.58±0.07** (0.43↑)	**91.22±0.14** (2.00↑)	**71.95±0.13** (1.51x↑)	**70.04±0.10** (1.61↑)	**65.69±0.34** (1.10↑)
MEST (EM&S) [3]	93.27±0.14	92.44±0.13	90.51±0.11	71.30±0.31	70.36±0.05	67.16±0.25
MEST (EM&S) + S^2-SAM (ours)	**93.39±0.17** (0.12↑)	**92.97±0.17** (0.53↑)	**91.32±0.18** (0.81↑)	**72.74±0.08** (1.44↑)	**71.85±0.09** (1.49↑)	**69.13±0.20** (1.97↑)

Detailed Experimental Results



Training Efficiency

- **Comparison**: S2-SAM vs. SAM in terms of training speed (Table 4).
- **Observation**: S2-SAM maintains throughput close to original training methods.

Table 4: Training speed of SAM [25] and S^2-SAM for different sparse training at 90% sparsity.

Methods	Training	Accuracy (%)	Throughput (↑)
GraSP	Original	68.10	**2148** imgs/s
	SAM	**68.95**	1021 imgs/s
	S^2-SAM	68.78	2132 imgs/s
RigL	Original	72.00	**3133** imgs/s
	SAM	**72.75**	1508 imgs/s
	S^2-SAM	72.44	3098 imgs/s
MEST (EM)	Original	73.60	**2981** imgs/s
	SAM	**74.88**	1398 imgs/s
	S^2-SAM	74.58	2977 imgs/s

Robustness to Perturbations

- **ImageNet-C Results**: Show improvement in model robustness (Table 5).
- **Implication**: Wider loss basin correlates with better handling of data corruption.

Table 5: Testing accuracy on ImageNet-C test set. We compare the results with and without S^2-SAM using 80% sparsity.

Methods	ImageNet-1K Accuracy (%)	ImageNet-C Accuracy (%)
SNIP	69.70	31.12
SNIP + S^2-SAM	**70.55** (0.85↑)	**34.87** (3.75↑)
GraSP	72.10	32.24
GraSP + S^2-SAM	**72.66** (0.56↑)	**35.17** (2.93↑)
MEST (EM)	75.70	33.87
MEST (EM) + S^2-SAM	**76.35** (0.65↑)	**36.98** (3.11↑)
RigL	74.60	33.68
RigL + S^2-SAM	**75.39** (0.79↑)	**36.80** (3.12↑)

Application to Dense Models

- **Results**: Applying S2-SAM on dense networks (Table 6).
- **Finding**: Effective even for models with lower parameter counts.

Table 6: Testing accuracy on dense model training. We compare original training with S^2-SAM in same settings.

Networks	Params. Count	Original Accuracy (%)	S^2-SAM Accuracy (%)
CIFAR-10			
ResNet-32	1.86M	94.58	**94.99** (0.41↑)
MobileNet-V2	2.30M	94.13	**94.55** (0.42↑)
VGG-19	20.03M	94.21	**94.48** (0.27↑)
ImageNet-1K			
EfficientNet-B0	5.30M	76.54	**77.10** (0.56↑)
ResNet-34	21.80M	74.09	**74.58** (0.49↑)
ResNet-50	25.50M	76.90	**77.32** (0.42↑)

Conclusion & Contributions

- **Contributions**:
- Identification of chaotic loss surfaces as a challenge in sparse training.
- Development of S2-SAM, a zero-cost, plug-and-play sharpness-aware minimization.
- Theoretical and experimental validation of S2-SAM's effectiveness.
- **Future Work**: Potential applications to dense training.

THANK YOU!

- See you at Vancouver Convention Center!
- Our poster session: Thu 12 Dec 4:30 p.m. PST — 7:30 p.m. PST