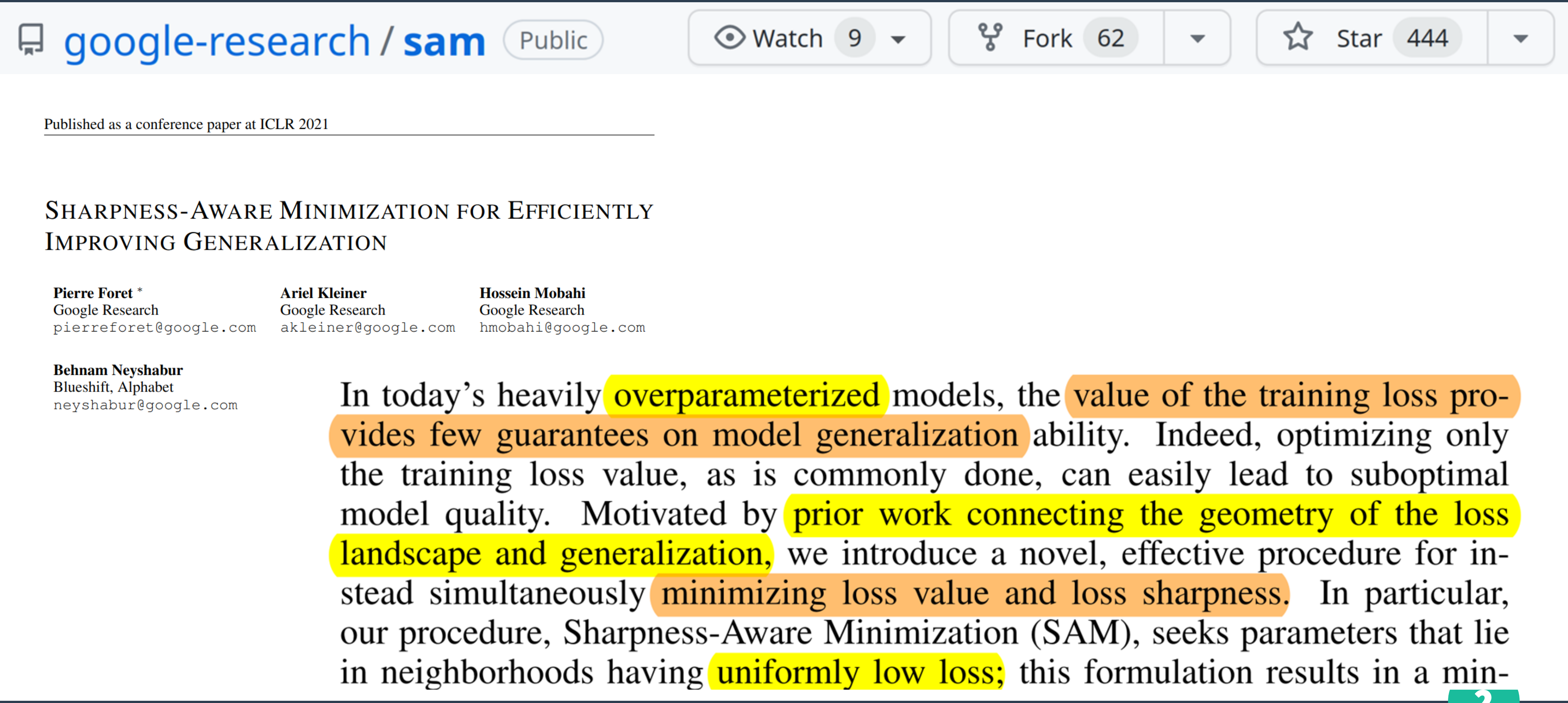
Sharpness aware minimization

Axel Böhm



google-research / sam

Published as a conference paper at ICLR 2021

Sharpness-Aware Minimization for Efficiently Improving Generalization

Pierre Foret *
Google Research
pierreforet@google.com

Ariel Kleiner
Google Research
akleiner@google.com

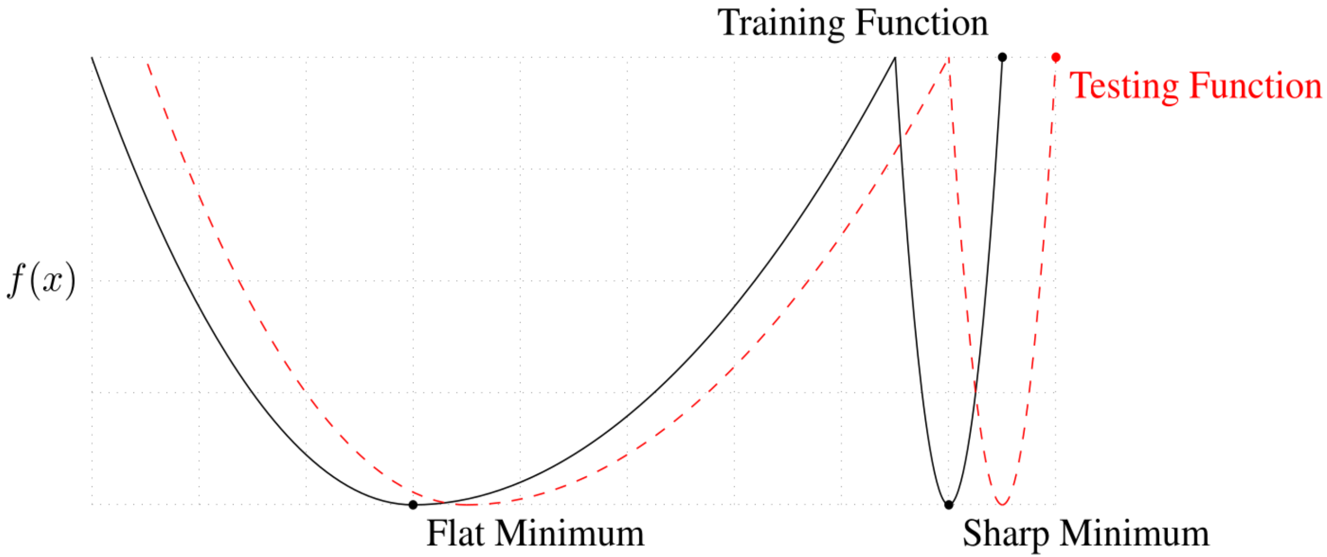
Hossein Mobahi
Google Research
hmobahi@google.com

Behnam Neyshabur
Blueshift, Alphabet
neyshabur@google.com

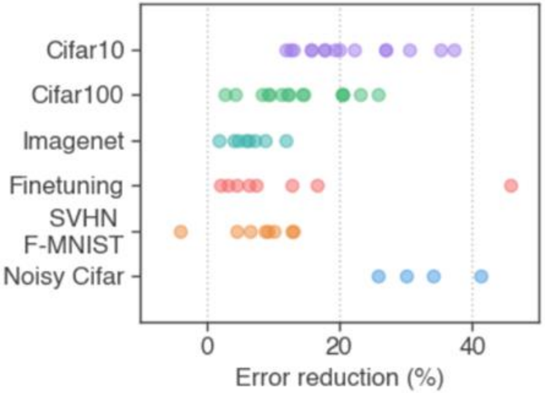
In today's heavily overparameterized models, the value of the training loss provides few guarantees on model generalization ability. Indeed, optimizing only the training loss value, as is commonly done, can easily lead to suboptimal model quality. Motivated by prior work connecting the geometry of the loss landscape and generalization, we introduce a novel, effective procedure for instead simultaneously minimizing loss value and loss sharpness. In particular, our procedure, Sharpness-Aware Minimization (SAM), seeks parameters that lie in neighborhoods having uniformly low loss; this formulation results in a min-

Why flatness?

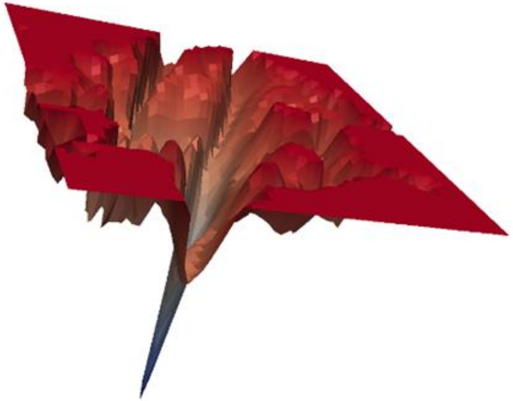
- description length
- Free Gibbs energy
- Bayesian learning
- Intuition (r.h.s.)



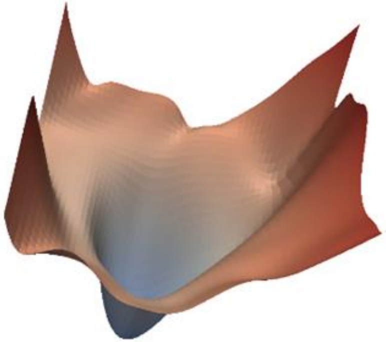
Extensive empirical evaluation



Error rate reduction obtained by switching to SAM. Each point is a different dataset / model / data augmentation

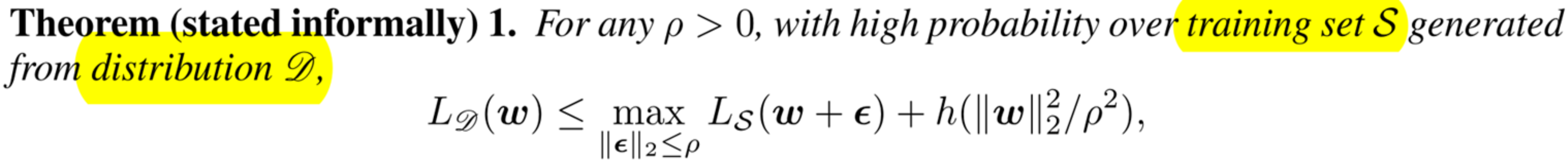


A sharp minimum to which a ResNet trained with SGD converged



A wide minimum to which the same ResNet trained with SAM converged.

The motivation



Theorem (stated informally) 1. *For any $\rho > 0$, with high probability over training set S generated from distribution $\mathscr{D}$,*

$$L_{\mathscr{D}}(\boldsymbol{w}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\boldsymbol{w} + \boldsymbol{\epsilon}) + h(\|\boldsymbol{w}\|_2^2/\rho^2),$$

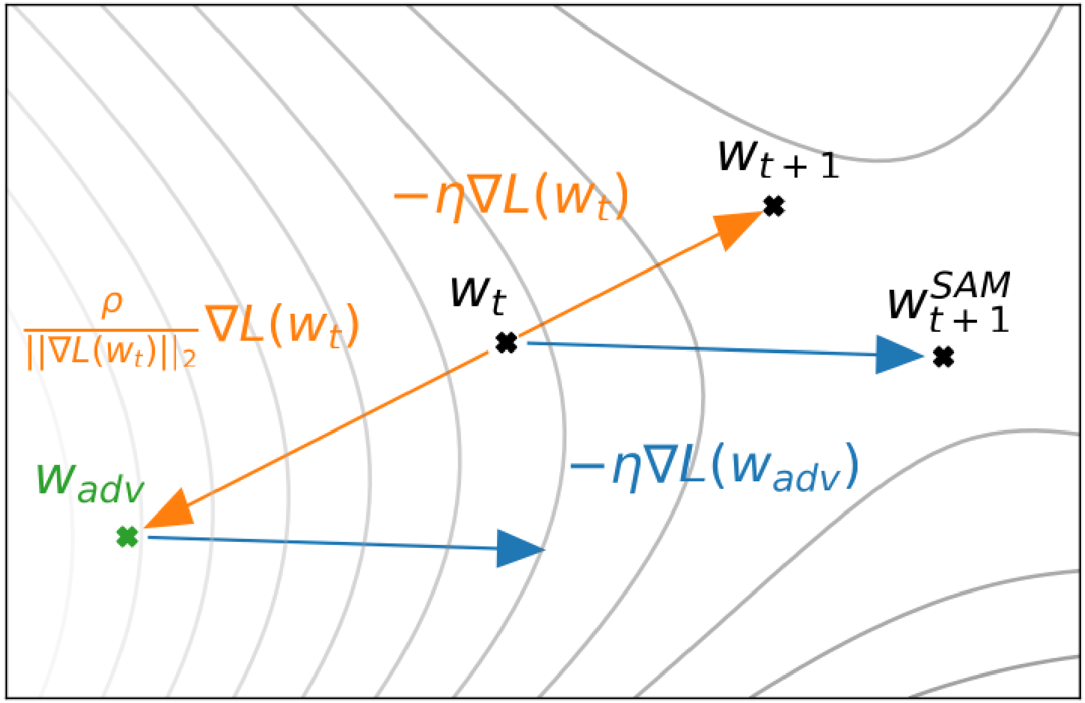
- PAC-Bayes (Probably Approximately Correct ‘84)
- Minimize the r.h.s. instead:

$$\min_{\boldsymbol{w}} L_S^{SAM}(\boldsymbol{w}) + \lambda||\boldsymbol{w}||_2^2 \quad \text{where} \quad L_S^{SAM}(\boldsymbol{w}) \triangleq \max_{||\boldsymbol{\epsilon}||_p \leq \rho} L_S(\boldsymbol{w} + \boldsymbol{\epsilon}),$$

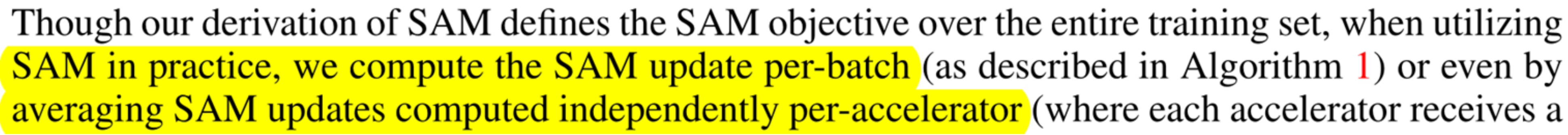
The algorithm

$$x(t+1) = x(t) - \eta \nabla L\left(x + \rho \frac{\nabla L(x)}{\|\nabla L(x)\|_2}\right)$$

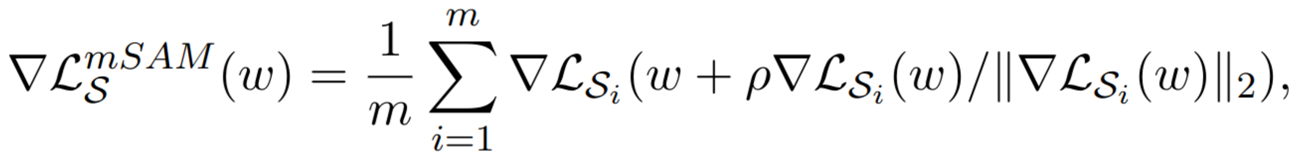
$$\boldsymbol{\epsilon}^*(\boldsymbol{w}) \triangleq \underset{\|\boldsymbol{\epsilon}\|_p \le \rho}{\arg\max}\, L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) \approx \underset{\|\boldsymbol{\epsilon}\|_p \le \rho}{\arg\max}\, L_{\mathcal{S}}(\boldsymbol{w}) + \boldsymbol{\epsilon}^T \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w}) = \underset{\|\boldsymbol{\epsilon}\|_p \le \rho}{\arg\max}\, \boldsymbol{\epsilon}^T \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w}).$$



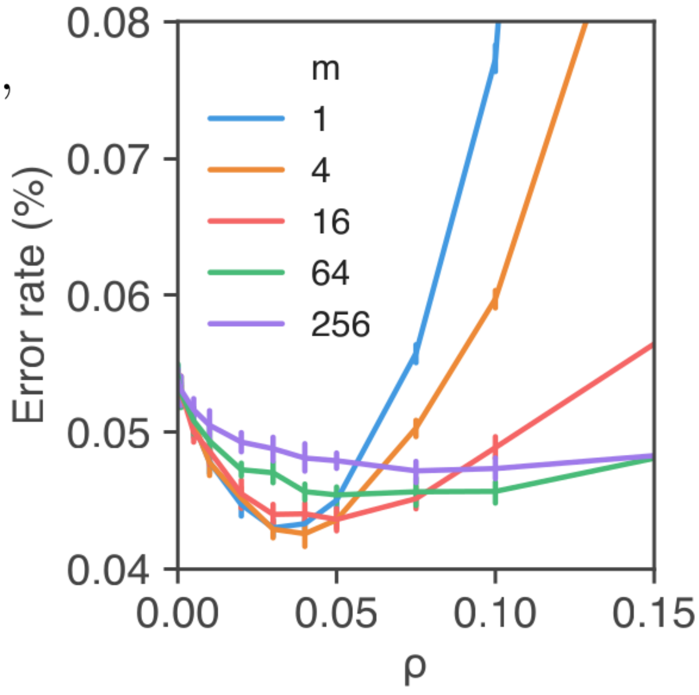
In implementation: mSAM

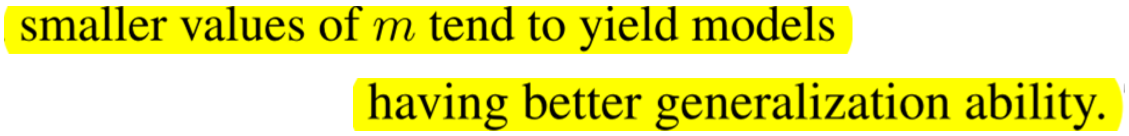


Though our derivation of SAM defines the SAM objective over the entire training set, when utilizing SAM in practice, we compute the SAM update per-batch (as described in Algorithm 1) or even by averaging SAM updates computed independently per-accelerator (where each accelerator receives a



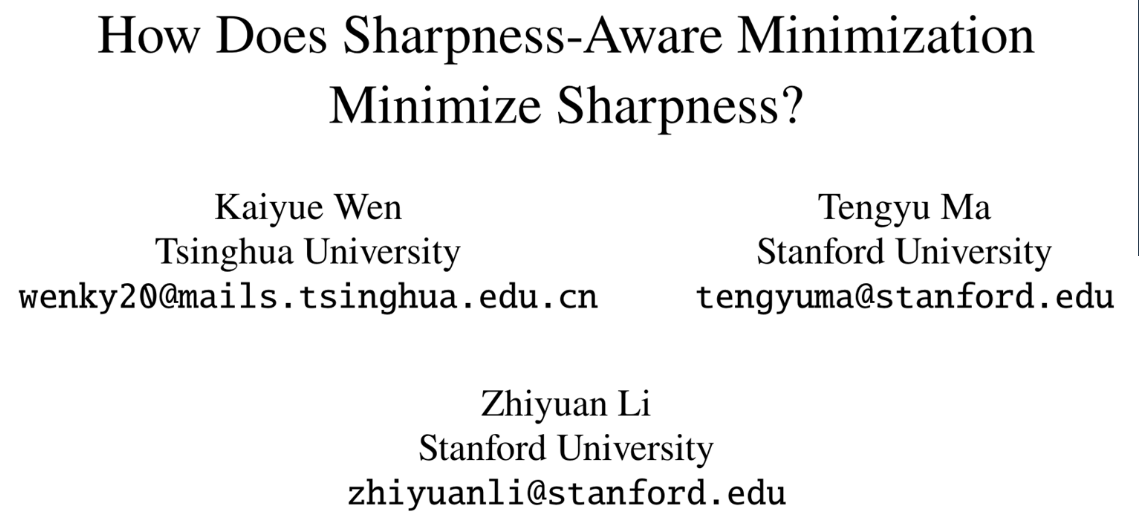
$$\nabla \mathcal{L}_{\mathcal{S}}^{mSAM}(w) = \frac{1}{m}\sum_{i=1}^{m} \nabla \mathcal{L}_{\mathcal{S}_i}(w + \rho \nabla \mathcal{L}_{\mathcal{S}_i}(w)/\|\nabla \mathcal{L}_{\mathcal{S}_i}(w)\|_2),$$





smaller values of m tend to yield models having better generalization ability.

Better understanding

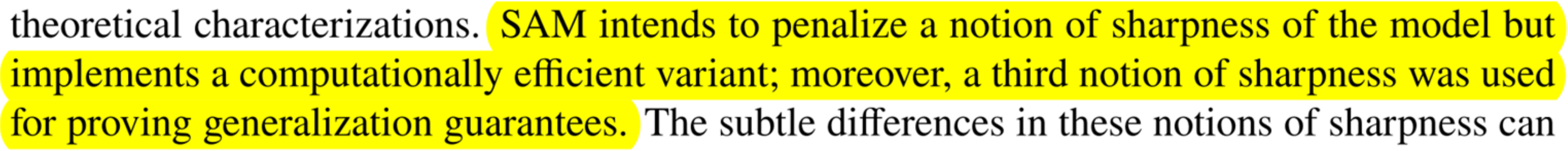


How Does Sharpness-Aware Minimization Minimize Sharpness?

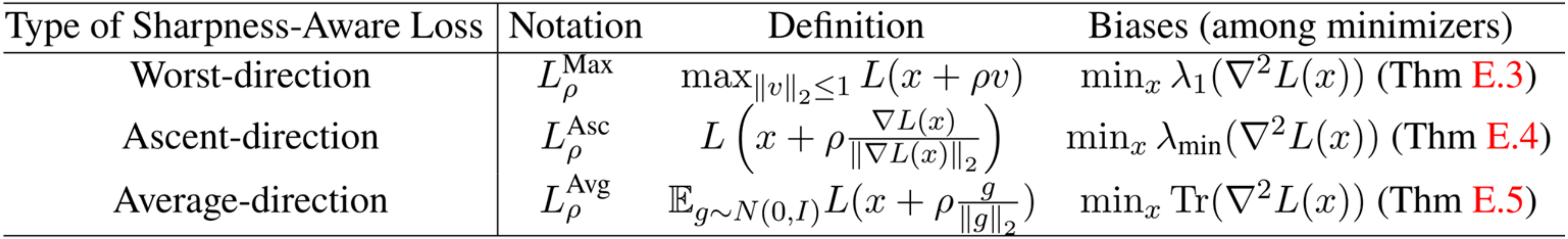
Kaiyue Wen
Tsinghua University
wenky20@mails.tsinghua.edu.cn

Tengyu Ma
Stanford University
tengyuma@stanford.edu

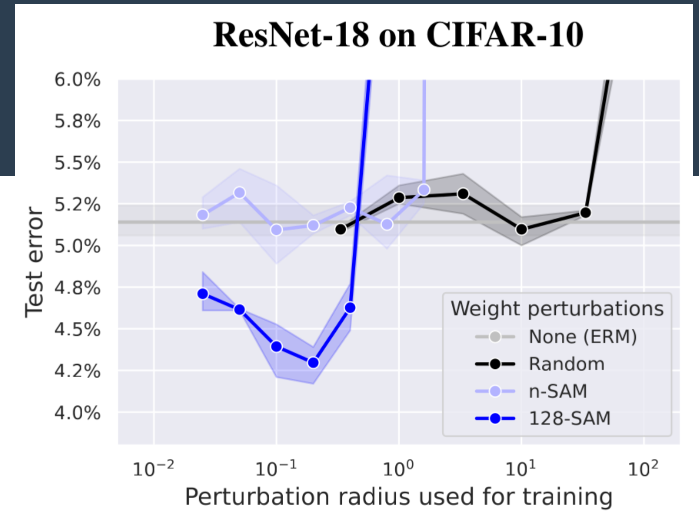
Zhiyuan Li
Stanford University
zhiyuanli@stanford.edu

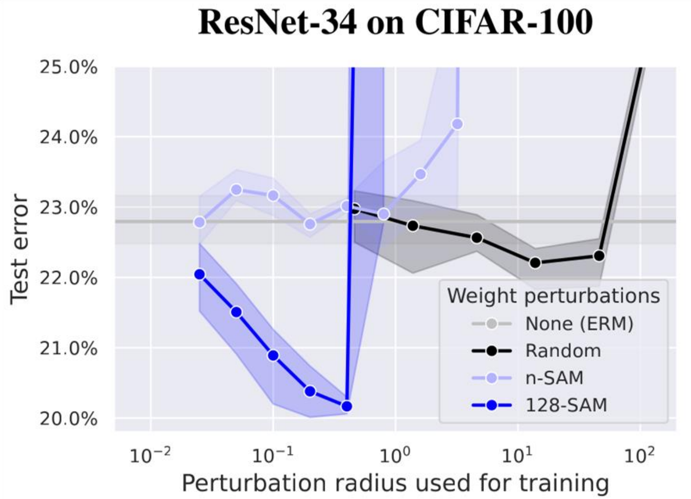


theoretical characterizations. SAM intends to penalize a notion of sharpness of the model but implements a computationally efficient variant; moreover, a third notion of sharpness was used for proving generalization guarantees. The subtle differences in these notions of sharpness can



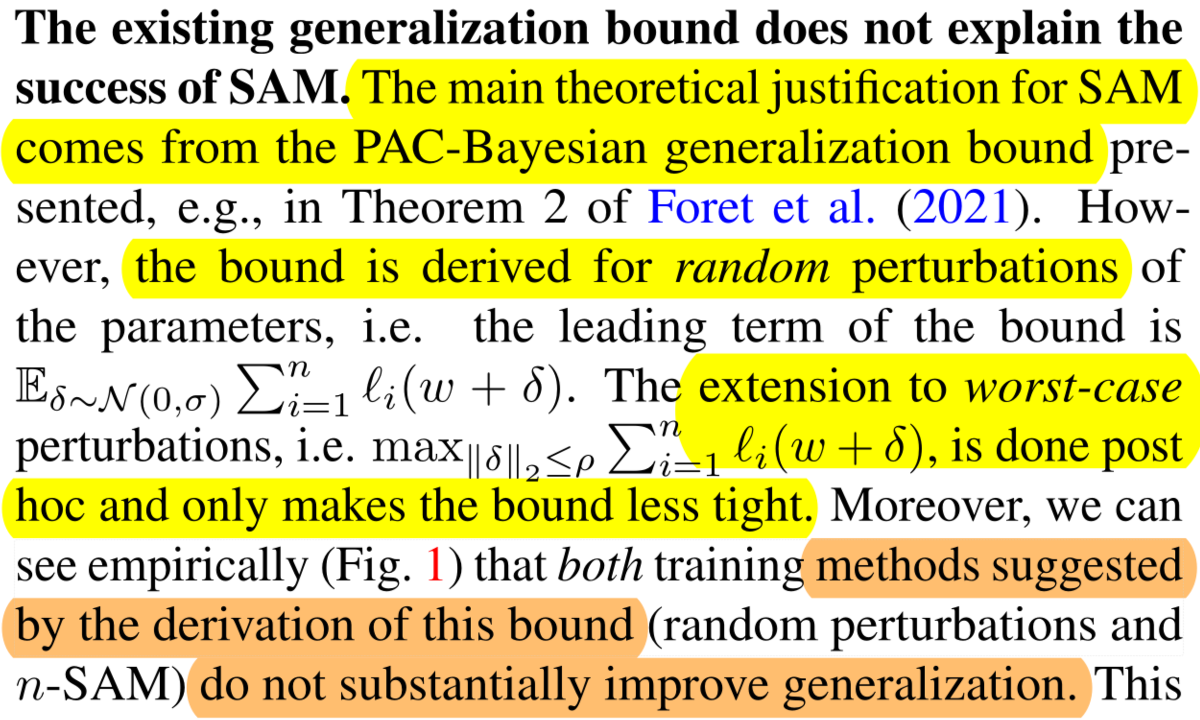
Type of Sharpness-Aware Loss	Notation	Definition	Biases (among minimizers)
Worst-direction	L_ρ^{Max}	$\max_{\\|v\\|_2 \leq 1} L(x+\rho v)$	$\min_x \lambda_1(\nabla^2 L(x))$ (Thm E.3)
Ascent-direction	L_ρ^{Asc}	$L\left(x+\rho\frac{\nabla L(x)}{\\|\nabla L(x)\\|_2}\right)$	$\min_x \lambda_{\min}(\nabla^2 L(x))$ (Thm E.4)
Average-direction	L_ρ^{Avg}	$\mathbb{E}_{g\sim N(0,I)} L(x+\rho\frac{g}{\\|g\\|_2})$	$\min_x \mathrm{Tr}(\nabla^2 L(x))$ (Thm E.5)


ResNet-18 on CIFAR-10


ResNet-34 on CIFAR-100

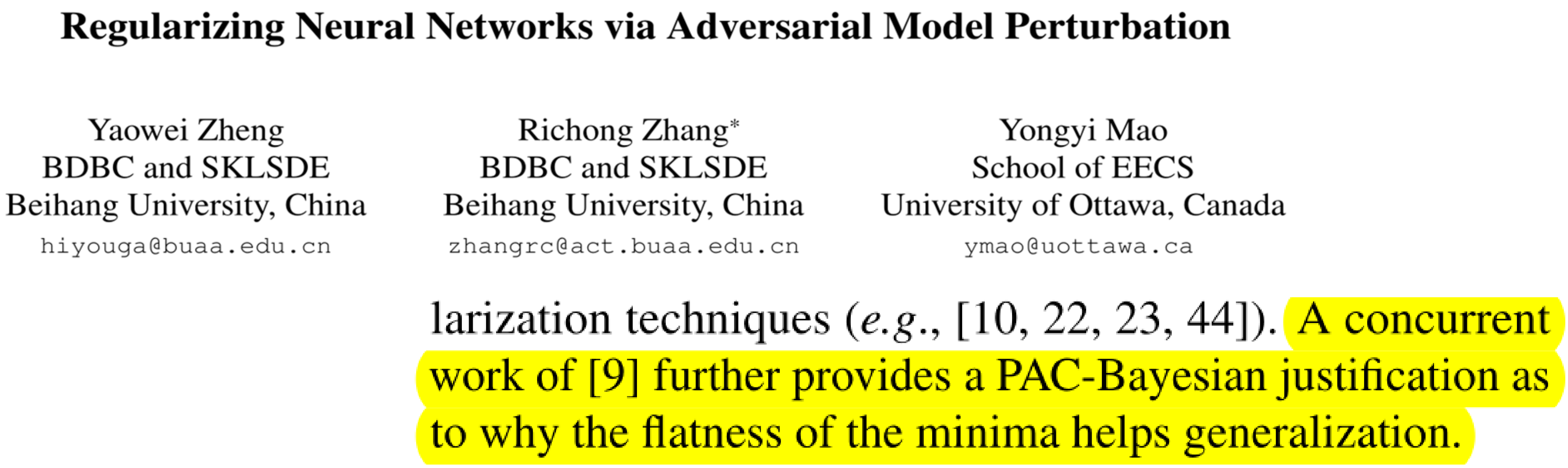
Towards Understanding Sharpness-Aware Minimization

Maksym Andriushchenko [1] Nicolas Flammarion [1]



The existing generalization bound does not explain the success of SAM. The main theoretical justification for SAM comes from the PAC-Bayesian generalization bound presented, e.g., in Theorem 2 of Foret et al. (2021). However, the bound is derived for *random* perturbations of the parameters, i.e. the leading term of the bound is $\mathbb{E}_{\delta\sim\mathcal{N}(0,\sigma)}\sum_{i=1}^{n}\ell_i(w+\delta)$. The extension to *worst-case* perturbations, i.e. $\max_{\|\delta\|_2\leq\rho}\sum_{i=1}^{n}\ell_i(w+\delta)$, is done post hoc and only makes the bound less tight. Moreover, we can see empirically (Fig. 1) that *both* training methods suggested by the derivation of this bound (random perturbations and n-SAM) do not substantially improve generalization. This

Related literature: A concurrent work



Regularizing Neural Networks via Adversarial Model Perturbation

Yaowei Zheng
BDBC and SKLSDE
Beihang University, China
hiyouga@buaa.edu.cn

Richong Zhang*
BDBC and SKLSDE
Beihang University, China
zhangrc@act.buaa.edu.cn

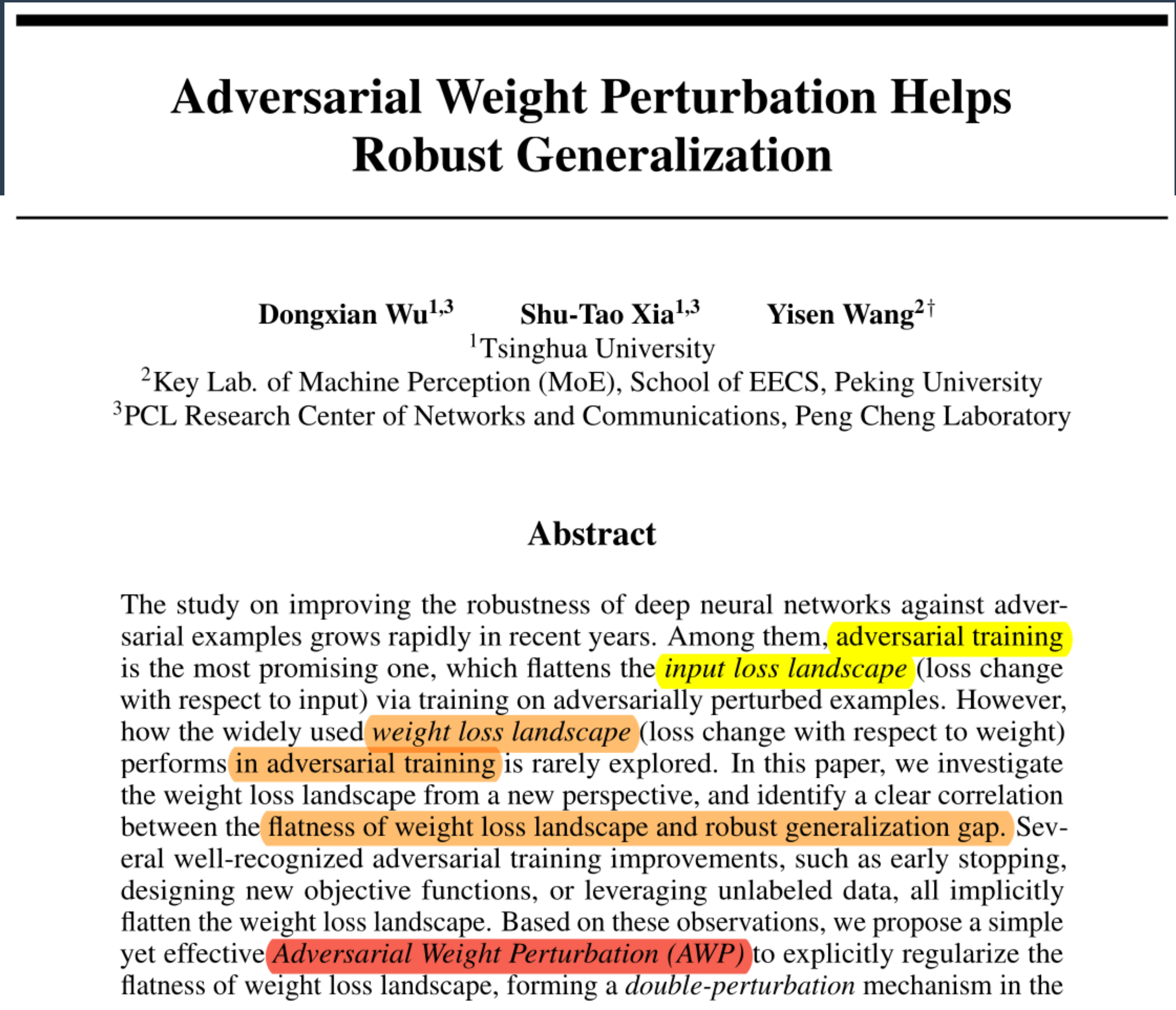
Yongyi Mao
School of EECS
University of Ottawa, Canada
ymao@uottawa.ca

larization techniques (*e.g.*, [10, 22, 23, 44]). A concurrent work of [9] further provides a PAC-Bayesian justification as to why the flatness of the minima helps generalization.

$$\mathcal{L}_{\text{AMP}}(\boldsymbol{\theta}) := \max_{\Delta:\|\Delta\|\le\epsilon} \mathcal{L}_{\text{ERM}}(\boldsymbol{\theta}+\Delta)$$

The prequel

- Basically SAM for adversarial training
- Predates all other papers



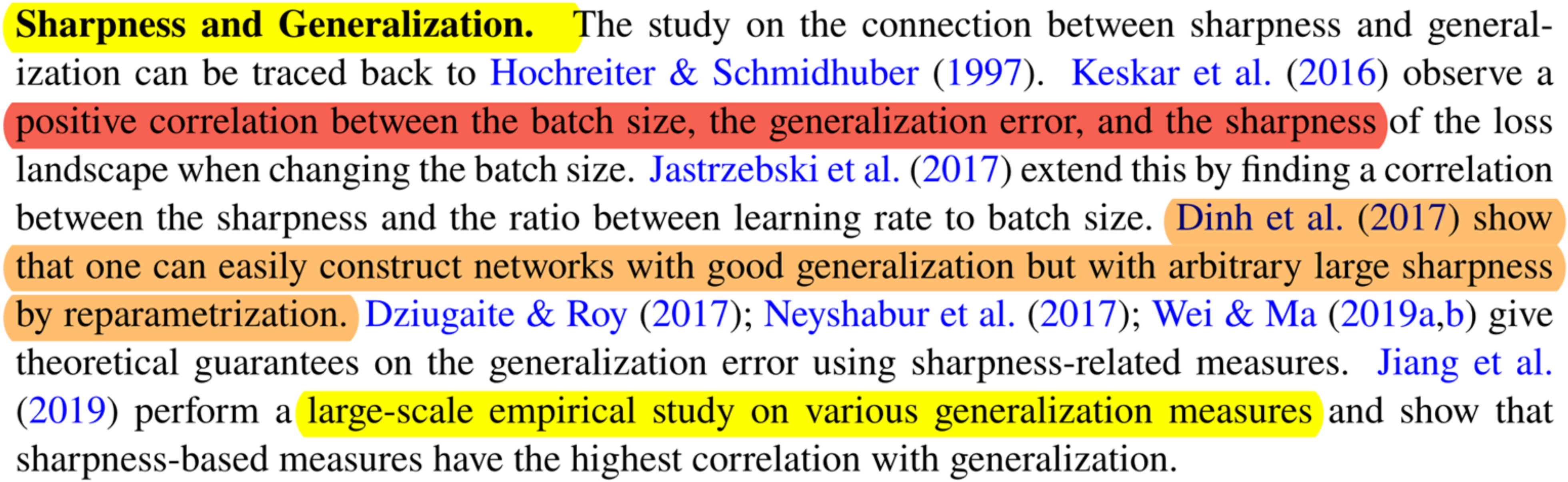
Adversarial Weight Perturbation Helps Robust Generalization

Dongxian Wu[1,3] Shu-Tao Xia[1,3] Yisen Wang[2†]
[1]Tsinghua University
[2]Key Lab. of Machine Perception (MoE), School of EECS, Peking University
[3]PCL Research Center of Networks and Communications, Peng Cheng Laboratory

Abstract

The study on improving the robustness of deep neural networks against adversarial examples grows rapidly in recent years. Among them, adversarial training is the most promising one, which flattens the *input loss landscape* (loss change with respect to input) via training on adversarially perturbed examples. However, how the widely used *weight loss landscape* (loss change with respect to weight) performs in adversarial training is rarely explored. In this paper, we investigate the weight loss landscape from a new perspective, and identify a clear correlation between the flatness of weight loss landscape and robust generalization gap. Several well-recognized adversarial training improvements, such as early stopping, designing new objective functions, or leveraging unlabeled data, all implicitly flatten the weight loss landscape. Based on these observations, we propose a simple yet effective *Adversarial Weight Perturbation (AWP)* to explicitly regularize the flatness of weight loss landscape, forming a *double-perturbation* mechanism in the

Generalization



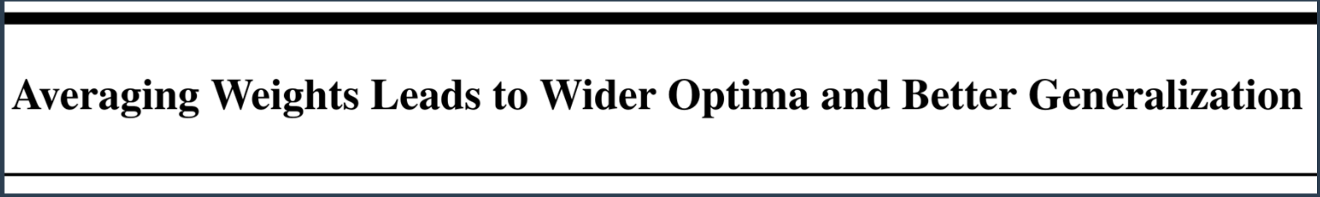
Sharpness and Generalization. The study on the connection between sharpness and generalization can be traced back to Hochreiter & Schmidhuber (1997). Keskar et al. (2016) observe a positive correlation between the batch size, the generalization error, and the sharpness of the loss landscape when changing the batch size. Jastrzebski et al. (2017) extend this by finding a correlation between the sharpness and the ratio between learning rate to batch size. Dinh et al. (2017) show that one can easily construct networks with good generalization but with arbitrary large sharpness by reparametrization. Dziugaite & Roy (2017); Neyshabur et al. (2017); Wei & Ma (2019a,b) give theoretical guarantees on the generalization error using sharpness-related measures. Jiang et al. (2019) perform a large-scale empirical study on various generalization measures and show that sharpness-based measures have the highest correlation with generalization.

SAM

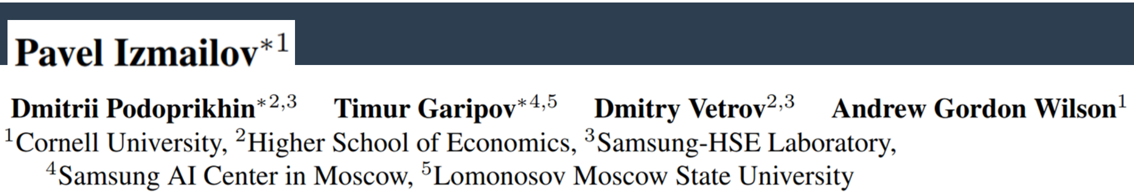
- Improve generalization via flatness
- Easy to implement, only 1.5x slower
- Many approximations / engineering
- Little understanding
- Useful in (semi)-supervised / adversarial learning, noisy labels, ...

Questions?

Bonus



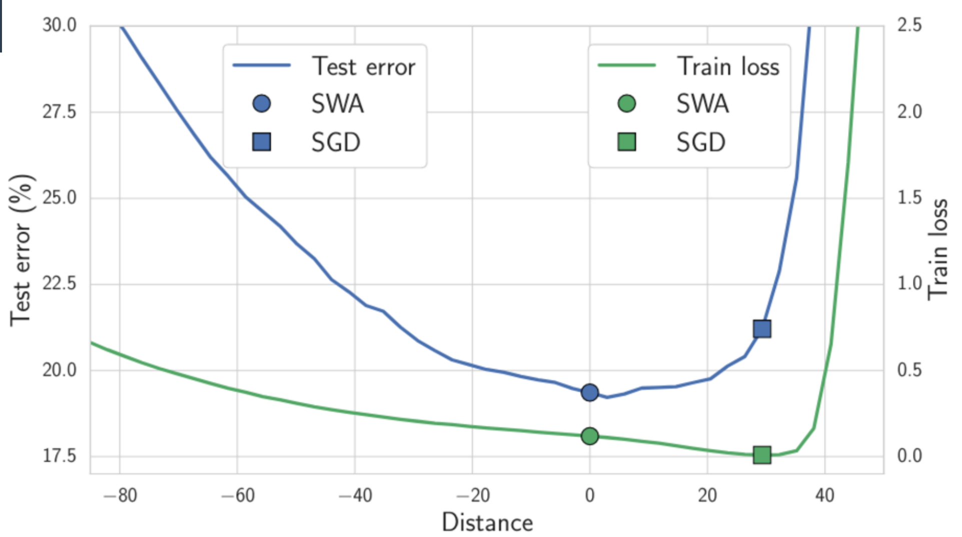
Averaging Weights Leads to Wider Optima and Better Generalization

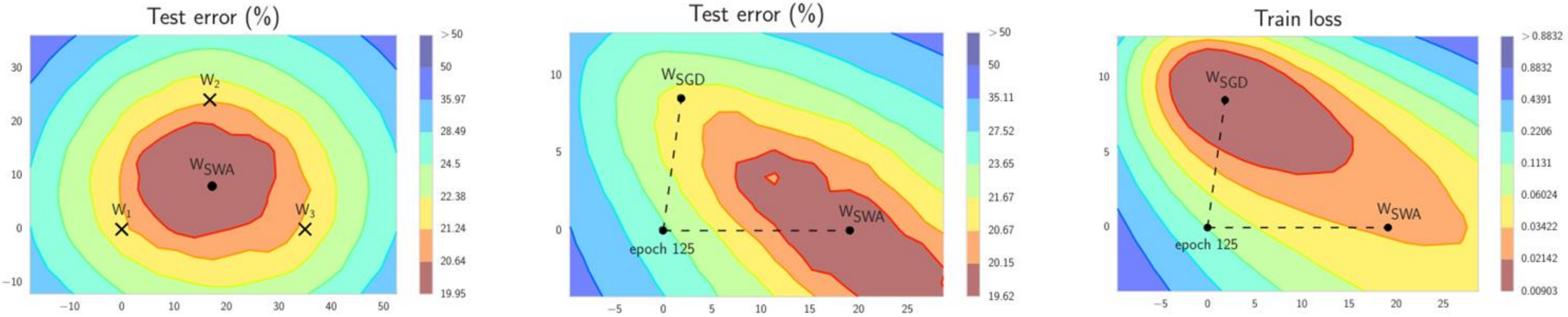


Pavel Izmailov*[1]

Dmitrii Podoprikhin*[2,3] **Timur Garipov***[4,5] **Dmitry Vetrov**[2,3] **Andrew Gordon Wilson**[1]

[1]Cornell University, [2]Higher School of Economics, [3]Samsung-HSE Laboratory,
[4]Samsung AI Center in Moscow, [5]Lomonosov Moscow State University


Test error
SWA
SGD
Train loss
SWA
SGD
Test error (%)
Train loss
Distance


Test error (%)
W2
W1
W3
WSWA
> 50
Test error (%)
WSGD
WSWA
epoch 125
> 50
Train loss
WSGD
WSWA
epoch 125
> 0.8832