

# SHARPNESS-AWARE MINIMIZATION FOR EFFICIENTLY IMPROVING GENERALIZATION

Адыгамов Ильяс, Гусева Полина, Кириллов Дмитрий,  
Дроздова Анастасия

НИУ Высшая школа экономики

06 октября 2021 г.

## 1 Доклад

- О чем эта работа?
- Теоретическая предпосылка
- Выведение метода
- Схема метода
- Эксперименты
- m-острота
- Спектр гессиана

## 2 Контекст

## 3 Рецензия

## 4 Эксперименты

# О чем эта работа?

- Обобщающая способность зависит от гладкости функционала в точке оптимума
- Можно добавить в функционал штраф за остроту
- Результат - заметно лучше обобщающая способность и новые SOTA в некоторых задачах

# Интуитивно понятный пример того, что делает метод

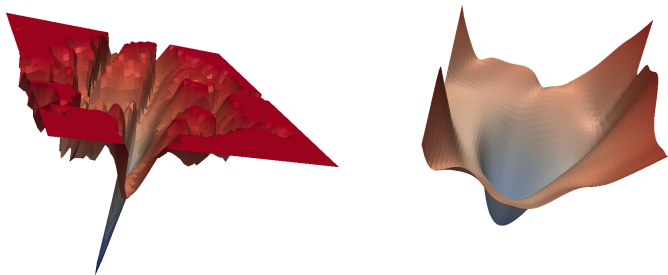


Рис.: Слева изображена поверхность функционала вокруг минимума для обычного SGD. Справа - для SAM

## Theorem (stated informally)

*For any  $\rho > 0$ , with high probability over training set  $S$  generated from distribution  $\mathcal{D}$ ,*

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

*where  $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is a strictly increasing function (under some technical conditions on  $L_{\mathcal{D}}(\mathbf{w})$ ).*

Переписав правую часть можно выделить "слагаемое негладкости"

$$\left[ \max_{\|\epsilon\|_2 \leq \rho} L_S(\mathbf{w} + \epsilon) - L_S(\mathbf{w}) \right] + L_S(\mathbf{w}) + h(\|\mathbf{w}\|_2^2 / \rho^2)$$

# Выведение метода

Запишем функцию потерь

$$\min_{\mathbf{w}} L_S^{SAM}(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \quad \text{где} \quad L_S^{SAM}(\mathbf{w}) \triangleq \max_{\|\epsilon\|_p \leq \rho} L_S(\mathbf{w} + \epsilon) \quad (1)$$

$\epsilon$  оптимальный можно выразить так

$$\hat{\epsilon}(\mathbf{w}) = \rho \operatorname{sign}(\nabla_{\mathbf{w}} L_S(\mathbf{w})) |\nabla_{\mathbf{w}} L_S(\mathbf{w})|^{q-1} / \left( \|\nabla_{\mathbf{w}} L_S(\mathbf{w})\|_q^q \right)^{1/p} \quad (2)$$

где  $1/p + 1/q = 1$ . Подставив это выражение в (1) и продифференцировав, получаем

$$\nabla_{\mathbf{w}} L_S^{SAM}(\mathbf{w}) \approx \nabla_{\mathbf{w}} L_S(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})} + \frac{d\hat{\epsilon}(\mathbf{w})}{d\mathbf{w}} \nabla_{\mathbf{w}} L_S(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})} \quad (3)$$

# Итоговый метод

- 1 Вычислить  $\nabla_{\mathbf{w}} L_{\mathcal{S}}(\mathbf{w})$
- 2 Вычислить  $\epsilon$  оптимальное

$$\hat{\epsilon}(\mathbf{w}) = \rho \operatorname{sign}(\nabla_{\mathbf{w}} L_{\mathcal{S}}(\mathbf{w})) |\nabla_{\mathbf{w}} L_{\mathcal{S}}(\mathbf{w})|^{q-1} / \left( \|\nabla_{\mathbf{w}} L_{\mathcal{S}}(\mathbf{w})\|_q^q \right)^{1/p}$$

- 3 Вычислить градиент (слагаемые второго порядка отброшены)

$$\nabla_{\mathbf{w}} L_{\mathcal{S}}^{SAM}(\mathbf{w}) \approx \nabla_{\mathbf{w}} L_{\mathcal{S}}(\mathbf{w})|_{\mathbf{w} + \hat{\epsilon}(\mathbf{w})}$$

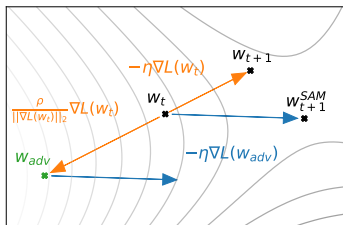


Рис.: Схема обновления параметров в методе SAM

- Авторы просто подменяют процесс оптимизации на SAM.
- Наборы данных: CIFAR- $\{10, 100\}$ , ImageNet, SVHN, Fashion-MNIST.
- Задачи: image classification from scratch, finetuning, learning with noisy labels.
- Каждый эксперимент запускается по 5 раз и берется среднее.
- Framework: JAX, Hardware: 8 x V100 and TPUv3 for ImageNet.
- Модели: WideResNet with ShakeShake, PyramidNet with ShakeDrop, ResNet-50, 101, 152 for ImageNet, EffecientNet-b7 and EffecientNet-L2 for finetuning.
- Аугментации.



Model	Epoch	SAM		Standard Training (No SAM)	
		Top-1	Top-5	Top-1	Top-5
ResNet-50	100	<b>22.5</b> $\pm 0.1$	6.28 $\pm 0.08$	22.9 $\pm 0.1$	6.62 $\pm 0.11$
	200	<b>21.4</b> $\pm 0.1$	5.82 $\pm 0.03$	22.3 $\pm 0.1$	6.37 $\pm 0.04$
	400	<b>20.9</b> $\pm 0.1$	5.51 $\pm 0.03$	22.3 $\pm 0.1$	6.40 $\pm 0.06$
ResNet-101	100	<b>20.2</b> $\pm 0.1$	5.12 $\pm 0.03$	21.2 $\pm 0.1$	5.66 $\pm 0.05$
	200	<b>19.4</b> $\pm 0.1$	4.76 $\pm 0.03$	20.9 $\pm 0.1$	5.66 $\pm 0.04$
	400	<b>19.0</b> $\pm <0.01$	4.65 $\pm 0.05$	22.3 $\pm 0.1$	6.41 $\pm 0.06$
ResNet-152	100	<b>19.2</b> $\pm <0.01$	4.69 $\pm 0.04$	20.4 $\pm <0.0$	5.39 $\pm 0.06$
	200	<b>18.5</b> $\pm 0.1$	4.37 $\pm 0.03$	20.3 $\pm 0.2$	5.39 $\pm 0.07$
	400	<b>18.4</b> $\pm <0.01$	4.35 $\pm 0.04$	20.9 $\pm <0.0$	5.84 $\pm 0.07$

Таблица: Доля ошибок для экспериментов на ImageNet.

## Как считать градиент SAM параллельно?

Авторы распределяют данные равномерно по GPU (каждый получает по  $m$  объектов) и на каждом считают градиент SAM независимо. Результаты усредняют.

Получившуюся меру авторы называют  *$m$ -остротой*

То есть

$$L_S^{SAM}(\mathbf{w}) \triangleq \max_{\|\epsilon\|_p \leq \rho} L_S(\mathbf{w} + \epsilon)$$

переходит в

$$L_S^{SAM}(\mathbf{w}) \triangleq \sum_{b \sim \text{Unif}^m(X)} \max_{\|\epsilon\|_p \leq \rho} L_b(\mathbf{w} + \epsilon)$$

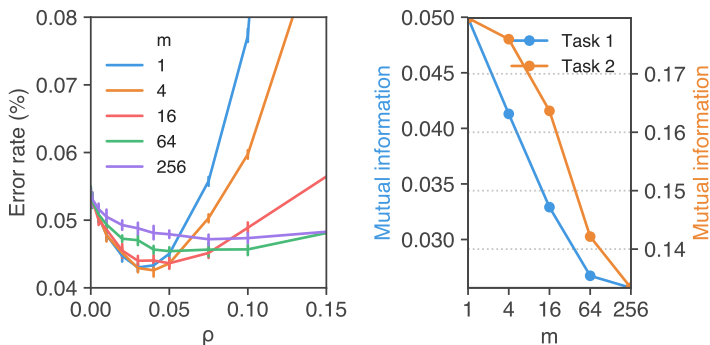


Рис.: (слева) Ошибка на тесте, как функция от  $\rho$  для разных значений  $m$ . (справа) Предсказывающая способность  $m$ -негладкости для обобщаемости при разных значениях  $m$  (большие значения говорят о большей корреляции). Левая шкала показывает значения для "Task 1" правая - для "Task 2".

# Спектр гессiana

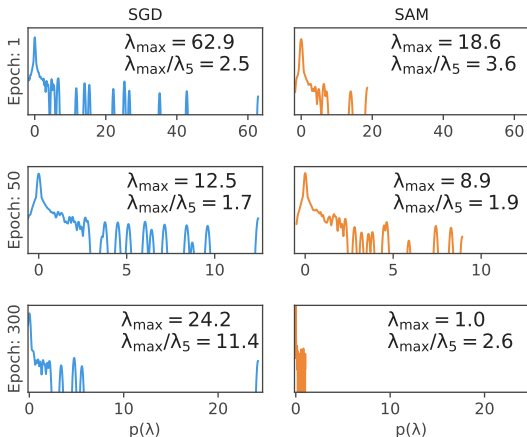


Рис.: Изменение спектра гессiana во время обучения для SGD(слева) и для SAM(справа).

- Авторы представили новый метод с хорошим теоретическим обоснованием.
- Подтвердили полезность метода экспериментально.
- Можно пробовать заменять классическую оптимизацию на SAM в своих задачах.

## 1 Доклад

- О чем эта работа?
- Теоретическая предпосылка
- Выведение метода
- Схема метода
- Эксперименты
- m-острота
- Спектр гессиана

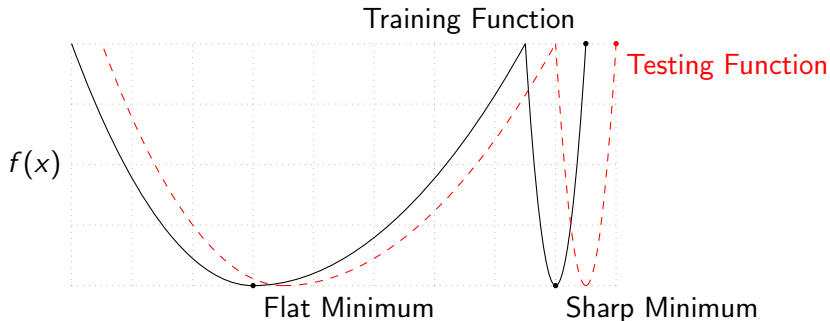
## 2 Контекст

## 3 Рецензия

## 4 Эксперименты

- Pierre Foret — магистр Barkley. Был одним из организаторов соревнования по оценке генерализации на NIPS 2020.
- Ariel Kleiner до значительного перерыва публиковал работы по высокопроизводительному машинному обучению и большим данным
- Hossein Mobahi и Behnam Neyshabur более заслуженные исследователи ( $h\text{-index } 22 \pm 1$ ). В последние годы увеличивалось количество исследование обобщающей способности моделей

# На что опирается



1


<sup>1</sup>Nitish Shirish Keskar и др. "On large-batch training for deep learning: Generalization gap and sharp minima". в: [arXiv preprint arXiv:1609.04836](https://arxiv.org/abs/1609.04836) (2016). 16/30



# На что опирается

- Масштабное сравнение [6] разных метрик генерализации, которое показало перспективность "sharpness" подхода
- В исследовании [1] была использована схожая техника вывода теоретических оценок

- ❶ Похожая работа [11], в которой с той же мотивацией демонстрируют, что небольшое изменение параметров сети может сильно испортить качество.
- ❷ SWA [5] собирает несколько наборов весов в процессе оптимизации и усредняет их
- ❸ В работе [12] получился весьма похожий шаг оптимизатора. Заметно больший теоретический анализ и меньше экспериментальных данных.

Репозиторий	★	
<a href="#">google-research/sam</a>	290	
<a href="#">davda54/sam</a>	700	
<a href="#">sayakpaul/Sharpness-Aware-Minimization-TensorFlow</a>	27	

- 60 работ цитируют SAM. Сам метод в них активно не используется, но все сравниваются с полученными результатами
- ASAM [9] предлагает модификацию инвариантную к шкалированию весов

## 1 Доклад

- О чем эта работа?
- Теоретическая предпосылка
- Выведение метода
- Схема метода
- Эксперименты
- m-острота
- Спектр гессиана

## 2 Контекст

## 3 Рецензия

## 4 Эксперименты

- Проведено **обширное эмпирическое исследование** того, как SAM влияет на качество предсказаний. Использовались разные архитектуры, наборы данных и методы регуляризации.
- Эксперименты показали, что метод действительно дает **значимое улучшение**.
- Статья написана четко и понятно, ее **легко читать**.
- Метод описан достаточно подробно, так что **результаты можно воспроизвести**.

- Нет сравнений с близкими работами [2, 5, 12] и прямыми конкурентами
- Аналогичный метод был предложен под названием extragradient method [8] и неоднократно появлялся в недавних работах по оптимизации [3]. Он в статье не упомянут.
- У метода нет теоретического обоснования, так как взятая за основу оценка держится только в очень специфичных условиях.
- Поведение метода во время обучения не изучено, поскольку эксперименты фокусируются только на улучшении метрик.

- Экспериментальная установка в статье описана подробно. В аппендиксе приведены подобранные гиперпараметры, код на JAX выложен в открытый доступ.
- Существуют неофициальные реализации, в которых попытались воспроизвести результаты в классификации CIFAR-10.

Реализация	Модель	С SAM	Без SAM
<a href="#">davda54</a>	WRN 16-8	97.1%	96.8%
<a href="#">sayakpaul</a>	ResNet20	80.5%	83.1%

- 1 Доклад
  - О чем эта работа?
  - Теоретическая предпосылка
  - Выведение метода
  - Схема метода
  - Эксперименты
  - m-острота
  - Спектр гессиана
- 2 Контекст
- 3 Рецензия
- 4 Эксперименты



- Наборы данных: FashionMNIST, CIFAR10
- Параметры SAM:  $p = 2$ ,  $\rho = 0.05$
- Косинусное расписание длины шага
- ConvNet:  $\alpha = 0.1$ (FMNIST),  $\alpha = 0.01$ (CIFAR), momentum=0.9, weight\_decay=0.0001
- ResNet:  $\alpha = 1.0$ (CIFAR), momentum=0.9, weight\_decay=0.0001

Набор данных	Модель	SGD	SAM
Fashion-MNIST	ConvNet	0.8730	0.8751
CIFAR10	ConvNet	0.7171	0.6952
CIFAR10	ResNet	0.7349	0.7579

- Графики функции потерь в области оптимума для сверточной сети, получены с помощью [10].

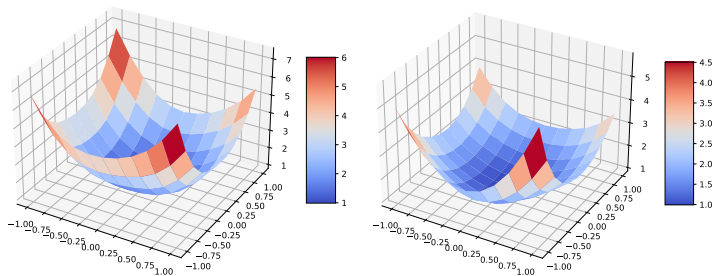


Рис.: Функция потерь в области оптимума. Слева: SGD, справа: SAM

- Графики функции потерь в области оптимума для ResNet.

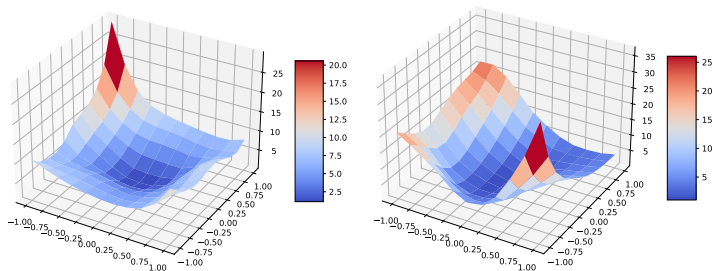


Рис.: Функция потерь в области оптимума. Слева: SGD, справа: SAM



Niladri S Chatterji, Behnam Neyshabur и Hanie Sedghi. “The intriguing role of module criticality in the generalization of deep networks”. в: *arXiv preprint arXiv:1912.00528* (2019).



Pratik Chaudhari и др. “Entropy-SGD: Biasing Gradient Descent Into Wide Valleys”. в: *CoRR abs/1611.01838* (2016). arXiv: 1611.01838. URL: <http://arxiv.org/abs/1611.01838>.



Jelena Diakonikolas и L. Orecchia. “Accelerated Extra-Gradient Descent: A Novel Accelerated First-Order Method”. в: *ITCS*. 2018.



Pierre Foret и др. “Sharpness-Aware Minimization for Efficiently Improving Generalization”. в: *CoRR abs/2010.01412* (2020). arXiv: 2010.01412. URL: <https://arxiv.org/abs/2010.01412>.



Pavel Izmailov и др. “Averaging Weights Leads to Wider Optima and Better Generalization”. в: *CoRR abs/1803.05407* (2018). arXiv: 1803.05407. URL: <http://arxiv.org/abs/1803.05407>.



Yiding Jiang и др. “Fantastic generalization measures and where to find them”. в: *arXiv preprint arXiv:1912.02178* (2019).



Nitish Shirish Keskar и др. “On large-batch training for deep learning: Generalization gap and sharp minima”. в: *arXiv preprint arXiv:1609.04836* (2016).



G. M. Korpelevich. “The extragradient method for finding saddle points and other problems”. в: 1976.



Jungmin Kwon и др. “ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks”. в: *CoRR abs/2102.11600* (2021). arXiv: 2102.11600. URL: <https://arxiv.org/abs/2102.11600>.



Hao Li и др. “Visualizing the Loss Landscape of Neural Nets”. в: *Neural Information Processing Systems*. 2018.



Ху Sun и др. “Exploring the vulnerability of deep neural networks: A study of parameter corruption”. в: *arXiv preprint arXiv:2006.05620* (2020).



Colin Wei и Tengyu Ma. “Improved Sample Complexities for Deep Networks and Robust Classification via an All-Layer Margin”. в: *CoRR abs/1910.04284* (2019). arXiv: [1910.04284](https://arxiv.org/abs/1910.04284). URL: <http://arxiv.org/abs/1910.04284>.