# Sharp-MAML: Sharpness-Aware Model-Agnostic Meta-Learning

*Momin Abbas** *Quan Xiao** *Lisha Chen** *Pin-Yu Chen* *Tianyi Chen*

**Rensselaer Polytechnic Institute**

**IBM Thomas J. Watson Research Center**

**ICML 2022**

# Meta-Learning: A Challenging Nested Problem

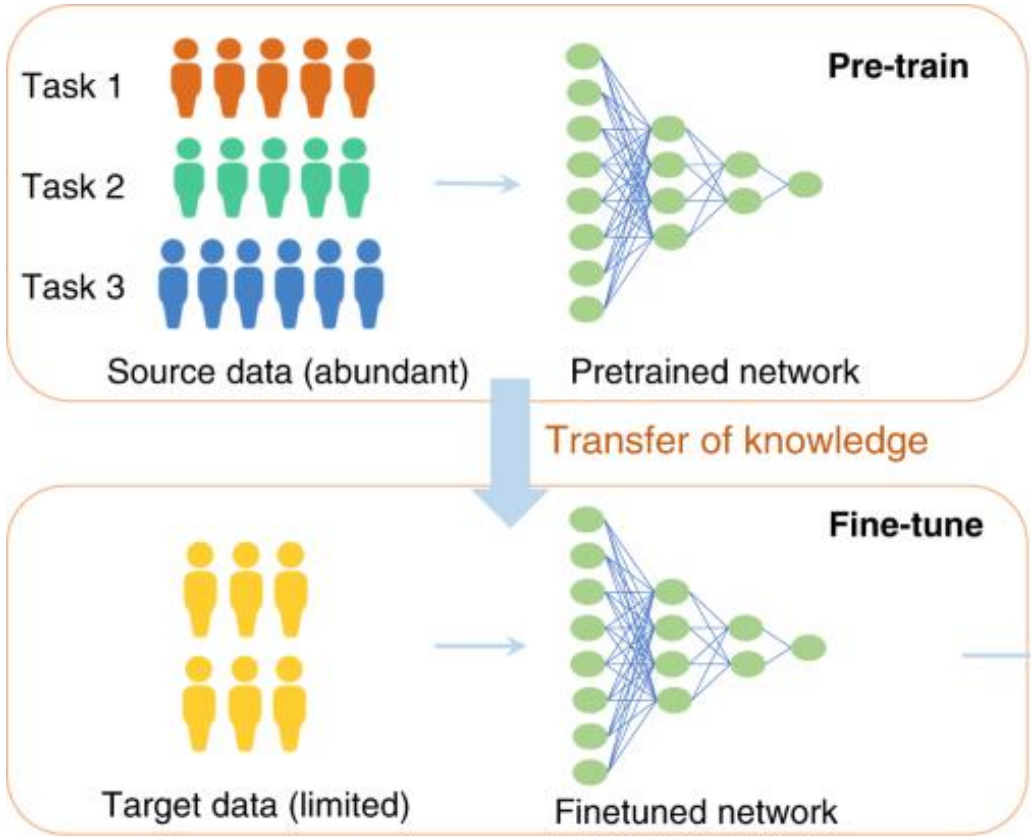



Image from Internet

## Meta learning

Model initialization, shared representation

$\min_{\theta}$ meta testing (learned $\theta_m^*$ under $\theta$)

$\theta_m^* \in \underset{\theta_m}{\operatorname{argmin}}$ fine-tuning $\theta$

# Nested Structure Yields Complex Loss Landscape

**MAML [Finn et al., 2017] problem:**

$$\min_{\theta} \; F(\theta) := \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\theta'_m(\theta); \mathcal{D}'_m) \quad \text{(upper)}$$

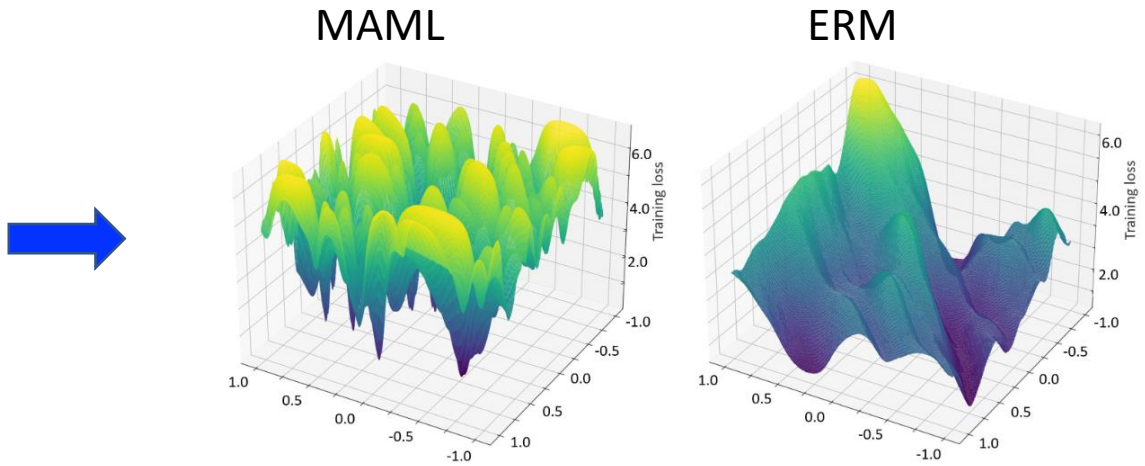
$$\text{s.t. } \theta'_m(\theta) = \theta - \beta_{\text{low}} \nabla_\theta \mathcal{L}(\theta; \mathcal{D}_m). \quad \text{(lower)}$$

# Nested Structure Yields Complex Loss Landscape

**MAML [Finn et al., 2017] problem:**

$$\min_{\theta} \; F(\theta) := \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(\theta'_m(\theta); \mathcal{D}'_m) \quad \text{(upper)}$$

$$\text{s.t. } \theta'_m(\theta) = \theta - \beta_{\text{low}} \nabla_\theta \mathcal{L}(\theta; \mathcal{D}_m). \quad \text{(lower)}$$




Lemma 1 (informal): MAML has more stationary points and local minimizers than ERM (i.e. a more complex loss landscape)

Prior works e.g. [Keskar et al., 2017] show that **sharp minima** yield **poor generalization** than wide minima

# Generalization And Sharpness Of Loss Landscape

Generalization gap : $\mathcal{L}_D(\theta) - \mathcal{L}_S(\theta)$

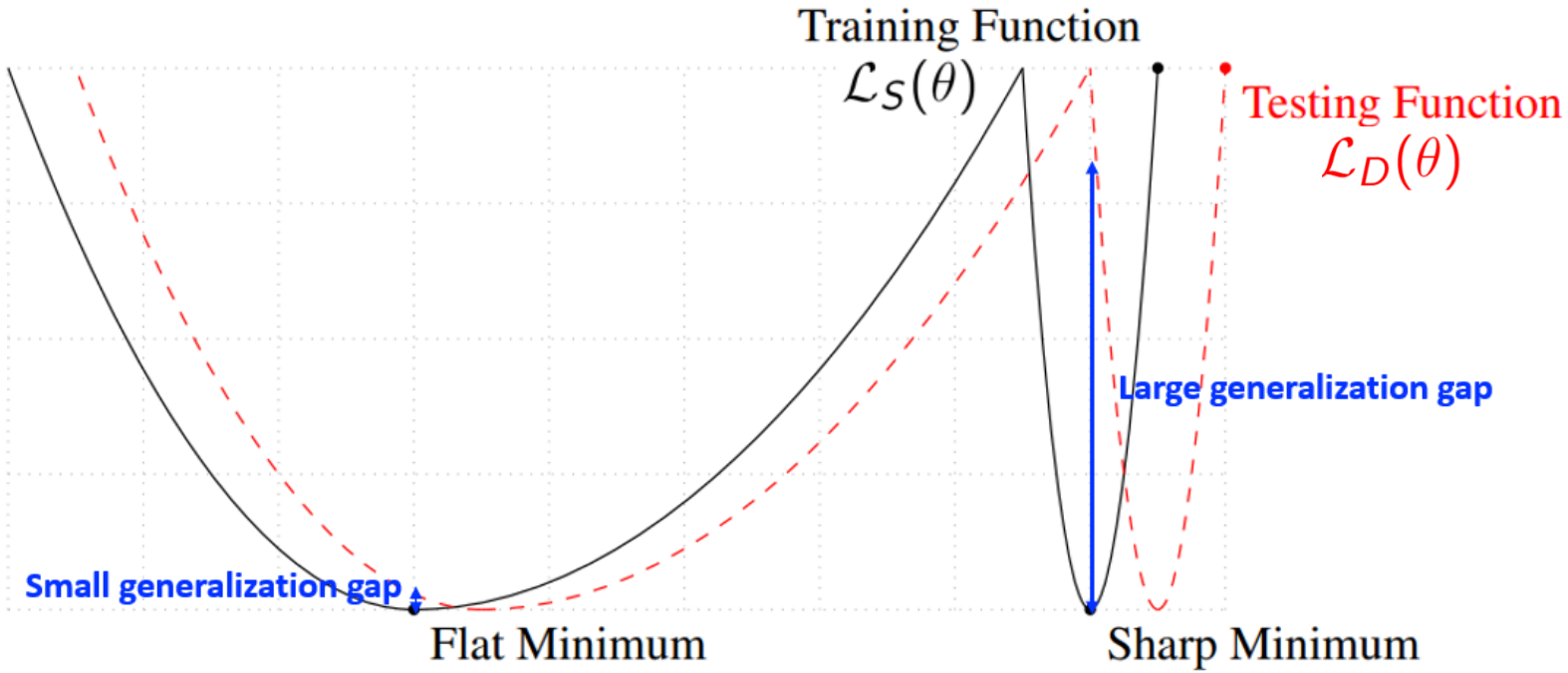
Sharpness : $\max_{||\epsilon||_p \leq \rho} \mathcal{L}_S(\theta + \epsilon) - \mathcal{L}_S(\theta)$

# Generalization And Sharpness Of Loss Landscape

Generalization gap : $\mathcal{L}_D(\theta) - \mathcal{L}_S(\theta)$

Sharpness : $\max_{||\epsilon||_p \leq \rho} \mathcal{L}_S(\theta + \epsilon) - \mathcal{L}_S(\theta)$

Figure: A Conceptual Sketch of Flat and Sharp Minima [Keskar et al., 2017]



# Sharpness-Aware Minimization (SAM)

**SAM [Foret et al., 2020]:**

$$\min_{\theta} \underbrace{\max_{||\epsilon||_p \leq \rho} \mathcal{L}_S(\theta + \epsilon)}_{\max_{||\epsilon||_p \leq \rho} \mathcal{L}_S(\theta + \epsilon) - \mathcal{L}_S(\theta) + \mathcal{L}_S(\theta)}$$

Minimize sharpness and training loss to improve the generalization performance

**Algorithm:**

1) compute SGD gradient $\nabla \mathcal{L}_S(\theta)$
2) compute epsilon $\epsilon$ using SGD gradient
3) compute SAM gradient $\nabla \mathcal{L}_S(\theta + \epsilon)$
4) update model $\theta$ by descending SAM gradient $\nabla \mathcal{L}_S(\theta + \epsilon)$

# Sharpness-Aware Minimization MAML (Sharp-MAML)

**Goal:** Improve the generalization performance of MAML

**Problem (Sharp-MAML$_{\text{up}}$):**

$$\min_{\theta} \max_{||\epsilon||_2 \leq \alpha_{\text{up}}} \sum_{i=1}^{M} \mathcal{L}(\theta_m^*(\theta+\epsilon); \mathcal{D}'_m) \quad \text{(upper)}$$

$$\text{s.t. } \theta_m^*(\theta) = \arg\min_{\theta_m} \mathcal{L}(\theta_m; \mathcal{D}_m) + \frac{\|\theta_m - \theta\|^2}{2\beta_{\text{low}}}, \; m = 1, ..., M. \quad \text{(lower)}$$

**Algorithm:**

1) compute upper MAML gradient
2) compute upper epsilon $\epsilon$ using upper MAML gradient
3) compute upper Sharp-MAML gradient using $\epsilon$
4) update upper model $\theta$ by descending Sharp-MAML gradient

# Sharpness-Aware Minimization MAML (Sharp-MAML)

**Goal:** Improve the generalization performance of MAML

**Problem (Sharp-MAML$_{\text{both}}$):**

$$\min_{\theta} \max_{||\epsilon||_2 \leq \alpha_{\text{up}}} \sum_{i=1}^{M} \mathcal{L}(\theta_m^*(\theta+\epsilon); \mathcal{D}'_m) \quad \text{(upper)}$$

$$\text{s.t. } \theta_m^*(\theta) = \arg\min_{\theta_m} \max_{||\epsilon_m||_2 \leq \alpha_{\text{low}}} \mathcal{L}(\theta_m+\epsilon_m; \mathcal{D}_m) + \frac{\|\theta_m - \theta\|^2}{2\beta_{\text{low}}}, \; m = 1, ..., M. \quad \text{(lower)}$$

# Sharp-MAML: Strong Empirical Performance

| ALGORITHMS | ACCURACY | TIME† |
|---|---|---|
| MAML (REPRODUCED) | 47.60% | x1 |
| SHARP-MAML$_{\text{low}}$ | 57.67% | x1.45 |
| SHARP-MAML$_{\text{up}}$ | 56.33% | x1.90 |
| SHARP-MAML$_{\text{both}}$ | 60.67% | x2.80 |
| SHARP-MAML$_{\text{low}}$ -ANIL | 57.33% | x1.19 |
| ESHARP-MAML$_{\text{low}}$ | 54.33% | x1.23 |
| ESHARP-MAML$_{\text{low}}$ -ANIL | 56.33% | x1.08 |

**Improved generalization + Computationally efficient**

- All Sharp-MAML variants *improve the generalization performance of MAML*
- Our computationally efficient versions, leveraging ESAM [Du et al., 2021] and ANIL [Raghu et al., 2019], *reduce computation significantly*

# Sharp-MAML: Optimization and Generalization Analysis

## Optimization Analysis:

Under some reasonable assumptions:

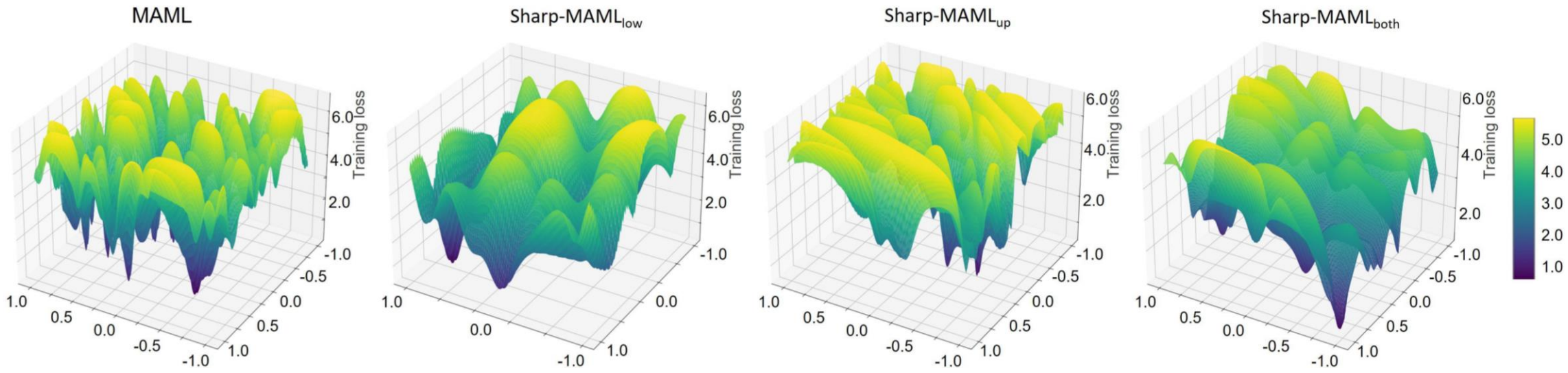
All variants of Sharp-MAML match the convergence rate of MAML

## Generalization Analysis:

Under some reasonable assumptions:

Sharp-MAML has smaller upper bound of generalization error than that of conventional MAML

# MAML vs Sharp-MAML: Loss Landscapes




**Sharp-MAML indeed seeks out landscapes that are smoother as compared to the landscape of original MAML**

# Concluding Remarks

- Nonconvex nested problems like MAML generalize better when used with generalization promoting SAM
- Sharp-MAML matches the convergence rate of MAML
- Sharp-MAML has smaller generalization error upper bound than MAML

Code: https://github.com/mominabbass/Sharp-MAML

# References


Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proc. International Conference on Machine Learning, Sydney, Australia, 2017.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. International Conference on Learning Representations, 2017

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In Proc. International Conference on Learning Representations, Virtual, April 2021.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. F. Efficient sharpness-aware minimization for improved training of neural networks. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Virtual, June 2021.

Raghu, A., Raghu, M., and Bengio, S. Rapid learning or feature reuse? towards understanding the effectiveness of maml. In Proc. International Conference on Learning Representations, Virtual, April 2020.