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

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

Rensselaer Polytechnic Institute IBM Research

Context and Motivation

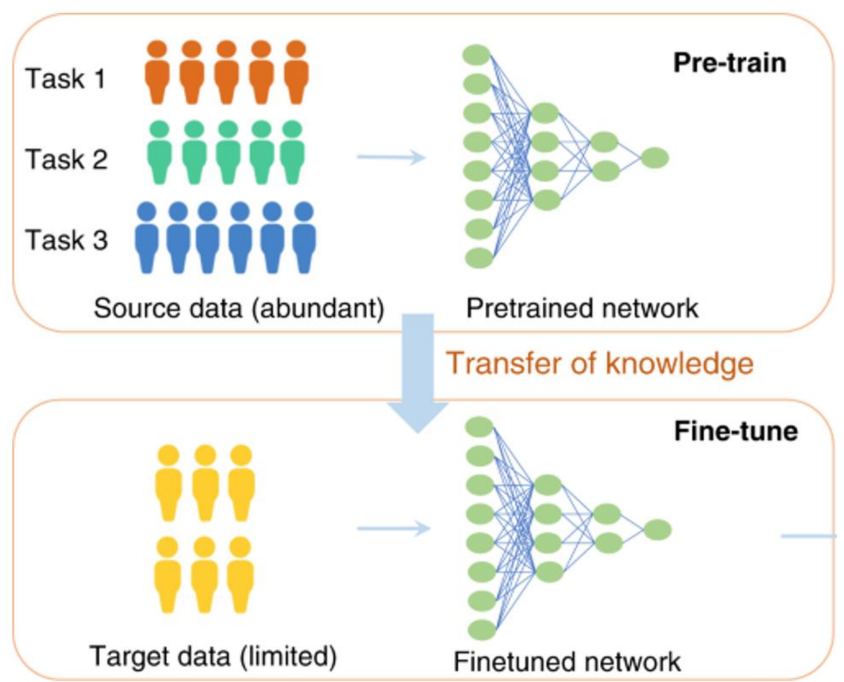


Figure source: https://www.nature.com/articles/s41416-021-01358-1

Meta learning

Model initialization, shared representation

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

$\theta_m^* \in \operatorname{argmin}_{\theta_m}$ fine-tuning θ

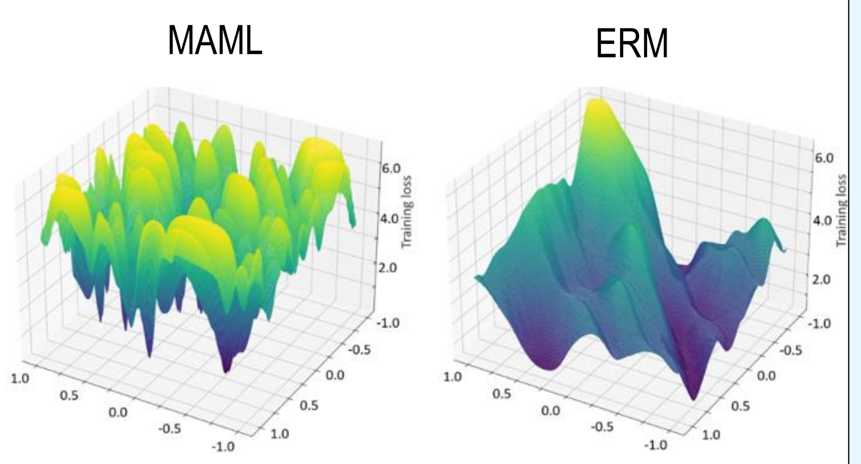
- ❑ Meta Learning is a challenging nested problem

Problem Formulation and Loss Landscape

MAML [Finn et al., 2017] problem

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

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



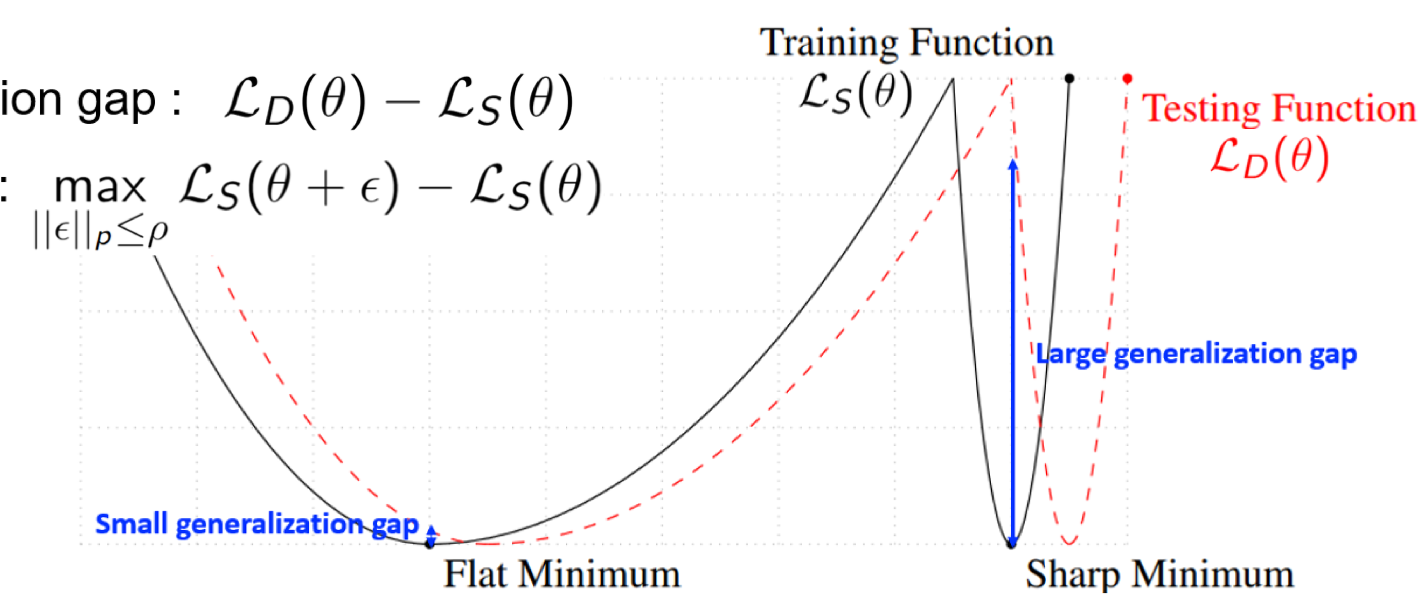
Lemma 1 (informal): MAML has more stationary points and local minimizers than ERM (i.e., a more complex loss landscape)

Generalization and Sharpness of Solutions

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

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

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



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

Sharp-MAML Algorithm

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

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

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

Algorithm:

for $t = 1, \cdots, T$ **do**

  for all tasks do

    Sample K examples from $\mathcal{D}_m$

    Evaluate stochastic gradient $\tilde{\nabla}\mathcal{L}(\theta^t; \mathcal{D}_m)$

    Compute perturbation $\epsilon_m(\theta^t) = \alpha_{\text{low}} \tilde{\nabla}\mathcal{L}(\theta^t; \mathcal{D}_m) / \|\tilde{\nabla}\mathcal{L}(\theta^t; \mathcal{D}_m)\|_2$

    Compute fine-tuned parameter $\tilde{\theta}_m^1(\theta^t) = \theta^t - \beta_{\text{low}} \nabla_{\theta^t}\mathcal{L}(\theta^t + \epsilon_m(\theta^t); \mathcal{D}_m)$

    Sample data from $\mathcal{D}'_m$ for meta-update

  Compute $\nabla_{\theta^t} \sum_{m=1}^{M} \mathcal{L}(\tilde{\theta}_m^1(\theta^t); \mathcal{D}'_m)$

  Compute perturbation $\epsilon(\theta^t) = \alpha_{\text{up}} \nabla_{\theta^t} \sum_{m=1}^{M} \mathcal{L}(\tilde{\theta}_m^1(\theta^t); \mathcal{D}'_m) / \|\nabla_{\theta^t} \sum_{m=1}^{M} \mathcal{L}(\tilde{\theta}_m^1(\theta^t); \mathcal{D}'_m)\|_2$

  Compute $\tilde{\theta}_m^2(\theta^t) = \theta^t + \epsilon(\theta^t) - \beta_{\text{low}} \nabla_{\theta^t}\mathcal{L}(\theta^t + \epsilon(\theta^t) + \epsilon_m(\theta^t); \mathcal{D}_m)$

  Update $\theta^{t+1} = \theta^t - \beta_{\text{up}} \nabla_{\theta^t} \sum_{m=1}^{M} \mathcal{L}(\tilde{\theta}_m^2(\theta^t); \mathcal{D}'_m)$

Optimization Analysis

Assumption (Informal): Assume $F(\theta)$ is Lipschitz continuous and smooth. Assume we can obtain unbiased estimators of $\nabla\mathcal{L}(\theta; \mathcal{D}_m), \nabla^2\mathcal{L}(\theta; \mathcal{D}_m), \nabla\mathcal{L}(\theta; \mathcal{D}'_m)$ and their variances are bounded.

Main theorem. If we choose stepsizes and perturbation radii

$$\beta_{\text{low}}, \beta_{\text{up}}, \alpha_{\text{up}} = \mathcal{O}\left(1/\sqrt{T}\right), \alpha_{\text{low}} = \mathcal{O}(1)$$

with some proper constants, we can get that the iterates generated by Sharp-MAML$_{\text{up}}$, Sharp-MAML$_{\text{low}}$ and Sharp-MAML$_{\text{both}}$ satisfy

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla F(\theta^t)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

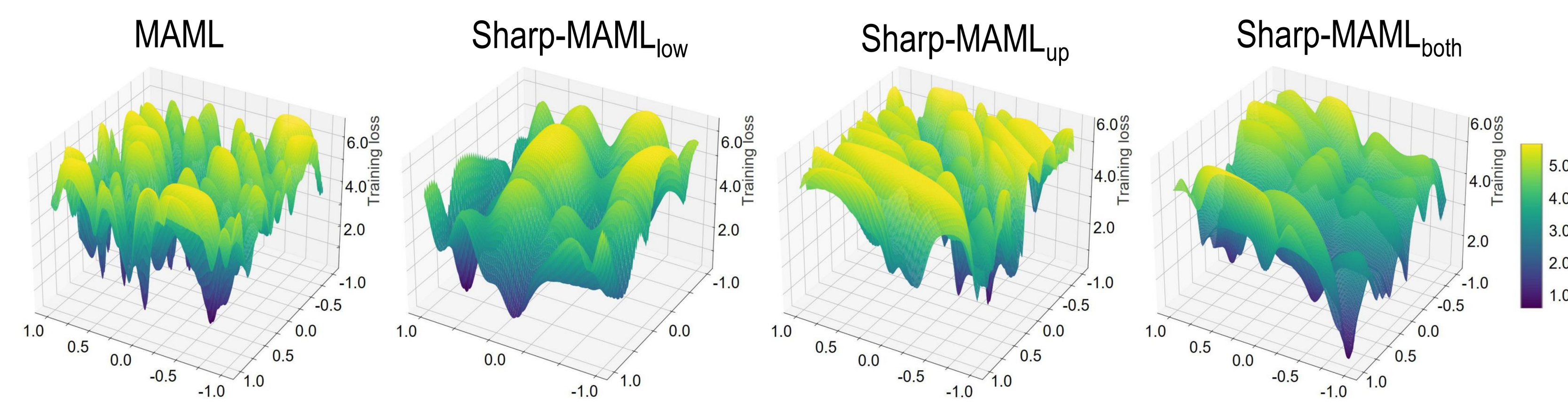
- ❑ Sharp-MAML matches the convergence rate of MAML

Empirical Comparison with MAML

Algorithms	Accuracy
Matching Nets	43.56%
MAML (reproduced)	47.13%
Sharp-MAML$_{\text{low}}$	48.87%
Sharp-MAML$_{\text{up}}$	49.03%
Sharp-MAML$_{\text{both}}$	49.60%

- ❑ Setting: 5-way 1-shot, conv-4-64 model, miniimagenet dataset
- ❑ All Sharp-MAML variants improve the generalization performance of MAML

Sharp-MAML Improves the Landscape of MAML



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

Generalization Analysis

Assumption (Informal): Assume $0 \le \mathcal{L}(\theta_m; \mathcal{D}) \le 1, \mathcal{D} \sim \mathcal{P}, |\mathcal{D}| = nM$.

Define $F(\theta; \mathcal{P}) = \mathbb{E}_{\mathcal{D}\sim\mathcal{P}}[F(\theta; \mathcal{D})]$. Let $\hat{\theta}$ be the stationary point of Sharp-MAML, satisfying $F(\hat{\theta}; \mathcal{P}) \le \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\alpha^2\mathbf{I})}\left[F(\hat{\theta}+\epsilon; \mathcal{P})\right]$.

Main theorem. If the lower-level algorithm used in Sharp-MAML is γ_A-uniformly stable, with high probability, the risk of Sharp-MAML$_{\text{up}}$ satisfies

$$F(\hat{\theta}; \mathcal{P}) \le \max_{\|\epsilon\|_2 \le \alpha} F(\hat{\theta}+\epsilon; \mathcal{D}) + \gamma_A + \tilde{\mathcal{O}}\left(\frac{1}{4nM}\right)$$

where $\tilde{\mathcal{O}}(\cdot)$ hides the polylogarithmic factor.

- ❑ Sharp-MAML has smaller generalization error upper bound than MAML

NSF MoDL-SCALE Grant 2134168 and the Rensselaer-IBM AI Research Collaboration