When MAML Can Adapt Fast And How to Assist When It Cannot

Séb Arnold — seb.arnold@usc.edu **Shariq Iqbal** — shariqiq@usc.edu **Fei Sha** — fsha@google.com

More information: sebarnold.net/projects/kfo

Summary

We take a closer look at Model Agnostic Meta-Learning (MAML) and show that it requires depth — shallow models fail because they lack parameters to shape the gradients during fast adaptation.

- Surprisingly, MAML fails to adapt on very simple tasks even with a model expressive enough to solve them perfectly; but, an over-parameterized model succeeds.
- Our analysis shows that this is because upper layers meta-learn update functions for the bottom layers.
- We propose three solutions to combat this issue:
    1. Using deeper non-linear models,
    2. Adding extra linear (collapsable) layers at the end of the model,
    3. Training with KFO (Kronecker-Factored Optimizer), a new meta-optimizer which scales to large deep networks.
- Empirically, we compare all three approaches and conclude that adding linear layers is a simple solution that almost matches meta-optimizers, while also enabling control of the model size post-adaptation.

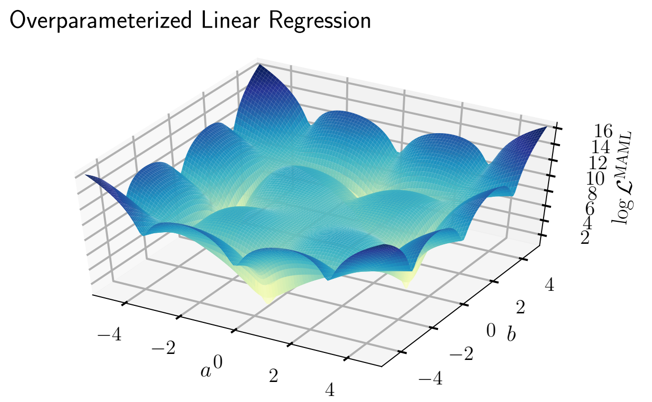
MAML: Model Agnostic Meta-Learning

The MAML [1] objective is simply expressed as:

$$\min_{\theta} \mathbb{E}_{\tau}\left[\mathcal{L}_{\tau}(\theta - \alpha \nabla_{\theta} \mathcal{L}_{\tau}(\theta))\right]$$

where:

- $\theta \triangleq$ the parameters to be learned,
- $\tau \triangleq$ a task index,
- $\mathcal{L}_{\mathcal{T}} \triangleq$ the loss associated with a task.



Intuition: MAML tries to meta-learn parameters that can be quickly adapt to any task from your training distribution.

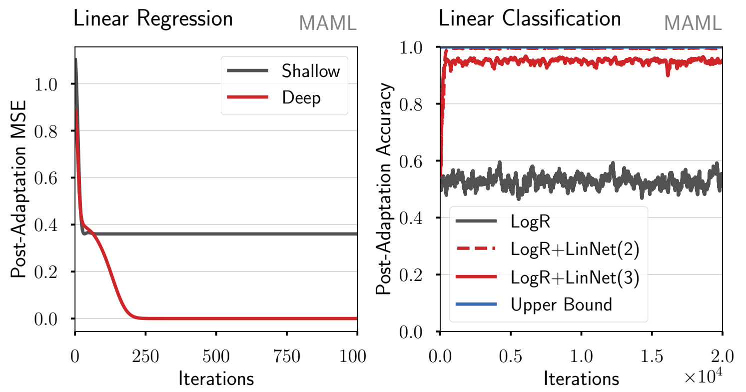
References

1. "Model Agnostic Meta-Learning", Finn et al., ICML 2017.
2. "Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML", Raghu et al., ICLR 2020.

Failure Mode

MAML fails to meta-learn with shallow models, even though they have sufficient capacity to solve all tasks.

However, meta-learning succeeds when over-parameterizing the models (with linear layers) without changing their original capacity.



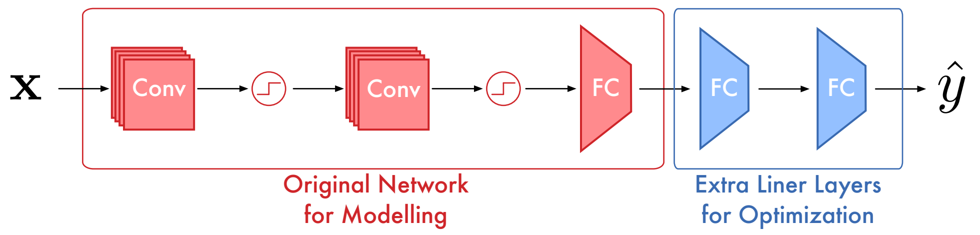
Why?

Insights

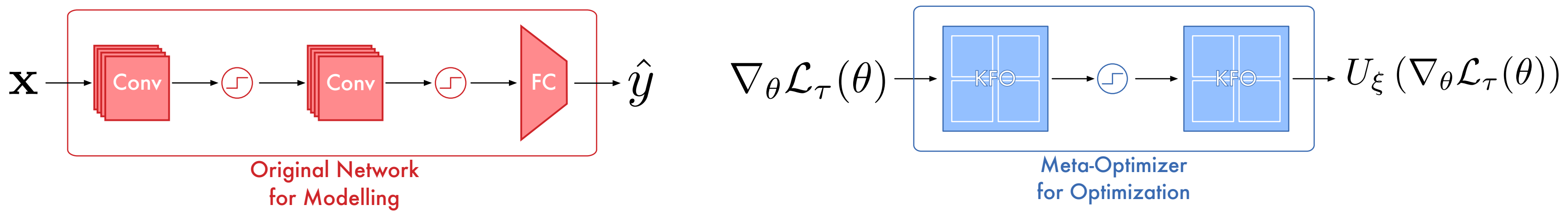
- Theoretical analysis on 1D shallow and deep models shows that:
    - deep models are required for meta-learning, because
    - the upper layers of the model facilitate (meta-)optimization.
- We can interpret those upper layers as "meta-optimizers that work from the inside" as they learn to modify the adaptation gradient of lower layers.
- We empirically verify this theory on linear & logistic regression, and with deep network architectures.

Solutions

- Use larger deeper models: current go-to solution, undesirable in compute-limited environments.
- Add extra linear layers on top of the mode: simple, universal, works decently but incurs small performance penalty.

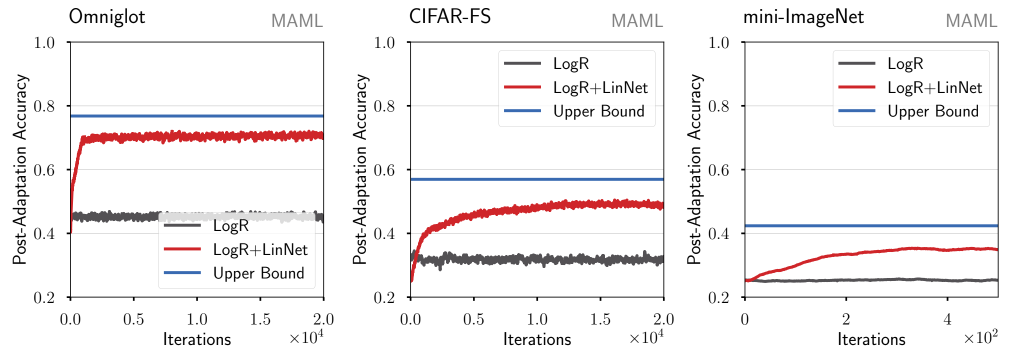


- Move optimization parameters to a KFO meta-optimizer: best performance, lightweight post-adaptation, but expensive during meta-training.



Empirical Results

Extra linear layers improve shallow meta-learning.



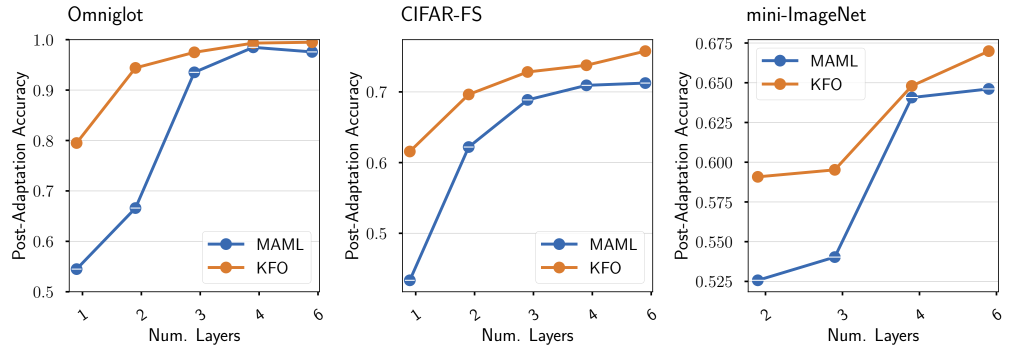
Extra linear layers improve deep meta-learning.

Method	MAML				MAML w/ LinNet			
CNN Layers	2	3	4	6	2	3	4	6
Omniglot	66.8	93.5	98.5	97.6	88.1	95.5	98.1	97.6
CIFAR-FS	62.2	68.9	70.9	71.3	66.1	71.1	74.4	71.9
mini-ImageNet	52.6	54.0	64.1	64.6	60.5	60.2	64.9	64.1

Meta-optimizers outperform MAML on 2-layer CNNs.

Dataset	MAML	MAML w/			
		MSGD	MC	T-Nets	META-KFO
Omniglot	66.6	74.07	94.63	92.27	**96.62**
CIFAR-FS	62.2	62.82	68.37	66.42	**69.64**
mini-ImageNet	52.6	**59.90**	58.95	58.47	59.08

Meta-optimizers are most effective with shallower models.



See our paper for more details, including:

- Theoretical analysis of 1D linear and logistic regression.
- Combining ANIL [2] with Meta-Optimizers.
- Why collapsing extra linear networks fails.