Unraveling Model-Agnostic Meta-Learning via The Adaptation Learning Rate

Yingtian Zou[†], Fusheng Liu[‡], Qianxiao Li[§]

[†] School of Computing,
[‡] Institute of Data Science,
[§] Department of Mathematics,

National University of Singapore

Introduction

From ERM to MAML

As is known to us, Empirical Risk Minimization (ERM) gives us a direct minimization algorithm over task distribution $\mathcal{D}(T)$. The multi-task problem can be solved by

$$\text{ERM} \Rightarrow \min_{\boldsymbol{w} \in \mathbb{R}^d} \mathbb{E}_{T \sim \mathcal{D}(T)} \left[\ell \left(\boldsymbol{w}, T \right) \right]$$

and Model-Agnostic Meta-Learning (Finn et al. [1])

$$\text{MAML} \Rightarrow \min_{\boldsymbol{w} \in \mathbb{R}^d} \mathbb{E}_{T \sim \mathcal{D}(T)} \left[\ell \left(\boldsymbol{w} - \alpha \nabla_w \ell(\boldsymbol{w}, T_{train}), T_{test} \right) \right]$$

Critical role of adaptation step size

As is known to us, Empirical Risk Minimization (ERM) gives us a direct minimization algorithm over task distribution $\mathcal{D}(T)$. The multi-task problem can be solved by

$$\text{ERM} \Rightarrow \min_{\boldsymbol{w} \in \mathbb{R}^d} \mathbb{E}_{T \sim \mathcal{D}(T)} \left[\ell\left(\boldsymbol{w} - 0 \cdot \nabla_w \ell(\boldsymbol{w}, T_{train}), T_{test}\right)\right]$$

and Model-Agnostic Meta-Learning (Finn et al. [1])

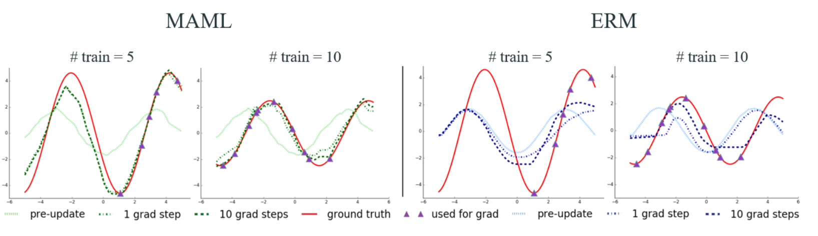
$$\text{MAML} \Rightarrow \min_{\boldsymbol{w} \in \mathbb{R}^d} \mathbb{E}_{T \sim \mathcal{D}(T)} [\ell\ (\underbrace{\boldsymbol{w} - \alpha \nabla_w \ell(\boldsymbol{w}, T_{train})}_{\text{inner loop}}, T_{test})]$$

α : adaptation learning rate/step size

MAML demonstrates the superiority on ...

Why we need meta-learning scheme? What's the superiority of Inner-Outer loops optimization (Adaptation)?

Because,



with the inner loop adaptation, desired model can quickly adapt to the new task.

Mixed linear regression setting

To clearly understand the mechanism, let's simplify it with a linear setting. Assume each task is to fit a linear function $\boldsymbol{y} = f_\Phi(X)$

- $X \in \mathbb{R}^{K \times d'}, \boldsymbol{y} \in \mathbb{R}^d$ are randomly sampled from data and label distributions respectively. Dataset X has independent rows.
- $f_\Phi(X) = \Phi(X)\boldsymbol{a}$ contains a (random) feature transformation $\Phi : \mathbb{R}^{d'} \mapsto \mathbb{R}^d$, and task optimum $\boldsymbol{a} \in \mathbb{R}^d$.

A model is required to be learned such that with one-step adaptation on each task, gets as minimum error as possible.

Objective functions under mixed linear setting

To this end, ERM has the objective function

$$\mathcal{L}_r(\boldsymbol{w}, K) = \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\mathcal{D}(\boldsymbol{a})} \mathop{\mathbb{E}}_{X\sim\mathcal{D}(\boldsymbol{x})} \frac{1}{K}\left\| \Phi(X)\boldsymbol{w} - \Phi(X)\boldsymbol{a} \right\|_2^2$$

while MAML optimizes the following objective

$$\begin{aligned} \mathcal{L}_m(\boldsymbol{w}, \alpha, K) &= \mathop{\mathbb{E}}_{\boldsymbol{a}\sim\mathcal{D}(\boldsymbol{a})} \mathop{\mathbb{E}}_{X\sim\mathcal{D}(\boldsymbol{x})} \frac{1}{K}\left\| \Phi(X)\boldsymbol{w}'(\alpha) - \Phi(X)\boldsymbol{a} \right\|_2^2 \\ \text{s.t. } \boldsymbol{w}'(\alpha) &= \left[\boldsymbol{w} - \frac{2\alpha}{K}\Phi(X)^\top \left(\Phi(X)\boldsymbol{w} - \Phi(X)\boldsymbol{a}\right)\right] \end{aligned}$$

Solutions of desired model

We sample N tasks to train the model with different algorithms, ERM and MAML. Empirical objective functions of these two algorithms yield different solutions

$$\boldsymbol{w}_{ERM} = \left(\sum_{i \in [N]} \Phi(X_i)^\top \Phi(X_i) \right)^{-1} \left(\sum_{j \in [N]} \Phi(X_j)^\top \Phi(X_j) \boldsymbol{a}_j \right)$$

$$\boldsymbol{w}_{MAML}(\alpha) = \left(\sum_{i \in [N]} C_i(\alpha)^\top C_i(\alpha) \right)^{-1} \left(\sum_{j \in [N]} C_j(\alpha)^\top C_j(\alpha) \boldsymbol{a}_i \right)$$

where

$$C_i(\alpha) = \Phi(X_i) \left[I - \frac{2\alpha}{K} \Phi(X_i)^\top \Phi(X_i) \right], C_i(\alpha) \in \mathbb{R}^{K \times d}$$

can be viewed as *adapted feature* of task i.

Asymptotics of α

Since the "adapted features" of MAML depends on α, then we can get some intuition from the asymptotics of α.

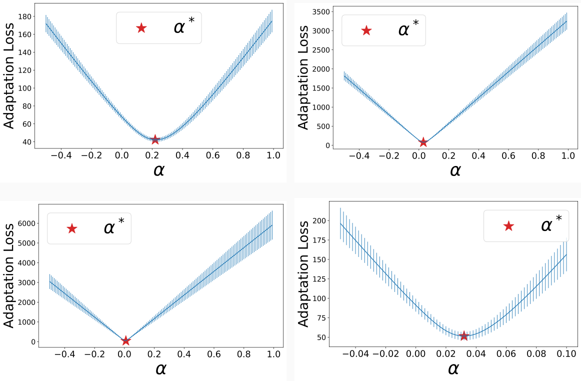
$$\text{Adapted feature:} \quad C(\alpha) = \Phi(X)\left[I - \frac{2\alpha}{K}\Phi(X)^\top \Phi(X)\right]$$

- $\alpha \to 0$, $\boldsymbol{w}_{MAML} \to \boldsymbol{w}_{ERM}$. Solution learned by MAML will be close to the one learned by ERM.
- $\alpha \to \infty$, $\|C(\alpha)\| \to \infty$. If α becomes large, the norm of adapted features will explode.
- What's the best $C(\alpha), \alpha \in [0, \infty)$ for MAML?

Optimal choice of α

What is the optimal choice?

Def. [Adaptation Loss] $L \triangleq \frac{1}{NK} \sum_{i=1}^{N} \| C_i(\alpha) \boldsymbol{w}_{MAML}(\alpha) - C_i(\alpha) \boldsymbol{a}_i \|_2^2$



A model is required to be learned such that with one-step adaptation step size $= \alpha^*$ on all N tasks, gets as minimum L as possible.

Step-by-step to find the α^*

- Step 1: Plug the global minimum $\boldsymbol{w}_{MAML}$ we solved before into the expected objective $\mathcal{L}_m(\cdot, \alpha, K)$.

Step-by-step to find the α^*

- Step 1: Plug the global minimum $\boldsymbol{w}_{MAML}$ we solved before into the expected objective $\mathcal{L}_m(\cdot, \alpha, K)$.

Note that the sampled N tasks leads to the randomness of $\boldsymbol{w}_{MAML}$. In order to get a deterministic value of α^*, we may take the expectation over $\boldsymbol{w}_{MAML}$ and try to minimize $\mathbb{E}_{\boldsymbol{w}_{MAML}}\mathcal{L}_m(\boldsymbol{w}_{MAML}, \alpha, K)$

$$\alpha^*(N, K) = \arg\min \mathbb{E}_{\boldsymbol{w}_{MAML}}\mathcal{L}_m(\boldsymbol{w}_{MAML}, \alpha, K)$$

$\Rightarrow$ shall be deemed as the optimal choice for the average case!

Step-by-step to find the α^*

- **Step 2:** Get the estimation of $\alpha^*(N, K)$.

Theorem (Optimal adaptation learning rate)

Under assumptions 1 and 2, we have as $N \to \infty$, $\alpha^*(N, K) \to \alpha^*_{lim}(K)$, *where*

$$\textbf{\textit{Estimator}} \quad \alpha^*_{lim}(K) = \frac{K \operatorname{tr}[\mathbb{E}_X[(\Phi(X)^\top \Phi(X))^2]]}{2 \operatorname{tr}[\mathbb{E}_X[(\Phi(X)^\top \Phi(X))^3]]}$$

$\Phi(X) \in \mathbb{R}^{K \times d}$, K *: sample size per task,* N *: number of tasks.*

Assumptions

- 1. $\mathbb{E}_{\boldsymbol{a} \sim \mathcal{D}(\boldsymbol{a})}[\boldsymbol{a}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{a}] = \sigma_a^2$ (Normalization).
- 2. With probability 1, $\Phi(X)^\top \Phi(X)$ has uniformly bounded eigenvalues between positive constants.

Step-by-step to find the α^*

- Step 3: Assess our estimation α^*_{lim} for α^*.

As number of tasks N becomes large, the estimation error between $\alpha^*_{lim}(K)$ and $\alpha^*(N, K)$ and the error in concentration will both be guaranteed.

Proposition (Concentration)

Under assumption 1 & 2, with probability $1 - \delta$, we have

$$\left| \mathcal{L}_m(\boldsymbol{w}_{MAML}, \alpha, K) - \mathbb{E}_{\boldsymbol{w}_{MAML}} \mathcal{L}_m(\boldsymbol{w}_{MAML}, \alpha, K) \right| \leq \frac{2L^2 \varepsilon_\alpha}{NK \log \delta}$$

where ε_α converges to some constants along α changing.

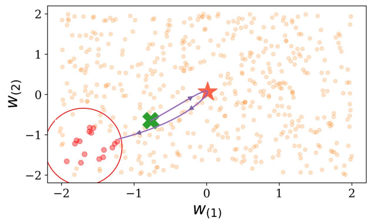
Role of α

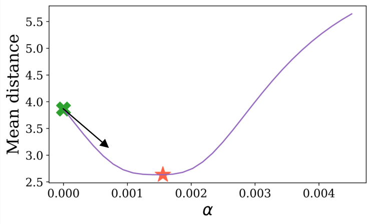
A simple experiment

If we create a dense task region (like increasing the weights of tasks, adding new tasks to the region etc.), what would happen to the global minima learned by different α?

A simple experiment

If we create a dense task region (like increasing the weights of tasks, adding new tasks to the region etc.), what would happen to the global minima learned by different α?





$\times$: position of $\boldsymbol{w}_{ERM} = \boldsymbol{w}_{MAML}(0)$, $\star$: position of $\boldsymbol{w}_{MAML}(\alpha^*)$, Curve : the trajectory.

Towards interpreting role of α statistically

To better understand the role of α, let's simplify the expression of α^*_{lim} we got from the previous theorem.

Corollary (statistical dependency)
With a feature mapping $\phi : \mathbb{R}^{d_x} \to \mathbb{R}^d$ *for each data* $\boldsymbol{x} \in \mathbb{R}^{d_x}$*, then*

$$\alpha^*_{lim} \in \left[\frac{2}{d}, \frac{d}{2}\right] \sigma^2(\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_K)), d \geq 2$$

Towards interpreting role of α statistically

To better understand the role of α, let's simplify the expression of α^*_{lim} we got from the previous theorem.

Corollary (statistical dependency)
With a feature mapping $\phi : \mathbb{R}^{d_x} \to \mathbb{R}^d$ *for each data* $\boldsymbol{x} \in \mathbb{R}^{d_x}$*, then*

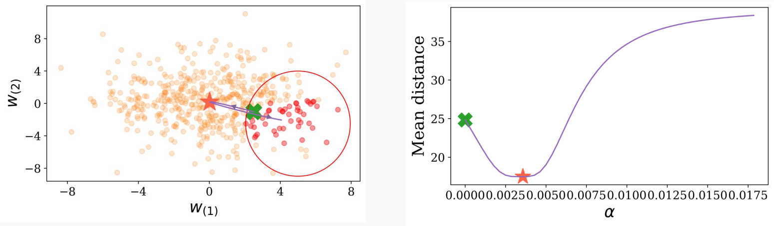
$$\alpha^*_{lim} \in \left[\frac{2}{d}, \frac{d}{2}\right] \sigma^2(\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_K)), d \geq 2$$

Example
Consider polynomial feature, $\phi(x) = (1, ..., x^{d-1})$, $x_1, ..., x_K \sim \mathcal{N}(0, \sigma)$, task optimum is a random vector from zero-mean distribution, then

$$\alpha^*_{lim} = \frac{\textsc{Poly}(\sigma^4)}{\textsc{Poly}(\sigma^6)} \to \frac{1}{\sigma^2} \text{ as } \sigma \to \infty$$

⋆ $\boldsymbol{w}_{MAML}(\alpha^*)$ chooses step size through inverse of data variance.

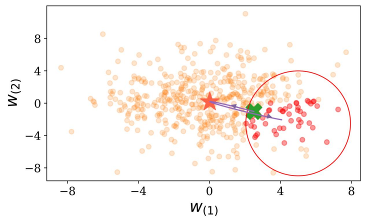
Towards interpreting role of α geometrically

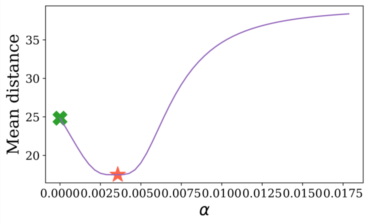


More evidence on shorter average distance of $\boldsymbol{w}_{MAML}$ to task optimum compared to $\boldsymbol{w}_{ERM}$, for $\alpha \in [0, \delta]$.

From the right figure we can see, solution learned by MAML leads to shorter solution mean distance than ERM as α increasing.

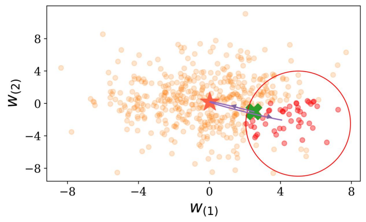
Towards interpreting role of α geometrically

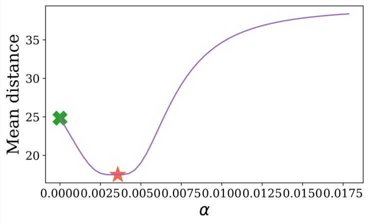




This phenomenon has also been explored in some papers Nichol et al. [3], Zhou et al. [4]. Or algorithm design based on the intuition, e.g. weight sharing across similar tasks.

Towards interpreting role of α geometrically





This phenomenon has also been explored in some papers Nichol et al. [3], Zhou et al. [4]. Or algorithm design based on the intuition, e.g. weight sharing across similar tasks.

So can we prove that the fast adaptability of MAML benefits from a shorter solution distance?

Towards interpreting role of α geometrically

Definition (Adaptation Distance)
Solution $\boldsymbol{w}^0_{\mathcal{A}}$ learned by algorithm $\mathcal{A}$, the average distance under t-step ($t \geq 0$) fast adaptation is

$$\mathcal{F}_t(\boldsymbol{w}^0_{\mathcal{A}}) \triangleq \mathop{\mathbb{E}}_{T \sim \mathcal{D}(T)} \|\boldsymbol{w}^t_{\mathcal{A},T} - \boldsymbol{a}_T\|^2, \ \boldsymbol{w}^t_{\mathcal{A},T} : \text{adapted param of task } T$$

Towards interpreting role of α geometrically

Definition (Adaptation Distance)
Solution $\boldsymbol{w}^0_{\mathcal{A}}$ learned by algorithm $\mathcal{A}$, the average distance under t-step ($t \geq 0$) fast adaptation is

$$\mathcal{F}_t(\boldsymbol{w}^0_{\mathcal{A}}) \triangleq \mathop{\mathbb{E}}_{T \sim \mathcal{D}(T)} \|\boldsymbol{w}^t_{\mathcal{A},T} - \boldsymbol{a}_T\|^2,\ \ \boldsymbol{w}^t_{\mathcal{A},T} : \text{adapted param of task } T$$

Theorem (Solution distance)

Under assumptions 1 and 2, $\exists p > 0, 0 < q < 1$ for any $\alpha \in [0, \delta]$ at number of step t, we have

$$\mathbb{E}_{T_1,\ldots,T_N}\left[\mathcal{F}_t(\boldsymbol{w}_{ERM}) - \mathcal{F}_t(\boldsymbol{w}_{MAML}(\alpha))\right] \geq \frac{\alpha p q^{2t}}{NK}$$

Conclusion

- As the step size of adaptation, α plays a central role in MAML. Here, we show a principled way to select the optimal value.
- From the statistical perspective, the optimal value α^* has relation to the inverse of data variance.
- From the geometric perspective, global minimum learned by MAML minimizes the solution distance in expectation.

Experiments

Estimation of optimal α

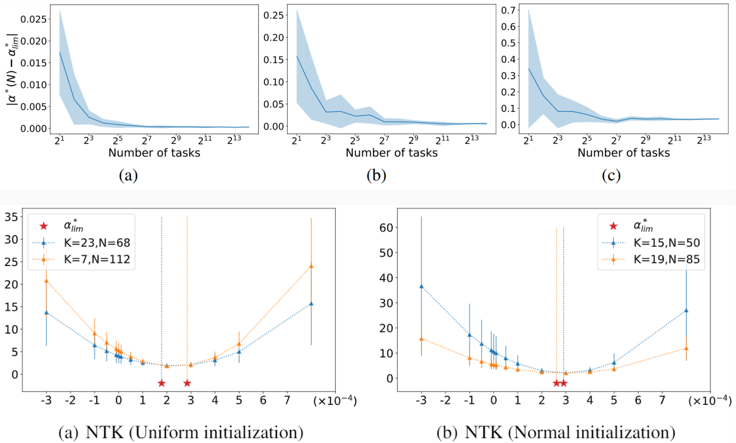


Figure 1: Upper row: estimation error | our theorem - true α^*| under different basis functions. Lower row: Estimation with NTK Jacot et al. [2]

Relation of data variance and α^*

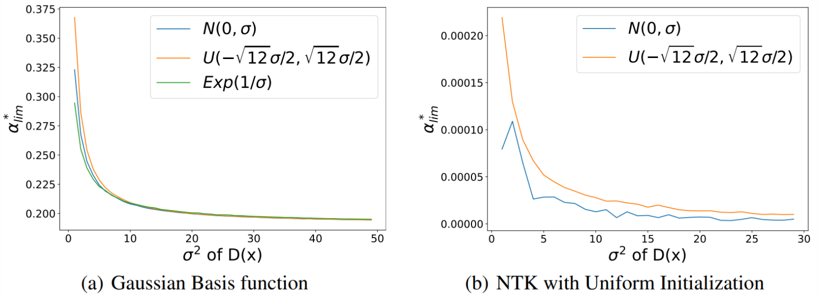
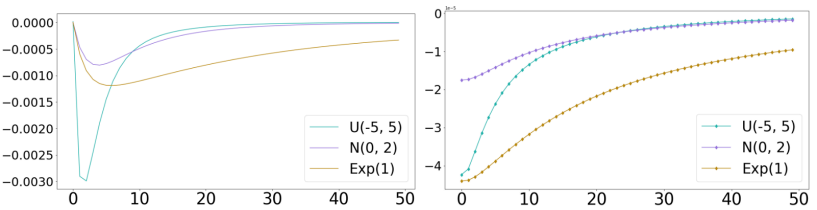


Figure 2: Value of α^* along the data variance σ^2.



(a) Adaptation Loss Gap (b) Fast Adaptation Distance Gap

Figure 3: (a) Loss difference $\mathcal{L}_{MAML} - \mathcal{L}_{ERM}$, (b) average solution distance gap $\mathcal{F}_t(\boldsymbol{w}_{MAML}(\alpha)) - \mathcal{F}_t(\boldsymbol{w}_{ERM})$.

References

[1] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[2] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.

[3] A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv*, 2018.

[4] P. Zhou, Y. Zou, X.-T. Yuan, J. Feng, C. Xiong, and S. Hoi. Task similarity aware meta learning: Theory-inspired improvement on maml. In *UAI*. PMLR, 2021.