Representation Learning with Model-Agnostic Meta-Learning (MAML)

Aryan Mokhtari
ECE Department, UT Austin

ITA Workshop 2022

May 24, 2022

Research Collaborators

Liam Collins

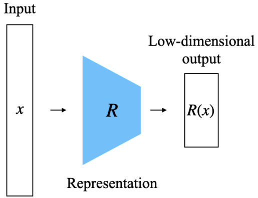
Sewoong Oh

Sanjay Shakkottai

L. Collins, A. Mokhtari, S. Oh, S. Shakkottai. "MAML and ANIL Provably Learn Representations", *ICML 2022*. [https://arxiv.org/abs/2202.03483]

Representation Learning

- A central goal of machine learning:
  $\Rightarrow$ **learn useful representations of data**.

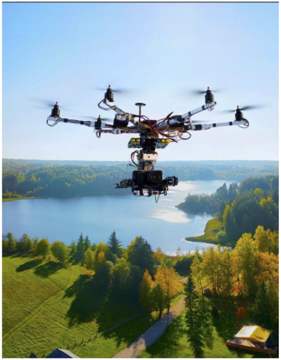


- Good representations are useful because they
  $\Rightarrow$ reduce data complexity
  $\Rightarrow$ enable generalizing models to *new tasks* quickly

- Quickly = with little (labeled) data and computation.

Representation Learning – Generalize to New Tasks

Training Data



Task in new Scenario

Image Credits: bit.ly/3i5m8ay, bit.ly/3w723ZY, bit.ly/3KHMQ5E, bit.ly/3i7pREJ, bit.ly/34I1ytT

- Training data in various scenarios; goal is to quickly adapt to the task in a new scenario

Meta-learning and Representation Learning

- *How* can we learn high-quality representations?
- *Meta-learning* methods have recently grown in popularity because they have yielded useful representations in practice.

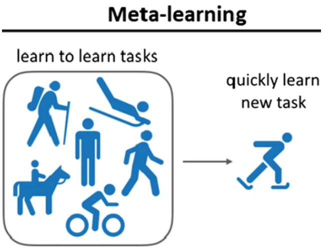


Image credit: https://meta-world.github.io, [HRJ21]

- Meta-learning leverages experience from learning a set of meta-training tasks to quickly solve new tasks.
- A popular approach is Model Agnostic Meta-Learning (MAML)

MAML and ERM

Traditional Supervised Learning, Empirical Risk Minimization (ERM)

- Set of tasks: $\mathcal{T} = \{\mathcal{T}_i\}_{i=1}^{i=n}$ coming from distribution p
- Select a model $\boldsymbol{\theta}^*_{train}$
- A new task $\mathcal{T}_{test}$ is revealed, drawn according to dist. p
- Performance: $f_{test}(\boldsymbol{\theta}^*_{train})$

Formally, the training objective is:

- $\min_{\boldsymbol{\theta}} \mathbb{E}_{i\sim p}[f_i(\boldsymbol{\theta})]$
- $\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta})$

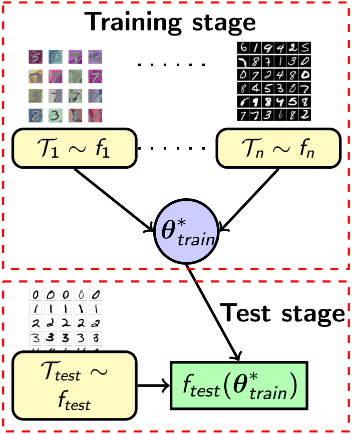


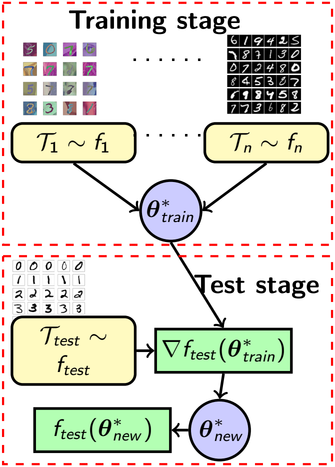
Image Credits: https://bit.ly/392pda9,

https://bit.ly/3EEIElq

MAML and ERM

Model-Agnostic Meta Learning [FAL17]

- What if we have budget to slightly update our model at test time?
- $\mathcal{T} = \{\mathcal{T}_i\}_{i=1}^{i=n}$ drawn from distribution p
- Select a model $\boldsymbol{\theta}^*_{train}$
- $\mathcal{T}_{test}$ is revealed, drawn based on p
- A few labeled samples of $\mathcal{T}_{test}$ given
- Performance: $f_{test}(\boldsymbol{\theta}^*_{train} - \alpha \nabla f_{test}(\boldsymbol{\theta}^*_{train}))$



Formally, the training objective is:

- $\min_{\boldsymbol{\theta}} \mathbb{E}_{i \sim p} \mathbb{E}_{(X_i, y_i) \sim D_i} [f_i(\boldsymbol{\theta} - \alpha \nabla f_i(\boldsymbol{\theta}; X_i, y_i))]$
- $\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{\theta} - \alpha \nabla f_i(\boldsymbol{\theta}; (X_i, y_i)))$

MAML Intuition: Adaptivity

- Empirical Risk Minimization (ERM): $\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta})$
- Gradient descent update for ERM.:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\beta}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{\theta}_t) \tag{1}$$

  Gradient evaluated at same $\boldsymbol{\theta}_t$ for all tasks $\implies$ **not adaptive**

- In contrast, for MAML the update is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\beta}{n}\sum_{i=1}^{n} (\boldsymbol{I} - \alpha\nabla^2 f_i(\boldsymbol{\theta}_t))\nabla f_i(\boldsymbol{\theta}_{t,i})$$

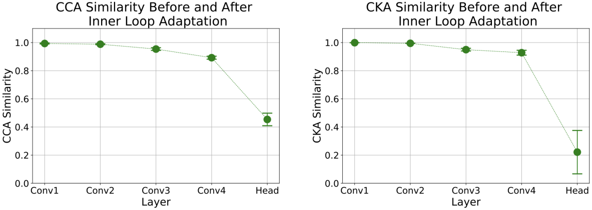
$$\text{where} \quad \boldsymbol{\theta}_{t,i} = \boldsymbol{\theta}_t - \alpha\nabla f_i(\boldsymbol{\theta}_t)$$

  $\boldsymbol{\theta}_{t,i}$ adapted to each task $\implies$ MAML finds an **adaptive** solution

- Seems like finding the right initialization for adaptation!

Empirical Observations of MAML

1. MAML learns models that can quickly solve new tasks [FAL17, AES19]
   - in image classification, sinusoid regression, reinforcement learning.
2. MAML seems to be learning a representation shared across tasks [RRBV20]
   - even though it is not designed for representation learning!



The representation learned by MAML does not change significantly when adapted to each task.
Figure credit: [RRBV20]

3. Can we formally prove this?

Representation learning in multi-task linear regression (1/2)

- Example: consider multi-task linear regression.
- Task i has ground-truth solution $\boldsymbol{\theta}_{*,i} \in \mathbb{R}^d$:

$$y_i \sim \boldsymbol{\theta}_{*,i}^\top \boldsymbol{x}_i + z_i$$

  - $\boldsymbol{x}_i$ is a random feature vector
  - $z_i \in \mathbb{R}$ is random, mean-zero noise

- Solving each task individually (i.e. finding a $\boldsymbol{\theta}_i \approx \boldsymbol{\theta}_{*,i}$ for each i) would require $\Omega(d)$ samples per task.

Can we do better using shared information across tasks?

Representation learning in multi-task linear regression (2/2)

- Now suppose the $\boldsymbol{\theta}_{*,i}$ lie in a shared k-dimensional subspace, $k \ll d$
- Let the columns of $\boldsymbol{B}_* \in \mathbb{R}^{d \times k}$ span this subspace, that is, for all tasks there exists $\boldsymbol{w}_{*,i} \in \mathbb{R}^k$ such that

$$\boldsymbol{\theta}_{*,i} = \boldsymbol{B}_* \boldsymbol{w}_{*,i}$$

  - $\boldsymbol{B}_*$ is the "ground-truth" representation

- If we know $\mathrm{col}(\boldsymbol{B}_*)$, we can solve new tasks with only $O(k)$ samples

Benefit of Representation Learning

$O(k)$ sample complexity much smaller than $\Omega(d)$

MAML for multi-task linear regression

- Loss function for task i at round t:

$$f_i(\boldsymbol{B}, \boldsymbol{w}) := \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_i, y_i}[(\langle \boldsymbol{B}\boldsymbol{w}, \boldsymbol{x}_i \rangle - y_i)^2]$$

- At each time t, we sample n distinct tasks (may differ across time)
- For each task, we collect $m_{in} + m_{out}$ data samples (approximates the expectation in the population loss in the inner and outer loops)

Algorithm (MAML)

- (Outer loop) For $t = 1, \ldots, T$:
  - Select n tasks satisfying diversity condition $(\text{span}(\{\boldsymbol{w}_{*,i}\}_{i\in[n]}) = \mathbb{R}^k)$
  - (Inner loop) For $i = 1, \ldots, n$:
    - Adapt: $\boldsymbol{w}_{t,i} = \boldsymbol{w}_t - \alpha\nabla_{\boldsymbol{w}} f_i(\boldsymbol{B}_t, \boldsymbol{w}_t)$, $\boldsymbol{B}_{t,i} = \boldsymbol{B}_t - \alpha\nabla_{\boldsymbol{B}} f_i(\boldsymbol{B}_t, \boldsymbol{w}_t)$
  - $\begin{bmatrix} \boldsymbol{w}_{t+1} \\ \bar{\boldsymbol{B}}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_t \\ \bar{\boldsymbol{B}}_t \end{bmatrix} - \frac{\beta}{n}\sum_{i=1}^n (\boldsymbol{I} - \alpha\nabla^2_{\boldsymbol{w},\bar{\boldsymbol{B}}} f_i(\boldsymbol{B}_t, \boldsymbol{w}_t)) \begin{bmatrix} \nabla_{\boldsymbol{w}} f_i(\boldsymbol{B}_{t,i}, \boldsymbol{w}_{t,i}) \\ \nabla_{\boldsymbol{B}} f_i(\boldsymbol{B}_{t,i}, \boldsymbol{w}_{t,i}) \end{bmatrix}$
    where $\bar{\boldsymbol{B}}$ is the column-wise vectorization of $\boldsymbol{B}$.

Our Main Results

- We consider the multi-task linear regression setting.

Main Results

- *Under standard assumptions, MAML (and variants) recover* $\mathrm{col}(\boldsymbol{B}_*)$ *exponentially fast when run on the task population losses.*
- *ANIL and FO-ANIL (simplified MAML variants) require* $m = O((\frac{d}{n} + 1)k^3) \ll d$ *samples per task to learn the ground-truth subspace.*
- *The key is that MAML and variants' adaptation of the head harnesses* ***task diversity*** *to improve the representation in all directions.*

- First results showing that MAML and variants provably learn effective representations.

Proof Intuition

- For FO-ANIL (a simplified version of MAML), we have

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t \underbrace{\left(\boldsymbol{I}_k - \frac{\beta}{n} \sum_{i=1}^{n} \boldsymbol{w}_{t,i} \boldsymbol{w}_{t,i}^\top \right)}_{\text{prior weight}} + \boldsymbol{B}_* \underbrace{\frac{\beta}{n} \sum_{i=1}^{n} \boldsymbol{w}_{*,t,i} \boldsymbol{w}_{t,i}^\top}_{\text{signal weight}}$$

- Suppose $\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{w}_{t,i} \boldsymbol{w}_{t,i}^\top$ is full rank (i.e., the $\boldsymbol{w}_{t,i}$'s are diverse), then:

Key observation

Prior weight reduces energy from $\boldsymbol{B}_t$, and signal weight boosts energy from $\boldsymbol{B}_*$ in all directions.
$\implies$ **Head adaptation** and **task diversity** are critical!

- The more diverse the adapted $\boldsymbol{w}_{t,i}$'s (i.e. smaller condition number of $\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{w}_{t,i} \boldsymbol{w}_{t,i}^\top$), the faster convergence rate.

Comparison to Empirical Risk Minimization (1/2)

$$\text{ERM:} \quad \min_{\boldsymbol{B},\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{B}, \boldsymbol{w})$$

- In this case we can show:

$$\boldsymbol{B}_{t+1} = \boldsymbol{B}_t \Big(\underbrace{\boldsymbol{I}_k - \beta \boldsymbol{w}_t \boldsymbol{w}_t^\top}_{\text{prior weight}} \Big) + \boldsymbol{B}_* \underbrace{\frac{\beta}{n} \sum_{i=1}^{n} \boldsymbol{w}_{*,t,i} \boldsymbol{w}_t^\top}_{\text{signal weight}}$$

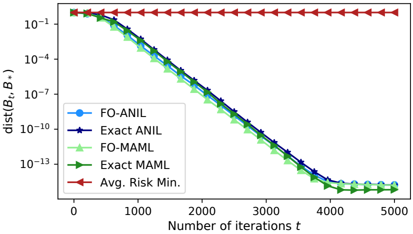
- The prior weight is rank $k-1$, while the signal weight is only rank 1.

Key observation

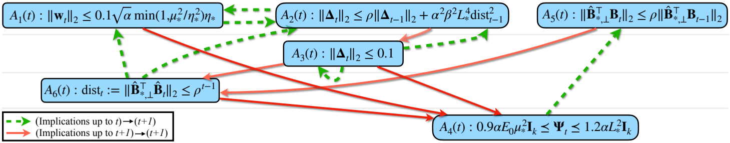
The representation can only move closer to $\mathrm{col}(\boldsymbol{B}_*)$ in **one** direction on each iteration $\rightarrow$ not clear that it can eventually reach $\mathrm{col}(\boldsymbol{B}_*)$.

Comparison to Empirical Risk Minimization (2/2)

- Empirically, ERM fails to learn col($\boldsymbol{B}_*$).



(2).pdf



(1).pdf

Inductive logic used in the proof for FO-ANIL.

Discussion

- We have obtained the first results showing that MAML and variants learn effective representations in any setting.
- Inner loop adaptation of the head is key to their ability to learn representations.
- Quantifies the benefits of diverse tasks in the training environment.
- Substantial sample complexity improvement can be achieve by learning representations

References

[FAL17] Chelsea Finn, Pieter Abbeel, Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Neural Networks, *International Conference on Machine Learning*, 2017.

[AES19] Antreas Antoniou, Harrison Edwards, Amos Storkey. How to Train Your MAML, *International Conference on Learning Representations*, 2019.

[RRBV20] Aniruddh Raghu, Maithra Raghu, Samy Bengio, Oriol Vinyals. Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML, *International Conference on Learning Representations*, 2020.

[HRJ21] Mike Huisman, Jan N. van Rijn, Aske Plaat. A Survey of deep Meta-Learning, *Artificial Intelligence Review* Volume 54, pages 4483âĂŞ4541, 2021.