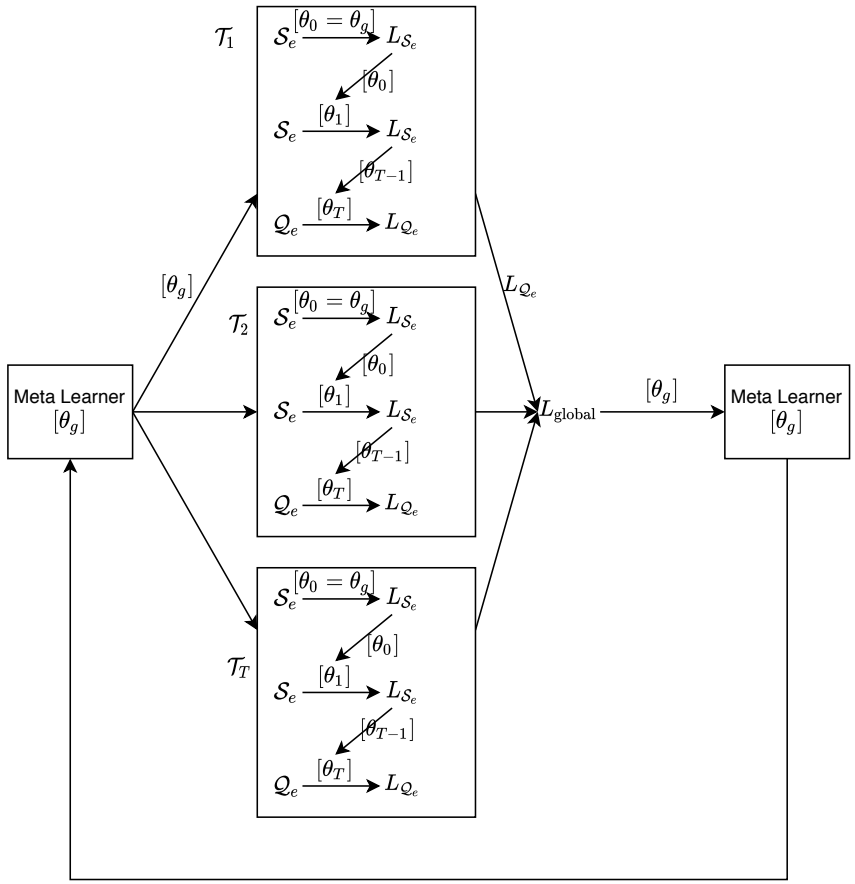
Few-Shot Learning

Outline

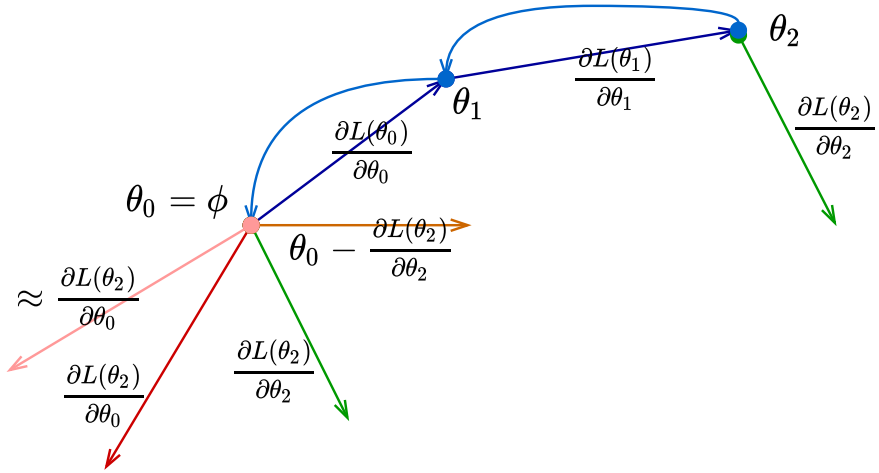
- Overview of meta-learning inner-loop and outer-loop updates
- Derivation of MAML's full outer-loop update
- Intuition of other model variants of outer-loop updates
  - Second-order and First-order MAML (Finn et al., ICML 2017)
  - Reptile (Nichol et al., 2018)
  - Implicit MAML (iMAML) (Rajeswaran et al., NeurIPS 2019)

Meta-Learning Inner and Outer Updates

- Meta-learning consists of 2 components
  - **Learner, inner-loop updates:** the learner tries to adapt the **global params** θ_g to **individual params** θ_e based on the loss of each individual task
  - **Meta-learner, outer-loop updates:**, the meta-learner tries to update the global params based on the loss of the individual task (then aggregate/average updates for multiple tasks)



Meta-Learning Inner and Outer Updates



- Given a task $\mathcal{T}_e$, the learner adapts the global params $\theta_0 = \theta_g$ to individual task params θ_T through a T updates, each update is based on the previous updated params and the task loss

$$\theta_t = \theta_{t-1} - \alpha \frac{\partial L(\theta_{t-1})}{\partial \theta_{t-1}}$$

- The meta-learner then tries to optimize the global params based on the updated individual task params θ_T and the task loss

$$\theta_0 = \theta_0 - \beta \frac{\partial L(\theta_T)}{\partial \theta_0} = \theta_0 - \beta \frac{\partial L(U^T(\theta_0))}{\partial \theta_0}$$

Inner-Loop Computing Gradients and Updating Params

- Computing gradient and updating params, unrolling

$$\theta_0 = \theta_g$$

$$\theta_1 = U_0(\theta_0) = \theta_0 - \alpha \frac{\partial L(\theta_0)}{\partial \theta_0}$$

$$\theta_2 = U_1(\theta_1) = \theta_1 - \alpha \frac{\partial L(\theta_1)}{\partial \theta_1}$$

$$\theta_2 = U_1(U_0(\theta_0)) = \theta_0 - \alpha(\frac{\partial L(\theta_1)}{\partial \theta_1} + \frac{\partial L(\theta_0)}{\partial \theta_0})$$

$$\theta_T = U^T(\theta_0) = (U_{T-1} \circ U_{T-2} \ldots \circ U_0)(\theta_0) = \theta_0 - \alpha \sum_{t=0}^{T-1} \frac{\partial L(\theta_t)}{\partial \theta_t}$$

Outer-Loop Computing Gradients and Updating Params

- Updating params

$$\theta_0 = \theta_0 - \beta \frac{\partial L(U^T(\theta_0))}{\partial \theta_0}$$

- Computing gradients, second-order (full) MAML, unrolling

$$\frac{\partial L(U^T(\theta_0))}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_T} \cdot \frac{\partial \theta_T}{\partial \theta_0}$$

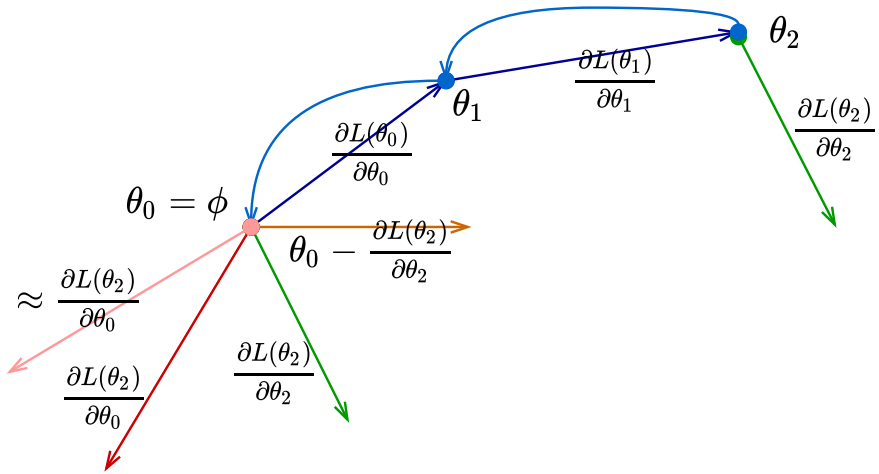
$$\frac{\partial L(U^2(\theta_0))}{\partial \theta_0} = \frac{\partial (L \circ U_1 \circ U_0)(\theta_0)}{\partial \theta_0}$$

$$= \frac{\partial (L \circ U_1 \circ U_0)(\theta_0)}{\partial (U_1 \circ U_0)(\theta_0)} \cdot \frac{\partial (U_1 \circ U_0)(\theta_0)}{\partial U_0(\theta_0)} \cdot \frac{\partial U_0(\theta_0)}{\partial \theta_0}$$

$$= \frac{\partial L(\theta_2)}{\partial \theta_2} \cdot \frac{\partial U_1(\theta_1)}{\partial \theta_1} \cdot \frac{\partial U_0(\theta_0)}{\partial \theta_0}$$

$$\frac{\partial L(U^T(\theta_0))}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_T} \cdot \prod_{t=0}^{T-1} \frac{\partial U_t(\theta_t)}{\partial \theta_t}$$

Model Variants of MAML Outer Updates



- Full MAML

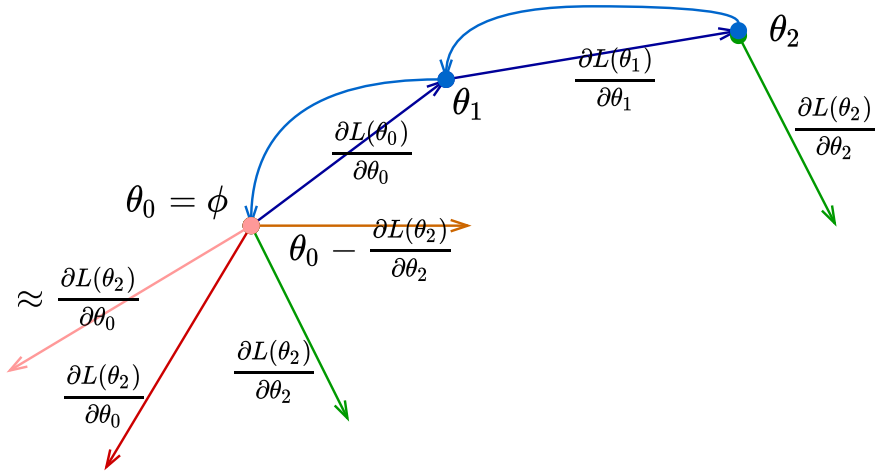
$$\frac{\partial L(\theta_T)}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_T} \cdot \frac{\partial \theta_T}{\partial \theta_0}$$

- First-order MAML

$$\frac{\partial L(\theta_T)}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_T}$$

  - only use first-order term, which is the last gradient of loss of the individual task

Model Variants of MAML Outer Updates



- Full MAML

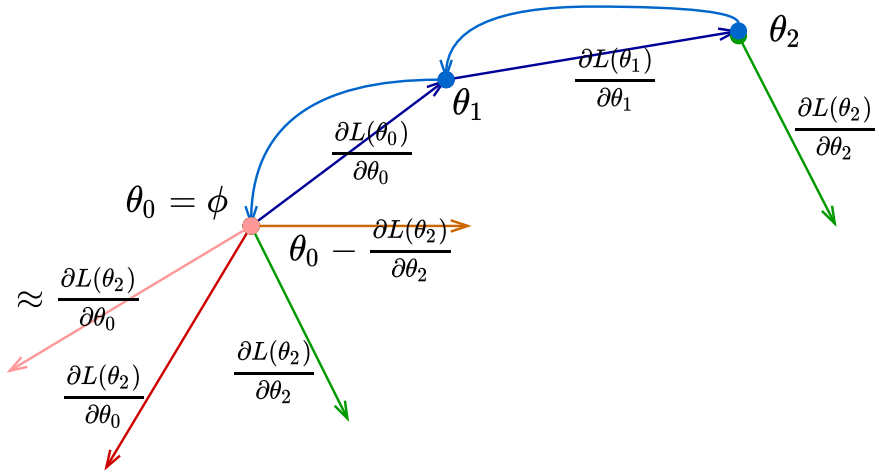
$$\frac{\partial L(\theta_T)}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_T} \cdot \frac{\partial \theta_T}{\partial \theta_0}$$

- Reptile

$$\frac{\partial L(\theta_T)}{\partial \theta_0} = \theta_0 - \frac{\partial L(\theta_T)}{\partial \theta_T}$$

  - only use first-order term,
  - average the task gradient with the global gradient

Model Variants of MAML Outer Updates



- Full MAML

$$\frac{\partial L(\theta_T)}{\partial \theta_0} = \frac{\partial L(\theta_T)}{\partial \theta_T} \cdot \frac{\partial \theta_T}{\partial \theta_0}$$

- Implicit MAML (iMAML)

$$\frac{\partial \theta_T}{\partial \theta_0} \approx \left(1 + \lambda \frac{\partial^2 L(\theta_T)}{\partial \theta_T^2}\right)^{-1}$$

  - approximate second-order term with the last gradient

Thank you !