THE UNIVERSITY OF TOKYO

ON THE CONDITIONS OF MAML CONVERGENCE

SHIRO TAKAGI[1] YOSHIHIRO NAGANO[1] YUKI YOSHIDA[1] MASATO OKADA[1]

[1]THE UNIVERSITY OF TOKYO

We derived the necessary conditions of inner learning rate α and meta-learning rate β for a simplified MAML to locally converge to local min from any point in the vicinity of the local min

We found that maximum possible β is larger when α is close to its maximum possible value

MAML AS NEGATIVE GRADIENT PENALTY

MAML

Update parameters to find a representation that can rapidly adapt to new tasks with a small quantity of data.

1. Inner loop

$$\boldsymbol{\theta}'_\tau = \boldsymbol{\theta} - \underbrace{\alpha}_{\text{Inner lr}} \nabla_{\boldsymbol{\theta}} L_\tau(\boldsymbol{\theta})$$

2. Meta-loop

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \underbrace{\beta}_{\text{Meta-lr}} \nabla_{\boldsymbol{\theta}} \sum_{\tau \sim P(\tau)} L_\tau(\boldsymbol{\theta}'_\tau)$$

[Finn et al. 2017]

Assumptions

1. Only one step is taken per update
2. Only one task is considered
3. Data are not resampled

Approximation

Ignore higher order derivative terms for simplicity.

$$\nabla_{\boldsymbol{\theta}} L_\tau(\boldsymbol{\theta}'_\tau) = \nabla_{\boldsymbol{\theta}} \theta'_\tau \frac{\partial L_\tau}{\partial \theta'_\tau} = (I - \alpha \nabla^2_{\boldsymbol{\theta}} L_\tau) \frac{\partial L_\tau}{\partial \theta'_\tau}$$

$$\approx \boldsymbol{g}_\tau(\boldsymbol{\theta}) - \alpha H_\tau(\boldsymbol{\theta}) \boldsymbol{g}_\tau(\boldsymbol{\theta})$$

Negative gradient penalty

An approximated MAML loss can be regarded as the loss with the negative gradient penalty.

$$\tilde{L}(\theta) = L(\theta') \approx L(\theta) - \frac{\alpha}{2} g(\theta)^\top g(\theta)$$

Then, update equation of $\boldsymbol{v}$ is

$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) - \beta \Lambda_{\tilde{H}} \boldsymbol{v}(t)$$

[LeCun et al. 1998]

To reach a minimum, β should satisfy the following condition:

$$\forall i, |\, 1 - \beta\, \lambda(H - \alpha H^2)_i | < 1$$

The necessary condition of β is:

$$\forall i, \quad \beta < \frac{2}{\lambda(H)_i - \alpha\, \lambda(H)_i^2}$$

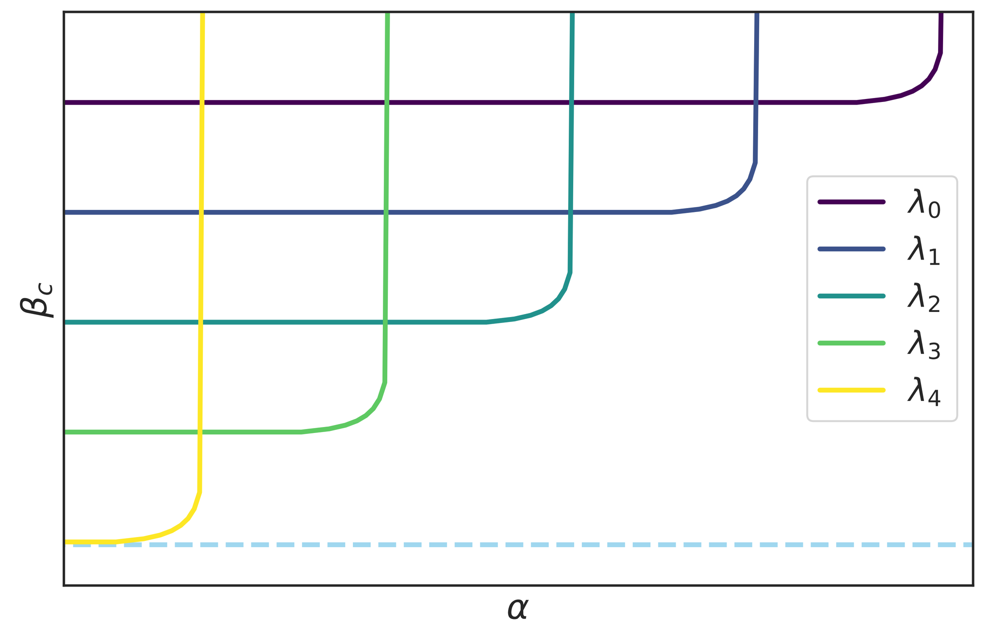
Condition for simplified MAML to locally converge

The condition for the simplified MAML to locally converge to local min from any point in the vicinity of the local min:

$$\forall i, \quad \alpha \le \frac{1}{\lambda(H)_i} \wedge \beta < \frac{2}{\lambda(H)_i - \alpha\, \lambda(H)_i^2}$$

As α approaches α_c, upper bound β_c diverges.

$\Rightarrow$ β_c is larger when α is close to α_c



CONVERGENCE CONDITION

Simplified MAML Loss

Taking the Taylor series for the second-order term at a fixed point $\boldsymbol{\theta}^*$, the simplified MAML loss is

$$\tilde{L}(\boldsymbol{\theta}) \approx \tilde{L}(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \tilde{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

Condition for inner learning rate α

Since the necessary condition $\boldsymbol{\theta}^*$ to be a local minimum is that all eigenvalues λ of the Hessian $\tilde{H}$ at $\boldsymbol{\theta}^*$ are non-negative, α should satisfy the following condition:

$$\forall i, \lambda(\tilde{H})_i \approx \lambda(H)_i - \alpha\, \lambda(H)_i^2 \ge 0$$

$$^* \tilde{H} = H - \alpha(T\boldsymbol{g} + H^2) \approx H - \alpha(H^2)$$

The necessary condition of α is:

$$\forall i, \quad \alpha \le \frac{1}{\lambda(H)_i}$$

Condition for meta-learning rate β

$\tilde{H}$ is diagonalizable: $\tilde{H} = P \Lambda_{\tilde{H}} P^\top$

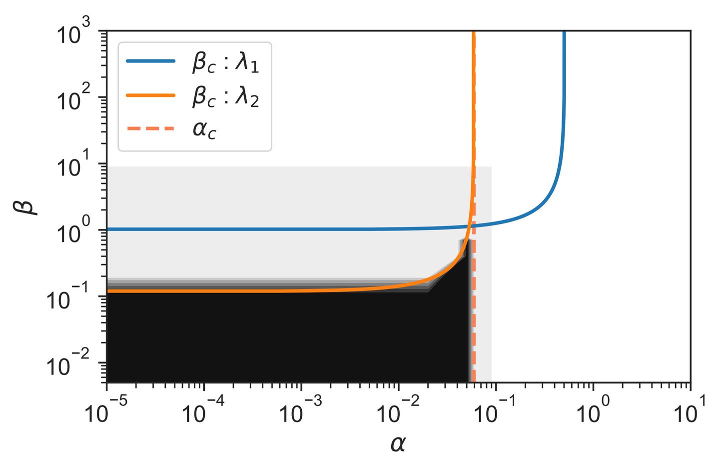
Denoting $P^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ by $\boldsymbol{v}$, the simplified MAML loss is

$$\tilde{L}(\boldsymbol{v}) \approx \tilde{L}(0) + \frac{1}{2}\boldsymbol{v}^\top \Lambda_{\tilde{H}} \boldsymbol{v}$$

EXPERIMENTS

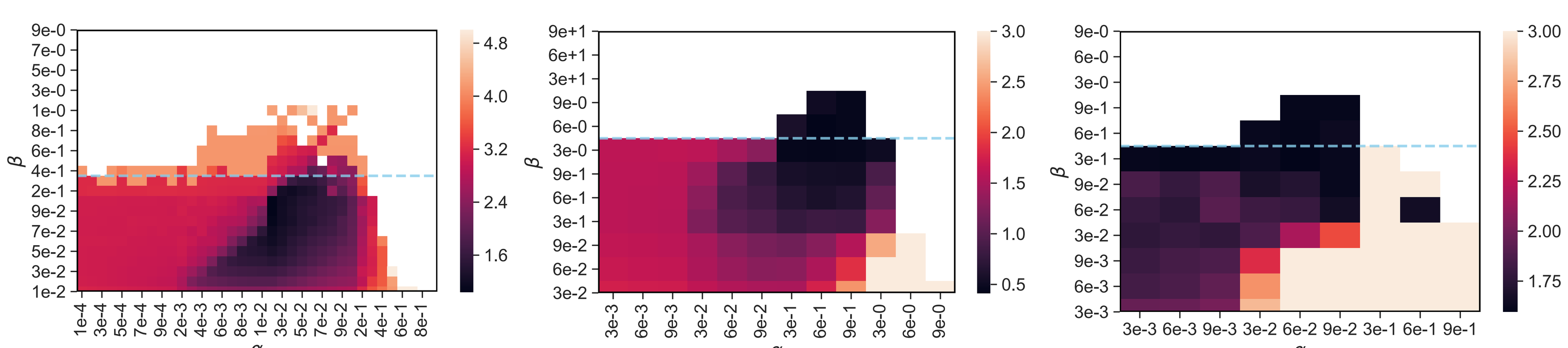
Linear regression

Linear regression with all assumptions being hold. Theoretical β_c and α_c match empirical ones.



Few-shot learning

Training error of (a) Sinusoid regression and (b) Omniglot and (c) MiniImagenet classification with various α and β.



(a) Sinusoid (b) Omninglot (c) MiniImagenet

The largest possible β is larger whenα is close to its maximum possible value.

$\Rightarrow$ Our theory explains the experimental result.