Meta-learning with negative learning rates

Alberto Bernacchia
MediaTek Research

Meta-learning

training data

Braque

Cezanne

test datapoint

By Braque or Cezanne?

Model agnostic meta-learning (MAML)

Learns the optimal starting point to adapt to multiple tasks

while not done **do**

Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$

for all $\mathcal{T}_i$ **do**

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$$

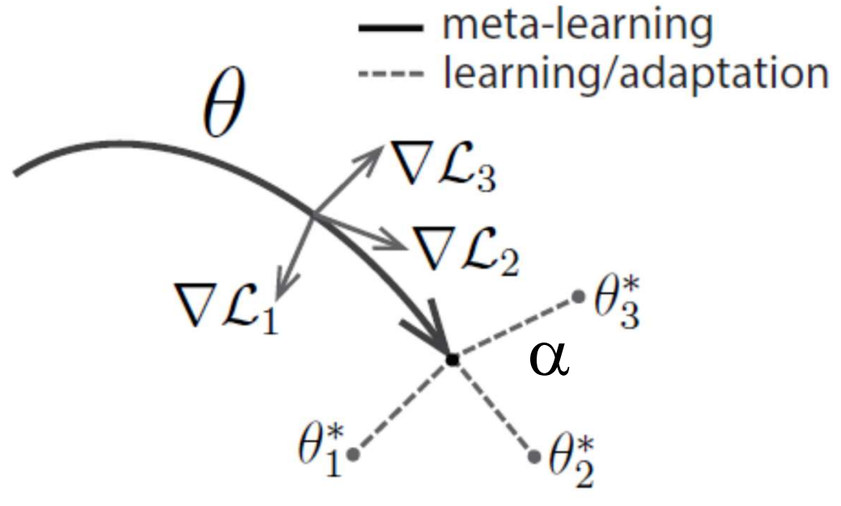
end for

Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$

end while

outer loop

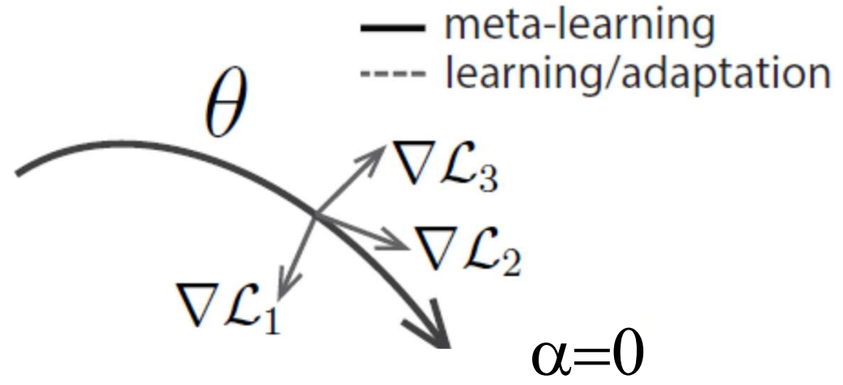
inner loop



Finn et al 2017 ICML

Model agnostic meta-learning (MAML)

Learns the optimal starting point to adapt to multiple tasks



outer loop

while not done **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for all $\mathcal{T}_i$ **do**
    $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
end while

inner loop

Method	Omniglot-20way-1shot	Omniglot-20way-5shot	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
MAML	93.7 ± 0.7	96.4 ± 0.1	46.9 ± 0.2	63.1 ± 0.4
ANIL	96.2 ± 0.5	98.0 ± 0.3	46.7 ± 0.4	61.5 ± 0.5
NIL	96.7 ± 0.3	98.0 ± 0.04	48.0 ± 0.7	62.2 ± 0.5

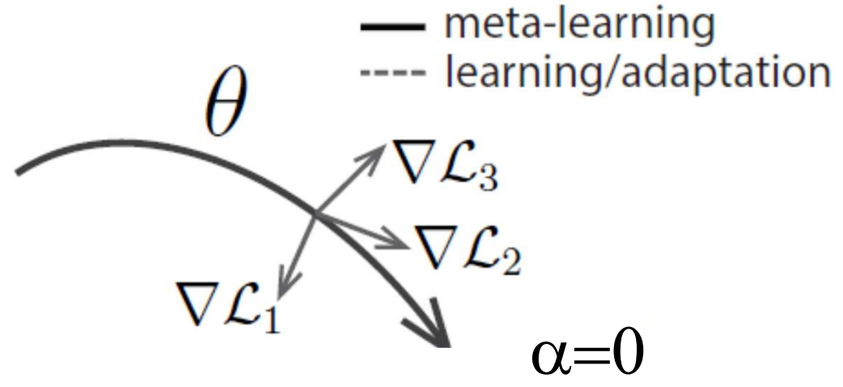
Raghu et al 2020 ICLR

Model agnostic meta-learning (MAML)

Learns the optimal starting point to adapt to multiple tasks

while not done **do**
  Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
  for all $\mathcal{T}_i$ **do**
    $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
  end for
  Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'})$
end while

outer loop



what is the optimal learning rate?

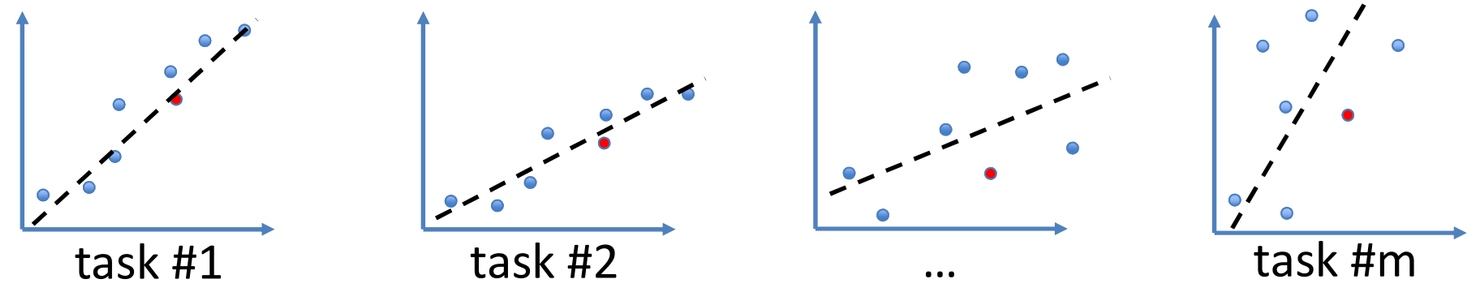
Method	Omniglot-20way-1shot	Omniglot-20way-5shot	MiniImageNet-5way-1shot	MiniImageNet-5way-5shot
MAML	93.7 ± 0.7	96.4 ± 0.1	46.9 ± 0.2	63.1 ± 0.4
ANIL	96.2 ± 0.5	98.0 ± 0.3	46.7 ± 0.4	61.5 ± 0.5
NIL	96.7 ± 0.3	98.0 ± 0.04	48.0 ± 0.7	62.2 ± 0.5

Finn et al 2017 ICML

Raghu et al 2020 ICLR

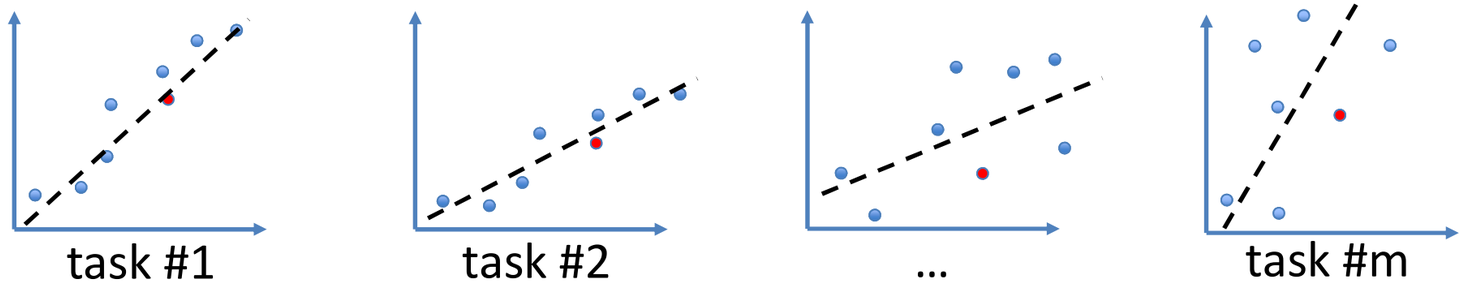
Mixed linear regression

$$y = \mathbf{x}^T \mathbf{w} + z$$

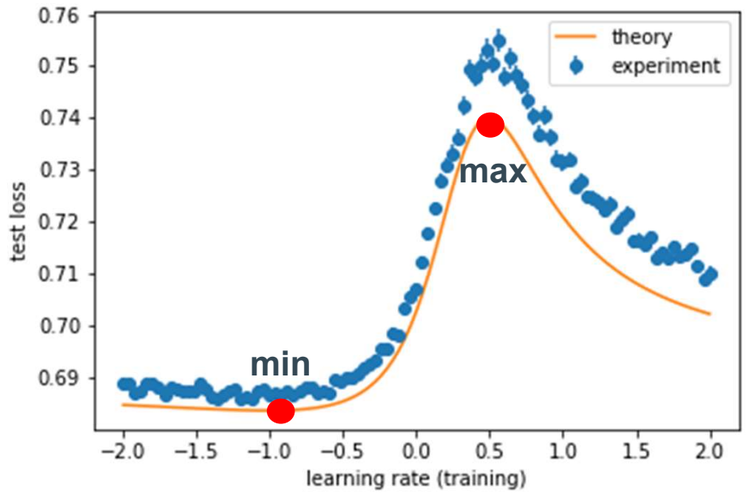


Mixed linear regression

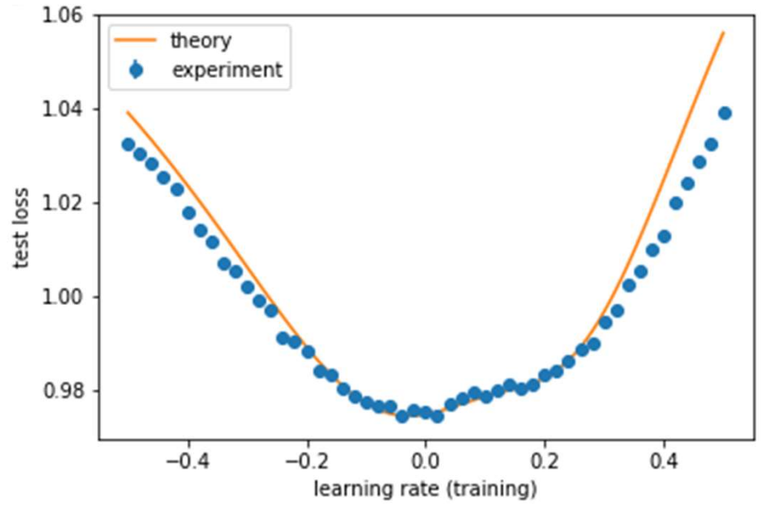
$$y = \mathbf{x}^T \mathbf{w} + z$$



overparameterized

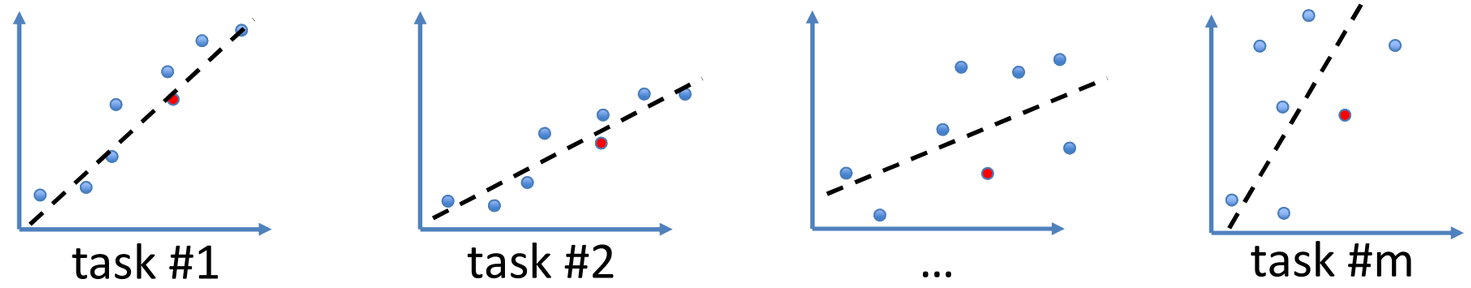


underparameterized

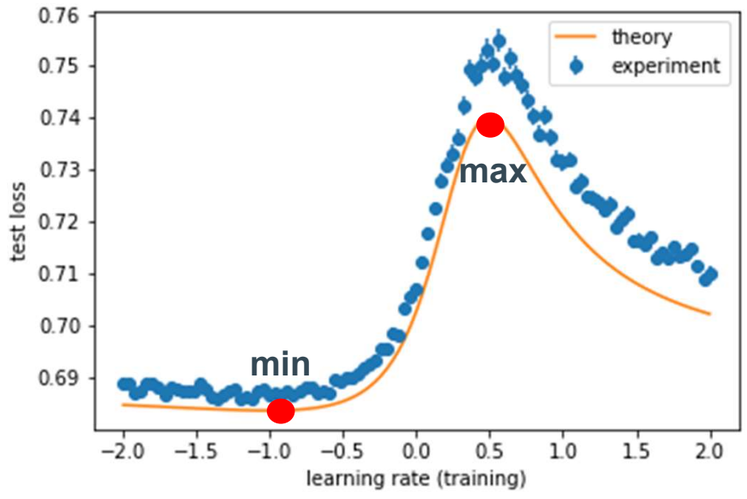


Mixed linear regression

$$y = \mathbf{x}^T \mathbf{w} + z$$

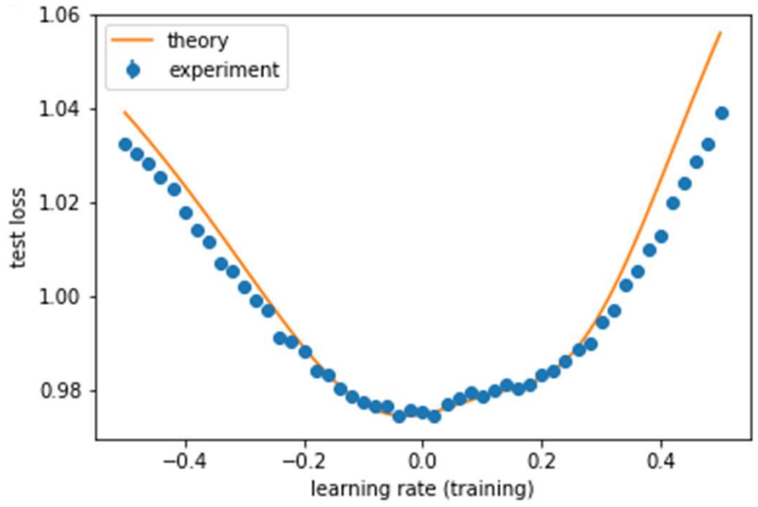


overparameterized



$$\alpha_t^{\pm} = -\frac{n_t + 1}{2p} \pm \sqrt{\left(\frac{n_t + 1}{2p}\right)^2 + \frac{n_t}{p}}$$

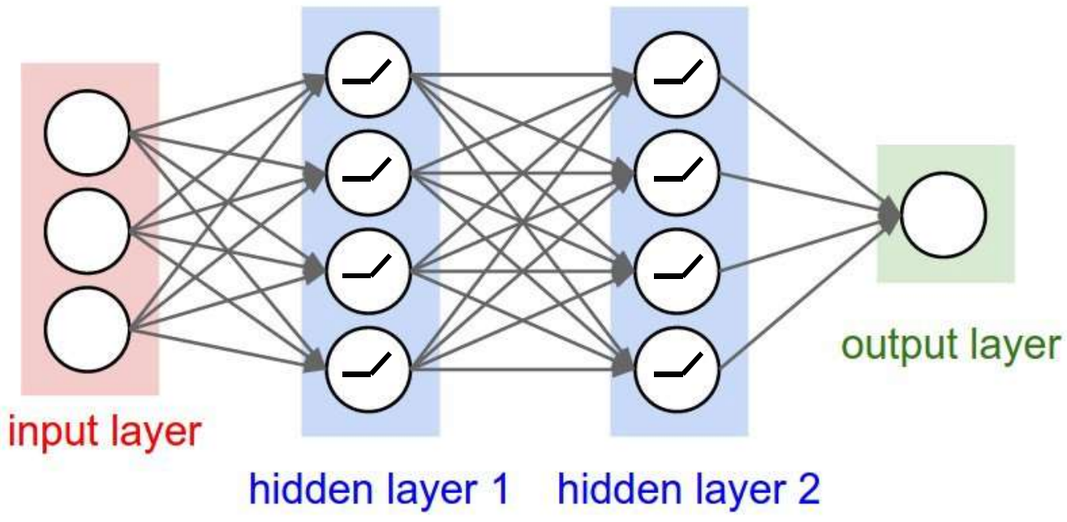
underparameterized

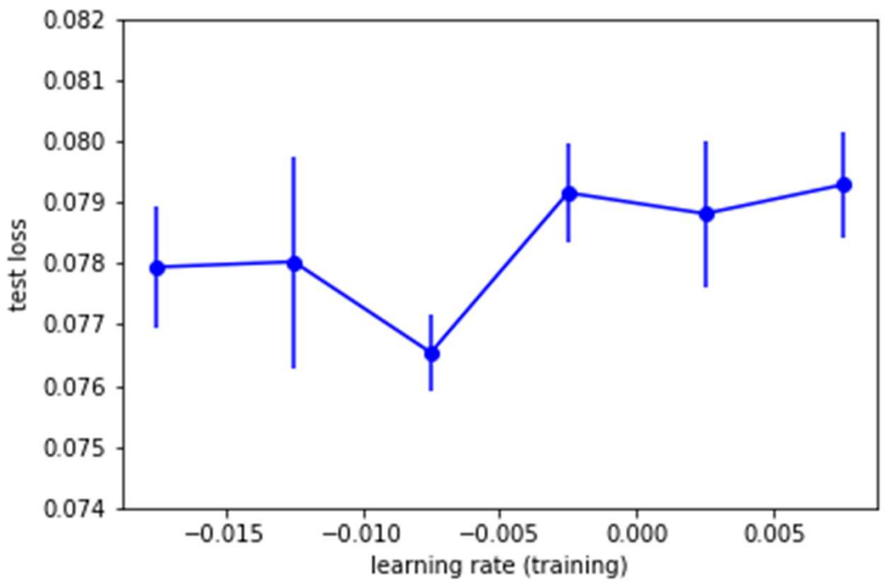


$$\left.\frac{\partial \overline{\mathcal{L}}^{test}}{\partial \alpha_t}\right|_{\alpha_t = 0} = \frac{\sigma^2 p}{n_v m}$$

Nonlinear regression

$$y = (\mathbf{w}^T \mathbf{x} + b)^2 + z$$





Summary

In simple problems (linear and quadratic regression)
optimal learning rate is negative

This result is counter-intuitive, but the learning rate of the inner loop does not have to be positive to guarantee convergence.

How general is this result?

What is the intuition behind this observation?