META-LEARNING WITH NEGATIVE LEARNING RATES

沈雯杰

LAMDA, Nanjing University

January 29, 2021

目录

目录

MAML

We auusume the existence of a distribution of tasks τ and, for each task, a distribution of data points $\mathcal{D}^r$ and a loss function $\mathcal{L}^\tau$. The loss function of the meta-learning problem, $\mathcal{L}^{\text{meta}}$, is defined as an average over both distributions of tasks and data points. The goal of meta-learning is to minimize the loss function with respect to a vector of meta-parameters ω

$$\mathcal{L}^{\text{meta}}(\boldsymbol{\omega}) = \mathop{\mathbb{E}}_{\tau} \mathop{\mathbb{E}}_{\mathcal{D}_t^\tau} \mathop{\mathbb{E}}_{\mathcal{D}_v^\tau} \mathcal{L}^\tau \left(\boldsymbol{\theta}^\tau \left(\boldsymbol{\omega}, \mathcal{D}_t^\tau \right); \mathcal{D}_v^\tau \right)$$

$$\mathcal{L}^{\text{meta}}(\boldsymbol{\omega}) = \frac{1}{m n_v} \sum_{i=1}^{m} \sum_{j=1}^{n_v} \mathcal{L} \left(\boldsymbol{\theta}^{(i)}(\boldsymbol{\omega}); \mathcal{D}_j^{(i)} \right)$$

MAML

In this work we consider the simple case of a single gradient step. Therefore, the inner loop of MAML is given by

$$\boldsymbol{\theta}^{(i)}(\boldsymbol{\omega}) = \boldsymbol{\omega} - \frac{\alpha_t}{n_t} \sum_{j=1}^{n_t} \left. \frac{\partial \mathcal{L}^{(i)}}{\partial \theta} \right|_{\boldsymbol{\omega}; \mathcal{D}_j^{(i)}}$$

目录

MIXED LINEAR REGRESSION

In mixed linear regression, each task is characterized by a different linear function, and a model is evaluated by the mean squared error loss function. We assume a generative model in the form of $y = \mathrm{x}^T \mathrm{w} + z$, where x is the input vector (of dimension p), y is the output (scalar), z is noise (scalar), and $\mathbf{w}$ is a vector of generating parameters (of dimension p)

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{w}_0, \frac{\nu^2}{p} I_p\right) \quad \mathbf{x} \sim \mathcal{N}\left(0, I_p\right) \quad y \mid \mathbf{x}, \mathbf{w} \sim \mathcal{N}\left(\mathbf{x}^T \mathbf{w}, \sigma^2\right)$$

目录

OVERPARAMETERIZED CASE

Let $p > n_v m$.

$$\overline{\mathcal{L}}^{\text{test}} = \frac{\sigma^2}{2}\left(1 + \frac{\alpha_r^2 p}{n_r}\right) +$$
$$+h^r\left[\frac{\nu^2}{2}\left(1 + \frac{n_v m}{p}\right) + \frac{1}{2}\left(1 - \frac{n_v m}{p}\right)|\omega_0 - \mathbf{w}_0|^2 + \frac{\sigma^2 n_v m}{2p}\frac{1+\frac{\alpha_t^2 p}{n_t}}{h^t}\right] + o\left(\xi^{-3/2}\right)$$

where we define the following expressions

$$h^t = (1-\alpha_t)^2 + \alpha_t^2\frac{p+1}{n_t}$$

$$h^r = (1-\alpha_r)^2 + \alpha_r^2\frac{p+1}{n_r}$$

Proof. The proof of this Theorem can be found in the Appendix, sections $7.3, 7.3.1$.

对于 α_r 来说，这个式子可以看成是两个二次函数的相加，一个最小值取在 0 的时候，一个取在 $\alpha_r = 1/\left(1 + (p+1)/n_r\right)$ 的时候，所以总的最小值在两者之间并且一定是正的。

对于 α_t 来说，通过求导可以发现它可能在负值处有最小值：

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

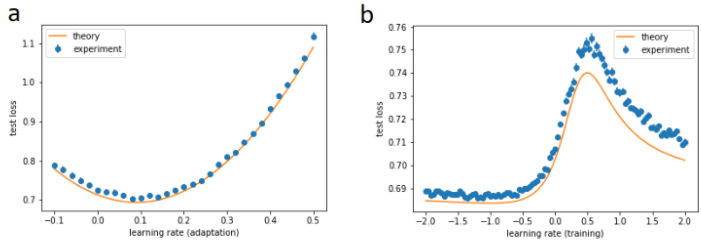

图:

In panel a) we set $\alpha_t = 0.2$, in panel b) we set $\alpha_r = 0.2$. In the experiments, each run is evaluated on 100 tasks of 50 data points each, and each point is an average over 100 runs (a) or 1000 runs (b).

UNDERPARAMETERIZED CASE

Theorem 2. Let $p < n_v$ m.

$$\overline{\mathcal{L}}^{\text{test}} = \frac{\sigma^2}{2}\left(1 + \frac{\alpha_r^2 p}{n_r}\right) + \frac{h^r \nu^2}{2} +$$
$$+ \frac{h^r}{2h^t}\frac{p}{n_v m}\left\{\sigma^2\left[h^t + \frac{\alpha_t^2}{n_t}\left[(n_v+1)\,g_1 + p g_2\right]\right] + \frac{\nu^2}{p}\left[(n_v+1)\,g_3 + p g_4\right]\right\} + o\left((m\xi)^{-3/2}\right)$$

同样的，对于 α_r 来说，这个式子可以看成是两个二次函数的相加，一个最小值取在 0 的时候，一个取在 $\alpha_r = 1/\left(1 + (p+1)/n_r\right)$ 的时候，所以总的最小值在两者之间并且一定是正的。

对于 α_t 来说，虽然不能直接求出导数为 0 的情况，但是可以得到 $\left.\frac{\partial \overline{\mathcal{L}}^{\text{test}}}{\partial \alpha_t}\right|_{\alpha_t=0} = \frac{\sigma^2 p}{n_v m} \geq 0$ 由此说明可能在负值处取到最小值。

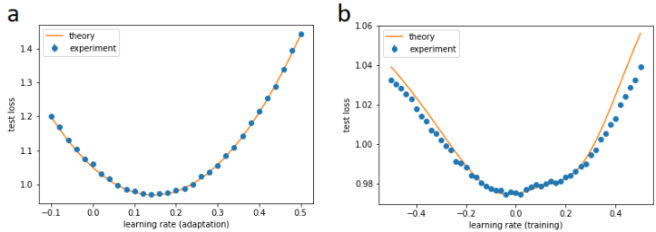

图:

Values of parameters: $n_t = 5, n_v = 25, n_r = 10, m = 40, p = 30, \sigma = 0.2, \nu = 0.2$. In panel a) we set $\alpha_t = 0.2$, in panel b) we set $\alpha_r = 0.2$. In the experiments, the model is evaluated on 100 tasks of 50 data points each, and each point is an average over 100 (a) or 1000 (b) runs.

参考文献

https://openreview.net/pdf?id=60j5LygnmD