Meta-Learning #3

February 11, 2022

Ozgur Taylan Turan

Generalization of MAML for Linear Problems

Problem

$$y = \mathbf{x}\mathbf{a}^{\mathsf{T}} + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma=1)}$$

- $\mathbf{a} \in \mathbb{R}^{D}$ represents the task
- Assume $\mathbf{a} \sim \mathcal{N}(m\mathbf{1}, c\boldsymbol{I})$, where $\mathbf{1} \in \mathbb{R}^{D\times 1}$ and $\boldsymbol{I} \in \mathbb{R}^{D\times D}$
- Training data $Z := (\mathbf{x}_i, y_i)_{i=1}^{N}$ are drawn from distribution distribution p_Z.
- $p_x \sim \mathcal{N}(\mathbf{0}, k\boldsymbol{I})$
- Expected Loss is $\mathcal{E} := \int\int\int (\mathbf{x}\hat{\mathbf{a}}_N(Z)^{\mathsf{T}} - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy p_Z dZ p_{\mathbf{a}} d\mathbf{a}$.

Models

$$\mathcal{M}(\mathbf{w}, \mathbf{b}, \mathbf{x}) := \mathbf{w}^{\mathsf{T}}\mathbf{x} + b$$
$$\mathcal{L} := ||\mathbf{y} - \mathcal{M}(\bar{\mathbf{w}}, \mathbf{X})||^2$$
$$\mathcal{L}_{\text{ridge}} := ||\mathbf{y} - \mathcal{M}(\bar{\mathbf{w}}, \mathbf{X})||^2 + \lambda||\bar{\mathbf{w}}||^2$$
$$\mathcal{L}_{\text{gen.ridge}} := ||\mathbf{y} - \mathcal{M}(\bar{\mathbf{w}}, \mathbf{X})||^2 + \lambda||\bar{\mathbf{w}} - \mathbf{h}||^2$$

Note that,

- $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & 1 \\ \vdots & \vdots \\ \mathbf{x}_n & 1 \end{bmatrix}_{N\times(d+1)}$, $\bar{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}_{(d+1)}$ and $\mathbf{h} \in \mathbb{R}^{D\times 1}$

Models

$$\mathcal{M}(\mathbf{w},\mathbf{b},\mathbf{x}) := \mathbf{w}^{\mathsf{T}}\mathbf{x} + b$$
$$\mathcal{L} := ||\mathbf{y} - \mathcal{M}(\bar{\mathbf{w}},\mathbf{X})||^2$$
$$\mathcal{L}_{\text{ridge}} := ||\mathbf{y} - \mathcal{M}(\bar{\mathbf{w}},\mathbf{X})||^2 + \lambda||\bar{\mathbf{w}}||^2$$
$$\mathcal{L}_{\text{gen.ridge}} := ||\mathbf{y} - \mathcal{M}(\bar{\mathbf{w}},\mathbf{X})||^2 + \lambda||\bar{\mathbf{w}} - \mathbf{h}||^2$$

- Linear: $\bar{\mathbf{w}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$
- Ridge: $\bar{\mathbf{w}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$
- GeneralRidge[1]: $\bar{\mathbf{w}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^{\mathsf{T}}\mathbf{y} + \lambda\mathbf{h})$
- GD: $\mathbf{w}_{niter+1} = \mathbf{w}_{niter} - l_r * \frac{2}{N}\sum_{i=1}^{N} \mathbf{x}_i\big((\mathbf{w}_{niter}^{\mathsf{T}}\mathbf{x}_i + b) - y_i\big)$ and $b_{niter+1} = b_{niter} - l_r * \frac{2}{N}\sum_{i=1}^{N}\big((\mathbf{w}_{niter}^{\mathsf{T}}\mathbf{x}_i + b) - y_i\big)$

[1] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil (2018). "Learning to learn around a common mean". In: *Advances in Neural Information Processing Systems* 2018-Decem.NeurIPS, pp. 10169–10179. ISSN: 10495258

Additional Info

- GD: starts from $\bar{\mathbf{w}}_{\text{opt}}$ and takes step with drawn training samples
- MAML: start GD from $\bar{\mathbf{w}} \sim \mathcal{N}(\bar{\mathbf{w}}_{\text{opt}}, 0.1\mathbf{I})$
- randomGD: start with a random initialization for $\bar{\mathbf{w}}$
- Bayes: Bayes error resulting from the $\bar{\mathbf{w}}_{\text{opt}}$

Aim

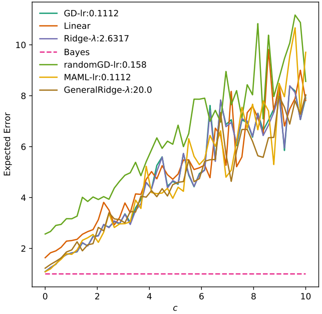
- See the effect of the MAML adaptation.
- See if Ridge Regression can compete with a task informed algorithm.

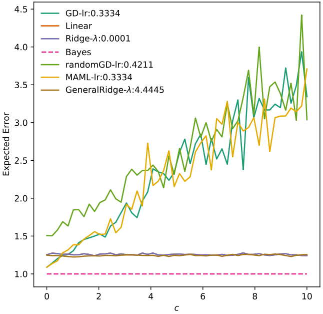
Experiments

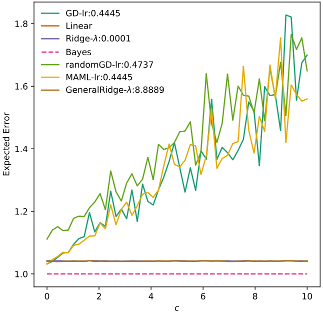
- $n_{iter} : [0, 90; 10]$, $\sigma : [0, 5; 50]$, $D : [1, 50; 50]$, $m : [0, 10; 50]$, $c : [0, 10, 50]$, $b : [1, 5, 50]$, $N : [1, 10; 50]$
- $\lambda : [0.0001, 20; 10]$, $l_r : [0.0001, 1, 10]$
- $N_{test} : 1000$, $N_a : 100$, $N_Z : 100$, $m = 0$, $c = 1$, $\sigma = 1$

Task Variance(c)

Increasing $N : 1 \to 10 \to 50$ and $D : 1$

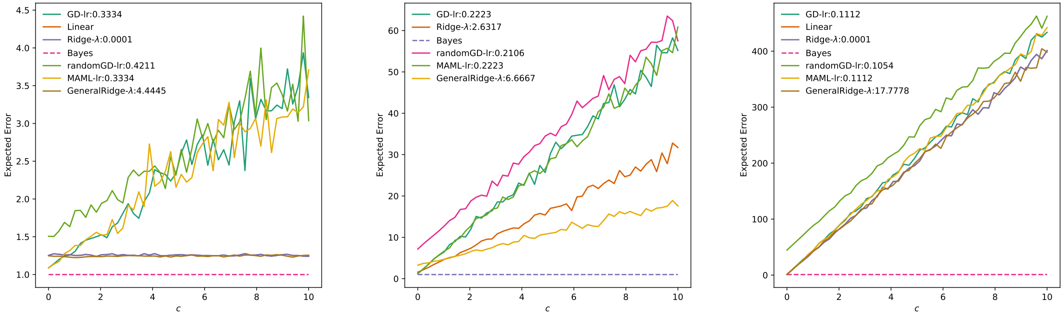






Task Variance(c)

Increasing $D : 1 \rightarrow 10 \rightarrow 50$ and $N : 10$



Next

- Investigation of that little area where the MAML performs better only over all the experimentation. Increasing dimensions. (Fresh out results!)
- Writing the Draft! (Started!)
- Non-linear experimentations (Next-week!) [Predicting the sine-wave with changing amplitude and the phase...]

Struggles

- Kernel Ridge for GeneralRidge like bias parameter?
- Why is it not common to gradient descent with Kernel Ridge regression?