Is Bayesian Model-Agnostic Meta Learning Better than Model-Agnostic Meta Learning, Provably?

Lisha Chen, Tianyi Chen

Rensselaer Polytechnic Institute

03/14/2022

Definition of meta learning

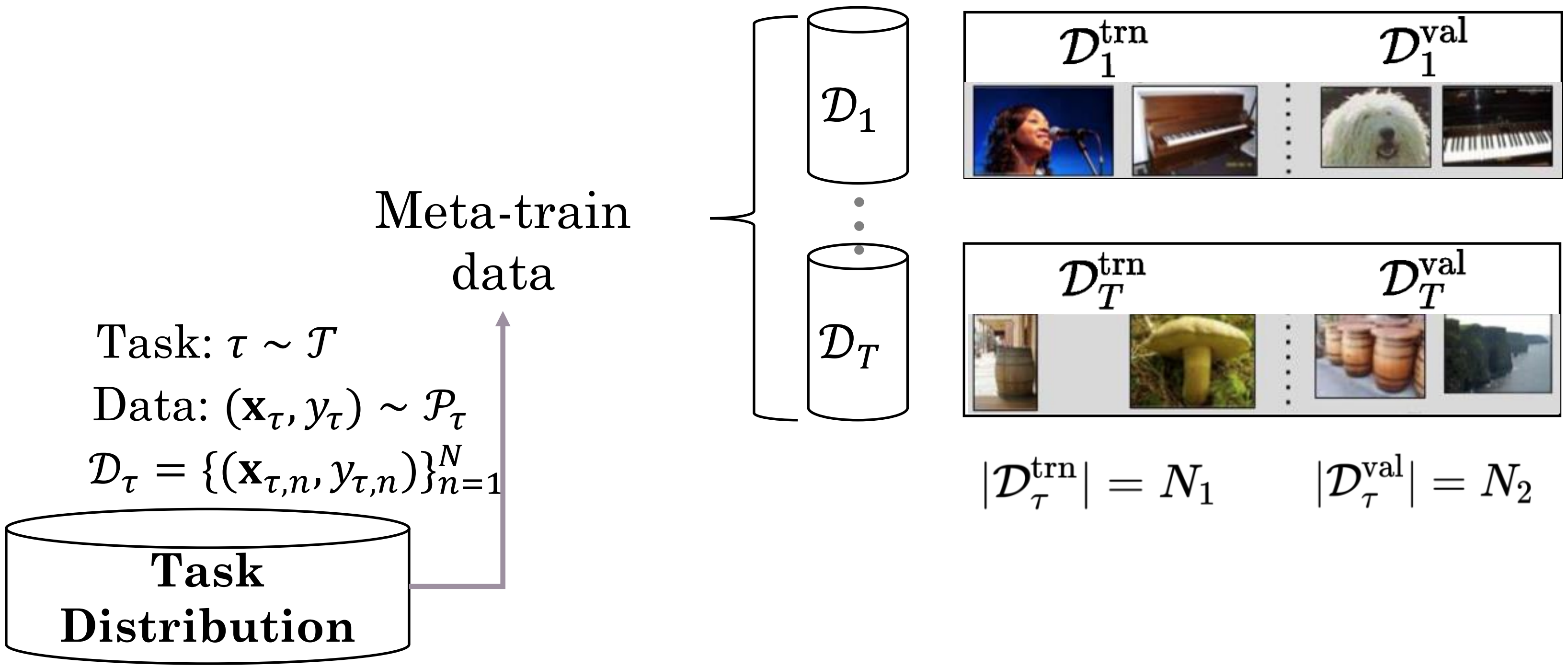
Meta-learning (learning to learn):

To learn a model that can well adapt or generalize to new tasks and new environments that have never been encountered during training time.

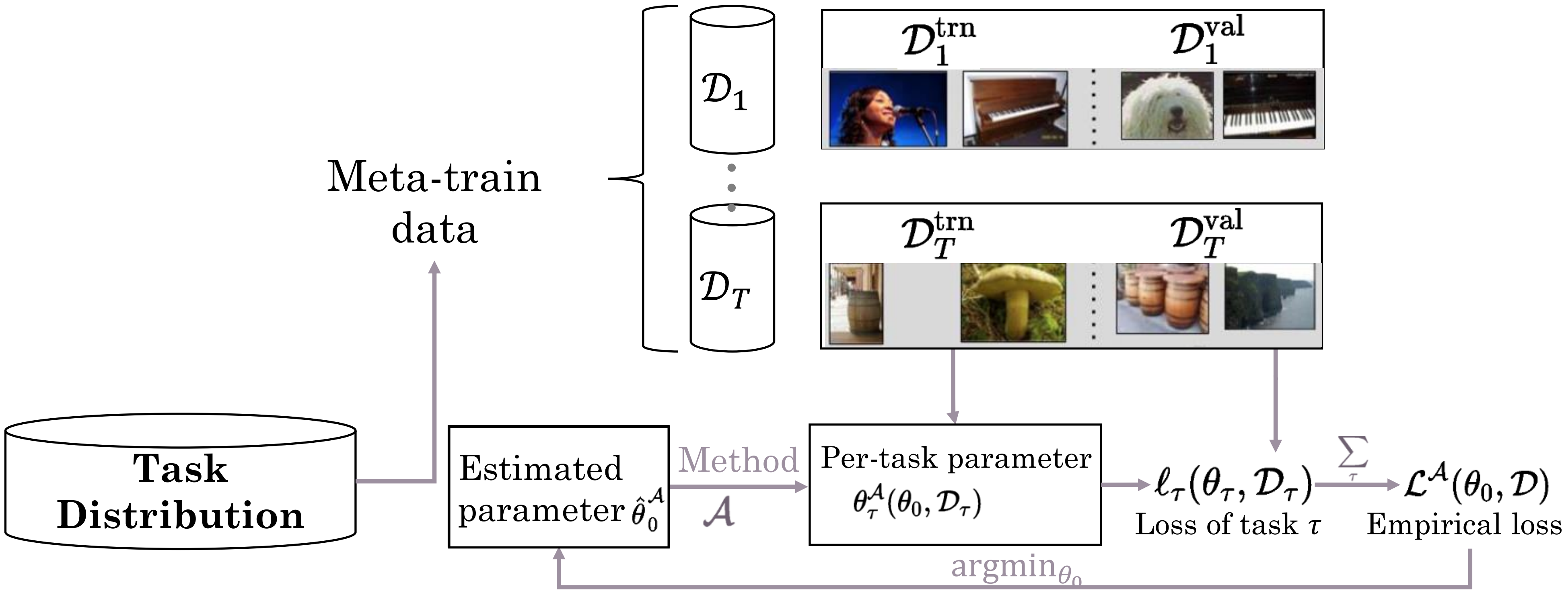
Prior work (incomplete)

[Bengio et al ’90]
[Maclaurin et al ’15]
[Vinyals et al ’16]
[Santoro et al ‘16]
[Wichrowska et al ’17]
[Finn et al ’17]

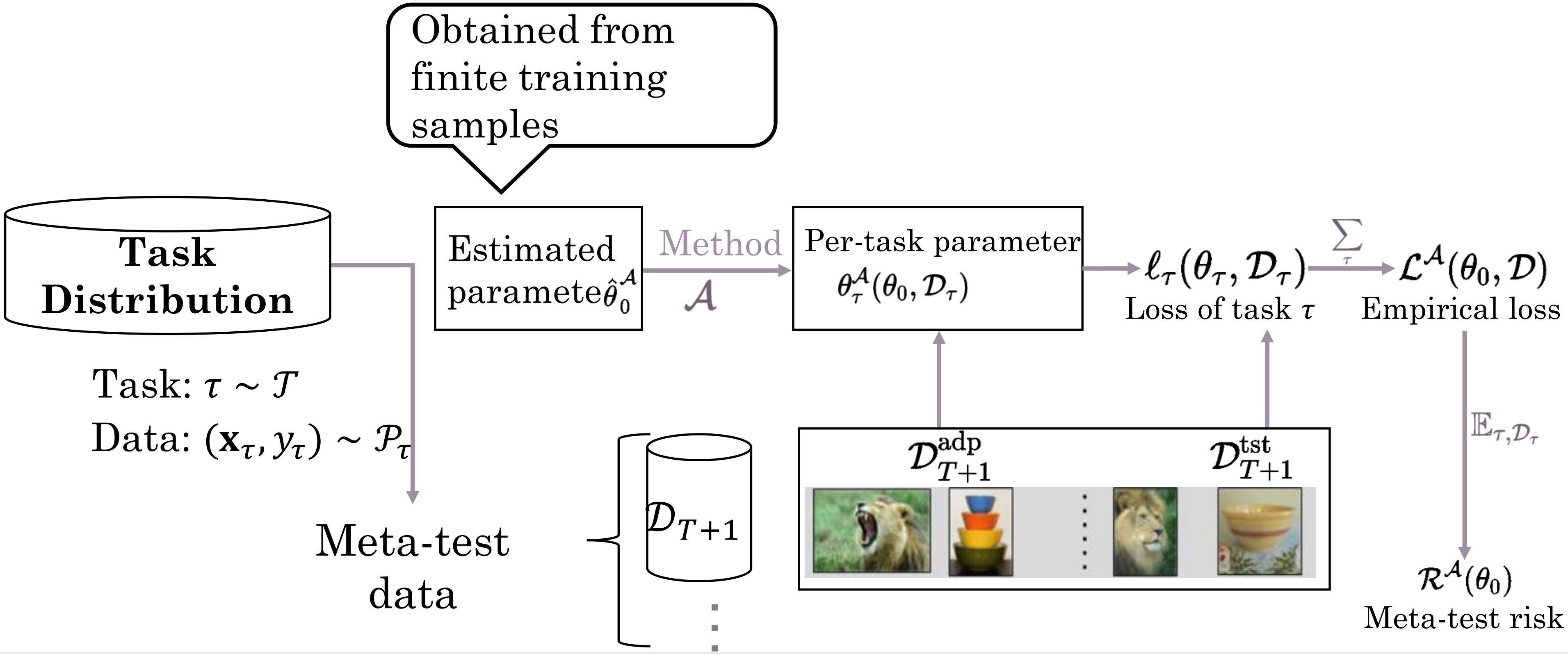
Meta learning setup



Meta learning setup

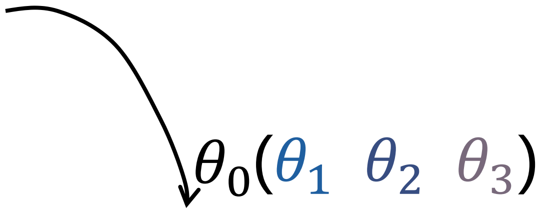


Meta learning setup

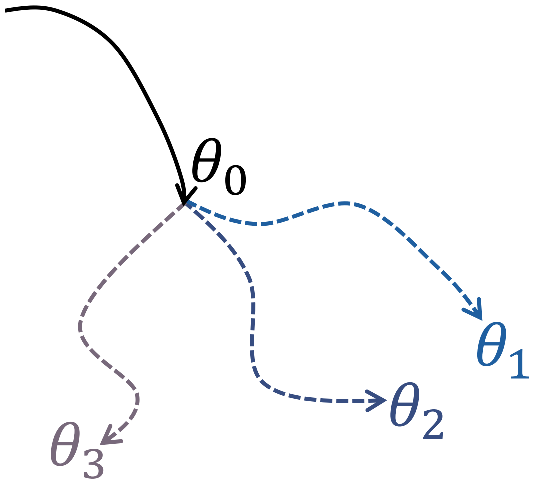


Basis of comparison

$$\hat{\theta}_{\tau}^{\text{er}}\left(\theta_0, \mathcal{D}_{\tau}^{\text{trn}}\right) = \theta_0$$



$$\hat{\theta}_{\tau}^{\mathcal{A}}\left(\theta_0, \mathcal{D}_{\tau}^{\text{trn}}\right) \neq \theta_0$$



$$\hat{\theta}_{\tau}^{\mathcal{A}}\left(\theta_0, \mathcal{D}_{\tau}^{\text{trn}}\right) =?$$

Baseline methods – ERM

General formulations (empirical loss)

ERM

$$\mathcal{L}^{\text{er}}(\theta_0, \mathcal{D}) = \frac{1}{T} \sum_{\tau=1}^{T} \ell_\tau(\theta_0, \mathcal{D}_{\tau,N})$$

$\theta_0(\theta_1 \;\; \theta_2 \;\; \theta_3)$

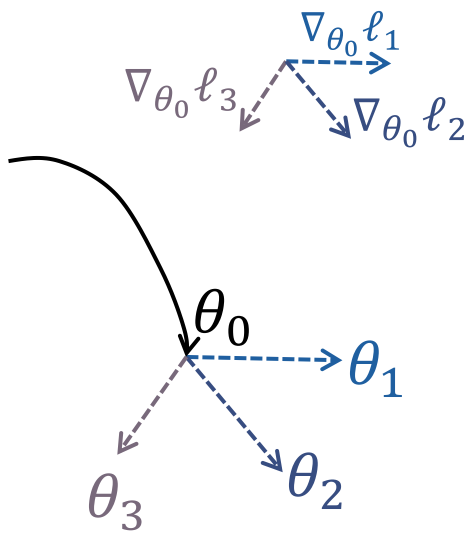
Baseline methods – MAML

General formulations (empirical loss)

MAML [Finn et al '17]

$$\mathcal{L}^{\text{ma}}(\theta_0, \mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau\left(\hat{\theta}_\tau^{\text{ma}}(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}}), \mathcal{D}_{\tau,N_2}^{\text{val}}\right)$$

$$\text{s.t. } \hat{\theta}_\tau^{\text{ma}}(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}}) = \theta_0 - \alpha\nabla_{\theta_0}\ell_\tau(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}})$$



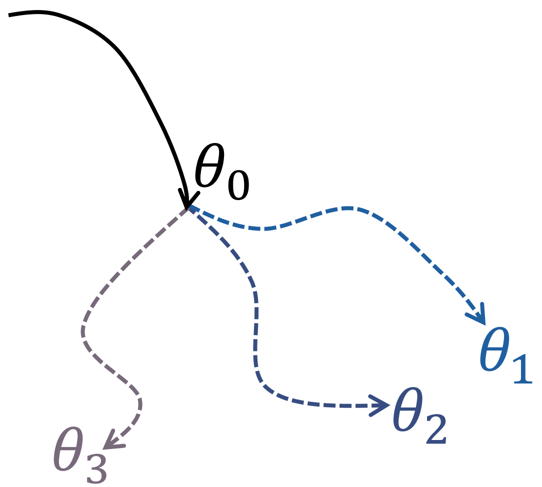
Baseline methods – iMAML

General formulations (empirical loss)

iMAML [Rajeswaran et al '19]

$$\mathcal{L}^{\text{im}}(\theta_0, \mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau\Big(\hat{\theta}_\tau^{\text{im}}\big(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}}\big), \mathcal{D}_{\tau,N_2}^{\text{val}}\Big)$$

$$\text{s.t. } \hat{\theta}_\tau^{\text{im}}\big(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}}\big) = \arg\min_{\theta_\tau} -\log p\big(\theta_\tau \mid \mathcal{D}_{\tau,N_1}^{\text{trn}}, \theta_0\big)$$

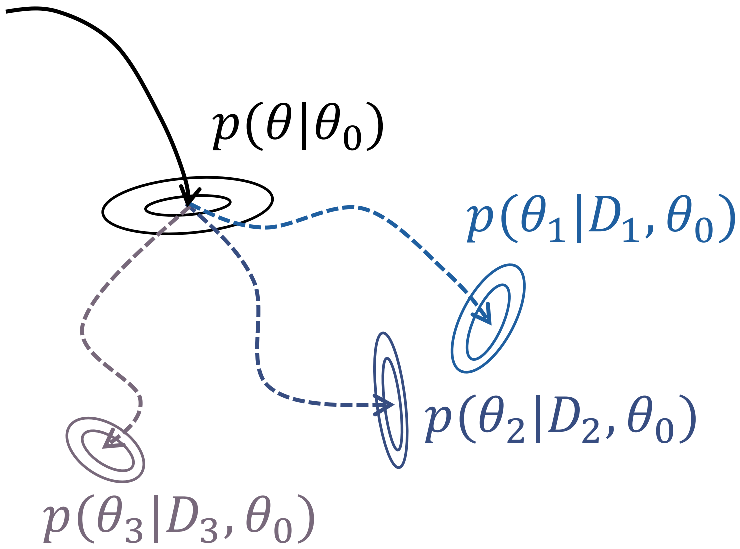


Baseline methods – BaMAML

General formulations (empirical loss)

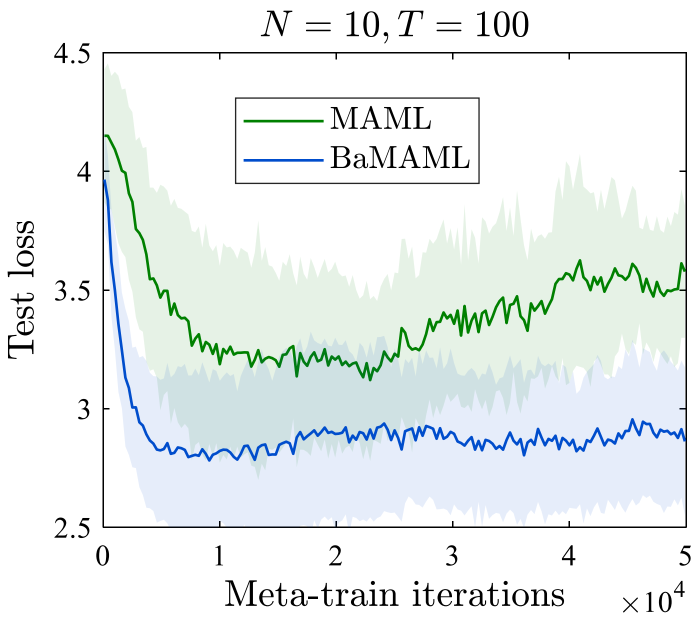
BaMAML

$$\mathcal{L}^{\mathrm{ba}}(\theta_0, \mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau\big(\hat{p}\big(\theta_\tau \mid \mathcal{D}^{\mathrm{trn}}_{\tau,N_1}, \theta_0\big), \mathcal{D}^{\mathrm{val}}_{\tau,N_2}\big)$$
$$\text{s.t. } \hat{p}\big(\theta_\tau \mid \mathcal{D}^{\mathrm{trn}}_{\tau,N_1}, \theta_0\big) = \mathop{\arg\min}_{q(\theta_\tau)\in\mathcal{Q}} D_{\mathrm{KL}}\big(q(\theta_\tau) \| p\big(\theta_\tau \mid \mathcal{D}^{\mathrm{trn}}_{\tau,N_1}, \theta_0\big)\big)$$



LLAMA [Grant et al '18]
PLATIPUS [Finn et al '18]
BMAML [Yoon et al '18]
VAMPIRE [Nguyen et al '20]

Compare the baseline methods



Sinusoidal regression

MiniImageNet classification accuracy

Method	1-shot 5-way
MAML	48.70 ± 1.84
iMAML	49.30 ± 1.88
BaMAML	51.54 ± 0.74

Empirically, BaMAML has better accuracy but is more challenging to solve than MAML.

Goal of this work

Questions:

- ❑ If and when is BaMAML better than MAML, provably?
- ❑ What are the decomposable factors that make BaMAML better?

Contributions:

First theoretical understanding for above questions.

A unified view

$$\ell_\tau(\theta_0, \mathcal{D}_\tau) = -\log p\left(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \theta_0, \mathcal{D}_\tau^{\text{trn}}\right)$$

$$= -\log \int \underbrace{p\left(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \theta_\tau\right)}_{\text{Likelihood}} \underbrace{p\left(\theta_\tau \mid \theta_0, \mathcal{D}_\tau^{\text{trn}}\right)}_{\text{Posterior}} d\theta_\tau$$

Bayes rule

$$p\left(\theta_\tau \mid \theta_0, \mathcal{D}_\tau^{\text{trn}}\right) = \frac{p(\mathcal{D}_\tau^{\text{trn}} \mid \theta_\tau) p(\theta_\tau \mid \theta_0)}{p(\mathcal{D}_\tau^{\text{trn}} \mid \theta_0)}$$

A unified view

$$\ell_\tau(\theta_0, \mathcal{D}_\tau) = -\log \int p\left(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \theta_\tau\right) p\left(\theta_\tau \mid \theta_0, \mathcal{D}_\tau^{\text{trn}}\right) d\theta_\tau$$

Point estimate ↓ e.g. MAML, iMAML

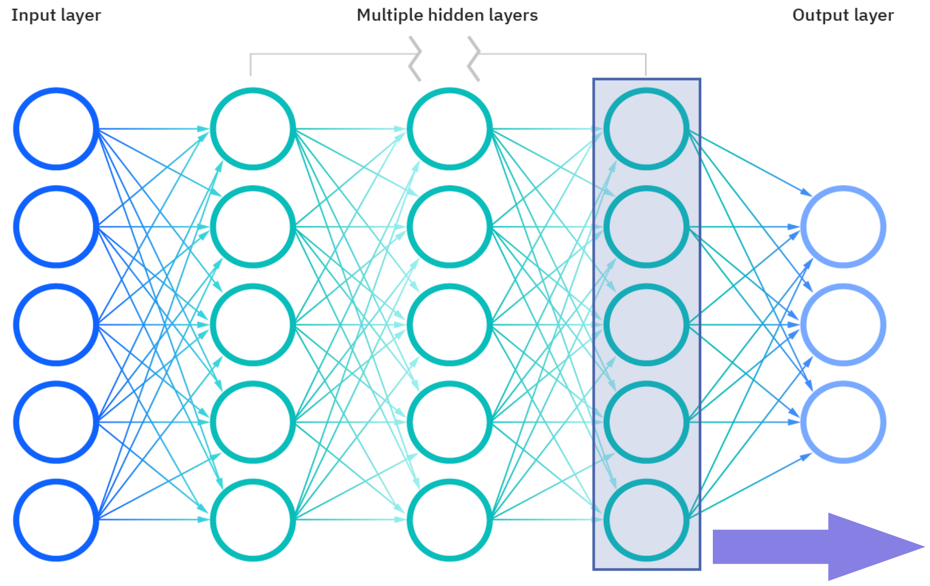
$$\delta\left(\theta_\tau - \hat{\theta}_\tau^{\mathcal{A}}\right)$$

↓

$$\ell_\tau(\theta_0, \mathcal{D}_\tau) = -\log p\left(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \hat{\theta}_\tau^{\mathcal{A}}(\theta_0, \mathcal{D}_\tau^{\text{trn}})\right)$$

Now we are ready to compare these methods in the same framework.

Meta linear regression – data model



Treat this feature as input $\mathbf{x}$

Prior work
[Denevi ’18]
[Gao et al ’20]
[Bai et al ’21]
…

Data model

$$y_\tau = \theta_\tau^{\text{gt}\top} \mathbf{x}_\tau + \epsilon_\tau, \text{ with } \quad \epsilon_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right) \quad \mathbf{Q}_\tau = \mathbb{E}\left[\mathbf{x}_\tau \mathbf{x}_\tau^\top \mid \tau\right].$$

Loss function under meta linear regression

$$y_\tau = \theta_\tau^{\text{gt}\top} \mathbf{x}_\tau + \epsilon_\tau, \text{ with } \quad \epsilon_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right) \quad \mathbf{Q}_\tau = \mathbb{E}\left[\mathbf{x}_\tau \mathbf{x}_\tau^\top \mid \tau\right].$$

Recall $\ell_\tau(\theta_0, \mathcal{D}_\tau) = -\log \int p\big(\mathbf{y}_\tau^{\text{val}} \mid \mathbf{X}_\tau^{\text{val}}, \theta_\tau\big) p\big(\theta_\tau \mid \theta_0, \mathcal{D}_\tau^{\text{trn}}\big) d\theta_\tau$

Assumption: Given $\hat{\theta}_\tau^{\mathcal{A}}$, $p(y_\tau \mid \mathbf{x}_\tau, \hat{\theta}_\tau^{\mathcal{A}}) = \mathcal{N}(\hat{\theta}_\tau^{\mathcal{A}\top} \mathbf{x}_\tau, \sigma_\tau^2)$

Bayes rule

$$p(\theta_\tau \mid \mathcal{D}_\tau^{\text{trn}}, \theta_0) \propto p\big(\mathcal{D}_\tau^{\text{trn}} \mid \theta_\tau\big) p(\theta_\tau \mid \theta_0),$$

Prior is Gaussian $p(\theta_\tau \mid \theta_0) \propto \exp\Big\{-\gamma \|\theta_\tau - \theta_0\|_2^2\Big\}$

γ is the weight of the prior

Basic assumptions

1. **(Bounded eigenvalues)** For any $\boldsymbol{\tau}, 0 < \underline{\boldsymbol{\lambda}} \leq \lambda(\mathbf{Q}_{\tau}) \leq \bar{\lambda}$

2. **(Ground truth task parameter distribution)**

   1) $\theta_{\tau}^{\mathrm{gt}}$ is independent of $\mathbf{X}_{\tau}$.

   2) the individual entries $\left\{\theta_{\tau,i}^{\mathrm{gt}}\right\}_{i\in[d],\tau\in[T]}$ are independent and $\mathcal{O}\left(R/\sqrt{d}\right)$-sub-Gaussian, where R is a constant.

   3) $\left\|\mathbb{E}\left[\theta_{\tau}^{\mathrm{gt}}\right]\right\| \leq M$.

Meta linear regression – Risk decomposition

Meta test risk decomposition

$$\mathcal{R}^{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}}) = \underbrace{\mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0^{\mathcal{A}})}_{\text{optimal population risk}} + \underbrace{\left\|\hat{\boldsymbol{\theta}}_0^{\mathcal{A}} - \boldsymbol{\theta}_0^{\mathcal{A}}\right\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}}_{\text{statistical error } \mathcal{E}^2_{\mathcal{A}}(\hat{\boldsymbol{\theta}}_0^{\mathcal{A}})}$$

$$\theta_0^{\mathcal{A}} := \arg\min_{\theta_0} \mathcal{R}^{\mathcal{A}}(\theta_0)$$

Analyze optimal population risk and statistical error separately.

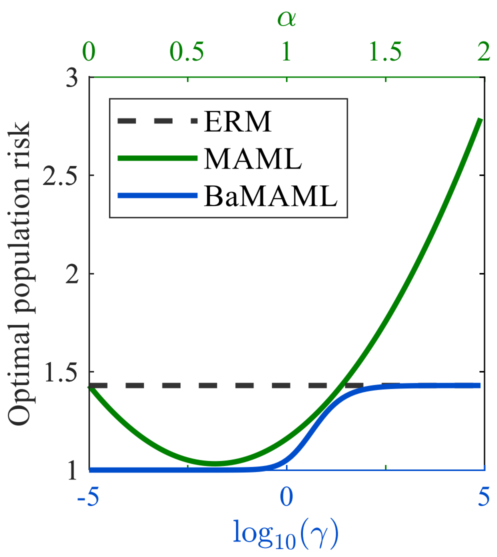
Comparison of optimal population risk

Theorem 1 (informal)

Under Assumptions 1-2,

ERM vs MAML
Can find α in a range that
$\mathcal{R}^{\mathrm{ma}}(\theta_0^{\mathrm{ma}}) < \mathcal{R}^{\mathrm{er}}(\theta_0^{\mathrm{er}})$.

MAML vs BaMAML
Can find γ in a range that
$\mathcal{R}^{\mathrm{ba}}(\theta_0^{\mathrm{ba}}) < \mathcal{R}^{\mathrm{ma}}(\theta_0^{\mathrm{ma}})$.



- Essentially,

$$\mathcal{R}^{\mathrm{er}}(\theta_0^{\mathrm{er}}) > \inf_{\alpha} \mathcal{R}^{\mathrm{ma}}(\theta_0^{\mathrm{ma}}; \alpha) > \inf_{\gamma} \mathcal{R}^{\mathrm{ba}}(\theta_0^{\mathrm{ba}}; \gamma)$$

- If α not properly chosen, MAML can be worse than ERM, but not for iMAML.
- Choice of γ reflects trade-off between adaptation speed and adaptation performance.

Precise characterization of statistical error

Assumption 3 (Linear centroid model).

1) $\mathbf{x}_\tau \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_\text{d})$
2) $\text{Cov}_{\theta_\tau^{\text{gt}}}\left[\theta_\tau^{\text{gt}}\right] = \frac{R^2}{d}\mathbf{I}_\text{d}.$

Implications

Assumption 3 (1) assumes $\mathbf{Q}_\tau = \mathbf{I}_d$, therefore $\mathbf{W}_\tau^{\mathcal{A}} = w_{\mathcal{A}}\mathbf{I}_d$.
This implies for different methods $\mathcal{A}, \theta_0^{\mathcal{A}} = \mathbb{E}_\tau\left[\theta_\tau^{\text{gt}}\right]$.

Precise characterization of statistical error

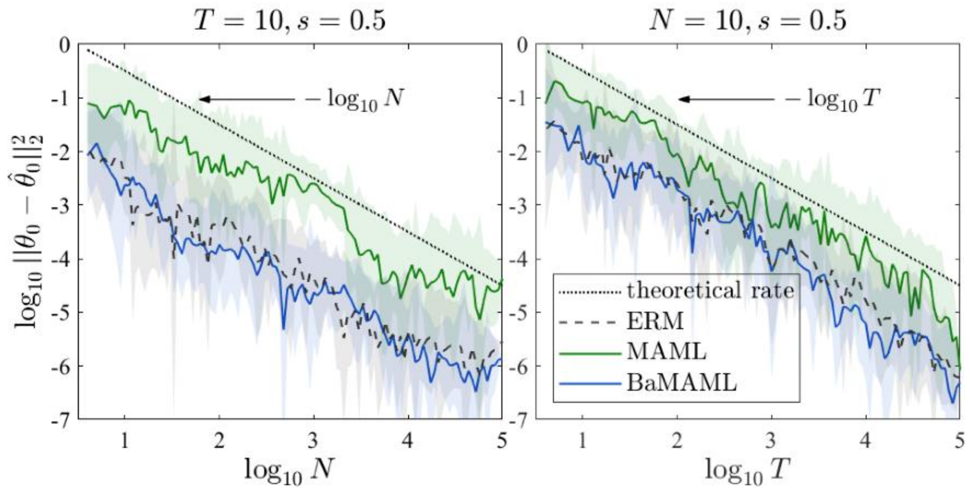
Theorem 2 (informal)

Define $C^{\mathcal{A}} := \frac{1}{d}\left\langle \mathbb{E}^{-2}\left[\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\right], \mathbb{E}\left[\left(\hat{\mathbf{W}}_{\tau,N}^{\mathcal{A}}\right)^2\right]\right\rangle$

ϱ as higher order term.

Under Assumptions 1-3, the following hold with high probability

$$w_{\mathcal{A}}\left\|\theta_0 - \hat{\theta}_0\right\|_2^2 = \frac{R^2}{T}\left(w_{\mathcal{A}}C^{\mathcal{A}} + \widetilde{\mathcal{O}}\left(\sqrt{\frac{d}{T}}\right) + \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right)\right) + \varrho$$



Now we are ready to quantify the dominating constant exactly.

Sharp comparison of statistical error

- Limits of dominating constants

$$\inf_{\substack{\alpha > 0 \\ s \in (0,1)}} \lim_{\substack{d, N \to \infty \\ d/N \to \eta}} C^{\mathrm{ma}} = \inf_{\substack{\gamma > 0 \\ s \in (0,1)}} \lim_{\substack{d, N \to \infty \\ d/N \to \eta}} C^{\mathrm{im}} = 1 + \eta.$$

$$\inf_{\substack{\gamma > 0 \\ s \in (0,1)}} \lim_{\substack{d, N \to \infty \\ d/N \to \eta}} C^{\mathrm{ba}} \begin{cases} = 1, & \eta \leq 1 \\ \leq \eta, & \eta > 1 \end{cases}$$

- Under linear centroid model, the dominating constant in the statistical error with optimally tuned hyperparameters satisfies,

BaMAML < MAML = iMAML

Concluding remarks

- BaMAML (and iMAML) has better adaptation flexibility than one-step MAML, leading to smaller optimal population risk.

- For statistical error, all methods have the same dependence on N, T, d, and their difference lies in the constant. BaMAML has better constant than point estimate (iMAML & MAML) due to model averaging.

- Under linear centroid model, the dominating constant in the statistical error with optimally tuned hyperparameters satisfies,

$$\text{BaMAML} \leq \text{MAML} = \text{iMAML}$$

Justify the theoretical benefits of BaMAML over MAML.

References

Lisha Chen and Tianyi Chen. ``Is Bayesian Model-Agnostic Meta Learning Better than Model-Agnostic Meta Learning, Provably?,” ***Proc. of AISTATS***, 2022.

Yu Bai et al. ``How Important is the Train-Validation Split in Meta-Learning?,” ***Proc. of ICML***, 2021.

Katelyn Gao and Ozan Sener. ``Modeling and Optimization Trade-off in Meta-learning,” ***Proc. of NeurIPS***, 2020.

Erin Grant et al. ``Recasting Gradient-Based Meta-Learning as Hierarchical Bayes,” ***Proc. of ICLR***, 2018.

Taesup Kim et al. ``Bayesian Model-Agnostic Meta-Learning,” ***Proc. of NIPS***, 2018.

Thank you!

Contact: chenl21@rpi.edu

Code: https://github.com/lisha-chen/Bayesian-MAML-vs-MAML