Is Bayesian model agnostic meta learning better than model agnostic meta learning, provably?

Lisha Chen
Tianyi Chen

MOTIVATION

Learning with big data

Challenges:

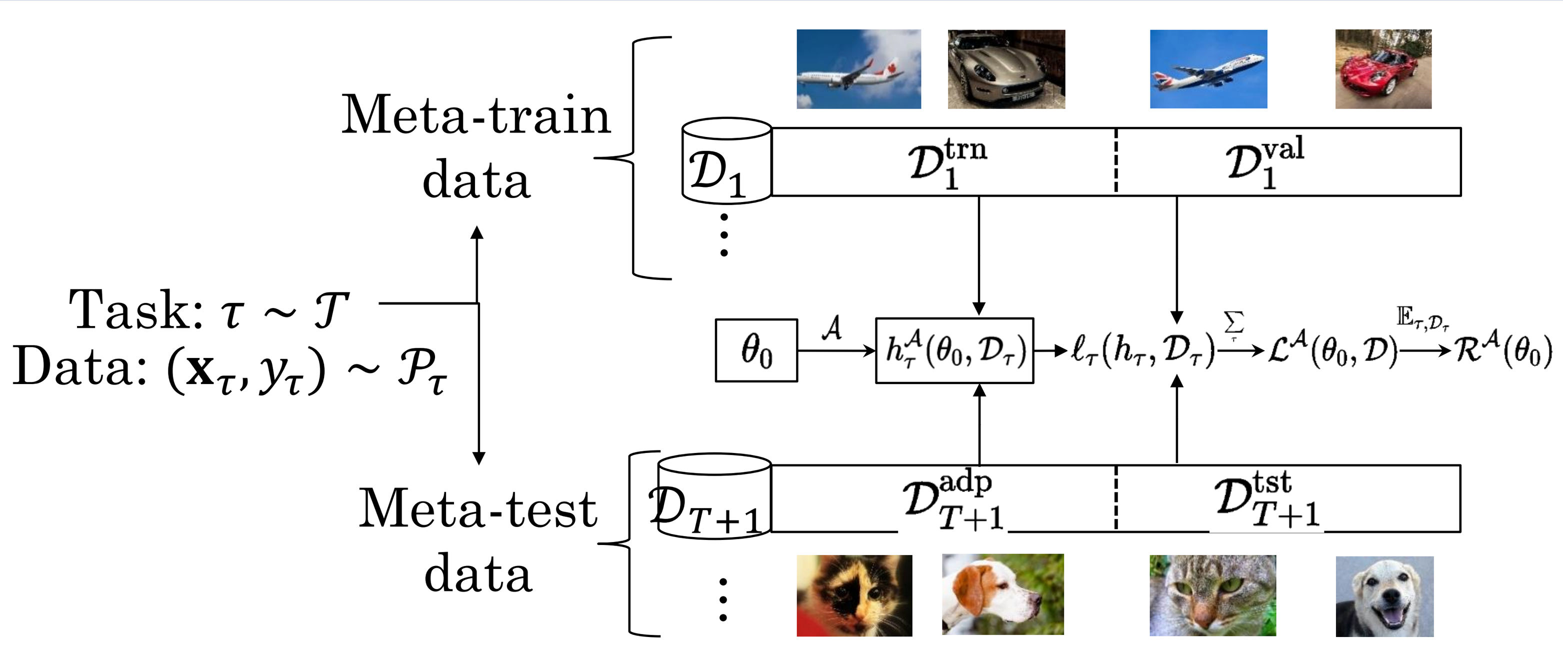
- Requires laborious data collection and/or annotation.
- May not generalize well to unseen domains.

Real world problems without large-scale supervised data

- e.g. medical decision making, personalization, etc.

We need methods without large-scale supervised data

META LEARNING SETUP



Empirical loss $\mathcal{L}^{\mathcal{A}}(\theta_0, \mathcal{D}) := \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau(\mathcal{A}(\theta_0, \mathcal{D}_\tau^{\text{trn}}), \mathcal{D}_\tau^{\text{val}})$.

Population risk $\mathcal{R}^{\mathcal{A}}(\theta_0) := \mathbb{E}_\tau\left[\mathbb{E}_{\mathcal{D}_\tau}\left[\ell_\tau(\mathcal{A}(\theta_0, \mathcal{D}_\tau^{\text{trn}}), \mathcal{D}_\tau^{\text{val}})\right]\right]$.

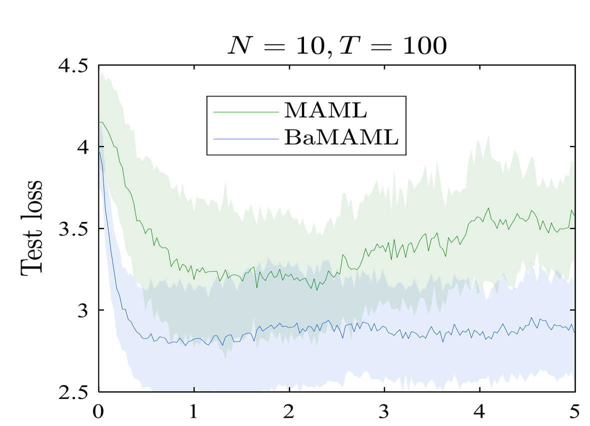
Probabilistic perspective

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

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

OUR GOAL

Bayesian model agnostic meta learning (BaMAML) exhibits better performance than MAML.



Sinusoidal regression

MiniImageNet classification

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

Theoretical understanding to this behavior is limited.

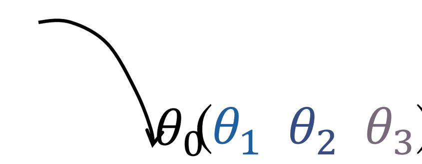
We want to study

Is Bayesian model agnostic meta learning better than model agnostic meta learning, provably?

BASELINES

ERM

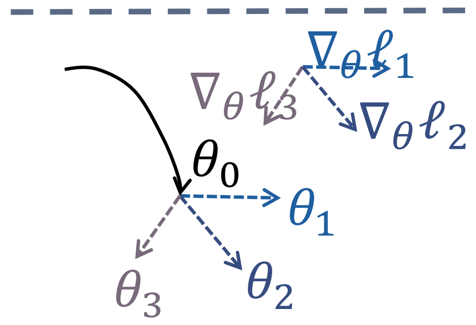
$$\mathcal{L}^{\text{er}}(\theta_0, \mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau(\theta_0, \mathcal{D}_{\tau,N})$$



MAML [Finn et al '17]

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

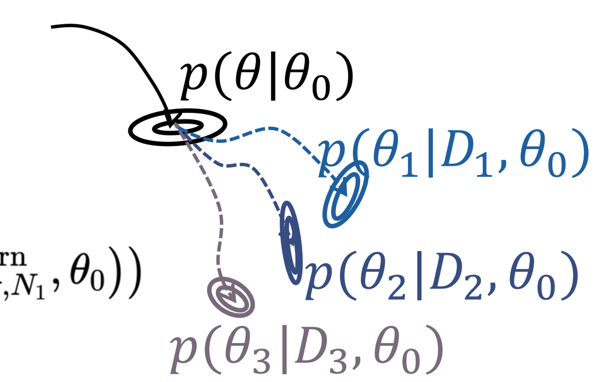
$$\text{s.t. } \hat{\theta}_\tau^{\text{ma}}(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}}) = \theta_0 - \alpha \nabla_{\theta_0} \ell_\tau(\theta_0, \mathcal{D}_{\tau,N_1}^{\text{trn}})$$



BaMAML [Grant et al '18, Yoon et al '18]

$$\mathcal{L}^{\text{ba}}(\theta_0, \mathcal{D}) = \frac{1}{T}\sum_{\tau=1}^{T} \ell_\tau(\hat{p}(\theta_\tau \mid \mathcal{D}_{\tau,N_1}^{\text{trn}}, \theta_0), \mathcal{D}_{\tau,N_2}^{\text{val}})$$

$$\text{s.t. } \hat{p}(\theta_\tau \mid \mathcal{D}_{\tau,N_1}^{\text{trn}}, \theta_0) = \arg\min_{q(\theta_\tau) \in \mathcal{Q}} D_{\text{KL}}\left(q(\theta_\tau) \| p(\theta_\tau \mid \mathcal{D}_{\tau,N_1}^{\text{trn}}, \theta_0)\right)$$



META LINEAR REGRESSION

Data model

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

Assumptions

1. Bounded eigenvalues of data covariance
2. Sub-gaussian ground truth task parameter
3. (Linear centroid model) 1) $\mathbf{x}_\tau \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$ 2) $\text{Cov}_{\theta_\tau^{\text{gt}}}[\theta_\tau^{\text{gt}}] = \frac{R^2}{d}\mathbf{I}_d$.

Meta test risk decomposition

$$\mathcal{R}^{\mathcal{A}}(\hat{\theta}_0^{\mathcal{A}}) = \underbrace{\mathcal{R}^{\mathcal{A}}(\theta_0^{\mathcal{A}})}_{\text{optimal popultation risk}} + \underbrace{\left\|\hat{\theta}_0^{\mathcal{A}} - \theta_0^{\mathcal{A}}\right\|^2_{\mathbb{E}_\tau[\mathbf{W}_\tau^{\mathcal{A}}]}}_{\text{statistical error } \mathcal{E}_{\mathcal{A}}^2(\hat{\theta}_0^{\mathcal{A}})}$$

POPULATION RISK

Theorem 1 (informal)

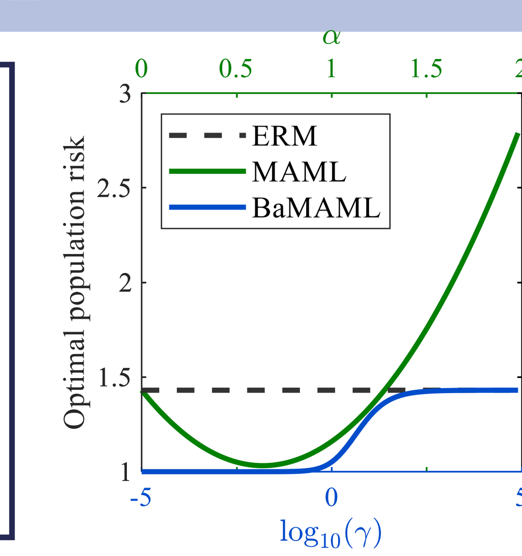
Under Assumptions 1-2,

ERM vs MAML

Can find α that $\mathcal{R}^{\text{ma}}(\theta_0^{\text{ma}}) < \mathcal{R}^{\text{er}}(\theta_0^{\text{er}})$.

MAML vs iMAML & BaMAML

Can find γ that $\mathcal{R}^{\text{ba}}(\theta_0^{\text{ba}}) < \mathcal{R}^{\text{ma}}(\theta_0^{\text{ma}})$.



- ❑ $\mathcal{R}^{\text{er}}(\theta_0^{\text{er}}) > \inf_\alpha \mathcal{R}^{\text{ma}}(\theta_0^{\text{ma}}; \alpha) > \inf_\gamma \mathcal{R}^{\text{ba}}(\theta_0^{\text{ba}}; \gamma)$
- ❑ If α not properly chosen, MAML can be worse than ERM, but not for BaMAML.
- ❑ Choice of γ reflects trade-off between adaptation speed & adaptation performance.

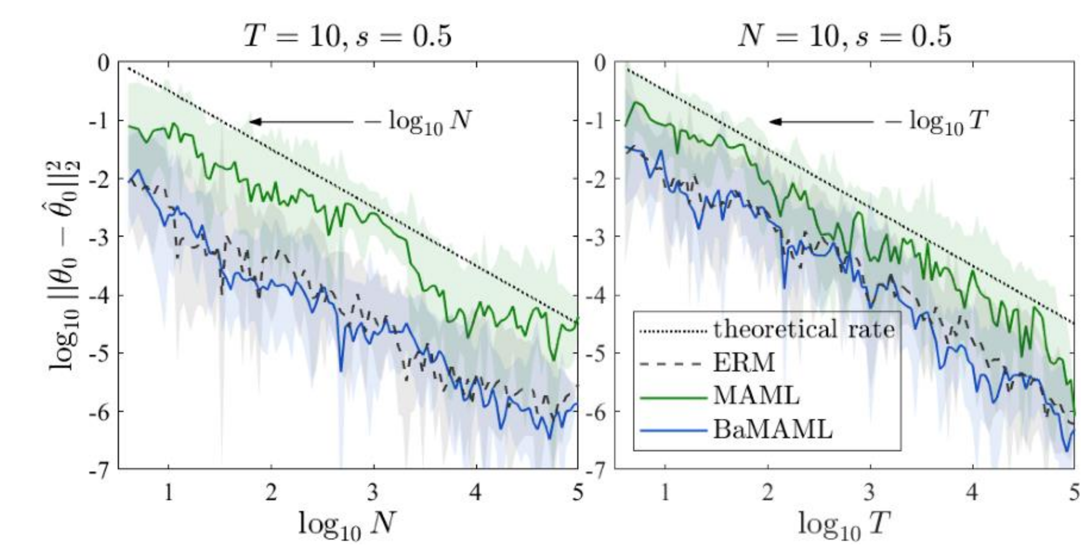
STATISTICAL ERROR

Theorem 2 (informal)

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

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

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



- ❑ Limits of dominating constants

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

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

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

BaMAML < MAML = iMAML

PROOF TECHNIQUE

To characterize statistical error

Sub-Gaussian concentration inequality

$$w_{\mathcal{A}}\left\|\theta_0 - \hat{\theta}_0\right\|_2^2 = \underbrace{\left\{w_{\mathcal{A}}\left\|\theta_0 - \hat{\theta}_0\right\|_2^2 - \mathbb{E}\left[w_{\mathcal{A}}\left\|\theta_0 - \hat{\theta}_0\right\|_2^2\right]\right\}}_{\text{Higher order based on Hanson-Wright inequality}} + \underbrace{\mathbb{E}\left[w_{\mathcal{A}}\left\|\theta_0 - \hat{\theta}_0\right\|_2^2\right]}_{\text{Contains the dominating constant}}$$

To compute the dominating constant

Stieltjes transform

Stieltjes form of the Marchenko-Pastur law

$$s(\omega_1, \omega_2) := \lim_{d,N \to \infty, d/N \to \eta} \frac{1}{d}\mathbb{E}\left[\text{tr}\left(\left(\omega_1 \mathbf{I}_d + \omega_2 \hat{\mathbf{Q}}_N\right)^{-1}\right)\right] \Rightarrow$$

Need this to compute the dominating constant

REFERENCES

Y. Bai et al. ``How Important is the Train-Validation Split in Meta-Learning?,'' *ICML*, 2021.
K. Gao and O. Sener. ``Modeling and Optimization Trade-off in Meta-learning,'' *NeurIPS*, 2020.
E. Grant et al. ``Recasting Gradient-Based Meta-Learning as Hierarchical Bayes,'' *ICLR*, 2018.
T. Kim et al. ``Bayesian Model-Agnostic Meta-Learning,'' *NIPS*, 2018.
C. Finn et al. ``Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,'' *ICML*, 2017.