

# 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]

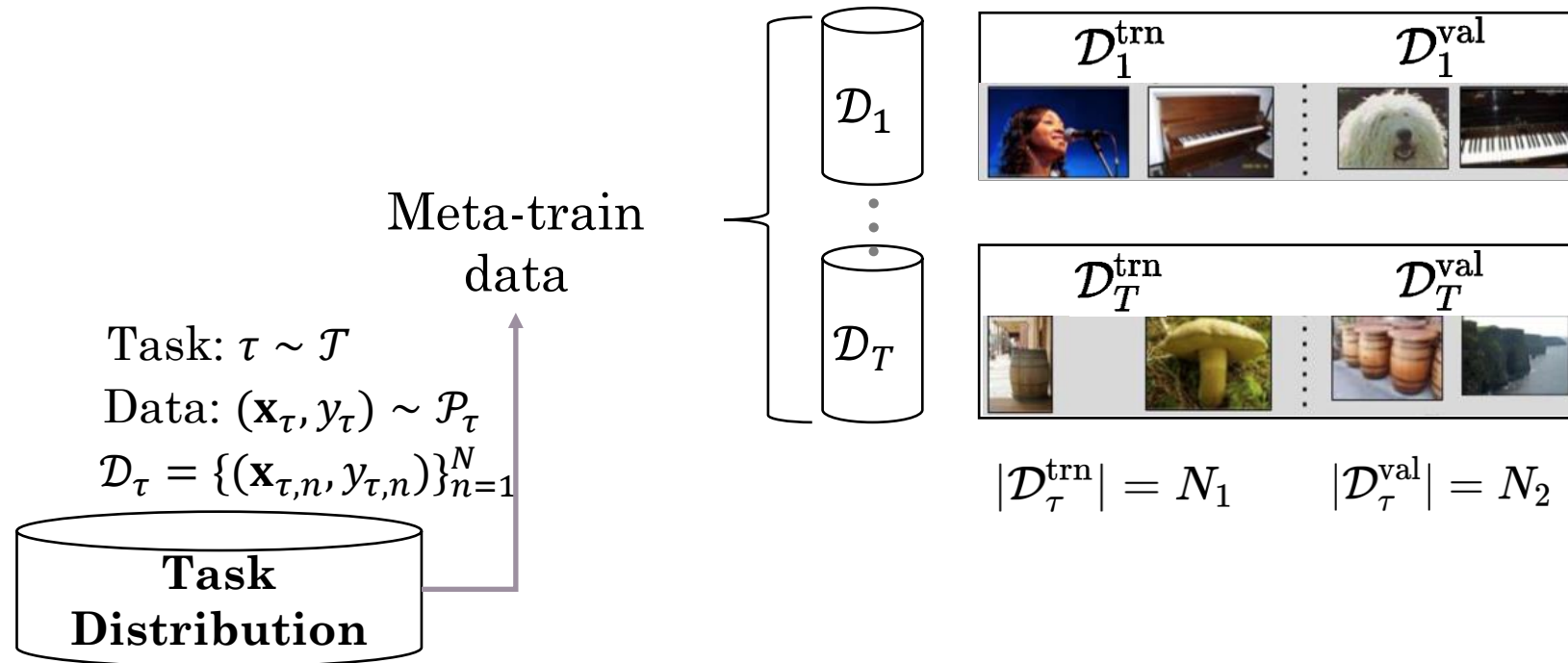
[Vinyals et al '16]

[Santoro et al '16]

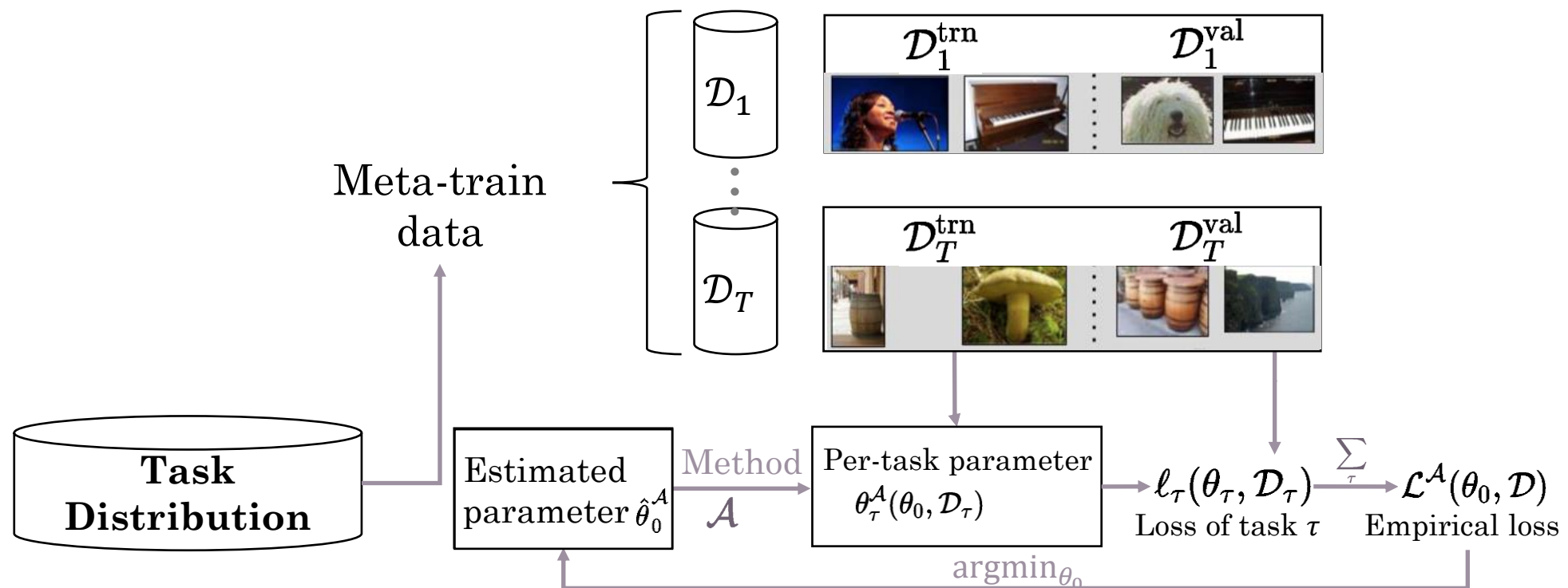
[Wichrowska et al '17]

[Finn et al '17]

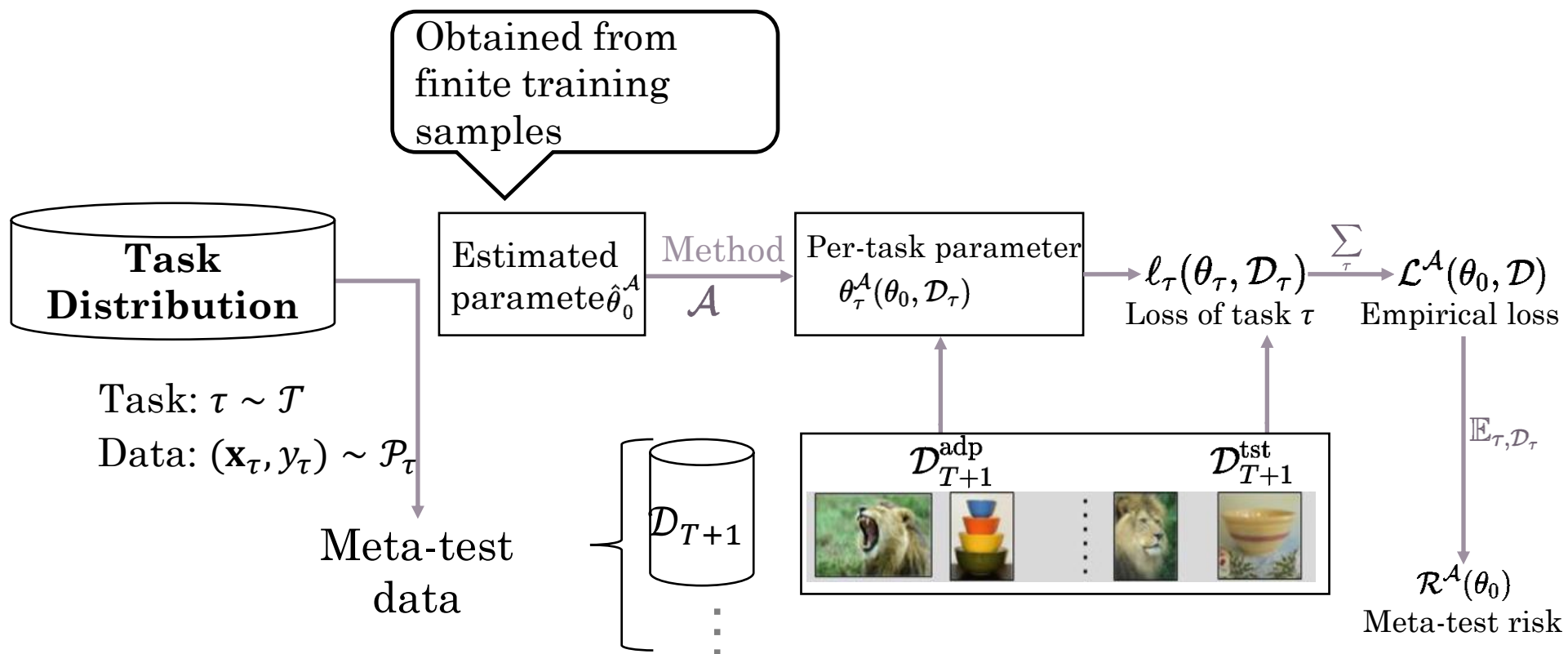
# Meta learning setup



# Meta learning setup

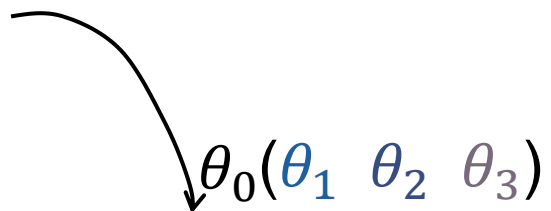


# Meta learning setup

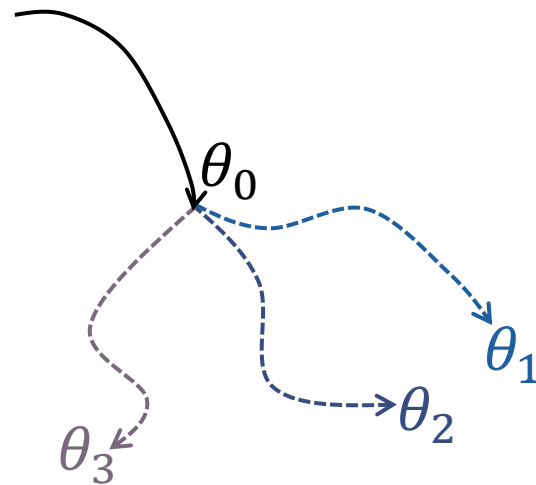


# Basis of comparison

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



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



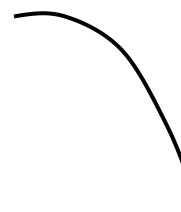
$$\hat{\theta}_{\tau}^{\mathcal{A}}(\theta_0, \mathcal{D}_{\tau}^{\text{trn}}) = ?$$

# Baseline methods – ERM

## General formulations (empirical loss)

ERM

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

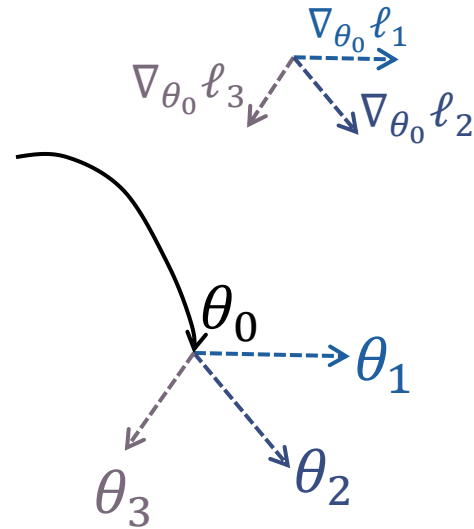
  $\theta_0(\theta_1 \ \theta_2 \ \theta_3)$

# Baseline methods – MAML

## General formulations (empirical loss)

MAML [Finn et al '17]

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



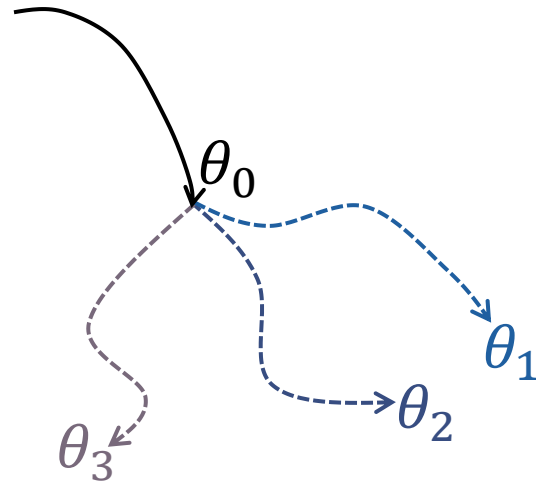


# Baseline methods – iMAML

## General formulations (empirical loss)

iMAML [Rajeswaran et al '19]

$$\mathcal{L}^{\text{im}}(\theta_0, \mathcal{D}) = \frac{1}{T} \sum_{\tau=1}^T \ell_{\tau} \left( \hat{\theta}_{\tau}^{\text{im}}(\theta_0, \mathcal{D}_{\tau, N_1}^{\text{trn}}), \mathcal{D}_{\tau, N_2}^{\text{val}} \right)$$
$$\text{s.t. } \hat{\theta}_{\tau}^{\text{im}}(\theta_0, \mathcal{D}_{\tau, N_1}^{\text{trn}}) = \arg \min_{\theta_{\tau}} -\log p(\theta_{\tau} \mid \mathcal{D}_{\tau, N_1}^{\text{trn}}, \theta_0)$$

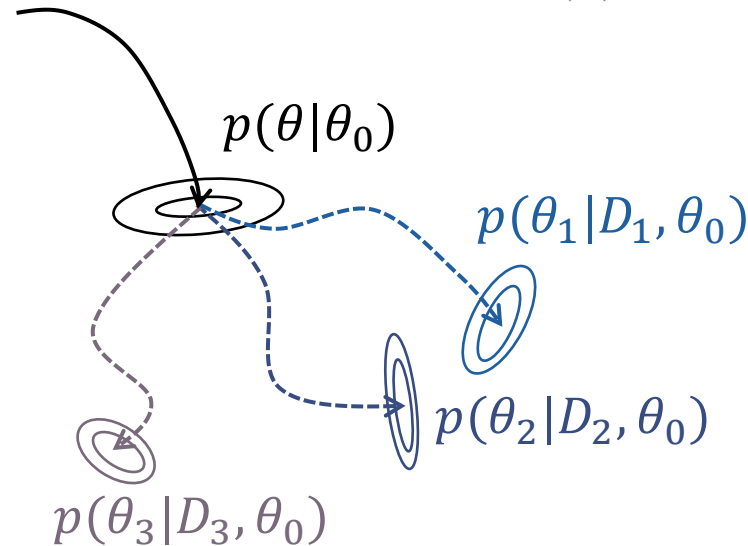


# Baseline methods – BaMAML

## General formulations (empirical loss)

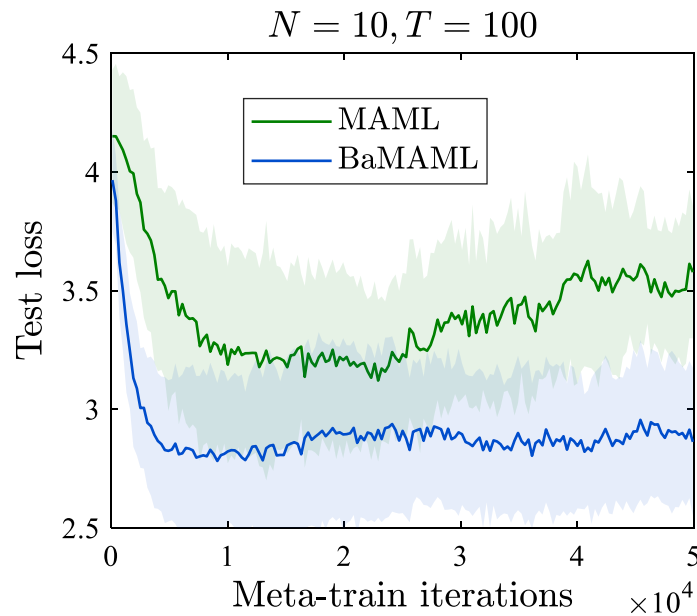
### BaMAML

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



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	$48.70 \pm 1.84$
iMAML	$49.30 \pm 1.88$
BaMAML	$51.54 \pm 0.74$

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

# Goal of this work

## Questions:

- ❑ 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

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

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

Posterior

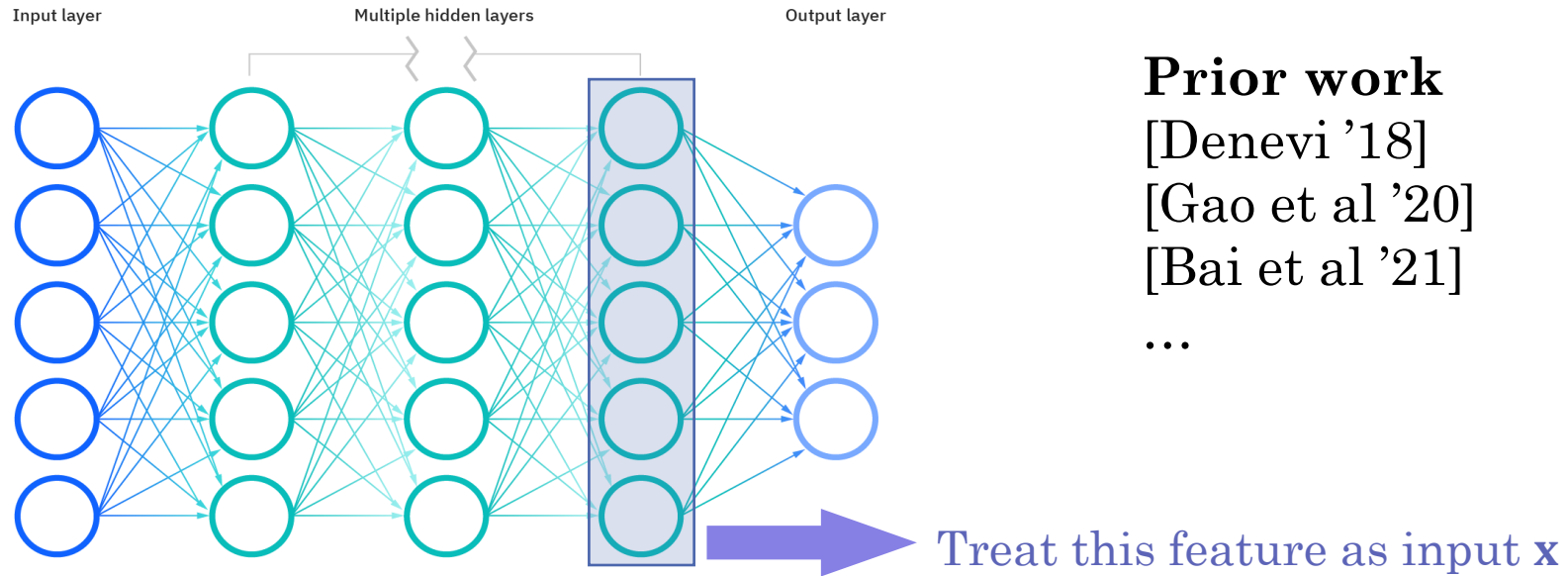


Bayes rule

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

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

# Meta linear regression – data model



Data model

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

# Loss function under meta linear regression

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

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

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

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

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

$\gamma$  is the weight of the prior

# Basic assumptions

1. **(Bounded eigenvalues)** For any  $\tau, 0 < \underline{\lambda} \leq \lambda(\mathbf{Q}_\tau) \leq \bar{\lambda}$
2. **(Ground truth task parameter)**
  - 1)  $\theta_\tau^{\text{gt}}$  is independent of  $\mathbf{X}_\tau$ .
  - 2) the individual entries  $\left\{ \theta_{\tau,i}^{\text{gt}} \right\}_{i \in [d], \tau \in [T]}$  are independent and  $\mathcal{O}(R/\sqrt{d})$ -sub-Gaussian, where  $R$  is a constant.
  - 3)  $\|\mathbb{E}[\theta_\tau^{\text{gt}}]\| \leq M$ .



# Meta linear regression – Risk decomposition

Meta test risk decomposition

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

$$\boldsymbol{\theta}_0^{\mathcal{A}} := \arg \min_{\boldsymbol{\theta}_0} \mathcal{R}^{\mathcal{A}}(\boldsymbol{\theta}_0)$$

Analyze optimal population risk and statistical error separately.

# Comparison of optimal population risk

## Theorem 1 (informal)

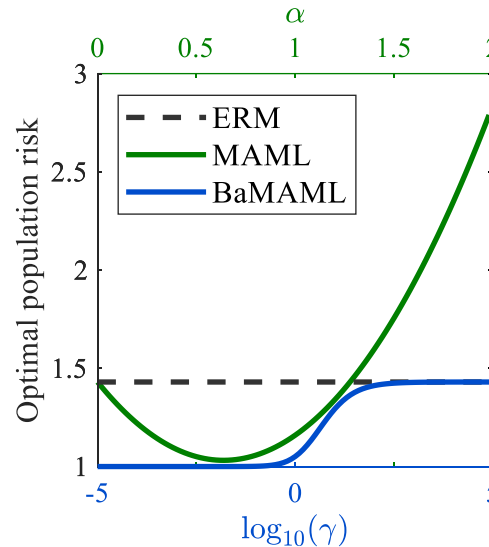
Under Assumptions 1-2,

### ERM vs MAML

Can find  $\alpha$  in a range that  
 $\mathcal{R}^{\text{ma}}(\theta_0^{\text{ma}}) < \mathcal{R}^{\text{er}}(\theta_0^{\text{er}})$ .

### MAML vs BaMAML

Can find  $\gamma$  in a range that  
 $\mathcal{R}^{\text{ba}}(\theta_0^{\text{ba}}) < \mathcal{R}^{\text{ma}}(\theta_0^{\text{ma}})$ .



□ Essentially,

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

□ If  $\alpha$  not properly chosen, MAML can be worse than ERM, but not for iMAML.

□ Choice of  $\gamma$  reflects trade-off between adaptation speed and adaptation performance.

# Precise characterization of statistical error

**Assumption 3 (Linear centroid model).**

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

## 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  $\mathcal{A}$ ,  $\theta_0^{\mathcal{A}} = \mathbb{E}_\tau[\theta_\tau^{\text{gt}}]$ .

# Precise characterization of statistical error

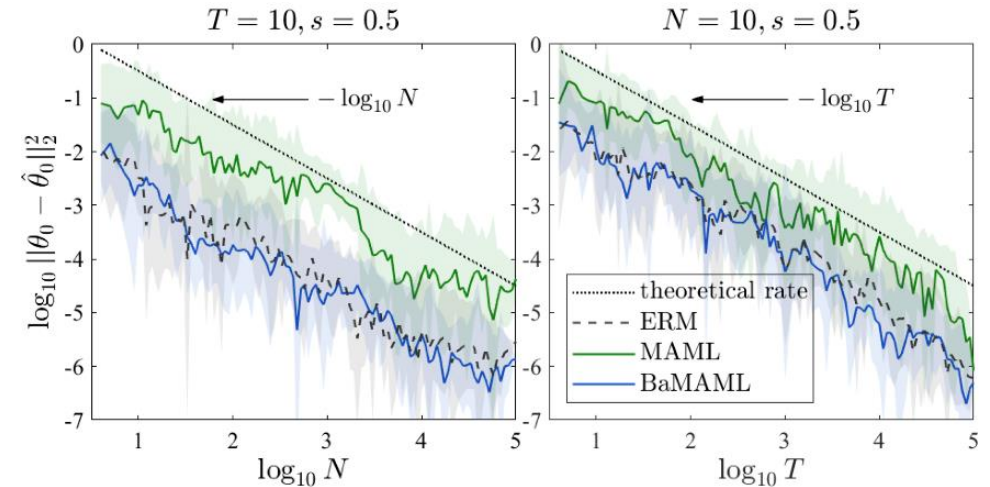
## Theorem 2 (informal)

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

$\varrho$  as higher order term.

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

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



Now we are ready to quantify the dominating constant exactly.

# Sharp comparison of statistical error

□ Limits of dominating constants

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

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

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

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

# 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>

