Bridging the Gap Between Practice and PAC-Bayes Theory in Few-Shot Meta-Learning

Nan Ding, Xi Chen, Tomer Levinboim,
Sebastian Goodman, Radu Soricut

Google AI

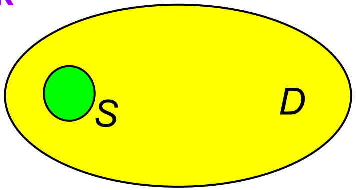
Outline

- Theory
  - Existing PAC-Bayesian bounds for Meta-Learning
  - Two new PAC-Bayesian bounds for Few-Shot Meta-Learning
- Practice
  - Connection to Reptile, MAML and PACOH
  - A new few-shot meta-learning algorithm: PACMAML
  - Empirical Results

Motivation

Supervised learning

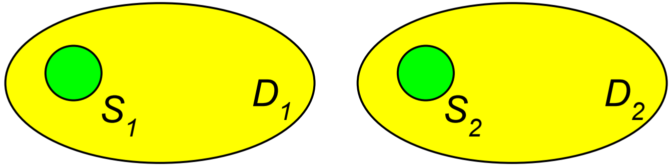
Target Task



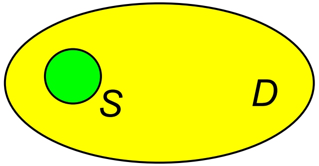
Theory:
Thm-1 [McAllester99]

Meta learning

Observed Tasks



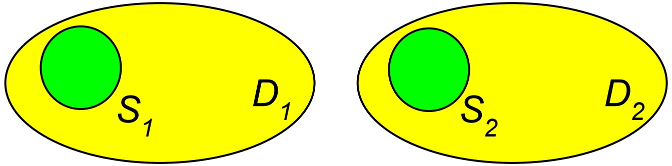
Target Task



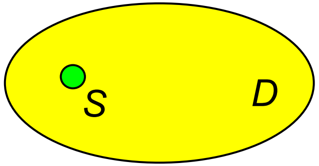
Theory:
Thm-2 [Pentina14, Rothfuss20]

Few-shot Meta learning

Observed Tasks



Target Task



Theory:
???

PAC-Bayes Bounds on Supervised Learning

Define: $S \sim D^m$, $z_i = (x_i, y_i)$, $h \sim Q(S, P)$

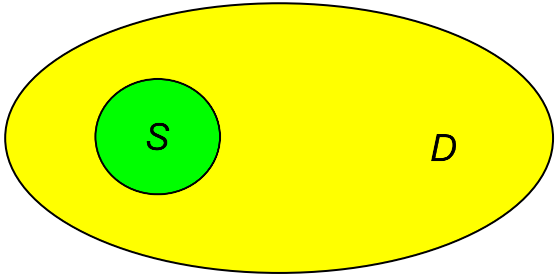
$$\hat{L}(h,S) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$$

$$\hat{L}(Q,S) = \mathbb{E}_{h\sim Q}\,\hat{L}(h,S)$$

Training loss of Q

$$L(Q,D) = \mathbb{E}_{h\sim Q}\,L(h,D)$$

Real loss of Q



Theorem 1 ([2, 12]) *Given a data distribution D, a hypothesis space H, a prior P, a confidence level $\delta \in (0,1]$, and $\beta > 0$, with probability at least $1-\delta$ over samples $S \sim D^m$, we have for all posterior Q,*

$$L(Q,D) \leq \hat{L}(Q,S) + \frac{1}{\beta}\left(D_{KL}(Q\|P) + \log\frac{1}{\delta}\right) + \frac{m}{\beta}\Psi\left(\frac{\beta}{m}\right) \qquad (1)$$

where $\Psi(\beta) = \log \mathbb{E}_{h\sim P}\,\mathbb{E}_{z\sim D} \exp(\beta(l(h,z) - L(h,D)))$.

Meta-Learning

Meta-training: $\mathcal{P}(P) \Rightarrow Q(P)$

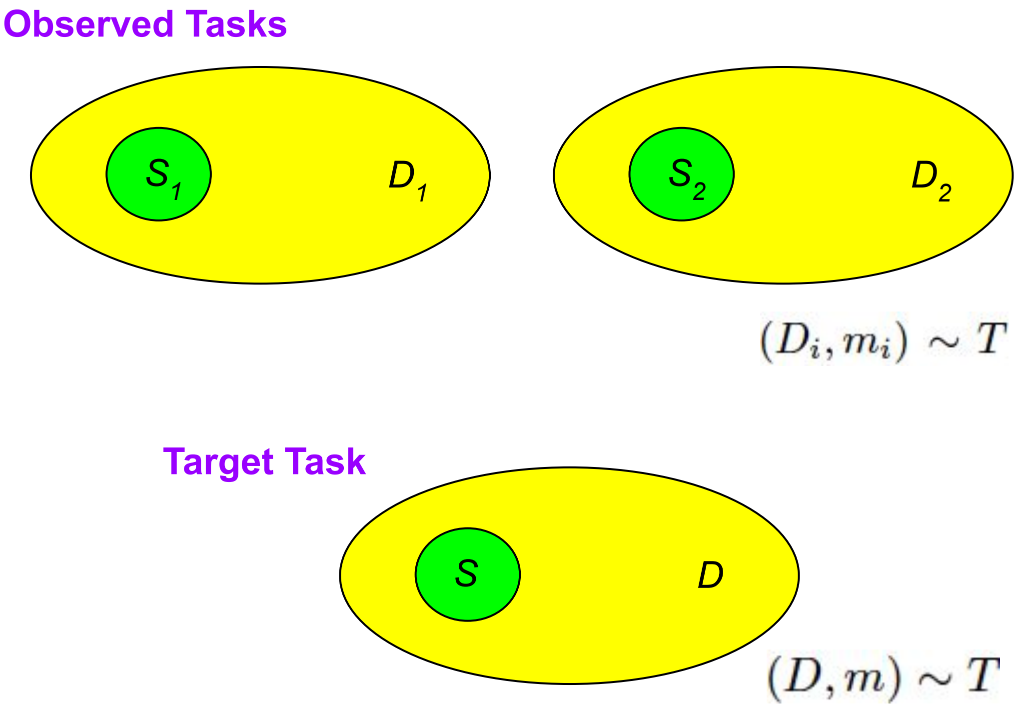
$S_i \sim D_i^{m_i}$

$(D_i, m_i) \sim T$

Meta-testing: $P \Rightarrow Q(S, P)$

$S \sim D^m$

$(D, m) \sim T$



Goal: learn $Q(P)$ over the prior distribution P for fast adaptation of the base-learner $Q(S, P)$ for the target task.

PAC-Bayes Bounds on Meta-Learning

Define:

$$\hat{R}(\mathcal{Q}, S_{i=1}^n) := \mathbb{E}_{P\sim\mathcal{Q}}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{L}(Q(S_i, P), S_i)\right]$$

Training loss of $\mathcal{Q}$

$$R(\mathcal{Q}, T) := \mathbb{E}_{P\sim\mathcal{Q}}\,\mathbb{E}_{(D,m)\sim T}\,\mathbb{E}_{S\sim D^m}\left[L(Q(S,P), D)\right]$$

Real loss of $\mathcal{Q}$

Theorem 2 ([18, 22]) *Given a task environment T and a set of n observed tasks $(D_i, m_i) \sim T$, let $\mathcal{P}$ be a fixed hyper-prior and $\lambda > 0$, $\beta > 0$, with probability at least $1-\delta$ over samples $S_1 \in D_1^{m_1}, \ldots, S_n \in D_n^{m_n}$, we have, for all base learner Q and all hyper-posterior $\mathcal{Q}$,*

$$R(\mathcal{Q}, T) \leq \hat{R}(\mathcal{Q}, S_{i=1}^n) + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}\,\|\,\mathcal{P})$$
$$+ \frac{1}{n\beta}\sum_{i=1}^{n}\mathbb{E}_{P\sim\mathcal{Q}}\left[D_{KL}(Q(S_i, P)\|P)\right] + C(\delta, \lambda, \beta, n, m_i). \quad (4)$$

Few-Shot Meta-Learning in Practice

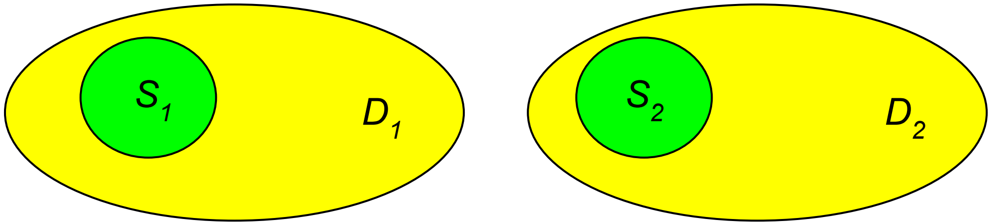
Theorem 2 assumes that m_i for the observed task and m for the target task come from the same task environment T.

Problem: This assumption makes the bounds loose when the number of training examples in the target tasks is limited (e.g., few-shot).

$$\mathbb{E}_{\tilde{T}}[m_i] \gg \mathbb{E}_T[m]$$

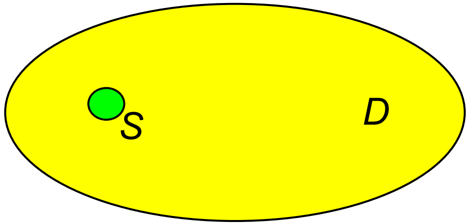
Question: can we benefit from more examples in the observed tasks?

Observed Tasks



$$(D_i, m_i) \sim \tilde{T}$$

Target Task



$$(D, m) \sim T$$

PAC-Bayes Bounds for Few-Shot Meta-Learning

One way is to use the same meta-training loss of Theorem 2:

$$\hat{R}(\mathcal{Q}, S_{i=1}^n) := \mathbb{E}_{P\sim\mathcal{Q}}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{L}(Q(S_i, P), S_i)\right]$$

Training loss of $\mathcal{Q}$

despite of the difference between the training sample sizes.

Theorem 3 *For a target task environment T and an observed task environment $\tilde{T}$ where $\mathbb{E}_{\tilde{T}}[D] = \mathbb{E}_T[D]$ and $\mathbb{E}_{\tilde{T}}[m] \geq \mathbb{E}_T[m]$, let $\mathcal{P}$ be a fixed hyper-prior and $\lambda > 0$, $\beta > 0$, then with probability at least $1-\delta$ over samples $S_1 \in D_1^{m_1}, \ldots, S_n \in D_n^{m_n}$ where $(D_i, m_i) \sim \tilde{T}$, we have, for all base learners Q and hyper-posterior $\mathcal{Q}$,*

$$\begin{aligned} R(\mathcal{Q}, T) \leq& \hat{R}(\mathcal{Q}, S_{i=1}^n) + \left(\frac{1}{\lambda} + \frac{1}{n\beta}\right) D_{KL}(\mathcal{Q}\,\|\,\mathcal{P}) \\ &+ \frac{1}{n\beta}\sum_{i=1}^{n}\mathbb{E}_{P\sim\mathcal{Q}}\left[D_{KL}(Q(S_i, P)\|P)\right] + C(\delta, \lambda, \beta, n, m_i) + \Delta_\lambda(\mathcal{P}, T, \tilde{T}), \end{aligned} \tag{5}$$

where $\Delta_\lambda(\mathcal{P}, T, \tilde{T}) = \frac{1}{\lambda}\log \mathbb{E}_{P\in\mathcal{P}}\, e^{\lambda(R(P,T) - R(P,\tilde{T}))}$.

PAC-Bayes Bounds for Few-Shot Meta-Learning

Reorganize the bound of Thm-3:

$$R(\mathcal{Q},T) \leq \hat{R}(\mathcal{Q},S_{i=1}^{n}) + \frac{1}{n\beta}\sum_{i=1}^{n}\mathbb{E}_{P\sim\mathcal{Q}}\left[D_{KL}(Q(S_i,P)\|P)\right]$$

W_1

$$+\left(\frac{1}{\lambda}+\frac{1}{n\beta}\right)D_{KL}(\mathcal{Q}\|\mathcal{P})$$

$\xi * D_{KL}$

$$+\Delta_\lambda(\mathcal{P},T,\tilde{T})$$

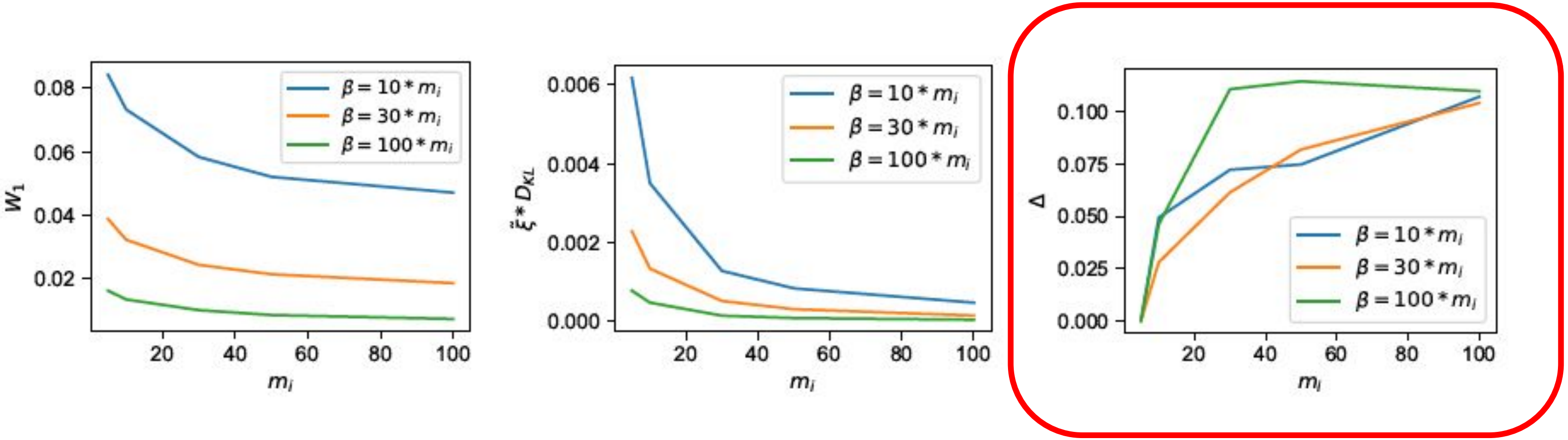
Δ

$$+C(\delta,\lambda,\beta,n,m_i)$$

Const. if $\boldsymbol{\lambda} \sim n$, $\boldsymbol{\beta} \sim m_i$

PAC-Bayes Bounds for Few-Shot Meta-Learning

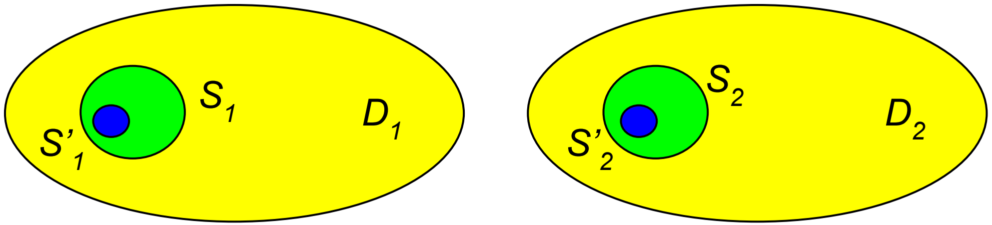
Theorem 3 introduces an additional penalty term Δ_λ , which grows monotonically as the sample difference between observed and target tasks are bigger.



***The number of samples m* of the target task is fixed as 5.**

PAC-Bayes Bounds for Few-Shot Meta-Learning

Can we get rid of Δ_λ in the bound?

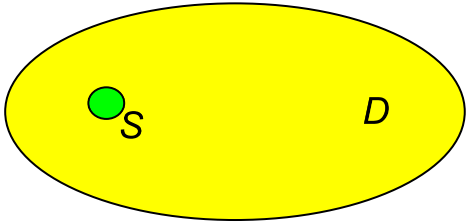


$(D_i, m_i) \sim \tilde{T}$ and $\mathbb{E}[m_i'] = \mathbb{E}_T[m] < \mathbb{E}_{\tilde{T}}[m_i]$

Inspired by MAML:

$$\mathbb{E}_{P\sim \mathcal{Q}}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{L}(Q(S_i', P), S_i)\right]$$

Training loss of $\mathcal{Q}$



$(D, m) \sim T$

- Subsample training examples S_i' for training the base-learner $Q(S_i', P)$.
- Use all training examples S_i to evaluate the meta-training loss of $Q(S_i', P)$ for training the meta-learner $\mathcal{Q}$.

PAC-Bayes Bounds for Few-Shot Meta-Learning

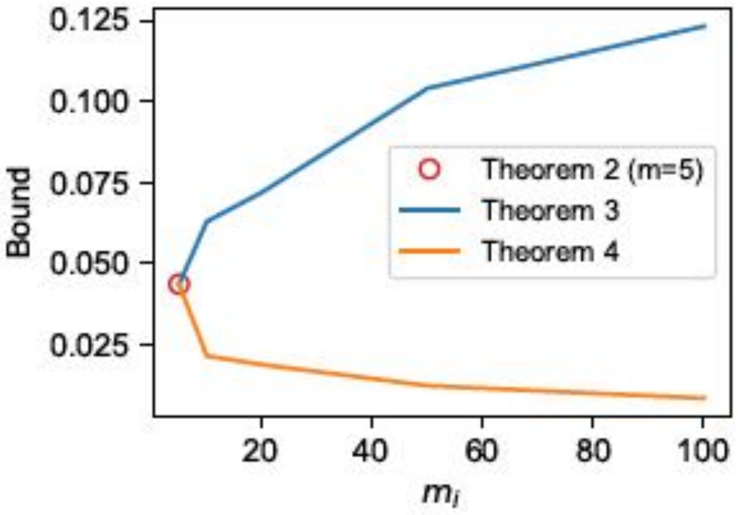
Theorem 4 *For a target task environment T and an observed task environment $\tilde{T}$ where $\mathbb{E}_{\tilde{T}}[D] = \mathbb{E}_T[D]$ and $\mathbb{E}_{\tilde{T}}[m] \geq \mathbb{E}_T[m]$, let $\mathcal{P}$ be a fixed hyper-prior and $\lambda > 0$, $\beta > 0$, then with probability at least $1 - \delta$ over samples $S_1 \in D_1^{m_1}, \ldots, S_n \in D_n^{m_n}$ where $(D_i, m_i) \sim \tilde{T}$, and subsamples $S'_1 \in D_1^{m'_1} \subset S_1, \ldots, S'_n \in D_n^{m'_n} \subset S_n$, where $\mathbb{E}[m'_i] = \mathbb{E}_T[m]$, we have, for all base learner Q and all hyper-posterior $\mathcal{Q}$,*

$$R(\mathcal{Q}, T) \leq \mathbb{E}_{P \sim \mathcal{Q}} \left[\frac{1}{n} \sum_{i=1}^{n} \hat{L}(Q(S'_i, P), S_i) \right] + \left(\frac{1}{\lambda} + \frac{1}{n\beta} \right) D_{KL}(\mathcal{Q} \| \mathcal{P})$$

$$+ \frac{1}{n\beta} \sum_{i=1}^{n} \mathbb{E}_{P \sim \mathcal{Q}} \left[D_{KL}(Q(S'_i, P) \| P) \right] + C(\delta, \lambda, \beta, n, m_i). \quad (6)$$

PAC-Bayes Bounds for Few-Shot Meta-Learning

The PAC-Bayesian bounds of Theorems 2, 3, 4 as evaluated over the Sinusoid dataset.



The number of samples m of the target task is fixed as 5.

Outline

- Theory
  - Existing PAC-Bayesian bounds for Meta-Learning
  - Two new PAC-Bayesian bounds for Few-Shot Meta-Learning
- **Practice**
  - **Connection to Reptile, MAML and PACOH**
  - **A new few-shot meta-learning algorithm: PACMAML**
  - **Empirical Results**

Connection to Reptile and MAML

When the hyper-posterior $\mathcal{Q}$ and base-learner Q both use the Delta-distribution:

$$\mathcal{P}(P) = \mathcal{N}(\mathbf{p} \,|\, 0, \sigma_0^2),\ \mathcal{Q}(P) = \delta(\mathbf{p} = \mathbf{p}_0),\ P(h_{\mathbf{v}}) = \mathcal{N}(\mathbf{v} \,|\, \mathbf{p}, \sigma^2),\ Q_i(h_{\mathbf{v}}) = \delta(\mathbf{v} = \mathbf{q}_i),$$

The PAC-Bayesian bound in Thm-3 and 4 reduces to the following (neglecting the constants):

$$PacB(\mathbf{p}_0) = \frac{1}{n}\sum_{i=1}^{n} \hat{L}(\mathbf{q}_i, S_i) + \frac{\tilde{\xi} \| \mathbf{p}_0 \|^2}{2\sigma_0^2} + \frac{1}{n\beta}\sum_{i=1}^{n} \frac{\| \mathbf{p}_0 - \mathbf{q}_i \|^2}{2\sigma^2}$$

As a result, one can show that using MAP estimation:

- **Theorem 3 ⇒ Reptile**
- **Theorem 4 ⇒ MAML**

Connection to PACOH

The optimal base-learner in the bound of Theorem-3 is the following Gibbs distribution:

$$Q^*(S_i, P)(h) = P(h)\exp(-\beta \hat{L}(h, S_i))/Z_\beta(S_i, P)$$

Plugging into the bound, yields:

$$R(\mathcal{Q}, T) \le \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}_{P\sim\mathcal{Q}}\left[-\frac{1}{\beta}\log Z_\beta(S_i, P)\right]}_{W_1} + \tilde{\xi} D_{KL}(\mathcal{Q}\,\|\,\mathcal{P}) + \Delta_\lambda + C \tag{10}$$

where $\tilde{\xi} = \frac{1}{\lambda} + \frac{1}{n\beta}$ and C is the same constant from the previous bounds. Since Δ_λ is independent of $\mathcal{Q}$ and can be neglected during inference or optimization of $\mathcal{Q}$, it reduces to the same PACOH objective as in [22].

PACMAML

For Theorem-4, we use

$$Q_i^\alpha(S_i', P)(h) = \frac{P(h)\exp(-\alpha \hat{L}(h, S_i'))}{Z_\alpha(S_i', P)}.$$

And yield the following PACMAML objective:

$$R(\mathcal{Q}, T) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P\sim\mathcal{Q}} \underbrace{\left[-\frac{1}{\beta}\log Z_\alpha(S_i', P) + \hat{L}^{\Delta}_{\frac{\alpha}{\beta}}(Q_i^\alpha, S_i, S_i')\right]}_{W_2} + \tilde{\xi} D_{KL}(\mathcal{Q} \,\|\, \mathcal{P}) + C.$$

where $\hat{L}^{\Delta}_{\frac{\alpha}{\beta}}(Q_i^\alpha, S_i, S_i') \triangleq \hat{L}(Q_i^\alpha, S_i) - \frac{\alpha}{\beta}\hat{L}(Q_i^\alpha, S_i')$.

Gradient Estimation of PACMAML

- The PACOH and PACMAML objectives do not have closed-form integration when the loss function is not the squared loss.
- Their gradient can be approximated using a Monte-Carlo approximation similar to the REINFORCE algorithm.

$$\frac{dW_1}{d\mathbf{p}} = -\frac{1}{\beta}\frac{d}{d\mathbf{p}}\log Z_\beta(S_i, \mathbf{p}) = \int Q_i^\beta(\mathbf{w}; S_i)\frac{\partial \hat{L}(\mathbf{p}+\mathbf{w}, S_i)}{\partial \mathbf{p}} d\,\mathbf{w},$$

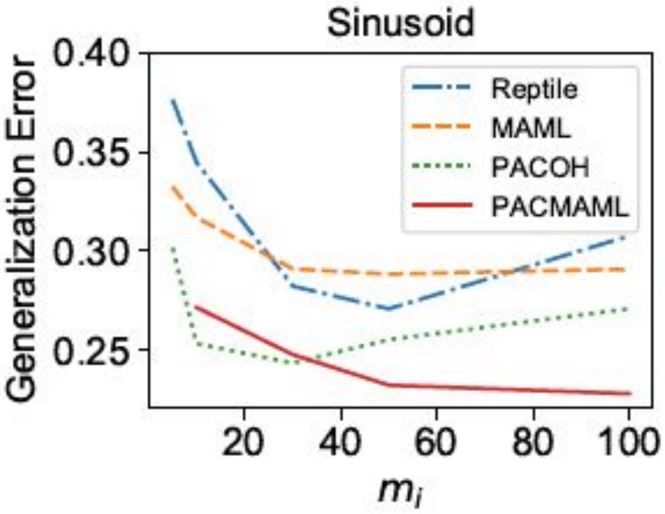
$$\frac{dW_2}{d\mathbf{p}} \simeq \int Q_i^\alpha(\mathbf{w}; S_i')\frac{\partial \hat{L}(\mathbf{p}+\mathbf{w}; S_i)}{\partial \mathbf{p}} d\,\mathbf{w} + \frac{\alpha}{\beta}\int \left(Q_i^\beta(\mathbf{w}; S_i) - Q_i^\alpha(\mathbf{w}; S_i')\right)\frac{\partial \hat{L}(\mathbf{p}+\mathbf{w}; S_i')}{\partial \mathbf{p}} d\,\mathbf{w}\,.$$

where, $Q_i^\beta(\mathbf{w}; S_i) \propto \mathcal{N}(\mathbf{w}\,|\mathbf{0}, \sigma^2)\exp(-\beta\hat{L}(\mathbf{p}+\mathbf{w}, S_i))$.

Experiments

Few-Shot Regression Problems

- Synthetic Sinusoid Task Environment
- Target tasks with m=5 shots
- Squared loss, closed form solution.



Reptile and PACOH both have a U shape that bend up with larger m_i.

MAML and PACMAML monotonically reduce the generalization error with larger m_i.

Experiments

Few-shot Image classification

- Mini-Imagenet (5 classes, k=1 shot per class, m=1x5=5)
- ANIL learning (base-learner only adapts the top layer.)

	FOMAML	MAML	BMAML	PACOH	PACMAML
$m_i = 10$	41.8 ± 0.9	47.3 ± 0.9	29.9 ± 0.9	31.2 ± 0.8	$\mathbf{47.8 \pm 0.9}$
$m_i = 20$	44.3 ± 0.9	48.0 ± 0.9	34.3 ± 0.9	37.0 ± 0.9	$\mathbf{49.1 \pm 0.9}$
$m_i = 40$	46.2 ± 1.0	47.8 ± 0.9	41.5 ± 0.9	41.6 ± 0.9	$\mathbf{48.9 \pm 0.9}$
$m_i = 80$	45.7 ± 0.9	48.1 ± 0.9	44.2 ± 0.9	44.6 ± 0.9	$\mathbf{50.1 \pm 0.9}$

Table 1: Averaged test accuracy and standard error in the ANIL setting.

Experiments

Few-shot Natural language inference

- 12 tasks covering entity typing, rating classification and text classification.
- k=4, 8, 16 shot data per class
- ANIL learning (v=6, 9, 11, 12, base-learner only adapts layers higher than v).

k	H-SMLMT [5]	MAML	BMAML	PACOH	PACMAML
4	48.61	48.21	47.27	50.47	**51.58**
8	52.92	53.52	52.08	54.83	**55.68**
16	57.90	57.38	56.53	58.22	**59.18**

	v=6	v=9	v=11	v=12
MAML	120G	57G	16G	4G
BMAML	121G	59G	19G	4G
PACMAML	33G	16G	8G	4G

Table 2: Top: Averaged test accuracy over the 12 NLI tasks. Bottom: The comparison of TPU memory (High Bandwidth Memory) usage with different adaptive layer thresholds v.

Conclusion

- Two PAC-Bayesian bounds for few-shot meta-learning.
- Using MAP approximation, the 1st bound leads to Reptile and the 2nd bound leads to MAML.
- With Gibbs posterior based base-learner, the 1st bound leads to PACOH. The 2nd PAC-Bayes bound leads to a new PACMAML algorithm.
- PACMAML outperforms existing meta-learning algorithms when evaluated on several benchmark few-shot tasks.