Fine-Grained Analysis of Stability and Generalization for Modern Meta Learning Algorithms

Jiechao Guan[1,3], **Yong liu**[2,3], **Zhiwu Lu**[2,3,*]

[1]School of Information, Renmin University of China, Beijing, China
[2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China

*Corresponding Author {2014200990, liuyonggsai, luzhiwu}@ruc.edu.cn

October 26, 2022

Presentation Outline

1. **Background**
   - Support/Query Episodic Training based Meta Learning Algorithms
   - The Relationship between Meta Learning and Single-Task Learning
   - Uniform Argument Stability of Meta Learning Algorithms
2. **Stability Bounds for Modern Meta Algorithms**
   - Tight Stability Bounds for Nonsmooth Functions
   - Tight Stability Bounds for Smooth Functions
3. **Generalization Bounds for Meta Learning**
   - Near Optimal Bound for Independent Episodes
   - Fast-Rate Bound for Independent Episodes
   - Generalization Bound for Dependent Episodes
4. **Experiment**
   - Convergence Analysis of our Generalization Bounds
5. **Conclusion**

Support/Query Episodic Training based Meta Learning

- A loss function $f : \mathcal{H} \times \mathcal{Z} \to [0, M](M > 0)$ is a function over the product space of hypothesis space $\mathcal{H}$ and sample space $\mathcal{Z}$.
- Meta learning aims to extract knowledge from n training tasks and apply it to the unseen task for fast adaptation.
- Meta learning theory assumes that the distributions $\{D_i\}_{i=1}^n$ associated with training tasks and the distribution D of unseen task are drawn from the same task *environment* τ, i.e., $D, D_i \sim \tau$.
- During meta-train process, a meta-sample $\mathbf{S} = \{S_i = S_i^{tr} \cup S_i^{ts}\}_{i=1}^n$ is available, where $S_i^{tr} \overset{\text{i.i.d.}}{\sim} D_i^K$ of size K is the support set, and $S_i^{ts} \overset{\text{i.i.d.}}{\sim} D_i^q$ of size q is the query set of the i-th training task.
- For any meta learning algorithm $\mathbf{A}$, it takes the meta-sample $\mathbf{S} = \{S_i\}_{i=1}^n$ as input and outputs an inner-task algorithm $\mathbf{A}(\mathbf{S}) : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$.

Transfer Error and Empirical Multi-Task Error for Meta Learning

- The performance of the learned inner-task algorithm is measured by the expectation of the generalization error w.r.t. the task environment τ, which is defined as the *transfer error* by [9, 2] as follows:

$$er(\mathbf{A}(\mathbf{S}), \tau) \triangleq \mathbb{E}_{D \sim \tau} \mathbb{E}_{S^{tr} \sim D^K} \mathbb{E}_{z \sim D} f(\mathbf{A}(\mathbf{S})(S^{tr}), z). \tag{1}$$

- The goal of meta learning theory is thus to give a bound on the transfer error, based on the *empirical multi-task error* on the meta-sample $\mathbf{S}$:

$$\hat{er}(\mathbf{A}(\mathbf{S}), \mathbf{S}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \hat{L}(\mathbf{A}(\mathbf{S})(S_i^{tr}), S_i^{ts}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{l}(\mathbf{A}(\mathbf{S}), S_i). \tag{2}$$

Our Motivation for Improved Generalization Bounds

* Key Point: Reveal the Equivalent Relationship between episodic meta learning and single-task learning

- The environment τ can define an induced distribution $\mathbf{D}_\tau \in \mathcal{M}_1(\mathcal{Z}^m)$, by setting $\mathbf{D}_\tau(F) = \mathbb{E}_{D\sim\tau} D^m(F)$ for any measurable set $F \subseteq \mathcal{Z}^m$.
- Define the estimator $\mathbf{l}(\mathbf{A}(\mathbf{S}), S) \triangleq \hat{L}(\mathbf{A}(\mathbf{S})(S^{tr}), S^{ts})$, where $S = S^{tr} \cup S^{ts}$, $S \overset{\text{i.i.d.}}{\sim} D^m$.
- Rewrite the transfer error as $er(\mathbf{A}(\mathbf{S}), \tau) = \mathbb{E}_{S\sim\mathbf{D}_\tau} \mathbf{l}(\mathbf{A}(\mathbf{S}), S)$.
- The training error $\mathbf{l}(\mathbf{A}(\mathbf{S}), S)$ is the unbiased version of the transfer error $er(\mathbf{A}(\mathbf{S}), \tau) = \mathbb{E}_{S\sim\mathbf{D}_\tau} \mathbf{l}(\mathbf{A}(\mathbf{S}), S)$.
- This is similar to the fact that, in single-task learning, the empirical error $f(A(S), z)$ is the unbiased version of the generalization error $L(A(S), D) = \mathbb{E}_{z\sim D} f(A(S), z)$.

Meta Learning and Single-Task Learning

Table 1: The equivalence relation between the notations of single-task learning and modern support/query (S/Q) episodic training based meta learning.

	Single-Task Learning	S/Q Training based Meta Learning
Sample	$z \in \mathcal{Z}$	$S = (z_1, ..., z_m) \in \mathcal{Z}^m$
Training Set	$S = (z_1, ..., z_m) \in \mathcal{Z}^m$	$\mathbf{S} = (S_1, ..., S_n) \in (\mathcal{Z}^m)^n$
Hypothesis	$h \in \mathcal{H}$	$A \in \mathcal{A}(\mathcal{H}, \mathcal{Z})$
Algorithm	$A \in \mathcal{A}(\mathcal{H}, \mathcal{Z})$	$\mathbf{A} \in \mathcal{A}(\mathcal{A}(\mathcal{H}, \mathcal{Z}), \mathcal{Z}^m)$
Learning Task	$D \in \mathcal{M}_1(\mathcal{Z})$	$\mathbf{D} \in \mathcal{M}_1(\mathcal{Z}^m)$, typically $\mathbf{D} = \mathbf{D}_\tau$ is induced by the environment $\tau \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$.
Loss Estimator	$f : \mathcal{H} \times \mathcal{Z} \to [0, M]$	$\mathbf{l} : \mathcal{A}(\mathcal{H}, \mathcal{Z}) \times \mathcal{Z}^m \to [0, M]$
Empirical Error	$\hat{L}(A(S), S) = \frac{1}{m}\sum_{i=1}^m f(A(S), z_i)$	$\hat{er}(\mathbf{A}(\mathbf{S}), \mathbf{S}) = \frac{1}{n}\sum_{i=1}^n \mathbf{l}(\mathbf{A}(\mathbf{S}), S_i)$
Expected Error	$L(A(S), D) = \mathbb{E}_{z\sim D} f(A(S), z)$	$er(\mathbf{A}(\mathbf{S}), \tau) = \mathbb{E}_{S\sim \mathbf{D}_\tau} \mathbf{l}(\mathbf{A}(\mathbf{S}), S)$
Probability Bound	$D^m\{S : L(A(S), D) \geq B(\delta, S)\} \leq \delta$	$\mathbf{D}^n\{\mathbf{S} : R(\mathbf{A}(\mathbf{S}), \tau) \geq \Pi(\delta, \mathbf{S})\} \leq \delta$

Uniform Argument Stability of Meta Algorithms

Definition 1.1

Given a meta learning algorithm $\mathbf{A}$, any neighboring meta samples $\mathbf{S}, \mathbf{S}'$, and any training episode $S \in \mathcal{Z}^m$, we define the *uniform argument stability* random variable of $\mathbf{A}$ as $\delta_{\mathbf{A}}(\mathbf{S}, \mathbf{S}'; S) = ||\mathbf{A}(\mathbf{S})(S) - \mathbf{A}(\mathbf{S}')(S)||$. $\mathbf{A}$ is defined as a uniform argument β-stable meta learning algorithm if for some $\beta > 0$, we have $\sup_{\mathbf{S}\simeq\mathbf{S}', S} \delta_{\mathbf{A}}(\mathbf{S}, \mathbf{S}'; S) \leq \beta$ or $\sup_{\mathbf{S}\simeq\mathbf{S}', S} \mathbb{E}_{\mathbf{A}}\delta_{\mathbf{A}}(\mathbf{S}, \mathbf{S}'; S) \leq \beta$, where $\mathbb{E}_{\mathbf{A}}$ denote the expectation w.r.t. the internal randomness of $\mathbf{A}$.

For a meta learning algorithm with SGD method, the internal randomness of $\mathbf{A}$ comes from the randomness of sampling at each iteration.

Note: Since the updated parameter
$w_{t+1} = \mathrm{Proj}_{\mathcal{W}}[w_t - \eta_t \partial_{w_t} \hat{L}(\mathbf{A}(\mathbf{S})(S^{tr}_{i_t}), S^{ts}_{i_t})]$ is related to the whole episode, we equivalently write $\hat{R}(\mathbf{A}(\mathbf{S})(S), S) \triangleq \hat{L}(\mathbf{A}(\mathbf{S})(S^{tr}), S^{ts})$ and the episode-level SGD update rule is: $w_{t+1} = \mathrm{Proj}_{\mathcal{W}}[w_t - \eta_t \partial_{w_t} R(w_t, S_{i_t})]$.

Pseudo Code of Episodic Meta Learning Algorithms

Algorithm 1 Episodic Meta Learning Algorithm

```
Input: training dataset $\mathbf{S} = \{S_i\}_{i=1}^n$ with $S_i = \{S_i^{tr}, S_i^{ts}\}$, # of iterations $T$, learning rates $\eta_t$ $(t \in [T])$.
Initialize: the parameters of DNN $w_1$.
for $t = 1$ to $T$ do
    Uniformly sample one of $n$ training episodes with replacement. Let $i_t$ be the episode index.
    $w_{t+1} = \mathrm{Proj}_{\mathcal{W}}\left(w_t - \eta_t \partial \hat{R}(w_t, S_{i_t})\right)$
end for
return $w_{T+1}$
```

For the metric-learning based ProtoNet [13] and MatchingNet [15] in classification, h_{w_t} is regarded as the feature extractor,

$$\hat{R}(w_t, S_{i_t}) = \frac{1}{q} \sum_{(x,y)\in S_{i_t}^{ts}} -\log \frac{\exp\{-\mathrm{d}(h_{w_t}(x), c_y)\}}{\sum_k \exp\{-\mathrm{d}(h_{w_t}(x), c_k)\}},$$

where $c_k = \frac{1}{Norm} \sum_{(x,y)\in S_{i_t}^{tr}, y=k} h_{w_t}(x)$ is the averaged vector of the sample features in $S_{i_t}^{tr}$ with the same class label k; $\mathrm{d}(\cdot,\cdot)$ is the distance between two feature vectors (e.g. the Euclidean [13] or Cosine distance [15]). For MetaOptNet [8]:

$$\hat{R}(w_t, S_{i_t}) = \frac{1}{q} \sum_{(x,y)\in S_{i_t}^{ts}} -\log \frac{\exp\{\lambda\langle h_{w_t}(x), \phi_y\rangle\}}{\sum_k \exp\{\lambda\langle h_{w_t}(x), \phi_k\rangle\}},$$

where $\{\phi_k\}_{k=1}^K$ are the parameters of the classifier returned by supervised learning algorithms (e.g. SVM) on the support set $S_{i_t}^{tr}$, $\langle,\rangle$ represents the inner product. For MAML [6]:

$$\hat{R}(w_t, S_{i_t}) = \frac{1}{q} \sum_{z\in S_{i_t}^{ts}} f(w_t - \frac{\alpha_t}{K} \sum_{z'\in S_{i_t}^{tr}} \partial f(w_t, z'), z).$$

Stability Bounds for Nonsmooth Functions with α-Hölder Continuous Subgradients

Theorem 2.1

$\forall$ fixed $S \in \mathcal{Z}^m$, let $\hat{R}(\cdot, S)$ be a convex and (α, G)-Hölder smooth function, where $\alpha \in [0,1)$. Let $\mathbf{A}$ be a meta learning algorithm with sampling-with-replacement SGD. Denote by w_j and w_j' the outputs after $j(j \in [T])$ steps of SGD on $\mathbf{S}$ and $\mathbf{S}^i$, respectively. Define $R_{\mathbf{S}}(w) = n^{-1}\sum_{i=1}^n \hat{R}(w, S_i)$, $\forall w \in \mathcal{W}$. Then $\forall S \in \mathcal{Z}^m$, $\mathbb{E}_{\mathbf{A}}\delta_{\mathbf{A}}(\mathbf{S}, \mathbf{S}'; S)$ is upper bounded by

$$\sqrt{2}c_\alpha\Big[\sum_{j=1}^{T}\eta_j^2\mathbb{E}\big[R_{\mathbf{S}}^{\frac{2\alpha}{1+\alpha}}(w_j) + R_{\mathbf{S}^i}^{\frac{2\alpha}{1+\alpha}}(w_j')\big]\Big]^{\frac{1}{2}} + \frac{2c_\alpha}{n}\sum_{j=1}^{T}\eta_j\big[\hat{R}^{\frac{\alpha}{1+\alpha}}(w_j, S_i) + \hat{R}^{\frac{\alpha}{1+\alpha}}(w_j, S_i')\big]. \quad (3)$$

In addition, if $\hat{R}(\cdot, S)$ is bounded by M and the step size $\eta_j = \eta$ $\forall j \in [T]$, we can obtain the lower and upper bounds of the uniform argument stability of $\mathbf{A}$: $c_\alpha M^{\frac{\alpha}{1+\alpha}}(\min\{1, \frac{T}{n}\}\eta\sqrt{T} + \frac{\eta T}{n}) \leq \sup_{\mathbf{S},\mathbf{S}',S}\mathbb{E}_{\mathbf{A}}\delta_{\mathbf{A}}(\mathbf{S}, \mathbf{S}'; S) \leq 4c_\alpha M^{\frac{\alpha}{1+\alpha}}\big(\min\{1, \frac{T}{n}\}\eta\sqrt{T} + \frac{\eta T}{n}\big)$.

Stability Bounds for Smooth Function

Theorem 2.2

$\forall$ fixed $S \in \mathcal{Z}^m$, let $\hat{R}(\cdot, S)$ be a G-smooth convex function. Let $\mathbf{A}$ be a meta learning algorithm with sampling-with-replacement SGD. Denote by w_j and w_j' the outputs after $j(j \in [T])$ steps of SGD on neighboring meta samples $\mathbf{S}$ and $\mathbf{S}^i$, respectively. Then $\forall S \in \mathcal{Z}^m$, $\eta_j \leq 2/G$,

$$\mathbb{E}_{\mathbf{A}}||\mathbf{A}(\mathbf{S})(S) - \mathbf{A}(\mathbf{S}^i)(S)|| \leq \frac{\sqrt{2G}}{n}\sum_{j=1}^{T}\eta_j \mathbb{E}_{\mathbf{A}}\left[\sqrt{\hat{R}(w_j, S_i)} + \sqrt{\hat{R}(w_j', S_i')}\right].$$

In addition, if $\hat{R}(\cdot, S)$ is bounded by M, we can obtain the lower and upper bounds of the uniform argument stability of $\mathbf{A}$: $\frac{1}{n}\sum_{j=1}^{T}\eta_j \leq \sup_{\mathbf{S},\mathbf{S}',S}\mathbb{E}_{\mathbf{A}}\delta_{\mathbf{A}}(\mathbf{S},\mathbf{S}';S) \leq \frac{2\sqrt{2MG}}{n}\sum_{j=1}^{T}\eta_j$.

Theorem 2.3

$\forall$ fixed $S \in \mathcal{Z}^m$, let $\hat{R}(\cdot, S)$ be a σ-Lipschitz and G-smooth function. Let $\mathbf{A}$ be a meta learning algorithm. Denote by w_j and w_j' the outputs after $j(j \in [T])$ steps of SGD on $\mathbf{S}$ and $\mathbf{S}^i$, respectively. Define the learning rate $\eta_j = \frac{a}{jG}$, $\forall j \in [T]$ with $a > 0$. Then $\forall S \in \mathcal{Z}^m$, the lower and upper stability bounds of $\mathbf{A}$ satisfy: $\frac{T^a}{6n^{1+a}} \leq \sup_{\mathbf{S},\mathbf{S}',S}\mathbb{E}_{\mathbf{A}}\delta_{\mathbf{A}}(\mathbf{S},\mathbf{S}';S) \leq \frac{11\ln(n)\sigma T^a}{n^{1+a}}$.

Near Optimal Transfer Error Bound for Meta Learning with Independent Episodes

Theorem 3.1

Let **A** $\in \mathcal{A}(\mathcal{A}(\mathcal{H},\mathcal{Z}),\mathcal{Z}^m)$ *be a uniform argument* β*-stable meta algorithm, i.e.,* $\sup_{\mathbf{S}\simeq\mathbf{S}',S}\mathbb{E}_{\mathbf{A}}||\mathbf{A}(\mathbf{S})(S)-\mathbf{A}(\mathbf{S}')(S)||\le\beta$*. For any* $S\in\mathcal{Z}^m$*, let* $\hat{R}(\cdot,S)$ *be* $[0,M]$*-valued, and satisfy one of the two following conditions:* **(1)** $\hat{R}(\cdot,S)$ *is convex and* (α,G)*-Hölder smooth* $(\alpha\in[0,1])$*;* **(2)** $\hat{R}(\cdot,S)$ *is* σ*-Lipschitz and* G*-smooth. Suppose* $\mathbb{P}_{\mathbf{A}}[\delta_{\mathbf{A}}(\mathbf{S},\mathbf{S}';S)>\beta]\le\delta_0$*. Then for any independent task environment* $\tau\in\mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$*, any* $\delta\in(0,1)$*, the following holds with probability at least* $1-\delta-\delta_0$ *over the draw of* **S** *and the internal randomness of* **A***:*

$$\sigma_\alpha\beta\ln\frac{1}{\delta}+\frac{M}{\sqrt{n}}\sqrt{\ln(1/\delta)}\lesssim er(\mathbf{A}(\mathbf{S}),\tau)-\hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S})\lesssim\sigma_\alpha\beta\ln\frac{n}{\delta}+\frac{M}{\sqrt{n}}\sqrt{\ln(1/\delta)}.$$

Remark 3.2

Our transfer error bound in Theorem 3.1 has three advantages over the bound in Theorem 2 from [2]: **(1)** *[2, Theorem 2] gives a high-probability upper bound of* $O(\sqrt{n}\gamma+M/\sqrt{n})$ *for transfer error, where* γ *is the uniform stability parameter and always scales as* $O(1/n)$*; in contrast, our upper bound of* $O(\beta\ln n+M/\sqrt{n})$ *is improved by replacing the* $\sqrt{n}$ *factor before the stability parameter with* $\ln n$*.* **(2)** *In [2], the uniform stability* $\gamma=O(T^{\frac{a}{a+1}}/n)$*, whereas our uniform argument stability* $\beta=O(T^a/n^{1+a})$ *is tighter when* $T^{\frac{a}{a+1}}\le n$*, i.e., when the uniform stability bound* $\gamma=O(T^{\frac{a}{a+1}}/n)$ *is non-vacuous.* **(3)** *Our high-probability transfer error bound of order* $O(1/\sqrt{n})$ *is near optimal.*

Theoretical Insights From the Optimal Bound

Remark 3.3

We uncover two limitations of stability-based meta learning theory: **(1)** *Recall the lower stability bound for meta learning algorithms with convex α-Hölder smooth function ($\alpha \in [0,1)$) in Theorem 2.1, we find that the lower transfer error bound in Theorem 3.1 is $er(\mathbf{A}(\mathbf{S}),\tau) - \hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S}) \gtrsim \sigma_\alpha \ln(1/\delta) c_\alpha M^{\frac{\alpha}{1+\alpha}}(\eta\sqrt{T} + \eta T/n)$ when $T \geq n$. This indicates that the lower transfer error bound is greater than a constant and will not converge to zero with the increase of n. Thus, the stability-based transfer error bound is vacuous and cannot provide asymptotic guarantees for convex Hölder smooth functions.* **(2)** *The stability-based transfer error bound of $O(1/\sqrt{n})$ in Theorem 3.1 is near optimal. Such result is consistant with the observation in [10, Section 2] that under the (i.i.d.) task environment assumption, the term $O(1/\sqrt{n})$ in the generalization bound is unavoidable. Thus, to obtain sharper generalization bounds for meta learning (e.g. the bound of $O(1/\sqrt{mn})$ or even $O(1/mn)$), we need to consider other stability notions (e.g. [4]), or suppose stronger task relatedness in the environment (e.g. [1, 7]), or even drop the task environment assumption (e.g. [3, 14]).*

The Benefits of Support/Query Training over Traditional ERM Training Strategy

Remark 3.4

Under the independent task environment assumption, we compare our bound of $O(1/\sqrt{n})$ via S/Q episodic training strategy with other transfer error bounds that are obtained via traditional ERM strategy over all samples in training tasks. In detail, the bound from [9, Theorems 2 and 6] via algorithmic stability analysis is of $O(1/m + 1/\sqrt{n})$; the bounds from [11, Theorem 1] and [12, Theorem 2] via PAC-Bayes analysis are of $O(1/\sqrt{n} + 1/\sqrt{m})$; the bound from [7, Theorem 5] via covering number analysis is of $O(1/\sqrt{nm} + 1/\sqrt{m})$. All of these bounds via ERM strategy involve a term $O(1/\sqrt{m})$, and such term can be large when m is relatively small (e.g. $m = 5$ or $m = 10$ in the few-shot learning setting). Thus, in terms of the tightness of transfer error bounds, the S/Q episodic training strategy is superior to the ERM strategy for meta learning, when $m << n$. Such result was also pointed out by [2] and is more rigorously demonstrated in this work.

Table 2: Different transfer error bounds. All bounds hold under the independent task environment assumption, with n training tasks and m samples per task. For the stability-based bounds in [9, 2], $\gamma_m = O(\frac{1}{m})$ represents the uniform stability of an inner-task algorithm. In our bounds, $\beta_n = O(\frac{1}{n})$ represents the uniform argument stability of a meta learning algorithm.

Existing Works	Object	Training Strategy	Generalization Gap	Bounds on Generalization Gap
[7, Theorem 5]	$\mathcal{H}$	T-ERM	$er(\mathcal{H},\tau) - \hat{er}(\mathcal{H},\mathbf{S})$	$O(\frac{c_1}{\sqrt{nm}} + \frac{c_2}{\sqrt{m}})$
[11, Theorem 1]	$\mathcal{Q}$	T-ERM	$er(\mathcal{Q},\tau) - \hat{er}(\mathcal{Q},\mathbf{S})$	$O(\frac{\mathrm{KL}(\mathcal{Q}\|\|\mathcal{P})}{\sqrt{n}} + \frac{\mathbb{E}_{P\sim\mathcal{Q}}\mathrm{KL}(Q(S_i,P)\|\|P)}{\sqrt{m}})$
[5, Theorem 3]	$\mathcal{Q}$	T-ERM	$er(\mathcal{Q},\tau) - \hat{er}(\mathcal{Q},\mathbf{S})$	$O(\sqrt{\frac{\mathrm{KL}(\mathcal{Q}\|\|\mathcal{P})}{n}} + \gamma_m)$
[9, Theorem 6]	$\mathbf{A}(\mathbf{S})$	T-ERM	$er(\mathbf{A}(\mathbf{S}),\tau) - \hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S})$	$O(\gamma_n\sqrt{n} + \frac{M}{\sqrt{n}} + \gamma_m)$
[2, Theorem 1]	$\mathbf{A}(\mathbf{S})$	S/Q	$er(\mathbf{A}(\mathbf{S}),\tau) - \hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S})$	$O(\gamma_n\sqrt{n} + \frac{M}{\sqrt{n}})$
Our Theorem 3.1	$\mathbf{A}(\mathbf{S})$	S/Q	$er(\mathbf{A}(\mathbf{S}),\tau) - \hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S})$	$O(\beta_n \ln n + \frac{M}{\sqrt{n}})$
Our Theorem 3.6	$\mathbf{A}(\mathbf{S})$	S/Q	$er(\mathbf{A}(\mathbf{S}),\tau) - \hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S})$	$O(\beta_n \ln n + \frac{M}{n})$

Fast-Rate Transfer Error Bound of $O(\frac{\ln n}{n})$ for Independent Episodes

Definition 3.5

(Polyak-Łojasiewicz [16]) Any function $f : \mathcal{W} \to \mathbb{R}$ satisfies the Polyak-Łojasiewicz (PL) condition on $\mathcal{W}$ with parameter $\mu > 0$ if for all $w \in \mathcal{W}$, $f(w) - f(w^*) \leq \frac{1}{2\mu}||\partial^0 f(w)||_2^2$, where w^* denotes the Euclidean projection of w onto the set of global minimizer of f in $\mathcal{W}$.

Theorem 3.6

Under the same conditions of Theorem 3.1, for any fixed $S \in \mathcal{Z}^m$, let $\hat{R}(\cdot, S)$ additionally satisfy Polyak-Łojasiewicz condition in Definition 3.5. Suppose $\mathbb{P}_{\mathbf{A}}[\delta_{\mathbf{A}}(\mathbf{S}, \mathbf{S}'; S) > \beta] \leq \delta_0$. Then, there exist $c > 0$, such that $\forall \tau \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$, and any $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta - \delta_0$ over the draw of $\mathbf{S}$ and the internal randomness of $\mathbf{A}$:

$$er(\mathbf{A}(\mathbf{S}), \tau) \leq (1 + \eta)\hat{er}(\mathbf{A}(\mathbf{S}), \mathbf{S}) + c(1 + 1/\eta)\Big(\sigma_\alpha \beta \ln n + \frac{M}{n}\Big) \ln \frac{1}{\delta}.$$

Forest Approximation and Forest Complexity

Definition 3.7

(Forest Approximation [17]) Given a graph Γ, a forest F, and a mapping $\phi : V(\Gamma) \to V(F)$, if $\phi(u) = \phi(v)$ or $\langle \phi(u), \phi(v) \rangle \in E(F)$ for any $\langle u, v \rangle \in E(\Gamma)$, we say that (ϕ, F) is a forest approximation of Γ. Let $\Phi(\Gamma)$ denote the set of forest approximations of Γ.

Definition 3.8

(Forest Complexity [17]) Given a graph Γ and any forest approximation $(\phi, F) \in \Phi(G)$ with F consisting of trees $\{T_i\}_{i \in [k]}$. Define $\lambda_{(\phi,F)} = \sum_{\langle u,v \rangle \in E(F)} \left(|\phi^{-1}(u)| + |\phi^{-1}(v)| \right)^2 + \sum_{i=1}^{k} \min_{u \in V(T_i)} |\phi^{-1}(u)|^2$. We call $\Lambda(\Gamma) = \min_{(\phi,F) \in \Phi(\Gamma)} \lambda_{(\phi,F)}$ the forest complexity of the graph $\Gamma = (V, E)$. Here, $\phi^{-1}(u)$ is the set of pre-images of the element u.

Transfer Error Bound for Meta Learning with Dependent Episodes

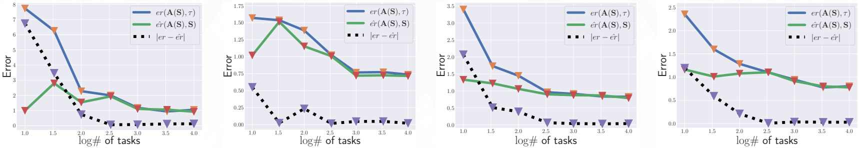
Theorem 3.9

Under the same conditions of Theorem 3.1, except that $\mathbf{S}$ *is a meta sample of size* n *with dependency graph* Γ. *Let the maximum degree of the graph* Γ *is* $\triangle$. *Suppose* $\mathbb{P}_{\mathbf{A}}[\delta_{\mathbf{A}}(\mathbf{S},\mathbf{S}';S) > \beta] \leq \delta_0$. *Then, for any environment* $\tau \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{Z}))$, *any* $\delta \in (0,1)$, *the following holds with probability at least* $1-\delta-\delta_0$ *over the draw of* $\mathbf{S}$ *and the internal randomness of* $\mathbf{A}$*:*

$$er(\mathbf{A}(\mathbf{S}),\tau) \leq \hat{er}(\mathbf{A}(\mathbf{S}),\mathbf{S}) + \sigma_\alpha\beta(\triangle+1) + \left(2\sigma_\alpha\beta + \frac{M}{n}\right)\sqrt{\frac{\Lambda(\Gamma)\ln 1/\delta}{2}},$$

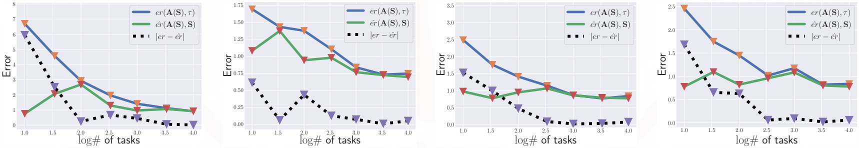
When $\mathbf{S}$ is an independent sample, the forest complexity $\Lambda(\Gamma) = n$, the maximum degree $\triangle = 0$, and the above forest-complexity based generalization bound degenerates to the bound for independent episodes.

Convergence Analysis of Transfer Error Bounds



(a) Bilevel, l_2 loss (b) Bilevel, l_1 loss (c) MAML, l_2 loss (d) MAML, l_1 loss

Figure 1: Convergence analysis of generalization gaps for independent tasks.



(a) Bilevel, l_2 loss (b) Bilevel, l_1 loss (c) MAML, l_2 loss (d) MAML, l_1 loss

Figure 2: Convergence analysis of generalization gaps for dependent tasks.

Conclusions and Future Works

Our contributions are four-fold:

(1) We provide matching lower and upper stability bounds for modern meta learning algorithms with general loss functions. The stability bound for nonsmooth convex functions implies that modern meta learning algorithms are not stable enough.

(2) We develop a near-optimal high-probability bound of $O(1/\sqrt{n})$ on the transfer error in meta learning. Such bound is also used to reveal the advantage of the S/Q episodic strategy for meta learning over the traditional ERM strategy.

(3) We derive a deformed generalization bound of $O(\ln n/n)$ with additional curvature condition of loss functions.

(4) We obtain the first generalization bound for meta learning with dependent episodes.

Thanks!

References

[1] Shai Ben-David and Reba Schuller. "Exploiting Task Relatedness for Mulitple Task Learning". In: *Conference on Learning Theory (COLT)*. 2003, pp. 567–580.

[2] Jiaxin Chen et al. "A Closer Look at the Training Strategy for Modern Meta-Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020, pp. 396–406.

[3] Simon S. Du et al. "Few-Shot Learning via Learning the Representation, Provably". In: *International Conference on Learning Representations (ICLR)*. 2021.

List of References

[4] Alireza Fallah, Aryan Mokhtari, and Asuman E. Ozdaglar. "Generalization of Model-Agnostic Meta-Learning Algorithms: Recurring and Unseen Tasks". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021.

[5] Alec Farid and Anirudha Majumdar. "Generalization Bounds for Meta-Learning via PAC-Bayes and Uniform Stability". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021, pp. 2173–2186.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *International Conference on Machine Learning (ICML)*. 2017, pp. 1126–1135.

List of References

[7] Jiechao Guan and Zhiwu Lu. "Task Relatedness-Based Generalization Bounds for Meta Learning". In: *International Conference on Learning Representations (ICLR)*. 2022.

[8] Kwonjoon Lee et al. "Meta-Learning With Differentiable Convex Optimization". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10657–10665.

[9] Andreas Maurer. "Algorithmic Stability and Meta-Learning". In: *Journal of Machine Learning Research (JMLR)* 6 (2005), pp. 967–994.

[10] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. "The Benefit of Multitask Representation Learning". In: *Journal of Machine Learning Research (JMLR)* 17 (2016), 81:1–81:32.

List of References

[11] Anastasia Pentina and Christoph H. Lampert. "A PAC-Bayesian bound for Lifelong Learning". In: *International Conference of Machine Learning (ICML)*. 2014, pp. 991–999.

[12] Jonas Rothfuss et al. "PACOH: Bayes-Optimal Meta-Learning with PAC-Guarantees". In: *International Conference on Machine Learning (ICML)*. 2021, pp. 9116–9126.

[13] Jake Snell, Kevin Swersky, and Richard Zemel. "Prototypical Networks for Few-shot Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2017, pp. 4077–4087.

[14] Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. "On the Theory of Transfer Learning: The Importance of Task Diversity". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020, pp. 7852–7862.

List of References

[15] Oriol Vinyals et al. "Matching Networks for One Shot Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2016, pp. 3630–3638.

[16] Hui Zhang. "New analysis of linear convergence of gradient-type methods via unifying error bound conditions". In: *Mathematical Programming* 180.1 (2020), pp. 371–416.

[17] Rui Ray Zhang et al. "McDiarmid-Type Inequalities for Graph-Dependent Variables and Stability Bounds". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2019, pp. 10889–10899.