Improving Generalization of Meta-Learning with Inverted Regularization at Inner-Level

Lianzhe Wang[1] · Shiji Zhou[1] · Shanghang Zhang[2] · Xu Chu[1] · Heng Chang[1] · Wenwu Zhu[1]
[1]Tsinghua University · [2]Peking University

Presenter: Lianzhe Wang
May 30, 2023
TUE-PM-354

One-Minute Paper Summary

(A Glance at the Poster)

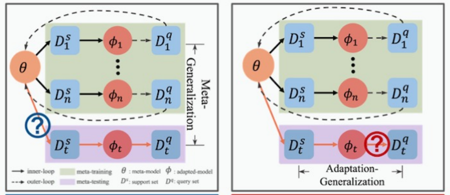
Improving Generalization of Meta-Learning with Inverted Regularization at Inner-Level

Lianzhe Wang[1] · Shiji Zhou[1] · Shanghang Zhang[2] · Xu Chu[1] · Heng Chang[1] · Wenwu Zhu[1]
[1]Tsinghua University · [2]Peking University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Background & Motivations

- Meta-learning, with its primary goal of achieving strong performance when adapting to **new tasks**, necessitates the meta-model to possess a **strong generalization ability**.
- For representative optimization-based meta-learning approaches, which formulate the meta-learning problem as a bi-level optimization problem, the generalization challenge unfolds in two dimensions:



Meta-generalization: the meta-model should generalize to unseen tasks.
- requiring performance consistency between training tasks and new tasks.

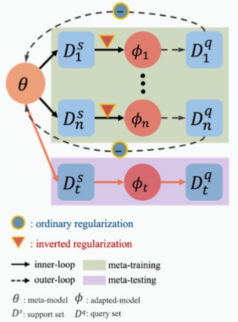
Adaptation-generalization: the adapted-model should generalize to the domain of a specific task.
- requiring performance consistency between task support data (often few-shot) and true data distribution of corresponding task.

Meta-generalization & **Adaptation-generalization**.

- Addressing the twofold generalization problem is challenging, particularly given the current absence of dedicated work specifically targeting adaptation-generalization.

(illustration Figure adapted from [1].)

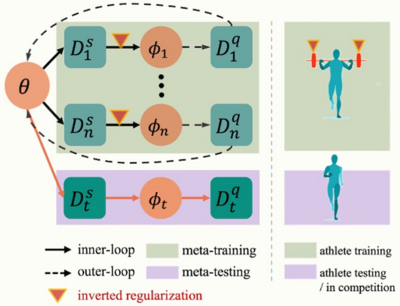
Method



Our Minimax-Meta Regularization method is designed to improve the generalization of bi-level meta-learning by combining two types of regularizations during training:

- an **ordinary regularization** at the outer-level, which encourages the meta-model to learn more generalized hypotheses. (for meta-generalization)
- * an **inverted regularization** at the inner-level, which intentionally increases the generalization difficulty of the adapted model. This forces the meta-model to learn meta-knowledge with better generalization. (for adaptation-generalization)

The Inner-level "Inverted" Regularization

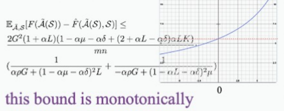


- The **inverted regularization** increases the generalization difficulty, which could typically be achieved by changing the sign of an **ordinary regularization** term (e.g., negative L1/L2-Norm, inverted entropy regularization, changing the inner-loss from $\hat{\mathcal{L}}(\theta, \mathcal{D}_i^S)$ to $\hat{\mathcal{L}}(\theta, \mathcal{D}_i^S) + \sigma^{in} Inverted\_Reg(\theta, \mathcal{D}_i^S)$).
- Intuitionally, at the inner-level, by making the adapted model more difficult to generalize to each training task domain, the meta-model is pushed to learn better-generalized meta-knowledge.
- This can be seen as a form of "adversarial training" or "loaded training" for the meta-model, enhancing its generalization performance.
- Importantly, the inverted regularization/"adversarial training" is only applied during training and not during testing. The meta-model is tested without extra burden after training.

Theoretical Results

- Taking **L2-Norm** and **MAML** as the example regularization and meta-learning algorithm, an analysis of inverted regularization at inner-level is conducted.
- Regularization parameter δ here can be either positive or negative to represent the ordinary or inverted regularization. We treat δ as a variable and analyze how its value would influence the generalization and test error bound.
- The result suggests that the inner-level regularization improves generalization and test error bound only when it's inverted.

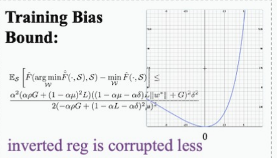
Generalization Error Bound:



this bound is monotonically increasing with respect to δ.

+

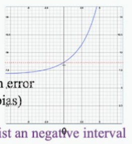
Training Bias Bound:



inverted reg is corrupted less compared to ordinary reg.

=

Test Error Bound: (combining generalization error and training bias)



there must exist an negative interval $[\delta^*, 0)$ in which all δ values improve the test performance.

Experimental Results

Table 1 & 2. Test accuracy of MAML with different types of regularization in the Mini-Imagenet 5-way MAML Few-shot Classification experiment Backbone: 48-48-48-48 conv.

(L2-Norm as regularization objective only).

Mini-Imagenet 5-way Few-shot Classification for MAML *(Reg Objective: L2-Norm)*				
Regularization Type	Outer Reg	Inner Reg	1-Shot	5-Shot
no regularization	-	-	49.58±0.45%	65.39±0.50%
regularize the outer-level	*Ordinary*	-	49.90±0.54%	66.47±1.21%
regularize the inner-level	-	*Ordinary*	49.28±0.37%	64.80±0.25%
invertedly regularize the inner-level	-	*Inverted*	49.92±0.42%	66.05±0.68%
Minimax-Meta Regularization	*Ordinary*	*Inverted*	**50.25±0.38%**	**68.17±0.92%**

(L2-Norm & Entropy combined regularization objective).

Mini-Imagenet 5-way Few-Shot Classification for MAML *(Reg Objective: L2-Norm & Entropy)*				
Regularization Type	Outer Reg	Inner Reg	1-Shot	5-Shot
no regularization	-	-	49.58±0.45%	65.39±0.50%
regularize the outer-level	*Ordinary*	-	50.23±0.67%	67.18±0.88%
regularize the inner-level	-	*Ordinary*	48.07±1.01%	64.32±0.35%
invertedly regularize the inner-level	-	*Inverted*	49.96±0.33%	65.91±0.41%
Minimax-Meta Regularization	*Ordinary*	*Inverted*	**50.85±0.37%**	**69.36±0.34%**

Table 3. Omniglot 20-way 1-shot experiment.

*the * indicates result generated in our experiment.*

Omniglot 20-way 1-Shot Classification	
	Accuracy
Meta-SGD	95.93±0.38%
Prototypical Net	96.00%
Meta-Networks	97.00%
GNN	97.40%
Relation Network	97.60±0.20%
R2-D2	96.24±0.05%
SNAIL	97.64±0.30%
TAML(Entropy)	95.62±0.50%
MAML *	94.20±0.41%
Minimax-MAML(ours)*	95.76±0.39%
MAML++ *	97.21±0.51%
Minimax-MAML++(ours)*	**97.77±0.06%**

Table 4. Mini-Imagenet 5-way few-shot experiment.

*the * indicates result generated in our experiment.*

Mini-Imagenet 5-way Few-Shot Classification			
Approach	Backbone	1-Shot Accuracy	5-Shot Accuracy
Meta-SGD	64-64-64-64	50.47±1.87%	64.03±0.94%
Prototypical Nets	64-64-64-64	49.42±0.78%	68.20±0.66%
GNN	64-96-128-256	50.33±0.36%	66.41±0.63%
R2-D2	64-64-64-64	49.50±0.20%	65.40±0.20%
LR-D2	96-192-384-512	51.90±0.20%	68.70±0.20%
MetaOptNet	64-64-64-64	53.23±0.59%	69.51±0.48%
TAML(Entropy)	64-64-64-64	51.73±1.88%	66.05±0.85%
MAML-Meta Dropout	32-32-32-32	51.93±0.67%	67.42±0.52%
MAML-MMCF	32-32-32-32	50.35±1.82%	64.91±0.96%
MAML *	64-64-64-64	50.20±1.65%	65.86±0.61%
Minimax-MAM(ours)*	64-64-64-64	51.70±0.42%	68.41±1.28%
MAML++ *	64-64-64-64	52.96±0.78%	70.02±0.55%
Minimax-MAML++(ours)*	64-64-64-64	**53.28±0.35%**	**71.70±0.23%**

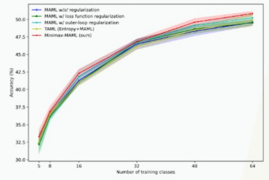


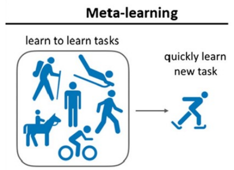
Figure 1. Mini-Imagenet 5-way 1-shot test accuracies (%) with varying training classes number.

References

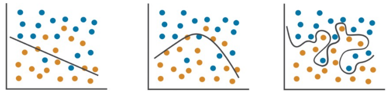
[1] Yao, Huaxiu, et al. "Improving Generalization in Meta-learning via Task Augmentation." *International Conference on Machine Learning*. PMLR, 2021.

Background: Meta-learning and Generalization

- **Meta-learning**:
  - Meta-learning represents a distinct class of machine learning algorithms.
  - It aims to enhance the learning ability of agents by learning from past experiences.
  - Meta-learning focuses on the performance of models on **new/unseen** tasks, enabling models to quickly adapt and achieve good performance on new tasks.



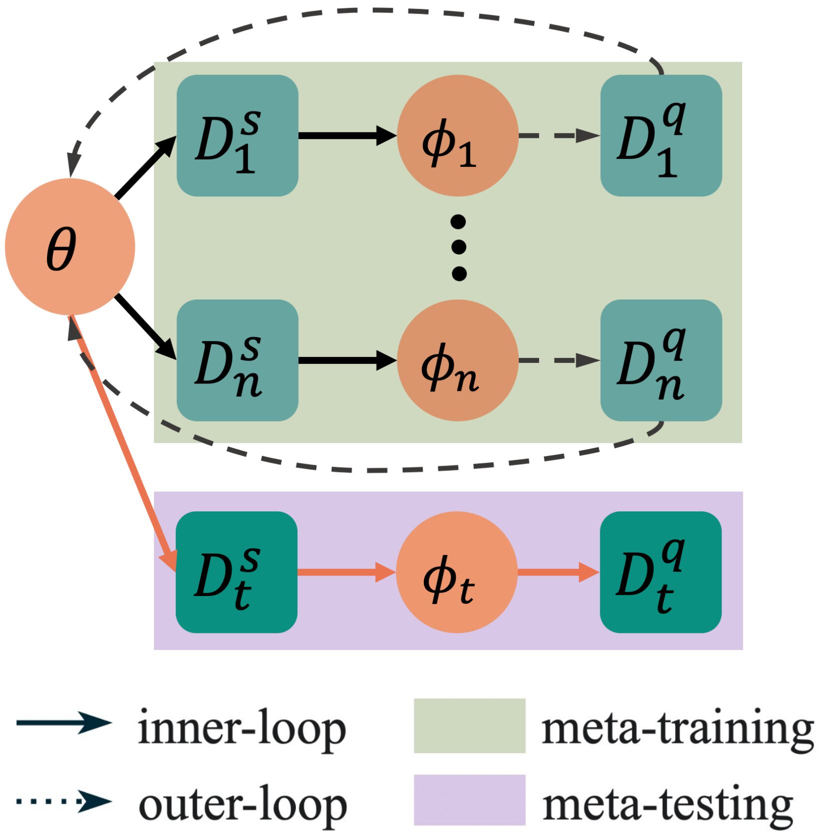
- **Generalization**:
  - Generalization ability is a critical characteristic for machine learning algorithms
  - It emphasizes the consistent performance of models on new/unseen tasks and training data, beyond their training performance.
  - It ultimately refers to the ability of a machine learning model to perform well on **unseen** or **new** data after being trained on a limited dataset.



For meta-learning algorithms, it is natural and crucial to consider their generalization ability.

(The goal of meta-learning is to equip models with the capability to adapt well to new tasks and consistently perform on new/unseen tasks.)

Background: Bi-level Learning Architecture of Optimization-based Meta-learning



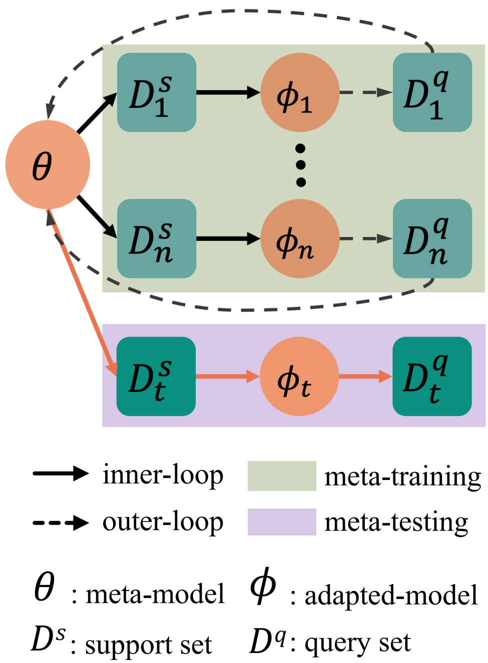
- Optimization-based approaches formulates the meta-learning problem as a **bi-level optimization** problem, to learn a meta-initialization that is well-generalized for individual new tasks. They have shown promising results in various domains, making them a widely adopted and representative choice in meta-learning research.

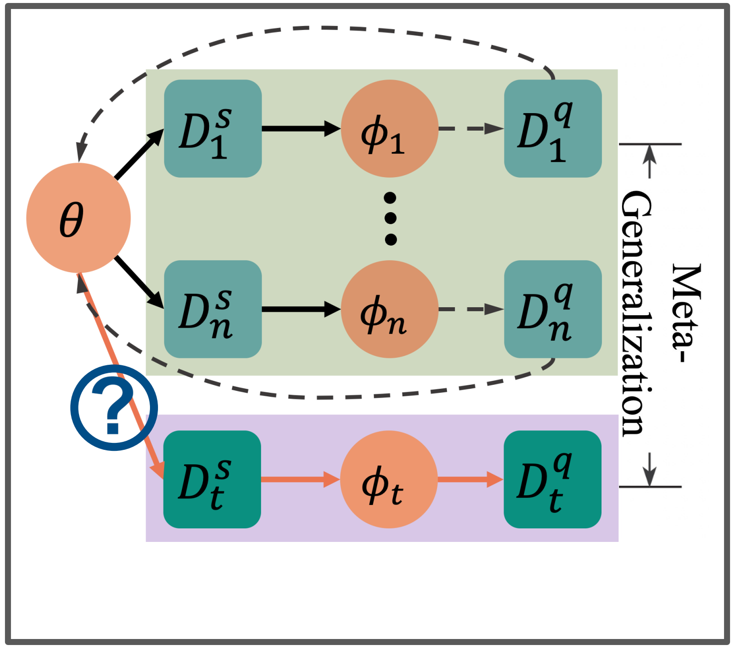
- During training, for i_{th} meta training step,
  - at the **inner-level** (inner-loop), a base model, which is initialized using the meta-model's parameters θ, adapts to i_{th} selected task by taking gradient descent steps over the support set D_i^S (usually few-shot data), obtaining adapted model ϕ_i.
  - at the **outer-level** (outer-loop), the meta-loss objective is calculated based on the adapted model ϕ_i 's predictions on task query set D_i^t. Then, the meta model θ 's parameter is optimized for the meta loss, helping to ensure that the meta model could effectively adapt to new task domain after simple update steps on minimal support data.

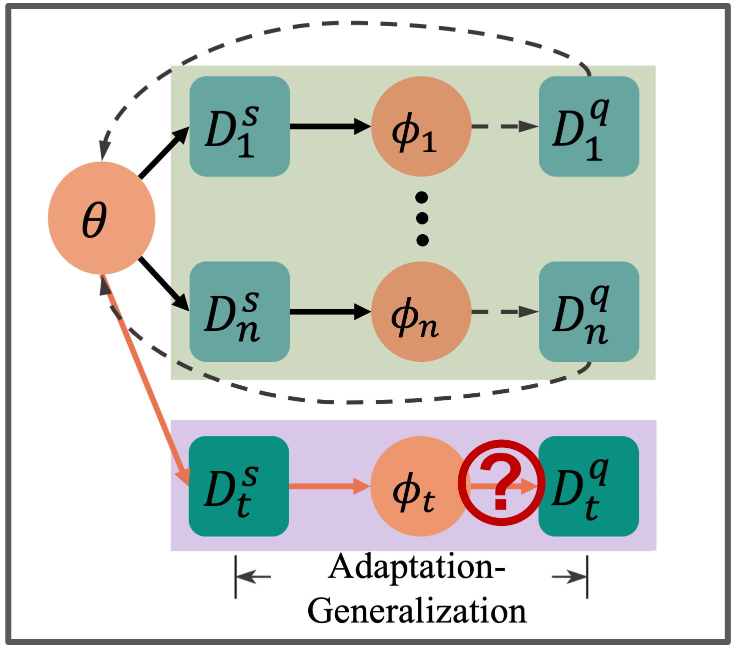
- This bi-level learning-to-learn process often involves optimizations for gradients-over-gradients, the final trained meta-model could be regarded as the model with good initialization to adapt to new tasks.

illustration Figure adapted from:
Yao, Huaxiu, et al. "Improving Generalization in Meta-learning via Task Augmentation." *International Conference on Machine Learning*. PMLR, 2021.

Background: Generalization Challenges of Bi-level Optimization-based Meta-Learning —— Meta-Generalization and Adaptation-Generalization







Meta-Generalization:
the meta-model should generalize to unseen tasks.
(requiring performance consistency between training tasks and new tasks.)

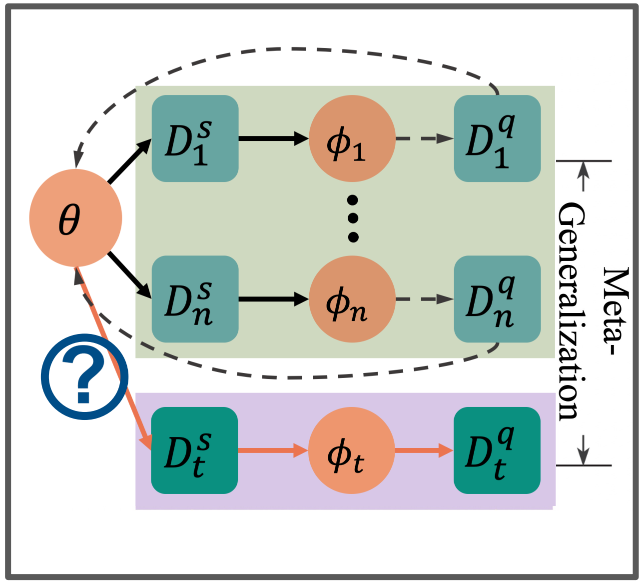
Adaptation-Generalization:
the adapted model should generalize to the domain of a specific task.
(requiring performance consistency between task support data and true data distribution of corresponding task.)

Limitations of Existing Methods
—— Focus on Meta-Generalization / Overlook Adaptation-Generalization

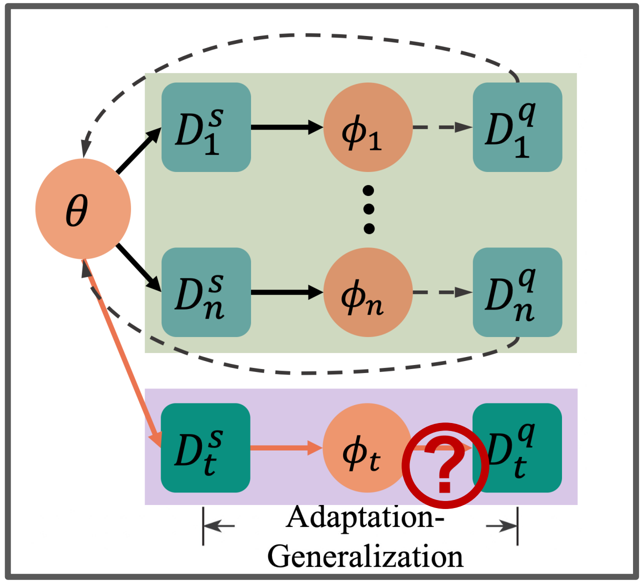
Methods addressing Meta-Generalization:

- Meta-Regularization.
- Meta-Augmentation.
- Task-Augmentation.
- Meta-Dropout.
- Meta-Memorization Analysis.
- Bayesian Methods.

…



Meta-Generalization:
the meta-model must generalize to unseen tasks.



Adaptation-Generalization:
the adapted model must generalize to the domain of a specific task.

Methods addressing Adapatation-Generalizaton:

?
(underexploring)

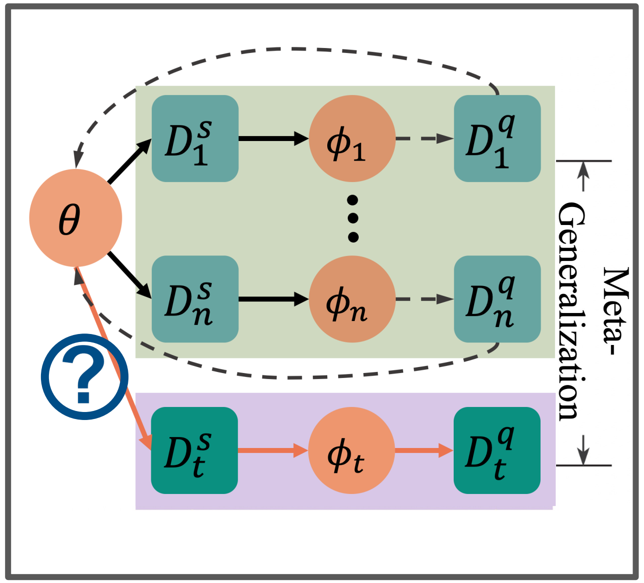
Limitations of Existing Methods

—— Focus on Meta-Generalization / Overlook Adaptation-Generalization

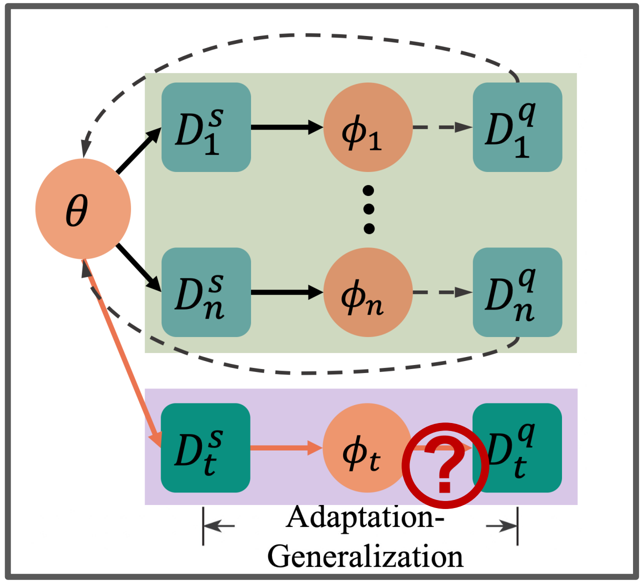
Methods addressing Meta-Generalization:

- Meta-Regularization.
- Meta-Augmentation.
- Task-Augmentation.
- Meta-Dropout.
- Meta-Memorization Analysis.
- Bayesian Methods.
- Minimax-Meta Regularization (ours)

…



Meta-Generalization:
the meta-model must generalize to unseen tasks.



Adaptation-Generalization:
the adapted model must generalize to the domain of a specific task.

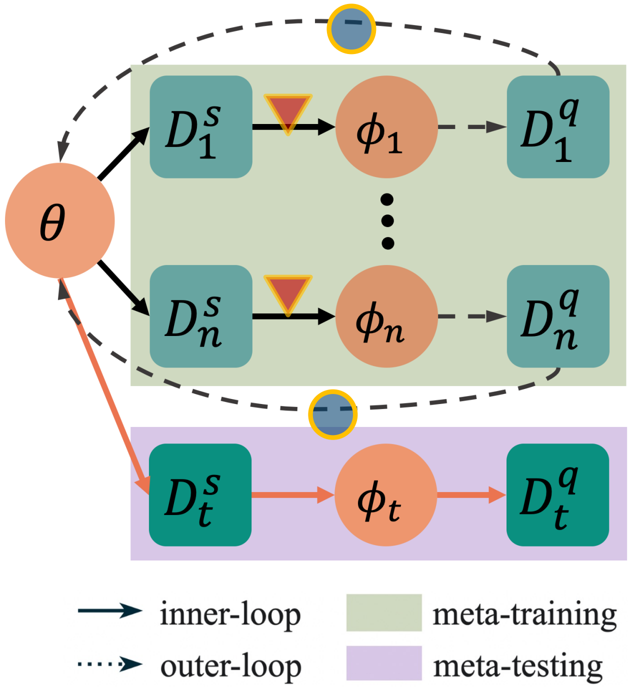
Methods addressing Adaptation-Generalization:

- **Minimax-Meta Regularization** (ours)

Method: Bi-level Minimax-Meta Regularization for Meta-learning
—— a Reg-framework Considering **both** Meta-Generalization and Adaptation-Generalization

Minimax-Meta Regularization method is designed to improve the generalization performance of bi-level meta-learning by combining two types of regularizations during training:

- an **ordinary regularization** at the outer-level to encourage the meta-model to learn more generalized hypotheses (for meta-generalization).
- *an **inverted regularization** at the inner-level to increase the generalization difficulty of adapted model, thus help the meta-model improve generalization during training (for adaptation-generalization).



: **ordinary regularization,** can be any classic regularization term, such as L1/L2-Norm or information entropy regularization, which encourages the meta-model to learn more conservative / generalized hypotheses.

: **inverted regularization,** on the contrast, should be adopting an inverted regularization term, which could typically be achieved by **changing the sign** of an ordinary regularization term (e.g., negative L1/L2-Norm, inverted entropy regularization), and this increases the adaptation difficulty and forces the meta-model to learn better-generalized hypotheses.

(note that inverted regularization is not adopted during testing phase.)

Method: Bi-level Minimax-Meta Regularization for Meta-learning
—— a Reg-framework Considering **both** Meta-Generalization and Adaptation-Generalization

Example Usage — Application to Model-Agnostic Meta-learning (MAML) Algorithm:

(Minimax-Meta Regularization is easy for implementation, oftentimes only involve modifications to inner- and outer-loss functions.)

inverted regularization at the inner-level :

ordinary regularization at the outer-level :

Algorithm 1 Minimax-MAML

Require: Datasets $\mathcal{S} = \{\mathcal{S}_i^{\text{in}}, \mathcal{S}_i^{\text{out}}\}_{i=1}^{m}$; total number of iterations T; regularization coefficients σ^{in} and σ^{out}.

Initialize the meta-model w^0
for $t = 0$ to $T-1$ **do**
  Randomly sample r tasks with indices stored in $\mathcal{B}_t$;
  for each sampled task $\mathcal{T}_i$ **do**
    Sample a support data batch $\mathcal{D}_i^{t,\text{ in}}$ from $\mathcal{S}_i^{\text{in}}$;
    Sample a query data batch $\mathcal{D}_i^{t,\text{ out}}$ from $\mathcal{S}_i^{\text{out}}$;
    (Inner-level) Compute per-task adapted parameters with gradient descent:

$$w'^{t}_i := w^t - \alpha \nabla_{w^t}\left(\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right) + \sigma^{in} Inverted\_Reg\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right)\right);$$

    (Outer-level) SGD step for meta-model, save per-task meta-weight for meta-update:

$$w_i^{t+1} := w^t - \beta_t \nabla_{w^t}\left(\hat{\mathcal{L}}\left(w'^{t}_i, \mathcal{D}_i^{t,\text{ out}}\right) + \sigma^{out} Ordinary\_Reg\left(w'^{t}_i, \mathcal{D}_i^{t,\text{ out}}\right)\right);$$

  end for
  Meta-update $w^{t+1} := \frac{1}{r}\sum_{i\in\mathcal{B}_t} w_i^{t+1}$
end for
Return: w^T

Method: Inner-level **Inverted** Regularization for Adaptation-Generalization

Inverted Regularization at Inner-level

- In Minimax-Meta Regularization, the inverted regularization is applied at the inner-level of meta-learning.

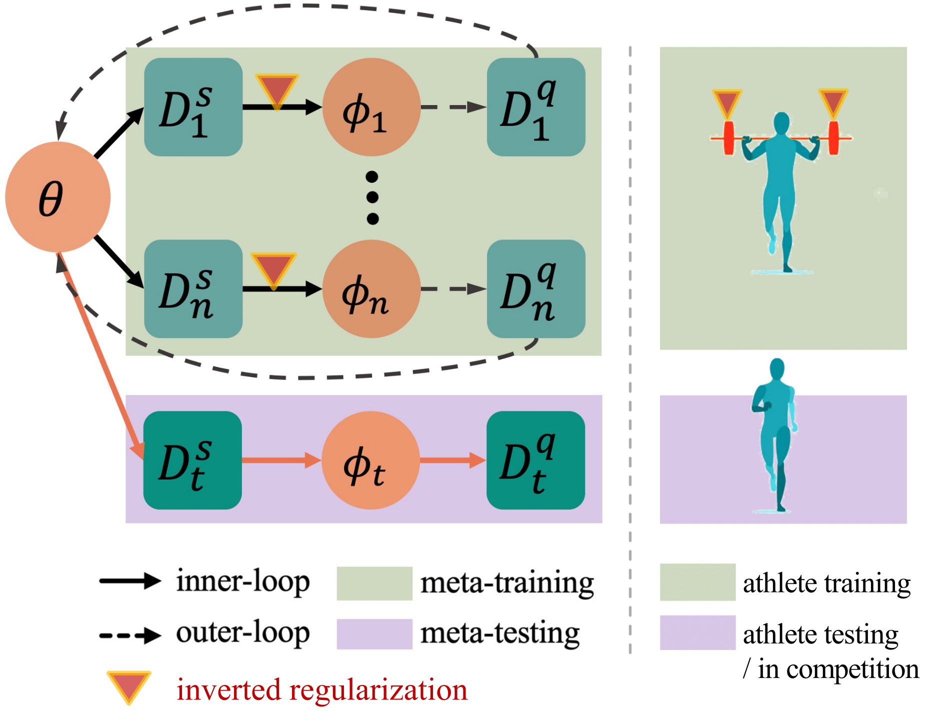


illustration of intuition for inverted regularization at inner-level: the "loaded training" athlete.

Intuition for Improving Adaptation-Generalization with Inverted Regularization at Inner-level

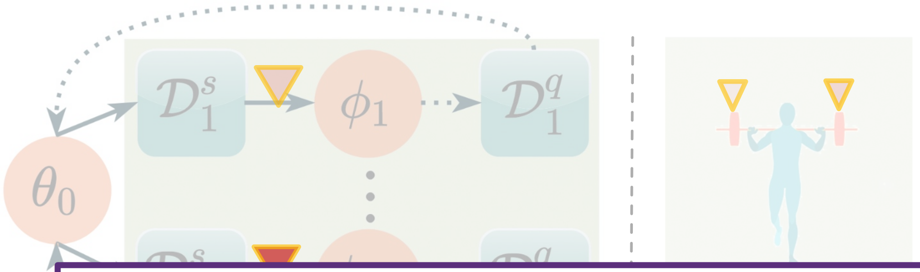
- The intuition behind using inverted regularization at the inner-level is to improve the meta-model's generalization by increasing adaptation difficulty during training.
- The inverted regularization term encourages the adapted model to learn more challenging and less generalizable hypotheses.
- By making the adapted model more difficult to fit the meta-support set, the meta-model is pushed to learn better-generalized meta-knowledge.
- This can be seen as a form of "adversarial training" or "loaded training" for the meta-model, enhancing its generalization performance.

- (Importantly, the "adversarial training" is only applied during training and not during meta-testing, allowing the meta-model to perform well in new environments without carrying the training burden)

Method: Inner-level **Inverted** Regularization for Adaptation-Generalization

Inverted Regularization at Inner-level

- In Minimax-Meta Regularization, the inverted regularization is applied at the inner-level of meta-learning.



However, the concept of using inverted regularization at the inner-level to improve generalization may seem either <u>too intuitional</u> or <u>counterintuitive</u> to some, it's crucial that we provide a <u>**theoretical analysis**</u> in the next section to support its utility.

Intuition for Improving Adaptation-Generalization with Inverted Regularization at Inner-level

- The intuition behind using inverted regularization at the inner-level is to <u>improve the meta-model's generalization</u> by <u>increasing adaptation difficulty</u> during training.
- The inverted regularization term ▼ encourages the <u>adapted model</u> to learn more <u>challenging</u> and <u>less generalizable</u> hypotheses.
- By making the <u>adapted model more difficult to fit the meta-support set</u>, the <u>meta-model is pushed to learn better-generalized meta-knowledge.</u>
- This can be seen as a form of <u>"adversarial training"</u> or <u>"loaded training"</u> for the meta-model, enhancing its generalization performance.

- (Importantly, the "adversarial training" is only applied during training and not during meta-testing, allowing the meta-model to perform well in new environments <u>without carrying the training burden</u>)

Theoretical Analysis: Inner-level Inverted Regularization Improves Generalization —— Preliminary

We provide an analysis of the effectiveness of inverted regularization in meta-learning by taking **L2-Norm regularization** at the inner-level of the **single-step MAML** algorithm as a typical example, which is very possible to generalize to other regularization.

The effectiveness would be proved by deriving generalization bound while inner-level regularization is adopted. The derivation is mainly based on a convex analysis approach, and the results are given under the strongly-convex loss assumptions.

Assumptions:

Assumption 1. *We assume the function $\ell(\cdot, z)$ satisfies the following properties for any $z \in \mathcal{Z}$:*

1. *(Strong convexity) $\ell(\cdot, z)$ is μ-strongly convex, i.e., $(\nabla \ell(w, z) - \nabla \ell(u, z))^T (w - u) \geq \mu \|w - u\|^2$;*
2. *(Lipschitz in function value) $\ell(\cdot, z)$ has gradients with norm bounded by G, i.e., $\|\nabla \ell(w, z)\| \leq G$;*
3. *(Lipschitz gradient) $\ell(\cdot, z)$ is L-smooth, i.e., $\|\nabla \ell(w, z) - \nabla \ell(u, z)\| \leq L \|w - u\|$;*
4. *(Lipschitz Hessian) $\ell(\cdot, z)$ has ρ-Lipschitz Hessian, i.e., $\left\|\nabla^2 \ell(w, z) - \nabla^2 \ell(u, z)\right\| \leq \rho \|w - u\|$*

MAML with inner-level L2-Norm regularization.

The updating rule of MAML get changed from

$$w_i^{t+1} := w^t - \beta_t \nabla_{w^t} \hat{\mathcal{L}}\left(w^t - \alpha \nabla \hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right), \mathcal{D}_i^{t,\text{ out}}\right)$$

to

$$w_i^{t+1} := w^t - \beta_t \nabla_{w^t} \hat{\mathcal{L}}\left(w^t - \alpha \nabla_{w^t}(\hat{\mathcal{L}}\left(w^t, \mathcal{D}_i^{t,\text{ in}}\right) \boxed{+ \frac{\delta}{2}\|w^t\|^2})), \mathcal{D}_i^{t,\text{ out}}\right)$$

where δ is the regularization parameter. Here δ can be either positive or negative to represent the ordinary or inverted regularization, respectively. We treat δ as a variable and analyze how its value would influence the generalization error.

Theoretical Analysis: Inner-level Inverted Regularization Improves Generalization —— **Generalization Error** and **Training Bias** with Regularization

Without Regularization, the Decomposition of Test Error

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A},\mathcal{S}}\left[F(\mathcal{A}(\mathcal{S})) - \min_{\mathcal{W}} F\right] \quad \text{(test error)} \; = \\
&\underbrace{\mathbb{E}_{\mathcal{A},\mathcal{S}}\left[\hat{F}(\mathcal{A}(\mathcal{S}),\mathcal{S}) - \min_{\mathcal{W}} \hat{F}(\cdot,\mathcal{S})\right]}_{\text{training error}} + \boxed{\underbrace{\mathbb{E}_{\mathcal{A},\mathcal{S}}[F(\mathcal{A}(\mathcal{S})) - \hat{F}(\mathcal{A}(\mathcal{S}),\mathcal{S})]}_{\text{generalization error}}} + \underbrace{\mathbb{E}_{\mathcal{S}}\left[\min_{\mathcal{W}} \hat{F}(\cdot,\mathcal{S})\right] - \min_{\mathcal{W}} F}_{\leq 0}
\end{aligned}
\tag{1}
$$

With Regularization, the Decomposition of Test Error:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}\left[F(\tilde{\mathcal{A}}(\mathcal{S})) - \min_{\mathcal{W}} F\right] \quad \text{(test error)} \; = \\
&\underbrace{\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}\left[\hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S}) - \hat{F}(\arg\min_{\mathcal{W}} \hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S})\right]}_{\text{training error}} + \boxed{\underbrace{\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}[F(\tilde{\mathcal{A}}(\mathcal{S})) - \hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S})]}_{\text{generalization error}}} + \underbrace{\mathbb{E}_{\mathcal{S}}\left[\min_{\mathcal{W}} \hat{F}(\cdot,\mathcal{S})\right] - \min_{\mathcal{W}} F}_{\leq 0} \\
&+ \boxed{\underbrace{\mathbb{E}_{\mathcal{S}}\left[\hat{F}(\arg\min_{\mathcal{W}} \hat{\tilde{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \min_{\mathcal{W}} \hat{F}(\cdot,\mathcal{S})\right]}_{\text{training bias}}}
\end{aligned}
\tag{2}
$$

The process of adding regularization often involves changes to the loss function during training.
In other words, if the model is obtained by a new regularized algorithm $\tilde{\mathcal{A}}$, it is usually optimized for a different function $\tilde{F}(\cdot)$ instead of the original $F(\cdot)$. As a result, we need to consider the **Training Bias** caused by this change of learning objective. Instead of directly adopting (1)'s decomposition, we need to further decompose the test error as in (2).

Theoretical Analysis: Inner-level Inverted Regularization Improves Generalization —— results and properties for **Generalization Error Bound**

Algorithm Stability Based Approach for Deriving the Generalization Error Bound.

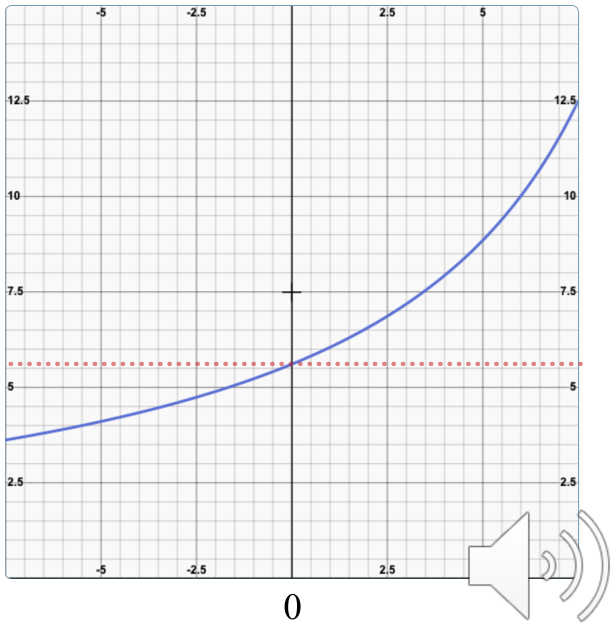
Key Lemma: If a meta-learning algorithm is (γ, K)-stable, the generalization error is then bounded by γ.
(Intuitively, the term "(γ, K)-stable" refers to a property where the error difference between two models trained on datasets with only K different samples is limited to γ, indicating that the models' performance is relatively stable even with limited data variability.)

Generalization Error Bound, Result:

$$\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}[F(\tilde{\mathcal{A}}(\mathcal{S})) - \hat{F}(\tilde{\mathcal{A}}(\mathcal{S}),\mathcal{S})] \leq$$
$$\frac{2G^2(1+\alpha L)(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}\left(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2 L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}\right)$$

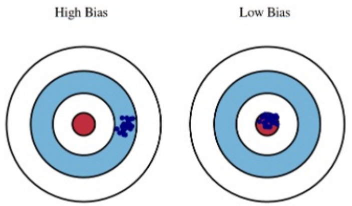
Properties with δ:

The generalization bound could be regarded as a function $GB(\delta)$, and its derivative $GB'(\delta)$ is always positive within the defined domain $\delta \in \left(-\infty, \frac{1}{2\alpha}\right)$.
It suggests that $GB(\delta)$ is monotonically increasing, implying that L2 regularization at the inner-level decreases the generalization bound of MAML only when it's inverted (i.e. $\delta < 0$). And ordinary regularization at the inner-level would increase the generalization bound.



Theoretical Analysis: Inner-level Inverted Regularization Improves Generalization —— results and properties for **Training Bias Bound**

Key Technique: We establish an equivalence relationship for training bias, enabling us to transform the analysis involving differences in optimal values between two functions into an analysis involving only a single function.



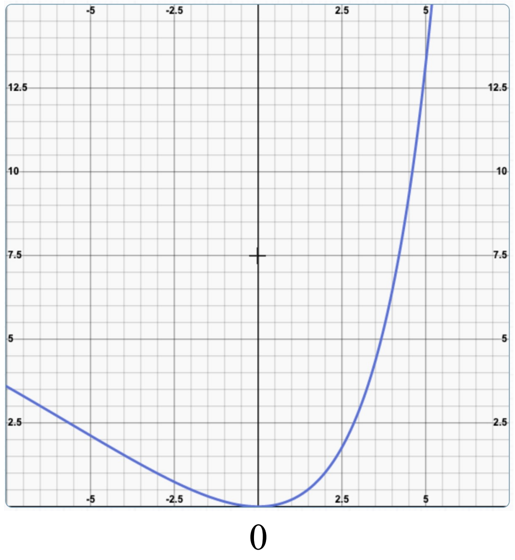
Training Bias Bound Result:

$$\mathbb{E}_{\mathcal{S}}\left[\hat{F}(\arg\min_{\mathcal{W}}\hat{\hat{F}}(\cdot,\mathcal{S}),\mathcal{S}) - \min_{\mathcal{W}}\hat{F}(\cdot,\mathcal{S})\right] \le \frac{\alpha^2(\alpha\rho G+(1-\alpha\mu)^2L)((1-\alpha\mu-\alpha\delta)L\|w^*\|+G)^2\delta^2}{2(-\alpha\rho G+(1-\alpha L-\alpha\delta)^2\mu)^2}$$

Properties with δ:

The training bias bound could also be regarded as a function $TB(\delta)$.

- We could observe that $TB(\delta) > TB(0) = 0$ for $\delta \neq 0$, which suggests that <u>training bias is inevitabl</u>e when regularization is adopted.
- Another important finding is that for any legal choice of $\delta_0 > 0$, we have <u>$TB(-\delta_0) < TB(\delta_0)$, which suggests that the inverted regularization has less corruption</u> to training bias bound at the inner-level than the ordinary regularization with the same coefficient magnitude.



Theoretical Analysis: Inner-level Inverted Regularization Improves Generalization —— Finally, the **Total Test Error** Bound.

Based on the previous analysis, we could obtain the **Total Test Error** bound by combining the **Generalization Error Bound** and **Training Bias Bound**.

Test Error Bound Result:

$$\mathbb{E}_{\tilde{\mathcal{A}},\mathcal{S}}\left[F(\tilde{\mathcal{A}}(\mathcal{S})) - \min_{\mathcal{W}} F\right] \leq \underbrace{\frac{2G^2(1+\alpha L)(1-\alpha\mu-\alpha\delta+(2+\alpha L-\alpha\delta)\alpha LK)}{mn}\left(\frac{1}{\alpha\rho G+(1-\alpha\delta-\alpha\mu)^2 L}+\frac{1}{-\alpha\rho G+(1-\alpha\delta-\alpha L)^2\mu}\right)}_{\text{generalization error bound } GB(\delta)}$$

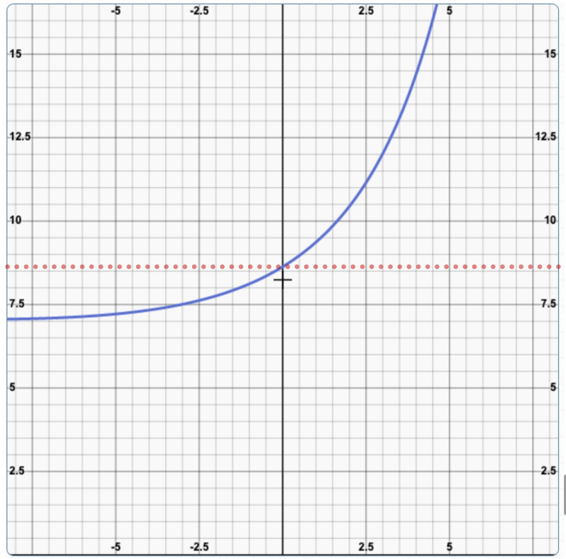
$$+\underbrace{\frac{\alpha^2(\alpha\rho G+(1-\alpha\mu)^2 L)((1-\alpha\mu-\alpha\delta)L\|w^*\|+G)^2\delta^2}{2(-\alpha\rho G+(1-\alpha L-\alpha\delta)^2\mu)^2}}_{\text{training bias bound } TB(\delta)}$$

Properties with δ:

The test error bound could be described by $TE(\delta) \coloneqq TB(\delta) + GB(\delta)$.

- When δ is positive, we have $TB(\delta) > TB(0)$ and $GB(\delta) > GB(0)$, which suggests ordinary regularization at the inner-level worsens the model's test error bound.
- Instead, for inverted regularization, since $TE'(0) = TB'(0) + GB'(0) = 0 + GB'(0) > 0$, there must be an interval $[\delta^*, 0)$ in which all values can be used as the inverted regularization parameter to decrease the test error bound.

These results are validated in the experiment section.



0

Bi-level Minimax-Meta Regularization: Applications and Experiments —— Few-shot Classification

Based on the preceding findings, we have observed that the bi-level Minimax-Meta Regularization, which combines inverted regularization at the inner level and ordinary regularization at the outer level, is a method that may exhibit improvements in both Adaptation-Generalization and Meta-Generalization.

In this section, we empirically validate this approach, starting with the classic **few-shot classification task**.

Experiment Settings:

Datasets

- **Mini-ImageNet** Dataset:
  - Derived from ImageNet, it consists of 600 instances from 100 classes.
  - We split the Mini-ImageNet dataset into 64 classes for training, 12 classes for validation, and 24 classes for testing.
- **Omniglot** Dataset:
  - Contains 1623 character classes with different alphabets.
  - Each class has 20 instances.
  - Divided into training, validation, and test sets, with 1150, 50, and 423 instances, respectively.

Methods

- Minimax-MAML with Norm Regularization
- Minimax-MAML with Norm & Entropy Combined Regularization
- Minimax + Other MAML variants. (e.g., MAML++, fo-MAML)

First Experiment: **Empirical Verification** for the Theoretical Results —— Few-shot Classification, Mini-ImageNet Dataset

The first experiment is for empirically verifying the insights & theoretical results that inner- and outer-level regularizations should be respectively be inverted and ordinary to improve generalization performance.

Note that our bi-level Minimax-Meta Regularization is a framework that is compatible with regularizations beyond L2-Norm. ↓

Mini-Imagenet 5-way Few-shot Classification for MAML *(Reg Objective: L2-Norm)*				
Regularization Type	Outer Reg	Inner Reg	1-Shot	5-Shot
no regularization	-	-	49.58±0.45%	65.39±0.50%
regularize the outer-level	*Ordinary*	-	49.90±0.54%	66.47±1.21%
regularize the inner-level	-	*Ordinary*	49.28±0.37%	64.80±0.25%
invertedly regularize the inner-level	-	*Inverted*	49.92±0.42%	66.05±0.68%
Minimax-Meta Regularization	*Ordinary*	*Inverted*	**50.25±0.38%**	**68.17±0.92%**

Mini-Imagenet 5-way Few-Shot Classification for MAML $\boxed{\textit{(Reg Objective: L2-Norm \& Entropy)}}$				
Regularization Type	Outer Reg	Inner Reg	1-Shot	5-Shot
no regularization	-	-	49.58±0.45%	65.39±0.50%
regularize the outer-level	*Ordinary*	-	50.23±0.67%	67.18±0.88%
regularize the inner-level	-	*Ordinary*	48.07±1.01%	64.32±0.35%
invertedly regularize the inner-level	-	*Inverted*	49.96±0.33%	65.91±0.41%
Minimax-Meta Regularization	*Ordinary*	*Inverted*	**50.85±0.37%**	**69.36±0.34%**

Observations:

Inner-level inverted regularization enhances the generalization performance.
Inner-level ordinary regularization impairs the generalization performance.
Outer-level ordinary regularization enhances the generalization performance.
The outer-level ordinary regularization and inner-level inverted regularization are compatible.
Inner-level inverted regularization and the outer-level ordinary regularization are suitable for combined regularizer

Few-Shot Classification: Comparing with Representative Approaches

Omniglot 20-way 1-shot experiment

Omniglot 20-way 1-Shot Classification	
Approach	Accuracy
Meta-SGD(Li et al., 2017)	95.93±0.38%
Prototypical Net(Snell et al., 2017)	96.00%
Meta-Networks(Munkhdalai et al., 2017)	97.00%
GNN(Garcia et al., 2018)	97.40%
Relation Network(Sung et al., 2018)	97.60±0.20%
R2-D2(Bertinetto et al., 2019)	96.24±0.05%
SNAIL(Mishra et al., 2018)	97.64±0.30%
TAML(Entropy)(Jamal et al., 2019)	95.62±0.50%
MAML(Finn et al., 2017)*	94.20±0.41%
Minimax-MAML(ours)*	95.76±0.39%
MAML++(Antoniou et al., 2018)*	97.21±0.51%
Minimax-MAML++(ours)*	**97.77±0.06%**

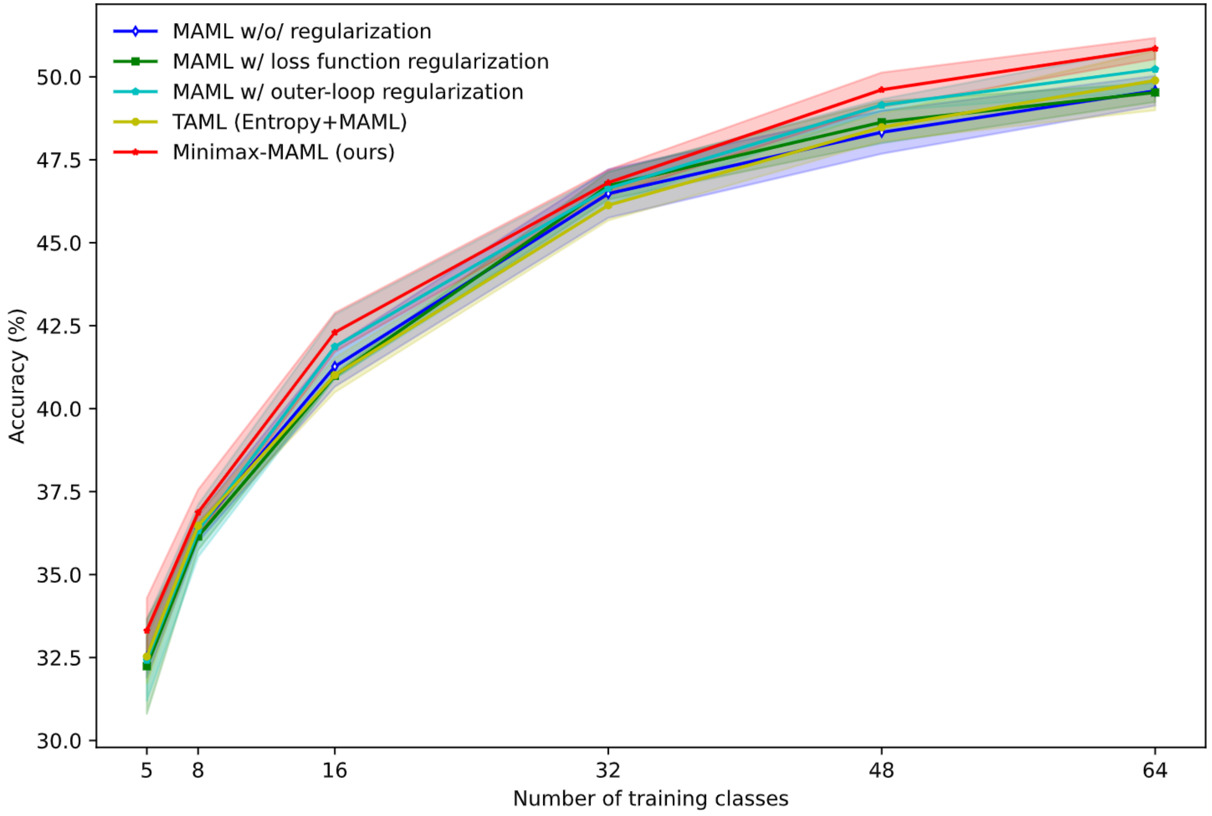
Mini-ImageNet 5-way few-shot experiment

Mini-Imagenet 5-way Few-Shot Classification			
Approach	Backbone	1-Shot Accuracy	5-Shot Accuracy
Meta-SGD(Li et al., 2017)	64-64-64-64	50.47±1.87%	64.03±0.94%
Prototypical Nets(Snell et al., 2017)	64-64-64-64	49.42±0.78%	68.20±0.66%
LLAMA(Grant et al., 2018)	64-64-64-64	49.40±1.83%	-
Meta-Networks(Munkhdalai et al., 2017)	64-64-64-64-64	49.21±0.96%	-
GNN(Garcia et al., 2018)	64-96-128-256	50.33±0.36%	66.41±0.63%
Relation Network(Sung et al., 2018)	64-96-128-256	50.44±0.82%	65.32±0.70%
R2-D2(Bertinetto et al., 2019)	64-64-64-64	49.50±0.20%	65.40±0.20%
LR-D2(Bertinetto et al., 2019)	96-192-384-512	51.90±0.20%	68.70±0.20%
MetaOptNet(Lee et al., 2019)	64-64-64-64	53.23±0.59%	69.51±0.48%
TAML(Entropy)(Jamal et al., 2019)	64-64-64-64	51.73±1.88%	66.05±0.85%
MAML-Meta Dropout(Lee et al., 2020)	32-32-32-32	51.93±0.67%	67.42±0.52%
MAML-MMCF(Yao et al., 2021)	32-32-32-32	50.35±1.82%	64.91±0.96%
MAML(Finn et al., 2017)*	64-64-64-64	50.20±1.65%	65.86±0.61%
Minimax-MAM(ours)*	64-64-64-64	51.70±0.42%	68.41±1.28%
MAML++(Antoniou et al., 2018)*	64-64-64-64	52.96±0.78%	70.02±0.55%
Minimax-MAML++(ours)*	64-64-64-64	**53.28±0.35%**	**71.70±0.23%**

Few-Shot Classification with Limited Tasks

- **Experiment Setup**
  - To further illustrate the generalization ability, we conducted a few-shot classification experiment with a limited number of training tasks.
  - For a dataset with M training classes available, there would be accordingly $\binom{M}{N}$ training tasks available. So we could restrict the number of training tasks by restricting the number of training classes.
  - We restricted the number of training classes to 48/32/16/8/5 to examine the effect of limited tasks.
  - Compared methods: *Meta-Minimax regularization*, *original MAML*, *MAML with outer-loop regularization*, *MAML with loss function regularization*, and *TAML (Entropy+MAML)*.

- **Results and Analysis**
  - Meta-Minimax regularization consistently outperforms other methods under the limited task number scenario.
  - Even with a very small number of tasks, Meta-Minimax regularization significantly improves accuracy compared to other methods.



Test accuracies (%) with varying training classes number. The shaded region denotes the 95% confidence interval.

Meta-Dataset Experiment with Larger Backbones

- **Experiment Motivation**
  - Test the applicability of our method to first-order methods and evaluate its performance on larger backbones and more complex datasets.
- **Experiment Setup**
  - Meta-Dataset is a dataset of datasets benchmark for meta-learning.
  - We conduct this experiment using first-order MAML (**fo-MAML**) and **ResNet-12** backbone on **Meta-Dataset**.
  - Experiment settings: Only training on ILSVRC training set, testing on ILSVRC testing set and 8 additional datasets.
  - Minimax-Meta Regularization was implemented for fo-MAML and compared with the original baseline version.
- **Results and Analysis**
  - Minimax-fo-MAML outperforms the baseline method on all 9 testing datasets.
  - The results indicate that Minimax-Meta Regularization improves generalization for first-order methods with larger backbones.

Method	ILSVRC (test)	Omniglot	Aircraft	Birds	Textures	QuickDraw	VGG Flower	Traffic	MSCOCO
fo-MAML	38.24±2.30	44.75±6.26	28.06±2.43	37.64±3.56	39.41±4.50	42.57±3.79	58.55±5.20	36.62±2.85	42.38±5.09
Minimax-fo-MAML(ours)	**40.53±1.54**	**68.43±3.53**	**30.95±2.97**	**41.09±0.40**	**45.12±1.41**	**51.57±2.68**	**66.23±0.89**	**38.83±2.71**	**45.15±0.85**

(Datasets except ILSVRC are only used for testing)

Conclusion

- In this paper, we have analyzed the challenges of meta-learning's generalization, including both meta-generalization and adaptation-generalization.
- We have proposed Minimax-Meta Regularization, a novel bi-level regularization-based approach that enhances both meta-generalization and adaptation-generalization.
- Theoretical analysis and extensive experiments have demonstrated the effectiveness of Minimax-Meta Regularization in improving the generalization performance of various meta-learning algorithms.
- Our work provides a new perspective on meta-learning's generalization and have the potential to contribute to the development of robust and effective meta-learning algorithms for real-world applications.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you!

Lianzhe

May 30, 2023