Large-Scale Meta-Learning with Continual Trajectory Shifting

JaeWoong Shin[1*], Hae Beom Lee[1*], Boqing Gong[3], Sung Ju Hwang[12]

[1]KAIST, South Korea

[2]AITRICS, South Korea

[3]Google, LA

[*]: Equal contribution

Motivation

Conventional meta-learnings are yet limited to few-shot learning, due to expensive computational cost.

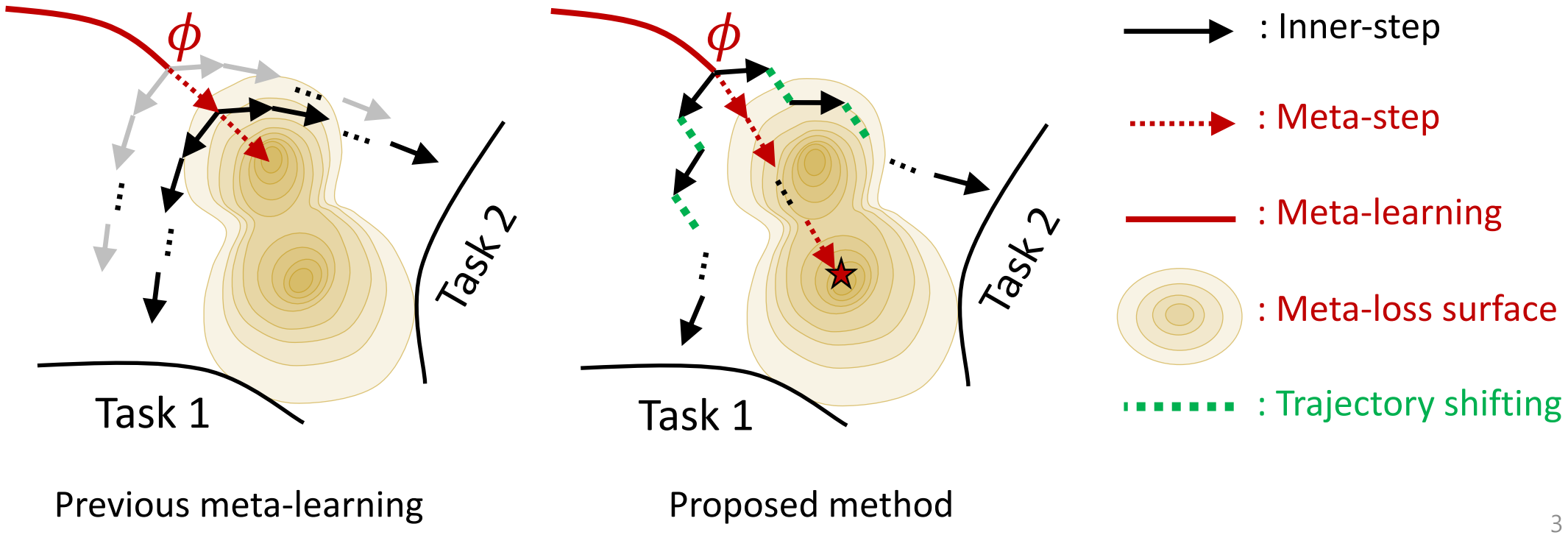
e.g.

- 2nd order meta-gradient
- bi-level optimization/discarded updates

Many-shot scenarios require longer inner optimization, which results in less frequent meta-updates.

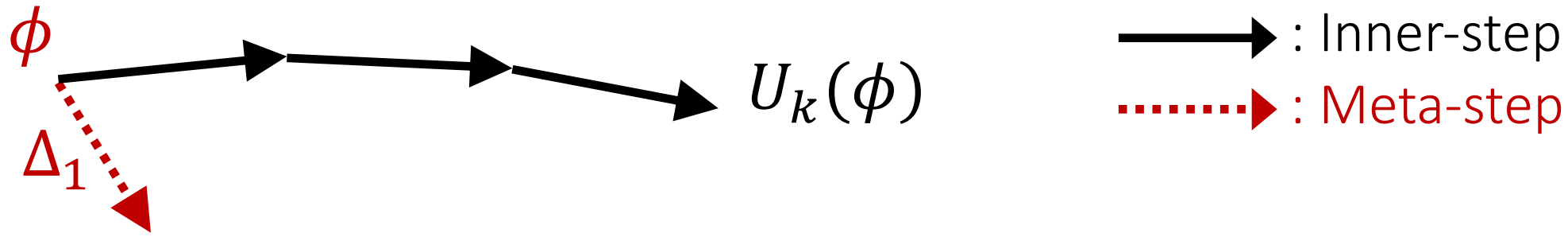
Concept

Increase the frequency of meta-updates even with the excessively long inner-optimization trajectories.



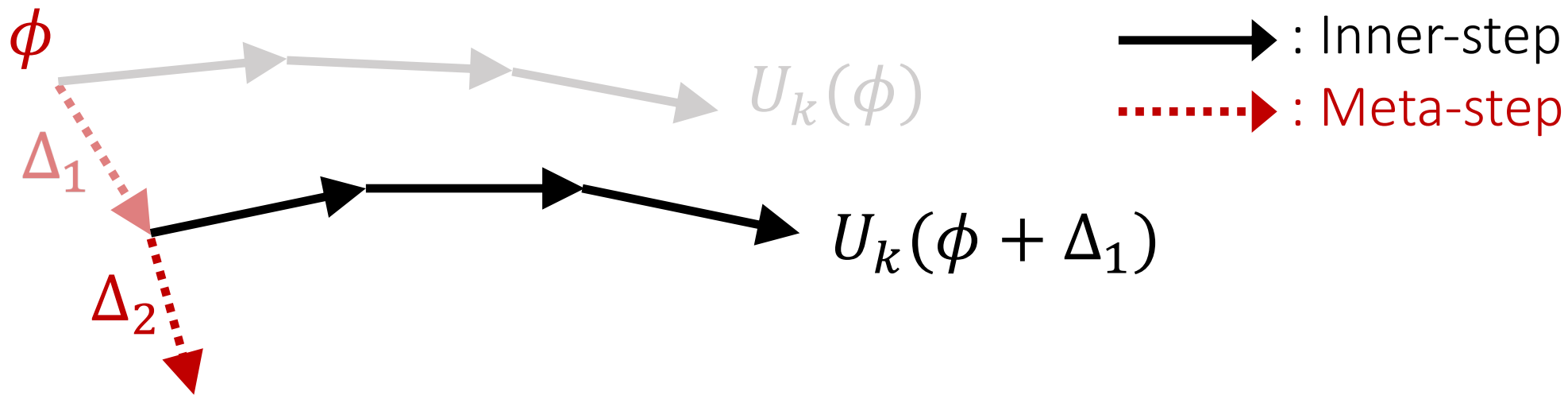
Previous Meta-Learning

Perform meta-updates after k inner optimization steps.



Previous Meta-Learning

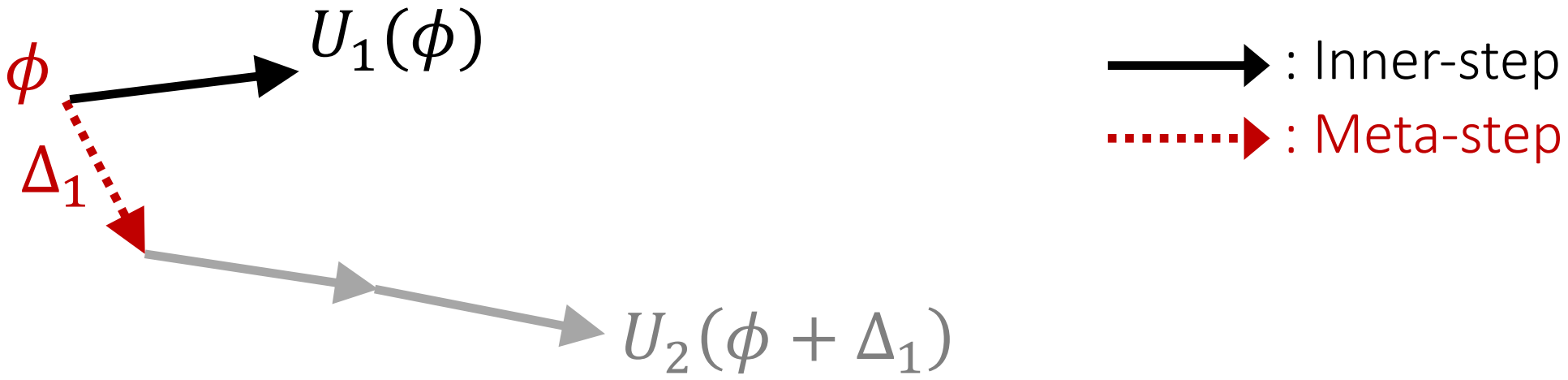
Perform meta-updates after k inner optimization steps.



Frequency of meta-updates: $\mathbf{1/k}$

Continual Trajectory Shifting

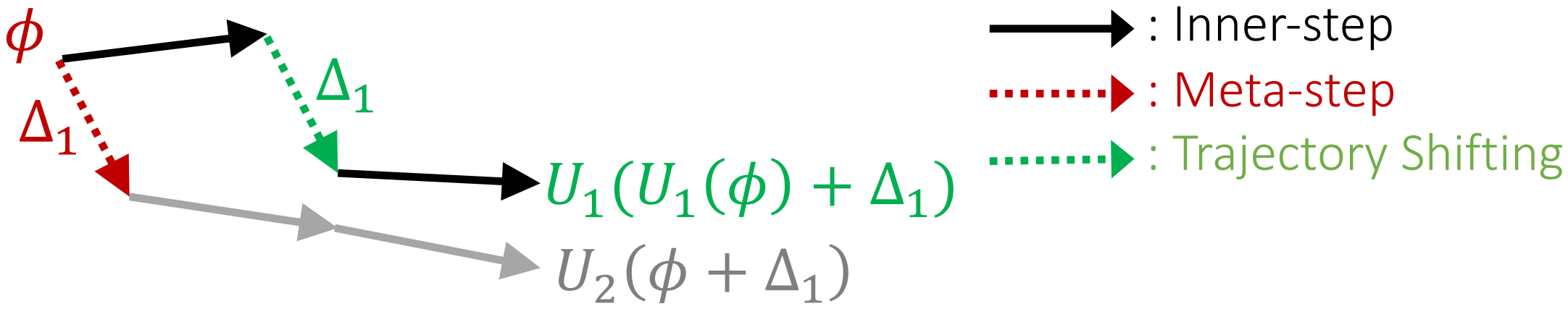
Interleave the meta-updates with the inner optimization processes.



Continual Trajectory Shifting

Interleave the meta-updates with the inner-optimization processes.

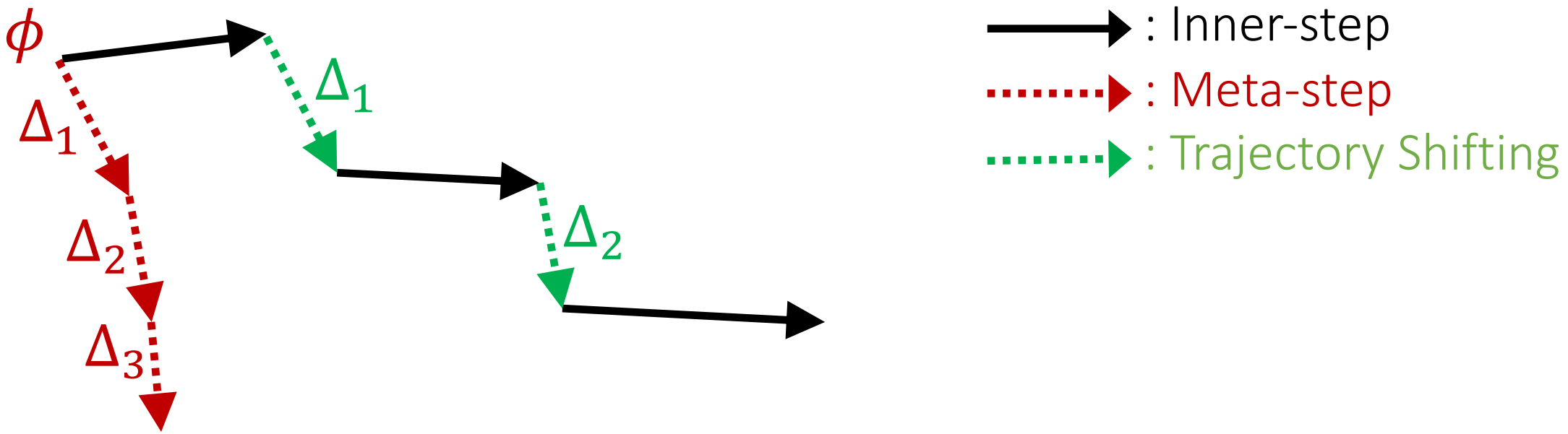
Then, estimate the required shift of the task-specific parameters.



Continual Trajectory Shifting

Interleave the meta-updates with the inner-optimization processes.

Then, estimate the required shift of the task-specific parameters.



Frequency of meta-updates: **1**

Continual Trajectory Shifting

From Talyor approximation and hessian approximation, we can derive:

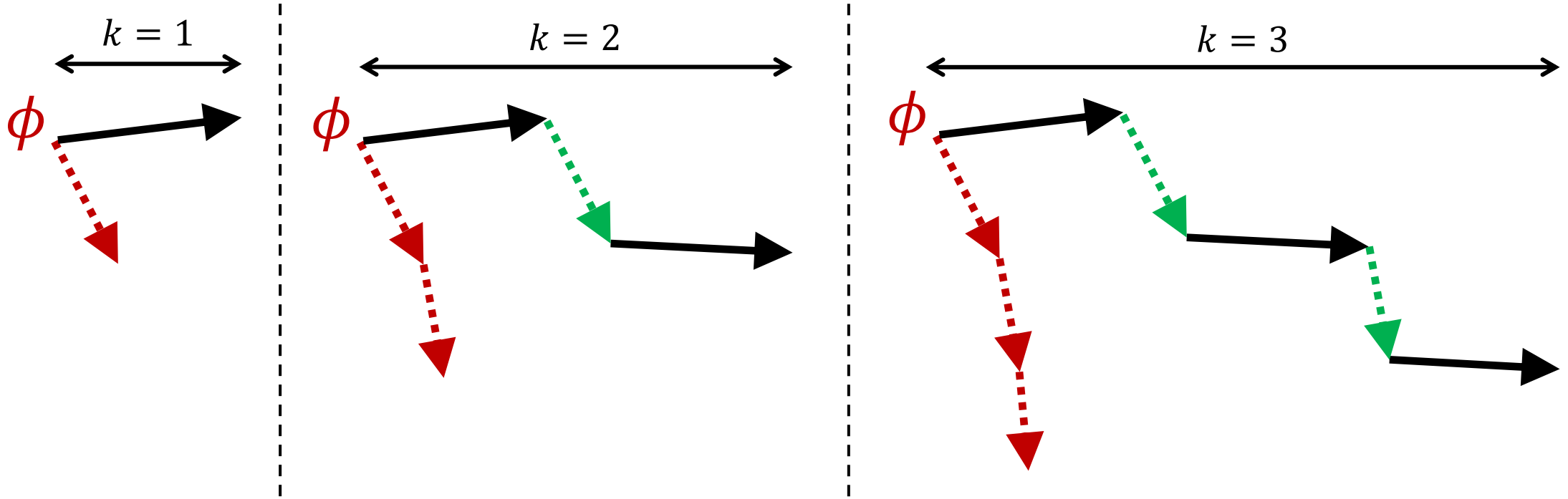
$$U_k(\phi + \Delta_1 + \cdots + \Delta_{k-1}) \approx U_1(\cdots U_1(U_1(\phi) + \Delta_1) \cdots + \Delta_{k-1})$$

However, there are some issues:

1. Approximation error will be accumulated as # of steps increases.
2. ReLU and max-pooling are not differentiable at a certain point.

Meta Curriculum Learning

The **inner-trajectory length** used to compute each meta-update **gradually increases** from 1 to maximum K.



Meta Curriculum Learning

When # of inner steps is small, information about task learning trajectories is limited due to short-horizon bias.

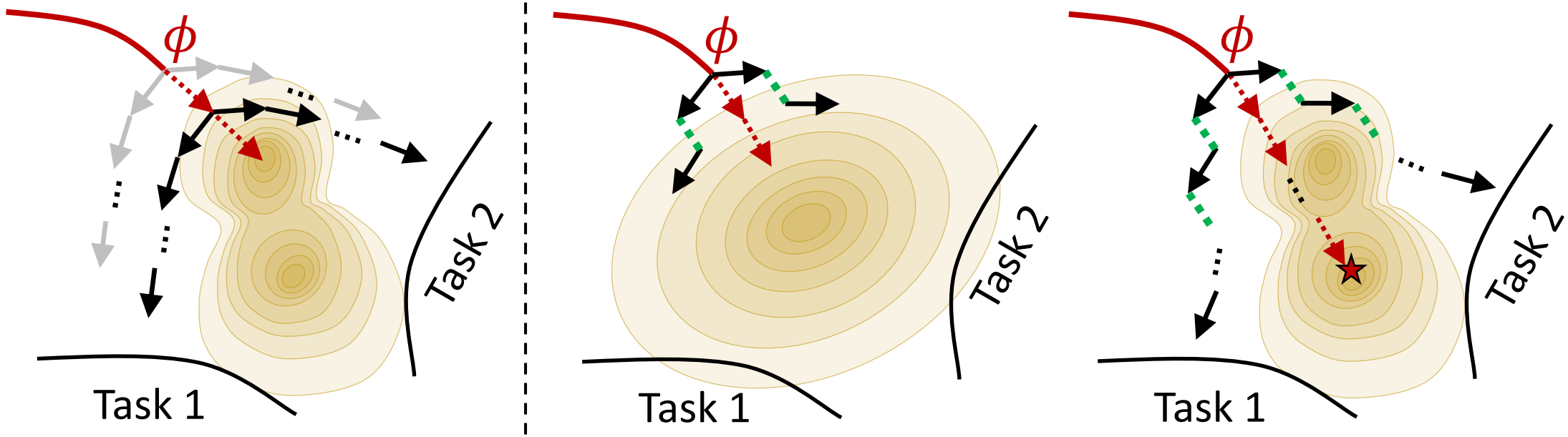
→ Simple meta-loss surface

When # of inner steps is large, local optimum can be reached from many in itializations, which results in many meta-level local optimum.

→ Complex meta-loss surface

Meta Curriculum Learning

Our algorithm goes through from the smooth to the complex meta-loss surface, thus it arrives at a better local optimum by the curriculum learning effect.



Previous meta-learning

Proposed method

Experiments

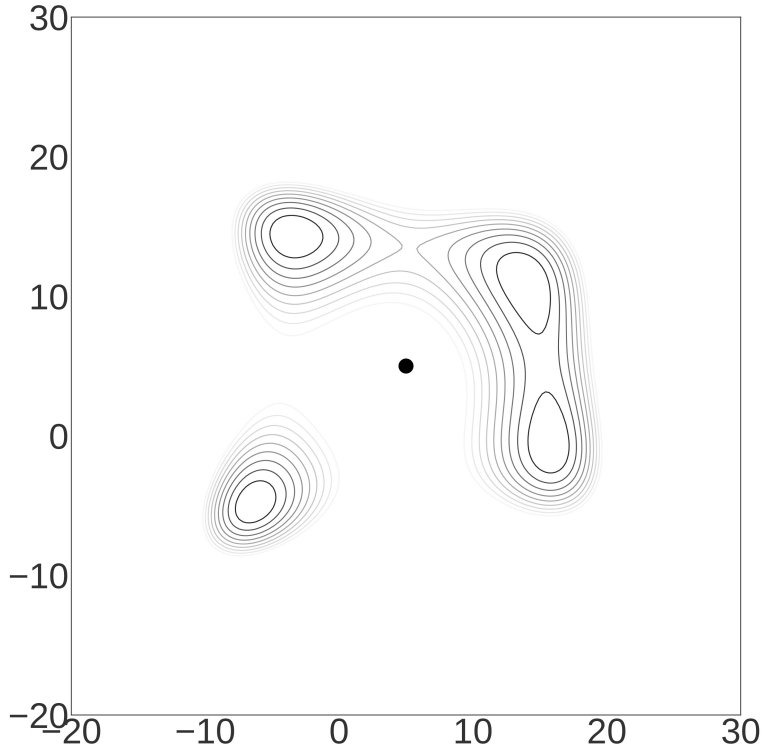
1. Synthetic experiments
   - How and why our method outperforms

2. Image classification
   - Realistic large-scale scenario

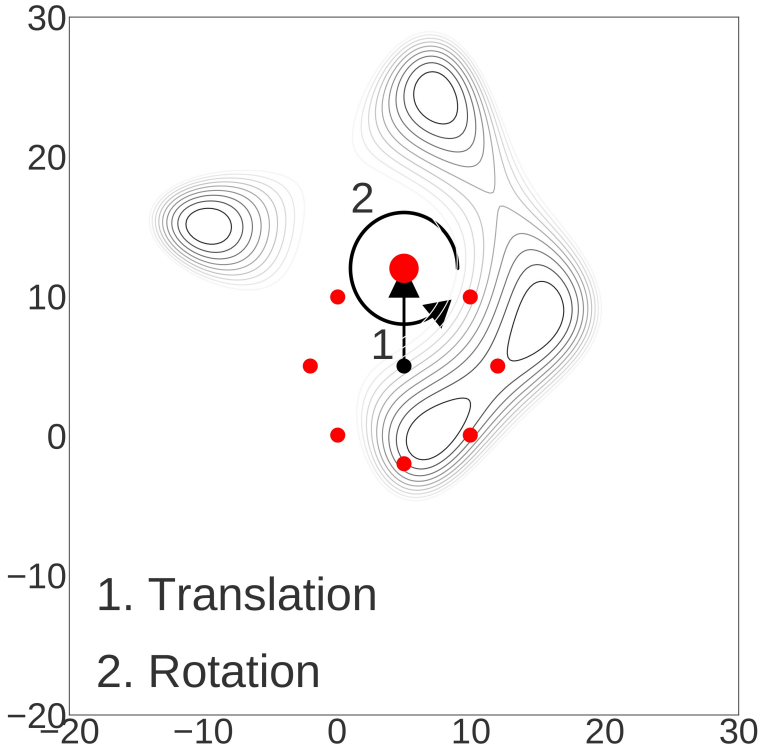
3. Improving on ImageNet Pre-trained Model

Synthetic Experiments

Set up 2D functions as tasks and perform meta-learning.



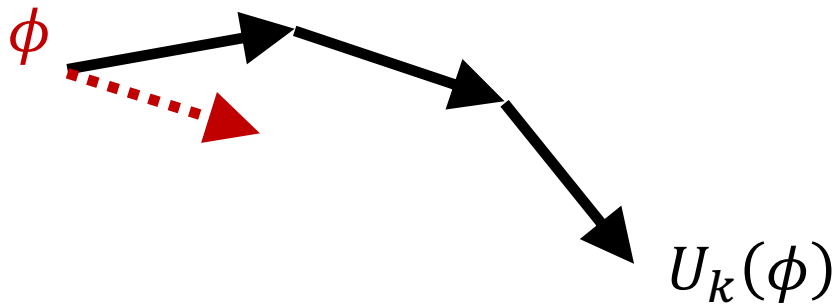
Template function



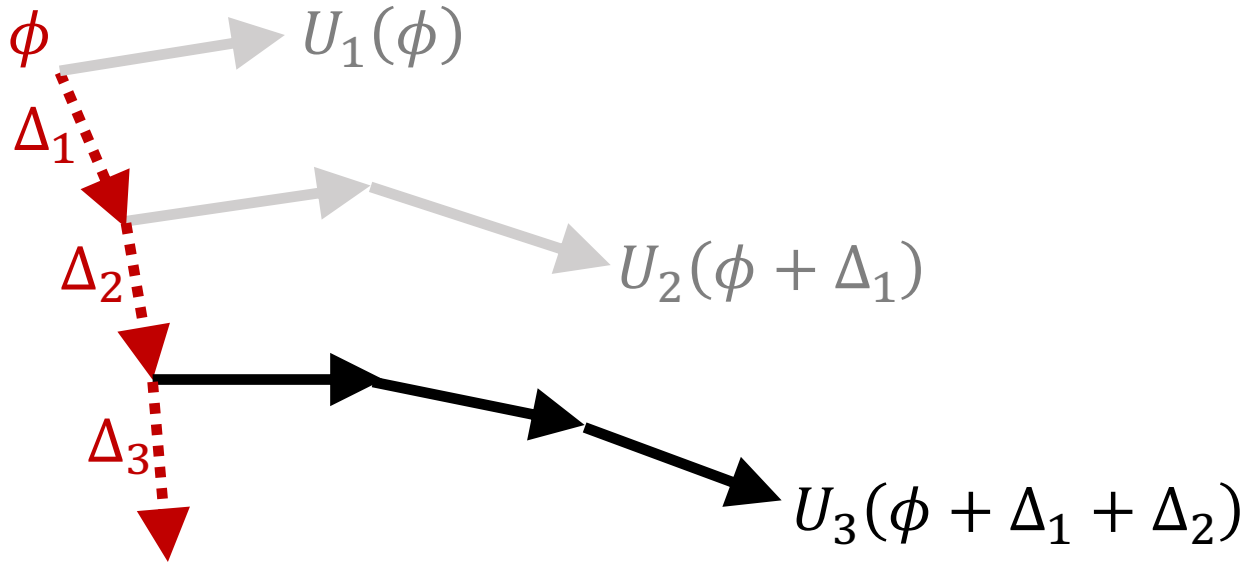
Loss function example

Synthetic Experiments - Baselines

Compare with two baselines.



Reptile[1]



Ours accurate

[1] Nichol et al. 18, On First-Order Meta-Learning Algorithms.

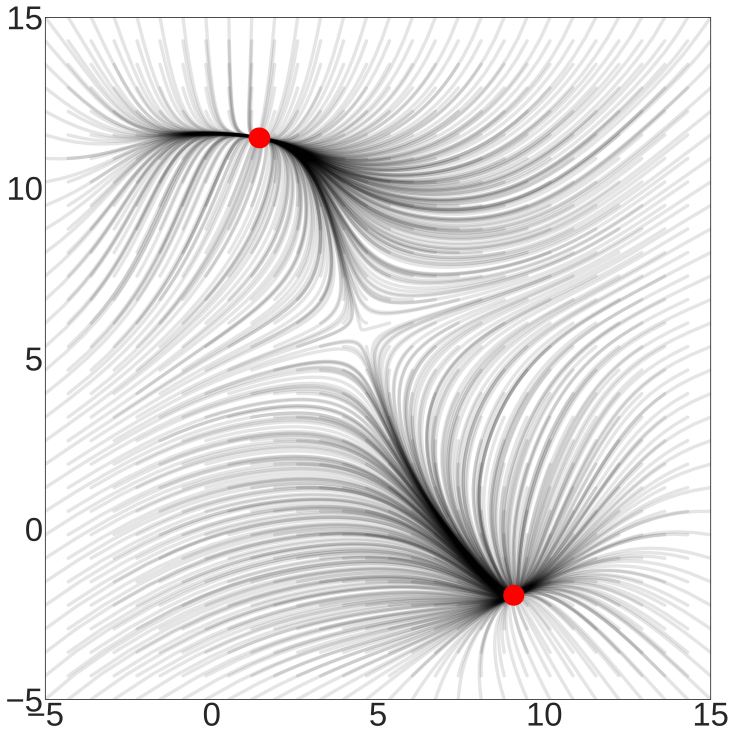
Synthetic Experiments - Baselines

Compare to baselines, our method is the most efficient.

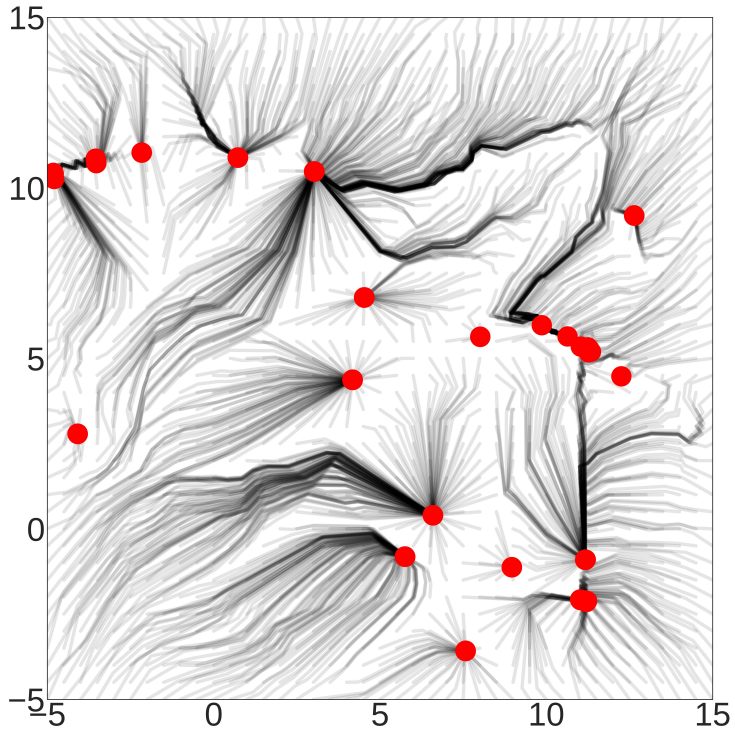
	Reptile	Ours Accurate	Ours
Length of inner-opt. trajectories	$K,..,K$	$1,2,..,K$	$K,..,K$
# Repetitions of inner-optimizations	MK	MK	M
# Total cumulative meta-updates	MK		
# Total cumulative inner-gradient steps	MK^2	$\frac{MK(K+1)}{2}$	MK

Synthetic Experiments - Results

The fewer number of inner steps shows the smoother meta-loss surface.



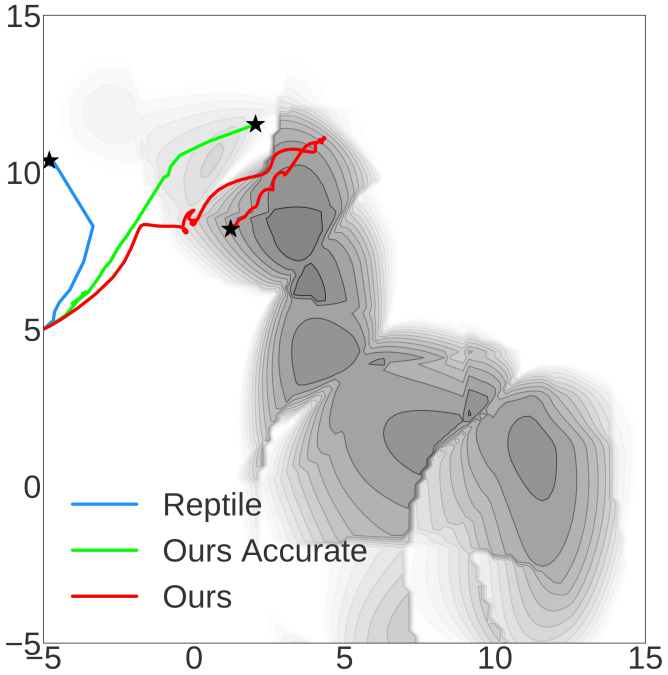
$K = 5$



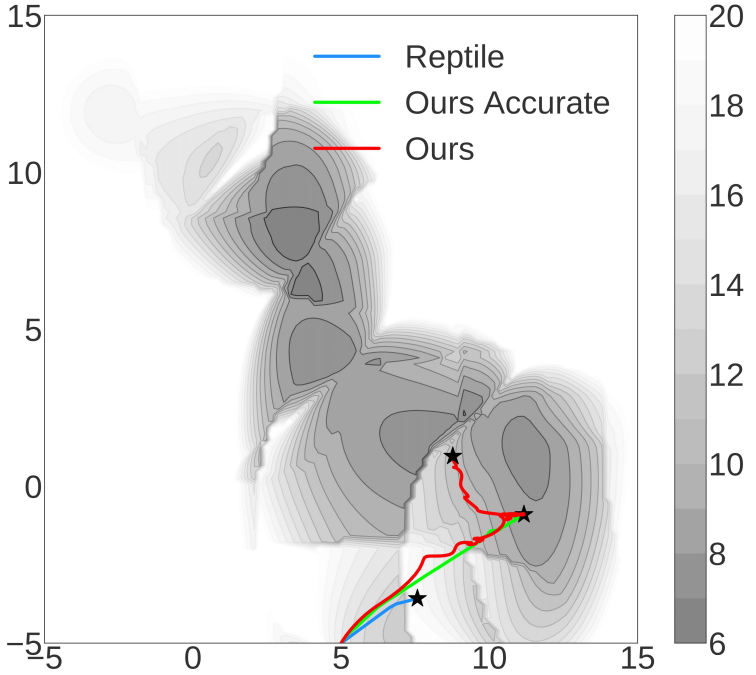
$K = 100$

Synthetic Experiments - Results

In the early stages, our approximation works well. Later, ours converges to good local optimum by meta-curriculum learning effects.



Starts at $(-5,5)$



Starts at $(5,-5)$

Image Classification - Datasets

Meta-train/test with realistic large-scale and heterogeneous datasets.

Meta-train with seven datasets:

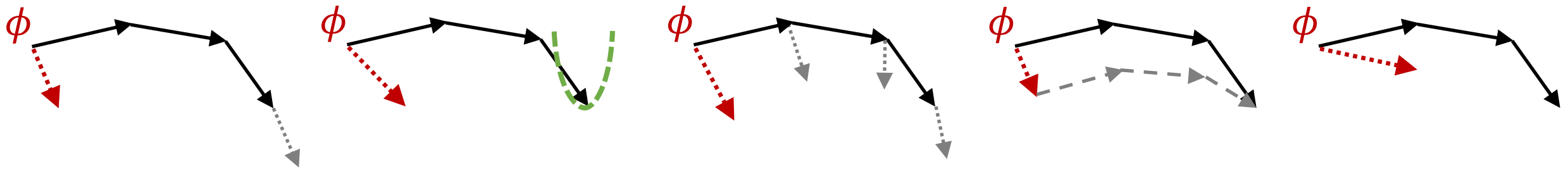
Tiny ImageNet, CIFAR-100, Stanford Dogs, Aircraft, CUB, Fashion-MNIST, and SVHN.

Meta-test with five datasets:

Stanford Cars, QuickDraw, VGG-Flowers, VGG-Pets, and STL10.

Image Classification - Baselines

Compare with two fine-tuning methods (from multi-task learning, Tiny ImageNet) and five first-order meta-learning algorithms.



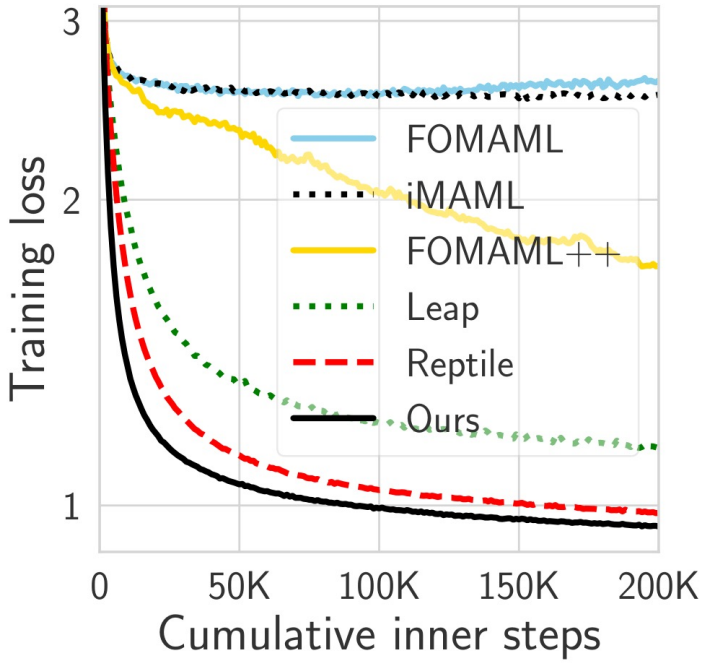
FOMAML[1] iMAML[2] FOMAML++[3] Leap[4] Reptile[5]

[1] Finn et al. 17 Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.
[2] Rajeswaran, et al. 19, Meta-learning with implicit gradients.
[3] Antoniou et al. 19, How to train your MAML.
[4] Flennerhag et al. 18, Transferring Knowledge across Learning Processes.
[5] Nichol et al. 18, On First-Order Meta-Learning Algorithms.

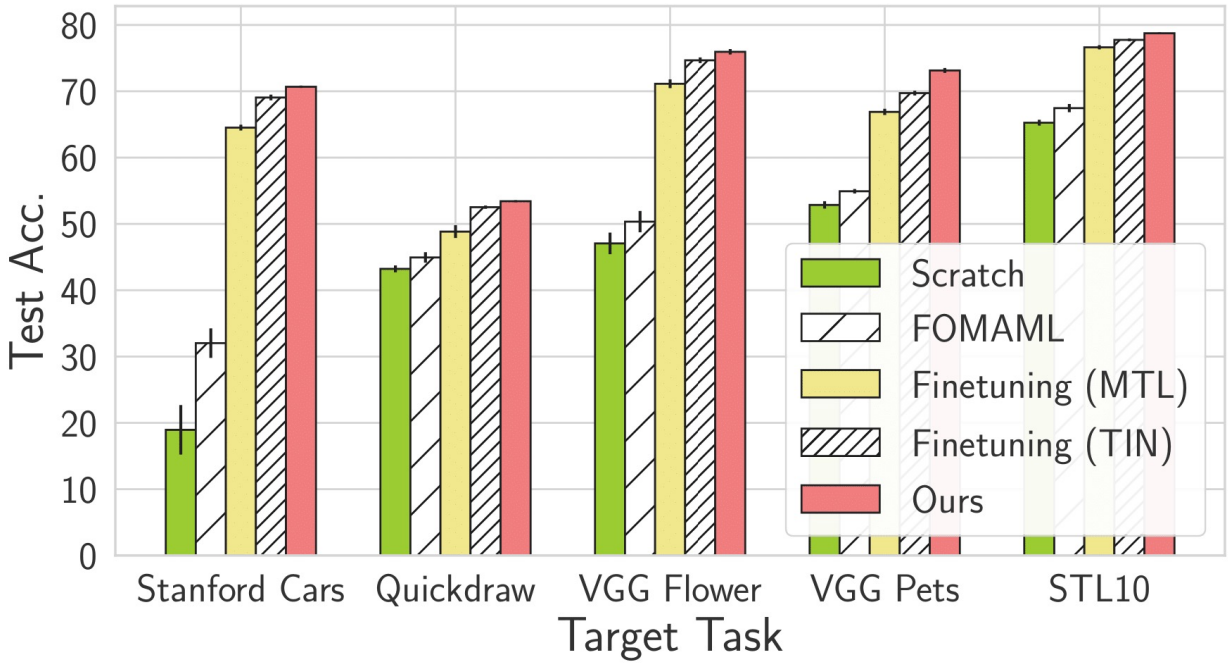
Image Classification - Results

Our method achieves faster meta-convergence in the meta-training stage.

Also, our method outperforms finetuning baselines.



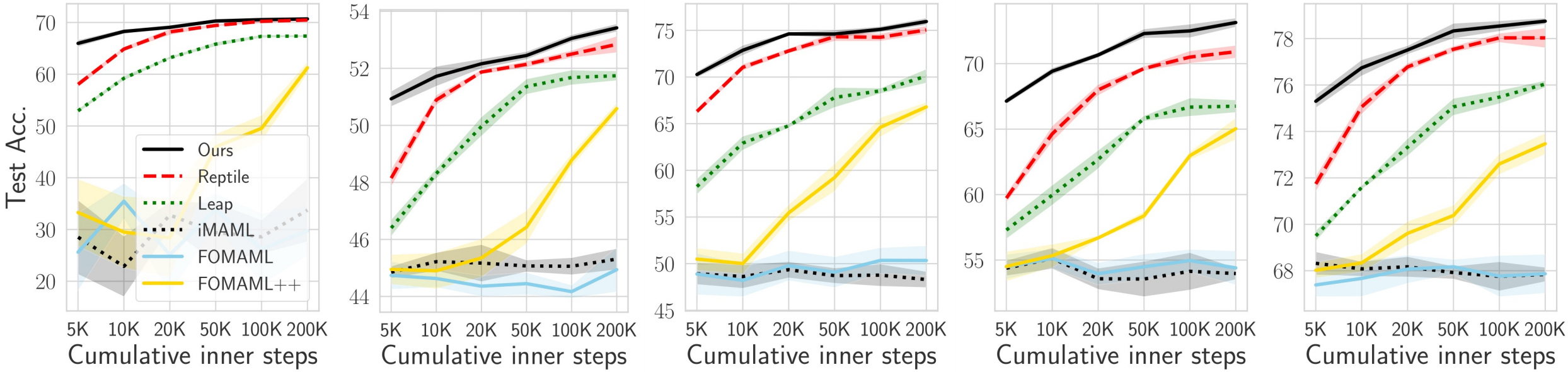
Meta-convergence



Meta-testing performance

Image Classification - Results

Our method outperforms meta-learning baselines, in terms of meta-convergence speed and performance.

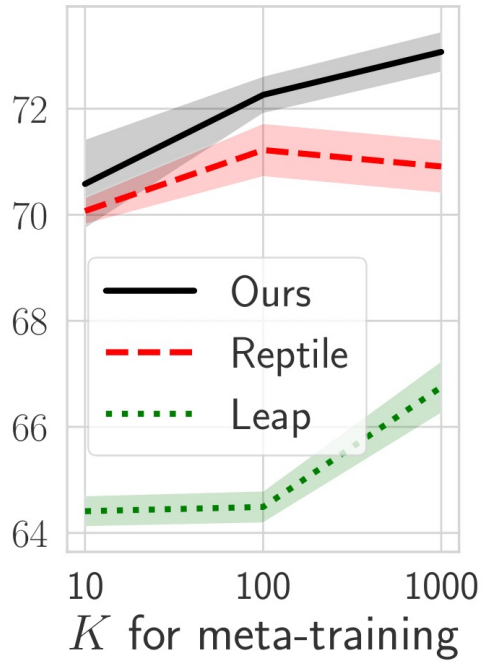


Stanford Cars | Quickdraw | VGG Flowers | VGG Pets | STL10

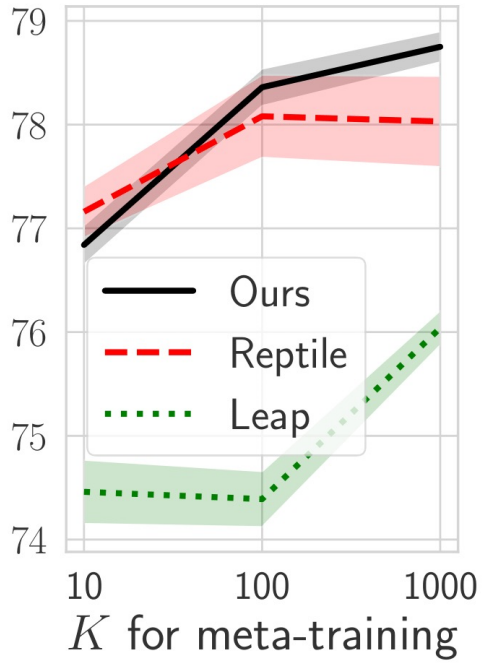
Note) x-axis is cumulative inner steps at meta-training, not training steps at meta-testing.

Image Classification - Results

Longer inner optimization significantly improves performance, especially on our method.



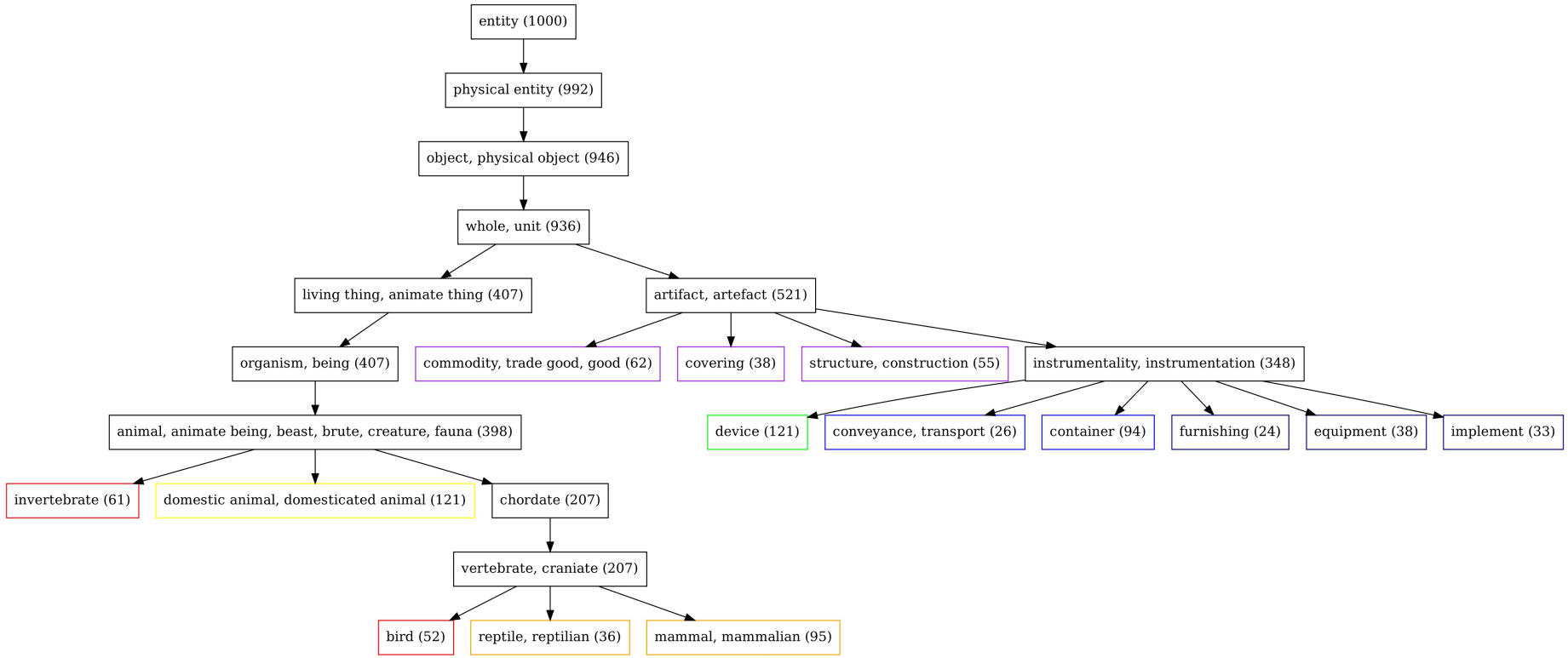
VGG Pets



STL10

Improving on ImageNet Pre-trained Model

Meta-train with a heterogeneous data distribution by class-wisely dividing the ImageNet dataset into 8 subsets based on the WordNet class hierarchy.



Improving on ImageNet Pre-trained Model

Our method outperforms ImageNet finetuning under limited data regime.

CIFAR100 SVHN VGG Pets Food CUB

Conclusion

- We tackled the challenging problem of large-scale meta-learning and showed that a larger number of inner-optimization steps capture the structure of large-scale meta-learning better.

- We improved the meta-learning efficiency with the continual trajectory shifting, which continuously shifts the inner-learning trajectories w.r.t. the frequent update of the initialization point.

- We demonstrated the meta-curriculum learning effect from synthetic experiments and validated that our method outperforms baselines on large-scale image classification in terms of both generalization performance and meta-convergence.

Large-Scale Meta-Learning with Continual Trajectory Shifting

JaeWoong Shin[1*], Hae Beom Lee[1*], Boqing Gong[3], Sung Ju Hwang[12]




https://github.com/JWoong148/ContinualTrajectoryShifting