CAp 2021
Conférence sur l'Apprentissage Automatique

VERS UNE MEILLEURE COMPREHENSION DES MÉTHODES DE MÉTA-APPRENTISSAGE À TRAVERS LA THÉORIE DE L'APPRENTISSAGE DE REPRESENTATIONS MULTI-TÂCHES

Quentin BOUNIOT

INTRODUCTION

INTRODUCTION

Apple

Blueberry

INTRODUCTION

Apple

Blueberry

?

Apple

Blueberry

?

INTRODUCTION


Apple


Blueberry


Da Vinci


Botero


?

INTRODUCTION

Apple

Blueberry

?

Da Vinci

Botero

?

INTRODUCTION

Apple

Blueberry

?

Da Vinci

Botero

?

Meta-learning = Learning to Learn

OUTLINE

- Meta-learning 101
- Multi-task Representation Learning Theory
- From Theory to Practice
- Take Home Message

META-LEARNING 101

- **What is Meta-learning ?**
  - ▶ A meta-learner trained on multiple tasks.
  - ▶ For each task, the meta-learner trains a learner.
  - ▶ The meta-learner is evaluated on new unseen tasks.

META-LEARNING 101

- **What is Meta-learning ?**
  - ▶ A meta-learner trained on multiple tasks.
  - ▶ For each task, the meta-learner trains a learner.
  - ▶ The meta-learner is evaluated on new unseen tasks.

▶ **Meta-Learning can be used for a lot of problems (classification, regression, RL, …)**

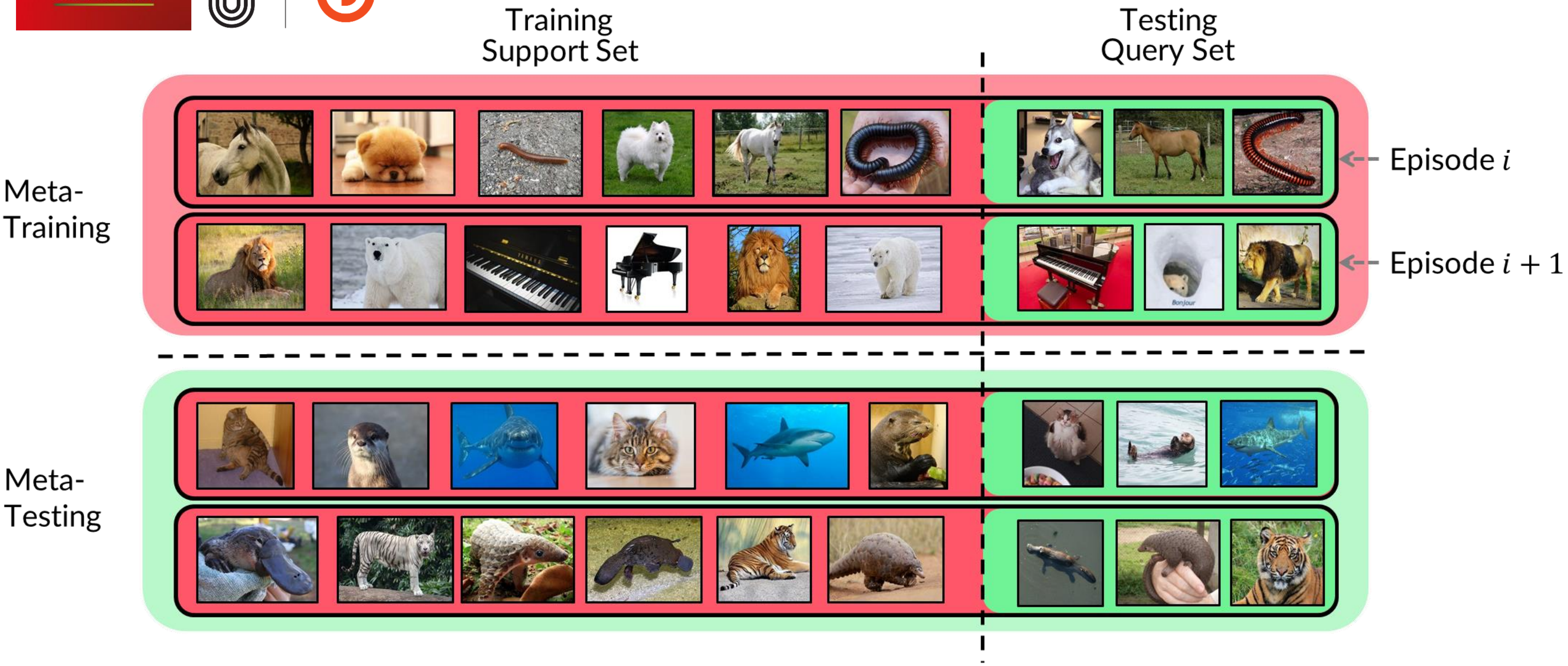
META-LEARNING 101

- **What is Meta-learning ?**
  - A meta-learner trained on multiple tasks.
  - For each task, the meta-learner trains a learner.
  - The meta-learner is evaluated on new unseen tasks.

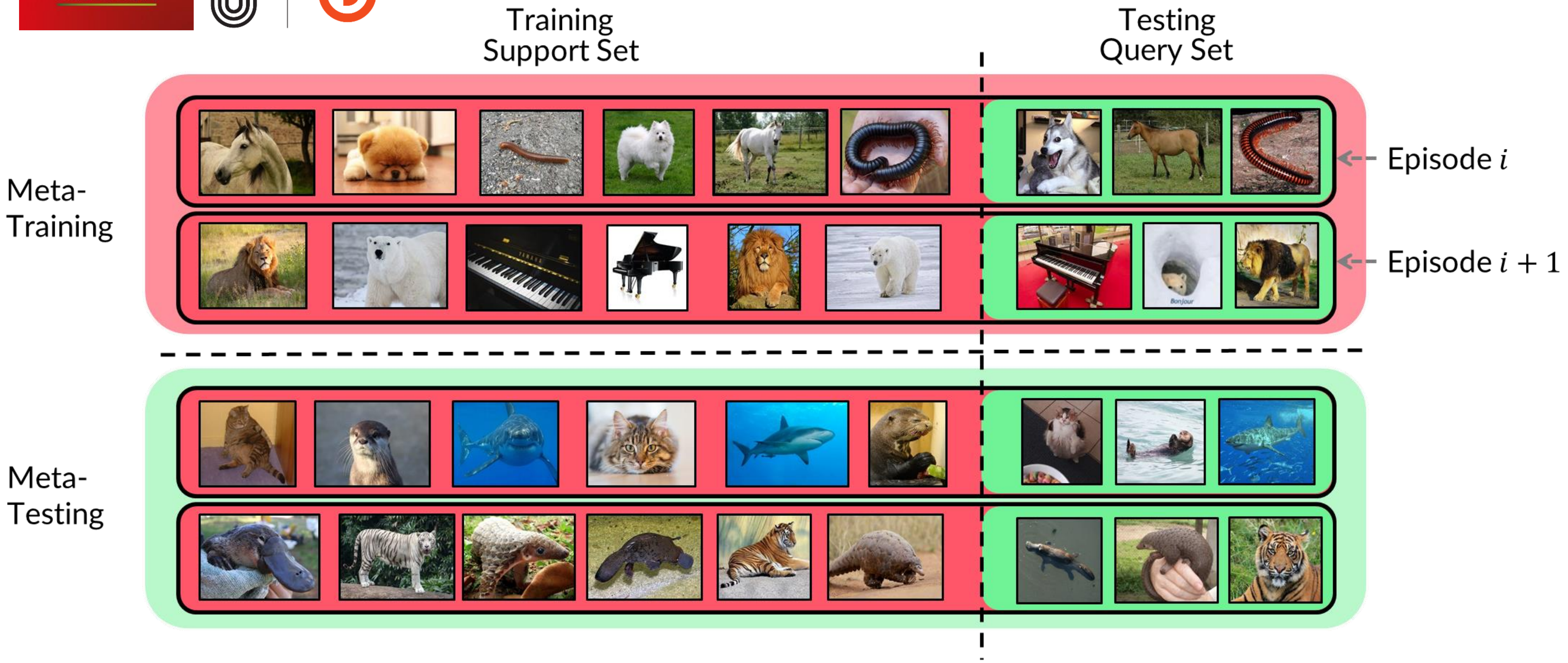
- **Meta-Learning can be used for a lot of problems (classification, regression, RL, …)**

- **How is it related to Few-shot Learning ?**
  - The meta-learner *learns to learn* a new task with few shots.

INTRODUCING EPISODES

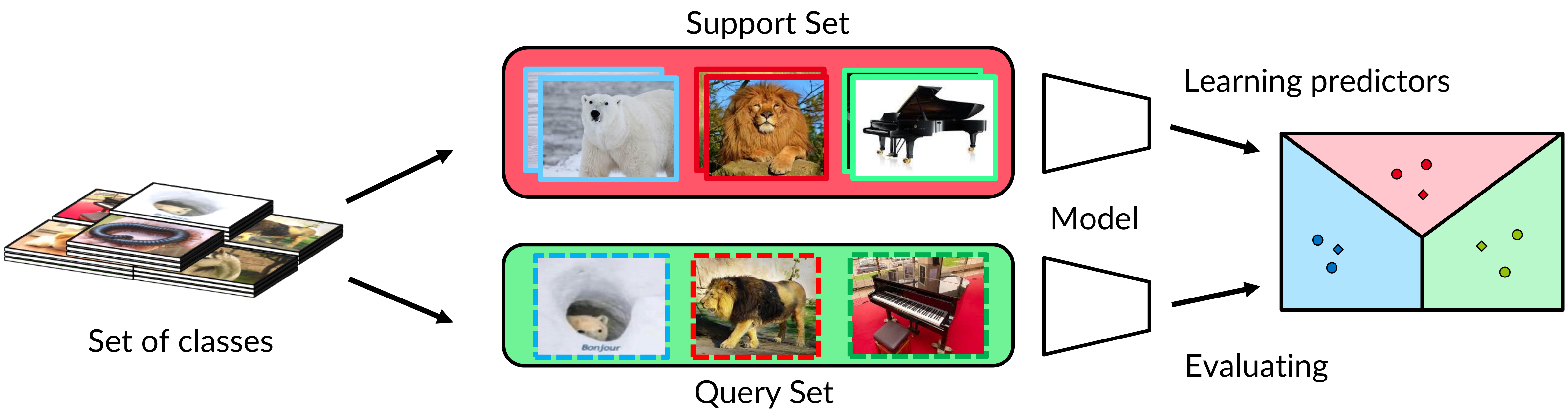


INTRODUCING EPISODES

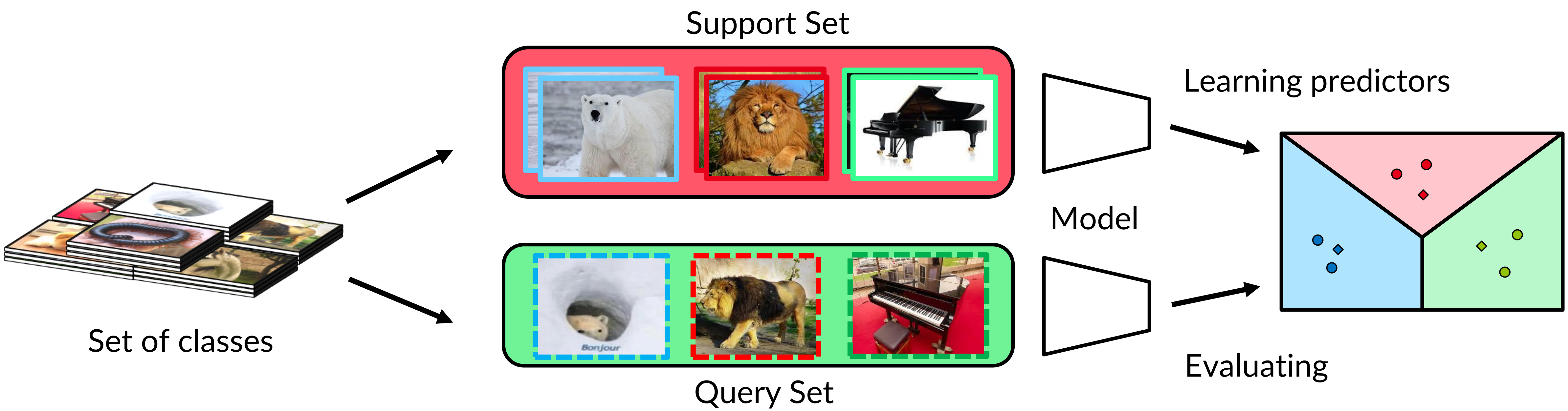


► ***N-way k-shot* episode**: task with N different classes and k images for each class.

EPISODIC TRAINING



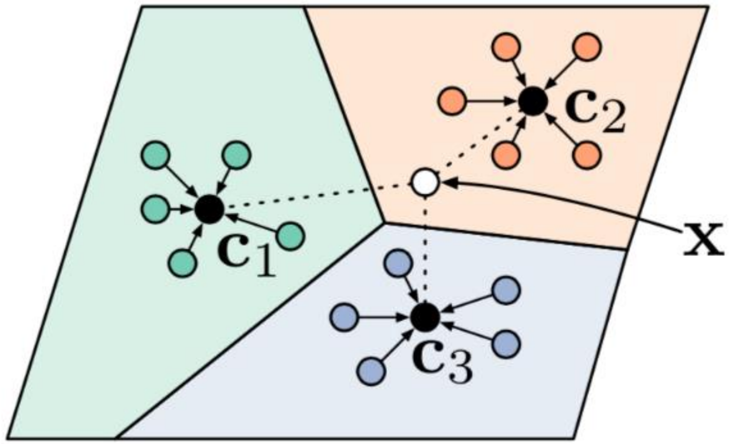
EPISODIC TRAINING



- **Disjoint** sets of classes between **meta-training** and **meta-testing** classes
- Construction of **episodes** from dataset
- **Non-overlapping** class labels between episodes

METHODS I: METRIC-BASED

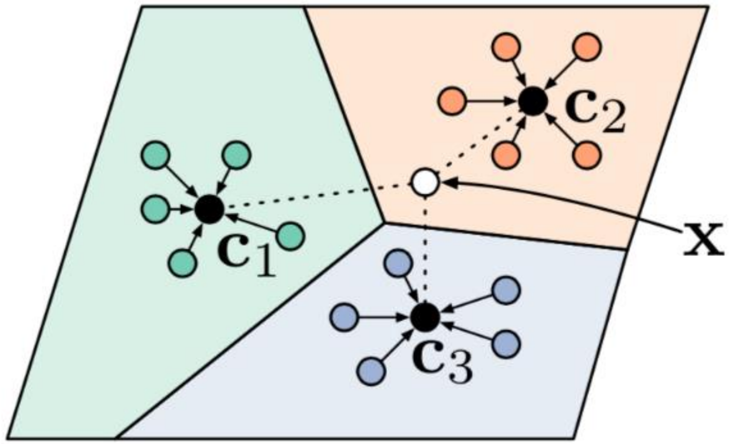
PROTOTYPICAL NETWORK (PROTONET)



Snell J. et al. (2017), *Prototypical Networks for Few-shot Learning.* In NeurIPS 2017.
Allen K. et al. (2019), *Infinite Mixture Prototypes for few-shot learning.* In ICML 2019.

METHODS I: METRIC-BASED

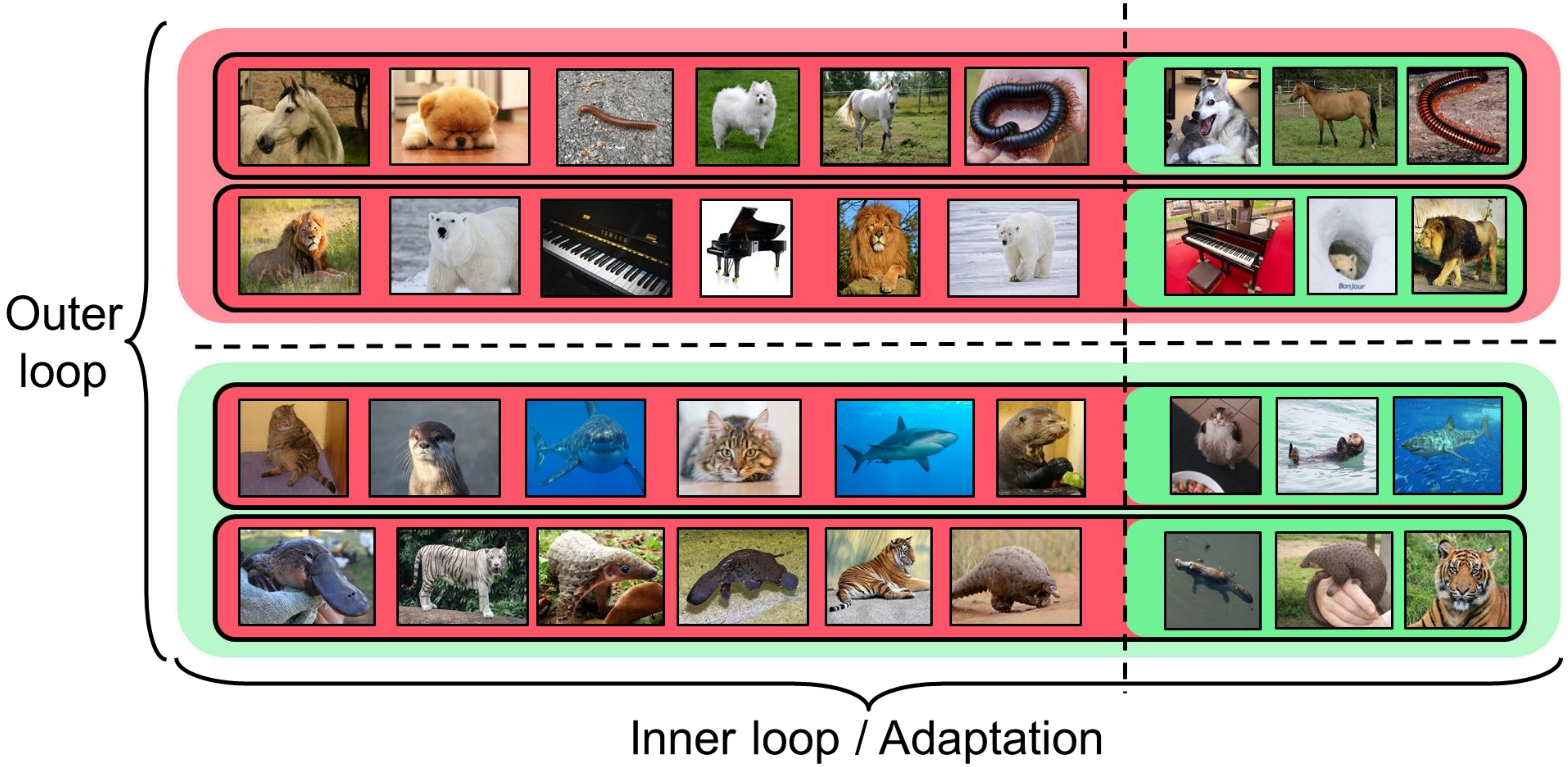
PROTOTYPICAL NETWORK (PROTONET)

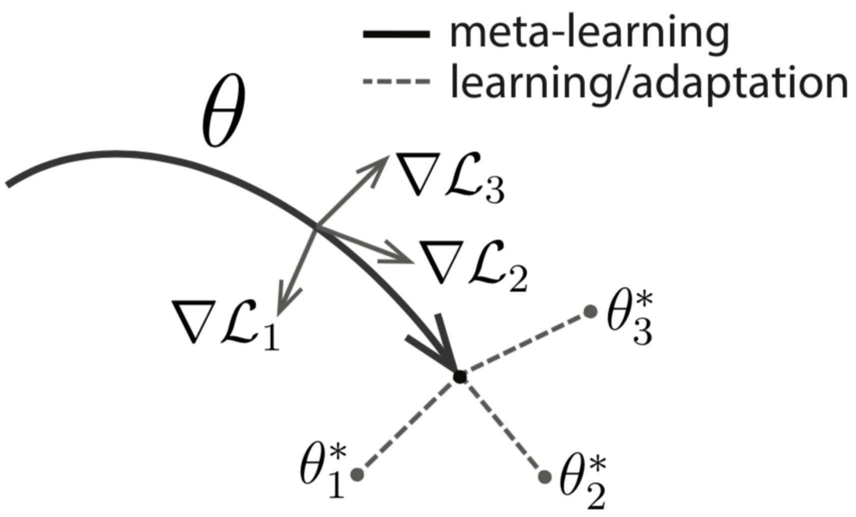


- **Embedding function** to encode query and support samples.
- Support samples fused into **prototypes** c_i for each class
- Probability distribution using **inverse of distances** to prototypes.
- **Contrastive loss** according to distance function.

Snell J. et al. (2017), *Prototypical Networks for Few-shot Learning.* In NeurIPS 2017.
Allen K. et al. (2019), *Infinite Mixture Prototypes for few-shot learning.* In ICML 2019.

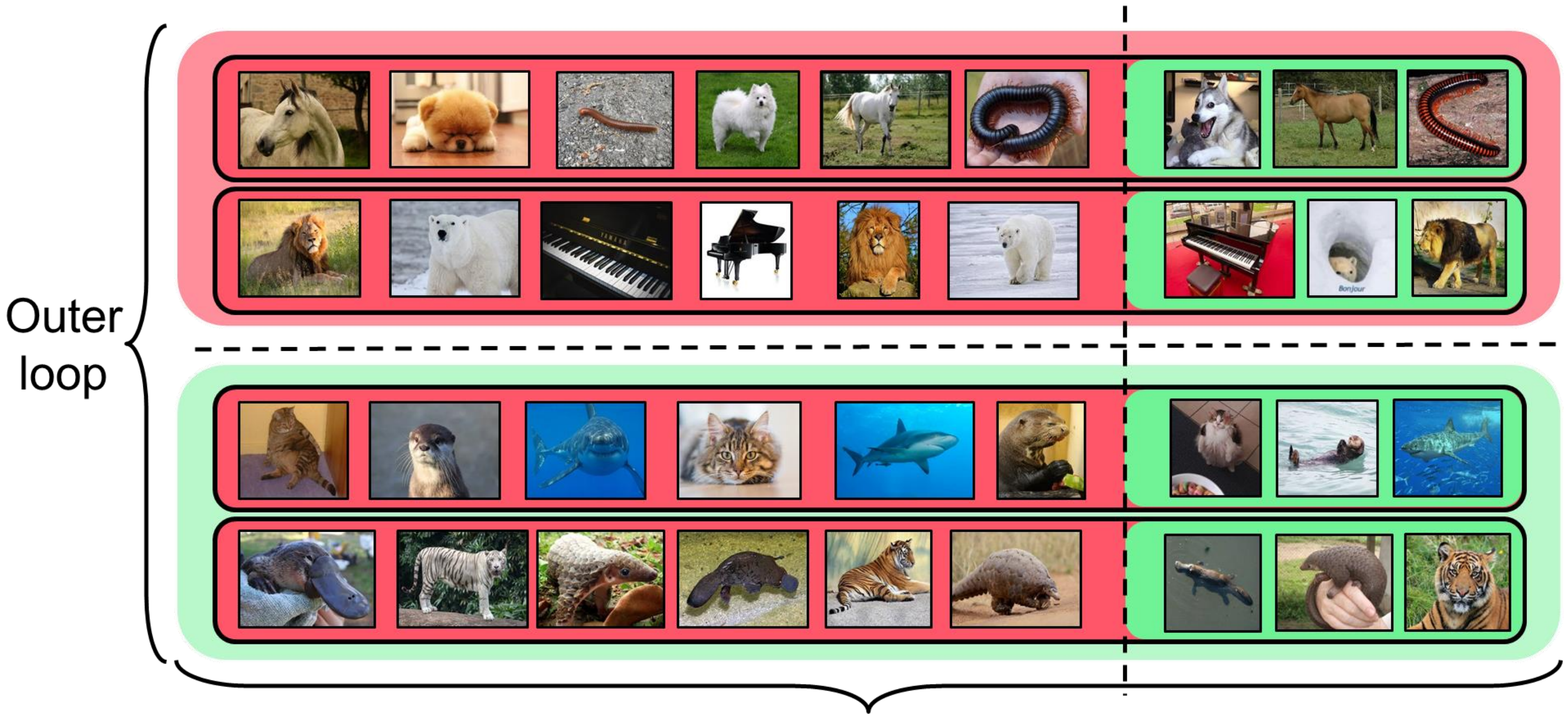
METHODS II: GRADIENT-BASED
MODEL AGNOSTIC META-LEARNING (MAML)

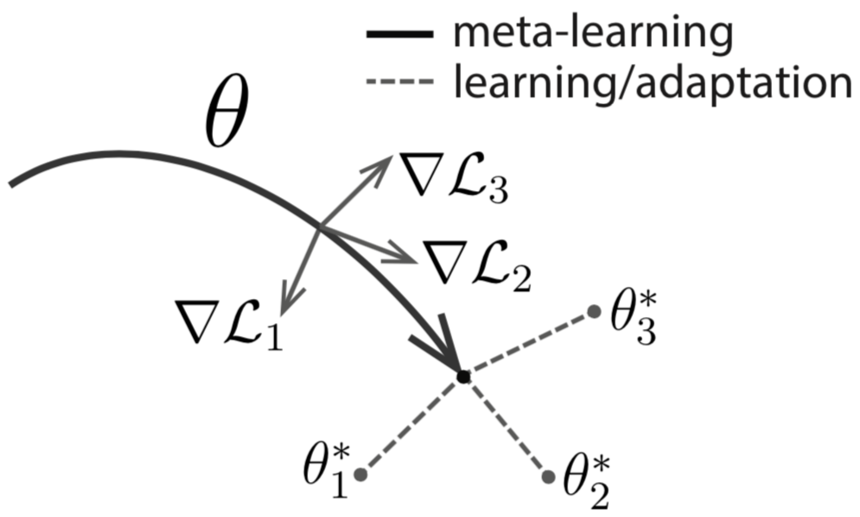




Finn C. et al. (2017), *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.* In ICML 2017
Park E. et Oliva J.B. (2019). *Meta-curvature.* In NeurIPS 2019

MODEL AGNOSTIC META-LEARNING (MAML)





- **Inner Loop:**
  - Performs a few gradient updates over the *k* labelled examples (the support set) of **current episode/task.**
- **Outer Loop:**
  - Updates the **initialization** of the parameters (often called the *meta-initialization*).

Finn C. et al. (2017), *Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks.* In ICML 2017
Park E. et Oliva J.B. (2019). *Meta-curvature.* In NeurIPS 2019

MULTI-TASK REPRESENTATION LEARNING THEORY

Quentin Bouniot, Ievgen Redko, Romaric Audigier, Angélique Loesch | CAp2021 | 14/06/2021

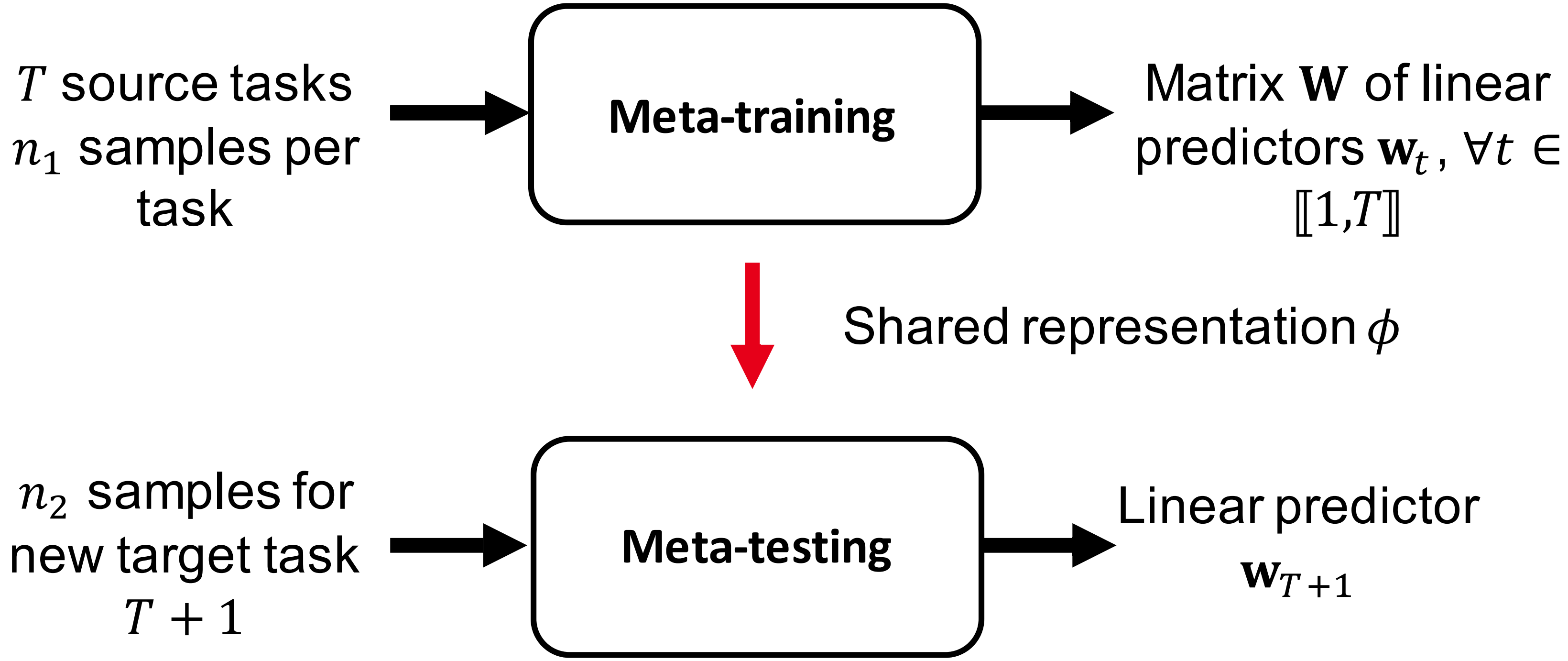
PRELIMINARY KNOWLEDGE

T source tasks
n_1 samples per task

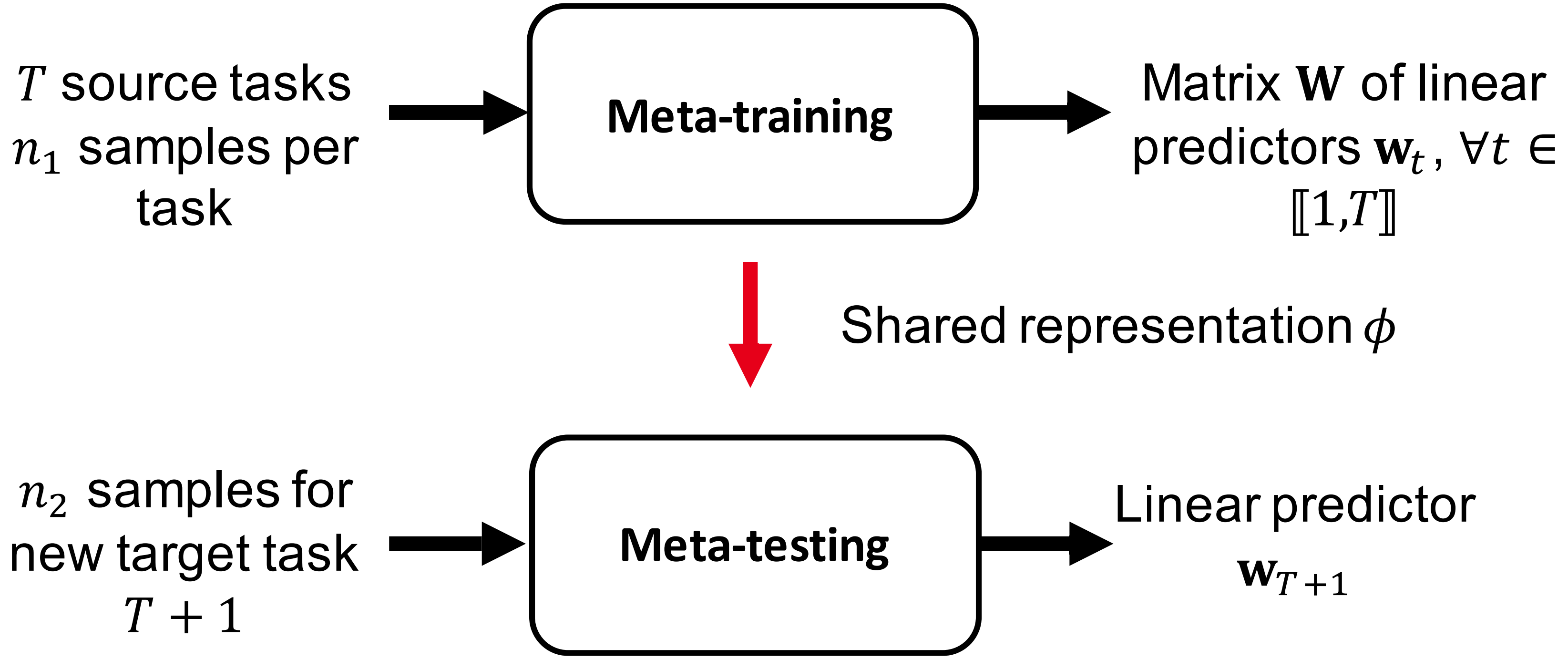
→ **Meta-training** →

Matrix $\mathbf{W}$ of linear predictors $\mathbf{w}_t$, $\forall t \in [\![1,T]\!]$

PRELIMINARY KNOWLEDGE



PRELIMINARY KNOWLEDGE



Goal: Minimize *excess risk* $\mathrm{ER} = \mathcal{L}(\hat{\phi}, \hat{\mathrm{w}}_{T+1}) - \mathcal{L}(\phi^*, \mathrm{w}_{T+1}^*)$

- True risk $\mathcal{L}$
- Optimal weights ϕ^*
- w_{T+1}^* ideal target linear predictor

IMPORTANT ASSUMPTIONS

Du S. et al. (2020), *Few-Shot Learning via Learning the Representation, Provably*. In ICRL 2021
Tripuraneni N. et al. (2020).Provable Meta-Learning of Linear Representations. In arXiv 2020.

IMPORTANT ASSUMPTIONS

- Assumption 1: **Diversity of the source tasks**

  - Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_T^*]$ cover all the directions evenly

  - Condition Number $\kappa(\mathbf{W}^*) = \frac{\sigma_{max}(\mathbf{W}^*)}{\sigma_{min}(\mathbf{W}^*)}$ should not increase with T

Du S. et al. (2020), *Few-Shot Learning via Learning the Representation, Provably*. In ICRL 2021
Tripuraneni N. et al. (2020).Provable Meta-Learning of Linear Representations. In arXiv 2020.

IMPORTANT ASSUMPTIONS

- Assumption 1: **Diversity of the source tasks**
  - ▶ Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \dots, \mathbf{w}_T^*]$ cover all the directions evenly
  - ▶ Condition Number $\kappa(\mathbf{W}^*) = \frac{\sigma_{max}(\mathbf{W}^*)}{\sigma_{min}(\mathbf{W}^*)}$ should not increase with T
- Assumption 2: **Constant classification margin**
  - ▶ Norm of the predictors $\|\mathbf{w}_t^*\|_{t \in [1,T]}$ should not increase with T

Du S. et al. (2020), *Few-Shot Learning via Learning the Representation, Provably*. In ICRL 2021
Tripuraneni N. et al. (2020).Provable Meta-Learning of Linear Representations. In arXiv 2020.

IMPORTANT ASSUMPTIONS

- Assumption 1: **Diversity of the source tasks**
  - ▶ Optimal predictors $\mathbf{W}^* = [\mathbf{w}_1^*, \dots, \mathbf{w}_T^*]$ cover all the directions evenly

▶ Condition Number $\kappa(\mathbf{W}^*) = \frac{\sigma_{max}(\mathbf{W}^*)}{\sigma_{min}(\mathbf{W}^*)}$ should not increase with T

- Assumption 2: **Constant classification margin**

▶ Norm of the predictors $\|\mathbf{w}_t^*\|_{t\in[1,T]}$ should not increase with T

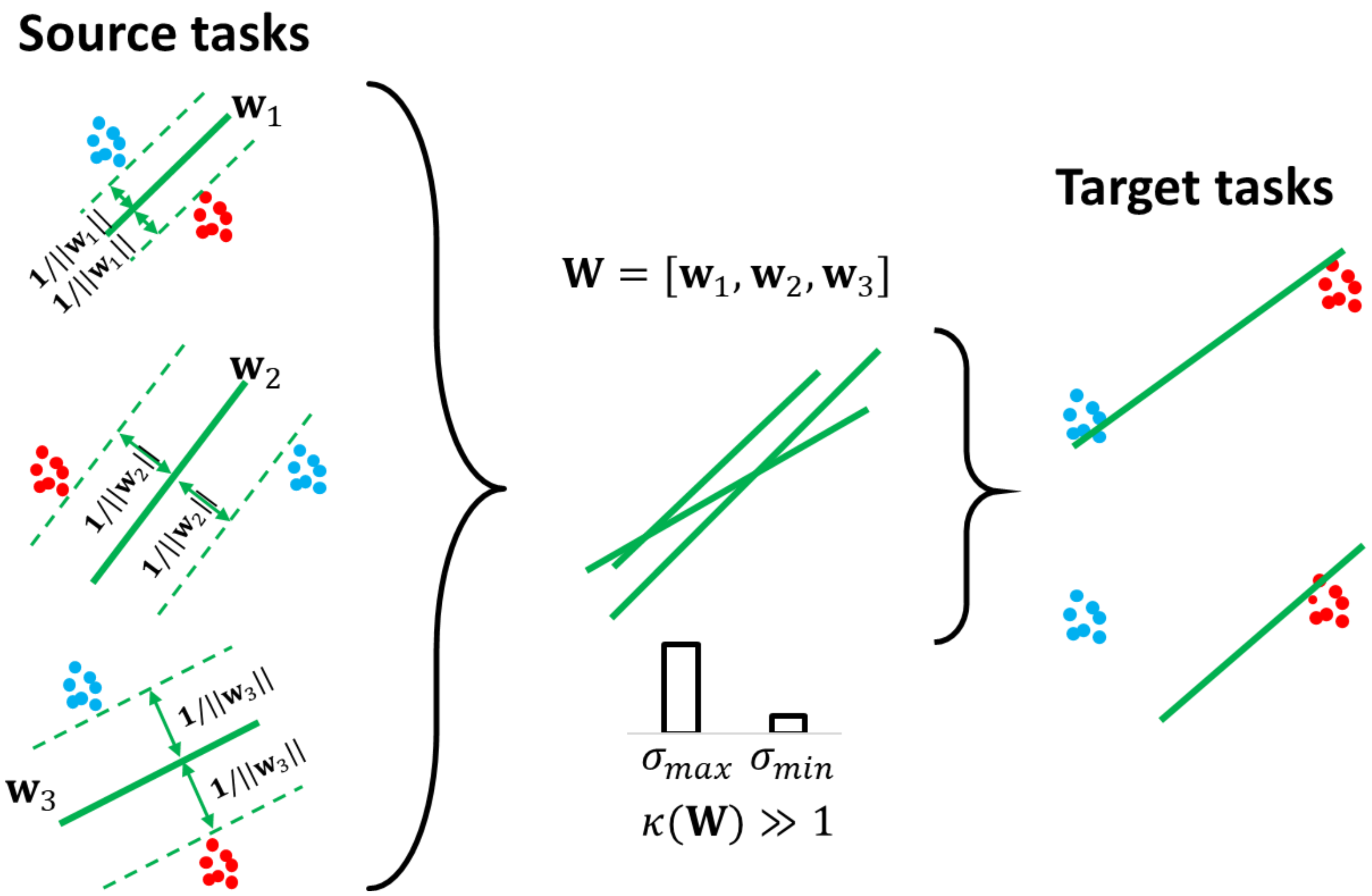
If satisfied, $\mathrm{ER}(\phi, \mathbf{w}_{T+1}) \leq O(\frac{1}{n_1 T} + \frac{1}{n_2})$

✓ All source and target data are useful to decrease the bound of *excess risk*

Du S. et al. (2020), *Few-Shot Learning via Learning the Representation, Provably*. In ICRL 2021
Tripuraneni N. et al. (2020).Provable Meta-Learning of Linear Representations. In arXiv 2020.

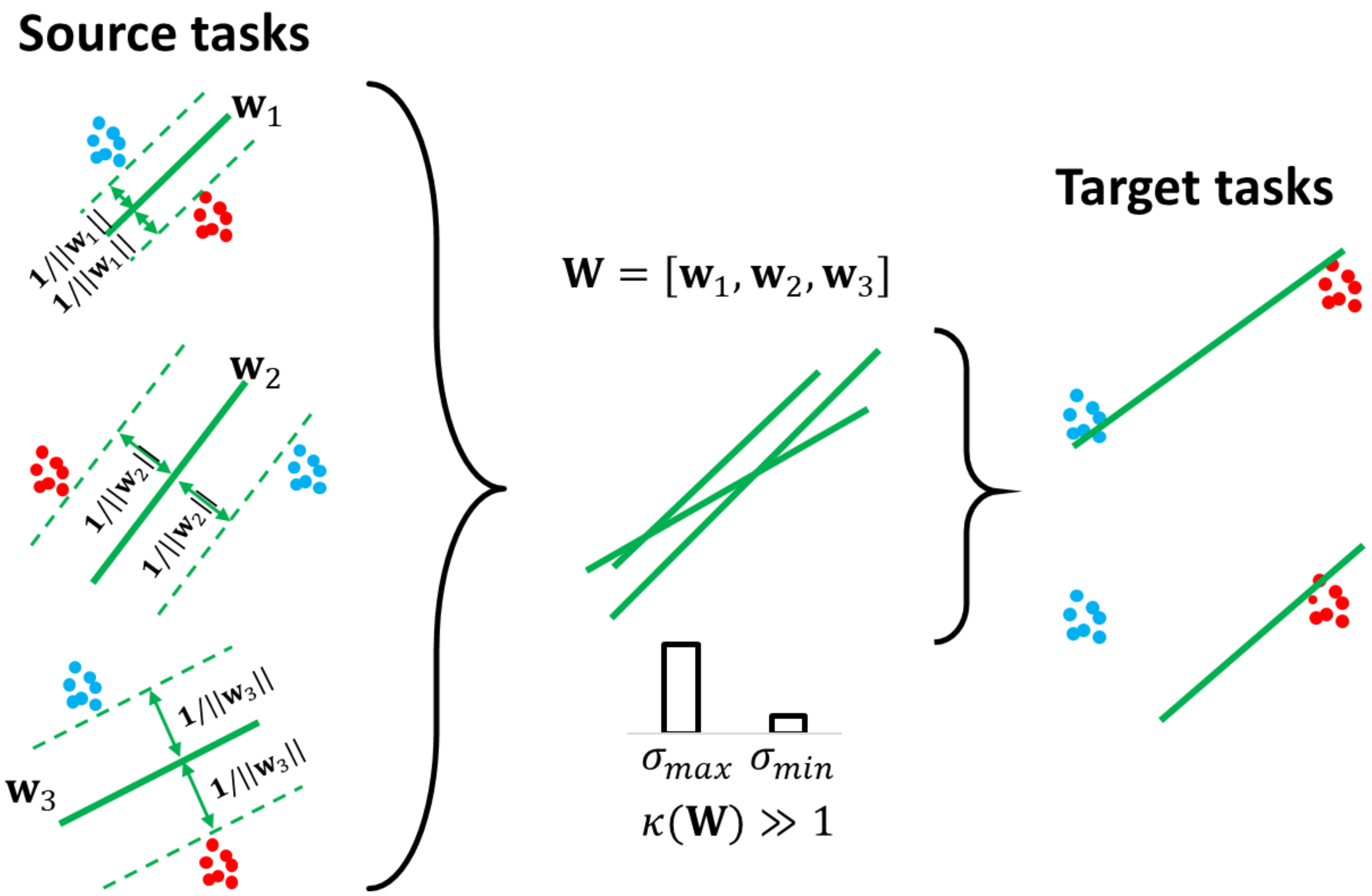
ILLUSTRATION

WHEN ASSUMPTIONS ARE NOT SATISFIED



ILLUSTRATION

WHEN ASSUMPTIONS ARE NOT SATISFIED



× Linear predictors cover **only part** of the space or **over-specialize** to the tasks

ILLUSTRATION

SATISFIED ASSUMPTIONS

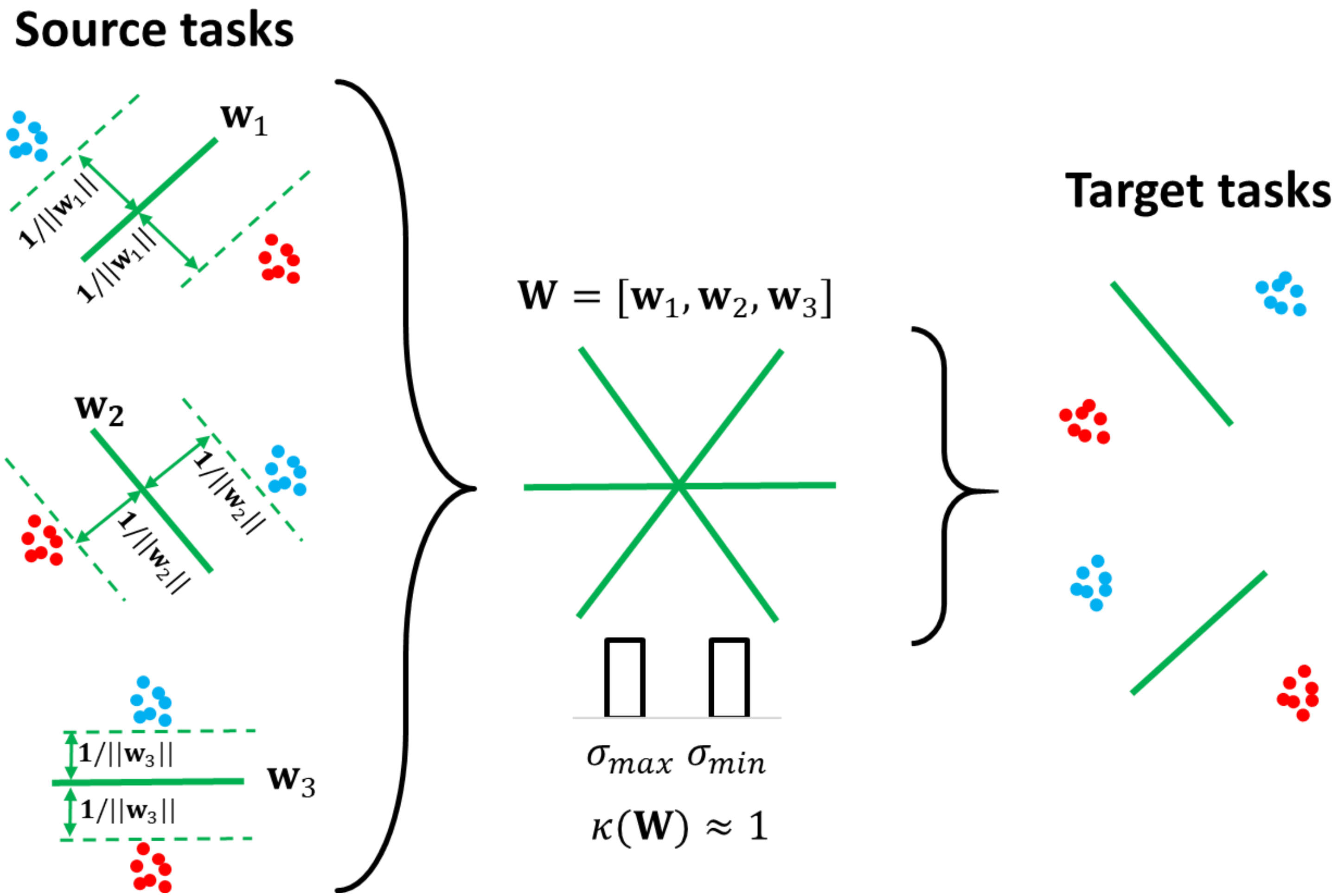
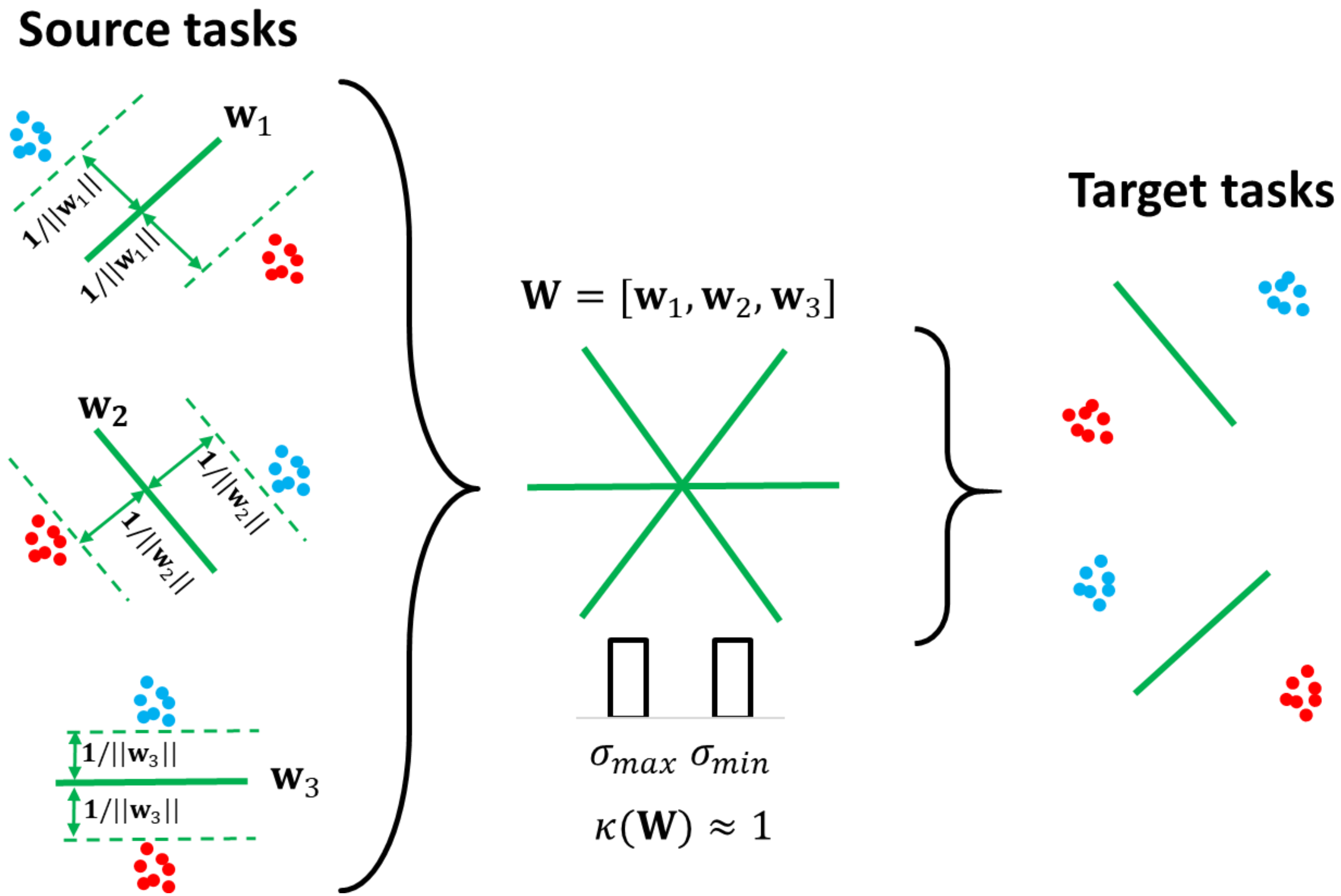


ILLUSTRATION
SATISFIED ASSUMPTIONS



- ✓ Satisfying assumption 1 makes sure that linear predictors are **complementary**
- ✓ Satisfying assumption 2 avoids **under- or over-specialization** to the tasks

FROM THEORY TO PRACTICE

CONTRIBUTIONS

CAN WE FORCE THE ASSUMPTIONS ?

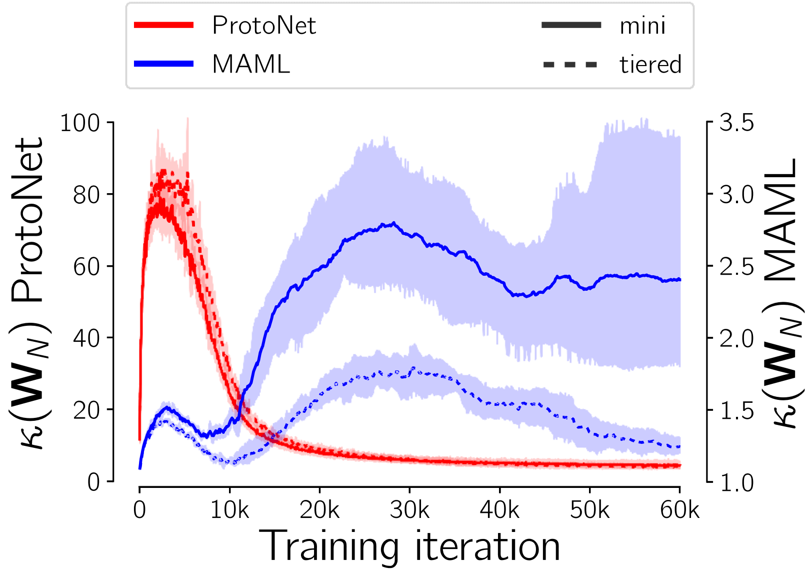
Given $\mathbf{W}^*$ *such that* $\kappa(\mathbf{W}^*) \gg 1$, *can we learn* $\widehat{\mathbf{W}}$ *with* $\kappa\left(\widehat{\mathbf{W}}\right) \approx 1$ *while solving the underlying classification problems equally well ?*

CAN WE FORCE THE ASSUMPTIONS ?

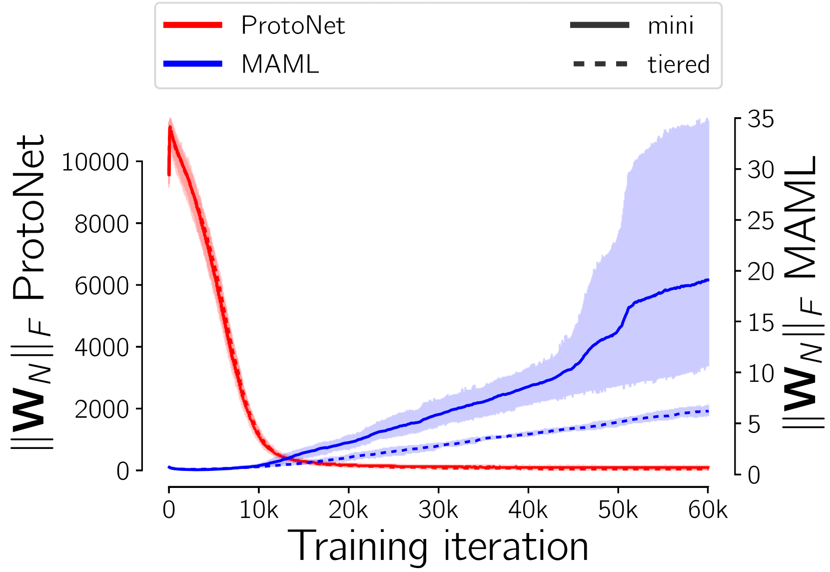
Given $\mathbf{W}^*$ *such that* $\kappa(\mathbf{W}^*) \gg 1$, *can we learn* $\widehat{\mathbf{W}}$ *with* $\kappa(\widehat{\mathbf{W}}) \approx 1$
while solving the underlying classification problems equally well ?

- ✓ Even when $\mathbf{W}^*$ does not satisfy the assumptions, it is **possible to learn** $\hat{\phi}$ to respect them

WHAT HAPPENS IN PRACTICE ?

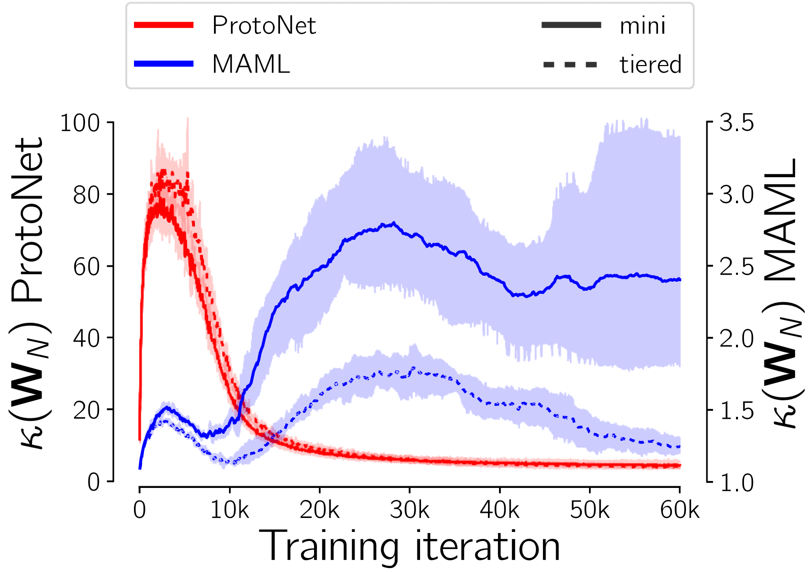


Monitoring the Condition Number

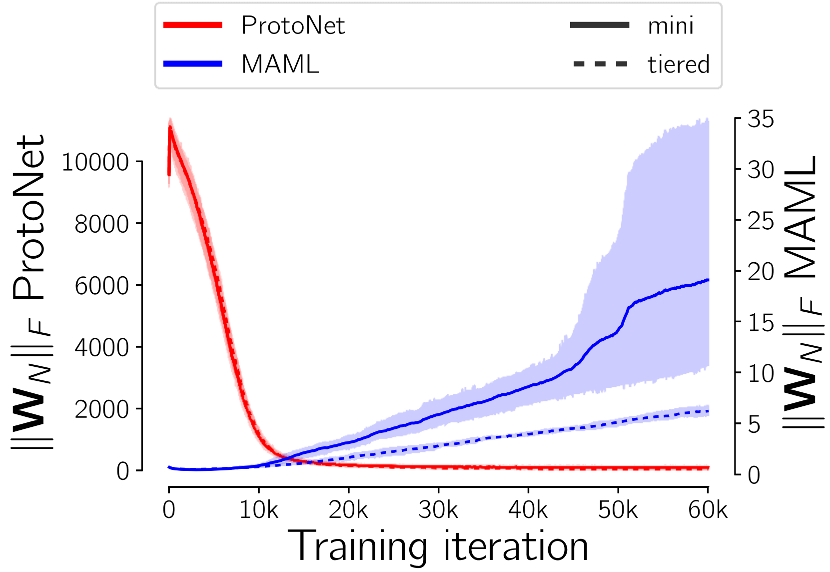


Monitoring the Norm

WHAT HAPPENS IN PRACTICE ?



Monitoring the Condition Number



Monitoring the Norm

- ▶ $\mathbf{W}_N$ restriction to N last predictors
- ✓ ProtoNet **naturally verifies** the assumptions
- × MAML **does not** verify the assumptions

WHY DOES IT HAPPEN ?

WHY DOES IT HAPPEN ?

- **Theorem** (Normalized ProtoNet):

$$\text{if } \forall i \ \|prototype_i\| = 1, \text{then } \exists \phi \in \arg\min loss \ \text{ such that } \kappa(\mathbf{W}^*) = 1$$

✓ Norm minimization is **enough** to obtain well-behaved condition number for ProtoNet.

WHY DOES IT HAPPEN ?

- **Theorem** (Normalized ProtoNet):

$$\text{if } \forall i \ \|prototype_i\| = 1, \text{then } \exists \phi \in \arg\min loss \ \text{ such that } \kappa(\mathbf{W}^*) = 1$$

✓ Norm minimization is **enough** to obtain well-behaved condition number for ProtoNet.

- **Proposition** (Condition Number for MAML):

$$\text{At iteration } i, \text{if } \sigma_{min} = 0 \text{ for last two tasks,} \\ \text{then } \kappa\left(\widehat{W}_2^{i+1}\right) \geq \kappa(\widehat{W}_2^{i})$$

× The condition number for MAML can **increase** between iterations.

FROM THEORY TO PRACTICE

- **Ensuring Assumption 1:** **Spectral** or **entropic** regularization

FROM THEORY TO PRACTICE

- **Ensuring Assumption 1:** **Spectral** or **entropic** regularization

$$\kappa(\mathbf{W}_N) = \frac{\sigma_{max}(\mathbf{W}_N)}{\sigma_{min}(\mathbf{W}_N)} \quad \text{or} \quad H_\sigma(\mathbf{W}_N) = \sum_{i=1}^{N} \text{softmax}\big(\sigma(\mathbf{W_N})\big)_i \cdot \log \text{softmax}\big(\sigma(\mathbf{W_N})\big)_i$$

FROM THEORY TO PRACTICE

- **Ensuring Assumption 1:** **Spectral** or **entropic** regularization

$$\kappa(\mathbf{W}_N) = \frac{\sigma_{max}(\mathbf{W}_N)}{\sigma_{min}(\mathbf{W}_N)} \quad \text{or} \quad H_\sigma(\mathbf{W}_N) = \sum_{i=1}^{N} \text{softmax}\big(\sigma(\mathbf{W_N})\big)_i \cdot \log \text{softmax}\big(\sigma(\mathbf{W_N})\big)_i$$

✓ Regularizing with $\kappa(\mathbf{W_N})$ *or* $H_\sigma(\mathbf{W}_N)$ leads to a **better coverage** of the searched space

FROM THEORY TO PRACTICE

- **Ensuring Assumption 1:** **Spectral** or **entropic** regularization

$$\kappa(\mathbf{W}_N) = \frac{\sigma_{max}(\mathbf{W}_N)}{\sigma_{min}(\mathbf{W}_N)} \quad \text{or} \quad H_\sigma(\mathbf{W}_N) = \sum_{i=1}^{N} \mathrm{softmax}\big(\sigma(\mathbf{W_N})\big)_i \cdot \log \mathrm{softmax}\big(\sigma(\mathbf{W_N})\big)_i$$

✓ Regularizing with $\kappa(\mathbf{W_N})$ *or* $H_\sigma(\mathbf{W}_N)$ leads to a **better coverage** of the searched space

- **Ensuring Assumption 2:** **Norm regularization or normalization** for linear predictors

FROM THEORY TO PRACTICE

- **Ensuring Assumption 1:** **Spectral** or **entropic** regularization

$$\kappa(\mathbf{W}_N) = \frac{\sigma_{max}(\mathbf{W}_N)}{\sigma_{min}(\mathbf{W}_N)} \quad \text{or} \quad H_\sigma(\mathbf{W}_N) = \sum_{i=1}^{N} \mathrm{softmax}\big(\sigma(\mathbf{W_N})\big)_i \cdot \log \mathrm{softmax}\big(\sigma(\mathbf{W_N})\big)_i$$

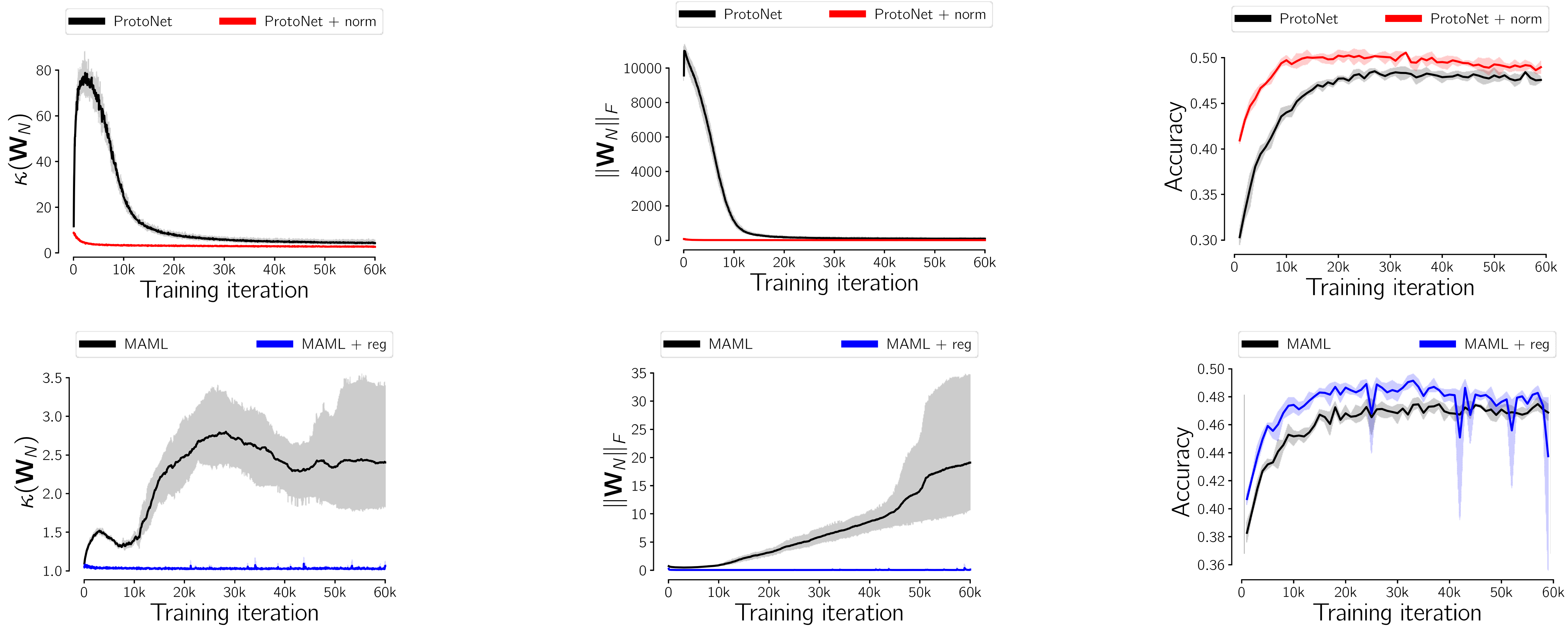
  ✓ Regularizing with $\kappa(\mathbf{W_N})$ *or* $H_\sigma(\mathbf{W}_N)$ leads to a **better coverage** of the searched space

- **Ensuring Assumption 2:** **Norm regularization or normalization** for linear predictors

  ✓ Normalizing predictors ensures **constant margin** that does not change with T

EXPERIMENTAL RESULTS

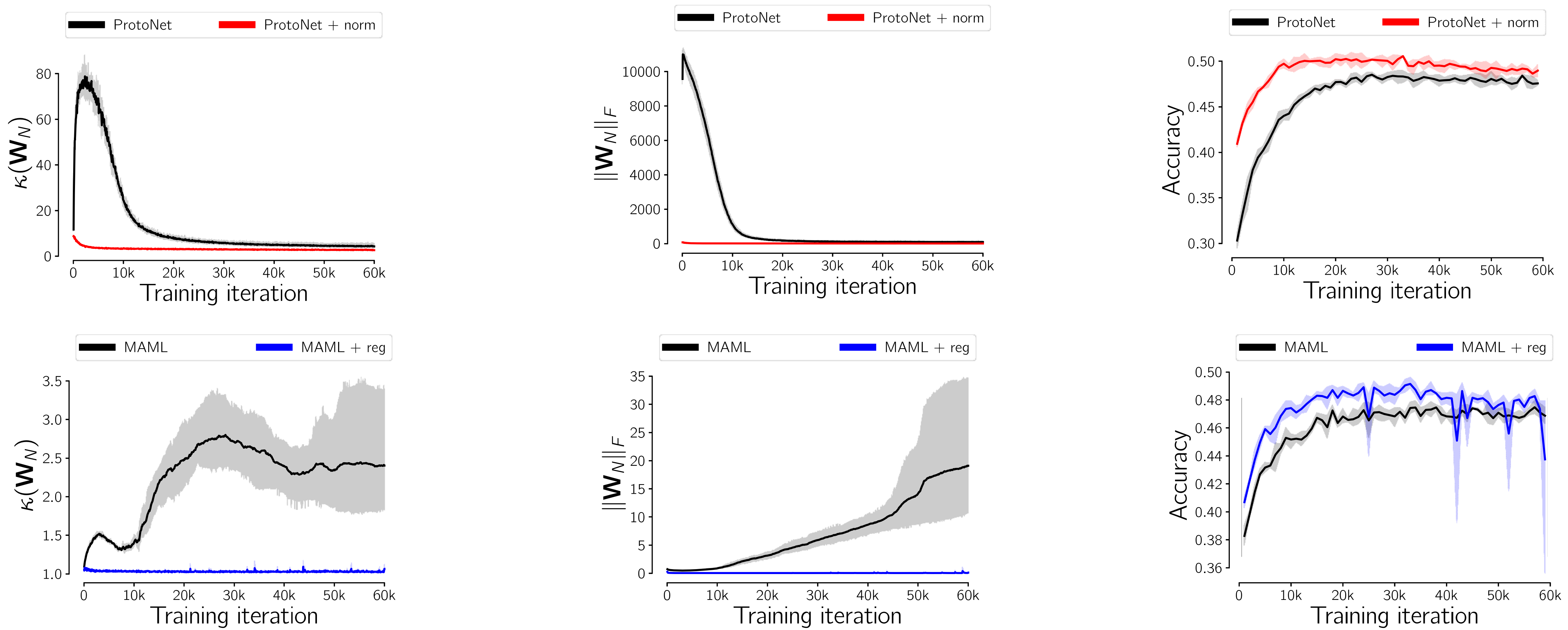
MONITORING THE CONDITION NUMBER AND THE NORM



Experiments on mini-ImageNet 5-way 1-shot

EXPERIMENTAL RESULTS

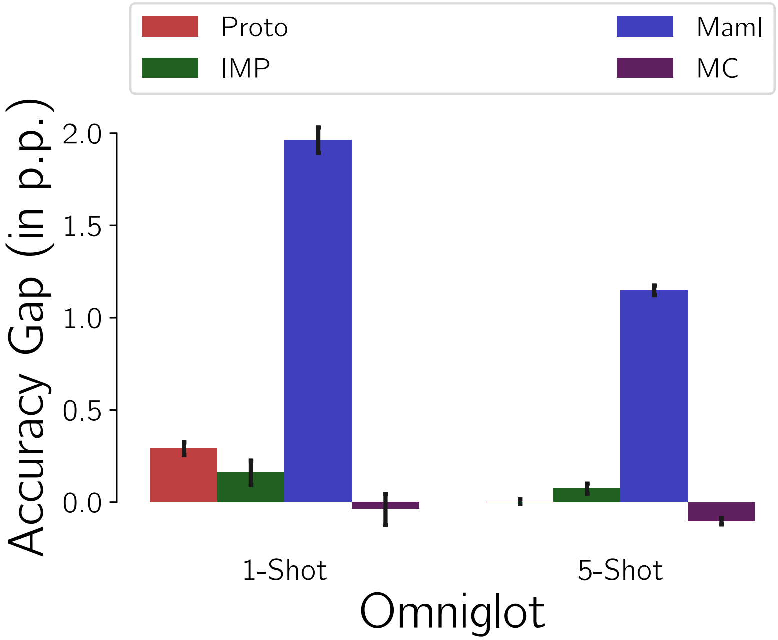
MONITORING THE CONDITION NUMBER AND THE NORM



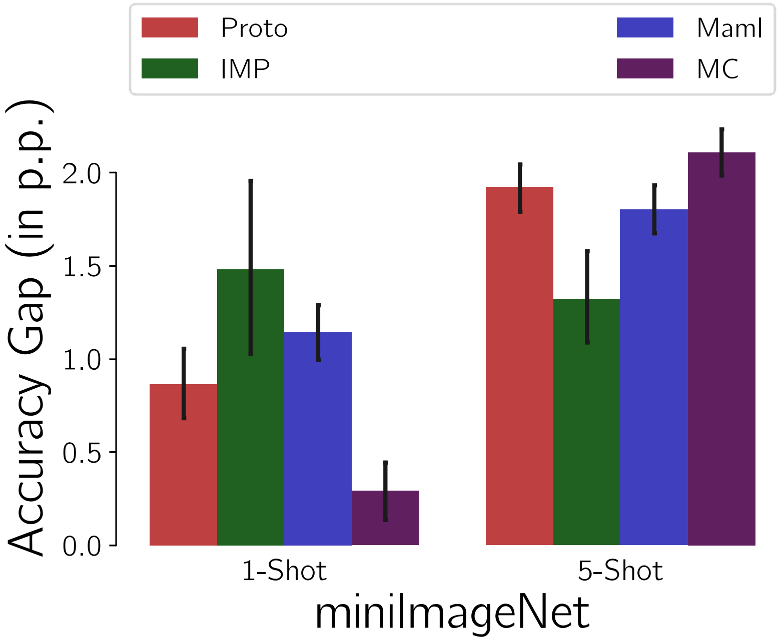
Experiments on mini-ImageNet 5-way 1-shot

✓ Our **regularization** and **normalization** have the intended effects.

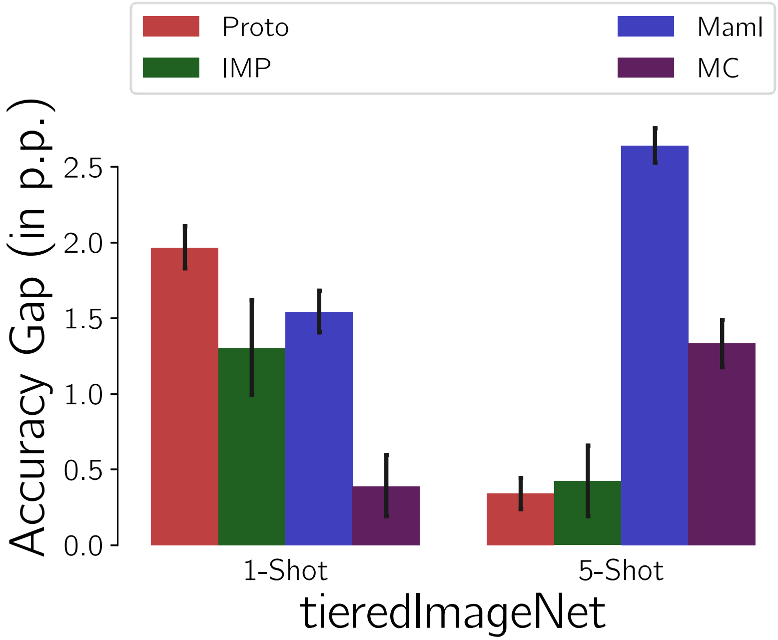
EXPERIMENTAL RESULTS
ACCURACY GAPS



20-way

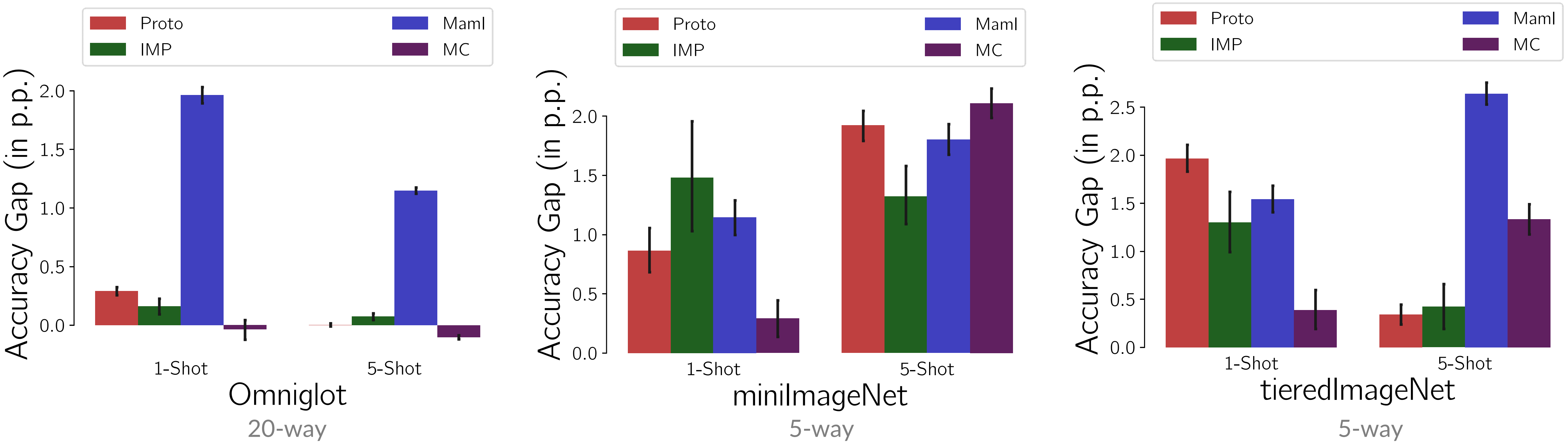


5-way



5-way

EXPERIMENTAL RESULTS
ACCURACY GAPS

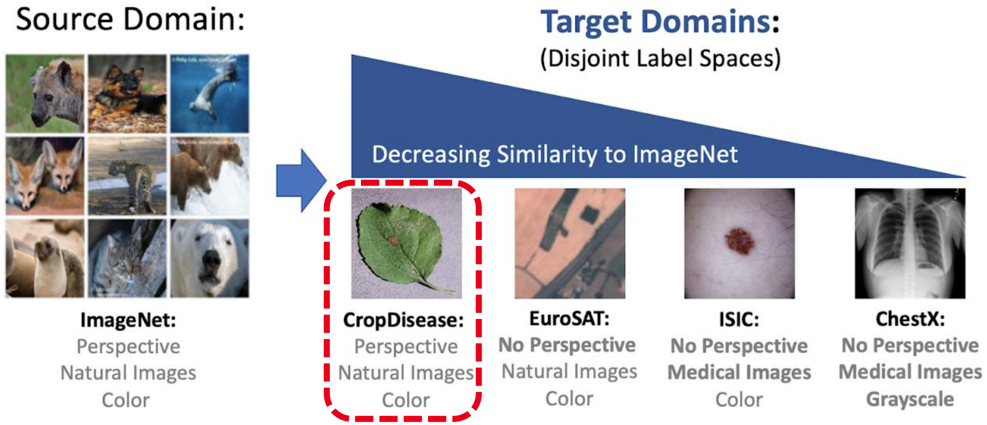


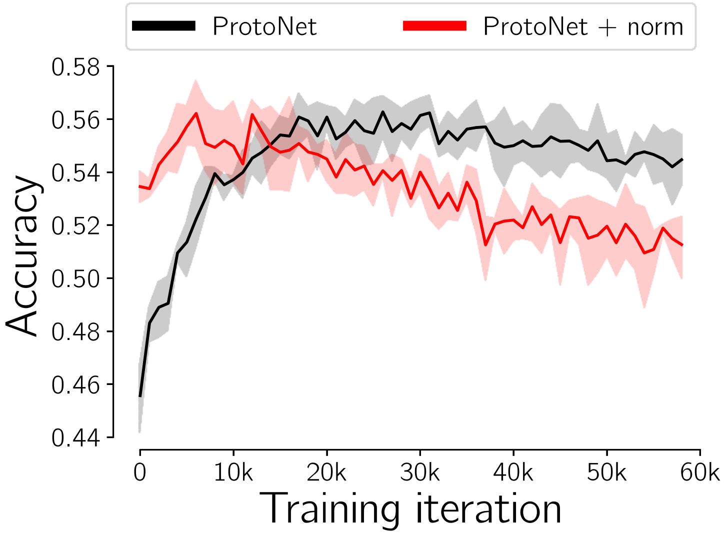
- ✓ **Statistically significant** improvement with our regularization and normalization.
- ✓ Enforcing the assumptions leads to **better generalization** when not verified naturally.

EXPERIMENTAL RESULTS:

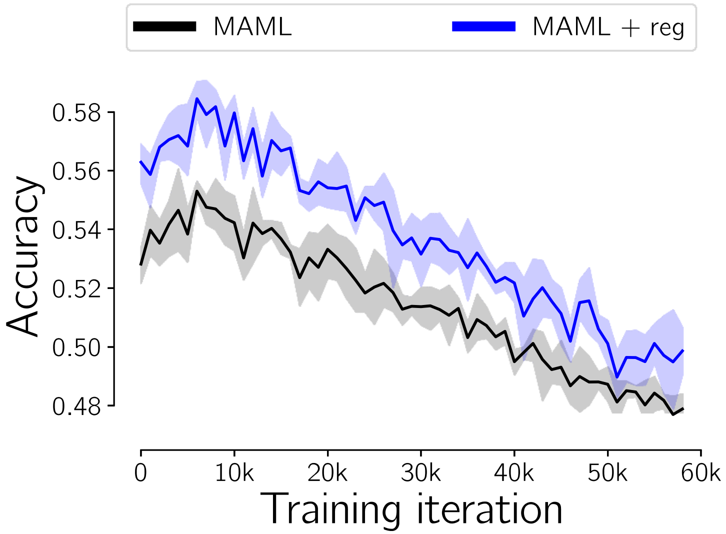
CROSS-DOMAIN

Guo et al. 2020.
A Broader Study of Cross-Domain Few-Shot Learning.
In ECCV 2020



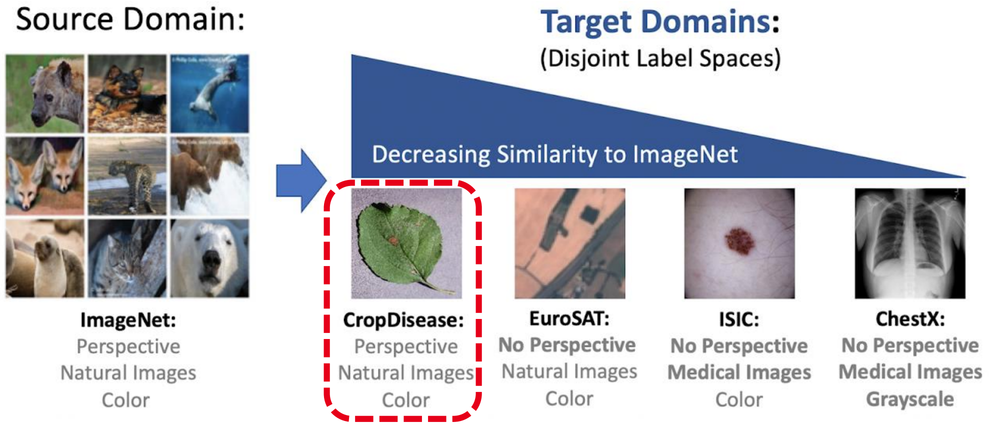


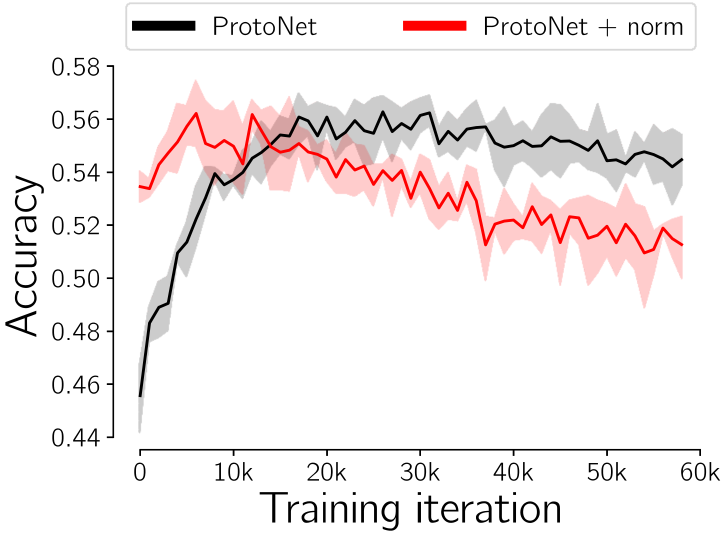
5-way
1-shot



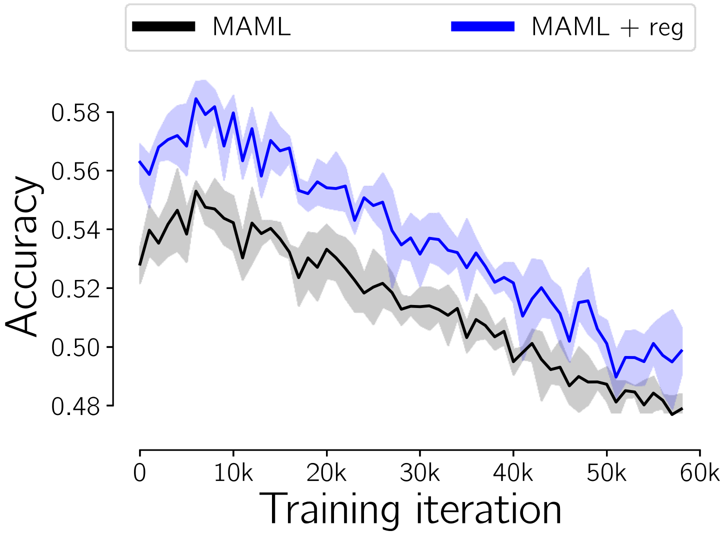
EXPERIMENTAL RESULTS:
CROSS-DOMAIN

Guo et al. 2020.
A Broader Study of Cross-Domain Few-Shot Learning.
In ECCV 2020





5-way
1-shot



- ✗ Improvement does **not** translate to cross-domain for *metric-based methods*.
- ✓ *Gradient-based methods* keep their accuracy gains.

TAKE HOME MESSAGE

TAKE HOME MESSAGE

- **Improving Few-Shot Learning Through Multi-Task Representation Learning Theory**
  - ✓ Connection between Meta-Learning and Multi-Task Representation Learning Theory
  - ✓ Explanations of why some meta-learning methods **naturally fulfill** theoretical assumptions of the best learning bounds.
  - ✓ **Practical ways** to enforce the assumptions which leads to **significant** performance improvements.

More details in arXiv paper:

Contact:

quentin.bouniot@cea.fr

@QBouniot

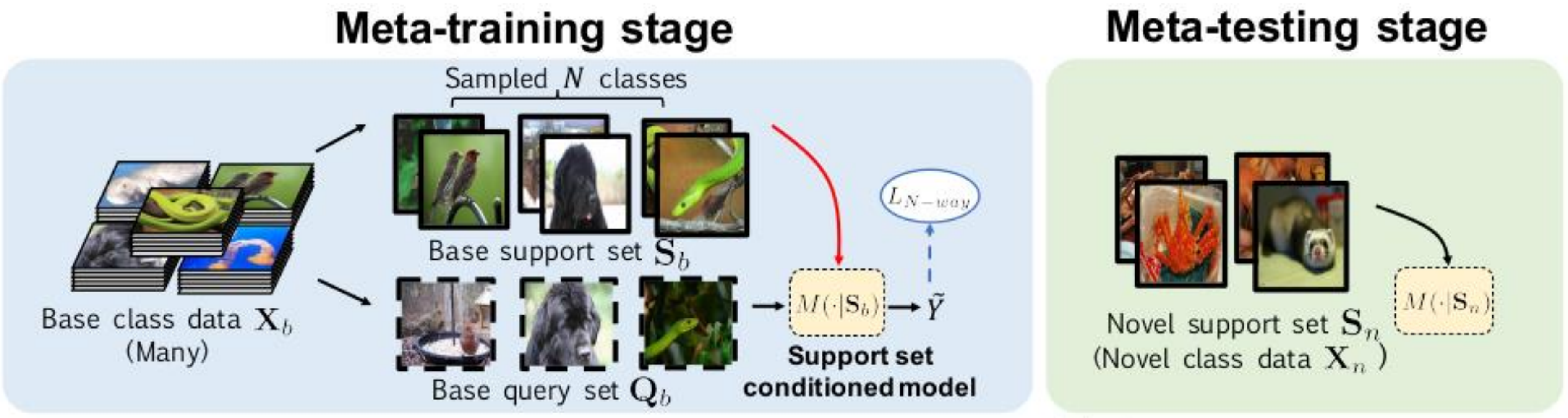
https://qbouniot.github.io

Thank you for listening !

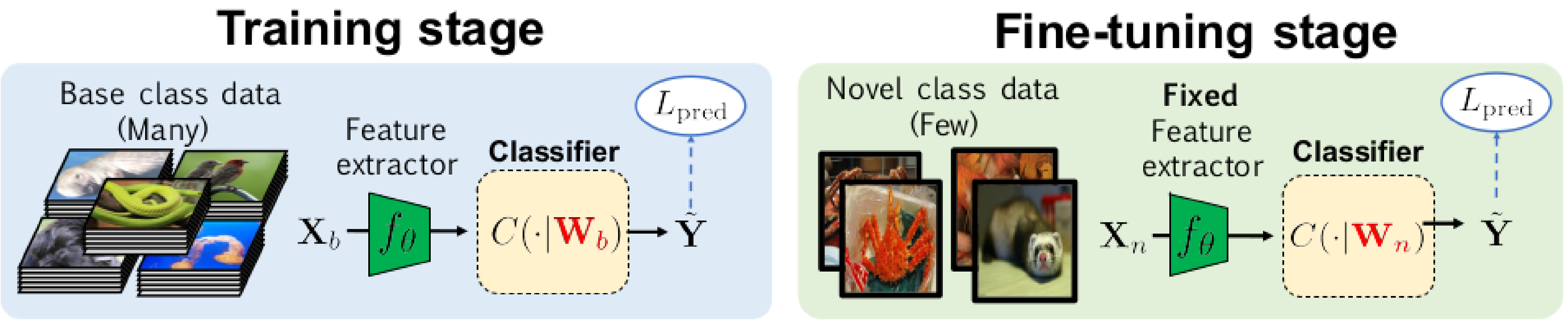
Commissariat à l’énergie atomique et aux énergies alternatives
Institut List | CEA SACLAY NANO-INNOV | BAT. 861 – PC142
91191 Gif-sur-Yvette Cedex - FRANCE
www-list.cea.fr

APPENDIX:
EPISODIC TRAINING VS REGULAR TRAINING

Episodic Training



Regular Training

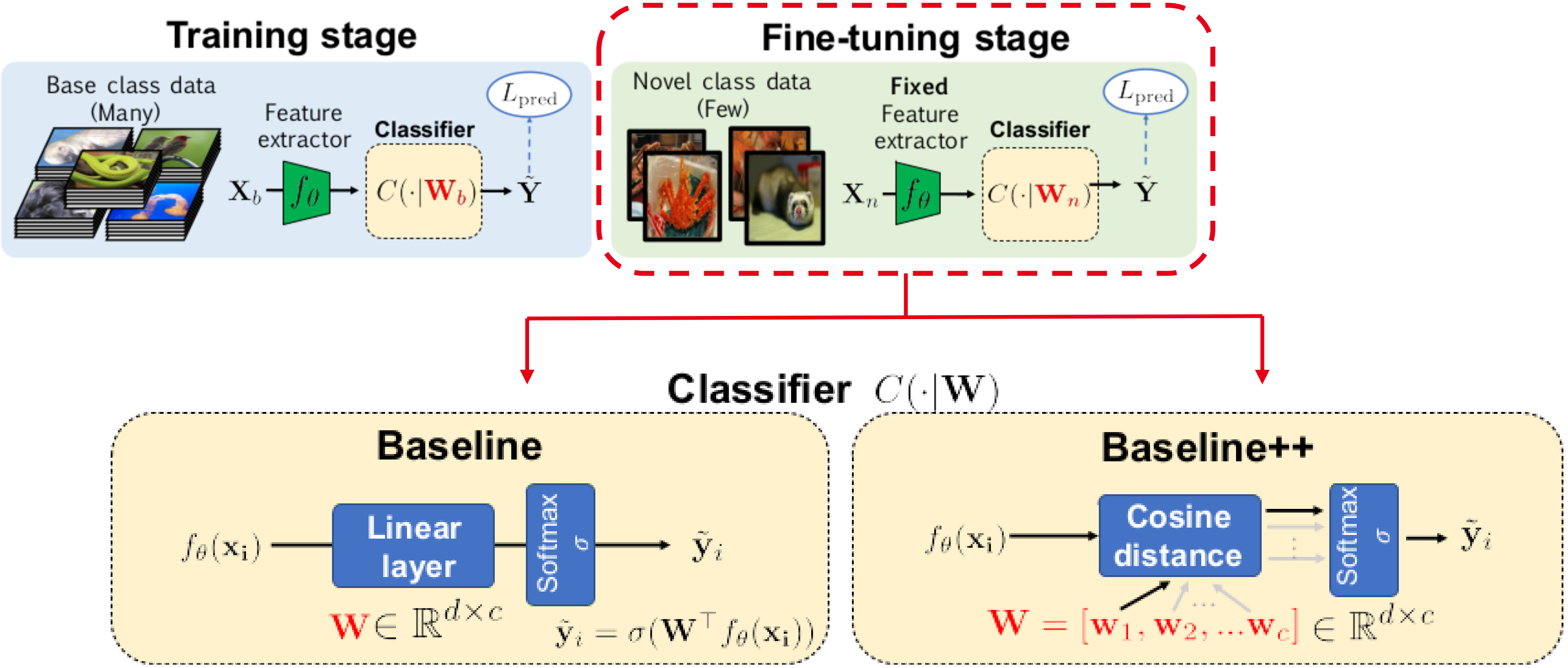


Learning a single episode

Chen W.-Y. et al., *A Closer Look at Few-Shot Classification.* In ICLR 2019

FINE-TUNING METHODS

Regular Training



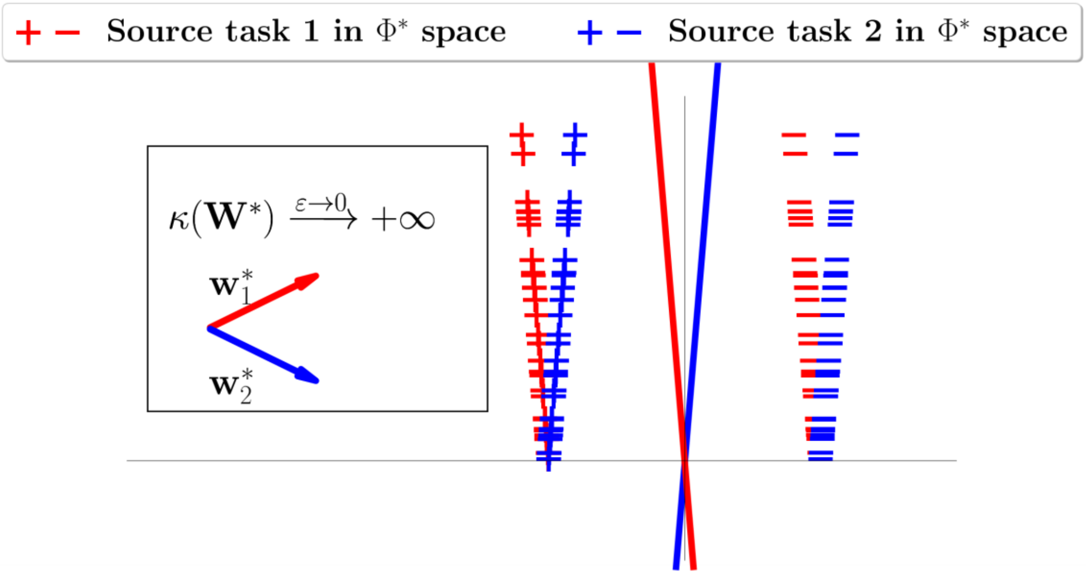
Adapted from [Chen19]

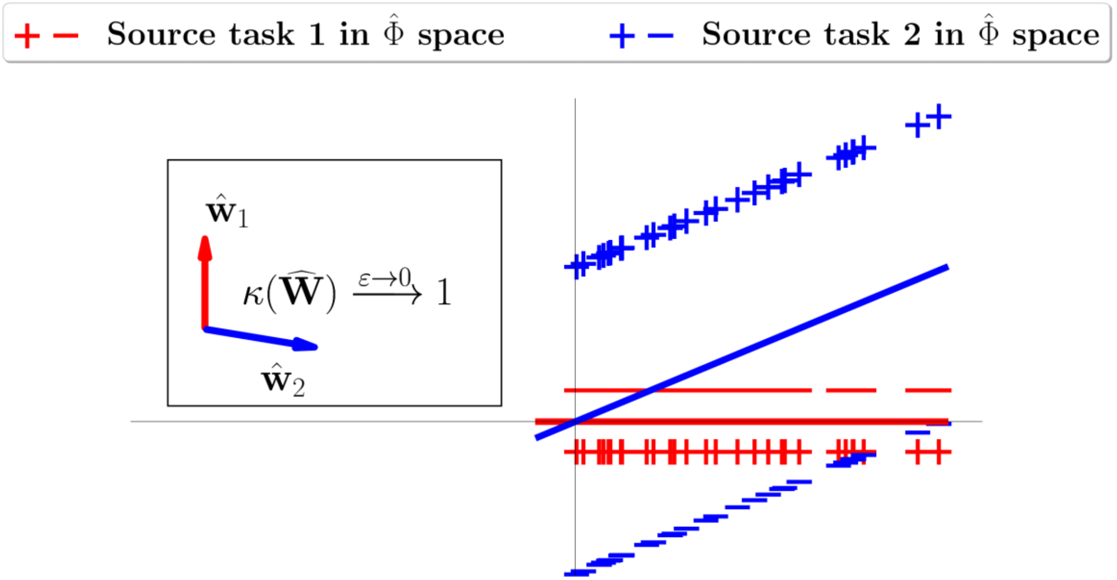
- ***Baseline*** uses a dot product in the classification layer followed by a softmax
- ***Cosine classifier*** (or ***Baseline++***) uses a cosine similarity followed by a softmax

Gidaris S. et Komodakis N., *Dynamic Few-Shot Visual Learning Without Forgetting.* In CVPR 2018

CAN WE FORCE THE ASSUMPTIONS ?

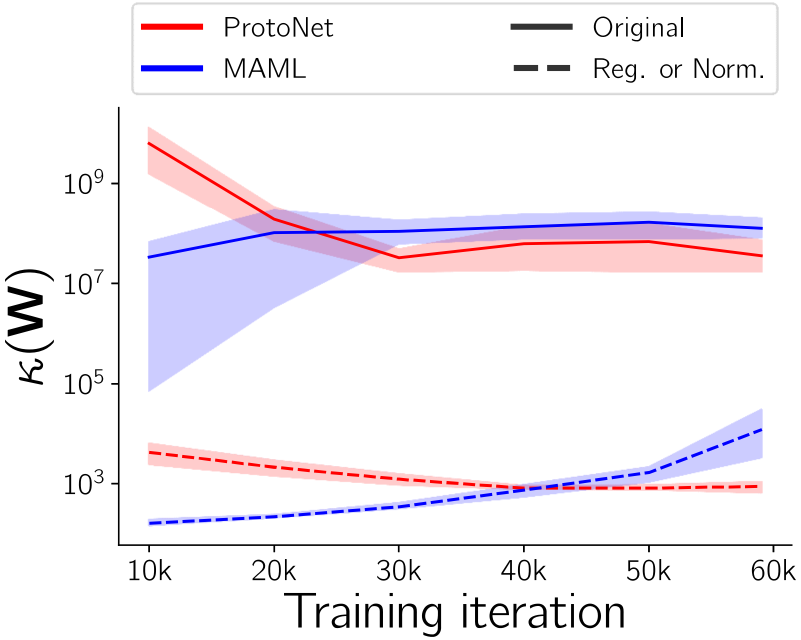
Given $\mathbf{W}^*$ *such that* $\kappa(\mathbf{W}^*) \gg 1$*, can we learn* $\widehat{\mathbf{W}}$ *with* $\kappa(\widehat{\mathbf{W}}) \approx 1$ *while solving the underlying classification problems equally well ?*





✓ Even when $\mathbf{W}^*$ does not satisfy the assumptions, it is **possible to learn** $\hat{\phi}$ to respect them

APPENDIX: CONDITION NUMBER OF ALL PREDICTORS



- ► $\kappa(\mathbf{W}_N)$ shows **dynamics** during training, but values are not comparable
- ► $\kappa(\mathbf{W})$ is **intractable to compute** during training.