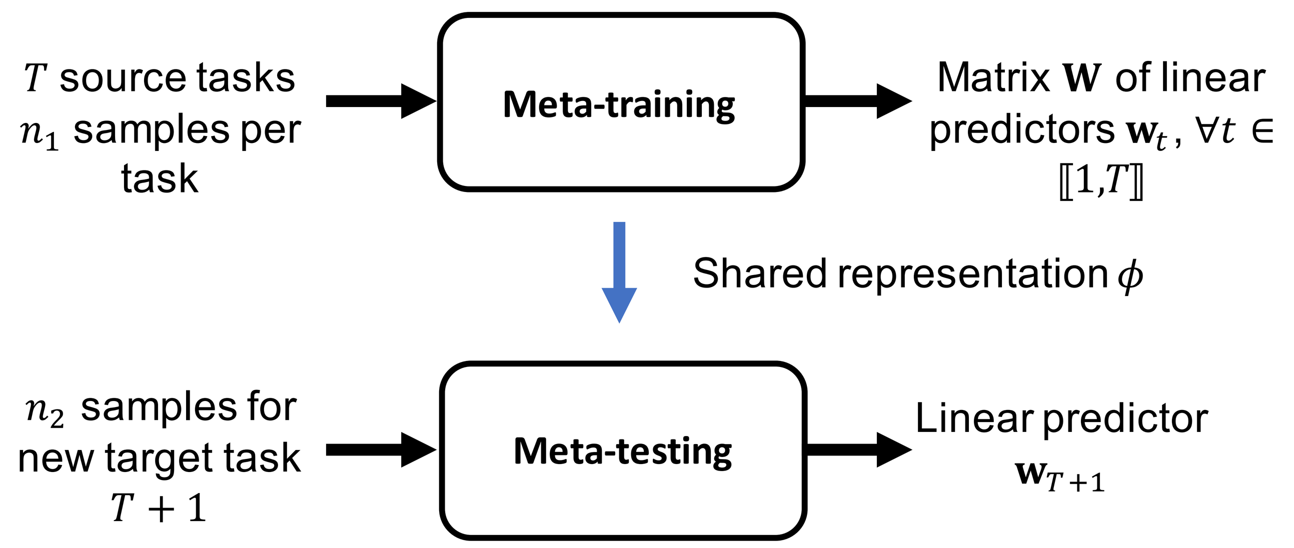
Vers une Meilleure Compréhension des Méthodes de Méta-Apprentissage à Travers la Théorie de l'Apprentissage de Représentations Multi-tâches

Quentin Bouniot †‡ Ievgen Redko ‡ Romaric Audigier † Angélique Loesch †

†Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France ‡Université de Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

Multi-Task Representation Learning (MTR)



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

with $\mathcal{L}$ the *risk*, ϕ^* the optimal representation and $\mathbf{w}^*_{T+1}$ the optimal linear predictor for task $T+1$.

When does MTR Provably Work?

Assumption 1: Diversity of the source tasks

The matrix of optimal predictors $\mathbf{W}^* = [\mathbf{w}^*_1, \ldots, \mathbf{w}^*_T]$ should cover all the directions evenly.

Assumption 2: Constant classification margin

The norm of the optimal predictors $\{\mathbf{w}^*_t\}_{t\in[\![1,T]\!]}$ should not increase with the number of tasks.

Learning bound [1, 2]

With these assumptions, we can derive:

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

Putting Theory to Work

Ensuring assumption 1.

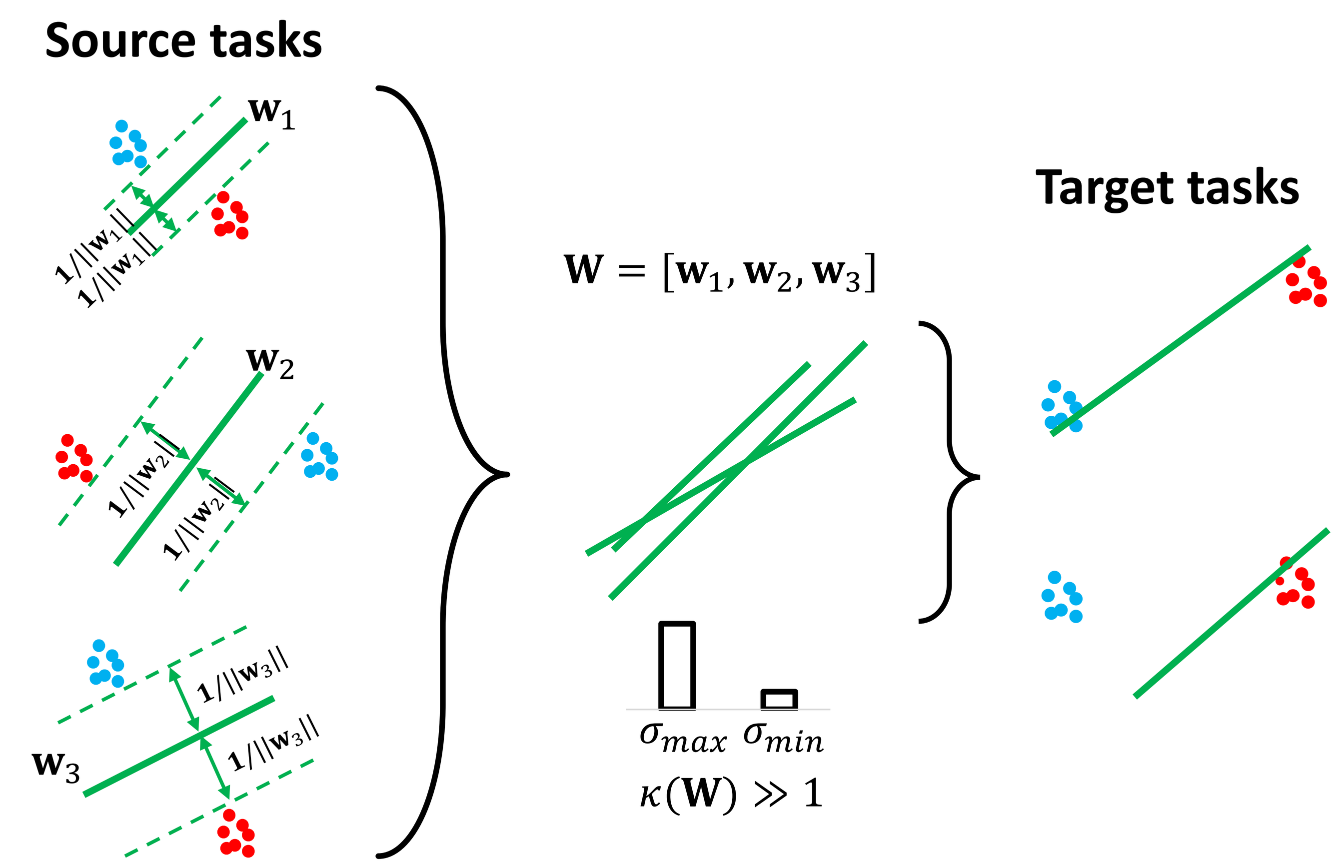
$$\kappa(\mathbf{W}) = \frac{\sigma_{max}(\mathbf{W})}{\sigma_{min}(\mathbf{W})} \quad \text{or} \quad H_\sigma(\mathbf{W}) = \sum_{i=1}^{N} \mathrm{softmax}(\sigma(\mathbf{W}))_i \cdot \log \mathrm{softmax}(\sigma(\mathbf{W}))_i$$

Adding κ or H_σ as a regularization term leads to a **better coverage** of representation space $\mathbb{R}^k$.

Ensuring assumption 2.

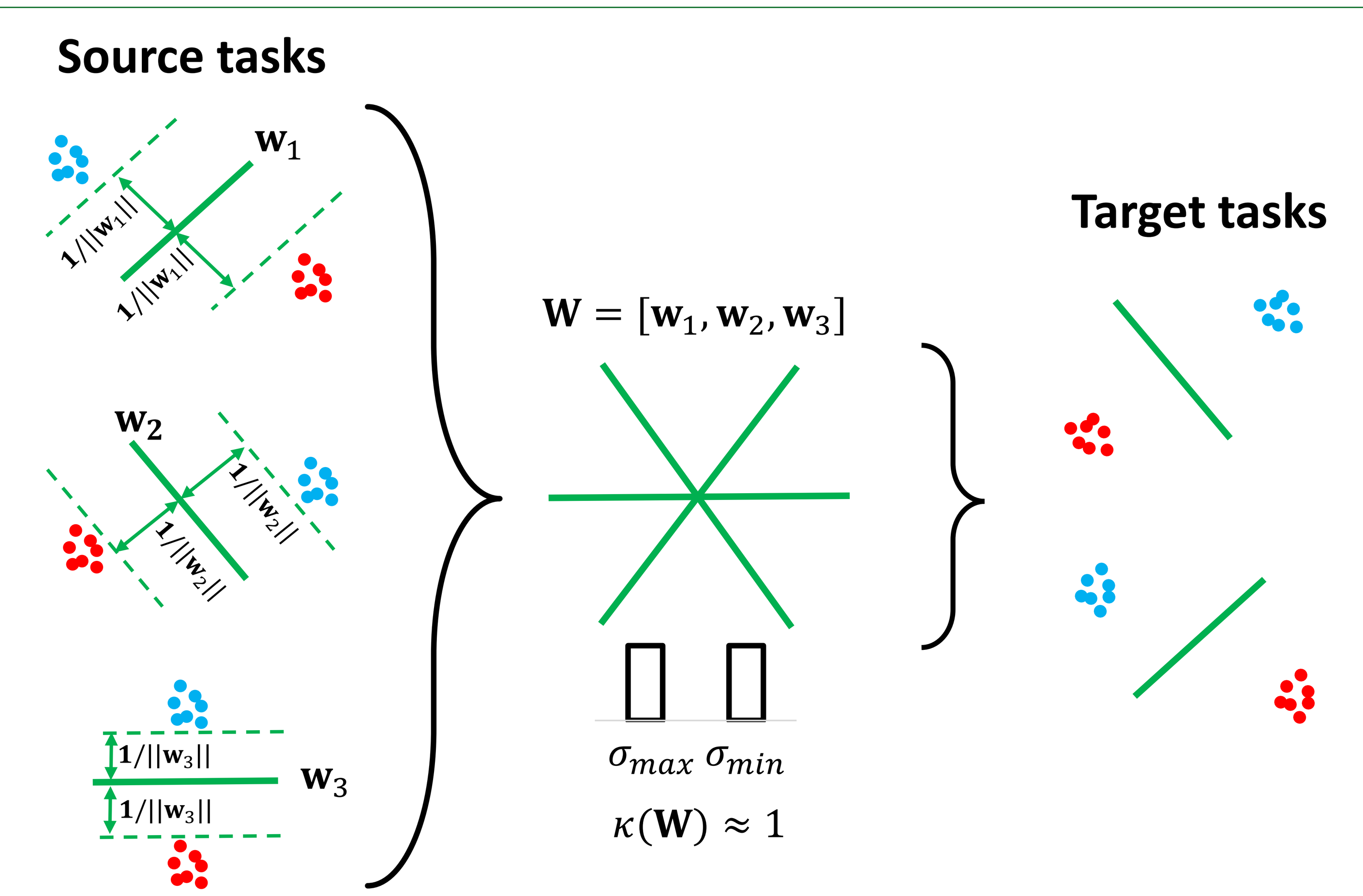
Regularizing the norm or **normalizing** the linear predictors

Without Regularization



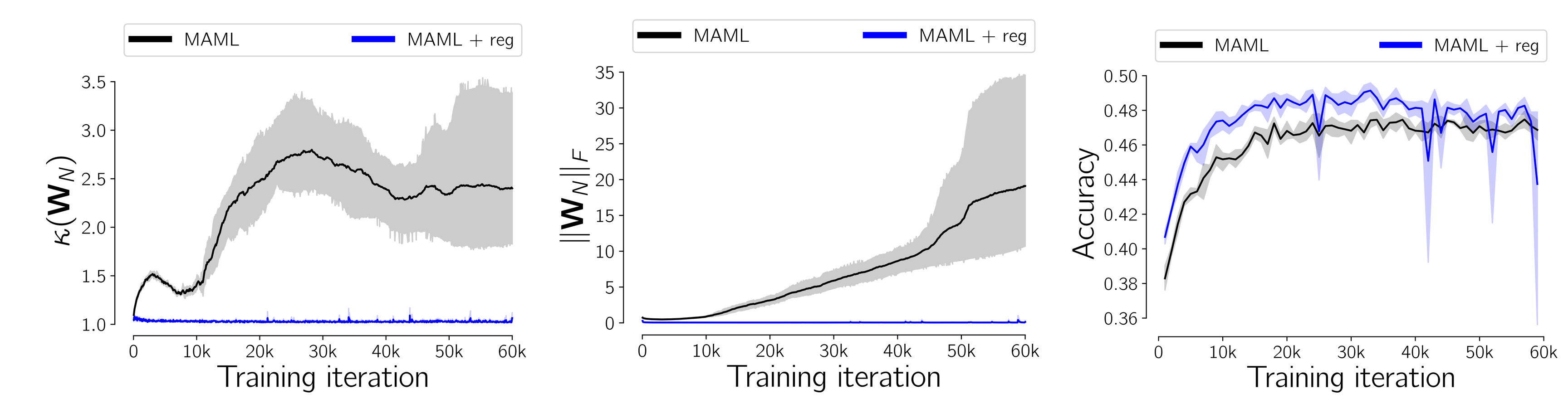
- Linear predictors can be **biased** from previous tasks and **cover a single part** of the space.
- With few examples per task, linear predictors can be **over-specialized**.

With Regularization

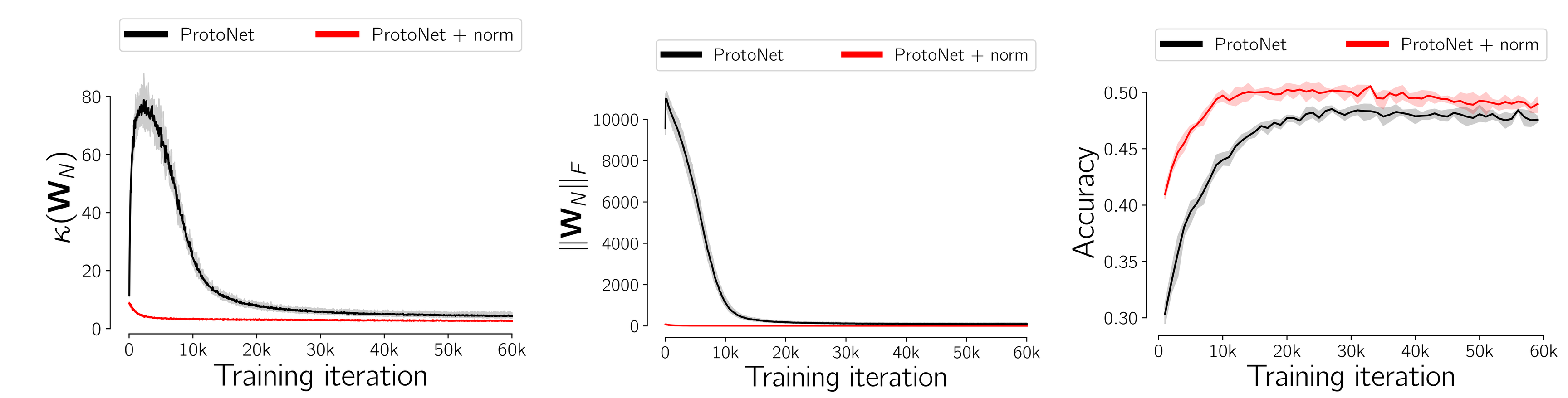


- Assumption 1 makes sure that linear predictors are **complementary** to each other.
- Assumption 2 avoids **over- or under-specialization**.

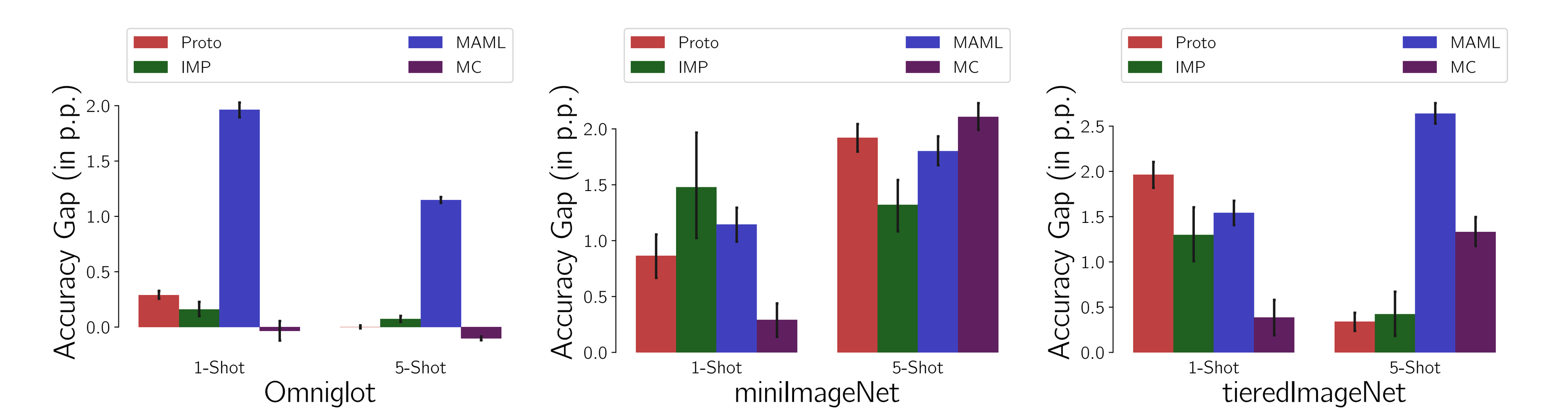
Practical Results



✗ MAML does not verify the assumptions.



✓ ProtoNet naturally verifies the assumptions.



✓ **Statistically significant improvement** with our regularization and normalization.

Take Home Message

- **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.

References

[1] Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representation*, 2021.

[2] Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable Meta-Learning of Linear Representations. In *arXiv:2002.11684*, 2020.