A Appendix

A.1 Notations

We introduce here several notations used throughout the main theoretical meta-learning papers [3, 10, 11]. We denote by $\mu_{\mathbb{X}_t}$ the marginal distribution of $\mathbf{x}_t$ and its covariance matrix by $\Sigma_t = \mathbb{E}_{\mathbf{x}\sim\mu_{\mathbb{X}_t}}[\mathbf{x}\mathbf{x}^T]$. We further use $\sigma_i(\cdot)$ to denote the i^{th} singular value of a matrix, let $\bar{R} = ||\hat{\phi}^*\mathbf{W}^*||_*/\sqrt{T}$, where $||\cdot||_*$ denotes the nuclear norm and $\hat{\phi}^*$ is the matrix of the transformation applied to the samples from $\mathbb{X}$. The point-wise and uniform covariance convergence refers to the fact that empirical covariance matrices converge to their true counterparts with the increasing number of samples. In [10], the authors further assume that random vectors have zero mean, *i.e.*, $\mathbb{E}_{\mathbf{x}\sim\mu_{\mathbb{X}_t}}[\mathbf{x}] = 0$ for all t and that $\mathbf{x} \sim \mu_{\mathbb{X}_t}$ can be written as $\Sigma_t^{1/2}\bar{\mathbf{x}}$ with $\bar{\mathbf{x}}$ having zero mean and identity covariance matrix. Finally, when considering a two-layer neural network (NN) with Rectifier Linear Unit (ReLU) activation function, the data generating model presented in Eq. 2 is modified by applying the ReLU activation to $\hat{\phi}(\cdot)$. This is denoted as a teacher network assumption. As, for the work of [11], we refer to the method of methods when using SVD to find the top k singular vectors of $\frac{1}{n_1}\sum_{t=1}^{T}\sum_{i=1}^{n_1} y_{t,i}^2\mathbf{x}_{t,i}\mathbf{x}_{t,i}^T$, while the linear regression stands for calculating the traditional closed-form solution on the transformed target task given by $\hat{\mathbf{w}}_{T+1} = (\sum_{i=1}^{n_2}\hat{\phi}(\mathbf{x}_{T+1,i})\hat{\phi}(\mathbf{x}_{T+1,i})^T)^{-1}\hat{\phi}^T\sum_{i=1}^{n_2}\mathbf{x}_{T+1,i}y_{T+1,i}$.

A.2 Detailed experimental setups

Common architecture For all methods, we use the common architecture used in [25] which consists of 4 modules

Omniglot [21] is a dataset of 20 instances of 1623 characters from 50 different alphabets. Each image was hand-drawn by different people. The images are resized to 28×28 pixels and the classes are augmented with rotations by multiples of 90 degrees.

miniImageNet [22] is a dataset made from randomly chosen classes and images taken from the ILSVRC-12 dataset [28]. The dataset consists of 100 classes and 600 images for each class. The images are resized to 84×84 pixels.

tieredImageNet [23] is also a subset of ILSVRC-12 dataset. However, unlike miniImageNet, training classes are semantically unrelated to testing classes. The dataset consists of $779,165$ images divided into 608 classes. Here again, the images are resized to 84×84 pixels.

A.3 Performance comparisons with according evaluation settings

Table 2 shows the performance of our reproduced methods, MAML[24], PROTONET[26], BASELINE[22] and BASELINE++[27], compared to the reported results for the according training and evaluation setting to validate our implementations. We can see that our performance are on par with corresponding reported results. We attribute the differences to minor variations in implementations. Table 3 provides the detailed performance of our reproduced methods with and without our regularization (or normalization for PROTONET). Theses results are summarized in Table 1 of our paper and discussions about them can be found in Section 4.

A.4 Ablative studies

In the following, we include ablative studies on the effect of each terms in our regularization scheme to complete results given in Section 4 of our paper. In Table 4, we compared the performance of our reproduced MAML without regularization, with a regularization on the ratio of singular values, on the norm of the linear predictors, and with both regularization terms on Omniglot. We can see that both regularization terms are important in the training and that using only a single term can be detrimental to the training results.

In Table 5, we report the performance of our reproduced PROTONET without normalization, with normalization and with both normalization and regularization on the entropy. We can see that further

Method	Dataset	Episodes	Reported	Reproduced
MAML	Omniglot	20-way 1-shot	$93.7^{*} \pm 0.7\%$	$91.72 \pm 0.29\%$
		20-way 5-shot	$96.4^{*} \pm 0.1\%$	$97.07 \pm 0.14\%$
	miniImageNet	5-way 1-shot	$46.47^{\dagger} \pm 0.82\%$	$47.93 \pm 0.83\%$
		5-way 5-shot	$62.71^{\dagger} \pm 0.71\%$	$64.47 \pm 0.69\%$
	tieredImageNet	5-way 1-shot	/	$50.08 \pm 0.91\%$
		5-way 5-shot	/	$67.5 \pm 0.79\%$
PROTONET	Omniglot	20-way 1-shot	$96.00^{\sharp}$	$95.56 \pm 0.10\%$
		20-way - 5-shot	$98.90^{\sharp}$	$98.80 \pm 0.04\%$
	miniImageNet	5-way 1-shot	$44.42^{\dagger} \pm 0.84\%$	$49.53 \pm 0.41\%$
		5-way 5-shot	$64.24^{\dagger} \pm 0.72\%$	$65.10 \pm 0.35\%$
	tieredImageNet	5-way 1-shot	/	$51.95 \pm 0.45\%$
		5-way 5-shot	/	$71.61 \pm 0.38\%$
BASELINE	Omniglot	20-way 1-shot	/	$78.18 \pm 0.43\%$
		20-way 5-shot	/	$95.34 \pm 0.15\%$
	miniImageNet	5-way 1-shot	$42.11^{\dagger} \pm 0.71\%$	$42.35 \pm 0.73\%$
		5-way 5-shot	$62.53^{\dagger} \pm 0.69\%$	$59.58 \pm 0.71\%$
	tieredImageNet	5-way 1-shot	/	$44.59 \pm 0.76\%$
		5-way 5-shot	/	$66.38 \pm 0.75\%$
BASELINE++	Omniglot	20-way 1-shot	/	$77.00 \pm 0.49\%$
		20-way 5-shot	/	$94.18 \pm 0.17\%$
	miniImageNet	5-way 1-shot	$48.24^{\dagger} \pm 0.75\%$	$48.06 \pm 0.76\%$
		5-way 5-shot	$66.43^{\dagger} \pm 0.63\%$	$65.00 \pm 0.68\%$
	tieredImageNet	5-way 1-shot	/	$52.70 \pm 0.87\%$
		5-way 5-shot	/	$71.58 \pm 0.74\%$

Table 2: Our reproduced performances compared to reported performances from the according evaluation settings. All accuracy results (in %) are averaged over 2400 test episodes and 4 different random seeds and are reported with 95% confidence interval. *: Results reported from [12]. †: Results reported from [25]. $^{\sharp}$: Results reported from [26].

enforcing a regularization on the singular values (through the entropy) does not help the training since PROTONET naturally learns to minimize the singular values of the prototypes.

In Table 6 and 7, we show the effect of regularization on different part of the training process of BASELINE and BASELINE++ respectively. The regularization used in training is limited to the ratio of singular values R_σ, whereas during finetuning, we regularize both the ratio R_σ and the norm $\|\mathbf{W}_N\|_F$. We can see that for BASELINE, similarly to MAML, both regularization terms are important *on miniImageNet and tieredImageNet*. For BASELINE++, on the other hand, learning with any of the regularization terms neither improves nor decreases performance in a statistically significant manner.

Method	Dataset	Episodes	without Reg./Norm.	with Reg./Norm.
MAML	Omniglot	1-shot	$91.72 \pm 0.29\%$	$\mathbf{95.67 \pm 0.20\%}$
		5-shot	$97.07 \pm 0.14\%$	$\mathbf{98.24 \pm 0.10\%}$
	miniImageNet	1-shot	$47.93 \pm 0.83\%$	$\mathbf{49.16 \pm 0.85\%}$
		5-shot	$64.47 \pm 0.69\%$	$\mathbf{66.43 \pm 0.69\%}$
	tieredImageNet	1-shot	$50.08 \pm 0.91\%$	$\mathbf{51.5 \pm 0.90\%}$
		5-shot	$67.5 \pm 0.79\%$	$\mathbf{70.16 \pm 0.76\%}$
PROTONET	Omniglot	1-shot	$95.56 \pm 0.10\%$	$\mathbf{95.89 \pm 0.10\%}$
		5-shot	$\mathbf{98.80 \pm 0.04\%}$	$\mathbf{98.80 \pm 0.04\%}$
	miniImageNet	1-shot	$49.53 \pm 0.41\%$	$\mathbf{50.29 \pm 0.41\%}$
		5-shot	$65.10 \pm 0.35\%$	$\mathbf{67.13 \pm 0.34\%}$
	tieredImageNet	1-shot	$51.95 \pm 0.45\%$	$\mathbf{54.05 \pm 0.45\%}$
		5-shot	$\mathbf{71.61 \pm 0.38\%}$	$\mathbf{71.84 \pm 0.38\%}$
BASELINE	Omniglot	1-shot	$\mathbf{86.85 \pm 0.36\%}$	$73.65 \pm 0.52\%$
		5-shot	$96.95 \pm 0.12\%$	$\mathbf{97.61 \pm 0.11\%}$
	miniImageNet	1-shot	$42.35 \pm 0.73\%$	$\mathbf{43.87 \pm 0.75\%}$
		5-shot	$59.58 \pm 0.71\%$	$\mathbf{61.24 \pm 0.71\%}$
	tieredImageNet	1-shot	$44.59 \pm 0.76\%$	$\mathbf{50.02 \pm 0.82\%}$
		5-shot	$66.38 \pm 0.75\%$	$\mathbf{68.30 \pm 0.74\%}$
BASELINE++	Omniglot	1-shot	$\mathbf{82.5 \pm 0.39\%}$	$75.21 \pm 0.47\%$
		5-shot	$\mathbf{95.49 \pm 0.15\%}$	$93.25 \pm 0.20\%$
	miniImageNet	1-shot	$\mathbf{48.06 \pm 0.76\%}$	$\mathbf{48.45 \pm 0.78\%}$
		5-shot	$\mathbf{65.00 \pm 0.68\%}$	$\mathbf{64.87 \pm 0.68\%}$
	tieredImageNet	1-shot	$\mathbf{52.70 \pm 0.87\%}$	$\mathbf{52.98 \pm 0.88\%}$
		5-shot	$\mathbf{71.58 \pm 0.74\%}$	$\mathbf{70.86 \pm 0.74\%}$

Table 3: Performance of several meta-learning algorithms without and with our regularization (or normalization in the case of PROTONET) to enforce the theoretical assumptions. All accuracy results (in %) are averaged over 2400 test episodes and 4 different seeds and are reported with 95% confidence interval. Episodes are 20-way classification for Omniglot and 5-way classification for miniImageNet and tieredImageNet.

Episodes	Reproduced	Ratio	Norm	Ratio + Norm
20-way 1-shot	$91.72 \pm 0.29\%$	$89.86 \pm 0.31\%$	$92.80 \pm 0.26\%$	$\mathbf{95.67 \pm 0.20\%}$
20-way 5-shot	$97.07 \pm 0.14\%$	$72.47 \pm 0.17\%$	$96.99 \pm 0.14\%$	$\mathbf{98.24 \pm 0.10\%}$

Table 4: Ablative study of the regularization parameter for MAML on Omniglot. All accuracy results (in %) are averaged over 2400 test episodes and 4 different random seeds and are reported with 95% confidence interval. Using both regularization terms is important.

Dataset	Episodes	Reproduced	Norm	Norm + Entropy
Omniglot	20-way 1-shot	$95.56 \pm 0.10\%$	$\mathbf{95.89 \pm 0.10}\%$	$91.90 \pm 0.14\%$
	20-way 5-shot	$\mathbf{98.80 \pm 0.04}\%$	$\mathbf{98.80 \pm 0.04}\%$	$96.40 \pm 0.07\%$
miniImageNet	5-way 1-shot	$49.53 \pm 0.41\%$	$\mathbf{50.29 \pm 0.41}\%$	$49.43 \pm 0.40\%$
	5-way 5-shot	$65.10 \pm 0.35\%$	$\mathbf{67.13 \pm 0.34}\%$	$65.71 \pm 0.35\%$
tieredImageNet	5-way 1-shot	$51.95 \pm 0.45\%$	$\mathbf{54.05 \pm 0.45}\%$	$53.54 \pm 0.44\%$
	5-way 5-shot	$\mathbf{71.61 \pm 0.38}\%$	$\mathbf{71.84 \pm 0.38}\%$	$70.30 \pm 0.40\%$

Table 5: Performance of PROTONET with and without our regularization on the entropy and/or normalization. All accuracy results (in %) are averaged over 2400 test episodes and 4 different random seeds and are reported with 95% confidence interval. Further enforcing regularization on the singular values can be detrimental to performance.

Dataset	Episodes	Reproduced	Reg. in training	Reg. in finetuning	Reg. in both
miniImageNet	5-way 1-shot	$42.35 \pm 0.73\%$	$43.12 \pm 0.73\%$	$43.32 \pm 0.76\%$	$\mathbf{43.87 \pm 0.75}\%$
	5-way 5-shot	$59.58 \pm 0.71\%$	$60.17 \pm 0.71\%$	$60.72 \pm 0.70\%$	$\mathbf{61.24 \pm 0.71}\%$
tieredImageNet	5-way 1-shot	$44.59 \pm 0.76\%$	$\mathbf{49.49 \pm 0.83}\%$	$45.78 \pm 0.75\%$	$\mathbf{50.02 \pm 0.82}\%$
	5-way 5-shot	$66.38 \pm 0.75\%$	$\mathbf{68.66 \pm 0.74}\%$	$66.19 \pm 0.74\%$	$\mathbf{68.30 \pm 0.74}\%$

Table 6: Ablative study on the effect of the regularization on different parts of training process of BASELINE. All accuracy results (in %) are averaged over 2400 test episodes and 4 random seeds and are reported with 95% confidence interval. Similarly to MAML, both regularization terms are important.

Dataset	Episodes	Reproduced	Reg. in training	Reg. in finetuning	Reg. in both
miniImageNet	5-way 1-shot	$48.06 \pm 0.76\%$	$47.83 \pm 0.78\%$	$48.66 \pm 0.79\%$	$48.45 \pm 0.78\%$
	5-way 5-shot	$65.00 \pm 0.68\%$	$64.71 \pm 0.68\%$	$65.35 \pm 0.68\%$	$64.87 \pm 0.68\%$
tieredImageNet	5-way 1-shot	$52.70 \pm 0.87\%$	$52.75 \pm 0.87\%$	$52.83 \pm 0.87\%$	$52.98 \pm 0.88\%$
	5-way 5-shot	$71.58 \pm 0.74\%$	$71.03 \pm 0.74\%$	$71.64 \pm 0.74\%$	$70.86 \pm 0.74\%$

Table 7: Ablative study on the effect of the regularization on different parts of training process of BASELINE++. All accuracy results (in %) are averaged over 2400 test episodes and 4 random seeds and are reported with 95% confidence interval. Similarly to PROTONET, further enforcing regularization does not improve nor decrease performance.