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 Review of the meta-learning theory

We now formulate the main results of the three main theoretical analyses of meta-learning provided in [3, 10, 11] in Table 2.

Paper	Assumptions	Φ	Bound
[3]	**A1**. $\forall t \in [[T+1]],\ \mu_t \sim \eta$	–	$O\left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{T}}\right)$
[10]	**A2.1**. $\forall t$, $\bar{\mathbf{x}}$ is ρ^2-subgaussian		
	A2.2. $\forall t \in [[T]], \exists c > 0 : \Sigma_t \succeq c\Sigma_{T+1}$	**A2.1-2.4**, linear, $k \ll d$	$O\left(\frac{kd}{cn_1T} + \frac{k}{n_2}\right)$
	A2.3. $\frac{\sigma_1(\mathbf{W}^*)}{\sigma_k(\mathbf{W}^*)} = O(1)$	**A2.3-2.5**, general, $k \ll d$	$O\left(\frac{\mathcal{C}(\Phi)}{n_1T} + \frac{k}{n_2}\right)$
	A2.4. $\mathbf{w}_{T+1}^* \sim \mu_{\mathbf{w}} : \Vert\,\mathbb{E}_{\mathbf{w}\sim\mu_{\mathbf{w}}}[\mathbf{w}\mathbf{w}^T]\Vert \leq O(\frac{1}{k})$	**A2.1,2.5,2.6**, linear + ℓ_2 regul., $k \gg d$	$\sigma\bar{R}\tilde{O}\left(\frac{\sqrt{\text{Tr}(\Sigma)}}{\sqrt{n_1T}} + \frac{\sqrt{\Vert\Sigma\Vert_2}}{\sqrt{n_2}}\right)$
	A2.5. $\forall t$, $p_t = p, \Sigma_t = \Sigma$	**A2.1,2.5,2.6,2.7**, two-layer NN (ReLUs+ ℓ_2 regul.)	$\sigma\bar{R}\tilde{O}\left(\frac{\sqrt{\text{Tr}(\Sigma)}}{\sqrt{n_1T}} + \frac{\sqrt{\Vert\Sigma\Vert_2}}{\sqrt{n_2}}\right)$
	A2.6. Point-wise+unif. cov. convergence		
	A2.7. Teacher network		
[11]	**A3.1**. $\forall t$, $\mathbf{x} \sim \mu_{\mathbb{X}_t}$ is ρ^2-subgaussian		
	A3.2. $\frac{\sigma_1(\mathbf{W}^*)}{\sigma_k(\mathbf{W}^*)} = O(1)$	**A1-4**, linear, $k \ll d$	$\tilde{O}\left(\frac{kd}{n_1T} + \frac{k}{n_2}\right)$
	A3.3. $\widehat{\mathbf{W}}$ learned using the Method of Moments		
	A3.4. $\mathbf{w}_{T+1}^*$ is learned using Linear Regression		

Table 2: Overview of main theoretical contributions related to meta-learning with their assumptions, considered classes of representations and the obtained bounds on the excess risk. Here $\tilde{O}(\cdot)$ hides logarithmic factors.

One may note that all the assumptions presented in this table can be roughly categorized into two groups. First one consists of the assumptions related to the data generating process (A1, A2.1, A2.4-7 and A3.1), technical assumptions required for the manipulated empirical quantities to be well-defined (A2.6) and assumptions specifying the learning setting (A3.3-4). We put them together as they are not directly linked to the quantities that we optimize over in order to solve the meta-learning problem. The second group of assumptions include A2.2 and A3.2: both defined as a measure of diversity between source tasks' predictors that are expected to cover all the directions of $\mathbb{R}^k$ evenly. This assumption is of primary interest as it involves the matrix of predictors optimized in Eq. 1 as thus one can attempt to force it in order for $\hat{\mathbf{W}}$ to have the desired properties. Finally, we note that assumption A3.2 related to the covariance dominance can be seen as being at the intersection between the two groups. On the one hand, this assumption is related to the population covariance and thus is related to the data generating process that is supposed to be fixed. On the other hand, we can think about a pre-processing step that precedes the meta-train step of the algorithm and transforms the source and target tasks' data so that their sample covariance matrices satisfy A3.2. While presenting a potentially interesting research direction, it is not clear how this can be done in practice especially under a constraint of the largest value of c required to minimize the bound.

A.3 Intuition behind the assumptions

An intuition behind the assumptions and the regularization terms can be seen in Fig. 3.

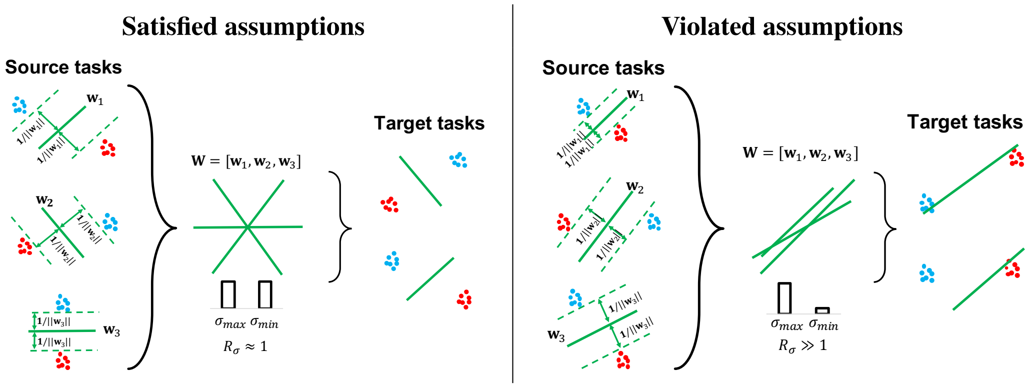

Figure 3: Illustration of the intuition behind the assumptions derived from the few-shot learning theory. **(left)** When the assumptions are satisfied, the linear predictors cover the embedding space evenly and their norm remains roughly constant on source tasks making them useful for a previously unseen task. **(right)** Lack of diversity and increasing norm of the linear predictors restrict them from being useful on the target task.

A.4 Existence of the subgradients of singular values functions

According to [30], in their theorem 7.1, subgradients of singular values function are defined for *absolutely symmetric functions*. In our case, we are computing the squared singular values $\sigma^2(W)$ and we retrieve the singular values by taking the square root before the ratio or the entropy. This means that effectively, we are computing $R_\sigma(W) = max(|\sigma(W)|)/min(|\sigma(W)|)$ and $H_\sigma(W) = -\sum_{i=1}^{N} \operatorname{softmax}(|\sigma(W)|)_i \cdot \log \operatorname{softmax}(|\sigma(W)|)_i$, which are both *absolutely symmetric functions*. Consequently, subgradients of both R_σ and H_σ are defined.

A.5 Detailed experimental setups

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 [31]. 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.6 Performance comparisons with according evaluation settings

Table 3 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 and we attribute the differences to minor variations in implementations such as data augmentation. Table 4 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.

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 3: 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].

A.7 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 5, 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.

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 4: 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.

In Table 6, 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 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 7 and 8, 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

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 5: 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 6: 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.

$\|\mathbf{W}_N\|_F$. 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.

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 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 MAML, both regularization terms are important.

In Tables 9 and 10, we show that by tuning the hyperparameters λ_1 and λ_2, we can adjust the strength of the regularization and improve the performance even when assumptions are naturally met. In these experiments, we considered $\lambda_1 = \lambda_2 = \lambda$.

A.8 More recent algorithms

In Table 11, we provide results with and without our regularization on a more recent meta-learning algorithm, Meta-Curvature [28]. In Figure 4, we show the evolution of $\|\mathbf{W}_N\|_F$ and R_σ during training. Similarly to MAML, Meta-Curvature violates both assumptions. We can see that our regularization is still effective for more recent algorithms.

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 8: 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.

Dataset	Episodes	Reproduced	$\lambda = 0$	$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$	$\lambda = 0.0001$
miniImageNet	5-way 1-shot	$49.53 \pm 0.41\%$	$\mathbf{50.29 \pm 0.41}\%$	$49.43 \pm 0.40\%$	$\mathbf{50.19 \pm 0.41}\%$	$\mathbf{50.44 \pm 0.42}\%$	$\mathbf{50.46 \pm 0.42}\%$	$\mathbf{50.45 \pm 0.42}\%$
	5-way 5-shot	$65.10 \pm 0.35\%$	$\mathbf{67.13 \pm 0.34}\%$	$65.71 \pm 0.35\%$	$66.69 \pm 0.36\%$	$66.69 \pm 0.34\%$	$\mathbf{67.2 \pm 0.35}\%$	$\mathbf{67.12 \pm 0.35}\%$
Omniglot	20-way 1-shot	$95.56 \pm 0.10\%$	$\mathbf{95.89 \pm 0.10}\%$	$91.90 \pm 0.14\%$	$94.38 \pm 0.12\%$	$95.60 \pm 0.10\%$	$95.7 \pm 0.10\%$	$95.77 \pm 0.10\%$
	20-way 5-shot	$98.80 \pm 0.04\%$	$98.80 \pm 0.04\%$	$96.40 \pm 0.07\%$	$97.93 \pm 0.05\%$	$98.62 \pm 0.04\%$	$98.76 \pm 0.04\%$	$\mathbf{98.91 \pm 0.03}\%$

Table 9: Ablative study on the strength of the regularization with normalized PROTONET. All accuracy results (in %) are averaged over 2400 test episodes and 4 random seeds and are reported with 95% confidence interval.

Model	Episodes	$\lambda = 0$	$\lambda = 1$	$\lambda = 0.1$	$\lambda = 0.01$	$\lambda = 0.001$	$\lambda = 0.0001$
BASELINE	20-way 1-shot	$86.85 \pm 0.36\%$	$73.65 \pm 0.52\%$	$84.27 \pm 0.37\%$	$\mathbf{87.51 \pm 0.33}\%$	$\mathbf{87.44 \pm 0.33}\%$	$\mathbf{87.58 \pm 0.32}\%$
	20-way 5-shot	$96.95 \pm 0.12\%$	$\mathbf{97.61 \pm 0.11}\%$	$97.10 \pm 0.12\%$	$97.26 \pm 0.11\%$	$97.14 \pm 0.11\%$	$97.23 \pm 0.11\%$
BASELINE++	20-way 1-shot	$\mathbf{82.5 \pm 0.39}\%$	$75.21 \pm 0.47\%$	$81.24 \pm 0.39\%$	$\mathbf{82.58 \pm 0.37}\%$	$\mathbf{82.36 \pm 0.39}\%$	$\mathbf{82.27 \pm 0.39}\%$
	20-way 5-shot	$\mathbf{95.49 \pm 0.15}\%$	$93.25 \pm 0.20\%$	$95.07 \pm 0.15\%$	$\mathbf{95.56 \pm 0.14}\%$	$\mathbf{95.46 \pm 0.14}\%$	$95.32 \pm 0.15\%$

Table 10: Ablative study on the strength of the regularization with BASELINE and BASELINE++. All accuracy results (in %) are averaged over 2400 test episodes and 4 random seeds and are reported with 95% confidence interval.

Episodes	Reproduced	Ratio + Norm
5-way 1-shot	$49.28 \pm 0.84\%$	$49.64 \pm 0.84\%$
5-way 5-shot	$63.74 \pm 0.69\%$	$65.67 \pm 0.69\%$

Table 11: Performance of Meta-Curvature with and without our regularization on miniImageNet. All accuracy results (in %) are averaged over 2400 test episodes and 4 different random seeds and are reported with 95% confidence interval.

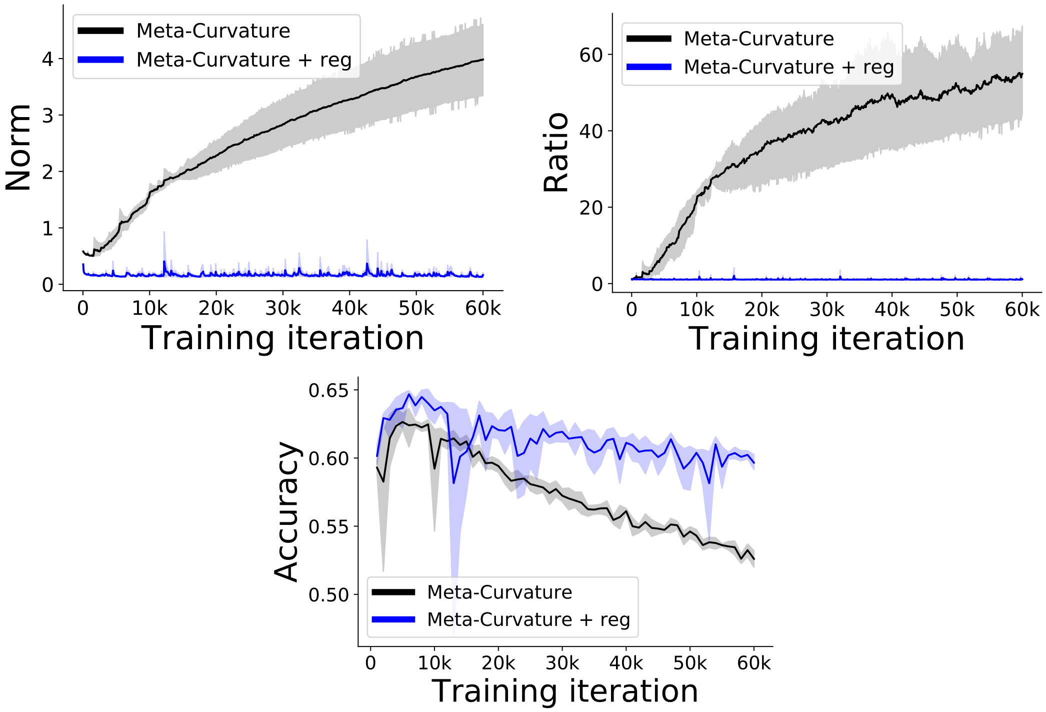

Figure 4: Evolution of $\|\mathbf{W}_N\|_F$ (*top-left*), R_σ (*top-right*) and validation accuracy (*bottom*) when training of Meta-Curvature on miniImageNet 5 with shots. All curves were averaged over 4 different random seeds. $\|\mathbf{W}_N\|_F$ and R_σ increase during training and violate Assumptions 1-2. With our regularization, $\|\mathbf{W}_N\|_F$ and R_σ are constant during training in accordance with theory and we achieve better generalization on the validation set.