Supplementary Material: Meta-Learning via Hypernetworks

A Experimental Setups

A.1 Few-shot Regression

We follow the same experiment described in [8, 1]. We perform few-shot regression on sinusoidal functions with amplitude and phase randomly sampled from [0.1,5.0] and [0,π]. In both the 5-shot and 10-shot setting, our meta model was trained on K simulated test datapoints, after training on K examples. At test time, similarly to [8], we obtain the MSE for one task by testing on 100 datapoints after the K-shot adaptation.

As outlined in [1],we used 1 gradient step with learning rate 0.0001 at training and test time in the inner loop. The outer loop was optimized with one gradient step using ADAM with learning rate 0.001. One training iteration consists of a batch of 25 tasks. The training phase ran for 70000 iterations and the best model was picked by early stopping on a held out validation set.

A.2 Few-shot Classification

We follow the same experimental setup as [1] on the MiniImageNet dataset. This dataset proposed by [26, 27] is separated into 64 training classes, 12 validation classes and 24 test classes. We used the same hyperparameters for the compressed,comparable and overparametrized model. For the 1-shot and 5-shot settings, we used batch sizes of 4 and 2 respectively. In both settings, MH uses gradient steps with learning rate 0.05 in the inner loop and learning rate 0.001 in the outer loop. Both were trained with 6 inner loop gradient steps and tested with 15 gradient steps. The outer loop optimization was done with normal SGD. The best models for 1-shot and 5-shot classification were chosen with early stopping on a validation set and were trained for 100 000 and 150 000 iterations respectively.

A.3 Frozen Features in Inner Loop

All models were trained following the experimental setup described in A.2. For the frozen features setting, we follow [2] and fix the templates as well as the task embedding at test time. The only parameters updated in the inner-loop are the weights of the head. If MH was learning good embeddings, the frozen features experiment would show similar performance as the normal setting one (this is the case for MAML). We can see a drastic difference between the two, showing that the hypernetwork is really helping during the adaptation phase.

B Derivation for NQM model analysis

B.1 Same task used in inner and outer loop

From (17), one can verify that

$$\frac{\partial \tilde{L}}{\partial \alpha_0} = \theta^\top (I - \gamma H \theta\theta^\top) H (I - \gamma \theta\theta^\top H)(\theta\alpha_0 - \epsilon) \tag{20}$$

$$\begin{aligned}\frac{\partial \tilde{L}}{\partial \theta} =& (I - \gamma H \theta\theta^\top) H (I - \gamma \theta\theta^\top H)(\theta\alpha_0 - \epsilon)\alpha_0^\top \\ & - \gamma H [(\theta\alpha_0 - \epsilon)(\theta\alpha_0 - \epsilon)^\top (I - \gamma H\theta\theta^\top) + (I - \gamma\theta\theta^\top H)(\theta\alpha_0 - \epsilon)(\theta\alpha_0 - \epsilon)^\top] H\theta\end{aligned} \tag{21}$$

By taking the expectation of these gradients over the task distribution, since ϵ is sampled from $\mathcal{N}(W^*, \Sigma)$, we get:

$$\mathbb{E}_\epsilon \frac{\partial \tilde{L}}{\partial \alpha_0} = \theta^\top (I - \gamma H \theta\theta^\top) H (I - \gamma \theta\theta^\top H)(\theta\alpha_0 - W^*) \tag{22}$$

$$\begin{aligned}\mathbb{E}_\epsilon \frac{\partial \tilde{L}}{\partial \theta} =& (I - \gamma H \theta\theta^\top) H (I - \gamma \theta\theta^\top H)(\theta\alpha_0 - W^*)\alpha_0^\top \\ & - \gamma H (\theta\alpha_0\alpha_0^\top\theta^\top + \Sigma - \theta\alpha_0 W^{*\top} - W^*\alpha_0^\top\theta^\top)(I - \gamma H\theta\theta^\top) H\theta \\ & - \gamma H (I - \gamma\theta\theta^\top H)(\theta\alpha_0\alpha_0^\top\theta^\top + \Sigma - \theta\alpha_0 W^{*\top} - W^*\alpha_0^\top\theta^\top) H\theta\end{aligned} \tag{23}$$

We can see that the expected gradients (23) and (24) becomes both 0 when $\gamma\theta\theta^\top H = I$.

B.2 Different task used in inner and outer loop

When computing the outer loop loss on a different task than that used in the inner loop, the noise appearing in each of the loss are independent. By denoting by ϵ_1, ϵ_2 the inner and outer loop tasks respectively, equations (21) and (22) become

$$\frac{\partial \tilde{L}}{\partial \alpha_0} = \theta^\top H(I - \gamma\theta^\top\theta^\top H)(\theta\alpha_0 - \epsilon_2 - \gamma\theta\theta^\top H(\theta\alpha_0 - \epsilon_1)) \tag{24}$$

$$\begin{aligned}\frac{\partial \tilde{L}}{\partial \theta} =& H(I - \gamma\theta^\top H\theta\theta^\top H)(\theta\alpha_0 - \epsilon_2 - \gamma\theta\theta^\top H(\theta\alpha_0 - \epsilon_1))\alpha_0^\top \\ &- \gamma H[(\theta\alpha_0 - \epsilon_1)(\theta\alpha_0 - \epsilon_2)^\top - \gamma(\theta\alpha_0 - \epsilon_1)(\theta\alpha_0 - \epsilon_1)^\top H\theta\theta^\top \\ &+ (\theta\alpha_0 - \epsilon_2)(\theta\alpha_0 - \epsilon_1)^\top - \gamma\theta\theta^\top H(\theta\alpha_0 - \epsilon_1)(\theta\alpha_0 - \epsilon_1)^\top]H\theta\end{aligned} \tag{25}$$

Taking the expectation of these gradient over the task distribution results in the following

$$\begin{aligned}\mathbb{E}_\epsilon \frac{\partial \tilde{L}}{\partial \alpha_0} =& \theta^\top H(I - \gamma\theta^\top\theta^\top H)(\theta\alpha_0 - W^* - \gamma\theta\theta^\top H(\theta\alpha_0 - W^*)) \\ =& \theta^\top(I - \gamma H\theta\theta^\top)H(I - \gamma\theta\theta^\top H)(\theta\alpha_0 - W^*)\end{aligned} \tag{26}$$

$$\begin{aligned}\mathbb{E}_\epsilon \frac{\partial \tilde{L}}{\partial \theta} =& (I - \gamma H\theta\theta^\top)H(I - \gamma\theta\theta^\top H)(\theta\alpha_0 - W^*)\alpha_0^\top \\ &- \gamma H[(\theta\alpha_0 - W^*)(\theta\alpha_0 - W^*)^\top(I - \gamma H\theta\theta^\top) + (I - \gamma\theta\theta^\top H)(\theta\alpha_0 - W^*)(\theta\alpha_0 - W^*)^\top]H\theta \\ &+ \gamma^2 H[\theta\theta^\top H(\Sigma - W^*W^{*\top})(\Sigma - W^*W^{*\top})H\theta\theta^\top]H\theta\end{aligned} \tag{27}$$

We can see that taking independent noise in the inner and outer loop result in some covariance terms in (24) to be replaced by $W^*W^{*\top}$, resulting in an expected gradient which does not vanish in general when $\gamma\theta\theta^\top H = I$.