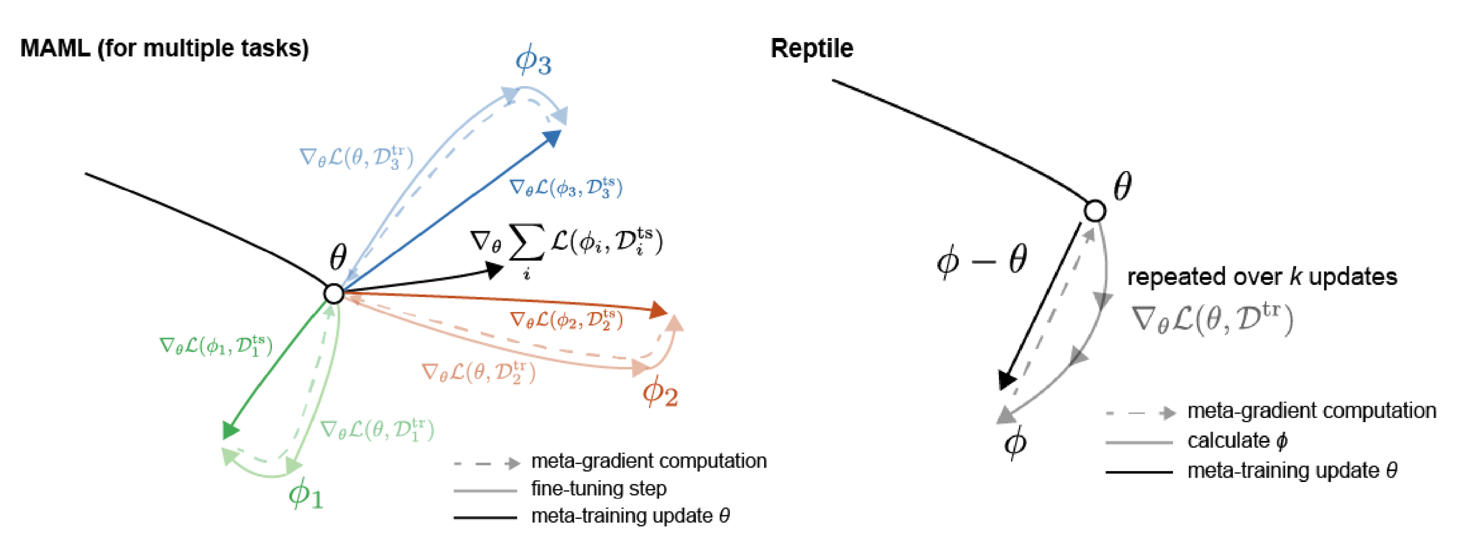
First-Order Approximations for Efficient Meta-Learning

B. King, A. Black, A. J. McLeay

UNIVERSITY OF CAMBRIDGE

Introduction

Following on from the promising performance of the Model Agnostic Meta Learning (MAML) algorithm [1], Reptile was written as a strong approximator for MAML that retains its performance whilst boasting much improved algorithmic complexity. This project seeks to replicate the experiments of the Reptile paper [2] and extend them into more complex contexts.



We can mathematically justify Reptile (and, as it happens, MAML) by thinking in terms of inner products of gradients after *k* inner-loop update steps: [2]

$$g_{\text{Reptile}} = -\frac{1}{\eta}(\phi_k - \theta) = \sum_{i=1}^{k} \mathcal{L}_i'(\phi_i) + \mathcal{L}_1'(\theta) \tag{1}$$

$$\approx \sum_{i=1}^{k} \mathcal{L}_i'(\theta) - \eta \sum_{i=1}^{k} \sum_{j=1}^{i-1} \mathcal{L}_i''(\theta)\mathcal{L}_j'(\theta) \tag{2}$$

$$E_\tau\left[g_{\text{Reptile}}\right] \approx kE_\tau\left[\mathcal{L}_1'(\theta)\right] - \frac{\eta}{4}k(k-1)E_\tau\left[\frac{\partial}{\partial\theta}\langle\mathcal{L}_1'(\theta)|\mathcal{L}_2'(\theta)\rangle\right] \tag{3}$$

Inner-Loop Optimiser Comparisons

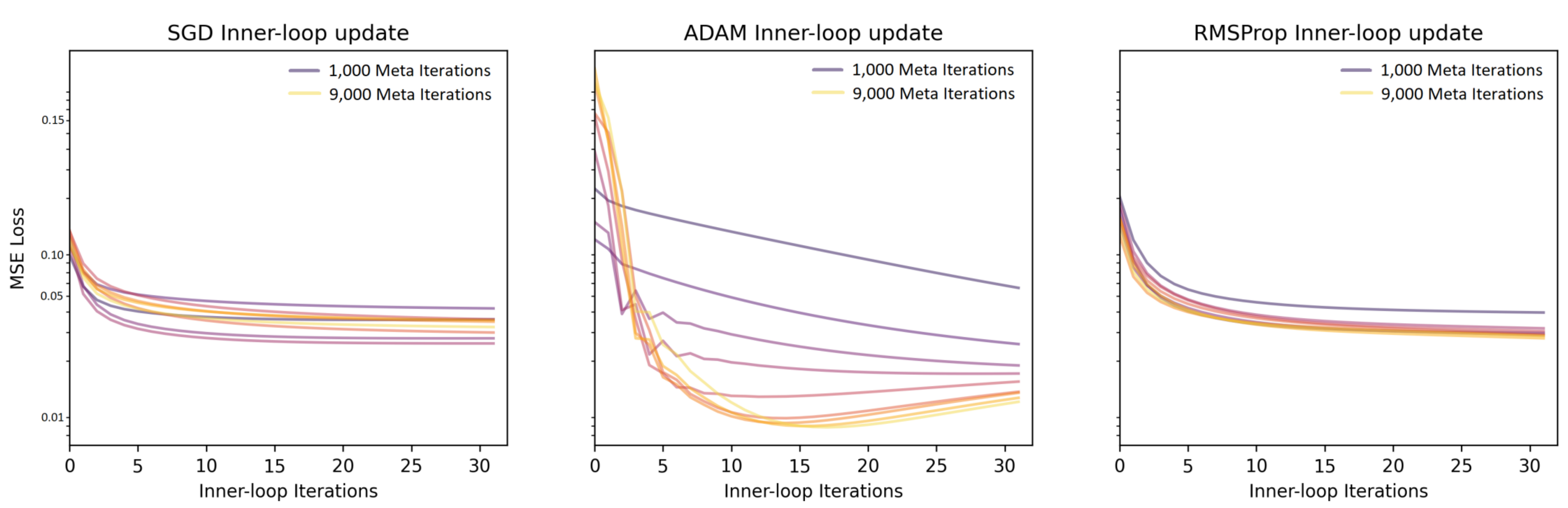


Figure 2: Using different optimisers for the rapid-learning updates on the sinusoid regression experiment from the original paper [1]. This concerns how φ_k is found in (1). Whilst ADAM improved learning, like [1], including momentum made learning more unstable and Reptile performed best when the momentum coefficient was near zero. Lastly, overfitting on the test task can be seen after about 20 ADAM inner-loop iterations when trained for many meta iterations.

Extensions to Few-Shot Regression

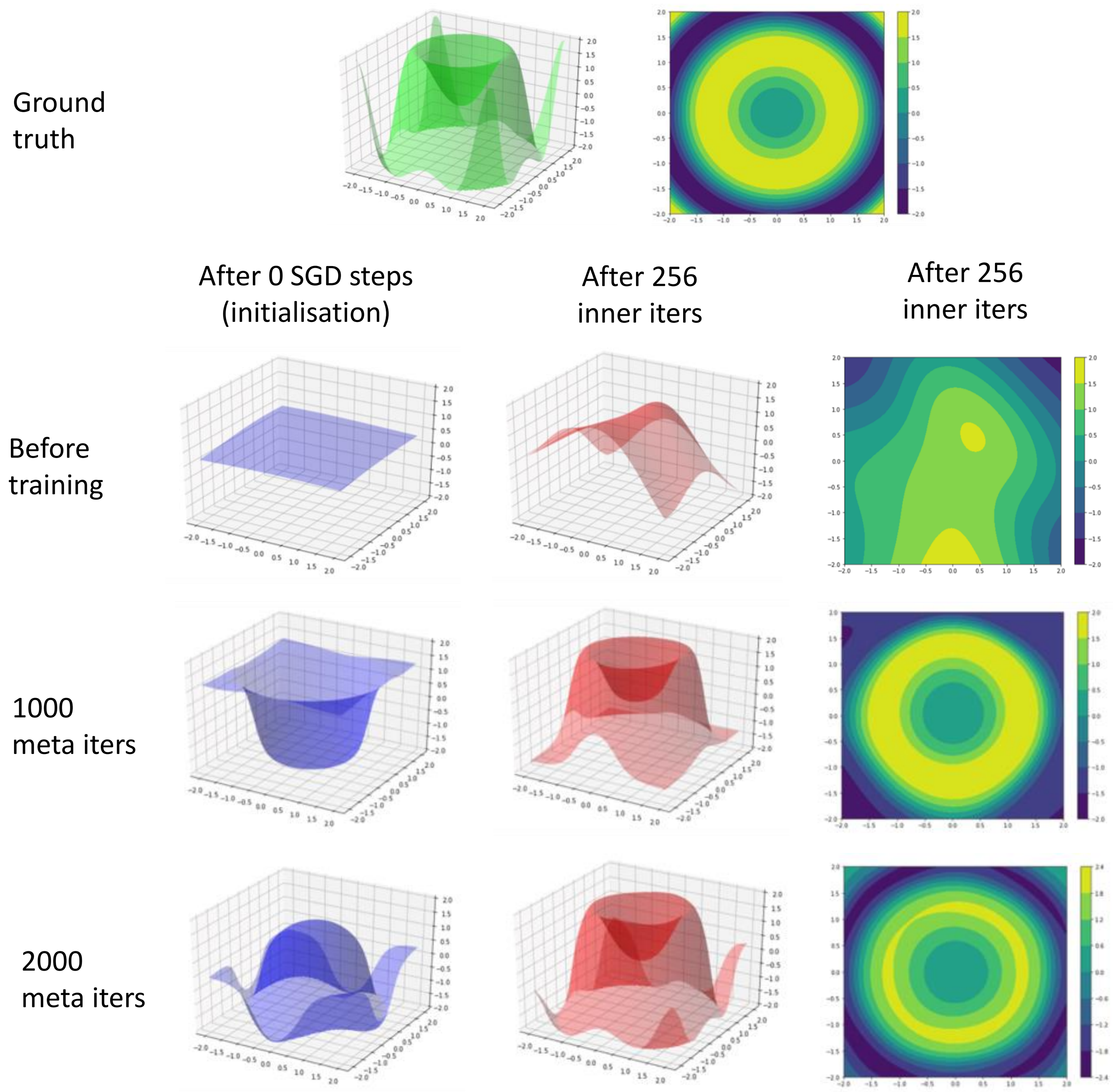


***Figure 3:** Results from a 2D version of the 1D few-shot regression sine wave example in the Reptile paper. The input space is scaled from 50 to 10K points, and the minibatch size is scaled from 10 to 20. The amplitude is fixed at 2.0 and the phase is varied from 0 to 2π. An SGD optimiser is used. After 1000 meta iterations an initialisation is learned that enables the 'bowl' shape of the function to be modelled. After 2000 epochs the learned initialisation enables the 'tails' of the function to also be modelled.*

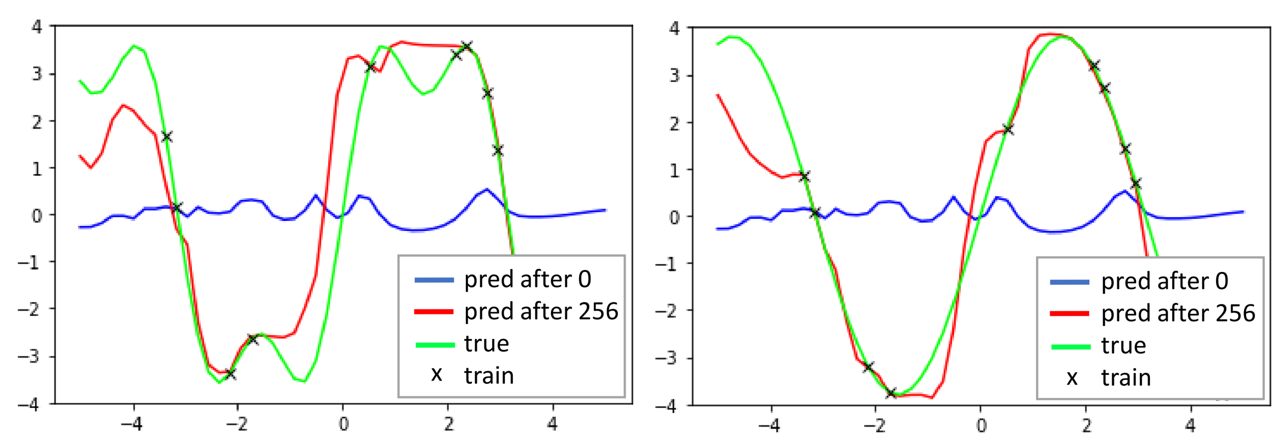


***Figure 4:** Results from adding an additional Fourier series function to the 1D few-shot regression sine wave example. The amplitude is randomly sampled from 0.1 to 5, the phase from 0 to 2π, and the function from the sine and Fourier series functions. The model learns an initialisation (in blue) that captures key differences between the two functions after 256 inner steps, despite the size of the network not being increased.*

References

[1] Finn, C. et al. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. International Conference on Machine Learning, 1126–1135
[2] Nichol, A., et al. (2018). On first-order meta-learning algorithms. ArXiv Preprint ArXiv:1803.02999.
[3] Zhao, B. (2021). Basics of few-shot learning with optimization-based meta-learning. In boyangzhao.github.io. https://boyangzhao.github.io/posts/few_shot_learning

Image Classification: Results and Extensions

Mini-Imagenet Accuracy			
Algorithm	1-shot 5-way	5-shot 5-way	10-shot 5-way
Reptile	51.6%	67.4%	72.6%
Reptile + Pretrain	49.4%	-	-
Reptile [2]	50.0%	66.0%	-

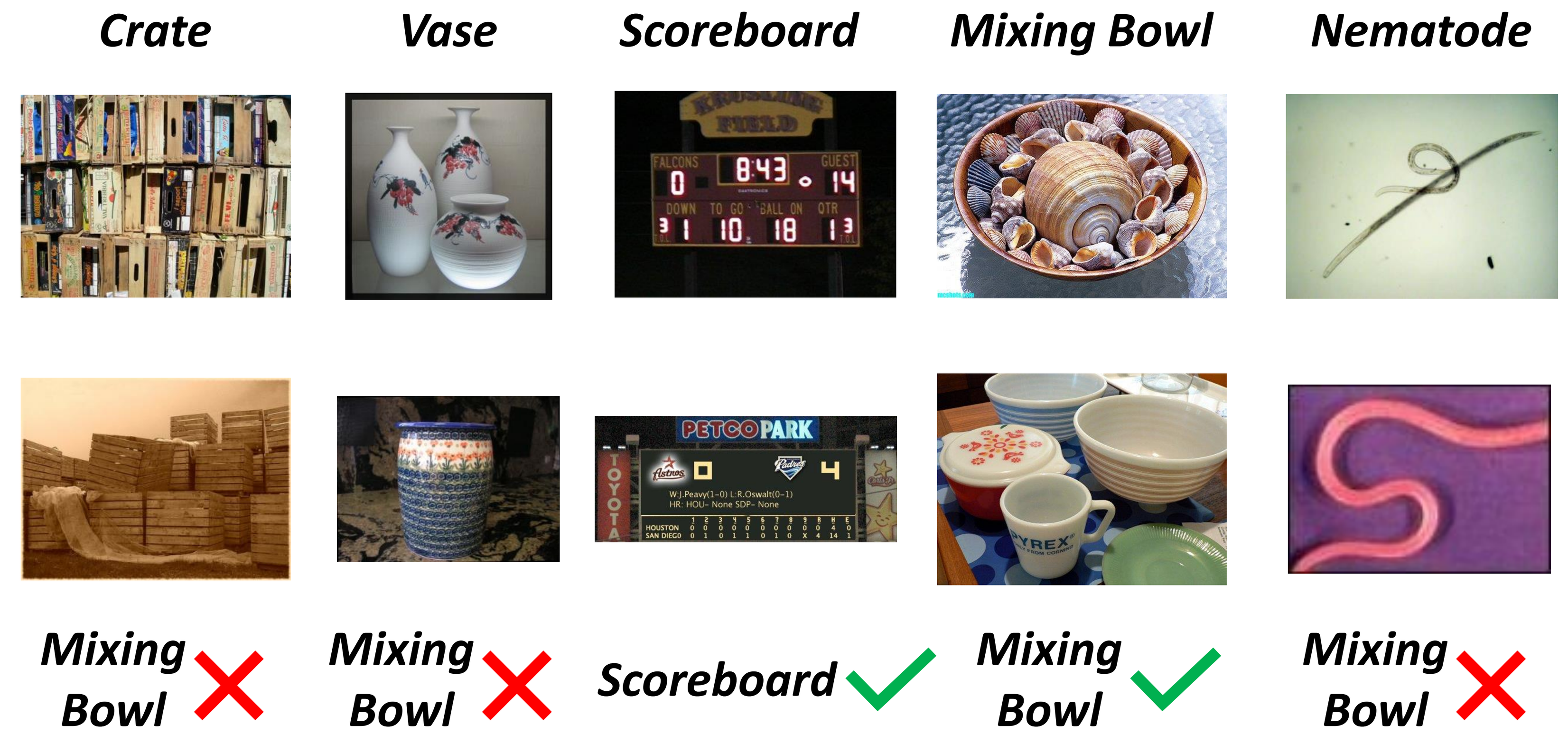
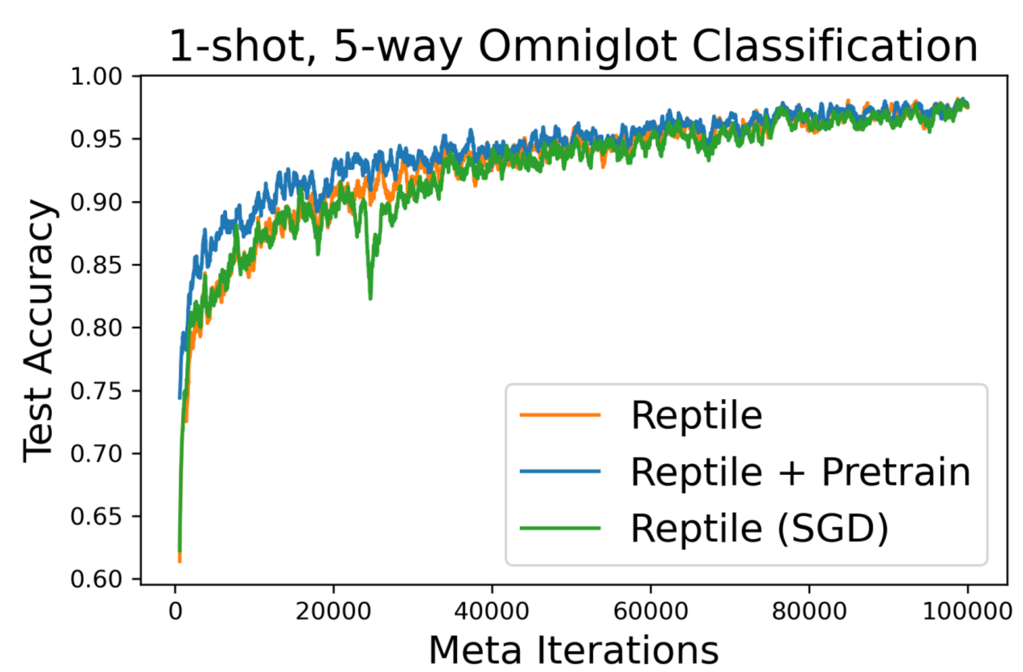


***Figure 5:** Example 5-way classifier: 1-shot images (top row) and test images (bottom row).*



Omniglot Accuracy	
Algorithm	1-shot 5-way
Reptile	97.7%
Reptile + Pretrain	97.7%
Reptile (SGD)	97.4%
Reptile [2]	97.7%
FOMAML	87.5%
FOMAML [1]	98.3%

***Figure 6:** Initializing weights from a pretrained classifier provides Reptile with an early, but unsustained, training advantage. Adopting SGD (instead of ADAM) for the inner-loop leads to momentarily suboptimal exploration but ultimately comparable performance.*

Conclusions

1) On 1D sinusoid regression, an ADAM inner-loop optimiser achieves lowest MSE but with less stability due to momentum.
2) Reptile learns few-shot approximations to 2D sine waves as well as mixtures of 1D function families (Fourier series and sine waves).
3) Reptile performs comparably to MAML [1] on few-shot image classification. Pretraining elevates Reptile performance during early training iterations, but does not produce a sustained advantage. However we did not extensively tune pretraining hyperparameters and further research could develop a more rigorous pretraining protocol.