On the Subspace Structure of Gradient-Based Meta-Learning

Gustaf Tegnér [*1] **Alfredo Reichlin** [*1] **Hang Yin** [1] **Mårten Björkman** [1] **Danica Kragic** [1]

Abstract

In this work we provide an analysis of the distribution of the post-adaptation parameters of Gradient-Based Meta-Learning (GBML) methods. Previous work has noticed how, for the case of image-classification, this adaption only takes place on the last layers of the network. We propose the more general notion that parameters are updated over a low-dimensional *subspace* of the same dimensionality as the task-space and show that this holds for regression as well. Furthermore, the induced subspace structure provides a method to estimate the intrinsic dimension of the space of tasks of common few-shot learning datasets.

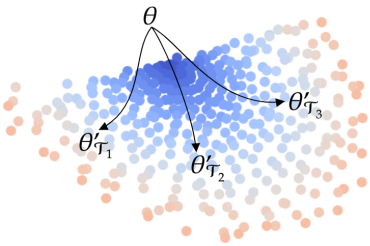

Figure 1: The space of task-adapted parameters for a sine regression task embedded in two-dimensional space. The polar-coordinate structure of the task is preserved in the space of task-adapted parameters.

1. Introduction

Humans possess an innate ability to draw upon past experiences to efficiently solve new problems. From being exposed to multiple environments, we are able to learn general concepts which can be leveraged to solve tasks in novel domains. This concept of *transfer* between problem domains is a core tenant of human intelligence. To endow this ability on intelligent learning systems, one can adhere to Multi-Task Learning (MTL), a paradigm that enables learning over multiple frameworks concurrently. A particular instance of MTL is Meta-Learning. These set of algorithms are designed to quickly learn and adapt to novel tasks. In general, this is done by exposing the learner to several tasks concurrently, allowing it to learn a common source of information between the tasks, as well as their differences. This ability to discriminate between tasks is at the core of meta learning models. A single task can be arbitrarily complex while a set of related tasks, instead, can form a well defined structure. In most cases, such a structure may be defined by only a small number of variables. That is, the space of related tasks can generally be assumed to be equipped with a very low *intrinsic* dimension. Exploiting effectively this structure enables the meta learner to make an efficient use of its acquired knowledge. Adapting to novel tasks then essentially amounts to finding and exploiting these intrinsic factors of variation in the task space.

To give an example of such a structure, imagine the regression problem of meta-learning a collection of tasks $\{\mathcal{T}_i\}_{i=1}^N$ defined as the sum of sine waves of varying amplitude and phase. Learning a single task $\mathcal{T}_i$ can be made arbitrarily complex by considering additional sine waves. However, when learning multiple tasks in conjunction it suffices to identify the common form of the tasks and then learning only the relatively simple task structure induced by the varying task parameters. Figure 1 depicts such a structure learnt by GBML. As shown, the task structure is *embedded* in the task-adapted parameters themselves.

In this paper we aim to study more explicitly this task structure through the lens of Gradient-Based Meta-Learning (GBML) (Finn et al., 2017). This particular family of meta learning methods adapt to a new task by learning how to generate a set of optimal parameters specific to that task. Recent work has shown how, in the context of classification, GBML requires to update only a subset of its parameters (Raghu et al., 2019; Oh et al., 2020). We expand on this perspective by demonstrating that GBML implicitly learns a subspace that reflects the inherent structure of the tasks for both classification and regression. To investigate this phenomena, we borrow notions from dimensionality estimation and study the distribution of the parameters found by GBML

[*]Equal contribution [1]Division of Robotics, Perception and Learning, EECS, KTH Royal Institute of Technology, Stockholm, Sweden. Correspondence to: Gustaf Tegnér <gustafte@kth.se>, Alfredo Reichlin <alfrei@kth.se>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

after adaptation for each one of the tasks considered. By representing a task as the parameters of a task-adapted neural network, we find a high-dimensional representation of it. We can then leverage representation learning techniques such as dimensionality reduction to estimate the intrinsic dimensionality of the task space. We empirically show that GBML finds a subspace of a dimensionality that is optimal for the set of tasks at hand. This, in turn, provides us with an insight into the inherent subspace structure of common datasets.

Our contributions are as follows:

- Empirical evidence that GBML methods naturally tend to a solution where the adapted parameters lie on a subspace of the smallest dimensionality possible given the meta-dataset.
- A method to analyze the dimensionality of the task-space of common datasets.

2. Related Work

Meta-Learning. The general idea of meta-learning tackles the setting of learning a new task given only a small set of labeled examples. To achieve this, one of the most prominent approaches is model agnostic meta learning (MAML) (Finn et al., 2017). It learns a common parameter initialization across tasks that lies close to the optimal solution. Extensions to MAML couple the optimization procedure with a learned set of parameters to increase its expressivity (Lee & Choi, 2018; Flennerhag et al., 2019; Park & Oliva, 2019).

Few-shot Classification. GBML for image classification has been extensively studied. (Raghu et al., 2019) showed that GBML models naturally converge to the solution of adapting only the parameters of the last layer of an image classifier for N-way K-shot classification. They suggest that the test performances are correlated with the quality of the representation learner represented by the first part of the network. This has been confirmed by (Oh et al., 2020) where they also propose to freeze the adaptation of the last layer to incentivize the meta learner to find a solution that makes use of more layers. Further analysis have compared the behavior of MAML with a contrastive learning framework where the learned features exhibit invariance to the samples used in adaptation (Kao et al., 2021; Goldblum et al., 2020; Li et al., 2018a).

Dimensionality Reduction A common accepted notion in machine learning is that high-dimensional data lies on a submanifold of a much lower dimension than the dimension of the ambient space. Dimensionality reduction techniques can be utilized to find this intrinsic dimensionality of data. Examples of such techniques are Principal Component Analysis (PCA) (Pearson, 1901) and Isomap (Tenenbaum et al., 2000). Isomap is a non-linear dimensionality reduction technique that aims to preserve geodesics of the original space in a space of dimension k. It works by minimizing the reconstruction error which is the difference between distances in the original space and embedded space w.r.t to the Isomap kernel. Previous work has utilized such techniques in an attempt to estimate the true intrinsic dimensionality of images (Pope et al., 2021).

Subspace Structure. Several works have investigated the geometry of the loss-landscape of neural networks (Li et al., 2018b) (Lengyel et al., 2020). In particular, it has been shown that there exists paths in the parameter space on which the loss remains low (Wortsman et al., 2021) (Garipov et al., 2018) (Frankle, 2020). Other type of structure, such as hyper-planes has been explored in (Lengyel et al., 2020). Recently, subspace structure has been considered in online reinforcement-learning by learning a subspace of policies which enables better generalization during testing (Gaya et al., 2021). In this work, we extend this analysis to the meta-learning setting by identifying sub-manifold structure in the learnt parameter space.

3. Preliminaries

The general formulation of gradient-based meta-learning follows from (Finn et al., 2017). We consider a set of tasks $\mathcal{T}_i$ sampled from a distribution $p(\mathcal{T})$ together with a parameterized function f_θ. Given a new task $\mathcal{T}_i$, GBML seeks to find a set of parameters θ such that applying one gradient step on the loss computed on a few data-points (D^S) of the task results in the optimal set of parameters $\theta'_{\mathcal{T}_i}$:

$$\theta'_{\mathcal{T}_i} = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i} \left(f_\theta, D^S \right) \tag{1}$$

To find such θ, GBML optimizes the expectation of the loss, computed on the rest of the task's data (D^Q), of the model f_θ over all the tasks after adaptation:

$$\mathcal{L}_{\text{GBML}} = \mathbb{E}_{p(\mathcal{T})} \left[\mathcal{L}_{\mathcal{T}} \left(f_{\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}}(f_\theta, D^S)}, D^Q \right) \right]. \tag{2}$$

Let $S_{\mathcal{T}} = \text{supp}(\mathcal{T})$ be the support of the distribution of tasks. GBML essentially learns a differentiable map $M : S_{\mathcal{T}} \to \Theta$ from the support to the space of parameters of a neural network. Let $\Theta_{\mathcal{T}} = M(S_{\mathcal{T}})$ denote the image of this map, that is the space $\Theta_{\mathcal{T}} \subseteq \Theta$ induced by the task-adapted parameters found through GBML. Since M is differentiable, we have that $\dim(\Theta_{\mathcal{T}}) \leq \dim(S_{\mathcal{T}})$. Assuming the tasks carry no redundancy, the dimensions are expected to be equal. The intrinsic complexity of the task can then be reduced to the study of the properties of $\Theta_{\mathcal{T}}$.

4. Evaluation Method

We aim to investigate the geometry of the parametric model space of gradient-based meta-learning methods for different meta-problems. We hypothesize that since the space of tasks holds an intrinsic dimension d, our task-adapted parametric model space should be embedded in low-dimensional space of the same dimension. To investigate this hypothesis, we make use of dimensionality reduction techniques. Given a set of task-adapted parameters $\theta_{\mathcal{T}_i} \in \Theta = \mathbb{R}^D$, we make the assumption that $\theta_{\mathcal{T}_i}$ is sampled from a lower-dimensional manifold $\mathcal{M} \subseteq \mathbb{R}^D$ of intrinsic dimension $\dim(\mathcal{M}) = d \ll D$. A common way to estimate the intrinsic dimensionality is to reduce the dimensions of $\theta_{\mathcal{T}_i}$ to a varying number of dimensions k. The intrinsic dimensionality would then correspond to the smallest k which still preserves the features of the original data space, measured by the reconstruction error. We utilize Principal Component Analysis (PCA) (Pearson, 1901) and Isomap (Tenenbaum et al., 2000) to estimate this intrinsic dimensionality.

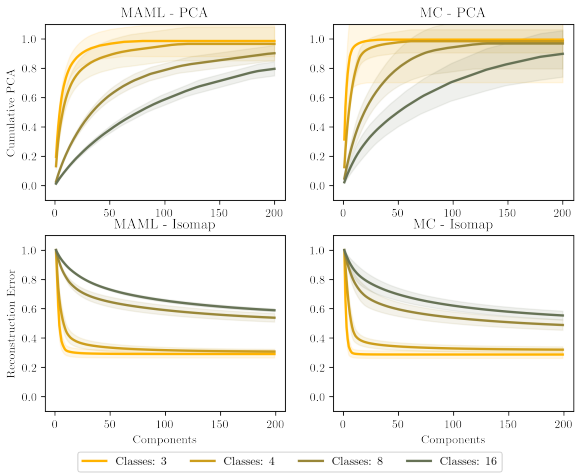

Figure 2: PCA and Isomap results for MAML and Meta-Curvature on the toy classification experiment.

5. Experiments

We conduct a number of experiments to analyze the parameter space that emerges when applying GBML to a number of regression and classification problems. For our experiments we used Model Agnostic Meta Learning (MAML) (Finn et al., 2017), one the most prominent members of the GBML family. To confirm that a structure of the parameters emerges independently of the expressiveness of the learner, we also used in the experiments Meta-Curvature (MC) (Park & Oliva, 2019). MC relies on learning a matrix G which conditions the gradients by $G\nabla\mathcal{L}_\theta$. We perform our analysis on a toy classification and regression task and furthermore investigate few-shot classification on miniImagenet and Omniglot datasets. For the toy experiments we consider a two layer neural network with ReLU activations and 40 hidden units. For miniImagenet and Omniglot, we use a 4 layer convolutional neural network with batch-normalization, ReLU activations and max-pooling. We train the models for 100 epochs with a meta-batch size of 32. For Isomap, we use a neighborhood size of 20.

5.1. Classification Task

In few-shot classification, it has been noted that GBML algorithms find a solution where only the last layers of the meta-learner gets updated during adaptation (Raghu et al., 2019). In our first experiment, we expand upon this insight by studying the entire space of task-adapted parameters for a toy N-way K-shot classification problem. To create this task, we construct a set of prototype classes $x_p \in \mathbb{R}^D$ that are equally spaced on a D-dimensional grid. We sample uniformly a subset of N prototypes, each assigned to a distinct class $c \in [1, N]$. From each prototype, we now sample a set of K examples $x_p^i \sim \mathcal{N}(x_p, \sigma)$. Adapting to a new task now involves changing the decision boundaries of a multinomial classification problem. We consider 5 different datasets with the number of classes $N \in \{3, 4, 8, 16\}$. To show that our method is architecture agnostic, we train our model by varying the number of hidden units in the last layer by $\{32, 64, 128, 256\}$.

Since we are interested in identifying the intrinsic dimension k from which no further increase in performance can be found, we normalize the Isomap scores by dividing them by the first (largest) value. The plots are shown in Figure 2. As the number of classes increase, the number of dimensions required for adaptation increase accordingly. For PCA, it can be seen that for lower number of classes, more weight is put on a single principal component. This is reflected in the Isomap results where using a lower number of classes correspond to a steeper curve and thus lower dimension. Furthermore, the results are consistent with different architectures as the error bars remain intact. This shows that subspace structure is agnostic to the model used and is entirely dependent on the task itself. We can note that for two classes, a one-dimensional embedded task space seems to be the predominant result. In the appendix, Figure 6 confirms that only a small subset of parameters in the last layers are updated for the binary classification problem while for other classes we note a larger change. While Figure 6 re-confirms the results of (Raghu et al., 2019), we raise the point that it is not only what *subset* of parameters are updated, but what *subspace* they lie on. Although a subset of parameters alters with choice of architecture, the subspace is instead only dependent on the complexity of the problem itself.

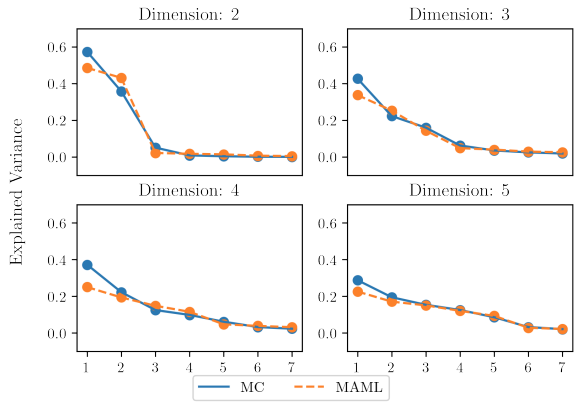

Figure 3: Explained variance of the first k principal components of the task-adapted parameters of the sine tasks. PCA correctly identifies that the number of significant components is correlated with the intrinsic dimensionality of the task.

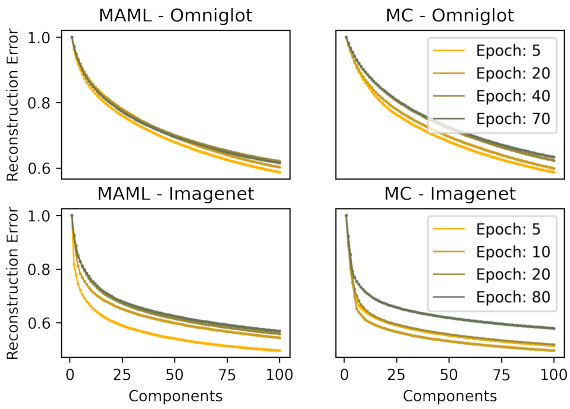

Figure 4: Estimated dimensionality from Isomap of Omniglot and miniImagenet.

5.2. Regression Task

Next we investigate if the same results holds true for regression tasks as well. We construct a regression task by considering the sum of sine waves with different amplitudes. To put it explicitly, for each task we sample a set of amplitudes $A_k \sim \mathcal{U}([0.1, 5.0])$ for $k = 1, \ldots, N$ and construct a regression task as the sum:

$$y = \sum_{k=1}^{N} A_k \sin(kx). \tag{3}$$

Here, the number of sines N defines the intrinsic dimension of the task space as the tasks vary only by the parameters A_k. For our experiments, we consider $N \in \{2, 3, 4, 5\}$. We investigate the explained variance of PCA to compute an estimate of the dimensionality of the parametric model space. The results for PCA are shown in Figure 3 while the ones for Isomap are included in the appendix (Figure 7). From the figure, the strongest principal components coincide with the intrinsic dimensionality of the task. The effect declines as the complexity of the task increases, since there is an increased ambiguity for the meta-learner to learn complex functions with a limited amount of support data. This confirms our hypothesis that the intrinsic dimensionality of the task-space is preserved in the space of parameters induced by the meta-learner.

5.3. Analysis of Common Datasets

We now turn to a more complex setting of estimating the intrinsic dimensionality of common classification tasks. We consider an N-way, K-shot few shot learning setting with data from miniImagenet and Omniglot. For both of these tasks we let $N = 5$ and $K = 5$ and train our model for 100 epochs. For the architecture, we consider the convolutional architecture specified in (Vinyals et al., 2016). Figure 4 depicts the reconstruction error of Isomap for MAML and Meta-Curvature for both the miniImagenet and Omniglot tasks. The dimensionality reduction is performed on different epochs throughout the training. The results for miniImagenet show both a higher variability during training and a seemingly lower estimated dimensionality than Omniglot which is reflected in their performance on the test set as well. In the appendix we elaborate upon this observation.

5.4. Reconstruction Experiment

To further substantiate our findings, we perform an experiment involving reconstructing the task from the task-adapted parameters. We perform dimensionality reduction using PCA on the task-adapted parameters to find a representation $z \in \mathcal{Z}$ of reduced dimension k. From this we train a learner $f : \mathcal{Z} \to \mathcal{T}$ in a supervised manner to reconstruct the task from the low-dimensional embedding. For the sine-experiment, we attempt to regress the amplitudes defined in Equation 3. For the classification experiments, we do not have access to the ground-truth task-parameters. In this case, we consider classifying an image from a specific task given the embedding of the learnt parameters for that task. We perform the experiments for varying dimension k and evaluate the results by considering the regression or classification performance. The results for the sine experiment are shown in Figure 5. As can be seen, the performance does not increase further as you increase the embedding size beyond the intrinsic dimension of the task. We evaluate the performance of $3, 8$ and 16 classes for the toy-classification experiment. The results can be seen in Figure 10 in the Appendix. However, we argue that the discrete nature of classification tasks poses a limit in the analysis of the dimensionality of the manifold.

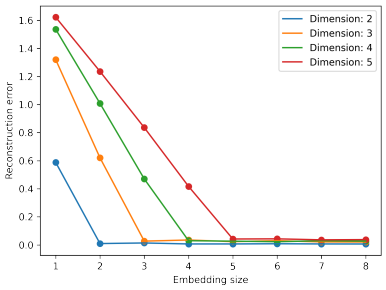

Figure 5: Regressing the task-parameters of the sinusoid task from the PCA embedding of the adapted parameters. The embedding size required to reach a low error is exactly the intrinsic dimension of the task.

6. Conclusion and Future Work

In this work we proposed a method for analyzing the intrinsic dimensionality of the task-space learnt by gradient-based meta-learning methods. We provided an empirical analysis that reveals that GBML learn a low-dimensional subspace on which it performs adaptation on. Furthermore, we related this observation to the generalization performance of the model. We believe our analysis adds a valuable contribution to recent work that employ subspace structure in GBML. A possible future line of work is applying our method on a larger set of models and datasets in order to gain further insight. As a second possible direction, the space of parameters can be studied using additional geometric techniques such as topological data analysis. In our experiments we observe that the dimension of the subspace changes during training and could possibly be correlated with performance. This opens up to future directions of research of further exploring this correlation. We hypothesize that if the dimensionality of the subspace is known a priori, developing a regularizer that enforces this could possibly improve performance for GBML methods.

7. Acknowledgements

We thank Giovanni Luca Marchetti, Vladislav Polianskii and Marco Moletta for useful discussions. This work has been supported by the European Research Council (BIRD: 884807), Swedish Research Council and Knut and Alice Wallenberg Foundation and H2020 CANOPIES.

References

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Flennerhag, S., Rusu, A. A., Pascanu, R., Visin, F., Yin, H., and Hadsell, R. Meta-learning with warped gradient descent. *arXiv preprint arXiv:1909.00025*, 2019.

Frankle, J. Revisiting" qualitatively characterizing neural network optimization problems". *arXiv preprint arXiv:2012.06898*, 2020.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Gaya, J.-B., Soulier, L., and Denoyer, L. Learning a subspace of policies for online adaptation in reinforcement learning. *arXiv preprint arXiv:2110.05169*, 2021.

Goldblum, M., Reich, S., Fowl, L., Ni, R., Cherepanova, V., and Goldstein, T. Unraveling meta-learning: Understanding feature representations for few-shot tasks. In *International Conference on Machine Learning*, pp. 3607–3616. PMLR, 2020.

Kao, C.-H., Chiu, W.-C., and Chen, P.-Y. Maml is a noisy contrastive learner. *arXiv preprint arXiv:2106.15367*, 2021.

Lee, Y. and Choi, S. Gradient-based meta-learning with learned layerwise metric and subspace. In *International Conference on Machine Learning*, pp. 2927–2936. PMLR, 2018.

Lengyel, D., Petangoda, J., Falk, I., Highnam, K., Lazarou, M., Kolbeinsson, A., Deisenroth, M. P., and Jennings, N. R. Genni: Visualising the geometry of equivalences for neural network identifiability. *arXiv preprint arXiv:2011.07407*, 2020.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018a.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018b.

Oh, J., Yoo, H., Kim, C., and Yun, S.-Y. Boil: Towards representation change for few-shot learning. *arXiv preprint arXiv:2008.08882*, 2020.

Park, E. and Oliva, J. B. Meta-curvature. *Advances in Neural Information Processing Systems*, 32, 2019.

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Wortsman, M., Horton, M. C., Guestrin, C., Farhadi, A., and Rastegari, M. Learning neural network subspaces. In *International Conference on Machine Learning*, pp. 11217–11227. PMLR, 2021.

A. Appendix

A.1. Classification

In figure 6, we empirically investigate the distribution of how the parameters change with different tasks. We take all the tasks in the test set and compute the mean absolute difference between all the parameters. This gives an indication of what parameters change between tasks. As can be seen in the figure, for classification, mostly the parameters in the later layers get updated which confirms the results of (Raghu et al., 2019).

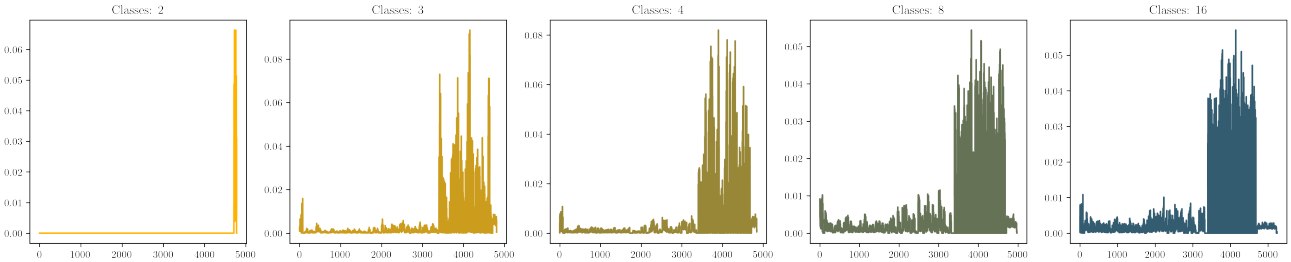

Figure 6: Mean absolute value of difference between parameters in the toy classification problem. The change in parameters is concentrated to the later layers. We note that for $N = 2$, only a small set of parameters are altered during adaptation.

A.2. Sine-Experiment

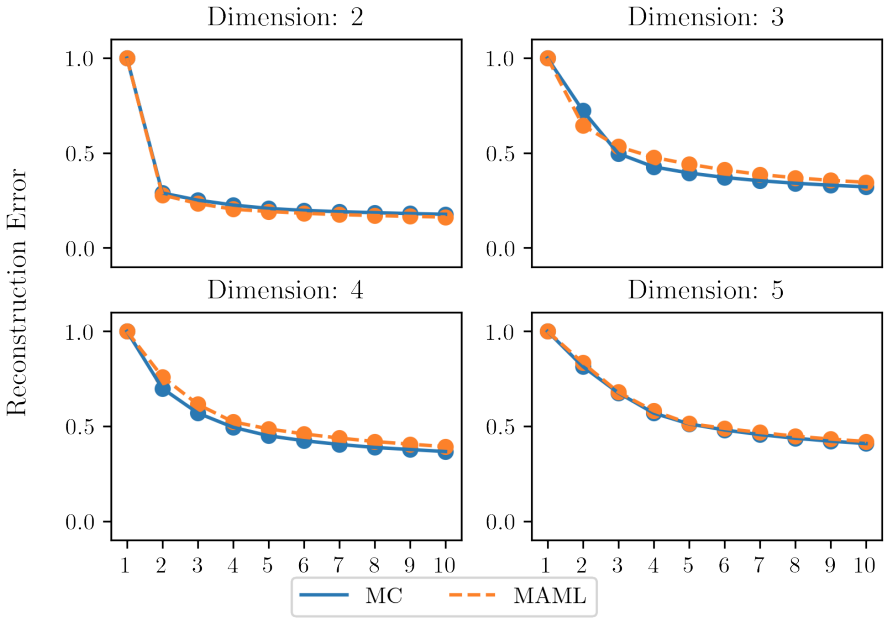

Figure 7: Isomap results for the sine experiment.

A.3. Omniglot

We also investigate few-shot classification on Omniglot. The results are depicted in figure 8.

A.4. miniImagenet

In Figure 9, we provide a comparison between the estimated dimension and the test-set performance of the models at different stages during the training. The estimated dimensionality varies during training with MAML seemingly starting

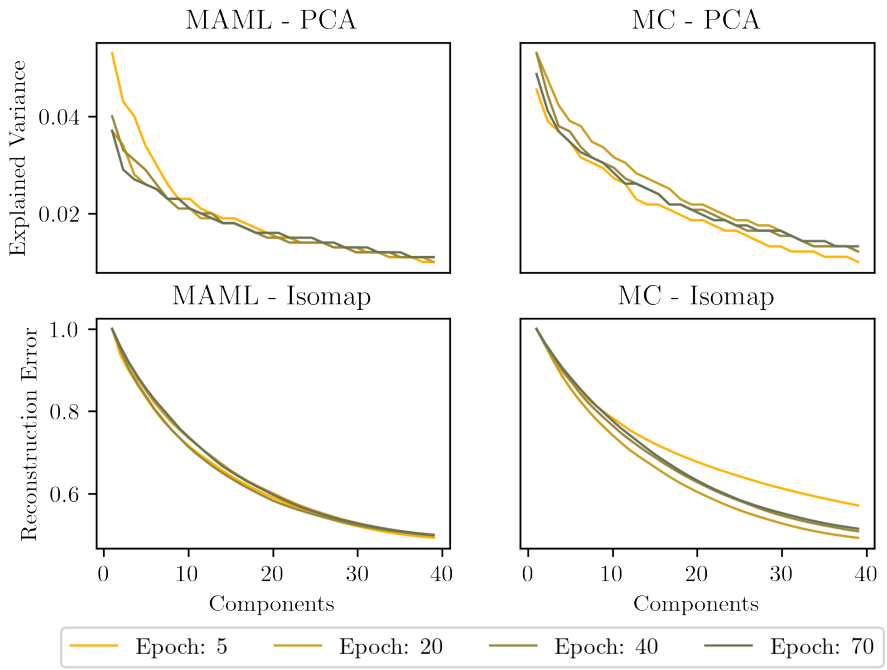

Figure 8: Omniglot results for PCA and Isomap.

off with a low-dimensional parameter space which grows larger in the first epochs then settles down again as performance increases. For meta-curvature, we can observe a clear overfitting on the test set with performance peaking at around epoch 10. This is reflected in the dimensionality estimation as Isomap tends to estimate a lower dimension for later epochs.

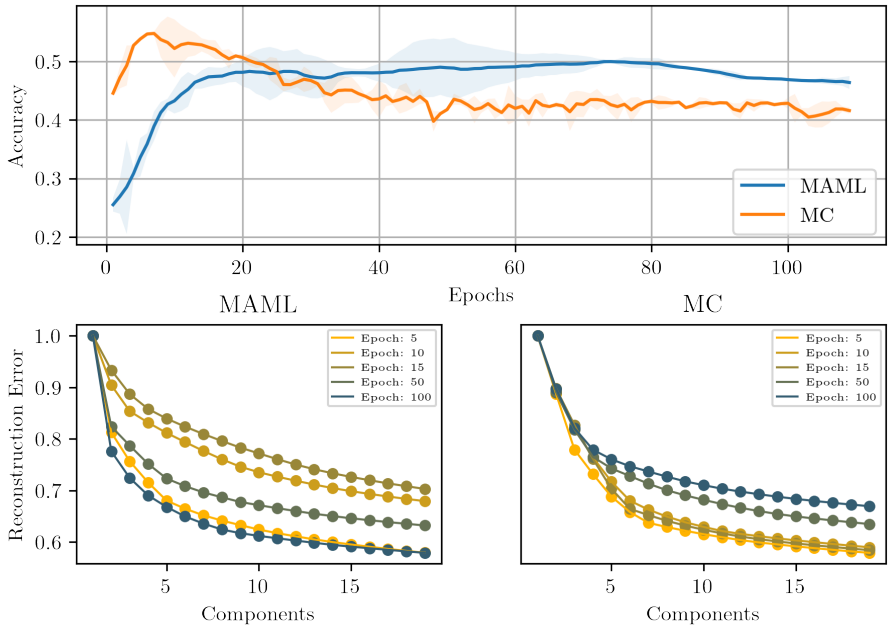

Figure 9: Performance of MAML and Meta-Curvature on a test-set of miniImagenet. Meta-Curvature begins to overfit after epoch 15 while MAML shows increasing performance.

A.5. Reconstruction Experiment - Toy Classification

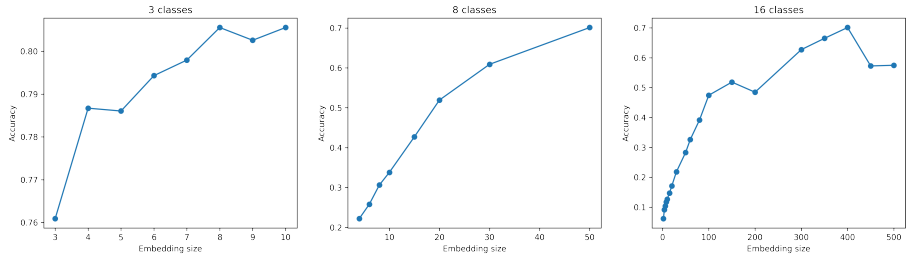

Figure 10: Accuracy of a trained classifier compared to embedding size for the toy-classification experiment.