Addressing Catastrophic Forgetting in Few-Shot Problems

Pauching Yap [1]   **Hippolyt Ritter** [1]   **David Barber** [1 2]

Abstract

Neural networks are known to suffer from catastrophic forgetting when trained on sequential datasets. While there have been numerous attempts to solve this problem in large-scale supervised classification, little has been done to overcome catastrophic forgetting in few-shot classification problems. We demonstrate that the popular gradient-based model-agnostic meta-learning algorithm (MAML) indeed suffers from catastrophic forgetting and introduce a Bayesian online meta-learning framework that tackles this problem. Our framework utilises Bayesian online learning and meta-learning along with Laplace approximation and variational inference to overcome catastrophic forgetting in few-shot classification problems. The experimental evaluations demonstrate that our framework can effectively achieve this goal in comparison with various baselines. As an additional utility, we also demonstrate empirically that our framework is capable of meta-learning on sequentially arriving few-shot tasks from a stationary task distribution.

1. Introduction

Few-shot classification (Miller et al., 2000; Li et al., 2004; Lake et al., 2011) focuses on learning to adapt to unseen classes (known as **novel classes**) with very few labelled examples from each class. Recent works show that meta-learning provides promising approaches to few-shot classification problems (Santoro et al., 2016; Finn et al., 2017; Ravi & Larochelle, 2017). Meta-learning or learning-to-learn (Schmidhuber, 1987; Thrun & Pratt, 1998) takes the learning process a level deeper – instead of learning from the labelled examples in the training classes (known as **base classes**), meta-learning learns the example-learning process. The training process in meta-learning that utilises the base

classes is called the **meta-training** stage, and the evaluation process that reports the few-shot performance on the novel classes is known as the **meta-evaluation** stage.

Despite being a promising solution to few-shot classification problems, meta-learning methods suffer from a limitation where a meta-learned model loses its few-shot classification ability on previous datasets as new ones arrive subsequently for meta-training. Some popular examples of the few-shot classification datasets are Omniglot (Lake et al., 2011), CIFAR-FS (Bertinetto et al., 2019) and *mini*ImageNet (Vinyals et al., 2016). A meta-learned model is restricted to perform few-shot classification on a specific dataset, in the sense that the base and novel classes have to originate from the same dataset distribution. The current practice to few-shot classify the novel classes from different datasets is to meta-learn a model for each dataset separately (Snell et al., 2017; Vinyals et al., 2016; Bertinetto et al., 2019). This paper considers meta-learning a single model for few-shot classification on multiple datasets with evident distributional shift that arrive sequentially for meta-training. Figure 1 gives an example of the sequential few-shot classification setting of concern.

Figure 1. An example of the sequential few-shot classification problems with evident dataset distributional shift: Omniglot → CIFAR-FS → *mini*ImageNet.

We introduce a recursive framework to train a model that is applicable to a broader scope of few-shot classification datasets by overcoming catastrophic forgetting. Bayesian online learning (BOL) (Opper, 1998) provides a principled framework for the posterior of the model parameters, while model-agnostic meta-learning (MAML) (Finn et al., 2017) finds a good model parameter initialisation (called **meta-parameters**) that can quickly few-shot adapt to novel classes. Our framework incorporates BOL and meta-learning to give a recursive formula for the posterior of the meta-parameters as new few-shot datasets arrive. Taking a MAP estimate in implementation leads to Laplace approx-

[1]Department of Computer Science, University College London, London, United Kingdom [2]Alan Turing Institute, London, United Kingdom. Correspondence to: Pauching Yap <p.yap@cs.ucl.ac.uk>.

imation, whereas using a KL-divergence leads to variational inference. Our work builds on Ritter et al. (2018a) that combine BOL and Laplace approximation, and Nguyen et al. (2018) that use variational inference with BOL to prevent forgetting in large-scale supervised classification.

Advantage of our framework: An important reason to employ BOL over non-Bayesian approaches such as regret-based methods in an online setting is that BOL provides a grounded framework that suggests using the previous posterior as the prior recursively. BOL implicitly keeps a memory on previous knowledge via the posterior, in contrast to recent online meta-learning methods that explicitly accumulate previous data in a task buffer (Finn et al., 2019; Zhuang et al., 2019). Explicitly keeping a memory on previous data often triggers an important question: *how should the carried-forward data be processed in future rounds, in order to accumulate knowledge?* Finn et al. (2019) update the meta-parameters at each iteration using data sampled from the accumulated task buffer. This defeats the purpose of online learning, which by definition means to update the parameters each round using only the new data encountered.

Disadvantage of memorising past data: Having to re-train on previous data to avoid forgetting also increases the training time as the data accumulate (Finn et al., 2019; He et al., 2019). Certainly one can clamp the amount of data at some maximal limit and sample from the buffer, but the final performance of such an algorithm would be dependent on the samples being informative and of good quality which may vary across different seed runs. In contrast to memorising the datasets, having an implicit memory via the posterior automatically deals with the question on how to process carried-forward data and allows a better knowledge accumulation process.

Below are the contributions we make in this paper:

- We develop the Bayesian online meta-learning (BOML) framework for sequential few-shot classification problems. Under this framework we introduce the algorithms Bayesian online meta-learning with Laplace approximation (BOMLA) and Bayesian online meta-learning with variational inference (BOMVI).

- We propose an approximation to the Fisher corresponding to BOMLA that carries the desirable block-diagonal Kronecker-factored structure.

- We demonstrate that BOML can overcome catastrophic forgetting in the sequential few-shot datasets setting with apparent distributional shift in the datasets.

- We demonstrate empirically that BOML can also continually learn to few-shot classify the novel classes in the sequential meta-training few-shot tasks setting.

2. Meta-Learning

Most meta-learning algorithms comprise an inner loop for example-learning and an outer loop that learns the example-learning process. Such algorithms often require sampling a meta-batch of tasks at each iteration, where a **task** from a stationary task distribution $p(\mathcal{T})$ is formed by sampling a subset of classes from the pool of base classes or novel classes during meta-training or meta-evaluation respectively. The N-way K-shot task, for instance, refers to sampling N classes and using K examples per class for few-shot quick adaptation.

An offline meta-learning algorithm learns a model only for a specific dataset $\mathfrak{D}$, which is divided into the set of base classes $\mathcal{D}$ and novel classes $\widehat{\mathcal{D}}$ for meta-training and meta-evaluation respectively. Upon completing meta-training on $\mathcal{D}$, the goal is to perform well on an unseen task $\widehat{\mathcal{D}}^*$ sampled from the novel set $\widehat{\mathcal{D}}$ after a quick adaptation on a small subset $\widehat{\mathcal{D}}^{*,S}$ (known as the **support set**) of $\widehat{\mathcal{D}}^*$. The performance of this unseen task is evaluated on the **query set** $\widehat{\mathcal{D}}^{*,Q}$, where $\widehat{\mathcal{D}}^{*,Q} = \widehat{\mathcal{D}}^* \backslash \widehat{\mathcal{D}}^{*,S}$. Since $\widehat{\mathcal{D}}$ is not accessible during meta-training, this support-query split is mimicked on the base set $\mathcal{D}$ for meta-training.

Model-agnostic meta-learning: Each updating step of the well-known meta-learning algorithm MAML (Finn et al., 2017) aims to improve the ability of the meta-parameters to act as a good model initialisation for a quick adaptation on unseen tasks. Each iteration of the MAML algorithm samples M tasks from the base class set $\mathcal{D}$ and runs a few steps of stochastic gradient descent (SGD) for an inner loop task-specific learning. The number of tasks sampled per iteration is known as the **meta-batch size**. For task m, the inner loop outputs the task-specific parameters $\tilde{\theta}^m$ from a k-step SGD quick adaptation on the objective $\mathcal{L}(\theta, \mathcal{D}^{m,S})$ with the support set $\mathcal{D}^{m,S}$ and initialised at θ:

$$\tilde{\theta}^m = SGD_k(\mathcal{L}(\theta, \mathcal{D}^{m,S})), \qquad (1)$$

where $m = 1, \ldots, M$. The outer loop gathers all task-specific adaptations to update the meta-parameters θ using the loss $\mathcal{L}(\tilde{\theta}^m, \mathcal{D}^{m,Q})$ on the query set $\mathcal{D}^{m,Q}$.

The overall MAML optimisation objective is

$$\arg\min_\theta \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}(SGD_k(\mathcal{L}(\theta, \mathcal{D}^{m,S})), \mathcal{D}^{m,Q}). \quad (2)$$

Like most offline meta-learning algorithms, MAML assumes a stationary task distribution during meta-training and meta-evaluation. Under this assumption, a meta-learned model is only applicable to a specific dataset distribution. When the model encounters a sequence of datasets with apparent distributional shift, it loses the few-shot classification ability on previous datasets as new ones arrive for

meta-training. Our work aims to meta-learn a single model for few-shot classification on multiple datasets that arrive sequentially for meta-training. We achieve this goal by incorporating meta-learning into the BOL framework to give the *Bayesian online meta-learning* (BOML) framework that considers the posterior of the meta-parameters.

3. Bayesian Online Meta-Learning Framework Overview

Our central contribution is to extend the benefits of meta-learning to the BOL scenario, thereby training models that can generalise across tasks whilst dealing with parameter uncertainty in the setting of sequentially arriving datasets.

In this setting, meta-training occurs sequentially on the datasets $\mathfrak{D}_1, \ldots, \mathfrak{D}_T$. Each dataset $\mathfrak{D}_i$ can be seen as a knowledge domain with an associated underlying task distribution $p(\mathcal{T}_i)$. A newly-arrived $\mathfrak{D}_{t+1}$ is separated into the base class set $\mathcal{D}_{t+1}$ and novel class set $\widehat{\mathcal{D}}_{t+1}$ for meta-training and meta-evaluation respectively, where the tasks in these two stages are drawn from the task distribution $p(\mathcal{T}_{t+1})$. Notationally, let $\mathcal{D}_{t+1}^S$ and $\mathcal{D}_{t+1}^Q$ denote the collection of support sets and query sets respectively from $\mathcal{D}_{t+1}$, so that $\mathcal{D}_{t+1} = \mathcal{D}_{t+1}^S \cup \mathcal{D}_{t+1}^Q$. Using Bayes' rule on the posterior gives the recursive formula

$$p(\theta|\mathcal{D}_{1:t+1})$$
$$\propto p(\mathcal{D}_{t+1}^S, \mathcal{D}_{t+1}^Q|\theta)\, p(\theta|\mathcal{D}_{1:t}) \tag{3}$$
$$= p(\mathcal{D}_{t+1}^Q|\theta, \mathcal{D}_{t+1}^S)\, p(\mathcal{D}_{t+1}^S|\theta)\, p(\theta|\mathcal{D}_{1:t}) \tag{4}$$
$$= \int p(\mathcal{D}_{t+1}^Q|\tilde{\theta})\, p(\tilde{\theta}|\theta, \mathcal{D}_{t+1}^S)\, d\tilde{\theta}\, p(\mathcal{D}_{t+1}^S|\theta)\, p(\theta|\mathcal{D}_{1:t}) \tag{5}$$

where Eq. (3) follows from the assumption that each dataset is independent given θ. Figure 2 illustrates the BOML process flow for meta-training and meta-evaluation as datasets arrive sequentially.

From the meta-learning perspective, the parameters $\tilde{\theta}$ introduced in Eq. (5) can be viewed as the task-specific parameters in MAML. There are various choices for the distribution $p(\tilde{\theta}|\theta, \mathcal{D}_{t+1}^S)$ in Eq. (5). In particular if we choose to set it as the deterministic function of taking several steps of SGD on loss $\mathcal{L}$ with the support set collection $\mathcal{D}_{t+1}^S$ and initialised at θ, we have

$$p(\tilde{\theta}|\theta, \mathcal{D}_{t+1}^S) = \delta(\tilde{\theta} - SGD_k(\mathcal{L}(\theta, \mathcal{D}_{t+1}^S))), \tag{6}$$

where $\delta(\cdot)$ is the Dirac delta function. This recovers the MAML inner loop with SGD quick adaptation in Eq. (1). The recursion given by Eq. (5) forms the basis of our approach and the remainder of this paper explains how we implement this.
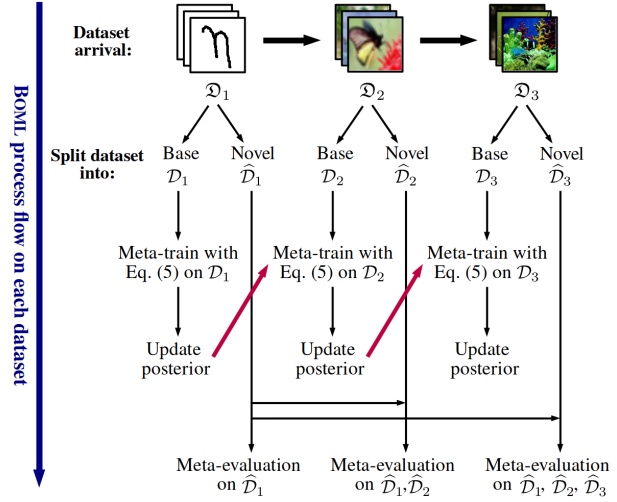

Figure 2. The BOML process flow for meta-training and meta-evaluation on an example sequence (Omniglot → CIFAR-FS → *mini*ImageNet) when each dataset arrives. The arrows in purple illustrate that the updated posterior is being brought forward for the next meta-training when a new dataset arrives.

4. Implementation

The posterior in Eq. (5) is typically intractable for modern neural network architectures. This leads to the requirement for a good approximate posterior. This section demonstrates how we arrive at the algorithms Bayesian online meta-learning with Laplace approximation (BOMLA) and Bayesian online meta-learning with variational inference (BOMVI) by implementing Laplace approximation and variational continual learning (VCL) respectively to the BOML posterior in Eq. (5). We give a mini tutorial in Appendix A on BOL, VCL and Laplace approximation.

4.1. BOMLA

As described in Appendix A.2, Laplace approximation justifies the use of a Gaussian approximate posterior by Taylor expanding the log-posterior around a mode up to the second order. The second order term corresponds to the log-probability of a Gaussian distribution. The BOML framework in Eq. (5) with a Gaussian approximate posterior q of mean and precision $\phi_t = \{\mu_t, \Lambda_t\}$ from the Laplace approximation gives a MAP estimate:

$$\theta^* = \arg\max_\theta \left\{ \log \bar{p}_\theta + \log p(\mathcal{D}_{t+1}^S|\theta) + r_\theta \right\}, \tag{7}$$

where

$$\bar{p}_\theta = \int p(\mathcal{D}_{t+1}^Q|\tilde{\theta}) p(\tilde{\theta}|\theta, \mathcal{D}_{t+1}^S)\, d\tilde{\theta},$$
$$r_\theta = -\frac{1}{2}(\theta - \mu_t)^T \Lambda_t (\theta - \mu_t).$$

For an efficient optimisation, we use the deterministic $\tilde{\theta}$ in Eq. (6) which leads to minimising the objective

$$f_{t+1}^{\text{BOMLA}}(\theta, \mu_t, \Lambda_t) = \bar{f}_\theta^{(1)} + \bar{f}_\theta^{(2)} - r_\theta, \tag{8}$$

where

$$\bar{f}_\theta^{(1)} = -\frac{1}{M} \sum_{m=1}^M \log p(\mathcal{D}_{t+1}^{m,Q}|\tilde{\theta}^m),$$

$$\bar{f}_\theta^{(2)} = -\frac{1}{M} \sum_{m=1}^M \log p(\mathcal{D}_{t+1}^{m,S}|\theta),$$

with $\tilde{\theta}^m = SGD_k(\mathcal{L}(\theta, \mathcal{D}_{t+1}^{m,S}))$ for $m = 1, \ldots, M$ and M denotes the number of tasks sampled per iteration. The first term $\bar{f}_\theta^{(1)}$ in Eq. (8) corresponds to the MAML objective in Eq. (2) with a cross-entropy loss, the second term $\bar{f}_\theta^{(2)}$ can be viewed as the pre-adaptation loss on the support set and the last term r_θ can be seen as a regulariser.

We discover that the Laplace approximation method provides a well-fitted meta-training framework for BOML in Eq. (5). Each updating step in the approximation procedure can be modified to correspond to the meta-parameters for few-shot classification, instead of the model parameters for large-scale supervised classification.

4.2. Hessian approximation

We calculate a block-diagonal Kronecker-factored Hessian approximation in order to update the precision Λ_t, as explained in Appendix A.3.

The Hessian matrix corresponding to the first term of the BOMLA objective in Eq. (8) is

$$\widetilde{H}_{t+1}^{ij} = \frac{1}{M} \sum_{m=1}^M -\frac{\partial^2}{\partial\theta^{(i)}\partial\theta^{(j)}} \log p(\mathcal{D}_{t+1}^{m,Q}|\tilde{\theta}^m))\Big|_{\theta=\mu_{t+1}}. \tag{9}$$

It is worth noting that the BOMLA Hessian deviates from the original Laplace approximation Hessian in Appendix A.2, and it is necessary to derive an adjusted approximation to the Hessian with some further assumptions.

The Hessian for a single data point can be approximated using the Fisher information matrix F to ensure its positive semi-definiteness (Martens & Grosse, 2015):

$$F = \mathbb{E}_{x,y}\left[\frac{d}{d\theta}\log p(y|x,\theta)\frac{d}{d\theta}\log p(y|x,\theta)^T\right]. \tag{10}$$

Each (x, y) pair for the Fisher in BOMLA is associated to a task m. The Fisher information matrix $\widetilde{F}$ corresponding to

the BOMLA Hessian in Eq. (9) for a single data point is

$$\widetilde{F} = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{x,y}\left[\left(\frac{\partial\tilde{\theta}^m}{\partial\theta}\right)\frac{d}{d\tilde{\theta}^m}\log p(y|x,\tilde{\theta}^m)\right. \\ \left. \times \frac{d}{d\tilde{\theta}^m}\log p(y|x,\tilde{\theta}^m)^T\left(\frac{\partial\tilde{\theta}^m}{\partial\theta}\right)^T\right]. \tag{11}$$

The additional Jacobian matrix $\frac{\partial\tilde{\theta}^m}{\partial\theta}$ breaks the Kronecker-factored structure described by Martens & Grosse (2015) for the original Fisher in Eq. (10).

The results in Finn et al. (2017) show that the first step of the quick adaptation in $\tilde{\theta}^m$ contributes the largest change to the meta-evaluation objective, and the remaining adaptation steps give a relatively small change to the objective. We can reasonably assume a one-step SGD quick adaptation $\tilde{\theta}^m = \theta - \nabla_\theta\mathcal{L}(\theta, \mathcal{D}_{t+1}^{m,S})$ to approximate the Fisher, although in other parts of the implementation we use a few-step SGD. By imposing this assumption, the (i, j)-th entry of the Jacobian term can be interpreted as

$$\left(\frac{\partial\tilde{\theta}^m}{\partial\theta}\right)^{ij} = I^{ij} - \frac{\partial^2(-\log p(\mathcal{D}_{t+1}^{m,S}|\theta))}{\partial\theta^{(i)}\partial\theta^{(j)}}, \tag{12}$$

where I is the corresponding identity matrix and the objective $\mathcal{L}$ involved is the negative log-likelihood. The Hessian for a single data point in the second term of Eq. (12) can be approximated by F in Eq. (10) via the usual block-diagonal Kronecker-factored approximation. Putting the Jacobian back into Eq. (11) and expanding the factors give terms that multiply two or more Kronecker products together. The detailed derivation of $\widetilde{F}$ is explained in Appendix A.3.1. We introduce the posterior-regularising hyperparameter λ when updating the precision: $\Lambda_{t+1} = \lambda\widetilde{H}_{t+1} + \Lambda_t$ and the rationale for introducing λ is explained in Appendix A.3.2. The pseudo-code of the BOMLA algorithm can be found in Appendix B.1.

4.3. BOMVI

The VCL framework (Nguyen et al., 2018) is directly applicable to BOML. This section demonstrates how we arrive at the BOMVI algorithm by implementing VCL to the posterior of the BOML framework in Eq. (5).

As described in Appendix A.4, VCL approximates the posterior by minimising the KL-divergence over some predetermined approximate posterior family $\mathcal{Q}$. Fitting the BOML posterior in Eq. (5) into the VCL framework gives the approximate posterior:

$$q(\theta|\phi_{t+1}) = \underset{q\in\mathcal{Q}}{\arg\min}\, D_{\text{KL}}\big(q(\theta|\phi)\big\|\check{q}_{\phi_t}\big), \tag{13}$$

where $\check{q}_{\phi_t} = \int p(\mathcal{D}_{t+1}^Q|\tilde{\theta})\,p(\tilde{\theta}|\theta, \mathcal{D}_{t+1}^S)d\tilde{\theta}\,p(\mathcal{D}_{t+1}^S|\theta)q(\theta|\phi_t)$. Similar to BOMLA, we use the deterministic $\tilde{\theta}$ in Eq. (6).

This leads to minimising the objective

$$f_{t+1}^{\text{BomVI}}(\phi, \phi_t) = \breve{f}_\phi^{(1)} + \breve{f}_\phi^{(2)} + \breve{r}_\phi, \qquad (14)$$

where

$$\breve{f}_\phi^{(1)} = -\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{q(\theta|\phi)} \big[\log p(\mathcal{D}_{t+1}^{m,Q} | \tilde{\theta}^m) \big],$$

$$\breve{f}_\phi^{(2)} = -\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{q(\theta|\phi)} \big[\log p(\mathcal{D}_{t+1}^{m,S} | \theta) \big],$$

$$\breve{r}_\phi = D_{\text{KL}}(q(\theta|\phi) \| q(\theta|\phi_t)),$$

with $\tilde{\theta}^m = SGD_k(\mathcal{L}(\theta, \mathcal{D}_{t+1}^{m,S}))$ for $m = 1, \ldots, M$ and M denotes the number of tasks sampled per iteration. We use a Gaussian mean-field $q(\theta|\phi_t) = \prod_{d=1}^{D} N(\mu_{t,d}, \sigma_{t,d}^2)$, where $\phi_t = \{\mu_{t,d}, \sigma_{t,d}\}_{d=1}^{D}$, $D = \dim(\theta)$ and the objective in Eq. (14) is minimised over ϕ. The pseudo-code of the BomVI algorithm can be found in Appendix B.1.

The term $\breve{f}_\phi^{(1)}$ in Eq. (14) is rather cumbersome to estimate in optimisation. To compute its Monte Carlo estimate, we have to generate samples $\theta_r \sim q$ for $r = 1, \ldots, R$, and run a quick adaptation on each sampled meta-parameters θ_r before evaluating their log-likelihoods. This is computationally intensive and gives an estimator with large variance. We propose a workaround by modifying the inner loop SGD quick adaptation, and the details can be found in Appendix B.2.

5. Related Work

5.1. Online Meta-Learning

Regret minimisation: The goal of this setting is to minimise the regret function, and the assumptions are made on the loss function rather than the task distribution. Recent works Finn et al. (2019) and Zhuang et al. (2019) belong to this category, where the aim is to compete with the best meta-learner and supersede it. These methods accumulate data as they arrive and meta-learn using all data acquired so far. Data accumulation is not desirable as the algorithmic complexity of training grows with the amount of data accumulated, and training time increases as new data arrive (Finn et al., 2019; He et al., 2019). The agent will eventually run out of memory for a long sequence of data. The BoML framework on the other hand is advantageous, as it only takes the current data and the posterior of the meta-parameters into consideration during optimisation. This gives a framework with an algorithmic complexity independent of the length of the dataset sequence.

Same underlying task distribution: Sequential tasks are assumed to originate from the same underlying task distribution $p(\mathcal{T})$ in this setting. Denevi et al. (2019) introduce

the online-within-online (OWO) and online-within-batch (OWB) settings, where OWO encounters tasks and examples within tasks sequentially while OWB encounters tasks sequentially but examples within tasks are in batch. Our work in the sequential datasets setting is novel in overcoming few-shot catastrophic forgetting, where the goal is to few-shot classify unseen tasks drawn from a sequence of distributions $p(\mathcal{T}_1), \ldots, p(\mathcal{T}_T)$ as explained in Section 3. He et al. (2019), Harrison et al. (2019) and Jerfel et al. (2019) look into continual meta-learning for a non-stationary task distribution where the task boundaries are unknown to the model. Jerfel et al. (2019) consider a latent task structure to adapt to the non-stationary task distribution.

5.2. Offline Meta-Learning

Previous meta-learning works attempt to solve few-shot classification problems in an offline setting, under the assumption of having a stationary task distribution during meta-training and meta-evaluation. A single meta-learned model is aimed to few-shot classify one specific dataset with all base classes of the dataset readily available in a batch for meta-training. There are two general frameworks for the offline meta-learning setting:

Probabilistic: The MAML algorithm can be cast into a probabilistic inference problem (Finn et al., 2018) or with a hierarchical Bayesian structure (Grant et al., 2018; Yoon et al., 2018). Yoon et al. (2018) use Stein Variational Gradient Descent (SVGD) for task-specific learning. Gordon et al. (2019) implement probabilistic inference by considering the posterior predictive distribution with amortised networks. Grant et al. (2018) discuss the use of a Laplace approximation in the task-specific inner loop to improve MAML using the curvature information. Although at first sight our work seems similar to Grant et al. (2018) due to the use of Laplace approximation, our work is clearly distinct in terms of goal and context. Grant et al. (2018) use Laplace approximation at the task-specific level, whilst we use Laplace approximation at the meta-level for the meta-parameters approximate posterior. The formulation in Grant et al. (2018) does not accumulate past experience, whereas our work enables few-shot learning on unseen tasks from multiple knowledge domains sequentially.

Non-probabilistic: Gradient-based meta-learning (Finn et al., 2017; Nichol et al., 2018; Rusu et al., 2019) updates the meta-parameters by accumulating the gradients of a meta-batch of task-specific inner loop updates. The meta-parameters will be used as a model initialisation for a quick adaptation on the novel tasks. Metric-based meta-learning (Koch et al., 2015; Vinyals et al., 2016; Snell et al., 2017) utilises the metric distance between labelled examples. This method assumes that base and novel classes are from the

same dataset distribution, and the metric distance estimations can be generalised to the novel classes upon meta-learning the base classes.

5.3. Continual Learning

Modern continual learning works (Goodfellow et al., 2013; Lee et al., 2017; Zenke et al., 2017) focus primarily on large-scale supervised learning, in contrast to our work that looks into continual few-shot classification across sequential datasets with evident distributional shift. Wen et al. (2018) utilise few-shot learning to improve on overcoming catastrophic forgetting via logit matching on a small sample from the previous tasks. The online learning element in our work is closely related to recent works that overcome catastrophic forgetting in large-scale supervised classification (Kirkpatrick et al., 2017; Zenke et al., 2017; Ritter et al., 2018a; Nguyen et al., 2018). In particular, our work builds on the online Laplace approximation method (Ritter et al., 2018a). Our work extends this to the meta-learning scenario to avoid forgetting in few-shot classification problems. Nguyen et al. (2018) provide an alternative of using variational inference instead of Laplace approximation for approximating the posterior. Our work utilises this approach to adapt the variational method to approximate the posterior of the meta-parameters by adjusting the KL-divergence objective.

6. Experiments

6.1. N-athlon

We implement BOMLA and BOMVI[1] to the 5-way 1-shot **triathlon** and **pentathlon** sequences. The experiment details and the datasets explanations are in Appendix C.1. We compare our algorithms to the following baselines:

Train-On-Everything (TOE): When a new dataset arrives for meta-training, we randomly re-initialise the meta-parameters and run MAML meta-training using *all datasets encountered so far*.

Sequential MAML: Upon the arrival of a new dataset, we run MAML to meta-train *only on the newly-arrived dataset*.

Follow The Meta-Leader (FTML): We introduce a slight modification to FTML (Finn et al., 2019) on its evaluation method, as FTML is not designed for few-shot learning on unseen tasks. In our experiments, we apply `Update-Procedure` in FTML to the data from unseen tasks, rather than the data from the same training task as in the original FTML.

---

[1] Implementation code is available at https://github.com/pauchingyap/boml

6.1.1. TRIATHLON

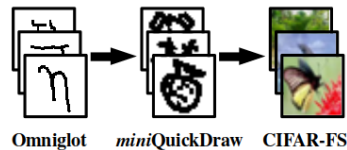This experiment considers the few-shot triathlon sequence as in Figure 3.

Figure 3. The triathlon 5-way 1-shot sequence in this experiment.

The distributional shift from Omniglot to *mini*QuickDraw is less drastic, compared to the shift from *mini*QuickDraw to CIFAR-FS. The result in Figure 4 shows that BOMLA and BOMVI are able to prevent catastrophic forgetting in both dataset transitions. BOMLA, in particular, is able to proceed to the *mini*QuickDraw meta-training phase with **almost no forgetting on Omniglot**. In other words, the meta-level pattern of Omniglot is retained throughout the meta-training period of *mini*QuickDraw. There is a small trade-off in the performance of CIFAR-FS as BOMLA and BOMVI avoid catastrophically forgetting Omniglot and *mini*QuickDraw. Sequential MAML gives a noticeable drop in the performance of Omniglot and *mini*QuickDraw when meta-training on CIFAR-FS. TOE is able to retain the few-shot performance as it has access to all previous datasets, whilst FTML gives a mixed performance. We elaborate on the result interpretation, the BOMLA-BOMVI comparison and the choice of λ along with the next experiment pentathlon, which resembles the setting of this experiment except with a more challenging dataset sequence.
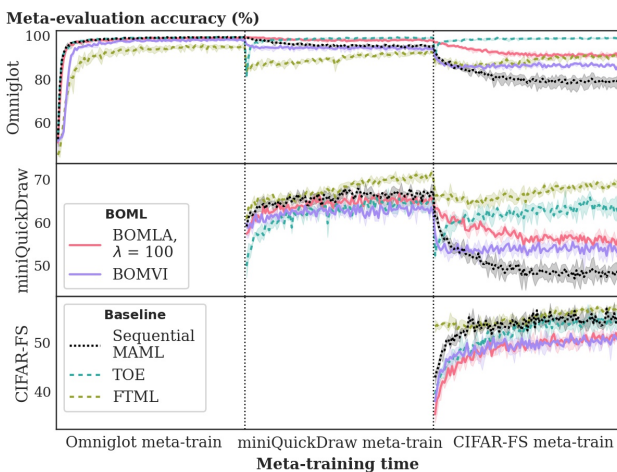

Figure 4. Meta-evaluation accuracy across 3 seed runs on each dataset along meta-training. Higher accuracy values in the off-diagonals indicate less forgetting.

6.1.2. PENTATHLON

We implement BOMLA and BOMVI to the more challenging pentathlon sequence as in Figure 5.

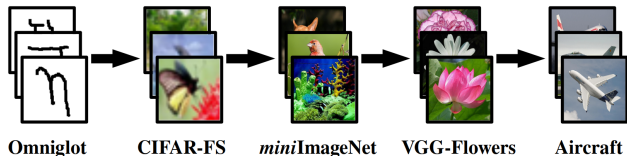Omniglot CIFAR-FS *mini*ImageNet VGG-Flowers Aircraft

Figure 5. The pentathlon 5-way 1-shot sequence in this experiment.

Figure 6 shows that BOMLA and BOMVI are able to prevent few-shot catastrophic forgetting in the pentathlon dataset sequence. TOE is also able to retain the few-shot performance as it has access to all datasets encountered so far.

Since TOE learns all datasets from random re-initialisation each time it encounters a new dataset, the meta-training time required to achieve a similarly good meta-evaluation performance is longer compared to other runs. Sequential MAML catastrophically forgets the previously learned datasets but has the best performance on new datasets compared to other runs. FTML gives a mixed performance on different datasets.

TOE and FTML can be memory-intensive as the dataset sequence becomes longer. They take the brute-force approach to prevent forgetting by memorising all datasets. Unlike TOE and FTML, our algorithms BOMLA and BOMVI only take the newly-arrived dataset and the posterior of the meta-parameters into consideration during optimisation. This gives a framework with an algorithmic complexity independent of the length of the dataset sequence.
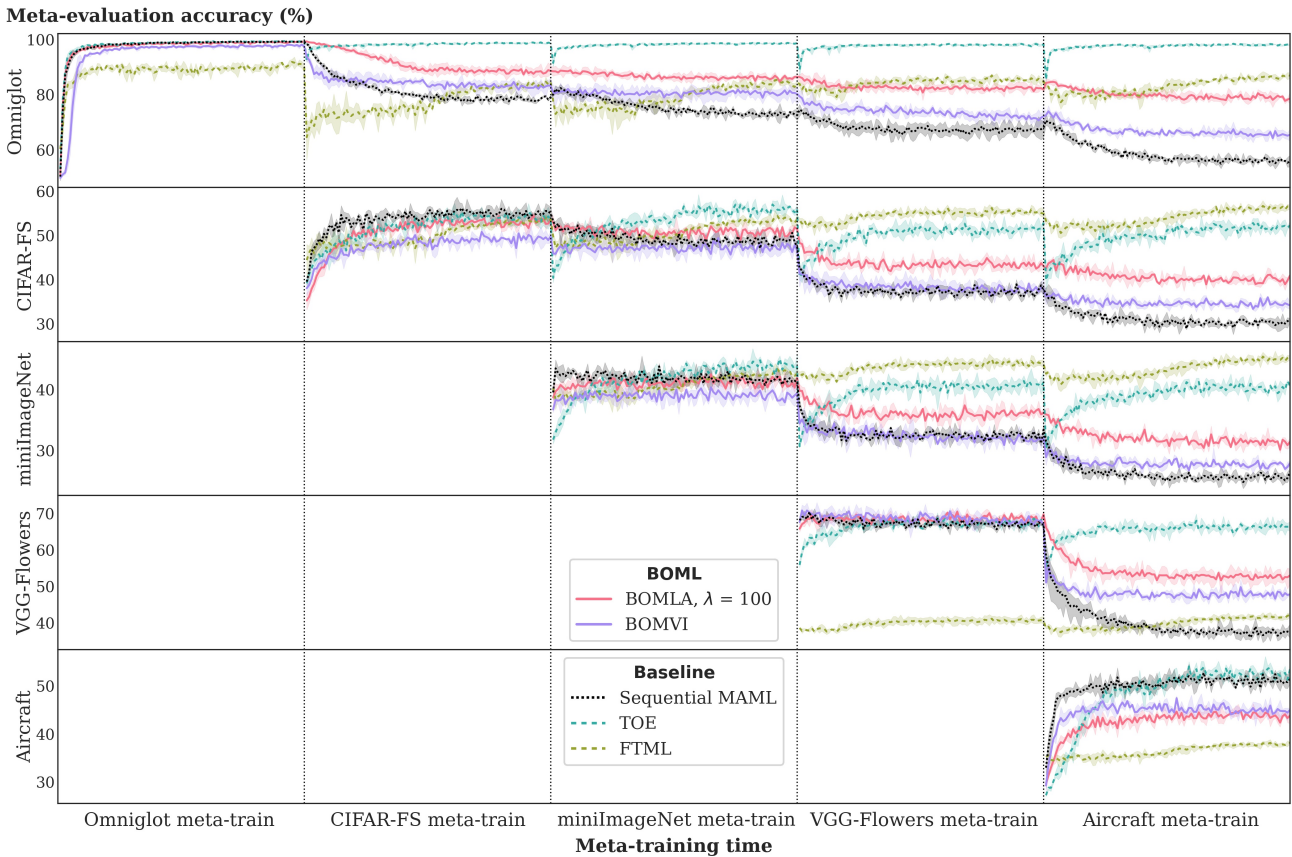

Figure 6. Meta-evaluation accuracy across 3 seed runs on each dataset along meta-training. Higher accuracy values indicate better results with less forgetting as we proceed to new datasets. BOMLA with $\lambda = 100$ gives good performance in the off-diagonal plots (retains performances on previously learned datasets), and has a minor performance trade-off in the diagonal plots (learns less well on new datasets). Sequential MAML gives better performance in the diagonal plots (learns well on new datasets) but worse performance in the off-diagonal plots (forgets previously learned datasets). BOMVI is also able to retain performance on previous datasets, although it may be unable to perform as good as BOMLA due to sampling and estimator variance.

BOMLA-BOMVI comparison: As shown in Figure 6, BOMLA with appropriate λ is superior to BOMVI in the performance. This is due to BOMLA having a better posterior approximation than BOMVI. Whilst BOMLA has a Gaussian approximate posterior with block-diagonal precision, BOMVI uses a Gaussian mean-field approximate posterior. Trippe & Turner (2017) compare the performances of variational inference with different covariance structures, and discover that variational inference with block-diagonal covariance performs worse than mean-field approximation. This is because the block-diagonal covariance in variational inference prohibits variance reduction methods such as local reparameterisation trick for Monte Carlo estimation. The variance of the Monte Carlo estimate has been proven problematic (Kingma et al., 2015; Trippe & Turner, 2017). We address this issue in Section 4.3 and Appendix B.2 specifically to the meta-learning setting by modifying the inner loop quick adaptation. We analyse the change in the approximate posterior covariance in Appendix C.2, as meta-training occurs sequentially on datasets from different knowledge domains.

Choosing λ: Tuning the posterior regulariser λ mentioned in Section 4.2 and Appendix A.3.2 corresponds to balancing between a smaller performance trade-off on a new dataset and less forgetting on previous datasets. We compare BOMLA with different λ values and BOMVI in Appendix C.3. A larger $\lambda = 1000$ results in a more concentrated Gaussian posterior and is therefore unable to learn new datasets well, but can better retain the performances on previous datasets. A smaller $\lambda = 1$ on the other hand gives a widespread Gaussian posterior and learns better on new datasets by sacrificing the performance on the previous datasets. In both triathlon and pentathlon experiments, the value $\lambda = 100$ gives the best balance between old and new datasets. Ideally we seek for a good performance on both old and new datasets, but in reality there is a trade-off between retaining performance on old datasets and learning well on new datasets due to posterior approximation errors.

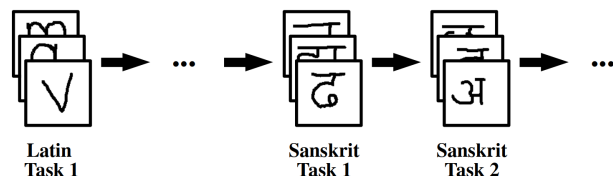### 6.2. Omniglot: Stationary Task Distribution

Figure 7. An example of the Omniglot task sequence for meta-training in this experiment.

In this experiment we demonstrate empirically that BOML can also continually learn to few-shot classify the novel classes in the sequential meta-training few-shot tasks setting, where all tasks originate from a stationary task distribution. This setting only involves **one dataset** $\mathfrak{D}$ with an associated underlying task distribution $p(\mathcal{T})$, where $\mathfrak{D}$ is separated into the base and novel class sets. In this setting, $\mathcal{D}_1, \ldots, \mathcal{D}_{t+1}$ denote the non-overlapping tasks formed from the base class set and they arrive sequentially for meta-training. We show the corresponding modifications of BOMLA and BOMVI under this setting in Appendix B.1 Algorithms 1 and 2.

We run the sequential tasks experiment on the Omniglot dataset. To increase the difficulty level, we split the dataset based on the alphabets (super-classes) instead of the characters (classes) as in Figure 7. The goal of this experiment is to classify the 5-way 5-shot novel tasks sampled from the meta-evaluation alphabets. The experimental details and the alphabet splits can be found in Appendix C.4. We compare our algorithms to the baselines **TOE**, **Sequential MAML** and **FTML** similar to the N-athlon experiments but in the sequential tasks setting.

As the tasks arrive sequentially for meta-training, Figure 8 shows that BOMLA and BOMVI can accumulate few-shot classification ability on the novel tasks over time. The knowledge acquired from previous meta-training tasks is carried forward in the form of a posterior, which is then used as a prior when a new task arrives for meta-training. Despite having access to all previous tasks, TOE shows no positive forward transfer in the meta-evaluation accuracy each time it encounters a new task. FTML and sequential MAML are inferior to BOMLA and BOMVI in the performance. BOMLA with $\lambda = 0.01$ gives the best performance in this experiment.

6.3. Discussion on Baselines

Finn et al. (2019) discover that TOE does not explicitly learn the structure across tasks, thus unable to fully utilise the data. The TOE performance in our Omniglot experiment is coherent with the TOE result in Finn et al. (2019). The result figures in Finn et al. (2019) show a TOE result similar to ours in the Omniglot experiment. In contrast, TOE in the N-athlon experiments performs well as it has access to drastically more data points than TOE in the Omniglot experiment, and samples numerous tasks from all previous datasets.

Sequential MAML in the N-athlon experiments suffers from catastrophic forgetting due to the apparent distributional shift in the datasets. The Omniglot experiment, on the other hand, has tasks originating from the same underlying distribution. As a result sequential MAML in this setting is able to accumulate few-shot ability, although it performs worse than BOMLA and BOMVI as shown in Figure 8 since there is only one task available at a time.
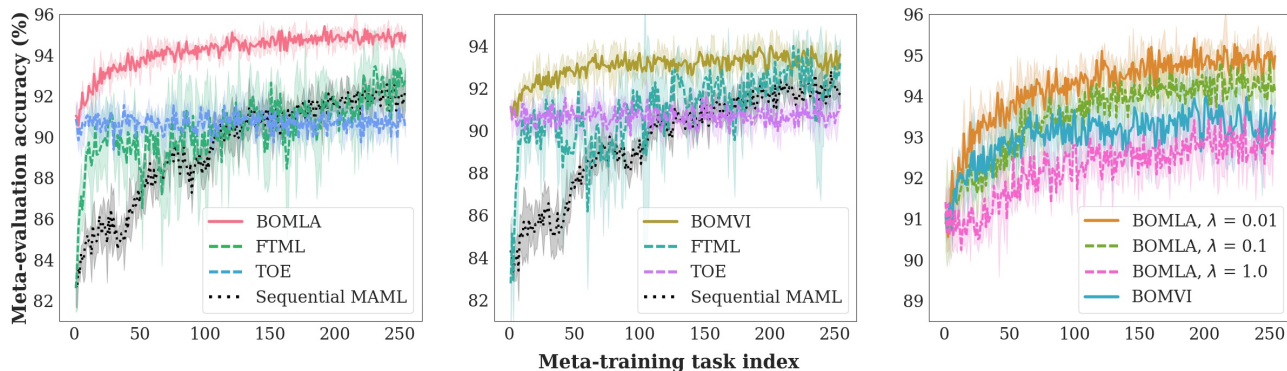
Figure 8. Meta-evaluation accuracy across 3 seed runs on the novel tasks along meta-training. **Left**: compares BOMLA to the baselines, **centre**: compares BOMVI to the baselines, **right**: compares BOMLA with different λ values to BOMVI.

Since the original FTML is not aimed for unseen few-shot tasks and does not deal with sequential datasets setting as in the N-athlon experiments, we have to modify FTML as described in Section 6.1. Sampling from previous tasks in the buffer is a key feature of the FTML algorithm. Certainly one can sample many tasks from the buffer to achieve perfect memory in the N-athlon experiments, but such a baseline setup has been taken into consideration by TOE. Therefore we choose to retain the online characteristic of the original FTML in our modified implementation.

7. Conclusion

We introduced the Bayesian online meta-learning (BOML) framework with two algorithms: BOMLA and BOMVI. Our framework can overcome catastrophic forgetting in few-shot classification problems on datasets with evident distributional shift. BOML merged the BOL framework and meta-learning via Laplace approximation or variational continual learning. We proposed the necessary adjustments in the Hessian and Fisher approximation for BOMLA, as we optimise the meta-parameters for few-shot classification instead of the usual model parameters in large-scale supervised classification. The experiments show that BOMLA and BOMVI are able to retain the few-shot classification ability when trained on sequential datasets with apparent distributional shift, resulting in the ability to perform few-shot classification on multiple datasets with a single meta-learned model. BOMLA and BOMVI are also able to continually learn to few-shot classify the novel tasks, as the meta-training tasks from a stationary distribution arrive sequentially for learning.

References

Bertinetto, L., Henriques, J. F., Torr, P., and Vedaldi, A. Meta-Learning with Differentiable Closed-Form Solvers. In *International Conference on Learning Representations*, 2019.

Botev, A., Ritter, H., and Barber, D. Practical Gauss-Newton Optimisation for Deep Learning. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Denevi, G., Stamos, D., Ciliberto, C., and Pontil, M. Online-Within-Online Meta-Learning. In *Advances in Neural Information Processing Systems 32*, 2019.

Denker, J. S. and LeCun, Y. Transforming Neural-Net Output Levels to Probability Distributions. In *Advances in Neural Information Processing Systems 3*, 1991.

Finn, C., Abbeel, P., and Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Finn, C., Xu, K., and Levine, S. Probabilistic Model-Agnostic Meta-Learning. In *Advances in Neural Information Processing Systems 31*, 2018.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online Meta-Learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An Empirical Investigation of Catastrophic

Forgetting in Gradient-Based Neural Networks. *arXiv preprint*, arXiv:1312.6211, 2013.

Gordon, J., Bronskill, J., Bauer, M., Nowozin, S., and Turner, R. Meta-Learning Probabilistic Inference for Prediction. In *International Conference on Learning Representations*, 2019.

Grant, E., Finn, C., Levine, S., Darrell, T., and Griffiths, T. Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. In *International Conference on Learning Representations*, 2018.

Grosse, R. and Martens, J. A Kronecker-Factored Approximate Fisher Matrix for Convolution Layers. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Ha, D. and Eck, D. A Neural Representation of Sketch Drawings. *arXiv preprint*, arXiv:1704.03477, 2017.

Harrison, J., Sharma, A., Finn, C., and Pavone, M. Continuous Meta-Learning without Tasks. *arXiv preprint*, arXiv:1912.08866, 2019.

He, X., Sygnowski, J., Galashov, A., Rusu, A. A., Teh, Y., and Pascanu, R. Task Agnostic Continual Learning via Meta Learning. *arXiv preprint*, arXiv:1906.05201, 2019.

Jerfel, G., Grant, E., Griffiths, T., and Heller, K. A. Reconciling Meta-Learning and Continual Learning with Online Mixtures of Tasks. In *Advances in Neural Information Processing Systems 32*, 2019.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

Kingma, D. P., Salimans, T., and Welling, M. Variational Dropout and the Local Reparameterization Trick. In *Advances in Neural Information Processing Systems 28*, 2015.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 2017.

Koch, G., Zemel, R., and Salakhutdinov, R. Siamese Neural Networks for One-Shot Image Recognition. In *32th International Conference on Machine Learning Deep Learning Workshop*, 2015.

Lake, B., Salakhutdinov, R., Gross, J., and Tenenbaum, J. One Shot Learning of Simple Visual Concepts. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

Lee, S., Kim, J., Jun, J., Ha, J., and Zhang, B. Overcoming Catastrophic Forgetting by Incremental Moment Matching. In *Advances in Neural Information Processing Systems 30*, 2017.

Li, F., Fergus, R., and Perona, P. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2004.

MacKay, D. J. C. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 1992.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-Grained Visual Classification of Aircraft. *arXiv preprint*, arXiv:1306.5151, 2013.

Martens, J. and Grosse, R. Optimizing Neural Networks with Kronecker-Factored Approximate Curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Miller, E. G., Matsakis, N. E., and Viola, P. A. Learning from One Example Through Shared Densities on Transforms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational Continual Learning. In *International Conference on Learning Representations*, 2018.

Nichol, A., Achiam, J., and Schulman, J. On First-Order Meta-Learning Algorithms. *arXiv preprint*, arXiv:1803.02999, 2018.

Nilsback, M. and Zisserman, A. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Opper, M. *A Bayesian Approach to Online Learning*. Cambridge University Press, 1998.

Osawa, K., Tsuji, Y., Ueno, Y., Naruse, A., Foo, C., and Yokota, R. Scalable and Practical Natural Gradient for Large-Scale Deep Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Ravi, S. and Beatson, A. Amortized Bayesian Meta-Learning. In *International Conference on Learning Representations*, 2019.

Ravi, S. and Larochelle, H. Optimization as a Model for Few-Shot Learning. In *International Conference on Learning Representations*, 2017.

Ritter, H., Botev, A., and Barber, D. Online Structured Laplace Approximations for Overcoming Catastrophic Forgetting. In *Advances in Neural Information Processing Systems 31*, 2018a.

Ritter, H., Botev, A., and Barber, D. A Scalable Laplace Approximation for Neural Networks. In *International Conference on Learning Representations*, 2018b.

Robbins, H. and Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 1951.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-Learning with Latent Embedding Optimization. In *International Conference on Learning Representations*, 2019.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. Meta-Learning with Memory-Augmented Neural Networks. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Schmidhuber, J. Evolutionary Principles in Self-Referential Learning. On Learning How to Learn: The Meta-Meta-Meta...-Hook. Diploma thesis, Institut für Informatik, Technische Universität München, 1987.

Snell, J., Swersky, K., and Zemel, R. Prototypical Networks for Few-Shot Learning. In *Advances in Neural Information Processing Systems 30*, 2017.

Thrun, S. and Pratt, L. *Learning to Learn: Introduction and Overview*. Springer, Boston, MA, 1998.

Trippe, B. L. and Turner, R. E. Overpruning in Variational Bayesian Neural Networks. In *Advances in Neural Information Processing Systems 30 – Advances in Approximate Bayesian Inference Workshop*, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems 29*, 2016.

Wen, J., Cao, Y., and Huang, R. Few-Shot Self Reminder to Overcome Catastrophic Forgetting. *arXiv preprint*, arXiv:1812.00543, 2018.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian Model-Agnostic Meta-Learning. In *Advances in Neural Information Processing Systems 31*, 2018.

Zenke, F., Poole, B., and Ganguli, S. Continual Learning through Synaptic Intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Zhuang, Z., Wang, Y., Yu, K., and Lu, S. No-Regret Non-Convex Online Meta-Learning. *arXiv preprint*, arXiv:1910.10196, 2019.