Efficient Variance Reduction for Meta-Learning

Hansi Yang [1] **James T. Kwok** [1]

Abstract

Meta-learning tries to learn meta-knowledge from a large number of tasks. However, the stochastic meta-gradient can have large variance due to data sampling (from each task) and task sampling (from the whole task distribution), leading to slow convergence. In this paper, we propose a novel approach that integrates variance reduction with first-order meta-learning algorithms such as Reptile. It retains the bilevel formulation which better captures the structure of meta-learning, but does not require storing the vast number of task-specific parameters in general bilevel variance reduction methods. Theoretical results show that it has fast convergence rate due to variance reduction. Experiments on benchmark few-shot classification data sets demonstrate its effectiveness over state-of-the-art meta-learning algorithms with and without variance reduction.

1. Introduction

Meta-learning (Hospedales et al., 2021), or learning to learn (Thrun & Pratt, 1998), aims to quickly learn new tasks by utilizing meta-knowledge from tasks that are already learned. This is especially useful for deep networks, as they typically have to train on a huge amount of labeled samples for each task. Meta-learning has been successfully used in various applications such as few-shot learning (Wang et al., 2020; Finn et al., 2017; Nichol et al., 2018), reinforcement learning (Clavera et al., 2019; Gupta et al., 2018), neural architecture search (Elsken et al., 2020), and semi-supervised learning (Shu et al., 2019; Ren et al., 2020). In this paper, we focus on a particularly well-known family of meta-learning algorithms which is based on the MAML (Finn et al., 2017) and its variants (such as Reptile (Nichol et al., 2018) and ANIL (Raghu et al., 2020)).

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. Correspondence to: James Kwok <jamesk@cse.ust.hk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Mathematically, meta-learning is often formulated as a bilevel optimization problem (Franceschi et al., 2018). The outer problem learns some meta-parameter useful to all the tasks; while the lower-level problem (one for each task) learns a task-specific model by adapting the meta-parameter. As meta-learning involves learning from a lot of tasks, variance of the stochastic meta-gradient comes from two sources: (i) variance of data samples for each task, and (ii) variance of task samples from the task distribution. This is different from simpler machine learning problems involving only one task, in which the variance comes only from the data samples. When the tasks are diverse, meta-learning algorithms can suffer from large variance of their updates, leading to slow convergence (Ghadimi & Lan, 2013; Ji et al., 2020).

To have faster convergence, a natural approach is to use variance reduction (Gower et al., 2020), which accelerates convergence by reducing the variance in the stochastic gradients. As is demonstrated by the classic variance reduction methods such as SVRG (Johnson & Zhang, 2013) and SARAH (Nguyen et al., 2017), this can achieve convergence rates faster than SGD both theoretically and empirically. However, the batch gradient has to be computed occasionally, which can still be very expensive when the training data set is large. To alleviate this problem, a recent variance reduction algorithm, STORM (Cutkosky & Orabona, 2019), proposes to perform variance reduction without the need for batch gradient, while still achieving the same convergence rate as previous variance reduction methods.

Very recently, Wang et al. (2021) made an initial attempt to use variance reduction in meta-learning. They proposed VFML, which integrates STORM into the first-order meta-learning algorithm Reptile. However, VFML ignores the bilevel structure in meta-learning. Moreover, a theoretical study on its variance reduction properties is lacking.

While classic variance reduction algorithms mainly focus on single-level stochastic optimization problems, there are recent extensions to bilevel stochastic optimization (Khanduri et al., 2021; Yang et al., 2021). In principle, these can be straightforwardly incorporated into meta-learning algorithms, by simply replacing the original gradients by their variance-reduced counterparts. However, during optimization, this requires storing the task-specific parameters for all

the tasks. When the number of tasks is large (as is typically the case) or the task model is huge (as in many deep learning models), the subsequent storage cost can be prohibitive.

In this paper, we propose an efficient variance reduction method that can be used with various meta-learning algorithms. The proposed family of variance-reduced variants is aware of the bilevel optimization structure in meta-learning, while removing the need for storing task-specific parameters in existing bilevel variance reduction methods. We show theoretically that it achieves a faster convergence rate due to variance reduction. Experiments on benchmark few-shot image classification data sets also demonstrate the effectiveness of the proposed method.

A summary of this paper's contributions is as follows: (i) we propose a novel variance reduction method which can be integrated into various meta-learning algorithms; (ii) we provide theoretical analysis demonstrating that the proposed method has a faster convergence rate; and (iii) extensive experiments on benchmark data sets show that it has faster convergence and better performance than existing meta-learning algorithms with and without variance reduction.

2. Related works

2.1. Meta-learning

Given a set of tasks $\mathcal{I}$, meta-learning (Hospedales et al., 2021) is commonly formulated as the following bilevel optimization problem:

$$\min_{\boldsymbol{w}} \quad \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{L}((\boldsymbol{w},\boldsymbol{\theta}^i(\boldsymbol{w})); \mathcal{D}^i_{\text{vald}}) \tag{1}$$

$$\text{s.t.} \quad \boldsymbol{\theta}^i \equiv \boldsymbol{\theta}^i(\boldsymbol{w}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}((\boldsymbol{w},\boldsymbol{\theta}); \mathcal{D}^i_{\text{tr}}), \tag{2}$$

where $\boldsymbol{w}$ is the meta-parameter shared among all tasks, $\boldsymbol{\theta}^i(\boldsymbol{w})$ is the parameter specific to task i, $\mathcal{D}^i_{\text{vald}}$ and $\mathcal{D}^i_{\text{tr}}$'s are the meta-validation and meta-training data, respectively, for task i, $\mathcal{L}((\boldsymbol{w},\boldsymbol{\theta});\mathcal{D}) = \mathbb{E}_{\xi\sim\mathcal{D}}[\ell((\boldsymbol{w},\boldsymbol{\theta});\xi)]$ is the loss of task i's model on data $\mathcal{D}$, and $\ell((\boldsymbol{w},\boldsymbol{\theta});\xi)$ is the loss on a stochastic sample ξ drawn from $\mathcal{D}$. Since our focus is on learning the meta-parameter, we will simplify notations in the sequel and use $\mathcal{L}^i(\boldsymbol{w};\mathcal{D})$ for $\mathcal{L}((\boldsymbol{w},\boldsymbol{\theta}^i);\mathcal{D})$ and $\ell(\boldsymbol{w};\xi^i)$ for $\ell((\boldsymbol{w},\boldsymbol{\theta}^i);\xi^i)$.

In this paper, we focus on the family of MAML algorithms (Finn et al., 2017), in which the meta-parameter is used as a meta-initialization for $\boldsymbol{\theta}^i$'s. The outer loop (1) finds a suitable meta-initialization, while the inner loop (2) adapts $\boldsymbol{w}$ to each task. In many cases, the inner problem is complex and cannot be explicitly solved. Instead of finding the exact minimizer for $\mathcal{L}^i$, MAML performs only a single-step SGD: $\boldsymbol{\theta}^i = \boldsymbol{w} - \alpha\nabla\ell(\boldsymbol{w};\xi^i)$, where ξ^i is sampled from the training data $\mathcal{D}^i_{\text{tr}}$ of task i. This can also be naturally extended to K-step SGD with $K > 1$, which allows better adaptation.

In general, bilevel optimization is expensive. While the meta-gradient on $\boldsymbol{w}$ can be obtained from the implicit function theorem, it requires computing the Hessian matrix or its inverse. To alleviate this problem, methods such as FO-MAML (Finn et al., 2017) and Reptile (Nichol et al., 2018) propose to approximate the meta-gradient by using only the first-order information. Specifically, FOMAML simply uses the model gradient after task-specific adaptation as the meta-gradient, while Reptile uses the average gradient during adaptation. In practice, Reptile usually has a better empirical performance than MAML and FOMAML.

Another problem with the bilevel formulation of meta-learning is that it requires storing the $\boldsymbol{\theta}^i$'s of all the tasks. This can lead to a huge storage cost when the number of tasks is large and/or each $\boldsymbol{\theta}^i$ has a large number of parameters. To address this problem, methods like MAML, FO-MAML and Reptile do not explicitly keep the $\boldsymbol{\theta}^i$'s for all tasks. Instead, they directly use the $\boldsymbol{\theta}^i$'s (as a function of $\boldsymbol{w}$) in the outer objective, leading to the single-level optimization problem: $\min_{\boldsymbol{w}} \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathcal{L}(\boldsymbol{w}-\alpha\nabla\mathcal{L}(\boldsymbol{w};\mathcal{D}^i_{\text{tr}});\mathcal{D}^i_{\text{vald}})$.

Domain randomized search (DRS) (Gao & Sener, 2020), also called joint training (Finn et al., 2019), is another way to avoid the bilevel formulation by simply optimizing the losses of all tasks altogether. While it shows good results in some applications (Gao & Sener, 2020), DRS does not consider multi-step task adaptation and can have inferior performance than the other meta-learning algorithms, as will also be demonstrated empirically in Section 4.

2.2. Variance Reduction in Stochastic Optimization

Variance reduction (Gower et al., 2020) has been commonly used to reduce the variance in stochastic gradients and thus accelerate optimization. Pioneering works such as SAG (Roux et al., 2012), SDCA (Shalev-Shwartz & Zhang, 2013) and SVRG (Johnson & Zhang, 2013) are only applicable to strongly-convex problems. More recently, variance reduction methods for general non-convex problems are also developed (Allen-Zhu & Hazan, 2016; Nguyen et al., 2017; Fang et al., 2018). However, they still require the occasional computation of the batch gradient, which can be expensive on large training sets.

Recently, Cutkosky & Orabona (2019) propose a momentum-based variance reduction algorithm called STORM (Algorithm 1). It computes a variance-reduced gradient (step 6) by using only the stochastic gradients at two successive iterates ($\boldsymbol{w}_t$ and $\boldsymbol{w}_{t-1}$) on the same stochastic sample ξ_t, without requiring the time-consuming batch gradient computation. Note that it reduces to standard SGD when all γ_t's are set to one. Moreover, STORM (and variant STORM+ (Levy et al., 2021)) has the same asymptotic convergence rate as other variance reduction methods.

Algorithm 1 STORM (Cutkosky & Orabona, 2019).

Input: $\boldsymbol{w}_0$, step-size $\{\eta_t\}$, decay parameter $\{\gamma_t\}$.
$\boldsymbol{c}_0 = \nabla \ell(\boldsymbol{w}_0; \xi_0)$
$\boldsymbol{w}_1 = \boldsymbol{w}_0 - \eta_0 \boldsymbol{c}_0$
for $t = 1$ **to** $T-1$ **do**
  sample ξ_t
  $\boldsymbol{c}_t = \nabla \ell(\boldsymbol{w}_t; \xi_t) + (1-\gamma_t)(\boldsymbol{c}_{t-1} - \nabla \ell(\boldsymbol{w}_{t-1}; \xi_t))$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{c}_t$
end for

Algorithm 2 Reptile (Nichol et al., 2018)

Input: $\boldsymbol{w}_0$, step-size $\{\eta_t\}$ and α, number of local steps K.
for $t = 0$ **to** $T-1$ **do**
  sample tasks $\mathcal{I}_t \subset \mathcal{I}$
  for $i \in \mathcal{I}_t$ **do**
    $\boldsymbol{u}_0^i = \boldsymbol{w}_t$
    for $k = 0$ **to** $K-1$ **do**
      obtain samples $\xi_{k,t}^i$ from support data of task i
      $\boldsymbol{u}_{k+1}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,t}^i)$
    end for
    $\boldsymbol{c}_t^i = \frac{1}{K\alpha}(\boldsymbol{w}_t - \boldsymbol{u}_K^i)$
  end for
  $\boldsymbol{c}_t = \frac{1}{|\mathcal{I}_t|}\sum_{i \in \mathcal{I}_t} \boldsymbol{c}_t^i$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{c}_t$
end for

Very recently, momentum-based variance reduction has also been extended to stochastic bilevel optimization. Examples include SUSTAIN (Khanduri et al., 2021), MRBO/VRBO (Yang et al., 2021), RSVRB (Guo et al., 2021), and VR-BiAdam (Huang & Huang, 2021). With the help of variance reduction, these methods achieve the same asymptotic convergence rate and are faster than bilevel algorithms without variance reduction (Ji et al., 2021; Chen et al., 2021).

3. Variance Reduction for Meta-Learning

Recall from Section 2.1 that meta-learning can be formulated as either a bilevel or single-level optimization problem. In both cases, a straightforward approach to reduce variance in the stochastic gradients is to use the corresponding variance reduction methods. For example, when using the bilevel optimization formulation, one can use the recent methods in (Khanduri et al., 2021; Yang et al., 2021). However, recall that this demands a lot of storage and can be infeasible when the number of tasks is large. Moreover, to avoid overfitting the often limited data available in each task, the inner loop typically performs only a small number of gradient descent steps. For effective variance reduction, a sufficiently large number of steps is usually required (Allen-Zhu & Hazan, 2016; Cutkosky & Orabona, 2019).

Algorithm 3 VR-Reptile (Variance-Reduced Reptile).

Input: initial weight $\boldsymbol{w}_0$, stepsizes $\{\eta_t\}$ and α, number of local steps K, decay parameter $\{\gamma_t\}$.
sample tasks $\mathcal{I}_0 \subset \mathcal{I}$
for $i \in \mathcal{I}_0$ **do**
  $\boldsymbol{u}_0^i = \boldsymbol{w}_0$
  for $k = 0$ **to** $K-1$ **do**
    obtain samples $\xi_{k,0}^i$ from support data of task i
    $\boldsymbol{u}_{k+1}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i; \xi_{k,0}^i)$
  end for
  $\tilde{\boldsymbol{c}}_0^i = \frac{1}{K\alpha}(\boldsymbol{w}_0 - \boldsymbol{u}_K^i)$
end for
$\tilde{\boldsymbol{c}}_0 = \frac{1}{|\mathcal{I}_0|}\sum_{i \in \mathcal{I}_0} \tilde{\boldsymbol{c}}_0^i$
$\boldsymbol{w}_1 = \boldsymbol{w}_0 - \eta_0 \tilde{\boldsymbol{c}}_0$
for $t = 1$ **to** $T-1$ **do**
  sample tasks $\mathcal{I}_t \subset \mathcal{I}$
  for $i \in \mathcal{I}_t$ **do**
    $\boldsymbol{u}_0^i = \boldsymbol{w}_t$
    $\boldsymbol{v}_0^i = \boldsymbol{w}_{t-1}$
    for $k = 0$ **to** $K-1$ **do**
      obtain samples $\xi_{k,t}^i$ from support data of task i
      $\boldsymbol{u}_{k+1}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i; \xi_{k,t}^i)$
      $\boldsymbol{v}_{k+1}^i = \boldsymbol{v}_k^i - \alpha \nabla \ell(\boldsymbol{v}_k^i; \xi_{k,t}^i)$
    end for
    $\tilde{\boldsymbol{d}}_{t-1}^i = \frac{1}{K\alpha}(\boldsymbol{w}_{t-1} - \boldsymbol{v}_K^i)$
    $\tilde{\boldsymbol{c}}_t^i = \frac{1}{K\alpha}(\boldsymbol{w}_t - \boldsymbol{u}_K^i)$
  end for
  $\tilde{\boldsymbol{d}}_{t-1} = \frac{1}{|\mathcal{I}_t|}\sum_{i \in \mathcal{I}_t} \tilde{\boldsymbol{d}}_{t-1}^i$
  $\tilde{\boldsymbol{c}}_t = \frac{1}{|\mathcal{I}_t|}\sum_{i \in \mathcal{I}_t} \tilde{\boldsymbol{c}}_t^i + (1-\gamma_t)(\tilde{\boldsymbol{c}}_{t-1} - \tilde{\boldsymbol{d}}_{t-1})$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \tilde{\boldsymbol{c}}_t$
end for

Alternatively, by formulating meta-learning as a single-level optimization problem, one can use recent variance reduction methods such as STORM. Very recently, an initial attempt in this direction is VFML (Wang et al., 2021), which integrates STORM into Reptile. However, it lacks a formal study on its theoretical properties. Experiments in Section 4 also show that VFML has inferior performance.

In this section, we propose a novel variance reduction algorithm for meta-learning that avoids the pitfalls of directly applying existing variance reduction methods in bilevel or single-level optimization. The idea is to keep utilizing the double-loop structure in the bilevel formulation, which is known to more accurately capture the structure in meta-learning (Gao & Sener, 2020), while maintaining the efficiency of single-level optimization formulation that does not require storing a vast number of task-specific parameters.

Algorithm 4 VFML (Wang et al., 2021).

Input: $\boldsymbol{w}_0$, stepsizes $\{\eta_t\}$ and α, number of local steps K, decay parameters $\{\beta_t\}$ and $\{\gamma_t\}$.
sample tasks $\mathcal{I}_{-1} \subset \mathcal{I}$
for $i \in \mathcal{I}_{-1}$ **do**
  $\boldsymbol{m}_0^i = \nabla \ell(\boldsymbol{w}_0, \xi_{0,-1}^i)$
end for
$\boldsymbol{m}_0 = \frac{1}{|\mathcal{I}_{-1}|} \sum_{i \in \mathcal{I}_{-1}} \boldsymbol{m}_0^i$
for $t = 0$ **to** $T-1$ **do**
  sample tasks $\mathcal{I}_t \subset \mathcal{I}$
  for $i \in \mathcal{I}_t$ **do**
    $\bar{\boldsymbol{u}}_0^i = \boldsymbol{w}_t$
    for $k = 0$ **to** $K-1$ **do**
      $\bar{\boldsymbol{u}}_{k+1}^i = \bar{\boldsymbol{u}}_k^i - \alpha(\gamma_t \nabla \ell(\bar{\boldsymbol{u}}_k^i, \xi_{k,t}^i) + (1-\gamma_t)\boldsymbol{m}_t)$
    end for
    $\bar{\boldsymbol{m}}_t^i = \nabla \ell(\boldsymbol{w}_t, \xi_{K,t}^i)$
    $\bar{\boldsymbol{c}}_t^i = \frac{1}{K\alpha}(\boldsymbol{w}_t - \bar{\boldsymbol{u}}_K^i)$
  end for
  $\bar{\boldsymbol{c}}_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \boldsymbol{c}_t^i$
  $\bar{\boldsymbol{m}}_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \bar{\boldsymbol{m}}_t^i$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \bar{\boldsymbol{c}}_t$
  for $i \in \mathcal{I}_t$ **do**
    $\boldsymbol{m}_{t+1}^i = \nabla \ell(\boldsymbol{w}_{t+1}, \xi_{K,t}^i)$
  end for
  $\boldsymbol{m}_{t+1} = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \boldsymbol{m}_{t+1}^i + (1-\beta_t)(\boldsymbol{m}_t - \bar{\boldsymbol{m}}_t)$
end for

3.1. Proposed Algorithm

In the following, for easy comparison with VFML, we focus on Reptile (Algorithm 2). The proposed integration, which will be called VR-Reptile (variance-reduced Reptile), is shown in Algorithm 3. The integration with other first-order meta-learning algorithms (such as MAML, FOMAML and BMG (Flennerhag et al., 2022)) are analogous and are presented in Appendix B.1.

The main part of Algorithm 3 is in steps 13-29. Recall that STORM computes a variance-reduced gradient estimate by using the stochastic gradients at two successive iterates (step 6 of Algorithm 1). Applying this on the stochastic gradient update $\boldsymbol{c}_t$ in the outer loop of Reptile (step 13 in Algorithm 2), we obtain the update in step 27 of Algorithm 3. Here, $\tilde{\boldsymbol{d}}_{t-1}$ is the stochastic gradient estimate (analogous to $\tilde{\boldsymbol{c}}_t$ in Algorithm 3 or $\boldsymbol{c}_t$ in Reptile) but evaluated at the previous iterate ($\boldsymbol{w}_{t-1}$) on the current batch of samples $\xi_{k,t}^i$. The components $\tilde{\boldsymbol{c}}_t^i$'s (resp. $\tilde{\boldsymbol{d}}_t^i$'s) in $\tilde{\boldsymbol{c}}_t$ (resp. $\tilde{\boldsymbol{d}}_t$), which correspond to the local update on each task i in the inner loop, are computed at steps 16-24. Again, $\tilde{\boldsymbol{c}}_t^i$'s are based on $\boldsymbol{w}_t$, while $\tilde{\boldsymbol{d}}_t^i$'s are based on $\boldsymbol{w}_{t-1}$. Note that the bilevel optimization structure in meta-learning is still preserved. Moreover, it can be easily seen that when all γ_t's are set to 1, VR-Reptile reduces to Reptile, in the same manner as STORM reduces to standard SGD in this case.

In the implementation, we do not need to store the task-specific model parameters ($\boldsymbol{u}_K^i$'s and $\boldsymbol{v}_K^i$'s), as they are used only once to update $\tilde{\boldsymbol{c}}_t^i$'s and $\tilde{\boldsymbol{d}}_{t-1}^i$'s. Indeed, neither $\tilde{\boldsymbol{c}}_t^i$'s nor $\tilde{\boldsymbol{d}}_{t-1}^i$'s have to be stored separately, as they only need to be summed in the updates of $\tilde{\boldsymbol{c}}_t$ and $\tilde{\boldsymbol{d}}_{t-1}$. Thus, this is much more space-efficient than a direct application of the variance reduction methods for stochastic bilevel optimization, which requires storing all the task-specific parameters.

Remark 3.1. Recall that VFML (shown in Algorithm 4) also aims at integrating STORM into Reptile. However, it is very different from the proposed Algorithm 3. While Algorithm 3 computes a variance-reduced gradient estimate $\tilde{\boldsymbol{c}}_t$ for the update of target meta-parameter $\boldsymbol{w}$ (step 27), VFML computes a STORM-like variance-reduced estimate of the intermediate gradient $\nabla \ell(\boldsymbol{u}_k^i; \xi_{k,t}^i)$ that is used only by the local variable $\boldsymbol{u}_{k+1}^i$ (step 12 in Algorithm 4). However, the number of gradient descent steps performed in the inner loop is typically small and not enough for effective variance reduction. Moreover, performing variance reduction on an intermediate gradient makes theoretical analysis of VFML difficult.

3.2. Theoretical Analysis

In Section 3.2.1, we first study the (gradient of the) loss function that Reptile and the proposed VR-Reptile are implicitly minimizing. Section 3.2.2 then shows that VR-Reptile has a faster convergence rates than vanilla Reptile due to variance reduction. The analysis can be easily extended to show that VR-MAML/VR-FOMAML/VR-BMG also have faster convergence rate than vanilla MAML/FOMAML/BMG.

3.2.1. Loss Implicitly used in Reptile

For Reptile (Algorithms 2), let $\xi_{0:K-1,t}^i \equiv \{\xi_{0,t}^i, \ldots, \xi_{K-1,t}^i\}$ be the K i.i.d. samples from the training data $\mathcal{D}_{\text{tr}}^i$ of task i. At step 13, the meta-parameter $\boldsymbol{w}_t$ is updated with $\eta_t \boldsymbol{c}_t$, in which $\boldsymbol{c}_t$ is an average over $\boldsymbol{c}_t^i$'s from the sampled tasks in $\mathcal{I}_t$ (step 12). Note that $\boldsymbol{c}_t^i = \frac{1}{K} \sum_{k=0}^{K-1} \nabla \ell(\boldsymbol{u}_k^i; \xi_{k,t}^i)$. As each gradient $\nabla \ell(\boldsymbol{u}_k^i; \xi_{k,t}^i)$ is a gradient field and thus path-independent by the gradient theorem (Rudin, 1976), summing them together means $\boldsymbol{c}_t^i$ is also path-independent and thus a gradient field from the converse of gradient theorem. Thus, each $\boldsymbol{c}_t^i$ can be considered as the stochastic gradient of some loss $\tilde{\ell}$ on samples $\xi_{0:K-1,t}^i$ (i.e., $\boldsymbol{c}_t^i \equiv \nabla \tilde{\ell}(\boldsymbol{w}_t; \xi_{0:K-1,t}^i)$). Let $\nabla \tilde{\mathcal{L}}^i(\boldsymbol{w}_t) = \mathbb{E}_{\xi_{0:K-1,t}^i \sim \mathcal{D}_{\text{tr}}^i}[\nabla \tilde{\ell}(\boldsymbol{w}_t; \xi_{0:K-1,t}^i)]$. The Reptile update can then be viewed as the stochastic gradient on an "implicit" loss $\tilde{\mathcal{L}}$ with $\nabla \tilde{\mathcal{L}}(\boldsymbol{w}_t) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla \tilde{\mathcal{L}}^i(\boldsymbol{w}_t)$.

When $K = 1$, as can be seen from Reptile (Al-

gorithms 2), $\nabla\tilde{\ell}(\boldsymbol{w}_t;\xi^i_{0,t}) \equiv \boldsymbol{c}^i_t = \nabla\ell(\boldsymbol{w}_t;\xi^i_{0,t})$. Hence, $\nabla\tilde{\mathcal{L}}^i(\boldsymbol{w}_t) = \mathbb{E}_{\xi^i_{0,t}\sim\mathcal{D}^i_{\text{tr}}}\nabla\tilde{\ell}(\boldsymbol{w}_t;\xi^i_{0,t}) = \mathbb{E}_{\xi^i_{0,t}\sim\mathcal{D}^i_{\text{tr}}}\nabla\ell(\boldsymbol{w}_t;\xi^i_{0,t}) = \nabla\mathcal{L}(\boldsymbol{w}_t;\mathcal{D}^i_{\text{tr}})$, which matches the gradient of task i's training loss. This agrees with the fact that Reptile reduces to SGD on loss $\mathcal{L}$ in this case.

Similarly, for VR-Reptile, when $K = 1$, we have:

$$\begin{aligned}\tilde{\boldsymbol{c}}^i_t =&\nabla\tilde{\ell}(\boldsymbol{w}_t;\xi^i_{0,t}) = \nabla\ell(\boldsymbol{u}^i_0;\xi^i_{0,t}) = \nabla\ell(\boldsymbol{w}_t;\xi^i_{0,t}),\\ \tilde{\boldsymbol{d}}^i_{t-1} =&\nabla\tilde{\ell}(\boldsymbol{w}_{t-1};\xi^i_{0,t}) = \nabla\ell(\boldsymbol{v}^i_0;\xi^i_{0,t}) = \nabla\ell(\boldsymbol{w}_{t-1};\xi^i_{0,t}),\end{aligned}$$

which are the stochastic gradients at $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t-1}$ with the same stochastic sample $\xi^i_{0,t}$. Since $\xi^i_{0,t}$'s are sampled from different tasks and $\tilde{\boldsymbol{d}}_{t-1}$ is the average over all $\tilde{\boldsymbol{d}}^i_{t-1}$'s, $\tilde{\boldsymbol{d}}_{t-1}$ becomes the stochastic gradient of $\mathcal{L}$ at $\boldsymbol{w}_{t-1}$. $\tilde{\boldsymbol{c}}_t$ in step 27 (which is the same as step 6 in STORM (Algorithm 1)) then becomes the variance-reduced gradient of $\mathcal{L}$ at $\boldsymbol{w}_t$. Thus, while Reptile reduces to SGD when $K = 1$, VR-Reptile reduces to (the faster) STORM in this case.

3.2.2. Convergence Properties

First, we introduce the following smoothness assumption on the loss ℓ, which is commonly used in stochastic optimization algorithms (Cutkosky & Orabona, 2019).

Assumption 3.2. ℓ is M-Lipschitz smooth w.r.t. $\boldsymbol{w}$ (i.e., for any $\boldsymbol{w}, \boldsymbol{w}'$ and ξ, $\|\nabla\ell(\boldsymbol{w};\xi)-\nabla\ell(\boldsymbol{w}';\xi)\| \le M\|\boldsymbol{w}-\boldsymbol{w}'\|$).

The following Proposition shows that the implicit loss $\tilde{\mathcal{L}}$ is also Lipschitz-smooth. All the proofs are in Appendix A.

Proposition 3.3. *$\tilde{\ell}$ is $\tilde{M}$-Lipschitz-smooth w.r.t. $\boldsymbol{w}_t$, where $\tilde{M} = (1+\alpha M)^K/(\alpha K)$.*

Corollary 3.4. *$\tilde{\mathcal{L}}^i(\boldsymbol{w})$ and $\tilde{\mathcal{L}}(\boldsymbol{w})$ are $\tilde{M}$-Lipschitz smooth.*

Next, we decompose the variance of $\boldsymbol{c}^i_t$ (which is equal to $\frac{1}{K}\sum_{k=0}^{K-1}\nabla\ell(\boldsymbol{u}^i_k;\xi^i_{k,t})$) into two parts.

$$\begin{aligned}&\mathbb{E}_i\mathbb{E}_{\xi^i_{0:K-1,t}\sim\mathcal{D}^i_{\text{tr}}}\|\boldsymbol{c}^i_t - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\\ =\;& \mathbb{E}_i\left[\mathbb{E}_{\xi^i_{0:K-1,t}\sim\mathcal{D}^i_{\text{tr}}}\|\boldsymbol{c}^i_t - \nabla\tilde{\mathcal{L}}^i(\boldsymbol{w}_t)\|^2\right]\\ &+\mathbb{E}_i\|\nabla\tilde{\mathcal{L}}^i(\boldsymbol{w}_t) - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2. \qquad (3)\end{aligned}$$

The first part is due to data sampling, while the second part is due to task sampling. To bound the first part, we assume that the stochastic variance of $\nabla\ell(\boldsymbol{w};\xi)$ is bounded. This is also commonly assumed in stochastic gradient methods (Ghadimi & Lan, 2013; Cutkosky & Orabona, 2019).

Assumption 3.5. For any $\boldsymbol{w}$ and task i, there exists a constant σ^2 such that $\mathbb{E}_{\xi\sim\mathcal{D}^i_{\text{tr}}}\|\nabla\ell(\boldsymbol{w};\xi) - \nabla\mathcal{L}^i(\boldsymbol{w})\|^2 \le \sigma^2$.

For simplicity, we assume the same σ^2 for all tasks. This can be easily extended to the case where different tasks have difference variance bounds.

The following Proposition bounds the variance of the first part in (3). When $K = 1$, it reduces to the condition in Assumption 3.5.

Proposition 3.6. *Define*

$$\tilde{\varsigma}^2 = \frac{2^K(1+M^2\alpha^2)^K - 1}{K(1+2M^2\alpha^2)}\frac{2M^2\alpha^2\sigma^2}{1+2M^2\alpha^2} + \frac{\sigma^2}{1+2M^2\alpha^2}.$$

For any $\boldsymbol{w}_t$ and task i, we have $\mathbb{E}_{\xi^i_{0:K-1,t}\sim\mathcal{D}^i_{tr}}\|\boldsymbol{c}^i_t - \nabla\tilde{\mathcal{L}}^i(\boldsymbol{w}_t)\|^2 \le \tilde{\varsigma}^2$. When $K = 1$, we have $\tilde{\varsigma} = \sigma$.

For effective meta-learning, we assume that the tasks should not differ too much, as in (Fallah et al., 2020; Ji et al., 2020).

Assumption 3.7. There exists a positive constant δ^2 such that for any $\boldsymbol{w}$ and two different tasks i, j, $\|\nabla\mathcal{L}^i(\boldsymbol{w}) - \nabla\mathcal{L}^j(\boldsymbol{w})\|^2 \le \delta^2$.

Proposition 3.8. *Define*

$$\begin{aligned}\tilde{\delta}^2 =&2\delta^2 + \frac{(1+4KM^2\alpha^2)^K}{4K^2M^2\alpha^2}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2)\\ &+8KM^2\alpha^2(\delta^2+2\sigma^2).\end{aligned}$$

For any $\boldsymbol{w}$ and tasks i, j, we have $\|\nabla\tilde{\mathcal{L}}^i(\boldsymbol{w}_t) - \nabla\tilde{\mathcal{L}}^j(\boldsymbol{w}_t)\|^2 \le \tilde{\delta}^2$.

This can then be used to bound the second term in (3).

Corollary 3.9. *If tasks i's are uniformly sampled from $\mathcal{I}$, then $\mathbb{E}_i\|\nabla\tilde{\mathcal{L}}^i(\boldsymbol{w}_t) - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 \le \tilde{\delta}^2$, where the expectation is taken over the set of training tasks.*

Combining Proposition 3.6 and Corollary 3.8, we have:

Corollary 3.10. *$\mathbb{E}_i\mathbb{E}_{\xi^i_0,\ldots,\xi^i_{K-1}\sim\mathcal{D}^i_{tr}}\|\boldsymbol{c}^i_t - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 \le \tilde{\sigma}^2$, where $\tilde{\sigma}^2 \equiv \tilde{\varsigma}^2 + \tilde{\delta}^2$.*

With a suitable α, $\tilde{M}$ can be upper-bounded by a constant not depending on K. Consequently, $\tilde{\delta}^2$ also grows linearly (but not exponentially) with K.

Corollary 3.11. *With $\alpha = \frac{1}{KM}$, $\tilde{M} = M(1+\frac{1}{K})^K \le eM$ and $\tilde{\delta}^2 = 2\delta^2 + \frac{(1+\frac{4}{K})^K}{4}(1+KM)(\delta^2+2\sigma^2) + \frac{8}{K}(\delta^2+2\sigma^2) \le 2\delta^2 + (\frac{e^4}{4}+8+\frac{eKM}{4})(\delta^2+2\sigma^2)$.*

The following Theorem shows convergence rate for Reptile on the implicit loss.

Theorem 3.12. *For any $\eta_0 > 0$, set $\eta_t = \frac{\eta_0}{(1+t)^{1/2}}$, then Reptile (Algorithm 2) satisfies:*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\right] \le \frac{\sqrt{2}G_1/\eta_0}{\sqrt{T}} + \frac{\sqrt{2}G_1/\eta_0}{T},$$

where $G_1 = \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0) - \tilde{\mathcal{L}}^*] + \tilde{M}\tilde{\sigma}^2\eta_0^2\ln(T+1)$ and $\tilde{\mathcal{L}}^* = \min_{\boldsymbol{w}}\tilde{\mathcal{L}}(\boldsymbol{w})$.

Table 1. Statistics for the data sets used.

		number of classes			
		training	validation	testing	#samples per class
Meta-Dataset	bird	64	16	20	60
	texture	30	7	10	120
	aircraft	64	16	20	100
	fungi	64	16	20	150
Mini-Imagenet		64	16	20	600

Theorem 3.12 matches the asymptotic convergence rate for SGD on problems with non-convex objectives (Ghadimi & Lan, 2013). The difference is that we prove the convergence w.r.t. $\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2$ (i.e., the gradient norm on the implicit loss) instead of $\|\nabla\mathcal{L}(\boldsymbol{w}_t)\|^2$.

The following Theorem shows that with the use of variance reduction, VR-Reptile has a faster convergence rate than Reptile. The bound also matches the convergence rate in STORM (Theorems 1 and 2 in (Cutkosky & Orabona, 2019)).

Theorem 3.13. *Let* $\eta_t = 1/(4\tilde{M}(t+(\frac{65}{28})^3)^{1/3})$ *and* $\gamma_{t+1} = \frac{65}{28}/(t+(\frac{65}{28})^3)^{1/3}$. *Algorithm 3 satisfies:*

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\right] \leq \frac{4\tilde{M}G_2}{T^{2/3}} + \frac{65}{7}\cdot\frac{\tilde{M}G_2}{T},$$

where $G_2 = 8\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0) - \tilde{\mathcal{L}}^*] + \frac{\tilde{\sigma}^2}{\tilde{M}}\left(\frac{65}{28} + \frac{4225}{392}\ln T\right)$.

Table 2. Number of outer-loop training iterations on the data sets.

		1-shot 5-way	5-shot 5-way
Meta-Dataset	Bird	60,000	20,000
	Texture	40,000	30,000
	Aircraft	20,000	20,000
	Fungi	60,000	20,000
Mini-Imagenet		60,000	60,000

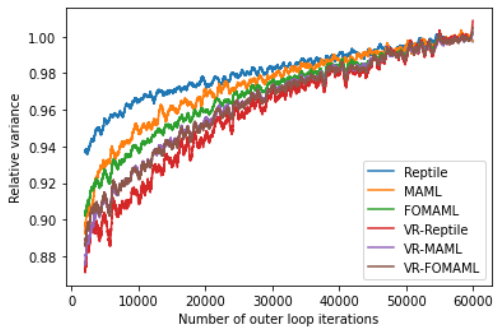


Figure 1. Variance of updates during training.

4. Experiments

As in (Finn et al., 2017; Nichol et al., 2018), we perform meta-learning experiments in the few-shot image classification setting. Experiments are performed on Mini-Imagenet (Vinyals et al., 2016; Ravi & Larochelle, 2017) and Meta-Dataset (Triantafillou et al., 2020). Meta-Dataset has been popularly used for meta-learning. It consists of image classification data sets from different domains. In this experiment, we use (i) bird, (ii) texture, (iii) aircraft, and (iv) fungi as in (Yao et al., 2019). While the Meta-Dataset focuses on fine-grained classification, Mini-Imagenet contains more diverse images. A summary of these data sets is in Table 1. Experiments are performed in the 1-shot 5-way and 5-shot 5-way settings.

Following (Finn et al., 2017; Nichol et al., 2018), we use the CONV4 model as base learner. It is a 4-layer CNN. Each layer contains 64 3×3 convolutional filters, followed by batch normalization, ReLU activation, and 2×2 max-pooling. For all data sets, the hyper-parameter settings follow Reptile (Nichol et al., 2018): we use vanilla SGD as the optimizer for the outer loop, and Adam for the inner loop. The learning rate for SGD is 1, and no momentum is used. The learning rate for Adam is 0.001, the first-order momentum weight is 0, and the second-order momentum weight is 0.99. The number of gradient descent steps K in the inner loop is 5. The number of iterations in the outer loop (T in Algorithm 3) is shown in Table 2.

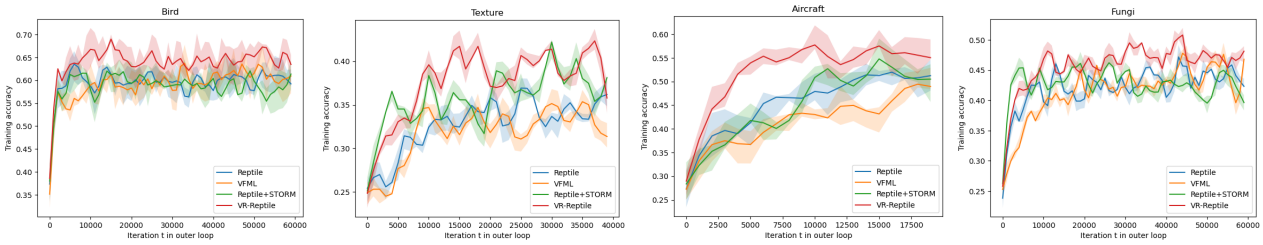
The following groups of meta-learning baselines are compared: (i) Standard single-level optimization methods, including (a) MAML, (b) FOMAML, (c) Reptile, (d) BMG, and (e) DRS (Gao & Sener, 2020), which optimizes the average loss of all tasks; (ii) Variants of the first group[1] (denoted MAML+STORM, FOMAML+STORM, Reptile+STORM, BMG+STORM and DRS+STORM), in which the stochastic gradients are replaced by the variance-reduced counterparts obtained with STORM; (iii) Variants of the first group integrated with the proposed method (denoted VR-MAML, VR-FOMAML, VR-Reptile and VR-BMG);[2] (iv) VFML (Wang

[1]These algorithms are shown in Appendix B.2.

[2]VR-MAML, VR-FOMAML and VR-BMG are shown in Appendix B.1. DRS does not use a bilevel structure, and so cannot be

Table 3. Classification accuracy (%) on mini-ImageNet.

	var reduction	single/bilevel	1-shot 5-way	5-shot 5-way
MAML	×	single	48.7±1.8	63.1±0.9
FOMAML	×	single	48.1±1.8	63.2±0.9
Reptile	×	single	50.0±0.3	66.0±0.6
BMG	×	single	50.7±0.5	65.6±0.6
DRS	×	single	24.5±0.8	30.4±0.6
ProtoNet	×	single	49.4±0.8	68.2±0.7
MAML+STORM	√	single	47.9±1.4	61.6±1.2
FOMAML+STORM	√	single	48.0±1.6	63.4±1.1
Reptile+STORM	√	single	49.9±0.3	66.2±0.3
BMG+STORM	√	single	46.7±0.6	60.9±0.8
DRS+STORM	√	single	24.7±1.1	30.3±0.7
VR-MAML	√	single	49.2±1.4	63.6±0.8
VR-FOMAML	√	single	48.3±1.2	63.4±0.6
VR-Reptile	√	single	50.4±0.4	67.6±0.8
VR-BMG	√	single	**51.4±0.3**	**68.4±0.6**
VFML	√	single	49.6±0.5	66.2±0.8
ANIL	×	bilevel	46.9±0.4	61.4±0.2
ANIL+SUSTAIN	√	bilevel	47.0±0.4	61.8±0.3
ANIL+MRBO	√	bilevel	47.2±0.5	62.0±0.2
ANIL+VRBO	√	bilevel	47.2±0.4	61.9±0.2



(a) Bird. (b) Texture. (c) Aircraft. (d) Fungi.

Figure 2. Training accuracy with number of outer-loop iterations on Meta-Dataset in 1-shot 5-way setting.

et al., 2021); (v) ANIL (Raghu et al., 2020), which uses the bilevel formulation for meta-learning. As the bilevel formulation needs to keep $\boldsymbol{\theta}^i$'s for all tasks i, it is infeasible to use the whole CONV4 model as $\boldsymbol{\theta}^i$.[3] To alleviate this problem, we only adapt parameters in the last layer of CONV4; (vi) variants of ANIL using a straightforward combination with variance reduction methods for bilevel optimization, including (a) SUSTAIN (Khanduri et al., 2021), (b) MRBO (Yang et al., 2021), and (c) VRBO (Yang et al., 2021).

For performance evaluation, we follow (Finn et al., 2017; Nichol et al., 2018) and report the average accuracy over 1,000 5-way classification tasks randomly sampled from its meta-testing set. Each method is repeated 3 times.

integrated with the proposed method.

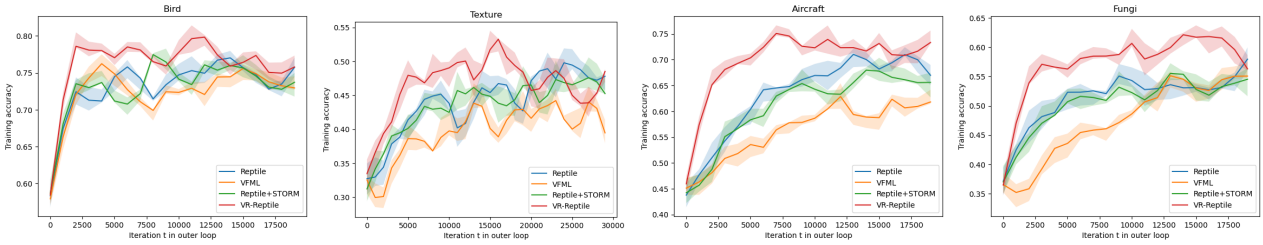
[3]For example, Mini-Imagenet has 64 classes, and about 7.6×10^6 5-way classification tasks in the meta-training set. Assume that the deep network has only 0.1M parameters (which is small), the task models take a total of $7.6 \times 10^6 \times 0.1$M=760G memory.

4.1. Results on Mini-Imagenet

Table 3 shows the testing accuracies of the various methods in the 1-shot and 5-shot settings. As can be seen, integrating meta-learning algorithms with any of the variance reduction methods generally leads to better performance. In particular, the proposed VR-BMG achieves the best overall performance. The superiority of the VR variants over the STORM variants demonstrates that explicitly considering the double-loop meta-learning structure is useful. On the other hand, ANIL and its variance-reduced variants, which are based on the bilevel formulation, perform less well than those using the single-level formulation. As ANIL can only adapt part of its model in order to be computationally

Table 4. Classification accuracy (%) on meta-testing set from Meta-Dataset in 1-shot 5-way classification.

	var reduction	single/bi-level	Bird	Texture	Aircraft	Fungi
MAML	×	single	57.44	33.74	**57.98**	41.80
FOMAML	×	single	56.14	31.48	56.94	39.70
Reptile	×	single	60.82	34.66	54.08	42.84
BMG	×	single	60.96	34.78	54.16	42.76
DRS	×	single	33.64	23.82	28.24	24.96
ProtoNet	×	single	60.78	34.66	56.62	40.24
MAML+STORM	√	single	57.16	33.78	57.52	41.88
FOMAML+STORM	√	single	55.94	31.52	57.16	39.48
Reptile+STORM	√	single	61.02	34.84	56.48	43.56
BMG+STORM	√	single	56.22	31.18	53.88	42.04
DRS+STORM	√	single	33.78	23.96	28.36	25.12
VR-MAML	√	single	57.62	34.04	57.88	41.66
VR-FOMAML	√	single	57.08	31.48	56.46	39.88
VR-Reptile	√	single	62.04	35.10	57.54	**45.34**
VR-BMG	√	single	**62.14**	**35.24**	57.68	45.26
VFML	√	single	61.32	34.32	53.94	42.88
ANIL	×	bilevel	56.82	32.68	56.84	41.84
ANIL+SUSTAIN	√	bilevel	56.78	32.74	56.88	41.92
ANIL+MRBO	√	bilevel	56.74	32.76	56.92	41.94
ANIL+VRBO	√	bilevel	56.96	32.90	56.94	42.06



(a) Bird. (b) Texture. (c) Aircraft. (d) Fungi.

Figure 3. Training accuracy with number of outer-loop iterations on Meta-Dataset in 5-shot 5-way setting.

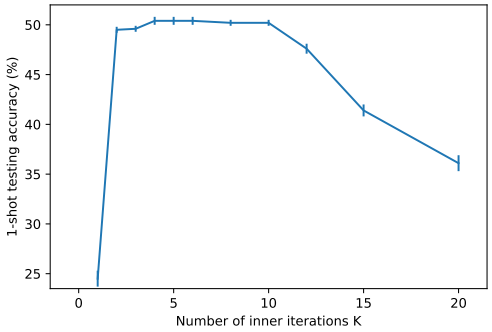


Figure 4. Testing accuracy (%) of VR-Reptile on 1-shot 5-way mini-ImageNet with different K's.

feasible, this greatly limits model flexibility, leading to inferior performance. DRS and DRS+STORM are also much inferior than the other baselines, suggesting that accurate modeling the structure of meta-learning is more important than the ease of optimization. Also, VFML generally has worse performance than VR-Reptile, as has been discussed in Remark 3.1.

Next, we demonstrate that the proposed method can reduce the variance. We focus on three gradient-based meta-learning algorithms, MAML/FOMAML/Reptile, and their variance-reduced variants VR-MAML/VR-FOMAML/VR-Reptile. Figure 1 shows $\frac{\mathbb{E}\|\tilde{\boldsymbol{c}}_t - \mathbb{E}[\tilde{\boldsymbol{c}}_t]\|^2}{\|\mathbb{E}[\tilde{\boldsymbol{c}}_t]\|^2}$, the variance of weight update $\tilde{\boldsymbol{c}}_t$ (relative to its squared norm), for different methods on Mini-ImageNet in the 1-shot 5-way setting. As expected, the update variance becomes smaller with variance reduction. MAML and FOMAML have smaller gradient variance than Reptile, which explains the smaller improvements for variance reduction on MAML/FOMAML than on Reptile.

Table 5. Classification accuracy (%) on meta-testing set from Meta-Dataset in 5-shot 5-way classification.

	var reduction	single/bilevel	Bird	Texture	Aircraft	Fungi
MAML	×	single	74.56	45.68	69.06	53.68
FOMAML	×	single	73.64	42.82	66.38	52.18
Reptile	×	single	74.60	43.26	66.46	52.88
BMG	×	single	74.52	43.74	66.64	53.02
DRS	×	single	53.34	33.28	41.06	37.64
ProtoNet	×	single	74.22	**49.86**	71.38	53.94
MAML+STORM	√	single	74.86	45.26	68.48	53.72
FOMAML+STORM	√	single	73.72	42.78	66.22	52.28
Reptile+STORM	√	single	75.24	44.60	67.48	52.54
BMG+STORM	√	single	73.16	42.32	66.46	51.74
DRS+STORM	√	single	53.48	33.42	41.12	37.72
VR-MAML	√	single	75.06	46.18	68.36	53.86
VR-FOMAML	√	single	74.28	43.28	66.98	52.16
VR-Reptile	√	single	76.48	46.94	**71.62**	54.24
VR-BMG	√	single	**76.56**	47.28	71.48	**54.38**
VFML	√	single	74.38	44.48	65.64	52.76
ANIL	×	bilevel	73.68	41.96	68.74	52.84
ANIL+SUSTAIN	√	bilevel	73.74	42.12	68.82	52.78
ANIL+MRBO	√	bilevel	73.78	42.18	68.78	52.86
ANIL+VRBO	√	bilevel	73.88	42.22	68.74	52.82

4.2. Results on Meta-Dataset

Tables 4 and 5 shows the testing accuracies in the 1-shot and 5-shot settings, respectively, for the four few-shot data sets in Meta-Dataset. The observations are generally similar to those on mini-ImageNet in Section 4.1. The integration of variance reduction into different meta-learning algorithms leads to best performance overall, as is demonstrated by VR-BMG and VR-Reptile. Moreover, meta-learning methods based on the single-level formulation have better performance than methods based on the bilevel formulation in general.

Figures 2 and 3 show the training accuracy with the number of outer-loop iterations (in (1)) for the 1-shot and 5-shot settings. To reduce clutterness, we only show results for Reptile and related methods (VFML, Reptile+STORM and VR-Reptile). As can be seen, VR-Reptile has much faster convergence than the other baselines, showing the benefits of variance reduction and verifies Theorem 3.13. On the other hand, Reptile+STORM does not show faster convergence as compared to Reptile. This is because the inner loop only involves a small number of gradient descent steps, while variance reduction methods like STORM typically require a sufficiently large number of steps to be effective. VFML converges even slower than Reptile in most cases. This can partly be attributed to its lack of theoretical study on convergence properties.

4.3. Ablation Study: Number of Inner Iterations K

Finally, we study the influence of K (number of inner-loop iterations) on VR-Reptile. We use the Mini-ImageNet data set under the 1-shot 5-way setting. The testing accuracies with different K's are shown in Figure 4. As can be seen, having a K too small or too large lead to inferior performance, and the performance with $4 \leq K \leq 10$ are very similar. To be consistent with (Finn et al., 2017; Nichol et al., 2018), we set $K = 5$ in all previous experiments.

5. Conclusion

In this paper, we propose a novel variance reduction method VR-Reptile to accelerate convergence of meta-learning. VR-Reptile utilizes the double-loop structure of meta-learning algorithms, but does not require storing the task-specific parameters. Theoretical results demonstrate that VR-Reptile has a faster convergence rate than Reptile due to variance reduction. Experiments on benchmark few-shot classification data sets demonstrate its superiority over meta-learning algorithms with and without variance reduction.

Acknowledgements

This research was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grant 16200021).

References

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, 2016.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, 2021.

Clavera, I., Nagabandi, A., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., and Finn, C. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *International Conference on Learning Representations*, 2019.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, 2019.

Elsken, T., Staffler, B., Metzen, J. H., and Hutter, F. Meta-learning of neural architectures for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, 2018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. In *International Conference on Machine Learning*, 2019.

Flennerhag, S., Schroecker, Y., Zahavy, T., van Hasselt, H., Silver, D., and Singh, S. Bootstrapped Meta-Learning. In *International Conference on Learning Representations*, 2022.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, 2018.

Gao, K. and Sener, O. Modeling and optimization trade-off in meta-learning. In *Advances in Neural Information Processing Systems*, 2020.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Gower, R. M., Schmidt, M., Bach, F., and Richtárik, P. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Guo, Z., Hu, Q., Zhang, L., and Yang, T. Randomized Stochastic Variance-Reduced Methods for Multi-Task Stochastic Bilevel Optimization. Technical Report arXiv:2105.02266, 2021.

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., and Levine, S. Meta-reinforcement learning of structured exploration strategies. In *Advances in Neural Information Processing Systems*, 2018.

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):1–20, 2021.

Huang, F. and Huang, H. BiAdam: Fast Adaptive Bilevel Optimization Methods. Technical Report arXiv:2106.11396, 2021.

Ji, K., Yang, J., and Liang, Y. Theoretical Convergence of Multi-Step Model-Agnostic Meta-Learning. Technical Report arXiv:2002.07836, 2020.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, 2021.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

Khanduri, P., Zeng, S., Hong, M., Wai, H. T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In *Advances in Neural Information Processing Systems*, 2021.

Levy, K. Y., Kavis, A., and Cevher, V. STORM+: Fully adaptive SGD with recursive momentum for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2021.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, 2017.

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

Raghu, A., Raghu, M., Bengio, S., and Vinyals, O. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In *International Conference on Learning Representations*, 2020.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

Ren, Z., Yeh, R., and Schwing, A. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*, 2012.

Rudin, W. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

Shalev-Shwartz, S. and Zhang, T. Accelerated mini-batch stochastic dual coordinate ascent. In *Advances in Neural Information Processing Systems*, 2013.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., and Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019.

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

Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Evci, U., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P.-A., and Larochelle, H. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

Wang, L., Huang, K., Ma, T., Gu, Q., and Huang, J. Variance-reduced first-order meta-learning for natural language processing tasks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. In *Advances in Neural Information Processing Systems*, 2021.

Yao, H., Wei, Y., Huang, J., and Li, Z. Hierarchically structured meta-learning. In *International Conference on Machine Learning*, 2019.

A. Proofs

A.1. Proof of Proposition 3.3

We first construct the following two sequences from $\boldsymbol{w}_t$ and $\boldsymbol{w}'_t$.

$$\begin{aligned}\boldsymbol{u}_0^i &= \boldsymbol{w}_t, \boldsymbol{u}_{k+1,t}^i = \boldsymbol{u}_k^i - \alpha\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i),\\ \boldsymbol{u}_0^{i'} &= \boldsymbol{w}'_t, \boldsymbol{u}_{k+1,t}^{i'} = \boldsymbol{u}_k^{i'} - \alpha\nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i).\end{aligned}$$

From the definition of $\nabla\tilde{\ell}$ in Section 3.2.1, we have:

$$\begin{aligned}\|\nabla\tilde{\ell}(\boldsymbol{w}_t, \xi_{0:K-1,t}^i) - \nabla\tilde{\ell}(\boldsymbol{w}'_t, \xi_{0:K-1,t}^i)\| =& \|\frac{1}{K}\sum_{k=0}^{K-1}\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i) - \frac{1}{K}\sum_{k=0}^{K-1}\nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i)\|\\ =& \|\frac{1}{K}\sum_{k=0}^{K-1}(\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i) - \nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i))\|\\ \leq& \frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i) - \nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i)\|\\ \leq& \frac{M}{K}\sum_{k=0}^{K-1}\|\boldsymbol{u}_k^i - \boldsymbol{u}_k^{i'}\|,\end{aligned}$$

where the last inequality comes from Assumption 3.2. Now we have to bound $\|\boldsymbol{u}_k^i - \boldsymbol{u}_k^{i'}\|$ for $k = 0, \ldots, K-1$ in terms of $\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|$. For $k = 0$, obviously we have $\|\boldsymbol{u}_0^i - \boldsymbol{u}_0^{i'}\| = \|\boldsymbol{w}_t - \boldsymbol{w}'_t\|$. For $k > 0$, we have the following induction:

$$\begin{aligned}\|\boldsymbol{u}_{k+1}^i - \boldsymbol{u}_{k+1}^{i'}\| =& \|\left(\boldsymbol{u}_k^i - \alpha\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i)\right) - \left(\boldsymbol{u}_k^{i'} - \alpha\nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i)\right)\|\\ =& \|(\boldsymbol{u}_k^i - \boldsymbol{u}_k^{i'}) - \left(\alpha\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i) - \alpha\nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i)\right)\|\\ \leq& \|\boldsymbol{u}_k^i - \boldsymbol{u}_k^{i'}\| + \alpha\|\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i) - \nabla\ell(\boldsymbol{u}_k^{i'}, \xi_{k,t}^i)\|\\ \leq& (1+\alpha M)\|\boldsymbol{u}_k^i - \boldsymbol{u}_k^{i'}\|,\end{aligned}$$

where the last inequality comes from Assumption 3.2. This then gives:

$$\|\boldsymbol{u}_k^i - \boldsymbol{u}_k^{i'}\| \leq (1+\alpha M)^k\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|, \quad k = 0, \ldots, K-1.$$

Summing over k, we have:

$$\begin{aligned}\|\nabla\tilde{\ell}(\boldsymbol{w}_t, \xi_{0:K-1,t}^i) - \nabla\tilde{\ell}(\boldsymbol{w}'_t, \xi_{0:K-1,t}^i)\| \leq& \frac{M}{K}\sum_{k=0}^{K-1}(1+\alpha M)^k\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|\\ =& \frac{M}{K}\frac{(1+\alpha M)^K - 1}{\alpha M}\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|\\ \leq& \frac{(1+\alpha M)^K}{\alpha K}\|\boldsymbol{w}_t - \boldsymbol{w}'_t\|.\end{aligned}$$

Setting $\tilde{M} = \frac{(1+\alpha M)^K}{\alpha K}$ then gives the desired result.

A.2. Proof of Proposition 3.6

From the definition of $\mathbb{E}_{\xi_{0,t}^i, \ldots, \xi_{K-1,t}^i \sim \mathcal{D}^i}[\boldsymbol{c}_t^i]$ in Section 3.2.1, we have:

$$\begin{aligned}\mathbb{E}_{\xi_{0,t}^i, \ldots, \xi_{K-1,t}^i \sim \mathcal{D}^i}[\boldsymbol{c}_t^i] &= \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi_{0,t}^i, \ldots, \xi_{K-1,t}^i \sim \mathcal{D}^i}[\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i)],\\ \boldsymbol{u}_0^i &= \boldsymbol{w}_t, \boldsymbol{u}_{k+1,t}^i = \boldsymbol{u}_k^i - \alpha\nabla\ell(\boldsymbol{u}_k^i, \xi_{k,t}^i)\end{aligned} \tag{4}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{c}^i_t-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{c}^i_t]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\frac{1}{K}\sum_{k=0}^{K-1}\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
\le&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2.
\end{aligned}$$

Now we need to bound $\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{k,t}}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{k,t}}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2$ for each $k=0,\dots,K-1$, for which we have:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)+\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)\|^2+\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2,
\end{aligned}\tag{5}$$

where the last equality is obtained from the fact that $\xi^i_{k,t}$ and $\boldsymbol{u}^i_k$ are independent, as all these stochastic samples $\xi^i_{0,t},\dots,\xi^i_{k,t}$ are i.i.d..

For the first term in (5), from Assumption 3.5, we have:

$$\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)\|^2=\mathbb{E}_{\xi^i_{k,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)\|^2\le\sigma^2.\tag{6}$$

For the second term in (5), when $k=0$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_0)-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{k,t}}[\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t}}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_0)-\mathbb{E}_{\xi^i_{0,t}}[\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t}}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_0)-\nabla\mathcal{L}^i(\boldsymbol{u}^i_0)\|^2=0.
\end{aligned}$$

When $k>0$, $\boldsymbol{u}^i_k$ depends on $\xi^i_{0,t},\dots,\xi^i_{k-1,t}$. Define $\bar{\boldsymbol{u}}^i_k$ that satisfies:

$$\bar{\boldsymbol{u}}^i_0=\boldsymbol{u}^i_0=\boldsymbol{w}_t,\bar{\boldsymbol{u}}^i_k=\qquad\qquad\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{k-1,t}\sim\mathcal{D}^i}[\boldsymbol{u}^i_k].\tag{7}$$

This gives:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\left(\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)\right)-\left(\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)]-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)\right)\|^2,
\end{aligned}$$

which can be seen as the variance of $\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)$ w.r.t. stochastic samples $\xi^i_{0,t},\dots,\xi^i_{k-1,t}$, as the other stochastic samples do not affect $\boldsymbol{u}^i_k$. With this, we have:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)\|^2-\|\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)]-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)\|^2\\
\le&\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)\|^2.
\end{aligned}$$

From Assumption 3.2, we have:

$$\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\nabla\mathcal{L}^i(\bar{\boldsymbol{u}}^i_k)\|^2\le M^2\mathbb{E}_{\xi^i_{0,t},\dots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_k-\bar{\boldsymbol{u}}^i_k\|^2.\tag{8}$$

For $\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_k-\bar{\boldsymbol{u}}^i_k\|^2$, from the definitions of $\boldsymbol{u}^i_k$ and $\bar{\boldsymbol{u}}^i_k$ in (4) and (7), we have:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_k-\bar{\boldsymbol{u}}^i_k\|^2\\
=&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_{k-1}-\alpha\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{u}^i_{k-1}-\alpha\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})]\|^2\\
\leq&2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_{k-1}-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{u}^i_{k-1}]\|^2\\
&+2\alpha^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})]\|^2\\
=&2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_{k-1}-\bar{\boldsymbol{u}}^i_{k-1}\|^2\\
&+2\alpha^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})]\|^2, \qquad (9)
\end{aligned}$$

where the last term is what we want to bound, except that here we have $k-1$ instead of k. Combining (6), (8) and (9) in (5), we have:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
\leq&\sigma^2+\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
\leq&\sigma^2+M^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_k-\bar{\boldsymbol{u}}^i_k\|^2\\
\leq&\sigma^2+2M^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_{k-1}-\bar{\boldsymbol{u}}^i_{k-1}\|^2\\
&+2M^2\alpha^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})]\|^2. \qquad (10)
\end{aligned}$$

Finally, we rewrite the two bounds in (9) and (10) as:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_k-\bar{\boldsymbol{u}}^i_k\|^2\\
\leq&2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_{k-1}-\bar{\boldsymbol{u}}^i_{k-1}\|^2+2\alpha^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})]\|^2,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
\leq&\sigma^2+2M^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{u}^i_{k-1}-\bar{\boldsymbol{u}}^i_{k-1}\|^2\\
&+2M^2\alpha^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})]\|^2.
\end{aligned}$$

Using the fact that

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})]\|^2\\
=&\mathbb{E}_{\xi^i_{0,t}}\|\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})-\mathbb{E}_{\xi^i_{0,t}}[\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})]\|^2=\mathbb{E}_{\xi^i_{0,t}}\|\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t})-\nabla\mathcal{L}^i(\boldsymbol{u}^i_0)\|^2\leq\sigma^2,
\end{aligned}$$

and

$$\mathbb{E}_{\xi^i_{0,t}}\|\boldsymbol{u}^i_0-\mathbb{E}_{\xi^i_{0,t}}[\boldsymbol{u}^i_0]\|^2=\mathbb{E}_{\xi^i_{0,t}}\|\boldsymbol{u}^i_0-\boldsymbol{u}^i_0\|^2=0,$$

we can obtain that:

$$\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\leq2^k(1+M^2\alpha^2)^k\frac{2M^2\alpha^2\sigma^2}{1+2M^2\alpha^2}+\frac{\sigma^2}{1+2M^2\alpha^2}.$$

Summing k from 0 to $K-1$, we have:

$$\begin{aligned}
&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\boldsymbol{c}^i_t-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{c}^i_t]\|^2\\
\leq&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}\|\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})-\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]\|^2\\
\leq&\frac{2^K(1+M^2\alpha^2)^K-1}{K(1+2M^2\alpha^2)}\frac{2M^2\alpha^2\sigma^2}{1+2M^2\alpha^2}+\frac{\sigma^2}{1+2M^2\alpha^2}.
\end{aligned}$$

Finally, define

$$\tilde{\varsigma}^2 = \frac{2^K(1+M^2\alpha^2)^K - 1}{K(1+2M^2\alpha^2)} \frac{2M^2\alpha^2\sigma^2}{1+2M^2\alpha^2} + \frac{\sigma^2}{1+2M^2\alpha^2},$$

and this concludes the proof.

A.3. Proof of Proposition 3.8

Similar to the proof of Proposition 3.6, we first obtain:

$$\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{c}^i_t] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})],$$

which then gives:

$$\begin{aligned}
&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{c}^i_t] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\boldsymbol{c}^j_t]\|^2\\
=&\|\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})] - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\ell((\boldsymbol{u}^j_k,\xi^j_{k,t})]\|^2\\
=&\|\frac{1}{K}\sum_{k=0}^{K-1}(\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)])\|^2\\
\leq&\frac{1}{K}\sum_{k=0}^{K-1}\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2. \qquad (11)
\end{aligned}$$

Next, we need to bound $\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2$ for each $k=0,\ldots,K-1$ and tasks $i,j\in\mathcal{I}$. Note that when $k=0$, we have:

$$\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2 = \|\nabla\mathcal{L}^i(\boldsymbol{w}_t) - \nabla\mathcal{L}^j(\boldsymbol{w}_t)\|^2 \leq \delta^2,$$

that is directly obtained from Assumption 3.7. For $k>0$, we have:

$$\begin{aligned}
&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2\\
=&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)]\\
&+\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2\\
\leq&2\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k) - \nabla\mathcal{L}^j(\boldsymbol{u}^i_k)]\|^2 + 2\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2\\
\leq&2\delta^2 + 2\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2, \qquad (12)
\end{aligned}$$

where the first term is bounded by Assumption 3.7. For the second term in (12), we have:

$$\begin{aligned}
&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2\\
=&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\left[\nabla\mathcal{L}^j\left(\boldsymbol{u}^i_{k-1} - \alpha\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})\right) - \nabla\mathcal{L}^j\left(\boldsymbol{u}^j_{k-1} - \alpha\nabla\ell(\boldsymbol{u}^j_{k-1},\xi^j_{k-1,t})\right)\right]\|^2\\
\leq&\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\|\nabla\mathcal{L}^j(\boldsymbol{u}^i_{k-1} - \alpha\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})) - \nabla\mathcal{L}^j(\boldsymbol{u}^j_{k-1} - \alpha\nabla\ell(\boldsymbol{u}^j_{k-1},\xi^j_{k-1,t}))\|^2\\
\leq&M^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\|\left(\boldsymbol{u}^i_{k-1} - \alpha\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t})\right) - \left(\boldsymbol{u}^j_{k-1} - \alpha\nabla\ell(\boldsymbol{u}^j_{k-1},\xi^j_{k-1,t})\right)\|^2\\
\leq&M^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\left(2\|\boldsymbol{u}^i_{k-1} - \boldsymbol{u}^j_{k-1}\|^2 + 2\alpha^2\|\nabla\ell(\boldsymbol{u}^i_{k-1},\xi^i_{k-1,t}) - \nabla\ell(\boldsymbol{u}^j_{k-1},\xi^j_{k-1,t})\|^2\right). \qquad (13)
\end{aligned}$$

For the first term in (13), consider $\|\boldsymbol{u}_k^i - \boldsymbol{u}_k^j\|^2$, we have:

$$\begin{aligned}
\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\boldsymbol{u}_k^i - \boldsymbol{u}_k^j\|^2 =& \mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|(\boldsymbol{u}_k^i - \boldsymbol{w}_t) - (\boldsymbol{u}_k^j - \boldsymbol{w}_t)\|^2 \\
=& \mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\sum_{l=0}^{k-1}\alpha(\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j))\|^2 \\
\leq& k\alpha^2\sum_{l=0}^{k-1}\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2. \qquad (14)
\end{aligned}$$

Substituting the first term in (13) by (14),

$$\begin{aligned}
&\|\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}_k^i)] - \mathbb{E}_{\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}[\nabla\mathcal{L}^j(\boldsymbol{u}_k^j)]\|^2 \\
\leq& M^2\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}[2\|\boldsymbol{u}_{k-1}^i - \boldsymbol{u}_{k-1}^j\|^2 + 2\alpha^2\|\nabla\ell(\boldsymbol{u}_{k-1}^i,\xi_{k-1,t}^i) - \nabla\ell(\boldsymbol{u}_{k-1}^j,\xi_{k-1,t}^j)\|^2] \\
\leq& 2(k-1)M^2\alpha^2\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\left[\sum_{l=0}^{k-2}\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2\right] \\
&+ 2kM^2\alpha^2\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}[\|\nabla\ell(\boldsymbol{u}_{k-1}^i,\xi_{k-1,t}^i) - \nabla\ell(\boldsymbol{u}_{k-1}^j,\xi_{k-1,t}^j)\|^2]. \qquad (15)
\end{aligned}$$

Now we need to bound $\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}[\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2]$ for any $l = 0,\ldots,K-1$, that appears in (15). First, we have:

$$\begin{aligned}
&\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2 \\
=&\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\mathcal{L}^i(\boldsymbol{u}_l^i) + \nabla\mathcal{L}^i(\boldsymbol{u}_l^i) - \nabla\mathcal{L}^j(\boldsymbol{u}_l^i) + \nabla\mathcal{L}^j(\boldsymbol{u}_l^i) - \nabla\mathcal{L}^j(\boldsymbol{u}_l^j) + \nabla\mathcal{L}^j(\boldsymbol{u}_l^j) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2 \\
\leq&4\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\mathcal{L}^i(\boldsymbol{u}_l^i)\|^2 + 4\|\nabla\mathcal{L}^i(\boldsymbol{u}_l^i) - \nabla\mathcal{L}^j(\boldsymbol{u}_l^i)\|^2 \\
&+ 4\|\nabla\mathcal{L}^j(\boldsymbol{u}_l^i) - \nabla\mathcal{L}^j(\boldsymbol{u}_l^j)\|^2 + 4\|\nabla\mathcal{L}^j(\boldsymbol{u}_l^j) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2. \qquad (16)
\end{aligned}$$

For the expectation, we have:

$$\begin{aligned}
&\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2 \\
\leq&\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}[4\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\mathcal{L}^i(\boldsymbol{u}_l^i)\|^2 + 4\|\nabla\mathcal{L}^i(\boldsymbol{u}_l^i) - \nabla\mathcal{L}^j(\boldsymbol{u}_l^i)\|^2 \\
&+ 4\|\nabla\mathcal{L}^j(\boldsymbol{u}_l^i) - \nabla\mathcal{L}^j(\boldsymbol{u}_l^j)\|^2 + 4\|\nabla\mathcal{L}^j(\boldsymbol{u}_l^j) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2] \\
\leq&4\delta^2 + 8\sigma^2 + 4M^2\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\boldsymbol{u}_l^i - \boldsymbol{u}_l^j\|^2 \\
\leq&4\delta^2 + 8\sigma^2 + 4lM^2\alpha^2\sum_{m=0}^{l-1}\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}_m^i,\xi_{m,t}^i) - \ell(\boldsymbol{u}_m^j,\xi_{m,t}^j)\|^2,
\end{aligned}$$

where the first and last terms are bounded by Assumption 3.5, the second term is bounded by Assumption 3.7, and the third term is obtained from Assumption 3.2. This also implies:

$$\begin{aligned}
&\sum_{m=0}^{l}\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}_m^i,\xi_{m,t}^i) - \nabla\ell(\boldsymbol{u}_m^j,\xi_{m,t}^j)\|^2 \\
=&\sum_{m=0}^{l-1}\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}_m^i,\xi_{m,t}^i) - \ell(\boldsymbol{u}_m^j,\xi_{m,t}^j)\|^2 \\
&+ \mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}_l^i,\xi_{l,t}^i) - \nabla\ell(\boldsymbol{u}_l^j,\xi_{l,t}^j)\|^2 \\
\leq&4\delta^2 + 8\sigma^2 + (1 + 4KM^2\alpha^2)\sum_{m=0}^{l-1}\mathbb{E}_{\xi_{0,t}^i,\ldots,\xi_{K-1,t}^i\sim\mathcal{D}^i,\xi_{0,t}^j,\ldots,\xi_{K-1,t}^j\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}_m^i,\xi_{m,t}^i) - \ell(\boldsymbol{u}_m^j,\xi_{m,t}^j)\|^2, \qquad (17)
\end{aligned}$$

as we always have $l < K$ by definition. To derive the final bound, consider $l = 0$ in (17):

$$\begin{aligned}
&\sum_{m=0}^{0} \mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j} \|\nabla\ell(\boldsymbol{u}^i_m,\xi^i_{m,t}) - \nabla\ell(\boldsymbol{u}^j_m,\xi^j_{m,t})\|^2 \\
=&\mathbb{E}_{\xi^i_{0,t}\sim\mathcal{D}^i,\xi^j_{0,t}\sim\mathcal{D}^j} \|\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t}) - \nabla\ell(\boldsymbol{u}^j_{0,t},\xi^j_{0,t})\|^2 \\
=&\mathbb{E}_{\xi^i_{0,t}\sim\mathcal{D}^i,\xi^j_{0,t}\sim\mathcal{D}^j} \|\nabla\ell(\boldsymbol{w}_t,\xi^i_{0,t}) - \nabla\mathcal{L}^i(\boldsymbol{w}_t) + \nabla\mathcal{L}^i(\boldsymbol{w}_t) - \nabla\mathcal{L}^j(\boldsymbol{w}_t) + \nabla\mathcal{L}^j(\boldsymbol{w}_t) - \nabla\ell(\boldsymbol{w}_t,\xi^j_{0,t})\|^2 \\
=&\mathbb{E}_{\xi^i_{0,t}\sim\mathcal{D}^i} \|\nabla\ell(\boldsymbol{w}_t,\xi^i_{0,t}) - \nabla\mathcal{L}^i(\boldsymbol{w}_t)\|^2 + \|\nabla\mathcal{L}^i(\boldsymbol{w}_t) - \nabla\mathcal{L}^j(\boldsymbol{w}_t)\|^2 + \mathbb{E}_{\xi^j_{0,t}\sim\mathcal{D}^j} \|\nabla\mathcal{L}^j(\boldsymbol{w}_t) - \nabla\ell(\boldsymbol{w}_t,\xi^j_{0,t})\|^2 \\
\leq&2\sigma^2 + \delta^2,
\end{aligned}$$

where we remove the stochastic samples $\xi^i_{1,t},\ldots,\xi^i_{K-1,t}$ and $\xi^j_{1,t},\ldots,\xi^j_{K-1,t}$ that are irrelevant to the expectations, as $\|\nabla\ell(\boldsymbol{u}^i_0,\xi^i_{0,t}) - \nabla\ell(\boldsymbol{u}^j_{0,t},\xi^j_{0,t})\|^2$ only depends on $\xi^i_{0,t},\xi^j_{0,t}$. This then gives:

$$\begin{aligned}
&\sum_{m=0}^{l} \mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j} \|\nabla\ell(\boldsymbol{u}^i_m,\xi^i_{m,t}) - \nabla\ell(\boldsymbol{u}^j_m,\xi^j_{m,t})\|^2 \\
\leq&(1+4KM^2\alpha^2)^l(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2) - \frac{\delta^2+2\sigma^2}{KM\alpha^2}.
\end{aligned} \tag{18}$$

Now we have:

$$\begin{aligned}
&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j((\boldsymbol{u}^j_k)]\|^2 \\
\leq&2(k-1)\alpha^2M^2\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\left[\sum_{l=0}^{k-2}\|\nabla\ell(\boldsymbol{u}^i_l,\xi^i_{l,t}) - \ell(\boldsymbol{u}^j_l,\xi^j_{l,t})\|^2\right] \\
&+2k\alpha^2M^2\left(4\delta^2+8\sigma^2+4M^2(k-1)\alpha^2\sum_{m=0}^{k-2}\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\|\nabla\ell(\boldsymbol{u}^i_m,\xi^i_{m,t}) - \ell(\boldsymbol{u}^j_m,\xi^j_{m,t})\|^2\right) \\
=&2(k-1)M^2\alpha^2\left(1+4(k-1)M^2\alpha^2\right)\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i,\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}\left[\sum_{m=0}^{k-2}\|\nabla\ell(\boldsymbol{u}^i_m,\xi^i_{m,t}) - \ell(\boldsymbol{u}^j_m,\xi^j_{m,t})\|^2\right] \\
&+8kM^2\alpha^2(\delta^2+2\sigma^2) \\
\leq&2(k-1)M^2\alpha^2\left(1+4(k-1)M^2\alpha^2\right)\left((1+4KM^2\alpha^2)^{k-2}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2) - \frac{\delta^2+2\sigma^2}{KM\alpha^2}\right) \\
&+8kM^2\alpha^2(\delta^2+2\sigma^2) \\
\leq&2(k-1)M^2\alpha^2(1+4KM^2\alpha^2)^{k-1}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2) + 8kM^2\alpha^2(\delta^2+2\sigma^2),
\end{aligned} \tag{19}$$

where the first inequality is due to (16), the second is due to (18), and the last is obtained by removing the $-\frac{\delta^2+2\sigma^2}{KM\alpha^2}$ term which is always negative. Using (19) in (12), we have:

$$\begin{aligned}
&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j(\boldsymbol{u}^j_k)]\|^2 \\
\leq&2\delta^2 + 2\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^j(\boldsymbol{u}^i_k)] - \mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j(\boldsymbol{u}^j_k)]\|^2 \\
\leq&2\delta^2 + 4(k-1)M^2\alpha^2(1+4KM^2\alpha^2)^{k-1}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2) + 16kM^2\alpha^2(\delta^2+2\sigma^2).
\end{aligned} \tag{20}$$

Using (20) in (11),

$$\begin{aligned}
&\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\boldsymbol{c}^i_t]-\mathbb{E}_{\xi^j_{0:K-1,t}}[\boldsymbol{c}^j_t]\|^2\\
=&\|\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\ell(\boldsymbol{u}^i_k,\xi^i_{k,t})]-\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\ell(\boldsymbol{u}^j_k,\xi^j_{k,t})]\|^2\\
\leq&\frac{1}{K}\sum_{k=0}^{K-1}\|\mathbb{E}_{\xi^i_{0,t},\ldots,\xi^i_{K-1,t}\sim\mathcal{D}^i}[\nabla\mathcal{L}^i(\boldsymbol{u}^i_k)]-\mathbb{E}_{\xi^j_{0,t},\ldots,\xi^j_{K-1,t}\sim\mathcal{D}^j}[\nabla\mathcal{L}^j(\boldsymbol{u}^j_k)]\|^2\\
\leq&2\delta^2+4M^2\alpha^2\frac{(1+4KM^2\alpha^2)^K}{16K^2M^4\alpha^4}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2)+8KM^2\alpha^2(\delta^2+2\sigma^2)\\
=&2\delta^2+\frac{(1+4KM^2\alpha^2)^K}{4K^2M^2\alpha^2}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2)+8KM^2\alpha^2(\delta^2+2\sigma^2).
\end{aligned}$$

Finally, define

$$\tilde{\delta}^2=2\delta^2+\frac{(1+4KM^2\alpha^2)^K}{4K^2M^2\alpha^2}(1+\frac{1}{KM\alpha^2})(\delta^2+2\sigma^2)+8KM^2\alpha^2(\delta^2+2\sigma^2),$$

which concludes the proof.

A.4. Proof of Theorem 3.12

First, we need the following Lemma.

Lemma A.1. *If $\eta_t\leq\frac{1}{2\tilde{M}}$ in Algorithm 2, then:*

$$\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1})\leq\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\frac{\eta_t}{2}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2+\frac{\eta_t^2\tilde{M}}{2}\tilde{\sigma}^2.$$

Proof. Since $\tilde{\mathcal{L}}(\boldsymbol{w}_t)$ is $\tilde{M}$-Lipschitz smooth, we have:

$$\begin{aligned}
\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1})\leq&\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t)+(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top(\boldsymbol{w}_{t+1}-\boldsymbol{w}_t)+\frac{\tilde{M}}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2]\\
=&\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\eta_t(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{c}_t+\frac{\eta_t^2\tilde{M}}{2}\|\boldsymbol{c}_t\|^2].
\end{aligned}$$

For Algorithm 2, we have $\boldsymbol{c}_t=\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}^i_t$. Since the tasks i are independently sampled, obviously we have $\mathbb{E}[\boldsymbol{c}_t]=\mathbb{E}_i[\boldsymbol{c}^i_t]=\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)$, which gives:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{c}_t\|^2=&\mathbb{E}\|\boldsymbol{c}_t-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)+\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\\
=&\mathbb{E}\|\boldsymbol{c}_t-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2+2\mathbb{E}(\boldsymbol{c}_t-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)+\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\\
\leq&\tilde{\sigma}^2+\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2,
\end{aligned}$$

where the last inequality comes from Corollary 3.10. Then we have

$$\begin{aligned}
\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1})\leq&\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\eta_t(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{c}_t+\frac{\eta_t^2\tilde{M}}{2}\|\boldsymbol{c}_t\|^2]\\
\leq&\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\eta_t\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2+\frac{\eta_t^2\tilde{M}}{2}(\tilde{\sigma}^2+\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2)]\\
=&\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\eta_t(1-\frac{\eta_t\tilde{M}}{2})\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2+\frac{\eta_t^2\tilde{M}}{2}\tilde{\sigma}^2].
\end{aligned}$$

Since $\eta_t \leq \frac{1}{2\tilde{M}}$, we have $\eta_t(1-\frac{\eta_t\tilde{M}}{2}) \geq \frac{\eta_t}{2}$, which gives:

$$\begin{aligned}\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1}) \leq& \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \eta_t(1-\frac{\eta_t\tilde{M}}{2})\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{\eta_t^2\tilde{M}}{2}\tilde{\sigma}^2] \\ \leq& \mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \frac{\eta_t}{2}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{\eta_t^2\tilde{M}}{2}\tilde{\sigma}^2,\end{aligned}$$

and this concludes the proof. □

In Lemma A.1, summing t from 0 to $T-1$ gives:

$$\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_T) \leq \mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_0) - \sum_{t=0}^{T-1}\frac{\eta_t}{2}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \sum_{t=0}^{T-1}\frac{\eta_t^2\tilde{M}}{2}\tilde{\sigma}^2,$$

i.e.,

$$\sum_{t=0}^{T-1}\frac{\eta_t}{2}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 \leq \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0) - \tilde{\mathcal{L}}(\boldsymbol{w}_T)] + \tilde{M}\tilde{\sigma}^2\eta_{g,0}^2\ln(T+1),$$

where we have used the fact that $\sum_{t=0}^{T-1}\eta_t^2 = \sum_{t=0}^{T-1}\frac{\eta_{g,0}^2}{1+t} \leq 2\eta_{g,0}^2\ln(T+1)$. Also, note that $\eta_t > \eta_t$ for any $t = 0,\ldots,T-1$ and $\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_T) \leq \tilde{\mathcal{L}}^*$, which gives:

$$\frac{\eta_t}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 \leq \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0) - \tilde{\mathcal{L}}^*] + \tilde{M}\tilde{\sigma}^2\eta_{g,0}^2\ln(T+1) = G,$$

where G denotes the right hand side for notational simplicity. To obtain the final result, we need to divide both sides by $\frac{T\eta_t}{2}$, for which we have:

$$\frac{2}{T\eta_t} = \frac{2}{T\eta_{g,0}}(1+T)^{1/2} \leq \frac{2}{T\eta_{g,0}}\frac{1}{2}(\sqrt{2}+\sqrt{2T}) = \frac{1}{\eta_{g,0}}(\frac{\sqrt{2}}{T}+\sqrt{\frac{2}{T}}),$$

which comes from the concavity of the square root function. Then we obtain:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 \leq \frac{2}{T\eta_t}G \leq \frac{\sqrt{2}G/\eta_{g,0}}{\sqrt{T}} + \frac{\sqrt{2}G/\eta_{g,0}}{T},$$

which concludes the proof.

A.5. Proof of Theorem 3.13

We extend Theorem 3.13 to the following which allows a flexible choice of η_t.

Theorem A.2. *For any $b > 0$, set*

$$\eta_t = \frac{b}{\tilde{M}\left((7b+\frac{1}{28b^2})^3+t\right)^{1/3}},\ \gamma_{t+1} = c\eta_t^2$$

for all $t = 0,\ldots,T-1$, where $c = \tilde{M}^2(28+\frac{1}{7b^3})$, then Reptile+STORM (Algorithm 3) satisfies:

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\right] \leq \frac{G_2\tilde{M}/b}{T^{2/3}} + \frac{G_2\tilde{M}(7+\frac{1}{28b^3})}{T},$$

where

$$G_2 = 8\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0) - \tilde{\mathcal{L}}^*] + \frac{(7+\frac{1}{28b^3})\tilde{\sigma}^2}{4\tilde{M}} + \frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T$$

and $\tilde{\mathcal{L}}^* = \min_{\boldsymbol{w}}\tilde{\mathcal{L}}(\boldsymbol{w})$.

First, we need the following Lemma.

Lemma A.3. *If $\eta_t \leq \frac{1}{4\tilde{M}}$ in Algorithm 3, then:*

$$\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1}) \leq \mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \frac{\eta_t}{4}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{3\eta_t}{4}\mathbb{E}\|\boldsymbol{e}_t\|^2,$$

where $\boldsymbol{e}_t = \boldsymbol{c}_t - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)$ and the expectation is taken over all tasks and training data.

Proof. Since $\tilde{\mathcal{L}}(\boldsymbol{w}_t)$ is $\tilde{M}$-Lipschitz smooth, we have:

$$\begin{aligned}\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1}) \leq& \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t) + (\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{\tilde{M}}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2] \\ =& \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \eta_t(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{c}_t + \frac{\eta_t^2\tilde{M}}{2}\|\boldsymbol{c}_t\|^2].\end{aligned}$$

Since $\boldsymbol{c}_t = \boldsymbol{e}_t + \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)$, we have:

$$\begin{aligned}(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{c}_t &= \|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + (\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{e}_t, \\ \|\boldsymbol{c}_t\|^2 = \|\boldsymbol{e}_t + \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 &\leq 2\|\boldsymbol{e}_t\|^2 + 2\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2,\end{aligned}$$

and so:

$$\begin{aligned}\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1}) \leq& \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \eta_t\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 - \eta_t(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{e}_t \\ &+ \eta_t^2\tilde{M}\|\boldsymbol{e}_t\|^2 + \eta_t^2\tilde{M}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2].\end{aligned}$$

For $(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{e}_t$, we have:

$$(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))^\top\boldsymbol{e}_t \geq -\frac{1}{2}(\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \|\boldsymbol{e}_t\|^2),$$

which then gives:

$$\begin{aligned}\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1}) \leq& \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \frac{\eta_t}{2}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{\eta_t}{2}\|\boldsymbol{e}_t\|^2 + \eta_t^2\tilde{M}\|\boldsymbol{e}_t\|^2 + \eta_t^2\tilde{M}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2] \\ =& \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \frac{\eta_t}{2}(1 - 2\eta_t\tilde{M})\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{\eta_t}{2}(1 + 2\eta_t\tilde{M})\|\boldsymbol{e}_t\|^2].\end{aligned}$$

Since $\eta_t \leq \frac{1}{4\tilde{M}}$, we have $-\frac{\eta_t}{2}(1 - 2\eta_t\tilde{M}) \leq -\frac{\eta_t}{4}$ and $\frac{\eta_t}{2}(1 + 2\eta_t\tilde{M}) \leq \frac{3\eta_t}{4}$. Combining these two gives:

$$\mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1}) \leq \mathbb{E}\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \frac{\eta_t}{4}\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{3\eta_t}{4}\mathbb{E}\|\boldsymbol{e}_t\|^2,$$

which concludes the proof. □

The following Lemma. bounds $\|\boldsymbol{e}_t\|^2$.

Lemma A.4. *For any $t > 0$, we have:*

$$\mathbb{E}\|\boldsymbol{e}_t\|^2 \leq \frac{2\gamma_t^2\sigma^2}{|\mathcal{I}_t|} + (1 - \gamma_t)^2(1 + 4\tilde{M}^2\eta_{t-1}^2)\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2 + 4(1 - \gamma_t)^2\tilde{M}^2\eta_{t-1}^2\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})\|^2.$$

Proof. From Algorithm 3, we have:

$$\begin{aligned}\boldsymbol{e}_t = \boldsymbol{c}_t - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t) =& \frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t) + (1 - \gamma_t)(\boldsymbol{c}_{t-1} - \boldsymbol{d}_t) \\ =& \gamma_t\left(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\right) + (1 - \gamma_t)\left(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t) + \boldsymbol{c}_{t-1} - \frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{d}_t^i\right) \\ =& \gamma_t\left(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\right) + (1 - \gamma_t)\left(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}(\boldsymbol{c}_t^i - \boldsymbol{d}_t^i) - (\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t) - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1}))\right) \\ &+ (1 - \gamma_t)(\boldsymbol{c}_{t-1} - \nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})),\end{aligned}$$

where the last term is exactly $(1-\gamma_t)\boldsymbol{e}_{t-1}$ and is independent of the first two terms. For $\mathbb{E}\|\boldsymbol{e}_t\|^2$, we have:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}_t\|^2 =&\mathbb{E}\|\gamma_t(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t))+(1-\gamma_t)(\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}(\boldsymbol{c}_t^i-\boldsymbol{d}_t^i)-(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})))\|^2\\&+(1-\gamma_t)^2\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2.\end{aligned}$$

Note that for any two vectors $\boldsymbol{w},\boldsymbol{u}$, we always have $\|\boldsymbol{w}+\boldsymbol{u}\|^2\le 2\|\boldsymbol{w}\|^2+2\|\boldsymbol{u}\|^2$. This implies:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}_t\|^2 =&2\gamma_t^2\mathbb{E}\|\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2+2(1-\gamma_t)^2\mathbb{E}\|\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}(\boldsymbol{c}_t^i-\boldsymbol{d}_t^i)-(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1}))\|^2\\&+(1-\gamma_t)^2\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2.\end{aligned}\tag{21}$$

For the first term in (21), since the tasks i are independently sampled, we have:

$$\mathbb{E}\|\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\boldsymbol{c}_t^i-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2=\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\boldsymbol{c}_t^i-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2\le\frac{\sigma^2}{|\mathcal{I}_t|}.$$

For the second term in (21), similarly we have:

$$\begin{aligned}\mathbb{E}\|\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}(\boldsymbol{c}_t^i-\boldsymbol{d}_t^i)-(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1}))\|^2=&\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|(\boldsymbol{c}_t^i-\boldsymbol{d}_t^i)-(\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1}))\|^2\\=&\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\boldsymbol{c}_t^i-\boldsymbol{d}_t^i\|^2-\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})\|^2\\\le&\frac{1}{|\mathcal{I}_t|}\sum_{i\in\mathcal{I}_t}\mathbb{E}\|\boldsymbol{c}_t^i-\boldsymbol{d}_t^i\|^2.\end{aligned}$$

For any $\|\boldsymbol{c}_t^i-\boldsymbol{d}_t^i\|^2$, Proposition 3.3 leads to $\|\boldsymbol{c}_t^i-\boldsymbol{d}_t^i\|^2\le\tilde{M}^2\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|^2$. Combining all these together, we have:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{e}_t\|^2\le&\frac{2\gamma_t^2\sigma^2}{|\mathcal{I}_t|}+2(1-\gamma_t)^2\tilde{M}^2\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|^2+(1-\gamma_t)^2\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2\\=&\frac{2\gamma_t^2\sigma^2}{|\mathcal{I}_t|}+2(1-\gamma_t)^2\tilde{M}^2\eta_{t-1}^2\|\boldsymbol{c}_{t-1}\|^2+(1-\gamma_t)^2\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2\\=&\frac{2\gamma_t^2\sigma^2}{|\mathcal{I}_t|}+2(1-\gamma_t)^2\tilde{M}^2\eta_{t-1}^2\|\boldsymbol{e}_{t-1}+\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})\|^2+(1-\gamma_t)^2\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2\\\le&\frac{2\gamma_t^2\sigma^2}{|\mathcal{I}_t|}+4(1-\gamma_t)^2\tilde{M}^2\eta_{t-1}^2\|\boldsymbol{e}_{t-1}\|^2+4(1-\gamma_t)^2\tilde{M}^2\eta_{t-1}^2\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})\|^2+(1-\gamma_t)^2\mathbb{E}\|\boldsymbol{e}_{t-1}\|^2\\=&\frac{2\gamma_t^2\sigma^2}{|\mathcal{I}_t|}+(1-\gamma_t)^2(1+4\tilde{M}^2\eta_{t-1}^2)\|\boldsymbol{e}_{t-1}\|^2+4(1-\gamma_t)^2\tilde{M}^2\eta_{t-1}^2\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})\|^2,\end{aligned}$$

which concludes the proof. □

Now we are ready to prove Theorem 3.13:

Proof for Theorem 3.13. From Lemma A.4, we first consider bounding $\mathbb{E}[\eta_t^{-1}\|\boldsymbol{e}_{t+1}\|^2-\eta_{t-1}^{-1}\|\boldsymbol{e}_t\|^2]$, i.e., the difference in variance, which is given by:

$$\begin{aligned}\mathbb{E}[\eta_t^{-1}\|\boldsymbol{e}_{t+1}\|^2-\eta_{t-1}^{-1}\|\boldsymbol{e}_t\|^2]\le&\frac{2\gamma_{t+1}^2\tilde{\sigma}^2}{\eta_t|\mathcal{I}_{t+1}|}+(1-\gamma_{t+1})^2\frac{1+4\tilde{M}^2\eta_t^2}{\eta_t}\|\boldsymbol{e}_t\|^2\\&+4(1-\gamma_{t+1})^2\tilde{M}^2\eta_t\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2-\frac{1}{\eta_{t-1}}\|\boldsymbol{e}_t\|^2.\end{aligned}$$

Recall that

$$\eta_t = \frac{b}{\tilde{M}\left((7b+\frac{1}{28b^2})^3+t\right)^{1/3}},\ \gamma_{t+1} = c\eta_t^2,$$

which implies that

$$\begin{aligned}\eta_t \leq & \eta_0 = \frac{1}{\tilde{M}(7+\frac{1}{28b^3})} < \frac{1}{4\tilde{M}},\\ \gamma_{t+1} \leq & c\eta_0^2 = \tilde{M}^2(28+\frac{1}{7b^3})\cdot\frac{1}{\tilde{M}^2(7+\frac{1}{28b^3})^2} = \frac{28+\frac{1}{7b^3}}{(7+\frac{1}{28b^3})^2} < 1,\end{aligned}$$

i.e., $\eta_t \leq \frac{1}{4\tilde{M}}$ and $0 < \gamma_{t+1} \leq 1$ always hold for $t \geq 0$. Thus, we have $0 \leq 1-\gamma_{t+1} < 1$, which gives:

$$\begin{aligned}\mathbb{E}[\eta_t^{-1}\|\boldsymbol{e}_{t+1}\|^2 - \eta_{t-1}^{-1}\|\boldsymbol{e}_t\|^2] \leq & \frac{2\gamma_{t+1}^2\tilde{\sigma}^2}{\eta_t|\mathcal{I}_{t+1}|} + (\frac{1-\gamma_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}} + 4\tilde{M}^2\eta_t)\|\boldsymbol{e}_t\|^2\\ & + 4\tilde{M}^2\eta_t\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2.\end{aligned}$$

Summing t from 0 to $T-1$, we have:

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}[\eta_t^{-1}\|\boldsymbol{e}_{t+1}\|^2 - \eta_{t-1}^{-1}\|\boldsymbol{e}_t\|^2] \leq & \sum_{t=0}^{T-1}\frac{2\gamma_{t+1}^2\tilde{\sigma}^2}{\eta_t|\mathcal{I}_{t+1}|} + \sum_{t=0}^{T-1}(\frac{1-\gamma_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}} + 4\tilde{M}^2\eta_t)\mathbb{E}\|\boldsymbol{e}_t\|^2\\ & + 4\tilde{M}^2\sum_{t=0}^{T-1}\eta_t\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2. \qquad (22)\end{aligned}$$

For the first term in (22), since $\eta_t = \frac{b}{\tilde{M}\left((7b+\frac{1}{28b^2})^3+t\right)^{1/3}}$ and $\gamma_{t+1} = c\eta_t^2$, we have:

$$\begin{aligned}\sum_{t=0}^{T-1}\frac{2\gamma_{t+1}^2\tilde{\sigma}^2}{\eta_t|\mathcal{I}_{t+1}|} = & \sum_{t=0}^{T-1}\frac{2c^2\eta_t^3\tilde{\sigma}^2}{|\mathcal{I}_{t+1}|}\\ = & \sum_{t=0}^{T-1}\frac{2c^2b^3\tilde{\sigma}^2}{\tilde{M}^3\left((7b+\frac{1}{28b^2})^3+t\right)|\mathcal{I}_{t+1}|}\\ \leq & \frac{2c^2b^3\tilde{\sigma}^2}{\tilde{M}^3}\cdot\sum_{t=0}^{T-1}\frac{1}{1+t} \leq \frac{2c^2b^3\tilde{\sigma}^2}{\tilde{M}^3}\ln T,\end{aligned}$$

where the last two inequalities come from $(7b+\frac{1}{28b^2})^3 > 2$ and $|\mathcal{I}_{t+1}| \geq 1$.

For the second term in (22), we first bound $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}$. Obviously, $x^{1/3}$ is a concave function, which gives $(x+y)^{1/3} \leq x^{1/3} + y\cdot x^{-2/3}/3$. Therefore, we have:

$$\begin{aligned}\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = & \frac{\tilde{M}}{b}\left(\left((7b+\frac{1}{28b^2})^3+t\right)^{1/3} - \left((7b+\frac{1}{28b^2})^3+(t-1)\right)^{1/3}\right)\\ \leq & \frac{\tilde{M}}{3b}\left((7b+\frac{1}{28b^2})^3+(t-1)\right)^{-2/3} = \frac{\tilde{M}}{3b\left((7b+\frac{1}{28b^2})^3+(t-1)\right)^{2/3}}.\end{aligned}$$

Since $(7b+\frac{1}{28b^2})^3 > 2$ for $b>0$, we have $\frac{1}{2}(7b+\frac{1}{28b^2})^3+\frac{t}{2} < (7b+\frac{1}{28b^2})^3 - 1 + t$ for any $t \geq 0$, which implies that:

$$\begin{aligned}\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}} \leq & \frac{\tilde{M}}{3b\left(\left((7b+\frac{1}{28b^2})^3-1\right)+t\right)^{2/3}} \\ \leq & \frac{\tilde{M}}{3b\left(\frac{1}{2}(7b+\frac{1}{28b^2})^3+\frac{t}{2}\right)^{2/3}} \\ = & \frac{2^{2/3}\tilde{M}}{3b\left((7b+\frac{1}{28b^2})^3+t\right)^{2/3}} = \frac{2^{2/3}\tilde{M}^3}{3b^3}\eta_t^2.\end{aligned}$$

Since $\eta_t \leq \frac{1}{4\tilde{M}}$, we have:

$$\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}} \leq \frac{2^{2/3}\tilde{M}^3}{3b^3}\eta_t^2 \leq \frac{2^{2/3}\tilde{M}^2}{12b^3}\eta_t < \frac{\tilde{M}^2}{7b^3}\eta_t, \tag{23}$$

where the last inequality comes from $12/2^{2/3} > 7$. For the $-\frac{\gamma_{t+1}}{\eta_t}+4\tilde{M}^2\eta_t$ term, we have:

$$-\frac{\gamma_{t+1}}{\eta_t}+4\tilde{M}^2\eta_t = (4\tilde{M}^2-c)\eta_t \leq (4\tilde{M}^2-\tilde{M}^2(28+\frac{1}{7b^3}))\eta_t = -24\tilde{M}^2\eta_t - \frac{\tilde{M}\eta_t}{7b^3}. \tag{24}$$

Combining (23) and (24), we have:

$$(\frac{1-\gamma_{t+1}}{\eta_t}-\frac{1}{\eta_{t-1}}+4\tilde{M}^2\eta_t)\|\boldsymbol{e}_t\|^2 \leq (\frac{\tilde{M}^2}{7b^3}\eta_t - 24\tilde{M}^2\eta_t - \frac{\tilde{M}\eta_t}{7b^3})\|\boldsymbol{e}_t\|^2 = -24\tilde{M}^2\eta_t\|\boldsymbol{e}_t\|^2.$$

Now we have:

$$\sum_{t=0}^{T-1}\mathbb{E}[\eta_t^{-1}\|\boldsymbol{e}_{t+1}\|^2-\eta_{t-1}^{-1}\|\boldsymbol{e}_t\|^2] \leq \frac{2c^2b^3\tilde{\sigma}^2}{\tilde{M}^3}\ln T - 24\tilde{M}^2\sum_{t=0}^{T-1}\eta_t\mathbb{E}\|\boldsymbol{e}_t\|^2 + 4\tilde{M}^2\sum_{t=0}^{T-1}\eta_t\mathbb{E}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2.$$

Dividing the sum by $32\tilde{M}^2$ on both sides gives:

$$\frac{1}{32\tilde{M}^2}\sum_{t=0}^{T-1}\mathbb{E}[\eta_t^{-1}\|\boldsymbol{e}_{t+1}\|^2-\eta_{t-1}^{-1}\|\boldsymbol{e}_t\|^2] \leq \frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T + \sum_{t=0}^{T-1}\mathbb{E}[\frac{\eta_t}{8}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_{t-1})\|^2 - \frac{3\eta_t}{4}\|\boldsymbol{e}_t\|^2].$$

Consider the potential function $\Phi_t = \tilde{\mathcal{L}}(\boldsymbol{w}_t) + \frac{1}{32\tilde{M}^2\eta_{t-1}}\|\boldsymbol{e}_t\|^2$. We have:

$$\begin{aligned}\mathbb{E}[\Phi_{t+1}-\Phi_t] = & \mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_{t+1})-\tilde{\mathcal{L}}(\boldsymbol{w}_t)] + \mathbb{E}[\frac{1}{32\tilde{M}^2\eta_t}\|\boldsymbol{e}_{t+1}\|^2 - \frac{1}{32\tilde{M}^2\eta_{t-1}}\|\boldsymbol{e}_t\|^2] \\ \leq & \mathbb{E}[-\frac{\eta_t}{4}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{3\eta_t}{4}\|\boldsymbol{e}_t\|^2 + \frac{1}{32\tilde{M}^2\eta_t}\|\boldsymbol{e}_{t+1}\|^2 - \frac{1}{32\tilde{M}^2\eta_{t-1}}\|\boldsymbol{e}_t\|^2],\end{aligned}$$

where the first part is bounded from Lemma A.3. Summing t from 0 to $T-1$ gives:

$$\begin{aligned}\mathbb{E}[\Phi_T-\Phi_0] \leq & \sum_{t=0}^{T-1}\mathbb{E}[-\frac{\eta_t}{4}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{3\eta_t}{4}\|\boldsymbol{e}_t\|^2 + \frac{1}{32\tilde{M}^2\eta_t}\|\boldsymbol{e}_{t+1}\|^2 - \frac{1}{32\tilde{M}^2\eta_{t-1}}\|\boldsymbol{e}_t\|^2] \\ \leq & \sum_{t=0}^{T-1}\mathbb{E}[-\frac{\eta_t}{4}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{3\eta_t}{4}\|\boldsymbol{e}_t\|^2] + \frac{1}{32\tilde{M}^2}\sum_{t=0}^{T-1}\mathbb{E}[\frac{1}{\eta_t}\|\boldsymbol{e}_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\|\boldsymbol{e}_t\|^2] \\ \leq & \sum_{t=0}^{T-1}\mathbb{E}[-\frac{\eta_t}{4}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 + \frac{3\eta_t}{4}\|\boldsymbol{e}_t\|^2] + \frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T + \sum_{t=0}^{T-1}\mathbb{E}[\frac{\eta_t}{8}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2 - \frac{3\eta_t}{4}\|\boldsymbol{e}_t\|^2] \\ = & \sum_{t=0}^{T-1}\mathbb{E}[-\frac{\eta_t}{8}\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2] + \frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T,\end{aligned}$$

which implies that:

$$\sum_{t=0}^{T-1}\mathbb{E}[\eta_t\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2] \leq 8\mathbb{E}[\Phi_0-\Phi_T]+\frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T$$
$$\leq 8\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0)-\tilde{\mathcal{L}}^*]+\frac{1}{4\tilde{M}^2\eta_{-1}}\mathbb{E}\|\boldsymbol{e}_0\|^2+\frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T.$$

Since η_t is decreasing w.r.t. t, we have $\eta_t > \eta_T$ for $t = 0, \ldots, T-1$. For η_{-1}, we have:

$$\frac{1}{\eta_{-1}} = \frac{\tilde{M}\left((7b+\frac{1}{28b^2})^3-1\right)^{1/3}}{b} \leq \tilde{M}(7+\frac{1}{28b^3}).$$

For $\mathbb{E}\|\boldsymbol{e}_0\|^2 = \mathbb{E}\|\boldsymbol{c}_0-\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_0)\|^2$, we have $\mathbb{E}\|\boldsymbol{e}_0\|^2 \leq \tilde{\sigma}^2$ from Proposition 3.10. Combining all these,

$$\eta_T\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2] \leq 8\mathbb{E}[\tilde{\mathcal{L}}(\boldsymbol{w}_0)-\tilde{\mathcal{L}}^*]+\frac{(7+\frac{1}{28b^3})\tilde{\sigma}^2}{4\tilde{M}}+\frac{c^2b^3\tilde{\sigma}^2}{16\tilde{M}^5}\ln T = G_2,$$

where G_2 denotes the whole right hand for notational simplicity. To obtain the final result, we need to divide both sides by $T\eta_T$, for which we have:

$$\frac{1}{T\eta_T} = \frac{(w+\tilde{\sigma}^2T)^{1/3}}{kT} \leq \frac{\tilde{\sigma}^{2/3}}{kT^{2/3}}+\frac{w^{1/3}}{kT},$$

and

$$\frac{1}{T\eta_T} = \frac{\tilde{M}\left((7b+\frac{1}{28b^2})^3+T\right)^{1/3}}{bT} \leq \frac{\tilde{M}}{bT^{2/3}}+\frac{\tilde{M}(7+\frac{1}{28b^3})}{T}.$$

Then we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\tilde{\mathcal{L}}(\boldsymbol{w}_t)\|^2] \leq \frac{G_2\tilde{M}/b}{T^{2/3}}+\frac{G_2\tilde{M}(7+\frac{1}{28b^3})}{T},$$

which concludes the proof. □

B. Proposed Variance-Reduced Variants for Popular Meta-Learning Algorithms

B.1. Integration with MAML and FOMAML

Besides the integration with Reptile as discussed in Section 3.1, the proposed method can also be integrated with other meta-learning algorithms such as MAML, FOMAML and BMG. The resultant algorithms, which will be called VR-MAML, VR-FOMAML and VR-BMG, respectively, are shown in Algorithm 5. Similar to VR-Reptile, we introduce additional $\boldsymbol{c}_t$ and $\boldsymbol{d}_{t-1}$, and apply similar variance reduction steps.

B.2. Combining STORM with Meta-learning

Algorithm 6 shows how STORM can be integrated into meta-learning algorithms, by replacing all stochastic gradients in these algorithms with the variance-reduced gradients in STORM.

Algorithm 5 VR-MAML/FOMAML/BMG (Variance-Reduced MAML/FOMAML/BMG)

Input: $\boldsymbol{w}_0$, stepsizes $\{\eta_t\}$ and α, number of local steps K, decay parameter $\{\gamma_t\}$ (1 means no variance reduction).
sample tasks $\mathcal{I}_0 \subset \mathcal{I}$
for $i \in \mathcal{I}_t$ **do**
  $\boldsymbol{u}_{0,0}^i = \boldsymbol{w}_0$
  for $k = 0$ **to** $K-1$ **do**
    obtain samples $\xi_{k,t}^i$ from support data of task i
    $\boldsymbol{u}_{k+1,0}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,0}^i)$
  end for
  if VR-BMG **then**
    for $k = K$ **to** $K+M-1$ **do**
      obtain samples $\xi_{k,0}^i$ from support data of task i
      $\boldsymbol{u}_{k+1,0}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,0}^i)$
    end for
  end if
  obtain samples $\xi_{K,0}^i$ from query data of task i
  if VR-MAML **then**
    $\tilde{\boldsymbol{c}}_0^i = \nabla_{\boldsymbol{w}_0} \ell(\boldsymbol{u}_k^i, \xi_{K,0}^i)$
  else if VR-FOMAML **then**
    $\tilde{\boldsymbol{c}}_0^i = \nabla \ell(\boldsymbol{u}_k^i, \xi_{K,0}^i)$
  else if VR-BMG **then**
    $\tilde{\boldsymbol{c}}_0^i = \nabla_{\boldsymbol{w}_0} (\frac{1}{2} \|\boldsymbol{u}_{K+M,0}^i - \boldsymbol{u}_{K,0}^i\|)$
  end if
end for
$\tilde{\boldsymbol{c}}_0 = \frac{1}{|\mathcal{I}_0|} \sum_{i \in \mathcal{I}_0} \boldsymbol{c}_0^i$
$\boldsymbol{w}_1 = \boldsymbol{w}_0 - \eta_0 \tilde{\boldsymbol{c}}_0$
for $t = 1$ **to** $T-1$ **do**
  sample tasks $\mathcal{I}_t \subset \mathcal{I}$
  for $i \in \mathcal{I}_t$ **do**
    $\boldsymbol{u}_0^i = \boldsymbol{w}_t, \boldsymbol{v}_{0,t-1}^i = \boldsymbol{w}_{t-1}$
    for $k = 0$ **to** $K-1$ **do**
      obtain samples $\xi_{k,t}^i$ from support data of task i
      $\boldsymbol{u}_{k+1,t}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,t}^i), \boldsymbol{v}_{k+1,t-1}^i = \boldsymbol{v}_{k,t-1}^i - \alpha \nabla \ell(\boldsymbol{v}_{k,t-1}^i, \xi_{k,t}^i)$
    end for
    if VR-BMG **then**
      for $k = K$ **to** $K+M-1$ **do**
        obtain samples $\xi_{k,t}^i$ from support data of task i
        $\boldsymbol{u}_{k+1,t}^i = \boldsymbol{u}_k^i - \alpha \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,t}^i), \boldsymbol{v}_{k+1,t-1}^i = \boldsymbol{v}_{k,t-1}^i - \alpha \nabla \ell(\boldsymbol{v}_{k,t-1}^i, \xi_{k,t}^i)$
      end for
    end if
    obtain samples $\xi_{K,t}^i$ from query data of task i
    if VR-MAML **then**
      $\tilde{\boldsymbol{d}}_{t-1}^i = \nabla_{\boldsymbol{w}_{t-1}} \ell(\boldsymbol{v}_{K,t-1}^i, \xi_{K,t}^i), \tilde{\boldsymbol{c}}_t^i = \nabla_{\boldsymbol{w}_t} \ell(\boldsymbol{u}_k^i, \xi_{K,t}^i)$
    else if VR-FOMAML **then**
      $\tilde{\boldsymbol{d}}_{t-1}^i = \nabla \ell(\boldsymbol{v}_{K,t-1}^i, \xi_{K,t}^i), \tilde{\boldsymbol{c}}_t^i = \nabla \ell(\boldsymbol{u}_k^i, \xi_{K,t}^i)$
    else if VR-BMG **then**
      $\tilde{\boldsymbol{d}}_{t-1}^i = \nabla_{\boldsymbol{w}_{t-1}} (\frac{1}{2} \|\boldsymbol{v}_{K+M,t-1}^i - \boldsymbol{v}_{K,t-1}^i\|), \tilde{\boldsymbol{c}}_t^i = \nabla_{\boldsymbol{w}_t} (\frac{1}{2} \|\boldsymbol{u}_{K+M,t}^i - \boldsymbol{u}_{K,t}^i\|)$
    end if
  end for
  $\tilde{\boldsymbol{d}}_{t-1} = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \tilde{\boldsymbol{d}}_{t-1}^i$
  $\tilde{\boldsymbol{c}}_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \tilde{\boldsymbol{c}}_t^i + (1-\gamma_t)(\tilde{\boldsymbol{c}}_{t-1} - \tilde{\boldsymbol{d}}_{t-1})$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \tilde{\boldsymbol{c}}_t$
end for

Algorithm 6 MAML/FOMAML/Reptile+STORM: meta-learning algorithms with stochastic gradients replaced by STORM.

Input: $\boldsymbol{w}_0$, step-size η_t and α, number of local steps K.
for $t = 0$ **to** $T-1$ **do**
  Sample tasks $\mathcal{I}_t \subset \mathcal{I}$
  for $i \in \mathcal{I}_t$ **do**
    $\boldsymbol{u}_0^i = \boldsymbol{w}_t$
    for $k = 0$ **to** $K-1$ **do**
      Obtain data samples $\xi_{k,t}^i$ from the support data of task i
      if $k > 0$ **then**
        $\boldsymbol{m}_{k,t}^i = \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,t}^i) + (1-\gamma_k)(\boldsymbol{m}_{k-1,t}^i - \nabla \ell(\boldsymbol{u}_{k-1}^i, \xi_{k,t}^i))$
      else
        $\boldsymbol{m}_{k,t}^i = \nabla \ell(\boldsymbol{u}_k^i, \xi_{k,t}^i)$
      end if
      $\boldsymbol{u}_{k+1,t}^i = \boldsymbol{u}_k^i - \alpha \boldsymbol{m}_{k,t}^i$
    end for
    Obtain data samples $\xi_{K,t}^i$ from the query data of task i
    $\boldsymbol{c}_t^i = \nabla_{\boldsymbol{w}_t} \ell(\boldsymbol{u}_k^i, \xi_{K,t}^i)$ (MAML), $\boldsymbol{c}_t^i = \nabla \ell(\boldsymbol{u}_k^i, \xi_{K,t}^i)$ (FOMAML), or $\boldsymbol{c}_t^i = \frac{1}{K\alpha}(\boldsymbol{w}_t - \boldsymbol{u}_k^i)$ (Reptile)
  end for
  $\boldsymbol{c}_t = \frac{1}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} \boldsymbol{c}_t^i$
  $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{c}_t$
end for