Convergence of First-Order Algorithms for Meta-Learning with Moreau Envelopes

Anonymous Authors[1]

Abstract

In this work, we consider the problem of minimizing the sum of Moreau envelopes of given functions, which has previously appeared in the context of meta-learning and personalized federated learning. In contrast to the existing theory that requires running subsolvers until a certain precision is reached, we only assume that a finite number of gradient steps is taken at each iteration. As a special case, our theory allows us to show the convergence of First-Order Model-Agnostic Meta-Learning (FO-MAML) to the vicinity of a solution of Moreau objective. We also study a more general family of first-order algorithms that can be viewed as a generalization of FO-MAML. Our main theoretical achievement is a theoretical improvement upon the inexact SGD framework. In particular, our perturbed-iterate analysis allows for tighter guarantees that improve the dependency on the problem's conditioning. In contrast to the related work on meta-learning, ours does not require any assumptions on the Hessian smoothness, and can leverage smoothness and convexity of the reformulation based on Moreau envelopes. Furthermore, to fill the gaps in the comparison of FO-MAML to the Implicit MAML (iMAML), we show that the objective of iMAML is neither smooth nor convex, implying that it has no convergence guarantees based on the existing theory.

1. Introduction

Efficient optimization methods for empirical risk minimization have helped the breakthroughs in many areas of machine learning such as computer vision (Krizhevsky et al., 2012) and speech recognition (Hinton et al., 2012). More recently, elaborate training algorithms have enabled fast progress in the area of meta-learning, also known as learning to learn (Schmidhuber, 1987). At its core lies the idea that one can find a model capable of retraining for a new task with just a few data samples from the task. Algorithmically, this corresponds to solving a bilevel optimization problem (Franceschi et al., 2018), where the inner problem corresponds to a single task, and the outer problem is that of minimizing the post-training error on a wide range of tasks.

The success of Model-Agnostic Meta-Learning (MAML) and its first-order version (FO-MAML) (Finn et al., 2017) in meta-learning applications has propelled the development of new gradient-based meta-learning methods. However, most new algorithms effectively lead to new formulations of meta-learning. For instance, iMAML (Rajeswaran et al., 2019) and proximal meta-learning (Zhou et al., 2019) define two MAML-like objectives with implicit gradients, while Reptile (Nichol et al., 2018) was proposed without defining any objective at all. These dissimilarities cause fragmentation of the field and make it particularly hard to have a clear comparison of meta-learning theory. Nonetheless, having a good theory helps to compare algorithms as well as identify and fix their limitations.

Unfortunately, for most of the existing methods, the theory is either incomplete as is the case with iMAML or even completely missing. In this work, we set out to at least partially mitigate this issue by proposing a new analysis for minimization of Moreau envelopes. We show that a general family of algorithms with multiple gradient steps is stable on this objective and, as a special case, we obtain results even for FO-MAML. Previously, FO-MAML was viewed as a heuristic to approximate MAML (Fallah et al., 2020), but our approach reveals that FO-MAML can be regarded as an algorithm for a the sum of Moreau envelopes. While both perspectives show only approximate convergence, the main justification for the sum of Moreau envelopes is that requires unprecedentedly mild assumptions. In addition, the Moreau formulation of meta-learning does not require Hessian information and is easily implementable by any

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

first-order optimizer, which Zhou et al. (2019) showed to give good empirical performance.

Due to the space constraints, we provide detailed literature comparison in Appendix A.1 and Table 1.

2. Background and mathematical formulation

Before we introduce the considered formulation of meta-learning, let us provide the problem background. As the notation in meta-learning varies, we correspond ours to that of works in the next subsection.

We assume that training is performed over n tasks with task losses $f_1, \dots, f_n$ and we will introduce *implicit* and *proximal* meta-losses $\{F_i\}$ in the next section. We denote by x the vector of parameters that we aim to train, which is often called *model*, *meta-model* or *meta-parameters* in the meta-learning literature, and *outer variable* in the bilevel literature. Similarly, given task i, we denote by z_i the *task-specific parameters* that are also called as *ground model*, *base-model*, or *inner variable*. We will use letters α, β, γ to denote scalar hyper-parameters such as stepsize or regularization coefficient. Given a function $\varphi(\cdot)$, we call the following function its *Moreau envelope*:

$$\Phi(x) = \min_{z \in \mathbb{R}^d} \left\{ \varphi(x) + \tfrac{1}{2\alpha} \|z - x\|^2 \right\},$$

where $\alpha > 0$ is some parameter. Given the Moreau envelope F_i of a task loss f_i, we denote by $z_i(x)$ the solution to the inner objective of F_i, i.e., $z_i(x) \overset{\text{def}}{=} \mathrm{argmin}_{z \in \mathbb{R}^d} \left\{ f_i(z) + \frac{1}{2\alpha} \|z - x\|^2 \right\}$.

Finally, let us introduce some standard function properties that are commonly used in the optimization literature.

Definition 1. *We say that a function $\varphi(\cdot)$ is* L-smooth *if its gradient is L-Lipschitz, i.e., for any $x, y \in \mathbb{R}^d$,*

$$\|\nabla \varphi(x) - \nabla \varphi(y)\| \le L \|x - y\|.$$

Definition 2. *Given a function $\varphi(\cdot)$, we call it* μ-strongly convex *if it satisfies for any $x, y \in \mathbb{R}^d$,*

$$\varphi(y) \ge \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \tfrac{\mu}{2} \|y - x\|^2.$$

If the property above holds with $\mu = 0$, we call φ to be convex. *If the property does not hold even with $\mu = 0$, we say that φ is* nonconvex.

2.1. Meta-learning objectives

Assume that we are given n tasks, and that the performance on task i is evaluated according to some loss function $f_i(x)$. MAML has been proposed as an algorithm for solving the following objective:

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x - \alpha \nabla f_i(x)), \tag{1}$$

where $\alpha > 0$ is a stepsize. Ignoring for simplicity minibatching, MAML update computes the gradient of a task meta-loss $\varphi_i(x) = f_i(x - \alpha \nabla f_i(x))$ through backpropagation and can be explicitly written as

$$x^{k+1} = x^k - \beta \left(\mathbf{I} - \alpha \nabla^2 f_i(x^k) \right) \nabla f_i(x^k - \alpha \nabla f_i(x^k)), \tag{MAML update}$$

where $\beta > 0$ is a stepsize, i is sampled uniformly from $\{1, \dots, n\}$ and $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix.

Algorithm 1 FO-MAML: First-Order MAML

```
Input: $x^0$, $\alpha, \beta > 0$
for $k = 0, 1, \dots$ do
    Sample a subset of tasks $T_k$
    for each sampled task $i$ in $T_k$ do
        $z_i^k = x^k - \alpha \nabla f_i(x^k)$
    end for
    $x^{k+1} = x^k - \beta \frac{1}{|T_k|} \sum_{i \in T_k} \nabla f_i(z_i^k)$
end for
```

Unfortunately, objective (1) might be nonsmooth and nonconvex even if the task losses $\{f_i\}$ are convex and smooth (Fallah et al., 2020). Moreover, generalizing this objective for more than one gradient step inside $f_i(\cdot)$ further deteriorates its smoothness properties and complicates the development of multistep methods.

To avoid differentiating through a graph, Rajeswaran et al. (2019) proposed an alternative objective iMAML that replaces the gradient step inside each function with an *implicit* gradient step. In particular, if we define $z_i(x) \overset{\text{def}}{=} \mathrm{argmin}_{z \in \mathbb{R}^d} \left\{ f_i(z) + \frac{1}{2\alpha} \|z - x\|^2 \right\}$, and the objective

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i\left(x - \alpha \nabla f_i(z_i(x))\right). \tag{iMAML objective}$$

However, it is not shown in (Rajeswaran et al., 2019) if the objective of iMAML is solvable. As a sign that the problem is rather ill-designed, we provide negative examples on the problem's convexity and smoothness. We are not aware of any result showing when the problem is convex or smooth.

Theorem 1. *There exists a convex function f with Lipschitz gradient and Lipschitz Hessian such that the iMAML meta-objective $\varphi(x) \overset{\text{def}}{=} f(z(x))$ is nonconvex, where $z(x) = x - \alpha \nabla f(z(x))$.*

Theorem 2. *There exists a convex function f with Lipschitz gradient and Lipschitz Hessian such that the iMAML meta-objective $\varphi(x) \overset{\text{def}}{=} f(z(x))$ is nonsmooth for any $\alpha > 0$, where $z(x) = x - \alpha \nabla f(z(x))$.*

2.2. Our objective: Moreau envelopes

We consider the following formulation of meta-learning

$$\min_{x\in\mathbb{R}^d} F(x) \overset{\text{def}}{=} \tfrac{1}{n}\sum_{i=1}^n F_i(x), \tag{2}$$
$$\text{where}\quad F_i(x) \overset{\text{def}}{=} \min_{z\in\mathbb{R}^d}\left\{f_i(z)+\tfrac{1}{2\alpha}\|z-x\|^2\right\},$$

and $\alpha > 0$ is a parameter controlling the level of adaptation to the problem. In other words, we seek to find a parameter vector x such that somewhere close to x there exists a vector z_i that verifies that $f_i(z)$ is sufficiently small. This formulation of meta-learning was first introduced by Zhou et al. (2019) and it has been used by Hanzely et al. (2020) and T. Dinh et al. (2020) to study personalization in federated learning.

We denote the solution to Problem (2) as $x^* \overset{\text{def}}{=} \arg\min_{x\in\mathbb{R}^d} F(x)$, and we can express the difficulty of (2) by gradient variance at the optimum,

$$\sigma_*^2 \overset{\text{def}}{=} \tfrac{1}{n}\sum_{i=1}^n \|\nabla F_i(x^*)\|^2. \tag{3}$$

Note that σ_* is always finite because it is defined on a single point, in contrast to the *maximum* gradient variance over all space, which might be infinite. For $i = 1, \dots, n$ we use the following variables for minimizers of meta-problems F_i:

$$z_i(x) \overset{\text{def}}{=} \operatorname*{argmin}_{z\in\mathbb{R}^d}\left\{f_i(z)+\tfrac{1}{2\alpha}\|z-x\|^2\right\}. \tag{4}$$

Notice that if $\alpha \to 0$, then $F_i(x) \approx f_i(x)$, and Problem (2) reduces to the well-known empirical risk minimization:

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \tfrac{1}{n}\sum_{i=1}^n f_i(x).$$

If, on the other hand, $\alpha \to +\infty$, the minimization problem in (2) becomes essentially independent of x and it holds $z_i(x) \approx \operatorname{argmin}_{z\in\mathbb{R}^d} f_i(z)$. Thus, one has to treat the parameter α as part of the objective that controls the similarity between the task-specific parameters. For notational simplicity, we keep α constant throughout the paper and do not explicitly write the dependence of $x^*, F, F_1, z_1, \dots, F_n, z_n$ on α.

Proposition 1 from (Beck, 2017) shows that convex, proper and closed f_i lead to differentiable and $\frac{1}{\alpha}$-smooth F_i. However, the tasks in meta-learning are often defined by training a neural network with nonconvex landscapes. Therefore, we refine Proposition 1 for such application and also improve the smoothness constant in the convex case.

Lemma 1. *Let function f_i be L-smooth.*
- *If f_i is nonconvex and $\alpha < \frac{1}{L}$, then F_i is $\frac{L}{1-\alpha L}$-smooth. If $\alpha \le \frac{1}{2L}$, then F_i is $2L$-smooth.*
- *If f_i is convex, then F_i is $\frac{L}{1+\alpha L}$-smooth. Moreover, for any α, it is L-smooth.*
- *If f_i is μ-strongly convex, then F_i is $\frac{\mu}{1+\alpha\mu}$-strongly convex. If $\alpha \le \frac{1}{\mu}$, then F_i is $\frac{\mu}{2}$-strongly convex.*

Whenever F_i is smooth, its gradient is given as in equation (7), i.e., $\nabla F_i(x) = \nabla f_i(z_i(x))$.

The takeaway message of Lemma 1 is that the optimization properties of F_i are always at least as good as those of f_i (up to constant factors). Furthermore, if f_i is convex but nonsmooth ($L \to +\infty$), F_i is still smooth with constant $\frac{1}{\alpha}$.

Finally, note that computing the exact gradient of F_i requires solving its inner problem as per equation (7). Even if the gradient of task $\nabla f_i(x)$ is easy to compute, we still cannot obtain $\nabla F_i(x)$ through standard differentiation or backpropagation. However, one can approximate $\nabla F_i(x)$ in various ways, as we will discuss later.

We can show that FO-MAML approximates SGD for objective (2) with error proportional to the stochastic gradient norm.

Lemma 2. *Let task losses f_i be L–smooth and $\alpha > 0$. Given i and $x \in \mathbb{R}^d$, we define recursively $z_{i,0} \overset{def}{=} x$ and $z_{i,j+1} \overset{def}{=} x - \alpha\nabla f_i(z_{i,j})$. Then, it holds for any $s \ge 0$*

$$\|\nabla f_i(z_{i,s}) - \nabla F_i(x)\| \le (\alpha L)^{s+1}\|\nabla F_i(x)\|.$$

In particular, the iterates of FO-MAML (Algorithm 1) satisfy for any k

$$\left\|\nabla f_i(z_i^k) - \nabla F_i(x^k)\right\| \le (\alpha L)^2\|\nabla F_i(x^k)\|.$$

3. Analysis of FO-MAML as inexact SGD

The prior literature views FO-MAML as an inexact version of MAML for Problem (1). Even though we are interested in Problem (2), one can use idea of inexact SGD to obtain convergence guarantees even for (2). We can write

$$\nabla f_i(z_i^k) = \nabla F(x^k) + \underbrace{\nabla F_i(x^k) - \nabla F(x^k)}_{\overset{\text{def}}{=}\xi_i^k \text{ (noise)}} + \underbrace{b_i^k}_{\text{bias}},$$

where it holds $\mathbb{E}[\xi_i^k] = 0$, and b_i^k is a bias vector that also depends on i but does not have zero mean. The best known guarantees (Ajalloeian and Stich, 2020) for inexact SGD are not applicable as they require independence of ξ_i^k and b_i^k. The analysis of Zhou et al. (2019) is not applicable either because their inexactness assumption requires the error to be smaller than a predefined constant ε, while the error in Lemma 2 can be unbounded. To resolve these issues, we provide a refined analysis in the next subsection.

Algorithm 2 FO-MuML: First-Order Multistep Meta-Learning (general formulation)

```
Input: $x^0$, $\beta > 0$, accuracy $\delta \ge 0$ or $\varepsilon \ge 0$.
for $k = 0, 1, \dots$ do
    Sample a subset of tasks $T_k$
    for each sampled task $i$ in $T_k$ do
        Find $z_i^k$ s.t. $\left\|\frac{1}{\alpha}\left(x^k - z_i^k\right) - \nabla F_i(x^k)\right\| \le \delta \left\|\nabla F_i(x^k)\right\|$    ▷ E.g., Lemma 2 finds such $\delta$.
    end for
    $x^{k+1} = x^k - \beta \frac{1}{|T_k|}\sum_{i\in T_k} \nabla f_i(z_i^k)$
end for
```

3.1. Strongly convex inexact SGD

For strongly convex objectives, we can modify analysis of Ajalloeian and Stich (2020) for our purpose.

Theorem 3. *Let losses $f_1, \dots, f_n$ be μ-strongly convex and L-smooth. If $|T_k| = \tau$ for all k, $\alpha \le \frac{1}{L}, \beta \le \frac{1}{20L}, \delta \le \frac{1}{4\sqrt{\kappa}}$ and $\kappa \overset{def}{=} \frac{L}{\mu}$, then iterates of Algorithm 2 satisfy*

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \le \left(1 - \frac{\beta\mu}{4}\right)^k \|x^0 - x^*\|^2 + \frac{16}{\mu}\left(\frac{2\delta^2}{\mu} + \frac{\beta}{\tau} + \beta\delta^2\right)\sigma_*^2. \tag{5}$$

In particular, if $\alpha \le \frac{1}{4\sqrt{\kappa}L}$, then the iterates x^k of FO-MAML (Algorithm 1) satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \le \left(1 - \frac{\beta\mu}{4}\right)^k \|x^0 - x^*\|^2 + \frac{16}{\mu}\left(\frac{2\alpha^2 L^2}{\mu} + \frac{\beta}{\tau} + \beta\right)\sigma_*^2. \tag{6}$$

Comparing (6) to the rate of vanilla SGD (studied in Gower et al. (2019)), the first term decreases exponentially as well. The second term decreases only if we decrease β and α. Decreasing β corresponds to using decreasing stepsizes in SGD, which is fine, but α defines objective, so it is fixed.

Choosing δ according to Lemma 2 in Algorithm 2 leads to better rate than of Algorithm 1, as **i)** it allows for larger α and **ii)** decreasing convergence neighborhood requires the inexactness parameter δ to go to 0.

3.2. Improved theory via virtual iterates

We show that the convergence theory can be improved by considering a sequence of virtual iterates that appear only in the analysis. The main difficulty of the analysis is we use $\{\nabla f_i\}$ instead of desired but inaccessible meta-gradients $\{\nabla F_i\}$. We aim to use virtual iterates to bridge this gap as

$$x^{k+1} = x^k - \frac{\alpha}{\tau}\sum_{i\in T_k} \nabla f_i(z_i^k) = x^k - \frac{\alpha}{\tau}\sum_{i\in T_k} \nabla F_i(y_i^k),$$

for some point y_i^k. Such y_i^k lead to $\nabla F_i(y_i^k) \approx \nabla F_i(x^k)$,

$$x^{k+1} = x^k - \frac{\alpha}{\tau}\sum_{i\in T_k} \nabla F_i(y_i^k) \approx x^k - \frac{\alpha}{\tau}\sum_{i\in T_k} \nabla F_i(x^k),$$

and it would allow us to better bound the bias. Fortunately, Moreau Envelopes (2) allows us to find such point easily.

Lemma 3. *For any points $z, y \in \mathbb{R}^d$ it holds $y = z + \alpha\nabla f_i(z)$ if and only if $z = y - \alpha\nabla F_i(y)$. Therefore, given z, for $y \overset{def}{=} z + \alpha\nabla f_i(z)$ holds $\nabla f_i(z) = \nabla F_i(y)$.*

This allows us to write

$$\nabla f_i(z_i^k) = \nabla F_i(y_i^k) = \nabla F(x^k) + \underbrace{\nabla F_i(x^k) - \nabla F(x^k)}_{\text{noise}} + \underbrace{\nabla F_i(y_i^k) - \nabla F_i(x^k)}_{\text{reduced bias}},$$

which leads to convergence to a smaller neighborhood, $\mathcal{O}\left(\frac{\frac{\beta}{\tau} + \alpha^2 L}{\mu}\right)$ in contrast to $\mathcal{O}\left(\frac{\beta + \kappa\alpha^2 L}{\mu}\right)$.

Theorem 4. *Consider the iterates of Algorithm 2 (with general δ) or Algorithm 1 (for which $\delta = \alpha L$). Let task losses be L–smooth and μ–strongly convex and let objective parameter satisfy $\alpha \le \frac{1}{\sqrt{6}L}$. Choose stepsize $\beta \le \frac{\tau}{4L}$, where $\tau = |T_k|$ is the batch size. Then we have*

$$\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] \le \left(1 - \frac{\beta\mu}{12}\right)^k \left\|x^0 - x^*\right\|^2 + \frac{6}{\mu}\left(\frac{\beta}{\tau} + 3\delta^2\alpha^2 L\right)\sigma_*^2.$$

3.3. Nonconvex convergence

Theorem 5. *Let variance of meta-loss gradients is uniformly bounded $\mathbb{E}\left[\|\nabla F_i(x) - \nabla F(x)\|^2\right] \le \sigma^2$, functions $f_1, \dots, f_n$ be L–smooth and F be lower bounded by $F^* > -\infty$. Assume $\alpha \le \frac{1}{4L}, \beta \le \frac{1}{16L}$. If we consider the iterates of Algorithm 1 (with $\delta = \alpha L$) or Algorithm 2 (with general δ), then*

$$\min_{t\le k} \mathbb{E}\left[\|\nabla F(x^t)\|^2\right] \le \frac{4}{\beta k}\mathbb{E}\left[F(x^0) - F^*\right] + 4(\alpha L)^2\delta^2\sigma^2 + 32\beta(\alpha L)^2\left(\frac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2.$$

The uniform bound assumption on meta-loss gradients variance is stronger than one in (3). Yet, it is very common in literature on stochastic optimization when studying convergence on nonconvex functions.

To make cconvergence radius smaller than some given target accuracy $\varepsilon > 0$, Algorithm 3 needs at most $s = \mathcal{O}(\log\frac{1}{\varepsilon})$ inner-loop steps. FO-MAML converges to a neighborhood of size $\mathcal{O}((\alpha L)^4)$.

References

Ajalloeian, A. and Stich, S. U. (2020). Analysis of SGD with biased gradient estimators. *arXiv preprint arXiv:2008.00051*. (Cited on pages 3, 4, and 7)

Antoniou, A., Edwards, H., and Storkey, A. J. (2018). How to train your MAML. In *International Conference on Learning Representations*. (Cited on page 7)

Bai, Y., Chen, M., Zhou, P., Zhao, T., Lee, J., Kakade, S., Wang, H., and Xiong, C. (2021). How important is the train-validation split in meta-learning? In *International Conference on Machine Learning*, pages 543–553. PMLR. (Cited on page 8)

Balcan, M.-F., Khodak, M., and Talwalkar, A. (2019). Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR. (Cited on page 8)

Beck, A. (2017). *First order methods in optimization*. MOS-SIAM Series on Optimization. (Cited on pages 3, 8, and 13)

Davis, D. and Drusvyatskiy, D. (2021). Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, pages 1–46. (Cited on page 9)

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092. PMLR. (Cited on pages 1, 2, 7, 8, and 9)

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR. (Cited on pages 1, 7, and 8)

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR. (Cited on page 7)

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. (2018). Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR. (Cited on pages 1 and 7)

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: general analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 5200–5209. PMLR. (Cited on page 4)

Hanzely, F., Hanzely, S., Horváth, S., and Richtárik, P. (2020). Lower bounds and optimal algorithms for personalized federated learning. *arXiv preprint*. (Cited on page 3)

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97. (Cited on page 1)

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. (2021). Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*. (Cited on page 7)

Ji, K., Yang, J., and Liang, Y. (2020). Theoretical convergence of multi-step model-agnostic meta-learning. *arXiv e-prints*, pages arXiv–2002. (Cited on page 8)

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. (2019). Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32. (Cited on page 8)

Konobeev, M., Kuzborskij, I., and Szepesvári, C. (2021). A distribution-dependent analysis of meta-learning. In *International Conference on Machine Learning*. PMLR. (Cited on page 7)

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105. (Cited on page 1)

Mania, H., Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., and Jordan, M. I. (2017). Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229. (Cited on page 9)

Mishchenko, K., Khaled, A., and Richtárik, P. (2020). Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320. (Cited on page 9)

Nichol, A., Achiam, J., and Schulman, J. (2018). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*. (Cited on pages 1, 7, and 8)

Planiden, C. and Wang, X. (2016). Strongly convex functions, Moreau envelopes, and the generic nature of convex functions with strong minimizers. *SIAM Journal on Optimization*, 26(2):1341–1364. (Cited on pages 12 and 13)

Poliquin, R. and Rockafellar, R. T. (1996). Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838. (Cited on page 13)

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. (2019). Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, pages 113–124. (Cited on pages 1, 2, 7, 8, and 11)

Schmidhuber, J. (1987). Evolutionary principles in self-referential learning. Diploma thesis, Technische Universität München. (Cited on page 1)

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified SGD with memory. *Advances in Neural Information Processing Systems*, 31:4447–4458. (Cited on page 9)

Sun, Y., Chen, T., and Yin, W. (2021). An optimal stochastic compositional optimization method with applications to meta learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3665–3669. IEEE. (Cited on page 7)

T. Dinh, C., Tran, N., and Nguyen, T. D. (2020). Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, 33. (Cited on page 3)

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. (2018). Bayesian model-agnostic meta-learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7343–7353. (Cited on page 7)

Zhou, P., Yuan, X., Xu, H., Yan, S., and Feng, J. (2019). Efficient meta learning via minibatch proximal update. *Advances in Neural Information Processing Systems*, 32:1534–1544. (Cited on pages 1, 2, 3, 7, and 8)

Zhou, P., Zou, Y., Yuan, X., Feng, J., Xiong, C., and Hoi, S. C. (2020). Task similarity aware meta learning: Theory-inspired improvement on MAML. In *4th Workshop on Meta-Learning at NeurIPS*. (Cited on page 7)

A. Content left out

Table of frequently used notation

For clarity, we provide a table of frequently used notation.

Notation	Meaning
f_i	The loss of task i
$F_i(x) = \min_z\{f_i(z) + \frac{1}{2\alpha}\|z - x\|^2\}$	Meta-loss
$F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x)$	Full meta loss
$z_i(x) = \operatorname{argmin}_z\{f_i(z) + \frac{1}{2\alpha}\|z - x\|^2\}$	The minimizer of regularized loss
L, μ	Smoothness and strong convexity constants of f_i
L_F	Smoothness constant of F
α	Objective parameter
β	Stepsize of the outer loop
γ, s	Stepsize and number of steps in the inner loop
δ	Precision of the proximal oracle

A.1. Related work

MAML (Finn et al., 2017) has attracted a lot of attention due to its success in practice. Many improvements have been proposed for MAML, for instance, (Zhou et al., 2020) suggested augmenting each group of tasks with its own global variable, and (Antoniou et al., 2018) proposed MAML++ that uses intermediate task losses with weights to improve the stability of MAML. (Rajeswaran et al., 2019) proposed iMAML that makes the objective optimizer-independent by relying on *implicit* gradients. Zhou et al. (2019) used a similar implicit objective to that of iMAML with an additional regularization term that, unlike iMAML, does not require inverting matrices. Reptile (Nichol et al., 2018) is an even simpler method that merely runs gradient descent on each sampled task. Based on generalization guarantees, (Zhou et al., 2020) also provided a trade-off between the optimization and statistical errors for a multi-step variant MAML, which shows that it may not improve significantly from increasing the number of gradient steps in the inner loop. We refer to (Hospedales et al., 2021) for a recent survey of the literature on meta-learning with neural networks.

On the theoretical side, the most relevant works to ours is that of (Zhou et al., 2019), whose main limitation is that it requires a high-precision solution of the inner problem in Moreau envelope at each iteration. Another relevant work that studied convergence of MAML and FO-MAML on the standard MAML objective is by (Fallah et al., 2020), but they do not provide any guarantees for the sum of Moreau envelopes and their assumptions are more stringent. Fallah et al. (2020) also study a Hessian-free variant of MAML, but its convergence guarantees still require posing assumptions on the Hessian Lipschitzness and variance.

Some works treat meta-learning as a special case of compositional optimization (Sun et al., 2021) or bilevel programming (Franceschi et al., 2018) and develop theory for the more general problem. Unfortunately, both approaches lead to worse dependence on the conditioning numbers of both inner and outer objective, and provide very pessimistic guarantees. Bilevel programming, even more importantly, requires computation of certain inverse matrices, which is prohibitive in large dimensions. One could also view minimization-based formulations of meta-learning as instances of empirical risk minimization, for which FO-MAML can be seen as instance of inexact (biased) SGD. For example, (Ajalloeian and Stich, 2020) analyzed SGD with deterministic bias and some of our proofs are inspired by theirs, except in our problem the bias is not deterministic. We will discuss the limitations of their approach in the section on inexact SGD.

Several works have also addressed meta-learning from the statistical perspective, for instance, Yoon et al. (2018) proposed a Bayesian variant of MAML, and Finn et al. (2019) analyzed convergence of MAML in online learning. Another example is the work of Konobeev et al. (2021) who studied the setting of linear regression with task-dependent solutions that are sampled from same normal distribution. These directions are orthogonal to ours, as we want to study the optimization properties of meta-learning.

Table 1: A summary of related work and conceptual differences to our approach. We mark as "N/A" unknown properties that have not been established in prior literature or our work. We say that F_i "Preserves convexity" if for convex f_i, F_i is convex as well, which implies that F_i has no extra local minima or saddle points. We say that F_i "Preserves smoothness" if its gradients are Lipschitz whenever the gradients of f_i are, which corresponds to more stable gradients. We refer to (Fallah et al., 2020) for the claims regarding nonconvexity and nonsmoothness of the MAML objective.

Algorithm	F_i: meta-loss of task i	Hessian-free	Arbitrary number of steps	No matrix inversion	Preserves convexity	Preserves smoothness	Reference
MAML	$f_i(x - \alpha\nabla f_i(x))$	✗	✗	✓	✗	✗	(Finn et al., 2017)
Multi-step MAML	$f_i(GD(f_i, x))$[1]	✗	✓	✓	✗	✗	(Finn et al., 2017) (Ji et al., 2020)
iMAML[2]	$f_i(z_i(x))$, where $z_i(x) = x - \alpha\nabla f_i(z_i(x))$	✗	✓	✗	✗ (Theorem 1)	✗ (Theorem 2)	(Rajeswaran et al., 2019)
Reptile	N/A[3]	✓	✓	✓	N/A	N/A	(Nichol et al., 2018)
FO-MAML (original)	$f_i(x - \alpha\nabla f_i(x))$	✓	✗	✓	✗	✗	(Finn et al., 2017)
Meta-MinibatchProx	$\min\limits_{x_i}\{f_i(x_i) + \frac{1}{2\alpha}\|x_i - x\|^2\}$	✓	✗[4]	✓	✓	✓	(Zhou et al., 2019)
FO-MuML (extended FO-MAML)	$\min\limits_{x_i}\{f_i(x_i) + \frac{1}{2\alpha}\|x_i - x\|^2\}$	✓	✓	✓	✓	✓	**This work**

[1] Multi-step MAML runs an inner loop with gradient descent applied to task loss f_i, so the objective of multi-step MAML is $F_i(x) = f_i(x_s(x))$, where $x_0 = x$ and $x_{j+1} = x_j - \alpha\nabla f_i(x_j)$ for $j = 0, \ldots, s-1$.

[2] To the best of our knowledge, iMAML is not guaranteed to work; Rajeswaran et al. (2019) studied only the approximation error for gradient computation, see the discussion in our special section on iMAML.

[3] Reptile was proposed as an algorithm on its own, without providing any optimization problem. This makes it hard to say how it affects smoothness and convexity. Balcan et al. (2019) and Khodak et al. (2019) studied convergence of Reptile on the average loss over the produced iterates, i.e., $F_i(x) = \frac{1}{m}\sum_{j=0}^{s} f_i(x_j)$, where $x_0 = x$ and $x_{j+1} = x_j - \alpha\nabla f_i(x_j)$ for $j = 0, \ldots, s-1$. Analogously to the loss of MAML, this objective seems nonconvex and nonsmooth.

[4] Zhou et al. (2019) assumed that the subproblems are solved to precision ε, i.e., x_i is found such that $\|\nabla f_i(x_i) + \frac{1}{\alpha}(x_i - x)\| \le \varepsilon$ with an absolute constant ε.

A.2. MAML objective remark

Sometimes, MAML update evaluates the gradient of φ_i using an additional data sample, but Bai et al. (2021) recently showed that this is often unnecessary.

iMAML objective remarks

The idea of iMAML is to optimize this objective during training so that at inference, given a new function f_{n+1} and solution x_{iMAML} of the problem above, one can find an approximate solution to $\min_{z\in\mathbb{R}^d}\left\{f_{n+1}(z) + \frac{1}{2\alpha}\|z - x_{\mathrm{iMAML}}\|^2\right\}$ and use it as a new model for task f_{n+1}.

Rajeswaran et al. (2019) proved, under some mild assumptions, that one can efficiently obtain an estimate of the gradient of $\varphi_i(x) \overset{\text{def}}{=} f_i\left(x - \alpha\nabla f_i(z_i(x))\right)$ with access only to gradients and Hessian-vector products of f_i, which rely on standard backpropagation operations. In particular, Rajeswaran et al. (2019) showed that

$$\nabla\varphi_i(x) = \left(\mathbf{I} + \alpha\nabla^2 f_i(z(x))\right)^{-1}\nabla f_i(z(x)),$$

where $\mathbf{I}$ is the identity matrix, and they proposed to run the conjugate gradient method to find $\nabla\varphi_i(x)$.

A.3. Moreau-Envelope objective remarks

Proposition 1 (Theorem 6.60 in (Beck, 2017))**.** *Let F_i and $z_i(x)$ be defined as in* (2) *and* (4)*. If f_i is convex, proper and closed, then F_i is differentiable and $\frac{1}{\alpha}$-smooth:*

$$\nabla F_i(x) = \frac{1}{\alpha}(x - z_i(x)) = \nabla f_i(z_i(x)), \tag{7}$$

$$\|\nabla F_i(x) - \nabla F_i(y)\| \le \frac{1}{\alpha}\|x - y\|. \tag{8}$$

Algorithm 3 FO-MuML (implementation according to Lemma 2)

```
1:  Input: $x^0$, number of steps $s$, $\alpha > 0$, $\beta > 0$
2:  for $k = 0, 1, \dots$ do
3:      Sample a subset of tasks $T_k$
4:      for each sampled task $i$ in $T_k$ do
5:          $z_{i,0}^k = x^k$
6:          for $l = 0, \dots, s-1$ do
7:              $z_{i,l+1}^k = x^k - \alpha \nabla f_i(z_{i,l}^k)$
8:          end for
9:          $z_i^k = z_{i,s}^k$
10:     end for
11:     $x^{k+1} = x^k - \beta \frac{1}{|T_k|} \sum_{i \in T_k} \nabla f_i(z_i^k)$
12: end for
```

Our refinement of Proposition 1, Lemma 1 is similar to Lemma 2.5 of (Davis and Drusvyatskiy, 2021), except their guarantee is a bit weaker because they consider more general assumptions.

A.4. Parametrization of the inner loop of Algorithm 3

Note that Algorithm 3 depends on only one parameter – β. We need to keep in mind that parameter α is fixed by the objective (2) and changing α shifts convergence neighborhood. Nevertheless, we can still investigate the case wehn α from (2) and α from Line 6 of Algorithm 3 are different, as we can see in the following remark.

Remark. *If we replace line 7 of Algorithm 3 by $z_{l+1}^k = x^k - \gamma \nabla f_i(z_{i,l}^k)$, we will have freedom to choose γ. However, if we choose stepsize $\gamma \neq \alpha$, then similar analysis to the proof of Lemma 2 yields*

$$\tfrac{1}{\gamma} \| z_{i,s}^k - (x^k - \gamma \nabla F_i(x^k)) \| \le \left((\gamma L)^s + |\alpha - \gamma| L \right) \| \nabla F_i(x^k) \|. \tag{9}$$

Note that in case $\gamma \neq \alpha$, we cannot set number of steps s to make the right-hand side of (9) smaller than $\delta \| \nabla F_i(x^k) \|$ when δ is small. In particular, increasing the number of local steps s will help only as long as $\delta > |\alpha - \gamma| L$.

This is no surprise, for the modified algorithm (using inner loop stepsize γ) will no longer be approximating $\nabla F_i(x^k)$. It will be exactly approximating $\nabla \tilde{F}_i(x^k)$, where $\tilde{F}_i(x) \overset{\text{def}}{=} \min_{z \in \mathbb{R}^d} \left\{ f_i(z) + \frac{1}{2\gamma} \| z - x \|^2 \right\}$ (see Lemma 2). Thus, choice of stepsize in the inner loop affects what implicit gradients do we approximate and also what objective we are minimizing.

A.5. Motivation for virtual iterate analysis

The literature on asynchronous optimization has established that getting gradient at a wrong point does not significantly worsen its rate of convergence (Mania et al., 2017). A similar analysis with additional virtual sequence was used in the so-called error-feedback for compression (Stich et al., 2018), where the goal of the sequence is to follow the path of *exact* gradients even if *compressed* gradients are used by the algorithm itself. Motivated by these observations, we set out to find a virtual sequence that could help us analyze FO-MAML.

The proof technique for this theorem also uses recent advances on the analysis of biased SGD methods by Mishchenko et al. (2020). In particular, we show that the three-point identity (provided in the Appendix) is useful for getting a tighter recursion.

A.6. Nonconvex analyses in literature

Our Theorem 5 is very similar to the one obtained by Fallah et al. (2020), except **i)** their convergence neighborhood depends on α as $\mathcal{O}(\alpha^2)$, whereas ours has better dependency $\mathcal{O}(\alpha^4)$, **ii)** ours theory does not require any assumptions on the Hessian smoothness and in addition **iii)** we study different objectives, as Fallah et al. (2020) does not consider Moreau envelopes.

A.7. Summary

We presented a new analysis of first-order meta-learning algorithms for minimization of Moreau envelopes. Our theory covers both nonconvex and strongly convex smooth losses and guarantees convergence of the family of methods covered by Algorithm 2. As a special case, all convergence bounds apply to Algorithm 3 with an arbitrary number of inner-loop steps. Compared to other results available in the literature, ours are more general as they hold with an arbitrary number of inner steps and do not require Hessian smoothness. The main theoretical difficulty we faced was the limitation of the inexact SGD framework, which we overcame by presenting a refined analysis using virtual iterates. As a minor contribution, we also pointed out that standard algorithms, such as SGD, are not immediately guaranteed to work on the iMAML objective, which might be nonconvex and nonsmooth even for convex and smooth losses. To show this, we presented examples of losses whose convexity and smoothness cease when the iMAML objective is constructed.

B. Proofs

B.1. Basic facts

For any vectors $a, b \in \mathbb{R}^d$ and scalar $\nu > 0$, Young's inequality states that

$$2\langle a, b\rangle \le \nu\|a\|^2 + \tfrac{1}{\nu}\|b\|^2. \tag{10}$$

Moreover, we have

$$\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2. \tag{11}$$

More generally, for a set of m vectors $a_1, \dots, a_m$ with arbitrary m, it holds

$$\Big\|\tfrac{1}{m}\sum_{i=1}^m a_i\Big\|^2 \le \tfrac{1}{m}\sum_{i=1}^m \|a_i\|^2. \tag{12}$$

For any random vector X we have

$$\mathbb{E}\left[\|X\|^2\right] = \|\mathbb{E}[X]\|^2 + \mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right]. \tag{13}$$

If f is L_f-smooth, then for any $x, y \in \mathbb{R}^d$, it is satisfied

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{L_f}{2}\|y - x\|^2. \tag{14}$$

Finally, for L_f-smooth and convex function f, it holds

$$f(x) \le f(y) + \langle \nabla f(x), x - y\rangle - \tfrac{1}{2L_f}\|\nabla f(x) - \nabla f(y)\|^2. \tag{15}$$

Proposition 2. *[Three-point identity] For any $u, v, w \in \mathbb{R}^d$, any f with its Bregman divergence $D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$, it holds*

$$\langle \nabla f(u) - \nabla f(v), w - v\rangle = D_f(v, u) + D_f(w, v) - D_f(w, u).$$

B.2. Proof of Theorem 1

Proof. The counterexample that we are going to use is given below:

$$\begin{aligned} f(x) &= \min\left\{\tfrac{1}{4}x^4 - \tfrac{1}{3}|x|^3 + \tfrac{1}{6}x^2, \tfrac{2}{3}x^2 - |x| + \tfrac{5}{12}\right\} \\ &= \begin{cases} \tfrac{1}{4}x^4 - \tfrac{1}{3}|x|^3 + \tfrac{1}{6}x^2, & \text{if } |x| \le 1, \\ \tfrac{2}{3}x^2 - |x| + \tfrac{5}{12}, & \text{otherwise.} \end{cases} \end{aligned}$$

See also Figure 1 for its numerical visualization.

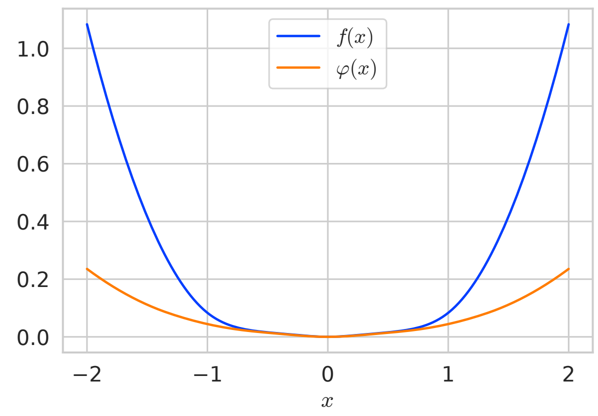

Figure 1: Values of functions f and φ.

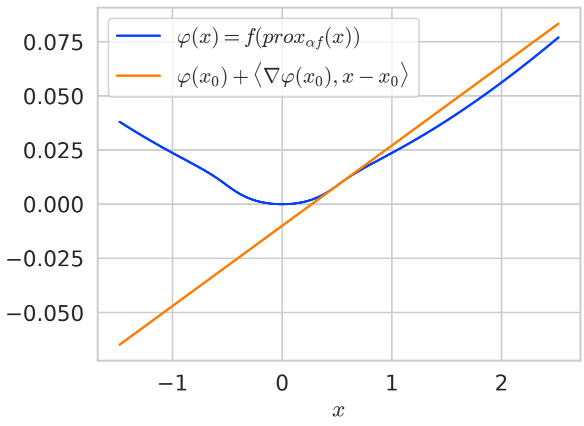

Figure 2: Illustration of nonconvexity: the value of φ goes below its tangent line from x_0, which means that φ is nonconvex at x_0.

It is straightforward to observe that this function is smooth and convex because its Hessian is

$$f''(x) = \begin{cases} 3x^2 - 2|x| + \frac{1}{3}, & \text{if } |x| \le 1, \\ \frac{4}{3}, & \text{otherwise.} \end{cases},$$

which is always nonnegative and bounded. However, the function $\varphi(x) = f(z(x))$ is not convex at point $x_0 = 0.4 + \alpha\nabla f(0.4)$, because its Hessian is negative, i.e., $\varphi''(x_0) < 0$, which we shall prove below. First of all, by definition of x_0, it holds that $0.4 = x_0 - \alpha\nabla f(0.4)$, which is equivalent to the definition of $z(x)$, implying $z(x_0) = 0.4$. Next, let us obtain the expression for the Hessian of φ. As shown in (Rajeswaran et al., 2019), it holds in general that

$$\nabla\varphi(x) = \frac{dz(x)}{dx}\nabla f(z(x)),$$

where $\frac{dz(x)}{dx}$ is the Jacobian matrix of the mapping $z(x)$. Differentiating this equation again, we obtain

$$\nabla^2\varphi(x) = \frac{d^2z(x)}{dx^2}\nabla f(z(x)) + \nabla^2 f(z(x))\frac{dz(x)}{dx}\left(\frac{dz(x)}{dx}\right)^\top.$$

Moreover, we can compute $\frac{d^2z(x)}{dx^2}$ by differentiating two times the equation $z(x) = x - \alpha\nabla f(z(x))$, which gives

$$\frac{dz(x)}{dx} = \mathbf{I} - \alpha\nabla^2 f(z(x))\frac{dz(x)}{dx},$$

where $\mathbf{I}$ is the identity matrix. Rearranging the terms in this equation yields

$$\frac{dz(x)}{dx} = (\mathbf{I} + \alpha\nabla^2 f(z(x)))^{-1}.$$

At the same time, if we do not rearrange and instead differentiate the equation again, we get

$$\frac{d^2z(x)}{dx^2} = -\alpha\nabla^2 f(z(x))\frac{d^2z(x)}{dx^2} - \alpha\nabla^3 f(z(x))\left[\frac{dz(x)}{dx}, \frac{dz(x)}{dx}\right],$$

where $\nabla^3 f(z(x))[\frac{dz(x)}{dx}, \frac{dz(x)}{dx}]$ denotes tensor-matrix-matrix product, whose result is a tensor too. Thus,

$$\frac{d^2z(x)}{dx^2} = -\alpha(\mathbf{I} + \alpha\nabla^2 f(z(x)))^{-1}\nabla^3 f(z(x))\left[\frac{dz(x)}{dx}, \frac{dz(x)}{dx}\right],$$

and, moreover,

$$\nabla^2\varphi(x) = -\alpha(\mathbf{I} + \alpha\nabla^2 f(z(x)))^{-1}\nabla^3 f(z(x))\left[\frac{dz(x)}{dx}, \frac{dz(x)}{dx}\right] + \nabla^2 f(z(x))\frac{dz(x)}{dx}\left(\frac{dz(x)}{dx}\right)^\top.$$

For any $x \in (0,1]$, our counterexample function satisfies $f''(x) = 3x^2 - 2x + \frac{1}{3}$ and $f'''(x) = 6x - 2$. Moreover, since $z(x_0) = 0.4$, we have $f''(z(x_0)) = \frac{1}{75}$, $f'''(z(x_0)) = \frac{2}{5}$, $\frac{dz(x)}{dx} = \frac{1}{1+\alpha/75}$, and

$$\varphi''(x) = -\frac{2\alpha}{5(1+\alpha/75)^3} + \frac{1}{75(1+\alpha/75)^2}.$$

It can be verified numerically that $\varphi''(x)$ is negative at x_0 for any $\alpha > \frac{75}{2249}$. Notice that this value of α is much smaller than the value of $\frac{1}{L} = \frac{3}{4}$, which can be obtained by observing that our counterexample satisfies $f''(x) \leq \frac{4}{3}$. □

Let us also note that obtaining nonconvexity of this objective for a fixed function and arbitrary α is somewhat challenging. Indeed, in the limit case $\alpha \to 0$, it holds that $\varphi(x)'' \to f''(x)$ for any x. If $f''(x) > 0$ then for a sufficiently small α it would also hold $\varphi''(x) > 0$. Finding an example that works for any α, thus, would require $f''(x_0) = 0$.

B.3. Proof of Theorem 2

Proof. Consider the following simple function

$$f(x) = \tfrac{1}{2}x^2 + \cos(x).$$

The Hessian of f is $f''(x) = 1 - \cos(x) \geq 0$, so it is convex. Moreover, it is apparent that the gradient and the Hessian of f are Lipschitz. However, we will show that the Hessian of φ is unbounded for any fixed $\alpha > 0$. To establish this, let us first derive some properties of $z(x)$. First of all, by definition $z(x)$ is the solution of $\alpha f'(z(x)) + (z(x) - x) = 0$, where by definition of f, it holds $f'(z(x)) = z(x) - \sin(z(x))$. Plugging it back, we get

$$(\alpha+1)z(x) - \alpha\sin(z(x)) = x.$$

Differentiating both sides with respect to x, we get $(\alpha+1)\frac{dz(x)}{dx} - \alpha\cos(z(x))\frac{dz(x)}{dx} = 1$ and

$$\frac{dz(x)}{dx} = \frac{1}{1+\alpha-\alpha\cos(z(x))}.$$

Thus, using the fact that $\varphi(x) = \varphi(z(x))$, we get

$$\varphi'(x) = \frac{d\varphi(x)}{dx} = \frac{df(z)}{dz}\frac{dz(x)}{dx} = \frac{z(x)-\sin(z(x))}{1+\alpha-\alpha\cos(z(x))}.$$

Denoting, for brevity, $z(x)$ as z, we differentiate this identity with respect to z and derive $\frac{d\varphi'(x)}{dz} = \frac{1+2\alpha-\alpha z\sin(z)-(1+2\alpha)\cos(z)}{(1+\alpha-\alpha\cos(z))^2}$. Therefore, for the Hessian of φ, we can produce an implicit identity,

$$\varphi''(x) = \frac{d^2\varphi(x)}{dx^2} = \frac{d\varphi'(x)}{dz}\frac{dz(x)}{dx} = \frac{1+2\alpha-\alpha z\sin(z)-(1+2\alpha)\cos(z)}{(1+\alpha-\alpha\cos(z))^3}.$$

The denominator of $\varphi''(x)$ satisfies $|1+\alpha-\alpha\cos(z)|^3 \leq (1+2\alpha)^3$, so it is bounded for any x. The numerator, on the other hand, is unbounded in terms of $z(x)$ since $|1+2\alpha-\alpha z\sin(z)-(1+2\alpha)\cos(z)| \geq \alpha|z\sin(z)| - 2(1+2\alpha)$. Therefore, $|\varphi''(x)|$ is unbounded. Moreover, $z(x)$ is itself unbounded, since the previously established identity for $z(x)$ can be rewritten as $|z(x)| = \left|\frac{1}{1+\alpha}x - \frac{\alpha}{1+\alpha}\sin(z(x))\right| \geq \frac{1}{1+\alpha}|x| - 1$. Therefore, $z(x)$ is unbounded, and since $\varphi''(x)$ grows with z, it is unbounded too. The unboundedness of $\varphi''(x)$ implies that φ is not L-smooth for any finite L. □

B.4. Proof of Lemma 1

Proof. The statement that F_i is $\frac{\mu}{1+\alpha\mu}$-strongly convex is proven as Lemma 2.19 in (Planiden and Wang, 2016), so we skip this part.

For nonconvex F_i and any $x \in \mathbb{R}^d$, we have by first-order stationarity of the inner problem that $\nabla F_i(x) = \nabla f_i(z_i(x))$, where $z_i(x) = \arg\min_z\{f_i(z) + \frac{1}{2\alpha}\|z-x\|^2\} = x - \alpha\nabla F_i(x)$. Therefore,

$$\begin{aligned}\|\nabla F_i(x) - \nabla F_i(y)\| &= \|\nabla f_i(z_i(x)) - \nabla f_i(z_i(y))\| \leq L\|z_i(x) - z_i(y)\| \\ &= L\|x - y - \alpha(\nabla F_i(x) - \nabla F_i(y))\| \\ &\leq L\|x-y\| + \alpha L\|\nabla F_i(x) - \nabla F_i(y)\|.\end{aligned}$$

Rearranging the terms, we get the desired bound:

$$\|\nabla F_i(x) - \nabla F_i(y)\| \le \tfrac{L}{1-\alpha L}\|x-y\|.$$

For convex functions, our proof of smoothness of F_i follows the exact same steps as the proof of Lemma 2.19 in (Planiden and Wang, 2016). Let f_i^* be the convex-conjugate of f_i. Then, it holds that $F_i = (f_i^* + \frac{\alpha}{2}\|\cdot\|^2)^*$, see Theorem 6.60 in (Beck, 2017). Therefore, $F_i^* = f_i^* + \frac{\alpha}{2}\|\cdot\|^2$. Since f_i is L-smooth, f_i^* is $\frac{1}{L}$-strongly convex. Therefore, F_i^* is $(\frac{1}{L}+\alpha)$-strongly convex, which, finally, implies that F_i is $\frac{1}{\frac{1}{L}+\alpha}$-smooth.

The statement $\frac{L}{1+\alpha L} \le L$ holds trivially since $\alpha > 0$. In case $\alpha \le \frac{1}{\mu}$, we get the constants from the other statements by mentioning that $\frac{\mu}{1+\alpha\mu} \ge \frac{\mu}{2}$.

The differentiability of F_i follows from Theorem 4.4 of Poliquin and Rockafellar (1996), who show differentiability assuming f_i is *prox-regular*, which is a strictly weaker property than L-smoothness, so it automatically holds under the assumptions of Lemma 1. □

B.5. Proof of Lemma 2

Lemma 2. Let task losses f_i be L–smooth and $\alpha > 0$. Given i and $x \in \mathbb{R}^d$, we define recursively $z_{i,0} = x$ and $z_{i,j+1} = x - \alpha\nabla f_i(z_{i,j})$. Then, it holds for any $s \ge 0$

$$\|\nabla f_i(z_{i,s}) - \nabla F_i(x)\| \le (\alpha L)^{s+1}\|\nabla F_i(x)\|.$$

In particular, the iterates of FO-MAML (Algorithm 1) satisfy for any k

$$\left\|\nabla f_i(z_i^k) - \nabla F_i(x^k)\right\| \le (\alpha L)^2\|\nabla F_i(x^k)\|.$$

Proof. First, observe that by eq. (7) it holds

$$z_i(x) = x - \alpha\nabla F_i(x) = x - \alpha\nabla f_i(z_i(x)).$$

For $s = 0$, the lemma's claim then follows from initialization, $z_{i,0} = x$, since

$$\|\nabla f_i(z_{i,s}) - \nabla F_i(x)\| = \|\nabla f_i(x) - \nabla f_i(z_i(x))\| \le L\|x - z_i(x)\| = \alpha L\|\nabla F_i(x)\|.$$

For $s > 0$, we shall prove the bound by induction. We have for any $l \ge 0$

$$\begin{aligned}\|z_{i,l+1} - (x - \alpha\nabla F_i(x))\| &= \alpha\|\nabla f_i(z_{i,l}) - \nabla F_i(x)\| = \alpha\|\nabla f_i(z_{i,l}) - \nabla f_i(z_i(x))\| \le \alpha L\|z_{i,l} - z_i(x)\| \\ &= \alpha L\|z_{i,l} - (x - \alpha\nabla F_i(x))\|.\end{aligned}$$

This proves the induction step as well as the lemma itself. □

Lemma 3 For any points $z, y \in \mathbb{R}^d$ it holds $y = z + \alpha\nabla f_i(z)$ if and only if $z = y - \alpha\nabla F_i(y)$. Therefore, given z, for $y \overset{\text{def}}{=} z + \alpha\nabla f_i(z)$ holds $\nabla f_i(z) = \nabla F_i(y)$.

Proof. The result follows immediately from the last statement of Lemma 1. □

Lemma 4. *If task losses $f_1, \dots, f_n$ are L-smooth and $\beta \le \frac{1}{L}$, then it holds*

$$\begin{aligned}\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k} g_i^k\Big\|^2 &\le \left(1 + 2(\alpha L)^{2s} + \tfrac{2}{|T|}\right)4L(F(x^k) - F(x^*)) + 4\left(\tfrac{1}{|T_k|} + (\alpha L)^{2s}\right)\sigma_*^2 &(16)\\ &\le 20L(F(x^k) - F(x^*)) + 4\left(\tfrac{1}{|T_k|} + \delta^2\right)\sigma_*^2. &(17)\end{aligned}$$

Proof. First, let us replace g_i^k with $\nabla F_i(x^k)$, which g_i^k approximates:

$$\begin{aligned}\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k} g_i^k\Big\|^2 &= \Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k) + \tfrac{1}{|T_k|}\sum_{i\in T_k}(g_i^k - \nabla F_i(x^k))\Big\|^2 \\ &\overset{(11)}{\le} 2\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\Big\|^2 + 2\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}(g_i^k - \nabla F_i(x^k))\Big\|^2 \\ &\overset{(12)}{\le} 2\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\Big\|^2 + \tfrac{2}{|T_k|}\sum_{i\in T_k}\|g_i^k - \nabla F_i(x^k)\|^2 \\ &\le 2\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\Big\|^2 + \tfrac{2}{|T_k|}\sum_{i\in T_k}\delta^2\|\nabla F_i(x^k)\|^2.\end{aligned}$$

Taking the expectation on both sides, we get

$$\mathbb{E}\left[\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k} g_i^k\Big\|^2\right] \overset{(13)}{\le} 2\|\nabla F(x^k)\|^2 + 2\mathbb{E}\left[\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k) - \nabla F(x^k)\Big\|^2\right] + \tfrac{2}{n}\sum_{i=1}^n \delta^2\|\nabla F_i(x^k)\|^2.$$

Moreover, each summand in the last term can be decomposed as

$$\|\nabla F_i(x^k)\|^2 \overset{(11)}{\le} 2\|\nabla F_i(x^*)\|^2 + 2\|\nabla F_i(x^k) - \nabla F_i(x^*)\|^2 \overset{(3)}{=} 2\sigma_*^2 + 2\|\nabla F_i(x^k) - \nabla F_i(x^*)\|^2.$$

Since F_i is convex and L-smooth, we have for any i

$$\|\nabla F_i(x^k) - \nabla F_i(x^*)\|^2 \le 2L(F_i(x^k) - F_i(x^*) - \langle\nabla F_i(x^*), x^k - x^*\rangle).$$

Averaging and using $\frac{1}{n}\sum_{i=1}^n \nabla F_i(x^*) = 0$, we obtain

$$\tfrac{1}{n}\sum_{i=1}^n \|\nabla F_i(x^k) - \nabla F_i(x^*)\|^2 \le 2L(F(x^k) - F(x^*)).$$

Thus,

$$\begin{aligned}\tfrac{2}{n}\sum_{i=1}^n \delta^2\|\nabla F_i(x^k)\|^2 &\le 4\delta^2\sigma_*^2 + 8L\delta^2(F(x^k) - F(x^*)) \\ &\le 4\delta^2\sigma_*^2 + 8L(F(x^k) - F(x^*)).\end{aligned} \tag{18}$$

Proceeding to another term in our initial bound, by independence of sampling $i \in T_k$ we have

$$\begin{aligned}\mathbb{E}\left[\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k) - \nabla F(x^k)\Big\|^2\right] &= \tfrac{1}{|T_k|}\tfrac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla F_i(x^k)\|^2\right] \\ &\overset{(11)}{\le} \tfrac{2}{|T_k|}\tfrac{1}{n}\sum_{i=1}^n \left(\mathbb{E}\left[\|\nabla F_i(x^k) - \nabla F_i(x^*)\|^2\right] + \mathbb{E}\left[\|\nabla F_i(x^*)\|^2\right]\right) \\ &\overset{(15)}{\le} \tfrac{2}{|T_k|}\left(2L(F(x^k) - F(x^*)) + \sigma_*^2\right) \\ &\le \tfrac{4L}{|T_k|}(F(x^k) - F(x^*)) + \tfrac{2}{|T_k|}\sigma_*^2.\end{aligned}$$

Finally, we also have $\|\nabla F(x^k)\|^2 \le 2L(F(x^k) - F(x^*))$. Combining all produced bounds, we get the claim

$$\Big\|\tfrac{1}{|T_k|}\sum_{i\in T_k} g_i^k\Big\|^2 \le \left(1 + 2\delta^2 + \tfrac{2}{|T|}\right)4L(F(x^k) - F(x^*)) + 4\left(\tfrac{1}{|T_k|} + \delta^2\right)\sigma_*^2. \tag{19}$$

□

B.6. Proof of Theorem 3

Theorem 3. Let task losses $f_1, \dots, f_n$ be L-smooth and μ-strongly convex. If $|T_k| = \tau$ for all k, $\alpha \le \frac{1}{L}, \beta \le \frac{1}{20L}$ and $\delta \le \frac{1}{4\sqrt{\kappa}}$, where $\kappa \overset{\text{def}}{=} \frac{L}{\mu}$, then the iterates of Algorithm 2 satisfy

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \le \left(1 - \frac{\beta\mu}{4}\right)^k \|x^0 - x^*\|^2 + \frac{16}{\mu}\left(\frac{2\delta^2}{\mu} + \frac{\beta}{\tau} + \beta\delta^2\right)\sigma_*^2.$$

Proof. For the iterates of Algorithm 2, we can write

$$x^{k+1} = x^k - \frac{\beta}{\tau}\sum_{i \in T_k} g_i^k.$$

We also have by Lemma 2 that

$$\|g_i^k - \nabla F_i(x^k)\|^2 \le (\alpha L)^2 \delta^2 \|\nabla F_i(x^k)\|^2 \le \delta^2 \|\nabla F_i(x^k)\|^2,$$

so let us decompose g_i^k into $\nabla F_i(x^k)$ and the approximation error:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \frac{2\beta}{\tau}\sum_{i \in T_k}\left\langle g_i^k, x^k - x^*\right\rangle + \beta^2 \left\|\frac{1}{\tau}\sum_{i \in T_k} g_i^k\right\|^2 \\ &= \|x^k - x^*\|^2 - \frac{2\beta}{\tau}\sum_{i \in T_k}\left\langle \nabla F_i(x^k), x^k - x^*\right\rangle + \frac{2\beta}{\tau}\sum_{i \in T_k}\left\langle \nabla F_i(x^k) - g_i^k, x^k - x^*\right\rangle + \beta^2 \left\|\frac{1}{\tau}\sum_{i \in T_k} g_i^k\right\|^2.\end{aligned}$$

First two terms can be upperbounded using strong convexity (recall that by Lemma 1, F_i is $\frac{\mu}{2}$-strongly convex):

$$\|x^k - x^*\|^2 - \frac{2\beta}{\tau}\sum_{i \in T_k}\left\langle \nabla F_i(x^k), x^k - x^*\right\rangle \le \left(1 - \frac{\beta\mu}{2}\right)\|x^k - x^*\|^2 - \frac{2\beta}{\tau}\sum_{i \in T_k}(F_i(x^k) - F_i(x^*)).$$

For the third term, we will need Young's inequality:

$$2\left\langle \nabla F_i(x^k) - g_i^k, x^k - x^*\right\rangle \overset{(10)}{\le} \frac{4}{\mu}\|\nabla F_i(x^k) - g_i^k\|^2 + \frac{\mu}{4}\|x^k - x^*\|^2 \le \frac{4}{\mu}\delta^2\|\nabla F_i(x^k)\|^2 + \frac{\mu}{4}\|x^k - x^*\|^2,$$

which we can scale by β and average over $i \in T_k$ to obtain

$$\frac{2\beta}{\tau}\sum_{i \in T_k}\left\langle \nabla F_i(x^k) - g_i^k, x^k - x^*\right\rangle \le \frac{4\beta\delta^2}{\mu}\frac{1}{\tau}\sum_{i \in T_k}\|\nabla F_i(x^k)\|^2 + \frac{\beta\mu}{4}\|x^k - x^*\|^2.$$

Plugging in upper bounds and taking expectation yields

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] &\le \left(1 - \frac{\beta\mu}{4}\right)\|x^k - x^*\|^2 - 2\beta(F(x^k) - F(x^*)) + \frac{4}{\mu}\beta\delta^2\frac{1}{n}\sum_{i=1}^n\|\nabla F_i(x^k)\|^2 + \beta^2\left\|\frac{1}{\tau}\sum_{i \in T_k} g_i^k\right\|^2 \\ &\overset{(17)}{\le} \left(1 - \frac{\beta\mu}{4}\right)\|x^k - x^*\|^2 - 2\beta(1 - 10\beta L)(F(x^k) - F(x^*)) + \frac{4}{\mu}\beta\delta^2\frac{1}{n}\sum_{i=1}^n\|\nabla F_i(x^k)\|^2 \\ &\quad + 4\beta^2\left(\frac{1}{\tau} + \delta^2\right)\sigma_*^2 \\ &\overset{(18)}{\le} \left(1 - \frac{\beta\mu}{4}\right)\|x^k - x^*\|^2 - 2\beta(1 - 10\beta L)(F(x^k) - F(x^*)) \\ &\quad + \frac{8}{\mu}\beta\delta^2\left(\sigma_*^2 + 2L(F(x^k) - F(x^*))\right) + 4\beta^2\left(\frac{1}{\tau} + \delta^2\right)\sigma_*^2 \\ &= \left(1 - \frac{\beta\mu}{4}\right)\|x^k - x^*\|^2 - 2\beta\left(1 - 10\beta L - \frac{8L}{\mu}\delta^2\right)(F(x^k) - F(x^*)) + \frac{8}{\mu}\beta\delta^2\sigma_*^2 + 4\beta^2\left(\frac{1}{\tau} + \delta^2\right)\sigma_*^2.\end{aligned}$$

By assumption $\beta \le \frac{1}{20L}, \delta \le \frac{1}{4\sqrt{\kappa}}$, we have $10\beta L \le \frac{1}{2}$ and $8\frac{L}{\mu}\delta^2 \le \frac{1}{2}$, so $1 - 10\beta L - \frac{8L}{\mu}\delta^2 \ge 0$, hence

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \le \left(1 - \frac{\beta\mu}{4}\right)\|x^k - x^*\|^2 + \frac{8}{\mu}\beta\delta^2\sigma_*^2 + 4\beta^2\left(\frac{1}{\tau} + \delta^2\right)\sigma_*^2.$$

Recurring this bound, which is a standard argument, we obtain the theorem's claim.

$$
\begin{aligned}
\mathbb{E}\left[\|x^k - x^*\|^2\right] &\le \left(1 - \tfrac{\beta\mu}{4}\right)^k \|x^0 - x^*\|^2 + \left(\tfrac{8}{\mu}\beta\delta^2\sigma_*^2 + 4\beta^2\left(\tfrac{1}{\tau} + \delta^2\right)\sigma_*^2\right) \frac{1-\left(1-\frac{\beta\mu}{4}\right)^k}{\frac{\beta\mu}{4}} \\
&\le \left(1 - \tfrac{\beta\mu}{4}\right)^k \|x^0 - x^*\|^2 + \tfrac{32}{\mu^2}\delta^2\sigma_*^2 + \tfrac{16}{\mu\tau}\beta\sigma_*^2 + \tfrac{16}{\mu}\beta\delta^2\sigma_*^2 \\
&\le \left(1 - \tfrac{\beta\mu}{4}\right)^k \|x^0 - x^*\|^2 + \tfrac{16}{\mu}\left(\tfrac{2\delta^2}{\mu} + \tfrac{\beta}{\tau} + \beta\delta^2\right)\sigma_*^2.
\end{aligned}
$$

□

B.7. Proof of Theorem 4

Theorem 4. Consider the iterates of Algorithm 2 (with general δ) or Algorithm 1 (for which $\delta = \alpha L$). Let task losses be L–smooth and μ–strongly convex and let objective parameter satisfy $\alpha \le \frac{1}{\sqrt{6}L}$. Choose stepsize $\beta \le \frac{\tau}{4L}$, where $\tau = |T_k|$ is the batch size. Then we have

$$
\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] \le \left(1 - \tfrac{\beta\mu}{12}\right)^k \left\|x^0 - x^*\right\|^2 + \frac{6\left(\frac{\beta}{\tau} + 3\delta^2\alpha^2 L\right)\sigma_*^2}{\mu}.
$$

Proof. Denote L_F, $\mu_F, \kappa_F = \frac{L_F}{\mu_F}$ smoothness constant, strong convexity constant, condition number of Meta-Loss functions $F_1, \dots, F_n$, respectively. We have

$$
\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &= \left\|x^k - x^* - \tfrac{\beta}{\tau}\sum_{i\in T_k}\nabla F_i(y_i^k)\right\|^2 \\
&= \left\|x^k - x^*\right\|^2 - \tfrac{2\beta}{\tau}\sum_{i\in T_k}\langle\nabla F_i(y_i^k), x^k - x^*\rangle + \beta^2\left\|\tfrac{1}{\tau}\sum_{i\in T_k}\nabla F_i(y_i^k)\right\|^2 \\
&\le \left\|x^k - x^*\right\|^2 - \tfrac{2\beta}{\tau}\sum_{i\in T_k}\langle\nabla F_i(y_i^k) - \nabla F_i(x^*), x^k - x^*\rangle + 2\beta^2\left\|\tfrac{1}{\tau}\sum_{i\in T_k}(\nabla F_i(y_i^k) - \nabla F_i(x^*))\right\|^2 \\
&\quad - \tfrac{2\beta}{\tau}\sum_{i\in T_k}\langle\nabla F_i(x^*), x^k - x^*\rangle + 2\beta^2\left\|\tfrac{1}{\tau}\sum_{i\in T_k}\nabla F_i(x^*)\right\|^2.
\end{aligned}
$$

Using Proposition 2, we rewrite the scalar product as $\langle\nabla F_i(y_i^k) - \nabla F_i(x^*), x^k - x^*\rangle = D_{F_i}(x^*, y_i^k) + D_{F_i}(x^k, x^*) - D_{F_i}(x^k, y_i^k)$, which gives

$$
\begin{aligned}
\left\|x^{k+1} - x^*\right\|^2 &\le \left\|x^k - x^*\right\|^2 - \tfrac{2\beta}{\tau}\sum_{i\in T_k}\left[D_{F_i}(x^*, y_i^k) + D_{F_i}(x^k, x^*) - D_{F_i}(x^k, y_i^k)\right] \\
&\quad + 2\beta^2\left\|\tfrac{1}{\tau}\sum_{i\in T_k}(\nabla F_i(y_i^k) - \nabla F_i(x^*))\right\|^2 - \tfrac{2\beta}{\tau}\sum_{i\in T_k}\langle\nabla F_i(x^*), x^k - x^*\rangle + 2\beta^2\left\|\tfrac{1}{\tau}\sum_{i\in T_k}\nabla F_i(x^*)\right\|^2.
\end{aligned}
$$

Since we sample T_k uniformly and $\{\nabla F_i(x^*)\}_{i\in T_k}$ are independent random vectors, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|^2\right] &\le \left\|x^k - x^*\right\|^2 + \tfrac{2\beta}{\tau}\mathbb{E}\left[\sum_{i\in T_k}\left[-D_{F_i}(x^*, y_i^k) - D_{F_i}(x^k, x^*) + D_{F_i}(x^k, y_i^k)\right]\right] \\
&\quad + \tfrac{2\beta^2}{\tau^2}\mathbb{E}\left[\sum_{i\in T_k}\left\|\nabla F_i(y_i^k) - \nabla F_i(x^*)\right\|^2\right] + \tfrac{2\beta^2}{\tau}\sigma_*^2.
\end{aligned}
$$

Next, we are going to use the following three properties of Bregman divergence:

$$\begin{aligned} -D_{F_i}(x^*, y_i^k) &\overset{(15)}{\le} -\tfrac{1}{2L_F} \left\|\nabla F_i(y_i^k) - \nabla F_i(x^*)\right\|^2 \\ -D_{F_i}(x^k, x^*) &\le -\tfrac{\mu_F}{2} \left\|x^k - x^*\right\|^2 \\ D_{F_i}(x^k, y_i^k) &\le \tfrac{L_F}{2} \left\|x^k - y_i^k\right\|^2. \end{aligned} \tag{20}$$

Moreover, using identity $y_i^k = z_i^k + \alpha \nabla F_i(y_i^k)$, we can bound the last divergence as

$$\begin{aligned} D_{F_i}(x^k, y_i^k) &\le \tfrac{L_F}{2} \left\|x^k - z_i^k - \alpha \nabla F_i(y_i^k)\right\|^2 \\ &= \tfrac{1}{2}\alpha^2 L_F \left\|\tfrac{1}{\alpha}(x^k - z_i^k) - \nabla F_i(y_i^k)\right\|^2 \\ &\le \tfrac{3}{2}\alpha^2 L_F \left(\left\|\tfrac{1}{\alpha}(x^k - z_i^k) - \nabla F_i(x^k)\right\|^2 + \left\|\nabla F_i(x^k) - \nabla F_i(x^*)\right\|^2 + \left\|\nabla F_i(x^*) - \nabla F_i(y_i^k)\right\|^2\right) \\ &\le \tfrac{3}{2}\alpha^2 L_F \left(\delta^2 \left\|\nabla F_i(x^k)\right\|^2 + \left\|\nabla F_i(x^k) - \nabla F_i(x^*)\right\|^2 + \left\|\nabla F_i(x^*) - \nabla F_i(y_i^k)\right\|^2\right), \end{aligned}$$

where the last step used the condition in Algorithm 2. Using inequality (11) on $\nabla F_i(x^k) = \nabla F_i(x^*) + (\nabla F_i(x^k) - \nabla F_i(x^*))$, we further derive

$$\begin{aligned} D_{F_i}(x^k, y_i^k) &\le \tfrac{3}{2}\alpha^2 L_F \left(2\delta^2 \left\|\nabla F_i(x^*)\right\|^2 + (1+2\delta^2) \left\|\nabla F_i(x^k) - \nabla F_i(x^*)\right\|^2 + \left\|\nabla F_i(x^*) - \nabla F_i(y_i^k)\right\|^2\right) \\ &\overset{(15)}{\le} \tfrac{3}{2}\alpha^2 L_F \left(2\delta^2 \left\|\nabla F_i(x^*)\right\|^2 + (1+2\delta^2) L_F D_{F_i}(x^k, x^*) + \left\|\nabla F_i(x^*) - \nabla F_i(y_i^k)\right\|^2\right). \end{aligned}$$

Assuming $\alpha \le \sqrt{\frac{2}{3}(1+2\delta^2)}\frac{1}{L_F}$ so that $1 - \frac{3}{2}\alpha^2 L_F^2 (1+2\delta^2) > 0$, we get

$$\begin{aligned} -D_{F_i}(x^k, x^*) + D_{F_i}(x^k, y_i^k) &\le -\left(1 - \tfrac{3}{2}\alpha^2 L_F^2 (1+2\delta^2)\right) D_{F_i}(x^k, x^*) \\ &\quad + \tfrac{3}{2}\alpha^2 L_F \left(2\delta^2 \left\|\nabla F_i(x^*)\right\|^2 + \left\|\nabla F_i(x^*) - \nabla F_i(y_i^k)\right\|^2\right) \\ &\overset{(20)}{\le} -\left(1 - \tfrac{3}{2}\alpha^2 L_F^2 (1+2\delta^2)\right) \tfrac{\mu_F}{2} \left\|x^k - x^*\right\|^2 \\ &\quad + \tfrac{3}{2}\alpha^2 L_F \left(2\delta^2 \left\|\nabla F_i(x^*)\right\|^2 + \left\|\nabla F_i(x^*) - \nabla F_i(y_i^k)\right\|^2\right). \end{aligned}$$

Plugging these inequalities yields

$$\begin{aligned} \mathbb{E}\left[\left\|x^{k+1} - x^*\right\|^2\right] &\le \left(1 - \beta\mu_F \left(1 - \tfrac{3}{2}\alpha^2 L_F^2 (1+2\delta^2)\right)\right) \left\|x^k - x^*\right\|^2 \\ &\quad + \tfrac{\beta}{\tau}\left(3\alpha^2 L_F + \tfrac{2\beta}{\tau} - \tfrac{1}{L_F}\right) \mathbb{E}\left[\sum_{i \in T_k} \left\|\nabla F_i(y_i^k) - \nabla F_i(x^*)\right\|^2\right] \\ &\quad + 2\beta\left(\tfrac{\beta}{\tau} + 3\alpha^2\delta^2 L_F\right)\sigma_*^2. \end{aligned}$$

Now, if $\alpha \le \frac{1}{\sqrt{6}L_F}$ and $\beta \le \frac{\tau}{4L_F}$, then $3\alpha^2 L_F + \frac{2\beta}{\tau} - \frac{1}{L_F} \le 0$, and consequently

$$\mathbb{E}\left[\left\|x^{k+1} - x^*\right\|^2\right] \le \left(1 - \beta\mu_F \left(1 - \tfrac{3}{2}\alpha^2 L_F^2 (1+2\delta^2)\right)\right) \left\|x^k - x^*\right\|^2 + 2\beta\left(\tfrac{\beta}{\tau} + 3\alpha^2\delta^2 L_F\right)\sigma_*^2.$$

We can unroll the recurrence to obtain the rate

$$\begin{aligned}\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] &\le \left(1 - \beta\mu_F\left(1 - \tfrac{3}{2}\alpha^2 L_F^2(1+2\delta^2)\right)\right)^k \left\|x^0 - x^*\right\|^2 \\ &\quad + \left(\sum_{i=0}^{k-1}\left(1 - \beta\mu_F\left(1 - \tfrac{3}{2}\alpha^2 L_F^2(1+2\delta^2)\right)\right)^i\right) 2\beta\left(\tfrac{\beta}{\tau} + 3\alpha^2\delta^2 L_F\right)\sigma_*^2 \\ &= \left(1 - \beta\mu_F\left(1 - \tfrac{3}{2}\alpha^2 L_F^2(1+2\delta^2)\right)\right)^k \left\|x^0 - x^*\right\|^2 \\ &\quad + \left(\frac{1-\left(1-\beta\mu_F\left(1-\frac{3}{2}\alpha^2 L_F^2(1+2\delta^2)\right)\right)^k}{1-\frac{3}{2}\alpha^2 L_F^2(1+2\delta^2)}\right)\tfrac{2}{\mu_F}\left(\tfrac{\beta}{\tau} + 3\alpha^2\delta^2 L_F\right)\sigma_*^2 \\ &\le \left(1 - \beta\mu_F\left(1 - \tfrac{3}{2}\alpha^2 L_F^2(1+2\delta^2)\right)\right)^k \left\|x^0 - x^*\right\|^2 + \frac{2\left(\frac{\beta}{\tau}+3\alpha^2\delta^2 L_F\right)\sigma_*^2}{\mu_F(1-\frac{3}{2}\alpha^2 L_F^2(1+2\delta^2))}.\end{aligned}$$

Choice of δ implies $0 \le \delta \le 1$; Proposition 1 yields $\frac{\mu}{2} \le \frac{\mu}{1+\alpha\mu} \le \mu_F$ and $L_F \le \frac{L}{1+\alpha L} \le L$, so we can simplify

$$\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] \le \left(1 - \tfrac{\beta\mu}{2}\left(1 - 5\alpha^2 L^2\right)\right)^k \left\|x^0 - x^*\right\|^2 + \frac{4\left(\frac{\beta}{\tau}+3\alpha^2 L\delta^2\right)\sigma_*^2}{\mu(1-2\alpha^2 L^2)}.$$

□

B.8. Proof of Theorem 5

Assumption 1. *We assume that the variance of meta-loss gradients is uniformly bounded by some σ^2, i.e.,*

$$\mathbb{E}\left[\|\nabla F_i(x) - \nabla F(x)\|^2\right] \le \sigma^2. \tag{21}$$

Theorem 5 Let Assumption 1 hold, functions $f_1, \dots, f_n$ be L–smooth and F be lower bounded by $F^* > -\infty$. Assume $\alpha \le \frac{1}{4L}, \beta \le \frac{1}{16L}$. If we consider the iterates of Algorithm 1 (with $\delta = \alpha L$) or Algorithm 2 (with general δ), then

$$\min_{t\le k} \mathbb{E}\left[\|\nabla F(x^t)\|^2\right] \le \tfrac{4}{\beta k}\mathbb{E}\left[F(x^0) - F^*\right] + 16\beta(\alpha L)^2\left(\tfrac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2.$$

Proof. We would like to remind the reader that for our choice of z_i^k and y_i^k, the following three identities hold. Firstly, by definition $y_i^k = z_i^k + \alpha\nabla f_i(z_i^k)$. Secondly, as shown in Lemma 3, $z_i^k = y_i^k - \alpha\nabla F_i(y_i^k)$. And finally, $\nabla f_i(z_i^k) = \nabla F_i(y_i^k)$. We will frequently use these identities in the proof.

Since functions $f_1, \dots, f_n$ are L-smooth and $\alpha \le \frac{1}{4L}$, functions $F_1, \dots, F_n$ are $(2L)$-smooth as per Lemma 1. Therefore, by smoothness of F, we have for the iterates of Algorithm 2

$$\begin{aligned}\mathbb{E}\left[F(x^{k+1})\right] &\overset{(14)}{\le} \mathbb{E}\left[F(x^k) + \langle\nabla F(x^k), x^{k+1} - x^k\rangle + L\|x^{k+1} - x^k\|^2\right] \\ &= \mathbb{E}\left[F(x^k) - \beta\left\langle\nabla F(x^k), \tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla f_i(z_i^k)\right\rangle + \beta^2 L\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla f_i(z_i^k)\right\|^2\right] \\ &= F(x^k) - \beta\|\nabla F(x^k)\|^2 + \beta\mathbb{E}\left[\left\langle\nabla F(x^k), \nabla F(x^k) - \tfrac{1}{n}\sum_{i=1}^n\nabla f_i(z_i^k)\right\rangle\right] \\ &\quad + \beta^2 L\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla f_i(z_i^k)\right\|^2\right] \\ &\overset{(11)}{\le} F(x^k) - \tfrac{\beta}{2}\|\nabla F(x^k)\|^2 + \tfrac{\beta}{2}\tfrac{1}{n}\sum_{i=1}^n\left\|\nabla F_i(x^k) - \nabla f_i(z_i^k)\right\|^2 + \beta^2 L\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla f_i(z_i^k)\right\|^2\right].\end{aligned}$$

Next, let us observe, similarly to the proof of Lemma 4, that the gradient approximation error satisfies

$$\begin{aligned}\left\|\nabla F_i(x^k) - \nabla f_i(z_i^k)\right\| &= \left\|\nabla F_i(x^k) - \nabla F_i(y_i^k)\right\| \le L\left\|x^k - y_i^k\right\| = L\left\|x^k - z_i^k - \alpha\nabla F_i(y_i^k)\right\| \\ &\le \alpha L\left\|\nabla F(x^k) - \nabla F_i(y_i^k)\right\| + \alpha L\left\|\tfrac{1}{\alpha}(x^k - z_i^k) - \nabla F_i(x^k)\right\| \\ &= \alpha L\left\|\nabla F(x^k) - \nabla f_i(z_i^k)\right\| + \alpha L\left\|\tfrac{1}{\alpha}(x^k - z_i^k) - \nabla F_i(x^k)\right\|.\end{aligned}$$

By rearranging and using our assumption on error δ as formulated in Algorithm 2, we have

$$\left\|\nabla F_i(x^k) - \nabla f_i(z_i^k)\right\| \le \tfrac{\alpha L}{1-\alpha L}\left\|\tfrac{1}{\alpha}(x^k - z_i^k) - \nabla F_i(x^k)\right\| \le \tfrac{\alpha L}{1-\alpha L}\delta\|\nabla F_i(x^k)\| \overset{\alpha\le\frac{1}{4L}}{\le} \tfrac{4}{3}\alpha L\delta\|\nabla F_i(x^k)\|.$$

Squaring this bound and averaging over i, we obtain

$$\begin{aligned}\tfrac{1}{n}\sum_{i=1}^n\left\|\nabla F_i(x^k) - \nabla f_i(z_i^k)\right\|^2 &\le \tfrac{16}{9}(\alpha L)^2\delta^2\tfrac{1}{n}\sum_{i=1}^n\|\nabla F_i(x^k)\|^2 \\ &= \tfrac{16}{9}(\alpha L)^2\delta^2\|\nabla F(x^k)\|^2 + \tfrac{16}{9}(\alpha L)^2\delta^2\tfrac{1}{n}\sum_{i=1}^n\|\nabla F_i(x^k) - \nabla F(x^k)\|^2 \\ &\overset{(21)}{\le} \tfrac{16}{9}(\alpha L)^2\delta^2\|\nabla F(x^k)\|^2 + \tfrac{16}{9}(\alpha L)^2\delta^2\sigma^2 \\ &\le \tfrac{1}{9}\|\nabla F(x^k)\|^2 + 2(\alpha L)^2\delta^2\sigma^2.\end{aligned}$$

For the other term in the smoothness upper bound, we can write

$$\begin{aligned}\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla f_i(z_i^k)\right\|^2\right] &= \mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k) + \tfrac{1}{|T_k|}\sum_{i\in T_k}(\nabla f_i(z_i^k) - \nabla F_i(x^k))\right\|^2\right] \\ &\overset{(11)}{\le} 2\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\right\|^2\right] + 2\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}(\nabla f_i(z_i^k) - \nabla F_i(x^k))\right\|^2\right] \\ &\overset{(12)}{\le} 2\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\right\|^2\right] + \tfrac{2}{|T_k|}\mathbb{E}\left[\sum_{i\in T_k}\|\nabla f_i(z_i^k) - \nabla F_i(x^k)\|^2\right] \\ &\le 2\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\right\|^2\right] + \mathbb{E}\left[\tfrac{32}{9}\tfrac{1}{|T_k|}\sum_{i\in T_k}(\alpha L)^2\delta^2\|\nabla F_i(x^k)\|^2\right].\end{aligned}$$

Using bias-variance decomposition, we get for the first term in the right-hand side

$$\begin{aligned}2\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k)\right\|^2\right] &\overset{(13)}{=} 2\|\nabla F(x^k)\|^2 + 2\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla F_i(x^k) - \nabla F(x^k)\right\|^2\right] \\ &= 2\|\nabla F(x^k)\|^2 + \tfrac{2}{|T_k|}\tfrac{1}{n}\sum_{i=1}^n\|\nabla F_i(x^k) - \nabla F(x^k)\|^2.\end{aligned}$$

Similarly, we simplify the second term using $\frac{32}{9} < 4$ and then obtain

$$\tfrac{32}{9}\mathbb{E}\left[\tfrac{1}{|T_k|}\sum_{i\in T_k}(\alpha L)^2\delta^2\|\nabla F_i(x^k)\|^2\right] \overset{(13)}{\le} 4(\alpha L)^2\delta^2\|\nabla F(x^k)\|^2 + \tfrac{4(\alpha L)^2\delta^2}{n}\sum_{i=1}^n\|\nabla F_i(x^k) - \nabla F(x^k)\|^2.$$

Thus, using $\alpha \le \frac{1}{4L}$ and $\delta \le 1$, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\tfrac{1}{|T_k|}\sum_{i\in T_k}\nabla f_i(z_i^k)\right\|^2\right] &\le 3\|\nabla F(x^k)\|^2 + \left(\tfrac{2}{|T_k|} + 4(\alpha L)^2\delta^2\right)\sum_{i=1}^n\|\nabla F_i(x^k) - \nabla F(x^k)\|^2 \\ &\overset{(21)}{\le} 3\|\nabla F(x^k)\|^2 + 4\left(\tfrac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2.\end{aligned}$$

Now we plug these inequalities back and continue:

$$\begin{aligned}\mathbb{E}\left[F(x^{k+1})\right] - F(x^k) &\le -\tfrac{\beta}{2}\|\nabla F(x^k)\|^2 + \tfrac{\beta}{18}\|\nabla F(x^k)\|^2 + \beta(\alpha L)^2\delta^2\sigma^2 \\ &\quad + 3\beta^2 L\|\nabla F(x^k)\|^2 + 4\beta^2 L\sigma^2\left(\tfrac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2 \\ &\overset{\beta\le\frac{1}{16L}}{\le} -\tfrac{\beta}{4}\|\nabla F(x^k)\|^2 + 4\beta^2 L\sigma^2\left(\tfrac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2 + \beta(\alpha L)^2\delta^2\sigma^2.\end{aligned}$$

Rearranging the terms and telescoping this bound, we derive

$$\begin{aligned}\min_{t\le k}\mathbb{E}\left[\|\nabla F(x^t)\|^2\right] &\le \tfrac{4}{\beta k}\mathbb{E}\left[F(x^0) - F(x^{k+1})\right] + 16\beta\left(\tfrac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2 + 4(\alpha L)^2\delta^2\sigma^2 \\ &\le \tfrac{4}{\beta k}\mathbb{E}\left[F(x^0) - F^*\right] + 16\beta\left(\tfrac{1}{|T_k|} + (\alpha L)^2\delta^2\right)\sigma^2 + 4(\alpha L)^2\delta^2\sigma^2.\end{aligned}$$

The result for Algorithm 1 is obtained as a special case with $\delta = \alpha L$. □