On the Stability and Generalization of Meta-Learning

Yunjuan Wang
Department of Computer Science
Johns Hopkins University
Baltimore, MD, 21218
ywang509@jhu.edu

Raman Arora
Department of Computer Science
Johns Hopkins University
Baltimore, MD, 21218
arora@cs.jhu.edu

Abstract

We focus on developing a theoretical understanding of meta-learning. Given multiple tasks drawn i.i.d. from some (unknown) task distribution, the goal is to find a good pre-trained model that can be adapted to a new, previously unseen, task with little computational and statistical overhead. We introduce a novel notion of stability for meta-learning algorithms, namely *uniform meta-stability*. We instantiate two uniformly meta-stable learning algorithms based on regularized empirical risk minimization and gradient descent and give explicit generalization bounds for convex learning problems with smooth losses and for weakly convex learning problems with non-smooth losses. Finally, we extend our results to stochastic and adversarially robust variants of our meta-learning algorithm.

1 Introduction

Traditional machine learning algorithms excel at generalizing, but they often require extensive training data and assume that both training and test data come from the same distribution or task. In real-world scenarios, large sets of training data from a single task are often lacking. Instead, training data may stem from diverse tasks with shared similarities, while test data come from entirely new tasks. The challenge is to rapidly adapt to these unseen tasks without the need to train from scratch.

To address this challenge, meta-learning, also referred to as learning-to-learn, has emerged as an effective approach. Meta-learning has gained significant attention recently [Hospedales et al., 2021], with applications spanning across various domains including computer vision [Nichol et al., 2018] and robotics [Al-Shedivat et al., 2017], ranging from few-shot classification [Snell et al., 2017], hyperparameter optimization [Franceschi et al., 2018], to personalized recommendation systems [Wang et al., 2022].

As the name suggests, meta-learning operates on two levels of abstraction to enhance learning over time. On an intra-task level, the learner needs to find models that perform well on individual tasks. On a meta-level, the learner needs to figure out useful meta-information, perhaps a prior over tasks, that relates different tasks and allows transferring and adaptation of knowledge to new unseen tasks efficiently (both in terms of statistical as well computational overhead). It is typical to represent such meta-information in the form of a pre-trained model, which we can represent using certain meta-parameters. Distinct from a standard setting, meta-learning involves training on a diverse set of tasks. At test time, we evaluate the performance of the pre-trained model on new unseen tasks while allowing it to adapt using a small sample on the test task.

An increasing body of empirical research is dedicated to advancing meta-learning algorithms, among which model-agnostic meta-learning (MAML) [Finn et al., 2017] stands out as a prominent approach. MAML is designed to find a good meta-parameter w which facilitates the learning of task-specific parameters through a single step of gradient descent. In particular, given a set of m tasks denoted as $\{\mathcal{D}_j\}_{j=1}^m$, MAML estimates the meta-parameter as $\mathrm{w} = \mathrm{argmin}_{\mathrm{w}} \frac{1}{m}\sum_{j=1}^m L(\mathrm{u}_j, \mathcal{D}_j)$, where task-specific parameters are computed as $\mathrm{u}_j = \mathrm{w} - \eta\nabla L(\mathrm{w}, \mathcal{D}_j)$.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2661

However, a notable limitation of MAML is that it requires computing second-order derivatives, which is computationally demanding for deep neural networks in practical applications. This computational complexity also poses a challenge for a theoretical understanding of MAML, an aspect that remains largely under-explored. To mitigate this challenge, several MAML variants have been proposed, including first-order MAML [Finn et al., 2017], Reptile [Nichol et al., 2018], and iMAML [Rajeswaran et al., 2019]. Owing to its success, MAML has been used for robust adversarial meta-learning [Yin et al., 2018, Goldblum et al., 2020, Wang et al., 2021, Collins et al., 2020], differential private meta-learning [Li et al., 2019], and personalized federated learning [Chen et al., 2018, Fallah et al., 2020].

Another popular framework for meta-learning is based on a "proximal" update, wherein the task-specific parameter are iteratively learned by minimizing the empirical loss and an ℓ_2 regularizer [Denevi et al., 2018, Zhou et al., 2019, Denevi et al., 2019a, 2020, Jiang et al., 2021]. Given a task $\mathcal{D}$ and a meta-parameter w, the task-specific parameter u are defined as $\mathrm{u} = \mathrm{argmin}_{\mathrm{u}} L(\mathrm{u}; \mathcal{D}) + \frac{\lambda}{2} \|\mathrm{u} - \mathrm{w}\|^2$. This regularization strategy ensures that the task-specific parameter remains close to the meta-parameter. A similar strategy has been explored in other contexts. For example, Kuzborskij and Orabona [2017] study the problem of hypothesis transfer learning and show a fast rate on the generalization error of a task-specific parameter u returned by regularized empirical risk minimization conditioned on a good meta-parameter w. Yet, it remains unclear how to ensure finding such a good meta-parameter, *provably*. Relatedly, Denevi et al. [2019b] study stochastic gradient descent with biased regularization for linear model and incrementally update the bias (meta-parameter). Concurrently, Zhou et al. [2019] proposed the Meta-Prox algorithm as a generic stochastic meta-learning approach. Specifically, given a set of meta-training tasks $\mathcal{D}_1, \ldots, \mathcal{D}_m$, the meta-parameter w is estimated by solving $\min_{\mathrm{w}} \sum_{j=1}^{m} \min_{\mathrm{u}} L(\mathrm{u}, \mathcal{D}_j) + \frac{\lambda}{2} \|\mathrm{u} - \mathrm{w}\|^2$. Zhou et al. [2019] argue that Meta-Prox is a generalization of MAML since the gradient descent update in MAML can be viewed as taking the first-order Taylor expansion of the objective, [Zhou et al., 2019, Section 3.1].

In this work, we adopt the framework of Zhou et al. [2019] to study meta-learning from a theoretical perspective. Given m tasks drawn i.i.d. from some (unknown) task distribution μ, our goal is to find a good pre-trained model (the meta-parameter) which can be adapted to a new unseen task, drawn i.i.d. from μ, at test time, using gradient descent. Our key contributions are as follows.

1. We introduce a novel notion of stability for meta-learning algorithms, namely uniform meta-stability. For $\bar{\beta}$ uniformly meta-stable algorithm, we bound the generalization gap by $\mathcal{O}(\bar{\beta} \log(mn/\delta) + \sqrt{\log(1/\delta)/(mn)})$.
2. We consider two variants of task-specific learning – based on regularized empirical risk minimization (RERM) and gradient descent (GD) – within our meta-learning framework. We apply our stability-based analysis to these variants to learning problems with convex, smooth losses and weakly convex, non-smooth losses. Our results are summarized in Table 1.

Algorithm	Loss	Conditions	Uniform meta-stability $\bar{\beta}$
Algo. 1 with RERM	convex, G-Lipschitz	$\gamma \leq \frac{1}{\lambda}$	$\frac{G^2}{\lambda m} + \frac{G^2}{\lambda n}$
Algo. 1 with RERM	convex, H-smooth, M-bounded	$\gamma \leq \frac{1}{\lambda}, \lambda \geq H$	$\frac{HM}{\lambda(2n-1)} + \frac{HM}{\lambda(m+1)}$
Algo. 1 with GD	convex, G-Lipschitz, H-smooth	$\eta \leq \frac{2}{H+2\lambda}, \gamma \leq \frac{1}{\lambda T}$	$\frac{G^2}{\lambda m} + \frac{G^2}{\lambda n}$
Algo. 1 with GD	ρ-weakly convex, G-Lipschitz	$\eta \leq \frac{1}{\lambda}, \gamma \leq \frac{1}{\lambda T}, \lambda \geq 2\rho$	$G^2\sqrt{\frac{\eta}{\lambda}} + \frac{G^2}{\lambda m} + \frac{G^2}{\lambda n}$
Algo. 3 with GD	ρ-weakly convex, G-Lipschitz	$\eta \leq \frac{1}{\lambda}, \gamma \leq \frac{1}{\lambda T}, \lambda \geq 2\rho$	$G^2\sqrt{\frac{\eta}{\lambda}} + \frac{G^2}{\lambda m} + \frac{G^2}{\lambda n}$, w.h.p.

Table 1: Bounds on uniform meta-stability $\bar{\beta}$ for different families of learning problems. Here, η is the step-size for GD for task-specific learning, γ is the step-size for GD for meta-parameter learning, m is the number of tasks during training, n is the number of training data for the task at test time.

3. We extend our results to stochastic and adversarially robust variants of our meta-learning algorithm.

1.1 Related Work

Algorithmic Stability Analysis. In many machine learning problems, standard learning theoretic tools, such as uniform convergence, do not apply since the associated complexity measures are unbounded or undefined (e.g., nearest neighbor classification), or yield guarantees that are not meaningful. Stability-based analysis is an alternative approach for obtaining generalization bounds

in such settings, introduced by Bousquet and Elisseeff [2002] and further developed in a long line of influential works [Elisseeff et al., 2005, Mukherjee et al., 2006, Shalev-Shwartz et al., 2010, Liu et al., 2017]. More recently, there have been significant breakthroughs in this field, with the work of Feldman and Vondrak [2018, 2019], Bousquet et al. [2020], Klochkov and Zhivotovskiy [2021], thereby improving the high probability bounds for uniformly stable learning algorithms beyond those established by Bousquet and Elisseeff [2002]. These results are complemented by Hardt et al. [2016], who provide the generalization bounds via algorithmic stability analysis of stochastic gradient for stochastic convex optimization with smooth loss functions. Subsequent work by Bassily et al. [2020] improves upon these results by removing the smoothness assumption, while Zhou et al. [2022], Lei [2023] advance the state-of-the-art by relaxing the convexity assumption.

Theoretical Guarantees for Meta-Learning. There has been significant progress in understanding the theoretical aspects of meta-learning, both in terms of convergence guarantees [Fallah et al., 2019, Ji et al., 2020, Mishchenko et al., 2023] and the generalization guarantees. The first generalization analysis can be traced back to Baxter [2000], who assumed that all tasks are sampled i.i.d. from the same task distribution. Subsequent works have enriched the guarantees through various learning theoretic constructs, including VC theory [Ben-David and Schuller, 2003, Maurer, 2009, Maurer et al., 2016], information-theoretic tools [Chen et al., 2021, Jose and Simeone, 2021, Jose et al., 2021, Rezazadeh et al., 2021, Hellström and Durisi, 2022], PAC-Bayes framework [Pentina and Lampert, 2014, Amit and Meir, 2018, Rothfuss et al., 2021, Farid and Majumdar, 2021, Liu et al., 2021, Ding et al., 2021, Rezazadeh, 2022, Riou et al., 2023, Zakerinia et al., 2024], etc. Other works that do not rely on the task distribution assumption instead choose to get a handle on the bound by defining certain metrics to measure either the task similarity [Du et al., 2020, Tripuraneni et al., 2020, Guan and Lu, 2021] or the divergence between the new tasks and the training sample for the training tasks [Fallah et al., 2021]. Finally, several works focus on the online meta-learning setting, also referred to as the lifelong learning [Pentina and Lampert, 2014, Balcan et al., 2019, Denevi et al., 2019a,b, Meunier and Alquier, 2021].

A prominent line of work, starting with that of Maurer [2005], focuses on giving theoretical guarantees for meta-learning via algorithmic stability analysis. More recently, Chen et al. [2020] establish connections between single-task learning with support/query (episodic) meta-learning algorithms, providing generalization gap of $\mathcal{O}(1/\sqrt{m})$ (where m is the number of tasks) for smooth functions that is independent of the sample size n – this was shown to be nearly optimal in Guan et al. [2022]. Subsequently, Fallah et al. [2021] show a bound of $\mathcal{O}(1/mn)$ for strongly convex functions and by leveraging a new notion of stability. Al-Shedivat et al. [2021] extend the result of Maurer [2005] to practical meta-learning algorithms for Lipschitz and smooth losses. Farid and Majumdar [2021] derive a PAC-Bayes bound to address the qualitatively different challenges of generalization within the task compared to that at the meta-level. Other relevant work includes analyzing the stability of bilevel optimization [Bao et al., 2021] and federated learning [Sun et al., 2024] for smooth functions.

2 Problem Setup and Preliminaries

Notation. Throughout the paper, we denote scalars and vectors with lowercase italics and lowercase bold Roman letters, respectively; e.g., u, u. We work in a Euclidean space and use $\|\cdot\|$ and $\|\cdot\|_2$ to denote the ℓ_2 norm. We use $[n]$ to represent the set $\{1, 2, \ldots, n\}$, and define $\mathrm{U}[n]$ to be the uniform distribution over $[n]$. Let $\Pi_{\mathcal{W}}$ be the Euclidean projection onto $\mathcal{W}$. We adopt the standard O-notation and use $\lesssim$ and $\mathcal{O}$ interchangeably. We use $\tilde{\mathcal{O}}$ to hide poly-logarithmic dependence on the parameters.

Let $\mathcal{X}, \mathcal{Y}$ denote the input and output spaces, respectively. Consider a supervised learning setting where each data point is denoted by $\mathrm{z} = (\mathrm{x}, y)$ drawn from some unknown distribution $\mathcal{D}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We consider a hypothesis space $\mathcal{H}$ (maps from $\mathcal{X} \to \mathcal{Y}$) parameterized by $\mathrm{w} \in \mathcal{W}$, where $\mathcal{W} \subseteq \mathbb{R}^d$ is a closed set with radius D. Let $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^+$ denote the loss function. We say that a loss function ℓ is $\underline{M\text{-bounded}}$ if $\forall \mathrm{w} \in \mathcal{W}, \forall \mathrm{z} \in \mathcal{D}, \ell(\mathrm{w}, \mathrm{z}) \leq M$; ℓ is $\underline{\mu\text{-strongly convex}}$ if $\forall \mathrm{w}_1, \mathrm{w}_2 \in \mathcal{W}, \forall \mathrm{z} \in \mathcal{D}$, $\ell(\mathrm{w}_1, \mathrm{z}) \geq \ell(\mathrm{w}_2, \mathrm{z}) + \langle \nabla \ell(\mathrm{w}_2, \mathrm{z}), \mathrm{w}_1 - \mathrm{w}_2 \rangle + \frac{\mu}{2} \|\mathrm{w}_1 - \mathrm{w}_2\|_2^2$; if $\mu = 0$, we say $\ell(\cdot, \mathrm{z})$ is convex. We say ℓ is $\underline{G\text{-Lipschitz continuous}}$ if $\forall \mathrm{w}_1, \mathrm{w}_2 \in \mathcal{W}, \forall \mathrm{z} \in \mathcal{D}, \|\ell(\mathrm{w}_1, \mathrm{z}) - \ell(\mathrm{w}_2, \mathrm{z})\|_2 \leq G \|\mathrm{w}_1 - \mathrm{w}_2\|_2$; ℓ is $\underline{H\text{-smooth}}$ if $, \forall \mathrm{w}_1, \mathrm{w}_2 \in \mathcal{W}, \forall \mathrm{z} \in \mathcal{D}, \|\nabla \ell(\mathrm{w}_1, \mathrm{z}) - \nabla \ell(\mathrm{w}_2, \mathrm{z})\|_2 \leq H \|\mathrm{w}_1 - \mathrm{w}_2\|_2$.

In a standard (single-task) learning setup, given a model w, the expected loss on task $\mathcal{D}$ and the empirical loss on a training sample $\mathcal{S}$ drawn i.i.d. from $\mathcal{D}$, are defined, respectively, as follows.

https://doi.org/10.52202/079017-2661

$$L(\mathrm{w}, \mathcal{D}) = \mathbb{E}_{\mathrm{z}\sim\mathcal{D}}\left[\ell(\mathrm{w}, \mathrm{z})\right]; \quad L(\mathrm{w}, \mathcal{S}) = \frac{1}{n}\sum_{z\in\mathcal{S}} \ell(\mathrm{w}, \mathrm{z}).$$

In a meta-learning framework, we consider distributions $\{\mathcal{D}_j\}_{j=1}^m$ associated with m different tasks that are drawn from some (unknown) task distribution μ. For each task j, we assume that the learner has access to n training examples drawn i.i.d. from $\mathcal{D}_j$, i.e., $\mathcal{S}_j = \left\{\mathrm{z}_j^i\right\}_{i=1}^n \sim \mathcal{D}_j^n$. We denote the cumulative training data as $\mathbf{S} = \{\mathcal{S}_j\}_{j=1}^m$, and refer to it as the meta-sample.

A meta-learning algorithm $\mathcal{A}$ takes the meta-sample $\mathbf{S}$ as input and outputs an algorithm $\mathcal{A}(\mathbf{S}) : (\mathcal{X}\times\mathcal{Y})^n \to \mathcal{H}$. The performance of the meta-algorithm $\mathcal{A}$ is measured in terms of its ability to generalize w.r.t. loss $\ell(\cdot)$ to a new (previously unseen) task from the task distribution μ; we also refer to it as the *transfer risk*:

$$L(\mathcal{A}(\mathbf{S}), \mu) = \mathbb{E}_{\mathcal{D}\sim\mu}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^n} L(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathcal{D}).$$

The goal of meta-learning is to learn a useful prior over tasks to help with rapid adaptation to new tasks. Formally, we pose the problem as learning a meta-model, parameterized by what we will refer to as meta-parameter w, that performs well on a variety of tasks. The hope is that the meta-parameter w can be adapted easily to a new task $\mathcal{D}\sim\mu$; in particular, that a task-specific model u can be quickly learned from a task-specific training set $\mathcal{S}\sim\mathcal{D}^n$ of size n using the following proximal update:

$$\mathrm{u} = \mathrm{argmin}_{\mathrm{u}\in\mathcal{W}}\, L(\mathrm{u}, \mathcal{S}) + \tfrac{\lambda}{2}\left\|\mathrm{u}-\mathrm{w}\right\|^2,$$

where $\lambda > 0$ is a regularization parameter.

Algorithm 1 Prox Meta-Learning Algorithm $\mathcal{A}$

Input: Meta-sample $\mathbf{S}=\{\mathcal{S}_j\}_{j=1}^m$, epochs T, K, step sizes γ, η, regularization parameter λ

$\mathrm{w}_1 = 0$.
for $t = 1, 2, \ldots, T$ **do**
  for $j = 1, \ldots, m$ **do**
    $\mathrm{u}(\mathrm{w}_t, \mathcal{S}_j) = \mathcal{A}_{\text{task}}(\mathrm{w}_t, \mathcal{S}_j, K, \eta, \lambda)$ % Using Algorithm 2
  end for
  Calculate the gradient, $\forall j\in[m]$, $\nabla F_{\mathcal{S}_j}(\mathrm{u}(\mathrm{w}_t, \mathcal{S}_j), \mathrm{w}_t) = -\lambda(\mathrm{u}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{w}_t)$.
  Update $\mathrm{w}_{t+1} = \mathrm{w}_t - \frac{\gamma}{m}\sum_{j=1}^m \nabla F_{\mathcal{S}_j}(\mathrm{u}(\mathrm{w}_t, \mathcal{S}_j), \mathrm{w}_t)$
  $\mathrm{w}_{t+1} = \Pi_{\mathcal{W}}(\mathrm{w}_{t+1})$
end for
return $@\mathcal{A}_{\text{task}}(\mathrm{w}_{T+1}, \cdot, K, \eta, \lambda)$

Algorithm 2 Task-specific Algorithm $\mathcal{A}_{\text{task}}$

Input: Pretrained model w, training data $\mathcal{S}$, #epochs K, step size η, reg. parameter λ

Option 1 (RERM):
$\mathrm{u}(\mathrm{w}, \mathcal{S}) = \mathrm{argmin}_{\mathrm{u}\in\mathcal{W}} L(\mathrm{u}, \mathcal{S}) + \frac{\lambda}{2}\left\|\mathrm{u}-\mathrm{w}\right\|^2$.
Option 2 (GD): Set $\mathrm{u}^{(1)}(\mathrm{w}, \mathcal{S}) = \mathrm{w}$
for $t = 1, 2, \ldots, K-1$ **do**
  $\mathrm{u}^{(k+1)}(\mathrm{w}, \mathcal{S}) = \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \eta(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) + \lambda(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{w}))$
  $\mathrm{u}^{(k+1)}(\mathrm{w}, \mathcal{S}) = \Pi_{\mathcal{W}}(\mathrm{u}^{(k+1)}(\mathrm{w}, \mathcal{S}))$
end for
return Option 1 (RERM): $\mathrm{u}(\mathrm{w}, \mathcal{S})$
Option 2 (GD): $\frac{1}{K}\sum_{k=1}^K \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S})$

The meta-parameter w itself is learned on the given meta-sample $\mathbf{S}$ by minimizing a regularized empirical loss averaged over tasks, where the regularization term penalizes the task-specific models in proportion to the ℓ_2 distance from the meta-parameter [Zhou et al., 2019]:

$$\widehat{\mathrm{w}} = \mathrm{argmin}_{\mathrm{w}\in\mathcal{W}} \frac{1}{m}\sum_{j=1}^m \min_{\mathrm{u}\in\mathcal{W}} F_{\mathcal{S}_j}(\mathrm{u}, \mathrm{w}) := \mathrm{argmin}_{\mathrm{w}\in\mathcal{W}} \frac{1}{m}\sum_{j=1}^m \min_{\mathrm{u}\in\mathcal{W}} \left[L(\mathrm{u}, \mathcal{S}_j) + \frac{\lambda}{2}\left\|\mathrm{u}-\mathrm{w}\right\|^2\right]. \quad (1)$$

The formulation above involves a bi-level optimization problem. The upper-level optimization involves finding the meta-parameter w which requires solving the lower-level optimization problem of finding task-specific model parameters u. We consider both Gradient Descent (GD) as well Regularized Empirical Risk Minimization (RERM) for task-specific learning (see Algorithm 2 for more details); for meta-learning we employ a gradient descent method (see Algorithm 1).

We would like to bound the transfer risk in terms of the *empirical multi-task risk*:

$$L(\mathcal{A}(\mathbf{S}), \mathbf{S}) = \frac{1}{m}\sum_{j=1}^m L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j), \mathcal{S}_j).$$

To do so, we rely on the stability of the meta-learning algorithm.

Stability of Meta-Learning Algorithm. Given a meta-sample $\mathbf{S} = \{\mathcal{S}_j\}_{j=1}^m$, define $\mathbf{S}^{(j)}$ to be the meta-sample obtained by replacing the training samples $\mathcal{S}_j$ for the j-th task, in $\mathbf{S}$, by another i.i.d. sample $\mathcal{S}_j' \sim \mathcal{D}_j^n$. We refer to $\mathbf{S}, \mathbf{S}^{(j)}$ as neighboring meta-samples. For a task-specific training sample $\mathcal{S} = \left\{\mathrm{z}^i\right\}_{i=1}^n$, let $\mathcal{S}^{(i)}$ denote the training data obtained by replacing the i-th example $\mathrm{z}^i \in \mathcal{S}$ by another example $\mathrm{z}' \sim \mathcal{D}$ drawn independently; we refer to $\mathcal{S}, \mathcal{S}^{(i)}$ as neighboring samples.

Theorem 2.1 (Maurer [2005])**.** Suppose the meta-algorithm $\mathcal{A}$ satisfies:

1. (Uniform Stability of Single-Task Learning) For any meta-sample $\mathbf{S}$ and any $\mathcal{S}, \mathcal{S}^{(i)}$,
$$\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathrm{z}) - \ell(\mathcal{A}(\mathbf{S})(\mathcal{S}^{(i)}), \mathrm{z})\right| \leq \beta.$$
2. (Uniform Stability of Meta-Learning) For any $\mathbf{S}, \mathbf{S}^{(j)}$ and any given training set $\mathcal{S} \sim \mathcal{D}$,
$$\left|L(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathcal{S}) - L(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}), \mathcal{S})\right| \leq \beta'.$$

Then, for M-bounded loss ℓ, with probability at least $1-\delta$, we have that

$$L(\mathcal{A}(\mathbf{S}), \mu) \lesssim L(\mathcal{A}(\mathbf{S}), \mathbf{S}) + (m\beta' + M)\sqrt{\log(1/\delta)/m} + \beta.$$

Theorem 2.1 follows using a simple extension of arguments in Bousquet and Elisseeff [2002]. By utilizing sharper bounds tailored for uniformly stable algorithms [Bousquet et al., 2020], a tighter bound can be achieved, as demonstrated in Theorem 2.2 below. A similar result was shown in Guan et al. [2022] for episodic training algorithms (except there is no β).

Theorem 2.2. Suppose the meta-algorithm $\mathcal{A}$ satisfies the same conditions as shown in Theorem 2.1. Then for M-bounded loss ℓ, with probability at least $1-\delta$, we have that

$$L(\mathcal{A}(\mathbf{S}), \mu) \lesssim L(\mathcal{A}(\mathbf{S}), \mathbf{S}) + \beta' \log(m) \log(1/\delta) + M\sqrt{\log(1/\delta)/m} + \beta.$$

3 Uniform Meta-Stability

Motivated by prior work (i.e., Theorem 2.1 and the definitions therein), we introduce a new notion of stability which measures the sensitivity of the learning algorithm as we replace both a task in the meta-sample as well as a single training example available for the task at test time.

Definition (Uniform Meta-Stability)**.** We say that a meta-learning algorithm $\mathcal{A}$ is $\bar{\beta}$-uniformly meta-stable if for any neighbouring meta-samples $\mathbf{S}, \mathbf{S}^{(j)}$, and neighboring samples $\mathcal{S}, \mathcal{S}^{(i)}$, for any task $\mathcal{D} \sim \mu$ and any $\mathrm{z} \sim \mathcal{D}$, we have that
$$\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathrm{z}) - \ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}), \mathrm{z})\right| \leq \bar{\beta}.$$

The definition above is rather natural. Intuitively, for a meta-learning algorithm to transfer well, we require that the learning algorithms, i.e., $\mathcal{A}(\mathbf{S})$ and $\mathcal{A}(\mathbf{S}')$, returned on two neighboring meta-samples, when trained on two neighboring samples return models that predict similarly. Our first result bounds the generalization gap in terms of the uniform meta-stability parameter.

Theorem 3.1. Consider a meta-learning problem for some M-bounded loss function ℓ and task distribution μ. Let $\mathbf{S}$ be a meta-sample consisting of training samples on m tasks each of size n, and let $\mathcal{S} \sim \mathcal{D}$ be a sample of size n on a previously unseen task $\mathcal{D} \sim \mu$. Then, for any β-uniformly meta-stable learning algorithm $\mathcal{A}$, we have that with probability $1-\delta$,

$$L(\mathcal{A}(\mathbf{S}), \mu) \lesssim L(\mathcal{A}(\mathbf{S}), \mathbf{S}) + \bar{\beta} \log(mn) \log(1/\delta) + M\sqrt{\log(1/\delta)/(mn)}.$$

The result above is a direct analogue of Theorem 2.1 with stability parameters β, β' both subsumed into a single meta-stability parameter. We do obtain a faster rate of convergence – as we instantiate concrete algorithms and specialize our results to specific problems in Section 4.1, we will see a notable improvement in rates from $1/\sqrt{m}$ to $1/m$, for $n > m$.

We conclude the section by presenting an alternate notion of algorithmic meta-stability and a basic result that directly bounds the generalization gap for the meta-learning problem.

Definition (On-Average Meta-Stability)**.** Let μ be an (unknown) underlying task distribution. We say that a meta-learning algorithm $\mathcal{A}$ is $\bar{\beta}$-on-average-replace-one-meta-stable if

$$\mathbb{E}_{\mathbf{S}\sim\{\mathcal{D}_j^n\}_{j=1}^m, (\mathcal{S}_j', \mathrm{z}_j')\sim\mathcal{D}_j^{n+1}, \{\mathcal{D}_j\}_{j=1}^m\sim\mu^m, j\sim\mathrm{U}[m], i\sim\mathrm{U}[n]} \left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}_j), \mathrm{z}_j^i) - \ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}), \mathrm{z}_j^i)\right| \leq \bar{\beta}.$$

Theorem 3.2. Let μ be an underlying task distribution. Given a meta-sample $\mathbf{S}$, test task $\mathcal{D} \sim \mu$, and $\mathcal{S} \sim \mathcal{D}^n$, for any $\bar{\beta}$ -on-average-replace-one-meta-stable meta-learning algorithm $\mathcal{A}$, we have that

$$\mathbb{E}_{\mathbf{S}\sim\{\mathcal{D}_j^n\}_{j=1}^m, \{\mathcal{D}_j\}_{j=1}^m\sim\mu^m}[L(\mathcal{A}(\mathbf{S}), \mu) - L(\mathcal{A}(\mathbf{S}), \mathbf{S})] \leq \bar{\beta}.$$

4 Bounding Transfer Risk

In this section, we consider a concrete meta-learning algorithm given in Algorithm 1.

https://doi.org/10.52202/079017-2661

4.1 Convex and Smooth Losses

We begin with meta-learning problems with convex, Lipschitz (and potentially smooth) losses.

Lemma 4.1. Assume that the loss function ℓ is convex and G-Lipschitz loss. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then, the following holds for Algorithm 1 with RERM for task-specific learning (i.e., Option 1 for Algorithm 2) $\forall T \geq 1$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]} \left\| \mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}) \right\| \leq \frac{G}{\lambda m} + \frac{2G}{\lambda n}.$$

Further, if ℓ is convex, M-bounded and H-smooth, then setting $\lambda \geq H$, $\gamma \leq \frac{1}{\lambda}$, we have $\forall T \geq 1$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]} \left\| \mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}) \right\| \leq \frac{2\sqrt{2HM}}{2\lambda n - H} + \frac{n}{2\lambda n - H}\frac{4\sqrt{2HM}}{(m+1)}.$$

We can now use the result above with Theorem 3.1 to get the following bound on the transfer risk.

Theorem 4.2. The following holds for Algorithm 1 with step-size $\gamma \leq \frac{1}{\lambda}$ on a given meta-sample $\mathbf{S}$, and RERM for task-specific learning (i.e., Option 1 for Algorithm 2), for all $T \geq 1$:

1. For convex, M-bounded, and G-Lipschitz loss functions, with probability at least $1-\delta$

$$L(\mathcal{A}(\mathbf{S}),\mu) \lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S}) + \left(\frac{G^2}{\lambda n} + \frac{G^2}{\lambda m}\right)\log(mn)\log(1/\delta) + \frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

2. For convex, M-bounded, and H-smooth loss functions ($H \leq \lambda$), with probability at least $1-\delta$

$$L(\mathcal{A}(\mathbf{S}),\mu) \lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S}) + \left(\frac{HM}{(2n-1)\lambda} + \frac{HM}{(m+1)\lambda}\right)\log(mn)\log(1/\delta) + \frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

Next, we give analogous results for GD for task-specific learning (i.e., Option 2 for Algorithm 2), albeit for smooth loss functions. Lemma 4.3 bounds the output sensitivity of the meta-learning algorithm. We use it with Theorem 3.1 to give the generalization guarantee in Theorem 4.4.

Lemma 4.3. Assume that the loss function is convex, G-Lipschitz and H-smooth. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then the following holds for Algorithm 1 with GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta \leq \frac{2}{H+2\lambda}$, for all $T \geq 1$ as long as we set $\gamma \leq \frac{1}{\lambda T}$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]} \left\| \mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}) \right\| \leq \frac{4eG}{\lambda m} + \frac{2G}{\lambda n}.$$

Theorem 4.4. Assume that the loss function is convex, M-bounded, G-Lipschitz and H-smooth. Suppose we run Algorithm 1 for T iterations with $\gamma \leq \frac{1}{\lambda T}$ on a given meta-sample $\mathbf{S}$, and GD for task-specific learning (Option 2, Algorithm 2) with $\eta \leq \frac{2}{H+2\lambda}$. Then, with probability at least $1-\delta$,

$$L(\mathcal{A}(\mathbf{S}),\mu) \lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S}) + \left(\frac{G^2}{\lambda m} + \frac{G^2}{\lambda n}\right)\log(mn)\log(1/\delta) + \frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

The results above show that meta-stable learning algorithms do not overfit. The bound on the generalization gap of $\tilde{\mathcal{O}}(\frac{1}{m} + \frac{1}{n} + \frac{1}{\sqrt{mn}})$ is tighter than what we would obtain using prior work. Indeed, we show that Theorem 2.2 yields a rate of $\tilde{\mathcal{O}}(\frac{1}{m} + \frac{1}{n} + \frac{1}{\sqrt{m}})$ (see Theorems C.2 and C.3 in Appendix), which is worse for all $m \leq n^2$. Notably, the bounds on the generalization gap are independent of the number of iterations of the meta learning Algorithm 1 and the number of iterations of GD for Algorithm 2. This holds since the objective we are minimizing is strongly convex (given the strongly convex regularizer), which ensures that the output sensitivity (in Lemmas 4.3 and 4.1 are independent of T and K. In itself, this should not be surprising since we only bound the generalization error in terms of the empirical error – the latter may not be small unless the algorithms have converged. To get a better handle on the generalization error we focus on excess (transfer) risk bounds in Section 4.3. But first we give a similar development for another important problem class.

4.2 Weakly Convex and Non-smooth Losses

Here, we focus on a more practical setting of learning problems with loss functions that are weakly convex and non-smooth. The notion of weak convexity is often used in non-convex optimization literature in a variety of problems including robust phase retrieval [Davis et al., 2020] and dictionary learning [Davis and Drusvyatskiy, 2019]; see Drusvyatskiy [2017] for an extended discussion.

Definition. A function $f(\mathrm{w})$ is ρ-*weakly convex* w.r.t. $\|\cdot\|$ if $f(\mathrm{w}) + \frac{\rho}{2}\|\mathrm{w}\|^2$ is convex in w.

The class of weakly convex functions is contained within the larger class of non-smooth functions and semi-smooth functions [Mifflin, 1977]. It includes convex functions and smooth functions with Lipschitz continuous gradient as special cases; $\rho < 0$ implies that the function is strongly convex. An important example from a practical perspective is that of training over-parameterized two-layer neural networks with smooth activation functions using a smooth loss [Richards and Rabbat, 2021]. We first bound the sensitivity of Algorithm 1 for weakly convex and non-smooth losses.

Lemma 4.5. Assume that the loss function is ρ-weakly convex and G-Lipschitz. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then the following holds for Algorithm 1 with $\lambda \geq 2\rho$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta \leq \frac{1}{\lambda}$, for all $T \geq 1$ as long as we set $\gamma \leq \frac{1}{\lambda T}$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]} \left\| \mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}) \right\| \leq (8eG + 2G)\sqrt{\frac{\eta}{\lambda}} + \frac{8eG}{\lambda m} + \frac{8G}{\lambda n}.$$

Using the result above in conjunction with Thm 3.1 gives the following bound on the transfer risk.

Theorem 4.6. Assume that the loss function is ρ-weakly convex, M-bounded, and G-Lipschitz. Suppose we run Algorithm 1 for T iterations with $\gamma \leq \frac{1}{\lambda T}$, $\lambda \geq 2\rho$ on a meta-sample $\mathbf{S}$, and GD for task-specific learning (Option 2, Algorithm 2) with $\eta \leq \frac{1}{\lambda}$, Then, with probability at least $1-\delta$,

$$L(\mathcal{A}(\mathbf{S}), \mu) \lesssim L(\mathcal{A}(\mathbf{S}), \mathbf{S}) + \left(G^2\sqrt{\frac{\eta}{\lambda}} + \frac{G^2}{\lambda m} + \frac{G^2}{\lambda n} \right) \log(mn) \log(1/\delta) + \frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

Proof of Theorem 4.6 follows from Lemma 4.5 and Theorem 3.1. A few remarks are in order.

For learning rate $\gamma \leq \frac{1}{\lambda T}$, Theorem 4.6 gives a rate of $\tilde{\mathcal{O}}(\sqrt{\eta} + \frac{1}{m} + \frac{1}{n} + \frac{1}{\sqrt{mn}})$ on the generalization gap. This naturally suggests setting $\eta = \frac{1}{\lambda K}$, where $K \geq \min\{m, n\}$ is the number of iterations of GD in task-specific learning. Then, similar to the discussion in Section 4.1, Theorem 4.6 gives a tighter bound, when $n > m$, than those derived using prior work (Theorem 2.2); we refer the reader to Theorem D.4 in the appendix for further details.

Our proof technique shares similarities with Bassily et al. [2020]. However, our result is not a straightforward application of theirs as we deal with a bi-level optimization problem and focus on weakly convex functions. It is worth noting that our results for weakly convex non-smooth losses require regularization parameter $\lambda \geq 2\rho$, which can be chosen in practice using cross-validation.

The work most related to ours is that of Guan et al. [2022]. However, our results are fundamentally different from theirs in several aspects. Firstly, the algorithms we study are different. Guan et al. [2022] focus on support/query (S/Q) training strategies (aka episodic training) where each task $\mathcal{S}_j$ is split into two non-overlapping parts – the support set $\mathcal{S}_j^{tr}$ for training the task-specific parameter and the query set $\mathcal{S}_j^{ts}$ for measuring the algorithm's performance [Vinyals et al., 2016]. The meta-parameter is learned by minimizing the loss computed over the query set. Such S/Q training strategy is popular for modern gradient-based meta-learning algorithm such as MAML for few-shot learning [Finn et al., 2017], where the optimization objective can be written as $\min_{\mathrm{w}} \frac{1}{m}\sum_{j=1}^{m} L(\mathrm{w} - \nabla L(\mathrm{w}, \mathcal{S}_j^{tr}), \mathcal{S}_j^{ts})$. One notable limitation is that Guan et al. [2022] assume that the loss function on the task level, e.g., $R(\mathrm{w}, \mathcal{S}_j) = L(\mathrm{w} - \nabla L(\mathrm{w}, \mathcal{S}_j^{tr}), \mathcal{S}_j^{ts})$, is convex or (Hölder) smooth. Such an assumption is highly impractical, as demonstrated by [Mishchenko et al., 2023, Theorem 1, Theorem 2], which provides several counterexamples where L is convex and smooth but R is neither convex nor smooth. In contrast, we directly deal with L being weakly convex and nonsmooth. Our approach requires a more involved proof that deals with stability of bi-level optimization. This is in stark contrast with Guan et al. [2022] who directly reduce the meta-learning problem to a single-task learning problem without considering the bi-level structure of the problem.

The work of Fallah et al. [2021] proposed a notion of stability similar to ours. The difference is that they consider S/Q training and define the stability by changing a mini-batch of samples in $\mathcal{S}_j^{tr}$ as

https://doi.org/10.52202/079017-2661

well as a single sample in $\mathcal{S}_j^{ts}$. Moreover, their focus is primarily on strongly convex losses. They discuss generalization to training tasks and unseen tasks separately, as they do not assume all tasks are sampled from the same task distribution. Another related work of Guan and Lu [2021] present a generalization bound of $\mathcal{O}(\sqrt{C/mn})$ under a task relatedness assumption, where C captures the logarithm of the covering number of hypothesis class that possibly depends on the dimension d. More recently, Riou et al. [2023] provide generalization bounds with a fast rate of $\mathcal{O}(\frac{1}{m}+\frac{1}{n})$, albeit under an additional extended Bernstein's condition.

4.3 Excess Transfer Risk

In the previous sections, we focused on establishing that meta-stable rules do not overfit to the meta-sample. In this Section, we focus on the question of whether meta-learning Algorithm 1 can achieve a small generalization error, i.e., are they guaranteed to transfer well on unseen tasks? We show that by focusing on the computational aspects, i.e., by bounding the optimization error in terms of the number of iterations. Furthermore, we give bounds on excess risk, wherein the benchmark is the performance of the best possible in-class predictor.

Let $\mathrm{u}_* = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} L(\mathrm{u},\mathcal{D})$, $\mathrm{u}_j^* = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} L(\mathrm{u},\mathcal{S}_j)$, $\forall j\in[m]$ be the optimal task-specific hypotheses for the unseen task and the given training tasks, respectively. Given a meta-algorithm $\mathcal{A}$, the excess transfer risk can be decomposed as follows:

$$\underbrace{L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{D})-L(\mathrm{u}_*,\mathcal{D})}_{\text{Excess Transfer Risk }\ \mathcal{E}_{\text{risk}}(\mathcal{A})} = \underbrace{L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{D})-\frac{1}{m}\sum_{j=1}^{m}L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathcal{S}_j)}_{\text{Generalization Gap }\mathcal{E}_{\text{gen}}(\mathcal{A})} + \underbrace{\frac{1}{m}\sum_{j=1}^{m}\left[L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathcal{S}_j)-L(\mathrm{u}_j^*,\mathcal{S}_j)\right]}_{\text{Optimization and Approximation Error }\mathcal{E}_{\text{opt+app}}(\mathcal{A})}$$
$$+\underbrace{\frac{1}{m}\sum_{j=1}^{m}\left[L(\mathrm{u}_j^*,\mathcal{S}_j)-L(\mathrm{u}_*,\mathcal{S}_j)\right]}_{\leq 0} + \underbrace{\frac{1}{m}\sum_{j=1}^{m}\left[L(\mathrm{u}_*,\mathcal{S}_j)-L(\mathrm{u}_*,\mathcal{D})\right]}_{\mathbb{E}_{\forall j\in[m],\mathcal{S}_j\sim\mathcal{D}_j^n,\mathcal{D}_j\sim\mu,\mathcal{D}\sim\mu}=0}.$$

To control excess risk, we need to bound $\mathcal{E}_{\text{gen}}(\mathcal{A})$ and $\mathcal{E}_{\text{opt+app}}(\mathcal{A})$ simultaneously. The bounds on the first term are presented in the previous section. Here, we focus on analyzing the second term.

Theorem 4.7. Assume that the loss ℓ is convex and G-Lipschitz. Define $\mathrm{u}_j^* = \operatorname{argmin}_{\mathrm{u}} L(\mathrm{u},\mathcal{S}_j), \forall j\in[m]$. Suppose we run Algorithm 1 for T iterations with step-size $\gamma=\frac{1}{\lambda T}$, and using GD for task-specific learning (i.e., Option 2 for Algorithm 2), to find an algorithm $\mathcal{A}(\mathbf{S})=\mathcal{A}_{\text{task}}(\mathrm{w}_{T+1},\cdot)$ which is then run on $\mathcal{S}_j$ for K iterations with step-size $\eta\leq\frac{1}{2\lambda}$. Then, we have that

$$L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathcal{S}_j)-\inf_{\mathrm{u}}L(\mathrm{u},\mathcal{S}_j)\lesssim\frac{D^2}{\eta K}+G^2\eta+GD\eta\lambda+\lambda\left\|\mathrm{w}_{T+1}-\widehat{\mathrm{w}}\right\|^2+\lambda\sigma^2$$

where $\widehat{\mathrm{w}}$ is defined in Equation (1). Here $\sigma^2:=\frac{1}{m}\sum_{j=1}^{m}\left\|\widehat{\mathrm{w}}-\mathrm{u}_j^*\right\|^2$ is the approximation error, and $\|\mathrm{w}_{T+1}-\widehat{\mathrm{w}}\|^2\lesssim\frac{1}{T}(D^2+\frac{D^2}{\lambda\eta K}+\frac{\eta(G+2\lambda D)^2}{\lambda})$ is the optimization error.

Finally, to bound the excess transfer risk for convex and non-smooth losses, we use Theorem 4.6 with Theorem 4.7 to get that in expectation over the sampling of data (meta-sample $\mathbf{S}$ and sample $\mathcal{S}$)

$$\mathbb{E}[\mathcal{E}_{\text{risk}}(\mathcal{A})]\leq\mathbb{E}[\mathcal{E}_{\text{gen}}(\mathcal{A})]+\mathbb{E}[\mathcal{E}_{\text{opt+app}}(\mathcal{A})]\lesssim G^2\sqrt{\frac{\eta}{\lambda}}+\frac{G^2}{\lambda m}+\frac{G^2}{\lambda n}+\frac{D^2}{\eta K}+G^2\eta+GD\eta\lambda+\frac{\lambda D^2}{T}+\eta(G+2\lambda D)^2+\lambda\sigma^2.$$

By properly choosing step size $\eta=\mathcal{O}\left(\frac{1}{\lambda K^{2/3}}\right)$, we obtain that the expected excess transfer risk decays at a rate of $\mathcal{O}(\frac{1}{\lambda K^{1/3}}+\frac{1}{\lambda m}+\frac{1}{\lambda n}+\frac{\lambda}{T}+\lambda\sigma^2)$. Similarly, for convex, Lipschitz and smooth losses, applying Theorem 4.4 with Theorem 4.7 and selecting $\eta=\mathcal{O}(\frac{1}{\lambda\sqrt{K}})$ results in an expected excess transfer risk of $\mathcal{O}(\frac{1}{\lambda\sqrt{K}}+\frac{1}{\lambda m}+\frac{1}{\lambda n}+\frac{\lambda}{T}+\lambda\sigma^2)$. Therefore, as K,T,m,n tend to infinity, the excess risk converges to σ^2. As σ represents the average distance between the optimal task-specific parameters u_j's and the optimal estimated meta-parameter $\widehat{\mathrm{w}}$, the excess risk is small when σ is small. It is also typical to set the regularization parameter λ inversely proportional to the sample size n (e.g., $\lambda=\mathcal{O}(1/\sqrt{n})$).

Denevi et al. [2019a] study the same algorithm as ours except in the online setting. However, the function classes they consider are limited to compositions of linear hypothesis classes with convex and closed losses. In contrast, our work considers a broader range of functions, encompassing not only convex, Lipschitz, and smooth functions but also weakly-convex and non-smooth functions. The

bound on expected excess risk shown in Denevi et al. [2019a] takes the form $\mathcal{O}(\frac{\text{Var}_m}{\sqrt{n}} + \frac{1}{\sqrt{m}})$, where Var_m captures the relatedness among the tasks sampled from the task environment. Unfortunately, this bound relies on a specific choice of $\lambda = \mathcal{O}\left(\frac{1}{\text{Var}_m}\sqrt{\frac{\log(n)}{n}}\right)$, which depends on Var_m – a quantity that is often not known a priori in practice. To compare with our work, set $K = n$, $T = m$, $\eta = \mathcal{O}(1/\sqrt{n})$, and $\lambda = \mathcal{O}(1/\sqrt{n})$.Then, applying Theorem 4.4 with Theorem 4.7, we obtain that $\mathbb{E}[\mathcal{E}_{\text{risk}}(\mathcal{A})] \lesssim \frac{\sqrt{n}}{m} + \frac{\max(1,\sigma^2)}{\sqrt{n}}$. Considering both Var_m and σ as constants, the bound on expected excess risk based on our analysis is tighter than that of Denevi et al. [2019a] when $n \lesssim m$, a common setting studied in meta-learning framework.

We also conduct a simple experiment to empirically verify the tightness of our generalization bounds, which we defer to Appendix A due to space limitations.

5 Implications of the Generalization Bounds

Next, we present stochastic and adversarially robust variants of the meta-learning Algorithm 1.

5.1 Proximal Meta-Learning with Stochastic Optimization

We adapt Algorithm 1 to utilize sampling-with-replacement where at each iteration we process the training set of a single task; see Algorithm 3 for more details. We show that with high probability the sensitivity of this stochastic meta-learning algorithm is bounded.

Lemma 5.1. Assume that the loss function is ρ-weakly convex and G-Lipschitz. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then, with probability at least $1-\exp\left(-T^2e^2/m^2\right)$, the following holds for Algorithm 3 with $\lambda \geq 2\rho$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta \leq \frac{1}{\lambda}$, for all $T \geq 1$ as long as we set $\gamma \leq \frac{1}{\lambda T}$,

$$\sup_{\mathbf{S},\mathcal{S},i\in[n],j\in[m]} \left\|\mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})\right\| \leq (8eG + 2G)\sqrt{\frac{\eta}{\lambda}} + \frac{8eG}{\lambda m} + \frac{8G}{\lambda n}.$$

5.2 Robust Adversarial Proximal Meta-Learning

We consider inference-time adversarial attacks with a general threat model $\mathcal{B} : \mathcal{X} \to 2^{\mathcal{X}}$. Specifically, given an input example $\mathrm{x} \in \mathcal{X}$, $\mathcal{B}(\mathrm{x}) \subseteq \mathbb{R}^d$ represents the set of all possible perturbations of x that an adversary can choose from. This includes the typical examples such as the L_p threat models that are often considered in practice, or a discrete set of designed transformations.

Algorithm 3 Stochastic Prox Meta-Learning

Input: Meta-sample $\mathbf{S} = \{\mathcal{S}_j\}_{j=1}^m$, epochs T, K, step size γ, η, regularization parameter λ.

$\mathrm{w}_1 = 0$.
for $t = 1, 2, \ldots, T$ **do**
  Sample $j_t \sim \mathrm{U}[m]$.
  $\mathrm{u}(\mathrm{w}_t, \mathcal{S}_{j_t}) = \mathcal{A}_{\text{task}}(\mathrm{w}_t, \mathcal{S}_{j_t}, K, \eta, \lambda)$.
  % Using Algorithm 2
  Calculate the gradient $\nabla F_{\mathcal{S}_{j_t}}(\mathrm{u}(\mathrm{w}_t, \mathcal{S}_{j_t}), \mathrm{w}_t) = -\lambda(\mathrm{u}(\mathrm{w}_t, \mathcal{S}_{j_t}) - \mathrm{w}_t)$.
  Update $\mathrm{w}_{t+1} = \mathrm{w}_t - \gamma \nabla F_{\mathcal{S}_{j_t}}(\mathrm{u}(\mathrm{w}_t, \mathcal{S}_{j_t}), \mathrm{w}_t)$.
  $\mathrm{w}_{t+1} = \Pi_{\mathcal{W}}(\mathrm{w}_{t+1})$.
end for
return $@\mathcal{A}_{\text{task}}(\mathrm{w}_{T+1}, \cdot, K, \eta, \lambda)$

Given a model parameter w, let $\tilde{\ell}(\mathrm{w}, \mathrm{z}) = \max_{\tilde{\mathrm{z}}\in\mathcal{B}(\mathrm{z})} \ell(\mathrm{w}, \tilde{\mathrm{z}})$ denote the adversarial loss. We adapt the standard meta-learning framework simply by considering the robust variant, $\tilde{\ell}$, of the standard loss ℓ. We denote the robust transfer risk and empirical robust multi-task risk as $L_{\text{rob}}(\mathcal{A}(\mathbf{S}), \mu)$ and $L_{\text{rob}}(\mathcal{A}(\mathbf{S}), \mathbf{S})$. Now, given meta-sample $\mathbf{S}$, the goal is to learn a robust prior (e.g., a pre-trained model) for rapid adaptation to and robust generalization on new tasks. We adopt the framework presented in Section 2 except we use robust loss for task-specific training; indeed, using GD (Option 2) on robust loss in Algorithm 2 yields adversarial training. We use Algorithm 1 for meta-learning. We now relate a loss function with its adversarially robust counterpart.

Proposition 5.2. Given a loss function $\ell(\cdot, \mathrm{z})$ and its adversarial counterpart $\tilde{\ell}(\cdot, \mathrm{z})$, the following holds: (1) If ℓ is G-Lipschitz (in its first argument), then $\tilde{\ell}$ is G-Lipschitz. (2) $\tilde{\ell}$ is **not** H-smooth even if ℓ is H-smooth. (3) If ℓ is H-smooth in w, then $\tilde{\ell}$ is H-weakly convex in w.

Using the result above with Theorem 3.1 yields the following bound on robust (transfer) risk.

https://doi.org/10.52202/079017-2661

Corollary 5.3. Assume that the loss ℓ is M-bounded and H-smooth. Suppose we run Algorithm 1 for T iterations with $\gamma \leq \frac{1}{\lambda T}, \eta \leq \frac{1}{\lambda}, \lambda > 2H$, and wherein task-specific learning Algorithm 2 (GD) is invoked with robust loss $\tilde{\ell}$, we have that with probability at least $1-\delta$,

$$L_{\text{rob}}(\mathcal{A}(\mathbf{S}), \mu) \lesssim L_{\text{rob}}(\mathcal{A}(\mathbf{S}), \mathbf{S}) + \left(G^2\sqrt{\frac{\eta}{\lambda}} + \frac{G^2}{\lambda m} + \frac{G^2}{\lambda n}\right)\log(mn)\log(1/\delta) + \frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

Note that prior work on robust adversarial meta-learning [Yin et al., 2018, Goldblum et al., 2020, Wang et al., 2021] focuses on empirical study of the problem; we present first theoretical guarantees.

6 Conclusion

In this paper, we introduce a novel notion of stability for meta-learning algorithms, namely uniform meta-stability, and offer a tighter bound on the generalization gap for the meta-learning problem compared to existing literature. We instantiate uniformly meta-stable learning algorithms and give generalization guarantees for both convex, smooth losses as well as weakly convex and non-smooth losses. Several avenues for further exciting research remain. For instance, it remains to be seen if our bounds are tight. Can we show lower bounds on the generalization error for meta-learning? Additionally, understanding how meta-learning relates to federated learning may offer insights on how to extend the theory to broader applications and inform the design of new algorithms. Finally, motivated by data privacy considerations, it would be interesting to extend our setup to privacy-preserving meta-learning, similar in spirit to the recent work of Zhou and Bassily [2022].

Acknowledgments and Disclosure of Funding

This research was supported, in part, by the DARPA GARD award HR00112020004, NSF CAREER award IIS-1943251, funding from the Institute for Assured Autonomy (IAA) at JHU, and the Spring'22 workshop on "Learning and Games" at the Simons Institute for the Theory of Computing. YW acknowledges the support of Amazon Fellowship.

References

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. *arXiv preprint arXiv:1710.03641*, 2017.

Maruan Al-Shedivat, Liam Li, Eric Xing, and Ameet Talwalkar. On data efficiency of meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1369–1377. PMLR, 2021.

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended pac-bayes theory. In *International Conference on Machine Learning*, pages 205–214. PMLR, 2018.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433. PMLR, 2019.

Fan Bao, Guoqiang Wu, Chongxuan Li, Jun Zhu, and Bo Zhang. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in neural information processing systems*, 34:4529–4541, 2021.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33: 4381–4391, 2020.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Shai Ben-David and Reba Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 567–580. Springer, 2003.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

Jiaxin Chen, Xiao-Ming Wu, Yanke Li, Qimai Li, Li-Ming Zhan, and Fu-lai Chung. A closer look at the training strategy for modern meta-learning. *Advances in Neural Information Processing Systems*, 33:396–406, 2020.

Qi Chen, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. *Advances in Neural Information Processing Systems*, 34:25878–25890, 2021.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, 33:18860–18871, 2020.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. *Advances in neural information processing systems*, 31, 2018.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575. PMLR, 2019a.

Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, and Massimiliano Pontil. Online-within-online meta-learning. *Advances in Neural Information Processing Systems*, 32, 2019b.

Giulia Denevi, Massimiliano Pontil, and Carlo Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. *Advances in Neural Information Processing Systems*, 33: 964–974, 2020.

Nan Ding, Xi Chen, Tomer Levinboim, Sebastian Goodman, and Radu Soricut. Bridging the gap between practice and pac-bayes theory in few-shot meta-learning. *Advances in Neural Information Processing Systems*, 34:29506–29516, 2021.

Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. arxiv preprint: 1908.10400, 2019.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Generalization of model-agnostic meta-learning algorithms: Recurring and unseen tasks. *Advances in Neural Information Processing Systems*, 34:5469–5480, 2021.

 https://doi.org/10.52202/079017-2661

Alec Farid and Anirudha Majumdar. Generalization bounds for meta-learning via pac-bayes and uniform stability. *Advances in neural information processing systems*, 34:2173–2186, 2021.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018.

Micah Goldblum, Liam Fowl, and Tom Goldstein. Adversarially robust few-shot learning: A meta-learning approach. *Advances in Neural Information Processing Systems*, 33:17886–17895, 2020.

Jiechao Guan and Zhiwu Lu. Task relatedness-based generalization bounds for meta learning. In *International Conference on Learning Representations*, 2021.

Jiechao Guan, Yong Liu, and Zhiwu Lu. Fine-grained analysis of stability and generalization for modern meta learning algorithms. *Advances in Neural Information Processing Systems*, 35: 18487–18500, 2022.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Fredrik Hellström and Giuseppe Durisi. Evaluated cmi bounds for meta learning: Tightness and expressiveness. *Advances in Neural Information Processing Systems*, 35:20648–20660, 2022.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.

Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

Weisen Jiang, James Kwok, and Yu Zhang. Effective meta-regularization by kernelized proximal regularization. *Advances in Neural Information Processing Systems*, 34:26212–26222, 2021.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021.

Sharu Theresa Jose, Osvaldo Simeone, and Giuseppe Durisi. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 68(1):474–501, 2021.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.

Ilja Kuzborskij and Francesco Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106:171–195, 2017.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 191–227. PMLR, 2023.

Jeffrey Li, Mikhail Khodak, Sebastian Caldas, and Ameet Talwalkar. Differentially private meta-learning. *arXiv preprint arXiv:1909.05830*, 2019.

Tianyu Liu, Jie Lu, Zheng Yan, and Guangquan Zhang. Pac-bayes bounds for meta-learning with data-dependent prior. *arXiv preprint arXiv:2102.03748*, 2021.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6 (6), 2005.

Andreas Maurer. Transfer bounds for linear feature learning. *Machine learning*, 75(3):327–350, 2009.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Dimitri Meunier and Pierre Alquier. Meta-strategy for learning tuning parameters with guarantees. *Entropy*, 23(10):1257, 2021.

Robert Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

Konstantin Mishchenko, Slavomir Hanzely, and Peter Richtárik. Convergence of first-order algorithms for meta-learning with moreau envelopes. *arXiv preprint arXiv:2301.06806*, 2023.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Anastasia Pentina and Christoph Lampert. A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR, 2014.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Yao-Feng Ren and Han-Ying Liang. On the best constant in marcinkiewicz–zygmund inequality. *Statistics & probability letters*, 53(3):227–233, 2001.

Arezou Rezazadeh. A unified view on pac-bayes bounds for meta-learning. In *International Conference on Machine Learning*, pages 18576–18595. PMLR, 2022.

Arezou Rezazadeh, Sharu Theresa Jose, Giuseppe Durisi, and Osvaldo Simeone. Conditional mutual information-based generalization bound for meta learning. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1176–1181. IEEE, 2021.

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR, 2021.

Charles Riou, Pierre Alquier, and Badr-Eddine Chérief-Abdellatif. Bayes meets bernstein at the meta level: an analysis of fast rates in meta-learning with pac-bayes. *arXiv preprint arXiv:2302.11709*, 2023.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pages 9116–9126. PMLR, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

 https://doi.org/10.52202/079017-2661

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR, 2024.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Chunyang Wang, Yanmin Zhu, Haobing Liu, Tianzi Zang, Jiadi Yu, and Feilong Tang. Deep meta-learning in recommendation systems: A survey. *arXiv preprint arXiv:2206.04415*, 2022.

Ren Wang, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Tsui-Wei Weng, Chuang Gan, and Meng Wang. On fast adversarial robustness adaptation in model-agnostic meta-learning. *arXiv preprint arXiv:2102.10454*, 2021.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459, 2022.

Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *Advances in neural information processing systems*, 34:26523–26535, 2021.

Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018.

Hossein Zakerinia, Amin Behjati, and Christoph H Lampert. More flexible pac-bayesian meta-learning by learning learning algorithms. *arXiv preprint arXiv:2402.04054*, 2024.

Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, and Jiashi Feng. Efficient meta learning via minibatch proximal update. *Advances in Neural Information Processing Systems*, 32, 2019.

Xinyu Zhou and Raef Bassily. Task-level differentially private meta learning. *Advances in Neural Information Processing Systems*, 35:20947–20959, 2022.

Yi Zhou, Yingbin Liang, and Huishuai Zhang. Understanding generalization error of sgd in nonconvex optimization. *Machine Learning*, pages 1–31, 2022.

Supplementary Material

A Experiments

In this section, we conduct a simple experiment to empirically verify our generalization bounds.

Setting. Following the experimental setting in Nichol et al. [2018] and Zhou et al. [2019], we consider a synthetic one-dimensional sine wave regression problem. The goal is to approximate the distribution of parameters of function $f(x;\alpha,\beta) = \alpha\sin(x+\beta)$. The task environment μ is a joint distribution $\mathcal{D}(\alpha,\beta)$ of the parameters α and β. We take $\mathcal{D}(\alpha,\beta)$ to be a product distribution of $\mathcal{D}(\alpha) = \mathrm{U}([-5,5]), \mathcal{D}(\beta) = \mathrm{U}([0,\pi])$. We generate the meta-sample by first sampling m training tasks, i.e., m pairs of (α,β) sampled independently from $\mathcal{D}(\alpha,\beta)$. For each of these m tasks, we sample $n = 10$ points, $x_1,\ldots,x_{10}$ uniformly on $[-5,5]$ and label them as $y_i = f(x_i;\alpha,\beta)$. Similarly, at test time we generate a new task from the task distribution and generate a training sample of size n (by sampling x's uniformly on the interval $[-5,5]$ and labeling them using $f(x;\alpha,\beta)$. We sample 1000 new tasks at test time. For each of the test task, we also generate an evaluation set of size 200, and use it to estimate the mean-squared error between the predictions of the learned model and the true labels. Our hypothesis class is a two layer network of width 40 and $tanh(\cdot)$ activation function. We run Algorithm 1 for $T = 100$ iterations with a step size of $\gamma = 0.1$ and regularization parameter $\lambda = 0.5$. Algorithm 2 (GD) is run for $K = 15$ iterations with step size $\eta = 0.02$. The experiment is conducted on a T4 GPU.

Results. We report the transfer risk, the average empirical risk (over tasks), and the generalization gap for different values of m and n in Figure 1. In the plot on the left, we fix $n = 10$, and vary the number of tasks m from 10 to 5000. In the plot in the middle, we fix $m = 1000$, and change the number of samples n from 5 to 1000. In the plot on the right, we choose $m = n$, and scale m and n simultaneously from 10 to 1000. We observe that in all of these three scenarios, as m (and/or n) increase, both the generalization gap as well as the transfer risk decrease. Moreover, the generalization gap decreases at rates approximately $\mathcal{O}(1/m + 1/n + 1/\sqrt{mn})$ for these three scenarios as suggested by our theoretical result.

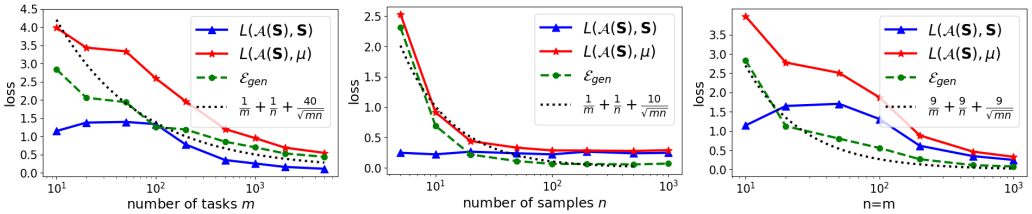

Figure 1: Error plots as a function of: (**left**) the number of tasks m for fixed $n = 10$; (**middle**) the number of training samples n on a test task for fixed $m = 1$K; (**right**) both m and n.

B Missing Proofs of Section 3

The following Lemmas and Theorems are used for proving Theorem 3.1.

Lemma B.1 (Bounded differences/McDiarmid's inequality)**.** Consider a function f of independent random variables $z_1,\ldots,z_n$ that take their value in $\mathcal{Z}$. Suppose that f satisfies the bounded differences property, namely, for any $i = 1,\ldots,n$ and any $z_1,\ldots,z_n,z_i' \in \mathcal{Z}$, it holds that

$$f(z_1,\ldots,z_n) - f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n) \leq \beta.$$

Then we have for any $p \geq 2$,

$$\|f(z_1,\ldots,z_n) - \mathbb{E}f(z_1,\ldots,z_n)\|_p \leq 2\sqrt{np}\beta.$$

Theorem B.2 (Marcinkiewicz-Zygmund's inequality Ren and Liang [2001])**.** Let $x_1,\ldots,x_n$ be independent centered random variables with a finite p-th moment for $p \geq 2$. Then

$$\left\|\sum_{i=1}^n x_i\right\|_p \leq 3\sqrt{2np}\left(\frac{1}{n}\sum_{i=1}^n \|x_i\|_p^p\right)^{\frac{1}{p}}$$

https://doi.org/10.52202/079017-2661

Theorem B.3. Let $\mathrm{Z} = (\mathrm{z}_1, \dots, \mathrm{z}_n)$ be a vector of independent random variables each taking values in $\mathcal{Z}$. Let $\mathbf{Z} = (\mathrm{Z}_1, \dots, \mathrm{Z}_m)$ be a vector of independent random vectors each taking values in $\mathcal{Z}^n$. Let $g_{j,i} : (\mathcal{Z}^n)^m \times \mathcal{Z}^n \to \mathbb{R}$ be some functions such that the following holds for any $i \in [n], j \in [m]$:

(1). $|\mathbb{E}[g_{j,i}(\mathbf{Z}, \mathrm{Z})|\mathrm{Z}_j, \mathrm{z}_i]| \leq M$ a.s.,

(2). $\mathbb{E}\left[g_{j,i}(\mathbf{Z}, \mathrm{Z})|\mathrm{Z}_{[m]\setminus\{j\}}, \mathrm{z}_{[n]\setminus\{i\}}\right] = 0$ a.s.,

(3). $g_{j,i}$ has a bounded difference $\bar{\beta}$ w.r.t. all variables except the (j,i)-th variable.

Then we have

$$\left\| \sum_{j=1}^{m} \sum_{i=1}^{n} g_{j,i}(\mathbf{Z}, \mathrm{Z}) \right\| \lesssim mn\bar{\beta} \log(mn) + M\sqrt{mn}.$$

Proof of Theorem B.3. The proof is an extension of [Bousquet et al., 2020, Theorem 4].

Without loss of generality, we suppose that $n = 2^k, m = 2^r$. Otherwise, we can add extra functions that equal to zero. Consider a sequence of partitions $\mathcal{C}_0, \dots, \mathcal{C}_k$ with $\mathcal{C}_0 = \{\{i\} : i \in [n]\}$, $\mathcal{C}_k = \{[n]\}$, and to get $\mathcal{C}_l$ from $\mathcal{C}_{l+1}$ we split each subset into $\mathcal{C}_{l+1}$ into two equal parts. We have

$$\mathcal{C}_0 = \left\{\{1\}, \dots, \left\{2^k\right\}\right\}, \mathcal{C}_1 = \left\{\{1,2\}, \{3,4\}, \dots, \left\{2^k - 1, 2^k\right\}\right\}, \mathcal{C}_k = \left\{\left\{1, \dots, 2^k\right\}\right\}.$$

By construction, we have $|\mathcal{C}_l| = 2^{k-l}$ and $|C| = 2^l$ for each $C \in \mathcal{C}_l$. For each $i \in [n]$ and $l = 0, \dots, k$, denote by $C^l(i) \in \mathcal{C}_l$ the only set from $\mathcal{C}_l$ that contains i. In particular, $C^0(i) = \{i\}$ and $C^k(i) = [n]$.

Similarly, we consider a sequence of partitions $\mathcal{E}_0, \dots, \mathcal{E}_r$ with $\mathcal{E}_0 = \{\{j\} : j \in [m]\}$, $\mathcal{E}_r = \{[m]\}$, and to get $\mathcal{E}_q$ from $\mathcal{E}_{q+1}$ we split each subset in $\mathcal{E}_{q+1}$ into two equal parts. We have

$$\mathcal{E}_0 = \{\{1\}, \dots, \{2^r\}\}, \mathcal{E}_1 = \{\{1,2\}, \{3,4\}, \dots, \{2^r - 1, 2^r\}\}, \mathcal{E}_r = \{\{1, \dots, 2^r\}\}.$$

By construction, we have $|\mathcal{E}_q| = 2^{r-q}$ and $|E| = 2^q$ for each $E \in \mathcal{E}_q$. For each $j \in [m]$ and $q = 0, \dots, r$, denote by $E^q(j) \in \mathcal{E}_q$ the only set from $\mathcal{E}_q$ that contains j. In particular, $E^0(j) = \{j\}$ and $E^r(j) = [m]$.

For each $i \in [n], j \in [m]$ and every $l = 0, \dots, k$, $q = 0, \dots, r$, consider the random variables

$$g_{j,i}^{q,l} = g_{j,i}^{q,l}(\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^q(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^l(i)}),$$

i.e., conditioned on $\mathrm{Z}_j, \mathrm{z}_i$ and all the vectors that are not in the same set as Z_j in the partition $\mathcal{E}_q$ and all the variables that are not in the same set as z_i in the partition $\mathcal{C}_l$. In particular, $g_{j,i}^{0,0} = g_{j,i}$, $g_{j,i}^{r,k} = \mathbb{E}[g_{j,i}|\mathrm{Z}_j, \mathrm{z}_i]$. We can write a telescopic sum as follows:

$$g_{j,i} - \mathbb{E}[g_{j,i}|\mathrm{Z}_j, \mathrm{z}_i] = \sum_{q=0}^{r-1} g_{j,i}^{q,0} - g_{j,i}^{q+1,0} + \sum_{l=0}^{k-1} g_{j,i}^{r,l} - g_{j,i}^{r,l+1},$$

and the total sum of interest satisfies by the triangle inequality

$$\left\| \sum_{j=1}^{m} \sum_{i=1}^{n} g_{j,i} \right\| \leq \left\| \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{E}\left[g_{j,i}|\mathrm{Z}_j, \mathrm{z}_i\right] \right\| + \sum_{q=0}^{r-1} \left\| \sum_{j=1}^{m} \sum_{i=1}^{n} g_{j,i}^{q,0} - g_{j,i}^{q+1,0} \right\| + \sum_{l=0}^{k-1} \left\| \sum_{j=1}^{m} \sum_{i=1}^{n} g_{j,i}^{r,l} - g_{j,i}^{r,l+1} \right\|.$$

Since $|\mathbb{E}[g_{j,i}|\mathrm{Z}_j, \mathrm{z}_i]| \leq M$ and $\mathbb{E}\left(\mathbb{E}\left[g_{j,i}|\mathrm{Z}_j, \mathrm{z}_i\right]\right) = 0$, by applying McDiarmid inequality in Lemma B.1, we have

$$\left\| \sum_{j=1}^{m} \sum_{i=1}^{n} \mathbb{E}\left[g_{j,i}|\mathrm{Z}_j, \mathrm{z}_i\right] \right\| \leq 4\sqrt{2mn}M. \tag{2}$$

We observe that

$$
\begin{aligned}
&g_{j,i}^{q+1,l+1}(\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{q+1}(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l+1}(i)}) \\
&= \mathbb{E}\left[g_{j,i}^{q+1,l}(\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{q+1}(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l}(i)})|\mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l+1}(i)}\right] \\
&\qquad\qquad \text{(The expectation is take w.r.t. the variable } \mathrm{z}_s, s \in C^{l+1}(i)\setminus C^l(i)) \\
&= \mathbb{E}\left[g_{j,i}^{q,l+1}(\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{q}(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l+1}(i)})|\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{q+1}(j)}\right] \\
&\qquad\qquad \text{(The expectation is take w.r.t. the variable } \mathrm{Z}_s, s \in E^{q+1}(j)\setminus E^q(j))
\end{aligned}
$$

As function $g_{j,i}^{q,l}$ preserves the bounded differences property, if we apply McDiarmid's inequality conditioned on $\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{q+1}(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l+1}(i)}$, we obtain a uniform bound

$$
\begin{aligned}
\left\|g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\|\left(\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{q+1}(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{0}(i)}\right) &\le 2\sqrt{2^{q+1}}\bar{\beta} \\
\left\|g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\|\left(\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^{r}(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l+1}(i)}\right) &\le 2\sqrt{2^{l+1}}\bar{\beta}
\end{aligned}
$$

as there are 2^l indices in $C^{l+1}(i)\setminus C^l(i)$ and 2^q indices in $E^{q+1}(j)\setminus E^q(j)$.

Now we focus on $\sum_{j\in E^q}\sum_{i\in C^0} g_{j,i}^{q,0} - g_{j,i}^{q+1,0}$ for $E^q \in \mathcal{E}_q$ and $\sum_{j\in E^r}\sum_{i\in C^l} g_{j,i}^{r,l} - g_{j,i}^{r,l+1}$ for $C^l \in \mathcal{C}_l$, respectively. Since $g_{j,i}^{q,0} - g_{j,i}^{q+1,0}$ for $j \in E^q, i \in C^0$ depends on $\mathrm{Z}_j, \mathrm{Z}_{[m]\setminus E^q(j)}, \mathrm{z}_i, \mathrm{z}_{[n]\setminus C^0(i)}$, the terms are independent and zero mean conditioned on $\mathrm{Z}_{[m]\setminus E^q(j)}$. Applying Theorem B.2, we have

$$
\begin{aligned}
&\left\|\sum_{j\in E^q}\sum_{i\in C^0} g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\|^2 (\mathrm{Z}_{[m]\setminus E^q}) \\
&\qquad \le 36\cdot 2^q \frac{1}{2^q}\sum_{j\in E^q}\sum_{i\in C^0}\left\|g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\|^2 (\mathrm{Z}_{[m]\setminus E^q})
\end{aligned}
$$

Integrating with respect to $(Z_{[m]\setminus E^q})$ and using $\left\|g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\| \le 2\sqrt{2^{q+1}}\bar{\beta}$, we have

$$
\left\|\sum_{j\in E^q}\sum_{i\in C^0} g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\| \le 6\sqrt{2^q}\times 2\sqrt{2^{q+1}}\bar{\beta} = 12\sqrt{2}\cdot 2^q\bar{\beta}.
$$

Applying triangle inequality over all sets $C^0 \in \mathcal{C}_0, E^q \in \mathcal{E}_q$ gives us that

$$
\begin{aligned}
\left\|\sum_{j\in[m]}\sum_{i\in[n]} g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\| &\le \sum_{E^q\in\mathcal{E}_q, C^0\in\mathcal{C}_0}\left\|\sum_{j\in E^q, i\in C^0} g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\| \\
&\le 2^{r+k-q}\times 12\sqrt{2}\cdot 2^q\bar{\beta} \\
&= 12\sqrt{2}\cdot 2^{r+k}\bar{\beta}.
\end{aligned}
$$

Similarly, $g_{j,i}^{r,l} - g_{j,i}^{r,l+1}$ for $j \in E^r, i \in C^l$ depends on $\mathrm{z}_i, \mathrm{z}_{[n]\setminus C^{l+1}(i)}$, the terms are independent and zero mean conditioned on $\mathrm{z}_{[n]\setminus C^{l+1}(i)}$. Applying Theorem B.2, we have

$$
\begin{aligned}
&\left\|\sum_{j\in E^r}\sum_{i\in C^l} g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\|^2 (\mathrm{z}_{[n]\setminus C^l}) \\
&\qquad \le 36\cdot 2^{l+r}\frac{1}{2^{l+r}}\sum_{j\in E^r}\sum_{i\in C^l}\left\|g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\|^2 (\mathrm{z}_{[n]\setminus C^l})
\end{aligned}
$$

Integrating with respect to $(\mathrm{z}_{[n]\setminus C^l})$ and using $\left\|g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\| \le 2\sqrt{2^{l+1}}\bar{\beta}$, we have

$$
\left\|\sum_{j\in E^r}\sum_{i\in C^l} g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\| \le 6\sqrt{2^{l+r}}\times 2\sqrt{2^{l+1}}\bar{\beta} = 12\sqrt{2}\cdot 2^{l+0.5r}\bar{\beta}.
$$

https://doi.org/10.52202/079017-2661

Applying triangle inequality over all sets $C^l \in \mathcal{C}_l, E^r \in \mathcal{E}_r$ gives us that

$$\begin{aligned}\left\|\sum_{j\in[m]}\sum_{i\in[n]} g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\| &\leq \sum_{E^r\in\mathcal{E}_r, C^l\in\mathcal{C}_l}\left\|\sum_{j\in E^r, i\in C^l} g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\| \\ &\leq 2^{k-l}\times 12\sqrt{2}\cdot 2^{l+0.5r}\bar{\beta} \\ &\leq 12\sqrt{2}\cdot 2^{r+k}\bar{\beta}.\end{aligned}$$

Recall that $2^k < 2n, 2^r < 2m$ due to the possible extension of the sample. Therefore we have

$$\begin{aligned}\sum_{q=0}^{r-1}\left\|\sum_{j=1}^{m}\sum_{i=1}^{n} g_{j,i}^{q,0} - g_{j,i}^{q+1,0}\right\| + \sum_{l=0}^{k-1}\left\|\sum_{j=1}^{m}\sum_{i=1}^{n} g_{j,i}^{r,l} - g_{j,i}^{r,l+1}\right\| &\leq 48\sqrt{2}mn\bar{\beta}(\lceil\log(m)\rceil + \lceil\log(n)\rceil) \\ &\lesssim mn\bar{\beta}\log(mn)\end{aligned}$$

Combined with Equation (2) get the required bound.

□

We now restate and prove Theorem 3.1.

Theorem 3.1. Consider a meta-learning problem for some M-bounded loss function ℓ and task distribution μ. Let $\mathbf{S}$ be a meta-sample consisting of training samples on m tasks each of size n, and let $\mathcal{S} \sim \mathcal{D}$ be a sample of size n on a previously unseen task $\mathcal{D} \sim \mu$. Then, for any β-uniformly meta-stable learning algorithm $\mathcal{A}$, we have that with probability $1-\delta$,

$$L(\mathcal{A}(\mathbf{S}), \mu) \lesssim L(\mathcal{A}(\mathbf{S}), \mathbf{S}) + \bar{\beta}\log(mn)\log(1/\delta) + M\sqrt{\log(1/\delta)/(mn)}.$$

Proof of Theorem 3.1. In order to make use of Theorem B.3, we consider the following functions:

$$g_{j,i} = g_{j,i}(\mathbf{Z}, \mathrm{Z}) = \mathbb{E}_{(\mathcal{S}'_j, \mathrm{z}'_j)\sim\mathcal{D}_j^{n+1}, \mathcal{D}_j\sim\mu}\mathbb{E}_{(\mathcal{S},\mathrm{z})\sim\mathcal{D}^{n+1},\mathcal{D}\sim\mu}\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}), \mathrm{z}) - \ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}), \mathrm{z}_j^i)$$

By the definition of uniform meta stability, we can write the following decomposition:

$$\begin{aligned}&|mn\left(L(\mathcal{A}(\mathbf{S}), \mu) - L(\mathcal{A}(\mathbf{S}), \mathcal{S})\right)| \\ &= \left|\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{(\mathcal{S},\mathrm{z})\sim\mathcal{D}^{n+1},\mathcal{D}\sim\mu}\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathrm{z}) - \ell(\mathcal{A}(\mathbf{S})(\mathcal{S}_j), \mathrm{z}_j^i)\right| \\ &= \left|\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{\substack{(\mathcal{S},\mathrm{z})\sim\mathcal{D}^{n+1},\\ \mathcal{D}\sim\mu}}\mathbb{E}_{\substack{(\mathcal{S}'_j,\mathrm{z}'_j)\sim\mathcal{D}_j^{n+1},\\ \mathcal{D}_j\sim\mu}}\Big(\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathrm{z}) - \ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}), \mathrm{z}_j^i)\right. \\ &\qquad \left.+\,\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}), \mathrm{z}_j^i) - \ell(\mathcal{A}(\mathbf{S})(\mathcal{S}_j), \mathrm{z}_j^i)\Big)\right| \\ &\leq \left\|\sum_{j=1}^{m}\sum_{i=1}^{n}\mathbb{E}_{\substack{(\mathcal{S},\mathrm{z})\sim\mathcal{D}^{n+1},\\ \mathcal{D}\sim\mu}}\mathbb{E}_{\substack{(\mathcal{S}'_j,\mathrm{z}'_j)\sim\mathcal{D}_j^{n+1},\\ \mathcal{D}_j\sim\mu}}\left(\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}), \mathrm{z}) - \ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}), \mathrm{z}_j^i)\right)\right\| + mn\bar{\beta} \\ &= \left\|\sum_{j=1}^{m}\sum_{i=1}^{n} g_{j,i}\right\| + mn\bar{\beta}\end{aligned}$$

Moreover, we have $\mathbb{E}\left[g_{j,i}|\mathcal{S}_1, \ldots, \mathcal{S}_{j-1}, \mathcal{S}_{j+1}, \ldots, \mathcal{S}_m, \mathrm{z}_1, \ldots, \mathrm{z}_{i-1}, \mathrm{z}_{i+1}, \ldots, \mathrm{z}_n\right] = 0$ and $|g_{j,i}| \leq 2M$ a.s. for $i \in [n], j \in [m]$. Applying Theorem B.3 as well as [Bousquet and Elisseeff, 2002, Lemma 1] achieves the results.

□

Theorem 3.2 can be directly proved by the definition of uniform meta-stability.

Theorem 3.2. Let μ be an underlying task distribution. Given a meta-sample $\mathbf{S}$, test task $\mathcal{D} \sim \mu$, and $\mathcal{S} \sim \mathcal{D}^n$, for any $\bar{\beta}$ -on-average-replace-one-meta-stable meta-learning algorithm $\mathcal{A}$, we have that

$$\mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m}[L(\mathcal{A}(\mathbf{S}),\mu) - L(\mathcal{A}(\mathbf{S}),\mathbf{S})] \leq \bar{\beta}.$$

Proof of Theorem 3.2. Since $\mathcal{S}$ and z' are both drawn i.i.d. from $\mathcal{D}$, and $\mathbf{S}$ and $\mathcal{S}_j'$ are both drawn i.i.d. from μ, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^n,\mathcal{D}\sim\mu}L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{D}) \\
&= \mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,(\mathcal{S}_j')\sim\mathcal{D}_j^n,\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m,j\sim\mathrm{U}[m]}L(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j),\mathcal{D}_j) \\
&= \mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,(\mathcal{S}_j',\mathrm{z}_j')\sim\mathcal{D}_j^{n+1},\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m,j\sim\mathrm{U}[m],i\sim\mathrm{U}[n]}\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}),\mathrm{z}_j^i)
\end{aligned}$$

as well as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m}\left[\frac{1}{m}\sum_{j=1}^m L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathcal{S}_j)\right] \\
&= \mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m,j\sim\mathrm{U}[m]}\left[L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathcal{S}_j)\right] \\
&= \mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,(\mathcal{S}_j',\mathrm{z}_j')\sim\mathcal{D}_j^{n+1},\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m,j\sim\mathrm{U}[m],i\sim\mathrm{U}[n]}\mathbb{E}_{\mathcal{S}\sim\mathcal{D}^n}\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathrm{z}_j^i)
\end{aligned}$$

As a result, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m}[L(\mathcal{A}(\mathbf{S}),\mu) - L(\mathcal{A}(\mathbf{S}),\mathbf{S})] \\
&= \mathbb{E}_{\mathbf{S}\sim\left\{\mathcal{D}_j^n\right\}_{j=1}^m,(\mathcal{S}_j',\mathrm{z}_j')\sim\mathcal{D}_j^{n+1},\{\mathcal{D}_j\}_{j=1}^m\sim\mu^m,j\sim\mathrm{U}[m],i\sim\mathrm{U}[n]}\left|\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}_j^{(i)}),\mathrm{z}_j^i) - \ell(\mathcal{A}(\mathbf{S})(\mathcal{S}_j),\mathrm{z}_j^i)\right| \\
&\leq \bar{\beta} \qquad \text{(By definition of } \bar{\beta} \text{ -on-average-replace-one-meta-stable)}
\end{aligned}$$

□

C Missing Proofs of Section 4.1

Lemma C.1 (Shalev-Shwartz and Ben-David [2014])**.** Given $\mathcal{S}$ and $\mathcal{S}^{(i)}$, for a fixed w, define $\mathrm{u}(\mathrm{w},\mathcal{S})$ and $\mathrm{u}(\mathrm{w},\mathcal{S}^{(i)})$ is achieved via Algo. 1 with Option 1 RERM. Then if ℓ is convex, G-Lipschitz, we have $\sup_{\mathcal{S},i\in[n]}\left\|\mathrm{u}(\mathrm{w},\mathcal{S}) - \mathrm{u}(\mathrm{w},\mathcal{S}^{(i)})\right\| \leq \frac{4G}{\lambda n}$. If ℓ is convex and H-smooth ($H \leq \frac{\lambda n}{2}$), we have $\left\|\mathrm{u}(\mathrm{w},\mathcal{S}) - \mathrm{u}(\mathrm{w},\mathcal{S}^{(i)})\right\| \leq \frac{\sqrt{8H}}{\lambda n}(\sqrt{\ell(\mathrm{w},\mathrm{z}_i)} + \sqrt{\ell(\mathrm{w},\mathrm{z}')})$.

Lemma 4.1. Assume that the loss function ℓ is convex and G-Lipschitz loss. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then, the following holds for Algorithm 1 with RERM for task-specific learning (i.e., Option 1 for Algorithm 2) $\forall T \geq 1$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]}\left\|\mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})\right\| \leq \frac{G}{\lambda m} + \frac{2G}{\lambda n}.$$

Further, if ℓ is convex, M-bounded and H-smooth, then setting $\lambda \geq H$, $\gamma \leq \frac{1}{\lambda}$, we have $\forall T \geq 1$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]}\left\|\mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})\right\| \leq \frac{2\sqrt{2HM}}{2\lambda n - H} + \frac{n}{2\lambda n - H}\frac{4\sqrt{2HM}}{(m+1)}.$$

Proof of Lemma 4.1. We slightly abuse the notation, at iteration t, define $\mathrm{w}_t = \mathcal{A}(\mathbf{S})$, $\mathrm{w}_t' = \mathcal{A}(\mathbf{S}^{(j)})$. Given w_{T+1}, define $\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}) = \mathcal{A}(\mathbf{S})(\mathcal{S}), \mathrm{u}(\mathrm{w}_{T+1}',\mathcal{S}^{(i)}) = \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$.

We first consider the setting where the loss ℓ is convex, G-Lipschitz. Recall that $F_{\mathcal{S}}(\mathrm{u},\mathrm{w}) = L(\mathrm{u},\mathcal{S}) + \frac{\lambda}{2}\left\|\mathrm{u}-\mathrm{w}\right\|^2$. If ℓ is convex, then $F_{\mathcal{S}}(\mathrm{u},\mathrm{w})$ is λ-strongly-convex w.r.t u. Define $\mathrm{u}(\mathrm{w},\mathcal{S}) = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} F_{\mathcal{S}}(\mathrm{u},\mathrm{w})$, $\mathrm{u}(\mathrm{w}',\mathcal{S}) = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} F_{\mathcal{S}}(\mathrm{u},\mathrm{w}')$. We have the following:

$$\begin{aligned}
F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{w}) - F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}) &\geq \lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S}) - \mathrm{u}(\mathrm{w}',\mathcal{S})\right\|^2 \\
F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}') - F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{w}') &\geq \lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S}) - \mathrm{u}(\mathrm{w}',\mathcal{S})\right\|^2
\end{aligned}$$

https://doi.org/10.52202/079017-2661

Sum the above gives us that

$$
\begin{aligned}
&2\lambda \left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S})\right\|^2 \\
&\le F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{w})-F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w})+F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}')-F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{w}') \\
&= \frac{\lambda}{2}\left(\left\|\mathrm{u}(\mathrm{w}',\mathcal{S})-\mathrm{w}\right\|^2-\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{w}\right\|^2+\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{w}'\right\|^2-\left\|\mathrm{u}(\mathrm{w}',\mathcal{S})-\mathrm{w}'\right\|^2\right) \\
&= \lambda\left\langle \mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{w}-\mathrm{w}'\right\rangle \\
&\le \lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S})\right\|\left\|\mathrm{w}-\mathrm{w}'\right\|
\end{aligned}
$$

This gives us that

$$
\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S})\right\| \le \frac{1}{2}\left\|\mathrm{w}-\mathrm{w}'\right\|. \tag{3}
$$

Similarly, define $\mathrm{u}(\mathrm{w}',\mathcal{S}') = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} F_{\mathcal{S}'}(\mathrm{u},\mathrm{w}')$, we have

$$
\begin{aligned}
F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathrm{w})-F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}) &\ge \lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\|^2 \\
F_{\mathcal{S}'}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}')-F_{\mathcal{S}'}(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathrm{w}') &\ge \lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\|^2
\end{aligned}
$$

Sum the above gives us that

$$
\begin{aligned}
&2\lambda \left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\|^2 \\
&\le F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathrm{w})-F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w})+F_{\mathcal{S}'}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}')-F_{\mathcal{S}'}(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathrm{w}') \\
&= L(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathcal{S})-L(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathcal{S})+L(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathcal{S}')-L(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathcal{S}') \\
&\quad+\frac{\lambda}{2}\left(\left\|\mathrm{u}(\mathrm{w}',\mathcal{S}')-\mathrm{w}\right\|^2-\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{w}\right\|^2+\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{w}'\right\|^2-\left\|\mathrm{u}(\mathrm{w}',\mathcal{S}')-\mathrm{w}'\right\|^2\right) \\
&\le 2G\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\|+\lambda\left\langle \mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathrm{w}-\mathrm{w}'\right\rangle \qquad (\ell \text{ is } G\text{-Lipschitz}) \\
&\le 2G\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\|+\lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\|\left\|\mathrm{w}-\mathrm{w}'\right\|
\end{aligned}
$$

This gives us that

$$
\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\right\| \le \frac{1}{2}\left\|\mathrm{w}-\mathrm{w}'\right\|+\frac{G}{\lambda} \tag{4}
$$

Finally, at iteration t, we have

$$
\begin{aligned}
\left\|\mathrm{w}_{t+1}-\mathrm{w}'_{t+1}\right\| &\le \left\|\mathrm{w}_t-\gamma\lambda\left(\mathrm{w}_t-\frac{1}{m}\sum_{j=1}^{m}\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j)\right)-\mathrm{w}'_t+\gamma\lambda\left(\mathrm{w}'_t-\frac{1}{m}\sum_{j=1}^{m}\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j)\right)\right\| \\
&\qquad\qquad \text{(Projection is non-expansive)} \\
&= \left\|(1-\gamma\lambda)(\mathrm{w}_t-\mathrm{w}'_t)+\gamma\lambda\frac{1}{m}\sum_{j=1}^{m}\left(\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j)-\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j)\right)\right\| \\
&\le (1-\gamma\lambda)\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\gamma\lambda\left(\frac{m-1}{m}\frac{1}{2}\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\frac{1}{m}\left(\frac{1}{2}\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\frac{G}{\lambda}\right)\right) \\
&\qquad\qquad \text{(Equation (3), (4))} \\
&= (1-\frac{\gamma\lambda}{2})\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\frac{\gamma G}{m}
\end{aligned}
$$

Choose $\gamma \le \frac{1}{\lambda}, \forall t$. Rearrange gives us that

$$
\frac{\left\|\mathrm{w}_{t+1}-\mathrm{w}'_{t+1}\right\|}{(1-\gamma\lambda/2)^{t+1}} \le \frac{\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|}{(1-\gamma\lambda/2)^{t}}+\frac{\gamma G}{m}\frac{1}{(1-\gamma\lambda/2)^{t+1}}
$$

Note that at initialization when $t=1$ we have $\|\mathrm{w}_1-\mathrm{w}'_1\|=0$. Telescoping from $t=1$ to $T+1$

$$
\frac{\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\|}{(1-\gamma\lambda/2)^{T}} \le \frac{\gamma G}{m}\sum_{t=1}^{T-1}\frac{1}{(1-\gamma\lambda/2)^{t+1}}
$$

Calculate gives us that

$$\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\| \leq \frac{2G}{\lambda m}$$

Similarly, define $\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} F_{\mathcal{S}}(\mathrm{u}, \mathrm{w}'_{T+1})$,
$\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}) = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} F_{\mathcal{S}^{(i)}}(\mathrm{u}, \mathrm{w}'_{T})$, we have

$$F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathrm{w}_{T+1}) - F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}), \mathrm{w}_{T+1}) \geq \lambda \left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\|^2$$
$$F_{\mathcal{S}^{(i)}}(\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}), \mathrm{w}'_{T+1}) - F_{\mathcal{S}^{(i)}}(\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathrm{w}'_{T+1}) \geq \lambda \left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\|^2$$

Sum the above gives us that

$$\begin{aligned}
&2\lambda \left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\|^2 \\
&\leq F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathrm{w}_{T+1}) - F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}), \mathrm{w}_{T+1}) \\
&\quad + F_{\mathcal{S}^{(i)}}(\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}), \mathrm{w}'_{T+1}) - F_{\mathcal{S}^{(i)}}(\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathrm{w}'_{T+1}) \\
&= L(\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathcal{S}) - L(\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}), \mathcal{S}) + L(\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}), \mathcal{S}^{(i)}) - L(\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathcal{S}^{(i)}) \\
&\quad + \frac{\lambda}{2}\Bigg(\left\|\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}) - \mathrm{w}_{T+1}\right\|^2 - \left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{w}_{T+1}\right\|^2 \\
&\qquad + \left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{w}'_{T+1}\right\|^2 - \left\|\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}) - \mathrm{w}'_{T+1}\right\|^2\Bigg) \\
&\leq \frac{2G}{n}\left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\| + \lambda\left\langle \mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}^{(i)}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}), \mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\rangle \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\ell \text{ is } G\text{-Lipschitz}) \\
&\leq \frac{2G}{n}\left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\| + \lambda\left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\|\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\|
\end{aligned}$$

This gives us that

$$\left\|\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\| \leq \frac{1}{2}\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\| + \frac{2G}{\lambda n} \leq \frac{G}{\lambda m} + \frac{2G}{\lambda n} \tag{5}$$

We now consider the surrogate loss ℓ is convex, non-negative and H-smooth. Note that such loss is also self-bounded. From a similar argument, we have

$$\left\|\mathrm{u}(\mathrm{w}, \mathcal{S}) - \mathrm{u}(\mathrm{w}', \mathcal{S})\right\| \leq \frac{1}{2}\left\|\mathrm{w} - \mathrm{w}'\right\|.$$

Moreover,

$$\begin{aligned}
&2\lambda \left\|\mathrm{u}(\mathrm{w}, \mathcal{S}) - \mathrm{u}(\mathrm{w}', \mathcal{S}')\right\|^2 \\
&\leq F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathrm{w}) - F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}, \mathcal{S}), \mathrm{w}) + F_{\mathcal{S}'}(\mathrm{u}(\mathrm{w}, \mathcal{S}), \mathrm{w}') - F_{\mathcal{S}'}(\mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathrm{w}') \\
&= L(\mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathcal{S}) - L(\mathrm{u}(\mathrm{w}, \mathcal{S}), \mathcal{S}) + L(\mathrm{u}(\mathrm{w}, \mathcal{S}), \mathcal{S}') - L(\mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathcal{S}') \\
&\quad + \frac{\lambda}{2}\left(\left\|\mathrm{u}(\mathrm{w}', \mathcal{S}') - \mathrm{w}\right\|^2 - \left\|\mathrm{u}(\mathrm{w}, \mathcal{S}) - \mathrm{w}\right\|^2 + \left\|\mathrm{u}(\mathrm{w}, \mathcal{S}) - \mathrm{w}'\right\|^2 - \left\|\mathrm{u}(\mathrm{w}', \mathcal{S}') - \mathrm{w}'\right\|^2\right) \\
&\leq (\|\nabla L(\mathrm{u}(\mathrm{w}, \mathcal{S}), \mathcal{S})\| + \|\nabla L(\mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathcal{S}')\|)\|\mathrm{u}(\mathrm{w}', \mathcal{S}') - \mathrm{u}(\mathrm{w}, \mathcal{S})\| + H\left\|\mathrm{u}(\mathrm{w}', \mathcal{S}') - \mathrm{u}(\mathrm{w}, \mathcal{S})\right\|^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\ell \text{ is } H\text{-smooth}) \\
&\quad + \lambda\left\langle \mathrm{u}(\mathrm{w}, \mathcal{S}) - \mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathrm{w} - \mathrm{w}'\right\rangle \\
&\leq \left(\sqrt{2HL(\mathrm{u}(\mathrm{w}, \mathcal{S}), \mathcal{S})} + \sqrt{2HL(\mathrm{u}(\mathrm{w}', \mathcal{S}'), \mathcal{S}')}\right)\left\|\mathrm{u}(\mathrm{w}', \mathcal{S}') - \mathrm{u}(\mathrm{w}, \mathcal{S})\right\| \\
&\quad + H\left\|\mathrm{u}(\mathrm{w}', \mathcal{S}') - \mathrm{u}(\mathrm{w}, \mathcal{S})\right\|^2 + \lambda\left\|\mathrm{u}(\mathrm{w}, \mathcal{S}) - \mathrm{u}(\mathrm{w}', \mathcal{S})\right\|\left\|\mathrm{w} - \mathrm{w}'\right\| \qquad (\ell \text{ is } H\text{-smooth})
\end{aligned}$$

https://doi.org/10.52202/079017-2661

which is equivalent as

$$\begin{aligned}\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}')\| &\le \frac{\sqrt{2HL(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathcal{S})}+\sqrt{2HL(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathcal{S}')}+\lambda\|\mathrm{w}-\mathrm{w}'\|}{2\lambda-H} \quad (\lambda\ge H)\\ &\le \frac{\sqrt{2H}}{\lambda}\left(\sqrt{L(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathcal{S})}+\sqrt{L(\mathrm{u}(\mathrm{w}',\mathcal{S}'),\mathcal{S}')}\right)+\|\mathrm{w}-\mathrm{w}'\| \qquad (6)\end{aligned}$$

Finally, at iteration t, we have

$$\begin{aligned}&\left\|\mathrm{w}_{t+1}-\mathrm{w}'_{t+1}\right\|\\ &\le \left\|\mathrm{w}_t-\gamma\lambda\left(\mathrm{w}_t-\frac{1}{m}\sum_{j=1}^{m}\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j)\right)-\mathrm{w}'_t+\gamma\lambda\left(\mathrm{w}'_t-\frac{1}{m}\sum_{j=1}^{m}\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j)\right)\right\|\\ &\qquad\qquad\text{(Projection is non-expansive)}\\ &= \left\|(1-\gamma\lambda)(\mathrm{w}_t-\mathrm{w}'_t)+\gamma\lambda\frac{1}{m}\sum_{j=1}^{m}\left(\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j)-\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j)\right)\right\|\\ &\le (1-\gamma\lambda)\|\mathrm{w}_t-\mathrm{w}'_t\|+\gamma\lambda\Bigg(\frac{m-1}{m}\frac{1}{2}\|\mathrm{w}_t-\mathrm{w}'_t\|\\ &\qquad+\frac{1}{m}\left(\frac{\sqrt{2H}}{\lambda}\left(\sqrt{L(\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j),\mathcal{S}_j)}+\sqrt{L(\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j),\mathcal{S}'_j)}\right)+\|\mathrm{w}_t-\mathrm{w}'_t\|\right)\Bigg)\\ &\qquad\qquad\text{(Equation (3), (6))}\\ &= \left(1-\frac{m+1}{2m}\gamma\lambda\right)\|\mathrm{w}_t-\mathrm{w}'_t\|+\frac{\gamma\sqrt{2H}}{m}\left(\sqrt{L(\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j),\mathcal{S}_j)}+\sqrt{L(\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j),\mathcal{S}'_j)}\right)\end{aligned}$$

Telescope gives us that

$$\begin{aligned}\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\| &\le \frac{\gamma\sqrt{2H\lambda}}{m}\sum_{t=1}^{T}\left(1-\frac{m+1}{2m}\gamma\lambda\right)^{T-t}\left(\sqrt{L(\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j),\mathcal{S}_j)}+\sqrt{L(\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j),\mathcal{S}'_j)}\right)\\ &\le \frac{4\sqrt{2HM}}{\lambda(m+1)}, \qquad (7)\end{aligned}$$

where the last line holds if we consider M-bounded loss. Otherwise, we have

$$\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\| \le \frac{2\sqrt{2H}}{\lambda(m+1)}\left(\sqrt{\max_{t\in[T]}L(\mathrm{u}(\mathrm{w}_t,\mathcal{S}_j),\mathcal{S}_j)}+\sqrt{\max_{t\in[T]}L(\mathrm{u}(\mathrm{w}'_t,\mathcal{S}'_j),\mathcal{S}'_j)}\right). \qquad (8)$$

Therefore, we have

$$\begin{aligned}&2\lambda\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)})\right\|^2\\ &\le F_{\mathcal{S}}(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{w})-F_{\mathcal{S}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w})+F_{\mathcal{S}^{(i)}}(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{w}')-F_{\mathcal{S}^{(i)}}(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{w}')\\ &= L(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S})-L(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathcal{S})+L(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathcal{S}^{(i)})-L(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S}^{(i)})\\ &\quad+\frac{\lambda}{2}\left(\left\|\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)})-\mathrm{w}\right\|^2-\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{w}\|^2+\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{w}'\|^2-\left\|\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)})-\mathrm{w}'\right\|^2\right)\\ &\le \frac{1}{n}\left(\ell(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{z}^i)-\ell(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{z}')+\ell(\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{z}')-\ell(\mathrm{u}(\mathrm{w}',\mathcal{S}),\mathrm{z}^i)\right)\\ &\quad+\lambda\left\langle\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{w}-\mathrm{w}'\right\rangle\\ &\le \frac{1}{n}\left(\sqrt{2H\ell(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{z}^i)}+\sqrt{2H\ell(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{z}')}\right)\left\|\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)})-\mathrm{u}(\mathrm{w},\mathcal{S})\right\|\\ &\quad+\frac{H}{n}\left\|\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)})-\mathrm{u}(\mathrm{w},\mathcal{S})\right\|^2+\lambda\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S})\|\,\|\mathrm{w}-\mathrm{w}'\| \qquad (\ell \text{ is } H\text{-smooth})\end{aligned}$$

Rearrange gives us,

$$\left\|\mathrm{u}(\mathrm{w},\mathcal{S})-\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)})\right\| \leq \frac{1}{2\lambda n-H}\left(\sqrt{2H\ell(\mathrm{u}(\mathrm{w},\mathcal{S}),\mathrm{z}^i)}+\sqrt{2H\ell(\mathrm{u}(\mathrm{w}',\mathcal{S}^{(i)}),\mathrm{z}')}\right) + \frac{\lambda n}{2\lambda n-H}\left\|\mathrm{w}-\mathrm{w}'\right\| \tag{9}$$

Plug in w_{T+1} and w'_{T+1} gives us that

$$\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\| \leq \frac{2\sqrt{2HM}}{2\lambda n-H}+\frac{n}{2\lambda n-H}\frac{4\sqrt{2HM}}{(m+1)}$$

□

If we apply Lemma 4.1 and Lemma C.1 with Theorem 2.2, we have the following theorem.

Theorem C.2. The following holds for Algorithm 1 with step-size $\gamma \leq \frac{1}{\lambda}$ on a given meta-sample $\mathbf{S}$, and RERM for task-specific learning (i.e., Option 1 for Algorithm 2), for all $T \geq 1$:

1. For convex, M-bounded, and G-Lipschitz loss functions, with probability at least $1-\delta$
$$L(\mathcal{A}(\mathbf{S}),\mu) \lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\frac{G^2}{\lambda m}\log(m)\log(1/\delta)+\frac{M}{\sqrt{m}}\sqrt{\log(1/\delta)}+\frac{G^2}{\lambda n}.$$
2. For convex, M-bounded, and H-smooth loss functions ($H \leq \lambda$), with probability at least $1-\delta$
$$L(\mathcal{A}(\mathbf{S}),\mu) \lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\frac{HM}{(m+1)\lambda}\log(m)\log(1/\delta)+\frac{M}{\sqrt{m}}\sqrt{\log(1/\delta)}+\frac{HM}{\lambda n}.$$

Proof of Theorem C.2. We slightly abuse the notation, at iteration t, define $\mathrm{w}_t = \mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t = \mathcal{A}(\mathbf{S}^{(j)})$, $\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}) = \mathcal{A}(\mathbf{S})(\mathcal{S})$, $\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}) = \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$. Apply Lemma 4.1 gives us that

$$\begin{aligned}
\left|L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{S})-L(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}),\mathcal{S})\right| &= \left|L(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S})-L(\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}),\mathcal{S})\right| \\
&\leq G\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S})\right\| && (G\text{-Lipschitz}) \\
&\leq \frac{G}{2}\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\| && (\text{Equation (3)}) \\
&\leq \frac{G^2}{\lambda m}
\end{aligned}$$

Apply Lemma C.1 gives us that

$$\begin{aligned}
\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}^{(i)}),\mathrm{z})\right| &= \left|\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})-\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&\leq G\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}^{(i)})\right\| \\
&\leq \frac{4G^2}{\lambda n}
\end{aligned}$$

Apply Theorem 2.2 with $\beta' = \frac{G^2}{\lambda m}$, $\beta = \frac{4G^2}{\lambda n}$ achieves the results.

Similarly, if the loss is M-bounded, convex, non-negative and H-smooth, we have

$$\begin{aligned}
&\left|L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{S})-L(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}),\mathcal{S})\right| \\
&= \left|L(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S})-L(\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}),\mathcal{S})\right| \\
&\leq \sqrt{2HL(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S})}\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S})\right\|+\frac{H}{2}\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S})\right\|^2 \\
& && (\ell \text{ is } H\text{-smooth}) \\
&\leq \sqrt{2HM}\frac{1}{2}\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\|+\frac{H}{8}\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\|^2 && (\text{Equation (7)}) \\
&\leq \frac{4HM}{(m+1)\lambda}+\frac{4H^2M}{(m+1)^2\lambda^2} \\
&\leq \frac{8HM}{(m+1)\lambda} && (\lambda \geq H)
\end{aligned}$$

https://doi.org/10.52202/079017-2661

Apply Lemma C.1 gives us that

$$\begin{aligned}
&\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&=\left|\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})-\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&\leq\sqrt{2H\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})}\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}^{(i)})\right\|+\frac{H}{2}\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}^{(i)})\right\|^2 \\
&\leq\frac{8HM}{\lambda n}+\frac{16H^2M}{\lambda^2n^2} \\
&\leq\frac{24HM}{\lambda n} && (\lambda\geq H)
\end{aligned}$$

Apply Theorem 2.2 with $\beta'=\frac{8HM}{(m+1)\lambda}$, $\beta=\frac{24HM}{\lambda n}$ achieves the results.

□

Theorem 4.2. The following holds for Algorithm 1 with step-size $\gamma\leq\frac{1}{\lambda}$ on a given meta-sample $\mathbf{S}$, and RERM for task-specific learning (i.e., Option 1 for Algorithm 2), for all $T\geq1$:

1. For convex, M-bounded, and G-Lipschitz loss functions, with probability at least $1-\delta$

$$L(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\left(\frac{G^2}{\lambda n}+\frac{G^2}{\lambda m}\right)\log(mn)\log(1/\delta)+\frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

2. For convex, M-bounded, and H-smooth loss functions ($H\leq\lambda$), with probability at least $1-\delta$

$$L(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\left(\frac{HM}{(2n-1)\lambda}+\frac{HM}{(m+1)\lambda}\right)\log(mn)\log(1/\delta)+\frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

Proof of Theorem 4.2. We slightly abuse the notation, at iteration t, define $\mathrm{w}_t=\mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t=\mathcal{A}(\mathbf{S}^{(j)})$, $\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})=\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})=\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$. For ℓ to be convex and G-Lipschitz, applying Lemma 4.1 gives us that

$$\begin{aligned}
\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}),\mathrm{z})\right|&=\left|\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})-\ell(\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&\leq G\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\| \\
&\leq\frac{G^2}{\lambda m}+\frac{2G^2}{\lambda n}.
\end{aligned}$$

Further apply Theorem 3.1 gives us the result. For ℓ to be convex, M-bounded and H-smooth, we have

$$\begin{aligned}
&\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&=\left|\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})-\ell(\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&\leq\sqrt{2H\ell(\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})}\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|+\frac{H}{2}\left\|\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|^2 \\
&\leq\sqrt{2HM}\left(\frac{2\sqrt{2HM}}{2\lambda n-H}+\frac{n}{2\lambda n-H}\frac{4\sqrt{2HM}}{(m+1)}\right)+\frac{H}{2}\left(\frac{2\sqrt{2HM}}{2\lambda n-H}+\frac{n}{2\lambda n-H}\frac{4\sqrt{2HM}}{(m+1)}\right)^2 \\
&\leq\frac{4HM}{(2n-1)\lambda}+\frac{8HM}{(m+1)\lambda}+\frac{8H^2M}{(2n-1)^2\lambda^2}+\frac{16H^2M}{(m+1)^2\lambda^2} \\
&\leq\frac{12HM}{(2n-1)\lambda}+\frac{24HM}{(m+1)\lambda} && (H\leq\lambda)
\end{aligned}$$

Apply Theorem 3.1 gives the results.

□

Lemma 4.3. Assume that the loss function is convex, G-Lipschitz and H-smooth. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then the following holds for Algorithm 1 with GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta \leq \frac{2}{H+2\lambda}$, for all $T \geq 1$ as long as we set $\gamma \leq \frac{1}{\lambda T}$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]} \left\| \mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}) \right\| \leq \frac{4eG}{\lambda m} + \frac{2G}{\lambda n}.$$

Proof of Lemma 4.3. We slightly abuse the notation, at iteration t, define $\mathrm{w}_t = \mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t = \mathcal{A}(\mathbf{S}^{(j)})$, $\mathrm{u}(\mathrm{w}_{T+1}, \mathcal{S}) = \mathcal{A}(\mathbf{S})(\mathcal{S})$, $\mathrm{u}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)}) = \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$. Recall that $F_{\mathcal{S}}(\mathrm{u}, \mathrm{w}) = L(\mathrm{u}, \mathcal{S}) + \frac{\lambda}{2}\|\mathrm{u} - \mathrm{w}\|^2$. If ℓ is convex, then $F_{\mathcal{S}}(\mathrm{u}, \mathrm{w})$ is λ-strongly-convex w.r.t u. If ℓ is H-smooth, then

$$\langle \nabla L(\mathrm{u}, \mathcal{S}) - \nabla L(\mathrm{v}, \mathcal{S}), \mathrm{u} - \mathrm{v} \rangle \geq \frac{1}{H} \|\nabla L(\mathrm{u}, \mathcal{S}) - \nabla L(\mathrm{v}, \mathcal{S})\|^2$$

Given $\mathcal{S}$, $\mathcal{S}'$, for any w, w', we have

$$\begin{aligned}
&\left\| \mathrm{u}^{(k+1)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}', \mathcal{S}') \right\| \\
&\leq \Bigg\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') - \eta \Big(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) + \lambda \left(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{w} \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Projection is non-expansive)} \\
&\qquad - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}'), \mathcal{S}') - \lambda \left(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') - \mathrm{w}' \right) \Big) \Bigg\| \\
&\leq (1 - \eta\lambda) \left\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') \right\| + \eta\lambda \|\mathrm{w} - \mathrm{w}'\| + 2\eta G
\end{aligned}$$

Given $\mathcal{S}$, for any w, w', we have

$$\begin{aligned}
&\left\| \mathrm{u}^{(k+1)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}', \mathcal{S}) \right\| \\
&\leq \Bigg\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) \qquad \text{(Projection is non-expansive)} \\
&\qquad - \eta \left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) + \lambda \Big(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{w} \Big) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}), \mathcal{S}) - \lambda \Big(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) - \mathrm{w}' \Big) \right) \Bigg\| \\
&= \Bigg\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) + \lambda\eta(\mathrm{w} - \mathrm{w}') \\
&\qquad - \eta \left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}), \mathcal{S}) + \lambda \Big(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) \Big) \right) \Bigg\| \\
&\leq \lambda\eta \|\mathrm{w} - \mathrm{w}'\| + \Bigg(\Bigg\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) \\
&\qquad - \eta \left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}), \mathcal{S}) + \lambda \Big(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) \Big) \right) \Bigg\|^2 \Bigg)^{1/2} \\
&\leq \lambda\eta \|\mathrm{w} - \mathrm{w}'\| + \Bigg((1 - \lambda\eta)^2 \left\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}) \right\|^2 \\
&\qquad + \left(\eta^2 - \frac{2\eta(1 - \eta\lambda)}{H} \right) \left\| \nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}), \mathcal{S}) \right\|^2 \Bigg)^{1/2} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (L \text{ is } H\text{-smooth}, \eta \leq \tfrac{2}{H+2\lambda})
\end{aligned}$$

https://doi.org/10.52202/079017-2661

$$\leq (1-\lambda\eta)\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S})\right\|+\lambda\eta\left\|\mathrm{w}-\mathrm{w}'\right\| \tag{10}$$

Combine the above two cases gives us that

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\left\|\mathrm{u}^{(k+1)}(\mathrm{w},\mathcal{S}_j)-\mathrm{u}^{(k+1)}(\mathrm{w}',\mathcal{S}_j')\right\| \\
&\leq \frac{1}{m}\sum_{j\neq i}^{m}\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}_j)-\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}_j')\right\|+\frac{1}{m}\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}_i)-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}_i')\right\| \\
&\leq (1-\lambda\eta)\frac{1}{m}\sum_{j=1}^{m}\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}_j)-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}_j')\right\|+\lambda\eta\left\|\mathrm{w}-\mathrm{w}'\right\|+\frac{2\eta G}{m}
\end{aligned}$$

Given $\mathrm{w}_t, \mathrm{w}_t'$, when $k=1$, $\mathrm{u}^{(1)}(\mathrm{w}_t,\mathcal{S}_j')=\mathrm{w}_t$, $\mathrm{u}^{(1)}(\mathrm{w}_t',\mathcal{S}_j)=\mathrm{u}^{(1)}(\mathrm{w}_t',\mathcal{S}_j')=\mathrm{w}_t'$. Telescoping gives us that

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m}\left\|\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S}_j)-\mathrm{u}^{(k)}(\mathrm{w}_t',\mathcal{S}_j')\right\| &\leq \left(1+(1-\lambda\eta)^{k-1}\right)\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{2G}{\lambda m} \\
&\leq 2\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{2G}{\lambda m}
\end{aligned} \tag{11}$$

Finally, at iteration t, we have

$$\begin{aligned}
&\left\|\mathrm{w}_{t+1}-\mathrm{w}_{t+1}'\right\| \\
&\leq \left\|\mathrm{w}_t-\gamma\lambda\left(\mathrm{w}_t-\frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S}_j)\right)-\mathrm{w}_t'+\gamma\lambda\left(\mathrm{w}_t'-\frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_t',\mathcal{S}_j')\right)\right\| \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Projection is non-expansive)} \\
&= \left\|(1-\gamma\lambda)(\mathrm{w}_t-\mathrm{w}_t')+\gamma\lambda\frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\left(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S}_j)-\mathrm{u}^{(k)}(\mathrm{w}_t',\mathcal{S}_j')\right)\right\| \\
&\leq (1-\gamma\lambda)\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\gamma\lambda\left(2\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{2G}{\lambda m}\right) \qquad\qquad \text{(Equation (11))} \\
&= (1+\gamma\lambda)\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{2\gamma G}{m}
\end{aligned}$$

Note that $\mathrm{w}_1=\mathrm{w}_1'=0$. Choosing $\gamma\leq\frac{1}{\lambda T}$ and telescoping gives us that

$$\left\|\mathrm{w}_{T+1}-\mathrm{w}_{T+1}'\right\|\leq\left(1+\frac{1}{T}\right)\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{2G}{m\lambda T}\leq\frac{2eG}{\lambda m} \tag{12}$$

where the inequality holds because $(1+\frac{1}{T})^T\leq e$. Similarly, we have

$$\begin{aligned}
&\left\|\mathrm{u}^{(k+1)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k+1)}(\mathrm{w}',\mathcal{S}^{(i)})\right\| \\
&\leq \Bigg\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)})-\eta\Big(\nabla L(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}),\mathcal{S})+\lambda\left(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{w}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Projection is non-expansive)} \\
&\qquad -\nabla L(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S}^{(i)})-\lambda\left(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)})-\mathrm{w}'\right)\Big)\Bigg\| \\
&\leq \eta\lambda\left\|\mathrm{w}-\mathrm{w}'\right\|+\Bigg\|(1-\eta\lambda)\left(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)})\right) \\
&\qquad +\eta\left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S})\right)
\end{aligned}$$

$$
\begin{aligned}
&+\eta\left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S}^{(i)})\right)\Bigg\|\\
&\leq \eta\lambda\left\|\mathrm{w}-\mathrm{w}'\right\|+\frac{2G\eta}{n}+\Bigg(\Bigg\|(1-\eta\lambda)\left(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)})\right)\\
&\quad+\eta\left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)}),\mathcal{S})\right)\Bigg\|^2\Bigg)^{1/2}\\
&\leq \eta\lambda\left\|\mathrm{w}-\mathrm{w}'\right\|+\frac{2G\eta}{n}+\Bigg((1-\lambda\eta)^2\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)})\right\|^2\\
&\quad+\left(\eta^2-\frac{2\eta(1-\eta\lambda)}{H}\right)\left\|\nabla L(\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}),\mathcal{S}^{(i)})\right\|^2\Bigg)^{1/2}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad(L\text{ is }H\text{-smooth},\ \eta\leq\tfrac{2}{H+2\lambda})\\
&\leq (1-\lambda\eta)\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}',\mathcal{S}^{(i)})\right\|+\eta\lambda\left\|\mathrm{w}-\mathrm{w}'\right\|+\frac{2G\eta}{n} \qquad (13)
\end{aligned}
$$

Therefore we have $\forall k\in[K-1]$,

$$
\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}^{(k+1)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|\leq\left(1+(1-\lambda\eta)^{k-1}\right)\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\|+\frac{2G}{\lambda n}.
$$

As $\mathrm{u}^{(1)}(\mathrm{w}_{T+1},\mathcal{S})=\mathrm{w}_{T+1}$, $\mathrm{u}^{(1)}(\mathrm{w}'_{T+1},\mathcal{S})=\mathrm{w}'_{T+1}$, plug in Equation (12), we have

$$
\begin{aligned}
&\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|\\
&\leq\min\left(2,1+\frac{1}{\lambda\eta K}\right)\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\|+\frac{2G}{\lambda n}\\
&\leq\min\left(2,1+\frac{1}{\lambda\eta K}\right)\frac{2eG}{\lambda m}+\frac{2G}{\lambda n}\\
&\leq\frac{4eG}{\lambda m}+\frac{2G}{\lambda n}
\end{aligned}
$$

□

The following theorem can be derived via Theorem 2.2 and Lemma 4.3.

Theorem C.3. Consider a meta-learning problem with convex, M-bounded, G-Lipschitz and H-smooth loss function. Then, after T iterations of Algorithm 1 with $\gamma\leq\frac{1}{\lambda T}$ on a given meta-sample $\mathbf{S}$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta\leq\frac{2}{H+2\lambda}$, we have with probability at least $1-\delta$,

$$
L(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\frac{G^2}{\lambda m}\log(m)\log(1/\delta)+\frac{M}{\sqrt{m}}\sqrt{\log(1/\delta)}+\frac{G^2}{\lambda n}.
$$

Proof of Theorem C.3. We slightly abuse the notation, at iteration t, define $\mathrm{w}_t=\mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t=\mathcal{A}(\mathbf{S}^{(j)})$, $\mathrm{u}^{(K)}(\mathrm{w}_{T+1},\mathcal{S})=\mathcal{A}(\mathbf{S})(\mathcal{S})$, $\mathrm{u}^{(K)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})=\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$. If the loss is M-bounded,

https://doi.org/10.52202/079017-2661

convex and G-Lipschitz, apply Equation (10) gives us

$$\begin{aligned}
&\left|L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{S})-L(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}),\mathcal{S})\right| \\
&=\left|L(\mathrm{u}^{(K)}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S})-L(\mathrm{u}^{(K)}(\mathrm{w}'_{T+1},\mathcal{S}),\mathcal{S})\right| \\
&\le G\left\|\mathrm{u}^{(K)}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}^{(K)}(\mathrm{w}'_{T+1},\mathcal{S})\right\| && (G\text{-Lipschitz}) \\
&\le G\left((1-\lambda\eta)\left\|\mathrm{u}^{(K)}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}^{(K)}(\mathrm{w}'_{T+1},\mathcal{S})\right\|+\lambda\eta\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\|\right) && (\text{Equation (10)}) \\
&\le G\left\|\mathrm{w}_{T+1}-\mathrm{w}'_{T+1}\right\| \\
&\le \frac{2eG^2}{\lambda m} && (\text{Equation (12)})
\end{aligned}$$

Given $\mathcal{S},\mathcal{S}^{(i)}$. For any w, by Equation (13), for all $k\in[K-1]$, we have

$$\left\|\mathrm{u}^{(k+1)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k+1)}(\mathrm{w},\mathcal{S}^{(i)})\right\|\le\frac{2G\eta}{n}+(1-\lambda\eta)\left\|\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w},\mathcal{S}^{(i)})\right\|\le\frac{2G}{\lambda n} \quad (\text{Telescope})$$

Therefore, we have

$$\begin{aligned}
&\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&=\left|\ell\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z}\right)-\ell\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}^{(i)}),\mathrm{z}\right)\right| \\
&\le G\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}^{(i)})\right\| \\
&\le\frac{2G^2}{\lambda n}
\end{aligned}$$

Apply Theorem 2.2 with $\beta'=\frac{2eG^2}{\lambda m}$, with $\beta=\frac{2G^2}{\lambda n}$ gives us the result.

□

Theorem 4.4. Assume that the loss function is convex, M-bounded, G-Lipschitz and H-smooth. Suppose we run Algorithm 1 for T iterations with $\gamma\le\frac{1}{\lambda T}$ on a given meta-sample $\mathbf{S}$, and GD for task-specific learning (Option 2, Algorithm 2) with $\eta\le\frac{2}{H+2\lambda}$. Then, with probability at least $1-\delta$,

$$L(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\left(\frac{G^2}{\lambda m}+\frac{G^2}{\lambda n}\right)\log(mn)\log(1/\delta)+\frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

Proof of Theorem 4.4. We denote $\mathrm{w}_t=\mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t=\mathcal{A}(\mathbf{S}^{(j)})$, $\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S})=\mathcal{A}(\mathbf{S})(\mathcal{S})$, $\mathrm{u}(\mathrm{w}_{T+1},\mathcal{S}^{(i)})=\mathcal{A}(\mathbf{S})(\mathcal{S}^{(i)})$. For ℓ to be convex and G-Lipschitz, applying Lemma 4.3 gives us that

$$\begin{aligned}
&\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}),\mathrm{z})\right| \\
&=\left|\ell\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z}\right)-\ell\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathrm{z}\right)\right| \\
&\le G\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\| \\
&\le\min\left(2,1+\frac{1}{\lambda\eta K}\right)\frac{2eG^2}{\lambda m}+\frac{2G^2}{\lambda n}.
\end{aligned}$$

Further apply Theorem 3.1 gives us the result. □

D Missing Proofs of Section 4.2

We start with a proposition that provide some equivalent characterizations of weak convexity.

Proposition D.1 (Proposition 2.1 in Davis and Grimmer [2019])**.** Suppose $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a closed function and $\rho > 0$, then the following are equivalent:

1. For any $\mathrm{w}_1 \in \mathbb{R}^d$, $f(\cdot) + \frac{\rho}{2} \|\cdot - \mathrm{w}_1\|$ is convex.
2. For any $\mathrm{w}_1, \mathrm{w}_2 \in \mathbb{R}^d$ with $g(\mathrm{w}_1) \in \partial f(\mathrm{w}_1)$, we have
$$f(\mathrm{w}_2) \geq f(\mathrm{w}_1) + \langle g(\mathrm{w}_1), \mathrm{w}_2 - \mathrm{w}_1 \rangle - \frac{\rho}{2} \|\mathrm{w}_2 - \mathrm{w}_1\|^2 .$$
3. For any $\mathrm{w}_1, \mathrm{w}_2 \in \mathbb{R}^d$ and $\lambda > 0$,
$$f(\lambda \mathrm{w}_1 + (1-\lambda)\mathrm{w}_2) \leq \lambda f(\mathrm{w}_1) + (1-\lambda) f(\mathrm{w}_2) + \frac{\rho\lambda(1-\lambda)}{2} \|\mathrm{w}_1 - \mathrm{w}_2\|^2 .$$

Lemma D.2. [Bassily et al. [2020]] Given $\mathcal{S}$ and $\mathcal{S}^{(i)}$, for a fixed w, consider $\mathrm{u}(\mathrm{w}, \mathcal{S})$ and $\mathrm{u}(\mathrm{w}, \mathcal{S}^{(i)})$ are achieved via Algo. 1 with gradient descent for K iterations. Then if ℓ is convex and G-Lipschitz, we have $\sup_{i \in [n]} \left\| \frac{1}{K} \sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \frac{1}{K} \sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}^{(i)}) \right\| \leq \frac{4GK\eta}{n} + 4G\eta\sqrt{K}$.

Below we provide our key Lemma D.3 for the stability analysis.

Lemma D.3. Consider a meta-learning problem with ρ-weakly convex and G-Lipschitz loss function. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then, after T iterations of Algorithm 1 with $\gamma \leq \frac{1}{\lambda T}$, $\lambda \geq 2\rho$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta \leq \frac{1}{\lambda}$,
$$\sup_{\mathbf{S}, j \in [m]} \left\| \mathrm{w}_{T+1} - \mathrm{w}'_{T+1} \right\| \leq 2eG\sqrt{\frac{\eta}{\lambda}} + \frac{2eG}{\lambda m}.$$

Proof of Lemma D.3. If ℓ is ρ-weakly convex, then from Proposition D.1 we have that
$$\langle \nabla \ell(\mathrm{u}) - \nabla \ell(\mathrm{v}), \mathrm{u} - \mathrm{v} \rangle \geq -\rho \|\mathrm{u} - \mathrm{v}\|^2 , \forall \mathrm{u}, \mathrm{v} \in \mathbb{R}^d$$

Given $\mathcal{S}$ and $\mathcal{S}'$, for any w and w′, we have

$$\begin{aligned}
\forall k \in [K-1], \quad & \left\| \mathrm{u}^{(k+1)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}', \mathcal{S}') \right\| \\
&\leq \Bigg\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') - \eta \Bigg(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}), \mathcal{S}) + \lambda \left(\mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{w} \right) \\
&\qquad \text{(Projection is non-expansive)} \\
&\qquad - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}'), \mathcal{S}) - \lambda \left(\mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') - \mathrm{w}' \right) \Bigg) \Bigg\| \\
&= (1 - \eta\lambda) \left\| \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') \right\| + \eta\lambda \|\mathrm{w} - \mathrm{w}'\| + 2\eta G \\
&\leq \left(1 + (1 - \eta\lambda)^k\right) \|\mathrm{w} - \mathrm{w}'\| + \frac{2G}{\lambda} \\
&\qquad \text{(Telescope, } \mathrm{u}^{(1)}(\mathrm{w}, \mathcal{S}) = \mathrm{w}, \mathrm{u}^{(1)}(\mathrm{w}', \mathcal{S}) = \mathrm{w}'.)
\end{aligned}$$

And therefore
$$\left\| \frac{1}{K} \sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}, \mathcal{S}) - \frac{1}{K} \sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}', \mathcal{S}') \right\| \leq \min \left(2, 1 + \frac{1}{\lambda\eta K} \right) \|\mathrm{w} - \mathrm{w}'\| + \frac{2G}{\lambda}$$

We now focus on the situation where we give $\mathcal{S}$ and $\mathcal{S}'$ with a fix w. For simplicity, we define $\delta_k = \left\| \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}) \right\|$. Note that $\delta_1 = \|\mathrm{w}_t - \mathrm{w}'_t\|$. We have
$$\delta_{k+1} = \left\| \mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}'_t, \mathcal{S}) \right\|$$

https://doi.org/10.52202/079017-2661

$$
\begin{aligned}
&\leq \Bigg\| \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}) - \eta \Bigg(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}), \mathcal{S}) + \lambda \left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{w}_t \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Projection is non-expansive)} \\
&\qquad - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}), \mathcal{S}) - \lambda \left(\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}) - \mathrm{w}'_t \right) \Bigg) \Bigg\| \\
&\leq \lambda\eta \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| + \Bigg\| (1 - \eta\lambda) \left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}) \right) \\
&\qquad - \eta \left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}), \mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}), \mathcal{S}) \right) \Bigg\| \\
&= \lambda\eta \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| + \Delta_k
\end{aligned}
$$

where we define

$$
\Delta_k = \left\| (1 - \eta\lambda) \left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}) \right) - \eta \left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}), \mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}), \mathcal{S}) \right) \right\|.
$$

We have that

$$
\begin{aligned}
\Delta_k^2 &= (1 - \eta\lambda)^2 \delta_k^2 + 4\eta^2 G^2 \\
&\quad - 2\eta(1 - \eta\lambda) \left\langle \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}), \nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}), \mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}), \mathcal{S}) \right\rangle \\
&\leq (1 - \eta\lambda)^2 \delta_k^2 + 4\eta^2 G^2 + 2\eta(1 - \eta\lambda)\rho\delta_k^2 && (\ell \text{ is } \rho\text{-weakly convex}) \\
&\leq (1 - \eta\lambda)\delta_k^2 + 4\eta^2 G^2 && (\eta \leq \tfrac{1}{\lambda}, \lambda \geq 2\rho)
\end{aligned}
$$

Therefore, we have

$$
\delta_{k+1}^2 + \lambda^2\eta^2 \left\| \mathrm{w}_t - \mathrm{w}'_t \right\|^2 - 2\lambda\eta\delta_{k+1} \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| \leq \Delta_k^2 \leq (1 - \eta\lambda)\delta_k^2 + 4\eta^2 G^2 \tag{14}
$$

Rearrange it gives us that

$$
\frac{\delta_{k+1}^2}{(1 - \eta\lambda)^{k+1}} + \frac{\lambda^2\eta^2 \left\| \mathrm{w}_t - \mathrm{w}'_t \right\|^2}{(1 - \eta\lambda)^{k+1}} - \frac{2\lambda\eta \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| \delta_{k+1}}{(1 - \eta\lambda)^{k+1}} \leq \frac{\delta_k^2}{(1 - \eta\lambda)^k} + \frac{4\eta^2 G^2}{(1 - \eta\lambda)^{k+1}}
$$

Telescoping from $k = 1$ to K gives us that

$$
\frac{\delta_{K+1}^2}{(1 - \eta\lambda)^{K+1}} + \sum_{k=1}^{K} \frac{\lambda^2\eta^2 \left\| \mathrm{w}_t - \mathrm{w}'_t \right\|^2}{(1 - \eta\lambda)^{k+1}} \leq \sum_{k=1}^{K} \frac{2\lambda\eta \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| \delta_{k+1}}{(1 - \eta\lambda)^{k+1}} + \sum_{k=1}^{K} \frac{4\eta^2 G^2}{(1 - \eta\lambda)^{k+1}}
$$

Thus

$$
\begin{aligned}
&\delta_{K+1}^2 + \lambda^2\eta^2 \left\| \mathrm{w}_t - \mathrm{w}'_t \right\|^2 \sum_{k=1}^{K} (1 - \eta\lambda)^{K-k} - 2\lambda\eta \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| \delta_{K+1} \\
&\leq \frac{4\eta G^2}{\lambda} + 2\lambda\eta \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| \sum_{k=1}^{K-1} \delta_{k+1} (1 - \eta\lambda)^{K-k} \\
&\leq \frac{4\eta G^2}{\lambda} + 2\lambda\eta(1 - \eta\lambda) \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| \sum_{k=1}^{K} \delta_k (1 - \eta\lambda)^{K-k}
\end{aligned} \tag{15}
$$

Now we start proving the following bound by induction:

$$
\delta_K \leq 2 \left\| \mathrm{w}_t - \mathrm{w}'_t \right\| + 2G\sqrt{\frac{\eta}{\lambda}}.
$$

This claim holds when $k = 1$. For the inductive step, we assume it holds for some $k \in [K]$ and prove the result for $k + 1$. We consider the following two cases. If $\delta_{k+1} \le \max_{s\in[k]} \delta_k$, induction automatically holds. Otherwise, $\delta_{k+1} > \max_{s\in[k]} \delta_s$. Applying Equation (15) gives us that

$$\begin{aligned}
&\delta_{k+1}^2 + \lambda^2\eta^2 \left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2 \sum_{j=1}^{k} (1-\eta\lambda)^{k-j} - 2\lambda\eta \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \delta_{k+1} \\
&\le \frac{4\eta G^2}{\lambda} + 2\lambda\eta(1-\eta\lambda) \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \sum_{j=1}^{k} \delta_k (1-\eta\lambda)^{k-j} \\
&\le \frac{4\eta G^2}{\lambda} + 2\lambda\eta(1-\eta\lambda) \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \delta_{k+1} \sum_{j=1}^{k} (1-\eta\lambda)^{k-j}
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\delta_{k+1}^2 + \lambda^2\eta^2 \left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2 \sum_{j=1}^{k} (1-\eta\lambda)^{k-j} \\
&\le \frac{4\eta G^2}{\lambda} + 2\lambda\eta \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \delta_{k+1} \left(1 + (1-\eta\lambda)\sum_{j=1}^{k} (1-\eta\lambda)^{k-j}\right)
\end{aligned}$$

Rearrange gives us that

$$\begin{aligned}
&\left(\delta_{k+1} - \lambda\eta \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \left(1 + (1-\eta\lambda)\sum_{j=1}^{k} (1-\eta\lambda)^{k-j}\right)\right)^2 \\
&\le \left(\lambda\eta \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \left(1 + (1-\eta\lambda)\sum_{j=1}^{k} (1-\eta\lambda)^{k-j}\right)\right)^2 + \frac{4\eta G^2}{\lambda} \\
&\quad - \lambda\eta \left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2 \left(1 - (1-\eta\lambda)^{k+1}\right)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\forall k \in [K-1] \quad &\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}'_t, \mathcal{S})\right\| \\
&\le 2G\sqrt{\frac{\eta}{\lambda}} + 2\left(\lambda\eta \left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \left(1 + (1-\eta\lambda)\sum_{j=1}^{k} (1-\eta\lambda)^{k-j}\right)\right) \\
&\le 2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + 2G\sqrt{\frac{\eta}{\lambda}}
\end{aligned}$$

And therefore

$$\left\|\frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) - \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S})\right\| \le 2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + 2G\sqrt{\frac{\eta}{\lambda}} \tag{16}$$

As a result,

$$\begin{aligned}
&\left\|\mathrm{w}_{t+1} - \mathrm{w}'_{t+1}\right\| \\
&\le \left\|\mathrm{w}_t - \gamma\lambda\left(\mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j)\right) - \mathrm{w}'_t + \gamma\lambda\left(\mathrm{w}'_t - \frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}'_j)\right)\right\| \\
&\hspace{25em}\text{(Projection is non-expansive)} \\
&= (1-\gamma\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \gamma\lambda\left\|\frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \frac{1}{m}\sum_{j=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}'_j)\right\|
\end{aligned}$$

https://doi.org/10.52202/079017-2661

$$\leq (1-\gamma\lambda)\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{m-1}{m}\gamma\lambda\left(2\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+2G\sqrt{\frac{\eta}{\lambda}}\right)+\frac{\gamma\lambda}{m}\left(2\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+\frac{2G}{\lambda}\right)$$
$$\leq (1+\gamma\lambda)\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+2G\gamma\sqrt{\eta\lambda}+\frac{2G\gamma}{m}$$

Telescoping gives us that

$$\left\|\mathrm{w}_{T+1}-\mathrm{w}_{T+1}'\right\|\leq(1+\gamma\lambda)^T\left(2G\sqrt{\frac{\eta}{\lambda}}+\frac{2G}{\lambda m}\right)$$

Choosing $\gamma\leq\frac{1}{\lambda T}$ gives us that

$$\left\|\mathrm{w}_{T+1}-\mathrm{w}_{T+1}'\right\|\leq(1+\frac{1}{T})^T\left(2G\sqrt{\frac{\eta}{\lambda}}+\frac{2G}{\lambda m}\right)\leq 2eG\sqrt{\frac{\eta}{\lambda}}+\frac{2eG}{\lambda m}$$

We remark that if we consider convex and non-smooth loss function by setting $\rho=0$, then follow a similar argument, Equation (14) can be replaced by

$$\delta_{k+1}^2+\lambda^2\eta^2\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|^2-2\lambda\eta\delta_{k+1}\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|\leq\Delta_k^2\leq(1-\eta\lambda)^2\delta_k^2+4\eta^2G^2$$

And therefore Equation (16) can be replaced by

$$\begin{aligned}
&\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t,\mathcal{S})-\mathrm{u}^{(k+1)}(\mathrm{w}_t',\mathcal{S})\right\|\\
&\leq 2G\sqrt{\frac{\eta}{\lambda}}+2\left(\lambda\eta\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|\left(1+(1-\eta\lambda)^2\sum_{j=1}^{k}(1-\eta\lambda)^{2k-2j}\right)\right)\\
&\leq\frac{2}{2-\eta\lambda}\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+2G\sqrt{\frac{\eta}{\lambda}}\\
&\leq 2\left\|\mathrm{w}_t-\mathrm{w}_t'\right\|+2G\sqrt{\frac{\eta}{\lambda}} && (\eta\lambda\leq 1)
\end{aligned}$$

and the rest follows.

□

Theorem D.4. Consider a meta-learning problem with ρ-weakly convex, M-bounded, G-Lipschitz loss function. Then, after T iterations of Algorithm 1 with $\gamma\leq\frac{1}{\lambda T}$, $\lambda\geq 2\rho$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta\leq\frac{1}{\lambda}$, we have with probability at least $1-\delta$,

$$L_(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\left(G^2\sqrt{\frac{\eta}{\lambda}}+\frac{G^2}{\lambda m}\right)\log(m)\log(1/\delta)+\frac{M}{\sqrt{m}}\sqrt{\log(1/\delta)}+\frac{G^2}{\lambda n}+G^2\eta\sqrt{K}.$$

By setting $\eta=\frac{1}{\lambda K}$, we have

$$L_(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\left(\frac{G^2}{\lambda\sqrt{K}}+\frac{G^2}{\lambda m}\right)\log(m)\log(1/\delta)+\frac{M}{\sqrt{m}}\sqrt{\log(1/\delta)}+\frac{G^2}{\lambda n}+\frac{G^2}{\lambda\sqrt{K}}.$$

Proof of Theorem D.4. If the loss is M-bounded and G-Lipschitz, apply Lemma D.3 gives us

$$\begin{aligned}
&\left|L(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathcal{S})-L(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}),\mathcal{S})\right|\\
&=\left|L\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S}\right)-L\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1}',\mathcal{S}),\mathcal{S}\right)\right|\\
&\leq G\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1}',\mathcal{S})\right\| && (G\text{-Lipschitz})\\
&\leq 2G\left\|\mathrm{w}_{T+1}-\mathrm{w}_{T+1}'\right\|+2G^2\sqrt{\frac{\eta}{\lambda}} && (\text{Equation (16)})\\
&\leq\left(4eG^2+2G^2\right)\sqrt{\frac{\eta}{\lambda}}+\frac{4eG^2}{\lambda m} && (\text{Lemma D.3})\\
&\leq\left(4eG^2+2G^2\right)\frac{1}{\lambda\sqrt{K}}+\frac{4eG^2}{\lambda m} && (\text{Set }\eta\leq\tfrac{1}{\lambda K})
\end{aligned}$$

On the other hand, applying Lemma D.2 gives us that

$$\begin{aligned}\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}^{(i)}),\mathrm{z})\right| &= \left|\ell(\mathrm{u}^{(K+1)}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z})-\ell(\mathrm{u}^{(K+1)}(\mathrm{w}_{T+1},\mathcal{S}^{(i)}),\mathrm{z})\right| \\ &\le G\left\|\mathrm{u}^{(K+1)}(\mathrm{w}_T,\mathcal{S})-\mathrm{u}^{(K+1)}(\mathrm{w}_T,\mathcal{S}^{(i)})\right\| \\ &\le \frac{4G^2K\eta}{n}+4G^2\eta\sqrt{K} \\ &\le \frac{4G^2}{\lambda n}+\frac{4G^2}{\lambda\sqrt{K}} \qquad (\text{Set } \eta\le\tfrac{1}{\lambda K})\end{aligned}$$

Plug back into Theorem 2.2 gives the result. □

Lemma 4.5. Assume that the loss function is ρ-weakly convex and G-Lipschitz. Let $\mathbf{S}$, $\mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}$, $\mathcal{S}^{(i)}$ the neighboring samples on a test task. Then the following holds for Algorithm 1 with $\lambda\ge 2\rho$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta\le\frac{1}{\lambda}$, for all $T\ge 1$ as long as we set $\gamma\le\frac{1}{\lambda T}$,

$$\sup_{\mathbf{S},\mathcal{S},j\in[m],i\in[n]}\left\|\mathcal{A}(\mathbf{S})(\mathcal{S})-\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})\right\|\le(8eG+2G)\sqrt{\frac{\eta}{\lambda}}+\frac{8eG}{\lambda m}+\frac{8G}{\lambda n}.$$

Proof of Lemma 4.5. We slightly abuse the notation, at outer iteration t, define $\mathrm{w}_t=\mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t=\mathcal{A}(\mathbf{S}^{(j)})$. Given w_t, at inner iteration k, define $\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S})=\mathcal{A}(\mathbf{S})(\mathcal{S})$, $\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)})=\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$. We now provide the upper bound on $\left\|\frac{1}{K}\sum_{k=1}^K\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^K\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|$. Recall that if ℓ is ρ-weakly convex, then we have

$$\langle\nabla\ell(\mathrm{u})-\nabla\ell(\mathrm{v}),\mathrm{u}-\mathrm{v}\rangle\ge-\rho\left\|\mathrm{u}-\mathrm{v}\right\|^2$$

We apply a similar procedure as Lemma D.3. For simplicity, we define $\delta_k=\left\|\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)})\right\|$. Note that $\delta_1=\|\mathrm{w}_t-\mathrm{w}'_t\|$. We have

$$\begin{aligned}\delta_{k+1} &= \left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t,\mathcal{S})-\mathrm{u}^{(k+1)}(\mathrm{w}'_t,\mathcal{S}^{(i)})\right\| \\ &= \Bigg\|\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)})-\eta\Bigg(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S}),\mathcal{S})+\lambda\left(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S})-\mathrm{w}_t\right) \\ &\quad -\nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)}),\mathcal{S}^{(i)})-\lambda\left(\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)})-\mathrm{w}'_t\right)\Bigg)\Bigg\| \\ &\le \lambda\eta\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\Bigg\|(1-\eta\lambda)\left(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)})\right) \\ &\quad -\eta\left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)}),\mathcal{S}^{(i)})\right)\Bigg\| \\ &= \lambda\eta\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\Delta_k\end{aligned}$$

where we define

$$\begin{aligned}\Delta_k=\Bigg\|&(1-\eta\lambda)\left(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)})\right) \\ &-\eta\left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t,\mathcal{S}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_t,\mathcal{S}^{(i)}),\mathcal{S}^{(i)})\right)\Bigg\|.\end{aligned}$$

We have that

$$\begin{aligned}\Delta_k^2 &= (1-\eta\lambda)^2\delta_k^2+4\eta^2G^2 \\ &\quad -2\eta(1-\eta\lambda)\left\langle\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S})-\nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathcal{S}^{(i)})\right\rangle\end{aligned}$$

https://doi.org/10.52202/079017-2661

$$
\begin{aligned}
&= (1-\eta\lambda)^2\delta_k^2 + 4\eta^2G^2 \\
&\quad - 2\eta(1-\eta\lambda)\Big\langle \mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}), \nabla L(\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathcal{S}) - \nabla L(\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathcal{S})\Big\rangle \\
&\quad - \frac{2\eta(1-\eta\lambda)}{n}\Big\langle \mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}) - \mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}), \nabla \ell(\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathrm{z}^i) - \nabla \ell(\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathrm{z}')\Big\rangle \\
&\le (1-\eta\lambda)^2\delta_k^2 + 4\eta^2G^2 + 2\eta(1-\eta\lambda)\rho\delta_k^2 + \frac{4G\eta(1-\eta\lambda)}{n}\delta_k \\
&\le (1-\eta\lambda)\delta_k^2 + 4\eta^2G^2 + \frac{4G\eta(1-\eta\lambda)}{n}\delta_k \qquad (\eta \le \tfrac{1}{\lambda}, \lambda \ge 2\rho)
\end{aligned}
$$

Therefore, we have

$$
\delta_{k+1}^2 + \lambda^2\eta^2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2 - 2\lambda\eta\delta_{k+1}\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| \le \Delta_k^2 \le (1-\eta\lambda)\delta_k^2 + 4\eta^2G^2 + \frac{4G\eta(1-\eta\lambda)}{n}\delta_k
$$

Rearrange it gives us that

$$
\begin{aligned}
&\frac{\delta_{k+1}^2}{(1-\eta\lambda)^{k+1}} + \frac{\lambda^2\eta^2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2}{(1-\eta\lambda)^{k+1}} - \frac{2\lambda\eta\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|\delta_{k+1}}{(1-\eta\lambda)^{k+1}} \\
&\qquad \le \frac{\delta_k^2}{(1-\eta\lambda)^k} + \frac{4\eta^2G^2}{(1-\eta\lambda)^{k+1}} + \frac{4G\eta(1-\eta\lambda)\delta_k}{n(1-\eta\lambda)^{k+1}}
\end{aligned}
$$

Telescoping from $k=1$ to K gives us that

$$
\begin{aligned}
&\frac{\delta_{K+1}^2}{(1-\eta\lambda)^{K+1}} + \sum_{k=1}^{K}\frac{\lambda^2\eta^2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2}{(1-\eta\lambda)^{k+1}} \\
&\qquad \le \sum_{k=1}^{K}\frac{2\lambda\eta\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|\delta_{k+1}}{(1-\eta\lambda)^{k+1}} + \sum_{k=1}^{K}\frac{4\eta^2G^2}{(1-\eta\lambda)^{k+1}} + \sum_{k=1}^{K}\frac{4G\eta(1-\eta\lambda)\delta_k}{n(1-\eta\lambda)^{k+1}}
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\delta_{K+1}^2 + \lambda^2\eta^2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2\sum_{k=1}^{K}(1-\eta\lambda)^{K-k} - 2\lambda\eta\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|\delta_{K+1} \\
&\le \frac{4\eta G^2}{\lambda} + 2\lambda\eta\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|\sum_{k=1}^{K-1}\delta_{k+1}(1-\eta\lambda)^{K-k} + \frac{4G\eta(1-\eta\lambda)}{n}\sum_{k=1}^{K}\delta_k(1-\eta\lambda)^{K-k} \\
&\le \frac{4\eta G^2}{\lambda} + \left(2\lambda\eta(1-\eta\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{4G\eta(1-\eta\lambda)}{n}\right)\sum_{k=1}^{K}\delta_k(1-\eta\lambda)^{K-k} \qquad (17)
\end{aligned}
$$

Now we start proving the following bound by induction:

$$
\delta_K \le 2G\sqrt{\frac{\eta}{\lambda}} + 4\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{8G(1-\eta\lambda)}{\lambda n} \qquad (18)
$$

This claim holds when $k=1$. For the inductive step, we assume it holds for some $k \in [K]$ and prove the result for $k+1$. We consider the following two cases. If $\delta_{k+1} \le \max_{s\in[k]}\delta_k$, induction automatically holds. Otherwise, $\delta_{k+1} > \max_{s\in[k]}\delta_s$. Applying Equation (17) gives us that

$$
\begin{aligned}
&\delta_{k+1}^2 + \lambda^2\eta^2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|^2\sum_{j=1}^{k}(1-\eta\lambda)^{k-j} - 2\lambda\eta\left\|\mathrm{w}_t - \mathrm{w}'_t\right\|\delta_{k+1} \\
&\le \frac{4\eta G^2}{\lambda} + \left(2\lambda\eta(1-\eta\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{4G\eta(1-\eta\lambda)}{n}\right)\delta_{k+1}\sum_{j=1}^{k}(1-\eta\lambda)^{k-j}
\end{aligned}
$$

which is equivalent to

$$\left(\delta_{k+1}-\left(2\lambda\eta\|\mathrm{w}_t-\mathrm{w}'_t\|+2\lambda\eta(1-\lambda\eta)\|\mathrm{w}_t-\mathrm{w}'_t\|\sum_{j=1}^{k}(1-\eta\lambda)^{k-j}+\frac{4G\eta(1-\eta\lambda)}{n}\sum_{j=1}^{k}(1-\lambda\eta)^{k-j}\right)\right)^2$$
$$\leq\frac{4\eta G^2}{\lambda}+\left(2\lambda\eta\|\mathrm{w}_t-\mathrm{w}'_t\|+2\lambda\eta(1-\lambda\eta)\|\mathrm{w}_t-\mathrm{w}'_t\|\sum_{j=1}^{k}(1-\eta\lambda)^{k-j}+\frac{4G\eta(1-\eta\lambda)}{n}\sum_{j=1}^{k}(1-\lambda\eta)^{k-j}\right)^2$$

Therefore, we have

$$\begin{aligned}\delta_{k+1}\leq&2G\sqrt{\frac{\eta}{\lambda}}+2\left(2\lambda\eta\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+2\lambda\eta(1-\lambda\eta)\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|\sum_{j=1}^{k}(1-\eta\lambda)^{k-j}\right.\\&\left.+\frac{4G\eta(1-\eta\lambda)}{n}\sum_{j=1}^{k}(1-\lambda\eta)^{k-j}\right)\\\leq&2G\sqrt{\frac{\eta}{\lambda}}+4\left\|\mathrm{w}_t-\mathrm{w}'_t\right\|+\frac{8G(1-\eta\lambda)}{\lambda n}\end{aligned}$$

Plug in Lemma D.3 gives us that

$$\delta_{k+1}\leq(8eG+2G)\sqrt{\frac{\eta}{\lambda}}+\frac{8eG}{\lambda m}+\frac{8G}{\lambda n}$$

Therefore we have

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|\leq(8eG+2G)\sqrt{\frac{\eta}{\lambda}}+\frac{8eG}{\lambda m}+\frac{8G}{\lambda n}$$

Moreover, setting $\eta=\frac{1}{\lambda K}$ gives us that

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|\leq\frac{8eG+2G}{\lambda\sqrt{K}}+\frac{8eG}{\lambda m}+\frac{8G}{\lambda n}$$

□

Theorem 4.6. Assume that the loss function is ρ-weakly convex, M-bounded, and G-Lipschitz. Suppose we run Algorithm 1 for T iterations with $\gamma\leq\frac{1}{\lambda T}$, $\lambda\geq2\rho$ on a meta-sample $\mathbf{S}$, and GD for task-specific learning (Option 2, Algorithm 2) with $\eta\leq\frac{1}{\lambda}$, Then, with probability at least $1-\delta$,

$$L(\mathcal{A}(\mathbf{S}),\mu)\lesssim L(\mathcal{A}(\mathbf{S}),\mathbf{S})+\left(G^2\sqrt{\frac{\eta}{\lambda}}+\frac{G^2}{\lambda m}+\frac{G^2}{\lambda n}\right)\log(mn)\log(1/\delta)+\frac{M\sqrt{\log(1/\delta)}}{\sqrt{mn}}.$$

Proof of Theorem 4.6. For ℓ to be G-Lipschitz, applying Lemma 4.1 gives us that

$$\begin{aligned}&\left|\ell(\mathcal{A}(\mathbf{S})(\mathcal{S}),\mathrm{z})-\ell(\mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)}),\mathrm{z})\right|\\&=\left|\ell\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S}),\mathrm{z}\right)-\ell\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)}),\mathrm{z}\right)\right|\\&\leq G\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1},\mathcal{S})-\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1},\mathcal{S}^{(i)})\right\|\\&\leq(8eG^2+2G^2)\sqrt{\frac{\eta}{\lambda}}+\frac{8eG^2}{\lambda m}+\frac{8G^2}{\lambda n}.\end{aligned}$$

Plug it back into Theorem 3.1 gives the result. □

https://doi.org/10.52202/079017-2661

Theorem D.5 (Restatement of Theorem 4.7)**.** Assume the loss ℓ is convex and G-Lipschitz. Define $\mathrm{u}_j^* = \operatorname{argmin}_{\mathrm{u}} L(\mathrm{u}, \mathcal{S}_j), \forall j \in [m]$. Suppose we run Algorithm 1 with GD for task-specific learning with $\gamma = \frac{1}{\lambda T}$ to find an algorithm $\mathcal{A}(\mathbf{S}) = \mathcal{A}_{\text{task}}(\mathrm{w}_{T+1}, \cdot)$ which is then run on $\mathcal{S}_j$ for K iterations with step-size $\eta \leq \frac{1}{\lambda}$. Then, we have that

$$L(\mathcal{A}(\mathbf{S})(\mathcal{S}_j), \mathcal{S}_j) - \inf_{\mathrm{u}} L(\mathrm{u}, \mathcal{S}_j) \leq \frac{D^2}{2\eta(1-\eta\lambda)K} + \frac{G^2\eta}{2(1-\eta\lambda)} + \frac{GD\eta\lambda}{1-\eta\lambda} + \frac{\lambda \left\|\mathrm{w}_{T+1} - \widehat{\mathrm{w}}\right\|^2 + \lambda\sigma^2}{(1-\eta\lambda)(2-\eta\lambda)}$$

where $\sigma^2 := \frac{1}{K}\sum_{j=1}^K \left\|\widehat{\mathrm{w}} - \mathrm{u}_j^*\right\|^2$, with $\widehat{\mathrm{w}}$ as defined in Equation (1). $\left\|\mathrm{w}_{T+1} - \widehat{\mathrm{w}}\right\|^2 \leq \frac{1}{T}\left(8D^2 + \frac{4D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{\lambda}\right) + \frac{2D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{2\lambda}$.

Proof of Theorem D.5. Recall the definition $\widehat{\mathrm{w}} = \operatorname{argmin}_{\mathrm{w}\in\mathcal{W}} \frac{1}{m}\sum_{j=1}^m \min_{\mathrm{u}}\left[L(\mathrm{u}; \mathcal{S}_j) + \frac{\lambda}{2}\left\|\mathrm{u} - \mathrm{w}\right\|^2\right]$, $\mathrm{u}^*(\mathrm{w}, \mathcal{S}) = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}}\left[L(\mathrm{u}; \mathcal{S}) + \frac{\lambda}{2}\left\|\mathrm{u} - \mathrm{w}\right\|^2\right]$, $\mathrm{u}_j^* = \operatorname{argmin}_{\mathrm{u}\in\mathcal{W}} L(\mathrm{u}, \mathcal{S}_j), \forall j \in [m]$. We slightly abuse the notation by defining $\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) = \mathcal{A}(\mathbf{S})(\mathcal{S}_j)$ at inner iteration k for given w_t. Then we have

$$\begin{aligned}
&\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\|^2 \\
&= \left\|\Pi_{\mathcal{W}}\left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \eta\left(\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j) + \lambda(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{w}_t)\right)\right) - \mathrm{u}_j^*\right\|^2 \\
&\leq \left\|(1-\eta\lambda)\left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right) + \eta\lambda\left(\mathrm{w}_t - \mathrm{u}_j^*\right) - \eta\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j)\right\|^2 \\
&\leq (1-\eta\lambda)^2\left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\|^2 + \eta^2\left\|\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j)\right\|^2 + \eta^2\lambda^2\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\|^2 \\
&\quad + 2\eta\lambda(1-\eta\lambda)\left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\|\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\| + \eta^2\lambda G\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\| \\
&\quad - 2\eta(1-\eta\lambda)\left\langle\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j), \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\rangle \\
&= \left((1-\eta\lambda)\left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\| + \eta\lambda\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\|\right)^2 + \eta^2\left\|\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j)\right\|^2 \\
&\quad + \eta^2\lambda G\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\| - 2\eta(1-\eta\lambda)\left\langle\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j), \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\rangle \\
&\leq \left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\|^2 + \frac{\eta\lambda}{2-\eta\lambda}\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\|^2 + \eta^2G^2 + \eta^2\lambda G\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\| \\
&\qquad\qquad ((a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2 \text{ with } p = \tfrac{(2-\eta\lambda)\eta\lambda}{(1-\eta\lambda)^2}) \\
&\quad - 2\eta(1-\eta\lambda)\left\langle\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j), \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\rangle
\end{aligned}$$

Rearrange it and telescope it gives us that

$$\begin{aligned}
&L\left(\frac{1}{K}\sum_{k=1}^K \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j\right) - L(\mathrm{u}_j^*, \mathcal{S}_j) \\
&\leq \frac{1}{K}\sum_{k=1}^K L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j) - L(\mathrm{u}_j^*, \mathcal{S}_j) && \text{(Jensen's inequality)} \\
&\leq \frac{1}{K}\sum_{k=1}^K \left\langle\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j), \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}_j^*\right\rangle && \text{(Convexity)} \\
&\leq \frac{\left\|\mathrm{u}_j^*\right\|^2 + \eta^2G^2K + \eta^2\lambda G\sum_{j=1}^K\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\| + \frac{\eta\lambda}{2-\eta\lambda}\sum_{j=1}^K\left\|\mathrm{w}_t - \mathrm{u}_j^*\right\|^2}{2\eta(1-\eta\lambda)K} \\
&\leq \frac{D^2}{2\eta(1-\eta\lambda)K} + \frac{G^2\eta}{2(1-\eta\lambda)} + \frac{2GD\eta\lambda}{1-\eta\lambda} + \frac{\lambda\left\|\mathrm{w}_t - \widehat{\mathrm{w}}\right\|^2 + \lambda\sigma^2}{(1-\eta\lambda)(2-\eta\lambda)} \quad (\sigma^2 = \tfrac{1}{K}\textstyle\sum_{k=1}^K\left\|\mathrm{u}_j^* - \widehat{\mathrm{w}}\right\|^2)
\end{aligned}$$

Follow from [Zhou et al., 2019, Theorem 1], we now control $\|\mathrm{w}_{t+1} - \widehat{\mathrm{w}}\|^2$. Define $\mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) = \operatorname{argmin}_{\mathrm{u}} F_{\mathcal{S}_j}(\mathrm{u}, \mathrm{w}_t) = \operatorname{argmin}_{\mathrm{u}} L(\mathrm{u}, \mathcal{S}_j) + \frac{\lambda}{2}\|\mathrm{u} - \mathrm{w}_t\|^2$. We start with the following:

$$
\begin{aligned}
\|\mathrm{w}_{t+1} - \widehat{\mathrm{w}}\|^2 &= \left\| \Pi_{\mathcal{W}} \left(\mathrm{w}_t - \gamma\lambda \left(\mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right) \right) - \widehat{\mathrm{w}} \right\|^2 \\
&\leq \left\| \mathrm{w}_t - \widehat{\mathrm{w}} - \gamma\lambda \left(\mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right) \right\|^2 \\
&\leq \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 - 2\gamma\lambda \left\langle \mathrm{w}_t - \widehat{\mathrm{w}}, \mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right\rangle \\
&\quad + \gamma^2\lambda^2 \left\| \mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right\|^2 \qquad (19)
\end{aligned}
$$

We now bound the latter two terms separately as follows:

$$
\begin{aligned}
&\left\| \mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right\|^2 \\
&= \left\| \mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) + \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left(\mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right) \right\|^2 \\
&\leq 2 \left\| \mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right\|^2 + \frac{2}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\| \left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right) \right\|^2 \\
&\leq 8D^2 + \frac{2}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\| \left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right) \right\|^2
\end{aligned}
$$

as well as

$$
\begin{aligned}
&\left\langle \mathrm{w}_t - \widehat{\mathrm{w}}, \mathrm{w}_t - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) \right\rangle \\
&= \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \langle \mathrm{w}_t - \widehat{\mathrm{w}}, \mathrm{w}_t - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \rangle \\
&\quad - \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\langle \mathrm{w}_t - \widehat{\mathrm{w}}, \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right\rangle \\
&\geq \frac{1}{m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\langle \mathrm{w}_t - \widehat{\mathrm{w}}, \frac{1}{\lambda}\nabla F_{\mathcal{S}_j}(\mathrm{u}^{(k)}, \mathrm{w}_t) \right\rangle - \frac{1}{2}\|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 \\
&\quad - \frac{1}{2m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\| \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right\|^2 \\
&\geq \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 - \frac{1}{2}\|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 - \frac{1}{2m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\| \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right\|^2 \\
&\qquad\qquad (\tfrac{1}{m}\textstyle\sum_{j=1}^{m} F_{\mathcal{S}_j}(\mathrm{u}, \mathrm{w}) \text{ is } \lambda\text{-strongly convex w.r.t. } \mathrm{w}) \\
&= \frac{1}{2}\|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 - \frac{1}{2m}\sum_{j=1}^{m} \frac{1}{K}\sum_{k=1}^{K} \left\| \mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right\|^2
\end{aligned}
$$

https://doi.org/10.52202/079017-2661

where the common term $\left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2$ can be controlled as follows:

$$
\begin{aligned}
&\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2 \\
&\leq \left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \eta \nabla F_{\mathcal{S}_j}(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathrm{w}_t) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2 \\
&\leq \left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2 - 2\eta \left\langle \nabla F_{\mathcal{S}_j}(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathrm{w}_t), \mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j) \right\rangle \\
&\qquad + \eta^2 \left\|\nabla F_{\mathcal{S}_j}(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathrm{w}_t)\right\|^2 \\
&\leq (1-2\eta\lambda) \left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2 + \eta^2 \left\|\nabla L(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j), \mathcal{S}_j) + \lambda(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{w}_t)\right\|^2 \\
&\qquad\qquad (F_{\mathcal{S}_j}(\mathrm{u}, \mathrm{w}) \text{ is } \lambda\text{-strongly convex w.r.t. u}) \\
&\leq (1-2\eta\lambda) \left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2 + \eta^2 (G + 2\lambda D)^2 \\
&\leq 4(1-2\eta\lambda)^k D^2 + \frac{\eta(G+2\lambda D)^2}{2\lambda} \qquad \text{(Telescoping)}
\end{aligned}
$$

Plug back into Equation (19) gives us that

$$
\begin{aligned}
&\|\mathrm{w}_{t+1} - \widehat{\mathrm{w}}\|^2 \\
&\leq \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 - 2\gamma\lambda \left(\frac{1}{2} \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 - \frac{1}{2m} \sum_{j=1}^{m} \frac{1}{K} \sum_{k=1}^{K} \left\|\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right\|^2 \right) \\
&\qquad + \gamma^2\lambda^2 \left(8D^2 + \frac{2}{m} \sum_{j=1}^{m} \frac{1}{K} \sum_{k=1}^{K} \left\| \left(\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_j) - \mathrm{u}^*(\mathrm{w}_t, \mathcal{S}_j)\right) \right\|^2 \right) \\
&\leq (1-\gamma\lambda) \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 + \gamma\lambda \left(\frac{1}{K} \sum_{k=1}^{K} 4(1-2\eta\lambda)^k D^2 + \frac{\eta(G+2\lambda D)^2}{2\lambda} \right) \\
&\qquad + \gamma^2\lambda^2 \left(8D^2 + \frac{1}{K} \sum_{k=1}^{K} 8(1-2\eta\lambda)^k D^2 + \frac{\eta(G+2\lambda D)^2}{\lambda} \right) \\
&\leq (1-\gamma\lambda) \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 + \gamma\lambda \left(\frac{2D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{2\lambda} \right) + \gamma^2\lambda^2 \left(8D^2 + \frac{4D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{\lambda} \right)
\end{aligned}
$$

Choosing $\gamma = \frac{1}{\lambda T}$ gives us that

$$
\begin{aligned}
&\|\mathrm{w}_{t+1} - \widehat{\mathrm{w}}\|^2 \\
&\leq \left(1 - \frac{1}{T}\right) \|\mathrm{w}_t - \widehat{\mathrm{w}}\|^2 + \frac{1}{T} \left(\frac{2D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{2\lambda} \right) + \frac{1}{T^2} \left(8D^2 + \frac{4D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{\lambda} \right) \\
&\leq \frac{\left(8D^2 + \frac{4D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{\lambda} \right)}{T} + \frac{2D^2}{\eta\lambda K} + \frac{\eta(G+2\lambda D)^2}{2\lambda} \qquad \text{(Telescope)}
\end{aligned}
$$

□

E Missing Proofs in Section 5

Theorem E.1 (Bennett's inequality)**.** Let $x_1, \ldots, x_n$ be independent r.v. with finite variance. Further assume $|x_i - \mathbb{E}x_i| \leq a$ a.s. for all i. Define $S_n = \sum_{i=1}^n [x_i - \mathbb{E}[x_i]]$ and $\sigma^2 = \sum_{i=1}^n \mathbb{E}\left(x_i - \mathbb{E}[x_i]\right)^2$. Then for any $t \geq 0$,

$$
\mathrm{P}\left(S_n > t\right) \leq \exp\left(-\frac{\sigma^2}{a^2} h\left(\frac{at}{\sigma^2} \right) \right),
$$

where $h(u) = (1+u)\log(1+u) - u$.

Lemma 5.1. Assume that the loss function is ρ-weakly convex and G-Lipschitz. Let $\mathbf{S}, \mathbf{S}^{(j)}$ denote neighboring meta-samples and $\mathcal{S}, \mathcal{S}^{(i)}$ the neighboring samples on a test task. Then, with probability at least $1-\exp\left(-T^2e^2/m^2\right)$, the following holds for Algorithm 3 with $\lambda \geq 2\rho$, and GD for task-specific learning (i.e., Option 2 for Algorithm 2) with $\eta \leq \frac{1}{\lambda}$, for all $T \geq 1$ as long as we set $\gamma \leq \frac{1}{\lambda T}$,

$$\sup_{\mathbf{S},\mathcal{S},i\in[n],j\in[m]} \left\|\mathcal{A}(\mathbf{S})(\mathcal{S}) - \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})\right\| \leq (8eG + 2G)\sqrt{\frac{\eta}{\lambda}} + \frac{8eG}{\lambda m} + \frac{8G}{\lambda n}.$$

Proof of Lemma 5.1. We slightly abuse the notation, at outer iteration t, define $\mathrm{w}_t = \mathcal{A}(\mathbf{S})$, $\mathrm{w}'_t = \mathcal{A}(\mathbf{S}^{(j)})$. Given w_t, at inner iteration k, define $\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}) = \mathcal{A}(\mathbf{S})(\mathcal{S})$, $\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}^{(i)}) = \mathcal{A}(\mathbf{S}^{(j)})(\mathcal{S}^{(i)})$. From a similar argument as Lemma D.3, $\forall k \in [K-1]$, we have

$$\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}'_t, \mathcal{S}')\right\| \leq 2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{2G}{\lambda}$$
$$\left\|\mathrm{u}^{(k+1)}(\mathrm{w}_t, \mathcal{S}) - \mathrm{u}^{(k+1)}(\mathrm{w}'_t, \mathcal{S})\right\| \leq 2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + 2G\sqrt{\frac{\eta}{\lambda}}.$$

Let us define $r_t = \mathbb{1}(\mathcal{S}_{j_t} \neq \mathcal{S}_{j_t})$. Note that at every step t, $\mathbb{E}_{\mathcal{A}}(r_t) = \frac{1}{m}$. Moreover, note that $\{r_t : t \in [T]\}$ is an independent sequence of Bernoulli random variables. As a result,

$$\begin{aligned}
&\left\|\mathrm{w}_{t+1} - \mathrm{w}'_{t+1}\right\| \\
&\leq \left\|\mathrm{w}_t - \gamma\lambda\left(\mathrm{w}_t - \frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_{j_t})\right) - \mathrm{w}'_t + \gamma\lambda\left(\mathrm{w}'_t - \frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}'_{j_t})\right)\right\| \\
&\qquad\qquad \text{(Projection is non-expansive)} \\
&= (1-\gamma\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \gamma\lambda\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_t, \mathcal{S}_{j_t}) - \frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_t, \mathcal{S}'_{j_t})\right\| \\
&\leq (1-\gamma\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \gamma\lambda(1-r_t)\left(2\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + 2G\sqrt{\frac{\eta}{\lambda}}\right) + \gamma\lambda r_t\left(\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{2G}{\lambda}\right) \\
&\leq (1 + (1-r_t)\gamma\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + 2G\gamma\sqrt{\eta\lambda} + 2G\gamma r_t \\
&\leq (1+\gamma\lambda)\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + 2G\gamma\sqrt{\eta\lambda} + 2G\gamma r_t
\end{aligned}$$

Telescoping gives us that

$$\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\| \leq 2G\sqrt{\frac{\eta}{\lambda}}(1+\gamma\lambda)^T + 2G\gamma\sum_{t=1}^{T}(1+\gamma\lambda)^{t-1}\, r_t$$

Further taking expectation w.r.t the randomness of the algorithm and gives us that

$$\mathbb{E}_{\mathcal{A}}\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\| \leq (1+\gamma\lambda)^T\left(2G\sqrt{\frac{\eta}{\lambda}} + \frac{2G}{\lambda m}\right)$$

Choosing $\gamma \leq \frac{1}{\lambda T}$ gives us that

$$\mathbb{E}_{\mathcal{A}}\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\| \leq (1+\frac{1}{T})^T\left(2G\sqrt{\frac{\eta}{\lambda}} + \frac{2G}{\lambda m}\right) \leq 2eG\sqrt{\frac{\eta}{\lambda}} + \frac{2eG}{\lambda m}$$

Plug this back into Equation (18) gives us that

$$\begin{aligned}
\mathbb{E}_{\mathcal{A}}\left\|\mathrm{u}^{(K+1)}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}^{(K+1)}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\| &\leq 2G\sqrt{\frac{\eta}{\lambda}} + 4\mathbb{E}_{\mathcal{A}}\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{8G(1-\eta\lambda)}{\lambda n} \\
&\leq (8eG + 2G)\sqrt{\frac{\eta}{\lambda}} + \frac{8eG}{\lambda m} + \frac{8G}{\lambda n}
\end{aligned}$$

https://doi.org/10.52202/079017-2661

Setting $\gamma \leq \frac{1}{\lambda T}$. We note that for each r_t has variance smaller than $\frac{1}{m}$. Define random variable $x_t := (1+\frac{1}{T})^{t-1} r_t$. We have

$$|x_t - \mathbb{E}[x_t]| = (1+\frac{1}{T})^{t-1}(r_t - \mathbb{E}[r_t]) \leq e\,|x_t - \mathbb{E}[x_t]| \leq (1+\frac{1}{T})^{t-1}\left(1-\frac{1}{m}\right) \leq e$$

$$\sum_{t=1}^{T} \mathbb{E}\,(x_i - \mathbb{E}[x_t])^2 \leq \sum_{t=1}^{T} (1+\frac{1}{T})^{2t-2}\frac{1}{m}\left(1-\frac{1}{m}\right) < \frac{Te^2}{m}$$

Hence by Bennett's inequality Theorem E.1, we have

$$\mathrm{P}\left[\sum_{t=1}^{T}(1+\frac{1}{T})^{t-1} r_t \geq \frac{1}{m}\sum_{t=1}^{T}(1+\frac{1}{T})^{t-1}\right] \leq \exp\left(-\frac{T^2e^2}{m^2}\right).$$

Therefore, with probability at least $1-\exp\left(-T^2e^2/m^2\right)$, we have

$$\left\|\mathrm{w}_{T+1} - \mathrm{w}'_{T+1}\right\| \leq 2G\sqrt{\frac{\eta}{\lambda}}(1+\frac{1}{T})^T + \frac{2G}{\lambda m T}\sum_{t=1}^{T}\left(1+\frac{1}{T}\right)^{t-1} \leq 2eG\sqrt{\frac{\eta}{\lambda}} + \frac{2eG}{\lambda m}$$

and therefore with probability at least $1-\exp\left(-T^2e^2/m^2\right)$, we have

$$\begin{aligned}\forall k \in [K-1], \left\|\mathrm{u}^{(K+1)}(\mathrm{w}_{T+1}, \mathcal{S}) - \mathrm{u}^{(K+1)}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\| &\leq 2G\sqrt{\frac{\eta}{\lambda}} + 4\left\|\mathrm{w}_t - \mathrm{w}'_t\right\| + \frac{8G(1-\eta\lambda)}{\lambda n} \\ &\leq (8eG+2G)\sqrt{\frac{\eta}{\lambda}} + \frac{8eG}{\lambda m} + \frac{8G}{\lambda n}\end{aligned}$$

By triangle inequality, we have with probability at least $1-\exp\left(-T^2e^2/m^2\right)$,

$$\left\|\frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}_{T+1}, \mathcal{S}) - \frac{1}{K}\sum_{k=1}^{K}\mathrm{u}^{(k)}(\mathrm{w}'_{T+1}, \mathcal{S}^{(i)})\right\| \leq (8eG+2G)\sqrt{\frac{\eta}{\lambda}} + \frac{8eG}{\lambda m} + \frac{8G}{\lambda n}$$

□

Proposition 5.2. Given a loss function $\ell(\cdot, \mathrm{z})$ and its adversarial counterpart $\tilde{\ell}(\cdot, \mathrm{z})$, the following holds: (1) If ℓ is G-Lipschitz (in its first argument), then $\tilde{\ell}$ is G-Lipschitz. (2) $\tilde{\ell}$ is **not** H-smooth even if ℓ is H-smooth. (3) If ℓ is H-smooth in w, then $\tilde{\ell}$ is H-weakly convex in w.

Proof of Proposition 5.2. Given $\mathrm{w}_1, \mathrm{w}_2$, define

$$\tilde{\mathrm{z}}_1 \in \operatorname*{argmax}_{\tilde{\mathrm{z}}\in\mathcal{B}(\mathrm{z})} \ell(\mathrm{w}_1, \tilde{\mathrm{z}})$$

$$\tilde{\mathrm{z}}_2 \in \operatorname*{argmax}_{\tilde{\mathrm{z}}\in\mathcal{B}(\mathrm{z})} \ell(\mathrm{w}_2, \tilde{\mathrm{z}}).$$

For the first item, it holds as

$$\begin{aligned}\left|\tilde{\ell}(\mathrm{w}_1, \mathrm{z}) - \tilde{\ell}(\mathrm{w}_2, \mathrm{z})\right| &= \left\|\ell(\mathrm{w}_1, \tilde{\mathrm{z}}_1) - \ell(\mathrm{w}_2, \tilde{\mathrm{z}}_2)\right\| \\ &= \max\left\{\left|\ell(\mathrm{w}_1, \tilde{\mathrm{z}}_1) - \ell(\mathrm{w}_2, \tilde{\mathrm{z}}_1)\right|, \left|\ell(\mathrm{w}_1, \tilde{\mathrm{z}}_2) - \ell(\mathrm{w}_2, \tilde{\mathrm{z}}_2)\right|\right\} \\ &\leq G\left\|\mathrm{w}_1 - \mathrm{w}_2\right\|.\end{aligned}$$

For the second item, the non-smoothness of the adversarial loss has been verified in Xing et al. [2021], Xiao et al. [2022]. For the third item, $\ell(\mathrm{w}, \mathrm{z})$ is H-smooth implies that $\ell(\mathrm{w}, \mathrm{z})$ is H-weakly convex, and further derive that $\tilde{\ell}(\mathrm{w}, \mathrm{z})$ is H-weakly convex because

$$\begin{aligned}\tilde{\ell}(\mathrm{w}_2, \mathrm{z}) &= \ell(\mathrm{w}_2, \tilde{\mathrm{z}}_2) \\ &\geq \ell(\mathrm{w}_2, \tilde{\mathrm{z}}_1) && \text{(By definition of } \tilde{\mathrm{z}}_1, \tilde{\mathrm{z}}_2) \\ &\geq \ell(\mathrm{w}_1, \tilde{\mathrm{z}}_1) + \langle g(\mathrm{w}_1, \tilde{\mathrm{z}}_1), \mathrm{w}_2 - \mathrm{w}_1\rangle - \frac{\rho}{2}\left\|\mathrm{w}_2 - \mathrm{w}_1\right\|^2 \\ & && (g(\mathrm{w}_1, \tilde{\mathrm{z}}_1) \in \partial\ell(\mathrm{w}_2, \tilde{\mathrm{z}}_1), \text{ apply Proposition D.1}) \\ &= \tilde{\ell}(\mathrm{w}_1, \mathrm{z}) + \langle \tilde{g}(\mathrm{w}_1, \mathrm{z}), \mathrm{w}_2 - \mathrm{w}_1\rangle - \frac{\rho}{2}\left\|\mathrm{w}_2 - \mathrm{w}_1\right\|^2 && (\text{Redefine } \tilde{g}(\mathrm{w}_1, \mathrm{z}) \in \partial\tilde{\ell}(\mathrm{w}_1, \mathrm{z}))\end{aligned}$$

□

NeurIPS Paper Checklist

A. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We further expand on the claims made in abstract and introduction in Section 3 and 4. The detailed proofs are provided in the Appendix.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

B. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We list the required assumptions in the statement of each theorems. Several limitations are described together with future work in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

C. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-2661

Answer: [Yes]

Justification: All the proofs are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

D. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Please see Section A in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

E. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Please see Section A in the Appendix. The code is provided in the supplementary file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

F. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please see Section A in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

G. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Please see Section A in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-2661

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

H. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Please see Section A in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

I. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The theoretical nature of the results means there are minimal ethical concerns.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

J. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Section 5.2 and 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

K. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

L. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-2661

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

M. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

N. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

O. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.