On the Global Optimality of Model-Agnostic Meta-Learning: Reinforcement Learning and Supervised Learning

Lingxiao Wang [1]   **Qi Cai** [1]   **Zhuoyan Yang** [2]   **Zhaoran Wang** [1]

Abstract

Model-agnostic meta-learning (MAML) formulates meta-learning as a bilevel optimization problem, where the inner level solves each subtask based on a shared prior, while the outer level searches for the optimal shared prior by optimizing its aggregated performance over all the subtasks. Despite its empirical success, MAML remains less understood in theory, especially in terms of its global optimality, due to the nonconvexity of the meta-objective (the outer-level objective). To bridge such a gap between theory and practice, we characterize the optimality gap of the stationary points attained by MAML for both reinforcement learning and supervised learning, where the inner-level and outer-level problems are solved via first-order optimization methods. In particular, our characterization connects the optimality gap of such stationary points with (i) the functional geometry of inner-level objectives and (ii) the representation power of function approximators, including linear models and neural networks. To the best of our knowledge, our analysis establishes the global optimality of MAML with nonconvex meta-objectives for the first time. [1]

1. Introduction

Meta-learning aims to find a prior that efficiently adapts to a new subtask based on past subtasks. One of the most popular meta-learning methods, namely model-agnostic meta-learning (MAML) (Finn et al., 2017a), is based on bilevel op-

timization, where the inner level solves each subtask based on a shared prior, while the outer level optimizes the aggregated performance of the shared prior over all the subtasks. In particular, MAML associates the solution to each subtask with the shared prior through one step of gradient descent based on the subtask data. Due to its model-agnostic property, MAML is widely adopted in reinforcement learning (Finn et al., 2017a;b; Xu et al., 2018; Nagabandi et al., 2018; Gupta et al., 2018; Yu et al., 2018; Mendonca et al., 2019) and supervised learning (Finn et al., 2017a; Li et al., 2017; Finn et al., 2018; Rakelly et al., 2018; Yoon et al., 2018).

Despite its popularity in empirical studies, MAML is scarcely explored theoretically. In terms of the global optimality of MAML, (Finn et al., 2019) show that the meta-objective is strongly convex assuming that the inner-level objective is strongly convex (in its finite-dimensional parameter). However, such an assumption fails to hold for neural function approximators, which leads to a gap between theory and practice. For nonconvex meta-objectives, (Fallah et al., 2019) characterize the convergence of MAML to a stationary point under certain regularity conditions. Meanwhile, (Rajeswaran et al., 2019) propose a variant of MAML that utilizes implicit gradients, which is also guaranteed to converge to a stationary point. However, the global optimality of such stationary points remains unclear. On the other hand, (Pentina & Lampert, 2014; Amit & Meir, 2017) establish PAC-Bayes bounds for the generalization error of two variants of MAML. However, such generalization guarantees only apply to the global optima of the two meta-objectives rather than their stationary points.

In this work, we characterize the global optimality of the ϵ-stationary points attained by MAML for both reinforcement learning (RL) and supervised learning (SL). For meta-RL, we study a variant of MAML, which associates the solution to each subtask with the shared prior, namely π_θ, through one step of proximal policy optimization (PPO) (Schulman et al., 2015; 2017) in the inner level of optimization. In the outer level of optimization, we maximize the expected total reward associated with the shared prior aggregated over all the subtasks. We prove that the ϵ-stationary point attained by such an algorithm is (approximately) globally optimal given that the function approximator has sufficient

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, USA [2]Department of Operations Research and Financial Engineering, Princeton University, USA. Correspondence to: Lingxiao Wang <lingxiaowang2022@u.northwestern.edu>.

[1]See https://arxiv.org/abs/2006.13182 for the full version.

representation power. For example, for the linear function approximator $\pi_\theta(s, a) \propto \exp(\phi(s, a)^\top \theta)$, the optimality gap of the ϵ-stationary point is characterized by the representation power of the linear class $\{\phi(\cdot, \cdot)^\top v : v \in \mathcal{B}\}$, where $\mathcal{B}$ is the parameter space (which is specified later). The core of our analysis is the functional one-point monotonicity (Facchinei & Pang, 2007) of the expected total reward $J(\pi)$ with respect to the policy π (Liu et al., 2019) for each subtask. Based on a similar notion of functional geometry in the inner level of optimization, we establish similar results on the optimality gap of meta-SL. Moreover, our analysis of both meta-RL and meta-SL allows for neural function approximators. More specifically, we prove that the optimality gap of the attained ϵ-stationary points is characterized by the representation power of the corresponding classes of overparameterized two-layer neural networks.

Challenge. We highlight that the bilevel structure of MAML makes it challenging for the analysis of its global optimality. In the simple case where the inner-level objective is strongly convex and smooth, (Finn et al., 2019) show that the meta-objective is also strongly convex assuming that the stepsize of inner-level optimization is sufficiently small.

- In practice, however, both the inner-level objective and the meta-objective can be nonconvex, which leads to a gap between theory and practice. For example, the inner-level objective of meta-RL is nonconvex even in the (infinite-dimensional) functional space of policies.

- Even assuming that the inner-level objective is convex in the (infinite-dimensional) functional space, nonlinear function approximators, such as neural networks, can make the inner-level objective nonconvex in the finite-dimensional space of parameters.

- Furthermore, even for linear function approximators, the bilevel structure of MAML can make the meta-objective nonconvex in the finite-dimensional space of parameters, especially when the stepsize of inner-level optimization is large.

In this work, we tackle all these challenges by analyzing the global optimality of both meta-RL and meta-SL for both linear and neural function approximators.

Contribution. Our contribution is three-fold. First, we propose a meta-RL algorithm and characterize the optimality gap of the ϵ-stationary point attained by such an algorithm for linear function approximators. Second, under an assumption on the functional convexity of the inner-level objective, we characterize the optimality gap of the ϵ-stationary point attained by meta-SL. Finally, we extend our optimality analysis for linear function approximators to handle overparameterized two-layer neural networks. To the best of our

knowledge, our analysis establishes the global optimality of MAML with nonconvex meta-objectives for the first time.

Related Work. Meta-learning is studied by various communities (Evgeniou & Pontil, 2004; Thrun & Pratt, 2012; Pentina & Lampert, 2014; Amit & Meir, 2017; Nichol et al., 2018; Nichol & Schulman, 2018; Khodak et al., 2019). See (Pan & Yang, 2009; Weiss et al., 2016) for the surveys of meta-learning and (Taylor & Stone, 2009) for a survey of meta-RL. Our work focuses on the model-agnostic formulation of meta-learning (MAML) proposed by (Finn et al., 2017a). In contrast to existing empirical studies, the theoretical analysis of MAML is relatively scarce. (Fallah et al., 2019) establish the convergence of three variants of MAML for nonconvex meta-objectives. (Rajeswaran et al., 2019) propose a variant of MAML that utilizes implicit gradients of the inner level of optimization and establish the convergence of such an algorithm. This line of work characterizes the convergence of MAML to the stationary points of the corresponding meta-objectives. Our work is complementary to this line of work in the sense that we characterize the global optimality of the stationary points attained by MAML. Meanwhile, (Finn et al., 2019) propose an online algorithm for MAML with regret guarantees, which rely on the strong convexity of the meta-objectives. In contrast, our work tackles nonconvex meta-objectives, which allows for neural function approximators, and characterizes the global optimality of MAML. (Mendonca et al., 2019) propose a meta-policy search method and characterize the global optimality for solving the subtasks under the assumption that the meta-objective is (approximately) globally optimal. Our work is complementary to their work in the sense that we characterize the global optimality of MAML in terms of optimizing the meta-objective. See also the concurrent work (Wang et al., 2020).

There is a large body of literature that studies the training and generalization of overparameterized neural networks for SL (Daniely, 2017; Jacot et al., 2018; Wu et al., 2018; Allen-Zhu et al., 2018a;b; Du et al., 2018a;b; Zou et al., 2018; Chizat & Bach, 2018; Li & Liang, 2018; Cao & Gu, 2019a;b; Arora et al., 2019; Lee et al., 2019; Bai & Lee, 2019). See (Fan et al., 2019) for a survey. In comparison, we study MAML with overparameterized neural networks for both RL and SL. The bilevel structure of MAML makes our analysis significantly more challenging than that of RL and SL.

Notation. We denote by $[n] = \{1, 2, ..., n\}$ the index set. Also, we denote by $x = ([x]_1^\top, \ldots, [x]_m^\top)^\top \in \mathbb{R}^{md}$ a vector in $\mathbb{R}^{md}$, where $[x]_k \in \mathbb{R}^d$ is the k-th block of x for $k \in [m]$. For a real-valued function f defined on $\mathcal{X}$, we denote by $\|f(\cdot)\|_{p,\nu} = \{\int_{\mathcal{X}} f^p(x) \, \mathrm{d}\nu(x)\}^{1/p}$ the $L_p(\nu)$-norm of f, where ν is a measure on $\mathcal{X}$. We write $\|f(\cdot)\|_{2,\nu} = \|f(\cdot)\|_\nu$ for notational simplicity and $\|f\|_{p,\nu} = \|f(\cdot)\|_{p,\nu}$ when the

variable is clear from the context. For a vector $\phi \in \mathbb{R}^n$, we denote by $\|\phi\|_2$ the ℓ_2-norm of ϕ.

2. Background

In this section, we briefly introduce reinforcement learning and meta-learning.

2.1. Reinforcement Learning

We define a Markov decision process (MDP) by a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \zeta)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, P is the Markov kernel, r is the reward function, which is possibly stochastic, $\gamma \in (0, 1)$ is the discount factor, and ζ is the initial state distribution over $\mathcal{S}$. In the sequel, we assume that $\mathcal{A}$ is finite. An agent interacts with the environment as follows. At each step t, the agent observes the state s_t of the environment, takes the action a_t, and receives the reward $r(s_t, a_t)$. The environment then transits into the next state according to the distribution $P(\cdot \,|\, s_t, a_t)$ over $\mathcal{S}$. We define a policy π as a mapping from $\mathcal{S}$ to distributions over $\mathcal{A}$. Specifically, $\pi(a \,|\, s)$ gives the probability of taking the action a at the state s. Given a policy π, we define for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ the corresponding state- and action-value functions V^π and Q^π as follows,

$$
V^\pi(s) = (1 - \gamma) \cdot \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r(s_t, a_t) \,\middle|\, s_0 = s \right],
$$
(2.1)
$$
Q^\pi(s, a) = (1 - \gamma) \cdot \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \cdot r(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a \right],
$$
(2.2)

where $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ and $a_t \sim \pi(\cdot \,|\, s_t)$ for all $t \geq 0$. Correspondingly, the advantage function A^π is defined as follows,

$$
A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.
$$
(2.3)

A policy π induces a state visitation measure ν_π on $\mathcal{S}$, which takes the form of

$$
\nu_\pi(s) = (1 - \gamma) \cdot \sum_{t=0}^\infty \gamma^t \cdot \mathbb{P}(s_t = s),
$$
(2.4)

where $s_0 \sim \zeta$, $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$, and $a_t \sim \pi(\cdot \,|\, s_t)$ for all $t \geq 0$. Correspondingly, we define the state-action visitation measure by $\sigma_\pi(s, a) = \pi(a \,|\, s) \cdot \nu_\pi(s)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, which is a probability distribution over $\mathcal{S} \times \mathcal{A}$. The goal of reinforcement learning is to find the optimal policy π^* that

maximizes the expected total reward $J(\pi)$, which is defined as

$$
J(\pi) = \mathbb{E}_{s \sim \zeta}\left[V^\pi(s) \right] = \mathbb{E}_{(s,a) \sim \sigma_\pi}\left[r(s, a) \right].
$$
(2.5)

When $\mathcal{S}$ is continuous, maximizing $J(\pi)$ over all possible π is computationally intractable. A common alternative is to parameterize the policy by π_θ with the parameter $\theta \in \Theta$, where Θ is the parameter space, and maximize $J(\pi_\theta)$ over $\theta \in \Theta$.

2.2. Meta-Learning

In meta-learning, the meta-learner is given a sample of learning subtasks $\{\mathcal{T}_i\}_{i \in [n]}$ drawn independently from the task distribution ι and a set of parameterized algorithms $\mathcal{A} = \{\mathscr{A}_\theta : \theta \in \Theta\}$, where Θ is the parameter space. Specifically, given θ, the algorithm $\mathscr{A}_\theta \in \mathcal{A}$ maps from a learning subtask $\mathcal{T}$ to its desired outcome. For example, an algorithm that solves reinforcement learning subtasks maps from an MDP $\mathcal{T} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \zeta)$ to a policy π, aiming at maximizing the expected total reward $J(\pi)$ defined in (2.5). As an example, given a hypothesis class $\mathcal{H}$, a distribution $\mathcal{D}$ over $\mathcal{Z}$, which is the space of data points, and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \mapsto \mathbb{R}$, a supervised learning subtask aims at minimizing the risk $\mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ over $h \in \mathcal{H}$. We denote the supervised learning subtask $\mathcal{T}$ by the tuple $(\mathcal{D}, \ell, \mathcal{H})$. Similarly, an algorithm that solves supervised learning subtasks maps from $\mathcal{T} = (\mathcal{D}, \ell, \mathcal{H})$ to a hypothesis $h \in \mathcal{H}$, aiming at minimizing the risk $R(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$ over $h \in \mathcal{H}$. In what follows, we denote by $H_\mathcal{T}$ the objective of a learning subtask $\mathcal{T}$. If $\mathcal{T}$ is a reinforcement learning subtask, we have $H_\mathcal{T} = J$, and if $\mathcal{T}$ is a supervised learning subtask, we have $H_\mathcal{T} = R$.

The goal of the meta-learner is to find $\theta^* \in \Theta$ that optimizes the population version of the meta-objective $\overline{L}(\theta)$, which is defined as

$$
\overline{L}(\theta) = \mathbb{E}_{\mathcal{T} \sim \iota}\left[H_\mathcal{T}\big(\mathscr{A}_\theta(\mathcal{T})\big) \right].
$$
(2.6)

To approximately optimize $\overline{L}$ defined in (2.6) based on the sample $\{\mathcal{T}_i\}_{i \in [n]}$ of subtasks, the meta-learner optimizes the following empirical version of the meta-objective,

$$
L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^n H_{\mathcal{T}_i}\big(\mathscr{A}_\theta(\mathcal{T}_i)\big).
$$
(2.7)

The algorithm $\mathscr{A}_{\theta^*}$ corresponding to the global optimum θ^* of (2.7) incorporates the past experience through the observed learning subtasks $\{\mathcal{T}_i\}_{i \in [n]}$, and therefore, facilitates the learning of a new subtask (Pentina & Lampert,

2014; Finn et al., 2017a; Amit & Meir, 2017; Yoon et al., 2018). As an example, in model-agnostic meta-learning (MAML) (Finn et al., 2017a) for supervised learning, the hypothesis class $\mathcal{H}$ is parameterized by h_θ with $\theta \in \Theta$, and the algorithm $\mathscr{A}_\theta$ performs one step of gradient descent with $\theta \in \Theta$ as the starting point. In this setting, MAML aims to find the globally optimal starting point θ^* by minimizing the following meta-objective by gradient descent,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} R_i\big(h_{\theta - \eta \cdot \nabla_\theta R_i(h_\theta)}\big),$$

where η is the learning rate of $\mathscr{A}_\theta$ and $R_i(h) = \mathbb{E}_{z \sim \mathcal{D}_i}[\ell(h, z)]$ is the risk of the supervised learning subtask $\mathcal{T}_i = (\mathcal{D}_i, \ell, \mathcal{H})$. Similarly, in MAML for reinforcement learning, the algorithm $\mathscr{A}_\theta$ performs, e.g., one step of policy gradient with θ as the starting point. We call π_θ the main effect in the sequel. MAML aims to find the globally optimal main effect π_{θ^*} by maximizing the following meta-objective by gradient ascent,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} J_i\big(\pi_{\theta + \eta \cdot \nabla_\theta J_i(\pi_\theta)}\big),$$

where η is the learning rate of $\mathscr{A}_\theta$ and J_i is the expected total reward of the reinforcement learning subtask $\mathcal{T}_i = (\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$.

3. Meta-Reinforcement Learning

In this section, we present the analysis of meta-reinforcement learning (meta-RL). We first define the detailed problem setup of meta-RL and propose a meta-RL algorithm. We then characterize the global optimality of the stationary point attained by such an algorithm. We refer the analysis of meta-RL with neural network parameterization to §A.2.

3.1. Problem Setup and Algorithm

In meta-RL, the meta-learner observes a sample of MDPs $\{(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)\}_{i \in [n]}$ drawn independently from a task distribution ι. We set the algorithm $\mathscr{A}_\theta$ that optimizes the policy to be one step of (a variant of) proximal policy optimization (PPO) (Schulman et al., 2015; 2017) starting from the main effect π_θ. More specifically, $\mathscr{A}_\theta$ solves the

following maximization problem,

$$\begin{aligned} &\mathscr{A}_\theta(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i) \\ &= \operatorname*{argmax}_\pi \mathbb{E}_{s \sim \nu_{i,\pi_\theta}}\Big[\langle Q_i^{\pi_\theta}(s, \cdot), \pi(\cdot \,|\, s)\rangle \quad (3.1) \\ &\quad - 1/\eta \cdot D_{\mathrm{KL}}\big(\pi(\cdot \,|\, s)\big\|\pi_\theta(\cdot \,|\, s)\big)\Big]. \end{aligned}$$

Here $\langle \cdot, \cdot \rangle$ is the inner product over $\mathbb{R}^{|\mathcal{A}|}$, η is the tuning parameter of $\mathscr{A}_\theta$, and $Q_i^{\pi_\theta}$, ν_{i,π_θ} are the action-value function and the state visitation measure, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ and the policy π_θ. Note that the objective in (3.1) has $D_{\mathrm{KL}}(\pi(\cdot \,|\, s)\|\pi_\theta(\cdot \,|\, s))$ in place of $D_{\mathrm{KL}}(\pi_\theta(\cdot \,|\, s)\|\pi(\cdot \,|\, s))$ compared with the original version of PPO (Schulman et al., 2015; 2017). As shown by (Liu et al., 2019), such a variant of PPO enjoys global optimality and convergence.

We parameterize the main effect π_θ as the following energy-based policy (Haarnoja et al., 2017) for all $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$\pi_\theta(a \,|\, s) = \frac{\exp\big(1/\tau \cdot \phi(s, a)^\top \theta\big)}{\sum_{a' \in \mathcal{A}} \exp\big(1/\tau \cdot \phi(s, a')^\top \theta\big)}, \quad (3.2)$$

where $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is the feature mapping, $\theta \in \mathbb{R}^d$ is the parameter, $\phi(\cdot, \cdot)^\top \theta$ is the energy function, and τ is the temperature parameter. The maximizer $\pi_{i,\theta} = \mathscr{A}_\theta(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ defined in (3.1) then takes the following form (Liu et al., 2019, Proposition 3.1) for all $s \in \mathcal{S}$,

$$\pi_{i,\theta}(\cdot \,|\, s) \propto \exp\big(1/\tau \cdot \phi(s, \cdot)^\top \theta + \eta \cdot Q_i^{\pi_\theta}(s, \cdot)\big). \quad (3.3)$$

The goal of meta-RL is to find the globally optimal main effect π_θ by maximizing the following meta-objective,

$$\begin{aligned} L(\theta) &= \frac{1}{n} \cdot \sum_{i=1}^{n} J_i(\pi_{i,\theta}), \\ &\text{where } \pi_{i,\theta} = \mathscr{A}_\theta(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i). \quad (3.4) \end{aligned}$$

Here J_i is the expected total reward defined in (2.5) corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ for all $i \in [n]$. To maximize $L(\theta)$, we use gradient ascent, which iteratively updates θ as follows,

$$\theta_{\ell+1} \leftarrow \theta_\ell + \alpha_\ell \cdot \nabla_\theta L(\theta_\ell), \quad \text{for } \ell = 0, 1, \ldots, T-1, \quad (3.5)$$

where $\nabla_\theta L(\theta_\ell)$ is the gradient of the meta-objective at θ_ℓ, α_ℓ is the learning rate at the ℓ-th iteration, and T is the number of iterations. It remains to calculate the gradient $\nabla_\theta L(\theta)$. To this end, we first define the state-action vis-

itation measures induced by the main effect π_θ, and then calculate $\nabla_\theta L(\theta)$ in closed form based on such state-action visitation measures.

Definition 3.1 (Visitation Measures of Main Effect). For all $i \in [n]$, given the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$ and the main effect π_θ, we denote by σ_{i,π_θ} the state-action visitation measure induced by the main effect π_θ. We further define the state-action visitation measure $\sigma_{i,\pi_\theta}^{(s,a)}$ initialized at $(s,a) \in \mathcal{S} \times \mathcal{A}$ as follows,

$$\sigma_{i,\pi_\theta}^{(s,a)}(s',a') = (1-\gamma_i) \cdot \sum_{t=0}^{\infty} \gamma_i^t \cdot \mathbb{P}(s_t = s', a_t = a'),$$
$$(3.6)$$

where $(s',a') \in \mathcal{S} \times \mathcal{A}$, $s_0 \sim P_i(\cdot \,|\, s, a)$, $s_{t+1} \sim P_i(\cdot \,|\, s_t, a_t)$, and $a_t \sim \pi_\theta(\cdot \,|\, s_t)$ for all $t \geq 0$.

In other words, given the transition kernel P_i and the discount factor γ_i, $\sigma_{i,\pi_\theta}^{(s,a)}$ is the state-action visitation measure induced by the main effect π_θ where the initial state distribution is given by $s_0 \sim P_i(\cdot \,|\, s, a)$. Based on the policy gradient theorem (Sutton & Barto, 2018), the following proposition calculates the gradient of the meta-objective L defined in (3.4) with respect to the parameter θ of the main effect π_θ.

Proposition 3.2 (Gradient of Meta-Objective). It holds for all $\theta \in \mathbb{R}^d$ that

$$\nabla_\theta L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta}}} \left[h_{i,\theta}(s,a) \cdot A_i^{\pi_{i,\theta}}(s,a) \right],$$
$$(3.7)$$

where the auxiliary function $h_{i,\theta}$ takes the form of

$$h_{i,\theta}(s,a) = 1/\tau \cdot \phi(s,a) \qquad (3.8)$$
$$+ \eta \cdot \gamma_i/\tau \cdot \mathbb{E}_{(s',a') \sim \sigma_{i,\pi_\theta}^{(s,a)}} \left[\phi(s',a') \cdot A_i^{\pi_\theta}(s',a') \right].$$

Here $A_i^{\pi_{i,\theta}}$ and $A_i^{\pi_\theta}$ are the advantage functions of the policy $\pi_{i,\theta}$ and the main effect π_θ, respectively, both corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$. Also, $\sigma_{i,\pi_\theta}^{(s,a)}$ is the state-action visitation measure induced by the main effect π_θ defined in Definition 3.1, and $\sigma_{\pi_{i,\theta}}$ is the state-action visitation measure induced by the policy $\pi_{i,\theta}$, both corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$.

Proof. See §D.1 for a detailed proof. □

In the sequel, we assume without loss of generality that the action-value function Q^π is available once we obtain the

policy π, and the expectations over state-action visitation measures in (3.7) and (3.8) of Theorem 3.2 are available once we obtain the policies $\{\pi_{i,\theta}\}_{i \in [n]}$ and the main effect π_θ. We summarize meta-RL in Algorithm 1. In practice, we can estimate the action-value functions by temporal difference learning (Sutton, 1988) and the expectations over the visitation measures by Monte Carlo sampling (Konda, 2002).

Algorithm 1 Meta-RL
___
Require: Sampled MDPs $\{(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)\}_{i \in [n]}$ from the task distribution τ, feature mapping ϕ, number of iterations T, learning rate $\{\alpha_\ell\}_{\ell \in [T]}$, temperature parameter $1/\tau$, tuning parameter τ, initial parameter θ_0.
1: **Initialization:**
2: **for** $\ell = 0, \ldots, T-1$ **do**
3:     **for** $i \in [n]$ **do**
4:         Update the policy: $\pi_{i,\theta_\ell}(\cdot \,|\, s) \propto \exp\big(1/\tau \cdot \phi(s,\cdot)^\top \theta_\ell + \eta \cdot Q_i^{\pi_{\theta_\ell}}(s,\cdot)\big)$.
5:         Compute the auxiliary function $h_{i,\theta_\ell}(s,a)$ via (3.8)
6:     **end for**
7:     Compute the gradient of meta-objective $\nabla_\theta L(\theta_\ell)$ based on the policies $\{\pi_{i,\theta_\ell}\}_{i \in [n]}$ and auxiliary functions $\{h_{i,\theta_\ell}\}_{i \in [n]}$ via (3.7).
8:     Update the parameter of the main effect: $\theta_{\ell+1} \leftarrow \theta_\ell + \alpha_\ell \cdot \nabla_\theta L(\theta_\ell)$.
9:     Update the main effect: $\pi_{\theta_{\ell+1}}(\cdot \,|\, s) \propto \exp\big(1/\tau \cdot \phi(s,\cdot)^\top \theta_{\ell+1}\big)$.
10: **end for**
11: **Output:** θ_T and π_{θ_T}.

3.2. Theoretical Results

In this section, we analyze the global optimality of the ϵ-stationary point attained by meta-RL (Algorithm 1). In the sequel, we assume that the reward functions $\{r_i\}_{i \in [n]}$ are upper bounded by an absolute constant $Q_{\max} > 0$ in absolute value. It then follows from (2.1) and (2.2) that $|V_i^\pi(s,a)|$ and $|Q_i^\pi(s,a)|$ are upper bounded by $Q_{\max}$ for all $i \in [n]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Here we define Q_i^π and V_i^π as the state- and action-value functions of the policy π, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$.

To analyze the global optimality of meta-RL, we define the following meta-visitation measures induced by the main effect π_θ.

Definition 3.3 (Meta-Visitation Measures). We define the joint meta-visitation measure ρ_{i,π_θ} over $(s',a',s,a) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{A}$ induced by the main effect π_θ and the policy

$\pi_{i,\theta}$ as follows,

$$\rho_{i,\pi_\theta}(s', a', s, a) = \sigma_{i,\pi_\theta}^{(s,a)}(s', a') \cdot \sigma_{\pi_{i,\theta}}(s, a). \quad (3.9)$$

We further define the meta-visitation measure ς_{i,π_θ} as the marginal distribution of the joint meta-visitation measure ρ_{i,π_θ} of $(s', a') \in \mathcal{S} \times \mathcal{A}$, that is,

$$\varsigma_{i,\pi_\theta}(s', a') = \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta}}} \left[\sigma_{i,\pi_\theta}^{(s,a)}(s', a') \right]. \quad (3.10)$$

In addition, for $(s', a') \in \mathcal{S} \times \mathcal{A}$ we define the mixed meta-visitation measure ϱ_{π_θ} over all the subtasks as follows,

$$\varrho_{\pi_\theta}(s', a') = \frac{1}{n} \cdot \sum_{i=1}^{n} \varsigma_{i,\pi_\theta}(s', a'). \quad (3.11)$$

In other words, the meta-visitation measure ς_{i,π_θ} is the state-action visitation measure induced by π_θ given the transition kernel P_i, the discount factor γ_i, and the initial state distribution $s_0 \sim \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta}}}[P_i(\cdot \mid s, a)]$.

In what follows, we impose an assumption on the meta-visitation measures defined in Definition 3.3.

Assumption 3.4 (Regularity Condition on Meta-Visitation Measures). We assume for all $\theta \in \mathbb{R}^d$ and $i \in [n]$ that

$$\mathbb{E}_{(s',a') \sim \varrho_{\pi_\theta}} \left[\left(d\sigma_{\pi_{i,\theta}} / d\varrho_{\pi_\theta}(s', a') \right)^2 \right] \leq C_0^2, \quad (3.12)$$

$$\mathbb{E}_{(s',a') \sim \varrho_{\pi_\theta}} \left[\left(d\varsigma_{i,\pi_\theta} / d\varrho_{\pi_\theta}(s', a') \right)^2 \right] \leq C_0^2, \quad (3.13)$$

where $C_0 > 0$ is an absolute constant . Here ς_{i,π_θ} and ϱ_{π_θ} are the meta-visitation measure and the mixed meta-visitation measure induced by the main effect π_θ, which are defined in (3.10) and (3.11) of Definition 3.3, respectively. Meanwhile, $\sigma_{\pi_{i,\theta}}$ is the state-action visitation measure induced by the policy $\pi_{i,\theta}$, which is defined in (2.4). Here $d\sigma_{\pi_{i,\theta}} / d\varrho_{\pi_\theta}$ and $d\varsigma_{i,\pi_\theta} / d\varrho_{\pi_\theta}$ are the Radon-Nikodym derivatives.

According to (3.11) of Definition 3.3, the upper bound in (3.12) of Assumption 3.4 holds if the $L_2(\varrho_{\pi_\theta})$-norms of $d\sigma_{\pi_{i,\theta}} / d\varsigma_{j,\pi_\theta}$ is upper bounded by C_0 for all $i, j \in [n]$. For $i = j$, note that $\pi_{i,\theta}$ is obtained by one step of PPO with π_θ as the starting point. Thus, for a sufficiently small tuning parameter η in (3.3), $\pi_{i,\theta}$ is close to π_θ. Hence, the assumption that $d\sigma_{\pi_{i,\theta}} / d\varsigma_{j,\pi_\theta}$ has an upper bounded $L_2(\varrho_{\pi_\theta})$-norm for all $i = j$ is a mild regularity condition. For $i \neq j$, to ensure the upper bound of the $L_2(\varrho_{\pi_\theta})$-norms of $d\sigma_{\pi_{i,\theta}} / d\varsigma_{j,\pi_\theta}$ in (3.12), Assumption 3.4 requires the task distribution ι to generate similar MDPs so that the

meta-visitation measures $\{\varsigma_{i,\pi_\theta}\}_{i \in [n]}$ are similar across all the subtasks indexed by $i \in [n]$. Similarly, to ensure the upper bound in (3.13), Assumption 3.4 also requires that the meta-visitation measures $\{\varsigma_{i,\pi_\theta}\}_{i \in [n]}$ are similar across all the subtasks indexed by $i \in [n]$.

The following theorem characterizes the optimality gap of the ϵ-stationary point attained by meta-RL (Algorithm 1). Let θ^* be a global maximizer of the meta-objective $L(\theta)$ defined in (3.4). For all $(s', a') \in \mathcal{S} \times \mathcal{A}$ and $\omega \in \mathbb{R}^d$, we define

$$f_\omega(s', a') = \left(\sum_{i=1}^{n} \frac{A_i^{\pi_{i,\omega}}(s', a')}{1 - \gamma_i} \cdot \frac{d\sigma_{\pi_{i,\theta^*}}}{d\varrho_{\pi_\omega}}(s', a') \right) \Big/ \left(\sum_{i=1}^{n} g_{i,\omega}(s', a') \cdot \frac{d\varsigma_{i,\pi_\omega}}{d\varrho_{\pi_\omega}}(s', a') \right), \quad (3.14)$$

where we defined $g_{i,\omega}$ as follows,

$$g_{i,\omega}(s', a') = 1/\tau \cdot A_i^{\pi_{i,\omega}}(s', a') \cdot (d\sigma_{\pi_{i,\omega}} / d\varsigma_{i,\pi_\omega})(s', a') + \gamma_i \cdot \eta/\tau \cdot G_{i,\pi_\omega}(s', a') \cdot A_i^{\pi_\omega}(s', a').$$

Here τ is the temperature parameter in (3.2), η is the tuning parameter defined in (3.1), $A_i^{\pi_{i,\omega}}$ and $A_i^{\pi_\omega}$ are the advantage functions of the policy $\pi_{i,\omega}$ and the main effect π_ω, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$, and G_{i,π_ω} is defined as follows,

$$G_{i,\pi_\omega}(s', a') = \mathbb{E}_{(s',a',s,a) \sim \rho_{i,\pi_\omega}} \left[A_i^{\pi_{i,\omega}}(s, a) \mid s', a' \right], \quad (3.15)$$

where ρ_{i,π_ω} is the joint meta-visitation measure defined in (3.9) of Definition 3.3.

Theorem 3.5 (Optimality Gap of ϵ-Stationary Point). Under Assumption 3.4, for all $R > 0$, $\omega \in \mathbb{R}^d$, and $\epsilon > 0$ such that

$$\nabla_\omega L(\omega)^\top v \leq \epsilon, \quad \forall v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}, \quad (3.16)$$

we have

$$L(\theta^*) - L(\omega) \leq R \cdot \epsilon + C \cdot \inf_{v \in \mathcal{B}_R} \|f_\omega(\cdot, \cdot) - \phi(\cdot, \cdot)^\top v\|_{\varrho_{\pi_\omega}}, \quad (3.17)$$

where $\mathcal{B}_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, $\overline{\gamma} = (\sum_{i=1}^{n} \gamma_i)/n$, and $C = 2C_0 \cdot Q_{\max}/\tau \cdot (1 + 2Q_{\max} \cdot \overline{\gamma} \cdot \eta)$. Here C_0 is defined in Assumption 3.4, τ is the temperature parameter

in (3.2), η is the tuning parameter defined in (3.1), and $Q_{\max}$ is the upper bound of the reward functions $\{r_i\}_{i \in [n]}$ in absolute value.

Proof. See §C.1 for a detailed proof. $\qquad\square$

By Theorem 3.5, the global optimality of the ϵ-stationary point ω hinges on the representation power of the linear class $\{\phi(\cdot)^\top \theta : \theta \in \mathcal{B}_R\}$. More specifically, if the function f_ω defined in (3.14) is well approximated by $\phi(\cdot)^\top \theta$ for a parameter $\theta \in \mathcal{B}_R$, then ω is approximately globally optimal.

4. Meta-Supervised Learning

In this section, we present the analysis of meta-supervised learning (meta-SL). We first define the detailed problem setup of meta-SL and present a meta-SL algorithm. We then characterize the global optimality of the stationary point attained by such an algorithm. We refer the analysis of meta-SL with neural network parameterization to §A.3.

4.1. Problem Setup and Algorithm

In meta-SL, the meta-learner observes a sample of supervised learning subtasks $\{(\mathcal{D}_i, \ell, \mathcal{H})\}_{i \in [n]}$ drawn independently from a task distribution ι. Specifically, each subtask $(\mathcal{D}_i, \ell, \mathcal{H})$ consists of a distribution $\mathcal{D}_i$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}$, a loss function $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, and a hypothesis class $\mathcal{H}$. Each hypothesis $h \in \mathcal{H}$ is a mapping from $\mathcal{X}$ to $\mathcal{Y}$. The goal of the supervised learning subtask $(\mathcal{D}_i, \ell, \mathcal{H})$ is to obtain the following hypothesis,

$$h_i^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, R_i(h) = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}_{z \sim \mathcal{D}_i}\big[\ell(h, z)\big], \quad (4.1)$$

where $R_i(h) = \mathbb{E}_{z \sim \mathcal{D}_i}[\ell(h, z)]$ is the risk of $h \in \mathcal{H}$. To approximately attain the minimizer defined in (4.1), we parameterize the hypothesis class $\mathcal{H}$ by $\mathcal{H}_\theta$ with a feature mapping $\phi : \mathcal{X} \mapsto \mathbb{R}^d$ as follows,

$$\mathcal{H}_\theta = \big\{h_\theta(\cdot) = \phi(\cdot)^\top \theta : \theta \in \mathbb{R}^d\big\}, \quad (4.2)$$

and minimize $R_i(h_\theta)$ over $\theta \in \mathbb{R}^d$. We set the algorithm $\mathscr{A}_\theta$ in (2.7), which solves $(\mathcal{D}_i, \ell, \mathcal{H})$, to be one step of gradient descent with the starting point θ, that is,

$$\mathscr{A}_\theta(\mathcal{D}_i, \ell, \mathcal{H}) = h_{\theta - \eta \cdot \nabla_\theta R_i(h_\theta)}. \quad (4.3)$$

Here η is the learning rate of $\mathscr{A}_\theta$. The goal of meta-SL is to minimize the following meta-objective,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} R_i(h_{\theta_i}), \quad \text{where } h_{\theta_i} = \mathscr{A}_\theta(\mathcal{D}_i, \mathcal{R}, \mathcal{H}). \quad (4.4)$$

To minimize $L(\theta)$ defined in (4.4), we adopt gradient descent, which iteratively updates θ_ℓ as follows,

$$\theta_{\ell+1} \leftarrow \theta_\ell - \alpha_\ell \cdot \nabla_\theta L(\theta_\ell), \quad \text{for } \ell = 0, 1, \ldots, T-1. \quad (4.5)$$

Here $\nabla_\theta L(\theta_\ell)$ is the gradient of the meta-objective at θ_ℓ, α_ℓ is the learning rate at the ℓ-th iteration, and T is the number of iterations. (Fallah et al., 2019) show that the update defined in (4.5) converges to an ϵ-stationary point of the meta-objective L under a smoothness assumption on L. In what follows, we characterize the optimality gap of such an ϵ-stationary point.

We first introduce the Fréchet differentiability of the risk R_i in (4.1).

Definition 4.1 (Fréchet Differentiability). Let $\mathcal{H}$ be a Banach space with the norm $\|\cdot\|_{\mathcal{H}}$. A functional $R : \mathcal{H} \mapsto \mathbb{R}$ is Fréchet differentiable at $h \in \mathcal{H}$ if it holds for a bounded linear operator $A : \mathcal{H} \mapsto \mathbb{R}$ that

$$\lim_{h_1 \in \mathcal{H},\, \|h_1\|_{\mathcal{H}} \to 0} \frac{|R(h + h_1) - R(h) - A(h_1)|}{\|h_1\|_{\mathcal{H}}} \to 0. \quad (4.6)$$

We define A as the Fréchet derivative of R at $h \in \mathcal{H}$, and write

$$D_h R(\cdot) = A(\cdot). \quad (4.7)$$

In what follows, we assume that the hypothesis class $\mathcal{H}$ with the $L_2(\rho)$-inner product is a Hilbert space, where ρ is a distribution over $\mathcal{X}$. Thus, following from the definition of the Fréchet derivative in Definition 4.1 and the Rieze representation theorem (Rudin, 2006), it holds for an $a_h \in \mathcal{H}$ that

$$D_h R(\cdot) = A(\cdot) = \langle \cdot, a_h \rangle_{\mathcal{H}}, \quad (4.8)$$

Here we denote by $\langle f, g \rangle_{\mathcal{H}} = \int_{\mathcal{X}} f(x) \cdot g(x) \, d\rho$ the $L_2(\rho)$-inner product. In what follows, we write

$$(\delta R / \delta h)(x) = a_h(x), \quad \forall x \in \mathcal{X}, h \in \mathcal{H}. \quad (4.9)$$

We refer to §B for an example of the Fréchet derivative defined in (4.9). We assume that $\mathcal{H}$ contains the parameterized hypothesis class $\mathcal{H}_\theta$ defined in (4.2), and impose the following assumption on the convexity and the Fréchet differentiability of the risk R_i in (4.1).

Assumption 4.2 (Convex and Differentiable Risk). We assume for all $i \in [n]$ that the risk R_i defined in (4.1) is convex and Fréchet differentiable on $\mathcal{H}$.

Assumption 4.2 is a mild regularity condition on the risk R_i, which holds for the risks induced by commonly used loss functions, such as the squared loss and the cross entropy loss. Specifically, the convexity of R_i holds if the loss function $\ell(h, z)$ is convex in $h \in \mathcal{H}$ for all $z \in \mathcal{Z}$ (Rockafellar, 1968).

The following proposition holds under Assumption 4.2.

Proposition 4.3 (Convex and Differentiable Risk (Ekeland & Temam, 1999)). Under Assumption 4.2, it holds for all $i \in [n]$ that

$$R_i(h_1) \geq R_i(h_2) + \langle \delta R_i / \delta h_2, h_1 - h_2 \rangle_{\mathcal{H}}, \quad \forall h_1, h_2 \in \mathcal{H}.$$

Proof. See (Ekeland & Temam, 1999) for a detailed proof. $\square$

We highlight that the convexity of the risks over $h \in \mathcal{H}$ does not imply the convexity of the meta-objective defined in (4.4). In contrast, Proposition 4.3 characterizes the functional geometry of the risk R_i in the Hilbert space $\mathcal{H}$ for all $i \in [n]$, which allows us to analyze the global optimality of meta-SL in the sequel.

4.2. Theoretical Results

In this section, we characterize the global optimality of the ϵ-stationary point attained by meta-SL defined in (4.5). Let θ^* be a global minimizer of the meta-objective $L(\theta)$ defined in (4.4), and ω be the ϵ-stationary point attained by meta-SL such that

$$\nabla_\omega L(\omega)^\top v \leq \epsilon, \quad \forall v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}. \tag{4.10}$$

Our goal is to upper bound the optimality gap $L(\omega) - L(\theta^*)$. To this end, we first define the mixed distribution $\mathcal{M}$ over all the distributions $\{\mathcal{D}_i\}_{i \in [n]}$ as follows,

$$\mathcal{M}(x, y) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathcal{D}_i(x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{4.11}$$

To simplify the notation, we write ω_i and θ_i^* as the parameters that correspond to the outputs of the algorithms $\mathscr{A}_\omega(\mathcal{D}_i, \ell, \mathcal{H})$ and $\mathscr{A}_{\theta^*}(\mathcal{D}_i, \ell, \mathcal{H})$, respectively. More specifically, according to (4.3), we have

$$\omega_i = \omega - \eta \cdot \nabla_\omega R_i(h_\omega),$$
$$\theta_i^* = \theta^* - \eta \cdot \nabla_{\theta^*} R_i(h_{\theta^*}), \quad \forall i \in [n], \tag{4.12}$$

where η is the learning rate of the algorithms $\mathscr{A}_\omega(\mathcal{D}_i, \ell, \mathcal{H})$ and $\mathscr{A}_{\theta^*}(\mathcal{D}_i, \ell, \mathcal{H})$.

The following theorem characterizes the optimality gap of the ϵ-stationary point ω attained by meta-SL. We define for all $(x, y, x') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{X}$ that

$$w(x, y, x') = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\delta R_i}{\delta h_{\omega_i}}(x') \cdot \frac{d\mathcal{D}_i}{d\mathcal{M}}(x, y), \tag{4.13}$$

$$u(x, y, x') = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\delta R_i}{\delta h_{\omega_i}}(x') \cdot \frac{h_{\omega_i}(x') - h_{\theta_i^*}(x')}{w(x, y, x')}, \tag{4.14}$$

$$\phi_{\ell,\omega}(x, y, x') = \left(I_d - \eta \cdot \nabla_\omega^2 \ell\big(\phi(x)^\top \omega, (x, y)\big) \right) \phi(x'), \tag{4.15}$$

where $d\mathcal{D}_i / d\mathcal{M}$ is the Radon-Nikodym derivative and $\delta R_i / \delta h_{\omega_i}$ is the Fréchet derivative defined in (4.9).

Theorem 4.4 (Optimality Gap of ϵ-Stationary Point). Let θ^* be a global minimizer of $L(\theta)$. Also, let ω be the ϵ-stationary point defined in (4.10). Let $\ell(h_\theta(x), (x, y))$ be twice differentiable with respect to all $\theta \in \mathbb{R}^d$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Under Assumption 4.2, it holds for all $R > 0$ that

$$L(\omega) - L(\theta^*) \tag{4.16}$$
$$\leq \underbrace{R \cdot \epsilon}_{\text{(i)}} + \underbrace{\|w\|_{\mathcal{M}\cdot\rho}}_{\text{(ii)}} \cdot \underbrace{\inf_{v \in \mathcal{B}_R} \|u(\cdot) - \phi_{\ell,\omega}(\cdot)^\top v\|_{\mathcal{M}\cdot\rho}}_{\text{(iii)}},$$

where we define $\mathcal{B}_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$ as the ball with radius R and

$$\|w\|_{\mathcal{M}\cdot\rho} = \left(\int w^2(x, y, x') \, d\mathcal{M}(x, y) \, d\rho(x') \right)^{1/2}$$

as the $L_2(\mathcal{M} \cdot \rho)$-norm of w.

Proof. See §C.2 for a detailed proof. $\square$

By Theorem 4.4, the optimality gap of the ϵ-stationary point ω hinges on the three terms on the right-hand side of (4.16). Here term (i) characterizes the deviation of the ϵ-stationary

point ω from a stationary point. Term (ii) characterizes the difficulty of all the subtasks sampled from the task distribution ι. Specifically, given the ϵ-stationary point ω, if the output h_{ω_i} of $\mathscr{A}_\omega(\mathcal{D}_i, \ell, \mathcal{H})$ well approximates the minimizer of the risk R_i in (4.1), then the Fréchet derivative $\delta R_i/\delta h_{\omega_i}$ defined in (4.9) is close to zero. Meanwhile, the Radon-Nikodym derivative $\mathrm{d}\mathcal{D}_i/\mathrm{d}\mathcal{M}$ characterizes the deviation of the distribution $\mathcal{D}_i$ from the mixed distribution $\mathcal{M}$ defined in (4.11), which is upper bounded if $\mathcal{D}_i$ is close to $\mathcal{M}$. Thus, term (ii) is upper bounded if h_{ω_i} well approximates the minimizer of R_i and $\mathcal{D}_i$ is close to $\mathcal{M}$ for all $i \in [n]$. Term (iii) characterizes the representation power of the feature mapping $\phi_{\ell,\omega}$ defined in (4.15). Specifically, if the function u defined in (4.14) of Theorem 4.4 is well approximated by $\phi_{\ell,\omega}(\cdot)^\top v$ for some $v \in \mathcal{B}_R$, then term (iii) is small. In conclusion, if the subtasks generated by the task distribution ι are sufficiently regular so that term (ii) is upper bounded, and the linear class $\{\phi_{\ell,\omega}(\cdot)^\top v : v \in \mathcal{B}_R\}$ has sufficient representation power, then ω is approximately globally optimal. See §B for a corollary of Theorem 4.4 when it is adapted to the squared loss.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018a.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended PAC-Bayes theory. *arXiv preprint arXiv:1711.01244*, 2017.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, 2019.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019a.

Cao, Y. and Gu, Q. A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019b.

Chizat, L. and Bach, F. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

Daniely, A. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, 2017.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

Ekeland, I. and Temam, R. *Convex analysis and variational problems*, volume 28. SIAM, 1999.

Evgeniou, T. and Pontil, M. Regularized multi–task learning. In *International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, 2004.

Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

Fallah, A., Mokhtari, A., and Ozdaglar, A. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. *arXiv preprint arXiv:1908.10400*, 2019.

Fan, J., Ma, C., and Zhong, Y. A selective overview of deep learning. *arXiv preprint arXiv:1904.05526*, 2019.

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

Finn, C., Yu, T., Zhang, T., Abbeel, P., and Levine, S. One-shot visual imitation learning via meta-learning. *arXiv preprint arXiv:1709.04905*, 2017b.

Finn, C., Xu, K., and Levine, S. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

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

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, 2017.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Khodak, M., Balcan, M.-F., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. *arXiv preprint arXiv:1902.10644*, 2019.

Konda, V. *Actor-Critic Algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-SGD: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Mendonca, R., Gupta, A., Kralev, R., Abbeel, P., Levine, S., and Finn, C. Guided meta-policy search. In *Advances in Neural Information Processing Systems*, 2019.

Nagabandi, A., Finn, C., and Levine, S. Deep online learning via meta-learning: Continual adaptation for model-based RL. *arXiv preprint arXiv:1812.07671*, 2018.

Nichol, A. and Schulman, J. Reptile: A scalable meta-learning algorithm. *arXiv preprint arXiv:1803.02999*, 2: 2, 2018.

Nichol, A., Achiam, J., and Schulman, J. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Pentina, A. and Lampert, C. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learning*, 2014.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, 2019.

Rakelly, K., Shelhamer, E., Darrell, T., Efros, A. A., and Levine, S. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*, 2018.

Rockafellar, R. Integrals which are convex functionals. *Pacific journal of mathematics*, 24(3):525–539, 1968.

Rudin, W. *Real and complex analysis*. McGraw-Hill, 2006.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

Thrun, S. and Pratt, L. *Learning to learn*. Springer Science & Business Media, 2012.

Wang, H., Sun, R., and Li, B. Global convergence and induced kernels of gradient-based meta-learning with neural nets. *arXiv preprint arXiv: 2006.14606*, 2020.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

Wu, L., Ma, C., and Weinan, E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, 2018.

Xu, K., Ratner, E., Dragan, A., Levine, S., and Finn, C. Learning a prior over intent via meta-inverse reinforcement learning. *arXiv preprint arXiv:1805.12573*, 2018.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. Bayesian model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 2018.

Yu, T., Abbeel, P., Levine, S., and Finn, C. One-shot hierarchical imitation learning of compound visuomotor tasks. *arXiv preprint arXiv:1810.11043*, 2018.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.

A. Neural Network Prameterization

In this section, we present the global optimality analysis of meta-RL and meta-SL with the overparameterized two-layer neural network parameterization, namely neural meta-RL and neural meta-SL, respectively. Specifically, for both neural meta-RL and neural meta-SL, we show that the global optimality of the attained ϵ-stationary points hinges on the representation power of the corresponding classes of overparameterized two-layer neural networks.

A.1. Neural Network

We first introduce the neural network parameterization. For $x \in \mathbb{R}^d$, $b = (b_1, \dots, b_m) \in \mathbb{R}^m$, and $W = ([W]_1^\top, \dots, [W]_m^\top) \in \mathbb{R}^{md}$, we define

$$f(x; b, W) = \frac{1}{\sqrt{m}} \cdot \sum_{r=1}^{m} b_r \cdot \sigma\left([W]_r^\top x\right), \tag{A.1}$$

where $\sigma(x) = x \cdot \mathbb{1}\{x > 0\}$ is the rectified linear unit (ReLU).

We set m to be divisible by two and initialize the parameter W with $[W_{\text{init}}]_r = [W_{\text{init}}]_{r+m/2} \sim N(0, I_d/d)$ for $r \in [m/2]$. Meanwhile, we initialize $b_r = 1$ and $b_{r+m/2} = -1$ for all $r \in [m/2]$. Such initialization (Bai & Lee, 2019) is almost equivalent to the independent and identical initialization of $[W_{\text{init}}]_r$ for $r \in [m]$ in our analysis, and ensures that $f(x; W_{\text{init}}) = 0$ for all $x \in \mathcal{X}$. In what follows, we fix b_r for all $r \in [m]$ and only optimize over W. We write $f(x; W) = f(x; b, W)$ in the sequel for notational simplicity. Note that $f(x; W)$ is almost everywhere differentiable with respect to W, and it holds for all $x \in \mathbb{R}^d$ and $W \in \mathbb{R}^{md}$ that $\nabla_W f(x; W) = ([\nabla_W f(x; W)]_1^\top, \dots, [\nabla_W f(x; W)]_m^\top)^\top \in \mathbb{R}^{md}$, where

$$\begin{aligned} \left[\nabla_W f(x; W)\right]_r &= \left[\phi_W(x)\right]_r \\ &= \frac{b_r}{\sqrt{m}} \cdot x \cdot \mathbb{1}\left\{[W]_r^\top x > 0\right\}, \quad \forall r \in [m]. \end{aligned} \tag{A.2}$$

Here we define the feature mapping as $\phi_W(x) = ([\phi_W(x)]_1^\top, \dots, [\phi_W(x)]_m^\top)^\top$ for all $x \in \mathbb{R}^d$ and $W \in \mathbb{R}^{md}$. It then follows from the definition of $f(x; W)$ in (A.1) that $f(x; W) = \phi_W(x)^\top W$. In the sequel, we denote by $\mathbb{E}_{\text{init}}$ the expectation with respect to the random initialization of the neural network.

A.2. Neural Meta-RL

In this section, we analyze the global optimality of the ϵ-stationary point attained by meta-RL when the main effect π_θ is parameterized by the neural network defined in (A.1). Without loss of generality, we assume that $\mathcal{S} \times \mathcal{A} \subseteq \mathbb{R}^d$ and $\|(s, a)\|_2 \leq 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Similar to (3.2), we parameterize the main effect π_θ as follows,

$$\pi_\theta(a \,|\, s) = \frac{\exp\{1/\tau \cdot f((s, a); \theta)\}}{\sum_{a' \in \mathcal{A}} \exp\{1/\tau \cdot f((s, a'); \theta)\}}, \tag{A.3}$$

where $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $f(\cdot; \theta)$ is the neural network defined in (A.1) with $W = \theta$ for all $\theta \in \mathbb{R}^{md}$. Correspondingly, given the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$, the maximizer $\pi_{i,\theta}$ defined in (3.1) takes the form of

$$\pi_{i,\theta}(\cdot \,|\, s) \propto \exp\left(1/\tau \cdot f((s, \cdot); \theta) + \eta \cdot Q_i^{\pi_\theta}(s, \cdot)\right), \tag{A.4}$$

where $s \in \mathcal{S}$ and $Q_i^{\pi_\theta}$ is the action-value function of π_θ corresponding the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$. Neural meta-RL maximizes the following meta-objective via gradient ascent with W_{init} as the starting point,

$$L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} J_i(\pi_{i,\theta}), \tag{A.5}$$

where $\pi_{i,\theta}$ is defined in (A.4), and $J_i(\pi_{i,\theta})$ is the expected total reward of $\pi_{i,\theta}$ corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$. In what follows, we analyze the global optimality of the ϵ-stationary point ω of the meta-objective L attained by neural meta-RL. Specifically, we define ω as follows,

$$
\begin{aligned}
&\nabla_\omega L(\omega)^\top (v - \omega) \leq \epsilon, \\
&\forall v \in \mathcal{B}_{\text{init}} = \{\theta \in \mathbb{R}^d : \|\theta - W_{\text{init}}\|_2 \leq R_T\}.
\end{aligned}
\tag{A.6}
$$

Here W_{init} is the initial parameter, and the radius R_T is the maximum trajectory length of T gradient ascent steps.

We impose the following regularity condition on the mixed meta-visitation measure ϱ_{π_θ} defined in (3.11) of Definition 3.3.

Assumption A.1 (Regularity Condition on ϱ_{π_θ}). We assume for all $y \in \mathbb{R}^d$, $u > 0$, and $\theta \in \mathbb{R}^{md}$ that

$$
\mathbb{E}_{(s,a) \sim \varrho_{\pi_\theta}} \left[\mathbb{1}\{ |y^\top(s,a)| \leq u \} \right] \leq c \cdot u / \|y\|_2,
$$

where $c > 0$ is an absolute constant.

Assumption A.1 is imposed to rule out the corner case where ϱ_{π_θ} has a point mass at a specific state action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. Similar assumptions arise in the analysis of RL with neural network parameterization (Cai et al., 2019; Liu et al., 2019).

The following corollary characterizes the optimality gap of the ϵ-stationary point defined in (A.6). Let θ^* be a global maximizer of the meta-objective $L(\theta)$ defined in (A.5). We define

$$
c_\omega(s, a) = f((s, a); \omega) + f_\omega(s, a),
\tag{A.7}
$$

where $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $f(\cdot; \omega)$ is the neural network defined in (A.1) with $W = \omega$ and f_ω is defined in (3.14).

Corollary A.2 (Optimality Gap of ϵ-Stationary Point). Under Assumptions 3.4 and A.1, for the ϵ-stationary point ω defined in (A.6), we have

$$
\mathbb{E}_{\text{init}}\left[L(\theta^*) - L(\omega) \right] \leq \underbrace{\epsilon}_{(i)} + \underbrace{\mathcal{O}(R_T^{3/2} \cdot m^{-1/4})}_{(iii)}
\tag{A.8}
$$
$$
+ \underbrace{C \cdot \mathbb{E}_{\text{init}}\left[\inf_{v \in \mathcal{B}_{\text{init}}} \left\| c_\omega(\cdot, \cdot) - f((\cdot, \cdot); v) \right\|_{\varrho_{\pi_\omega}} \right]}_{(ii)}
$$

where $C = 2C_0 \cdot Q_{\max}/\tau \cdot (1 + 2Q_{\max} \cdot \overline{\gamma} \cdot \eta)$, $\overline{\gamma} = (\sum_{i=1}^n \gamma_i)/n$, C_0 is defined in Assumption 3.4, and $\mathcal{B}_{\text{init}}$ is the parameter space defined in (A.6).

Proof. See §C.4 for a detailed proof. ∎

By Corollary A.2, the global optimality of the ϵ-stationary point ω is upper bounded by the three terms on the right-hand side of (A.8). Here term (i) characterizes the deviation of ω from a stationary point. Term (ii) characterizes the representation power of neural networks. Specifically, if the function c_ω defined in (A.7) is well approximated by the neural network defined in (A.1) with a parameter from the parameter space $\mathcal{B}_{\text{init}}$, then term (ii) is small. Term (iii) is the linearization error of the neural networks, which characterizes the deviation of a neural network from its first-order Taylor expansion at the initial parameter W_{init}. Such an error is small for a sufficiently large width m, that is, if the neural network is overparameterized. In conclusion, if the class of overparameterized two-layer neural networks with the parameter space $\mathcal{B}_{\text{init}}$ has sufficient representation power, then the ϵ-stationary point ω attained by neural meta-RL is approximately globally optimal.

A.3. Neural Meta-SL

In this section, we analyze the global optimality of the ϵ-stationary point attained by neural meta-SL associated with the squared loss, where we parameterize the hypothesis $h_\theta(\cdot) = f(\cdot; \theta)$ by the neural network defined in (A.1). Specifically, neural meta-SL minimizes the meta-objective L defined in (4.4) via gradient descent defined in (4.5) with W_{init} as the starting point. We analyze the global optimality of the ϵ-stationary point ω attained by neural meta-SL, which is defined as follows,

$$
\nabla_\omega L(\omega)^\top (\omega - v) \leq \epsilon,
$$
$$
\forall v \in \mathcal{B}_{\mathrm{init}} = \{\theta \in \mathbb{R}^d : \|\theta - W_{\mathrm{init}}\|_2 \leq R_T\}. \tag{A.9}
$$

Here R_T is the maximum trajectory length of T gradient descent steps. In what follows, we set $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. Similar to Assumption A.1, we impose the following regularity condition on the distribution ρ that defines the Hilbert space in (4.8).

Assumption A.3 (Regularity Condition on ρ). We assume for an absolute constant $c > 0$ that

$$
\mathbb{E}_{x \sim \rho}\Big[\mathbb{1}\big\{|x^\top y| \leq u\big\}\Big] \leq c \cdot u/\|y\|_2, \quad \forall y \in \mathbb{R}^d, u > 0.
$$

Such an assumption holds if the probability density function of ρ is upper bounded by an absolute constant. Under Assumption A.3, the following corollary characterizes the optimality gap of the ϵ-stationary point ω defined in (A.9). We define

$$
K_{\omega,\eta} = \mathbb{E}_{x \sim \rho}\big[I_{md} - 2\eta \cdot \phi_\omega(x)\phi_\omega(x)^\top\big],
$$
$$
\mathcal{B}_0 = K_{\omega,\eta}(\omega - \mathcal{B}_{\mathrm{init}}) + W_{\mathrm{init}}, \tag{A.10}
$$
$$
\overline{u}(x) = f(x; W_{\mathrm{init}}) + \sum_{i=1}^n \Big(\frac{\delta R_i}{\delta h_{\omega_i}}(x) \cdot \big(h_{\omega_i}(x) - h_{\theta_i^*}(x)\big)\Big)
$$
$$
\Big/ \Big(\sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x)\Big), \tag{A.11}
$$

where $f(\cdot; W_{\mathrm{init}})$ and ϕ_ω are the neural network and the feature mapping defined in (A.1) with $W = W_{\mathrm{init}}$ and (A.2) with $W = \omega$, respectively, $\mathcal{B}_{\mathrm{init}}$ is the parameter space defined in (A.9), W_{init} is the initial parameter, and ω_i, θ_i^* are the parameters defined in (4.12). We further define the average risk $\overline{R}$ as follows,

$$
\overline{R} = \frac{1}{n} \cdot \sum_{i=1}^n R_i^{1/2}(h_{\omega_i})
$$
$$
= \frac{1}{n} \cdot \sum_{i=1}^n \Big\{\mathbb{E}_{(x,y) \sim \mathcal{D}_i}\Big[\big(y - h_{\omega_i}(x)\big)^2\Big]\Big\}^{1/2}. \tag{A.12}
$$

Corollary A.4 (Optimality Gap of ϵ-Stationary Point). We denote by $\overline{\mathcal{D}}_i$ the marginal distribution of $\mathcal{D}_i$ over $\mathcal{X}$. Let $\overline{\mathcal{D}}_i = \rho$ for all $i \in [n]$ and $|y| \leq Y_{\max}$ for all $y \in \mathcal{Y}$. Under Assumptions 4.2 and A.3, for the squared loss $\ell(h, (x, y)) = (h(x) - y)^2$ and ω defined in (A.9), we have

$$
\mathbb{E}_{\mathrm{init}}\big[L(\omega) - L(\theta^*)\big] \leq \epsilon + \mathbb{E}_{\mathrm{init}}\Big[2\overline{R} \cdot \inf_{v \in \mathcal{B}_0} \|\overline{u}(\cdot) - f(\cdot; v)\|_\rho\Big]
$$
$$
+ \mathcal{O}(G_T^{3/2} \cdot m^{-1/4}), \tag{A.13}
$$

where $G_T = (1 + \eta) \cdot R_T + \eta \cdot Y_{\max}$, R_T is the maximum trajectory length in (A.9), and η is the learning rate of $\mathscr{A}_\omega$ in

(4.12).

Proof. See §C.5 for a detailed proof. □

B. Meta-SL with Squared Loss

In this section, we analyze the global optimality of meta-SL with the squared loss. The optimality gap characterized in Theorem 4.4 has a more straightforward interpretation when restricted to meta-SL with the squared loss, which is defined as

$$\ell\big(h, (x, y)\big) = \big(h(x) - y\big)^2, \quad \forall h \in \mathcal{H}, \ (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{B.1}$$

The following proposition calculates the Fréchet derivative $\delta R_i/\delta h$ defined in (4.9) for the squared loss.

Proposition B.1. We denote by $\overline{\mathcal{D}}_i$ the marginal distribution of $\mathcal{D}_i$ over $\mathcal{X}$. Let $\overline{\mathcal{D}}_i = \rho$ for all $i \in [n]$. For the squared loss ℓ defined in (B.1) and $R_i = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\ell(h, (x, y))]$, it holds that

$$(\delta R_i/\delta h)(x') = 2\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[h(x) - y \mid x = x'\big], \quad \forall h \in \mathcal{H}, \ x' \in \mathcal{X}. \tag{B.2}$$

Proof. See §D.3 for a detailed proof. □

By Proposition B.1, we obtain from Jensen's inequality that

$$\|\delta R_i/\delta h_{\omega_i}\|_\rho^2 \leq 4\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\Big[\big(h_{\omega_i}(x) - y\big)^2\Big] = 4R_i(h_{\omega_i}).$$

Meanwhile, recall that the function w defined in (4.13) is a weighted average over the Fréchet derivatives $\{\delta R_i/\delta h_{\omega_i}\}_{i\in[n]}$. Hence, the $L_2(\rho)$-norm of the function w characterizes the difficulty of subtasks by aggregating the risks $R_i(h_{\omega_i})$.

The following corollary characterizes the the optimality gap of the ϵ-stationary point ω attained by meta-SL, which is defined in (4.10). We define

$$K_\eta = \mathbb{E}_{x\sim\rho}\big[I_d - 2\eta \cdot \phi(x)\phi(x)^\top\big], \tag{B.3}$$

$$u(x') = \left(\sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(h_{\omega_i}(x') - h_{\theta_i^*}(x')\big)\right) \Big/ \left(\sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x')\right), \tag{B.4}$$

$$\overline{R} = \frac{1}{n} \cdot \sum_{i=1}^n R_i^{1/2}(h_{\omega_i}) = \frac{1}{n}\sum_{i=1}^n \Big\{\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\Big[\big(y - h_{\omega_i}(x)\big)^2\Big]\Big\}^{1/2}, \tag{B.5}$$

where ω_i and θ_i^* are the parameters defined in (4.12) and η is the learning rate of $\mathscr{A}_\omega$ in (4.12).

Corollary B.2. We denote by $\overline{\mathcal{D}}_i$ the marginal distribution of $\mathcal{D}_i$ over $\mathcal{X}$. Let $\overline{\mathcal{D}}_i = \rho$ for all $i \in [n]$. Under Assumption 4.2, for the squared loss ℓ defined in (B.1) and $R > 0$, we have

$$L(\omega) - L(\theta^*) \leq R \cdot \epsilon + 2\overline{R} \cdot \inf_{v\in\mathcal{B}} \|u - (K_\eta \cdot \phi)^\top (R \cdot v)\|_\rho.$$

Proof. See §C.3 for a detailed proof. □

By Corollary B.2, the optimality gap $L(\omega) - L(\theta^*)$ hinges on the average risk $\overline{R}$, the representation power of the feature ϕ, and the kernel K_η defined in (B.3). Note that

$$\inf_{v\in\mathcal{B}} \|u - (K_\eta \cdot \phi)^\top (R \cdot v)\|_\rho = \inf_{v\in R\cdot K_\eta\cdot\mathcal{B}} \|u - \phi^\top v\|_\rho,$$

where we write $R \cdot K_\eta \cdot \mathcal{B} = \{v \in \mathbb{R}^d : v = R \cdot K_\eta \cdot u, u \in \mathcal{B}\}$. Hence, if $\overline{R}$ is upper bounded and $\phi(\cdot)^\top \theta$ well approximates the function u defined in (B.4) for a parameter $\theta \in R \cdot K_\eta \cdot \mathcal{B}$, then the ϵ-stationary point ω attained by meta-SL is approximately globally optimal.

C. Proof of Main Result

In this section, we present the proofs of the main results.

C.1. Proof of Theorem 3.5

Proof. By Lemma E.3, it holds for all $i \in [n]$ that

$$J_i(\pi_{i,\theta^*}) - J_i(\pi_{i,\omega}) = (1 - \gamma_i)^{-1} \cdot \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta^*}}} \left[A_i^{\pi_{i,\omega}}(s, a) \right], \tag{C.1}$$

where $A_i^{\pi_{i,\omega}}$ and $\sigma_{\pi_{i,\theta^*}}$ are the advantage function and the state-action visitation measure of the policies π_{i,θ^*} and $\pi_{i,\omega}$, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$. Meanwhile, note that

$$\nabla_\omega L(\omega)^\top v \leq \epsilon, \quad \forall v \in \mathbb{R}^d, \ \|v\|_2 \leq 1. \tag{C.2}$$

Thus, combining (C.1) and (C.2), we obtain for all $v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ and $R > 0$ that

$$\begin{aligned}
L(\theta^*) - L(\omega) &\leq R \cdot \epsilon - R \cdot \nabla_\omega L(\omega)^\top v + \frac{1}{n} \cdot \sum_{i=1}^n J_i(\pi_{i,\theta^*}) - J_i(\pi_{i,\omega}) \\
&= R \cdot \epsilon - R \cdot \nabla_\omega L(\omega)^\top v + \frac{1}{n} \cdot \sum_{i=1}^n (1 - \gamma_i)^{-1} \cdot \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta^*}}} \left[A_i^{\pi_{i,\omega}}(s, a) \right].
\end{aligned} \tag{C.3}$$

In what follows, we upper bound the right-hand side of (C.3). By Proposition 3.2, we have the following lemma, which calculates $\nabla_\omega L(\omega)$.

Lemma C.1. It holds for all $\theta \in \mathbb{R}^d$ that

$$\nabla_\theta L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{E}_{(s',a') \sim \varsigma_{i,\pi_\theta}} \left[g_{i,\theta}(s', a') \cdot \phi(s', a') \right], \tag{C.4}$$

where ς_{i,π_θ} is the meta-visitation measure defined in (3.10) of Definition 3.3 and

$$g_{i,\theta}(s', a') = 1/\tau \cdot A_i^{\pi_{i,\theta}}(s', a') \cdot (\mathrm{d}\sigma_{\pi_{i,\theta}} / \mathrm{d}\varsigma_{i,\pi_\theta})(s', a') + \gamma_i \cdot \eta/\tau \cdot G_{i,\pi_\theta}(s', a') \cdot A_i^{\pi_\theta}(s', a'). \tag{C.5}$$

Here $A_i^{\pi_{i,\theta}}$ and $A_i^{\pi_\theta}$ are the advantage functions of $\pi_{i,\theta}$ and π_θ, respectively, corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$, and G_{i,π_θ} is defined as follows,

$$G_{i,\pi_\theta}(s', a') = \mathbb{E}_{(s',a',s,a) \sim \rho_{i,\pi_\theta}} \left[A_i^{\pi_{i,\theta}}(s, a) \,\middle|\, s', a' \right], \tag{C.6}$$

where ρ_{i,π_θ} is the joint meta-visitation measure defined in (3.9) of Definition 3.3.

Proof. See §D.2 for a detailed proof. □

By Lemma C.1, we obtain for all $v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ that

$$\nabla_\omega L(\omega)^\top v = \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{E}_{(s',a') \sim \varsigma_{i,\pi_\omega}} \left[g_{i,\omega}(s',a') \cdot \phi(s',a')^\top v \right], \tag{C.7}$$

where $g_{i,\omega}$ is defined in (C.5) of Lemma C.1 with $\theta = \omega$. By plugging (C.7) into (C.3), we obtain that

$$
\begin{aligned}
L(\theta^*) - L(\omega) &\leq R \cdot \epsilon + \frac{1}{n} \cdot \sum_{i=1}^n (1 - \gamma_i)^{-1} \cdot \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta^*}}} \left[A_i^{\pi_{i,\omega}}(s,a) \right] \\
&\qquad - R \cdot \mathbb{E}_{(s',a') \sim \varsigma_{i,\pi_\omega}} \left[g_{i,\omega}(s',a') \cdot \phi(s',a')^\top v \right] \\
&\leq R \cdot \epsilon + \mathbb{E}_{(s',a') \sim \varrho_{\pi_\omega}} \Bigg[\frac{1}{n} \cdot \sum_{i=1}^n (1 - \gamma_i)^{-1} \cdot A_i^{\pi_{i,\omega}}(s',a') \cdot \frac{\mathrm{d}\sigma_{\pi_{i,\theta^*}}}{\mathrm{d}\varrho_{\pi_\omega}}(s',a') \\
&\qquad - \frac{1}{n} \cdot \sum_{i=1}^n g_{i,\omega}(s',a') \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}}(s',a') \cdot \phi(s',a')^\top (R \cdot v) \Bigg],
\end{aligned}
\tag{C.8}
$$

where $\mathrm{d}\sigma_{i,\pi_{\theta^*}} / \mathrm{d}\varrho_{\pi_\omega}$ and $\mathrm{d}\varsigma_{i,\pi_\omega} / \mathrm{d}\varrho_{\pi_\omega}$ are the Radon-Nikodym derivatives, and ϱ_{π_ω} is the mixed meta-visitation measure defined in (3.11) of Definition 3.3. By the Cauchy-Schwartz inequality, we obtain from (C.8) that

$$L(\theta^*) - L(\omega) \leq R \cdot \epsilon + \underbrace{\left\| \frac{1}{n} \cdot \sum_{i=1}^n g_{i,\omega} \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}} \right\|_{\varrho_{\pi_\omega}}}_{H} \cdot \|f_\omega(\cdot,\cdot) - \phi(\cdot,\cdot)^\top (R \cdot v)\|_{\varrho_{\pi_\omega}}, \tag{C.9}$$

where we define for all $(s',a') \in \mathcal{S} \times \mathcal{A}$ that

$$f_\omega(s',a') = \left(\frac{1}{n} \cdot \sum_{i=1}^n \frac{A_i^{\pi_{i,\omega}}(s',a')}{1 - \gamma_i} \cdot \frac{\mathrm{d}\sigma_{\pi_{i,\theta^*}}}{\mathrm{d}\varrho_{\pi_\omega}}(s',a') \right) \Big/ \left(\frac{1}{n} \cdot \sum_{i=1}^n g_{i,\omega}(s',a') \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}}(s',a') \right). \tag{C.10}$$

It remains to upper bound the term H in (C.9). By the definition of $g_{i,\omega}$ in (C.5) of Lemma C.1, we have

$$
\begin{aligned}
g_{i,\omega}(s',a') \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}}(s',a') &= 1/\tau \cdot A_i^{\pi_{i,\omega}}(s',a') \cdot \frac{\sigma_{\pi_{i,\omega}}}{\varrho_{\pi_\omega}}(s',a') \\
&\qquad + \gamma_i \cdot \eta/\tau \cdot A_{i,\omega}(s',a') \cdot A_i^{\pi_\omega}(s',a') \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}}(s',a'),
\end{aligned}
\tag{C.11}
$$

which holds for all $i \in [n]$ and $(s',a') \in \mathcal{S} \times \mathcal{A}$. By the assumption that the rewards are upper bounded by $Q_{\max}$, we have

$$|A_i^{\pi_\omega}(s,a)| \leq 2Q_{\max}, \quad |A_i^{\pi_{i,\omega}}(s,a)| \leq 2Q_{\max}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \ i \in [n]. \tag{C.12}$$

Meanwhile, by the definition of $A_{i,\omega}$ in (C.6) of Lemma C.1, we further have

$$|A_{i,\omega}(s',a')| \leq \mathbb{E}_{(s',a',s,a) \sim \rho_{i,\pi_\omega}} \left[|A_i^{\pi_{i,\omega}}(s,a)| \, \big| \, s',a' \right] \leq 2Q_{\max}, \quad \forall (s',a') \in \mathcal{S} \times \mathcal{A}, \ i \in [n]. \tag{C.13}$$

Combining (C.11), (C.12), and (C.13), we have

$$\left\| g_{i,\omega} \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}} \right\|_{\varrho_{\pi_\omega}} \leq 2Q_{\max}/\tau \cdot \left\| \frac{\mathrm{d}\sigma_{\pi_{i,\omega}}}{\mathrm{d}\varrho_{\pi_\omega}} \right\|_{\varrho_{\pi_\omega}} + 4Q_{\max}^2 \cdot \gamma_i \cdot \eta/\tau \cdot \left\| \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}} \right\|_{\varrho_{\pi_\omega}}, \quad \forall i \in [n]. \tag{C.14}$$

Thus, following from Assumption 3.4 and (C.14), we obtain that

$$\left\| \frac{1}{n} \cdot \sum_{i=1}^n g_{i,\omega} \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}} \right\|_{\varrho_{\pi_\omega}} \leq \frac{1}{n} \cdot \sum_{i=1}^n \left\| g_{i,\omega} \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}} \right\|_{\varrho_{\pi_\omega}} \leq 2C_0 \cdot Q_{\max}/\tau \cdot (1 + 2Q_{\max} \cdot \overline{\gamma} \cdot \eta), \tag{C.15}$$

where we define $\overline{\gamma} = (\sum_{i=1}^n \gamma_i)/n$. Finally, by plugging (C.15) into (C.9), we have for all $v \in \mathcal{B}$ that

$$L(\theta^*) - L(\omega) \leq R \cdot \epsilon + 2C_0 \cdot Q_{\max}/\tau \cdot (1 + 2Q_{\max} \cdot \overline{\gamma} \cdot \eta) \cdot \|f_\omega(\cdot, \cdot) - \phi(\cdot, \cdot)^\top (R \cdot v)\|_{\varrho_{\pi_\omega}}, \tag{C.16}$$

where f_ω is defined in (C.10) and $\overline{\gamma} = (\sum_{i=1}^n \gamma_i)/n$. By taking the infimum over $v \in \mathcal{B}$ on the right-hand side of (C.16), we complete the proof of Theorem 3.5. □

C.2. Proof of Theorem 4.4

Proof. By Assumption 4.2 and Proposotion 4.3, we have

$$R_i(h_{\theta_2}) - R_i(h_{\theta_1}) \leq \langle \delta R_i/\delta h_{\theta_2}, h_{\theta_2} - h_{\theta_1} \rangle_{\mathcal{H}}, \quad \forall i \in [n], \theta_1, \theta_2 \in \mathbb{R}^d. \tag{C.17}$$

Meanwhile, by the definition of meta-objective in (4.4), we have

$$L(\omega) - L(\theta^*) = \frac{1}{n} \cdot \sum_{i=1}^n R_i(h_{\omega_i}) - R_i(h_{\theta_i^*}). \tag{C.18}$$

Recall that ω_i and θ_i^* are defined as follow,

$$\omega_i = \omega - \eta \cdot \nabla_\omega R_i(h_\omega), \quad \theta_i^* = \theta^* - \eta \cdot \nabla_{\theta^*} R_i(h_{\theta^*}), \quad \forall i \in [n].$$

By plugging (C.17) into (C.18) with $\theta_1 = \theta_i^*$ and $\theta_2 = \omega_i$ for $i \in [n]$, we have

$$L(\omega) - L(\theta^*) \leq \frac{1}{n} \cdot \sum_{i=1}^n \langle \delta R_i/\delta h_{\omega_i}, h_{\omega_i} - h_{\theta_i^*} \rangle_{\mathcal{H}}. \tag{C.19}$$

Thus, combining (C.19) and the definition of the ϵ-stationary point ω in (A.6), we have

$$L(\omega) - L(\theta^*) \leq R \cdot \epsilon - \nabla_\omega L(\omega)^\top (R \cdot v) + \frac{1}{n} \cdot \sum_{i=1}^n \langle \delta R_i/\delta h_{\omega_i}, h_{\omega_i} - h_{\theta_i^*} \rangle_{\mathcal{H}}, \tag{C.20}$$

which holds for all $R > 0$ and $v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$.

It suffices to upper bound the right-hand side of (C.20). To this end, we first compute the gradient $\nabla_\omega L(\omega)$. By the chain rule, we obtain for all R_i defined in (4.1) and $v \in \mathbb{R}^d$ that

$$\nabla_\omega R_i(h_{\omega_i})^\top v = \langle \delta R_i/\delta h_{\omega_i}, (\mathrm{d}h_{\omega_i}/\mathrm{d}\omega)^\top v \rangle_{\mathcal{H}}, \quad \forall i \in [n]. \tag{C.21}$$

Meanwhile, by the definition of ω_i in (4.12) and the parameterization of the hypothesis defined in (4.2), we obtain from the chain rule that

$$(\mathrm{d}h_{\omega_i}/ \mathrm{d}\omega)(\cdot) = \big(I_d - \eta \cdot \nabla_\omega^2 R_i(h_\omega)\big)\phi(\cdot), \tag{C.22}$$

where I_d is the identity matrix of shape $d \times d$. By the Leibniz integral rule, we have

$$\nabla_\omega^2 R_i(h_\omega) = \nabla_\omega^2 \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\Big[\ell\big(\phi(x)^\top \omega, (x,y)\big)\Big] = \int_{\mathcal{X}\times\mathcal{Y}} \nabla_\omega^2 \ell\big(\phi(x)^\top\omega, (x,y)\big)\, \mathrm{d}\mathcal{D}_i(x,y). \tag{C.23}$$

In what follows, we write

$$N(\omega, x, y) = I_d - \eta \cdot \nabla_\omega^2 \ell\big(\phi(x)^\top\omega, (x,y)\big) \tag{C.24}$$

for notational simplicity. By plugging (C.22) and (C.23) into (C.21), we obtain that

$$\nabla_\omega R_i(h_{\omega_i})^\top v = \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(N(\omega, x, y)\phi(x')\big)^\top v\, \mathrm{d}\mathcal{D}_i(x,y)\, \mathrm{d}\rho(x'). \tag{C.25}$$

Thus, by the definition of meta-objective in (4.4), we have for all $v \in \mathbb{R}^d$ that

$$\begin{aligned} \nabla_\omega L(\omega)^\top v &= \frac{1}{n} \cdot \sum_{i=1}^n \nabla_\omega R_i(h_{\omega_i})^\top v \\ &= \frac{1}{n} \cdot \sum_{i=1}^n \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(N(\omega, x, y)\phi(x')\big)^\top v\, \mathrm{d}\mathcal{D}_i(x,y)\, \mathrm{d}\rho(x'). \end{aligned} \tag{C.26}$$

By plugging (C.26) into (C.20), we have

$$\begin{aligned} &L(\omega) - L(\theta^*) \\ &\leq R \cdot \epsilon + \int_{\mathcal{X}} \frac{1}{n} \cdot \sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(h_{\omega_i}(x') - h_{\theta_i^*}(x')\big)\, \mathrm{d}\rho(x') \\ &\quad - \frac{1}{n} \cdot \sum_{i=1}^n \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(N(\omega, x, y)\phi(x')\big)^\top (R \cdot v)\, \mathrm{d}\mathcal{D}_i(x,y)\, \mathrm{d}\rho(x'), \end{aligned} \tag{C.27}$$

which holds for all $v \in \mathcal{B}$. Meanwhile, it holds for all $i \in [n]$ that

$$\begin{aligned} &\int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(N(\omega, x, y)\phi(x')\big)^\top (R \cdot v)\, \mathrm{d}\mathcal{D}_i(x,y)\, \mathrm{d}\rho(x') \\ &= \int_{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot (\mathrm{d}\mathcal{D}_i/\mathrm{d}\mathcal{M})(x, y) \cdot \big(N(\omega, x, y)\phi(x')\big)^\top (R \cdot v)\, \mathrm{d}\mathcal{M}(x, y)\, \mathrm{d}\rho(x'), \end{aligned} \tag{C.28}$$

where $\mathcal{M}$ is the mixed distribution defined in (4.11) and $\mathrm{d}\mathcal{D}_i/\mathrm{d}\mathcal{M}$ is the Radon-Nikodym derivative. Thus, by plugging

(C.28) into (C.27), we obtain for all $R > 0$ and $v \in \mathcal{B}$ that

$$L(\omega) - L(\theta^*)$$
$$\leq R \cdot \epsilon + \int_{\mathcal{X}} \frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot \left(h_{\omega_i}(x') - h_{\theta_i^*}(x')\right) \mathrm{d}\rho(x') \tag{C.29}$$
$$- \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{X}} \frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot (\mathrm{d}\mathcal{D}_i / \mathrm{d}\mathcal{M})(x, y) \cdot \left(N(\omega, x, y)\phi(x')\right)^{\top} (R \cdot v) \, \mathrm{d}\mathcal{M}(x, y) \, \mathrm{d}\rho(x')$$
$$\leq R \cdot \epsilon + \|w\|_{\mathcal{M} \cdot \rho} \cdot \left(\int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{X}} u(x, y, x') - \left(N(\omega, x, y)\phi(x')\right)^{\top} (R \cdot v) \, \mathrm{d}\mathcal{M}(x, y) \, \mathrm{d}\rho(x')\right)^{1/2},$$

where the second inequality follows from the Cauchy-Schwartz inequality and N is defined in (C.24). Here we define

$$w(x, y, x') = \frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot (\mathrm{d}\mathcal{D}_i / \mathrm{d}\mathcal{M})(x, y),$$

$$u(x, y, x') = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot \left(h_{\omega_i}(x') - h_{\theta_i^*}(x')\right)\right) \bigg/ w(x, y, x'),$$

and we define $\|w\|_{\mathcal{M} \cdot \rho} = (\int w^2(x, y, x') \, \mathrm{d}\mathcal{M}(x, y) \, \mathrm{d}\rho(x'))^{1/2}$ the $L_2(\mathcal{M} \cdot \rho)$-norm of w. Thus, by taking the infimum on the right-hand side of (C.29) over $v \in \mathcal{B}$ and setting $\phi_{\ell,\omega}(x, y, x') = N(\omega, x, y)\phi(x')$ for all $(x, y, x') \in \mathcal{X} \times \mathcal{Y} \times \mathcal{X}$, we complete the proof of Theorem 4.4. □

C.3. Proof of Corollary B.2

Proof. The proof is similar to that of Theorem 4.4 in §C.2. By (C.20) in the proof of Theorem 4.4, we have

$$L(\omega) - L(\theta^*) \leq R \cdot \epsilon - \nabla_\omega L(\omega)^{\top} (R \cdot v) + \frac{1}{n} \cdot \sum_{i=1}^{n} \langle \delta R_i / \delta h_{\omega_i}, h_{\omega_i} - h_{\theta_i^*} \rangle_{\mathcal{H}}. \tag{C.30}$$

In what follows, we upper bound the right-hand side of (C.30). To this end, we first compute the gradient $\nabla_\omega L(\omega)$. By (C.25) in the proof of Theorem 4.4 in §C.2, we have

$$\nabla_\omega R_i(h_{\omega_i})^{\top} v = \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \left(N(\omega, x, y)\phi(x')\right)^{\top} v \, \mathrm{d}\mathcal{D}_i(x, y) \, \mathrm{d}\rho(x'), \tag{C.31}$$

where

$$N(\omega, x, y) = I_d - \eta \cdot \nabla_\omega^2 \ell\left(\phi(x)^{\top}\omega, (x, y)\right). \tag{C.32}$$

Note that for $\ell(h, (x, y)) = (y - h(x))^2$, we have

$$\nabla_\omega^2 \ell\left(\phi(x)^{\top}\omega, (x, y)\right) = 2\phi(x)\phi(x)^{\top}, \tag{C.33}$$

which does not depend on y. Thus, for $\overline{\mathcal{D}}_i(x) = \int_{\mathcal{Y}} \mathcal{D}_i(x, y) \, \mathrm{d}y = \rho(x)$, we obtain from (C.32) and (C.33) that

$$\int_{\mathcal{X} \times \mathcal{Y}} N(\omega, x, y) \, \mathrm{d}\mathcal{D}_i(x, y) = I_d - \eta \cdot \int_{\mathcal{X}} 2\phi(x)\phi(x)^{\top} \, \mathrm{d}\rho(x) = \mathbb{E}_{x \sim \rho}\left[I_d - 2\eta \cdot \phi(x)\phi(x)^{\top}\right]. \tag{C.34}$$

By further plugging (C.34) into (C.31), we obtain that

$$\nabla_\omega R_i(h_{\omega_i})^\top v = \int_{\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(K_\eta \phi(x')\big)^\top v \, d\rho(x'), \tag{C.35}$$

where we define

$$K_\eta = \mathbb{E}_{x \sim \rho}\big[I_d - 2\eta \cdot \phi(x)\phi(x)^\top\big]. \tag{C.36}$$

Thus, by the definition of the meta-objective in (4.4), it holds for all $v \in \mathbb{R}^d$ that

$$\nabla_\theta L(\omega)^\top v = \frac{1}{n} \cdot \sum_{i=1}^n \nabla_\omega R_i(h_{\omega_i})^\top v = \frac{1}{n} \cdot \sum_{i=1}^n \int_{\mathcal{X}} (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(K_\eta \phi(x')\big)^\top v \, d\rho(x'), \tag{C.37}$$

where K_η is defined in (C.36). By plugging (C.37) into (C.30), we have

$$L(\omega) - L(\theta^*)$$
$$\leq R \cdot \epsilon + \int_{\mathcal{X}} \frac{1}{n} \cdot \sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(h_{\omega_i}(x') - h_{\theta_i^*}(x')\big) \, d\rho(x')$$
$$- \int_{\mathcal{X}} \frac{1}{n} \cdot \sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(K_\eta \phi(x')\big)^\top (R \cdot v) \, d\rho(x'), \tag{C.38}$$

which holds for all $v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$. By the Cauchy-Schwartz inequality, we obtain from (C.38) that

$$L(\omega) - L(\theta^*) \leq R \cdot \epsilon + \|w\|_\rho \cdot \|u - (K_\eta \phi)^\top (R \cdot v)\|_\rho, \tag{C.39}$$

which holds for all $v \in \mathcal{B}$. Here we define for all $x' \in \mathcal{X}$ that

$$w(x') = \frac{1}{n} \cdot \sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x'),$$
$$u(x') = \bigg(\frac{1}{n} \cdot \sum_{i=1}^n (\delta R_i/\delta h_{\omega_i})(x') \cdot \big(h_{\omega_i}(x') - h_{\theta_i^*}(x')\big)\bigg) \bigg/ w(x'). \tag{C.40}$$

It remains to upper bound the norm $\|w\|_\rho$, where w is defined in (C.40). By Proposition B.1, it holds for all $x' \in \mathcal{X}$ that

$$(\delta R_i/\delta h)(x') = 2\mathbb{E}_{(x,y) \sim \mathcal{D}_i}\big[h(x) - y \mid x = x'\big]. \tag{C.41}$$

Thus, by the fact that $\overline{\mathcal{D}}_i(x) = \int_{\mathcal{Y}} \mathcal{D}_i(x,y) \, dy = \rho(x)$, we obtain from (C.41) that

$$\|\delta R_i/\delta h\|_\rho^2 = 4 \int \Big\{\mathbb{E}_{(x,y) \sim \mathcal{D}_i}\big[h(x) - y \mid x = x'\big]\Big\}^2 d\rho(x')$$
$$\leq 4 \int \big(h(x) - y\big)^2 d\mathcal{D}_i(y \mid x) \, d\rho(x)$$
$$= 4\mathbb{E}_{(x,y) \sim \mathcal{D}_i}\Big[\big(y - h(x)\big)^2\Big], \tag{C.42}$$

where the second inequality follows from Jensen's inequality, and we denote by $\mathcal{D}_i(y \mid x)$ the conditional distribution of y given x for $(x, y) \sim \mathcal{D}_i$. Thus, following from (C.42) and the definition of w in (C.40), we obtain that

$$\|w\|_\rho \leq \frac{1}{n} \sum_{i=1}^n \|\delta R_i / \delta h_{\omega_i}\|_\rho \leq \frac{2}{n} \cdot \sum_{i=1}^n \left\{ \mathbb{E}_{(x,y) \sim \mathcal{D}_i} \left[(y - h_{\omega_i}(x))^2 \right] \right\}^{1/2}. \tag{C.43}$$

Finally, by plugging (C.43) into (C.40), we have

$$L(\omega) - L(\theta^*) \leq R \cdot \epsilon + 2\overline{R} \cdot \|u - (K_\eta \phi)^\top (R \cdot v)\|_\rho, \tag{C.44}$$

which holds for all $v \in \mathcal{B}$. Here we define K_η and u in (C.36) and (C.40), respectively, and we define $\overline{R}$ as follows,

$$\overline{R} = \frac{1}{n} \sum_{i=1}^n R_i^{1/2}(h_{\omega_i}) = \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}_{(x,y) \sim \mathcal{D}_i} \left[(y - h_{\omega_i}(x))^2 \right] \right\}^{1/2}.$$

Thus, by taking the infimum over $v \in \mathcal{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ on the right-hand side of (C.44), we complete the proof of Corollary B.2. $\qquad \square$

C.4. Proof of Corollary A.2

Proof. The proof hinges on the following lemma, which is adapted from (Cai et al., 2019).

Lemma C.2 (Linearization Error (Cai et al., 2019)). Under Assumption A.1, it holds for all $\omega_0, \omega_1, \omega_2 \in \mathcal{B} = \{\theta \in \mathbb{R}^{md} : \|\theta - W_{\text{init}}\|_2 \leq R\}$ that

$$\mathbb{E}_{\text{init}} \left[\|\phi_{\omega_0}(\cdot, \cdot)^\top \omega_2 - \phi_{\omega_1}(\cdot, \cdot)^\top \omega_2\|_{\varrho_{\pi_\theta}}^2 \right] = \mathcal{O}(R^3 \cdot m^{-1/2}),$$

where ϱ_{π_θ} is the mixed visitation measure defined in (3.11) of Definition 3.3.

Proof. See §D.4 for a detailed proof. $\qquad \square$

It holds that $\nabla_\omega f((s, a); \omega) = \phi_\omega(s, a)$ almost everywhere for $(s, a) \in \mathcal{S} \times \mathcal{A}$. Here ϕ_ω is the feature mapping defined in (A.2) with $W = \omega$. Hence, following from similar analysis to that in the proof of Theorem 3.5 in §C.1, we obtain that

$$L(\theta^*) - L(\omega) \leq \epsilon + C \cdot \|f_\omega(\cdot, \cdot) - \phi_\omega(\cdot, \cdot)^\top (v - \omega)\|_{\varrho_{\pi_\omega}}, \tag{C.45}$$

which holds for all $v \in \mathcal{B}_{\text{init}}$. Here $C = 2C_0 \cdot Q_{\max}/\tau \cdot (1 + 2Q_{\max} \cdot \overline{\gamma} \cdot \eta)$, C_0 is defined in Assumption 3.4, and $\overline{\gamma} = (\sum_{i=1}^n \gamma_i)/n$. Meanwhile, we define

$$f_\omega(s', a') = \left(\sum_{i=1}^n \frac{A_i^{\pi_{i,\omega}}(s', a')}{1 - \gamma_i} \cdot \frac{\mathrm{d}\sigma_{\pi_{i,\theta^*}}}{\mathrm{d}\varrho_{\pi_\omega}}(s', a') \right) \bigg/ \left(\sum_{i=1}^n g_{i,\omega}(s', a') \cdot \frac{\mathrm{d}\varsigma_{i,\pi_\omega}}{\mathrm{d}\varrho_{\pi_\omega}}(s', a') \right), \tag{C.46}$$

where $g_{i,\omega}$ is defined in (C.5) of Lemma C.1. In what follows, we define $v_0 \in \mathcal{B}_{\text{init}}$ as follows,

$$v_0 \in \underset{v \in \mathcal{B}_{\text{init}}}{\operatorname{argmin}} \|f_\omega(\cdot, \cdot) + \phi_\omega(\cdot, \cdot)^\top \omega - \phi_v(\cdot, \cdot)^\top v\|_{\varrho_{\pi_\omega}}. \tag{C.47}$$

It then holds from (C.45) that

$$L(\theta^*) - L(\omega) \le \epsilon + C \cdot \big\| f_\omega(\cdot, \cdot) - \phi_\omega(\cdot, \cdot)^\top (v_0 - \omega) \big\|_{\varrho_{\pi_\omega}} \tag{C.48}$$

$$\le \epsilon + C \cdot \big\| f_\omega(\cdot, \cdot) + \phi_\omega(\cdot, \cdot)^\top \omega - \phi_{v_0}(\cdot, \cdot)^\top v_0 \big\|_{\varrho_{\pi_\omega}} + \big\| \big(\phi_\omega(\cdot, \cdot) - \phi_{v_0}(\cdot, \cdot) \big)^\top v_0 \big\|_{\varrho_{\pi_\omega}}.$$

Note that $v_0, \omega \in \mathcal{B}_{\text{init}}$. Thus, following from Lemma C.2, upon taking expectation of (C.48) over the random initialization, we obtain that

$$\mathbb{E}_{\text{init}} \big[L(\theta^*) - L(\omega) \big]$$
$$\le \epsilon + C \cdot \mathbb{E}_{\text{init}} \big[\| f_\omega(\cdot, \cdot) + \phi_\omega(\cdot, \cdot)^\top \omega - \phi_{v_0}(\cdot, \cdot)^\top v_0 \|_{\varrho_{\pi_\omega}} \big] + \mathcal{O}(R_T^{3/2} \cdot m^{-1/4}). \tag{C.49}$$

Note that by the definition of the neural network and the feature mapping in (A.1) and (A.2), respectively, we have

$$f\big((\cdot, \cdot); v_0\big) = \phi_{v_0}(\cdot, \cdot)^\top v_0, \quad f\big((\cdot, \cdot); \omega\big) = \phi_\omega(\cdot, \cdot)^\top \omega.$$

Thus, by plugging the definition of v_0 in (C.47) into (C.49) and setting $c_\omega(\cdot, \cdot) = f((\cdot, \cdot); \omega) + f_\omega(\cdot, \cdot)$, we have

$$\mathbb{E}_{\text{init}} \big[L(\theta^*) - L(\omega) \big] \le \epsilon + C \cdot \mathbb{E}_{\text{init}} \Big[\inf_{v \in \mathcal{B}} \big\| c_\omega(\cdot, \cdot) - f\big((\cdot, \cdot); v\big) \big\|_{\varrho_{\pi_\omega}} \Big] + \mathcal{O}(R_T^{3/2} \cdot m^{-1/4}),$$

which completes the proof of Corollary A.2. □

C.5. Proof of Corollary A.4

Proof. The proof is similar to that of Corollary B.2 in §C.3. Similar to (C.20) in the proof of Theorem 4.4, we have

$$L(\omega) - L(\theta^*) \le \epsilon - \nabla_\omega L(\omega)^\top (\omega - v) + \frac{1}{n} \cdot \sum_{i=1}^n \langle \delta R_i / \delta h_{\omega_i}, h_{\omega_i} - h_{\theta_i^*} \rangle_{\mathcal{H}}, \tag{C.50}$$

which holds for all $v \in \mathcal{B}_{\text{init}}$. It then suffices to upper bound the right-hand side of (C.50). To this end, we first calculate the gradient $\nabla_\omega L(\omega)$. Similar to §C.3, we have

$$\nabla_\omega R_i(h_{\omega_i})^\top (\omega - v)$$
$$= \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \big(N(\omega, x, y) \phi_{\omega_i}(x') \big)^\top (\omega - v) \, \mathrm{d}\mathcal{D}_i(x, y) \, \mathrm{d}\rho(x'), \tag{C.51}$$

where ϕ_{ω_i} is the feature mapping defined in (A.2) with $W = \omega_i$ and

$$N(\omega, x, y) = I_{md} - \eta \cdot \nabla_\omega^2 \ell\big(h_\omega, (x, y)\big). \tag{C.52}$$

Note that for $\ell(h, (x, y)) = (y - h(x))^2$, we have

$$\nabla_\omega^2 \ell\big(h_\omega, (x, y)\big) = 2\big(\nabla_\omega h_\omega(x)\big)\big(\nabla_\omega h_\omega(x)\big)^\top + 2\big(h_\omega(x) - y\big)\nabla_\omega^2 h_\omega(x). \tag{C.53}$$

Meanwhile, by the parameterization $h_\omega(x) = f(x; \omega)$ defined in (A.1), we obtain that $\nabla_\omega h_\omega(x) = \phi_\omega(x)$ and $\nabla_\omega^2 h_\omega(x) = 0$, which holds for $x \in \mathcal{X}$ almost everywhere. Thus, it follows from (C.33) that

$$\nabla_\omega^2 \ell\big(h_\omega, (x, y)\big) = 2\phi_\omega(x)\phi_\omega(x)^\top \tag{C.54}$$

for $x \in \mathcal{X}$ almost everywhere. Moreover, for a fixed x, (C.54) holds for $\omega \in \mathbb{R}^{md}$ almost everywhere. Here ϕ_ω is the feature mapping defined in (A.2) with $W = \omega$. Hence, to obtain (C.54) for all $x \in \mathcal{X}$, we set the second order derivative of the neural network with respect to the parameter to be zero in the optimization of meta-objective when it is infinite (which occurs with zero probability and does not affect the convergence of meta-SL). By plugging (C.52) and (C.54) into (C.51), we have

$$\nabla_\omega R_i(h_{\omega_i})^\top v = \int_{\mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \phi_{\omega_i}(x')^\top K_{\omega,\eta}(\omega - v) \, d\rho(x'), \tag{C.55}$$

where we define

$$K_{\omega,\eta} = I_{md} - \eta \cdot \mathbb{E}_{x \sim \rho}[2\phi_\omega(x)\phi_\omega(x)^\top]. \tag{C.56}$$

In the sequel, we define

$$\phi_0(x) = \phi_{W_{\text{init}}}(x) \tag{C.57}$$

for notational simplicity, where W_{init} is the initial parameter of the neural network, and $\phi_{W_{\text{init}}}$ is the feature mapping defined in (A.2) with $W = W_{\text{init}}$. It then follows from (C.55) that

$$\nabla_\omega R_i(h_{\omega_i})^\top (\omega - v) = \int_{\mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \phi_0(x')^\top K_{\omega,\eta}(\omega - v) \, d\rho(x') \tag{C.58}$$

$$+ \int_{\mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \left(\phi_{\omega_i}(x') - \phi_0(x')\right)^\top K_{\omega,\eta}(\omega - v) \, d\rho(x').$$

Thus, by the definition of meta-objective in (4.4), it follows from (C.58) that

$$\nabla_\omega L(\omega)^\top (\omega - v) = \frac{1}{n} \cdot \sum_{i=1}^{n} \nabla_\omega R_i(h_{\omega_i})^\top (\omega - v)$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{\mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \phi_0(x')^\top K_{\omega,\eta}(\omega - v) \, d\rho(x') + \frac{1}{n} \cdot \sum_{i=1}^{n} P_i, \tag{C.59}$$

where we define ϕ_0 in (C.57) and we define

$$P_i = \int_{\mathcal{X}} (\delta R_i / \delta h_{\omega_i})(x') \cdot \left(\phi_{\omega_i}(x') - \phi_0(x')\right)^\top K_{\omega,\eta}(\omega - v) \, d\rho(x'). \tag{C.60}$$

Similar to (C.39) in the proof of Corollary B.2 in §C.3, we obtain from (C.50) and (C.59) that

$$L(\omega) - L(\theta^*) \leq \epsilon + \int_{\mathcal{X}} w(x') \cdot \left(u(x') - \phi_0(x')^\top K_{\omega,\eta}(\omega - v)\right) d\rho(x') + \frac{1}{n} \cdot \sum_{i=1}^{n} P_i \tag{C.61}$$

for all $v \in \mathcal{B}_{\text{init}}$. Here we define for all $x' \in \mathcal{X}$ that

$$
w(x') = \frac{1}{n} \cdot \sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x'),
$$

$$
u(x') = \left(\sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \cdot \left(h_{\omega_i}(x') - h_{\theta_i^*}(x') \right) \right) \Big/ \left(\sum_{i=1}^{n} (\delta R_i / \delta h_{\omega_i})(x') \right). \tag{C.62}
$$

In what follows, we fix $v \in \mathcal{B}_{\text{init}}$ as follows,

$$
v \in \operatorname*{argmin}_{\theta \in \mathcal{B}_{\text{init}}} \left\| u(\cdot) - f \left(\cdot ; K_{\omega,\eta}(\omega - \theta) + W_{\text{init}} \right) \right\|_{\rho}. \tag{C.63}
$$

Meanwhile, we define

$$
s = K_{\omega,\eta}(\omega - v) + W_{\text{init}}, \tag{C.64}
$$

where v is defined in (C.63). Note that $f(x; W_{\text{init}}) = \phi_0(x)^\top W_{\text{init}} = 0$ for all $x \in \mathcal{X}$ by the initialization of neural networks. It then follows from (C.61) that

$$
L(\omega) - L(\theta^*) \leq \epsilon + \int_{\mathcal{X}} w(x') \cdot \left(u(x') - \phi_0(x')^\top \left(K_{\omega,\eta}(\omega - v) + W_{\text{init}} \right) \right) d\rho(x') + \frac{1}{n} \cdot \sum_{i=1}^{n} P_i
$$

$$
\leq \epsilon + \int_{\mathcal{X}} w(x') \cdot \left(u(x') - \phi_s(x')^\top s \right) d\rho(x') + P_0 + \frac{1}{n} \cdot \sum_{i=1}^{n} P_i, \tag{C.65}
$$

where we define P_i and s in (C.60) and (C.64), respectively, and we define

$$
P_0 = \int_{\mathcal{X}} w(x') \cdot \left(\phi_0(x') - \phi_s(x') \right)^\top s \, d\rho(x'). \tag{C.66}
$$

By further plugging (C.63) into (C.65), we have

$$
L(\omega) - L(\theta^*) \leq \epsilon + \|w\|_{\rho} \cdot \inf_{\theta \in \mathcal{B}_0} \|u(\cdot) - f(\cdot; \theta)\|_{\rho} + P_0 + \frac{1}{n} \cdot \sum_{i=1}^{n} P_i, \tag{C.67}
$$

where we define

$$
\mathcal{B}_0 = K_{\omega,\eta}(\omega - \mathcal{B}_{\text{init}}) + W_{\text{init}}. \tag{C.68}
$$

Following from (C.43) in the proof of Corollary B.2 in §C.3, we obtain that

$$
\|w\|_{\rho} \leq 2\overline{R} = \frac{2}{n} \sum_{i=1}^{n} R_i^{1/2}(h_{\omega_i}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}_{(x,y) \sim \mathcal{D}_i} \left[\left(y - h_{\omega_i}(x) \right)^2 \right] \right\}^{1/2}. \tag{C.69}
$$

Upper Bounding P_0 and P_i for $i \in [n]$. It remains to upper bound the terms P_0 and P_i in (C.67) for $i \in [n]$. By the

Cauchy-Schwartz inequality and the definition of P_0 in (C.66), we have

$$\mathbb{E}_{\text{init}}[P_0] \leq \left\{ \mathbb{E}_{\text{init}}\big[\|w\|_\rho^2\big] \cdot \mathbb{E}_{\text{init}}\big[\|\phi_s(\cdot)^\top s - \phi_0(\cdot)^\top s\|_\rho^2\big] \right\}^{1/2}. \tag{C.70}$$

Meanwhile, for $K_{\omega,\eta}$ and s defined in (C.56) and (C.64), respectively, we have

$$\begin{aligned}
\|s - W_{\text{init}}\|_2 &\leq \|\omega - v\|_2 + 2\eta \cdot \left\| \mathbb{E}_{x\sim\rho}\big[\phi_\omega(x)\phi_\omega(x)^\top(\omega - v)\big] \right\|_2 \\
&\leq \|\omega - v\|_2 + 2\eta \cdot \mathbb{E}_{x\sim\rho}\big[\|\phi_\omega(x)\phi_\omega(x)^\top(\omega - v)\|_2\big] \\
&\leq (1 + 2\eta) \cdot \|\omega - v\|_2 \leq (2 + 4\eta) \cdot R_T,
\end{aligned} \tag{C.71}$$

where the first inequality follows from the triangle inequality, the second inequality follows from Jensen's inequality, the third inequality follows from the fact that $\|\phi_\omega(x)\|_2 \leq 1$ for all $x \in \mathcal{X}$, and the last inequality follows from the fact that $\omega, v \in \mathcal{B}_T$. Hence, by (C.71) and Lemma C.2, we obtain that

$$\mathbb{E}_{\text{init}}\big[\|\phi_s(\cdot)^\top s - \phi_0(\cdot)^\top s\|_\rho^2\big] = \mathcal{O}\big((1 + 2\eta)^3 \cdot R_T^3 \cdot m^{-1/2}\big). \tag{C.72}$$

By further plugging (C.72) and (C.69) into (C.70), we have

$$\mathbb{E}_{\text{init}}[P_0] = \mathcal{O}\big(\overline{C} \cdot (1 + 2\eta)^{3/2} \cdot R_T^{3/2} \cdot m^{-1/4}\big). \tag{C.73}$$

Here we define

$$\overline{C} = \big\{ \mathbb{E}_{\text{init}}[4\overline{R}^2] \big\}^{1/2}, \tag{C.74}$$

where $\overline{R}$ is defined in (C.69).

Similarly, for P_i defined in (C.60), we have

$$\mathbb{E}_{\text{init}}[P_i] \leq \left\{ \mathbb{E}_{\text{init}}\big[\|\delta R_i/\delta h_{\omega_i}\|_\rho^2\big] \cdot \underbrace{\mathbb{E}_{\text{init}}\big[\|\phi_{\omega_i}(\cdot)^\top K_{\omega,\eta}(\omega - v) - \phi_0(\cdot)^\top K_{\omega,\eta}(\omega - v)\|_\rho^2\big]}_{U} \right\}^{1/2}. \tag{C.75}$$

We first upper bound the term U in (C.75). By the definition of ω_i in (4.12), we have

$$\|\omega_i - W_{\text{init}}\|_2 \leq \|\omega - W_{\text{init}}\|_2 + \eta \cdot \|\nabla_\omega R_i(h_\omega)\|_2. \tag{C.76}$$

Meanwhile, by the definition of the risk in (4.1) and the definition of the squared loss in (B.1), we have

$$\|\nabla_\omega R_i(h_\omega)\|_2 = \left\| \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[2\big(h_\omega(x) - y\big)\phi_\omega(x)\big] \right\|_2 \leq \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[2|h_\omega(x) - y| \cdot \|\phi_\omega(x)\|_2\big], \tag{C.77}$$

where the first equality follows from the neural network parameterization of h_ω and the definition of feature mapping in (A.2), and the second inequality follows from Jensen's inequality. Following from the fact that $\|\phi_\omega(x)\|_2 \leq 1$ for all $x \in \mathcal{X}$ and the assumption that $|y| \leq Y_{\max}$ for all $y \in \mathcal{Y}$, we further have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[2|h_\omega(x) - y| \cdot \|\phi_\omega(x)\|_2\big] \leq 2Y_{\max} + 2\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[|h_\omega(x)|\big]. \tag{C.78}$$

Note that $f(x; W_{\text{init}}) = \phi_0(x)^\top W_{\text{init}} = 0$ by the initialization. Hence, we have for all $x \in \mathcal{X}$ that

$$
\begin{aligned}
|h_\omega(x)| &= |f(x; \omega) - f(x; W_{\text{init}})| \\
&\leq \sup_{\theta \in \mathbb{R}^{md}} \|\nabla_\theta f(x; \theta)\|_2 \cdot \|\omega - W_{\text{init}}\|_2 \\
&\leq \|\omega - W_{\text{init}}\|_2 \leq R_T,
\end{aligned}
\tag{C.79}
$$

where the first equality follows from the neural network parameterization of the hypothesis h_ω, the second inequality follows from the fact that $\|\nabla_\theta f(x; \theta)\|_2 = \|\phi_\theta(x)\|_2 \leq 1$ for all $x \in \mathcal{X}$, and the last inequality follows from the fact that $\omega \in \mathcal{B}_T$. By plugging (C.79) into (C.78), we have

$$
\mathbb{E}_{(x,y) \sim \mathcal{D}_i} \big[2|h_\omega(x) - y| \cdot \|\phi_\omega(x)\|_2 \big] \leq 2Y_{\max} + 2R_T.
\tag{C.80}
$$

By further plugging (C.77) into (C.80) into (C.76), we obtain for all $i \in [n]$ that

$$
\|\omega_i - W_{\text{init}}\|_2 \leq \big((1 + 2\eta) \cdot R_T + 2\eta \cdot Y_{\max} \big).
\tag{C.81}
$$

Meanwhile, similar to (C.71), we obtain for $\omega, v \in \mathcal{B}_T$ that

$$
\|K_{\omega, \eta}(\omega - v)\|_2 \leq (2 + 4\eta) \cdot \|\omega - v\|_2 \leq (4 + 8\eta) R_T.
\tag{C.82}
$$

Finally, by Assumption A.3 with Lemma C.2, we obtain for U in (C.75) that

$$
\begin{aligned}
U &\leq 2\mathbb{E}_{\text{init}} \Big[\big\| \big(\phi_{\omega_i}(\cdot) - \phi_0(\cdot) \big)^\top \big(K_{\omega, \eta}(\omega - v) + W_{\text{init}} \big) \big\|_\rho^2 \Big] + 2\mathbb{E}_{\text{init}} \Big[\| \phi_{\omega_i}(\cdot)^\top W_{\text{init}} - \phi_0(\cdot)^\top W_{\text{init}} \|_\rho^2 \Big] \\
&= \mathcal{O}\Big(\big((1 + \eta) \cdot R_T + \eta \cdot Y_{\max} \big)^3 \cdot m^{-1/2} \Big).
\end{aligned}
\tag{C.83}
$$

Meanwhile, by (C.42) in the proof of Corollary B.2 in §C.3, we obtain that

$$
\mathbb{E}_{\text{init}} \big[\| \delta R_i / \delta h_{\omega_i} \|_\rho^2 \big] \leq 4\mathbb{E}_{\text{init}} \big[R_i(h_{\omega_i}) \big].
\tag{C.84}
$$

Thus, by plugging (C.83) and (C.84) into (C.75), we have

$$
\mathbb{E}_{\text{init}}[P_i] = \mathcal{O}(C_i \cdot R_1^{3/2} \cdot m^{-1/4}),
\tag{C.85}
$$

where we define C_i and R_1 as follows,

$$
C_i = 2\Big\{ \mathbb{E}_{\text{init}} \big[R_i(h_{\omega_i}) \big] \Big\}^{1/2}, \qquad R_1 = (1 + \eta) \cdot R_T + \eta \cdot Y_{\max}.
\tag{C.86}
$$

Finally, by plugging (C.69), (C.73), and (C.85) into (C.67), we conclude that

$$
\begin{aligned}
\mathbb{E}_{\text{init}} \big[L(\omega) - L(\theta^*) \big] &\leq \epsilon + \mathbb{E}_{\text{init}} \Big[2\overline{R} \cdot \inf_{\theta \in \mathcal{B}_0} \|u(\cdot) - f(\cdot; \theta)\|_\rho \Big] \\
&\quad + \mathcal{O}\big(\overline{C} \cdot (1 + 2\eta)^{3/2} \cdot R_T^{3/2} \cdot m^{-1/4} + D_0 \cdot R_1^{3/2} \cdot m^{-1/4} \big).
\end{aligned}
\tag{C.87}
$$

Here $\mathcal{B}_0$ and $\overline{C}$ are the constants defined in (C.68) and (C.74), respectively, $R_1 = (1 + \eta) \cdot R_T + \eta \cdot Y_{\max}$, and $D_0 = \frac{1}{n} \cdot \sum_{i=1}^n C_i$, where C_i is the constant defined in (C.86). Thus, by setting $G_T = (1 + \eta) \cdot R_T + \eta \cdot Y_{\max}$, we obtain from

(C.87) that

$$\mathbb{E}_{\text{init}}\big[L(\omega) - L(\theta^*)\big] \leq \epsilon + \mathbb{E}_{\text{init}}\Big[2\overline{R} \cdot \inf_{v \in \mathcal{B}_0} \|\overline{u}(\cdot) - f(\cdot; v)\|_\rho\Big] + \mathcal{O}(G_T^{3/2} \cdot m^{-1/4}),$$

which completes the proof of Corollary A.4. □

D. Proof of Auxiliary Result

In this section, we present the proofs fo the auxiliary results.

D.1. Proof of Proposition 3.2

Proof. By Lemma E.2, which corresponds to the policy gradient theorem (Sutton & Barto, 2018), we have

$$\nabla_\theta J_i(\pi_{i,\theta}) = \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta}}}\big[\nabla_\theta \log \pi_{i,\theta}(a \,|\, s) \cdot A_i^{\pi_{i,\theta}}(s,a)\big], \tag{D.1}$$

where J_i and $A_i^{\pi_{i,\theta}}$ are the expected total reward and the advantage function of the policy $\pi_{i,\theta}$ corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$, respectively, and $\sigma_{\pi_{i,\theta}}$ is the state-action visitation measure induced by the policy $\pi_{i,\theta}$. By plugging the form of $\pi_{i,\theta}$ in (3.3) into (D.1), we obtain that

$$\nabla_\theta J_i(\pi_{i,\theta}) = \mathbb{E}_{(s,a) \sim \sigma_{\pi_{i,\theta}}}\Big[\big(1/\tau \cdot \phi(s,a) + \eta \cdot \nabla_\theta Q_i^{\pi_\theta}(s,a)\big) \cdot A_i^{\pi_{i,\theta}}(s,a)\Big]. \tag{D.2}$$

Here $Q_i^{\pi_\theta}$ is the state-action value function of the main effect π_θ corresponding to the MDP $(\mathcal{S}, \mathcal{A}, P_i, r_i, \gamma_i, \zeta_i)$. Applying Lemma E.2 again, we obtain that

$$\begin{aligned}
\nabla_\theta Q_i^{\pi_\theta}(s,a) &= \nabla_\theta\Big((1 - \gamma_i) \cdot r_i(s,a) + \gamma_i \cdot \mathbb{E}_{s' \sim P_i(\cdot \,|\, s,a)}\big[V_i^{\pi_\theta}(s')\big]\Big) \\
&= \gamma_i \cdot \mathbb{E}_{(s',a') \sim \sigma_{i,\pi_\theta}^{(s,a)}}\big[\nabla_\theta \log \pi_\theta(a' \,|\, s') \cdot A_i^{\pi_\theta}(s',a')\big] \\
&= \gamma_i \cdot \mathbb{E}_{(s',a') \sim \sigma_{i,\pi_\theta}^{(s,a)}}\Big[\big(1/\tau \cdot \phi(s',a')\big) \cdot A_i^{\pi_\theta}(s',a')\Big].
\end{aligned} \tag{D.3}$$

Here the last equality follows from the parameterization of π_θ in (3.2) and $\sigma_{i,\pi_\theta}^{(s,a)}$ is the state-action visitation measure of the main effect π_θ, which is defined in Definition 3.1. By plugging (D.3) into (D.2), we complete the proof of Proposition 3.2. □

D.2. Proof of Lemma C.1

Proof. Following from Proposition 3.2, we have

$$\begin{aligned}
\nabla_\theta L(\theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}_{(s',a',s,a) \sim \rho_{i,\pi_\theta}} \big[&\gamma_i \cdot \eta/\tau \cdot \phi(s',a') \cdot A_i^{\pi_\theta}(s',a') \cdot A_i^{\pi_{i,\theta}}(s,a) \\
&+ 1/\tau \cdot \phi(s,a) \cdot A_i^{\pi_{i,\theta}}(s,a)\big],
\end{aligned} \tag{D.4}$$

where ρ_{i,π_θ} is the joint meta-visitation measure defined in (3.9) of Definition 3.3. Meanwhile, it holds that

$$\begin{aligned}
\mathbb{E}_{(s',a',s,a) \sim \rho_{i,\pi_\theta}}\big[&\phi(s',a') \cdot A_i^{\pi_\theta}(s',a') \cdot A_i^{\pi_{i,\theta}}(s,a)\big] \\
&= \mathbb{E}_{(s',a') \sim \varsigma_{i,\pi_\theta}}\big[\phi(s',a') \cdot G_{i,\pi_\theta}(s',a') \cdot A_i^{\pi_\theta}(s',a')\big],
\end{aligned} \tag{D.5}$$

where ς_{i,π_θ} is the meta-visitation measure defined in (3.10) of Definition 3.3, and G_{i,π_θ} is defined as follows,

$$G_{i,\pi_\theta}(s',a') = \mathbb{E}_{(s',a',s,a)\sim\rho_{i,\pi_\theta}}\big[A_i^{\pi_{i,\theta}}(s,a)\,\big|\,s',a'\big]. \tag{D.6}$$

Here ρ_{i,π_θ} is the joint meta-visitation measure defined in (3.9) of Definition 3.3. By plugging (D.5) into (D.4), we obtain that

$$
\begin{aligned}
\nabla_\theta L(\theta) &= \frac{1}{n}\cdot\sum_{i=1}^n \mathbb{E}_{(s,a)\sim\sigma_{\pi_{i,\theta}}}\big[1/\tau\cdot\phi(s,a)\cdot A_i^{\pi_{i,\theta}}(s,a)\big] \\
&\qquad\qquad + \mathbb{E}_{(s',a')\sim\varsigma_{i,\pi_\theta}}\big[\gamma_i\cdot\eta/\tau\cdot\phi(s',a')\cdot G_{i,\pi_\theta}(s',a')\cdot A_i^{\pi_\theta}(s',a')\big] \\
&= \frac{1}{n}\cdot\sum_{i=1}^n \mathbb{E}_{(s',a')\sim\varsigma_{i,\pi_\theta}}\Big[\phi(s',a')\cdot\big(1/\tau\cdot A_i^{\pi_{i,\theta}}(s',a')\cdot(\mathrm{d}\sigma_{\pi_{i,\theta}}/\mathrm{d}\varsigma_{i,\pi_\theta})(s',a') \\
&\qquad\qquad\qquad\qquad + \gamma_i\cdot\eta/\tau\cdot G_{i,\pi_\theta}(s',a')\cdot A_i^{\pi_\theta}(s',a')\big)\Big],
\end{aligned}
$$

where $A_{i,\theta}$ is defined in (D.6) and $\mathrm{d}\sigma_{\pi_{i,\theta}}/\mathrm{d}\varsigma_{i,\pi_\theta}$ is the Radon-Nikodym derivative. Thus, we complete the proof of Lemma C.1. $\qquad\square$

D.3. Proof of Proposition B.1

Proof. It suffices to prove for all $h \in \mathcal{H}$ and $\delta R_i/\delta h$ defined in (B.2) that the operator

$$A_h(\cdot) = \langle\cdot, \delta R_i/\delta h\rangle_\mathcal{H} \tag{D.7}$$

is the Fréchet derivative of R_i at $h \in \mathcal{H}$ defined in Definition 4.1. For all $h_1 \in \mathcal{H}$, we have

$$
\begin{aligned}
&R_i(h_1) - R_i(h) - A_h(h_1 - h) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \big(h_1(x) - y\big)^2 - \big(h(x) - y\big)^2 \,\mathrm{d}\mathcal{D}_i(x,y) - \int_\mathcal{X} \big(h_1(x) - h(x)\big)\cdot(\delta R_i/\delta h)(x)\,\mathrm{d}\rho(x), \tag{D.8}
\end{aligned}
$$

where the equality follows from the definition of A_h in (D.7). Meanwhile, for $\delta R_i/\delta h$ defined in (B.2) and $\overline{\mathcal{D}}_i = \rho$, we have

$$\int_\mathcal{X} \big(h_1(x) - h(x)\big)\cdot(\delta R_i/\delta h)(x)\,\mathrm{d}\rho(x) = \int_\mathcal{X} 2\big(h_1(x) - h(x)\big)\cdot\big(h(x) - y\big)\,\mathrm{d}\mathcal{D}_i(x,y). \tag{D.9}$$

By plugging (D.9) into (D.8), we obtain that

$$
\begin{aligned}
&R_i(h_1) - R_i(h) - A_h(h_1 - h) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \big(h_1(x) - y\big)^2 - \big(h(x) - y\big)^2 - 2\big(h_1(x) - h(x)\big)\cdot\big(h(x) - y\big)\,\mathrm{d}\mathcal{D}_i(x,y) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \big(h_1(x) - h(x)\big)\cdot\big(h_1(x) + h(x) - 2y\big) - 2\big(h_1(x) - h(x)\big)\cdot\big(h(x) - y\big)\,\mathrm{d}\mathcal{D}_i(x,y) \\
&= \int_\mathcal{X} \big(h_1(x) - h(x)\big)^2\,\mathrm{d}\rho(x) = \|h_1 - h\|_\mathcal{H}^2. \tag{D.10}
\end{aligned}
$$

Hence, by (D.10), we have

$$\lim_{\|h_1-h\|_{\mathcal{H}}\to 0} \frac{|R_i(h_1) - R_i(h) - A_h(h_1 - h)|}{\|h_1 - h\|_{\mathcal{H}}} = \lim_{\|h_1-h\|_{\mathcal{H}}\to 0} \|h_1 - h\|_{\mathcal{H}} = 0.$$

Thus, following from the definition of the Fréchet derivative in Definition 4.1, we conclude that A_h defined in (D.7) is the Fréchet derivative of R_i at $h \in \mathcal{H}$, which completes the proof of Proposition B.1. □

D.4. Proof of Lemma C.2

Proof. In what follows, we write $\phi_0 = \phi_{W_{\text{init}}}$ for notational simplicity, where $\phi_{W_{\text{init}}}$ is the feature mapping defined in (A.2) with $W = W_{\text{init}}$. Note that

$$\|\phi_{\omega_0}(\cdot,\cdot)^\top \omega_2 - \phi_{\omega_1}(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}$$
$$\leq 2\|\phi_{\omega_0}(\cdot,\cdot)^\top \omega_2 - \phi_0(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}} + 2\|\phi_{\omega_1}(\cdot,\cdot)^\top \omega_2 - \phi_0(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}, \tag{D.11}$$

where the inequality follows from the fact that $\|f(\cdot) + g(\cdot)\|^2_{\varrho_{\pi_\theta}} \leq 2\|f(\cdot)\|^2_{\varrho_{\pi_\theta}} + 2\|g(\cdot)\|^2_{\varrho_{\pi_\theta}}$. We now upper bound the right-hand side of (D.11) under the expectation with respect to the random initialization. In the sequel, we write $\phi_\omega(\cdot) = ([\phi_\omega(\cdot)]_{\text{U}}^\top, [\phi_\omega(\cdot)]_{\text{L}}^\top)^\top$ and $\phi_0(\cdot) = ([\phi_0(\cdot)]_{\text{U}}^\top, [\phi_0(\cdot)]_{\text{L}}^\top)^\top$, respectively, where $[\phi_0(\cdot)]_{\text{U}} = ([\phi_0(\cdot)]_1^\top, \ldots, [\phi_0(\cdot)]_{m/2}^\top)^\top$ and $[\phi_0(\cdot)]_{\text{L}} = ([\phi_0(\cdot)]_{m/2+1}^\top, \ldots, [\phi_0(\cdot)]_m^\top)^\top$. Similarly, we write $\omega = ([\omega]_{\text{U}}^\top, [\omega]_{\text{L}}^\top)^\top$ and $W_{\text{init}} = ([W_{\text{init}}]_{\text{U}}^\top, [W_{\text{init}}]_{\text{L}}^\top)^\top$, respectively. Note that for $\|\omega - W_{\text{init}}\|_2 \leq R$, we have

$$\|\omega - W_{\text{init}}\|_2^2 = \|[\omega]_{\text{U}} - [W_{\text{init}}]_{\text{U}}\|_2^2 + \|[\omega]_{\text{L}} - [W_{\text{init}}]_{\text{L}}\|_2^2 \leq R^2. \tag{D.12}$$

Hence, we obtain from (D.12) that $\|[\omega]_{\text{U}} - [W_{\text{init}}]_{\text{U}}\|_2 \leq R$ and $\|[\omega]_{\text{L}} - [W_{\text{init}}]_{\text{L}}\|_2 \leq R$. Meanwhile, note that $[W_{\text{init}}]_{\text{U}} = ([W_{\text{init}}]_1^\top, \ldots, [W_{\text{init}}]_{m/2}^\top)^\top$, where $[W_{\text{init}}]_r \sim N(0, I_d/d)$ are sampled independently for all $r \in [m/2]$. Thus, by Lemma E.1, under Assumption A.1, we obtain for $\omega_0, \omega_2 \in \mathcal{B} = \{\theta \in \mathbb{R}^{md} : \|\theta - W_{\text{init}}\|_2 \leq R\}$ that

$$\mathbb{E}_{\text{init}}\left[\left\|\left[\phi_{\omega_0}(\cdot)\right]_{\text{U}}^\top [\omega_2]_{\text{U}} - \left[\phi_0(\cdot)\right]_{\text{U}}^\top [\omega_2]_{\text{U}}\right\|^2_{\varrho_{\pi_\theta}}\right] = \mathcal{O}(R^3 \cdot m^{-1/2}). \tag{D.13}$$

Similarly, we have

$$\mathbb{E}_{\text{init}}\left[\left\|\left[\phi_{\omega_0}(\cdot)\right]_{\text{L}}^\top [\omega_2]_{\text{L}} - \left[\phi_0(\cdot)\right]_{\text{L}}^\top [\omega_2]_{\text{L}}\right\|^2_{\varrho_{\pi_\theta}}\right] = \mathcal{O}(R^3 \cdot m^{-1/2}). \tag{D.14}$$

Following from (D.13) and (D.14), we have

$$\mathbb{E}_{\text{init}}\left[\|\phi_{\omega_0}(\cdot,\cdot)^\top \omega_2 - \phi_0(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}\right]$$
$$= \mathbb{E}_{\text{init}}\left[\left\|\left[\phi_{\omega_0}(\cdot)\right]_{\text{U}}^\top [\omega_2]_{\text{U}} - \left[\phi_{\omega_0}(\cdot)\right]_{\text{U}}^\top [\omega_2]_{\text{U}} + \left[\phi_{\omega_0}(\cdot)\right]_{\text{L}}^\top [\omega_2]_{\text{L}} - \left[\phi_{\omega_0}(\cdot)\right]_{\text{L}}^\top [\omega_2]_{\text{L}}\right\|^2_{\varrho_{\pi_\theta}}\right]$$
$$\leq 2\mathbb{E}_{\text{init}}\left[\left\|\left[\phi_{\omega_0}(\cdot)\right]_{\text{U}}^\top [\omega_2]_{\text{U}} - \left[\phi_0(\cdot)\right]_{\text{U}}^\top [\omega_2]_{\text{U}}\right\|^2_{\varrho_{\pi_\theta}}\right] + 2\mathbb{E}_{\text{init}}\left[\left\|\left[\phi_{\omega_0}(\cdot)\right]_{\text{L}}^\top [\omega_2]_{\text{L}} - \left[\phi_0(\cdot)\right]_{\text{L}}^\top [\omega_2]_{\text{L}}\right\|^2_{\varrho_{\pi_\theta}}\right]$$
$$= \mathcal{O}(R^3 \cdot m^{-1/2}), \tag{D.15}$$

where the second inequality follows from the fact that $\|f(\cdot) + g(\cdot)\|^2_{\varrho_{\pi_\theta}} \leq 2\|f(\cdot)\|^2_{\varrho_{\pi_\theta}} + 2\|g(\cdot)\|^2_{\varrho_{\pi_\theta}}$. Similarly, we have

$$\mathbb{E}_{\text{init}}\left[\|\phi_{\omega_1}(\cdot,\cdot)^\top \omega_2 - \phi_0(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}\right] = \mathcal{O}(R^3 \cdot m^{-1/2}). \tag{D.16}$$

Finally, by plugging (D.15) and (D.16) into (D.11), we have

$$
\mathbb{E}_{\mathrm{init}}\big[\|\phi_{\omega_0}(\cdot,\cdot)^\top \omega_2 - \phi_{\omega_1}(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}\big]
$$
$$
\leq 2\mathbb{E}_{\mathrm{init}}\big[\|\phi_{\omega_0}(\cdot,\cdot)^\top \omega_2 - \phi_0(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}\big] + 2\mathbb{E}_{\mathrm{init}}\big[\|\phi_{\omega_1}(\cdot,\cdot)^\top \omega_2 - \phi_0(\cdot,\cdot)^\top \omega_2\|^2_{\varrho_{\pi_\theta}}\big]
$$
$$
= \mathcal{O}(R^3 \cdot m^{-1/2}),
$$

which concludes the proof of Lemma C.2. $\qquad\square$

E. Auxiliary Lemma

In this section, we present the auxiliary lemmas.

Lemma E.1 (Linearization Error (Cai et al., 2019))**.** Let $\|x\| \leq 1$ for all $x \in \mathcal{X}$ and $[W_{\mathrm{init}}]_r \sim N(0, I_d/d)$ be sampled independently for all $r \in [m]$. For parameters $\omega, \omega' \in \mathcal{B}_{\mathrm{init}} = \{\theta \in \mathbb{R}^{md} : \|\theta - W_{\mathrm{init}}\|_2 \leq R\}$ and the distribution ρ over $\mathcal{X}$ such that Assumption A.3 holds, we have

$$
\mathbb{E}_{\mathrm{init}}\big[\|\phi_\omega(\cdot)^\top \omega' - \phi_{W_{\mathrm{init}}}(\cdot)^\top \omega'\|^2_\rho\big] = \mathcal{O}(R^3 \cdot m^{-1/2}).
$$

Proof. See (Cai et al., 2019) for a detailed proof. $\qquad\square$

Lemma E.2 (Policy Gradient (Sutton & Barto, 2018))**.** Let π_θ be the parameterized policy with the parameter θ. It holds that

$$
\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s \sim \nu_{\pi_\theta}}\big[\langle \pi_\theta(\cdot \,|\, s), Q^{\pi_\theta}(s, \cdot)\rangle\big]
$$
$$
= \mathbb{E}_{(s,a) \sim \sigma_{\pi_\theta}}\big[\nabla_\theta \log \pi_\theta(a \,|\, s) \cdot Q^{\pi_\theta}(s, a)\big],
$$

where ν_{π_θ} is the state visitation measure defined in (2.4) with $\pi = \pi_\theta$, and $\sigma_{\pi_\theta}(\cdot, \cdot) = \pi_\theta(\cdot \,|\, \cdot) \cdot \nu_{\pi_\theta}(\cdot)$ is the corresponding state-action visitation measure induce by π_θ.

Proof. See (Sutton & Barto, 2018) for a detailed proof. $\qquad\square$

Lemma E.3 (Performance Difference (Kakade & Langford, 2002))**.** It holds for all policies π and $\widetilde{\pi}$ that

$$
J(\widetilde{\pi}) - J(\pi) = (1 - \gamma)^{-1} \cdot \mathbb{E}_{(s,a) \sim \sigma_{\widetilde{\pi}}}\big[A^\pi(s, a)\big],
$$

where $\sigma_{\widetilde{\pi}}$ is the state-action visitation measure induced by $\widetilde{\pi}$.

Proof. See (Kakade & Langford, 2002) for a detailed proof. $\qquad\square$