Data-Driven Knowledge Transfer in Batch Q^* Learning

Elynn Chen[♭] Xi Chen[♮] * Wenbo Jing[♯]
[♭,♮,♯]Stern School of Business, New York University.

April 2, 2024

Abstract

In data-driven decision-making in marketing, healthcare, and education, it is desirable to utilize a large amount of data from existing ventures to navigate high-dimensional feature spaces and address data scarcity in new ventures. We explore knowledge transfer in dynamic decision-making by concentrating on batch stationary environments and formally defining task discrepancies through the lens of Markov decision processes (MDPs). We propose a framework of Transferred Fitted Q-Iteration algorithm with general function approximation, enabling the direct estimation of the optimal action-state function Q^* using both target and source data. We establish the relationship between statistical performance and MDP task discrepancy under sieve approximation, shedding light on the impact of source and target sample sizes and task discrepancy on the effectiveness of knowledge transfer. We show that the final learning error of the Q^* function is significantly improved from the single task rate both theoretically and empirically.

Keywords: Transfer Learning; Stationary Markov Decision Process; Offline Reinforcement Learning; Sieve Approximation; Fitted Q Iteration

*Corresponding author. Email: *xchen3@stern.nyu.edu*

1 Introduction

Data-driven sequential decision-making is gaining widespread prominence in real-world applications, including marketing (Liu, 2023), healthcare (Komorowski et al., 2018), and education (Rafferty et al., 2016). A primary challenge in these areas is managing high-dimensional feature spaces, especially when personalizing services or navigating complex domains. Furthermore, societal applications often face a significant data scarcity issue when venturing into new locations, targeting different population groups, or introducing new products or services. Data scarcity, marked by high dimensionality or a lack of historical data, demands innovative methods for data aggregation and automatic knowledge transfer. To tackle this crucial challenge, we introduce a knowledge transfer framework designed for data-driven sequential decision-making. This method can accelerate learning in a specific decision-making task by utilizing related source tasks from large-scale observational, simulated, or offline datasets.

The formal study of data-driven sequential decision-making is conducted within the broad framework of *reinforcement learning* (RL) (Sutton and Barto, 2018). Within RL, numerous model assumptions and methods exist for estimation and decision-making. In this paper, we focus on sample-transferred estimation of the optimal action-value function, i.e., the Q^* function, for stationary *Markov decision processes* (MDPs), employing the classic framework of *Fitted Q-Iteration* (FQI) (Ernst et al., 2005).

The literature lacks a thorough examination of *transfer learning* (TL) for RL with Q^* estimation that is supported by theoretical guarantees. We pioneer this investigation by first delineating the transferred RL problem within the framework of MDPs, where we formally define the *RL task discrepancy* based on differences in reward functions and transition probabilities. To facilitate a transferred algorithm that aims at direct estimation of

the Q^* function, we derive an important theoretical result. This result explicates the relationship between task discrepancy and the divergence of Q^* functions of different MDPs. Given that reward functions and transition probabilities can be readily estimated in practical applications, this also guarantees the tangible efficacy of transferring MDP tasks that display minor discrepancies. This precise definition of the TL for RL, along with our theoretical finding, lays the groundwork for both the development of an estimation procedure and subsequent theoretical analysis.

Based on the formal characterization we've established, we introduce a general framework of Transferred FQI algorithm (Algorithm 1). It is, in essence, an *iterative fixed-point algorithm* with knowledge transfer built on general function approximations. While it adopts the core transfer-learning concept of initial learning of commonalities followed by adjustments for idiosyncratic biases, the Transferred FQI algorithm notably diverges from current transfer algorithms applied in supervised or unsupervised learning in two principal respects. First, the Transferred FQI adopts an iterative approach to knowledge transfer, in contrast to the one-off nature of transferred learning in supervised or unsupervised settings. This necessitates meticulous attention to mitigate estimation biases and transfer-induced errors across each iteration – a challenge not previously tackled by non-iterative transferred learning algorithms in existing literature. Second, due to its self-iterative nature, the pseudo-response in FQI is not observable and needs to be re-constructed for each task in each iteration using the estimators obtained in the previous step. Therefore, to enable the benefit of transfer learning, we take extra steps to simultaneously build improved estimators for both the target and source tasks. This contrasts with the conventional focus of transferred supervised learning algorithms, which predominantly aim to refine estimations solely for the target task.

For theoretical analysis, we instantiate the general framework using semi-parametric

sieve approximation, which is widely employed in societal applications, and establish theoretical guarantees. The theoretical analysis framework developed here for an iterative fixed-point algorithm with knowledge transfer generally applies to other types of function approximation and similarity characterization. The result we establish for the Transferred FQI algorithm provides several important insights into the practice of TL for offline RL. Firstly, for a given task difference, it spells out the necessary numbers of target and source samples to correct the task-difference bias and to fully enjoy the benefit of transfer learning, which implies an artful tug-of-war between the source and the target tasks when task discrepancy exists: on the one hand, one wishes to have a large size of source data to improve the estimation accuracy of the common component shared by the source and target tasks; on the other hand, one needs to control the ratio between the size of source data over that of the target data to make sure that the bias induced from the task difference can be corrected with limited target samples. We show that, with the help of transfer learning, the target RL task with sample size n_0 can enjoy the convergence rate of $\sqrt{p/n_{\mathcal{K}}}$, ignoring the logarithm term, where $n_{\mathcal{K}}$ is the total sample size of all tasks.

1.1 Literature and Organization

This paper is situated at the intersection of two bodies of literature: batch reinforcement learning and transfer learning. The literature on reinforcement learning is broad and vast. The readers are referred to Sutton and Barto (2018) for comprehensive reviews of RL. We review only the most relevant studies with theoretical guarantees.

Batch Reinforcement Learning. We work under the setting of batch reinforcement learning (Shi et al., 2022; Yan et al., 2022; Li et al., 2024), where a sufficient amount of *source* data – usually a set of transitions sampled from the *source* MDP – is available.

FQI is an iterative framework that is the prototype of many batch RL algorithms. For example, the empirically successful Deep Q-Network (DQN) can be reduced to the neural FQI algorithm (Riedmiller, 2005) under the technique of experience replay and function approximation with ReLU deep neural network. Murphy (2005) and Munos and Szepesvári (2008) established the finite sample bounds for FQI for a general class of regression functions. Various variations of FQI have been studied in the literature. For example, Chen and Jiang (2019) and Xie and Jiang (2020, 2021) studied the necessity of assumptions for polynomial sample complexity and developed an algorithm under relaxed assumptions; and Fan et al. (2020) studied the Deep Q-Network (DQN) algorithm from both algorithmic and statistical perspectives.

All the current literature in FQI considers only a single RL task. Quite differently, the present paper considers the assistance of multiple RL source tasks to a target task with rigorous formulation, estimation algorithm, and theoretical guarantees. The iterative nature of FQI, together with semi-parametric function approximation and penalization, brings new challenges to the problem. In the trivial case where the source tasks are empty, the statistical rate of our proposed algorithm (Algorithm 1) recovers the statistical convergence rate of non-parametric estimation on a single RL task. When the source sample size is larger than the target sample size, i.e., $n_{\mathcal{K}} \gtrsim n_0$, the statistical rate becomes sharper as long as the task discrepancy satisfies a mild condition. Therefore, for a fixed target task, our proposed algorithm achieves improvement by transferring as long as the source tasks are sufficiently similar to the target task and the source sample size is sufficiently large compared to the target sample size.

Transfer Learning. On the methodological and theoretical front, the broad framework of transfer learning has been investigated under conditional shifts and marginal shifts between the source and target domains. A prominent example of conditional drift is *posterior drift*

(Li et al., 2022), where the conditional distribution of the response Y given predictors X of the target task differs from that of the source task. Prominent examples of marginal drift include *covariate shift* (Wang, 2023), where the marginal distribution of predictors X changes, and *label shift* (Maity et al., 2022), where the marginal distribution of the response Y changes.

The present paper is most related to the literature of transfer learning under *posterior drifts*, which has been studied in different contexts across a spectrum of supervised learning (SL) problems, including classification (Cai and Wei, 2021), high-dimensional linear regression (Li et al., 2022), and generalized linear models (Li et al., 2023; Tian and Feng, 2023), and there are also applications to problems in unsupervised learning (UL) scenarios (Li et al., 2023). Moreover, this work is also related to other multi-dataset problems such as multi-task learning (see, e.g., Duan and Wang (2023) and references therein) and multi-distribution learning (see, e.g., Zhang et al. (2023)).

This paper uniquely explores TL within offline RL contexts, particularly focusing on *posterior drifts*. Our objective is to estimate the optimal Q^* function through sample transfers directly. Unlike supervised learning, offline RL estimation of Q^* involves no direct observation but instead seeks to approximate the fixed point of the population-level Bellman optimality equation through iterative updates of a sample-level version. This process requires a sophisticated de-biasing method in each iteration to counter sequential bias from task variances. Additionally, our theoretical exploration into TL using semi-parametric sieve approximation presents novel insights into RL-based transfer learning, uncovering phenomena not previously identified in supervised or unsupervised learning contexts. A thorough discussion is provided in Section 3.3.

Transfer Learning for RL. The RL framework, inclusive of various elements within an MDP, leads to empirical TL studies in deep RL adopting different assumptions about

task similarities across MDP components, resulting in diverse research focuses. For instance, learning from demonstration assumes identical source and target MDPs (Ma et al., 2019). In contrast, policy transfer research often considers variations in state and action spaces (Yin and Pan, 2017) or reward functions (Barreto et al., 2017), while reward shaping studies presuppose differences in reward functions defined by a specific function (Vecerik et al., 2017). Federated multi-task RL assumes different tasks have different reward functions while sharing the same transition kernel (Yang et al., 2023). Representation transfer research posits that state, action, or reward spaces can be divided into orthogonal, task-invariant subspaces, facilitating knowledge transfer across domains. For recent reviews on TL in deep RL, see works like Zhu et al. (2023) and references therein.

Theoretical RL research has begun to rigorously address TL with formal proofs, concentrating on non-stationary finite-horizon MDPs. Under linear MDP settings with varying reward functions, Chen et al. (2022) introduced a transfer algorithm leveraging backward-style dynamic programming and one-step least-square regression, contrasting with our iterative approach. Agarwal et al. (2023) explored representation transfer, assuming a low-rank transition model, whereas we avoid specific assumptions about transition probabilities. Our theoretical contributions, highlighting the impact of task discrepancies on TL's statistical benefits in RL, enrich this body of work and enhance the understanding of TL in sequential decision-making.

Organization. The rest of this paper is organized as follows. Section 2 introduces the formulation of transfer learning in the stationary sequential decision setting, including a formal characterization of task discrepancy defined on MDPs. Section 3 presents batched Q^* learning with knowledge transfer under general function approximation, instantiated with sieve function approximation. Section 4 establishes theoretical guarantees for the proposed algorithm under both a transition homogeneous setting in Section 4.1 and a

generalized transition heterogeneous setting in Section 4.2. Sections 5.1 and 5.2 present empirical results with synthetic and real datasets, respectively. Computational details, proofs, and additional information are relegated to the supplementary material.

2 Statistical Framework

Mathematical Framework for RL under a Stationary Environment. We focus on data-driven decision-making, or reinforcement learning, under a stationary environment. The mathematical model for studying this type of problem is the discounted *Markov Decision Process* (MDP) which is characterized by a tuple $\mathcal{M} = \{\mathcal{X}, \mathcal{A}, P, r, \gamma\}$. We specifically focus on the setting with finite action space $\mathcal{A}$, i.e., $\mathcal{A} = \{1, 2, \ldots, m\}$ for a constant m. For a fixed trajectory index $i = 1, 2, \ldots, I$, at time $t = 0, 1, \ldots T$, an agent observes the current system state $\boldsymbol{X}_{i,t}$ supported on the state space $\mathcal{X}$, chooses a decision $A_{i,t}$ supported on the action space $\mathcal{A}$, transits to the next state $\boldsymbol{X}_{i,t+1}$ according to the system transition probability $P(\cdot \,|\, \boldsymbol{X}_{i,t}, A_{i,t})$, and receives an immediate reward $R_{i,t} = r(\boldsymbol{X}_{i,t}, A_{i,t}) + \eta_{i,t}$, where $r(\boldsymbol{x}, a)$ is a reward function, and $\eta_{i,t}$ denotes a zero-mean noise.

An agent's decision-making rule is characterized by a policy $\pi(a \mid \boldsymbol{x})$ that defines a distribution over actions conditional on states. A policy $\pi(a \mid \boldsymbol{x}) : \mathcal{X} \mapsto \mathcal{P}(\mathcal{A})$ is a function that maps the state space $\mathcal{X}$ to probability mass functions on the action space $\mathcal{A}$. It satisfies $\pi(a \mid \boldsymbol{x}) \geq 0$ for any $a \in \mathcal{A}$, $\boldsymbol{x} \in \mathcal{X}$, and $\sum_{a \in \mathcal{A}} \pi(a \mid \boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in \mathcal{X}$. Under policy π, at time t, a decision maker chooses action $A_{i,t} = j$ at state $\boldsymbol{X}_{i,t} = \boldsymbol{x}$ with probability $\pi(j \mid \boldsymbol{x})$. The trajectory i is then given by $\mathcal{H}_i = \{(\boldsymbol{X}_{i,0}, A_{i,0}), \cdots, (\boldsymbol{X}_{i,T}, A_{i,T})\}$, where the horizon T can be either finite or infinite. Let ν denote the distribution of the initial state $\boldsymbol{X}_{i,0}$. The trajectory distribution $\mathcal{P}_\pi$ given policy π is then given by

$$\mathcal{P}_\pi(\mathcal{H}_i) = \nu(\boldsymbol{X}_{i,0}) \prod_{t=0}^{T-1} \pi\left(A_{i,t} | \boldsymbol{X}_{i,t}\right) P\left(\boldsymbol{X}_{i,t+1} | \boldsymbol{X}_{i,t}, A_{i,t}\right). \tag{1}$$

The goal of RL is to learn an optimal policy that maximizes the discounted accumulative reward, or *return*, defined as

$$R_T = \sum_{t=0}^{T} \gamma^t R_{i,t}.$$

The horizon T can be either finite (episodic) or infinite (continuing), and the discount factor $\gamma \in [0, 1)$ reflects a trade-off between immediate and future rewards. If $\gamma = 0$, the decision maker chooses actions that maximize the immediate reward $R_{i,0}$. As γ increases, the decision maker puts more weight on future rewards. The value $T \cdot (1 - \gamma)^{-1}$ can be viewed as the effective horizon length of an MDP.

Given a policy π and a discount factor $\gamma \in [0, 1)$, the *state-value function* is the expectation of the total return starting from a state $\boldsymbol{x}$:

$$V^{\pi}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{H} \sim \mathcal{P}_{\pi}(\mathcal{H})} \left[\sum_{t=0}^{T} \gamma^t R_{i,t} \middle| \boldsymbol{X}_{i,0} = \boldsymbol{x} \right], \tag{2}$$

where the expectation is taken under the trajectory distribution generated by policy π on MDP $\mathcal{M}$. The *action-value function* or *Q-function* of a given policy π is defined as the expectation of the accumulated discounted rewards starting from a state $\boldsymbol{x}$ with action a:

$$Q^{\pi}(\boldsymbol{x}, a) = \mathbb{E}_{\mathcal{H} \sim \mathcal{P}_{\pi}(\mathcal{H})} \left[\sum_{t=0}^{T} \gamma^t R_{i,t} \middle| \boldsymbol{X}_{i,0} = \boldsymbol{x}, A_{i,0} = a \right], \tag{3}$$

where the expectation is taken by assuming that the dynamic system follows the given policy π afterward. The *optimal action-value function* Q^* is defined as

$$Q^*(\boldsymbol{x}, a) = \sup_{\pi} Q^{\pi}(\boldsymbol{x}, a), \quad \forall (\boldsymbol{x}, a) \in \mathcal{X} \times \mathcal{A}. \tag{4}$$

where the supremum is taken over all policies. One important property of Q^* is that it satisfies the *Bellman optimal equation*:

$$\mathbb{E}\Big[R_{i,t} + \gamma \max_{a' \in \mathcal{A}} Q^*(\boldsymbol{X}_{i,t+1}, a') - Q^*(\boldsymbol{X}_{i,t}, A_{i,t}) \,\Big|\, \boldsymbol{X}_{i,t}, A_{i,t} \Big] = 0. \tag{5}$$

For any given action-value function $Q : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$, the *greedy policy* π^Q is defined as the policy that selects the action with the largest Q-value, i.e.,

$$\pi^{Q}(a \mid \boldsymbol{x}) = 0 \quad \text{if} \quad a \notin \underset{a'}{\operatorname{argmax}}\, Q(\boldsymbol{x}, a'). \tag{6}$$

The goal of this paper is to study knowledge transfer under posterior shifts in learning the Q^* function from offline datasets. Once a better estimator of the Q^* function is constructed with the assistance from source data, a better optimal policy π^* can be derived as any policy that is greedy with respect to the estimated Q^*.

The Target and Source RL Data. Transferred RL aims to improve the learning on a target RL task by leveraging data from similar source RL tasks. We consider the case where we have abundant source data from offline observational data or simulated data, while the target task only has a small amount of offline data. Specifically, we have a target task and K source tasks, which are characterized by MDPs $\mathcal{M}^{(k)} = \{\mathcal{X}, \mathcal{A}, P^{(k)}, r^{(k)}, \gamma\}$ for $k \in \{0\} \cup [K]$. The target RL task of interest is referred to as the 0-th task and denoted by a superscript "(0)," while the source RL tasks are denoted by a superscript "(k)," for $k \in [K]$.

Without loss of generality, we assume the horizon length of all tasks is the same, denoted as T. For each task $k \in \{0\} \cup [K]$, we collect $I^{(k)}$ i.i.d. trajectories of length T, denoted as $\{(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}, R_{i,t}^{(k)}, \boldsymbol{X}_{i,t+1}^{(k)})\}$, $1 \le i \le I^{(k)}$, $0 \le t \le T-1$. We also assume that the trajectories in different tasks are independent.

Single-task RL considers each task $k \in \{0\} \cup [K]$ separately and defines the underlying *true response* of interest at step t as

$$Y_{i,t}^{(k)} := R_{i,t}^{(k)} + \gamma \cdot \max_{a' \in \mathcal{A}} Q^{*(k)}(\boldsymbol{X}_{i,t+1}^{(k)}, a'), \tag{7}$$

where $Q^{*(k)}$ denotes the optimal action-value function of task k. According to (5), we have

$$Q^{*(k)}(\boldsymbol{x}, a) = \mathbb{E}\left[Y_{i,t}^{(k)} \mid \boldsymbol{X}_{i,t}^{(k)} = \boldsymbol{x},\ A_{i,t}^{(k)} = a\right], \quad \text{for} \quad k \in \{0\} \cup [K], \tag{8}$$

which provides a moment condition for the estimation of $Q^{*(k)}(\boldsymbol{x}, a)$. If $Y_{i,t}^{(k)}$ is directly

observable, then $Q^{*(k)}(\boldsymbol{x}, a)$ can be estimated via regression. However, what we observe in the RL setting is only a "partial response" $R_{i,t}^{(k)}$. The other component of $Y_{i,t}^{(k)}$, as shown in the second term on the RHS of (7), depends on the unknown Q^* function and future observations. As will be discussed in detail in Section 3, we estimate $Q^{*(k)}(\boldsymbol{x}, a)$ in an iterative fashion.

Similarity Measure for Transferring between Different MDPs. The study of transfer learning necessitates a formal characterization of task similarity or difference. Since MDPs are characterized by tuples $\mathcal{M}^{(k)} = \left\{\mathcal{X}, \mathcal{A}, P^{(k)}, r^{(k)}, \gamma\right\}$ for $k \in \{0\} \cup [K]$, we characterize the similarity between the target and the source tasks through the difference between the reward functions $r^{(k)}$ and the transition probabilities $P^{(k)}$. Specifically, we denote $\rho^{(k)}$ as the density of the transition kernel $P^{(k)}$ and assume it exists almost everywhere for all k. The discrepancy between any source task k and the target task 0 is quantified by

$$\delta_r^{(k)}(\boldsymbol{x}, a) := r^{(k)}(\boldsymbol{x}, a) - r^{(0)}(\boldsymbol{x}, a), \tag{9}$$

$$\delta_\rho^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a) := \rho^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a) - \rho^{(0)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a). \tag{10}$$

Since our estimation target is the Q^* function, we next establish the relationship between the difference of the Q^* function and task discrepancy defined on MDP tuples.

Lemma 2.1 (Difference of Q^*)**.** *Let the difference between the optimal action-value functions across different tasks be defined as*

$$\delta_Q^{(k)}(\boldsymbol{x}, a) := Q^{*(k)}(\boldsymbol{x}, a) - Q^{*(0)}(\boldsymbol{x}, a). \tag{11}$$

Assume that the reward functions $r^{(k)}(\boldsymbol{x}, a)$ are uniformly upper bounded by a constant $R_{\max}$. Then we have

$$\sup_{\boldsymbol{x},a} \left|\delta_Q^{(k)}(\boldsymbol{x}, a)\right| \leq \frac{1}{1-\gamma} \sup_{\boldsymbol{x},a} \left|\delta_r^{(k)}(\boldsymbol{x}, a)\right| + \frac{\gamma R_{\max}}{(1-\gamma)^2} \int_{\mathcal{X}} \sup_{\boldsymbol{x},a} \left|\delta_\rho^{(k)}(\boldsymbol{x}' \mid \boldsymbol{x}, a)\right| \mathrm{d}\boldsymbol{x}'. \tag{12}$$

Lemma 2.1 shows that the magnitude of the difference of Q^* functions can be upper bounded by that of δ_r and δ_ρ, which theoretically guarantees the transferability across RL tasks that are similar in reward functions and transition kernels for optimal Q^* learning.

Indeed, the premise of transfer learning is that the differences δ_r and δ_ρ are "small" in the sense that the bias incurred from different source tasks can be "easily" corrected even with a small amount of target data, which will be shown later in the theoretical analysis in Section 4. For a better illustration, the magnitude of δ_r and δ_ρ will be instantiated under a specific function class and will show up in the statistical learning rate of our proposed algorithm, where we provide a formal quantification for how "small" the differences should be to benefit from transferring. As the rewards and transition pairs for all of the stages are *directly observable*, it can be *verified* in practice whether similarity assumptions to be imposed on (9) and (10) are satisfied (Silver et al., 2021).

3 Batch Q^* Learning with Knowledge Transfer

3.1 Transferred FQI with General Function Approximation

The proposed knowledge transfer algorithm is based on the framework of *Fitted Q-Iteration* (FQI) due to its wide applications in offline RL. The framework of FQI aims to minimize the *Bellman error* (BE) by bootstrapping and semi-gradient method. Recall that the Bellman optimal equation is

$$\mathbb{E}\left[R_{i,t}+\gamma\max_{a'\in\mathcal{A}}Q^*\left(\boldsymbol{X}_{i,t+1},a'\right)\big|\boldsymbol{X}_{i,t},A_{i,t}\right]=Q^*\left(\boldsymbol{X}_{i,t},A_{i,t}\right). \tag{13}$$

The FQI proceeds iteratively with a function class $\widehat{Q}\big(\boldsymbol{x},a;\boldsymbol{\beta}\big)$ parameterized by $\boldsymbol{\beta}$ to approximate $Q^*\left(\boldsymbol{x},a\right)$. In the τ-th iteration, given an estimator $\widehat{Q}\big(\boldsymbol{x},a;\widehat{\boldsymbol{\beta}}_{\tau-1}\big)$, the FQI computes $Y_{i,t}^{\tau}=R_{i,t}+\gamma\max_{a'\in\mathcal{A}}\widehat{Q}\big(\boldsymbol{X}_{i,t+1},a';\widehat{\boldsymbol{\beta}}_{\tau-1}\big)$ as a pseudo-response variable and regresses $\{Y_{i,t}^{\tau}\}$ on

$\{(\boldsymbol{X}_{i,t}, A_{i,t})\}$ to obtain an updated estimator $\widehat{Q}\big(\boldsymbol{x}, a; \widehat{\boldsymbol{\beta}}_\tau\big)$. Specifically,

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{I} \sum_{t=0}^{T-1} \big(Y_{i,t}^\tau - \widehat{Q}\left(\boldsymbol{X}_{i,t}, A_{i,t}; \boldsymbol{\beta}\right)\big)^2. \tag{14}$$

Based on the iterative framework of FQI, we develop the Transferred FQI algorithm (Algorithm 1) to apply knowledge transfer across different batch RL tasks. Suppose that we obtain samples $\mathcal{S}^{(k)} = \big\{\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}, R_{i,t}^{(k)}, \boldsymbol{X}_{i,t+1}^{(k)}\big)\big\}$ for $1 \le i \le I^{(k)}, 0 \le t \le T-1$ that are randomly sampled from the k-th task, where $k = 0$ represents the target task and $k \in [K]$ represents the k-th source task. As illustrated in Section 2, the samples are assumed to be i.i.d. across trajectory i but correlated across t for the same i. For each k, we randomly divide $\mathcal{S}^{(k)}$ into Υ disjoint subsets $\mathcal{S}_1^{(k)}, \mathcal{S}_2^{(k)}, \ldots, \mathcal{S}_\Upsilon^{(k)}$ and use $n_k = I^{(k)}T/\Upsilon$ to denote the sample size of each subset. The subsets $\{\mathcal{S}_\tau^{(k)}\}_{k=0}^K$ contain the data we used in the τ-th iteration of our proposed algorithm, where $\tau = 1, 2, \ldots, \Upsilon$.

On the population level, there exists a center of the Q^* functions $\{Q^{*(k)}\}_{k=0}^K$, defined as

$$W^* = \underset{W}{\operatorname{argmin}}\, \mathbb{E}\bigg[\sum_{i,t,k} \Big(Y_{i,t}^{(k)} - W\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\Big)^2\bigg],$$

where $Y_{i,t}^{(k)}$ is the true response defined by (7). It is straightforward to derive that

$$W^* = (n_{\mathcal{K}})^{-1} \sum_{k=0}^{K} n_k \cdot Q^{*(k)} = Q^{*(0)} + (n_{\mathcal{K}})^{-1} \sum_{k=0}^{K} n_k \cdot \delta_Q^{(k)}, \tag{15}$$

where $n_{\mathcal{K}} := \sum_{k=0}^K n_k$, and $\delta_Q^{(k)}$ is the defined in (11). The weighted average $\delta^{(0)} := (n_{\mathcal{K}})^{-1} \sum_k n_k \cdot \delta_Q^{(k)}$ characterizes the bias of the center W^* from the target function $Q^{*(0)}$. If each $\delta_Q^{(k)}$ is "sufficiently small," then their weighted average is also small, and we expect to learn the bias $\delta^{(0)}$ well even with a small amount of target data.

To estimate the optimal action-value function Q^*, we further need an *approximating space* $\mathcal{Q}$, a well-defined function class on $\mathcal{X} \times \mathcal{A}$. Given a general approximating space $\mathcal{Q}$, we denote the projection of the optimal action-value function $Q^{*(k)}$ and the center W^* on $\mathcal{Q}$ by $\widehat{Q^{*(k)}}$ and $\widehat{W^*}$, respectively. The above equations also hold for the projections

$\widehat{Q^{*(k)}}$ and $\widehat{W^*}$, so we estimate $\widehat{W^*}$ by minimizing the empirical L_2 loss in Step I (Equation (16)) of Algorithm 1. Since an informative source task must be similar to the target task, the approximating space for the bias $\delta^{(0)}$, denoted as $\mathcal{Q}' \subset \mathcal{Q}$, is often more restrictive such that we can employ this restrictive structure to estimate $\delta^{(0)}$ well even with a small amount of target data. In literature, restrictive structures that characterize task similarity include the sparse condition (Li et al., 2023), smoothness condition (Cai and Wei, 2021), polynomial order (Cai and Pu, 2024), and RKHS norms (Wang et al., 2023). In Step II of Algorithm 1, we denote the restrictive structure imposed by $\mathcal{Q}'$ as a norm $\|\cdot\|_{\mathcal{Q}'}$ and minimize a $\|\cdot\|_{\mathcal{Q}'}$-penalized objective in (18) to obtain an estimator for the bias $\delta^{(k)}$ on each task. At the end of each iteration, the optimal action-value functions $Q^{*(k)}$ are estimated by combining the center estimator (16) with the bias-correction estimator (18) on each task, where, in particular, $\widehat{Q}^{(0)}$ is our goal estimator for the target task. These estimators are further refined as the iterations proceed.

3.2 Transferred FQI with Sieve Function Approximation

Algorithm 1 establishes a general framework for transferred FQI. With different applications, the approximating space $\mathcal{Q}$ can be chosen according to the norm of the domain. For example, for language and vision tasks, neural networks are usually chosen for the approximating space $\mathcal{Q}$ due to the intrinsic data structure and the availability of massive training data. However, semi-parametric sieve approximation has been proven to offer better approximation and sensible interpretations for applications of business, economics, and finance (Chen, 2007). To propel data-driven decision in societal applications, we focus our attention on sieve function approximation hereafter.

Now we instantiate Algorithm 1 with sieve function approximation. The approximating space $\mathcal{Q}$ is chosen to be linear combinations of sieve basis functions, i.e., $\mathcal{Q}$ contains

Algorithm 1: TransFQI: Transferred Fitted Q-Iteration (General)

Input:

- Target data $\mathcal{S}^{(0)} = \big\{\big(\boldsymbol{X}_{i,t}^{(0)}, A_{i,t}^{(0)}, R_{i,t}^{(0)}, \boldsymbol{X}_{i,t+1}^{(0)}\big)\big\}$, $1 \le i \le I^{(0)}$, $0 \le t \le T-1$;
- Informative source data $\mathcal{S}^{(k)} = \big\{\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}, R_{i,t}^{(k)}, \boldsymbol{X}_{i,t+1}^{(k)}\big)\big\}$, $1 \le i \le I^{(k)}$, $0 \le t \le T-1$, $1 \le k \le K$;
- An approximation space $\mathcal{Q}$ for Q^* and a function class $\mathcal{Q}'$ for the difference δ;
- Initial estimators $\widehat{Q}_0^{(k)} \in \mathcal{Q}$;
- Maximum number of iterations Υ;
- Regularization parameters $\lambda_\delta^{(k)}$ for $k \in \{0\} \cup [K]$.

Output: An estimator $\widehat{Q}_\Upsilon^{(0)}$ and the greedy policy $\pi_\Upsilon^{(0)}$ accordingly.

For $k = 0, 1, \ldots, K$, randomly divide $\mathcal{S}^{(k)}$ into Υ disjoint subsets $\mathcal{S}_1^{(k)}, \mathcal{S}_2^{(k)}, \ldots, \mathcal{S}_\Upsilon^{(k)}$;
for $\tau = 1, 2, \ldots, \Upsilon$ **do**

Compute $Y_{i,t}^{(k),\tau} = R_{i,t}^{(k)} + \gamma \cdot \max_{a' \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}\big(\boldsymbol{X}_{i,t+1}^{(k)}, a'\big)$ for all $(i,t) \in \mathcal{S}_\tau^{(k)}$ and $k \in \{0\} \cup [K]$.

Step I. Compute an aggregated estimator for all tasks:

$$\widehat{W}_\tau = \operatorname*{argmin}_{W \in \mathcal{Q}} \Big(\frac{1}{2n_{\mathcal{K}}} \sum_{k=0}^{K} \sum_{(i,t) \in \mathcal{S}_\tau^{(k)}} \Big[Y_{i,t}^{(k),\tau} - W\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big) \Big]_2^2 \Big). \tag{16}$$

Step II. Compute a corrected target estimator for each task:

for $k = 0, 1, \ldots, K$ **do**

Obtain

$$\widehat{Q}_\tau^{(k)} = \widehat{W}_\tau + \widehat{\delta}_\tau^{(k)}, \tag{17}$$

where

$$\widehat{\delta}_\tau^{(k)} = \operatorname*{argmin}_{\delta \in \mathcal{Q}'} \Big(\frac{1}{2n_k} \sum_{(i,t) \in \mathcal{S}_\tau^{(k)}} \Big[Y_{i,t}^{(k),\tau} - \widehat{W}_\tau(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}) - \delta(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}) \Big]_2^2 + \lambda_\delta^{(k)} \|\delta\|_{\mathcal{Q}'} \Big), \tag{18}$$

and $\|\cdot\|_{\mathcal{Q}'}$ is a function norm that imposes structures on the task difference.

functions of the form $\widehat{Q}(\boldsymbol{x}, a; \boldsymbol{\beta}) = \boldsymbol{\xi}(\boldsymbol{x}, a)^\top \boldsymbol{\beta}$, where

$$\boldsymbol{\xi}(\boldsymbol{x}, a) := \big[\boldsymbol{\phi}^\top(\boldsymbol{x}) \mathbb{I}(a = 1), \boldsymbol{\phi}^\top(\boldsymbol{x}) \mathbb{I}(a = 2), \ldots, \boldsymbol{\phi}^\top(\boldsymbol{x}) \mathbb{I}(a = m)\big]^\top, \tag{19}$$

and $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_p(\cdot))^\top$ denotes a set of pre-selected sieve basis functions such as B-splines or wavelets.

With sieve approximation, Algorithm 1 is instantiated in several aspects. Firstly, the initial estimator is characterized by $\widehat{Q}_0^{(k)}(\boldsymbol{x}, a) = \boldsymbol{\xi}(\boldsymbol{x}, a)^\top \widehat{\boldsymbol{\beta}}_0^{(k)}$, where $\widehat{\boldsymbol{\beta}}_0^{(k)}$ is an (mp)-dimensional initial coefficient vector for the k-th task. In the τ-th iteration, after computing the pseudo response variable $\{Y_{i,t}^{(k),\tau}\}$ for all of the tasks, the aggregated estimator $\widehat{W}_\tau(\boldsymbol{x}, a) = \boldsymbol{\xi}(\boldsymbol{x}, a)^\top \widehat{\boldsymbol{w}}_\tau$, where $\widehat{\boldsymbol{w}}_\tau$ is obtained from sieve-instantiated Equation (16) in Step I, that is,

$$\widehat{\boldsymbol{w}}_\tau = \underset{\boldsymbol{w}\in\mathbb{R}^{mp}}{\operatorname{argmin}} \frac{1}{2n_{\mathcal{K}}} \sum_{k=0}^{K} \sum_{(i,t)\in\mathcal{S}_\tau^{(k)}} \left[Y_{i,t}^{(k),\tau} - \boldsymbol{\xi}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)^\top \boldsymbol{w} \right]_2^2. \tag{20}$$

Due to the linearity of sieve spaces, the aggregated estimator $\widehat{\boldsymbol{w}}_\tau$ can be viewed as an estimator for a weighted average of the underlying parameters of the K tasks, which is biased from the parameter of each task. In Step II, we estimate the bias of $\widehat{\boldsymbol{w}}_\tau$ on each task by an ℓ_1-regularized estimator $\widehat{\boldsymbol{\delta}}_\tau^{(k)}$, as we use ℓ_1 distance to measure the difference across tasks under sieve function approximation. That is,

$$\widehat{\boldsymbol{\delta}}_\tau^{(k)} = \underset{\boldsymbol{\delta}\in\mathbb{R}^{mp}}{\operatorname{argmin}} \left(\frac{1}{2n_k} \sum_{(i,t)\in\mathcal{S}_\tau^{(k)}} \left[Y_{i,t}^{(k),\tau} - \boldsymbol{\xi}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)^\top (\widehat{\boldsymbol{w}}_\tau + \boldsymbol{\delta}) \right]_2^2 + \lambda_\delta^{(k)} \|\boldsymbol{\delta}\|_1 \right). \tag{21}$$

Then we correct the aggregated estimator by $\widehat{\boldsymbol{\beta}}_\tau^{(k)} = \widehat{\boldsymbol{w}}_\tau + \widehat{\boldsymbol{\delta}}_\tau^{(k)}$, and the corrected estimator $\widehat{\boldsymbol{\beta}}_\tau^{(k)}$ is input into the next iteration for further refinement.

3.3 Distinction of Transferred FQI from Transferred SL

In the realm of RL, FQI represents a pivotal algorithmic approach for estimating the optimal action-value function Q^*. When considering transferred FQI within the context of transfer learning, it becomes evident that several key distinctions arise compared to algorithms designed for transferred supervised learning (SL).

Firstly, while transferred SL often relies on non-iterative regression-based estimation methods, FQI adopts a fundamentally different approach. Rather than relying on re-

gression, FQI addresses the problem through an iterative fixed-point solving process. This iterative nature not only distinguishes FQI from conventional regression-based methods but also introduces novel challenges and opportunities within the transfer learning paradigm. Consequently, our transferred FQI algorithm and its associated theoretical properties introduce new contributions to the landscape of transfer learning, particularly within the dynamic domain of RL.

Moreover, in transferred SL, the primary focus typically revolves around the target task, without the need to construct better estimators for the source tasks. However, the nature of RL introduces additional complexities, as the response of the optimal action-value function is not directly observable. Consequently, in transferred RL scenarios, the construction of pseudo responses for the source tasks becomes indispensable. Despite the primary emphasis on the target task, our algorithm adeptly handles the construction of pseudo responses for the source tasks as well. This dual focus not only underscores the versatility of our approach but also enables the derivation of estimators for the Q^* function of the source tasks as byproducts. This holistic perspective enhances the scope and applicability of the proposed transferred FQI algorithm, extending its utility beyond traditional transfer learning methodologies.

4 Theory

In this section, we establish statistical guarantees for our proposed transferred FQI algorithm (Algorithm 1) under sieve function approximation. As defined in Section 2, the dataset of task $k \in \{0\} \cup [K]$ contains $I^{(k)}$ independent trajectories, denoted by $\left\{\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}, R_{i,t}^{(k)}, \boldsymbol{X}_{i,t+1}^{(k)}\right)\right\}$ $(1 \leq i \leq I^{(k)}, 0 \leq t \leq T-1)$, each with length T.

4.1 Theoretical Results for Transition Homogeneous Tasks

To clearly present our theoretical results, we first focus on the transition homogeneous setting, where the state-action variables are assumed to share the same distribution, i.e., the transition $P^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)$ are the same across $k \in \{0\} \cup [K]$. In contrast, the reward functions $r^{(k)}(\boldsymbol{x}, a)$ are different across different tasks.

We first introduce some common regularity conditions for the theoretical development.

Assumption 4.1. *For all tasks $k \in \{0\} \cup [K]$, assume the following conditions hold:*

(a) Assume that the state-action chain $\big\{\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\big\}_{t=0}^{T-1}$ satisfies time-homogeneous Markov property. Specifically, given $\boldsymbol{X}_{i,t}^{(k)} = \boldsymbol{x}$ and $A_{i,t}^{(k)} = a$, assume that the distribution of the next state $\boldsymbol{X}_{i,t+1}^{(k)}$ is determined by a time-invariant transition kernel $P(\cdot \,|\, \boldsymbol{x}, a)$ with density $\rho(\cdot \,|\, \boldsymbol{x}, a)$ almost everywhere. Moreover, assume that there exists a behavior policy $b(\cdot \,|\, \boldsymbol{x})$ such that $\mathbb{P}\big(A_{i,t}^{(k)} = a \mid \boldsymbol{X}_{i,t}^{(k)} = \boldsymbol{x}\big) = b(a \,|\, \boldsymbol{x})$.

(b) Assume that the Markov chain $\big\{\boldsymbol{X}_{i,t}^{(k)}\big\}_{t=0}^{T-1}$ has a unique stationary distribution with a density μ almost everywhere. Let ν denote the probability density of the initial state $\boldsymbol{X}_{i,0}^{(k)}$. Assume that μ and ν are bounded away from 0 and ∞. Furthermore, assume that the Markov chain $\big\{\boldsymbol{X}_{i,t}^{(k)}\big\}_{t=0}^{T-1}$ is geometrically ergodic if $T \to \infty$.

(c) Let $\boldsymbol{\Sigma} := \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1} \boldsymbol{\xi}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\boldsymbol{\xi}^{\top}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\Big]$, where $\boldsymbol{\xi}(\boldsymbol{x}, a)$ is the basis function defined in (19). Assume that there exists a constant $c_\Sigma \geq 1$ such that $c_\Sigma^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_\Sigma$ for all k.

(d) Assume that the reward $R_{i,t}^{(k)} = r^{(k)}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big) + \eta_{i,t}^{(k)}$, where the noise $\eta_{i,t}^{(k)}$ is σ_η^2-sub-Gaussian with a constant $\sigma_\eta > 0$ that does not depend on i, t, k.

In Assumption 4.1, the condition (a) ensures that, for any i, k, $\big\{\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\big\}$ is a time-homogeneous MDP on $\mathcal{X} \times \mathcal{A}$ with the transition kernel $P(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)b(a' \,|\, \boldsymbol{x}')$, and thus

$\{\boldsymbol{X}_{i,t}^{(k)}\}$ is an MDP with the transition kernel $\sum_{a\in\mathcal{A}} P(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)b(a \,|\, \boldsymbol{x})$. The condition (b) further assumes that the latter MDP has a stationary distribution with bounded density and is geometrically ergodic, i.e., the stationarity can be attained at a geometric rate when $T \to \infty$. The condition (c) assumes the invertibility and boundedness of the population covariance matrices on all tasks. These conditions are consistent with Assumptions (A2) and (A3) in Shi et al. (2022), which are necessary to ensure the concentration properties between the empirical and population quantities. The condition (d) assumes the sub-Gaussianity of the noise in the reward, which is also standard to guarantee the concentration properties.

We then introduce the notion of Hölder κ-smooth function, which is a generalization of Lipschitz continuity and is widely used to characterize the regularity of functions.

Definition 4.2 (Hölder κ-smooth functions)**.** *Let $f(\cdot)$ be an arbitrary function on $\mathcal{X} \in \mathbb{R}^d$. For a d-tuple $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)$ of non-negative integers, let D^α denote the differential operator*

$$D^{\boldsymbol{\alpha}} f(\boldsymbol{x}) = \frac{\partial^{\|\boldsymbol{\alpha}\|_1} f(\boldsymbol{x})}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}},$$

where $\boldsymbol{x} = (x_1, \cdots, x_d)^\top$. The class of κ-smooth functions is defined as

$$\Lambda(\kappa, c) = \Big\{ f : \sup_{\|\boldsymbol{\alpha}\|_1 \le \lfloor\kappa\rfloor} \sup_{\boldsymbol{x}\in\mathcal{X}} |D^{\boldsymbol{\alpha}} f(\boldsymbol{x})| \le c \ \text{ and } \sup_{\|\boldsymbol{\alpha}\|_1 = \lfloor\kappa\rfloor} \sup_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \boldsymbol{x}_1 \neq \boldsymbol{x}_2} \frac{|D^{\boldsymbol{\alpha}} f(\boldsymbol{x}_1) - D^{\boldsymbol{\alpha}} f(\boldsymbol{x}_2)|}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^{\kappa - \lfloor\kappa\rfloor}} \le c \Big\}.$$

With a set of typical sieve basis functions $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1, \ldots, \phi_p)$ such as B-splines and wavelets, the Hölder κ-smooth functions satisfy the following property (Huang, 1998): for any function $f(\cdot) \in \Lambda(\kappa, c)$, there exist coefficients $\boldsymbol{\beta}$ such that

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left| f(\boldsymbol{x}) - \boldsymbol{\beta}^\top \boldsymbol{\phi}(\boldsymbol{x}) \right| \le C p^{-\kappa/d}, \tag{22}$$

for some positive constant C. We then assume the following regularity condition on the reward functions and the transition kernels to ensure the Hölder κ-smoothness of the Q^*

function.

Assumption 4.3. *Assume that there exist some constants κ, $c > 0$ such that $r^{(k)}(\cdot, a)$ and $\rho(\boldsymbol{x}' \,|\, \cdot, a)$ belong to $\Lambda(\kappa, c)$ for any $a \in \mathcal{A}$, $k \in \{0\} \cup [K]$, and $\boldsymbol{x}' \in \mathcal{X}$. In particular, there exists a uniform upper bound $R_{\max}$ on the reward functions $r^{(k)}(\boldsymbol{x}, a)$, i.e., $\sup_{\boldsymbol{x},a,k} \left| r^{(k)}(\boldsymbol{x}, a) \right| \leq R_{\max}$.*

Lemma 4.4. *Under Assumption 4.3, there exists some constant $c' > 0$ such that the optimal action-value function $Q^{*(k)}(\cdot, a)$ belongs to the class $\Lambda(\kappa, c')$ for any task $k \in \{0\} \cup [K]$ and any action $a \in \mathcal{A}$. In particular, $\sup_{\boldsymbol{x},a,k} \left| Q^{*(k)}(\boldsymbol{x}, a) \right| \leq R_{\max}/(1-\gamma)$.*

Lemma 4.4 ensures that, under Assumption 4.3, the optimal $Q^{*(k)}$ function is κ-smooth for all tasks and, in particular, uniformly upper bounded in ℓ_∞ norm. Using the sieve approximation, we further characterize the transferability across different tasks. Recall that $\boldsymbol{\xi}(\boldsymbol{x}, a) := \left[\boldsymbol{\phi}^\top(\boldsymbol{x})\mathbb{I}(a=1), \boldsymbol{\phi}^\top(\boldsymbol{x})\mathbb{I}(a=2), \ldots, \boldsymbol{\phi}^\top(\boldsymbol{x})\mathbb{I}(a=m)\right]^\top$. From (22) and Assumption 4.3, we have that there exist coefficients $\left\{\boldsymbol{\beta}_r^{(k)}\right\}$ such that

$$\sup_{\boldsymbol{x},a,k} \left| r^{(k)}(\boldsymbol{x}, a) - \boldsymbol{\xi}^\top(\boldsymbol{x}, a)\boldsymbol{\beta}_r^{(k)} \right| \leq C p^{-\kappa/d},$$

for some constant C. The similarity between $r^{(k)}(\boldsymbol{x}, a)$ and $r^{(0)}(\boldsymbol{x}, a)$ is manifested through their projections on the sieve space. Specifically, we use

$$h_r := \max_{k \in [K]} \left\| \boldsymbol{\beta}_r^{(k)} - \boldsymbol{\beta}_r^{(0)} \right\|_1 \tag{23}$$

to measure the discrepancy between K source tasks and the target task. The level h_r quantifies the difference between the target and source tasks. As long as h_r is small enough, the source tasks are sufficiently informative to improve the estimation performance on the target task, which is shown by the following theorem for the theoretical property of Algorithm 1 under transition homogeneity:

Theorem 4.5. *Suppose Assumptions 4.1 and 4.3 hold. Further assume that the sample sizes n_0 and $n_{\mathcal{K}}$ satisfy $\frac{p \log^2 n_{\mathcal{K}}}{n_{\mathcal{K}}} + h_r \sqrt{\frac{\log p}{n_0}} = o(1)$. By choosing the initial estimators such*

that $\sup_{\boldsymbol{x},a,k} \widehat{Q}_0^{(k)}(\boldsymbol{x},a) \le R_{\max}/(1-\gamma)$ *and* $\sup_k \left\|\widehat{\boldsymbol{\beta}}_0^{(k)} - \widehat{\boldsymbol{\beta}}_0^{(0)}\right\|_1 \le h_r$ *and choosing the tuning parameter* $\lambda_\delta^{(k)} = c_\delta\sqrt{\frac{\log p}{n_k}}$ *for some sufficiently large constant* c_δ*, it holds that*

$$\mathbb{E}\left|Q^* - \widehat{Q}_\Upsilon^{(0)}\right| = \mathcal{O}_p\Bigg(\frac{\gamma}{(1-\gamma)^2}\Big[\underbrace{p^{-\kappa/d}}_{\substack{\text{function-}\\ \text{approximation}\\ \text{bias}}} + \underbrace{\sqrt{\frac{p\log^2 n_{\mathcal{K}}}{n_{\mathcal{K}}}}}_{\substack{\text{commonality}\\ \text{estimation error}}} + \underbrace{\sqrt{h_r}\left(\frac{\log p}{n_0}\right)^{1/4}}_{\text{task-difference bias}}\Big] + \underbrace{\frac{\gamma^{\Upsilon+1}R_{\max}}{(1-\gamma)^2}}_{\text{algorithmic error}}\Bigg), \tag{24}$$

where $\widehat{Q}_\Upsilon^{(0)}$ *is the output estimator of Algorithm 1, and the expectation is taken over* $(\boldsymbol{x},a) \sim \mu(\boldsymbol{x})b(a\,|\,\boldsymbol{x})$.

Theorem 4.5 establishes that the convergence rate of the Q^* estimator contains a statistical rate (the first three terms) and an algorithmic rate (the last term). The statistical rate is weighted by $\frac{\gamma}{(1-\gamma)^2}$, which is incurred by the iterative nature of the fixed-point solution. The algorithmic error is due to the error of the initial estimator, which decreases exponentially as the iterations proceed since $\gamma < 1$. By carefully choosing the number of iterative steps Υ, the statistical error will dominate the algorithmic error.

We now focus on discussing the statistical rates in the brackets in (24). The first term $p^{-\kappa/d}$ is the function approximation bias, instantiated as the bias of non-parametric estimation in (22). Due to the careful design of our algorithm, this estimation bias does not accumulate over different tasks. It is consistent with the bias agreed in the literature on sieve approximation. We conjecture that this function approximation error will be instantiated by other estimation biases if we consider different function classes $\mathcal{Q}$ and approximation techniques, such as neural network function approximation. We leave the detailed analysis of other function approximation methods for future works.

The second term $\sqrt{\frac{p\log^2 n_{\mathcal{K}}}{n_{\mathcal{K}}}}$ represents the statistical convergence rate of estimating the shared commonality between the target and the source tasks, i.e., the center W^* in (15), demonstrating itself in the order of standard deviation. For sieve approximation,

by choosing the number of basis functions $p > n_{\mathcal{K}}^{\frac{d}{2\kappa+d}}(\log n_{\mathcal{K}})^{\frac{-2d}{2\kappa+d}}$, this standard deviation term dominates function-approximation bias term $p^{-\kappa/d}$. Compared to the single task convergence rate $\sqrt{p/n_0}$, the commonality estimation error is improved into the convergence rate when the target sample size is as large as $n_{\mathcal{K}}$ (ignoring the logarithm term), which shows the advantage of transfer learning.

The third term $\sqrt{h_r}\left(\frac{\log p}{n_0}\right)^{1/4}$ is the task-difference bias. It is incurred by task discrepancy and mathematically characterized by h_r defined in (23). Note that although the theoretical conditions for the initial estimators in Theorem 4.5 are related to h_r, they can be easily satisfied in practice without knowing h_r, for example, by setting $\widehat{\boldsymbol{\beta}}_0^{(k)} = 0$ for all k, and the constant c_δ can be chosen by cross-validation.

The result in Theorem 4.5 provides several important insights into the practice of TL for offline RL. Firstly, for a given task difference h_r, it spells out the necessary numbers of target and source samples to correct the task-difference bias and to fully enjoy the benefit of transfer learning. Specifically, for a given h_r, the task-difference bias can be dominated by the commonality estimation error if practitioners collect $n_0 \gtrsim n_{\mathcal{K}}^2 h_r^2 p^{-2}$ samples for the target task, ignoring the logarithm term. This implies an artful tug-of-war between the source and the target tasks when task discrepancy exists: on the one hand, one wishes to have a large size of source data to improve the estimation accuracy of the common component shared by the source and target tasks; on the other hand, one needs to control the relative sample sizes of the source and target data such that the bias induced from the task difference can be corrected with n_0 target samples. The growth of the required n_0 is proportional to $n_{\mathcal{K}}^2 p^{-2}$ for any given h_r, which is not restrictive in the sense that the number of p is growing and $n_{\mathcal{K}}^2 p^{-2}$ can be much smaller than $n_{\mathcal{K}}$. Since we only require $\frac{p\log^2 n_{\mathcal{K}}}{n_{\mathcal{K}}} = o(1)$ in Theorem 4.5, the growth of p can be as fast as $n_{\mathcal{K}}/\log^2 n_{\mathcal{K}}$. As a result, with the help of transfer learning, the target RL task with sample size n_0 much smaller

than $n_{\mathcal{K}}$ can enjoy the convergence rate of $\sqrt{p/n_{\mathcal{K}}}$.

In cases where source tasks offer no data, meaning $n_{\mathcal{K}} = n_0$ and $h_r = 0$, the performance of our algorithm matches the statistical convergence rate of single-task non-parametric RL estimation, with a rate of $\mathcal{O}_p\Big(\frac{\gamma}{(1-\gamma)^2}\big(\sqrt{\frac{p\log^2 n_0}{n_0}}+p^{-\kappa/d}\big)\Big)$. As the source sample size increases relative to the target, specifically when $n_{\mathcal{K}} \gtrsim n_0$ and $h_r \lesssim \frac{p}{\sqrt{n_0 \log p}}$, our algorithm's accuracy improves, demonstrating the value of transferring knowledge from similar and sufficiently large source tasks. The condition on task discrepancy, $h_r \lesssim \frac{p}{\sqrt{n_0 \log p}}$, is mild, since the basis number p is allowed to be as large as $n_{\mathcal{K}}$, which allows the upper bound of h_r to be of the order $\frac{n_{\mathcal{K}}}{\sqrt{n_0}}$, ignoring the logarithm term.

4.2 Theoretical Results for Transition Heterogeneous Tasks

In this section, we generalize the results in Section 4.1 to transition heterogeneous tasks, allowing the distributions of $(\boldsymbol{X}, A)$ to differ in different tasks. Therefore, both the transition kernels $P^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)$ and the reward functions $r^{(k)}(\boldsymbol{x}, a)$ are different across tasks. The theoretical results explicitly depend on the differences between the target and the source tasks defined by (9) and (10). We first extend Assumption 4.1 to the heterogeneous setting.

Assumption 4.6. *For all tasks $k \in \{0\} \cup [K]$, assume the following conditions hold:*

(a') *Given $\boldsymbol{X}_{i,t}^{(k)} = \boldsymbol{x}$ and $A_{i,t}^{(k)} = a$, assume that the distribution of the next state $\boldsymbol{X}_{i,t+1}^{(k)}$ is determined by a time-invariant transition kernel $P^{(k)}\big(\cdot \,|\, \boldsymbol{x}, a\big)$ with density $\rho^{(k)}(\cdot \,|\, \boldsymbol{x}, a)$ almost everywhere. Moreover, assume that there exists a behavior policy $b^{(k)}(\cdot \,|\, \boldsymbol{x})$ such that $\mathbb{P}\big(A_{i,t}^{(k)} = a \,|\, \boldsymbol{X}_{i,t}^{(k)} = \boldsymbol{x}\big) = b^{(k)}(a \,|\, \boldsymbol{x})$ for all i, t.*

(b') *Assume that the Markov chain $\big\{\boldsymbol{X}_{i,t}^{(k)}\big\}_{t=0}^{T-1}$ has a unique stationary distribution with a density $\mu^{(k)}$ almost everywhere. Let $\nu^{(k)}$ denote the probability density of the initial*

state $\boldsymbol{X}_{i,0}^{(k)}$. *Assume that* $\mu^{(k)}$ *and* $\nu^{(k)}$ *are bounded away from 0 and* ∞. *Furthermore, assume that the Markov chain* $\{\boldsymbol{X}_{i,t}^{(k)}\}_{t=0}^{T-1}$ *is geometrically ergodic if* $T \to \infty$.

(c') *Let* $\boldsymbol{\Sigma}^{(k)} := \frac{1}{T}\mathbb{E}\Big[\sum_{t=0}^{T-1} \boldsymbol{\xi}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\boldsymbol{\xi}^{\top}\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\Big]$, *where* $\boldsymbol{\xi}(\boldsymbol{x}, a)$ *is the basis function defined in* (19). *Assume that there exists a constant* $c_\Sigma \geq 1$ *such that* $c_\Sigma^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}^{(k)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{(k)}) \leq c_\Sigma$ *for all* k.

(d) *The same as condition (d) in Assumption 4.1.*

Assumption 4.6 is parallel to Assumption 4.1, allowing the transition kernel $\rho^{(k)}$, the behavior policy $b^{(k)}$, the initial distribution $\nu^{(k)}$, and the invariant distribution $\mu^{(k)}$ to be different in different tasks. Similar to Assumption 4.1, the condition (a') ensures that, for any i, $\big\{\big(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\big)\big\}$ is a time-homogeneous MDP on $\mathcal{X} \times \mathcal{A}$ with the transition kernel $P^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)b^{(k)}(a' \,|\, \boldsymbol{x}')$, and thus $\big\{\boldsymbol{X}_{i,t}^{(k)}\big\}$ is an MDP with the transition kernel $\sum_{a\in\mathcal{A}} b^{(k)}(a \,|\, \boldsymbol{x})P^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)$. The condition (b') further assumes that the latter MDP has a stationary distribution $\mu^{(k)}$, which is allowed to be different for different k. The condition (c') requires a uniform boundedness for the eigenvalues of the population covariance matrices on all tasks. Additionally, we also require the Hölder κ-smoothness of the reward functions and transition kernels on all of the tasks.

Assumption 4.7. *Assume that there exist some constants* κ, $c > 0$ *such that* $r^{(k)}(\cdot, a)$ *and* $\rho^{(k)}(\boldsymbol{x}' \,|\, \cdot, a)$ *belong to* $\Lambda(\kappa, c)$ *for any* $a \in \mathcal{A}$, $k \in \{0\} \cup [K]$, *and* $\boldsymbol{x}' \in \mathcal{X}$. *In particular, there exists a uniform upper bound* $R_{\max}$ *on the reward functions, i.e.,* $\sup_{\boldsymbol{x},a,k} \big|r^{(k)}(\boldsymbol{x}, a)\big| \leq R_{\max}$.

We then characterize the transferability across different tasks in the heterogeneous setting. By (22) and Assumption 4.7, there exist coefficients $\big\{\boldsymbol{\beta}_r^{(k)}\big\}$ and $\big\{\boldsymbol{\beta}_\rho^{(k)}(\boldsymbol{x}')\big\}$ such that

$$\sup_{\boldsymbol{x},a,k} \Big|r^{(k)}(\boldsymbol{x}, a) - \boldsymbol{\xi}^{\top}(\boldsymbol{x}, a)\boldsymbol{\beta}_r^{(k)}\Big| \leq Cp^{-\kappa/d},$$

and

$$\sup_{\boldsymbol{x},a,\boldsymbol{x}',k} \left| \rho^{(k)}\left(\boldsymbol{x}' \,|\, \boldsymbol{x}, a\right) - \boldsymbol{\xi}^{\top}(\boldsymbol{x}, a)\boldsymbol{\beta}_{\rho}^{(k)}(\boldsymbol{x}') \right| \leq Cp^{-\kappa/d},$$

for some constant C. Similar to the transition homogeneous setting, the similarity of $r^{(k)}$ and $\rho^{(k)}$ is measured through their projections on the sieve space. Specifically, define

$$h := \max_{k} \left[\left\| \boldsymbol{\beta}_{r}^{(k)} - \boldsymbol{\beta}_{r}^{(0)} \right\|_1 + \int_{\mathcal{X}} \left\| \boldsymbol{\beta}_{\rho}^{(k)}(\boldsymbol{x}') - \boldsymbol{\beta}_{\rho}^{(0)}(\boldsymbol{x}') \right\|_1 \mathrm{d}\boldsymbol{x}' \right], \tag{25}$$

which is a generalization of h_r defined in (23) in the homogeneous setting.

In addition, since Q^* is approximated by $\boldsymbol{\xi}(\boldsymbol{x}, a)^{\top}\boldsymbol{\beta}_{Q}^{(k)}$, we also need to characterize the discrepancy between the population covariance matrices of $\boldsymbol{\xi}(\boldsymbol{x}, a)$. Let $\overline{\boldsymbol{\Sigma}} = \sum_{k=0}^{K} \frac{n_k}{n_{\mathcal{K}}} \boldsymbol{\Sigma}^{(k)}$, where $\boldsymbol{\Sigma}^{(k)}$ is defined in condition (c') in Assumption 4.6. Then we define

$$C_{\Sigma} := 1 + \max_{k} \left\| \overline{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\Sigma}^{(k)} - \overline{\boldsymbol{\Sigma}}) \right\|_1. \tag{26}$$

The quantity C_{Σ} is similar to the heterogeneity constant defined in Li et al. (2022). However, Li et al. (2022) characterized the differences between the covariance matrices $\boldsymbol{\Sigma}^{(k)}$ and $\boldsymbol{\Sigma}^{(0)}$ directly, while we characterize the differences between $\boldsymbol{\Sigma}^{(k)}$ and their weighted average for technical simplicity.

We are ready to present the theoretical property of Algorithm 1 in the transition heterogeneous setting.

Theorem 4.8. *Suppose Assumptions 4.6, 4.7, and the other conditions in Theorem 4.5 hold. Then we have*

$$\mathbb{E}\left| Q^* - \widehat{Q}_{\Upsilon}^{(0)} \right| = \mathcal{O}_p\Bigg(\frac{\gamma}{(1-\gamma)^2} \Bigg[\underbrace{p^{-\kappa/d}}_{\substack{\text{function-}\\ \text{approximation}\\ \text{bias}}} + \underbrace{\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}}}_{\substack{\text{commonality}\\ \text{estimation error}}} + \underbrace{\sqrt{C_{\Sigma} h}\left(\frac{\log p}{n_0}\right)^{1/4}}_{\text{task-difference bias}} \Bigg] + \underbrace{\frac{\gamma^{\Upsilon+1} R_{\max}}{(1-\gamma)^2}}_{\text{algorithmic error}} \Bigg), \tag{27}$$

where the expectation is taken over $(\boldsymbol{x}, a) \sim \mu^{(0)}(\boldsymbol{x}) b^{(0)}(a \,|\, \boldsymbol{x})$.

Compared to the convergence rate in the homogeneous setting (Theorem 4.5), Theorem

4.8 indicates that Algorithm 1 achieves a similar convergence rate for the transition heterogeneous tasks. Concretely, the function-approximation bias, commonality estimation error, and algorithmic error terms (i.e., the first, second, and fourth terms) are of the same rate, while the third term, which is due to the discrepancy among the tasks, is characterized by $C_\Sigma h$ instead of h_r.

Similar to Theorem 4.5, we have that, when Υ is sufficiently large, the target sample size n_0 needed to correct the task-difference bias is $n_{\mathcal{K}}^2 C_\Sigma^2 h^2 p^{-2}$, where p is allowed to grow as fast as $n_{\mathcal{K}}/\log^2 n_{\mathcal{K}}$. We note that Theorem 4.8 is consistent with Theorem 4.5 since $C_\Sigma = 1$ and $h = h_r$ for the transition homogeneous setting. As $h > h_r$ and $C_\Sigma > 1$ when the transition kernels are different in different tasks, the discrepancy $C_\Sigma h > h_r$, which indicates that more target samples are required to attain the desired rate $\sqrt{p\log^2 n_{\mathcal{K}}/n_{\mathcal{K}}}$ in the transition heterogeneous setting than the homogeneous one.

In this study, we do not specify a structure for the transition probability $P^{(k)}$ nor delve into its transfer learning aspects. However, recent research (Lu et al., 2021; Cheng et al., 2022; Agarwal et al., 2023) highlights the benefits of leveraging $P^{(k)}$'s shared low-rank structure to enhance the estimation of the target task's Q^* functions. A promising avenue for future investigation is integrating $P^{(k)}$'s low-rank structure into Transferred FQI, particularly in settings with heterogeneous transitions.

5 Empirical Studies

5.1 Simulations

In this section, we use simulation studies to show the advantage of our proposed Algorithm 1 in RL. We choose the state space $\mathcal{X}$ to be $[-1, 1]^3$, set the action space $\mathcal{A}$ to be $\{-1, +1\}$, and generate each trajectory by the following equation:

$$\boldsymbol{x}_{t+1} = \frac{3}{4}\begin{pmatrix} a_t & & \\ & -a_t & \\ & & a_t \end{pmatrix}\boldsymbol{x}_t + \epsilon_{x,t},$$

where $\boldsymbol{x}_0 \sim \mathcal{N}(0, I_3)$, $a_t \sim$ Bernoulli(0.5), and $\epsilon_{x,t} \sim \mathcal{N}(0, I_3/4)$. For the convenience of constructing sieve basis, we use a tanh function to map each state vector $\boldsymbol{x}_t$ onto $[-1, 1]$. This setting is partly adapted from the two-dimensional setting in Shi et al. (2022). We consider a quadratic reward function:

$$r_t = (A_t) * (\boldsymbol{x}_t^\top \boldsymbol{C} \boldsymbol{x}_t) + \epsilon_{r,t},$$

where $\epsilon_{r,t} \sim \mathcal{N}(0, I_3/4)$ and $\boldsymbol{C}$ is a random matrix to be specified.

Based on the settings above, we generate a target task and a source task, with index $k = 0$ and 1, respectively. The number of trajectories in task k is denoted by $I^{(k)}$, and we fix the length of each trajectory to be 5. In practice, we choose $I^{(0)} = 20$ and let $I^{(1)}$ range from 10 to 80. For the target task, the diagonal entries of the matrix $\boldsymbol{C}^{(0)}$ are independently drawn from $\mathcal{N}(0, 1)$, while the off-diagonal entries are independently drawn from $\mathcal{N}(0, 1/4)$. For the source task, we set $\boldsymbol{C}^{(1)} = \boldsymbol{C}^{(0)} + \boldsymbol{C}_\delta$, where each entry of $\boldsymbol{C}_\delta$ is independently drawn from $\mathcal{N}\big(0, \sigma_C^2\big)$. The parameter σ_C controls the level of discrepancy between the target task and the source task, whose possible values are set as 0.25, 0.5, 0.75, and 1.0 in our experiments.

To verify the effect of our proposed two-step algorithm, we compare the following three methods:

(1) "no-transfer": Fitted Q-Iteration on the target task without transferring any information from the source task.

(2) "one-step": Fitted Q-Iteration on the aggregated data, without correcting the difference (i.e., only do Step I in Algorithm 1 for each iteration).

(3) "two-step": our proposed method (Algorithm 1).

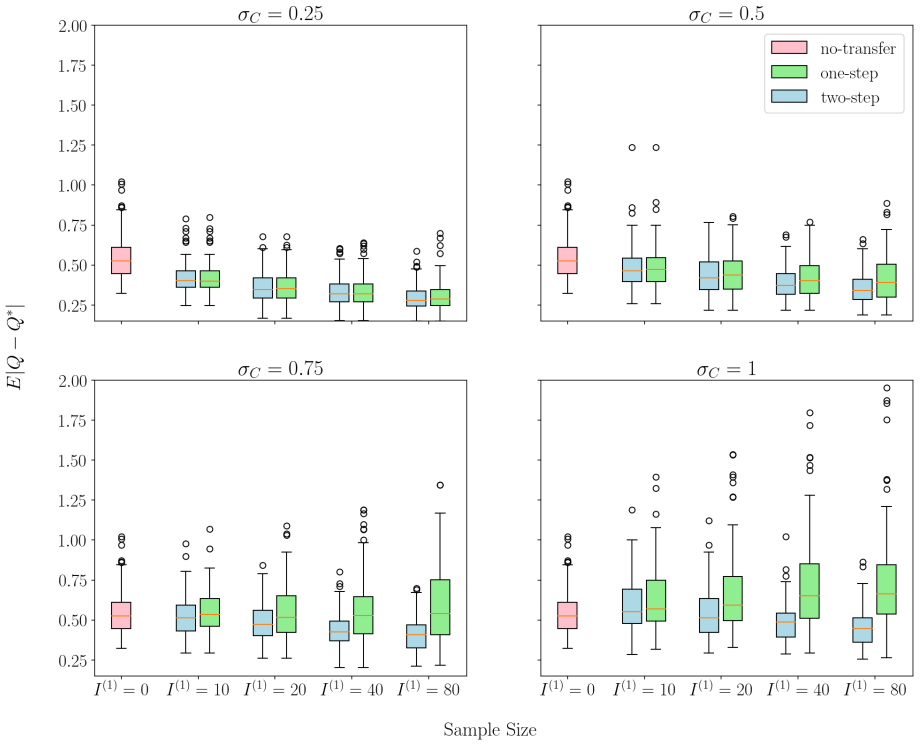

Figure 1: Boxplots of the estimation errors for the Q^* function with $I_0 = 20$ and different source sample sizes I_1. The parameter σ_C on the top of each subfigure indicates the standard deviation of the difference matrix $\boldsymbol{C}_\delta$.

The basis function $\boldsymbol{\phi}(\boldsymbol{x})$ is a set of three-dimensional B-splines on $[-1, 1]^3$ for all methods. For each k, each entry of the initial estimator $\widehat{\boldsymbol{\beta}}_0^{(k)}$ is independently drawn from $\mathcal{N}(0, 0.01)$, and the regularization parameters $\lambda_\delta^{(k)}$ are determined through cross validation. To evaluate the performance of each method, we estimate the expected estimation error $\mathbb{E}\left|\widehat{Q}(\boldsymbol{x}, a) - Q^*(\boldsymbol{x}, a)\right|$ by taking an average over 200 values of $(\boldsymbol{x}, a)$ that are independently drawn from the true distribution of $(\boldsymbol{x}_0, a_0)$. Since it is hard to compute the Q^* function in closed form, we also use Monte Carlo approximations to obtain the true value of Q^* for any fixed $(\boldsymbol{x}, a)$. The results over 100 independent runs are displayed by boxplots in Figure 1.

As shown in Figure 1, when the task discrepancy is small (i.e., $\sigma_C = 0.25$), the one-step and the two-step methods both significantly outperform the "no-transfer" FQI method, and their performance improves as the source sample size $I^{(1)}$ increases. However, the

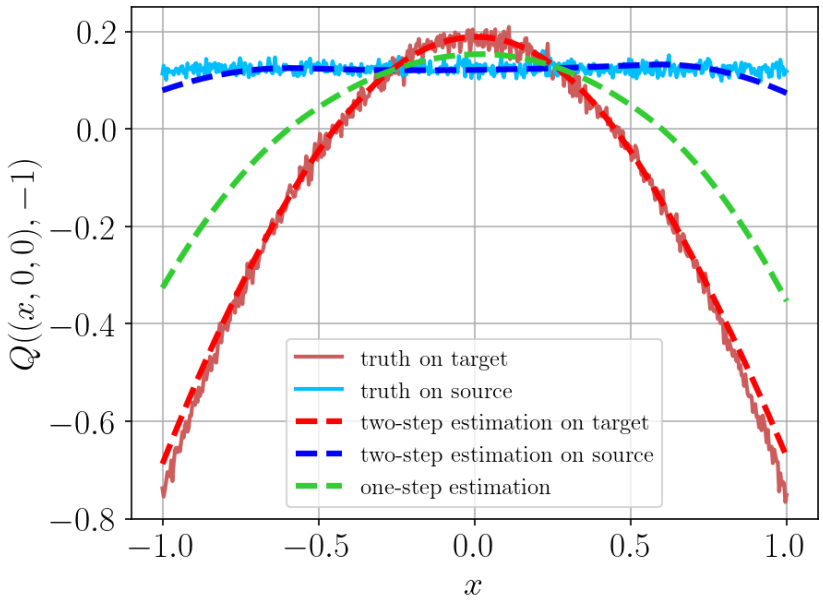

Figure 2: The estimated functions of the one-step method and the two-step method for the optimal action-value function $Q^*((x,0,0),-1)$. The legend "truth" represents the true Q^* function, approximated by Monte Carlo simulation, and "two-step estimation" represents the estimated Q^* function obtained by the two-step method. For clear presentation, the estimated Q^* function of the "no-transfer" method is omitted since its scale is too large. The matrices in the reward function are fixed to be $\boldsymbol{C}^{(0)} = \mathrm{diag}(1,0,0)$ and $\boldsymbol{C}^{(1)} = \mathrm{diag}(0,0,0)$.

one-step method's performance deteriorates as the task discrepancy σ_C grows larger. Particularly, when $\sigma_C \geq 0.75$, it performs even worse than the "no-transfer" method due to the significant bias induced by the source task. On the contrary, the proposed two-step method consistently performs well with sufficiently large source sample size, highlighting the importance of the second step in our proposed algorithm, which corrects the bias of the center estimator obtained in the first step.

To illustrate the learning outcomes of various methods, we juxtapose the estimated Q^* functions' marginal versions against the true Q^* function for $\boldsymbol{x} = (x,0,0)$ with x ranging between $[-1,1]$ and $a = -1$. For the target task, we set $\boldsymbol{C}^{(0)} = \mathrm{diag}(1,0,0)$ and $I^{(0)} = 40$, while for the source task, $\boldsymbol{C}^{(1)} = \mathrm{diag}(0,0,0)$ and $I^{(1)} = 40$. The comparison, as shown in Figure 2, reveals that Transferred FQI accurately estimates the true Q^* values for both tasks, in contrast to the one-step method, which yields less accurate intermediate estimates

far from either task.

5.2 Real Data Analysis

In this section, we apply the proposed Algorithm 1 on Medical Information Mart for Intensive Care (MIMIC-III) dataset (Johnson et al., 2016) to illustrate the benefit of knowledge transfer in Q^* learning.

MIMIC-III is a large, publicly available database of medical records containing de-identified health-related data about patients admitted to critical care units at a large tertiary care hospital. In particular, we follow the procedure of Komorowski et al. (2018) to select the data of the adult sepsis patients and extract a set of 43 features for characterizing each patient, including demographics, Elixhauser premorbid status, vital signs, laboratory values, fluids and vasopressors received. To save the computation time, we further compute the 20 top principal components of the features, which explains 99.9% of the total variance, as state variables, i.e., $\boldsymbol{X}_{i,t} \in \mathbb{R}^{20}$. The two action variables of interest are the total volume of intravenous (IV) fluids and the maximum dose of vasopressors administrated over each period. Each action variable is discretized into three levels (low, medium, and high); hence, there are nine possible action combinations. For the rewards $R_{i,t}$, we follow Prasad et al. (2017) and Komorowski et al. (2018) to assign rewards to each state based on the health measurement and mortality of the patient. A higher reward $R_{i,t}$ indicates a better physical condition of the patient i after the action $A_{i,t}$ taken at time t.

The final processed dataset contains 5,923 observations from 400 patients, including 1,391 observations for female patients and 4,136 observations for male patients. We consider the female group as the target task ($k = 0$) and the male group as the source task ($k = 1$) and compare the performance of our proposed transferred FQI (Algorithm 1) with that of FQI on the target set without transferring. We set the discount parameter $\gamma = 0.6$,

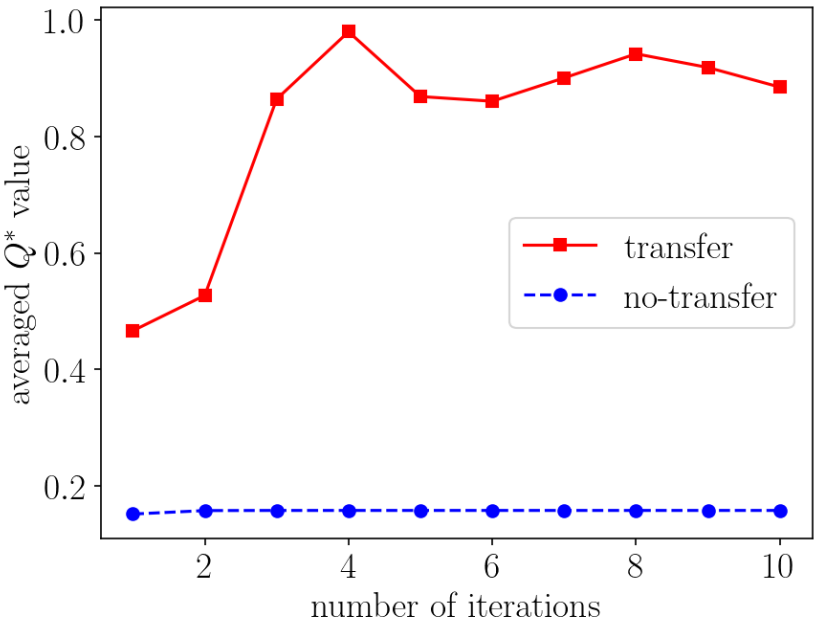

Figure 3: Averaged Q^* values estimated by the transferred FQI (Algorithm 1) and FQI on the target data without transferring.

construct 20-dimensional B-splines as basis functions $\boldsymbol{\phi}(\boldsymbol{x})$, and determine the penalization parameter $\lambda_\delta^{(k)}$ by cross-validation. Figure 3 presents the estimated Q^* function value averaged over all observations in the target dataset as the iterations proceed. The optimal value obtained by our proposed algorithm is clearly larger than simple FQI without transferring, especially when the number of iterations is large enough, which verifies the advantage of knowledge transfer in FQI.

6 Conclusion

This paper introduces a novel knowledge transfer framework for sequential decision-making problems formulated as RL tasks. We first formally define the notion of task discrepancy and establish a theoretical relationship between task discrepancy and the discrepancy in optimal action-value functions, based on which we proposed the Transferred Fitted Q-Iteration (Transferred FQI) algorithm, an iterative procedure that jointly learns the com-

monalities and corrects for biases across tasks within a general function approximation framework. The Transferred FQI algorithm is further instantiated with sieve function approximation, where our theoretical analysis reveals several important insights into the practice of transfer learning for offline RL. Concretely, we quantify the required target sample size to overcome task discrepancy bias and benefit from transfer.

Overall, this work opens up new frontiers in transfer learning for RL by providing a rigorous iterative algorithm, theoretical analysis, and insights into the interplay between source knowledge, task discrepancy, and target sample size requirements. The framework complements prior RL and TL literature while introducing novel perspectives on iterative knowledge transfer during batch Q^* estimation. Potential future work can build upon these foundations to develop new transfer approaches under other function approximation methods, such as neural networks.

References

Agarwal, A., Y. Song, W. Sun, K. Wang, M. Wang, and X. Zhang (2023). Provable benefits of representational transfer in reinforcement learning. In *Proceedings of Thirty Sixth Conference on Learning Theory*, Volume 195.

Barreto, A., W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. P. van Hasselt, and D. Silver (2017). Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, Volume 30.

Cai, T. T. and H. Pu (2024). Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*.

Cai, T. T. and H. Wei (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics 49*(1), 100–128.

Chen, E. Y., M. I. Jordan, and S. Li (2022). Transferred Q-learning. *arXiv preprint arXiv:2202.04709*.

Chen, J. and N. Jiang (2019). Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Cheng, Y., S. Feng, J. Yang, H. Zhang, and Y. Liang (2022). Provable benefit of multitask representation learning in reinforcement learning. In *Advances in Neural Information Processing Systems*, Volume 35.

Duan, Y. and K. Wang (2023). Adaptive and robust multi-task learning. *The Annals of Statistics 51*(5), 2015–2039.

Ernst, D., P. Geurts, and L. Wehenkel (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research 6*(18), 503–556.

Fan, J., Z. Wang, Y. Xie, and Z. Yang (2020). A theoretical analysis of deep Q-learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, Volume 120.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics 26*(1), 242–272.

Johnson, A. E., T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data 3*(1), 1–9.

Komorowski, M., L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine 24*(11), 1716–1720.

Li, G., L. Shi, Y. Chen, Y. Chi, and Y. Wei (2024). Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics 52*(1), 233–260.

Li, S., T. T. Cai, and H. Li (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(1), 149–173.

Li, S., T. T. Cai, and H. Li (2023). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association 118*(543), 2171–2183.

Li, S., L. Zhang, T. T. Cai, and H. Li (2023). Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 1–12.

Liu, X. (2023). Dynamic coupon targeting using batch deep reinforcement learning: An application to livestream shopping. *Marketing Science 42*(4), 637–658.

Lu, R., G. Huang, and S. S. Du (2021). On the power of multitask representation learning in linear MDP. *arXiv preprint arXiv:2106.08053*.

Ma, Y., Y.-X. Wang, and B. Narayanaswamy (2019). Imitation-regularized offline learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, Volume 89.

Maity, S., Y. Sun, and M. Banerjee (2022). Minimax optimal approaches to the label shift problem in non-parametric settings. *The Journal of Machine Learning Research 23*(1), 15698–15742.

Munos, R. and C. Szepesvári (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research 9*(27), 815–857.

Murphy, S. A. (2005). A generalization error for Q-learning. *Journal of Machine Learning Research 6*(37), 1073–1097.

Prasad, N., L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt (2017). A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*.

Rafferty, A. N., E. Brunskill, T. L. Griffiths, and P. Shafto (2016). Faster teaching via POMDP planning. *Cognitive Science 40*(6), 1290–1332.

Riedmiller, M. (2005). Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In *Machine Learning: ECML 2005*, Volume 3720.

Shi, C., S. Zhang, W. Lu, and R. Song (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(3), 765–793.

Shi, L., G. Li, Y. Wei, Y. Chen, and Y. Chi (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *Proceedings of the 39th International Conference on Machine Learning*, Volume 162.

Silver, D., S. Singh, D. Precup, and R. S. Sutton (2021). Reward is enough. *Artificial Intelligence 299*, 103535.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Tian, Y. and Y. Feng (2023). Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association 118*(544), 2684–2697.

Vecerik, M., T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. Riedmiller (2017). Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Wang, C., C. Wang, X. He, and X. Feng (2023). Minimax optimal transfer learning for kernel-based nonparametric regression. *arXiv preprint arXiv:2310.13966*.

Wang, K. (2023). Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*.

Xie, T. and N. Jiang (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, Volume 124.

Xie, T. and N. Jiang (2021). Batch value-function approximation with only realizability. In *Proceedings of the 38th International Conference on Machine Learning*, Volume 139.

Yan, Y., G. Li, Y. Chen, and J. Fan (2022). Model-based reinforcement learning is minimax-optimal for offline zero-sum Markov games. *arXiv preprint arXiv:2206.04044*.

Yang, T., S. Cen, Y. Wei, Y. Chen, and Y. Chi (2023). Federated natural policy gradient methods for multi-task reinforcement learning. *arXiv preprint arXiv:2311.00201*.

Yin, H. and S. Pan (2017). Knowledge transfer for deep reinforcement learning with hierarchical experience replay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 31.

Zhang, Z., W. Zhan, Y. Chen, S. S. Du, and J. D. Lee (2023). Optimal multi-distribution learning. *arXiv preprint arXiv:2312.05134*.

Zhu, Z., K. Lin, A. K. Jain, and J. Zhou (2023). Transfer learning in deep reinforcement learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*(11), 13344–13362.

SUPPLEMENTARY MATERIAL of

"Data-Driven Knowledge Transfer in Batch Q^* Learning"

Appendix A Notations

Let lowercase letter x, boldface letter $\boldsymbol{x}$, capital letter X, boldface capital letter $\boldsymbol{X}$, and calligraphic letter $\mathcal{X}$ represent scalar, vector, random variable, random vector (or matrix), and set (or space), respectively. We use the notation $[N]$ to refer to the positive integer set $\{1, \ldots, N\}$ for $N \in \mathbb{Z}_+$. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some constant C that does not depend on n, and write $a \asymp b$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Moreover, for random variable sequences $\{X_n\}$ and $\{Y_n\}$, we let $X_n = \mathcal{O}_p(Y_n)$ if X_n/Y_n is bounded in probability. We let C, c, C_0, c_0, ... denote generic constants, where the uppercase and lowercase letters represent large and small constants, respectively. The actual values of these generic constants may vary from time to time.

As a convention, in this paper, we use upper case letters, such as I and T, as the sizes of dimensions (fixed or growing). We use the corresponding lowercase letters, e.g., i and t, as the running indices.

We use $\lambda_{\min}(\boldsymbol{X})$ and $\lambda_{\max}(\boldsymbol{X})$ to denote its smallest and largest eigenvalues of matrix $\boldsymbol{X}$. We use the following matrix norms: maximum norm $\|\boldsymbol{X}\|_{\max} \triangleq \max_{ij} |x_{ij}|$, ℓ_1-norm $\|\boldsymbol{X}\|_1 \triangleq \max_j \sum_i |x_{ij}|$, ℓ_∞-norm $\|\boldsymbol{X}\|_\infty \triangleq \max_i \sum_j |x_{ij}|$, and ℓ_2-norm $\|\boldsymbol{X}\|_2 \triangleq \lambda_{\max}(\boldsymbol{X})$.

Appendix B Technical Proof for Theoretical Results

Proof for Lemma 2.1. By the Bellman optimality equation, we have

$$Q^{*(k)}(\boldsymbol{x}, a) = r^{(k)}(\boldsymbol{x}, a) + \gamma \int_{\mathcal{X}} \left[\max_{a' \in \mathcal{A}} Q^{*(k)}(\boldsymbol{x}', a') \right] \rho^{(k)}(\boldsymbol{x}' \mid \boldsymbol{x}, a) \, \mathrm{d}\boldsymbol{x}',$$

which implies that

$$\begin{aligned}\delta_Q^{(k)}(\boldsymbol{x},a) &= \delta_r^{(k)}(\boldsymbol{x},a) + \gamma\int_{\mathcal{X}}\left[\max_{a'\in\mathcal{A}} Q^{*(k)}(\boldsymbol{x}',a')\right]\delta_\rho^{(k)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\,\mathrm{d}\boldsymbol{x}' \\ &\quad + \gamma\int_{\mathcal{X}}\left[\max_{a'\in\mathcal{A}} Q^{*(k)}(\boldsymbol{x}',a') - \max_{a'\in\mathcal{A}} Q^{*(0)}(\boldsymbol{x}',a')\right]\rho^{(0)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\,\mathrm{d}\boldsymbol{x}' \\ &\le \delta_r^{(k)}(\boldsymbol{x},a) + \frac{\gamma R_{\max}}{1-\gamma}\int_{\mathcal{X}}\delta_\rho^{(k)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\,\mathrm{d}\boldsymbol{x}' \\ &\quad + \gamma\int_{\mathcal{X}}\max_{a'\in\mathcal{A}}\delta_Q^{(k)}(\boldsymbol{x}',a')\rho^{(0)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\,\mathrm{d}\boldsymbol{x}',\end{aligned}$$

where we use

$$\sup_{\boldsymbol{x},a} Q^{*(k)}(\boldsymbol{x},a) = \sup_{\boldsymbol{x},a,\pi}\mathbb{E}_{\mathcal{H}\sim\mathcal{P}_\pi(\mathcal{H})}\left[\sum_{t=0}^{T}\gamma^t R_{i,t}\,\middle|\,\boldsymbol{X}_{i,0}=\boldsymbol{x}, A_{i,0}=a\right] \le \sum_{t=0}^{T}\gamma^t R_{\max} = \frac{R_{\max}}{1-\gamma}. \tag{28}$$

Taking supreme over $\boldsymbol{x}\in\mathcal{X}$ and $a\in\mathcal{A}$ on both sides leads to

$$\sup_{\boldsymbol{x},a}\delta_Q^{(k)}(\boldsymbol{x},a) \le \sup_{\boldsymbol{x},a}\delta_r^{(k)}(\boldsymbol{x},a) + \frac{\gamma R_{\max}}{1-\gamma}\int_{\mathcal{X}}\sup_{\boldsymbol{x},a}\delta_\rho^{(k)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\,\mathrm{d}\boldsymbol{x}' + \gamma\sup_{\boldsymbol{x},a}\delta_Q^{(k)}(\boldsymbol{x},a),$$

hence

$$\sup_{\boldsymbol{x},a}\delta_Q^{(k)}(\boldsymbol{x},a) \le \frac{1}{1-\gamma}\sup_{\boldsymbol{x},a}\delta_r^{(k)}(\boldsymbol{x},a) + \frac{\gamma R_{\max}}{(1-\gamma)^2}\int_{\mathcal{X}}\sup_{\boldsymbol{x},a}\delta_\rho^{(k)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\,\mathrm{d}\boldsymbol{x}'.$$

Proof for Lemma 4.4. Note that Assumption 4.3 assumes that the constants κ, c, and $R_{\max}$ are uniform for all k, then the results directly follow from Lemma 1 in Shi et al. (2022) and Equation (28).

Proof of Theorems 4.5 and 4.8. We first restate the assumptions on the distribution of the data in task k, i.e., $\mathcal{S}^{(k)} = \left\{\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}, R_{i,t}^{(k)}, \boldsymbol{X}_{i,t+1}^{(k)}\right)\right\}$ for $i\in[I^{(k)}]$, $t\in\{0\}\cup[T]$.

The initial state $\boldsymbol{X}_{1,0}^{(k)}$ has a distribution $\nu^{(k)}$. For any $0\le t\le T$, the probability distribution of the action $A_{1,t}^{(k)}$ conditional on $\boldsymbol{X}_{1,t}^{(k)}=\boldsymbol{x}$ is given by an underlying behavior policy $b^{(k)}(\cdot\,|\,\boldsymbol{x})$. Given $\boldsymbol{X}_{1,t}^{(k)}=\boldsymbol{x}, A_{1,t}^{(k)}=a$, the distribution of the next state $\boldsymbol{X}_{1,t+1}^{(k)}$ is determined by a transition kernel $P^{(k)}\big(\,\cdot\,|\,\boldsymbol{x},a\big)$. Suppose that the density $\rho^{(k)}(\boldsymbol{x}'\,|\,\boldsymbol{x},a) =$

$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}'}P^{(k)}\big(\boldsymbol{x}' \,|\, \boldsymbol{x}, a\big)$ exists and is continuous for almost every $(\boldsymbol{x}, a)$. Also, a reward $R_{i,t}^{(k)} = r^{(k)}\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right) + \eta_{i,t}^{(k)}$ is observed for any t. The tuple set $\left\{\left(\boldsymbol{X}_{1,t}^{(k)}, A_{1,t}^{(k)}, R_{1,t}^{(k)}, \boldsymbol{X}_{1,t+1}^{(k)}\right)\right\}_{t=0}^{T-1}$ forms a trajectory, and there are multiple i.i.d. trajectories $\left\{\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}, R_{i,t}^{(k)}, \boldsymbol{X}_{i,t+1}^{(k)}\right)\right\}$ in one task. We further split $\mathcal{S}^{(k)}$ into τ subsets, $\left\{\mathcal{S}_\tau^{(k)}\right\}_{\tau\in[\Upsilon]}$, each with sample size n_k, and use subset $\mathcal{S}_\tau^{(k)}$ in iteration τ, which ensures that the data used in different iterations are independent. Assumption 4.6 ensures that, for any i, $\left\{\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right)\right\}_{t=0}^{T-1}$ is a Markov chain on $\mathcal{X}\times\mathcal{A}$ with the transition kernel $P^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)b^{(k)}(a' \,|\, \boldsymbol{x}')$, and $\left\{\left(\boldsymbol{X}_{i,t}^{(k)}\right)\right\}_{t=0}^{T-1}$ is also a Markov chain with the transition kernel $\sum_{a\in\mathcal{A}} b^{(k)}(a \,|\, \boldsymbol{x})P^{(k)}(\boldsymbol{x}' \,|\, \boldsymbol{x}, a)$. The latter chain has an invariant distribution $\mu^{(k)}(\boldsymbol{x})$, which implies that the former chain has an invariant distribution $\mu^{(k)}(\boldsymbol{x})b^{(k)}(a \,|\, \boldsymbol{x})$.

Define the Bellman optimality operator $\mathcal{T}^{(k)}$ on the k-th task by

$$\mathcal{T}^{(k)}Q\left(\boldsymbol{x}, a\right) = r^{(k)}\left(\boldsymbol{x}, a\right) + \gamma\cdot\mathbb{E}\left[\max_{a'\in\mathcal{A}} Q\left(\boldsymbol{x}', a'\right) \,\middle|\, \boldsymbol{x}' \sim P^{(k)}\left(\cdot \,|\, \boldsymbol{x}, a\right)\right]. \tag{29}$$

For a function $f(\boldsymbol{x}, a)$ defined on $\mathcal{S}\times\mathcal{A}$ and a policy function $\omega(\cdot \mid \boldsymbol{x})$, define the operator P^ω by

$$P^\omega f(\boldsymbol{x}, a) := \mathbb{E}\left[f(\boldsymbol{x}', a') \,|\, \boldsymbol{x}' \sim P^{(0)}(\cdot \,|\, \boldsymbol{x}, a), a' \sim \omega(\cdot \,|\, \boldsymbol{x}')\right].$$

Let Q^* be the optimal action-value function on the target dataset. The iterative algorithm gives a series of estimators $\widehat{Q}_\tau^{(0)}$ for Q^* and finally an estimated optimal policy π_Υ. By the proof of Theorem 6.1 in Fan et al. (2020), it holds that

$$\begin{aligned}
&\mathbb{E}_{\mu^{(0)},b^{(0)}}\left|Q^* - Q^{\pi_\Upsilon}\right| \\
&\lesssim \sum_{\tau=0}^{\Upsilon-1}\sum_{j=0}^{\infty}\gamma^{\Upsilon-\tau+j}\sup_{\omega_1,\ldots,\omega_{\Upsilon-\tau+j}}\mathbb{E}_{\mu^{(0)},b^{(0)}}\left[\left(P^{\omega_{\Upsilon-\tau+j}}P^{\omega_{\Upsilon-\tau+j-1}}\cdots P^{\omega_1}\right)\left|\mathcal{T}^{(0)}\widehat{Q}_\tau^{(0)} - \widehat{Q}_{\tau+1}^{(0)}\right|\right] \\
&\quad + \frac{\gamma^{\Upsilon+1}}{(1-\gamma)^2}R_{\max},
\end{aligned} \tag{30}$$

where the supreme is taken over all possible policies $\omega_1, \ldots, \omega_{\Upsilon-\tau+j}$, and $\mathbb{E}_{\mu^{(0)},b^{(0)}}$ denotes the expectation with respect to the invariant distribution $\mu^{(0)}(\boldsymbol{x})b^{(0)}(a \,|\, \boldsymbol{x})$. In the following proof, we focus on bounding $\left|\mathcal{T}^{(0)}\widehat{Q}_{\tau-1}^{(0)} - \widehat{Q}_\tau^{(0)}\right|(\boldsymbol{x}, a)$ for a fixed $\tau\in[\Upsilon]$, where $\widehat{Q}_\tau^{(k)}(\boldsymbol{x}, a) =$

$\boldsymbol{\xi}(\boldsymbol{x}, a)^\top \widehat{\boldsymbol{\beta}}_\tau^{(k)}$ in our algorithm.

Let $V_{\max} = R_{\max}/(1-\gamma)$, which is an upper bound for the optimal Q^* function since

$$Q^*(\boldsymbol{x}, a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[r^{(0)}(\boldsymbol{X}_{i,t}, A_{i,t}) \mid \boldsymbol{X}_{i,0} = x, A_{i,t} = a] \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = V_{\max}.$$

We first assume that $\left\|\widehat{Q}_{\tau-1}^{(0)}\right\|_\infty \leq V_{\max}$ and $\sup_k \left\|\widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)}\right\|_1 \leq h$, which are satisfied for $\tau = 1$ by assumption.

Recall that in Algorithm 1, for each k, we compute

$$Y_{i,t}^{(k),\tau} = R_{i,t}^{(k)} + \gamma \cdot \max_{a \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}\left(\boldsymbol{X}_{i,t+1}^{(k)}, a\right).$$

By definition, we have $\mathbb{E}\left[Y_{i,t}^{(k),\tau} \mid \boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right] = \left(\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}\right)\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right)$ for any $\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right) \in \mathcal{X} \times \mathcal{A}$. Thus, $\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}$ can be viewed as the underlying truth of the regression problem defined in (14), where the covariates and responses are $\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right)$ and $Y_{i,t}^{(k),\tau}$, respectively. Therefore, we rewrite

$$Y_{i,t}^{(k),\tau} = \mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right) + e_{i,t}^{(k),\tau}, \tag{31}$$

where $\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)} : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ is an unknown regression function to be estimated. Recall that $R_{i,t}^{(k)} = r^{(k)}\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right) + \eta_{i,t}^{(k)}$ and

$$\left(\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}\right)\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right) = r^{(k)}\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right) + \gamma \mathbb{E}\left[\max_{a \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}\left(\boldsymbol{x}', a\right) \middle| \boldsymbol{x}' \sim P^{(k)}\left(\cdot \mid \boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right)\right].$$

Hence, the regression noise term $e_{i,t}^{(k),\tau}$ is zero-mean sub-Gaussian, since

$$e_{i,t}^{(k),\tau} = \eta_{i,t}^{(k)} + \gamma\left(1 - \mathbb{E}\right)\left[\max_{a \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}\left(\boldsymbol{x}', a\right) \middle| \boldsymbol{x}' \sim P^{(k)}\left(\cdot \mid \boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)}\right)\right],$$

where $\eta_{i,t}^{(k)}$ is zero-mean sub-Gaussian, and the second term is bounded.

In each step τ, we use $\widehat{Q}_\tau^{(k)}(\boldsymbol{x}, a) = \boldsymbol{\xi}^\top(\boldsymbol{x}, a)\widehat{\boldsymbol{\beta}}_\tau^{(k)}$ to estimate $\left(\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}\right)(\boldsymbol{x}, a)$, where

$$\boldsymbol{\xi}(\boldsymbol{x}, a) := \left[\boldsymbol{\phi}^\top(\boldsymbol{x})\mathbb{I}(a=1), \boldsymbol{\phi}^\top(\boldsymbol{x})\mathbb{I}(a=2), \ldots, \boldsymbol{\phi}^\top(\boldsymbol{x})\mathbb{I}(a=m)\right]^\top,$$

and $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_p(\cdot))^\top$ is a set of sieve basis functions. By definition,

$$\left(\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}\right)(\boldsymbol{x}, a) = r^{(k)}(\boldsymbol{x}, a) + \gamma \int_{\mathcal{X}} \left[\max_{a' \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}\left(\boldsymbol{x}', a'\right)\right] \rho^{(k)}\left(\boldsymbol{x}' \mid \boldsymbol{x}, a\right) \mathrm{d}\boldsymbol{x}'.$$

The estimation is guaranteed by the following property: For any function $f(\boldsymbol{x}, a)$ that satisfies $f(\cdot, a) \in \Lambda(\kappa, c)$ for all $a \in \mathcal{A}$, there exists a set of vectors $\{\boldsymbol{\beta}_a \in \mathbb{R}^p\}_{a\in\mathcal{A}}$ that

$$\sup_{\boldsymbol{x}\in\mathcal{X}, a\in\mathcal{A}} \left|f(\boldsymbol{x}, a) - \boldsymbol{\beta}_a^\top \boldsymbol{\phi}(\boldsymbol{x})\right| \le Cp^{-\kappa/d},$$

for some positive constant C. Let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top\right)^\top$, and then we have

$$\sup_{\boldsymbol{x}\in\mathcal{X}, a\in\mathcal{A}} \left|f(\boldsymbol{x}, a) - \boldsymbol{\beta}^\top \boldsymbol{\xi}(\boldsymbol{x}, a)\right| \le Cp^{-\kappa/d}.$$

By Assumption 4.7 and the definition of h, there exist $\left\{\boldsymbol{\beta}_r^{(k)}\right\}_{k\in\mathcal{K}}$, $\left\{\boldsymbol{\beta}_\rho^{(k)}(\boldsymbol{x}')\right\}_{\boldsymbol{x}'\in\mathcal{X}, k\in\mathcal{K}}$ such that

$$\sup_{\boldsymbol{x}, a, k} \left|r^{(k)}(\boldsymbol{x}, a) - \boldsymbol{\xi}^\top(\boldsymbol{x}, a)\boldsymbol{\beta}_r^{(k)}\right| \le Cp^{-\kappa/d},$$
$$\sup_{\boldsymbol{x}, a, \boldsymbol{x}', k} \left|\rho^{(k)}(\boldsymbol{x}' \mid \boldsymbol{x}, a) - \boldsymbol{\xi}^\top(\boldsymbol{x}, a)\boldsymbol{\beta}_\rho^{(k)}(\boldsymbol{x}')\right| \le Cp^{-\kappa/d},$$

and

$$\int_{\mathcal{X}} \left\|\boldsymbol{\beta}_\rho^{(k)}(\boldsymbol{x}') - \boldsymbol{\beta}_\rho^{(0)}(\boldsymbol{x}')\right\|_1 \mathrm{d}\boldsymbol{x}' \le h, \quad \left\|\boldsymbol{\beta}_r^{(k)} - \boldsymbol{\beta}_r^{(0)}\right\|_1 \le h,$$

Define

$$\boldsymbol{\beta}_\tau^{(k)} := \boldsymbol{\beta}_r^{(k)} + \gamma \int \left[\max_{a'\in\mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}(\boldsymbol{x}', a')\right] \boldsymbol{\beta}_\rho^{(k)}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'.$$

Since $\max_{a'\in\mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}(\boldsymbol{x}', a') \lesssim V_{\max}$, it follows that

$$\sup_{\boldsymbol{x}, a, k} \left|\left(\mathcal{T}^{(k)}\widehat{Q}_{\tau-1}^{(k)}\right)(\boldsymbol{x}, a) - \boldsymbol{\xi}^\top(\boldsymbol{x}, a)\boldsymbol{\beta}_\tau^{(k)}\right| \le C'p^{-\kappa/d}, \tag{32}$$

for some absolute constant C'.

Moreover, note that

$$\begin{aligned}\boldsymbol{\beta}_\tau^{(k)} - \boldsymbol{\beta}_\tau^{(0)} = \boldsymbol{\beta}_r^{(k)} - \boldsymbol{\beta}_r^{(0)} &+ \gamma \int \left[\max_{a'\in\mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}(\boldsymbol{x}', a')\right] \left(\boldsymbol{\beta}_\rho^{(k)}(\boldsymbol{x}') - \boldsymbol{\beta}_\rho^{(0)}(\boldsymbol{x}')\right) \mathrm{d}\boldsymbol{x}' \\ &+ \gamma \int \left[\max_{a'\in\mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}(\boldsymbol{x}', a') - \max_{a'\in\mathcal{A}} \widehat{Q}_{\tau-1}^{(0)}(\boldsymbol{x}', a')\right] \boldsymbol{\beta}_\rho^{(0)}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'.\end{aligned} \tag{33}$$

The ℓ_1-norm of the first two terms on the RHS of (33) can be bounded by a constant times h. For the third term, we first have

$$
\begin{aligned}
&\int_{\mathcal{X}} \left| \max_{a' \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(k)}(\boldsymbol{x}', a') - \max_{a' \in \mathcal{A}} \widehat{Q}_{\tau-1}^{(0)}(\boldsymbol{x}', a') \right| \mathrm{d}\boldsymbol{x}' \\
&\leq \int_{\mathcal{X}} \left| \max_{a'} \boldsymbol{\xi}^{\top}(\boldsymbol{x}', a') \left(\widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)} \right) \right| \mathrm{d}\boldsymbol{x}' \\
&\leq \sqrt{\left(\widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)} \right)^{\top} \int_{\mathcal{X}} \left[\boldsymbol{\Phi}(\boldsymbol{x}') \boldsymbol{\Phi}^{\top}(\boldsymbol{x}') \right] \mathrm{d}\boldsymbol{x}' \left(\widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)} \right)} \\
&\lesssim \left\| \widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)} \right\|_2,
\end{aligned}
$$

where $\boldsymbol{\Phi}(\boldsymbol{x}) := \left[\boldsymbol{\phi}^{\top}(\boldsymbol{x}), \boldsymbol{\phi}^{\top}(\boldsymbol{x}), \ldots, \boldsymbol{\phi}^{\top}(\boldsymbol{x}) \right]^{\top}$ is the concatenation of $|\mathcal{A}|$ copies of the basis functions $\boldsymbol{\phi}(\boldsymbol{x})$. The last inequality follows from Lemma 2 in Shi et al. (2022), which shows that

$$
\lambda_{\max} \left[\int_{\mathcal{X}} \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^{\top}(\boldsymbol{x}) \mathrm{d}x \right] < \infty.
$$

Then we obtain that the ℓ_1-norm of the third term of (33) is upper bounded by a constant times $\gamma \left\| \widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)} \right\|_2$. Therefore,

$$
\left\| \boldsymbol{\beta}_{\tau}^{(k)} - \boldsymbol{\beta}_{\tau}^{(0)} \right\|_1 \lesssim h + \gamma \left\| \widehat{\boldsymbol{\beta}}_{\tau-1}^{(k)} - \widehat{\boldsymbol{\beta}}_{\tau-1}^{(0)} \right\|_2 \lesssim h, \forall k.
$$

Without loss of generality, we can assume that $\boldsymbol{\beta}_{\tau}^{(k)} = \boldsymbol{\beta}_{\tau} + \boldsymbol{\delta}_{\tau}^{(k)}$ with $\left\| \boldsymbol{\delta}_{\tau}^{(k)} \right\|_1 \leq h$ for $k = 0, 1, \ldots, K$.

Till now, we have already established the formal definition and properties of the parameter of interest $\boldsymbol{\beta}_{\tau}^{(k)}$. In the sequel, we will provide an upper bound for the ℓ_2 error $\left\| \widehat{\boldsymbol{\beta}}_{\tau}^{(k)} - \boldsymbol{\beta}_{\tau}^{(k)} \right\|_2$, which is closely related to $\left| \mathcal{T}^{(0)} \widehat{Q}_{\tau-1}^{(0)} - \widehat{Q}_{\tau}^{(0)} \right|$. We first simplify some notations for a clear presentation.

Let $f_{\tau}^{(k)} = \mathcal{T}^{(k)} \widehat{Q}_{\tau-1}^{(k)}$. Define the k-th sample matrix $\boldsymbol{Z}_{\tau}^{(k)} \in \mathbb{R}^{n_k \times mp}$ such that the rows of $\boldsymbol{Z}^{(k)}$ are $\boldsymbol{\xi}^{\top}\left(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)} \right)$, for $(i,t) \in \mathcal{S}_{\tau}^{(k)}$. Likewise, define

$$
\boldsymbol{y}_{\tau}^{(k)} := \left[Y_{i_1,0}^{(k),\tau}, \ldots, Y_{i_1,T-1}^{(k),\tau}, Y_{i_2,0}^{(k),\tau}, \ldots, Y_{i_{n_k/T},T-1}^{(k),\tau} \right]^{\top}, (i_1, \ldots, i_{n_k/T} \in \mathcal{S}_{\tau}^{(k)})
$$

$$
\boldsymbol{f}_{\tau}^{(k)} := \left[f_{\tau}^{(k)}\left(\boldsymbol{X}_{i_1,0}^{(k)}, \boldsymbol{A}_{i_1,0}^{(k)} \right), \ldots, f_{\tau}^{(k)}\left(\boldsymbol{X}_{i_{n_k/T},T-1}^{(k)}, \boldsymbol{A}_{i_{n_k/T},T-1}^{(k)} \right) \right]^{\top},
$$

and $\boldsymbol{e}_{\tau}^{(k)} := \left[e_{i_1,0}^{(k),\tau}, \ldots, e_{i_{n_k/T},T-1}^{(k),\tau} \right]^{\top}$. Using these notations, we have

$$\boldsymbol{y}_\tau^{(k)} = \boldsymbol{Z}^{(k)}\boldsymbol{\beta}_\tau^{(k)} + \left[\boldsymbol{f}_\tau^{(k)} - \boldsymbol{Z}^{(k)}\boldsymbol{\beta}_\tau^{(k)}\right] + \boldsymbol{e}_\tau^{(k)},$$

where $\boldsymbol{e}_\tau^{(k)}$ is sub-Gaussian, and by (32),

$$\left\|\boldsymbol{f}_\tau^{(k)} - \boldsymbol{Z}^{(k)}\boldsymbol{\beta}_\tau^{(k)}\right\|_\infty \leq C' p^{-\kappa/d}.$$

Moreover, define $n_{\mathcal{K}} = \sum_{0\leq k\leq K} n_k$, $\widehat{\boldsymbol{\Sigma}}^{(k)} := \left(\boldsymbol{Z}^{(k)}\right)^\top \boldsymbol{Z}^{(k)}/n_k$, $\widehat{\boldsymbol{\Sigma}} := \sum_{0\leq k\leq K} \alpha_k \widehat{\boldsymbol{\Sigma}}^{(k)}$ with $\alpha_k = n_k/n_{\mathcal{K}}$, $\overline{\boldsymbol{\Sigma}}^{(k)} := \mathbb{E}\left[\widehat{\boldsymbol{\Sigma}}^{(k)}\right]$, $\overline{\boldsymbol{\Sigma}} := \mathbb{E}\left[\widehat{\boldsymbol{\Sigma}}\right]$, and

$$C_\Sigma := 1 + \max_k \left\|\overline{\boldsymbol{\Sigma}}^{-1}\left(\overline{\boldsymbol{\Sigma}}^{(k)} - \overline{\boldsymbol{\Sigma}}\right)\right\|_1.$$

By definition,

$$\overline{\boldsymbol{\Sigma}}^{(k)} = \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \boldsymbol{\xi}(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)})\boldsymbol{\xi}^\top(\boldsymbol{X}_{i,t}^{(k)}, A_{i,t}^{(k)})\right].$$

By Assumption 4.6, there exists a constant $c_\Sigma \geq 1$ such that $c_\Sigma^{-1} < \lambda_{\min}\left(\overline{\boldsymbol{\Sigma}}^{(k)}\right) < \lambda_{\max}\left(\overline{\boldsymbol{\Sigma}}^{(k)}\right) < c_\Sigma$ for all k and thus, $c_\Sigma^{-1} < \lambda_{\min}\left(\overline{\boldsymbol{\Sigma}}\right) < \lambda_{\max}\left(\overline{\boldsymbol{\Sigma}}\right) < c_\Sigma$. By the proof of Lemma 4 in Shi et al. (2022), there exists a constant $c'_\Sigma \geq 1$ such that $c_\Sigma'^{-1} < \lambda_{\min}\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right) < \lambda_{\max}\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right) < c'_\Sigma$ for all $k \in \mathcal{K}$ and $c_\Sigma'^{-1} < \lambda_{\min}\left(\widehat{\boldsymbol{\Sigma}}\right) < \lambda_{\max}\left(\widehat{\boldsymbol{\Sigma}}\right) < c'_\Sigma$, with probability tending to one. Without less of generality, we suppose $c_\Sigma = c'_\Sigma$. Therefore, we have that $\left\|\widehat{\boldsymbol{\Sigma}}^{-1}\right\|_2 = O_{\mathbb{P}}(1)$ and $\sup_k \left\|\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1}\right\|_2 = O_{\mathbb{P}}(1)$.

We first establish the convergence rate for Step I. Let

$$\boldsymbol{w}_\tau := \boldsymbol{\beta}_\tau + \boldsymbol{\delta}_\tau, \tag{34}$$

where

$$\boldsymbol{\delta}_\tau := \overline{\boldsymbol{\Sigma}}^{-1}\left(\sum_{0\leq k\leq K} \alpha_k \overline{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\delta}_\tau^{(k)}\right). \tag{35}$$

We will first prove that

$$\|\boldsymbol{w}_\tau - \widehat{\boldsymbol{w}}_\tau\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}} + p^{-\kappa/d}\right). \tag{36}$$

For simplicity, we omit the subscript τ in the following proof when there is no ambiguity.

Recall that $\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta} + \boldsymbol{\delta}^{(k)}$ and

$$\boldsymbol{y}^{(k)} = \boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)} + \left[\boldsymbol{f}^{(k)} - \boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right] + \boldsymbol{e}^{(k)}.$$

Then the error of the estimator $\widehat{\boldsymbol{w}}$ in Step I can be decomposed as

$$\begin{aligned}\widehat{\boldsymbol{w}}-\boldsymbol{w} &= \widehat{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{n_{\mathcal{K}}}\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\boldsymbol{y}^{(k)}\right)-\boldsymbol{\beta}-\overline{\boldsymbol{\Sigma}}^{-1}\left(\sum_{0\le k\le K}\alpha_k\overline{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\delta}^{(k)}\right)\\ &= \underbrace{\left(\widehat{\boldsymbol{\Sigma}}^{-1}\sum_{0\le k\le K}\alpha_k\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\delta}^{(k)}-\boldsymbol{\delta}\right)}_{\boldsymbol{E}_1}+\underbrace{\widehat{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{n_{\mathcal{K}}}\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\left[\boldsymbol{f}^{(k)}-\boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right]\right)}_{\boldsymbol{E}_2}\\ &\quad+\underbrace{\widehat{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{n_{\mathcal{K}}}\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\boldsymbol{e}^{(k)}\right)}_{\boldsymbol{E}_3}.\end{aligned} \tag{37}$$

Then we separately bound $\boldsymbol{E}_1, \boldsymbol{E}_2$ and $\boldsymbol{E}_3$.

Bound on $\boldsymbol{E}_1$. Note that

$$\boldsymbol{E}_1=\widehat{\boldsymbol{\Sigma}}^{-1}\left[\left(\overline{\boldsymbol{\Sigma}}-\widehat{\boldsymbol{\Sigma}}\right)\boldsymbol{\delta}+\Big(\sum_{0\le k\le K}\alpha_k\left(\widehat{\boldsymbol{\Sigma}}^{(k)}-\overline{\boldsymbol{\Sigma}}^{(k)}\right)\boldsymbol{\delta}^{(k)}\Big)\right].$$

By the proof of Lemma 3 and Lemma 4 in Shi et al. (2022), we have

$$\left\|\overline{\boldsymbol{\Sigma}}-\widehat{\boldsymbol{\Sigma}}\right\|_2=O_{\mathbb{P}}\left(\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}}\right),\left\|\overline{\boldsymbol{\Sigma}}^{(k)}-\widehat{\boldsymbol{\Sigma}}^{(k)}\right\|_2=O_{\mathbb{P}}\left(\sqrt{\frac{p\log^2(n_k)}{n_k}}\right).$$

Therefore,

$$\begin{aligned}\|\boldsymbol{E}_1\|_2 &\le \left\|\widehat{\boldsymbol{\Sigma}}^{-1}\right\|_2\left[\left\|\overline{\boldsymbol{\Sigma}}-\widehat{\boldsymbol{\Sigma}}\right\|_2\|\boldsymbol{\delta}\|_2+\sum_{0\le k\le K}\alpha_k\left\|\left(\widehat{\boldsymbol{\Sigma}}^{(k)}-\overline{\boldsymbol{\Sigma}}^{(k)}\right)\right\|_2\left\|\boldsymbol{\delta}^{(k)}\right\|_2\right]\\ &\lesssim \sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}}\left(1+\sum_{0\le k\le K}\sqrt{\frac{n_k}{n_{\mathcal{K}}}}\right)\\ &\lesssim \sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}}.\end{aligned} \tag{38}$$

Bound on $\boldsymbol{E}_2$. Since $\sup_k\left\|\boldsymbol{f}^{(k)}-\boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right\|_\infty\le C'p^{-\kappa/d}$, we have that for any vector $\boldsymbol{v}\in\mathbb{R}^{mp}$,

$$
\begin{aligned}
&\frac{1}{n_{\mathcal{K}}}\boldsymbol{v}^{\top}\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\left[\boldsymbol{f}^{(k)}-\boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right]\\
&=\frac{1}{n_{\mathcal{K}}}\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\boldsymbol{v}\right)^{\top}\left[\boldsymbol{f}^{(k)}-\boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right]\\
&\le\frac{Cp^{-\kappa/d}}{n_{\mathcal{K}}}\sum_{0\le k\le K}\left\|\boldsymbol{Z}^{(k)}\boldsymbol{v}\right\|_1\\
&\le Cp^{-\kappa/d}\sqrt{\boldsymbol{v}^{\top}\left(\frac{1}{n_{\mathcal{K}}}\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\boldsymbol{Z}^{(k)}\right)\boldsymbol{v}}\\
&\le Cp^{-\kappa/d}\sqrt{\left\|\widehat{\boldsymbol{\Sigma}}\right\|_2}\left\|\boldsymbol{v}\right\|_2,
\end{aligned}
$$

where the last inequality follows from the Cauchy-Schwartz inequality. Therefore, we have that

$$
\left\|\boldsymbol{E}_2\right\|_2\lesssim p^{-\kappa/d}.
$$

Bound on $\boldsymbol{E}_3$. We have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\boldsymbol{e}^{(k)}\right\|_2^2\right]\\
&=\mathbb{E}\left[\left(\sum_{0\le k\le K}\sum_{i,t}\boldsymbol{\xi}\left(\boldsymbol{X}_{i,t}^{(k)},A_{i,t}^{(k)}\right)e_{i,t}^{(k)}\right)^{\top}\left(\sum_{0\le k\le K}\sum_{i,t}\boldsymbol{\xi}\left(\boldsymbol{X}_{i,t}^{(k)},A_{i,t}^{(k)}\right)e_{i,t}^{(k)}\right)\right]\\
&=\mathbb{E}\left[\sum_{0\le k\le K}\sum_{i=1}^{n_k}\boldsymbol{\xi}^{\top}\left(\boldsymbol{X}_{i,t}^{(k)},A_{i,t}^{(k)}\right)\boldsymbol{\xi}\left(\boldsymbol{X}_{i,t}^{(k)},A_{i,t}^{(k)}\right)\left(e_{i,t}^{(k)}\right)^2\right]\\
&\le(n_{\mathcal{K}})\sup_{\boldsymbol{x}}\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2\sup_k\mathbb{E}\left[\left(e_{i,t}^{(k)}\right)^2\right],
\end{aligned}
$$

where the second equality follows from the independence between $e_{i,t}^{(k)}$ with different (i,t) or different k. Lemma 2 in Shi et al. (2022) shows that $\sup_{\boldsymbol{x}}\left\|\boldsymbol{\phi}(\boldsymbol{x})\right\|_2^2\lesssim p$. Since $e_{i,t}^{(k)}$ is sub-Gaussian with a uniform sub-Gaussian parameter, we obtain that

$$
\mathbb{E}\left[\left\|\sum_{0\le k\le K}\left(\boldsymbol{Z}^{(k)}\right)^{\top}\boldsymbol{e}^{(k)}\right\|_2^2\right]\lesssim(n_{\mathcal{K}})p.
$$

Then by Markov's inequality, we have

$$\|\boldsymbol{E}_3\|_2 = \mathcal{O}_p\left(\sqrt{\frac{p}{n_{\mathcal{K}}}}\right).$$

Combining the bounds on $\boldsymbol{E}_1, \boldsymbol{E}_2$ and $\boldsymbol{E}_3$, we obtain that

$$\|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|_2 = \mathcal{O}_p\left(\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}} + p^{-\kappa/d}\right). \tag{39}$$

Now we turn to the analysis of Step II. The quantity $\boldsymbol{\delta}$ defined in (35) is a weighted sum of the sparse difference vectors $\boldsymbol{\delta}^{(k)}$ between the tasks. By the definition of $\boldsymbol{\delta}$ and C_Σ, it is easy to show that $\|\boldsymbol{\delta}\|_1 \le C_\Sigma h$. By the definition of $\widehat{\boldsymbol{\delta}}^{(k)}$ in Step II, for a fixed k, define $\boldsymbol{v}^{(k)} = \boldsymbol{\delta}^{(k)} - \boldsymbol{\delta}$, then it holds that

$$\frac{1}{2n_k}\left\|\boldsymbol{y}^{(k)} - \boldsymbol{Z}^{(k)}\left(\widehat{\boldsymbol{w}} + \widehat{\boldsymbol{\delta}}^{(k)}\right)\right\|_2^2 + \lambda_\delta\left\|\widehat{\boldsymbol{\delta}}^{(k)}\right\|_1 \le \frac{1}{2n_k}\left\|\boldsymbol{y}^{(k)} - \boldsymbol{Z}^{(k)}\left(\widehat{\boldsymbol{w}} + \boldsymbol{v}^{(k)}\right)\right\|_2^2 + \lambda_\delta\left\|\boldsymbol{v}^{(k)}\right\|_1.$$

With some algebra, the above equation is equivalent to

$$\begin{aligned}\frac{1}{2}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big)^\top\widehat{\boldsymbol{\Sigma}}^{(k)}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big) \le{}& \lambda_\delta\left(\left\|\boldsymbol{v}^{(k)}\right\|_1 - \left\|\widehat{\boldsymbol{\delta}}^{(k)}\right\|_1\right)\\ &+ \frac{1}{n_k}\left(\boldsymbol{e}^{(k)}\right)^\top\boldsymbol{Z}^{(k)}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big) - \big(\widehat{\boldsymbol{w}} - \boldsymbol{w}\big)^\top\widehat{\boldsymbol{\Sigma}}^{(k)}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big)\\ &+ \frac{1}{n_k}\left[\boldsymbol{f}^{(k)} - \boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right]^\top\boldsymbol{Z}^{(k)}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big).\end{aligned}$$

Since $\boldsymbol{e}^{(k)}$ is sub-Gaussian, it is standard to show that

$$\left\|\frac{1}{n_k}(\boldsymbol{e}^{(k)})^\top\boldsymbol{Z}^{(k)}\right\|_\infty = O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{n_k}}\right).$$

By taking $\lambda_\delta = c_\lambda\sqrt{\frac{\log p}{n_k}}$ for sufficiently large c_λ, we have

$$\frac{1}{n_k}(\boldsymbol{e}^{(k)})^\top\boldsymbol{Z}^{(k)}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big) \le \frac{\lambda_\delta}{2}\left\|\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\right\|_1,$$

with probability tending to one.

Next, since $c_\Sigma^{-1} \le \lambda_{\min}\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right) \le \lambda_{\max}\left(\widehat{\boldsymbol{\Sigma}}^{(k)}\right) \le c_\Sigma$, we have

$$\frac{1}{2}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big)^\top\widehat{\boldsymbol{\Sigma}}^{(k)}\big(\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\big) \ge \frac{1}{2c_\Sigma}\left\|\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\right\|_2^2,$$

and

$$-\left(\widehat{\boldsymbol{w}}-\boldsymbol{w}\right)^{\top}\widehat{\boldsymbol{\Sigma}}^{(k)}\left(\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right) \leq c_{\Sigma}\left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2 \leq 4c_{\Sigma}^2\left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2^2+\frac{1}{4c_{\Sigma}}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2^2.$$

Finally, by the proof of the bound on $\boldsymbol{E}_2$, we have that

$$\frac{1}{n_k}\left[\boldsymbol{f}^{(k)}-\boldsymbol{Z}^{(k)}\boldsymbol{\beta}^{(k)}\right]^{\top}\boldsymbol{Z}^{(k)}\left(\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right) \leq Cp^{-\kappa/d}c_{\Sigma}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2.$$

Combining the above inequalities leads to

$$\begin{aligned}&\frac{1}{4c_{\Sigma}}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2^2\\ &\leq \lambda_{\delta}\left(\left\|\boldsymbol{v}^{(k)}\right\|_1-\left\|\widehat{\boldsymbol{\delta}}^{(k)}\right\|_1\right)+\frac{\lambda_{\delta}}{2}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_1+4c_{\Sigma}^2\left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2^2+Cp^{-\kappa/d}c_{\Sigma}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2.\end{aligned}$$

Using the fact that $\left\|\widehat{\boldsymbol{\delta}}^{(k)}\right\|_1 \geq \left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_1-\left\|\boldsymbol{v}^{(k)}\right\|_1$, we obtain

$$\frac{1}{4c_{\Sigma}}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2^2 \leq 2\lambda_{\delta}\left\|\boldsymbol{v}^{(k)}\right\|_1-\frac{\lambda_{\delta}}{2}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_1+4c_{\Sigma}^2\left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2^2+Cp^{-\kappa/d}c_{\Sigma}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2.$$

Among the three positive terms $2\lambda_{\delta}\left\|\boldsymbol{v}^{(k)}\right\|_1$, $4c_{\Sigma}^2\left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2^2$ and $Cp^{-\kappa/d}c_{\Sigma}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2$ on the RHS of the above inequality, if

(a) $2\lambda_{\delta}\left\|\boldsymbol{v}^{(k)}\right\|_1$ is the dominant one, then we have

$$\frac{1}{4c_{\Sigma}}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2^2 \leq 6\lambda_{\delta}\left\|\boldsymbol{v}^{(k)}\right\|_1,$$

and

$$\frac{\lambda_{\delta}}{2}\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_1 \leq 6\lambda_{\delta}\left\|\boldsymbol{v}^{(k)}\right\|_1,$$

which implies that

$$\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2 \lesssim \min\left\{\left\|\boldsymbol{v}^{(k)}\right\|_1, \sqrt{\lambda_{\delta}\left\|\boldsymbol{v}^{(k)}\right\|_1}\right\}.$$

(b) $4c_{\Sigma}^2\left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2^2$ is the dominant one, then we have

$$\left\|\widehat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{v}^{(k)}\right\|_2 \lesssim \left\|\widehat{\boldsymbol{w}}-\boldsymbol{w}\right\|_2.$$

(c) $Cp^{-\kappa/d}c_\Sigma \left\|\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\right\|_2$ is the dominant one, then we have

$$\left\|\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\right\|_2 \lesssim p^{-\kappa/d}.$$

Recall that $\boldsymbol{v}^{(k)} = \boldsymbol{\delta}^{(k)} - \boldsymbol{\delta}$. Since $\left\|\boldsymbol{v}^{(k)}\right\|_1 \lesssim C_\Sigma h$ and

$$\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}} + p^{-\kappa/d}\right),$$

we finally obtain

$$\left\|\widehat{\boldsymbol{\delta}}^{(k)} - \boldsymbol{v}^{(k)}\right\|_2 = O_{\mathbb{P}}\left(\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}} + p^{-\kappa/d} + C_\Sigma h \wedge \sqrt{C_\Sigma h\sqrt{\frac{\log p}{n_k}}}\right).$$

This is true for any k. Since $\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{w}} + \widehat{\boldsymbol{\delta}}^{(0)} - \boldsymbol{\beta} - \boldsymbol{\delta}^{(0)} = \widehat{\boldsymbol{w}} - \boldsymbol{\beta} - \boldsymbol{\delta} + \widehat{\boldsymbol{\delta}}^{(0)} - \boldsymbol{\delta}^{(0)} + \boldsymbol{\delta}$, the proof is completed.

It is straightforward to show by induction that

$$\sup_{\tau\geq 1}\left\|\widehat{\boldsymbol{\beta}}_\tau^{(0)} - \boldsymbol{\beta}_\tau^{(0)}\right\|_2 = \mathcal{O}_p\left(\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}} + p^{-\kappa/d} + C_\Sigma h \wedge \sqrt{C_\Sigma h}\left(\frac{\log p}{n_0}\right)^{1/4}\right).$$

Recall that

$$\begin{aligned}
&\mathbb{E}_{\mu^{(0)},b^{(0)}}\left|Q^* - Q^{\pi_\Upsilon}\right| \\
&\lesssim \sum_{\tau=0}^{\Upsilon-1}\sum_{j=0}^{\infty}\gamma^{\Upsilon-\tau+j}\sup_{\omega_1,\ldots,\omega_{\Upsilon-\tau+j}}\mathbb{E}_{\mu^{(0)},b^{(0)}}\left[\left(P^{\omega_{\Upsilon-\tau+j}}P^{\omega_{\Upsilon-\tau+j-1}}\cdots P^{\omega_1}\right)\left|\mathcal{T}^{(0)}\widehat{Q}_\tau^{(0)} - \widehat{Q}_{\tau+1}^{(0)}\right|\right] \\
&\quad + \frac{\gamma^{\Upsilon+1}}{(1-\gamma)^2}R_{\max},
\end{aligned} \tag{40}$$

where $\omega_1,\ldots,\omega_{\Upsilon-\tau}$ are policy functions and $\mathbb{E}_{\mu,b}$ denotes the expectation with respect to the invariant distribution $\mu(\boldsymbol{x})b(a\,|\,\boldsymbol{x})$. Denote the marginal distribution of the second-to-the-last state $(P^{\omega_{\Upsilon-\tau+j-2}}\cdots P^{\omega_1})(\boldsymbol{X},A)$, $(\boldsymbol{X},A)\sim\mu^{(0)}(\boldsymbol{x})b^{(0)}(a\,|\,\boldsymbol{x})$ by $\widetilde{\mu}$. Since $\rho^{(0)}$ is upper-bounded and $\mu^{(0)}$ is lower bounded, we have that

$$
\begin{aligned}
&\mathbb{E}_{\mu^{(0)},b^{(0)}}\left[(P^{\omega_{\Upsilon-\tau+j}}P^{\omega_{\Upsilon-\tau+j-1}}\cdots P^{\omega_1})\left|\mathcal{T}^{(0)}\widehat{Q}_\tau^{(0)}-\widehat{Q}_{\tau+1}^{(0)}\right|\right]\\
&=\int_{\mathcal{X}}\sum_{a'\in\mathcal{A}}\left|\mathcal{T}^{(0)}\widehat{Q}_\tau^{(0)}-\widehat{Q}_{\tau+1}^{(0)}\right|(\boldsymbol{x}',a')\omega_{\Upsilon-\tau+j}(a'\mid\boldsymbol{x}')\int_{\mathcal{X}}\sum_{a\in\mathcal{A}}\rho^{(0)}(\boldsymbol{x}'\mid\boldsymbol{x},a)\omega_{\Upsilon-\tau+j-1}(a\mid\boldsymbol{x})\widetilde{\mu}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\\
&\lesssim\int_{\mathcal{X}}\sum_{a'\in\mathcal{A}}\left|\mathcal{T}^{(0)}\widehat{Q}_\tau^{(0)}-\widehat{Q}_{\tau+1}^{(0)}\right|(\boldsymbol{x}',a')\omega_{\Upsilon-\tau+j}(a'\mid\boldsymbol{x}')\mu^{(0)}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'\\
&\lesssim\int_{\mathcal{X}}\sum_{a'\in\mathcal{A}}\left[\boldsymbol{\xi}^\top(\boldsymbol{x}',a')\left(\widehat{\boldsymbol{\beta}}_\tau^{(0)}-\boldsymbol{\beta}_\tau^{(0)}\right)+p^{-\kappa/d}\right]\omega_{\Upsilon-\tau+j}(a'\mid\boldsymbol{x}')\mu^{(0)}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'\\
&\lesssim p^{-\kappa/d}+\sup_{\tau\geq1}\left\|\widehat{\boldsymbol{\beta}}_\tau^{(0)}-\boldsymbol{\beta}_\tau^{(0)}\right\|_2.
\end{aligned}
\tag{41}
$$

Adding that

$$
\sum_{\tau=0}^{\Upsilon-1}\sum_{j=0}^{\infty}\gamma^{\Upsilon-\tau+j}\leq\frac{\gamma}{(1-\gamma)^2},
\tag{42}
$$

we finally obtain

$$
\begin{aligned}
&\mathbb{E}_{\mu^{(0)},b^{(0)}}\left|Q^*-Q^{\pi_\Upsilon}\right|\\
&=\mathcal{O}_p\left(\frac{\gamma}{(1-\gamma)^2}\left[p^{-\kappa/d}+\sqrt{\frac{p\log^2(n_{\mathcal{K}})}{n_{\mathcal{K}}}}+C_\Sigma h\wedge\sqrt{C_\Sigma h}\left(\frac{\log p}{n_0}\right)^{1/4}\right]+\frac{\gamma^{\Upsilon+1}}{(1-\gamma)^2}R_{\max}\right).
\end{aligned}
\tag{43}
$$