Off-Policy Estimation of Long-Term Average Outcomes with Applications to Mobile Health

Peng Liao
Department of Statistics, University of Michigan

Predrag Klasnja
School of Information, University of Michigan

Susan Murphy
Department of Statistics, Harvard University

Abstract

Due to recent advancements in mobile device and sensing technology, health scientists are increasingly interested in developing mobile health (mHealth) treatments that are delivered to individuals at moments in which they are most effective in influencing the individual's behavior. The mHealth intervention policies, also called just-In-time adaptive Interventions, are decision rules that map an individual's context to a treatment at each of many time points. Many mHealth interventions are designed for longer-term use, however their long-term efficacy is not well understood. In this work, we provide an approach for conducting inference about the long-term performance of one or more such policies using historical data collected under a possibly different policy. Our performance measure is the average of proximal outcomes (rewards) over a long time period should the particular mHealth policy be followed. We model the relative value function by a nonparametric function class and develop a coupled, penalized estimator of the average reward. We show that the proposed estimator is asymptotically normal when the number of trajectories goes to infinity. This work is motivated by HeartSteps, an mHealth physical activity intervention.

Keywords: sequential decision making, policy evaluation, markov decision process, reinforcement learning

1 Introduction

Due to the recent advancement in mobile device and sensing technology, health scientists are more and more interested in developing mobile health (mHealth) interventions. In mHealth, mobile devices (e.g., wearables and smartphones) are used to deliver interventions to individuals as they go about their daily lives. In general, there are two types of mHealth treatments. Most are "pull" treatments that reside on the individual's mobile device and allow the individual to access treatment content as needed. This work focuses on the second type, the "push" treatment, typically in the form of a notification or a text message that appears on a mobile device. There is a wide variety of possible treatment messages (behavioral, cognitive, motivational, reminders and so on). These treatments are generally intended to impact a near time, proximal outcome, such as stress or behaviors such as physical activity over some subsequent minutes/hours. The mHealth intervention policies, often called just-in-time adaptive interventions in the mHealth literature (Nahum-Shani et al. 2018), are decision rules that map the individual's current context (e.g., the location, time, social activity, stress level and urge to smoke) to a particular treatment at each of many time points. Many mHealth interventions are designed for long-term use in chronic disease management (Lee et al. 2018). The vast majority of current mHealth interventions deploy expert-derived policies with limited use of data evidence (for an example see Kizakevich et al. (2014)), however the long-term efficacy of these policies on the health behavior is not well understood. An important first step toward developing data-based, effective mHealth interventions is to properly measure the long-term performance of these policies. In this work, we provide an approach for conducting inference about the optimality of one or more mHealth policies of interest. Our optimality criteria is the long-term average of proximal outcomes should a particular mHealth policy be followed. We develop a flexible method to estimate the performance of an mHealth policy using a historical dataset in which the treatments are decided by a possibly different policy.

This work is motivated by HeartSteps (Klasnja et al. 2015), an mHealth physical activity intervention. To design this intervention, we are conducting a series of studies. The first, already completed, study was for 42 days. The last study will be for one year. Here we focus on the intervention component involving contextually-tailored activity suggestion messages. These messages may be delivered at each of five user-specified times per day. While in the first study there were $42 \times 5 = 210$ time points per user, in the year-long study there will be about 2,000 time points per user. The proximal outcome is the step count in the 30 minutes following each of the five times per day. Our goal is to use the data collected from the first 42-day study to predict and estimate the long-term average of proximal outcomes for a variety of policies that could be used to decide whether or not to send the contextually tailored activity suggestion in the year-long study. The 42-day study is a Micro-Randomized Trial (MRT) (Klasnja et al. 2015, Liao et al. 2016). In an MRT, a known stochastic policy, also called a behavior policy, is used to decide when and which type of treatment to provide at each time point. A partial list of MRTs in the field or completed can be found at the website of The Methodology Center [1]. From an experimental point of view, the stochastic behavior policy is used to conduct sequential randomizations within each user. That is, the stochastic behavior policy sequentially randomizes each user among the treatment options. In this work we focus on settings in which the behavior policy is stochastic and known; this is the case with MRTs by design.

The rest of the article is organized as follows. Section 2 provides the background of Markov Decision Processes, a probabilistic model for sequential decision making. In Section 3, we review the related work. In Section 4, we present our method to estimate the average reward. Section 5 presents our main theoretical results, including the global convergence rate and the asymptotic distribution of estimated average reward. In Section 6, we describe a procedure to select the tuning parameters and apply the proposed method to simulated data to examine the coverage probability of the confidence intervals in various settings. A case study using data from the 42 day MRT of HeartSteps is presented in Section 7. We end with a discussion of future work in Section 8.

[1] `https://www.methodology.psu.edu/ra/adap-inter/mrt-projects`

2 Markov Decision Process

We use Markov Decision Progress (MDP) to model the sequential decision-making process in the infinite horizon setting (Howard 1960, Puterman 1994, Sutton & Barto 2018). We use $t \in \mathbb{N}^+$ to index the decision time. Let $S_t \in \mathcal{S}$ be the state and $A_t \in \mathcal{A}$ be the action at time t. We assume that the reward (or the proximal outcome), denoted by $R_{t+1} \in \mathbb{R}$, is a known function of (S_t, A_t, S_{t+1}), which is used to measure the effectiveness of choosing the current action. In this work, we focus on the case of continuous outcome; see Section 8 for a discussion about other types of outcomes. The data-generating process is assumed to be Markovian, i.e., for $t \geq 1$, $S_{t+1} \perp \{S_1, A_1, \ldots, S_{t-1}, A_{t-1}\} \,|\, \{S_t, A_t\}$. Furthermore, we assume this conditional distribution is time-homogeneous. Denote the transition kernel by P, so that given a measurable set B in the state space, $\mathcal{S}$, $P(B|s,a) = \Pr(S_{t+1} \in B | S_t = s, A_t = a)$, which does not depend on t because of the assumed time homogeneity. Denote by $p(s'|s,a)$ the transition density with respect to some reference measure on $\mathcal{S}$ (e.g., the counting measure when $\mathcal{S}$ is discrete). We use $r(s,a)$ to denote the conditional expectation of reward given state and action, i.e., $r(s,a) = \mathbb{E}\left[R_{t+1} | S_t = s, A_t = a\right]$.

We consider assessing the class of time-stationary, stochastic policies that takes the state as input and outputs a probability distribution on the action space, $\mathcal{A}$. We use π to denote such a policy, so that $\pi(a|s)$ is the probability of choosing the action, a, at the state, s. Note that the policy, π, can be different from the behavior policy that is used to choose the action in the training data. We define the average reward of the policy, π, by

$$\eta^\pi(s) = \lim_{t^* \to \infty} \mathbb{E}_\pi \left[\frac{1}{t^*} \sum_{t=1}^{t^*} R_{t+1} \,\middle|\, S_1 = s \right], \tag{1}$$

where the expectation, $\mathbb{E}_\pi$, is taken over the trajectory $\{S_1, A_1, S_2, \ldots, S_{t^*}, A_{t^*}, S_{t^*+1}\}$ in which the actions are selected according to the policy π, that is, the likelihood in the expectation is given by $\mathbb{1}_{\{S_1 = s\}} \prod_{t=1}^{t^*} \pi(A_t|S_t) p(S_{t+1} \mid S_t, A_t)$. The policy π induces a Markov chain on the state with the transition kernel, $P^\pi(\cdot|s) = \sum_a \pi(a|s) P(\cdot|s,a)$. Suppose for now the state space, $\mathcal{S}$, is finite. It is well known (Puterman 1994) that when the markov chain P^π is irreducible and aperiodic, the average reward in (1) is well-defined and independent of the initial state, that is, for all $s \in \mathcal{S}$,

$$\eta^\pi(s) = \eta^\pi = \sum_{s,a} \pi(a|s) d^\pi(s) r(s,a), \tag{2}$$

where $d^\pi(s)$ is the stationary distribution (the existence of d^π is guaranteed by irreducibility and aperiodicity of P^π; see Puterman (1994)). The above result can be generalized to general state spaces, e.g., $\mathcal{S} \subset \mathbb{R}^d$, with more involved conditions on the transition kernel, P^π, analogous to irreducibility and aperiodic in the finite state case (see, for example, Hernández-Lerma & Lasserre (1999), chap. 7).

By the definition, the average reward, η^π, is the limit of the average of the rewards over a finite time horizon, i.e., $\mathbb{E}_\pi[\frac{1}{t^*}\sum_{t=1}^{t^*} R_{t+1} \mid S_1 = s]$. In fact, one can show that $\sup_{s \in S} |\mathbb{E}_\pi[(1/t^*)\sum_{t=1}^{t^*} R_{t+1} | S_1 = s] - \eta^\pi| = O(1/t^*)$, where the leading constant depends on the mixing time of P^π (see Theorem 7.5.10 in Hernández-Lerma & Lasserre (1999)). We propose to conduct inference about the long-term performance of each policy, π, based on its average reward, η^π, which can be viewed as an asymptotic surrogate of the average of finite rewards collected over a long period of time.

Recall in the HeartSteps example, the action is whether or an activity suggestion message is delivered and the reward (i.e., the proximal outcome) is the 30-min step count following each decision time. The average reward, η^π, is a proxy to the average of the 30-min step counts when the policy π is used to determine whether to send the activity suggestion messages over a long time period. We emphasize that the target quantity, η^π, does not depend on the length of trajectory (i.e., the study duration in an MRT) in the dataset at hand, which might be determined by other considerations (e.g., the budget).

In the following we will focus on the setting where the average reward, η^π, does not depend on the initial state (e.g., the induced markov chain is irreducible). Obviously, for this assumption to hold, any time-invariant information is not allowed to include in the state. Motivated by our mHealth application, in Supplement A we present a generalization that allows the average reward to depend on time-invariant variables. In the case of mHealth, time-invariant variables might be gender, baseline severity, genetics and so on. Another important assumption in the average reward framework is that the underlying dynamic is time-stationary (e.g., the expected reward, $\mathbb{E}[R_{t+1}|S_t = s, A_t = a]$, does not depend on time t). In mHealth interventions, non-stationarity primarily occurs due to the unobserved aspects of the current context (e.g., user's engagement and/or burden). Therefore, it is critical to collect as much information as possible (via self-report or wearable sensors) in order to sufficiently represent user's context.

We develop methods to conduct inference about the average reward η_π of a pre-specified target policy, π, using data collected under possibly different behavior policies. An alternate, and more common, evaluation measure is based on an expected discounted sum of rewards, $\mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} R_{t+1} \mid S_1 = s\right]$, with the discount rate, $\gamma \in [0, 1)$. When the discount rate,γ, is small (e.g., $\gamma = 0.5$), the discounted sum of rewards focuses only on finitely many near-term rewards. Note that even with a large discount rate of $\gamma = 0.99$, the rewards at time $t = 100$ have a weight of 0.37 and the rewards at time $t = 200$ have a weight of 0.13. Recall our motivating mHealth intervention is being designed to optimize the overall physical activity in one year. From a scientific point of view, the rewards in the distant future are as important as the near-term ones, especially when considering the effect of habituation and burden. With this in mind, we opt for the long-term average reward, which can be viewed as a proxy for the (undiscounted) average of rewards over a long period of time. On the other hand, note that as $\gamma \to 1$, the above conditional expectation of the sum of the discounted rewards normalized by the constant, $1/(1-\gamma)$, converges to the average reward, η_π (Mahadevan 1996). Because the Bellman operator in the discounted setting is a contraction (Sutton & Barto 2018), researchers focused on a discounted sum of rewards to ensure online computational stability and simplify associated convergence arguments. However, as we shall see below, consideration of the average reward is not problematic in the batch (i.e., off-line) setting.

3 Related Work

The evaluation of a given target policy using data collected from different policy (the behavior policy) is called off-policy evaluation. This has been widely studied in both the statistical and reinforcement learning (RL) literature. Many authors evaluate and contrast policies in terms of the expected sum of rewards over a finite number of time points (Murphy et al. 2001, Chakraborty & Moodie 2013, Jiang & Li 2015). However because these methods use products of weights with probabilities from the behavior policy in the denominator, the extension to problems with a large number of time points often suffers from a large variance (Thomas & Brunskill 2016, Jiang & Li 2015).

The most common off-policy evaluation methods for infinite-horizon problems (i.e., a large number of time points) focus on a discounted sum of rewards and are thus based in some way on the value function (in the discounted reward setting $\mathbb{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_{t+1}\,|\,S_1=s\right]$ considered as a function of s is the value function). Farahmand et al. (2016) proposed a regularized version of Least Square Temporal Difference (Bradtke et al. 1996) was proposed and statistical properties studied. They used a non-parametric model to estimate the value function and derived the convergence rate when training data consists of i.i.d. transition sample that consists of current state, action and next state. From a technical point of view, our estimation method is similar to Farahmand et al. (2016), albeit focused on the average reward; most importantly our method relaxes the assumption that Bellman operator can be modeled correctly for each candidate relative value function and only assumes the data consists of i.i.d. samples of trajectories. Luckett et al. (2019) also focused on the discounted reward setting. They evaluated policies, π based on an average of $\mathbb{E}_\pi\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_{t+1}\,|\,S_1=s\right]$ with respect to a pre-selected reference distribution for s. While the reference distribution can be naturally chosen as the distribution of the initial state (Farajtabar et al. 2018, Liu et al. 2018, Luckett et al. 2019, Thomas & Brunskill 2016), choosing a "right" discount rate, γ, can be non-trivial, at least in mHealth. They assumed a parametric model for the value function and developed a regularized estimating equation. In computer science literature, there also exists many off-policy evaluation methods for the discounted reward setting . We refer the interested reader to the recent works by Farajtabar et al. (2018) and Kallus & Uehara (2019) and references therein.

Closest to the setting of this work is the recent work by Murphy et al. (2016) and Liu et al. (2018). Murphy et al. (2016) considered the average reward setting. They assumed a linear model for the value function and constructed the estimating equations to estimate the average reward. However the linearity assumption of the value function is unlikely to hold in practice and difficult to validate (e.g., the value function involves the infinite sum of the rewards). Our method allows the use of a non-parametric model for the value function to increase robustness. Liu et al. (2018) also considered the average reward and proposed an estimator of the average reward based on estimating the ratio of the stationary distribution under the target policy divided by the stationary distribution under the behavior policy. However they do not provide confidence intervals or other inferential methods besides an estimator of the average reward. In addition, they assume that the behavior policy is Markovian and time-stationary. In mHealth the behavior policy can be determined by an algorithm based on accruing data and thus violates this assumption (Liao et al. 2018, Dempsey et al. 2017).

4 Estimator for Off-Policy Evaluation

We consider the setting where we have observations of n trajectories:

$$\mathcal{D}_n = \{\mathcal{D}^i\}_{i=1}^n = \{S_1^i, A_1^i, S_2^i, \ldots, S_T^i, A_T^i, S_{T+1}^i\}_{i=1}^n.$$

For simplicity, we assume the length of trajectory, T, is non-random and identical for each trajectory. Each trajectory $\mathcal{D}^i$ is an independent, identically distributed (i.i.d.) copy of $\mathcal{D} = \{S_1, A_1, S_2, \ldots, S_{T+1}\}$, which is assumed to follow a MDP specified in Section 2. The actions, $\{A_t\}_{t=1}^T$, are selected according to a behavior policy, $\pi_b = \{\pi_1^b, \ldots, \pi_T^b\}$, i.e., $A_t \sim \pi_t^b(\cdot \mid H_t)$, where $H_t = \{S_1, A_1, \ldots, S_t\}$ is the history collected up to time t. As would be the case for an MRT, throughout we assume that the behavior policy is known and the behavior policy results in strictly positive randomization probabilities, e.g., $\pi_t^b(a|H_t) \geq p_{\min} > 0$ for all $a \in \mathcal{A}$, H_t and $t \leq T$. In the following, the expectation $\mathbb{E}$ without the subscript is with respect to the distribution of the trajectory, $\mathcal{D}$, induced by the behavior policy, π_b.

Consider a target policy π, possibly different from the behavior policy π_b used in the training data. Throughout we only consider the target policies that are Markovian (i.e., only depends on the current state) and time-invariant (i.e., the mapping does not vary with time). Note that for each trajectory in the training data, $\mathcal{D}_n$, we only observe the first T time points. Our goal is to learn the average of the rewards over a long period of time which can be much long than T and we approximate this by the average reward η^π. Below we introduce the estimator for η^π.

In this work, we follow the so-called "model-free" approach (i.e., does not require modeling the state transition kernel) to estimate the average reward. Our estimator is based on the Bellman equations, also known as the Poisson equation (Puterman 1994), to characterize the average reward. We define the relative value function, Q^π, by

$$Q^\pi(s,a) = \mathbb{E}_\pi\Big[\sum_{t=1}^{\infty}(R_{t+1} - \eta^\pi)\,\Big|\, S_1 = s, A_1 = a\Big]. \tag{3}$$

It is easy to verify by definition that (η^π, Q^π) is the solution of the Bellman equation:

$$\mathbb{E}_\pi\left[R_{t+1} + Q(S_{t+1}, A_{t+1}) \,|\, S_t = s, A_t = a\right] = \eta + Q(s,a),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{4}$$

Furthermore, if the state space is finite and the transition probability matrix induced by the target policy is irreducible and aperiodic, the Bellman equations (4) uniquely identifies the average reward and identifies the relative value function Q^π up to a constant (see Puterman (1994), p. 343 for details). That is, the set of solutions of the Bellman equations (4) is given by $\{(\eta^\pi, Q) : Q = Q^\pi + c\mathbf{1}, c \in \mathbb{R}, \mathbf{1}(s,a) = 1\}$. This result can be generalized to general state spaces, e.g., the infinite discrete or continuous state space (see chap. 7 in Hernández-Lerma & Lasserre (1999)). The key requirement for our method is that the average reward is a constant and can be uniquely identified by solving Bellman equations (4) up to a constant.

Assumption 1. *The average reward of the target policy π is independent of state and satisfies (2). (η^π, Q^π) is the unique solution of Bellman equations (4) up to a constant for Q^π. The stationary distribution of the induced transition kernel, P^π, exists.*

As the focus of this work is to estimate the average reward, we only need to identify one specific version of Q^π in solving the Bellman equation. Define the shifted relative value function by $\tilde{Q}^\pi(s,a) = Q^\pi(s,a) - Q^\pi(s^*, a^*)$ for a specific state-action pair (s^*, a^*). Obviously $\tilde{Q}^\pi(s^*, a^*) = 0$ and $\tilde{Q}^\pi(s_1, a_1) - \tilde{Q}^\pi(s_2, a_2) = Q^\pi(s_1, a_1) - Q^\pi(s_2, a_2)$, e.g., the difference in the relative value remains the same. By restricting the function class such that $Q(s^*, a^*) = 0$, the solution of Bellman equations (4) is unique and given by $(\eta^\pi, \tilde{Q}^\pi)$.

In the following, we use $\mathcal{Q}$ to denote a vector space of functions on the state-action space $\mathcal{S} \times \mathcal{A}$ such that $Q(s^*, a^*) = 0$ for all $Q \in \mathcal{Q}$, in which we will assume $\tilde{Q}^\pi \in \mathcal{Q}$. The Bellman operator $\mathcal{T}_\pi$ with respect to the target policy π is given by

$$\mathcal{T}_\pi(s, a; Q) = \mathbb{E}\big[R_{t+1} + \sum_{a'} \pi(a'|S_{t+1}) Q(S_{t+1}, a') \,|\, S_t = s, A_t = a\big]. \tag{5}$$

Note that the above conditional expectation does not depend on the behavior policy as we condition on the current state and action. The Bellman error at (s, a) with respect to (η, Q) and π is defined as $\mathcal{T}_\pi(s, a; Q) - \eta - Q(s, a)$. This error is 0 at $\eta = \eta^\pi$, $Q = \tilde{Q}^\pi$. That is, from the Bellman equations (4) we have $\mathcal{T}_\pi(s, a; \tilde{Q}^\pi) = \eta^\pi + \tilde{Q}^\pi(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

The Bellman operator involves the (unknown) transition kernel, that is, the conditional distribution of the next state given the current state and action. Suppose for now that one have access to the transition kernel. Since the Bellman error is 0 at $\eta = \eta^\pi$, $Q = \tilde{Q}^\pi$, a natural way to estimate $(\eta^\pi, \tilde{Q}^\pi)$ is to minimize the empirical squared Bellman error, i.e.,

$$\min_{(\eta, Q) \in \mathbb{R} \times \mathcal{Q}} \mathbb{P}_n \Big\{ \frac{1}{T} \sum_{t=1}^{T} [\mathcal{T}_\pi(S_t, A_t; Q) - \eta - Q(S_t, A_t)]^2 \Big\}, \tag{6}$$

where $\mathbb{P}_n f(\mathcal{D}) = (1/n) \sum_{i=1}^{n} f(\mathcal{D}^i)$ is the empirical mean over the training data, $\mathcal{D}_n$, for a function of the trajectory, f. Obviously, this is not a feasible estimator as we don't know the transition kernel and thus $\mathcal{T}_\pi(S_t, A_t; Q)$ is unknown. A natural idea is to replace the Bellman operator by its sample counterpart, i.e., replace $\mathcal{T}_\pi(S_t, A_t; Q)$ by $R_{t+1} + \sum_{a'} \pi(a'|S_{t+1}) Q(S_{t+1}, a')$ in the objective function of (6). Unlike the regression problem, in which the dependent variable is fully observed, the dependent variable here is $R_{t+1} + \sum_{a'} \pi(a'|S_{t+1}) Q(S_{t+1}, a')$ which involves the conditional expectation of the unknown relative value function, Q. As a result, the natural plug-in estimator is biased; see Antos et al. (2008) for a similar discussion in the discounted reward setting. This motivates a coupled estimator in which we use the estimated Bellman error to form the objective function.

On the other hand, we note that even when $Q \in \mathcal{Q}$, the output of the Bellman operator is not necessarily in the function space, $\mathbb{R} \oplus \mathcal{Q} = \{c + Q : c \in \mathbb{R}, Q \in \mathcal{Q}\}$. To see this, use the fact that $\mathbb{E}[R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})\tilde{Q}^\pi(S_{t+1}, a')|S_t = s, A_t = a] - \eta^\pi - \tilde{Q}^\pi(s,a) = 0$ to obtain

$$\mathcal{T}_\pi(s, a; Q) = \eta^\pi + \tilde{Q}^\pi(s,a) - \mathbb{E}\Big[\sum_{a'} \pi(a'|S_{t+1})(\tilde{Q}^\pi - Q)(S_{t+1}, a') \,|\, S_t = s, A_t = a\Big].$$

Depending on the complexity of the transition kernel, the last term is unlikely to be in $\mathcal{Q}$ for every $Q \in \mathcal{Q}$. Assuming the last term is in $\mathcal{Q}$ for every $Q \in \mathcal{Q}$ is a much stronger than requiring $\tilde{Q}^\pi \in \mathcal{Q}$. Instead, we only need to form an approximation to the Bellman error. We introduce a second function class, $\mathcal{G}$, to form a projected error:

$$g^*_\pi(\cdot, \cdot; \eta, Q) = \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}(\mathcal{T}_\pi(S_t, A_t; Q) - \eta - Q(S_t, A_t) - g(S_t, A_t))^2\Big]. \tag{7}$$

In practice, one can use the same function class for both $\mathcal{Q}$ and $\mathcal{G}$, except that we require members of $\mathcal{Q}$ to satisfy $Q(s^*, a^*) = 0$; a similar constraint need not be placed on the members of $\mathcal{G}$. And in addition we do not assume that the error $\mathcal{T}_\pi(\cdot, \cdot; \eta, Q) - \eta - Q(\cdot, \cdot)$ is in function class, $\mathcal{G}$. The key reason why the projected Bellman error from (7) allows us to identify $(\eta^\pi, \tilde{Q}^\pi)$ is that $g^*_\pi(\cdot, \cdot; \eta^\pi, \tilde{Q}^\pi) = 0$.

We now formally introduce the coupled estimator for $(\eta^\pi, \tilde{Q}^\pi)$. In what follows, for ease of notation we use $\hat{Q}^\pi_n$ to denote the estimator of the shifted relative value function, $\tilde{Q}^\pi$. Recall that the Bellman operator, $\mathcal{T}_\pi(s, a; Q)$ in (5) is conditional expectation of $R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1}, a')$ given $S_t = s, A_t = a$. The estimator $(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)$ is found by simultaneously minimizing

$$\begin{aligned}\hat{g}_{n,\pi}(\cdot, \cdot; \eta, Q) = \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{P}_n\Big\{\frac{1}{T}\sum_{t=1}^{T}\big(R_{t+1} + &\sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1}, a'),\\ &- \eta - Q(S_t, A_t) - g(S_t, A_t)\big)^2\Big\} + \mu_n J_2^2(g),\end{aligned} \tag{8}$$

and

$$(\hat{\eta}^\pi_n, \hat{Q}^\pi_n) = \operatorname*{argmin}_{(\eta, Q) \in \mathbb{R} \times \mathcal{Q}} \mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T}\hat{g}^2_{n,\pi}(S_t, A_t; \eta, Q)\Big] + \lambda_n J_1^2(Q), \tag{9}$$

where $J_1 : \mathcal{Q} \to \mathbb{R}^+$ and $J_2 : \mathcal{G} \to \mathbb{R}^+$ are two regularizers, and λ_n and μ_n are tuning parameters. We can see that for every (η, Q), $\hat{g}_{n,\pi}(\cdot, \cdot; \eta, Q)$ is a penalized estimator of the projected Bellman error $g^*_\pi(\cdot,\cdot;\eta,Q)$ in (7). On the other hand, the objective function in (9) is a plug-in version of (6) where we replace the Bellman error by $\hat{g}_{n,\pi}(\cdot,\cdot;\eta,Q)$. The penalty term $\lambda_n J_1^2(Q)$ is used to balance between the model fitting, i.e., the squared estimated Bellman error and the complexity of the relative value function, measured by $J_1(Q)$. Similarly, $\mu_n J_2^2(g)$ is used to control the overfitting in estimating the Bellman error when the function class, $\mathcal{G}$, is complex. In the case where the function space is k-th order Sobolev space, the regularizer is typically defined by the k-th order derivative to capture the smoothness of function. In the case where the function space is Reproducing Kernel Hilbert Space (RKHS), the regularizer is the endowed norm. In Supplement D, we provide a closed-form solution of the estimator when both $\mathcal{Q}$ and $\mathcal{G}$ are RKHSs.

So far we have focused on evaluating a single target policy. In practice, one might want to compare the target policy to some reference policy or contrast multiple target policies of interest. Suppose we are interested in K different target policies, $\{\pi_j\}_{j=1}^K$. The above procedure (9) can be applied to estimate $\{\eta^{\pi_j}\}_{j=1}^K$. In the next section, we will provide the result of the joint asymptotic distribution of the estimated average rewards, $\{\hat{\eta}_n^{\pi_j}\}_{j=1}^K$. This can be used, for example, to construct the confidence interval of the difference of the average rewards between two policies.

5 Theoretical Results

In this section, we first derive the global rate of convergence for $(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)$ in (9) and derive the asymptotic distribution of $\hat{\eta}_n^\pi$ for a single policies. We then extend to the case of multiple policies. For any state-action function $f(s,a,s')$ and distribution ν on $\mathcal{S}\times\mathcal{A}$, denote the $L_2(\nu)$ norm by $\|f\|_\nu^2 = \int f^2(s,a)d\nu(s,a)$. If the norm does not have a subscript, then the expectation is with respect to the average state-action distribution in the trajectory $\mathcal{D}$, that is, $\|f\|^2 = \mathbb{E}[(1/T)\sum_{t=1}^T f^2(S_t,A_t)]$.

We first state two standard assumptions used in the non-parametric regression literature (Györfi et al. 2006). Recall that the shifted value function is defined as $\tilde{Q}^\pi = Q^\pi - Q^\pi(s^*,a^*)$.

Assumption 2. *The reward is uniformly bounded:* $|R_{t+1}| \leq R_{\max} < \infty$ *for all* $t \geq 1$. *The shifted relative value function is bounded:* $|\tilde{Q}^{\pi}(s, a)| \leq Q_{\max}$ *for all* $s \in \mathcal{S}$.

Assumption 3. *The function class,* $\mathcal{Q}$, *satisfies (i)* $Q(s^*, a^*) = 0$ *and* $\|Q\|_\infty \leq Q_{\max}$ *for all* $Q \in \mathcal{Q}$ *and (ii)* $\tilde{Q}^{\pi} \in \mathcal{Q}$.

The assumption of a bounded reward is mainly to simplify the proof and can be relaxed to the sub-Gaussian case, that is, the error $R_{t+1} - r(S_t, A_t)$ is sub-Gaussian for all $t \leq T$. The boundedness assumption of the relative value function can be ensured by assuming certain smoothness assumptions on the transition distribution (Ortner & Ryabko 2012) or assuming geometric convergence to the stationary distribution (Hernández-Lerma & Lasserre 1999). The boundedness assumption for the function class, $\mathcal{Q}$, is used to simplify the proof; a truncation argument can be used to avoid this assumption.

Recall that in 7 we introduce a second function class, $\mathcal{G}$, to approximate the Bellman error. We make the following assumption about the complexity of $\mathcal{G}$ to ensure that we can consistently estimate $(\eta^{\pi}, \tilde{Q}^{\pi})$. Recall $g^*_{\pi}(\cdot, \cdot; \eta, Q)$ is the projected Bellman error, given in 7.

Assumption 4. *The function class,* $\mathcal{G}$, *satisfies (i)* $0 \in \mathcal{G}$ *and* $\|g\|_\infty \leq G_{\max}$ *for all* $g \in \mathcal{G}$ *and (ii)* $\kappa = \inf\big\{\|g^*_{\pi}(\cdot, \cdot; \eta, Q)\| : \|\mathcal{T}_{\pi}(\cdot, \cdot; Q) - \eta - Q(\cdot, \cdot)\| = 1, \eta \in \mathbb{R}, Q \in \mathcal{Q}\big\} > 0$.

The value of κ measures how well the function class, $\mathcal{G}$, approximates the Bellman error for all (η, Q) in which $\eta \in \mathbb{R}$ and $Q \in \mathcal{Q}$. The condition of a strictly positive κ ensures the estimator (9) based on minimizing the projected Bellman error onto the space, $\mathcal{G}$, is able to identify the true parameters $(\eta^{\pi}, \tilde{Q}^{\pi})$. This is similar to the eigenvalue condition (Assumption 5) in Luckett et al. (2019) but they are essentially using the same function class for $\mathcal{Q}$ and $\mathcal{G}$. Note that unlike in Farahmand et al. (2016), here we do not assume $\mathcal{T}_{\pi}(\cdot, \cdot; Q) - \eta - Q(\cdot, \cdot) \in \mathcal{G}$ for every (η, Q). If this were the case, then we would have that $g^*_{\pi}(\cdot, \cdot; \eta, Q) = \mathcal{T}_{\pi}(\cdot, \cdot; Q) - \eta - Q(\cdot, \cdot)$ and thus $\kappa = 1$.

Below we make assumptions on the complexity of the function classes, $\mathcal{Q}$ and $\mathcal{G}$. These assumptions are satisfied for common function classes, for example RKHS and Sobolev spaces (Van de Geer 2000, Zhao et al. 2016, Steinwart & Christmann 2008, Györfi et al. 2006). We denote by $N(\epsilon, \mathcal{F}, \|\cdot\|)$ the ϵ-covering number of a set of functions $\mathcal{F}$ with respect to $\|\cdot\|$.

Assumption 5. *(i) The regularization functional* J_1 *and* J_2 *are pseudo norms and induced by the inner products* $J_1(\cdot,\cdot)$ *and* $J_2(\cdot,\cdot)$*, respectively. For all* $(\eta, Q) \in \mathbb{R} \times Q$*, there exists constants* C_1, C_2 *such that* $J_2(g_\pi^*(\cdot,\cdot;\eta,Q)) \le C_1 + C_2 J_1(Q)$.

(ii) Let $\mathcal{Q}_M = \{c + Q : |c| \le R_{\max}, Q \in \mathcal{Q}, J_1(Q) \le M\}$ *and* $\mathcal{G}_M = \{g : g \in \mathcal{G}, J_2(g) \le M\}$. *There exists constants* C_3 *and* $\alpha \in (0,1)$ *such that for any* $\epsilon, M > 0$,

$$\max\Big\{\log N(\epsilon, \mathcal{G}_M, \|\cdot\|_\infty), \log N(\epsilon, \mathcal{Q}_M, \|\cdot\|_\infty)\Big\} \le C_3\left(\frac{M}{\epsilon}\right)^{2\alpha}.$$

The upper bound on $J_2(g_\pi^*(\cdot,\cdot;\eta,Q))$ is realistic when the transition kernel is sufficiently smooth; see Farahmand et al. (2016) for an example of MDP satisfying this condition. We use a common $\alpha \in (0,1)$ for both $\mathcal{Q}$ and $\mathcal{G}$ in (ii) to simply the proof.

Now we are ready to state the theorem about the convergence rate for $(\hat\eta_n^\pi, \hat Q_n^\pi)$ in terms of the Bellman error.

Theorem 1 (Global Convergence Rate)**.** *Let* $(\hat\eta_n^\pi, \hat Q_n^\pi)$ *be the estimator defined in (9). Suppose Assumptions 1-5 hold and the tuning parameters* (λ_n, μ_n) *satisfy* $\tau^{-1} n^{-\frac{1}{1+\alpha}} \le \mu_n \le \tau\lambda_n$ *for some constant* $\tau > 0$*. Then the followings hold with probability at least* $1 - \delta$,

$$\begin{aligned}&\big\|\mathcal{T}_\pi(\cdot,\cdot;\hat Q_n^\pi) - \hat\eta_n^\pi - \hat Q_n^\pi(\cdot,\cdot)\big\|^2 \lesssim \kappa^{-2}\lambda_n(1 + J_1^2(\tilde Q^\pi))(1 + \log(1/\delta)),\\ &J_1(\hat Q_n^\pi) \lesssim 1 + \log(1/\delta) + J_1^2(\tilde Q^\pi),\end{aligned}$$

where the leading constant only depends only on $(\tau, R_{\max}, Q_{\max}, G_{\max}, C_1, C_2, C_3, \alpha)$.

In Lemma B.3 in Supplement B, we show that up to a constant, $|\hat\eta_n^\pi - \eta^\pi| \lesssim \|\mathcal{T}_\pi(\cdot,\cdot;\hat Q_n^\pi) - \hat\eta_n^\pi - \hat Q_n^\pi(\cdot,\cdot)\|^2$ and thus when $\lambda_n = o_P(1)$ we see that $\hat\eta_n^\pi$ is a consistent estimator of η^π. When the tuning parameters are chosen such that $\lambda_n \asymp \mu_n$ and $\lambda_n \asymp n^{-1/(1+\alpha)}$, the Bellman error at $(\hat\eta_n^\pi, \hat Q_n^\pi)$ has the optimal rate of convergence, i.e., $\|\mathcal{T}_\pi(\cdot,\cdot;\hat\eta_n^\pi, \hat Q_n^\pi) - \hat\eta_n^\pi - \hat Q_n^\pi(\cdot,\cdot)\|^2 = O_P(n^{-1/(1+\alpha)})$. The proof of Theorem 1 is provided in Supplement B.

In what follows, we focus on the asymptotic distribution of the estimator of average reward. In particular, we show that under certain conditions the estimated average reward, $\hat\eta_n^\pi$, is $\sqrt{n}$-consistent and asymptotically normal. Below we introduce some additional notations. Let $d^\pi(s,a) = \pi(a|s)d^\pi(s)$ be density of the stationary distribution of state-action under the target policy, π. For each $t \geq 1$, denote by $d_t(s,a)$ the density of the state-action pair in the trajectory, $\mathcal{D}$, under the behavior policy. Let $\bar{d}_T(s,a)$ be the average density over T decision times. Motivated by the least favorable direction in partial linear regression problems (Van de Geer 2000, Zhao et al. 2016), we define the direction $e^\pi(s,a)$ by

$$e^\pi(s,a) = \frac{d^\pi(s,a)/\bar{d}_T(s,a)}{\int (d^\pi(\tilde{s},\tilde{a})/\bar{d}_T(\tilde{s},\tilde{a}))d^\pi(\tilde{s},\tilde{a})d\tilde{s}d\tilde{a}}. \tag{10}$$

The above-defined direction e^π is used to control the bias $(\hat{\eta}_n^\pi - \eta^\pi)$ caused by the penalization on the non-parametric relative value function in the estimator (9). This is akin to partial linear regression problem, $Y = f(Z) + X^\top \beta + \epsilon$, in which the analog of $e^\pi(s,a)$ is the residual $x - \mathbb{E}[X|Z=z]$; see Donald et al. (1994), Van de Geer (2000) for an analysis of bias in the regression problem. In our setting, the direction $e^\pi(s,a)$ satisfies the following orthogonality property: for any state-action function Q,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Big[\big(Q(S_t,A_t) - \sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1},a')\big)e^\pi(S_t,A_t)\Big]\right] = 0, \tag{11}$$

by noting that $\int Q(s,a)d^\pi(s,a)dsda = \int \sum_{a'}\pi(a'|s')Q(s',a')P(s'|s,a)d^\pi(s,a)dsdads'$. Note that e^π is a scaled version of $d^\pi(s,a)/\bar{d}_T(s,a)$, the ratio between the stationary distribution under target policy, π, and the distribution in the trajectory, $\mathcal{D}$. The denominator is the expectation of the ratio under the stationary distribution and it is greater than 1:

$$\begin{aligned}\int \frac{d^\pi(s,a)}{\bar{d}_T(s,a)}d^\pi(s,a)dsda &= \int \left(\frac{d^\pi(s,a)}{\bar{d}_T(s,a)}\right)^2 \bar{d}_T(s,a)dsda \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\frac{d^\pi(S_t,A_t)}{\bar{d}_T(S_t,A_t)}\right)^2\right] = \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{d^\pi(S_t,A_t)}{\bar{d}_T(S_t,A_t)}\right) + 1,\end{aligned}$$

where in the last equality the variance is with respect to the distribution of state-action pairs collected from the behavior policy and $\mathbb{E}[(1/T)\sum_{t=1}^{T} d^\pi(S_t,A_t)/\bar{d}_T(S_t,A_t)] = 1$.

Next define $q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=1}^{\infty}(1 - e^\pi(S_t,A_t)) \,|\, S_1 = s, A_1 = a\right]$. Note q^π has a similar structure to that of the relative value function in (3) with the reward at time t being $(1 - e^\pi(S_t,A_t))$ and the average reward being 0 ($\int e^\pi(s,a)d^\pi(s,a)dsda = 1$ by definition). Similarly, $q^\pi(s,a)$ satisfies a Bellman-like equation:

$$q^\pi(s,a) = 1 - e^\pi(s,a) + \mathbb{E}\Big[\sum_{a'}\pi(S_{t+1},a')q^\pi(S_{t+1},a') \,|\, S_t = s, A_t = a\Big]. \tag{12}$$

We make the following smoothness assumption about e^π and q^π, akin to the assumptions used in partially linear regression literature (Van de Geer 2000, Zhao et al. 2016).

Assumption 6. *The shifted function $\tilde{q}^\pi = q^\pi - q^\pi(s^*, a^*) \in \mathcal{Q}$ and $e^\pi \in \mathcal{G}$.*

Recall that in Assumption 3 we restrict $Q(s^*, a^*) = 0$ for all $Q \in \mathcal{Q}$. Thus we consider the shifted function $\tilde{q}^\pi$ in the assumption above. The condition, $\tilde{q}^\pi \in \mathcal{Q}$, is imposed to ensure the $\sqrt{n}$ rate of convergence and asymptotic normality of $\hat{\eta}_n^\pi$. On the other hand, unlike in the regression setting we further assume e^π is smooth enough (i.e., $e^\pi \in \mathcal{G}$) due to not knowing the conditional expectation $\mathbb{E}_\pi\left[\tilde{q}^\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]$. Note that this assumption (i.e., $e^\pi \in \mathcal{G}$) would hold automatically if we assume the much stronger condition that $\mathcal{T}_\pi(\cdot, \cdot; Q) - \eta - Q(\cdot, \cdot) \in \mathcal{G}$ for every η and $Q \in \mathcal{Q}$ (i.e., $\kappa = 1$ in Assumption 4).

Our last assumption is a contraction-type property. This is imposed to control the variance of the remainder term caused by estimation of the nuisance function Q^π.

Assumption 7. *Define by $\mathcal{P}^\pi f(s, a) = \mathbb{E}_\pi[f(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$ the conditional expectation operator. Let $\mu^\pi(f) = \int f(s, a) d^\pi(s, a) ds da$ be the expectation of a state-action function, f, under stationary distribution induced by π. There exists constants, $C_4 > 0$ and $0 \leq \beta < 1$, such that for $f \in L_2$ and $t \geq 1$,*

$$\|(\mathcal{P}^\pi)^t(f) - \mu^\pi(f)\| \leq C_4 \|f\| \beta^t. \tag{13}$$

The parameter, β, in Assumption 7 is akin to the discount factor, γ, in the discounted reward setting. Intuitively, this is related to the "mixing rate" of the Markov chain induced by the target policy π. A similar assumption was imposed in Van Roy (1998) (Assumption 7.2 on p. 99). Now we are ready to present our main result, the asymptotic normality of the estimated average reward, $\hat{\eta}_n^\pi$

Theorem 2 (Asymptotic Dsitribution)**.** *Suppose the conditions in Theorem 1 hold. In addition, suppose Assumption 6 and 7 hold and $\lambda_n = a_n n^{-1/2}$ with $a_n \to 0$. The estimator $\hat{\eta}_n^\pi$ in (8) is $\sqrt{n}$-consistent and asymptotically normal: $\sqrt{n}(\hat{\eta}_n^\pi - \eta^\pi) \Rightarrow \boldsymbol{N}(0, \sigma^2)$, where*

$$\sigma^2 = \mathrm{Var}\left(\frac{1}{T}\sum_{t=1}^{T} \frac{d^\pi(S_t, A_t)}{\bar{d}_T(S_t, A_t)} \big(R_{t+1} + \sum_{a'} Q^\pi(S_{t+1}, a') - \eta^\pi - Q^\pi(S_t, A_t)\big)\right).$$

From Theorem 2, we see the variance in estimating the average reward parameter η^π depends on the ratio between the stationary distribution of state-action pair induced by the target policy and the average state-action distribution in the training data. When these two distributions are close, the variance of estimating the average reward gets small. In the special case where the target policy is same as the behavior policy (i.e., on-policy evaluation) and the states in the training data follows the stationary distribution (e.g., when the length of trajectory is sufficiently large), one should expect to see the smallest variance. A similar definition of ratio is used in Liu et al. (2018) to build an estimator of η^π. Although here we focus only on the asymptotic property of $\hat{\eta}_n^\pi$ for large n (n is the number of i.i.d. trajectories), one can see that increasing length of trajectory, T, reduces the variance, as inside of the variance, σ^2, is an average over T decision time points.

Now we present the result for evaluating a class of policies, $\Pi = \{\pi_1, \ldots, \pi_K\}$. Denote by $\hat{\eta}_n^{\pi_j}$ the estimated average reward of the policy, π_j, using (9).

Corollary 1 (Multiple Policies)**.** *Suppose the conditions in Theorem 1 and 2 hold for each time t and policy $\pi \in \Pi$. Let $\epsilon_t^\pi = \frac{d^\pi(S_t,A_t)}{\bar{d}_T(S_t,A_t)}(R_{t+1}+\sum_{a'}\pi(a'|S_{t+1})Q^\pi(S_{t+1},a')-\eta^\pi-Q^\pi(S_t,A_t))$ for each $\pi \in \Pi$. The estimated average rewards jointly converge in distribution to a multivariate Gaussian distribution:*

$$\begin{pmatrix} \sqrt{n}(\hat{\eta}_n^{\pi_1} - \eta^{\pi_1}) \\ \vdots \\ \sqrt{n}(\hat{\eta}_n^{\pi_K} - \eta^{\pi_K}) \end{pmatrix} \Rightarrow \boldsymbol{MVN}(0, \Sigma),$$

where the covariance matrix $\Sigma = (\mathbb{E}[(1/T^2)\sum_{t=1}^T \epsilon_t^{\pi_i}\epsilon_t^{\pi_j}])_{i,j=1}^K$.

To construct inference, we need to estimate the asymptotic variance Σ. For each π, we denote the plug-in estimation of ϵ_t^π (defined in Corollary 1) by $\hat{\epsilon}_t^\pi$ in which we plug in $(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)$ and an estimator of ratio, $d^\pi(s,a)/\bar{d}_T(s,a)$. We then estimate the asymptotic variance, Σ, by $\hat{\Sigma}_n = (\mathbb{P}_n[(\frac{1}{T}\sum_{t=1}^T \hat{\epsilon}_t^{\pi_i})(\frac{1}{T}\sum_{t=1}^T \hat{\epsilon}_t^{\pi_j})])_{i,j=1}^K$. We can estimate the ratio as follows. Taking expectation of both sides of (10) gives $d^\pi(s,a)/\bar{d}_T(s,a) = e^\pi(s,a)/\mathbb{E}[(1/T)\sum_{t=1}^T e^\pi(S_t,A_t)]$. Given the estimator $\hat{e}_n^\pi(\cdot,\cdot)$ of $e^\pi(\cdot,\cdot)$, we estimate the ratio by $\hat{e}_n^\pi(s,a)/\mathbb{P}_n[(1/T)\sum_{t=1}^T \hat{e}_n^\pi(S_t,A_t)]$.

Motivated by the orthogonality (11) and the expression (12), we construct the estimator of $e^\pi(\cdot,\cdot)$ by $\hat{e}^\pi_n(\cdot,\cdot) = \tilde{g}_{n,\pi}(\cdot,\cdot;\hat{q}^\pi_n)$, where $\hat{q}^\pi_n(\cdot,\cdot) = \operatorname{argmin}_{q\in\mathcal{Q}} \mathbb{P}_n[(1/T)\sum_{t=1}^T \tilde{g}^2_{n,\pi}(S_t,A_t;q)] + \tilde{\lambda}_n J_1^2(q)$ and $\tilde{g}_{n,\pi}(\cdot,\cdot;q) = \operatorname{argmin}_{g\in\mathcal{G}} \mathbb{P}_n\{(1/T)\sum_{t=1}^T [1-q(S_t,A_t)+\sum_{a'}\pi(a'|S_{t+1})q(S_{t+1},a')-g(S_t,A_t)]^2\} + \tilde{\mu}_n J_2^2(g)$ for each $q\in\mathcal{Q}$. See Supplement C for additional details about the estimator $\hat{e}^\pi_n$. Here $(\tilde{\lambda}_n,\tilde{\mu}_n)$ are some tuning parameters. In Supplement D, we provide a closed-form solution for the estimator of the asymptotic variance when $\mathcal{Q}$ and $\mathcal{G}$ are RKHSs

6 Simulation

In this section, we conduct a simulation study to evaluate the performance of the method. The generative model is given as follows. We follow the state generative model in Luckett et al. (2019). Specifically, the state $S_t = (S_{t,1}, S_{t,2})$ is two-dimensional vector and the action $A_t \in \{0,1\}$ is binary. Given the current state S_t and action A_t, the next state $S_{t+1} = (S_{t+1,1}, S_{t+1,2})$ are generated by $S_{t+1,1} = (3/4)(2A_t-1)S_{t,1} + (1/4)S_{t,1}S_{t,2} + N(0,0.5^2)$ and $S_{t+1,2} = (3/4)(1-2A_t)S_{t,2} + (1/4)S_{t,1}S_{t,2} + N(0,0.5^2)$. Note that receiving a treatment $(A_t = 1)$ increases the value of $S_{t,1}$ while decreases $S_{t,2}$. The reward is generated by $R_{t+1} = S_{t+1,1} + (1/2)S_{t+1,2} + (1/4)(2A_t-1)$. For each trajectory in the training data, we initiate the state variables as independent standard normal random variables and the behavior policy is to choose $A_t = 1$ with a fixed probability 0.5. We evaluate and compare two natural policies: the "always treat" policy, $\pi_1(a|s) = 1$, and "no treatment" policy, $\pi_2(a|s) = 0$.

In the implementation, we use RKHS with the radial basis function (RBF) kernel to construct the functional classes, $\mathcal{Q}$ and $\mathcal{G}$; the details of how to modify an arbitrary RKHS such that the value at (s^*,a^*) is zero can be found in Supplement D. The bandwidth parameter in the RBF kernel is chosen by the median heuristic. Recall that the estimator (9) involves two tuning parameters, (λ_n,μ_n). Following the idea in Farahmand & Szepesvári (2011), we select these tuning parameters as follows. We first split the dataset into a training set $\mathcal{D}^{\text{trn}}$ and a validation set, $\mathcal{D}^{\text{val}}$. For each candidate value of tuning parameters, (λ,μ), the training set $\mathcal{D}^{\text{trn}}$ is used to form the estimator by (8) and (9). Denote the corresponding estimator by $(\hat{\eta}^\pi(\lambda,\mu), \hat{Q}^\pi(\cdot,\cdot;\lambda,\mu))$. Then the temporal difference

(TD) error, $R + \sum_{a'} \hat{Q}^{\pi}(S', a'; \lambda, \mu) - \hat{\eta}^{\pi}(\lambda, \mu) - \hat{Q}^{\pi}(S, A; \lambda, \mu)$, is calculated for each transition sample (S, A, S', R) in the validation set, $\mathcal{D}^{\text{val}}$. Recall that the Bellman error is 0 at (η^{π}, Q^{π}). We use the validation set $\mathcal{D}^{\text{val}}$ to fit a model for the Bellman error with respect to $(\hat{\eta}^{\pi}(\lambda, \mu), \hat{Q}^{\pi}(\cdot, \cdot; \lambda, \mu))$ and denote the estimated Bellman error by $\hat{f}(\cdot, \cdot; \lambda, \mu)$. Note that this step is essentially a regression problem (i.e., the dependent variable is the TD error and independent variables are the current state and action). Finally, we choose the tuning parameters (λ, μ) that minimizes the squared estimated Bellman error, $\sum_{(S,A) \in \mathcal{D}^{\text{val}}} \hat{f}^2(S, A; \lambda, \mu)$. The final estimator of η^{π} is then calculated with the optimal tuning parameters using the entire dataset. In the simulation, we use (1/2) of data for the training set and (1/2) for the validation set and we use Gaussian Process regression to estimate the Bellman error in the validation step.

We consider different scenarios of the number of trajectories, $n \in \{25, 40\}$, and the length of each trajectory, $T \in \{25, 50, 75\}$. Under each scenario, we generate 500 simulated dataset and for each dataset we construct the 95% confidence interval (CI) of $\eta^{\pi_1}, \eta^{\pi_2}$ and $\eta^{\pi_1} - \eta^{\pi_2}$. The coverage probability of each CI is calculated over 500 repetitions. The simulation result is reported in Table 1. When the number of trajectories is small (i.e., $n = 25$), the simulated coverage probability is slightly smaller than the claimed value, 0.95, especially when the length of trajectory, T, is small . It can be seen that the coverage probability slightly improves when T increases. When $n = 40$, the coverage probability becomes closer to 0.95 as desired. Overall, the simulation result demonstrates the validity of the inference and the selection procedure for the tuning parameters. It also suggests that it is necessary to perform a small-sample correction when both n and T are small. This is left for future work.

7 Case Study: HeartSteps

We apply the method to the data collected in the first study in HeartSteps (Klasnja et al. 2015, Liao et al. 2016, Klasnja et al. 2018). Below we refer to this study by HS1 for simplicity. HS1 is a 42-day MRT with 44 healthy sedentary adults. We focused on the activity message intervention component. There were five user-specified times in a day which were

	n	T	Coverage Prob.	MAD	n	T	Coverage Prob.	MAD
	25	25	0.926	0.0702	40	25	0.944	0.0546
Case 1	25	50	0.930	0.0535	40	50	0.944	0.0427
	25	75	0.938	0.0438	40	75	0.948	0.0346
	25	25	0.934	0.0368	40	25	0.928	0.0313
Case 2	25	50	0.946	0.0261	40	50	0.940	0.0224
	25	75	0.922	0.0222	40	75	0.942	0.0185
	25	25	0.932	0.0761	40	25	0.946	0.0612
Case 3	25	50	0.928	0.0598	40	50	0.948	0.0461
	25	75	0.932	0.0480	40	75	0.948	0.0388

Table 1: Coverage probability of the 95% CI and MAD (mean absolute deviation) over 500 repetitions. Case 1: policy evaluation of π_1. Case 2: policy evaluation of π_2. Case 3: policy comparison between π_1 and π_2.

roughly separated by 2.5 hours and corresponded to the user's morning commute, mid-day, mid-afternoon, evening commute, and post-dinner times. At each decision time, a contextually-tailored activity suggestion message was sent with a fixed probability 0.6 only if the participants were considered to be available for treatment. For example, the participants were considered unavailable when they were currently physically active (e.g., walking or running) or driving. The activity messages were intended to motivate walking. Each participant wore a Jawbone wrist tracker and the minute-level step count data was recorded.

We constructed the state based on the participant's step count data (e.g., the 30-min step count prior to the decision time and the total step count from yesterday), location, temperature and number of notifications received over the last 7 days. We also included in the state the time slot index in the day (1 to 5) and the indicator measuring how the step count varies at the current time slot over the last 7 days. The reward was formed by the log transformation of the total step count collected 30 min after the decision time. The log transformation was performed as the step count data is highly noisy and positively skewed; see Klasnja et al. (2018). The step count data might be missing because the Jawbone tracker recorded data only when there were steps occurred. We used the same imputation procedure as in Klasnja et al. (2018). The state related to the step count were constructed based on the imputed step counts. The variable in the state was chosen to be predictive of the reward. In particular, each variable was selected, at the significance level of 0.05, based on a marginal Generalized Estimating Equation (GEE) analysis. In the analysis, we excluded seven participants' data as in the primary analysis in Klasnja et al. (2018) (three due to technical issues and four due to early dropout). In addition, from the 37 participants' data we excluded the decision times when participants were traveling abroad or experiencing technical issues or when the reward (post 30-min step count) was considered as missing; see Klasnja et al. (2018).

We evaluated three target policies. The first policy, π_{nothing}, is "do nothing". The second policy, π_{always}, is the "always treat" policy. Recall that in HeartSteps the activity message can only be sent when the participant is available. So here the "always treat" policy refers to always send the message whenever the participant is available. The third policy, π_{location}, is based on the location. Specifically, we consider the policy that sends the activity message when the participant is at either home or work location and available. This policy is of interest because people at home or work are in a more structured environment and thus might be able to better respond to an activity suggestion message as compared with at other locations. Among the decision times that were available for treatment in the dataset, there were about 44% of them that occurred while the participants were at home or work location. Thus the policy, π_{location}, is different to "always treat" policy, π_{always}.

In the implementation, we used the RKHS with the radial basis function kernel to form the function classes, $\mathcal{Q}$ and $\mathcal{G}$. The tuning parameters were selected based on the procedure described in Section 6. The estimated average reward of the location-based policy, π_{location}, is 3.155 with the 95% CI, $[2.893, 3.417]$, which is better than the "do nothing" policy. Specifically, the estimated average reward of π_{nothing} is 2.962 and the 95% CI of the difference, $\eta^{\pi_{\text{location}}} - \eta^{\pi_{\text{nothing}}}$, is $[-0.016, 0.402]$. Translating back to the raw step count as in

Klasnja et al. (2018), the location-based policy is able to increase the average 30-min step count roughly by 22% ($\exp(3.16 - 2.96) - 1 = 1.22$), corresponding to 55 steps (the mean post-decision time step count is 248 across all decision times in the dataset). However if we compare the "always treat" policy ($\hat{\eta}^{\pi_{\text{always}}} = 3.127$, 95% CI $[2.840, 3.413]$) with the location-based policy, π_{location}, we see no indication that providing treatment only at home/work is better than always providing treatment (the 95% CI for $\eta^{\pi_{\text{location}}} - \eta^{\pi_{\text{always}}}$ is [-0.161, 0.217]). Recall that the sample size for this study is $n = 37$ thus this non-significant finding may be due to the small sample.

8 Discussion

In this work we developed a flexible method to conduct inference about the the long-term average outcomes for given target policies using data collected from a possibly different behavioral policy. We believe that this is an important first step towards developing data-based just-in-time adaptive interventions. Below we discuss some directions for future research.

In many MRT studies, the natural choice of the proximal outcome to assess the effectiveness of the intervention is binary. For example, in the Substance Abuse Research Assistance study (Rabbi et al. 2018), the proximal outcome was whether the user completed a daily survey. An interesting open question is how to extend the method to the binary reward setting, which would require carefully choosing the model to represent the value function and/or the loss functions used in estimating the Bellman error and solving the Bellman equations.

Non-stationarity occurs mainly because of the unobserved aspects of the current context (e.g.,the engagement and/or burden) in many mHealth applications. It will be interesting to generalize the current framework of average reward to incorporate the non-stationarity detected in the observed trajectory. Alternatively, one can consider evaluating the treatment policy in the indefinite horizon setting where there is an absorbing state (akin to the user disengaging from the mobile app) and thus we aim to conduct inference about the expected total rewards until the absorbing state is reached.

We focused on evaluating and contrasting multiple pre-specified treatment policies. An important next step is to extend the method to learn the optimal policy that would lead to the largest long-term average reward and to develop the inferential methods to assess the usefulness of certain variables in the policy.

Acknowledgment

This work was supported by National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health under award number R01AA23187, National Institute on Drug Abuse (NIDA) of the National Institutes of Health under award numbers P50DA039838 and R01DA039901, National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number U54EB020404, National Cancer Institute (NCI) of the National Institutes of Health under award number U01CA229437, and National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health under award number R01HL125440. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Antos, A., Szepesvári, C. & Munos, R. (2008), 'Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path', *Machine Learning* **71**(1), 89–129.

Bradtke, S. J., Barto, A. G. & Kaelbling, P. (1996), Linear least-squares algorithms for temporal difference learning, *in* 'Machine Learning', pp. 22–33.

Chakraborty, B. & Moodie, E. (2013), *Statistical methods for dynamic treatment regimes*, Springer.

Dempsey, W., Liao, P., Kumar, S. & Murphy, S. A. (2017), 'The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments', *arXiv preprint arXiv:1711.03587* .

Donald, S. G., Newey, W. K. et al. (1994), 'Series estimation of semilinear models', *Journal of Multivariate Analysis* **50**(1), 30–40.

Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C. & Mannor, S. (2016), 'Regularized policy iteration with nonparametric function spaces', *The Journal of Machine Learning Research* **17**(1), 4809–4874.

Farahmand, A.-m. & Szepesvári, C. (2011), 'Model selection in reinforcement learning', *Machine learning* **85**(3), 299–332.

Farajtabar, M., Chow, Y. & Ghavamzadeh, M. (2018), More robust doubly robust off-policy evaluation, *in* J. Dy & A. Krause, eds, 'Proceedings of the 35th International Conference on Machine Learning', Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, Stockholmsmässan, Stockholm Sweden, pp. 1447–1456.

Györfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2006), *A distribution-free theory of nonparametric regression*, Springer Science & Business Media.

Hernández-Lerma, O. & Lasserre, J. B. (1999), *Further topics on discrete-time Markov control processes*, Vol. 42, Springer.

Howard, R. A. (1960), 'Dynamic programming and markov processes.'.

Jiang, N. & Li, L. (2015), 'Doubly robust off-policy value evaluation for reinforcement learning', *arXiv preprint arXiv:1511.03722* .

Kallus, N. & Uehara, M. (2019), Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning, *in* 'Advances in Neural Information Processing Systems', pp. 3320–3329.

Kizakevich, P. N., Eckhoff, R., Weger, S., Weeks, A., Brown, J., Bryant, S., Bakalov, V., Zhang, Y., Lyden, J. & Spira, J. (2014), 'A personal health information toolkit for health intervention research', *Stud Health Technol Inform* **199**, 35–39.

Klasnja, P., Hekler, E., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A. & Murphy, S. (2015), 'Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions.', *Health Psychology* **34**(S), 1220.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B. & Murphy, S. A. (2018), 'Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of heartsteps', *Annals of Behavioral Medicine* .

Lee, J.-A., Choi, M., Lee, S. A. & Jiang, N. (2018), 'Effective behavioral intervention strategies using mobile health applications for chronic disease management: a systematic review', *BMC medical informatics and decision making* **18**(1), 12.

Liao, P., Dempsey, W., Sarker, H., Hossain, S. M., al'Absi, M., Klasnja, P. & Murphy, S. (2018), 'Just-in-time but not too much: Determining treatment timing in mobile health', *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* **2**(4), 179.

Liao, P., Klasjna, P., Tewari, A. & Murphy, S. (2016), 'Micro-randomized trials in mhealth', *Statistics in Medicine* **35**(12), 1944–71.

Liu, Q., Li, L., Tang, Z. & Zhou, D. (2018), Breaking the curse of horizon: Infinite-horizon off-policy estimation, *in* 'Advances in Neural Information Processing Systems', pp. 5356–5366.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. & Kosorok, M. R. (2019), 'Estimating dynamic treatment regimes in mobile health using v-learning', *Journal of the American Statistical Association* (just-accepted), 1–39.

Mahadevan, S. (1996), 'Average reward reinforcement learning: Foundations, algorithms, and empirical results', *Machine learning* **22**(1-3), 159–195.

Murphy, S. A., Deng, Y., Laber, E. B., Maei, H. R., Sutton, R. S. & Witkiewitz, K. (2016), 'A batch, off-policy, actor-critic algorithm for optimizing the average reward', *arXiv preprint arXiv:1607.05047* .

Murphy, S. A., van der Laan, M. J., Robins, J. M. & Group, C. P. P. R. (2001), 'Marginal mean models for dynamic regimes', *Journal of the American Statistical Association* **96**(456), 1410–1423.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A. & Murphy, S. A. (2018), 'Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support', *Annals of Behavioral Medicine* **52**(6), 446–462.

Ortner, R. & Ryabko, D. (2012), Online regret bounds for undiscounted continuous reinforcement learning, *in* 'Advances in Neural Information Processing Systems', pp. 1763–1771.

Puterman, M. L. (1994), 'Markov decision processes: Discrete stochastic dynamic programming'.

Rabbi, M., Kotov, M. P., Cunningham, R., Bonar, E. E., Nahum-Shani, I., Klasnja, P., Walton, M. & Murphy, S. (2018), 'Toward increasing engagement in substance use data collection: development of the substance abuse research assistant app and protocol for a microrandomized trial using adolescents and emerging adults', *JMIR research protocols* **7**(7), e166.

Steinwart, I. & Christmann, A. (2008), *Support vector machines*, Springer Science & Business Media.

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

Thomas, P. & Brunskill, E. (2016), Data-efficient off-policy policy evaluation for reinforcement learning, *in* 'International Conference on Machine Learning', pp. 2139–2148.

Van de Geer, S. (2000), *Empirical Processes in M-estimation*, Vol. 6, Cambridge university press.

Van Roy, B. (1998), Learning and value function approximation in complex decision processes, PhD thesis, Massachusetts Institute of Technology.

Zhao, T., Cheng, G. & Liu, H. (2016), 'A partially linear framework for massive heterogeneous data', *Annals of statistics* **44**(4), 1400.

Supplementary Material: Off-Policy Estimation of Long-Term Average Outcomes with Applications to Mobile Health

A Incorporating time-invariant variables

In mobile health application, often the baseline demographic information are collected, e.g., gender and occupation. However including time-invariant information into the states would violate Assumption 1. In the followings, we extend the results to incorporate the baseline information. Let $S_t = (Z, X_t)$ where $Z \in \mathcal{Z}$ is the time-invariant baseline information and $X_t \in \mathcal{X}$ is the time-varying variables. We consider a target policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. Note that the input of the target policy can depend on the baseline. We generalize Assumption 1 with respect to the target policy π as follows.

Assumption 1a. *We assume that for all $z \in \mathcal{Z}$, the induced markov chain on $\mathcal{X}$ by target policy π is irreducible and aperiodic for all $z \in \mathcal{Z}$. The density of stationary distribution exists for each $z \in \mathcal{Z}$ is denoted by $d^\pi(x|z)$.*

Under this assumption, the average reward defined in (1) can be shown to be a function of baseline variable z only, i.e., $\eta^\pi(s) = \eta^\pi(z)$ and the relative value function can be identified up to a function of z by solving Bellman equations: for all $s = (z, x)$ and a,

$$\mathbb{E}_\pi \left[R_{t+1} + Q(S_{t+1}, A_{t+1}) \,|\, S_t = s, A_t = a\right] = \eta(z) + Q(s, a). \tag{A.1}$$

Furthermore, we assume that the average reward follows a linear model.

Assumption 1b. *The average reward at the baseline $z \in \mathcal{Z}$ is $\eta^\pi(z) = f(z)^\top \beta^\pi$, where $f(z)$ is a feature vector of length p.*

Define the shifted relative value function $\tilde{Q}^\pi(s, a) = Q^\pi(s, a) - Q^\pi((z, x^*), a^*)$ for some reference time-varying state x^* and action a^*. Similar to the estimator presented in Section

4, we can estimate β^π and $\tilde{Q}^\pi$ by

$$(\hat{\beta}_n^\pi, \hat{Q}_n^\pi) = \operatorname*{argmin}_{(\beta,Q)\in\mathbb{R}^p\times\mathcal{Q}} \mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T}\hat{g}_n^2(S_t, A_t; \beta, Q)\Big] + \lambda_n J_1^2(Q), \tag{A.2}$$

where

$$\begin{aligned}\hat{g}_{\pi,n}(\cdot,\cdot;\beta,Q) = \operatorname*{argmin}_{g\in\mathcal{G}} \mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T}\big(R_{t+1} + \sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1}, a') \\ - f(Z)^\top\beta - Q(S_t, A_t) - g(S_t, A_t)\big)^2\Big] + \mu_n J_2^2(g). \end{aligned}\tag{A.3}$$

Replacing Assumption 1 in Theorem 1 by the Assumptions 1a and 1b above, the global convergence rate of the estimator $(\hat{\beta}_n^\pi, \hat{Q}_n^\pi)$ can be derived similarly and thus skipped here.

To obtain the asymptotic distribution of $\hat{\beta}_n^\pi$, we generalize the definition of e^π as follows. For any state $s = (z, x)$, define the vector

$$e^\pi(s,a) = (e_1^\pi(s,a), \ldots, e_p^\pi(s,a))^\top = \frac{d^\pi(x,a|z)/\bar{d}_T(x,a|z)}{\int(d^\pi(x,a|z)/\bar{d}_T(x,a|z))d^\pi(x,a|z)dxda} f(z) \in \mathbb{R}^p,$$

where $d^\pi(x,a|z) = \pi(a|x,z)d^\pi(x|z)$ is the density of stationary distribution of the time-varying states and action induced by the target policy π when starting at the baseline z and similarly, $\bar{d}_T(x,a|z)$ is the density of the average distribution of the time-varying states and action in the trajectory $\mathcal{D}$ if starting the baseline z. Under a similar set of conditions in Theorem 2, it can be shown that

$$\sqrt{n}(\hat{\beta}_n^\pi - \beta^\pi) \Rightarrow \mathbf{MVN}(0, (\boldsymbol{U}^\pi)^{-1}\boldsymbol{V}^\pi(\boldsymbol{U}^\pi)^{-\top}),$$

where $\boldsymbol{V}^\pi = \mathrm{Cov}[\sum_{t=1}^T(R_{t+1} + \sum_{a'}\pi(a'|S_{t+1})Q^\pi(S_{t+1}, a') - f(Z)^\top\beta^\pi - Q^\pi(S_t, A_t))e^\pi(S_t, A_t)]$ and $\boldsymbol{U}^\pi = \mathbb{E}[\sum_{t=1}^T e^\pi(S_t, A_t)f(Z)^\top]$.

Similar to the estimate of asymptotic variance in Section 5, we can form a sandwich estimator, $(\hat{\boldsymbol{U}}_n^\pi)^{-1}\hat{\boldsymbol{V}}_n^\pi(\hat{\boldsymbol{U}}_n^\pi)^{-\top}$ by plugging in the estimated value of $\hat{\beta}_n^\pi, \hat{Q}_n^\pi$ and $\hat{e}_n^\pi$. More specifically, we first estimate e^π by $\hat{e}_n^\pi = (\hat{g}_{n,1}(\cdot,\cdot;\hat{q}_{n,1}), \ldots, \hat{g}_{n,p}(\cdot,\cdot;\hat{q}_{n,p}))^\top$, where

$$\hat{q}_{n,k}(\cdot,\cdot) = \operatorname*{argmin}_{q\in\mathcal{Q}} \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\hat{g}_{n,k}^2(S_t, A_t; q)\big] + \tilde{\lambda}_n J_1^2(q),$$

and for each $k = 1, \dots, p$,

$$\hat{g}_{n,k}(\cdot,\cdot;q) = \operatorname*{argmin}_{g\in\mathcal{G}} \mathbb{P}_n[\frac{1}{T}\sum_{t=1}^{T}(f_k(Z) - q(S_t, A_t) + \sum_{a'} \pi(a'|S_{t+1})q(S_{t+1}, a') - g(S_t, A_t))^2] + \tilde{\mu}_n J_2^2(g).$$

Then $\hat{\boldsymbol{U}}_n^\pi = \mathbb{P}_n[\sum_{t=1}^T \hat{e}_n^\pi(S_t, A_t) f(Z)^\top]$, $\hat{\boldsymbol{V}}_n^\pi = \mathbb{P}_n[\big(\sum_{t=1}^T \hat{\delta}_t^\pi \hat{e}_n^\pi(S_t, A_t)\big)\big(\sum_{t=1}^T \hat{\delta}_t^\pi \hat{e}_n^\pi(S_t, A_t)\big)^\top]$ where $\hat{\delta}_t^\pi = R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})\hat{Q}_n^\pi(S_{t+1}, a') - f(Z)^\top \hat{\beta}_n^\pi - \hat{Q}_n^\pi(S_t, A_t)$ is the plug-in TD error. In Supplement D, we provide closed-form formula of $(\hat{\beta}_n^\pi, \hat{Q}_n^\pi, \hat{e}_n^\pi)$ when using RKHS.

B Technical proofs

The proof of global rate of convergence, Theorem 1, borrows the proof techniques developed in Farahmand et al. (2016) that controls the additional randomness in the objective function in (9) due to the estimated Bellman error operator $\hat{g}_{n,\pi}$ (5). The main differences are (1) we allow the approximation error when using the function space $\mathcal{G}$ to approximate Bellman error and (2) we have multiple trajectories instead of i.i.d. samples of transition. Due to (2), Theorem 16 in Farahmand et al. (2016) is not applicable as it requires conditioning on the current state-action pair. In proving the asymptotic distribution, Theorem 2, we follow a similar template in the classic partially linear regression setting (Van de Geer 2000, Zhao et al. 2016) to control the bias caused by the estimation error of nuisance functions Q^π.

In the rest of this section, for simplicity we use $\mathcal{E}_\pi(\cdot,\cdot;\eta,Q) = \mathcal{T}_\pi(\cdot,\cdot;Q) - \eta - Q(\cdot,\cdot)$ to denote the Bellman error at (η, Q) with respect to the target policy π. When there is no confusion, we omit the dot notation for brevity when referring the function of states and action, for example $\mathcal{E}_\pi(\eta, Q) = \mathcal{E}_\pi(\cdot,\cdot;\eta,Q)$ and $g_\pi^*(\eta, Q) = g_\pi^*(\cdot,\cdot;\eta,Q)$. We use $\delta_t^\pi(\eta, Q) = R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1}, a') - \eta - Q(S_t, A_t)$ to denote the Temporal Difference (TD) error at time t at (η, Q) with respect to the target policy π. For any function $f(s, a, s')$, denote the L_2 norm and the empirical 2-norm by $\|f\|^2 = \mathbb{E}[1/T\sum_{t=1}^T f^2(S_t, A_t, S_{t+1})]$ and $\|f\|_n^2 = \mathbb{P}_n[1/T\sum_{t=1}^T f^2(S_t, A_t, S_{t+1})]$, respectively.

B.1 Proof of Theorem 1

Proof of Theorem 1. By the definition of κ in Assumption 4,

$$\begin{aligned}
&\|\mathcal{T}_\pi(\cdot,\cdot;\hat{Q}_n^\pi) - \hat{\eta}_n^\pi - \hat{Q}_n^\pi(\cdot,\cdot)\|^2 \\
&\leq \frac{1}{\kappa^2}\|g_\pi^*(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2 \leq \frac{2}{\kappa^2}\left(\|g_\pi^*(\hat{\eta}_n^\pi,\hat{Q}_n^\pi) - \hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2 + \|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2\right). \qquad \text{(B.1)}
\end{aligned}$$

We first consider the second term in the parenthesis. Using Lemma B.2 together with the condition on the tuning parameters (i.e.,$\tau^{-1}n^{-\frac{1}{1+\alpha}} \leq \mu_n \leq \tau\lambda_n$), the following holds with probability at least $1-\delta$:

$$\begin{aligned}
\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2 &\lesssim n^{-\frac{1}{1+\alpha}} + \mu_n(1+J_1^2(\tilde{Q}^\pi)) + \lambda_n J_1^2(\tilde{Q}^\pi) + (n\lambda_n^\alpha)^{-1} \\
&\quad + (n\mu_n^\alpha)^{-1} + (n\mu_n^{\alpha/(1+\alpha)})^{-1} + \log(1/\delta)(n^{-1} + (n\mu_n^{\alpha/(1+\alpha)})^{-1}) \\
&\leq \tau^2\lambda_n + \tau\lambda_n(1+J_1^2(\tilde{Q}^\pi)) + \lambda_n J_1^2(\tilde{Q}^\pi) + \tau^{2+2\alpha}\lambda_n \\
&\quad + \tau^{2+\alpha}\lambda_n + \tau^{2+\alpha/(1+\alpha)}\lambda_n + \log(1/\delta)(\tau^2\lambda_n + \tau^{2+\alpha/(1+\alpha)}\lambda_n) \\
&\lesssim \lambda_n(1+J_1^2(\tilde{Q}^\pi))(1+\log(1/\delta)), \qquad \text{(B.2)}
\end{aligned}$$

where the leading constant depends only on $(R_{\max}, Q_{\max}, G_{\max}, C_1, C_2, C_3, \alpha, \tau)$ (same holds for all $\lesssim$ in the rest of the proof).

We now turn to the first term. From Lemma B.1, there exists some constant depending only on $(R_{\max}, Q_{\max}, G_{\max}, C_3, \alpha)$ such that with probability $1-\delta$,

$$\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi) - g_\pi^*(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2 \lesssim \mu_n(1+J_1^2(\hat{Q}_n^\pi) + J_2^2(g_\pi^*(\hat{\eta}_n^\pi,\hat{Q}_n^\pi))) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n}.$$

Using Assumption 5, this can be further bounded by

$$\begin{aligned}
\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi) - g_\pi^*(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2 &\lesssim \mu_n(1+J_1^2(\hat{Q}_n^\pi) + C_1^2 + C_2^2 J_2^2(\hat{Q}_n^\pi)) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n} \\
&\lesssim \mu_n(1+J_1^2(\hat{Q}_n^\pi)) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n}. \qquad \text{(B.3)}
\end{aligned}$$

To bound $J_1^2(\hat{Q}_n^\pi)$, the optimizing property of the estimators $(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)$ in (9) implies that

$$\lambda_n J_1^2(\hat{Q}_n^\pi) \leq \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^T \hat{g}_{\pi,n}^2(S_t,A_t;\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\right] + \lambda_n J_1^2(\hat{Q}_n^\pi)$$

$$
\begin{aligned}
&\leq \mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}\hat{g}^2_{\pi,n}(S_t,A_t;\eta^\pi,\tilde{Q}^\pi)\right]+\lambda_n J_1^2(\tilde{Q}^\pi)\\
&=\mathbb{P}_n\left[\frac{1}{T}\sum_{t=1}^{T}(\hat{g}_{\pi,n}(S_t,A_t;\eta^\pi,\tilde{Q}^\pi)-g^*_\pi(S_t,A_t;\eta^\pi,\tilde{Q}^\pi))^2\right]+\lambda_n J_1^2(\tilde{Q}^\pi)\\
&\lesssim \mu_n(1+J_1^2(\tilde{Q}^\pi))+\frac{1}{n\mu_n^\alpha}+\frac{1}{n}+\frac{\log(1/\delta)}{n}+\lambda_n J_1^2(\tilde{Q}^\pi),
\end{aligned}
$$

where we use $g^*_\pi(\eta^\pi, Q^\pi) = 0$ in the third line and the last inequality follows by Lemma B.1 and the fact that $J_2(g^*_\pi(\eta^\pi, Q^\pi)) = 0$. As a result, we have

$$
J_1^2(\hat{Q}^\pi_n)\lesssim J_1^2(\tilde{Q}^\pi)+\frac{\mu_n}{\lambda_n}(1+J_1^2(\tilde{Q}^\pi))+\frac{1}{n\lambda_n\mu_n^\alpha}+\frac{1}{n\lambda_n}+\frac{\log(1/\delta)}{n\lambda_n}.
$$

Combining with (B.3) gives

$$
\begin{aligned}
\|\hat{g}_{\pi,n}(\hat{\eta}^\pi_n,\hat{Q}^\pi_n)-g^*_\pi(\hat{\eta}^\pi_n,\hat{Q}^\pi_n)\|^2 &\lesssim \mu_n(1+J_1^2(\tilde{Q}^\pi))+\frac{1}{n\mu_n^\alpha}+\frac{1}{n}+\frac{\log(1/\delta)}{n}\\
&\quad+\mu_n\left(\frac{\mu_n}{\lambda_n}(1+J_1^2(\tilde{Q}^\pi))+\frac{1}{n\lambda_n\mu_n^\alpha}+\frac{1}{n\lambda_n}+\frac{\log(1/\delta)}{n\lambda_n}\right).
\end{aligned}
$$

Using the condition on the tuning parameters ($\tau^{-1}n^{-\frac{1}{1+\alpha}} \leq \mu_n \leq \tau\lambda_n$) and absorbing the lower order terms, the complexity of $\hat{Q}^\pi_n$ is then bounded by

$$
J_1^2(\hat{Q}^\pi_n)\lesssim J_1^2(\tilde{Q}^\pi)+\tau(1+J_1^2(\tilde{Q}^\pi))+\tau^{2+\alpha}+\tau^2 n^{-\frac{\alpha}{1+\alpha}}(1+\log(1/\delta))\lesssim 1+\log(1/\delta)+J_1^2(\tilde{Q}^\pi)
$$

On the other hand, we have

$$
\begin{aligned}
&\|\hat{g}_{\pi,n}(\hat{\eta}^\pi_n,\hat{Q}^\pi_n)-g^*_\pi(\hat{\eta}^\pi_n,\hat{Q}^\pi_n)\|^2\\
&\lesssim \tau\lambda_n(1+J_1^2(\tilde{Q}^\pi))+\tau^{2+\alpha}\lambda_n+\frac{\log(1/\delta)}{n}+\tau^2\lambda_n(1+J_1^2(\tilde{Q}^\pi))+\tau^{3+\alpha}\lambda_n+\frac{\tau}{n}+\frac{\tau\log(1/\delta)}{n}\\
&\lesssim \lambda_n(1+J_1^2(\tilde{Q}^\pi))(1+\log(1/\delta)). \qquad \text{(B.4)}
\end{aligned}
$$

Combing (B.1),(B.2), and (B.4), we see that with probability at least $1-2\delta$,

$$
\|\mathcal{T}_\pi(\cdot,\cdot;\hat{Q}^\pi_n)-\hat{\eta}^\pi_n-\hat{Q}^\pi_n(\cdot,\cdot)\|^2\lesssim \kappa^{-2}\lambda_n(1+J_1^2(\tilde{Q}^\pi))(1+\log(1/\delta)).
$$

□

Lemma B.1. *Let $g_\pi^*(\eta, Q)$ be the projected Bellman error operator defined in (7) and $\hat{g}_{\pi,n}(\eta, Q)$ be the estimated Bellman error in (5) with the tuning parameter μ_n. Suppose Assumptions 2, 3,4 and 5 hold. Then with probability at least $1-\delta$, the followings hold for all $\eta \in [-R_{\max}, R_{\max}]$, $Q \in \mathcal{Q}$:*

$$\|\hat{g}_{\pi,n}(\eta, Q) - g_\pi^*(\eta, Q)\|^2 \lesssim \mu_n(1 + J_1^2(Q) + J_2^2(g_\pi^*(\eta, Q))) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n}$$

$$J_2(\hat{g}_{\pi,n}(\eta, Q)) \lesssim 1 + J_1(Q) + J_2(g_\pi^*(\eta, Q)) + \sqrt{\frac{1}{n\mu_n^{\alpha+1}}} + \sqrt{\frac{1}{n\mu_n}} + \sqrt{\frac{\log(1/\delta)}{n\mu_n}}$$

$$\|\hat{g}_{\pi,n}(\eta, Q) - g_\pi^*(\eta, Q)\|_n^2 \lesssim \mu_n(1 + J_1^2(Q) + J_2^2(g_\pi^*(\eta, Q))) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n},$$

where the leading constant depends only on $R_{\max}, Q_{\max}, G_{\max}, C_3$ and α.

Proof of Lemma B.1. We start with decomposing the error:

$$\|\hat{g}_{\pi,n}(\eta, Q) - g_\pi^*(\eta, Q)\|^2 = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[(\hat{g}_{\pi,n}(S_t, A_t; \eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))^2\right]$$

$$= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[(\hat{g}_{\pi,n}(S_t, A_t; \eta, Q) - \delta_t^\pi(\eta, Q) + \delta_t^\pi(\eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))^2\right]$$

$$= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[(\delta_t^\pi(\eta, Q) - \hat{g}_{\pi,n}(S_t, A_t; \eta, Q))^2\right] + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[(\delta_t^\pi(\eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))^2\right]$$

$$+ \frac{2}{T}\sum_{t=1}^T \mathbb{E}\left[(\hat{g}_{\pi,n}(S_t, A_t; \eta, Q) - \delta_t^\pi(\eta, Q))(\delta_t^\pi(\eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))\right].$$

Since $\sum_{t=1}^T \mathbb{E}\left[(\mathcal{E}_\pi(S_t, A_t; \eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))g(S_t, A_t)\right] = 0$ for all $g \in \mathcal{G}$ due to the optimizing property of g_π^* in (7), the last term above can be written as

$$\frac{2}{T}\sum_{t=1}^T \mathbb{E}\Big[\big(\hat{g}_{\pi,n}(S_t, A_t; \eta, Q) - g_\pi^*(S_t, A_t; \eta, Q)$$
$$+ g_\pi^*(S_t, A_t; \eta, Q) - \delta_t^\pi(\eta, Q)\big)\big(\delta_t^\pi(\eta, Q) - g_\pi^*(S_t, A_t; \eta, Q)\big)\Big]$$

$$= \frac{2}{T}\sum_{t=1}^T \mathbb{E}\left[(g_\pi^*(S_t, A_t; \eta, Q) - \delta_t^\pi(\eta, Q))(\delta_t^\pi(\eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))\right]$$

$$= -\frac{2}{T}\sum_{t=1}^T \mathbb{E}\left[(\delta_t^\pi(\eta, Q) - g_\pi^*(S_t, A_t; \eta, Q))^2\right].$$

As a result, we have

$$\|\hat{g}_{\pi,n}(\eta,Q)-g^*_\pi(\eta,Q)\|^2=\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}(\delta^\pi_t(\eta,Q)-\hat{g}_{\pi,n}(S_t,A_t;\eta,Q))^2-(\delta^\pi_t(\eta,Q)-g^*_\pi(S_t,A_t;\eta,Q))^2\Big].$$

For $g_1,g_2\in\mathcal{G},\eta\in\mathbb{R},Q\in\mathcal{Q}$, we define the following functions f^π_1,f^π_2 of a trajectory $\mathcal{D}$ and functional $\mathbf{J}$ by

$$\begin{aligned}
f^\pi_1(g_1,g_2,\eta,Q)&:\mathcal{D}\mapsto\frac{1}{T}\sum_{t=1}^{T}(\delta^\pi_t(\eta,Q)-g_1(S_t,A_t))^2-(\delta^\pi_t(\eta,Q)-g_2(S_t,A_t))^2\\
f^\pi_2(g_1,g_2,\eta,Q)&:\mathcal{D}\mapsto\frac{1}{T}\sum_{t=1}^{T}(\delta^\pi_t(\eta,Q)-g_2(S_t,A_t))(g_1(S_t,A_t)-g_2(S_t,A_t)).
\end{aligned}$$

With these notations, we have

$$\begin{aligned}
&\|\hat{g}_{\pi,n}(\eta,Q)-g^*_\pi(\eta,Q)\|^2=\mathbb{E}f^\pi_1(\hat{g}_{\pi,n}(\eta,Q),g^*_\pi(\eta,Q),\eta,Q)\\
&\|\hat{g}_{\pi,n}(\eta,Q)-g^*_\pi(\eta,Q)\|^2_n=\mathbb{P}_n\left[f^\pi_1(\hat{g}_{\pi,n}(\eta,Q),g^*_\pi(\eta,Q),\eta,Q)+2f^\pi_2(\hat{g}_{n,\pi}(\eta,Q),g^*_\pi(\eta,Q),\eta,Q)\right].
\end{aligned}$$

We introduce the following decomposition for each pair of (η,Q):

$$\begin{aligned}
&\|\hat{g}_{\pi,n}(\eta,Q)-g^*_\pi(\eta,Q)\|^2+\|\hat{g}_{\pi,n}(\eta,Q)-g^*_\pi(\eta,Q)\|^2_n+\mu_nJ^2_2(\hat{g}_{\pi,n}(\eta,Q))\\
&=I_1(\eta,Q)+I_2(\eta,Q),
\end{aligned}$$

where

$$\begin{aligned}
I_1(\eta,Q)&=3\mathbb{P}_nf^\pi_1(\hat{g}_{\pi,n}(\eta,Q),g^*_\pi(\eta,Q),\eta,Q)+\mu_n[3J^2_2(\hat{g}_{\pi,n}(\eta,Q))+2J^2_2(g^*_\pi(\eta,Q))+2J^2_1(Q)]\\
I_2(\eta,Q)&=(\mathbb{P}_n+P)f^\pi_1(\hat{g}_{\pi,n}(\eta,Q),g^*_\pi(\eta,Q),\eta,Q)+\mu_nJ^2_2(\hat{g}_{\pi,n}(\eta,Q))\\
&\quad+2\mathbb{P}_nf^\pi_2(\hat{g}_{n,\pi}(\eta,Q),g^*_\pi(\eta,Q),\eta,Q)-I_1(\eta,Q).
\end{aligned}$$

For the first term, the optimizing property of $\hat{g}_{\pi,n}(\eta,Q)$ implies that

$$\begin{aligned}
\frac{1}{3}I_1(\eta,Q)&=\mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T}(\delta^\pi_t(\eta,Q)-\hat{g}_{\pi,n}(S_t,A_t;\eta,Q))^2-(\delta^\pi_t(\eta,Q)-g^*_\pi(S_t,A_t;\eta,Q))^2\Big]\\
&\quad+\mu_nJ^2_2(\hat{g}_{\pi,n}(\eta,Q))+\frac{2}{3}\mu_nJ^2_2(g^*_\pi(\eta,Q))+\frac{2}{3}\mu_nJ^2_1(Q)\\
&=\mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T}(\delta^\pi_t(\eta,Q)-\hat{g}_{\pi,n}(S_t,A_t;\eta,Q))^2\Big]+\mu_n\mathbf{J}^2(\hat{g}_{\pi,n}(\eta,Q))
\end{aligned}$$

$$
- \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}(\delta_t^\pi(\eta,Q) - g_\pi^*(S_t,A_t;\eta,Q))^2\big] + \frac{2}{3}\mu_n J_2^2(g_\pi^*(\eta,Q)) + \frac{2}{3}\mu_n J_1^2(Q)
$$
$$
\leq \frac{5}{3}\mu_n J_2^2(g_\pi^*(\eta,Q)) + \frac{2}{3}\mu_n J_1^2(Q).
$$

Thus, $I_1(\eta,Q) \leq 5\mu_n J_2^2(g_\pi^*(\eta,Q)) + 2\mu_n J_1^2(Q)$ holds for all (η,Q).

Next we derive the uniform bound of $I_2(\eta,Q)$ over all (η,Q). We apply the peeling device together with the exponential inequality of the relative deviation of the empirical process (Györfi et al. (2006), thm. 19.3). Note that $\mathbb{E}[f_2^\pi(\hat{g}_{n,\pi}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})] = 0$. We can then write $I_2(\eta,Q)$ as

$$
\begin{aligned}
I_2(\eta,Q) &= (\mathbb{P}_n + P)f_1^\pi(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) + \mu_n J_2^2(\hat{g}_{\pi,n}(\eta,Q)) \\
&\qquad + 2\mathbb{P}_n f_2^\pi(\hat{g}_{n,\pi}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) - 3\mathbb{P}_n f_1^\pi(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) \\
&\qquad - \mu_n[3J_2^2(\hat{g}_{\pi,n}(\eta,Q)) + 2J_2^2(g_\pi^*(\eta,Q)) + 2J_1^2(Q)] \\
&= 2(P-\mathbb{P}_n)(f_1^\pi - f_2^\pi)(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) - P(f_1^\pi - f_2^\pi)(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) \\
&\quad - 2\mu_n[J_2^2(\hat{g}_{\pi,n}(\eta,Q)) + J_2^2(g_\pi^*(\eta,Q)) + J_1^2(Q)].
\end{aligned}
$$

For simplicity, we introduce $f^\pi = f_1^\pi - f_2^\pi$, given by

$$
f^\pi(g_1,g_2,\eta,Q) : \mathcal{D} \mapsto \frac{1}{T}\sum_{t=1}^{T}(g_2 - g_1)(S_t,A_t)\cdot\left(3\delta_t^\pi(\eta,Q) - 2g_2(S_t,A_t) - g_1(S_t,A_t)\right),
$$

and the functional

$$
\mathbf{J}^2(g_1,g_2,Q) = J_2^2(g_1) + J_2^2(g_2) + J_1^2(Q),
$$

for any $g_1,g_2 \in \mathcal{G}$ and $Q \in \mathcal{Q}$. Define the interval $B = [-R_{\max}, R_{\max}]$. Fix some $t > 0$.

$$
\begin{aligned}
&\Pr\left(\exists(\eta,Q) \in B\times\mathcal{Q}, I_2(\eta,Q) > t\right) \\
&= \sum_{l=0}^{\infty}\Pr\Big(\exists(\eta,Q) \in B\times\mathcal{Q},\ 2\mu_n\mathbf{J}^2(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),Q) \in\ [2^l t\mathbb{1}_{\{l\neq 0\}}, 2^{l+1}t), \\
&\qquad\qquad 2(P-\mathbb{P}_n)f^\pi(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) > Pf^\pi(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),\eta,Q) \\
&\qquad\qquad\qquad + 2\mu_n\mathbf{J}^2(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),Q) + t\Big) \\
&\leq \sum_{l=0}^{\infty}\Pr\Big(\exists(\eta,Q) \in B\times\mathcal{Q},\ 2\mu_n\mathbf{J}^2(\hat{g}_{\pi,n}(\eta,Q), g_\pi^*(\eta,Q),Q) \leq 2^{l+1}t,
\end{aligned}
$$

$$2(P-\mathbb{P}_n)f^\pi(\hat{g}_{\pi,n}(\eta,Q),g_\pi^*(\eta,Q),\eta,Q) > Pf^\pi(\hat{g}_{\pi,n}(\eta,Q),g_\pi^*(\eta,Q),\eta,Q)+2^l t\Big)$$

$$\leq \sum_{l=0}^{\infty} \Pr\left(\sup_{h\in\mathcal{F}_l} \frac{(P-\mathbb{P}_n)h(\mathcal{D})}{Ph(\mathcal{D})+2^l t} > \frac{1}{2}\right),$$

where the function class $\mathcal{F}_l = \{f^\pi(g, g_\pi^*(\eta,Q),\eta,Q) : J_2^2(g) \leq \frac{2^l t}{\mu_n}, J_2^2(g_\pi^*(\eta,Q)) \leq \frac{2^l t}{\mu_n}, J_1^2(Q) \leq \frac{2^l t}{\mu_n}, \eta \in B, Q \in \mathcal{Q}, \pi \in \Pi\}$.

Next we verify the conditions (A1-A4) in Theorem 19.3 in Györfi et al. (2006) with $\mathcal{F} = \mathcal{F}_l$, $\epsilon = 1/2$ and $\eta = 2^l t$ to get an exponential inequality for each term in the summation. The conditions (A1) and (A2) are easy to verify using Assumptions 2, 3 and 4. For (A1), it is easy to see that for any $h = f^\pi(g, g_\pi^*(\eta,Q),\eta,Q) \in \mathcal{F}_l$,

$$\|f^\pi(g, g_\pi^*(\eta,Q),\eta,Q)\|_\infty \leq 6G_{\max}(2R_{\max}+2Q_{\max}+3G_{\max}),$$

and thus (A1) is satisfied with $K_1 = 6G_{\max}(2R_{\max}+2Q_{\max}+3G_{\max})$. For (A2), recall $f^\pi = f_1^\pi - f_2^\pi$ and thus

$$\mathbb{E}[f^\pi(g, g_\pi^*(\eta,Q),\eta,Q)^2] \leq 2\mathbb{E}[f_1^\pi(g, g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})^2] + 2\mathbb{E}[f_2^\pi(g, g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})^2].$$

For the first term of RHS:

$$\begin{aligned}
&\mathbb{E}[f_1^\pi(g, g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})^2] \\
&= \mathbb{E}[(\frac{1}{T}\sum_{t=1}^{T}(\delta_t^\pi(\eta,Q)-g(S_t,A_t))^2 - (\delta_t^\pi(\eta,Q)-g_\pi^*(S_t,A_t;\eta,Q))^2)^2] \\
&= \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}(2\delta_t^\pi(\eta,Q)-g(S_t,A_t)-g_\pi^*(S_t,A_t;\eta,Q))^2\,(g_\pi^*(S_t,A_t;\eta,Q)-g(S_t,A_t))^2] \\
&\leq (2(2R_{\max}+2Q_{\max})+2G_{\max})^2\,\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}(g_\pi^*(S_t,A_t;\eta,Q)-g(S_t,A_t))^2] \\
&= 4\,(2R_{\max}+2Q_{\max}+G_{\max})^2\,\mathbb{E}[f^\pi(g, g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})],
\end{aligned}$$

and the second term:

$$\begin{aligned}
&\mathbb{E}[f_2^\pi(g, g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})^2] \\
&= \mathbb{E}[(\frac{1}{T}\sum_{t=1}^{T}(\delta_t^\pi(\eta,Q)-g_\pi^*(S_t,A_t;\eta,Q))(g(S_t,A_t)-g_\pi^*(S_t,A_t;\eta,Q)))^2]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}(\delta_t^\pi(\eta,Q)-g_\pi^*(S_t,A_t;\eta,Q))^2(g(S_t,A_t)-g_\pi^*(S_t,A_t;\eta,Q))^2] \\
&\leq (2R_{\max}+2Q_{\max}+G_{\max})^2\,\mathbb{E}[\frac{1}{T}\sum_{t=1}^{T}(g_\pi^*(S_t,A_t;\eta,Q)-g(S_t,A_t))^2] \\
&= (2R_{\max}+2Q_{\max}+G_{\max})^2\,\mathbb{E}[f^\pi(g,g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})],
\end{aligned}
$$

where we use again the fact that $\mathbb{E}[f_2^\pi(\hat{g}_{n,\pi}(\eta,Q),g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})]=0$. Putting together, we have shown that

$$
\mathbb{E}[f^\pi(g,g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})^2] \leq 10\,(2R_{\max}+2Q_{\max}+G_{\max})^2\,\mathbb{E}[f^\pi(g,g_\pi^*(\eta,Q),\eta,Q)(\mathcal{D})].
$$

Thus (A2) is satisfied with $K_2 = 10\,(2R_{\max}+2Q_{\max}+G_{\max})^2$

Recall that $\epsilon = 1/2$ and $\eta = 2^l t$. To ensure the assumption (A3) holds for every l, i.e., $\sqrt{n}\epsilon\sqrt{1-\epsilon}\sqrt{\eta} \geq 288\max(K_1,\sqrt{2K_2})$, it is enough to ensure the inequality holds when $l=0$:

$$
\sqrt{n}(1/2)^{3/2}\sqrt{t} \geq 288\max(K_1,\sqrt{2K_2})
$$

That is, $t \geq c_1 n^{-1}$ where $c_1 = 8(288\max(K_1,\sqrt{2K_2}))^2$.

Next we verify (A4). Recall $\mathcal{Q}_M = \{c+Q : |c| \leq R_{\max}, Q\in\mathcal{Q}, J_1(Q)\leq M\}$ and $\mathcal{G}_M = \{g : g\in\mathcal{G}, J_2(g)\leq M\}$. It is straightforward to verify that with $M=\sqrt{\frac{2^l t}{\mu_n}}$,

$$
\begin{aligned}
&N(\epsilon(12R_{\max}+12Q_{\max}+21G_{\max}),\mathcal{F},\|\cdot\|_n) \\
&\leq N(\epsilon,\mathcal{Q}_M,\|\cdot\|_\infty)N^2(\epsilon,\mathcal{G}_M,\|\cdot\|_\infty)N(\epsilon,\mathcal{V}(\mathcal{Q}_M,\pi),\|\cdot\|_\infty)
\end{aligned}
$$

where $\mathcal{V}(\mathcal{Q}_M,\pi) = \{s\mapsto\sum_a\pi(a|s)Q(s,a) : Q\in\mathcal{Q}_M\}$ is a class of state-only function depending on the target policy π and the function class $\mathcal{Q}$. Let $c_2 = 12R_{\max}+12Q_{\max}+21G_{\max}$. As a result of the entropy condition in Assumption 5, we have

$$
\begin{aligned}
&\log N(\epsilon,\mathcal{F},\|\cdot\|_n) \\
&\leq \log N(\epsilon/c_2,\mathcal{Q}_M,\|\cdot\|_\infty)+2\log N(\epsilon/c_2,\mathcal{G}_M,\|\cdot\|_\infty)+\log N(\epsilon/c_2,\mathcal{V}(\mathcal{Q}_M,\pi),\|\cdot\|_\infty) \\
&\leq 4C_3(c_2M/\epsilon)^{2\alpha} = 4c_2^{2\alpha}C_3\left(\frac{2^l t}{\mu_n}\right)^\alpha\epsilon^{-2\alpha} = c_3\left(\frac{2^l t}{\mu_n}\right)^\alpha\epsilon^{-2\alpha},
\end{aligned}
$$

where C_3 is the constant introduced in Assumption 5 and $c_3 = 4c_2^{2\alpha}C_3$. Now we see the condition (A4) can be satisfied if the following inequality holds for all $x \geq 2^l t/8$,

$$\frac{\sqrt{n}(1/2)^2 x}{96\sqrt{2}\max(K_1, 2K_2)} \geq \int_0^{\sqrt{x}} \sqrt{c_3}\left(\frac{2^l t}{\mu_n}\right)^{\alpha/2} u^{-\alpha} du = x^{\frac{1-\alpha}{2}} \sqrt{c_3}\left(\frac{2^l t}{\mu_n}\right)^{\alpha/2}.$$

Or equivalently

$$x^{\frac{1+\alpha}{2}} \geq 4 \cdot 96\sqrt{2}\max(K_1, 2K_2)\sqrt{c_3}\left(\frac{2^l t}{\mu_n}\right)^{\alpha/2} n^{-1/2}.$$

Clearly it is enough to ensure the inequality holds when x is at the minimum (i.e., $x = 2^l t/8$), that is,

$$\begin{aligned}
&(2^l t/8)^{\frac{1+\alpha}{2}} \geq 4 \cdot 96\sqrt{2}\max(K_1, 2K_2)\sqrt{c_3}\left(\frac{2^l t}{\mu_n}\right)^{\alpha/2} n^{-1/2} \\
\iff &(2^l t)^{1/2} \geq 8^{\frac{1+\alpha}{2}} \cdot 4 \cdot 96\sqrt{2}\max(K_1, 2K_2)\sqrt{c_3}(\mu_n^\alpha n)^{-1/2}.
\end{aligned}$$

The above holds for all $l \geq 0$ as long as

$$t \geq \left(8^{\frac{1+\alpha}{2}} \cdot 4 \cdot 96\sqrt{2}\max(K_1, 2K_2)\sqrt{c_3}\right)^2 (\mu_n^\alpha n)^{-1} \geq c_4(\mu_n^\alpha n)^{-1},$$

where $c_4 = c_3 18(32)^4 \max(K_1^2, 4K_2^2)$. To summarize, the conditions (A1-A4) in Theorem 19.3 in Györfi et al. (2006) with the choice $\mathcal{F} = \mathcal{F}_l$, $\epsilon = 1/2$ and $\eta = 2^l t$ will hold for every $l \geq 0$ when $t \geq c_1 n^{-1}, t \geq \mu_n$ and $t \geq c_4(\mu_n^\alpha n)^{-1}$, which implies that

$$\begin{aligned}
&\Pr\left(\exists(\eta, Q) \in B \times \mathcal{Q}, I_2(\eta, Q) > t\right) \\
&\leq \sum_{l=0}^{\infty} \Pr\left(\sup_{h \in \mathcal{F}_l} \frac{(P - \mathbb{P}_n)h(\mathcal{D})}{Ph(\mathcal{D}) + 2^l t} > \frac{1}{2}\right) \\
&\leq \sum_{l=0}^{\infty} 60\exp\left(-\frac{n 2^l t (1/2)^3}{128 \cdot 2304 \max(K_1^2, K_2)}\right) \\
&= \sum_{l=0}^{\infty} 60\exp\left(-\frac{n 2^l t}{c_5}\right) \leq \sum_{k=1}^{\infty} 60\exp\left(-\frac{nkt}{c_5}\right) \leq \frac{60\exp(-\frac{nt}{c_5})}{1 - \exp(-\frac{nt}{c_5})},
\end{aligned}$$

where we define $c_5 = 8 \cdot 128 \cdot 2304 \max(K_1^2, K_2)$. Fix some $\delta > 0$, when $t \geq \log(120/\delta)c_5 n^{-1}$, we have both $\exp(-\frac{nt}{c_5}) \leq 1/2$ and $120\exp(-nt/c_5) \leq \delta$ and hence

$$\Pr\left(\exists(\eta, Q) \in B \times \mathcal{Q}, I_2(\eta, Q) > t\right) \leq \frac{60\exp(-\frac{nt}{c_5})}{1 - \exp(-\frac{nt}{c_5})} \leq 120\exp\left(-\frac{nt}{c_5}\right) \leq \delta.$$

Collecting all the conditions on t and combining with the bound of $I_1(\eta, Q)$, we have shown that with probability at least $1-\delta$, the following holds for all $(\eta, Q) \in B \times \mathcal{Q}$:

$$
\begin{aligned}
&\|\hat{g}_{\pi,n}(\eta, Q) - g_\pi^*(\eta, Q)\|^2 + \|\hat{g}_{\pi,n}(\eta, Q) - g_\pi^*(\eta, Q)\|_n^2 + \mu_n J_2^2(\hat{g}_{\pi,n}(\eta, Q)) \\
&\leq 5\mu_n J_2^2(g_\pi^*(\eta, Q)) + 2\mu_n J_1^2(Q) + c_1 n^{-1} + \mu_n + c_4(\mu_n^\alpha n)^{-1} + \log(120/\delta) c_5 n^{-1} \\
&= (1 + 5J_2^2(g_\pi^*(\eta, Q)) + 2J_1^2(Q))\mu_n + (c_1 + \log(120)c_5 + \log(1/\delta)c_5)n^{-1} + c_4(\mu_n^\alpha n)^{-1} \\
&\leq K\left((1 + J_2^2(g_\pi^*(\eta, Q)) + J_1^2(Q))\mu_n + \frac{1+\log(1/\delta)}{n} + \frac{1}{n\mu_n^\alpha}\right),
\end{aligned}
$$

where the leading constant can be chosen by $K = 5 + 8(288\max(K_1, \sqrt{2K_2}))^2 + 6\cdot 8\cdot 128\cdot 2304\max(K_1^2, K_2) + 4C_3\cdot 2(18\cdot(32)^4\cdot\max(K_1^2, 4K_2^2))(12R_{\max} + 12Q_{\max} + 21G_{\max})^{2\alpha}$, which depends on $R_{\max}, Q_{\max}, G_{\max}, C_3$ and α. $\square$

Lemma B.2. *Suppose the conditions in Lemma B.1 and Assumptions 3, 5 hold. Let $(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)$ be the estimator in (9) with tuning parameter λ_n and $\hat{g}_{\pi,n}(\eta, Q)$ be the estimated Bellman error operator in (8) with the tuning parameter μ_n. Up to some constant that depends only on $R_{\max}, Q_{\max}, G_{\max}, C_1, C_2, C_3$ and α, the following holds with probability at least $1-\delta$:*

$$
\begin{aligned}
\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|^2 \lesssim\ & n^{-\frac{1}{1+\alpha}} + \mu_n(1 + J_1^2(\tilde{Q}^\pi)) + \lambda_n J_1^2(\tilde{Q}^\pi) + (n\lambda_n^\alpha)^{-1} \\
& + (n\mu_n^\alpha)^{-1} + (n\mu_n^{\alpha/(1+\alpha)})^{-1} + \log(1/\delta)(n^{-1} + (n\mu_n^{\alpha/(1+\alpha)})^{-1}).
\end{aligned}
$$

Proof of Lemma B.2. Fix some $\delta > 0$. For the ease of notation, define a functional

$$
f(g) : \mathcal{D} \to \frac{1}{T}\sum_{t=1}^T g(S_t, A_t)^2, g \in \mathcal{G}.
$$

With this functional, we decompose the error by

$$
\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|^2 + \|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|_n^2 = (P + \mathbb{P}_n) f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) = I_1 + I_{n,2},
$$

where

$$
\begin{aligned}
I_1 &= 3\big(\mathbb{P}_n f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) + (2/3)\lambda_n J_1^2(\hat{Q}_n^\pi)\big) \\
I_2 &= (\mathbb{P}_n + P) f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) - I_1.
\end{aligned}
$$

Without loss of generality, we assume the average reward estimates $\hat{\eta}_n^\pi \in [-R_{\max}, R_{\max}]$ (otherwise one can first show the consistency and use the truncation argument). For the first term I_1, Assumption 3, the optimizing property (9) and the in-sample error bound in Lemma B.1 with the leading constant C (depending on only $R_{\max}, Q_{\max}, G_{\max}, C_3$ and α) imply that under the choice of (μ_n, λ_n) specified in the condition, the following holds with probability at least $1-\delta$,

$$
\begin{aligned}
I_1 &\leq 3\mathbb{P}_n f(\hat{g}_{\pi,n}(\eta^\pi, \tilde{Q}^\pi)) + 3\lambda_n J_1^2(\tilde{Q}^\pi) \\
&= 3\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \hat{g}_{\pi,n}^2(S_t, A_t; \eta^\pi, \tilde{Q}^\pi)\big] + 3\lambda_n J_1^2(\tilde{Q}^\pi) \\
&= 3\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} (\hat{g}_{\pi,n}(S_t, A_t; \eta^\pi, \tilde{Q}^\pi) - g_\pi^*(S_t, A_t; \eta^\pi, \tilde{Q}^\pi))^2\big] + 3\lambda_n J_1^2(\tilde{Q}^\pi) \\
&= 3\|\hat{g}_{\pi,n}(\eta^\pi, \tilde{Q}^\pi) - g_\pi^*(\eta^\pi, \tilde{Q}^\pi)\|_n^2 + 3\lambda_n J_1^2(\tilde{Q}^\pi) \\
&\leq 3C\Big(\mu_n(1 + J_1^2(\tilde{Q}^\pi)) + \frac{p}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n}\Big) + 3\lambda_n J_1^2(\tilde{Q}^\pi),
\end{aligned}
$$

where in the second equality we use $g_\pi^*(\eta^\pi, \tilde{Q}^\pi) = 0$ from Assumption 4.

The second term I_2 can be written as

$$
\begin{aligned}
I_2 &= (\mathbb{P}_n + P) f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) - 3(\mathbb{P}_n f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) + (2/3)\lambda_n J_1^2(\hat{Q}_n^\pi)) \\
&= 2(P - \mathbb{P}_n) f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) - P f(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) - 2\lambda_n J_1^2(\hat{Q}_n^\pi).
\end{aligned}
$$

Define the constant $\beta(n, \mu_n, \delta) = \sqrt{\frac{1}{n\mu_n^{\alpha+1}}} + \sqrt{\frac{1}{n\mu_n}} + \sqrt{\frac{\log(1/\delta)}{n\mu_n}}$. Using the uniform bound on the complexity (i.e., $J_2(\hat{g}_{\pi,n}(\eta, Q))$) in Lemma B.1 and Assumption 5, we see that with probability at least $1-\delta$,

$$
\begin{aligned}
J_2(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) &\leq C(1 + J_1(\hat{Q}_n^\pi) + J_2(g_\pi^*(\hat{\eta}_n^\pi, \hat{Q}_n^\pi) + \beta(n, \mu_n, \delta)) \\
&\leq C(1 + J_1(\hat{Q}_n^\pi) + C_1 + C_2 J_1(\hat{Q}_n^\pi) + \beta(n, \mu_n, \delta)) \\
&= C(1 + C_1 + (1 + C_2) J_1(\hat{Q}_n^\pi) + \beta(n, \mu_n, \delta)) \\
&\leq C(1 + C_1 + C_2)(1 + J_1(\hat{Q}_n^\pi) + \beta(n, \mu_n, \delta)) \\
&= c_1(1 + J_1(\hat{Q}_n^\pi) + \beta(n, \mu_n, \delta)),
\end{aligned}
$$

where we introduce $c_1 = C(1 + C_1 + C_2)$; recall that C_1, C_2 are constants specified in Assumption 5 and C is the leading constant in Lemma B.1. For simplicity we denote this event by $E = \{J_2(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) \leq c_1(1 + J_1(\hat{Q}^\pi_n) + \beta(n, \mu_n, \delta))\}$. Now we have $\Pr(I_2 > t) \leq \Pr(\{I_2 > t\} \cap E) + \delta$ and we bound the first term using peeling device on $2\lambda_n J_1^2(\hat{Q}^\pi_n)$ in I_2. In particular,

$$
\begin{aligned}
&\Pr(\{I_2 > t\} \cap E) = \sum_{l=0}^{\infty} \Pr\left(\{I_2 > t,\ 2\lambda_n J_1^2(\hat{Q}^\pi_n) \in [2^l t \mathbb{1}_{\{t\neq 0\}}, 2^{l+1}t)\} \cap E\right) \\
&\leq \sum_{l=0}^{\infty} \Pr\Big(\ 2(P - \mathbb{P}_n) f(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) > Pf(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) + 2\lambda_n J_1^2(\hat{Q}^\pi_n) + t, \\
&\qquad\qquad 2\lambda_n J_1^2(\hat{Q}^\pi_n) \in [2^l t \mathbb{1}_{\{t\neq 0\}}, 2^{l+1}t), J_2(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) \leq c_1(1 + J_1(\hat{Q}^\pi_n) + \beta(n, \mu_n, \delta))\Big) \\
&\leq \sum_{l=0}^{\infty} \Pr\Big(\ 2(P - \mathbb{P}_n) f(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) > Pf(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) + 2^l t \mathbb{1}_{\{t\neq 0\}} + t, \\
&\qquad\qquad 2\lambda_n J_1^2(\hat{Q}^\pi_n) \leq 2^{l+1}t, J_2(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) \leq c_1(1 + \sqrt{(2^l t)/\lambda_n} + \beta(n, \mu_n, \delta))\Big) \\
&\leq \sum_{l=0}^{\infty} \Pr\Big(\ 2(P - \mathbb{P}_n) f(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) > Pf(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) + 2^l t, \\
&\qquad\qquad J_2(\hat{g}_{\pi,n}(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)) \leq c_1(1 + \sqrt{(2^l t)/\lambda_n} + \beta(n, \mu_n, \delta))\Big) \\
&\leq \sum_{l=0}^{\infty} \Pr\left(\sup_{h \in \mathcal{F}_l} \frac{(P - \mathbb{P}_n)h(\mathcal{D})}{Ph(\mathcal{D}) + 2^l t} > \frac{1}{2}\right),
\end{aligned}
$$

where $\mathcal{F}_l = \{f(g) : J_2(g) \leq c_1(1 + \sqrt{(2^l t)/\lambda_n} + \beta(n, \mu_n, \delta)), g \in \mathcal{G}\}$. In what follows we verify the conditions (A1-A4) in Theorem 19.3 in Györfi et al. (2006) with $\mathcal{F} = \mathcal{F}_l$, $\epsilon = 1/2$ and $\eta = 2^l t$ to get an exponential inequality for each term in the summation.

The first three conditions are straightforward to verify. For (A1), it is easy to see that $|f(g)(\mathcal{D})| = |\frac{1}{T}\sum_{t=1}^T g(S_t, A_t)^2| \leq G_{\max}^2$. Thus (A1) holds with $K_1 = G_{\max}^2$. For (A2), we have $Pf^2(g) \leq G_{\max}^2 Pf(g)$ and thus (A2) holds by choosing $K_2 = G_{\max}^2$. The condition (A3), i.e., $\sqrt{n}\epsilon\sqrt{1-\epsilon}\sqrt{\eta} \geq 288 \max\{2K_1, \sqrt{2K_2}\}$, becomes

$$\sqrt{n}(1/2)^{3/2}\sqrt{2^l t} \geq 288 \max\{2G_{\max}^2, \sqrt{2}G_{\max}\}.$$

So this holds for all $l \geq 0$ as long as $t \geq c_2/n$, where $c_2 = (8 \cdot 288 \max\{2G_{\max}^2, \sqrt{2}G_{\max}\})^2$.

We now verify the final condition (A4). First, we obtain an upper bound $\mathcal{N}_2(u, \mathcal{F}_l; \|\cdot\|_\infty)$ for all possible realization of trajectories. For any $g_1, g_2 \in \mathcal{G}$,

$$\frac{1}{n}\sum_{i=1}^{n}\left[f(g_1)(\mathcal{D}_i) - f(g_2)(\mathcal{D}_i)\right]^2 \leq 4G_{\max}^2\|g_1 - g_2\|_n^2.$$

Thus applying Assumption 5 implies that for some constant c_3, the metric entropy for each l is bounded by

$$\begin{aligned}
&\log N(u, \mathcal{F}_l, \|\cdot\|_\infty) \\
&\leq \log N\left(\frac{u}{2G_{\max}}, \{g : J_2(g) \leq c_1(1 + \sqrt{(2^l t)/\lambda_n} + \beta(n, \mu_n, \delta)), g \in \mathcal{G}\}, \|\cdot\|_\infty\right) \\
&\leq C_3\left(\frac{c_1(1 + \sqrt{(2^l t)/\lambda_n} + \beta(n, \mu_n, \delta))}{u/(2G_{\max})}\right)^{2\alpha} \leq c_3\left(1 + \left(\frac{2^l t}{\lambda_n}\right)^{\alpha} + \beta(n, \mu_n, \delta)^{2\alpha}\right)u^{-2\alpha},
\end{aligned}$$

where the constant C_3 in the last inequality is given in Assumption 5 and $c_3 = (2G_{\max}c_1)^2 C_3$. Now we just need to ensure for all $x \geq \eta/8 = 2^l t/8$ and $l \geq 0$:

$$\frac{\sqrt{n}(1/2)^2 x}{96\sqrt{2}\cdot 2G_{\max}^2} \geq \int_0^{\sqrt{x}} \sqrt{c_3}\left(1 + \left(\frac{2^l t}{\lambda_n}\right)^{\alpha} + \beta(n, \mu_n, \delta)^{2\alpha}\right)^{1/2} u^{-\alpha} du$$

Note that $\int_0^{\sqrt{x}} u^{-\alpha} du = x^{\frac{1-\alpha}{2}}$. The above equality is equivalent with the following:

$$\frac{(1/2)^2}{96\sqrt{2}\cdot 2G_{\max}^2}\sqrt{n}x^{\frac{1+\alpha}{2}} \geq \sqrt{c_3}\left(1 + \left(\frac{2^l t}{\lambda_n}\right)^{\alpha} + \beta(n, \mu_n, \delta)^{2\alpha}\right)^{1/2}.$$

Using the inequality $(a+b)^{1/2} \leq \sqrt{a} + \sqrt{b}$ and the fact that LHS is increasing function of x, it's enough to ensure that the following three inequalities hold:

$$\begin{aligned}
\frac{1}{4\cdot 96\sqrt{2}\cdot 2G_{\max}^2}\sqrt{n}(2^l t/8)^{\frac{1+\alpha}{2}} &\geq \sqrt{c_3} \\
\frac{1}{4\cdot 96\sqrt{2}\cdot 2G_{\max}^2}\sqrt{n}(2^l t/8)^{\frac{1+\alpha}{2}} &\geq \sqrt{c_3}\left(\frac{2^l t}{\lambda_n}\right)^{\alpha/2} \\
\frac{1}{4\cdot 96\sqrt{2}\cdot 2G_{\max}^2}\sqrt{n}(2^l t/8)^{\frac{1+\alpha}{2}} &\geq \sqrt{c_3}\beta(n, \mu_n, \delta)^{\alpha}.
\end{aligned}$$

It is easy to see that each of the above inequalities can be satisfied for all $l \geq 0$ by choosing a large enough t. Specifically, the first one will hold for all $l \geq 0$ if

$$t \geq 8(c_3 4\cdot 96\sqrt{2}\cdot 2G_{\max}^2)^{\frac{1}{1+\alpha}} n^{-1/(1+\alpha)} := c_4 n^{-1/(1+\alpha)}$$

Similarly, the second holds for all l if

$$t \geq 8^{1+\alpha} c_3 (4 \cdot 96\sqrt{2} \cdot 2G_{\max}^2)^2 (n\lambda_n^\alpha)^{-1} := c_5 (n\lambda_n^\alpha)^{-1}$$

The third one holds for all $l \geq 0$ whenever t satisfies

$$t \geq 8\big((4 \cdot 96\sqrt{2} \cdot 2G_{\max}^2)^2 c_3\big)^{\frac{1}{1+\alpha}} \cdot n^{-1/(1+\alpha)} \beta^{\frac{2\alpha}{1+\alpha}}(n, \mu_n, \delta, p) = c_4 n^{-1/(1+\alpha)} \beta(n, \mu_n, \delta)^{\frac{2\alpha}{1+\alpha}}.$$

Note that

$$\begin{aligned} n^{-\frac{1}{1+\alpha}} \beta(n, \mu_n, \delta)^{\frac{2\alpha}{1+\alpha}} &= n^{-\frac{1}{1+\alpha}} \left[n^{-1/2} \mu_n^{-(1+\alpha)/2} + (1 + \sqrt{\log(1/\delta)})(n\mu_n)^{-1/2} \right]^{\frac{2\alpha}{1+\alpha}} \\ &\leq \frac{1}{n\mu_n^\alpha} + \frac{1 + \log^{\frac{\alpha}{1+\alpha}}(1/\delta)}{n\mu_n^{\alpha/(1+\alpha)}}. \end{aligned}$$

Thus to satisfy the third condition, it is enough to assume that

$$t \geq c_4 \left[\frac{1}{n\mu_n^\alpha} + \frac{1 + \log^{\frac{\alpha}{1+\alpha}}(1/\delta)}{n\mu_n^{\alpha/(1+\alpha)}} \right].$$

Putting all together, all conditions (A1) to (A4) would be satisfied for all $l \geq 0$ when

$$t \geq c_2 n^{-1} + c_4 n^{-1/(1+\alpha)} + c_5 (n\lambda_n^\alpha)^{-1} + c_4 \left[\frac{1}{n\mu_n^\alpha} + \frac{1 + \log^{\frac{\alpha}{1+\alpha}}(1/\delta)}{n\mu_n^{\alpha/(1+\alpha)}} \right].$$

We can now apply Theorem 19.3 in (Györfi et al. 2006) for each l-th term:

$$\Pr(\{I_2 > t\} \cap E) \leq \sum_{l=0}^{\infty} \Pr\left(\sup_{h \in \mathcal{F}_l} \frac{(P - \mathbb{P}_n)h(\mathcal{D})}{Ph(\mathcal{D}) + 2^l t} > \frac{1}{2} \right) \leq \frac{60 \exp(-\frac{nt}{c_6})}{1 - \exp(-\frac{nt}{c_6})},$$

where $c_6 = 8 \cdot 128 \cdot 2304 \max(G_{\max}^4, G_{\max}^2)$. When $t \geq \log(120/\delta) c_6 n^{-1}$, we have both $\exp(-\frac{nt}{c_6}) \leq 1/2$ and $120 \exp(-nt/c_6) \leq \delta$ and thus

$$\Pr(I_2 > t) \leq \delta + \frac{60 \exp(-\frac{nt}{c_5})}{1 - \exp(-\frac{nt}{c_5})} \leq \delta + 120 \exp(-nt/c_5) \leq 2\delta.$$

Collecting all conditions on t and combing with the bound on I_1, we have shown that with probability at least $1 - 2\delta$,

$$\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|^2 + \|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|_n^2$$

$$\leq 3C\Big(\mu_n(1+J_1^2(\tilde{Q}^\pi)) + \frac{1}{n\mu_n^\alpha} + \frac{1}{n} + \frac{\log(1/\delta)}{n}\Big) + 3\lambda_n J_1^2(\tilde{Q}^\pi)$$
$$+ c_2 n^{-1} + c_5 n^{-\frac{1}{1+\alpha}} + c_5(n\lambda_n^\alpha)^{-1} + c_5\left[\frac{1}{n\mu_n^\alpha} + \frac{1+\log^{\frac{\alpha}{1+\alpha}}(1/\delta)}{n\mu_n^{\alpha/(1+\alpha)}}\right] + \log(120/\delta)c_6 n^{-1}.$$

Note that all of the constants specified throughout the proof, e.g., $c_1,\ldots,c_6$ depends only on $R_{\max}, Q_{\max}, G_{\max}, \{C_i\}_{i=1}^3$ and α. Thus we have up to such a constant

$$\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|^2 + \|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi,\hat{Q}_n^\pi)\|_n^2 \lesssim n^{-\frac{1}{1+\alpha}} + \mu_n(1+J_1^2(\tilde{Q}^\pi)) + \lambda_n J_1^2(\tilde{Q}^\pi) + (n\lambda_n^\alpha)^{-1}$$
$$+ (n\mu_n^\alpha)^{-1} + (n\mu_n^{\alpha/(1+\alpha)})^{-1} + \log(1/\delta)(n^{-1} + (n\mu_n^{\alpha/(1+\alpha)})^{-1})$$

□

B.2 Proof of Theorem 2

Proof of Theorem 2. Let the objective function in (9) be $L_n(\eta, Q)$:

$$L_n(\eta,Q) = \mathbb{P}_n\big[(1/T)\sum_{t=1}^T \hat{g}_{\pi,n}^2(S_t,A_t;\eta,Q)\big] + \lambda_n J_1^2(Q).$$

Recall the definition of $\tilde{q}^\pi$ in Assumption 6 and by assumption we have $\tilde{q}^\pi \in \mathcal{Q}$. Consider the directional derivative:

$$\begin{aligned}
&\frac{d}{du}L_n(\eta-u, Q+u\tilde{q}^\pi)|_{u=0}\\
&= \frac{d}{du}\mathbb{P}_n\big[(1/T)\sum_{t=1}^T \hat{g}_{\pi,n}^2(S_t,A_t;\eta-u,Q+u\tilde{q}^\pi)\big] + \lambda_n J_1^2(Q+u\tilde{q}^\pi)|_{u=0}\\
&= 2\mathbb{P}_n\big[(1/T)\sum_{t=1}^T \hat{g}_{\pi,n}(S_t,A_t;\eta,Q)\times\frac{d}{du}\hat{g}_{\pi,n}(S_t,A_t;\eta-u,Q+u\tilde{q}^\pi)|_{u=0}\big] + 2\lambda_n J_1(Q,\tilde{q}^\pi).
\end{aligned}$$

By using first-order optimization condition of (8), one can show that for any (s,a),

$$\frac{d}{du}\hat{g}_{\pi,n}(s,a;\eta-u,Q+u\tilde{q}^\pi)|_{u=0} = \hat{e}_n^\pi(s,a),$$

where $\hat{e}_n^\pi = \hat{e}_n^\pi(\cdot,\cdot)$ is given by

$$\hat{e}_n^\pi = \operatorname*{argmin}_{g\in\mathcal{G}} \mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^T(1-\tilde{q}^\pi(S,A)+\sum_{a'}\pi(a'|S')\tilde{q}^\pi(S',a')-g(S_t,A_t))^2\Big] + \mu_n J_2^2(g). \quad \text{(B.5)}$$

See the derivation of (B.5) at the end of the proof. Based on the relationship between e^π and q^π in (12), one can view $\hat{e}_n^\pi$ as an infeasible estimator of e^π (infeasible as it requires the access to $\tilde{q}^\pi$). Since $(\hat{\eta}_n^\pi, \hat{Q}_n^\pi) = \operatorname{argmin}_{\eta,Q} L_n(\eta, Q)$, we have $\frac{d}{du} L_n(\hat{\eta}_n^\pi - t, \hat{Q}_n^\pi + t\tilde{q}^\pi)\big|_{t=0} = 0$, or equivalently

$$\mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T} \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}_n^\pi, \hat{Q}_n^\pi)\hat{e}_n^\pi(S_t, A_t)\Big] + \lambda_n J_1(\hat{Q}_n^\pi, \tilde{q}^\pi) = 0. \tag{B.6}$$

From Theorem 1, we have $J_1(\hat{Q}_n^\pi) = O_P(1)$ and since $\lambda_n = o(n^{-1/2})$ by assumption we have

$$\lambda_n J_1(\hat{Q}_n^\pi, \tilde{q}^\pi) = O_P(n^{-1/2}). \tag{B.7}$$

We decompose the first term into

$$\begin{aligned}
&\mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T} \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}_n^\pi, \hat{Q}_n^\pi)\hat{e}_n^\pi(S_t, A_t)\Big] \\
&= \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}_n^\pi, \hat{Q}_n^\pi)e^\pi(S_t, A_t)\big] + \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}_n^\pi, \hat{Q}_n^\pi)(\hat{e}_n^\pi - e^\pi)(S_t, A_t)\big].
\end{aligned}$$

Note that $e^\pi \in \mathcal{G}$ by Assumption 6. Apply the same technique used in proving Lemma B.1, one can show that the empirical error bound $\|\hat{e}_n^\pi - e^\pi\|_n^2 = O_P(\mu_n)$. In Lemma B.2, we have shown $\|\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|_n^2 = O_P(\lambda_n)$. Since by assumption $\mu_n = O(\lambda_n)$ and $\lambda_n = o(n^{-1/2})$, we have $\mu_n = o(n^{-1/2})$. Apply Cauchy inequality gives that

$$\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}_n^\pi, \hat{Q}_n^\pi)(\hat{e}_n^\pi - e^\pi)(S_t, A_t)\big] = o_P(n^{-1/2}). \tag{B.8}$$

On the other hand, the directional derivative along $e^\pi \in \mathcal{G}$ of the objective function in (8) at $(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)$ implies that

$$\begin{aligned}
&\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}_n^\pi, \hat{Q}_n^\pi)e^\pi(S_t, A_t)\big] \\
&= \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \delta_t^\pi(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)e^\pi(S_t, A_t)\big] - 2\mu_n J_2(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi), e^\pi),
\end{aligned}$$

where we use the notation for the temporal difference error at time t, i.e., $\delta_t^\pi(\eta, Q) = R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1}, a') - \eta - Q(S_t, A_t)$. Note that Lemma B.1 implies that $J_2(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)) =$

$O_P(1)$ since $\mu_n^{-1} = O(n^{1/(1+\alpha)})$ by assumption. Thus the second term $\mu_n J_2(\hat{g}_{\pi,n}(\hat{\eta}_n^\pi, \hat{Q}_n^\pi), e^\pi) = o_P(n^{-1/2})$. Now we consider the first term above. Plugging-in the the temporal difference error at the true value $(\eta^\pi, \tilde{Q}^\pi)$ implies that

$$
\begin{aligned}
&\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\delta_t^\pi(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)e^\pi(S_t, A_t)\big] \\
&= \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}(\delta_t^\pi(\eta^\pi, \tilde{Q}^\pi) + (\delta_t^\pi(\hat{\eta}_n^\pi, \hat{Q}_n^\pi) - \delta_t^\pi(\eta^\pi, \tilde{Q}^\pi))e^\pi(S_t, A_t)\big] \\
&= \mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}\delta_t^\pi(\eta^\pi, Q^\pi)e^\pi(S_t, A_t)\big] - (\hat{\eta}_n^\pi - \eta^\pi)\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T}e^\pi(S_t, A_t)\big] + \mathrm{Rem},
\end{aligned}
$$

where the last term, Rem is given by

$$
\mathrm{Rem} = \mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T}\big[(\tilde{Q}^\pi - \hat{Q}_n^\pi)(S_t, A_t) - \sum_{a'}\pi(a'|S_{t+1})(\tilde{Q}^\pi - \hat{Q}_n^\pi)(S_{t+1}, a')\big]e^\pi(S_t, A_t)\Big].
$$

For $Q \in \mathcal{Q}$, define the function of trajectory $f^\pi(Q)(\mathcal{D})$ by

$$
f^\pi(Q) : \mathcal{D} \to \frac{1}{T}\sum_{t=1}^{T}\big[Q(S_t, A_t) - \sum_{a'}\pi(a'|S_{t+1})Q(S_{t+1}, a')\big]e^\pi(S_t, A_t).
$$

Note that we can write $Pf^\pi(Q) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[h(S_t, A_t)e^\pi(S_t, A_t)\right]$, where the function $h(s, a) = Q(s, a) - \sum_{a'}\mathbb{E}\left[\pi(a'|S_{t+1})Q(S_{t+1}, a')|S_t = a, A_t = a\right]$ is mean 0 under stationary distribution. Using the orthogonality in (11), we have $Pf^\pi(Q) = 0$. The remainder term can then be written as

$$
\mathrm{Rem} = -\sqrt{n}(\mathbb{G}_n f(\hat{Q}_n^\pi) - \mathbb{G}_n f(\tilde{Q}^\pi)),
$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process. Note that $J_1(\hat{Q}_n^\pi) = O_P(1)$ and the sup-norm metric condition (5) implies that the bracketing entropy integral of the function class $\mathcal{F} = \{f(Q) : Q \in \mathcal{Q}, J(Q) \leq M\}$ is finite for all M. Using Assumptions 6 and 7 and Lemma B.3, we can show that up to a constant K

$$
P(f(\hat{Q}_n^\pi) - f(\tilde{Q}^\pi))^2 \leq K\|\mathcal{E}_\pi(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\|^2. \tag{B.9}
$$

See the derivation of (B.9) and the constant K at the end of the proof. By Theorem 1 and the condition on λ_n, we have $\|\mathcal{E}_\pi(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)\| = O_P(\lambda_n) = o_P(1)$. We conclude that

$P(f(\hat{Q}^\pi_n) - f(\tilde{Q}^\pi))^2 = o_P(1)$. The asymptotic equicontinuity (Van de Geer (2000), thm. 19.5) implies that $|\mathbb{G}_n f(\hat{Q}^\pi_n) - \mathbb{G}_n f(\tilde{Q}^\pi)| = o_P(1)$ and thus Rem $= o_P(n^{-1/2})$. Now by plugging (B.7 and (B.8) into the equation (B.6), we have shown that

$$(\hat{\eta}^\pi_n - \eta^\pi)\mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} e^\pi(S_t, A_t)\big] = \mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \delta^\pi_t(\eta^\pi, Q^\pi)e^\pi(S_t, A_t)\big] + o_P(n^{-1/2}). \tag{B.10}$$

Recall the definition of e^π in (10). Taking the expectation on both sides implies that

$$\frac{d^\pi(s,a)}{\bar{d}_T(s,a)} = \frac{e^\pi(s,a)}{\mathbb{E}[(1/T)\sum_{t=1}^{T} e^\pi(S_t, A_t)]}.$$

Obviously we have $\mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T} \delta^\pi_t(\eta^\pi, Q^\pi)e^\pi(S_t, A_t)\big] = O_P(n^{-1/2})$ and $\mathbb{P}_n[\frac{1}{T}\sum_{t=1}^{T} e^\pi(S_t, A_t)] = \mathbb{E}[\frac{1}{T}\sum_{t=1}^{T} e^\pi(S_t, A_t)] + o_P(1)$. Consequently, re-arranging (B.10) implies that $\sqrt{n}(\hat{\eta}^\pi_n - \eta^\pi) = \mathbb{G}_n[(1/T)\sum_{t=1}^{T} \frac{d^\pi(S_t,A_t)}{\bar{d}_T(S_t,A_t)}\delta^\pi_t(S_t, A_t; \eta^\pi, Q^\pi)] + o_P(1)$ and we get the desired result by using Slutsky's theorem.

Derivation of Equation (B.5) Note that for all $g \in \mathcal{G}$, $\hat{g}_{\pi,n}(\eta, Q)$ satisfies

$$\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \left(\delta(S_t, A_t, S_{t+1}, R_{t+1}; \eta, Q) - \hat{g}_{\pi,n}(S_t, A_t; \eta, Q)\right) g(S_t, A_t)\big] = \mu_n J_2(g, \hat{g}_{\pi,n}(\eta, Q)).$$

Now for all t and $g \in \mathcal{G}$, we have

$$\mathbb{P}_n\big[(1/T)\sum_{t=1}^{T} \Big[\delta(S_t, A_t, S_{t+1}, R_{t+1}; \hat{\eta}^\pi_n - t, \hat{Q}^\pi_n + t\tilde{q}^\pi) - \hat{g}_{\pi,n}(S_t, A_t; \hat{\eta}^\pi_n - t, \hat{Q}^\pi_n + t\tilde{q}^\pi)\Big] g(S_t, A_t)\big]$$
$$= \mu_n J_2(g, \hat{g}_{\pi,n}(\hat{\eta}^\pi_n - t, \hat{Q}^\pi_n + t\tilde{q}^\pi)).$$

Taking derivative w.r.t. t at $t = 0$ of above gives

$$\mathbb{P}_n\big[\frac{1}{T}\sum_{t=1}^{T}[1 - \tilde{q}^\pi(S_t, A_t) + \sum_{a'} \pi(a'|S_{t+1})\tilde{q}^\pi(S_{t+1}, a') - \hat{e}^\pi_n(S_t, A_t)]g(S_t, A_t)\big] = \mu_n J_2(g, \hat{e}^\pi_n),$$

where we use the fact that

$$\frac{d}{du}\delta(S, A, S', R; \hat{\eta}^\pi_n - t, \hat{Q}^\pi_n + t\tilde{q}^\pi) = 1 - \tilde{q}^\pi(S, A) + \sum_{a'} \pi(a'|S')\tilde{q}^\pi(S', a').$$

And thus we can see that $\hat{e}^\pi_n$ solves

$$\hat{e}^\pi_n = \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{P}_n\Big[\frac{1}{T}\sum_{t=1}^{T} \big(1 - \tilde{q}^\pi(S_t, A_t) + \sum_{a'} \pi(a'|S_{+1})\tilde{q}^\pi(S_{t+1}, a') - g(S_t, A_t)\big)^2\Big] + \mu_n J_2^2(g).$$

Derivation of Inequality (B.9) Define the operator $U^\pi(s,a,s';Q) = \sum_{a'} \pi(a'|s')Q(s',a') - Q(s,a)$. For simplicity, we denote $U^\pi(\cdot,\cdot,\cdot;Q^\pi)$ by U^π and $U^\pi(\cdot,\cdot,\cdot;\hat{Q}^\pi_n) = \hat{U}^\pi_n$. By Assumption 6, we have $P(f(\hat{Q}^\pi_n) - f(\tilde{Q}^\pi))^2 \le G^2_{\max}\|\hat{U}^\pi_n - U^\pi\|^2$. Below we provide the bound for $\|\hat{U}^\pi_n - U^\pi\|^2$. Note the norm here is under the distribution of the transition samples (state, action and next state) in the trajectory.

First we note that $U^\pi(s,a,s';Q) = U^\pi(s,a,s';Q+c)$, i.e., shifting by a constant does not change the value. Thus, for $\bar{Q}^\pi_n = \hat{Q}^\pi_n - \mu_\pi(\hat{Q}^\pi_n)$ we have $\|\hat{U}^\pi_n - U^\pi\| = \|U^\pi(\bar{Q}^\pi_n) - U^\pi\|$ and applying Lemma B.4 with $Q = \bar{Q}^\pi_n$ gives

$$\|\hat{U}^\pi_n - U^\pi\| \le \sqrt{1 + (1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty)(1/p_{\min})}\|\bar{Q}^\pi_n - Q^\pi\|. \tag{B.11}$$

On the other hand, since $\bar{Q}^\pi$ is mean 0 under stationary distribution by construction, we can apply Lemma B.5 with $\tilde{Q} = \bar{Q}^\pi_n$ and get

$$\|\bar{Q}^\pi_n - Q^\pi\| \le 2\left(1 + C_4\beta/(1-\beta)\right)\|(\mathcal{I} - \mathcal{P}^\pi)(\bar{Q}^\pi_n - Q^\pi)\|.$$

Clearly the functional $(\mathcal{I} - \mathcal{P}^\pi)$ is invariant to a constant shift, i.e. $(\mathcal{I} - \mathcal{P}^\pi)c = 0$ for any constant function c. We can change back to the original estimates on RHS and apply Lemma B.3 to obtain

$$\begin{aligned}
\|\bar{Q}^\pi_n - Q^\pi\| &\le 2\left(1 + C_4\beta/(1-\beta)\right)\|(\mathcal{I} - \mathcal{P}^\pi)(\hat{Q}^\pi_n - Q^\pi)\|_\nu \\
&\le 2\left(1 + C_4\beta/(1-\beta)\right)(1 + \sqrt{1+\sigma^2_\pi})\|\mathcal{E}_\pi(\hat{\eta}^\pi_n, \hat{Q}^\pi_n)\|.
\end{aligned} \tag{B.12}$$

Combing the two bounds in (B.11) and (B.12) gives the desired inequality (B.9) with the constant $K = 4G^2_{\max}(1 + p^{-1}_{\min}(1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty))\left(1 + C_4\beta/(1-\beta)\right)^2(1 + \sqrt{1+\sigma^2_\pi})^2$.

$\square$

Proof of Collorary 1. For each target policy $\pi \in \Pi$, we can apply the same argument as in the proof of Theorem 2 and get

$$\begin{pmatrix} \sqrt{n}(\hat{\eta}^{\pi_1}_n - \eta^{\pi_1}) \\ \vdots \\ \sqrt{n}(\hat{\eta}^{\pi_K}_n - \eta^{\pi_K}) \end{pmatrix} = \sqrt{n}\mathbb{P}_n \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T \frac{d^{\pi_1}(S_t,A_t)}{\bar{d}_T(S_t,A_t)}\delta^{\pi_1}_t(S_t,A_t;\eta^\pi,Q^\pi) \\ \vdots \\ \frac{1}{T}\sum_{t=1}^T \frac{d^{\pi_K}(S_t,A_t)}{\bar{d}_T(S_t,A_t)}\delta^{\pi_K}_t(S_t,A_t;\eta^\pi,Q^\pi) \end{pmatrix} + o_P(1)$$

Recall $\epsilon_t^\pi = \frac{d^\pi(S_t,A_t)}{d_T(S_t,A_t)}(R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q^\pi(S_{t+1},a') - \eta^\pi - Q^\pi(S_t,A_t))$. Note that for any i,j, we have $\mathbb{E}[(\frac{1}{T}\sum_{t=1}^T \epsilon_t^{\pi_i})(\frac{1}{T}\sum_{t=1}^T \epsilon_t^{\pi_j})] = (1/T^2)\mathbb{E}[\sum_{t=1}^T \epsilon_t^{\pi_i}\epsilon_t^{\pi_j}] + (1/T^2)\mathbb{E}[\sum_{t\neq s} \epsilon_t^{\pi_i}\epsilon_t^{\pi_j}] = (1/T^2)\mathbb{E}[\sum_{t=1}^T \epsilon_t^{\pi_i}\epsilon_t^{\pi_j}]$. The result is followed by applying Slutsky's theorem.

□

Lemma B.3. *Suppose Assumption 1 holds. Then, for all $(\eta, Q) \in \mathbb{R} \times \mathcal{Q}$, $|\eta - \eta^\pi| \leq \sqrt{1+\sigma_\pi^2}\|\mathcal{E}_\pi(\eta,Q)\|$ and $\|(\mathcal{I} - \mathcal{P}^\pi)(Q - Q^\pi)\| \leq (1 + \sqrt{1+\sigma_\pi^2})\|\mathcal{E}_\pi(\eta,Q)\|$, where $\mathcal{I}$ is the identity operator.*

Proof of Lemma B.3. Note that the Bellman error can be written as

$$
\begin{aligned}
&\mathcal{E}_\pi(s,a;\eta,Q) = \mathbb{E}\left[R_{t+1} + \sum_{a'} \pi(a'|S_{t+1})Q(S_{t+1},a') - \eta - Q(s,a) \,|\, S_t = s, A_t = a\right] \\
&= (\eta^\pi - \eta) + (Q^\pi - Q)(s,a) - \mathcal{P}^\pi(Q^\pi - Q)(s,a) \\
&= (\eta^\pi - \eta)e^\pi(s,a) + (\eta^\pi - \eta)u^\pi(s,a) + (Q^\pi - Q)(s,a) - \mathcal{P}^\pi(Q^\pi - Q)(s,a) \\
&= (\eta^\pi - \eta)e^\pi(s,a) + (\eta^\pi - \eta)(q^\pi(s,a) - \mathcal{P}^\pi q^\pi(s,a)) + (Q^\pi - Q)(s,a) - \mathcal{P}^\pi(Q^\pi - Q)(s,a) \\
&= (\eta^\pi - \eta)e^\pi(s,a) + h(s,a) - \mathcal{P}^\pi h(s,a),
\end{aligned}
$$

where we introduce $h = Q^\pi - Q + (\eta^\pi - \eta)q^\pi$. Using the orthogonality property in (11), we have $\|\mathcal{E}_\pi(\eta,Q)\|^2 = (\eta - \eta^\pi)^2\|e^\pi\|^2 + \|(\mathcal{I} - \mathcal{P}^\pi)h\|^2$ and thus $|\eta - \eta^\pi| \leq \|e^\pi\|^{-1}\|\mathcal{E}_\pi(\eta,Q)\|$. Furthermore, we have

$$
\begin{aligned}
&\|(\mathcal{I} - \mathcal{P}^\pi)(Q - Q^\pi)\| = \|\mathcal{E}_\pi(\eta,Q) + (\eta - \eta^\pi)\| \\
&\leq \|\mathcal{E}_\pi(\eta,Q)\| + |\eta - \eta^\pi| \leq (1 + \|e^\pi\|^{-1})\|\mathcal{E}_\pi(\eta,Q)\|.
\end{aligned}
$$

Note that $\|e^\pi\| = \|w^\pi\|/(1+\sigma_\pi^2) = (1+\sigma_\pi^2)^{-1/2}$ (since $\|w^\pi\|^2 = 1 + \sigma_\pi^2$) and thus we have

$$
\begin{aligned}
&|\eta - \eta^\pi| \leq \sqrt{1+\sigma_\pi^2}\|\mathcal{E}_\pi(\eta,Q)\| \\
&\|(\mathcal{I} - \mathcal{P}^\pi)(Q - Q^\pi)\| \leq (1 + \sqrt{1+\sigma_\pi^2})\|\mathcal{E}_\pi(\eta,Q)\|.
\end{aligned}
$$

□

Lemma B.4. *For $t = 1, \dots, T+1$, denote by ν_t the distribution of states S_t at time t in the trajectory $\mathcal{D}$ and $\bar{\nu}_t = (1/t)\sum_{s=1}^{t} \nu_s$ be the average distribution up to time t. For any state-action function Q, we have*

$$\|U^\pi(Q) - U^\pi\| \le \sqrt{1 + \frac{1}{p_{\min}}\big(1 + \frac{1}{T}\Big\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\Big\|_\infty\big)}\|Q - Q^\pi\|.$$

Proof of Lemma B.4.

$$\begin{aligned}
\|U^\pi(Q) - U^\pi\| &= \sqrt{\mathbb{E}[(1/T)\sum_{t=1}^{T}(U^\pi(S_t, A_t, S_{t+1}; Q) - U^\pi(S_t, A_t, S_{t+1}))^2]} \\
&\le \sqrt{\mathbb{E}[(1/T)\sum_{t=1}^{T}(Q(S_t, A_t) - Q^\pi(S_t, A_t))^2]} \\
&\quad + \sqrt{\mathbb{E}\Big[(1/T)\sum_{t=1}^{T}\big(\sum_{a}\pi(a|S_{t+1})(Q(S_{t+1}, a) - Q^\pi(S_{t+1}, a'))\big)^2\Big]}.
\end{aligned} \tag{B.13}$$

Note that the second term involves the distribution of the last states, S_{T+1} which is not present in the first term. It is easy to see that $\big\|\frac{d\bar{\nu}_{T+1}}{d\bar{\nu}_T}\big\|_\infty \le 1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty$. Thus we have

$$\begin{aligned}
&\mathbb{E}[(1/T)\sum_{t=1}^{T}(\sum_{a'}\pi(a'|S_{t+1})(Q(S_{t+1}, a') - Q^\pi(S_{t+1}, a')))^2] \\
&\le (1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty)\mathbb{E}[(1/T)\sum_{t=1}^{T}(\sum_{a}\pi(a|S_t)(Q(S_t, a) - Q^\pi(S_t, a)))^2] \\
&\le (1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty)(1/T)\sum_{t=1}^{T}\mathbb{E}\big[\sum_{a}\pi(a|S_t)(Q(S_t, a) - Q^\pi(S_t, a))^2\big] \\
&\le (1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty)(1/p_{\min})(1/T)\sum_{t=1}^{T}\mathbb{E}\big[\sum_{a}\pi_t^b(a|H_t)(Q(S_t, a) - Q^\pi(S_t, a))^2\big] \\
&= (1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty)(1/p_{\min})\mathbb{E}[(1/T)\sum_{t=1}^{T}(Q(S_t, A_t) - Q^\pi(S_t, A_t))^2] \\
&= (1 + (1/T)\|\frac{d\nu_{T+1}}{d\bar{\nu}_T}\|_\infty)(1/p_{\min})\|Q - Q^\pi\|^2.
\end{aligned}$$

We obtain the desired bound by combining the above with (B.13). $\square$

Lemma B.5. *Suppose Assumption 7 holds. Then, for any state-action function $\tilde{Q}$ such that $\mu_\pi(\tilde{Q}) = 0$, we have $\|\tilde{Q} - Q^\pi\| \le 2\left(1 + C_4\beta/(1-\beta)\right)\|(\mathcal{I} - \mathcal{P}^\pi)(\tilde{Q} - Q^\pi)\|$*

Proof of Lemma B.5. Let $\mathcal{P}^\pi_t := (\mathcal{P}^\pi)^t$. Pick t such that $C_4\beta^t \le 1/2$, then we have

$$\begin{aligned}
\|\tilde{Q} - Q^\pi\| &\le \|(\mathcal{I} - \mathcal{P}^\pi_t)(\tilde{Q} - Q^\pi)\| + \|\mathcal{P}^\pi_t(\tilde{Q} - Q^\pi)\| \\
&\le \|(\mathcal{I} - \mathcal{P}^\pi)(\tilde{Q} - Q^\pi)\| + C_4\beta^t\|\tilde{Q} - Q^\pi\| \\
&\le \|(\mathcal{I} - \mathcal{P}^\pi)(\tilde{Q} - Q^\pi)\| + (1/2)\|\tilde{Q} - Q^\pi\|.
\end{aligned}$$

Now we have

$$\begin{aligned}
\|\tilde{Q} - Q^\pi\| &\le 2\|(\mathcal{I} - \mathcal{P}^\pi_t)(\tilde{Q} - Q^\pi)\| \\
&= 2\|(\mathcal{I} - \mathcal{P}^\pi_1 + \mathcal{P}^\pi_1 - \mathcal{P}^\pi_2 + \cdots + \mathcal{P}^\pi_{t-1} - P^\pi_t)(\tilde{Q} - Q^\pi)\| \\
&\le 2(\|(\mathcal{I} - \mathcal{P}^\pi_1)(\tilde{Q} - Q^\pi)\| + \|(\mathcal{P}^\pi_1 - \mathcal{P}^\pi_2)(\tilde{Q} - Q^\pi)\| + \cdots \|(\mathcal{P}^\pi_{t-1} - P^\pi_t)(\tilde{Q} - Q^\pi)\|).
\end{aligned}$$

For simplicity, let $h = (\mathcal{I} - \mathcal{P}^\pi)(\tilde{Q} - Q^\pi)$. Clearly we have $\mu_\pi(h) = 0$. Now for each k,

$$\|(\mathcal{P}^\pi_{k-1} - \mathcal{P}^\pi_k)(\tilde{Q} - Q^\pi)\| = \|\mathcal{P}^\pi_{k-1}(\mathcal{I} - \mathcal{P}^\pi)(\tilde{Q} - Q^\pi)\| = \|\mathcal{P}^\pi_{k-1}h\| \le C_4\|h\|\beta^{k-1}.$$

Thus

$$\begin{aligned}
\|\tilde{Q} - Q^\pi\| &\le 2(\|h\| + C_4\|h\|\beta + C_4\|h\|\beta^2 + \cdots + C_4\|h\|\beta^{t-1}) \\
&\le 2(\|h\| + C_4\|h\|\frac{\beta}{1-\beta}) = \|h\|\left(2 + 2C_4\beta/(1-\beta)\right).
\end{aligned}$$

□

C Rationale of the ratio estimator

Recall that the estimator $\hat{e}^\pi_n = \tilde{g}_{n,\pi}(\hat{q}^\pi_n)$ where $\hat{q}^\pi_n$ is given by

$$\hat{q}^\pi_n = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathbb{P}_n[(1/T)\sum_{t=1}^{T} \tilde{g}^2_{n,\pi}(S_t, A_t; q)] + \tilde{\lambda}_n J_1^2(q), \tag{C.1}$$

and for any $q \in \mathcal{Q}$, $\tilde{g}_{n,\pi}(q) = \text{argmin}_{g\in\mathcal{G}} \mathbb{P}_n\{(1/T)\sum_{t=1}^{T}[1-q(S_t,A_t)+\sum_{a'}\pi(a'|S_{t+1})q(S_{t+1},a')-g(S_t,A_t)]^2\} + \tilde{\mu}_n J_2^2(g)$. The estimator $\tilde{q}^\pi$ solves a coupled optimization, similar to $(\hat{\eta}_n^\pi, \hat{Q}_n^\pi)$ in (9). Below we argue the rationale behind this estimator. First, re-arranging (12) gives

$$e^\pi(s,a) = 1 - \tilde{q}^\pi(s,a) + \mathbb{E}\Big[\sum_{a'}\pi(S_{t+1},a')\tilde{q}^\pi(S_{t+1},a') \mid S_t = s, A_t = a\Big]. \tag{C.2}$$

The orthogonality of e^π in (11) implies that

$$\mathbb{E}\Big[(1/T)\sum_{t=1}^{T}\Big(1-\tilde{q}^\pi(S_t,A_t)+\mathbb{E}\big[\sum_{a'}\pi(S_{t+1},a')\tilde{q}^\pi(S_{t+1},a') \mid S_t, A_t\big]\Big)\times \Big(q(S_t,A_t)-\mathbb{E}\big[\sum_{a'}\pi(a'|S_{t+1})q(S_{t+1},a')|S_t,A_t\big]\Big)\Big] = 0.$$

Under the assumption that $\tilde{q}^\pi \in Q$, it is straightforward to verify that $\tilde{q}^\pi$ minimizes $J(q) = \mathbb{E}[(1/T)\sum_{t=1}^{T}\big(1-q(S_t,A_t)+\mathbb{E}[\sum_{a'}\pi(a'|S_{t+1})q(S_{t+1},a')|S_t,A_t]\big)^2]$ over $q \in \mathcal{Q}$. Clearly, $\tilde{g}_{n,\pi}(q)$ approximates $1 - q(s,a) + \mathbb{E}[\sum_{a'}\pi(a'|S_{t+1})q(S_{t+1},a')|S_t = s, A_t = a]$ as a function of (s,a). The first term in the objective function of (A.3), i.e., $\mathbb{P}_n[(1/T)\sum_{t=1}^{T}\tilde{g}_{n,\pi}^2(S_t,A_t;q)]$, is a proxy of $J(q)$ and thus $\tilde{q}_n^\pi$ approximates $\tilde{q}^\pi$. Based on (C.2), we can see that $\tilde{g}_{n,\pi}(\tilde{q}_n^\pi)$ estimates e^π.

D Computation

Below we derive the closed-form solution of the estimators as well as the asymptotic variance when $\mathcal{Q}$ and $\mathcal{G}$ are both Reproducing Kernel Hilbert Space (RKHS). We consider the general setting discussed in Appendix where there is time-invariant information in the state and the average reward is modeled as $f(z)^\top\beta$ for some feature vector f. Setting $f = 1$ gives the special case where there is no time-invariant state. Without loss of generality, we denote the training data simply by $\mathcal{D} = \{Z_h, X_h, A_h, X_h', R_h\}_{h=1}^{N}$ where h indexes the tuple of transition sample with baseline in the training set, Z_h is the corresponding baseline, X_h and X_h' is the current and next time-varying state and R_h is the reward. Let $U_h = (Z_h, X_h, A_h)$ be the state-action pair, $S_h' = (Z_h, X_h')$ and $U_h' = (Z_h, X_h, A_h, X_h')$.

It is more common to form the RKHS for the function of state. Suppose the kernel function for the state function is given by $k_0(s_1, s_2)$, $s_1, s_2 \in \mathcal{S}$. To incorporate the action (assuming binary below for simplicity), one can define $k((s_1, a_1), (s_2, a_2)) = \mathbb{1}_{\{a_1=a_2\}} k_0(s_1, s_2)$. That is, we model each arm separately with each $Q(\cdot, a)$ in the RKHS with kernel k_0. Alternatively, one can also model the baseline value $Q(s, 0)$ and the difference, i.e., $Q(s, 1) - Q(s, 0)$ with two kernels k_0, k_1. In this case, one can define the kernel by $k((s_1, a_1), (s_2, a_2)) = k_0(s_1, s_2) + \mathbb{1}_{\{a_1=a_2=1\}} k_1(s_1, s_2)$.

Recall that we need to restrict the function space $\mathcal{Q}$ such that $Q((z, x^*), a^*) = 0$ for all $Q \in \mathcal{Q}$. For an arbitrary kernel function k_0 on $\mathcal{S} \times \mathcal{A}$, we can always transform it into k such that this assumption is satisfied. In particular, define $k(U_1, U_2) = k_0(U_1, U_2) - k_0((Z_1, x^*, a^*), U_2) - k_0(U_1, (Z_2, x^*, a^*)) + k_0((Z_1, x^*, a^*), (Z_2, x^*, a^*))$. It can be seen that the induced RKHS by k satisfies the condition. Suppose the kernel function for $\mathcal{Q}$ and $\mathcal{G}$ are given by $k(\cdot, \cdot), l(\cdot, \cdot)$. Denote the inner product by $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$, respectively.

Note that the minimizer (A.3) is a regression problem. And it is well known that the solution is given by $\hat{g}_n(\cdot, \cdot; \eta, Q) = \sum_{h=1}^N l(U_i, \cdot)\theta_h(\eta, Q)$ where $\theta(\eta, Q) = (L + \mu I_N)^{-1} \delta_N^\pi(\eta, Q)$, where $\mu = \mu_n N$ and $\delta_N^\pi(\eta, Q) = (\delta^\pi(U_h'; \eta, Q))_{h=1}^N$ is the vector of TD error where $\delta(U_h'; \eta, Q) = R_h + \sum_{a'} \pi(a'|S_h') Q(S_h', a') - f(Z_h)^\top \beta - Q(S_h, A_h)$. In addition, each TD error can be written as $\delta(U'; \beta, Q) = R - f(Z)^\top \beta - \langle Q, \tilde{k}_{U'}^\pi \rangle_{\mathcal{G}}$, where

$$\tilde{k}_{U'}^\pi(\cdot) = k(U, \cdot) - \sum_{a'} \pi(a'|S') k((S', a'), \cdot) \in \mathcal{Q}.$$

We show that $\hat{Q}_n^\pi$ in (A.2) must stay in the linear span: $\{\sum_{h=1}^N \alpha_h \tilde{k}_{U_h'}^\pi(\cdot) : \alpha_h \in \mathbb{R}, h = 1, \ldots, N\}$. To see this, suppose $Q = Q_0 + \Delta$ where $Q_0 = \sum_{h=1}^N \alpha_h \tilde{k}_{U_h'}^\pi$ and $\Delta \in \mathcal{Q}$ satisfies $\langle \Delta, \tilde{k}_{U_h'}^\pi \rangle_{\mathcal{Q}} = 0$ for all $h = 1, \ldots, N$. For the first term in (A.2), denoted by $L(\beta, Q)$, we have $L(\beta, Q) = L(\beta, Q_0)$, using the fact that each TD error is unchanged by adding Δ. And the second term, we have $\|Q\|_{\mathcal{Q}}^2 = \|Q_0\|_{\mathcal{Q}}^2 + \|\Delta\|_{\mathcal{Q}}^2$ due to the orthogonality of Δ. Thus the minimizer must have $\Delta = 0$. Using this finite representer property, we can find $(\hat{\beta}, \hat{\alpha})$ by

$$(\hat{\beta}^\pi, \hat{\alpha}^\pi) = \operatorname*{argmin}_{\beta \in \mathbb{R}^p, \alpha \in \mathbb{R}^N} (R_N - F\beta - \tilde{K}^\pi \alpha)^\top M (R_N - F\beta - \tilde{K}^\pi \alpha) + \lambda \alpha^\top \tilde{K}^\pi \alpha,$$

where $R_N = (R_h)_{h=1}^N$, $\tilde{K}^\pi = (\langle \tilde{k}_{U_h'}^\pi, \tilde{k}_{U_k'}^\pi \rangle_{\mathcal{Q}})_{k,h=1}^N$, $M = (L + \mu I_N)^{-1} L^2 (L + \mu I_N)^{-1}$, $F =$

$(f(Z_h))_{h=1}^N$ and $\lambda = \lambda_n N$. Note that we have $\tilde{K}\,[h,k]$ can be calculated by

$$\langle \tilde{k}^\pi_{U'_h}, \tilde{k}^\pi_{U'_k} \rangle_{\mathcal{Q}} = k(U_h, U_k) - \sum_{a'} \pi(a'|S'_h) k((S'_h, a'), U_k) - \sum_{a'} \pi(a'|S'_k) k((S'_k, a'), U_h)$$
$$+ \sum_{a'_h} \sum_{a'_k} \pi(a'_h|S'_h) \pi(a'_k|S'_k) k((S'_h, a'_h), (S'_k, a'_k))$$

Taking derivative implies $(\hat{\beta}^\pi, \hat{\alpha}^\pi)$ solves

$$F^\top M F \beta = F^\top M (R_N - \tilde{K}^\pi \alpha)$$
$$(M\tilde{K}^\pi + \lambda I_N)\alpha = M(R_N - F\beta)$$

Similarly, we can find the closed-form solution of $\hat{e}^\pi = (\hat{e}^\pi_k)_{k=1}^p$: for each k, $\hat{e}_k = \sum_{h=1}^N \theta_h l(U_h, \cdot)$ where $\theta = (\theta_h)_{h=1}^N = (L + \mu I_N)^{-1}(F_k - \tilde{K}^\pi \hat{\alpha}^\pi)$ and $\hat{\alpha}^\pi$ solves $(M\tilde{K}^\pi + \lambda I_N)\alpha = MF_k$. Here F_k is the k-the column of F.