HYBRID RL FOR OPTIMAL DYNAMIC TREATMENT REGIMES: IMPACT OF UNMEASURED CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid Reinforcement Learning (RL), where an agent learns from both an offline dataset and online explorations in an unknown environment, has attracted significant recent interest. Existing hybrid RL studies have largely focused on two cases for the offline data-generating process, either imposing no restriction on the degree of unmeasured confounding or assuming that such confounding is entirely absent. Many applications, however, lie between these extremes, with unmeasured confounding present but bounded in magnitude. In this paper, we investigate how the degree of unmeasured confounding in the offline data impacts online learning for optimal dynamic treatment regimes. We show that the offline data never hurt the guarantee of the purely online setting. Moreover, the guarantee improves monotonically as the assumed degree of unmeasured confounding decreases. When the offline data are fully unconfounded, the guarantee is at least as strong as those in the purely online, unrestricted, and bounded unmeasured confounding settings.

1 INTRODUCTION

A dynamic treatment regime (DTR) (Robins, 2004; Kosorok & Moodie, 2015; Bareinboim et al., 2024) consists of a sequence of decision rules, one per stage of intervention, each mapping a patient's history to a treatment assignment. These regimes are particularly effective for managing chronic disorders and provide an attractive framework for personalized decision-making in longitudinal settings. Unlike standard episodic RL, the underlying process is not Markov: earlier stages directly affect future treatments and outcomes, even when all covariates (i.e., states) and treatments are observed. Importantly, in many clinical applications the patient's covariates are only partially observed, for instance, health-related factors such as disease severity or frailty are often unmeasured (Mayer et al., 2019). These unmeasured health factors can also violate the Markov property and introduce unobserved (a.k.a. unmeasured) confounding among treatments, covariates, and outcomes.

The strength of unmeasured confound factors can vary across (clinical) settings. Under unrestricted (i.e., arbitrary) confounding, no information is available on how strongly unmeasured factors influence clinical decisions (Manski, 1990); under bounded confounding, domain experts (e.g., epidemiologists and physicians) may provide knowledge of how strongly unmeasured factors influence treatments (Cinelli & Hazlett, 2020); and in the fully unconfounded case, treatments do not depend on unmeasured factors, such as when the treatments are randomized (Lavori & Dawson, 2004).

A substantial literature has developed methods for estimating optimal DTRs in the offline setting, including A-learning (Murphy, 2003), Q-learning (Murphy, 2005a), outcome-weighted learning approaches (Zhao et al., 2015), and Bayesian approaches (Murray et al., 2018). All of these works assume no unmeasured confounding (i.e. the best case). Recently, a growing literature has studied online and hybrid learning of DTRs, in tabular settings with offline dataset (Zhang & Bareinboim, 2019; 2020) and in the purely online setting (Hu & Kallus, 2022). These works considered the offline data as having arbitrary confounding (i.e. the worst case). No works have investigated online DTR learning with offline data subject to different levels of unmeasured confounding.

In this paper, we study the problem of learning optimal policies for DTRs in a hybrid setting, under varying degrees of unmeasured confounding in the offline dataset. More specifically, our contributions are as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of regret guarantees stated up to logarithmic factors $\widetilde{\mathcal{O}}(\cdot)$. POMPDP and MDP settings are different than the DTR setting but are included for reference. Notation: T episodes; no. of stages K[1]; feature dimension d_K at stage K; N_{off} offline trajectories; per-stage spaces $\mathcal{S}_k, \mathcal{A}_k$; estimated offline widths W_k^{off} (Section 3); $W_{k,\text{off}}^{\Gamma}$ (Section 4); $\|\mathbf{C}\|_1$ the causal bound widths; offline coverage coefficients γ_{off} and c_{off}; margin exponent α; the state variables after graphical reduction $\mathbf{S}'$; $\mathcal{D}_k$ and $\mathbf{C}_k$ are joint state-action domain and causal bounds width related with $\mathbf{S}'$; $\mathcal{X}_{\text{off}}, \mathcal{X}_{\text{on}}$ a partition of the state action space; Δ the information gain from the online episodes relative to the confounded offline data; $\mathbf{O}$ is the space of observations. † At $\Gamma = 1$ the offline data are unconfounded. $*$ Considers settings that allow point identification. § Considers Bayesian regret. UC denotes unmeasured (arbitrary or bounded) confounding. ‡ Linearity is assumed for the optimal Q-functions. N/A denotes the DTR setting without offline data.

Model	Algorithm	Func. Class	Setting	UC	Regret
POMDP	(Azizi et al., 2024)	Tabular	Hybrid	✓	$\lvert\mathbf{O}\rvert\sqrt{\max_k \lvert\mathcal{A}_k\rvert K^3 T}$ §
MDP	(Wang et al., 2021a)	Linear	Hybrid	✓*	$\Delta \cdot K^2\sqrt{T}$
	(Tan et al., 2024)	Linear	Hybrid	✗	$\inf_{\mathcal{X}_{\text{off}},\mathcal{X}_{\text{on}}}\{c_{\text{off}} T\sqrt{\frac{K^3}{N_{\text{off}}}} + \sqrt{K^3 T}\}$
DTR	(Zhang & Bareinboim, 2019)	Tabular	Hybrid	✓	$\sqrt{T} + \min\{\sqrt{T\prod_k \lvert\mathcal{S}_k\rvert\lvert\mathcal{A}_k\rvert}, \lvert\lvert\mathbf{C}\rvert\rvert_1 T\}$
	(Zhang & Bareinboim, 2020)	Tabular	Hybrid	✓	$\sqrt{T} + \sum_{S_k\in\mathbf{S}'} \min\{\sqrt{T\lvert\mathcal{D}_k\rvert}, \lvert\lvert\mathbf{C}_k\rvert\rvert_1 T\}$
	(Hu & Kallus, 2022)	Linear ‡	Online	N/A	$T^{(1-\alpha)/2}$ [2]
	Corollary 1	Linear	Online	N/A	$d_K^{3/2} K\sqrt{T}$
	Theorem 1	Linear	Hybrid	✓	$\sqrt{TK} + \min\{d_K^{3/2} K\sqrt{T},\ T\sum_k W_k^{\text{off}}\}$
	Theorem 2	Linear	Hybrid	✓†	$\sqrt{TK} + \min\{d_K^{3/2} K\sqrt{T},\ T\sum_k W_{k,\text{off}}^{\Gamma}\}$
	Theorem 3	Linear	Hybrid	✗	$\sqrt{TK} + \min\{d_K^{3/2} K\sqrt{T}, T\sum_k W_{k,\text{off}}^{\Gamma=1}, TK\sqrt{\frac{\gamma_{\text{off}}}{N_{\text{off}}}}\}$

(i) We develop hybrid optimistic Q-learning algorithms for optimal dynamic treatment regimes, covering offline dataset that are arbitrarily confounded, boundedly confounded, or fully unconfounded. We show that the offline dataset never degrade the purely online guarantee.

(ii) For unrestricted confounding, Theorem 1 improves on existing hybrid DTR results in its dependence on the cardinality of the state and action spaces (Table 1). For bounded confounding, Theorem 2 reduces exactly to Theorem 1 when the unmeasured confounding is arbitrary, and the guarantee improves as the assumed degree of confounding decreases (Remark 2). For fully unconfounded offline data, Theorem 3 is never worse than Theorems 1 and 2, with a further improvement available when the offline feature vectors span the full feature space (Remark 3).

We also validate our proposed methods on synthetic data (see Appendix H).

1.1 Related Works

The closest prior works are (Zhang & Bareinboim, 2019; 2020), who first studied online learning of the optimal DTR from offline observational data subject to unrestricted confounding. Our work improves on theirs in three aspects. (1) They use model-based methods for the tabular settings, which require estimating all transition probabilities and mean rewards. We note that since DTR environments are not Markov, the estimates for transition probabilities and the mean rewards are conditioned on the full state-action sequence history. We impose a linear structure on the Q-function, and our regret improves on theirs in its dependence on the cardinality of state-action space for low-dimensional settings. (2) They treat the causal bounds on the transition kernels and mean rewards as exact inputs to their regret analysis. Effectively, they consider infinite offline data while we consider only a finite offline data set. Thus, unlike in their work, we need to account for the uncertainty of the causal bounds learned from the offline data in our algorithm and analysis. (3) In terms of confounding, they only investigated the worst-case setting of unrestricted confounding. We consider

[1] The number of stages K in DTRs is typically small, often $K = 2$ (Chakraborty et al., 2010; Kasari et al., 2014; Laber et al., 2014), while general RL episodes can be large even for simple problems.

[2] (Hu & Kallus, 2022) achieves $\log(T)$ when $\alpha \geqslant 1$ by assuming a margin condition.

that worst-case setting as well as the intermediate and more benign settings of bounded confounding and fully uncounfounded offline data, respectively, covering the full range of confounding levels.

Hu & Kallus (2022) studied the two-armed two-stage DTR bandit problem in a purely online setting. Their method builds on Q-learning and they impose a margin condition on the gaps between optimal Q-functions at each stage. We do not adopt such an assumption; when extended to K stages with $|\mathcal{A}|$ treatments at every stage, it requires $\mathcal{O}(|\mathcal{A}|^K)$ inequalities to hold which grows exponentially with the horizon K. In the Markov decision process (MDP) setting, Wang et al. (2021a) studied episodic linear MDPs with an offline dataset, in settings where the transition probabilities and expected rewards are point-identified from the observational data via a backdoor or frontdoor adjustment. In our Sections 3 and 4, these quantities are only partially identified. In the partially observable MDP (POMDP) setting, Azizi et al. (2024) studied a hybrid setting with offline data. They model the state as consisting of an observed component and an unobserved one, with the latter acting as an unmeasured confounder. Their method adopts a Bayesian framework and assumes that the unmeasured confounders are i.i.d. across stages, which we do not. See Appendix B for more related works.

Preliminaries. We use capital letters to denote variables (A) and small letters for their values (a). Let $\mathcal{A}$ represent the domain of A and $|\mathcal{A}|$ its cardinality. We write $P(a)$ to represent the probabilities $P(A = a)$. $\overline{A}_k$ stands for a sequence $\{A_1, \dots, A_k\}$ ($\varnothing$ if $k < 1$), and $\overline{\mathcal{A}}_k$ represents its domain, i.e., $\mathcal{A}_1 \times \cdots \times \mathcal{A}_k$. A structural causal model (SCM) (Pearl, 2009, Ch. 7) M is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$, where $\boldsymbol{U}$ is a set of exogenous (unobserved or latent) variables, $\boldsymbol{V}$ is a set of endogenous (observed) variables, $\boldsymbol{F}$ is a set of structural functions, and $P(\boldsymbol{U})$ is a distribution over the latent variables. An intervention, denoted by $do(A_i = a)$ ($do(a)$, for short), represents a model manipulation where the values of a variables $A_i \in \boldsymbol{V}$ are set fixed to a regardless of how their values are ordinarily determined by $f_i \in \boldsymbol{F}$. For the outcome Y, we use the counterfactual distribution $P(Y_{A_i=a} = y)$ ($P(Y_a)$, for short) to denote the causal effect of the intervention $do(A_i = a)$ on the outcome Y.

2 Problem Formulation

In this section, we formalize the problem of online optimization in DTRs with (finite) offline observational data. We start by defining DTRs using the semantics of structural causal models (SCMs).

2.1 Dynamic Treatment Regimes

A dynamic treatment regime (DTR) (Murphy et al., 2001; Murphy, 2003; Bareinboim et al., 2024) is a tuple $\langle \mathcal{M}^*, \Pi, Y \rangle$. The environment $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$ is a structural causal model with endogenous variables $\boldsymbol{V} = \{S_1, \dots, S_K, A_1, \dots, A_K, Y\}$, where S_k and A_k denote the state and action at stage k, respectively, Y is the terminal outcome, and K is the number of stages. Both the state space $\{\mathcal{S}_k\}_{k=1}^K$ and action space $\{\mathcal{A}_k\}_{k=1}^K$ are finite. Values of exogenous variables $\boldsymbol{U}$ are drawn from the distribution $P(\boldsymbol{U})$. The policy space Π is a collection of treatment policies $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, where each decision rule $\pi_k : \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1} \to \Delta(\mathcal{A}_k)$ maps the history at stage k to a distribution over actions. A policy $\boldsymbol{\pi}$ defines a sequence of stochastic interventions (Pearl, 2009; Correa & Bareinboim, 2020) $do(A_1 \sim \pi_1(A_1|s_1)), ..., do(A_K \sim \pi_K(A_K|\bar{s}_K, \bar{a}_{K-1}))$. $Y \in [0, 1]$ is the outcome at the final stage K. Figure 1 gives a graphical representation of a two-stage DTR.

For each stage $k < K$, the optimal Q and value functions are defined by the backward recursion (Murphy, 2003; Schulte et al., 2014; Tsiatis et al., 2019), specifically,

$$Q_k^*(\bar{s}_k, \bar{a}_k) = \sum_{s_{k+1} \in \mathcal{S}_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \max_{a_{k+1} \in \mathcal{A}} Q_{k+1}^*(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}), \quad (1)$$

$$V_k^*(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \mathcal{A}} Q_k^*(\bar{s}_k, \bar{a}_{k-1}, a_k). \quad (2)$$

$P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$ is the transition distribution at stage k under the intervention $do(\bar{A}_k = \bar{a}_k)$. At the terminal stage $k = K$, the expected reward in $\mathcal{M}^*$ is given by $Q_K^*(\bar{s}_K, \bar{a}_K) = \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$, where $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ is the expected outcome under $do(\bar{A}_K = \bar{a}_K)$.

We study the problem of learning an optimal DTR by combining offline data with online interaction. Let $\mathcal{D}^{\text{off}} = \{(\bar{s}_K^{(n)}, \bar{a}_K^{(n)}, Y^{(n)})\}_{n=1}^{N_{\text{off}}}$ consists of N_{off} trajectories drawn i.i.d. from the observational distribution induced in $\mathcal{M}^*$ by a behavior policy $\boldsymbol{\pi}_b = \{\pi_{b,1}, \pi_{b,2}, ..., \pi_{b,K}\}$. The

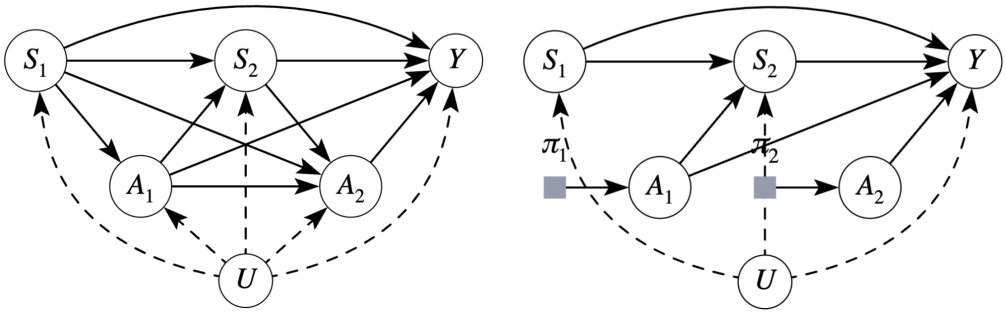


Figure 1: Causal diagrams of: (a) a two-stage DTR; (b) a two-stage DTR under sequential interventions $do(A_1 \sim \pi_1(A_1|S_1), A_2 \sim \pi_2(A_2|S_1, A_1, S_2))$. Note that the system is non-Markovian even when there is no unmeasured confounding ($\boldsymbol{U} = \varnothing$).

learner interacts with the environment $\mathcal{M}^*$ for T episodes. At each episode $t \in [T]$, the agent follows a policy $\boldsymbol{\pi}_t = \{\pi_{t,1}, \pi_{t,2}, ..., \pi_{t,K}\}$, then observes and stores the resulting trajectory $(s_1^t, a_1^t; s_2^t, a_2^t; \ldots; s_K^t, a_K^t, y_K^t)$. The goal for the agent is to optimize an unknown SCM ($\mathcal{M}^*$) that maximizes the cumulative rewards. The performance of the agent is evaluated through the (pseudo) regret, that is, $R(T) = \sum_{t=1}^{T} \left[V_1^*(s_1^t) - V_1^{\boldsymbol{\pi}_t}(s_1^t)\right]$, where $V_1^{\boldsymbol{\pi}_t}(s_1^t)$ is the expected reward obtained by starting from s_1^t and following $\boldsymbol{\pi}_t$ through all K stages of $\mathcal{M}^*$. See Example 1 in Appendix C for a medical treatment example.

2.2 Linear Value Function Approximation

Throughout this paper, we consider linear value function approximation, where the terminal mean outcome and the Bellman backups of value functions are linear in known features. The standard approach in linear RL is to impose linearity directly on the transition law and the terminal mean outcome, as in (Jin et al., 2020b). In a DTR, the feature vectors are history-dependent; this forces the feature dimension d_k to be at least $\prod_{i=1}^{k} |\mathcal{S}_i||\mathcal{A}_i|$ at every stage k (see Lemma 1 for more details), which grows exponentially with the number of stages, making the regret bound prohibitive. We therefore adapt the optimistic closure condition of (Wang et al., 2021b) to our DTR setting, imposing it over a specified class of value functions, permitting smaller feature dimensions.

Definition 1. For the known feature maps $\psi_k : \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k \to \mathbb{R}^{d_k}$, the class of linear functions at stage $k \in [K]$ of the DTR is $\mathcal{Q}_k := \left\{(\bar{s}_k, \bar{a}_k) \mapsto \langle \psi_k(\bar{s}_k, \bar{a}_k)^{3}, \theta \rangle : \ \theta \in \mathbb{R}^{d_k}, \ \|\theta\|_2 \leqslant \sqrt{d_k}\right\}$.

Following (Wang et al., 2021b; Agarwal et al., 2023), we define an enlarged function class $\mathcal{Q}_k^{\text{enl}}$ that contains optimistic functions built from $\mathcal{Q}_k$. Let $\|x\|_A := \sqrt{x^\top A x}$ for a positive semi-definite matrix $A \succeq 0$, and $\mathfrak{P}_k := \left\{\{\vartheta \in \mathbb{R}^{d_k} : L(\bar{s}_k, \bar{a}_k) \leqslant \langle \psi_k, \vartheta \rangle \leqslant U(\bar{s}_k, \bar{a}_k), \ \forall (\bar{s}_k, \bar{a}_k)\} \ \middle| \ L, U : \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k \to [-\infty, \infty]\right\}$, which is a class of closed convex subsets of $\mathbb{R}^{d_k}$ and contains $\mathbb{R}^{d_k}$. Define

$$\mathcal{Q}_k^{\text{enl}} := \Big\{(\bar{s}_k, \bar{a}_k) \mapsto 0 \vee \big(1 \wedge (\langle \psi_k, \theta \rangle + \|\psi_k\|_A) \wedge (\langle \psi_k, \theta' \rangle + \|\psi_k\|_{A'}) \wedge \sup_{\vartheta \in \mathcal{P}} \langle \psi_k, \vartheta \rangle\big) \ \Big| \ \theta, \theta' \in \mathbb{R}^{d_k}, \ A, A' \succeq 0, \ \mathcal{P} \in \mathfrak{P}_k\Big\}, \tag{3}$$

where $a \wedge b := \min\{a, b\}$, $a \vee b := \max\{a, b\}$. The class $\mathcal{Q}_k^{\text{enl}}$ contains every function in $\mathcal{Q}_k$ (Definition 1) with values in $[0, 1]$, by taking $\theta' = \theta$, $A = A' = 0$ and $L = U = \langle \psi_k, \theta \rangle$. The value function class is defined as $\mathcal{V}_k := \{(\bar{s}_k, \bar{a}_{k-1}) \mapsto \max_{a_k \in \mathcal{A}_k} Q(\bar{s}_k, (\bar{a}_{k-1}, a_k)) : \ Q \in \mathcal{Q}_k^{\text{enl}}\}$.

Assumption 1. A DTR $\langle \mathcal{M}^*, \Pi, Y \rangle$ admits linear value function approximation if the feature maps satisfy $\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2 \leqslant 1$ for every $k \in [K]$ and $(\bar{s}_k, \bar{a}_k)$, and the following two conditions hold.

(i) For stage K, the outcome satisfies $Y \in [0, 1]$, and $(\bar{s}_K, \bar{a}_K) \mapsto \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ belongs to $\mathcal{Q}_K$.

(ii) (*Optimistic closure, adapted from* (Wang et al., 2021b).) For every stage $k \in [K-1]$ and $V \in \mathcal{V}_{k+1}$, $(\bar{s}_k, \bar{a}_k) \mapsto \mathbb{E}_{\bar{a}_k}\left[V\left((\bar{s}_k, S_{k+1}), \bar{a}_k\right) \mid \bar{s}_k\right]$ belongs to $\mathcal{Q}_k$.

[3]For simplicity, we write $\psi_k(\bar{s}_k, \bar{a}_k)$ as ψ_k when the arguments are clear from the context.

In Assumption 1, condition (i) is realizability of the terminal mean reward, and condition (ii) is the optimistic closure assumption of (Wang et al., 2021b), adapted to the DTR setting. Setting their link function to the identity, our enlarged function class $\mathcal{Q}_k^{\text{enl}}$ reduces to theirs, up to normalization, when $\theta' = \theta$, $A' = A$ and $\mathcal{P} = \mathbb{R}^{d_k}$.

Fact 1 (Optimistic closure in tabular DTRs)**.** If $\psi_k(\bar{s}_k, \bar{a}_k) = e_{(\bar{s}_k, \bar{a}_k)} \in \mathbb{R}^{d_k}$ is the standard-basis feature map on $\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$ for every $k \in [K]$, then Assumption 1 holds.

The following result shows that optimistic closure is a strictly weaker condition.

Proposition 1. *For DTRs with fixed feature maps, optimistic closure is strictly weaker than the linear DTR assumption, which adapts the linear MDP assumption of (Jin et al., 2020b) to histories.*

See Appendix C for proofs. In practice, feature maps can be chosen from knowledge of the dynamics rather than value functions (Chang et al., 2022); those making the transition law linear satisfy optimistic closure (Proposition 1), with one-hot features on histories as the extreme (tabular) case.

3 Causal Offline-Online Learning under Unrestricted Confounding

In this section, we introduce the methods and analysis for learning the optimal DTR with finite offline data. We impose no quantitative restrictions on the dependence of the behavior policy $\boldsymbol{\pi}_b$ on the unmeasured variables $\boldsymbol{U}$, conditioned on the observed history (Zhang & Bareinboim, 2019; 2020) The offline treatment assignment mechanism may therefore be arbitrarily confounded. We start by constructing a confidence set for the optimal Q functions from the offline dataset.

3.1 Offline Confidence Set

Under unrestricted unmeasured confounding, the transition distributions and the expected reward are not point identified in general (Pearl, 2009; Hernán & Robins, 2020). To address this challenge, we consider a partial identification approach (Manski, 1990; Balke & Pearl, 1997) for bounding $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$ and $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$. Let $\mathcal{H} := \bigcup_{k=1}^{K} (\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1} \cup \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k)$ denote the collection of histories. Define the offline bonus $b_{\text{off}} = \sqrt{\log(2|\mathcal{H}|/\delta_1)/2N_{\text{off}}}$ for any $\delta_1 \in (0, 1]$. For a trajectory $h \in \mathcal{H}$, let $N_{\text{off}}(h)$ denote the number of offline trajectories consistent with h, and set the empirical frequency $\widehat{\mathbb{P}}(h) = N_{\text{off}}(h)/N_{\text{off}}$. Define the empirical mean rewards $\widehat{m}_K(\bar{s}_K, \bar{a}_K) = \frac{1}{N_{\text{off}}} \sum_{i=1}^{N_{\text{off}}} Y_i \mathbf{1}\{\bar{S}_{i,K} = \bar{s}_K, \bar{A}_{i,K} = \bar{a}_K\}$. Define a confidence region for variables $(\mathbb{P}, m_K)$[4] as

$$\begin{aligned}\mathcal{B}^{\text{off}} = \Big\{ &(\mathbb{P}, m_K) : |\mathbb{P}(h) - \widehat{\mathbb{P}}(h)| \leqslant b_{\text{off}} \ \forall h \in \mathcal{H};\ |m_K(\bar{s}_K, \bar{a}_K) - \widehat{m}_K(\bar{s}_K, \bar{a}_K)| \leqslant b_{\text{off}} \ \forall (\bar{s}_K, \bar{a}_K); \\ &0 < \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_{k-1}) \quad \forall k \in [K-1]; \\ &0 < \mathbb{P}(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_{K-1}); \quad 0 \leqslant m_K(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_K) \Big\}. \end{aligned} \tag{4}$$

Then, leveraging the results in (Zhang & Bareinboim, 2019, Theorem 5 and Corollary 1) and accounting for finite sample uncertainty, we obtain the estimated lower bound $L_K(\bar{s}_K, \bar{a}_K)$ and upper bound $U_K(\bar{s}_K, \bar{a}_K)$ on the optimal Q-functions at stage K as follows:

$$L_K = \max\left\{ \inf_{\mathcal{B}^{\text{off}}} \frac{m_K(\bar{s}_K, \bar{a}_K)}{\zeta(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})}, 0 \right\}; U_K = \min\left\{ \sup_{\mathcal{B}^{\text{off}}} 1 - \frac{\mathbb{P}(\bar{s}_K, \bar{a}_K) - m_K(\bar{s}_K, \bar{a}_K)}{\zeta(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})}, 1 \right\} \tag{5}$$

where $\zeta(\bar{s}_{k+1}, \bar{a}_k; \mathbb{P}) = \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) - \mathbb{P}(\bar{s}_k, \bar{a}_k) + \zeta(\bar{s}_k, \bar{a}_{k-1})$ with $\zeta(s_1) = \mathbb{P}(s_1)$. For $k \in [K-1]$ the estimated bounds on the optimal Q functions are obtained by backward recursion,

$$L_k = \min_{p_k \in \widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k)} \sum_{s \in \mathcal{S}_k} p_k(s) \max_{a \in \mathcal{A}_k} L_{k+1}; \quad U_k = \max_{p_k \in \widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k)} \sum_{s \in \mathcal{S}_k} p_k(s) \max_{a \in \mathcal{A}_k} U_{k+1}, \tag{6}$$

where $\widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k) = \Big\{ p \in \Delta(\mathcal{S}_{k+1}) : \inf_{\mathcal{B}^{\text{off}}} \frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\zeta(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})} \leqslant p(s_{k+1}) \leqslant \sup_{\mathcal{B}^{\text{off}}} \big(1 - \frac{\mathbb{P}(\bar{s}_k, \bar{a}_k) - \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\zeta(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})}\big) \ \forall s_{k+1} \in \mathcal{S}_{k+1} \Big\}$ is the the estimated set of feasible transition laws.

[4] $(\mathbb{P}, m_K)$ is the optimization variable ranging over $\mathcal{B}^{\text{off}}$, and $\widehat{\mathbb{P}}(h)$ and $\widehat{m}_K(\bar{s}_K, \bar{a}_K)$ are defined above.

Proposition 2. *Suppose Assumption 1 holds and $P(\bar{s}_K, \bar{a}_K, y) > 0$ for all $(\bar{s}_K, \bar{a}_K, y)$. Fix $\delta_1 \in (0, 1]$, and let $L_k(\bar{s}_k, \bar{a}_k), U_k(\bar{s}_k, \bar{a}_k), k \in [K]$ be the estimated bounds in Equations (5) and (6). Let*

$$\widehat{C}_k^{\text{off}} = \left\{ \theta_k \in \mathbb{R}^{d_k} : L_k(\bar{s}_k, \bar{a}_k) \leqslant \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k \leqslant U_k(\bar{s}_k, \bar{a}_k), \ \forall (\bar{s}_k, \bar{a}_k) \right\}. \tag{7}$$

Then we have $P\big(\forall\, k \in [K],\ \theta_k^* \in \widehat{C}_k^{\text{off}}\big) \geqslant 1 - \delta_1$.

See proof in Appendix D.1. We next introduce a quantity measuring uncertainty in the offline data.

Definition 2 (Offline width)**.** For $k \in [K]$, the worst-case offline confidence bound width is $W_k^{\text{off}} := \max_{(\bar{s}_k, \bar{a}_k) \in \overline{\mathcal{S}}_k \times \overline{\mathcal{A}}_k} \big[U_k(\bar{s}_k, \bar{a}_k) - L_k(\bar{s}_k, \bar{a}_k)\big]$.

3.2 Learning Algorithm and Theoretical Analysis

In this section, we first lay out the main components of our algorithm. See Algorithm 1 for pseudocode. The algorithm builds on two standard techniques: least-squares value iteration (LSVI) (Bradtke & Barto, 1996; Osband et al., 2016; Jin et al., 2020b) and optimism in the face of uncertainty (OFU) (Auer et al., 2002; Abbasi-yadkori et al., 2011). It combines the pre-estimated offline confidence sets $\widehat{\mathcal{C}}_k^{\text{off}}$ with online data in two phases. *(i) Backward pass:* we estimate $\{\widehat{\theta}_k^t\}_{k=1}^K$, construct the online confidence set $\{\widehat{\mathcal{C}}_k^t\}_{k=1}^K$, and recompute the Q-functions optimistically. *(ii) Forward pass:* treatments are selected greedily and the terminal outcome is revealed at the end of the episode.

Unlike the existing hybrid RL literature (Song et al., 2023; Wagenmaker & Pacchiano, 2023; Tan et al., 2024), because of unmeasured confounding the offline trajectories generated by the same $\mathcal{M}^*$ cannot be pooled into the regression that produces $\{\widehat{\theta}_k^t\}_{k=1}^K$. The parameters are estimated from online samples alone, and the offline data enter Algorithm 1 only through the offline confidence set $\widehat{\mathcal{C}}_k^{\text{off}}$. Computationally, the maximum over the ellipsoid $\widehat{\mathcal{C}}_k^t$ has the closed form $\langle \psi_k, \widehat{\theta}_k^t \rangle + \beta_k \|\psi_k\|_{(\Lambda_k^t)^{-1}}$, and the maximum over the polytope $\widehat{\mathcal{C}}_k^{\text{off}}$ is a linear program, so both branches of Equation 11 are efficient. We present the following results (see Appendix D.2 for the proof):

Theorem 1. *Under Assumption 1, let $\chi := \log\big(\frac{4(c_\beta+1)d_K KT}{\delta}\big)$ and absolute constants $C_1 > C > 0, c_\beta > 0, \lambda > 0$. Set the confidence radii as $\beta_k = C d_K \sqrt{\chi} + \sqrt{\lambda d_k}$. For any $\delta, \delta_1 \in (0, 1]$, then with probability at least $1 - \delta - 2\delta_1$, Algorithm 1 achieves the following result:*

$$R(T) \leqslant 2\sqrt{2TK \log(2/\delta_1)} + \min\left\{ 2C_1 d_K^{3/2} K \sqrt{T \cdot \chi \, \log\big(\lambda + T\big)},\ T \sum_{k=1}^K W_k^{\text{off}} \right\}. \tag{8}$$

Corollary 1 (Purely online setting (see Algorithm 3 in Appendix A.1.))**.** *Under Assumption 1, let χ, C_1, C, c_β, λ and β_k be as in Theorem 1. Then for any $\delta, \delta_1 \in (0, 1]$, with probability at least $1 - \delta - \delta_1$, Algorithm 1 without offline data achieves*

$$R(T) \leqslant 2\sqrt{2TK \log(2/\delta_1)} \ + \ 2C_1 \, d_K^{3/2} \, K \sqrt{T \cdot \chi \, \log(1 + T)}. \tag{9}$$

Remark 1. Corollary 1 is the first branch of the minimum in Theorem 1, so the offline data can never worsen the guarantee and strictly improves it whenever $T \sum_{k=1}^K W_k^{\text{off}} < 2C_1 d_K^{3/2} K \sqrt{T\chi \log(\lambda + T)}$. The worst-case offline width decomposes as $W_k^{\text{off}} = W_k + \widetilde{\mathcal{O}}(1/\sqrt{N_{\text{off}}})$. Here $W_k = \lim_{N_{\text{off}} \to \infty} W_k^{\text{off}}$ is the width of the causal bounds on the optimal Q-function for the true observational distribution, with the bounds for $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$ and $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ exactly those of (Zhang & Bareinboim, 2019, Theorem 5; Corollary 1). Under arbitrary confounding the optimal Q-function is only partially identified, so W_k is not reducible by collecting more offline data, while $\widetilde{\mathcal{O}}(1/\sqrt{N_{\text{off}}})$ accounts for estimation uncertainty and vanishes as $N_{\text{off}} \to \infty$.

4 Causal Offline-Online Learning under Bounded Confounding

In Section 3, the behavior policy π_b generating the offline dataset is allowed to depend on the unmeasured variables arbitrarily. In practice, however, offline trajectories are often collected from an

Algorithm 1 Offline–Online Optimistic Q-Learning for Dynamic Treatment Regimes

Require: Regularization parameter λ, confidence radii $\beta_k,\ \forall k \in [K]$, , and precomputed offline confidence sets $\widehat{\mathcal{C}}_k^{\text{off}},\ \forall k \in [K]$.

1: **for** episode $t = 1, \ldots, T$ **do**
2: Receive the initial state s_1^t.
3: **for** step $k = K, K-1, \ldots, 1$ **do**
4: Accumulate all past features at step k: $\Lambda_k^t \leftarrow \sum_{\tau=1}^{t-1} \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)^\top + \lambda I$.
5: Estimate the stage parameters:

$$\widehat{\theta}_K^t \leftarrow (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K Y^\tau; \quad \widehat{\theta}_k^t \leftarrow (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k \max_{a_{k+1}} \widehat{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1}). \tag{10}$$

6: Construct the online confidence set: $\widehat{\mathcal{C}}_k^t \leftarrow \left\{\theta \in \mathbb{R}^{d_k} : \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \leqslant \beta_k\right\}$.
7: Define the optimistic Q-function, $\forall\ (\bar{s}_k, \bar{a}_k)$:

$$\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \leftarrow \max\Big\{ \min\big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ \max_{\theta \in \widehat{\mathcal{C}}_k^{\text{off}}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ 1\big\},\ 0\Big\}. \tag{11}$$

8: **for** step $k = 1, \ldots, K$ **do**
9: Choose treatment $a_k^t \leftarrow \arg\max_{a_k} \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k)$ and observe the next state s_{k+1}^t.
10: Observe the outcome y^t at terminal stage K.

observational study whose treatment assignment mechanism may be subject only to bounded unmeasured confounding. For example, in clinical studies, physicians may use undocumented information, such as subtle symptoms or patient preferences, reflecting latent factors $\boldsymbol{U}$ that affect both treatment assignment and outcomes. Because these signals provide clinically meaningful information about the unmeasured variable $\boldsymbol{U}$, physicians can specify a bound on the extent to which $\boldsymbol{U}$ influences treatment assignment (Hsu & Small, 2013; Kallus et al., 2019; Cinelli & Hazlett, 2020). In this section, we use a marginal sensitivity model (MSM) (Rosenbaum, 1987; Tan, 2006; Kallus & Zhou, 2018; 2020) to control the level of dependence of treatment assignment on unobserved confounders.

4.1 Sensitivity Model

Recall from Section 2 that the offline dataset $\mathcal{D}^{\text{off}} = \{(\bar{s}_K^{(n)}, \bar{a}_K^{(n)}, Y^{(n)})\}_{n=1}^{N^{\text{off}}}$ consists of N^{off} i.i.d. trajectories drawn from the distribution induced by $(\mathcal{M}^\star, \boldsymbol{\pi}_b)$: $\boldsymbol{u} \sim P(\cdot)$; $s_1 \sim P(\cdot \mid \boldsymbol{u})$; $a_k \sim \pi_{b,k}(\cdot \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{u})$; $s_{k+1} \sim P(\cdot \mid \bar{s}_k, \bar{a}_k, \boldsymbol{u})$; $Y \sim P(\cdot \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u})$ [5]. Here $\pi_{b,k}(\cdot \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{u})$ is the behavior policy, in which the clinician actually acted on and cannot be estimated from $\mathcal{D}^{\text{off}}$. We write $\pi_{b,k}(a_k \mid \bar{s}_k, \bar{a}_{k-1}) = \mathbb{E}[\pi_{b,k}(a_k \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{U}) \mid \bar{s}_k, \bar{a}_{k-1}]$ for the propensity an analyst can estimate from $\mathcal{D}^{\text{off}}$. Following (Kallus & Zhou, 2018; 2020), we make an assumption on the degree of confounding in treatment assignment:

Assumption 2 (Sequential marginal sensitivity model)**.** Suppose $\nu \leqslant \pi_{b,k}(a_k \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{u}) \leqslant 1-\nu$ for some $\nu > 0$. There exists $\Gamma \in [1, \infty)$ such that for every stage $k \in [K]$, every $(\bar{s}_k, \bar{a}_k)$, and every $\boldsymbol{u} \in \text{supp}(\boldsymbol{U}) = \{\boldsymbol{u} : P(\boldsymbol{U} = \boldsymbol{u}) > 0\}$,

$$\Gamma^{-1} \leqslant \frac{\pi_{b,k}\big(a_k \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{u}\big)/\big(1 - \pi_{b,k}(a_k \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{u})\big)}{\pi_{b,k}\big(a_k \mid \bar{s}_k, \bar{a}_{k-1}\big)/\big(1 - \pi_{b,k}(a_k \mid \bar{s}_k, \bar{a}_{k-1})\big)} \leqslant \Gamma. \tag{12}$$

In the existing literature (Kallus & Zhou, 2018; Kallus et al., 2019; Frauen et al., 2023), Γ is typically treated as a prespecified parameter informed by domain knowledge. It is not identified from the offline dataset $\mathcal{D}^{\text{off}}$. Taking $\Gamma = 1$ forces the two propensities to coincide, so treatment assignment satisfies sequential randomization; letting $\Gamma \to \infty$ imposes no restriction and recovers the unrestricted confounding of Section 3. Regarding the causal graph interpretation: when $\Gamma = 1$, treatment assignment satisfies $A_k \perp\!\!\!\perp U \mid (\bar{S}_k, \bar{A}_{k-1})$, so the edge $U \to A_k$ is omitted in Figure 2a. For any $\Gamma > 1$ the topology in Figure 2a is unchanged: the edge $U \to A_k$ is present for every $\Gamma > 1$.

[5] The offline data are observational, unlike the interventional online data (Zhang & Bareinboim, 2019; 2020; Wang et al., 2021a).

4.2 Learning Algorithm and Theoretical Analysis

We first derive causal bounds on $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$ and $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ under Assumption 2. Because of limited pages, the construction and its proof are deferred to Proposition 6 (in Appendix E.1). We note that the obtained causal bounds on $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$ and $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ in Proposition 6 are never wider than those of (Zhang & Bareinboim, 2019, Theorem 5; Corollary 1) used in Section 3. Moreover, they are nested in Γ: at $\Gamma = 1$ they collapse to the g-formula (Robins, 1986) and as $\Gamma \to \infty$ they recover (Zhang & Bareinboim, 2019, Theorem 5; Corollary 1) exactly. To account for finite-sample uncertainty, we construct the offline confidence set in Proposition 3. For reasons of space, the construction of the estimated causal bounds $L_k^\Gamma(\bar{s}_k, \bar{a}_k), U_k^\Gamma(\bar{s}_k, \bar{a}_k), k \in [K]$, is given in Appendix A.2 (eqs. (17) to (29)) and the proof is in Appendix E.1.

Proposition 3. *Under Assumptions 1 and 2, suppose* $P(\bar{s}_K, \bar{a}_K, y) > 0, \forall(\bar{s}_K, \bar{a}_K, y)$. *Fix* $\delta_1 \in (0, 1]$, *let* $L_k^\Gamma(\bar{s}_k, \bar{a}_k), U_k^\Gamma(\bar{s}_k, \bar{a}_k), k \in [K]$ *be the estimated causal bounds (eqs. (17) to (29)). Let*

$$\widehat{C}^\Gamma_{k,\text{off}} = \big\{ \theta_k \in \mathbb{R}^{d_k} : L_k^\Gamma(\bar{s}_k, \bar{a}_k) \leqslant \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k \leqslant U_k^\Gamma(\bar{s}_k, \bar{a}_k) \quad \forall(\bar{s}_k, \bar{a}_k) \big\}. \tag{13}$$

Then $P\big(\forall\, k \in [K],\ \theta_k^* \in \widehat{C}^\Gamma_{k,\text{off}}\big) \geqslant 1 - \delta_1$.

For Proposition 3, with probability at least $1 - \delta_1$, $\widehat{C}^\Gamma_{k,\text{off}} \subseteq \widehat{C}^{\Gamma'}_{k,\text{off}}$ for every $1 \leqslant \Gamma \leqslant \Gamma' < \infty$ and every $k \in [K]$. At $\Gamma = \infty$ it reduces to that of Proposition 2, i.e. $\widehat{C}^\infty_{k,\text{off}} = \widehat{C}^{\text{off}}_k$, therefore $\widehat{C}^\Gamma_{k,\text{off}} \subseteq \widehat{C}^{\text{off}}_k$ for every $\Gamma \in [1, \infty)$. Under bounded confounding, the algorithm remains unchanged from Section 3; only the offline confidence set changes, with $\widehat{C}^{\text{off}}_k$ replaced by $\widehat{C}^\Gamma_{k,\text{off}}$. For convenience, we restate it as Algorithm 4 in Appendix A.1. Under the sensitivity model, the worst-case offline width is defined as follows.

Definition 3 (Offline width under the sensitivity model)**.** For $k \in [K]$, the worst-case offline width under the sensitivity model is $W^\Gamma_{k,\text{off}} := \max_{(\bar{s}_k, \bar{a}_k) \in \overline{\mathcal{S}}_k \times \overline{\mathcal{A}}_k} \big[U_k^\Gamma(\bar{s}_k, \bar{a}_k) - L_k^\Gamma(\bar{s}_k, \bar{a}_k)\big]$.

With that, we obtain the following result under bounded confounding; the proof is in Appendix E.2.

Theorem 2. *Let Assumptions 1 and 2 hold. Let* χ, C_1, C, c_β, λ *and* β_k *be as in Theorem 1. Then for any* $\delta, \delta_1 \in (0, 1]$, *with probability at least* $1 - \delta - 2\delta_1$, *Algorithm 4 (in Appendix A.1) achieves*

$$R(T) \leqslant 2\sqrt{2TK\log(2/\delta_1)} + \min\Big\{2C_1 d_K^{3/2} K\sqrt{T \cdot \chi\, \log(\lambda + T)},\ T\sum_{k=1}^K W^\Gamma_{k,\text{off}}\Big\}. \tag{14}$$

Remark 2. Theorem 2 has the same form as Theorem 1, the only difference is that W^{off}_k is replaced by $W^\Gamma_{k,\text{off}}$. Since $\widehat{C}^\Gamma_{k,\text{off}} \subseteq \widehat{C}^{\text{off}}_k$ for every $\Gamma \geqslant 1$, we have $W^\Gamma_{k,\text{off}} \leqslant W^{\text{off}}_k$, therefore Theorem 2 is never weaker than Theorem 1, and at $\Gamma = \infty$ it reduces to Theorem 1. Here, the worst-case offline width $W^\Gamma_{k,\text{off}}$ depends on Γ and decomposes as $W^\Gamma_{k,\text{off}} = W_k(\Gamma) + \widetilde{\mathcal{O}}(1/\sqrt{N_{\text{off}}})$. As in Remark 1, $W_k(\Gamma) = \lim_{N_{\text{off}} \to \infty} W^\Gamma_{k,\text{off}}$ is the width of the causal bounds on Q^* and does not depend on N_{off}. Unlike Remark 1, $W_k(\Gamma)$ varies with Γ: it decreases as Γ decreases, recovering the W_k of Remark 1 at $\Gamma = \infty$ and equaling 0 at $\Gamma = 1$. Since the offline bound $T\sum_{k=1}^K W^\Gamma_{k,\text{off}}$ is one argument of the minimum in Equation 14, a smaller Γ makes this bound smaller and allows it to be selected over a longer episode.

5 Causal Offline-Online Learning under Fully Unconfounding

In Sections 3 and 4, the mechanism generating the offline trajectories is subject to unrestricted and bounded confounding, respectively. In this section, we consider that the behavior policy $\boldsymbol{\pi}_b$ is independent of the unmeasured variables, so that treatment assignment satisfies sequential ignorability, as holds by design in randomized experiments (Lavori & Dawson, 2004; Murphy, 2005b; Lavori & Dawson, 2014). Sections 3 and 4 use the offline data only to construct the offline confidence set constraining $\widehat{Q}^t_k$. However, when the offline data are fully unconfounded ($\Gamma = 1$ in Section 4), they can be pooled directly into the estimation of $\{\widehat{\theta}^t_k\}_{k=1}^K$, replacing the design matrix Λ^t_k by $\widetilde{\Lambda}^t_k$. See Algorithm 2 for pseudocode. Algorithm 2 accordingly differs from Algorithms 1 and 4 in one place:

Algorithm 2 Offline–Online Optimistic Q-Learning for DTRs under Fully Unconfounding

Require: Regularization parameter λ, confidence radii β_k and $\widetilde{\beta}_k$, $\forall k \in [K]$, offline dataset $\mathcal{D}^{\text{off}} = \{(\bar{s}_K^{(n)}, \bar{a}_K^{(n)}, Y^{(n)})\}_{n=1}^{N_{\text{off}}}$, precomputed offline confidence sets $\widehat{C}_{k,\text{off}}^{\Gamma=1}$, $\forall k \in [K]$ of Proposition 3.

1: **for** episode $t = 1, \ldots, T$ **do**
2: Receive the initial state s_1^t.
3: **for** step $k = K, K-1, \ldots, 1$ **do**
4: Form design matrices: $\Lambda_k^t \leftarrow \lambda I + \sum_{\tau=1}^{t-1} \psi_k^\tau (\psi_k^\tau)^\top$; $\widetilde{\Lambda}_k^t \leftarrow \lambda I + \sum_{\iota=1}^{N_{\text{off}}+t-1} \psi_k^\iota (\psi_k^\iota)^\top$.
5: Estimate the online-only and pooled parameters at the terminal and non-terminal stages:

$$\widehat{\theta}_K^t \leftarrow (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K^\tau Y^\tau; \qquad \widehat{\theta}_k^t \leftarrow (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \max_{a_{k+1}} \widetilde{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1}), \tag{15}$$

$$\widetilde{\theta}_K^t \leftarrow (\widetilde{\Lambda}_K^t)^{-1} \sum_{\iota=1}^{N_{\text{off}}+t-1} \psi_K^\iota Y^\iota; \; \widetilde{\theta}_k^t \leftarrow (\widetilde{\Lambda}_k^t)^{-1} \sum_{\iota=1}^{N_{\text{off}}+t-1} \psi_k^\iota \max_{a_{k+1}} \widetilde{Q}_{k+1}^t(\bar{s}_{k+1}^\iota, \bar{a}_k^\iota, a_{k+1}). \tag{16}$$

6: Construct confidence sets: $\widehat{\mathcal{C}}_k^t \leftarrow \{\theta : \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \leqslant \beta_k\}$, $\widetilde{\mathcal{C}}_k^t \leftarrow \{\theta : \|\theta - \widetilde{\theta}_k^t\|_{\widetilde{\Lambda}_k^t} \leqslant \widetilde{\beta}_k\}$.
7: Define the optimistic Q-function, $\forall\, (\bar{s}_k, \bar{a}_k)$:

$$\widetilde{Q}_k^t(\bar{s}_k, \bar{a}_k) \leftarrow \max\Big\{ \min\Big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), \max_{\theta \in \widetilde{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), \max_{\theta \in \widehat{C}_{k,\text{off}}^{\Gamma=1}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), 1 \Big\}, 0 \Big\}.$$

8: **for** step $k = 1, \ldots, K$ **do**
9: Choose treatment $a_k^t \leftarrow \arg\max_{a_k} \widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k)$ and observe the next state s_{k+1}^t.
10: Observe the outcome y^t at terminal stage K.

the offline trajectories additionally enter the regression in Equation 16, which contributes a third branch to the optimistic Q-function. The offline confidence set is retained, now at $\Gamma = 1$.

Before discussing the analysis, we first define a quantity measuring the quality of $\boldsymbol{\pi}_b$.

Definition 4 (Offline all-policy concentrability coefficient)**.** The offline all-policy concentrability coefficient of the behavior policy $\boldsymbol{\pi}_b$ is $\gamma_{\text{off}} := \max_{k \in [K]} 1/\lambda_{\min}(\Sigma_k)$, where $\Sigma_k = \mathbb{E}_{(\bar{S}_k, \bar{A}_k) \sim P}[\psi_k(\bar{S}_k, \bar{A}_k)\psi_k(\bar{S}_k, \bar{A}_k)^\top]$ and P is the observational distribution induced by the environment together with $\boldsymbol{\pi}_b$.

The quantity γ_{off} coincides with the partial offline all-policy concentrability coefficient of (Tan et al., 2024, Eq. (5)) when taking the partition that assigns the entire space to the offline component. It measures how well the observational distribution, from which the offline trajectories are sampled, covers the feature space at every stage $k \in [K]$. The smaller γ_{off} is, the stronger the coverage is.

Theorem 3. *Let Assumptions 1, 2 with $\Gamma = 1$ hold. Let χ, C, C_1, c_β, λ and β_k be as in Theorem 1, let γ_{off} be as in Definition 4, and let $W_{k,\text{off}}^{\Gamma=1}$ be as in Definition 3 with $\Gamma = 1$. For any $\delta, \delta_1 \in (0, 1]$, suppose $8\gamma_{\text{off}} \log\left(Kd_K/\delta_1\right) \leqslant N_{\text{off}} \leqslant T^{d_K}$; let $\widetilde{\chi} = \log\left(\frac{4(c_\beta+1)d_K K(N_{\text{off}}+T)}{\delta}\right)$ and $\widetilde{\beta}_k = Cd_K\sqrt{\widetilde{\chi}} + \sqrt{\lambda d_k}$. Then, with probability at least $1 - \delta - 3\delta_1$, Algorithm 2 achieves*

$$R(T) \leqslant 2\sqrt{2TK \log \frac{2}{\delta_1}} + \min\Big\{ 2C_1 d_K^{3/2} K \sqrt{T\chi \log(\lambda + T)}, C_1 d_K KT \sqrt{\tfrac{2\widetilde{\chi}\gamma_{\text{off}}}{N_{\text{off}}}}, T \sum_{k=1}^{K} W_{k,\text{off}}^{\Gamma=1} \Big\}.$$

Remark 3. The purely online branch of Theorem 3 (i.e. $2\sqrt{2TK\log(2/\delta_1)}$) is identical to that of Theorems 1 and 2. This regret bound is never worse than those of Theorems 1 and 2. At $\Gamma = 1$, the setting of Section 4 reduces to the fully unconfounded setting of Section 5, but the guarantees do not (as offline data was only used for confidence bound construction). Theorem 3 includes the causal bound branch $T\sum_{k=1}^{K} W_{k,\text{off}}^{\Gamma=1}$ in Theorem 2 and adds the pooled branch $C_1 d_K KT\sqrt{\frac{2\widetilde{\chi}\gamma_{\text{off}}}{N_{\text{off}}}}$. Since the bound is a minimum, the offline data are informative as soon as either branch is small. The causal bound branch requires $P(\bar{s}_k, \bar{a}_k) > 0$ at every $(\bar{s}_k, \bar{a}_k) \in \overline{\mathcal{S}}_k \times \overline{\mathcal{A}}_k$, since a single sequence $(\bar{s}_k, \bar{a}_k)$ that the behavior policy never generates gives $T\sum_{k=1}^{K} W_{k,\text{off}}^{\Gamma=1} \geqslant T$, no matter how many offline trajectories are collected. However, a sequence $(\bar{s}_k, \bar{a}_k)$ that the behavior policy never generates does not by itself make γ_{off} infinite. The features $\psi_k(\bar{s}_k, \bar{a}_k)$ at the sequences carrying positive probability may still span $\mathbb{R}^{d_k}$, so that $\lambda_{\min}(\Sigma_k) > 0$, which is all the pooled branch requires.

AI USE STATEMENT

We used large language models (LLMs) in preparing this paper for the following purposes: (1) polishing the writing; (2) helping to search for related works; (3) assisted with checking the proofs; and (4) assisting with code debugging. The authors take full responsibility for the paper's content.

ETHICS STATEMENT

This work is mostly theoretical and all experiments use synthetic data. No human subjects were involved, no patient records or other personally identifiable data were used, and no data set is released. We declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

All assumptions are stated explicitly, and complete proofs of all propositions, theorems, and lemmas are provided in the appendices. Details of the experiments are described in Appendix H. We will release the code upon publication.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo*q*l: Towards optimal regret in model-free rl with nonlinear function approximation. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 987–1063. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/agarwal23a.html.

Philip Amortila, Dylan Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16051–16096, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/462eec462264e54382a780faa658bb20-Paper-Conference.pdf.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002. doi: 10.1023/A:1013689704352.

Oussama Azizi, Philip Boeken, Onno Zoeter, Frans A Oliehoek, and Matthijs T. J. Spaan. Leveraging diverse offline data in POMDPs with unobserved confounders. In *Seventeenth European Workshop on Reinforcement Learning*, 2024. URL https://openreview.net/forum?id=uSheVlIgzc.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, pp. 1171–1176, 1997. URL https://doi.org/10.2307/2965583.

E. Bareinboim, J. Zhang, and S. Lee. An introduction to causal reinforcement learning. Technical Report R-65, Causal Artificial Intelligence Lab, Columbia University, December 2024.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996. doi: 10.1007/BF00114723.

David A Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1116–1126. PMLR, 18–24 Jul 2021.

Bibhas Chakraborty, Susan Murphy, and Victor Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343, 2010. doi: 10.1177/0962280209105013.

Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning Bellman complete representations for offline policy evaluation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/chang22b.html.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 02 2020. ISSN 1369-7412. doi: 10.1111/rssb.12348. URL https://doi.org/10.1111/rssb.12348.

Juan D. Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10093–10100, 2020. doi: 10.1609/aaai.v34i06.6567.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., 2023. doi: 10.52202/075280-1765. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7f8b8bc8ebac661c442c4dafd5d98c08-Paper-Conference.pdf.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

Jesse Y. Hsu and Dylan S. Small. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 2013. ISSN 0006-341X. doi: 10.1111/biom.12101. URL https://doi.org/10.1111/biom.12101.

Yichun Hu and Nathan Kallus. DTR bandit: Learning to make response-adaptive decisions with low regret, 2022. URL https://arxiv.org/abs/2005.02791.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete POMDPs. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18530–18539. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d783823cc6284b929c2cd8df2167d212-Paper.pdf.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020b. URL https://proceedings.mlr.press/v125/jin20a.html.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/3a09a524440d44d7f19870070a5ad42f-Paper.pdf.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fd4f21f2556dad0ea8b7a5c04eabebda-Paper.pdf.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2281–2290. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/kallus19a.html.

Connie Kasari, Ann Kaiser, Kelly Goods, Jennifer Nietfeld, Pamela Mathy, Rebecca Landa, Susan Murphy, and Daniel Almirall. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(6):635–646, 2014. ISSN 0890-8567. doi: https://doi.org/10.1016/j.jaac.2014.01.019.

Michael R. Kosorok and Erica E. M. Moodie. *Adaptive Treatment Strategies in Practice*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015. doi: 10.1137/1.9781611974188.

Eric B. Laber, Daniel J. Lizotte, Min Qian, William E. Pelham, and Susan A. Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1): 1225–1272, 2014. doi: 10.1214/14-EJS920.

Philip W Lavori and Ree Dawson. Dynamic treatment regimes: practical design considerations. *Clinical Trials*, 1(1):9–20, 2004. doi: 10.1191/1740774504cn002oa. URL https://journals.sagepub.com/doi/abs/10.1191/1740774504cn002oa. PMID: 16281458.

Philip W. Lavori and Ree Dawson. Introduction to dynamic treatment strategies and sequential multiple assignment randomization. *Clinical Trials*, 11(4):393–399, 2014. doi: 10.1177/1740774514527651.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 1990. URL http://www.jstor.org/stable/2006592.

Sophie E. Mayer, Hung-Jui Tan, Sharon Peacock Hinton, Hanna K. Sanoff, Til Stürmer, Laura L. Hester, Keturah R. Faurot, Michele Jonsson Funk, and Jennifer L. Lund. Comparison of Medicare claims-based proxy measures of poor function and associations with treatment receipt and mortality in older colon cancer patients. *Medical Care*, 57(4):286–294, 2019. doi: 10.1097/MLR.0000000000001073.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 05 2003. ISSN 1369-7412. doi: 10.1111/1467-9868.00389. URL https://doi.org/10.1111/1467-9868.00389.

Susan A. Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6 (37):1073–1097, 2005a. URL http://jmlr.org/papers/v6/murphy05a.html.

Susan A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481, 2005b. doi: 10.1002/sim.2022.

Susan A. Murphy, Mark J. van der Laan, James M. Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001. URL http://www.jstor.org/stable/3085909.

Thomas A. Murray, Ying Yuan, and Peter F. Thall. A Bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association*, 113 (523):1255–1267, 2018. doi: 10.1080/01621459.2017.1340887.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18819–18831. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/da21bae82c02d1e2b8168d57cd3fbab7-Paper.pdf.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2377–2386, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/osband16.html.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

Judea Pearl and James Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11*, San Francisco, CA, 1995. Morgan Kaufmann.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 1986. ISSN 0270-0255. doi: https://doi.org/10.1016/0270-0255(86)90088-6.

James M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, Lecture Notes in Statistics. Springer, New York, NY, 2004.

Paul R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987. doi: 10.1093/biomet/74.1.13. URL https://doi.org/10.1093/biomet/74.1.13.

A. John Rush, Maurizio Fava, Stephen R. Wisniewski, et al. Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. *Controlled Clinical Trials*, pp. 119–142, 2004. doi: 10.1016/S0197-2456(03)00112-0.

Soroush Saghafian. Ambiguous dynamic treatment regimes: A reinforcement learning approach. *Management Science*, 70(9):5667–5690, 2024. doi: 10.1287/mnsc.2022.00883.

Phillip J. Schulte, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, 2014. doi: 10.1214/13-STS450.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable Markov decision processes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20057–20094. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shi22f.html.

Yuda Song, Yifei Zhou, Ayush Sekhari, Drew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yyBis80iUuU.

Kevin Tan, Wei Fan, and Yuting Wei. Hybrid reinforcement learning breaks sample size barriers in linear MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120038–120077. Curran Associates, Inc., 2024. doi: 10.52202/079017-3815. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d9251dc20346dffe9b6db86dcc6f8cc9-Paper-Conference.pdf.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. URL http://www.jstor.org/stable/27639777.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10276–10283, 2020. doi: 10.1609/aaai.v34i06.6590.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230, May 2015. ISSN 1935-8237. doi: 10.1561/2200000048. URL https://doi.org/10.1561/2200000048.

Anastasios A. Tsiatis, Marie Davidian, Shannon T. Holloway, and Eric B. Laber. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman & Hall/CRC, Boca Raton, FL, 2019.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 578–592. Curran Associates, Inc., 2022. doi: 10.52202/068431-0042. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/03d7e13f0092405804f3a381ade8f3f0-Paper-Conference.pdf.

Andrew Wagenmaker and Aldo Pacchiano. Leveraging offline data in online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35300–35338. PMLR, 2023. URL https://proceedings.mlr.press/v202/wagenmaker23a.html.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b0b79da57b95837f14be95aaa4d54cf8-Paper.pdf.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=CBmJwzneppz.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e61eaa38aed621dd776d0e67cfeee366-Paper.pdf.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8252831b9fce7a49421e622c14ce0f65-Paper.pdf.

Junzhe Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11012–11022. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zhang20a.html.

Ying-Qi Zhao, Donglin Zeng, Eric B. Laber, and Michael R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015. doi: 10.1080/01621459.2014.937488.

CONTENTS

1 Introduction **1**

1.1 Related Works 2

2 Problem Formulation **3**

2.1 Dynamic Treatment Regimes 3

2.2 Linear Value Function Approximation 4

3 Causal Offline-Online Learning under Unrestricted Confounding **5**

3.1 Offline Confidence Set 5

3.2 Learning Algorithm and Theoretical Analysis 6

4 Causal Offline-Online Learning under Bounded Confounding **6**

4.1 Sensitivity Model 7

4.2 Learning Algorithm and Theoretical Analysis 8

5 Causal Offline-Online Learning under Fully Unconfounding **8**

A Preliminaries **16**

A.1 Algorithm Pseudocode 16

A.2 Offline confidence set construction of the causal bounds (for Proposition 3) 17

B Additional Related Works **18**

C Supplementary Material for Section 2 **19**

D Proof for Section 3 **23**

D.1 Proof for Offline Confidence Set 23

D.2 Proof for the Regret Bound 25

E Proof for Section 4 **36**

E.1 Proof for Offline Confidence Set 36

E.2 Proof for Regret Bound 42

F Proof for Section 5 **44**

G Auxiliary Results **50**

H Numerical Experiments **52**

A PRELIMINARIES

Due to space constraints, Algorithms 3 and 4, for the purely online and bounded confounding settings respectively, together with the construction of the offline confidence set under the sensitivity model in Proposition 3, were deferred to here.

A.1 ALGORITHM PSEUDOCODE

Algorithm 3 Online Optimistic Q-Learning for DTRs (Algorithm 1 without offline data)

Require: Regularization parameter λ, confidence radii $\beta_k,\ \forall k \in [K]$, .

for episode $t = 1, \ldots, T$ **do**
  Receive the initial state s_1^t.
  for step $k = K, K-1, \ldots, 1$ **do**
    Accumulate all past features at step k: $\Lambda_k^t \leftarrow \sum_{\tau=1}^{t-1} \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)^\top + \lambda I$.
    Estimate the stage parameters:

$$\widehat{\theta}_K^t \leftarrow (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K(\bar{s}_K^\tau, \bar{a}_K^\tau) Y^\tau; \quad \widehat{\theta}_k^t \leftarrow (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau) \max_{a_{k+1}} \widehat{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1}).$$

    Construct the online confidence set: $\widehat{\mathcal{C}}_k^t \leftarrow \left\{\theta \in \mathbb{R}^{d_k} : \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \leqslant \beta_k\right\}$.
    Define the optimistic Q-function, $\forall\ (\bar{s}_k, \bar{a}_k)$:

$$\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \leftarrow \max\Big\{ \min\Big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ 1\Big\},\ 0\Big\}.$$

  for step $k = 1, \ldots, K$ **do**
    Choose treatment $a_k^t \leftarrow \arg\max_{a_k} \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k)$.
    Observe the next state s_{k+1}^t.
  Observe the outcome y^t at terminal stage K.

Algorithm 4 Offline-Online Optimistic Q-Learning for DTRs under Bounded Confounding (Γ-version of Algorithm 1)

Require: Regularization parameter λ, confidence radii $\beta_k,\ \forall k \in [K]$, , and precomputed offline confidence sets $\widehat{\mathcal{C}}_{k,\text{off}}^\Gamma, \forall k \in [K]$.

for episode $t = 1, \ldots, T$ **do**
  Receive the initial state s_1^t.
  for step $k = K, K-1, \ldots, 1$ **do**
    Accumulate all past features at step k: $\Lambda_k^t \leftarrow \sum_{\tau=1}^{t-1} \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)^\top + \lambda I$.
    Estimate the stage parameters:

$$\widehat{\theta}_K^t \leftarrow (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K(\bar{s}_K^\tau, \bar{a}_K^\tau) Y^\tau; \quad \widehat{\theta}_k^t \leftarrow (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau) \max_{a_{k+1}} \widehat{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1}).$$

    Construct the online confidence set: $\widehat{\mathcal{C}}_k^t \leftarrow \left\{\theta \in \mathbb{R}^{d_k} : \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \leqslant \beta_k\right\}$.
    Define the optimistic Q-function, $\forall\ (\bar{s}_k, \bar{a}_k)$:

$$\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \leftarrow \max\Big\{ \min\Big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ \max_{\theta \in \widehat{\mathcal{C}}_{k,\text{off}}^\Gamma} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ 1\Big\},\ 0\Big\}.$$

  for step $k = 1, \ldots, K$ **do**
    Choose treatment $a_k^t \leftarrow \arg\max_{a_k} \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k)$.
    Observe the next state s_{k+1}^t.
  Observe the outcome y^t at terminal stage K.

A.2 Offline confidence set construction of the causal bounds (for Proposition 3)

The construction of the offline confidence set in Proposition 3 uses the following definitions:

$$\begin{aligned}\mathcal{B}^{\text{off}} = \Big\{&(\mathbb{P}, m_K) : |\mathbb{P}(h) - \widehat{\mathbb{P}}(h)| \leqslant b_{\text{off}} \forall h \in \mathcal{H};\ |m_K(\bar{s}_K, \bar{a}_K) - \widehat{m}_K(\bar{s}_K, \bar{a}_K)| \leqslant b_{\text{off}}\ \forall (\bar{s}_K, \bar{a}_K);\\ &0 < \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_{k-1}) \quad \forall k \in [K-1];\\ &0 < \mathbb{P}(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_{K-1}); \quad 0 \leqslant m_K(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_K)\Big\} \end{aligned} \tag{17}$$

$$B_k^-(\mathbb{P}) = \prod_{i=1}^{k} \frac{\pi_{b,i}(\mathbb{P})}{\Gamma(1-\pi_{b,i}(\mathbb{P})) + \pi_{b,i}(\mathbb{P})},\ B_0^- = 1 \tag{18}$$

$$B_k^+(\mathbb{P}) = \prod_{i=1}^{k} \frac{\Gamma\pi_{b,i}(\mathbb{P})}{1-\pi_{b,i}(\mathbb{P}) + \Gamma\pi_{b,i}(\mathbb{P})},\ B_0^+ = 1 \tag{19}$$

$$\zeta^\Gamma(\bar{s}_{k+1}, \bar{a}_k; \mathbb{P}) = \min\Big\{\frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-(\mathbb{P})},\ \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) - \mathbb{P}(\bar{s}_k, \bar{a}_k) + \zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})\Big\} \tag{20}$$

$$\underline{\mu}(\mathbb{P}, m_K) = \max\Big\{\frac{m_K}{B_K^+(\mathbb{P})\,\zeta^\Gamma(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})},\ 1 - \frac{B_{K-1}^+(\mathbb{P})}{B_K^-(\mathbb{P})}\frac{\mathbb{P}(\bar{s}_K, \bar{a}_K) - m_K}{\mathbb{P}(\bar{s}_K, \bar{a}_{K-1})}\Big\} \tag{21}$$

$$\overline{\mu}(\mathbb{P}, m_K) = \min\Big\{1 - \frac{\mathbb{P}(\bar{s}_K, \bar{a}_K) - m_K}{B_K^+(\mathbb{P})\,\zeta^\Gamma(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})},\ \frac{B_{K-1}^+(\mathbb{P})}{B_K^-(\mathbb{P})}\frac{m_K}{\mathbb{P}(\bar{s}_K, \bar{a}_{K-1})}\Big\} \tag{22}$$

$$L_K^\Gamma(\bar{s}_K, \bar{a}_K) = \max\Big\{\inf_{\mathcal{B}^{\text{off}}} \underline{\mu}(\mathbb{P}, m_K),\ 0\Big\} \tag{23}$$

$$U_K^\Gamma(\bar{s}_K, \bar{a}_K) = \min\Big\{\sup_{\mathcal{B}^{\text{off}}} \overline{\mu}(\mathbb{P}, m_K),\ 1\Big\} \tag{24}$$

$$\underline{\tau}_k(s_{k+1}; \mathbb{P}) = \frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{B_k^+(\mathbb{P})\,\zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})}; \tag{25}$$

$$\overline{\tau}_k(s_{k+1}; \mathbb{P}) = \min\Big\{1 - \frac{\mathbb{P}(\bar{s}_k, \bar{a}_k) - \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})},\ \frac{B_{k-1}^+(\mathbb{P})}{B_k^-(\mathbb{P})}\frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\mathbb{P}(\bar{s}_k, \bar{a}_{k-1})}\Big\} \tag{26}$$

$$\widehat{\mathcal{P}}_k^\Gamma(\bar{s}_k, \bar{a}_k) = \Big\{p \in \Delta(\mathcal{S}_{k+1}) : \inf_{\mathcal{B}^{\text{off}}} \underline{\tau}_k(s_{k+1}; \mathbb{P}) \leqslant p(s_{k+1}) \leqslant \sup_{\mathcal{B}^{\text{off}}} \overline{\tau}_k(s_{k+1}; \mathbb{P}) \forall s_{k+1} \in \mathcal{S}_{k+1}\Big\} \tag{27}$$

$$L_k^\Gamma(\bar{s}_k, \bar{a}_k) = \min_{p \in \widehat{\mathcal{P}}_k^\Gamma} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} L_{k+1}^\Gamma(\bar{s}_{k+1}, \bar{a}_{k+1}) \tag{28}$$

$$U_k^\Gamma(\bar{s}_k, \bar{a}_k) = \max_{p \in \widehat{\mathcal{P}}_k^\Gamma} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} U_{k+1}^\Gamma(\bar{s}_{k+1}, \bar{a}_{k+1}), \tag{29}$$

And $\pi_{b,i}(\mathbb{P}) = \frac{\mathbb{P}(\bar{s}_i, \bar{a}_i)}{\mathbb{P}(\bar{s}_i, \bar{a}_{i-1})}$ and $\zeta^\Gamma(s_1; \mathbb{P}) = \mathbb{P}(s_1)$.

1. We detail below useful equations that will be used throughout the appendix.

$$\Lambda_k^t = \sum_{\tau < t} \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)^\top + \lambda I, \tag{30}$$

$$V_k^*(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \mathcal{A}_k} Q_k^*(\bar{s}_k, \bar{a}_{k-1}, a_k), \tag{31}$$

$$V_k^{\boldsymbol{\pi}}(\bar{s}_k, \bar{a}_{k-1}) = \mathbb{E}_{a_k \sim \pi_k(\cdot|\bar{s}_k, \bar{a}_{k-1})}\big[Q_k^{\boldsymbol{\pi}}(\bar{s}_k, \bar{a}_{k-1}, a_k)\big], \tag{32}$$

$$Q_k^*(\bar{s}_k, \bar{a}_k) = \mathbb{E}_{\bar{a}_k}\Big[\max_{a_{k+1} \in \mathcal{A}_{k+1}} Q_{k+1}^*\big((\bar{s}_k, S_{k+1}), \bar{a}_k, a_{k+1}\big) \,\Big|\, \bar{s}_k\Big], \tag{33}$$

$$Q_k^{\boldsymbol{\pi}}(\bar{s}_k, \bar{a}_k) = \mathbb{E}_{\bar{a}_k}\Big[V_{k+1}^{\boldsymbol{\pi}}\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \,\Big|\, \bar{s}_k\Big] \tag{34}$$

$$= \mathbb{E}_{\bar{a}_k}\Big[\mathbb{E}_{a_{k+1} \sim \pi_{k+1}(\cdot|(\bar{s}_k, S_{k+1}), \bar{a}_k)}\big[Q_{k+1}^{\boldsymbol{\pi}}\big((\bar{s}_k, S_{k+1}), \bar{a}_k, a_{k+1}\big)\big] \,\Big|\, \bar{s}_k\Big]. \tag{35}$$

2. For convenience in the proofs, we restate Assumption 1 explicitly; by Definition 1, it is equivalent to the feature maps satisfying $\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2 \leqslant 1$ for every $k \in [K]$ and $(\bar{s}_k, \bar{a}_k)$, together with the following two conditions.

(i) For the terminal stage K, the outcome satisfies $Y \in [0,1]$ and there exists an unknown vector $w_K \in \mathbb{R}^{d_K}$ with $\|w_K\|_2 \leqslant \sqrt{d_K}$ such that, for every $(\bar{s}_K, \bar{a}_K) \in \bar{\mathcal{S}}_K \times \bar{\mathcal{A}}_K$,

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] = \langle \psi_K(\bar{s}_K, \bar{a}_K), w_K \rangle. \tag{36}$$

(ii) (*Optimistic closure, adapted from* Wang et al., 2021b.) For each stage $k \in [K-1]$ and every $V \in \mathcal{V}_{k+1}$, there exists an unknown vector $\theta_V \in \mathbb{R}^{d_k}$ with $\|\theta_V\|_2 \leqslant \sqrt{d_k}$ such that, for every $(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$,

$$\mathbb{E}_{\bar{a}_k}\big[V\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \,\big|\, \bar{s}_k\big] = \langle \psi_k(\bar{s}_k, \bar{a}_k), \theta_V \rangle. \tag{37}$$

3. We next introduce the constants, function class, and good event used in Lemma 9.

Let χ be as in Theorem 1 and $C > 0$ a constant, and define

$$\widetilde{\chi} := \log\Big(\frac{4(c_\beta + 1) d_K K (N_{\mathrm{off}} + T)}{\delta}\Big), \quad \widetilde{\mathcal{C}} := C d_K \sqrt{\widetilde{\chi}} + \sqrt{\lambda d_K}, \quad \widetilde{B} := (C+1) d_K \sqrt{\widetilde{\chi}},$$

and the confidence radii $\widetilde{\beta}_k := C d_K \sqrt{\widetilde{\chi}} + \sqrt{\lambda d_k}$ and $\beta_k := C d_K \sqrt{\chi} + \sqrt{\lambda d_k}$, so that $\beta_k \leqslant \widetilde{\beta}_k \leqslant \widetilde{B}$ for every $k \in [K]$.

Definition 5 (Pooled value-function class for Algorithm 2)**.** For $k < K$, let $\widetilde{\mathcal{V}}_{k+1}^{\mathrm{pool}}$ be the class of functions

$$V(\bar{s}_{k+1}, \bar{a}_k) = \max\Big\{ \max_{a_{k+1} \in \mathcal{A}_{k+1}} \min\Big\{1, \langle \omega_{k+1}, \psi_{k+1} \rangle + \beta \|\psi_{k+1}\|_{\Lambda^{-1}}, \langle \widetilde{\omega}_{k+1}, \psi_{k+1} \rangle + \widetilde{\beta} \|\psi_{k+1}\|_{\widetilde{\Lambda}^{-1}}, \max_{\theta \in \widehat{C}_{k+1,\mathrm{off}}^{\Gamma=1}} \theta^\top \psi_{k+1} \Big\},$$

where the parameters satisfy $\omega_{k+1}, \widetilde{\omega}_{k+1} \in \mathbb{R}^{d_{k+1}}$ with $\|\omega_{k+1}\|_2, \|\widetilde{\omega}_{k+1}\|_2 \leqslant \sqrt{d_{k+1}(N_{\mathrm{off}} + T)/\lambda}$, $\beta, \widetilde{\beta} \in [0, \widetilde{B}]$, and $\lambda_{\min}(\Lambda), \lambda_{\min}(\widetilde{\Lambda}) \geqslant \lambda$. Here $\max_{\theta \in \widehat{C}_{k+1,\mathrm{off}}^{\Gamma=1}} \theta^\top \psi_{k+1}$ is a fixed function of $(\bar{s}_{k+1}, \bar{a}_{k+1})$ taking values in $[0,1]$.

Taking $(\theta, A) = (\omega_{k+1}, \beta^2 \Lambda^{-1})$, $(\theta', A') = (\widetilde{\omega}_{k+1}, \widetilde{\beta}^2 \widetilde{\Lambda}^{-1})$ and $\mathcal{P} = \widehat{C}_{k+1,\mathrm{off}}^{\Gamma=1} \in \mathfrak{P}_{k+1}$ in Equation 3 gives $\widetilde{\mathcal{V}}_{k+1}^{\mathrm{pool}} \subseteq \mathcal{V}_{k+1}$. Define the following good event

$$\begin{aligned}
\widetilde{\mathcal{E}}_1 := &\Big\{\forall k < K,\ \forall t \in [T],\ \forall V \in \widetilde{\mathcal{V}}_{k+1}^{\mathrm{pool}} :\ \Big\| \sum_{\iota=1}^{N_{\mathrm{off}}+t-1} \psi_k^\iota \big(V(\bar{s}_{k+1}^\iota, \bar{a}_k^\iota) - \mathbb{E}[V(\bar{s}_{k+1}^\iota, \bar{a}_k^\iota) \mid \mathcal{G}_{k,\iota-1}]\big) \Big\|_{(\widetilde{\Lambda}_k^t)^{-1}} \leqslant C d_K \sqrt{\widetilde{\chi}} \Big\} \\
&\cap \Big\{\forall t \in [T] :\ \Big\| \sum_{\iota=1}^{N_{\mathrm{off}}+t-1} \psi_K^\iota \big[Y^\iota - \mathbb{E}[Y^\iota \mid \mathcal{G}_{K,\iota-1}]\big] \Big\|_{(\widetilde{\Lambda}_K^t)^{-1}} \leqslant C d_K \sqrt{\widetilde{\chi}} \Big\} \\
&\cap \Big\{\forall k < K,\ \forall t \in [T],\ \forall V \in \widetilde{\mathcal{V}}_{k+1}^{\mathrm{pool}} :\ \Big\| \sum_{\tau=1}^{t-1} \psi_k^\tau \big(V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}[V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) \mid \mathcal{F}_{k,\tau-1}]\big) \Big\|_{(\Lambda_k^t)^{-1}} \leqslant C d_K \sqrt{\chi} \Big\} \\
&\cap \Big\{\forall t \in [T] :\ \Big\| \sum_{\tau=1}^{t-1} \psi_K^\tau \big[Y^\tau - \mathbb{E}[Y^\tau \mid \mathcal{F}_{K,\tau-1}]\big] \Big\|_{(\Lambda_K^t)^{-1}} \leqslant C d_K \sqrt{\chi} \Big\}.
\end{aligned} \tag{38}$$

B Additional Related Works

Dynamic Treatment Regimes. Structural nested mean models with g-estimation (Robins, 2004) estimate optimal decisions through models for treatment effects, while Q-learning (Murphy, 2005a; Chakraborty et al., 2010) recursively models stage-specific Q-functions through backward induction. A-learning (Murphy, 2003) focuses more directly on treatment contrasts. Value-search methods directly optimize over a class of treatment regimes (Zhao et al., 2015). Bayesian methods place probability models on the treatment-response process and quantify uncertainty in the resulting optimal regime (Murray et al., 2018). Closer to our bounded-confounding setting, (Saghafian, 2024)

introduces ambiguous DTRs, which evaluate a regime over a set of candidate causal models to account for time-varying unmeasured confounders.

Hybrid Reinforcement Learning. Hybrid methods incorporate offline data to improve subsequent online learning. Xie et al. (2021) study policy fine-tuning, where an online learner has access to a reference policy close to the optimal one. Song et al. (2023) run fitted Q-iteration on pooled offline and online data. Wagenmaker & Pacchiano (2023) characterize the number of online samples needed to learn a near-optimal policy given an offline dataset in linear MDPs. Amortila et al. (2024) use density-ratio function approximation for online exploration and obtain hybrid guarantees through a reduction from hybrid to offline RL.

Confounded Reinforcement Learning. When point identification is unavailable, existing work is largely offline and focuses mainly on policy evaluation. Sensitivity-analysis approaches derive bounds on the value of a fixed target policy under marginal sensitivity models or related restrictions on unmeasured confounding (Kallus & Zhou, 2020; Namkoong et al., 2020; Bruns-Smith, 2021). Partially observable models provide a different framework for representing latent variables. Sample-efficient reinforcement learning is possible under structural conditions such as undercompleteness or observability in POMDP (Jin et al., 2020a; Uehara et al., 2022), while proximal and minimax approaches address off-policy evaluation when an unobserved state simultaneously affects treatment and outcomes (Tennenholtz et al., 2020; Shi et al., 2022).

C Supplementary Material for Section 2

Example 1 (Adapted from (Rush et al., 2004))**.** A two-stage depression treatment vignette (see Figure 1). S_1 is the baseline severity; A_1 represents a treatment that continues the current medication or prescribes a different antidepressant; S_2 is the updated severity; A_2 is the same treatment as A_1 at the second stage; and Y is the final remission status. The confounder U collects unmeasured factors, adherence, and patient preference, that affect both states, decisions, and outcome. The goal is to learn a two-stage regime maximizing the probability of remission.

Fact 1 (Optimistic closure in tabular DTRs)**.** If $Y \in [0,1]$ and $\psi_k(\bar{s}_k, \bar{a}_k) = e_{(\bar{s}_k, \bar{a}_k)} \in \mathbb{R}^{d_k}$ is the standard-basis feature map on $\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$ for every $k \in [K]$, then Assumption 1 holds.

Proof. With the standard-basis feature map, $d_k = |\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k| = \prod_{i=1}^{k} |\mathcal{S}_i||\mathcal{A}_i|$ and $\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2 = \|e_{(\bar{s}_k, \bar{a}_k)}\|_2 = 1$ for every $(\bar{s}_k, \bar{a}_k)$ and $k \in [K]$. Then, for any $\theta \in \mathbb{R}^{d_k}$,

$$\big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta \big\rangle \ = \ \theta\big[(\bar{s}_k, \bar{a}_k)\big], \tag{39}$$

the $(\bar{s}_k, \bar{a}_k)$-th coordinate of θ. Hence every function on $\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$ is linear in ψ_k.

For the terminal stage $k = K$, define $w_K \in \mathbb{R}^{d_K}$ coordinate-wise by $w_K\big[(\bar{s}_K, \bar{a}_K)\big] := \mathbb{E}_{\bar{a}_K}\big[Y \mid \bar{s}_K\big]$. By Equation 39,

$$\big\langle \psi_K(\bar{s}_K, \bar{a}_K),\ w_K \big\rangle \ = \ w_K\big[(\bar{s}_K, \bar{a}_K)\big] \ = \ \mathbb{E}_{\bar{a}_K}\big[Y \mid \bar{s}_K\big] \qquad \forall (\bar{s}_K, \bar{a}_K).$$

Since $Y \in [0,1]$, every coordinate of w_K lies in $[0,1]$, so $\|w_K\|_2 \leqslant \sqrt{d_K}$.

For the non terminal stage, fix $k \in [K-1]$ and $V \in \mathcal{V}_{k+1}$, and define $\theta_V \in \mathbb{R}^{d_k}$ coordinatewise by

$$\theta_V\big[(\bar{s}_k, \bar{a}_k)\big] \ := \ \mathbb{E}_{\bar{a}_k}\big[V\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\big].$$

By Equation 39,

$$\big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_V \big\rangle \ = \ \theta_V\big[(\bar{s}_k, \bar{a}_k)\big] \ = \ \mathbb{E}_{\bar{a}_k}\big[V\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\big] \qquad \forall (\bar{s}_k, \bar{a}_k),$$

which is the required identity. Since V takes values in $[0,1]$ by Equation 3, every coordinate of θ_V lies in $[0,1]$, so $\|\theta_V\|_2 \leqslant \sqrt{d_k}$. ■

Assumption 3 (Adapted from (Jin et al., 2020b))**.** A DTR $\langle \mathcal{M}^*, \Pi, Y \rangle$ is linear with known feature maps $\psi_k : \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k \to \mathbb{R}^{d_k}$ satisfying $\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2 \leqslant 1$ for all $(\bar{s}_k, \bar{a}_k)$, $k \in [K]$, if the following two conditions hold.

(i) For the terminal stage $k = K$, the outcome satisfies $Y \in [0,1]$ and there exists an unknown vector $w_K \in \mathbb{R}^{d_K}$ with $\|w_K\|_2 \leqslant \sqrt{d_K}$ such that, for every $(\bar{s}_K, \bar{a}_K) \in \bar{\mathcal{S}}_K \times \bar{\mathcal{A}}_K$, $\mathbb{E}_{\bar{a}_K}\big[\, Y \mid \bar{s}_K \big] = \big\langle \psi_K(\bar{s}_K, \bar{a}_K), w_K \big\rangle$.

(ii) For each stage $k \in [K-1]$, there exist unknown signed measures $\mu_k = (\mu_k^{(1)}, \ldots, \mu_k^{(d_k)})$ on $\bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k$ with $\Big\| \big(\sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} \big| \mu_k^{(i)}\big(\{(\bar{s}', \bar{a}')\}\big) \big| \big)_{i=1}^{d_k} \Big\|_2 \leqslant \sqrt{d_k}$ such that, for every $(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$ and $(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k$, $P_{\bar{a}_k}\big((\bar{s}', \bar{a}') \;\big|\; \bar{s}_k\big) = \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \mu_k\big(\{(\bar{s}', \bar{a}')\}\big) \big\rangle$.

Remark 4 (Why Assumption 3(ii) is linear over next histories)**.** Suppose linearity were imposed over the next state alone, as in the linear MDP of Jin et al. (2020b), with feature vector $(\psi_k(\bar{s}_k, \bar{a}_k))$ evaluated at the conditioning history and the signed measure at the next state: there exist unknown signed measures $\nu_k = (\nu_k^{(1)}, \ldots, \nu_k^{(d_k)})$ on $\mathcal{S}_{k+1}$ with $\Big\| \big(\sum_{s' \in \mathcal{S}_{k+1}} |\nu_k^{(i)}(\{s'\})| \big)_{i=1}^{d_k} \Big\|_2 \leqslant \sqrt{d_k}$ such that, for every $(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$ and $s_{k+1} \in \mathcal{S}_{k+1}$, $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) = \big\langle \psi_k(\bar{s}_k, \bar{a}_k), \nu_k(\{s_{k+1}\}) \big\rangle$. This does not suffice in general, because value functions in a DTR may depend on the whole history. For $V : \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k \to [0,1]$,

$$\begin{aligned}\mathbb{E}_{\bar{a}_k}\big[V\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\big] &= \sum_{s_{k+1} \in \mathcal{S}_{k+1}} V\big((\bar{s}_k, s_{k+1}), \bar{a}_k\big)\, P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \\ &= \Big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \sum_{s_{k+1} \in \mathcal{S}_{k+1}} V\big((\bar{s}_k, s_{k+1}), \bar{a}_k\big)\, \nu_k(\{s_{k+1}\}) \Big\rangle,\end{aligned}$$

and $\sum_{s_{k+1} \in \mathcal{S}_{k+1}} V\big((\bar{s}_k, s_{k+1}), \bar{a}_k\big)\, \nu_k(\{s_{k+1}\}) \Big\rangle$ depends on $(\bar{s}_k, \bar{a}_k)$ whenever V depends on the past history. The backup is then in general not of the form $\langle \psi_k(\bar{s}_k, \bar{a}_k), \theta_V \rangle$ with a fixed θ_V. Condition (ii) therefore imposes linearity over next histories, under which every backup is linear in ψ_k with the fixed coefficient $\theta_V = \sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} V(\bar{s}', \bar{a}')\, \mu_k\big(\{(\bar{s}', \bar{a}')\}\big)$.

Lemma 1. *For every DTR and feature maps $\{\psi_k\}_{k \in [K]}$ satisfying Assumption 3, $d_k \geqslant \prod_{i=1}^{k} |\mathcal{S}_i||\mathcal{A}_i|$ at every stage $k \in [K-1]$.*

Proof. Let $V : \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k \to [0,1]$ be any bounded value function and set

$$\theta_V \;:=\; \sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} V(\bar{s}', \bar{a}')\, \mu_k\big(\{(\bar{s}', \bar{a}')\}\big) \in \mathbb{R}^{d_k}, \tag{40}$$

which does not depend on specific sequence $(\bar{s}_k, \bar{a}_k)$.

$$\begin{aligned}&\mathbb{E}_{\bar{a}_k}\big[V\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\big] \\ &\quad= \sum_{s_{k+1} \in \mathcal{S}_{k+1}} V\big((\bar{s}_k, s_{k+1}), \bar{a}_k\big)\, P_{\bar{a}_k}\big(\{(\bar{s}_k, s_{k+1}, \bar{a}_k)\} \mid \bar{s}_k\big) && (41)\\ &\quad= \sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} V(\bar{s}', \bar{a}')\, P_{\bar{a}_k}\big(\{(\bar{s}', \bar{a}')\} \mid \bar{s}_k\big) && (42)\\ &\quad= \sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} V(\bar{s}', \bar{a}') \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \mu_k\big(\{(\bar{s}', \bar{a}')\}\big) \big\rangle && (43)\\ &\quad= \Big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} V(\bar{s}', \bar{a}')\, \mu_k\big(\{(\bar{s}', \bar{a}')\}\big) \Big\rangle && (44)\\ &\quad= \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_V \big\rangle. && (45)\end{aligned}$$

Equation 41 is the definition of the conditional expectation. Equation 42 holds because $P_{\bar{a}_k}\big(\{(\bar{s}', \bar{a}')\} \mid \bar{s}_k\big) = 0$ unless $\bar{s}' = (\bar{s}_k, s_{k+1})$ for some $s_{k+1} \in \mathcal{S}_{k+1}$ and $\bar{a}' = \bar{a}_k$. Equation 43 follows by Assumption 3(ii). Equation 44 holds by the bilinearity of the inner product. Equation 45 follows by Equation 40.

Fix $(\bar{s}'_k, \bar{a}'_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$ and let $V' : \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k \to [0,1]$ be

$$V'\big(\bar{s}_k, s_{k+1}, \bar{a}_k\big) \;:=\; \begin{cases} 1, & (\bar{s}_k, \bar{a}_k) = (\bar{s}'_k, \bar{a}'_k), \\ 0, & \text{otherwise}, \end{cases} \qquad \forall s_{k+1} \in \mathcal{S}_{k+1}.$$

Since V' does not depend on s_{k+1},

$$\mathbb{E}_{\bar{a}_k}\big[V'\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\big] \;=\; \begin{cases} 1, & (\bar{s}_k, \bar{a}_k) = (\bar{s}'_k, \bar{a}'_k), \\ 0, & \text{otherwise.} \end{cases}$$

Then Equation 45 gives

$$\big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_{V'} \big\rangle \;=\; \begin{cases} 1, & (\bar{s}_k, \bar{a}_k) = (\bar{s}'_k, \bar{a}'_k), \\ 0, & \text{otherwise,} \end{cases} \qquad \forall (\bar{s}_k, \bar{a}_k). \tag{46}$$

Next, we show that $\{\psi_k(\bar{s}_k, \bar{a}_k)\}_{(\bar{s}_k, \bar{a}_k)}$ is linearly independent, a consequence of Assumption 3(ii), not a property that feature maps have in general.

Fist, by definition, suppose scalars $c_{(\bar{s}_k, \bar{a}_k)}$ satisfy $\sum_{(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k} c_{(\bar{s}_k, \bar{a}_k)}\, \psi_k(\bar{s}_k, \bar{a}_k) = 0$. For any $(\bar{s}'_k, \bar{a}'_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$,

$$\begin{aligned} 0 \;&=\; \Big\langle \sum_{(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k} c_{(\bar{s}_k, \bar{a}_k)}\, \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_{V'} \Big\rangle \\ &=\; \sum_{(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k} c_{(\bar{s}_k, \bar{a}_k)} \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_{V'} \big\rangle \\ &=\; c_{(\bar{s}'_k, \bar{a}'_k)}, \end{aligned}$$

where the last equality is Equation 46.

Hence $\{\psi_k(\bar{s}_k, \bar{a}_k)\}_{(\bar{s}_k, \bar{a}_k)}$ are $\prod_{i=1}^{k} |\mathcal{S}_i||\mathcal{A}_i|$ linearly independent vectors in $\mathbb{R}^{d_k}$, so $d_k \geqslant \prod_{i=1}^{k} |\mathcal{S}_i||\mathcal{A}_i|$. ■

Proposition 1. *For DTRs with fixed feature maps, optimistic closure is strictly weaker than the linear DTR assumption (Assumption 3), which adapts the linear MDP assumption of Jin et al. (2020b) to histories.*

Proof. To prove this argument, we need to show the following:

(a) Every DTR satisfying Assumption 3 also satisfies Assumption 1 with the same feature maps and dimensions.

(b) The converse fails: there is a DTR with $K = 2$, binary states and actions, and feature maps with $d_1 = d_2 = 1$ that satisfies Assumption 1 but is not a linear DTR for any choice of signed measures.

(a). The feature maps and the bound $\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2 \leqslant 1$ are common to both assumptions, and Assumption 3(i) coincides with Assumption 1(i) at the terminal stage.

For the non-terminal stages, fix $k \in [K-1]$ and $V \in \mathcal{V}_{k+1}$, and set

$$\theta_V \;=\; \sum_{(\bar{s}', \bar{a}') \in \bar{\mathcal{S}}_{k+1} \times \bar{\mathcal{A}}_k} V(\bar{s}', \bar{a}')\, \mu_k\big(\{(\bar{s}', \bar{a}')\}\big) \;\in\; \mathbb{R}^{d_k}.$$

By Equation 45 which use only the transition linearity in Assumption 3(ii), we have,

$$\mathbb{E}_{\bar{a}_k}\big[V\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\big] \;=\; \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_V \big\rangle.$$

Which is the optimistic closure condition in Assumption 1(ii).

For the norm bound in the non-terminal stage, V takes values in $[0, 1]$ by Equation 3, so $|\theta_V^{(i)}| \leqslant \sum_{(\bar{s}', \bar{a}')} \big|\mu_k^{(i)}\big(\{(\bar{s}', \bar{a}')\}\big)\big|$, and the normalization in Assumption 3(ii) gives $\|\theta_V\|_2 \leqslant \sqrt{d_k}$. Hence optimistic closure (in Assumption 1(ii)) holds.

(b). We first construct the witness.

- *Stages and spaces.* $K = 2$ and $\mathcal{S}_1 = \mathcal{S}_2 = \mathcal{A}_1 = \mathcal{A}_2 = \{0, 1\}$.
- *Transitions.* S_2 is uniform on $\mathcal{S}_2$ under every intervention, i.e., $P_{a_1}(s_2 \mid s_1) = 1/2$ for all (s_1, a_1, s_2).

- *Outcome.* $Y \sim \text{Bernoulli}(1/2)$ under every intervention, so $\mathbb{E}_{\bar{a}_2}[Y \mid \bar{s}_2] = 1/2$ for every $(\bar{s}_2, \bar{a}_2)$.
- *Features.* $d_1 = d_2 = 1$ and $\psi_1(s_1, a_1) \equiv \psi_2(\bar{s}_2, \bar{a}_2) \equiv 1$, so $\|\psi_k\|_2 = 1$.

Assumption 1 holds.

- For terminal stage $k = 2$. $Y \in [0,1]$, and with $w_2 := 1/2$ we have $\mathbb{E}_{\bar{a}_2}[Y \mid \bar{s}_2] = 1/2 = \langle \psi_2(\bar{s}_2, \bar{a}_2), w_2 \rangle$ and $|w_2| \leqslant 1 = \sqrt{d_2}$.
- For non-terminal stage $k = 1$. Since $d_2 = 1$ and $\psi_2(\bar{s}_2, \bar{a}_2) \equiv 1$, each term in Equation 3, namely $\theta + \sqrt{A}$, $\theta' + \sqrt{A'}$ and $\sup_{\vartheta \in \mathcal{P}} \vartheta$, does not depend on $(\bar{s}_2, \bar{a}_2)$. Hence every $Q \in \mathcal{Q}_2^{\text{enl}}$ is constant, and so every $V \in \mathcal{V}_2$ is a constant $c \in [0,1]$ by the truncation $0 \vee (1 \wedge \cdot)$. Therefore, $\mathbb{E}_{\bar{a}_1}[c \mid \bar{s}_1] = c = \langle \psi_1(s_1, a_1), \theta_V \rangle$ with $\theta_V := c$, and $|\theta_V| \leqslant 1 = \sqrt{d_1}$.

Assumption 3 does not hold. We look at the non-terminal stage $k = 1$.

Suppose there exist signed measures μ_1 on $\bar{\mathcal{S}}_2 \times \bar{\mathcal{A}}_1$ satisfying Assumption 3(ii), that is, for all (s_1, a_1),

$$P_{a_1}(\cdot \mid s_1) \;=\; \langle \psi_1(s_1, a_1), \mu_1(\cdot) \rangle. \tag{47}$$

Since $d_1 = 1$ and $\psi_1(s_1, a_1) = 1$, by Equation 47, we have

$$P_{a_1=0}(\cdot \mid s_1 = 0) = P_{a_1=1}(\cdot \mid s_1 = 0) = P_{a_1=0}(\cdot \mid s_1 = 1) = P_{a_1=1}(\cdot \mid s_1 = 1) = \mu_1(\cdot).$$

Since the input of $P_{a_1}(\cdot \mid s_1)$ is a next history $\big((s_1', s_2'), a_1'\big) \in \bar{\mathcal{S}}_2 \times \bar{\mathcal{A}}_1$, so $P_{a_1}\big((s_1', s_2'), a_1' \mid s_1\big)$ is the probability that the next history equals $\big((s_1', s_2'), a_1'\big)$ when starting from s_1 under $do(A_1 = a_1)$.

And we know that S_2 is uniform, $P_{a_1}(\cdot \mid s_1)$ puts mass $1/2$ on each of the two next histories $\big((s_1, 0), a_1\big)$ and $\big((s_1, 1), a_1\big)$, specifically, $P_{a_1}((s_1, 0), a_1 \mid s_1) = P_{a_1}((s_1, 1), a_1 \mid s_1) = 1/2$.

In particular, at the next history $\big((s_1, 0), a_1\big) = \big((0, 0), 0\big)$,

$$\begin{aligned}
\frac{1}{2} \;&=\; P_{a_1=0}\big(\{((0,0),0)\} \;\big|\; s_1 = 0\big) && (48)\\
&=\; \big\langle \psi_1(0,0),\; \mu_1\big(\{((0,0),0)\}\big) \big\rangle && (49)\\
&=\; \big\langle \psi_1(1,1),\; \mu_1\big(\{((0,0),0)\}\big) \big\rangle && (50)\\
&=\; P_{a_1=1}\big(\{((0,0),0)\} \;\big|\; s_1 = 1\big) && (51)\\
&=\; 0.
\end{aligned}$$

Equation 48 holds since S_2 is uniform; Equation 49 and Equation 51 follow by Assumption 3(ii) at the pairs $(s_1, a_1) = (0,0)$ and $(1,1)$, respectively; Equation 50 holds since $\psi_1(0,0) = \psi_1(1,1) = 1$; and the last equality holds since the next history $\big((0,0),0\big)$ cannot follow $(s_1, a_1) = (1,1)$, which is a contradiction.

This finishes the proof. ■

Proposition 4. *Under Assumption 1, there exist vectors $\{\theta_k^*\}_{k=1}^K$ with $\theta_k^* \in \mathbb{R}^{d_k}$ and $\|\theta_k^*\|_2 \leqslant \sqrt{d_k}$ such that, for all $k \in [K]$ and all $(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$, $Q_k^*(\bar{s}_k, \bar{a}_k) = \langle \psi_k(\bar{s}_k, \bar{a}_k), \theta_k^* \rangle$.*

Proof. We argue by backward induction on k, constructing θ_k^* explicitly at each stage.

(a) Base case: $k = K$. Set $\theta_K^* = w_K$, where w_K is the vector in Assumption 1(i). Then, for every $(\bar{s}_K, \bar{a}_K) \in \bar{\mathcal{S}}_K \times \bar{\mathcal{A}}_K$, we have

$$Q_K^*(\bar{s}_K, \bar{a}_K) \;=\; \mathbb{E}_{\bar{a}_K}\big[Y \mid \bar{s}_K\big] \;=\; \big\langle \psi_K(\bar{s}_K, \bar{a}_K),\; w_K \big\rangle \;=\; \big\langle \psi_K(\bar{s}_K, \bar{a}_K),\; \theta_K^* \big\rangle, \tag{52}$$

where the second equality is Assumption 1(i). The norm bound $\|\theta_K^*\|_2 = \|w_K\|_2 \leqslant \sqrt{d_K}$ holds by construction by Assumption 1(i).

(b) Induction step: $k < K$. Suppose the claim holds at stage $k+1$, that is, there exists $\theta_{k+1}^* \in \mathbb{R}^{d_{k+1}}$ with $\|\theta_{k+1}^*\|_2 \leqslant \sqrt{d_{k+1}}$ and

$$Q_{k+1}^*(\bar{s}_{k+1}, \bar{a}_{k+1}) \;=\; \big\langle \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_{k+1}),\; \theta_{k+1}^* \big\rangle \qquad \forall (\bar{s}_{k+1}, \bar{a}_{k+1}). \tag{53}$$

We first check that $V^*_{k+1} \in \mathcal{V}_{k+1}$. By Equation 53, $Q^*_{k+1} \in \mathcal{Q}_{k+1}$, and since $Y \in [0,1]$ by Assumption 1(i), backward induction gives $Q^*_{k+1} \in [0,1]$. Hence $Q^*_{k+1} \in \mathcal{Q}^{\text{enl}}_{k+1}$, taking $\theta = \theta' = \theta^*_{k+1}$, $A = A' = 0$ and $L = U = Q^*_{k+1}$ in Equation 3, and therefore $V^*_{k+1} = \max_{a_{k+1}} Q^*_{k+1} \in \mathcal{V}_{k+1}$.

Then, for every $(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k$,

$$Q^*_k(\bar{s}_k, \bar{a}_k) = \mathbb{E}_{\bar{a}_k}\left[V^*_{k+1}\big((\bar{s}_k, S_{k+1}), \bar{a}_k\big) \mid \bar{s}_k\right] \tag{54}$$

$$= \left\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_{V^*_{k+1}} \right\rangle \tag{55}$$

$$= \left\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta^*_k \right\rangle. \tag{56}$$

Equation 54 is the Bellman equation for Q^*_k; Equation 55 follows by Assumption 1(ii) applied to $V^*_{k+1} \in \mathcal{V}_{k+1}$; and Equation 56 is by setting $\theta^*_k := \theta_{V^*_{k+1}}$.

The norm bound $\|\theta^*_k\|_2 \leqslant \sqrt{d_k}$ holds by construction, again by Assumption 1(ii).

Proceeding backward from $k = K$ to $k = 1$ produces the vectors $\{\theta^*_k\}_{k=1}^K$ with the stated properties. ■

D Proof for Section 3

D.1 Proof for Offline Confidence Set

Proposition 2. *Under Assumption 1, suppose $P(\bar{s}_K, \bar{a}_K, y) > 0$ for all $(\bar{s}_K, \bar{a}_K, y)$. Fix $\delta_1 \in (0,1]$, and let $L_k(\bar{s}_k, \bar{a}_k), U_k(\bar{s}_k, \bar{a}_k)$, $k \in [K]$, be the estimated causal bounds above. Let*

$$\widehat{C}^{\text{off}}_k = \Big\{ \theta_k \in \mathbb{R}^{d_k} : \ L_k(\bar{s}_k, \bar{a}_k) \ \leqslant \ \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k \ \leqslant \ U_k(\bar{s}_k, \bar{a}_k) \ \ \forall (\bar{s}_k, \bar{a}_k) \Big\}.$$

Then $P\big(\forall\, k \in [K],\ \theta^*_k \in \widehat{C}^{\text{off}}_k\big) \geqslant 1 - \delta_1$.

Proof. *Step 1. Estimated causal bounds.* Let $\mathcal{H} := \bigcup_{k=1}^K \left(\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k \bigcup \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1}\right)$ denote the collection of histories with $|\mathcal{H}| = \mathcal{O}(\sum_{k=1}^K \prod_{i=1}^k |\mathcal{S}_i||\mathcal{A}_i|)$. For any $\delta_1 \in (0,1]$, we set the bonus as

$$b_{\text{off}} = \sqrt{\frac{\log(2|\mathcal{H}|/\delta_1)}{2N_{\text{off}}}}.$$

We write $\mathbb{P}^\star$[6] and $m^\star_K$ for the following two functions on histories, the true unknown quantities underlying the offline data:

$$\mathbb{P}^\star(\bar{s}_k, \bar{a}_k) = P\left(\bar{S}_k = \bar{s}_k,\ \bar{A}_k = \bar{a}_k\right),$$

$$m^\star_K(\bar{s}_K, \bar{a}_K) = \mathbb{E}\left[Y\, \mathbf{1}\left\{\bar{S}_K = \bar{s}_K,\ \bar{A}_K = \bar{a}_K\right\}\right].$$

And $\widehat{\mathbb{P}}$ and $\widehat{m}_K$ are the empirical estimates:

$$\widehat{\mathbb{P}}(\bar{s}_k, \bar{a}_k) = \frac{1}{N_{\text{off}}} \sum_{i=1}^{N_{\text{off}}} \mathbf{1}\{\bar{S}_{i,k} = \bar{s}_k,\ \bar{A}_{i,k} = \bar{a}_k\}, \tag{57}$$

$$\widehat{m}_K(\bar{s}_K, \bar{a}_K) = \frac{1}{N_{\text{off}}} \sum_{i=1}^{N_{\text{off}}} Y_i\, \mathbf{1}\{\bar{S}_{i,K} = \bar{s}_K,\ \bar{A}_{i,K} = \bar{a}_K\}. \tag{58}$$

We define a confidence box for variables $(\mathbb{P}, m_K)$ as

$$\mathcal{B}^{\text{off}} = \Big\{(\mathbb{P}, m_K) : \ |\mathbb{P}(h) - \widehat{\mathbb{P}}(h)| \leqslant b_{\text{off}}\ \forall h \in \mathcal{H};\ |m_K(\bar{s}_K, \bar{a}_K) - \widehat{m}_K(\bar{s}_K, \bar{a}_K)| \leqslant b_{\text{off}} \ \ \forall (\bar{s}_K, \bar{a}_K);$$
$$0 < \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_{k-1}) \quad \forall k \in [K-1];$$

[6] Here, $\mathbb{P}$ denotes the history probability function, $P(\cdot)$ denotes the probability of an event and $P_{\bar{a}_k}$ the interventional probability.

$$0 < \mathbb{P}(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_{K-1}); \quad 0 \leqslant m_K(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_K)\Big\}, \tag{59}$$

where the pair $(\mathbb{P}, m_K)$ is the optimization variable ranging over $\mathcal{B}^{\text{off}}$. Here, $\mathbb{P}$ is a family of probabilities indexed by the histories $h \in \mathcal{H}$:

$$\mathbb{P} = \big(\mathbb{P}(h)\big)_{h\in\mathcal{H}} = \Big(\mathbb{P}(s_1),\, \mathbb{P}(s_1, a_1),\, \mathbb{P}(\bar{s}_2, a_1),\, \mathbb{P}(\bar{s}_2, \bar{a}_2),\, \ldots\Big).$$

Likewise, m_K is a family indexed by the terminal histories: $m_K = \Big(m_K(\bar{s}_K, \bar{a}_K)\Big)_{(\bar{s}_K, \bar{a}_K)}$.

Let $\mathcal{E}^{\text{off}}$ be the good event for all stages in the offline dataset, i.e.,

$$\mathcal{E}^{\text{off}} := \Big\{\big|\widehat{\mathbb{P}}(h) - \mathbb{P}^\star(h)\big| \leqslant b_{\text{off}}\ \forall h \in \mathcal{H}\Big\} \bigcap \Big\{\big|\widehat{m}_K(\bar{s}_K, \bar{a}_K) - m_K^\star(\bar{s}_K, \bar{a}_K)\big| \leqslant b_{\text{off}}\ \forall (\bar{s}_K, \bar{a}_K)\Big\}. \tag{60}$$

Then on the event $\mathcal{E}^{\text{off}}$, we have $(\mathbb{P}^\star, m_K^\star) \in \mathcal{B}^{\text{off}}$.

We then express the causal bounds of (Zhang & Bareinboim, 2019, Theorem 5, Corollary 1) as functions of $(\mathbb{P}, m_K)$. Write

$$\zeta(\bar{s}_{k+1}, \bar{a}_k; \mathbb{P}) = \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) - \mathbb{P}(\bar{s}_k, \bar{a}_k) + \zeta(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P}), \qquad k \in [K-1], \tag{61}$$

with $\zeta(s_1; \mathbb{P}) = \mathbb{P}(s_1)$,

For the stage K, we define

$$\underline{\mu}(\mathbb{P}; m_K) = \frac{m_K(\bar{s}_K, \bar{a}_K)}{\zeta(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})}, \quad \overline{\mu}(\mathbb{P}; m_K) = 1 - \frac{\mathbb{P}(\bar{s}_K, \bar{a}_K) - m_K(\bar{s}_K, \bar{a}_K)}{\zeta(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})}. \tag{62}$$

By (Zhang & Bareinboim, 2019, Corollary 1), we have, $\underline{\mu}(\mathbb{P}^\star; m_K^\star) \leqslant \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \leqslant \overline{\mu}(\mathbb{P}^\star; m_K^\star)$.

For the stage $k \in [K-1]$, we define

$$\underline{\tau}_k(s_{k+1}; \mathbb{P}) = \frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\zeta(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})} \tag{63}$$

$$\overline{\tau}_k(s_{k+1}; \mathbb{P}) = 1 - \frac{\mathbb{P}(\bar{s}_k, \bar{a}_k) - \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\zeta(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})}. \tag{64}$$

By (Zhang & Bareinboim, 2019, Theorem 5), we have $\underline{\tau}_k(s_{k+1}; \mathbb{P}^*) \leqslant P_{\bar{a}_k}(s_{k+1}|\bar{s}_k) \leqslant \overline{\tau}_k(s_{k+1}; \mathbb{P}^*)$. And we further define a transition set as

$$\widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k) = \big\{p \in \Delta(\mathcal{S}_{k+1}) : \inf_{\mathcal{B}^{\text{off}}} \underline{\tau}_k(s_{k+1}; \mathbb{P}) \leqslant p(s_{k+1}) \leqslant \sup_{\mathcal{B}^{\text{off}}} \overline{\tau}_k(s_{k+1}; \mathbb{P})\ \ \forall s_{k+1} \in \mathcal{S}_{k+1}\big\}, \tag{65}$$

Then set the estimated causal bounds for stage $k \in [K]$:

$$L_K(\bar{s}_K, \bar{a}_K) = \max\Big\{\inf_{\mathcal{B}^{\text{off}}} \underline{\mu}(\mathbb{P}, m_K),\ 0\Big\}, \tag{66}$$

$$U_K(\bar{s}_K, \bar{a}_K) = \min\Big\{\sup_{\mathcal{B}^{\text{off}}} \overline{\mu}(\mathbb{P}, m_K),\ 1\Big\}, \tag{67}$$

$$L_k(\bar{s}_k, \bar{a}_k) = \min_{p\in\widehat{\mathcal{P}}_k} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} L_{k+1}\big(\bar{s}_k, s_{k+1}, \bar{a}_k, a_{k+1}\big), \tag{68}$$

$$U_k(\bar{s}_k, \bar{a}_k) = \max_{p\in\widehat{\mathcal{P}}_k} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} U_{k+1}\big(\bar{s}_k, s_{k+1}, \bar{a}_k, a_{k+1}\big). \tag{69}$$

If $\widehat{\mathcal{P}}_k$ is empty, $L_k(\bar{s}_k, \bar{a}_k) = 0$ and $U_k(\bar{s}_k, \bar{a}_k) = 1$.

Step 2. Concentration. By Hoeffding's inequality (Lemma 11) and a union bound over $\mathcal{H}$, $P(\mathcal{E}^{\text{off}}) \geqslant 1 - \delta_1$. On $\mathcal{E}^{\text{off}}$ we have $(\mathbb{P}^\star, m_K^\star) \in \mathcal{B}^{\text{off}}$, then,

$$\inf_{\mathcal{B}^{\text{off}}} \underline{\tau}_k(s_{k+1}; \mathbb{P}) \leqslant \underline{\tau}_k(s_{k+1}; \mathbb{P}^\star), \qquad \overline{\tau}_k(s_{k+1}; \mathbb{P}^\star) \leqslant \sup_{\mathcal{B}^{\text{off}}} \overline{\tau}_k(s_{k+1}; \mathbb{P}), \tag{70}$$

Then, by eqs. (65) and (70),

$$P_{\bar{a}_k}(\cdot \mid \bar{s}_k) \in \widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k). \tag{71}$$

The same thing applied to $\underline{\mu}(\mathbb{P}; m_K), \overline{\mu}(\mathbb{P}; m_K)$ gives $L_K \leqslant \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \leqslant U_K$.

Step 3. Backward propagation to Q_k^*. We show by backward induction on k that

$$L_k(\bar{s}_k, \bar{a}_k) \leqslant Q_k^*(\bar{s}_k, \bar{a}_k) \leqslant U_k(\bar{s}_k, \bar{a}_k). \tag{72}$$

(a). Base case, at $k = K$ this is the terminal statement of Step 2, since $Q_K^*(\bar{s}_K, \bar{a}_K) = \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ by Proposition 4.

(b). Induction step, assuming Equation 72 at stage $k+1$, i.e., $L_{k+1}(\bar{s}_{k+1}, \bar{a}_{k+1}) \leqslant Q_{k+1}^*(\bar{s}_{k+1}, \bar{a}_{k+1}) \leqslant U_{k+1}(\bar{s}_{k+1}, \bar{a}_{k+1})$. Then, we have,

$$Q_k^*(\bar{s}_k, \bar{a}_k) = \sum_{s_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \max_{a_{k+1}} Q_{k+1}^* \tag{73}$$

$$\leqslant \sum_{s_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \max_{a_{k+1}} U_{k+1} \tag{74}$$

$$\leqslant \max_{p \in \widehat{\mathcal{P}}_k} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} U_{k+1} \tag{75}$$

$$= U_k(\bar{s}_k, \bar{a}_k), \tag{76}$$

where Equation 73 follows by Equation 1; Equation 74 uses the induction hypothesis; Equation 75 is because $P_{\bar{a}_k}(\cdot \mid \bar{s}_k) \in \widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k)$ in Equation 71; and Equation 76 is by Equation 69.

The lower bound follows identically, with $\min$ in place of $\max$ and L_{k+1} in place of U_{k+1}.

Step 4. Conclusion. By Assumption 1 and Proposition 4, $Q_k^*(\bar{s}_k, \bar{a}_k) = \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k^*$, so Equation 72 gives $L_k \leqslant \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k^* \leqslant U_k$ at every history and every $k \in [K]$.

Then, on $\mathcal{E}^{\text{off}}$, we have the following holds,

$$P\left(\forall(\bar{s}_k, \bar{a}_k) : L_k(\bar{s}_k, \bar{a}_k) \leqslant \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k^* \leqslant U_k(\bar{s}_k, \bar{a}_k)\right) \geqslant 1 - \delta_1. \tag{77}$$

And the estimated confidence set at stage $k \in [K]$ is:

$$\widehat{C}_k^{\text{off}} = \left\{\theta_k \in \mathbb{R}^{d_k} : L_k(\bar{s}_k, \bar{a}_k) \leqslant \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k \leqslant U_k(\bar{s}_k, \bar{a}_k),\ \forall(\bar{s}_k, \bar{a}_k)\right\}.$$

This finishes the proof. ■

D.2 Proof for the Regret Bound

Theorem 1. *Under Assumption 1, let $\chi := \log\left(\frac{4(c_\beta+1)d_K KT}{\delta}\right)$ and absolute constants $C_1 > 0$, $C > 0$, $c_\beta \geqslant C + 1$. Set the confidence radii as $\beta_k = Cd_K\sqrt{\chi} + \sqrt{\lambda d_k}$. For any $\delta \in (0, 1]$, $\delta_1 \in (0, 1]$, then with probability at least $1 - \delta - 2\delta_1$, Algorithm 1 achieves the following result:*

$$R(T) \leqslant 2\sqrt{2TK\log(2/\delta_1)} + \min\left\{2C_1 d_K^{3/2} K\sqrt{T \cdot \chi \log(\lambda + T)},\ T\sum_{k=1}^{K} W_k^{\text{off}}\right\}.$$

Proof. *Step 1. Bound for the estimated linear parameter.* For any non-terminal stage $k \in [K-1]$, the ridge estimator is

$$\|\widehat{\theta}_k^t\|_2 = \left\|(\Lambda_k^t)^{-1}\sum_{\tau=1}^{t-1}\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\max_{a_{k+1}}\widehat{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1})\right\|_2 \tag{78}$$

$$\leqslant \frac{1}{\sqrt{\lambda}}\left\|\sum_{\tau=1}^{t-1}\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\max_{a_{k+1}}\widehat{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1})\right\|_{(\Lambda_k^t)^{-1}}. \tag{79}$$

$$\leqslant \frac{1}{\sqrt{\lambda}}\sqrt{\left(\sum_{\tau=1}^{t-1}\left[\max_{a_{k+1}}\widehat{Q}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau, a_{k+1})\right]^2\right)\operatorname{tr}\left((\Lambda_k^t)^{-1}\sum_{\tau=1}^{t-1}\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)^\top\right)} \tag{80}$$

$$= \frac{1}{\sqrt{\lambda}} \sqrt{\Big(\sum_{\tau=1}^{t-1} \Big[\max_{a_{k+1}} \widehat{Q}_{k+1}^{t}(\bar{s}_{k+1}^{\tau}, \bar{a}_{k}^{\tau}, a_{k+1}) \Big]^2 \Big) \operatorname{tr}\left((\Lambda_k^t)^{-1}(\Lambda_k^t - \lambda I)\right)} \tag{81}$$

$$\leqslant \sqrt{\frac{d_k}{\lambda}} \sqrt{\sum_{\tau=1}^{t-1} \Big[\max_{a_{k+1}} \widehat{Q}_{k+1}^{t}(\bar{s}_{k+1}^{\tau}, \bar{a}_{k}^{\tau}, a_{k+1}) \Big]^2} \tag{82}$$

$$\leqslant \sqrt{\frac{d_k t}{\lambda}}. \tag{83}$$

Equation 78 follows by Equation 10; Equation 79 is because $\Lambda_k^t \succeq \lambda I$; Equation 80 comes from Cauchy–Schwarz inequality; Equation 81 is by the definition of Λ_k^t; Equation 82 follows by

$$\operatorname{tr}\left((\Lambda_k^t)^{-1}(\Lambda_k^t - \lambda I)\right) \;=\; \operatorname{tr}\left(I - \lambda(\Lambda_k^t)^{-1}\right) \;\leqslant\; d_k;$$

Equation 83 holds since $\max_{a_{k+1}} \widehat{Q}_{k+1}^{t}(\bar{s}_{k+1}^{\tau}, \bar{a}_{k}^{\tau}, a_{k+1}) < 1, \forall t \in [T], k \in [K-1]$.

Similarly, we could show that the same results hold for the terminal stage $k = K$, i.e., $\|\widehat{\theta}_K^t\|_2 \leqslant \sqrt{d_K t/\lambda}$.

Step 2. Concentration bounds. Firstly, we bound the empirical noise term $\widehat{V}_{k+1}^{t}(\bar{s}_{k+1}^{\tau}, \bar{a}_{k}^{\tau}) - \mathbb{E}_{\bar{a}_k^{\tau}}[\widehat{V}_{k+1}^{t}(\bar{S}_{k+1}, \bar{a}_{k}^{\tau}) \mid \bar{s}_{k}^{\tau}]$, which is the difference between sample next-state value and conditional expected next-state value. In the algorithm, $\widehat{V}_{k+1}^{t}$ is computed from the data. And, we need a uniform bound over the whole value-function class $\mathcal{V}_{k+1}$.

Fix a stage $k < K$. Assumption 1(ii) is imposed over the class $\mathcal{V}_{k+1}$ built from Equation 3. For the concentration argument it suffices to control the subfamily of $\mathcal{V}_{k+1}$ that Algorithm 1 can produce,

$$\widetilde{\mathcal{V}}_{k+1} := \Big\{ V : \; V(\bar{s}_{k+1}, \bar{a}_k) = \max\Big\{ \max_{a_{k+1} \in \mathcal{A}_{k+1}} \min\Big\{ 1, \; \langle \omega, \, \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \rangle \\ + \beta \|\psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1})\|_{\Lambda^{-1}}, \; \max_{\theta \in \widehat{\mathcal{C}}_{k+1}^{\mathrm{off}}} \theta^\top \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \Big\}, \; 0 \Big\} \Big\},$$

where the parameters satisfy

$$\omega \in \mathbb{R}^{d_{k+1}}, \quad \|\omega\|_2 \leqslant \sqrt{d_{k+1} T/\lambda}, \quad \beta \in [0, B], \quad \Lambda \in \mathbb{R}^{d_{k+1} \times d_{k+1}}, \quad \lambda_{\min}(\Lambda) \geqslant \lambda.$$

The offline term $\max_{\theta \in \widehat{\mathcal{C}}_{k+1}^{\mathrm{off}}} \theta^\top \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1})$ is deterministic conditioned on the offline dataset, which is independent of the online trajectories; the offline term does not depend on (ω, β, Λ) and is therefore not a parameter of $\widetilde{\mathcal{V}}_{k+1}$, only (ω, β, Λ) vary.

Note that $\widetilde{\mathcal{V}}_{k+1} \subseteq \mathcal{V}_{k+1}$: since $\beta\|\psi\|_{\Lambda^{-1}} = \|\psi\|_{\beta^2\Lambda^{-1}}$, every $V \in \widetilde{\mathcal{V}}_{k+1}$ is of the form of Equation 3 with $\theta = \theta' = \omega$, $A = A' = \beta^2\Lambda^{-1}$ and $\mathcal{P} = \widehat{\mathcal{C}}_{k+1}^{\mathrm{off}} \in \mathfrak{P}_{k+1}$ (i.e., $L = L_{k+1}$, $U = U_{k+1}$), maximized over a_{k+1}; hence Assumption 1(ii) applies on $\widetilde{\mathcal{V}}_{k+1}$. Moreover $\widehat{V}_{k+1}^{t} \in \widetilde{\mathcal{V}}_{k+1}$ by taking $\omega = \widehat{\theta}_{k+1}^{t}$, $\beta = \beta_{k+1} \leqslant B$ and $\Lambda = \Lambda_{k+1}^{t}$.

For a fixed value function $V \in \widetilde{\mathcal{V}}_{k+1}$, define the martingale

$$M_\tau(V) := V(\bar{s}_{k+1}^{\tau}, \bar{a}_k^{\tau}) - \mathbb{E}_{\bar{a}_k^{\tau}}\left[V(\bar{S}_{k+1}, \bar{a}_k^{\tau}) \mid \bar{s}_k^{\tau}\right].$$

Then $\mathbb{E}_{\bar{a}_k^{\tau}}[M_\tau(V) \mid \bar{s}_k^{\tau}] = 0$. Since $0 \leqslant V \leqslant 1$, we have $|M_\tau(V)| \leqslant 1$. Thus, for any fixed $V \in \widetilde{\mathcal{V}}_{k+1}$, the sequence $\{\psi_k^\tau M_\tau(V)\}_{\tau=1}^{t-1}$ is a self-normalized martingale-difference sequence with respect to $\{\mathcal{F}_{k,\tau}\}_\tau$. And $\{\mathcal{F}_{k,\tau}\}_\tau$ is the filtration such that $\psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau)$ is $\mathcal{F}_{k,\tau-1}$-measurable, i.e., $\mathcal{F}_{k,\tau-1} := \sigma\Big(\{(\bar{s}_K^i, \bar{a}_K^i, Y^i)\}_{i<\tau} \bigcup \{(\bar{s}_k^\tau, \bar{a}_k^\tau)\}\Big)$.

Let $\mathcal{N}_{k+1,\epsilon}$ denote the ϵ-covering number of $\widetilde{\mathcal{V}}_{k+1}$ under the metric

$$\operatorname{dist}(V, V') := \sup_{(\bar{s}_{k+1}, \bar{a}_k) \in \overline{\mathcal{S}}_{k+1} \times \overline{\mathcal{A}}_k} |V(\bar{s}_{k+1}, \bar{a}_k) - V'(\bar{s}_{k+1}, \bar{a}_k)|,$$

for $V, V' \in \widetilde{\mathcal{V}}_{k+1}$. By the uniform self-normalized concentration inequality from Lemma 14, for any $\delta > 0$, with probability at least $1 - \delta/(2(K-1))$, for all $t \in [T]$ and all $V \in \widetilde{\mathcal{V}}_{k+1}$,

$$\Big\| \sum_{\tau=1}^{t-1} \psi_k^\tau \left\{ V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}\left[V(\bar{S}_{k+1}^\tau, \bar{a}_k^\tau) \mid \mathcal{F}_{k,\tau-1} \right] \right\} \Big\|^2_{(\Lambda_k^t)^{-1}} \leqslant 4\Big[\frac{d_k}{2} \log\left(\frac{t+\lambda}{\lambda} \right) + \log\left(\frac{2(K-1)\mathcal{N}_{k+1,\epsilon}}{\delta} \right) \Big] + \frac{8t^2\epsilon^2}{\lambda}. \tag{84}$$

Then the remaining task is to control the covering number $\mathcal{N}_{k+1,\epsilon}$.

The class $\widetilde{\mathcal{V}}_{k+1}$ is parametrized by (ω, β, Λ) with $\|\omega\|_2 \leqslant \sqrt{d_{k+1}T/\lambda}$, $\beta \in [0, B]$ and $\lambda_{\min}(\Lambda) \geqslant \lambda$. Since V depends on (ω, β, Λ) through a composition of maps that are non-expansive in the supremum norm, any ϵ-cover of the parameters induces an ϵ-cover of $\widetilde{\mathcal{V}}_{k+1}$. By covering-number bound from Lemma 15, the ϵ-covering number satisfies

$$\log \mathcal{N}_{k+1,\epsilon} \leqslant d_K \log\left(1 + \frac{4\sqrt{d_K T/\lambda}}{\epsilon} \right) + d_K^2 \log\left(1 + \frac{8 d_K^{1/2} B^2}{\lambda \epsilon^2} \right), \tag{85}$$

where Equation 85 holds since $d_k \leqslant d_K,\ \forall k \in [K]$.

Now choose $\epsilon = \frac{d_K}{T}$, $\lambda = 1$, and $B = c_\beta d_K \sqrt{\chi}$ with $\chi = \log(\frac{4(c_\beta+1)d_K KT}{\delta})$ and constant $c_\beta > 0$. Then by eqs. (84) and (85)

$$\left\| \sum_{\tau=1}^{t-1} \psi_k^\tau \left[V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\left[V(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau \right] \right] \right\|_{(\Lambda_k^t)^{-1}} \leqslant C d_K \sqrt{\log\left(\frac{4(c_\beta+1)d_K KT}{\delta} \right)},$$

for some absolute constant $C > 0$. Since the above bound holds uniformly over all $V \in \widetilde{\mathcal{V}}_{k+1}$, and $\widehat{V}_{k+1}^t \in \widetilde{\mathcal{V}}_{k+1}$, we may take $V = \widehat{V}_{k+1}^t$ even though $\widehat{V}_{k+1}^t$ is data-dependent.

Taking a union bound over the $K-1$ non-terminal stages, with probability at least $1 - \delta/2$, simultaneously for all $k < K$ and all $t \in [T]$

$$\left\| \sum_{\tau=1}^{t-1} \psi_k^\tau \left[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau \right] \right] \right\|_{(\Lambda_k^t)^{-1}} \leqslant C d_K \sqrt{\log\left(\frac{4(c_\beta+1)d_K KT}{\delta} \right)}. \tag{86}$$

Step 3. Optimism bounds. The online confidence set is $\widehat{\mathcal{C}}_k^t = \{\theta \in \mathbb{R}^{d_k} : \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \leqslant \beta_k\}$, for simplicity, we denote it as $\mathcal{B}_k^t$.

For $k < K$, define an online event as following:

$$\mathcal{E}_1 := \Big\{ \forall k < K, \forall t \in [T], \forall V \in \widetilde{\mathcal{V}}_{k+1} : \Big\| \sum_{\tau=1}^{t-1} \psi_k^\tau \Big(V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\left[V(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau \right] \Big) \Big\|_{(\Lambda_k^t)^{-1}} \leqslant \rho \Big\}$$
$$\bigcap \Big\{ \forall t \in [T] : \Big\| \sum_{\tau=1}^{t-1} \psi_K^\tau \left[Y^\tau - \mathbb{E}\left[Y^\tau \mid \mathcal{F}_{K,\tau-1} \right] \right] \Big\|_{(\Lambda_K^t)^{-1}} \leqslant \rho \Big\}. \tag{87}$$

For simplicity, we write $\rho = C d_K \sqrt{\log\left(\frac{4(c_\beta+1)d_K KT}{\delta} \right)}$. By Lemmas 14 and 15, and the standard self-normalized concentration inequality for terminal stage K, we then have $P(\mathcal{E}_1) \geqslant 1 - \delta$.

By Equation 60, the offline good event is the following:

$$\mathcal{E}^{\text{off}} := \Big\{ \big| \widehat{P}(h) - P(h) \big| \leqslant b_{\text{off}}\ \forall h \in \mathcal{H}_P \Big\} \bigcap \Big\{ \big| \widehat{m}_K(\bar{s}_K, \bar{a}_K) - m_K(\bar{s}_K, \bar{a}_K) \big| \leqslant b_{\text{off}}\ \forall (\bar{s}_K, \bar{a}_K) \Big\},$$

Then, by the offline construction, we have $P(\mathcal{E}^{\text{off}}) \geqslant 1 - \delta_1$. Set

$$\mathcal{E} := \mathcal{E}_1 \bigcap \mathcal{E}^{\text{off}}. \tag{88}$$

Therefore, $P(\mathcal{E}) \geqslant 1 - \delta - \delta_1$.

Then on the good event $\mathcal{E}$, by Lemma 2, 3, and 4, $\forall k \in [K],\ \forall t \in [T],\ \forall (\bar{s}_k, \bar{a}_k)$, we have

$$\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \ \geqslant \ Q_k^*(\bar{s}_k, \bar{a}_k). \tag{89}$$

Step 4: Recursive formula for estimation error. Along the $\boldsymbol{\pi}_t$-trajectory $(s_1^t, a_1^t, \bar{s}_2^t, a_2^t, \ldots)$, for stage $k < K$, define

$$\zeta_{k+1}^t := \mathbb{E}_{\bar{a}_k^t}\left[\widehat{V}_{k+1}^t(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) - V_{k+1}^{\boldsymbol{\pi}_t}(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) \,\middle|\, \bar{s}_k^t\right] - \left[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^t, \bar{a}_k^t) - V_{k+1}^{\boldsymbol{\pi}_t}(\bar{s}_{k+1}^t, \bar{a}_k^t)\right] \tag{90}$$

and $\zeta_{K+1}^t = 0$ for stage K. By construction ζ_{k+1}^t is a martingale difference.

At stage k, the algorithm selects decision

$$a_k^t = \pi_{t,k}(\bar{s}_k^t, \bar{a}_{k-1}^t) := \arg\max_{a_k} \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k). \tag{91}$$

Then, we have,

$$\widehat{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) \ = \ \max_{a_k} \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k) \ = \ \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k^t) \ = \ \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t). \tag{92}$$

Now consider the true value under the policy $\boldsymbol{\pi}_t$:

$$\begin{aligned}
V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t) \ &= \ \mathbb{E}_{a_k \sim \pi_{t,k}(\cdot \mid \bar{s}_k, \bar{a}_{k-1})}\left[Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k, \bar{a}_{k-1}, a_k)\right] && (93)\\
&= \ \sum_{a_k \in \mathcal{A}} \pi_{t,k}\left(a_k \mid \bar{s}_k^t, \bar{a}_{k-1}^t\right) Q_k^{\boldsymbol{\pi}_t}\left(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k\right) && (94)\\
&= \ Q_k^{\boldsymbol{\pi}_t}\left(\bar{s}_k^t, \bar{a}_{k-1}^t, a_k^t\right) && (95)\\
&= \ Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t). && (96)
\end{aligned}$$

Equation 93 is the definition of the value function under policy $\pi_{t,k}$; Equation 94 comes from the expectation over policy $\pi_{t,k}$; Equation 95 is by Equation 91.

Combining Equations (92) and (96) gives

$$\widehat{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t) = \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t), \tag{97}$$

To bound $\widehat{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t)$, it is sufficient to bound $\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t)$.

(a). For the online case,

$$\begin{aligned}
\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) \ &= \ \max\left\{\min\left\{\max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k^t, \bar{a}_k^t), \max_{\theta \in \widehat{\mathcal{C}}_k^{\text{off}}} \theta^\top \psi_k(\bar{s}_k^t, \bar{a}_k^t), 1\right\}, 0\right\} && (98)\\
&\leqslant \ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k^t, \bar{a}_k^t) && (99)\\
&= \ \widehat{\theta}_k^{t\top} \psi_k(\bar{s}_k^t, \bar{a}_k^t) + \max_{\theta \in \widehat{\mathcal{C}}_k^t} (\theta - \widehat{\theta}_k^t)^\top \psi_k(\bar{s}_k^t, \bar{a}_k^t) && (100)\\
&\leqslant \ \widehat{\theta}_k^{t\top} \psi_k(\bar{s}_k^t, \bar{a}_k^t) + \max_{\theta \in \widehat{\mathcal{C}}_k^t} \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}} && (101)\\
&\leqslant \ \widehat{\theta}_k^{t\top} \psi_k(\bar{s}_k^t, \bar{a}_k^t) + \beta_k \|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}}, && (102)
\end{aligned}$$

where Equation 98 follows by Equation 11; Equation 99 holds since $\max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k^t, \bar{a}_k^t) \geqslant 0$, which follows from Lemma 2 together with Assumption 1(ii) for $k < K$ and with $Q_K^* \geqslant 0$ for $k = K$; Equation 100 holds since $\theta = \widehat{\theta}_k^t + \left(\theta - \widehat{\theta}_k^t\right)$; Equation 101 follows by Cauchy-Schwarz inequality, i.e., $(\theta - \widehat{\theta}_k^t)^\top \psi_k \leqslant \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \|\psi_k\|_{(\Lambda_k^t)^{-1}}$; Equation 102 is from the online confidence set $\widehat{\mathcal{C}}_k^t \leftarrow \left\{\theta \in \mathbb{R}^{d_k} : \|\theta - \widehat{\theta}_k^t\|_{\Lambda_k^t} \leqslant \beta_k\right\}$.

Then, we have,

$$\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t)$$

$$\leqslant \left\langle \psi_k(\bar{s}_k^t, \bar{a}_k^t), \hat{\theta}_k^t \right\rangle - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) + \beta_k \|\psi_k\|_{(\Lambda_k^t)^{-1}} \tag{103}$$

$$\leqslant \mathbb{E}_{\bar{a}_k^t}\left[(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) \mid \bar{s}_k^t\right] + \mathcal{C}\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}} + \beta_k \|\psi_k\|_{(\Lambda_k^t)^{-1}} \tag{104}$$

$$\leqslant \mathbb{E}_{\bar{a}_k^t}\left[(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) \mid \bar{s}_k^t\right] + (\mathcal{C} + \beta_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}}. \tag{105}$$

Equation 103 follows by Equation 102; Equation 104 is by Proposition 5.

(b). For the offline case, we have,

$$\begin{aligned}
&\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\
&= \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^*(\bar{s}_k^t, \bar{a}_k^t) + Q_k^*(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\
&\leqslant U_k - L_k + Q_k^*(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) && (106) \\
&\leqslant U_k - L_k + \mathbb{E}_{\bar{a}_k^t}\big[\max_{a_{k+1} \in \mathcal{A}} Q_{k+1}^*(\bar{s}_{k+1}^t, \bar{a}_k^t, a_{k+1}) \,|\bar{s}_k^t\big] - \mathbb{E}_{\bar{a}_k^t}\big[V_{k+1}^{\boldsymbol{\pi}_t}(\bar{s}_{k+1}^t, \bar{a}_k^t) \mid \bar{s}_k^t\big] && (107) \\
&= U_k - L_k + \mathbb{E}_{\bar{a}_k^t}\big[V_{k+1}^*(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \,|\bar{s}_k^t\big] - \mathbb{E}_{\bar{a}_k^t}\big[V_{k+1}^{\boldsymbol{\pi}_t}(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \mid \bar{s}_k^t\big] && (108) \\
&\leqslant U_k - L_k + \mathbb{E}_{\bar{a}_k^t}\big[\widehat{V}_{k+1}^t(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \,|\bar{s}_k^t\big] - \mathbb{E}_{\bar{a}_k^t}\big[V_{k+1}^{\boldsymbol{\pi}_t}(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \mid \bar{s}_k^t\big]. && (109)
\end{aligned}$$

Equation 106 holds because both $\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t)$ and $Q_k^*(\bar{s}_k^t, \bar{a}_k^t)$ lie in the band $[L_k(\bar{s}_k^t, \bar{a}_k^t), U_k(\bar{s}_k^t, \bar{a}_k^t)]$:

- Since $\theta_k^* \in \widehat{\mathcal{C}}_k^{\text{off}}$ on $\mathcal{E}^{\text{off}}$, and $Q_k^*(\bar{s}_k^t, \bar{a}_k^t) = \left\langle \psi_k(\bar{s}_k^t, \bar{a}_k^t), \theta_k^* \right\rangle$ by Proposition 4, we have $Q_k^*(\bar{s}_k^t, \bar{a}_k^t) \in [L_k(\bar{s}_k^t, \bar{a}_k^t), U_k(\bar{s}_k^t, \bar{a}_k^t)]$ by Equation 77,
- By Equations (11) and (77), we have $\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) \leqslant \max_{\theta \in \widehat{\mathcal{C}}_k^{\text{off}}} \theta^\top \psi_k \leqslant U_k(\bar{s}_k^t, \bar{a}_k^t)$,
- By Equations (77) and (89), we have $\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) \geqslant Q_k^*(\bar{s}_k^t, \bar{a}_k^t) \geqslant L_k(\bar{s}_k^t, \bar{a}_k^t)$;

Equation 107 follows by Equations (33) and (34); Equation 108 is from Equation 31; and Equation 109 comes from Equation 77 and the following:

$$\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k) = \max_{a_{k+1}} \widehat{Q}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k, a_{k+1}) \geqslant \max_{a_{k+1}} Q_{k+1}^*(\bar{S}_{k+1}, \bar{a}_k, a_{k+1}) = V_{k+1}^*(\bar{S}_{k+1}, \bar{a}_k).$$

Then, by eqs. (105) and (109)

$$\begin{aligned}
&\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\
&\leqslant \big(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t}\big)(\bar{s}_{k+1}^t, \bar{a}_k^t) + \zeta_{k+1}^t + \min\Big\{(\mathcal{C} + \beta_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}}, U_k - L_k\Big\} && (110) \\
&= \big(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t}\big)(\bar{s}_{k+1}^t, \bar{a}_k^t) + \zeta_{k+1}^t + b_k^t, && (111)
\end{aligned}$$

where Equation 110 follows by Equation 90; and Equation 111 is by denoting $b_k^t := \min\big\{(\mathcal{C} + \beta_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}}, U_k - L_k\big\}$.

Step 5: Finalize the results. By eqs. (97) and (111), we have,

$$\widehat{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t) \leqslant \big(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t}\big)(\bar{s}_{k+1}^t, \bar{a}_k^t) + \zeta_{k+1}^t + b_k^t. \tag{112}$$

By Equation 89, we have the following,

$$V_1^*(s_1^t) - V_1^{\boldsymbol{\pi}_t}(s_1^t) \leqslant \widehat{V}_1^t(s_1^t) - V_1^{\boldsymbol{\pi}_t}(s_1^t). \tag{113}$$

Then, we look at Equation 112 for $k = 1, 2, \ldots, K$,

$$\begin{aligned}
\widehat{V}_1^t(s_1^t) - V_1^{\boldsymbol{\pi}_t}(s_1^t) &\leqslant \widehat{V}_2^t(\bar{s}_2^t, \bar{a}_1^t) - V_2^{\boldsymbol{\pi}_t}(\bar{s}_2^t, \bar{a}_1^t) + \zeta_2^t + b_1^t \\
&\leqslant \big(\widehat{V}_3^t(\bar{s}_3^t, \bar{a}_2^t) - V_3^{\boldsymbol{\pi}_t}(\bar{s}_3^t, \bar{a}_2^t) + \zeta_3^t + b_2^t\big) + \zeta_2^t + b_1^t \\
&\vdots
\end{aligned}$$

$$\leqslant \widehat{V}_{K+1}^t(\bar{s}_{K+1}^t, \bar{a}_K^t) - V_{K+1}^{\boldsymbol{\pi}_t}(\bar{s}_{K+1}^t, \bar{a}_K^t) + \sum_{k=1}^{K} \left(\zeta_{k+1}^t + b_k^t\right)$$

$$= \sum_{k=1}^{K} \zeta_{k+1}^t + \sum_{k=1}^{K} b_k^t \tag{114}$$

$$= \sum_{k=2}^{K} \zeta_k^t + \sum_{k=1}^{K} b_k^t. \tag{115}$$

Equation 114 follows by $\widehat{V}_{K+1}^t(\bar{s}_{K+1}^t, \bar{a}_K^t) - V_{K+1}^{\boldsymbol{\pi}_t}(\bar{s}_{K+1}^t, \bar{a}_K^t) = 0$ and Equation 115 is by reindexing.

Summing over episodes,

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} \left[V_1^*(s_1^t) - V_1^{\boldsymbol{\pi}_t}(s_1^t)\right] \\ &\leqslant \sum_{t=1}^{T}\sum_{k=2}^{K} \zeta_k^t + \sum_{t=1}^{T}\sum_{k=1}^{K} b_k^t. \end{aligned} \tag{116}$$

Equation 116 follows by Equation 113 and Equation 115.

(a). Bounding $\sum_{t=1}^{T}\sum_{k=2}^{K} \zeta_k^t$. By Equation 90, we have,

$$\zeta_k^t = \mathbb{E}_{\bar{a}_{k-1}^t}[\widehat{V}_k^t(\bar{s}_{k-1}^t, S_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_{k-1}^t, S_k^t, \bar{a}_{k-1}^t)|\bar{s}_{k-1}^t] - [\widehat{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t)].$$

In Step 2, we write the filtration as $\mathcal{F}_{k,t-1} := \sigma\Big(\{(\bar{s}_K^i, \bar{a}_K^i, Y^i)\}_{i<t} \bigcup \{(\bar{s}_k^t, \bar{a}_k^t)\}\Big)$. It is the σ-algebra generated by the past episodes together with the current episode's trajectory through stage k (state s_k^t observed and action a_k^t taken).

Then conditioned on $\mathcal{F}_{k,t-1}$, the functions $\widehat{V}_{k+1}^t, V_{k+1}^{\boldsymbol{\pi}_t}$ and the point $(\bar{s}_k^t, \bar{a}_k^t)$ are fixed, and the only remaining randomness is $S_{k+1}^t \sim P_{\bar{a}_k^t}(\cdot \mid \bar{s}_k^t)$. Hence

$$\begin{aligned} \mathbb{E}[\zeta_{k+1}^t \mid \mathcal{F}_{k,t-1}] = \mathbb{E}_{\bar{a}_k^t}[(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}, \bar{a}_k^t)|\bar{s}_k^t] \\ - \mathbb{E}[(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t)|\mathcal{F}_{k,t-1}] = 0. \end{aligned}$$

Therefore $\{\zeta_{k+1}^t\}_{(t,k):\, t\in[T],\, k\in[K-1]}$ is a martingale-difference sequence adapted to $\{\mathcal{F}_{k,t-1}\}$ (with $\zeta_{K+1}^t = 0$).

Since $\widehat{V}_{k+1}^t, V_{k+1}^{\boldsymbol{\pi}_t} \in [0,1]$, we have $|\widehat{V}_{k+1}^t(\bar{s}_{k+1}^t, \bar{a}_k^t) - V_{k+1}^{\boldsymbol{\pi}_t}(\bar{s}_{k+1}^t, \bar{a}_k^t)| \leqslant 1$ and hence $|\zeta_{k+1}^t| \leqslant 2$. The sum $\sum_{t=1}^{T}\sum_{k=2}^{K} \zeta_k^t$ has $T(K-1)$ summands, so by Azuma-Hoeffding inequality (Lemma 12), with probability at least $1-\delta_1$,

$$\sum_{t=1}^{T}\sum_{k=2}^{K} \zeta_k^t \;\leqslant\; 2\sqrt{2T(K-1)\log(2/\delta_1)} \tag{117}$$

(b). Bound $\sum_{t=1}^{T}\sum_{k=1}^{K} b_k^t$. Firstly, denoting the worst-case offline width (Definition 2)

$$W_k^{\text{off}} := \sup_{(\bar{s}_k, \bar{a}_k)} \big(U_k(\bar{s}_k, \bar{a}_k) - L_k(\bar{s}_k, \bar{a}_k)\big), \tag{118}$$

Then, we have,

$$\sum_{t=1}^{T} b_k^t = \sum_{t=1}^{T} \min\left\{(\mathcal{C}+\beta_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}},\; U_k - L_k\right\} \tag{119}$$

$$\leqslant \min\left\{(\beta_k + \mathcal{C})\sum_{t=1}^{T} \|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}},\; W_k^{\text{off}} T\right\} \tag{120}$$

$$\leqslant \min\Big\{(\beta_k+\mathcal{C})\,2\sqrt{T\cdot d_k\log\big(\lambda+\frac{T}{d_k}\big)},\ W_k^{\text{off}}T\Big\} \tag{121}$$

$$\leqslant \min\Big\{2C_1 d_K\sqrt{T\cdot d_k\,\chi\,\log\big(\lambda+\frac{T}{d_k}\big)},\ W_k^{\text{off}}T\Big\} \tag{122}$$

$$\leqslant \min\Big\{2C_1 d_K^{3/2}\sqrt{T\cdot \chi\,\log\big(\lambda+\frac{T}{d_k}\big)},\ W_k^{\text{off}}T\Big\}. \tag{123}$$

Equation 119 is from Equation 111; Equation 120 follows by Equation 118; Equation 121 holds since

$$\sum_{t=1}^{T}\|\psi_k(\bar{s}_k^t,\bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}} \leqslant \sqrt{T\cdot\sum_{t=1}^{T}\|\psi_k(\bar{s}_k^t,\bar{a}_k^t)\|^2_{(\Lambda_k^t)^{-1}}} \leqslant 2\sqrt{T\cdot d_k\log\big(\lambda+\frac{T}{d_k}\big)}$$

from the elliptical potential bound (Lemma 13) and the Cauchy-Schwarz inequality; Equation 122 is because

$$\begin{aligned}\mathcal{C} &= \rho+\sqrt{\lambda d_K} = Cd_K\sqrt{\chi}+\sqrt{\lambda d_K},\\ \beta_k &= Cd_K\sqrt{\chi}+\sqrt{\lambda d_k},\end{aligned}$$

where $\mathcal{C}$ is the constant of Proposition 5 and $\chi = \log\big(\frac{4(c_\beta+1)d_KKT}{\delta}\big)$; Equation 123 holds since $d_k\leqslant d_K$.

Combining Equations (116), (117) and (123), on the good event $\mathcal{E}$, with probability at least $1-\delta-2\delta_1$

$$\begin{aligned}R(T) &\leqslant 2\sqrt{2TK\log(2/\delta_1)}+\sum_{k=1}^{K}\min\Big\{2C_1d_K^{3/2}\sqrt{T\,\chi\,\log\big(\lambda+\tfrac{T}{d_k}\big)},\ T\,W_k^{\text{off}}\Big\}\\ &\leqslant 2\sqrt{2TK\log(2/\delta_1)}+\min\Big\{2C_1d_K^{3/2}K\sqrt{T\,\chi\,\log(\lambda+T)},\ T\sum_{k=1}^{K}W_k^{\text{off}}\Big\}\\ &= \widetilde{\mathcal{O}}\Big(\sqrt{KT}+\min\Big\{\sqrt{d_K^3K^2T},\ T\sum_{k=1}^{K}W_k^{\text{off}}\Big\}\Big),\end{aligned} \tag{124}$$

which finishes the proof. ■

Proposition 5. *Recall the good event $\mathcal{E}_1$ (Equation 87) as*

$$\begin{aligned}\mathcal{E}_1 := \Big\{\forall k<K,\ \forall t\in[T],\ \forall V\in\widetilde{\mathcal{V}}_{k+1}:\ \Big\|\sum_{\tau=1}^{t-1}\psi_k^\tau\Big(V(\bar{s}_{k+1}^\tau,\bar{a}_k^\tau)-\mathbb{E}_{\bar{a}_k^\tau}\big[V(\bar{S}_{k+1},\bar{a}_k^\tau)\mid\bar{s}_k^\tau\big]\Big)\Big\|_{(\Lambda_k^t)^{-1}}\leqslant\rho\Big\}\\ \bigcap\Big\{\forall t\in[T]:\Big\|\sum_{\tau=1}^{t-1}\psi_K^\tau\big[Y^\tau-\mathbb{E}\big[Y^\tau\mid\mathcal{F}_{K,\tau-1}\big]\big]\Big\|_{(\Lambda_K^t)^{-1}}\leqslant\rho\Big\}.\end{aligned} \tag{125}$$

$\rho = Cd_K\sqrt{\log\big(\frac{4(c_\beta+1)d_KKT}{\delta}\big)}$ and $\mathcal{C}:=\rho+\sqrt{\lambda d_K}$. Then, on $\mathcal{E}_1$, for every policy $\boldsymbol{\pi}$, every k, every t, and every $(\bar{s}_k,\bar{a}_k)$,

$$\big\langle\psi_k(\bar{s}_k,\bar{a}_k),\hat{\theta}_k^t\big\rangle - Q_k^{\boldsymbol{\pi}}(\bar{s}_k,\bar{a}_k)\leqslant\mathbb{E}_{\bar{a}_k}\Big[(\widehat{V}_{k+1}^t-V_{k+1}^{\boldsymbol{\pi}})(\bar{s}_k,s_{k+1},\bar{a}_k)\mid\bar{s}_k\Big]+\mathcal{C}\|\psi_k(\bar{s}_k,\bar{a}_k)\|_{(\Lambda_k^t)^{-1}}.$$

The conditional expectation is over $s_{k+1}\sim P_{\bar{a}_k}(\cdot\mid\bar{s}_k)$.

Proof. Fix $k<K$, $t\in[T]$, a policy $\boldsymbol{\pi}$ and a history $(\bar{s}_k,\bar{a}_k)$.

By Equation 11, $\widehat{V}_{k+1}^t=\max_{a_{k+1}}\widehat{Q}_{k+1}^t$ belongs to $\widetilde{\mathcal{V}}_{k+1}\subseteq\mathcal{V}_{k+1}$ (Step 2 of the proof of Theorem 1), so Assumption 1(ii) provides $\theta_{\widehat{V}_{k+1}^t}\in\mathbb{R}^{d_k}$ with $\|\theta_{\widehat{V}_{k+1}^t}\|_2\leqslant\sqrt{d_k}$ such that, for every $(\bar{s}_k,\bar{a}_k)$,

$$\mathbb{E}_{\bar{a}_k}\big[\widehat{V}_{k+1}^t\big((\bar{s}_k,S_{k+1}),\bar{a}_k\big)\mid\bar{s}_k\big] = \big\langle\psi_k(\bar{s}_k,\bar{a}_k),\,\theta_{\widehat{V}_{k+1}^t}\big\rangle. \tag{126}$$

Then, we have,

$$\langle \psi_k, \widehat{\theta}_k^t \rangle - Q_k^{\boldsymbol{\pi}}(\bar{s}_k, \bar{a}_k) \;=\; \underbrace{\langle \psi_k,\; \widehat{\theta}_k^t - \theta_{\widehat{V}_{k+1}^t} \rangle}_{(a)} \;+\; \underbrace{\langle \psi_k,\; \theta_{\widehat{V}_{k+1}^t} \rangle - Q_k^{\boldsymbol{\pi}}(\bar{s}_k, \bar{a}_k)}_{(b)}. \tag{127}$$

For term (a), we have,

$$\begin{aligned}
&\langle \psi_k,\; \widehat{\theta}_k^t - \theta_{\widehat{V}_{k+1}^t} \rangle \\
&\leqslant \; \left|\langle \psi_k,\; \widehat{\theta}_k^t - \theta_{\widehat{V}_{k+1}^t} \rangle\right| \\
&= \; \left|\left\langle \psi_k,\; (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) \;-\; (\Lambda_k^t)^{-1} \Lambda_k^t\, \theta_{\widehat{V}_{k+1}^t} \right\rangle\right| && (128) \\
&= \; \left|\left\langle \psi_k,\; (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) \;-\; (\Lambda_k^t)^{-1} \Big[\sum_{\tau=1}^{t-1} \psi_k^\tau (\psi_k^\tau)^\top + \lambda I \Big] \theta_{\widehat{V}_{k+1}^t} \right\rangle\right| && (129) \\
&= \; \left|\left\langle \psi_k,\; (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \big[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - (\psi_k^\tau)^\top \theta_{\widehat{V}_{k+1}^t} \big] \;-\; \lambda (\Lambda_k^t)^{-1} \theta_{\widehat{V}_{k+1}^t} \right\rangle\right| \\
&= \; \left|\left\langle \psi_k,\; (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \big[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\big[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau \big] \big] \;-\; \lambda (\Lambda_k^t)^{-1} \theta_{\widehat{V}_{k+1}^t} \right\rangle\right| && (130) \\
&\leqslant \; \left|\left\langle \psi_k,\; (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \big[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\big[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau \big] \big] \right\rangle\right| \\
&\quad + \; \lambda \left|\langle \psi_k,\; (\Lambda_k^t)^{-1} \theta_{\widehat{V}_{k+1}^t} \rangle\right| && (131) \\
&\leqslant \; \|\psi_k\|_{(\Lambda_k^t)^{-1}} \Big\| \sum_{\tau=1}^{t-1} \psi_k^\tau \big[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\big[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau \big] \big] \Big\|_{(\Lambda_k^t)^{-1}} \\
&\quad + \; \lambda \, \|\psi_k\|_{(\Lambda_k^t)^{-1}} \, \|\theta_{\widehat{V}_{k+1}^t}\|_{(\Lambda_k^t)^{-1}} && (132) \\
&\leqslant \; \rho \, \|\psi_k\|_{(\Lambda_k^t)^{-1}} \;+\; \sqrt{\lambda}\, \|\theta_{\widehat{V}_{k+1}^t}\|_2 \, \|\psi_k\|_{(\Lambda_k^t)^{-1}} && (133) \\
&\leqslant \; \big(\rho + \sqrt{\lambda d_k}\big) \, \|\psi_k\|_{(\Lambda_k^t)^{-1}}. && (134)
\end{aligned}$$

Equation 128 follows by Equation 10; Equation 129 follows by Equation 30; Equation 130 is because $(\psi_k^\tau)^\top \theta_{\widehat{V}_{k+1}^t} = \mathbb{E}_{\bar{a}_k^\tau}\big[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau\big]$ by Equation 126; Equation 131 follows by the triangle inequality; Equation 132 follows by the Cauchy-Schwarz inequality; Equation 133 follows since $\widehat{V}_{k+1}^t \in \widetilde{\mathcal{V}}_{k+1}$ and the first event in Equation 125 holds uniformly over $\widetilde{\mathcal{V}}_{k+1}$, together with $\lambda \|\theta_{\widehat{V}_{k+1}^t}\|_{(\Lambda_k^t)^{-1}} \leqslant \sqrt{\lambda}\, \|\theta_{\widehat{V}_{k+1}^t}\|_2$; and Equation 134 follows by $\|\theta_{\widehat{V}_{k+1}^t}\|_2 \leqslant \sqrt{d_k}$ in Assumption 1(ii).

For term (b), we have

$$\begin{aligned}
\langle \psi_k,\; \theta_{\widehat{V}_{k+1}^t} \rangle - Q_k^{\boldsymbol{\pi}}(\bar{s}_k, \bar{a}_k) \;&=\; \mathbb{E}_{\bar{a}_k}\big[\widehat{V}_{k+1}^t \mid \bar{s}_k\big] - \mathbb{E}_{\bar{a}_k}\big[V_{k+1}^{\boldsymbol{\pi}} \mid \bar{s}_k\big] && (135) \\
&=\; \mathbb{E}_{\bar{a}_k}\Big[\big(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}}\big)\big(\bar{s}_k, S_{k+1}, \bar{a}_k\big) \;\Big|\; \bar{s}_k \Big]. && (136)
\end{aligned}$$

Equation 135 follows by Equations (34) and (126); and Equation 136 follows by linearity of expectation.

Combining eqs. (127), (134) and (136) with $\mathcal{C} = \rho + \sqrt{\lambda d_K}$ gives the claim.

For $k = K$, Assumption 1(i) gives $Q_K^{\boldsymbol{\pi}} = \langle \psi_K, w_K \rangle$, so term (b) goes away, and the bound on term (a) follows by the same argument with Y^τ in place of $\widehat{V}_{k+1}^t$, w_K in place of $\theta_{\widehat{V}_{k+1}^t}$, and the second event in Equation 125 in place of the first. ■

Lemma 2. *Suppose the good event $\mathcal{E}$ in Equation 88 holds. Let $\beta_k = \rho + \sqrt{\lambda d_k}$ for all $k \in [K]$, $\rho = C d_K \sqrt{\log\left(\frac{4(c_\beta+1) d_K K T}{\delta}\right)}$. Then, for every $t \in [T]$, we have:*

(a) *at the terminal stage* $k = K$, $\theta_K^* \in \mathcal{B}_K^t$;

(b) *at every non-terminal stage* $k < K$, $\theta_{\widehat{V}_{k+1}^t} \in \mathcal{B}_k^t$ *where* $\theta_{\widehat{V}_{k+1}^t}$ *is the vector supplied by Assumption 1(ii) for* $\widehat{V}_{k+1}^t \in \mathcal{V}_{k+1}$.

Proof. This lemma shows that the constructed online confidence set contains the reward parameter at stage K and the linear representation of the optimistic value at stage $k \in [K-1]$.

For any $k \in [K]$ and $t \in [T]$, we prove the two claims separately.

(a) Terminal stage. At the terminal stage, the regression target is Y^t. By Assumption 1, there exists $\theta_K^* = w_K \in \mathbb{R}^{d_K}$ such that

$$\mathbb{E}_{\bar{a}_K^t}\left[Y^t \mid \bar{s}_K^t\right] = \left\langle \psi_K^t(\bar{s}_K^t, \bar{a}_K^t),\ \theta_K^* \right\rangle \tag{137}$$

The terminal ridge estimator is $\widehat{\theta}_K^t = (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K^\tau Y^\tau$. Then, we have,

$$\begin{aligned}
\widehat{\theta}_K^t - \theta_K^* &= (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K^\tau Y^\tau - \theta_K^* \\
&= (\Lambda_K^t)^{-1} \sum_{\tau=1}^{t-1} \psi_K^\tau \left[Y^\tau - (\psi_K^\tau)^\top \theta_K^*\right] - \lambda (\Lambda_K^t)^{-1} \theta_K^*. \qquad (138)
\end{aligned}$$

Equation 138 is by using $\theta_K^* = (\Lambda_K^t)^{-1}\Lambda_K^t \theta_K^*$ and Equation 30. Taking the Λ_K^t-norm gives

$$\begin{aligned}
\|\widehat{\theta}_K^t - \theta_K^*\|_{\Lambda_K^t} &\leqslant \left\| \sum_{\tau=1}^{t-1} \psi_K^\tau \left[Y^\tau - (\psi_K^\tau)^\top \theta_K^*\right] \right\|_{(\Lambda_K^t)^{-1}} + \lambda \|(\Lambda_K^t)^{-1} \theta_K^*\|_{\Lambda_K^t} \\
&= \left\| \sum_{\tau=1}^{t-1} \psi_K^\tau \left[Y^\tau - \mathbb{E}_{\bar{a}_K^\tau}\left[Y^\tau \mid \bar{s}_K^\tau\right]\right] \right\|_{(\Lambda_K^t)^{-1}} + \lambda \|(\Lambda_K^t)^{-1} \theta_K^*\|_{\Lambda_K^t} \qquad (139) \\
&\leqslant \rho + \sqrt{\lambda} \|\theta_K^*\|_2. \qquad (140) \\
&\leqslant \rho + \sqrt{\lambda d_K}. \qquad (141)
\end{aligned}$$

Equation 139 follows by Equation 137; Equation 140 follows from the second event in Equation 125 for the first term, and from $\lambda \|(\Lambda_K^t)^{-1}\theta_K^*\|_{\Lambda_K^t} = \lambda \|\theta_K^*\|_{(\Lambda_K^t)^{-1}} \leqslant \sqrt{\lambda}\|\theta_K^*\|_2$ for the second term, since $\Lambda_K^t \succeq \lambda I$; and Equation 141 follows from $\|\theta_K^*\|_2 \leqslant \sqrt{d_K}$ in Proposition 4. Therefore,

$$\theta_K^* \in \mathcal{B}_K^t.$$

(b) Non-terminal stages. Fix any $k < K$. By Assumption 1(ii) applied to $\widehat{V}_{k+1}^t \in \mathcal{V}_{k+1}$, we have

$$\mathbb{E}_{\bar{a}_k^\tau}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau\right] = \left\langle \psi_k^\tau, \theta_{\widehat{V}_{k+1}^t} \right\rangle. \tag{142}$$

Therefore,

$$\begin{aligned}
\widehat{\theta}_k^t - \theta_{\widehat{V}_{k+1}^t} &= (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \left[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - (\psi_k^\tau)^\top \theta_{\widehat{V}_{k+1}^t}\right] - \lambda (\Lambda_k^t)^{-1} \theta_{\widehat{V}_{k+1}^t} \qquad (143) \\
&= (\Lambda_k^t)^{-1} \sum_{\tau=1}^{t-1} \psi_k^\tau \left[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau\right]\right] - \lambda (\Lambda_k^t)^{-1} \theta_{\widehat{V}_{k+1}^t}. \\
&\qquad (144)
\end{aligned}$$

Equation 143 follows from Equation 10, together with $\theta_{\widehat{V}_{k+1}^t} = (\Lambda_k^t)^{-1}\Lambda_k^t \theta_{\widehat{V}_{k+1}^t}$, and Equation 30; and Equation 144 comes from Equation 142.

Taking the Λ_k^t-norm, we obtain,

$$\|\widehat{\theta}_k^t - \theta_{\widehat{V}_{k+1}^t}\|_{\Lambda_k^t} \leqslant \left\| \sum_{\tau=1}^{t-1} \psi_k^\tau \left[\widehat{V}_{k+1}^t(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau\right]\right] \right\|_{(\Lambda_k^t)^{-1}}$$

$$+ \sqrt{\lambda}\,\|\theta_{\widehat{V}^t_{k+1}}\|_2. \tag{145}$$

Equation 145 follows by the triangle inequality together with $\lambda\|(\Lambda^t_k)^{-1}\theta_{\widehat{V}^t_{k+1}}\|_{\Lambda^t_k} \leqslant \sqrt{\lambda}\|\theta_{\widehat{V}^t_{k+1}}\|_2$.

Since $\widehat{V}^t_{k+1} \in \widetilde{\mathcal{V}}_{k+1}$ by Equation 11 and the first event in Equation 125 holds uniformly over $\widetilde{\mathcal{V}}_{k+1}$, we have.

$$\left\|\sum_{\tau=1}^{t-1} \psi^\tau_k \left[\widehat{V}^t_{k+1}(\bar{s}^\tau_{k+1}, \bar{a}^\tau_k) - \mathbb{E}_{\bar{a}^\tau_k}\left[\widehat{V}^t_{k+1}(\bar{S}_{k+1}, \bar{a}^\tau_k) \mid \bar{s}^\tau_k\right]\right]\right\|_{(\Lambda^t_k)^{-1}} \leqslant \rho. \tag{146}$$

For the second term, Assumption 1(ii) gives $\|\theta_{\widehat{V}^t_{k+1}}\|_2 \leqslant \sqrt{d_k}$.

Consequently, by eqs. (145) and (146), the following holds,

$$\|\widehat{\theta}^t_k - \theta_{\widehat{V}^t_{k+1}}\|_{\Lambda^t_k} \leqslant \rho + \sqrt{\lambda d_k} \leqslant \beta_k.$$

Therefore, $\theta_{\widehat{V}^t_{k+1}} \in \mathcal{B}^t_k$. ■

Lemma 3. *On the good event $\mathcal{E}$ in Equation 88, fix $k < K$ and $t \in [T]$. Suppose the inductive optimism hypothesis holds:*

$$\widehat{Q}^t_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \geqslant Q^*_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}), \quad \forall(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}). \tag{147}$$

Let $\theta_{\widehat{V}^t_{k+1}}$ be the vector supplied by Assumption 1(ii) for $\widehat{V}^t_{k+1} \in \mathcal{V}_{k+1}$. Then, we have $\theta_{\widehat{V}^t_{k+1}} \in \widehat{\mathcal{C}}^{\text{off}}_k$.

Proof. By Assumption 1(ii), we have $\left\langle \psi_k(\bar{s}_k, \bar{a}_k), \theta_{\widehat{V}^t_{k+1}} \right\rangle = \mathbb{E}_{\bar{a}_k}[\widehat{V}^t_{k+1}(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k]$. Therefore, to show $\theta_{\widehat{V}^t_{k+1}} \in \widehat{\mathcal{C}}^{\text{off}}_k$, it is enough to prove that, for every $(\bar{s}_k, \bar{a}_k)$,

$$L_k(\bar{s}_k, \bar{a}_k) \leqslant \mathbb{E}_{\bar{a}_k}\left[\widehat{V}^t_{k+1}(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k\right] \leqslant U_k(\bar{s}_k, \bar{a}_k).$$

(a) Lower bound. By the induction hypothesis (Equation 147), we have,

$$\widehat{V}^t_{k+1}(\bar{s}_{k+1}, \bar{a}_k) = \max_{a_{k+1}} \widehat{Q}^t_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \geqslant \max_{a_{k+1}} Q^*_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) = V^*_{k+1}(\bar{s}_{k+1}, \bar{a}_k). \tag{148}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}_{\bar{a}_k}\left[\widehat{V}^t_{k+1}(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k\right] &\geqslant \mathbb{E}_{\bar{a}_k}\left[V^*_{k+1}(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k\right] \\
&= \sum_{s_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)\, V^*_{k+1}((\bar{s}_k, s_{k+1}), \bar{a}_k) && (149) \\
&\geqslant \sum_{s_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \max_{a_{k+1}} L_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) && (150) \\
&\geqslant L_k(\bar{s}_k, \bar{a}_k). && (151)
\end{aligned}$$

Equation 150 follows from $Q^*_{k+1} \geqslant L_{k+1}$ in Equation 77 on the good event $\mathcal{E}$; and Equation 151 follows from Equation 71, since $P_{\bar{a}_k}(\cdot \mid \bar{s}_k) \in \widehat{\mathcal{P}}_k(\bar{s}_k, \bar{a}_k)$, together with Equation 68.

(b) Upper bound. By Equation 11, we have,

$$\widehat{Q}^t_{k+1}(\bar{s}_{k+1}, \bar{a}_{k+1}) = \max\left\{\min\left\{\max_{\theta \in \widehat{\mathcal{C}}^t_{k+1}} \theta^\top \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_{k+1}),\ \max_{\theta \in \widehat{\mathcal{C}}^{\text{off}}_{k+1}} \theta^\top \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_{k+1}),\ 1\right\},\ 0\right\}. \tag{152}$$

On the event $\mathcal{E}$ we have $\theta^*_{k+1} \in \widehat{\mathcal{C}}^{\text{off}}_{k+1}$ by Proposition 2, hence

$$\max_{\theta \in \widehat{\mathcal{C}}^{\text{off}}_{k+1}} \theta^\top \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \ \geqslant\ Q^*_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \ \geqslant\ 0. \tag{153}$$

Thus,

$$\mathbb{E}_{\bar{a}_k}\left[\widehat{V}^t_{k+1}(\bar{S}_{k+1},\bar{a}_k)\mid \bar{s}_k\right] = \mathbb{E}_{\bar{a}_k}\left[\max_{a_{k+1}}\widehat{Q}^t_{k+1}(\bar{s}_{k+1},\bar{a}_k,a_{k+1})\mid \bar{s}_k\right] \tag{154}$$

$$\leqslant \mathbb{E}_{\bar{a}_k}\left[\max_{a_{k+1}}\max_{\theta\in\widehat{\mathcal{C}}^{\text{off}}_{k+1}}\theta^\top\psi_{k+1}(\bar{s}_{k+1},\bar{a}_k,a_{k+1})\mid \bar{s}_k\right] \tag{155}$$

$$\leqslant \mathbb{E}_{\bar{a}_k}\left[\max_{a_{k+1}}U_{k+1}(\bar{s}_k,S_{k+1},\bar{a}_k,a_{k+1})\mid \bar{s}_k\right] \tag{156}$$

$$= \sum_{s_{k+1}}P_{\bar{a}_k}(s_{k+1}\mid \bar{s}_k)\left(\max_{a_{k+1}}U_{k+1}(\bar{s}_k,s_{k+1},\bar{a}_k,a_{k+1})\right) \tag{157}$$

$$\leqslant \max_{P\in\widehat{\mathcal{P}}_k(\bar{s}_k,\bar{a}_k)}\sum_{s_{k+1}}P(s_{k+1})\left(\max_{a_{k+1}}U_{k+1}(\bar{s}_k,s_{k+1},\bar{a}_k,a_{k+1})\right) \tag{158}$$

$$= U_k(\bar{s}_k,\bar{a}_k). \tag{159}$$

Equation 154 follows by the definition $\widehat{V}^t_{k+1} = \max_{a_{k+1}}\widehat{Q}^t_{k+1}$; Equation 155 follows by Equation 152 together with Equation 153, which ensures the truncation at 0 preserves the inequality; Equation 156 follows by the definition of $\widehat{\mathcal{C}}^{\text{off}}_{k+1}$, since every $\theta\in\widehat{\mathcal{C}}^{\text{off}}_{k+1}$ satisfies $\psi^\top_{k+1}\theta\leqslant U_{k+1}$; Equation 157 is the definition of conditional expectation; Equation 158 follows by Equation 71; and Equation 159 holds by Equation 69.

Combining Equations (151) and (159), we obtain $L_k(\bar{s}_k,\bar{a}_k)\leqslant\psi_k(\bar{s}_k,\bar{a}_k)^\top\theta_{\widehat{V}^t_{k+1}}\leqslant U_k(\bar{s}_k,\bar{a}_k)$, which finishes the proof. ■

Lemma 4. *On the good event $\mathcal{E}$ in Equation 88, for every $k\in[K]$, $t\in[T]$, and $(\bar{s}_k,\bar{a}_k)$, we have*

$$\widehat{Q}^t_k(\bar{s}_k,\bar{a}_k)\geqslant Q^*_k(\bar{s}_k,\bar{a}_k).$$

Proof. We prove this lemma by backward induction. Firstly since $Y\in[0,1]$, backward induction gives $V^*_{k+1}\in[0,1]$ and hence

$$0\leqslant Q^*_k(\bar{s}_k,\bar{a}_k)\leqslant 1,\qquad \forall k\in[K],\ \forall(\bar{s}_k,\bar{a}_k). \tag{160}$$

(a) Base case: $k=K$. By Lemma 2 (a), we have, $\theta^*_K\in\mathcal{B}^t_K$, also, by $\mathcal{E}^{\text{off}}$ (Equation 60), we have $\theta^*_K\in\widehat{\mathcal{C}}^{\text{off}}_K$. Therefore θ^*_K is in both online and offline confidence set.

By Equation 11, we have the following estimated value function,

$$\widehat{Q}^t_K(\bar{s}_K,\bar{a}_K)\leftarrow\max\Big\{\min\big\{\max_{\theta\in\widehat{\mathcal{C}}^t_K}\theta^\top\psi_K(\bar{s}_K,\bar{a}_K),\ \max_{\theta\in\widehat{\mathcal{C}}^{\text{off}}_K}\theta^\top\psi_K(\bar{s}_K,\bar{a}_K),\ 1\big\},\ 0\Big\}.$$

For the notation $\mathcal{B}^t_K:=\widehat{\mathcal{C}}^t_K$. Hence,

$$\max_{\theta\in\widehat{\mathcal{C}}^t_K}\theta^\top\psi_K(\bar{s}_K,\bar{a}_K)\geqslant\left\langle\psi_K(\bar{s}_K,\bar{a}_K),\ \theta^*_K\right\rangle = Q^*_K, \tag{161}$$

$$\max_{\theta\in\widehat{\mathcal{C}}^{\text{off}}_K}\theta^\top\psi_K(\bar{s}_K,\bar{a}_K)\geqslant\left\langle\psi_K(\bar{s}_K,\bar{a}_K),\ \theta^*_K\right\rangle = Q^*_K. \tag{162}$$

Therefore, by Equations (160) to (162), we have $\widehat{Q}^t_K(\bar{s}_K,\bar{a}_K)\geqslant Q^*_K(\bar{s}_K,\bar{a}_K)$.

(b) Induction step: $k<K$. Assume the claim holds at stage $k+1$, that is, for any $t\in[T]$, and $(\bar{s}_{k+1},\bar{a}_{k+1})$, the following holds

$$\widehat{Q}^t_{k+1}(\bar{s}_{k+1},\bar{a}_{k+1})\geqslant Q^*_{k+1}(\bar{s}_{k+1},\bar{a}_{k+1}). \tag{163}$$

Then

$$\widehat{V}^t_{k+1}(\bar{s}_{k+1},\bar{a}_k)=\max_{a_{k+1}}\widehat{Q}^t_{k+1}(\bar{s}_{k+1},\bar{a}_{k+1})\geqslant\max_{a_{k+1}}Q^*_{k+1}(\bar{s}_{k+1},\bar{a}_{k+1})=V^*_{k+1}(\bar{s}_{k+1},\bar{a}_k). \tag{164}$$

Therefore,

$$\mathbb{E}_{\bar{a}_k}\left[\widehat{V}^t_{k+1}(\bar{S}_{k+1},\bar{a}_k)\mid\bar{s}_k\right]\geqslant\mathbb{E}_{\bar{a}_k}\left[V^*_{k+1}(\bar{S}_{k+1},\bar{a}_k)\mid\bar{s}_k\right]$$

$$= Q_k^*(\bar{s}_k, \bar{a}_k). \tag{165}$$

By Lemma 2 (b), we have $\theta_{\widehat{V}_{k+1}^t} \in \mathcal{B}_k^t$, and by Lemma 3, using the induction hypothesis, we also have $\theta_{\widehat{V}_{k+1}^t} \in \widehat{\mathcal{C}}_k^{\text{off}}$.

By Equation 11, we have the following estimated value function,

$$\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \leftarrow \max\Big\{ \min\big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), \max_{\theta \in \widehat{\mathcal{C}}_k^{\text{off}}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), 1\big\}, 0\Big\}.$$

Hence,

$$\max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k) \geqslant \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_{\widehat{V}_{k+1}^t} \big\rangle, \quad \max_{\theta \in \widehat{\mathcal{C}}_k^{\text{off}}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k) \geqslant \big\langle \psi_k(\bar{s}_k, \bar{a}_k),\ \theta_{\widehat{V}_{k+1}^t} \big\rangle.$$

By Assumption 1(ii), we have $\mathbb{E}_{\bar{a}_k}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k\right] = \big\langle \psi_k(\bar{s}_k, \bar{a}_k), \theta_{\widehat{V}_{k+1}^t} \big\rangle$. Then,

$$\max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k) \geqslant \mathbb{E}_{\bar{a}_k}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k\right] \geqslant Q_k^*, \tag{166}$$

$$\max_{\theta \in \widehat{\mathcal{C}}_k^{\text{off}}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k) \geqslant \mathbb{E}_{\bar{a}_k}\left[\widehat{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k\right] \geqslant Q_k^*. \tag{167}$$

Therefore, by Equations (160), (166) and (167), we have $\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \geqslant Q_k^*(\bar{s}_k, \bar{a}_k)$. ■

Corollary 1 (Purely online setting)**.** *Under Assumptions 1, let χ, C_1, C, c_β and β_k be as in Theorem 1. Then for any $\delta, \delta_1 \in (0, 1]$, with probability at least $1 - \delta - \delta_1$, Algorithm 1 without offline data achieves (see Algorithm 3)*

$$R(T) \leqslant 2\sqrt{2TK \log(2/\delta_1)} + 2C_1\, d_K^{3/2}\, K\sqrt{T \cdot \chi \log(1+T)}.$$

Proof. The proof follows that of Theorem 1 (Appendix D.2), with a single modification in Step 2. Since the value iterate no longer contains the offline upper confidence value, the class $\widetilde{\mathcal{V}}_{k+1}$ is replaced by

$$\widetilde{\mathcal{V}}_{k+1}^{\text{on}} := \Big\{ V :\ V(\bar{s}_{k+1}, \bar{a}_k) = \max\Big\{ \max_{a_{k+1} \in \mathcal{A}_{k+1}} \min\big\{1,\ \langle \omega, \psi_{k+1} \rangle + \beta \|\psi_{k+1}\|_{\Lambda^{-1}}\big\},\ 0 \Big\} \Big\},$$

with the same parameter ranges. $\widetilde{\mathcal{V}}_{k+1}^{\text{on}} \subseteq \mathcal{V}_{k+1}$, taking $\theta = \theta' = \omega$, $A = A' = \beta^2 \Lambda^{-1}$ and $\mathcal{P} = \mathbb{R}^{d_{k+1}} \in \mathfrak{P}_{k+1}$ (i.e., $L \equiv -\infty$, $U \equiv +\infty$) in Equation 3, so Assumption 1(ii) applies on $\widetilde{\mathcal{V}}_{k+1}^{\text{on}}$. The covering bound Equation 85 is unchanged, since the cover is taken over (ω, β, Λ) only.

The remainder of the proof is then read with $\widetilde{\mathcal{V}}_{k+1}^{\text{on}}$ in place of $\widetilde{\mathcal{V}}_{k+1}$ throughout the event $\mathcal{E}_1$ in Equation 87, Lemma 2 and Lemma 4, and with the inner minimum over the two confidence sets in Equation 11 replaced by the online branch alone.

For the failure probability, the offline confidence set no longer depends on the offline sample, so the event $\mathcal{E}^{\text{off}}$ of Equation 60 is not needed and the good event Equation 88 reduces to $\mathcal{E} = \mathcal{E}_1$, with $P(\mathcal{E}) \geqslant 1 - \delta$. ■

E Proof for Section 4

E.1 Proof for Offline Confidence Set

Proposition 6. *Under Assumptions 1 and 2, for any history $(\bar{s}_K, \bar{a}_K)$ with $P(\bar{s}_K, \bar{a}_K) > 0$, write*

$$\alpha_i^- = \frac{\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{\Gamma - (\Gamma - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}; \alpha_i^+ = \frac{\Gamma \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{1 + (\Gamma - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}; B_k^- = \prod_{i=1}^k \alpha_i^-; B_k^+ = \prod_{i=1}^k \alpha_i^+.$$

with $B_0^- = B_0^+ = 1$. Define

$$\zeta^\Gamma(\bar{s}_{k+1}, \bar{a}_k) = \min\Big\{ \frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-},\ P(\bar{s}_{k+1}, \bar{a}_k) - P(\bar{s}_k, \bar{a}_k) + \zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}) \Big\}, \tag{168}$$

with $\zeta^{\Gamma}(s_1) = P(s_1)$. *Then for every* $k \in [K-1]$,

$$\frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^+ \, \zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1})} \leqslant P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \leqslant \min\Big\{1 - \frac{P(\bar{s}_k, \bar{a}_k) - P(\bar{s}_{k+1}, \bar{a}_k)}{\zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1})}, \; \frac{P(\bar{s}_{k+1}, \bar{a}_k)\, B_{k-1}^+}{B_k^- \, P(\bar{s}_k, \bar{a}_{k-1})}\Big\},$$

and at the terminal stage,

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \geqslant \max\Big\{\frac{\mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K] P(\bar{s}_K, \bar{a}_K)}{B_K^+ \zeta^{\Gamma}(\bar{s}_K, \bar{a}_{K-1})}, \; 1 - \frac{(1 - \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K])\pi_{b,K}(a_K \mid \bar{s}_K, \bar{a}_{K-1}) B_{K-1}^+}{B_K^-}\Big\}$$

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \leqslant \min\Big\{1 - \frac{(1 - \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K]) P(\bar{s}_K, \bar{a}_K)}{B_K^+ \zeta^{\Gamma}(\bar{s}_K, \bar{a}_{K-1})}, \; \frac{\mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K] \pi_{b,K}(a_K \mid \bar{s}_K, \bar{a}_{K-1}) B_{K-1}^+}{B_K^-}\Big\}.$$

Proof. This proposition bounds two interventional quantities: the transition kernel $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$, $\forall k \in [K-1]$ and the terminal reward $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$. It uses only the observational distribution $P(\bar{s}_K, \bar{a}_K, y)$ and the restriction that Assumption 2 places on treatment assignment. The difficulty is that Assumption 2 restricts the propensity $\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u})$, while the interventional quantities $P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k)$ and $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$ are given by the truncated factorization, which involve only transition kernels.

Step 1. Relations between the interventional and observational distribution. Fix one history $(\bar{s}_K, \bar{a}_K)$ with $P(\bar{s}_K, \bar{a}_K) > 0$. By the truncated factorization Pearl (2009), the joint density of $(\boldsymbol{u}, \bar{s}_k)$ under $do(\bar{a}_{k-1})$ is

$$P_{\bar{a}_{k-1}}(\boldsymbol{u}, \bar{s}_k) = P(\boldsymbol{u})\, P(s_1 \mid \boldsymbol{u}) \prod_{i=1}^{k-1} P\big(s_{i+1} \mid \bar{s}_i, \bar{a}_i, \boldsymbol{u}\big), \tag{169}$$

and consequently

$$P_{\bar{a}_k}(\boldsymbol{u}, \bar{s}_{k+1}) = P_{\bar{a}_{k-1}}(\boldsymbol{u}, \bar{s}_k)\, P\big(s_{k+1} \mid \bar{s}_k, \bar{a}_k, \boldsymbol{u}\big) \leqslant P_{\bar{a}_{k-1}}(\boldsymbol{u}, \bar{s}_k). \tag{170}$$

Marginalizing eqs. (169) and (170) gives,

$$P_{\bar{a}_{k-1}}(\bar{s}_k) = \sum_{\boldsymbol{u}} P_{\bar{a}_{k-1}}(\boldsymbol{u}, \bar{s}_k), \qquad P_{\bar{a}_k}(\bar{s}_{k+1}) = \sum_{\boldsymbol{u}} P_{\bar{a}_k}(\boldsymbol{u}, \bar{s}_{k+1}). \tag{171}$$

By the chain rule of probability and eqs. (169) to (171), we have,

$$P(\bar{s}_k, \bar{a}_k) = \sum_{\boldsymbol{u}} P_{\bar{a}_{k-1}}(\boldsymbol{u}, \bar{s}_k) \prod_{i=1}^{k} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big), \tag{172}$$

$$P(\bar{s}_{k+1}, \bar{a}_k) = \sum_{\boldsymbol{u}} P_{\bar{a}_k}(\boldsymbol{u}, \bar{s}_{k+1}) \prod_{i=1}^{k} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big). \tag{173}$$

Finally, since an intervention on a_k does not affect $\bar{s}_k$, we have

$$P_{\bar{a}_k}\big(s_{k+1} \mid \bar{s}_k\big) = \frac{P_{\bar{a}_k}(\bar{s}_{k+1})}{P_{\bar{a}_k}(\bar{s}_k)} = \frac{P_{\bar{a}_k}(\bar{s}_{k+1})}{P_{\bar{a}_{k-1}}(\bar{s}_k)}. \tag{174}$$

Step 2. Convert the sensitivity model into bounds on the numerator and denominator of Equation 174. By Lemma 5, we have $\alpha_i^- \leqslant \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}) \leqslant \alpha_i^+$ for every $\boldsymbol{u} \in \mathrm{supp}(\boldsymbol{U})$, hence

$$B_k^- \leqslant \prod_{i=1}^{k} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big) \leqslant B_k^+. \tag{175}$$

By eqs. (171), (173) and (175), we have

$$\frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^+} \leqslant P_{\bar{a}_k}(\bar{s}_{k+1}) \leqslant \frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-}. \tag{176}$$

The same argument gives

$$\frac{P(\bar{s}_k, \bar{a}_{k-1})}{B_{k-1}^+} \leqslant P_{\bar{a}_{k-1}}(\bar{s}_k) \leqslant \frac{P(\bar{s}_k, \bar{a}_{k-1})}{B_{k-1}^-}. \tag{177}$$

Step 3. Bound on the numerator of Equation 174. By Equations (171) and (173),

$$P_{\bar{a}_k}(\bar{s}_{k+1}) = P(\bar{s}_{k+1}, \bar{a}_k) + \sum_{\boldsymbol{u}} P_{\bar{a}_k}(\boldsymbol{u}, \bar{s}_{k+1})\Big(1 - \prod_{i=1}^{k} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big)\Big). \tag{178}$$

By Equations (171) and (172),

$$\sum_{\boldsymbol{u}} P_{\bar{a}_{k-1}}(\boldsymbol{u}, \bar{s}_k)\Big(1 - \prod_{i=1}^{k} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big)\Big) = P_{\bar{a}_{k-1}}(\bar{s}_k) - P(\bar{s}_k, \bar{a}_k). \tag{179}$$

Therefore, combining Equations (170), (178) and (179), we have

$$P_{\bar{a}_k}(\bar{s}_{k+1}) \;\leqslant\; P(\bar{s}_{k+1}, \bar{a}_k) - P(\bar{s}_k, \bar{a}_k) + P_{\bar{a}_{k-1}}(\bar{s}_k). \tag{180}$$

Next, we show by induction on k that

$$P_{\bar{a}_{k-1}}(\bar{s}_k) \;\leqslant\; \zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1}). \tag{181}$$

Base case: for $k = 1$ no intervention precedes s_1, so $P(s_1) = \zeta^{\Gamma}(s_1)$ with equality.

Induction step: assume Equation 181 holds at stage k, i.e., $P_{\bar{a}_{k-1}}(\bar{s}_k) \leqslant \zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1})$. Then, we have

$$\begin{aligned}
P_{\bar{a}_k}(\bar{s}_{k+1}) \;&\leqslant\; \min\Big\{\frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-},\; P(\bar{s}_{k+1}, \bar{a}_k) - P(\bar{s}_k, \bar{a}_k) + P_{\bar{a}_{k-1}}(\bar{s}_k)\Big\} && (182)\\
&\leqslant\; \min\Big\{\frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-},\; P(\bar{s}_{k+1}, \bar{a}_k) - P(\bar{s}_k, \bar{a}_k) + \zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1})\Big\} && (183)\\
&=\; \zeta^{\Gamma}(\bar{s}_{k+1}, \bar{a}_k). && (184)
\end{aligned}$$

Equation 182 follows by Equations (176) and (180); Equation 183 comes from induction hypothesis; and Equation 184 is from Equation 168.

Step 4. Bound the transition kernel $P_{\bar{a}_k}\big(s_{k+1} \mid \bar{s}_k\big)$. For the lower bound, combining Equations (174), (176) and (181), we have

$$P_{\bar{a}_k}\big(s_{k+1} \mid \bar{s}_k\big) \;\geqslant\; \frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^+\, \zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1})}.$$

For the upper bound, we have

$$\begin{aligned}
P_{\bar{a}_k}\big(s_{k+1} \mid \bar{s}_k\big) \;&\leqslant\; 1 - \frac{P(\bar{s}_k, \bar{a}_k) - P(\bar{s}_{k+1}, \bar{a}_k)}{P_{\bar{a}_{k-1}}(\bar{s}_k)} && (185)\\
&\leqslant\; 1 - \frac{P(\bar{s}_k, \bar{a}_k) - P(\bar{s}_{k+1}, \bar{a}_k)}{\zeta^{\Gamma}(\bar{s}_k, \bar{a}_{k-1})}. && (186)
\end{aligned}$$

Equation 185 follows by dividing Equation 180 by $P_{\bar{a}_{k-1}}(\bar{s}_k)$; and Equation 186 is from Equation 181.

Also, combining Equations (174), (176) and (177) gives

$$P_{\bar{a}_k}\big(s_{k+1} \mid \bar{s}_k\big) \;\leqslant\; \frac{P(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-} \cdot \frac{B_{k-1}^+}{P(\bar{s}_k, \bar{a}_{k-1})}. \tag{187}$$

Taking the minimum of Equations (186) and (187) proves the stated transition bounds.

Step 5. Bound the terminal reward $\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$.

$$\begin{aligned}
\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \;&=\; \sum_{\boldsymbol{u}} \mathbb{E}_{\bar{a}_K}\left[Y \mid \bar{s}_K, \boldsymbol{u}\right] P_{\bar{a}_K}\left(\boldsymbol{u} \mid \bar{s}_K\right) && (188)\\
&=\; \sum_{\boldsymbol{u}} \mathbb{E}\left[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}\right] P_{\bar{a}_K}\left(\boldsymbol{u} \mid \bar{s}_K\right) && (189)
\end{aligned}$$

$$= \frac{\sum_{\boldsymbol{u}} \mathbb{E}\left[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}\right] P_{\bar{a}_{K-1}}(\boldsymbol{u}, \bar{s}_K)}{P_{\bar{a}_{K-1}}(\bar{s}_K)}. \tag{190}$$

Equation 188 is by the law of total expectation; Equation 189 holds by the stability of the structural mechanisms: the intervention $do(\bar{a}_K)$ leaves that generating Y unchanged, so $\mathbb{E}_{\bar{a}_K}\left[Y \mid \bar{s}_K, \boldsymbol{u}\right] = \mathbb{E}\left[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}\right]$; and Equation 190 is because a_K does not affect $\bar{s}_K$.

We next look at $\mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K]\, P(\bar{s}_K, \bar{a}_K)$.

$$\begin{aligned}\mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K]\, P(\bar{s}_K, \bar{a}_K) &= \sum_{\boldsymbol{u}} \mathbb{E}\left[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}\right] P\left(\boldsymbol{u}, \bar{s}_K, \bar{a}_K\right). \\ &= \sum_{\boldsymbol{u}} P_{\bar{a}_{K-1}}(\boldsymbol{u}, \bar{s}_K) \prod_{i=1}^{K} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big)\, \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}],\end{aligned} \tag{191}$$

$$(1 - \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K])P(\bar{s}_K, \bar{a}_K) = \sum_{\boldsymbol{u}} P_{\bar{a}_{K-1}}(\boldsymbol{u}, \bar{s}_K) \prod_{i=1}^{K} \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big)\, \big(1 - \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}]\big). \tag{192}$$

Equation 191 is by the chain rule of $P\left(\boldsymbol{u}, \bar{s}_K, \bar{a}_K\right)$ and Equation 169 at stage K.

Then, by Equations (175) and (191), we have,

$$\frac{\mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K]\, P(\bar{s}_K, \bar{a}_K)}{B_K^+} \leqslant \sum_{\boldsymbol{u}} P_{\bar{a}_{K-1}}(\boldsymbol{u}, \bar{s}_K)\, \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}] \leqslant \frac{\mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K]\, P(\bar{s}_K, \bar{a}_K)}{B_K^-}. \tag{193}$$

And by eqs. (175) and (192), we have

$$\frac{(1 - \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K])P(\bar{s}_K, \bar{a}_K)}{B_K^+} \leqslant P_{\bar{a}_{K-1}}(\bar{s}_K) - \sum_{\boldsymbol{u}} P_{\bar{a}_{K-1}}(\boldsymbol{u}, \bar{s}_K)\, \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K, \boldsymbol{u}] \leqslant \frac{(1 - \mathbb{E}[Y \mid \bar{s}_K, \bar{a}_K])P(\bar{s}_K, \bar{a}_K)}{B_K^-}. \tag{194}$$

For the lower bounds, by Equations (181), (190) and (193), we have

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \;\geqslant\; \frac{\mathbb{E}[Y \mid \bar{s}_K,\, \bar{a}_K]\, P(\bar{s}_K,\, \bar{a}_K)}{B_K^+\, \zeta^{\Gamma}(\bar{s}_K,\, \bar{a}_{K-1})}. \tag{195}$$

By Equations (190) and (194), also by the chain rule, $P(\bar{s}_K, \bar{a}_K) = \pi_{b,K}(a_K \mid \bar{s}_K, \bar{a}_{K-1})\, P(\bar{s}_K, \bar{a}_{K-1})$, and Equation 177 at $k = K$ gives $P_{\bar{a}_{K-1}}(\bar{s}_K) \geqslant P(\bar{s}_K, \bar{a}_{K-1})/B_{K-1}^+$. Hence $\frac{P(\bar{s}_K, \bar{a}_K)}{P_{\bar{a}_{K-1}}(\bar{s}_K)} \leqslant \pi_{b,K}(a_K \mid \bar{s}_K, \bar{a}_{K-1})\, B_{K-1}^+$, we have

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \;\geqslant\; 1 \;-\; \frac{(1 - \mathbb{E}[Y \mid \bar{s}_K,\, \bar{a}_K])\, \pi_{b,K}(a_K \mid \bar{s}_K,\, \bar{a}_{K-1})\, B_{K-1}^+}{B_K^-}. \tag{196}$$

For the upper bounds, by Equations (181), (190) and (194),

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \;\leqslant\; 1 \;-\; \frac{(1 - \mathbb{E}[Y \mid \bar{s}_K,\, \bar{a}_K])\, P(\bar{s}_K,\, \bar{a}_K)}{B_K^+\, \zeta^{\Gamma}(\bar{s}_K,\, \bar{a}_{K-1})}. \tag{197}$$

By Equations (190) and (193) with $\frac{P(\bar{s}_K, \bar{a}_K)}{P_{\bar{a}_{K-1}}(\bar{s}_K)} \leqslant \pi_{b,K}(a_K \mid \bar{s}_K, \bar{a}_{K-1})\, B_{K-1}^+$, we have

$$\mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \;\leqslant\; \frac{\mathbb{E}[Y \mid \bar{s}_K,\, \bar{a}_K]\, \pi_{b,K}(a_K \mid \bar{s}_K,\, \bar{a}_{K-1})\, B_{K-1}^+}{B_K^-}. \tag{198}$$

Combining Equations (195) to (198) finishes the proof. ■

Proposition 3. *Under Assumptions 1 and 2. Suppose $P(\bar{s}_K, \bar{a}_K, y) > 0$ for all $(\bar{s}_K, \bar{a}_K, y)$. Fix $\delta_1 \in (0, 1]$, and let $L_k^{\Gamma}(\bar{s}_k, \bar{a}_k), U_k^{\Gamma}(\bar{s}_k, \bar{a}_k), k \in [K]$ be the estimated causal bounds above. Let*

$$\widehat{C}_{k,\text{off}}^{\Gamma} = \Big\{\, \theta_k \in \mathbb{R}^{d_k} \,:\, L_k^{\Gamma}(\bar{s}_k, \bar{a}_k) \;\leqslant\; \psi_k(\bar{s}_k, \bar{a}_k)^{\top}\theta_k \;\leqslant\; U_k^{\Gamma}(\bar{s}_k, \bar{a}_k)\ \forall (\bar{s}_k, \bar{a}_k) \Big\}. \tag{199}$$

Then $P\big(\forall\, k \in [K],\ \theta_k^* \in \widehat{C}_{k,\text{off}}^{\Gamma}\big) \geqslant 1 - \delta_1$.

Proof. Step 1. Estimated causal bounds. Let $\mathcal{H}$ be the history collection of Section 3, $\mathcal{H} = \bigcup_{k=1}^{K} \left(\overline{\mathcal{S}}_k \times \overline{\mathcal{A}}_k \cup \overline{\mathcal{S}}_k \times \overline{\mathcal{A}}_{k-1} \right)$. Let $\widehat{m}_K(\bar{s}_K, \bar{a}_K)$ and $\widehat{\mathbb{P}}(\bar{s}_k, \bar{a}_k)$ be the same in Equation 58 and Equation 57. i.e.,

$$\widehat{m}_K(\bar{s}_K, \bar{a}_K) = \frac{1}{N_{\text{off}}} \sum_{i=1}^{N_{\text{off}}} Y_i \, \mathbf{1}\{\bar{S}_{i,K} = \bar{s}_K, \bar{A}_{i,K} = \bar{a}_K\}$$

$$\widehat{\mathbb{P}}(\bar{s}_k, \bar{a}_k) = \frac{1}{N_{\text{off}}} \sum_{i=1}^{N_{\text{off}}} \mathbf{1}\{\bar{S}_{i,k} = \bar{s}_k, \bar{A}_{i,k} = \bar{a}_k\},$$

and the bonus term is the same as in Section 3, $b_{\text{off}} = \sqrt{\frac{\log(2|\mathcal{H}|/\delta_1)}{2N_{\text{off}}}}$.

We use the same confidence box in Equation 59,

$$\begin{aligned}\mathcal{B}^{\text{off}} = \Big\{ (\mathbb{P}, m_K) : \; & |\mathbb{P}(h) - \widehat{\mathbb{P}}(h)| \leqslant b_{\text{off}} \; \forall h \in \mathcal{H}; \; |m_K(\bar{s}_K, \bar{a}_K) - \widehat{m}_K(\bar{s}_K, \bar{a}_K)| \leqslant b_{\text{off}} \;\; \forall (\bar{s}_K, \bar{a}_K); \\ & 0 < \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_k) \leqslant \mathbb{P}(\bar{s}_k, \bar{a}_{k-1}) \quad \forall k \in [K-1]; \\ & 0 < \mathbb{P}(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_{K-1}); \quad 0 \leqslant m_K(\bar{s}_K, \bar{a}_K) \leqslant \mathbb{P}(\bar{s}_K, \bar{a}_K) \Big\},\end{aligned}$$

where the pair $(\mathbb{P}, m_K)$ is the optimization variable ranging over $\mathcal{B}^{\text{off}}$.

Let $\mathcal{E}^{\text{off}}$ be the good event defined in Equation 60 for the offline dataset. We denote the pair $(\mathbb{P}^\star, m_K^\star)$ as the true but unknown population pair that generated the offline data. Specifically,

$$\mathbb{P}^\star(\bar{s}_k, \bar{a}_k) = P(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k); \; m_K^\star(\bar{s}_K, \bar{a}_K) = \mathbb{E}[Y \mathbf{1}\{\bar{S}_K = \bar{s}_K, \; \bar{A}_K = \bar{a}_K\}].$$

Under $\mathcal{E}^{\text{off}}$, we have $(\mathbb{P}^\star, m_K^\star) \in \mathcal{B}^{\text{off}}$.

We further write the bounds of Proposition 6 as functions of $(\mathbb{P}, m_K)$,

$$\alpha_i^-(\mathbb{P}) = \frac{\pi_{b,i}(\mathbb{P})}{\Gamma(1 - \pi_{b,i}(\mathbb{P})) + \pi_{b,i}(\mathbb{P})}; \quad \alpha_i^+(\mathbb{P}) = \frac{\Gamma \pi_{b,i}(\mathbb{P})}{1 - \pi_{b,i}(\mathbb{P}) + \Gamma \pi_{b,i}(\mathbb{P})}; \tag{200}$$

with $B_k^-(\mathbb{P}) = \prod_{i=1}^k \alpha_i^-(\mathbb{P})$, $B_k^+(\mathbb{P}) = \prod_{i=1}^k \alpha_i^+(\mathbb{P})$, and $B_0^- = B_0^+ = 1$.

The observed propensities are $\pi_{b,i}(\mathbb{P}) = \frac{\mathbb{P}(\bar{s}_i, \bar{a}_i)}{\mathbb{P}(\bar{s}_i, \bar{a}_{i-1})}$ and

$$\zeta^\Gamma(\bar{s}_{k+1}, \bar{a}_k; \mathbb{P}) = \min \left\{ \frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{B_k^-(\mathbb{P})}, \; \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k) - \mathbb{P}(\bar{s}_k, \bar{a}_k) + \zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P}) \right\}, \tag{201}$$

with $\zeta^\Gamma(s_1; \mathbb{P}) = \mathbb{P}(s_1)$.

We denote the transition bounds as follows:

$$\underline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}) = \frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{B_k^+(\mathbb{P}) \, \zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})}, \tag{202}$$

$$\overline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}) = \min \left\{ 1 - \frac{\mathbb{P}(\bar{s}_k, \bar{a}_k) - \mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\zeta^\Gamma(\bar{s}_k, \bar{a}_{k-1}; \mathbb{P})}, \frac{B_{k-1}^+(\mathbb{P})}{B_k^-(\mathbb{P})} \cdot \frac{\mathbb{P}(\bar{s}_{k+1}, \bar{a}_k)}{\mathbb{P}(\bar{s}_k, \bar{a}_{k-1})} \right\}. \tag{203}$$

Similarly, we denote the terminal bounds as follows:

$$\underline{\mu}^\Gamma(\mathbb{P}, m_K) := \max \left\{ \frac{m_K}{B_K^+(\mathbb{P}) \, \zeta^\Gamma(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})}, \; 1 - \frac{B_{K-1}^+(\mathbb{P})}{B_K^-(\mathbb{P})} \cdot \frac{\mathbb{P}(\bar{s}_K, \bar{a}_K) - m_K}{\mathbb{P}(\bar{s}_K, \bar{a}_{K-1})} \right\} \tag{204}$$

$$\overline{\mu}^\Gamma(\mathbb{P}, m_K) := \min \left\{ 1 - \frac{\mathbb{P}(\bar{s}_K, \bar{a}_K) - m_K}{B_K^+(\mathbb{P}) \, \zeta^\Gamma(\bar{s}_K, \bar{a}_{K-1}; \mathbb{P})}, \; \frac{B_{K-1}^+(\mathbb{P})}{B_K^-(\mathbb{P})} \cdot \frac{m_K}{\mathbb{P}(\bar{s}_K, \bar{a}_{K-1})} \right\}. \tag{205}$$

Where Equations (204) and (205) follow by $\pi_{b,K}(\mathbb{P}) = \frac{\mathbb{P}(\bar{s}_K, \bar{a}_K)}{\mathbb{P}(\bar{s}_K, \bar{a}_{K-1})}$.

Then, define the transition set as

$$\widehat{\mathcal{P}}_k^\Gamma(\bar{s}_k, \bar{a}_k) = \left\{ p \in \Delta(\mathcal{S}_{k+1}) : \; \inf_{\mathcal{B}^{\text{off}}} \underline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}) \leqslant p(s_{k+1}) \leqslant \sup_{\mathcal{B}^{\text{off}}} \overline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}) \; \forall s_{k+1} \in \mathcal{S}_{k+1} \right\}. \tag{206}$$

Following the above, we set the estimated causal bounds at stage $k < K$ and the terminal stage as:

$$L_K^\Gamma(\bar{s}_K, \bar{a}_K) = \max\left\{\inf_{\mathcal{B}^{\text{off}}} \underline{\mu}^\Gamma(\mathbb{P}, m_K),\ 0\right\}, \quad U_K^\Gamma(\bar{s}_K, \bar{a}_K) = \min\left\{\sup_{\mathcal{B}^{\text{off}}} \overline{\mu}^\Gamma(\mathbb{P}, m_K),\ 1\right\}, \tag{207}$$

$$L_k^\Gamma(\bar{s}_k, \bar{a}_k) = \min_{p \in \widehat{\mathcal{P}}_k^\Gamma} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} L_{k+1}^\Gamma, \quad U_k^\Gamma(\bar{s}_k, \bar{a}_k) = \max_{p \in \widehat{\mathcal{P}}_k^\Gamma} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} U_{k+1}^\Gamma. \tag{208}$$

If $\widehat{\mathcal{P}}_k^\Gamma(\bar{s}_k, \bar{a}_k) = \varnothing$ we set $L_k^\Gamma = 0$ and $U_k^\Gamma = 1$, as in Section 3.

Step 2. Concentration. By Hoeffding's inequality (Lemma 11) and a union bound over $\mathcal{H}$, we have $\Pr(\mathcal{E}^{\text{off}}) \geqslant 1 - \delta_1$ exactly as in Section 3.

Since on $\mathcal{E}^{\text{off}}$, we have $(\mathbb{P}^\star, m_K^\star) \in \mathcal{B}^{\text{off}}$, then,

$$\inf_{\mathcal{B}^{\text{off}}} \underline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}) \ \leqslant\ \underline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}^\star), \qquad \overline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}^\star) \ \leqslant\ \sup_{\mathcal{B}^{\text{off}}} \overline{\tau}_k^\Gamma(s_{k+1}; \mathbb{P}), \tag{209}$$

and similarly holds for $\underline{\mu}^\Gamma(\mathbb{P}, m_K), \overline{\mu}^\Gamma(\mathbb{P}, m_K)$. Then by Equations (206) and (209) and Proposition 6, we have

$$P_{\bar{a}_k}(\cdot \mid \bar{s}_k) \ \in\ \widehat{\mathcal{P}}_k^\Gamma(\bar{s}_k, \bar{a}_k). \tag{210}$$

The same chaining gives $L_K^\Gamma \leqslant \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K] \leqslant U_K^\Gamma$.

Step 3. Backward propagation to Q_k^*. We show by backward induction on k that

$$L_k^\Gamma(\bar{s}_k, \bar{a}_k) \ \leqslant\ Q_k^*(\bar{s}_k, \bar{a}_k) \ \leqslant\ U_k^\Gamma(\bar{s}_k, \bar{a}_k). \tag{211}$$

Base case. At $k = K$ this is the terminal statement of Step 2 (last line above), since $Q_K^*(\bar{s}_K, \bar{a}_K) = \mathbb{E}_{\bar{a}_K}[Y \mid \bar{s}_K]$.

Induction step. Assume Equation 211 at stage $k+1$, i.e., $L_{k+1}^\Gamma(\bar{s}_{k+1}, \bar{a}_{k+1}) \leqslant Q_{k+1}^*(\bar{s}_{k+1}, \bar{a}_{k+1}) \leqslant U_{k+1}^\Gamma(\bar{s}_{k+1}, \bar{a}_{k+1})$. Then, we have,

$$\begin{aligned}
Q_k^*(\bar{s}_k, \bar{a}_k) &= \sum_{s_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \max_{a_{k+1}} Q_{k+1}^* && (212)\\
&\leqslant \sum_{s_{k+1}} P_{\bar{a}_k}(s_{k+1} \mid \bar{s}_k) \max_{a_{k+1}} U_{k+1}^\Gamma && (213)\\
&\leqslant \max_{p \in \widehat{\mathcal{P}}_k^\Gamma} \sum_{s_{k+1}} p(s_{k+1}) \max_{a_{k+1}} U_{k+1}^\Gamma && (214)\\
&= U_k^\Gamma(\bar{s}_k, \bar{a}_k), && (215)
\end{aligned}$$

where Equation 212 is by Equation 1; Equation 213 is from the induction hypothesis; Equation 214 follows by Equation 210; and Equation 214 comes from Equation 208.

The lower bound follows identically, with $\min$ in place of $\max$ and L_{k+1}^Γ in place of U_{k+1}^Γ.

Step 4. Conclusion. By Proposition 4, $Q_k^*(\bar{s}_k, \bar{a}_k) = \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k^*$. Substituting into Equation 211 gives,

$$L_k^\Gamma(\bar{s}_k, \bar{a}_k) \ \leqslant\ \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k^* \ \leqslant\ U_k^\Gamma(\bar{s}_k, \bar{a}_k)$$

Then, on $\mathcal{E}^{\text{off}}$, we have the following holds,

$$P\left(\forall(\bar{s}_k, \bar{a}_k) :\ L_k^\Gamma(\bar{s}_k, \bar{a}_k) \ \leqslant\ \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k^* \ \leqslant\ U_k^\Gamma(\bar{s}_k, \bar{a}_k)\right) \ \geqslant\ 1 - \delta_1.$$

And the estimated confidence set is:

$$\widehat{C}_{k,\text{off}}^\Gamma = \left\{\theta_k \in \mathbb{R}^{d_k} :\ L_k^\Gamma(\bar{s}_k, \bar{a}_k) \ \leqslant\ \psi_k(\bar{s}_k, \bar{a}_k)^\top \theta_k \ \leqslant\ U_k^\Gamma(\bar{s}_k, \bar{a}_k),\ \forall(\bar{s}_k, \bar{a}_k)\right\}.$$

Which finishes the proof. ■

Lemma 5. *Under Assumptions 2, fix a stage $i \in [K]$, a history $(\bar{s}_i, \bar{a}_i)$ with $P(\bar{s}_i, \bar{a}_{i-1}) > 0$, and an action $a_i \in \mathcal{A}_I$. Write $\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})$ for the observed propensity. Then, for every $\boldsymbol{u} \in \mathrm{supp}(\boldsymbol{U})$,*

$$\alpha_i^- = \frac{\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{\Gamma - (\Gamma-1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})} \leqslant \pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big) \leqslant \frac{\Gamma \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{1 + (\Gamma-1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})} = \alpha_i^+. \tag{216}$$

Moreover $\alpha_i^- \leqslant \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}) \leqslant \alpha_i^+$, with equality if and only if $\Gamma = 1$.

Proof. This lemma converts the odds-ratio restriction Equation 12 into an explicit interval for the full-information propensity $\pi_{b,i}\big(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}\big)$, expressed in terms of the observed propensity $\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})$.

Let

$$\mathrm{OR} = \frac{\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u})}{1 - \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u})} \cdot \frac{1 - \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})} \tag{217}$$

denote the odds ratio appearing in Equation 12. Assumption 2 states that $\mathrm{OR} \in [\Gamma^{-1}, \Gamma]$ for every $\boldsymbol{u} \in \mathrm{supp}(\boldsymbol{U})$.

Solve Equation 217, we have

$$\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}) = \frac{\mathrm{OR}\, \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{1 + (\mathrm{OR} - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}. \tag{218}$$

Regarded Equation 218 as a function of OR on $(0, \infty)$ and take a derivative, we have

$$\begin{aligned}
&\frac{\partial}{\partial \mathrm{OR}} \frac{\mathrm{OR}\, \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{1 + (\mathrm{OR} - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})} \\
&= \frac{\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})\big(1 + (\mathrm{OR} - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})\big) - \mathrm{OR}\, \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})^2}{\big(1 + (\mathrm{OR} - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})\big)^2} \\
&= \frac{\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})(1 - \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}))}{\big(1 + (\mathrm{OR} - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})\big)^2} > 0.
\end{aligned} \tag{219}$$

Equation 219 is because $\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}) \in (0, 1)$. Hence $\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u})$ is strictly increasing in OR, and substituting the two endpoints $\mathrm{OR} = \Gamma^{-1}$ and $\mathrm{OR} = \Gamma$ into Equation 218 yields

$$\frac{\Gamma^{-1}\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{1 + (\Gamma^{-1} - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})} \leqslant \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}, \boldsymbol{u}) \leqslant \frac{\Gamma \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}{1 + (\Gamma - 1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})}.$$

Multiplying the numerator and denominator of $\frac{\Gamma^{-1}\pi_{b,i}(a_i|\bar{s}_i,\bar{a}_{i-1})}{1+(\Gamma^{-1}-1)\pi_{b,i}(a_i|\bar{s}_i,\bar{a}_{i-1})}$ by Γ gives $\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})/\big(\Gamma - (\Gamma-1)\pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1})\big) = \alpha_i^-$, which proves Equation 216.

If $\Gamma = 1$, $\alpha_i^- = \pi_{b,i}(a_i \mid \bar{s}_i, \bar{a}_{i-1}) = \alpha_i^+$; if $\Gamma > 1$ then $\Gamma^{-1} < 1 < \Gamma$ and both inequalities are strict. ■

E.2 Proof for Regret Bound

Theorem 2. *Let Assumptions 1 and 2 hold. Let χ, C_1, C, c_β, λ and β_k be as in Theorem 1. Then for any $\delta, \delta_1 \in (0, 1]$, with probability at least $1 - \delta - 2\delta_1$, Algorithm 4 achieves*

$$R(T) \leqslant 2\sqrt{2TK\log(2/\delta_1)} + \min\Big\{2C_1 d_K^{3/2} K\sqrt{T \cdot \chi\, \log(\lambda + T)},\ T\sum_{k=1}^{K} W_{k,\mathrm{off}}^{\Gamma}\Big\}.$$

Proof. The algorithm is unchanged from Section 3, except that the offline confidence set $\widehat{\mathcal{C}}_k^{\mathrm{off}}$ is replaced by $\widehat{\mathcal{C}}_{k,\mathrm{off}}^{\Gamma}$. See Algorithm 4 for convenience. We follow Steps 1–5 of the proof of Theorem 1. We only write the modifications in the following.

Step 1. Bound for the estimated linear parameter. Unchanged. $\|\widehat{\theta}_k^t\|_2 \leqslant \sqrt{\frac{d_k t}{\lambda}}$.

Step 2. Concentration bounds. The argument is unchanged, with $\widetilde{\mathcal{V}}_{k+1}$ replaced by $\widetilde{\mathcal{V}}_{k+1}^{\Gamma}$, defined identically except that the offline term is $\max_{\theta \in \widehat{\mathcal{C}}_{k+1,\text{off}}^{\Gamma}} \theta^\top \psi_{k+1}$, which, like its counterpart in Section 3, is deterministic given the offline data and does not depend on (ω, β, Λ). Since $\widehat{\mathcal{C}}_{k+1,\text{off}}^{\Gamma} \in \mathfrak{P}_{k+1}$ (take $L = L_{k+1}^{\Gamma}$ and $U = U_{k+1}^{\Gamma}$), every $V \in \widetilde{\mathcal{V}}_{k+1}^{\Gamma}$ is of the form Equation 3 with $\theta = \theta' = \omega$, $A = A' = \beta^2 \Lambda^{-1}$ and $\mathcal{P} = \widehat{\mathcal{C}}_{k+1,\text{off}}^{\Gamma}$, so $\widetilde{\mathcal{V}}_{k+1}^{\Gamma} \subseteq \mathcal{V}_{k+1}$; the covering bound Equation 85 is unchanged, since the cover is taken over (ω, β, Λ) only. The event is

$$\mathcal{E}_1 := \Big\{ \forall k < K,\ \forall t \in [T],\ \forall V \in \widetilde{\mathcal{V}}_{k+1}^{\Gamma} : \Big\| \sum_{\tau=1}^{t-1} \psi_k^\tau \Big(V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}_{\bar{a}_k^\tau}\big[V(\bar{S}_{k+1}, \bar{a}_k^\tau) \mid \bar{s}_k^\tau\big] \Big) \Big\|_{(\Lambda_k^t)^{-1}} \leqslant \rho \Big\}$$
$$\bigcap \Big\{ \forall t \in [T] : \Big\| \sum_{\tau=1}^{t-1} \psi_K^\tau \left[Y^\tau - \mathbb{E}\left[Y^\tau \mid \mathcal{F}_{K,\tau-1} \right] \right] \Big\|_{(\Lambda_K^t)^{-1}} \leqslant \rho \Big\}. \tag{220}$$

Where $\rho = C d_K \sqrt{\log\left(\frac{4(c_\beta+1) d_K K T}{\delta}\right)}$ and $P(\mathcal{E}_1) \geqslant 1 - \delta$.

Step 3. Optimism bounds. Unchanged. We work on the good event in Equation 88, i.e.,

$$\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}^{\text{off}}, \qquad \Pr(\mathcal{E}) \geqslant 1 - \delta - \delta_1.$$

Then on the good event $\mathcal{E}$, $\forall k \in [K],\ \forall t \in [T],\ \forall (\bar{s}_k, \bar{a}_k)$, we have

$$\widehat{Q}_k^t(\bar{s}_k, \bar{a}_k) \geqslant Q_k^*(\bar{s}_k, \bar{a}_k).$$

Step 4. Recursive formula for estimation error. By Equation 97, we have

$$\widehat{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t) = \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t). \tag{221}$$

Similar as the step 4 in Theorem 1, we give two bounds on the right-hand side and take their minimum.

(a) For the online branch (unchanged)

$$\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \leqslant \mathbb{E}_{\bar{a}_k^t}\left[(\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) \mid \bar{s}_k^t \right] + (\mathcal{C} + \beta_k) \|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}}, \tag{222}$$

and

$$\begin{aligned} \mathcal{C} &= \rho + \sqrt{\lambda d_K} = C d_K \sqrt{\chi} + \sqrt{\lambda d_K}; \\ \beta_k &= C d_K \sqrt{\chi} + \sqrt{\lambda d_k}, \end{aligned}$$

where $\mathcal{C}$ and β_k are as in Theorem 1.

(b) For the offline branch, we have,

$$\begin{aligned} &\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\ &= \widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^*(\bar{s}_k^t, \bar{a}_k^t) + Q_k^*(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\ &\leqslant U_k^{\Gamma} - L_k^{\Gamma} + Q_k^*(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \qquad (223) \\ &\leqslant U_k^{\Gamma} - L_k^{\Gamma} + \mathbb{E}_{\bar{a}_k^t}\left[\widehat{V}_{k+1}^t(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \,|\bar{s}_k^t \right] - \mathbb{E}_{\bar{a}_k^t}\left[V_{k+1}^{\boldsymbol{\pi}_t}(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \mid \bar{s}_k^t \right], \qquad (224) \end{aligned}$$

where Equation 223 is because both $\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t)$ and $Q_k^*(\bar{s}_k^t, \bar{a}_k^t)$ lie in the band $[L_k^{\Gamma}(\bar{s}_k^t, \bar{a}_k^t), U_k^{\Gamma}(\bar{s}_k^t, \bar{a}_k^t)]$; Equation 224 follows by Equation (109).

Combining Equation 222 and Equation 224, we have

$$\widehat{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \leqslant (\widehat{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_{k+1}^t, \bar{a}_k^t) + \zeta_{k+1}^t + b_k^t,$$

where $b_k^t := \min\Big\{\big(\mathcal{C}+\beta_k\big)\big\|\psi_k(\bar{s}_k^t,\bar{a}_k^t)\big\|_{(\Lambda_k^t)^{-1}},\ U_k^\Gamma - L_k^\Gamma\Big\}$.

Step 5. Finalize. Unchanged except we replace W_k^{off} with $W_{k,\text{off}}^\Gamma$ and $W_{k,\text{off}}^\Gamma = \max_{(\bar{s}_k,\bar{a}_k)\in\overline{\mathcal{S}}_k\times\overline{\mathcal{A}}_k}\big[U_k^\Gamma(\bar{s}_k,\bar{a}_k) - L_k^\Gamma(\bar{s}_k,\bar{a}_k)\big]$. Then, on $\mathcal{E}$ and with probability at least $1-\delta-2\delta_1$,

$$\begin{aligned} R(T) &\leqslant 2\sqrt{2TK\log(2/\delta_1)} + \sum_{k=1}^K \min\Big\{2C_1 d_K^{3/2}\sqrt{T\,\chi\,\log\big(\lambda+\tfrac{T}{d_k}\big)},\ T\,W_{k,\text{off}}^\Gamma\Big\} \\ &\leqslant 2\sqrt{2TK\log(2/\delta_1)} + \min\Big\{2C_1 d_K^{3/2}K\sqrt{T\,\chi\,\log(\lambda+T)},\ T\sum_{k=1}^K W_{k,\text{off}}^\Gamma\Big\} \\ &= \widetilde{\mathcal{O}}\Big(\sqrt{KT} + \min\Big\{\sqrt{d_K^3K^2T},\ T\sum_{k=1}^K W_{k,\text{off}}^\Gamma\Big\}\Big). \end{aligned}$$

Which finishes the proof. ■

F Proof for Section 5

Theorem 3. *Let Assumptions 1 and 2 with $\Gamma = 1$ hold. Let χ, C, C_1, c_β, λ and β_k be as in Theorem 1, let γ_{off} be as in Definition 4, and let $W_{k,\text{off}}^{\Gamma=1}$ be as in Definition 3 with $\Gamma=1$. For any $\delta,\delta_1\in(0,1]$, suppose $8\gamma_{\text{off}}\log\big(Kd_K/\delta_1\big)\leqslant N_{\text{off}}\leqslant T^{d_K}$; let $\widetilde{\chi}=\log\big(\frac{4(c_\beta+1)d_KK(N_{\text{off}}+T)}{\delta}\big)$ and $\widetilde{\beta}_k = Cd_K\sqrt{\widetilde{\chi}}+\sqrt{\lambda d_k}$. Then, with probability at least $1-\delta-3\delta_1$, Algorithm 2 achieves*

$$R(T) \leqslant 2\sqrt{2TK\log(2/\delta_1)} + \min\Big\{2C_1d_K^{3/2}K\sqrt{T\,\chi\,\log(\lambda+T)},\ C_1d_KKT\sqrt{\tfrac{2\widetilde{\chi}\,\gamma_{\text{off}}}{N_{\text{off}}}},\ T\sum_{k=1}^K W_{k,\text{off}}^{\Gamma=1}\Big\}.$$

Proof. We follow Steps 1–5 of the proof of Theorem 1, indicating only the modifications. Throughout we work on the good event

$$\widetilde{\mathcal{E}} := \widetilde{\mathcal{E}}_1 \bigcap \mathcal{E}_{\text{off}}^{\text{cov}} \bigcap \mathcal{E}^{\text{off}}, \qquad \Pr(\widetilde{\mathcal{E}}) \geqslant 1-\delta-2\delta_1, \tag{225}$$

where $\Pr(\widetilde{\mathcal{E}}_1)\geqslant 1-\delta$ by Lemma 9(ii), $\Pr(\mathcal{E}_{\text{off}}^{\text{cov}})\geqslant 1-\delta_1$ by Lemma 8, and $\Pr(\mathcal{E}^{\text{off}})\geqslant 1-\delta_1$ by Proposition 3.

Step 1. Bound for the estimated linear parameter. By Lemma 9(i), for all $k\in[K]$ and $t\in[T]$, $\|\widetilde{\theta}_k^t\|_2\leqslant\sqrt{d_k(N_{\text{off}}+t)/\lambda}$ and $\|\widehat{\theta}_k^t\|_2\leqslant\sqrt{d_kt/\lambda}$.

Step 2. Concentration bounds. The iterate $\widetilde{V}_{k+1}^t = \max_{a_{k+1}}\widetilde{Q}_{k+1}^t$ lies in the subfamily $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}}$ of Equation 3 obtained with $\theta=\omega$, $A=\beta^2\Lambda^{-1}$ (online branch), $\theta'=\widetilde{\omega}$, $A'=\widetilde{\beta}^2\widetilde{\Lambda}^{-1}$ (pooled branch), and $\mathcal{P}=\widehat{C}_{k+1,\text{off}}^{\Gamma=1}\in\mathfrak{P}_{k+1}$ (i.e., $L=L_{k+1}^{\Gamma=1}$, $U=U_{k+1}^{\Gamma=1}$), so so $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}}\subseteq\mathcal{V}_{k+1}$ and Assumption 1(ii) applies on $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}}$. The class is parametrized by the two triples (ω,β,Λ) and $(\widetilde{\omega},\widetilde{\beta},\widetilde{\Lambda})$, so its covering number is the product of two copies of Equation 85. By Lemma 9(ii), with probability at least $1-\delta$, simultaneously for all $k<K$ and all $t\in[T]$,

$$\Big\|\sum_{\iota=1}^{N_{\text{off}}+t-1}\psi_k^\iota\Big[\widetilde{V}_{k+1}^t(\bar{s}_{k+1}^\iota,\bar{a}_k^\iota) - \mathbb{E}_{\bar{a}_k^\iota}\big[\widetilde{V}_{k+1}^t(\bar{S}_{k+1},\bar{a}_k^\iota)\mid\bar{s}_k^\iota\big]\Big]\Big\|_{(\widetilde{\Lambda}_k^t)^{-1}} \leqslant Cd_K\sqrt{\widetilde{\chi}},$$

and at the terminal stage, for all $t\in[T]$,

$$\Big\|\sum_{\iota=1}^{N_{\text{off}}+t-1}\psi_K^\iota\big[Y^\iota - \mathbb{E}[Y^\iota\mid\mathcal{G}_{K,\iota-1}]\big]\Big\|_{(\widetilde{\Lambda}_K^t)^{-1}} \leqslant Cd_K\sqrt{\widetilde{\chi}}.$$

By the second and third components of $\widetilde{\mathcal{E}}_1$, the same bounds hold for the online-only sums, with $\sum_{\tau=1}^{t-1}$ in place of $\sum_{\iota=1}^{N_{\text{off}}+t-1}$, $\mathcal{F}_{k,\tau-1}$ in place of $\mathcal{G}_{k,\iota-1}$, $(\Lambda_k^t)^{-1}$ in place of $(\widetilde{\Lambda}_k^t)^{-1}$, and radius $Cd_K\sqrt{\chi}$; the latter uses $N_{\text{off}}\leqslant T^{d_K}$.

Step 3. Optimism bounds. On $\widetilde{\mathcal{E}}$, Lemma 10 gives

$$\widetilde{Q}_k^t(\bar{s}_k, \bar{a}_k) \;\geqslant\; Q_k^*(\bar{s}_k, \bar{a}_k) \qquad \text{for all } k \in [K],\ t \in [T],\ (\bar{s}_k, \bar{a}_k), \tag{226}$$

using $\theta_K^* \in \widetilde{\mathcal{C}}_K^t \bigcap \widehat{\mathcal{C}}_K^t$ and $\theta_{\widetilde{V}_{k+1}^t} \in \widetilde{\mathcal{C}}_k^t \bigcap \widehat{\mathcal{C}}_k^t$ from Lemma 9(iv) for the two regression branches, and $\theta_k^* \in \widehat{C}_{k,\text{off}}^{\Gamma=1}$ from Proposition 3 for the third.

Step 4. Recursive formula for estimation error. By Equation 97, we have

$$\widetilde{V}_k^t(\bar{s}_k^t, \bar{a}_{k-1}^t) - V_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_{k-1}^t) = \widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t). \tag{227}$$

Similar as the Step 4 in Theorem 1, we give three bounds on the right-hand side and take their minimum. By Algorithm 2

$$\widetilde{Q}_k^t(\bar{s}_k, \bar{a}_k) \leftarrow \max\Big\{ \min\Big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), \max_{\theta \in \widetilde{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k), \max_{\theta \in \widehat{C}_{k,\text{off}}^{\Gamma=1}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ 1\Big\},\ 0\Big\}.$$

(a) For the online branch $\max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k)$ (unchanged)

$$\widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \;\leqslant\; \mathbb{E}_{\bar{a}_k^t}\Big[(\widetilde{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) \mid \bar{s}_k^t\Big] + (\mathcal{C} + \beta_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}}.$$

(b) For the pooled branch $\max_{\theta \in \widetilde{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k)$,

$$\widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \;\leqslant\; \mathbb{E}_{\bar{a}_k^t}\Big[(\widetilde{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_k^t, S_{k+1}^t, \bar{a}_k^t) \mid \bar{s}_k^t\Big] + (\widetilde{\mathcal{C}} + \widetilde{\beta}_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\widetilde{\Lambda}_k^t)^{-1}},$$

using Lemma 9(iii) where

$$\widetilde{\mathcal{C}} = \widetilde{\rho} + \sqrt{\lambda d_K} \;=\; C d_K \sqrt{\widetilde{\chi}} + \sqrt{\lambda d_K}; \quad \widetilde{\beta}_k = C d_K \sqrt{\widetilde{\chi}} + \sqrt{\lambda d_k}; \quad \widetilde{\chi} = \log\Big(\tfrac{4(c_\beta+1) d_K K (N_{\text{off}}+T)}{\delta}\Big).$$

(c) For the offline branch $\max_{\theta \in \widehat{C}_{k,\text{off}}^{\Gamma=1}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k)$, we have,

$$\begin{aligned}
&\widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\
&\quad= \widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^*(\bar{s}_k^t, \bar{a}_k^t) + Q_k^*(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \\
&\quad\leqslant U_k^{\Gamma=1} - L_k^{\Gamma=1} + Q_k^*(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) &(228)\\
&\quad\leqslant U_k^{\Gamma=1} - L_k^{\Gamma=1} + \mathbb{E}_{\bar{a}_k^t}\big[\widetilde{V}_{k+1}^t(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \,|\bar{s}_k^t\big] - \mathbb{E}_{\bar{a}_k^t}\big[V_{k+1}^{\boldsymbol{\pi}_t}(S_{k+1}^t, \bar{s}_k^t, \bar{a}_k^t) \mid \bar{s}_k^t\big], &(229)
\end{aligned}$$

where Equation 228 is because both $\widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t)$ and $Q_k^*(\bar{s}_k^t, \bar{a}_k^t)$ lie in the band $[L_k^{\Gamma=1}(\bar{s}_k^t, \bar{a}_k^t), U_k^{\Gamma=1}(\bar{s}_k^t, \bar{a}_k^t)]$; and Equation 229 follows by Equation (109).

Combining (a), (b), and (c), we have

$$\widetilde{Q}_k^t(\bar{s}_k^t, \bar{a}_k^t) - Q_k^{\boldsymbol{\pi}_t}(\bar{s}_k^t, \bar{a}_k^t) \;\leqslant\; (\widetilde{V}_{k+1}^t - V_{k+1}^{\boldsymbol{\pi}_t})(\bar{s}_{k+1}^t, \bar{a}_k^t) + \widetilde{\zeta}_{k+1}^t + b_k^t,$$

where $b_k^t := \min\Big\{(\mathcal{C} + \beta_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\Lambda_k^t)^{-1}},\ (\widetilde{\mathcal{C}} + \widetilde{\beta}_k)\|\psi_k(\bar{s}_k^t, \bar{a}_k^t)\|_{(\widetilde{\Lambda}_k^t)^{-1}},\ U_k^{\Gamma=1} - L_k^{\Gamma=1}\Big\}$.

Step 5. Finalize the results. We bound $\sum_{t=1}^T b_k^t$ as follows:

$$\sum_{t=1}^T b_k^t \;\leqslant\; \min\Big\{2C_1 d_K \sqrt{T\, d_k\, \chi\, \log\big(\lambda + \tfrac{T}{d_k}\big)},\ C_1 d_K\, T \sqrt{\tfrac{2\widetilde{\chi}\,\gamma_{\text{off}}}{N_{\text{off}}}},\ T\, W_{k,\text{off}}^{\Gamma=1}\Big\}.$$

Then, with probability at least $1 - \delta - 3\delta_1$,

$$\begin{aligned}
R(T) &\leqslant 2\sqrt{2TK\log(2/\delta_1)} + \sum_{k=1}^K \min\Big\{2C_1 d_K \sqrt{T\, d_k\, \chi\, \log\big(\lambda + \tfrac{T}{d_k}\big)},\ C_1 d_K\, T \sqrt{\tfrac{2\widetilde{\chi}\,\gamma_{\text{off}}}{N_{\text{off}}}},\ T\, W_{k,\text{off}}^{\Gamma=1}\Big\} \\
&\leqslant 2\sqrt{2TK\log(2/\delta_1)} + \min\Big\{2C_1 d_K^{3/2} K \sqrt{T\,\chi\,\log(\lambda+T)},\ C_1 d_K K T \sqrt{\tfrac{2\widetilde{\chi}\,\gamma_{\text{off}}}{N_{\text{off}}}},\ T \sum_{k=1}^K W_{k,\text{off}}^{\Gamma=1}\Big\} \\
&= \widetilde{\mathcal{O}}\Big(\sqrt{KT} + \min\Big\{\sqrt{d_K^3 K^2 T},\ d_K K T \sqrt{\tfrac{\gamma_{\text{off}}}{N_{\text{off}}}},\ T \sum_{k=1}^K W_{k,\text{off}}^{\Gamma=1}\Big\}\Big),
\end{aligned}$$

This finishes the proof. ■

Lemma 6. *Under Assumptions 1 and 2 with $\Gamma = 1$, for offline trajectory n, we have*

(a). at the terminal stage, $\mathbb{E}\Big[Y^{(n)} \mid \bar{s}_K^{(n)}, \bar{a}_K^{(n)}\Big] = \Big\langle \psi_K^{(n)}(\bar{s}_K^{(n)}, \bar{a}_K^{(n)}), \theta_K^* \Big\rangle$.

(b). At every nonterminal stage $k < K$ and every $V_{k+1} \in \mathcal{V}_{k+1}$ (built from Equation 3),

$$\mathbb{E}\left[V_{k+1}\left(\bar{s}_k^{(n)}, S_{k+1}^{(n)}, \bar{a}_k^{(n)}\right) \mid \bar{s}_k^{(n)}, \bar{a}_k^{(n)}\right] = \Big\langle \psi_k(\bar{s}_k^{(n)}, \bar{a}_k^{(n)}), \theta_{V_{k+1}} \Big\rangle,$$

where $\theta_{V_{k+1}}$ is the vector in Assumption 1(ii).

Proof. This lemma shows that, under ful unconfounding, the offline observations satisfy the same conditional regression equations as the online ones, so that regressing the offline outcomes on $\psi_k^{(i)}$ targets the same parameter as the online regression.

Since Assumption 2 holds with $\Gamma = 1$, by Equation 12, we have

$$\pi_{b,k}\left(a_k \mid \bar{s}_k, \bar{a}_{k-1}, \boldsymbol{u}\right) = \pi_{b,k}\left(a_k \mid \bar{s}_k, \bar{a}_{k-1}\right),$$

so that treatment assignment is independent of $\boldsymbol{U}$ given the observed history. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[V_{k+1}\left(\bar{s}_k^{(n)}, S_{k+1}^{(n)}, \bar{a}_k^{(n)}\right) \,\middle|\, \bar{S}_k^{(n)} = \bar{s}_k^{(n)}, \bar{A}_k^{(n)} = \bar{a}_k^{(n)}\right] \\
&= \sum_{s_{k+1} \in \mathcal{S}_{k+1}} V_{k+1}\big(\bar{s}_k^{(n)}, s_{k+1}, \bar{a}_k^{(n)}\big)\, P\big(S_{k+1}^{(n)} = s_{k+1} \mid \bar{s}_k^{(n)}, \bar{a}_k^{(n)}\big) && (230)\\
&= \sum_{s_{k+1} \in \mathcal{S}_{k+1}} V_{k+1}\big(\bar{s}_k^{(n)}, s_{k+1}, \bar{a}_k^{(n)}\big)\, P_{\bar{a}_k^{(n)}}\big(s_{k+1} \mid \bar{s}_k^{(n)}\big) && (231)\\
&= \mathbb{E}_{\bar{a}_k^{(n)}}\Big[V_{k+1}\big((\bar{s}_k^{(n)}, S_{k+1}), \bar{a}_k^{(n)}\big) \mid \bar{s}_k^{(n)}\Big] && (232)\\
&= \Big\langle \psi_k(\bar{s}_k^{(n)}, \bar{a}_k^{(n)}),\ \theta_{V_{k+1}} \Big\rangle, && (233)
\end{aligned}$$

where Equation 230 is by the definition of the conditional expectation; Equation 231 holds because $A_i \perp\!\!\!\perp \boldsymbol{U} \mid (\bar{S}_i, \bar{A}_{i-1})$ for every $i \leqslant k$ by Assumption 2 with $\Gamma = 1$, so every back-door path from $\bar{A}_k$ to S_{k+1} is blocked by $\bar{S}_k$ and the sequential back-door criterion (Pearl & Robins, 1995) identifies the interventional kernel with the observational one; Equation 232 is the definition of the interventional conditional expectation; and Equation 233 follows by Assumption 1(ii) applied to $V_{k+1} \in \mathcal{V}_{k+1}$.

The same argument at stage K gives $\mathbb{E}\big[Y^{(n)} \mid \bar{s}_K^{(n)}, \bar{a}_K^{(n)}\big] = \mathbb{E}_{\bar{a}_K^{(n)}}\big[Y^{(n)} \mid \bar{s}_K^{(n)}\big] = \big\langle \psi_K(\bar{s}_K^{(n)}, \bar{a}_K^{(n)}),\ \theta_K^* \big\rangle$, the first equality by Assumption 2 with $\Gamma = 1$ and the second by Assumption 1(i) with $\theta_K^* = w_K$ as constructed in the proof of Proposition 4. This finishes the proof. ■

Lemma 7. *Fix $k < K$ and $V_{k+1} \in \mathcal{V}_{k+1}$ (from Equation 3), and let $\theta_{V_{k+1}}$ be the vector from Assumption 1(ii). Then, for every index ι of the pooled sequence (combining offline and online data),*

$$\mathbb{E}\Big[V_{k+1}\big(\bar{s}_k^{\iota}, S_{k+1}^{\iota}, \bar{a}_k^{\iota}\big) - \big\langle \psi_k^{\iota}, \theta_{V_{k+1}} \big\rangle \Big| \mathcal{G}_{k,\iota-1}, \bar{s}_k^{\iota}, \bar{a}_k^{\iota}\Big] = 0;\ \Big|V_{k+1}\big(\bar{s}_k^{\iota}, S_{k+1}^{\iota}, \bar{a}_k^{\iota}\big) - \big\langle \psi_k^{\iota}, \theta_{V_{k+1}} \big\rangle\Big| \leqslant 1.$$

At the terminal stage $k = K$, the same statements hold with $Y^{\iota} - \langle \psi_K^{\iota}, \theta_K^* \rangle$ in place of the above.

Proof. This lemma shows that the pooled sequence is a bounded martingale difference sequence with respect to $\{\mathcal{G}_{k,\iota}\}_{\iota}$, which is what the self-normalized concentration bound of Lemma 14 requires. The offline and online terms have conditional mean zero, which allows the offline and online samples to be treated as a single sequence.

Throughout this section the pooled sequence is ordered offline-first: index $\iota \leqslant N_{\text{off}}$ refers to offline trajectory ι, and $\iota = N_{\text{off}} + \tau$ to online episode τ. For $k \in [K]$ we write

$$\mathcal{G}_{k,\iota-1} := \sigma\Big(\big\{(\bar{s}_K^{j}, \bar{a}_K^{j}, Y^{j})\big\}_{j<\iota} \bigcup \big\{(\bar{s}_k^{\iota}, \bar{a}_k^{\iota})\big\}\Big), \tag{234}$$

the σ-algebra generated by all completed trajectories preceding ι in the pooled sequence, together with the current trajectory's history through stage k. In particular $\psi_k^\iota = \psi_k(\bar{s}_k^\iota, \bar{a}_k^\iota)$ is $\mathcal{G}_{k,\iota-1}$-measurable.

(1). Offline indices $\iota \leqslant N_{\text{off}}$. For the offline trajectory ι, and Assumption 2 holds with $\Gamma = 1$.

$$\begin{aligned}
&\mathbb{E}\Big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \,\Big|\, \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\Big] \\
&\quad = \mathbb{E}\big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) \mid \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\big] - \mathbb{E}\big[\big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \mid \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\big] \\
&\quad = \mathbb{E}\big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) \mid \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\big] - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \qquad (235) \\
&\quad = \mathbb{E}\big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) \mid \bar{s}_k^\iota, \bar{a}_k^\iota\big] - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \qquad (236) \\
&\quad = \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \qquad (237) \\
&\quad = 0.
\end{aligned}$$

Equation 235 is because ψ_k^ι is $\mathcal{G}_{k,\iota-1}$-measurable and $\theta_{V_{k+1}}$ is deterministic, so $\langle \psi_k^\iota, \theta_{V_{k+1}} \rangle$ equals its own conditional expectation; Equation 236 is because the offline trajectories are i.i.d., so S_{k+1}^ι is independent of $\mathcal{G}_{k,\iota-1}$ given $(\bar{s}_k^\iota, \bar{a}_k^\iota)$; and Equation 237 follows by Lemma 6(b).

At stage K, the same argument with Lemma 6(a) gives $\mathbb{E}[Y^\iota \mid \mathcal{G}_{K,\iota-1}, \bar{s}_K^\iota, \bar{a}_K^\iota] = \langle \psi_K^\iota, \theta_K^* \rangle$.

(2). Online indices $\iota = N_{\text{off}} + \tau$. At an online index, every action in $\bar{a}_k^\iota$ is computed by the algorithm from previous episodes' data and the current partial trajectory, so $\bar{a}_k^\iota$ is measurable with respect to $\sigma(\mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota)$. We have,

$$\begin{aligned}
&\mathbb{E}\Big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \,\Big|\, \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\Big] \\
&\quad = \mathbb{E}\big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) \mid \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\big] - E\Big[\big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \,\Big|\, \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\Big] \\
&\quad = \mathbb{E}\big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) \mid \mathcal{G}_{k,\iota-1}, \bar{s}_k^\iota, \bar{a}_k^\iota\big] - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \qquad (238) \\
&\quad = \mathbb{E}\big[V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota) \mid \bar{s}_k^\iota, \bar{a}_k^\iota\big] - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \qquad (239) \\
&\quad = \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle - \big\langle \psi_k^\iota, \theta_{V_{k+1}} \big\rangle \;=\; 0. \qquad (240)
\end{aligned}$$

Equation 238 is because $\psi_k(\bar{s}_k^\iota, \bar{a}_k^\iota)$ is $\mathcal{G}_{k,\iota-1}$-measurable and equals its own conditional expectation; Equation Equation 239 holds because $S_{k+1}^\iota \mid \mathcal{G}_{k,\iota-1} \sim P_{\bar{a}_k^\iota}(\cdot \mid \bar{s}_k^\iota)$, that is, the conditional law of the next state depends on $\mathcal{G}_{k,\iota-1}$ only through the current history $(\bar{s}_k^\iota, \bar{a}_k^\iota)$; and Equation 240 follows by Assumption 1(ii) applied to $V_{k+1} \in \mathcal{V}_{k+1}$.

At the terminal stage, the same argument gives $\mathbb{E}[Y^\iota \mid \mathcal{G}_{K,\iota-1}, \bar{s}_K^\iota, \bar{a}_K^\iota] = \langle \psi_K^\iota, \theta_K^* \rangle$.

(3). Boundedness. Both $V_{k+1}(\bar{s}_k^\iota, S_{k+1}^\iota, \bar{a}_k^\iota)$ and $\langle \psi_k^\iota, \theta_{V_{k+1}} \rangle$ take values in $[0,1]$: the first since $V_{k+1} \in [0,1]$, and the second since $\langle \psi_k^\iota, \theta_{V_{k+1}} \rangle = \mathbb{E}_{\bar{a}_k^\iota}\big[V_{k+1} \mid \bar{s}_k^\iota\big]$ by Assumption 1(ii), an average of values in $[0,1]$. The same argument applies at stage K because $Y^\iota \in [0,1]$. ■

Lemma 8. *Suppose Assumption 1 holds and $N_{\text{off}} \geqslant 8\gamma_{\text{off}} \log\big(K d_K/\delta_1\big)$. For $k \in [K]$, write*

$$\Lambda_k^{\text{off}} \;=\; \sum_{n=1}^{N_{\text{off}}} \psi_k\big(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}\big)\, \psi_k\big(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}\big)^\top.$$

Let Σ_k be the stage-k feature second moment of Definition 4. Define the event $\mathcal{E}_{\text{off}}^{\text{cov}} := \big\{\Lambda_k^{\text{off}} \succeq \frac{N_{\text{off}}}{2\gamma_{\text{off}}} I \text{ for all } k \in [K]\big\}$. Then $P(\mathcal{E}_{\text{off}}^{\text{cov}}) \geqslant 1 - \delta_1$, and on $\mathcal{E}_{\text{off}}^{\text{cov}}$, for all $k \in [K]$, $t \in [T]$, and $(\bar{s}_k, \bar{a}_k)$,

$$\|\psi_k(\bar{s}_k, \bar{a}_k)\|_{(\widetilde{\Lambda}_k^t)^{-1}} \;\leqslant\; \sqrt{\tfrac{2\gamma_{\text{off}}}{N_{\text{off}}}}.$$

Proof. This lemma shows that the offline sample alone already covers every feature direction, so that $\lambda_{\min}(\widetilde{\Lambda}_k^t) \geqslant \lambda + \frac{N_{\text{off}}}{2\gamma_{\text{off}}}$ at every episode $t \in [T]$, including $t = 1$, before any online data have been collected.

Since $\Lambda_k^{\text{off}} = \sum_{n=1}^{N_{\text{off}}} \psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}) \psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)})^\top$ is the offline Gram matrix at stage k, then the design matrix $\widetilde{\Lambda}_k^t$ of Algorithm 2 is the realization of $\lambda I + \Lambda_k^{\text{off}} + \sum_{\tau=1}^{t-1} \psi_k^\tau (\psi_k^\tau)^\top$ at the observed trajectories.

Fix a stage $k \in [K]$. The randomness in Λ_k^{off} enters through the state–action pairs $(\bar{S}_k^{(n)}, \bar{A}_k^{(n)})$ visited at stage k by the offline trajectories, drawn under the environment together with the behavior policy π_b. Since the offline trajectories are i.i.d., the summands are i.i.d.; each is a rank-one outer product, hence Λ_k^{off} is positive semidefinite. By Assumption 1, we have,

$$\left\| \psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}) \psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)})^\top \right\|_{\text{op}} = \left\| \psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}) \right\|_2^2 \leqslant 1. \tag{241}$$

Then, we have,

$$\begin{aligned}
\mathbb{E}[\Lambda_k^{\text{off}}] &= \sum_{n=1}^{N_{\text{off}}} \mathbb{E}\Big[\psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}) \, \psi_k(\bar{S}_k^{(n)}, \bar{A}_k^{(n)})^\top \Big] && (242) \\
&= N_{\text{off}} \sum_{(\bar{s}_k, \bar{a}_k) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_k} P(\bar{s}_k, \bar{a}_k) \, \psi_k(\bar{s}_k, \bar{a}_k) \, \psi_k(\bar{s}_k, \bar{a}_k)^\top && (243) \\
&= N_{\text{off}} \, \Sigma_k. && (244)
\end{aligned}$$

Here Equation 242 follows by the linearity of expectation; Equation 243 is because the N_{off} trajectories are i.i.d. draws from the observational distribution, so $(\bar{S}_k^{(n)}, \bar{A}_k^{(n)}) \sim P(\cdot, \cdot)$ for every n and all summands share the same expectation; and Equation 244 by the definition of Σ_k.

By Definition 4 we have $\lambda_{\min}(\Sigma_k) \geqslant 1/\gamma_{\text{off}}$, hence $\lambda_{\min}\big(\mathbb{E}[\Lambda_k^{\text{off}}]\big) \geqslant N_{\text{off}}/\gamma_{\text{off}}$ by Equation 244.

Therefore,

$$\begin{aligned}
P\Big(\lambda_{\min}\big(\Lambda_k^{\text{off}}\big) \leqslant \tfrac{N_{\text{off}}}{2\gamma_{\text{off}}}\Big) &\leqslant P\Big(\lambda_{\min}\big(\Lambda_k^{\text{off}}\big) \leqslant \tfrac{\lambda_{\min}\left(\mathbb{E}[\Lambda_k^{\text{off}}]\right)}{2}\Big) && (245) \\
&\leqslant d_k \, e^{-\lambda_{\min}\left(\mathbb{E}[\Lambda_k^{\text{off}}]\right)/8} && (246) \\
&\leqslant d_k e^{-N_{\text{off}}/(8\gamma_{\text{off}})}. && (247)
\end{aligned}$$

Here Equation 245 is because $N_{\text{off}}/(2\gamma_{\text{off}}) \leqslant \lambda_{\min}(\mathbb{E}[\Lambda_k^{\text{off}}])/2$; Equation 246 holds from Equation 255 by the matrix Chernoff bound (Lemma 16); Equation 247 follows by Equation 244.

We want the coverage event to hold at all K stages simultaneously with probability at least $1 - \delta_1$, so it suffices that each stage fails with probability at most δ_1/K. By the assumed burn-in condition, $N_{\text{off}}/(8\gamma_{\text{off}}) \geqslant \log(Kd_K/\delta_1)$, so $d_k e^{-N_{\text{off}}/(8\gamma_{\text{off}})} \leqslant d_k \cdot \delta_1/(Kd_K) \leqslant \delta_1/K$. A union bound over $k \in [K]$ gives $P(\mathcal{E}_{\text{off}}^{\text{cov}}) \geqslant 1 - \delta_1$.

Then, fix one episode $t \in [T]$. Every online term $\psi_k^\tau (\psi_k^\tau)^\top$ is positive semidefinite, so on $\mathcal{E}_{\text{off}}^{\text{cov}}$,

$$\widetilde{\Lambda}_k^t \;\geq\; \lambda I + \Lambda_k^{\text{off}} \;\geq\; \Big(\lambda + \tfrac{N_{\text{off}}}{2\gamma_{\text{off}}}\Big) I, \tag{248}$$

where Equation 248 follows by $\lambda_{\min}(\Lambda_k^{\text{off}}) \geqslant N_{\text{off}}/(2\gamma_{\text{off}})$.

Finally, for any $(\bar{s}_k, \bar{a}_k)$, we have,

$$\begin{aligned}
\|\psi_k(\bar{s}_k, \bar{a}_k)\|^2_{(\widetilde{\Lambda}_k^t)^{-1}} &\leqslant \|\psi_k(\bar{s}_k, \bar{a}_k)\|^2_{(\Lambda_k^{\text{off}} + \lambda I)^{-1}} && (249) \\
&\leqslant \frac{\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2^2}{\lambda + N_{\text{off}}/(2\gamma_{\text{off}})} && (250) \\
&\leqslant \frac{2\gamma_{\text{off}}}{N_{\text{off}}}. && (251)
\end{aligned}$$

Here Equation 249 is because $(\Lambda_k^{\text{off}} + \lambda I)^{-1} \succeq (\widetilde{\Lambda}_k^t)^{-1} \succ 0$; Equation 250 follows by Equation 248; Equation 251 holds since $\|\psi_k(\bar{s}_k, \bar{a}_k)\|_2 \leqslant 1$ and $\lambda > 0$.

Taking square roots gives the result. ■

Lemma 9. *Suppose Assumptions 1 and 2 with $\Gamma = 1$ hold, and $N_{\text{off}} \leqslant T^{d_K}$. Let $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}}$ be as in Definition 5 and $\widetilde{\mathcal{E}}_1$ as in Equation 38. Then:*

(i) For all $k \in [K]$ and $t \in [T]$, $\|\widetilde{\theta}_k^t\|_2 \leqslant \sqrt{d_k(N_{\text{off}}+t)/\lambda}$ and $\|\widehat{\theta}_k^t\|_2 \leqslant \sqrt{d_k t/\lambda}$. Hence, for $k < K$, $\widetilde{V}_{k+1}^t \in \widetilde{\mathcal{V}}_{k+1}^{\text{pool}}$ with $(\omega_{k+1}, \beta, \Lambda) = (\widehat{\theta}_{k+1}^t, \beta_{k+1}, \Lambda_{k+1}^t)$ and $(\widetilde{\omega}_{k+1}, \widetilde{\beta}, \widetilde{\Lambda}) = (\widetilde{\theta}_{k+1}^t, \widetilde{\beta}_{k+1}, \widetilde{\Lambda}_{k+1}^t)$.

(ii) $P(\widetilde{\mathcal{E}}_1) \geqslant 1-\delta$.

(iii) On $\widetilde{\mathcal{E}}_1$, with $\widetilde{V}_{k+1}^t$ in place of $\widehat{V}_{k+1}^t$, the bound of Proposition 5 holds for the online estimate $\widehat{\theta}_k^t$, and, with $(\widetilde{\mathcal{C}}, \widetilde{\Lambda}_k^t)$ in place of $(\mathcal{C}, \Lambda_k^t)$, for the pooled estimate $\widetilde{\theta}_k^t$.

(iv) On $\widetilde{\mathcal{E}}_1$, for all $t \in [T]$: $\theta_K^* \in \widetilde{\mathcal{C}}_K^t \cap \widehat{\mathcal{C}}_K^t$, and $\theta_{\widetilde{V}_{k+1}^t} \in \widetilde{\mathcal{C}}_k^t \cap \widehat{\mathcal{C}}_k^t$ for every $k < K$.

Proof. The four guarantees are the results of Theorem 1 applied to the pooled regression Equation 16, whose sums run over $\iota \leqslant N_{\text{off}} + t - 1$ rather than $\tau \leqslant t-1$.

Three things change relative to that analysis.

- The effective sample size becomes $N_{\text{off}} + T$: the log factor $\chi \to \widetilde{\chi}$, and the discretization level $\epsilon = d_K/(N_{\text{off}}+T)$ update accordingly, which gives the radius $Cd_K\sqrt{\widetilde{\chi}}$ for the pooled sums. The online-only components keep the radius $Cd_K\sqrt{\chi}$ since $N_{\text{off}} \leqslant T^{d_K}$.
- The value-function class becomes $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}}$ in place of the class $\widetilde{\mathcal{V}}_{k+1}$ of Theorem 1. Since $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}} \subseteq \mathcal{V}_{k+1}$ and $\widetilde{V}_{k+1}^t$ belongs to it, Assumption 1(ii) supplies $\theta_{\widetilde{V}_{k+1}^t}$ and the uniform bound may be instantiated at $V = \widetilde{V}_{k+1}^t$.
- The class $\widetilde{\mathcal{V}}_{k+1}^{\text{pool}}$ carries two bonus terms rather than one, so it is parametrized by the two triples $(\omega_{k+1}, \beta, \Lambda)$ and $(\widetilde{\omega}_{k+1}, \widetilde{\beta}, \widetilde{\Lambda})$. Lemma 15 with $m = 2$, $g = \max_{\theta \in \widehat{C}_{k+1,\text{off}}^{\Gamma=1}} \theta^\top \psi_{k+1}$, $L = \sqrt{d_{k+1}(N_{\text{off}}+T)/\lambda}$ and $B = \widetilde{B}$ bounds its covering number. The function g is fixed given the independent offline sample that constructs it and adds no parameters, so it leaves the covering number unchanged.

Concretely, the pooled statements in (i)–(iv) are Steps 1–2 of Theorem 1's proof with $t-1$ replaced by $N_{\text{off}}+t-1$ and χ by $\widetilde{\chi}$. Parts (iii)–(iv) then follow from the proofs of Proposition 5 and Lemma 2 under the respective substitutions, using Lemma 7 for the offline indices. ■

Lemma 10. *Define the good event*

$$\widetilde{\mathcal{E}} := \widetilde{\mathcal{E}}_1 \bigcap \mathcal{E}_{\text{off}}^{\text{cov}} \bigcap \mathcal{E}^{\text{off}}.$$

Then $P(\widetilde{\mathcal{E}}) \geqslant 1 - \delta - 2\delta_1$ and on $\widetilde{\mathcal{E}}$,

$$\widetilde{Q}_k^t(\bar{s}_k, \bar{a}_k) \;\geqslant\; Q_k^*(\bar{s}_k, \bar{a}_k) \qquad \text{for all } k \in [K],\ t \in [T],\ (\bar{s}_k, \bar{a}_k).$$

Proof. This lemma establishes optimism: on the good event, the algorithm's estimate $\widetilde{Q}_k^t$ is never less than Q_k^* at any stage, episode, or history. This is what converts regret into estimation error.

By Lemma 9(ii), $P(\widetilde{\mathcal{E}}_1) \geqslant 1-\delta$; by Lemma 8, $P(\mathcal{E}_{\text{off}}^{\text{cov}}) \geqslant 1-\delta_1$; and by Proposition 3, $P(\mathcal{E}^{\text{off}}) \geqslant 1-\delta_1$. A union bound gives $P(\widetilde{\mathcal{E}}) \geqslant 1-\delta-2\delta_1$.

We now prove the optimism. For any epsiode $t \in [T]$ and work on the event $\widetilde{\mathcal{E}}$. As in the proof of Lemma 4, we prove the optimism for all $k \in [K]$ by backward induction on k. Recall from Equation 16

$$\widetilde{Q}_k^t(\bar{s}_k, \bar{a}_k) = \max\Big\{ \min\Big\{ \max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ \max_{\theta \in \widetilde{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ \max_{\theta \in \widehat{C}_{k,\text{off}}^{\Gamma=1}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k),\ 1 \Big\},\ 0 \Big\}.$$

Since $\max\{x, 0\} \geqslant x$, it suffices to show that each of the four arguments of the inner minimum is at least $Q_k^*(\bar{s}_k, \bar{a}_k)$.

(i) The constant term. Since outcomes lie in $[0, 1]$, we have $Q_k^*(\bar{s}_k, \bar{a}_k) \leqslant 1$ at every stage.

(ii) Base case $k = K$. Here $Q_K^*(\bar{s}_K, \bar{a}_K) = \langle \theta_K^*, \psi_K(\bar{s}_K, \bar{a}_K) \rangle$, and $\theta_K^* \in \widetilde{\mathcal{C}}_K^t \bigcap \widehat{\mathcal{C}}_K^t$ by Lemma 9(iv), so

$$\max_{\theta \in \widehat{\mathcal{C}}_K^t} \theta^\top \psi_K(\bar{s}_K, \bar{a}_K) \geqslant (\theta_K^*)^\top \psi_K(\bar{s}_K, \bar{a}_K) = Q_K^*(\bar{s}_K, \bar{a}_K), \tag{252}$$

and the same holds with $\widetilde{\mathcal{C}}_K^t$ in place of $\widehat{\mathcal{C}}_K^t$. By Proposition 3 at $\Gamma = 1$, $\theta_K^* \in \widehat{C}_{K,\text{off}}^{\Gamma=1}$ on $\mathcal{E}^{\text{off}}$, so the same holds with $\widehat{C}_{K,\text{off}}^{\Gamma=1}$ as well. Together with (i), all four arguments of the minimum dominate Q_K^*.

(iii) Inductive step $k < K$. The induction hypothesis is $\widetilde{Q}_{k+1}^t(\bar{s}_{k+1}, \bar{a}_{k+1}) \geqslant Q_{k+1}^*(\bar{s}_{k+1}, \bar{a}_{k+1})$ for stage $k + 1$, then

$$\widetilde{V}_{k+1}^t(\bar{s}_{k+1}, \bar{a}_k) = \max_{a_{k+1}} \widetilde{Q}_{k+1}^t(\bar{s}_{k+1}, \bar{a}_{k+1}) \geqslant \max_{a_{k+1}} Q_{k+1}^*(\bar{s}_{k+1}, \bar{a}_{k+1}) = V_{k+1}^*(\bar{s}_{k+1}, \bar{a}_k). \tag{253}$$

Recall $Q_k^*(\bar{s}_k, \bar{a}_k) = \mathbb{E}_{\bar{a}_k}[V_{k+1}^*(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k]$, and that $\langle \theta_{\widetilde{V}_{k+1}^t}, \psi_k \rangle = \mathbb{E}_{\bar{a}_k}[\widetilde{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k]$ by Assumption 1(ii), applied to $\widetilde{V}_{k+1}^t \in \widetilde{\mathcal{V}}_{k+1}^{\text{pool}} \subseteq \mathcal{V}_{k+1}$ (Lemma 9). Since $\theta_{\widetilde{V}_{k+1}^t} \in \widetilde{\mathcal{C}}_k^t \bigcap \widehat{\mathcal{C}}_k^t$ by Lemma 9(iv), we have,

$$\max_{\theta \in \widehat{\mathcal{C}}_k^t} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k) \geqslant \mathbb{E}_{\bar{a}_k}[\widetilde{V}_{k+1}^t(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k] \geqslant \mathbb{E}_{\bar{a}_k}[V_{k+1}^*(\bar{S}_{k+1}, \bar{a}_k) \mid \bar{s}_k] = Q_k^*(\bar{s}_k, \bar{a}_k), \tag{254}$$

where the second inequality follows by Equation 253; the same chain holds with $\widetilde{\mathcal{C}}_k^t$ in place of $\widehat{\mathcal{C}}_k^t$, since $\theta_{\widetilde{V}_{k+1}^t}$ lies in both. For the third branch, Proposition 3 at $\Gamma = 1$ gives $\theta_k^* \in \widehat{C}_{k,\text{off}}^{\Gamma=1}$ on $\mathcal{E}^{\text{off}}$, so $\max_{\theta \in \widehat{C}_{k,\text{off}}^{\Gamma=1}} \theta^\top \psi_k(\bar{s}_k, \bar{a}_k) \geqslant (\theta_k^*)^\top \psi_k(\bar{s}_k, \bar{a}_k) = Q_k^*(\bar{s}_k, \bar{a}_k)$, without appeal to the induction hypothesis. Together with (i), all four arguments of the minimum dominate Q_k^*, which completes the induction.

Then by (i), (ii), (iii), $\widetilde{Q}_k^t(\bar{s}_k, \bar{a}_k) \geqslant Q_k^*(\bar{s}_k, \bar{a}_k)$ for all $k \in [K]$ and $t \in [T]$. ■

G Auxiliary Results

Lemma 11 (Hoeffding's Inequality). *Let $X_1, ..., X_n$ be independent random variables on $\mathbb{R}$ such that X_i is bounded in the interval $[a_i, b_i]$. Let $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\mathbb{P}\Big(S_n - \mathbb{E}[S_n] \geqslant t\Big) \leqslant \exp\Big(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

$$\mathbb{P}\Big(S_n - \mathbb{E}[S_n] \leqslant -t\Big) \leqslant \exp\Big(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

Lemma 12 (Azuma-Hoeffding Inequality). *Let $X_0, X_1, \ldots, X_n$ be a martingale sequence of random variables such that for all i, there exists a constant c_i such that $|X_i - X_{i-1}| < c_i$, then*

$$\mathbb{P}\big(X_n - X_0 \geqslant t\big) \leqslant \exp\Big(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\Big)$$

$$\mathbb{P}\big(X_n - X_0 \leqslant -t\big) \leqslant \exp\Big(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\Big).$$

Lemma 13 (Elliptical Potential Lemma). *Let $\{u_\tau\}_{\tau=1}^\infty$ be a sequence in $\mathbb{R}^n$ such that $||u_\tau|| \leqslant B$ for all $\tau \in \mathbb{N}$. Further let $\{\lambda_\tau\}_{\tau=1}^\infty$ be an non-decreasing sequence in $\mathbb{R}^+$ s.t. $\lambda_1 = 1$. For $t \geqslant 1$ define $V_t = \sum_{\tau=1}^{t-1} u_\tau u_\tau^\intercal + \lambda_t I_n$. Then:*

$$\sum_{t=1}^T ||u_t||_{V_t^{-1}}^2 \leqslant 2n(1 + B^2) \log\Big(\lambda_T + \frac{TB^2}{n}\Big).$$

Lemma 14 (Adapted from Jin et al., 2020b, Lemma D.4). *Fix any stage $k \in [K-1]$. Let $\{\mathcal{F}_{k,\tau}\}_{\tau \geqslant 0}$ be the corresponding filtration such that $\psi_k^\tau := \psi_k(\bar{s}_k^\tau, \bar{a}_k^\tau) \in \mathbb{R}^{d_k}$ is $\mathcal{F}_{k,\tau-1}$-measurable, and write $\Lambda_k^t = \lambda I + \sum_{\tau=1}^{t-1} \psi_k^\tau (\psi_k^\tau)^\top$. Let $\mathcal{U}_{k+1}$ be any class of functions $V : \overline{\mathcal{S}}_{k+1} \times \overline{\mathcal{A}}_k \to \mathbb{R}$ satisfying*

$$\sup_{(\bar{s}_{k+1}, \bar{a}_k) \in \overline{\mathcal{S}}_{k+1} \times \overline{\mathcal{A}}_k} |V(\bar{s}_{k+1}, \bar{a}_k)| \leqslant 1, \qquad \forall V \in \mathcal{U}_{k+1}.$$

Then for any $\delta > 0$, with probability at least $1-\delta$, simultaneously for all $t \geqslant 1$ and all $V \in \mathcal{U}_{k+1}$, we have

$$\Big\| \sum_{\tau=1}^{t-1} \psi_k^\tau \left\{ V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) - \mathbb{E}\left[V(\bar{s}_{k+1}^\tau, \bar{a}_k^\tau) \mid \mathcal{F}_{k,\tau-1} \right] \right\} \Big\|^2_{(\Lambda_k^t)^{-1}} \leqslant 4\Big[\frac{d_k}{2} \log\left(\frac{t+\lambda}{\lambda} \right) + \log\left(\frac{\mathcal{N}_{k+1,\epsilon}}{\delta} \right) \Big] + \frac{8t^2\epsilon^2}{\lambda},$$

where $\mathcal{N}_{k+1,\epsilon}$ is the ϵ-covering number of $\mathcal{U}_{k+1}$ with respect to the distance

$$\mathrm{dist}(V, V') := \sup_{(\bar{s}_{k+1}, \bar{a}_k) \in \overline{\mathcal{S}}_{k+1} \times \overline{\mathcal{A}}_k} |V(\bar{s}_{k+1}, \bar{a}_k) - V'(\bar{s}_{k+1}, \bar{a}_k)|.$$

Lemma 15 (Adapted from Jin et al., 2020b, Lemma D.6). *Fix a stage $k < K$ and an integer $m \geqslant 1$. Let $g : \overline{\mathcal{S}}_{k+1} \times \overline{\mathcal{A}}_{k+1} \to [0,1]$ be a fixed function, and let $\mathcal{U}_{k+1}$ denote the class of functions mapping $\overline{\mathcal{S}}_{k+1} \times \overline{\mathcal{A}}_k$ to $\mathbb{R}$ of the parametric form*

$$\mathcal{U}_{k+1} := \Big\{ V : V(\bar{s}_{k+1}, \bar{a}_k) = \max\Big\{ \max_{a_{k+1} \in \mathcal{A}_{k+1}} \min\Big\{ 1,\ g(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}), \min_{i \in [m]} \Big(\big\langle \omega^{(i)},\ \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \big\rangle + \beta^{(i)} \big\| \psi_{k+1}(\bar{s}_{k+1}, \bar{a}_k, a_{k+1}) \big\|_{(\Lambda^{(i)})^{-1}} \Big) \Big\},\ 0 \Big\} \Big\},$$

where the parameters satisfy $\|\omega^{(i)}\|_2 \leqslant L$, $\beta^{(i)} \in [0, B]$ and $\lambda_{\min}(\Lambda^{(i)}) \geqslant \lambda$ for every $i \in [m]$. Let $\mathcal{N}_{k+1,\epsilon}$ be the ϵ-covering number of $\mathcal{U}_{k+1}$ under $\mathrm{dist}(V, V') = \sup |V - V'|$. Then

$$\log \mathcal{N}_{k+1,\epsilon} \ \leqslant\ m \left[d_{k+1} \log\Big(1 + \tfrac{4L}{\epsilon} \Big) \ +\ d_{k+1}^2 \log\Big(1 + \tfrac{8 d_{k+1}^{1/2} B^2}{\lambda \epsilon^2} \Big) \right].$$

Lemma 16 (Matrix Chernoff (Tropp, 2015, Theorem 5.1.1)). *Consider a finite sequence $\{Z_k\}$ of independent, random, Hermitian matrices with common dimension n. Assume that*

$$0 \leqslant \lambda_{\min}(Z_k) \qquad \text{and} \qquad \lambda_{\max}(Z_k) \leqslant L$$

for each index k.

Introduce the random matrix $Y = \sum_k Z_k$. Define the minimum eigenvalue $\lambda_{\min}$ and maximum eigenvalue $\lambda_{\max}$ of the expectation $\mathbb{E}[Y]$:

$$\lambda_{\min} = \lambda_{\min}(\mathbb{E}[Y]) = \lambda_{\min}\left(\sum_k \mathbb{E}[Z_k] \right), \quad \lambda_{\max} = \lambda_{\max}(\mathbb{E}[Y]) = \lambda_{\max}\left(\sum_k \mathbb{E}[Z_k] \right).$$

Then, for $r > 0$,

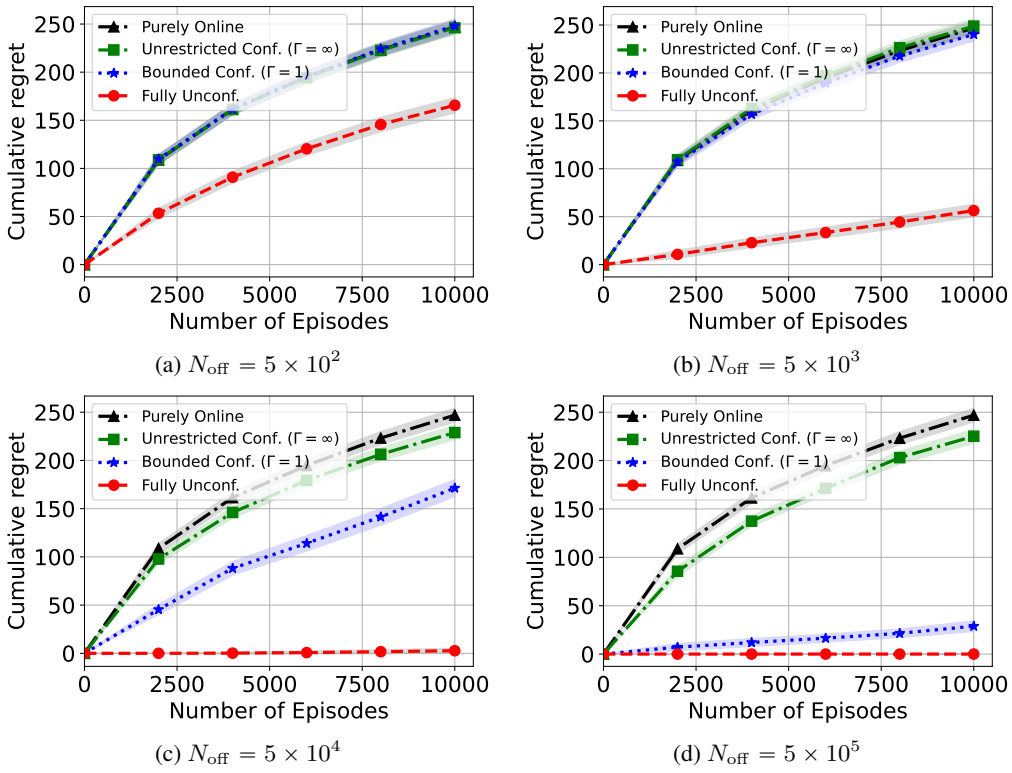
$$\mathbb{E}[\lambda_{\min}(Y)] \geqslant \frac{1 - e^{-r}}{r} \lambda_{\min} - \frac{1}{r} L \log n, \quad \mathbb{E}[\lambda_{\max}(Y)] \leqslant \frac{e^r - 1}{r} \lambda_{\max} + \frac{1}{r} L \log n.$$

Furthermore,

$$\mathbb{P}\{\lambda_{\min}(Y) \leqslant (1-\varepsilon)\lambda_{\min}\} \leqslant n \left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}} \right]^{\lambda_{\min}/L}, \qquad \text{for } \varepsilon \in [0,1), \tag{255}$$

$$\mathbb{P}\{\lambda_{\max}(Y) \geqslant (1+\varepsilon)\lambda_{\max}\} \leqslant n \left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \right]^{\lambda_{\max}/L}, \qquad \text{for } \varepsilon \geqslant 0. \tag{256}$$

H NUMERICAL EXPERIMENTS



(a) $N_{\text{off}} = 5 \times 10^2$ (b) $N_{\text{off}} = 5 \times 10^3$

(c) $N_{\text{off}} = 5 \times 10^4$ (d) $N_{\text{off}} = 5 \times 10^5$

Figure 2: Cumulative regret for purely online (Algorithm 3); unrestricted confounding ($\Gamma = \infty$) (Algorithm 1); bounded confounding ($\Gamma = 1$) (Algorithm 4); and fully unconfounded (Algorithm 2), at offline sample sizes $N_{\text{off}} = 5 \times 10^2$, 5×10^3, 5×10^4, 5×10^5.

We evaluate the empirical performance of the proposed methods on a synthetic dataset, described below.

Data set description. Following the random DTRs experiments of (Zhang & Bareinboim, 2019), we use $K = 2$ stages with binary states $S_k \in \{0, 1\}$, binary treatments $A_k \in \{0, 1\}$, and a binary outcome $Y \in \{0, 1\}$; we add a binary confounder $U \in \{0, 1\}$ with $P(U = 1) = \frac{1}{2}$. Writing $\sigma(\cdot)$ for the logistic function, we generated the synthetic dataset from

$$P(S_1 = 1 \mid U = u) = \sigma\big(b_{s_1} + 0.5\,(2u - 1)\big),$$
$$\pi_b(A_k = 1 \mid \bar{s}_k, \bar{a}_{k-1}, U = u) = \sigma\big(b_k^{\text{beh}}(\bar{s}_k, \bar{a}_{k-1}) + \kappa\,(2u - 1)\big),$$
$$P(S_{k+1} = 1 \mid \bar{s}_k, \bar{a}_k, U = u) = \sigma\big(b_k^{\text{tr}}(\bar{s}_k, \bar{a}_k) + 0.8\,(2u - 1)\big),$$
$$P(Y = 1 \mid \bar{s}_K, \bar{a}_K, U = u) = \sigma\big(b^{y}(\bar{s}_K, \bar{a}_K) + 1.2\,(2u - 1)\big).$$

The baseline logits are drawn from $b_{s_1} \sim \mathcal{N}(0, 1)$, $b_k^{\text{beh}} \sim \mathcal{N}(0, 0.3^2)$, $b_k^{\text{tr}} \sim \mathcal{N}(0, 1)$ and $b^{y} \sim \mathcal{N}(0, 1.5^2)$ per environment.

We set $\kappa = 0$, so that $\Gamma_{\text{true}} = 1$: the offline data used by all four algorithms have no unmeasured confounding.

Algorithm. We evaluate four different algorithms: purely online (Algorithm 3); unrestricted confounding (Algorithm 1); bounded confounding (Algorithm 4); fully unconfounded (Algorithm 2). To our knowledge, no public code is available for Zhang & Bareinboim (2019; 2020).

Metric. We evaluated the algorithms using cumulative regret. Since the synthetic model is known and finite, $V_1^{\star}$ and $V_1^{\boldsymbol{\pi}_t}$ are computed exactly by backward induction under the true interventional kernel.

Implementation details. Each algorithm is run over 8 environments with 10 runs per environment. We set $\lambda = 1$ and $\delta = \delta_1 = 0.05$ throughout. The feature vector $\psi_k(\bar{s}_k, \bar{a}_k)$ is the one-hot indicator of the history–action pair, so $d_1 = 4$ and $d_2 = 16$. We report results for $N_{\text{off}} = 5 \times 10^2$, 5×10^3, 5×10^4, 5×10^5 and episodes $T \in \{2000, 4000, 6000, 8000, 10000\}$. We report the mean and standard error of the regret.

Results. Figure 2 shows cumulative regret for the four algorithms at each offline sample size $N_{\text{off}} = 5 \times 10^2$, 5×10^3, 5×10^4, 5×10^5 over episodes $T \in \{2000, 4000, 6000, 8000, 10000\}$. The purely online curve is unchanged across the four figures. It therefore serves as a common reference for the three hybrid algorithms: unrestricted confounding, bounded confounding, and fully unconfounded. Three observations follow:

- First, across all four offline sample sizes, none of the three hybrid algorithms incurs more regret than the purely online algorithm empirically. For small N_{off}, the algorithms for unrestricted and bounded confounding perform nearly identically. (see Figure 2a and 2b).
- Second, the regret is ordered as: fully unconfounded $\leqslant$ bounded confounding $\leqslant$ unrestricted confounding in every figure. Bounded confounding and unrestricted confounding are close at $N_{\text{off}} = 5 \times 10^2$ and separate as N_{off} grows.
- Third, the regret of the three hybrid algorithms decreases as N_{off} increases.