Reinforcement Learning with Action-Triggered Observations

Alexander Ryabchenko * Wenlong Mou *

Abstract

We study reinforcement learning problems where state observations are stochastically triggered by actions, a constraint common in many real-world applications. This framework is formulated as Action-Triggered Sporadically Traceable Markov Decision Processes (ATST-MDPs), where each action has a specified probability of triggering a state observation. We derive tailored Bellman optimality equations for this framework and introduce the action-sequence learning paradigm in which agents commit to executing a sequence of actions until the next observation arrives. Under the linear MDP assumption, value-functions are shown to admit linear representations in an induced action-sequence feature map. Leveraging this structure, we propose off-policy estimators with statistical error guarantees for such feature maps and introduce ST-LSVI-UCB, a variant of LSVI-UCB adapted for action-triggered settings. ST-LSVI-UCB achieves regret $\widetilde{O}(\sqrt{Kd^3(1-\gamma)^{-3}})$, where K is the number of episodes, d the feature dimension, and γ the discount factor (per-step episode non-termination probability). Crucially, this work establishes the theoretical foundation for learning with sporadic, action-triggered observations while demonstrating that efficient learning remains feasible under such observation constraints.

1 Introduction

Reinforcement Learning (RL) addresses sequential decision-making problems where an agent interacts with an unknown environment to maximize rewards. As the environment changes in response to the agent's actions, it is typically expected for the agent to receive immediate feedback. However, in many real-world scenarios, observations may be delayed, intermittently available, or costly to obtain. While Partially Observable Markov Decision Processes (POMDPs) [Ast65] offer a general framework for limited observability, they often lack specificity for scenarios where observation availability directly depends on agent's actions.

To close this gap, we propose a novel RL framework characterized by "action-triggered observations," where each action a has an associated probability $\beta(a) \in [0,1]$ of revealing the new state after execution. A policy must therefore simultaneously optimize actions in the absence of immediate state feedback and strategically decide when to trigger observations to reduce uncertainty. This process involves executing sequences of actions across multiple consecutive rounds without environmental feedback until a state observation occurs — an event we define as a "data-burst." We formalize this framework as Action-Triggered Sporadically Traceable Markov Decision Processes (ATST-MDPs). The main goal of this work is to develop theoretical foundations for optimal learning under this observation mechanism.

The ATST-MDP framework with data-bursts captures several actively studied observation mechanisms in RL, addressing practical information constraints of real-world environments:

1. **Active sensing:** [Sat+17; SR23; KSJ23]. Agents may employ specialized sensing actions with varying observation probabilities to reduce state uncertainty. For instance, in medical scenarios, practitioners

^{*}University of Toronto and Vector Institute.

pair treatment decisions with diagnostic tests of different invasiveness levels, which may themselves affect patient state — creating a trade-off between timely diagnosis and last-resort interventions.

- 2. **Paid observations:** [NFB21; Bel+20; Wan+25]. Actions may include explicit decisions to purchase feedback (receiving no observation otherwise), affecting rewards through additional costs. For example, in marketing operations, companies execute promotional campaigns and then choose whether to purchase detailed market penetration studies to assess campaign effectiveness and market response.
- 3. **Intermittent feedback:** [HS17; KTO18; CL25]. Data-bursts, occurring with fixed probability each round, may be guided by independent external events. This corresponds to scenarios with limited observability due to unreliable sensors or communication channels that only sporadically provide environmental data, e.g., an autonomous vehicles navigating through dense fog with intermittent visibility.

Motivated by practical considerations, our work focuses on the theoretical underpinnings: a precise formulation of RL with action-triggered observations, a structural analysis of optimal policies for when and how to trigger state observations to maximize rewards, and rigorous regret guarantees for episodic learning.

Related work. Our framework overlaps with several well-studied settings, yet none directly capture action-triggered observations. Although the absence of state feedback superficially resembles RL with observation delays [KE03; Wal+09; Lio23], the *delays* in ATST-MDPs are endogenous, induced by the agent's actions, whereas classical delays are exogenous. Goal-conditioned RL [Sch+15; And+18] provides observations only upon goal attainment (state-triggered feedback), which is orthogonal to our action-triggered mechanism. Many POMDP formulations [PGT03; SV10; CYW24] model belief updates under partial observability; however, existing work generally does not exploit the structure induced by action-triggered observations. A more detailed discussion of related work can be found in Appendix A.

Our contributions and paper organization:

- In Section 2, we formally introduce ATST-MDPs, derive the associated Bellman optimality equations, and introduce an action-sequence perspective via a novel action-sequence value-function.
- In Section 3, under the Linear MDP assumption, the action-sequence value-function is shown to admit a linear representation in an induced action-sequence feature map. We provide efficient off-policy estimation guarantees for this feature map in Subsection 3.1.
- In Section 4, we propose ST-LSVI-UCB, an algorithm for episodic learning with geometrically distributed horizon lengths in linear ATST-MDPs, achieving $\widetilde{O}(\sqrt{Kd^3(1-\gamma)^{-3}})$ regret with high probability, provided sufficiently accurate estimation of the action-sequence feature map. We stress that regret here is measured against the optimal policy operating under the same observation constraints, not the infeasible policy with full observability, comparison to which would generally lead to linear regret.

2 Problem Setting

We introduce classical RL concepts and notation in Subsection 2.1 and then define our ATST-MDP in Subsection 2.2 as a special MDP on the augmented state space. Analysis of its value-functions, including a novel action-sequence value-function, is presented in Subsection 2.3.

2.1 Preliminaries and Notation

Markov Decision Processes and Discounted Returns. A discrete-time discounted Markov Decision Process (MDP) is a 5-tuple $(S, \mathcal{A}, \mathbb{P}, r, \gamma)$, where S, \mathcal{A} are measurable state and action spaces respectively, $\mathbb{P}(. \mid s, a) \in \Delta_S$ defines the transition probability measure over the next states given current state s and taken action $a, r : S \times \mathcal{A} \to [0, 1]$ is a deterministic reward function, and $\gamma \in (0, 1)$ – discount factor. We assume \mathcal{A} is a finite set of cardinality A, whereas S may contain infinitely many elements.

The agent's objective is to maximize expected discounted returns. For a deterministic policy $\pi: \mathcal{S} \to \mathcal{A}$, the state-action value-function $Q^{\pi}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as the expected discounted return when starting from state s, executing action a, and following policy π thereafter:

$$Q^{\pi}(s,a) = \mathbb{E}_{s_{h+1} \sim \mathbb{P}(.|s_h,a_h), a_{h+1} = \pi(s_{h+1})} \left[\sum_{h=1}^{\infty} \gamma^{h-1} r(s_h, a_h) \middle| s_1 = s, a_1 = a \right].$$

The corresponding state value-function $V^{\pi}: \mathcal{S} \to \mathbb{R}$ is defined as $V^{\pi}(s) = Q^{\pi}(s, \pi(s))$. There exists an optimal policy π^* satisfying $V^*(s) := V^{\pi^*}(s) = \sup_{\pi} V^{\pi}(s)$ for every state s (e.g., see [Put94]).

Discounting via Geometric Horizon. The state-action value-function can be equivalently formulated using a geometric horizon interpretation by considering an episode of random length $H_{\gamma} \sim \text{Geom}(1-\gamma)$:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\left.\sum_{h=1}^{H_{\gamma}} r(s_h, a_h)\right| s_1 = s, a_1 = a\right].$$

Discounting factor γ serves as a fixed per-step episode non-termination probability (e.g., see [Man+23]).

Augmented State Space. In situations where state observations may be unavailable for several consecutive rounds, the augmented state space $\mathcal{X} = \mathcal{S} \times \mathcal{A}^{<\mathbb{N}}$, provides a natural framework for reasoning under uncertainty. Each augmented state $x = (s_1; a_1, \ldots, a_{\Delta})$ consists of the last observed state s_1 followed by a finite sequence of $\Delta \in \mathbb{N} \cup \{0\}$ actions taken since then, capturing the distribution of the current state $s_{\Delta+1}$. The belief function $b: \mathcal{X} \to \Delta_{\mathcal{S}}$ represents this distribution as $s_{\Delta+1} \sim b(.|x)$. For $\Delta \geqslant 1$, we have

$$b(s|x) = \int_{\mathcal{S}^{\Delta-1}} \mathbb{P}(s|s_{\Delta}, a_{\Delta}) \prod_{i=2}^{\Delta} \mathbb{P}(s_i|s_{i-1}, a_{i-1}) ds_i. \tag{1}$$

Augmented states are actively used in RL with delays (e.g., see [Bou+21]).

Linear MDPs. When modeling complex environments with potentially large or continuous state spaces, structural properties can be exploited to enable efficient learning. Linear MDPs represent a fundamental class of RL problems where both transition dynamics and reward functions exhibit linearity in a feature space. As is standard in the field, we define the following linear MDP structure:

Assumption 2.1 (Linear MDP). There exists a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$:

$$\mathbb{P}(\cdot|s,a) = \langle \phi(s,a), \mu(\cdot) \rangle, \quad r(s,a) = \langle \phi(s,a), \theta \rangle,$$

where $\mu: \mathcal{S} \to \mathbb{R}^d$ consists of d finite signed measures over \mathcal{S} and $\theta \in \mathbb{R}^d$. Additionally, we require $\sup_{s,a} \|\phi(s,a)\|_2 \leq 1$, $\|\theta\|_2 \leq \sqrt{d}$, and $\||\mu|(\mathcal{S})\|_2 \leq \sqrt{d}$.

In the linear MDP framework, the feature vectors $\phi(s, a)$ are known to the learner, while the vectors (μ, θ) are unknown. This framework is widely used in the study of RL with function approximations. As shown in [Jin+19], it encompasses standard RL settings including tabular MDPs and simplex feature spaces.

Notation. Let \oplus denote concatenation in a general sense, e.g. $(x,y)=x\oplus y$ and $(x,y)\oplus z=(x,y,z)$. Let $\delta_{ij}=\mathbb{I}(i=j)$. For $n\in\mathbb{N}$, let $[n]=\{1,...,n\}$. For vector $\boldsymbol{x}\in\mathbb{R}^D$, matrix $M\in\mathbb{R}^{D\times D}$, and $q\in[1,\infty]$, let $\|\boldsymbol{x}\|_q$ denote l_q -norm, $\|M\|_q-l_q$ to l_q operator norm, $\lambda_{\min}(M)$ – minimal eigenvalue of matrix M.

For convenience, for every symbol $\mathbf{z} \in \{\beta, \bar{\beta}, M\}$, we consider short-hand notation $\mathbf{z}_a = \mathbf{z}(a)$ for all $a \in \mathcal{A}$.

2.2 Introducing Action-Triggered Sporadically Traceable MDPs

The Action-Triggered Sporadically Traceable Markov Decision Process (ATST-MDP) extends the traditional MDP by incorporating action-dependent probabilities of data-bursts. We define an ATST-MDP as the 6-tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma, \beta)$, where $\beta: \mathcal{A} \to [0, 1]$ assigns to each action $a \in \mathcal{A}$ the probability that executing this action will trigger a data-burst. State observations occur through data-bursts, which are critical events in this framework: when and only when a data-burst occurs is the current state of the environment revealed to the agent. For convenience, let $\bar{\beta}(a) = 1 - \beta(a)$ denote the probability that a does not invoke a data-burst.

From the agent's perspective in an ATST-MDP, the interaction dynamics work as follows. At any point in time, the agent's knowledge is represented by an augmented state $x \in \mathcal{X}$, consisting of the last observed state and actions executed since this observation. The true environmental state is unknown to the agent, unless $x \in \mathcal{S}$. When the agent executes action $a \in \mathcal{A}$, one of two outcomes occurs: with probability $\beta(a)$, a data-burst is triggered, and the agent observes the actual current environmental state $s \sim b(\cdot|x \oplus a)$; with probability $\bar{\beta}(a)$, no data-burst occurs, and the agent updates its augmented state to $x \oplus a$. In general, function β need not be known to the agent beforehand, but it can be for certain applications, e.g., Example 2.3.

For effective learning, agents require access to reward information. Our model specifies that during a databurst, the agent receives the total accumulated reward. While some specialized applications might allow for complete trajectories of state-reward pairs to be revealed during data-bursts, our formulation addresses the more general case where such detailed history is unavailable, but outcomes are periodically measurable.

The following concrete examples illustrate the types of problems ATST-MDPs allow us to analyze.

Example 2.2 (Faulty communication channel). In scenarios where the state observations occur with fixed probability β^* every round (e.g., due to a faulty environment-to-agent communication channel), we can set $\beta(a) = \beta^*$ for all $a \in A$. In each round, with probability $1 - \beta^*$ the augmented state grows by one action, and with probability β^* we remove uncertainty by obtaining the current state in S.

Example 2.3 (Paid observations). Consider a Linear MDP (Assumption 2.1) where the agent has the option to observe the current state at a price. Each action a has two versions with identical transition dynamics: a_1 (triggers observation for a cost $c(s, a) \in [0, c_{\max}]$) and a_0 (no observation). We can model this by setting $\beta(a_i) = \delta_{i1}$ and extending feature, reward, and measure vectors to \mathbb{R}^{d+1} : for $i \in \{0, 1\}$, we define

$$\phi'(s, a_i) = \frac{1}{\sqrt{d+1}} \begin{bmatrix} \phi(s, a)\sqrt{d} \\ 1 - \frac{\delta_{i1}c(s, a)}{c_{\max}} \end{bmatrix}, \quad \boldsymbol{\theta'} = \frac{1}{1+c_{\max}} \sqrt{\frac{d+1}{d}} \begin{bmatrix} \boldsymbol{\theta} \\ c_{\max}\sqrt{d} \end{bmatrix}, \quad \boldsymbol{\mu'}(.) = \sqrt{\frac{d+1}{d}} \begin{bmatrix} \boldsymbol{\mu}(.) \\ 0 \end{bmatrix}.$$

To preserve the Linear MDP structure, we consider a scaled and shifted but equivalent reward function $r'(s,a_i) = \frac{r(s,a) + c_{\max} - \delta_{i1}c(s,a)}{1 + c_{\max}} \in [0,1]$, which incorporates observation costs.

Example 2.4 (Reset-to-observe). Consider a Linear MDP to which we add a "restart" action a^* , whose execution always triggers a data-burst and transitions the environment to a random state according to probability measure $\lambda(.)$ over S, while all standard actions do not provide observations. We can model this

with $\beta(a) = \mathbb{I}(a = a^*)$ and extending feature, reward, and measure vectors to \mathbb{R}^{d+1} as follows:

$$\phi'(s,a) = \begin{bmatrix} \phi(s,a) \cdot \mathbb{I}(a \neq a^*) \\ \mathbb{I}(a = a^*) \end{bmatrix}, \quad \theta' = \begin{bmatrix} \theta \\ 0 \end{bmatrix}, \quad \mu'(.) = \begin{bmatrix} \mu(.) \\ \lambda(.) \end{bmatrix}.$$

While these examples, rooted in real-world observation constraints, provide compelling motivation for studying ATST-MDP, they introduce additional structures beyond the core framework. Our paper focuses on the most general ATST-MDP setting without additional assumptions beyond Assumption 2.1 in later sections, providing theoretical results supported by rigorous proofs in the appendix.

2.3 Value-Functions and Optimality in the Augmented State Space

With runtime information about the current state in ATST-MDPs represented by an augmented state from $\mathcal{X} = \mathcal{S} \times \mathcal{A}^{<\mathbb{N}}$, it is natural to consider augmented policies $\pi : \mathcal{X} \to \mathcal{A}$ and appropriate value-functions on the augmented state space. For each augmented policy π , we define a value-function $Q^{\pi} : \mathcal{X} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}]$ as the expected cumulative discounted reward when starting from augmented state $x \in \mathcal{X}$ (with hidden initial state $s_1 \sim b(.|x)$), executing action a, and following policy π thereafter:

$$Q^{\pi}(x,a) = \mathbb{E}\left[r(s_1,a) + \sum_{h=2}^{\infty} \gamma^{h-1} r(s_h, \pi(x_h)) \middle| x_1 = x, a_1 = a\right],$$

where expectation is over $s_1 \sim b(.|x), \ s_{n+1} \sim \mathbb{P}(.|s_n,a_n), \ x_{n+1} \sim \begin{cases} s_{n+1} & \text{with probability } \beta(a_n) \\ x_n \oplus a_n & \text{otherwise} \end{cases}$.

The state value-function $V^{\pi}: \mathcal{X} \to [0, \frac{1}{1-\gamma}]$ is similarly defined as $V^{\pi}(x) = Q^{\pi}(x, \pi(x))$. Building on these definitions, we can establish a key recursive relationship for these value-functions.

Theorem 2.5. Under augmented policy $\pi: \mathcal{X} \to \mathcal{A}$, the action value-function satisfies:

$$Q^{\pi}(x,a) = \mathbb{E}_{s \sim b(\cdot|x)} \left[r(s,a) \right] + \gamma \beta(a) \mathbb{E}_{s' \sim b(\cdot|x \oplus a)} \left[V^{\pi}(s') \right] + \gamma \bar{\beta}(a) V^{\pi}(x \oplus a).$$

This theorem directly connects to the classical Bellman equation framework in RL theory. For the set of measurable functions $\mathcal{V} = \{V : \mathcal{X} \to [0, \frac{1}{1-\gamma}]\}$, we obtain the Bellman optimality operator $\mathbb{T} : \mathcal{V} \to \mathcal{V}$ as

$$\mathbb{T}V(x) = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{s \sim b(.|x)} \left[r(s, a) \right] + \gamma \beta(a) \mathbb{E}_{s' \sim b(.|x \oplus a)} \left[V(s') \right] + \gamma \bar{\beta}(a) V(x \oplus a) \right\},\,$$

which turns out to be a γ -contraction. Applying the Banach fixed-point theorem, it follows that there exists an optimal augmented policy π^* such that $V^*(x) = \sup_{\pi} V^{\pi}(x)$ for all $x \in \mathcal{X}$. The proofs of these claims and Theorem 2.5 are provided in Appendix B.

Introducing action-sequence value-function. A key property of augmented policies in ATST-MDPs is that the sequence of actions selected between data-bursts is obtained by repeatedly applying the policy to the augmented state that grows by appending each selected action. This recursive process allows us to define $a^{\pi}(x)$, the sequence of actions generated by a policy π at a state x, as follows:

Definition 2.6. For
$$\pi: \mathcal{X} \to \mathcal{A}$$
 and $n \in \mathbb{N}$, let $\pi^{(n)}: \mathcal{X} \to \mathcal{A}$ be inductively defined as $\pi^{(1)} = \pi$ and $\pi^{(n+1)}(x) = \pi(x \oplus (\pi^{(1)}(x), \dots, \pi^{(n)}(x)))$. Then, let $\boldsymbol{a}^{\pi}(x) = (\pi^{(1)}(x), \pi^{(2)}(x), \dots)$.

As a novel concept, we define action-sequence value-function $K^{\pi}: \mathcal{X} \times \mathcal{A}^{\mathbb{N}} \to [0, \frac{1}{1-\gamma}]$ as the expected cumulative discounted reward when starting from augmented state $x \in \mathcal{X}$ and following sequence $a = (a_1, a_2, ...) \in \mathcal{A}^{\mathbb{N}}$ until the next data-burst and policy π thereafter. Notably, $V^{\pi}(x) = K^{\pi}(x, \mathbf{a}^{\pi}(x))$.

To formalize this mathematically, let $b_h \in \{0,1\}$ denote the occurrence of a data-burst at round h (where $b_h|a_h \sim \text{Ber}(\beta(a_h))$) and define $T_{\text{DB}} = \min\{h \in \mathbb{N} : b_h = 1\}$ as the first round with a data-burst, so that:

$$K^{\pi}(x, \boldsymbol{a}) = \mathbb{E}_{s_1 \sim b(.|x)} \left[\left. \sum_{h=1}^{T_{\text{DB}}} \gamma^{h-1} r(s_h, a_h) + \gamma^{T_{\text{DB}}} V^{\pi}(s_{T_{\text{DB}}+1}) \right| x_1 = x, \ (a_i)_{i=1}^{T_{\text{DB}}} = \boldsymbol{a}_{1:T_{\text{DB}}} \right].$$

For clearer analysis, we can decompose this function into two components. Let R(x, a) denote the expected discounted reward until the next data-burst. Additionally, for every function $V: \mathcal{S} \to \mathbb{R}$ (or $V: \mathcal{X} \to \mathbb{R}$), let $\mathbb{P}V(x,a)$ denote the expected discounted value of V at the state observed at the next data-burst (0 if $T_{\mathrm{DB}} = \infty$). Formally, we have

$$R(x, \boldsymbol{a}) = \mathbb{E}_{s_1 \sim b(.|x)} \left[\sum_{h=1}^{T_{\text{DB}}} \gamma^{h-1} r(s_h, a_h) \middle| x_1 = x, (a_i)_{i=1}^{T_{\text{DB}}} = \boldsymbol{a}_{1:T_{\text{DB}}} \right]$$

$$\mathbb{P}V(x, \boldsymbol{a}) = \mathbb{E}_{s_1 \sim b(.|x)} \left[\gamma^{T_{\text{DB}}} V(s_{T_{\text{DB}}+1}) \middle| x_1 = x, (a_i)_{i=1}^{T_{\text{DB}}} = \boldsymbol{a}_{1:T_{\text{DB}}} \right].$$

This formulation yields a clear decomposition $K^{\pi} = R + \mathbb{P}V^{\pi}$.

3 Linear ATST-MDPs

Here, we explore the properties of ATST-MDPs under Assumption 2.1, with proofs provided in Appendix C. For every action $a \in \mathcal{A}$, define its *action-matrix* as $M(a) = \int_{\mathcal{S}} \boldsymbol{\mu}(s) \phi(s, a)^{\top} ds$. Then, we extend the feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ to all augmented state $x = (s_1; a_1, ..., a_{\Delta}) \in \mathcal{X} \setminus \mathcal{S}$ as follows:

$$\phi(x)^{\top} = \phi(s_1, a_1)^{\top} \prod_{i=2}^{\Delta} M(a_i).$$

This extension enables us to establish crucial linear properties of belief distributions (1):

Lemma 3.1 (Linearity of belief). For all $x \in \mathcal{X} \setminus \mathcal{S}$, $b(.|x) = \langle \phi(x), \mu(.) \rangle$ and $\|\phi(x)\|_2 \leq 1$. Moreover, for every map $V : \mathcal{S} \to [0, 1/(1-\gamma)]$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$, it holds that

$$\mathbb{E}_{s \sim b(.|x)}\left[r(s,a)\right] = \langle \phi(x \oplus a), \theta \rangle, \quad \text{and} \quad \mathbb{E}_{s' \sim b(.|x \oplus a)}\left[V(s')\right] = \langle \phi(x \oplus a), \mathbf{v} \rangle,$$

where vector $\mathbf{v} = \int V(s) d\boldsymbol{\mu}(s)$ satisfies $\|\mathbf{v}\|_2 \leqslant \frac{\sqrt{d}}{1-\gamma}$.

Leveraging this result and conditioning on the first data-burst time T_{DB} , the components in the decomposition $K^{\pi} = R + \mathbb{P}V^{\pi}$ can be written as

$$R(x, a \oplus \mathbf{a}) = \phi(x \oplus a)^{\top} \left(\beta(a)I + \bar{\beta}(a)M_1(\mathbf{a})\right) \boldsymbol{\theta},$$

$$\mathbb{P}V^{\pi}(x, a \oplus \mathbf{a}) = \phi(x \oplus a)^{\top} \left(\beta(a)I + \bar{\beta}(a)M_2(\mathbf{a})\right) \gamma \boldsymbol{v}^{\pi},$$

where type 1 and 2 action-sequence matrices $M_1(a), M_2(a) \in \mathbb{R}^{d \times d}$ are respectively defined as

$$M_1(\mathbf{a}) = I + \sum_{k=1}^{\infty} \left[\gamma^k (\prod_{i=1}^{k-1} \bar{\beta}(a_i)) (\prod_{i=1}^k M(a_i)) \right],$$
 (2a)

$$M_2(\boldsymbol{a}) = \sum_{k=1}^{\infty} \left[\gamma^k (\prod_{i=1}^{k-1} \bar{\beta}(a_i)) \beta(a_k) (\prod_{i=1}^k M(a_i)) \right].$$
 (2b)

To integrate these components into a unified representation, we introduce a new *action-sequence feature* $map \ \psi : \mathcal{X} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{2d}$, defined as follows:

$$\psi(x, a \oplus \mathbf{a})^{\top} = \frac{1}{2} \phi(x \oplus a)^{\top} \left(\beta(a) I_{12} + \bar{\beta}(a) M_{12}(\mathbf{a}) \right), \tag{3}$$

where $I_{12} = [(1-\gamma)I \ \gamma I] \in \mathbb{R}^{d \times 2d}$ and $M_{12}(\boldsymbol{a}) = [(1-\gamma)M_1(\boldsymbol{a}) \ \gamma M_2(\boldsymbol{a})] \in \mathbb{R}^{d \times 2d}$.

This feature map construction leads to our main result for the action-sequence value-function:

Theorem 3.2 (Linearity of K^{π}). Define $\mathbf{v}_{12}^{\pi} = 2 \begin{bmatrix} \theta/(1-\gamma) \\ \mathbf{v}^{\pi} \end{bmatrix} \in \mathbb{R}^{2d}$, where $\mathbf{v}^{\pi} = \int_{\mathcal{S}} V^{\pi}(s) d\boldsymbol{\mu}(s)$. Then, for every $x \in \mathcal{X}$ and sequence $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$:

$$K^{\pi}(x, \boldsymbol{a}) = \langle \boldsymbol{\psi}(x, \boldsymbol{a}), \, \boldsymbol{v}_{12}^{\pi} \rangle.$$

Moreover, it holds that $\sup_{x,a} \|\psi(x,a)\|_2 \leqslant 1$ and $\|v_{12}^{\pi}\|_2 \leqslant \frac{4\sqrt{d}}{1-\gamma}$.

Theorem 3.2 shows that $K^{\pi}(x, a)$, though defined over an infinite action-sequence, is fully captured by the inner product of a fixed vector and a bounded feature map in \mathbb{R}^{2d} . Given access to this feature map, one can use regression techniques to learn K^* , as we demonstrate for episodic learning in Section 4.

3.1 Feature Map Estimation and Off-Policy Learning of Action-Matrices

Having established the theoretical foundation for linear representation of action-sequence value-functions, we now address practical implementation challenges regarding computation. While the linearity result of Theorem 3.2 is theoretically elegant, requiring exact knowledge of action-matrices is an unrealistic assumption in practical settings, even though it is less demanding than knowing the full transition dynamics $\mu(.)$.

This subsection addresses critical questions: Can we effectively approximate the action-sequence feature map ψ on domain $\mathcal{S} \times \mathcal{A}^{\mathbb{N}}$ when we only have estimates of action-matrices \widehat{M}_a and observation probabilities $\widehat{\beta}_a$? Also, can reliable estimates be obtained from off-policy data? We answer both questions affirmatively.

To formalize the notion of an acceptable approximation of our feature map, we define ϵ -admissibility:

Definition 3.3. For $\epsilon \geqslant 0$, function $\hat{\psi}: \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{2d}$ is said to be an ϵ -admissible estimation of ψ in (3) if the following three conditions hold: $\sup_{s,\boldsymbol{a}} \|(\hat{\psi} - \psi)(s,\boldsymbol{a})\|_2 \leqslant \epsilon$, $\sup_{s,\boldsymbol{a}} \|\hat{\psi}(s,\boldsymbol{a})\|_2 \leqslant 1$, and $\hat{\psi}(s,\cdot)$ is continuous with respect to the product topology on $\mathcal{A}^{\mathbb{N}}$ and the standard topology on \mathbb{R}^{2d} for every $s \in \mathcal{S}$.

The following theorem describes construction of ϵ -admissible estimations, given estimates for action-matrices and data-burst probabilities. In particular, this confirms that ψ is a 0-admissible estimation of itself.

Theorem 3.4. Assume estimates $\widehat{M}_a \in \mathbb{R}^{2d \times 2d}$ and $\widehat{\beta}_a \in [0,1]$ satisfy $\sup_{a \in \mathcal{A}} \|\widehat{M}_a - M_a\|_2 \leqslant \varepsilon$ and $\sup_{a \in \mathcal{A}} |\widehat{\beta}_a - \beta_a| \leqslant \varepsilon_{\beta}$ for some $\varepsilon \in [0,\frac{1-\gamma}{2\sqrt{d}}]$ and $\varepsilon_{\beta} \in [0,1]$. Let $\widehat{\psi}: \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{2d}$ be the estimated action-sequence feature map obtained from (3) by replacing action-matrices M_a and data-burst probabilities β_a with their estimates \widehat{M}_a , $\widehat{\beta}_a$ in computation. Then, it holds that $\sup_{s,a} \|(\widehat{\psi} - \psi)(s,a)\|_2 \leqslant \frac{16d}{1-\gamma}(\varepsilon + \varepsilon_{\beta}/\sqrt{d})$. Moreover, function $\widetilde{\psi}(s,a) = \frac{\widehat{\psi}(s,a)}{1+16d(\varepsilon + \varepsilon_{\beta}/\sqrt{d})/(1-\gamma)}$ is a $\frac{32d(\varepsilon + \varepsilon_{\beta}/\sqrt{d})}{1-\gamma}$ -admissible estimation of ψ .

This theorem guarantees admissibility of the normalized feature map $\widetilde{\psi}$ given uniform bounds on the action-matrix and data-burst probability estimation errors. Notably, the proof of the theorem shows that errors in estimating M_a and β_a propagate in a controlled manner through the infinite-horizon feature map construction, thanks to the special algebraic structure of matrices M_a . See Appendix C.2 for the proof.

Off-policy data model. To demonstrate that the assumptions in Theorem 3.4 can be satisfied in practice, we consider a standard off-policy sampling approach for data collection. We collect N samples from a distribution \mathcal{D} over $\mathcal{S} \times \mathcal{A}$, creating a dataset is $\{s_n, a_n, s'_n, b_n\}_{n=1}^N$, where (s_n, a_n) are drawn i.i.d from \mathcal{D} , states s'_n are sampled independently from the true transition dynamics $\mathbb{P}(.|s_n, a_n)$, and observation indicators $b_n \in \{0, 1\}$ sampled independently based on the true data-burst probabilities, i.e. $b_n | a_n \sim \text{Ber}(\beta_{a_n})$.

We assume that distribution \mathcal{D} provides sufficient exploration of the feature space, formalized by requiring that its second moment matrix $\Sigma = \mathbb{E}[\phi(s_1, a_1)\phi(s_1, a_1)^{\top}]$ is positive definite. The minimum eigenvalue $\lambda_{\min}(\Sigma) > 0$ quantifies the quality of this exploration. Additionally, we assume that either the true probabilities β_a are known or that each action is sampled with positive probability: $p_{\min} = \inf_{a \in \mathcal{A}} \mathbb{E}[\mathbb{I}(a_1 = a)] > 0$.

For action-matrices, we employ ridge estimators with parameter $\lambda > 0$: $\widehat{M}_a^{\lambda} = (X^{\top}X + \lambda I_d)^{-1}X^{\top}Y_a$, where $X, Y_a \in \mathbb{R}^{N \times d}$ have rows $\phi(s_n, a_n)$, $\phi(s'_n, a)$ respectively. And, for data-burst probabilities, we either employ empirical mean estimators: $\widehat{\beta}_a = \frac{\sum_{n=1}^N b_n \mathbb{I}(a_n = a)}{\sum_{n=1}^N \mathbb{I}(a_n = a)}$, or assume that true β_a are known.

Lemmas 3.5 and 3.6 provide high-probability uniform bounds on the estimation errors for \widehat{M}_a^1 and $\widehat{\beta}_a$, with their proofs presented in Appendix C.3.

Lemma 3.5. There exists absolute constant $C \geqslant 1$ such that for all $p \in (0,1)$ and $N \geqslant \frac{4C^2d\log(2Ad/p)}{\lambda_{\min}(\Sigma)^2}$, by choosing $\lambda = 1$, ridge estimators \widehat{M}_a^{λ} satisfy

$$\mathbb{P}\left(\sup_{a\in\mathcal{A}}\|\widehat{M}_a^{\lambda} - M_a\|_2 \leqslant 4C\sqrt{\frac{d\log(2Ad/p)}{N\lambda_{\min}(\Sigma)^2}}\right) \geqslant 1 - p.$$

Lemma 3.6. For all $p \in (0,1)$ and $N \ge 1$, empirical mean estimators $\hat{\beta}_a$ satisfy

$$\mathbb{P}\left(\sup_{a\in\mathcal{A}}|\widehat{\beta}_a - \beta_a| \leqslant \sqrt{\frac{12\ln(3A/p)}{Np_{\min}}}\right) \geqslant 1 - p.$$

Therefore, both action-matrices and data-burst probabilities can be effectively estimated from off-policy data, with estimation error decreasing at the standard statistical rate of $O(1/\sqrt{N})$, when $\lambda_{\min}(\Sigma)$, $p_{\min} > 0$.

Combining these lemmas with Theorem 3.4, we immediately obtain a complete practical framework for estimating the action-sequence feature map using off-policy data, as follows.

Corollary 3.7. Consider action-sequence feature map estimation procedure of Theorem 3.4. Let $\widetilde{\psi}_{\text{off-policy}}^{M,\beta}$ denote the estimated feature map computed using estimates \widehat{M}_a^1 and $\widehat{\beta}_a$ constructed from $N_{M,\beta}$ off-policy data points. Similarly, let $\widetilde{\psi}_{\text{off-policy}}^M$ denote the estimated feature map computed using true data-burst probabilities β_a and estimates \widehat{M}_a^1 constructed from N_M data points.

There exists an absolute constant c > 0 such that for all $p \in (0,1)$ and $\varepsilon \in (0,1)$, the following holds:

- If $N_{M,\beta} \geqslant c \cdot \frac{d^3 \log(2Ad/p)}{\varepsilon^2(1-\gamma)^2 \min\{\lambda_{\min}(\Sigma)^2, d^2p_{\min}\}}$, then $\widetilde{\psi}_{\text{off-policy}}^{M,\beta}$ is ε -admissible with probability at least 1-p.
- If $N_M \geqslant c \cdot \frac{d^3 \log(2Ad/p)}{\varepsilon^2 (1-\gamma)^2 \lambda_{\min}(\Sigma)^2}$, then $\widetilde{\psi}_{\text{off-policy}}^M$ is ε -admissible with probability at least 1-p.

In this corollary, the dataset requirement for joint estimation of M_a , β_a has at least linear dependence on the action space size, whereas the requirement for estimating only M_a scales logarithmically with A. This is because $p_{\min} \leq 1/A$, making the gap unavoidable since the estimation of each β_a relies on N/A data points on average. In contrast, the condition $\lambda_{\min}(\Sigma) > 0$ is relatively easy to satisfy, even when the support of $\mathcal D$ is restricted to d points in $\mathcal S \times \mathcal A$ whose feature maps form a non-singular basis. Thus, the assumption that β_a are known is highly valuable for action-sequence feature map estimation, and it is plausible for many real-world applications, e.g., the "paid observations" in Example 2.3.

Episodic Learning with Geometric Horizons

This section explores episodic reinforcement learning in a linear ATST-MDP (Figure 1), where the agent interacts with an environment over K episodes. Each episode k has random length of $H^k \sim \text{Geom}(1-\gamma)$ rounds, or equivalently, episode termination occurs independently with probability $1-\gamma$ each round. At the start of each episode, the agent selects a policy and executes actions according to it, observing the new state and total reward only during action-triggered data-bursts or episode termination (an implicit data-burst).

We allow the agent to select a burst-dependent policy $\pi = (\pi_u)_{u=1}^{\infty}$, where each deterministic policy π_u : $\mathcal{X} \to \mathcal{A}$ governs actions until the u-th data-burst, at which point the agent switches to the following π_{u+1} . This approach generalizes stationary policies considered in previous sections to a more powerful class of adaptive strategies. The linearity properties (e.g., Theorem 3.2) extend to burst-dependent policies, with V^{π} and K^{π} defined as expected total discounted rewards under this policy-switching mechanism.

Episodic Learning under ATST-MDP

For each episode $k = 1, 2, \dots, K$:

The environment initializes total reward $G_0^k=0$. The agent selects a burst-dependent policy π^k based on data from previous episodes.

The adversary selects an initial state s_1^k and reveals it to the agent as augmented state $x_1^k = s_1^k$.

For rounds $h = 1, 2, \ldots$:

- 1. The agent executes a_h^k determined by π^k , incurring unobserved reward $r_h^k = r(s_h^k, a_h^k)$. The environment updates $G_h^k = G_{h-1}^k + r_h^k$ and samples next state $s_{h+1}^k \sim \mathbb{P}(.|s_h^k, a_h^k)$.
- 2. **Episode termination** occurs with probability 1γ : the environment reveals pair (\varnothing, G_h^k) .
- 3. *Data-burst* occurs with probability $\beta(a_h^k)$: the environment reveals pair (s_{h+1}^k, G_h^k) . 4. The agent updates $x_{h+1}^k = s_{h+1}^k$ if data-burst occurred, and as $x_{h+1}^k = x_h^k \oplus a_h^k$ otherwise.

Figure 1: Execution protocol of the ATST-MDP over K episodes with geometric horizons.

To formalize episode termination, we introduce a termination state \varnothing reached with probability $1-\gamma$ each round. For all value-functions V, K, and Q, we define $V(\emptyset) = K(\emptyset, \mathbf{a}) = Q(\emptyset, a) = 0$.

In each episode $k \in [K]$, observation history can be presented as tuples $(\mathbf{s}_u^k, a_u^k, R_u^k, \mathbf{s}_{u+1}^k)_{u=1}^{B^k}$ corresponding to data-bursts. Here, B^k represents the number of data-bursts (including termination) in episode k; $\mathbf{s}_u^k \in \mathcal{S}$ denote observed states, with $\mathbf{s}_{B^k+1}^k = \varnothing$; $\boldsymbol{a}_u^k = \boldsymbol{a}^{\pi_u^k}(\mathbf{s}_u^k) \in \mathcal{A}^{\mathbb{N}}$ are sequences of actions that would be played until the next data-burst based on policy π^k from state \mathbf{s}_u^k ; and $R_u^k \ge 0$ are aggregated rewards for rounds between observing \mathbf{s}_{u}^{k} and \mathbf{s}_{u+1}^{k} .

The agent's objective is to minimize the total (expected) regret $\mathcal{R}_K = \sum_{k=1}^K (V^*(s_1^k) - V^{\pi^k}(s_1^k))$, i.e., the shortfall in the player's expected cumulative reward compared to that of the optimal augmented policy $\pi^*: \mathcal{X} \to \mathcal{A}$ in this ATST-MDP, where expectation is taken over the stochastic dynamics of each episode. It is worth noting that under ATST-MDP, π^* is the optimal policy that balances blind decision-making without state observations and the cost of acquiring new information through data-bursts. This policy is generally different from the optimal policy in the underlying MDP, which always has access to the current state. π^* is the natural benchmark for evaluating learning algorithms in this setting, as it represents the best possible performance given the constraints of sporadic observations.

4.1 Algorithm

Our Algorithm 1 (ST-LSVI-UCB) is based on the Least-Squares Value Iteration with Uniform Confidence Bound of Jin et al. [Jin+19], which we adapt to handle sporadic traceability and geometric horizons. This algorithm requires access to an ϵ -admissible estimation map $\hat{\psi}$, with regret bounds depending on this ϵ .

To stabilize computations, we use the **effective horizon** parameter H which serves three purposes: limit value iteration steps to H, cap the number of data-burst used in learning to $\min\{B^k,H\}$ per episode, and bound accumulated rewards as $\overline{R}_u^k = \min\{R_u^k,H\}$. Although separate parameters of similar magnitude could be employed, we simplify analysis by using the common parameter H.

Given this parameter, we define the **effective history** at the start of episode k as $\mathcal{H}^k = \left(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau}, R^{\tau}, \mathbf{s}_N^{\tau}\right)_{\tau=1}^{N^k}$, consisting of at most H first data-burst-tuples from each episode (i.e., $N^k = \sum_{k'=1}^{k-1} \min\{B^{k'}, H\}$), with $\mathbf{s}_N^{\tau} \in \mathcal{S} \cup \{\varnothing\}$ denoting the next revealed state after \mathbf{s}^{τ} . Also, let $\hat{\psi}^{\tau} = \hat{\psi}(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})$ and $\psi^{\tau} = \psi(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})$, with the latter being unknown to the agent and used solely in our theoretical analysis.

Algorithm 1 ST-LSVI-UCB

```
Input: estimation feature map \hat{\psi}: \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{2d}, discount factor \gamma.
Parameters: effective horizon H, regularizers \lambda and \rho.
  1: for episode k = 1, \dots, K do
               Compile observations from episodes 1, ..., k-1 into effective history \mathcal{H}^k.
               Compute \Lambda^k = \lambda I + \sum_{\tau=1}^{N^k} \hat{\psi}^{\tau} (\hat{\psi}^{\tau})^{\top}.
Initialize K_u^k(x, \boldsymbol{a}) = \frac{1}{1-\gamma} for all (x, \boldsymbol{a}) \in \mathcal{X} \times \mathcal{A}^{\mathbb{N}} and u \geqslant H.
  3:
  4:
               for u = H - 1, ..., 1 do
  5:
                      Compute \boldsymbol{w}_{u}^{k} = (\Lambda^{k})^{-1} \sum_{\tau=1}^{N^{k}} \Big[ \min\{R^{\tau}, H\} + \max_{\boldsymbol{a}} K_{u+1}^{k}(\mathbf{s}_{N}^{\tau}, \boldsymbol{a}) \Big].
  6:
                      Set K_u^k(s, \boldsymbol{a}) = \min \Big\{ \frac{1}{1-\gamma}, \langle \hat{\boldsymbol{\psi}}(s, \boldsymbol{a}), \boldsymbol{w}_u^k \rangle + \rho \|\hat{\boldsymbol{\psi}}(s, \boldsymbol{a})\|_{(\Lambda^k)^{-1}} \Big\}.
  7:
  8:
               end for
               Initialize counter u=1 and receive the initial state \mathbf{s}_1^k.
  9:
               while episode k continues do
 10:
                      Select action-sequence \boldsymbol{a}_u^k \in \arg\max_{\boldsymbol{a} \in \mathcal{A}^{\mathbb{N}}} K_u^k(\mathbf{s}_u^k, \boldsymbol{a}).
 11:
                      Play actions from a_u^k until either:
 12:
                           (1) a data-burst occurs, revealing \mathbf{s}_{u+1}^k and R_u^k, or (2) the episode terminates (\mathbf{s}_{u+1}^k = \varnothing) and R_u^k is revealed.
 13:
 14:
                      Increment u \leftarrow u + 1 and break if episode terminated.
 15:
               end while
 16:
 17: end for
```

At a high level, ST-LSVI-UCB performs two passes over all rounds. The first pass performs backward value-iteration, computing parameters \boldsymbol{w}_u^k that form the K-value-functions $K_u^k: \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to [0, (1-\gamma)^{-1}]$. Using the estimation map $\hat{\psi}$, these functions aim to approximate optimal $K^*(s, \boldsymbol{a}) = \langle \psi(s, \boldsymbol{a}), v_{12}^{\pi^*} \rangle$. The second pass executes the greedy policy by selecting action-sequences \boldsymbol{a}_u^k , maximizing $K_u^k(\mathbf{s}_u^k, \cdot)$, that the agent follows until the next data-burst. Only the second pass involves actual interaction with the environment.

The optimization in lines 6 and 11 requires computing $\max_{\boldsymbol{a}\in\mathcal{A}^{\mathbb{N}}}K_u^k(\mathbf{s}_u^k,\boldsymbol{a})$. Despite infinite-dimensionality of action-sequence space $\mathcal{A}^{\mathbb{N}}$, this problem can be framed as optimizing a convex function over $\widehat{\Psi}_{\mathbf{s}} = \{\widehat{\psi}(\mathbf{s},\boldsymbol{a}): \boldsymbol{a}\in\mathcal{A}^{\mathbb{N}}\}$ - a complicated but compact set in \mathbb{R}^{2d} for ϵ -admissible $\widehat{\psi}$. Crucially, the compactness guarantees the existence of a maximizing action-sequence. The γ -discounting and lower bound on $\inf_a \beta(a)$

provide computational advantages, as action influence decays exponentially with time until a data-burst. In practice, we can approximately solve such optimization problems by truncating the horizons and solving the finite-dimensional problem using gradient-based methods, though the worst-case computational complexity is still exponential. Our theoretical analysis therefore assumes access to an optimization oracle.

4.2 Theoretical Guarantees

Now we are ready to present the main result for episodic learning. We assume that the approximate feature map $\hat{\psi}$ used in Algorithm 1 is ϵ -admissible, with $\epsilon \leqslant \sqrt{(1-\gamma)/K}$. According to Corollary 3.7, this can be achieved using an off-policy estimation procedure. Given a confidence parameter $p \in (0,1)$ and number of episodes K, we set the parameters in Algorithm 1 as

$$H = \lceil \frac{\log(K(1-\gamma)^{-1})}{1-\gamma} \rceil + 1, \quad \lambda = 1, \quad \text{and} \quad \rho = c \cdot dH\sqrt{\iota} \quad \text{for} \quad \iota = \log(2dKH/p),$$

where c is an absolute constant.

Theorem 4.1 (ST-LSVI-UCB regret guarantee). There exists an absolute constant $c \ge 1$, such that under the above setup, with probability at least 1 - p, the total regret of Algorithm 1 is at most

$$\widetilde{O}(\sqrt{d^3K(1-\gamma)^{-3}\iota^2} + d^2(1-\gamma)^{-2}\iota + \epsilon \cdot \sqrt{d^2K^3(1-\gamma)^{-5}\iota}),$$

where \widetilde{O} omits polylogarithmic factors independent of $\log(1/p)$.

The proof is provided in Appendix D. Notably, for sufficiently small ϵ , the regret bound matches the classical $\widetilde{O}(\sqrt{K(1-\gamma)^{-3}})$ rate for MDPs (e.g., see [Man+23]). In particular, this optimal rate is attained when $\epsilon = O((1-\gamma)/K)$, while the general guarantee holds under the milder condition $\epsilon < \sqrt{(1-\gamma)/K}$. Corollary 3.7 shows that this level of estimation accuracy can be achieved from off-policy data using $\widetilde{O}(K^2d^3/(1-\gamma)^4)$ samples, with high probability.

5 Discussion and Future Work

This work introduces ATST-MDPs, a novel framework that captures the challenges of reinforcement learning in environments where state observability is action-triggered and sporadic. Our theoretical contributions include new Bellman optimality equations for this setting, a linear structure in the induced action-sequence feature map, and rigorous approximation guarantees for learning feature maps from off-policy data. We also design and analyze ST-LSVI-UCB, an algorithm that provably achieves low regret in episodic learning under geometric horizons, provided access to an accurate estimation of the action-sequence feature map.

Several interesting questions remain open for future research. First, ST-LSVI-UCB assumes access to an optimization oracle over action-sequences, a computationally demanding requirement in general. Designing efficient approximation schemes, such as restricting to finite-depth action trees or developing tractable surrogate objectives, would significantly enhance practical applicability. Second, while we establish off-policy methods for estimating action-matrices and data-burst probabilities, a fully online algorithm that adaptively refines these estimates during learning would provide a more robust and practical solution.

Additionally, ATST-MDPs offer a novel perspective on RL with stochastic delays (e.g., [Bou+21]). Classical models treat delays as *exogenous*; here they are *endogenous*, with actions shaping the distribution of observation times. A unifying view allows *round-dependent* data-burst probabilities $\beta_t(a)$: when β_t is

action-independent, one recovers some exogenous delay models. Analyzing how different delay-generation mechanisms affect learning and regret presents a promising research direction.

Overall, our results establish a foundation for learning under action-triggered state-dependent observations, and the flexibility of our formulation opens pathways toward addressing information constraints across a wide range of sequential decision-making problems.

References

- [Ast65] K. J. Astrom. "Optimal Control of Markov Processes with Incomplete State Information". In: *Journal of Mathematical Analysis and Applications* 10.1 (1965), pp. 174–205.
- [Sat+17] Y. Satsangi, S. Whiteson, F. A. Oliehoek, and M. T. J. Spaan. "Exploiting submodular value functions for scaling up active perception". In: *Autonomous Robots* 42.2 (Aug. 2017), pp. 209–233. ISSN: 1573-7527.
- [SR23] J. Shang and M. S. Ryoo. Active Vision Reinforcement Learning under Limited Visual Observability. 2023.
- [KSJ23] M. Krale, T. D. Simao, and N. Jansen. *Act-Then-Measure: Reinforcement Learning for Partially Observable Environments with Active Measuring*. 2023.
- [NFB21] H. A. Nam, S. L. Fleming, and E. Brunskill. "Reinforcement learning with state observation costs in action-contingent noiselessly observable markov decision processes". *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393.
- [Bel+20] C. Bellinger, R. Coles, M. Crowley, and I. Tamblyn. *Active Measure Reinforcement Learning for Observation Cost Minimization*. 2020.
- [Wan+25] T. Wang, J. Liu, B. Lee, Z. Wu, and Y. Wu. OCMDP: Observation-Constrained Markov Decision Process. 2025.
- [HS17] M. Hausknecht and P. Stone. *Deep Recurrent Q-Learning for Partially Observable MDPs*. 2017.
- [KTO18] R. Klíma, K. Tuyls, and F. A. Oliehoek. "Model-Based Reinforcement Learning under Periodical Observability". *AAAI Spring Symposia*. 2018.
- [CL25] G. Chen and S.-C. Liew. *Intermittently Observable Markov Decision Processes*. 2025.
- [KE03] K. Katsikopoulos and S. Engelbrecht. "Markov decision processes with delays and asynchronous cost collection". In: *IEEE Transactions on Automatic Control* 48.4 (2003), pp. 568–574.
- [Wal+09] T. J. Walsh, A. Nouri, L. Li, and M. L. Littman. "Learning and planning in environments with delayed feedback". In: *Autonomous Agents and Multi-Agent Systems* 18 (2009), pp. 83–105.
- [Lio23] P. Liotet. Delays in Reinforcement Learning. 2023.
- [Sch+15] T. Schaul, D. Horgan, K. Gregor, and D. Silver. "Universal Value Function Approximators". Proceedings of the 32nd International Conference on Machine Learning. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 1312–1320.
- [And+18] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba. *Hindsight Experience Replay*. 2018.

- [PGT03] J. Pineau, G. Gordon, and S. Thrun. "Point-based value iteration: an anytime algorithm for POMDPs". *International Joint Conference on Artificial Intelligence*. 2003.
- [SV10] D. Silver and J. Veness. "Monte-Carlo Planning in Large POMDPs". Advances in Neural Information Processing Systems. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc., 2010.
- [CYW24] Q. Cai, Z. Yang, and Z. Wang. Reinforcement Learning from Partial Observation: Linear Function Approximation with Provable Sample Efficiency. 2024.
- [Put94] M. L. Puterman. "Discounted Markov Decision Problems". In: *Markov Decision Processes*. John Wiley & Sons, Ltd, 1994. Chap. 6, pp. 142–276. ISBN: 9780470316887.
- [Man+23] D. Mandal, G. Radanovic, J. Gan, A. Singla, and R. Majumdar. *Online Reinforcement Learning with Uncertain Episode Lengths*. 2023.
- [Bou+21] Y. Bouteiller, S. Ramstedt, G. Beltrame, C. Pal, and J. Binas. "Reinforcement Learning with Random Delays". *International Conference on Learning Representations*. 2021.
- [Jin+19] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. *Provably Efficient Reinforcement Learning with Linear Function Approximation*. 2019.
- [KLC98] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. "Planning and Acting in Partially Observable Stochastic Domains". In: *Artif. Intell.* 101 (1998), pp. 99–134.
- [SS12] T. Smith and R. Simmons. Heuristic Search Value Iteration for POMDPs. 2012.
- [Lio+22] P. Liotet, D. Maran, L. Bisi, and M. Restelli. *Delayed Reinforcement Learning by Imitation*. 2022.
- [Sel+14] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori. "Prediction with Limited Advice and Multiarmed Bandits with Paid Observations". *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research. Bejing, China: PMLR, June 2014, pp. 280–287.
- [AB10] J.-Y. Audibert and S. Bubeck. "Regret Bounds and Minimax Policies under Partial Monitoring". In: *Journal of Machine Learning Research* 11.94 (2010), pp. 2785–2836.
- [Ver11] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. 2011.
- [Tro15] J. A. Tropp. An Introduction to Matrix Concentration Inequalities. 2015.
- [APS11] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger. Vol. 24. Curran Associates, Inc., 2011.

A Additional Related Work

POMDPs and planning under partial observability. Classical work on decision making with incomplete state information is captured by POMDPs; see the survey of [KLC98] and subsequent algorithmic advances such as point-based value iteration (PBVI) [PGT03] and heuristic search value iteration (HSVI) [SS12]. Recent progress includes statistical and computational guarantees for learning and planning in partially observed settings [CYW24]. Much of this line is theoretical and algorithmic, with empirical validations on standard POMDP benchmarks; deep implementations typically combine belief updates with function approximation, but the core guarantees are model-based and non-neural.

RL with delayed observations (and augmented states). Early formulations analyze delayed MDPs and augmented-state reductions that stack the last observed state with a queue of intervening actions [KE03; Wal+09]. More recent work examines random delays in deep RL, showing robustness and performance trade-offs under synthetic and real latency processes [Bou+21], and explores imitation/learning pipelines that must handle delayed feedback [Lio+22]. This area mixes theory (augmented-state equivalence, stability) with empirical deep RL; implementations often use standard neural agents (e.g., DQN/actor-critic) evaluated under injected delays.

Goal-conditioned reinforcement learning. Goal-conditioned RL provides observations (and learning signals) when goals are achieved. Universal Value Function Approximators (UVFA) [Sch+15] parametrize value functions by goals, and Hindsight Experience Replay (HER) [And+18] augments replay with achieved goals to improve sample efficiency. These works are predominantly empirical deep RL (CNN/RNN policies and value functions on robotics and navigation tasks), with limited formal regret analysis.

Paid observations and information acquisition. Another related line studies decision making when observations incur explicit costs. In RL, agents may choose when to acquire measurements or labels, trading reward for information [Bel+20; NFB21; Wan+25]. In online learning, closely related "label-efficient" and budgeted feedback models investigate how querying constraints affect regret [Sel+14; AB10]. This area blends theoretical formulations (budget/constraint design, regret) with empirical demonstrations; deep implementations appear mainly in application-driven studies.

Intermittent observations and unreliable sensing. A practical motif is intermittently available observations due to sensing/communication failures. Deep Recurrent Q-Learning (DRQN) [HS17] tackles partial observability (flickering screen) by replacing feedforward policies with RNNs, showing empirical gains under dropped observations. Subsequent empirical studies examine control with sporadic measurements or packet loss [KTO18]. More recent formulations introduce intermittently observable MDPs with modeling/algorithmic structure beyond ad-hoc masking [CL25]. This line is largely empirical deep RL.

Active sensing and perception. Active perception frames sensing as a decision problem: agents select actions that improve informativeness while pursuing task reward. Active-perception POMDPs [Sat+17] formalize this, and recent deep RL approaches study active vision and act-then-measure protocols that interleave task actions with targeted measurements [SR23; KSJ23]. These works are primarily empirical and use deep neural networks (vision backbones with policy/value heads), sometimes with recurrent modules for memory; theoretical analysis focuses on tractable planning surrogates and approximate belief updates rather than regret.

B Augmented Policies: Proofs

In this section, we prove existence of the optimal augmented policy $\pi^* : \mathcal{X} \to \mathcal{A}$. The argument follows by classic application of the Banach fixed-point theorem for the Bellman optimality operator (e.g., see [Put94]). First, we restate and prove Theorem 2.5.

Theorem 2.5 (Restated). *Under augmented policy* $\pi: \mathcal{X} \to \mathcal{A}$, the action value-function satisfies:

$$Q^{\pi}(x,a) = \mathbb{E}_{s \sim b(.|x)} \left[r(s,a) \right] + \gamma \beta(a) \mathbb{E}_{s' \sim b(.|x \oplus a)} \left[V^{\pi}(s') \right] + \gamma \bar{\beta}(a) V^{\pi}(x \oplus a).$$

Proof. $Q^{\pi}(x,a)$ is the expected return when starting from x, taking action a, and following π thereafter. The term $\mathbb{E}_{s \sim b(.|x)}\left[r(s,a)\right]$ is the expected immediate reward for executing action a. After executing a, the environment proceeds to an augmented state that depends on whether a data-burst occurs:

- with probability $\beta(a)$, the next state $s' \sim b(\cdot \mid x \oplus a)$ matches the next augmented state, and the continuation value is $V^{\pi}(s')$;
- with probability $\bar{\beta}(a)$, no new state is observed, the next augmented state is $x \oplus a$, and the continuation value is $V^{\pi}(x \oplus a)$.

Taking expectations and discounting yields the result.

Theorem B.1. Let \mathcal{M} be an ATST-MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma, \beta)$. Define the space of measurable value-functions $\mathcal{V} = \{V : \mathcal{X} \to [0, \frac{1}{1-\gamma}]\}$, and the Bellman optimality operator $\mathbb{T} : \mathcal{V} \to \mathcal{V}$ as

$$\mathbb{T}V(x) = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}_{s \sim b(.|x)} \left[r(s, a) \right] + \gamma \beta(a) \mathbb{E}_{s' \sim b(.|x \oplus a)} \left[V(s') \right] + \gamma \bar{\beta}(a) V(x \oplus a) \right\}.$$

Then, \mathbb{T} is a γ -contraction, meaning that for all $V,U\in\mathcal{V}$, we have $\|\mathbb{T}V-\mathbb{T}U\|_{\infty}\leqslant\gamma\,\|V-U\|_{\infty}$.

Proof. For all $V \in \mathcal{V}$, let function $Q_V : \mathcal{X} \times \mathcal{A} \to [0, \frac{1}{1-\gamma}]$ be given by

$$Q_{V}(x,a) = \mathbb{E}_{s \sim b(.|x)} \left[r(s,a) \right] + \gamma \beta(a) \mathbb{E}_{s' \sim b(.|x \oplus a)} \left[V(s') \right] + \gamma \bar{\beta}(a) V(x \oplus a),$$

so that the Bellman optimality operator satisfies $\mathbb{T}V(x) = \max_a Q_V(x,a)$.

Fix arbitrary $V, U \in \mathcal{V}$. For every $x \in \mathcal{X}$, we can write

$$\begin{aligned} \left| \mathbb{T}V(x) - \mathbb{T}U(x) \right| &= \left| \max_{a} Q_{V}(x, a) - \max_{a} Q_{U}(x, a) \right| \\ &\leq \max_{a} \left| Q_{V}(x, a) - Q_{U}(x, a) \right| \\ &= \max_{a} \left| \gamma \beta(a) \mathbb{E}_{s' \sim b(.|x \oplus a)} \left[(V - U)(s') \right] + \gamma \bar{\beta}(a) (V - U)(x \oplus a) \right| \\ &\leq \max_{a} \left(\gamma \beta(a) \left\| V - U \right\|_{\infty} + \gamma \bar{\beta}(a) \left\| V - U \right\|_{\infty} \right) \\ &= \gamma \left\| V - U \right\|_{\infty}. \end{aligned}$$

Thus, \mathbb{T} is indeed a γ -contraction on \mathcal{V} .

Corollary B.2. Under the conditions of Theorem B.1, there exists an optimal policy $\pi^*: \mathcal{X} \to \mathcal{A}$ that achieves $V^{\pi^*}(x) = \sup_{\pi} V^{\pi}(x)$ for every $x \in \mathcal{X}$.

Proof. It is easy to verify that function $V^*(x) := \sup_{\pi} V^{\pi}(x)$ has to be a fixed-point of \mathbb{T} by Theorem 2.5. From Theorem B.1 and Banach fixed-point theorem, we conclude that V^* is the unique fixed-point of \mathbb{T} . Consider any policy $\pi^* : \mathcal{X} \to \mathcal{A}$ such that for all $x \in \mathcal{X}$:

$$\pi^*(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ \mathbb{E}_{s \sim b(.|x)} \left[r(s,a) \right] + \gamma \beta(a) \mathbb{E}_{s' \sim b(.|x \oplus a)} \left[V^*(s') \right] + \gamma \bar{\beta}(a) V^*(x \oplus a) \right\}.$$

Then, $V^{\pi^*} = V^*$ because π^* always selects an action that attains the supremum in the Bellman equation. The reasoning follows [Put94].

Additionally, we provide formulas for R and $\mathbb{P}V$, obtained by conditioning on T_{DB} .

Lemma B.3. For all $x \in \mathcal{X}$ and $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$, it holds that

$$R(x, \boldsymbol{a}) = \sum_{h=1}^{\infty} \gamma^{h-1} \left(\prod_{i=1}^{h-1} \bar{\beta}(a_i) \right) \mathbb{E}_{s \sim b(.|\tilde{x}_h)} \left[r(s, a_h) \right],$$

$$\mathbb{P}V(x, \boldsymbol{a}) = \sum_{h=1}^{\infty} \gamma^h \left(\prod_{i=1}^{h-1} \bar{\beta}(a_i) \right) \beta(a_h) \mathbb{E}_{s' \sim b(.|\tilde{x}_{h+1})} \left[V^{\pi}(s') \right].$$

where $\widetilde{x}_h = x \oplus (a_i)_{i=1}^{h-1} \in \mathcal{X}$ for every $h \in \mathbb{N}$.

Proof. Let $\mathbb{P}(.|\boldsymbol{a})$ denote the probability measure of T_{DB} over $\mathbb{N} \cup \{\infty\}$ when the agents commits to playing sequence of actions $\boldsymbol{a} = (a_1, a_2, ...) \in \mathcal{A}^{\mathbb{N}}$. Then, it holds that $\mathbb{P}(T_{\mathrm{DB}} \geqslant h \mid \boldsymbol{a}) = \prod_{i=1}^{h-1} \bar{\beta}(a_i)$ and $\mathbb{P}(T_{\mathrm{DB}} = h \mid \boldsymbol{a}) = (\prod_{i=1}^{h-1} \bar{\beta}(a_i))\beta(a_h)$ for all $h \in \mathbb{N}$.

Then, by conditioning on T_{DB} , we can write

$$R(x, \boldsymbol{a}) = \mathbb{E}_{s_{1} \sim b(.|x)} \left[\sum_{h=1}^{T_{\text{DB}}} \gamma^{h-1} r(s_{h}, a_{h}) \middle| x_{1} = x, (a_{i})_{i=1}^{T_{\text{DB}}} = \boldsymbol{a}_{1:T_{\text{DB}}} \right]$$

$$= \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{s \sim b(.|x \oplus (a_{1}, ..., a_{h-1}))} [r(s, a_{k})] \cdot \mathbb{P}(T_{\text{DB}} \geqslant h | \boldsymbol{a})$$

$$= \sum_{h=1}^{\infty} \gamma^{h-1} \left(\prod_{i=1}^{h-1} \bar{\beta}(a_{i}) \right) \mathbb{E}_{s \sim b(.|\tilde{x}_{h})} [r(s, a_{h})],$$

$$\mathbb{P}V(x, \boldsymbol{a}) = \mathbb{E}_{s_{1} \sim b(.|x)} \left[\gamma^{T_{\text{DB}}} V(s_{T_{\text{DB}}+1}) \middle| x_{1} = x, (a_{i})_{i=1}^{T_{\text{DB}}} = \boldsymbol{a}_{1:T_{\text{DB}}} \right]$$

$$= \sum_{h=1}^{\infty} \gamma^{h} \mathbb{E}_{s' \sim b(.|x \oplus (a_{1}, ..., a_{h}))} [V(s')] \cdot \mathbb{P}(T_{\text{DB}} = h | \boldsymbol{a})$$

$$= \sum_{h=1}^{\infty} \gamma^{h} \left(\prod_{i=1}^{h-1} \bar{\beta}(a_{i}) \right) \beta(a_{h}) \mathbb{E}_{s' \sim b(.|\tilde{x}_{h+1})} [V^{\pi}(s')].$$

Thus, both formulas are correct.

C Linear ATST-MDPs: Proofs

C.1 Linearity of Belief and Action-Sequence Value-Function

In this subsection, we prove: Lemma 3.1 and Theorem 3.2.

Lemma 3.1 (Restated). For all $x \in \mathcal{X} \setminus \mathcal{S}$, $b(.|x) = \phi(x)^{\top} \mu(.)$ and $\|\phi(x)\|_2 \leq 1$. Moreover, for every map $V : \mathcal{S} \to [0, 1/(1-\gamma)]$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$, it holds that

$$\mathbb{E}_{s \sim b(.|x)}\left[r(s,a)\right] = \langle \phi(x \oplus a), \theta \rangle, \quad \text{and} \quad \mathbb{E}_{s' \sim b(.|x \oplus a)}\left[V(s')\right] = \langle \phi(x \oplus a), v \rangle,$$

where vector $\mathbf{v} = \int V(s) d\mathbf{\mu}(s)$ satisfies $\|\mathbf{v}\|_2 \leqslant \frac{\sqrt{d}}{1-\gamma}$.

Proof. We prove these claims separately:

1. Linearity of belief: Fix $x \in \mathcal{X} \setminus \mathcal{S}$ and let $x = (s_1; a_1, ..., a_{\Delta})$. Then, the belief b(.|x) satisfies

$$\begin{aligned} b(.|x) &= \int_{\mathcal{S}^{\Delta-1}} \left[\prod_{i=2}^{\Delta} \mathbb{P}(s_i|s_{i-1}, a_{i-1}) \right] \mathbb{P}(s|s_{\Delta}, a_{\Delta}) \, ds_i \\ &= \int_{\mathcal{S}^{\Delta-1}} \left[\prod_{i=2}^{\Delta} \phi(s_{i-1}, a_{i-1})^{\top} \boldsymbol{\mu}(s_i) \right] \, \phi(s_{\Delta}, a_{\Delta})^{\top} \boldsymbol{\mu}(.) \, ds_i \\ &= \phi(s_1, a_1)^{\top} \left[\prod_{i=2}^{\Delta} \left(\int_{\mathcal{S}} \boldsymbol{\mu}(s_i) \phi(s_i, a_i)^{\top} ds_i \right) \right] \boldsymbol{\mu}(.) \\ &= \langle \phi(x), \, \boldsymbol{\mu}(.) \rangle. \end{aligned}$$

2. **Norm bound:** From Assumption 2.1, $\sup_{s,a} \|\phi(s,a)\|_2 \le 1$. Consider any $x \in \mathcal{X} \setminus \mathcal{S}$ and $a \in \mathcal{A}$. Then, using linearity of belief, we can write

$$\boldsymbol{\phi}(x \oplus a)^{\top} = \boldsymbol{\phi}(x)^{\top} M(a) = \int_{S} \boldsymbol{\phi}(x)^{\top} \boldsymbol{\mu}(s) \boldsymbol{\phi}(s, a)^{\top} ds = \mathbb{E}_{s \sim b(.|x)} \boldsymbol{\phi}(s, a)^{\top},$$

from which the result follows by Jensen's inequality due to convexity of l^2 -norm

$$\|\phi(x \oplus a)\|_2 = \|\mathbb{E}_{s \sim b(.|x)}\phi(s, a)\|_2 \le \mathbb{E}_{s \sim b(.|x)} \|\phi(s, a)\|_2 \le 1.$$

3. Linearity of expected reward and value-function: From Assumption 2.1, $r(s, a) = \phi(s, a)^{\top} \theta$. Now, for all $(x, a) \in (\mathcal{X} \setminus \mathcal{S}) \times \mathcal{A}$, we have:

$$\mathbb{E}_{s \sim b(.|x)} \left[r(s, a) \right] = \int_{\mathcal{S}} \phi(x)^{\top} \boldsymbol{\mu}(s) \phi(s, a)^{\top} \boldsymbol{\theta} \, ds = \phi(x)^{\top} M(a) \, \boldsymbol{\theta} = \phi(x \oplus a)^{\top} \, \boldsymbol{\theta}.$$

Similarly, for all $x \in \mathcal{X} \setminus \mathcal{S}$, it holds that

$$\mathbb{E}_{s \sim b(.|x)} \left[V(s) \right] = \int_{\mathcal{S}} \boldsymbol{\phi}(x)^{\top} \boldsymbol{\mu}(s) V(s) \, ds = \boldsymbol{\phi}(x)^{\top} \boldsymbol{v},$$

where $m{v} = \int_{\mathcal{S}} m{\mu}(s) V(s) ds$ satisfies $\| m{v} \|_2 \leqslant \sup_s |V(s)| \cdot \| |m{\mu}|(\mathcal{S}) \|_2 \leqslant \frac{\sqrt{d}}{1-\gamma}$.

Theorem 3.2 (Restated). Define $\mathbf{v}_{12}^{\pi} = 2 \begin{bmatrix} \theta/(1-\gamma) \\ \mathbf{v}^{\pi} \end{bmatrix} \in \mathbb{R}^{2d}$, where $\mathbf{v}^{\pi} = \int_{\mathcal{S}} V^{\pi}(s) d\boldsymbol{\mu}(s)$. Then, for every $x \in \mathcal{X}$ and sequence $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$:

$$K^{\pi}(x, \boldsymbol{a}) = \langle \boldsymbol{\psi}(x, \boldsymbol{a}), \, \boldsymbol{v}_{12}^{\pi} \rangle.$$

Moreover, it holds that $\sup_{x, a} \|\psi(x, a)\|_2 \leqslant 1$ and $\|v_{12}^{\pi}\|_2 \leqslant \frac{4\sqrt{d}}{1-\gamma}$.

Proof. Follows immediately from the following Theorem C.1, we prove linearity in ψ for both R and $\mathbb{P}V^{\pi}$ in the decomposition $K^{\pi} = R + \mathbb{P}V^{\pi}$.

Theorem C.1 (Linearity of R and $\mathbb{P}V$ with respect to ψ). For every $x \in \mathcal{X}$, sequence $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$, and function $V: \mathcal{S} \to [0, (1-\gamma)^{-1}]$, it holds that

$$R(x, \boldsymbol{a}) = \boldsymbol{\psi}(x, \boldsymbol{a})^{\top} \begin{bmatrix} 2\boldsymbol{\theta}/(1-\gamma) \\ \mathbf{0}_d \end{bmatrix} \quad \text{and} \quad \mathbb{P}V(x, \boldsymbol{a}) = \boldsymbol{\psi}(x, \boldsymbol{a})^{\top} \begin{bmatrix} \mathbf{0}_d \\ 2\boldsymbol{v} \end{bmatrix},$$

where $\mathbf{v} = \int_{\mathcal{S}} V(s) d\mu(s)$ satisfies $\|\mathbf{v}\|_2 \leqslant \frac{\sqrt{d}}{1-\gamma}$. Moreover, $\sup_{x,a} \|\psi(x,a)\|_2 \leqslant 1$.

Proof. Using Lemmas B.3 and 3.1, we write

$$\begin{split} R(x,a\oplus\boldsymbol{a}) &= \mathbb{E}_{s\sim b(.|x)}[r(s,a)] + \bar{\beta}(a) \sum_{k=1}^{\infty} \gamma^k \left(\prod_{i=1}^{k-1} \bar{\beta}(a_i)\right) \underset{s\sim b(.|x\oplus(a,a_1,\ldots,a_{k-1}))}{\mathbb{E}} \left[r(s,a_k)\right] \\ &= \phi(x\oplus a)^{\top} \boldsymbol{\theta} + \bar{\beta}(a) \sum_{k=1}^{\infty} \gamma^k \left(\prod_{i=1}^{k-1} \bar{\beta}(a_i)\right) \phi(x\oplus(a,a_1,\ldots,a_k))^{\top} \boldsymbol{\theta} \\ &= \phi(x\oplus a)^{\top} \left(I + \bar{\beta}(a) \sum_{k=1}^{\infty} \gamma^k (\prod_{i=1}^{k-1} \bar{\beta}(a_i)) \left(\prod_{i=1}^{k} M(a_i)\right)\right) \boldsymbol{\theta} \\ &= \phi(x\oplus a)^{\top} \left(\beta(a)I + \bar{\beta}(a)M_1(a)\right) \boldsymbol{\theta} \\ &= \frac{1}{2}\phi(x\oplus a)^{\top} \left(\beta(a)\cdot(1-\gamma)I + \bar{\beta}(a)\cdot(1-\gamma)M_1(a)\right) \left(2\boldsymbol{\theta}/(1-\gamma)\right) \\ &= \psi(x,a)^{\top} \begin{bmatrix} 2\boldsymbol{\theta}/(1-\gamma) \\ 0_d \end{bmatrix}, \\ \mathbb{P}V(x,a\oplus a) &= \beta(a)\gamma \underset{s\sim b(.|x\oplus a)}{\mathbb{E}} V(s) + \bar{\beta}(a)\gamma \underset{k=1}{\sum} \gamma^k \left(\prod_{i=1}^{k-1} \bar{\beta}(a_i)\right) \beta(a_k) \underset{s\sim b(.|x\oplus(a,a_{1:k})))}{\mathbb{E}} V(s) \\ &= \beta(a)\gamma \phi(x\oplus a)^{\top} v + \bar{\beta}(a)\gamma \underset{k=1}{\sum} \gamma^k \left(\prod_{i=1}^{k-1} \bar{\beta}(a_i)\right) \beta(a_k) \phi(x\oplus(a,\ldots,a_k))^{\top} v \\ &= \phi(x\oplus a)^{\top} \left(\beta(a)\gamma I + \bar{\beta}(a)\gamma \underset{k=1}{\sum} \gamma^k \left(\prod_{i=1}^{k-1} \bar{\beta}(a_i)\right) \beta(a_k) \left(\prod_{i=1}^{k} M(a_i)\right)\right) v \\ &= \phi(x\oplus a)^{\top} \left(\beta(a)\gamma I + \bar{\beta}(a)\gamma M_2(a)\right) v \\ &= \psi(x,a)^{\top} \begin{bmatrix} 0_a \\ 2v \end{bmatrix}. \end{split}$$

To bound the l_2 -norm, we write

$$\|\psi(x, a \oplus a)\|_{2} \leq \frac{1-\gamma}{2} \cdot \|\phi(x \oplus a) + \bar{\beta}(a) \sum_{k=1}^{\infty} \gamma^{k} (\prod_{i=1}^{k-1} \bar{\beta}(a_{i})) \phi(x \oplus a, a_{1}, ..., a_{k}))\|_{2}$$

$$+ \frac{1}{2} \|\beta(a)\gamma \phi(x \oplus a) + \bar{\beta}(a)\gamma \sum_{k=1}^{\infty} \gamma^{k} (\prod_{i=1}^{k-1} \bar{\beta}(a_{i})) \beta(a_{k}) \phi(x \oplus (a, ..., a_{k}))\|_{2}$$

$$\stackrel{\text{(a)}}{\leq} \left(\frac{1-\gamma}{2} \cdot (1 + \sum_{k=1}^{\infty} \gamma^{k}) + \frac{\gamma}{2} \cdot (\beta(a) + \bar{\beta}(a) \sum_{k=1}^{\infty} (\sum_{i=1}^{k-1} \bar{\beta}(a_{i})) \beta(a_{k}))\right)$$

$$\leq \left(\frac{1-\gamma}{2} \cdot \frac{1}{1-\gamma} + \frac{\gamma}{2} \cdot 1\right) = \frac{1+\gamma}{2} \leq 1.$$

where (a) uses the fact that $\sup_{x'} \left\| \phi(x') \right\|_2 \leqslant 1.$

C.2 Approximation of the Action-Sequence Feature Map: Proofs

In this subsection, we prove Theorem 3.4. A key technical tool is Lemma C.2 provided below.

Theorem 3.4 (Restated). Assume $\widehat{M}_a \in \mathbb{R}^{2d \times 2d}$ and $\widehat{\beta}_a \in [0,1]$ satisfy $\sup_{a \in \mathcal{A}} \|\widehat{M}_a - M_a\|_2 \leqslant \varepsilon$ and $\sup_{a \in \mathcal{A}} |\widehat{\beta}_a - \beta_a| \leqslant \varepsilon_{\beta}$ for some $\varepsilon \in [0,\frac{1-\gamma}{2\sqrt{d}}]$ and $\varepsilon_{\beta} \in [0,1]$. Let $\widehat{\psi}: \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{2d}$ be the estimated action-sequence feature map obtained from (3) by replacing action-matrices M_a and data-burst probabilities β_a with their estimates \widehat{M}_a , $\widehat{\beta}_a$ in computation. Then, it holds that $\sup_{s,\boldsymbol{a}} \|(\widehat{\psi} - \psi)(s,\boldsymbol{a})\|_2 \leqslant \frac{16d}{1-\gamma}(\varepsilon + \varepsilon_{\beta}/\sqrt{d})$.

Moreover, function
$$\widetilde{\psi}(s, \boldsymbol{a}) = \frac{\widehat{\psi}(s, \boldsymbol{a})}{1 + 16d(\varepsilon + \varepsilon_{\beta}/\sqrt{d})/(1 - \gamma)}$$
 is a $\frac{32d(\varepsilon + \varepsilon_{\beta}/\sqrt{d})}{1 - \gamma}$ -admissible estimation of ψ .

At the core of the proof is the following more general lemma, which bounds the estimation error in the feature vector ψ using that of action-matrices.

Lemma C.2. Assume estimates \widehat{M}_a satisfy $\sup_{a \in \mathcal{A}} \|\widehat{M}_a - M_a\|_2 \leqslant \varepsilon$ and define norm-corrected estimates $\widehat{M}_a^c = \widehat{M}_a/(1 + \varepsilon \sqrt{d})$. Also, suppose that estimates $\widehat{\beta}_a \in [0,1]$ satisfy $\sup_{a \in \mathcal{A}} |\widehat{\beta}_a - \beta_a| \leqslant \varepsilon_{\beta}$. Let $\widehat{\psi}, \widehat{\psi}_c : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{2d}$ be the estimated action-sequence feature maps obtained from ψ by replacing M_a, β_a with their estimates \widehat{M}_a (or \widehat{M}_a^c) and $\widehat{\beta}_a$, respectively. Then, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}^{\mathbb{N}}$, it holds that

$$\|(\widehat{\boldsymbol{\psi}}_c - \boldsymbol{\psi})(s, \boldsymbol{a})\|_2 \le \frac{4d^2}{1 - \gamma} \cdot (\varepsilon + \varepsilon_{\beta}/d^{3/2}).$$

Moreover, if $\varepsilon < (1/\gamma - 1)/\sqrt{d}$, then it holds that

$$\|(\widehat{\psi} - \psi)(s, \boldsymbol{a})\|_2 \le \frac{4d(1-\gamma)}{(1-\gamma(1+\varepsilon\sqrt{d}))^2} \cdot (\varepsilon + \varepsilon_\beta/\sqrt{d}).$$

Taking this lemma as given, let us prove Theorem 3.4.

Proof of Theorem 3.4. For $\varepsilon \in [0, \frac{1-\gamma}{2\sqrt{d}}]$, we have $\varepsilon < \frac{1/\gamma - 1}{\sqrt{d}}$. So, by the second case of Lemma C.2,

$$\sup_{s,\boldsymbol{a}} \|(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})(s,\boldsymbol{a})\|_{2} \leqslant \frac{4d(1-\gamma)}{(1-\gamma(1+\varepsilon\sqrt{d}))^{2}} \cdot (\varepsilon + \varepsilon_{\beta}/\sqrt{d}) \leqslant \frac{16d}{1-\gamma} \cdot (\varepsilon + \varepsilon_{\beta}/\sqrt{d}),$$

which proves the first statement. Now, we have to show that $\widetilde{\psi}$ is $\frac{32d(\varepsilon+\varepsilon_{\beta}/\sqrt{d})}{1-\gamma}$ -admissible estimation of ψ . Let $\epsilon_2=\frac{16d(\varepsilon+\varepsilon_{\beta}/\sqrt{d})}{1-\gamma}$. Then, for every s, \boldsymbol{a} write following

$$\|(\widetilde{\psi} - \psi)(s, \mathbf{a})\|_{2} \leq \frac{\|(\widehat{\psi} - \psi)(s, \mathbf{a})\|_{2}}{1 + \epsilon_{2}} + \frac{\epsilon_{2} \|\psi(s, \mathbf{a})\|_{2}}{1 + \epsilon_{2}} \leq 2\varepsilon_{2} = \frac{32d(\varepsilon + \varepsilon_{\beta}/\sqrt{d})}{1 - \gamma},$$
$$\|\widetilde{\psi}(s, \mathbf{a})\|_{2} \leq \frac{\|(\widehat{\psi} - \psi)(s, \mathbf{a})\|_{2}}{1 + \epsilon_{2}} + \frac{\|\psi(s, \mathbf{a})\|_{2}}{1 + \epsilon_{2}} \leq \frac{\epsilon_{2}}{1 + \epsilon_{2}} + \frac{1}{1 + \epsilon_{2}} = 1.$$

So, we only have to show continuity of $\widetilde{\psi}(s,.)$ with respect to the product topology on $\mathcal{A}^{\mathbb{N}}$ and the standard topology on \mathbb{R}^{2d} . This follows from the formula of $\widehat{\psi}$, which is based on the γ -discounted summation of matrix products. Each term is bounded in operator norm as shown by Lemma C.4:

$$\gamma^n \| \prod_{i=1}^n \widehat{M}_{a_i} \|_2 \leqslant \gamma^n \cdot \sqrt{d} (1 + \varepsilon \sqrt{d})^n \leqslant \sqrt{d} \cdot \left(\frac{1+\gamma}{2}\right)^n,$$

where exponent term $\frac{1+\gamma}{2} \in (0,1)$ ensures convergence and therefore continuity for $\hat{\psi}$ and $\tilde{\psi}$.

C.2.1 Proof of Lemma C.2

The following lemmas are used to prove Lemma C.2.

Lemma C.3. For all $n \in \mathbb{N}$ and $a_1, \ldots a_n \in \mathcal{A}$, it holds that $\| \prod_{i=1}^n M_{a_i} \|_2 \leq \sqrt{d}$.

Proof. Using the Linear MDP Assumption 2.1, we can write

$$\prod_{i=1}^{n} M_{a_i} = \int_{\mathcal{S}} \boldsymbol{\mu}(s) \boldsymbol{\phi}(s, a_1)^{\top} \prod_{i=2}^{n} M_{a_i} ds = \int_{\mathcal{S}} \boldsymbol{\mu}(s) \boldsymbol{\phi}((s; a_1, ..., a_n))^{\top} ds.$$

Then, by spectral-Frobenius inequality, it follows that

$$\|\|\prod_{i=1}^{n} M_{a_{i}}\|\|_{2} \leq \sqrt{\sum_{i \in [d]} \|\int_{\mathcal{S}} \mu_{i}(s) \phi((s; a_{1}, ..., a_{n}))^{\top} ds \|_{2}^{2}}$$

$$\leq \sqrt{\sum_{i \in [d]} (|\mu_{i}|(\mathcal{S}))^{2} \cdot \sup_{x \in \mathcal{X}} \|\phi(x)\|_{2}^{2}}$$

$$= \||\mu|(\mathcal{S})\|_{2} \cdot \sup_{x \in \mathcal{X}} \|\phi(x)\|_{2} \leq \sqrt{d},$$

where the final inequality follows from Assumption 2.1 and Lemma 3.1.

Lemma C.4. Suppose that for every $a \in \mathcal{A}$, estimate $\widehat{M}_a \in \mathbb{R}^{d \times d}$ satisfies $||M_a - \widehat{M}_a||_2 \leq \varepsilon$. Then, for all $n \in \mathbb{N}$ and $a_1, \ldots a_n \in \mathcal{A}$, it holds that $|||\prod_{i=1}^n \widehat{M}_{a_i}||_2 \leq \sqrt{d}(1 + \varepsilon\sqrt{d})^n$.

Proof. Let $E_a=M_a-\widehat{M}_a$ so that $\widehat{M}_a=M_a+E_a$ and $|\!|\!| E_a|\!|\!|_2\leqslant \varepsilon.$

Also, let $X_a^0 = M_a$ and $X_a^1 = E_a$. Then, we can write

$$\begin{split} \|\|\prod_{i=1}^{n} \widehat{M}_{a_{i}}\|_{2} &= \|\|\prod_{i=1}^{n} (M_{a_{i}} + E_{a_{i}})\|_{2} \\ &\leq \sum_{\boldsymbol{b} \in \{0,1\}^{n}} \|\|\prod_{i=1}^{n} X_{a_{i}}^{b_{i}}\|_{2} \\ &\stackrel{\text{(a)}}{\leq} \sum_{\boldsymbol{b} \in \{0,1\}^{n}} \left(\sqrt{d} \prod_{i=1}^{n} [\mathbb{I}(b_{i} = 0) + \mathbb{I}(b_{i} = 1) \cdot \|E_{a_{i}}\|_{2} \sqrt{d}]\right) \\ &\leq \sqrt{d} \cdot \sum_{\boldsymbol{b} \in \{0,1\}^{n}} (\varepsilon \sqrt{d})^{\|\boldsymbol{b}\|_{1}} \\ &= \sqrt{d} (1 + \varepsilon \sqrt{d})^{n}. \end{split}$$

where (a) follows by bounding consecutive blocks of neighbouring X_a^0 matrices as $\|X_{a_l}^0X_{a_{l+1}}^0\dots X_{a_r}^0\|_2 \leqslant \sqrt{d}$ using Lemma C.3 and pairing each such block (except maybe one) with a neighbouring matrix X_a^1 , which has $\|X_a^1\|_2 = \|E_a\|_2 \leqslant \sqrt{d}$.

Lemma C.5. Let $\varepsilon \in [0,1)$. Suppose matrices $A, B \in \mathbb{R}^{d \times d}$ satisfy $|||A|||_2 \leqslant \sqrt{d}$ and $|||A - B|||_2 \leqslant \varepsilon$. Then, $B' = B/(1 + \varepsilon\sqrt{d})$ satisfies $|||A - B'||_2 \leqslant 2\varepsilon$.

Proof. Let $A' = A/(1 + \varepsilon \sqrt{d})$. Using the triangle inequality, we can write

$$|||A - B'|||_2 \leqslant |||A - A'|||_2 + |||A' - B'|||_2 \leqslant \frac{\varepsilon \sqrt{d}}{1 + \varepsilon \sqrt{d}} \cdot |||A|||_2 + \frac{1}{1 + \varepsilon \sqrt{d}} \cdot |||A - B|||_2 \leqslant 2d\varepsilon.$$

Lemma C.6. Under the conditions of Lemma C.4, let $\widehat{M}_a^c = \widehat{M}_a/(1+\varepsilon\sqrt{d})$. Then, we have

$$\| \prod_{i=1}^{n} \widehat{M}_{a_i} - \prod_{i=1}^{n} M_{a_i} \|_2 \leqslant d(1 + \varepsilon \sqrt{d})^{n-1} n\varepsilon, \tag{4}$$

$$\|\|\prod_{i=1}^{n} \widehat{M}_{a_i}^c - \prod_{i=1}^{n} M_{a_i}\|\|_2 \le 2d^2 n\varepsilon.$$
 (5)

Proof. To show (4), we write

$$\begin{aligned} \| \prod_{i=1}^{n} \widehat{M}_{a_{i}} - \prod_{i=1}^{n} M_{a_{i}} \|_{2} &\leq \sum_{k=1}^{n} \| (\prod_{i=1}^{k-1} \widehat{M}_{a_{i}}) (\widehat{M}_{a_{k}} - M_{a_{k}}) (\prod_{i=k+1}^{n} M_{a_{i}}) \|_{2} \\ &\leq \sum_{k=1}^{n} \| \prod_{i=1}^{k-1} \widehat{M}_{a_{i}} \|_{2} \| \widehat{M}_{a_{k}} - M_{a_{k}} \|_{2} \| \prod_{i=k+1}^{n} M_{a_{i}} \|_{2} \\ &\stackrel{\text{(a)}}{\leq} \sum_{k=1}^{n} \left(\sqrt{d} (1 + \sqrt{d}\varepsilon)^{k-1} \cdot \varepsilon \cdot \sqrt{d} \right) \leq d (1 + \varepsilon \sqrt{d})^{n-1} n\varepsilon \end{aligned}$$

where (a) follows from Lemmas C.3 and C.4.

Similarly, to prove (5), we write

$$\begin{split} \| \prod_{i=1}^{n} \widehat{M}_{a_{i}}^{c} - \prod_{i=1}^{n} M_{a_{i}} \|_{2} & \leq \sum_{k=1}^{n} \| (\prod_{i=1}^{k-1} \widehat{M}_{a_{i}}^{c}) \left(\widehat{M}_{a_{k}}^{c} - M_{a_{k}} \right) \left(\prod_{i=k+1}^{n} M_{a_{i}} \right) \|_{2} \\ & \leq \sum_{k=1}^{n} \| \prod_{i=1}^{k-1} \widehat{M}_{a_{i}}^{c} \|_{2} \| \widehat{M}_{a_{k}}^{c} - M_{a_{k}} \|_{2} \| \prod_{i=k+1}^{n} M_{a_{i}} \|_{2} \\ & \leq \sum_{k=1}^{n} (\sqrt{d} \cdot 2d\varepsilon \cdot \sqrt{d}) = 2d^{2} n\varepsilon, \end{split}$$

where (b) follows from Lemmas C.3, C.4, and C.5.

Lemma C.7. Let sequences $(a_i)_{i=1}^{\infty}$, $(b_i)_{i=1}^{\infty}$ with values in [0,1] be such that $\sup_{i\in\mathbb{N}}|a_i-b_i|\leqslant \varepsilon$ for some $\varepsilon\in[0,1]$. Let $\bar{a}_i=1-a_i$ and $\bar{b}_i=1-b_i$ for every $i\in\mathbb{N}$. Then, it holds that

$$\forall n \in \mathbb{N}, \quad \left| \prod_{i=1}^{n} b_i - \prod_{i=1}^{n} a_i \right| \leqslant n\varepsilon, \tag{6}$$

$$\forall \gamma \in (0,1), \quad \sum_{k=1}^{\infty} \gamma^k | (\prod_{i=1}^{k-1} \overline{b}_i) b_k - (\prod_{i=1}^{k-1} \overline{a}_i) a_k | \leqslant \frac{2\varepsilon}{1-\gamma}. \tag{7}$$

Proof. To prove (6) for arbitrary $n \in \mathbb{N}$, we simply write:

$$|\prod_{i=1}^{n} b_{i} - \prod_{i=1}^{n} a_{i}| \leq \sum_{k=1}^{n} |\prod_{i=1}^{k-1} a_{i} \prod_{i=k}^{n} b_{i} - \prod_{i=1}^{k} a_{i} \prod_{i=k+1}^{n} b_{i}|$$

$$= \sum_{k=1}^{n} |b_{k} - a_{k}| \prod_{i=1}^{k-1} a_{i} \prod_{i=k+1}^{n} b_{i}$$

$$\leq n\varepsilon.$$

To prove (7) for arbitrary $\gamma \in (0,1)$, consider the finite supremum over all appropriate pairs of sequences:

$$S = \sup_{\boldsymbol{a}, \boldsymbol{b} \in [0,1]^{\mathbb{N}}: \sup_{i} |a_{i} - b_{i}| \leq \varepsilon} \sum_{k=1}^{\infty} \gamma^{k} |(\prod_{i=1}^{k-1} \bar{b}_{i}) b_{k} - (\prod_{i=1}^{k-1} \bar{a}_{i}) a_{k}| \leq \sum_{k=1}^{\infty} \gamma^{k} = \frac{1}{1-\gamma}$$

with intention to show that $S \leqslant \frac{2\varepsilon}{1-\gamma}$. Then, for all $a, b \in [0,1]^{\mathbb{N}}$ such that $\sup_i |a_i - b_i| \leqslant \varepsilon$, we can write:

$$\begin{split} \sum_{k=1}^{\infty} \gamma^{k} | (\prod_{i=1}^{k-1} \overline{b}_{i}) b_{k} - (\prod_{i=1}^{k-1} \overline{a}_{i}) a_{k} | & \leq \gamma |b_{1} - a_{1}| + \sum_{k=2}^{\infty} \gamma^{k} |\overline{b}_{1} - \overline{a}_{1}| \cdot | (\prod_{i=2}^{k-1} \overline{b}_{i}) b_{k} | \\ & + \sum_{k=2}^{\infty} \gamma^{k} |\overline{a}_{1}| \cdot | (\prod_{i=2}^{k-1} \overline{b}_{i}) b_{k} - (\prod_{i=2}^{k-1} \overline{a}_{i}) a_{k} | \\ & \leq \varepsilon \cdot (1 + \sum_{k=1}^{\infty} (\prod_{i=1}^{k-1} \overline{b}_{i+1}) b_{k+1}) \\ & + \gamma \cdot \sum_{k=1}^{\infty} \gamma^{k} | (\prod_{i=1}^{k-1} \overline{b}_{i+1}) b_{k+1} - (\prod_{i=1}^{k-1} \overline{a}_{i+1}) a_{k+1} | \\ & \leq 2\varepsilon + \gamma S. \end{split}$$

Therefore, it holds that $S \leq 2\varepsilon + \gamma S$ and so $S \leq \frac{2\varepsilon}{1-\gamma}$.

Proof of Lemma C.2. Let $\hat{\beta}_a = 1 - \hat{\beta}_a \in [0, 1]$ to ease notation. From Lemma C.3, it follows that matrices $M_1(\boldsymbol{a}), M_2(\boldsymbol{a})$ from (2) satisfy

$$|||M_1(\boldsymbol{a})||_2 \leqslant 1 + \sum_{k=1}^{\infty} \gamma^k (\prod_{i=1}^{k-1} \bar{\beta}_{a_i}) |||(\prod_{i=1}^k M_{a_i})|||_2 \leqslant \sum_{k=0}^{\infty} \gamma^k \sqrt{d} \leqslant \frac{\sqrt{d}}{1-\gamma},$$
(8a)

$$|||M_2(\boldsymbol{a})|||_2 \leqslant \sum_{k=1}^{\infty} \gamma^k (\prod_{i=1}^{k-1} \bar{\beta}_{a_i}) \beta_{a_k} ||| (\prod_{i=1}^k M_{a_i}) |||_2 \leqslant \sum_{k=1}^{\infty} (\prod_{i=1}^{k-1} \bar{\beta}_{a_i}) \beta_{a_k} \sqrt{d} \leqslant \sqrt{d}.$$
 (8b)

Part 1: We prove the result for $\hat{\psi}$ first. Suppose $\varepsilon \in [0, (1-1/\gamma)/\sqrt{d})$, so that $\gamma(1+\varepsilon\sqrt{d}) \in [0,1)$.

Let $\widehat{M}_1(\boldsymbol{a})$, $\widehat{M}_2(\boldsymbol{a})$ denote estimates for matrices $M_1(\boldsymbol{a})$, $M_1(\boldsymbol{a})$ computed using estimates \widehat{M}_a , $\widehat{\beta}_a$.

Note that for all $c \in [0,1)$, $\sum_{n=0}^{\infty} c^n n = \frac{c}{(1-c)^2}$ and $\sup_n c^n n \leq \frac{1}{1-c}$. Then, using Lemmas C.3, C.6, and C.7, we can write:

$$\begin{split} \|\widehat{M}_{1}(\boldsymbol{a}) - M_{1}(\boldsymbol{a})\|_{2} &\leqslant \sum_{k=1}^{\infty} \gamma^{k} \| (\prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}}) (\prod_{i=1}^{k} \widehat{M}_{a_{i}}) - (\prod_{i=1}^{k-1} \overline{\beta}_{a_{i}}) (\prod_{i=1}^{k} M_{a_{i}}) \|_{2} \\ &\leqslant \sum_{k=1}^{\infty} \gamma^{k} \| \prod_{i=1}^{k} \widehat{M}_{a_{i}} - \prod_{i=1}^{k} M_{a_{i}} \|_{2} \\ &+ \sum_{k=1}^{\infty} \gamma^{k} \left| \prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}} - \prod_{i=1}^{k-1} \overline{\beta}_{a_{i}} \right| \| \prod_{i=1}^{k} M_{a_{i}} \|_{2} \\ &\leqslant \sum_{k=1}^{\infty} \gamma^{k} (1 + \varepsilon \sqrt{d})^{k-1} k \varepsilon d + \sum_{k=1}^{\infty} \gamma^{k} k \varepsilon_{\beta} \sqrt{d} \\ &= \frac{\gamma(1+\varepsilon \sqrt{d})}{(1-\gamma(1+\varepsilon \sqrt{d}))^{2}} \cdot \frac{\varepsilon d}{1+\varepsilon \sqrt{d}} + \frac{\gamma}{(1-\gamma)^{2}} \cdot \varepsilon_{\beta} \sqrt{d} \\ &\leqslant \frac{d\gamma}{(1-\gamma(1+\varepsilon \sqrt{d}))^{2}} (\varepsilon + \varepsilon_{\beta}/\sqrt{d}), \\ \|\widehat{M}_{2}(\boldsymbol{a}) - M_{2}(\boldsymbol{a})\|_{2} &\leqslant \sum_{k=1}^{\infty} \gamma^{k} \| (\prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}}) \widehat{\beta}_{a_{k}} (\prod_{i=1}^{k} \widehat{M}_{a_{i}}) - (\prod_{i=1}^{k-1} \overline{\beta}_{a_{i}}) \beta_{a_{k}} (\prod_{i=1}^{k} M_{a_{i}}) \|_{2} \\ &\leqslant \sup_{k \in \mathbb{N}} \left(\gamma^{k} \cdot \| \prod_{i=1}^{k} \widehat{M}_{a_{i}} - \prod_{i=1}^{k} M_{a_{i}} \|_{2} \right) \\ &+ \sum_{k=1}^{\infty} \gamma^{k} \left| (\prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}}) \widehat{\beta}_{a_{k}} - (\prod_{i=1}^{k-1} \overline{\beta}_{a_{i}}) \beta_{a_{k}} \right| \| \prod_{i=1}^{k} M_{a_{i}} \|_{2} \\ &\leqslant \sup_{k \in \mathbb{N}} \gamma^{k} (1 + \varepsilon \sqrt{d})^{k-1} k \varepsilon d \\ &+ \sum_{k=1}^{\infty} \gamma^{k} \left| (\prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}}) \widehat{\beta}_{a_{k}} - (\prod_{i=1}^{k-1} \overline{\beta}_{a_{i}}) \beta_{a_{k}} \right| \sqrt{d} \\ &\leqslant \frac{2d}{1-\gamma(1+\varepsilon\sqrt{d})} \cdot \frac{\varepsilon d}{1+\varepsilon\sqrt{d}} + \frac{2}{1-\gamma} \cdot \varepsilon_{\beta} \sqrt{d} \\ &\leqslant \frac{2d}{1-\gamma(1+\varepsilon\sqrt{d})} \cdot (\varepsilon + \varepsilon_{\beta}/\sqrt{d}). \end{split}$$

From (3), we have that

$$\psi(s, a \oplus \mathbf{a})^{\top} = \frac{1}{2} \phi(s \oplus a)^{\top} (\beta_a I_{12} + \bar{\beta}_a M_{12}(\mathbf{a})),$$

where $I_{12} = [(1-\gamma)I \ \gamma I] \in \mathbb{R}^{d \times 2d}$ and $M_{12}(\boldsymbol{a}) = [(1-\gamma)M_1(\boldsymbol{a}) \ \gamma M_2(\boldsymbol{a})] \in \mathbb{R}^{d \times 2d}$.

Then, using the fact that $\|\phi(s,a)\|_2 \le 1$, it follows that

$$\begin{split} \|(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi})(s, a \oplus \boldsymbol{a})\|_{2} &\leq \frac{1}{2} \| \left[(1 - \gamma)(\widehat{M}_{1} - M_{1})(\boldsymbol{a}) \ \gamma(\widehat{M}_{2} - M_{2})(\boldsymbol{a}) \right]^{\top} \|_{2} \\ &+ \frac{1}{2} \left| \widehat{\beta}_{a} - \beta_{a} \right| \cdot \| \left[(1 - \gamma)(I - M_{1}(\boldsymbol{a})) \ \gamma(I - M_{2}(\boldsymbol{a})) \right]^{\top} \|_{2} \\ &\stackrel{\text{(a)}}{\leq} \frac{1}{2} (1 - \gamma) \| (\widehat{M}_{1} - M_{1})(\boldsymbol{a}) \|_{2} + \frac{1}{2} \gamma \| (\widehat{M}_{2} - M_{2})(\boldsymbol{a}) \|_{2} \\ &+ \frac{1}{2} \varepsilon_{\beta} (1 - \gamma) (1 + \sqrt{d}/(1 - \gamma)) + \frac{1}{2} \varepsilon_{\beta} \gamma(1 + \sqrt{d}) \\ &\stackrel{\text{(b)}}{\leq} \frac{2 d(1 - \gamma)}{(1 - \gamma(1 + \varepsilon\sqrt{d}))^{2}} \cdot (\varepsilon + \varepsilon_{\beta}/\sqrt{d}) + 2\varepsilon_{\beta} \sqrt{d} \\ &\leq \frac{4 d(1 - \gamma)}{(1 - \gamma(1 + \varepsilon\sqrt{d}))^{2}} \cdot (\varepsilon + \varepsilon_{\beta}/\sqrt{d}) \end{split}$$

where (a) follows from (8) and (b) from the bounds on $\|\widehat{M}_1(a) - M_1(a)\|_2$ and $\|\widehat{M}_2(a) - M_2(a)\|_2$ above.

Part 2: Here, we will prove the result for $\hat{\psi}_c$ using similar approach. Suppose $\varepsilon \in [0,1)$.

Let $\widehat{M}_1^c(\boldsymbol{a})$, $\widehat{M}_2^c(\boldsymbol{a})$ denote estimates for matrices $M_1(\boldsymbol{a})$, $M_1(\boldsymbol{a})$ computed using estimates \widehat{M}_a^c , $\widehat{\beta}_a$. Using Lemmas C.3, C.6, and C.7, we write:

$$\|\widehat{M}_{1}^{c}(\boldsymbol{a}) - M_{1}(\boldsymbol{a})\|_{2} \leqslant \sum_{k=1}^{\infty} \gamma^{k} \| \prod_{i=1}^{k} \widehat{M}_{a_{i}}^{c} - \prod_{i=1}^{k} M_{a_{i}} \|_{2}$$

$$+ \sum_{k=1}^{\infty} \gamma^{k} \left| \prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}} - \prod_{i=1}^{k-1} \overline{\beta}_{a_{i}} \right| \| \prod_{i=1}^{k} M_{a_{i}} \|_{2}$$

$$\leqslant \sum_{k=1}^{\infty} \gamma^{k} 2d^{2}k\varepsilon + \sum_{k=1}^{\infty} \gamma^{k}k\varepsilon_{\beta}\sqrt{d}$$

$$\leqslant \frac{2d\gamma}{(1-\gamma)^{2}} \cdot (d\varepsilon + \varepsilon_{\beta}/\sqrt{d}),$$

$$\|\widehat{M}_{2}^{c}(\boldsymbol{a}) - M_{2}(\boldsymbol{a})\|_{2} \leqslant \sup_{k \in \mathbb{N}} \left(\gamma^{k} \cdot \| \prod_{i=1}^{k} \widehat{M}_{a_{i}}^{c} - \prod_{i=1}^{k} M_{a_{i}} \|_{2} \right)$$

$$+ \sum_{k=1}^{\infty} \gamma^{k} \left| (\prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}}) \widehat{\beta}_{a_{k}} - (\prod_{i=1}^{k-1} \overline{\beta}_{a_{i}}) \beta_{a_{k}} \right| \| \prod_{i=1}^{k} M_{a_{i}} \|_{2}$$

$$\leqslant \sup_{k \in \mathbb{N}} \gamma^{k} 2d^{2}k\varepsilon + \sum_{k=1}^{\infty} \gamma^{k} \left| (\prod_{i=1}^{k-1} \widehat{\beta}_{a_{i}}) \widehat{\beta}_{a_{k}} - (\prod_{i=1}^{k-1} \overline{\beta}_{a_{i}}) \beta_{a_{k}} \right| \sqrt{d}$$

$$\leqslant \frac{2d^{2}\varepsilon}{1-\gamma} + \frac{2\varepsilon_{\beta}\sqrt{d}}{1-\gamma} \leqslant \frac{2d}{1-\gamma} \cdot (d\varepsilon + \varepsilon_{\beta}/\sqrt{d}).$$

As in Part 1, we conclude that

$$\begin{split} \|(\widehat{\boldsymbol{\psi}}_{c} - \boldsymbol{\psi})(s, a \oplus \boldsymbol{a})\|_{2} &\leq \frac{1}{2} \| \left[(1 - \gamma)(\widehat{M}_{1}^{c} - M_{1})(\boldsymbol{a}) \ \gamma(\widehat{M}_{2}^{c} - M_{2})(\boldsymbol{a}) \right]^{\top} \|_{2} \\ &+ \frac{1}{2} \left| \widehat{\beta}_{a} - \beta_{a} \right| \cdot \| \left[(1 - \gamma)(I - M_{1}(\boldsymbol{a})) \ \gamma(I - M_{2}(\boldsymbol{a})) \right]^{\top} \|_{2} \\ &\leq \frac{1}{2} (1 - \gamma) \| (\widehat{M}_{1}^{c} - M_{1})(\boldsymbol{a}) \|_{2} + \frac{1}{2} \gamma \| (\widehat{M}_{2}^{c} - M_{2})(\boldsymbol{a}) \|_{2} \\ &+ \frac{1}{2} \varepsilon_{\beta} (1 - \gamma)(1 + \sqrt{d}/(1 - \gamma)) + \frac{1}{2} \varepsilon_{\beta} \gamma(1 + \sqrt{d}) \\ &\stackrel{\text{(c)}}{\leq} \frac{2d}{1 - \gamma} \cdot (d\varepsilon + \varepsilon_{\beta}/\sqrt{d}) + 2\varepsilon_{\beta} \sqrt{d} \\ &\leq \frac{4d}{1 - \gamma} \cdot (d\varepsilon + \varepsilon_{\beta}/\sqrt{d}), \end{split}$$

where (c) follows from the bounds on $\|\widehat{M}_1^c(a) - M_1(a)\|_2$ and $\|\widehat{M}_2^c(a) - M_2(a)\|_2$ above.

This concludes the proof of both statements.

C.3 Off-policy Evaluation

In this subsection, we prove Lemma 3.5, which will follow from Lemma C.8, provided below. We also prove Lemma 3.6. Corollary 3.7 follows immediately from these lemmas, by setting $\varepsilon_{\beta} = \varepsilon \sqrt{d}$ small enough in Theorem 3.4 and picking dataset size in Lemmas 3.5 and 3.6 large enough for the resulting uniform bounds to hold with probabilities 1 - p/2 each.

For the sake of notation, let $\boldsymbol{x}^{(n)} := \boldsymbol{\phi}(s_n, a_n)$ and $\boldsymbol{y}_a^{(n)} := \boldsymbol{\phi}(s_n', a)$, so that $X, Y_a \in \mathbb{R}^{N \times d}$ have rows $\boldsymbol{x}^{(n)}, \boldsymbol{y}_a^{(n)}$ respectively. Then, $\Sigma = \mathbb{E}[\frac{1}{N}X^\top X] = \mathbb{E}[\sum_{n=1}^N \boldsymbol{x}^{(n)}(\boldsymbol{x}^{(n)})^\top]$.

Recall that we consider ridge estimators $\widehat{M}_a = (X^\top X + \lambda I_d)^{-1} X^\top Y_a$.

Observe that $\mathbb{E}[\boldsymbol{y}_a^{(n)} \mid s_n, a_n] = M_a^{\top} \boldsymbol{x}_n$ and $\|\boldsymbol{y}_a^{(n)}\|_2 \leqslant 1$ almost surely. Moreover, for $\boldsymbol{z}_a^{(n)} := \boldsymbol{y}_a^{(n)} - M_a \boldsymbol{x}_n$, it holds that $\|\boldsymbol{z}_a^{(n)}\|_2 \leqslant 2$. In the matrix form, we consider $Z_a := Y_a - XM_a$.

Lemma 3.5 (Restated). There exists absolute constant $C \ge 1$ such that for all $p \in (0,1)$ and $N \ge \frac{4C^2d\log(2Ad/p)}{\lambda_{\min}(\Sigma)^2}$, by choosing $\lambda = 1$, with probability at least 1 - p, it holds that

$$\sup_{a \in \mathcal{A}} \|\widehat{M}_a^{\lambda} - M_a\|_2 \le 4C\sqrt{\frac{d \log(2Ad/p)}{N\lambda_{\min}(\Sigma)^2}}.$$

Proof. We will show that this claim holds for the same $C \ge 1$ as in Lemma C.8.

Fix arbitrary $p \in (0,1)$ and $N \geqslant \frac{4C^2d\log(2Ad/p)}{\lambda_{\min}(\Sigma)^2}$. As $\lambda_{\min}(\Sigma) \leqslant \|\Sigma\|_2 \leqslant 1$, for this N, it holds that $\mathbb{P}(\mathcal{E}) \geqslant 1-p$, where \mathcal{E} denotes the event from Lemma C.8.

Conditioned on event \mathcal{E} , for every $a \in \mathcal{A}$, it holds that

$$\begin{split} \|\widehat{M}_{a}^{\lambda} - M_{a}\|_{2} & \leq \|(X^{\top}X + \lambda I_{d})^{-1}X^{\top}Z_{a} - \lambda(X^{\top}X + \lambda I_{d})^{-1}M_{a}\|_{2} \\ & \leq \|(X^{\top}X + \lambda I_{d})^{-1}\|_{2} \|X^{\top}Z_{a}\|_{2} + \lambda \|(X^{\top}X + \lambda I_{d})^{-1}\|_{2} \|M_{a}\|_{2} \\ & \leq \frac{\|X^{\top}Z_{a}\|_{2} + \lambda\sqrt{d}}{\lambda_{\min}(X^{\top}X) + \lambda} \leq \frac{C\sqrt{N\log(2Ad/p)} + \sqrt{d}}{N\lambda_{\min}(\Sigma) - C\sqrt{Nd\log(2/p)}} \\ & \leq \frac{2C\sqrt{Nd\log(2Ad/p)}}{N\lambda_{\min}(\Sigma)/2} = 4C\sqrt{\frac{d\log(2Ad/p)}{N\lambda_{\min}(\Sigma)^{2}}}. \end{split}$$

Note that we use the fact that $||M_a||_2 \leq \sqrt{d}$ from Lemma C.3.

Lemma C.8 (Concentration). There exists an absolute constant C such that for all $p \in (0,1)$ and $N \ge C^2 \cdot d \log(2Ad/p)$, event $\mathcal{E} = \mathcal{E}_X \cap (\cap_{a \in \mathcal{A}} \mathcal{E}_a)$, where

$$\mathcal{E}_X: \qquad \lambda_{\min}(X^\top X) \geqslant N\lambda_{\min}(\Sigma) - C\sqrt{Nd\log(2/p)},$$

$$\mathcal{E}_a: \qquad \|X^\top Z_a\|_2 \leqslant C\sqrt{N\log(2Ad/p)},$$

occurs with probability at least 1 - p.

Proof. It will suffice to show that there exists constant C such that for every $N \ge C^2 \cdot d \log(2Ad/p)$, it holds that $\mathbb{P}(\mathcal{E}_X) \ge 1 - \frac{p}{2}$ and $\mathbb{P}(\mathcal{E}_a) \ge 1 - \frac{p}{2A}$ for all $a \in \mathcal{A}$.

Part 1: Observe that rows in matrix X are independent sub-Gaussian vectors that are uniformly bounded in l_2 -norm by 1, because $\sup_{s,a} \|\phi(s,a)\|_2 \le 1$. Using Theorem C.9, fix absolute constants C_1 and c_1 so that

$$\forall N \in \mathbb{N}, \ \forall t \geqslant 0, \ \mathbb{P}\left(\|X^\top X - N\Sigma\|_2 \leqslant N \max\{\delta, \delta^2\}\right) \geqslant 1 - 2\exp(-c_1t^2) \quad \text{for } \delta = \frac{C_1\sqrt{d} + t}{\sqrt{N}}.$$

Then, we claim that $\mathbb{P}(\mathcal{E}_X) \geqslant 1 - \frac{p}{2}$ if we select $C \geqslant C_1 + \sqrt{2/c_1}$.

Note that the minimal eigenvalue of $X^{T}X$ can be bounded from below as follows:

$$\lambda_{\min}(\boldsymbol{X}^{\top}\boldsymbol{X}) \geqslant \lambda_{\min}(\boldsymbol{N}\boldsymbol{\Sigma}) - \|\boldsymbol{X}^{\top}\boldsymbol{X} - \boldsymbol{N}\boldsymbol{\Sigma}\|_{2}.$$

So, by setting $t = \sqrt{\log(4/p)/c_1}$, we obtain that, for all $N \ge C \cdot d \log(2/p)$, it holds that

$$\begin{split} \mathbb{P}(\mathcal{E}_X) \geqslant \mathbb{P}\left(\| X^\top X - N \Sigma \|_2 \leqslant C \cdot \sqrt{N d \log(2/p)} \right) \\ \geqslant \mathbb{P}\left(\| X^\top X - N \Sigma \|_2 \leqslant N \cdot \frac{C_1 \sqrt{d} + t}{\sqrt{N}} \right) \\ \geqslant 1 - 2 \exp(-c_1 t^2) = 1 - \frac{p}{2}. \end{split}$$

Part 2: We claim that $\mathbb{P}(\mathcal{E}_a) \ge 1 - \frac{p}{2A}$ for every action $a \in \mathcal{A}$ if we select $C \ge 8$.

Observe that for every action $a \in \mathcal{A}$, $Z_a^\top X = \sum_{n=1}^N S_a^{(n)}$, where matrices $S_a^{(n)} := \mathbf{z}_a^{(n)} (\mathbf{x}^{(n)})^\top$ are independent and satisfy the following properties:

$$\begin{split} &\text{Uniformly bounded:} \quad \|S_a^{(n)}\|_2 = \|\pmb{z}_a^{(n)}\|_2 \, \|\pmb{x}^{(n)}\|_2 \leqslant 2 \\ &\text{Centered:} \quad \mathbb{E}[S_a^{(n)}] = \mathbb{E}\left[\mathbb{E}[\pmb{z}_a^{(n)} \mid \pmb{x}^{(n)}](\pmb{x}^{(n)})^\top\right] = \mathbb{E}[\pmb{0}(\pmb{x}^{(n)})^\top] = 0_{d \times d}. \end{split}$$

Moreover, it holds that

$$\begin{split} & \| \mathbb{E} [S_a^{(n)}(S_a^{(n)})^\top] \|_2 \leqslant \mathbb{E} \left[\| \boldsymbol{x}^{(n)} \|_2^2 \cdot \mathbb{E} \left[\| \boldsymbol{z}_a^{(n)}(\boldsymbol{z}_a^{(n)})^\top \|_2 \, \middle| \, \boldsymbol{x}^{(n)} \right] \right] \leqslant 4, \\ & \| \mathbb{E} [(S_a^{(n)})^\top S_a^{(n)}] \|_2 \leqslant \mathbb{E} \left[\| \boldsymbol{x}^{(n)}(\boldsymbol{x}^{(n)})^\top \|_2 \cdot \mathbb{E} \left[\| \boldsymbol{z}_a^{(n)} \|_2^2 \middle| \, \boldsymbol{x}^{(n)} \right] \right] \leqslant 4, \end{split}$$

which implies that the variance statistic of the sum satisfies

$$\nu(Z_a^\top X) \leqslant \textstyle \sum_{n=1}^N \max \left\{ \| \mathbb{E}[S_a^{(n)}(S_a^{(n)})^\top] \|_2, \ \| \mathbb{E}[(S_a^{(n)})^\top S_a^{(n)}] \|_2 \right\} \leqslant 4N.$$

By Theorem C.10, we have that

$$\forall t \geqslant 0, \quad \mathbb{P}(\|X^{\top} Z_a\|_2 \geqslant t) \leqslant 2d \cdot \exp\left(\frac{-t^2/2}{4N + 2t/3}\right) \leqslant 2d \cdot \exp\left(\frac{-t^2/8}{N + t}\right).$$

So, for $N \ge C^2 \cdot \log(2Ad/p)$, fixing $t = \sqrt{16N \log(4Ad/p)} \le N$, yields

$$\mathbb{P}(\mathcal{E}_a) \geqslant \mathbb{P}\left(\|X^\top Z_a\|_2 \leqslant t\right) \geqslant 1 - 2d \cdot \exp\left(\frac{-t^2/8}{N+t}\right) \geqslant 1 - 2d \cdot \exp\left(\frac{-t^2}{16N}\right) = 1 - \frac{p}{2A}.$$

Conclusion: To sum up, the choice of the absolute constant $C = \max\{C_1 + \sqrt{2/c_1}, 8\}$ guarantees that for all $p \in (0,1)$ and $N \ge C^2 \cdot d \log(2Ad/p)$, it holds that $\mathbb{P}(\mathcal{E}) \ge 1 - p$.

Theorem C.9 (Theorem 5.39 (5.40) from [Ver11]). Let A be $N \times d$ matrix whose rows A_i are independent sub-Gaussian vectors in \mathbb{R}^d with common second moment matrix Σ . Let $K := \max_{i \in [N]} \|A_i\|_{\psi_2}$ denote the maximal sub-Gaussian norm among the rows. Then, there exist constants c and C that depend only on the value of K, such that, for every $t \ge 0$, the following inequality holds with probability at least $1 - 2 \exp(-ct^2)$:

$$\| \frac{1}{N} A^{\top} A - \Sigma \|_2 \leqslant \max\{\delta, \delta^2\} \quad \textit{where} \quad \delta = \frac{C\sqrt{d} + t}{\sqrt{N}}.$$

Theorem C.10 (Theorem 6.1.1 (Matrix Bernstein) from [Tro15]). Let $S_1, ..., S_n$ be independent \mathbb{R} -valued centered random matrices with common dimensions $d_1 \times d_2$, and suppose that for some $L \ge 0$, it holds that $||S_k||_2 \le L$ for every $k \in [n]$ almost surely. Consider their sum $Z := \sum_{k=1}^n S_k$ and let $\nu(Z)$ denote the variance statistic of the sum:

$$\nu(Z) := \max \left\{ \| \mathbb{E}[ZZ^\top] \|_2, \| \mathbb{E}[Z^\top Z] \|_2 \right\}.$$

Then, for all $t \ge 0$, it holds that

$$\mathbb{P}(|||Z||_2 \ge t) \le (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\nu(Z) + Lt/3}\right).$$

Lemma 3.6 (Restated). For all $p \in (0,1)$, empirical mean estimators $\hat{\beta}_a$ satisfy

$$\mathbb{P}\left(\sup_{a\in\mathcal{A}}|\widehat{\beta}_a - \beta_a| \leqslant \sqrt{\frac{12\ln(3A/p)}{Np_{\min}}}\right) \geqslant 1 - p.$$

Proof. For every $a \in \mathcal{A}$, let $N_a = \sum_{n=1}^N \mathbb{I}(a_n = a)$ and $S_a = \sum_{n=1}^N b_n \mathbb{I}(a_n = a)$, so that $\widehat{\beta}_a = S_a/N_a$. Also, let $p_a = \mathbb{E}[\mathbb{I}(a_1 = a)]$, so that $p_{\min} = \inf_{a \in \mathcal{A}} p_a$.

By Multiplicative Chernoff Bound, for fixed $a \in \mathcal{A}$ and arbitrary $\varepsilon \in (0,1)$, we have

$$\mathbb{P}(N_a \leqslant \frac{1}{2}Np_a) \leqslant \exp(-Np_a/8),$$

$$\mathbb{P}(|S_a - N_a\beta_a| = |(N - S_a) - N_a\bar{\beta}_a| \geqslant \varepsilon N_a \max\{\beta_a, \bar{\beta}_a\}|N_a) \leqslant 2\exp(-\varepsilon^2 N_a \max\{\beta_a, \bar{\beta}_a\}/3),$$

which allows us to write

$$\mathbb{P}(|\widehat{\beta}_{a} - \beta_{a}| \geq \varepsilon) = \mathbb{P}(|S_{a} - N_{a}\beta_{a}| \geq \varepsilon N_{a})$$

$$\leq \mathbb{P}(|S_{a} - N_{a}\beta_{a}| \geq \varepsilon N_{a}|N_{a} > \frac{1}{2}Np_{a}) + \mathbb{P}(N_{a} \leq \frac{1}{2}Np_{a})$$

$$\leq 2\exp(-\varepsilon^{2}Np_{a}\max\{\beta_{a}, \overline{\beta}_{a}\}/6) + \exp(-Np_{a}/8)$$

$$\leq 3\exp(-\varepsilon^{2}Np_{\min}/12).$$

Therefore, by the uniform confidence bound, for every $p \in (0, 1)$, it indeed holds that

$$\mathbb{P}\left(\sup_{a\in\mathcal{A}}|\widehat{\beta}_a - \beta_a| \leqslant \sqrt{\frac{12\ln(3A/p)}{Np_{\min}}}\right) \geqslant 1 - p.$$

D Episodic Learning: Proofs

In this section, we prove Theorem 4.1. Our proof adapts the approach of Jin et al. [Jin+19] for ATST-MDPs with geometric horizons.

For notational convenience, let $\mathbf{s}_u^k = \emptyset$ for all $k \in [K]$ and $u > B^k + 1$. Let $\overline{R}^\tau = \min\{R^\tau, H\}$.

For burst-dependent policy $\pi=(\pi_u)_{u=1}^\infty$ and $n\in\mathbb{N}$, let $\pi_{(n)}=(\pi_{u+n-1})_{u=1}^\infty$ denote the burst-dependent policy obtained by shifting the original policy by n-1 data-bursts ahead. Then, we introduce notation $K_u^{\boldsymbol{\pi}}=K^{\boldsymbol{\pi}_{(u)}}$ and $V_u^{\boldsymbol{\pi}}=V^{\boldsymbol{\pi}_{(u)}}$.

D.1 Some Technical Lemmas

In this sections, we state some technical lemmas used in the proof of the main result. The proofs of these lemmas are deferred to later subsections.

First, we need the following lemma, which bounds the growth of the estimator's norm.

Lemma D.1 (Bound for \boldsymbol{w}_{u}^{k}). For all $(k, u) \in [K] \times [H-1]$, $\|\boldsymbol{w}_{u}^{k}\|_{2} \leq 4\sqrt{dkH^{3}/\lambda}$.

Proof. For every vector $v \in \mathbb{R}^{2d}$, we have

$$\begin{aligned} |\boldsymbol{v}^{\top}\boldsymbol{w}_{u}^{k}| &= \left|\boldsymbol{v}^{\top}(\Lambda^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [\overline{R}^{\tau} + \sup_{\boldsymbol{a}} K_{u+1}^{k} (\mathbf{s}_{N}^{\tau}, \boldsymbol{a})] \right| \\ &\leq \sum_{\tau=1}^{N^{k}} |\boldsymbol{v}^{\top}(\Lambda^{k})^{-1} \hat{\boldsymbol{\psi}}^{\tau}| \cdot 2H \\ &\leq 2H \cdot \sqrt{\left[\sum_{\tau=1}^{N^{k}} \|\boldsymbol{v}\|_{(\Lambda^{k})^{-1}}^{2}\right] \left[\sum_{\tau=1}^{N^{k}} \|\hat{\boldsymbol{\psi}}^{\tau}\|_{(\Lambda^{k})^{-1}}^{2}\right]} \\ &\leq 2H \cdot \|\boldsymbol{v}\|_{2} \sqrt{kH/\lambda} \cdot \sqrt{2d}, \end{aligned}$$

where the last step follows from the fact that $N^k \leqslant kH$ and Fact D.9.

Based on this lemma, we can establish the following concentration result.

Lemma D.2. Under the setting of Theorem 4.1, let c_{ρ} be the constant parameterizing ρ (i.e., $\rho = c_{\rho} \cdot dH \sqrt{\iota}$). There exists an absolute constant C, independent of c_{ρ} , such that for all fixed $p \in [0,1]$, if we let \mathcal{E} denote the event that

$$\forall (k, u) \in [K] \times [H - 1]: \quad \left\| \sum_{\tau=1}^{N^k} \widehat{\boldsymbol{\psi}}^{\tau} [V_{u+1}^k(\mathbf{s}_N^{\tau}) - \mathbb{P} V_{u+1}^k(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})] \right\|_{(\Lambda^k)^{-1}} \leqslant C \cdot \frac{d}{1 - \gamma} \sqrt{\chi},$$

$$\forall k \in [K]: \quad \left\| \sum_{\tau=1}^{N^k} \widehat{\boldsymbol{\psi}}^{\tau} [\overline{R}^{\tau} - \mathbb{E} [\overline{R}^{\tau} | \mathbf{s}^{\tau}, \boldsymbol{a}^{\tau}]] \right\|_{(\Lambda^k)^{-1}} \leqslant C \cdot H d^{1/2} \sqrt{\iota}$$

where $\chi = \log(2(c_{\rho} + 1)dKH/p)$, then $\mathbb{P}(\mathcal{E}) \geqslant 1 - p/2$

See Section D.3 for the proof of this lemma.

To further simplify the notations, we let $\epsilon_2 = \epsilon \cdot 5\rho\sqrt{KH}$. Note that $\epsilon_2 \geqslant \epsilon \|\boldsymbol{w}_u^k\|_2 + \epsilon\rho$ by Lemma D.1. This constant will be used throughout the rest of the proof. Also, let $\psi_u^k = \psi(\mathbf{s}_u^k, \boldsymbol{a}_u^k)$ be equal to $\mathbf{0} \in \mathbb{R}^{2d}$ when $\mathbf{s}_u^k = \varnothing$.

We also need the following two lemmas. The first lemma provides lower bounds on the estimated action-sequence value-functions on the event that the concentration bounds hold true.

Lemma D.3 (UCB). Under the setting of Theorem 4.1., conditioned on event E from Lemma D.2,

$$K_u^k(s, \boldsymbol{a}) \geqslant K^*(s, \boldsymbol{a}) - (H - u) \cdot \epsilon_2$$

for all $(s, \boldsymbol{a}, u, k) \in \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \times [H] \times [K]$.

Additionally, we need the following lemma, which provides a recursive relation on a term arising from the error decomposition.

Lemma D.4 (Recursive formula). For $k \in [K]$, $u \in [H]$, we define

- $\delta_u^k = V_u^k(\mathbf{s}_u^k) V_u^{\boldsymbol{\pi}^k}(\mathbf{s}_u^k),$
- $\bullet \ \zeta_{u+1}^k = \mathbb{E}[\delta_{u+1}^k \mid \mathbf{s}_u^k, \boldsymbol{a}_u^k] \delta_{u+1}^k.$

Then, conditioned on the event \mathcal{E} , we have that for every $(k, u) \in [K] \times [H-1]$:

$$\delta_u^k \le \delta_{u+1}^k + \zeta_{u+1}^k + 2\rho \|\psi_u^k\|_{(\Lambda^k)^{-1}} + \epsilon_2.$$

See Section D.4 for the proof of Lemma D.3 and D.4.

D.2 Proof of Theorem 4.1

Given lemmas in Section D.1, we are ready to prove Theorem 4.1. To start with, let us recall the statement of the theorem.

Theorem 4.1 (Restated). Suppose Algorithm 1 is executed with ϵ -admissible feature map $\widehat{\psi}$ for $\epsilon \leqslant \sqrt{(1-\gamma)/K}$. There exists an absolute constant $c \geqslant 1$, such that, for all fixed $p \in (0,1)$, if we set $H = \lceil \frac{\log(K(1-\gamma)^{-1})}{1-\gamma} \rceil + 1$, $\lambda = 1$, and $\rho = c \cdot dH\sqrt{\iota}$ with $\iota = \log(2dKH/p)$, then with probability at least 1-p, the total regret is at most

$$\widetilde{O}(\sqrt{d^3K(1-\gamma)^{-3}\iota^2} + d^2(1-\gamma)^{-2}\iota + \epsilon \cdot \sqrt{d^2K^3(1-\gamma)^{-5}\iota}).$$

Proof. We condition on the event \mathcal{E} from Lemma D.2, which occurs with probability at least 1 - p/2. Then, using Lemmas D.3 and D.4 and the choice of ϵ_2 , we can write:

$$\mathcal{R}_{K} = \sum_{k=1}^{K} \left[V^{*}(\mathbf{s}_{1}^{k}) - V_{1}^{\pi_{k}}(\mathbf{s}_{1}^{k}) \right] \leqslant \sum_{k=1}^{K} (\delta_{1}^{k} + H\epsilon_{2})$$

$$\leqslant \sum_{k=1}^{K} \sum_{u=1}^{H} \zeta_{u}^{k} + \sum_{k=1}^{K} \delta_{H}^{k} + 2\rho \sum_{k=1}^{K} \sum_{u=1}^{H-1} \|\psi_{u}^{k}\|_{(\Lambda^{k})^{-1}} + 2KH\epsilon_{2}$$

$$\leqslant \sum_{k=1}^{K} \sum_{u=1}^{H} \zeta_{u}^{k} + \sum_{k=1}^{K} \delta_{H}^{k} + 2\rho \sum_{k=1}^{K} \sum_{u=1}^{H-1} \|\hat{\psi}_{u}^{k}\|_{(\Lambda^{k})^{-1}} + 4KH\epsilon_{2}.$$

• To bound the first component, we use Azuma-Hoeffding for the martingale difference sequence $\{\zeta_u^k\}_{u,k}$ (ordered chronologically with respect to rounds/episodes and including $B^k < u \leqslant H$ with $\mathbf{s}_u^k = \varnothing$), which satisfies $|\zeta_u^k| \leqslant \frac{2}{1-\gamma}$. For all $t \geqslant 0$, we have

$$\mathbb{P}\left(\sum_{k=1}^{K} \sum_{u=1}^{H} \zeta_u^k \leqslant t\right) \geqslant 1 - \exp\left(\frac{-t^2}{8KH(1-\gamma)^{-2}}\right).$$

Hence, with probability at least 1 - p/4, we have that

$$\sum_{k=1}^{K} \sum_{u=1}^{H} \zeta_u^k \le \sqrt{8KH(1-\gamma)^{-2}} \cdot \sqrt{\log(4/p)}.$$

• To bound the second component, observe that for each $k \in [K]$

$$\delta_H^k = V_H^k(\mathbf{s}_H^k) - V_H^{\pi^k}(\mathbf{s}_H^k) \leqslant \frac{\mathbb{I}(s_H^k \neq \varnothing)}{1 - \gamma} - 0 \leqslant \frac{\mathbb{I}(H^k \geqslant H)}{1 - \gamma},$$

and use Chernoff inequality for binary indicators $\mathbb{I}(H^k \geqslant H)$. For all $\delta \geqslant 1$, it holds that

$$\mathbb{P}\left(\sum_{k=1}^{K} \mathbb{I}(H^k \geqslant H) > (1+\delta)K\gamma^{H-1}\right) \leqslant \left(\frac{e^{-\delta}}{(1+\delta)^{1+\delta}}\right)^{K\gamma^{H-1}} \\ \leqslant \exp\left(\frac{-\delta^2 K\gamma^{H-1}}{2+\delta}\right) \leqslant \exp(-\delta K\gamma^{H-1}/3).$$

Then, by Fact D.7, with probability at least 1 - p/4, by setting $\delta = \frac{3 \log(4/p)}{K \gamma^{H-1}} \geqslant 1$, it holds that

$$\sum_{k=1}^{K} \delta_{H}^{k} \leq (1+\delta)K\gamma^{H-1}(1-\gamma)^{-1}$$

$$\leq (K\gamma^{H-1} + 3\log(4/p))(1-\gamma)^{-1}$$

$$\leq 6\log(4/p)(1-\gamma)^{-1}.$$

• To bound the third component, let $\Lambda^k_u = \Lambda^k + \sum_{u'=1}^{u-1} \hat{\psi}^k_{u'} (\hat{\psi}^k_{u'})^{\top}$. Then, write the following

$$\begin{split} \sum_{k=1}^K \sum_{u=1}^H \| \hat{\psi}_u^k \|_{(\Lambda^k)^{-1}} & \leqslant \sqrt{H} \cdot \sum_{k=1}^K \sqrt{\sum_{u=1}^H \| \hat{\psi}_u^k \|_{(\Lambda^k)^{-1}}^2} \\ & \stackrel{\text{(a)}}{\leqslant} \sqrt{H} \cdot \sum_{k=1}^K \sqrt{\sum_{u=1}^H 2 \| \hat{\psi}_u^k \|_{(\Lambda^k)^{-1}}^2} \\ & + \sqrt{H} \cdot \sum_{k=1}^K \mathbb{I}(\det(\Lambda^{k+1}) > 2 \det(\Lambda^k)) \sqrt{H/\lambda} \\ & \leqslant \sqrt{2KH} \cdot \sqrt{\sum_{k=1}^K \sum_{u=1}^H (\hat{\psi}_u^k)^\top (\Lambda_u^k)^{-1} \hat{\psi}_u^k} \\ & + \sqrt{H^2/\lambda} \cdot \sum_{k=1}^K \mathbb{I}(\det(\Lambda^{k+1}) > 2 \det(\Lambda^k)) \\ & \stackrel{\text{(b)}}{\leqslant} \sqrt{2KH} \cdot \sqrt{2 \log \left(\frac{\det(\Lambda^{K+1})}{\det(\Lambda^1)}\right)} + \sqrt{H^2\lambda^{-1}} \cdot \log_2 \left(\frac{\det(\Lambda^{K+1})}{\det(\Lambda^1)}\right) \\ & \stackrel{\text{(c)}}{\leqslant} 4\sqrt{KH} \cdot \sqrt{d \log(2KH)} + 4H \cdot d \log(2KH), \end{split}$$

where (a) follows from Fact D.8, (b) from Fact D.10, and (c) from the following inequality

$$\frac{\det(\Lambda^{K+1})}{\det(\Lambda^1)} \leqslant \left(\frac{\lambda_{\max}(\Lambda^{K+1})}{\lambda_{\min}(\Lambda^1)}\right)^{2d} \leqslant \left(\frac{\lambda + KH}{\lambda}\right)^{2d} = (1 + KH)^{2d} \leqslant (2KH)^{2d}.$$

In conclusion, we have that with probability at least 1 - p:

$$\mathcal{R}(K) \leqslant \sqrt{8KH(1-\gamma)^{-2}} \cdot \sqrt{\log(4/p)}$$

$$+ 6\log(4/p)(1-\gamma)^{-1}$$

$$+ 2\rho \cdot \left(4\sqrt{KH} \cdot \sqrt{d\log(2KH)} + 4H \cdot d\log(2KH)\right)$$

$$+ 4KH \cdot 5\epsilon\rho\sqrt{KH}$$

$$\leqslant c_1 \cdot \sqrt{d^3KH^3\iota^2} + c_2 \cdot d^2H^2\iota + c_3 \cdot \epsilon KH \cdot \sqrt{d^2KH^3\iota},$$

for some absolute constants c_1, c_2, c_3 .

D.3 Proof of Lemma D.2

In Theorem 4.1, we have $H = \left\lceil \frac{\log(K(1-\gamma)^{-1})}{1-\gamma} \right\rceil + 1$, $\lambda = 1$, and $\iota = \log(2dKH/p)$.

From Lemma D.1, $\|\boldsymbol{w}_u^k\|_2 \leq 4\sqrt{dkH^3/\lambda}$. Hence, by combining Lemmas D.12 and D.13 for function class $\mathcal{V}(4\sqrt{dkH^3/\lambda},\rho,\lambda)$, we show that for all $\varepsilon>0$, with probability at least 1-p/4: for all $(k,u)\in[K]\times[H-1]$,

$$\left\| \sum_{\tau=1}^{N^k} \widehat{\boldsymbol{\psi}}^{\tau} [V_{u+1}^k(\mathbf{s}_N^{\tau}) - \mathbb{P} V_{u+1}^k(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})] \right\|_{(\Lambda^k)^{-1}}^2 \leqslant \frac{4}{(1-\gamma)^2} \left[d \log \frac{kH+\lambda}{\lambda} + 2d \log \left(1 + \frac{16\sqrt{dkH^3}}{\varepsilon\sqrt{\lambda}} \right) + 4d^2 \log \left(1 + \frac{16\rho^2\sqrt{d}}{\varepsilon^2\lambda} \right) + \log \left(\frac{4}{p} \right) \right] + \frac{8k^2H^2\varepsilon^2}{\lambda}.$$

We set $\lambda=1$ and $\rho=c_{\rho}\cdot dH\sqrt{\iota}$ and pick $\varepsilon=\frac{d}{(1-\gamma)kH}$. Then, there clearly exists absolute constant $C_1>0$, independent of c_{ρ} , such that

$$\left\| \sum_{\tau=1}^{N^k} \widehat{\boldsymbol{\psi}}^\tau [V_{u+1}^k(\mathbf{s}_N^\tau) - \mathbb{P} V_{u+1}^k(\mathbf{s}^\tau, \boldsymbol{a}^\tau)] \right\|_{(\Lambda^k)^{-1}}^2 \leqslant C_1 \cdot \frac{d^2}{(1-\gamma)^2} \log(2(c_\rho+1)dKH/p).$$

For the second part, we will use the concentration of self-normalized process, where $\overline{R}^{\tau}|\mathbf{s}^{\tau}, \mathbf{a}^{\tau} \in [0, H]$ is a H-sub-Gaussian. By applying Theorem D.11, we can find absolute constant $C_2 > 0$ independent of c_{ρ} such that with probability at least 1 - p/4: for all $k \in [K]$,

$$\left\| \sum_{\tau=1}^{N^k} \hat{\psi}^{\tau} [\overline{R}^{\tau} - \mathbb{E}[\overline{R}^{\tau} | \mathbf{s}^{\tau}, \boldsymbol{a}^{\tau}] \right\|_{(\Lambda^k)^{-1}}^2 \leq 4H^2 \left[d \log \left(\frac{kH + \lambda}{\lambda} \right) + \log \left(\frac{4}{p} \right) \right] \leq C_2 \cdot H^2 d \log(2kH/p).$$

Finally, set $C = \sqrt{\max\{C_1, C_2\}}$ to finish the proof.

D.4 Proof of Lemmas D.3 and D.4

The proof relies on the following technical lemma.

Lemma D.5. Under the setting of Theorem 4.1, there exists an absolute constant $c_{\rho} \ge 1$ such that for $\rho = c_{\rho} \cdot dH \sqrt{\iota}$ and arbitrary burst-dependent policy π , on the event \mathcal{E} from Lemma D.2, for all $(x, \boldsymbol{a}, k, u) \in \mathcal{X} \times \mathcal{A}^{\mathbb{N}} \times [K] \times [H-1]$:

$$\langle \boldsymbol{\psi}(x,\boldsymbol{a}), \boldsymbol{w}_{u}^{k} \rangle - K_{u}^{\boldsymbol{\pi}}(x,\boldsymbol{a}) = \mathbb{P}(V_{u+1}^{k} - V_{u+1}^{\boldsymbol{\pi}})(x,\boldsymbol{a}) + \Delta_{u}^{k}(x,\boldsymbol{a}),$$

where $\Delta_u^k(x, \boldsymbol{a})$ satisfies $|\Delta_u^k(x, \boldsymbol{a})| \leq \rho \|\psi(x, \boldsymbol{a})\|_{(\Lambda^k)^{-1}}$.

See Section D.4.1 for the proof of this lemma. Taking this lemma as given, let us now proceed with the proofs of Lemma D.3 and D.4.

Proof of Lemma D.3. We set $K_H^k(s, \mathbf{a}) = \frac{1}{1-\gamma} \geqslant K^*(s, \mathbf{a})$. Moreover, for all $u \in [H-1]$, we have that

$$K_{u}^{k}(s, \boldsymbol{a}) = \langle \hat{\boldsymbol{\psi}}(s, \boldsymbol{a}), \boldsymbol{w}_{u}^{k} \rangle + \rho \| \hat{\boldsymbol{\psi}}(s, \boldsymbol{a}) \|_{(\Lambda^{k})^{-1}}$$

$$\geqslant \langle \boldsymbol{\psi}(s, \boldsymbol{a}), \boldsymbol{w}_{u}^{k} \rangle + \rho \| \boldsymbol{\psi}(s, \boldsymbol{a}) \|_{(\Lambda^{k})^{-1}} - (\epsilon \| \boldsymbol{w}_{u}^{k} \|_{2} + \rho \epsilon / \sqrt{\lambda})$$

$$\stackrel{\text{(a)}}{\geqslant} K^{*}(s, \boldsymbol{a}) + \mathbb{P}(V_{u+1}^{k} - V^{*})(s; \boldsymbol{a}) - \epsilon_{2}$$

$$\geqslant K^{*}(s, \boldsymbol{a}) + \inf_{s', \boldsymbol{a'}} (K_{u+1}^{k} - K^{*})(s', \boldsymbol{a'}) - \epsilon_{2},$$

where (a) follows from Lemmas D.5 and the choice of ϵ_2 .

Then, the statement follows by trivial induction over u from u = H to u = 1.

Proof of Lemma D.4. We can write the following by Lemma D.5 for all s, a:

$$K_{u}^{k}(s, \boldsymbol{a}) - K_{u}^{\boldsymbol{\pi}^{k}}(s, \boldsymbol{a}) = \langle \hat{\boldsymbol{\psi}}(s, \boldsymbol{a}), \boldsymbol{w}_{u}^{k} \rangle + \rho \| \hat{\boldsymbol{\psi}}(s, \boldsymbol{a}) \|_{(\Lambda^{k})^{-1}} - \langle \boldsymbol{\psi}(s, \boldsymbol{a}), \boldsymbol{w}_{u}^{\boldsymbol{\pi}^{k}} \rangle$$

$$\leq \langle \boldsymbol{\psi}(s, \boldsymbol{a}), \boldsymbol{w}_{u}^{k} \rangle + \rho \| \boldsymbol{\psi}(s, \boldsymbol{a}) \|_{(\Lambda^{k})^{-1}} - \langle \boldsymbol{\psi}(s, \boldsymbol{a}), \boldsymbol{w}_{u}^{\boldsymbol{\pi}^{k}} \rangle + \epsilon_{2}$$

$$\leq \mathbb{P}(V_{u+1}^{k} - V_{u+1}^{\boldsymbol{\pi}^{k}})(s, \boldsymbol{a}) + 2\rho \| \boldsymbol{\psi}(s, \boldsymbol{a}) \|_{(\Lambda^{k})^{-1}} + \epsilon_{2}.$$

From the choice of π^k , we have that

$$\begin{split} \delta_{u}^{k} &= K_{u}^{k}(\mathbf{s}_{u}^{k}, \boldsymbol{a}_{u}^{k}) - K_{u}^{\boldsymbol{\pi}^{k}}(\mathbf{s}_{u}^{k}, \boldsymbol{a}_{u}^{k}) \\ &\leq \mathbb{P}(V_{u+1}^{k} - V_{u+1}^{\boldsymbol{\pi}^{k}})(\mathbf{s}_{u}^{k}, \boldsymbol{a}_{u}^{k}) + 2\rho \|\boldsymbol{\psi}(\mathbf{s}_{u}^{k}, \boldsymbol{a}_{u}^{k})\|_{(\Lambda^{k})^{-1}} + \epsilon_{2} \\ &= \delta_{u+1}^{k} + \zeta_{u+1}^{k} + 2\rho \|\boldsymbol{\psi}_{u}^{k}\|_{(\Lambda^{k})^{-1}} + \epsilon_{2}. \end{split}$$

Note that this holds even when $\mathbf{s}_u^k = \varnothing$, as $0 \leqslant \epsilon_2$.

D.4.1 Proof of Lemma D.5

We first state and prove the following lemma.

Lemma D.6 (Burst-dependent version of Theorem 3.2). *Under Assumption 2.1, for arbitrary burst-dependent policy* $\boldsymbol{\pi} = (\pi_u)_{u=1}^{\infty}$ *and* $u \in \mathbb{N}$ *, it holds that: for all* $(x, a) \in \mathcal{X} \times \mathcal{A}^{\mathbb{N}}$ *,*

$$K_u^{\boldsymbol{\pi}}(x, \boldsymbol{a}) = \langle \boldsymbol{\psi}(x, \boldsymbol{a}), \, \boldsymbol{w}_u^{\boldsymbol{\pi}} \rangle,$$

where
$$\boldsymbol{w}_{u}^{\pi}=2\begin{bmatrix} \boldsymbol{\theta}/(1-\gamma) \\ \int_{\mathcal{S}}V_{u+1}^{\pi}(s)d\boldsymbol{\mu}(s) \end{bmatrix}$$
 satisfies $\|\boldsymbol{w}_{u}^{\pi}\|\leqslant \frac{4\sqrt{d}}{1-\gamma}$.

Proof. Follows by decomposition $K_u^{\pi} = R + \mathbb{P}V_{u+1}^{\pi}$ and Theorem C.1.

Now we turn to the proof of Lemma D.5. As $(\psi^{\tau})^{\top} w_u^{\pi} = K_u^{\pi}(\mathbf{s}^{\tau}, \mathbf{a}^{\tau})$ by Lemma D.6, we have the following

$$\begin{split} \boldsymbol{w}_{u}^{k} - \boldsymbol{w}_{u}^{\pi} &= (\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [\overline{R}^{\tau} + V_{u+1}^{k}(\mathbf{s}_{N}^{\tau})] - \boldsymbol{w}_{u}^{\pi} \\ &= (\boldsymbol{\Lambda}^{k})^{-1} \left\{ -\lambda \boldsymbol{w}_{u}^{\pi} + \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [\overline{R}^{\tau} + V_{u+1}^{k}(\mathbf{s}_{N}^{\tau}) - K_{u}^{\pi}(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})] \right\} \\ &+ (\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} (\boldsymbol{\psi}^{\tau} - \hat{\boldsymbol{\psi}}^{\tau})^{\top} \boldsymbol{w}_{u}^{\pi} \\ &= \underbrace{-\lambda (\boldsymbol{\Lambda}^{k})^{-1} \boldsymbol{w}_{u}^{\pi}}_{q_{1}} + \underbrace{(\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [V_{u+1}^{k}(\mathbf{s}_{N}^{\tau}) - \mathbb{P}V_{u+1}^{k}(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})]}_{q_{2}} \\ &+ \underbrace{(\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [\mathbb{P}(V_{u+1}^{k} - V_{u+1}^{\pi})(\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau})] + \underbrace{(\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [\overline{R}^{\tau} - \mathbb{E}[\overline{R}^{\tau}|\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau}]]}_{q_{3}} \\ &+ \underbrace{(\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} [\mathbb{E}[\overline{R}^{\tau}|\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau}] - \mathbb{E}[R^{\tau}|\mathbf{s}^{\tau}, \boldsymbol{a}^{\tau}]]}_{q_{5}} + \underbrace{(\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau} (\boldsymbol{\psi}^{\tau} - \hat{\boldsymbol{\psi}}^{\tau})^{\top} \boldsymbol{w}_{u}^{\pi}}_{q_{6}}. \end{split}$$

We bound these six components separately. Note that

$$\begin{aligned} |\psi(x, \boldsymbol{a})^{\top} (\Lambda^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\psi}^{\tau}| &\leq \sum_{\tau=1}^{N^{k}} |\psi(x, \boldsymbol{a})^{\top} (\Lambda^{k})^{-1} \hat{\psi}^{\tau}| \\ &\leq \left[\sum_{\tau=1}^{N^{k}} \|\psi(x, \boldsymbol{a})\|_{(\Lambda^{k})^{-1}}^{2} \right]^{1/2} \left[\sum_{\tau=1}^{N^{k}} \|\hat{\psi}^{\tau}\|_{(\Lambda^{k})^{-1}}^{2} \right]^{1/2} \\ &\leq \sqrt{kH} \|\psi(x, \boldsymbol{a})\|_{(\Lambda^{k})^{-1}} \cdot \sqrt{d} \\ &= \sqrt{dkH} \cdot \|\psi(x, \boldsymbol{a})\|_{(\Lambda^{k})^{-1}} .\end{aligned}$$

• To bound q_1 , using Lemma D.6, write

$$\begin{split} |\langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{q}_1 \rangle| &\leqslant \lambda \, \|\boldsymbol{w}_u^{\boldsymbol{\pi}}\|_{(\Lambda^k)^{-1}} \, \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\Lambda^k)^{-1}} \\ &\leqslant \sqrt{\lambda} \, \|\boldsymbol{w}_u^{\boldsymbol{\pi}}\|_2 \, \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\Lambda^k)^{-1}} \leqslant \frac{4\sqrt{d\lambda}}{1-\gamma} \cdot \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\Lambda^k)^{-1}} \,. \end{split}$$

• To bound q_2 and q_4 , we use event $\mathcal E$ so that

$$|\langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{q}_2 + \boldsymbol{q}_4 \rangle| \leqslant C \cdot dH \sqrt{\chi} \cdot ||\boldsymbol{\psi}(x, \boldsymbol{a})||_{(\Lambda^k)^{-1}},$$

for some absolute constant C > 0 independent of c_{ρ} .

• To bound q_3 , using Theorem C.1, observe that for some vector v such that $||v||_2 \leqslant \frac{8\sqrt{d}}{1-\gamma}$:

$$\mathbb{P}(V_{n+1}^k - V_{n+1}^{\pi})(x; \boldsymbol{a}) = \langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{v} \rangle.$$

Then, we can write

$$\langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{q}_{3} \rangle = \langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{v} \rangle - \underbrace{\lambda \boldsymbol{\psi}(x, \boldsymbol{a})^{\top} (\Lambda^{k})^{-1} \boldsymbol{v}}_{c_{1}} + \underbrace{\boldsymbol{\psi}(x, \boldsymbol{a})^{\top} (\Lambda^{k})^{-1} \sum_{\tau=1}^{N^{k}} \widehat{\boldsymbol{\psi}}^{\tau} (\boldsymbol{\psi}^{\tau} - \widehat{\boldsymbol{\psi}}^{\tau})^{\top} \boldsymbol{v}}_{c_{2}},$$

where c_1, c_2 can be bounded as follows:

$$|c_{1}| \leq \sqrt{\lambda} \|\boldsymbol{v}\|_{2} \|\boldsymbol{\psi}(x,\boldsymbol{a})\|_{(\Lambda^{k})^{-1}} \leq \frac{8\sqrt{d\lambda}}{1-\gamma} \cdot \|\boldsymbol{\psi}(x,\boldsymbol{a})\|_{(\Lambda^{k})^{-1}}$$

$$|c_{2}| \leq |\boldsymbol{\psi}(x,\boldsymbol{a})^{\top} (\Lambda^{k})^{-1} \sum_{\tau=1}^{N^{k}} \hat{\boldsymbol{\psi}}^{\tau}| \cdot \epsilon \|\boldsymbol{v}\|_{2}$$

$$\leq \sqrt{dkH} \cdot \|\boldsymbol{\psi}(x,\boldsymbol{a})\|_{(\Lambda^{k})^{-1}} \cdot \epsilon \cdot \frac{8\sqrt{d}}{1-\gamma} \leq 8\sqrt{\epsilon^{2}d^{2}kH(1-\gamma)^{-2}} \cdot \|\boldsymbol{\psi}(x,\boldsymbol{a})\|_{(\Lambda^{k})^{-1}}.$$

• To bound q_5 , note that, as rewards are bounded to [0,1], we have

$$|\mathbb{E}[\overline{R}^{\tau}|\mathbf{s}^{\tau},\boldsymbol{a}^{\tau}] - \mathbb{E}[R^{\tau}|\mathbf{s}^{\tau},\boldsymbol{a}^{\tau}]]| \leqslant \gamma^{H}(1-\gamma)^{-1}.$$

By Fact D.7, for $H \geqslant \frac{\log(K(1-\gamma)^{-1})}{1-\gamma}$, $\gamma^H \leqslant \frac{1}{\sqrt{KH}}$, so we have

$$\begin{aligned} |\langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{q}_5 \rangle| &\leqslant \frac{\gamma^H}{1 - \gamma} \cdot |\boldsymbol{\psi}(x, \boldsymbol{a})^\top (\Lambda^k)^{-1} \sum_{\tau = 1}^{N^k} \widehat{\boldsymbol{\psi}}^\tau | \\ &\leqslant \frac{\sqrt{dkH}}{(1 - \gamma)\sqrt{KH}} \cdot \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\Lambda^k)^{-1}} \leqslant dH \cdot \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\Lambda^k)^{-1}} \end{aligned}$$

• To bound q_6 , we write

$$\begin{aligned} |\langle \boldsymbol{\psi}(x, \boldsymbol{a}), \boldsymbol{q}_{6} \rangle| &\leq \epsilon \|\boldsymbol{w}_{u}^{\boldsymbol{\pi}}\|_{2} \cdot |\boldsymbol{\psi}(x, \boldsymbol{a})^{\top} (\boldsymbol{\Lambda}^{k})^{-1} \sum_{\tau=1}^{N^{k}} \widehat{\boldsymbol{\psi}}^{\tau}| \\ &\leq \epsilon \cdot \frac{4\sqrt{d}}{1-\gamma} \cdot \sqrt{dkH} \cdot \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\boldsymbol{\Lambda}^{k})^{-1}} \\ &\leq 4\sqrt{\epsilon^{2} d^{2} k H (1-\gamma)^{-2}} \cdot \|\boldsymbol{\psi}(x, \boldsymbol{a})\|_{(\boldsymbol{\Lambda}^{k})^{-1}} \end{aligned}$$

To sum up, for our choice of $\lambda=1$ and $\epsilon\leqslant\sqrt{\frac{1-\gamma}{K}}$ we have that

$$\Delta_u^k(x, \boldsymbol{a}) \leq (25 + C) \cdot dH \sqrt{\chi} \cdot ||\psi(x, \boldsymbol{a})||_{(\Lambda^k)^{-1}}.$$

Finally, observe that c_{ρ} appears in χ only under the logarithm and C is an absolute constant. Therefore, we can select c_{ρ} as an absolute constant large enough such that for $\iota \geqslant \log(2)$, $c_{\rho} \cdot \sqrt{\iota} \geqslant (25 + C)\sqrt{\iota + \log(c_{\rho} + 1)}$, i.e. $\rho = c_{\rho} \cdot dH\sqrt{\iota} \geqslant (25 + C)dH\sqrt{\chi}$ for all K, H, d, p.

D.5 Some Basic Facts

In this section, we collect some basic algebraic facts used in the proofs.

Fact D.7. For $n \geqslant \frac{\log(K(1-\gamma)^{-1})}{1-\gamma}$ it holds that $\gamma^n \leqslant \min\{\frac{1-\gamma}{K}, \frac{1}{n(1-\gamma)}\} \leqslant \frac{1}{\sqrt{Kn}}$.

Proof. As $\log(1/x) \ge 1 - x$ for x > 0, we can write

$$\gamma^n = \exp\left(-H\log(1/\gamma)\right) \leqslant \exp\left(-\log(K(1-\gamma)^{-1})\right) = \tfrac{1-\gamma}{K}.$$

Moreover, as $1/x \ge e^{-x}$ for x > 0, we also have

$$\gamma^n = \exp\left(-n\log(1/\gamma)\right) \leqslant \frac{1}{n\log(1/\gamma)} \leqslant \frac{1}{n(1-\gamma)}$$
.

The final inequality follows trivially.

Fact D.8. Let $A, B \in \mathbb{R}^{d \times d}$ be positive definite matrices and $\mathbf{x} \in \mathbb{R}^d$. If $A \geq B$, then

$$\|\boldsymbol{x}\|_A \leqslant \|\boldsymbol{x}\|_B \sqrt{\frac{\det(A)}{\det(B)}}.$$

Fact D.9. Let $(\mathbf{x}_n)_{n=1}^N$ be an \mathbb{R}^D -valued sequence and $\lambda > 0$. Then, for $\Lambda_N = \lambda I + \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^{\top}$, it holds that

$$\sum_{n=1}^{N} \|x_n\|_{(\Lambda_N)^{-1}}^2 \leqslant D.$$

Proof. Proof is exactly the same as in Lemma D.1 from [Jin+19].

Fact D.10 ([APS11]). Let $(x_n)_{n=1}^{\infty}$ be an \mathbb{R}^D -valued sequence such that $||x_n||_2 \leq 1$ for every $n \in \mathbb{N}$. Let $\Lambda_0 \in \mathbb{R}^{D \times D}$ satisfy $\lambda_{\min}(\Lambda_0) \geq 1$ and define $\Lambda_N = \Lambda_0 + \sum_{n=1}^N x_n x_n^{\top}$ for every $n \in \mathbb{N}$. Then, it holds that: for all $N \in \mathbb{N}$,

$$\log \left[\frac{\log(\Lambda_N)}{\log(\Lambda_0)} \right] \leqslant \sum_{n=1}^N \|\boldsymbol{x}_n\|_{\Lambda_{n-1}^{-1}}^2 \leqslant 2 \log \left[\frac{\log(\Lambda_N)}{\log(\Lambda_0)} \right].$$

D.6 Concentration Inequalities

Theorem D.11 (Self-Normalized Bound for Vector-Valued Martingales, [APS11]). Let $\{\varepsilon_{\tau}\}_{\tau=1}^{\infty}$ be a \mathbb{R} -valued stochastic process with corresponding filtration $\{\mathcal{F}_{\tau}\}_{\tau=0}^{\infty}$, such that $\varepsilon|\mathcal{F}_{\tau-1}$ be zero-mean and σ -sub-Gaussian for every $\tau \geq 1$. Let $\{\zeta_{\tau}\}_{\tau=1}^{\infty}$ be an \mathbb{R}^{D} -valued stochastic process where $\zeta_{\tau} \in \mathcal{F}_{\tau-1}$. Let $\Lambda \in \mathbb{R}^{D \times D}$ be a positive definite matrix and define $\Lambda_{N} = \lambda I + \sum_{\tau=1}^{N} \zeta_{\tau} \zeta_{\tau}^{\top}$ for $N \geq 1$. Then, for all $\delta > 0$, with probability at least $1 - \delta$, it holds that

$$\forall N \geqslant 0: \qquad \left\| \sum_{\tau=1}^{N} \zeta_{\tau} \varepsilon_{\tau} \right\|_{(\Lambda_{N})^{-1}} \leqslant 2\sigma^{2} \log \left(\frac{\det(\Lambda_{N})^{1/2} \det(\Lambda)^{-1/2}}{\delta} \right).$$

Lemma D.12. Let $V \subset \mathbb{R}^S$ be an arbitrary function class such that, for every $V \in V$, $\sup_s |V(s)| \leq \frac{1}{1-\gamma}$. Let $\{s_\tau\}_{\tau=1}^{\infty}$ be a stochastic process on state space S with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^{\infty}$. Let $\{\zeta_\tau\}_{\tau=1}^{\infty}$ be an \mathbb{R}^D -valued stochastic process where $\zeta_\tau \in \mathcal{F}_{\tau-1}$ and $\|\zeta_\tau\|_2 \leq 1$. Let $\Lambda_N = \lambda I + \sum_{\tau=1}^N \zeta_\tau \zeta_\tau^\intercal$. Then, for all $\delta > 0$, with probability at least $1 - \delta$, it holds that for all $N \geq 0$ and $V \in V$

$$\left\| \sum_{\tau=1}^{N} \zeta_{\tau} \left\{ V(s_{\tau}) - \mathbb{E}[V(s_{\tau}) \mid \mathcal{F}_{\tau-1}] \right\} \right\|_{(\Lambda_{N})^{-1}}^{2} \leq \frac{4}{(1-\gamma)^{2}} \left[\frac{D}{2} \log \left(\frac{N+\lambda}{\lambda} \right) + \log \left(\frac{N_{\varepsilon}}{\delta} \right) \right] + \frac{8N^{2} \varepsilon^{2}}{\lambda},$$

where $\mathcal{N}_{\varepsilon}$ is the ε -covering number of \mathcal{V} with respect to $\operatorname{dist}(V,V')=\sup_{s}|V(s)-V'(s)|$.

Proof. The result follows by applying Theorem D.11 for each element in the ε -covering and using the union bound for the left-hand side, as was done in the proof of Lemma D.4 from [Jin+19].

Lemma D.13 (Covering number bound, [Jin+19]). Let $\zeta: \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^D$ be an arbitrary state-action-sequence feature map, such that $\sup_{s,\boldsymbol{a}} \|\zeta(s,\boldsymbol{a})\|_2 \leq 1$. For $L,B,\lambda>0$, let $\mathcal{V}(L,B,\lambda)$ denote the following parametric class of mappings from \mathcal{S} to $[0,\frac{1}{1-\gamma}]$:

$$\left\{V(.) = \min\{\tfrac{1}{1-\gamma}, \sup_{\boldsymbol{a} \in \mathcal{A}^{\mathbb{N}}} \boldsymbol{\zeta}(., \boldsymbol{a})^{\top} \boldsymbol{w} + \rho \, \|\boldsymbol{\zeta}(., \boldsymbol{a})\|_{\Lambda^{-1}}\} : \|\boldsymbol{w}\|_{2} \leqslant L, \rho \in [0, B], \Lambda \geq \lambda I\right\}.$$

Then, the covering number $\mathcal{N}_{\varepsilon}$ of $\mathcal{V}(L, B, \lambda)$ with respect to $\operatorname{dist}(V, V') = \sup_{s \in \mathcal{S}} |V(s) - V'(s)|$ satisfies

$$\log \mathcal{N}_{\varepsilon} \leq D \log(1 + 4L/\varepsilon) + D^2 \log \left(1 + 8D^{1/2}B^2/(\lambda \varepsilon^2)\right).$$

Proof. Accounting for the fact that we use a different feature map $\zeta : \mathcal{S} \times \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^{D}$, the proof follows similarly to Lemma D.6 from [Jin+19].