Confidence Intervals for the Return Process in Markov Decision Processes

Abstract

In this work, we derive confidence intervals for the return process in discounted reward Markov Decision Processes with continuous state and action spaces. These confidence bounds depend only on the statistics of the value function, which may be derived using dynamic programming. In the two special cases of MDPs with uniformly bounded value functions and MDPs with linear structures, simpler confidence intervals are provided for the return process. Finally, we study the effect of epistemic uncertainty on the derived confidence intervals. Numerical examples are provided to show how these bounds may be used in practice.

1 INTRODUCTION

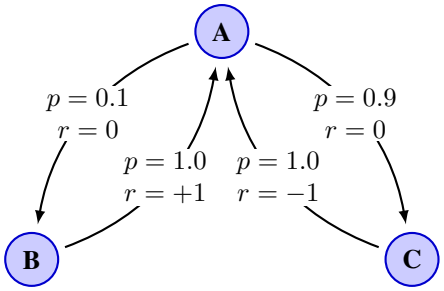
Uncertainty is ubiquitous in data-driven applications: a finite number of samples, noise in the environment and observation, and model mismatch all contribute to unavoidable uncertainty. Confidence intervals (CIs) are a fundamental tool in mathematical statistics for quantifying uncertainty from a finite set of data. For a chosen level δ, these bounds identify an interval that contains the quantity of interest with probability at least $1-\delta$. These bounds are used extensively in statistical studies related to clinical trials [Jennison and Turnbull, 1999], hypothesis testing theory [Lehmann and Romano, 2005], econometrics [Imbens and Rubin, 2015], reliability engineering [Nelson, 2021], and risk-sensitive decision making [Howard and Matheson, 1972]. Despite their importance, confidence intervals are less studied within sequential decision-making frameworks; in particular, there are very few works that characterize CIs for the return process in Markov Decision Processes (MDPs).

MDP is a mathematical framework widely used to model sequential decision-making under uncertainty. This model serves as the theoretical basis for reinforcement learning (RL) algorithms. In many high-stakes applications, such as healthcare, finance, and safety-critical systems, optimizing for optimal performance alone does not provide sufficient guarantees. As a result, there has been substantial effort in the literature to establish performance guarantees beyond the expectation of returns. These efforts may be categorized into distributional reinforcement learning (e.g., Bellemare et al. [2017], Dabney et al. [2018b,a], Rowland et al. [2019], Sobel [1982], Bellemare et al. [2023]), risk-averse and safe reinforcement learning (e.g., Whittle [1990], Castro et al. [2012], Tamar et al. [2014], García and Fernández [2015], Chow et al. [2018]), chance-constrained MDPs (e.g., Delage and Mannor [2010]), and asymptotic analysis of Markov reward processes (e.g., Hernández-Lerma and Lasserre [1996, 2012], Meyn and Tweedie [2012]).

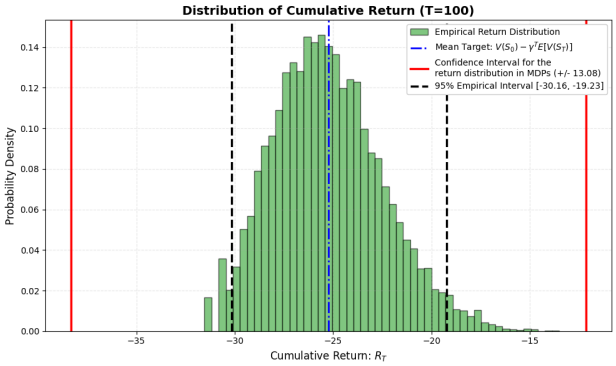
In this work, we choose a different viewpoint and establish CIs for the return process and treat these intervals as performance guarantees. Traditionally, CIs are used in sequential decision-making problems for characterizing the epistemic uncertainty over the unknown mean and integrating the CIs in the reinforcement learning algorithms (e.g., the class of upper-confidence bounds in Lai and Robbins [1985]). The other theoretical use case for CIs is in the study of off-policy evaluation of RL algorithms (e.g., in Hanna et al. [2016], Dai et al. [2020], Shi et al. [2024], Thomas et al. [2015]). We, on the other hand, focus on deriving CIs for the return process. Our analysis is motivated by the fact that in many applications only a single trajectory of return is observed by the agent and therefore, high-probability CIs might be more beneficial to the agent compared to the asymptotic mean of return (i.e., value function).

The sample path behavior of the return process is mostly studied in the average reward MDP framework (e.g. Hernández-Lerma and Lasserre [2012]). In such a setup, for a fixed policy π, the problem reduces to a Markov reward process. For this process, the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) are derived for the return process. Such results are also derived for the functionals of Markov chains [Meyn and Tweedie, 2012].

Submitted to the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026). **To be used for reviewing only**.



(a) MDP Dynamics: Skewed Policy (10/90)



(b) 95 percentile of return at $T = 100$ versus confidence intervals.

Figure 1: (a) The transition dynamics, where State A transitions to B with probability 0.1 and C with 0.9. (b) The empirical distribution of the cumulative return R_T at $T = 100$. The theoretical Freedman bound (solid red lines) is centered around the expected target $V(S_0) - \gamma^T \mathbb{E}[V(S_T)]$ (blue dash-dotted line) and successfully envelopes the empirical 95% interval (black dashed lines).

A martingale approach for average and discounted reward setup is presented in Sayedana et al. [2024], but the analysis is restricted to finite-state setup.

In the discounted reward setting, significant effort has been directed toward learning the distribution of the *asymptotic discounted return* in the distributional RL (DRL) framework [Bellemare et al., 2023]. In general, there are two shortcomings in the DRL framework: (i) it often requires a large number of computations to apply a distributional Bellman update on the distributions and (ii) it only describes the *asymptotic* distribution of return. In this work, we address these two problems by providing CIs for the return distribution over any finite time. To illustrate the difference, consider the Markov reward process illustrated in Figure 1a. The return distribution at the step $T = 100$ is plotted along with the CIs derived in this paper. The distribution at this finite time is approximated using Monte Carlo simulation. The CIs derived in our work are plotted for reference. Although these CIs are not exactly matching the 95% empirical percentile of the distribution (which is common for theoretical CIs), they are computed using only the statistics of the value function (i.e., via dynamic programming algorithms).

Contributions: We make the following contributions. We provide novel CIs for the return process in the general *continuous state and action* MDPs under mild assumptions. We derive CIs that are analogous to Bernstein's inequality, which use variance of the value function and may be potentially tighter than previously established bounds. We study the simplification of these CIs in the case of linear MDPs. At last, we study the effect of epistemic uncertainty on the derived CIs.

Notation: For a sequence of random variables $\{S_t\}_{t\geq 0}$, we use $S_{0:t}$ as shorthand for $\{S_0, \dots, S_t\}$, and $s_{0:t}$ for realization of $S_{0:t}$. We denote the σ-field generated by $S_{0:t}$ as $\sigma(S_{0:t})$. The notation $S \sim \rho$ denotes that the random variable S is sampled from the distribution ρ. δ_x denotes a Dirac delta distribution centered on x.

Given a finite set $\mathcal{S}$, let $|\mathcal{S}|$ denote its cardinality and let $\Delta(\mathcal{S})$ denote the set of probability measures on $\mathcal{S}$. For a function $f : \mathcal{S} \to \mathbb{R}$, sup-norm $\|f\|_\infty$ is defined as $\|f\|_\infty = \sup_{s\in\mathcal{S}} |f(s)|$, and the span of the function $\mathrm{sp}(f)$ is defined as $\mathrm{sp}(f) := \sup_{s\in\mathcal{S}} f(s) - \inf_{s\in\mathcal{S}} f(s)$

2 PROBLEM FORMULATION

2.1 DESCRIPTION OF ENVIRONMENT

A discounted MDP is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where

- $\mathcal{S}$ denotes the state space. We assume $\mathcal{S} \subseteq \mathbb{R}^p$. The state at time t is denoted by $S_t \in \mathcal{S}$.
- $\mathcal{A}$ denotes the action space. We assume $\mathcal{A} \subseteq \mathbb{R}^q$. The action at time t is denoted by $A_t \in \mathcal{A}$.
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the transition kernel, i.e., for any measurable set $\mathcal{I} \subset \mathcal{S}$, we have

$$\mathbb{P}(S_{t+1} \in \mathcal{I} | S_t = s_t, A_t = a_t) = P(\mathcal{I} | s_t, a_t).$$

- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the instantaneous reward function.
- γ is the discount factor.

A function $\pi : \mathcal{S} \to \mathcal{A}$ is called a (deterministic stationary) policy and it determines the agent's action at time t. We denote the space of all such policies by Π. A fixed policy $\pi \in \Pi$ induces a return process (cumulative reward up to time T) on the probability space defined as

$$R_T^\pi := \sum_{t=0}^{T-1} \gamma^t r(S_t, A_t), \quad \text{where,} \quad A_t = \pi(S_t).$$

Let $R_\infty^\pi = \lim_{T\to\infty} R_T^\pi$ denote the asymptotic return. The mean of the asymptotic return is defined as the state value function $V^\pi : \mathcal{S} \to \mathbb{R}$ given by:

$$V^\pi(s) := \mathbb{E}^\pi \left[R_\infty^\pi \middle| S_0 = s \right], \quad \forall s \in \mathcal{S}, \tag{1}$$

where $\mathbb{E}^\pi$ is the expectation operator over all random variables induced by π. In general, the existence of the value function in non-compact state and action spaces is not guaranteed and further sufficient conditions are required for the existence of the value function. However, in this work, we restrict our attention to the set of policies for which the value function *exists* and satisfies the Bellman equation.

Assumption 1. *For the MDP $\mathcal{M}$, we assume there exists a non-empty set $\Pi_{\mathscr{B}} \subseteq \Pi$, such that for any policy $\pi \in \Pi_{\mathscr{B}}$, the corresponding value function V^π exists and satisfies the following fixed-point equation, known as the Bellman equation.*

$$V^\pi(s) = r(s, \pi(s)) + \gamma \int_{\mathcal{S}} V^\pi(s') P(ds'|s, \pi(s)), \quad (2)$$

for all $s \in \mathcal{S}$.

In many RL applications the reward function is uniformly bounded over states and that would imply Assumption 1. However, in certain continuous control problems with unbounded cost functions, Assumption 1 is imposed to guarantee that there at least exists a policy π that induces a bounded cumulative cost. This assumption is closely related to the notion of stabilizability for models with state-dependent cost functions. We impose Assumption 1 to ensure that our results apply to this class of continuous-control problems.

The optimal value function $V^* : \mathcal{S} \to \mathbb{R}$ is defined as $V^*(s) = \sup_{\pi \in \Pi} V^\pi(s)$, for all $s \in \mathcal{S}$. A policy π^* is called optimal if $V^{\pi^*}(s) = V^*(s)$, for all $s \in \mathcal{S}$. Under sufficient conditions mentioned in Hernández-Lerma and Lasserre [1996, Theorem 4.2.3], there exists an optimal policy $\pi^* \in \Pi_{\mathscr{B}}$.

2.2 STATISTICS OF THE VALUE FUNCTION

To simplify the notation, let $P_\pi(ds'|s) \coloneqq P(ds'|s, \pi(s))$. For any function $f : \mathcal{S} \to \mathbb{R}$, we define the conditional expectation and conditional variance with respect to the measure induced by $P_\pi(\cdot|s)$ as follows:

$$\mathbb{E}_\pi[f(S_+)|s] \coloneqq \int_{\mathcal{S}} f(s') P_\pi(ds'|s),$$
$$\mathbb{V}_\pi[f(S_+)|s] \coloneqq \mathbb{E}_\pi\left(\Big[f(S_+) - \mathbb{E}_\pi[f(S_+)|s] \Big]^2 \Big| s \right).$$

We denote the reachable set under policy π by $\mathscr{R}_\pi$, given by

$$\mathscr{R}_\pi \coloneqq \big\{ s \in \mathcal{S} : \forall \text{ open } U \ni s, \mathbb{P}^\pi(\exists t \geq 0 : S_t \in U) > 0 \big\},$$

where $\mathbb{P}^\pi$ is the probability measure induced by following the policy π. For any policy $\pi \in \Pi_{\mathscr{B}}$, let $\{S_t\}_{t \geq 0}$ denote the sequence of induced states. We define the *value innovation* sequence as:

$$N^\pi_{t+1} \coloneqq V^\pi(S_{t+1}) - \mathbb{E}_\pi[V^\pi(S_{t+1})|S_t], \quad (3)$$

where $V^\pi(\cdot)$ is the value function corresponding to the policy π. We define two key statistics for the value innovation process using $V^\pi(\cdot)$: (i) Maximum absolute deviation with respect to P_π defined as

$$K^\pi \coloneqq \sup_{s \in \mathscr{R}_\pi} \operatorname*{ess\,sup}_{s' \sim P_\pi(\cdot|s)} \big| V^\pi(s') - \mathbb{E}_\pi[V^\pi(S')|s] \big|, \quad (4)$$

and (ii) maximum conditional standard deviation defined as

$$\bar{\sigma}^\pi = \sup_{s \in \mathscr{R}_\pi} \sqrt{\mathbb{V}_\pi[V^\pi(S_+)|s]}. \quad (5)$$

Notice that by definition, the value innovation sequence has zero mean conditioned on the current state, i.e., $\mathbb{E}_\pi[N^\pi_{t+1} \mid S_t] = 0$ and $|N^\pi_t| \leq K^\pi$ almost surely for all $t \in \mathbb{N}$.

3 MAIN RESULTS

We present CIs for the return process under three sets of assumptions. We first derive CIs under the boundedness assumption of K^π. We further simplify the CIs under the uniform boundedness assumption on the value function. By relaxing both of these assumptions, we derive bounds analogous to the Azuma–Hoeffding inequality by only assuming the value innovation process to be a sub-Gaussian process. Finally, in the linear MDP framework, CIs are further simplified and are rewritten in terms of the statistics of the feature map and MMD metric between distributions.

3.1 GENERAL STATE AND ACTION SPACES

For a fixed policy $\pi \in \Pi_{\mathscr{B}}$, let $\{S_t\}_{t \geq 0}$ denote the sequence of states induced by following the policy π. By the definition of value function, we have $\mathbb{E}\big[R^\pi_T + \gamma^T V^\pi(S_T) - V^\pi(S_0)\big] = 0$. We derive CIs for the return process R^π_T centered around the process $V^\pi(S_0) - \gamma^T V^\pi(S_T)$.

Assumption 2. *Let $\Pi_K \subseteq \Pi_{\mathscr{B}}$ denote the set of policies π such that the corresponding maximum absolute deviation K^π is finite. We assume Π_K is non-empty.*

Notice that this assumption does not necessarily imply that the value function $V^\pi(s)$ (or reward function) is uniformly bounded across states and it might potentially hold for unbounded value functions. It only requires that the maximum absolute deviation of the value function $V^\pi(\cdot)$ with respect to measure P_π to be uniformly bounded. The following theorem establishes CIs as a function of K^π.

Theorem 1. *For any policy $\pi \in \Pi_K$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have the following confidence interval:*

$$\Big| R^\pi_T - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T) \big) \Big| \leq K^\pi \sqrt{2 h_T \log \frac{2}{\delta}}, \quad (6)$$

where $h_T = \frac{\gamma^2 - \gamma^{2T+2}}{1 - \gamma^2}$.

The proof is presented in Appendix B.1.

The CI in the previous theorem is derived using only the statistic K^π. This bound is analogous to the Azuma–Hoeffding inequality. We now derive a new CI based on both $\bar{\sigma}^\pi$ and K^π which is analogous to Bernstein inequality. Depending on the relationship between $\bar{\sigma}^\pi$ and K^π either of these CIs might be tighter.

Theorem 2. *For any policy $\pi \in \Pi_K$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have the following confidence interval:*

$$\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| \leq \bar{\sigma}^\pi \sqrt{2h_T \log\frac{2}{\delta} + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\frac{2}{\delta}\right)^2} + \frac{K^\pi}{3}\log\frac{2}{\delta}, \tag{7}$$

where h_T is defined in Theorem 1.

The proof is presented in Appendix B.2.

The CIs derived in the previous theorems characterize the return process R_T^π for any finite time T. By imposing the following assumption on the growth rate of the value function $V^\pi(\cdot)$, we establish CIs for the limiting random variable R_∞^π.

Assumption 3. *Let $\Pi_E \subseteq \Pi_K$ denote the set of policies π such that the corresponding value function satisfies $\lim_{T\to\infty}\gamma^T V^\pi(S_T) = 0$, almost surely. We assume Π_E is non-empty.*

Above assumption is related to the notion of equalizing policies studied in Karatzas and Sudderth [2010]. By imposing Assumption 3, we establish CIs for the limiting return random variable R_∞^π.

Corollary 1. *For any policy $\pi \in \Pi_E$, we have the following confidence intervals:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left|R_\infty^\pi - V^\pi(S_0)\right| \leq K^\pi \sqrt{\frac{2\gamma^2}{1-\gamma^2}\log\frac{2}{\delta}}. \tag{8}$$

2. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left|R_\infty^\pi - V^\pi(S_0)\right| \leq \bar{\sigma}^\pi \sqrt{\frac{2\gamma^2}{1-\gamma^2}\log\frac{2}{\delta} + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\frac{2}{\delta}\right)^2} + \frac{K^\pi}{3}\log\frac{2}{\delta}. \tag{9}$$

The proof is presented in Appendix B.3.

In the next corollary, we derive simplified CIs assuming the uniform boundedness of the value function. In many standard MDP settings, uniformly bounded value functions are induced naturally. When the value function is bounded, we can derive CIs for the return process R_T^π centered around $V^\pi(S_0)$ instead of $V^\pi(S_0) - \gamma^T V^\pi(S_T)$.

Assumption 4. *Given a policy $\pi \in \Pi_{\mathscr{B}}$, we assume the corresponding value function is uniformly bounded, i.e., there exists a constant $V_{\max} > 0$ such that $\|V^\pi\|_\infty \leq V_{\max}$.*

Corollary 2. *Under Assumption 4, for any policy $\pi \in \Pi_{\mathscr{B}}$, we have the following confidence intervals:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left|R_T^\pi - V^\pi(S_0)\right| \leq K^\pi \sqrt{2h_T \log\frac{2}{\delta}} + \gamma^T V_{\max}. \tag{10}$$

2. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left|R_T^\pi - V^\pi(S_0)\right| \leq \bar{\sigma}^\pi \sqrt{2h_T \log\frac{2}{\delta} + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\frac{2}{\delta}\right)^2} + \frac{K^\pi}{3}\log\frac{2}{\delta} + \gamma^T V_{\max}. \tag{11}$$

The proof is presented in Appendix B.4.

Remark 1. *Assumption 4 implies Assumption 2, since a uniformly bounded value function guarantees that the maximum absolute deviation is bounded (i.e., $K^\pi \leq 2V_{\max}$). As a result, we obtain the results of Theorems 1 and 2 by imposing Assumption 4. Furthermore, Assumption 4 trivially implies Assumption 3 as uniform boundedness of the value function implies $\lim_{T\to\infty}\gamma^T V^\pi(S_T) = 0$. As a result, we obtain the results of Corollary 1 by imposing Assumption 4.*

Remark 2. *In the environments where the instantaneous reward function is uniformly bounded by a constant $R_{\max}$ (i.e., $\|r\|_\infty \leq R_{\max}$), Assumption 4 is trivially satisfied since $V_{\max} \leq \frac{R_{\max}}{1-\gamma}$. As a result, in such environments, both Corollary 1 and Corollary 2 hold.*

3.2 UNBOUNDED ENVIRONMENTS WITH SUB-GAUSSIAN VALUE INNOVATIONS

While Assumption 2 requires the absolute deviation K^π to be strictly bounded, many continuous control environments exhibit unbounded noise (e.g., Gaussian transition dynamics). To handle such environments, we can relax the strict boundedness requirement by assuming that the value innovation process is a sequence of sub-Gaussian random variables. Before formalizing our assumption, we briefly recall the standard definition of a sub-Gaussian random variable.

Definition 1. *A random variable X with mean $\mu = \mathbb{E}[X]$ is said to be σ-sub-Gaussian if its moment generating function satisfies:*

$$\mathbb{E}\big[\exp(\lambda(X-\mu))\big] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \quad \forall \lambda \in \mathbb{R}. \quad (12)$$

The constant $\sigma > 0$ is referred to as the variance proxy.

Recall that the value innovation sequence satisfies $|N_t^\pi| \leq K^\pi$ almost surely for all $t \in \mathbb{N}$. In the previous section, this property was used in the derivation of the results. In this section, instead of bounding the absolute deviation of $\{N_t^\pi\}_{t\geq 0}$, we bound its conditional moment generating function using the sub-Gaussian property.

Assumption 5. *Given a policy $\pi \in \Pi_{\mathscr{B}}$, we assume there exists a uniform variance proxy $\nu^\pi > 0$ such that the value innovation sequence $\{N_t^\pi\}_{t\geq 0}$ is conditionally ν^π-sub-Gaussian. That is, for all $\lambda \in \mathbb{R}$ and almost all $s \in \mathscr{R}_\pi$:*

$$\mathbb{E}_\pi\left[\exp(\lambda N_{t+1}^\pi) \;\middle|\; s\right] \leq \exp\left(\frac{\lambda^2(\nu^\pi)^2}{2}\right). \quad (13)$$

This assumption allows us to establish a CI which does not require the boundedness of K^π in Assumption 2 replacing it with sub-Gaussianity of the value innovation sequence.

Theorem 3. *Under Assumption 5, for any policy $\pi \in \Pi_{\mathscr{B}}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have the following confidence interval:*

$$\left|R_T^\pi - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big)\right| \leq \nu^\pi \sqrt{2h_T \log\frac{2}{\delta}}, \quad (14)$$

where $h_T = \frac{\gamma^2-\gamma^{2T+2}}{1-\gamma^2}$.

The proof is presented in Appendix B.5.

3.3 LINEAR MARKOV DECISION PROCESSES

The confidence intervals derived in the previous section depend on key statistics of the value function: K^π, $\bar{\sigma}^\pi$, and ν^π. However, in a continuous state and action MDP, computing these statistics is not a trivial task. In this section, we derive upper bounds for K^π and $\bar{\sigma}^\pi$ in the linear MDP framework. These bounds are derived by exploiting the structure of the linear MDPs.

Consider an MDP $\mathcal{M}$. Let $\phi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ be a feature map satisfying $r(s,a) = \langle \phi(s,a), \theta\rangle$ for a fixed vector $\theta \in \mathbb{R}^d$. Let $\mathcal{B}$ denote the set of all measurable subsets of $\mathcal{S}$, and let $\mu : \mathcal{B} \to \mathbb{R}^d$ denote another feature map satisfying $P(\mathcal{I}|s,a) = \langle \phi(s,a), \mu(\mathcal{I})\rangle$ for all $\mathcal{I} \in \mathcal{B}$. We impose the following standard boundedness assumption on the feature map.

Assumption 6. *$\mathcal{M}$ is a linear MDP and for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $\|\phi(s,a)\|_2 \leq 1$.*

This assumption bounds the reward function and consequently implies the boundedness assumption imposed in Assumption 4. To simplify the notation, let $\phi^\pi(s) := \phi(s,\pi(s))$. A standard result in the linear MDP framework (see, for example, Gabbianelli et al. [2023]) establishes that for any policy $\pi \in \Pi_{\mathscr{B}}$, the value function is linear in the features, i.e.,

$$V^\pi(s) = \langle \phi^\pi(s), \mathbf{w}^\pi\rangle,$$

where $\mathbf{w}^\pi \in \mathbb{R}^d$ is the weight vector for policy π.

The boundedness of the value function and this linear structure allow us to define statistics in the feature space analogous to the statistics defined for the value function.

Definition 2. *We define the maximum absolute deviation of the feature map w.r.t. P_π as K_ϕ^π, and the maximum conditional standard deviation of the feature map w.r.t. P_π as $\bar{\sigma}_\phi^\pi$, given by:*

$$K_\phi^\pi := \sup_{s\in\mathscr{R}_\pi} \operatorname*{ess\,sup}_{s'\sim P_\pi(\cdot|s)} \left\|\phi^\pi(s') - \mathbb{E}_\pi[\phi^\pi(S')|s]\right\|_2,$$

$$\bar{\sigma}_\phi^\pi := \sup_{s\in\mathscr{R}_\pi} \sqrt{\mathbb{E}_\pi\left[\left\|\phi^\pi(S_+) - \mathbb{E}_\pi[\phi^\pi(S_+)|s]\right\|_2^2 \;\middle|\; s\right]}.$$

The statistics K_ϕ^π and $\bar{\sigma}_\phi^\pi$ depend strictly on the mappings ϕ, π and the measure P_π, making them independent of the weights $\mathbf{w}^\pi$. We can bound the statistics $(K^\pi, \bar{\sigma}^\pi)$ using their feature map counterparts, resulting in CIs on the return process that separates the dependence on the policy (i.e., $\|\mathbf{w}^\pi\|_2$) from the stochasticity of the environment.

Theorem 4. *Consider a linear MDP where $V^\pi(s) = \langle \phi^\pi(s), \mathbf{w}^\pi\rangle$. For any policy $\pi \in \Pi_{\mathscr{B}}$, the statistics of the value function satisfy the following upper bounds:*

$$K^\pi \leq \|\mathbf{w}^\pi\|_2 K_\phi^\pi, \quad \textit{and} \quad \bar{\sigma}^\pi \leq \|\mathbf{w}^\pi\|_2 \bar{\sigma}_\phi^\pi.$$

As a result, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:

$$\begin{aligned}&\left|R_T^\pi - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big)\right| \\ &\quad\leq \|\mathbf{w}^\pi\|_2 \left(\bar{\sigma}_\phi^\pi \sqrt{2h_T \log\frac{2}{\delta} + \left(\frac{K_\phi^\pi}{3\bar{\sigma}_\phi^\pi}\log\frac{2}{\delta}\right)^2}\right. \\ &\qquad\left. + \frac{K_\phi^\pi}{3}\log\frac{2}{\delta}\right), \end{aligned} \quad (15)$$

where $h_T = \frac{\gamma^2-\gamma^{2T+2}}{1-\gamma^2}$.

The proof is presented in Appendix B.6.

3.3.1 Computing Statistics via RKHS and MMD

In the linear MDP framework, the value function $V^\pi(\cdot)$ belongs to a Reproducing Kernel Hilbert Space (RKHS). By exploiting the geometry of this space, we may derive upper-bounds for the statistics K^π and $\bar{\sigma}^\pi$ in terms of the induced kernel and Maximum Mean Discrepancy (MMD) metric.

Definition 3 (Aronszajn, 1950)**.** *Let $\mathcal{S}$ be a non-empty set. A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{S} \to \mathbb{R}$ is called a Reproducing Kernel Hilbert Space (RKHS) if there exists a symmetric, positive definite kernel function $k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ such that:*

1. *For all $s \in \mathcal{S}$, the function $k(\cdot, s)$ belongs to $\mathcal{H}$.*
2. *For all $s \in \mathcal{S}$ and all $f \in \mathcal{H}$, the point evaluation is given by the inner product: $f(s) = \langle f, k(\cdot, s)\rangle_{\mathcal{H}}$.*

Definition 4. *Let k be a positive definite kernel on $\mathcal{S}$ with RKHS $\mathcal{H}$. For probability measures P and Q on $\mathcal{S}$, the Maximum Mean Discrepancy (MMD) is defined as following:*

$$\mathrm{MMD}_{\mathcal{H}}(P, Q) \coloneqq \sup_{\|f\|_{\mathcal{H}} \leq 1} \Big(\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]\Big).$$

It is established in Gretton et al. [2012] that MMD may be rewritten in terms of the kernel as following

$$\begin{aligned}\mathrm{MMD}^2_{\mathcal{H}}(P, Q) &= \mathbb{E}[k(X, X')] \\ &\quad + \mathbb{E}[k(Y, Y')] - 2\mathbb{E}[k(X, Y)], \end{aligned} \tag{16}$$

where $X, X' \overset{\text{i.i.d.}}{\sim} P$ and $Y, Y' \overset{\text{i.i.d.}}{\sim} Q$.

Lemma 1. *Under Assumption 6, for any policy $\pi \in \Pi_{\mathscr{B}}$, we define $\kappa^\pi(s, s') \coloneqq \langle \phi^\pi(s), \phi^\pi(s')\rangle$. Let $\mathcal{H}_{\kappa^\pi}$ denote the RKHS associated with κ^π. Then V^π is also an element of RKHS $\mathcal{H}_{\kappa^\pi}$.*

Proof is presented in Appendix B.7.

The following theorem establishes a closed-form solution for computing K^π_ϕ and $\bar{\sigma}^\pi_\phi$ in terms of the MMD metric.

Theorem 5. *We have following expressions for K^π_ϕ and $\bar{\sigma}^\pi_\phi$:*

$$K^\pi_\phi = \sup_{s\in\mathscr{R}_\pi} \operatorname*{ess\,sup}_{s'\sim P_\pi(\cdot|s)} \sqrt{\mathrm{MMD}^2_{\mathcal{H}}(\delta_{s'}, P_\pi(\cdot|s))} \tag{17}$$

$$\bar{\sigma}^\pi_\phi = \sup_{s\in\mathscr{R}_\pi} \sqrt{\mathbb{E}_{s'\sim P_\pi(\cdot|s)}\big[\mathrm{MMD}^2_{\mathcal{H}}(\delta_{s'}, P_\pi(\cdot|s))\big]} \tag{18}$$

The proof is presented in Appendix B.8.

We can use the equivalent description of the MMD metric in (16) to compute K^π_ϕ and $\bar{\sigma}^\pi_\phi$. The implication of these characteristics is twofold: (i) they show how the Markov transition kernel affects the CIs derived in this paper, and (ii) they pave the way to use sampling algorithms using (16) for estimating statistics K^π and $\bar{\sigma}^\pi$.

4 CONFIDENCE INTERVALS WITH EPISTEMIC UNCERTAINTY

In many practical scenarios, the agent cannot exactly compute the value function and must estimate it using data sampled from the environment. We refer to the uncertainty resulting from not exactly knowing the environment as *epistemic uncertainty*. In this section, we study how epistemic uncertainty (i.e., the estimation error) impacts the confidence intervals derived in the previous sections.

For a fixed policy $\pi \in \Pi_{\mathscr{B}}$, we assume the agent estimates V^π using an estimation algorithm (e.g., TD learning). Let $\hat{V}^\pi_n$ denote the estimated value function using n samples for each state $s \in \mathcal{S}$. We assume the estimation process is asymptotically consistent and there exist CIs for the finite-time estimation error. Specifically, there exists N_0 such that for all $n \geq N_0$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have:

$$\|\hat{V}^\pi_n - V^\pi\|_\infty \leq U_\epsilon(n, \delta), \tag{19}$$

where $\lim_{n\to\infty} U_\epsilon(n, \delta) = 0$.

Since the true value function V^π is unknown, the true statistics K^π and $\bar{\sigma}^\pi$ are also unknown. Let $\hat{K}^\pi_n$ and $\hat{\bar{\sigma}}^\pi_n$ denote the estimated (empirical) statistics computed by substituting V^π with $\hat{V}^\pi_n$ into definitions (4) and (5).

To simplify the notation, let $U_a(T, K, \delta)$ denote the confidence interval derived in Theorem 1 and $U_\ell(T, \sigma, K, \delta)$ denote the confidence interval derived in Theorem 2:

$$U_a(T, K, \delta) \coloneqq K\sqrt{2h_T \log\frac{2}{\delta}}, \tag{20}$$

$$\begin{aligned}U_\ell(T, \sigma, K, \delta) \coloneqq \sigma\sqrt{2h_T \log\frac{2}{\delta} + \left(\frac{K}{3\sigma}\log\frac{2}{\delta}\right)^2} \\ + \frac{K}{3}\log\frac{2}{\delta}.\end{aligned} \tag{21}$$

The following theorem characterizes the impact of epistemic uncertainty $U_\epsilon(n, \delta)$ on the CIs found in Theorems 1 and 2.

Theorem 6. *For any policy $\pi \in \Pi_K$, let $\hat{V}^\pi_n$ be the estimated value function and let $\hat{K}^\pi_n$ and $\hat{\bar{\sigma}}^\pi_n$ be its corresponding statistics. Under Assumption 2, we have following CIs:*

1. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\begin{aligned}&\Big|R^\pi_T - \big(\hat{V}^\pi_n(S_0) - \gamma^T \hat{V}^\pi_n(S_T)\big)\Big| \\ &\leq \underbrace{U_\ell\Big(T, \hat{\bar{\sigma}}^\pi_n + U_\epsilon\Big(n, \frac{\delta}{2}\Big), \hat{K}^\pi_n + 2U_\epsilon\Big(n, \frac{\delta}{2}\Big), \frac{\delta}{2}\Big)}_{\textit{Confidence Interval Affected by Estimation Error}} \\ &\quad + \underbrace{U_\epsilon\Big(n, \frac{\delta}{2}\Big)(1 + \gamma^T)}_{\textit{Epistemic Uncertainty}}.\end{aligned} \tag{22}$$

2. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\begin{aligned}\Big|R_T^\pi - \big(\hat{V}_n^\pi(S_0) - \gamma^T \hat{V}_n^\pi(S_T)\big)\Big| \\ \leq \underbrace{U_a\Big(T, \hat{K}_n^\pi + 2U_\epsilon\Big(n, \frac{\delta}{2}\Big), \frac{\delta}{2}\Big)}_{\textit{Confidence Interval Affected by Estimation Error}} \\ + \underbrace{U_\epsilon\Big(n, \frac{\delta}{2}\Big)(1+\gamma^T)}_{\textit{Epistemic Uncertainty}}. \end{aligned} \tag{23}$$

The proof is presented in Appendix B.9.

5 NUMERICAL EXAMPLES

In this section, we present two numerical simulations to compare the empirical behavior of the return process with the theoretical CIs. The examples are for both continuous and discrete state MDPs.

Example 1. *Consider an MDP with state space $\mathcal{S} = \{-3, \dots, 3\}$ and action space $\mathcal{A} = \{0, 1\}$, where the reward function is defined as:*

$$r(s,a) = -\big(s^2 + \lambda \cdot \mathbb{1}_{\{a=1\}}\big). \tag{24}$$

Here, the quadratic term s^2 penalizes the agent for deviating from the center state $s = 0$, and $\lambda > 0$ represents the cost of exerting control effort. The state transition kernel under the action $a = 0$ is illustrated in Figure 2(a). The state transition kernel under the action $a = 1$ is illustrated in Figure 2(b). The trajectories of the return process, alongside the theoretical CIs, are shown in Figure 2(c).

The following example is an MDP modeling the power allocation in the remote estimation problem [Chakravorty and Mahajan, 2018].

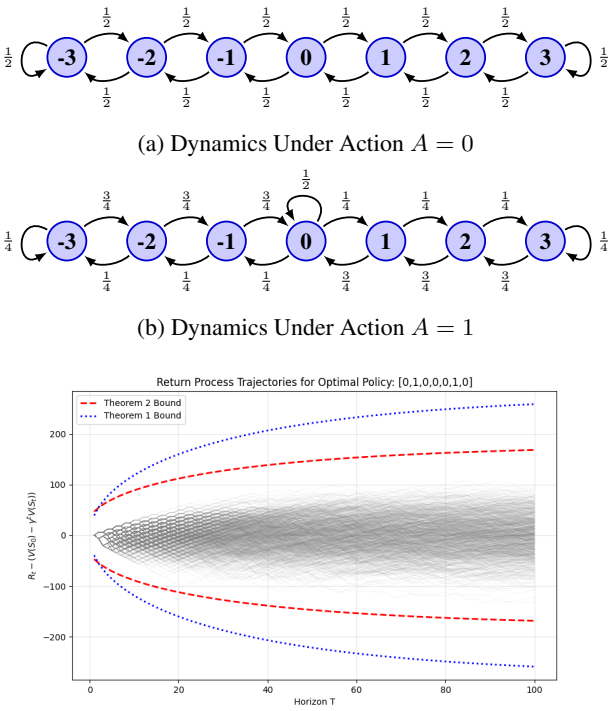
Example 2. *Consider an MDP with state space $\mathcal{S} = [-B, B]$, action space $\mathcal{A} = \{0, 1\}$, and the dynamics given by*

$$S_{t+1} = \begin{cases} [S_t + W_t]_{-B}^{B}, & \text{if } A_t = 0 \\ [W_t]_{-B}^{B}, & \text{if } A_t = 1 \end{cases}, \tag{25}$$

where $[x]_{-B}^{B} = clip(x, -B, B)$. We assume that $\{W_t\}_{t\geq 1}$ is an i.i.d. process with a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. The reward function, defined as the negative per-step cost, is given by

$$r(s,a) = -\big(\lambda a + (1-a)s^2\big), \tag{26}$$

where λ is the cost of choosing action $a = 1$. For this example, the trajectories of the return process, alongside the theoretical CIs are shown in Figure 3.



(a) Dynamics Under Action $A = 0$

(b) Dynamics Under Action $A = 1$

(c) $1,000$ trajectories of the return process alongside the theoretical CIs.

Figure 2: Trajectories of the return process for Example 1 under the optimal policy. The simulation consists of $1,000$ independent trajectories evaluated for $T = 100$, with discount factor $\gamma = 0.99$, control cost $\lambda = 4.0$, and failure probability $\delta = 0.05$. Simulated trajectories for $R_T^\pi - (V^\pi(S_0) - \gamma^T V^\pi(S_T))$ are plotted alongside the theoretical CIs derived in Theorems 1 and 2.

As demonstrated by these examples, the confidence intervals derived in this work capture the deviation of the return process R_T^π from the process $V^\pi(S_0) - \gamma^T V^\pi(S_T)$ over any finite time T. While the established CIs may not necessarily capture the exact $(1-\delta)$ percentile of the empirical distribution, they offer a rigorous, mathematically sound guarantee for the trajectory-level behavior of the return process. Importantly, they achieve this without requiring computationally expensive distributional updates or exhaustive Monte Carlo rollouts.

6 DISCUSSION ON THE RESULTS

6.1 PRACTICAL IMPLICATIONS

In this section, we explain the scenarios in which the CIs derived in this paper may be used in practical applications. Consider a decision-making agent (e.g., an RL algorithm) that will be deployed in a high-stakes *real-world* environment. In this framework, the cost/reward incurred by the

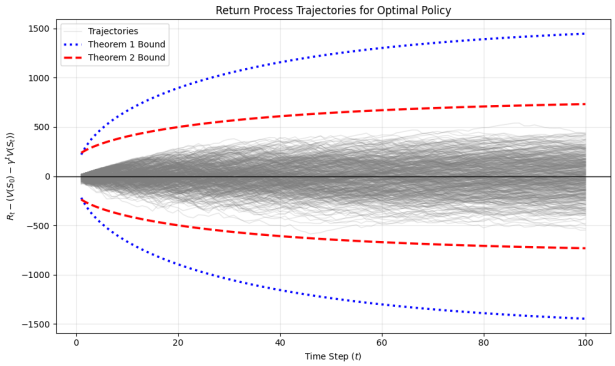


Figure 3: Trajectories of the return process for Example 2 under the optimal policy. The simulation consists of 500 independent trajectories evaluated over a horizon of $T = 100$. The environment parameters are set to state bound $B = 10.0$, communication cost $\lambda = 100.0$, noise standard deviation $\sigma = 2.5$, and discount factor $\gamma = 0.99$, with a failure probability of $\delta = 0.05$. The optimal policy and value functions are computed by discretizing the state space into 400 equal intervals. Simulated trajectories for $R_T^\pi - (V^\pi(S_0) - \gamma^T V^\pi(S_T))$ are plotted alongside the theoretical CIs from Theorems 1 and 2.

agent has real consequences for the agent. We assume the agent goes through two phases: (i) a development phase in which the agent is designed, trained, and tested in a simulated or highly controlled environment, where costs/rewards incurred by the agent are not high-stakes, and (ii) a deployment phase in which the agent is deployed in the real world with high-stakes costs/rewards. When the agent is deployed in the real world, we assume it operates for a long period of time and restarting the environment is either impossible or costly. At the time of the agent's deployment in the environment, we would like to answer the following question concretely. *Given the trained agent, can we guarantee that the return process will not be less than a threshold with high probability?*

Such a question may be posed as the safety or robustness guarantee for the agent before the real-world deployment. The CIs intervals derived in this paper may serve as guarantees required for the deployment of decision-making agents in the high-stakes environments. Although these bounds may not capture the tail behavior of the return process accurately, they provide solid conservative guarantees for the return process, and they are simpler to compute compared to the alternative methods (e.g., Monte Carlo simulation and distributional RL).

6.2 RELEVANT WORK

The results established in this paper are closely related to two sets of studies: sample path properties of Markov reward processes and distributional RL. In this section, we elaborate on the difference between the current work and these two sets of studies.

Distributional RL: The goal of Distributional RL (DRL) [Bellemare et al., 2023] is to find the distribution of the limiting return random variable R_∞^π. This is mainly done by iteratively applying the distributional Bellman operator on an approximation of the distribution of R_∞^π. There are two main distinctions between the CIs derived in this work and the distributional RL framework. First, contrary to DRL, CIs hold for the process R_T^π for any time T and not just in the asymptotic limit. Second, CIs may be computed using the knowledge of the environment and using the dynamic programming algorithms, while the DRL requires iterative application of the distributional Bellman update. As a result, CIs may be derived with a lower computational cost compared to the DRL.

Markov Reward Processes: The majority of the results characterizing the concentration of the return process in MDPs are developed for the average reward framework and they are established in the asymptotic regime (e.g., in Hernández-Lerma and Lasserre [2012], Meyn and Tweedie [2012]). Such results establish asymptotic convergence guarantees such as the Law of Large Numbers, the Law of Iterated Logarithm and the Central Limit Theorem. CIs in this paper are establishing *finite-time* CIs for the *discounted MDP* framework. To the best of our knowledge, finite-time CIs for the return process in this setup are reported only in [Sayedana et al., 2024]. Compared to the aforementioned paper, we derive CIs under much weaker assumptions on the MDP and provide Bernstein-type CIs which, to the best of our knowledge, are not reported in the literature before.

7 CONCLUSION

In this paper, we establish CIs for the return process in Markov decision processes. We derive these bounds under various assumptions on the structure of the MDP. By imposing the boundedness assumption on the maximum absolute deviation of the value function, we derive upper bounds analogous to the Azuma and Bernstein inequalities. Furthermore, under the assumption that the value innovation sequence is a sub-Gaussian process, we derive CIs analogous to Azuma-Hoeffding inequality. In the framework of linear MDPs, we provide further simplifications for the established CIs. In particular, we exploit the properties of the value function to derive CIs in terms of the MMD metric between two distributions. Moreover, we study the effect of epistemic uncertainty on the established CIs, characterizing the sensitivity of CIs to the model mismatch. In two numerical examples we compare the empirical distribution of the return process with the established CIs. We believe the results of this paper pave the way to having sample-path guarantees for decision-making in high-stakes environments.

References

Nathan Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950. URL https://api.semanticscholar.org/CorpusID:54040858.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 06–11 Aug 2017.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Dotan Di Castro, Aviv Tamar, and Shie Mannor. Policy gradients with variance related risk criteria, 2012. URL https://arxiv.org/abs/1206.6404.

Jhelum Chakravorty and Aditya Mahajan. Sufficient conditions for the value function and optimal strategy to be even and quasi-convex. *IEEE Transactions on Automatic Control*, 63(11):3858–3864, 2018. doi: 10.1109/TAC.2018.2800796.

Yinlam Chow, Ofir Nachum, Edgar A. Duéñez-Guzmán, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *CoRR*, abs/1805.07708, 2018. URL http://arxiv.org/abs/1805.07708.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1096–1105, 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. *Advances in neural information processing systems*, 33:9398–9411, 2020.

Erick Delage and Shie Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/40605970.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Germano Gabbianelli, Gergely Neu, Nneka Okolo, and Matteo Papini. Offline primal-dual reinforcement learning for linear mdps, 2023. URL https://arxiv.org/abs/2305.12944.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, January 2015. ISSN 1532-4435.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Josiah P Hanna, Peter Stone, and Scott Niekum. High confidence off-policy evaluation with models. *arXiv preprint arXiv:1606.06126*, 2016.

Onésimo Hernández-Lerma and Jean B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*, volume 42 of *Applications of Mathematics*. Springer Science & Business Media, Berlin, Heidelberg, 2012.

Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Stochastic Modelling and Applied Probability. Springer, New York, NY, 1996. ISBN 978-0-387-94579-8. doi: 10.1007/978-1-4612-0729-0.

Ronald A. Howard and James E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18 (7):356–369, March 1972. doi: 10.1287/mnsc.18.7.356. URL https://ideas.repec.org/a/inm/ormnsc/v18y1972i7p356-369.html.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, 1st edition, 1999. doi: 10.1201/9780367805326.

Ioannis Karatzas and William D Sudderth. Two characterizations of optimality in dynamic programming. *Applied Mathematics and Optimization*, 61(3):421–434, 2010.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Springer Science & Business Media, Cambridge, 2012.

Wayne B. Nelson. Statistical methods for reliability data, second edition,. *Technometrics*, 63(3):437–440, 2021. doi: 10.1080/00401706.2021.1945328.

Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications and coding (second edition), 2015. URL https://arxiv.org/abs/1212.4663.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5528–5536, Long Beach, California, USA, 2019. PMLR.

Borna Sayedana, Peter E Caines, and Aditya Mahajan. Concentration of cumulative reward in Markov decision processes. *arXiv preprint arXiv:2411.18551*, 2024.

Chengchun Shi, Jin Zhu, Ye Shen, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded Markov decision process. *Journal of the American Statistical Association*, 119(545):273–284, 2024.

Matthew J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19 (4):794–802, 1982.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling, 2014. URL https://arxiv.org/abs/1404.3862.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Ramon Van Handel. Probability in high dimension. Technical report, 2014.

Peter Whittle. Risk-sensitive optimal control. 1990. URL https://api.semanticscholar.org/CorpusID:116978296.

Proof Of The Results

A BACKGROUND ON MARTINGALES

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *filtration* $\{\mathcal{F}_t\}_{t\geq 0}$ is a non-decreasing family of sub-sigma fields of $\mathcal{F}$. A random sequence $\{X_t\}_{t\geq 0}$ is called *integrable* if $\mathbb{E}[|X_t|] < \infty$ for all $t \geq 0$. A random sequence $\{X_t\}_{t\geq 0}$ is called *adapted* to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if X_t is $\mathcal{F}_t$-measurable for all $t \geq 0$.

Definition 5 (Martingale)**.** *An integrable sequence $\{X_t\}_{t\geq 0}$ adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is called a martingale if*

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t, \quad \textit{a.s.} \quad \forall t \geq 0.$$

Definition 6 (Martingale Difference Sequence)**.** *Let $\{c_t\}_{t\geq 1}$ be a sequence of real numbers and C be a positive real number. A real integrable sequence $\{Y_t\}_{t\geq 1}$ adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is called:*

1. Martingale Difference Sequence (MDS) *if*

$$\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0, \quad \textit{a.s.} \quad \forall t \geq 1.$$

2. Sequentially bounded MDS *with respect to the sequence $\{c_t\}_{t\geq 1}$ if it is an MDS and*

$$|Y_t| \leq c_t, \quad \textit{a.s.} \quad \forall t \geq 1.$$

3. Uniformly bounded MDS *with respect to the constant C if it is an MDS and*

$$|Y_t| \leq C, \quad \textit{a.s.} \quad \forall t \geq 1.$$

There is a unique MDS corresponding to a martingale and vice versa. In particular, given a martingale $\{X_t\}_{t\geq 0}$, the corresponding MDS $\{Y_t\}_{t\geq 1}$ is defined as

$$Y_t := X_t - X_{t-1}, \quad \forall t \geq 1.$$

Moreover, given an MDS $\{Y_t\}_{t\geq 1}$, the corresponding martingale sequence $\{X_t\}_{t\geq 0}$ is defined as

$$X_0 = 0, \quad X_T = \sum_{t=1}^{T} Y_t, \quad \forall T \geq 1.$$

Consider a martingale $\{X_t\}_{t\geq 0}$ such that $\{X_t^2\}_{t\geq 0}$ is integrable. The *increasing process* $\{A_t\}_{t\geq 1}$ associated with the sequence $\{X_t^2\}_{t\geq 0}$ is defined as

$$A_1 = \mathbb{E}[X_1^2|\mathcal{F}_0] - X_1^2, \quad A_t = \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] - X_{t-1}^2 + A_{t-1}, \quad \forall t \geq 2.$$

Let $\{Y_t\}_{t\geq 0}$ be the MDS corresponding to $\{X_t\}_{t\geq 0}$. Then, we can express $\{A_t\}_{t\geq 0}$ in terms of $\{Y_t^2\}_{t\geq 0}$. In particular, we have

$$\begin{aligned} A_t &= \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] - X_{t-1}^2 + A_{t-1} \\ &= \mathbb{E}[X_{t-1}^2|\mathcal{F}_{t-1}] + 2\mathbb{E}[Y_t|\mathcal{F}_{t-1}]X_{t-1} + \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}] - X_{t-1}^2 + A_{t-1} \\ &= \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}] + A_{t-1}. \end{aligned}$$

As a result, we have

$$A_T = \sum_{t=1}^{T} \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}], \quad \forall T \geq 1.$$

Therefore, we sometimes say that $\{A_t\}_{t\geq 1}$ is the predictable quadratic variation associated with $\{Y_t^2\}_{t\geq 0}$.

Martingale sequences are an important class of stochastic processes. Both asymptotic and non-asymptotic concentrations of martingale sequences have been well studied. In Section A.1, we present the non-asymptotic concentration of martingales with uniformly bounded MDS.

A.1 NON-ASYMPTOTIC CONCENTRATION

A.1.1 Freedman Inequality

A foundational non-asymptotic concentration result for martingale sequences is Freedman's inequality, which provides a tight bound when the conditional variance is small.

Theorem 7 (Freedman, 1975)**.** *Let $\{Y_t\}_{t\geq 1}$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ such that $|Y_t| \leq K$ almost surely for all t. Let the predictable quadratic variation be:*

$$\Sigma_T^2 \coloneqq \sum_{t=1}^{T} \mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}].$$

Then, for any $\epsilon > 0$ and any $\nu > 0$, we have:

$$\mathbb{P}\left(\sum_{t=1}^{T} Y_t \geq \epsilon \quad \textit{and} \quad \Sigma_T^2 \leq \nu \right) \leq \exp\left(-\frac{\epsilon^2}{2\nu + \frac{2K\epsilon}{3}} \right).$$

In the analysis of this paper, we require a two-sided bound. The derivation of the two-sided Freedman inequality is presented in the following corollary.

Corollary 3. *Let $\{Y_t\}_{t\geq 1}$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ such that $|Y_t| \leq K$ almost surely. Let $S_T = \sum_{t=1}^{T} Y_t$ and let the conditional variance sum be bounded such that $\Sigma_T^2 \leq \nu$ almost surely.*

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have:

$$|S_T| \leq \sqrt{2\nu \log\left(\frac{2}{\delta}\right) + \left(\frac{K}{3}\log\left(\frac{2}{\delta}\right)\right)^2} + \frac{K}{3}\log\left(\frac{2}{\delta}\right).$$

Proof. We begin by applying Theorem 7 to the martingale S_T. For any $\epsilon > 0$:

$$\mathbb{P}\left(S_T \geq \epsilon \text{ and } \Sigma_T^2 \leq \nu\right) \leq \exp\left(-\frac{\epsilon^2}{2\nu + \frac{2K\epsilon}{3}} \right). \tag{27}$$

To bound the lower tail, we define the symmetric sequence $M_T = -S_T$. Because $\mathbb{E}[-Y_t \mid \mathcal{F}_{t-1}] = 0$, M_T is also a martingale. Its difference sequence, $-Y_t$, satisfies the same uniform bound $|-Y_t| \leq K$. Furthermore, the conditional variance remains identical since $(-Y_t)^2 = Y_t^2$. Applying Theorem 7 to M_T yields:

$$\mathbb{P}\left(S_T \leq -\epsilon \text{ and } \Sigma_T^2 \leq \nu\right) \leq \exp\left(-\frac{\epsilon^2}{2\nu + \frac{2K\epsilon}{3}} \right). \tag{28}$$

The event $\{|S_T| \geq \epsilon\}$ is the union of the disjoint events $\{S_T \geq \epsilon\}$ and $\{S_T \leq -\epsilon\}$. Assuming the variance bound $\Sigma_T^2 \leq \nu$ holds deterministically, we can apply the union bound to (27) and (28) to obtain the two-sided deviation probability:

$$\mathbb{P}\left(|S_T| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\nu + \frac{2K\epsilon}{3}}\right). \tag{29}$$

To obtain the high-probability bound, we equate the right-hand side failure probability to δ:

$$\delta = 2\exp\left(-\frac{\epsilon^2}{2\nu + \frac{2K\epsilon}{3}}\right) \implies \log\left(\frac{2}{\delta}\right) = \frac{\epsilon^2}{2\nu + \frac{2K\epsilon}{3}}. \tag{30}$$

Rearranging this expression yields the following quadratic equation in terms of ϵ:

$$\epsilon^2 - \frac{2K}{3}\log\left(\frac{2}{\delta}\right)\epsilon - 2\nu\log\left(\frac{2}{\delta}\right) = 0. \tag{31}$$

Applying the quadratic formula and taking the positive root (since $\epsilon > 0$), we obtain:

$$\epsilon = \frac{K}{3}\log\left(\frac{2}{\delta}\right) + \sqrt{\left(\frac{K}{3}\log\left(\frac{2}{\delta}\right)\right)^2 + 2\nu\log\left(\frac{2}{\delta}\right)}.$$

Therefore, the event $|S_T| \geq \epsilon$ occurs with probability at most δ, meaning its complement $|S_T| < \epsilon$ occurs with probability at least $1 - \delta$. This completes the proof. □

A.1.2 Azuma-Hoeffding Inequality

While Freedman's inequality utilizes the predictable quadratic variation to provide tighter bounds for martingales with small conditional variances, the Azuma-Hoeffding inequality provides a simpler, variance-independent bound depending only on the maximum absolute deviation of the martingale differences.

Theorem 8 (Raginsky and Sason, 2015)**.** *Let $\{X_t\}_{t=1}^T$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^T$. Suppose there exists a sequence of positive constants $\{c_t\}_{t=1}^T$ such that for all $t \in \{1, \ldots, T\}$, we have $|X_t| \leq c_t$ almost surely. Let $S_T = \sum_{t=1}^T X_t$. Then, for any $\epsilon > 0$:*

$$\mathbb{P}\left(|S_T| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^T c_t^2}\right). \tag{32}$$

By rewriting the statement of Theorem 8 in terms of a failure probability δ, we get the following equivalent high-probability form, which is directly applicable to our analysis.

Corollary 4. *Let $\{X_t\}_{t=1}^T$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^T$, such that $|X_t| \leq c_t$ almost surely. Let $S_T = \sum_{t=1}^T X_t$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have:*

$$|S_T| \leq \sqrt{2\left(\sum_{t=1}^T c_t^2\right)\log\left(\frac{2}{\delta}\right)}. \tag{33}$$

Proof. The result follows by equating the probability bound in Theorem 8 to the failure probability δ and solving for the deviation ϵ:

$$\delta = 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^T c_t^2}\right) \implies \epsilon = \sqrt{2\left(\sum_{t=1}^T c_t^2\right)\log\left(\frac{2}{\delta}\right)}.$$

Substituting this ϵ back into the complement event $|S_T| \leq \epsilon$ yields the stated high-probability bound. □

B PROOF OF MAIN RESULTS FOR THE DISCOUNTED REWARD SETUP

B.0.1 Preliminary Results

We first present a few preliminary lemmas. To simplify the notation, we define the value innovation sequence.

Definition 7. *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ be defined as $\mathcal{F}_t := \sigma(S_{0:t}, A_{0:t})$. For any policy $\pi \in \Pi_{\mathscr{B}}$, let V^π denote the corresponding discounted value function. We define the value innovation sequence $\{N_t^\pi\}_{t\geq 1}$ as follows:*

$$N_t^\pi := V^\pi(S_t) - \mathbb{E}_\pi[V^\pi(S_t) \mid S_{t-1}], \quad \forall t \geq 1, \tag{34}$$

where $\{S_t\}_{t\geq 1}$ denotes the random sequence of states following the policy π.

Lemma 2. *The value innovation sequence $\{\gamma^t N_t^\pi\}_{t\geq 1}$ is a Martingale Difference Sequence (MDS) with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$.*

Proof. By the definition of $\{\mathcal{F}_t\}_{t\geq 0}$, we have that S_{t-1} is $\mathcal{F}_{t-1}$-measurable. By the Markov property, we have:

$$\begin{aligned}
\mathbb{E}\left[\gamma^t N_t^\pi \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\gamma^t \left(V^\pi(S_t) - \mathbb{E}_\pi[V^\pi(S_t) \mid S_{t-1}]\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \gamma^t \mathbb{E}\left[V^\pi(S_t) \mid \mathcal{F}_{t-1}\right] - \gamma^t \mathbb{E}_\pi\left[V^\pi(S_t) \mid S_{t-1}\right] \\
&= \gamma^t \mathbb{E}_\pi\left[V^\pi(S_t) \mid S_{t-1}\right] - \gamma^t \mathbb{E}_\pi\left[V^\pi(S_t) \mid S_{t-1}\right] = 0,
\end{aligned}$$

which shows that $\{\gamma^t N_t^\pi\}_{t\geq 1}$ is an MDS with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. □

We now present a martingale decomposition for the return process R_T^π for any policy $\pi \in \Pi_{\mathscr{B}}$.

Lemma 3. *Given any policy $\pi \in \Pi_{\mathscr{B}}$, we can rewrite the return process R_T^π as follows:*

$$R_T^\pi = \sum_{t=1}^{T} \gamma^t N_t^\pi + V^\pi(S_0) - \gamma^T V^\pi(S_T). \tag{35}$$

Proof. Since $\pi \in \Pi_{\mathscr{B}}$, the Bellman equation implies that along the trajectory of states $\{S_t\}_{t=0}^T$ induced by the policy π, the instantaneous reward can be written as:

$$r(S_t, \pi(S_t)) = V^\pi(S_t) - \gamma \mathbb{E}_\pi\left[V^\pi(S_{t+1}) \mid S_t\right].$$

By substituting this into the definition of the discounted return, we get the following telescoping sum:

$$\begin{aligned}
R_T^\pi &= \sum_{t=0}^{T-1} \gamma^t r(S_t, \pi(S_t)) \\
&= \sum_{t=0}^{T-1} \gamma^t \left[V^\pi(S_t) - \gamma \mathbb{E}_\pi\left[V^\pi(S_{t+1}) \mid S_t\right]\right] \\
&\overset{(a)}{=} \sum_{t=0}^{T-1} \gamma^t V^\pi(S_t) - \sum_{t=0}^{T-1} \gamma^{t+1} \mathbb{E}_\pi\left[V^\pi(S_{t+1}) \mid S_t\right] + \gamma^T V^\pi(S_T) - \gamma^T V^\pi(S_T) \\
&\overset{(b)}{=} \sum_{t=0}^{T-1} \gamma^{t+1} \left[V^\pi(S_{t+1}) - \mathbb{E}_\pi\left[V^\pi(S_{t+1}) \mid S_t\right]\right] + V^\pi(S_0) - \gamma^T V^\pi(S_T) \\
&\overset{(c)}{=} \sum_{t=0}^{T-1} \gamma^{t+1} N_{t+1}^\pi + V^\pi(S_0) - \gamma^T V^\pi(S_T) \\
&= \sum_{t=1}^{T} \gamma^t N_t^\pi + V^\pi(S_0) - \gamma^T V^\pi(S_T),
\end{aligned}$$

where (a) follows from adding and subtracting the term $\gamma^T V^\pi(S_T)$, (b) follows from shifting the index of the first summation and pairing the terms evaluated at $t+1$, and (c) follows from the definition of the value innovation sequence $\{N_t^\pi\}_{t\geq 1}$. □

B.1 PROOF OF THEOREM 1

A variant of this result was previously established in Sayedana et al. [2024, Theorem 37] under stronger assumptions (finite state and action spaces and uniformly bounded rewards). For completeness, we prove this result under the weaker assumptions imposed in this paper, i.e., Assumption 2.

Proof. By Lemma 3, for any policy $\pi \in \Pi_K$, we have the following decomposition for the return process:

$$R_T^\pi = \sum_{t=1}^{T} \gamma^t N_t^\pi + V^\pi(S_0) - \gamma^T V^\pi(S_T).$$

Rearranging the terms, we obtain:

$$\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| = \left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right|. \tag{36}$$

To bound the right-hand side, we apply Azuma-Hoeffding inequality in Corollary 4. By Lemma 2, the sequence $\{\gamma^t N_t^\pi\}_{t=1}^T$ is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^T$. Under Assumption 2, the absolute deviation of the value innovation sequence is bounded by K^π almost surely. As a result, we have:

$$\left|\gamma^t N_t^\pi\right| \le \gamma^t K^\pi, \quad \text{a.s.} \quad \forall t \in \mathbb{N}. \tag{37}$$

Let $c_t = \gamma^t K^\pi$. We have:

$$\sum_{t=1}^{T} c_t^2 = \sum_{t=1}^{T} \left(\gamma^t K^\pi\right)^2 = (K^\pi)^2 \sum_{t=1}^{T} \gamma^{2t} = (K^\pi)^2 \frac{\gamma^2 - \gamma^{2T+2}}{1-\gamma^2} = (K^\pi)^2 h_T, \tag{38}$$

where the last equality follows by the geometric series formula.

Applying Corollary 4 with these bounds, we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right| \le \sqrt{2\big((K^\pi)^2 h_T\big) \log\left(\frac{2}{\delta}\right)} = K^\pi \sqrt{2 h_T \log\left(\frac{2}{\delta}\right)}. \tag{39}$$

Substituting (39) into (36) implies the final CI:

$$\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| \le K^\pi \sqrt{2 h_T \log\left(\frac{2}{\delta}\right)}. \tag{40}$$

□

B.2 PROOF OF THEOREM 2

Proof. By Lemma 3, for any policy $\pi \in \Pi_K$, we have the following decomposition for the reward process R_T^π.

$$R_T^\pi = \sum_{t=1}^{T} \gamma^t N_t^\pi + V^\pi(S_0) - \gamma^T V^\pi(S_T).$$

Rearranging terms, we obtain

$$\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| = \left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right|. \tag{41}$$

To bound the right-hand side, we apply the two-sided Freedman inequality in Corollary 3. By Lemma 2, the sequence $\{\gamma^t N_t^\pi\}_{t\geq 1}$ is a martingale difference sequence adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Furthermore, by Assumption 2 the value innovation sequence is bounded by K^π almost surely. As a result, we have

$$|\gamma^t N_t^\pi| \leq \gamma^t K^\pi \leq K^\pi, \quad \forall t \geq 1.$$

In addition, we must bound the predictable quadratic variation. By the definition of the maximum conditional standard deviation $\bar{\sigma}^\pi$, we know that $\mathbb{V}_\pi[V^\pi(S_+)|s] \leq (\bar{\sigma}^\pi)^2$ for all reachable states $s \in \mathscr{R}_\pi$. Therefore, the predictable quadratic variation sequence satisfies:

$$\sum_{t=1}^{T} \mathbb{E}\left[(\gamma^t N_t^\pi)^2 \;\middle|\; \mathcal{F}_{t-1}\right] \leq \sum_{t=1}^{T} \gamma^{2t}(\bar{\sigma}^\pi)^2 = \frac{\gamma^2 - \gamma^{2T+2}}{1-\gamma^2}(\bar{\sigma}^\pi)^2.$$

Let $h_T \coloneqq \frac{\gamma^2 - \gamma^{2T+2}}{1-\gamma^2}$. Applying Corollary 3 with the uniform range bound K^π and the variance bound $h_T(\bar{\sigma}^\pi)^2$, we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right| \leq \sqrt{2h_T(\bar{\sigma}^\pi)^2 \log\left(\frac{2}{\delta}\right) + \left(\frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right)\right)^2} + \frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right). \tag{42}$$

By combining (41) and (42), and factoring $\bar{\sigma}^\pi$ out the square root, we get the following:

$$\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| \leq \bar{\sigma}^\pi \sqrt{2h_T \log\left(\frac{2}{\delta}\right) + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\left(\frac{2}{\delta}\right)\right)^2} + \frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right).$$

□

B.3 PROOF OF COROLLARY 1

Proof. By Lemma 3, for any finite horizon T, we have the following martingale decomposition for the return process:

$$R_T^\pi - V^\pi(S_0) = \sum_{t=1}^{T} \gamma^t N_t^\pi - \gamma^T V^\pi(S_T). \tag{43}$$

We analyze the limit of this expression as $T \to \infty$. By definition, the discounted return converges to the asymptotic return, $\lim_{T\to\infty} R_T^\pi = R_\infty^\pi$ almost surely. Furthermore, under Assumption 3, we have

$$\lim_{T\to\infty} \gamma^T V^\pi(S_T) = 0, \quad \text{a.s.} \tag{44}$$

Therefore, by taking the limit as $T \to \infty$ in (43), we have

$$R_\infty^\pi - V^\pi(S_0) = \sum_{t=1}^{\infty} \gamma^t N_t^\pi, \tag{45}$$

Furthermore, notice that

$$h_\infty \coloneqq \lim_{T\to\infty} h_T = \lim_{T\to\infty} \frac{\gamma^2 - \gamma^{2T+2}}{1-\gamma^2} = \frac{\gamma^2}{1-\gamma^2}. \tag{46}$$

Proof of Part 1: By letting $T \to \infty$ in the statement of Theorem 1, we get

$$|R_\infty^\pi - V^\pi(S_0)| \leq K^\pi \sqrt{2h_\infty \log\frac{2}{\delta}} = K^\pi \sqrt{\frac{2\gamma^2}{1-\gamma^2}\log\frac{2}{\delta}}. \tag{47}$$

Proof of Part 2: By letting $T \to \infty$ in the statement of Theorem 2, we get

$$|R_\infty^\pi - V^\pi(S_0)| \leq \sqrt{2h_\infty(\bar{\sigma}^\pi)^2 \log\left(\frac{2}{\delta}\right) + \left(\frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right)\right)^2} + \frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right)$$
$$= \bar{\sigma}^\pi \sqrt{\frac{2\gamma^2}{1-\gamma^2}\log\frac{2}{\delta} + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\frac{2}{\delta}\right)^2} + \frac{K^\pi}{3}\log\frac{2}{\delta}. \tag{48}$$

□

B.4 PROOF OF COROLLARY 2

Proof. By Lemma 3, for any policy $\pi \in \Pi_{\mathscr{B}}$, we have the following decomposition for the return process R_T^π:

$$R_T^\pi = \sum_{t=1}^{T} \gamma^t N_t^\pi + V^\pi(S_0) - \gamma^T V^\pi(S_T).$$

Rearranging the terms and applying the triangle inequality yields

$$|R_T^\pi - V^\pi(S_0)| \leq \left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right| + \left|\gamma^T V^\pi(S_T)\right|. \tag{49}$$

Under Assumption 4, the value function is uniformly bounded by $V_{\max}$, meaning $|V^\pi(s)| \leq V_{\max}$ for all states $s \in \mathcal{S}$. As a result, the second term in (49) is bounded by:

$$\left|\gamma^T V^\pi(S_T)\right| \leq \gamma^T V_{\max}. \quad \text{a.s.} \tag{50}$$

Proof of Part 1: From Theorem 1, under Assumption 2, the sum of the martingale difference sequence is bounded with probability at least $1-\delta$ by:

$$\left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right| \leq K^\pi \sqrt{2h_T \log\frac{2}{\delta}}. \tag{51}$$

Substituting (51) and (50) into (49), we obtain with probability at least $1-\delta$:

$$|R_T^\pi - V^\pi(S_0)| \leq K^\pi \sqrt{2h_T \log\frac{2}{\delta}} + \gamma^T V_{\max}. \tag{52}$$

Proof of Part 2: Similarly, from Theorem 2, under Assumption 2, the sum of the martingale difference sequence is bounded with probability at least $1-\delta$ by:

$$\left|\sum_{t=1}^{T} \gamma^t N_t^\pi\right| \leq \bar{\sigma}^\pi \sqrt{2h_T \log\left(\frac{2}{\delta}\right) + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\left(\frac{2}{\delta}\right)\right)^2} + \frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right). \tag{53}$$

Substituting (53) and (50) into (49), we obtain with probability at least $1-\delta$:

$$|R_T^\pi - V^\pi(S_0)| \leq \bar{\sigma}^\pi \sqrt{2h_T \log\left(\frac{2}{\delta}\right) + \left(\frac{K^\pi}{3\bar{\sigma}^\pi}\log\left(\frac{2}{\delta}\right)\right)^2}$$
$$+ \frac{K^\pi}{3}\log\left(\frac{2}{\delta}\right) + \gamma^T V_{\max}. \tag{54}$$

□

B.5 PROOF OF THEOREM 3

Proof. The Azuma-Hoeffding inequality for sub-Gaussian martingale difference sequences is established in Van Handel [2014]. We do not directly use this result and provide a direct derivation for clarity and completeness in this paper.

By the martingale decomposition established in Lemma 3, we know that for any policy $\pi \in \Pi_{\mathscr{B}}$, we have:

$$R_T^\pi - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big) = \sum_{t=1}^{T} \gamma^t N_t^\pi. \tag{55}$$

Let $M_T \coloneqq \sum_{t=1}^{T} \gamma^t N_t^\pi$. We seek to bound $\mathbb{P}(M_T \geq \epsilon)$ for any $\epsilon > 0$. We apply the Chernoff bound method. By Markov's inequality applied to the exponential function, for any $\lambda > 0$, we have:

$$\mathbb{P}(M_T \geq \epsilon) = \mathbb{P}\Big(\exp(\lambda M_T) \geq \exp(\lambda\epsilon)\Big) \leq \exp(-\lambda\epsilon)\mathbb{E}\Big[\exp(\lambda M_T)\Big]. \tag{56}$$

We analyze the moment generating function (MGF) using the tower property of conditional expectation:

$$\mathbb{E}\Big[\exp(\lambda M_T)\Big] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T} \gamma^t N_t^\pi\right) \,\middle|\, \mathcal{F}_{T-1}\right]\right]. \tag{57}$$

Because the sum up to $T-1$ is measurable with respect to $\mathcal{F}_{T-1}$, we can factor it out of the inner expectation:

$$\mathbb{E}\Big[\exp(\lambda M_T)\Big] = \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T-1} \gamma^t N_t^\pi\right) \mathbb{E}\left[\exp\left(\lambda\gamma^T N_T^\pi\right) \,\middle|\, \mathcal{F}_{T-1}\right]\right]. \tag{58}$$

By Assumption 5, the value innovation sequence N_T^π is conditionally ν^π-sub-Gaussian. Consequently, the scaled term $\gamma^T N_T^\pi$ is conditionally $(\gamma^T\nu^\pi)$-sub-Gaussian. This provides the following bound on the conditional MGF:

$$\mathbb{E}\left[\exp\left(\lambda\gamma^T N_T^\pi\right) \,\middle|\, \mathcal{F}_{T-1}\right] \leq \exp\left(\frac{\lambda^2\gamma^{2T}(\nu^\pi)^2}{2}\right). \tag{59}$$

Substituting this bound back into our expectation and iterating this unrolling process backward to $t = 1$ results in:

$$\mathbb{E}\Big[\exp(\lambda M_T)\Big] \leq \exp\left(\frac{\lambda^2(\nu^\pi)^2}{2} \sum_{t=1}^{T} \gamma^{2t}\right). \tag{60}$$

By the definition of h_T, we get:

$$\mathbb{E}\Big[\exp(\lambda M_T)\Big] \leq \exp\left(\frac{\lambda^2(\nu^\pi)^2 h_T}{2}\right). \tag{61}$$

Substituting (61) back into (56) results in:

$$\mathbb{P}(M_T \geq \epsilon) \leq \exp\left(-\lambda\epsilon + \frac{\lambda^2(\nu^\pi)^2 h_T}{2}\right). \tag{62}$$

To obtain the tightest bound, we minimize the right-hand side with respect to λ. The quadratic function in the exponent is minimized at $\lambda = \frac{\epsilon}{(\nu^\pi)^2 h_T}$. Substituting this optimal λ yield:

$$\mathbb{P}(M_T \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2(\nu^\pi)^2 h_T}\right). \tag{63}$$

By symmetry, applying the exact same argument to $-M_T$ bound the lower tail, giving $\mathbb{P}(-M_T \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2(\nu^\pi)^2 h_T}\right)$. Taking the union bound over both tails provides the two-sided inequality:

$$\mathbb{P}(|M_T| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2(\nu^\pi)^2 h_T}\right). \tag{64}$$

To convert this to a high-probability bound, we equate the failure probability to δ:

$$\delta = 2\exp\left(-\frac{\epsilon^2}{2(\nu^\pi)^2 h_T}\right). \tag{65}$$

Solving for ϵ yields:

$$\epsilon = \nu^\pi \sqrt{2h_T \log\left(\frac{2}{\delta}\right)}. \tag{66}$$

Therefore, with probability at least $1-\delta$, we have $|M_T| \le \nu^\pi \sqrt{2h_T \log(2/\delta)}$. Substituting (55) back into this expression concludes the proof. □

B.6 PROOF OF THEOREM 4

Proof. The proof consists of two parts. First, we establish the upper bounds for the scalar statistics K^π and $\bar{\sigma}^\pi$. Then, we substitute these bounds into the concentration inequality established in Theorem 2.

Part 1: Bounding the Statistics. Recall the definition of the maximum absolute deviation K^π:

$$K^\pi = \sup_{s\in\mathscr{R}_\pi} \operatorname*{ess\,sup}_{s'\sim P_\pi(\cdot|s)} \left|V^\pi(s') - \mathbb{E}_\pi[V^\pi(S_+)|s]\right|. \tag{67}$$

Substituting the linear value function $V^\pi(x) = \langle \phi^\pi(x), \mathbf{w}^\pi\rangle$ and using the linearity of the expectation operator, we have:

$$\begin{aligned}
\left|V^\pi(s') - \mathbb{E}_\pi[V^\pi(S_+)|s]\right| &= \left|\langle \phi^\pi(s'), \mathbf{w}^\pi\rangle - \mathbb{E}_\pi[\langle \phi^\pi(S_+), \mathbf{w}^\pi\rangle \mid s]\right| \\
&= \left|\langle \phi^\pi(s'), \mathbf{w}^\pi\rangle - \langle \mathbb{E}_\pi[\phi^\pi(S_+)|s], \mathbf{w}^\pi\rangle\right| \\
&= \left|\langle \phi^\pi(s') - \mathbb{E}_\pi[\phi^\pi(S_+)|s], \mathbf{w}^\pi\rangle\right|.
\end{aligned} \tag{68}$$

Applying the Cauchy-Schwarz inequality ($|\langle u, v\rangle| \le \|u\|_2\|v\|_2$), we obtain:

$$\left|V^\pi(s') - \mathbb{E}_\pi[V^\pi(S_+)|s]\right| \le \|\phi^\pi(s') - \mathbb{E}_\pi[\phi^\pi(S_+)|s]\|_2 \|\mathbf{w}^\pi\|_2. \tag{69}$$

By taking the essential supremum over s' and the supremum over $s \in \mathscr{R}_\pi$ from both sides, we get:

$$K^\pi \le \|\mathbf{w}^\pi\|_2 K^\pi_\phi. \tag{70}$$

Next, recall the definition of the maximum conditional standard deviation $\bar{\sigma}^\pi$. For any state $s \in \mathscr{R}_\pi$, the conditional variance is:

$$\mathbb{V}_\pi[V^\pi(S_+)|s] = \mathbb{E}_\pi\left[\left(V^\pi(S_+) - \mathbb{E}_\pi[V^\pi(S_+)|s]\right)^2 \;\middle|\; s\right]. \tag{71}$$

Following the exact same substitution and linearity steps as above, we get:

$$\begin{aligned}
\mathbb{V}_\pi[V^\pi(S_+)|s] &= \mathbb{E}_\pi\left[\langle \phi^\pi(S_+) - \mathbb{E}_\pi[\phi^\pi(S_+)|s], \mathbf{w}^\pi\rangle^2 \;\middle|\; s\right] \\
&\overset{(a)}{\le} \mathbb{E}_\pi\left[\left\|\phi^\pi(S_+) - \mathbb{E}_\pi[\phi^\pi(S_+)|s]\right\|_2^2 \|\mathbf{w}^\pi\|_2^2 \;\middle|\; s\right] \\
&= \|\mathbf{w}^\pi\|_2^2 \mathbb{E}_\pi\left[\left\|\phi^\pi(S_+) - \mathbb{E}_\pi[\phi^\pi(S_+)|s]\right\|_2^2 \;\middle|\; s\right],
\end{aligned} \tag{72}$$

where (a) follows from the Cauchy-Schwarz inequality. By taking the square root of both sides and then the supremum over $s \in \mathscr{R}_\pi$, we get:

$$\bar{\sigma}^\pi \le \|\mathbf{w}^\pi\|_2 \bar{\sigma}^\pi_\phi. \tag{73}$$

Part 2: The Confidence Interval. From Theorem 2, we know that for any $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| \leq \sqrt{2h_T(\bar{\sigma}^\pi)^2 \log\frac{2}{\delta} + \left(\frac{K^\pi}{3}\log\frac{2}{\delta}\right)^2} + \frac{K^\pi}{3}\log\frac{2}{\delta}. \tag{74}$$

Let $f(x,y) = \sqrt{Ax^2+By^2} + Cy$ for strictly positive constants A, B, C. The function $f(x,y)$ is monotonically increasing in each of its arguments x and y. Therefore, substituting the upper bounds $x = \bar{\sigma}^\pi \leq \|\mathbf{w}^\pi\|_2 \bar{\sigma}_\phi^\pi$ and $y = K^\pi \leq \|\mathbf{w}^\pi\|_2 K_\phi^\pi$ preserves the inequality:

$$\begin{aligned}&\left|R_T^\pi - \left(V^\pi(S_0) - \gamma^T V^\pi(S_T)\right)\right| \\ &\leq \sqrt{2h_T\left(\|\mathbf{w}^\pi\|_2\bar{\sigma}_\phi^\pi\right)^2 \log\frac{2}{\delta} + \left(\frac{\|\mathbf{w}^\pi\|_2 K_\phi^\pi}{3}\log\frac{2}{\delta}\right)^2} + \frac{\|\mathbf{w}^\pi\|_2 K_\phi^\pi}{3}\log\frac{2}{\delta}.\end{aligned} \tag{75}$$

Factoring out $\|\mathbf{w}^\pi\|_2$ and $\bar{\sigma}_\phi^\pi$ results in:

$$\|\mathbf{w}^\pi\|_2\left(\bar{\sigma}_\phi^\pi\sqrt{2h_T\log\frac{2}{\delta} + \left(\frac{K_\phi^\pi}{3\bar{\sigma}_\phi^\pi}\log\frac{2}{\delta}\right)^2} + \frac{K_\phi^\pi}{3}\log\frac{2}{\delta}\right). \tag{76}$$

□

B.7 PROOF OF LEMMA 1

Proof. Under Assumption 6, the underlying environment is a linear MDP. A fundamental property of linear MDPs is that the state-action value functions can be represented as linear combinations of the feature map. Specifically, as shown by Gabbianelli et al. [2023], for any policy π, there exists a weight vector $\mathbf{w}^\pi \in \mathbb{R}^d$ such that $Q^\pi(s,a) = \langle\phi(s,a), \mathbf{w}^\pi\rangle$.

For any deterministic policy π, the state-value function is evaluated as $V^\pi(s) = Q^\pi(s,\pi(s))$. Substituting the linear representation of the action-value function, we obtain:

$$V^\pi(s) = \langle\phi(s,\pi(s)), \mathbf{w}^\pi\rangle = \langle\phi^\pi(s), \mathbf{w}^\pi\rangle.$$

Next, we analyze the policy-induced kernel $k^\pi(s,s') = \langle\phi^\pi(s), \phi^\pi(s')\rangle$. Because it is defined via a standard Euclidean inner product in $\mathbb{R}^d$, k^π is symmetric and positive semi-definite. By the Moore-Aronszajn theorem, the unique RKHS $\mathcal{H}_{k^\pi}$ associated with k^π consists exactly of functions that can be expressed in the form $f(s) = \langle\phi^\pi(s), \mathbf{w}_f\rangle$ for some parameter vector $\mathbf{w}_f \in \mathbb{R}^d$. Since $V^\pi(s)$ precisely matches this functional form with $\mathbf{w}_f = \mathbf{w}^\pi$, it immediately follows that $V^\pi \in \mathcal{H}_{k^\pi}$. □

B.8 PROOF OF THEOREM 5

Proof. Let $\mathcal{H}$ be the Reproducing Kernel Hilbert Space (RKHS) associated with the kernel $k^\pi(x,y) = \langle\phi^\pi(x), \phi^\pi(y)\rangle$. To simplify the notation, we define the conditional mean embedding of the transition distribution $P_\pi(\cdot|s)$ in the feature space as:

$$\mu_P^\pi(s) \coloneqq \mathbb{E}_{S_+\sim P_\pi(\cdot|s)}[\phi^\pi(S_+)]. \tag{77}$$

Proof of Part 1 (K_ϕ^π): Recall from Definition 2 that the absolute deviation evaluated at a specific transition (s,s') is given by $\|\phi^\pi(s') - \mu_P^\pi(s)\|$. Let $\Delta^2(s,s') \coloneqq \|\phi^\pi(s') - \mu_P^\pi(s)\|_2^2$, we have:

$$\begin{aligned}\Delta^2(s,s') &\coloneqq \|\phi^\pi(s') - \mu_P^\pi(s)\|_2^2 \\ &= \langle\phi^\pi(s') - \mu_P^\pi(s), \phi^\pi(s') - \mu_P^\pi(s)\rangle \\ &= \langle\phi^\pi(s'), \phi^\pi(s')\rangle - 2\langle\phi^\pi(s'), \mu_P^\pi(s)\rangle + \langle\mu_P^\pi(s), \mu_P^\pi(s)\rangle.\end{aligned} \tag{78}$$

We evaluate each of the three terms in (78) using the kernel trick and the linearity of the inner product:

(i) The first term is simply the kernel evaluated at s':

$$\langle \phi^\pi(s'), \phi^\pi(s') \rangle = k^\pi(s', s').$$

(ii) For the second term, we pass the inner product through the expectation:

$$\begin{aligned} \langle \phi^\pi(s'), \mu_P^\pi(s) \rangle &= \Big\langle \phi^\pi(s'), \mathbb{E}_{S_+}[\phi^\pi(S_+)] \Big\rangle \\ &= \mathbb{E}_{S_+}[k^\pi(s', S_+)]. \end{aligned}$$

(iii) For the third term, we use the property of independent copies. Let $\tilde{S}_+$ and S_+ be i.i.d. samples drawn from $P_\pi(\cdot|s)$. We write the squared norm of the expectation as the expectation over the joint distribution of these independent copies:

$$\begin{aligned} \langle \mu_P^\pi(s), \mu_P^\pi(s) \rangle &= \Big\langle \mathbb{E}_{S_+}[\phi^\pi(S_+)], \mathbb{E}_{\tilde{S}_+}[\phi^\pi(\tilde{S}_+)] \Big\rangle \\ &= \mathbb{E}_{S_+, \tilde{S}_+}[k^\pi(S_+, \tilde{S}_+)]. \end{aligned}$$

Substituting (i), (ii), and (iii) back into (78) results in:

$$\Delta^2(s, s') = k^\pi(s', s') - 2\mathbb{E}_{S_+}[k^\pi(s', S_+)] + \mathbb{E}_{S_+, \tilde{S}_+}[k^\pi(S_+, \tilde{S}_+)]. \tag{79}$$

By (16), this is equal to $\mathrm{MMD}^2_{\mathcal{H}}(\delta_{s'}, P_\pi(\cdot|s))$. Taking the square root and then taking the essential supremum over s' and the supremum over $s \in \mathscr{R}_\pi$ completes the proof for K^π_ϕ.

Proof of Part 2 ($\bar{\sigma}^\pi_\phi$): Let the conditional variance of the feature map given a state $s \in \mathscr{R}_\pi$ be defined as

$$\mathbb{V}^2_\phi(s) \coloneqq \mathbb{E}_{S_+}\left[\|\phi^\pi(S_+) - \mu_P^\pi(s)\|^2_{\mathcal{H}} \;\middle|\; s \right]. \tag{80}$$

Notice that the term inside the expectation is exactly the squared distance $\Delta^2(s, S_+)$ we analyzed in Part 1. Therefore, the conditional variance is simply the expected value of the squared MMD over the transition distribution:

$$\mathbb{V}^2_\phi(s) = \mathbb{E}_{S_+ \sim P_\pi(\cdot|s)}\Big[\Delta^2(s, S_+)\Big] = \mathbb{E}_{S_+ \sim P_\pi(\cdot|s)}\Big[\mathrm{MMD}^2_{\mathcal{H}}(\delta_{S_+}, P_\pi(\cdot|s))\Big]. \tag{81}$$

□

B.9 PROOF OF THEOREM 6

To prove Theorem 6, we first prove a lemma that quantifies how the finite-time estimation error of the value function affects the statistics K^π and $\bar{\sigma}^\pi$.

Lemma 4. *Suppose the value function estimation error satisfies $\|\hat{V}^\pi_n - V^\pi\|_\infty \leq \epsilon$, almost surely. Then, we have the following upper-bounds:*

$$\begin{aligned} K^\pi &\leq \hat{K}^\pi_n + 2\epsilon, \\ \bar{\sigma}^\pi &\leq \hat{\bar{\sigma}}^\pi_n + \epsilon. \end{aligned}$$

Proof of Lemma 4. Let $Z(s) \coloneqq V^\pi(s) - \hat{V}^\pi_n(s)$ denote the estimation error function. By assumption, $\|Z\|_\infty \leq \epsilon$, implying $|Z(s)| \leq \epsilon$ for all $s \in \mathcal{S}$.

Bound on K^π: By definition, $K^\pi = \sup_{s \in \mathscr{R}_\pi} \operatorname{ess\,sup}_{s'} \left| V^\pi(s') - \mathbb{E}_\pi[V^\pi(S_+) \mid s] \right|$. Substituting $V^\pi = \hat{V}^\pi_n + Z$ and applying the triangle inequality, we get:

$$\Big| V^\pi(s') - \mathbb{E}_\pi[V^\pi(S_+) \mid s] \Big| \leq \Big| \hat{V}^\pi_n(s') - \mathbb{E}_\pi[\hat{V}^\pi_n(S_+) \mid s] \Big| + \Big| Z(s') - \mathbb{E}_\pi[Z(S_+) \mid s] \Big|.$$

The error term is bounded by $|Z(s')| + \mathbb{E}_\pi[|Z(S_+)| \mid s] \leq \epsilon + \epsilon = 2\epsilon$. Taking the supremum over all transitions gives $K^\pi \leq \hat{K}^\pi_n + 2\epsilon$.

Bound on $\bar{\sigma}^\pi$: Using the triangle inequality for standard deviations (Minkowski's inequality), we have:

$$\sqrt{\mathbb{V}_\pi[V^\pi(S_+) \mid s]} \leq \sqrt{\mathbb{V}_\pi[\hat{V}_n^\pi(S_+) \mid s]} + \sqrt{\mathbb{V}_\pi[Z(S_+) \mid s]}.$$

Since $Z(S_+)$ is bounded by ϵ, its variance is bounded by ϵ^2, so its standard deviation is at most ϵ. Taking the supremum over all states yields $\bar{\sigma}^\pi \leq \hat{\bar{\sigma}}_n^\pi + \epsilon$. □

We are now ready to prove the main theorem.

Proof of Theorem 6. Let $\mathcal{E}_1$ be the event that the estimation error is upper-bounded by $U_\epsilon\big(n, \frac{\delta}{2}\big)$. By the consistency assumption of our estimator and setting the failure probability to $\delta/2$, we get:

$$\mathbb{P}(\mathcal{E}_1) \coloneqq \mathbb{P}\Big(\|\hat{V}_n^\pi - V^\pi\|_\infty \leq U_\epsilon\big(n, \tfrac{\delta}{2}\big)\Big) \geq 1 - \frac{\delta}{2}. \tag{82}$$

Let $\epsilon \coloneqq U_\epsilon(n, \frac{\delta}{2})$. Lemma 4 implies that conditioned on $\mathcal{E}_1$, we have $K^\pi \leq \hat{K}_n^\pi + 2\epsilon$ and $\bar{\sigma}^\pi \leq \hat{\bar{\sigma}}_n^\pi + \epsilon$.

Furthermore, conditioned on $\mathcal{E}_1$, the difference between the true expected boundary values and the estimated boundary values is strictly bounded by:

$$\begin{aligned}\Big|\big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big) - \big(\hat{V}_n^\pi(S_0) - \gamma^T \hat{V}_n^\pi(S_T)\big)\Big| &\leq |V^\pi(S_0) - \hat{V}_n^\pi(S_0)| + \gamma^T |V^\pi(S_T) - \hat{V}_n^\pi(S_T)| \\ &\leq \epsilon + \gamma^T \epsilon = \epsilon(1 + \gamma^T).\end{aligned} \tag{83}$$

We now prove each part of the theorem separately.

Proof of Part 1: Let $\mathcal{E}_{2,\ell}$ be the event that the true return concentrates around the true expected value difference as in Theorem 2. By setting the failure probability to $\delta/2$ in Theorem 2, we have $\mathbb{P}(\mathcal{E}_{2,\ell}) \geq 1 - \delta/2$, where under this event:

$$\Big|R_T^\pi - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big)\Big| \leq U_\ell\Big(T, \bar{\sigma}^\pi, K^\pi, \tfrac{\delta}{2}\Big). \tag{84}$$

By the union bound, the joint event $\mathcal{E}_1 \cap \mathcal{E}_{2,\ell}$ holds with probability at least $1 - (\delta/2 + \delta/2) = 1 - \delta$. Conditioned on this joint event, we decompose the total error using the triangle inequality:

$$\begin{aligned}\Big|R_T^\pi - \big(\hat{V}_n^\pi(S_0) - \gamma^T \hat{V}_n^\pi(S_T)\big)\Big| &\leq \Big|R_T^\pi - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big)\Big| \\ &\quad + \Big|\big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big) - \big(\hat{V}_n^\pi(S_0) - \gamma^T \hat{V}_n^\pi(S_T)\big)\Big| \\ &\leq U_\ell\Big(T, \bar{\sigma}^\pi, K^\pi, \tfrac{\delta}{2}\Big) + \epsilon(1 + \gamma^T).\end{aligned}$$

Because the bound function $U_\ell(T, \sigma, K, \delta)$ is monotonically increasing with respect to both σ and K, we can substitute the empirical upper bounds derived from $\mathcal{E}_1$. Substituting $\bar{\sigma}^\pi \leq \hat{\bar{\sigma}}_n^\pi + \epsilon$ and $K^\pi \leq \hat{K}_n^\pi + 2\epsilon$ gives the final result for Part 1.

Proof of Part 2: Let $\mathcal{E}_{2,a}$ be the event that the return concentrates around the true expected value difference as in Theorem 1. By setting the failure probability to $\delta/2$ in Theorem 1, we have $\mathbb{P}(\mathcal{E}_{2,a}) \geq 1 - \delta/2$, where under this event:

$$\Big|R_T^\pi - \big(V^\pi(S_0) - \gamma^T V^\pi(S_T)\big)\Big| \leq U_a\Big(T, K^\pi, \tfrac{\delta}{2}\Big). \tag{85}$$

By the union bound, the joint event $\mathcal{E}_1 \cap \mathcal{E}_{2,a}$ holds with probability at least $1 - (\delta/2 + \delta/2) = 1 - \delta$. Conditioned on this joint event, by applying the triangle inequality, we have:

$$\Big|R_T^\pi - \big(\hat{V}_n^\pi(S_0) - \gamma^T \hat{V}_n^\pi(S_T)\big)\Big| \leq U_a\Big(T, K^\pi, \tfrac{\delta}{2}\Big) + \epsilon(1 + \gamma^T).$$

The function $U_a(T, K, \delta)$ is monotonically increasing with respect to K. Substituting the empirical upper bound $K^\pi \leq \hat{K}_n^\pi + 2\epsilon$ derived from $\mathcal{E}_1$, gives the final result for Part 2. □