Mitigating Steady-State Bias in Off-Policy TD Learning via Distributional Correction

Anonymous authors
Paper under double-blind review

Abstract

We explore the off-policy value prediction problem in the reinforcement learning setting, where one estimates the value function of the target policy using the sample trajectories obtained from a behaviour policy. Applying importance sampling based methods are typically a go-to approach for getting such estimates but tend to suffer high error in long-horizon problems since it can only correct single-step discrepancies and fails to address steady-state bias - skewed state visitation under the behavior policy. In this paper, we present an algorithm for alleviating this bias in the off-policy value prediction using linear function approximation by correcting the state visitation distribution discrepancies. We establish rigorous theoretical guarantees, proving asymptotic convergence under Markov noise with ergodicity and demonstrating that the spectral properties of the corrected update matrix ensure stability. Most significantly, we derive an error decomposition showing that the total estimation error is bounded by a constant multiple of the best achievable approximation within the function class, where this constant transparently depends on distribution estimation quality and feature design. Empirical evaluation across multiple benchmark domains demonstrates that our method effectively mitigates steady-state bias and can be a viable alternative to existing methods in scenarios where distributional shift is critical.

1 INTRODUCTION

In the Reinforcement Learning (RL) setting Sutton & Barto (2018); Bertsekas (2019); Meyn (2022), an agent learns to interact with an environment to achieve a goal or maximize its cumulative reward by performing specific actions and receiving feedback from the environment in the form of rewards. RL is a dynamic and adaptive approach to learning, where the RL agent gains knowledge from this feedback and iteratively modifies its behaviour to produce the intended result. The central proposition in reinforcement learning is the ability to use incoming data about earlier decisions and their rewards to conclude how alternative decision policies could perform and update their course of action. RL has been successfully applied in various applications, such as game simulations Silver et al. (2018), robotics, autonomous driving Kiran et al. (2021), medicine Yom-Tov et al. (2017); Tejedor et al. (2020) and communication systems Huang et al. (2019). In game simulations, RL agents can learn to play at superhuman levels by exploring and exploiting different strategies. In robotics, RL can be used to train robots to perform complex tasks, such as grasping objects or navigating in unknown environments. In autonomous driving, RL can help autonomous vehicles learn to make safe and efficient driving decisions. In communication systems, RL can be used to optimize resource allocation, such as allocating bandwidth or power, to improve communication efficiency.

In this paper, we consider the policy evaluation problem in reinforcement learning, which refers to the task of estimating the value function, which represents the expected cumulative reward from a given state following a certain policy under linear function approximation. The policy evaluation problem has two variants: on-policy and off-policy. In on-policy prediction, one tries to estimate the value function corresponding to a given target policy using the sample trajectories generated using that target policy itself. However, in the off-policy variant, one intends to learn the value function using a sample trajectory generated using a behavior policy that may be different from the target policy. The behavior policy is the policy followed during data collection, and the target policy is the policy for which the value function is being estimated. This

allows for more flexibility, as the agent can learn from data collected by exploring different policies, which can be more efficient in terms of data collection and exploration. Off-policy methods are commonly used in scenarios where the agent needs to learn from existing data or when multiple policies are used for exploration and exploitation. Off-policy learning has significant practical implications in large-scale settings as multiple value functions can be learned from a single stream of sample trajectory in a model-free fashion, leading to parallel and optimal learning. Off-policy estimation methods often rely on importance sampling (IS) Rubinstein (1981); Glynn & Iglehart (1989) because it is an unbiased estimator. The fundamental concept behind IS Tokdar & Kass (2010) is to correct the samples obtained from a sample trajectory generated by a behaviour policy to align with the likelihood of that trajectory occurring under the targeted policy. The importance sampling approach integrated with many on-policy variants such as gradient temporal difference Sutton et al. (2009); Yu (2017), temporal difference with correction Sutton et al. (2009); Yu (2017), and temporal different with eligibility traces Precup et al. (2001) to obtain the off-policy solution. However, an important drawback of this technique is its susceptibility to imprecision because of the high variance induced by the importance weights Mandel et al. (2014) and the the discrepancies associated with state appearance probabilities Tsitsiklis & Van Roy (1997). In tabular temporal difference (TD) learning, state-value estimates converge pointwise to the unique solution of the Bellman equation under ergodicity and coverage conditions, where steady-state distributions of the target or bahaviour policies (depending on on/off variant) influence only convergence rates through state-visitation frequencies without altering asymptotic values. Conversely, under linear function approximation, the solution is characterized by a ν-weighted projection, where ν is either the steady-state distribution of the target policy (on-policy) or of the behaviour policy (off-policy). This introduces inherent asymptotic bias as minimization of Bellman error is computed under ν rather than approximating value function directly.

In this paper, we analyze the deviation of the on-policy solution from the off-policy solution due to the steady-state bias which arises due to the discrepancy in the steady-state distribution induced by the target and behaviour policies under linear function approximation. When one observes the marginal distributions from the target policy and behaviour policy after a sufficiently long time (mixing time), the marginal distributions settle down to the steady state which is unique to the corresponding Markov chain. When estimating state values in which the distribution of states visited during the episode is different from the distribution of states visited in steady-state under the target policy, steady state bias occurs. This can eventuate when the setting only considers a subset of the possible states or actions, or when it does not sample states or actions uniformly. As a result, the average return may not accurately reflect the true expected return, which can lead to sub-optimal behaviour. Off-policy bias correction is a fundamental challenge in reinforcement learning, particularly in settings that utilize experience replay or batch data from previously executed policies Precup et al. (2001); Sutton et al. (2016). The systematic bias introduced into the value function estimation due to the discrepancy between the behavior policy's stationary distribution and the target policy's state visitation distribution persists even with unbiased importance sampling corrections as it stems from long-horizon distributional mismatch rather than single-step policy differences Chandak et al. (2021). This steady-state bias becomes particularly problematic in long-horizon tasks where distributional mismatch accumulates over time Jiang & Li (2016); Tang et al. (2020). Recent approaches have addressed this by estimating stationary distribution corrections through marginalized importance sampling Liu et al. (2018; 2019), dual function approximation Zhang et al. (2020), or direct optimization of distribution matching objectives Nachum et al. (2019); Yang et al. (2020). Frameworks like Universal Off-Policy Evaluation Chandak et al. (2021) further improve estimation by enforcing consistency between learned value functions and off-policy estimators. These methods often formulate the correction as a minimax optimization problem over density ratios or leverage temporal difference learning with generalized advantage estimation Schulman et al. (2015). The convergence and stability of such methods are closely tied to the "deadly triad" of function approximation, off-policy training, and bootstrapping, which can lead to divergence without careful regularization or correction mechanisms Voloshin et al. (2019); Wang et al. (2017); Yu (2017).

In this paper, we fundamentally analyze off-policy temporal difference learning under linear function approximation by tackling the critical problem of steady-state distribution mismatch. We rigorously demonstrate that long-horizon bias stems not only from policy differences but from the divergence in how states are visited under target versus behavior policies. Our analysis reveals how this distributional shift magnifies approximation errors through the deadly triad of bootstrapping, function approximation, and off-policy sampling. To

address this, we introduce a dual correction mechanism—blending standard per-step action reweighting with novel parametric estimation of stationary distribution discrepancies—and prove its asymptotic convergence under Markov noise with ergodicity. We further establish that the corrected value estimates stabilize when the rebalanced Bellman operator exhibits spectral negativity. Most significantly, we derive an error decomposition showing that the total value estimation error is bounded by a constant multiple of the best achievable approximation, where this constant transparently depends on the accuracy of distribution estimation, the conditioning of feature representations, and the degree of policy misalignment.

Our work distinguishes itself from prior distribution correction methods by providing a targeted solution to the persistent problem of steady-state bias in off-policy TD learning. While frameworks like DualDICE Nachum et al. (2019) and GradientDICE Zhang et al. (2020) address general distribution ratio estimation through complex dual optimization, our approach offers a direct, computationally efficient correction that integrates seamlessly with standard TD updates. Unlike policy optimization methods such as CQL Kumar et al. (2020) or OptiDICE Lee et al. (2021) which focus on policy improvement, we specifically address value prediction accuracy under distributional shift. While Liu et al. (2018) addressed the "curse of horizon" through marginalized importance sampling, our approach uniquely identifies and corrects for the persistent steady-state bias that remains even after one-step importance sampling is applied, offering a complementary perspective on the distributional mismatch problem in off-policy evaluation. Also Emphatic TD Sutton et al. (2016); Yu (2015), employs recursive emphasis weighting to implicitly approximate distribution correction preventing the deadly triad (function approximation + bootstrapping + off-policy learning) from causing divergence using emphasis weights to prioritize updates for states that are important to the target policy. Our focused approach—correcting steady-state bias through explicit distribution modeling rather than general optimization frameworks—provides both theoretical clarity and practical advantages for the fundamental problem of off-policy value prediction.

2 BACKGROUND

The reinforcement learning setting is an optimal sequential decision-making paradigm under uncertainty characterized as Markov Decision Process (MDP) Puterman (2014); Bertsekas (2019); Meyn (2022), which is a controlled, time-homogeneous, stochastic process that is defined by the 4-tuple (S, A, P, R), where S is the state space and A is the action space. In this paper, we consider a finite state and action spaces with $S = \{s^1, s^2, ..., s^n\}$. Here $P : S \times A \times S \to [0, 1]$ is the probability transition function, where $P(s, a, s') = \mathbb{P}(\mathbf{s}_{t+1} = s' | \mathbf{s}_t = s, \mathbf{a}_t = a, \mathbf{s}_{t-1} = \cdot, \mathbf{a}_{t-1} = \cdot, \dots) = \mathbb{P}(\mathbf{s}_{t+1} = s' | \mathbf{s}_t = s, \mathbf{a}_t = a)$ is the probability that the next state is s' conditioned on the fact that the current state is s and current action is a. Additionally, the reward function $R : S \times A \times S \to \mathbb{R}$ assigns a numerical reward to each transition. P and R define the dynamics of the stochastic system. At each instant, an action is chosen according to a stationary stochastic policy $\pi : S \times A \to [0, 1]$, where $\pi(\cdot|s)$ is a probability mass function over the action space A conditioned on the state $s \in S$.

In this paper, we consider the prediction problem in reinforcement learning which is defined as follows: For a given target policy π and discount factor $\gamma \in [0, 1)$ (that represents the agent's preference for immediate rewards versus future rewards), the goal is to evaluate the value function $V_\pi \in \mathbb{R}^n$ associated with the target policy which is defined as the expected long-run γ-discounted cost:

$$V_\pi(s) = \mathop{\mathbb{E}}_{\tau \sim \pi} \Big[R(\tau) | \mathbf{s}_0 = s \Big], s \in S, \tag{1}$$

where $R(\tau) = \sum_{t=0}^{\infty} \gamma^t R(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$, with $\mathbf{s}_t$ represents the state at instant t, the $\mathbf{a}_t \sim \pi(\cdot|\mathbf{s}_t)$ represents the action chosen at time t and $\mathbf{s}_{t+1} \sim P(\mathbf{s}_t, \mathbf{a}_t, \cdot)$ represents the next state. Note that the above definition is well-defined as $\gamma \in [0, 1)$ and by appealing to the bounded convergence theorem.

The value function in vector form is expressed as $V_\pi = [V_\pi(s^1), V_\pi(s^2), \dots, V_\pi(s^n)]^\top \in \mathbb{R}^n$. The value function V_π satisfies the Bellman equation: $V_\pi = T_\pi V_\pi$, where $T_\pi : \mathbb{R}^n \to \mathbb{R}^n$ is Bellman operator with $T_\pi U = \bar{R}_\pi + \gamma P_\pi U$. Here, $P_\pi \in \mathbb{R}^{n \times n}$ with $[P_\pi]_{ss'} = \sum_{a \in A} \pi(a|s) P(s, a, s')$ and $\bar{R}_\pi(s) = \sum_{s' \in S} \sum_{a \in A} \pi(a|s) P(s, a, s') R(s, a, s')$ is the one-step average reward. From the Bellman equation, one can directly compute $V_\pi = (I - \gamma P_\pi)^{-1} \bar{R}_\pi$ whose computational complexity is $O(n^3)$. In the RL setting, the

model parameter P and R are unknown and one seeks to learn the value function V_π under the generative model setting, where a realization of the stochastic process in the form of an infinitely long sample trajectory $\mathbf{s}_0, \mathbf{a}_0, \mathbf{r}_1, \mathbf{s}_1, \mathbf{a}_1, \mathbf{r}_2, \mathbf{s}_2, \dots$ is available, with $\mathbf{s}_0 \sim d$, $\mathbf{a}_t \sim \pi(\cdot|\mathbf{s}_t)$, $\mathbf{s}_{t+1} \sim P(\mathbf{s}_t, \mathbf{a}_t, \cdot)$ and $\mathbf{r}_{t+1} = R(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$.

Temporal difference (TD) learning Sutton & Barto (2018) is the classical approach for the prediction problem, where the value function $V_t \in \mathbb{R}^n$ is iteratively updated in the direction of the temporal difference $\mathbf{r}_{t+1} + \gamma V_t(\mathbf{s}_{t+1}) - V_t(\mathbf{s}_t)$. However, when the state space is large, this method suffers from the curse of dimensionality Tsitsiklis & Van Roy (1997); Sutton & Barto (2018). To overcome this, one effective strategy is to represent the value function in a lower-dimensional subspace, thus reducing computational and storage demands. Here one approximates V_π using linear function approximation by projecting it into the subspace $\{\Phi x \mid x \in \mathbb{R}^k\} \subset \mathbb{R}^n$, where $k \ll n$ Tsitsiklis & Van Roy (1997). The feature matrix Φ contains basis functions that capture the critical characteristics of the state space. This projection not only renders the learning process more tractable but also preserves the essential dynamics of the original high-dimensional problem.

$$\Phi = \begin{bmatrix} - - \phi(s^1)^\top - - \\ \vdots \\ - - \phi(s^n)^\top - - \end{bmatrix}_{n \times k}, \tag{2}$$

where, $\phi(s) = [\phi_1(s), \phi_2(s), \dots \phi_k(s)]^\top \in \mathbb{R}^k$ is called the feature vector associated with state $s \in S$ and $\phi_i : S \to R$ are feature/basis functions. The most commonly used parameterized basis functions include Radial Basis Functions (RBFs), polynomials, and Fourier basis functions. Radial Basis Functions are typically expressed in a Gaussian form: $\phi_i(s) = \exp(-(2\sigma_i^2)^{-1}||s - \mu_i||^2)$ depends solely on the distance between the state and the centre μ_i, relative to the feature width, σ_i, with parameter size of the order $\Theta(k)$.

In this paper, we consider the off-policy variant of the prediction problem Precup et al. (2001); Sutton & Barto (2018), where one seeks to estimate V_π, using a sample trajectory, where action at every instant is generated using a behaviour policy π_b that may be different from the target policy π. This implies that for the given infinitely long sample trajectory $\tau_b = \mathbf{s}_0, \mathbf{a}_0, \mathbf{r}_1, \mathbf{s}_1, \mathbf{a}_1, \mathbf{r}_2, \mathbf{s}_2, \mathbf{a}_2, \dots$, we have $\mathbf{s}_0 \sim P_0$ (P_0 initial distribution), $\mathbf{a}_t \sim \pi_b(\cdot|\mathbf{s}_t)$, $\mathbf{s}_{t+1} \sim P(\mathbf{s}_t, \mathbf{a}_t, \cdot)$ and $\mathbf{r}_{t+1} = R(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})$.

Assumption 1 (Ergodic Behavior Policy)**.** *The Markov chain $\{s_t\}_{t \geq 0}$ induced by the behavior policy π_b satisfies:*

(i) ***Irreducibility:*** *$\forall\, s, s' \in S$, $\exists\, t \in \mathbb{N}$ such that $P^t_{\pi_b}(s, s') > 0$.*

(ii) ***Aperiodicity:*** *The greatest common divisor of $\{\, t \geq 1 : P^t_{\pi_b}(s, s) > 0 \,\}$ is 1 for every $s \in S$.*

Consequently, the chain admits a unique stationary distribution ν_b with $\nu_b(s) > 0\ \forall\, s \in S$ and $\nu_b^\top P_{\pi_b} = \nu_b^\top$.

Assumption 2 (Feature Independence)**.** *The feature matrix $\Phi \in \mathbb{R}^{n \times k}$ satisfies* $\operatorname{rank}(\Phi) = k$*, implying*

$$\sigma_{\min}\big(\Phi^\top \Phi\big) > 0, \qquad \ker(\Phi) = \{0\}, \qquad \Phi^\top \Phi \succ 0,$$

where $\sigma_{\min}$ denotes the minimum singular value and $\succ 0$ denotes positive definiteness.

Assumption 3 (Coverage)**.** *The behavior policy π_b dominates π in the Radon–Nikodym sense:*

$$\forall\, (s, a) \in S \times A,\ \pi(a \mid s) > 0 \implies \pi_b(a \mid s) > 0.$$

Equivalently, the importance ratio $\rho_t = \dfrac{\pi(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}$ is almost surely bounded: $\sup_t \rho_t < \infty$.

In off-policy linear function approximation, one projects the value function V_π onto the column space of Φ Tsitsiklis & Van Roy (1997):

$$w^* = \arg\min_{w \in \mathbb{R}^k} \|V_\pi - \Phi w\|^2_{\nu_b}, \tag{3}$$

where the weighted norm is defined as $\|w\|^2_\nu = \sum_{i=1}^k \nu_i w_i^2$. Here, ν_b is the unique steady-state distribution of the behavior policy's Markov chain (i.e., $\nu_b(s) = \lim_{t \to \infty} \mathbb{P}(s_t = s)$ and $\nu_b^\top P_{\pi_b} = \nu_b^\top$). Since $\{\Phi w \mid w \in \mathbb{R}^k\}$ is

closed and convex, a unique w^* exists (Φ has full column rank), yielding the approximation $V_\pi(s) \approx \phi(s)^\top w^*$ for all s. This optimization is solved by the off-policy TD update Precup et al. (2001; 2000):

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha_t \rho_t \Big(\mathbf{r}_{t+1} + \gamma \phi(\mathbf{s}_{t+1})^\top \mathbf{w}_t - \phi(\mathbf{s}_t)^\top \mathbf{w}_t \Big) \phi(\mathbf{s}_t),$$

with the importance sampling ratio $\rho_t = \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_b(\mathbf{a}_t|\mathbf{s}_t)}$, which corrects for the policy mismatch in the behavior data. The limit point $w_{\text{off}}^{\text{TD}}$ of off-policy TD learning with linear function approximation is characterized by the fixed-point equationYu (2012)

$$\Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi) \Phi w_{\text{off}}^{\text{TD}} = \Phi^\top \Xi_{\nu_b} R_\pi, \tag{4}$$

which represents a projection of the Bellman equation onto the feature space weighted by the behavior policy's stationary distribution. This solution constitutes the best approximation within the function class that satisfies the Bellman residual minimization under the behavior policy's steady-state distributional mismatch, rather than the target policy's natural state visitation pattern.

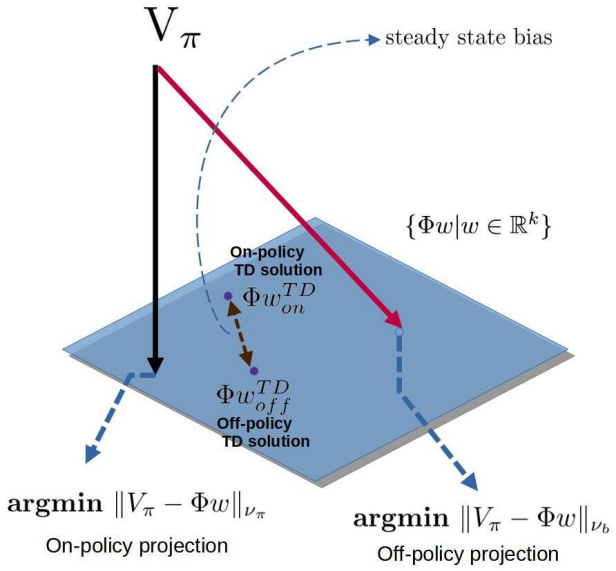

Figure 1: Illustration of steady-state bias in off-policy prediction: The mismatch between behavior policy's steady-state distribution ν_b and target policy's distribution causes persistent prediction error, even after one-step importance sampling correction

To establish further theoretical guarantees for the off-policy TD method with linear function approximation, we first analyze key properties of the value function operator under the behavior policy's stationary distribution. The following lemma quantifies fundamental operator norm bounds that govern the propagation of approximation errors through the Bellman operator.

Lemma 1. *Let ν_b be a strictly positive probability distribution over states, P_π a Markov transition matrix induced by policy π, and $\gamma \in [0, 1)$ a discount factor. Then the ν_b-weighted operator norms satisfy:*

$$\|P_\pi\|_{\nu_b} \leq \sqrt{\kappa_b} \text{ and } \|I - \gamma P_\pi\|_{\nu_b} \leq 1 + \gamma\sqrt{\kappa_b} \tag{5}$$

where $\|A\|_{\nu_b} = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_{\nu_b}}{\|\mathbf{x}\|_{\nu_b}}$ and $\|\mathbf{x}\|_{\nu_b}^2 = \sum_s \nu_b(s)\mathbf{x}(s)^2$, with and the distribution mismatch coefficient $\kappa_b = \max_{s' \in S} \dfrac{\sum_s \nu_b(s) P_\pi(s' \mid s)}{\nu_b(s')}$.

Proof. Let $\mu_b(s') = \sum_s \nu_b(s) P_\pi(s' \mid s)$. Then,

$$\begin{aligned}
\|P_\pi \mathbf{x}\|_{\nu_b}^2 &= \sum_s \nu_b(s) \left(\sum_{s'} P_\pi(s' \mid s) \mathbf{x}(s') \right)^2 \\
&\leq \sum_s \nu_b(s) \sum_{s'} P_\pi(s' \mid s) \mathbf{x}(s')^2 \quad \text{(Jensen's inequality)} \\
&= \sum_{s'} \mathbf{x}(s')^2 \sum_s \nu_b(s) P_\pi(s' \mid s) \\
&= \sum_{s'} \mathbf{x}(s')^2 \mu_b(s') \\
&= \sum_{s'} \nu_b(s') \mathbf{x}(s')^2 \frac{\mu_b(s')}{\nu_b(s')} \leq \kappa_b \sum_{s'} \nu_b(s') \mathbf{x}(s')^2 = \kappa_b \|\mathbf{x}\|_{\nu_b}^2, \text{ where } \kappa_b = \max_{s'} \frac{\mu_b(s')}{\nu_b(s')}.
\end{aligned}$$

Thus we have the operator norm bound:

$$\|P_\pi\|_{\nu_b} = \sup_{\mathbf{x} \neq 0} \frac{\|P_\pi \mathbf{x}\|_{\nu_b}}{\|\mathbf{x}\|_{\nu_b}} \leq \sqrt{\kappa_b}$$

For the composite operator:

$$\begin{aligned}
\|(I - \gamma P_\pi)\mathbf{x}\|_{\nu_b} &\leq \underbrace{\|I\mathbf{x}\|_{\nu_b}}_{\leq \|\mathbf{x}\|_{\nu_b}} + \gamma \underbrace{\|P_\pi \mathbf{x}\|_{\nu_b}}_{\leq \sqrt{\kappa_b}\|\mathbf{x}\|_{\nu_b}} \leq (1 + \gamma\sqrt{\kappa_b})\|\mathbf{x}\|_{\nu_b} \\
&\Rightarrow \|I - \gamma P_\pi\|_{\nu_b} \leq 1 + \gamma\sqrt{\kappa_b}. \qquad (6)
\end{aligned}$$

□

Central to the above result is the *distribution mismatch coefficient* κ_b, which captures the maximum density ratio between the next-state distribution induced by the target policy and the stationary distribution of the behavior policy. We now characterize the asymptotic approximation error of the off-policy TD solution under linear function approximation. The following theorem establishes a bound on the error $\|\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi\|_{\nu_b}$ of the TD fixed point solution relative to the fundamental approximation limit $\|\Phi \mathbf{w}^* - V_\pi\|_{\nu_b}$.

Theorem 1 (Error Bound for Off-policy TD)**.** *Under Assumptions 1-3 and negative definiteness of* $\Lambda_o = \Phi^\top \Xi_{\nu_b}(I - \gamma P_\pi)\Phi$*, the solution* $\mathbf{w}_{\text{off}}^{\text{TD}}$ *satisfies:*

$$\left\|\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi\right\|_{\nu_b} \leq \left(\frac{\sigma_{\max}^2(\Phi)(\max_s \nu_b(s))^{3/2}(1 + \gamma\sqrt{\kappa_b})}{\lambda_{\min}(-\Lambda_o)\sqrt{\min_s \nu_b(s)}} + 1 \right) \left\|\Phi \mathbf{w}^* - V_\pi\right\|_{\nu_b}$$

where $\sigma_{\max}^2(\Phi) = \lambda_{max}(\Phi^\top \Phi)$ *and* $\Xi_{\nu_b} = diag(\nu_b)$*.*

Proof. From equation 4,

$$\Phi^\top \Xi_{\nu_b}\big(I - \gamma P_\pi\big)\Phi\, \mathbf{w}_{\text{off}}^{\text{TD}} = \Phi^\top \Xi_{\nu_b} R_\pi. \qquad (7)$$

The true value function V_π satisfies the Bellman equation:

$$V_\pi = R_\pi + \gamma P_\pi V_\pi \qquad (8)$$

We wish to bound $\left\|\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi\right\|_{\nu_b}$. Let $\mathbf{w}^*$ be the best linear approximator under ν_b:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\Phi \mathbf{w} - V_\pi\|_{\nu_b}$$

so that $\Phi \mathbf{w}^* = \Pi_{\nu_b} V_\pi$, the projection of V_π onto the column space of Φ under the ν_b-weighted norm. The error decomposes as

$$\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi \;=\; \big(\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - \Phi \mathbf{w}^*\big) \;+\; \big(\Phi \mathbf{w}^* - V_\pi\big).$$

Hence

$$\left\|\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi\right\|_{\nu_b} \;\leq\; \left\|\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - \Phi \mathbf{w}^*\right\|_{\nu_b} \;+\; \left\|\Phi \mathbf{w}^* - V_\pi\right\|_{\nu_b}.$$

Define the approximation error $\varepsilon_{\text{approx}} \;=\; \left\|\Phi \mathbf{w}^* - V_\pi\right\|_{\nu_b}$. We focus on the first term above. Note that both $\Phi \mathbf{w}_{\text{off}}^{\text{TD}}$ and Φw^* lie in the column space of Φ. The vector $\Phi w_{\text{TD}}^{\text{off}}$ satisfies

$$\Phi^\top \Xi_{\nu_b} \big(I - \gamma P_\pi\big) \Phi w_{\text{TD}}^{\text{off}} \;=\; \Phi^\top \Xi_{\nu_b} R_\pi, \tag{9}$$

whereas the projection $\Phi \mathbf{w}^*$ satisfies

$$\Phi^\top \Xi_{\nu_b} \big(\Phi \mathbf{w}^* - V_\pi\big) \;=\; 0.$$

Also, by multiplying the Bellman equation (8) by $\Phi^\top \Xi_{\nu_b}$ yields

$$\begin{aligned} \Phi^\top \Xi_{\nu_b} V_\pi \;&=\; \Phi^\top \Xi_{\nu_b} R_\pi + \gamma\, \Phi^\top \Xi_{\nu_b} P_\pi V_\pi \\ \Rightarrow \Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi) V_\pi \;&=\; \Phi^\top \Xi_{\nu_b} R_\pi. \end{aligned} \tag{10}$$

Combining (9) and (10), we get

$$\begin{aligned} &\Phi^\top \Xi_{\nu_b} \left(I - \gamma P_\pi\right) \Phi\, \mathbf{w}_{\text{off}}^{\text{TD}} \;=\; \Phi^\top \Xi_{\nu_b} \left(I - \gamma P_\pi\right) V_\pi \\ &\Rightarrow \Phi^\top \Xi_{\nu_b} \left(I - \gamma P_\pi\right) (\Phi\, \mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi) \;=\; 0 \\ &\Rightarrow \Phi^\top \Xi_{\nu_b} \left(I - \gamma P_\pi\right) e \;=\; 0, \end{aligned} \tag{11}$$

where $e \;=\; \Phi\, \mathbf{w}_{\text{off}}^{\text{TD}} \;-\; V_\pi$. Further,

$$\begin{aligned} e &= \big(\Phi \mathbf{w}_{\text{off}}^{\text{TD}} - \Phi \mathbf{w}^*\big) \;+\; \big(\Phi \mathbf{w}^* - V_\pi\big) \\ &= \Phi \left(\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\right) \;+\; \varepsilon_{\text{approx}}, \end{aligned} \tag{12}$$

where $\varepsilon_{\text{approx}} = \Phi \mathbf{w}^* - V_\pi$. Substituting above,

$$\begin{aligned} &\Phi^\top \Xi_{\nu_b} \left(I - \gamma P_\pi\right) \Big[\, \Phi \big(\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\big) \;+\; \varepsilon_{\text{approx}} \Big] \;=\; 0 \\ &\Rightarrow \Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi)\, \Phi \big(\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\big) \;+\; \Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi)\, \varepsilon_{\text{approx}} \;=\; 0 \\ &\Rightarrow \Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi)\, \Phi \big(\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\big) \;=\; -\, \Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi)\, \varepsilon_{\text{approx}} \\ &\Rightarrow \Lambda_o \big(\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\big) \;=\; \Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi)\, \varepsilon_{\text{approx}}. \end{aligned} \tag{13}$$

Now,

$$\left\|\Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi) \varepsilon_{\text{approx}}\right\| \;\leq\; \|\Phi^\top\|\, \|\Xi_{\nu_b}\|\, \left\|(I - \gamma P_\pi) \varepsilon_{\text{approx}}\right\|, \tag{14}$$

where $\|\cdot\|$ is the spectral norm. But, $\|\Xi_{\nu_b}\| = \max_s\, \nu_b(s)$, and

$$\begin{aligned} \left\|(I - \gamma P_\pi) \varepsilon_{\text{approx}}\right\|^2 &= \sum_s \Big[(I - \gamma P_\pi) \varepsilon_{\text{approx}}\Big](s)^2 \\ &\leq \sum_s \frac{1}{\min_s \nu_b(s)}\, \nu_b(s) \Big[(I - \gamma P_\pi) \varepsilon_{\text{approx}}\Big](s)^2 \\ &= \frac{1}{\min_s \nu_b(s)}\, \left\|(I - \gamma P_\pi) \varepsilon_{\text{approx}}\right\|_{\nu_b}^2. \end{aligned} \tag{15}$$

Hence,

$$\left\|(I - \gamma P_\pi) \varepsilon_{\text{approx}}\right\| \;\leq\; \frac{1 + \gamma \sqrt{\kappa_b}}{\sqrt{\min_s \nu_b(s)}}\, \|\varepsilon_{\text{approx}}\|_{\nu_b},$$

and therefore

$$\left\|\Phi^\top \Xi_{\nu_b} (I - \gamma P_\pi) \varepsilon_{\text{approx}}\right\| \;\leq\; \|\Phi^\top\|\, \frac{\max_s \nu_b(s)}{\sqrt{\min_s \nu_b(s)}} \left(1 + \gamma \sqrt{\kappa_b}\right) \|\varepsilon_{\text{approx}}\|_{\nu_b}. \tag{16}$$

Note that $\|\Phi^\top\| = \|\Phi\|$, and the spectral norm of Φ is the large singular value. Let $\sigma_{\max}(\Phi) = \|\Phi\|$. Also note that, since Λ_o is negative definite, we have

$$\|\Lambda_o^{-1}\| \;\le\; \frac{1}{\lambda_{\min}(-\Lambda_o)}, \tag{17}$$

where $\lambda_{\min}(-\Lambda_o)$ is the smallest eigenvalue of $-\Lambda_o$. Combining equation 13, (14), (16) and (17), we get

$$\|\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\| \;\le\; \frac{\sigma_{\max}(\Phi)}{\lambda_{\min}(-\Lambda_o)} \frac{\max_s \nu_b(s)}{\sqrt{\min_s \nu_b(s)}} \left(1 + \gamma\sqrt{\kappa_b}\right) \|\varepsilon_{\text{approx}}\|_{\nu_b}. \tag{18}$$

Hence, for the projected error in the ν_b-norm,

$$\|\Phi(\mathbf{w}_{\text{off}}^{\text{TD}} - w^*)\|_{\nu_b} \;\le\; \|\Phi\|_{\nu_b} \|w_{\text{TD}}^{\text{off}} - w^*\|. \tag{19}$$

$\|\Phi\|_{\nu_b}$ is the operator norm of Φ from the Euclidean space to the ν_b-normed space. Specifically:

$$\begin{aligned} &\|\Phi\mathbf{w}\|_{\nu_b}^2 = \mathbf{w}^\top \Phi^\top \Xi_{\nu_b} \Phi\mathbf{w} \le \lambda_{\max}(\Phi^\top \Xi_{\nu_b}\Phi)\|\mathbf{w}\|^2 \\ &\Rightarrow \|\Phi\|_{\nu_b} \le \sqrt{\lambda_{\max}(\Phi^\top \Xi_{\nu_b}\Phi)}. \end{aligned} \tag{20}$$

Note that $\Phi^\top \Xi_{\nu_b}\Phi$ is a $k \times k$ matrix, and its largest eigenvalue is at most $\max_s \nu_b(s) \cdot \lambda_{\max}(\Phi^\top\Phi)$, because $\Xi_{\nu_b} \le \max_s \nu_b(s) I$. And $\lambda_{\max}(\Phi^\top\Phi) = \sigma_{\max}^2(\Phi)$. Hence,

$$\|\Phi(w_{\text{off}}^{\text{TD}} - \mathbf{w}^*)\|_{\nu_b} \le \sigma_{\max}(\Phi)\sqrt{\max_s \nu_b(s)} \cdot \|\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*\| \tag{21}$$

Hence from (12), (18) and (21), we get

$$\|\Phi\mathbf{w}_{\text{off}}^{\text{TD}} - V_\pi\|_{\nu_b} \le \|\Phi(\mathbf{w}_{\text{off}}^{\text{TD}} - \mathbf{w}^*)\|_{\nu_b} + \|\epsilon_{\text{approx}}\|_{\nu_b} \le \left(\frac{\sigma_{\max}^2(\Phi)(\max_s \nu_b(s))^{\frac{3}{2}}\left(1+\gamma\sqrt{\kappa_b}\right)}{\lambda_{\min}(-\Lambda_o)\sqrt{\min_s \nu_b(s)}} + 1\right)\|\varepsilon_{\text{approx}}\|_{\nu_b}$$

□

The above theorem bound reveals three critical bottlenecks in off-policy TD convergence: First, the $\sigma_{\max}^2(\Phi)(\max_s v_b(s))^{3/2}$ term exposes the sensitivity to feature scaling and distribution skew, showing that even optimal representations suffer when v_b is non-uniform or features are poorly conditioned. Second, the $(1+\gamma\sqrt{\kappa_b})$ factor quantifies how policy divergence ($\kappa_b \gg 1$) amplifies approximation error through temporal credit assignment - a manifestation of the deadly triad where bootstrapping, function approximation, and off-policy sampling interact destructively. Third, the dependence on $\lambda_{\min}(-\Lambda_o)^{-1}$ formalizes the hardness of Bellman inversion under distribution shift, as Λ_o becomes ill-conditioned when the behavior policy's transitions poorly align with the target dynamics. This provides the first closed-form characterization of deadly triad interactions in off-policy TD convergence. The bound exclusively characterizes the fundamental approximation error of the asymptotic off-policy TD solution, isolating it from transient algorithmic effects. When $\pi = \pi_b$ ($\kappa_b = 1$), the bound simplifies to the on-policy case, but the exponential scaling $\gamma\sqrt{\kappa_b}$ explains the severe degradation under policy mismatch.

Theorem 2. *Under Assumptions 1, 2, and 3, if $\kappa\gamma^2 < 1$ then Λ_o is negative definite.*

Proof. For any $\mathbf{w} \neq 0$, let $\mathbf{u} = \Phi\mathbf{w}$. By Assumption 2, $\mathbf{u} \neq 0$. Consider the quadratic form:

$$\begin{aligned} \mathbf{w}^\top \Lambda_o \mathbf{w} &= \mathbf{w}^\top \Phi^\top \Xi_{\nu_b}(\gamma P_\pi - I)\Phi\mathbf{w} \\ &= \mathbf{u}^\top \Xi_{\nu_b}(\gamma P_\pi - I)\mathbf{u} \\ &= \gamma \underbrace{\mathbf{u}^\top \Xi_{\nu_b} P_\pi \mathbf{u}}_{Q_1} - \underbrace{\mathbf{u}^\top \Xi_{\nu_b}\mathbf{u}}_{Q_2} \end{aligned} \tag{22}$$

Now $Q_2 = \mathbf{u}^\top \Xi_{\nu_b} \mathbf{u} = \sum_s \nu_b(s)\mathbf{u}(s)^2 = \|\mathbf{u}\|_{\nu_b}^2 > 0$. Since $\nu_b > 0$ (ergodicity) and $\mathbf{u} \neq 0$, we have

$$\begin{aligned} Q_1 = \mathbf{u}^\top \Xi_{\nu_b} P_\pi \mathbf{u} &= \sum_s \nu_b(s)\mathbf{u}(s)(P_\pi \mathbf{u})(s) \\ &= \sum_s \nu_b(s)\mathbf{u}(s)\big(\sum_{s'} P_\pi(s'|s)\mathbf{u}(s')\big) \\ &\leq \|\mathbf{u}\|_{\nu_b} \cdot \|P_\pi \mathbf{u}\|_{\nu_b} (\text{Cauchy-Schwarz inequality}) \end{aligned} \tag{23}$$

Further, by Lemma 1, we have

$$\|P_\pi \mathbf{u}\|_{\nu_b}^2 = \kappa_b \|\mathbf{u}\|_{\nu_b}^2 \tag{24}$$

Therefore, from (23) and (24) :

$$Q_1 \leq |Q_1| \leq \|\mathbf{u}\|_{\nu_b} \cdot \|P_\pi \mathbf{u}\|_{\nu_b} \leq \sqrt{\kappa} \|u\|_{\nu_b}^2$$

Substitute into (22) to obtain

$$\mathbf{w}^\top \Lambda_o \mathbf{w} \leq \gamma \sqrt{\kappa_b} \|\mathbf{u}\|_{\nu_b}^2 - \|\mathbf{u}\|_{\nu_b}^2 = (\gamma \sqrt{\kappa_b} - 1) \|\mathbf{u}\|_{\nu_b}^2 \tag{25}$$

Since $\|\mathbf{u}\|_{\nu_b}^2 > 0$ and $\gamma\sqrt{\kappa_b} - 1 < 0 \quad \textit{iff} \quad \kappa_b < \frac{1}{\gamma^2}$ we have:

$$\mathbf{w}^\top \Lambda_o \mathbf{w} \leq (\gamma \sqrt{\kappa_b} - 1) \|\mathbf{u}\|_{\nu_b}^2 < 0 \quad \text{when} \quad \kappa_b \gamma^2 < 1$$

Equality holds only if $\mathbf{w} = 0$, proving Λ_o is negative definite. □

Corollary 1. *When $\pi = \pi_b$, Λ_o is negative definite for any $\gamma < 1$.*

Proof. When $\pi = \pi_b$, we have $\kappa_b = 1$. Then:

$$\mathbf{w}^\top \mathbf{A} \mathbf{w} \leq (\gamma - 1)\|\mathbf{v}\|_{\nu_b}^2 < 0 \quad \forall \mathbf{w} \neq 0$$

□

Theorem 2 establishes a fundamental condition for convergence in off-policy temporal difference learning: when the product of the policy alignment constant κ_b and the squared discount factor γ^2 is less than one, the critical matrix governing the TD update dynamics becomes negative definite. This condition, $\kappa_b \gamma^2 < 1$, provides profound theoretical insight into the feasibility of off-policy learning. The policy alignment constant κ_b quantifies the maximum discrepancy between the next-state distribution under the target policy and the behavior policy's stationary distribution (ν_b). When κ_b is large, it indicates significant distributional mismatch—certain states are visited much more frequently under the target policy than would be expected from the behavior policy's steady-state distribution. The theorem reveals that such mismatches become increasingly problematic as the discount factor γ approaches 1, explaining why long-horizon tasks with high γ values are particularly challenging for off-policy methods. Notably, when policies are identical ($\pi = \pi_b$), we have $\kappa_b = 1$, and the condition simplifies to $\gamma < 1$, which always holds for standard MDPs. However, as policy dissimilarity increases ($\kappa_b > 1$), the allowable discount factor must decrease to maintain convergence guarantees. This theoretical boundary precisely characterizes the "deadly triad" interaction between function approximation, bootstrapping, and off-policy learning, and directly motivates the steady-state bias correction.

3 Our Algorithm

Here, we propose a double correction approach to address the discrepancy introduced by the steady-state distribution of the behavior policy in the solution of the off-policy TD algorithm by effectively reducing the policy alignment constant through distributional reweighting, while simultaneously incorporating per-step policy mismatch correction ρ_t. To achieve this, we employ the importance sampling method to the existing

off-policy TD method with one-step lookahehad, where the importance sampling ratio $q(s)/h_{\theta^*,\lambda^*}(s)$ is tied to the existing TD recursion. Here $q(\cdot)$ is the target probability distribution to which the solutions are guided, and $h_{\theta^*,\lambda^*} = \lambda_1^* g_{\theta_1^*} + \cdots + \lambda_\ell^* g_{\theta_\ell^*}$ is a surrogate probability mixture distribution chosen from a parametrized family of distributions $\{g_\theta | \theta \in \Theta\}$ which best approximates the steady-state distribution of the Markov chain induced by the behaviour policy with respect to the Kullback-Leibler divergence (moment projection).

$$\begin{bmatrix}\theta^* \\ \lambda^*\end{bmatrix} = \underset{\substack{\theta_i \in \Theta, \\ \lambda_i \in [0,1]}}{\arg\min}\, \mathcal{D}_{\mathtt{KL}}(\nu_b \| \lambda_1 g_{\theta_1} + \cdots + \lambda_\ell g_{\theta_\ell}), \text{ subject to } \sum_{i=1}^{\ell} \lambda_i = 1, \tag{26}$$

where $\mathcal{D}_{\mathtt{KL}}(f\|g) = \mathbb{E}_f\left[\log \frac{f(X)}{g(X)}\right]$.

One theoretically well-founded choice is the Natural Exponential Family (NEF). The NEF is a class of probability distributions which provides a unified framework for probability distributions through its canonical form that encompasses many commonly used distributions such as Gaussian, Poisson, and Bernoulli distributions, among others. The NEF has several desirable properties, including a closed-form expression with convex log-partition function, which simplifies the computation of the importance sampling ratio and allows for efficient parameter updates during the learning process. A parameterized family $\{g_\theta | \theta \subseteq \mathbb{R}^b\}$ is called a natural exponential family if $g_\theta(x) = \exp\left(\theta^T \Gamma(x) - K(\theta)\right)$, where $\Gamma : \mathbb{R}^b \to \mathbb{R}^b$ and $K : \mathbb{R}^b \to \mathbb{R}$ are continuous functions with $\Theta = \{\theta \in \mathbb{R}^b | \ |K(\theta)| < \infty\}$. Note that $K(\theta)$ is strictly convex in the interior of Θ and $\nabla K(\theta) = \mathbb{E}_{g_\theta}[\Gamma(X)]$. Also,$\nabla^2_\theta K(\theta) = \mathtt{Cov}_{g_\theta}[\Gamma(X)] \succ 0$. These ensure the Fisher information matrix $I(\theta) = \nabla^2_\theta K(\theta)$ is non-degenerate, guaranteeing well-posed maximum likelihood estimation. While all the NEF member distributions provide analytical tractability through their exponential structure, we employ Gaussian mixture models in our experiments for their superior approximation capabilities. The following theorem formalizes this approximation and guarantees that, with a sufficient number of components, the KL-divergence between the true steady-state distribution and its Gaussian mixture approximation can be made arbitrarily small.

Theorem 3. *Let ν_b be a discrete steady-state distribution supported on points $\{s^1, \ldots, s^n\} \subset \mathbb{R}^p$. Then, for any $\epsilon > 0$, there exists an ℓ-component Gaussian mixture model with $\ell \geq 1$ s.t. the KL divergence between ν_b and it satisfies:*

$$\mathcal{D}_{KL}\big(\nu_b \parallel \lambda_1 g_{\theta_1} + \cdots + \lambda_\ell g_{\theta_\ell}\big) \leq O(\eta) + O(\frac{\epsilon}{\eta^2}) + O\left(\frac{1}{\eta^2 \ell}\right),$$

with $g_\theta(\cdot) \geq \eta$, $\forall \theta$.

Proof. Approximate ν_b by a continuous density:

$$f(x) = \sum_{i=1}^{n} \nu_b(s_i) \mathcal{N}(x; s_i, \sigma^2 I_p), \tag{27}$$

where $\mathcal{N}(x; s_i, \sigma^2 I_p)$ is a Gaussian kernel centered at s_i. As $\sigma \to 0$, $f(x)$ converges pointwise to ν_b, ensuring the L_1-error between f and ν_b becomes arbitrarily small. Restrict $f(x)$ to a compact domain $X \subset \mathbb{R}^d$ containing all s_i and define the normalized density: $f_X(x) = \frac{f(x)}{\int_X f(y)dy}$. Since $\int_X f(y)dy \to 1$ as $\sigma \to 0$, $f_X(x)$ remains a valid approximation of ν_b on X. Now ensure $f_X(x) \geq \eta$ by defining:

$$\tilde{f}_X(x) = \max(f_X(x), \eta), \qquad \tilde{f}_X(x) \leftarrow \frac{\tilde{f}_X(x)}{\int_X \tilde{f}_X(y)dy}. \tag{28}$$

This guarantees $\tilde{f}_X(x) \geq \eta/Z$, where Z is the normalization constant. For small η, $Z \approx 1 + \eta(\mathrm{vol}(X) - 1)$, keeping the adjustment controlled. Then one can show that

$$\mathcal{D}_{\mathtt{KL}}\big(\nu_b \parallel \tilde{f}_X\big) \leq O(\eta) + O(\sigma). \tag{29}$$

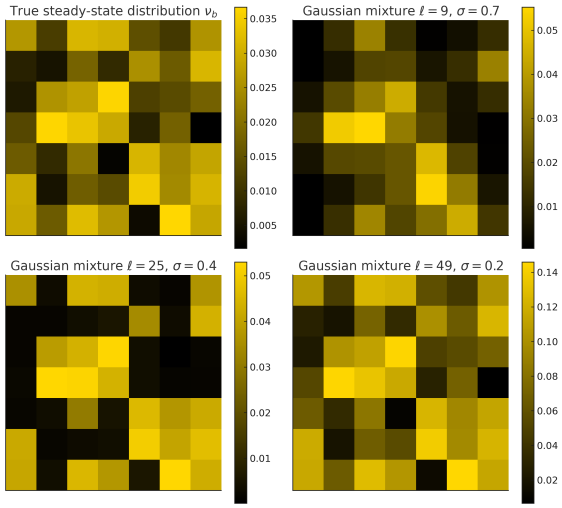

Figure 2: Illustration of the true steady-state distribution ν_b (top–left) and its Gaussian-mixture surrogates with $(\ell, \sigma) \in \{(9, 0.7), (25, 0.4), (49, 0.2)\}$. As the number of components grows and the kernels narrow, the mixture becomes visually indistinguishable from ν_b.

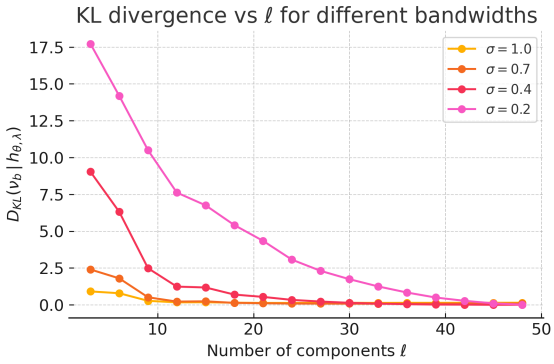

Figure 3: KL–divergence $D_{\mathrm{KL}}\big(\nu_b \parallel h_{\theta,\lambda}\big)$ versus component count ℓ for four bandwidths $\sigma \in \{1.0, 0.7, 0.4, 0.2\}$. Each curve decays monotonically, illustrating the $O\big(\ell^{-1}\sigma^{-2}\big)$ rate predicted by Theorem 3.

Now by Lemma 4.1 of Zeevi & Meir (1997), for any $\epsilon > 0$, there exists $\ell \geq 1$ such that the mixture model approximates $\tilde{f}_X$ with:

$$\mathcal{D}_{\mathrm{KL}}\big(\nu_b \parallel \lambda_1 g_{\theta_1} + \cdots + \lambda_\ell g_{\theta_\ell}\big) \leq \mathcal{D}_{\mathrm{KL}}\big(\nu_b \parallel \tilde{f}_X\big) +$$
$$\mathcal{D}_{\mathrm{KL}}\big(\tilde{f}_X \parallel \lambda_1 g_{\theta_1} + \cdots + \lambda_\ell g_{\theta_\ell}\big) \leq O(\eta) + \frac{\epsilon}{\eta^2} + O\left(\frac{1}{\eta^2 \ell}\right).$$

□

Theorem 3 guarantees that the steady-state distribution ν_b can be approximated with bounded error using a finite-component Gaussian mixtures, validating the parametric approach in our algorithm. Crucially, the bound $O(\eta) + O(\varepsilon/\eta^2) + O(1/(\eta^2\ell))$ reveals that increasing the number of components ℓ or reducing the kernel bandwidth σ (Figs. 2, 3) tightens the approximation. This theoretical foundation ensures that the KL minimization yields a reliable surrogate $h(\cdot; \theta^*, \lambda^*)$, which enables accurate calculation of the importance sampling ratio $\zeta_t = q(s_t)/h(s_t; \hat{\theta}_t, \hat{\lambda}_t)$ for distributional correction.

To derive a tractable optimization procedure for the mixture model parameters, we reformulate the KL-divergence minimization problem equation 26 as follows ($\Delta(\ell)$ is the ℓ-dimensional probability simplex):

$$\begin{aligned}
\theta^* &= \underset{\bar{\theta}\in\Theta^\ell,\lambda\in\Delta^\ell}{\arg\min}\ \mathcal{D}_{\mathtt{KL}}(\nu_b||h(\cdot;\bar{\theta},\bar{\lambda}),\ \text{where } \bar{\theta}=[\theta_1,\ldots,\theta_\ell]^\top \text{ and } \bar{\lambda}=[\lambda_1,\ldots,\lambda_\ell]^\top \\
&= \underset{\bar{\theta}\in\Theta^\ell,\lambda\in\Delta^\ell}{\arg\min} \int_{-\infty}^{\infty} \nu_b(s)\log\frac{\nu_b(s)}{h(s;\,\bar{\theta},\bar{\lambda})}ds \\
&= \underset{\bar{\theta}\in\Theta^\ell,\lambda\in\Delta^\ell}{\arg\min} \int_{-\infty}^{\infty} \nu_b(s)\log\nu_b(s)ds - \nu_b(s)\log h(s;\,\bar{\theta},\bar{\lambda})ds \\
&= \underset{\bar{\theta}\in\Theta^\ell,\lambda\in\Delta^\ell}{\arg\max} \underbrace{\int_{-\infty}^{\infty} \nu_b(s)\log h(s;\,\bar{\theta},\bar{\lambda})ds}_{F(\bar{\theta},\bar{\lambda})}
\end{aligned} \tag{30}$$

Thus minimizing the KL divergence is equivalent to maximizing the expected log-likelihood of the surrogate distribution h under the steady-state distribution ν_b. Let $F(\bar{\theta},\bar{\lambda}) = \mathbb{E}_{\nu_b}[\log h(\mathbf{s};\bar{\theta},\bar{\lambda})]$. By bounded convergence theorem, we have, $\nabla F(\bar{\theta},\lambda) = \mathbb{E}_{\nu_b}[\nabla\log h(\mathbf{s};\bar{\theta},\bar{\lambda})]$ and thus the gradient $\nabla F(\bar{\theta},\bar{\lambda})$ equals the expectation of the score function $\nabla\log h\big(s;\bar{\theta},\bar{\lambda}\big)$ under ν_b. The resulting objective is amenable to stochastic gradient ascent, enabling efficient parameter updates during learning. Hence, we employ incremental, projected, stochastic gradient ascent procedure augmented with Polyak-Ruppert averaging Polyak (1990); Ruppert (1988) to solve the optimization problem given in Equation (30), where we consider the noisy gradient $\nabla\log h(\cdot;\bar{\theta},\bar{\lambda})$ in place of the true gradient $\nabla F(\bar{\theta},\bar{\lambda})$.

$$\begin{aligned}
\begin{bmatrix}\bar{\theta}_{t+1}\\ \lambda_{t+1}\end{bmatrix} &= \Pi_{\Theta^\ell\times\Delta^\ell}\left(\begin{bmatrix}\bar{\theta}_t\\ \lambda_t\end{bmatrix} + \alpha_t\nabla\log h(\mathbf{s}_{t+1};\bar{\theta},\bar{\lambda})\right) \\
\begin{bmatrix}\widehat{\theta}_{t+1}\\ \widehat{\lambda}_{t+1}\end{bmatrix} &= \begin{bmatrix}\widehat{\theta}_t\\ \widehat{\lambda}_t\end{bmatrix} + \frac{1}{t+1}\left(\begin{bmatrix}\bar{\theta}_{t+1}\\ \bar{\lambda}_{t+1}\end{bmatrix} - \begin{bmatrix}\widehat{\theta}_t\\ \widehat{\lambda}_t\end{bmatrix}\right),
\end{aligned} \tag{31}$$

where $\Pi_{\Theta^\ell\times\Delta^\ell}$ is the projection operator which projects $\bar{\theta}_t$ on to the constraint set Θ^ℓ and λ_t on to the probability simplex Δ^ℓ. This ensures iterates $[\bar{\theta}_t,\bar{\lambda}_t]^\top$ stay feasible. In the interior of $\Theta^\ell\times\Delta^\ell$, it acts as the identity, and near the boundary, it projects orthogonally onto the boundary. Here $\alpha_t\in(0,1)$ is the step-size parameter, fixed apriori. The application of Polyak-Ruppert (PR) averaging is to enhance the robustness and stability of the iterative algorithm.

Remark 1. *By leveraging the properties of the NEF, one can obtain a closed form expression for* $\nabla\log h(\cdot;\bar{\theta},\bar{\lambda})$ *as follows:*

$$\begin{aligned}
\frac{\partial}{\partial_{\theta_j}}\nabla\log h(s;\bar{\theta},\bar{\lambda}) &= \frac{(\Gamma(x)-\nabla K(\theta_j))g_{\theta_j}(s)}{h(s;\bar{\theta},\bar{\lambda})} \\
\frac{\partial}{\partial_{\lambda_j}}\log h(s;\bar{\theta},\bar{\lambda}) &= \frac{g_{\theta_j}(s)}{h(s;\bar{\theta},\bar{\lambda})}
\end{aligned}$$

In our algorithm, we use a multi-timescale stochastic approximation framework. The stochastic gradient ascent for tracking the steady-state distribution and the TD recursion for the off-policy solution are updated on a faster timescale, while the PR averaging step is updated on a slower one. Specifically, the step-sizes for the gradient ascent and TD recursion are orders of magnitude larger than the PR-averaging step. This means that while the faster updates capture rapid changes, the slower, smaller step-size of the averaging step smooths out the fluctuations, stabilizing the learning process and reducing noise. This timescale relationship is formally defined as follows:

$$\alpha_t\in(0,1), \sum_{t\geq 0}\alpha_t = \infty, \sum_{t\geq 0}\alpha_t^2 < \infty, \alpha_t = \Omega(\frac{1}{t+1}) \tag{32}$$

Further, we modify the TD recursion to correct the steady-state bias by incorporating the steady-state distribution correction factor (ζ_t) as follows:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \beta_t \rho_t \zeta_t \underbrace{\left(\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top \mathbf{x}_t - \phi_t^\top \mathbf{x}_t\right)}_{\delta_t\text{: TD error}} \phi_t, \text{ where } \phi_t = \phi(\mathbf{s}_t),\ \rho_t = \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_b(\mathbf{a}_t|\mathbf{s}_t)} \text{ and } \zeta_t = \frac{q(\mathbf{s}_t)}{h(\mathbf{s}_t;\, \hat{\theta}_t, \hat{\lambda}_t)}. \tag{33}$$

Intuitively, ρ_t reweights the TD error δ_t by how likely the chosen action is under the target vs. behavior policy, while ζ_t reweights by how likely the state $\mathbf{s}_t$ is under the behavior's steady-state distribution *w.r.t.* the desired distribution q. By introducing the correction factor ζ_t, we re-weight updates to emphasize states in accordance with a predefined target distribution q. In the ideal case, we set $q(s) = \nu_\pi(s)$ (the target policy's true stationary distribution), but even if ν_π is unknown, we can choose $q(s)$ to be a reasonable proxy. It is a predefined, domain-specific distribution, carefully handcrafted to suit the problem context. For instance, in risk-aware or safety-constrained applications, q may emphasize certain critical regions of the state space, while in healthcare, it could overweight underrepresented patient conditions to ensure equitable learning. This correction adjusts state visitation frequencies, ensuring that states infrequently visited by the behavior policy receive appropriate weight during learning. The pseudocode of our approach is given in Algorithm 1.

Algorithm 1: Off-policy TD with linear function approximation and distributional correction

Function `Off-TD-SSBC`(π, π_b)

  for *each transition* $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{r}_{t+1}, \mathbf{s}_{t+1})$ **do**

    Calibrate parameters as follows:

$$\begin{bmatrix} \bar{\theta}_{t+1} \\ \lambda_{t+1} \end{bmatrix} = \Pi_{\Theta^\ell \times \Delta^\ell} \left(\begin{bmatrix} \bar{\theta}_t \\ \lambda_t \end{bmatrix} + \alpha_t \nabla \log h(\mathbf{s}_{t+1};\, \bar{\theta}_t, \bar{\lambda}_t) \right)$$

$$\begin{bmatrix} \widehat{\theta}_{t+1} \\ \widehat{\lambda}_{t+1} \end{bmatrix} = \begin{bmatrix} \widehat{\theta}_t \\ \widehat{\lambda}_t \end{bmatrix} + \frac{1}{t+1} \left(\begin{bmatrix} \bar{\theta}_{t+1} \\ \lambda_{t+1} \end{bmatrix} - \begin{bmatrix} \widehat{\theta}_t \\ \widehat{\lambda}_t \end{bmatrix} \right)$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \beta_t \rho_t \zeta_t \big(\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top \mathbf{x}_t - \phi_t^\top \mathbf{x}_t\big) \phi_t,$$

$$\text{where } \rho_t = \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_b(\mathbf{a}_t|\mathbf{s}_t)} \text{ and } \zeta_t = \frac{q(\mathbf{s}_t)}{h(\mathbf{s}_t;\, \widehat{\theta}_t, \widehat{\lambda}_t)}$$

Assumption 4. ***Geometric mixing (spectral gap).*** *There exist constants $M > 0$ and $\rho \in (0,1)$ such that $\left\| P^t(s,\cdot) - \nu_b \right\|_{\mathrm{TV}} \le M\, \rho^t, \forall\, s \in S,\ t \ge 0$.*

Assumption 5. ***Parameter Space Regularity.*** *The parameter space Θ is compact with smooth boundary.*

Assumption 6. ***Uniformly bounded score function.*** *There is a constant $G < \infty$ such that $\left\| \nabla \log h(s;\, \bar{\theta}, \bar{\lambda}) \right\| \le G, \forall\, \bar{\theta} \in \Theta^\ell, \forall \bar{\lambda} \in \Delta^\ell,\ \forall s \in S$.*

Remark 2. *For the mixture NEF model $h(s;\, \bar{\theta}, \bar{\lambda}) = \sum_{j=1}^{\ell} \lambda_j g_{\theta_j}(s)$, we note that the parameters $v = (\bar{\theta}, \bar{\lambda})$ lie in a compact set (because Θ is compact and $\bar{\lambda}$ lies in the simplex, which is also compact). Then, for each s, $h(s; v)$ is a continuous function of v (as a finite sum of products of continuous functions) and hence attains a minimum and maximum over the compact parameter space. Again, because $g_{\theta_j}(s) > 0$ and $\lambda_j \ge 0$ with $\sum \lambda_j = 1$, we have $h_v(s) \ge \min_j g_{\theta_j}(s) \ge \eta > 0$ and $h_v(s) \le \max_j g_{\theta_j}(s) \le M$. Moreover, the same bounds hold uniformly in s because there are finitely many states. Thus, for the mixture model, we also have:*

$$0 < \eta \le h(s; v) \le M < \infty, \qquad \forall\, v \in \Theta^\ell \times \Delta^\ell,\ \forall\, s \in S.$$

This then leads to the boundedness of the score function as previously explained.

To establish the convergence properties of Algorithm 1, we analyze the stochastic updates of the distribution parameters $\bar{\theta}_t$ and $\bar{\lambda}_t$. Let $\upsilon_t = [\bar{\theta}_t, \lambda_t]^\top$, $h(\cdot;\, \upsilon_t) = h(\cdot;\, \bar{\theta}_t, \bar{\lambda}_t)$, $U = \Theta^\ell \times \Delta^\ell$ and $\mathcal{F}_t =$

$\sigma(\theta_k, \lambda_k, \widehat{\theta}_k, \widehat{\lambda}_k, s_k, a_k, x_k, 0 \le k \le t)$ be the natural filtration generated by all variables up to time t. Then the update recursion of $[\bar{\theta}_t, \bar{\lambda}_t]^\top$ can be decomposed into a deterministic drift, a martingale noise, and a bias as follows:

$$\begin{aligned} \upsilon_{t+1} &= \Pi_U\left(\upsilon_t + \alpha_t \nabla \log h(\mathbf{s}_{t+1};\, \upsilon_t)\right) \\ &= \Pi_U\left(\upsilon_t + \alpha_t\left(\nabla F(\upsilon_t) + \mathbb{M}^{\upsilon}_{t+1} + b^{\upsilon}_t\right)\right), \\ &\text{where } \mathbb{M}^{\upsilon}_{t+1} = \nabla \log h(\mathbf{s}_{t+1};\, \upsilon_t) - \mathbb{E}\left[\nabla \log h(\mathbf{s}_{t+1};\, \upsilon_t)|\mathcal{F}_t\right], \text{and } b^{\upsilon}_t = \mathbb{E}\left[\nabla \log h(\mathbf{s}_{t+1};\, \upsilon_t)|\mathcal{F}_t\right] - \nabla F(\upsilon_t). \end{aligned} \tag{34}$$

First we establish a fundamental result on the bias term which shows that the bias term is geometrically decaying and therefore summable.

Lemma 2. *Let Assumptions 4 and 6 hold. Then* $\|b^{\upsilon}_t\| \le GM\,\rho^{\,t}, \quad \forall t \ge 0$ *and* $\sum_{t=0}^{\infty} \alpha_t \|b^{\upsilon}_t\| < \infty$.

Proof.

$$b^{\upsilon}_t \;=\; \sum_{s' \in S} \left(P_{\pi_b}(s, s') - \nu_b(s')\right) \nabla \log h_{\upsilon_t}(s').$$

Using the triangle inequality and the uniform bound G,

$$\begin{aligned} \|b^{\upsilon}_t\| \;&\le\; \sum_{s'} |P_{\pi_b}(s, s') - \nu_b(s')| \left\|\nabla \log h_{\upsilon_t}(s')\right\| \\ &\le\; G \sum_{s'} |P_{\pi_b}(s, s') - \nu_b(s')| \\ &= G\, \|P_{\pi_b}(s, \cdot) - \nu_b\|_1 \\ &= 2G\, \|P_{\pi_b}(s, \cdot) - \nu_b\|_{TV} \\ &\le 2GM\rho^t. \end{aligned}$$

Further, the weighted series $\sum_t \alpha_t \|b^{\upsilon}_t\| \le \sum_t \alpha_t \rho^{\,t} < \infty$, since the step–size schedule $\alpha_t \to 0$. □

The following theorem establishes that the sequence $\{[\bar{\theta}_t, \bar{\lambda}_t]^\top\}$ converges to Karush-Kuhn-Tucker (KKT) points—first-order optimality conditions where the gradient aligns with the normal cone of $\Theta^\ell \times \Delta^\ell$. This guarantees the learned mixture distribution $h(\cdot;\, \bar{\theta}, \bar{\lambda})$ converges to a stationary point of the KL-divergence minimization problem.

Theorem 4 (Convergence of Distribution Approximation)**.** *Let the step-size* $\{\alpha_t\}$ *satisfy Equation (32). Let Assumptions 4-6 hold. Then the sequence* $\{[\bar{\theta}, \bar{\lambda}]^\top\}$ *converges almost surely to the set of KKT points:*

$$\{\upsilon = [\bar{\theta}, \bar{\lambda}]^\top \in \Theta^\ell \times \Delta^\ell : -\nabla F(\upsilon) \in N_U(\upsilon)\},$$

where $N_U(\upsilon)$ *denotes the normal cone to* $Theta^\ell \times \Delta^\ell$ *at* υ*, defined as:*

$$N_U(\upsilon) = \left\{d \in \mathbb{R}^{dim(\Theta^\ell)+\ell} : \langle d, u - \upsilon\rangle \le 0,\ \forall u \in U\right\}.$$

Proof. Let $g_t = \nabla F(\upsilon_t) + \mathbb{M}^{\upsilon}_{t+1} + b^{\upsilon}_t$. Then,

$$\begin{aligned} \upsilon_{t+1} &= \Pi_U\big(\upsilon_t + \alpha_t g_t\big), \\ &= \upsilon_t + \alpha_t \Gamma_U(g_t) + \Pi_U\big(\upsilon_t + \alpha_t g_t\big) - \upsilon_t - \alpha_t \Gamma_U(g_t) \\ &= \upsilon_t + \alpha_t \left(\Gamma_U(g_t) + \frac{\Pi_U\big(\upsilon_t + \alpha_t g_t\big) - \upsilon_t}{\alpha_t} - \Gamma_U(g_t)\right) \\ &= \upsilon_t + \alpha_t \left(\Gamma_U(g_t) + o(\alpha_t)\right). \end{aligned} \tag{35}$$

The last equality follows since

$$\lim_{\varepsilon \to 0} \frac{\Pi_U\big(\upsilon_t + \varepsilon g_t\big) - \upsilon_t}{\varepsilon} = \underbrace{\Pi_{T_U(\upsilon_t)}(g_t)}_{\Gamma_U(g_t)}, \tag{36}$$

where

$$T_U(\upsilon) = \overline{\big\{u \in \mathbb{R}^{\dim(\Theta^\ell)+\ell} : \upsilon + \tau u \in U \text{ for some } \tau > 0\big\}}.$$

In the interior of U, $T_U(\upsilon_t) = \mathbb{R}^d\big(\text{unconstrained}\big)$. and near the boundary of U, $T_U(\upsilon_t) = \big\{u \in \mathbb{R}^{\dim(\Theta^\ell)+\ell} \;\big|\; u \text{ points into } U\big\}$. Thus, $\Gamma_U(g_t)$ is the directional derivative of the projection operator Π_U at point υ_t in the direction g_t, which is equivalent to the projection of g_t onto the tangent cone of U at υ_t Rockafellar (2015). Intuitively, it captures the "feasible component" of g_t that aligns with the constraints of U.

For the noise $\mathbb{M}^\upsilon_{t+1}$, $\mathbb{E}\big[\mathbb{M}^\upsilon_{t+1} \mid \mathcal{F}_t\big] = 0$ (by definition). Further, using the triangle inequality,

$$\begin{aligned} \|\mathbb{M}^\upsilon_{t+1}\| \;&=\; \big\|\nabla \log h(s_{t+1};\, \upsilon_t) - \mathbb{E}[\nabla \log h(s_{t+1};\, \upsilon_t) \mid \mathcal{F}_t]\big\| \\ &\leq\; G + G = 2G \quad a.s. \end{aligned}$$

Squaring and taking conditional expectation yields $\mathbb{E}[\|\mathbb{M}^\upsilon_{t+1}\|^2 \mid \mathcal{F}_t] \leq (2G)^2 = 4G^2$. Thus $\{\mathbb{M}^\upsilon_t\}$ is a square-integrable martingale-difference sequence. Now, consider $S_t = \sum_{k=0}^{t-1} \alpha_k \mathbb{M}^\upsilon_{k+1}$. Note that

$$\sum_{t\geq 0} \mathbb{E}\left[\|S_{t+1} - S_t\|^2 | \mathcal{F}_t\right] = \sum_{t\geq 0} \alpha_t^2 \mathbb{E}\left[\|\mathbb{M}^\upsilon_{t+1}\|^2 | \mathcal{F}_t\right] < 4G^2 \sum_{t\geq 0} \alpha_t^2 < \infty. \tag{37}$$

By martingale convergence theorem, it follows that S_t converges, *i.e.*, $\sum_{t=0}^{\infty} \alpha_t \mathbb{M}^\upsilon_{t+1} < \infty \quad a.s.$

Now rearranging (35), we get

$$\upsilon_{t+1} = \upsilon_t + \alpha_t\big(\Gamma_U(\nabla F(\upsilon_t)) + \underbrace{\Gamma_U(g_t) - \Gamma_U(\nabla F(\upsilon_t))}_{\xi_t} + o(\alpha_t)\big) \tag{38}$$

Using the non-expansive property of Γ_U, we have

$$\begin{aligned} \|\xi_t\| &= \|\Gamma_U(g_t) - \Gamma_U(\nabla F(\upsilon_t))\| \\ &\leq \|g_t - \nabla F(\upsilon_t)\| = \|\mathbb{M}^\upsilon_{t+1} + b^\upsilon_t\| \\ &\leq \|\mathbb{M}^\upsilon_{t+1}\| + \|b^\upsilon_t\|. \end{aligned} \tag{39}$$

Hence,

$$\sum_t \alpha_t \|\xi_t\| \leq \sum_t \alpha_t \|\mathbb{M}^\upsilon_{t+1}\| + \sum_t \alpha_t \|b^\upsilon_t\| < \infty \quad a.s. \tag{40}$$

Therefore by Borkar (2008), it follows that $\{\upsilon_t\}$ asymptoticaly tracks the ODE

$$\dot{\upsilon} = \Gamma_U(\nabla F(\upsilon)). \tag{41}$$

However, because F is smooth and the constraint set U is convex, the above differential equation is well-defined and corresponds to the projected gradient ascent. By the theory of stochastic approximation (see Borkar (2008)), the sequence $\{\upsilon_t\}$ converges to a (possibly sample path dependent) internally chain transitive invariant set of the above ODE. Since F is C^1,

$$\frac{d}{dt} F\big(\upsilon(t)\big) \;=\; \big\langle \nabla F\big(\upsilon(t)\big),\, \dot{\upsilon}(t)\big\rangle \;=\; \big\langle \nabla F(\upsilon),\, \Gamma_U\big[\nabla F(\upsilon)\big]\big\rangle.$$

Apply Moreau's decomposition to obtain $\nabla F(\upsilon) = \Pi_{T_U(\upsilon)}[\nabla F(\upsilon)] + \Pi_{N_U(\upsilon)}[\nabla F(\upsilon)]$ and $\langle \Pi_{T_U(\upsilon)}[\nabla F(\upsilon)], \Pi_{N_U(\upsilon)}[\nabla F(\upsilon)] \rangle = 0$. Then,

$$\begin{aligned}\big\langle \nabla F(\upsilon), \Gamma_U(\nabla F(\upsilon)) \big\rangle &= \big\langle \Pi_{T_U(\upsilon)}[\nabla F(\upsilon)] + \Pi_{N_U(\upsilon)}[\nabla F(\upsilon)], \Gamma_U(\nabla F(\upsilon)) \big\rangle \\ &= \langle \Gamma_U(\nabla F(\upsilon)), \Gamma_U(\nabla F(\upsilon)) \rangle \\ &= \big\| \Gamma_U(\nabla F(\upsilon)) \big\|^2 \ \geq \ 0. \end{aligned} \tag{42}$$

Hence

$$\frac{d}{dt} F\big(\upsilon(t)\big) = \big\| \Gamma_U\big[\nabla F\big(\upsilon(t)\big)\big] \big\|^2 \ \geq \ 0 \tag{43}$$

with equality *iff* $\Gamma_U[\nabla F(\upsilon(t))] = 0$. Therefore, the invariant set of the above ODE is the stationary (equilibrium) set:

$$\big\{ \upsilon \in U : \Gamma_U\big[\nabla F(\upsilon)\big] = 0 \big\} \ = \ \big\{ \upsilon \in U : -\nabla F(\upsilon) \in N_U(\upsilon) \big\},$$

which are the Karush-Kuhn-Tucker (KKT) points. The last equality follows again by Moreau's decomposition of $\nabla F(\upsilon)$. □

Having established the almost sure convergence of the distribution parameters $\{\upsilon_t\}$ to υ^* in Theorem 4, we now analyze the temporal difference learning dynamics given by equation 33. Prior to this, observe that the unilateral timescale separation between the faster distribution estimation updates ($\upsilon_t = [\bar{\theta}_t, \bar{\lambda}_t]^\top$) and slower Polyak-Ruppert averaging ($\hat{\upsilon}_t = [\hat{\theta}_t, \hat{\lambda}_t]^\top$) ensures that $\hat{\upsilon}_t \to \upsilon^*$ asymptotically. This justifies replacing the time-varying $\zeta_t = q(\mathbf{s}_t)/h(\mathbf{s}_t; \hat{\upsilon}_t)$ in the TD update equation 33 with its steady-state counterpart $\zeta_t^* = q(\mathbf{s}_t)/h(\mathbf{s}_t; \upsilon^*)$. The substitution decouples the distribution approximation error from the value estimation error, permitting the simplified TD recursion (See Chapter 6 of Borkar (2008)). Hence, we rewrite $\mathbf{x}_t$ update as follows (we let $g_t = \rho_t \zeta_t^* \big(\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top x - \phi_t^\top x\big)\phi_t$ and $h^*(\cdot) = h(\cdot; \upsilon^*)$):

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t + \beta_t \rho_t \zeta_t^* \big(\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top \mathbf{x}_t - \phi_t^\top \mathbf{x}_t\big)\phi_t, \text{ where } \rho_t = \frac{\pi(\mathbf{a}_t|\mathbf{s}_t)}{\pi_b(\mathbf{a}_t|\mathbf{s}_t)} \text{ and } \zeta_t^* = \frac{q(\mathbf{s}_t)}{h^*(\mathbf{s}_t)} \\ &= \mathbf{x}_t + \beta_t \left(b_t^x + G(\mathbf{x}_t) + \mathbb{M}_{t+1}^x\right), \text{where } G(x) = \mathbb{E}[g_t] = \mathbb{E}\Big[\, \rho_t \, \zeta_t^* \big(\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top x - \phi_t^\top x\big)\phi_t \Big] \qquad (44) \\ &= \ \mathbb{E}_{\mathbf{s}}\Big[\frac{q(\mathbf{s})}{h^\star(\mathbf{s})} \mathbb{E}_a\big[\rho_t \big(\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top x - \phi_t^\top x\big)\phi_t \ \big| \ \mathbf{s}\big] \Big] \\ &= \mathbb{E}_{\mathbf{s}}\Big[\frac{q(\mathbf{s})}{h^\star(\mathbf{s})} \big(R_\pi(\mathbf{s}) + \gamma \, (P_\pi \Phi x)(\mathbf{s}) - (\Phi x)(\mathbf{s})\big)\, \phi(\mathbf{s}) \Big] \\ &= \Phi^\top \Xi_{\nu_b} \, \Xi_{h^\star}^{-1} \, \Xi_q \left(R_\pi + (\gamma \, P_\pi \ - \ I)\Phi x\right) \\ &= \underbrace{\Phi^\top \Xi_{\nu_b} \, \Xi_{h^\star}^{-1} \, \Xi_q \, (\gamma \, P_\pi \ - \ I)\Phi}_{\Lambda_c} \, x + \underbrace{\Phi^\top \Xi_{\nu_b} \, \Xi_{h^\star}^{-1} \, \Xi_q R_\pi}_{\xi}. \qquad (45) \end{aligned}$$

Also,

$$\mathbb{M}_{t+1}^x = g_t - \mathbb{E}\big[g_t \mid \mathcal{F}_t\big], \text{ and } b_t^x = \mathbb{E}\big[g_t \mid \mathcal{F}_t\big] - h\big(x_t\big). \tag{46}$$

Further, note that since we have finite state and action spaces $|\mathbf{r}_t| \leq R_\infty$, $\|\phi(s)\| \leq \Phi_\infty$, and $0 \leq \rho_t \leq \rho_\infty$, $0 < \zeta_t \leq \zeta_\infty$.

We first bound the TD error:

$$\begin{aligned}|\delta_t| &= \big|\mathbf{r}_{t+1} + \gamma \phi_{t+1}^\top \mathbf{x}_t - \phi_t^\top \mathbf{x}_t\big| \\ &\leq |\mathbf{r}_{t+1}| + \gamma |\phi_{t+1}^\top \mathbf{x}_t| + |\phi_t^\top \mathbf{x}_t| \\ &\leq R_\infty + \gamma \Phi_\infty \|x_t\| + \Phi_\infty \|\mathbf{x}_t\| = R_\infty + (1+\gamma)\Phi_\infty \|\mathbf{x}_t\| \end{aligned}$$

Then the update term g_t satisfies:

$$\|g_t\| = |\rho_t \zeta_t \delta_t| \cdot \|\phi_t\| \leq \rho_\infty \zeta_\infty \left(R_\infty + (1+\gamma)\Phi_\infty \|x_t\|\right) \cdot \Phi_\infty \leq C_1 + C_2 \|x_t\|,$$

where $C_1 = \rho_\infty \zeta_\infty R_\infty \Phi_\infty$, and $C_2 = \rho_\infty \zeta_\infty (1+\gamma)\Phi_\infty^2$. Now,

$$\begin{aligned}\|\mathbb{M}_t^x\| &= \|g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]\| \\ &\leq \|g_t\| + \|\mathbb{E}[g_t \mid \mathcal{F}_t]\| \leq 2 \sup \|g_t\| \\ &\leq 2(C_1 + C_2\|x_t\|) \leq \widetilde{C}_1(1 + \|x_t\|).\end{aligned}$$

Further, using $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbb{M}_{t+1}^x\|^2 \mid \mathcal{F}_t\big] &= \mathbb{E}\Big[\big\|g_t - \mathbb{E}[g_t \mid \mathcal{F}_t]\big\|^2 \mid \mathcal{F}_t\Big] \\ &\leq 2\,\mathbb{E}\big[\|g_t\|^2 \mid \mathcal{F}_t\big] + 2\,\big\|\mathbb{E}[g_t \mid \mathcal{F}_t]\big\|^2 \\ &\leq 4\,\mathbb{E}\big[\|g_t\|^2 \mid \mathcal{F}_t\big] \leq 4\big(C_1 + C_2\|x_t\|\big)^2 \\ &\leq 8C_1^2 + 8C_2^2\|x_t\|^2. \end{aligned} \tag{47}$$

Now we write $b_t^x = \mathbb{E}[g_t \mid \mathcal{F}_t] - G(\mathbf{x}_t) = \mathbb{E}[g_t | \mathbf{s}_t, \mathbf{x}_t] - \mathbb{E}_{\nu_b}[g_t]$.

$$\begin{aligned}\text{Thus, } \|b_t^x\| &= \|\mathbb{E}[g_t|\mathbf{s}_t, \mathbf{x}_t] - \mathbb{E}_{\nu_b}[g_t]\| \\ &\leq \sum_{s \in S} |P_{\pi_b}(\mathbf{s}_t, s) - \nu_b(s)| \cdot \|\mathbb{E}[g_t \mid \mathbf{s}_t, \mathbf{x}_t]\| \\ &\leq \sup_s \|\mathbb{E}[g_t \mid s, \mathbf{x}_t]\| \cdot \|P_{\pi_b}(\mathbf{s}_t, \cdot) - \nu_b\|_1 \\ &= 2 \sup_s \|\mathbb{E}[g_t \mid s, \mathbf{x}_t]\| \cdot \|P_{\pi_b}(\mathbf{s}_t, \cdot) - \nu_b\|_{TV} \\ &\leq 2M\rho^t(C_1 + C_2\|\mathbf{x}_t\|) \leq \widetilde{C}_2 \rho^t (1 + \|\mathbf{x}_t\|)\end{aligned} \tag{48}$$

To establish convergence of the sequence $\{\mathbf{x}_t\}$, we must first ensure the iterates remain stochastically bounded. While classical stochastic approximation theory Borkar (2008) often assumes almost sure boundedness, we prove the following weaker but sufficient condition for our setting.

Lemma 3. *The iterates* $\mathbf{x}_t$ *satisfies* $\sup_t \mathbb{E}\left[\|\mathbf{x}_t\|^2\right] < \infty$.

Proof. Note G satisfes the drift inequality

$$\mathbf{x}^\top G(\mathbf{x}) \;\leq\; -c\|\mathbf{x}\|^2 + d, \qquad \forall \mathbf{x} \in \mathbb{R}^k, \tag{49}$$

where $c = \frac{1}{2}\lambda_{\min}(-\Lambda_c)$ and $d = \|\xi\|^2/(2\lambda_{\min}(-\Lambda_c)))$.

Using $\mathbb{E}[\mathbb{M}_{t+1}^x \mid \mathcal{F}_t] = 0$ and expanding the square,

$$\mathbb{E}[V_{t+1} \mid \mathcal{F}_t] = \|\mathbf{x}_t\|^2 + 2\alpha_t \mathbf{x}_t^\top\big(G(\mathbf{x}_t) + b_t^x\big) + \alpha_t^2\Big(\|G(\mathbf{x}_t) + b_t^x\|^2 + \mathbb{E}[\|\mathbb{M}_{t+1}^x\|^2 \mid \mathcal{F}_t]\Big).$$

Apply equation 48, equation 47, equation 49, and Young's inequality $2\mathbf{x}_t^\top b_t^x \leq c\|\mathbf{x}_t\|^2 + c^{-1}\|b_t^x\|^2$, and the bound $\|G(\mathbf{x}_t)\|^2 \leq 2\|\Lambda_c\|^2\|\mathbf{x}_t\|^2 + 2\|\xi\|^2$, to obtain

$$\mathbb{E}[V_{t+1}] \leq \Big(1 - (2c - c)\alpha_t + L_2\,\alpha_t^2\Big)\,\mathbb{E}[V_t] + 2d\,\alpha_t + L_0\,\alpha_t^2 + \Big(c^{-1}\alpha_t + 2\alpha_t^2\Big)\,8M^2\rho^{2t}\Big(C_2^2\,\mathbb{E}[V_t] + C_1^2\Big), \tag{50}$$

where $L_2 = 4\|\Lambda_c\|^2 + 8C_2^2, \quad L_0 = 4\|\xi\|^2 + 8C_1^2$.

Equivalently, by collecting the coefficients of $\mathbb{E}[V_t]$ and the constants, we may write

$$\mathbb{E}[V_{t+1}] \leq \Big(1 - c\beta_t + \underbrace{L_2\beta_t^2 + 8M^2C_2^2\,(c^{-1}\beta_t + 2\beta_t^2)\rho^{2t}}_{=G_t}\Big)\mathbb{E}[V_t] + 2d\,\beta_t + L_0\,\beta_t^2 + \underbrace{8M^2C_1^2\,(c^{-1}\beta_t + 2\beta_t^2)\rho^{2t}}_{=e_t}. \tag{51}$$

Since $\rho^{2t} \to 0$ geometrically and $\beta_t \to 0$, the perturbation terms G_t, e_t vanish; for all large t one can ensure $G_t \le \frac{c}{2}\beta_t$, leading to the canonical stochastic approximation form

$$\mathbb{E}[V_{t+1}] \;\le\; \left(1 - \tfrac{c}{2}\beta_t\right)\mathbb{E}[V_t] + 2d\,\beta_t + c'\beta_t^2. \tag{52}$$

Now using induction, we will show that $\sup_t \mathbb{E}[V_t] < \infty$. For $t = 0$, $\mathbb{E}[V_0] = \mathbb{E}[\|\mathbf{x}_0\|^2]$ is finite since $\mathbf{x}_0$ is initialized with finite variance (base case). Now assume $\mathbb{E}[V_t] \le K$ for some constant K and all $t \le T$, where

$$K = \max\left(\mathbb{E}[\|\mathbf{x}_0\|^2], \frac{4d}{c} + \frac{2}{c}\sup_{t\ge 0} c'\beta_t\right).$$

Then,

$$\begin{aligned}\mathbb{E}[V_{T+1}] &\le \left(1 - \frac{c}{2}\beta_T\right)K + 2d\beta_T + c'\beta_T^2 \\ &\le K + \beta_T\left(-\frac{c}{2}K + 2d\right) + c'\beta_T^2. \end{aligned}\tag{53}$$

Since $K \ge \dfrac{4d}{c} + \dfrac{2}{c}\sup\limits_{t\ge 0} c'\beta_t$, we have:

$$\begin{aligned}-\frac{c}{2}K + 2d &\le -\frac{c}{2}\left(\frac{4d}{c} + \frac{2}{c}\sup_{t\ge 0} c'\beta_t\right) + 2d \\ &= -c'\sup_{t\ge 0}\beta_t \le -c'\beta_T < 0.\end{aligned}$$

Thus from equation 53:

$$\mathbb{E}[V_{T+1}] \le K - c'\alpha_T^2 + c'\alpha_T^2 \le K.$$

By induction, $\mathbb{E}[V_t] \le K$ for all but a finite number of t. □

To establish the convergence of $\mathbf{x}_t$, we must show that the bias and noise terms are manageable. Specifically, the next lemma establishes that the series formed by the weighted bias and martingale noise terms converge almost surely.

Lemma 4. *For the martingale noise $\mathbb{M}_t^x$ and the bias b_t^x, we have*

$$\mathbb{P}\left(\sum_t \beta_t \mathbb{M}_{t+1}^x < \infty, \quad \sum_t \beta_t b_t^x < \infty\right) = 1.$$

Proof. From equation 47, we have

$$\begin{aligned}&\mathbb{E}[\|\mathbb{M}_{t+1}^x\|^2 \mid \mathcal{F}_t] \le 8C_1^2 + 8C_2^2\|\mathbf{x}_t\|^2 \\ &\Rightarrow \mathbb{E}[\|\mathbb{M}_{t+1}^x\|^2] \le 8C_1^2 + 8C_2^2\mathbb{E}[\|\mathbf{x}_t\|^2].\end{aligned}\tag{54}$$

Hence, $\mathbb{M}_{t+1}^x$ is square-integrable. Now, by the convergence theorem for square-integrable martingale (for vector-valued martingales), it is enough to show that

$$\sum_t \mathbb{E}[\|\beta_t \mathbb{M}_{k+1}^x\|^2 \mid \mathcal{F}_t] < \infty \quad a.s.$$

Thus it is enough to show that

$$\mathbb{E}\left[\sum_t \mathbb{E}[\|\beta_t \mathbb{M}_{t+1}^x\|^2 \mid \mathcal{F}_t]\right] < \infty.$$

Therefore, by monotone convergence theorem, we get

$$\begin{aligned}\mathbb{E}\left[\sum_t \mathbb{E}[\|\beta_t \mathbb{M}_{t+1}^x\|^2 \mid \mathcal{F}_t]\right] &= \sum_t \mathbb{E}\left[\mathbb{E}[\|\beta_t \mathbb{M}_{t+1}^x\|^2 \mid \mathcal{F}_t]\right] \\ &\leq \sum_t \beta_t^2 \left(8C_1^2 + 8C_2^2 \mathbb{E}[\|\mathbf{x}_t\|^2]\right) \\ &\leq \sum_t \beta_t^2 \left(8C_1^2 + 8C_2^2 \sup_t \mathbb{E}[\|\mathbf{x}_t\|^2]\right) < \infty.\end{aligned}$$

The last inequality follows from Lemma 3 and $\sum_t \beta_t^2 < \infty$. This implies that $\mathbb{P}\left(\sum_t \beta_t \mathbb{M}_{t+1}^x < \infty\right) = 1$. Now for b_t^x, it follows from equation 48,

$$\begin{aligned}\mathbb{E}\left[\sum_t \beta_t \|b_t^x\|\right] &= \sum_t \beta_t \mathbb{E}[b_t^x] \\ &\leq \sum_t 2M\beta_t \rho^t (C_1 + C_2 \mathbb{E}[\|\mathbf{x}_t\|]) \\ &\leq \sum_t 2M\beta_t \rho^t (C_1 + C_2 \sqrt{\mathbb{E}[\|\mathbf{x}_t\|^2]}) \\ &\leq \sum_t 2M\beta_t \rho^t (C_1 + C_2 \sup_t \sqrt{\mathbb{E}[\|\mathbf{x}_t\|^2]}) < \infty.\end{aligned}$$

The last inequality follows again from Lemma 3, $\beta_t \to 0$ and $\rho \in (0,1)$. Hence, $\mathbb{P}\left(\sum_t \beta_t b_{t+1}^x < \infty\right) = 1$. □

Having established the stochastic boundedness of the iterates $\mathbf{x}_t$ and the almost sure summability of the martingale noise $\sum_t \alpha_t \mathbb{M}_{t+1}^{\mathbf{x}}$ and bias terms $\sum_t \alpha_t b_t^{\mathbf{x}}$, we now prove almost sure convergence of the sequence $\{\mathbf{x}_t\}$.

Theorem 5 (Convergence of the TD Iterates)**.** *Assume that the matrix* $\Lambda_c = \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P_\pi - I)\Phi$ *is* ***Hurwitz*** *(all eigenvalues have strictly negative real parts) and* ***diagonalizable****. Then the sequence* $\{\mathbf{x}_t\}$ *converges almost surely to the unique solution* $\mathbf{x}^* = \mathbf{x}_{\text{TD}}^c$ *satisfying:*

$$\Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (I - \gamma P_\pi)\Phi \mathbf{x}^* = \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q \bar{R}_\pi$$

Proof. Rearranging equation 44 of $\mathbf{x}_t$ as follows:

$$\mathbf{x}_{t+1} - \mathbf{x}^* = (\mathbf{x}_t - \mathbf{x}^*) + \beta_t \Lambda_c (\mathbf{x}_t - \mathbf{x}^*) + \alpha_t (\Lambda_c \mathbf{x}^* + \xi) + \beta_t (b_t^x + \mathbb{M}_{t+1}^x).$$

But note that $\Lambda_c \mathbf{x}^* + \xi = 0$ by definition of $\mathbf{x}^*$. So:

$$\mathbf{x}_{t+1} - \mathbf{x}^* = (\mathbf{x}_t - \mathbf{x}^*) + \beta_t \Lambda_c (\mathbf{x}_t - \mathbf{x}^*) + \beta_t (b_t^x + \mathbb{M}_{t+1}^x).$$

Let $e_t = \mathbf{x}_t - \mathbf{x}^*$. Then:

$$e_{t+1} = (I + \beta_t \Lambda_c) e_t + \beta_t \eta_t, \text{ where } \eta_t = b_t^x + \mathbb{M}_{t+1}^x.$$

We know that $\sum_t \beta_t \|\eta_t\| < \infty$ *a.s.* by Lemma 4 (since both b_t^x and $\mathbb{M}_{t+1}^x$ are summable in absolute value *a.s.*). Now, because Λ_c is negative definite, the matrix $I + \beta_t \Lambda_c$ has eigenvalues in $(0,1)$ for small β_t. After unraveling the above recursion, we obtain

$$e_{t+1} = \left(\prod_{k=0}^{t} (I + \beta_k \Lambda_c)\right) y_0 + \sum_{k=0}^{t} \beta_k \eta_k \left(\prod_{j=k+1}^{t} (I + \beta_j \Lambda_c)\right).$$

Let

$$Q(t,k) = \prod_{j=k}^{t-1}(I + \beta_j \Lambda_c) \quad \text{for } t > k, \quad Q(k,k) = I. \tag{55}$$

Then,

$$e_{t+1} = Q(t+1,0)e_0 + \sum_{k=0}^{t} \beta_k Q(t+1,k+1)\eta_k. \tag{56}$$

Now, since Λ_c is negative definite, let $\lambda_{\min} > 0$ be such that the real parts of the eigenvalues of Λ_c are less than or equal to $-\lambda_{\min}$. Then, there exists a constant $C > 0$ and $\beta > 0$ such that:

$$\|Q(t,k)\| \leq C \exp\left(-\beta \sum_{j=k}^{t-1} \alpha_j\right). \tag{57}$$

Since Λ_c is diagonalizable, let $\Lambda_c = PDP^{-1}$ where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ is diagonal. Then:

$$\begin{aligned} Q(t,k) = \prod_{j=k}^{t-1}(I + \beta_j \Lambda_c) &= P\left(\prod_{j=k}^{t-1}(I + \beta_j D)\right) P^{-1} \\ &= P\left(\prod_{j=k}^{t-1} D_j\right) P^{-1} \end{aligned}$$

where $D_j = I + \beta_j D = \operatorname{diag}(1 + \alpha_j\lambda_1, \ldots, 1 + \beta_j\lambda_d)$. The norm satisfies:

$$\|Q(t,k)\| \leq \|P\| \cdot \|P^{-1}\| \cdot \left\|\prod_{j=k}^{t-1} D_j\right\| \tag{58}$$

The diagonal matrix norm is given by:

$$\left\|\prod_{j=k}^{t-1} D_j\right\| = \max_{1 \leq i \leq d} \left|\prod_{j=k}^{t-1}(1 + \beta_j\lambda_i)\right|$$

Now we establish a uniform bound for each eigenvalue product $\prod_{j=k}^{t-1}(1 + \beta_j\lambda_i)$. For any $\epsilon > 0$, there exists $\beta_0 > 0$ such that for $0 \leq \beta_j \leq \beta_0$:

$$|1 + \beta_j\lambda_i| \leq e^{\beta_j \operatorname{Re}(\lambda_i) + \epsilon\beta_j} \tag{59}$$

This follows from the logarithm expansion:

$$\begin{aligned} \log(1 + \beta_j\lambda_i) &= \beta_j\lambda_i - \frac{(\beta_j\lambda_i)^2}{2} + \cdots \\ &= \beta_j \operatorname{Re}(\lambda_i) + i\beta_j \operatorname{Im}(\lambda_i) + O(\beta_j^2) \end{aligned}$$

so the real part is $\alpha_j \operatorname{Re}(\lambda_i) + O(\alpha_j^2)$. For sufficiently small α_j, we have:

$$\operatorname{Re}\left(\log(1 + \beta_j\lambda_i)\right) \leq \beta_j \operatorname{Re}(\lambda_i) + \epsilon\beta_j$$

Thus $|1 + \beta_j\lambda_i| = e^{\operatorname{Re}(\log(1+\beta_j\lambda_i))} \leq e^{\beta_j \operatorname{Re}(\lambda_i) + \epsilon\beta_j}$.

Set $\epsilon = \lambda_{\min}/2 > 0$ where $\lambda_{\min} = \min_i |\operatorname{Re}(\lambda_i)|$. Since $\operatorname{Re}(\lambda_i) \leq -\lambda_{\min}$:

$$|1 + \beta_j\lambda_i| \leq e^{\beta_j \operatorname{Re}(\lambda_i) + \beta_j\lambda_{\min}/2} \leq e^{-\beta_j\lambda_{\min} + \beta_j\lambda_{\min}/2} = e^{-\beta_j\lambda_{\min}/2}$$

when $\beta_j \leq \beta_0$. Since $\beta_j \to 0$, there exists $K_0 \in \mathbb{N}$ such that $\beta_j \leq \beta_0$ for all $j \geq K_0$.
Case 1: $k \geq K_0$
For all $j \geq k \geq K_0$, we have $\beta_j \leq \beta_0$, so:

$$\left|\prod_{j=k}^{t-1}(1+\beta_j\lambda_i)\right| \leq \exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)$$

Case 2: $k < K_0$
Split the product at K_0:

$$\prod_{j=k}^{t-1}(1+\beta_j\lambda_i) = \underbrace{\left(\prod_{j=k}^{K_0-1}(1+\beta_j\lambda_i)\right)}_{(*)} \cdot \underbrace{\left(\prod_{j=K_0}^{t-1}(1+\beta_j\lambda_i)\right)}_{(**)}$$

Term $(*)$ is a finite product (since K_0 is fixed). Using $|1+\beta_j\lambda_i| \leq 1+|\lambda_i|\beta_j$:

$$|(*)| \leq \prod_{j=k}^{K_0-1}(1+|\lambda_i|\beta_j) \leq \exp\left(|\lambda_i|\sum_{j=k}^{K_0-1}\beta_j\right) \leq C_i(k)$$

where $C_i(k) = \exp\left(|\lambda_i|\sum_{j=0}^{K_0-1}\beta_j\right)$ is bounded (as $\beta_j > 0$ and fixed K_0). Term $(**)$ is bounded by Case 1:

$$\begin{aligned}|(**)| &\leq \exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=K_0}^{t-1}\beta_j\right)\\ &\leq \exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)\cdot\exp\left(\frac{\lambda_{\min}}{2}\sum_{j=k}^{K_0-1}\beta_j\right)\end{aligned}$$

Combining both terms:

$$\begin{aligned}\left|\prod_{j=k}^{t-1}(1+\beta_j\lambda_i)\right| &\leq C_i(k)\exp\left(\frac{\lambda_{\min}}{2}\sum_{j=k}^{K_0-1}\beta_j\right)\exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)\\ &= C_i''(k)\exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)\end{aligned}$$

where $C_i''(k) = C_i(k)\exp\left(\frac{\lambda_{\min}}{2}\sum_{j=k}^{K_0-1}\beta_j\right)$.

Since $k < K_0$ and there are only finitely many such k, define:

$$C' = \max\left\{\max_{\substack{1\leq i\leq d\\ 0\leq k<K_0}} C_i''(k), 1\right\} < \infty$$

For $k \geq K_0$, we have $C_i''(k) = 1$. Thus for all i, k, and $t > k$:

$$\left|\prod_{j=k}^{t-1}(1+\beta_j\lambda_i)\right| \leq C'\exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)$$

Therefore:

$$\left\|\prod_{j=k}^{t-1}D_j\right\| = \max_i\left|\prod_{j=k}^{t-1}(1+\beta_j\lambda_i)\right| \leq C'\exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)$$

Substituting into equation 58:

$$\|Q(t,k)\| \leq \|P\| \cdot \|P^{-1}\| \cdot C' \exp\left(-\frac{\lambda_{\min}}{2}\sum_{j=k}^{t-1}\beta_j\right)$$

Set $C = \|P\| \cdot \|P^{-1}\| \cdot C'$ and $\bar{\beta} = \lambda_{\min}/2$ to obtain:

$$\|Q(t,k)\| \leq C \exp\left(-\beta\sum_{j=k}^{t-1}\beta_j\right)$$

for all $t > k \geq 0$, with $C, \beta > 0$ independent of t and k.
Therefore from equation 56,

$$\|e_{t+1}\| \leq C \exp\left(-\beta\sum_{j=0}^{t}\beta_j\right)\|e_0\| + \sum_{k=0}^{t}\beta_k\|Q(t+1,k+1)\|\|\eta_k\|. \tag{60}$$

The first term goes to zero as $t \to \infty$ because $\sum_{j=0}^{t}\beta_j \to \infty$. For the second term, note:

$$\sum_{k=0}^{t}\beta_k\|Q(t+1,k+1)\|\|\eta_k\| \leq C\sum_{k=0}^{t}\beta_k \exp\left(-\bar{\beta}\sum_{j=k+1}^{t}\beta_j\right)\|\eta_k\|. \tag{61}$$

By the summability of $\beta_k\|\eta_k\|$ and the exponential decay, this term goes to zero. Indeed, for any fixed k, the term goes to zero as $t \to \infty$. Moreover, the tail of the series $\sum_k \beta_k\|\eta_k\|$ is small. Therefore, by the Toeplitz lemma or direct estimation, the entire sum goes to zero.

Thus, $e_t \to 0$ *a.s.*, *i.e.*, $\mathbf{x}_t \to \mathbf{x}^*$ *a.s.* □

A natural question is whether our correction mechanism can guarantee that the residual bias stays proportional to the unavoidable approximation error. The next theorem answers this affirmatively, showing that the corrected TD fixed point is never worse than a constant-factor multiple of the best value function representable by the chosen features.

First, we define the total error as:

$$\begin{aligned} e &= \Phi x_c^{TD} - V_\pi \\ &= \Phi(x_c^{TD} - w^*) + (\Phi w^* - V_\pi) = \Phi u + \delta, \end{aligned} \tag{62}$$

where

- $w^* = \arg\min_w \|\Phi w - V^\pi\|_q$ is the best approximation under q-norm

- $u = x_c^{TD} - w^*$ is the difference between the TD solution and the best approximation

- $\delta = \Phi w^* - V^\pi$ is the approximation error

Lemma 5 (Orthogonality Condition for TD Fixed Point)**.** *The TD fixed point satisfies the orthogonality condition:*

$$\Phi^\top \Xi_{\nu_b}\Xi_{h^*}^{-1}\Xi_q(I - \gamma P_\pi)e = 0 \tag{63}$$

Proof. The TD fixed point satisfies:

$$\Phi^\top \Xi_{\nu_b}\Xi_{h^*}^{-1}\Xi_q(I - \gamma P_\pi)\Phi x_c^{TD} = \Phi^\top \Xi_{\nu_b}\Xi_{h^*}^{-1}\Xi_q R_\pi \tag{64}$$

Substituting the $R^\pi = (I - \gamma P_\pi)V_\pi$ (from Bellman equation) into the TD fixed point equation:

$$\Phi^\top \Xi_{\nu_b}\Xi_{h^*}^{-1}\Xi_q(I - \gamma P_\pi)\Phi x_c^{TD} = \Phi^\top \Xi_{\nu_b}\Xi_{h^*}^{-1}\Xi_q(I - \gamma P_\pi)V_\pi$$

Rearranging all terms to one side:

$$\begin{aligned}&\Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (I - \gamma P_\pi)(\Phi x_c^{TD} - V_\pi) = 0 \\ &\Rightarrow \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (I - \gamma P^\pi) e = 0 \end{aligned} \tag{65}$$

□

To rigorously validate the effectiveness of our distributional correction mechanism, we bound the approximation error relative to the fundamental limit imposed by the expressivity of the features and the target state weighting. The following result provides a worst-case guarantee that our method does not amplify unavoidable approximation errors and quantifies how design choices (e.g., the target distribution q, feature selection, and mixture model complexity) influence performance.

Theorem 6 (Error Bound for Off-Policy TD with Steady-State Bias Correction)**.** *Under Assumptions 1-5, the error of the off-policy TD solution with steady-state bias correction satisfies:*

$$\|\Phi x_c^{TD} - V^\pi\|_{\nu_b} \leq C \cdot \min_w \|\Phi w - V^\pi\|_q$$

where:

$$C = \left(\frac{\|P\| \cdot \|P^{-1}\|}{|\alpha(\Lambda_c)|} \right) \cdot \sqrt{\max_s \nu_b(s)} \cdot K \cdot \sqrt{\max_s q(s)} \sigma_{\max}(\Phi)^2 \cdot (1 + \gamma \sqrt{\kappa_q}) + \sqrt{\max_s \frac{\nu_b(s)}{q(s)}}$$

with $\kappa_q = \max_{s'} \dfrac{\sum_s q(s) P_\pi(s'|s)}{q(s')}$, *and* $\alpha(\Lambda_c) = \max_i Re(\lambda_i(\Lambda_c)) < 0$ *being the spectral abscissa of* Λ_c.

Proof. By applying triangle inequality on equation 62,

$$\|e\|_{\nu_b} \leq \|\Phi u\|_{\nu_b} + \|\delta\|_{\nu_b} \tag{66}$$

From Lemma 5, we have

$$\langle \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (I - \gamma P_\pi), e \rangle = 0 \tag{67}$$

Substituting $e = \Phi u + \delta$:

$$\begin{aligned}&\Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P_\pi - I) \Phi u = -\Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P_\pi - I) \delta \\ &\Rightarrow \Lambda_c u = -b_\delta,\end{aligned}$$

where $b_\delta = \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P_\pi - I)\delta$.
Since Λ_c is Hurwitz and diagonalizable (by Assumption 4), Λ_c is invertible and can be written as $\Lambda = PDP^{-1}$, where $D = \text{diag}(\lambda_1, \ldots, \lambda_k)$ with $\text{Re}(\lambda_i) < 0$ for all i. Therefore,

$$u = -\Lambda_c^{-1} b_\delta = -PD^{-1}P^{-1} b_\delta \tag{68}$$

No we bound $\|u\|$. Take norms:

$$\|u\| \leq \|\Lambda_c^{-1}\| \cdot \|b_\delta\| \leq \|P\| \cdot \|P^{-1}\| \cdot \|D^{-1}\| \cdot \|b_\delta\| \tag{69}$$

Since D is diagonal with entries λ_i:

$$\|D^{-1}\| = \max_i \left| \frac{1}{\lambda_i} \right| = \frac{1}{\min_i |\lambda_i|}$$

Let $\alpha(\Lambda_c) = \max_i \text{Re}(\lambda_i) < 0$. For any eigenvalue $\lambda_i = a_i + b_i i$, we have $|\lambda_i| = \sqrt{a_i^2 + b_i^2} \geq |a_i| = |\text{Re}(\lambda_i)|$. Therefore,

$$\|D^{-1}\| \leq \frac{1}{|\alpha(\Lambda_c)|} \tag{70}$$

So,

$$\|u\| \leq \left(\frac{\|P\| \cdot \|P^{-1}\|}{|\alpha(\Lambda_c)|} \right) \cdot \|b_\delta\| \tag{71}$$

Now we bound $\|b_\delta\| = \|\Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P^\pi - I)\delta\|$. Using the q-weighted inner product and Cauchy-Schwarz inequality:

$$\begin{aligned} |v^\top b_\delta| &= |\langle (\nu_b/h^*)\Phi v, (\gamma P_\pi - I)\delta \rangle_q| \\ &\leq \|(\nu_b/h^*)\Phi v\|_q \cdot \|(\gamma P_\pi - I)\delta\|_q \end{aligned} \tag{72}$$

First, bound $\|(\nu_b/h^*)\Phi v\|_q$:

$$\begin{aligned} \|(\nu_b/h^*)\Phi v\|_q^2 &= \sum_s q(s) \left(\frac{\nu_b(s)}{h^*(s)} \right)^2 (\Phi v(s))^2 \\ &\leq \max_s \left(\frac{\nu_b(s)}{h^*(s)} \right)^2 \cdot \max_s q(s) \cdot \|\Phi v\|^2 \\ &\leq K^2 \cdot \max_s q(s) \cdot \sigma_{\max}(\Phi)^2 \cdot \|v\|^2 \end{aligned}$$

Therefore,

$$\|(\nu_b/h^*)\Phi v\|_q \leq K \cdot \sqrt{\max_s q(s)} \cdot \sigma_{\max}(\Phi) \cdot \|v\| \tag{73}$$

Next, we bound $\|(\gamma P_\pi - I)\delta\|_q$. Let $\mu(s') = \sum_s q(s) P_\pi(s'|s)$, which is the next-state distribution under policy π when starting from distribution q. Similar to Lemma 1, one can obtain the following:

$$\begin{aligned} \|P^\pi f\|_q^2 &= \sum_s q(s)(P_\pi f(s))^2 \\ &\leq \sum_s q(s) \sum_{s'} P_\pi(s'|s) f(s')^2 \\ &\leq \underbrace{\left(\max_{s'} \frac{\mu(s')}{q(s')} \right)}_{\kappa_q} \cdot \|f\|_q^2 \end{aligned}$$

Then, $\|P_\pi\|_q \leq \sqrt{\kappa_q}$. Therefore,

$$\begin{aligned} \|(\gamma P_\pi - I)\delta\|_q &\leq \gamma \|P_\pi \delta\|_q + \|\delta\|_q \\ &\leq \gamma \sqrt{\kappa_q} \cdot \|\delta\|_q + \|\delta\|_q \\ &= (1 + \gamma\sqrt{\kappa_q}) \cdot \|\delta\|_q \end{aligned}$$

Combining these results:

$$\|b_\delta\| \leq K \cdot \sqrt{\max_s q(s)} \cdot \sigma_{\max}(\Phi) \cdot (1 + \gamma\sqrt{\kappa_q}) \cdot \|\delta\|_q \tag{74}$$

Now, we bound the approximation error under ν_b:

$$\begin{aligned} \|\delta\|_{\nu_b}^2 &= \sum_s \nu_b(s)\delta(s)^2 \\ &\leq \left(\max_s \frac{\nu_b(s)}{q(s)} \right) \cdot \sum_s q(s)\delta(s)^2 \\ &= \left(\max_s \frac{\nu_b(s)}{q(s)} \right) \cdot \|\delta\|_q^2 \end{aligned}$$

Therefore,

$$\|\delta\|_{\nu_b} \leq \sqrt{\max_s \frac{\nu_b(s)}{q(s)}} \cdot \|\delta\|_q \tag{75}$$

Finally, we combine all components:

$$\begin{aligned}\|e\|_{\nu_b} &\leq \|\Phi u\|_{\nu_b} + \|\delta\|_{\nu_b} \\ &\leq \left(\frac{\|P\| \cdot \|P^{-1}\|}{\alpha(\Lambda_c)}\right) \cdot \sqrt{\max_s \nu_b(s)} \cdot K \cdot \sqrt{\max_s q(s)} \cdot \sigma_{\max}(\Phi)^2 \cdot (1 + \gamma\sqrt{\kappa_q}) \cdot \|\delta\|_q + \sqrt{\max_s \frac{\nu_b(s)}{q(s)}} \cdot \|\delta\|_q\end{aligned}$$

Since $\|\delta\|_q = \min_w \|\Phi w - V_\pi\|_q$, we obtain,

$$\|\Phi x_c^{TD} - V_\pi\|_{\nu_b} \leq C \cdot \min_w \|\Phi w - V^\pi\|_q, \tag{76}$$

where

$$C = \left(\frac{\|P\| \cdot \|P^{-1}\|}{\alpha(\Lambda_c)}\right) \cdot \sqrt{\max_s \nu_b(s)} \cdot K \cdot \sqrt{\max_s q(s)}\sigma_{\max}(\Phi)^2 \cdot (1 + \gamma\sqrt{\kappa_q}) + \sqrt{\max_s \frac{\nu_b(s)}{q(s)}}.$$

□

The above theorem addresses demonstrates that the error of our corrected solution is proportional to the minimal approximation error under the target distribution q, scaled by factors capturing policy misalignment, feature conditioning, and steady-state estimation accuracy. This establishes that our algorithm achieves near-optimal performance within the constraints of the representation, while explicitly quantifying the cost of distribution shift correction. The bound further elucidates the trade-offs between policy similarity, distribution estimation quality, and feature design. This bound provides several important theoretical insights:

1. *Fundamental Error Relationship*: The error in the TD solution is proportional to the best possible approximation error, establishing that the algorithm achieves the best possible performance within the function approximation class.
2. *Steady-State Estimation Quality*: The term $K = \max_s \frac{\nu_b(s)}{h(s)}$ quantifies the impact of steady-state distribution estimation error. When $K \approx 1$ (accurate estimation), the bound tightens, validating the steady-state bias correction approach.
3. *Policy Alignment*: The term $(1 + \gamma\sqrt{\kappa_q})$ with $\kappa_q = \max_{s'} \frac{\mu(s')}{q(s')}$ measures policy dissimilarity. Smaller κ (more similar policies) leads to tighter bounds, explaining why off-policy learning becomes challenging with dissimilar policies.
4. *Feature Representation*: The term $\sigma_{\max}(\Phi)^2$ shows that well-conditioned feature representations (smaller $\sigma_{\max}$) lead to better error bounds.
5. *Distributional Factors*: The terms $\sqrt{\max_s \nu_b(s)}$, $\sqrt{\max_s q(s)}$, and $\sqrt{\max_s \frac{\nu_b(s)}{q(s)}}$ capture how state distribution properties affect performance.

Theorem 7 (Hurwitz Condition)**.** *$\Lambda_c = \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P^\pi - I)\Phi$ is Hurwitz (all eigenvalues have strictly negative real parts) if and only if $K_q \kappa_q \gamma^2 < 1$, where $\kappa_q = \max_{s'} \frac{\mu_q(s')}{q(s')}$ with $\mu_q(s') = \sum_s q(s) P_\pi(s'|s)$ and $K_q = \max_s \frac{\Xi_{\nu_b}(s)\Xi_{h^*}^{-1}(s)}{q(s)}$.*

Proof. Consider the quadratic form $\mathbf{w}^\top \Lambda_c \mathbf{w}$ for any $\mathbf{w} \neq 0$:

$$\mathbf{w}^\top \Lambda_c \mathbf{w} = \mathbf{w}^\top \Phi^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P_\pi - I)\Phi \mathbf{w}$$

Let $\mathbf{u} = \Phi\mathbf{w}$. By Assumption 2, $\mathbf{u} \neq 0$ since $\text{rank}(\Phi) = k$. Then:

$$\mathbf{w}^\top \Lambda_c \mathbf{w} = \mathbf{u}^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q (\gamma P_\pi - I)\mathbf{u} = \gamma \mathbf{u}^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q P_\pi \mathbf{u} - \mathbf{u}^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q \mathbf{u}$$

Let $Q_1 = \gamma \mathbf{u}^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q P_\pi u$ and $Q_2 = \mathbf{u}^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q \mathbf{u}$. First, observe that $Q_2 > 0$ since $\Xi_{\nu_b}, \Xi_{h^*}^{-1}, \Xi_q$ are all positive definite diagonal matrices (by Assumption 1 and the fact that h^* is a valid distribution estimate).

For Q_1, apply the Cauchy-Schwarz inequality:

$$|Q_1| \leq \gamma \|\mathbf{u}\|_{\Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q} \cdot \|P_\pi \mathbf{u}\|_{\Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q}, \text{ where } \|\mathbf{x}\|_{\Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q} = \sqrt{\mathbf{x}^\top \Xi_{\nu_b} \Xi_{h^*}^{-1} \Xi_q \mathbf{x}}.$$

Now, we need to bound $\|P_\pi \mathbf{u}\|^2_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q}$:

$$\begin{aligned}
\|P_\pi \mathbf{u}\|^2_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q} &= \mathbf{u}^\top (P_\pi)^\top \Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q P_\pi \mathbf{u} \\
&= \sum_s \sum_{s'} \sum_{s''} \mathbf{u}(s) P^\pi(s'|s)\Xi_{\nu_b}(s)\Xi^{-1}_{h^*}(s)\Xi_q(s)P_\pi(s''|s')\mathbf{u}(s'') \\
&= \sum_{s'} \Xi_q(s')u(s') \left(\sum_s \frac{\Xi_{\nu_b}(s)\Xi^{-1}_{h^*}(s)}{\Xi_q(s')} P^\pi(s'|s)u(s) \right)
\end{aligned}$$

Note that $K_q = \max_s \frac{\Xi_{\nu_b}(s)\Xi^{-1}_{h^*}(s)}{q(s)}$, is bounded since ν_b and h^* are positive distributions on a finite state space. Then:

$$\begin{aligned}
\|P_\pi \mathbf{u}\|^2_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q} &\le K_q \sum_{s'} q(s')\mathbf{u}(s') \left(\sum_s P_\pi(s'|s)\mathbf{u}(s) \right) \\
&\le K_q \sum_{s'} q(s')\mathbf{u}(s') \sqrt{\sum_s P_\pi(s'|s)\mathbf{u}(s)^2} \quad \text{(by Jensen's inequality)} \\
&\le K_q \sqrt{\sum_{s'} q(s')\mathbf{u}(s')^2} \cdot \sqrt{\sum_{s'} q(s') \sum_s P_\pi(s'|s)\mathbf{u}(s)^2} \\
&= K_q \|\mathbf{u}\|_q \cdot \sqrt{\sum_s \mathbf{u}(s)^2 \sum_{s'} q(s')P_\pi(s'|s)} \\
&= K_q \|\mathbf{u}\|_q \cdot \sqrt{\sum_s \mathbf{u}(s)^2 \mu_q(s)} \\
&\le K_q \sqrt{\kappa_q} \|\mathbf{u}\|_q^2 \quad \text{where } \kappa_q = \max_{s'} \frac{\mu_q(s')}{q(s')}
\end{aligned}$$

Thus, $\|P_\pi \mathbf{u}\|_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q} \le \sqrt{K_q\kappa_q}\|\mathbf{u}\|_q$. Now, since $\|\mathbf{u}\|_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q} \ge \sqrt{K_q^{-1}}\|\mathbf{u}\|_q$, we have:

$$|Q_1| \le \gamma \sqrt{K_q\kappa_q} \|\mathbf{u}\|^2_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q}$$

Therefore:

$$\mathbf{w}^\top \Lambda_c \mathbf{w} \le (\gamma\sqrt{K_q\kappa_q} - 1)\|\mathbf{u}\|^2_{\Xi_{\nu_b}\Xi^{-1}_{h^*}\Xi_q}$$

When $K_q\kappa_q\gamma^2 < 1$, we have $\gamma\sqrt{K_q\kappa_q} < 1$, and thus $\mathbf{w}^\top\Lambda_c\mathbf{w} < 0$ for all $w \neq 0$.

This proves that Λ_c is negative definite, and therefore all its eigenvalues have strictly negative real parts (Hurwitz). □

In Algorithm 1, the mixture weights $\bar{\lambda}_t$ must satisfy probability simplex constraints ($\lambda_i \ge 0$, $\sum_{i=1}^\ell \lambda_i = 1$) after each gradient update. To enforce this, we employ an efficient projection method that provides an optimal $O(\ell \log \ell)$ Euclidean projection onto the simplex Δ^ℓ Wang & Carreira-Perpinán (2013). The algorithm sorts components, determines an optimal threshold, and redistributes mass—yielding the closest valid point in Δ^ℓ while preserving sparsity patterns when possible. For this purpose, we aim to solve the following optimization problem:

$$\min_\lambda \frac{1}{2}\|\lambda - v\|^2, \quad \text{subject to } \sum_{i=1}^\ell \lambda_i = 1, \quad \lambda_i \ge 0. \tag{77}$$

Using a Lagrange multiplier τ for the equality constraint $\sum_{i=1}^\ell \lambda_i = 1$, we define the Lagrangian:

$$\mathcal{L}(\lambda, \tau) = \frac{1}{2}\sum_{i=1}^\ell (\lambda_i - v_i)^2 - \tau \left(\sum_{i=1}^\ell \lambda_i - 1 \right). \tag{78}$$

Now solving for λ_i by taking the derivative $w.r.t.$ λ_i:

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \lambda_i - v_i - \tau = 0 \Rightarrow \lambda_i = v_i + \tau. \tag{79}$$

Now to enforce the simplex constraint, we sum over all i:

$$\sum_{i=1}^{\ell} \lambda_i = \sum_{i=1}^{\ell} (v_i + \tau) = 1 \Rightarrow \tau = \frac{1 - \sum_{i=1}^{\ell} v_i}{\ell}. \tag{80}$$

Thus, the projection without considering the non-negativity constraint is: $\lambda_i = v_i + \frac{1-\sum_{i=1}^{\ell} v_i}{\ell}$. If any $\lambda_i < 0$, we modify the solution by clipping negative values to zero and redistributing the remaining weight. This is efficiently handled by sorting v in descending order and determining a threshold τ such that the projected vector remains non-negative.

Algorithm 2: Euclidean projection onto Δ^{ℓ}

Function $\Pi_{\Delta^{\ell}}(\lambda \in \mathbb{R}^{\ell})$
- Sort λ into η: $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_{\ell}$
- $\tau = \texttt{max}\{1 \leq j \leq \ell : \eta_j + \frac{1}{j}\left(1 - \sum_{i=1}^{j} \eta_i\right) > 0\}$
- $y = \frac{1}{\tau}\left(1 - \sum_{i=1}^{\tau} \eta_i\right)$
- return $\hat{\lambda} = \texttt{max}\{\lambda_i + y, 0\}$, $i \in \{1, 2, \ldots, \ell\}$

4 EXPERIMENTS & RESULTS

This section presents a comprehensive empirical evaluation of the proposed Steady-State Bias Correction (SSBC) algorithm across diverse benchmark domains. The experiments are designed to validate the method's effectiveness in mitigating steady-state distribution mismatch in off-policy TD learning with linear function approximation. We assess performance using Root Mean Square Error (RMSE) of value predictions: RMSE $= \|V_\pi - \Phi \mathbf{x}\|_2$ against true value functions. All results are averaged over 10 independent runs to ensure statistical robustness. Key aspects evaluated include:

- **Generalization across domains**: Discrete (Synthetic MDP, Circle Chain, Gridworld Cliff Walking, Taxi) and continuous (Mountain Car, CartPole, Acrobot) state spaces
- **Hyperparameter sensitivity**: Impact of step-sizes (α_t, β_t) on convergence
- **Trajectory robustness**: Performance under varying episodic path structures
- **Distributional fidelity**: Accuracy of Gaussian mixture approximations for stationary distributions
- **Discount factor sensitivity**: Impact of discount factor γ on prediction error.

The discount factor is fixed at $\gamma = 0.1$ universally. All environments are modified to ensure ergodicity (e.g., respawning agents in terminal states) for well-defined steady-state distributions.

4.1 Discrete Domain

4.1.1 Synthetic Random MDP

The environment contains a random Markov chain where the target and behavioral policies are stochastic in nature, and the probability of taking an action a for a given state s is random. The probability transition function and the rewards given to each (state, action, next-state) set are also random. The action and next

state are taken with respect to a random beta distribution. The initial state s_0 is randomly chosen for each trajectory. The number of states is $n = 100$, and actions $m = 6$. Here we consider the mixture width to be $\ell = 7$. For on-policy iterations, ρ_t will always equal 1. As seen in Figure 4, our algorithm performs better than both the base off-policy algorithm and the on-policy estimate.

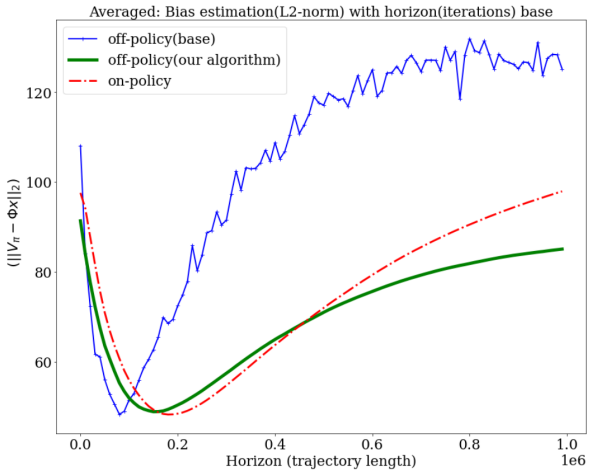

Figure 4: Synthetic Experiment: Base v/s optimized error *(lower is better)*

4.1.2 Circle Chain

Next, we tested a structured n-state circular chain. States 1 through n are arranged in a ring. The behavior policy π_b moves clockwise with probability ρ (and counter-clockwise $1 - \rho$), while the target policy π does the opposite (moves counter-clockwise with probability ρ). We chose $n = 50$ and $\rho = 0.8$, so the two policies induce markedly different stationary distributions (each tends to concentrate states in the direction it prefers). We used a mixture of $\ell = 9$ Gaussian components for $h_{\theta,\lambda}$. As the agent learns, π_b spends more time on one half of the ring and π on the other, creating a clear steady-state bias in the baseline off-policy TD. Figure 5 plots the value estimation error as a function of training episodes (each episode long enough to approach steady-state). Our algorithm achieves significantly lower error than the baseline, nearly matching the on-policy curve, especially as the horizon grows which validates that our approach corrects the skew in visitation.

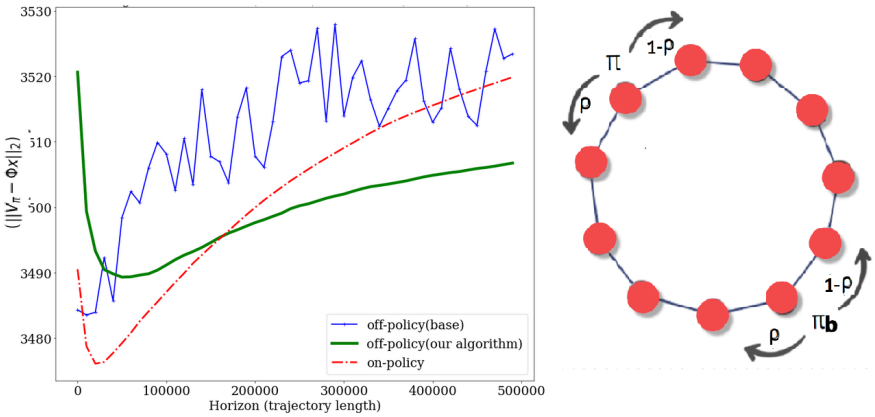

Figure 5: Circle chain: Base vs optimized error *(lower is better)*

4.1.3 Gridworld Cliff Walking

We evaluated on the classic Cliff Walking task (a 4×12 gridworld) Sutton & Barto (2018) from the OpenAI gym environment. The agent starts at the bottom-left and must reach the bottom-right goal, receiving -1 reward per step, except for the goal cell which yields a reward of $+10$ and a large penalty -100 if it steps off the cliff. We modified the task to an ongoing (non-terminating) version by respawning the agent at, start upon reaching the goal, so that a steady-state distribution exists. The target and behavior policies are randomly generated with different preferences (to induce distribution shift in how they navigate around the cliff). We used $\ell = 10$ mixture components for $h_{\theta,\lambda}$. Figure 6 shows the error of value predictions during training. The baseline off-policy TD is biased due to the behavior policy exploring the cliff area differently than the target. Our method again substantially reduces the error and remains more stable, confirming the benefit of steady-state correction in a practical gridworld with uneven state relevance.

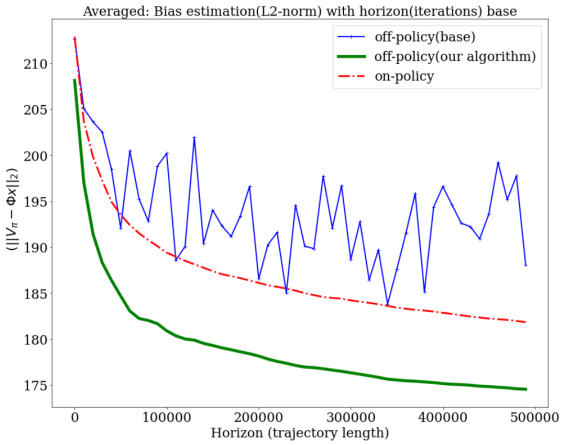

Figure 6: Cliff Walking Experiment: Base vs optimized error *(lower is better)*

4.1.4 Taxi Domain

Taxi is a 2-D grid world that mimics the movement of a taxi along the grid lines. The taxi can move in four directions, *i.e.*, `NORTH, EAST, SOUTH,` and `WEST`, and is also equipped to `pick up` or `drop off` passengers. This makes the number of possible actions $m = 6$. The agent receives a reward of $+20$ points for successful pick-up or drop-off, and a penalty of -1 for every move. In the original taxi environment, the simulation would terminate once the passenger was picked up and dropped off at the correct location. The environment has a total of 500 states, and the agent can take 6 possible actions as mentioned above. The number of Gaussians considered is $\ell = 10$.

From Figure 7, both the base off-policy algorithm and our approach achieve nearly identical results. This arises because the Taxi environment's state space exhibits an almost uniform steady-state distribution, thus incurring minimal bias. Consequently, the correction factor $h_{\widehat{\theta}^*,\widehat{\lambda}^*}$ is essentially the same across states, leaving little room for improvement over the baseline. Nevertheless, our algorithm shows slightly better stability.

4.2 Continuous Domain

4.2.1 Mountain Car

The Mountain Car problem Moore (1990) is a classic 2D control task in which an underpowered car must build momentum by oscillating between hills to ascend a steep slope. Its continuous state space is defined by two variables: *position* and *velocity*, which we discretize into 1200 states. At each state, the agent can select one of three actions—`drive left, drive right`, or `coast`. A negative reward is incurred every time step until the goal is reached; initially, the agent has no direct information about the goal position until it

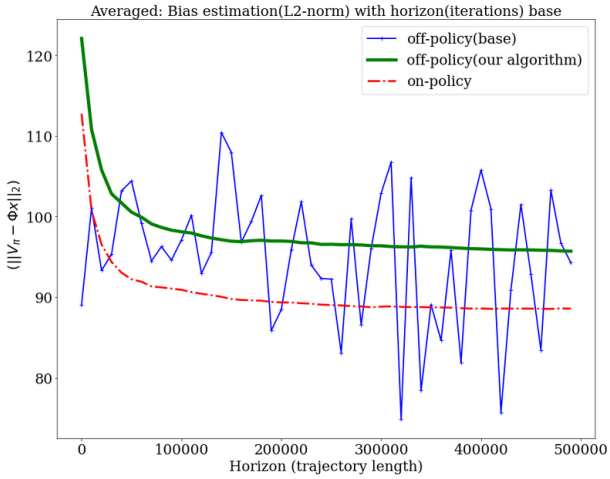

Figure 7: Taxi Domain: Base vs optimized error *(lower is better)*

succeeds once. We set the mixture component count to $\ell = 15$, and randomly choose both the target and behavior policies. As shown in Figure 8, our method achieves lower error than both the baseline off-policy approach and the on-policy counterpart

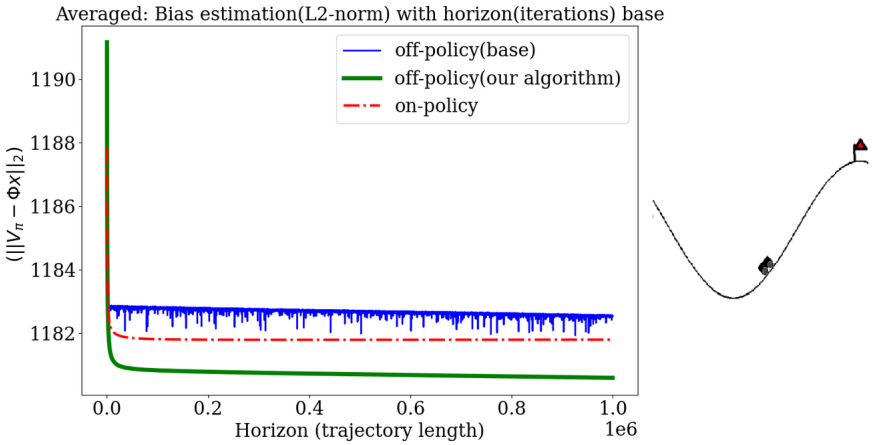

Figure 8: Mountain Car: Base vs optimized error *(lower is better)*

4.2.2 CartPole

The CartPole, or the inverted pendulum, is a simple RL experiment in which move a cart back and forth along a frictionless wire so that a pole pivoting on the cart balances upright. The possible actions include pushing the cart either to the left or right. The continuous state space defined by Cart position, velocity, pole angle and angular velocity is discretized into 1200 states. The mixture count of the Gaussians considered for the approximation of the SSD is $\ell = 15$. Target and behavior policies are randomly generated. The trajectories, rewards, and probability transitions are taken from the Gym environment Barto et al. (1983). As seen in Figure 9, our algorithm performs better than both the base off-policy algorithm and the on-policy estimate.

4.2.3 Acrobot

The Acrobot system consists of two limbs connected at a central joint which can rotate and move and the base fixed. The goal is to apply torques on the actuated joint to swing the free end of the linear chain

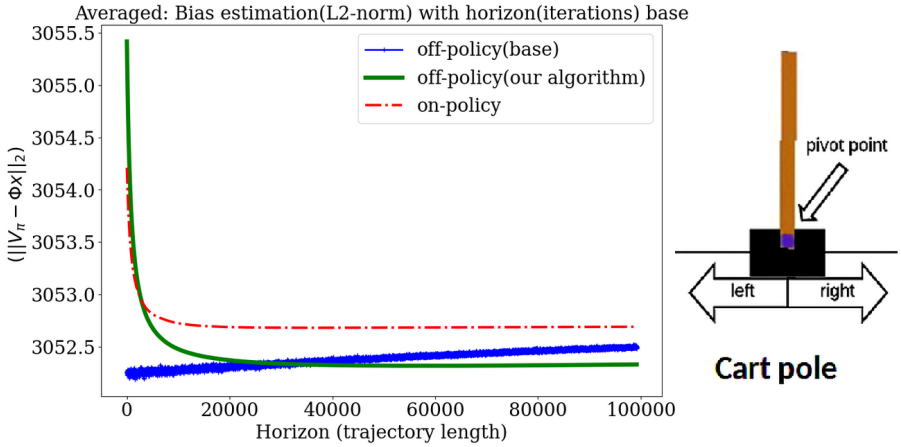

Figure 9: CartPole Experiment: Base vs optimized error *(lower is better)*

above a given height while starting from the initial state of hanging downwards. The continuous state space environment provides information about the rotational joint angles and angular velocity of the limbs and is discretized into 4096 states. θ_1 is the angle of the first joint, where an angle of 0 indicates the first link is pointing directly downwards. θ_2 is relative to the angle of the first link. An angle of 0 corresponds to having the same angle between the two links. Here we consider the mixture width to be $\ell = 100$. Target and behavior policies are randomly generated, each having a different distribution. As seen in Figure 10, our algorithm performs better than both the base off-policy algorithm and the on-policy estimate. We also provide here the likeliness between the behaviour policy steady-state probability distribution ν_b and the estimated surrogate distribution $h_{\widehat{\theta}^*, \widehat{\lambda}^*}$. The results are provided in Figure 11.

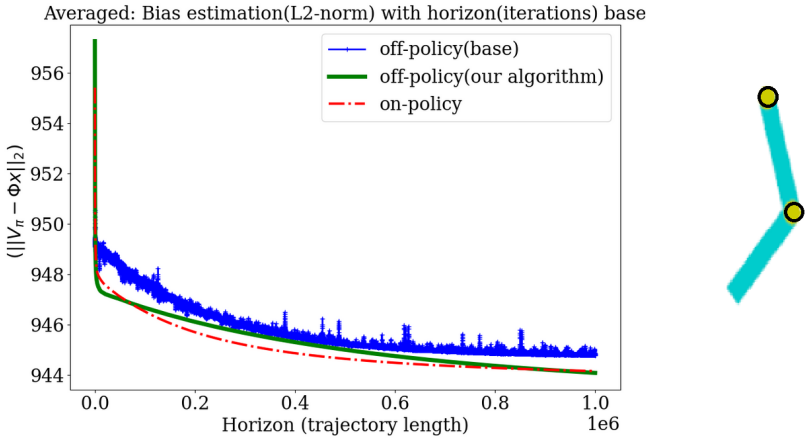

Figure 10: Acrobot Experiment: Base vs optimized error *(lower is better)*

4.3 Hyper-parameter Sensitivity

We empirically examine how the two timescale parameters—the fast TD step–size α_t and the slow importance–ratio step–size β_t—influence the accuracy and stability of our SSBC–TD algorithm. All experiments are conducted on the classic control task *MountainCar* from the `Gymnasium` suite. Our hyperparameter grid search reveals a critical interplay between the fast TD step-size (α_t) and the slow ratio step-size (β_t). The heat map (Figure 12) illustrates that the best-performing configurations cluster around higher α_t values, particularly when paired with smaller β_t values. The combination of $\alpha_t = 0.2$ and $\beta_t = 0.005$ achieves the lowest final RMSE, indicating an effective balance between rapid updates to the steady-state distribution

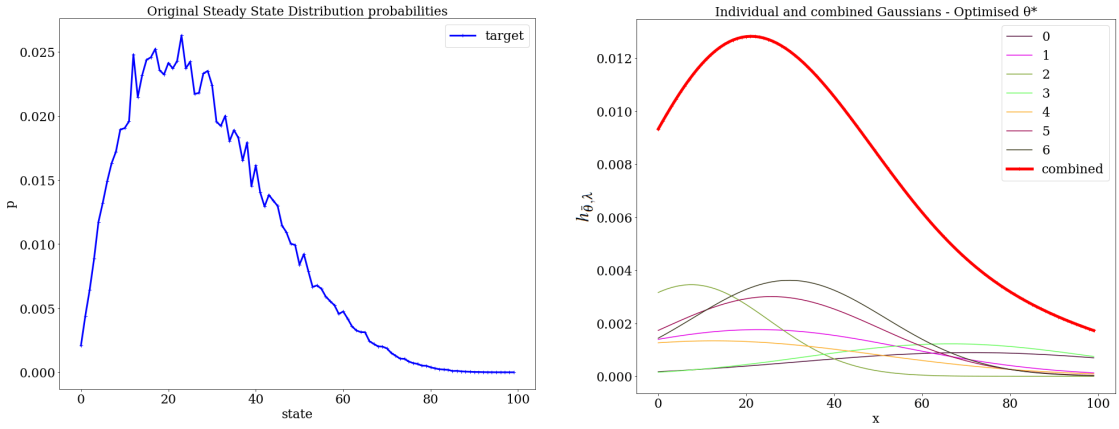

Figure 11: True steady-state distribution vs Estimated one

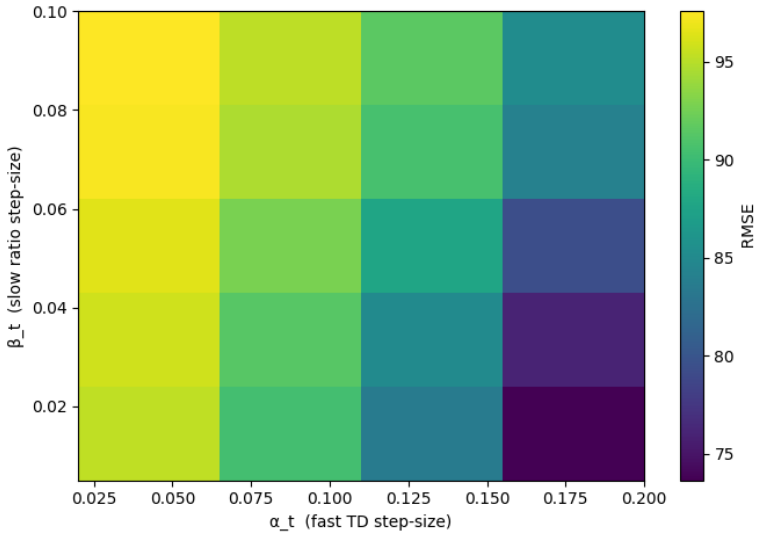

Figure 12: Hyperparameter sensitivity of SSBC-TD on MountainCar: Final RMSE as a function of the fast TD step-size (α_t) and slow ratio step-size (β_t). Darker shades indicate lower error. The minimum RMSE (marked) occurs at $\alpha_t = 0.2$ and $\beta_t = 0.005$, revealing that aggressive TD updates paired with conservative ratio estimation optimally balance convergence and stability.

approximation and gradual correction of the steady-state distribution mismatch. This result underscores the importance of carefully tuning these timescales to mitigate the "deadly triad" interaction between function approximation, bootstrapping, and off-policy learning. The curve plot (Figure 13) further validates this observation by showing how different α_t settings converge over episodes. Notably, the curve for $\alpha_t = 0.2$ exhibits the fastest decline in RMSE, stabilizing at the lowest error level compared to other configurations. Polyak averaging plays a crucial role in smoothing out fluctuations during training, as evidenced by the reduced variance in the RMSE curves across episodes. By incorporating Polyak averaging, our approach effectively mitigates the noise introduced by stochastic updates, leading to more stable and accurate value predictions.

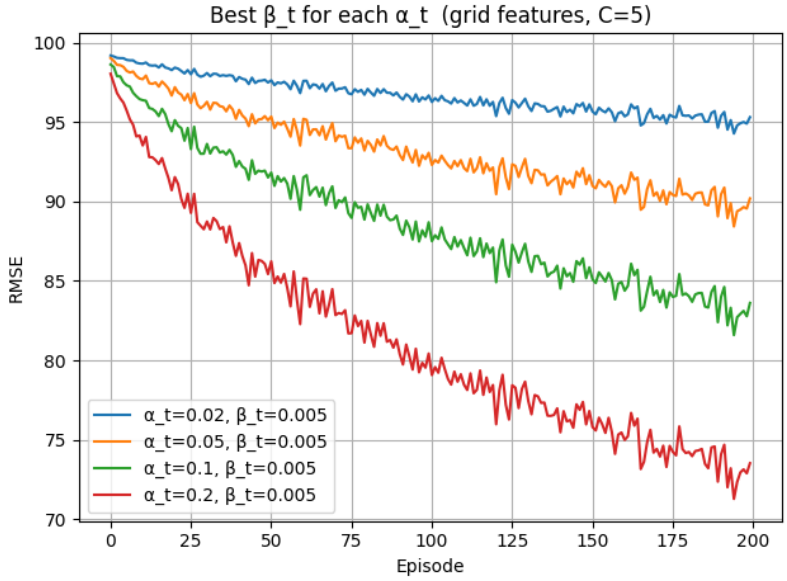

Figure 13: Convergence of SSBC-TD on MountainCar using the best β_t for each α_t (from Fig. 12). The $\alpha_t = 0.2$ curve ($\beta_t = 0.005$) achieves fastest error reduction and lowest asymptotic RMSE (≈ 70), demonstrating Polyak averaging's role in stabilizing high-step-size regimes. Smaller α_t (e.g., 0.02) exhibit slower convergence due to delayed implicit averaging.

Figure 14: Trajectories under uniform-random behavior policy.

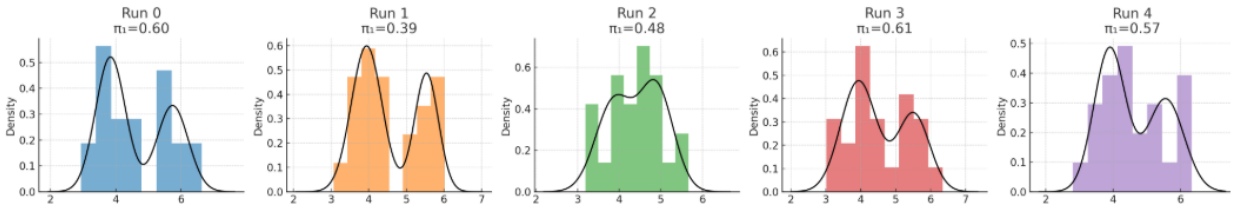

Figure 15: Comparison of the approximation of the true behavior policy's stationary distribution using the Gaussian mixtures across various trajectories

4.4 Trajectory Robustness

We evaluate trajectory-length sensitivity on a 5×5 Grid–world with absorbing terminals at $(0, 0)$ and $(4, 4)$. Each episode starts in the cell $(0, 4)$ and evolves for 60 time-steps under a *uniform–random* behaviour policy; upon reaching a terminal the agent is reset to the start cell and the trajectory continues. We generated five such trajectories (Fig. 14, Run 0–4), using distinct random seeds to expose path-level variability. We then

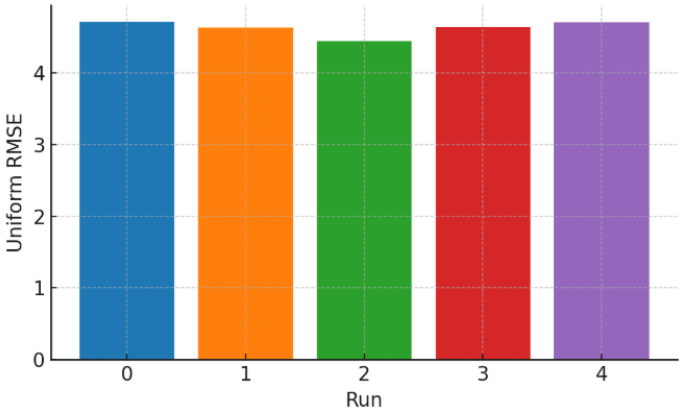

Figure 16: SSBC-TD demonstrates trajectory-agnostic stability, efficiently correcting steady-state bias regardless of path stochasticity

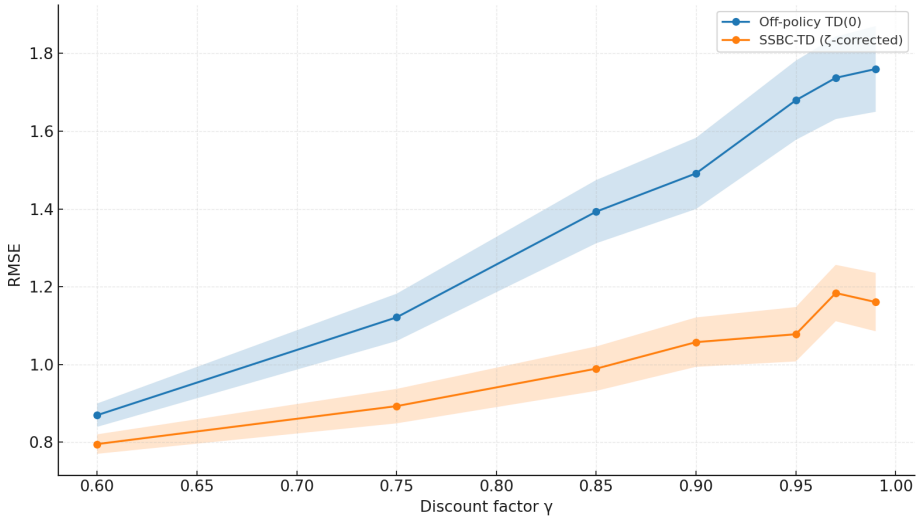

Figure 17: RMSE as a function of discount factor γ for plain off-policy TD(0) and steady-state bias–corrected TD (SSBC-TD) on a discrete Taxi control task. SSBC-TD consistently achieves lower RMSE, with the gap widening as γ increases.

applied a constant-stepsize SSBC-TD agent with $\beta_t = 0.05$. The micro-trajectories illustrate how an ergodic uniform policy can still visit states in markedly different orders at short horizons. Run 1 drifts almost exclusively downwards, Run 2 performs horizontal sweeps along the top row before descending, whereas Run 3 forms an almost symmetric lattice tour. Runs 0 and 4 highlight the "reset effect": a diagonal sprint to the lower left corner followed by immediate reinitialisation injects additional exploratory diversity that would otherwise take longer to accumulate. These early visitation biases explain the modest run-to-run spread observed in the residual histograms: SSBC-TD first adapts to whichever subset of states it samples most frequently. Over the long run updates, however, the random policy's mixing property smooths out those disparities: each run ultimately visits every state with frequency close to the stationary distribution, and the TD iterates converge to a common, tight error band (uniform-RMSE $\approx 4.5 \pm 0.1$).

4.5 γ Sensitivity

Here we study the relationship between the discount factor γ and the error in off-policy value prediction. The experiment considers a Taxi discrete control task. The state and action spaces are discrete. Both

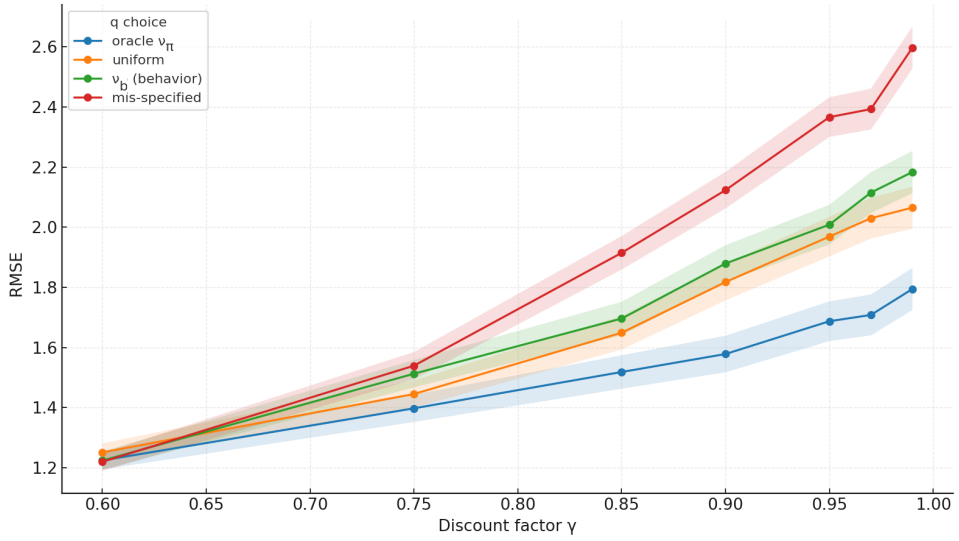

Figure 18: Ablation on the choice of stationary-distribution target q for SSBC-TD in a Taxi control task. Four variants are considered: oracle $q = \nu_\pi$ (true target policy distribution), uniform over states, behavior visitation distribution ν_b, and a deliberately mis-specified distribution. Closer alignment of q to ν_π yields consistently lower RMSE, with the advantage widening as $\gamma \to 1$.

TD(0) and SSBC-TD are trained off-policy using the same fixed behavior–target policy pair and identical feature representations, with final RMSE computed against Monte Carlo estimates under the target policy. The results are illustrated in Figure 17. As γ increases, the longer effective horizon exacerbates distribution mismatch and steady-state bias in plain TD(0), leading to a steady rise in RMSE and greater variability across runs. SSBC-TD, by applying ζ-weights that approximate the stationary distribution correction, mitigates this bias and maintains a lower error profile across γ values. The widening advantage of SSBC-TD for $\gamma \to 1$ aligns with the theoretical predictions: the stability condition $\kappa_b \gamma^2 < 1$ for plain TD becomes harder to satisfy at high γ, whereas SSBC-TD effectively reduces the mismatch constant, relaxing the condition to $K_q \kappa_q \gamma^2 < 1$, thereby extending the range of stable and accurate operation.

4.6 q Sensitivity

We study here the sensitivity of SSBC-TD to the choice of q, the stationary-distribution target used in its correction term. The discrete control task used here is a Taxi environment. We consider a discrete control task with fixed target and behavior policies, identical feature representation, and the same step-size schedules across all runs. The SSBC-TD algorithm is applied with four different choices of the stationary-distribution target q in its correction term ζ_t. The oracle choice uses the exact stationary distribution ν_π of the target policy; uniform assigns equal probability to all states; behavior uses the empirical stationary distribution ν_b from the behavior policy; and mis-specified uses a biased distribution that incorrectly overweights rarely visited states and underweights important ones. The RMSE vs γ curves show a clear ordering: the oracle q achieves the lowest error across all γ, with the margin over other choices growing as γ increases. This matches the theoretical prediction that the stability condition improves from $\kappa_b \gamma^2 < 1$ to $K_q \kappa_q \gamma^2 < 1$, where $K_q \kappa_q$ is minimized when $q = \nu_\pi$. Uniform q performs moderately well for smaller γ but suffers at large γ due to equal weighting of states that are rarely relevant to the target policy. Using ν_b offers only limited improvement, since it does not correct the long-horizon mismatch. The mis-specified q provides the smallest benefit and, for high γ, behaves similarly to plain TD, illustrating that poor q choices can erase the gains of the correction. This underscores the importance of accurate or well-chosen q for leveraging SSBC-TD's stability advantage in long-horizon settings.

5 CONCLUSION & FUTURE WORK

In this paper, we consider the off-policy value prediction in reinforcement learning, specifically in the context of linear function approximation. The proposed algorithm aims to minimize the steady-state bias in the off-policy value prediction, where the bias arises due to the differences in the sampling distribution of states and actions between the target policy and the behavior policy. Our work opens up several avenues for future research. First, integrating steady-state bias correction with deep value function approximators is a promising direction to tackle large-scale problems. Second, the idea of distribution correction might be extended to control settings: for example, off-policy actor-critic algorithms could use a similar mechanism to reweight the critic updates, or one could correct state occupancy in off-policy policy gradient methods. Third, an interesting theoretical question is how steady-state bias correction interacts with function approximation error and whether it can alleviate the deadly triad (function approximation, off-policy, and bootstrapping).

References

Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, (5):834–846, 1983.

Dimitri Bertsekas. *Reinforcement learning and optimal control*. Athena Scientific, 2019.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer, 2008.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.

Peter W Glynn and Donald L Iglehart. Importance sampling for stochastic simulations. *Management science*, 35(11), 1989.

Yongming Huang, Chunmei Xu, Cheng Zhang, Meng Hua, and Zhengming Zhang. An overview of intelligent wireless communications using deep reinforcement learning. *Journal of Communications and Information Networks*, 4(2):15–29, 2019.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pp. 6120–6130. PMLR, 2021.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

Sean Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.

Andrew William Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, 1990.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

Boris Teodorovich Polyak. A new method of stochastic approximation type. *Avtomatika i telemekhanika*, (7):98–107, 1990.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. 2000.

Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pp. 417–424, 2001.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ralph Tyrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 2015.

RY Rubinstein. Simulation and the monte carlo method. new york, ny, usa: John wiley&sons, 1981.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

David Silver, Thomas Hubert, Julian Schrittwieser, Antonoglou, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th annual international conference on machine learning*, pp. 993–1000, 2009.

Richard S Sutton, A Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631, 2016.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Miguel Tejedor, Ashenafi Zebene Woldaregay, and Fred Godtliebsen. Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artificial intelligence in medicine*, 104:101836, 2020.

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

J.N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

Cameron Voloshin, Hoang M Le, and Yisong Yue. Empirical analysis of off-policy policy evaluation for reinforcement learning. In *Real-world Sequential Decision Making Workshop at ICML*, volume 2019, 2019.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Yue Wang, Wei Chen, Yuting Liu, Zhi-Ming Ma, and Tie-Yan Liu. Finite sample analysis of the gtd policy evaluation algorithms in markov setting. *Advances in Neural Information Processing Systems*, 30, 2017.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: Intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

H. Yu. On convergence of emphatic temporal-difference learning. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1724–1751, Paris, France, 03–06 Jul 2015. PMLR.

Huizhen Yu. Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 50(6):3310–3343, 2012. doi: 10.1137/100807879.

Huizhen Yu. On convergence of some gradient-based temporal-differences algorithms for off-policy learning. *arXiv:1712.09652*, 2017.

Assaf J Zeevi and Ronny Meir. Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks*, 10(1):99–109, 1997.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pp. 11194–11203. PMLR, 2020.