

Limiting-Kernel $Q(\lambda)$: Bridging Short and Long Horizons

Tolga Ok¹, Arman Sharifi Kolarijani^{1,2}, Peyman Mohajerin Esfahani^{1,3}, and Mohamad Amin Sharifi Kolarijani¹

¹Delft University of Technology, The Netherlands

²Alpha Brain Technologies, The Netherlands

³University of Toronto, Canada

ABSTRACT. In reinforcement learning, more accurate policy evaluation has been frequently shown to improve downstream policy optimization performance. For tractability, most value estimators rely on n -step truncation and thereby omit the long-horizon behavior of the transition dynamics. Methods that instead capture this behavior through the global structure of the dynamics can converge faster, but their computational and memory requirements limit their scalability to large or continuous state spaces. To reconcile these limitations, we introduce Limiting-Kernel $Q(\lambda)$ (LKQL), an off-policy value estimator that uses the limiting kernel to approximate the long-horizon terms omitted by truncation. LKQL has the same order of complexity as n -step estimators and integrates directly into both on- and off-policy actor-critic algorithms. We prove that, under aperiodicity and in the near-on-policy regime, the operator underlying LKQL improves the policy evaluation convergence rate over its truncated counterpart for sufficiently large n , and that LKQL itself converges almost surely to the optimal values in finite Markov decision processes (MDPs) under a fixed behavior policy. On the MuJoCo continuous-control benchmark, we show that LKQL improves over n -step baselines in most settings, particularly on long-horizon tasks.

1. Introduction

Policy evaluation underlies much of value-based reinforcement learning. Given a Markov decision process (MDP), it entails finding the value function Q^π of a policy π , typically by solving the Bellman equation $Q = \mathcal{T}^\pi Q$, where $\mathcal{T}^\pi$ denotes the Bellman operator. In finite MDPs, the convergence rate of the overall algorithm is governed by how accurately each evaluation step solves the Bellman equation [42, 45]. In MDPs with large or continuous state spaces, policy evaluation instead determines the target values to which a parameterized value function is regressed. The suboptimality of the resulting greedy policy is in turn bounded by the accuracy of these targets [18, 46]. Empirically, more accurate targets are consistently found among the most effective design choices [24, 19, 23]. Together, these results motivate more accurate approximations of the full evaluation step, provided they remain computationally tractable and estimable from *rollouts*, i.e., trajectories sampled from the MDP.

Correspondence to: Tolga Ok <t.ok@tudelft.nl>.

This work was partially supported by the European Research Council (ERC) project TRUST-949796, the Horizon Europe Pathfinder Open project RELIEVE-101099481, and the NSERC Discovery grant RGPIN-2025-06544.

Existing approximations of the evaluation step build on the structure of its exact solution, which admits the closed form

$$Q^\pi = (\mathcal{I} - \gamma \mathcal{P}^\pi)^{-1} r = \sum_{k \geq 0} (\gamma \mathcal{P}^\pi)^k r, \quad (1)$$

where $\gamma \in (0, 1)$ is the discount factor, $\mathcal{P}^\pi$ is the transition operator of the chain induced by π , and r is the reward function. We refer to the inverse in (1) as the *preconditioner*. Applying this preconditioner to obtain Q^π , however, amounts to solving a linear system whose dimension is the number of state-action pairs.

One common approximation therefore truncates the series in (1) after n terms. This is an appealing choice since the truncated sum admits a computationally inexpensive estimator from a single rollout, at no cost beyond computing its temporal-difference (TD) errors. Estimators of this form, known as *return-based* or λ -return estimators, underpin a wide range of deep reinforcement learning algorithms. This includes the n -step returns of A3C [36], Rainbow [24], and R2D2 [25]; the generalized advantage estimation (GAE) of PPO [47, 48]; the V-trace targets of IMPALA [16]; and the off-policy returns including Retrace [37], Tree-Backup [41], Peng’s $Q(\lambda)$ [39, 27], and Harutyunyan’s $Q(\lambda)$ [22]. Moreover, these estimators update the values of only the sampled state-action pairs at each iteration. This locality renders them readily compatible with the stochastic-gradient training of parameterized value functions.

A second line of work instead approximates the preconditioner as a whole, leveraging the global structure of the underlying chain. This structure appears as the leading eigenvectors (eigenfunctions) of the transition operator in rank-one corrections [3] and Deflated Dynamics Value Iteration (DDVI) [29], an estimate of a stationary distribution in rank-one Q-learning (R1-QL) [49], and a direct estimate of the exact preconditioner in Zap Q-learning [14, 9]. Although their mechanisms and assumptions differ, these methods generally do not scale to MDPs with large or continuous state spaces, as they either require global updates over the entire state-action space at each iteration or rely on matrix operations that scale superlinearly with the number of parameters.

In this paper, we propose a *tractable* spectral approximation of the preconditioner, that is,

- (i) *estimable* from sampled rollouts, and
- (ii) *scalable* to large or continuous state spaces, with the same order of per-sample (i.e., rollout) complexity as return-based estimators.

Our approximation builds on the long-horizon behavior of the transition dynamics. For a chain whose recurrent classes are aperiodic, the powers $(\mathcal{P}^\pi)^k$ converge geometrically to the *limiting kernel* (LK), $\lim_{k \rightarrow \infty} (\mathcal{P}^\pi)^k$ [20, §4.4.2]. The LK thus provides an increasingly accurate approximation of these powers in the *tail* of the series in (1), particularly for terms with large k . This suggests a natural extension of n -step truncation, which we call *tail completion*, that retains the first n terms in (1) while approximating the truncated terms using the LK.

Based on *tail completion*, we develop a value estimator for *off-policy* learning, where the objective is to estimate the value function Q^π of a target policy π from rollouts generated by a behavior policy μ . Specifically, we build upon Harutyunyan’s $Q(\lambda)$ (HQL) [22], a non-conservative off-policy operator commonly implemented through n -step truncation. For HQL, the tail approximation reduces to a

single correction, which we call the *LK term*. The LK term tracks the long-run average of the TD errors and can be estimated alongside the value function, preserving the per-sample complexity and locality of return-based estimators.

The resulting estimator admits a realization entirely by TD learning, coupling the standard backup for the value functions with a second backup for the LK term. We term this realization the *double backup*, from which we derive our value learning algorithm, *Limiting-Kernel $Q(\lambda)$* (LKQL). In practice, each backup maintains its own function approximation, one for the values and one for the LK term, allowing LKQL to integrate directly into on- and off-policy actor-critic frameworks.

In summary, our contributions are as follows.

- (i) **Analysis of the truncated operator.** We derive explicit convergence rates for the n -step Harutyunyan operator in both policy evaluation and control (Lemma 2.1), characterizing how the truncation length n governs both the rate and region of convergence.
- (ii) **A limiting-kernel operator with provable rate improvement.** We introduce the LK preconditioner and establish convergence guarantees for the resulting LK Harutyunyan operator, characterizing the regime in which it improves on the rate of its truncated counterpart (Theorem 3.2, Corollary 3.3, and Theorem 3.5).
- (iii) **A scalable long-horizon extension of n -step estimators.** We introduce LKQL and prove that it converges almost surely to the optimal values under a fixed behavior policy and i.i.d. initial state-action pairs (Theorem 3.6). On the MuJoCo continuous-control benchmark, we demonstrate that LKQL improves over the baseline value estimators (namely, GAE [47] in on-policy learning and truncated HQL (THQL) in off-policy learning), particularly on long-horizon tasks (Section 4).

The remainder of the paper develops these contributions in turn. Section 2 reviews the operator preliminaries and analyzes the truncated operator. Section 3 develops the LK operators and LKQL. Section 4 introduces the actor-critic integration of LKQL and reports the empirical results. Section 5 concludes the paper by discussing the limitations and possible extensions of the current work. All the technical proofs are provided in Appendix A.

Notation. We write $\mathbb{N} \triangleq \{1, 2, \dots\}$ and $\mathbb{N}_0 \triangleq \{0, 1, 2, \dots\}$ for the positive and non-negative integers, respectively. $\mathbb{R}$ denotes the set of real numbers. We use calligraphic letters, as in $\mathcal{P}$ or $\mathcal{C}$, for matrices/operators on the space of Q -functions, with $\mathcal{I}$ denoting the identity matrix/operator. For a vector $V \in \mathbb{R}^d$, we use $\|V\|_p$ to denote its p -norm for $p \in \{1, 2, \infty\}$. For a matrix $\mathcal{M} \in \mathbb{R}^{d \times m}$, we use $\|\mathcal{M}\|_\infty$ for its operator ∞ -norm, that is, the maximum absolute row sum of $\mathcal{M}$. For any finite set $\mathbf{S}$, $\Delta(\mathbf{S})$ denotes the probability simplex over $\mathbf{S}$. For any set $\mathbf{S}$, $\mathbf{1} \in \mathbb{R}^{\mathbf{S}}$ is the all-ones vector/function with $\mathbf{1}(s) = 1$ for all $s \in \mathbf{S}$. Finally, $\mathbb{E}$ denotes the expectation operator.

2. Preliminaries

2.1. MDPs, policy evaluation, and control

In this paper, we consider an infinite-horizon discounted MDP $\langle \mathbf{S}, \mathbf{A}, \mathbb{P}, r, \gamma \rangle$ consisting of (i) a finite state space $\mathbf{S}$, (ii) a finite action space $\mathbf{A}$, (iii) a transition probability kernel $\mathbb{P}$ such that

$\mathbb{P}(\cdot|s, a) \in \Delta(\mathbf{S})$ gives the distribution of the next state for each state-action pair $(s, a) \in \mathbf{S} \times \mathbf{A}$, (iv) a reward function $r: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$, which is bounded and represents the immediate reward received for each state-action pair, and (v) a discount factor $\gamma \in (0, 1)$.

Let $\mathbf{Q} \triangleq \{Q \mid Q: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}\}$ denote the set of (Q-)functions that map $(s, a) \in \mathbf{S} \times \mathbf{A}$ to real values, and $\mathbf{\Pi} \triangleq \{\pi \mid \pi: \mathbf{S} \rightarrow \Delta(\mathbf{A})\}$ denote the set of stochastic policies that map states to distributions over actions. We now define several key functions and operators.

Given a Q-function $Q \in \mathbf{Q}$, we define the set $\mathbf{\Pi}_Q$ of greedy policies with respect to Q by

$$\mathbf{\Pi}_Q \triangleq \{\pi \in \mathbf{\Pi} \mid \pi(a|s) > 0 \Rightarrow a \in \arg \max_{b \in \mathbf{A}} Q(s, b), \forall (s, a) \in \mathbf{S} \times \mathbf{A}\}.$$

We use π_Q to denote a generic member of $\mathbf{\Pi}_Q$. We note that, in our analysis, the particular choice of greedy policy is inconsequential.

Given a policy $\pi \in \mathbf{\Pi}$, the *transition operator* $\mathcal{P}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ is defined by

$$\mathcal{P}^\pi Q(s, a) \triangleq \sum_{(s', a') \in \mathbf{S} \times \mathbf{A}} \pi(a'|s') \mathbb{P}(s'|s, a) Q(s', a'), \quad \forall (s, a) \in \mathbf{S} \times \mathbf{A}.$$

That is, $\mathcal{P}^\pi$ propagates Q-functions through the Markov chain induced by π (in the reverse direction with respect to $\mathbb{P}$).

The two main problems concerning an MDP are then as follows:

- (i) *Policy evaluation*: Given a policy $\pi \in \mathbf{\Pi}$, the goal is to find the state-action values of π defined by $Q^\pi \triangleq \sum_{k \geq 0} (\gamma \mathcal{P}^\pi)^k r = (\mathcal{I} - \gamma \mathcal{P}^\pi)^{-1} r$. In particular, the value Q^π of π admits the characterization $\mathcal{T}^\pi Q^\pi = Q^\pi$, i.e., the fixed point of the *Bellman operator* $\mathcal{T}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ defined by $\mathcal{T}^\pi Q \triangleq \gamma \mathcal{P}^\pi Q + r$.
- (ii) *Control*: The goal is to find an optimal policy $\pi^* \in \mathbf{\Pi}$ with the corresponding optimal value function

$$Q^*(s, a) \triangleq \sup_{\pi \in \mathbf{\Pi}} Q^\pi(s, a), \quad \forall (s, a) \in \mathbf{S} \times \mathbf{A},$$

such that $Q^{\pi^*} = Q^*$. The optimal value Q^* admits the characterization $\mathcal{T} Q^* = Q^*$, i.e., the fixed point of the *Bellman optimality operator* $\mathcal{T}: \mathbf{Q} \rightarrow \mathbf{Q}$ defined by $\mathcal{T} Q \triangleq \mathcal{T}^{\pi_Q} Q$.

Both Bellman operators are γ -contractions under the ∞ -norm, and thus repeated application of these operators to any Q-function converges to their respective fixed points $(\mathcal{T}^\pi)^k Q \rightarrow Q^\pi$ and $(\mathcal{T})^k Q \rightarrow Q^*$ as $k \rightarrow \infty$ for all $Q \in \mathbf{Q}$. This is indeed the basis for the corresponding value iteration algorithms.

Although our primary focus is on off-policy value learning, we first introduce the classical algorithms for the control problem, since they provide the operator view from which we construct the preconditioned operators underlying our value learning algorithms. The two classical algorithms for solving the control problem of an MDP are

$$\textbf{value iteration (VI): } Q_+ = \mathcal{T} Q = Q + \mathcal{I}(\mathcal{T} Q - Q),$$

$$\textbf{policy iteration (PI): } Q_+ = \lim_{k \rightarrow \infty} (\mathcal{T}^{\pi_Q})^k Q = Q + (\mathcal{I} - \gamma \mathcal{P}^{\pi_Q})^{-1} (\mathcal{T} Q - Q).$$

Both VI and PI admit geometric convergence to Q^* in the ∞ -norm at the same rate γ . This rate, however, does not capture the practical speed of PI. In finite MDPs, exact PI terminates after finitely

many iterations [45] and typically reaches a neighborhood of Q^* in far fewer iterations than VI. The speed-up in PI is attributed to its “preconditioner” $(\mathcal{I} - \gamma\mathcal{P}^{\pi_Q})^{-1}$ [50], which plays the role of the inverse Jacobian when PI is viewed as Newton’s method for the Bellman equation [38]. This preconditioner fully propagates the Bellman residual $(\mathcal{T}Q - Q)$ along the Markov chain induced by π_Q , thereby performing a *complete* policy evaluation step. By contrast, VI employs the identity preconditioner $\mathcal{I}$ and hence performs only a *partial* policy evaluation step. This dichotomy motivates the search for a preconditioner that captures sufficient structure of the transition dynamics to accelerate convergence without requiring full policy evaluation.

Several algorithms interpolate between VI and PI. These algorithms mainly fall under the umbrella of *Generalized Policy Iteration* (GPI) algorithms [52, §4.6]. To elaborate, for a policy π , let us define the n -step, λ -discounted preconditioner operator $\mathcal{C}_{n,\lambda}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ by

$$\mathcal{C}_{n,\lambda}^\pi \triangleq \sum_{k=0}^{n-1} (\lambda\gamma\mathcal{P}^\pi)^k, \quad \pi \in \Pi, \quad n \in \mathbb{N}, \quad \lambda \in (0, 1].$$

Let us also define the corresponding *on-policy* operator $\mathcal{T}_{n,\lambda}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ by

$$\mathcal{T}_{n,\lambda}^\pi Q \triangleq Q + \mathcal{C}_{n,\lambda}^\pi (\mathcal{T}^\pi Q - Q), \quad \pi \in \Pi, \quad n \in \mathbb{N}, \quad \lambda \in (0, 1]. \quad (2)$$

Observe that $\mathcal{C}_{1,\lambda}^\pi = \mathcal{I}$ (and hence $\mathcal{T}_{1,\lambda}^\pi = \mathcal{T}^\pi$) and $\mathcal{C}_{\infty,\lambda}^\pi = (\mathcal{I} - \lambda\gamma\mathcal{P}^\pi)^{-1}$ for all $\lambda \in (0, 1]$. Then, two canonical examples of GPI are

$$\textbf{modified policy iteration (MPI):} \quad Q_+ = \mathcal{T}_{n,1}^{\pi_Q} Q,$$

$$\textbf{\lambda-policy iteration (LPI):} \quad Q_+ = \mathcal{T}_{\infty,\lambda}^{\pi_Q} Q.$$

The MPI operator essentially performs a truncated *policy evaluation* step determined by n , whereas LPI discounts the contribution of the Bellman residual by a factor of λ .

In model-free RL, we only have access to sampling oracles (sample-based estimators) for single-step operators (i.e., with identity preconditioner) such as $\mathcal{T}^\pi$ (e.g., SARSA [52, §6.4]) and $\mathcal{T}$ when $\arg \max_{a \in \mathbf{A}} Q(\cdot, a)$ is available (e.g., Q-learning [55]). In n -step on-policy algorithms, one uses the same policy (π in the case of policy evaluation or π_Q in the case of control) for both the preconditioner $\mathcal{C}_{n,\lambda}^\pi$ and the Bellman residual $(\mathcal{T}^\pi Q - Q)$, as can be seen in (2). However, in the off-policy setting, the transition operator $\mathcal{P}^\pi$ of the target policy π or its sampling oracle is unavailable or undesirable (e.g., for better exploration). Instead, we use the transition operator $\mathcal{P}^\mu$ of the behavior policy μ to propagate the Bellman residual $(\mathcal{T}^\pi Q - Q)$ of the target policy π . In order to account for the discrepancy between μ and π , one may also modify the transition operator $\mathcal{P}^\mu$ by including a correction term. To be precise, following Munos et al. [37], for a correction function $c: \mathbf{S} \times \mathbf{A} \rightarrow \mathbb{R}$, we define the *corrected* transition operator $\mathcal{P}^{c,\mu}: \mathbf{Q} \rightarrow \mathbf{Q}$ by

$$\mathcal{P}^{c,\mu} Q(s, a) \triangleq \sum_{(s', a') \in \mathbf{S} \times \mathbf{A}} c(s', a') \mu(a'|s') \mathbb{P}(s'|s, a) Q(s', a'), \quad \forall (s, a) \in \mathbf{S} \times \mathbf{A}.$$

Depending on the choice of c , off-policy operators fall into two categories: conservative and non-conservative. In the conservative category, a canonical choice for c is the *importance sampling* (IS) ratio $c(s, a) = \pi(a|s)/\mu(a|s)$ when both distributions are known. IS-based estimators, although unbiased, suffer from high variance [37]. To address this issue, many low-variance alternatives for c have been proposed without hindering value learning [37, 41]. Non-conservative operators, on the other hand,

use a fixed c , such as $c = \mathbf{1} \in \mathbf{Q}$, i.e., the all-ones function. As a result, these operators do not require knowledge of μ and are less restrictive. The Harutyunyan [22] and Peng & Williams [39, 27] operators are examples of non-conservative off-policy operators.

2.2. Truncated Harutyunyan operator

In this work, we focus on *non-conservative* n -step off-policy operators that utilize the preconditioners defined over a behavior policy. To this end, for a behavior policy μ and a target policy π , we define the n -step Harutyunyan operator $\mathcal{R}_{n,\lambda}^{\mu,\pi}: \mathbf{Q} \rightarrow \mathbf{Q}$ by

$$\mathcal{R}_{n,\lambda}^{\mu,\pi} Q \triangleq Q + \mathcal{C}_{n,\lambda}^{\mu}(\mathcal{T}^{\pi} Q - Q), \quad \mu, \pi \in \mathbf{\Pi}, \quad n \in \mathbb{N}, \quad \lambda \in (0, 1]. \quad (3)$$

We next define the corresponding Harutyunyan iterations that form the value update algorithms by

$$\begin{aligned} \mathbf{HI}(\pi) : \quad Q_{\ell+1} &= \mathcal{R}_{n,\lambda}^{\mu,\pi} Q_{\ell}, \quad \ell \in \mathbb{N}_0, \quad Q_0 \in \mathbf{Q}; & (\text{policy evaluation}) \\ \mathbf{HI} : \quad Q_{\ell+1} &= \mathcal{R}_{n,\lambda}^{\mu,\pi} Q_{\ell}, \quad \ell \in \mathbb{N}_0, \quad Q_0 \in \mathbf{Q}. & (\text{control}) \end{aligned} \quad (4)$$

We note that the *non-truncated* operator $\mathcal{R}_{\infty,\lambda}^{\mu,\pi}$ corresponds to Harutyunyan's original $Q(\lambda)$ (HQL) [22], which introduces the so-called *off-policy correction* terms. In this work, we focus on the generic n -step truncated operator in (3) as it is widely used in practice [36, 24, 16, 25]. We start with the convergence guarantees for the truncated operator. For this purpose, following Harutyunyan et al. [22], we define the maximum statewise difference $\epsilon \in [0, 2]$ between two policies as

$$\epsilon \triangleq \max_{s \in \mathbf{S}} \|\mu(\cdot|s) - \pi(\cdot|s)\|_1, \quad \mu, \pi \in \mathbf{\Pi}.$$

Moreover, given scalars γ and λ , define the ϵ -dependent coefficient

$$\eta_{\epsilon} \triangleq \frac{\gamma(1 + \lambda(\epsilon - 1))}{1 - \lambda\gamma},$$

and the corresponding boundary values

$$\eta_0 \triangleq \frac{\gamma(1 - \lambda)}{1 - \lambda\gamma} \quad \text{and} \quad \eta_2 \triangleq \frac{\gamma(1 + \lambda)}{1 - \lambda\gamma}.$$

The preceding coefficient characterizes the geometric convergence rates of the original Harutyunyan operators (with $n = \infty$) to their corresponding fixed points. In policy evaluation, the iteration converges to Q^{π} with rate η_{ϵ} if $\epsilon < \frac{1-\gamma}{\gamma\lambda}$. In particular, η_0 is the fastest rate corresponding to the *on-policy* setting with $\epsilon = 0$. In control, the iteration converges to $Q^{\star}$ with rate η_2 if $\lambda < \frac{1-\gamma}{2\gamma}$ [22]. The following lemma characterizes the convergence rate of the truncated, n -step Harutyunyan operators.

Lemma 2.1 (Convergence rate of $\mathbf{HI}(\pi)$ & $\mathbf{HI}$). *Consider the n -step Harutyunyan operator (3) and the corresponding iterations in (4).*

- (Policy evaluation) *The iterates of $\mathbf{HI}(\pi)$ satisfy $\|Q_{\ell} - Q^{\pi}\|_{\infty} \leq \alpha_{\epsilon}^{\ell} \|Q_0 - Q^{\pi}\|_{\infty}$ for all $\ell \in \mathbb{N}_0$, where*

$$\alpha_{\epsilon} \triangleq (\gamma\lambda)^{n-1}(\gamma - \eta_{\epsilon}) + \eta_{\epsilon}.$$

In particular, for $n \geq 2$, $\alpha_e < 1$ and hence $Q_\ell \rightarrow Q^\pi$ if

$$\epsilon < \epsilon_{\max} \triangleq \frac{1-\gamma}{\gamma\lambda} \cdot \frac{1-(\gamma\lambda)^n}{1-(\gamma\lambda)^{n-1}}. \quad (5)$$

- (Control) The iterates of HI satisfy $\|Q_\ell - Q^*\|_\infty \leq \beta_c^\ell \|Q_0 - Q^*\|_\infty$ for all $\ell \in \mathbb{N}_0$, where

$$\beta_c \triangleq \eta_2 - \frac{(1+\gamma)(\gamma\lambda)^n}{1-\gamma\lambda}.$$

In particular, $\beta_c < 1$ and hence $Q_\ell \rightarrow Q^*$ if

$$\lambda < \frac{1-\gamma}{2\gamma} + \frac{(1+\gamma)(\gamma\lambda)^n}{2\gamma}. \quad (6)$$

Remark 2.2 (Convergence of HI). The implicit condition (6) can be characterized explicitly as

$$\lambda < \lambda_{\max} \triangleq \begin{cases} 1/\gamma, & n = 1, \\ \bar{x}/\gamma \in (\frac{1-\gamma}{2\gamma}, \frac{1-\gamma}{\gamma(1+\gamma)}], & n \geq 2, \end{cases} \quad (7)$$

where, for $n \geq 2$, $\bar{x} > 0$ uniquely solves the equation $\sum_{i=1}^{n-1} x^i = \frac{1-\gamma}{1+\gamma}$. Observe that the upper bound $\lambda_{\max}$ is a decreasing function of n . Moreover, $\lambda_{\max} > 1$, i.e., the condition becomes trivial, if any of the following conditions holds:

$$(i) \ n = 1, \quad (ii) \ \gamma \leq \frac{1}{3}, \quad (iii) \ n < \frac{\log(3\gamma - 1) - \log(1 + \gamma)}{\log \gamma} \quad \text{for } \gamma > \frac{1}{3}.$$

Some remarks are in order regarding the preceding result. First, observe that the extreme values of n recover the standard convergence results as expected. For $n = 1$, the rates α_e for policy evaluation and β_c for control are both equal to γ . Moreover, when $n = 1$, the convergence bounds become trivial (i.e., $\epsilon_{\max} = \infty$ and $\lambda_{\max} = \gamma^{-1} > 1$), recovering the convergence behavior of the standard value iteration (indeed, $\mathcal{R}_{1,\lambda}^{\mu,\pi} = \mathcal{T}^\pi$ for all $\mu \in \mathbf{\Pi}$ and $\lambda \in (0, 1]$). On the other hand, as $n \rightarrow \infty$, we have $\alpha_e \rightarrow \eta_\epsilon$ and $\epsilon_{\max} \rightarrow \frac{1-\gamma}{\gamma\lambda}$ for policy evaluation and $\beta_c \rightarrow \eta_2$ and $\lambda_{\max} \rightarrow \frac{1-\gamma}{2\gamma}$ for control, recovering the convergence properties of the original non-truncated operators.

An interesting observation is that for both policy evaluation and control, truncation improves the stability of the iteration by increasing the upper bounds $\epsilon_{\max}$ and $\lambda_{\max}$ in the conditions (5) and (7) for convergence. Indeed, similar to the role of λ , the truncation length n controls the reliance on Bellman residuals from future steps. Hence, shortening n leads to a more stable iteration that can handle a larger policy discrepancy.

Regarding the convergence rate, in control, we have that $\gamma \leq \beta_c < \eta_2$ for all $n \in \mathbb{N}$, indicating that truncation always improves the convergence rate. However, in policy evaluation, we have two regimes based on the policy discrepancy. If $\epsilon < 1 - \gamma$, then $\eta_\epsilon < \alpha_e \leq \gamma$ for all $n \in \mathbb{N}$, meaning that truncation worsens the convergence rate. In contrast, if $\epsilon > 1 - \gamma$, then $\gamma \leq \alpha_e < \eta_\epsilon$ for all $n \in \mathbb{N}$, implying that truncation improves the convergence rate. This can be explained by the fact that as ϵ vanishes, the iteration becomes on-policy with convergence rate $\eta_0 < \gamma$ for all $\lambda \in (0, 1]$. At the operator level, Lemma 2.1 thus provides explicit convergence rates for the truncated n -step Harutyunyan operator in both policy evaluation and control.

2.3. Limiting kernel, aperiodicity, and unichains

We finish this section by defining the limiting kernel and reviewing aperiodicity and unichains. So, let us fix a policy $\nu \in \Pi$ and consider the transition operator $\mathcal{P}^\nu$ of the Markov chain induced by ν . We define the *limiting kernel (LK)* $\mathcal{P}_\infty^\nu: \mathbf{Q} \rightarrow \mathbf{Q}$ to be the corresponding Cesàro limit [42, App. A.4], that is,

$$\mathcal{P}_\infty^\nu \triangleq \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{j < k} (\mathcal{P}^\nu)^j.$$

We note that the limiting kernel exists for every finite chain. Moreover, it is idempotent and absorbs $\mathcal{P}^\nu$ from either side [5, §3.2], that is,

$$\mathcal{P}_\infty^\nu \mathcal{P}_\infty^\nu = \mathcal{P}_\infty^\nu \mathcal{P}^\nu = \mathcal{P}^\nu \mathcal{P}_\infty^\nu = \mathcal{P}_\infty^\nu.$$

Let us also define the finite-horizon limiting-kernel residual by

$$\delta_n^\nu \triangleq \|(\mathcal{P}^\nu)^n - \mathcal{P}_\infty^\nu\|_\infty \in [0, 2], \quad n \in \mathbb{N}_0.$$

Note that $\delta_n^\nu \leq 2$, independent of ν and n . This upper bound, however, can be significantly improved for a certain class of chains, as we discuss next.

The transition operator $\mathcal{P}^\nu$ is called *aperiodic* if every recurrent class of the induced chain is aperiodic. In this case, the powers of $\mathcal{P}^\nu$ converge as $k \rightarrow \infty$ [20, §4.3.5], and the limiting kernel is equivalently given by

$$\mathcal{P}_\infty^\nu = \lim_{k \rightarrow \infty} (\mathcal{P}^\nu)^k.$$

For an aperiodic chain, the limiting kernel consists of stationary distributions of the chain on the recurrent classes. That is, $\mathcal{P}_\infty^\nu Q(s, a)$ is a weighted combination of the stationary averages of Q across the reachable recurrent classes from (s, a) . It follows from Gallager [20, §4.4.2] that for an aperiodic $\mathcal{P}^\nu$, we can define the *mixing rate* $\sigma_\nu \in (0, 1)$ and *mixing constant* $c_\nu < \infty$ such that

$$\delta_n^\nu \leq c_\nu \sigma_\nu^n, \quad \forall n \in \mathbb{N}_0. \quad (8)$$

We note that $\sigma_\nu \in (\sigma_\nu^*, 1)$, where σ_ν^* is the largest modulus among the eigenvalues of $\mathcal{P}^\nu$ other than 1, with $\sigma_\nu^* \triangleq 0$ if no such eigenvalue exists. Similarly, we can define the *mixing time* $n_\nu \in \mathbb{N}$ of $\mathcal{P}^\nu$ by

$$n_\nu \triangleq \min\{n \in \mathbb{N} \mid \delta_n^\nu < 1\}. \quad (9)$$

We call the chain induced by ν a *unichain* if it has a single recurrent class, possibly together with transient states [20, Def. 4.3.2]. When the induced chain is a unichain, the limiting kernel $\mathcal{P}_\infty^\nu$ reduces to a rank-one operator, given by $\mathcal{P}_\infty^\nu Q \triangleq \mathbf{1} \mathbb{E}_{(s,a) \sim d_\nu}[Q(s, a)]$, where $d_\nu \in \Delta(\mathbf{S} \times \mathbf{A})$ is the unique stationary distribution. If the unichain is also aperiodic, then $(\mathcal{P}^\nu)^k Q \rightarrow \mathbf{1} \mathbb{E}_{(s,a) \sim d_\nu}[Q(s, a)]$ as $k \rightarrow \infty$ for every $Q \in \mathbf{Q}$.

3. Tail-completed operators

In the previous section, we defined the truncated off-policy policy evaluation operator $\mathcal{R}_{n,\lambda}^{\mu,\pi}$ in (3) for a given target policy π and behavior policy μ . In this section, we present *tail-completed* versions of these operators using *limiting-kernel* preconditioners, together with the corresponding convergence guarantees and acceleration over the truncated operators under some conditions.

3.1. Limiting-kernel preconditioner

Many on- and off-policy operators build their preconditioner by approximating the PI preconditioner via n -step truncation and/or averaging, as we discussed in the previous section. In contrast, operator splitting (OS) and the corresponding OS-VI algorithm of Rakhsha et al. [43] leverage the decomposition of complete *policy evaluation*, i.e., the computation of $\mathcal{C}_{\infty,1}^\pi$, through inversion of “simpler” surrogates. To be precise, given a policy π and an approximate transition operator $\hat{\mathcal{P}}^\pi$, OS-VI introduces the *Varga operator* $\mathcal{V}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ defined by

$$\mathcal{V}^\pi Q \triangleq (\mathcal{I} - \gamma \hat{\mathcal{P}}^\pi)^{-1} (r + \gamma(\mathcal{P}^\pi - \hat{\mathcal{P}}^\pi)Q). \quad (10)$$

That is, $\mathcal{P}^\pi$ is replaced by the approximation $\hat{\mathcal{P}}^\pi$ in the preconditioner of $\mathcal{V}^\pi$. Notably, under $\pi = \pi_Q$, the Varga operator reduces to VI, LPI, and PI for $\hat{\mathcal{P}}^{\pi_Q} = 0$, $\hat{\mathcal{P}}^{\pi_Q} = \lambda \mathcal{P}^{\pi_Q}$, and $\hat{\mathcal{P}}^{\pi_Q} = \mathcal{P}^{\pi_Q}$, respectively.

The choice of $\hat{\mathcal{P}}^\pi$ is crucial for the convergence rate of the Varga operator. However, since the preconditioner involves inverting $(\mathcal{I} - \gamma \hat{\mathcal{P}}^\pi)$ and must be applicable in a model-free setting, the choice of $\hat{\mathcal{P}}^\pi$ must ensure both tractable invertibility and estimability. The former condition motivated the low-rank approximations of $\mathcal{P}^\pi$ in DDVI [29], which approximate $\mathcal{P}^\pi$ via its leading spectral components to invert the preconditioner efficiently. Ensuring estimability further constrains the choice of approximation. For instance, R1-QL [49] forms a rank-one surrogate from a power-method estimate of a stationary distribution of $\mathcal{P}^\pi$.

We instead approximate $\mathcal{P}^\pi$ by its limiting kernel $\mathcal{P}_\infty^\pi$. We show that this approximation satisfies both conditions put forward previously, namely, tractable invertibility and estimability from the samples. We defer the discussion on estimation to a later section and focus on the resulting operator and its convergence properties below.

Observe that, since the limiting kernel $\mathcal{P}_\infty^\pi$ is idempotent, the inversion in the preconditioner with $\mathcal{P}_\infty^\pi$ admits the closed form

$$(\mathcal{I} - \gamma \mathcal{P}_\infty^\pi)^{-1} = \mathcal{I} + \frac{\gamma}{1 - \gamma} \mathcal{P}_\infty^\pi.$$

Then, substituting $\hat{\mathcal{P}}^\pi$ with $\mathcal{P}_\infty^\pi$ in the Varga operator (10) yields the *limiting-kernel Varga operator* $\tilde{\mathcal{V}}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ given by

$$\tilde{\mathcal{V}}^\pi Q \triangleq Q + \left(\mathcal{I} + \frac{\gamma}{1 - \gamma} \mathcal{P}_\infty^\pi \right) (\mathcal{T}^\pi Q - Q).$$

As we discussed above, when the chain induced by π is a unichain, the limiting kernel reduces to a rank-one operator. With $\pi = \pi_Q$, the corresponding rank-one LK Varga operator is thus the same as the R1-VI operator of Sharifi Kolarijani et al. [49]. We note that the R1-VI algorithm of Sharifi Kolarijani et al. [49] approximates the stationary distribution of the current policy-induced transition matrix by a single step of the power method in each iteration and is guaranteed to converge with at least the same geometric rate γ as VI.

The limiting-kernel approximation enjoys two advantages. First, let $\mathcal{F}: \mathbf{Q} \rightarrow \mathbf{Q}$ be any operator such that $\mathcal{F}Q$ admits a sample-based estimator, e.g., the Bellman residual $\mathcal{F}Q = \mathcal{T}^\pi Q - Q$. Then, $\mathcal{P}_\infty^\pi \mathcal{F}Q$ can also be estimated from rollouts. We elaborate on this in Section 3.3. Second, it is

compatible with n -step truncations. Given a policy π , the infinite sum in the preconditioner $\mathcal{C}_{\infty,\lambda}^\pi$ can be split into two parts as follows

$$\mathcal{C}_{\infty,\lambda}^\pi = \sum_{k=0}^{n-2} (\lambda\gamma\mathcal{P}^\pi)^k + (\lambda\gamma\mathcal{P}^\pi)^{n-1}(\mathcal{I} - \lambda\gamma\mathcal{P}^\pi)^{-1}.$$

Approximating $(\mathcal{I} - \lambda\gamma\mathcal{P}^\pi)^{-1}$ by $(\mathcal{I} - \lambda\gamma\mathcal{P}_\infty^\pi)^{-1}$, we derive the *limiting-kernel (LK) preconditioner* $\tilde{\mathcal{C}}_{n,\lambda}^\pi: \mathbf{Q} \rightarrow \mathbf{Q}$ given by

$$\tilde{\mathcal{C}}_{n,\lambda}^\pi \triangleq \mathcal{C}_{n,\lambda}^\pi + \frac{(\lambda\gamma)^n}{1 - \lambda\gamma} \mathcal{P}_\infty^\pi, \quad n \in \mathbb{N}, \lambda \in (0, 1]. \quad (11)$$

Observe that the effect of the limiting-kernel modification diminishes as $n \rightarrow \infty$. Indeed, $\tilde{\mathcal{C}}_{\infty,\lambda}^\pi = \mathcal{C}_{\infty,\lambda}^\pi$ for all $\lambda \in (0, 1]$. This modification can be interpreted as *tail completion*, since the powers of $\mathcal{P}^\pi$ in the tail omitted by $\mathcal{C}_{n,\lambda}^\pi$ are approximated by $\mathcal{P}_\infty^\pi$ in $\tilde{\mathcal{C}}_{n,\lambda}^\pi$, that is,

$$\tilde{\mathcal{C}}_{n,\lambda}^\pi = \sum_{k=0}^{n-1} (\lambda\gamma\mathcal{P}^\pi)^k + \sum_{l \geq n} (\lambda\gamma\mathcal{P}_\infty^\pi)^l. \quad (12)$$

If the policy π is such that $\mathcal{P}^\pi$ is aperiodic, then $(\mathcal{P}^\pi)^n$ approaches $\mathcal{P}_\infty^\pi$ as n increases, which makes the limiting-kernel approximation in (11) increasingly accurate for larger n . As a result, if n is large enough, the LK preconditioner $\tilde{\mathcal{C}}_{n,\lambda}^\pi$ is strictly closer to $\mathcal{C}_{\infty,\lambda}^\pi$ than the truncated preconditioner $\mathcal{C}_{n,\lambda}^\pi$ in terms of the operator ∞ -norm. The following lemma formalizes this statement.

Lemma 3.1 (Preconditioner tail completion). *Let $\pi \in \mathbf{\Pi}$ and suppose that $\mathcal{P}^\pi$ is aperiodic with mixing rate $\sigma_\pi \in (0, 1)$ and mixing time $n_\pi \in \mathbb{N}$. Define $\delta_n \triangleq \|\mathcal{C}_{n,\lambda}^\pi - \mathcal{C}_{\infty,\lambda}^\pi\|_\infty$ and $\tilde{\delta}_n \triangleq \|\tilde{\mathcal{C}}_{n,\lambda}^\pi - \mathcal{C}_{\infty,\lambda}^\pi\|_\infty$. Then, $\tilde{\delta}_n/\delta_n = \mathcal{O}(\sigma_\pi^n)$. In particular, $\tilde{\delta}_n < \delta_n$ for all $n \geq n_\pi$.*

If the chain induced by π is a unichain and $\lambda = 1$, the LK preconditioner (11) reduces to the preconditioner (gain matrix) introduced by Sharifi Kolarijani et al. [49, Rem. 3.4] for rank-one modified policy iteration (R1-MPI). In contrast, the LK formulation accommodates behavior chains with multiple recurrent classes and admits a scalable sample-based estimator, which we develop in Section 3.3.

3.2. Off-policy operators

Using the LK preconditioner introduced in (11), we now define the corresponding *limiting-kernel (LK) Harutyunyan operator* $\tilde{\mathcal{R}}_{n,\lambda}^{\mu,\pi}: \mathbf{Q} \rightarrow \mathbf{Q}$ for a behavior policy μ and a target policy π by

$$\tilde{\mathcal{R}}_{n,\lambda}^{\mu,\pi} Q \triangleq Q + \left(\mathcal{C}_{n,\lambda}^\mu + \frac{(\lambda\gamma)^n}{1 - \lambda\gamma} \mathcal{P}_\infty^\mu \right) (\mathcal{T}^\pi Q - Q), \quad \mu, \pi \in \mathbf{\Pi}, n \in \mathbb{N}, \lambda \in (0, 1].$$

This is the extension of the n -step Harutyunyan operator $\mathcal{R}_{n,\lambda}^{\mu,\pi}$ with LK preconditioning. For the rest of this section, we omit the on-policy operator and its policy evaluation analysis, as the off-policy operator reduces to the on-policy operator when $\mu = \pi$ (i.e., $\epsilon = 0$).

Intuitively, the n -step Harutyunyan operator in HI, through $\mathcal{C}_{n,\lambda}^\mu$, propagates the next $n - 1$ Bellman errors through the Markov chain induced by the behavior policy μ . The LK preconditioner, in turn, performs tail completion by approximating the omitted tail with the limiting kernel from step n onward, propagating the Bellman error according to the stationary structure of the recurrent classes

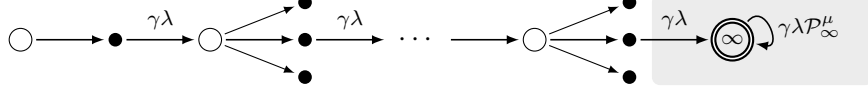


FIGURE 1. Backup diagram of the LK Harutyunyan operator. Hollow circles denote states, and filled circles denote actions. The first n steps follow the chain induced by μ . The shaded region depicts tail completion with the LK, and the omitted tail is replaced by stationary propagation from step n onward. Removing the LK term recovers the n -step Harutyunyan operator.

reachable from each state-action pair. Figure 1 illustrates the corresponding backup diagram of the LK Harutyunyan operator, depicting the resulting target value estimator.

We now introduce the corresponding iterations and analyze their convergence properties. To that end, let us define the LK iterations

$$\begin{aligned} \mathbf{LKHI}(\pi) : \quad Q_{\ell+1} &= \tilde{\mathcal{R}}_{n,\lambda}^{\mu,\pi} Q_\ell, \quad \ell \in \mathbb{N}_0, Q_0 \in \mathbf{Q}; & (\text{policy evaluation}) \\ \mathbf{LKHI} : \quad Q_{\ell+1} &= \tilde{\mathcal{R}}_{n,\lambda}^{\mu,\pi Q_\ell} Q_\ell, \quad \ell \in \mathbb{N}_0, Q_0 \in \mathbf{Q}. & (\text{control}) \end{aligned} \quad (13)$$

Our first result concerns the convergence of $\mathbf{LKHI}(\pi)$ for policy evaluation under a generic behavior policy μ . Recall that δ_n^μ denotes the n -step limiting-kernel residual for the policy μ , and α_e is the convergence rate of $\mathbf{HI}(\pi)$ established in Lemma 2.1.

Theorem 3.2 (Convergence of $\mathbf{LKHI}(\pi)$). *Consider the iterates of $\mathbf{LKHI}(\pi)$ in (13). We have $\|Q_\ell - Q^\pi\|_\infty \leq \tilde{\alpha}_e^\ell \|Q_0 - Q^\pi\|_\infty$ for all $\ell \in \mathbb{N}_0$, where $\tilde{\alpha}_e \triangleq \eta_e + (\gamma\lambda)^n \delta_n^\mu$. In particular, $\tilde{\alpha}_e < 1$ and hence $Q_\ell \rightarrow Q^\pi$ if*

$$\epsilon < \tilde{\epsilon}_{\max} \triangleq \frac{1-\gamma}{\gamma\lambda} - \frac{1-\gamma\lambda}{\gamma\lambda} (\gamma\lambda)^n \delta_n^\mu.$$

Moreover, we have $\tilde{\alpha}_e < \alpha_e$ if and only if $\delta_n^\mu < \frac{1-\gamma-\epsilon}{1-\gamma\lambda}$.

The preceding theorem characterizes the convergence rate $\tilde{\alpha}_e$ and the upper bound $\tilde{\epsilon}_{\max}$ on the policy discrepancy for $\mathbf{LKHI}(\pi)$. Unlike $\mathbf{HI}(\pi)$, whose convergence rate α_e can be smaller or larger than η_e , $\mathbf{LKHI}(\pi)$ has a convergence rate that is the same as or slower than that of the non-truncated operator, as $\tilde{\alpha}_e \geq \eta_e$. Similarly, it admits a range of policy discrepancies that is no wider than that of the non-truncated operator, as $\tilde{\epsilon}_{\max} \leq \frac{1-\gamma}{\gamma\lambda}$. However, compared to $\mathbf{HI}(\pi)$, which uses the truncated operator $\mathcal{C}_{n,\lambda}^\mu$, $\mathbf{LKHI}(\pi)$ can achieve a strictly smaller convergence rate $\tilde{\alpha}_e < \alpha_e$ in the regime $\epsilon \in [0, 1-\gamma)$, provided that the residual δ_n^μ is sufficiently small. Recall that, for policy evaluation, the non-truncated preconditioner $\mathcal{C}_{\infty,\lambda}^\mu$ yields a strictly smaller convergence rate than the truncated counterpart $\mathcal{C}_{n,\lambda}^\mu$ in the same regime for ϵ . Correspondingly, the LK preconditioner $\tilde{\mathcal{C}}_{n,\lambda}^\mu$, as an approximation of $\mathcal{C}_{\infty,\lambda}^\mu$, can only inherit this improvement over the same regime.

Note that the condition on the residual δ_n^μ ensuring $\tilde{\alpha}_e < \alpha_e$ may not hold for any $n \in \mathbb{N}$ for a generic behavior policy μ . However, for an *aperiodic* behavior transition operator $\mathcal{P}^\mu$, this is always the case for large enough n , given $\epsilon < 1-\gamma$. Indeed, by Lemma 3.1, for a large enough n , $\tilde{\mathcal{C}}_{n,\lambda}^\mu$ is a strictly better approximation of $\mathcal{C}_{\infty,\lambda}^\mu$ than $\mathcal{C}_{n,\lambda}^\mu$. The following corollary characterizes this.

Corollary 3.3 (Convergence of LKHI(π) with aperiodicity). *Consider the setup of Theorem 3.2. Assume that $\mathcal{P}^\mu$ is aperiodic with mixing rate $\sigma_\mu \in (0, 1)$ and mixing constant $c_\mu > 0$. Also, assume that $\epsilon < 1 - \gamma$. Then, $\tilde{\alpha}_e < \alpha_e$ if $n > \log(c_\mu(1 - \gamma\lambda)(1 - \gamma - \epsilon)^{-1}) / \log(\sigma_\mu^{-1})$.*

The preceding result indicates that the greatest improvement is for $\epsilon = 0$, that is, for the on-policy setting. The following remark discusses this in more detail.

Remark 3.4 (Convergence of on-policy LKHI(π) with aperiodicity). *Under the assumption of aperiodicity, the greatest improvement in the convergence rate of LKHI(π) over HI(π) is achieved in the on-policy setting where $\mu = \pi$ with $\epsilon = 0$. In particular, for $\epsilon = 0$ and $\lambda = 1$, LKHI(π) has a strictly smaller convergence rate than HI(π) for every $n \geq n_\pi$, with n_π being the mixing time of $\mathcal{P}^\pi$. This is expected in the on-policy setting, where tail completion approximates full policy evaluation of the target policy.*

Next, we analyze the convergence of LKHI in control.

Theorem 3.5 (Convergence of LKHI). *Consider the iterates of LKHI in (13). We have*

$$\|Q_\ell - Q^*\|_\infty \leq \tilde{\beta}_c^\ell \|Q_0 - Q^*\|_\infty$$

for all $\ell \in \mathbb{N}_0$, where $\tilde{\beta}_c \triangleq \eta_2$. In particular, $\tilde{\beta}_c < 1$ and hence $Q_\ell \rightarrow Q^*$ if $\lambda < \frac{1-\gamma}{2\gamma}$.

The preceding theorem shows that LKHI has the same convergence rate and converges under the same conditions as the non-truncated Harutyunyan iteration. In contrast to policy evaluation, the preceding bound does not establish a rate improvement for LKHI over HI. This stems from the discontinuous greedy map $Q \mapsto \pi_Q$, under which small changes in Q can perturb π_Q arbitrarily, leaving the policy discrepancy ϵ unconstrained across iterations. The analysis is therefore forced to accommodate the worst case $\epsilon = 2$, at which the LK and non-truncated preconditioners share the same operator norm and yield the common rate η_2 .

Beyond the convergence properties established above, the applicability of LKHI hinges on whether it admits a tractable sample-based estimator from rollouts of the behavior chain. In the following subsection, we construct such an estimator under a fixed behavior policy μ and establish almost-sure convergence of the resulting learning algorithm to Q^* .

3.3. Value estimation with a double backup

In the previous subsection, we showed that LKHI converges to optimal Q-values with the same rate as the non-truncated Harutyunyan iteration (that is, HQL) in control. Here, we develop the corresponding learning algorithm, which we call *Limiting-Kernel Q* ($\mathbf{LKQL}$). Our sample-based realization of the LKQL update depends on estimating an approximation of the limiting kernel $\mathcal{P}_\infty^\mu$ of the behavior policy $\mu \in \Pi$. To be precise, for $\tau \in [0, 1)$, we consider the approximation $\mathcal{A}_\tau^\mu$ of $\mathcal{P}_\infty^\mu$ defined by

$$\mathcal{A}_\tau^\mu \triangleq (1 - \tau)(\mathcal{I} - \tau\mathcal{P}^\mu)^{-1}.$$

Observe that $\lim_{\tau \rightarrow 1} \mathcal{A}_\tau^\mu = \mathcal{P}_\infty^\mu$. Equipped with $\mathcal{A}_\tau^\mu$, the estimation of the LK term reduces to a standard TD learning problem along a sampled rollout following policy μ , which we develop next.

To this end, we introduce an auxiliary function $U \in \mathbf{Q}$ tracking the long-run average TD error and defined by the fixed-point equation

$$U = (1 - \tau)(\mathcal{T}Q - Q) + \tau \mathcal{P}^\mu U. \quad (14)$$

Notice that the preceding equation is analogous to a Bellman equation with “reward” $(1 - \tau)(\mathcal{T}Q - Q)$ and “discount” τ under the transition operator $\mathcal{P}^\mu$. As $\tau \rightarrow 1$, U provides a more accurate approximation of $\mathcal{P}_\infty^\mu(\mathcal{T}Q - Q)$, matching the additional term introduced within the LK operator. A crucial observation is that the fixed point of (14) is unique and can be estimated by standard TD methods under the same sampling conditions as the original value estimation problem.

Let $\rho \in \Delta(\mathbf{S} \times \mathbf{A})$ be a state-action reset (i.e., initial) distribution. At each iteration $k \in \mathbb{N}_0$, given a sample rollout $(s_{k,0}, a_{k,0}, r_{k,0}, \dots, s_{k,n}, a_{k,n})$ with $(s_{k,0}, a_{k,0}) \sim \rho$ and $a_{k,i} \sim \mu(\cdot | s_{k,i})$ for $i \in \{1, \dots, n\}$, we define the TD errors by

$$\delta_{k,j}^Q \triangleq r_{k,j} + \gamma V(s_{k,j+1}) - Q(s_{k,j}, a_{k,j}), \quad j \in \{0, 1, \dots, n-1\}, \quad (15)$$

where $V(s) \triangleq \mathbb{E}_{a \sim \pi_Q}[Q(s, a)]$. The equation in (14) suggests the following TD error for U

$$\delta_{k,j}^U \triangleq (1 - \tau) \delta_{k,j}^Q + \tau U(s_{k,j+1}, a_{k,j+1}) - U(s_{k,j}, a_{k,j}), \quad j \in \{0, 1, \dots, n-1\}. \quad (16)$$

Intuitively, when U updates on a faster timescale than Q , it remains approximately equilibrated to its Q -dependent fixed point throughout learning [40]. Furthermore, the TD estimate of U admits a natural n -step and λ -return extension. However, we defer this extension to the empirical studies in Section 4.2 and focus on the one-step estimator in our analysis.

For reset distribution ρ , behavior policy μ , step sizes $\alpha_k, \beta_k > 0$, and TD errors $\{\delta_{k,j}^Q, \delta_{k,j}^U\}_{j=0}^{n-1}$ defined along a sampled rollout by (15) and (16), respectively, at each iteration $k \in \mathbb{N}_0$, we define the LKQL update at $(s, a) = (s_{k,0}, a_{k,0})$ by

$$\text{LKQL} : \begin{cases} Q_{k+1}(s, a) = Q_k(s, a) + \alpha_k \left(\sum_{j=0}^{n-1} (\gamma \lambda)^j \delta_{k,j}^Q + \frac{(\gamma \lambda)^n}{1 - \gamma \lambda} U_k(s, a) \right), \\ U_{k+1}(s, a) = U_k(s, a) + \beta_k \delta_{k,0}^U. \end{cases} \quad (17)$$

Despite the spectral structure of the LK preconditioner introduced in Section 3.1, the update rule in (17) updates Q at a single state-action pair per step, akin to Q -learning. This contrasts with prior value-learning schemes that estimate the full preconditioner [14] or a low-rank approximation [29, 49], both of which update Q at *all* state-action pairs simultaneously. As a result, LKQL readily extends to continuous state-action spaces, which we study in Section 4.2.

While Theorem 3.5 establishes operator-level convergence using the exact limiting kernel $\mathcal{P}_\infty^\mu$, the LKQL update (17) replaces that with the auxiliary iterate U , estimated via TD learning with discount factor τ . It then remains to verify that this approximation and the resulting update scheme preserve convergence to Q^* . In a restricted setting with a fixed behavior policy μ and i.i.d. initial state-action pairs, the following theorem establishes an almost-sure convergence guarantee.

Theorem 3.6 (Convergence of LKQL). *Consider the iterates $\{(Q_k, U_k)\}_{k \geq 0}$ of LKQL (17) in a finite MDP with parameters $n \in \mathbb{N}$, $\lambda \in (0, 1]$, $\tau \in [0, 1)$, initialization $Q_0 \in \mathbf{Q}$, $U_0 = 0 \in \mathbf{Q}$, and TD errors (15) and (16) computed at each iteration $k \in \mathbb{N}_0$ along a rollout generated from reset distribution $\rho \in \Delta(\mathbf{S} \times \mathbf{A})$ and behavior policy $\mu \in \mathbf{\Pi}$. Assume further that*

- (1) the step sizes $\{\alpha_k\}, \{\beta_k\}$ are positive and satisfy $\sum_k \alpha_k = \sum_k \beta_k = \infty$, $\sum_k (\alpha_k^2 + \beta_k^2) < \infty$, and $\alpha_k/\beta_k \rightarrow 0$;
- (2) the initial state-action pairs $z_k \triangleq (s_{k,0}, a_{k,0})$ are sampled from ρ independently of past samples, and $\rho(s, a) > 0$ for all (s, a) ;
- (3) $\lambda < \frac{1-\gamma}{2\gamma}$.

Then, $Q_k \rightarrow Q^*$ and $U_k \rightarrow 0$ almost surely.

Notably, the convergence guarantee holds for every fixed $\tau \in [0, 1)$. The *coverage* assumption on ρ guarantees that every state-action pair is visited infinitely often, the standard condition for asynchronous stochastic approximation [54, 6]. By contrast, the i.i.d. assumption on $(s_{k,0}, a_{k,0})$ is an analytical idealization. For instance, replay buffers in off-policy methods [17, 57] and parallel environments in multi-actor methods [51] reduce temporal dependence and are commonly modeled as i.i.d. sampling oracles in finite-time analyses. We use the latter setup, i.e., parallel sampling environments, in our experiments, which we discuss in Section 4. Related literature also accommodates Markovian sampling [4, 6, 26], showing that exact independence is not fundamental to this style of analysis.

Beyond these sampling assumptions, we follow prior off-policy value-learning analyses [34, 11, 10, 33, 14, 30, 31, 56] and assume the behavior policy is fixed throughout the iterations. Recent works relax this restriction by analyzing Q -dependent behavior policies [35, 32], a direction we leave to future work.

4. Evaluation

Section 3 established convergence guarantees for the LK operator and the LKQL algorithm in finite MDPs. We now assess how far these results carry from finite MDPs to those with large or continuous state spaces. In Section 4.1, we compare LKHI(π) against HI(π) in operator norm and LKQL against truncated HQL (THQL) in value error for control, where the relevant operator norms and Q^* can be computed exactly. In Section 4.2, we demonstrate that LKQL, used as the target for a parameterized value function, is effective on the MuJoCo continuous-control benchmark [53].

4.1. Finite MDP

Under the exact model, Theorem 3.2 predicts that tail completion can tighten the policy evaluation rate of LKHI(π) (13) relative to HI(π) (4). We evaluate the ∞ -norms of their error operators across 15 randomly generated Garnet MDPs with 50 states and 5 actions [2]. For each MDP, we compute the difference

$$\|\mathcal{I} + \mathcal{C}_{n,\lambda}^\mu(\gamma\mathcal{P}^\pi - \mathcal{I})\|_\infty - \|\mathcal{I} + \tilde{\mathcal{C}}_{n,\lambda}^\mu(\gamma\mathcal{P}^\pi - \mathcal{I})\|_\infty.$$

Figure 2 (left) shows the average of this difference across MDPs as ϵ and λ vary, with fixed rollout length $n = 5$. As shown, LKHI(π) attains the lower rate at small ϵ (near-on-policy) and large λ , the regime Theorem 3.2 identifies. The ranking reverses toward larger ϵ , where HI(π) is tighter, and the two operators nearly coincide at small λ .

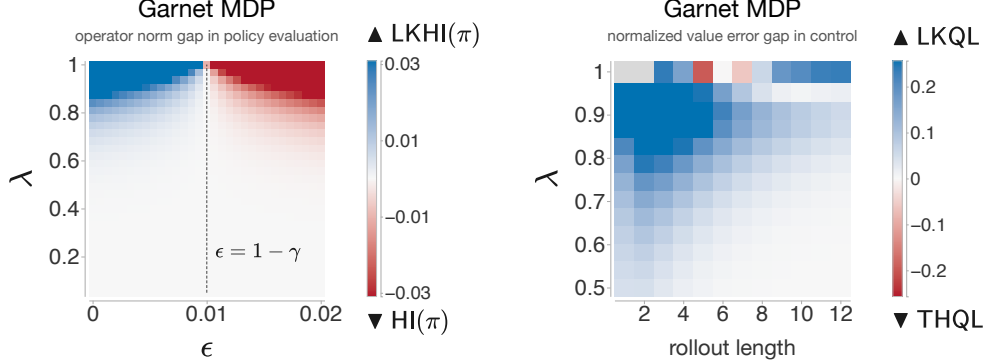


FIGURE 2. Garnet MDP results. Across both panels, blue marks the regime where the limiting-kernel method improves over its truncated counterpart. Left: Exact-model policy evaluation, showing the operator-norm difference between $\text{HI}(\pi)$ and $\text{LKHI}(\pi)$ over ϵ and λ , averaged over 15 random Garnet MDPs (Theorem 3.2). The dashed line $\epsilon = 1 - \gamma$ marks the theoretical boundary below which $\text{LKHI}(\pi)$ improves on $\text{HI}(\pi)$ for sufficiently large n . Right: Model-free control, showing the normalized value-error difference between truncated HQL (THQL) and LKQL after $T = 8000$ iterations from a common initialization Q_0 (averaged over 32 seeds), over rollout length n and λ (Theorem 3.6).

In learning from samples, Theorem 3.6 guarantees that LKQL (17) converges almost surely to Q^* under a fixed behavior policy and i.i.d. initial state-action pairs. Figure 2 (right) shows, over rollout length n and λ , the normalized value-error advantage of LKQL over THQL, given by

$$(\|Q_T - Q^*\|_\infty - \|\tilde{Q}_T - Q^*\|_\infty) / \|Q_0 - Q^*\|_\infty,$$

where Q_T and $\tilde{Q}_T$ are the final THQL and LKQL iterates, respectively, after $T = 8000$ iterations from the same initialization Q_0 and using the same step sizes α_k for all k . For $\lambda < 1$, LKQL attains the lower error at every rollout length. The gap is largest for large λ and short n and narrows as n grows, consistent with the diminishing contribution of the LK term. Only at $\lambda = 1$ is LKQL not uniformly better across rollout lengths, diverging at $n \in \{1, 2\}$ and yielding a higher error than THQL at $n \in \{5, 6, 7\}$.

4.2. Continuous control

We now extend LKQL to the continuous-control setting. As discussed in Section 3.3, many of the design choices in LKQL, including the choice of the limiting kernel for approximating the exact preconditioner and the TD estimator for the LK term, make such an extension possible.

To this end, we parameterize the value function, the LK term, and the policy and denote them by

$$V_\phi: \mathbf{S} \rightarrow \mathbb{R}, \quad U_\omega: \mathbf{S} \rightarrow \mathbb{R}, \quad \text{and} \quad \pi_\theta: \mathbf{S} \rightarrow \Delta(\mathbf{A}),$$

respectively, each modeled by a separate neural network. The parameters are then updated using the corresponding loss functions, which we explain in detail later in this section. The on- and off-policy estimators share the same sampling and parameter update scheme. Algorithm 1 summarizes the resulting procedure, which mirrors fitted Q-iteration [15, 44] and actor-critic frameworks [36, 1, 12].

Algorithm 1 Actor-critic implementation of LKQL

```

1: initialize:  $V_\phi$  (and  $P_\phi$ ),  $U_\omega$ , and  $\pi_\theta$ 
2: for iteration  $k = 0, 1, 2, \dots$  do
3:   sample: rollouts from  $N$  parallel actors to form  $\mathbf{D}$ 
4:   compute: targets  $G_V$  (or  $G_Q$ ) and  $G_U$  on  $\mathbf{D}$ 
5:   for iteration  $i = 1, \dots, K$  do
6:     draw: minibatch  $\mathbf{B}$  from  $\mathbf{D}$ 
7:     update:  $\phi$ ,  $\omega$ , and  $\theta$  on  $\mathbf{B}$ 
8:   end for
9: end for

```

That is, each outer iteration (over k) collects a single rollout from each of the N parallel instances of the behavior policy, which we refer to interchangeably as “actors” (the behavior policy coincides with the target policy in on-policy learning). An inner loop (over i) then performs K minibatch updates on the resulting batch. Parenthesized terms in Algorithm 1 apply to the off-policy estimator, where “and” indicates additions and “or” indicates substitutions.

On-policy estimator. Here, we use the actor-critic model of PPO [48] as the backbone and focus on the value (critic) updates. For the policy updates, we retain PPO’s clipped surrogate objective. The LKQL update (17) then translates to the loss functions

$$L(\phi) \triangleq \mathbb{E}_{s \sim \mathbf{D}} [(G_V(s) - V_\phi(s))^2], \quad \text{and} \quad L(\omega) \triangleq \mathbb{E}_{s \sim \mathbf{D}} [(G_U(s) - U_\omega(s))^2], \quad (18)$$

where $\mathbf{D}$ is the minibatch distribution of states sampled from the rollouts, and G_V and G_U are the respective target functions. Crucially, the target functions are computed once before the inner loop and kept fixed throughout the parameter updates in the inner loop, hence reflecting the fitted Q-iteration scheme.

For notational simplicity, the definitions below assume that the truncation horizon n equals the rollout length, although the two are separate hyperparameters in general (see Appendix B). At iteration k , for a rollout generated by an actor and starting from state $s_{k,0}$, we define the state-value estimate (i.e., the return, or target value) by

$$G_V(s_{k,t}) \triangleq \sum_{j=0}^{n-t-1} (\gamma\lambda)^j \delta_{k,t+j}^V + \frac{(\gamma\lambda)^{n-t}}{1-\gamma\lambda} U_\omega(s_{k,t}) + V_\phi(s_{k,t}), \quad t \in \{0, 1, \dots, n-1\}, \quad (19)$$

and the corresponding V-TD errors by

$$\delta_{k,t+j}^V \triangleq r_{k,t+j} + \gamma V_\phi(s_{k,t+j+1}) - V_\phi(s_{k,t+j}), \quad j \in \{0, 1, \dots, n-t-1\}. \quad (20)$$

Similarly, with a separate trace coefficient $\lambda_U \in (0, 1]$ for the LK term, we define the target for U_ω by

$$G_U(s_{k,t}) \triangleq \sum_{j=0}^{n-t-1} (\tau\lambda_U)^j \delta_{k,t+j}^U + U_\omega(s_{k,t}), \quad t \in \{0, 1, \dots, n-1\},$$

and the corresponding U-TD errors, computed using the V-TD errors (20), by

$$\delta_{k,t+j}^U \triangleq (1-\tau) \delta_{k,t+j}^V + \tau U_\omega(s_{k,t+j+1}) - U_\omega(s_{k,t+j}), \quad j \in \{0, 1, \dots, n-t-1\}. \quad (21)$$

For the policy loss $L(\theta)$, we follow the clipped surrogate objective of PPO with an entropy regularization term, using the advantage term $A(s_{k,t}, a_{k,t}) \triangleq G_V(s_{k,t}) - V_\phi(s_{k,t})$. Combining the policy loss with the loss functions in (18), we obtain the final loss function

$$L(\phi, \omega, \theta) \triangleq L(\phi) + \beta_U L(\omega) + \beta_\pi L(\theta), \quad (22)$$

where β_U and β_π are the respective loss coefficients. Using π_θ for the sample step and the loss function in (22) for the update step, Algorithm 1 summarizes the resulting on-policy LKQL extension of PPO.

Off-policy estimator. We now turn to the off-policy LKQL estimator, in which we borrow the Q -value function architecture of the Normalized Advantage Function (NAF) [21]. Specifically, we represent Q -values as the sum of the state value V_ϕ and a quadratic advantage centered at the deterministic target policy $\bar{\pi}_\theta: \mathbf{S} \rightarrow \mathbf{A}$, that is,

$$Q_{\theta, \phi}(s, a) \triangleq V_\phi(s) - \frac{1}{2}(a - \bar{\pi}_\theta(s))^\top P_\phi(s)(a - \bar{\pi}_\theta(s)),$$

where $P_\phi(s)$ is a positive-definite matrix. In this representation, with $\pi_\theta(\cdot|s)$ defined as the Dirac distribution at $\bar{\pi}_\theta(s)$, the relation $V_\phi(s) = \mathbb{E}_{a \sim \pi_\theta}[Q_{\theta, \phi}(s, a)]$ holds explicitly for any θ and $s \in \mathbf{S}$, hence rendering the value learning scheme off-policy.

Due to the inherent relation $\bar{\pi}_\theta(s) = \arg \max_a Q_{\theta, \phi}(s, a)$, the policy improvement step is implicit, and we define only the Q -value loss for updating the parameters ϕ and θ together by

$$L(\phi, \theta) \triangleq \mathbb{E}_{(s, a) \sim \mathbf{D}}[(G_Q(s, a) - Q_{\theta, \phi}(s, a))^2].$$

Above, $\mathbf{D}$ is again the minibatch distribution, and G_Q is the return, given by

$$G_Q(s_{k,t}, a_{k,t}) \triangleq \sum_{j=0}^{n-t-1} (\gamma\lambda)^j \delta_{k,t+j}^Q + \frac{(\gamma\lambda)^{n-t}}{1-\gamma\lambda} U_\omega(s_{k,t}) + Q_{\theta, \phi}(s_{k,t}, a_{k,t}), \quad t \in \{0, 1, \dots, n-1\},$$

with Q -TD errors

$$\delta_{k,t+j}^Q \triangleq r_{k,t+j} + \gamma V_\phi(s_{k,t+j+1}) - Q_{\theta, \phi}(s_{k,t+j}, a_{k,t+j}), \quad j \in \{0, \dots, n-t-1\}. \quad (23)$$

Although the value update is off-policy, the update for the LK term remains on-policy under the behavior policy μ used to define $\mathcal{A}_t^\mu$. We keep U_ω as a function of state. The loss function $L(\omega)$, the target G_U , and the U -TD errors for the LK term are the same as in the on-policy estimator, except that the Q -TD errors (23) replace the V -TD errors in computing the U -TD errors (21). The final loss function is then given by

$$L(\phi, \omega, \theta) \triangleq L(\phi, \theta) + \beta_U L(\omega), \quad (24)$$

where β_U is the loss coefficient for the LK term. The off-policy loss (24) drives the update step of Algorithm 1, while the sample step draws actions as $a = \bar{\pi}_\theta(s) + P_\phi(s)^{-1/2} w$ for $w \sim \mathcal{N}(0, I)$.

Experiments. LKQL extends two standard λ -return baselines, namely, GAE [47] for on-policy learning and THQL [22] for off-policy learning. Both are recovered from Algorithm 1 by setting $U_\omega \equiv 0$. Specifically, the baselines use the PPO backbone in the on-policy case and the NAF backbone in the off-policy case. We benchmark LKQL against these baselines on Hopper, Walker2d, and HalfCheetah from the MuJoCo continuous-control benchmark [53]. We train each method for $T = 500$ outer iterations (see Algorithm 1) with $N = 16$ parallel actors and several rollout lengths.

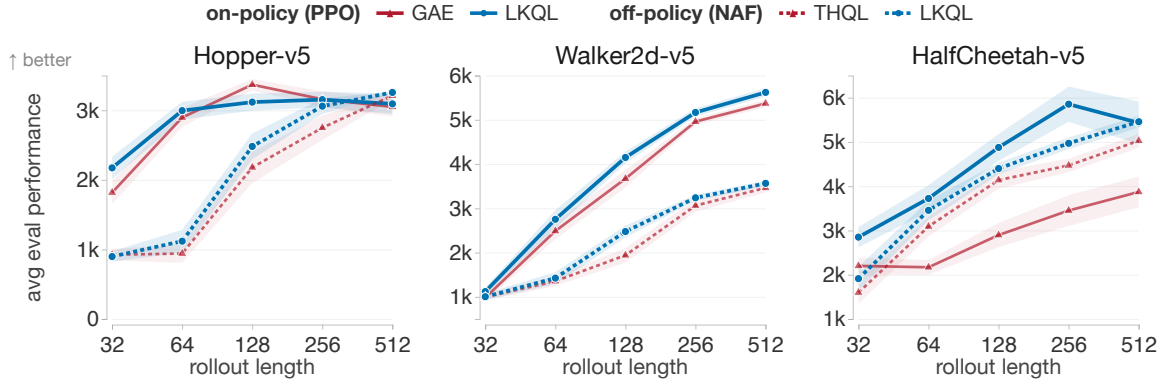


FIGURE 3. MuJoCo continuous-control results. Final evaluation performance (mean $\pm$ s.e.m. over 32 seeds) versus rollout length, comparing LKQL (blue) against its baseline value estimator (red) within each actor-critic backbone: GAE in PPO (on-policy, solid) and truncated HQL (THQL) in NAF (off-policy, dashed).

Figure 3 reports the final evaluation performance (cumulative episodic reward), averaged over $S = 32$ seeds (runs). Full hyperparameters and architectures are detailed in Appendix B. The baselines are tuned over their respective hyperparameter sets, and LKQL inherits the resulting best settings while tuning only its own hyperparameters, namely, the LK trace coefficient λ_U , the LK-term discount factor τ , the n -step truncation horizon, and the optimizer parameters for ω .

Figure 3 shows that backups with LK tail completion achieve higher returns than the baselines across most tested rollout lengths on Hopper, Walker2d, and HalfCheetah in both on- and off-policy learning. The largest improvement occurs on HalfCheetah. Unlike Hopper and Walker2d, which terminate early when the simulated body falls, HalfCheetah has no termination condition. Its episodes run the full length, giving it the longest effective horizon. This longer effective horizon in turn increases the relevance of the long-run behavior modeled by the LK term, which we argue accounts for the larger improvement here. On this task, the improvement over the baseline is also more pronounced in the on-policy setting than in the off-policy setting, consistent with the acceleration being greatest near the on-policy regime. See Remark 3.4 and Figure 2.

5. Conclusion

λ -return algorithms are often realized via n -step truncation [36, 47, 24, 16], which discards the tail of the non-truncated preconditioner and thereby yields a short-horizon update. The LK Harutyunyan operator augments the n -step truncation by approximating the omitted tail using the limiting kernel (LK) of the behavior transition operator. We showed that, for an aperiodic behavior chain and in the near-on-policy regime, the proposed operator improves the policy evaluation convergence rate over the truncated HQL (THQL) operator for sufficiently large n .

LKQL implements this operator through the double backup. In practice, each backup maintains a separate value function, one for the critic and one for the LK term, allowing straightforward integration into existing actor-critic frameworks. Together, the two backups bridge long and short horizons

within a single value estimator. We analyzed this estimator on two fronts, convergence and empirical performance. We proved in Section 3.3 that it converges almost surely to the optimal value function Q^* in finite MDPs under a fixed behavior policy and i.i.d. initial state-action pairs. Empirically, in Section 4.2, we showed that, on the MuJoCo continuous-control benchmark, it outperforms GAE in the on-policy setting and THQL in the off-policy setting across most rollout lengths.

The LK extension of the truncated λ -return operators applies beyond THQL and GAE, our off-policy and on-policy baselines, respectively. For instance, setting $\mu = \pi$, the HQL operator can be rewritten as

$$\mathcal{R}_{\infty,\lambda}^{\pi,\pi}Q = \mathcal{C}_{\infty,\lambda}^{\pi}(r + (1 - \lambda)\gamma\mathcal{P}^{\pi}Q). \quad (25)$$

However, replacing $\mathcal{C}_{\infty,\lambda}^{\pi}$ with $\mathcal{C}_{\infty,\lambda}^{\mu}$ in (25) does not recover the HQL operator $\mathcal{R}_{\infty,\lambda}^{\mu,\pi}$. Instead, the operator defined by $\mathcal{S}_{\infty,\lambda}^{\mu,\pi}Q \triangleq \mathcal{C}_{\infty,\lambda}^{\mu}(r + (1 - \lambda)\gamma\mathcal{P}^{\pi}Q)$ recovers *Peng & Williams' $Q(\lambda)$* (PQL) [39, 27]. In a similar fashion, one can define the *n-step Peng & Williams operator* $\mathcal{S}_{n,\lambda}^{\mu,\pi} : \mathbf{Q} \rightarrow \mathbf{Q}$ by

$$\mathcal{S}_{n,\lambda}^{\mu,\pi}Q \triangleq Q + \mathcal{C}_{n,\lambda}^{\mu}(\mathcal{T}^{\lambda\mu+(1-\lambda)\pi}Q - Q), \quad \mu, \pi \in \mathbf{\Pi}, n \in \mathbb{N}, \lambda \in (0, 1].$$

We note that, by changing the target policy, the limiting-kernel approximation readily extends to the truncated PQL operator $\mathcal{S}_{n,\lambda}^{\mu,\pi}$. Beyond PQL, conservative off-policy operators such as Retrace [37], with $c(s, a) = \min\{1, \pi(a|s)/\mu(a|s)\}$, and Tree-Backup [41], with $c(s, a) = \pi(a|s)$, give rise to corrected operators $\mathcal{P}^{c,\mu}$. Extending the double-backup scheme to these operators would require replacing $\mathcal{P}^{\mu}$ in the preconditioner with $\mathcal{P}^{c,\mu}$ and deriving a suitable tail approximation together with the corresponding auxiliary estimators. This nontrivial extension is left for future work.

Besides these potential extensions, we acknowledge several limitations of LKQL. One such limitation is that our convergence analysis of LKQL in control assumes a fixed behavior policy. We believe this assumption can be relaxed using the techniques developed by Meyn [35] and Liu et al. [32]. On the practical side, LKQL introduces a second critic for the LK term, modeled as a separate network with the same hidden sizes as the critic. This increases both the memory footprint and the computational cost of the critic updates. Additionally, estimating the LK term introduces the hyperparameters τ and λ_U , along with those of the optimizer used for its critic.

Appendix A. Technical Proofs

In this section, we provide the proofs for the main results. We begin by collecting notation and identities shared across all proofs.

A.1. Notation and Preliminaries

Following Harutyunyan et al. [22], for a behavior policy μ and a target policy π , we define the decomposition

$$\gamma\mathcal{P}^{\pi} - \mathcal{I} = \mathcal{X} + \mathcal{W}, \quad (26a)$$

where

$$\mathcal{X} \triangleq \gamma\lambda\mathcal{P}^{\mu} - \mathcal{I}, \quad \mathcal{W} \triangleq \gamma\lambda(\mathcal{P}^{\pi} - \mathcal{P}^{\mu}) + \gamma(1 - \lambda)\mathcal{P}^{\pi}. \quad (26b)$$

Above, the operator $\mathcal{X}$ captures the component aligned with the behavior chain (and hence the part of the Bellman residual absorbed by the preconditioner via the telescoping identity, as we shall see), while $\mathcal{W}$ captures the off-policy mismatch and the λ -discounting. We further define the scalar

$$\xi \triangleq \frac{(\gamma\lambda)^n}{1-\gamma\lambda}.$$

Note that ξ is the ∞ -norm of the LK tail completion operator $\frac{(\gamma\lambda)^n}{1-\gamma\lambda}\mathcal{P}_\infty^\mu$ in (11).

The following identities are used repeatedly and referenced by name in subsequent proofs:

$$\mathcal{I} + \mathcal{C}_{n,\lambda}^\mu \mathcal{X} = (\gamma\lambda\mathcal{P}^\mu)^n, \quad (\text{telescoping}) \quad (27a)$$

$$\mathcal{P}_\infty^\mu \mathcal{P}^\mu = \mathcal{P}^\mu \mathcal{P}_\infty^\mu = \mathcal{P}_\infty^\mu, \quad (\text{absorption}) \quad (27b)$$

$$\mathcal{P}^\pi \mathbf{1} = \mathcal{P}^\mu \mathbf{1} = \mathcal{P}_\infty^\mu \mathbf{1} = \mathbf{1}. \quad (\text{row-sum}) \quad (27c)$$

We also note that, for any non-negative finite operator $\mathcal{A}$, we have (see Puterman [42, §C.1])

$$\|\mathcal{A}\|_\infty = \max_i \sum_j \mathcal{A}_{ij} = \|\mathcal{A}\mathbf{1}\|_\infty. \quad (28)$$

In particular, using the row-sum identity (27c), we have

$$\|\mathcal{P}^\pi\|_\infty = \|\mathcal{P}^\mu\|_\infty = \|\mathcal{P}_\infty^\mu\|_\infty = 1.$$

Similarly, for the preconditioner $\mathcal{C}_{n,\lambda}^\mu$, using (28) and the row-sum identity (27c), we obtain

$$\|\mathcal{C}_{n,\lambda}^\mu\|_\infty = \left\| \sum_{k=0}^{n-1} (\lambda\gamma\mathcal{P}^\mu)^k \right\|_\infty = \sum_{k=0}^{n-1} (\lambda\gamma)^k = \frac{1-(\gamma\lambda)^n}{1-\gamma\lambda}. \quad (29)$$

Also, by defining $\mathcal{C}_{0,\lambda}^\mu \triangleq 0$, the construction of the preconditioner $\mathcal{C}_{n,\lambda}^\mu$ satisfies the recursion

$$\mathcal{C}_{n+1,\lambda}^\mu = \mathcal{I} + \gamma\lambda\mathcal{P}^\mu \mathcal{C}_{n,\lambda}^\mu, \quad \forall n \in \mathbb{N}_0. \quad (30)$$

Finally, recall that $\epsilon \triangleq \max_{s \in \mathbf{S}} \|\mu(\cdot|s) - \pi(\cdot|s)\|_1$ denotes the discrepancy between behavior and target policies. We have (see the proof of Harutyunyan et al. [22, Lem. 1])

$$\|\mathcal{P}^\pi - \mathcal{P}^\mu\|_\infty \leq \epsilon.$$

Using the preceding inequality and the definition of $\mathcal{W}$ in (26b), we can also write

$$\|\mathcal{W}\|_\infty \leq \gamma\lambda\epsilon + \gamma(1-\lambda) = (1-\gamma\lambda)\eta_\epsilon. \quad (31)$$

A.2. Proof of Lemma 2.1

Policy evaluation. Let $E_\ell \triangleq Q_\ell - Q^\pi$ be the value error of the ℓ -th iteration of HI(π). Since $\mathcal{T}^\pi Q = \gamma\mathcal{P}^\pi Q + r$ and $\mathcal{T}^\pi Q^\pi = Q^\pi$, the Bellman residual factors as

$$\mathcal{T}^\pi Q - Q = (\mathcal{T}^\pi Q - \mathcal{T}^\pi Q^\pi) - (Q - Q^\pi) = (\gamma\mathcal{P}^\pi - \mathcal{I})(Q - Q^\pi).$$

Then, using the definition of $\mathcal{R}_{n,\lambda}^{\mu,\pi}$, we can write $E_{\ell+1} = \mathcal{M}_n E_\ell$, where $\mathcal{M}_n \triangleq \mathcal{I} + \mathcal{C}_{n,\lambda}^\mu (\gamma\mathcal{P}^\pi - \mathcal{I})$ is the error operator of HI(π). Below, we show that $m_n \triangleq \|\mathcal{M}_n\|_\infty \leq \alpha_e$, from which $\|E_\ell\|_\infty \leq \alpha_e^\ell \|E_0\|_\infty$ follows.

First, using the decomposition (26) and applying the telescoping identity (27a), the error operator can be equivalently characterized as $\mathcal{M}_n = (\gamma\lambda\mathcal{P}^\mu)^n + \mathcal{C}_{n,\lambda}^\mu\mathcal{W}$. Then, the recursion (30) for the preconditioner implies that $\mathcal{M}_n$ satisfies

$$\mathcal{M}_{n+1} = \mathcal{W} + \gamma\lambda\mathcal{P}^\mu\mathcal{M}_n, \quad \forall n \in \mathbb{N}.$$

Taking norms in the preceding recursion and using (31) then gives

$$m_{n+1} \leq \|\mathcal{W}\|_\infty + \gamma\lambda m_n \leq (1 - \gamma\lambda)\eta_\epsilon + \gamma\lambda m_n, \quad \forall n \in \mathbb{N}. \quad (32)$$

For $n = 1$, we have

$$m_1 = \|\mathcal{M}_1\|_\infty = \|\mathcal{W} + \gamma\lambda\mathcal{P}^\mu\|_\infty = \|\gamma\mathcal{P}^\pi\|_\infty = \gamma = \alpha_\epsilon.$$

For $n \geq 2$, iterating (32) also yields

$$m_n \leq (\gamma\lambda)^{n-1}m_1 + (1 - \gamma\lambda)\eta_\epsilon \sum_{j=0}^{n-2} (\gamma\lambda)^j = (\gamma\lambda)^{n-1}(\gamma - \eta_\epsilon) + \eta_\epsilon = \alpha_\epsilon.$$

This completes the first part of the proof. For $n \geq 2$, we now consider the condition $\alpha_\epsilon < 1$ to derive the bound on ϵ . Starting from the definition of α_ϵ and isolating η_ϵ yields

$$\alpha_\epsilon = (\gamma\lambda)^{n-1}(\gamma - \eta_\epsilon) + \eta_\epsilon < 1 \iff \eta_\epsilon(1 - (\gamma\lambda)^{n-1}) < 1 - \gamma(\gamma\lambda)^{n-1}.$$

Substituting η_ϵ then gives

$$\gamma(1 + \lambda(\epsilon - 1))(1 - (\gamma\lambda)^{n-1}) < (1 - \gamma\lambda)(1 - \gamma(\gamma\lambda)^{n-1}).$$

Rearranging terms yields $\epsilon < \epsilon_{\max}$ as claimed.

Control. Now, let $E_\ell \triangleq Q_\ell - Q^*$ be the value error of the ℓ -th iteration of HI for the control problem. Below, we show that $\|E_{\ell+1}\|_\infty \leq \beta_c \|E_\ell\|_\infty$, from which the first claim in the statement of the lemma follows. Recall that π_Q is greedy with respect to Q , and hence $\mathcal{T}^{\pi_Q}Q = \mathcal{T}Q$. Using $\mathcal{T}Q^* = Q^*$ and the recursion (30) for the preconditioner (with $\mathcal{C}_{0,\lambda}^\mu \triangleq 0$), the value error can be written as

$$\begin{aligned} E_{\ell+1} &= \mathcal{R}_{n,\lambda}^{\mu,\pi_Q}Q_\ell - Q^* = Q_\ell + \mathcal{C}_{n,\lambda}^\mu(\mathcal{T}Q_\ell - Q_\ell) - Q^* \\ &= \mathcal{C}_{n,\lambda}^\mu((\mathcal{T}Q_\ell - \mathcal{T}Q^*) - (Q_\ell - Q^*)) + (Q_\ell - Q^*) \\ &= \mathcal{C}_{n,\lambda}^\mu(\mathcal{T}Q_\ell - \mathcal{T}Q^*) - \gamma\lambda\mathcal{P}^\mu\mathcal{C}_{n-1,\lambda}^\mu E_\ell. \end{aligned}$$

Taking norms and using the fact that $\mathcal{T}$ is a γ -contraction, we have

$$\|E_{\ell+1}\|_\infty \leq \left(\gamma \|\mathcal{C}_{n,\lambda}^\mu\|_\infty + \gamma\lambda \|\mathcal{P}^\mu\|_\infty \|\mathcal{C}_{n-1,\lambda}^\mu\|_\infty \right) \|E_\ell\|_\infty.$$

Finally, using the bound (29) (which also holds for $n = 0$ since $\mathcal{C}_{0,\lambda}^\mu \triangleq 0$ by our definition), we obtain

$$\|E_{\ell+1}\|_\infty \leq \frac{\gamma(1 - (\gamma\lambda)^n) + \gamma\lambda - (\gamma\lambda)^n}{1 - \gamma\lambda} \|E_\ell\|_\infty = \left[\eta_2 - \frac{(1 + \gamma)(\gamma\lambda)^n}{1 - \gamma\lambda} \right] \|E_\ell\|_\infty = \beta_c \|E_\ell\|_\infty.$$

We next consider the condition $\beta_c < 1$ to derive the corresponding inequality for λ . The first characterization of β_c above yields

$$\beta_c < 1 \iff \gamma(1 + \lambda) - (1 + \gamma)(\gamma\lambda)^n < 1 - \gamma\lambda.$$

Rearranging then gives the implicit inequality

$$\lambda < \frac{1-\gamma}{2\gamma} + \frac{(1+\gamma)(\gamma\lambda)^n}{2\gamma}.$$

This completes the proof.

A.3. Proof of Lemma 3.1

Let $\Delta_k^\pi \triangleq (\mathcal{P}^\pi)^k - \mathcal{P}_\infty^\pi$ for $k \in \mathbb{N}_0$, so that the finite-horizon limiting-kernel residual is given by $\delta_k^\pi = \|\Delta_k^\pi\|_\infty$. The residuals from truncation and tail completion are given by

$$\begin{aligned}\delta_n &\triangleq \|\mathcal{C}_{\infty,\lambda}^\pi - \mathcal{C}_{n,\lambda}^\pi\|_\infty = \left\| \sum_{k \geq n} (\gamma\lambda)^k (\mathcal{P}^\pi)^k \right\|_\infty, \\ \tilde{\delta}_n &\triangleq \|\mathcal{C}_{\infty,\lambda}^\pi - \tilde{\mathcal{C}}_{n,\lambda}^\pi\|_\infty = \left\| \sum_{k \geq n} (\gamma\lambda)^k \Delta_k^\pi \right\|_\infty,\end{aligned}$$

where the second equality follows from the representation (12). Now, observe that, by the non-negative operator norm identity (28) and the row-sum identity (27c), we obtain

$$\delta_n = \left\| \sum_{k \geq n} (\gamma\lambda)^k (\mathcal{P}^\pi)^k \mathbf{1} \right\|_\infty = \left\| \sum_{k \geq n} (\gamma\lambda)^k \mathbf{1} \right\|_\infty = \sum_{k \geq n} (\gamma\lambda)^k = \frac{(\gamma\lambda)^n}{1-\gamma\lambda}.$$

For $\tilde{\delta}_n$, the triangle inequality and the aperiodicity bound (8) give

$$\tilde{\delta}_n \leq \sum_{k \geq n} (\gamma\lambda)^k \delta_k^\pi \leq c_\pi \sum_{k \geq n} (\gamma\lambda\sigma_\pi)^k = \frac{c_\pi (\gamma\lambda\sigma_\pi)^n}{1-\gamma\lambda\sigma_\pi}.$$

Dividing the bound on $\tilde{\delta}_n$ by the expression for δ_n , we obtain

$$\frac{\tilde{\delta}_n}{\delta_n} \leq \frac{c_\pi(1-\gamma\lambda)}{1-\gamma\lambda\sigma_\pi} \cdot \sigma_\pi^n = \mathcal{O}(\sigma_\pi^n).$$

It remains to show the strict inequality for $n \geq n_\pi$. The absorption identity (27b) gives $\Delta_{k+1}^\pi = \Delta_k^\pi \mathcal{P}^\pi$, and hence

$$\delta_{k+1}^\pi \leq \delta_k^\pi \|\mathcal{P}^\pi\|_\infty = \delta_k^\pi.$$

That is, δ_k^π is non-increasing in k . By the definition of n_π in (9), the residual δ_k^π drops below one at $k = n_\pi$, and by monotonicity it stays below one for every $k \geq n_\pi$. Substituting this strict bound gives

$$\tilde{\delta}_n \leq \sum_{k \geq n} (\gamma\lambda)^k \delta_k^\pi < \sum_{k \geq n} (\gamma\lambda)^k = \delta_n, \quad \forall n \geq n_\pi.$$

A.4. Proof of Theorem 3.2

Let $E_\ell \triangleq Q_\ell - Q^\pi$ be the value error of the ℓ -th iteration of LKHI(π). Using the same argument as in the policy evaluation part of the proof of Lemma 2.1, we obtain $E_{\ell+1} = \tilde{\mathcal{M}}_n E_\ell$, where $\tilde{\mathcal{M}}_n \triangleq \mathcal{I} + \tilde{\mathcal{C}}_{n,\lambda}^\mu (\gamma\mathcal{P}^\pi - \mathcal{I})$ is the error operator. Below, we show that $\|\tilde{\mathcal{M}}_n\|_\infty \leq \tilde{\alpha}_e$, from which the first claim in the statement of the theorem follows.

Combining the identity $\tilde{\mathcal{C}}_{n,\lambda}^\mu = \mathcal{C}_{n,\lambda}^\mu + \xi\mathcal{P}_\infty^\mu$ in (11) with the decomposition (26), we can expand the error operator into four terms as follows

$$\tilde{\mathcal{M}}_n = (\mathcal{I} + \mathcal{C}_{n,\lambda}^\mu \mathcal{X}) + \mathcal{C}_{n,\lambda}^\mu \mathcal{W} + \xi\mathcal{P}_\infty^\mu \mathcal{X} + \xi\mathcal{P}_\infty^\mu \mathcal{W}. \quad (33)$$

Now, observe that the telescoping identity (27a) collapses the first term to $(\gamma\lambda\mathcal{P}^\mu)^n$, while the absorption identity (27b) reduces the third term to $-(\gamma\lambda)^n\mathcal{P}_\infty^\mu$. Summing these two terms exposes the mixing residual $\Delta_n^\mu \triangleq (\mathcal{P}^\mu)^n - \mathcal{P}_\infty^\mu$, and (33) reduces to

$$\widetilde{\mathcal{M}}_n = (\gamma\lambda)^n \Delta_n^\mu + \mathcal{C}_{n,\lambda}^\mu \mathcal{W} + \xi \mathcal{P}_\infty^\mu \mathcal{W}. \quad (34)$$

Applying the triangle inequality and using $\delta_n^\mu = \|\Delta_n^\mu\|_\infty$ by definition and the bounds (29) and (31), we obtain

$$\begin{aligned} \|\widetilde{\mathcal{M}}_n\|_\infty &\leq (\gamma\lambda)^n \delta_n^\mu + \frac{1 - (\gamma\lambda)^n}{1 - \gamma\lambda} \cdot (1 - \gamma\lambda)\eta_\epsilon + \frac{(\gamma\lambda)^n}{1 - \gamma\lambda} \cdot (1 - \gamma\lambda)\eta_\epsilon \\ &= (\gamma\lambda)^n \delta_n^\mu + \eta_\epsilon = \widetilde{\alpha}_e. \end{aligned}$$

For convergence, we require $\widetilde{\alpha}_e < 1$, which is equivalent to $\eta_\epsilon < 1 - (\gamma\lambda)^n \delta_n^\mu$. Substituting η_ϵ , we arrive at

$$\gamma(1 - \lambda) + \gamma\lambda\epsilon < (1 - \gamma\lambda)(1 - (\gamma\lambda)^n \delta_n^\mu),$$

which rearranges to $\epsilon < \widetilde{\epsilon}_{\max}$ as claimed.

Finally, comparing against the HI(π) rate α_e from Lemma 2.1, we have

$$\alpha_e - \widetilde{\alpha}_e = (\gamma\lambda)^{n-1}(\gamma - \eta_\epsilon) - (\gamma\lambda)^n \delta_n^\mu = (\gamma\lambda)^{n-1}((\gamma - \eta_\epsilon) - \gamma\lambda\delta_n^\mu).$$

Using the identity $\gamma - \eta_\epsilon = \gamma\lambda(1 - \gamma - \epsilon)/(1 - \gamma\lambda)$, we obtain

$$\widetilde{\alpha}_e < \alpha_e \iff \delta_n^\mu < \frac{1 - \gamma - \epsilon}{1 - \gamma\lambda}.$$

This completes the proof.

A.5. Proof of Corollary 3.3

The result follows from Theorem 3.2 and the aperiodicity bound (8). Indeed, if

$$n > \frac{\log\left(\frac{c_\mu(1-\gamma\lambda)}{1-\gamma-\epsilon}\right)}{\log(1/\sigma_\mu)} = \frac{\log\left(\frac{1-\gamma-\epsilon}{c_\mu(1-\gamma\lambda)}\right)}{\log(\sigma_\mu)},$$

then

$$\log\left(\frac{1 - \gamma - \epsilon}{c_\mu(1 - \gamma\lambda)}\right) > n \log(\sigma_\mu),$$

and we have

$$\delta_n^\mu \leq c_\mu \sigma_\mu^n < \frac{1 - \gamma - \epsilon}{1 - \gamma\lambda}.$$

A.6. Proof of Theorem 3.5

Let $E_\ell \triangleq Q_\ell - Q^*$ be the value error of the ℓ -th iteration of LKHI for the control problem. Below, we show that $\|E_{\ell+1}\|_\infty \leq \eta_2 \|E_\ell\|_\infty$, from which the first claim in the statement of the theorem

follows. Recall that π_Q is greedy with respect to Q , and hence $\mathcal{T}^{\pi_Q} Q = \mathcal{T} Q$. Then, using the fact that $\mathcal{T} Q^* = Q^*$, we have

$$\begin{aligned} \|E_{\ell+1}\|_\infty &= \|\tilde{\mathcal{R}}_{n,\lambda}^{\mu,\pi_Q} Q_\ell - Q^*\|_\infty = \|Q_\ell + \tilde{\mathcal{C}}_{n,\lambda}^\mu (\mathcal{T} Q_\ell - Q_\ell) - Q^*\|_\infty \\ &= \|\tilde{\mathcal{C}}_{n,\lambda}^\mu ((\mathcal{T} Q_\ell - \mathcal{T} Q^*) - (Q_\ell - Q^*)) + (Q_\ell - Q^*)\|_\infty \\ &= \|(\mathcal{I} - \tilde{\mathcal{C}}_{n,\lambda}^\mu)(Q_\ell - Q^*) + \tilde{\mathcal{C}}_{n,\lambda}^\mu (\mathcal{T} Q_\ell - \mathcal{T} Q^*)\|_\infty \\ &\leq \|\tilde{\mathcal{C}}_{n,\lambda}^\mu - \mathcal{I}\|_\infty \|E_\ell\|_\infty + \|\tilde{\mathcal{C}}_{n,\lambda}^\mu\|_\infty \|\mathcal{T} Q_\ell - \mathcal{T} Q^*\|_\infty. \end{aligned}$$

Then, using the fact that $\mathcal{T}$ is a γ -contraction, we arrive at

$$\|E_{\ell+1}\|_\infty \leq \left(\|\tilde{\mathcal{C}}_{n,\lambda}^\mu - \mathcal{I}\|_\infty + \gamma \|\tilde{\mathcal{C}}_{n,\lambda}^\mu\|_\infty \right) \|E_\ell\|_\infty. \quad (35)$$

Thus, the problem reduces to computing the ∞ -norms of $\tilde{\mathcal{C}}_{n,\lambda}^\mu$ and $\tilde{\mathcal{C}}_{n,\lambda}^\mu - \mathcal{I}$. Observe that both operators are non-negative. In particular, using the row-sum identity (27c), we have

$$\tilde{\mathcal{C}}_{n,\lambda}^\mu \mathbf{1} = \left(\sum_{k=0}^{n-1} (\lambda \gamma \mathcal{P}^\mu)^k + \xi \mathcal{P}_\infty^\mu \right) \mathbf{1} = \left(\sum_{k=0}^{n-1} (\lambda \gamma)^k + \xi \right) \mathbf{1} = \frac{1}{1 - \gamma \lambda} \mathbf{1}.$$

Similarly,

$$(\tilde{\mathcal{C}}_{n,\lambda}^\mu - \mathcal{I}) \mathbf{1} = \left(\sum_{k=1}^{n-1} (\lambda \gamma \mathcal{P}^\mu)^k + \xi \mathcal{P}_\infty^\mu \right) \mathbf{1} = \left(\sum_{k=1}^{n-1} (\lambda \gamma)^k + \xi \right) \mathbf{1} = \frac{\gamma \lambda}{1 - \gamma \lambda} \mathbf{1},$$

where the sum is empty when $n = 1$. Then, the non-negative operator norm identity (28) implies that $\|\tilde{\mathcal{C}}_{n,\lambda}^\mu\|_\infty = 1/(1 - \gamma \lambda)$ and $\|\tilde{\mathcal{C}}_{n,\lambda}^\mu - \mathcal{I}\|_\infty = \gamma \lambda / (1 - \gamma \lambda)$. Substituting these norms into (35), we arrive at

$$\|E_{\ell+1}\|_\infty \leq \left(\frac{\gamma \lambda}{1 - \gamma \lambda} + \frac{\gamma}{1 - \gamma \lambda} \right) \|E_\ell\|_\infty = \eta_2 \|E_\ell\|_\infty.$$

Finally, requiring $\eta_2 < 1$ for convergence rearranges to $\lambda < (1 - \gamma)/(2\gamma)$, which completes the proof.

A.7. Proof of Theorem 3.6

We use the standard two-timescale stochastic approximation theorem based on the ODE method [6, §8.1]. For completeness, we state the theorem below.

Theorem A.1 (Two-timescale SA with ODE limit). [6, §8.1] *Let $\{\mathbf{F}_k\}_{k \geq 0}$ be a filtration, and consider iterates $(X_k, Y_k) \in \mathbb{R}^d \times \mathbb{R}^m$ satisfying*

$$\begin{aligned} X_{k+1} &= X_k + \alpha_k (H(X_k, Y_k) + M_{k+1}^X), \\ Y_{k+1} &= Y_k + \beta_k (G(X_k, Y_k) + M_{k+1}^Y). \end{aligned}$$

Assume:

(i) *The step sizes are positive and satisfy*

$$\sum_k \alpha_k = \sum_k \beta_k = \infty, \quad \sum_k (\alpha_k^2 + \beta_k^2) < \infty, \quad \alpha_k / \beta_k \rightarrow 0.$$

(ii) *The maps H and G are Lipschitz.*

(iii) *$\{M_{k+1}^X\}$ and $\{M_{k+1}^Y\}$ are martingale-difference noise terms with respect to $\{\mathbf{F}_k\}$ and, for some $K < \infty$,*

$$\mathbb{E}[\|M_{k+1}^X\|^2 + \|M_{k+1}^Y\|^2 \mid \mathbf{F}_k] \leq K(1 + \|X_k\|^2 + \|Y_k\|^2).$$

- (iv) The iterates are almost surely bounded, that is, $\sup_k (\|X_k\| + \|Y_k\|) < \infty$.
- (v) For every fixed X , the ODE $\dot{Y}(t) = G(X, Y(t))$ has a unique globally asymptotically stable equilibrium $Y^*(X)$, and Y^* is Lipschitz.
- (vi) The ODE $\dot{X}(t) = H(X(t), Y^*(X(t)))$ has a unique globally asymptotically stable equilibrium X^* .

Then $X_k \rightarrow X^*$ and $Y_k \rightarrow Y^*(X^*)$ almost surely.

We now apply Theorem A.1 with $X_k = Q_k$ and $Y_k = U_k$. Let $\{\mathbf{F}_k\}_{k \geq 0}$ be the filtration generated by (Q_0, U_0) and the samples collected before iteration k . Then Q_k and U_k are $\mathbf{F}_k$ -measurable. Define the drift functions $g, h : \mathbf{Q} \times \mathbf{Q} \rightarrow \mathbf{Q}$ for each state-action pair $z \triangleq (s, a)$ by

$$\begin{aligned} g(Q, U)(z) &\triangleq (1 - \tau)(\mathcal{T}Q - Q)(z) + \tau \mathcal{P}^\mu U(z) - U(z), \\ h(Q, U)(z) &\triangleq (\mathcal{C}_{n, \lambda}^\mu(\mathcal{T}Q - Q))(z) + \xi U(z). \end{aligned}$$

Let $\mathcal{B}_\rho : \mathbf{Q} \rightarrow \mathbf{Q}$ denote the positive diagonal operator with $\mathcal{B}_\rho X(z) \triangleq \rho(z)X(z)$, where ρ is the reset distribution. By Assumption 2 in Theorem 3.6, the initial state-action pair $z_k \triangleq (s_{k,0}, a_{k,0})$ is independent of $\mathbf{F}_k$ and has distribution ρ . The coverage assumption on ρ implies that the diagonal entries of $\mathcal{B}_\rho$ lie in $(0, 1]$. Thus, the expected drifts of the two-timescale recursion are $H \triangleq \mathcal{B}_\rho h$ and $G \triangleq \mathcal{B}_\rho g$.

The following two lemmas establish global asymptotic stability for the equilibria of the fast and slow ODEs, which we use below to verify assumptions (iv)–(vi).

Lemma A.2 (Fast ODE global asymptotic stability). *For any $V \in \mathbf{Q}$ and diagonal operator $\mathcal{B} : \mathbf{Q} \rightarrow \mathbf{Q}$ with diagonal entries in $(0, 1]$, the ODE*

$$\dot{U}(t) = \mathcal{B}(V + \tau \mathcal{P}^\mu U(t) - U(t))$$

has the unique globally asymptotically stable equilibrium $U^\dagger = (\mathcal{I} - \tau \mathcal{P}^\mu)^{-1}V$.

Proof. Recall that $\tau \in [0, 1)$ and $\mathcal{P}^\mu$ is stochastic. It follows that, for fixed V , the map $U \mapsto V + \tau \mathcal{P}^\mu U$ is a τ -contraction in ∞ -norm with fixed point $U^\dagger$. Applying the argument in Borkar [6, §6.4] for diagonally scaled ODEs, we conclude that $U^\dagger$ is the unique globally asymptotically stable equilibrium. $\square$

Lemma A.3 (Slow ODE global asymptotic stability). *Let $\hat{\mathcal{T}} : \mathbf{Q} \rightarrow \mathbf{Q}$ be γ -Lipschitz in ∞ -norm with fixed point $Q^\dagger$, and $\mathcal{B} : \mathbf{Q} \rightarrow \mathbf{Q}$ be a diagonal operator with diagonal entries in $(0, 1]$. If $\lambda < (1 - \gamma)/(2\gamma)$, the ODE*

$$\dot{Q}(t) = \mathcal{B}(\mathcal{C}_{n, \lambda}^\mu + \xi \mathcal{A}_\tau^\mu)(\hat{\mathcal{T}}Q(t) - Q(t))$$

has $Q^\dagger$ as the unique globally asymptotically stable equilibrium.

Proof. The map $Q \mapsto Q + (\mathcal{C}_{n, \lambda}^\mu + \xi \mathcal{A}_\tau^\mu)(\hat{\mathcal{T}}Q - Q)$ is an η_2 -contraction in ∞ -norm with fixed point $Q^\dagger$, using the same arguments as in the proof of Theorem 3.5. In particular, observe that the operator $\mathcal{A}_\tau^\mu$ shares the same norm identities used there for $\mathcal{P}_\infty^\mu$. Furthermore, by Borkar and Soumyanath [7, Thm. 4.2], for any ∞ -norm contraction $\mathcal{F}$ with fixed point X^* and any scalar $a > 0$, the ODE

$\dot{X} = a(\mathcal{F}X - X)$ has X^* as its unique globally asymptotically stable equilibrium. Combining this with the argument in Borkar [6, §6.4] for diagonally scaled ODEs, we conclude that $Q^\dagger$ is globally asymptotically stable. $\square$

In what follows, we verify that assumptions (i)–(vi) of Theorem A.1 hold for the LKQL recursion.

(i) Step sizes. Assumption (i) coincides with Assumption 1 of Theorem 3.6 and therefore holds trivially.

(ii) Lipschitzness. With $\mu \in \Pi$ fixed, the operators $\mathcal{P}^\mu$, $\mathcal{C}_{n,\lambda}^\mu$, and $\mathcal{A}_\tau^\mu \triangleq (1 - \tau)(\mathcal{I} - \tau\mathcal{P}^\mu)^{-1}$ are constant non-negative linear maps with norms

$$\|\mathcal{P}^\mu\|_\infty = \|\mathcal{A}_\tau^\mu\|_\infty = 1, \quad \|\mathcal{C}_{n,\lambda}^\mu\|_\infty = \frac{1 - (\gamma\lambda)^n}{1 - \gamma\lambda} \leq \frac{1}{1 - \gamma\lambda}.$$

The Bellman optimality operator $\mathcal{T}$ is a γ -contraction in ∞ -norm. Hence, the Bellman residual operator $\mathcal{T} - \mathcal{I}$ is $(1 + \gamma)$ -Lipschitz in ∞ -norm.

Slow drift H. For any $(Q, U), (Q', U') \in \mathbf{Q} \times \mathbf{Q}$, using $\|\mathcal{B}_\rho\|_\infty \leq 1$, we have

$$\|H(Q, U) - H(Q', U')\|_\infty \leq \frac{1 + \gamma}{1 - \gamma\lambda} \|Q - Q'\|_\infty + \xi \|U - U'\|_\infty.$$

Fast drift G. Similarly, using $\|\mathcal{P}^\mu\|_\infty = 1$,

$$\|G(Q, U) - G(Q', U')\|_\infty \leq (1 - \tau)(1 + \gamma) \|Q - Q'\|_\infty + (1 + \tau) \|U - U'\|_\infty.$$

Fast equilibrium U^* . The fast equilibrium $U^*(Q) = \mathcal{A}_\tau^\mu(\mathcal{T}Q - Q)$ satisfies the Lipschitz bound

$$\|U^*(Q) - U^*(Q')\|_\infty \leq \|\mathcal{A}_\tau^\mu\|_\infty (1 + \gamma) \|Q - Q'\|_\infty = (1 + \gamma) \|Q - Q'\|_\infty.$$

This verifies (ii) and the Lipschitz hypothesis on $Y^*(\cdot) = U^*(\cdot)$ in (v).

(iii) Noise. Write the LKQL update (17) in full-vector form

$$\begin{aligned} Q_{k+1} &= Q_k + \alpha_k \Delta_{k+1}^Q, & \Delta_{k+1}^Q(z_k) &\triangleq \sum_{j=0}^{n-1} (\gamma\lambda)^j \delta_{k,j}^Q + \xi U_k(z_k), \\ U_{k+1} &= U_k + \beta_k \Delta_{k+1}^U, & \Delta_{k+1}^U(z_k) &\triangleq \delta_{k,0}^U, \end{aligned}$$

where Δ_{k+1}^Q and Δ_{k+1}^U are sparse update vectors supported only at z_k , so $\Delta_{k+1}^Q(z) = \Delta_{k+1}^U(z) = 0$ for $z \neq z_k$. For every $z \in \mathbf{S} \times \mathbf{A}$, conditioning on $\mathbf{F}_k$ and applying the Markov property gives

$$\begin{aligned} \mathbb{E} [\Delta_{k+1}^Q(z) \mid \mathbf{F}_k] &= \rho(z) \left[\sum_{j=0}^{n-1} (\gamma\lambda)^j ((\mathcal{P}^\mu)^j(\mathcal{T}Q_k - Q_k))(z) + \xi U_k(z) \right] \\ &= \rho(z) h(Q_k, U_k)(z) = H(Q_k, U_k)(z), \\ \mathbb{E} [\Delta_{k+1}^U(z) \mid \mathbf{F}_k] &= \rho(z) [(1 - \tau)(\mathcal{T}Q_k - Q_k)(z) + \tau(\mathcal{P}^\mu U_k)(z) - U_k(z)] \\ &= \rho(z) g(Q_k, U_k)(z) = G(Q_k, U_k)(z). \end{aligned}$$

Consequently, the full-vector conditional expectations satisfy

$$\mathbb{E} [\Delta_{k+1}^Q \mid \mathbf{F}_k] = H(Q_k, U_k), \quad \mathbb{E} [\Delta_{k+1}^U \mid \mathbf{F}_k] = G(Q_k, U_k).$$

We therefore define the martingale-difference noise terms by

$$M_{k+1}^Q \triangleq \Delta_{k+1}^Q - H(Q_k, U_k) \quad \text{and} \quad M_{k+1}^U \triangleq \Delta_{k+1}^U - G(Q_k, U_k),$$

so the LKQL recursion takes the form of Theorem A.1 with drifts $H = \mathcal{B}_\rho h$ and $G = \mathcal{B}_\rho g$. By construction, $\mathbb{E}[M_{k+1}^Q | \mathbf{F}_k] = \mathbb{E}[M_{k+1}^U | \mathbf{F}_k] = 0$.

We next bound the noise second moments. Each TD error satisfies $|\delta_{k,j}^Q| \leq \|r\|_\infty + (1 + \gamma) \|Q_k\|_\infty$. Hence, for some finite constants L_1, L_2 , we obtain

$$\begin{aligned} \|\Delta_{k+1}^Q\|_2 &= |\Delta_{k+1}^Q(z_k)| \leq \frac{\|r\|_\infty + (1 + \gamma) \|Q_k\|_\infty}{1 - \gamma\lambda} + \xi \|U_k\|_\infty \leq L_1(1 + \|Q_k\|_\infty + \|U_k\|_\infty), \\ \|\Delta_{k+1}^U\|_2 &= |\Delta_{k+1}^U(z_k)| \leq (1 - \tau)|\delta_{k,0}^Q| + (1 + \tau) \|U_k\|_\infty \leq L_2(1 + \|Q_k\|_\infty + \|U_k\|_\infty). \end{aligned}$$

Combining these bounds with the linear growth of H and G established in (ii), we obtain

$$\mathbb{E}[\|M_{k+1}^Q\|_2^2 + \|M_{k+1}^U\|_2^2 | \mathbf{F}_k] \leq K(1 + \|Q_k\|_\infty^2 + \|U_k\|_\infty^2) \leq K(1 + \|Q_k\|_2^2 + \|U_k\|_2^2),$$

for some $K < \infty$, where the second inequality uses $\|\cdot\|_\infty \leq \|\cdot\|_2$. This verifies (iii).

(iv) Boundedness. To establish boundedness of the iterates, we apply the two-timescale stability theorem of Lakshminarayanan and Bhatnagar [28]. This theorem guarantees a.s. boundedness $\sup_k(\|Q_k\|_\infty + \|U_k\|_\infty) < \infty$ under conditions (i)–(iii) of Theorem A.1 together with the following two additional scaling conditions:

- (a) the rescaled fast drift $G_c(Q, U) \triangleq G(cQ, cU)/c$ converges uniformly on compacts as $c \rightarrow \infty$ to a limit G_∞ , and the ODE $\dot{U}_\infty = G_\infty(Q, U)$ has a unique globally asymptotically stable equilibrium $U_\infty^*(Q)$ that is Lipschitz in Q with $U_\infty^*(0) = 0$;
- (b) the rescaled slow drift at the fast equilibrium, $H_c(Q) \triangleq H(cQ, cU_\infty^*(Q))/c$, converges uniformly on compacts as $c \rightarrow \infty$ to a limit H_∞ , and the ODE $\dot{Q}_\infty = H_\infty(Q)$ has the origin as its unique globally asymptotically stable equilibrium.

Conditions (i)–(iii) are verified above. We verify (a) and (b) next.

For $c \geq 1$, scaling Q by c leaves all linear operators in H and G unchanged and only affects the reward term in $\mathcal{T}$. In particular, we have $\frac{\mathcal{T}(cQ)}{c} = \frac{r}{c} + \gamma \mathcal{P}^{\pi_Q} Q$ with $\lim_{c \rightarrow \infty} \frac{\mathcal{T}(cQ)}{c} = \gamma \mathcal{P}^{\pi_Q} Q$. The map $Q \mapsto \gamma \mathcal{P}^{\pi_Q} Q$ is γ -Lipschitz in ∞ -norm and has the unique fixed point 0. This follows from the same argument as in (ii), applied to the reward-free part of $\mathcal{T}$.

Scaled fast equilibrium. Since $\mathcal{T}(cQ)/c \rightarrow \gamma \mathcal{P}^{\pi_Q} Q$ uniformly on compacts as $c \rightarrow \infty$, the rescaled fast drift converges uniformly on compacts to

$$G_\infty(Q, U) \triangleq \mathcal{B}_\rho((1 - \tau)(\gamma \mathcal{P}^{\pi_Q} Q - Q) + \tau \mathcal{P}^\mu U - U).$$

For fixed Q , applying Lemma A.2 with $V = (1 - \tau)(\gamma \mathcal{P}^{\pi_Q} Q - Q)$ yields the unique globally asymptotically stable equilibrium $U_\infty^*(Q) = \mathcal{A}_\tau^\mu(\gamma \mathcal{P}^{\pi_Q} Q - Q)$. Observe that $U_\infty^*(0) = 0$ and that U_∞^* is Lipschitz in Q by the same arguments as for U^* in (ii), with $\mathcal{T}$ replaced by $\gamma \mathcal{P}^{\pi_Q}$.

Scaled slow ODE. Once again, using the uniform convergence $\mathcal{T}(cQ)/c \rightarrow \gamma \mathcal{P}^{\pi_Q} Q$ on compacts as $c \rightarrow \infty$, and substituting $U_\infty^*(Q)$ into the rescaled slow drift, we have $H_c(Q) \rightarrow H_\infty(Q)$ uniformly on compacts, where

$$H_\infty(Q) \triangleq \mathcal{B}_\rho(\mathcal{C}_{n,\lambda}^\mu + \xi \mathcal{A}_\tau^\mu)(\gamma \mathcal{P}^{\pi_Q} Q - Q).$$

TABLE 1. Hyperparameters and their ranges for the continuous-control experiments

Symbol	Description	Value
γ	discount factor	0.99
λ	return trace coefficient	0.95
n	truncation	[4, rollout length]
–	LK-term truncation	rollout length
λ_U	LK-term trace coefficient	[0.5, 0.999]
τ	LK-term discount factor	[0.5, 0.9999]
β_U	LK-term loss coefficient	1 (fixed)
–	LK-term learning rate	$[10^{-5}, 10^{-1}]$ (constant)
ϵ_{clip}	PPO clipping parameter	0.2
β_π	policy loss coefficient	2
–	entropy coefficient	0
–	actor-critic learning rate	3×10^{-4} (linearly annealed)
–	gradient-norm clip	0.5
–	NAF exploration noise scale	$\{0.1, 0.2, 0.25\}$
N	parallel actors	16
T	outer iterations	500
–	minibatches per epoch	32
–	minibatch size	rollout length/2
–	epochs per iteration	10
S	seeds	32
–	rollout length	$\{32, 64, 128, 256, 512\}$

Since $\gamma\mathcal{P}^{\pi_Q}$ is γ -Lipschitz in ∞ -norm and has fixed point 0, Lemma A.3 applied with $\widehat{\mathcal{T}} = \gamma\mathcal{P}^{\pi_Q}$ and $Q^\dagger = 0$ implies that the origin is the unique globally asymptotically stable equilibrium of the scaled slow ODE $\dot{Q}_\infty(t) = H_\infty(Q_\infty(t))$.

Conditions (i)–(iii), (a), and (b) allow us to apply the stability theorem of Lakshminarayanan and Bhatnagar [28], which gives $\sup_{k \geq 0} (\|Q_k\|_\infty + \|U_k\|_\infty) < \infty$ almost surely, verifying (iv).

(v) Fast-scale ODE. For $V \triangleq (1 - \tau)(\mathcal{T}Q - Q)$ and the fast ODE $\dot{U}(t) = \mathcal{B}_\rho(V + \tau\mathcal{P}^\mu U(t) - U(t))$, Lemma A.2 yields the unique globally asymptotically stable equilibrium $U^*(Q) = \mathcal{A}_\tau^\mu(\mathcal{T}Q - Q)$. Together with the Lipschitzness of $Q \mapsto U^*(Q)$ shown in (ii), this verifies (v).

(vi) Slow-scale ODE. On the slow timescale, the fast iterate is equilibrated at $U^*(Q)$, giving the limiting slow ODE $\dot{Q}(t) = \mathcal{B}_\rho(\mathcal{C}_{n,\lambda}^\mu + \xi\mathcal{A}_\tau^\mu)(\mathcal{T}Q(t) - Q(t))$. For $\widehat{\mathcal{T}} = \mathcal{T}$ and $Q^\dagger = Q^*$, Lemma A.3 yields Q^* as the unique globally asymptotically stable equilibrium, verifying (vi).

Finally, combining (i)–(vi), we have $Q_k \rightarrow Q^*$ and $U_k \rightarrow U^*(Q^*) = 0$ almost surely.

Appendix B. Experimental Setup

We implement the algorithms in JAX [8] and evaluate on Hopper-v5, Walker2d-v5, and HalfCheetah-v5 from MuJoCo [53], simulated with its JAX-based MJX engine. LKQL and its baselines share the same actor-critic backbone and sampling loop (Algorithm 1), where the baselines are recovered exactly by disabling the LK term ($U_\omega \equiv 0$), yielding GAE in the on-policy setting and THQL in the off-policy setting. We use the `rlax` package [13] to compute the truncated returns for both the values and the LK term.

Architectures. All networks are multilayer perceptrons with tanh activations, layer normalization, and orthogonal initialization. The on-policy backbone parameterizes the policy π_θ with hidden sizes (64, 64) and a state-independent log standard deviation, and the value V_ϕ with (128, 128). The off-policy NAF backbone parameterizes the state value V_ϕ with (128, 128), the deterministic mean $\bar{\pi}_\theta$ with (64, 64), and the (diagonal) precision P_ϕ with (128, 128). In both estimators, the LK term U_ω is a separate network with the same hidden sizes as the critic, (128, 128), mapping the observation to a scalar.

Optimization. We optimize with Adam ($\epsilon_{\text{Adam}} = 10^{-5}$) and clip gradients to a global norm of 0.5. The actor and critic share a single optimizer with learning rate 3×10^{-4} annealed linearly to zero over training, whereas the LK term U_ω uses a separate optimizer with a constant learning rate, a choice motivated by the two-timescale analysis of Theorem 3.6. Each run uses $N = 16$ parallel actors and $T = 500$ outer iterations. Each iteration collects a rollout per actor and performs 10 epochs of updates with 32 minibatches per epoch. We report the final return of the deterministic policy, averaged over $S = 32$ seeds. The rollout length is varied over the set $\{32, 64, 128, 256, 512\}$. Since each iteration collects (rollout length $\times N$) transitions, the total number of environment steps scales with the rollout length.

Hyperparameters. Table 1 lists the shared settings. The baselines are tuned over the rollout length and, for NAF, the exploration noise scale. LKQL inherits the resulting best settings and tunes only its own hyperparameters, namely, the LK trace coefficient λ_U , the LK-term discount factor τ , the n -step truncation horizon, and the optimizer parameters for ω .

References

- [1] Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., Gelly, S., and Bachem, O. (2021). What matters in on-policy reinforcement learning? A large-scale empirical study. In *International Conference on Learning Representations*.
- [2] Archibald, T. W., McKinnon, K. I. M., and Thomas, L. C. (1995). On the generation of Markov decision processes. *Journal of the Operational Research Society*, 46:354–361.
- [3] Bertsekas, D. P. (1995). Generic rank-one corrections for value iteration in Markovian decision problems. *Operations Research Letters*, 17:111–119.
- [4] Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*.
- [5] Boldi, P., Lonati, V., Santini, M., and Vigna, S. (2006). Graph fibrations, graph isomorphism, and PageRank. *RAIRO—Theoretical Informatics and Applications*, 40:227–253.

- [6] Borkar, V. S. (2023). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2nd edition.
- [7] Borkar, V. S. and Soumyanath, K. (1997). An analog scheme for fixed point computation—part I: theory. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44:351–355.
- [8] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Katariya, Y., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: composable transformations of Python+NumPy programs.
- [9] Chen, S., Devraj, A. M., Lu, F., Bušić, A., and Meyn, S. (2020). Zap Q-learning with nonlinear function approximation. In *Advances in Neural Information Processing Systems*.
- [10] Chen, Z., Clarke, J.-P., and Maguluri, S. T. (2023). Target network and truncation overcome the deadly triad in Q-learning. *SIAM Journal on Mathematics of Data Science*, 5:1078–1101.
- [11] Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. (2022). Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623.
- [12] Cobbe, K., Hilton, J., Klimov, O., and Schulman, J. (2021). Phasic policy gradient. In *International Conference on Machine Learning*.
- [13] DeepMind, Babuschkin, I., Baumli, K., Bell, A., Bhupatiraju, S., Bruce, J., Buchlovsky, P., Budden, D., Cai, T., Clark, A., Danihelka, I., Dedieu, A., Fantacci, C., Godwin, J., Jones, C., Hemsley, R., Hennigan, T., Hessel, M., Hou, S., Kapturowski, S., Keck, T., Kemaev, I., King, M., Kunesch, M., Martens, L., Merzic, H., Mikulik, V., Norman, T., Papamakarios, G., Quan, J., Ring, R., Ruiz, F., Sanchez, A., Sartran, L., Schneider, R., Sezener, E., Spencer, S., Srinivasan, S., Stanojević, M., Stokowiec, W., Wang, L., Zhou, G., and Viola, F. (2020). The DeepMind JAX Ecosystem.
- [14] Devraj, A. M. and Meyn, S. (2017). Zap Q-learning. In *Advances in Neural Information Processing Systems*.
- [15] Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556.
- [16] Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. (2018). IMPALA: Scalable distributed Deep-RL with importance weighted actor-learner architectures. In *International Conference on Machine Learning*.
- [17] Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep Q-learning. In *Conference on Learning for Dynamics and Control*.
- [18] Farahmand, A.-M., Munos, R., and Szepesvári, C. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*.
- [19] Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., and Dabney, W. (2020). Revisiting fundamentals of experience replay. In *International Conference on Machine Learning*.
- [20] Gallager, R. G. (2013). *Stochastic Processes: Theory for Applications*. Cambridge University Press.
- [21] Gu, S., Lillicrap, T., Sutskever, I., and Levine, S. (2016). Continuous deep Q-learning with model-based acceleration. In *International Conference on Machine Learning*.
- [22] Harutyunyan, A., Bellemare, M. G., Stepleton, T., and Munos, R. (2016). $Q(\lambda)$ with off-policy corrections. In *International Conference on Algorithmic Learning Theory*.
- [23] Hernández-García, J. F. and Sutton, R. S. (2019). Understanding multi-step deep reinforcement learning: A systematic study of the DQN target. *arXiv preprint 1901.07510*.
- [24] Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. (2018). Rainbow: Combining improvements in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*.

- [25] Kapturowski, S., Ostrovski, G., Quan, J., Munos, R., and Dabney, W. (2019). Recurrent experience replay in distributed reinforcement learning. In *International Conference on Learning Representations*.
- [26] Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. (2022). Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*.
- [27] Kozuno, T., Tang, Y., Rowland, M., Munos, R., Kapturowski, S., Dabney, W., Valko, M., and Abel, D. (2021). Revisiting Peng’s $Q(\lambda)$ for modern reinforcement learning. In *International Conference on Machine Learning*.
- [28] Lakshminarayanan, C. and Bhatnagar, S. (2017). A stability criterion for two timescale stochastic approximation schemes. *Automatica*, 79:108–114.
- [29] Lee, J., Rakhsha, A., Ryu, E. K., and Farahmand, A.-M. (2025). Deflated dynamics value iteration. *Transactions on Machine Learning Research*.
- [30] Lim, H.-D. and Lee, D. (2024). Regularized Q-learning. In *Advances in Neural Information Processing Systems*.
- [31] Lim, H.-D. and Lee, D. (2025). Understanding the theoretical properties of projected Bellman equation, linear Q-learning, and approximate value iteration. *arXiv preprint 2504.10865*.
- [32] Liu, X., Xie, Z., and Zhang, S. (2025). Linear Q-learning does not diverge in L^2 : Convergence rates to a bounded set. In *International Conference on Machine Learning*.
- [33] Maei, H. R., Szepesvári, C., Bhatnagar, S., and Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *International Conference on Machine Learning*.
- [34] Melo, F. S. and Ribeiro, M. I. (2007). Convergence of Q-learning with linear function approximation. In *European Control Conference*.
- [35] Meyn, S. (2024). The projected Bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 69:8323–8337.
- [36] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*.
- [37] Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. (2016). Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*.
- [38] Ok, T., Sharifi Kolarijani, A., Sharifi Kolarijani, M. A., and Mohajerin Esfahani, P. (2025). Control and reinforcement learning through the lens of optimization: An algorithmic perspective. *arXiv preprint 2512.07377*.
- [39] Peng, J. and Williams, R. J. (1994). Incremental multi-step Q-learning. In *International Conference on Machine Learning*.
- [40] Perkins, S. and Leslie, D. S. (2012). Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems*, 2:409–446.
- [41] Precup, D., Sutton, R. S., and Singh, S. (2000). Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*.
- [42] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.
- [43] Rakhsha, A., Wang, A., Ghavamzadeh, M., and Farahmand, A.-M. (2022). Operator splitting value iteration. In *Advances in Neural Information Processing Systems*.
- [44] Riedmiller, M. (2005). Neural fitted Q iteration—first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*.

- [45] Scherrer, B. (2016). Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41:758–774.
- [46] Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. (2015). Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16:1629–1676.
- [47] Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*.
- [48] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint 1707.06347*.
- [49] Sharifi Kolarijani, A., Ok, T., Mohajerin Esfahani, P., and Sharifi Kolarijani, M. A. (2025). Rank-one modified value iteration. In *International Conference on Machine Learning*.
- [50] Sharifi Kolarijani, M. A. and Mohajerin Esfahani, P. (2026). From optimization to control: Quasi-policy iteration. *IEEE Transactions on Automatic Control*, 71:3880–3893.
- [51] Shen, H., Zhang, K., Hong, M., and Chen, T. (2023). Towards understanding asynchronous advantage actor-critic: Convergence and linear speedup. *IEEE Transactions on Signal Processing*, 71:2579–2594.
- [52] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition.
- [53] Todorov, E., Erez, T., and Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.
- [54] Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202.
- [55] Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.
- [56] Wu, Z., Greenwald, A., and Parr, R. (2025). A unifying view of linear function approximation in off-policy RL through matrix splitting and preconditioning. In *Advances in Neural Information Processing Systems*.
- [57] Zhang, S., Li, H., Wang, M., Liu, M., Chen, P.-Y., Lu, S., Liu, S., Murugesan, K., and Chaudhury, S. (2023). On the convergence and sample complexity analysis of deep Q-networks with ε -greedy exploration. In *Advances in Neural Information Processing Systems*.