Revisiting Peng's Q(λ) for Modern Reinforcement Learning

Tadashi Kozuno [1 *] **Yunhao Tang** [2 *] **Mark Rowland** [3] **Rémi Munos** [4] **Steven Kapturowski** [3] **Will Dabney** [3] **Michal Valko** [4] **David Abel** [3]

Abstract

Off-policy multi-step reinforcement learning algorithms consist of conservative and non-conservative algorithms: the former actively cut traces, whereas the latter do not. Recently, Munos et al. (2016) proved the convergence of conservative algorithms to an optimal Q-function. In contrast, non-conservative algorithms are thought to be unsafe and have a limited or no theoretical guarantee. Nonetheless, recent studies have shown that non-conservative algorithms empirically outperform conservative ones. Motivated by the empirical results and the lack of theory, we carry out theoretical analyses of Peng's Q(λ), a representative example of non-conservative algorithms. We prove that *it also converges to an optimal policy* provided that the behavior policy slowly tracks a greedy policy in a way similar to conservative policy iteration. Such a result has been conjectured to be true but has not been proven. We also experiment with Peng's Q(λ) in complex continuous control tasks, confirming that Peng's Q(λ) often outperforms conservative algorithms despite its simplicity. These results indicate that Peng's Q(λ), which was thought to be unsafe, is a theoretically-sound and practically effective algorithm.

1. Introduction

Q-learning is a canonical algorithm in reinforcement learning (RL) (Watkins, 1989). It is a single-step algorithm, in that it only uses individual transitions to update value estimates. Many *multi-step* generalisations of Q-learning have been proposed, which allow temporally-extended trajectories to be used in the updating of values (Bertsekas & Ioffe, 1996; Watkins, 1989; Peng & Williams, 1994; 1996; Precup et al., 2000; Harutyunyan et al., 2016; Munos et al., 2016; Rowland et al., 2020), potentially leading to more efficient credit assignment. Indeed, multi-step algorithms have often been observed to outperform single-step algorithms for control in a variety of RL tasks (Mousavi et al., 2017; Harb & Precup, 2017; Hessel et al., 2018; Barth-Maron et al., 2018; Kapturowski et al., 2018; Daley & Amato, 2019).

However, using multi-step algorithms for RL comes with both theoretical and practical difficulties. The discrepancy between the policy that generated the data to be learnt from (the *behavior policy*) and the policy being learnt about (the *target policy*) can lead to complex, non-convergent behavior in these algorithms, and so must be considered carefully. There are two main approaches to deal with this discrepancy (cf. Table 1). *Conservative methods* ensure convergence is guaranteed no matter what behavior policy is used, typically by truncating the trajectories used for learning. By contrast, *non-conservative methods* typically do not truncate trajectories, and as a result do not come with generic convergence guarantees. Nevertheless, non-conservative methods have consistently been found to outperform conservative methods in practical large-scale applications. Thus, there is a clear gap in our understanding about non-conservative methods; why do they so work well in practice, but lack the guarantees of their conservative counterparts?

In this paper, we address this question by studying a representative non-conservative algorithm, Peng's Q(λ) (Peng & Williams, 1994; 1996, PQL), in more realistic learning settings. Our results show that while PQL does not learn optimal policies under arbitrary behavior policies, a convergence guarantee can be recovered if the behavior policy tracks the target policy, as is often the case in practice. This represents a closing of the gap between the strong empirical performance of non-conservative methods and their previous lack of theoretical guarantees.

More concretely, our primary theoretical contributions bring new understanding to PQL, and are summarized as follows:

- A proof that PQL with a *fixed* behavior policy converges to a "biased" (i.e., different from Q^*) fixed-point.
- Analysis of the quality of the resulting policy.

[*]Equal contribution [1]Independent Researcher, Okayama, Japan (Now at the University of Alberta) [2]Columbia University, NY, USA [3]DeepMind, London, UK [4]DeepMind, Paris, France. Correspondence to: Tadashi Kozuno <tadashi.kozuno@gmail.com>, Yunhao Tang <yt2541@columbia.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Table 1. List of off-policy multi-step algorithms for control. Harutyunyan's Q(λ), Tree-backup, Watkins' Q(λ), and Peng's Q(λ) are abbreviated as HQL, TBL, WQL, and PQL, respectively (cf. Section 3.2 for details of the algorithms). Conservative column indicates if an algorithm is conservative or not (cf. Section 4). Convergence column indicates the convergence of algorithms to any fixed point, whereas Convergence to Q^* column indicates the convergence of algorithms to the optimal Q-function Q^*. ✓ indicates new results in the present paper. PQL converges to a biased fixed-point when the behavior policy is fixed. It converges to Q^* when a behavior policy is updated appropriately. (An exact condition is given in Section 5.)

Algorithm	Conservative	Convergence	Convergence to Q^*
α-TRACE (ROWLAND ET AL., 2020)	NO	?	?
C-TRACE (ROWLAND ET AL., 2020)	NO	?	?
HQL (HARUTYUNYAN ET AL., 2016)	NO	✓(WITH SMALL λ)	✓(WITH SMALL λ)
RETRACE (MUNOS ET AL., 2016)	YES	✓	✓
TBL (PRECUP ET AL., 2000)	YES	✓	✓
UNCORRECTED n-STEP RETURN	NO	?	?
WQL (WATKINS, 1989)	YES	✓	✓
PQL (PENG & WILLIAMS, 1994)	NO	✓ (BIASED)	✓ (CF. CAPTION)

- Convergence of PQL to an optimal policy when using appropriate behavior policy updates.
- Error propagation analysis when using approximations.

In addition to these theoretical insights, we validate the empirical performance of PQL through extensive experiments. Our focus is on continuous control tasks, where one encounters many technical challenges that do not exist in discrete control tasks (cf. Section 7.2). They are also accessible to a wider range of readers. We show that PQL can be easily extended to popular off-policy actor-critic algorithms such as DDPG, TD3 and SAC (Lillicrap et al., 2016; Fujimoto et al., 2018; Haarnoja et al., 2018). Over a large subset of tasks, PQL consistently outperforms other conservative and non-conservative baseline alternatives.

2. Notation and Definitions

For a finite set $\mathbf{A}$ and an arbitrary set $\mathbf{B}$, we let $\Delta_{\mathbf{A}}$ and $\mathbf{B}^{\mathbf{A}}$ be the probability simplex over $\mathbf{A}$ and the set of all mappings from $\mathbf{A}$ to $\mathbf{B}$, respectively.

Markov Decision Processes (MDP). We consider an MDP defined by a tuple $\langle \mathbf{X}, \mathbf{A}, \mathcal{P}, \mathcal{P}_0, \mathcal{R}, \gamma \rangle$, where $\mathbf{X}$ is the finite state space, $\mathbf{A}$ the finite action space, $\mathcal{P} : \mathbf{X} \times \mathbf{A} \to \Delta_{\mathbf{X}}$ the state transition probability kernel, $\mathcal{P}_0 \in \Delta_{\mathbf{X}}$ the initial state distribution, $\mathcal{R}$ the (conditional) reward distribution, and $\gamma \in [0, 1)$ the discount factor (Puterman, 1994). We let $r \in \mathbb{R}^{\mathbf{X} \times \mathbf{A}}$ be a reward function defined by $r(x, a) := \int r' \mathcal{R}(\mathrm{d}r'|x, a)$.

On the Finiteness of the State and Action Spaces. While we assume both $\mathbf{X}$ and $\mathbf{A}$ to be finite, most of theoretical results in the paper hold in continuous state spaces with appropriate measure-theoretic considerations. The finiteness assumption on the action space is necessary to guarantee the existence of the optimal policy (Puterman, 1994). In Appendix B, we discuss assumptions necessary to extend our theoretical results to continuous action spaces.

Policy and Value Functions. Suppose a policy $\pi : \mathbf{X} \to \Delta_{\mathbf{A}}$. We consider the standard RL setup where an agent interacts with an environment, generating a sequence of state-action-reward tuples $(X_t, A_t, R_t)_{t \geq 0}$ with A_t being an action sampled from some policy; throughout, we denote random variables by upper cases. Define $G = \sum_{t=0}^{\infty} \gamma^t R_t$ as the cumulative return. The state-value and Q-functions are defined by $V^\pi(x) := \mathbb{E}[G|X_0 = x, \pi]$ and $Q^\pi(x, a) := \mathbb{E}[G|X_0 = x, A_0 = a, \pi]$, respectively, where the conditioning by π means $A_t \sim \pi(\cdot|X_t)$.

Evaluation and Control. Two key tasks in RL are evaluation and control. The problem of evaluation is to learn the Q-function of a fixed policy. The aim in the control setting is to learn an optimal policy π_* defined as to satisfy $V^{\pi_*} := V^* \geq V^\pi, \forall \pi$ (the inequality is point-wise, i.e., $V^*(x) \geq V^\pi(x)$ for all $x \in \mathbf{X}$). Similarly to V^*, we let Q^* denote the optimal Q-function Q^{π_*}. As a greedy policy with respect to Q^* is optimal, it suffices to learn Q^*. In this paper, we are particularly interested in the off-policy control setting, where an agent collects data with a behavior policy μ, which is not necessarily the agent's current policy π. On-policy settings are a special case where $\pi = \mu$.

3. Multi-step RL Algorithms and Operators

Operators play a crucial role in RL since all value-based RL algorithms (exactly or approximately) update a Q-function based on the recursion $Q_{k+1} := \mathcal{O}_k Q_k$, where $\mathcal{O}_k : \mathbb{R}^{\mathbf{X} \times \mathbf{A}} \to \mathbb{R}^{\mathbf{X} \times \mathbf{A}}$ is an operator that characterizes

each algorithm. In this section, we review multi-step RL algorithms and their operators.

Basic Operators. Assume we have a fixed policy π. With an abuse of notations, we define operators $\pi : \mathbb{R}^{\mathbf{X}\times\mathbf{A}} \to \mathbb{R}^{\mathbf{X}}$ and $\mathcal{P} : \mathbb{R}^{\mathbf{X}} \to \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ by

$$(\pi Q)(x) := \textstyle\sum_{a\in\mathbf{A}} \pi(a|x)Q(x,a), \text{ and}$$
$$(\mathcal{P}V)(x,a) := \textstyle\sum_{y\in\mathbf{X}} \mathcal{P}(y|x,a)V(y)$$

for any $Q \in \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ and $V \in \mathbb{R}^{\mathbf{X}}$, respectively (hereafter, we omit "for any..." in definitions of operators for brevity). We define their composite $\mathcal{P}^\pi := \mathcal{P}\pi$. As a result, the Bellman operator $\mathcal{T}^\pi : \mathbb{R}^{\mathbf{X}\times\mathbf{A}} \to \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ is defined by $\mathcal{T}^\pi Q := r + \gamma\mathcal{P}^\pi Q$. For a function $Q \in \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$, we let $\mathbf{G}(Q)$ be the set of all greedy policies[1] with respect to Q. The Bellman optimality operator $\mathcal{T}$ is defined by $\mathcal{T}Q = \mathcal{T}^{\pi_Q}Q$ with $\pi_Q \in \mathbf{G}(Q)$[2]. Q-learning approximates the value iteration (VI) updates $Q_{k+1} := \mathcal{T}Q_k$.

3.1. On-policy Multi-step Operators for Control

We first introduce on-policy multi-step operators for control.

Modified Policy Iteration (MPI). MPI uses the recursion $Q_{k+1} := \mathcal{T}_n^{\pi_k}Q_k$ for Q-function updates (Puterman & Shin, 1978), where $\pi_k \in \mathbf{G}(Q_k)$. The n-step return operator $\mathcal{T}_n^\pi : \mathbb{R}^{\mathbf{X}\times\mathbf{A}} \to \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ is defined by $\mathcal{T}_n^\pi Q := (\mathcal{T}^\pi)^n Q$.

λ-Policy Iteration (λ-PI). λ-PI uses the recursion $Q_{k+1} := \mathcal{T}_\lambda^{\pi_k}Q_k$ for Q-function updates (Bertsekas & Ioffe, 1996), where $\pi_k \in \mathbf{G}(Q_k)$. The λ-return operator $\mathcal{T}_\lambda^\pi : \mathbb{R}^{\mathbf{X}\times\mathbf{A}} \to \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ is defined as

$$\begin{aligned}\mathcal{T}_\lambda^\pi Q &:= (1-\lambda)\textstyle\sum_{n=1}^\infty \lambda^{n-1}\mathcal{T}_n^\pi Q \\ &= Q + (\mathcal{I} - \gamma\lambda\mathcal{P}^\pi)^{-1}(\mathcal{T}^\pi Q - Q),\end{aligned}$$

where $(\mathcal{I} - \gamma\lambda\mathcal{P}^\pi)^{-1} := \sum_{t=0}^\infty (\gamma\lambda\mathcal{P}^\pi)^t$, and $\lambda \in [0,1]$.

3.2. Off-policy Multi-step Operators for Control

Next, we explain off-policy multi-step operators for control. We note that on-policy algorithms in the last subsection can be converted to off-policy versions by using importance sampling (Precup et al., 2000; Casella & Berger, 2002).

Uncorrected n-step Return. For a sequence of behavior policies $(\mu)_{k\geq 0}$, the uncorrected n-step return algorithm uses the recursion $Q_{k+1} := \mathcal{N}_n^{\mu_k,\pi_k}Q_k$ for Q-function updates (Hessel et al., 2018; Kapturowski et al., 2018), where $\pi_k \in \mathbf{G}(Q_k)$. Here, the uncorrected n-step return operator $\mathcal{N}_n^{\mu,\pi}$ is defined for any policies π and μ by

$$\mathcal{N}_n^{\mu,\pi}Q := (\mathcal{T}^\mu)^{n-1}\mathcal{T}^\pi Q.$$

[1]Note that there may be multiple greedy policies due to ties.

[2]Note that this definition is independent of the choice of π_Q.

Peng's Q(λ) (PQL) For a sequence of behavior policies $(\mu)_{k\geq 0}$, PQL uses the recursion $Q_{k+1} := \mathcal{N}_\lambda^{\mu_k,\pi_k}Q_k$ for Q-function updates (Peng & Williams, 1994; 1996), where $\pi_k \in \mathbf{G}(Q_k)$. Here, the PQL operator $\mathcal{N}_\lambda^{\mu,\pi}$ is defined for any policies π and μ by

$$\mathcal{N}_\lambda^{\mu,\pi}Q := (1-\lambda)\textstyle\sum_{n=1}^\infty \lambda^{n-1}\mathcal{N}_n^{\mu,\pi}Q\,, \qquad (1)$$

where $\lambda \in [0,1]$. Note that *PQL is a generalization of λ-PI* because it reduces to λ-PI when $\mu_k = \pi_k$. In other words, PQL is λ-PI with one additional degree of freedom in μ_k.

General Retrace. We next introduce a general version of the Retrace operator (Munos et al., 2016), from which other operators are obtained as special cases.

For a behavior policy μ and a target policy π, we let $\mathcal{P}^{c\mu} : \mathbb{R}^{\mathbf{X}\times\mathbf{A}} \to \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ be an operator defined by

$$(\mathcal{P}^{c\mu}Q)(x,a) := \sum_{(y,b)\in\mathbf{X}\times\mathbf{A}} \mathcal{P}(y|x,a)c(y,b)\mu(b|y)Q(y,b),$$

where c is an arbitrary non-negative function over $\mathbf{X}\times\mathbf{A}$ whose choice depends on an algorithm. Note that for any n, $((\mathcal{P}^{c\mu})^n Q)(x,a)$ can be estimated off-policy with data collected under the behavior policy μ.

A general Retrace operator $\mathcal{R}_\lambda^{c\mu,\pi} : \mathbb{R}^{\mathbf{X}\times\mathbf{A}} \to \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ is obtained by replacing $\mathcal{P}^\pi$ of $(\mathcal{I} - \gamma\lambda\mathcal{P}^\pi)^{-1}$ in the λ-return operator $\mathcal{T}_\lambda^\pi$ with $\mathcal{P}^{c\mu}$. Concretely,

$$\mathcal{R}_\lambda^{c\mu,\pi}Q := Q + (\mathcal{I} - \gamma\lambda\mathcal{P}^{c\mu})^{-1}(\mathcal{T}^\pi Q - Q).$$

The general Retrace algorithm updates its Q-function by $Q_{k+1} := \mathcal{R}_\lambda^{c_k\mu_k,\pi_k}Q_k$, where $(c_k)_{k\geq 0}$ is a sequence of arbitrary non-negative functions over $\mathbf{X}\times\mathbf{A}$, $(\mu_k)_{k\geq 0}$ is an arbitrary sequence of behavior policies, and $(\pi_k)_{k\geq 0}$ is a sequence of target policies that depends on an algorithm. Given the choices of c_k and π_k in Table 2, we recover a few known algorithms (Watkins, 1989; Peng & Williams, 1994; 1996; Precup et al., 2000; Harutyunyan et al., 2016; Munos et al., 2016; Rowland et al., 2020).

The general Retrace algorithm is off-policy as $(\mathcal{R}_\lambda^{c_k\mu_k,\pi_k}Q_k)(x_0,a_0)$ can be estimated off-policy by the following estimator given a trajectory $(x_t,a_t,r_t)_{t\geq 0}$ collected under μ_k:

$$Q_k(x_0,a_0) + \sum_{t=0}^\infty \Big(\prod_{u=1}^t c(x_u,a_u)\Big)\gamma^t\lambda^t\delta_t, \qquad (2)$$

where $\prod_{u=1}^0 c(x_u,a_u) := 1$, and δ_t is the TD error $r_t + \gamma(\pi_k Q_k)(x_{t+1}) - Q_k(x_t,a_t)$ at time step t.

4. Conservative and Non-conservative Multi-step RL Algorithms

Munos et al. (2016) showed that the following conditions suffice for the convergence of the general Retrace to Q^*:

Table 2. Choices of c_k and π_k in off-policy multi-step operators for control. See Section 3.2 for details. The same abbreviations as those in Table 1 are used. For brevity, we defined $\pi_{Q_k} \in \mathbf{G}(Q_k)$. We denote $\pi_k(a|x)/\mu_k(a|x)$ by $\rho_k(x,a)$ and $(1-\alpha)+\alpha\pi_{Q_k}(a|x)/\mu_k(a|x)$ by $\widetilde{\rho}_k(x,a)$. α-trace and C-trace look the same in the table, but C-trace adaptively changes α so that the trace length matches to a target trace length.

Algorithm	c_k	π_k
α-TRACE	$\min\{1, \widetilde{\rho}_k\}$	$\alpha\pi_{Q_k} + (1-\alpha)\mu_k$
C-TRACE	$\min\{1, \widetilde{\rho}_k\}$	$\alpha\pi_{Q_k} + (1-\alpha)\mu_k$
HQL	1	π_{Q_k}
RETRACE	$\min\{1, \rho_k\}$	ANY
TBL	π_k	ANY
WQL	$\min\{1, \rho_k\}$	π_{Q_k}
PQL	1	$\lambda\pi_{Q_k} + (1-\lambda)\mu_k$

1. $c_k(x,a) \in [0, \pi_k(a|x)/\mu_k(a|x)]$ for any k and $(x,a) \in \mathbf{X}\times\mathbf{A}$.
2. π_k satisfies some greediness condition, such as ε-greediness with decreasing ε as k increases; cf. Munos et al. (2016) for further details.

We call algorithms that satisfy the first condition *conservative* algorithms for reasons to be explained below. Otherwise, we call the algorithms *non-conservative*. See Table 1 for the classification of algorithms. The uncorrected n-step return algorithm can also be viewed as a non-conservative algorithm with *non-Markovian traces* that depend also on the past.

Conservativeness, Theoretical Guarantees, and Empirical Performance of Algorithms. Recall that in the general Retrace update estimator (2), the effect of the TD error δ_t is attenuated by $\prod_{u=1}^{t} c(x_u, a_u)$ in addition to $\gamma^t\lambda^t$. Hence, from the backward view (Sutton & Barto, 1998), the first condition intuitively requires that *the trace must be cut* if a sub-trajectory $(x_0, a_0, \ldots, x_t, a_t)$ is unlikely under π_k relative to μ_k. As a result, conservative algorithms only carry out *safe* updates to Q-functions.

As shown in (Munos et al., 2016), such conservative updates enable a convergence guarantee of general conservative algorithms. However, Rowland et al. (2020) observed that it often results in frequent trace cuts, and conservative algorithms usually benefit less from multi-step updates.

In contrast, non-conservative algorithms accumulate TD errors without carefully cutting traces. As a result, non-conservative algorithms might perform poorly. As we show later (Proposition 5), it is the case at least for Harutyunyan's Q(λ) (Harutyunyan et al. (2016), HQL), an instance of non-conservative algorithms, when a behavior policy is fixed. Nonetheless, non-conservative algorithms are known to perform well in practice (Hessel et al., 2018; Kapturowski et al., 2018; Daley & Amato, 2019). To understand its reason, it is important to characterize what kind of updates to the behavior policy entail the convergence of the overall algorithm. In the following sections, we take a step forward along this direction. We establish the convergence guarantee of PQL under two setups: **(1)** when the behavior policy is fixed; **(2)** when the behavior policy is updated in an appropriate way.

5. Theoretical Analysis of Peng's Q(λ)

In this section, we analyze Peng's Q(λ). We start with the *exact case* where there is no update errors in value functions. Later, we will consider the *approximate case* when accounting for update errors. The following lemma is particularly useful in theoretical analyses as well as practical implementations.

Lemma 1 (Harutyunyan et al., 2016)**.** *The PQL operator can be rewritten in the following forms:*

$$\mathcal{N}_\lambda^{\mu,\pi} Q = Q + (\mathcal{I} - \gamma\lambda\mathcal{P}^\mu)^{-1}\Big(\mathcal{T}^{\lambda\mu+(1-\lambda)\pi}Q - Q\Big)$$
$$= (\mathcal{I} - \gamma\lambda\mathcal{P}^\mu)^{-1}(r + \gamma(1-\lambda)\mathcal{P}^\pi Q)\,.$$

Proof. This is proven in (Harutyunyan et al., 2016), but we provided a proof in Appendix C for completeness. □

5.1. Exact Case with a Fixed Behavior Policy

We now analyze PQL with a fixed behavior policy μ. While the behavior policy is not fixed in a practical situation, the analysis shows a trade-off between bias and convergence rate. This trade-off is analogous to the bias-contraction-rate trade-off of off-policy multi-step algorithms for policy evaluation (Rowland et al., 2020) and sheds some light on important properties of PQL.

Concretely, we analyze the following algorithm:

$$\pi_k \in \mathbf{G}(Q_k) \text{ and } Q_{k+1} := \mathcal{N}_\lambda^{\mu,\pi_k} Q_k. \tag{3}$$

Harutyunyan et al. (2016) has proven that a fixed point of the PQL operator coincides with the unique fixed point of $\lambda\mathcal{T}^\mu + (1-\lambda)\mathcal{T}$, which is guaranteed to exist since $\lambda\mathcal{T}^\mu + (1-\lambda)\mathcal{T}$ is a contraction with modulus γ under L^∞-norm (see Appendix A for details about the contraction and other notions).

The existence of a fixed point does not imply the convergence of PQL, and we need to show that the distance between Q_k and the fixed point is decreasing. With the following theorem, we show that PQL does converge.

Theorem 2. *Let $\pi_\dagger$ be a policy such that $Q^{\lambda\mu+(1-\lambda)\pi_\dagger} \geq Q^{\lambda\mu+(1-\lambda)\pi}$ for any policy π, where the inequality is pointwise. Then, $\pi_\dagger \in \mathbf{G}(Q^{\lambda\mu+(1-\lambda)\pi_\dagger})$, and Q_k of PQL (3) uniformly converges to $Q^{\lambda\mu+(1-\lambda)\pi_\dagger}$ with the rate β^k, where $\beta := \gamma(1-\lambda)/(1-\gamma\lambda)$.*

Proof. See Appendix E. □

We build intuitions about the bias-convergence-rate trade-off implied in Theorem 2. When λ increases, the fixed point is $Q^{\lambda\mu+(1-\lambda)\pi_\dagger}$, whose bias against Q^* arguably increases; at the same time, the contraction rate β decreases, so that the contraction is faster.

Remark 1. *In Section 7.6 of (Sutton & Barto, 1998), it is conjectured that PQL with a fixed policy would converge to a hybrid of Q^μ and Q^*. Theorem 2 gives an answer to this conjecture and shows that Sutton & Barto (1998)'s conjecture is not necessarily true. Rather, the theorem shows that PQL converges to the Q-function of the best policy among policies of the form $\lambda\mu + (1-\lambda)\pi$.*

5.2. Approximate Case with a Fixed Behavior Policy

In practice, value-update errors are inevitable due to e.g., finite-sample estimations and function approximation errors. In this subsection, we provide the error propagation analysis of PQL with a fixed behavior policy. As we will see, the analysis depicts a trade-off between fixed point bias and error tolerance.

We analyze the following algorithm:

$$\pi_k \in \mathbf{G}(Q_k) \text{ and } Q_{k+1} := \mathcal{N}_\lambda^{\mu,\pi_k} Q_k + \varepsilon_k \,,$$

where $\varepsilon_k \in \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$ denotes the value-update error at iteration k. For simplicity, we use $\rho_k := \lambda\mu + (1-\lambda)\pi_k$ and $\rho_\dagger := \lambda\mu + (1-\lambda)\pi_\dagger$ in this subsection.

Remark 2. *We emphasize that ε_k should be rather understood as $Q_{k+1} - \mathcal{N}_\lambda^{\mu,\pi_k} Q_k$,the difference between the function Q_{k+1} at $k+1$-th iteration and the ideal update $\mathcal{N}_\lambda^{\mu,\pi_k} Q_k$. In practice, Q_{k+1} is obtained by first constructing a sample estimate $\mathcal{N}_\lambda^{\mu,\pi_k} Q_k$ and then fitting a parametric model to it by, for example, the square loss minimization. As a result, ε_k typically consists of estimation error and function approximation error.*

In Section 5.1, we showed $\lim_{k\to\infty} Q_k = Q^{\lambda\mu+(1-\lambda)\pi_\dagger}$ when $\varepsilon_k(x,a) = 0$ at every $(x,a) \in \mathbf{X}\times\mathbf{A}$, and $\pi_\dagger \in \mathbf{G}(Q^{\lambda\mu+(1-\lambda)\pi_\dagger})$. Therefore, π_k is an approximation to $\pi_\dagger$, and thus it is natural to define $V^{\rho_\dagger} - V^{\rho_k}$ as the loss of using π_k rather than $\pi_\dagger$. The following theorem provides an upper bound for the loss.

Theorem 3. *For any K, the following holds:*

$$\|V^{\rho_\dagger} - V^{\rho_K}\|_\infty \le O(\beta^K) + \frac{2}{1-\gamma}\sum_{k=0}^{K-1} \beta^{K-k-1}\|\varepsilon_k\|_\infty \,,$$

where $\|\cdot\|_\infty$ is the L_∞-norm defined for any real-valued function f by $\|f\|_\infty := \max_v |f(v)|$.

Proof. See Appendix G. □

Remark 3. *In Theorem 3, we provide an upper bound of the L_∞-norm of $V^{\rho_\dagger} - V^{\rho_K}$. As a result, the L_∞-norm of ε_k appears in the upper-bound. However in Appendix G, we prove a point-wise upper-bound of $V^{\rho_\dagger} - V^{\rho_K}$, from which an L_p-norm upper-bound can be obtained in a straightforward way (e.g., see Lemma 6 of Scherrer et al. (2015)). For simplicity, we present the L_∞-norm upper-bound.*

As we have already explained the bias-convergence-rate trade-off, for now we ignore the $O(\beta^K)$ term and focus on the error term. For simplicity, we assume $\|\varepsilon_k\|_\infty = \varepsilon$ for every k. Then,

$$\frac{2}{1-\gamma}\sum_{k=0}^{K-1} \beta^{K-k-1}\|\varepsilon_k\|_\infty = O\left(\frac{1-\gamma\lambda}{(1-\gamma)^2}\varepsilon\right),$$

In contrast, an analogous result of λ-PI is $O(\varepsilon/(1-\gamma)^2)$ (Scherrer, 2013). When $\lambda = 0$, these results coincide, which is expected since both λ-PI and PQL degenerate to value iteration. When $\lambda = 1$, PQL's error dependency is $O(\varepsilon/(1-\gamma))$, which is significantly better than $O(\varepsilon/(1-\gamma)^2)$. However in this case, PQL is completely biased and converges to Q^μ. At intermediate values of λ, PQL achieves a trade-off between error tolerance with bias by changing λ.

5.3. Approximate Case with Behavior Policy Updates

Previously, we have analyzed PQL with a fixed behavior policy. However, in practice, the behavior policy is updated along with the target policy. Besides, value-update errors are inevitable in complex tasks. As a result, PQL may behave quite differently in a practical scenario. This motivates our analysis for the following algorithm:[3]

$$\begin{aligned} Q_{k+1} &:= \mathcal{N}_\lambda^{\mu_k,\pi_k} Q_k + \varepsilon_k \\ \mu_k &:= \alpha\pi_k + (1-\alpha)\mu_{k-1}, \end{aligned} \tag{4}$$

where $\pi_k \in \mathbf{G}(Q_k)$, and $\alpha \in [1-\lambda, 1]$. Note that when $\alpha = 1$, this algorithm reduces to λ-PI as a special case. Though this behavior policy update closely resembles to that of conservative policy iteration (Kakade & Langford, 2002), here we require $\alpha \ge 1-\lambda$.

This algorithm has the following performance guarantee.

Theorem 4. *For any K, the following holds:*

$$\|V^* - V^{\pi_K}\|_\infty \le O(\zeta^K) + \frac{2}{1-\gamma}\sum_{l=0}^{K-1} \zeta^{K-l-1}\|\varepsilon_l\|_\infty \,,$$

where $\zeta := 1-\alpha+\alpha\gamma$. Hence, PQL with behavior policy updates converges to the optimal policy with the rate ζ^K.

[3] This algorithm updates the behavior policy after each application of the PQL operator. In Appendix F, we analyze a case where the behavior policy is updated after multiple applications of the PQL operator.

Proof. See Appendix H. □

Remark 4. *As noted in Remark 3, we can derive an L_p-norm upper-bound of $V^* - V^{\pi_K}$ given its point-wise upper-bound. In Appendix H, we actually provide its point-wise upper-bound.*

Remark 5. *Our results differ from existing work on off-policy learning and multi-step methods in several important respects. Munos et al. (2016) also analyse multi-step off-policy algorithms, but focus on identifying conditions for algorithms to be* conservative*, and provide asymptotic convergence guarantees of Q_k in the tabular setting under* any *choice of behaviour policies. In contrast, our analysis focuses on* non-conservative *algorithms, and provides a finite-time error bound for $V^* - V^{\pi_k}$ that incorporates approximation in the algorithm steps. In this regard, these results bear similarity with earlier analysis of multi-step* on-policy *algorithms such as λ-policy iteration (Scherrer, 2013). The central distinction is that our analysis clarifies what conditions are sufficient in non-conservative off-policy algorithms to retain convergence of V^{π_k} to V^*.*

The first term on the right hand side shows the convergence of PQL with behavior policy updates in an exact case, i.e., $\|\varepsilon_k\|_\infty = 0$ for any k. It states that the fastest convergence rate is γ^K (achieved when $\alpha = 1$), which is the same as the convergence rate of VI (Munos, 2005), policy iteration (Munos, 2003), MPI (Scherrer et al., 2012; 2015), and λ-PI (Scherrer, 2013). When $\alpha \neq 1$, the convergence rate coincides with that of conservative policy iteration (Scherrer, 2014). However we are not aware of a similar result of conservative λ-PI, which would be an analogue of PQL considered here. Theorem 4 also provides the error dependency of PQL (the second term on the right hand side). It coincides with the previous result of the above algorithms when $\alpha = 1$, as one would expect, since PQL with $\alpha = 1$ is precisely λ-PI. Nonetheless PQL allows some degree of off-policiness when $\alpha \neq 1$.

5.4. Oscillatory Behavior of HQL

In this section, we have proven the convergence of exact PQL (i.e., no value-update errors). However, the following proposition shows that exact HQL, an instance of non-conservative algorithms, does not converge in an MDP when the behavior policy is fixed. Nonetheless, in the same MDP, setting the behavior policy μ_k to a greedy policy $\pi_k \in \mathbf{G}(Q_k)$ guarantees the convergence.

Proposition 5. *There is an MDP such that when exact HQL is run with a fixed policy $\mu_k = \mu$ for all k, $\lambda = 1$, and $Q_0 = Q^\mu$, HQL's Q-function Q_k oscillates between two functions, and its greedy policy π_k oscillate between optimal and sub-optimal policies. Contrarily, if $\mu_k \in \mathbf{G}(Q_k)$, HQL converges to an optimal policy.*

Proof. A proof of the first claim is given in Appendix D. The second claim immediately follows by noting that if $\mu_k = \pi_k \in \mathbf{G}(Q_k)$, HQL is λ-PI, which is known to converge (Bertsekas & Ioffe, 1996). □

While this result is specialized to HQL, it sheds light on an important aspect of non-conservative algorithms in general:

While non-conservative algorithms may perform poorly when the behavior policy is fixed, they may converge to Q^* when the behavior policy is updated.

The above captures a critical aspect of how algorithms behave in practice, where the behavior policy is continuously updated.

6. Deep RL Implementations

We next show that Peng's Q(λ) can be conveniently implemented with established off-policy deep RL algorithms. Our experiments focus on continuous control problems where the action space $\mathbf{A} = [-1, 1]^m$. A primary motivation for considering continuous control benchmarks (e.g., (Brockman et al., 2016; Tassa et al., 2020)) is that they are usually more accessible to a wider RL research community, compared to challenging discrete control benchmarks such as Atari games (Bellemare et al., 2013).

6.1. Off-policy Actor-critic Algorithms

Off-policy actor-critic algorithms maintain a policy $\pi_\theta(a|x)$ with parameter θ and a Q-function critic $Q_\phi(x, a)$ with parameter ϕ. For the policy, a popular choice is the point mass distribution $\pi_\theta(a|x) = \delta(a - \pi_\theta(x))$, where $\pi_\theta(x) \in \mathbb{R}^{\mathbf{A}}$ (Lillicrap et al., 2016; Fujimoto et al., 2018; Barth-Maron et al., 2018). The algorithm collects data with an exploratory behavior policy μ and saves tuples (x_t, a_t, r_t) into a replay buffer $\mathcal{D}$. At each training iteration, the critic $Q_\phi(x, a)$ is updated by minimizing squared errors against a Q-function target $\mathbb{E}_D\left[(Q_\phi(x, a) - Q_{\text{target}}(x, a))^2\right]$. The policy is updated via the deterministic policy gradient $\theta \leftarrow \theta + \alpha \mathbb{E}_\mu\left[\nabla_\theta Q_\phi(x, \pi_\theta(x))\right]$ (Silver et al., 2014). See further details in Appendix J.

6.2. Implementations of Multi-step Operators

While approximate estimates to $\mathcal{T}Q(x, a)$ are arguably the simplest to implement, it only myopically looks ahead for one step. Usually, the learning can be significantly sped up when the targets are constructed with multi-step operators. (See, e.g, empirical examples in (Hessel et al., 2018; Barth-Maron et al., 2018; Kapturowski et al., 2018) and theoretical insights in (Rowland et al., 2020)) For example, the uncorrected n-step operator is estimated as follows (Hessel et al., 2018): given a n-step trajectory $(x_i, a_i, r_i)_{i=0}^n$,

the target at (x_0, a_0) is computed as $Q_{\text{target}}(x_0, a_0) = \sum_{i=0}^{n-1} \gamma^i r_0 + \gamma^n Q_{\phi^-}(x_n, \pi_{\theta^-}(x_n))$. Similar estimates could be derived for all multi-step operators introduced in Section 3, especially Peng's Q(λ). We present full details in Appendix J.

Desirable empirical properties of Peng's Q(λ). The estimates of Peng's Q(λ) do not require importance sampling ratios $\frac{\pi(a|x)}{\mu(a|x)}$. This is especially valuable for continuous control, where the policy could be deterministic, in which case algorithms such as Retrace (Munos et al., 2016) cuts traces immediately. Even when policies are stochastic and traces based on IS ratios are not cut immediately, prior work suggests that the trace cuts are usually pessimistic especially for high-dimensional action space (see, e.g., (Wang et al., 2017) for implementation techniques to mitigate the issue).

7. Experiments

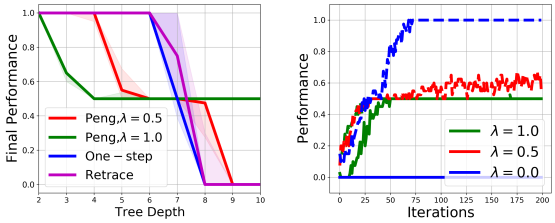
To build better intuitions about Peng's Q(λ), we start with tabular examples in Section 7.1. We will see that the empirical properties of Peng's Q(λ) echo the theoretical analysis in previous sections. In Section 7.2, we evaluate Peng's Q(λ) in the deep RL contexts. We combine Peng's Q(λ) with baseline deep RL algorithms and compare its performance against alternative operators.

7.1. A tabular example

Tree MDP. We consider toy examples with a tree MDP of depth D. The MDPs are binary trees, with each node corresponding to a state. Starting from any non-leaf state, the two actions $a \in \{L, R\}$ transition the agent to one of its child nodes with probability one. Each episode lasts for D steps and the agent always starts at the root node. The rewards are zero everywhere except $r = 1$ at the leftmost leaf node and $r = 0.5$ at the rightmost leaf node. The behavior policy μ is $\mu(L|x) = 0.3, \mu(R|x) = 0.7$ for all states x.

Note that there is a sub-optimal policy of collecting $r = 0.5$ at the rightmost leaf. The behavior policy is by design biased towards taking right moves, such that it is easy for the agent to learn the sub-optimal policy. The optimal policy is to take left moves and collect $r = 1$. Throughout training, we optimize the target policy π while fixing the behavior policy μ. This echos the theoretical setup in Section 5.2. See Appendix J for further details on the setup.

Results. In Figure 1(a), we show the performance of different algorithms after $10,000$ iterations as a function of the MDP's tree depth D. When $D = 2$, all algorithms achieve the optimal performance; when $\lambda = 1$, as D increases, the fixed point bias of Peng's Q(λ) hurts the performance drastically. This is less severe for $\lambda = 0.5$, whose performance decays less quickly. On the other hand, both Retrace and the one-step operator learn the optimal policy even for $D \leq 6$. However, when D increases, it becomes difficult to sample the optimal trajectory. As such, the sparse rewards make it difficult to learn meaningful Q-functions in a reasonable amount of time, unless the return signals get propagated effectively (i.e,. do not cut traces). This is shown in Figure 1(a), where Peng's Q(λ) with $\lambda = 1$ is the only method that finds a policy with non-zero expected return.



(a) Final performance (b) Learning curves

Figure 1. Performance on tree MDPs. Figure(a) shows how performance (after $10,000$ iterations) changes as a function of tree depth D; Figure(b) shows the learning curves of different operators.

Similar observations are made in Figure 1(b), where we compare Peng's Q(λ) for various λ under $D = 10$ (solid lines) and $D = 5$ (dotted lines). Small λ corresponds to less bias in the Q-function fixed points and should asymptotically converge to higher performance, but its ineffective reward signal propagation hinders policy improvement in a reasonable time when D is large; on the other hand, large λ suffers sub-optimality when D is small, but its initial policy improvement is substantially expedited when the D is large.

7.2. Deep RL experiments

Evaluations. We evaluate performance over environments with a number of different physics simulation backends, such as MuJoCo (Todorov et al., 2012) based DeepMind (DM) control suite (Tassa et al., 2020) and an open sourced simulator Bullet physics (Coumans & Bai, 2016–2019). Due to space limit, below we only show results for DM control suite and provide a more complete set of evaluations in Appendix J.

Baseline comparison. We use TD3 (Fujimoto et al., 2018) as the base algorithm.[4] We compare with a few multi-step baselines: **(1)** one-step (also the base algorithm); **(2)**

[4]TD3 reasonably echoes the theoretical setup in Section 5.3: both the behavior and target policies are near-greedy policy (slowly) trying to maximize Q_k, and Q-value updates are performed by using data obtained by following the behavior policy and stored in the replay buffer.

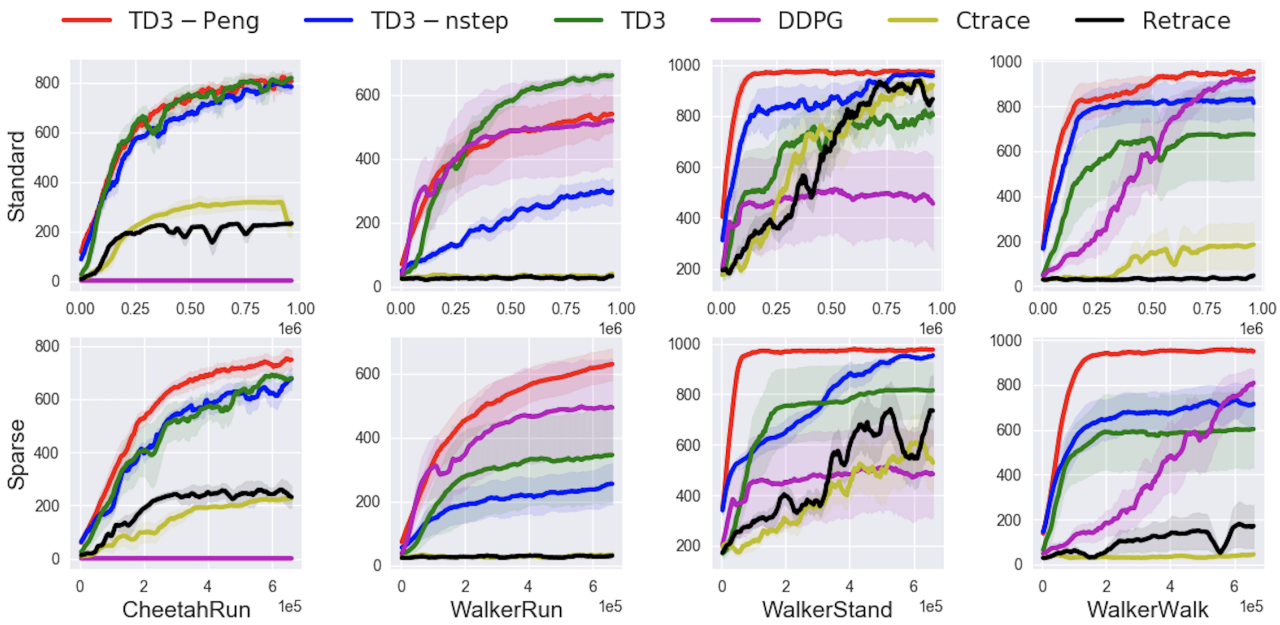


Figure 2. Evaluation of baseline algorithms over standard DM control domains. The first row shows results on standard benchmarks; the second row shows results on sparse reward variants of the benchmarks. Four task names are labeled at the bottom. In each plot, x-axis shows the number of training steps and y-axis shows the performance. In standard benchmarks, Peng's Q(λ) generally performs more stably than other algorithms; in sparse reward benchmarks, Peng's Q(λ) outperforms all other algorithms across all presented tasks.

Uncorrected n-step with a fixed n; **(3)** Peng's Q(λ) with a fixed λ; **(4)** Retrace and C-trace. Among all baselines, uncorrected n-step operator is the most commonly used non-conservative operator while Retrace is a representative conservative operator. See Appendix J for more details. All algorithms are trained with a fixed number of steps and results are averaged across 5 random seeds.

Standard benchmark results. In the top row of Figure 2, we show evaluations on standard benchmarks. Across most tasks, Peng's Q(λ) performs more stably than other baseline algorithms. We see that Peng's Q(λ) learns generally as fast as other baselines, and in some cases significantly faster than others. Note that though Peng's Q(λ) does not necessarily obtain the best learning performance *per each task*, it consistently ranks as the top two algorithms (with ties). This is in contrast to baseline algorithms whose performance rank might vary drastically across tasks. For example, the one-step TD3 performs well in CheetahRun while performs poorly in WalkerWalk. Also, both Ctrace and Retrace generally significantly perform more poorly. We provide further analysis in Appendix J.

Sparse rewards results. In the bottom row of Figure 2, we show evaluations on sparse reward variants of the benchmark tasks. See details on these environments in Appendix J. Sparse rewards are challenging for deep RL algorithms, as it is more difficult to numerically propagate learning signals across time steps. Accordingly, sparse rewards are natural benchmarks for operator-based algorithms. Across all tasks, Peng's Q(λ) consistently outperforms other baselines. In a few cases, uncorrected n-step also outperforms the baseline TD3 – we speculate that this is because the former propagates the learning signal more efficiently, which is critical for sparse rewards. Compared to uncorrected n-step, Peng's Q(λ) seems to achieve a better trade-off between efficient propagation of learning signals and fixed point biases, which leads to relatively stable and consistent performance gains across all selected benchmark tasks.

7.3. Additional deep RL experiments

Maximum-entropy RL. In Appendix I, we show how Peng's Q(λ) could be extended to maximum-entorpy RL (Ziebart et al., 2008; Fox et al., 2016; Haarnoja et al., 2017; 2018). We combine multi-step operators with maximum-entropy deep RL algorithms such as SAC (Haarnoja et al., 2018) and show performance gains over benchmark tasks. See Appendix J for further details.

Ablation study on λ. In Appendix J, we provide an ablation study on the effect of λ. We show that the performance of Peng's Q(λ) depends on the choice of λ. Nevertheless, we find that a single λ can usually lead to fairly uniform

performance gains across a large number of benchmarks.

8. Conclusion

In this paper, we have studied the non-conservative off-policy algorithm Peng's Q(λ), and shown that while in the worst case its convergence guarantees are less strong than conservative algorithms such as Retrace, convergence guarantees to the optimal policy are recovered when the behavior policy closely tracks the target policy. This has important consequences for deep RL theory and practice, as this condition often holds when agents are trained through replay buffers, and serves to close the gap between the strong empirical performance observed with non-conservative algorithms in deep RL, and their previous lack of theory.

We expect this to have several important consequences for deep RL theory and practice. Firstly, these results make clear that the *degree* of off-policyness is an important quantity that has real impact on the success of deep RL algorithms, and incorporating quantities related to this into the analysis of off-policy algorithms will be important for developing theoretical understanding of deep RL. Secondly, these findings add weight to growing empirical work highlighting that quantities such as replay buffer size and replay ratio are crucial to the success of deep RL agents (Zhang & Sutton, 2017; Daley & Amato, 2019; Fedus et al., 2020), and deserve further attention.

We believe the analysis presented in this paper is an important step towards a deeper understanding of non-conservative methods, and there are several open questions suitable for future work. For example, the convergence guarantee in Theorem 4 requires $\alpha \geq 1 - \lambda$. However we conjecture that this assumption can be lifted. Besides, while we did not analyze the concentrability coefficients of PQL, Scherrer (2014) reports that conservative policy iteration, which is analogous to PQL, has a better concentrability coefficients. Finally, careful error propagation analyses of gap-increasing algorithms (Azar et al., 2012; Kozuno et al., 2019) and policy-update-regularized algorithms (Vieillard et al., 2020) show a slow update of policies confer the stability against errors on algorithms. In PQL with behavior policy updates, we expect a similar result when α takes an intermediate value.

Acknowledgement

We are grateful to all reviewers for their time spent on reviewing this paper and writing feedback. TK was supported by JSPS KAKENHI Grant Numbers 16H06563. TK thanks Prof. Kenji Doya, Dongqi Han, and Ho Ching Chiu at Okinawa Institute of Science and Technology (OIST) for their valuable comments. TK is also grateful to the research support of OIST to the Neural Computation Unit, where TK partially conducted this research. In particular, TK is thankful for OIST's Scientific Computation and Data Analysis section, which maintains a cluster we used for many of our experiments. YHT acknowledges the computational support from Google Cloud Platform.

References

Achiam, J. Spinning Up in Deep Reinforcement Learning. 2018.

Asadi, K. and Littman, M. L. An Alternative Softmax Operator for Reinforcement Learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

Azar, M. G., Gómez, V., and Kappen, H. J. Dynamic policy programming. *Journal of Machine Learning Research*, 13(103):3207–3245, 2012.

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., TB, D., Muldal, A., Heess, N., and Lillicrap, T. Distributed distributional deterministic policy gradients. In *Proceedings of the International Conference on Learning Representations*, 2018.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Bertsekas, D. P. and Ioffe, S. Temporal differences-based policy iteration and applications in neuro-dynamic programming. Technical Report LIDS-P-2349, Lab. for Info. and Decision Systems Report, MIT, Cambridge, Massachusetts, 1996.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

Casella, G. and Berger, R. L. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

Coumans, E. and Bai, Y. PyBullet, a Python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016–2019.

Daley, B. and Amato, C. Reconciling λ-returns with experience replay. In *Advances in Neural Information Processing Systems*, 2019.

Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., and Dabney, W. Revisiting fundamentals of experience replay. In *Proceedings of the International Conference on Machine Learning*, 2020.

Fox, R., Pakman, A., and Tishby, N. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.

Fujimoto, S., Van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *Proceedings of the International Conference on Machine Learning*, 2018.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the International Conference on Machine Learning*, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning*, 2018.

Harb, J. and Precup, D. Investigating recurrence and eligibility traces in deep Q-networks. *arXiv preprint arXiv:1704.05495*, 2017.

Harutyunyan, A., Bellemare, M. G., Stepleton, T., and Munos, R. Q(λ) with off-policy corrections. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 2016.

Hasselt, H. V. Double Q-learning. In *Advances in Neural Information Processing Systems*, 2010.

Hessel, M., Modayil, J., van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M. G., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2002.

Kapturowski, S., Ostrovski, G., Quan, J., Munos, R., and Dabney, W. Recurrent experience replay in distributed reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Kozuno, T., Uchibe, E., and Doya, K. Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2019.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2016.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Mousavi, S. S., Schukat, M., Howley, E., and Mannion, P. Applying Q(λ)-learning in deep reinforcement learning to play Atari games. In *AAMAS Workshop on Adaptive Learning Agents*, 2017.

Munos, R. Error bounds for approximate policy iteration. In *Proceedings of the International Conference on Machine Learning*, 2003.

Munos, R. Error bounds for approximate value iteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2005.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, 2016.

Oh, J., Guo, Y., Singh, S., and Lee, H. Self-imitation learning. In *Proceedings of the International Conference on Machine Learning*, 2018.

Peng, J. and Williams, R. J. Incremental multi-step Q-learning. In *Proceedings of the International Conference on Machine Learning*, 1994.

Peng, J. and Williams, R. J. Incremental multi-step Q-learning. *Machine learning*, 22(1):283–290, March 1996.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the International Conference on Machine Learning*, 2000.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Puterman, M. L. and Shin, M. C. Modified policy iteration algorithms for discounted Markov decision problems. *Management Science*, 24(11):1127–1137, 1978.

Rowland, M., Dabney, W., and Munos, R. Adaptive trade-offs in off-policy learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020.

Scherrer, B. Performance bounds for λ policy iteration and application to the game of Tetris. *Journal of Machine Learning Research*, 14(1):1181–1227, 2013.

Scherrer, B. Approximate policy iteration schemes: A comparison. In *Proceedings of the International Conference on Machine Learning*, 2014.

Scherrer, B., Gabillon, V., Ghavamzadeh, M., and Geist, M. Approximate modified policy iteration. In *Proceedings of the International Conference on Machine Learning*, 2012.

Scherrer, B., Ghavamzadeh, M., Gabillon, V., Lesner, B., and Geist, M. Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16:1629–1676, 2015.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, 2014.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 1 edition, 1998.

Tassa, Y., Tunyasuvunakool, S., Muldal, A., Doron, Y., Liu, S., Bohez, S., Merel, J., Erez, T., Lillicrap, T., and Heess, N. dm_control: Software and Tasks for Continuous Control, 2020.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2012.

Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., and Geist, M. Leverage the average: an analysis of KL regularization in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. Sample efficient actor-critic with experience replay. In *Proceedings of the International Conference on Learning Representations*, 2017.

Watkins, C. J. C. H. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, Cambridge, UK, May 1989.

Zhang, S. and Sutton, R. S. A deeper look at experience replay. In *NeurIPS Workshop on Deep Reinforcement Learning*, 2017.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2008.

A. Preliminaries for Theoretical Analyses

In this appendix, we explain important notions we used in our theoretical analyses.

Contraction and Monotonicity of Operators. An operator $\mathcal{O}$ from a normed space $(\mathbf{F}, \|\cdot\|)$ to another normed space $(\mathbf{F}', |\cdot|)$ is said to be a contraction if there is a constant $c \in [0, 1)$ such that $\|\mathcal{O}f - \mathcal{O}g\| \leq c|f - g|$. This constant c is sometimes called as modulus. For example, $\mathcal{T} : (\mathbb{R}^{\mathbf{X}\times\mathbf{A}}, \|\cdot\|_\infty) \to (\mathbb{R}^{\mathbf{X}\times\mathbf{A}}, \|\cdot\|_\infty)$ is a contraction with modulus γ. In the main text, we usually meant a contraction under $\|\cdot\|_\infty$ and did not always mention which norm is considered.

A related notion is a non-expansion. If an operator $\mathcal{O}$ satisfies only $\|\mathcal{O}f - \mathcal{O}g\| \leq |f - g|$, it is said to be a non-expansion. For example, $\mathcal{P}$ is a non-expansion, as proven later.

Monotonicity is probably the most important property in our analyses. An operator $\mathcal{O}$ is said to be monotone if $\mathcal{O}f \geq \mathcal{O}g$ for any f and g satisfying $f \geq g$. For example, $\mathcal{P}$ is monotone: if $V \geq V'$ (point-wisely, i.e., $V(x) \geq V'(x)$ at every x), $\mathcal{P}V - \mathcal{P}V'$ holds too, as one can easily confirm from

$$(\mathcal{P}V - \mathcal{P}V')(x, a) = \mathbb{E}[V(X_1) - V'(X_1)|X_0 = x, A_0 = a] \geq 0.$$

Let $\mathbf{1} \in \mathbf{F}$ be a constant function taking 1 everywhere. If a linear operator $\mathcal{O} : (\mathbf{F}, \|\cdot\|_\infty) \to (\mathbf{F}', \|\cdot\|_\infty)$ is monotone and satisfies $\mathcal{O}\mathbf{1} = c\mathbf{1}$ with a scalar c, we have $\|\mathcal{O}f - \mathcal{O}g\|_\infty \leq c\|f - g\|_\infty$. Indeed,

$$\mathcal{O}f - \mathcal{O}g = \mathcal{O}(f - g) \leq \mathcal{O}\|f - g\|_\infty\mathbf{1} = c\|f - g\|_\infty\mathbf{1} \text{ and } \mathcal{O}f - \mathcal{O}g \geq -c\|f - g\|_\infty\mathbf{1}$$

imply $\|\mathcal{O}f - \mathcal{O}g\|_\infty \leq c\|f - g\|_\infty$. Thus, $\mathcal{P}$ is non-expansive as $\mathcal{P}\mathbf{1} = \mathbf{1}$. Note that $(1 - \gamma\lambda)(\mathcal{I} - \gamma\lambda\mathcal{P}^\pi)^{-1}$ is also a non-expansive operator for any π, as one can easily confirm.

B. On an Extension of Theoretical Results to Continuous Action Spaces

In this appendix, we explain how to extend our theoretical results to a case where both the state and action spaces are continuous. We mainly follow Appendix B in (Puterman, 1994). We ask interested readers to refer to the textbook.

Notation. Let $\mathbf{S}$ and $\mathbf{S}'$ be Polish spaces. We denote by $\mathbf{B}(\mathbf{S}; c)$ the set of all Borel-measurable functions from $\mathbf{S}$ to a bounded closed interval $[-c, c]$, where $c \in [0, \infty)$; throughout this appendix, the Borel σ-algebra is always considered. We denote by $\mathbf{P}(\mathbf{S})$ the set of all Borel probability measures on $\mathbf{S}$. We say that a real-valued function f on $\mathbf{S}$ is upper semicontinuous (usc) at a point p^* if $\limsup_{n\to\infty} f(p_n) \leq f(p^*)$ for any sequence of points $(p_n)_{n\geq 0}$ converging to p^*. We say that f is usc if it is usc at any point. We denote by $\mathbf{U}(\mathbf{S}; c)$ the set of all usc functions from $\mathbf{S}$ to a bounded closed interval $[-c, c]$, where $c \in [0, \infty)$. We say that a stochastic kernel $q : \mathbf{S} \to \mathbf{P}(\mathbf{S}')$ is continuous if $\lim_{n\to\infty} \int f(p')q(dp'|p_n) = \int f(p')q(dp'|p)$ for any bounded continuous function f and any sequence of points $(p_n)_{n\geq 0}$ converging to p.

Main Discussion. We impose the following assumption on MDPs. It is necessary to guarantee that all functions in the analyses are usc, as we shall explain soon.

Assumption 6. *The state and action spaces are compact subsets of finite-dimensional Euclidean spaces equipped with Borel σ-algebras. The reward function r is an usc function bounded by r_{max}, and the state transition probability kernel $\mathcal{P}$ is continuous.*

We first explain that there exists an optimal policy that is a measurable function from the state space $\mathbf{X}$ to the action space $\mathbf{A}$. Let $V_{max} := r_{max}/(1 - \gamma)$. We denote by $\mathcal{M} : \mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max}) \to \mathbf{U}(\mathbf{X}; V_{max})$ the max operator defined by $(\mathcal{M}Q)(x) := \max_{a\in\mathbf{A}} Q(x, a)$ for any $Q \in \mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$. Theorem B.5 in Puterman (1994) guarantees that $\mathcal{M}Q$ is usc. Furthermore, Proposition B.4 in Puterman (1994) guarantees that $\mathcal{P}\mathcal{M}Q$ is usc. It is easy to confirm that both $\mathcal{M}Q$ and $\mathcal{P}\mathcal{M}Q$ are bounded by V_{max}. Since a sum of usc functions is again usc (Puterman, 1994, Proposition B.1.a), $r + \gamma\mathcal{P}\mathcal{M}Q = \mathcal{T}Q$ belongs to $\mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$. Suppose the recursion $Q_{k+1} := \mathcal{T}Q_k$. Proposition B.1.e in Puterman (1994) guarantees that $\lim_{k\to\infty} Q_k = Q^*$ is usc. Proposition B.4 in Puterman (1994) guarantees that there exists a measurable function $\pi_* : \mathbf{X} \to \mathbf{A}$ such that $Q^*(x, \pi_*(x)) = \max_{a\in\mathbf{A}} Q^*(x, a)$. Accordingly, there exists an optimal policy that is a measurable function from $\mathbf{X}$ to $\mathbf{A}$.

From the above discussion, it is easy to confirm that all Q_k in the exact version of PQL (3) belong to $\mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$ given that the behavior policy μ is continuous. Therefore, the proof of Theorem 2 in Appendix E is valid under the assumption that μ is continuous. We note that it is a weak assumption because the behavior policy μ is often continuous in practice. Indeed, an action distribution $\mu(\cdot|x)$ is frequently a normal distribution whose mean and diagonal covariance matrix are continuous functions of a state x expressed by, for example, neural networks. As a result, as long as all elements of the diagonal covariance matrix are bounded from below by some constant, the probability density function of $\mu(\cdot|x)$ is bounded. Therefore, the dominated convergence theorem can be used to show that μ is continuous. When there is an element of the diagonal covariance matrix converging to 0, this argument does not hold. However, it is a pathological case that usual implementations, such as SpinningUp (Achiam, 2018), try to avoid by value clipping.

For other theoretical results, we need two additional assumptions: (i) all behavior policies μ and μ_k are continuous, and (ii) all error functions ε_k belong to $\mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$. As for the assumption (i), it is a weak assumption as noted above. (See also the following paragraph on the relaxation of π_k's exact greediness.) As for the assumption (ii), it is also a weak assumption: because Q_{k+1} approximates $(\mathcal{N}_\lambda^\mu)^k Q_0 \in \mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$, there is no strong reason to use a function approximator that does not belong to $\mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$; using a function approximator belonging to $\mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$ guarantees that $\varepsilon_k = Q_{k+1} - \mathcal{N}_\lambda^\mu Q_k$ belongs to $\mathbf{U}(\mathbf{X} \times \mathbf{A}; V_{max})$. Similar arguments can be made even when the behavior policy is updated, and we can conclude that these assumptions are weak.

We finally mention how to relax the exact greedy assumption that $\pi_k \in \mathbf{G}(Q_k)$. When the action space is continuous, it is not feasible to find an exact greedy policy even if Q_k is continuous. In addition, it is often the case that a policy is expressed by a neural network. However, it is relatively straightforward to extend our theoretical analyses to a case where this exact greedy assumption is relaxed to a δ_k-greedy assumption, that is, $\pi_k Q_k \geq \mathcal{M} Q_k - \delta_k$, where $\delta_k \in \mathbf{U}(\mathbf{X}; V_{max})$. A similar near-greedy condition is found in, for example, Scherrer (2014).

C. A Proof of Lemma 1 (Different Forms of the PQL Operator)

In this appendix, we prove Lemma 1, which provides the following forms of the PQL operator:

$$\begin{aligned}\mathcal{N}_\lambda^{\mu,\pi} Q &= Q + (\mathcal{I} - \gamma\lambda\mathcal{P}^\mu)^{-1}\Big(\mathcal{T}^{\lambda\mu+(1-\lambda)\pi} Q - Q\Big) \\ &= (\mathcal{I} - \gamma\lambda\mathcal{P}^\mu)^{-1}(r + \gamma(1-\lambda)\mathcal{P}^\pi Q).\end{aligned}$$

We first recall the original PQL operator (1): $\mathcal{N}_\lambda^{\mu,\pi} Q := (1-\lambda)\sum_{n=0}^\infty \lambda^n (\mathcal{T}^\mu)^n \mathcal{T}^\pi Q$. Note that each term in the sum can be rewritten as $(\mathcal{T}^\mu)^n \mathcal{T}^\pi Q = \sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r + \gamma^{n+1}(\mathcal{P}^\mu)^n \mathcal{P}^\pi Q$. Therefore,

$$\begin{aligned}(1-\lambda)\sum_{n=0}^\infty \lambda^n (\mathcal{T}^\mu)^n \mathcal{T}^\pi Q &= (1-\lambda)\sum_{n=0}^\infty \lambda^n \left[\sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r + \gamma^{n+1}(\mathcal{P}^\mu)^n \mathcal{P}^\pi Q\right] \\ &= (1-\lambda)\sum_{n=0}^\infty \lambda^n \sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r + \sum_{n=0}^\infty \lambda^n \gamma^{n+1}(1-\lambda)(\mathcal{P}^\mu)^n \mathcal{P}^\pi Q.\end{aligned}$$

Note that

$$\begin{aligned}(1-\lambda)\sum_{n=0}^\infty \lambda^n \sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r &= \sum_{n=0}^\infty \lambda^n \sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r - \sum_{n=0}^\infty \lambda^{n+1} \sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r \\ &= \sum_{n=0}^\infty \lambda^n \sum_{m=0}^n \gamma^m (\mathcal{P}^\mu)^m r - \sum_{n=1}^\infty \lambda^n \sum_{m=0}^{n-1} \gamma^m (\mathcal{P}^\mu)^m r \\ &= \sum_{n=0}^\infty \lambda^n \gamma^n (\mathcal{P}^\mu)^n r.\end{aligned}$$

Consequently,

$$(1-\lambda)\sum_{n=0}^\infty \lambda^n (\mathcal{T}^\mu)^n \mathcal{T} Q = \sum_{n=0}^\infty \lambda^n \gamma^n (\mathcal{P}^\mu)^n (r + \gamma(1-\lambda)\mathcal{P}^\pi Q) = (\mathcal{I} - \gamma\lambda\mathcal{P}^\mu)^{-1}(r + \gamma(1-\lambda)\mathcal{P}^\pi Q).$$

The right hand side can be rewritten as follows:

$$\begin{aligned}(\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(r+\gamma(1-\lambda)\mathcal{P}^{\pi}Q) &= (\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(\lambda\mathcal{T}^{\mu}Q+(1-\lambda)\mathcal{T}^{\pi}Q-\lambda\mathcal{P}^{\mu}Q)\\ &= (\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(\lambda\mathcal{T}^{\mu}Q+(1-\lambda)\mathcal{T}^{\pi}Q-Q+(\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})Q)\\ &= Q+(\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(\lambda\mathcal{T}^{\mu}Q+(1-\lambda)\mathcal{T}^{\pi}Q-Q).\end{aligned}$$

This concludes the proof.

D. A Proof of Proposition 5 (HQL's Oscillation)

In this appendix, we prove that under a certain circumstance, HQL oscillates. We prove it by using an example shown in Figure 3. In this MDP, there are two types of states $\mathbf{X}_1 = \{x|x = 1, 2, \ldots\}$ and $\mathbf{X}_2 = \{x'|x = 1, 2, \ldots\}$. We denote a state in $\mathbf{X}_1$ by x and a state in $\mathbf{X}_2$ by x'. There are two actions go and $exit$. When an agent chooses go at x, it moves to $x + 1$ with a reward of -1. When an agent chooses $exit$ at x, it moves to x' with a reward of 1. At x', any action results in a state transition to the same state x' with a reward of 1. Therefore, an agent must $exit$ from x as soon as possible.

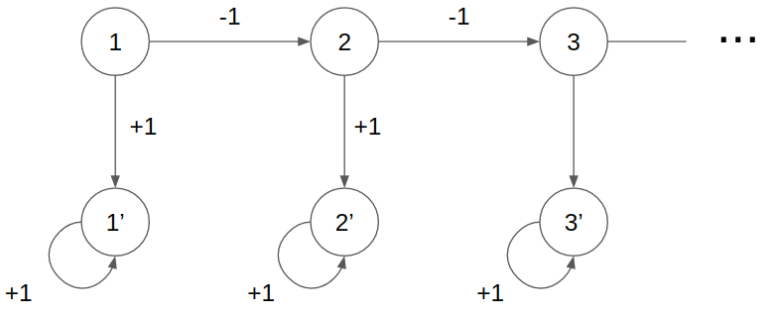


Figure 3. An MDP in which HQL may oscillate.

We assume that $\lambda = 1$, $\gamma > 0.5$, μ chooses go everywhere, $Q_0(x, exit) = Q^{\mu}(x, exit) = 1/(1-\gamma)$, and $Q_0(x, go) = Q^{\mu}(x, go) + \delta = -1/(1-\gamma) + \delta$ with $\delta > 2/(1-\gamma)$. For other state-action pairs, $Q_0 = Q^{\mu}$. As a result, $\pi_0 = \mu$. (At a state $x' \in \mathbf{X}$, any policy is effectively the same as μ.)

Step 1. HQL's update can be rewritten as follows (Harutyunyan et al., 2016):

$$Q_{k+1} := Q_k + (\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(\mathcal{T}^{\pi_k}Q_k - Q_k) = Q_k + (\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(\mathcal{T}^{\mu}Q_k - Q_k + \gamma\mathcal{P}(\pi_k-\mu)Q_k).$$

Since $\pi_0 = \mu$, and μ chooses go everywhere (that is, $A_t = go$ for every t in the following equations), we deduce that

$$Q_1(x, go) = Q_0(x, go) + \sum_{t=0}^{\infty}\gamma^t\mathbb{E}[(\mathcal{T}^{\mu}Q_0)(X_t, A_t) - Q_0(X_t, A_t)|X_0 = x, A_0 = go, \mu] = Q^{\mu}(x, go) = -\frac{1}{1-\gamma}.$$

Besides, $Q_1(x, exit) = Q_0(x, exit) = Q^{\mu}(x, exit) = 1/(1-\gamma)$. Accordingly, $\arg\max_a Q_1(x, a) = exit$, and $Q_1 = Q^{\mu}$.

Step 2. Let us consider what happens at the next iteration. Since μ chooses go everywhere (that is, $A_t = go$ for every t in the following equations), we deduce that

$$\begin{aligned}Q_2(x, go) &= Q_1(x, go) + \sum_{t=0}^{\infty}\gamma^t\mathbb{E}[(\mathcal{T}^{\mu}Q_1)(X_t, A_t) - Q_1(X_t, A_t) + \gamma(\pi_1Q_1 - \mu Q_1)(X_{t+1})|X_0 = x, A_0 = go, \mu]\\ &= Q^{\mu}(x, go) + \sum_{t=0}^{\infty}\gamma^t\mathbb{E}[\gamma(\pi_1Q_1 - \mu Q_1)(X_{t+1})|X_0 = x, A_0 = go, \mu]\\ &= Q^{\mu}(x, go) + \frac{2\gamma}{(1-\gamma)^2} > \frac{1}{1-\gamma} = Q^{\mu}(x, exit).\end{aligned}$$

Besides, $Q_2(x, exit) = Q^{\mu}(x, exit) = 1/(1-\gamma)$. Accordingly, $\arg\max_a Q_1(x, a) = go$.

Step 3. Now, note that by setting δ in Step 1 to be $2\gamma/(1-\gamma)^2$, the situation is completely the same as the one we considered in Step 1. Accordingly, $\arg\max_a Q_3(x,a) = exit$, and $Q_3 = Q^\mu$. The situation of the next iteration is completely the same as the one we considered in Step 2. This argument can be repeated forever, and thus, Q_k (as well as π_k) oscillates.

E. A Proof of Theorem 2 (PQL's Convergence with a Fixed Behavior Policy)

We define $\mathcal{N}_\lambda^\mu$ as an operator such that $\mathcal{N}_\lambda^\mu Q = \mathcal{N}_\lambda^{\mu,\pi_Q} Q$ for any $Q \in \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$, where $\pi_Q \in \mathbf{G}(Q)$. This operator is analogous to $\mathcal{T}$, whereas $\mathcal{N}_\lambda^{\mu,\pi}$ is analogous to $\mathcal{T}^\pi$.

From Lemma 1, we deduce that $\mathcal{N}_\lambda^\mu Q - \mathcal{N}_\lambda^\mu Q' = (1-\lambda)(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}(\mathcal{T}Q - \mathcal{T}Q')$ for any $Q, Q' \in \mathbb{R}^{\mathbf{X}\times\mathbf{A}}$. Because $(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}$ is linear and monotonic, and satisfies $(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}\mathbf{1} = \mathbf{1}/(1-\gamma\lambda)$, we have that $\|\mathcal{N}_\lambda^\mu Q - \mathcal{N}_\lambda^\mu Q'\|_\infty \le (1-\lambda)\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty/(1-\gamma\lambda)$. As noted in Appendix A, $\mathcal{T}$ is a contraction with modulus γ. Therefore, $\|\mathcal{N}_\lambda^\mu Q - \mathcal{N}_\lambda^\mu Q'\|_\infty \le \gamma(1-\lambda)\|Q-Q'\|_\infty/(1-\gamma\lambda) = \beta\|Q-Q'\|_\infty$. Combining this with Banach's fixed point theorem (Puterman, 1994), it is proven that PQL with a fixed behavior policy converges to a unique fixed point with the rate β^k.

Let Q_{fixed} and π_{fixed} be the fixed point and a greedy policy with respect to the fixed point, respectively. (It will turn out to be $Q_{\text{fixed}} = Q^{\lambda\mu+(1-\lambda)\pi_\dagger}$ and $\pi_{\text{fixed}} = \pi_\dagger$.) As noted in Section 5.1, Q_{fixed} is the fixed point of $\lambda\mathcal{T}^\mu + (1-\lambda)\mathcal{T}$. It is easy to confirm that it is also the fixed point of $\mathcal{T}^{\lambda\mu+(1-\lambda)\pi_{\text{fixed}}}$ as $\pi_{\text{fixed}} \in \mathbf{G}(Q_{\text{fixed}})$. Therefore, $Q_{\text{fixed}} = Q^{\lambda\mu+(1-\lambda)\pi_{\text{fixed}}}$.

As $\pi_{\text{fixed}} \in \mathbf{G}(Q_{\text{fixed}})$, $Q_{\text{fixed}} = \mathcal{T}^{\lambda\mu+(1-\lambda)\pi_{\text{fixed}}} Q_{\text{fixed}} \ge \mathcal{T}^{\lambda\mu+(1-\lambda)\pi} Q_{\text{fixed}}$ for any policy π. Therefore, for any positive integer n, we have that $Q_{\text{fixed}} \ge (\mathcal{T}^{\lambda\mu+(1-\lambda)\pi})^n Q_{\text{fixed}}$. As a result, $Q^{\lambda\mu+(1-\lambda)\pi_{\text{fixed}}} = Q_{\text{fixed}} \ge Q^{\lambda\mu+(1-\lambda)\pi}$ for any π. This implies that π_{fixed} is $\pi_\dagger$.

F. Double-loop PQL

In this appendix, we analyze PQL in which $\mathcal{N}_\lambda^{\mu_k}$ is applied multiple times to Q_k, and then, the current behavior policy μ_k is updated to μ_{k+1}. (See Appendix E for the definition of $\mathcal{N}_\lambda^\mu$.) Concretely, we consider the following algorithm:

$$\mu_k \in \mathbf{G}_{\delta_k}(Q_k)\ \text{, and}\ Q_{k+1} := (\mathcal{N}_\lambda^{\mu_k})^\infty Q_k + \varepsilon_k\,, \tag{5}$$

where $\delta_k \in \mathbb{R}^{\mathbf{X}}$ is a non-negative function over $\mathbf{X}$, and $\mathbf{G}_{\delta_k}(Q_k)$ is the set of δ_k-greedy policies π defined by $\pi Q_k \ge \pi' Q_k - \delta_k$ for a greedy policy $\pi' \in \mathbf{G}(Q_k)$. Here, we used a shorthand notation $(\mathcal{N}_\lambda^{\mu_k})^\infty Q_k := \lim_{n\to\infty}(\mathcal{N}_\lambda^{\mu_k})^n Q_k$. Note that this algorithm involves a double-loop structure: in the inner loop $\mathcal{N}_\lambda^{\mu_k}$ is repeatedly applied to $(\mathcal{N}_\lambda^{\mu_k})^n Q_k$, and in the outer loop the Q-function and policies are updated. Hence, we call this algorithm as a doule-loop PQL.

There are two main differences from approximate PQL with behavior policy updates (4): first, the behavior policy is required to be near-greedy rather than a mixture policy; second, the Q-function is updated to $(\mathcal{N}_\lambda^{\mu_k})^\infty Q_k + \varepsilon_k$ rather than $\mathcal{N}_\lambda^{\mu_k} Q_k + \varepsilon_k$. As for the first difference, we think that the behavior policy update in (4) is more practical, but we are unsure if Theorem 4 can be extended to double-loop PQL. As for the second difference, this Q-function update is an abstraction of a situation where $\mathcal{N}_\lambda^{\mu_k}$ is applied only finitely many times, and Q_{k+1} deviates from $(\mathcal{N}_\lambda^{\mu_k})^\infty Q_k$ as a result. Because it is impossible to compute $(\mathcal{N}_\lambda^{\mu_k})^\infty Q_k$ in a practical situation, this abstraction is necessary. We note that other errors such as function approximation errors can be also included to ε_k.

For this algorithm, we have the following guarantee.

Proposition 7. *For any non-negative integer k, the following holds:*

$$Q^* - Q^{\mu_{k+1}} \le \frac{2\gamma}{1-\gamma}\sum_{j=0}^{k}\gamma^{k-j}\|\varepsilon_j\|_\infty \mathbf{1} + \frac{\gamma(1+\gamma)}{1-\gamma}\sum_{j=0}^{k}\gamma^{k-j}\|\delta_{j+1}\|_\infty \mathbf{1} + \gamma^{k+1}\|Q^* - Q^{\mu_0}\|_\infty\,.$$

Thus, if $\|\delta_k\|_\infty \to 0$ and $\|\varepsilon_k\|_\infty \to 0$, then $Q^{\mu_k} \to Q^*$.

Proof. First let us prove that $Q^{\mu_k} - \|\varepsilon_k\|_\infty \mathbf{1} \le Q_{k+1} \le Q^{\mu_{k+1}} + \frac{1+\gamma}{1-\gamma}\|\varepsilon_k\|_\infty \mathbf{1} + \frac{\gamma}{1-\gamma}\|\delta_k\|_\infty \mathbf{1}$. By definition of Q_{k+1},

$$Q_{k+1} - \varepsilon_k = (\mathcal{N}_\lambda^{\mu_k})^\infty Q_k \ge Q^{\lambda\mu_k+(1-\lambda)\pi} \implies Q_{k+1} \ge Q^{\lambda\mu_k+(1-\lambda)\pi} - \|\varepsilon_k\|_\infty \mathbf{1}$$

for any policy π, where the first inequality follows from Theorem 2. Now, setting $\pi = \mu_k$ yields $Q_{k+1} \geq Q^{\mu_k} - \|\varepsilon_k\|_\infty \mathbf{1}$. Next, recall that $Q_{k+1} - \varepsilon_k = (\mathcal{N}_\lambda^{\mu_k})^\infty Q_k$ is a fixed point of $\lambda \mathcal{T}^{\mu_k} + (1-\lambda)\mathcal{T}$. Accordingly,

$$Q_{k+1} - \varepsilon_k = \lambda \mathcal{T}^{\mu_k}(Q_{k+1} - \varepsilon_k) + (1-\lambda)\mathcal{T}(Q_{k+1} - \varepsilon_k) \leq \mathcal{T}(Q_{k+1} - \varepsilon_k) \leq \mathcal{T} Q_{k+1} + \gamma \|\varepsilon_k\|_\infty \mathbf{1}\,,$$

where the last inequality follows from the monotonicity of $\mathcal{T}$ and $-\varepsilon_k \leq \|\varepsilon_k\|_\infty$. Furthermore, from the fact that $\mu_{k+1} \in \mathbf{G}_{\delta_{k+1}}(Q_{k+1})$, we deduce that $Q_{k+1} - \varepsilon_k \leq \mathcal{T}^{\mu_{k+1}} Q_{k+1} + \gamma \|\varepsilon_k\|_\infty \mathbf{1} + \gamma \|\delta_k\|_\infty \mathbf{1}$. This implies that $Q_{k+1} \leq \mathcal{T}^{\mu_{k+1}} Q_{k+1} + (1+\gamma)\|\varepsilon_k\|_\infty \mathbf{1} + \gamma \|\delta_k\|_\infty \mathbf{1}$. By induction on k and the monotinicity of $\mathcal{T}^{\mu_{k+1}}$, we deduce that

$$Q_{k+1} \leq Q^{\mu_{k+1}} + \frac{1+\gamma}{1-\gamma}\|\varepsilon_k\|_\infty \mathbf{1} + \frac{\gamma}{1-\gamma}\|\delta_k\|_\infty \mathbf{1}\,.$$

Now we have

$$\begin{aligned}
Q^* - Q^{\mu_{k+1}} &= \gamma \mathcal{P}^{\pi_*} Q^* - \gamma \mathcal{P}^{\pi_*} Q_{k+1} + \underbrace{\gamma \mathcal{P}^{\pi_*} Q_{k+1} - \gamma \mathcal{P}^{\mu_{k+1}} Q_{k+1}}_{\leq \gamma \|\delta_{k+1}\|_\infty \mathbf{1}} + \gamma \mathcal{P}^{\mu_{k+1}} Q_{k+1} - \gamma \mathcal{P}^{\mu_{k+1}} Q^{\mu_{k+1}} \\
&\leq \gamma \mathcal{P}^{\pi_*}(Q^* - Q_{k+1}) + \gamma \|\delta_{k+1}\|_\infty \mathbf{1} + \gamma \mathcal{P}^{\mu_{k+1}} \underbrace{(Q_{k+1} - Q^{\mu_{k+1}})}_{\leq \frac{1+\gamma}{1-\gamma}\|\varepsilon_k\|_\infty \mathbf{1} + \frac{\gamma}{1-\gamma}\|\delta_{k+1}\|_\infty \mathbf{1}} \\
&\leq \frac{\gamma(1+\gamma)}{1-\gamma}\|\varepsilon_k\|_\infty \mathbf{1} + \frac{\gamma(1+\gamma)}{1-\gamma}\|\delta_{k+1}\|_\infty \mathbf{1} + \gamma \mathcal{P}^{\pi_*}(Q^* - Q_{k+1}) \\
&\leq \frac{2\gamma}{1-\gamma}\|\varepsilon_k\|_\infty \mathbf{1} + \frac{\gamma(1+\gamma)}{1-\gamma}\|\delta_{k+1}\|_\infty \mathbf{1} + \gamma \mathcal{P}^{\pi_*}(Q^* - Q^{\mu_k})
\end{aligned}$$

By induction on k, we see that

$$Q^* - Q^{\mu_{k+1}} \leq \frac{2\gamma}{1-\gamma}\sum_{j=0}^{k} \gamma^{k-j}\|\varepsilon_j\|_\infty \mathbf{1} + \frac{\gamma(1+\gamma)}{1-\gamma}\sum_{j=0}^{k}\gamma^{k-j}\|\delta_{j+1}\|_\infty \mathbf{1} + (\gamma \mathcal{P}^{\pi_*})^{k+1}(Q^* - Q^{\mu_0})\,.$$

By upper-bounding $Q^* - Q^{\mu_0}$ by $\|Q^* - Q^{\mu_0}\|_\infty$, the claimed result is obtained. □

G. A Proof of Theorem 3 (PQL's Error Propagation with a Fixed Behavior Policy)

Here, we provide the error propagation analysis of PQL with a fixed behavior policy. While the behavior policy is not fixed in a practical situation, the error propagation analysis of PQL with a fixed behavior policy shows the trade-off between bias and convergence rate of PQL. This result is analogous to trade-offs explained in (Rowland et al., 2020) and sheds some light on a fundamental property of PQL.

Definition and Notation. We first recall our problem setting: (approximate) PQL updates its Q-function by

$$\pi_k \in \mathbf{G}(Q_k) \text{ and } Q_{k+1} := \mathcal{N}_\lambda^{\mu, \pi_k} Q_k + \varepsilon_k,$$

We know that $\varepsilon_k(x,a) = 0$ guarantees the convergence of Q_k to $Q^{\lambda\mu + (1-\lambda)\pi_\dagger}$, where $\pi_\dagger$ is $\arg\max_\pi Q^{\lambda\mu+(1-\lambda)\pi}$. (See Section 5.1.) Therefore, π_K is an approximation of $\pi_\dagger$, and thus, it is natural to define a loss of using the policy π_k rather than $\pi_\dagger$ by $V^{\rho_\dagger} - V^{\rho_K}$.

We define the following notations:

- $\rho_\dagger := \lambda\mu + (1-\lambda)\pi_\dagger$
- $\rho_k := \lambda\mu + (1-\lambda)\pi_k$
- $d_k := Q^{\rho_\dagger} - Q_k$
- $b_k := Q_k - \mathcal{T}^{\rho_k} Q_k$

- $\mathcal{A}^\dagger := \gamma(1-\lambda)(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}\mathcal{P}^{\pi_\dagger}$
- $\mathcal{A}_k := \gamma(1-\lambda)\mathcal{P}^{\pi_k}(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}$
- $\beta := \gamma(1-\lambda)/(1-\gamma\lambda)$

Note that $\mathcal{A}_k$ is a contraction with respect to L_∞-norm $\|\cdot\|_\infty$ with modulus β. (See Appendix C.)

Proofs. Now we start proofs. The main strategy is the following: we first decompose $V^{\rho_\dagger} - V^{\rho_K}$ (≥ 0 since $\pi_\dagger$ is a policy such that $Q^{\lambda\mu+(1-\lambda)\pi_\dagger} \geq Q^{\lambda\mu+(1-\lambda)\pi}$ for any policy π) to $V^{\rho_\dagger} - \rho_K Q_K$ and $\rho_K Q_K - V^{\rho_K} = \rho_K(Q_K - Q^{\rho_K})$; then we note that $V^{\rho_\dagger} - \rho_K Q_K \leq \rho_\dagger(Q^{\rho_\dagger} - Q_K)$ because of $\pi_K \in \mathbf{G}(Q_K)$ and $\rho_K = \lambda\mu + (1-\lambda)\pi_K$; these results tell us that we need upper bounds of $Q^{\rho_\dagger} - Q_K$ and $Q_K - Q^{\rho_K}$, which we shall derive.

We first prove an upper bound of $d_K = Q^{\rho_\dagger} - Q_K$.

Lemma 8. *For any non-negative integer K, the following holds:*

$$d_K \leq \left(\mathcal{A}^\dagger\right)^K d_0 + \sum_{k=0}^{K-1} \left(\mathcal{A}^\dagger\right)^{K-k-1} \varepsilon_k \,,$$

where $\sum_{l=0}^{-1} f_l := 0\mathbf{1}$ for any sequence of functions $(f_l)_{l\geq 0}$.

Proof. From Lemma 1, we may deduce that

$$\begin{aligned}
d_K &= Q^{\rho_\dagger} - (\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}(r + \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}Q_{K-1}) - \varepsilon_{K-1} \\
&= (\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}[Q^{\rho_\dagger} - \gamma\lambda\mathcal{P}^\mu Q^{\rho_\dagger} - r - \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}Q_{K-1}] - \varepsilon_{K-1} \\
&= (\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}(\gamma\mathcal{P}^{\rho_\dagger}Q^{\rho_\dagger} - \gamma\lambda\mathcal{P}^\mu Q^{\rho_\dagger} - \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}Q_{K-1}) - \varepsilon_{K-1} \\
&= \gamma(1-\lambda)(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}(\mathcal{P}^{\pi_\dagger}Q^{\rho_\dagger} - \mathcal{P}^{\pi_{K-1}}Q_{K-1}) - \varepsilon_{K-1} \,,
\end{aligned}$$

where the last line follows from $\rho_\dagger = \lambda\mu + (1-\lambda)\pi_\dagger$. Because $\pi_{K-1} \in \mathbf{G}(Q_{K-1})$, we have $\mathcal{P}^{\pi_{K-1}}Q_{K-1} \geq \mathcal{P}^{\pi_\dagger}Q_{K-1}$. Furthermore, since $(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1} = \sum_{t=0}^\infty \gamma^t\lambda^t(\mathcal{P}^\mu)^t$ is monotone, $(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}\mathcal{P}^{\pi_{K-1}}Q_{K-1} \geq (\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}\mathcal{P}^{\pi_\dagger}Q_{K-1}$. As a result,

$$d_K \leq \gamma(1-\lambda)(\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)^{-1}\mathcal{P}^{\pi_\dagger}(Q^{\rho_\dagger} - Q_{K-1}) - \varepsilon_{K-1} = \mathcal{A}^\dagger d_{K-1} - \varepsilon_{K-1} \,.$$

By induction on K, the claim is proven. □

We next prove an upper bound for $Q_K - Q^{\rho_K}$. To this end, note that

$$Q_K - Q^{\rho_K} = (\mathcal{I}-\gamma\mathcal{P}^{\rho_K})^{-1}(Q_K - \mathcal{T}^{\rho_K}Q_K) = (\mathcal{I}-\gamma\mathcal{P}^{\rho_K})^{-1}b_K.$$

Therefore, we need an upper bound for b_K, which is given below.

Lemma 9. *For any non-negative integer K, the following holds:*

$$b_K \leq \mathcal{A}_{K-1}\cdots\mathcal{A}_0 b_0 + \sum_{k=0}^{K-1} \mathcal{A}_{K-1}\cdots\mathcal{A}_{k+1}(\mathcal{I}-\gamma\mathcal{P}^{\rho_k})\varepsilon_k \,,$$

where $\mathcal{A}_{K-1}\cdots\mathcal{A}_K := \mathcal{I}$ and $\sum_{l=0}^{-1} f_l := 0\mathbf{1}$ for any sequence of functions $(f_l)_{l\geq 0}$..

Proof. By a simple calculation, and $\pi_K \in \mathbf{G}(Q_K)$,

$$b_K = Q_K - r - \gamma\lambda\mathcal{P}^\mu Q_K - \gamma(1-\lambda)\mathcal{P}^{\pi_K}Q_K \leq (\mathcal{I}-\gamma\lambda\mathcal{P}^\mu)Q_K - r - \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}Q_K \,.$$

From Lemma 1, we may deduce that

$$\begin{aligned} b_K &\leq \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}Q_{K-1} + (\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})\varepsilon_{K-1} - \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}Q_K \\ &= \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}(Q_{K-1}-Q_K) + (\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})\varepsilon_{K-1} \\ &= \gamma(1-\lambda)\mathcal{P}^{\pi_{K-1}}(\mathcal{I}-\gamma\lambda\mathcal{P}^{\mu})^{-1}(Q_{K-1}-\mathcal{T}^{\rho_{K-1}}Q_{K-1}) + (\mathcal{I}-\gamma\mathcal{P}^{\rho_{K-1}})\varepsilon_{K-1} \\ &= \mathcal{A}_{K-1}b_{K-1} + (\mathcal{I}-\gamma\mathcal{P}^{\rho_{K-1}})\varepsilon_{K-1}\,. \end{aligned}$$

By induction on K, the claim is proven. □

Now we are ready to prove an upper bound for $V^{\rho_\dagger} - V^{\rho_K}$. It is easy to derive the following two inequalities from the monotonicity of $\mathcal{A}_k$ and $\mathcal{A}^\dagger$:

$$\begin{aligned} b_K &\leq \mathcal{A}_{K-1}\cdots\mathcal{A}_0\|b_0\|_\infty\mathbf{1} + (1+\gamma)\sum_{k=0}^{K-1}\mathcal{A}_{K-1}\cdots\mathcal{A}_{k+1}\|\varepsilon_k\|_\infty\mathbf{1} \\ &= \beta^K\|b_0\|_\infty\mathbf{1} + (1+\gamma)\sum_{k=0}^{K-1}\beta^{K-k-1}\|\varepsilon_k\|_\infty\mathbf{1}\,, \end{aligned}$$

and

$$\begin{aligned} d_K &\leq \left(\mathcal{A}^\dagger\right)^K\|d_0\|_\infty\mathbf{1} + \sum_{k=0}^{K-1}\left(\mathcal{A}^\dagger\right)^{K-k-1}\|\varepsilon_k\|_\infty\mathbf{1} \\ &= \beta^K\|d_0\|_\infty\mathbf{1} + \sum_{k=0}^{K-1}\beta^{K-k-1}\|\varepsilon_k\|_\infty\mathbf{1}\,. \end{aligned}$$

Note that

$$\begin{aligned} V^{\rho_\dagger} - V^{\rho_K} &= \rho_\dagger Q^{\rho_\dagger} - \rho_K Q_K + \rho_K Q_K - V^{\rho_K} \\ &\leq \rho_\dagger(Q^{\rho_\dagger} - Q_K) + \rho_K Q_K - V^{\rho_K} \\ &= \rho_\dagger d_K + \rho_K(\mathcal{I}-\gamma\mathcal{P}^{\rho_K})^{-1}b_K. \end{aligned}$$

Therefore, we may deduce that

$$V^{\rho_\dagger} - V^{\rho_K} \leq \beta^K\left(\|d_0\|_\infty + \frac{\|b_0\|_\infty}{1-\gamma}\right)\mathbf{1} + \frac{2}{1-\gamma}\sum_{k=0}^{K-1}\beta^{K-k-1}\|\varepsilon_k\|_\infty\mathbf{1},$$

where we used $1+(1+\gamma)/(1-\gamma) = 2/(1-\gamma)$. Because $V^{\rho_\dagger} - V^{\rho_K} \geq 0$ and the right hand side is independent of a state,

$$\|V^{\rho_\dagger} - V^{\rho_K}\|_\infty \leq \beta^K\left(\|d_0\|_\infty + \frac{\|b_0\|_\infty}{1-\gamma}\right) + \frac{2}{1-\gamma}\sum_{k=0}^{K-1}\beta^{K-k-1}\|\varepsilon_k\|_\infty,$$

This concludes the proof.

H. A Proof of Theorem 4 (PQL's Error Propagation with Behavior Policy Updates)

Here we provide error propagation analysis of PQL with behavior policy updates. We prove a bound tighter than the one provided in Theorem 4 and derive the one in Theorem 4 as a corollary.[5] Concretely, we derive the following bound:

$$\|V^* - V^{\pi_K}\|_\infty \leq \zeta^K\|Q^* - Q_0\|_\infty + \frac{\zeta^K}{1-\gamma}\|b_0\|_\infty + \sum_{l=0}^{K-1}\frac{2\zeta^{K-l-1}}{1-\gamma}\|\varepsilon_l\|_\infty\,,$$

where $\zeta := 1-\alpha+\alpha\gamma$. It is not difficult to confirm that $\zeta \leq \eta$ when $\alpha \geq 1-\lambda$. Therefore, this bound implies the bound in Theorem 4. (Note that $0.5 \leq \max\{\lambda, 1-\gamma\lambda\} \leq 1$ and it is negligible in the O notation.)

[5] We found this tighter bound after the submission of the main paper. We intend to include this result in a camera ready version of the main paper if the paper is accepted.

Definition and Notation. We first recall our problem setting: (approximate) PQL updates its Q-function by

$$\pi_k \in \mathbf{G}(Q_k)\,,\ \mu_k = \alpha\pi_k + (1-\alpha)\mu_{k-1} \text{ and } Q_{k+1} := \mathcal{N}_\lambda^{\mu_k,\pi_k} Q_k + \varepsilon_k,$$

where μ_{-1} is arbitrary.

We define the following notations, some of which differ from those defined in Appendix G:

- $\rho_k := \lambda\mu_k + (1-\lambda)\pi_k$
- $b_k := Q_k - \mathcal{T}^{\rho_k} Q_k$
- $d_k := Q^* - Q_k$
- $\mathcal{A}_k := \gamma(1-\lambda)\mathcal{P}^{\pi_k}(\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_k})^{-1}$
- $\beta := \gamma(1-\lambda)/(1-\gamma\lambda)$
- $\mathcal{P}^* := \mathcal{P}^{\pi_*}$
- $\mathcal{T}^* := \mathcal{T}^{\pi_*}$

Here we highlighted (by red color texts) differences from the definitions in Appendix G. Note that $\mathcal{A}_k$ is still a contraction with modulus β.

Proofs. Now we start the proof. The main strategy is the almost same as the one we used in Appendix G: we first decompose $V^* - V^{\pi_K}$ to two components $V^* - \pi_K Q_K$ and $\pi_K Q_K - V^{\pi_K}$, and then, we show an upper bound to each of them.

We first prove an upper bound for b_k, which turns out to be useful later.

Lemma 10. *For any non-negative integer k, the following holds:*

$$b_k \le \mathcal{A}_{k-1}\cdots\mathcal{A}_0 b_0 + \sum_{l=0}^{k-1} \mathcal{A}_{k-1}\cdots\mathcal{A}_{l+1}(\mathcal{I} - \gamma\mathcal{P}^{\rho_l})\varepsilon_l\,,$$

where $\mathcal{A}_{-1}\cdots\mathcal{A}_0 = \mathcal{I}$, and $\sum_{l=0}^{-1} f_l := 0\mathbf{1}$ for any sequence of functions $(f_l)_{l\ge 0}$.

Proof. Because $\pi_k \in \mathbf{G}(Q_k)$,

$$\mathcal{T}^{\rho_k} Q_k = \lambda\mathcal{T}^{\mu_k} Q_k + (1-\lambda)\mathcal{T}^{\pi_k} Q_k \ge \mathcal{T}^{\mu_k} Q_k.$$

Therefore, $b_k \le Q_k - \mathcal{T}^{\mu_k} Q_k$. By the assumption on μ_k,

$$\begin{aligned}
\mathcal{T}^{\mu_k} Q_k &= r + \gamma(1-\alpha)\mathcal{P}^{\mu_{k-1}} Q_k + \gamma\alpha\mathcal{P}^{\pi_k} Q_k \\
&= r + \gamma\lambda\mathcal{P}^{\mu_{k-1}} Q_k + \gamma(1-\lambda)\mathcal{P}^{\pi_k} Q_k + \gamma(\alpha - (1-\lambda))(\mathcal{P}^{\pi_k} Q_k - \mathcal{P}^{\mu_{k-1}} Q_k) \\
&\ge r + \gamma\lambda\mathcal{P}^{\mu_{k-1}} Q_k + \gamma(1-\lambda)\mathcal{P}^{\pi_k} Q_k \\
&\ge r + \gamma\lambda\mathcal{P}^{\mu_{k-1}} Q_k + \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}} Q_k\,,
\end{aligned}$$

where the third line follows since $\alpha \ge 1-\lambda$. Consequently

$$b_k \le Q_k - r - \gamma\lambda\mathcal{P}^{\mu_{k-1}} Q_k - \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}} Q_k = (\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_{k-1}})Q_k - r - \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}} Q_k\,.$$

From Lemma 1, we may deduce that

$$\begin{aligned}
b_k &\le r + \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}} Q_{k-1} + (\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_{k-1}})\varepsilon_{k-1} - r - \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}} Q_k \\
&= \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}}(Q_{k-1} - Q_k) + (\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_{k-1}})\varepsilon_{k-1} \\
&= \mathcal{A}_{k-1}(Q_{k-1} - \mathcal{T}^{\rho_{k-1}} Q_{k-1}) - \gamma(1-\lambda)\mathcal{P}^{\pi_{k-1}}\varepsilon_{k-1} + (\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_{k-1}})\varepsilon_{k-1} \\
&= \mathcal{A}_{k-1}\underbrace{(Q_{k-1} - \mathcal{T}^{\rho_{k-1}} Q_{k-1})}_{=b_{k-1}} + (\mathcal{I} - \gamma\mathcal{P}^{\rho_{k-1}})\varepsilon_{k-1},
\end{aligned}$$

where the last line follows from the definition of $\rho_{K-1} = \lambda\mu_{K-1} + (1-\lambda)\pi_{K-1}$. Therefore, by induction on k, we may deduce that

$$b_k \leq \mathcal{A}_{k-1}\cdots\mathcal{A}_0 b_0 + \sum_{l=0}^{k-1} \mathcal{A}_{k-1}\cdots\mathcal{A}_{l+1}(\mathcal{I} - \gamma\mathcal{P}^{\rho_l})\varepsilon_l\,.$$

This concludes the proof. □

We use a simple corollary of this lemma, derived based on the monotonicity of $(\mathcal{A}_k)_{k\geq 0}$ and $(\mathcal{P}^{\rho_k})_{k\geq 0}$.

Corollary 10.1. *For any non-negative integer k, the following holds:*

$$b_k \leq \beta^k \|b_0\|_\infty \mathbf{1} + (1+\gamma)\sum_{l=0}^{k-1} \beta^{k-l-1}\|\varepsilon_l\|_\infty \mathbf{1} := \bar{b}_k\,.$$

where $\sum_{l=0}^{-1} f_l := 0\mathbf{1}$ for any sequence of functions $(f_l)_{l\geq 0}$.

We next prove an upper bound for $Q^* - Q_K$.

Lemma 11. *For any non-negative integer K, the following holds:*

$$d_K \leq \zeta^K \|Q^* - Q_0\|_\infty + \sum_{l=0}^{K-1} \zeta^{K-1-l}\left(\frac{1-\alpha(1-\gamma\lambda)}{1-\gamma\lambda}\bar{b}_l + \|\varepsilon_l\|_\infty \mathbf{1}\right),$$

where $\zeta := 1-\alpha+\gamma\alpha$, $\sum_{l=0}^{-1} f_l := 0\mathbf{1}$ for any sequence of functions $(f_l)_{l\geq 0}$, and $\bar{b}_l$ is defined in Corollary 10.1.

Proof. We note that

$$\begin{aligned} Q_K &= Q_{K-1} + (\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_{K-1}})^{-1}(\mathcal{T}^{\rho_{K-1}}Q_{K-1} - Q_{K-1}) + \varepsilon_{K-1} \\ &= \mathcal{T}^{\rho_{K-1}}Q_{K-1} - \gamma\lambda\mathcal{P}^{\mu_{K-1}}(\mathcal{I} - \gamma\lambda\mathcal{P}^{\mu_{K-1}})^{-1} b_{K-1} + \varepsilon_{K-1}\,. \end{aligned}$$

Let us focus on deriving a lower bound of $\mathcal{T}^{\rho_{K-1}}Q_{K-1}$. From the definition of ρ_{K-1} and μ_{K-1},

$$\begin{aligned} \mathcal{T}^{\rho_{K-1}}Q_{K-1} &= (1-\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1} + \lambda\mathcal{T}^{\mu_{K-1}}Q_{K-1} \\ &= (1-\lambda+\alpha\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1} + (1-\alpha)\lambda\mathcal{T}^{\mu_{K-2}}Q_{K-1} \\ &= (1-\lambda+\alpha\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1} - (1-\alpha)(1-\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1} + (1-\alpha)Q_{K-1} \\ &\qquad + (1-\alpha)[\lambda\mathcal{T}^{\mu_{K-2}}Q_{K-1} + (1-\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1} - Q_{K-1}] \\ &= \alpha\mathcal{T}^{\pi_{K-1}}Q_{K-1} + (1-\alpha)Q_{K-1} - (1-\alpha)[Q_{K-1} - \lambda\mathcal{T}^{\mu_{K-2}}Q_{K-1} - (1-\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1}]\,. \end{aligned}$$

Recall that the first step of proving Lemma 10 is showing that $b_k \leq Q_k - \lambda\mathcal{T}^{\mu_{k-1}}Q_k - (1-\lambda)\mathcal{T}^{\pi_k}Q_k$. Therefore the upper bound of b_{K-1} in the lemma can serve as an upper bound of $Q_{K-1} - \lambda\mathcal{T}^{\mu_{K-2}}Q_{K-1} - (1-\lambda)\mathcal{T}^{\pi_{K-1}}Q_{K-1}$ too. Accordingly,

$$\begin{aligned} Q^* - Q_K &\leq Q^* - \alpha\mathcal{T}^{\pi_{K-1}}Q_{K-1} - (1-\alpha)Q_{K-1} + \frac{1-\alpha(1-\gamma\lambda)}{1-\gamma\lambda}\bar{b}_{K-1} + \|\varepsilon_{K-1}\|_\infty\mathbf{1} \\ &\leq [(1-\alpha)\mathcal{I} + \alpha\gamma\mathcal{P}^*](Q^* - Q_{K-1}) + \frac{1-\alpha(1-\gamma\lambda)}{1-\gamma\lambda}\bar{b}_{K-1} + \|\varepsilon_{K-1}\|_\infty\mathbf{1}\,. \end{aligned}$$

By induction on K, we deduce that

$$\begin{aligned} Q^* - Q_K &\leq [(1-\alpha)\mathcal{I} + \alpha\gamma\mathcal{P}^*]^K(Q^* - Q_0) + \sum_{l=0}^{K-1}\zeta^{K-1-l}\left(\frac{1-\alpha(1-\gamma\lambda)}{1-\gamma\lambda}\bar{b}_l + \|\varepsilon_l\|_\infty\mathbf{1}\right) \\ &\leq \zeta^K\|Q^* - Q_0\|_\infty\mathbf{1} + \sum_{l=0}^{K-1}\zeta^{K-1-l}\left(\frac{1-\alpha(1-\gamma\lambda)}{1-\gamma\lambda}\bar{b}_l + \|\varepsilon_l\|_\infty\mathbf{1}\right). \end{aligned}$$

This concludes the proof. □

Now we are ready to prove an upper bound for $V^* - V^{\pi_K}$. Note that from Corollary 10.1 and Lemma 11

$$\begin{aligned}
V^* - V^{\pi_K} &= \pi_* Q^* - \pi_K Q_K + \pi_K Q_K - V^{\pi_K} \\
&\leq \pi_*(Q^* - Q_K) + \pi_K Q_K - \pi_K Q^{\pi_K} \\
&= \pi_* d_K + \pi_K (\mathcal{I} - \gamma \mathcal{P}^{\pi_K})^{-1} b_K \\
&\leq \zeta^K \|Q^* - Q_0\|_\infty \mathbf{1} + \sum_{l=0}^{K-1} \zeta^{K-1-l} \left(\frac{1-\alpha(1-\gamma\lambda)}{1-\gamma\lambda} \bar{b}_l + \|\varepsilon_l\|_\infty \mathbf{1} \right) + \frac{1}{1-\gamma} \bar{b}_K \,.
\end{aligned}$$

We simplify $\sum_{l=0}^{K-1} \zeta^{K-1-l} \bar{b}_l$ as follows:

$$\begin{aligned}
\sum_{l=0}^{K-1} \zeta^{K-1-l} \bar{b}_l &= \sum_{l=0}^{K-1} \zeta^{K-1-l} \beta^l \|b_0\|_\infty \mathbf{1} + \sum_{l=0}^{K-1} \zeta^{K-l-1} (1+\gamma) \sum_{m=0}^{l-1} \beta^{l-m-1} \|\varepsilon_m\|_\infty \mathbf{1} \\
&= \frac{\zeta^K - \beta^K}{\zeta - \beta} \|b_0\|_\infty \mathbf{1} + (1+\gamma) \sum_{m=0}^{K-2} \sum_{l=m+1}^{K-1} \zeta^{K-l-1} \beta^{l-m-1} \|\varepsilon_m\|_\infty \mathbf{1} \\
&= \frac{\zeta^K - \beta^K}{\zeta - \beta} \|b_0\|_\infty \mathbf{1} + (1+\gamma) \sum_{m=0}^{K-2} \sum_{l=0}^{K-m-2} \zeta^{K-m-l-2} \beta^l \|\varepsilon_m\|_\infty \mathbf{1} \\
&= \frac{\zeta^K - \beta^K}{\zeta - \beta} \|b_0\|_\infty \mathbf{1} + (1+\gamma) \sum_{m=0}^{K-2} \frac{\zeta^{K-m-1} - \beta^{K-m-1}}{\zeta - \beta} \|\varepsilon_m\|_\infty \mathbf{1} \,,
\end{aligned}$$

where the last line follows from

$$\sum_{l=0}^{K-m-2} \zeta^{K-m-l-2} \beta^l = \zeta^{K-m-2} \sum_{l=0}^{K-m-2} \left(\frac{\beta}{\zeta} \right)^l = \zeta^{K-m-2} \frac{1 - \left(\frac{\beta}{\zeta} \right)^{K-m-1}}{1 - \frac{\beta}{\zeta}} = \frac{\zeta^{K-m-1} - \beta^{K-m-1}}{\zeta - \beta} \,.$$

Using this result and

$$\zeta - \beta = 1 - \alpha + \alpha\gamma - \frac{\gamma(1-\lambda)}{1-\gamma\lambda} = 1 - \frac{\gamma(1-\lambda)}{1-\gamma\lambda} - \alpha(1-\gamma) = \frac{1-\gamma}{1-\gamma\lambda} - \alpha(1-\gamma) = \frac{(1-\gamma)(1-\alpha(1-\gamma\lambda))}{1-\gamma\lambda} \,,$$

we deduce that

$$\begin{aligned}
V^* - V^{\pi_K} &\leq \zeta^K \|Q^* - Q_0\|_\infty \mathbf{1} + \frac{\zeta^K - \beta^K}{1-\gamma} \|b_0\|_\infty \mathbf{1} \\
&\quad + (1+\gamma) \sum_{l=0}^{K-2} \frac{\zeta^{K-l-1} - \beta^{K-l-1}}{1-\gamma} \|\varepsilon_l\|_\infty \mathbf{1} + \sum_{l=0}^{K-1} \zeta^{K-1-l} \|\varepsilon_l\|_\infty \mathbf{1} + \frac{1}{1-\gamma} \bar{b}_K \\
&= \zeta^K \|Q^* - Q_0\|_\infty \mathbf{1} + \frac{\zeta^K}{1-\gamma} \|b_0\|_\infty \mathbf{1} + \sum_{l=0}^{K-1} \frac{2\zeta^{K-l-1}}{1-\gamma} \|\varepsilon_l\|_\infty \mathbf{1} \,,
\end{aligned}$$

where we used $1 + (1+\gamma)/(1-\gamma) = 2/(1-\gamma)$. Because $V^* - V^{\pi_K} \geq 0$ and the right hand side is independent of a state,

$$\|V^* - V^{\pi_K}\|_\infty \leq \zeta^K \|Q^* - Q_0\|_\infty + \frac{\zeta^K}{1-\gamma} \|b_0\|_\infty + \sum_{l=0}^{K-1} \frac{2\zeta^{K-l-1}}{1-\gamma} \|\varepsilon_l\|_\infty .$$

This concludes the proof.

I. Details on Maximum-entropy RL

The maximum-entropy RL (Ziebart et al., 2008; Fox et al., 2016; Asadi & Littman, 2017; Haarnoja et al., 2017; 2018) formulates that the agent maximizes both cumulative rewards and entropy at the same time. In particular, for a fixed $\alpha > 0$,

let $G_{\text{ent}}(x,a)$ be $\sum_{t=0}^{\infty}\gamma^t(R_t+\alpha H_t)$ conditional on $X_0=x, A_0=a$ where H_t is the entropy of policy $\pi(\cdot|X_t)$. Define the maximum-entropy Q-function $Q^{\pi}_{\text{ent}}(x,a) := \mathbb{E}[r(X_0,A_0)+\gamma G_{\text{ent}}(X_1,A_1)|X_0=x,A_0=a]$. It is then possible to define Bellman operators as well as their multi-step variants as in Section 3. Due to space limit, we postpone their details in Appendix I.

It is straightforward to extend off-policy Q(λ) actor-critic algorithm to the formulation of maximum-entropy RL (Fox et al., 2016; Haarnoja et al., 2017). Maximum-entropy actor-critic algorithms also maintain a Q-function $Q_\phi(x,a)$ along with a stochastic policy $\pi_\theta(a|x)$. With off-policy data $(x_t,a_t,r_t)_{t=0}^{\infty}$, one could modify Equation 7 to recursively compute the Q-function targets as

$$\hat{Q}_i = r_i + \gamma\hat{V}_{\text{ent}}(x_{i+1}) + +\gamma\lambda\left(\hat{Q}_{i+1}-\hat{V}_{\text{ent}}(x_{i+1})\right), \tag{6}$$

where the value target $\hat{V}_{\text{ent}}(x_{i+1}) = Q_{\phi^-}(x_{i+1},\pi_{\theta^-}(x_{i+1})) + \alpha_{\text{td}}H(\pi_{\theta^-}(\cdot|x_{i+1}))$. Contrasting Equation 6 and Equation 7, the major difference is that the Q-function target is augmented with an entropy bonus $\alpha_{\text{td}}H(\pi_{\theta^-}(\cdot|x_{i+1}))$. Given a batch of data $(x_0^{(j)},a_0^{(j)})_{j=1}^B$, The policy is updated via gradient ascent $\theta\leftarrow\theta+\nabla_\theta\frac{1}{B}\sum_{j=1}^V Q_\phi(x_0^{(j)},\pi_\theta(x_0^{(j)})+\alpha_{\text{pol}}H(\pi_\theta(\cdot|x_0^{(j)}))$. See Appendix I for the pseudocode of the full algorithm.

In theory, here, one should set $\alpha_{\text{pol}}=\alpha_{\text{td}}=\alpha$ to ensure that the fixed point is unbiased when the collected data are on-policy $\mu=\pi$. However, in practice, we find that large α_{td} tends to destabilize the update. In particular, when setting $\alpha_{\text{pol}}=\alpha_{\text{td}}=0.1$ chosen as the default hyper-parameter, multi-step SAC does not learn stably. We hypothesize that this is because when $\alpha_{\text{td}}>0$, an entropy bonus term is added to the target Q-function at each step (over $n\geq 1$ steps), whose numerical scale makes it much more difficult to learn a proper Q-function.

Instead, we find that a stable alternative is to set $\alpha_{\text{td}}=0$ except at the last time step, where $\alpha_{\text{td}}=\alpha=0.2$. This greatly stablizes the update as the intermediate entropy bonus is effectively removed. It is of interest to study how such bonus term affects the performance of multi-step algorithms and how to align the practice more consistently with theory.

J. Experiments

J.1. Further details on implementations of Peng's Q(λ)

Generic off-policy actor-critic deep RL algorithms. We provide pseudocode for generic off-policy actor-critic deep RL algorithms in Algorithm 1. These algorithms maintain a Q-function critic $Q_\phi(x,a)$ and a policy $\pi_\theta(x)$. In general, The algorithm collects data with an exploratory behavior policy μ and saves tuples (x_t,a_t,r_t) into a replay buffer $\mathcal{D}$. At each training iteration, the critic $Q_\phi(x,a)$ is updated by minimizing squared errors against a Q-function target $\mathbb{E}_D\left[(Q_\phi(x,a)-Q_{\text{target}}(x,a))^2\right]$. The policy is updated via the deterministic policy gradient $\theta\leftarrow\theta+\alpha\mathbb{E}_\mu\left[\nabla_\theta Q_\phi(x,\pi_\theta(x))\right]$ (Silver et al., 2014).

Now, we focus on the definition of targets $Q_{\text{target}}(x,a)$. Given the transitions (x,a,r,x'), one popular choice (see, e.g., (Lillicrap et al., 2016; Fujimoto et al., 2018)) is to compute the target as $Q_{\text{target}}(x,a)=r+\gamma Q_{\phi^-}(x',\pi_{\theta^-}(x'))$ where θ^-,ϕ^- are delayed copies of θ,ϕ respectively (Mnih et al., 2015). An interpretation is that since the policy follows the deterministic gradient through $Q_\phi(x,a)$, it serves as an approximate greedy operator $\pi_\theta(x)\approx\arg\max_a Q_\phi(x,a)$. Note that when $\mathbf{A}$ is continuous, the exact greedy operation $\max_a Q_\phi(x,a)$ is not tractable. In this sense, the above update is an approximate stochastic estimate of the Bellman operator $\mathcal{T}Q(x,a)$.

Algorithm 1 Off-policy Q(λ) actor-critic algorithm

Require: policy $\pi_\theta(x)$, critic $Q_\phi(x,a)$, target parameters θ^-,ϕ^- and learning rate α
while not converged **do**
1. Collect partial trajectories $(x_t,a_t,r_t)_{t=1}^T$ under behavior policy μ.
2. Samples B partial trajectories each of length n from the replay buffer $\mathcal{D}$.
3. Construct Q(λ) targets $Q_{\text{targ}}^{(j)}$. Gradient descent update on critic $\phi\leftarrow\phi-\alpha\frac{1}{B}\nabla_\phi\sum_{j=1}^B(Q_\phi(x_0^{(j)},a_0^{(j)})-Q_{\text{targ}}^{(j)})^2$.
4. Gradient ascent on policy $\theta\leftarrow\theta+\nabla_\theta\frac{1}{B}\sum_{j=1}^B Q_\phi(x_0^{(j)},\pi_\theta(x_0^{(j)}))$.
5. Update the target parameters $\theta^-\leftarrow\theta,\phi^-\leftarrow\phi$.

end while

Recursive computations of Q-function targets. The target value defined by the Q(λ) operator could be computed recursively. In particular, given an infinite trajectory $(x_0, a_0, r_0, x_1, a_1, r_1, ...)$. Assume that we have a Q-function critic $Q_\phi(x, a)$. Let $\hat{Q}_i$ be the target value estimate at time step i, then

$$\hat{Q}_i = r_i + \gamma \max_a Q_\phi(x_i, a) + \gamma\lambda \left(\hat{Q}_{i+1} - \max_a Q_\phi(x_i, a)\right).$$

For continuous action space where computing $\max_a Q(x_i, a)$ is difficult, we propose to replace $\max_a Q_\phi(x, a) \approx Q_\phi(x, \pi_\theta(x))$. In addition, in practice, it is not feasible to generate trajectories of an infinite length. For a partial trajectory $(x_0, a_0, r_0, x_1, a_1, r_1, ...x_n)$ of length n, we bootstrap the Q-function value at the end of the trajectory as $\hat{Q}_n = Q_{\phi^-}(x_i, \pi_{\theta(x_i)^-})$. Then the target at (x_0, a_0) can be recursively computed as

$$\hat{Q}_i = r_i + \gamma Q_{\phi^-}(x_{i+1}, \pi_{\theta^-}(x_{i+1})) + \gamma\lambda \left(\hat{Q}_{i+1} - Q_{\phi^-}(x_{i+1}, \pi_{\theta^-}(x_{i+1}))\right). \tag{7}$$

J.2. Implementations and algorithms for continuous control in deep RL

Implementation code base. We adapt the base implementations in OpenAI SpinningUp (Achiam, 2018). All algorithmic variants adopt default hyper-parameters from the code base. These include learning rates, batch size, replay buffer size, target network update rules, as well as other missing hyper-parameters.

Deep deterministic policy gradient (DDPG). DDPG (Lillicrap et al., 2016) maintains a deterministic policy network $\pi_\theta(a|x) \equiv \pi_\theta(x)$ and a Q-function critic $Q_\phi(x, a)$. The algorithm explores by executing a perturbed policy $a = \epsilon + \pi_\theta(x)$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for $\sigma = 0.1$, and then saves the data (x, a, r, x') into a replay buffer $\mathcal{D}$. At training time, the behavior data is sampled uniformly from the replay buffer $(x_i, a_i, r_i, x'_i)_{i=0}^{B-1} \sim \mathcal{U}(\mathcal{D})$ with $B = 100$. The critic is updated via TD(0), by minimizing: $\frac{1}{B}\sum_{i=0}^{B-1}(Q_\phi(x_i, a_i) - Q_{\text{target}}(x_i, a_i))^2$ where $Q_{\text{target}}(x_i, a_i) = r_i + \gamma Q_{\phi'}(x'_i, \pi_{\theta'}(x'_i))$, where θ', ϕ' are delayed versions of θ, ϕ respectively (Mnih et al., 2015). The policy is updated by maximizing $\frac{1}{B}\sum_{i=0}^{B-1} Q_\phi(x_i, \pi_\theta(x_i))$ with respect to θ. Both parameters θ, ϕ are trained with the Adam optimizer (Kingma & Ba, 2015) with learning rate $\alpha = 10^{-4}$. We adopt other default hyper-parameters in (Achiam, 2018), for details, please refer to the code base.

Twin-delayed deep deterministic policy gradient (TD3). TD3 (Fujimoto et al., 2018) adopts the same training pipeline and architectures as DDPG. TD3 also adopts two critic networks $Q_{\phi_1}(x, a), Q_{\phi_2}(x, a)$ with parameters ϕ_1, ϕ_2, in order to minimize the over-estimation bias (Hasselt, 2010).

Soft actor-critic (SAC). SAC (Haarnoja et al., 2018) adopts the same training pipeline and architecture as DDPG and TD3. However, the critical difference is that SAC augments the reward functions with state-wise entropy to discourage the policy from collapsing to a deterministic distribution. It also maintains two networks to counter the over-estimation bias as TD3. Please see Appendix I for further backgrounds regarding maximum-entropy RL.

J.3. Further details on baseline operators (algorithms)

Uncorrected n-step. We implement uncorrected n-step as one of the baseline algorithms (Hessel et al., 2018). This implements the target Q-functions as $\hat{Q}_i = \sum_{j=i}^{i+n-1} \gamma^{j-i} r_j + \gamma^n \max_a Q_\phi(x_{i+n}, a)$ where Q_ϕ is the Q-function network. It is *uncorrected* because there is no importance sampling ratios that adjust the discrepancy between the π and μ. In continuous control, the maximization operation is replaced by the output of the policy network, i.e. $Q_{\phi^-}(x_{i+n}, \pi_{\theta^-}(x_{i+n}))$. When $n = 1$, we recover the one-step baseline of a vanilla baseline algorithm.

Peng's Q(λ). As briefly discussed in the main paper, we implement a version Peng's Q(λ) with finite horizon n. This means that the recursive computation of target defined in Eqn 7 holds until the n-th step, where $\hat{Q}_{i+n} = Q_{\phi^-}(x_{i+n}, \pi_\theta(x_{i+n}))$. This is because in practice, trajectories are always truncated and of finite lengths, which implies that at the end of trajectories we need to bootstrap directly from the learned Q-functions.

Retrace. We implement Retrace (Munos et al., 2016) as a baseline algorithm for comparison. Retrace computes the Q-function target recursively as

$$\hat{Q}_i = r_i + \gamma Q_{\phi^-}(x_{i+1}, \pi_{\theta^-}(x_{i+1})) + \gamma c_i \left(\hat{Q}_{i+1} - Q_{\phi^-}(x_{i+1}, a_{i+1})\right). \tag{8}$$

Here, the trace coefficient $c_i = \lambda \min(\frac{\pi_\theta(a_i|x_i)}{\mu(a_i|x_i)}, \bar{c})$ where $\bar{c}$ is the truncation level. By default, $\lambda = \bar{c} = 1$. The motivation is that the variance is controlled by truncating the importance sampling ratio. As a result of the update, TD3 is not directly compatible with the update because it requires π, μ to be both stochastic. We implement a version of TD3 with a stochastic actor: $\pi_\theta(a|x) = \tanh\left(\mu_\theta(x) + \sigma_\theta(x) \cdot \epsilon\right)$, where $\epsilon \sim \mathcal{N}(0, \mathbb{I})$ and $\tanh(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x)) \in (-1, 1)$. The log probability $\ln \pi(a|x)$ is still tractable and can be analytically computed (see, e.g., similar computations in (Haarnoja et al., 2018)). The behavior policy μ is implemented as $\mu(a|x) = \tanh\left(\mu_\theta(x) + \sigma \cdot \epsilon\right)$ with a fixed standard deviation parameter $\sigma = 0.1$. These hyper-parameters are chosen such that they match the scale of action perturbation in the original TD3 implementation.

Ctrace. Ctrace (Rowland et al., 2020) is an adaptive off-policy learning algorithm based on Retrace. Its main idea is to adjust the target policy at evaluation time. Instead of evaluating Q^π, the target Q-function is changed to $Q^{\alpha\pi + (1-\alpha)\mu}$ where $\alpha \in [0, 1]$ is a trainable coefficient that interpolates target policy and behavior policy. By changing α, Ctrace achieves a trade-off between fixed point bias (against Q^π) and contraction rate. We always adapt α such that the contraction rate of the overall operator matches a particular value Γ. Since we implement a version of Ctrace with finite horizon n, we use the following modified definition of the contraction rate so that the contraction rate ranges from 0 to 1 regardless of n: $1 - \frac{1-\gamma}{1-\gamma^n}\mathbb{E}[\sum_{t=0}^{n-1} \gamma^t \prod_{s=1}^{t}((1-\alpha) + \alpha\rho_s)]$, where $\rho_s := \pi_\theta(a_s|x_s)/\mu(a_s|x_s)$. Throughout experiments, we set $\Gamma = 0.7$. See (Rowland et al., 2020) for more comprehensive description of the algorithm.

Tree-backup. Similar to Retrace, algorithms such as tree-backup (Precup et al., 2000) also preserve the unbiased fixed point of the operator as Q^π. Tree-backup adopts the same recursive computation as Retrace in Eqn 8 except that the trace coefficient is $c_i = \pi_\theta(a_i|x_i)$. However, the tree-backup algorithm was developed for discrete action space alone, where the probability $\pi_\theta(a_i|x_i) \in [0, 1]$. For continuous control tasks, this is not true because $\pi_\theta(a|x)$ is a density. We observe that naive implementations of tree-backup algorithm leads to very unstable update because of the numerical scale of $\ln \pi_\theta(a|x)$. Empirically, we find that the performance of tree-backup to be very poor on continuous control tasks and we do not include the results.

J.4. Further details on the toy example

At each iteration t of the algorithm, we maintain a Q-function table $Q^{(t)}(x, a)$. Given a sampled trajectory $(x_t, a_t, r_t)_{t=0}^{D-1}$, the operator (e.g. Retrace or Peng's Q(λ)) constructs targets $Q_{\text{target}}(x, a)$. The Q-functions are updated as $Q^{(t+1)}(x, a) \leftarrow (1-\alpha)Q^{(t)}(x, a) + \alpha Q_{\text{target}}(x, a)$. Then the policy is updated as $\pi^{(t)} \leftarrow (1-\alpha)\pi + \alpha\pi_g(Q^{(t)}(x, a))$ where $\pi_g(Q^{(t)}(x, a))$ is the greedy policy with respect to $Q^{(t)}(x, a)$. Throughout experiments, the learning rate is fixed $\alpha = 0.1$.

When computing the target Q-functions $Q_{\text{target}}(x, a)$, we apply the recursive computations introduced in previous sections. This is applied to all state-action pairs along sampled trajectories. At each iteration, the algorithm collects $N = 1$ trajectory from the MDP.

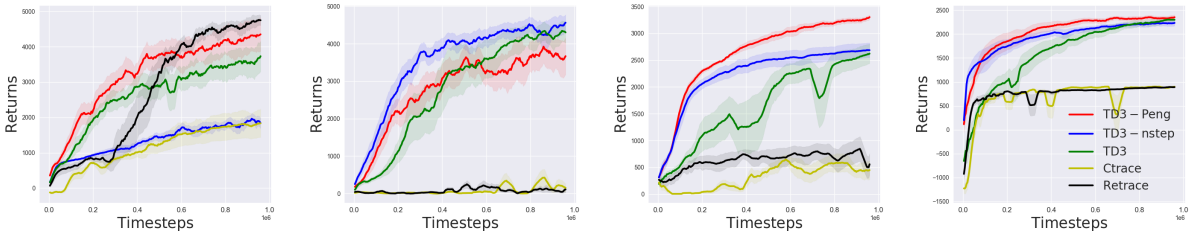
J.5. Additional evaluations on standard benchmarks

Detailed hyper-parameters. In the main paper, we use $n = 5$ for all multi-step algorithms to cap the length of the partial trajectories. For Peng's Q(λ), we set $\lambda = 0.9$ throughout the experiments.

Further results. See Figure 4 for additional experiments on evaluations over standard benchmarks. We further evaluate TD3 variants over tasks from Bullet physics (B) and OpenAI gym (G). Throughout the experiments, we use $n = 5$ for all multi-step algorithms to cap the length of the partial trajectories. For Peng's Q(λ), we set $\lambda = 0.7$. Overall, Peng's Q(λ) performs fairly stably, though it does not perform the best per task. Interestingly, Retrace performs fairly well on Ant(G), which is in sharp contrast to its relatively poor performance across other tasks. We no longer include DDPG as a baseline as it is generally considered a slightly less competitive baseline compared to TD3.

J.6. Additional evaluations on sparse rewards benchmarks

Sparse rewards. We implement delayed rewards as a form of sparse rewards. Delayed reward environment tests algorithms' capability to tackle delayed feedback in the form of sparse rewards (Oh et al., 2018). In particular, a standard benchmark environment returns dense reward r_t at each step t. Consider accumulating the reward over d consecutive steps and return the sum at the end k steps, i.e. $r'_t = 0$ if $t \bmod k \neq 0$ and $r'_t = \sum_{\tau=t-d+1}^{t} r_\tau$ if $t \bmod d = 0$. Throughout the



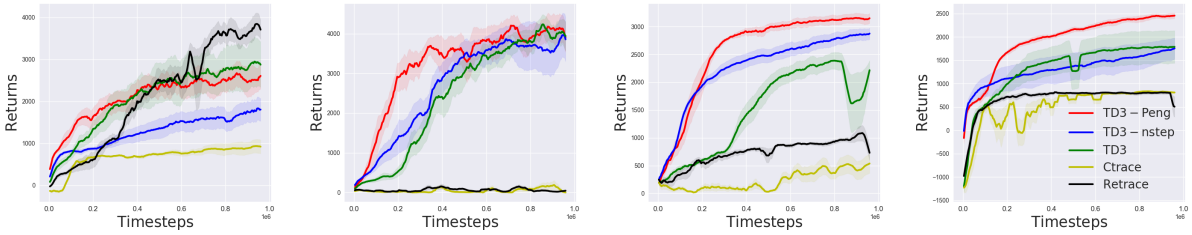
(a) Ant(G) (b) Walker(G) (c) Ant(B) (d) HalfCheetah(B)

Figure 4. Evaluation of TD3 baselines over continuous control domains. Each curve corresponds to a baseline algorithm averaged over 5 random seeds. (B) denotes tasks from Bullet physics and (G) denotes tasks from OpenAI gym.

experiments, we set $d = 3$.

Detailed hyper-parameters. We use $n = 5$ for all multi-step algorithms to cap the length of the partial trajectories. For Peng's Q(λ), we set $\lambda = 0.7$ throughout the experiments.

Further results. See Figure 5 for additional experiments on evaluations over standard benchmarks. We further evaluate TD3 variants over tasks from Bullet physics (B) and OpenAI gym (G). Throughout the experiments, we use $n = 5$ for all multi-step algorithms to cap the length of the partial trajectories. For Peng's Q(λ), we set $\lambda = 0.7$. Overall, Peng's Q(λ) performs fairly stably, though it does not perform the best per task. Interestingly, consistent with results in Figure 4, Retrace performs well in Ant(G) with sparse rewards.



(a) Ant(G) (b) Walker(G) (c) Ant(B) (d) HalfCheetah(B)

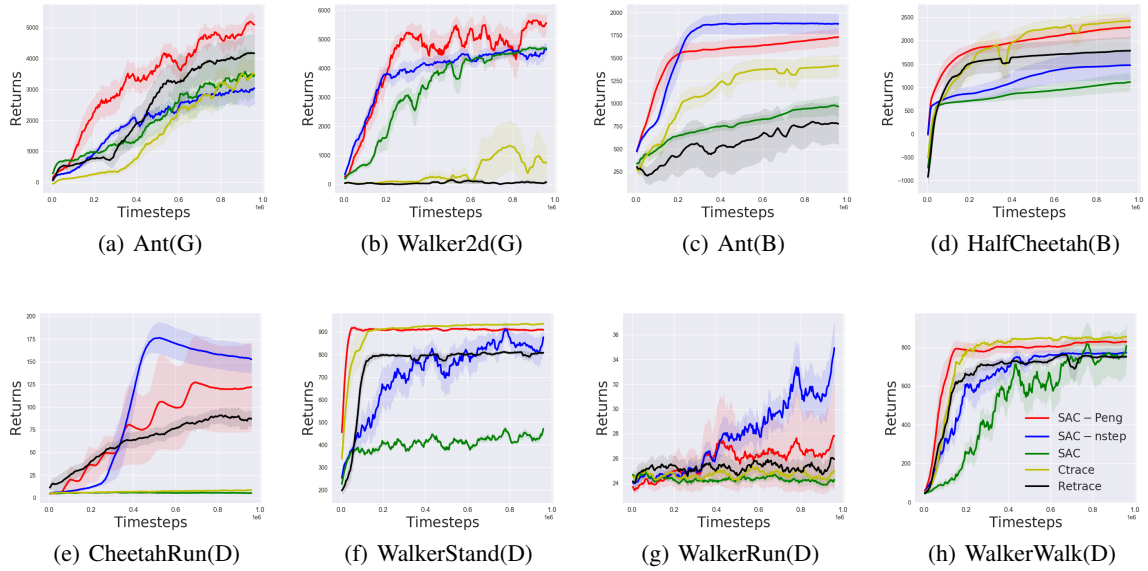
Figure 5. Evaluation of TD3 baselines over continuous control domains with sparse rewards. Each curve corresponds to a baseline algorithm averaged over 5 random seeds. (B) denotes tasks from Bullet physics and (G) denotes tasks from OpenAI gym.

J.7. Experiment results on maximum-entropy RL

We build on soft actor-critic (SAC) (Haarnoja et al., 2018) and evaluate algorithmic variants over standard benchmark tasks. For Peng's Q(λ), we use $\lambda = 0.7$. In Figure 6 we show the results across all selected benchmark tasks. Peng's Q(λ) generally performs more stably than other baselien variants. This is highlighted by the fact that Peng's Q(λ) always ranks as the top two baselines per each task. As an additional empirical observation, we find that SAC generally performs not as well as TD3 on DM control suites. We speculate that this might be because throughout the experiments we use $\alpha = 0.2$. An adaptive entropy coefficient might further improve the performance.

J.8. Ablation on λ

In Figure 7, we show the ablation study on the sensitivity of Peng's Q(λ) to its only hyper-parameter λ. We choose $\lambda \in \{0.3, 0.5, 0.7, 0.9\}$ and examine the performance of the resulting algorithms over DM control suite (sparse rewards). Overall, we see that the best hyper-parameter is achieved $\lambda \approx 0.7$. When λ deviates from this value, its performance is still relatively robust. When λ decreases, we see its performance degrades more drastically than when it increases. Finally, it is

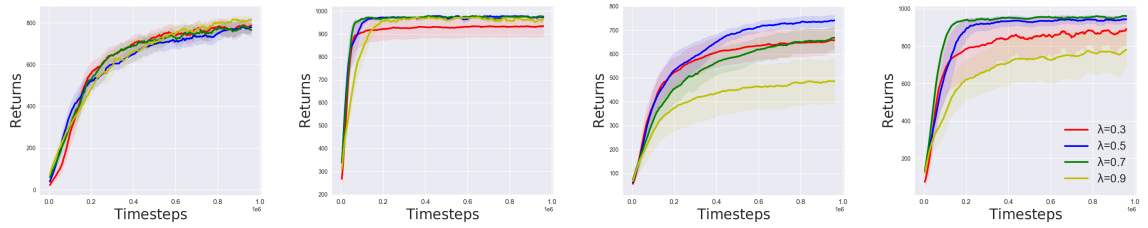


(a) Ant(G) (b) Walker2d(G) (c) Ant(B) (d) HalfCheetah(B)

(e) CheetahRun(D) (f) WalkerStand(D) (g) WalkerRun(D) (h) WalkerWalk(D)

Figure 6. Evaluation of soft actor-critic (SAC) variants over standard continuous control domains. Each curve corresponds to a baseline algorithm averaged over 5 random seeds. We consider tasks from gym (G), bullet physics (B) and DM control suite (D).

worth noting that across all our previous evaluations, we always select $\lambda \in \{0.7, 0.9\}$ and adopt a single λ for benchmark tasks with the same simulation backend. This shows the robustness of Peng's Q(λ) in practical applications.



(a) CheetahRun(D) (b) WalkerStand(D) (c) WalkerRun(D) (d) WalkerStand(D)

Figure 7. Ablation study on the sensitivity of Peng's Q(λ) to the hyper-parameter λ. Each curve corresponds to a choice of λ averaged over 5 random seeds.