Revisiting Exploration-Conscious Reinforcement Learning

Lior Shani*  
Technion, Israel

Yonathan Efroni *  
Technion, Israel

Shie Mannor  
Technion, Israel

Abstract

The objective of Reinforcement Learning is to learn an optimal policy by performing actions and observing their long term consequences. Unfortunately, acquiring such a policy can be a hard task. More severely, since one cannot tell if a policy is optimal, there is a constant need for exploration. This is known as the Exploration-Exploitation trade-off. In practice, this trade-off is resolved by using some inherent exploration mechanism, such as the $\epsilon$-greedy exploration, while still trying to learn the optimal policy. In this work, we take a different approach. We define a surrogate optimality objective: an optimal policy with respect to the exploration scheme. As we show throughout the paper, although solving this criterion does not necessarily lead to an optimal policy, the problem becomes easier to solve. We continue by analyzing this notion of optimality, devise algorithms derived from this approach, which reveal connections to existing work, and test them empirically on tabular and deep Reinforcement Learning domains.

1 Introduction

Emerging from the field of optimal control, the main goal of Reinforcement Learning (RL) (Sutton, Barto, and others 1998) is to find an optimal policy for a given decision problem. If the dynamics of the environment is unknown and number of states is large, finding such a policy can be a hard task. A major difficulty arises due to the Exploration-Exploitation dilemma, which characterizes the omnipresent tension between exploring new actions and exploiting the so-far acquired knowledge.

Considerable line of work has been devoted, in past decades, for the difficulty in dealing with this trade-off. Specifically, algorithms that explicitly handle the Exploration-Exploitation trade-off were developed for tabular RL problems (Kearns and Singh 2002; Brafman and Tennenholtz 2002; Jaksch, Ortner, and Auer 2010; Osband, Russo, and Van Roy 2015; Fortunato et al. 2017). However, generalizing the results of these works in approximate RL, i.e., when using function approximation, yet remains an open problem. On the practical side, recent works have combined more advanced exploration schemes in approximate RL (e.g., (Bellemare, Dabney, and Munos 2017; Fortunato et al. 2017)), inspired by the theory of tabular RL. Yet, the more traditional and simpler $\epsilon$-greedy and soft-max exploration schemes (Sutton, Barto, and others 1998; Asadi and Littman 2016) are still very useful and popular in practice (Mnih et al. 2015; Mnih et al. 2016), especially due to their simplicity.

The latter exploration schemes share some common properties. First, they all fix some exploration parameter beforehand, e.g., $\epsilon$ or the ‘inverse temperature’ $\beta$, for $\epsilon$-greedy and soft-max exploration schemes, respectively. By doing so, the balance between exploring and exploiting is set. Second, all these exploration schemes perform the exploration using some random policy, and exploits using the current estimate of the optimal policy. In this work, we follow a different approach, when using these fixed exploration schemes. The approach which exploits by using an estimate of the optimal policy w.r.t. the exploration mechanism.

In previous literature this approach is referred to as being exploration-conscious (John 1994; Littman and others 1997). In fact, it is the reason for the improved performance of Sarsa and Expected-Sarsa over Q-learning in training, when an $\epsilon$-greedy exploration is used (e.g. (Sutton, Barto, and others 1998) [Example 6.6: Cliff Walking], (Van Seijen 2002)). In these line of works, exploration-conscious information was integrated in Stochastic Approximation, Q-based methods, and the training performance was shown to be superior.

Here, we take a different approach. Instead of the latter, we start by defining an exploration-conscious optimization criterion in policy space; calculate the optimal policy s.t. with some probability an explorative policy is played. If the explorative policy is played w.p. $\epsilon$ and is uniform on the action space, this criterion yields the optimal policy w.r.t. $\epsilon$-greedy exploration. Due to the generality of the criterion, we need not to focus on Q-based methods, and, in fact, policy-search methods and other techniques can be used to solve the optimization criterion as well.

We continue and analyze the properties of the suggested criterion. Specifically, we show that solving it amounts to solving a surrogate MDP. This simple observation allows us to establish new results in this setup, and, most importantly, a policy improvement result. We conclude by suggesting algorithms to solve the devised criterion, and empirically studying their properties in tabular and deep RL, while emphasizing the relation to the theoretical results.

*Equal contribution.
2 Preliminaries

Our framework is the infinite-horizon discounted Markov Decision Process (MDP). An MDP is defined as the 5-tuple \((S, A, P, R, \gamma)\) where \(S\) is a finite state space, \(A\) is a finite action space, \(P \equiv P(s'|s, a)\) is a transition kernel, \(R \equiv r(s, a) \in [0, R_{\text{max}}]\) is a bounded reward function, and \(\gamma \in (0, 1)\) is a discount factor. Let \(\pi : S \rightarrow \mathcal{P}(A)\) be a stationary policy, where \(\mathcal{P}(A)\) is a probability distribution on \(A\). Let \(v^\pi \in \mathbb{R}^{|S|}\) be the value of a policy \(\pi\), defined in state \(s\) as \(v^\pi(s) \equiv \mathbb{E}_\pi^s[t=0] \sum_{t=0}^\infty \gamma^t r(s_t, \pi(s_t))\), where \(E_{\pi}^s\) denotes expectation w.r.t. the distribution induced by \(\pi\) and conditioned on the event \(\{s_0 = s\}\). For brevity, we occasionally denote the reward and value at time \(t\) by \(r_t \equiv r(s_t, \pi(s_t))\) and \(v_t \equiv v(s_t)\). It is known that \(v^\pi = \sum_{t=0}^\infty \gamma^t P^\pi T^\pi v\), with the component-wise values \([P^\pi]_{s, s'} \triangleq \sum_{a} \pi(a | s) P(s' | s, a)\) and \([r^\pi]_{s} \triangleq r(s, \pi(s))\). Furthermore, the \(Q\)-function of \(\pi\) is given by \(q^\pi(s, a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) v^\pi(s')\), and represents the value of taking the action \(a\) from state \(s\) and then using the policy \(\pi\).

The main RL goal is to find a policy \(\pi^*\) yielding the optimal value such that

\[
\pi^* \in \arg \max_{\pi} \mathbb{E}^\pi \left[ \sum_{t=0}^\infty \gamma^t r(s_t, \pi(s_t)) \right],
\]

and the optimal value is \(v^* = v^{\pi^*}\).

To achieve this goal the following classical operators are defined (with equalities holding component-wise):

\[
\forall v, \pi, T^\pi v = r + \gamma P^\pi v,
\]

\[
\forall v, T v = \max_{\pi} T^\pi v,
\]

\[
\forall v, G(v) = \{ \pi : T^\pi v = T v \},
\]

where \(T^\pi\) is a linear operator, \(T\) is the optimal Bellman operator and both \(T^\pi\) and \(T\) are \(\gamma\)-contraction mappings w.r.t. the max norm. It is known that the unique fixed points of \(T^\pi\) and \(T\) are \(v^\pi\) and \(v^*\), respectively. The set \(G(v)\) is the standard set of 1-step greedy policies w.r.t. \(v\). Furthermore, given \(v\), the set \(G(v)\) coincides with that of stationary optimal policies. In other words, every policy that is 1-step greedy w.r.t. \(v^*\) is optimal and vice versa. It is also useful to define the optimal Bellman operator

\[
T q(s, a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) \max_{a'} q(s', a'),
\]

which is a \(\gamma\)-contraction, with fixed point \(q^\pi\).

In this work, the use of mixture policies is abundant. We denote the \(\alpha \in [0, 1]\)-convex mixture of policies \(\pi_1\) and \(\pi_2\) by

\[
\pi^\alpha(\pi_1, \pi_2) \triangleq (1-\alpha)\pi_1 + \alpha \pi_2.
\]

When the policies \(\pi_1, \pi_2\) can be deduced from the text, we occasionally use the simpler notation \(\pi^\alpha\). Importantly, \(\pi^\alpha(\pi_1, \pi_2)\) can be interpreted as a stochastic policy s.t. with probability (w.p) \((1-\alpha)\) the agent acts with \(\pi_1\) and w.p \(\alpha\) acts with \(\pi_2\).

3 The \(\alpha\)-optimal criterion

In this section, we define the notion of \(\alpha\)-optimal policy w.r.t. a policy, \(\pi_0\). We then claim that finding an \(\alpha\)-optimal policy can be done by solving a surrogate MDP. We continue by defining the surrogate MDP, characterizing its properties and use them to prove basic properties of the \(\alpha\)-optimal policy.

Let \(\alpha \in [0, 1]\). We define \(\pi^\alpha_{\alpha_0}\) to be the \(\alpha\)-optimal policy w.r.t. \(\pi_0\), and is contained in the following set,

\[
\pi^\alpha_{\alpha_0} \in \arg \max_{\pi} \alpha \mathbb{E}^{\pi^\alpha}(\pi^\alpha_0) \left[ \sum_{t=0}^\infty \gamma^t r(s_t, \pi^\alpha(\pi^\alpha_0, \pi_0)(s_t)) \right],
\]

where \(\pi^\alpha(\pi^\alpha_0, \pi_0)\) is the \(\alpha\)-convex mixture of \(\pi^\alpha_0\) and \(\pi_0\) \((6)\).

For brevity, we omit the subscript \(\pi_0\), and denote the \(\alpha\)-optimal policy by \(\pi^\alpha_\alpha\) throughout the rest of the paper. The \(\alpha\)-optimal value (w.r.t. \(\pi^\alpha_\alpha\)) is \(v^\alpha(\pi^\alpha_\alpha, \pi_0)\), the value of the policy \(\pi^\alpha(\pi^\alpha_\alpha, \pi_0)\).

Optimization problem \((7)\) can be viewed as optimizing over a restricted set of policies: all policies that are a convex combination of \(\pi_0\) with a fixed \(\alpha\). Naturally, we can consider in \((7)\) a state-dependent \(\alpha(s)\) as well, and some of the results in this work will consider this scenario.

In other words, \(\pi^\alpha_\alpha\) is the best policy one can act with, if it plays w.p. \((1-\alpha)\) according to \(\pi^\alpha_\alpha\), and w.p. \(\alpha\) according to \(\pi_0\). The relation to the \(\epsilon\)-greedy exploration setup becomes clear by letting \(\pi_0\) be the uniform distribution on the actions for all states, and set \(\alpha = \epsilon\) instead of \(\alpha\). Then, \(\pi^\alpha_\alpha\) is optimal w.r.t. the \(\epsilon\)-greedy exploration scheme; the policy would have the largest accumulated reward, relatively to all other policies, when acting in an \(\epsilon\)-greedy fashion w.r.t. it.

The defined objective lets the algorithm designer choose \(\pi_0\) as she wishes, and not necessarily take it as the uniform distribution. By doing so, as we further discuss in this work, \(\pi_0\) can incorporate prior-knowledge.

We choose to name the policy as the \(\alpha\)- and not \(\epsilon\)-optimal to prevent confusion with other frameworks. The \(\epsilon\)-optimal policy is a notation used in the context of PAC-MDP type of analysis \((\text{Strehl, Li, and Littman} 2009)\), and has a different meaning than the objective in this work \((7)\).

3.1 The \(\alpha\)-optimal Bellman operator, \(\alpha\)-optimal policy and policy improvement

In the previous section, we defined the \(\alpha\)-optimal policy and the \(\alpha\)-optimal value, \(\pi^\alpha_\alpha\) and \(v^\alpha(\pi^\alpha_\alpha, \pi_0)\), respectively. We start this section by observing that problem \((7)\) can be viewed as solving a surrogate MDP. We define the Bellman operators of the surrogate MDP, and use them to prove an important improvement property of the \(\alpha\)-optimal policy, \(\pi^\alpha_\alpha\).

In objective \((7)\), a surrogate MDP is implicitly defined. We denote the surrogate MDP by \(\mathcal{M}_\alpha\). Its dynamics and reward are given by

\[
P_\alpha = (1-\alpha) P + \alpha P^{\pi_0}, \quad r_\alpha = (1-\alpha) r + \alpha r^{\pi_0},
\]

and rest of its ingredients are similar to \(\mathcal{M}\), i.e., \(\mathcal{M}_\alpha = (S, A, P, R, \gamma)\). We denote the value of a policy
π on \( M_\alpha \) by \( v^*_\pi \), and the optimal value on \( M_\alpha \) by \( v^*\). The following lemma relates the value of a policy \( \pi \), measured on \( M \) and \( M_\alpha \) (see proof in Appendix B).

**Lemma 1.** For any policy \( \pi, v^\pi_\alpha = v^{\pi^*_\alpha}(\pi, \pi_0, \alpha) \).

The fixed-policy and optimal Bellman operators of \( M_\alpha \) are denoted by \( T^\pi_\alpha \) and \( T^*_\alpha \), respectively. Again, for brevity we omit \( \pi_0 \) from the definitions. Notice that \( T^\pi_\alpha \) and \( T^*_\alpha \) are \( \gamma \)-contractions as being Bellman operators of a \( \gamma \)-discounted MDP. The following Lemma relates \( T^\pi_\alpha \) and \( T^*_\alpha \) to the Bellman operators of the original MDP, \( M \). Furthermore, it stresses a non-trivial relation between the \( \alpha \)-optimal policy \( \pi^*_\alpha \) and the \( \alpha \)-optimal value, \( v^{*\pi}(\pi_0, \alpha) \).

**Lemma 2.** The following claims hold for any policy \( \pi^* \):
1. \( T^\pi_\alpha = (1-\alpha)T^\pi + \alpha T^\pi_0 \). Its fixed point is \( v^*_\alpha = v^{\pi^*_\alpha}(\pi, \pi_0) \).
2. \( T^* = (1-\alpha)T + \alpha T^\pi_0 \). Its fixed point is \( v^*_0 = v^{\pi^*_0}(\pi_0, \pi_0) \).
3. An \( \alpha \)-optimal policy is an optimal policy of \( M_\alpha \), and is greedy w.r.t. \( v^*_\alpha, \pi^*_\alpha \in \mathcal{G}(v^*_\alpha) = \{ \pi : T^\pi v^*_\alpha = T^*_\alpha \} \).

**Proof.** Let \( v \in \mathbb{R}^{|S|} \) and consider the surrogate MDP, \( M_\alpha \). Its fixed policy Bellman operator (see 2) is given by:
\[
T^\pi_\alpha v = R^\pi + \gamma P^\pi_\alpha v
= (1-\alpha)T^\pi v + \alpha T^\pi_0 v + \alpha (v^\pi_0 + \gamma P^\pi v)
= (1-\alpha)T^\pi v + \alpha T^\pi_0 v.
\]
(9)

The second relation is by plugging \( P^\pi_\alpha, r^\pi_\alpha \) from (8), and rearranging. The fixed point \( v^*_\alpha \) is the value of \( \pi \) measured in \( M_\alpha \). Due to Lemma 1, \( v^*_\alpha = v^{\pi^*_\alpha}(\pi, \pi_0) \).

The optimal Bellman operator of \( M_\alpha \) is (see 3):
\[
T^*_\alpha v = \max_{\pi} T^\pi v
= \max_{\pi} (1-\alpha)T^\pi v + \alpha T^\pi_0 v
= (1-\alpha)T^\pi v + \alpha T^\pi_0 v.
\]
(10)

where the second relation holds by (9). The fixed point of \( T^*_\alpha \) is, by construction, \( v^*_\alpha \), the optimal value on \( M_\alpha \). Moreover, \( v^*_\alpha \) is the optimal value of a policy on \( M_\alpha \). By Lemma 1 the policy that achieves the optimal value on \( M_\alpha \) achieves the \( \alpha \)-optimal value, \( v^{\pi^*_\alpha}(\pi, \pi_0) = v^{\pi^*_\alpha}(\pi_0, \pi_0) \).

Since \( M_\alpha \) is an MDP, its optimal policy is in the greedy set w.r.t. \( v^*_\alpha \). Thus, \( \pi^*_\alpha \) satisfies the following important property.

**Proposition 3.** Let \( \alpha \in [0, 1], \beta \in [0, \alpha] \), let \( \pi_0 \) be a policy, and let \( \pi^*_\alpha \) be the \( \alpha \)-optimal policy w.r.t. \( \pi_0 \). Then,
\[
v^{\pi_0} \leq v^{\pi^*_\alpha}(\pi_0, \pi_0) \leq v^{\pi^*_\alpha}(\pi_0, \pi_0) \leq v^{\pi^*_\alpha}(\pi_0, \pi_0),
\]
with equality if and only if \( \pi_0 = \pi^*_\alpha \).

The central observation from the above proposition is that \( \pi^*_\alpha \), the \( \alpha \)-optimal policy w.r.t. \( \pi_0 \), is strictly better, on the original MDP \( M \), than both \( \pi_0 \) and \( \pi^*_\alpha(\pi_0, \pi_0) \), unless \( \pi_0 \) is an optimal policy (see proof in Appendix B).

Remark 4. Consider the entropy of a policy \( \pi \) in state \( s \),
\[
H(\pi(s); \pi(s)) = \sum_{a \in A} \pi(a \mid s) \log \pi(a \mid s).
\]
Thus, solving (7) is related to the popular entropy regularization term, added in policy gradients methods (e.g. \cite{nachum2017qpg, dai2018softmax}). There, \( \lambda \in \mathbb{R}_+ \) controls the entropy regularization. Even so, fixing \( \lambda \) does not imply fixing the entropy to a certain value. Interestingly, taking the approach devised here (7), allows the algorithm designer easily set the minimal entropy to a certain value, \( c \): by setting \( \alpha \) and \( \pi_0 \) s.t \( \max_{\pi} \alpha(s) H(\pi_0(s); \pi_0(s)) \geq c \).

### 3.2 Bias analysis

In this section we upper bound the difference between the value of the \( \alpha \)-optimal and optimal policy. We refer to this difference as the bias. Ultimately, we would like the bias to be as small as possible; when it is small, considering the defined objective (7) does not cause a great loss relatively to the optimal performance.

Interestingly, the bias can be bounded by using the following ‘Lipschitz’ property of the optimal value function, \( v^* \).

**Definition 1.** Let \( v^* \) be the optimal value of an MDP, \( M \). We define \( L(s) = \| v^*(s) - T^\pi^*(s) \| \geq 0 \), to be the Lipschitz constant w.r.t. \( \pi^*_0 \) of the MDP at state \( s \). We further define the upper bound on the Lipschitz constant \( L \leq \max_{s \in S} L(s) \).

Intuitively, \( L(s) \) quantifies a degree of ‘smoothness’ of the optimal value function. A small value of \( L(s) \) indicates that, from state \( s \), if we act according to \( \pi_0 \) once and then

---

\(^1\)We have \( H(\pi^*; s) \geq (1 - \alpha) H(\pi^*_\alpha; s) + \alpha H(\pi^*_0; s) \), and \( \forall s \in S, H(\pi_0; s) = 0 \), since \( \pi_0 \) can be chosen as a deterministic optimal policy due to Lemma 2.
continue playing the optimal policy, we do not suffer great loss. Large values of \(L(s)\) indicate that using \(p_0\) from state \(s\) leads to an irreparable outcome (e.g., falling off a cliff).

The Lipschitzness w.r.t. \(p_0\) of an MDP is central to the bias of the \(\alpha\)-optimal policy, as the following proposition suggests (see proof in Appendix D).

**Proposition 5.** Let \(\forall s \in S, \alpha(s) \in [0, 1]\), be a state-dependent function. Let \(\pi^*_\alpha\) be the \(\alpha\)-optimal policy, and \(L(s)\) the MDP Lipschitz constant, both relatively to \(p_0\). Define \(B(\alpha) \triangleq \max_s \alpha(s)L(s)\). The following bounds hold,

\[
\|\pi^* - \pi^*_\alpha\| \leq \frac{B(\alpha)}{1 - \gamma}.
\]

If \(\forall s \in S, \alpha(s) = \alpha \in [0, 1]\) then \(B(\alpha) = \alpha L\) (see Definition [7]). Furthermore, this bound is tight.

We can continue to bound \(B(\alpha)\) and get a result that has an explicit dependency on \(R_{\max}\), similarly to related results, e.g., (Petrik and Scherrer 2009) [Theorem 2] (see Appendix D). Theorem 6 suggests (see proof in Appendix D).

**Proposition 6.** Let \(\pi^*_\alpha\) be an \(\alpha\)-optimal policy w.r.t. \(\alpha\). Then,

\[
\|\pi^* - \pi^*_\alpha\| \leq \frac{\alpha L}{1 - \gamma} + \frac{2(1 - \alpha)\gamma \delta}{1 - \gamma}.
\]

The proof is obtained by adding and subtracting \(\pi^*_\alpha\) to the LHS, using the triangle inequality and use Propositions 5 and 6.

Next, we derive performance bounds for the 'model-based' setup. Meaning, we assume an approximate model from which we solve an approximate \(\alpha\)-optimal policy, e.g., by using a planning algorithm. First, we define the notion of approximate model.

**Definition 2.** An MDP \(\hat{M} = (\bar{S}, \bar{A}, \bar{P}, \bar{R}, \gamma)\) is a \((\Delta_r, \Delta_p)\) approximation of an MDP \(M = (S, A, P, R, \gamma)\) if,

\[
\max_{s,a} |\bar{r}(s, a) - \bar{r}(s, a)| = \Delta_r,
\]

\[
\max_{s,a} \sum_{s'} |\bar{P}(s' | s, a) - P(s' | s, a)| = \Delta_p.
\]

This definition is well known (Strehl, Li, and Littman 2009) [Lemma 12], and defines the notion of a 'close' MDP. Since the approximation error is uniform on \((s, a)\)-pairs it does not quantify a possible knowledge on reward and dynamics of a specific policy. We now define a new notion, an approximate model w.r.t. a policy \(\pi\).

**Definition 3.** An MDP \(\hat{M} = (\bar{S}, \bar{A}, \bar{P}, \bar{R}, \gamma)\) is a \((\Delta^\pi_r, \Delta^\pi_p)\) approximation of an MDP \(M = (S, A, P, R, \gamma)\) w.r.t. a policy \(\pi\), if,

\[
\max_{s} |\bar{r}^\pi(s) - \bar{r}(s)| = \Delta^\pi_r,
\]

\[
\max_{s} \sum_{s'} |\bar{P}^\pi(s' | s) - P^\pi(s' | s)| = \Delta^\pi_p.
\]

Observe that if \(\hat{M}\) is a \((\Delta_r, \Delta_p)\) approximation of \(M\), then it also is a \((\Delta^\pi_r, \Delta^\pi_p)\) approximation of \(M\) w.r.t. \(\pi\), for any \(\pi\) (see Appendix E). Meaning, assuming an approximate MDP is a stronger assumption than assuming an approximate MDP w.r.t. a specific policy. More intuitively, when an agent interacts with an environment using a base policy \(\pi_0\) it is reasonable to assume the knowledge she acquired on \(r^\pi_{\hat{M}}\) and \(P^\pi_{\hat{M}}\) is more reliable than knowledge on the reward and dynamics of other policies, which are possibly 'far' from \(\pi_0\).

Using these definitions, we bound the performance of \(\pi^*(\hat{\pi}_\alpha, \pi_0)\) relatively to the optimal value.

**Theorem 7.** Assume the setup as in Proposition 6. Then, the performance relatively to the optimal policy is bounded by,

\[
\left\|\pi^* - \pi^*(\hat{\pi}_\alpha, \pi_0)\right\| \leq \frac{\alpha L}{1 - \gamma} + \frac{2(1 - \alpha)\gamma \delta}{1 - \gamma}.
\]

The proof is obtained by adding and subtracting \(\pi^*_\alpha\) to the LHS, using the triangle inequality and Propositions 5 and 6.
Both Theorem 7 and 8 share a similar structure (as long, $\Delta_{\alpha}^{} < \Delta_{\alpha}^{}$). The first term in both bounds is the bias of the $\alpha$-optimal value relatively to the optimal one. As $\alpha$ increases the bias increases as well. The second term represents sensitivity to errors, and is generally reduced as $\alpha$ increases. This formalizes the following argument; although the mixture policy $\pi^{\alpha}(\pi^{\alpha*}, \pi_0)$ is not optimal, and considering it introduces a bias, it handles errors more favorably than an approximated optimal policy.

Interestingly, under the assumptions of Theorem 8, using $\pi^{\alpha}(\pi^{\alpha*}, \pi_0)$ is particularly better when the ratio $\Delta_{\alpha}^{} / \Delta_{\alpha}^{}$ is small. i.e., the estimated model under policy $\pi_0$ is better than the estimated model.

These bounds stress the importance of prior knowledge on the decision problem, when using the optimization criteria 7. Specifically, choosing $\pi_0$ such that the MDP Lipschitz constant is small (see Definition 1), allows to use bigger $\alpha$, while still keeping the bias small. Simultaneously, by using a bigger $\alpha$ value, the sensitivity to errors is reduced, according to the performance bounds of this section.

5 Algorithms

In this section, we analyse possible approaches to solve the $\alpha$-optimal policy w.r.t. $\pi_0$. Two algorithms are suggested, based on the equivalent views of the problem; (7) can be solved in both i) the original and ii) the surrogate MDP. We find that i) leads to exploration conscious Q-learning as in (John 1994; Littman and others 1997). Whereas ii) leads us to formulate a new variation of Q-learning, which updates all action entries given a single sample.

Both algorithms facilitate the knowledge of the exploration policy $\pi_0$ and the probability to use it, $\alpha$, in their updating rules, and their convergence is proved. Furthermore, although the algorithms do not converge to the same Q value, we show that, in the limit, the result in similar policy, which is the $\alpha$-greedy policy.

We start by defining the following q-function,

$$q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)}(s, a) \triangleq r(s, a) + \gamma \sum_{s'} P(s' | s, a) v^{\pi^{\alpha}}(s').$$

Indeed, $q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)}$ is the q-function of the policy $\pi^{\alpha}(\pi^{\alpha*}, \pi_0)$.

Algorithm 1 Expected $\alpha$-Q-Learning

1: Initialize:
\hspace{1cm} $\alpha \in [0, 1], q_0, \text{learning rate } \eta_t, t = 0$
2: while $q_t$ not converged do
3: $a^{chosen} \leftarrow \arg \max_a q_t(s_t, a)$
4: $X_t \sim \text{Bernoulli}(1 - \alpha)$
5: $a^{env} \leftarrow \{ a \sim \pi_0(\cdot | s), \text{if } X_t = 0 \}$
6: Act with $a^{env}$ and receive $(s_{t+1}, r_{t+1}, s_t)$
7: $y_t \leftarrow r_t + \gamma(1 - \alpha)v_t(s_{t+1}) + \gamma \alpha v^\pi(s_t)$
8: $q_{t+1}(s_t, a^{env}) \leftarrow (1 - \eta_t)q_t(s_t, a^{env}) + \eta_t y_t$
9: $t \leftarrow t + 1$
10: end while
11: return $\pi \in \arg \max_a q_t(\cdot, a)$

Algorithm 2 Boosted $\alpha$-Q-Learning

1: Initialize:
\hspace{1cm} $\alpha \in [0, 1], q_0, q, \text{learning rate } \eta_t, t = 0$
2: while $q$ not converged do
3: $a^{chosen} \leftarrow \arg \max_a q(s_t, a)$
4: $X_t \sim \text{Bernoulli}(1 - \alpha)$
5: $a^{env} \leftarrow \{ a \sim \pi_0(\cdot | s), \text{if } X_t = 0 \}$
6: Act with $a^{env}$ and receive $(s_{t+1}, r_{t+1}, s_t)$
7: $y_{E,t} \leftarrow r_t + \gamma(1 - \alpha)v_{E,t}(s_{t+1}) + \gamma \alpha v^\pi(s_t)$
8: $q_{E,t+1}(s_t, a^{env}) \leftarrow (1 - \eta_t)q_{E,t}(s_t, a^{env}) + \eta_t y_{E,t}$
9: for $\alpha \in \mathcal{A}$ do
10: if $\alpha = a^{chosen}$ then
11: $y^\alpha_t \leftarrow r_t + \gamma v_t(s_{t+1})$
12: else
13: $y^\alpha_t \leftarrow X_t q_{E,t}(s_t, \bar{a}) + (1-X_t) (r_t + \gamma v_t(s_{t+1}))$
14: end if
15: $q_{t+1}(s_t, \bar{a}) \leftarrow (1 - \eta) q_{t}(s_t, \bar{a}) + \eta y^\alpha_t$
16: end for
17: end while
18: return $\pi \in \arg \max_a q_\alpha(\cdot, a)$

on the MDP $\mathcal{M}_\alpha$. We continue by defining $T^{\pi^{\alpha}}_{\alpha}$, the optimal Bellman operator of the surrogate MDP $\mathcal{M}_\alpha^{}$ (5):

$$T^{\pi^{\alpha}}_{\alpha} q(s, a) \triangleq r(s, a) + \gamma \sum_s P(s' | s, a) \max_{a'} q(s', a'),$$

where $r, \alpha, P, \pi^{\alpha}$, the reward and dynamics of $\mathcal{M}_\alpha$. By Lemma 10, the fixed point of $T^{\pi^{\alpha}}_{\alpha}$ is $q^{\pi^{\alpha}}_{\alpha}$, the optimal $\pi^{\alpha}$-function of $\mathcal{M}_\alpha$, and the $\alpha$-optimal policy is simply $\pi^{\alpha}_{\alpha}(s) \in \arg \max_a q^{\pi^{\alpha}}_{\alpha}(s, a)$. Notice that, as usual, we have $v^{\pi^{\alpha}}_{\alpha} = \max_a q^{\pi^{\alpha}}_{\alpha}(\cdot, a)$. Furthermore, the following lemma relates $q^{\pi^{\alpha}}_{\alpha}$ and $q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)}$ (12) (see proof in Appendix H).

Lemma 9, $q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)} = (1 - \alpha)q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)} + \alpha \Pi_0 q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)}(s, a')$.

Due to this lemma, and since $\Pi_0 \Pi_0 q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)}$ is solely a function of the states, the $\alpha$-optimal policy is also given by $\pi^{\alpha}_{\alpha}(s) \in \arg \max_a q^{\pi^{\alpha}(\pi^{\alpha*}, \pi_0)}(s, a)$.

We now describe the Expected $\alpha$-Q-learning algorithm (see Algorithm 1), also given in (John 1994; Littman and others 1997). We use the notation

$$u_t(s_t) = \max_a q_t(s_t, a')$$

Algorithm 1 is a Stochastic Approximation (SA) algorithm, based on the operator

$$T^{\pi^{\alpha}}_{\alpha} q(s, a) \triangleq r(s, a) + \gamma(1 - \alpha) \sum_{a'} P(s' | s, a) \max_{a'} q(s', a') + \gamma \alpha \sum_{s', a'} P(s' | s, a) \pi_0(a' | s') q(s', a').$$
Lemma 10. The operator $T^{eq}_α$ is a $γ$ contraction, and its fixed point is $q^{π∗}((s,π_0))$.

This lemma leads to the convergence proof of the Expected $α$-Q-learning algorithm, which follows standard proofs by using (Bertsekas and Tsitsiklis 1995)(Proposition 4.4) (see proof in Appendix I).

Theorem 11. Consider the process described in Algorithm 2. Assume the sequence $\{q_α^n\}_{n=0}^∞$ satisfies $∀s ∈ S, ∀a ∈ A$, $\sum_{t=0}^∞ q_α^n(s_t = s, a_t = a) = ∞$, and $\sum_{t=0}^∞ q_α^n(s_t = s, a_t^{env} = a) < ∞$. Then, the sequence $\{q_α^n\}_{n=0}^∞$ converges w.p 1 to $q^{π∗}((s,π_0))$.

Indeed, due to Lemma 9, we can obtain the $α$-optimal policy directly from $q^{π∗}((s,π_0))$; the output of Algorithm 1. Differently from $q$-learning, Expected $α$-Q-learning learns the optimal policy w.r.t. the played exploration scheme; w.p $α$ the exploration policy $π_0$ is performed in the learning phase.

Instead of formulating a SA algorithm based on $T^{eq}_α$, another SA algorithm can be formulated, by using the operator $T_q^n$.[13]. Since $T_q^n$ is the optimal Bellman operator of $M_α$, the updating should simply take the form of

\[ y_t ≜ r_t + γv(q(s_{t+1}),1) \]
\[ q(s,a^{chosen}) ≜ (1-γ)q(s,a^{chosen}) + γq(s_{t+1}) \]

See that the updated state-action entry is $(s,q^{chosen})$, differently from the update in line 3 of Algorithm 1. There, the updated entry is $(s,a^{env})$. This approach does not utilize a meaningful knowledge we have at our disposal; when the exploration policy $π_0$ is played, i.e., when $X_t = 0$, $(r_t,s_{t+1})$ can be used to update all the action entries from the current state. In this case, all actions from the state $s_t$ can be updated using this information.

This reasoning leads us to formulate Boosted $α$-Q-learning (see Algorithm 2). The Boosted $α$-Q-learning updates two $q$-functions, $q_E$ and $q$. The first, $q_E$, has the same update as in Expected $α$-Q-learning, and thus converges (w.p 1) to $q^{π∗}((s,π_0))$. The second $q$-function, updates the chosen greedy action using the naive updating equations ([16], [17]), when the exploration policy is not played ($X_t = 1$). By bootstrapping on $q_E$, the algorithm updates all other actions when the exploration policy $π_0$ is played ($X_t = 0$). Using (Singh et al. 2000)[Lemma 1], the convergence of Boosted $α$-Q-learning to $(q^{π∗}((s,π_0))$, $q^*_α$) is established (see proof in Appendix I).

Theorem 12. Consider the process described in Algorithm 2. Assume the sequence $\{q_α^n\}_{n=0}^∞$ satisfies $∀s ∈ S, ∀a ∈ A$, $\sum_{t=0}^∞ q_α^n(s_t = s, a_t = a) = ∞$, and $\sum_{t=0}^∞ q_α^n(s_t = s, a_t^{env} = a) < ∞$. Then, the sequences $\{q_E^n\}_{n=0}^∞$ and $\{q_α^n\}_{n=0}^∞$ converges w.p 1 to $q^{π∗}((s,π_0))$ and $q^*_α$, respectively.

The central novelty of Boosted $α$-Q-learning relatively to other $q$-based algorithms (e.g, Expected $α$-Q-learning, Q-learning, etc.), is its update of the entire action set. In $q$-based algorithms, given a sample, a single state-action pair is updated, unlike the update in Algorithm 2.

6 Relation to existing work

The $γ$-greedy exploration scheme is possibly the most simple strategy to perform exploration in RL. Deep neural networks accomplished impressive progress on implementing RL algorithm for complex tasks, while using the $γ$-greedy exploration. For example, in (Mnih et al. 2015) (Wang et al. 2015) (Mnih et al. 2016), the authors succeeded in playing Atari games using Q-learning. While in these works the $γ$-greedy exploration scheme is used, the objective is to find the optimal policy. As we stress in this work, it is occasionally better to set the objective to be the $α$-optimal policy; which is optimal w.r.t. $γ$-greedy exploration. Furthermore, some other exploration schemes, such as Boltzmann exploration (Sutton, Barto, and others 1998), (Stadie, Levine, and Abbeel 2015), have been proposed in the past, and are claimed to work better. An interesting future work is to study an optimal policy w.r.t. these classes of exploration schemes (Dai et al. 2018).

Expected $α$-Q-learning highlights a relation to algorithms analysed in (John 1994; Littman and others 1997) and to Expected-Sarsa (ES) (Van Seijen et al. 2009). The focus of (John 1994; Littman and others 1997) is exploration-conscious $q$-based methods. Here we put the focus on the exploration-conscious policy, study some of its properties, and continued by deriving new approach for solving it. In ES, when setting the ‘estimation policy’ (Van Seijen et al. 2009) to be $π = (1-α_t)π_E + α_tπ_0$, we get similar updating equations as in lines [17] and similarly to (John 1994; Littman and others 1997). However, in ES $α_t$ decays to zero, and the optimal policy is obtained in the infinite time limit. Here, we offer and analyze the approach of setting $α_t$ to a constant value. This would be of interest especially when a ‘good’ mechanism for decaying $α_t$ lacks, which is usually the case; such decay mechanism is chosen by trial-and-error, and is not clear how it should be chosen.

Lastly, [7] can be understood as defining a surrogate, easier problem to solve, rather than finding the optimal policy [1]. In this sense, it offers an alternative approach to biasing the problem by lowering the discount-factor, i.e., solve a surrogate MDP with $γ < γ_*$ (Petrik and Scherrer 2009; Jiang et al. 2015). Interestingly, the introduced bias when solving [7] is proportional to a local property of $v^∗$, $L(s)$, that can be estimated using prior-knowledge on the MDP. Whereas in (Jiang, Singh, and Tewari 2016) the authors prove that solving a problem with a lower discount-factor introduces a bias proportional to $κ_γ = \max_{s,s′} |v_γ^*(s) - v_γ^*(s′)|$, a non-local term [7], which is hard to estimate without solving the $γ$-discounted MDP.

More importantly, the performance of $π_0$, the optimal policy of the surrogate MDP, $M_α$, is assured to improve when tested on the original MDP, $M$, due to Proposition 3. This comes in contrast to the performance of the optimal policy when a smaller discount factor is used; when tested on the original $γ$-discounted MDP performance degradation might occur.

2To calculate it, all $(s,s′)$ pairs should be considered
7 Experiments

In this section we test the theory and algorithms suggested in the paper. Throughout this section we use an adaptation of the Cliff-Walking maze (Sutton, Barto, and others 1998). The agent starts at the bottom-left side of a maze, and needs to get to the bottom-right side goal state with value +1 (see Appendix A). If the agent falls off the cliff, the episode terminates with reward −1. When the agent visits any of the three middle states next to the cliff, it gets a reward of 0.01 ⋅ (1 − γ). In all experiments we used γ = 0.99.

7.1 Bias analysis

We analyzed the behavior of the bias bound in Proposition 5 on the Cliff-Walking maze and on random MDPs. We drew random MDPs following (Asadi and Littman 2016). We compared the bound and the calculated bias \( \|v^* - v^*\|, \|v^* - v^{\pi_0}\| \) with respect to different \( \alpha \) values.

As can be seen in Figure 1 the distance between the bound and the real bias grows linearly in \( \alpha \), which is an interesting phenomenon. Both results demonstrate that the bias in Proposition 5 is generically smaller than the upper bound. Moreover, the results reaffirm the improvement guarantee in Proposition 5 by showing that in the exact case, the \( \alpha \)-optimal policy \( \pi^*_\alpha \) is better than \( \pi^\alpha(\pi^*_\alpha, \pi_0) \).

7.2 Exploration Consciousness in a tabular setting

In Figure 2 we tested Expected \( \alpha \)-Q-learning, Boosted \( \alpha \)-Q-learning, and compared their performance to Q-learning in the presence of \( \epsilon \)-greedy exploration, with \( \epsilon = \alpha \). We used the previously discussed Cliff-Walking maze (see Appendix A). In the Cliff Walking scenario, the optimal policy is walking on the edge of the cliff. While in a noiseless environment it is a wanted behaviour, in the presence of \( \epsilon \)-greedy exploration this degrades the performance of an agent acting according to the optimal policy. The \( \alpha \)-optimal policy takes into account the risk due to exploration, and therefore learns faster. Figure 2 shows that the online training behaviour of both Expected-\( \alpha \)-Q-learning and Boosted-\( \alpha \)-Q-learning, expectedly, outperforms Q-learning. The lack of exploration consciousness mechanism makes the Q-learning agent use the estimated optimal policy without being aware of the exploration consequences. Thus, it converges slower. This suggests that the exploration consciousness of the agent can be crucial in many cases, where otherwise this exploratory behaviour can do the opposite and hurt the exploration process itself.

Figure 2 stresses the typical behaviour of the \( \alpha \)-optimality criterion. It is easier to approximate the \( \pi^\alpha(\pi^*_\alpha, \pi_0) \) than the optimal policy. Hence, the true value of the approximated policy improves faster using the \( \alpha \)-optimal algorithms. However, as the bound in Proposition 5 suggests, the value of the learned policy is biased w.r.t \( \psi^* \). Combining the two, the designer should balance between this consciousness and the bias it inflicts upon the optimality of the agent.

Finally, as suggested by the policy improvement proposition 5 acting greedily w.r.t the approximated value attains better performance than acting \( \alpha \)-greedy. Such improvement is not guaranteed while the value had not yet converged to \( \psi^*_\alpha \). However, the results suggest that in areas of the state space where the agent performs well over the mixture policy, it is worth considering using the greedy policy. To conclude, when possible, it is worth evaluating both policies.

The next experiment shows that we can incorporate prior knowledge in order to decrease the bias induced by \( \alpha \)-optimality as predicted by Theorem 7. We used a variant we termed T-Cliff-Walking (See Appendix B). This scenario requires more exploration because of the bottleneck state between the two sides of the maze. Thus, we used \( \epsilon = 0.3 \). However, the \( \alpha \)-optimal policy in such case is to stay at the left part of the maze. We used the prior knowledge that the \( L(s) \) close to the barrier is high. Thus, according to Proposition 5 we injected this knowledge through the choice of \( \alpha \), i.e., we chose a state-wise exploration scheme with \( \alpha(s) = 0.1 \) in the passage and the two states around it. We used \( \alpha(s) = 0.3 \) for all other states. The results in Figure 3 suggests that using prior knowledge by setting \( \alpha(s) \), can increase the performance of the proposed algorithms, through reducing some of the bias.

7.3 Exploration Consciousness in the Atari 2600 domain

We tested the \( \alpha \)-optimal criterion in the more complex function approximation setting. As a test case we used four Atari 2600 games (5) from the Arcade Learning Environ-
Figure 3: (Left) $\alpha=0.3$. (Right) $\alpha(s)$ from prior knowledge.

In this paper, we revisited the notion of an agent being conscious to the exploration processes (John 1994; Littman and others 1997) and argued in its favor. We offered an exploration-conscious optimality criterion w.r.t. the $\epsilon$-greedy exploration. This criterion was used implicitly in the past, but did not receive the proper attention in our view.

Analysis of the offered optimality criterion formally stated an expected performance trade-off. Exploration-conscious optimal policy will not be as good as the optimal one. However, it is less sensitive to approximation errors. Such an approach, biasing the objective to get an easier optimization problem, is well known, and considered in RL by means of lowering the discount factor (Petrik and Scherrer 2009; Jiang et al. 2015). The policy improvement result (Proposition 3) highlights an important advantage of using the techniques discussed in this work, relatively to the latter; policy improvement is not guaranteed when biasing using a lower-discount factor.

The empirical results of this work emphasized the possible practical advantage, both in tabular and Deep RL, of being exploration-conscious. Furthermore, we stressed the importance of prior knowledge when the offered criterion is used. If the bias term is too large (Proposition 6), being exploration-conscious is expected to perform poorly. Yet, when used, on proper problems, it is expected to improve the performance.

Acknowledgments

The authors thank Chen Tessler and Nadav Merlis for helpful discussions.

References

[Asadi and Littman 2016] Asadi, K., and Littman, M. L. 2016. An alternative softmax operator for reinforcement learning. arXiv preprint arXiv:1612.05628.

[Bellemare et al. 2013] Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research 47:253–279.

[Bellemare, Dabney, and Munos 2017] Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A distributional perspective on reinforcement learning. arXiv preprint arXiv:1707.06887.

[Bertsekas and Tsitsiklis 1995] Bertsekas, D. P., and Tsitsiklis, J. N. 1995. Neuro-dynamic programming: an overview. In Decision and Control, 1995., Proceedings of the 34th IEEE Conference on, volume 1, 560–564. IEEE.
[Brafman and Tennenholtz 2002] Brafman, R. I., and Tennenholtz, M. 2002. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research* 3(Oct):213–231.

[Dai et al. 2018] Dai, B.; Shaw, A.; Li, L.; Xiao, L.; He, N.; Liu, Z.; Chen, J.; and Song, L. 2018. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, 1133–1142.

[Dhariwal et al. 2017] Dhariwal, P.; Hesse, C.; Klimov, O.; Nichol, A.; Plappert, M.; Radford, A.; Schulman, J.; Sidor, S.; Wu, Y.; and Zhokhov, P. 2017. Openai baselines. https://github.com/openai/baselines

[Efroni et al. 2018] Efroni, Y.; Dalal, G.; Scherrer, B.; and Mannor, S. 2018. Beyond the one-step greedy approach in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, 1386–1395.

[Fortunato et al. 2017] Fortunato, M.; Azar, M. G.; Piot, B.; Menick, J.; Osband, I.; Graves, A.; Mnih, V.; Munos, R.; Hassabis, D.; Pietquin, O.; et al. 2017. Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*.

[Hessel et al. 2017] Hessel, M.; Modayil, J.; Van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M.; and Silver, D. 2017. Rainbow: Combining improvements in deep reinforcement learning. *arXiv preprint arXiv:1710.02298*.

[Jackson, Ortner, and Auer 2010] Jackson, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(Apr):1563–1600.

[Jiang et al. 2015] Jiang, N.; Kulesza, A.; Sing, S.; and Lewis, R. 2015. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, 1181–1189. International Foundation for Autonomous Agents and Multiagent Systems.

[Jiang, Singh, and Tewari 2016] Jiang, N.; Singh, S. P.; and Tewari, A. 2016. On structural properties of mdps that bound loss due to shallow planning. In *IJCAI*, 1640–1647.

[John 1994] John, G. H. 1994. When the best move isn’t optimal: Q-learning with exploration. Citeseer.

[Kearns and Singh 2002] Kearns, M., and Singh, S. 2002. Near-optimal reinforcement learning in polynomial time. *Machine learning* 49(2-3):209–232.

[Littman and others 1997] Littman, M. L., et al. 1997. Generalized markov decision processes: Dynamic-programming and reinforcement-learning algorithms.

[Mnih et al. 2015] 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. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

[Mnih et al. 2016] Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 1928–1937.

[Nachum et al. 2017] Nachum, O.; Norouzi, M.; Xu, K.; and Schuurmans, D. 2017. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2775–2785.

[Osband, Russo, and Van Roy 2013] Osband, I.; Russo, D.; and Van Roy, B. 2013. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, 3003–3011.

[Petrik and Scherrer 2009] Petrik, M., and Scherrer, B. 2009. Biasing approximate dynamic programming with a lower discount factor. In *Advances in neural information processing systems*, 1265–1272.

[Puterman 1994] Puterman, M. L. 1994. Markov decision processes. j. *Wiley and Sons*.

[Singh et al. 2000] Singh, S.; Jaakkola, T.; Littman, M. L.; and Szepesvári, C. 2000. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning* 38(3):287–308.

[Stadie, Levine, and Abbeel 2015] Stadie, B. C.; Levine, S.; and Abbeel, P. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*.

[Strehl, Li, and Littman 2009] Strehl, A. L.; Li, L.; and Littman, M. L. 2009. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research* 10(Nov):2413–2444.

[Sutton, Barto, and others 1998] Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*. MIT press.

[Van Hasselt, Guez, and Silver 2016] Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double q-learning. In *AAAI*, volume 2, 5. Phoenix, AZ.

[Van Seijen et al. 2009] Van Seijen, H.; Van Hasselt, H.; Whiteson, S.; and Wiering, M. 2009. A theoretical and empirical analysis of expected sarsa. In *Adaptive Dynamic Programming and Reinforcement Learning, 2009. ADPRL’09. IEEE Symposium on*, 177–184. IEEE.

[Wang et al. 2015] Wang, Z.; Schaul, T.; Hessel, M.; Van Hasselt, H.; Lanctot, M.; and De Freitas, N. 2015. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*.
Appendix A  Experimental details

In this section we will discuss some technicalities that are related to the experiments done in this paper.

A.1 Random MDP

We draw MDPs from the same distribution noted by [Asadi and Littman 2016]. However, in order to extend the generality of the results, we extended the sizes of the state and action spaces. We used $|S| \in [10, 99]$ and $|A| \in [2, 9]$. Then, in order to calculate the bound, we calculated $L$ using the value-iteration procedure and set the bound to be $\alpha L (1 - \gamma)$.

A.2 Cliff Walking

We used the exact scenarios in Figure 4. The size of the cliff is $(h, w) = (4, 12)$. Because we used a $\gamma = 0.99$, we have added the small reward (green states) in order to create some small bias between the optimal and the $\alpha$-optimal policy. The maximal reward in this example is $r_{\max} = 1 - \gamma$. Hence, the small reward given is $0.01 r_{\max}$. We used an $\alpha = \epsilon = 0.1$ and performed 2,000 runs for each of the algorithms. The test error was evaluated with high precision using the fixed value iteration procedure.

In the second experiment, we first checked to see that that $\epsilon = 0.1$ performed bad. Then, we raised the $\epsilon$ value. The bottleneck passage between to sides of the maze, creates a scenario where high exploration is needed. In this scenario, the small positive reward is moved to the first three states of the cliff, so it will not collide with the behaviour of the obstacle. Note that we could have used the same states in the first experiment, but we ran it beforehand.
Appendix B  Proof of Lemma 1

For any policy π the following equalities hold.

\[ v^\pi_\alpha = (I - \gamma P^\pi_\alpha)^{-1}v^\pi_\alpha = (I - \gamma((1 - \alpha)P^\pi + \alpha P^{\pi_0}))^{-1}(1 - \alpha)v^\pi_\alpha + \alpha v^\pi_0 = (I - \gamma P^{\pi_0}(\pi, \pi_0))^{-1}v^{\pi_0}(\pi, \pi_0). \]

Appendix C  Proof of Theorem 3

For completeness we give two useful lemmas that are in use. The first one has several instances in the literature.

**Lemma 13.** Let \( v^\pi \) and \( v^{\pi'} \) be the corresponding values of the policies \( \pi \) and \( \pi' \). Then,

\[ v^{\pi'} - v^\pi = (I - \gamma P^{\pi'})^{-1}(T^{\pi'}v^{\pi'} - v^\pi) \]  

**Proof.**

\[ v^{\pi'} - v^\pi = (I - \gamma P^{\pi'})^{-1}v^{\pi'} - v^\pi \]

\[ = (I - \gamma P^{\pi'})^{-1}(v^{\pi'} + \gamma P^{\pi'}v^\pi - v^\pi) \]

\[ = (I - \gamma P^{\pi'})^{-1}(T^{\pi'}v^\pi - v^\pi). \]

The following Lemma has several instances in previous literature:

**Lemma 14.** Let \( \pi \) be any policy and \( \pi_{1-step} \in \mathcal{G}(v^\pi) \). Then,

\[ v^\pi \leq v^{\pi^*}(\pi_{1-step}, \pi), \]

where the inequality is strict in at least one-component if \( \pi \neq \pi^* \), if \( \pi \) is not the optimal policy.

**Proof.**

\[ v^{\pi^*}(\pi_{1-step}, \pi) - v^\pi = (I - \gamma P^{\pi^*}(\pi_{1-step}, \pi))^{-1}(T^{\pi^*}(\pi_{1-step}, \pi)v^{\pi^*} - v^\pi), \]

where the first relation holds due to Lemma 13. See that,

\[ T^{\pi^*}(\pi_{1-step}, \pi)v^{\pi^*} - v^\pi \]

\[ = (1 - \alpha)T^{\pi_{1-step}}v^{\pi^*} + \alpha T^\pi v^{\pi^*} - v^\pi \]

\[ = (1 - \alpha)T^{\pi_{1-step}}v^{\pi^*} + \alpha v^{\pi^*} - v^\pi \]

\[ = (1 - \alpha)(T^{\pi_{1-step}}v^\pi - v^{\pi^*}) = (1 - \alpha)(Tv^\pi - v^{\pi^*}) \]

Plugging into (19) yields,

\[ v^{\pi^*}(\pi_{1-step}, \pi) - v^\pi \]

\[ = (1 - \alpha)(I - \gamma P^{\pi^*}(\pi_{1-step}, \pi))^{-1}(Tv^\pi - v^{\pi^*}). \]

We have that \( P^{\pi^*}(\pi_{1-step}, \pi) \) does not have any zero eigenvalues.

We now prove the result. The first relation holds almost by construction. We have that,

\[ v^{\pi^*}(\pi_{1-step}, \pi_0) = \max_{\pi'} v^{\pi^*}(\pi', \pi_0) \geq v^{\pi^*}(\pi_0, \pi_0) = v^{\pi_0} \]

where the first relation is due to the definition of the \( \pi \)-optimal value, the second relation holds by definition and the third relation holds since

\[ \pi_0(\pi_0, \pi_0) = (1 - \epsilon)\pi_0 + \epsilon \pi_0 = \pi_0. \]

As long as \( \pi_0 \neq \pi^* \), the policy \( \pi_{1-step} \in \mathcal{G}(v^{\pi_0}) \) achieves strict improvement in (20). Meaning,

\[ v^{\pi^*}(\pi_{1-step}, \pi_0) > v^{\pi_0}. \]

This means that the improvement in (20) is strict as long as \( \pi_0 \neq \pi^* \). If \( \pi_0 \) is not optimal we have that

\[ v^{\pi_0} \leq v^{\pi^*}(\pi_{1-step}, \pi_0) \leq v^{\pi^*}(\pi_0, \pi_0). \]

The first relation is strict due to Lemma 14 and the second relation holds by the definition of the \( \pi \)-optimal policy.

We now prove the second relation of the lemma. Let \( \beta \in [0, \alpha] \). Then,

\[ v^{\pi^*}(\pi_{1-step}, \pi_0) - v^* \]

\[ = (I - \gamma P^{\pi^*}(\pi_{1-step}, \pi_0))^{-1}(T^{\pi^*}(\pi_{1-step}, \pi_0)v^* - v^*), \]

We have that,

\[ T^{\pi^*}(\pi_{1-step}, \pi_0)v^* - v^* \]

\[ = T^{\pi^*}(\pi_{1-step}, \pi_0)v^* - \alpha T^{\pi_0}v^* + \beta T^{\pi_0}v^* - (1 - \alpha)T^\pi v^* + \alpha T^{\pi_0}v^* \]

\[ = (\alpha - \beta)(T^\pi v^* - T^{\pi_0}v^*). \]

where in the last relation we used \( T^{\pi_0}v^* = T^\pi v^* \) (see Lemma 2). Plugging into (21) yields,

\[ v^{\pi^*}(\pi_{1-step}, \pi_0) - v^* \]

\[ = (\alpha - \beta)(I - \gamma P^{\pi^*}(\pi_{1-step}, \pi_0))^{-1}(T^\pi v^* - T^{\pi_0}v^*). \]

We have that \( (I - \gamma P^{\pi^*}(\pi_{1-step}, \pi_0))^{-1} \geq 0 \) since it is a \( \gamma \)-discounted sum of stochastic matrices, and \( T^\pi v^* \geq T^{\pi_0}v^* \) with equality if and only if \( \pi_0 \) is optimal; if and only if \( \pi_0 \) is optimal \( v^{\pi_0} = v^* \) due to the first part of this proof.

To conclude, by setting \( \beta = 0 \) we get the third inequality.

Appendix D  Proof of Proposition 5

We have that for any \( s \in S \),

\[ v^* - v^*_{\alpha}(s) \]

\[ = (Tv^* - T_{\alpha}v^*)(s) + (T_{\alpha}v^* - T_{\alpha}v^*_{\alpha}(s)) \]

\[ \leq \|Tv^* - T_{\alpha}v^*\| + \|T_{\alpha}v^* - T_{\alpha}v^*_{\alpha}\| \]

\[ \leq \|Tv^* - T_{\alpha}v^*\| + \|v^* - v^*_{\alpha}\|. \]

in the last relation we used the fact that \( T_{\alpha} \) is a \( \gamma \) contraction in the max-norm. Moreover, we have that for any \( s \in S \),

\[ T^\pi v^*(s) - T_{\alpha}v^*(s) \]

\[ = Tv^* - (1 - \alpha)(Tv^*(s) - \alpha T^\pi v^*(s)) \]

\[ = \alpha(s)(Tv^*(s) - T^\pi v^*(s)) \]

\[ = \alpha(s)(Tv^*(s) - T_{\alpha}v^*(s)) = \alpha(s)L(s). \]
In the third relation we used the fact that $Tv^* = v^*$ component-wise, since $v^*$ is the fixed-point of $T$. Thus, we see that,

$$\|Tv^* - T\alpha v^*\| = \max_s \alpha(s) L(s) = B(\alpha),$$

and that $L(s) \geq 0$ since $v^*(s) - T^\pi v^*(s) \geq 0$. By taking the max-norm on (22), which is possible since it is positive, and simple algebraic manipulation we conclude the result.

We can continue and bound the above to get the bound in (23), which is less right. We have that,

$$|Tv^* - T^\pi v^*| \leq \sum_a |\pi^*(a | s) - \pi_0(a | s)|$$

$$\times \left( r(s,a) + \gamma \sum_{s'} P(s' | s,a)v^*(s') \right),$$

where the first relation is by using the triangle inequality, and then use $|a \cdot b| \leq |a| \cdot |b|$. We further have that,

$$r(s,a) + \gamma \sum_{s'} P(s' | s,a)v^*(s') \leq \frac{R_{\max}}{1 - \gamma}. $$

Thus, continuing from (3), we can further bound (24),

$$|Tv^* - T^\pi v^*| \leq \frac{R_{\max}}{1 - \gamma} \sum_a |\pi^*(a | s) - \pi_0(a | s)|.$$

Thus,

$$\alpha(s)(Tv^* - T^\pi v^*)(s) \leq \max_a \alpha(s) \left\| \pi^* - \pi_0 \right\|_{TV} \frac{R_{\max}}{1 - \gamma}$$

where $\left\| \pi^* - \pi_0 \right\|_{TV}(s) = \sum_a |\pi^*(a | s) - \pi_0(a | s)|$, is the total variation of $\pi^*$ and $\pi_0$ in state $s$.

Finally, the bound is proved tight by an example which attain it as described below:

**Proof.** For the MDP described in figure 7 it is easy to see that for the uniform $\pi_0$:

$$v^* - v_\pi^* = \frac{1}{1 - \gamma} \left( 1 - \alpha/2 \right) \frac{1 - \gamma}{1 - \gamma} = \alpha/2 \frac{1 - \gamma}{1 - \gamma}$$

Next:

$$\frac{\alpha}{1 - \gamma} \left\| v^*(s) - \sum_a \pi_0(a | s) q^*(s,a) \right\|$$

$$= \frac{\alpha}{1 - \gamma} \left\| 1 - \frac{1}{1 - \gamma} - \frac{1/2}{1 - \gamma} \right\|$$

$$= \frac{\alpha/2}{1 - \gamma}$$

Figure 7: One State MDP that attains the bound in Proposition 6.

**Appendix E  Proof of Proposition 6**

We start by deriving the bound.

$$\left\| v^*_\alpha - v_\pi^*(\hat{\pi}^*, \pi_0) \right\| = \left\| T_\alpha v^*_\alpha - T_\alpha v_\pi^*(\hat{\pi}^*, \pi_0) \right\|$$

$$\leq \left\| T_\alpha v^*_\alpha - (1 - \alpha) T_\hat{\pi}^*_\alpha - \alpha T^\pi v_\pi^* \right\|$$

$$+ \left\| T_\hat{\pi}^*_\alpha v_\pi^*(\hat{\pi}^*, \pi_0) - (1 - \alpha) T_\hat{\pi}^*_\alpha - \alpha T^\pi v_\pi^* \right\|. \quad (25)$$

Notice that the term added and subtracted is different then the one in (Bertsekas and Tsitsiklis 1995). This difference allows us to get tighter bound then for the considered setup.

We continue by analyzing each term in the above. Starting from the first term in (25).

$$\left\| T_\alpha v^*_\alpha - (1 - \alpha) T_\hat{\pi}^*_\alpha - \alpha T^\pi v_\pi^* \right\|$$

$$\leq (1 - \alpha) \left\| T_\alpha v^*_\alpha - T_\hat{\pi}^*_\alpha \right\|$$

$$+ \alpha \left\| T_\pi v_\pi^*(\hat{\pi}^*, \pi_0) - T^\pi v_\pi^* \right\|$$

$$\leq (1 - \alpha) \left\| T_\beta^{\hat{\pi}^*}(\hat{\pi}^*, \pi_0) - T_\hat{\pi}^*_\alpha \right\|$$

$$+ \alpha \left\| T_\pi v_\pi^*(\hat{\pi}^*, \pi_0) - T^\pi v_\pi^* \right\|$$

$$\leq (1 - \alpha) \gamma \left\| v_\pi^{\hat{\pi}^*, \pi_0} - v_\alpha^* \right\| + \alpha \gamma \left\| v_\pi^{\hat{\pi}^*, \pi_0} - v_\alpha^* \right\|$$

$$= (1 - \alpha) \gamma \left\| v_\pi^{\hat{\pi}^*, \pi_0} - v_\alpha^* \right\| + \alpha \gamma \left\| v_\pi^{\hat{\pi}^*, \pi_0} - v_\alpha^* \right\|$$

$$\leq (1 - \alpha) \gamma \delta + \gamma \left\| v_\pi^{\hat{\pi}^*, \pi_0} - v_\alpha^* \right\|. \quad (26)$$

The first relation is by plugging $T_\hat{\pi}^*$ due to Lemma 2 and the triangle inequality. The second relation is since $T_\hat{\pi}^* = T_\hat{\pi}^*$. The third relation holds since $T_\hat{\pi}^*$ and $T^\pi$ are $\gamma$ contractions in the max-norm.

Plugging (26) and (27) to bound both of the terms in (26). Then,

$$\left\| v^*_\alpha - v_\pi^*(\hat{\pi}^*, \pi_0) \right\| \leq 2(1 - \alpha) \gamma \delta + \gamma \left\| v_\pi^{\hat{\pi}^*, \pi_0} - v_\alpha^* \right\|.$$

By simple algebraic manipulation we conclude the result.

Finally, we prove that this bound is tight (see that different MDP then in (Bertsekas and Tsitsiklis 1995) is used).
Observe at the MDP described in Figure 8. The policy \( \pi_n^* \) is to always choose action \( a_1 \). Hence,
\[
v_\alpha^* = \sum_{n=0}^{\infty} \gamma^n \left[ \gamma \delta (1 - \alpha) - \gamma \delta \alpha \right] = \frac{\gamma \delta (1 - \alpha)}{1 - \gamma}
\]
Now, given value estimation \( \hat{v}_\alpha^* \), such that \( \hat{v}_\alpha^*(s_0) = \delta \), \( \hat{v}_\alpha^*(s_1) = -\delta \), taking always \( a_1 \) is an \( \alpha \)-greedy policy with respect to \( \hat{v}_\alpha^* \):
\[
(1 - \alpha)(\gamma \delta + \gamma \hat{v}_\alpha^*(s_1)) + \alpha(\gamma \delta + \gamma \hat{v}_\alpha^*(s_0)) = 0 = (1 - \alpha)(\gamma \delta + \gamma \hat{v}_\alpha^*(s_0)) + \alpha(\gamma \delta + \gamma \hat{v}_\alpha^*(s_1))
\]
Hence,
\[
v^{\pi_\alpha}(\hat{v}_\alpha^*, \pi_0) = \sum_{n=0}^{\infty} \gamma^n \left[ -\gamma \delta (1 - \alpha) + \gamma \delta \alpha \right] = \frac{\gamma \delta (\alpha - 1)}{1 - \gamma}
\]
Simple arithmetics show that this MDP attains the upper bound.

**Appendix F** A note on approximate model w.r.t. a policy

In this section we elaborate on the relation between Definition 2 and 3. We show that a \((\Delta_r, \Delta_P)\) approximate model results in \((\Delta_r, \Delta_P)\) approximate model w.r.t. any policy \( \pi \). This emphasizes the first assumption is stronger than the other. More interestingly, it is natural to assume one can acquire better approximation for \( \pi \) when the agent interacts with the environment using this policy.

Assume \( \hat{M} \) is a \((\Delta_r, \Delta_P)\) approximation of \( M \). Then we build the following approximate model w.r.t. \( \pi \),
\[
\hat{r}^\pi = \sum_a \pi(a \mid s) \hat{r}(s, a)
\]
\[
\hat{P}^\pi = \sum_a \pi(a \mid s) \hat{P}(s' \mid s, a).
\]
We prove it is a \((\Delta_r, \Delta_P)\) approximation w.r.t. \( \pi \). For any \( s \in S \) we have,
\[
|r^\pi(s) - \hat{r}^\pi(s)| = \left| \sum_a \pi(a \mid s) (r(s, a) - \hat{r}(s, a)) \right| \leq \sum_a \pi(a \mid s) |r(s, a) - \hat{r}(s, a)| \leq \sum_a \pi(a \mid s) \Delta_r = \Delta_r,
\]
where the first relation is by the triangle inequality and the last relation is by assuming a \((\Delta_r, \Delta_P)\) approximate model. Thus, \( \Delta^\pi = \max_s |r^\pi(s) - \hat{r}^\pi(s)| \leq \Delta_r \). Moreover, for any \( s \in S \),
\[
\sum_a |P^\pi(s' \mid s) - \hat{P}^\pi(s' \mid s)| \leq \sum_a \sum_{s'} \pi(a \mid s) \left| P^\pi(s' \mid s, a) - \hat{P}^\pi(s' \mid s, a) \right| \leq \sum_a \pi(a \mid s) \Delta_P = \Delta_P,
\]
where the first relation is by the triangle inequality, and the third relation is by assuming a \((\Delta_r, \Delta_P)\) approximate model. Thus, \( \Delta_P^\pi = \max_s \sum_{s'} |P^\pi(s' \mid s) - \hat{P}^\pi(s' \mid s)| \leq \Delta_P \).

**Appendix G** Proof of Theorem 8

We start by proving a useful lemma, equivalently to [Jiang et al. 2015][Size. 3]. In the following proof we denote by lower subscript the value on the corresponding MDP. Meaning, \( v_M^\pi \) and \( v_\hat{M}^\pi \) are the value functions of policy \( \pi \) measured on MDP \( M \) and \( \hat{M} \), respectively.

**Lemma 15.** Let \( M, \hat{M} \) be MDPs. Let \( \pi^*_M, \pi^*_\hat{M} \) be the \( \alpha \)-optimal policy w.r.t. \( \pi_0 \) on \( M, \hat{M} \), respectively. Then,
\[
0 \leq v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, s_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) \leq \| v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) \| + \| v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) \|
\]

**Proof.** We have that,
\[
v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) \\
\leq v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) + v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) \\
\leq v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) + v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0)
\]

The first relation holds since \( \pi^*_\hat{M} \) is the \( \alpha \)-optimal policy w.r.t. \( \pi_0 \) on \( \hat{M} \); by construction (7), \( \pi^*(\hat{s}^*_\hat{M}, \pi_0) \) has bigger value than \( \pi^*(\hat{s}^*_\hat{M}, \pi_0) \), where \( \pi^* \) is a policy. Thus,
\[
v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) - v_{\hat{M}}^{\pi^*_\hat{M}}(s^*_\hat{M}, \pi_0) \geq 0.
\]

Similarly, since \( \pi^*_M \) is the \( \alpha \)-optimal policy w.r.t. \( \pi_0 \) on \( M \), \( v_{\hat{M}}^{\pi^*_M}(s^*_M, \pi_0) - v_{\hat{M}}^{\pi^*_M}(s^*_M, \pi_0) \geq 0. \)

\[
\square
\]
Using this lemma, we can bound the difference between the performance of \(\pi^\alpha(\pi_\alpha^*, \pi_0)\) relatively to the \(\alpha\)-optimal value.

**Proposition 16.** Let \(\alpha \in [0, 1]\). Assume \(\hat{M}\) is jointly a \((\Delta_{\hat{r}}, \Delta_{\hat{p}})\) approximate model and \((\Delta_{\pi^0}, \Delta_{\pi^0})\) approximate model w.r.t. \(\pi_0\). Let \(\pi_\alpha^*, \hat{\pi}_\alpha^*\) be an \(\alpha\)-optimal policy w.r.t. \(\pi_0\), calculated in \(M, \hat{M}\) respectively. Then,

\[
\left\|u^\pi(\pi_\alpha^*, \pi_0) - u^\pi(\hat{\pi}_\alpha^*, \pi_0)\right\| \leq \frac{2\Delta_{\hat{M}}}{1 - \gamma} \left(1 - \alpha \left(1 - \frac{\Delta_{\pi_0}}{\Delta_{\hat{M}}}\right)\right),
\]

where \(\Delta_{\hat{M}} = \Delta_{\hat{r}} + \gamma R_{\max} \Delta_{\hat{p}}, \Delta_{\pi_0} = \Delta_{\pi^0} + \gamma R_{\max} \Delta_{\pi^0} \).

**Proof.** Lemma [15] translates the problem of bounding the value difference of policies \(\pi_\alpha^*\) and \(\hat{\pi}_\alpha^*\) measured on \(M, \hat{M}\) into bounding terms of the form \(\left\|u^\pi(\pi_\alpha^*, \pi_0) - u^\pi(\hat{\pi}_\alpha^*, \pi_0)\right\|\); the difference of the value of \(\pi'\) measured on \(M, \hat{M}\). For any policy \(\pi'\), we have that

\[
\left\|u^\pi(\pi', \pi_0) - u^\pi(\hat{\pi}', \pi_0)\right\| = \left\|T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| \\
= \left\|T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| \\
\leq \left\|T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| \\
+ \gamma \left\|u^\pi(\pi', \pi_0) - u^\pi(\hat{\pi}', \pi_0)\right\|. \tag{28}
\]

In the first relation we used the fact the \(u^\pi\) is the fixed point of \(\pi\), in the second relation, we added and subtracted \(T_{\hat{M}} u^\pi(\pi', \pi_0)\) and in the third relation we used the fact that \(T_{\hat{M}}\) is a \(\gamma\)-contraction for any \(\gamma\)-discounted MDP.

We now bound the first term in (28). For any \(s \in S\) we have that,

\[
\left\|T_{\hat{M}} u^\pi(\pi', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| (s) \\
\leq (1 - \alpha)T_{\hat{M}} u^\pi(\pi', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| (s) \\
+ \alpha T_{\hat{M}} u^\pi(\pi', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| (s). \tag{29}
\]

Where the first relation holds since \(u^\pi(\pi', \pi_0)\) is an \(\alpha\)-convex combination of policies \(\pi'\).

We bound both of the terms in (29). For the first term we have that,

\[
\left\|T_{\hat{M}} u^\pi(\pi', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| (s) \\
= \sum_a \pi(a \mid s) \left(\Delta_{\hat{r}}(s, a) + \gamma \sum_{s'} \Delta_{\hat{p}}(s', s, a)^0 u_{\hat{M}}(\hat{\pi}^0(s', \pi_0))\right) \\
\leq \sum_a \pi(a \mid s) \left(\Delta_{\hat{r}}(s, a) + \gamma \sum_{s'} \Delta_{\hat{p}}(s', s, a)^0 u_{\hat{M}}(\hat{\pi}^0(s', \pi_0))\right) \\
\leq \sum_a \pi(a \mid s) \left(\Delta_{\hat{r}} + \gamma R_{\max} \gamma \Delta_{\hat{p}}\right) \\
\leq \sum_a \pi(a \mid s) \left(\Delta_{\hat{r}} + \gamma R_{\max} \gamma \Delta_{\hat{p}}\right) \\
\leq \sum_a \pi(a \mid s) \left(\Delta_{\hat{r}} + \gamma R_{\max} \gamma \Delta_{\hat{p}}\right) \\
\leq \sum_a \pi(a \mid s) \left(\Delta_{\hat{r}} + \gamma R_{\max} \gamma \Delta_{\hat{p}}\right) \\
\Delta_{\hat{r}} + \gamma R_{\max} \gamma \Delta_{\hat{p}} = \Delta_{\hat{M}}. \tag{30}
\]

In the first relation we used the definition of the fixed policy Bellman operator and defined the state-action reward and transition difference,

\[
\Delta_{\hat{r}}(s, a) = r(s, a) - \hat{r}(s, a), \\
\Delta_{\hat{p}}(s', s, a) = \hat{P}(s' \mid s, a) - \hat{P}(s' \mid s, a).
\]

In the second relation we used the triangle inequality. In the third relation we used the bound \(\Delta_{\hat{r}}(s, a) \leq \Delta_{\hat{r}}\), by assumption, and \(u(\hat{\pi}^0) (s) \leq R_{\max} \gamma \Delta_{\hat{p}}\). In the forth relation we used \(\sum_s \Delta_{\hat{p}}(s' \mid s, a) \leq \Delta_{\hat{p}}\), by assumption, and in the last relation, the fact that \(\pi(a \mid s)\) is a probability function which sums to one.

Using similar analysis for the second term in (29), while using the assumption on \(M\) which is also \((\Delta_{\pi^0}, \Delta_{\pi^0})\) w.r.t. \(\pi_0\), we obtain that for any \(s \in S\),

\[
\left\|T_{\hat{M}} u^\pi(\pi', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| (s) \leq \Delta_{\pi_0} + \gamma R_{\max} \gamma \Delta_{\pi_0}, \tag{31}
\]

Since both (30) and (32) holds uniformly on all states, we have that

\[
\left\|T_{\hat{M}} u^\pi(\pi', \pi_0) - T_{\hat{M}} u^\pi(\hat{\pi}', \pi_0)\right\| (s) \\
\leq (1 - \alpha)\Delta_{\hat{M}} + \alpha \Delta_{\pi_0}. \tag{33}
\]

Thus, by plugging this result in (28) an simple algebraic manipulation we obtain that for any policy \(\pi'\),

\[
\left\|u^\pi(\pi', \pi_0) - u^\pi(\hat{\pi}', \pi_0)\right\| \leq \frac{\Delta_{\hat{M}}}{1 - \gamma} \left(1 - \alpha \left(1 - \frac{\Delta_{\pi_0}}{\Delta_{\hat{M}}}\right)\right). \tag{33}
\]

By using Lemma [15] and (33), that holds for any policy, and specifically for both \(\pi_\alpha^*\) and \(\hat{\pi}_\alpha^*\), we conclude the result. \(\square\)

The proof is similar to the one in Theorem [7] while using Proposition [16] instead of [6].
We conclude the proof of Theorem 8 using the aforementioned lemma.

\[
\|v^* - v^{\pi_0}(\pi_0, \pi_0)\| \leq \|v^* - \nu^*_\alpha\| + \|v^* - \nu^{\pi_0}(\pi_0, \pi_0)\| \\
\leq \frac{\alpha L}{1 - \gamma} + \|v^* - \nu^{\pi_0}(\pi_0, \pi_0)\| \\
= \frac{\alpha L}{1 - \gamma} + \|v^M(\pi_0, \pi_0) - \nu^M(\pi_0, \pi_0)\| \\
\leq \frac{\alpha L}{1 - \gamma} + \frac{2\Delta M}{1 - \gamma} \left(1 - \alpha \left(1 - \frac{\Delta_0}{\Delta M}\right)\right).
\]

The first relation holds by the triangle inequality, the second relation by Proposition 6, the third relation by the equivalence of values due to Lemma 1 and the forth relation holds by Lemma 16.

**Appendix H Supplementary material for Section 5**

In this section we give the proofs for the statements of Section 5.

The proof of Lemma 9 is given as follows.

**Proof.** By using the definition of \( T^\pi_{\alpha} \) [13], and due to \( v^*_\alpha = \max_\alpha q^*_\alpha(\cdot, a) \), we have that,

\[
q^*_\alpha(s, a) = T^\pi_{\alpha} q^*_\alpha(s, a) \\
= r(s, a) + \gamma \sum_{a'} P_a(s' | s, a) \max_{\alpha'} q^*_{\alpha'}(s', a') \\
(1 - \alpha)(r(s, a) + \gamma \sum_{a'} P_a(s' | s, a) v^*_{\alpha'}(s')) \\
+ \alpha \sum_a \pi(a' | s)(r(s, a')) + \gamma \sum_{a'} P(a' | s, a') v^*_{\alpha'}(s')) \\
= (1 - \alpha)q^{\pi_0}(\pi_0, \pi_0)(s, a) + \alpha \sum_a \pi(a' | s)q^{\pi_0}(\pi_0, \pi_0)(s, a'),
\]

where in the last relation we used (12).

The proof of Lemma 10 is given as follows.

**Proof.** It is easy to verify this operator is a \( \gamma \)-contraction using standard arguments [Bertsekas and Tsitsiklis 1995]. We prove that the fixed point of \( T^\pi_{\alpha} \) is \( q^{\pi_0}(\pi_0, \pi_0) \). First, by using the max operator w.r.t. the action on the result in Lemma 9 we get

\[
v^*_\alpha = (1 - \alpha) \max_{\alpha} q^{\pi_0}(\pi_0, \pi_0)(\cdot, a) + \alpha \Pi_0 q^{\pi_0}(\pi_0, \pi_0).
\]

Consider the definition of \( q^{\pi_0}(\pi_0, \pi_0) \) [12]. We have that,

\[
q^{\pi_0}(\pi_0, \pi_0)(s, a) = r(s, a) + \gamma \sum_{a'} P(s' | s, a) v^*(s') \\
= r(s, a) + \gamma (1 - \alpha) \sum_{a'} P(s' | s, a) \max_{a'} q^{\pi_0}(\pi_0, \pi_0)(s', a') \\
+ \gamma \alpha \sum_{a'} P(s' | s, a) \pi_0(a') q^{\pi_0}(\pi_0, \pi_0)(s', a') \\
= T^E q^{\pi_0}(\pi_0, \pi_0)(s, a),
\]

where the first relation holds by plugging [34] and the third relation holds by identifying the operator \( T^E \).

The proof of Lemma 11 is given as follows.

**Proof.** For brevity, we omit the superscript \( q^E \) and write as \( q_t \). The updating equations of Algorithm 3 can be written as

\[
q_{t+1}(s, a^n) = (1 - \eta_t) q_t(s, a^n) + \eta_t T^E q_t(s, a^n) - w_t,
\]

where

\[
w_t = r_t + \gamma (1 - \alpha) v(s_{t+1}) \\
+ \gamma \alpha v^{\pi_0}(s_{t+1}) - T^E q_t(s, a^n),
\]

and

\[
v(s_{t+1}) = \max_{a'} q(s_{t+1}, a')
\]

\[
v^{\pi_0}(s_{t+1}) = \sum_{a'} \pi_0(a') | s_{t+1}) q(s_{t+1}, a').
\]

We let \( \mathcal{F}_t = \{ \mathcal{H}_{t-1}, s_t, a_{t+1}^{env}, X_t, q_0^{ch} \}, \) where \( \mathcal{H}_{t-1} \) is the entire history until and including time \( t - 1 \), i.e., the filtration includes both the chosen action, before deciding whether to act with it or according to \( \pi_0 \), and the acted action.

We have that,

\[
E \left[ r_t + \gamma (1 - \alpha) q(s_{t+1}, a_{t+1}^{env}) | \mathcal{F}_t \right] \\
= r(s_t, a_{t+1}^{env}) + \gamma (1 - \alpha) \sum_{a'} P(s' | s, a_{t+1}^{env}) q(s', a') \\
+ \gamma \alpha \sum_{a'} P(s' | s, a_{t+1}^{env}) \pi_0(a') q(s', a'),
\]

and \( \mathbb{E}[w_t | \mathcal{F}_t] = 0 \). It is also easy to see that \( \mathbb{E}[w^2_t | \mathcal{F}_t] \leq A + B ||Q||^2_\infty \).

Thus, according to [Bertsekas and Tsitsiklis 1995] Proposition 4.4 the process converges to the fixed point contraction operator \( T^E q^{\pi_0}(\pi_0, \pi_0) \) (see Lemma 10).

**Appendix I Proof of Theorem 12**

We will use the following result [Singh et al 2000] [Lemma 1].

**Lemma 17.** Consider a stochastic process \( (\alpha_t, \Delta_t, \Delta_t, f_t) \), \( t \geq 0 \), where \( \alpha_t, \Delta_t, f_t : X \to \mathbb{R} \) satisfy the equations

\[
\Delta_{t+1}(x) = (1 - \alpha_t(x)) \Delta_t(x) + \alpha_t(x) f_t(x),
\]

\( x \in X, \ t = 0, 1, 2, \ldots \) (35)
Let $\mathcal{F}_t$ be a sequence of increasing $\sigma$-fields such that $\alpha_n$ and $\Delta_n$ are $\mathcal{F}_n$-measurable, $t = 1, 2, \ldots$. Assume that the following hold:

1. The set $X$ is finite.
2. $0 \leq \alpha_t(x) \leq 1$, $\sum_t \alpha_t(x) = \infty$, $\sum_t \alpha_t^2(x) < \infty$ w.p. 1.
3. $|E[ f_t(\cdot) \mid \mathcal{F}_t]| \leq \kappa |\Delta_t| + c_t$, where $\kappa \in [0,1)$ and $c_t$ converges to zero w.p. 1.
4. $\text{Var} \{ F_t(\cdot) \mid \mathcal{F}_t \} \leq K(1 + |\Delta_t|)^2$, where $K$ is some constant.

Then, $\Delta_t$ converges to zero with probability one (w.p. 1).

Observe that $q_{E,t}$ has updating rule as in Expected $\alpha$-Q-Learning (see Algorithm [5]), and is independent of $q$. Due to the assumptions that $\forall s \in S, \forall a \in A$

$$\sum_{t=0}^{\infty} \eta_t(s_t = s, a_t^{env} = a) = \infty,$$

we get that the sequence $\{q_{E,t}\}_{t=0}^{\infty}$ converges to $q^{\pi_n}(\pi_n, \pi_0)$ w.p. 1.

We now show the updating of $\eta$ in Algorithm [2] to have the form of [35]. Define the following difference

$$\Delta_t(s, a) = q_t(s, a) - q^*_a(s, a),$$

and consider the filtration $\mathcal{H}_t = \{H_{t-1}, s_t, a_{\text{chosen}}\}$.

By decreasing $q^*_a(s, a)$ from both side of the updating equations of $q$ in Algorithm [2] we obtain for any $a \in A$,

$$\Delta_{t+1}(s_t, a) = (1 - \eta_t)\Delta_t(s_t, a) f_t(s_t, a).$$

If $\bar{a} = a_{\text{chosen}}$ then,

$$f_t(s_t, \bar{a}) = r_t + \gamma v_t(s_{t+1}) - q^*_a(s, a),$$

whereas for $\bar{a} \neq a_{\text{chosen}}$,

$$f_t(s_t, \bar{a}) = X_t q^*(\pi_n, \pi_0)(s_t, \bar{a}) + \left(1 - X_t\right)(r_t + \gamma v_t(s_{t+1})) + X_t q_E(s_t, \bar{a}) - q^*(\pi_n, \pi_0)(s_t, \bar{a}) - q^*_a(s, a).$$

We now show that for all action entries $\bar{a} \in A$,

$$E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t] \leq \kappa |\Delta_t(s_t, \bar{a})| + c_t,$$

and $c_t$ converges to zero w.p. 1.

If $\bar{a} = a_{\text{chosen}}$ then,

$$E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t] = (1 - \alpha)q_t(s_t, \bar{a}) + \gamma \sum_{s'} P(s' \mid s_t, \bar{a}) \max_{a'} q_t(s', a')$$

$$\alpha(r^{\pi_0}(s_t) + \gamma \sum_{s'} P^{\pi_0}(s' \mid s_t) \max_{a'} q_t(s', a')) - q^*_a(s, a)$$

$$= T_0^q q_t(s_{t+1}, a') - q^*_a(s, a).$$

Thus, for this case,

$$|E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t]| = ||T_0^q q_t(s_{t+1}, a') - q^*_a(s, a)||$$

$$= ||T_0^q q_t(s_{t+1}, a') - q^*_a(s, a)||$$

$$\leq \gamma||q_t(s_{t+1}, a') - q^*_a(s, a)||,$$

meaning, $c_t = 0$ for this entry. We now turn to the case $\bar{a} \neq a_{\text{chosen}}$.

$$E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t] = (1 - \alpha)q^{\pi_n}(\pi_n, \pi_0)(s_t, \bar{a}) - q^*_a(s, \bar{a})$$

$$+ \alpha(r^{\pi_0} + \gamma \sum_{s'} P^{\pi_0}(s' \mid s) \max_{a'} q_t(s', a'))$$

$$+ (1 - \alpha)(q_E(s_t, \bar{a}) - q^{\pi_n}(\pi_n, \pi_0)(s_t, \bar{a})).$$

Define

$$c_t = (1 - \alpha)(q_{E,t}(s_t, \bar{a}) - q^{\pi_n}(\pi_n, \pi_0)(s_t, \bar{a})).$$

See that $c_t$ converges to zero w.p. 1, since $q_{E,t}$ converges to $q^{\pi_n}(\pi_n, \pi_0)$. Furthermore, using Lemma [7] we have that

$$(1 - \alpha)q^{\pi_n}(\pi_n, \pi_0)(s_t, \bar{a}) - q^*_a(s, \bar{a})$$

$$= - \alpha(r^{\pi_0} + \gamma \sum_{s'} P^{\pi_0}(s' \mid s) \max_{a'} q^*_a(s', a')).$$

Thus,

$$E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t] = - \alpha(r^{\pi_0} + \gamma \sum_{s'} P^{\pi_0}(s' \mid s) \max_{a'} q^*_a(s', a'))$$

$$+ \alpha(r^{\pi_0} + \gamma \sum_{s'} P^{\pi_0}(s' \mid s) \max_{a'} q_t(s', a') + c_t$$

$$= \alpha\gamma \sum_{s'} P^{\pi_0}(s' \mid s)(\max_{a'} q_t(s', a') - \max_{a'} q^*_a(s', a')) + c_t$$

$$= \alpha\gamma \sum_{s'} P^{\pi_0}(s' \mid s)\max_{a'} (|q_t(s', a') - q^*_a(s', a')| + c_t$$

$$= \alpha\gamma \max_{s', a'} |q_t - q^*_a| + c_t$$

Where in the first relation we applied Lemma [9]. By showing similar result for $-E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t]$, we conclude that

$$E[f_t(s_t, \bar{a}) \mid \mathcal{F}_t] \leq \alpha\gamma \max_{s', a'} |q_t - q^*_a| + c_t,$$

where $c_t$ converges to zero w.p. 1. The Var$f_t(\cdot, \cdot)$ can be bounded by $K(1 + |\Delta_t|)^2$, since the reward is bounded and $\sum_{t=0}^{\infty} \eta_t^2(s_t = s, a_t^{env} = a) < \infty$.

We conclude that all conditions of Lemma [17] are satisfied for each $\bar{a} \in A$ and, thus, Lemma [17] establishes the convergence of the procedure.