Exploration Conscious Reinforcement Learning Revisited

Lior Shani [*][1] Yonathan Efroni [*][1] Shie Mannor [1]

Abstract

The Exploration-Exploitation tradeoff arises in Reinforcement Learning when one cannot tell if a policy is optimal. Then, there is a constant need to explore new actions instead of exploiting past experience. In practice, it is common to resolve the tradeoff by using a fixed exploration mechanism, such as ϵ-greedy exploration or by adding Gaussian noise, while still trying to learn an optimal policy. In this work, we take a different approach and study exploration-conscious criteria, that result in optimal policies with respect to the exploration mechanism. Solving these criteria, as we establish, amounts to solving a surrogate Markov Decision Process. We continue and analyze properties of exploration-conscious optimal policies and characterize two general approaches to solve such criteria. Building on the approaches, we apply simple changes in existing tabular and deep Reinforcement Learning algorithms and empirically demonstrate superior performance relatively to their non-exploration-conscious counterparts, both for discrete and continuous action spaces.

1. Introduction

The main goal of Reinforcement Learning (RL) (Sutton et al., 1998) is to find an optimal policy for a given decision problem. A major difficulty arises due to the Exploration-Exploitation tradeoff, which characterizes the omnipresent tension between exploring new actions and exploiting the so-far acquired knowledge. Considerable line of work has been devoted for dealing with this tradeoff. Algorithms that explicitly balance between exploration and exploitation were developed for tabular RL (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Jaksch et al., 2010; Osband et al., 2013). However, generalizing these results to approximate RL, i.e, when using function approximation, remains an open problem. On the practical side, recent works combined more advanced exploration schemes in approximate RL (e.g, Bellemare et al. (2016); Fortunato et al. (2017)), inspired by the theory of tabular RL. Nonetheless, even in the presence of more advanced mechanisms, ϵ-greedy exploration is still applied (Bellemare et al., 2017; Dabney et al., 2018; Osband et al., 2016). More generally, the traditional and simpler ϵ-greedy scheme (Sutton et al., 1998; Asadi & Littman, 2016) in discrete RL, and Gaussian action noise in continuous RL, are still very useful and popular in practice (Mnih et al., 2015; 2016; Silver et al., 2014; Schulman et al., 2017; Horgan et al., 2018), especially due to their simplicity.

These types of exploration schemes share common properties. First, they all fix some exploration parameter beforehand, e.g, ϵ, the 'inverse temperature' β, or the action variance σ for the ϵ-greedy, soft-max and Gaussian exploration schemes, respectively. By doing so, the balance between exploration and exploitation is set. Second, they all explore using a random policy, and exploit using current estimate of the *optimal policy*. In this work, we follow a different approach, when using these fixed exploration schemes: exploiting by using an estimate of the optimal policy w.r.t. the *exploration mechanism*.

Exploration-Consciousness is the main reason for the improved performance of on-policy methods like Sarsa and Expected-Sarsa (Van Seijen et al., 2009) over Q-learning during training (Sutton et al., 1998)[Example 6.6: Cliff Walking]. Imagine a simple Cliff-Walking problem: The goal of the agent is to reach the end without falling of the cliff, where the optimal policy is to go alongside the cliff. While using a fixed-exploration scheme, playing a near optimal policy which goes alongside the cliff will lead to a significant sub-optimal performance. This, in turn, will hurt the acquisition of new experience needed to learn the optimal policy. However, learning to act optimally w.r.t. the exploration scheme can mitigate this difficultly; the agent learns to reach the goal while keeping a safe enough distance from the cliff.

In the past, tabular q-learning-like exploration-conscious algorithms were suggested (John, 1994; Littman et al., 1997; Van Seijen et al., 2009). Here we take a different approach, and focus on exploration conscious *policies*. The main contributions of this work are as follows:

[*]Equal contribution [1]Department of Electrical Engineering, Technion, Haifa, Israel. Correspondence to: Lior Shani <shanlior@gmail.com>, Yonathan Efroni <jonathan.efroni@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

- We define exploration-consciousness optimization criteria, for discrete and continuous actions spaces. The criteria are interpreted as finding an optimal policy within a restricted set of policies. Both, we show, can be reduced to solving a surrogate MDP. The surrogate MDP approach, to the best of our knowledge, is a new one, and serves us repeatedly in this work.

- We formalize a bias-error sensitivity tradeoff. The solutions are biased w.r.t. the optimal policy, yet, are less sensitive to approximation errors.

- We establish two fundamental approaches to practically solve Exploration-Conscious optimization problems. Based on these, we formulate algorithms in discrete and continuous action spaces, and empirically test the algorithms on the Atari and MuJoCo domains.

2. Preliminaries

Our framework is the infinite-horizon discounted Markov Decision Process (MDP). An MDP is defined as the 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (Puterman, 1994), where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a compact space, $P \equiv P(s'|s,a)$ is a transition kernel, $R \equiv r(s,a) \in [0, R_{\max}]$ is a bounded reward function, and $\gamma \in [0,1)$. Let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ be a stationary policy, where $\mathcal{P}(\mathcal{A})$ is a probability distribution on $\mathcal{A}$, and denote Π as the set of deterministic policies, $\pi \in \Pi : \mathcal{S} \to \mathcal{A}$. Let $v^\pi \in \mathbb{R}^{|\mathcal{S}|}$ be the value of a policy π, defined in state s as $v^\pi(s) \equiv \mathbb{E}^\pi_{|s}[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)]$, where $a_t \sim \pi(s_t)$, and $\mathbb{E}^\pi_{|s}$ denotes expectation w.r.t. the distribution induced by π and conditioned on the event $\{s_0 = s\}$. It is known that $v^\pi = \sum_{t=0}^\infty \gamma^t (P^\pi)^t r^\pi = (I - \gamma P^\pi)^{-1} r^\pi$, with the component-wise values $[P^\pi]_{s,s'} \triangleq \mathbb{E}_{a\sim\pi}[P(s' \mid s,a)]$ and $[r^\pi]_s \triangleq \mathbb{E}_{a\sim\pi}[r(s,a)]$. Furthermore, the q-function of π is given by $q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) v^\pi(s')$, and represents the value of taking an action a from state s and then using the policy π.

Usually, the goal is to find π^* yielding the optimal value, $\pi^* \in \arg\max_{\pi\in\Pi} \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)]$, and the optimal value is $v^* = v^{\pi^*}$. It is known that optimal deterministic policy always exists (Puterman, 1994). To achieve this goal the following classical operators are defined (with equalities holding component-wise). $\forall v, \pi :$

$$T^\pi v = r^\pi + \gamma P^\pi v, \; Tv = \max_\pi T^\pi v, \tag{1}$$

$$\mathcal{G}(v) = \{\pi : T^\pi v = Tv\}, \tag{2}$$

where T^π is a linear operator, T is the optimal Bellman operator and both T^π and T are γ-contraction mappings w.r.t. the max norm. It is known that the unique fixed points of T^π and T are v^π and v^*, respectively. $\mathcal{G}(v)$ is the standard set of 1-step greedy policies w.r.t. v. Furthermore, given v^*, the set $\mathcal{G}(v^*)$ coincides with that of stationary optimal policies. It is also useful to define the q-optimal Bellman operator, which is a γ-contraction, with fixed point q^*.

$$T^q q(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} q(s',a'), \tag{3}$$

In this work, the use of *mixture policies* is abundant. We denote the $\alpha \in [0,1]$-convex mixture of policies $\pi_1, \; \pi_2$ by $\pi^\alpha(\pi_1,\pi_2) \triangleq (1-\alpha)\pi_1 + \alpha\pi_2$. Importantly, $\pi^\alpha(\pi_1,\pi_2)$ can be interpreted as a stochastic policy s.t with w.p $(1-\alpha)$ the agent acts with π_1 and w.p α acts with π_2.

3. The α-optimal criterion

In this section, we define the notion of α-optimal policy w.r.t. a policy, π_0. We then claim that finding an α-optimal policy can be done by solving a *surrogate* MDP. We continue by defining the surrogate MDP, and analyze some basic properties of the α-optimal policy.

Let $\alpha \in [0,1]$. We define π^*_{α,π_0} to be the α-optimal policy w.r.t. π_0, and is contained in the following set,

$$\pi^*_{\alpha,\pi_0} \in \arg\max_{\pi'\in\Pi} \mathbb{E}^{\pi^\alpha(\pi',\pi_0)} \left[\sum_{t=0} \gamma^t r(s_t,a_t)) \right], \tag{4}$$

or, $\pi^*_{\alpha,\pi_0} \in \arg\max_{\pi'} v^{\pi^\alpha(\pi',\pi_0)}$, where $a_t \sim \pi^\alpha(\pi',\pi_0)$ and $\pi^\alpha(\pi',\pi_0)$ is the α-convex mixture of π' and π_0, and thus a probability distribution. For brevity, we omit the subscript π_0, and denote the α-optimal policy by π^*_α throughout the rest of the paper. The α-optimal value (w.r.t. π_0) is $v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$, the value of the policy $\pi^\alpha(\pi^*_\alpha,\pi_0)$. In the following, we will see the problem is equivalent to solving a surrogate MDP, for which an optimal deterministic policy is known to exist. Thus, there is no loss optimizing over the set of deterministic policies Π.

Optimization problem (4) can be viewed as optimizing over a restricted set of policies: all policies that are a convex combination of π_0 with a fixed α. Naturally, we can consider in (4) a state-dependent $\alpha(s)$ as well, and some of the results in this work will consider this scenario. In other words, π^*_α is the best policy an agent can act with, if it plays w.p $(1-\alpha)$ according to π^*_α, and w.p α according to π_0, where π_0 can be any policy. The relation to the ϵ-greedy exploration setup becomes clear when π_0 is a uniform distribution on the actions, and set $\alpha = \epsilon$ instead of α. Then, π^*_α is optimal w.r.t. the ϵ-greedy exploration scheme; the policy would have the largest accumulated reward, relatively to all other policies, when acting in an ϵ-greedy fashion w.r.t. it.

We choose to name the policy as the α- and not ϵ-optimal to prevent confusion with other frameworks. The ϵ-optimal policy is a notation used in the context of PAC-MDP type of analysis (Strehl et al., 2009), and has a different meaning than the objective in this work (4).

3.1. The α-optimal Bellman operator, α-optimal policy and policy improvement

In the previous section, we defined the α-optimal policy and the α-optimal value, π^*_α and $v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$, respectively. We start this section by observing that problem (4) can be viewed as solving a *surrogate MDP*, denoted by $\mathcal{M}_\alpha$. We define the Bellman operators of the surrogate MDP, and use them to prove an important improvement property.

Define the surrogate MDP as $\mathcal{M}_\alpha = (\mathcal{S}, \mathcal{A}, P_\alpha, R_\alpha, \gamma)$.

$$\begin{aligned} &\forall a \in \mathcal{A},\ r_\alpha(s,a) = (1-\alpha) r(s,a) + \alpha r^{\pi_0}(s), \\ &P^\pi_\alpha(s' \mid s,a) = (1-\alpha) P(s' \mid s,a) + \alpha P^{\pi_0}(s' \mid s), \end{aligned} \quad (5)$$

are its reward and dynamics, and rest of its ingredients are similar to $\mathcal{M}$. We denote the value of a policy π on $\mathcal{M}_\alpha$ by v^π_α, and the optimal value on $\mathcal{M}_\alpha$ by v^*_α. The following simple lemma relates the value of a policy π, measured on $\mathcal{M}$ and $\mathcal{M}_\alpha$ (see proof in Appendix D).

Lemma 1. *For any policy π, $v^\pi_\alpha = v^{\pi^\alpha(\pi,\pi_0)}$. Thus, an optimal policy on $\mathcal{M}_\alpha$ is the α-optimal policy π^*_α (4).*

The fixed-policy and optimal Bellman operators of $\mathcal{M}_\alpha$ are denoted by T^π_α and T_α, respectively. Again, for brevity we omit π_0 from the definitions. Notice that T^π_α and T_α are γ-contractions as being Bellman operators of a γ-discounted MDP. The following Lemma relates T^π_α and T_α to the Bellman operators of the original MDP, $\mathcal{M}$. Furthermore, it stresses a non-trivial relation between the α-optimal policy π^*_α and the α-optimal value, $v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$.

Proposition 2. *The following claims hold for any policy π:*

1. *$T^\pi_\alpha = (1-\alpha)T^\pi + \alpha T^{\pi_0}$, with fixed point $v^\pi_\alpha = v^{\pi^\alpha(\pi,\pi_0)}$.*
2. *$T_\alpha = (1-\alpha)T + \alpha T^{\pi_0}$, with fixed point $v^*_\alpha = v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$.*
3. *An α-optimal policy is an optimal policy of $\mathcal{M}_\alpha$ and is greedy w.r.t. v^*_α, $\pi^*_\alpha \in \mathcal{G}(v^*_\alpha) = \{\pi' : T^{\pi'} v^*_\alpha = T v^*_\alpha\}$.*

In previous works, e.g. (Asadi & Littman, 2016), the operator $(1-\epsilon)T + \epsilon T^{\pi_0}$ was referred to as the ϵ-greedy operator. Lemma 2 shows this operator is T_α (with $\alpha = \epsilon$), the optimal Bellman operator of the defined surrogate MDP $\mathcal{M}_\alpha$. This lemma leads to the following important property.

Proposition 3. *Let $\alpha \in [0,1)$, $\beta \in [0,\alpha]$, π_0 be a policy, and π^*_α be the α-optimal policy w.r.t π_0. Then, $v^{\pi_0} \le v^{\pi^\alpha(\pi^*_\alpha,\pi_0)} \le v^{\pi^\beta(\pi^*_\alpha,\pi_0)}$, with equality iff $v^{\pi_0} = v^*$.*

The first relation $v^{\pi_0} \le v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$, $\pi^\alpha(\pi^*_\alpha,\pi_0)$ is better than π_0, is trivial and holds by definition (4). The non-trivial statement is the second one. It asserts that given π^*_α, it is worthwhile to use the mixture policy $\pi^\beta(\pi^*_\alpha,\pi_0)$ with $\beta < \alpha$; use π_0 with smaller probability. Specifically, better performance, compared to $\pi^\alpha(\pi^*_\alpha,\pi_0)$, is assured when using the deterministic policy π^*_α, by setting $\beta = 0$.

In section 6, we demonstrate the empirical consequences of the improvement lemma, which, to our knowledge, has not yet been stated. Furthermore, the improvement lemma is unique to the defined optimization criterion (4). We will show that alternative definitions of exploration conscious criteria does not necessarily have this property. Moreover, one can use Proposition 3 to generalize the notion of the 1-step greedy policy (2), as was done in Efroni et al. (2018) with multiple-step greedy improvement. We leave studying this generalization and its Policy Iteration scheme for future work, and focus on solving (4) a single time.

3.2. Performance bounds in the presence of approximations

We now consider an approximate setting and quantify a bias - error sensitivity tradeoff in $\pi^\alpha(\hat{\pi}^*_\alpha,\pi_0)$, where $\hat{\pi}^*_\alpha$ is an approximated α-optimal policy. We formalize an intuitive argument; as α increases the bias relatively to the optimal policy increases. Yet, the sensitivity to errors decreases, since the agent uses π_0 w.p. α regardless of errors.

Definition 1. *Let v^* be the optimal value of an MDP, $\mathcal{M}$. We define $L(s) \triangleq v^*(s) - T^{\pi_0} v^*(s) \ge 0$, to be the Lipschitz constant w.r.t. π_0 of the MDP at state s. We further define the upper bound on the Lipschitz constant $L \triangleq \max_s L(s)$.*

Definition 1 defines the 'Lipschitz' property of the optimal value, v^*. Intuitively, $L(s)$ quantifies a degree of 'smoothness' of the *optimal value*. A small value of $L(s)$ indicates that if one acts according to π_0 once and then continue playing the optimal policy from state s, a great loss will not occur. Large values of $L(s)$ indicate that using π_0 from state s leads to an irreparable outcome (e.g, falling off a cliff). The following theorem formalizes a bias-error sensitivity tradeoff. As α increases, the bias increases, while the sensitivity to errors decreases (see proof in Appendix H).

Theorem 4. *Let $\alpha \in [0,1]$. Assume $\hat{v}^*_\alpha$ is an approximate α-optimal value s.t $\|v^*_\alpha - \hat{v}^*_\alpha\| = \delta$ for some $\delta \ge 0$. Let $\hat{\pi}^*_\alpha$ be the greedy policy w.r.t. $\hat{v}^*_\alpha$, $\hat{\pi}^*_\alpha \in \mathcal{G}(\hat{v}^*_\alpha)$. Then, the performance relatively to the optimal policy is bounded by,*

$$\left\| v^* - v^{\pi^\alpha(\hat{\pi}^*_\alpha,\pi_0)} \right\| \le \underbrace{\frac{\alpha L}{1-\gamma}}_{\text{Bias}} + \underbrace{\frac{2(1-\alpha)\gamma\delta}{1-\gamma}}_{\text{Sensitivity}}.$$

When the bias of the α-optimal value relatively to the optimal one is small, solving (4) does not lead to a great loss relatively to the optimal performance. The bias can be bounded by the 'Lipschitz' property L of the MDP. For a state dependent $\alpha(s)$, the bias bound changes to be dependent on $\max_s \alpha(s)L(s)$. This highlights the importance of prior knowledge when using (4). Choosing π_0 (possibly state-wise) s.t. $\max_s \alpha(s)L(s)$ is small, allows to use a bigger α, while the bias is small. The sensitivity term upper bounds the performance of $\pi^\alpha(\hat{\pi}^*_\alpha,\pi_0)$ relatively to the α-optimal value, and is less sensitive to errors as α increase.

The bias term is derived by using the structure of $\mathcal{M}_\alpha$, and is not a direct application of the Simulation Lemma (Kearns & Singh, 2002; Strehl et al., 2009); applying it would lead to a bias of $\frac{\alpha R_{\max}}{(1-\gamma)^2}$. For the sensitivity term, we generalize (Bertsekas & Tsitsiklis, 1995)[Proposition 6.1] (see Appendix G). There, a $(1-\alpha)$ factor does not exists.

4. Exploration-Conscious Continuous Control

The α-greedy approach from Section 3 relies on an exploration mechanism which is fixed beforehand: π_0 and α are fixed, and an optimal policy w.r.t. them is being calculated (4). However, in continuous control RL algorithms, such as DDPG and PPO (Lillicrap et al., 2015; Schulman et al., 2017), different approach is used. Usually, a policy is being learned, and the exploration noise is injected by perturbing the policy, e.g., by adding to it a Gaussian noise.

We start this section by defining an exploration-conscious optimality criterion that captures such perturbation for the simple case of Gaussian noise. Then, results from Section 3 are adapted to the newly defined criterion, while highlighting commonalities and differences relatively to (4). As in Section 3, we define an appropriate surrogate MDP and we show it can be solved by the usual machinery of Bellman operators. Unlike Section 3, we show that improvement when decreasing the stochasticity does not generally hold. Finally, we prove a similar bias-error sensitivity result: As σ grows, the bias increases, but the sensitivity term decreases.

Instead of restricting the set of policies to the one defined in (4), we restrict our set of policies to be the set of Gaussian policies with a fixed σ^2 variance. Formally, we wish to find the optimal deterministic policy $\mu^*_\sigma : \mathcal{S} \to \mathcal{A}$ in this set,

$$\mu^*_\sigma \in \arg\max_{\mu\in\Pi} \mathbb{E}^{\pi_{\mu,\sigma}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right], \tag{6}$$

where $\pi_{\mu,\sigma}(\cdot \mid s) = \mathcal{N}(\mu(s),\sigma^2)$, is a Gaussian policy with mean $\mu(s)$ and a fixed variance σ^2. We name μ^*_σ and π^*_σ as the mean and σ-optimal policy, respectively. As in (4), we show in the following that solving (6) is equivalent for solving a surrogate MDP. Thus, optimal policy can always be found in the deterministic class of policies Π; mixture of Gaussians would not lead to a better performance in (6).

Similarly to (5), we define a surrogate MDP $\mathcal{M}_\sigma$ w.r.t. to the Gaussian noise and relate it to values of Gaussian policies on the original MDP $\mathcal{M}$. Then, we characterize its Bellman operators and thus establish it can be solved using Dynamic Programming. Define the surrogate MDP as $\mathcal{M}_\sigma = (\mathcal{S},\mathcal{A},P_\sigma,R_\sigma,\gamma)$. For every $a\in\mathcal{A}$,

$$\begin{aligned} r_\sigma(s,a) &= \int_{\mathcal{A}} \mathcal{N}(a';a,\sigma) r(s,a')da', \\ P_\sigma(s' \mid s,a) &= \int_{\mathcal{A}} \mathcal{N}(a';a,\sigma) P(s' \mid s,a')da', \end{aligned} \tag{7}$$

are its reward and dynamics, and denote a value of a policy on $\mathcal{M}_\sigma$ by v^μ_σ. The following results correspond to Lemma 1 and Proposition 2 for the class of Gaussian policies.

Lemma 5. *For any policy π, $v^\pi_{\mu,\sigma} = v^\mu_\sigma$. Thus, an optimal policy on $\mathcal{M}_\sigma$ is the mean optimal policy μ^*_σ (6).*

Proposition 6. *Let π be a mixture of Gaussian policies. Then, the following holds:*

1. *$T^\mu_\sigma = \mathbb{E}^{\pi\sim\pi_{\mu,\sigma}}\, T^\pi$, with fixed point $v^\mu_\sigma = v^{\pi_{\mu,\sigma}}$.*
2. *$T_\sigma = \max_{\mu\in\tilde{\mathcal{A}}} \mathbb{E}^{\pi\sim\pi_{\mu,\sigma}}\, T^\pi$, with fixed point $v^*_\sigma = v^{\pi_{\mu^*_\sigma,\sigma}}$.*
3. *The mean σ-optimal policy μ^*_σ is an optimal policy of $\mathcal{M}_\sigma$ and, $\mu^*_\sigma \in \{\mu : T^{\pi_{\mu,\sigma}} v^*_\sigma = \max_\mu T^{\pi_{\mu,\sigma}} v^*_\sigma\}$.*

Surprisingly, given a σ-optimal policy mean μ^*_σ, an improvement is not assured when lowering the stochasticity by decreasing σ in $\pi_{\mu^*_\sigma,\sigma}$. This comes in contrast to Proposition 3 and highlights its uniqueness (proof in Appendix J).

Proposition 7. *Let $0 \le \sigma' < \sigma$ and let μ^*_σ be the mean σ-optimal policy. There exists an MDP s.t $v^{\pi_{\mu^*,\sigma}} \not\leq v^{\pi_{\mu^*,\sigma'}}$.*

Definition 2. *Let $\mathcal{M}$ be a continuous action space MDP. Assume that exists $L_r,\ L_p \ge 0$, s.t. $\forall s \in \mathcal{S}$, $\forall a_1,a_2 \in \mathcal{A}$, $|r(s,a_1) - r(s,a_2)| \le L_r \left\|a_1 - a_2\right\|_1$ and $\left\|p(\cdot|s,a_1) - p(\cdot|s,a_2)\right\|_{TV} \le L_p \left\|a_1-a_2\right\|_1$. The Lipschitz constant of $\mathcal{M}$ is $\mathcal{L} \triangleq (1-\gamma)L_r + \gamma L_p R_{max}$.*

The following theorem quantifies a bias-error sensitivity tradeoff in σ, similarly to Theorem 4 (see Appendix K).

Theorem 8. *Let $\mathcal{M}$ be an MDP with Lipschitz constant $\mathcal{L}$ and let $\sigma \in \mathbb{R}^{|\mathcal{A}|}_+$. Let v^*_σ be the σ-optimal value of $\mathcal{M}_\sigma$. Let $\hat{v}^*_\sigma$ be an approximation of v^*_σ s.t. $\|v^*_\sigma - \hat{v}^*_\sigma\| = \delta$ for $\delta \ge 0$. Let $\mu^*_\sigma, \hat{\mu}^*_\sigma \in \mathbb{R}^{\mathcal{A}}$ be the greedy mean policy w.r.t. v^*_σ and $\hat{v}^*_\sigma$ respectively. Let $\|\cdot\|_{\sigma^{-2}}$ is the σ^{-2}-weighted euclidean norm. Then,*

$$\left\|v^* - v^{\hat{\pi}^*_\sigma}\right\| \le \underbrace{\frac{\mathcal{L}\left\|\sigma\right\|_1}{2\left(1-\gamma\right)^2}}_{\text{Bias}} + \underbrace{\frac{\gamma\delta\min\{\frac{1}{2}\left\|\mu^*_\sigma - \hat{\mu}^*_\sigma\right\|_{\sigma^{-2}}, 2\}}{1-\gamma}}_{\text{Sensitivity}}.$$

5. Algorithms

In this section, we offer two fundamental approaches to solve exploration conscious criteria using sample-based algorithms: the *Expected* and *Surrogate* approaches. For both, we formulate converging, q-learning-like, algorithms. Next, by adapting DDPG, we show the two approaches can be used in exploration-conscious continuous control as well.

Consider any fixed exploration scheme. Generally, these schemes operate in two stages: (i) Choose a greedy action, a_{chosen}. (ii) Based on a_{chosen} and some randomness generator, choose an action to be applied on the environment, a_{env}. E.g., for ϵ-greedy exploration, w.p. $1-\alpha$ the agent acts with

a_{chosen}, otherwise, with a random uniform policy. While in RL the common update rules use a_{env}, the saved experience is $(s, a_{\text{env}}, r, s')$, in the following we motivate the use of a_{chosen}, and view the data as $(s, a_{\text{chosen}}, a_{\text{env}}, r, s')$.

The two approaches characterized in the following are based on two, inequivalent, ways to define the q-function. For the *Expected* approach the q-function is defined as usual: $q^\pi(s,a)$ represents the value obtained when **taking an action** $a = a_{\text{env}}$ and then acting with π, meaning a is the action chosen in step (ii). Alternatively, for the *Surrogate* approach, the q-function is defined on the 'Surrogate' MDP, i.e., the exploration is viewed as stochasticity of the environment. Then, $q^\pi_\alpha(s,a)$ is the value obtained when a is the action of step (i), i.e., **choosing action** $a = a_{\text{chosen}}$.

5.1. Exploration Conscious Q-Learning

We focus on solving the α-optimal policy (4), and formulate q-learning-like algorithms using the two aforementioned approaches. The *Expected* α-optimal q-function is,

$$q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s,a) \triangleq r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) v^*_\alpha(s') \tag{8}$$

Indeed, $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$ is the usually defined q-function of the policy $\pi^\alpha(\pi^*_\alpha,\pi_0)$ on an MDP $\mathcal{M}$. Here, the action a represents the actual performed action, a_{env}. By relating $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$ to v^*_α it can be easily verified that $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$ satisfies the fixed point equation (see Appendix L),

$$\begin{aligned} & q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s,a) = \\ & r(s,a) + \gamma(1-\alpha) \sum_{s'} P(s' \mid s,a) \max_{a'} q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s',a') \\ & + \gamma\alpha \sum_{s',a'} P(s' \mid s,a) \pi_0(a' \mid s') q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s',a'). \end{aligned} \tag{9}$$

Alternatively, consider the optimal q-function of the surrogate MDP $\mathcal{M}_\alpha$ (5). It satisfies the fixed-point equation

$$q^*_\alpha(s,a) \triangleq r_\alpha(s,a) + \gamma \sum_{s'} P_\alpha(s' \mid s,a) \max_{a'} q^*_\alpha(s',a').$$

The following lemma formalizes the relation between the two q-functions, and shows they are related by a function of the state, and not of the action.

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

The α-optimal policy π^*_α is also an optimal policy of $\mathcal{M}_\alpha$ (Lemma 1). Thus, it is greedy w.r.t. q^*_α, the optimal q of $\mathcal{M}_\alpha$. By Proposition 2.3 it is also greedy w.r.t. $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$, i.e.,

$$\pi^*_\alpha(s) \in \arg\max_{a'} q^*_\alpha(s,a') = \arg\max_{a'} q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s,a').$$

Lemma 9 describes this fact by different means; the two q-functions are related by a function of the state and, thus, the greedy action w.r.t. each is equal. Furthermore, it stresses the fact that the two q-function are not equal.

Before describing the algorithms, we define the following notation for any $q(s,a)$,

$$v(s) = \max_{a'} q(s,a'), \;\; v^\pi(s) = \sum_{a'} \pi(a' \mid s) q(s,a').$$

We now describe the Expected α-Q-learning algorithm (see Algorithm 1), also given in (John, 1994; Littman et al., 1997), and re-interpret it in light of the previous discussion.

The fixed point equation (9), leads us to define the operator T^{Eq}_α for which $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)} = T^{Eq}_\alpha q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$. Expected α-Q-learning (Alg. 1) is a Stochastic Approximation (SA) alg. based on the operator T^{Eq}_α. Given a sample of the form $(s, a_{\text{chosen}}, a_{\text{env}}, r, s')$, it updates $q(s, a_{\text{env}})$ by

$$(1-\eta) q(s, a_{\text{env}}) + \eta \left(r_t + \gamma((1-\alpha) v(s_{t+1}) + \alpha v^{\pi_0}(s_{t+1})) \right) \tag{10}$$

Algorithm 1 Expected α-Q-Learning

```
Initialize: $\alpha \in [0,1]$, $\pi_0$, $q$, learning rate $\eta_t$.
  for $t = 0, 1, ...$ do
    $a_{\text{chosen}} \leftarrow \arg\max_a q_t(s_t, a)$
    $X_t \sim Bernoulli(1-\alpha)$
    $a_{\text{env}} = \begin{cases} a_{\text{chosen}}, & \text{if } X_t = 1 \\ a \sim \pi_0(\cdot \mid s), & \text{if } X_t = 0 \end{cases}$
    $r_t, s_{t+1} \leftarrow ACT(a_{\text{env}})$
    $y_t \leftarrow r_t + \gamma(1-\alpha) v_t(s_{t+1}) + \gamma\alpha v^{\pi_0}_t(s_{t+1})$
    $q(s_t, a_{\text{env}}) \leftarrow (1-\eta_t)\, q(s_t, a_{\text{env}}) + \eta_t y_t$
  end for
  return: $\pi \in \arg\max_a q(\cdot, a)$
```

Its convergence proof is standard and follows by showing T^{Eq}_α is a γ-contraction and using (Bertsekas & Tsitsiklis, 1995)[Proposition 4.4] (see proof in Appendix L.1).

We now turn to describe an alternative algorithm, which operates on the surrogate MDP, $\mathcal{M}_\alpha$, and converges to q^*_α. Naively, given a sample $(s, a_{\text{chosen}}, r, s')$, regular q-learning on $\mathcal{M}_\alpha$ can be used by updating $q(s, a_{\text{chosen}})$ as,

$$(1-\eta_t) q(s, a_{\text{chosen}}) + \eta_t (r_t + \gamma v(s_{t+1})), \tag{11}$$

Yet, this approach does not utilize a meaningful knowledge; when the exploration policy π_0 is played, i.e., when $X_t = 0$, the sample (r_t, s_{t+1}) can be used to update all the action entries from the current state. These entries are also affected by the policy π_0. In fact, we cannot prove the convergence of the naive update based on current techniques; if the greedy action is repeatedly chosen, 'infinitely often' visit in all (s,a) pairs cannot be guaranteed.

This reasoning leads us to formulate Surrogate α-Q-learning (see Algorithm 2). The Surrogate α-Q-learning updates two q-functions, q and q_α. The first, q, has the same update

Algorithm 2 Surrogate α-Q-Learning

```
**Initialize:** $\alpha \in [0,1],\ \pi_0,\ q_\alpha, q$, learning rate $\eta_t$.
  **for** $t = 0, 1, ...$ **do**
    $a_{\text{chosen}} \leftarrow \arg\max_a q(s_t, a)$
    $X_t \sim Bernoulli(1-\alpha)$
    $a_{\text{env}} = \begin{cases} a_{\text{chosen}},\ \text{if } X_t = 1 \\ a \sim \pi_0(\cdot \mid s),\ \text{if } X_t = 0 \end{cases}$
    $r_t, s_{t+1} \leftarrow ACT(a_{\text{env}})$
    **for** $\bar{a} \in \mathcal{A}$ **do**
      $y_t^{\bar{a}} = \begin{cases} r_t + \gamma v_\alpha(s_{t+1}),\ \bar{a} = a_{\text{chosen}} \\ X_t q(s_t, \bar{a}) + (1 - X_t)\left(r_t + \gamma v_\alpha(s_{t+1})\right), \text{o.w} \end{cases}$
      $q_\alpha(s_t, \bar{a}) \leftarrow (1-\eta)\, q_\alpha(s_t, \bar{a}) + \eta y_t^{\bar{a}}$
    **end for**
    $y_t \leftarrow r_t + \gamma(1-\alpha) v(s_{t+1}) + \gamma\alpha v^{\pi_0}(s_{t+1})$
    $q(s_t, a_{\text{env}}) \leftarrow (1-\eta_t) q(s_t, a_{\text{env}}) + \eta_t y_t$
  **end for**
  **return** $\pi \in \arg\max_a q_\alpha(\cdot, a)$
```

as in Expected α-Q-learning, and thus converges (w.p 1) to $q^{\pi^\alpha(\pi^*_\alpha, \pi_0)}$. The second, q_α, updates the chosen greedy action using equation (11), when the exploration policy is not played ($X_t = 1$). By bootstrapping on q, 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 Surrogate α-Q-learning to $(q^{\pi^\alpha(\pi^*_\alpha, \pi_0)}, q^*_\alpha)$ is established (see proof in Appendix L.2). Interestingly, and unlike other q-learning algorithms (e.g, Expected α-Q-learning, Q-learning, etc.), Surrogate α-Q-learning updates the entire action set given a single sample. For completness, we state the convergence result for both algorithms.

Theorem 10. *Consider the processes described in Alg. 1, 2. Assume $\{\eta_t\}_{t=0}^\infty$ satisfies $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$, $\sum_{t=0}^\infty \eta_t = \infty$, and $\sum_{t=0}^\infty \eta_t^2 < \infty$, where $\eta_t \equiv \eta_t(s_t = s, a_{\text{env},t} = a)$. Then, for both 1, 2 the sequence $\{q_n\}_{n=0}^\infty$ converges w.p. 1 to $q^{\pi^\alpha(\pi^*_\alpha, \pi_0)}$, and for 2, $\{q_{\alpha,n}\}_{n=0}^\infty$ converges w.p. 1 to q^*_α.*

5.2. Continuous Control

Building on the two approaches for solving Exploration Conscious criteria, we suggest two techniques to find an optimal Gaussian policy (6) using gradient based Deep RL (DRL) algorithms, and specifically, DDPG (Lillicrap et al., 2015). Nonetheless, the techniques are generalizable to other actor-critic, DRL algorithms (Schulman et al., 2017).

Assume we wish to find an optimal Gaussian policy by parameterizing its mean $\mu(\phi)$. Nachum et al. (2018)[Eq. 13] showed the gradient of the value w.r.t. ϕ is similar to Silver et al. (2014),

$$\nabla_\phi v^{\pi_{\mu,\sigma}} = \int_{\mathcal{S}} \partial_a q_\sigma^{\pi_{\pi_{\mu,\sigma}}}(s,a) \nabla_\phi \mu^\theta(s) d\rho^{\pi_{\mu,\sigma}}(s), \quad (12)$$

where $q_\sigma^\mu(s,a) = r_\sigma(s,a) + \gamma \int_{\mathcal{S}} p_\sigma(s' \mid s,a) v^{\pi_{\mu,\sigma}}(s') ds'$, is the q-function of the surrogate MDP. In light of previous section, we interpret q_σ^μ as the q-function of the surrogate MDP's $\mathcal{M}_\sigma$ (7). Furthermore, we have the following relation between the surrogate and expected q-functions, $q_\sigma^\mu(s,a) = \int_{a' \in \mathcal{A}} \mathcal{N}(a' \mid a, \sigma) q^{\pi_{\mu,\sigma}}(s,a') da'$, from which it is easy to verify that (see Appendix L.3),

$$\nabla_u q_\sigma^{\pi_{\mu,\sigma}}(s,b) = \int_{\mathcal{A}} \mathcal{N}(b \mid a, \sigma) \nabla_b q^{\pi_{\mu,\sigma}}(s,b) db. \quad (13)$$

Thus, we can update the actor in two inequivalent ways, by using gradients on the surrogate MDP's q-function (12), or by using gradients of the expected q-function (13).

The updates of the critic, q_σ^μ or $q^{\pi_{\mu,\sigma}}$, can be done using the same notion that led to the two forms of updates in (11)-(10). When using Gaussian noise, one performs the two stages defined in Section 5, where a_{chosen} is the output of the actor $\mu(s)$, and $a_{\text{env}} \sim \mathcal{N}(a_{\text{chosen}}, \sigma)$. Then, the sample $(s, a_{\text{chosen}}, a_{\text{env}}, r, s')$ is obtained by interacting with the environment. Based on the the fixed policy TD-error defined in (11), we define the following loss function, for learning q_σ^μ, q-function of the fixed policy μ over $\mathcal{M}_\sigma$,

$$\left(q_\sigma^\theta(s, a_{\text{chosen}}) - r - \gamma q_\sigma^{\theta-}(s', \mu^{\phi-}(s'))\right)^2.$$

On the other hand, we can define a loss function derived from the fixed-policy TD-error defined in (10), for learning $q^{\pi_{\mu,\sigma}}$, the q-function of the Gaussian policy with mean and variance μ, σ^2 over $\mathcal{M}$,

$$\left(q^\theta(s, a_{\text{env}}) - r - \gamma \int_{\mathcal{A}} \mathcal{N}(b \mid \mu^{\phi-}(s'), s') q^{\theta-}(s', b) db\right)^2.$$

6. Experiments

In this section, we test the theory and algorithms [1] suggested in this work. In all experiments we used $\gamma = 0.99$. The tested DRL algorithms in this section (See Appendix B) are simple variations of DDQN (Van Hasselt et al., 2016) and DDPG (Lillicrap et al., 2015), without any parameter tuning, and based on Section 5. For example, for the surrogate approach in both DDQN and DDPG we merely save $(s, a_{\text{chosen}}, r, s')$ instead of $(s, a_{\text{env}}, r, s')$ in the replay buffer (see Section 5 for definitions of $a_{\text{env}}, a_{\text{chosen}}$).

We observe a significant improved empirical performance, both in **training** and **evaluation** for both the surrogate and expected approaches relatively to the baseline performance. The improved training performance is predictable; the learned policy is optimal w.r.t. the noise which is being played. In large portion of the results, the exploration-conscious criteria leads to better performance in evaluation.

6.1. Exploration Consciousness with Prior Knowledge

We use an adaptation of the Cliff-Walking maze (Sutton et al., 1998) we term T-Cliff-Walking (see Appendix C).

[1] Implementation of the proposed algorithms can be found in https://github.com/shanlior/ExplorationConsciousRL.

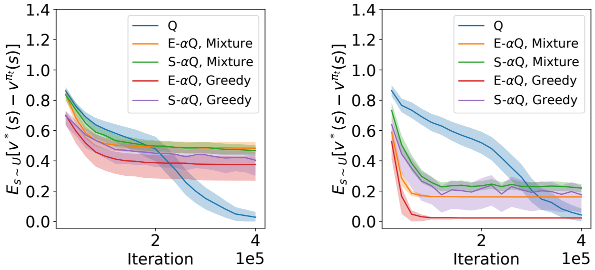

Figure 1. T-Cliff-Walking for the expected (E) and surrogate (S) approaches. (Left) α=0.3. (Right) $\alpha(s)$ from prior knowledge.

The agent starts at the bottom-left side of a maze, and needs to get to the bottom-right side goal state with value $+1$. If the agent falls off the cliff, the episode terminates with reward -1. When the agent visits any of the first three steps on top of the cliff, it gets a reward of $0.01 \cdot (1 - \gamma)$.

We tested Expected α-Q-learning, Surrogate α-Q-learning, and compared their performance to Q-learning in the presence of ϵ-greedy exploration. Figure 1 stresses the typical behaviour of the α-optimality criterion. It is easier to approximate $\pi^\alpha(\pi^*_\alpha, \pi_0)$ than the optimal policy. Further, by being exploration-consciousness, the value of the approximated policy improves faster using the α-optimal algorithms; it learns faster which regions to avoid. As Proposition 4 suggests, the value of the learned policy is biased w.r.t v^*. Next, as suggested by Proposition 3, acting greedily w.r.t. the approximated value attains better performance. Such improvement is not guaranteed while the value had not yet converged to v^*_α. However, the empirical results suggest that if the agent performs well over the mixture policy, it is worth using the greedy policy.

We show that it is possible to incorporate prior knowledge to decrease the bias caused by being Exploration-Conscious. The T-Cliff-Walking example demands high exploration, $\alpha = \epsilon = 0.3$, because of the bottleneck state between the two sides of the maze. The α-optimal policy in such case is to stay at the left part of the maze. We used the prior knowledge that $L(s)$ close to the barrier is high. The knowledge was injected through the choice of α, i.e., we chose a state-wise exploration scheme with $\alpha(s) = \epsilon(s) = 0.1$ in the passage and the two states around it, and $\alpha(s) = 0.3$ elsewhere, for all three algorithms. The results in Figure 1 suggests that using prior knowledge to set $\alpha(s)$, can increase the performance by reducing the bias. In contrast, such prior knowledge does not help the baseline q-learning.

6.2. Exploration Consciousness in Atari

We tested the α-optimal criterion in the more complex function approximation setting (see Appendix Alg. 3, 4). We used five Atari 2600 games (5) from the ALE (Bellemare et al., 2013). We chose games that resemble the Cliff Walking scenario, where the wrong choice of action can lead to a sudden termination of the episode. Thus, being unaware of the exploration strategy can lead to poor training results. We used the same deep neural network as in DQN (Mnih et al., 2015), using the openAI Baselines implementation (Dhariwal et al., 2017), without *any parameter tuning*, except for the update equations. We chose to use the Double-DQN variant of DQN (Van Hasselt et al., 2016) for simplicity and generality. Nonetheless, changing the optimality criterion is orthogonal to any of the suggested add-ons to DQN (Hessel et al., 2017). We used $\alpha = \epsilon = 0.01$ in the train phase, and $\epsilon = 0.001$ in the evaluation phase. For the *surrogate* version, we used a naive implementation based on equation (11).

Table 1. Train and Test rewards for the Atari 2600 environment, with 90% confidence interval

	Game	DDQN	Expected α-DDQN	Surrogate α-DDQN
Train	Breakout	350$\pm$4	356$\pm$6	**357**$\pm$4
	FishingDer	-45$\pm$9	-35$\pm$27	**-8**$\pm$8
	Frostbite	1191$\pm$171	794$\pm$158	**1908**$\pm$162
	Qbert	13221$\pm$565	13431$\pm$178	**14240**$\pm$225
	Riverraid	8602$\pm$205	8811$\pm$645	**11476**$\pm$79
Test	Breakout	**402**$\pm$14	390$\pm$5	392$\pm$5
	FishingDer	-37$\pm$15	-19$\pm$34	**-3**$\pm$19
	Frostbite	1720$\pm$191	1638$\pm$292	**2686**$\pm$278
	Qbert	15627$\pm$497	15780$\pm$206	**16082**$\pm$338
	Riverraid	9049$\pm$443	9491$\pm$802	**12846**$\pm$241

Table 1 shows that our method improves upon using the optimal criterion. That is, while bias exists, the algorithm still converges to a better policy. This result holds both on the exploratory training regime and the evaluation regime. Again, acting greedy w.r.t. the approximation of the α-optimal policy proved beneficial: The evaluation phase results surpasses the train phase results as shown in the table, and the training figures in Appendix (2). The evaluation is usually done with an $\epsilon = 0.001 > 0$. Proposition 3 put formal grounds for using smaller ϵ in the evaluation phase than in the training phase; improvement is assured. Being accurate is extremely important in most Atari games, so Exploration-Consciousness can also hurt the performance. Still, one can use prior knowledge to overcome this obstacle.

6.3. Exploration Consciousness in MuJoCo

We tested the Expected σ-DDPG (5) and Surrogate σ-DDPG (6) on continuous control tasks from the MuJoCo environment (Todorov et al., 2012). We used the OpenAI implementation of DDPG as the baseline, where we only changed the update equations to match our proposed algorithms. We used the default hyper-parameters, and independent Gaussian noise with $\sigma = 0.2$, for all tasks and

Table 2. Train and Test rewards for the MuJoCo environment.

	Game	DDPG	Expected σ-DDPG	Surrogate σ-DDPG
Train	Ant	809±47	**1013**±49	993±110
	HalfCheetah	2255±804	2634±828	**3848**±248
	Hopper	1864±139	1866±132	**2566**±155
	Humanoid	1281±142	1416±155	**1703**±272
	InPendulum	694±109	882±33	**998**±3
	Walker	1722±170	2144±145	**2587**±214
Test	Ant	1611±120	**1924**±126	1754±184
	HalfCheetah	2729±936	3147±986	**4579**±298
	Hopper	**3099**±113	3071±50	3037±78
	Humanoid	1688±223	1994±389	**2154**±408
	InPendulum	999±2	1000±0	**1000**±0
	Walker	3031±298	3315±147	**3501**±240

algorithms. The results in Table 2 were averaged over 10 different seeds. The performance of the σ-optimal variants superseded the baseline DDPG, for most of the training and test results. Interestingly, although improvement is not guaranteed (Proposition 7), the σ-optimal policy improved when using μ^ϕ deterministically, i.e., in the test phase. This suggests that improvement can be expected on certain scenarios, although that generally it is not guaranteed. We also found that the training process was faster using the σ-optimal algorithms, as can be seen in the learning curves in Appendix 3. Interestingly, again, the surrogate approach proved superior.

7. Relation to existing work

Lately, several works have tackled the exploration problem for deep RL. In some, like Bootstrapped-DQN (see appendix [D.1] in (Osband et al., 2016)), the authors still employ an ϵ-greedy mechanism on top of their methods. Moreover, methods like Distributional-DQN (Bellemare et al., 2017; Dabney et al., 2018) and the state-of-the-art Ape-X DQN (Horgan et al., 2018), still uses ϵ-greedy and Gaussian noise, for discrete and continuous actions, respectively. Hence, all the above works are applicable for the α-optimal criterion by using the simple techniques described in Section 5.

Existing on-policy methods produce variants of Exploration-Consciousness. In TRPO and A3C (Schulman et al., 2015; Mnih et al., 2016), the exploration is *implicitly injected* into the agent policy through entropy regularization, and the agent improves upon the value of the explorative policy. Simple derivation shows the α-greedy and the Gaussian approaches are both equivalent to regularizing the entropy to be higher than a certain value by setting α or σ appropriately.

Expected α-Q-learning highlights a relation to algorithms analysed in (John, 1994; Littman et al., 1997) and to Expected-Sarsa (ES) (Van Seijen et al., 2009). The focus of (John, 1994; Littman et al., 1997) is exploration-conscious q-based methods. In ES, when setting the 'estimation policy' (Van Seijen et al., 2009) to be $\pi = (1-\alpha_t)\pi_{\mathcal{G}} + \alpha_t\pi_0$, we get similar updating equations as in lines 1-1, and similarly to (John, 1994; Littman et al., 1997). However, in ES α_t decays to zero, and the *optimal policy* is obtained in the infinite time limit. In (Nachum et al., 2018), the authors offer a gradient based mechanism for updating the mean and variance of the actor. Here, we offer and analyze the approach of setting α_t and σ_t to a constant value. This would be of interest especially when a 'good' mechanism for decaying α_t and σ_t lacks; the decay mechanism is usually chosen by trial-and-error, and is not clear how it should be set.

Lastly, (4) and (6) can be understood as defining a 'surrogate problem', rather than finding an optimal policy. In this sense, it offers an alternative approach to biasing the problem by lowering the discount-factor, i.e., solve a surrogate MDP with $\bar{\gamma} < \gamma$ (Petrik & Scherrer, 2009; Jiang et al., 2015). Interestingly, the introduced bias when solving (4) is proportional to a *local property* of v^*, $L(s)$, that can be estimated using prior-knowledge on the MDP, where solving an MDP with $\bar{\gamma}$ introduces a bias proportional to a *non-local* term, which is harder to estimate. More importantly, the performance of an α-optimal policy π^*_α is assured to improve when tested on the original MDP $\mathcal{M}$ (Proposition 3), while the performance of an optimal policy in an MDP with $\bar{\gamma}$ might decline when tested on $\mathcal{M}$ with γ-discounting.

8. Summary

In this paper, we revisited the notion of an agent being conscious to an exploration process. To our view, this notion did not receive the proper attention, though it is implicitly and repeatedly used.

We started by formally defining *optimal policy* w.r.t. an exploration mechanism (4), (6). This expanded the view on exploration-conscious q-learning (John, 1994; Littman et al., 1997) to a more general one, and lead us to derive new algorithms, as well as re-interpreting existing ones (Van Seijen et al., 2009). We formulated the surrogate MDP notion, which helped us to establish that exploration-conscious criteria can be solved by Dynamic Programming, or, more generally, by an MDP solver. From the practical side, based on the theory, we tested DRL algorithms – by simply modifying existing ones, with no further hyper-parameter tuning – and empirically showed their superiority.

Although a bias - error sensitivity tradeoff was formulated, we did not prove (4), (6) are easier to solve than an MDP. We believe proving whether the claim is true is of interest. Furthermore, analyzing more exploration-conscious criteria, e.g., exploration-conscious w.r.t. Ornstein-Uhlenbeck noise, is of interest, as well as defining a unified framework for exploration-conscious criteria.

Acknowledgments

We would like to thank Chen Tessler, Nadav Merlis and Tom Zahavy for helpful discussions.

References

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

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

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.

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

Bertsekas, D. P. and Tsitsiklis, J. N. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pp. 560–564. IEEE, 1995.

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

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

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

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

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., Van Hasselt, H., and Silver, D. Distributed prioritized experience replay. *arXiv preprint arXiv:1803.00933*, 2018.

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

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

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

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

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. *arXiv preprint arXiv:1509.02971*, 2015.

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

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, 2015.

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

Nachum, O., Norouzi, M., Tucker, G., and Schuurmans, D. Smoothed action value functions for learning gaussian policies. *arXiv preprint arXiv:1803.02348*, 2018.

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

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pp. 4026–4034, 2016.

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

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

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *ICML*, 2014.

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

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

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*. MIT press, 1998.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pp. 5026–5033. IEEE, 2012.

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

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

A. Training graphs for the Atari and MuJoCo experiments

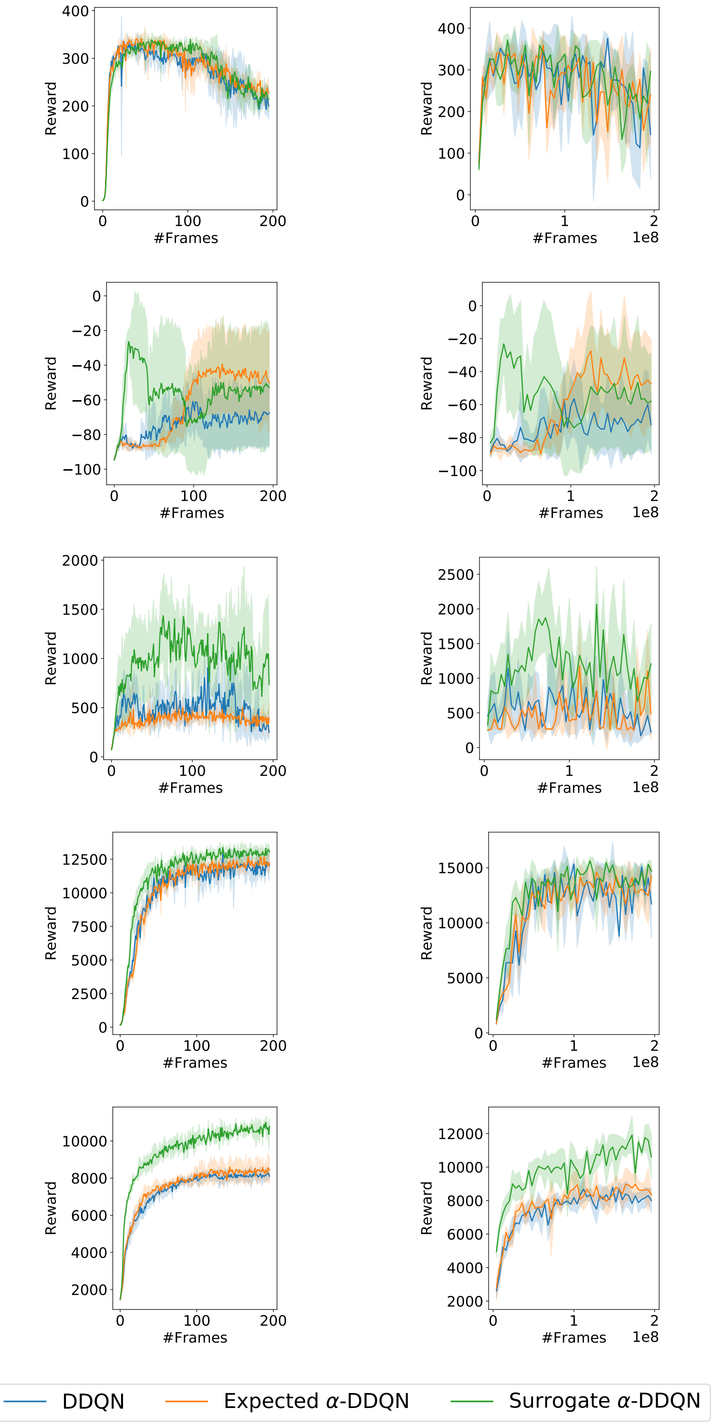

Figure 2. Simulation results for the Atari 2600 environment: From up to bottom: Breakout, Fishing Derby, Frostbite, Qbert and Riverraid. (Left) Training. (Right) Test.

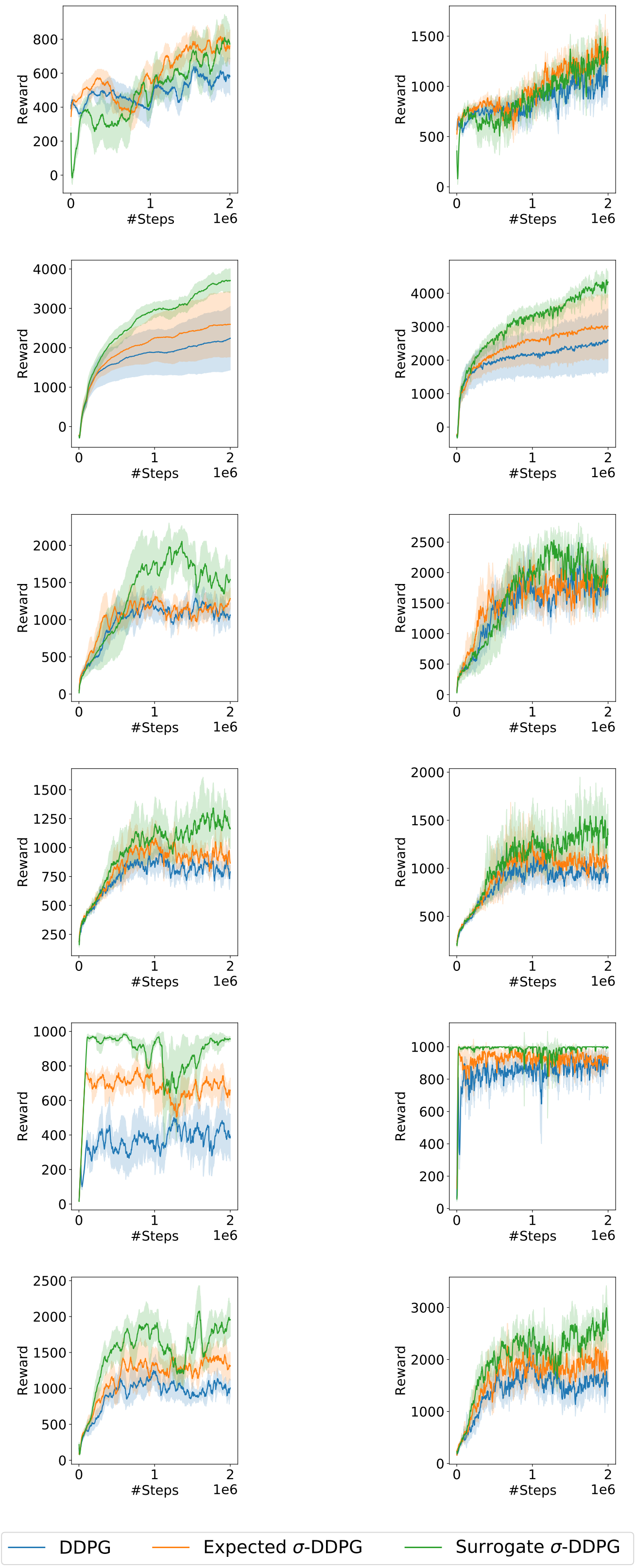

Figure 3. Simulation results for the MuJoCo environment: From up to bottom: Ant, HalfCheetah, Hopper, Humanoid, InvertedPendulum and Walker2d. (Left) Training. (Right) Test.

B. Deep RL Exploration Conscious algorithms

The algorithms in this section are the adjusted DDQN (Van Hasselt et al., 2016) and DDPG (Lillicrap et al., 2015) to solve the α-optimal and σ-optimal policies, respectively. For the surrogate approach the change is merely the gathered data; the action a_{chosen} is saved and not a_{env}. For the expected approach, the expectation is calculated by an explicit averaging Algorithm 3 or by simple sampling technique Algorithm 5.

Algorithm 3 Expected α-DDQN

```
Initialize: Network parameters $\theta$, $\theta^- \leftarrow \theta$
  Replay buffer $R$, Target network update time $N^-$
  for episode= $1, M$ do
    for $t = 1, T$ do do
      $a_{\text{chosen}} \leftarrow \arg\max_a q(s_t, a|\theta)$
      $X_t \sim Bernoulli(1-\alpha)$
      $a_{\text{env}} = \begin{cases} a_{\text{chosen}}, \text{ if } X_t = 1 \\ a \sim \pi_0(\cdot \mid s), \text{ if } X_t = 0 \end{cases}$
      $r_t, s_{t+1} \leftarrow ACT(a_{\text{env}})$
      Store $(s_t, a_{\text{env}}, r_t, s_{t+1})$ in $R$
      Sample $N$ tuples $(s_i, a^i_{\text{env}}, r_i, s'_i)$ from $R$
      $a_i \leftarrow \arg\max_a q(s'_i, a|\theta)$
      $v_i \leftarrow (1-\alpha) q(s'_i, a_i|\theta^-) + \alpha v^{\pi_0}(s'_i|\theta^-)$
      $y_i \leftarrow r_i + \gamma v_i$
      Minimize $L = \frac{1}{N}\sum_i \left(y_i - q(s_i, a^i_{\text{env}}|\theta\right)^2$
      Update $\theta^- \leftarrow \theta$ every $N^-$ steps
    end for
  end for
return $\pi \in \arg\max_a q(\cdot, a)$
```

Algorithm 4 Surrogate α-DDQN

```
Initialize: Network parameters $\theta$, $\theta^- \leftarrow \theta$
  Replay buffer $R$, Target network update time $N^-$
  for episode= $1, M$ do
    for $t = 1, T$ do do
      $a_{\text{chosen}} \leftarrow \arg\max_a q_\alpha(s_t, a|\theta)$
      $X_t \sim Bernoulli(1-\alpha)$
      $a_{\text{env}} = \begin{cases} a_{\text{chosen}}, \text{ if } X_t = 1 \\ a \sim \pi_0(\cdot \mid s), \text{ if } X_t = 0 \end{cases}$
      $r_t, s_{t+1} \leftarrow ACT(a_{\text{env}})$
      Store $(s_t, a_{\text{chosen}}, r_t, s_{t+1})$ in $R$
      Sample $N$ tuples $(s_i, a^i_{\text{chosen}}, r_i, s'_i)$ from $R$
      $a_i \leftarrow \arg\max_a q_\alpha(s'_i, a|\theta)$
      $y_i \leftarrow r_i + \gamma q_\alpha(s'_i, a_i|\theta^-)$
      Minimize $L = \frac{1}{N}\sum_i \left(y_i - q_\alpha(s_i, a^i_{\text{chosen}}|\theta\right)^2$
      Update $\theta^- \leftarrow \theta$ every $N^-$ steps
    end for
  end for
return $\pi \in \arg\max_a q_\alpha(\cdot, a)$
```

Algorithm 5 Expected σ-DDPG

```
Initialize: Critic and Actor networks $q(s, a|\theta)$, $\mu(s|\phi)$
  Target networks weights: $\theta^- \leftarrow \theta$ and $\phi^- \leftarrow \phi$
  Replay buffer $R$, Target network update time $N^-$
  for episode= $1, M$ do
    Initialize random markovian exploration process $\mathcal{N}$
    Receive initial observation state $s_1$
    for $t = 1, T$ do do
      $a_{\text{env}} \leftarrow \mu(s_t|\phi) + \mathcal{N}_t$
      $r_t, s_{t+1} \leftarrow ACT(a_t)$
      Store $(s_t, a_{\text{env}}, r_t, s_{t+1}, \mathcal{N}_t)$ in $R$
      Sample $N$ transitions $(s_i, a_i, r_i, s'_i, \mathcal{N}_i)$ from $R$
      Sample $D_1$ noise terms $n_j$ given $\mathcal{N}_i$
      $y_i \leftarrow r_i + \gamma \frac{1}{D_1}\sum_j q(s'_i, \mu(s'_i) + n_j|\phi^-)|\theta^-)$
      Critic Loss: $L = \frac{1}{N}\sum_i (y_i - q(s_i, a_i|\theta)^2$
      Sample $D_2$ noise terms $n_j$ given $\mathcal{N}_i$
      Approximate gradient policy gradient:
        $\nabla_\pi q^\pi(s_i) \approx \frac{1}{D_2}\sum_j \nabla_a q(s_i, a|\theta)\big|_{a=\mu(s_i)+n_j}$
      Update actor using policy gradient:
        $\nabla_\phi V \approx \frac{1}{N}\sum_i \nabla_\pi q^\pi(s_i) \nabla_\phi \mu(s_i|\phi)$
      Update target networks every $N^-$ steps
    end for
  end for
return $\mu(\cdot|\phi)$
```

Algorithm 6 Surrogate σ-DDPG

```
Initialize: Critic and Actor networks $q_\sigma(s, a|\theta)$, $\mu(s|\phi)$
  Target networks weights: $\theta^- \leftarrow \theta$ and $\phi^- \leftarrow \phi$
  Replay buffer $R$, Target network update time $N^-$
  for episode= $1, M$ do
    Initialize random markovian exploration process $\mathcal{N}$
    Receive initial observation state $s_1$
    for $t = 1, T$ do do
      $a_{\text{chosen}} \leftarrow \mu(s_t|\phi)$
      $a_{\text{env}} \leftarrow a_{\text{chosen}} + \mathcal{N}_t$
      $r_t, s_{t+1} \leftarrow ACT(a_{\text{env}})$
      Store $(s_t, a_{\text{chosen}}, r_t, s_{t+1})$ in $R$
      Sample $N$ transitions $(s_i, a_i, r_i, s'_i)$ from $R$
      $y_i \leftarrow r_i + \gamma q_\sigma(s'_i, \mu(s'_i|\phi^-))|\theta^-)$
      Critic Loss: $L = \frac{1}{N}\sum_i (y_i - q_\sigma(s_i, a_i|\theta)^2$
      Update actor using policy gradient:
      $\nabla_\phi V = \frac{1}{N}\sum_i \nabla_a q_\sigma(s_i, a|\theta)\big|_{a=\mu(s_i)} \nabla_\phi \mu(s_i|\phi)$
      Update target networks every $N^-$ steps
    end for
  end for
return $\mu(\cdot|\phi)$
```

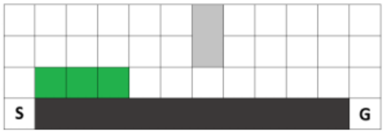

Figure 4. T-Cliff-Walking: The bright gray area is an impenetrable barrier. The cliff is colored in dark gray. The green states are with a small reward of $0.01 \cdot (1-\gamma)$.

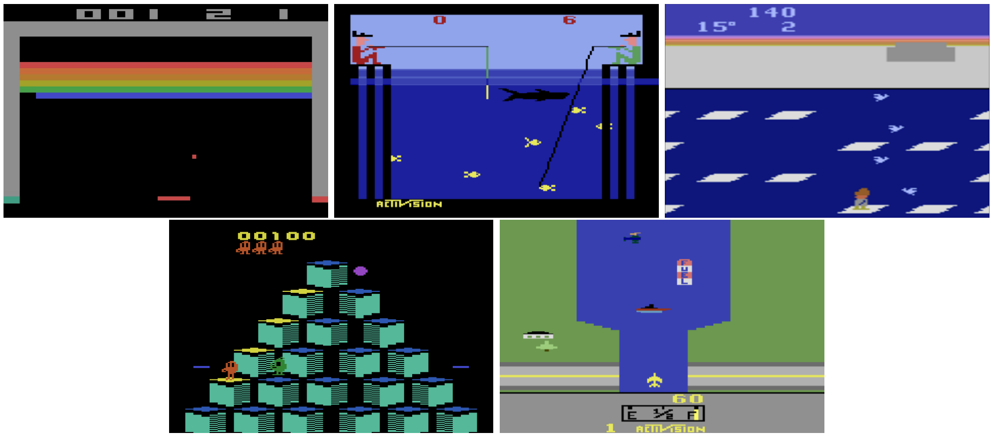

Figure 5. Atari games. From left to right: Breakout, Fishing Derby, Frostbite, Qbert, Riverraid.

C. Experimental details

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

C.1. Cliff Walking

We used the T-Cliff-Walking scenario in Figure 4: The size of the cliff is $(h, w) = (4, 12)$. We added small reward of $0.01 r_{max}$ (green states) in order to create some small bias between the optimal and the α-optimal policy. The maximal reward in this example is $r_{max} = 1 - \gamma$. We first checked to see that that $alpha = \epsilon = 0.1$ performed bad. Then, we raised the ϵ value. The bottleneck passage between to sides of the maze, creates a scenario where high exploration is needed. We performed 2,000 runs for each of the algorithms. Finally, the test error was evaluated with high precision using the fixed value iteration procedure.

D. Proof of Lemma 1

For any policy π the following equalities hold.

$$\begin{aligned}
v_\alpha^\pi &= (I - \gamma P_\alpha^\pi)^{-1} r_\alpha^\pi \\
&= (I - \gamma((1-\alpha)P^\pi + \alpha P^{\pi_0}))^{-1}((1-\alpha)r^\pi + \alpha r^{\pi_0}) \\
&= (I - \gamma P^{\pi^\alpha(\pi,\pi_0)})^{-1} r^{\pi^\alpha(\pi,\pi_0)} = v^{\pi^\alpha(\pi,\pi_0)}.
\end{aligned}$$

E. Proof of Proposition 2

Proof. Let $v \in \mathbb{R}^{|\mathcal{S}|}$ and consider the surrogate MDP, $\mathcal{M}_\alpha$. Its fixed policy Bellman operator (see (1)) is given by:

$$\begin{aligned}
T_\alpha^\pi v &= r_\alpha^\pi + \gamma P_\alpha^\pi v \\
&= (1-\alpha)(r^\pi + \gamma P^\pi v) + \alpha(r^{\pi_0} + \gamma P^{\pi_0} v) \\
&= (1-\alpha)T^\pi v + \alpha T^{\pi_0} v. \qquad (14)
\end{aligned}$$

The second relation is by plugging $P_\alpha^\pi, r_\alpha^\pi$ from (5), and rearranging. The fixed point of T_α^π is v_α^π, the value of π measured in $\mathcal{M}_\alpha$. Due to Lemma 1, $v_\alpha^\pi = v^{\pi^\alpha(\pi,\pi_0)}$.

The optimal Bellman operator of $\mathcal{M}_\alpha$ is (see (1)):

$$\begin{aligned} T_\alpha v &= \max_\pi T^\pi_\alpha v \\ &= \max_\pi (1-\alpha)T^\pi v + \alpha T^{\pi_0} v \\ &= (1-\alpha)\max_\pi T^\pi v + \alpha T^{\pi_0} v = (1-\alpha)T + \alpha T^{\pi_0}, \end{aligned} \tag{15}$$

where the second relation holds by (14). The fixed point of T_α is, by construction, v^*_α, the optimal value on $\mathcal{M}_\alpha$. Moreover, v^*_α is the optimal value of a policy on $\mathcal{M}_\alpha$. By Lemma 1, the policy that achieves the optimal value on $\mathcal{M}_\alpha$ achieves the α-optimal value, $\max_{\pi'} v^{\pi^\alpha(\pi',\pi_0)} = v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$. Thus, this policy is the α-optimal policy, π^*_α, and $v^*_\alpha = v^{\pi^*_\alpha} = v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$.

Since $\mathcal{M}_\alpha$ is an MDP, its optimal policy is in the greedy set w.r.t. v^*_α (see (2)). Thus,

$$\begin{aligned} \pi^*_\alpha &\in \{\pi : T^\pi_\alpha v^*_\alpha = T_\alpha v^*_\alpha\} \\ &= \{\pi : (1-\alpha)T^\pi v^*_\alpha + \alpha T^{\pi_0} v^*_\alpha = (1-\alpha)T v^*_\alpha + \alpha T^{\pi_0} v^*_\alpha\} \\ &= \{\pi : T^\pi v^*_\alpha = T v^*_\alpha\} = \mathcal{G}(v^*_\alpha). \end{aligned}$$

□

F. Proof of Theorem 3

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

Lemma 11. *Let v^π and $v^{\pi'}$ be the correspondsing values of the policies π and π'. Then,*

$$v^{\pi'} - v^\pi = (I - \gamma P^{\pi'})^{-1}(T^{\pi'} v^\pi - v^\pi) \tag{16}$$

Proof.

$$\begin{aligned} v^{\pi'} - v^\pi &= (I - \gamma P^{\pi'})^{-1} r^{\pi'} - v^\pi \\ &= (I - \gamma P^{\pi'})^{-1}(r^{\pi'} + \gamma P^{\pi'} v^\pi - v^\pi) \\ &= (I - \gamma P^{\pi'})^{-1}(T^{\pi'} v^\pi - v^\pi). \end{aligned}$$

□

The following Lemma has several instrances in previous literature:

Lemma 12. *Let π be any policy and $\pi_{1-\text{step}} \in \mathcal{G}(v^\pi)$. Then,*

$$v^\pi \leq v^{\pi^\alpha(\pi_{1-\text{step}},\pi)},$$

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

Proof.

$$v^{\pi^\alpha(\pi_{1-\text{step}},\pi)} - v^\pi = (I - \gamma P^{\pi^\alpha(\pi_{1-\text{step}},\pi)})^{-1}(T^{\pi^\alpha(\pi_{1-\text{step}},\pi)} v^\pi - v^\pi),$$

where the first relation holds due to Lemma 11. See that,

$$\begin{aligned} T^{\pi^\alpha(\pi_{1-\text{step}},\pi)} v^\pi - v^\pi &= (1-\alpha)T^{\pi_{1-\text{step}}} v^\pi + \alpha T^\pi v^\pi - v^\pi \\ &= (1-\alpha)T^{\pi_{1-\text{step}}} v^\pi + \alpha v^\pi - v^\pi \\ &= (1-\alpha)\left(T^{\pi_{1-\text{step}}} v^\pi - v^\pi\right) = (1-\alpha)\left(T v^\pi - v^\pi\right) \end{aligned}$$

Plugging it into (F) yields,

$$v^{\pi^\alpha(\pi_{1-\text{step}},\pi)} - v^\pi = (1-\alpha)(I - \gamma P^{\pi^\alpha(\pi_{1-\text{step}},\pi)})^{-1}(T v^\pi - v^\pi).$$

We have that $P^{\pi^\alpha(\pi_{1-\text{step}},\pi)})^{-1} \geq 0$ since it is a γ-discounted weighted sum of stochastic matrices. Furthermore,

$$v^\pi = T^\pi v^\pi \leq T v^\pi,$$

where the last inequality is strict at least in one component if $v^\pi \neq v^*$, i.e, if $\pi \neq \pi^*$. □

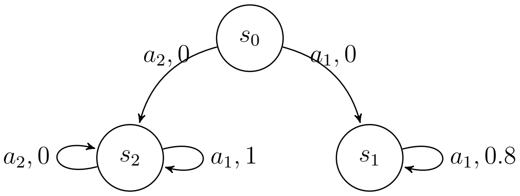

Figure 6. Counter exmple for an MDP with no monotonous improvement for the α-optimal criterion 3.

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

$$v^{\pi^\alpha(\pi^*_\alpha,\pi_0)} = \max_{\pi'} v^{\pi^\alpha(\pi',\pi_0)} \geq v^{\pi^\alpha(\pi_0,\pi_0)} = v^{\pi_0} \tag{17}$$

where the first relation is due to the definition of the α-optimal value (4), the second relation holds by definition and the third relation holds since

$$\pi_\alpha(\pi_0, \pi_0) = (1-\epsilon)\pi_0 + \epsilon\pi_0 = \pi_0.$$

As long as $\pi_0 \neq \pi^*$, the policy $\pi_{1-\text{step}} \in \mathcal{G}(v^{\pi_0})$ acheives strict improvement in (17). Meaning,

$$v^{\pi^\alpha(\pi_{1-\text{step}},\pi_0)} \geq v^{\pi_0}.$$

This means that the improvement in (17) is strict as long as $\pi_0 \neq \pi^*$.If π_0 is not optimal we have that

$$v^{\pi_0} \leq v^{\pi^\alpha(\pi_{1-\text{step}},\pi_0)} \leq v^{\pi^\alpha(\pi^*_\alpha,\pi_0)}.$$

The first relation is strict due to Lemma 12, and the second relation holds by the definition of the α-optimal policy.

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

$$v^{\pi^\beta(\pi^*_\alpha,\pi_0)} - v^*_\alpha = (I - \gamma P^{\pi^\beta(\pi^*_\alpha,\pi_0)})^{-1}(T^{\pi^\beta(\pi^*_\alpha,\pi_0)} v^*_\alpha - v^*_\alpha).$$

We have that,

$$\begin{aligned} T^{\pi^\beta(\pi^*_\alpha,\pi_0)} v^*_\alpha - v^*_\alpha &= T^{\pi^\beta(\pi^*_\alpha,\pi_0)} v^*_\alpha - T_\alpha v^*_\alpha \\ &= (1-\beta) T^{\pi^*_\alpha} v^*_\alpha + \beta T^{\pi_0} v^*_\alpha - (1-\alpha) T v^*_\alpha - \alpha T^{\pi_0} v^*_\alpha \\ &= (\alpha - \beta)\left(T v^*_\alpha - T^{\pi_0} v^*_\alpha\right), \end{aligned}$$

where in the last relation we used $T^{\pi^*_\alpha} v^*_\alpha = T v^*_\alpha$ (see Proposition 2). Plugging into (F) yields,

$$v^{\pi^\beta(\pi^*_\alpha,\pi_0)} - v^*_\alpha = (\alpha - \beta)(I - \gamma P^{\pi^\beta(\pi^*_\alpha,\pi_0)})^{-1}\left(T v^*_\alpha - T^{\pi_0} v^*_\alpha\right).$$

We have that $(I - \gamma P^{\pi^\beta(\pi^*_\alpha,\pi_0)})^{-1} \geq 0$ since it is a γ-discounted sum of stochastic matrices, and $T v^*_\alpha \geq T^{\pi_0} v^*_\alpha$ with equality if and only if π_0 is optimal; if and only if π_0 is optimal $v^*_\alpha = v^*$ due to the first part of this proof.

F.1. Counter example for monotonous improvement for the α-optimal criterion

In this section, we give a counter example that proves that the improvement in Proposition 3 is not monotonous w.r.t β. Let the MDP given in Figure 6 be a γ-discounted MDP for some $\gamma \in (0, 1)$. Let π_0 be a deterministic policy which always chooses action a_2. For $\alpha = 0.25$, It is easy to verify that $v^*_\alpha(s_1) = 0.8$ and $v^*_\alpha(s_2) = (1 - \alpha) = 0.75$. Now, $q^*_\alpha(s_0, a_1) = \gamma(1 - \alpha)v^*_\alpha(s_1) + \alpha v^*_\alpha(s_2) = 0.7875\gamma$, and $q^*_\alpha(s_0, a_2) = 0.75\gamma$. Thus, the α-optimal policy on s_0 is to choose a_1, and $v^*_\alpha(s_0) = 0.7875\gamma$.

Now, we consider acting according to the mixture policy $\pi^\beta(\pi^*_\alpha, \pi_0)$ for some $\beta < \alpha$. For the greedy policy, i.e. $\beta = 0$, we get that $v^{\pi^0(\pi^*_\alpha,\pi_0)} = v^{\pi^*_\alpha} = 0.8\gamma$. For $\beta = 0.1$, we get that $v^{\pi^{0.1}(\pi^*_\alpha,\pi_0)} = \gamma(0.9 \cdot 0.8 + 0.1 \cdot (0.9 \cdot 1)) = 0.81\gamma$. To conclude, as the lemma 3 suggests, we get improvement for both inspected β, i.e. $v^*_\alpha < v^{\pi^*_\alpha}$ and $v^*_\alpha < v^{\pi^{0.1}(\pi^*_\alpha,\pi_0)}$. However, the improvement does not increase monotonically as we decrease β, as $v^{\pi^*_\alpha} = 0.8\gamma < 0.81\gamma = v^{\pi^{0.1}(\pi^*_\alpha,\pi_0)}$.

G. Generalization of (Bertsekas & Tsitsiklis, 1995)[Proposition 6.1] for any policy class

In this section, we prove a generalization of (Bertsekas & Tsitsiklis, 1995)[Proposition 6.1] for any class of policies.

Proposition 13. *Let σ a set of fixed parameters of some distribution class. Assume $\hat{v}^*_\sigma$ is an approximate σ-optimal value s.t. $\|v^*_\sigma - \hat{v}^*_\sigma\| = \delta$ for some $\delta > 0$. Then,*

$$\left\|v^*_\sigma - v^{\hat{\pi}^*_\sigma}\right\| \leq \frac{\gamma\delta\|\pi^*_\sigma - \hat{\pi}^*_\sigma\|_{TV}}{1-\gamma}.$$

Proof.

$$\begin{aligned}
v^*_\sigma - v^{\hat{\pi}^*_\sigma}_\sigma &= T_\sigma v^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma} = T_\sigma v^*_\sigma - T_\sigma \hat{v}^*_\sigma + T_\sigma \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma} \\
&= T_\sigma v^*_\sigma - T_\sigma \hat{v}^*_\sigma + T^{\hat{\pi}^*_\sigma} v^*_\sigma - T^{\hat{\pi}^*_\sigma} v^*_\sigma + T^{\hat{\pi}^*_\sigma} \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} \hat{v}^*_\sigma + T_\sigma \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma} \\
&= \left(T_\sigma v^*_\sigma - T_\sigma \hat{v}^*_\sigma\right) + \left(T^{\hat{\pi}^*_\sigma} \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^*_\sigma\right) + \left(T^{\hat{\pi}^*_\sigma} v^*_\sigma - T^{\hat{\pi}^*_\sigma} \hat{v}^*_\sigma + T_\sigma \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma}\right) \\
&\overset{(a)}{\leq} \left(T^{\pi^*_\sigma} v^*_\sigma - T^{\pi^*_\sigma} \hat{v}^*_\sigma\right) + \left(T^{\hat{\pi}^*_\sigma} \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^*_\sigma\right) + \left(T^{\hat{\pi}^*_\sigma} v^*_\sigma - T^{\hat{\pi}^*_\sigma} \hat{v}^*_\sigma + T_\sigma \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma}\right) \\
&\overset{(b)}{=} \gamma P^{\pi^*_\sigma}\left(v^*_\sigma - \hat{v}^*_\sigma\right) - \gamma P^{\hat{\pi}^*_\sigma}\left(v^*_\sigma - \hat{v}^*_\sigma\right) + \left(T^{\hat{\pi}^*_\sigma} v^*_\sigma - T_\sigma \hat{v}^*_\sigma + T_\sigma \hat{v}^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma}\right) \\
&= \gamma\left(P^{\pi^*_\sigma} - P^{\hat{\pi}^*_\sigma}\right)\left(v^*_\sigma - \hat{v}^*_\sigma\right) + \left(T^{\hat{\pi}^*_\sigma} v^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma}\right)
\end{aligned}$$

Where (a) is due to the fact that for any v and π, $T^\pi_\sigma \leq T_\sigma v$, and (b) is due to the definition of the σ-greedy operator.

Taking the max-norm,

$$\begin{aligned}
\left|v^*_\sigma(s) - v^{\hat{\pi}^*_\sigma}(s)\right| &\leq \gamma\left|\left(\left(P^{\pi^*_\sigma} - P^{\hat{\pi}^*_\sigma}\right)\left(v^*_\sigma - \hat{v}^*_\sigma\right)\right)(s)\right| + \left|\left(T^{\hat{\pi}^*_\sigma} v^*_\sigma - T^{\hat{\pi}^*_\sigma} v^{\hat{\pi}^*_\sigma}\right)(s)\right| \\
&\leq \gamma\left|\sum_{s',a} p\left(s|s',a\right)\left(\pi^*_\sigma\left(a|s'\right) - \hat{\pi}^*_\sigma\left(a|s'\right)\right)\left(v^*_\sigma\left(s'\right) - \hat{v}^*_\sigma\left(s'\right)\right)\right| + \gamma\left\|v^*_\sigma - v^{\hat{\pi}^*_\sigma}\right\| = \\
&\leq \gamma \text{max}_{s'}\left|\sum_a \left(\pi^*_\sigma\left(a|s'\right) - \hat{\pi}^*_\sigma\left(a|s'\right)\right)\left(v^*_\sigma\left(s'\right) - \hat{v}^*_\sigma\left(s'\right)\right)\right| + \gamma\left\|v^*_\sigma - v^{\hat{\pi}^*_\sigma}\right\| \\
&\leq \gamma\left\|v^*_\sigma - \hat{v}^*_\sigma\right\|\left\|\pi^*_\sigma - \hat{\pi}^*_\sigma\right\|_{TV} + \gamma\left\|v^*_\sigma - v^{\hat{\pi}^*_\sigma}\right\|
\end{aligned}$$

Where the $\|\cdot\|_{TV}$ accounts for the maximal total-variation distance over all states. Finally,

$$\left\|v^*_\sigma - v^{\hat{\pi}^*_\sigma}\right\| \leq \frac{\gamma\delta\|\pi^*_\sigma - \hat{\pi}^*_\sigma\|_{TV}}{1-\gamma}.$$

□

Finally, this bound is a generalization of (Bertsekas & Tsitsiklis, 1995)[Proposition 6.1], for any class of distributions. Notice that the total variation distance is not bigger than 2, which is the case of two different deterministic policies. This leads back to the familiar bound.

H. Proof of Theorem 4: Bias-Error Sensitivity in the α-greedy case

In order to prove the theorem, we first prove the following two propositions 14,15. Then, we plug the results in the following triangle inequality:

$$\left\|v^* - v^{\pi^\alpha(\hat{\pi}^*_\alpha, \pi_0)}\right\| \leq \|v^* - v^*_\alpha\| + \left\|v^*_\alpha - v^{\pi^\alpha(\hat{\pi}^*_\alpha, \pi_0)}\right\|$$

Proposition 14. *Let $\forall s \in \mathcal{S},\ \alpha(s) \in [0,1]$, be a state-dependent function. Let π^*_α be the α-optimal policy, and $L(s)$ the MDP Lipschitz constant, both relatively to π_0. Define $B(\alpha) \triangleq \max_s \alpha(s)L(s)$. The following bounds hold,*

$$\left\|v^* - v^{\pi^*_\alpha}\right\| \leq \left\|v^* - v^{\pi^\alpha(\pi_g, \pi_0)}\right\| \leq \frac{B(\alpha)}{1-\gamma},$$

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

Proof. We have that for any $s \in \mathcal{S}$,

$$\begin{aligned} v^* - v^*_\alpha(s) &= (Tv^* - T_\alpha v^*)(s) + (T_\alpha v^* - T_\alpha v^*_\alpha)(s) \\ &\le \|Tv^* - T_\alpha v^*\| + \|T_\alpha v^* - T_\alpha v^*_\alpha\| \\ &\le \|Tv^* - T_\alpha v^*\| + \gamma \left\| v^* - v^*_\alpha \right\|, \end{aligned}$$

in the last relation we used the fact that T_α is a γ contraction in the max-norm. Moreover, we have that for any $s \in \mathcal{S}$,

$$\begin{aligned} Tv^*(s) - T_\alpha v^*(s) &= Tv^* - (1-\alpha(s))Tv^*(s) - \alpha(s)T^{\pi_0}v^*(s) \\ &= \alpha(s)\left(Tv^*(s) - T^{\pi_0}v^*(s)\right) \\ &= \alpha(s)\left(v^*(s) - T^{\pi_0}v^*(s)\right) = \alpha(s)L(s). \end{aligned} \tag{18}$$

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 $L(s) \ge 0$ since $v^*(s) - T^{\pi_0}v^*(s) \ge 0$. By taking the max-norm on (14), 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 (14), which is less tight. We have that,

$$\begin{aligned} |Tv^* - T^{\pi_0}v^*|(s) &= |T^{\pi^*}v^* - T_\alpha v^*|(s) \\ &\le \sum_a |\pi^*(a \mid s) - \pi_0(a \mid s)| \times \left| r(s,a) + \gamma \sum_{s'} P(s' \mid s,a)v^*(s') \right|, \end{aligned} \tag{19}$$

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

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

Thus, continuing from (14), we can further bound (19),

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

Thus,

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

where $\left\| \pi^* - \pi_0 \right\|_{TV}(s) = \sum_a |\pi^*(a \mid s) - \pi_0(a \mid s)|$, is the total variation of π^* and π_0 in state s.

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

For the MDP described in figure 7, it is easy to see that for the uniform π_0:

$$v^* - v^{\pi^*_\alpha} = \frac{1}{1-\gamma} - \frac{1-\alpha/2}{1-\gamma} = \frac{\alpha/2}{1-\gamma}$$

Next:

$$\frac{\alpha}{1-\gamma} \left\| v^*(s) - \sum_a \pi_0(a|s)q^*(s,a) \right\| = \frac{\alpha}{1-\gamma} \left\| \frac{1}{1-\gamma} - \frac{1/2}{1-\gamma} - \frac{\gamma/2}{1-\gamma} \right\| = \frac{\alpha/2}{1-\gamma}$$

□

Proposition 15. *Let $\alpha \in [0,1]$. Assume $\hat{v}^*_\alpha$ is an approximate α-optimal value s.t $\|v^*_\alpha - \hat{v}^*_\alpha\| = \delta$ for some $\delta \ge 0$. Let π_g be the greedy policy w.r.t. v, $\hat{\pi}^*_\alpha \in \mathcal{G}(\hat{v}^*_\alpha)$. Then*

$$\left\| v^*_\alpha - v^{\pi^\alpha(\hat{\pi}^*_\alpha, \pi_0)} \right\| \le \frac{2(1-\alpha)\gamma\delta}{1-\gamma}$$

Furthermore, there exists some $\delta_0 > 0$ such that if $\delta < \delta_0$, then $\hat{\pi}^*_\alpha = \pi^*_\alpha$, and this bound is tight.

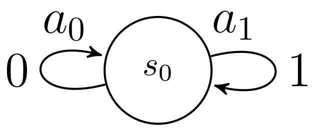

Figure 7. One State MDP that attains the bound in Proposition 14

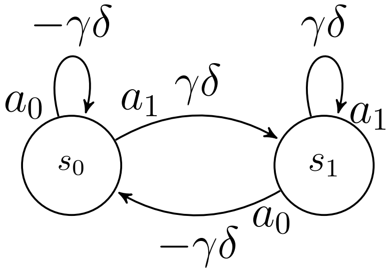

Figure 8. Two State MDP that attains the bound in Proposition 15 over a uniform π_0.

Proof. First, notice that for any two α-greedy policies, $\pi^\alpha(\pi_1,\pi_0), \pi^\alpha(\pi_2,\pi_0)$,

$$\begin{aligned}\|\pi^\alpha(\pi_1,\pi_0)-\pi^\alpha(\pi_2,\pi_0)\|_{TV} &= \|(1-\alpha)\pi_1+\alpha\pi_0-(1-\alpha)\pi_2-\alpha\pi_0\|_{TV}\\ &=(1-\alpha)\,\|\pi_1-\pi_2\|_{TV}\\ &\le 2(1-\alpha)\end{aligned}$$

Where the last transition is due to the fact that for the total- variation between distributions is always smaller than 2, which is the case of two different deterministic policies. Plugging in the result in Proposition 13, we get the required bound.

Finally, we prove that this bound is tight (see that different MDP then in (Bertsekas & Tsitsiklis, 1995) is used). Observe at the MDP described in Figure 8. The policy π^*_α is to always choose action a_1. Hence,

$$v^*_\alpha = \sum_{n=0}^{\infty}\gamma^n\left[\gamma\delta(1-\frac{\alpha}{2})-\gamma\delta\frac{\alpha}{2}\right]=\frac{\gamma\delta(1-\alpha)}{1-\gamma}$$

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 α-greedy policy with respect to $\hat{v}^*_\alpha$:

$$(1-\frac{\alpha}{2})(\gamma\delta+\gamma\hat{v}^*_\alpha(s_1))+\frac{\alpha}{2}(-\gamma\delta+\gamma\hat{v}^*_\alpha(s_0))=0=(1-\frac{\alpha}{2})(-\gamma\delta+\gamma\hat{v}^*_\alpha(s_0))+\frac{\alpha}{2}(\gamma\delta+\gamma\hat{v}^*_\alpha(s_1))$$

Hence,

$$v^{\pi^\alpha(\hat{\pi}^*_\alpha,\pi_0)}=\sum_{n=0}^{\infty}\gamma^n\left[-\gamma\delta(1-\frac{\alpha}{2})+\gamma\delta\frac{\alpha}{2}\right]=\frac{\gamma\delta(\alpha-1)}{1-\gamma}$$

Simple arithmetic show that this MDP attains the upper bound. □

I. Proof of Proposition 6

In this section, we will prove Proposition 6. First, we define the sufficient conditions for an MDP on which Proposition 6 is true:

Definition 3. *An MDP $\mathcal{M}=(\mathcal{S},\mathcal{A},P,R,\gamma)$ is a bounded continuous MDP if the following holds:*

1. *$\mathcal{A}$ is a metric space, s.t. $\mathcal{A}=\mathbb{R}^{|\mathcal{A}|}$*
2. *$\forall s\in\mathcal{S}$ and $a\in\mathcal{A}$, the state-wise reward function is positive, continuous, and bounded $r(s,a)$*
3. *$\forall s\in\mathcal{S}$, the state-wise reward function $r(s,a)$ is continuous in $a\in\mathcal{A}$*

4. *$\forall s, s' \in \mathcal{S}$, the transition probability density function $p(s'|s,a)$ is continuous in $a \in \mathcal{A}$.*

Furthermore, we assume $\mathcal{S}$ is finite. Yet, we believe it is possible to extend our result to continuous space as well. This we leave for future work.

Next, we state again the definition the optimal policy with respect to the Gaussian noise:

$$\mu^*_\sigma \in \arg\max_{\mu \in \tilde{\mathcal{A}}} \mathbb{E}^{\pi_{\mu,\sigma}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right], \tag{20}$$

Where the optimization is restricted to $\tilde{\mathcal{A}}$, a compact subset of $\mathcal{A}$.

We are now state again our main theorem regarding the σ-optimal optimization criterion:

Lemma 16. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be a bounded continuous MDP (see (3)). Let $\mathcal{N}(\mu, \sigma)$ be the Gaussian measure with mean $\mu \in \mathbb{R}^n$ and $\sigma \geq 0$ and let $\tilde{\mathcal{A}} \subset \mathcal{A}$ be a compact metric space. Then, the following claims hold:*

1. *$T^\mu_\sigma = \mathbb{E}^{\pi \sim \pi_{\mu,\sigma}} T^\pi$, with fixed point $v^\mu_\sigma = v^{\pi_{\mu,\sigma}}$.*
2. *$T_\sigma = \max_{\mu \in \tilde{\mathcal{A}}} \mathbb{E}^{\pi \sim \pi_{\mu,\sigma}} T^\pi$, with fixed point $v^*_\sigma = v^{\pi_{\mu^*_\sigma,\sigma}}$.*
3. *A σ-optimal policy is an optimal policy of $\mathcal{M}_\sigma$ and is Gaussian w.r.t. v^*_α, $\mu^*_\sigma \in \mathcal{N}_\sigma(v^*_\sigma) = \{\mu : T^{\pi_{\mu,\sigma}} v^*_\sigma = \max_\mu T^{\pi_{\mu,\sigma}} v^*_\sigma\}$.*

Proof. We define the surrogate MDP $\mathcal{M}_\sigma$ to have the following reward and dynamics,

$$\begin{aligned} r_\sigma(s,a) &= \int \mathcal{N}(a' \mid a, \sigma) r(s, a') da', \\ p_\sigma(s' \mid s, a) &= \int \mathcal{N}(a' \mid a, \sigma) p(s' \mid s, a') da'. \end{aligned}$$

Notice that

$$\begin{aligned} \sum_{s'} p_\sigma(s' \mid s, a) &= \int \mathcal{N}(a' \mid a, \sigma) \sum_{s'} p(s' \mid s, a') da' \\ &= \int \mathcal{N}(a' \mid a, \sigma) da' = 1. \end{aligned}$$

First, we show that the surrogate MDP $\mathcal{M}_\sigma$ is equivalent to a Gaussian policy on $\mathcal{M}$. More specifically, we show that the fixed policy bellman operator for a deterministic policy on $\mathcal{M}_\sigma$ is equivalent to the bellman operator of a Gaussian policy on $\mathcal{M}$. Then, we show similar relation for the bellman optimality operator.

Lemma 17. *The following claims hold:*

1. *The fixed-policy bellman operator on $\mathcal{M}_\sigma$, T^μ_σ and $T^{\pi_{\mu,\sigma}}$ are equivalent.*
2. *The bellman operator on $\mathcal{M}_\sigma$, T_σ and $\max_\mu T^{\pi_{\mu,\sigma}}$ are equivalent.*

Proof.

$$\begin{aligned} T^\mu_\sigma v &= r^\mu_\sigma + \gamma p^\mu_\sigma v \\ &= \mathbb{E}^{\pi \sim \pi_{\mu,\sigma}} r^\pi + \mathbb{E}^{\pi \sim \pi_{\mu,\sigma}} p^\pi v \\ &= \mathbb{E}^{\pi \sim \pi_{\mu,\sigma}} r^\pi + \gamma p^\pi v \\ &= \mathbb{E}^{\pi \sim \pi_{\mu,\sigma}} T^\pi \end{aligned}$$

The second relation holds directly from taking the maximum over both sides. □

By Lemma 17, the connection between operators is stated for both (1) and (2). By the definition of $\mathcal{M}_\sigma$, for any Gaussian policy with mean μ, $\pi_{\mu,\sigma}$, it holds that $v_\sigma^\mu = v_{\mu,\sigma}$.

Next, we prove the second relation. Again, we start by proving the following Lemma:

Lemma 18. *There exists a σ-optimal Gaussian policy*

Proof. The functions $r_\sigma(s,a)$ and $p_\sigma(s'|s,a)$ are defined as the expectation of $r(s,\cdot)$ and $p(s'|s,\cdot)$ on the Gaussian measure with mean a respectively. For every $s,s' \in \mathcal{S}$, define the integrand $g(a,\mu) = \phi(a|\mu,\sigma)f(s',s,a)$, where $f(s',s,a)$ represents $r(s,a)$ or $p(s'|s,a)$. The derivative of $\phi_\mu(a|\mu,\sigma)$ exists $\forall\mu \in \mathbb{R}^{|\mathcal{A}|}$. Thus, (a) $g_\mu(a,\mu)$ exists $\forall\mu \in \mathbb{R}^{|\mathcal{A}|}$. Next, For all $s,s' \in \mathcal{S}$, $r(s,a)$ and $p(s'|s,a)$ are continuous and bounded in a. $\forall\mu$, the Gaussian function is lebesgue-integrable function of a. Thus, (b) $\forall\mu, g(a,\mu)$ is a Lebesgue-integrable function of a. Now, there exist $c > 0$, such that, $|\phi_\mu| \leq c|a-\mu|\phi(a|\mu,\sigma)$. Furthermore, $f(s',s,a)$ is bounded. Hence, there exists $C > 0$, such that, $|g_\mu(a,\mu)| \leq C|a-\mu|\phi(a|\mu,\sigma) \triangleq h(a,\mu)$. Then, $\forall\mu$, we can take an open ball of radius r, $B_r(\mu)$. Define, $t(a) = \max_{x\in B_r(\mu)} h(a,x)$. t is integrable for every $a \in \mathcal{A}$ by construction. In other words, (c) there is an integrable function $t : A \to \mathbb{R}$ such that $|g_\mu(a,\mu)| \leq t(a)$ for all $\mu \in B_r(\mu)$.

Finally, From (a),(b) and (c), by the Dominated convergence theorem, Leibniz integral rule applies, which means that $r_\sigma(s,a)$ and $p_\sigma(s'|s,a)$ are differentiable in $a \in \mathcal{A}$, and thus continuous in $a \in \mathcal{A}$, for every $s,s' \in \mathcal{S}$.

Now, (1) let $\mathcal{M}_\sigma$ be the surrogate MDP, and assume the state space is discrete. (2) For all $s,s' \in \mathcal{S}$, $r_\sigma(s,a)$ and $p_\sigma(s'|s,a)$ are continuous in a. (3) By the definition of the optimality criterion, we consider only actions $a \in \mathcal{A}$. Hence, the action space of $\mathcal{M}_\sigma$ is compact.

Then, by theorem [6.2.10] in (Puterman, 1994), there exist an optimal deterministic policy for the surrogate MDP, $\mathcal{M}_\sigma$.

By the definition of the $\mathcal{M}_\sigma$ and Lemma 17, a deterministic policy μ in $\mathcal{M}_\sigma$ is equivalent to a Gaussian policy $\pi_{\mu,\sigma}$ on $\mathcal{M}$. Denote the optimal deterministic policy on the surrogate MDP as μ_σ^*. Thus, the policy $\pi_{\mu_\sigma^*,\sigma}$ is an σ-optimal Gaussian policy on $\mathcal{M}$.

□

Finally, we show that solving the surrogate MDP is equivalent to solving (20) T_σ is the greedy bellman operator on the surrogate MDP. Therefore, it is a γ-contraction. Thus, (a) by the Banach fixed point theorem and Theorem [6.2.2] in (Puterman, 1994), v_σ^* is the unique solution to the optimality equation, $T_\sigma v_\sigma^* = v_\sigma^*$. (b) By Lemma 18, there exists a deterministic optimal policy. Combining (a) and (b), we get that the greedy policy w.r.t. v_σ^*, μ_σ^*, is an optimal policy in the surrogate MDP. By transforming back to the original MDP we get that $\pi_\sigma^* = \pi_{\mu_\sigma^*,\sigma}$:

$$\begin{aligned}
\mu_\sigma^* &\in \{\mu : T_\sigma^\mu v_\sigma^* = T_\sigma v_\sigma^*\} \\
&= \{\mu : \mathbb{E}^{\pi\sim\pi_{\mu,\sigma}} T^\pi v_\sigma^* = \max_\mu \mathbb{E}^{\pi\sim\pi_{\mu,\sigma}} T^\pi v_\sigma^*\} \\
&= \mathcal{N}_\sigma(v_\sigma^*).
\end{aligned}$$

□

I.1. MDP with bounded action space

In this section we explain how to apply the σ-optimal criterion to an MDP with bounded action space. Let $\mathcal{M}$ be a bounded continuous MDP with a compact action-space $\mathcal{A}$. Proposition 6 demands the action space to be defined on the support of the Gaussian measure. Thus, we need to formalize how the Gaussian noise which is defined over $\mathbb{R}^{|\mathcal{A}|}$ operates on the bounded action set $\mathcal{A}$. Intuitively, we choose to project any action chosen outside the action set $a \notin \mathcal{A}$ onto the action set boundary. Formally, the noise operates on the extended MDP, $\mathcal{M}_{ext}$, as defined here.

Definition 4. *For a bounded continuous MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, we define the extended MDP, $\mathcal{M}_{ext}$, with action space $\mathcal{A}_{ext} = \mathbb{R}^{|\mathcal{A}|}$, such that:*

1. $R_{ext}(s,a) = R(s, \mathcal{P}_\mathcal{A}(a))$, *for all* $s \in \mathcal{S}$.
2. $P_{ext}(s,a) = P(s, \mathcal{P}_\mathcal{A}(a))$, *for all* $s,s' \in \mathcal{S}$

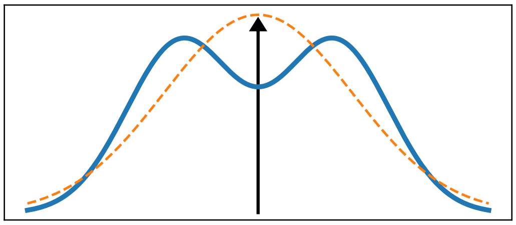

Figure 9. Illustration of a typical case where there is no improvement: (Blue) The state-action value function as a function of the action taken. (Orange) The σ-optimal policy is with $\mu_\sigma^* = 0$ due to the smoothing effect of the Gaussian policy. (Black) A deterministic policy around the μ_σ^*. It can be easily seen that decreasing the noise degrades the performance of the agent.

Where, $\mathcal{P}_{\mathcal{A}}(a)$ is the orthogonal projection of the action a onto the set $\mathcal{A}$

The MDP $\mathcal{M}_{ext}$ is a bounded continuous MDP, with action space $\mathbb{R}^{|\mathcal{A}|}$. Therefore, by 6, it is possible to find the optimal policy w.r.t. the σ-optimal criterion, over any bounded action space. Finally, most naturally, one can apply the criterion to the original action space $\mathcal{A}$.

J. No Improvement in Continuous Control

We give here the proof, the improvement is not always guaranteed in the continuous case.

Proposition 19. *Let $0 \leq \sigma' < \sigma$ and let μ^* be the σ-optimal policy. There exists an MDP such that $v^{\pi_{\mu^*,\sigma}} > v^{\pi_{\mu^*,\sigma'}}$. Decreasing the stochasticity can hurt the performance of the agent, and improvement is not guaranteed.*

Proof. Let $\mathcal{M}$ be a one-state MDP, with the following reward: $r(u) = \frac{1}{2}\frac{1}{\sqrt{\pi}}e^{-(u-1)^2} + \frac{1}{2}\frac{1}{\sqrt{\pi}}e^{-(u+1)^2}$. The expected reward under a Gaussian policy with μ and $\sigma = 1$ is: $r^\pi = \frac{1}{2}\frac{1}{\sqrt{3\pi}}e^{-(\mu-1)^2/3} + \frac{1}{2}\frac{1}{\sqrt{3\pi}}e^{-(\mu+1)^2/3}$. It is easy to calculate that the maximum of r^π is attained when $\mu = 0$ and its value lower bounded by 0.23. Hence, the σ-optimal policy with $\sigma = 1$ is $\pi(u|s) = \mathcal{N}(0,1)$. However, acting greedily w.r.t the mean of the σ-optimal, i.e., acting always with $u = 0$ can be upper bounded by 0.21. Thus, $r^{\pi_\sigma^*} > r^{\pi^{\mu_\sigma},0}$ □

An illustration of such a case is given in figure J.

While in the general case there is no improvement, it is easy to verify that a sufficient condition for improvement is that the state-wise variance of the $q^{\pi_\sigma^*}$ w.r.t. every smaller noise level, $\tilde{\sigma} < \sigma$, is less than the noise level itself:

$$\frac{\mathbb{E}_{a\sim\pi_{\mu_\sigma^*,\tilde{\sigma}}(\cdot|s)}\left[\left(a-\mu_\sigma^*\left(s\right)\right)^2 q^{\pi_\sigma^*}\left(s,a\right)\right]}{\mathbb{E}_{a\sim\pi_{\mu_\sigma^*,\tilde{\sigma}}(\cdot|s)} q^{\pi_\sigma^*}\left(s,a\right)} \leq \tilde{\sigma}^2.$$

K. Proof of Theorem 8: Bias-Error Sensitivity in the Gaussian case

In this section we prove a bias-error sensitivity result for the Gaussian noise case, similarly to 4. Theorem 8 exhibits a Bias-Sensitivity trade-off w.r.t. the noise parameter σ. When σ grows, the bias increases in $\|\sigma\|_1$, but the sensitivity term decreases. In the limit where σ goes to infinity, the approximation error tend to zero. In the other limit, where the noise reduces to zero, we return to the case of a greedy optimal policy. Indeed, as the bound shows, we get an unbiased solution, and the sensitivity term reduces to the classical bound of Bertsekas & Tsitsiklis (1995). Unsurprisingly, we get a better sensitivity bound only when there is a sufficient overlap between the two policies.

In order to prove the theorem, we will first prove two propositions: A bias proposition 20 and a sensitivity proposition 21. Then, we plug the results in the following triangle inequality:

$$\left\|v^* - v^{\hat{\mu},\sigma}\right\| \leq \left\|v^* - v_\sigma^*\right\| + \left\|v_\sigma^* - v^{\hat{\mu},\sigma}\right\|$$

First, we derive the bias proposition,

Proposition 20. *Let $\sigma \geq 0$ and let π^*_σ be the σ-optimal policy. Assume an MDP $\mathcal{M}$ is Lipschitz, i.e., there exists $L_r \geq 0$ and $L_p \geq 0$, such that, $\forall s, s' \in \mathcal{S}$ and $\forall a_1, a_2 \in \mathcal{A}$, $|r(s,a_1) - r(s,a_1)| < L_r \left\|a_1 - a_2\right\|_1$ and $|p(s'|s,a_1) - p(s'|s,a_1)| < L_p \left\|a_1 - a_2\right\|_1$. Then, the following holds,*

$$\|v^* - v^*_\sigma\| \leq \sqrt{\frac{2}{\pi}} \frac{(1-\gamma)\, L_r + \gamma L_p R_{max}}{(1-\gamma)^2} \sigma$$

Proof.

$$\begin{aligned}
\|v^* - v^*_\sigma\| &= \|v^* - T_\sigma v^*_\sigma\| \\
&\leq \|v^* - T_\sigma v^*\| + \|T_\sigma v^* - T_\sigma v^*_\sigma\| \\
&\leq \|v^* - T_\sigma v^*\| + \gamma \left\|v^* - v^*_\sigma\right\|
\end{aligned}$$

Where the inequality is due to the fact that T_σ is a γ-contraction. Simple algebra gives $\|v^* - v^*_\sigma\| \leq \frac{\|v^* - T_\sigma v^*\|}{1-\gamma}$

Next, we bound the nominator:

$$\begin{aligned}
v^*(s) - (T_\sigma v^*)(s) &= (T^* v^*)(s) - (T_\sigma v^*)(s) \\
&= \max_a r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') - \max_\mu \int \mathcal{N}(a|\mu,\sigma) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') \right] da \\
&\leq r(s,a^*) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a^*) v^*(s') - \int \mathcal{N}(a|a^*,\sigma) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') \right] da \\
&= \int \mathcal{N}(a|a^*,\sigma) \left[(r(s,a^*) - r(s,a)) + \gamma \sum_{s' \in \mathcal{S}} \left(p(s'|s,a^*) - p(s'|s,a)\right) v^*(s') \right] da \\
&\leq \int \mathcal{N}(a|a^*,\sigma) \left[(r(s,a^*) - r(s,a)) + \gamma \sum_{s' \in \mathcal{S}} \left|p(s'|s,a^*) - p(s'|s,a)\right| v^*(s') \right] da \\
&\leq \int \mathcal{N}(a|a^*,\sigma) \left[(r(s,a^*) - r(s,a)) + \frac{\gamma R_{max}}{1-\gamma} \sum_{s' \in \mathcal{S}} \left|p(s'|s,a^*) - p(s'|s,a)\right| \right] da \\
&\leq \int \mathcal{N}(a|a^*,\sigma) \left[L_r \left\|a^* - a\right\|_1 + \gamma \left\|p(\cdot \mid s,a^*) - p(\cdot \mid s,a)\right\|_{TV} \frac{R_{max}}{1-\gamma} \right] da \\
&\leq \int \mathcal{N}(a|a^*,\sigma) \left[L_r \left\|a^* - a\right\|_1 + \gamma L_p \left\|a^* - a\right\|_1 \frac{R_{max}}{1-\gamma} \right] da \\
&= \left(L_r + \gamma L_p \frac{R_{max}}{1-\gamma} \right) \int \mathcal{N}(a|a^*,\sigma) \left\|a^* - a\right\|_1 da \\
&= \left(L_r + \gamma L_p \frac{R_{max}}{1-\gamma} \right) \sqrt{\frac{2}{\pi}} \left\|\sigma\right\|_1
\end{aligned}$$

Where the first transition is due to $a^* \in \arg\max r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s')$, and the last is due to the absolute first moment of the Gaussian distribution.

We get,

$$\|v^* - T_\sigma v^*\| \leq \sqrt{\frac{2}{\pi}} \left(L_r + \gamma L_p \frac{R_{max}}{1-\gamma} \right) \left\|\sigma\right\|_1$$

Finally, combining the two results gives:

$$\|v^* - v^*_\sigma\| \leq \sqrt{\frac{2}{\pi}} \frac{(1-\gamma) L_r + \gamma L_p R_{max}}{(1-\gamma)^2} \left\|\sigma\right\|_1$$

□

Finally, we prove the following sensitivity proposition using:

Proposition 21. *Let $\sigma \in \mathbb{R}_+^{|\mathcal{A}|}$. Assume $\hat{v}_\sigma^*$ is an approximate σ-optimal value s.t. $\|v_\sigma^* - \hat{v}_\sigma^*\| = \delta$ for some $\delta > 0$. Let $\mu_\sigma^*, \hat{\mu}_\sigma^* \in \mathbb{R}^{\mathcal{A}}$ be the greedy mean policy w.r.t. v_σ^* and $\hat{v}_\sigma^*$ respectively. Then,*

$$\|v_\sigma^* - v^{\pi_{\hat{\mu}_\sigma^*,\sigma}}\| \leq \frac{1}{2}\frac{\gamma\delta\min\{\|\mu_\sigma^* - \hat{\mu}_\sigma^*\|_{\sigma^{-2}}, 4\}}{1-\gamma},$$

where $\|\cdot\|_{\sigma^{-2}}$ is the σ^{-2}-weighted euclidean norm.

Proof. First, notice that the total variation distance is not bigger than 2, which is the case of two different deterministic policies, as seen in (Bertsekas & Tsitsiklis, 1995)[Proposition 6.1]. Next, the Kullback-Leibler divergence between two Gaussian distributions with the same variance σ is $\frac{1}{2}\|\mu_\sigma^* - \hat{\mu}_\sigma^*\|_{\sigma^{-2}}^2$, where $\|\cdot\|_{\sigma^{-2}}$ is the σ^{-2}-weighted euclidean norm. Finally, by using Pinsker's inequality to bound the total variation distance, and plugging in the closed form of the Kullback-Leibler divergence, one gets the required result. □

L. Supplementary material for Section 5

In this section we give the proofs for the algorithms proposed in Section 5.1.

The proof of Lemma 9 is given as follows:

Proof. By using the definition of T_α^{Eq}, and due to $v_\alpha^* = \max_a q_\alpha^*(\cdot, a)$, we have that,

$$\begin{aligned}
q_\alpha^*(s,a) &= T_\alpha^{Eq} q_\alpha^*(s,a) \\
&= r_\alpha(s,a) + \gamma\sum_{s'} P_\alpha(s' \mid s,a)\max_{a'} q_\alpha^*(s',a') \\
&= (1-\alpha)\left(r(s,a) + \gamma\sum_{s'} P(s' \mid s,a) v_\alpha^*(s')\right) + \alpha\sum_a \pi(a' \mid s)\left(r(s,a') + \gamma\sum_{s'} P(s' \mid s,a') v_\alpha^*(s')\right) \\
&= (1-\alpha) q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(s,a) + \alpha\sum_a \pi_0(a' \mid s) q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(s,a'),
\end{aligned}$$

where in the last relation we used (8). □

We now prove the following lemma:

Lemma 22. *The operator T_α^{Eq} is a γ-contraction, and its fixed point is $q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}$*

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

$$v_\alpha^* = (1-\alpha)\max_a q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(\cdot,a) + \alpha\Pi_0 q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}. \tag{21}$$

Consider the definition of $q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}$ (8). We have that,

$$\begin{aligned}
q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(s,a) &= r(s,a) + \gamma\sum_{s'} P(s' \mid s,a) v_\alpha^*(s') \\
&= r(s,a) + \gamma(1-\alpha)\sum_{s'} P(s' \mid s,a)\max_{a'} q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(s',a') \\
&\quad + \gamma\alpha\sum_{s',a'} P(s' \mid s,a)\pi_0(a' \mid s') q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(s',a') \\
&= T_\alpha^{Eq} q^{\pi^\alpha(\pi_\alpha^*,\pi_0)}(s,a),
\end{aligned}$$

where the first relation holds by plugging (21) and the third relation holds by identifying the operator T_α^{Eq}. □

L.1. Convergence of Expected α-Q-Learning

Now, we move on to prove the convergence of Expected α-Q-Learning:

Theorem 23. *Consider the process described in Algorithm 1. Assume the sequence $\{\eta_t\}_{t=0}^{\infty}$ satisfies $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$, $\sum_{t=0}^{\infty} \eta_t(s_t = s, a_{\mathrm{env},t} = a) = \infty$, and $\sum_{t=0}^{\infty} \eta_t^2(s_t = s, a_{\mathrm{env},t} = a) < \infty$. Then, the sequence $\{q_n\}_{n=0}^{\infty}$ converges w.p 1 to $q^{\pi^{\alpha}(\pi_{\alpha}^*, \pi_0)}$.*

Proof. The updating equations of Algorithm 1 can be written as

$$q_{n+1}(s, a^{env}) = (1 - \eta_t) q_n(s, a_{\mathrm{env}}) + \eta_t (T_{\alpha}^{Eq} q_n(s, a_{\mathrm{env}}) - w_t),$$

where

$$w_t = r_t + \gamma(1-\alpha) v(s_{t+1}) + \gamma\alpha v^{\pi_0}(s_{t+1}) - T_{\alpha}^{Eq} q_t(s, a_{\mathrm{env}}),$$

and

$$v(s_{t+1}) = \max_{a'} q(s_{t+1}, a'), \qquad v^{\pi_0}(s_{t+1}) = \sum_{a'} \pi_0(a' \mid s_{t+1}) q(s_{t+1}, a').$$

We let $\mathcal{F}_t = \{\mathcal{H}_{t-1}, s_t, a_{\mathrm{env}}, X_t, a_{\mathrm{chosen}}, r_t\}$, 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 π_0, and the acted action.

We have that,

$$\begin{aligned} &\mathbb{E}\left[r_t + \gamma(1-\alpha) \max_a q(s_{t+1}, a_{\mathrm{env}})(s_{t+1}) \mid \mathcal{F}_t\right] = \\ &= r(s_t, a_{\mathrm{env}}) + \gamma(1-\alpha) \sum_{s'} P(s' \mid s, a_{\mathrm{env}}) \max_{a'} q(s', a') + \gamma\alpha \sum_{s', a'} P(s' \mid s_t, a_{\mathrm{env}}) \pi_0(a' \mid s') q(s', a'), \end{aligned}$$

and $\mathbb{E}[w_t \mid \mathcal{F}_t] = 0$. It is also easy to see that $\mathbb{E}\left[w_t^2 \mid \mathcal{F}_t\right] \le A + B||Q||_{\infty}^2$.

Thus, according to (Bertsekas & Tsitsiklis, 1995)[Proposition 4.4] the process converges to the fixed point contraction operator T_{α}^{Eq}, $q^{\pi^{\alpha}(\pi_{\alpha}^*, \pi_0)}$ (see Lemma 22). □

L.2. Convergence of Surrogate α-Q-Learning

In this section, we prove the convergence of Surrogate α-Q-Learning:

Theorem 24. *Consider the process described in Algorithm 2. Assume the sequence $\{\eta_t\}_{t=0}^{\infty}$ satisfies $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$, $\sum_{t=0}^{\infty} \eta_t(s_t = s, a_{\mathrm{env},t} = a) = \infty$, and $\sum_{t=0}^{\infty} \eta_t^2(s_t = s, a_{\mathrm{env},t} = a) < \infty$. Then, the sequences $\{q_n\}_{n=0}^{\infty}$ and $\{q_{\alpha,n}\}_{n=0}^{\infty}$ converges w.p 1 to $q^{\pi^{\alpha}(\pi_{\alpha}^*, \pi_0)}$ and q_{α}^*, respectively.*

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

Lemma 25. *Consider a stochastic process $(\alpha_t, \Delta_t, \Delta_t, f_t)$, $t \ge 0$, where α_t, Δ_t, $f_t : X \to \mathbb{R}$ satisfy the equations*

$$\begin{aligned} &\Delta_{t+1}(x) = (1 - \alpha_t(x)) \Delta_t(x) + \alpha_t(x) f_t(x), \\ &x \in X, \quad t = 0, 1, 2, .. \end{aligned} \tag{22}$$

Let $\mathcal{F}_t$ be a sequence of increasing σ-fields such that α_0 and Δ_0 are $\mathcal{F}_0$-measurable, $t = 1, 2, ...$ Assume that the following hold:

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

Then, Δ_t converges to zero with probability 1.

Observe that q_t has updating rule as in Expected α-Q-Learning (see Algorithm 1), and is independent of q_α. Due to the assumptions that $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$

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

we get that the sequence $\{q_t\}_{t=0}^{\infty}$ converges to $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$ w.p 1.

We now manipulate the updating of q in Algorithm 2 to have the form of (22). Define the following difference

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

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

By decreasing $q^*_\alpha(s,a)$ from both sides of the updating equations of q in Algorithm 2, we obtain for any $a \in \mathcal{A}$,

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

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

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

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

$$\begin{aligned} f_t(s_t,\bar{a}) =& X_t q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s_t,\bar{a}) + (1-X_t)(r_t + \gamma v_{\alpha,t}(s_{t+1})) \\ &+ X_t(q_t(s_t,\bar{a}) - q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s_t,\bar{a})) - q^*_\alpha(s_t,\bar{a}). \end{aligned}$$

We now show that for all action entries $\bar{a} \in \mathcal{A}$, $\mathbb{E}\left[f_t(s_t,\bar{a}) \mid \mathcal{F}_t\right] || \leq \kappa ||\Delta_t(s_t,\bar{a})|| + c_t$, and c_t converges to zero w.p. 1.

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

$$\begin{aligned} \mathbb{E}\left[f_t(s_t,\bar{a}) \mid \mathcal{F}_t\right] &= (1-\alpha)(r(s_t,\bar{a}) + \gamma \sum_{s'} P(s' \mid s_t,\bar{a}) \max_{a'} q_{\alpha,t}(s',a')) \\ &\quad + \alpha(r^{\pi_0}(s_t) + \gamma \sum_{s'} P^{\pi_0}(s' \mid s_t) \max_{a'} q_{\alpha,t}(s',a')) - q^*_\alpha(s,a) \\ &= T_\alpha q_{\alpha,t}(s_{t+1},a')) - q^*_\alpha(s,a). \end{aligned}$$

Thus, for this case,

$$\begin{aligned} ||\mathbb{E}\left[f_t(s_t,\bar{a}) \mid \mathcal{F}_t\right]|| &= ||T_\alpha q_{\alpha,t}(s_{t+1},a')) - q^*_\alpha(s,a)|| \\ &= ||T_\alpha q_{\alpha,t}(s_{t+1},a')) - q^*_\alpha(s,a)|| \\ &\leq \gamma ||q_{\alpha,t}(s_{t+1},a')) - q^*_\alpha(s,a)||, \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}\left[f_t(s_t,\bar{a}) \mid \mathcal{F}_t\right] &= (1-\alpha) q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s_t,\bar{a}) - q^*_\alpha(s,\bar{a}) \\ &\quad + \alpha(r^{\pi_0} + \gamma \sum_{s'} P^{\pi_0}(s' \mid s) \max_{a'} q_{\alpha,t}(s',a')) \\ &\quad + (1-\alpha)(q_t(s_t,\bar{a}) - q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s_t,\bar{a})). \end{aligned}$$

Define

$$c_t \triangleq (1-\alpha)(q_t(s_t,\bar{a}) - q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s_t,\bar{a})).$$

See that c_t converges to zero w.p. 1, since q_t converges to $q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}$. Furthermore, using Lemma 9, we have that

$$(1-\alpha) q^{\pi^\alpha(\pi^*_\alpha,\pi_0)}(s_t,\bar{a}) - q^*_\alpha(s,\bar{a}) = -\alpha(r^{\pi_0} + \gamma \sum_{s'} P^{\pi_0}(s' \mid s) \max_{a'} q^*_\alpha(s',a')).$$

Thus,

$$\begin{aligned}\mathbb{E}\left[f_t(s_t,\bar{a}) \mid \mathcal{F}_t\right] &= -\alpha(r^{\pi_0} + \gamma\sum_{s'} P^{\pi_0}(s' \mid s)\max_{a'} q^*_\alpha(s',a')) + \alpha(r^{\pi_0} + \gamma\sum_{s'} P^{\pi_0}(s' \mid s)\max_{a'} q_{\alpha,t}(s',a')) + c_t \\ &= \alpha\gamma\sum_{s'} P^{\pi_0}(s' \mid s)(\max_{a'} q_{\alpha,t}(s',a') - \max_{a'} q^*_\alpha(s',a')) + c_t \\ &= \alpha\gamma\sum_{s'} P^{\pi_0}(s' \mid s)|(\max_{a'} q_{\alpha,t}(s',a') - \max_{a'} q^*_\alpha(s',a'))| + c_t \\ &= \alpha\gamma\sum_{s'} P^{\pi_0}(s' \mid s)\max_{a'} |(q_{\alpha,t}(s',a') - q^*_\alpha(s',a'))| + c_t \\ &= \alpha\gamma\max_{s',a'} ||q_{\alpha,t} - q^*_\alpha|| + c_t\end{aligned}$$

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

$$\mathbb{E}\left[f_t(s_t,\bar{a}) \mid \mathcal{F}_t\right] \le \alpha\gamma\max_{s',a'} ||q_{\alpha,t} - q^*_\alpha|| + c_t,$$

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

We conclude that all conditions of Lemma 25 are satisfied for each $\bar{a} \in \mathcal{A}$ and, thus, Lemma 25 establishes the convergence of the procedure.

L.3. Proof of the gradients' equivalence in section 5.2

Proof.

$$\begin{aligned}\nabla_u q^\pi_\sigma(s,u) &= \nabla_u \int\limits_A \mathcal{N}(u'|u,\sigma)\, q^\pi(s,u')\, du' \\ &= \int\limits_A q^\pi(s,u')\, \nabla_u\mathcal{N}(u'|u,\sigma)\, du' \\ &= -\int\limits_A q^\pi(s,u')\, \nabla_{u'}\mathcal{N}(u'|u,\sigma)\, du' \\ &= -\, q^\pi(s,u')\,\mathcal{N}(u'|u,\sigma)|_{-\infty}^{\infty} + \int\limits_A \mathcal{N}(u'|u,\sigma)\, \nabla_{u'} q^\pi(s,u')\, du' \\ &= \int\limits_A \mathcal{N}(u'|u,\sigma)\, \nabla_{u'} q^\pi(s,u')\, du'\end{aligned}$$

Where we used integration by parts.

□