Journal of Machine Learning Research 22 (2021) 1-83 Submitted 8/19; Revised 12/20; Published 4/21

Safe Policy Iteration: A Monotonically Improving Approximate Policy Iteration Approach

Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
DEIB, Politecnico di Milano
Milano, Italy

Matteo Pirotta PIROTTA@FB.COM
Facebook AI Research
Paris, France

Daniele Calandriello DANIELE.CALANDRIELLO@IIT.IT
Istituto Italiano di Tecnologia
Genova, Italy

Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
DEIB, Politecnico di Milano
Milano, Italy

Editor: Peter Auer

Abstract

This paper presents a study of the policy improvement step that can be usefully exploited by approximate policy–iteration algorithms. When either the policy evaluation step or the policy improvement step returns an approximated result, the sequence of policies produced by policy iteration may not be monotonically increasing, and oscillations may occur. To address this issue, we consider safe policy improvements, i.e., at each iteration, we search for a policy that maximizes a lower bound to the policy improvement w.r.t. the current policy, until no improving policy can be found. We propose three safe policy–iteration schemas that differ in the way the next policy is chosen w.r.t. the estimated greedy policy. Besides being theoretically derived and discussed, the proposed algorithms are empirically evaluated and compared on some chain-walk domains, the prison domain, and on the Blackjack card game.

Keywords: Reinforcement Learning, Approximate Dynamic Programming, Approximate Policy Iteration, Policy Oscillation, Policy Chattering, Markov Decision Process

1. Introduction

Markov Decision Processes (MDPs) are widely used to model sequential decision-making problems under uncertainty (Puterman, 2014). In the last decades, a large body of research from control theory, operation research, and artificial intelligence has been devoted to the solution of MDPs. When a model of the environment is available, MDPs can be solved by dynamic programming algorithms or linear programming. On the contrary, when no or little prior knowledge about the model is known or when the problem is too complex for an exact solution, approximate methods need to be considered, like those studied

in the Reinforcement Learning (RL, Sutton and Barto, 2018) and Approximate Dynamic Programming (ADP, Bertsekas, 2011).

In this paper, we focus on approaches derived from Policy Iteration (PI, Howard, 1960), one of the two main classes of dynamic programming algorithms to solve MDPs. PI is an iterative algorithm that alternates between two steps: *policy evaluation* and *policy improvement*. At each iteration, the current policy π_k is evaluated computing the action–value function Q^{π_k} and the new policy π_{k+1} is generated by taking the greedy policy w.r.t. Q^{π_k}, i.e., the policy that in each state takes the best action according to Q^{π_k}. Policy iteration generates a sequence of monotonically improving policies that reaches the optimal policy in a finite number of iterations (Ye, 2011; Scherrer, 2013a).

When either Q^{π_k} or the corresponding greedy policy π_{k+1} cannot be computed exactly, *Approximate Policy Iteration* (API, Bertsekas, 2011) algorithms need to be considered. A large number of methods tackling this problem have been proposed in the literature (Scherrer, 2014). The standard API (Bertsekas and Tsitsiklis, 1996) simply computes the greedy policy w.r.t. to the estimated value function $\widehat{Q}^{\pi_k}$. However, in this case, the approximately greedy policy π_{k+1} may perform worse than π_k, leading, thus, to policy oscillation phenomena (Bertsekas, 2011; Wagner, 2011). Empirically, the value Q^{π_k} rapidly improves in the initial iterations, then gets stuck or oscillates without any further policy improvement (named stationary phase Munos, 2003). Most API studies and algorithms focus on reducing the approximation error in the policy evaluation step (Lagoudakis and Parr, 2003a; Munos, 2005; Lazaric et al., 2010; Gabillon et al., 2011), and, then, perform policy improvement by taking the relative greedy policy. However, the quality of the sequence of generated policies may oscillate or diverge when the policy evaluation is approximated, independently of the policy evaluation method (Bertsekas and Tsitsiklis, 1996; Bertsekas, 2011).

Almost all the API algorithms intrinsically implement a *generalized* policy iteration scheme (Sutton and Barto, 2018) because the improvement of the policy is performed over an incomplete estimate of the value functions. This idea was used in Scherrer et al. (2012); Scherrer (2013b) to generalize over Value Iteration (VI) and PI methods at the cost of additional free parameters.

It has been pointed out that the key source of this oscillation phenomena is the discontinuity introduced by the greedy improvement (De Farias and Van Roy, 2000; Perkins and Pendrith, 2002; Perkins and Precup, 2002). Approximate Linear Programming (de Farias and Roy, 2003) solves the RL problem in one shot, but typically assumes the knowledge of the transition model and the approximation comes from the fact that the value function is represented as a function approximator (e.g., linear in a vector of known features). As noticed in the early stages of RL (Singh et al., 1994), stochastic policies may represent the solution to many issues. A class of approaches deals with the oscillation phenomena by proposing converging algorithms that exploit smaller updates (*soft updates*) in the space of stochastic policies, instead of iterating on a sequence of greedy policies computed on approximated action–value functions (Perkins and Precup, 2002; Kakade and Langford, 2002; Lagoudakis and Parr, 2003b; Wagner, 2011; Azar et al., 2012). The idea is that the action–value function of a policy π can produce a good estimate of the performance of another policy π' when the two policies give rise to similar state distributions. This condition can be guaranteed when the policies themselves are similar. Incremental policy updates are also considered in the related class of policy gradient algorithms (e.g., Sutton et al., 1999;

Kakade, 2001; Peters et al., 2005). These methods share a common rationale based on managing the trade–off between jumping to the greedy policy according to the currently estimated action–value function and remaining close to the current policy avoiding too uncertain updates. From an intuitive point of view, the more we trust our value function estimate, the more we can move far from the current policy. This very simple idea has been developed in several works and for different purposes (e.g., Pirotta et al., 2013a; Abbasi-Yadkori et al., 2016; Ghavamzadeh et al., 2016; Papini et al., 2017; Metelli et al., 2018) and it represents the theoretical grounding of some of the most successful RL algorithms (e.g., Schulman et al., 2015).

Other works focus on the "optimistic" or "modified" PI approach. This variant of policy iteration is based on an approximate evaluation of the preceding policy obtained by applying the Bellman operator a finite number of times. While Scherrer et al. (2012) have derived a convergence and finite samples analysis for the "optimistic" policy iteration generalization of the classification–based policy iteration, Wagner (2013) has investigated the connection between optimistic policy iteration and natural actor–critic algorithms. They have shown that the natural actor–critic algorithm for Gibbs policy is a special case of optimistic policy iteration. In addition, they suggested that it is possible to get convergence guarantees for PI approaches exploiting the theory behind gradient methods. However, they proved that, while having the potential of overcoming policy oscillation, Gibbs soft–greedy value function approaches never converge to the optimal policy.

Another research line focuses on the exploitation of non–stationary policy sequences (Scherrer and Lesner, 2012). The authors propose an algorithm that, at each iteration, approximates the value function of a policy that loops over the last m greedy policies generated by the algorithm (possibly also considering all the policies generated from the beginning of the algorithm). Thus, the resulting policy is non–stationary and has a regularizing effect on the learning process. They show that, by employing non–stationary policies, it is possible to obtain better convergence rates (Scherrer and Lesner, 2012; Lesner and Scherrer, 2013). The methods have also been applied to the case of modified PI (Lesner and Scherrer, 2015).

Recently, one research line has imported the classical idea of delaying the backup operation, successfully applied in the famous TD methods (Sutton and Barto, 2018), to the PI framework by defining the *multi–step greedy* improvements (Efroni et al., 2018a). These works overcome the 1–step greedy update by defining an h–greedy policy ($h \geq 1$) as the policy that, from every state, is optimal for h time steps. This new improvement operator amounts to solve an h–horizon optimal control problem, reducing to the standard 1–step operator for $h = 1$, converges to the optimal policy in a number of iterations that decreases with h (Efroni et al., 2018a). Similarly to the TD(λ) case, it is possible to mix different horizons using a parameter to tradeoff between the 1–step greedy improvement and more far–sighted updates (Efroni et al., 2018a). The idea is then extended to the approximate setting in Efroni et al. (2018b).

In this paper, we limit our scope to the classical API approaches, built on stationary policies[1] and, following the approach of Conservative Policy Iteration (CPI, Kakade and Langford, 2002), which adopts soft updates to avoid oscillation phenomena, and recently extended to use deep architectures (Vieillard et al., 2020). We extend CPI Kakade

1. In the case of infinite–horizon γ–discounted MDPs, it is known that there exists a stationary optimal policy.

and Langford (2002) by introducing a tighter lower bound on the performance improvement, that allows designing API algorithms useful both in model–free contexts and when a restricted subset of policies is considered. These algorithms produce a sequence of monotonically improving policies and are characterized by a faster-improving rate compared to CPI. Furthermore, we devise an update schema that makes use of per–state combination coefficients and a novel generalization, not present in Pirotta et al. (2013b), that employs per–state–action coefficients.

The main contributions of this paper are theoretical, algorithmic, and experimental consisting in:

1. the introduction of new, more general lower bounds on the policy improvement, tighter than the one presented in Pirotta et al. (2013b) (Section 3);

2. the proposal of three approximate policy–iteration algorithms whose performance improvement moves toward the estimated greedy policy by maximizing the policy improvement bounds. The first two of them have already been presented in Pirotta et al. (2013b), while the third and more general is novel and presented here in a unified framework (Section 4);

3. a complete PAC analysis of the approximate version of the presented algorithms, with finite–sample improvement guarantees for the single iteration (Section 5);

4. an empirical evaluation and comparison of the proposed algorithms with related approaches (Section 6).

The rest of the paper is organized as follows. Section 2 introduces notation and the necessary background. Section 3 derives the bounds on the difference between the performance of two policies and provides the policy improvement bounds. Based on these bounds, we present the exact algorithms in Section 4 and the approximated in Section 5. In Section 6, the algorithms are empirically evaluated and compared with other approximate policy–iteration algorithms on several variants of the chain–walk domain (Lagoudakis and Parr, 2003a), Prison environment (Azar et al., 2012), and in a simplified version of the Blackjack (Dutech et al., 2005).

2. Preliminaries

In this section, we report the essential background that will be employed in the remainder of the paper.

Markov Decision Processes A discrete–time finite Markov Decision Process (MDP, Puterman, 2014) is defined as a 6–tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{P}$ is a Markovian transition model where $\mathcal{P}(s'|s,a)$ is the probability of making a transition to state s' when taking action a from state s, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, such that $\mathcal{R}(s,a)$ is the expected immediate reward for the state–action pair (s,a), $\gamma \in [0,1)$ is the discount factor for future rewards, and μ is the initial state distribution. The policy of an agent is characterized by a density distribution $\pi(a|s)$ that specifies the probability of taking action a in state s. When the policy is deterministic, with little abuse of notation, we use $\pi(s)$ to denote the action

prescribed in state s. We consider infinite–horizon problems where the future rewards are exponentially discounted with γ. For each state s, we define the utility of following a stationary policy π as:

$$V^{\pi}(s) = \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{+\infty} \gamma^t \mathcal{R}(s_t, a_t) | s_0 = s \right].$$

It is known that V^{π} is the unique solution of the following recursive (Bellman) equation:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) V^{\pi}(s') \right).$$

Policies can be ranked by their expected discounted reward starting from the state distribution :

$$J^{\pi} = \sum_{s \in \mathcal{S}} \mu(s) V^{\pi}(s) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{R}(s,a),$$

where $d^{\pi}(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | \pi, \mathcal{M})$ is the γ–discounted future state distribution for a starting state distribution (Sutton et al., 1999). Solving an MDP means to find a policy π^* that maximizes the expected long–term reward: $\pi^* \in \arg\max_{\pi \in \Pi^{\mathrm{SR}}} J^{\pi}$, where Π^{SR} is the set of stationary Markovian randomized policies. For any MDP, there exists at least one deterministic optimal policy that simultaneously maximizes $V^{\pi}(s)$, $\forall s \in \mathcal{S}$, i.e., exists $\pi^* \in \Pi^{\mathrm{SD}}$, where Π^{SD} is the set of stationary Markovian deterministic policies (Puterman, 2014). For control purposes, it is more convenient to consider the action–value function $Q^{\pi}(s,a)$, i.e., the value of taking action a in state s and following a policy π thereafter:

$$Q^{\pi}(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) \sum_{a' \in \mathcal{A}} \pi(a'|s') Q^{\pi}(s',a').$$

Given $Q^{\pi}(s,a)$, we define a greedy policy as $\pi^+(s) \in \arg\max_{a \in \mathcal{A}} Q^{\pi}(s,a)$. Furthermore, we define the advantage function as:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s),$$

that quantifies the convenience of performing action a in state s instead of following policy π. Furthermore, for each state s, we define the advantage of a policy π' over policy π as $A_{\pi}^{\pi'}(s) = \sum_{a \in \mathcal{A}} \pi'(a|s) A^{\pi}(s,a)$ and, following what done by Kakade and Langford (2002), we define its expected value w.r.t. the γ–discounted state distribution d^{π} as $\mathbb{A}_{\pi}^{\pi'} = \sum_{s \in \mathcal{S}} d^{\pi}(s) A_{\pi}^{\pi'}(s)$.

Notation Vectors are assumed to be columns and are denoted with lowercase bold letters, like $\mathbf{v}$; while matrices are denoted with upper case bold letters, like $\mathbf{M}$. For brevity, in the following, we will use matrix notation, where $\mathbf{I}$ denotes the identity matrix and $\mathbf{e}$ is a column vector of all ones (with sizes apparent from the context). Given a (column) vector $\mathbf{v}$, $\mathbf{v}^{\mathtt{T}}$ denotes the corresponding row vector. Whenever necessary, a d–dimensional vector $\mathbf{v}$ will be treated as a $d \times 1$ matrix, and, symmetrically, a row vector $\mathbf{v}^{\mathtt{T}}$ will be treated as a $1 \times d$ matrix. Given a matrix $\mathbf{M}$, $\mathbf{M}^{\mathtt{T}}$ denotes its transpose, and, given a non–singular square matrix $\mathbf{M}$, $\mathbf{M}^{-1}$ denotes its inverse. For brevity, $\mathbf{M}^{-\mathtt{T}} = \left(\mathbf{M}^{\mathtt{T}}\right)^{-1}$. For a vector $\mathbf{v}$, we indicate with $\mathbf{v}(i)$ its i–th component and for a matrix $\mathbf{M}$, we indicate

with $\mathbf{M}(i,j)$ the component at row i and column j. Let $p \in [1,\infty)$, we define the L_p–norm of a d–dimensional vector $\mathbf{v}$ as $\|\mathbf{v}\|_p^p = \sum_{i=1}^{d} |\mathbf{v}(i)|^p$. The L_∞–norm of $\mathbf{v}$ is given by $\|\mathbf{v}\|_\infty = \max_{i\in\{1,\dots,d\}} |\mathbf{v}(i)|$. Moreover, we define the span seminorm of $\mathbf{v}$ as $\mathrm{sp}(\mathbf{v}) = \max_{i\in\{1,\dots,d\}} \mathbf{v}(i) - \min_{i\in\{1,\dots,d\}} \mathbf{v}(i)$. We consider the L_p–norms of matrices induced by the corresponding vector norm, defined as $\|\mathbf{M}\|_p = \sup_{\mathbf{v}\,:\,\|\mathbf{v}\|_p \leq 1} \|\mathbf{M}\mathbf{v}\|_p$. In particular, the L_1–norm $\|\mathbf{M}\|_1$ of a matrix $\mathbf{M}$ is its maximum absolute column sum, while its L_∞–norm $\|\mathbf{M}\|_\infty$ is its maximum absolute row sum. It follows that $\|\mathbf{M}\|_1 = \left\|\mathbf{M}^\intercal\right\|_\infty$ (Petersen and Pedersen, 2012). If $\mathbf{v}$ is a probability column vector of size n and $\mathbf{M}$ is an $n \times m$ matrix, we indicate with $\|\mathbf{M}\|_{p,\mathbf{v}}$ the expectation of the L_p–norm of the columns of $\mathbf{M}$ taken under $\mathbf{v}$, i.e., $\|\mathbf{M}\|_{p,\mathbf{v}} = \sum_{i=1}^{n} \mathbf{v}(i) \left(\sum_{j=1}^{m} |\mathbf{M}(i,j)|^p\right)^{1/p}$.

Matrix Notation for MDPs Using matrix notation, we can rewrite previous equations as follows (e.g., Puterman, 2014; Wang et al., 2007):

$$\begin{aligned}
\mathbf{v}^\pi &= \boldsymbol{\pi}\left(\mathbf{r} + \gamma \mathbf{P}\mathbf{v}^\pi\right) = \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{v}^\pi = \left(\mathbf{I} - \gamma \mathbf{P}^\pi\right)^{-1} \mathbf{r}^\pi \\
\mathbf{q}^\pi &= \mathbf{r} + \gamma \mathbf{P}\boldsymbol{\pi}\mathbf{q}^\pi = \mathbf{r} + \gamma \mathbf{P}\mathbf{v}^\pi \\
\mathbf{d}^\pi &= (1-\gamma)\boldsymbol{\mu} + \gamma \mathbf{P}^{\pi\intercal}\mathbf{d}^\pi = (1-\gamma)(\mathbf{I} - \gamma \mathbf{P}^\pi)^{-\intercal}\boldsymbol{\mu} \\
J^\pi &= \boldsymbol{\mu}^\intercal \mathbf{v}^\pi = \boldsymbol{\mu}^\intercal \left(\mathbf{I} - \gamma \mathbf{P}^\pi\right)^{-1} \mathbf{r}^\pi = \frac{1}{1-\gamma} \mathbf{d}^{\pi\intercal}\mathbf{r}^\pi \\
\mathbb{A}_\pi^{\pi'} &= \mathbf{d}^{\pi\intercal}\boldsymbol{\pi'}\mathbf{a}^\pi = \mathbf{d}^{\pi\intercal}\mathbf{a}_\pi^{\pi'},
\end{aligned} \tag{1}$$

where J^π and $\mathbb{A}_\pi^{\pi'}$ are scalars, $\mathbf{v}^\pi$, $\mathbf{r}^\pi$, $\mathbf{d}^\pi$, $\boldsymbol{\mu}$, and $\mathbf{a}_\pi^{\pi'}$ are vectors of size $|\mathcal{S}|$, $\mathbf{q}^\pi$, $\mathbf{r}$, and $\mathbf{a}^\pi$ are vectors of size $|\mathcal{S}||\mathcal{A}|$, $\mathbf{P}$ is a stochastic matrix of size $(|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|)$ that contains the transition model of the process $\mathbf{P}((s,a),s') = \mathcal{P}(s'|s,a)$, $\boldsymbol{\pi}$ is a stochastic matrix of size $(|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|)$ that describes policy π:

$$\boldsymbol{\pi}(s,(s',a)) = \begin{cases} \pi(a|s) & \text{if } s' = s \\ 0 & \text{otherwise} \end{cases},$$

and $\mathbf{P}^\pi = \boldsymbol{\pi}\mathbf{P}$ is a stochastic matrix $|\mathcal{S}| \times |\mathcal{S}|$ that represents the state transition matrix under policy π, i.e., $\mathcal{P}^\pi(s'|s) = \sum_{a\in\mathcal{A}} \mathcal{P}(s'|s,a)\pi(a|s)$.

3. Bound on Policy Improvement

This section is devoted to the study of the performance improvement $J^{\pi'} - J^\pi$ of a policy π' over a policy π given the policy advantage function $A_\pi^{\pi'}$. Specifically, we will present two lower bounds of the improvement $J^{\pi'} - J^\pi$. The first bound (Theorem 3) is tighter, but it is hard to optimize due to the presence of quantities that are typically unknown. For this reason, it will be employed with USPI only. The second bound (Corollary 4) is a relaxation of the previous one, allows a more straightforward optimization and it will be used for SSPI and SASPI. The presented bounds are tighter compared to that by Kakade and Langford (2002) and Pirotta et al. (2013b).[2] As we will see, $A_\pi^{\pi'}$ can provide a good estimate of $J^{\pi'}$ only when the two policies π and π' visit the states with similar probabilities,

2. A better bound allows faster improving rates while preserving the property of having a monotonically improving sequence of policies.

i.e., $d^{\pi'} \simeq d^{\pi}$. The following lemma provides an upper bound to the difference between the two γ–discounted future state distributions.

Lemma 1. *Let π and π' be two stationary policies for an infinite horizon MDP $\mathcal{M}$. The L_1–norm of the difference between their γ–discounted future state distributions under starting state distribution μ can be upper bounded as follows:*

$$\left\|\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}\right\|_1 \leq \frac{\gamma}{1-\gamma} \left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_{1,\mathbf{d}^{\pi}}.$$

Proof To prove the lemma, we rewrite the difference $\mathbf{d}^{\pi'^{\mathtt{T}}} - \mathbf{d}^{\pi^{\mathtt{T}}}$ as follows:

$$\begin{aligned}
\mathbf{d}^{\pi'^{\mathtt{T}}} - \mathbf{d}^{\pi^{\mathtt{T}}} &= (1-\gamma)\boldsymbol{\mu}^{\mathtt{T}} + \gamma \mathbf{d}^{\pi'^{\mathtt{T}}} \mathbf{P}^{\pi'} - \left((1-\gamma)\boldsymbol{\mu}^{\mathtt{T}} + \gamma \mathbf{d}^{\pi^{\mathtt{T}}} \mathbf{P}^{\pi}\right) \\
&= \gamma \mathbf{d}^{\pi'^{\mathtt{T}}} \mathbf{P}^{\pi'} - \gamma \mathbf{d}^{\pi^{\mathtt{T}}} \mathbf{P}^{\pi} \\
&= \gamma \left(\mathbf{d}^{\pi'^{\mathtt{T}}} - \mathbf{d}^{\pi^{\mathtt{T}}}\right) \mathbf{P}^{\pi'} + \gamma \mathbf{d}^{\pi^{\mathtt{T}}} \left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right) \\
&= \gamma \mathbf{d}^{\pi^{\mathtt{T}}} \left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right) \left(\mathbf{I} - \gamma \mathbf{P}^{\pi'}\right)^{-1},
\end{aligned}$$

where the first equality follows from the convergence of Neumann series (e.g., Wang et al., 2007; Pirotta et al., 2013b). We restate here the result:

$$\begin{aligned}
\mathbf{d}^{\pi} &= (1-\gamma) \left[\sum_{t=0}^{+\infty} (\gamma \boldsymbol{\pi} \mathbf{P})^t\right]^{\mathtt{T}} \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\mu} + \left[\sum_{t=1}^{+\infty} (\gamma \mathbf{P}^{\pi})^t\right]^{\mathtt{T}} \boldsymbol{\mu} \\
&= (1-\gamma)\boldsymbol{\mu} + \left[\sum_{\tau=0}^{+\infty} (\gamma \mathbf{P}^{\pi})^{\tau+1}\right]^{\mathtt{T}} \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\mu} + (\gamma \mathbf{P}^{\pi})^{\mathtt{T}} \left[\sum_{\tau=0}^{+\infty} (\gamma \mathbf{P}^{\pi})^{\tau}\right]^{\mathtt{T}} \boldsymbol{\mu} \\
&= (1-\gamma)\boldsymbol{\mu} + \gamma \mathbf{P}^{\pi^{\mathtt{T}}} \mathbf{d}^{\pi}.
\end{aligned}$$

It is worth to notice that the inverse of matrix $\mathbf{I} - \gamma \mathbf{P}^{\pi'}$ exists for any $\gamma < 1$ since $\mathbf{P}^{\pi'}$ has the maximum eigenvalue equal to 1 being a stochastic matrix (Suzuki, 1976).

By recalling that $\mathbf{d}^{\pi^{\mathtt{T}}} \left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right)$ is a row vector, we derive the following inequality that will be employed for the L_1–norm bound:

$$\begin{aligned}
\left\|\mathbf{d}^{\pi^{\mathtt{T}}} \left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right)\right\|_{\infty} &= \sum_{s' \in \mathcal{S}} \left|\sum_{s \in \mathcal{S}} d^{\pi}(s) \left(\mathcal{P}^{\pi'}(s'|s) - \mathcal{P}^{\pi}(s'|s)\right)\right| \\
&\leq \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{s' \in \mathcal{S}} \left|\mathcal{P}^{\pi'}(s'|s) - \mathcal{P}^{\pi}(s'|s)\right| \\
&= \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{s' \in \mathcal{S}} \left|\sum_{a \in \mathcal{A}} \mathcal{P}(s'|s,a) \left(\pi'(a|s) - \pi(a|s)\right)\right| \\
&\leq \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \left|\pi'(a|s) - \pi(a|s)\right| \underbrace{\sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a)}_{=1} \qquad \text{(P.1)}
\end{aligned}$$

$$\leq \sum_{s \in \mathcal{S}} d^{\pi}(s) \left\| \pi'(\cdot|s) - \pi(\cdot|s) \right\|_1, \tag{P.2}$$

where line (P.1) derives from pushing the absolute value inside the summation and line (P.2) is obtained from the definition of L_1–norm. From the first equation, the bound on the L_1–norm follows recalling that $\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}$ is a column vector and, consequently, $\mathbf{d}^{\pi'\mathtt{T}} - \mathbf{d}^{\pi\mathtt{T}}$ is a row vector:

$$\begin{aligned}
\left\| \mathbf{d}^{\pi'} - \mathbf{d}^{\pi} \right\|_1 &= \left\| \mathbf{d}^{\pi'\mathtt{T}} - \mathbf{d}^{\pi\mathtt{T}} \right\|_\infty \\
&\leq \gamma \left\| \mathbf{d}^{\pi\mathtt{T}} \left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi} \right) \right\|_\infty \left\| \left(\mathbf{I} - \gamma \mathbf{P}^{\pi'} \right)^{-1} \right\|_\infty \\
&\leq \frac{\gamma}{1-\gamma} \left\| \boldsymbol{\pi'} - \boldsymbol{\pi} \right\|_{1, \mathbf{d}^{\pi}}.
\end{aligned}$$

For this part of the proof, we have exploited the consistency of the L_∞–norm. In the last equality, we have used the notion that $\mathbf{d}^{\pi}$ is a probability vector and observed that $\left\| \left(\mathbf{I} - \gamma \mathbf{P}^{\pi'} \right)^{-1} \right\|_\infty = \frac{1}{1-\gamma}$. ■

As a further step to prove the main theorem, it is useful to rewrite the difference between the performance of policy π' and the one of policy π as a function of the policy advantage function $A_{\pi}^{\pi'}$.

Lemma 2. *(Kakade and Langford, 2002) For any stationary policies π and π' and any starting state distribution μ:*

$$J^{\pi'} - J^{\pi} = \frac{1}{1-\gamma} \mathbf{d}^{\pi'\mathtt{T}} \mathbf{a}_{\pi}^{\pi'}.$$

Unfortunately, computing the improvement of policy π' w.r.t. to π using the previous lemma is very expensive, since it requires estimating $d^{\pi'}$ for each candidate π'. In the following, we will provide a bound to the policy improvement and we will show how it is possible to find a policy π' that optimizes its value.

Theorem 3. *For any stationary policies π and π' and any starting state distribution μ, given any baseline policy π_b, the difference between the performance of π' and the one of π can be lower bounded as follows:*

$$J^{\pi'} - J^{\pi} \geq \frac{1}{1-\gamma} \mathbf{d}^{\pi_b\mathtt{T}} \mathbf{a}_{\pi}^{\pi'} - \frac{\gamma}{(1-\gamma)^2} \left\| \boldsymbol{\pi'} - \boldsymbol{\pi_b} \right\|_{1, \mathbf{d}^{\pi_b}} \frac{\mathrm{sp}\left(\mathbf{a}_{\pi}^{\pi'} \right)}{2}.$$

Proof The proof can be obtained starting from Lemma 2:

$$\begin{aligned}
(1-\gamma) \left(J^{\pi'} - J^{\pi} \right) = \mathbf{d}^{\pi'\mathtt{T}} \mathbf{a}_{\pi}^{\pi'} &= \mathbf{d}^{\pi_b\mathtt{T}} \mathbf{a}_{\pi}^{\pi'} + \left(\mathbf{d}^{\pi'\mathtt{T}} - \mathbf{d}^{\pi_b\mathtt{T}} \right) \mathbf{a}_{\pi}^{\pi'} \\
&\geq \mathbf{d}^{\pi_b\mathtt{T}} \mathbf{a}_{\pi}^{\pi'} - \left| \left(\mathbf{d}^{\pi'} - \mathbf{d}^{\pi_b} \right)^{\mathtt{T}} \mathbf{a}_{\pi}^{\pi'} \right|
\end{aligned} \tag{P.3}$$

$$\geq \mathbf{d}^{\pi_b\mathtt{T}}\mathbf{a}_\pi^{\pi'} - \left\|\mathbf{d}^{\pi'} - \mathbf{d}^{\pi_b}\right\|_1 \frac{\mathrm{sp}\left(\mathbf{a}_\pi^{\pi'}\right)}{2} \tag{P.4}$$

$$\geq \mathbf{d}^{\pi_b\mathtt{T}}\mathbf{a}_\pi^{\pi'} - \frac{\gamma}{1-\gamma}\left\|\boldsymbol{\pi'} - \boldsymbol{\pi_b}\right\|_{1,\mathbf{d}^{\pi_b}} \frac{\mathrm{sp}\left(\mathbf{a}_\pi^{\pi'}\right)}{2}. \tag{P.5}$$

Statement (P.3) is a simple mathematical manipulation ($a + b \geq a - |b|,\ \forall a, b \in \mathbb{R}$), while the inequality (P.4) follows from Lemma 23 since $\mathbf{c} = \mathbf{d}^{\pi'} - \mathbf{d}^{\pi_b}$ is a vector satisfying $\mathbf{c}^\mathtt{T}\mathbf{e} = (\mathbf{d}^{\pi'} - \mathbf{d}^{\pi_b})^\mathtt{T}\mathbf{e} = 1 - 1 = 0$. The theorem is proved in (P.5) by exploiting the bound in Corollary 1.

■

The theorem is presented for a general *baseline* policy π_b that, ideally, is employed to collect the samples. In principle, π_b can be different from both π and π', although, typically, we select $\pi_b = \pi$. The bound is the sum of two terms: the advantage of policy π' over policy π averaged according to the distribution induced by policy π_b and a penalization term that is a function of the discrepancy between policy π' and policy π_b and the range of variability of the advantage function $A_\pi^{\pi'}$.[3]

Remark 1 (Comparison with Pirotta et al., 2013b)**.** The bound presented in Theorem 3 strictly improves Theorem 3.5 in Pirotta et al. (2013b), since the L_∞–norm between the policies π' and π has been replaced with an expectation taken w.r.t. to the γ–discounted stationary distribution d^π of the state–wise L_1–norms. Indeed:

$$\left\|\boldsymbol{\pi'} - \boldsymbol{\pi_b}\right\|_{1,\mathbf{d}^{\pi_b}} \leq \left\|\boldsymbol{\pi'} - \boldsymbol{\pi_b}\right\|_\infty.$$

Since the bound provided by Pirotta et al. (2013b) was already tight (being an improvement over CPI), it follows that our bound is tight as well. A similar derivation was previously provided in Achiam et al. (2017, Corollary 1) and Metelli et al. (2018, Corollary 3.1). Now we introduce a looser but simplified version of the bound in Theorem 3 that will be useful later.

Corollary 4. *For any stationary policies π and π' and any starting state distribution μ, the difference between the performance of π' and the one of π can be lower bounded as follows:*

$$J^{\pi'} - J^\pi \geq \frac{1}{1-\gamma}\mathbb{A}_\pi^{\pi'} - \frac{\gamma}{(1-\gamma)^2}\left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_\infty^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2}.$$

For completeness, we mention that a different (looser) bound on the policy difference in norm can be obtained by using Pinsker's inequality (Csiszár and Körner, 2011) stating that

$$\left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_{1,\mathbf{d}^\pi} = \mathbb{E}_{s\sim\mathbf{d}^{\pi_b}}\left[\|\pi(\cdot|s) - \pi_b(\cdot|s)\|_1\right] \leq \sqrt{\frac{1}{2}\mathbb{E}_{s\sim\mathbf{d}^{\pi_b}}\left[D_{\mathrm{KL}}(\pi(\cdot|s)\|\pi_b(\cdot|s))\right]}.$$

This bound was initially used in Pirotta et al. (2013a) to derive a simplified lower–bound for parametrized policies and it has been adopted frequently in the literature after that (e.g., Schulman et al., 2015; Achiam et al., 2017; Papini et al., 2017).

3. We tried to keep Theorem 3 as general as possible to favor its reuse in different contexts. Nonetheless, in the following, we will consider the particular case where π_b is equal to π. The possibility of selecting a suitable $\pi_b \neq \pi$ opens new and interesting lines of research that are out of the scope of this paper.

4. Exact Safe Policy Iteration

As we have already mentioned, Conservative Policy Iteration (CPI) approach successfully aims at overcoming policy degradation issues in approximate contexts. Indeed, while PI algorithms, moving from a greedy policy to another, guarantees to improve the performance at each iteration until convergence to the optimal policy only when no approximation is involved, the CPI algorithm performs a more conservative improvement step to ensure monotonically increasing policy performance even in the approximate setting. In this framework, following the approach of CPI (Kakade and Langford, 2002), we propose a new set of techniques, called Safe Policy Iteration (SPI) algorithms (Pirotta et al., 2013b).

The idea is to produce a sequence of monotonically improving policies and stop when no improvement can be guaranteed. The policy improvement step is a trade-off between the current policy π and a target policy $\overline{\pi}$ according to $\pi' = \alpha\overline{\pi} + (1-\alpha)\pi$, where the trade-off coefficient $\alpha \in [0,1]$ results from a maximization of a lower bound on the policy improvement. The main benefit of exploiting SPI methods instead of CPI one is quantifiable in a faster convergence to the optimal solution, due to a maximization of a better lower bounds w.r.t. that of CPI.

In the following, we analyze the exact case (in which the value functions are known without approximation) and we propose three safe policy-iteration algorithms: Unique–parameter Safe Policy Iteration (USPI), per–State–parameter Safe Policy Iteration (SSPI), and per–State–Action–parameter Safe Policy Iteration (SASPI). The main differences between the three algorithms lie in: i) the set of policies that they consider in the policy improvement step and ii) the policy improvement bounds (Section 3) employed. USPI (Section 4.1) employs a single coefficient α and its value is selected by maximizing the bound presented in Theorem 3. SSPI (Section 4.2) and SASPI (Section 4.3), instead, allow for a larger policy improvement space, considering different coefficients for each state $\alpha(s)$ and for each state-action pair $\alpha(s,a)$, respectively. Moreover, differently from USPI, they make use of the looser bound of Corollary 4 to select the value of the coefficients.

4.1 Unique–parameter Safe Policy Improvement

Following the approach proposed in CPI, USPI iteratively updates the current policy using a safe policy improvement. Given the current policy π and a target policy $\overline{\pi}$ (which may be different from the greedy policy), we define the update rule of the policy improvement step as:

$$\pi' = \alpha\overline{\pi} + (1-\alpha)\pi,$$

where $\alpha \in [0,1]$ is the scalar trade–off coefficient. It can easily be shown that if $A_\pi^{\overline{\pi}}(s) \geq 0$ for all s, then π' is not worse than π for any α. This condition always holds when the target policy $\overline{\pi}$ is the greedy policy π^+. Nevertheless, we will show that the greedy target policy is not always the optimal choice. In general, it is always possible to find an improving step–size α whenever the target policy $\overline{\pi}$ belongs to the set $\{\overline{\pi} \in \Pi^{\mathrm{SR}} | \mathbb{A}_\pi^{\overline{\pi}} \geq 0\}$. At each iteration, we seek the α that yields the maximal performance improvement in the worst case. For this reason, α is chosen to maximize the bound in Theorem 3. By taking $\pi_b = \pi$, the value of α that maximizes this lower bound is given by the following corollary.

Corollary 5. *If $\mathbb{A}_\pi^{\overline{\pi}} \geq 0$, then, using $\alpha^* = \frac{(1-\gamma)\mathbb{A}_\pi^{\overline{\pi}}}{\gamma\|\overline{\boldsymbol{\pi}}-\boldsymbol{\pi}\|_{1,\mathbf{d}^\pi}\operatorname{sp}(\mathbf{a}_\pi^{\overline{\pi}})}$, we set $\alpha = \min\{1, \alpha^*\}$, so that when $\alpha^* \leq 1$ we can guarantee the following policy improvement:*

$$J^{\pi'} - J^\pi \geq \frac{(\mathbb{A}_\pi^{\overline{\pi}})^2}{2\gamma\left\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\right\|_{1,\mathbf{d}^\pi}\operatorname{sp}(\mathbf{a}_\pi^{\overline{\pi}})}$$

and when $\alpha^* > 1$, we perform a full update towards the target policy $\overline{\pi}$, i.e., we set $\alpha = 1$ so that $\pi' = \overline{\pi}$. In such a case, the policy improvement is given by Theorem 3 by setting $\pi_b = \pi$ and $\pi' = \overline{\pi}$.

Proof Setting $\boldsymbol{\pi_b} = \boldsymbol{\pi}$ and $\boldsymbol{\pi'} = \alpha\overline{\boldsymbol{\pi}} + (1-\alpha)\boldsymbol{\pi}$, we can manipulate the bound in Theorem 3. Let us consider the following derivation of the individual terms involved Theorem 3:

$$\begin{aligned}
\mathbf{d}^{\pi\mathtt{T}}\mathbf{a}_\pi^{\pi'} &= \mathbf{d}^{\pi\mathtt{T}}\left(\boldsymbol{\pi'} - \boldsymbol{\pi}\right)\mathbf{q}^\pi = \mathbf{d}^{\pi\mathtt{T}}\Big(\alpha\overline{\boldsymbol{\pi}} + (1-\alpha)\,\boldsymbol{\pi} - \boldsymbol{\pi}\Big)\mathbf{q}^\pi \\
&= \alpha\mathbf{d}^{\pi\mathtt{T}}\left(\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\right)\mathbf{q}^\pi = \alpha\mathbf{d}^{\pi\mathtt{T}}\mathbf{a}_\pi^{\overline{\pi}} = \alpha\mathbb{A}_\pi^{\overline{\pi}}
\end{aligned}$$

$$\begin{aligned}
\operatorname{sp}\Big(\mathbf{a}_\pi^{\pi'}\Big) &= \max_{s,s'\in\mathcal{S}}\left\{\left|\mathbf{a}_\pi^{\pi'}(s) - \mathbf{a}_\pi^{\pi'}(s')\right|\right\} \\
&= \max_{s,s'\in\mathcal{S}}\left\{\left|\sum_{a\in\mathcal{A}}\Big(\pi'(a|s) - \pi(a|s)\Big)Q^\pi(s,a) - \sum_{a\in\mathcal{A}}\Big(\pi'(a|s') - \pi(a|s')\Big)Q^\pi(s',a)\right|\right\} \\
&= \alpha\max_{s,s'\in\mathcal{S}}\left\{\left|\sum_{a\in\mathcal{A}}\Big(\overline{\pi}(a|s) - \pi(a|s)\Big)Q^\pi(s,a) - \sum_{a\in\mathcal{A}}\Big(\overline{\pi}(a|s') - \pi(a|s')\Big)Q^\pi(s',a)\right|\right\} \\
&= \max_{s,s'\in\mathcal{S}}\left\{\left|\mathbf{a}_\pi^{\overline{\pi}}(s) - \mathbf{a}_\pi^{\overline{\pi}}(s')\right|\right\} = \alpha\operatorname{sp}\big(\mathbf{a}_\pi^{\overline{\pi}}\big)
\end{aligned}$$

$$\|\pi'(\cdot|s) - \pi(\cdot|s)\|_1 = \alpha\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1, \quad \forall s \in \mathcal{S}.$$

By plugging the terms derived above in Theorem 3 we obtain:

$$J^{\pi'} - J^\pi \geq \alpha\frac{1}{1-\gamma}\mathbb{A}_\pi^{\overline{\pi}} - \alpha^2\frac{\gamma}{(1-\gamma)^2}\left\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\right\|_{1,\mathbf{d}^\pi}\frac{\operatorname{sp}\big(\mathbf{a}_\pi^{\overline{\pi}}\big)}{2}. \tag{P.6}$$

The term α^* is the value of α that maximizes the above bound, i.e., the value that sets the partial derivative w.r.t. α to zero, as the bound is a quadratic function. By putting α^* in place of α in the last bound we derive the guaranteed performance improvement. ■

The pseudocode of USPI is reported in Algorithm 1. The algorithm takes as input the MDP $\mathcal{M}$, the target policy space $\overline{\Pi} \subseteq \Pi^{\mathrm{SR}}$ and a policy chooser PC. The target policy space $\overline{\Pi}$ is a (finite) set of policies from which the target policy $\overline{\pi}$ is selected. A standard choice for $\overline{\Pi}$ is the set of all deterministic policies, i.e., $\overline{\Pi} = \Pi^{\mathrm{SD}}$. The policy chooser PC is a function that takes as input the MDP $\mathcal{M}$, the target policy space $\overline{\Pi}$ and the current policy π and provides as output a target policy. We will discuss in the following possible implementations of PC. Thus, the goal of SPI algorithms is to terminate with a policy π such that for all $\overline{\pi} \in \overline{\Pi}$, $\mathbb{A}_\pi^{\overline{\pi}} \leq 0$.

From Corollary 4 and Corollary 5 it is straightforward to introduce a simplified USPI.

Algorithm 1 Exact USPI.

input: MDP $\mathcal{M}$, target policy space $\overline{\Pi}$, policy chooser PC
Initialize π
$\overline{\pi} \leftarrow \mathrm{PC}(\mathcal{M}, \overline{\Pi}, \pi)$
while $\mathbb{A}_\pi^{\overline{\pi}} > 0$ **do**
  $\alpha \leftarrow \min\left\{1, \frac{(1-\gamma)\mathbb{A}_\pi^{\overline{\pi}}}{\gamma \|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi} \operatorname{sp}(\mathbf{a}_\pi^{\overline{\pi}})}\right\}$
  $\boldsymbol{\pi} \leftarrow \alpha\overline{\boldsymbol{\pi}} + (1-\alpha)\boldsymbol{\pi}$
  $\overline{\pi} \leftarrow \mathrm{PC}(\mathcal{M}, \overline{\Pi}, \pi)$
end while
return $\boldsymbol{\pi}$

Corollary 6. *If $\mathbb{A}_\pi^{\overline{\pi}} \geq 0$, then, using $\alpha^* = \frac{(1-\gamma)\mathbb{A}_\pi^{\overline{\pi}}}{\gamma \|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty^2 \|\mathbf{q}^\pi\|_\infty}$, we set $\alpha = \min\{1, \alpha^*\}$, so that when $\alpha^* \leq 1$ we can guarantee the following policy improvement:*

$$J^{\pi'} - J^\pi \geq \frac{(\mathbb{A}_\pi^{\overline{\pi}})^2}{2\gamma \|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty^2 \|\mathbf{q}^\pi\|_\infty}$$

and when $\alpha^* > 1$, we perform a full update towards the target policy $\overline{\pi}$, i.e., we set $\alpha = 1$ so that $\pi' = \overline{\pi}$. In such a case, the policy improvement is given by Corollary 4 by setting $\pi' = \overline{\pi}$.

Remark 2 (Comparison with Conservative Policy Iteration)**.** Using the notation introduced in this paper, we report the bound proposed in Conservative Policy Iteration CPI (refer to Theorem 4.1 in Kakade and Langford, 2002 or Corollary 7.2.2 in Kakade, 2003) to be compared with SPI (Equation P.6):

$$J^{\pi'} - J^\pi \geq \alpha \frac{1}{1-\gamma} \mathbb{A}_\pi^{\overline{\pi}} - \alpha^2 \frac{2\gamma}{(1-\gamma)(1-\gamma(1-\alpha))} \left\|\mathbf{a}_\pi^{\overline{\pi}}\right\|_\infty .$$

Since $\left\|\mathbf{a}_\pi^{\overline{\pi}}\right\|_\infty$ is unknown, the bound that is employed by the algorithm is obtained by observing that $\left\|\mathbf{a}_\pi^{\overline{\pi}}\right\|_\infty \leq 1/(1-\gamma)$ and $0 \leq \alpha \leq 1$:[4]

$$J^{\pi'} - J^\pi \geq \alpha \frac{\mathbb{A}_\pi^{\overline{\pi}}}{1-\gamma} - \alpha^2 \frac{2\gamma}{(1-\gamma)^3},$$

which yields the optimal coefficient $\alpha^* = \frac{(1-\gamma)^2 \mathbb{A}_\pi^{\overline{\pi}}}{4\gamma}$ and performance improvement is given by (refer to Corollary 4.2 in Kakade and Langford, 2002 or Corollary 7.2.3 in Kakade, 2003):

$$J^{\pi'} - J^\pi \geq \frac{(1-\gamma)(\mathbb{A}_\pi^{\overline{\pi}})^2}{8\gamma}. \tag{2}$$

The only difference between such bound and the one of USPI (see Corollary 5) is in the denominator. Since $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi} \operatorname{sp}\big(\mathbf{a}_\pi^{\overline{\pi}}\big) \leq \frac{4}{1-\gamma}$, the improvement guaranteed by USPI is

4. In Kakade and Langford (2002) and Kakade (2003), the bound that is actually optimized is a slightly relaxed version in which the γ term at the numerator is bounded with 1.

Algorithm 2 Greedy Policy Chooser (GPC).

input: MDP $\mathcal{M}$, current policy $\boldsymbol{\pi}$, target policy space $\overline{\Pi}$
for each $\pi^\dagger \in \overline{\Pi}$ **do**
  Compute $\mathbb{A}_\pi^{\pi^\dagger}$
end for
return $\arg\max_{\pi^\dagger \in \overline{\Pi}} \mathbb{A}_\pi^{\pi^\dagger}$

not worse than the one of CPI. From the tightness of CPI bound, it follows that also USPI bound is tight. In general, the difference between the two approaches can be much larger whenever π is not completely different from $\overline{\pi}$ (i.e., $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi} < 2$) and/or the values of the advantage function are not spread from the theoretical minimum to the theoretical maximum (i.e., $\mathrm{sp}\big(\mathbf{a}_\pi^{\overline{\pi}}\big) < \frac{2}{1-\gamma}$). In particular, using policy iteration algorithms without approximation, where $\overline{\pi}$ is the greedy policy π^+, as the sequence of policies approaches the optimal policy, the discrepancy between the current policy π and the greedy policy π^+ decreases and so happens for the advantage values $A_\pi^{\pi^+}$, thus allowing USPI to guarantee much larger improvements than CPI (whose convergence is only asymptotic, being its coefficient $\alpha = \frac{(1-\gamma)^2 \mathbb{A}_\pi^{\pi^+}}{4\gamma}$ always less than 1).

Remark 3 (Target Policy Selection)**.** So far we have not specified how to select the target policy $\overline{\pi} \in \overline{\Pi}$. The Greedy Policy Chooser (GPC, Algorithm 2) selects, at each iteration, as target policy the greedy policy π^+, i.e., the one that maximizes $\mathbb{A}_\pi^{\pi^+}$. While π^+ is the best target for CPI, it might not be optimal for USPI. This consideration comes from the analysis of the policy performance bounds. While the greedy policy maximizes the bound of CPI (Equation 2), as π^+ is the target policy that yields the maximum advantage, π^+ may not be optimal for the bound of USPI due to the penalization term $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi}$. Indeed, the USPI bound trades off between the expected advantage and the distance between the target and current policy. In practice, when the approximation of Q^π is involved, the GPC might produce frequent switching among several target policies that might slow down the algorithm since the distance term $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi}$ remains high.

As a heuristic, we can employ a *persistent* version of the GPC, similarly to what was proposed in Metelli et al. (2018) (Section 4.3). This new policy chooser takes as input also the target policy at the previous iteration $\overline{\boldsymbol{\pi}}$ and selects between the greedy policy and the target policy at the previous iteration, the one that yields a higher performance improvement.

4.1.1 Convergence Guarantees for USPI

In this section, we discuss the convergence properties of USPI. The issue of convergence has been treated for CPI (Kakade and Langford, 2002) and USPI (Pirotta et al., 2013b) when considering a stopping condition of the form $\mathbb{A}_\pi^{\overline{\pi}} \leq \frac{\kappa}{1-\gamma}$, where $\kappa > 0$ is a user–defined threshold. In this case, both CPI and USPI terminate in $O\left(\frac{1}{\kappa^2}\right)$ iterations. It was also proved that when following a fixed target policy, USPI improves the convergence rate of CPI, being able to terminate in $O\left(\frac{1}{\kappa}\right)$ iterations (Pirotta et al., 2013b).

Our contribution to the convergence analysis consists in analyzing the case $\kappa = 0$, as in Algorithm 1. Of course, in this case, the convergence guarantees of CPI are vacuous. This section is organized as follows. We start by proving that USPI (and CPI) converges asymptotically under some assumptions on the γ–discounted state distribution (Assumption 1). Then, we show that, when the optimal policy is unique (Assumption 2), USPI converges to the optimal policy in a finite number of iterations (Theorem 11). The proofs of all the presented results are reported in Appendix A.2.

We start with the following lemma, which extends the Corollary 4.5 in Kakade and Langford (2002), and relates the expected advantage to the performance difference.

Lemma 7. *Let $\pi, \pi' \in \Pi^{SR}$ be two arbitrary policies and π^+ be a greedy policy induced by Q^π. Then, the expected advantage $\mathbb{A}_\pi^{\pi^+}$ can be lower bounded as:*

$$\frac{\mathbb{A}_\pi^{\pi^+}}{1-\gamma} \geq \left\| \frac{\mathbf{d}^{\pi'}}{\mathbf{d}^\pi} \right\|_\infty^{-1} \left(J^{\pi'} - J^\pi \right), \tag{3}$$

where $\frac{\mathbf{d}^{\pi'}}{\mathbf{d}^\pi}$ is the vector obtained by the element–wise division between $\mathbf{d}^{\pi'}$ and $\mathbf{d}^\pi$.

This result is very general since policy π' is chosen arbitrarily. Clearly, the bound is meaningful when $J^{\pi'} > J^\pi$ as we know that $\mathbb{A}_\pi^{\pi^+} \geq 0$. A straightforward choice of π' is, of course, a greedy policy π^+. In this case, we are able to lower bound the expected advantage function $\mathbb{A}_\pi^{\pi^+}$ in terms of the performance gap itself:

$$\frac{\mathbb{A}_\pi^{\pi^+}}{1-\gamma} \geq \left\| \frac{\mathbf{d}^{\pi^+}}{\mathbf{d}^\pi} \right\|_\infty^{-1} \left(J^{\pi^+} - J^\pi \right).$$

Remark 4 (On the γ–discounted stationary distribution)**.** How can we ensure that $\left\| \frac{\mathbf{d}^{\pi'}}{\mathbf{d}^\pi} \right\|_\infty < +\infty$ is satisfied? A sufficient condition for $\left\| \frac{\mathbf{d}^{\pi'}}{\mathbf{d}^\pi} \right\|_\infty < +\infty$ is that for any policy π and for any state s, we have $d^\pi(s) > 0$. In particular, if the distribution of the initial state is positive $\mu(s) > 0$ for all states $s \in \mathcal{S}$ the condition is satisfied, indeed $\mathbf{d}^\pi = (1-\gamma)\boldsymbol{\mu} + \gamma \mathbf{P}^{\pi\mathrm{T}} \mathbf{d}^\pi \geq (1-\gamma)$. When $d^\pi(s) > 0$ for every s and policy π, it admits for every state s a positive minimum over the set of Markovian stationary policies. This is a consequence of the fact that $\mathbf{d}^\pi$ is a continuous function w.r.t. the policy $\boldsymbol{\pi}$ (Corollary 1 provides the Lipschitz continuity of $\mathbf{d}^\pi$ w.r.t. $\boldsymbol{\pi}$) and the set of Markovian stationary policies Π^{SR} is compact. Moreover, if we consider finite state spaces $d^\pi(s)$ admits a positive minimum also over the state space, that we will denote with Δ_d. Therefore, under this assumption, we can provide the bound: $\left\| \frac{\mathbf{d}^{\pi'}}{\mathbf{d}^\pi} \right\|_\infty \leq \frac{1}{\Delta_d}$. From now on, we are going to make the following assumption.

Assumption 1. *For all $\pi \in \Pi^{SR}$ it holds that $\Delta_d > 0$, where $\Delta_d = \min_{s \in \mathcal{S}} \{d^\pi(s)\}$.*

Assumption 1 requires that each state is visited a (discounted) number of times at least equal to $\Delta_d > 0$. A sufficient condition is that the initial state distribution μ visits with non–zero probability each state of the MDP. In such case, $\Delta_d \geq (1-\gamma) \min_{s \in \mathcal{S}} \{\mu(s)\}$.

Theorem 8 (Convergence Bound)**.** *Under Assumption 1, USPI (and CPI) with GPC and termination condition $\mathbb{A}_{\pi}^{\pi^+} \leq 0$ asymptotically converges to the optimal policy, i.e., let $N > 0$ and π_N be the policy visited by USPI (or CPI) at iteration $N > 0$, it holds that:*

$$J^* - J^{\pi_N} \leq \frac{8\gamma}{N\Delta_d^2(1-\gamma)^3}.$$

Proof Let us consider a step of USPI starting from policy π_i and getting to policy π_{i+1}. For Corollary 6, by noticing that $\|\mathbf{q}^\pi\|_\infty \leq 1/(1-\gamma)$ for any $\pi \in \Pi^{\mathrm{SR}}$, the performance improvement is given by:

$$J^{\pi_{i+1}} - J^{\pi_i} \geq \frac{(1-\gamma)\left(\mathbb{A}_{\pi_i}^{\pi_i^+}\right)^2}{2\gamma\left\|\boldsymbol{\pi}_i^+ - \boldsymbol{\pi}_i\right\|_\infty^2} \geq \frac{(1-\gamma)\left(\mathbb{A}_{\pi_i}^{\pi_i^+}\right)^2}{8\gamma}, \tag{P.7}$$

where π_i^+ is a deterministic greedy policy. We now define the performance gap w.r.t. the optimal policy $\Delta_i = J^* - J^{\pi_i}$. By changing the sign on both sides of Equation (P.7), summing J_μ^*, and recalling the definition of Δ_i we get the following inequality:

$$J^* - J^{\pi_{i+1}} \leq J^* - J^{\pi_i} - \frac{(1-\gamma)\left(\mathbb{A}_{\pi_i}^{\pi_i^+}\right)^2}{8\gamma}$$
$$\Delta_{i+1} \leq \Delta_i - \frac{(1-\gamma)\left(\mathbb{A}_{\pi_i}^{\pi_i^+}\right)^2}{8\gamma}.$$

We now determine the convergence bound. Using Lemma 7 choosing $\pi' = \pi^*$ we can lower bound the expected advantage function $\mathbb{A}_{\pi_i}^{\pi_i^+}$ and write:

$$\Delta_{i+1} \leq \Delta_i - \left\|\frac{\mathbf{d}^{\pi_i^+}}{\mathbf{d}^{\pi_i}}\right\|_\infty^{-2} \frac{(1-\gamma)^3.(\Delta_i)^2}{8\gamma}$$

Let us now consider the following expression:

$$\frac{1}{\Delta_{i+1}} - \frac{1}{\Delta_i} = \frac{\Delta_i - \Delta_{i+1}}{\Delta_{i+1}\Delta_i} \geq \frac{\Delta_i - \Delta_{i+1}}{\Delta_i^2} \geq \left\|\frac{\mathbf{d}^{\pi_i^+}}{\mathbf{d}^{\pi_i}}\right\|_\infty^{-2} \frac{(1-\gamma)^3}{8\gamma},$$

where we simply exploited the monotonicity property of Δ_i due to the guaranteed performance improvement. We can now sum over i and exploiting the telescopic property we get:

$$\frac{1}{\Delta_N} \geq \frac{1}{\Delta_0} + \frac{(1-\gamma)^3}{8\gamma}\sum_{i=0}^{N-1}\left\|\frac{\mathbf{d}^{\pi_i^+}}{\mathbf{d}^{\pi_i}}\right\|_\infty^{-2} \geq N\frac{(1-\gamma)^3}{8\gamma}\min_{i\in\{0,1,\dots,N-1\}}\left\|\frac{\mathbf{d}^{\pi_i^+}}{\mathbf{d}^{\pi_i}}\right\|_\infty^{-2}.$$

Solving for Δ_N we have:

$$\Delta_N = J^* - J^{\pi_N} \leq \frac{8\gamma}{N(1-\gamma)^3}\max_{i\in\{0,1,\dots,N-1\}}\left\|\frac{\mathbf{d}^{\pi_i^+}}{\mathbf{d}^{\pi_i}}\right\|_\infty^{2} \leq \frac{8\gamma}{N\Delta_d^2(1-\gamma)^3}.$$

Thus, USPI converges asymptotically with convergence bound of order $O(N^{-1})$. Notice that in the Equation (P.7) we upper-bounded the policy distance with 2 and thus, we considered the same setting as CPI. As a consequence, this result applies as is to CPI. ■

Remark 5 (On the Convergence Bound)**.** The convergence bound we derived in Theorem 8 has a polynomial dependence on the iteration number N, i.e., $O\left(\frac{1}{N(1-\gamma)^3}\right)$. This appears to be suboptimal compared to PI and VI both having a convergence bound that depends exponentially on N, i.e., $O\left(\frac{\gamma^N}{1-\gamma}\right)$ (Puterman, 2014). In addition, also CPI with a constant learning rate $\alpha \in [0, 1]$ achieves an exponential convergence $O\left(\frac{(1-\alpha-\gamma\alpha)^N}{1-\gamma}\right)$ (Scherrer, 2014). It is not surprising that these algorithms allow for a better convergence bound. Indeed, in the *exact setting*, having access to the true greedy policy, we can safely perform a complete improvement step, i.e., setting $\alpha = 1$. CPI and SPI are meant to be employed in the *approximate setting* (Section 5) when only an approximately greedy policy is available and, consequently, we cannot fully trust it.

We now prove that USPI converges to the optimal policy in a finite number of steps when using GPC. We outline the steps of the proof. First, we need to guarantee that after a finite number of steps, USPI selects an optimal policy as target policy (Lemma 9). This follows by observing that the performance difference between an optimal policy and the second-best deterministic policy is finite and by applying Theorem 8 to bound the number of steps. Second, we need to ensure that when selecting an optimal policy as target, USPI converges to it in a finite number of steps. This is the most delicate part of the proof, as the finite convergence is a consequence of the interaction between the expected advantage $\mathbb{A}_{\pi}^{\pi^*}$ and the distance $\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty$. It must happen that the distance decreases at least as fast as the advantage when $\boldsymbol{\pi} \to \boldsymbol{\pi}^*$ (Lemma 10). With GPC, this can be guaranteed only in the presence of a unique (deterministic) optimal policy. Therefore, in the presence of multiple optimal policies, switching between one and another might prevent finite convergence. We are unable to guarantee that, in the presence of multiple optimal policies, GPC keeps selecting the same optimal policy; thus, we restrict our attention to the case in which the optimal policy is unique. In the following, we denote with Π^* for the set of optimal policies.

Lemma 9. *Assume the same setting as Theorem 8. Let $\Delta_J = J^* - \max_{\pi \in \Pi^{SD} \setminus \Pi^*} \{J^\pi\}$ be the performance gap between the optimal policies and the second-best deterministic policy, where $\Pi^* = \{\pi \in \Pi^{SD} : J^\pi = J^*\}$. Then, USPI (and CPI) with GPC selects an optimal policy as target policy after a finite number of iterations.*

Clearly, once we select an optimal policy as target policy, we will never select a subpotimal policy as target later as it could only decrease the performance. We can now prove that when following a deterministic optimal policy as target policy, the expected advantage $\mathbb{A}_{\pi}^{\pi^*}$ can be lower bounded by a function of the distance $\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty$.

Lemma 10. *Assume the same setting as Lemma 9. It π^* is a deterministic optimal policy, then, there exists a constant $\Delta_+ > 0$ such that:*

$$\mathbb{A}_{\pi}^{\pi^*} \geq \frac{\Delta_d \Delta_+}{2} \|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty . \tag{4}$$

So far, we did not exploit the assumption on the uniqueness of the optimal policy. In the following theorem, the assumption is crucial.

Assumption 2. *The optimal policy π^* is unique.*

Lemma 10 shows that, apart from constants, the expected advantage $\mathbb{A}_{\pi}^{\pi^*}$ decreases at most as fast as the distance $\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty$. We can exploit this result, together with the uniqueness of the optimal policy, to prove that USPI converges in a finite number of iterations.

Theorem 11 (Finite Convergence)**.** *Under Assumption 1 and 2, USPI with GPC and termination condition $\mathbb{A}_{\pi}^{\pi^+} \leq 0$ converges to the optimal policy in a finite number of iterations.*

Proof First of all, we know from Lemma 9 that the algorithm will select the (unique) optimal policy π^* as target policy after a finite number of iterations, say N_1. Thus, for $i > N_1$, we have that $J^{\pi_i} \geq J^* - \Delta_J$ and moreover:

$$J^{\pi_{i+1}} - J^{\pi_i} \geq \frac{(1-\gamma)\Big(\mathbb{A}_{\pi_i}^{\pi^*}\Big)^2}{2\gamma\big\|\boldsymbol{\pi}^* - \boldsymbol{\pi}_i\big\|_\infty^2} \geq \frac{(1-\gamma)\Delta_d^2\Delta_+^2\big\|\boldsymbol{\pi}^* - \boldsymbol{\pi}_i\big\|_\infty^2}{8\gamma\big\|\boldsymbol{\pi}^* - \boldsymbol{\pi}_i\big\|_\infty^2} = \frac{(1-\gamma)\Delta_d^2\Delta_+^2}{8\gamma},$$

where the first inequality is obtained from Equation (P.7) and the second inequality from Lemma 10. Since at each iteration the performance improves by a finite quantity, the algorithm will need additional N_2 iterations to fill the gap Δ_J between the performance of the policy π_{N_1} and the performance of the optimal policy π^*:

$$N_2\frac{(1-\gamma)\Delta_d^2\Delta_+^2}{8\gamma} \geq \Delta_J \implies N_2 \geq \frac{8\gamma\Delta_J}{\Delta_d^2\Delta_+^2(1-\gamma)}.$$

Consequently, the algorithm will converge in $N_1 + N_2$ iterations. ■

4.2 per–State–parameter Safe Policy Improvement

The USPI approach aims at finding the convex combination between a starting policy π and a target policy $\overline{\pi}$ that maximizes the bound on the performance improvement (either Theorem 3 or Corollary 4). In this section, we consider a more general kind of update, where the new policy π' is generated using different convex combination coefficients for each state:

$$\pi'(a|s) = \alpha(s)\overline{\pi}(a|s) + \big(1-\alpha(s)\big)\pi(a|s), \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \tag{5}$$

where $\alpha(s) \in [0,1], \forall s \in \mathcal{S}$. We name the resulting algorithm as *per–State–parameter Safe Policy Iteration* (SSPI).[5] When per-state parameters are exploited, the bound in Theorem 3 requires solving two dependent maximization problems over the state space that do not admit a simple solution. Therefore, to compute the values $\alpha(s)$, we consider the simplified bound from Corollary 4. We can state the following result.

5. SSPI was called Multiple–parameter Safe Policy Improvement (MSPI) in Pirotta et al. (2013b). The reason for the change of the name is due to the fact that we will present another approach exploiting multiple parameters (Section 4.3).

Corollary 12. *Let $\mathcal{S}_\pi^{\overline{\pi}}$ be the subset of states where the advantage of policy $\overline{\pi}$ over policy π and d^π are positive: $\mathcal{S}_\pi^{\overline{\pi}} = \{s \in \mathcal{S} : d^\pi(s) A_\pi^{\overline{\pi}}(s) > 0\}$. The bound in Corollary 4 is optimized by taking $\alpha(s) = 0$, $\forall s \notin \mathcal{S}_\pi^{\overline{\pi}}$ and $\alpha(s) = \min\left\{1, \frac{\Upsilon^*}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right\}$, $\forall s \in \mathcal{S}_\pi^{\overline{\pi}}$, where $\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 = \sum_{a \in \mathcal{A}} |\overline{\pi}(a|s) - \pi(a|s)|$ and Υ^* is the value that maximizes the following function:*

$$B(\Upsilon) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \min\left\{1, \frac{\Upsilon}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right\} d^\pi(s) A_\pi^{\overline{\pi}}(s) - \Upsilon^2 \frac{\gamma}{(1-\gamma)^2} \frac{\|\mathbf{q}^\pi\|_\infty}{2}.$$

Proof This proof starts from the transformation of several terms involved in Corollary 4 exploiting the definition of π' (see Equation 5). The average advantage $\mathbb{A}_\pi^{\overline{\pi}}$ can be stated as follows for all $s \in \mathcal{S}$:

$$\begin{aligned} d^\pi(s) A_\pi^{\pi'}(s) &= d^\pi(s) \sum_{a \in \mathcal{A}} \big(\pi'(a|s) - \pi(a|s)\big) Q^\pi(s,a) \\ &= d^\pi(s) \sum_{a \in \mathcal{A}} \alpha(s) \big(\overline{\pi}(a|s) - \pi(a|s)\big) Q^\pi(s,a) \\ &= \alpha(s) d^\pi(s) A_\pi^{\overline{\pi}}(s). \end{aligned}$$

Exploiting the definition of L_∞–norm of a matrix, we can write:

$$\left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_\infty = \sup_{s \in \mathcal{S}} \left\{ \sum_{a \in \mathcal{A}} \left| \pi'(a|s) - \pi(a|s) \right| \right\} = \sup_{s \in \mathcal{S}} \left\{ \alpha(s) \left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1 \right\}.$$

We can now restate the bound of Corollary 4 into the proposed framework:

$$\begin{aligned} J^{\pi'} - J^\pi \geq \frac{1}{1-\gamma} &\sum_{s \in \mathcal{S}} \alpha(s) d^\pi(s) A_\pi^{\overline{\pi}}(s) \\ &- \frac{\gamma}{(1-\gamma)^2} \sup_{s \in \mathcal{S}} \left\{ \alpha(s) \left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1 \right\}^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2}. \end{aligned} \tag{P.8}$$

The optimal values of $\alpha(s)$ do not admit a closed–form solution but can be computed iteratively. Given a state s with negative advantage $A_\pi^{\overline{\pi}}$, the larger $\alpha(s)$ is, the lower will be the bound on the policy improvement as expressed in Equation (P.8) (or in Corollary 4), so the optimal choice for these states is to set $\alpha(s) = 0$. Similarly, if $d^\pi(s) = 0$, state s does not have any contribution to the bound, so we can set $\alpha(s) = 0$.

Given these conditions, we define $\mathcal{S}_\pi^{\overline{\pi}} = \{s \in \mathcal{S} : d^\pi(s) A_\pi^{\overline{\pi}}(s) > 0\}$. Then, Υ denotes the L_∞–norm of the difference of the policies over $\mathcal{S}_\pi^{\overline{\pi}}$:

$$\Upsilon = \sup_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \left\{ \alpha(s) \left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1 \right\},$$

we can now introduce the following condition:

$$\alpha(s) \left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1 \leq \Upsilon, \quad \forall s \in \mathcal{S}_\pi^{\overline{\pi}}. \tag{P.9}$$

Algorithm 3 Exact SSPI.

input: MDP $\mathcal{M}$, target policy space $\overline{\Pi}$, policy chooser PC
Initialize π
$\overline{\pi} \leftarrow \mathrm{PC}(\mathcal{M}, \overline{\Pi}, \pi)$
$\Upsilon^* \leftarrow \mathrm{FBO}(\mathcal{M}, \pi, \overline{\pi})$ (see Alg. 4)
while $\Upsilon^* > 0$ **do**
  $\alpha(s) \leftarrow \min\left\{1, \frac{\Upsilon^*}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right\},\ \forall s \in \mathcal{S}$
  $\pi(a|s) \leftarrow \alpha(s)\overline{\pi}(a|s) + (1-\alpha(s))\,\pi(a|s),\ \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$
  $\overline{\pi} \leftarrow \mathrm{PC}(\mathcal{M}, \overline{\Pi}, \pi)$
  $\Upsilon^* \leftarrow \mathrm{FBO}(\mathcal{M}, \pi, \overline{\pi})$
end while

If we suppose to fix Υ, the previous relationship and the knowledge of $\alpha \in [0,1]$ impose the following equivalence:

$$\alpha(s) = \min\left\{1, \frac{\Upsilon}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right\}, \quad \forall s \in \mathcal{S}_\pi^{\overline{\pi}}.$$

Consider Equation (P.8), once the supremum is fixed, we cannot do better than set the other coefficients $\alpha(s)$ to the maximum feasible value that does not make $\alpha(s)\,\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1$ exceed the supremum. As mentioned before, states with negative advantage play as opponents, their influence is minimized by putting $\alpha(s) = 0$ in correspondence of such states.

Function $B(\Upsilon)$ is obtained by manipulating the bound (P.8) using these considerations. As a result, the optimization of the bound over the set of $|\mathcal{S}|$ coefficients $\alpha(s)$ has been translated into the maximization of the univariate function $B(\Upsilon)$. However, since the superior Υ is not known a priori, an iterative approach has to be carried out. Once the optimal value Υ^* is obtained, the following rule can be applied:

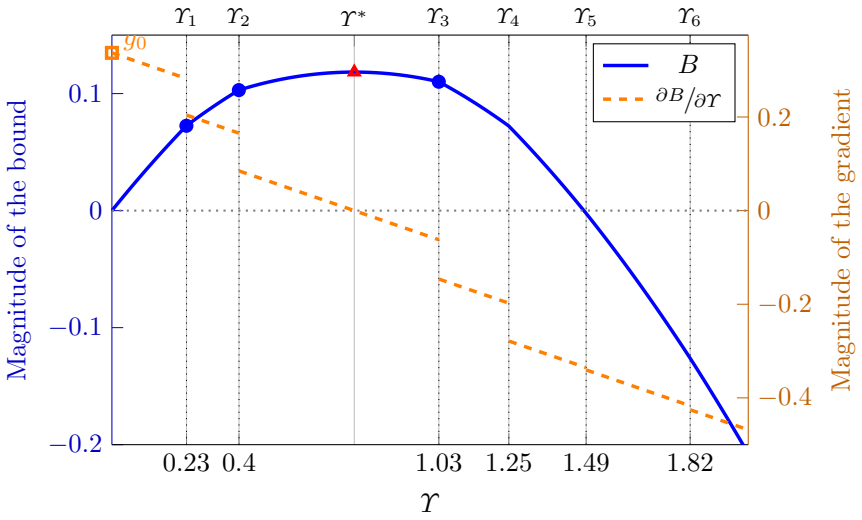
$$\alpha(s) = \begin{cases} \min\left\{1, \frac{\Upsilon^*}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}\right\} & \forall s \in \mathcal{S}_\pi^{\overline{\pi}} \\ 0 & \text{otherwise} \end{cases}.$$

■

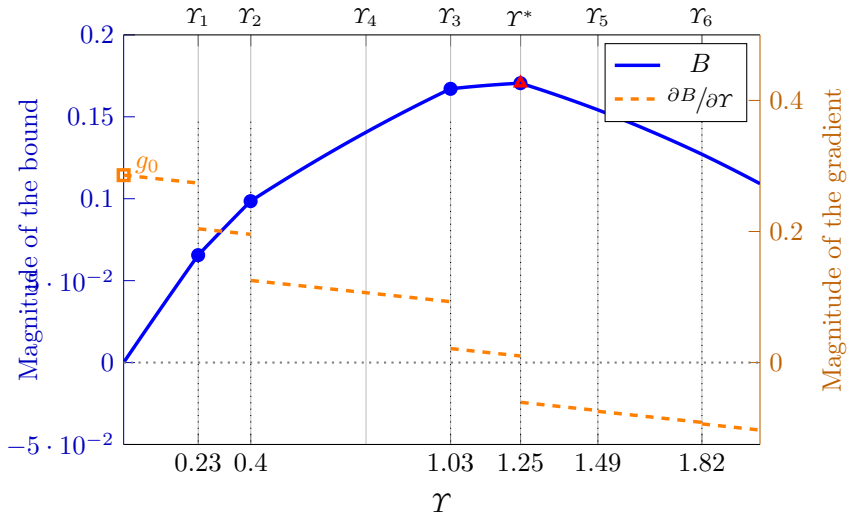
The pseudocode of SSPI is reported in Algorithm 3. The algorithm stops as the optimal budget Υ^* becomes zero, i.e., when in all states the advantage $A_\pi^{\overline{\pi}}(s) \leq 0$.

4.2.1 Computing Υ^*

Differently from USPI, the coefficients of SSPI cannot be computed in closed form due to their dependency from Υ^*, whose value requires the maximization of a function with discontinuous derivative (Corollary 12). This search formalizes the trade-off between increasing the probability budget Υ, and incur in a larger penalty while obtaining a gain by moving further towards the target policy. In order to solve this problem, we consider the graph in Figure 1, where we can see that function B is a continuous quadratic piecewise function, whose derivative is a discontinuous linear piecewise function. It is important to underline that all the pieces of the partial derivative of B have the same slope $m = -\frac{\gamma\|\mathbf{q}^\pi\|_\infty}{(1-\gamma)^2}$. Suppose



(a)



(b)

Figure 1: Bound B and its derivative. Blue–filled circles are set in correspondence with the discontinuities, whereas the blue triangle represents the maximum value of B. The gradient of B is depicted by the dashed brown piecewise linear function where the red square represents g_0, its evaluation in $\Upsilon = 0$.

Algorithm 4 Computing Υ^* (Forward Bound Optimizer - FBO)

input: MDP $\mathcal{M}$, current policy π, target policy $\overline{\pi}$
initialize: $i \leftarrow 1$, $m \leftarrow -\frac{\gamma \|\mathbf{q}^{\pi}\|_{\infty}}{(1-\gamma)^2}$ $\Upsilon_0, q_0 \leftarrow \text{FJP}(\mathcal{M}, \overline{\pi}, \pi)$, $g_0 \leftarrow q_0$
while $\Upsilon_i < 2$ **do**
  $\Upsilon_i, q_i \leftarrow \text{FJP}(\mathcal{M}, \overline{\pi}, \pi)$ (see Alg. 5)
  $g_i \leftarrow g_{i-1} + m \cdot (\Upsilon_i - \Upsilon_{i-1})$
  if $g_i \leq 0$ **then**
    return $\Upsilon_i - \frac{g_i}{m}$
  end if
  $g_i \leftarrow g_i - q_{i-1} + q_i$
  if $g_i \leq 0$ **then**
    return Υ_i
  end if
  $i \leftarrow i + 1$
end while
return Υ_i

we are given a function FJP (Find Jump Point) that returns the coordinates (Υ, q) of the next discontinuity point of the derivative of B. Then, the maximization of B can be computed using an iterative algorithm like the one proposed in Algorithm 4, Forward Bound Optimizer (FBO).[6] The idea is to start from $\Upsilon = 0$ and to search for the zero–crossing value of the derivative of B by running over the discontinuity points. The algorithm stops when either the derivative of B becomes negative or when we reach the maximum value of Υ, i.e., $\Upsilon = 2$ (the last return in Algorithm 4). When the derivative becomes negative, two different cases may happen: (i) the derivative equals zero at some value of Υ (as it happens in Figure 1a), which is the case of the first return in Algorithm 4; (ii) the derivative becomes negative in correspondence of a discontinuity without taking the value of zero (the second return in Algorithm 4), i.e., the maximum falls on an angular point of B (see Figure 1b). Notice that, as we presented it, Algorithm 4 can be used to optimize any function with linear piecewise derivative, provided that all pieces have the same slope.

Clearly, we need to be able to determine the discontinuity points of the derivative, i.e., we need to specify function FJP. For this purpose, we write down explicitly the derivative:

$$\frac{\partial}{\partial \Upsilon} B(\Upsilon) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_{\Upsilon}} \frac{d^{\pi}(s) A^{\overline{\pi}}_{\pi}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} - \Upsilon \frac{\gamma \|\mathbf{q}^{\pi}\|_{\infty}}{(1-\gamma)^2} = g(\Upsilon) + m\Upsilon, \tag{6}$$

where $\mathcal{S}_{\Upsilon} = \left\{ s \in \mathcal{S}^{\overline{\pi}}_{\pi} : \|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1 > \Upsilon \right\}$ is the set of all the states in which the coefficient $\alpha(s)$ is dependent on Υ. Since the derivative is non-negative at $\Upsilon = 0$, and it is monotonically decreasing, B is guaranteed to have a unique maximum. The discontinuity points correspond to values of Υ for which some state $\overline{s}$ saturates its coefficient to 1, so that, for larger values Υ, the coefficient $\alpha(\overline{s})$ does not depend on Υ anymore, thus disappearing from the derivative whose value changes discontinuously with a jump equal to $\frac{d^{\pi}(\overline{s}) A^{\overline{\pi}}_{\pi}(\overline{s})}{(1-\gamma)\|\overline{\pi}(\cdot|\overline{s}) - \pi(\cdot|\overline{s})\|_1}$. The procedure for finding the discontinuity points (FJP) is formalized in Algorithm 5.

6. Differently from Pirotta et al. (2013b), we decided to keep the optimization of the bound (FBO) and the identification of the discontinuity points (FJP) separated so that we can reuse FBO in Section 4.3.

Algorithm 5 Computing of the jump points for SSPI (Find Jump Point - FJP)

input: MDP $\mathcal{M}$, current policy π, target policy $\overline{\pi}$
initialize: $t \leftarrow 0$, $\mathcal{S}_\pi^{\overline{\pi}} \leftarrow \{s \in \mathcal{S} : d^\pi(s) A_\pi^{\overline{\pi}}(s) > 0\}$, $\Upsilon_0 \leftarrow 0$, $q_0 \leftarrow \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \frac{d^\pi(s) A_\pi^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}$
Sort states in $\mathcal{S}_\pi^{\overline{\pi}}$ so that $i < j \implies \|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1 \leq \|\overline{\pi}(\cdot|s_j) - \pi(\cdot|s_j)\|_1$
yield Υ_0, q_0
while $\mathcal{S}_\pi^{\overline{\pi}} \neq \{\}$ **do**
  $t \leftarrow t + 1$
  $\Upsilon_t \leftarrow \|\overline{\pi}(\cdot|s_t) - \pi(\cdot|s_t)\|_1$
  $q_t \leftarrow q_{t-1} - \frac{d^\pi(s_t) A_\pi^{\overline{\pi}}(s_t)}{(1-\gamma)\|\overline{\pi}(\cdot|s_t) - \pi(\cdot|s_t)\|_1}$
  $\mathcal{S}_\pi^{\overline{\pi}} \leftarrow \mathcal{S}_\pi^{\overline{\pi}} \setminus \{s_t\}$
  yield Υ_t, q_t
end while
yield 2, $-\infty$

The computational complexity of FJP is dominated by the cost of computing the L^1-norm between the policies and the cost of sorting the states according to the discrepancy between the current policy π and the target policy $\overline{\pi}$, that is $O\left(|\mathcal{S}||\mathcal{A}| + |\mathcal{S}| \log |\mathcal{S}|\right)$.

Remark 6 (On the policy space of SSPI)**.** When using per–state coefficients $\alpha(s)$ the space of policies accessible, by combining the target policies $\overline{\Pi}$, is larger than that obtainable with a single coefficient α. Clearly, it is possible to enhance $\overline{\Pi}$ with additional policies so that even USPI, with a unique coefficient α, can represent the same policies as SSPI. However, this transformation would produce, in the worst case, exponential growth in the number of target policies. Consider, for instance, the set of target policies $\overline{\Pi} = \{\pi_i(s) = a_i : \forall s \in \mathcal{S}, i \in \{1, ..., |\mathcal{A}|\}\}$. Thus, $\overline{\Pi}$ contains all the deterministic policies that perform the same action in all states. Consequently, $|\overline{\Pi}| = |\mathcal{A}|$. Using per–state coefficients $\alpha(s)$, we are able to represent all Markovian randomized policies Π^{SR}. Those are the policies accessible by SSPI. Instead, starting with $\overline{\Pi}$, USPI can represent just a subset of those. For USPI to represent all Markovian randomized policies, we need to consider as target policy space the set of all Markovian deterministic policies Π^{SD}, whose cardinality is $|\mathcal{A}|^{|\mathcal{S}|}$. We can generalize the rationale to all the target policy spaces made up of deterministic policies. Let $\mathcal{A}_{\overline{\Pi}}(s) = \{a \in \mathcal{A} : \exists \pi \in \overline{\Pi}, \pi(s) = a\}$ be the set of all actions that are prescribed in state s by the policies in $\overline{\Pi}$. The transformation of the policy space is obtained as follows:

$$\widetilde{\Pi} = \left\{\pi(s) = a : \forall a \in \mathcal{A}_{\overline{\Pi}}(s), \forall s \in \mathcal{S}\right\} \tag{7}$$

It is worth noting that the cardinality of $\widetilde{\Pi}$ is given by $\prod_{s \in \mathcal{S}} |\mathcal{A}_{\overline{\Pi}}(s)| \leq |\mathcal{A}|^{|\mathcal{S}|}$.

Remark 7 (Comparing USPI and SSPI)**.** Although SSPI maximizes over a set of policies that is a very large superset of the policies considered by USPI, it may happen that the policy improvement bound found by SSPI is smaller than the one of USPI. The reason is that the former optimizes the bound in Corollary 4 that is looser than the bound in Theorem 3 optimized by the latter. Finally, notice that, following the same procedure described in Remark 4 and constraining SSPI to use a single α for all the states (so that the SSPI improvement is bounded by $\frac{\mathbb{A}_{\overline{\pi}}^2}{2\gamma \|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty^2 \|\mathbf{q}^\pi\|_\infty}$), we can prove, as done with USPI, that the improvement of SSPI is never worse than the one of CPI.

4.3 per–State–Action–parameter Safe Policy Improvement

We can further generalize SSPI by considering an update scheme in which the new policy is generated using different convex combination coefficients for each state–action pair:

$$\pi'(a|s) = \alpha(s,a)\overline{\pi}(a|s) + \big(1-\alpha(s,a)\big)\pi(a|s), \tag{8}$$

where $\alpha(s,a) \in [0,1],\ \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. Note that, in order to ensure a valid probability distribution we need to impose that, $\forall s \in \mathcal{S}$, $\sum_{a\in\mathcal{A}} \alpha(s,a)\left(\overline{\pi}(a|s) - \pi(a|s)\right) = 0$. As for SSPI, the bound in Theorem 3 cannot be optimized easily, thus we consider again the simplified bound in Corollary 4.

The novel idea of this improvement scheme called *per–State–Action–parameter Safe Policy Improvement* (SASPI), consists in the fact that, for each state, we can move probability across the actions. For a given state s and action a, we define the probability increment induced by coefficient $\alpha(s,a)$ as $\Delta(s,a) = \pi'(a|s) - \pi(a|s) = \alpha(s,a)\left(\overline{\pi}(a|s) - \pi(a|s)\right)$. Clearly, we cannot change the probability arbitrarily, as we need to satisfy the following constraints:

$$\forall s \in \mathcal{S}, \quad \sum_{a\in\mathcal{A}} \Delta(s,a) = 0, \tag{9}$$

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \quad \begin{cases} 0 \le \Delta(s,a) \le \overline{\pi}(a|s) - \pi(a|s) & \text{if } \overline{\pi}(a|s) \ge \pi(a|s) \\ \overline{\pi}(a|s) - \pi(a|s) \le \Delta(s,a) \le 0 & \text{otherwise} \end{cases}. \tag{10}$$

The constraint (9) ensures that the resulting policy $\pi'(a|s) = \pi(a|s) + \Delta(s,a)$ is a valid probability distribution for all $s \in \mathcal{S}$, while constraint (10) guarantees that the chosen $\Delta(s,a)$ realizes a convex combination of the entries of the current policy π and those of the target policy $\overline{\pi}$. As a consequence, for each state $s \in \mathcal{S}$ we can partition the actions into three sets according to the sign of $\overline{\pi}(a|s) - \pi(a|s)$. $\mathcal{A}_s^{\uparrow} = \{a \in \mathcal{A} : \overline{\pi}(a|s) > \pi(a|s)\}$ is the set of the actions whose probability can only be increased, $\mathcal{A}_s^{\downarrow} = \{a \in \mathcal{A} : \overline{\pi}(a|s) < \pi(a|s)\}$ is the set of the actions whose probability can only be decreased, and if $\mathcal{A}_s^{=} = \{a \in \mathcal{A} : \overline{\pi}(a|s) = \pi(a|s)\}$ is the set of the actions whose probability cannot change whatever $\alpha(s,a)$ we pick.

To solve the problem of determining the optimal values of the $\Delta(s,a)$, we adopt an approach similar to that of SSPI and we introduce a budget $\Upsilon = \|\boldsymbol{\pi'} - \boldsymbol{\pi}\|_{\infty}$, with π' as defined in Equation (8). Notice that Υ can be spent independently in each state s, by definition of L^{∞}-norm. Suppose we are able to find the optimal budget value Υ^*, our optimization problem consists of maximizing the bound in Corollary 4 over the probability increments $\Delta(s,a)$ having a fixed budget Υ and fulfilling the constraints (9) and (10). Ideally, we would like to increase the probability of the actions with high Q^{π} and decrease the probability of the actions with low Q^{π}. Notice that, in order to satisfy (9), for each state, the amount of probability we add must coincide with the amount of probability we subtract across all actions. Thus, given a budget Υ we can increase (resp. decrease) the probability of the actions by $\Upsilon/2$ at most. In order to define the update rule, let $\rho_s^{\pi} : \mathcal{A} \to \{1, 2, \ldots, |\mathcal{A}|\}$ be an ordering of the actions for each state, such that if $\rho_s^{\pi}(a) < \rho_s^{\pi}(a') \implies Q^{\pi}(s,a) \le Q^{\pi}(s,a')$. We define the following quantities:

$$G^{\uparrow}(s,a) = \sum_{\substack{a'\in\mathcal{A}_s^{\uparrow}: \\ \rho_s^{\pi}(a') > \rho_s^{\pi}(a)}} \left(\overline{\pi}(a'|s) - \pi(a'|s)\right),$$

$$G^{\downarrow}(s,a) = \sum_{\substack{a' \in \mathcal{A}_s^{\downarrow}: \\ \rho_s^{\pi}(a') < \rho_s^{\pi}(a)}} \left(\pi(a'|s) - \overline{\pi}(a'|s)\right).$$

Given an action $a \in \mathcal{A}$, $G^{\uparrow}(s,a)$ represents the amount by which the total probability of all the actions with Q^{π} larger than $Q^{\pi}(s,a)$ can be increased. Symmetrically, for an action $a \in \mathcal{A}$, $G^{\downarrow}(s,a)$ represents the amount by which the total probability of all the actions with Q^{π} smaller than $Q^{\pi}(s,a)$ can be decreased. Note that it is not always convenient to spend Υ completely in every state. Indeed, it might be the case that in order to spend it all, we have to increase the probability of actions with low Q^{π} and decrease the probability of actions with high Q^{π}, which is clearly inconvenient. For this reason, we define the *expendable budget* for an action $a \in \mathcal{A}$ in a state $s \in \mathcal{S}$ as:

$$\Upsilon(s,a) = \begin{cases} \max\left\{0, \min\left\{\overline{\pi}(a|s) - \pi(a|s), G^{\downarrow}(s,a) - G^{\uparrow}(s,a)\right\}\right\} & \text{if } a \in \mathcal{A}_s^{\uparrow} \\ \max\left\{0, \min\left\{\pi(a|s) - \overline{\pi}(a|s), G^{\uparrow}(s,a) - G^{\downarrow}(s,a)\right\}\right\} & \text{if } a \in \mathcal{A}_s^{\downarrow} \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

To grasp the intuition behind the definition of expendable budget, consider an action $a \in \mathcal{A}_s^{\uparrow}$. We have two conditions to satisfy in order to define the budget $\Upsilon(s,a)$. First, for a we can increase its probability by at most $\overline{\pi}(a|s) - \pi(a|s)$. However, this might not be convenient depending on how much the probability of actions with Q^{π} smaller than $Q^{\pi}(s,a)$ can be decreased, i.e., $G^{\downarrow}(s,a)$. The best way of moving probability across actions consists of increasing the probability of actions in decreasing order of Q^{π} and decreasing the probability of actions in increasing order of Q^{π}. Thus, the second condition can be stated as follows. Given an action $a \in \mathcal{A}_s^{\uparrow}$, recalling that we have increased the probability of all actions with Q^{π} higher than $Q^{\pi}(s,a)$ as much as possible, i.e., by $G^{\uparrow}(s,a)$, the budget we have at our disposal is at most $G^{\downarrow}(s,a) - G^{\uparrow}(s,a)$. Therefore, to define $\Upsilon(s,a)$ we take the minimum between the two cases, which leads to Equation (11). A similar rationale holds for actions in $\mathcal{A}_s^{\downarrow}$. Clearly, for actions $a \in \mathcal{A}_s^{=}$ we have $\Upsilon(s,a) = 0$. We can also define the expendable budget for a state $s \in \mathcal{S}$ as: $\Upsilon(s) = \sum_{a \in \mathcal{A}} \Upsilon(s,a)$. If $\Upsilon \leq \Upsilon(s)$ we can define the two *active* actions, i.e., those of which we are currently increasing ($a_s^{\uparrow}$) and decreasing ($a_s^{\downarrow}$) the probability:

$$a_s^{\uparrow} = \underset{\substack{a \in \mathcal{A}_s^{\uparrow}: \\ G^{\uparrow}(s,a) \leq \frac{\Upsilon}{2}}}{\arg\max} \left\{\rho_s^{\pi}(a)\right\}, \qquad a_s^{\downarrow} = \underset{\substack{a \in \mathcal{A}_s^{\downarrow}: \\ G^{\downarrow}(s,a) \leq \frac{\Upsilon}{2}}}{\arg\max} \left\{\rho_s^{\pi}(a)\right\}. \tag{12}$$

We can now state the following optimality condition.

Corollary 13. *Let $\mathcal{S}_{\pi}^{\overline{\pi}} = \{s \in \mathcal{S} : d^{\pi}(s) > 0\}$ and Υ^* be the value that maximizes the following function:*

$$B(\Upsilon) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_{\pi}^{\overline{\pi}}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \Delta(s,a,\Upsilon) Q^{\pi}(s,a) - \frac{\gamma}{(1-\gamma)^2} \Upsilon^2 \frac{\|\mathbf{q}^{\pi}\|_{\infty}}{2},$$

where

$$\Delta(s,a,\Upsilon) = \begin{cases} \max\left\{0, \min\left\{\Upsilon(s,a), \frac{\Upsilon}{2} - G^{\uparrow}(s,a)\right\}\right\} & \text{if } a \in \mathcal{A}_s^{\uparrow} \\ -\max\left\{0, \min\left\{\Upsilon(s,a), \frac{\Upsilon}{2} - G^{\downarrow}(s,a)\right\}\right\} & \text{if } a \in \mathcal{A}_s^{\downarrow} \\ 0 & \text{if } a \in \mathcal{A}_s^{=} \end{cases}.$$

We set $\Delta(s,a,\Upsilon)=0$ *when* $d^\pi(s)=0$. *Then, the bound in Corollary 4 is optimized by taking*

$$\alpha(s,a)=\begin{cases}0 & \text{if } \overline{\pi}(a|s)=\pi(a|s)\\ \frac{\Delta(s,a,\Upsilon^*)}{\overline{\pi}(a|s)-\pi(a|s)} & \text{otherwise}\end{cases}$$

Proof First note that if $d^\pi(s)=0$, state s does not have any contribution in the bound value, thus we can set $\Delta(s,a)=0$ and restrict our analysis to the states in $\mathcal{S}_\pi^{\overline{\pi}}$. We now evaluate the contribution of each action to the bound in Corollary 4:

$$\begin{aligned}J^{\pi'}-J^\pi \geq& \frac{1}{1-\gamma}\mathbf{d}^{\pi\mathtt{T}}\mathbf{a}_\pi^{\pi'}-\frac{\gamma}{(1-\gamma)^2}\left\|\boldsymbol{\pi'}-\boldsymbol{\pi}\right\|_\infty^2\frac{\|\mathbf{q}^\pi\|_\infty}{2}\\ =&\frac{1}{1-\gamma}\sum_{s\in\mathcal{S}_\pi^{\overline{\pi}}}d^\pi(s)\sum_{a\in\mathcal{A}}\left[\pi'(a|s)Q^\pi(s,a)-\pi(a|s)Q^\pi(s,a)\right]-\frac{\gamma}{(1-\gamma)^2}\left\|\boldsymbol{\pi'}-\boldsymbol{\pi}\right\|_\infty^2\frac{\|\mathbf{q}^\pi\|_\infty}{2}\\ =&\frac{1}{1-\gamma}\sum_{s\in\mathcal{S}_\pi^{\overline{\pi}}}d^\pi(s)\sum_{a\in\mathcal{A}}\Delta(s,a)Q^\pi(s,a)-\frac{\gamma}{(1-\gamma)^2}\Upsilon^2\frac{\|\mathbf{q}^\pi\|_\infty}{2},\end{aligned}$$

where we denoted with $\Delta(s,a)=\pi'(a|s)-\pi(a|s)$ and $\Upsilon^2=\|\boldsymbol{\pi'}-\boldsymbol{\pi}\|_\infty^2$. Consider a fixed budget Υ. Since we can spend Υ independently in each state, we can reason for a generic state $s\in\mathcal{S}$. In particular, $\Upsilon/2$ can be used to increase the probability of some actions in $\mathcal{A}_s^\uparrow$ and $\Upsilon/2$ to decrease the probability of some actions in $\mathcal{A}_s^\downarrow$. The best we can do is to start increasing the probabilities of actions starting from the one with the highest Q^π value and, at the same time, decreasing the probabilities of actions starting from the one with the lowest Q^π value, until we ran out of budget. Thus, for an action $a\in\mathcal{A}_s^\uparrow$ we increase its probability as much as possible, i.e., by $\Upsilon(s,a)$. But we need to limit the increment if we do not have enough budget. Thus, if $\Upsilon(s,a)>\Upsilon/2-G^\uparrow(s,a)$ we need to set the increase to zero. Summing up, we have:

$$\Delta(s,a)=\max\left\{0,\min\left\{\Upsilon(s,a),\frac{\Upsilon}{2}-G^\uparrow(s,a)\right\}\right\},\quad\forall s\in\mathcal{S},\forall a\in\mathcal{A}_s^\uparrow. \tag{P.10}$$

Analogously, for all actions $a\in\mathcal{A}_s^\downarrow$ we decrease the probability as much as possible, provided that we have enough budget:

$$\Delta(s,a)=-\max\left\{0,\min\left\{\Upsilon(s,a),\frac{\Upsilon}{2}-G^\downarrow(s,a)\right\}\right\},\quad\forall s\in\mathcal{S},\forall a\in\mathcal{A}_s^\downarrow. \tag{P.11}$$

For the actions $a\in\mathcal{A}_s^=$ we need to set $\Delta(s,a)=0$. Recalling that $\alpha(s,a)=\frac{\Delta(s,a)}{\overline{\pi}(a|s)-\pi(a|s)}$ we get the result. ■

Algorithm 6 provides the pseudocode of SASPI. Similarly to SSPI, the termination condition is $\Upsilon^*=0$, i.e., when no improvement can be obtained with the target policy $\overline{\pi}$.

4.3.1 Computing Υ^*

Like for SASPI, we face the problem of computing Υ^*, which requires the maximization of a function with linear discontinuous derivative. Note that, once again, the slope of each

Algorithm 6 Exact SASPI.

input: MDP $\mathcal{M}$, target policy space $\overline{\Pi}$, policy chooser PC
Initialize π
$\overline{\pi} \leftarrow \mathrm{PC}(\mathcal{M}, \overline{\Pi}, \pi)$
$\Upsilon^* \leftarrow \mathrm{FBO}(\mathcal{M}, \overline{\pi}, \pi)$ (see Alg. 4)
while $\Upsilon^* > 0$ **do**
  Compute $\Upsilon(s,a)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$

$$\Delta(s,a) \leftarrow \begin{cases} \max\left\{0, \min\left\{\Upsilon(s,a), \frac{\Upsilon}{2} - G^{\uparrow}(s,a)\right\}\right\} & \text{if } a \in \mathcal{A}_s^{\uparrow}, \\ -\max\left\{0, \min\left\{\Upsilon(s,a), \frac{\Upsilon}{2} - G^{\downarrow}(s,a)\right\}\right\} & \text{if } a \in \mathcal{A}_s^{\downarrow}, \\ 0 & \text{if } a \in \mathcal{A}_s^{=}, \end{cases} \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

  $\alpha(s,a) \leftarrow \frac{\Delta(s,a)}{\overline{\pi}(a|s) - \pi(a|s)} \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$
  $\pi(a|s) \leftarrow \alpha(s,a)\overline{\pi}(a|s) + (1 - \alpha(s,a))\,\pi(a|s), \ \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$
  $\overline{\pi} \leftarrow \mathrm{PC}(\mathcal{M}, \overline{\Pi}, \pi)$
  $\Upsilon^* \leftarrow \mathrm{FBO}(\mathcal{M}, \overline{\pi}, \pi)$
end while

piece is the same and equal to $m = -\frac{\gamma \|\mathbf{q}\|_\infty}{(1-\gamma)^2}$. For this reason, provided that we are able to compute the coordinates of the discontinuity points (by using a properly defined FJP), we can employ Algorithm 4 to find Υ^*. Let us now write the explicit expression of the derivative of the bound:

$$\frac{\partial}{\partial \Upsilon} B(\Upsilon) = \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\Upsilon} \frac{1}{2} d^\pi(s) \left(Q^\pi(s, a_s^{\uparrow}) - Q^\pi(s, a_s^{\downarrow}) \right) - \Upsilon \frac{\gamma \left\| \mathbf{q}^\pi \right\|_\infty}{(1-\gamma)^2} = g(\Upsilon) + m\Upsilon, \quad (13)$$

where $\mathcal{S}_\Upsilon = \{s \in \mathcal{S} : \Upsilon(s) \leq \Upsilon\}$ is the set of states for which we still have budget (non-saturated) and thus their contribution in the summation is dependent on Υ. In such a case, there exist two active actions, as defined in Equation (12): $a_s^{\uparrow}$ of which we are increasing the probability and $a_s^{\downarrow}$ of which we are decreasing the probability. The probability of the other actions are either saturated to the maximum value or kept unchanged and thus, independent of Υ. Similarly to what happens in SSPI, the derivative is non negative at $\Upsilon = 0$, and it is monotonically decreasing since both $Q^\pi(s, a_s^{\uparrow})$ and $-Q^\pi(s, a_s^{\downarrow})$ are decreasing functions of Υ. Therefore, B is guaranteed to have a unique maximum since actions $a_s^{\uparrow}$ and $a_s^{\downarrow}$ are considered in decreasing and increasing order of Q^π, respectively. The discontinuity points correspond to values of Υ for which either one state saturates, i.e., Υ reaches $\Upsilon(s)$, or $a_s^{\uparrow}$ or $a_s^{\downarrow}$ change. In order to find these discontinuity points, we can adopt Algorithm 7. The idea of the algorithm is to go through state–action pairs sorted according to the budget at which they are going to saturate. This information is provided by $G^{\uparrow}(s,a)$ for actions candidate to increase their probability and by $G^{\downarrow}(s,a)$ for actions candidate to decrease their probability. Therefore we consider two orderings $\rho^{\uparrow}$ and $\rho^{\downarrow}$ in which the state–action pairs are sorted according to $G^{\uparrow}(s,a)$ and $G^{\downarrow}(s,a)$, respectively. At each iteration t we consider the pair (s,a) that will saturate sooner (the first if). For this pair, two situations might happen: (i) it is convenient to perform the update, i.e., the action a_t whose probability is going to be increased (resp. decreased) has higher Q^π than the action we have just decreased (resp. increased) the probability $a_{s_t}^{\downarrow}$; (ii) the update is not convenient. In case (ii), we need to declare the state s_t as saturated and remove it from the set $\mathcal{S}_{\overline{\pi}}^{\overline{\pi}}$. The

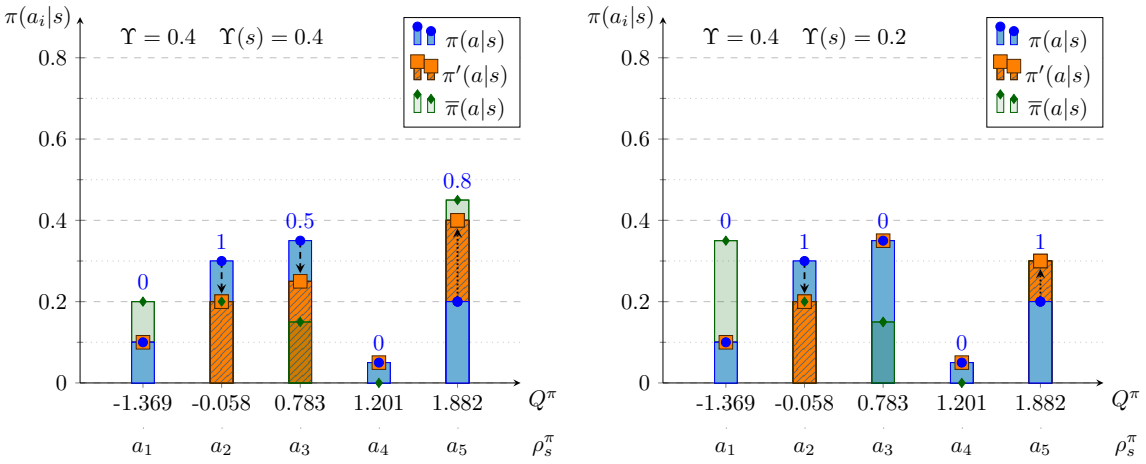


Figure 2: SASPI policy update.

computational complexity of Algorithm 7 is dominated by the computation of the orderings $\rho^{\uparrow}$ and $\rho^{\downarrow}$, which costs $O\left(|\mathcal{S}||\mathcal{A}|\log|\mathcal{A}||\mathcal{S}|\right)$, as the computation of $G^{\uparrow}(s,a)$ and $G^{\downarrow}(s,a)$ has cost $O\left(|\mathcal{S}||\mathcal{A}|\log|\mathcal{A}|\right)$ and the loop is executed at most $|\mathcal{S}||\mathcal{A}|$ times and at each iteration the cost is constant. In the following, we report a couple of examples of policy update using SASPI.

Example 1. *The initial policy (blue area with dotted mark), the updated policy (orange area with square mark) and the target policy (green area with diamond mark) are depicted in Figure 2. Actions are ordered according to ρ_s^π, i.e., in ascending order according to their Q–values. We fix a budget $\Upsilon = 0.4$. At the top, we show a case in which we are able to spend the whole budget in state s, i.e., $\Upsilon(s) = 0.4$. We start from the best a_5 and the worst a_1 actions. We try to increase the probability of a_5 and decrease that of a_1, but we cannot as $\overline{\pi}(a_1|s) > \pi(a_1|s)$, so we move to action a_2. We decrease the probability of a_2 by 0.1 and we increase the probability of a_5 by the same amount. Now, we move to action a_3 and we decrease its probability by 0.1 as well while increasing that of a_5 by the same amount. Since we have no further budget, we stop. At the bottom, we show a case where the expendable budget $\Upsilon(s) = 0.2$ is less than Υ. Again we start with a_1 and a_5. Similar to the case on the left, we have to move to a_2. Now, we can increase the probability of a_5 by 0.1 and meanwhile reduce the probability of a_2 by the same amount. We could now decrease the probability of a_3 but we have not enough probability on a_4 to compensate. Thus, we stop. The updates $\Delta(s,a)$ are reported in the figure as dashed arrows. Coefficients $\alpha(s,a_i)$ are drawn above the bars.*

Remark 8 (SASPI vs SSPI)**.** It is worth noting that the more flexible update rule introduced by SASPI shows an advantage over SSPI only when considering problems with more than two actions. Indeed, when $\mathcal{A} = \{a_1, a_2\}$ we have that $\alpha(s,a_1) = \alpha(s,a_2)$ since from Constraint (9) we have:

$$0 = \Delta(s,a_1) + \Delta(s,a_2) = \alpha(s,a_1)\left(\overline{\pi}(a_1|s) - \pi(a_1|s)\right) + \alpha(s,a_2)\left(\overline{\pi}(a_2|s) - \pi(a_2|s)\right)$$

Algorithm 7 Computing of the jump points for SASPI (Find Jump Point - FJP)

input: MDP $\mathcal{M}$, current policy π, target policy $\overline{\pi}$
initialize: $t \leftarrow 0$, $\mathcal{S}_\pi^{\overline{\pi}} \leftarrow \{s \in \mathcal{S} : d^\pi(s) > 0\}$, $\frac{\Upsilon_0}{2} \leftarrow 0$, $i^\uparrow \leftarrow 0$, $i^\downarrow \leftarrow 0$
Compute $G^\uparrow(s,a)$ and $G^\downarrow(s,a)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$
Compute the two orderings $\rho^\uparrow$ and $\rho^\downarrow$ such that s.t. $i < j \implies G^\uparrow(s_{\rho_i^\uparrow}, a_{\rho_i^\uparrow}) \leq G^\uparrow(s_{\rho_j^\uparrow}, a_{\rho_j^\uparrow})$ and $i < j \implies G^\downarrow(s_{\rho_i^\downarrow}, a_{\rho_i^\downarrow}) \leq G^\uparrow(s_{\rho_j^\downarrow}, a_{\rho_j^\downarrow})$
Compute $a_s^\uparrow = \arg\max_{a \in \mathcal{A}} \{Q^\pi(s,a)\}$ and $a_s^\downarrow = \arg\min_{a \in \mathcal{A}} \{Q^\pi(s,a)\}$, $\forall s \in \mathcal{S}$
$q_0 \leftarrow \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \frac{1}{2} d^\pi(s) \left(Q^\pi(s, a_s^\uparrow) - Q^\pi(s, a_s^\downarrow)\right)$
yield Υ_0, q_0
while $\mathcal{S}_\pi^{\overline{\pi}} \neq \{\}$ **do**
  $t \leftarrow t + 1$
  if $s_t \in \mathcal{S}_\pi^{\overline{\pi}}$ **then**
    if $G^\uparrow(s_{\rho_{,i\uparrow}^\uparrow} a_{\rho_{i\uparrow}^\uparrow}) \leq G^\downarrow(s_{\rho_{i\downarrow}^\downarrow}, a_{\rho_{i\downarrow}^\downarrow})$ **then**
      $s_t, a_t \leftarrow s_{\rho_{i\uparrow}^\uparrow}, a_{\rho_{i\uparrow}^\uparrow}$
      if $Q^\pi(s_t, a_t) > Q^\pi(s_t, a_{s_t}^\downarrow)$ **then**
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(Q^\pi(s_t, a_{s_i t}^\uparrow) - Q^\pi(s_t, a_t)\right)$
        $a_{s_t}^\uparrow \leftarrow a_t$
      else
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(Q^\pi(s_t, a_{s_t}^\uparrow) - Q^\pi(s_t, a_{s_t}^\downarrow)\right)$
        $\mathcal{S}_\pi^{\overline{\pi}} \setminus \{s_t\}$
      end if
      $\frac{\Upsilon_t}{2} \leftarrow G^\uparrow(s_t, a_t)$
      $i^\uparrow \leftarrow i^\uparrow + 1$
    else
      $s_t, a_t \leftarrow s_{\rho_{i\downarrow}^\downarrow}, a_{\rho_{i\downarrow}^\downarrow}$
      if $Q^\pi(s_t, a_t) < Q^\pi(s_t, a_{s_t}^\downarrow)$ **then**
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(Q^\pi(s_t, a_t) - Q^\pi(s_t, a_{s_t}^\downarrow)\right)$
        $a_{s_t}^\downarrow \leftarrow a_t$
      else
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(Q^\pi(s_t, a_{s_t}^\uparrow) - Q^\pi(s_t, a_{s_t}^\downarrow)\right)$
        $\mathcal{S}_\pi^{\overline{\pi}} \setminus \{s_t\}$
      end if
      $\frac{\Upsilon_t}{2} \leftarrow G^\downarrow(s_t, a_t)$
      $i^\downarrow \leftarrow i^\downarrow + 1$
    end if
  end if
  yield Υ_t, q_t
end while
yield 2, $-\infty$

$$= (\alpha(s,a_1) - \alpha(s,a_2))\,(\overline{\pi}(a_1|s) - \pi(a_1|s)) \implies \alpha(s,a_1) = \alpha(s,a_2).$$

As a consequence, the coefficient α depends on the state only, like in SSPI.

Remark 9 (Optimality of USPI, SSPI and SASPI)**.** In general, selecting a greedy policy π^+ as target policy for USPI and SSPI is not optimal in terms of bound value, i.e., there might exist a target policy $\overline{\pi} \neq \pi^+$ that allows reaching higher values of the bound. Indeed,

as proved in Appendix B, the policy that maximizes globally the bound, as defined in Corollary 29, might be outside the space of representable policies given the USPI and SSPI update rule. On the contrary, this policy is always representable with the update rule of SASPI provided that we select π^+ as target policy. Thus, SASPI is optimal in terms of bound value using the GPC (Greedy Policy Chooser, Algorithm 2).

Remark 10 (On the Convergence of SSPI and SASPI)**.** We have not provided a specific result for the convergence of SSPI and SASPI. It is worth noting that these two algorithms converge to the optimal policy under the same assumptions enforced for USPI (Section 4.1.1). Indeed, the convergence proof (Theorem 8) employs the performance improvement produced by USPI as in Corollary 6, i.e., USPI with the simplified bound. SSPI and SASPI yield a larger policy improvement compared to that of Corollary 6 and, consequently, the schema of the proof applies straightforwardly.

5. Approximate Safe Policy Iteration

The exact algorithms proposed in the previous sections are impractical in applications where the state space is very large (or even continuous) or when the state–transition model is unknown. In this section, we move to the approximate setting in which the quantities involved in the bounds need to be estimated from samples. We start introducing the approximate versions of the policy choosers (Section 5.1), then we present the approximate versions of the three algorithms we introduced in Section 4: aUSPI (Section 5.2), aSSPI (Section 5.3), and aSASPI (Section 5.4). For each of them, we present the algorithm, a sample complexity analysis and, when possible, we discuss the convergence properties.

Since accurate estimates of L_∞–norms ($\mathrm{sp}(\mathbf{a})$ for USPI and $\|\mathbf{q}^\pi\|_\infty$ for SSPI, and SASPI) need many samples, for approximate settings, we consider a further simplified bound obtained from Corollary 4 by observing that $\|\mathbf{q}^\pi\|_\infty \leq \frac{1}{1-\gamma}$:

$$J^{\pi'} - J^\pi \geq \frac{1}{1-\gamma}\mathbb{A}_\pi^{\pi'} - \frac{\gamma}{2(1-\gamma)^3}\left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_\infty^2 . \tag{14}$$

In this way, the only value that needs to be estimated is $\mathbb{A}_\pi^{\pi'}$. Following Kakade (2003), the goal of this section is to provide an analysis of the SPI algorithms presented when dealing with tabular domains.

In the following, we will denote with the "hat" the estimated/approximated quantities. We will employ this notation for both the value functions (e.g., $\widehat{Q}^\pi$, $\widehat{A}_\pi^{\widehat{\pi}}$) and the policies (e.g., $\widehat{\pi}$). Approximation involves essentially two parts of the SPI algorithms. First, the policy chooser needs samples to select in an approximate manner a suitable target policy. We will discuss in Section 5.1 how to extend the Greedy Policy Chooser (GPC) to account for samples. Second, once a target policy is available, we need to perform the policy update computing the convex combination coefficients α. This latter phase heavily depends on the SPI algorithm and we will discuss how to perform it for a fixed target policy in the Sections 5.2–5.4. The proofs of all the presented results are reported in Appendix A.3.

Algorithm 8 Approximate Greedy Policy Chooser ($\widehat{\text{GPC}}$).

input: current policy π, target policy space $\overline{\Pi}$
initialize: $N \leftarrow \left\lceil \frac{32|\mathcal{A}|^2}{9\epsilon^2} \left(\log(2|\overline{\Pi}|) + \log \frac{1}{\delta}\right) \right\rceil$, $T \leftarrow \left\lceil \log_\gamma \frac{\epsilon}{24} \right\rceil$
Compute a biased sample set $\{(s_i, a_i, \widehat{q}_i)\}_{i=1}^N$.
for each $\pi^\dagger \in \overline{\Pi}$ **do**
    Construct the estimate $\widehat{\mathbb{Q}}_\pi^{\pi^\dagger} = \frac{|\mathcal{A}|}{N} \sum_{i=1}^N \widehat{q}_i^\pi \pi^\dagger(a_i|s_i)$
end for
return $\arg\max_{\pi^\dagger \in \overline{\Pi}} \left\{ \widehat{\mathbb{Q}}_\pi^{\pi^\dagger} \right\}$

5.1 Approximate Policy Choosers

Since we are in model–free setting, we need a new policy chooser that, at each iteration, given a policy π, returns the target policy $\overline{\pi} \in \overline{\Pi}$ selected using samples. We will discuss in this section how to adapt the Greedy Policy Chooser (GPC) to the approximate framework. Let us first focus on the requirements for a generic approximate policy chooser $\widehat{\text{PC}}$, which outputs a target policy $\overline{\pi}$. Given an accuracy level $\epsilon > 0$ and an (exact) policy chooser PC, the goal of the corresponding approximate policy chooser $\widehat{\text{PC}}$ consists in returning an approximate target policy $\widehat{\overline{\pi}}$ such that with probability at least $1 - \delta$, we have that the difference between the expected advantage $\mathbb{A}_\pi^{\overline{\pi}}$ of the true target policy $\overline{\pi}$ and the expected advantage $\mathbb{A}_\pi^{\widehat{\overline{\pi}}}$ of the approximate target policy $\widehat{\overline{\pi}}$ is bounded:

$$\left| \mathbb{A}_\pi^{\overline{\pi}} - \mathbb{A}_\pi^{\widehat{\overline{\pi}}} \right| \leq \frac{\epsilon}{1-\gamma}, \tag{15}$$

where $\overline{\pi}$ is the target policy returned by the exact policy chooser PC. Note that equivalently we can write that

$$\widehat{\overline{\pi}} \in \arg\max_{\pi^\dagger \in \overline{\Pi}} \widehat{\mathbb{A}}_\pi^{\pi^\dagger},$$

where $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ is the estimated expected advantage function of a candidate target policy $\pi^\dagger$. Considering the GPC, it is worth noting that the requirement at Equation (15) is not dissimilar from the typical requirement employed in API, i.e., the fact that the improvement step is made by using an *approximately greedy policy* (Scherrer, 2014). Indeed, we can rewrite the difference of the expected advantages as $\mathbb{A}_\pi^{\overline{\pi}} - \mathbb{A}_\pi^{\widehat{\overline{\pi}}} = \mathbf{d}^{\pi\intercal} \left(\overline{\boldsymbol{\pi}} - \widehat{\overline{\boldsymbol{\pi}}} \right) \mathbf{q}^\pi$. Similar conditions on the approximate greedy policy can be found in some fundamental works on API (e.g., Lagoudakis and Parr, 2003a; Lazaric et al., 2016). The pseudocode of approximate GPC ($\widehat{\text{GPC}}$) is reported in Algorithm 8. The following lemma provides a sufficient condition to meet the previous requirement.

Lemma 14. *If with probability at least $1 - \delta$, simultaneously for all $\pi^\dagger \in \overline{\Pi}$ it holds that:*

$$\left| \widehat{\mathbb{Q}}_\pi^{\pi^\dagger} - \mathbb{Q}_\pi^{\pi^\dagger} \right| \leq \frac{\epsilon}{2(1-\gamma)}, \tag{16}$$

where $\mathbb{Q}_\pi^{\pi^\dagger} = \mathbf{d}^{\pi\intercal} \boldsymbol{\pi}^\dagger \mathbf{q}^\pi$, then with probability at least $1 - \delta$ it holds that:

$$\left| \mathbb{A}_\pi^{\overline{\pi}} - \mathbb{A}_\pi^{\widehat{\overline{\pi}}} \right| \leq \frac{\epsilon}{1-\gamma},$$

where $\overline{\pi}$ is the target policy returned by GPC and $\widehat{\overline{\pi}}$ is the policy returned by $\widehat{GPC}$ (Algorithm 8).

Proof Let us denote with $\overline{\pi}$ the policy returned by GPC and with $\widehat{\overline{\pi}}$ the policy returned by $\widehat{\text{GPC}}$. We consider the following sequence of inequalities:

$$\begin{aligned}\mathbb{A}_{\pi}^{\overline{\pi}} - \mathbb{A}_{\pi}^{\widehat{\overline{\pi}}} &= \mathbb{A}_{\pi}^{\overline{\pi}} - \mathbb{A}_{\pi}^{\widehat{\overline{\pi}}} + \widehat{\mathbb{A}}_{\pi}^{\widehat{\overline{\pi}}} - \widehat{\mathbb{A}}_{\pi}^{\widehat{\overline{\pi}}} \\ &\leq \mathbb{A}_{\pi}^{\overline{\pi}} - \widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} + \widehat{\mathbb{A}}_{\pi}^{\widehat{\overline{\pi}}} - \mathbb{A}_{\pi}^{\widehat{\overline{\pi}}} \\ &\leq 2 \max_{\pi^{\dagger} \in \overline{\Pi}} \left\{ \left| \mathbb{A}_{\pi}^{\pi^{\dagger}} - \widehat{\mathbb{A}}_{\pi}^{\pi^{\dagger}} \right| \right\},\end{aligned}$$

where we exploited the fact that $\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} \leq \widehat{\mathbb{A}}_{\pi}^{\widehat{\overline{\pi}}}$, as $\widehat{\text{GPC}}$ returns the policy $\widehat{\overline{\pi}}$ that maximizes $\widehat{\mathbb{A}}_{\pi}^{\widehat{\overline{\pi}}}$. We now observe that for any $\pi^{\dagger} \in \overline{\Pi}$, $\mathbb{A}_{\pi}^{\pi^{\dagger}} = \mathbb{Q}_{\pi}^{\pi^{\dagger}} - \mathbb{Q}_{\pi}^{\pi}$. The result then follows from the hypothesis. ■

Thus, the problem translates into computing an $\frac{\epsilon}{2(1-\gamma)}$–accurate estimation of $\mathbb{Q}_{\pi}^{\pi^{\dagger}}$ for all $\pi^{\dagger} \in \overline{\Pi}$. To this purpose, we present the following sampling procedure.

GPC Sampling Procedure The policy chooser generates a dataset $\{(s_i, a_i, \widehat{q}_i)\}_{i=1}^{N}$ that for each state-action pair (s_i, a_i) reports an estimation of the Q-function of policy π, denoted by $\widehat{q}_i$. The state s_i must be extracted according to the γ–discounted future state distribution d^{π}. In order to generate a sample s_i from d^{π}, we draw a state s_0 from μ and then we follow the policy π. Since γ represents the probability of continuing the simulation, at each step, the obtained state s_i is accepted with probability γ and with probability $1-\gamma$ the simulation ends. In this way, the state s_i is extracted according to the distribution d^{π} (see Thomas, 2014). The simulation is constrained to terminate in at most T steps. This condition has the effect of introducing a bias, but, in the meantime, it ensures a termination even when γ approaches 1 (Algorithm 14). The sample $s_i \sim d^{\pi}$ is then used to build the single sample $(s_i, a_i, \widehat{q}_i)$. The associated action a_i is drawn uniform in the action space $\mathcal{A}$. In order to compute the approximate Q–value $\widehat{q}_i$ for the state-action pair (s_i, a_i) a unique run under policy π is simulated. The estimation is obtained by executing action a_i in state s_i and, then, following policy π. Starting from time $t = 0$, the simulation is repeated T–steps and the final value $\widehat{q}_i$ is obtained as: $\widehat{q}_i = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$, where r_{t+1} is the immediate reward at time step t (Algorithm 15).

Once we have computed the biased sample set $\{(s_i, a_i, \widehat{q}_i)\}_{i=1}^{N}$ we can estimate the future state-action value function $\mathbb{Q}_{\pi}^{\pi^{\dagger}}$ of a stationary target policy $\pi^{\dagger}$ w.r.t. the current policy π as the weighted average of all $\widehat{q}_i$ according to the policy $\pi^{\dagger}$:

$$\widehat{\mathbb{Q}}_{\pi}^{\pi^{\dagger}} = \frac{|\mathcal{A}|}{N} \sum_{i=1}^{N} \pi^{\dagger}(a_i|s_i) \cdot \widehat{q}_i.$$

For the approximate greedy policy chooser $\widehat{\text{GPC}}$ it suffices to return the policy $\pi^{\dagger} \in \Pi^{\text{SR}}$ that maximizes the sample–based Q-function $\widehat{\mathbb{Q}}_{\pi}^{\pi^{\dagger}}$. We can easily obtain an approximate version of $\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}}$, recalling that the average advantage $\mathbb{A}_{\pi}^{\overline{\pi}}$ can be written as follows:

$$\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} = \widehat{\mathbb{Q}}_{\pi}^{\overline{\pi}} - \widehat{\mathbb{V}}^{\pi} = \widehat{\mathbb{Q}}_{\pi}^{\overline{\pi}} - \widehat{\mathbb{Q}}_{\pi,}^{\pi}.$$

Indeed, if we consider as greedy policy the target policy $\overline{\pi}$, previous equation permits the computation of $\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}}$ without additional costs, because the estimates $\widehat{\mathbb{Q}}_{\pi}^{\overline{\pi}}$ and $\widehat{\mathbb{Q}}_{\pi}^{\pi}$ are computed during maximization process.

The following lemma gives a theoretical guarantee to the estimation process and to the quality of the estimated quantity. The result is obtained as a straightforward adaptation of Lemma 7.3.4 in Kakade (2003). Note that, in order to get an $\frac{\epsilon}{1-\gamma}$–accurate estimation of the average advantage $\mathbb{A}_{\pi}^{\overline{\pi}}$, we need to consider the contribution of $\widehat{\mathbb{Q}}_{\pi}^{\pi^\dagger}$ and $\widehat{\mathbb{Q}}_{\pi}^{\pi}$ that must be $\frac{\epsilon}{2(1-\gamma)}$–accurate.

Lemma 15. *Let $\overline{\Pi} \subseteq \Pi^{SR}$ be a class of stationary policies for an infinite horizon MDP. Let*

$$T = \left\lceil \log_\gamma \frac{\epsilon}{24} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{32|\mathcal{A}|^2}{9\epsilon^2} \left(\log(2\left|\overline{\Pi}\right|) + \log \frac{1}{\delta} \right) \right\rceil.$$

Upon input of a policy π, GPC Sampling Procedure constructs a function $\widehat{\mathbb{Q}}_{\pi}^{\pi^\dagger}$ such that with probability $1-\delta$, simultaneously for all $\pi^\dagger \in \overline{\Pi}$

$$\left| \widehat{\mathbb{Q}}_{\pi}^{\pi^\dagger} - \mathbb{Q}_{\pi}^{\pi^\dagger} \right| \leq \frac{\epsilon}{2(1-\gamma)}.$$

Proof GPC Sampling Procedure has a form of bias due to the finite horizon T employed to generate samples from d^π and to estimate Q^π. For the sake of the analysis, let us define the T-horizon γ-discounted stationary distribution as:

$$d_T^\pi(s) = \frac{1-\gamma}{1-\gamma^T} \sum_{t=0}^{T-1} \gamma^t \Pr\left(s_t = s | \pi, \mathcal{M}\right),$$

and the T-horizon action-value function:

$$Q_T^\pi(s,a) = \mathop{\mathbb{E}}_{\substack{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t) \\ a_{t+1}\sim\pi(\cdot|s_{t+1})}} \left[\sum_{t=0}^{T-1} \gamma^t \mathcal{R}(s_t,a_t) | s_0 = s, a_0 = a \right],$$

and finally, we define:

$$\mathbb{Q}_{\pi,T}^{\pi^\dagger} = \mathop{\mathbb{E}}_{\substack{s\sim d_T^\pi \\ a\sim\pi(\cdot|s)}} \left[Q_T^\pi(s,a) \right].$$

Given the horizon truncation, our estimator is unbiased for $\mathbb{Q}_{\pi,T}^{\pi^\dagger}$. Therefore, we decompose the difference into two terms:

$$\left| \widehat{\mathbb{Q}}_{\pi}^{\pi^\dagger} - \mathbb{Q}_{\pi}^{\pi^\dagger} \right| \leq \underbrace{\left| \widehat{\mathbb{Q}}_{\pi}^{\pi^\dagger} - \mathbb{Q}_{\pi,T}^{\pi^\dagger} \right|}_{(E_1)} + \underbrace{\left| \mathbb{Q}_{\pi,T}^{\pi^\dagger} - \mathbb{Q}_{\pi}^{\pi^\dagger} \right|}_{(E_2)}.$$

We want to obtain that, with a probability of $1-\delta$, the largest deviation of the estimation from the true value is at most $E_1 + E_2 = \frac{\epsilon}{2(1-\gamma)}$. This can be achieved by setting $E_1 = \frac{3\epsilon}{8(1-\gamma)}$

and $E_2 = \frac{\epsilon}{8(1-\gamma)}$.[7] Let us start with E_1 and fix any strategy $\pi^\dagger \in \overline{\Pi}$. The crucial observation is that, the value $\widehat{q}_i$ generated by different trajectories are independent. This independence implies that we can apply Heöffding's bound for the deviation of an estimate from its mean. Note that the values $|\mathcal{A}|\widehat{q}_i \ \pi^\dagger(a_i|s_i)$ involved in the estimation of $\widehat{\mathbb{Q}}_\pi^{\pi^\dagger}$ are limited in the interval $\left[0, \frac{|\mathcal{A}|}{1-\gamma}\right]$ (see GPC Sampling Procedure). Hoeffding's bound implies that:

$$\Pr\left(\left|\widehat{\mathbb{Q}}_\pi^{\pi^\dagger} - \mathbb{Q}_{\pi,T}^{\pi^\dagger}\right| \geq \frac{3\epsilon}{8(1-\gamma)}\right) \leq 2e^{-\frac{2\left(\frac{3\epsilon}{8(1-\gamma)}\right)^2 N^2}{\sum_{i=1}^N (b_i - a_i)^2}}$$
$$= 2e^{-\frac{18\epsilon^2 N^2 (1-\gamma)^2}{64(1-\gamma)^2 N|\mathcal{A}|^2}} = 2e^{-\frac{9\epsilon^2 N}{32|\mathcal{A}|^2}}.$$

So far, we have restricted our attention to a fixed policy $\pi^\dagger$. Exploiting the union bound we have that the probability that any $\pi^\dagger \in \overline{\Pi}$ deviates by more than $\frac{3\epsilon}{8(1-\gamma)}$ from its mean is bounded by $2|\overline{\Pi}|e^{-\frac{9\epsilon^2 N}{32|\mathcal{A}|^2}} \leq \delta$. Solving the equation for N, we obtain:

$$e^{-\frac{9\epsilon^2 N}{32|\mathcal{A}|^2}} \leq \frac{\delta}{2|\overline{\Pi}|} \implies \frac{9\epsilon^2 N}{32|\mathcal{A}|^2} \geq -\log\frac{\delta}{2|\overline{\Pi}|} \implies N \geq \frac{32|\mathcal{A}|^2}{9\epsilon^2}\left(\log(2\left|\overline{\Pi}\right|) + \log\frac{1}{\delta}\right).$$

We now focus on E_2 and we further decompose it, highlighting the contribution of the finite horizon T:

$$\left|\mathbb{Q}_{\pi,T}^{\pi^\dagger} - \mathbb{Q}_\pi^{\pi^\dagger}\right| = \left|\mathop{\mathbb{E}}_{\substack{s\sim d_T^\pi \\ a\sim\pi(\cdot|s)}}[Q_T^\pi(s,a)] - \mathop{\mathbb{E}}_{\substack{s\sim d^\pi \\ a\sim\pi(\cdot|s)}}[Q^\pi(s,a)] \pm \mathop{\mathbb{E}}_{\substack{s\sim d^\pi \\ a\sim\pi(\cdot|s)}}[Q_T^\pi(s,a)]\right|$$
$$\leq \left|\mathop{\mathbb{E}}_{\substack{s\sim d_T^\pi \\ a\sim\pi(\cdot|s)}}[Q_T^\pi(s,a)] - \mathop{\mathbb{E}}_{\substack{s\sim d^\pi \\ a\sim\pi(\cdot|s)}}[Q_T^\pi(s,a)]\right| - \left|\mathop{\mathbb{E}}_{\substack{s\sim d^\pi \\ a\sim\pi(\cdot|s)}}[Q^\pi(s,a) - Q_T^\pi(s,a)]\right|$$
$$\leq \frac{1}{1-\gamma}\left\|\mathbf{d}^\pi - \mathbf{d}_T^\pi\right\|_1 + \left\|\mathbf{q}^\pi - \mathbf{q}_T^\pi\right\|_\infty \leq \frac{3\gamma^T}{1-\gamma}.$$

where we exploited the fact that $\left\|\mathbf{q}_T^\pi\right\|_\infty \leq \frac{1}{1-\gamma}$ and we used Lemma 26 and Lemma 27 to obtain the last line from the last but one. The value of T is obtained by solving $\frac{3\gamma^T}{1-\gamma} = \frac{\epsilon}{8(1-\gamma)}$. Hence, with a probability of $1-\delta$, the deviation from the true mean is $\frac{\epsilon}{2(1-\gamma)}$. ■

The computation of the arg max in Algorithm 8 might yield multiple solutions, especially when there are states that are never visited in the collected trajectories. Since the improvement ensured by the SPI algorithms is inversely proportional to the distance between the current policy $\boldsymbol{\pi}$ and the target policy $\boldsymbol{\pi}^\dagger$ (Corollary 6), ties are broken by selecting the policy minimizing $\left\|\boldsymbol{\pi} - \boldsymbol{\pi}^\dagger\right\|_\infty$ among those maximizing $\widehat{\mathbb{Q}}_{\pi,\mu}^{\pi^\dagger}$. This corresponds to set $\pi^\dagger(\cdot|s) = \pi(\cdot|s)$ in all states s that are never visited.

7. Other decompositions of the error are clearly possible.

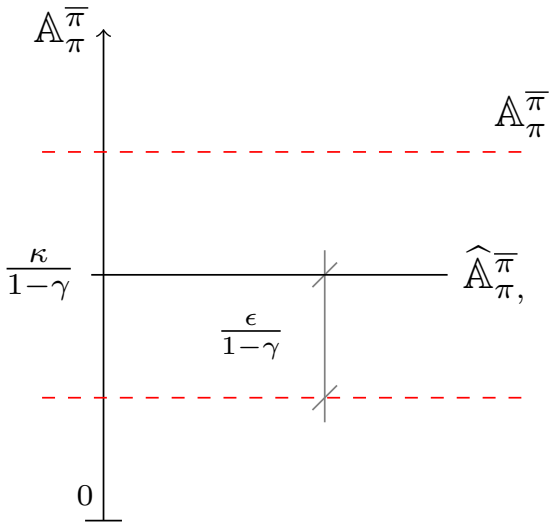


Figure 3: Average advantage estimation. The graph shows the relationship between the approximated and the real average advantage. At every iteration, the approximate average advantage of policy $\overline{\pi}$ satisfies the condition $\widehat{\mathbb{A}}_\pi^{\overline{\pi}} > \frac{\kappa}{1-\gamma}$, that is, $\mathbb{A}_\pi^{\overline{\pi}} > \frac{\kappa-\epsilon}{1-\gamma}$. When the algorithm terminates ($\widehat{\mathbb{A}}_\pi^{\overline{\pi}} \leq \frac{\kappa}{1-\gamma}$), the policy π returned by the algorithm guarantees by construction that, for all $\pi^\dagger \in \Pi^{\text{SR}}$, $\mathbb{A}_\pi^{\pi^\dagger} < \frac{\kappa+\epsilon}{1-\gamma}$ because $\overline{\pi}$ is the best policy in $\overline{\Pi}$.

5.2 approximate Unique–parameter Safe Policy Iteration (aUSPI)

In this section, we discuss how to deal with approximation when considering USPI. Let us start with some considerations about the requirements we need to enforce on the estimates. In the exact case, the USPI algorithm stops when the advantage function becomes zero. When approximation is involved, we need to take into account also the accuracy level ϵ in order to define a proper threshold. Suppose we are provided with an $\frac{\epsilon}{1-\gamma}$–accurate approximation of $\mathbb{A}_\pi^{\overline{\pi}}$, which threshold should we use for the termination condition of the SPI algorithms? If we use as threshold the value $\frac{\kappa}{1-\gamma}$, for some $\kappa > 0$, we have that: i) the algorithm stops when: $\mathbb{A}_\pi^{\overline{\pi}} \leq \frac{\kappa+\epsilon}{1-\gamma}$; ii) the algorithm continues when $\mathbb{A}_\pi^{\overline{\pi}} \geq \frac{\kappa-\epsilon}{1-\gamma}$ (Figure 3). For sure, we require that $\kappa - \epsilon \geq 0$ because we want to avoid following target policies with a negative advantage. We select $\kappa = \epsilon$.

With this assumption and using the bound in Equation (14), Corollary 5 can be restated in approximate setting as follows.

Corollary 16. *Assume to have an $\frac{\epsilon}{1-\gamma}$-accurate $\mathbb{A}_\pi^{\overline{\pi}}$, named $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$. If $\widehat{\mathbb{A}}_\pi^{\overline{\pi}} \geq \frac{\epsilon}{1-\gamma}$, then, using*

$$\alpha^* = \frac{(1-\gamma)^2 \left(\widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \frac{\epsilon}{1-\gamma}\right)}{\gamma \left\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\right\|_\infty^2}$$

we set $\alpha = \min(1, \alpha^*)$, so that when $\alpha^* \leq 1$ we can guarantee the following policy improvement:

Algorithm 9 Approximate USPI.

input: target policy space $\overline{\Pi}$, approximate policy chooser $\widehat{\mathrm{PC}}$, accuracy ϵ
Initialize π
$\overline{\pi} \leftarrow \widehat{\mathrm{PC}}(\overline{\Pi}, \pi)$
while $\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} \geq \frac{\epsilon}{1-\gamma}$ **do**
  $\alpha \leftarrow \min\left\{1, \frac{(1-\gamma)^2\left(\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} - \frac{\epsilon}{1-\gamma}\right)}{\gamma\|\overline{\boldsymbol{\pi}}-\boldsymbol{\pi}\|_{\infty}^2}\right\}$
  $\boldsymbol{\pi} \leftarrow \alpha\overline{\boldsymbol{\pi}} + (1-\alpha)\boldsymbol{\pi}$
  $\overline{\pi} \leftarrow \widehat{\mathrm{PC}}(\overline{\Pi}, \pi)$
end while
return $\boldsymbol{\pi}$

$$J^{\pi'} - J^{\pi} \geq \frac{(1-\gamma)\left(\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} - \frac{\epsilon}{1-\gamma}\right)^2}{2\gamma\,\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{\infty}^2},$$

and when $\alpha^* > 1$, we perform a full update towards the target policy $\overline{\pi}$ with a policy improvement equal to the one specified in Equation (14)*.*

Proof Let us consider the single component derivations in the proof of Theorem 5. The bound in Equation (14) can be rewritten as:

$$J^{\pi'} - J^{\pi} \geq \frac{\alpha}{1-\gamma}\mathbb{A}_{\pi}^{\overline{\pi}} - \alpha^2\frac{\gamma}{2(1-\gamma)^3}\,\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{\infty}^2\,. \tag{P.12}$$

Suppose now to have access to an $\frac{\epsilon}{1-\gamma}$-accurate estimation of $\mathbb{A}_{\pi}^{\overline{\pi}}$, called $\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}}$. We can exploit such assumption to derive the optimal α and the guaranteed policy performance improvement. We set in the worst-case scenario by considering

$$\mathbb{A}_{\pi}^{\overline{\pi}} = \widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} - \frac{\epsilon}{1-\gamma}, \tag{P.13}$$

that is, we suppose to have an overestimate of $\mathbb{A}_{\pi}^{\overline{\pi}}$ with maximum error. By replacing (P.13) in the bound (P.12) we obtain:

$$J^{\pi'} - J^{\pi} \geq \frac{\alpha}{1-\gamma}\left(\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}} - \frac{\epsilon}{1-\gamma}\right) - \alpha^2\frac{\gamma}{2(1-\gamma)^3}\,\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{\infty}^2\,. \tag{P.14}$$

The term α^* is the value of α that maximizes this bound, i.e. the value that set the partial derivative w.r.t. α to zero. By putting α^* in place of α in the last bound, we derive the guaranteed performance improvement. ■

The general algorithm for the approximated version of USPI, called Approximate USPI (aUSPI), is described in Algorithm 9. The skeleton of the algorithm is very similar to the exact one: starting with an initial policy π at random, iteratively select a target policy $\overline{\pi}$ and compute a conservative update toward it. The trade-off coefficient α is the one

that maximizes the lower bound in Equation 14, i.e., $\alpha = \frac{(1-\gamma)^2(\widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \frac{\epsilon}{1-\gamma})}{\gamma \|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty^2}$. The procedure continues until the average advantage $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ remains greater than the threshold $\frac{\epsilon}{1-\gamma}$. What changes with respect to the exact case is that the quantities involved in the algorithm are approximated. For this reason, both the update rule and the terminal condition are arranged to consider the worst case, i.e., the $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ is decreased by the maximal estimation error $\frac{\epsilon}{1-\gamma}$. We can now show that using an adaptive accuracy and threshold, aUSPI, when selecting the greedy policy as target ($\widehat{\text{GPC}}$), converges to the global optimum.

Theorem 17 (Finite Convergence)**.** *There exists a constant $\eta > 0$ depending on the MDP and a value of accuracy ϵ, such that under Assumptions 1 and 2, aUSPI with $\widehat{GPC}$ with condition $\widehat{\mathbb{A}}_\pi^{\pi^+} > \frac{\kappa}{1-\gamma}$ for keeping updating the policy, with approximations $\left|\mathbb{A}_\pi^{\pi^+} - \widehat{\mathbb{A}}_\pi^{\pi^+}\right| < \frac{\epsilon}{1-\gamma}$ converges to the optimal policy π^* in a finite number of steps when using as $\kappa = \epsilon + \eta \|\boldsymbol{\pi}^+ - \boldsymbol{\pi}\|_\infty$.*

Proof The idea of the proof is to show that, with a suitable choice of ϵ and η, the algorithm does not stop until the optimal policy is reached. Since $\widehat{\mathbb{A}}_\pi^{\pi^+}$ is an $\frac{\epsilon}{1-\gamma}$-accurate estimation of $\mathbb{A}_\pi^{\pi^+}$ and the algorithm stops as soon as $\widehat{\mathbb{A}}_\pi^{\pi^+} \leq \frac{\kappa}{1-\gamma}$, we have that at each iteration the performance improvement can be lower bounded by a finite quantity, unless $\pi = \pi^+$, i.e., unless we have reached the optimal policy:

$$J^{\pi'} - J^\pi \geq \frac{(1-\gamma)\left(\widehat{\mathbb{A}}_\pi^{\pi^+} - \frac{\epsilon}{1-\gamma}\right)^2}{2\gamma \left\|\boldsymbol{\pi}^+ - \boldsymbol{\pi}\right\|_\infty^2} \geq \frac{(1-\gamma)\left(\frac{\kappa-\epsilon}{1-\gamma}\right)^2}{2\gamma \left\|\boldsymbol{\pi}^+ - \boldsymbol{\pi}\right\|_\infty^2} = \frac{\eta^2}{2\gamma(1-\gamma)} > 0, \tag{P.15}$$

where the first inequality comes from Corollary 16, the second inequality from the condition $\widehat{\mathbb{A}}_\pi^{\pi^+} > \frac{\kappa}{1-\gamma}$, and the third inequality from the definition of $\kappa = \epsilon + \eta \|\boldsymbol{\pi}^+ - \boldsymbol{\pi}\|_\infty$. Recalling that $J^\pi = \frac{1}{1-\gamma}\mathbf{d}^{\pi\intercal}\mathbf{r}^\pi$ and $\|\mathbf{r}^\pi\|_\infty \leq 1$ since $\mathcal{R}(s,a) \in [0,1]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have:

$$J^{\pi'} - J^\pi \leq \frac{1}{1-\gamma}\left(\mathbf{d}^{\pi'} - \mathbf{d}^\pi\right)^\intercal \mathbf{r}^\pi \leq \frac{1}{1-\gamma}\left\|\mathbf{d}^{\pi'} - \mathbf{d}^\pi\right\|_1 \leq \frac{\gamma}{(1-\gamma)^2}\left\|\boldsymbol{\pi}' - \boldsymbol{\pi}\right\|_\infty. \tag{P.16}$$

From the previous two inequalities, if $\pi \neq \pi^*$, we can immediately lower bound the distance between the greedy policy π^+ and the current policy π, i.e., the minimum distance between two consecutive policies visited by the algorithm:

$$\frac{\eta^2}{2\gamma(1-\gamma)} \leq J^{\pi'} - J^\pi \leq \frac{\gamma}{(1-\gamma)^2}\left\|\boldsymbol{\pi}' - \boldsymbol{\pi}\right\|_\infty \implies \left\|\boldsymbol{\pi}' - \boldsymbol{\pi}\right\|_\infty \geq \frac{(1-\gamma)\eta^2}{2} > 0. \tag{P.17}$$

The immediate consequence is that we will always make finite jumps in the space of policies, similarly to what happens with performance. At each iteration, the algorithm achieves a finite performance improvement, it will stop in a finite number of iterations N.[8]

Let us recall that the algorithm terminates $\mathbb{A}_\pi^{\pi^+} \leq \widehat{\mathbb{A}}_\pi^{\pi^+} + \frac{\epsilon}{1-\gamma} \leq \frac{\kappa+\epsilon}{1-\gamma}$. Therefore, by using Lemma 7 and recalling the definition of κ, we can state the following condition that must hold at iteration N:

$$(1-\gamma)\Delta_d\left(J^* - J^{\pi_N}\right) \leq \mathbb{A}_{\pi_N}^{\pi^+} \leq \frac{2\epsilon + \eta\left\|\boldsymbol{\pi}^+ - \boldsymbol{\pi}_{\boldsymbol{N}}\right\|_\infty}{(1-\gamma)} \implies J^* - J^{\pi_N} \leq \frac{2(\eta+\epsilon)}{(1-\gamma)^2\Delta_d}, \tag{P.18}$$

8. We are not interested, in this phase, in bounding the number of iterations N.

having observed that $\|\boldsymbol{\pi}^+ - \boldsymbol{\pi}_N\|_\infty \leq 2$. Up to now, we have proven that our algorithm stops in a finite number of iterations with a performance gap w.r.t. the optimal performance of at most $\frac{2(\eta+\epsilon)}{(1-\gamma)^2\Delta_d}$. We will prove that, under certain conditions on η and ϵ we can also achieve the finite convergence to the optimal policy. For this purpose, we need first to guarantee that our algorithm will not stop before having exceeded the performance gap between the optimal policy and the second-best policy Δ_J (see Lemma 9):

$$\frac{2(\eta+\epsilon)}{(1-\gamma)^2\Delta_d} \leq \Delta_J \implies \eta \leq \frac{1}{2}(1-\gamma)^2\Delta_d\Delta_J - \epsilon. \tag{P.19}$$

It remains to prove is that the algorithm will not stop until it reaches the optimal policy. Thus, it must happen that our threshold $\frac{\kappa}{1-\gamma}$ is always larger than the minimum advantage we can see during the path. As we know, from Lemma 10, that $\mathbb{A}_\pi^{\pi^*} \geq \frac{\Delta_d\Delta_+}{2}\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty$ we consider the condition:

$$\frac{\Delta_d\Delta_+}{2}\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty \geq \frac{1}{1-\gamma}\left(\epsilon + \eta\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty\right) \implies \eta \leq \frac{1}{2}(1-\gamma)\Delta_d\Delta_+ - \frac{\epsilon}{\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty}.$$

By applying Equation (P.17), we enforce the stricter condition on η:

$$\eta \leq \frac{1}{2}(1-\gamma)\Delta_d\Delta_+ - \frac{2\epsilon}{(1-\gamma)\eta^2} \tag{P.20}$$

In order to complete the proof of the theorem we need to show that there exist $\eta > 0$ and $\epsilon > 0$ so that they satisfy the conditions at Equations (P.19) and (P.20). To this purpose, we make the following choices of $\epsilon > 0$ and $\eta > 0$:

$$\eta = \frac{1}{4}(1-\gamma)^2\Delta_d \min\{\Delta_J, \Delta_+\},$$
$$\epsilon = \frac{\eta^3}{2}.$$

It is immediate to prove that they fulfill both Equations (P.19) and (P.20). Concerning Equation (P.19), we have:

$$\begin{aligned}\eta = \frac{1}{4}(1-\gamma)^2\Delta_d \min\{\Delta_J, \Delta_+\} &\leq \frac{1}{2}(1-\gamma)^2\Delta_d\Delta_J - \frac{1}{4}(1-\gamma)^2\Delta_d \min\{\Delta_J, \Delta_+\} \\ &\leq \frac{1}{2}(1-\gamma)^2\Delta_d\Delta_J - \frac{1}{2}\left(\frac{1}{4}(1-\gamma)^2\Delta_d \min\{\Delta_J, \Delta_+\}\right)^3 \\ &= \frac{1}{2}(1-\gamma)^2\Delta_d\Delta_J - \frac{\eta^3}{2} = \frac{1}{2}(1-\gamma)^2\Delta_d\Delta_J - \epsilon,\end{aligned}$$

having observed that $\frac{1}{4}(1-\gamma)^2\Delta_d \min\{\Delta_J, \Delta_+\} < 1$. Regarding Equation (P.20), instead, we have:

$$\begin{aligned}\eta = \frac{1}{4}(1-\gamma)^2\Delta_d \min\{\Delta_J, \Delta_+\} &\leq \frac{1}{4}(1-\gamma)\Delta_d \min\{\Delta_J, \Delta_+\} \\ &\leq \frac{1}{2}(1-\gamma)\Delta_d\Delta_+ - \frac{1}{4}(1-\gamma)\Delta_d \min\{\Delta_J, \Delta_+\}\end{aligned}$$

$$= \frac{1}{2}(1-\gamma)\Delta_d\Delta_+ - \frac{\eta}{1-\gamma}$$
$$= \frac{1}{2}(1-\gamma)\Delta_d\Delta_+ - \frac{2\epsilon}{(1-\gamma)\eta^2}.$$

■

Theorem 17 proves that it is possible to set a non-zero value of ϵ and η such that aUSPI converges to the optimal policy in a finite number of steps. Unfortunately, those values depend on some unknown quantities (Δ_d, Δ_J and Δ_+) related, informally, to the "difficulty" of the task.[9] Previous literature (e.g., Kakade and Langford, 2002; Scherrer, 2014) typically analyzes the convergence to an *approximately optimal* policy in terms of the error ϵ of the estimation of the relevant quantities (e.g., the advantage function). To the best of our knowledge, this is the first analysis that proves a finite convergence to the *optimal policy*, even admitting a positive error ϵ.

It is worth noting that when setting $\epsilon = 0$, we reduce to the exact case. Therefore, this convergence result, which exploits a termination condition different from Theorem 17, holds also for the exact case, considering $\mathbb{A}_\pi^{\overline{\pi}} \leq \frac{\kappa}{1-\gamma}$ as condition to keep updating the policy, with $\kappa = \eta \left\| \boldsymbol{\pi}^+ - \boldsymbol{\pi} \right\|_\infty$.

aUSPI Sampling Procedure We now show how to obtain an $\frac{\epsilon}{1-\gamma}$-accurate estimate of $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$. The sampling procedure is very similar to that presented for the approximate GPC (GPC Sampling Procedure). We generate a sequence of N state-action pairs (s_i, a_i), where $s_i \sim d^\pi$ and $a_i \sim \pi(\cdot|s_i)$. For each of them, we generate a single rollout of length T, which is used to estimate the $Q^\pi(s_i, a_i)$ (Algorithm 15). In the end, we get a dataset $\{(s_i, a_i, \widehat{q}_i)\}_{i=1}^N$, that can be used to estimate the expected advantage function as:

$$\widehat{\mathbb{A}}_\pi^{\overline{\pi}} = \frac{1}{N}\sum_{i=1}^{N} \left(\overline{\pi}(a_i|s_i) - \pi(a_i|s_i)\right) \widehat{q}_i. \tag{17}$$

The following result provides the length T and the number of samples N to obtain an $\frac{\epsilon}{1-\gamma}$-accurate estimate.

Lemma 18. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_\gamma \frac{\epsilon}{12} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{16}{9\epsilon^2} \log \frac{2}{\delta} \right\rceil.$$

The aUSPI Sampling Procedure construct a function $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ such that with probability $1-\delta$:

$$\left| \widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \mathbb{A}_\pi^{\overline{\pi}} \right| \leq \frac{\epsilon}{1-\gamma}.$$

9. Δ_d is the minimum (discounted) probability with which a state is visited under any policy and it is related to the ergodicity properties of the MDP. Δ_J is the minimum performance gap between an optimal policy and any other deterministic suboptimal policy. Finally, Δ_+ is defined in Lemma 25 and it is somehow related to the action gap.

Remark 11 (On the Sample Complexity)**.** Let us make some considerations on these results. Provided that a target policy $\overline{\pi}$ is given, aUSPI needs $N = \lceil \frac{16}{9\epsilon^2} \log \frac{2}{\delta} \rceil$ trajectories each of length $T = \lceil \log_\gamma \frac{\epsilon}{12} \rceil$ to estimate $\mathbb{A}_\pi^{\overline{\pi}}$ with an $\frac{\epsilon}{1-\gamma}$ accuracy level. Therefore, the total number of samples needed in a single iteration is NT. We call this quantity *per–iteration sample complexity*. Instead, if the target policy has to be computed as well, we need to add the samples needed by the approximate policy chooser. Clearly, this number of samples provides a theoretical guarantee for the single iteration only. If the algorithm is run for I iterations, then a straightforward way to make the guarantee hold for all iterations is to rescale the confidence level by I, i.e., for all iterations $i = 1, 2, ..., I$ we set $\delta_i = \frac{\delta}{I}$. Instead, if we seek for asymptotic convergence, and thus I might be infinite, we select a non–uniform schedule for δ. The classical approach consists in setting $\delta_i = \frac{6\delta}{\pi^2 i^2}$, so that $\sum_{i=1}^{+\infty} \delta_i = \delta$.

5.3 approximate per–State–parameter Safe Policy Iteration (aSSPI)

As we have seen in Section 4.2, the optimal learning steps $\alpha(s)$ for the SSPI cannot be computed in closed form and an iterative procedure must be carried out. The iterative procedure aims at maximizing the bound in Corollary 12. Nevertheless, differently from aUSPI, we need to estimate two elements: the set of states with non-negative advantage $\mathcal{S}_\pi^{\overline{\pi}}$ and the derivative of the bound as in Equation (6). In this section, we derive the number of samples needed so that, in high probability, aSSPI performs a policy update that improves the performance. First of all, we observe that any subset of $\mathcal{S}_\pi^{\overline{\pi}}$ is a conservative choice, as we would avoid updating states with a positive advantage. Therefore, we seek an estimation of a subset of $\mathcal{S}_\pi^{\overline{\pi}}$. For this purpose, we need to have an estimation of the advantage $A_\pi^{\overline{\pi}}(s)$ in every state $s \in \mathcal{S}$. If we are able to provide an $\frac{\epsilon}{2(1-\gamma)}$-accurate[10] estimation $\widehat{A}_\pi^{\overline{\pi}}(s)$ of $A_\pi^{\overline{\pi}}(s)$, then we know that $\widehat{A}_\pi^{\overline{\pi}} - \frac{\epsilon}{2(1-\gamma)}$ is a lower bound of $A_\pi^{\overline{\pi}}$. If we require $\widehat{A}_\pi^{\overline{\pi}} \geq \frac{\epsilon}{2(1-\gamma)}$ we are guaranteed that $A_\pi^{\overline{\pi}}$ is non-negative. Therefore, we define the approximate set as:

$$\widehat{\mathcal{S}}_\pi^{\overline{\pi}} = \left\{ s \in \mathcal{S} : \widehat{A}_\pi^{\overline{\pi}}(s) \geq \frac{\epsilon}{2(1-\gamma)} \right\}.$$

In this way, we can guarantee that $\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \subseteq \mathcal{S}_\pi^{\overline{\pi}}$. We now discuss how to obtain an $\frac{\epsilon}{2(1-\gamma)}$-accurate estimation of $A_\pi^{\overline{\pi}}(s)$.

aSSPI Sampling Procedure For each state $s \in \mathcal{S}$, we generate N trajectories of length T for each action $a \in \mathcal{A}$, getting a set of independent estimates of $Q^\pi(s,a)$ for all $a \in \mathcal{A}$, i.e., $\left\{ \{(a, \widehat{q}_i(s,a))\}_{a \in \mathcal{A}} \right\}_{i=1}^N$ (Algorithm 16). Then we estimate the advantage function as:

$$\widehat{A}_\pi^{\overline{\pi}}(s) = \frac{1}{N} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} \left(\overline{\pi}(a|s) - \pi(a|s) \right) \widehat{q}_i(s,a). \tag{18}$$

We now provide the following PAC-bound on the number of samples N needed and provide a value for T.

Lemma 19. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_\gamma \frac{\epsilon}{8} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{128}{9\epsilon^2} \log \frac{2|\mathcal{S}|}{\delta} \right\rceil.$$

10. The choice of $\frac{\epsilon}{2(1-\gamma)}$ as accuracy threshold will be clarified later.

The aSSPI Sampling Procedure constructs a function $A_\pi^{\overline{\pi}}(s)$ such that with probability $1-\delta$, simultaneously for all $s \in \mathcal{S}$:

$$\left|\widehat{A}_\pi^{\overline{\pi}}(s) - A_\pi^{\overline{\pi}}(s)\right| \leq \frac{\epsilon}{2(1-\gamma)}.$$

Once we have an estimation of $\mathcal{S}_\pi^{\overline{\pi}}$ we can use it to compute the derivative of the bound. Note that in this case, we want to optimize a statistical lower bound of the derivative. To avoid estimating also $\|\mathbf{q}^\pi\|_\infty$ from samples, we consider the slope equal to $m = -\frac{\gamma}{(1-\gamma)^3}$. Since the derivative is a piece-wise linear decreasing function, a lower bound would intersect the zero axes at a smaller value of Υ, allowing a smaller budget and being, therefore, more conservative. Since we already estimated the advantage functions for each state $\widehat{A}_\pi^{\overline{\pi}}(s)$, we can reuse them for the derivative estimation. Suppose we have at our disposal a set of M states $\{s_i\}_{i=1}^M$ sampled independently from d^π, we can estimate the constant term in the bound derivative as:

$$\widehat{g} = \frac{1}{M(1-\gamma)} \sum_{i=1}^{M} I\left(s_i \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s_i)}{\|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1}. \tag{19}$$

We now provide a value for M to guarantee an $\frac{\epsilon}{(1-\gamma)^2}$–accurate estimate.[11]

Lemma 20. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$M = \left\lceil \frac{8}{\epsilon^2} \log \frac{2}{\delta} \right\rceil.$$

Under the assumptions of Lemma 19, with probability $1-2\delta$ it holds that:

$$\widehat{g} - g \leq \frac{\epsilon}{(1-\gamma)^2}.$$

The aSSPI Sampling Procedure and Lemma 20 allow us to compute the per–iteration sample complexity. Indeed, we sample N trajectories for every state and action pair, each of length T, in order to have an estimate of $A_\pi^{\overline{\pi}}(s)$ for all $s \in \mathcal{S}$. Then, we need M samples to compute g, taking a sample mean under d^π. Thus, overall we need $NT|\mathcal{A}||\mathcal{S}| + M$ samples. Algorithm 10 reports the pseudocode of the FJP function for aSSPI.

5.4 approximate per–State–Action–parameter Safe Policy Iteration (aSASPI)

Similarly to the case of SSPI, it is necessary to carry out an iterative procedure in order to optimize the bound in Corollary 13. In this section, we derive the number of samples needed so that aSASPI moves the policy in a direction that improves the performance. Differently from aSSPI, we have several elements we have to estimate: i) the Q-function for every state–action pair; ii) the set of non–saturated states $\mathcal{S}_\pi^{\overline{\pi}}$; iii) the active actions $a_s^\uparrow$ and $a_s^\downarrow$ for each non–saturated state; iv) the expectation under d^π to compute the bound derivative. We are going to partition the $\frac{\epsilon}{1-\gamma}$ accuracy over these four sources of error.

11. In this case, we seek for an $\frac{\epsilon}{(1-\gamma)^2}$–accurate estimate instead of an $\frac{\epsilon}{1-\gamma}$–accurate because g ranges in the interval $\left[0, \frac{1}{(1-\gamma)^2}\right]$.

Algorithm 10 Computing of the jump points for aSSPI (Find Jump Point - $\widehat{\text{FJP}}$)

input: current policy π, target policy $\overline{\pi}$, accuracy ϵ, confidence δ
initialize: $t \leftarrow 0$, $M \leftarrow \lceil \frac{8}{\epsilon^2} \log \frac{4}{\delta} \rceil$, $N \leftarrow \lceil \frac{128}{9\epsilon^2} \log \frac{4|\mathcal{S}|}{\delta} \rceil$, $T \leftarrow \lceil \log_\gamma \frac{\epsilon}{8} \rceil$, $\Upsilon_0 \leftarrow 0$
Compute an estimation of $A_\pi^{\overline{\pi}}(s)$ for all $s \in \mathcal{S}$ using N trajectories of length T
$\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \leftarrow \{s \in \mathcal{S} : \widehat{A}_\pi^{\overline{\pi}}(s) > \frac{\epsilon}{2(1-\gamma)}\}$
$q_0 \leftarrow \frac{1}{M(1-\gamma)} \sum_{i=1}^{M} I\left(s_i \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s_i)}{\|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1} - \frac{\epsilon}{(1-\gamma)^2}$
Sort states in $\widehat{\mathcal{S}}_\pi^{\overline{\pi}}$ so that $i < j \implies \|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1 \leq \|\overline{\pi}(\cdot|s_j) - \pi(\cdot|s_j)\|_1$
yield Υ_0, q_0
while $\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \neq \{\}$ **do**
  $t \leftarrow t + 1$
  $\Upsilon_t \leftarrow \|\overline{\pi}(\cdot|s_t) - \pi(\cdot|s_t)\|_1$
  $q_t \leftarrow q_{t-1} - \frac{1}{M(1-\gamma)} \sum_{i=1}^{M} I(s_i = s_t) \frac{\widehat{A}_\pi^{\overline{\pi}}(s_i)}{\|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1}$
  $\mathcal{S}_\pi^{\overline{\pi}} \leftarrow \mathcal{S}_\pi^{\overline{\pi}} \setminus \{s_t\}$
  yield Υ_t, q_t
end while
yield 2, $-\infty$

We start by showing how to estimate the Q-function, by using the following sampling procedure.

aSASPI Sampling Procedure For each state–action pair, we generate N trajectories of length T by executing policy π and for each of them, we compute the cumulative sum of the rewards. In this way we get a set of approximations $\left\{\left(s, a, \{\widehat{q}_i(s,a)\}_{i=1}^N\right)\right\}_{s\in\mathcal{S}, a\in\mathcal{A}}$ (Algorithm 17). Then we can estimate Q^π as:

$$\widehat{Q}^\pi(s,a) = \frac{1}{N} \sum_{i=1}^{N} \widehat{q}_i(s,a). \tag{20}$$

The following result provides the number of samples N and the rollout length T to have an $\frac{\epsilon}{12(1-\gamma)}$–accurate estimate of Q^π.

Lemma 21. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_\gamma \frac{\epsilon}{48} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{128}{\epsilon^2} \log \frac{2|\mathcal{S}||\mathcal{A}|}{\delta} \right\rceil.$$

The sampling procedure constructs a function $\widehat{Q}_\pi(s,a)$ such that with probability $1-\delta$, simultaneously for all $s \in \mathcal{S}$:

$$\left|\widehat{Q}^\pi(s,a) - Q^\pi(s,a)\right| \leq \frac{\epsilon}{12(1-\gamma)}$$

Given the $\frac{\epsilon}{12(1-\gamma)}$–accurate estimation of $\widehat{Q}^\pi$ we can state some consideration about the other quantities we have to estimate. First, given a budget Υ, consider the active actions $a_s^\uparrow$ and $a_s^\downarrow$ and the corresponding approximations $\widehat{a}_s^\uparrow$ and $\widehat{a}_s^\downarrow$ induced by $\widehat{Q}(s,a)$. Let us focus on $a_s^\uparrow$. If $a_s^\uparrow \neq \widehat{a}_s^\uparrow$ it means that we performed an incorrect ordering of the actions but this

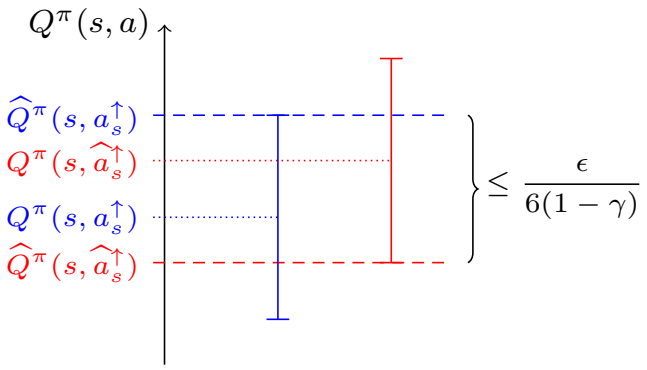


Figure 4: Representation of the configuration of the confidence intervals for the estimated Q^π. When $a_s^\uparrow \neq \widehat{a}_s^\uparrow$ it must be that we have misestimated the value functions. Precisely, we have $Q^\pi(s,a_s^\uparrow) > Q^\pi(s,\widehat{a}_s^\uparrow)$ but $\widehat{Q}^\pi(s,a_s^\uparrow) < \widehat{Q}^\pi(s,\widehat{a}_s^\uparrow)$. Thus, the maximum error is given by the length of the interval, i.e., $\frac{\epsilon}{6(1-\gamma)}$.

happens only when the confidence intervals of $\widehat{Q}(s,a_s^\uparrow)$ and $\widehat{Q}(s,\widehat{a}_s^\uparrow)$ overlap. Thanks to Lemma 21, for every state $s \in \mathcal{S}$ and budget Υ we have:

$$\left|\widehat{Q}^\pi(s,a_s^\uparrow) - \widehat{Q}^\pi(s,\widehat{a}_s^\uparrow)\right| \leq \frac{\epsilon}{6(1-\gamma)}, \tag{21}$$

since two intervals cannot overlap for more than $\frac{\epsilon}{6(1-\gamma)}$, i.e., the length of the confidence interval (see Figure 4). The same holds for $a_s^\downarrow$. We now define the set of active states as:

$$\widehat{\mathcal{S}}_\pi^{\overline{\pi}} = \left\{ s \in \mathcal{S} : \widehat{Q}^\pi(s,\widehat{a}_s^\uparrow) - \widehat{Q}^\pi(s,\widehat{a}_s^\downarrow) \geq \frac{\epsilon}{2(1-\gamma)} \right\}.$$

Thus, in order to remove a state s from $\widehat{\mathcal{S}}_\pi^{\overline{\pi}}$ we require $\widehat{Q}^\pi(s,\widehat{a}_s^\uparrow) - \widehat{Q}^\pi(s,\widehat{a}_s^\downarrow) \leq \frac{\epsilon}{2(1-\gamma)}$. Indeed, with this requirement, we can ensure that:

$$\begin{aligned}
\left|\widehat{Q}^\pi(s,\widehat{a}_s^\uparrow) - \widehat{Q}^\pi(s,\widehat{a}_s^\downarrow)\right| &\leq \left|\widehat{Q}^\pi(s,a_s^\uparrow) - \widehat{Q}^\pi(s,a_s^\downarrow)\right| + \frac{\epsilon}{3(1-\gamma)} \\
&\leq \left|Q^\pi(s,a_s^\uparrow) - Q^\pi(s,a_s^\downarrow)\right| + \frac{\epsilon}{6(1-\gamma)} + \frac{\epsilon}{3(1-\gamma)} \\
&= \left|Q^\pi(s,a_s^\uparrow) - Q^\pi(s,a_s^\downarrow)\right| + \frac{\epsilon}{2(1-\gamma)}
\end{aligned} \tag{22}$$

Once we have all these guarantees, all it takes is to estimate the constant term in the derivative of the bound:

$$\widehat{g} = \frac{1}{M(1-\gamma)} \sum_{i=1}^{M} I\left(s_i \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \left(\widehat{Q}^\pi(s_i,\widehat{a}_{s_i}^\uparrow) - \widehat{Q}^\pi(s_i,\widehat{a}_{s_i}^\downarrow)\right). \tag{23}$$

Approach	N	M	T	Per–iteration sample complexity
aUSPI	$\lceil \frac{16}{9\epsilon^2} \log \frac{2}{\delta} \rceil$	-	$\lceil \log_\gamma \frac{\epsilon}{12} \rceil$	NT
aSSPI	$\lceil \frac{128}{9\epsilon^2} \log \frac{4\lvert\mathcal{S}\rvert}{\delta} \rceil$	$\lceil \frac{8}{\epsilon^2} \log \frac{4}{\delta} \rceil$	$\lceil \log_\gamma \frac{\epsilon}{8} \rceil$	$NT\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert + M$
aSASPI	$\lceil \frac{128}{\epsilon^2} \log \frac{4\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\delta} \rceil$	$\lceil \frac{288}{121\epsilon^2} \log \frac{4}{\delta} \rceil$	$\lceil \log_\gamma \frac{\epsilon}{48} \rceil$	$NT\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert + M$

Table 1: Per–iteration sample complexity for the presented approximate versions of the SPI algorithms.

Lemma 22. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$M = \left\lceil \frac{288}{121\epsilon^2} \log \frac{2}{\delta} \right\rceil.$$

Under the assumptions of Lemma 19, the sampling procedure constructs a function $\widehat{g}$ such that with probability $1 - 2\delta$:

$$\widehat{g} - g \leq \frac{\epsilon}{(1-\gamma)^2}.$$

Algorithm 11 reports the pseudocode of the FJP function for aSASPI. The sampling procedure and Lemma 22 allow us to compute the per–iteration sample complexity. Indeed, we sample N trajectories for every state and action pair, each of length T, in order to have an estimate of $Q^\pi(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then, we need M samples to compute g, taking a sample mean under d^π. Thus, overall we need $NT|\mathcal{A}||\mathcal{S}| + M$ samples. Table 1 reports a comparison of the per–iteration sample complexity of the approximate SPI algorithms presented so far.

Remark 12 (Theoretical Guarantees Comparison)**.** The theoretical guarantees that we provide in this section aim at bounding the number of samples needed to ensure that, at each iteration, we perform a policy update that yields a performance improvement with high probability. These guarantees can be extended to multiple iterations, as done in Kakade (2003), to prove that the SPI algorithms stop with an expected average advantage that satisfies $\mathbb{A}_\pi^{\pi^+} \leq \frac{\kappa}{1-\gamma}$, where $\kappa > 0$ is a threshold. This kind of requirements is quite different compared to several employed in the literature (e.g., Azar et al., 2012, 2011; Sidford et al., 2018; Wainwright, 2019a,b), where the goal is to provide the conditions for which the learned policy is κ–optimal, leading to a bound on the difference of value functions $\|\mathbf{q}^* - \widehat{\mathbf{q}}\|$. The two conditions are clearly related (see for instance Lemma 7), but, we believe, not directly comparable. Furthermore, while for aUSPI we have provided the convergence to an optimal policy in a finite number of iterations (Theorem 17) under specific choices of the hyperparameters, we cannot directly extend this result for aSSPI and aSASPI. Unfortunately, without further specifications on the value of ϵ they could stop short of reaching the optimal policy.

6. Applications of SPI techniques

This section is devoted to the analysis of the performances of the algorithms in a discrete environment. The tests are mainly intended to give empirical supports to the theoretical

Algorithm 11 Computing of the jump points for aSASPI (FJP)

input: MDP $\mathcal{M}$, current policy π, target policy $\overline{\pi}$, accuracy ϵ, confidence δ
initialize: $t \leftarrow 0$, $M \leftarrow \lceil \frac{288}{121\epsilon^2} \log \frac{4}{\delta} \rceil$, $N \leftarrow \lceil \frac{128}{\epsilon^2} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta} \rceil$, $T \leftarrow \lceil \log_\gamma \frac{\epsilon}{48} \rceil$, $\frac{\Upsilon_0}{2} \leftarrow 0$, $i^\uparrow, i^\downarrow \leftarrow 0$
Compute $G^\uparrow(s,a)$ and $G^\downarrow(s,a)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$
Compute the two orderings $\rho^\uparrow$ and $\rho^\downarrow$ such that s.t. $i < j \implies G^\uparrow(s_{\rho_i^\uparrow}, a_{\rho_i^\uparrow}) \leq G^\uparrow(s_{\rho_j^\uparrow}, a_{\rho_j^\uparrow})$ and $i < j \implies G^\downarrow(s_{\rho_i^\downarrow}, a_{\rho_i^\downarrow}) \leq G^\uparrow(s_{\rho_j^\downarrow}, a_{\rho_j^\downarrow})$
Compute $a_s^\uparrow = \arg\max_{a \in \mathcal{A}} \left\{ \widehat{Q}^\pi(s,a) \right\}$ and $a_s^\downarrow = \arg\min_{a \in \mathcal{A}} \left\{ \widehat{Q}^\pi(s,a) \right\}$, $\forall s \in \mathcal{S}$
$\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \leftarrow \{ s \in \mathcal{S} : \widehat{Q}^\pi(s, a_s^\uparrow) - \widehat{Q}^\pi(s, a_s^\downarrow) > \frac{\epsilon}{2(1-\gamma)} \}$
$q_0 \leftarrow \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} \frac{1}{2} d^\pi(s) \left(Q^\pi(s, a_s^\uparrow) - Q^\pi(s, a_s^\downarrow) \right) - \frac{\epsilon}{1-\gamma}$
yield Υ_0, q_0
while $\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \neq \{\}$ **do**
  $t \leftarrow t + 1$
  if $s_t \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}$ **then**
    if $G^\uparrow(s_{\rho^\uparrow_{,i^\uparrow}} a_{\rho_{i^\uparrow}^\uparrow}) \leq G^\downarrow(s_{\rho_{i^\downarrow}^\downarrow}, a_{\rho_{i^\downarrow}^\downarrow})$ **then**
      $s_t, a_t \leftarrow s_{\rho_{i^\uparrow}^\uparrow}, a_{\rho_{i^\uparrow}^\uparrow}$
      if $\widehat{Q}^\pi(s_t, a_t) > \widehat{Q}^\pi(s_t, a_{s_t}^\downarrow) + \frac{\epsilon}{2(1-\gamma)}$ **then**
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(\widehat{Q}^\pi(s_t, a_{s_i t}^\uparrow) - \widehat{Q}^\pi(s_t, a_t) \right)$
        $a_{s_t}^\uparrow \leftarrow a_t$
      else
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(\widehat{Q}^\pi(s_t, a_{s_t}^\uparrow) - \widehat{Q}^\pi(s_t, a_{s_t}^\downarrow) \right)$
        $\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \setminus \{s_t\}$
      end if
      $\frac{\Upsilon_t}{2} \leftarrow G^\uparrow(s_t, a_t)$
      $i^\uparrow \leftarrow i^\uparrow + 1$
    else
      $s_t, a_t \leftarrow s_{\rho_{i^\downarrow}^\downarrow}, a_{\rho_{i^\downarrow}^\downarrow}$
      if $\widehat{Q}^\pi(s_t, a_t) < \widehat{Q}^\pi(s_t, a_{s_t}^\downarrow) - \frac{\epsilon}{2(1-\gamma)}$ **then**
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(\widehat{Q}^\pi(s_t, a_t) - \widehat{Q}^\pi(s_t, a_{s_t}^\downarrow) \right)$
        $a_{s_t}^\downarrow \leftarrow a_t$
      else
        $q_t \leftarrow q_{t-1} - \frac{1}{2} d^\pi(s_t) \left(\widehat{Q}^\pi(s_t, a_{s_t}^\uparrow) - \widehat{Q}^\pi(s_t, a_{s_t}^\downarrow) \right)$
        $\widehat{\mathcal{S}}_\pi^{\overline{\pi}} \setminus \{s_t\}$
      end if
      $\frac{\Upsilon_t}{2} \leftarrow G^\downarrow(s_t, a_t)$
      $i^\downarrow \leftarrow i^\downarrow + 1$
    end if
  end if
  yield Υ_t, q_t
end while
yield 2, $-\infty$

results derived in the previous section, illustrating the peculiar properties of the proposed algorithms. Moreover, the application offers some practical aspects of the implementation

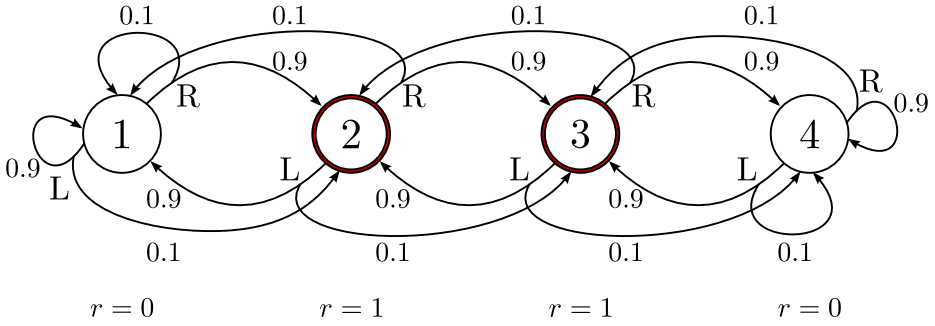


Figure 5: The problematic chain identified in Koller and Parr (2000). While states with double circle have a reward equal to 1, other states have null reward. Edges are decorated with the caption of the action ("left" or "right") and with the probability of success.

of the Safe Policy Iteration (SPI) methods and illustrates the competitiveness of the SPIs relative to the CPI.

We start considering exact settings and then we move to the approximate scenario. In the latter case, we consider two sources of approximation: 1) we assume to have access to a biased estimation of the action–value function $\widehat{Q}^{\pi}$ obtained through samples that induces an approximate estimation of the greedy target policy; 2) we assume to have access to a limited policy space that does not contain the optimal policy. In both the settings, results show the superiority of SPI approaches w.r.t. to CPI in terms of computational performances.

6.1 Chain-Walk Domain

We have chosen the chain walk problem (Lagoudakis and Parr, 2003a) for its simplicity that makes the comparison with other approaches straightforward and particular instructional. The chain walk domain is modeled as an N-state chain (numbered from 1 to N). The chain is traversed performing two actions, "left" (L) and "right" (R). Each action induces a transition into the associated direction and to the opposite one with probability p and $1 - p$ (in the following experiments, p is set to 0.9), respectively. Reward +1 is assigned only when the agent enters one of the two states located at a distance of $N/4$ from the boundaries, otherwise the reward is 0. The starting state distribution is assumed uniform over state space. None of the states is final and, thus, the chain can be potentially covered infinite times. The starting state distribution is unknown, but we consider having access to a reset distribution μ, which is assumed uniform over state space in any configuration. Figure 5 shows an interesting 4-states chain walk domain.

6.1.1 Experiments in the Exact Case

We start the analysis by considering the case in which no approximation is involved (so that $\overline{\pi} = \pi^{+}$). To give an idea of how the SPI algorithms work, in Figure 6, we compare their performance with the ones of policy iteration (PI) and conservative policy iteration (CPI) on a single run using a chain with 50 states and $\gamma = 0.9$. All the algorithms have been

initialized with the policy uniform over actions. The graph shows, for each algorithm, the value of J^π, the coefficient α (since SSPI and SASPI have several α coefficients, we plot the average of $\alpha(s)$ and $\alpha(s,a)$ respectively), the expected advantage function $\mathbb{A}_\pi^{\overline{\pi}}$ and the policy dissimilarity $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi}$ as a function of the number of iterations. As expected (since no approximation is involved), PI converges to the optimal policy in only one iteration. At the opposite end, CPI (whose convergence to the optimal policy is asymptotic) has a very slow performance improving rate when compared to the other algorithms. All SPI algorithms converge to the optimal policy in a finite number of iterations. USPI reaches the optimal policy in 44 iterations, while SSPI and SASPI take more than 100 iterations. It is worth noting that, since the chain domain has only two actions, the behavior of SSPI and SASPI is exactly the same, as mentioned in Section 4.3. The faster convergence of USPI w.r.t. SSPI and SASPI, although not theoretically proved, has been empirically verified in many different versions of the chain-walk domain obtained by varying the discount factor and the number of states. We can explain this behavior by recalling that USPI exploits a better bound w.r.t. the one of SSPI and SASPI, and, in the exact context, the advantage of choosing different convex combination coefficients for each state is not enough for SSPI and SASPI (at least in this domain) to attain the same improving rate of USPI. In order to obtain a fair comparison with SSPI and SASPI we considered a simplified version of USPI, USPI-simp, that exploits the same bound used by SSPI and SASPI. We can see that USPI-simp converges to the optimal policy in almost 300 iterations. Figure 6 also displays how the values of the convex combination coefficients change over the iterations for CPI, USPI, SSPI, SASPI, and USPI-simp. As expected, the value of α for CPI is always very low and decreases with iterations. On the other hand, the coefficients for the SPI algorithms start to increase when the current policy approaches the greedy one. This is also justified by the quick drop of the expected advantage and policy dissimilarity experimented by the SPI algorithms.

We further analyze the performance of the considered algorithms for different state-space dimensions (N) and for different values of the discount factor γ. The algorithms are tested over multiple runs, in particular 10 runs are performed starting from random policies. Figure 7 shows the behavior of the algorithms in terms of the distance between the performance of the policy at each iteration and the optimal performance (Error). It can be seen that CPI is always outperformed by the SPI algorithms. At the same time, the USPI achieves a significantly higher learning behavior than SSPI,[12] that leads to faster convergence to the optimal performance.

6.1.2 Experiments in the Approximate Case

To give a complete overview of the performance of the algorithms, we have moved to an approximate framework. We consider the error induced by the estimation of the value function via a set of samples. The dataset is built according to the technique explained in Section 5, using the theoretical values for the number of transitions to collect (i.e., N, T and M). In the approximate scenario, the improvement rate of Approximate SSPI (aSSPI) is always faster than the one of Approximate USPI (aUSPI). Figure 8 shows a comparison between aCPI, aUSPI, and aSSPI the 4-states chain walk domain, with 0.5 discount factor.

12. SASPI is not represented as its behavior is equal to that of SSPI.

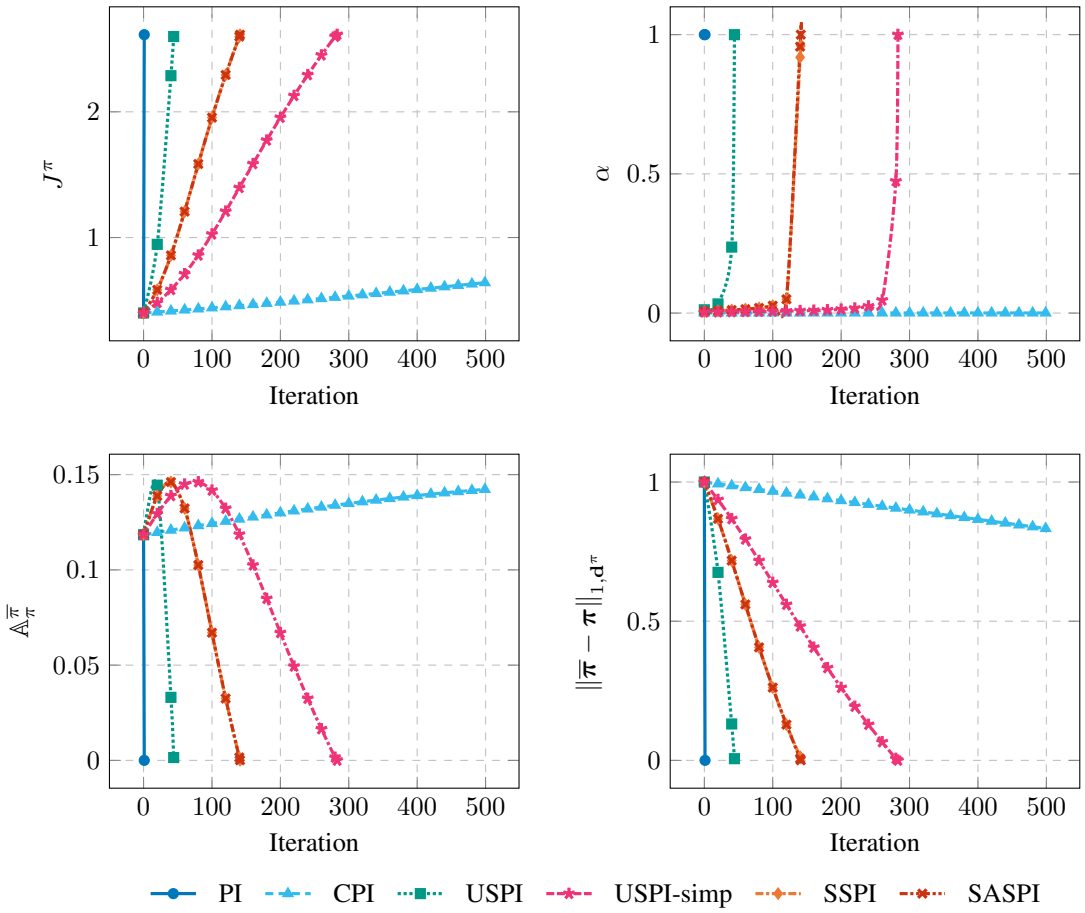


Figure 6: Score J^π, coefficient α, expected advantage $\mathbb{A}_\pi^{\overline{\pi}}$ and policy dissimilarity $\|\overline{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_{1,\mathbf{d}^\pi}$ as a function of iterations. The underline domain consists of a discounted (0.9) chain with 50 states.

All the algorithms have been initialized with the same starting policy (LLRR), that is, the complementary policy to the optimal one.

Each algorithm was run with two different values for the approximation error ϵ, i.e., 0.05 and 0.1, and with an estimate probability $\delta = 0.1$. The higher is the accuracy required (small values of ϵ), the larger is the number of iterations needed by the algorithms to converge. However, notice that the rate of improvement is higher for smaller values of ϵ. The reason is that low values of ϵ imply a more accurate estimate of the advantage function, thus allowing the algorithms to take larger update steps. Moreover, ϵ has a direct impact on the quality of the results (in terms of score J^π) because it controls the terminal condition. A high value of ϵ reflects in a wide terminal region, that is, few iterations and low performances. The advantages of small ϵ comes at the price of significantly increasing of the number of samples that at each iteration are used to obtain more accurate estimates of the Q-function. As

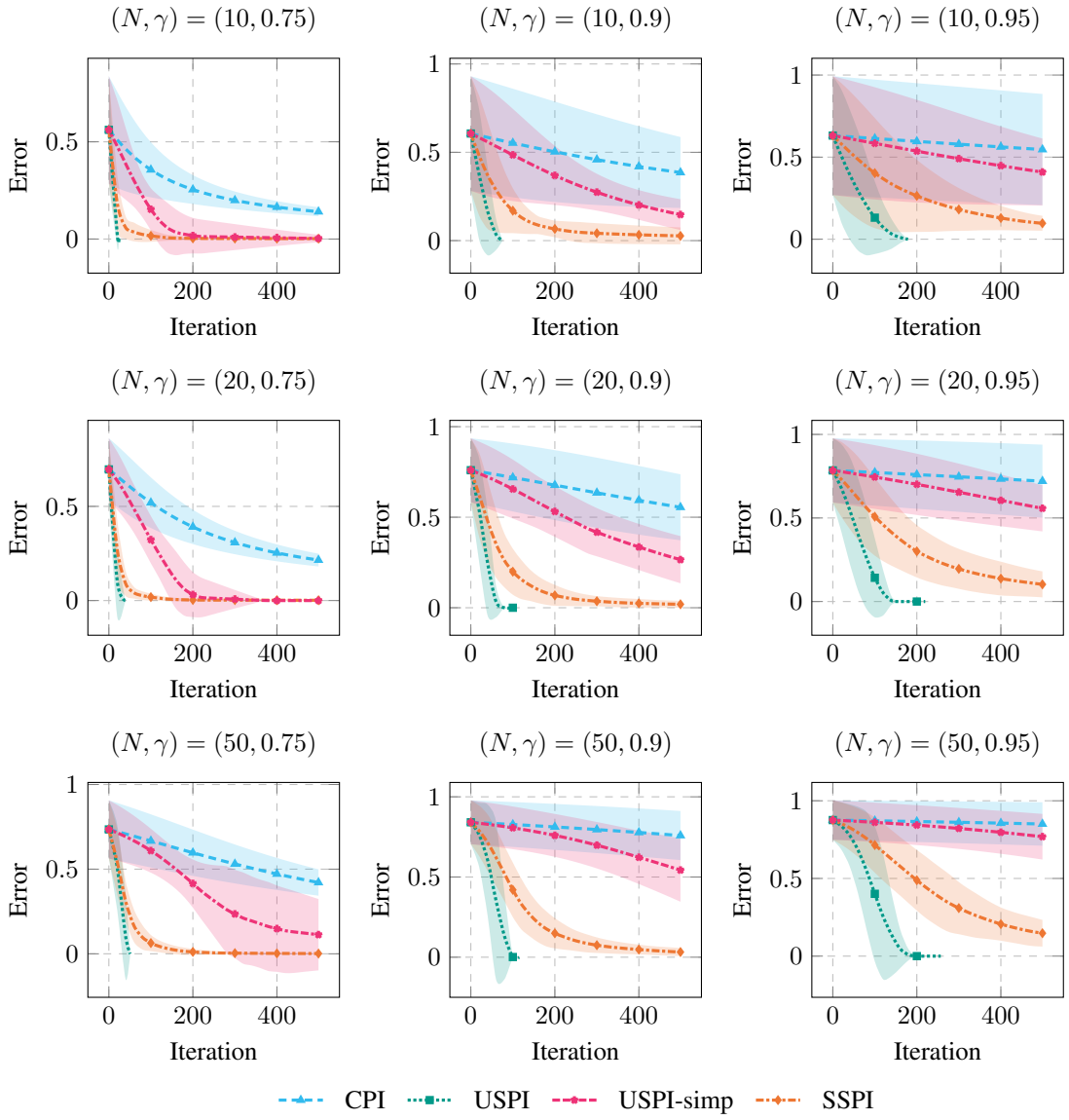


Figure 7: Error trend of policy w.r.t. the optimal performance J^π in different N-states chain walk domains with different γ values. 10 runs, 95% confidence intervals.

expected, aCPI takes much longer to converge w.r.t. both the approximated SPI algorithms, and aSSPI is faster than aUSPI.

These considerations are also supported by the results reported in Table 2, where for each algorithm, the average number of iterations (Table 2a) and the number of collected transitions (Table 2b) are reported for different values of ϵ in a 4-states domain, varying

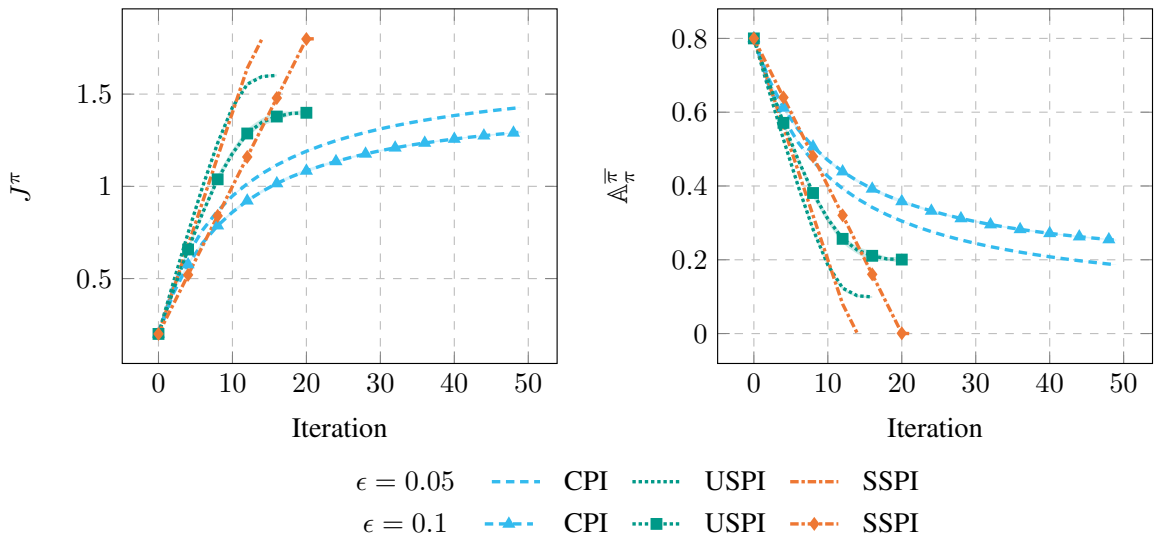


Figure 8: Score J^π and average advantage $\mathbb{A}_\pi^{\overline{\pi}}$ as a function of iterations for different values of the accuracy level ϵ. The underline domain consists of a discounted (0.5) chain with 4 states with fixed initial policy LLRR.10 runs, 95% confidence intervals.

also the discount factor (0.5 and 0.65).[13] Notice how quickly the number of iterations increases for aCPI as ϵ decreases, while for the approximated SPI algorithms such growth is significantly slower. Recall that, unlike the exact case, USPI and SSPI in the approximate scenario optimize the same lower bound on the policy performance (Section 5.2). Finally, it is worth noting that aSSPI is able to reach the optimal performance for all values of ϵ tested.

6.1.3 Experiments with Linear Approximations

To complete the analysis in the approximate framework, we have considered the problematic scenario obtained setting $N = 4$ (Koller and Parr, 2000). In this configuration, according to the previous definition, reward +1 is assigned only to the two central states, other states get 0. The reward vector is represented by (0, +1, +1, 0). When the distribution is assumed uniform over $\mathcal{S}$, the optimal policy is RRLL. In Koller and Parr (2000) it was shown that there exists a configuration where an approximate policy iteration, using an approximated state-value function for evaluation and an exact improvement, oscillates between two non-optimal policies. The mentioned configuration was obtained starting from the policy RRRR and representing the state-value function as a linear combination of the following 3 basis functions:

$$\phi(s) = \begin{pmatrix} 1 \\ s \\ s^2 \end{pmatrix}, \quad s \in \{1, 2, 3, 4\} \tag{24}$$

13. When using the theoretical values for the number of samples, higher values of γ require to reduce ϵ, making the number of transitions to collect for each iteration increase further.

			aCPI		aUSPI		aSSPI	
N	γ	ϵ	It	J^π	It	J^π	It	J^π
4	0.5	0.05	261.9 ± 20.32	1.593 ± 0.002	16.5 ± 0.5	1.601 ± 0.002	15.0 ± 0.0	1.8 ± 0.0
4	0.5	0.075	144.6 ± 19.845	1.482 ± 0.007	18.0 ± 1.183	1.497 ± 0.004	17.4 ± 0.49	1.8 ± 0.0
4	0.5	0.1	107.0 ± 9.37	1.378 ± 0.005	19.7 ± 1.1	1.399 ± 0.007	21.6 ± 0.49	1.8 ± 0.0
4	0.5	0.125	82.2 ± 10.458	1.271 ± 0.009	20.9 ± 1.64	1.293 ± 0.007	28.0 ± 0.0	1.8 ± 0.0
4	0.65	0.05	415.5 ± 27.156	2.142 ± 0.005	46.5 ± 2.291	2.156 ± 0.006	43.0 ± 0.0	2.571 ± 0.0
4	0.65	0.075	237.0 ± 22.996	1.915 ± 0.011	51.8 ± 4.423	1.945 ± 0.016	59.0 ± 0.0	2.571 ± 0.0
4	0.65	0.1	151.2 ± 17.503	1.679 ± 0.026	50.8 ± 3.124	1.72 ± 0.017	94.0 ± 0.0	2.571 ± 0.0
4	0.65	0.125	118.6 ± 16.62	1.466 ± 0.029	49.0 ± 2.408	1.491 ± 0.023	248.9 ± 0.7	2.571 ± 0.0

(a) Algorithm iterations and performances (sample mean ± standard deviation of the mean estimation).

N	γ	ϵ	aCPI	aUSPI	aSSPI
4	0.5	0.05	1936144	1936144	13266165
4	0.5	0.075	802175	802175	5427302
4	0.5	0.1	434812	434812	2863348
4	0.5	0.125	262550	262550	1742839
4	0.65	0.05	1936144	1936144	13266165
4	0.65	0.075	802175	802175	5427302
4	0.65	0.1	434812	434812	2863348
4	0.65	0.125	262550	262550	1742839

(b) Number of transitions collected per iteration.

Table 2: Results in approximate settings for the 4-states chain walk averaged over 10 runs for all the algorithms. Initial policies were stochastic policies chosen at random.

Approximated state-value function was computed using weighted least square where weights were represented by the stationary distribution. Starting from policy RRRR, policy iteration oscillates between non-optimal policies RRRR and LLLL. The same problem was addressed by Lagoudakis and Parr (2003a) where the same basis functions, repeated for each action, were used to approximate the action-value function:

$$\phi(s,a) = \begin{pmatrix} I(a=L)\cdot 1 \\ I(a=L)\cdot s \\ I(a=L)\cdot s^2 \\ I(a=R)\cdot 1 \\ I(a=R)\cdot s \\ I(a=R)\cdot s^2 \end{pmatrix}, \quad s \in \{1,2,3,4\}\,,\ a \in \{L,R\} \tag{25}$$

where I is the indicator function. They showed that Least-Square Policy Iteration (LSPI) was able to find the optimal policy in a few iterations. In our experiments, we approach the same problem using, where necessary, the set of features (25) to approximate the action-value function. The main difference w.r.t. approach in Koller and Parr (2000) arises from the fact that we use model-based least square method whereas they use (sample-based) Least-Square Temporal Difference for Q-functions (LSTD-Q) approximator. For this setting, we

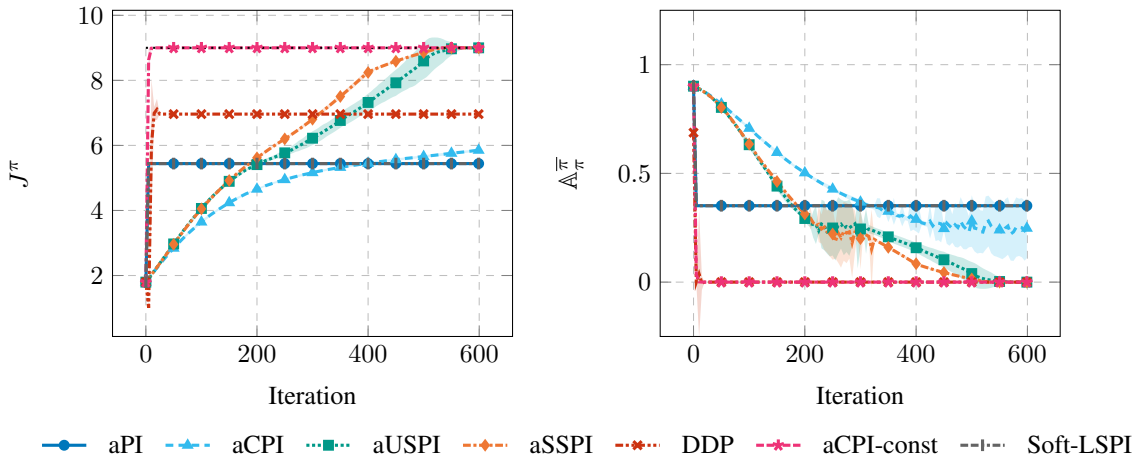


Figure 9: Score J^π and average advantage $\mathbb{A}_\pi^{\overline{\pi}}$ as a function of iterations when considering linear approximators for the value function. The underline domain consists of a discounted (0.9) chain with 4 states with fixed initial policy RRRR. A dotted line is drawn in correspondence of the value of the optimal policy. 10 runs, 95% confidence intervals.

considered additional baselines: Dynamic Policy Programming (DPP, Azar et al., 2012), CPI with constant learning rate (aCPI-const, Scherrer, 2014), and Softened-LSPI (Soft-LSPI, Pérolat et al., 2016). For aCPI-const we crossvalidated the learning rate α among the values 0.01, 0.05, 0.1, and 0.5. Similarly, for DPP we consider the best value of the inverse temperature parameter η among 0.1, 1, and 10.

The results of the experiment can be summarized as follows. Policy iteration algorithm is confirmed to oscillate between policy LRRR and policy RLLL which both have the same suboptimal performance. Similar behavior is displayed by Soft-LSPI, which tends to prefer too large learning rates, importing the same problems of aPI. aCPI does not suffer from the approximation and slowly converges (at infinity) to the optimal policy. On the other side, the proposed algorithms aUSPI and aSSPI are able to reach the optimal policy in a finite number of iterations without any loss of performance. aCPI-const with a learning rate set to 0.5 is able to reach the optimal performance very quickly with no oscillation phenomena. Instead, DPP, with an inverse temperature equal to 10, converges to a suboptimal policy. Furthermore, it should be noted that aCPI-const and DPP require the specification of an additional hyperparameter (the learning rate α and the inverse temperature η), whereas the SPI algorithms are hyperparameter free. For all algorithms, we used $N = 1000$ samples and a horizon of $T = 20$ for estimating the value function. For aSSPI, we used 1000 samples with and horizon 20 for estimating the advantage function in each state and 100 samples for estimating the derivative of the bound to be optimized. Figure 9 presents a general overview of the trend of the tested algorithms.

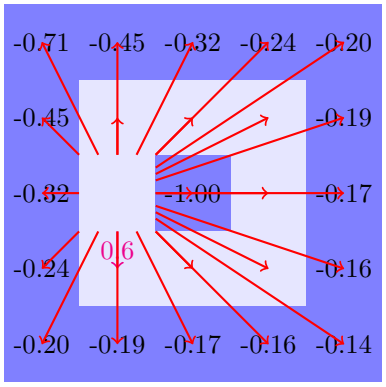


Figure 10: The prison domain. The reward is depicted for wall states (dark cells) because all the actions receive the same signal. Arrows represent the transitions associated to the "SOUTH" action with a noise level of 0.4.

6.2 The Prison

The domain is a particular implementation of the classical grid world domain (Sutton and Barto, 2018). This variant (Azar et al., 2012) is a M–squared grid world surrounded by absorbing states (wall states). Each state s is determined by the coordinate of the cell $c_s = (x, y)$, for some $x, y = 1, \dots, M$. The wall states that surround the grid have a reward that grows proportionally to the distance from the bottom–right corner:

$$R(s, a) = -\frac{1}{\|c_s\|_2}, \quad \forall a \in \mathcal{A}.$$

There is an additional absorbing state located at the center of the grid that has a reward of -1. All the remaining inner states have a zero reward signal. Four discrete actions $\mathcal{A} = \{\text{NORTH}, \text{SOUTH}, \text{WEST}, \text{EAST}\}$ are available in each state s. Taking an action a in an inner (non–absorbing) state s causes a transition in the direction corresponding to the action with a probability p, and a random move in a state $s' \neq s$ with a probability inversely proportional to the Euclidean distance $\frac{1}{\|c_s - c_{s'}\|_2}$. Due to the presence of absorbing states surrounding the grid, the optimal behavior is to survive in the grid, avoiding also the central wall. This domain is interesting due to the presence of absorbing states with negative reward and to the presence of noise that causes many transitions from inner states. Furthermore, having more than two actions lends itself to appreciate the advantages of SASPI over SSPI.

6.2.1 Experiments in the Exact Case

We first consider a set of experiments on the prison domain with $M = 5$ and $\gamma = 0.9$, in the exact case. In Figure 11 we can see a behavior that preserves the ordering, in learning speed, of the considered algorithms, compared to the chain domain. Indeed, PI converges to the optimal policy in a small number of iterations and among the SPI and CPI algorithms, USPI performs best. Again, we have the confirmation that the tighter bound exploited by USPI is enough to overcome the limitation of the unique coefficient α. Similarly, CPI is the

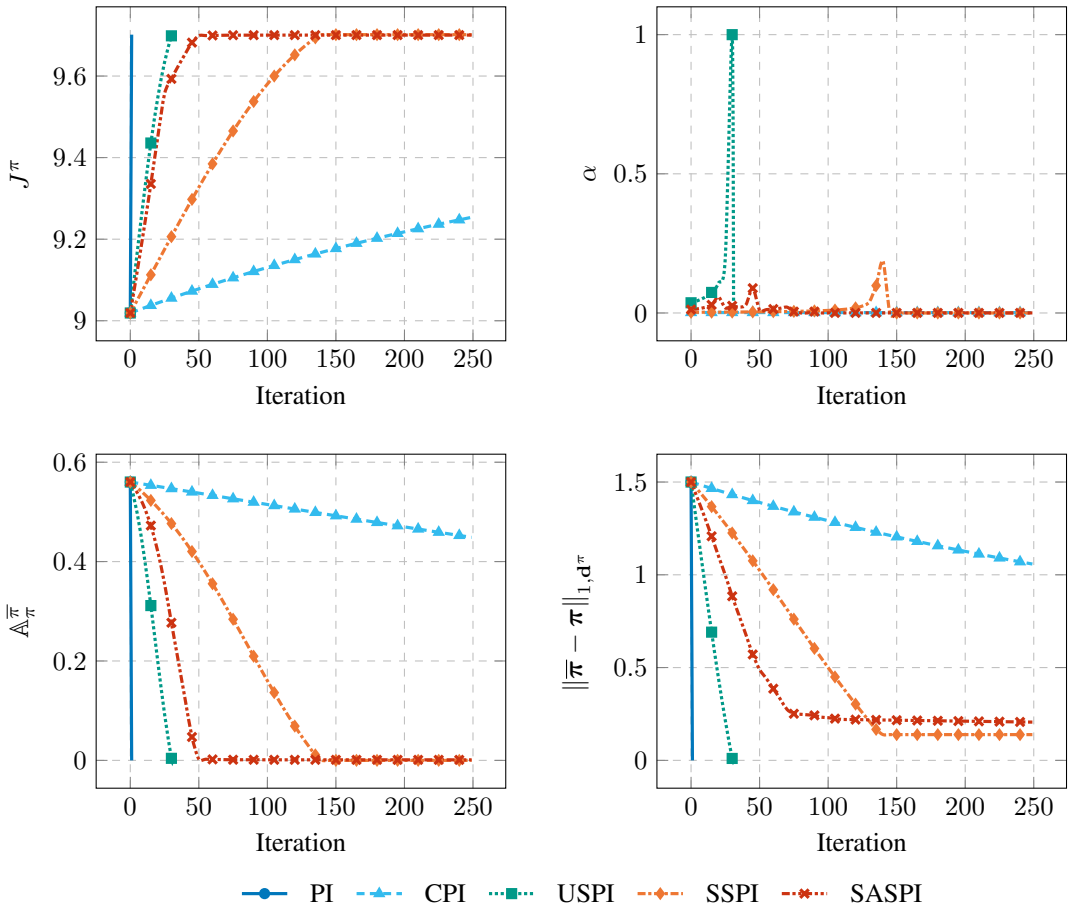


Figure 11: Score J^π, coefficient α, expected advantage $\mathbb{A}_\pi^{\overline{\pi}}$ and policy dissimilarity $\|\overline{\pi} - \pi\|_{1,\mathbf{d}^\pi}$ as a function of iterations. The underline domain consists of a discounted (0.9) prison environment in the exact setting.

worst performing algorithm. Between the latter two, we can now appreciate a difference in the learning curve. Since the latter considers a more flexible update rule, it is able to reach the optimal performance sooner.

6.2.2 Experiments in the Approximate Case

Compared to the Chain, the Prison domain poses significant challenges to the usage of the approximate versions of the SPI algorithms with the theoretical values of the hyperparameters. The reason of these limitations reside in the large number of deterministic policies to test, i.e., $|\mathcal{A}|^{|\mathcal{S}|} = 4^{M^2} \simeq 1.13 \cdot 10^{15}$ for $M = 5$. Even if we disregard the terminal states, as the action is those states has no effect, we reduce to $4^8 = 65536$ policies. For this reason, we decided to run the considered algorithms without the theoretical values of the hyperparametes. More specifically, we employed 200 samples with horizon 50 for the policy chooser, we used 50 with horizon 50 samples for estimating the state advantage function

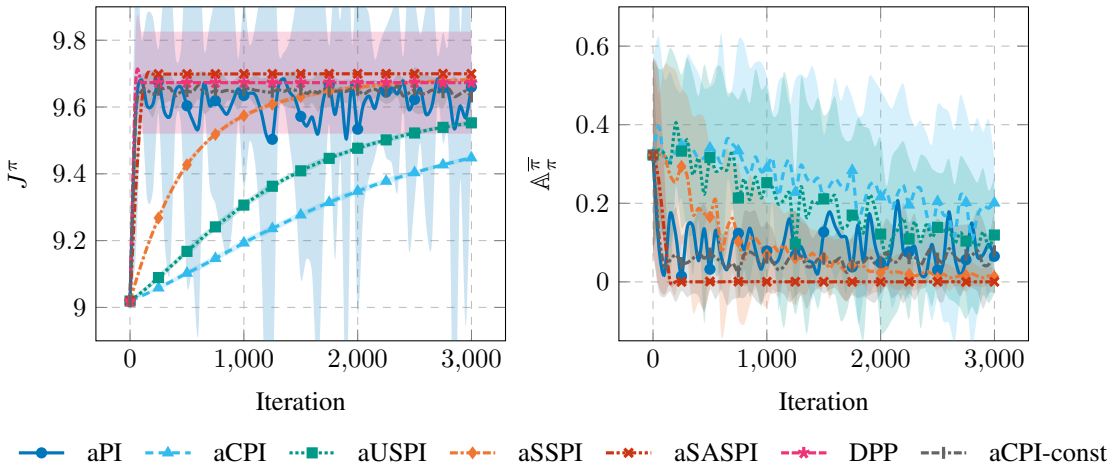


Figure 12: Score J^π and expected advantage $\mathbb{A}_\pi^{\overline{\pi}}$ as a function of iterations. The underline domain consists of a discounted (0.9) prison environment in the approximate setting. 10 runs, 95% confidence intervals.

and the Q-function for aSSPI and aSASPI, respectively, and 50 samples for computing the bound derivative. Therefore, the number of collected transitions for aPI, aCPI, and aUSPI is 10000, while for aSSPI and aSASPI we need to collect 260050 transitions (refer to Table 1 for the computation). Figure 12 shows the behavior of the compared algorithms. First of all, we easily notice that aPI displays a considerable oscillation phenomenon, this is due to the fact that it is continuously switching the policy between the optimal and some suboptimal policies. Instead, CPI and USPI converge slowly towards the optimal policy with a small advantage of USPI in terms of learning speed, but they are both outperformed by aSSPI. We can see that aSASPI is able to outperform all the other SPI algorithms and reach the optimal behavior with remarkable stability. aCPI-const with $\alpha = 0.05$, as well as DPP with $\eta = 1$ reach very quickly a good performance, although aCPI-const displays a little instability. However, both aCPI-const and DPP converge to a policy with slightly lower performance compared to that reached by aSASPI.

6.3 BlackJack Card Game

The BlackJack is a card game where the player seeks to beat the dealer by obtaining a total score greater than the dealer's one without exceeding 21 (refer to Dutech et al. (2005) for more details). Each card counts as its numerical value (2 through 10) except for aces and figures. The Jack, Queen, and King are worth 10, whereas the ace may value as either 1 or 11. The value of the ace is hand such that it produces the highest value equal to or less than 21. A hand is called *soft* when the ace is counted as 11. The set of cards is composed by 6 decks each one is a standard 52–cards deck.

At the beginning of the game, the dealer deals two cards to each player, including himself. One card is faced up and the other is faced down. The player checks his two cards and chooses to receiver a new card (*hit*) or to stop (*stand*). The player may ask for more

cards as long as he does not *bust*, i.e., the sum of the card values does not overcome 21. When all the players go bust or stops, is the turn of the dealer. In the beginning, the player can also decide to *double* the score, i.e., he does not ask for any card and let the dealer play. In this case, any score achieved by the player is doubled.

In Pirotta et al. (2013b) a simplified version, made of 260 states, of the blackjack game by removing advanced actions as "doubling", "splitting", etc was considered. We now extend those results by allowing the usage of the double action. The game is composed of a player and a dealer. The state of the game is defined by three components: the sum of the cards of the player (2 to 20), the dealer's faced-up card (1 to 10) and the soft hand flag. The player is forced to play "stand" action on blackjack and on 21. Moreover, the soft hand flag is irrelevant when player's value is greater than 11. As a consequence, the cardinality of the state space is 260. The action space is composed of the three actions: hit (H), stand (S), and double (D). The rewards assigned to the player are +1 for winning (+1.5 for blackjack), -1 for loosing and 0 for every hit. All rewards are doubled when the double action is performed. Rewards have been scaled to fit the interval $[0, 1]$.

To evaluate the performance of the algorithms, we have exploited the simplified Black-Jack model with discount factor equal to 0.8^{14} and "stands on soft 17" strategy for the dealer. The evaluation measure is the estimated player edge, i.e., the average reward over multiple runs. We have been able to define a configuration where an approximate policy iteration, using a sample-based policy evaluation step and an exact improvement, oscillates between two non-optimal policies. This configuration has been obtained by limiting the policy space to three policies. Two of them are the same employed in Pirotta et al. (2013b): they select the best action (S) when player's value is equal to 20 and opposite actions for the other states (π_S selects S and π_H selects H). States with dealer's value is at least 9 are treated in an opposite way: policy π_S selects H and policy π_H chooses S. We introduce a new policy π_D, which performs the double (D) action when the player's value is 9 and the dealer's value is between 3 and 6 or the player's value is 10 and the dealer's value is between 2 and 9. In all other cases, π_D behaves like π_H. To summarize, the policies are defined according to the following rules:

$$\pi_S = \begin{cases} H, & \text{if } \textit{dealer's value} \text{ is greater or equal to 9} \\ H, & \text{if } \textit{player's value} \text{ is less than 1} \\ S, & \text{otherwise} \end{cases}$$

$$\pi_H = \begin{cases} S, & \text{if } \textit{dealer's value} \text{ is greater or equal to 9} \\ S, & \text{if } \textit{player's value} \text{ is greater than 19} \\ H, & \text{otherwise} \end{cases}$$

$$\pi_D = \begin{cases} D, & \text{if } \textit{player's value} \text{ is 9 and } \textit{dealer's value} \text{ between 3 and 6} \\ D, & \text{if } \textit{player's value} \text{ is 10 and } \textit{dealer's value} \text{ between 2 and 9} \\ S, & \text{if } \textit{player's value} \text{ is greater than 19} \\ H, & \text{otherwise} \end{cases}$$

14. We employed this value of the discount factor to mitigate the over-conservative behavior of the dissimilarity penalization.

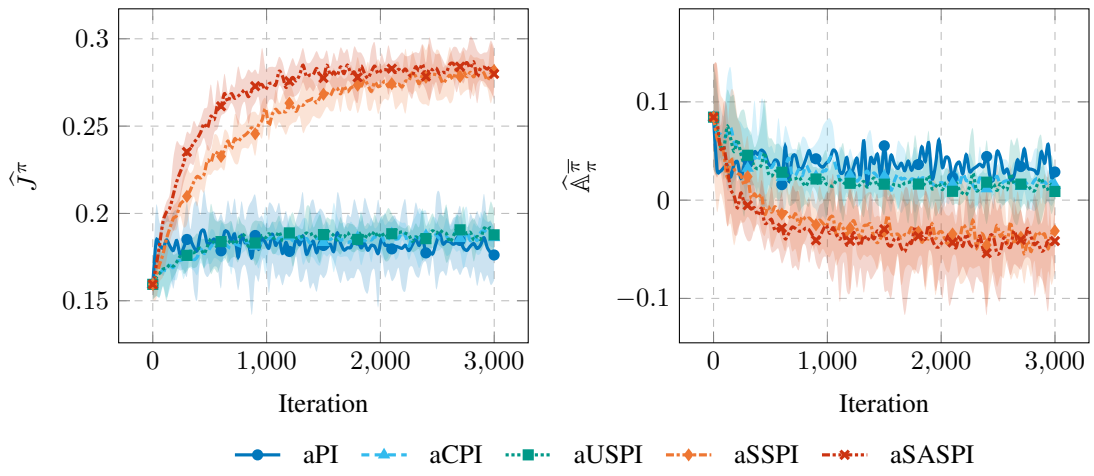


Figure 13: Estimated score $\widehat{J}^{\pi}$ and estimated expected advantage $\widehat{\mathbb{A}}_{\pi}^{\overline{\pi}}$ as a function of iterations. The underline domain consists of a discounted (0.8) blackjack environment. 5 runs, 95% confidence intervals.

Policy π_H has been chosen as initial policy. The goal of this experiment is to show the advantages of aSSPI and aSASPI when having access to a very limited target policy space. Indeed, by leveraging on the freedom granted by per–state and per–state–action combination coefficients, the effective set of policies representable by aSSPI and aSASPI is larger compared to that of aCPI and aUSPI.

Figure 13 reports the performance of the policies obtained by aPI, aCPI, aUSPI, aSSPI, and aSASPI algorithms using 1000 samples with horizon 100 for the policy chooser, we used 100 with horizon 100 samples for estimating the state advantage function and the Q-function for aSSPI and aSASPI respectively and 100 samples for computing the bound derivative. While aPI oscillates among the three policies, other algorithms do not get stuck and converge towards better policies. In particular, aCPI and aUSPI converge to a mixture of the three policies. It is worth underlining that, in this highly stochastic domain, the aSSPI and aSASPI are able to exploit the flexibility given by the multiple convex coefficients and to converge faster than aUSPI and aCPI, getting to policies that are not representable by aCPI and aUSPI. In principle, we could exploit the transformation presented in Remark 6 to allow aCPI and aUSPI to have access to the same set of policies as aSSPI and aSASPI. However, in this specific case, the number of target policies needed would be $\simeq 1.43e+73$, which is clearly prohibitive.[15]

7. Discussion and Conclusions

In this section, we discuss the contributions of this paper and we present some future research directions to overcome the limitations of the proposed approaches.

15. Notice that $|\mathcal{A}|^{|\mathcal{S}|} \simeq 1.13e+124$.

This paper builds upon (Pirotta et al., 2013b) and extends the work on the theoretical, algorithmic, and empirical sides. The main theoretical achievement, which was identified as a crucial point in Pirotta et al. (2013b), is the proof of the convergence of the SPI algorithms in a finite number of iterations in the exact setting (Section 4.1), opposed to the (only) asymptotic convergence of CPI. We have shown that the finite convergence also extends to the approximate setting, as long as the stopping threshold is chosen in an adaptive way (Section 5.2). From the algorithmic point of view, we introduce a novel and more general update rule based on per–state–action convex combination coefficients (SASPI, Section 4.3), that we presented in a unified framework together with SSPI (Section 4.2), which was already provided in Pirotta et al. (2013b). Moreover, we presented a complete PAC analysis of the single iteration of the considered algorithms (Section 5) that extends the one of Pirotta et al. (2013b) and Kakade and Langford (2002), allowing for the analysis of the newly presented approaches. Finally, the empirical validation extends the results of Pirotta et al. (2013b) by presenting the behavior of the new algorithms in the previous domains and the introduction of a new domain, the prison, in which the advantage of SASPI is clearly visible.

Although proved to be effective in terms of learning speed, aSSPI and aSASPI pose significant challenges in their application to domains with a large number of states and actions. Indeed, they require to approximate the advantage function and the Q-function (possibly using independent samples) for each state and each state–action pair. To alleviate this issue, it is possible to consider a slightly modified version of the algorithms, where the state–action space is split into subregions (using state–action aggregation) and all the state–action pairs in a region share the advantage function or Q-function. By changing the size of these subregions, we can generate several different situations that range from the original aSSPI and aSASPI approach (no aggregation) to the aUSPI one (where all the state and actions are associated with the same coefficient).

A research direction, already mentioned in Pirotta et al. (2013b), but not investigated in this paper, is to exploit the proposed bounds to perform approximate policy iteration in the off–policy case, i.e., when the samples have been initially collected (once for all) following some exploration strategy. In this case, we can use the bound in Theorem 3 where π_b is the exploration strategy.

Acknowledgments

...

Appendix A. Additional Proofs and Derivations

In this appendix, we report the proofs of the results we omitted in the main paper.

A.1 Proofs of Section 3

Corollary 4. *For any stationary policies π and π' and any starting state distribution μ, the difference between the performance of π' and the one of π can be lower bounded as follows:*

$$J^{\pi'} - J^{\pi} \geq \frac{1}{1-\gamma}\mathbb{A}_{\pi}^{\pi'} - \frac{\gamma}{(1-\gamma)^2}\left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_{\infty}^2 \frac{\|\mathbf{q}^{\pi}\|_{\infty}}{2}.$$

Proof The proof comes from a lower bound to the bound in Theorem 3 when $\pi_b = \pi$. Such lower bound involves two upper bounds. First, we perform the upper bound:

$$\left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_{1,\mathbf{d}^{\pi}} \leq \left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_{\infty}.$$

Second, upper bound the quantity $\frac{\mathrm{sp}\left(\mathbf{a}_{\pi}^{\pi'}\right)}{2}$:

$$\begin{aligned}
\frac{\mathrm{sp}\left(\mathbf{a}_{\pi}^{\pi'}\right)}{2} &\leq \left\|\mathbf{a}_{\pi}^{\pi'}\right\|_{\infty} = \left\|\left(\boldsymbol{\pi'} - \boldsymbol{\pi}\right)\mathbf{q}^{\pi}\right\|_{\infty} \\
&= \max_{s\in\mathcal{S}}\left\{\left|\sum_{a\in\mathcal{A}}\left(\pi'(a|s) - \pi(a|s)\right)Q^{\pi}(s,a)\right|\right\} \\
&\leq \max_{s\in\mathcal{S}}\left\{\left\|\pi'(\cdot|s) - \pi(\cdot|s)\right\|_1 \frac{\mathrm{sp}(Q^{\pi}(s,\cdot))}{2}\right\} \\
&\leq \max_{s\in\mathcal{S}}\left\{\left\|\pi'(\cdot|s) - \pi(\cdot|s)\right\|_1\right\}\max_{s\in\mathcal{S}}\left\{\frac{\mathrm{sp}(Q^{\pi}(s,\cdot))}{2}\right\} \\
&\leq \left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_{\infty}\frac{\|\mathbf{q}^{\pi}\|_{\infty}}{2},
\end{aligned}$$

where the last inequality follows from the positiveness of Q^{π} and from norm properties. The Corollary is proven by introducing this bound in place of $\frac{\mathrm{sp}\left(\mathbf{a}_{\pi}^{\pi'}\right)}{2}$ in Theorem 3 and replacing π_b with a generic π.

■

A.2 Proofs of Section 4

Lemma 2. *(Kakade and Langford, 2002) For any stationary policies π and π' and any starting state distribution μ:*

$$J^{\pi'} - J^{\pi} = \frac{1}{1-\gamma} \mathbf{d}^{\pi'^{\mathsf{T}}} \mathbf{a}_{\pi}^{\pi'}.$$

Proof The reader may refer to (Neu et al., 2017) and the references therein. We report the proof for completeness:

$$\begin{aligned}
(1-\gamma)J^{\pi'} &= (1-\gamma)\boldsymbol{\mu}^{\mathsf{T}}\mathbf{v}^{\pi'} = (1-\gamma)\boldsymbol{\mu}^{\mathsf{T}}\left(\mathbf{I}-\gamma\mathbf{P}^{\pi}\right)^{-1}\mathbf{r}^{\pi} = \mathbf{d}^{\pi'^{\mathsf{T}}}\mathbf{r}^{\pi'} \\
&= \mathbf{d}^{\pi'^{\mathsf{T}}}\mathbf{r}^{\pi'} + \left((1-\gamma)\boldsymbol{\mu}^{\mathsf{T}} + \gamma\mathbf{d}^{\pi'^{\mathsf{T}}}\mathbf{P}^{\pi'}\right)\mathbf{v}^{\pi} - \mathbf{d}^{\pi'^{\mathsf{T}}}\mathbf{v}^{\pi} \\
&= \mathbf{d}^{\pi'^{\mathsf{T}}}\left(\mathbf{r}^{\pi'} + \gamma\mathbf{P}^{\pi'}\mathbf{v}^{\pi} - \mathbf{v}^{\pi}\right) + (1-\gamma)\boldsymbol{\mu}^{\mathsf{T}}\mathbf{v}^{\pi} \\
&= \mathbf{d}^{\pi'^{\mathsf{T}}}\left(\mathbf{q}_{\pi}^{\pi'} - \mathbf{v}^{\pi}\right) + (1-\gamma)\boldsymbol{\mu}^{\mathsf{T}}\mathbf{v}^{\pi} = \mathbf{d}^{\pi'^{\mathsf{T}}}\mathbf{a}_{\pi}^{\pi'} + (1-\gamma)J^{\pi}.
\end{aligned}$$

■

We now present a general result for vectors that we will employ for the proof of Theorem 3.

Lemma 23. *(Haviv and Van der Heyden, 1984, Corollary 2.4) For any vector* $\mathbf{d}$ *and any vector* $\mathbf{c}$ *such that* $\mathbf{c}^T\mathbf{e} = 0$*,*

$$\left|\mathbf{c}^T\mathbf{d}\right| \leq \frac{1}{2}\left\|\mathbf{c}\right\|_1 \mathrm{sp}(\mathbf{d}).$$

Lemma 7. *Let* $\pi, \pi' \in \Pi^{SR}$ *be two arbitrary policies and* π^+ *be a greedy policy induced by* Q^{π}*. Then, the expected advantage* $\mathbb{A}_{\pi}^{\pi^+}$ *can be lower bounded as:*

$$\frac{\mathbb{A}_{\pi}^{\pi^+}}{1-\gamma} \geq \left\|\frac{\mathbf{d}^{\pi'}}{\mathbf{d}^{\pi}}\right\|_{\infty}^{-1}\left(J^{\pi'} - J^{\pi}\right), \tag{3}$$

where $\frac{\mathbf{d}^{\pi'}}{\mathbf{d}^{\pi}}$ *is the vector obtained by the element–wise division between* $\mathbf{d}^{\pi'}$ *and* $\mathbf{d}^{\pi}$*.*

Proof The lemma follows from the following decomposition:

$$\begin{aligned}
\mathbb{A}_{\pi}^{\pi^+} &= \mathbf{d}^{\pi\mathsf{T}}\mathbf{A}_{\pi}^{\pi^+} \\
&= \sum_{s\in\mathcal{S}} d^{\pi}(s) A_{\pi}^{\pi^+}(s) \\
&= \sum_{s\in\mathcal{S}} d^{\pi}(s) \max_{a\in\mathcal{A}} A^{\pi}(s,a) \\
&= \sum_{s\in\mathcal{S}} \frac{d^{\pi}(s)}{d^{\pi'}(s)} d^{\pi'}(s) \max_{a\in\mathcal{A}} A^{\pi}(s,a) \\
&\geq \left(\min_{s\in\mathcal{S}} \frac{d^{\pi}(s)}{d^{\pi'}(s)}\right)\left(\sum_{s\in\mathcal{S}} d^{\pi'}(s) \max_{a\in\mathcal{A}} A^{\pi}(s,a)\right) && \text{(P.21)} \\
&\geq \left(\min_{s\in\mathcal{S}} \frac{d^{\pi}(s)}{d^{\pi'}(s)}\right)\left(\sum_{s\in\mathcal{S}} d^{\pi'}(s) \sum_{a\in\mathcal{A}} \pi'(a|s) A^{\pi}(s,a)\right) && \text{(P.22)}
\end{aligned}$$

$$= \left(\max_{s \in \mathcal{S}} \frac{d^{\pi'}(s)}{d^{\pi}(s)} \right)^{-1} \sum_{s \in \mathcal{S}} d^{\pi'}(s) \sum_{a \in \mathcal{A}} \pi'(a|s) A^{\pi}(s,a) \tag{P.23}$$

$$= \left\| \frac{\mathbf{d}^{\pi'}}{\mathbf{d}^{\pi}} \right\|_{\infty}^{-1} \mathbf{d}^{\pi' \intercal} \mathbf{A}_{\pi}^{\pi'}$$

$$\geq \left\| \frac{\mathbf{d}^{\pi'}}{\mathbf{d}^{\pi}} \right\|_{\infty}^{-1} (1-\gamma)\big(J^{\pi'} - J^{\pi}\big), \tag{P.24}$$

where (P.21) follows by observing that $\sum_i \mathbf{x}(i)\mathbf{y}(i) \geq \min_i \mathbf{x}(i) \sum_i \mathbf{y}(i)$, for $\mathbf{x}$ and $\mathbf{y}$ vectors with non negative components. (P.22) follows from the fact that $\max_{a \in \mathcal{A}} A^{\pi}(s,a) \geq \sum_{a \in \mathcal{A}} \pi'(a|s) A^{\pi}(s,a)$ for any policy π'. (P.23) derives from simply observing that $\min_i \mathbf{x}(i) = (\max_i \{1/\mathbf{x}(i)\})^{-1}$. Finally, (P.24) is an application of Lemma 2. ■

Lemma 9. *Assume the same setting as Theorem 8. Let $\Delta_J = J^* - \max_{\pi \in \Pi^{SD} \setminus \Pi^*} \{J^{\pi}\}$ be the performance gap between the optimal policies and the second-best deterministic policy, where $\Pi^* = \{\pi \in \Pi^{SD} : J^{\pi} = J^*\}$. Then, USPI (and CPI) with GPC selects an optimal policy as target policy after a finite number of iterations.*

Proof The result is a straightforward application of Theorem 8:

$$J^* - J^{\pi_N} \leq \frac{8\gamma}{N \Delta_d^2 (1-\gamma)^3} \leq \Delta_J \quad \implies \quad N \geq \frac{8\gamma}{\Delta_J \Delta_d^2 (1-\gamma)^3}.$$

■

In the following lemma, we prove that Q^{π} is Lipschitz continuous w.r.t. to π across all state-action pairs. We will employ this property in the proof of Lemma 25.

Lemma 24. *The Q-function $Q^{\pi}(s,a)$ is Lipschitz continuous over the space of Markovian stationary randomized policies Π^{SR} in L_∞-norm, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For all $\pi, \pi' \in \Pi^{SR}$ and for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, it holds that:*

$$\left| Q^{\pi'}(s,a) - Q^{\pi}(s,a) \right| \leq \frac{\gamma}{(1-\gamma)^2} \left\| \boldsymbol{\pi'} - \boldsymbol{\pi} \right\|_{\infty}.$$

Proof As for the γ-stationary distribution we look at Q^{π} as a function $Q^{\bullet}(s,a) : \Pi \to \mathbb{R}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Therefore, we have:

$$\left| Q^{\pi'}(s,a) - Q^{\pi}(s,a) \right| \leq \left\| \mathbf{q}^{\pi'} - \mathbf{q}^{\pi} \right\|_{\infty}.$$

We now provide a bound for $\left\| \mathbf{q}^{\pi'} - \mathbf{q}^{\pi} \right\|_{\infty}$. By exploiting the Bellman equation we can write:

$$\mathbf{q}^{\pi'} - \mathbf{q}^{\pi} = \mathbf{r} + \gamma \mathbf{P}^{\pi'} \mathbf{q}^{\pi'} - \mathbf{r} - \gamma \mathbf{P}^{\pi} \mathbf{q}^{\pi}$$

$$
\begin{aligned}
&= \gamma \mathbf{P}^{\pi'} \mathbf{q}^{\pi'} - \gamma \mathbf{P}^{\pi'} \mathbf{q}^{\pi'} \pm \gamma \mathbf{P}^{\pi} \mathbf{q}^{\pi} \\
&= \gamma \mathbf{P}^{\pi'} \left(\mathbf{q}^{\pi'} - \mathbf{q}^{\pi} \right) + \gamma \left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi} \right) \mathbf{q}^{\pi}.
\end{aligned}
$$

Now, we can take the L^{∞}–norm and we obtain:

$$
\begin{aligned}
\left\| \mathbf{q}^{\pi'} - \mathbf{q}^{\pi'} \right\| &\leq \gamma \left\| \mathbf{P}^{\pi'} \right\|_{\infty} \left\| \mathbf{q}^{\pi'} - \mathbf{q}^{\pi} \right\|_{\infty} + \gamma \left\| \mathbf{P}^{\pi'} - \mathbf{P}^{\pi} \right\|_{\infty} \left\| \mathbf{q}^{\pi} \right\|_{\infty} \\
&\leq \gamma \left\| \mathbf{q}^{\pi'} - \mathbf{q}^{\pi} \right\|_{\infty} + \frac{\gamma}{1-\gamma} \left\| \mathbf{P}^{\pi'} - \mathbf{P}^{\pi} \right\|_{\infty} \\
&\leq \frac{\gamma}{(1-\gamma)^2} \left\| \mathbf{P}^{\pi'} - \mathbf{P}^{\pi} \right\|_{\infty} \\
&\leq \frac{\gamma}{(1-\gamma)^2} \left\| \mathbf{P} \right\|_{\infty} \left\| \boldsymbol{\pi'} - \boldsymbol{\pi} \right\|_{\infty} \\
&= \frac{\gamma}{(1-\gamma)^2} \left\| \boldsymbol{\pi'} - \boldsymbol{\pi} \right\|_{\infty}.
\end{aligned}
$$

■

The following technical result is employed in the proof of Lemma 10, i.e., to prove that the expected advantage can be lower–bounded by a function of the distance between the current policy π and the greedy policy π^+.

Lemma 25. *Assume the same setting as Lemma 9. Let π^* be an optimal policy and $\pi_0 \in \Pi$ a suboptimal policy such that $J^* - \Delta_J \leq J^{\pi_0} \leq J^*$. Let $\eta \in [0,1]$ and $\pi = \eta\pi^* + (1-\eta)\pi_0$. Then, there exists a state $s \in \mathcal{S}$ and a constant $\Delta_+ > 0$ independent of η, such that the function:*

$$
g_s(\pi) = \sum_{a \in \mathcal{A} \setminus \{a_s^+\}} \pi_0(a|s) \left(Q^{\pi}(s, a_s^+) - Q^{\pi}(s,a) \right) \geq \Delta_+, \tag{26}
$$

where $a_s^+ \in \arg\max_{a \in \mathcal{A}} \{Q^{\pi}(s,a)\}$ is a greedy action w.r.t. $Q^{\pi}(s,a)$.

Proof The proof is divided in two parts. We first prove that for any $s \in \mathcal{S}$, g_s is continuous over the set of policies $\widetilde{\Pi} = \{\eta\pi^* + (1-\eta)\pi_0 : \eta \in [0,1]\}$ and then we show that there exists a state s^- in which g_{s^-} it is strictly positive for any $\eta \in [0,1]$. Being $\widetilde{\Pi}$ a compact set, from Weierstrass theorem, it follows that g_{s^-} admits a positive minimum.

Continuity of g_s. Since the dependence on η is only in the term $Q^{\pi}(s, a_s^+) - Q^{\pi}(s,a)$, all it takes is to prove that it is a continuous function over $\widetilde{\Pi}$. By recalling that the greedy action a_s^+ is the one that maximizes $Q^{\pi}(s,\cdot)$ we have:

$$
Q^{\pi}(s, a_s^+) - Q^{\pi}(s,a) = \max_{a \in \mathcal{A}} \{Q^{\pi}(s,a)\} - Q^{\pi}(s,a).
$$

Since $Q^{\pi}(s,a)$ is continuous for Lemma 24 and the maximum of a continuous function is continuous we have that $Q^{\pi}(s, a_s^+) - Q^{\pi}(s,a)$ is continuous. $g_s(\pi)$ is obtained by applying transformations that preserve continuity.

Positivity g_s. By definition $g_s(\pi) \geq 0$. Suppose by contradiction that $g_s(\pi) = 0$ for all $s \in \mathcal{S}$. Thus, in all states $s \in \mathcal{S}$ and for all actions $a \in \mathcal{A} \setminus \{a_s^+\}$ we have that either $\pi_0(a|s) = 0$ or $Q^{\pi}(s, a_s^+) = Q^{\pi}(s,a)$. If in all states $s \in \mathcal{S}$ and for all actions $a \in \mathcal{A} \setminus \{a_s^+\}$

we have that $\pi_0(a|s) = 0$, then π_0 gives probability one to greedy actions, i.e., $\pi_0 = \pi^+$. Note that π^+ is optimal, since we are in the setting of Lemma 9, and thus π_0 is optimal, which is a contradiction. Thus, there must exist a state at least one state s^- and at least one action a_s^- such that $\pi_0(a_s^-|s^-) > 0$. Now, suppose that for all those states s^- and for all those actions a_s^-, we had $Q^\pi(s, a_s^+) = Q^\pi(s^-, a_s^-)$. It means that a_s^- is greedy and thus optimal. Therefore, π_0 is optimal, which is a contradiction. ■

Lemma 10. *Assume the same setting as Lemma 9. It π^* is a deterministic optimal policy, then, there exists a constant $\Delta_+ > 0$ such that:*

$$\mathbb{A}_\pi^{\pi^*} \geq \frac{\Delta_d \Delta_+}{2} \left\| \boldsymbol{\pi}^* - \boldsymbol{\pi} \right\|_\infty . \tag{4}$$

Proof The Lemma follows from the following decomposition:

$$\begin{aligned}
\mathbb{A}_\pi^{\pi^*} &= \mathbf{d}^{\pi\mathtt{T}} \mathbf{A}_\pi^{\pi^*} \\
&= \mathbf{d}^{\pi\mathtt{T}} \left(\boldsymbol{\pi}^* - \boldsymbol{\pi}\right) \mathbf{q}^\pi \\
&= (1-\eta) \mathbf{d}^{\pi\mathtt{T}} \left(\boldsymbol{\pi}^* - \boldsymbol{\pi}_0\right) \mathbf{q}^\pi && \text{(P.25)} \\
&= (1-\eta) \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \left(\pi^*(a|s) - \pi_0(a|s)\right) Q^\pi(s,a) \\
&= (1-\eta) \sum_{s \in \mathcal{S}} d^\pi(s) \left(\left(1 - \pi_0(a_s^*|s)\right) Q^\pi(s, a_s^*) - \sum_{a \in \mathcal{A} \setminus \{a_s^*\}} \pi_0(a|s) Q^\pi(s,a) \right) && \text{(P.26)} \\
&= (1-\eta) \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A} \setminus \{a_s^*\}} \pi_0(a|s) \left(Q^\pi(s, a_s^*) - Q^\pi(s,a)\right) && \text{(P.27)} \\
&\geq (1-\eta) \min_{s \in \mathcal{S}} \{d^\pi(s)\} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A} \setminus \{a_s^*\}} \pi_0(a|s) \left(Q^\pi(s, a_s^*) - Q^\pi(s,a)\right) && \text{(P.28)} \\
&\geq (1-\eta) \min_{s \in \mathcal{S}} \{d^\pi(s)\} \max_{s \in \mathcal{S}} \left\{ \sum_{a \in \mathcal{A} \setminus \{a_s^*\}} \pi_0(a|s) \left(Q^\pi(s, a_s^*) - Q^\pi(s,a)\right) \right\} && \text{(P.29)} \\
&\geq (1-\eta) \min_{s \in \mathcal{S}} \{d^\pi(s)\} \max_{s \in \mathcal{S}} \{g_s(\pi)\} && \text{(P.30)} \\
&\geq (1-\eta) \Delta_d \Delta_+ && \text{(P.31)} \\
&= \Delta_d \Delta_+ \frac{\left\| \boldsymbol{\pi}^* - \boldsymbol{\pi} \right\|_\infty}{\left\| \boldsymbol{\pi}^* - \boldsymbol{\pi}_0 \right\|_\infty} && \text{(P.32)} \\
&\geq \frac{\Delta_d \Delta_+}{2} \left\| \boldsymbol{\pi}^* - \boldsymbol{\pi} \right\|_\infty , && \text{(P.33)}
\end{aligned}$$

We now explain the steps of the derivation. (P.25) follows from observing that $\boldsymbol{\pi} = \eta \boldsymbol{\pi}^* + (1-\eta)\boldsymbol{\pi}_0$ and so $\boldsymbol{\pi}^* - \boldsymbol{\pi} = (1-\eta)\left(\boldsymbol{\pi}^* - \boldsymbol{\pi}_0\right)$. (P.26) is obtained by observing that the optimal policy is unique and deterministic; whereas (P.27) derives from observing that $1 - \pi_0(a_s^*|s) = \sum_{a \in \mathcal{A} \setminus \{a_s^*\}} \pi_0(a|s)$. (P.28) is obtained from $\sum_i \mathbf{x}(i)\mathbf{y}(i) \geq \min_i \mathbf{x}(i) \sum_i \mathbf{y}(i)$, for $\mathbf{x}$ and $\mathbf{y}$ vectors with non negative components; while (P.29) exploits the trivial relation $\sum_i \mathbf{y}(i) \geq \max_i \mathbf{y}(i)$. (P.30) simply exploits the definition of $g_s(\pi)$ as defined in

Lemma 25 and (P.31) applies Lemma 25. (P.32) derives from the fact that $\|\boldsymbol{\pi}^* - \boldsymbol{\pi}\|_\infty = (1-\eta)\|\boldsymbol{\pi}^* - \boldsymbol{\pi}_0\|_\infty$ and (P.33) by bounding $\|\boldsymbol{\pi}^* - \boldsymbol{\pi}_0\|_\infty \leq 2$. ■

A.3 Proofs of Section 5

The following result upper–bounds the error, in L_1-norm, when approximating the infinite-horizon γ-discounted stationary distribution with a finite horizon T.

Lemma 26. *Let $\mathbf{d}^\pi$ be the infinite-horizon γ-discounted stationary distribution and $\mathbf{d}^\pi_T$ be the T-horizon γ-discounted stationary distribution, then it holds that:*

$$\|\mathbf{d}^\pi - \mathbf{d}^\pi_T\|_1 \leq 2\gamma^T.$$

Proof We use the definition of d^π and d^π_T:

$$\begin{aligned}
\|\mathbf{d}^\pi - \mathbf{d}^\pi_T\|_1 &= \sum_{s\in\mathcal{S}} \left| (1-\gamma) \sum_{t=0}^{+\infty} \gamma^t \Pr(s_t = s|\pi, \mathcal{M}) - \frac{1-\gamma}{1-\gamma^T} \sum_{t=0}^{H-1} \gamma^t \Pr(s_t = s|\pi, \mathcal{M}) \right| \\
&\leq (1-\gamma) \sum_{s\in\mathcal{S}} \left| \sum_{t=T}^{+\infty} \gamma^t \Pr(s_t = s|\pi, \mathcal{M}) \right| + \frac{1-\gamma}{1-\gamma^T} \gamma^T \sum_{s\in\mathcal{S}} \left| \sum_{t=0}^{T-1} \gamma^t \Pr(s_t = s|\pi, \mathcal{M}) \right| \\
&\leq (1-\gamma) \sum_{t=T}^{+\infty} \gamma^t \sum_{s\in\mathcal{S}} \Pr(s_t = s|\pi, \mathcal{M}) + \frac{1-\gamma}{1-\gamma^T} \gamma^T \sum_{t=0}^{T-1} \gamma^t \sum_{s\in\mathcal{S}} \Pr(s_t = s|\pi, \mathcal{M}) \\
&= (1-\gamma) \sum_{t=T}^{+\infty} \gamma^t + \frac{1-\gamma}{1-\gamma^T} \gamma^T \sum_{t=0}^{T-1} \gamma^t \\
&= (1-\gamma) \frac{\gamma^T}{1-\gamma} + \frac{1-\gamma}{1-\gamma^T} \gamma^T \frac{1-\gamma^T}{1-\gamma} = 2\gamma^T,
\end{aligned}$$

where we exploited the fact that $\sum_{s\in\mathcal{S}} \Pr(s_t = s|\pi, \mathcal{M}) = 1$ for every t and we used the properties of the geometric series. ■

We now bound the error introduced when considering finite trajectories of horizon T in the estimation of the action-value function.

Lemma 27. *Let $\mathbf{q}^\pi$ be the infinite-horizon action-value function and $\mathbf{q}^\pi_T$ be the T-horizon action-value function, then it holds that:*

$$\|\mathbf{q}^\pi - \mathbf{q}^\pi_T\|_\infty \leq \frac{\gamma^T}{1-\gamma}.$$

Proof We exploit the definition of the involved quantities. For any state s and action a it holds that:

$$Q^\pi(s,a) - Q^\pi_T(s,a) = \mathop{\mathbb{E}}_{\substack{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t) \\ a_{t+1}\sim\pi(\cdot|s_{t+1})}} \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) | s_0 = s, a_0 = a \right]$$

$$
\begin{aligned}
&- \mathop{\mathbb{E}}_{\substack{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)\\ a_{t+1}\sim\pi(\cdot|s_{t+1})}} \left[\sum_{t=0}^{T-1}\gamma^t\mathcal{R}(s_t,a_t)|s_0=s,a_0=a\right] \\
&= \mathop{\mathbb{E}}_{\substack{s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)\\ a_{t+1}\sim\pi(\cdot|s_{t+1})}} \left[\sum_{t=T}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)|s_0=s,a_0=a\right] \\
&\leq \|\mathbf{r}\|_\infty \sum_{t=T}^{\infty}\gamma^t \leq \frac{\gamma^T}{1-\gamma},
\end{aligned}
$$

where we exploited the fact that $\|\mathbf{r}\|_\infty \leq 1$. ■

Lemma 18. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_\gamma \frac{\epsilon}{12} \right\rceil \qquad \textit{and} \qquad N = \left\lceil \frac{16}{9\epsilon^2}\log\frac{2}{\delta}\right\rceil.$$

The aUSPI Sampling Procedure construct a function $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ such that with probability $1-\delta$:

$$\left|\widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \mathbb{A}_\pi^{\overline{\pi}}\right| \leq \frac{\epsilon}{1-\gamma}.$$

Proof The proof is analogous to that of Lemma 15. First, we define the expected advantage function computed with horizon T:

$$\mathbb{A}_{\pi,T}^{\overline{\pi}} = \mathbb{Q}_{\pi,T}^{\overline{\pi}} - \mathbb{Q}_{\pi,T}^{\pi}.$$

Also in this case our estimation $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ is unbiased for $\mathbb{A}_{\pi,T}^{\overline{\pi}}$. We now partition the error highlighting the bias terms:

$$\left|\widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \mathbb{A}_\pi^{\overline{\pi}}\right| \leq \underbrace{\left|\widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \mathbb{A}_{\pi,T}^{\overline{\pi}}\right|}_{(E_1)} + \underbrace{\left|\mathbb{A}_{\pi,T}^{\overline{\pi}} - \mathbb{A}_\pi^{\overline{\pi}}\right|}_{(E_2)}.$$

We now require that $E_1 = \frac{3\epsilon}{4(1-\gamma)}$ and $E_2 = \frac{\epsilon}{4(1-\gamma)}$, so that $E_1 = E_2 = \frac{\epsilon}{1-\gamma}$. Then, we apply Hoeffding's inequality to bound the deviation E_1, by observing that the terms $(\overline{\pi}(a|s) - \pi(a|s))\,\widehat{q}_i$ in the estimation $\widehat{\mathbb{A}}_\pi^{\overline{\pi}}$ are all independent and belong to $\left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]$. Therefore, we have:

$$
\begin{aligned}
\Pr\left(\left|\widehat{\mathbb{A}}_\pi^{\overline{\pi}} - \mathbb{A}_{\pi,T}^{\overline{\pi}}\right| \geq \frac{3\epsilon}{4(1-\gamma)}\right) &\leq 2e^{-\frac{2\left(\frac{3\epsilon}{4(1-\gamma)}\right)^2 N^2}{\sum_{i=1}^N (b_i-a_i)^2}} \\
&= 2e^{-\frac{18\epsilon^2N^2(1-\gamma)^2}{32(1-\gamma)^2N}} = 2e^{-\frac{9\epsilon^2N}{16}}.
\end{aligned}
$$

From which we get:

$$e^{-\frac{9\epsilon^2N}{16}} \leq \frac{\delta}{2} \implies \frac{9\epsilon^2N}{16} \leq -\log\frac{\delta}{2} \implies N \geq \frac{16}{9\epsilon^2}\log\frac{2}{\delta}.$$

For E_2 we observe that:

$$\begin{aligned}
\left|\mathbb{A}^{\overline{\pi}}_{\pi,T} - \mathbb{A}^{\overline{\pi}}_{\pi}\right| &= \left| \mathop{\mathbb{E}}_{s\sim d^{\pi}_{T}} \left[A^{\overline{\pi}}_{\pi,T}(s)\right] - \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[A^{\overline{\pi}}_{\pi}(s)\right] \pm \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[A^{\overline{\pi}}_{\pi,T}(s)\right] \right| \\
&\leq \left| \mathop{\mathbb{E}}_{s\sim d^{\pi}_{T}} \left[A^{\overline{\pi}}_{\pi,T}(s)\right] - \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[A^{\overline{\pi}}_{\pi,T}(s)\right] \right| + \left| \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[A^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi,T}(s)\right] \right| \\
&\leq \frac{1}{1-\gamma} \left\|\mathbf{d}^{\pi} - \mathbf{d}^{\pi}_{T}\right\|_1 + \left| \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[\sum_{a\in\mathcal{A}} \left(\overline{\pi}(a|s) - \pi(a|s)\right) \left(Q^{\pi}(s,a) - Q^{\pi}_{T}(s,a)\right) \right] \right| \\
&\leq \frac{2\gamma^T}{1-\gamma} + \frac{1}{2} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[\left\|\pi(\cdot|s) - \pi(\cdot|s)\right\|_1 \mathrm{sp}(Q^{\pi}(s,\cdot) - Q^{\pi}_{T}(s,\cdot)) \right] \\
&\leq \frac{2\gamma^T}{1-\gamma} + \frac{\gamma^T}{1-\gamma} = \frac{3\gamma^T}{1-\gamma},
\end{aligned}$$

where we exploited Lemma 23 in the last but one line and Lemma 27 in the last line. Thus, by selecting $T = \lceil \log_{\gamma} \frac{\epsilon}{12} \rceil$ we have that the bias E_2 is bounded by $\frac{\epsilon}{4(1-\gamma)}$. ■

Lemma 19. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_{\gamma} \frac{\epsilon}{8} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{128}{9\epsilon^2} \log \frac{2|\mathcal{S}|}{\delta} \right\rceil.$$

The aSSPI Sampling Procedure constructs a function $A^{\overline{\pi}}_{\pi}(s)$ such that with probability $1-\delta$, simultaneously for all $s \in \mathcal{S}$:

$$\left| \widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi}(s) \right| \leq \frac{\epsilon}{2(1-\gamma)}.$$

Proof Similarly to the proof of Lemma 15, we define the expected advantage function computed with horizon T:

$$A^{\overline{\pi}}_{\pi,T}(s) = \sum_{a\in\mathcal{A}} \left(\overline{\pi}(a|s) - \pi(a|s)\right) Q^{\pi}_{T}(s,a),$$

where Q^{π}_{T} is the action-value function cut at horizon T. Once again, our estimator $\widehat{A}^{\overline{\pi}}_{\pi}$ is unbiased for $A^{\overline{\pi}}_{\pi,T}$. Therefore, we partition the error into:

$$\left| \widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi}(s) \right| \leq \underbrace{\left| \widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi,T}(s) \right|}_{(E_1)} + \underbrace{\left| A^{\overline{\pi}}_{\pi,T}(s) - A^{\overline{\pi}}_{\pi}(s) \right|}_{(E_2)}.$$

We set $E_1 = \frac{3\epsilon}{8(1-\gamma)}$ and $E_2 = \frac{\epsilon}{8(1-\gamma)}$ so that $E_1 + E_2 = \frac{\epsilon}{2(1-\gamma)}$. Then, we apply Hoeffding's inequality to bound the deviation E_1, by observing that $\sum_{a\in\mathcal{A}} \left(\overline{\pi}(a|s) - \pi(a|s)\right) \widehat{q}_i(s,a)$ involved in the estimation $\widehat{A}^{\overline{\pi}}_{\pi}(s)$ are all independent and belong to $\left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]$. Therefore, we have:

$$\begin{aligned}
\Pr\left(\left| \widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi,T}(s) \right| \geq \frac{3\epsilon}{8(1-\gamma)} \right) &\leq 2e^{-\frac{2\left(\frac{3\epsilon}{8(1-\gamma)}\right)^2 N^2}{\sum_{i=1}^{N} (b_i - a_i)^2}} \\
&= 2e^{-\frac{18\epsilon^2 N^2 (1-\gamma)^2}{256(1-\gamma)^2 N}} = 2e^{-\frac{9\epsilon^2 N}{128}}.
\end{aligned}$$

Taking the union bound over the state space $\mathcal{S}$ we have to solve for N

$$2|\mathcal{S}|e^{-\frac{9\epsilon^2 N}{128}} \leq \delta \implies N \geq \frac{128}{9\epsilon^2}\log\frac{2|\mathcal{S}|}{\delta}.$$

We now manage the bias E_2 due to the finite horizon:

$$\begin{aligned}\left|A^{\overline{\pi}}_{\pi,T}(s) - A^{\overline{\pi}}_{\pi}(s)\right| &= \left|\sum_{a\in\mathcal{A}}\left(\overline{\pi}(a|s) - \pi(a|s)\right)\left(Q^{\pi}_{T}(s,a) - Q^{\pi}(s,a)\right)\right| \\ &\leq \frac{1}{2}\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1 \operatorname{sp}(Q^{\pi}_{T}(s,\cdot) - Q^{\pi}(s,\cdot)) \\ &\leq \left\|\mathbf{q}^{\pi} - \mathbf{q}^{\pi}_{T}\right\|_\infty \leq \frac{\gamma^T}{1-\gamma},\end{aligned}$$

where we exploited Lemma 23 in the second line and Lemma 27 in the last line. Thus, by selecting $T = \lceil \log_\gamma \frac{\epsilon}{8} \rceil$ we have that the bias E_2 is bounded by $\frac{\epsilon}{8(1-\gamma)}$. ■

The following is an intermediate result that is employed for the proof of Lemma 20.

Lemma 28. *Let $\pi, \overline{\pi} \in \Pi$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_\gamma \frac{\epsilon}{8} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{128}{9\epsilon^2}\log\frac{2|\mathcal{S}|}{\delta} \right\rceil.$$

The sampling procedure constructs a function $A^{\overline{\pi}}_{\pi}(s)$ such that with probability $1-\delta$, simultaneously for all $s \in \mathcal{S}$:

$$\frac{\left|\widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi}(s)\right|}{\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1} \leq \frac{\epsilon}{2(1-\gamma)}.$$

Proof The proof is not dissimilar from that of Lemma 18. We partition the error into:

$$\frac{\left|\widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi}(s)\right|}{\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1} \leq \underbrace{\frac{\left|\widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi,T}(s)\right|}{\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1}}_{(E_1)} + \underbrace{\frac{\left|A^{\overline{\pi}}_{\pi,T}(s) - A^{\overline{\pi}}_{\pi}(s)\right|}{\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1}}_{(E_2)}.$$

We set $E_1 = \frac{3\epsilon}{8(1-\gamma)}$ and $E_2 = \frac{\epsilon}{8(1-\gamma)}$ so that $E_1 + E_2 = \frac{\epsilon}{2(1-\gamma)}$. Then, we apply Hoeffding's inequality to bound the deviation E_1. We first bound uniformly the terms involved in the sample mean:

$$\frac{\left|\sum_{a\in\mathcal{A}}\left(\overline{\pi}(a|s) - \pi(a|s)\right)\widehat{q}_i(s,a)\right|}{\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1} \leq \frac{1}{1-\gamma}. \tag{P.34}$$

Therefore, we can apply Hoeffding's inequality with ranges $\left[-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}\right]$:

$$\begin{aligned}\Pr\left(\frac{\left|\widehat{A}^{\overline{\pi}}_{\pi}(s) - A^{\overline{\pi}}_{\pi,T}(s)\right|}{\left\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\right\|_1} \geq \frac{3\epsilon}{8(1-\gamma)}\right) &\leq 2e^{-\frac{2\left(\frac{3\epsilon}{8(1-\gamma)}\right)^2 N^2}{\sum_{i=1}^{N}(b_i-a_i)^2}} \\ &= 2e^{-\frac{18\epsilon^2 N^2(1-\gamma)^2}{256(1-\gamma)^2 N}} = 2e^{-\frac{9\epsilon^2 N}{128}}.\end{aligned}$$

Performing a union bound over the state space $\mathcal{S}$ and solving for N, we get the first result. Now we manage E_2:

$$\begin{aligned}\frac{\left|A^{\overline{\pi}}_{\pi,T}(s) - A^{\overline{\pi}}_{\pi}(s)\right|}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} &= \frac{\left|\sum_{a\in\mathcal{A}}\left(\overline{\pi}(a|s) - \pi(a|s)\right)\left(Q^{\pi}_T(s,a) - Q^{\pi}(s,a)\right)\right|}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \\ &\leq \frac{1}{2}\operatorname{sp}(Q^{\pi}_T(s,\cdot) - Q^{\pi}(s,\cdot)) \\ &\leq \frac{1}{2}\left\|\mathbf{q}^{\pi} - \mathbf{q}^{\pi}_T\right\|_{\infty} \leq \frac{\gamma^T}{2(1-\gamma)},\end{aligned}$$

where we exploited Lemma 23 in the second line and Lemma 27 in the last line. Thus, by selecting $T = \lceil \log_{\gamma} \frac{\epsilon}{4} \rceil$ we have that the bias E_2 is bounded by $\frac{\epsilon}{8(1-\gamma)}$. Even more so for $T = \lceil \log_{\gamma} \frac{\epsilon}{8} \rceil$ as required by the statement. ■

Lemma 20. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$M = \left\lceil \frac{8}{\epsilon^2} \log \frac{2}{\delta} \right\rceil.$$

Under the assumptions of Lemma 19, with probability $1 - 2\delta$ it holds that:

$$\widehat{g} - g \leq \frac{\epsilon}{(1-\gamma)^2}.$$

Proof First, we rephrase g to highlight the presence of expectations:

$$g = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[I\left(s \in \mathcal{S}^{\overline{\pi}}_{\pi}\right) \frac{A^{\overline{\pi}}_{\pi}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right]. \tag{P.35}$$

Second, we consider another quantity $\widetilde{g}$ defined as the expected value under d^{π} of the approximate advantage function and indicator function, estimated as in Lemma 19:

$$\widetilde{g} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[I\left(s \in \widehat{\mathcal{S}}^{\overline{\pi}}_{\pi}\right) \frac{\widehat{A}^{\overline{\pi}}_{\pi}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right]. \tag{P.36}$$

Third, we define the expected value under d^{π}_T of the approximate advantage function and indicator function:

$$\widetilde{g}_T = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}_T} \left[I\left(s \in \widehat{\mathcal{S}}^{\overline{\pi}}_{\pi}\right) \frac{\widehat{A}^{\overline{\pi}}_{\pi}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right]. \tag{P.37}$$

Now, we consider the following decomposition: $\widehat{g} - g = \underbrace{\widehat{g} - \widetilde{g}_T}_{(E_1)} + \underbrace{\widetilde{g}_T - \widetilde{g}}_{(E_2)} + \underbrace{\widetilde{g} - g}_{(E_3)}$. We partition the error in the following way:

$$E_1 = \frac{\epsilon}{4(1-\gamma)^2}, \quad E_2 = \frac{\epsilon}{4(1-\gamma)^2}, \quad E_3 = \frac{\epsilon}{2(1-\gamma)^2}.$$

Let us start with E_1, by observing that $\widehat{g}$ is an unbiased estimator for $\widetilde{g}_T$:

$$\widehat{g} - \widetilde{g}_T = \frac{1}{M(1-\gamma)} \sum_{i=1}^{M} I\left(s_i \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s_i)}{\|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1} - \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_T^\pi} \left[I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right].$$

If we rename $f(s) = \frac{1}{1-\gamma} I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1}$, we are comparing a sample mean $\widehat{g}$ and the corresponding expectation $\widetilde{g}$. Notice that, although $f(s)$ are random variables as well they are independent of the samples $s_i \sim d_T^\pi$. First, we provide a uniform bound of $f(s_i)$ and then we apply Hoeffding's inequality. When $s \notin \widehat{\mathcal{S}}_\pi^{\overline{\pi}}$, we have $f(s_i) = 0$. Otherwise, by definition of $\widehat{\mathcal{S}}_\pi^{\overline{\pi}}$, we have $\widehat{A}_\pi^{\overline{\pi}}(s_i) \geq \frac{\epsilon}{2(1-\gamma)}$ and, in particular, $f(s_i) \geq 0$. We now upper bound $f(s_i)$:

$$\begin{aligned} f(s_i) &\leq \frac{1}{1-\gamma} \frac{\widehat{A}_\pi^{\overline{\pi}}(s_i)}{\|\overline{\pi}(\cdot|s_i) - \pi(\cdot|s_i)\|_1} \\ &= \frac{1}{1-\gamma} \frac{\left|\sum_{a\in\mathcal{A}} \left(\overline{\pi}(a|s_i) - \pi(a|s_i)\right) q_i(s,a)\right|}{\sum_{a\in\mathcal{A}} |\overline{\pi}(a|s_i) - \pi(a|s_i)|} \\ &\leq \frac{1}{(1-\gamma)^2} \frac{\sum_{a\in\mathcal{A}} |\overline{\pi}(a|s_i) - \pi(a|s_i)|}{\sum_{a\in\mathcal{A}} |\overline{\pi}(a|s_i) - \pi(a|s_i)|} = \frac{1}{(1-\gamma)^2}. \end{aligned}$$

Thus for all s_i we have that $f(s_i)$ ranges in $\left[0, \frac{1}{(1-\gamma)^2}\right]$. Therefore, we can apply Hoeffding's bound:

$$\Pr\left(|\widehat{g} - \widetilde{g}_T| \geq \frac{\epsilon}{4(1-\gamma)^2}\right) \leq 2\exp\left(-\frac{2\left(\frac{\epsilon}{4(1-\gamma)^2}\right)^2 M}{\frac{1}{(1-\gamma)^4}}\right) \leq 2\exp\left(-\frac{\epsilon^2 M}{8}\right),$$

from which we get the value of M. Let us now consider E_2:

$$\widetilde{g}_T - \widetilde{g} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_T^\pi} \left[I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right] - \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^\pi} \left[I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \frac{\widehat{A}_\pi^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right].$$

We can bound the absolute value of the difference by recalling that $|f(s_i)| \leq \frac{1}{(1-\gamma)^2}$ and using Lemma 26:

$$|\widetilde{g}_T - \widetilde{g}| \leq \frac{1}{(1-\gamma)^2} \|\mathbf{d}^\pi - \mathbf{d}_T^\pi\|_1 \leq \frac{2\gamma^T}{(1-\gamma)^2}.$$

Then, by selecting $T = \lceil \log_\gamma \frac{\epsilon}{8} \rceil$, we are guaranteed that $|\widehat{g}_T - \widetilde{g}| \leq \frac{\epsilon}{4(1-\gamma)^2}$ Let us now consider E_3. First observe that with probability at least $1-\delta$, $\widehat{S}_\pi^{\overline{\pi}}$ is a subset of $S_\pi^{\overline{\pi}}$ and

thus $I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right) \leq I\left(s \in \mathcal{S}_{\pi}^{\overline{\pi}}\right)$. Therefore, we have:

$$\begin{aligned}
\widetilde{g} - g &= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right) \frac{\widehat{A}_{\pi}^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right] - \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[I\left(s \in \mathcal{S}_{\pi}^{\overline{\pi}}\right) \frac{A_{\pi}^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right] \\
&\leq \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right) \frac{\widehat{A}_{\pi}^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right] - \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right) \frac{A_{\pi}^{\overline{\pi}}(s)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right] \\
&= \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[\frac{I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right)}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \left(\widehat{A}_{\pi}^{\overline{\pi}}(s) - A_{\pi}^{\overline{\pi}}(s) \right) \right] \\
&\leq \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}} \left[\frac{\left| \widehat{A}_{\pi}^{\overline{\pi}}(s) - A_{\pi}^{\overline{\pi}}(s) \right|}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \right].
\end{aligned}$$

This term can be treated in way similar to what done in Lemma 19. Lemma 28 proofs that, with probability at least $1-\delta$ and estimating $\widehat{A}_{\pi}^{\overline{\pi}}(s)$ with N samples and horizon T as defined in Lemma 19 we have that, simultaneously for all states $s \in \mathcal{S}$:

$$\frac{\left| \widehat{A}_{\pi}^{\overline{\pi}}(s) - A_{\pi}^{\overline{\pi}}(s) \right|}{\|\overline{\pi}(\cdot|s) - \pi(\cdot|s)\|_1} \leq \frac{5\epsilon}{12(1-\gamma)} \implies \widetilde{g} - g \leq \frac{\epsilon}{2(1-\gamma)^2}.$$

Observing that we combined two results that hold with probability at least $1-\delta$, the statement holds with probability at least $1-2\delta$. ■

Lemma 21. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$T = \left\lceil \log_{\gamma} \frac{\epsilon}{48} \right\rceil \qquad \text{and} \qquad N = \left\lceil \frac{128}{\epsilon^2} \log \frac{2|\mathcal{S}||\mathcal{A}|}{\delta} \right\rceil.$$

The sampling procedure constructs a function $\widehat{Q}_{\pi}(s,a)$ such that with probability $1-\delta$, simultaneously for all $s \in \mathcal{S}$:

$$\left| \widehat{Q}^{\pi}(s,a) - Q^{\pi}(s,a) \right| \leq \frac{\epsilon}{12(1-\gamma)}$$

Proof Similarly to the proof of Lemma 15, we partition the error into $E_1 = \frac{\epsilon}{16(1-\gamma)}$ and $E_2 = \frac{\epsilon}{48(1-\gamma)}$, so that $E_1 = E_2 = \frac{\epsilon}{12(1-\gamma)}$. Then, we apply Hoeffding's inequality to bound the deviation, by observing that the terms $\widehat{q}_i(s,a)$ involved in the estimation $\widehat{A}_{\pi}^{\overline{\pi}}(s)$ are all independent and belong to $\left[0, \frac{1}{1-\gamma}\right]$. Therefore, we have:

$$\begin{aligned}
\Pr\left(\left| \widehat{Q}^{\pi}(s,a) - Q_T^{\pi}(s,a) \right| \geq \frac{\epsilon}{16(1-\gamma)} \right) &\leq 2e^{-\frac{2\left(\frac{\epsilon}{(1-\gamma)}\right)^2 N^2}{\sum_{i=1}^{N}(b_i - a_i)^2}} \\
&= 2e^{-\frac{2\epsilon^2 N^2 (1-\gamma)^2}{256(1-\gamma)^2 N}} = 2e^{-\frac{\epsilon^2 N}{128}}.
\end{aligned}$$

Taking the union bound over the state–action space $\mathcal{S} \times \mathcal{A}$ and solving for N, we get:

$$2|\mathcal{S}||\mathcal{A}|e^{-\frac{\epsilon^2 N}{128}} \leq \delta \implies N \geq \frac{128}{\epsilon^2} \log \frac{2|\mathcal{S}||\mathcal{A}|}{\delta}.$$

We now manage the bias introduced by the truncation on length T:

$$|Q_T^\pi(s,a) - Q^\pi(s,a)| \leq \frac{\gamma^T}{1-\gamma}.$$

By selecting $T = \lceil \log_\gamma \frac{\epsilon}{48} \rceil$ we have that the bias E_2 is bounded by $\frac{\epsilon}{48(1-\gamma)}$. ■

Lemma 22. *Let $\pi, \overline{\pi} \in \Pi^{SR}$ be two Markovian stationary policies for the same MDP $\mathcal{M}$. Let*

$$M = \left\lceil \frac{288}{121\epsilon^2} \log \frac{2}{\delta} \right\rceil.$$

Under the assumptions of Lemma 19, the sampling procedure constructs a function $\widehat{g}$ such that with probability $1 - 2\delta$:

$$\widehat{g} - g \leq \frac{\epsilon}{(1-\gamma)^2}.$$

Proof The proof is analogous to that of Lemma 20. Let us first rephrase g:

$$g = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d^\pi} \left[I\left(s \in \mathcal{S}_\pi^{\overline{\pi}}\right) \left(Q^\pi(s, a_s^\uparrow) - Q^\pi(s, a_s^\downarrow) \right) \right].$$

We now define $\widetilde{g}$ and $\widetilde{g}_T$ in a way analogous to Lemma 20:

$$\widetilde{g} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d^\pi} \left[I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \left(\widehat{Q}^\pi(s, \widehat{a}_s^\uparrow) - \widehat{Q}^\pi(s, \widehat{a}_s^\downarrow) \right) \right],$$

$$\widetilde{g} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s \sim d_T^\pi} \left[I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \left(\widehat{Q}^\pi(s, \widehat{a}_s^\uparrow) - \widehat{Q}^\pi(s, \widehat{a}_s^\downarrow) \right) \right].$$

Again we decompose the error into: $\widehat{g} - g = \underbrace{\widehat{g} - \widetilde{g}_T}_{(E_1)} + \underbrace{\widetilde{g}_T - \widetilde{g}}_{(E_2)} + \underbrace{\widetilde{g} - g}_{(E_3)}$ and we partition the error in the following way:

$$E_1 = \frac{11\epsilon}{24(1-\gamma)^2}, \quad E_2 = \frac{\epsilon}{24(1-\gamma)^2}, \quad E_3 = \frac{1\epsilon}{2(1-\gamma)^2}.$$

Concerning E_1, having observed that $I\left(s \in \widehat{\mathcal{S}}_\pi^{\overline{\pi}}\right) \left(\widehat{Q}^\pi(s, \widehat{a}_s^\uparrow) - \widehat{Q}^\pi(s, \widehat{a}_s^\downarrow) \right) \in \left[0, \frac{1}{1-\gamma}\right]$, we can apply Hoeffding's inequality, as in Lemma 20:

$$\begin{aligned} \Pr\left(|\widehat{g} - \widetilde{g}_T| \geq \frac{11\epsilon}{24(1-\gamma)^2} \right) &\leq 2\exp\left(-\frac{2\left(\frac{11\epsilon}{24(1-\gamma)^2}\right)^2 M}{\frac{1}{(1-\gamma)^4}} \right) \\ &\leq 2\exp\left(-\frac{242\epsilon^2 M}{576} \right) \end{aligned}$$

$$= 2\exp\left(-\frac{121\epsilon^2 M}{288}\right),$$

from which, we get the same value of M. For E_2, similarly to Lemma 20, we have:

$$|\widetilde{g}_T - \widetilde{g}| \leq \frac{1}{(1-\gamma)^2}\left\|\mathbf{d}^{\pi} - \mathbf{d}_T^{\pi}\right\|_1 \leq \frac{2\gamma^T}{(1-\gamma)^2},$$

obtaining the horizon value $T = \lceil \log_\gamma \frac{\epsilon}{48} \rceil$. Finally, for E_3 we have:

$$\begin{aligned}
\widetilde{g} - g &= \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}}\left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right)\left(\widehat{Q}^{\pi}(s,\widehat{a}_s^{\uparrow}) - \widehat{Q}^{\pi}(s,\widehat{a}_s^{\downarrow})\right)\right] \\
&\quad - \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}}\left[I\left(s \in \mathcal{S}_{\pi}^{\overline{\pi}}\right)\left(Q^{\pi}(s,a_s^{\uparrow}) - Q^{\pi}(s,a_s^{\downarrow})\right)\right] \\
&\leq \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}}\left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right)\left(\widehat{Q}^{\pi}(s,\widehat{a}_s^{\uparrow}) - \widehat{Q}^{\pi}(s,\widehat{a}_s^{\downarrow})\right)\right] \\
&\quad - \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}}\left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right)\left(Q^{\pi}(s,a_s^{\uparrow}) - Q^{\pi}(s,a_s^{\downarrow})\right)\right] \\
&\leq \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}}\left[I\left(s \in \widehat{\mathcal{S}}_{\pi}^{\overline{\pi}}\right)\left|\left(\widehat{Q}^{\pi}(s,\widehat{a}_s^{\uparrow}) - \widehat{Q}^{\pi}(s,\widehat{a}_s^{\downarrow})\right) - \left(Q^{\pi}(s,a_s^{\uparrow}) - Q^{\pi}(s,a_s^{\downarrow})\right)\right|\right] \\
&\leq \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}}\left[\left|\left(\widehat{Q}^{\pi}(s,\widehat{a}_s^{\uparrow}) - \widehat{Q}^{\pi}(s,\widehat{a}_s^{\downarrow})\right) - \left(Q^{\pi}(s,a_s^{\uparrow}) - Q^{\pi}(s,a_s^{\downarrow})\right)\right|\right] \leq \frac{\epsilon}{2(1-\gamma)^2},
\end{aligned}$$

where we used Equation (22) in the last line. Putting all together and recalling that we used two events that hold with probability at least $1-\delta$ we get the result. ■

Appendix B. Global Safe Policy Improvement

In this appendix, we further investigate the optimization of the lower bound of Corollary 4, without constraining the choice of a target policy. We name this approach Global Safe Policy Iteration (GSPI) and will show that SASPI when selecting a greedy policy π^+ as target policy allows obtaining the global optimum of the bound.

The maximization of the bound in Corollary 4 can be formulated as a Quadratic Programming (QP) problem over the policy space Π:

$$\begin{aligned}
\max_{\boldsymbol{\pi'}\in\Pi} \quad & \mathbf{d}^{\pi\mathrm{T}}\left(\boldsymbol{\pi'} - \boldsymbol{\pi}\right)\mathbf{q}^{\pi} - \frac{\gamma}{2(1-\gamma)^2}\left\|\mathbf{q}^{\pi}\right\|_{\infty} z^2 \\
\text{s.t.} \quad & \boldsymbol{\pi'}_{ij} \geq 0, \quad \forall i,j \\
& \sum_j \boldsymbol{\pi'}_{ij} = 1, \quad \forall i \\
& \sum_j \left|\boldsymbol{\pi'}_{ij} - \boldsymbol{\pi}_{ij}\right| \leq z, \quad \forall i
\end{aligned}$$

However, previous formulation is not suited for a QP solver, it is necessary to introduce additional auxiliary variables in order to remove absolute values. After this manipulation we

obtain a QP problem with $2|\mathcal{S}||\mathcal{A}|$ variables. While QP problems can be solved using interior point approaches (Boyd and Vandenberghe, 2004), we are going to present an iterative algorithm that is able to attain the optimal solution and presents lower computational complexity.

We adopt an approach analogous to that of SSPI and SASPI. If we were able to compute the optimal budget Υ^*, then we can move probability across actions provided that we satisfy $\|\boldsymbol{\pi'} - \boldsymbol{\pi}\|_\infty \leq \Upsilon^*$. The natural choice consists in trying to move as much probability as possible on the action with highest Q^π, i.e., to one of the greedy actions a_s^+, while taking away probability from the suboptimal actions, starting from the worst action and following an increasing order. The amount of probability we can move around is bounded by $Y/2$, that we assume fixed. Let us start by stating the following optimality condition.

Corollary 29. *Let $\mathcal{S}_\pi^{\overline{\pi}}$ be the subset of states $d^\pi(s)$ is positive: $\mathcal{S}_\pi^{\overline{\pi}} = \{s \in \mathcal{S} : d^\pi(s) > 0\}$. Let $\mathcal{A}_s^+ = \{a \in \mathcal{A} : Q^\pi(s,a) = \max_{a' \in \mathcal{A}} Q^\pi(s,a')\}$ be the set of greedy actions in state $s \in \mathcal{S}$ and let $a_s^+ \in \mathcal{A}_s^+$. The bound in Corollary 4 is optimized any of the policies π' defined as:*

$$\pi'(a|s) = \begin{cases} \min\left\{\frac{\Upsilon^*}{2}, 1 - \pi\left(a|s\right)\right\} & \text{if } a = a_s^+ \\ \max\left\{0, \min\left\{\pi(a|s), \frac{\Upsilon}{2} - G^{\downarrow}\left(s,a\right)\right\}\right\} & \text{otherwise} \end{cases},$$

where Υ^* is the value that maximizes the following function:

$$\begin{aligned} B(\Upsilon) = &\frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} d^\pi(s) \min\left\{\frac{\Upsilon}{2}, 1 - \pi\left(a_s^+ | \, s\right)\right\} Q^\pi(s, a_s^+) \\ &- \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}_\pi^{\overline{\pi}}} d^\pi(s) \sum_{a \in \mathcal{A} \setminus \{a_s^+\}} \max\left\{0, \min\left\{\pi\left(a | \, s\right), \frac{\Upsilon}{2} - G^{\uparrow}(s,a)\right\}\right\} Q^\pi(s,a) \\ &- \frac{\gamma}{(1-\gamma)^2} \Upsilon^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2}. \end{aligned}$$

Proof First of all we observe that if $d^\pi(s) = 0$, state s has no contribution to the bound, thus we restrict our attention to the states in $\mathcal{S}_\pi^{\overline{\pi}}$. Fix a budget Υ, as for SASPI, for every state we can use $\Upsilon/2$ to increase the probability of some actions in $\mathcal{A}_s^{\uparrow}$ and $\Upsilon/2$ to decrease the probability of some actions in $\mathcal{A}_s^{\downarrow}$. As already mentioned, the best we can do is to move as much probability as possible to the action with highest Q^π, i.e., to one of the greedy actions $a_s^+ \in \mathcal{A}_s^+$.[16] We need to compensate this operation by removing probability from actions in $\mathcal{A}_s^{\downarrow}$ in increasing order of Q^π. Therefore, we can rewrite Corollary 4 as:

$$\begin{aligned} J^{\pi'} - J^\pi \geq &\frac{1}{1-\gamma} \mathbf{d}^{\pi\intercal} \mathbf{a}_\pi^{\pi'} - \frac{\gamma}{(1-\gamma)^2} \left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_\infty^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2} \\ = &\frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \left[\pi'(a|s) Q^\pi(s,a) - \pi(a|s) Q^\pi(s,a)\right] - \frac{\gamma}{(1-\gamma)^2} \left\|\boldsymbol{\pi'} - \boldsymbol{\pi}\right\|_\infty^2 \|\mathbf{q}^\pi\|_\infty \\ = &\frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} d^\pi(s) \left[\Delta(s, a_s^+) Q^\pi(s, a_s^+) + \sum_{a \in \mathcal{A} \setminus \{a_s^+\}} \Delta(s,a) Q^\pi(s,a)\right] - \frac{\gamma}{(1-\gamma)^2} \Upsilon^2 \frac{\|\mathbf{q}^\pi\|_\infty}{2}, \end{aligned}$$

16. For simplicity, we consider only deterministic greedy policies. Clearly, any mixture of them maximizes the bound too.

Algorithm 12 Exact GSPI.

input: MDP $\mathcal{M}$
Initialize π
$\pi^+ \leftarrow \text{GPC}(\mathcal{M}, \Pi^{\text{SD}}, \pi)$
$\Upsilon^* \leftarrow \text{FBO}(\mathcal{M}, \pi^+, \pi)$
while $\Upsilon^* > 0$ **do**
Compute $G_s^{\downarrow}(a),\ \forall s \in \mathcal{S},\ \forall a \in \mathcal{A}$

$$\pi'(a|s) \leftarrow \begin{cases} \min\left\{\frac{\Upsilon^*}{2}, 1 - \pi\left(a|s\right)\right\} & \text{if } a = a_s^+ \\ \max\left\{0, \min\left\{\pi(a|s), \frac{\Upsilon}{2} - G^{\downarrow}\left(s,a\right)\right\}\right\} & \text{otherwise} \end{cases}, \ \forall s \in \mathcal{S},\ \forall a \in \mathcal{A}$$

$\Upsilon^* \leftarrow \text{FBO}(\mathcal{M}, \pi^+, \pi)$
end while

where we denoted with $\Delta(s,a) = \pi'(a|s) - \pi(a|s)$ and $\Upsilon^2 = \|\boldsymbol{\pi'} - \boldsymbol{\pi}\|_\infty^2$. In order to assign as much probability as possible to one greedy action we need to set:

$$\Delta(s, a_s^+) = \min\left\{\frac{\Upsilon}{2}, 1 - \pi\left(a_s^+|s\right)\right\}, \quad \forall s \in \mathcal{S}. \tag{P.38}$$

As explained before, an equivalent decrement must be partitioned over the non–greedy actions according to the ordering in ρ_s^π. The operation is summarized by the following equation:

$$\Delta(s,a) = -\max\left\{0, \min\left\{\pi(a|s), \frac{\Upsilon}{2} - G^{\downarrow}\left(s,a\right)\right\}\right\}, \quad \forall s \in \mathcal{S}, \forall a \neq a_s^+.$$

■

The pseudocode of GSPI is reported in Algorithm 12. Therefore, the only real problem is now to find the optimal budget Υ^*. This search formulates, once again, the trade-off between further increasing the budget Υ, and incur in a larger penalty, while obtaining a gain by exploiting the gap between $Q^\pi(s, a_s^+)$ and the smallest $Q^\pi(s,a)$ that still has probabilities to reduce. As the worst actions have their probability reduced to zero and saturates, this gap becomes smaller. It is straightforward to see that GSPI is equivalent to SASPI when selecting as target policy a greedy policy π^+. For this reason, the computation of the optimal budget Υ^* can be performed by employing the same FBO (Algorithm 4) and FJP (Algorithm 7) functions. However, we are able to provide for this specific scenario a simplified algorithm, as reported in Algorithm 13. The computational complexity of Algorithm 13 is dominated by the sorting of state-action pairs and thus it is $O(|\mathcal{S}||\mathcal{A}| \log |\mathcal{S}||\mathcal{A}|)$.

Figure 14 reports an example of policy update.

Remark 13 (Global Optimality of GSPI)**.** The optimality of the Υ^* computed with Algorithm 4 is clearly guaranteed by the derivative. For a fixed Υ^* computing the best policy $\boldsymbol{\pi'}$ is also straightforward. The only detail left to guarantee that $\boldsymbol{\pi'}$ is indeed the policy that maximizes the bound in Corollary 4 is to note that all possible $\Upsilon \in [0, 2]$ are considered by the algorithm, and that for every fixed Υ the updates are carried out on the

Algorithm 13 Computing of the jump points for GSPI (Find Jump Point - FJP)

input: MDP $\mathcal{M}$, current policy π
Initialize $t \leftarrow 0$, $\frac{\Upsilon_0}{2} \leftarrow 0$
Compute $G^{\downarrow}(s,a)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$
Sort the state-action pairs in increasing order of $G^{\downarrow}_s(a)$, i.e., $i < j \implies G^{\downarrow}_{s_i}(a_i) \leq G^{\downarrow}_{s_j}(a_j)$
Compute $a^{\uparrow}_s = \arg\max_{a \in \mathcal{A}} \{Q^{\pi}(s,a)\}$ and $a^{\downarrow}_s = \arg\min_{a \in \mathcal{A}} \{Q^{\pi}(s,a)\}$, $\forall s \in \mathcal{S}$
$q_0 \leftarrow \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}^{\overline{\pi}}_{\pi}} d^{\pi}(s) \left(Q^{\pi}(s, a^{\uparrow}_s) - Q^{\pi}(s, a^{\downarrow}_s)\right)$
yield Υ_0, q_0
while $\mathcal{S}^{\overline{\pi}}_{\pi} \neq \{\}$ **do**
  yield Υ_t, q_t
  $t \leftarrow t + 1$
  if $s_t \in$ **then**
    if $Q^{\pi}(s_t, a_t) < Q^{\pi}(s_t, a^{\downarrow}_{s_t})$ **then**
      $q_t \leftarrow q_{t-1} - d^{\pi}(s_t) \left(Q^{\pi}(s_t, a^{\downarrow}_{s_t}) - Q^{\pi}(s_t, a_t)\right)$
      $a^{\downarrow}_{s_t} \leftarrow a_t$
    else
      $q_t \leftarrow q_{t-1} - d^{\pi}(s_t) \left(Q^{\pi}(s_t, a^{\downarrow}_{s_t}) - Q^{\pi}(s_t, a_t)\right)$
      $\mathcal{S}^{\overline{\pi}}_{\pi} \leftarrow \mathcal{S}^{\overline{\pi}}_{\pi} \setminus \{s_t\}$
    end if
    $\frac{\Upsilon_t}{2} \leftarrow G^{\downarrow}(s_t, a_t)$
  end if
  yield Υ_t, q_t
end while
yield $2, -\infty$

single state-action level. This implies that all possible stochastic policies, and therefore all policies, are candidates for GSPI, and that the one selected is truly the one that maximizes the improvement. As an example, USPI tunes only a single α parameter, and cannot fully optimize the bound by using single state-actions update to always find the best gain when choosing a larger $\|\boldsymbol{\pi}' - \boldsymbol{\pi}\|_{\infty}$.

Example 2. *The initial policy (blue area with dotted mark) and the updated policy (orange striked area with square mark) are depicted in Figure 14. Actions are ordered according to ρ^{π}_s, i.e., in ascending order according to their Q–values. Suppose that $\Delta(s, a^{+}_s) = 0.6 \leq 1 - \pi^{+}(a^{+}_s|s)$. Starting from the worse action (a_1), we have to compensate for the increment. Neither action a_1 nor action a_2 are able to consume the budget ($\Delta(s,a_1) = -0.1$, $\Delta(s,a_2) = -0.3$). When action a_3 is faced the available budget is $\Delta(s, a^{+}_s) - G^{\pi}(s,a_3) = \Delta(s, a^{+}_s) - (\pi(a_1|s) + \pi(a_2|s)) = 0.2$. Since the probability of action a_3, the decrement $\Delta(s,a_3)$ is equal to the remaining budget (-0.2). The updates $\Delta(s,a)$ are reported in figure as dashed arrows. The relationship between the value $\Delta(s,a)$ and the coefficient $\alpha(s,a)$ of the convex combination between π and π^{+} is a factor $(\pi^{+}(a|s) - \pi(a|s))^{-1}$. Coefficients $\alpha(s,a_i)$ are drawn above the bars.*

Appendix C. Sampling Procedures

In this appendix, we report the pseudocodes of the sampling procedures employed in Section 5.

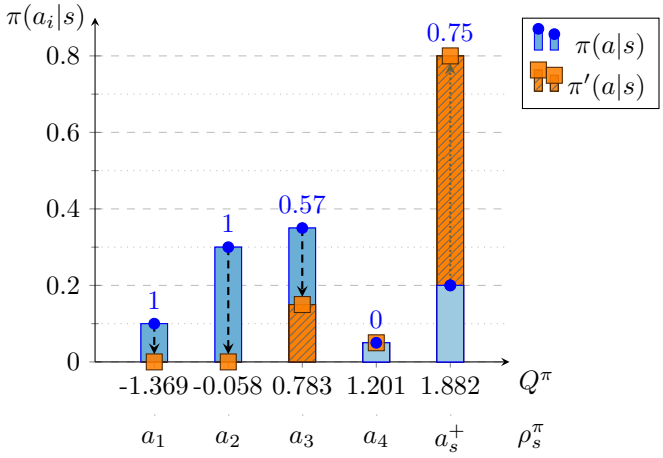


Figure 14: GSPI policy update.

Algorithm 14 d^π Sampling Procedure (d^π-sample).

input: policy π, number of samples N
Initialize $\mathcal{D} \leftarrow \{\}$
while $|\mathcal{D}| < N$ **do**
  $s_0 \sim \mu$
  for $j = 0, \ldots, T-1$ **do**
    $X \sim \text{Ber}(\gamma)$
    if $X = 0$ **then**
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{s_j\}$
      break
    end if
    $a_j \sim \pi(\cdot|s_j)$
    $s_{j+1} \sim \mathcal{P}(\cdot|s_j, a_j)$
  end for
end while
return $\mathcal{D}$

Algorithm 15 GPC and aUSPI Sampling Procedure (aUSPI-sample).

input: policy π, number of samples N
Initialize $\mathcal{D} \leftarrow \{\}$
for $i = 1, \dots, N$ **do**
  $s_i \sim d^{\pi}$-sample() ▷ Sample the state
  $a_i \sim \pi(\cdot|s_i)$ ▷ Sample the action
  $\widehat{q}_i \leftarrow 0$
  $s_{i,0} \leftarrow s_i$
  $a_{i,0} \leftarrow a_i$
  for $j = 0, \dots, T-1$ **do** ▷ Generate the rollout
    $\widehat{q}_i \leftarrow \widehat{q}_i + \gamma^j \mathcal{R}(s_{i,j}, a_{i,j})$
    $s_{i,j+1} \sim \mathcal{P}(\cdot|s_{i,j}, a_{i,j})$
    $a_{i,j+1} \leftarrow \pi(\cdot|s_{i,j+1})$
  end for
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_i, a_i, \widehat{q}_i)\}$
end for
return $\mathcal{D}$

Algorithm 16 aSSPI Sampling Procedure (aSSPI-sample).

input: policy π, number of samples N
Initialize $\mathcal{D} \leftarrow \{\}$
for $s \in \mathcal{S}$ **do**
  $\mathcal{D}(s) \leftarrow \{\}$
  for $i = 1, \dots, N$ **do**
    $\mathcal{D}_i(s) \leftarrow \{\}$
    for $a \in \mathcal{A}$ **do**
      $\widehat{q}_i(s, a) \leftarrow 0$
      $s_{i,0} \leftarrow s$
      $a_{i,0} \leftarrow a$
      for $j = 0, \dots, T-1$ **do**
        $\widehat{q}_i(s, a) \leftarrow \widehat{q}_i(s, a) + \gamma^j \mathcal{R}(s_{i,j}, a_{i,j})$
        $s_{i,j+1} \sim \mathcal{P}(\cdot|s_{i,j}, a_{i,j})$
        $a_{i,j+1} \leftarrow \pi(\cdot|s_{i,j+1})$
      end for
      $\mathcal{D}_i(s) \leftarrow \{(a, \widehat{q}_i(s, a))\}$
    end for
    $\mathcal{D}(s) \leftarrow \mathcal{D}(s) \cup \{\mathcal{D}_i(s)\}$
  end for
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s, \mathcal{D}(s))\}$
end for
return $\mathcal{D}$

Algorithm 17 aSASPI Sampling Procedure (aSASPI-sample).

input: policy π, number of samples N
Initialize $\mathcal{D} \leftarrow \{\}$
for $s \in \mathcal{S}$ **do**
  for $a \in \mathcal{A}$ **do**
    $\mathcal{D}(s,a) \leftarrow \{\}$
    for $i = 1, \ldots, N$ **do**
      $\widehat{q}_i(s,a) \leftarrow 0$
      $s_{i,0} \leftarrow s$
      $a_{i,0} \leftarrow a$
      for $j = 0, \ldots, T-1$ **do**
        $\widehat{q}_i(s,a) \leftarrow \widehat{q}_i(s,a) + \gamma^j \mathcal{R}(s_{i,j}, a_{i,j})$
        $s_{i,j+1} \sim \mathcal{P}(\cdot|s_{i,j}, a_{i,j})$
        $a_{i,j+1} \leftarrow \pi(\cdot|s_{i,j+1})$
      end for
      $\mathcal{D}(s,a) \leftarrow \mathcal{D}(s,a) \cup \{\widehat{q}_i(s,a)\}$
    end for
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s, a, \mathcal{D}(s,a))\}$
  end for
end for
return $\mathcal{D}$

References

Yasin Abbasi-Yadkori, Peter L. Bartlett, and Stephen J. Wright. A fast and reliable policy improvement algorithm. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 1338–1346. JMLR.org, 2016.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 22–31. PMLR, 2017.

Mohammad Gheshlaghi Azar, Rémi Munos, Mohammad Ghavamzadeh, and Hilbert J. Kappen. Speedy q-learning. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2411–2419, 2011.

Mohammad Gheshlaghi Azar, Vicenç Gómez, and Hilbert J. Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13:3207–3245, 2012.

Dimitri P Bertsekas. Approximate policy iteration: A survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996. ISBN 1886529108.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

Imre Csiszár and János Körner. *Information Theory - Coding Theorems for Discrete Memoryless Systems, Second Edition*. Cambridge University Press, 2011. ISBN 978-0-51192188-9.

Daniela Pucci de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

Daniela Pucci De Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization theory and Applications*, 105(3):589–608, 2000.

Alain Dutech, Timothy Edmunds, Jelle Kok, Michail Lagoudakis, Michael Littman, Martin Riedmiller, Bryan Russell, Bruno Scherrer, Richard Sutton, Stephan Timmer, et al. Reinforcement learning benchmarks and bake-offs II. *Advances in Neural Information Processing Systems (NIPS)*, 17:6, 2005.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. Beyond the one-step greedy approach in reinforcement learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1386–1395. PMLR, 2018a.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. Multiple-step greedy policies in approximate and online reinforcement learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 5244–5253, 2018b.

Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, and Bruno Scherrer. Classification-based policy iteration with a critic. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1049–1056. Omnipress, 2011.

Mohammad Ghavamzadeh, Marek Petrik, and Yinlam Chow. Safe policy improvement by minimizing robust baseline regret. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2298–2306, 2016.

Moshe Haviv and Ludo Van der Heyden. Perturbation bounds for the stationary probabilities of a finite Markov chain. *Advances in Applied Probability*, 16(4):804–818, 1984.

Ronald A Howard. Dynamic programming and Markov processes. 1960.

Sham M. Kakade. A natural policy gradient. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 1531–1538. MIT Press, 2001.

Sham M. Kakade. *On the sample complexity of reinforcement learning.* PhD thesis, University of London London, England, 2003.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pages 267–274. Morgan Kaufmann, 2002.

Daphne Koller and Ronald Parr. Policy iteration for factored mdps. In Craig Boutilier and Moisés Goldszmidt, editors, *UAI '00: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Stanford University, Stanford, California, USA, June 30 - July 3, 2000*, pages 326–334. Morgan Kaufmann, 2000.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003a.

Michail G. Lagoudakis and Ronald Parr. Reinforcement learning as classification: Leveraging modern classifiers. In Tom Fawcett and Nina Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 424–431. AAAI Press, 2003b.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of a classification-based policy iteration algorithm. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 607–614. Omnipress, 2010.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *J. Mach. Learn. Res.*, 17:19:1–19:30, 2016.

Boris Lesner and Bruno Scherrer. Tight performance bounds for approximate modified policy iteration with non-stationary policies. *CoRR*, abs/1304.5610, 2013.

Boris Lesner and Bruno Scherrer. Non-stationary approximate modified policy iteration. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1567–1575. JMLR.org, 2015.

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. Configurable Markov decision processes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3491–3500, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Rémi Munos. Error bounds for approximate policy iteration. In Tom Fawcett and Nina Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 560–567. AAAI Press, 2003.

Rémi Munos. Error bounds for approximate value iteration. In Manuela M. Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1006–1011. AAAI Press / The MIT Press, 2005.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *CoRR*, abs/1705.07798, 2017.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Adaptive batch size for safe policy gradients. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3591–3600, 2017.

Theodore J. Perkins and Mark D. Pendrith. On the existence of fixed points for q-learning and sarsa in partially observable domains. In Claude Sammut and Achim G. Hoffmann, editors, *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pages 490–497. Morgan Kaufmann, 2002.

Theodore J. Perkins and Doina Precup. A convergent form of approximate policy iteration. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 1595–1602. MIT Press, 2002.

Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. Softened approximate policy iteration for Markov games. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1860–1868. JMLR.org, 2016.

Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Natural actor-critic. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*, pages 280–291. Springer, 2005.

K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, nov 2012. Version 20121115.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1394–1402, 2013a.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 307–315. JMLR.org, 2013b.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Bruno Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 386–394, 2013a.

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

Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1314–1322. JMLR.org, 2014.

Bruno Scherrer and Boris Lesner. On the use of non-stationary policies for stationary infinite-horizon Markov decision processes. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1835–1843, 2012.

Bruno Scherrer, Victor Gabillon, Mohammad Ghavamzadeh, and Matthieu Geist. Approximate modified policy iteration. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. PMLR, 2012.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1889–1897. JMLR.org, 2015.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 5192–5202, 2018.

Satinder P. Singh, Tommi S. Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. In Gerald Tesauro, David S. Touretzky, and Todd K. Leen, editors, *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 361–368. MIT Press, 1994.

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

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063. The MIT Press, 1999.

Noboru Suzuki. On the convergence of neumann series in Banach space. *Mathematische Annalen*, 220(2):143–146, 1976.

Philip Thomas. Bias in natural actor-critic algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*,

volume 32 of *JMLR Workshop and Conference Proceedings*, pages 441–448. JMLR.org, 2014.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Deep conservative policy iteration. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6070–6077. AAAI Press, 2020.

Paul Wagner. A reinterpretation of the policy oscillation phenomenon in approximate policy iteration. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 2573–2581, 2011.

Paul Wagner. Optimistic policy iteration and natural actor-critic: A unifying view and a non-optimality result. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 1592–1600, 2013.

Martin J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. *CoRR*, abs/1905.06265, 2019a.

Martin J. Wainwright. Variance-reduced q-learning is minimax optimal. *CoRR*, abs/1906.04697, 2019b.

Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE, 2007.

Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4): 593–603, 2011.