HAL
open science

Convergence of Online and Approximate Multiple-Step Lookahead Policy Iteration

Yonathan Efroni, Gal Dalal, Bruno Scherrer, Shie Mannor

To cite this version:

Yonathan Efroni, Gal Dalal, Bruno Scherrer, Shie Mannor. Convergence of Online and Approximate Multiple-Step Lookahead Policy Iteration. EWRL 2018 - 14th European workshop on Reinforcement Learning, Oct 2018, Lille, France. hal-01927977

HAL Id: hal-01927977

https://inria.hal.science/hal-01927977v1

Submitted on 20 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Convergence of Online and Approximate Multiple-Step Lookahead Policy Iteration

Yonathan Efroni*
JONATHAN.EFRONI@GMAIL.COM

Gal Dalal*
GALD@CAMPUS.TECHNION.AC.IL

Bruno Scherrer†
BRUNO.SCHERRER@INRIA.FR

Shie Mannor*
SHIE@EE.TECHNION.AC.IL

Abstract

Multiple-step lookahead policies have demonstrated high empirical competence in Reinforcement Learning, via the use of Monte Carlo Tree Search or Model Predictive Control. In a recent work Efroni et al. (2018), multiple-step greedy policies and their use in vanilla Policy Iteration algorithms were proposed and analyzed. In this work, we study multiple-step greedy algorithms in more practical setups. We begin by highlighting a counter-intuitive difficulty, arising with soft-policy updates: even in the absence of approximations, and contrary to the 1-step-greedy case, monotonic policy improvement is not guaranteed unless the update stepsize is sufficiently large. Taking particular care about this difficulty, we formulate and analyze online and approximate algorithms that use such a multi-step greedy operator.

1. Introduction

The use of the 1-step policy improvement in Reinforcement Learning (RL) was theoretically investigated under several frameworks, e.g., Policy Iteration (PI) Puterman (1994), approximate PI Bertsekas and Tsitsiklis (1995); Kakade and Langford (2002); Munos (2003), and Actor-Critic Konda and Borkar (1999); its practical uses are abundant Schulman et al. (2015); Mnih et al. (2016); Silver et al. (2017b). However, single-step based improvement is not necessarily the optimal choice. It was, in fact, empirically demonstrated that multiple-step greedy policies can perform conspicuously better. Notable examples arise from the integration of RL and Monte Carlo Tree Search Browne et al. (2012); Tesauro and Galperin (1997); Sheppard (2002); Bouzy and Helmstetter (2004); Silver et al. (2017b,a) or Model Predictive Control Negenborn et al. (2005); Ernst et al. (2009); Tamar et al. (2017).

∗. Department of Electrical Engineering, Technion, Israel Institute of Technology
†. INRIA, Villers les Nancy, France

©2018 .
License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/.

Recent work Efroni et al. (2018) provided guarantees on the performance of the multiple-step greedy policy and generalizations of it in PI. Here, we establish it in the two practical contexts of online and approximate PI. With this objective in mind, we begin by highlighting a specific difficulty: *softly updating* a policy with respect to (w.r.t.) a multiple-step greedy policy does not necessarily result in improvement of the policy (Section 4). We find this property intriguing since monotonic improvement is guaranteed in the case of soft updates w.r.t. the 1-step greedy policy, and is central to the analysis of many RL algorithms Konda and Borkar (1999); Kakade and Langford (2002); Schulman et al. (2015). We thus engineer some algorithms to circumvent this difficulty and provide some non-trivial performance guarantees, that support the interest of using multi-step greedy operators. These algorithms assume access to a generative model (Section 5) or to an approximate multiple-step greedy policy (Section 6).

2. Preliminaries

Our framework is the infinite-horizon discounted Markov Decision Process (MDP). An MDP is defined as the 5-tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ Puterman (1994), where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P \equiv P(s'|s,a)$ is a transition kernel, $R \equiv r(s,a) \in [0, R_{\max}]$ is a reward function, and $\gamma \in (0,1)$ is a discount factor. Let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ be a stationary policy, where $\mathcal{P}(\mathcal{A})$ is a probability distribution on $\mathcal{A}$. Let $v^\pi \in \mathbb{R}^{|\mathcal{S}|}$ be the value of a policy π, defined in state s as $v^\pi(s) \equiv \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t r(s_t, \pi(s_t))]$. For brevity, we respectively denote the reward and value at time t by $r_t \equiv r(s_t, \pi_t(s_t))$ and $v_t \equiv v(s_t)$. It is known that $v^\pi = \sum_{t=0}^\infty \gamma^t (P^\pi)^t r^\pi = (I - \gamma P^\pi)^{-1} r^\pi$, with the component-wise values $[P^\pi]_{s,s'} \triangleq P(s' \mid s, \pi(s))$ and $[r^\pi]_s \triangleq r(s, \pi(s))$. Lastly, let

$$q^\pi(s,a) = \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t r(s_t, \pi(s_t)) \mid s_0 = s, a_0 = a]. \tag{1}$$

Our goal is to find a policy π^* yielding the optimal value v^* such that

$$v^* = \max_\pi (I - \gamma P^\pi)^{-1} r^\pi = (I - \gamma P^{\pi^*})^{-1} r^{\pi^*}. \tag{2}$$

This goal can be achieved using the three classical operators (equalities hold component-wise):

$$\forall v, \pi,\ T^\pi v = r^\pi + \gamma P^\pi v,$$
$$\forall v,\ Tv = \max_\pi T^\pi v,$$
$$\forall v,\ \mathcal{G}(v) = \{\pi : T^\pi v = Tv\},$$

where T^π is a linear operator, T is the optimal Bellman operator and both T^π and T are γ-contraction mappings w.r.t. the max norm. It is known that the unique fixed points of T^π and T are v^π and v^*, respectively. The set $\mathcal{G}(v)$ is the standard set of 1-step greedy policies w.r.t. v.

3. The h- and κ-Greedy Policies

In this section, we bring forward necessary definitions and results on two classes of multiple-step greedy policies: h- and κ-greedy Efroni et al. (2018). Let $h \in \mathbb{N}\backslash\{0\}$. The h-greedy

policy π_h outputs the first optimal action out of the sequence of actions solving a non-stationary, h-horizon control problem as follows:

$$\forall s \in \mathcal{S},\ \pi_h(s) \in \arg\max_{\pi_0} \max_{\pi_1,..,\pi_{h-1}} \mathbb{E}^{\pi_0 \cdots \pi_{h-1}} \left[\sum_{t=0}^{h-1} \gamma^t r(s_t, \pi_t(s_t)) + \gamma^h v(s_h) \mid s_0 = s \right].$$

Since the h-greedy policy can be represented as the 1-step greedy policy w.r.t. $T^{h-1}v$, the set of h-greedy policies w.r.t. v, $\mathcal{G}_h(v)$, can be formally defined as follows:

$$\begin{aligned} \forall v, \pi,\ T_h^\pi v &= T^\pi T^{h-1} v, \\ \forall v,\ \mathcal{G}_h(v) &= \{\pi : T_h^\pi v = T^h v\}. \end{aligned}$$

Let $\kappa \in [0, 1]$. The set of κ-greedy policies w.r.t. a value function v, $\mathcal{G}_\kappa(v)$, is defined using the following operators:

$$\begin{aligned} \forall v, \pi,\ T_\kappa^\pi v &= (I - \kappa\gamma P^\pi)^{-1}(r^\pi + (1-\kappa)\gamma P^\pi v) \\ \forall v,\ T_\kappa v &= \max_{\pi'} T_\kappa^{\pi'} v = \max_{\pi'} (I - \kappa\gamma P^{\pi'})^{-1}(r^{\pi'} + (1-\kappa)\gamma P^{\pi'} v) \\ \forall v,\ \mathcal{G}_\kappa(v) &= \{\pi : T_\kappa^\pi v = T_\kappa v\}. \end{aligned} \tag{3}$$

Remark 1 *A comparison of* (2) *and* (3) *reveals that finding the κ-greedy policy is equivalent to solving a $\kappa\gamma$-discounted MDP with a shaped reward $r_{v,\kappa}^\pi \stackrel{def}{=} r^\pi + (1-\kappa)\gamma P^\pi v$.*

In (Efroni et al., 2018, Proposition 11), the κ-greedy policy was explained to be interpolating over all geometrically κ-weighted h-greedy policies. It was also shown that for $\kappa = 0$, the 1-step greedy policy is restored, while for $\kappa = 1$, the κ-greedy policy is the optimal policy.

Both T_κ^π and T_κ are ξ_κ contraction mappings, where $\xi_\kappa = \frac{\gamma(1-\kappa)}{1-\gamma\kappa} \in [0, \gamma]$. Their respective fixed points are v^π and v^*. For brevity, where there is no risk of confusion, we shall denote ξ_κ by ξ. Moreover, in Efroni et al. (2018) it was shown that both the h- and κ-greedy policies w.r.t. v^π are strictly better then π, unless $\pi = \pi^*$.

Next, let the κ-optimal q-function be defined as follows,

$$q_\kappa^\pi(s, a) = \max_{\pi'} \mathbb{E}^{\pi'} [\sum_{t=0}^{\infty} (\kappa\gamma)^t (r(s_t, \pi'(s_t)) + \gamma(1-\kappa) v^\pi(s_{t+1}) \mid s_0 = s, a_0 = a]. \tag{4}$$

The latter is the optimal q-function of the surrogate, $\gamma\kappa$-discounted MDP with v^π-shaped reward (see Remark 1). Thus, we can obtain a κ-greedy policy, $\pi_\kappa \in \mathcal{G}_\kappa(v^\pi)$, directly from q_κ^π :

$$\pi_\kappa(s) \in \arg\max_a q_\kappa^\pi(s, a),\ \forall s \in \mathcal{S}.$$

See that the greedy policy w.r.t. $q_{\kappa=0}^\pi(s, a)$ is the 1-step greedy policy since $q_{\kappa=0}^\pi(s, a) = q^\pi(s, a)$.

4. Multi-step Policy Improvement and Soft Updates

In this section, we focus on policy improvement of multiple-step greedy policies, performed with soft updates. Soft updates of the 1-step greedy policy have proved necessary and beneficial in prominent algorithms Konda and Borkar (1999); Kakade and Langford (2002); Schulman et al. (2015). Here, we begin by describing an intrinsic difficulty in selecting the step-size parameter $\alpha \in (0, 1]$ when updating with multiple-step greedy policies such as of h- and κ-greedy. Specifically, denote by π' such multiple-step greedy policy w.r.t. v^π. Then, $\pi_{new} = (1-\alpha)\pi + \alpha\pi'$ is not necessarily better than π (see Appendix A for the proof).

Theorem 2 *For any MDP, let π be a policy and v^π its value. Let $\pi_\kappa \in \mathcal{G}_\kappa(v^\pi)$ and $\pi_h \in \mathcal{G}_h(v^\pi), \alpha \in (0, 1]$ where $\kappa \in [0, 1]$ and an integer $h > 1$. Consider the mixture policies*

$$\pi(\alpha, \kappa) \stackrel{def}{=} (1-\alpha)\pi + \alpha\pi_\kappa,$$
$$\pi(\alpha, h) \stackrel{def}{=} (1-\alpha)\pi + \alpha\pi_h.$$

We have the following equivalences:

1. *The inequality $v^{\pi(\alpha,\kappa)} \geq v^\pi$ holds for all MDPs if and only if $\alpha \in [\kappa, 1]$.*

2. *The inequality $v^{\pi(\alpha,h)} \geq v^\pi$ holds for all MDPs if and only if $\alpha = 1$.*

The above inequalities hold entry-wise, with strict inequality in at least one entry unless $v^\pi = v^*$.

Theorem 2 guarantees monotonic improvement for the 1-step greedy policy as a special case when $\kappa = 0$. Hence, we get that for any $\alpha \in (0, 1]$, the mixture of any policy π and the 1-step greedy policy w.r.t. v^π is monotonically better then π. To the best of our knowledge, this result was not explicitly stated anywhere. Instead, it appeared within proofs of several famous results, e.g, (Konda and Borkar, 1999, Lemma 5.4), (Kakade and Langford, 2002, Corollary 4.2), and (Scherrer and Geist, 2014, Theorem 1).

In the rest of the paper, we shall focus on the κ-greedy policy and extend it to the online and the approximate cases. The discovery that the κ-greedy policy w.r.t. v^π is not necessarily strictly better than π will guide us in appropriately devising algorithms.

5. Online κ-Policy Iteration with Cautious Soft Updates

In Efroni et al. (2018), it was shown that using the κ-greedy policy in the improvement stage leads to a converging PI procedure – the κ-PI algorithm. This algorithm repeats i) solving the optimal policy of small-horizon surrogate MDPs with shaped reward, and ii) calculating the value of the optimal policy and use it to shape the reward of next iteration. Here, we devise a practical version of κ-PI, which is model-free, online and runs in two timescales, i.e, performs i) and ii) simultaneously.

Algorithm 1 Two-Timescale Online κ-Policy-Iteration

1: **for** $n = 0, \dots$ **do**
2: sample transition (s_n, a_n, r_n, s'_n)
3: # Fast-timescale updates
4: update $q_{n+1}(s_n, a_n)$ using TD(0) with stepsize α_n
5: $\delta_{\kappa,n} = r_n + \gamma(1-\kappa)v^{\pi}_n(s'_n) + \kappa\gamma \max_{a'} q_{\kappa,n}(s'_n, a') - q_{\kappa,n}(s_n, a_n)$
6: $q_{\kappa,n+1}(s_n, a_n) \leftarrow q_{\kappa,n}(s_n, a_n) + \alpha_n \delta_{\kappa,n}$
7: # Slow-timescale updates
8: $\pi_{n+1}(s_n) \leftarrow \pi_n(s_n) + \beta_n(b_{s_n}(q_{n+1}, q_{\kappa,n+1}, \pi_n) - \pi_n(s_n))$
9: **end for**
10: **return:** π

The method is depicted in Algorithm 1. It is similar to the asynchronous PI analyzed in Perkins and Leslie (2013), except for two major differences. First, the fast timescale tracks both $q^{\pi}, q^{\pi}_{\kappa}$ and not just q^{π}. Thus, it enables access to *both* the κ-greedy and 1-step-greedy policies. The 1-step greedy policy is attained via the $q^{\pi}(s,a)$ estimate, which is plugged into a q-learning Watkins and Dayan (1992) update rule for obtaining the κ-greedy policy. The latter essentially solves the surrogate, $\kappa\gamma$-discounted, MDP (see Remark 1). The second difference is in the slow timescale; there, the policy is updated using a new operator, b_s, as defined below. To better understand this operator, first notice that in Stochastic Approximation methods such as Algorithm 1, the policy is improved using soft updates with decaying stepsizes. However, as Theorem 2 states, monotonic improvement is not guaranteed below a certain stepsize value. Hence, for $q, q_{\kappa} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and policy π, we set $b_s(q, q_{\kappa}, \pi)$ to be the κ-greedy policy only when assured to have improvement:

$$b_s(q, q_{\kappa}, \pi) = \begin{cases} a_{\kappa}(s) & \text{if } q(s, a_{\kappa}) \geq v^{\pi}(s), \\ a_{1-step}(s) & \text{else,} \end{cases}$$

where $a_{\kappa}(s) \stackrel{\text{def}}{=} \arg\max_a q_{\kappa}(s,a)$, $a_{1-step}(s) \stackrel{\text{def}}{=} \arg\max_a q(s,a)$, and $v^{\pi}(s) = \sum_a \pi(a \mid s) q(s,a)$.

We respectively denote the state and state-action-pair visitation counters after the n-th time-step by $\nu_n(s) \stackrel{\text{def}}{=} \sum_{k=1}^{n} \mathbb{1}_{s=s_k}$ and $\phi_n(s,a) \stackrel{\text{def}}{=} \sum_{k=1}^{n} \mathbb{1}_{(s,a)=(s_k,a_k)}$. The stepsize sequences $\mu_f(\cdot), \mu_s(\cdot)$ satisfy the common assumption (B2) in Perkins and Leslie (2013), among which $\lim_{n\to\infty} \mu_s(n)/\mu_f(n) \to 0$. The second moments of $\{r_n\}$ are assumed to be bounded. Furthermore, let ν be some measure over the state space, s.t. $\forall s \in \mathcal{S},\ \nu(s) > 0$. Then, we assume to have a generative model $\mathbb{G}(\nu, \pi)$, from which we sample state s, sample an action $a \sim \pi(s)$, apply action a and receive reward r and next state s'.

The fast-timescale update rules in lines 4 and 6 can be jointly written as a sum of $H^{\pi}_{\kappa}(q, q_{\kappa})$ and a martingale difference noise.

Definition 3 *Let $q, q_{\kappa} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Then the mapping $H^{\pi}_{\kappa} : \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$ is defined as follows $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.*

$$H^{\pi}_{\kappa}(q(s,a), q_{\kappa}(s,a)) \stackrel{def}{=} \begin{bmatrix} r(s,a) + \gamma \mathbb{E}_{s',a^{\pi}} q(s', a^{\pi}) \\ r(s,a) + \gamma(1-\kappa)\mathbb{E}_{s',a^{\pi}} q(s', a^{\pi}) + \kappa\gamma \mathbb{E}_{s'} \max_{a'} q_{\kappa}(s', a') \end{bmatrix},$$

where $s' \sim P(\cdot \mid s, a), a^{\pi} \sim \pi(s')$.

The following lemma shows that, given a fixed π, H_κ^π is a contraction, equivalently to (Perkins and Leslie, 2013, Lemma 5.3) (see Appendix B for the proof).

Lemma 4 *H_κ^π is a γ-contraction in the max-norm. Its unique fixed point is $[\ q^\pi, q_\kappa^\pi\ ]^\top$, as defined in* (1) *and* (4)*.*

Finally, after several intermediate results in Appendix C and relaying on Lemma 4, we establish convergence of Algorithm 1.

Theorem 5 *The coupled process $(q_n, q_{\kappa,n}, \pi_n)$ in Algorithm 1 converges to the limit (q^*, q^*, π^*), where q^* is the optimal q-function and π^* is the optimal policy.*

For $\kappa = 1$, the fast-timescale update rule in line 6 corresponds to that of q-learning Watkins and Dayan (1992). For that κ, Algorithm 1 uses an estimated optimal q-function to update the current policy when improvement is assured. For $\kappa < 1$, the estimated κ-optimal q-function (see (4)) is used, again with the 'cautions' policy update.

6. Approximate κ-Policy Iteration with Hard Updates

Theorem 2 establishes the conditions required for guaranteed monotonic improvement of softly-updated multiple-step greedy policies. The algorithm in Section 5 then accounts for these conditions to ensure convergence. Contrarily, in this section, we derive and study an algorithm that perform hard policy-updates. Specifically, we generalize the Policy Search by Dynamic Programming (PSDP) algorithm (Bagnell et al., 2004; Scherrer, 2014). Our result exhibits a performance tradeoff in the choice of κ, with optimal performance bound achieved in intermediate $\kappa \in [0, 1]$ values.

We denote $\mathcal{G}_{\kappa,\delta,\nu}(v)$ as the set of approximate κ-greedy policies w.r.t. v, with δ approximation error under some probability measure ν.

Definition 6 (Approximate κ-greedy policy) *Let $v : \mathcal{S} \to \mathbb{R}$ be a value function, $\delta \geq 0$ a real number and ν a distribution over $\mathcal{S}$. A policy $\pi \in \mathcal{G}_{\kappa,\delta,\nu}(v)$ if $\nu T_\kappa^\pi v \geq \nu T_\kappa v - \delta$.*

The approximate κ-greedy oracle assumed here is less restrictive than the one assumed in Efroni et al. (2018). There, a uniform error over states was assumed, whereas here, the error is defined w.r.t. a specific measure, ν.

We follow the line of work of (Munos, 2003, 2007; Farahmand et al., 2010; Scherrer, 2014; Lazaric et al., 2016) and use *concentrability coefficients* to specify our performance bounds. This allows a direct comparison of the algorithms proposed here with previously studied approximate 1-step greedy algorithms. Our bounds includes the concentrability coefficient $C^{\pi^*(1)}$, e.g, Scherrer (2014, Definition 1), as well as two new coefficients $C_\kappa^{\pi^*}$ and $C_\kappa^{\pi^*(1)}$, defined as follows.

Algorithm 2 κ-PSDP

initialize $\kappa \in [0,1], \nu, \delta, v^{\pi_0}, \Pi = [\ ]$
$v \leftarrow v^{\pi_0}$
for $k = 1, ..$ **do**
  $\pi_k \leftarrow \mathcal{G}_{\kappa,\nu,\delta}(v)$
  $v \leftarrow T_\kappa^{\pi_k} v$
  $\Pi \leftarrow$Append(Π, π_k)
end for
return Π

Definition 7 (Concentrability coefficients) *Let μ, ν be measures over $\mathcal{S}$. Let $\{c^{\pi^*}(i)\}_{i=0}^{\infty}$ be the sequence of the smallest values in $[1,\infty)\cup\{\infty\}$ such that for every i, $\mu\left(P^{\pi^*}\right)^i \leq c^{\pi^*}(i)\nu$. Let $C^{\pi^*(1)}(\mu,\nu) = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i c^{\pi^*}(i)$.*

Next, let $C_\kappa^{\pi^*(1)}(\mu,\nu) = \frac{\xi}{\gamma}C^{\pi^*(1)}(\mu,\nu) + (1-\xi)\kappa c(0)$. Also, let $C_\kappa^{\pi^*}(\mu,\nu) \in [1,\infty)\cup\{\infty\}$ be the smallest value s.t. $d_{\kappa,\mu}^{\pi^*} \leq C_\kappa^{\pi^*}(\mu,\nu)\nu$, where $d_{\kappa,\mu}^{\pi^*} = (1-\xi)\mu(I - \xi D_\kappa^{\pi^*}P^{\pi^*})^{-1}$ is a probability measure and $D_\kappa^{\pi} = (1-\kappa\gamma)(I-\kappa\gamma P^{\pi})^{-1}$ is a stochastic matrix.

In the definition above, ν is the measure according to which the approximate improvement is guaranteed, while μ specifies the distribution on which one measures the loss $\mathbb{E}_{s\sim\mu}[v^*(s) - v^{\pi_k}(s)] = \mu(v^* - v^{\pi_k})$ that we wish to bound. From Definition 7 it holds that $C_{\kappa=0}^{\pi^*}(\mu,\nu) = C^{\pi^*}(\mu,\nu)$; the latter was defined in, e.g, Scherrer (2014, Definition 1).

The following proposition sheds light on the behavior of $C_\kappa^{\pi^*}(\mu,\nu)$; it shows that under certain constructions, $C_\kappa^{\pi^*}(\mu,\nu)$ decreases[1] as κ increases (see proof in Appendix D).

Proposition 8 *Let $\nu(\alpha) = (1-\alpha)\nu + \alpha\mu$. Then, for all $\kappa' > \kappa$, there exists $\alpha^* \in (0,1)$ such that $C_{\kappa'}^{\pi^*}(\mu,\nu(\alpha^*)) \leq C_\kappa^{\pi^*}(\mu,\nu)$. The inequality is strict for $C_\kappa^{\pi^*}(\mu,\nu) > 1$. For $\mu = \nu$ this implies that $C_\kappa^{\pi^*}(\nu,\nu)$ is a decreasing function of κ.*

6.1 κ-Policy Search by Dynamic Programming

The κ-PSDP depicted in Algorithm 2 returns a *sequence of deterministic policies*, Π. Given this sequence, we build a stochastic, non-stationary policy by successively running N_k steps of $\Pi[k]$, followed by N_{k-1} steps of $\Pi[k-1]$, etc, where $\{N_i\}_{i=1}^k$ are i.i.d. geometric random variables with parameter $1-\kappa$. Once this process reaches π_0, it runs π_0 indefinitely. We shall refer to this non-stationary policy as $\sigma_{\kappa,k}$, the value of this policy is $v^{\sigma_{\kappa,k}} = T_\kappa^{\Pi[k]}T_\kappa^{\Pi[k-1]}\dots T_\kappa^{\Pi[1]}v^{\pi_0}$. This algorithm generalizes the PSDP from Scherrer (2014). The 1-step improvement is replaced with the κ-greedy improvement, and unlike the PSDP the returned policy is random for $\kappa > 0$. Its performance bound is given in the following theorem (see proof in Appendix E).

1. A smaller coefficient is obviously better. The best value for any concentrability coefficient is 1.

Theorem 9 *Let $\sigma_{\kappa,k}$ be the policy at the k-th iteration of κ-PSDP and δ be the error as defined in Definition 6. Then $\mu(v^* - v^{\sigma_{\kappa,k}}) \leq \frac{C_\kappa^{\pi^*(1)}(\mu,\nu)}{1-\xi}\delta + \xi^k \frac{R_{\max}}{1-\gamma}$. Also, let $k = \left\lceil \frac{\log \frac{R_{max}}{\delta(1-\gamma)}}{1-\xi} \right\rceil$. Then $\mu(v^* - v^{\sigma_{\kappa,k}}) \leq \frac{C_\kappa^{\pi^*}(\mu,\nu)}{(1-\xi)^2} \log\left(\frac{R_{\max}}{(1-\gamma)\delta}\right)\delta + \delta$.*

For $\kappa = 1$ this bound is tighter than $\kappa = 0$. The reason is that $C^{\pi^*(1)}(\mu,\nu) > (1-\gamma)c(0)$ (see Definition 7), and thus $\frac{C^{\pi^*(1)}(\mu,\nu)}{1-\gamma}\delta > c(0)\delta$. Also, from continuity, this behavior is interpolated for a region around $\kappa = 1$, i.e., we have that κ-PSDP is generally better than PSDP. More interestingly, under the constructions in Proposition 8, the second form of bound reveals a strict improvement of the bound as κ increases. By recalling that δ is expected to be *monotonically increasing* function of κ — solving an MDP with larger horizon is 'harder' (Jiang et al., 2015) — we see the bounds demonstrate a performance tradeoff as a function of κ.

An additional advantage of this new algorithm over PSDP is reduced space complexity. This can be seen, e.g., from the $1-\xi$ in the denominator in the choice of k in the second part of Theorem 9. It shows that, since ξ is a strictly decreasing function of κ, performance is preserved with significantly fewer iterations by increasing κ. Since the size of stored policy Π is linearly dependent on the number of iterations, larger κ improves space efficiency.

7. Discussion and Future Work

In this work, we introduced and analyzed online and approximate PI methods, generalized to the κ-greedy policy, an instance of a multiple-step greedy policy. Doing so, we discovered two intriguing properties compared to the well-studied 1-step greedy policy, which we believe can be impactful in designing state-of-the-art algorithms. First, successive application of multiple-step greedy policies with a soft, stepsize-based update does not guarantee improvement; see Theorem 2. To mitigate this caveat, we designed an online PI with a 'cautious' improvement operator; see Section 5.

The second property we find intriguing stemmed from analyzing κ generalizations of known approximate hard-update PI methods. In Section 6, we revealed a performance tradeoff in κ, which can be interpreted as a tradeoff between short-horizon bootstrap bias and long-rollout variance. This corresponds to the known λ tradeoff in the famous TD(λ).

The two characteristics above lead to new compelling questions. The first regards improvement operators: would a non-monotonically improving PI scheme necessarily not converge to the optimal policy? Our attempts to generalize existing proof techniques to show convergence in such cases have fallen behind. Specifically, in the online case, Lemma 5.4 in Konda and Borkar (1999) does not hold with multiple-step greedy policies; similar issues arise when trying to form a κ-CPI algorithm via, e.g., an attempt to generalize Corollary 4.2 in Kakade and Langford (2002). Another research question regards the choice of the parameter κ given the tradeoff it poses. One possible direction for answering it could be investigating the concentrability coefficients further and attempting to characterize them for specific MDPs, either theoretically or via estimation. Lastly, a next indisputable step would be to empirically evaluate implementations of the algorithms presented here.

References

J Andrew Bagnell, Sham M Kakade, Jeff G Schneider, and Andrew Y Ng. Policy search by dynamic programming. In *Advances in neural information processing systems*, pages 831–838, 2004.

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

Bruno Bouzy and Bernard Helmstetter. Monte-carlo go developments. In *Advances in computer games*, pages 159–174. Springer, 2004.

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. Beyond the one step greedy approach in reinforcement learning. *arXiv preprint arXiv:1802.03654*, 2018.

Damien Ernst, Mevludin Glavic, Florin Capitanescu, and Louis Wehenkel. Reinforcement learning versus model predictive control: a comparison on a power system problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):517–529, 2009.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pages 568–576, 2010.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

S.M. Kakade and J. Langford. Approximately Optimal Approximate Reinforcement Learning. In *International Conference on Machine Learning*, pages 267–274, 2002.

Vijaymohan R Konda and Vivek S Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123, 1999.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Analysis of classification-based policy iteration algorithms. *The Journal of Machine Learning Research*, 17(1): 583–612, 2016.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pages 560–567. AAAI Press, 2003.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Rudy R Negenborn, Bart De Schutter, Marco A Wiering, and Hans Hellendoorn. Learning-based model predictive control for markov decision processes. *IFAC Proceedings Volumes*, 38(1):354–359, 2005.

Steven Perkins and David S Leslie. Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems*, 2(2):409–446, 2013.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

Bruno Scherrer. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pages 1314–1322, 2014.

Bruno Scherrer. Improved and Generalized Upper Bounds on the Complexity of Policy Iteration. INFORMS, February 2016. doi: 10.1287/moor.2015.0753. Markov decision processes ; Dynamic Programming ; Analysis of Algorithms.

Bruno Scherrer and Matthieu Geist. Local policy search in a convex space and conservative policy iteration as boosted policy search. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

Brian Sheppard. World-championship-caliber scrabble. *Artificial Intelligence*, 134(1-2): 241–275, 2002.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017a.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017b.

Aviv Tamar, Garrett Thomas, Tianhao Zhang, Sergey Levine, and Pieter Abbeel. Learning from the hindsight plan—episodic mpc improvement. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 336–343. IEEE, 2017.

Gerald Tesauro and Gregory R Galperin. On-line policy improvement using monte-carlo search. In *Advances in Neural Information Processing Systems*, pages 1068–1074, 1997.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Appendix A. Proof of Theorem 2

We start with a generalization of a useful lemma; its original version appeared in, e.g., (Scherrer, 2016, Lemma 10).

Lemma 10 *Let v be a value function, π a policy, and $\kappa \in [0,1]$. Then*

$$T^\pi_\kappa v - v = (I - \kappa\gamma P^\pi)^{-1}(T^\pi v - v).$$

Proof The proof is a straightforward generalization of the proof in (Scherrer, 2016, Lemma 10), and (Kakade and Langford, 2002, Remark 6.1).

$$\begin{aligned}
T^\pi_\kappa v - v &= (I - \kappa\gamma P^\pi)^{-1}(r^\pi + (1-\kappa)\gamma P^\pi v) - v \\
&= (I - \kappa\gamma P^\pi)^{-1}(r^\pi + (1-\kappa)\gamma P^\pi v - (I - \kappa\gamma P^\pi)v) \\
&= (I - \kappa\gamma P^\pi)^{-1}(r^\pi + \gamma P^\pi v - v) \\
&= (I - \kappa\gamma P^\pi)^{-1}(T^\pi v - v).
\end{aligned}$$

■

This elementary lemma relates the 'κ-advantage' to the 1-step advantage and is useful to prove Theorem 2 and some following results.

First, since $\pi(\alpha,\kappa) = (1-\alpha)\pi + \alpha\pi_\kappa$, we have that

$$\begin{aligned}
P^{\pi(\alpha,\kappa)} &= (1-\alpha)P^\pi + \alpha P^{\pi_\kappa}, \\
r^{\pi(\alpha,\kappa)} &= (1-\alpha)r^\pi + \alpha r^{\pi_\kappa};
\end{aligned}$$

thus, since v^π is the fixed-point of T^π,

$$T^{\pi(\alpha,\kappa)}v^\pi = (1-\alpha)T^\pi v^\pi + \alpha T^{\pi_\kappa} v^\pi = (1-\alpha)v^\pi + \alpha T^{\pi_\kappa} v^\pi. \tag{5}$$

Using this, we now prove the first statement of Theorem 2.

$$\begin{aligned}
v^{\pi(\alpha,\kappa)} - v^\pi &= (I - \gamma P^{\pi(\alpha,\kappa)})^{-1}(T^{\pi(\alpha,\kappa)}v^\pi - v^\pi) \\
&= \alpha(I - \gamma P^{\pi(\alpha,\kappa)})^{-1}(T^{\pi_\kappa}v^\pi - v^\pi) \\
&= \alpha(I - \gamma P^{\pi(\alpha,\kappa)})^{-1}(I - \kappa\gamma P^{\pi_\kappa})(I - \kappa\gamma P^{\pi_\kappa})^{-1}(T^{\pi_\kappa}v^\pi - v^\pi) \\
&= \alpha(I - \gamma P^{\pi(\alpha,\kappa)})^{-1}(I - \kappa\gamma P^{\pi_\kappa})(T^{\pi_\kappa}_\kappa v^\pi - v^\pi) \\
&= \alpha(I - \gamma P^{\pi(\alpha,\kappa)})^{-1}(I - \gamma P^{\pi(\alpha,\kappa)} + \gamma(P^{\pi(\alpha,\kappa)} - \kappa P^{\pi_\kappa}))(T^{\pi_\kappa}_\kappa v^\pi - v^\pi) \\
&= \alpha(I + \gamma(I - \gamma P^{\pi(\alpha,\kappa)})^{-1}((1-\alpha)P^\pi + (\alpha-\kappa)P^{\pi_\kappa})(T^{\pi_\kappa}_\kappa v^\pi - v^\pi).
\end{aligned} \tag{6}$$

For the first relation we use Lemma 10 with $\kappa = 1$ and the fact that, by definition, $T^{\pi(\alpha,\kappa)}_{\kappa=1} v^{\pi(\alpha,\kappa)} = v^{\pi(\alpha,\kappa)}$. For the second relation we use (5), for the fourth we again use Lemma 10, and for the last relation we use that $P^{\pi(\alpha,\kappa)} - \kappa P^{\pi_\kappa} = (1-\alpha)P^\pi + (\alpha-\kappa)P^{\pi_\kappa}$.

Next, we show that for $\alpha \geq \kappa$, all terms in (6) are component-wise bigger than or equal to zero. First, using a Taylor expansion, $(I - \gamma P^{\pi(\alpha,\kappa)})^{-1} = \sum_t \gamma^t (P^{\pi(\alpha,\kappa)})^t \geq 0$ component-wise, since it is a weighted sum of transition matrices with positive weights. The same applies for $(1-\alpha)P^\pi + (\alpha-\kappa)P^{\pi_\kappa}$, when $\alpha \geq \kappa$. Thus, for $\alpha \geq \kappa$,

$$(I + \gamma(I - \gamma P^{\pi(\alpha,\kappa)})^{-1}((1-\alpha)P^\pi + (\alpha-\kappa)P^{\pi_\kappa}) \geq 0$$

component-wise. Lastly, since $\pi_\kappa \in \mathcal{G}_\kappa(v^\pi)$, $v^\pi = T_\kappa^\pi v^\pi \leq T_\kappa v^\pi = T_\kappa^{\pi_\kappa} v^\pi$, with equality holding if and only if $v^\pi = v^*$ (Efroni et al., 2018, Lemma 3). Thus, $T_\kappa^{\pi_\kappa} v^\pi - v^\pi \geq 0$. This concludes the proof for the first statement, for the κ-greedy policy.

For the κ-greedy policy part of the proof for the second statement, we now provide more details on the counterexample presented in Section 4. For convenience, we bring the MDP example here again in Fig. 1. Consider the mixture of the "hesitant" and "confident" policies: $\pi(\alpha, \kappa = 1) = (1-\alpha)\pi_0 + \alpha\pi(\alpha, \kappa = 1)$. It can be shown that its value is

$$v^{\pi(\alpha,\kappa=1)}(s_0) = \frac{\gamma\alpha}{1-\gamma(1-\alpha)} v^{\pi(\alpha,\kappa=1)}(s_1),$$
$$v^{\pi(\alpha,\kappa=1)}(s_1) = \gamma\frac{-c(1-\alpha)+\alpha}{1-\gamma}.$$

Thus, we deduce that for any $\alpha \in (0,1)$ and

$$c > \frac{\alpha}{1-\alpha}, \tag{7}$$

$v^{\pi(\alpha,\kappa=1)}(s_0) < v^\pi(s_0) = 0$, i.e, the mixture policy, $\pi(\alpha, \kappa = 1)$, is not strictly better then π_0.

We now find the conditions to ensure that the κ-greedy policy w.r.t. v^{π_0} is the optimal policy; this will generalize the above construction, made for $\kappa = 1$, to any $\kappa \in [0,1]$. Observe that for any $c > 0$ and κ it holds that $\pi_\kappa(s_1) = a_1 = \pi^*(s_1)$, where $\pi_\kappa \in \mathcal{G}_\kappa(v^{\pi_0})$. Thus, we solely need to consider the policy which is different than π^* at state s_0, $\tilde{\pi}(s_0) = a_0 \neq \pi^*(s_0)$ and $\tilde{\pi}(s_1) = \pi^*(s_1)$. To find which condition ensures the κ-greedy policy w.r.t. v^{π_0} is π^* (and not $\tilde{\pi}$), we require

$$T_\kappa^{\pi^*} v^{\pi_0}(s_0) \geq T_\kappa^{\tilde{\pi}} v^{\pi_0}(s_0). \tag{8}$$

Satisfying this condition insures that $\pi^* \in \mathcal{G}_\kappa(v^{\pi_0})$. By definition,

$$\begin{aligned}
&T_\kappa^{\pi^*} v^{\pi_0}(s_0) = \mathbb{E}^{\pi^*}\left[\sum_t (\kappa\gamma)^t (r(s_t, \pi^*(s_t)) + \gamma(1-\kappa)v^{\pi_0}(s_{t+1}) \mid s_{t=0} = s_0\right] \\
=&(\kappa\gamma)^0 \left(\gamma(1-\kappa)v^{\pi_0}(s_1))\right) + (\kappa\gamma)^1 \left(\gamma(1-\kappa)v^{\pi_0}(s_2)\right) + \sum_{t=2}^{\infty}(\kappa\gamma)^t(1 + v^{\pi_0}(s_2)) \\
=&(\kappa\gamma)^0 \left(\gamma(1-\kappa)(-\frac{\gamma c}{1-\gamma})\right) + (\kappa\gamma)^1 \left(\gamma(1-\kappa)\frac{1}{1-\gamma}\right) + \sum_{t=2}^{\infty}(\kappa\gamma)^t(1 + \gamma(1-\kappa)\frac{1}{1-\gamma}) \\
=&\gamma(1-\kappa)(-\frac{\gamma c}{1-\gamma}) + \kappa\gamma\frac{\gamma}{1-\gamma}.
\end{aligned} \tag{9}$$

Similarly, and since $\tilde{\pi}(s_0) = a_0$, we have that

$$T_\kappa^{\tilde{\pi}} v^{\pi_0}(s_0) = 0 \tag{10}$$

Plugging (9) and (10) into (8), we get the condition

$$c \leq \frac{\kappa}{1-\kappa}. \tag{11}$$

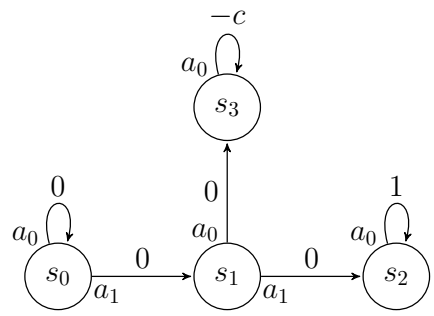

Figure 1: The Tightrope Walking MDP used in the proof of Theorem 2. This class of MDPs is parametrized by $c > 0$.

To finalize the counterexample and show that strict policy improvement is not guaranteed, we choose c such that both (7) and (11) are satisfied. Such feasible choice exists when $\alpha < \kappa$, due to the monotonicity of $\frac{x}{1-x}$.

The monotonic improvement of $\pi(\alpha, h)$ for $\alpha = 1$ was proved in (Efroni et al., 2018, Lemma 1). To build the counter example, again consider the Tightrope MDP. Let π_0 be the 'hesitant' policy. For any $\gamma \in (0, 1)$, $h > 1$, it holds that $\pi^* \in \mathcal{G}_h(v^{\pi_0})$. Thus, it suffices to satisfy (7) alone to show that $\pi(\alpha, h) = (1-\alpha)\pi_0 + \alpha\pi^*$ is not monotonically better then π. Large enough c value ensures that.

Appendix B. Proof of Lemma 4

We start by showing the contraction property of H^π_κ. Let (s, a) be a fixed state-action pair, $Q_1, Q_2 \in \mathbb{R}^{2|\mathcal{S}\times\mathcal{A}|}$. For any state-action pair (s, a), $Q_i(s, a)$ is a two-component vector. We denote its first component by $q_i(s, a)$ and its second component by $q_{i,\kappa}(s, a)$. See that

$$||q_1 - q_2||_\infty \leq ||Q_1 - Q_2||_\infty \tag{12}$$

$$||q_{1,\kappa} - q_{2,\kappa}||_\infty \leq ||Q_1 - Q_2||_\infty. \tag{13}$$

Taking a component-wise absolute value, we have that

$$\begin{aligned}
&|H^\pi_\kappa Q_1 - H^\pi_\kappa Q_2|(s, a) = \\
&|H^\pi_\kappa(q_1, q_{1,\kappa}) - H^\pi_\kappa(q_2, q_{2,\kappa})|(s, a) = \\
&\gamma \begin{bmatrix} |\mathbb{E}_{s',a^\pi}\left[q_1(s', a^\pi)) - q_2(s', \pi(s'))\right]| \\ |(1-\kappa)\mathbb{E}_{s',a^\pi}\left[q_1(s', a^\pi) - q_2(s', a^\pi))\right] + \kappa\mathbb{E}_{s'}[\max_{a'} q_{1,\kappa}(s', a') - \max_{a'} q_{2,\kappa}(s', a')]| \end{bmatrix},
\end{aligned}$$

where $s' \sim P(\cdot \mid s, a), a^\pi \sim \pi(s')$.

Let us focus on the first component of the above vector. We have that

$$\gamma|\mathbb{E}_{s',a^\pi}\left[q_1(s', a^\pi) - q_2(s', a^\pi)\right]| \leq \gamma||q_1 - q_2||_\infty \leq \gamma||Q_1 - Q_2||_\infty,$$

where we used the standard bound, $|\mathbb{E}[X]| \leq ||X||_\infty$ and (12). Similarly, for the second component, we have that

$$\begin{aligned}
&\gamma \left| \left((1-\kappa)\mathbb{E}_{s',a^\pi}\left[q_1(s',a^\pi) - q_2(s',a^\pi)\right] + \kappa \mathbb{E}_{s',a}[\max_{a'} q_{1,\kappa}(s',a') - \max_{a'} q_{2,\kappa}(s',a')] \right) \right| \\
\leq &\gamma \left((1-\kappa)|\mathbb{E}_{s',a^\pi}\left[q_1(s',a^\pi) - q_2(s',a^\pi)\right]| + \kappa \mathbb{E}_{s',a}[|\max_{a'} q_{1,\kappa}(s',a') - \max_{a'} q_{2,\kappa}(s',a')|] \right) \\
\leq &\gamma \left((1-\kappa)|\mathbb{E}_{s',a^\pi}\left[q_1(s',a^\pi) - q_2(s',a^\pi)\right]| + \kappa \mathbb{E}_{s',a'}[\max_{a'} |q_{1,\kappa}(s',a') - q_{2,\kappa}(s',a')|] \right) \\
\leq &\gamma \left((1-\kappa)||q_1 - q_2||_\infty + \kappa ||q_{1,\kappa} - q_{2,\kappa}||_\infty \right) \\
\leq &\gamma \left((1-\kappa)||Q_1 - Q_2||_\infty + \kappa ||Q_1 - Q_2||_\infty \right) = \gamma ||Q_1 - Q_2||_\infty,
\end{aligned}$$

where for the first relation we used the triangle inequality, for the second we used the standard bound $|\max_{x\in\mathcal{X}} f(x) - \max_{x\in\mathcal{X}} g(x)| \leq \max_{x\in\mathcal{X}} |f(x) - g(x)|$, for the third we used the bound $|\mathbb{E}[X]| \leq ||X||_\infty$, and for the last (12)-(13).

By taking the sup-norm on both sides, we get that

$$||H_\kappa^\pi Q_1 - H_\kappa^\pi Q_2||_\infty \leq \gamma ||Q_1 - Q_2||_\infty;$$

i.e., the operator H_κ^π is a γ contraction mapping in the max-norm.

We now show that the fixed-point of H_κ is (q^π, q_κ^π), i.e., $H_\kappa(q^\pi, q_\kappa^\pi) = (q^\pi, q_\kappa^\pi)$. For ease, we rewrite its form as in Definition 3. For any s, a we have that,

$$H_\kappa^\pi(q^\pi(s,a), q_\kappa^\pi(s,a)) \stackrel{\text{def}}{=} \begin{bmatrix} r(s,a) + \gamma \mathbb{E}_{s',a^\pi} q^\pi(s',a^\pi) \\ r(s,a) + \gamma(1-\kappa)\mathbb{E}_{s',a^\pi} q^\pi(s',a^\pi) + \kappa\gamma \mathbb{E}_{s'} \max_{a'} q_\kappa^\pi(s',a') \end{bmatrix}, \quad (14)$$

where $s' \sim P(\cdot \mid s,a), a^\pi \sim \pi(s')$.

It is clear that the the first component of (14) is $q^\pi(s)$, since it satisfies,

$$q_\kappa^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s',a^\pi} q^\pi(s',a^\pi).$$

Since q_κ^π is optimal q-function of the $\kappa\gamma$-discounted, with a shaped reward (see Remark 1)

$$r_{\text{eff}}(s,a,s') = r(s,a) + \gamma(1-\kappa)v^\pi(s'),$$

it is the solution of the following fixed-point equation,

$$\begin{aligned}
q_\kappa^\pi(s,a) &= \mathbb{E}_{s'}[r_{\text{eff}}(s,a,s') \mid s,a] + \gamma\kappa \mathbb{E}_{s'}[\max_{a'} q_\kappa^\pi(s',a') \mid s,a] \\
&= \mathbb{E}_{s'}[r(s,a) + \gamma(1-\kappa)v^\pi(s') \mid s,a] + \gamma\kappa \mathbb{E}_{s'}[\max_{a'} q_\kappa^\pi(s',a') \mid s,a] \\
&= \mathbb{E}_{s',a\sim\pi(s')}[r(s,a) + \gamma(1-\kappa)q^\pi(s',a') \mid s,a] + \gamma\kappa \mathbb{E}_{s'}[\max_{a'} q_\kappa^\pi(s',a') \mid s,a].
\end{aligned}$$

The final equation is the second-component of $H_\kappa(q^\pi, q_\kappa^\pi)(s,a)$ of (14), and indeed, q_κ^π solves the equation.

Appendix C. Proof of Theorem 5

The proof of Theorem 5 follows the proof in (Perkins and Leslie, 2013, Section 5.1), with several generalizations given below.

C.1 Lipschitzness of the Slow Time Scale Fixed-Point

Before following the main lemmas in Perkins and Leslie (2013) and showing they hold for Online κ-PI (Algorithm 1), we shall show that the solution of the fast-time scale ODE (found using a fixed-point argument), $[q^\pi, q^\pi_\kappa]$, is Lipschitz-continuous in the slow time-scale iterate, π.

Lemma 11 *Let $\pi : \mathcal{S}\times\mathcal{A} \to [0,1]$ be a stochastic policy. For any π_1, π_2 and $q_1, q_2 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, let*

$$||\pi_1 - \pi_2||_\infty \stackrel{def}{=} \max_s \sum_a |\pi_1(a \mid s) - \pi_2(a \mid s)|,$$
$$||q_1 - q_2||_\infty \stackrel{def}{=} \max_{s,a} |q_1(s,a) - q_2(s,a)|.$$

Then q^π and q^π_κ are Lipschitz-continuous in π in the max-norm; i.e.,

$$||q^{\pi_1} - q^{\pi_2}||_\infty \leq L_a||\pi_1 - \pi_2||_\infty,$$
$$||q^{\pi_1}_\kappa - q^{\pi_2}_\kappa||_\infty \leq L_b||\pi_1 - \pi_2||_\infty,$$

where $L_a, L_b > 0$, are functions of $\gamma, \kappa, R_{\max}$.

Proof We start by proving that $||v^{\pi_1} - v^{\pi_2}||_\infty \leq L||\pi_1 - \pi_2||_\infty$, i.e, v^π is Lipschitz in π.

$$\begin{aligned}
||v^{\pi_1} - v^{\pi_2}||_\infty &= ||T^{\pi_1}v^{\pi_1} - T^{\pi_2}v^{\pi_2}||_\infty \\
&\leq ||T^{\pi_1}v^{\pi_1} - T^{\pi_1}v^{\pi_2} + T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}||_\infty \\
&\leq ||T^{\pi_1}v^{\pi_1} - T^{\pi_1}v^{\pi_2}||_\infty + ||T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}||_\infty \\
&\leq \gamma||v^{\pi_1} - v^{\pi_2}||_\infty + ||T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}||_\infty, \qquad (15)
\end{aligned}$$

where the last relation is due to the fact T^{π_1} is a γ-contraction. We continue by calculating $|T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}|(s)$.

$$|T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}|(s) \leq |\sum_a (\pi_1(a \mid s) - \pi_2(a \mid s))r(s,a)| + \gamma|\sum_{s'}(P^{\pi_1}_{s',s} - P^{\pi_2}_{s',s})v^{\pi_2}(s')|. \qquad (16)$$

We bound each term in (16). The first term can be bounded by,

$$\begin{aligned}
|\sum_a (\pi_1(a \mid s) - \pi_2(a \mid s))r(s,a)| &\leq \sum_a |(\pi_1(a \mid s) - \pi_2(a \mid s))||r(s,a)| \\
&\leq R_{\max} \max_s \sum_a |(\pi_1(a \mid s) - \pi_2(a \mid s))| \\
&= R_{\max}||\pi_1 - \pi_2||_\infty. \qquad (17)
\end{aligned}$$

In the first relation we used the triangle inequality and in the second inequality the fact that $|r(s,a)|$ is bounded by $R_{\max}$.

The second term in (16) can be bounded by,

$$
\begin{aligned}
|\sum_{s'}(P^{\pi_1}_{s',s} - P^{\pi_2}_{s',s})v^{\pi_2}(s')| &= |\sum_{s',a} P(s' \mid s,a)(\pi_1(a \mid s) - \pi_2(a \mid s))v^{\pi_2}(s')| \\
&\leq \sum_a \sum_{s'} P(s' \mid s,a)|(\pi_1(a \mid s) - \pi_2(a \mid s))v^{\pi_2}(s')| \\
&\leq \sum_a \sum_{s'} P(s' \mid s,a)|(\pi_1(a \mid s) - \pi_2(a \mid s))||v^{\pi_2}(s')| \\
&\leq \sum_a \sum_{s'} P(s' \mid s,a)|(\pi_1(a \mid s) - \pi_2(a \mid s))|\frac{R_{\max}}{1-\gamma} \\
&= \sum_a |(\pi_1(a \mid s) - \pi_2(a \mid s))|\frac{R_{\max}}{1-\gamma}\sum_{s'} P(s' \mid s,a) \\
&= \sum_a |(\pi_1(a \mid s) - \pi_2(a \mid s))|\frac{R_{\max}}{1-\gamma} \\
&\leq \max_s \sum_a |(\pi_1(a \mid s) - \pi_2(a \mid s))|\frac{R_{\max}}{1-\gamma} = \frac{R_{\max}}{1-\gamma}||\pi_1 - \pi_2||_\infty
\end{aligned} \tag{18}
$$

In the first relation we used the triangle inequality, in the forth relation we used the fact that for any π and s, $v^\pi(s) \leq \frac{R_{\max}}{1-\gamma}$, and in the fifth relation the fact that for any s and a, $P(s' \mid s,a)$ is a probability function, thus sums to one.

Using (17), (18) to bound (16) yields that for any s,

$$
|T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}|(s) \leq \frac{R_{\max}}{1-\gamma}||\pi_1 - \pi_2||_\infty.
$$

Thus, $||T^{\pi_1}v^{\pi_2} - T^{\pi_2}v^{\pi_2}||_\infty \leq \frac{R_{\max}}{1-\gamma}||\pi_1 - \pi_2||_\infty$. Plugging this bound into (15) and rearranging yields,

$$
||v^{\pi_1} - v^{\pi_2}||_\infty \leq \frac{R_{\max}}{(1-\gamma)^2}||\pi_1 - \pi_2||_\infty, \tag{19}
$$

giving that $L = \frac{R_{\max}}{(1-\gamma)^2}$.

We continue by analysing $||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty$. We remind the reader that $T_\kappa v^\pi$ satisfies the following fixed-point equation:

$$
\begin{aligned}
T_\kappa v^\pi(s) &= \max_a \left[r(s,a) + \gamma(1-\kappa)\sum_{s'} P(s' \mid s,a)v^\pi(s') + \kappa\gamma \sum_{s'} P(s' \mid s,a)(T_\kappa v^\pi)(s') \right] \\
&\stackrel{\text{def}}{=} \bar{T}^\pi_\kappa T_\kappa v^\pi(s),
\end{aligned}
$$

where we defined the 'optimal' Bellman operator of the surrogate MDP to be $\bar{T}^\pi_\kappa$ (see Remark 1). Furthermore, since this operator is the optimal Bellman operator of a $\kappa\gamma$-discounted MDP, it is a $\kappa\gamma$ contraction mapping. We now use a similar technique as the above to show $||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty \leq L_\kappa||\pi_1 - \pi_2||_\infty$, i.e, $T_\kappa v^\pi$ is Lipschitz in π.

$$
\begin{aligned}
||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty &= ||\bar{T}^{\pi_1}_\kappa T_\kappa v^{\pi_1} - \bar{T}^{\pi_2}_\kappa T_\kappa v^{\pi_2}||_\infty \\
&\leq ||\bar{T}^{\pi_1}_\kappa T_\kappa v^{\pi_1} - \bar{T}^{\pi_1}_\kappa T_\kappa v^{\pi_2}||_\infty + ||\bar{T}^{\pi_1}_\kappa T_\kappa v^{\pi_2} - \bar{T}^{\pi_2}_\kappa T_\kappa v^{\pi_2}||_\infty \\
&\leq \kappa\gamma||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty + ||\bar{T}^{\pi_1}_\kappa T_\kappa v^{\pi_2} - \bar{T}^{\pi_2}_\kappa T_\kappa v^{\pi_2}||_\infty.
\end{aligned}
$$

We now bound the second term.

$$
\begin{aligned}
|\bar{T}_\kappa^{\pi_1} T_\kappa v^{\pi_2} - \bar{T}_\kappa^{\pi_2} T_\kappa v^{\pi_2}|(s) &\leq \max_a \gamma(1-\kappa)|\sum_{s'} P(s' \mid s,a)(v^{\pi_1} - v^{\pi_2})(s')| \\
&\leq \max_a \gamma(1-\kappa) \sum_{s'} P(s' \mid s,a)|v^{\pi_1} - v^{\pi_2}|(s') \\
&\leq \max_a \gamma(1-\kappa) \sum_{s'} P(s' \mid s,a)||v^{\pi_1} - v^{\pi_2}||_\infty = \gamma(1-\kappa)||v^{\pi_1} - v^{\pi_2}||_\infty,
\end{aligned}
$$

where we used the definition of $\bar{T}_\kappa^\pi$ and the identity $|\max_{x\in\mathcal{X}} f(x) - \max_{x\in\mathcal{X}} g(x)| \leq \max_{x\in\mathcal{X}} |f(x) - g(x)|$ in the first relation and the triangle inequality in the second.

Using (19), we have

$$
\begin{aligned}
||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty &\leq \frac{\gamma(1-\kappa)}{1-\kappa\gamma}||v^{\pi_1} - v^{\pi_2}||_\infty \\
&\leq \frac{\gamma(1-\kappa)}{1-\kappa\gamma}\frac{R_{\max}}{(1-\gamma)^2}||\pi_1 - \pi_2||_\infty.
\end{aligned}
$$

These results transform to results on q^π and q_κ^π as follows. Starting with q^π,

$$
\begin{aligned}
|q^{\pi_1} - q^{\pi_2}|(s,a) &= |r(s,a) + \gamma \sum_{s'} P(s' \mid s,a)v^{\pi_1} - r(s,a) - \gamma \sum_{s'} P(s' \mid s,a)v^{\pi_2}| \\
&= \gamma|\sum_{s'} P(s' \mid s,a)(v^{\pi_1} - v^{\pi_2})| \leq \gamma||v^{\pi_1} - v^{\pi_2}||_\infty.
\end{aligned}
$$

By taking the max-norm on both sides we get the result since $||v^{\pi_1} - v^{\pi_2}||_\infty$ was shown to be Lipschitz in π.

Next, for q_κ^π we have

$$
\begin{aligned}
&|q_\kappa^{\pi_1} - q_\kappa^{\pi_2}|(s,a) \\
=&|\gamma(1-\kappa)\sum_{s'} P(s' \mid s,a)(v^{\pi_1}(s') - v^{\pi_2}(s')) + \kappa\gamma \sum_{s'} P(s' \mid s,a)(T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2})(s')| \\
\leq&\gamma(1-\kappa)||v^{\pi_1}(s') - v^{\pi_2}(s')||_\infty + \kappa\gamma||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty.
\end{aligned}
$$

By taking the max-norm on both sides we get the result since, as shown above, both $||v^{\pi_1} - v^{\pi_2}||_\infty$ and $||T_\kappa v^{\pi_1} - T_\kappa v^{\pi_2}||_\infty$ are Lipschitz in π. Finally, since the vector space is finite (due to the finite state and action space), all L_p norms are equivalent. Thus, the Lipschitzness result applies in any L_p norm as well. ■

C.2 Improvement Step

Here, we prove an equivalent lemma to (Perkins and Leslie, 2013, Lemma 5.4) which shows that the mean value of the process improves. Denote $b_s \equiv b_s(q^\pi, q_\kappa^\pi, \pi)$ as the policy defined in the Algrorithm 1. By using Lemma 10 and setting $\kappa = 0$ we have that

$$
v^{(1-\alpha)\pi+\alpha b_s} - v^\pi = \alpha(I - \gamma P^{(1-\alpha)\pi+\alpha b_s})^{-1}(T^{b_s} v^\pi - v^\pi).
$$

Thus, by taking the limit $\alpha \to 0$ we have

$$\begin{aligned}\lim_{\alpha\to 0}(v^{(1-\alpha)\pi+\alpha b_s} - v^\pi) &= \alpha \nabla_\pi v^\pi (b_s - \pi) \\ &= \alpha \ \langle \nabla_\pi v^\pi, \Delta\pi \rangle \\ &= \alpha (I - \gamma P^\pi)^{-1}(T^{b_s} v^\pi - v^\pi) + \mathcal{O}(\alpha^2) \geq 0,\end{aligned}$$

where the last inequality is since $T^{b_s} v^\pi - v^\pi \geq 0$ by construction and $(I - \gamma P^\pi)^{-1} \geq 0$ component-wise. We thus get that

$$\frac{1}{\alpha} \lim_{\alpha\to 0}(v^{(1-\alpha)\pi+\alpha b_s} - v^\pi) = \langle \nabla_\pi v^\pi, \Delta\pi \rangle \geq 0.$$

C.3 Convergence of the Algorithm

We define the same Lyapunov function as defined in (Perkins and Leslie, 2013, Lemma 5.5). Due to previous section it is indeed a Lyapunov function since its derivative is negative and the function is bigger than 0 by construction. The presence of the Lyapunov function leads to the convergence of the policy to the optimal policy, similarly to (Perkins and Leslie, 2013, Corollary 5.6), which leads to the convergence of q^π to q^*. Lastly, since $T_\kappa v^* = v^*$ (Efroni et al., 2018, Lemma 4) we have that,

$$\begin{aligned}q_\kappa^{\pi^*}(\pi') &= r^{\pi'} + \gamma(1-\kappa)P^{\pi'} v^* + \kappa\gamma P^{\pi'} T_\kappa v^* \\ &= r^{\pi'} + \gamma(1-\kappa)P^{\pi'} v^* + \kappa\gamma P^{\pi'} v^* \\ &= r^{\pi'} + \gamma P^{\pi'} v^* = q^*(\pi').\end{aligned}$$

which concludes the proof.

Appendix D. Proof of Lemma 8

We first prove a useful lemma that relates the (unnormalized) future distribution, measured in different κ scales.

Lemma 12 *For any policy π and $\kappa, \kappa' \in [0,1]$,*

$$(I - \xi_{\kappa'} D_{\kappa'}^\pi P^\pi)^{-1} = \frac{\kappa' - \kappa}{1-\kappa} I + \frac{1-\kappa'}{1-\kappa}(I - \xi_\kappa D_\kappa^\pi P^\pi)^{-1}.$$

Proof We prove the lemma by using the definition and by some algebraic manipulations.

$$\begin{aligned}(I - \xi_{\kappa'} D_{\kappa'}^\pi P^\pi)^{-1} &= (I - \gamma(1-\kappa')(I - \kappa\gamma' P^\pi)^{-1} P^\pi)^{-1} \\ &= ((I - \kappa\gamma' P^\pi)^{-1}(I - \kappa\gamma' P^\pi - \gamma(1-\kappa')P^\pi))^{-1} \\ &= (I - \gamma P^\pi)^{-1}(I - \gamma\kappa' P^\pi) \\ &= (I - \gamma P^\pi)^{-1} - \kappa'\gamma P^\pi (I - \gamma P^\pi)^{-1} \\ &= (I - \gamma P^\pi)^{-1} - \kappa'(I + \gamma P^\pi (I - \gamma P^\pi)^{-1} - I) \\ &= (I - \gamma P^\pi)^{-1} - \kappa'((I - \gamma P^\pi)^{-1} - I) \\ &= \kappa' I + (1-\kappa')(I - \gamma P^\pi)^{-1}\end{aligned}$$

We see that the following relation holds for any $\kappa \in [0,1]$,

$$(I - \gamma P^\pi)^{-1} = \frac{1}{1-\kappa}((I - \xi_\kappa D_\kappa^\pi P^\pi)^{-1} - \kappa I).$$

Plugging this relation into the previous one we get,

$$\begin{aligned}(I - \xi_{\kappa'} D_{\kappa'}^\pi P^\pi)^{-1} &= \kappa' I + (1-\kappa')(I - \gamma P^\pi)^{-1} \\ &= \kappa' I + \frac{1-\kappa'}{1-\kappa}((I - \xi_\kappa D_\kappa^\pi P^\pi)^{-1} - \kappa I) \\ &= \frac{\kappa' - \kappa}{1-\kappa} I + \frac{1-\kappa'}{1-\kappa}(I - \xi_\kappa D_\kappa^\pi P^\pi)^{-1}.\end{aligned}$$

■

We are now ready to prove Lemma 8. Assume a constant $C_\kappa^{\pi^*}(\mu,\nu) < \infty$ such that,

$$d_{\kappa,\mu}^{\pi^*} = (1-\xi)\mu(I - \xi D_\kappa^{\pi^*})^{-1} < C_\kappa^{\pi^*}(\mu,\nu)\nu. \tag{20}$$

Given that, we shall calculate $C_{\kappa'}^{\pi^*}(\mu,\nu)$ where $\kappa' > \kappa$.

$$\begin{aligned}d_{\kappa',\mu}^{\pi^*} &= (1-\xi_{\kappa'})\mu(I - \xi D_{\kappa'}^{\pi^*})^{-1} \\ &= (1-\xi_{\kappa'})\left(\frac{\kappa'-\kappa}{1-\kappa}\mu + \frac{1-\kappa'}{1-\kappa}\mu((I - \xi_\kappa D_\kappa^\pi P^\pi)^{-1})\right) \\ &\leq \frac{1-\xi_{\kappa'}}{1-\kappa}\left((\kappa'-\kappa)\mu + \frac{1-\kappa'}{1-\xi_\kappa} C_\kappa^{\pi^*}(\mu,\nu)\nu\right) \\ &= \frac{1-\xi_{\kappa'}}{1-\kappa}(\kappa'-\kappa + \frac{1-\kappa'}{1-\xi_\kappa} C_\kappa^{\pi^*}(\mu,\nu))\,(\alpha^*\mu + (1-\alpha^*)\nu) \stackrel{\text{def}}{=} C_{\kappa'}^{\pi^*}(\mu,\nu(\alpha))\nu(\alpha),\end{aligned}$$

where we used Lemma 12 in the first line, Equation 20 in the second line, and defined $\alpha^* = (1 + \frac{1-\kappa'}{(1-\xi_\kappa)(\kappa'-\kappa)} C_\kappa^{\pi^*}(\mu,\nu)))^{-1} \in (0,1)$ and $C_{\kappa'}^{\pi^*}(\mu,\nu(\alpha^*)) = \frac{1-\xi_{\kappa'}}{1-\kappa}(\kappa' - \kappa + \frac{1-\kappa'}{1-\xi_\kappa} C_\kappa^{\pi^*}(\mu,\nu))$. By plugging the expressions of $\xi_\kappa, \xi_{\kappa'}$ we see that,

$$\begin{aligned}C_{\kappa'}^{\pi^*}(\mu,\nu(\alpha^*)) - C_\kappa^{\pi^*}(\mu,\nu) &= \frac{1-\xi_{\kappa'}}{1-\kappa}(\kappa'-\kappa + (\frac{1-\kappa'}{1-\xi_\kappa} - \frac{1-\kappa}{1-\xi_{\kappa'}})C_\kappa^{\pi^*}(\mu,\nu)) \\ &= \frac{1-\xi_{\kappa'}}{1-\kappa}(\kappa'-\kappa)(1 - C_\kappa^{\pi^*}(\mu,\nu)).\end{aligned} \tag{21}$$

Since $C_\kappa^{\pi^*}(\mu,\nu) \geq 1$ and $\frac{1-\xi_{\kappa'}}{1-\kappa}(\kappa'-\kappa) > 0$ we get that $C_{\kappa'}^{\pi^*}(\mu,\nu(\alpha^*)) - C_\kappa^{\pi^*}(\mu,\nu) \leq 0$, where the inequality is strict for $C_\kappa^{\pi^*}(\mu,\nu) > 1$. Finally, since for $\mu = \nu$ it holds that $\nu(\alpha^*) = (1-\alpha^*)\nu + \alpha^*\nu = \nu$ for, we get that $C_\kappa^{\pi^*}(\nu,\nu)$ is a decreasing function of κ.

Appendix E. Proof of Theorem 9

We first prove two technical lemmas.

Lemma 13 *Let π be a policy, $\kappa \in [0,1]$, $\gamma \in (0,1)$ and $i \in \mathbb{N}\backslash\{0\}$. Then*

$$(\xi D_\kappa^\pi P^\pi)^i = \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} \gamma^{t+1}(1-\kappa)^i \kappa^{t-(i-1)} (P^\pi)^{t+1},$$

where, as also given in Definition 7, $D_\kappa^\pi = (1-\kappa\gamma)(I-\kappa\gamma P^\pi)^{-1}$.

Proof First, for any $x \in \mathbb{R}$ s.t $|x| < 1$ and $i \in \mathbb{N}\backslash\{0\}$ we have that,

$$(1-x)^{-i} = \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} x^{t-(i-1)}.$$

Since it holds that $||\gamma\kappa P^\pi|| = \gamma\kappa < 1$, where $||\cdot||$ is the spectral norm of the matrix, we can use the same Taylor expansion when replacing x with $\gamma\kappa P^\pi$. Thus,

$$(I-\gamma\kappa P^\pi)^{-i} = \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} (\gamma\kappa)^{t-(i-1)} (P^\pi)^{t-(i-1)}. \tag{22}$$

Since $D_\kappa^\pi = (1-\kappa\gamma)(I-\kappa\gamma P^\pi)^{-1}$ and any matrix commutes with any function of itself we have that,

$$(\xi D_\kappa^\pi P^\pi)^i = \gamma^i (1-\kappa)^i (D_\kappa^\pi P^\pi)^i = \gamma^i (1-\kappa)^i ((I-\kappa\gamma P^\pi)^{-1})^i (P^\pi)^i.$$

By using (22) and packing the terms we conclude the proof.

$$\begin{aligned}(\xi D_\kappa^\pi P^\pi)^i &= \gamma^i (1-\kappa)^i (I-\kappa\gamma P^\pi)^{-i} (P^\pi)^i \\ &= \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} \gamma^{t+1}(1-\kappa)^i \kappa^{t-(i-1)} (P^\pi)^{t+1}\end{aligned}$$

■

Lemma 14 *Let $\kappa \in [0,1]$, $\gamma \in (0,1), n \in \mathbb{N} \cup \{\infty\}$ and $f : \mathbb{N} \to \mathbb{R}$. Then*

$$\begin{aligned}&\sum_{l=0}^{\infty} \sum_{i=1}^{n-1} \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} \gamma^{t+l+1} \kappa^{t-(i-1)} (1-\kappa)^i f(t+1+l) \\ \leq &(1-\kappa) \sum_{l=0}^{\infty} \sum_{t=0}^{n-2} \gamma^{t+l+1} f(t+1+l) + g(\kappa)(1-\kappa)\kappa \sum_{l=0}^{\infty} \sum_{t=n-1}^{\infty} \gamma^{t+l+1} f(t+1+l),\end{aligned}$$

where $g(\kappa)$ is a bounded function of κ. When $n \to \infty$ the second term vanishes.

Proof We start by exchanging the summation indices i and t. In order to do so, we decouple the summation to two sums. The range of the indices of the first sum is $t \in \{0, .., n-2\}$

and $i \in \{1, .., t+1\}$ and the range of the indices of the second sum is $t \in \{n-1, .., \infty\}$ and $i \in \{1, .., n-1\}$

$$
\begin{aligned}
&\sum_{l=0}^{\infty}\sum_{i=1}^{n-1}\sum_{t=i-1}^{\infty}\frac{t!}{(i-1)!(t-(i-1))!}\gamma^{t+l+1}\kappa^{t-(i-1)}(1-\kappa)^i f(t+1+l) \\
=&\sum_{l=0}^{\infty}\sum_{t=0}^{n-2}\gamma^{t+l+1}f(t+1+l)\sum_{i=1}^{t+1}\frac{t!}{(i-1)!(t-(i-1))!}\kappa^{t-(i-1)}(1-\kappa)^i \qquad (23)\\
&+\sum_{l=0}^{\infty}\sum_{t=n-1}^{\infty}\gamma^{t+l+1}f(t+1+l)\sum_{i=1}^{n-1}\frac{t!}{(i-1)!(t-(i-1))!}\kappa^{t-(i-1)}(1-\kappa)^i. \qquad (24)
\end{aligned}
$$

Let us bound the first sum first (23),

$$
\begin{aligned}
&\sum_{l=0}^{\infty}\sum_{t=0}^{n-2}\gamma^{t+l+1}f(t+1+l)\sum_{i=1}^{t+1}\frac{t!}{(i-1)!(t-(i-1))!}\kappa^{t-(i-1)}(1-\kappa)^i \\
=&\sum_{l=0}^{\infty}\sum_{t=0}^{n-2}\gamma^{t+l+1}f(t+1+l)\sum_{i=0}^{t}\frac{t!}{i!(t-i)!}\kappa^{t-i}(1-\kappa)^{i+1} \\
=&(1-\kappa)\sum_{l=0}^{\infty}\sum_{t=0}^{n-2}\gamma^{t+l+1}f(t+1+l),
\end{aligned}
$$

where in the first line we changed the index summation $i \leftarrow i-1$ and in the second line we used the binomial identity $\sum_{i=0}^{t}\frac{t!}{i!(t-i)!}\kappa^{t-i}(1-\kappa)^i = (1-\kappa+\kappa)^t = 1$.

In order to bound the second term (24) we define the following function, $\tilde{g} : [n-1, \infty) \to \mathbb{R}$,

$$
\tilde{g}(t) \stackrel{\text{def}}{=} \sum_{i=0}^{n-2}\frac{t!}{i!(t-i)!}\kappa^{t-i}(1-\kappa)^i.
$$

The function $\tilde{g}(t)$ is a sum of polynomial terms multiplied by a geometric decaying term, κ^t. Thus, this function is bounded from above, i.e, exists $t^* \in [n-1, \infty)$ such that $\tilde{g}(t) \leq \tilde{g}(t^*),\ \forall t \in [n-1, \infty)$. For such t^*, by construction, we have that

$$
\begin{aligned}
\sum_{i=1}^{n-1}\frac{t!}{(i-1)!(t-(i-1))!}\kappa^{t-(i-1)}(1-\kappa)^i &= (1-\kappa)\sum_{i=0}^{n-2}\frac{t!}{i!(t-i))!}\kappa^{t-i}(1-\kappa)^i \\
&\leq (1-\kappa)\sum_{i=0}^{n-2}\frac{t^*!}{i!(t^*-i)!}\kappa^{t^*-i}(1-\kappa)^i \\
&= (1-\kappa)\kappa^{t^*-(n-2)}\sum_{i=0}^{n-2}\frac{t^*!}{i!(t^*-i)!}\kappa^{(n-2)-i}(1-\kappa)^i \\
&\leq (1-\kappa)\kappa\sum_{i=0}^{n-2}\frac{t^*!}{i!(t^*-i)!}\kappa^{(n-2)-i}(1-\kappa)^i
\end{aligned}
$$

where the last line holds since for $\kappa \in [0,1],\ t^* \in [n-1,\infty)$ it holds that $\kappa^{t^*-(n-2)} \leq \kappa$. We now define $g(\kappa) \stackrel{\text{def}}{=} \sum_{i=0}^{n-2} \frac{t^*!}{i!(t^*-i)!}\kappa^{(n-2)-i}(1-\kappa)^i$, and observe that it is a bounded function of $\kappa \in [0,1]$, since it is a sum of positive powers of κ. Thus, (24) is bounded by

$$\begin{aligned}&\sum_{l=0}^{\infty}\sum_{t=n-1}^{\infty}\gamma^{t+l+1}f(t+1+l)\sum_{i=1}^{n-1}\frac{t!}{(i-1)!(t-(i-1))!}\kappa^{t-(i-1)}(1-\kappa)^i\\ \leq &g(\kappa)(1-\kappa)\kappa\sum_{l=0}^{\infty}\sum_{t=n-1}^{\infty}\gamma^{t+l+1}f(t+1+l)\end{aligned}$$

Finally, for the case $n=\infty$ observe we can repeat the same analysis we did for the first term (23) without the need to decouple to two sums. Thus, for this case, the bound on the first term, with $n=\infty$, bounds the expression.

■

Lemma 15 *Let $\kappa \in [0,1]$. For any sequence of policies $\{\pi_{k-i}\}_{i=0}^{k-1}$, optimal policy π^*, and error vectors which satisfy $\nu\bar{\delta}_i \leq \delta$, ,*

$$\sum_{i=0}^{k-1}\mu(\xi D_\kappa^{\pi^*}P^{\pi^*})^i\bar{\delta}_i \leq \frac{1-\kappa\gamma}{1-\gamma}C_\kappa^{\pi^*(1)}(\mu,\nu)\delta \tag{25}$$

and

$$\sum_{i=0}^{k-1}\mu(\xi D_\kappa^{\pi^*}P^{\pi^*})^i\bar{\delta}_i \leq k\frac{1-\kappa\gamma}{1-\gamma}C_\kappa^{\pi^*}(\mu,\nu)\delta. \tag{26}$$

Proof We begin proving the first statement. For $i > k-1$, we define vectors $\bar{\delta}_i$ s.t. $\nu\bar{\delta}_i \leq \delta$. Thus,

$$\sum_{i=0}^{k-1}\mu(\xi D_\kappa^{\pi^*}P^{\pi^*})^i\bar{\delta}_i \leq \mu\bar{\delta}_0 + \sum_{i=1}^{\infty}\mu(\xi D_\kappa^{\pi^*}P^{\pi^*})^i\bar{\delta}_i. \tag{27}$$

For the first term in (27),

$$\mu\bar{\delta}_0 \leq c(0)\nu\bar{\delta}_0 \leq c(0)\delta, \tag{28}$$

where we used Definition 7 and then Definition 6.

For the second term in (27), we have

$$
\begin{aligned}
&\sum_{i=1}^{\infty} \mu(\xi D_\kappa^{\pi^*} P^{\pi^*})^i \bar{\delta}_i \\
=&\sum_{i=1}^{\infty} \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} \gamma^{t+1}(1-\kappa)^i \kappa^{t-(i-1)} \mu (P^{\pi^*})^{t+1} \bar{\delta}_i \\
\leq&\sum_{i=1}^{\infty} \sum_{t=i-1}^{\infty} \frac{t!}{(i-1)!(t-(i-1))!} \gamma^{t+1}(1-\kappa)^i \kappa^{t-(i-1)} c^{\pi^*}(t+1)\delta \\
\leq&(1-\kappa)\sum_{t=0}^{\infty} \gamma^{t+1} c^{\pi*}(t+1)\delta \\
=&(1-\kappa)\sum_{t=0}^{\infty} \gamma^{t} c^{\pi*}(t)\delta - (1-\kappa)c(0)\delta = \frac{(1-\kappa)C^{\pi^*(1)}(\mu,\nu)}{1-\gamma}\delta - (1-\kappa)c(0)\delta. \qquad (29)
\end{aligned}
$$

For the first relation we apply Lemma 13, for the second we use the definition of $\{c^{\pi^*}(i)\}_{i=0}^{\infty}$ and use $\nu\bar{\delta}_i \leq \delta$. For the third relation we apply Lemma 14 with $n = \infty$, $f(\cdot) = c^{\pi^*}(\cdot)$ and drop the l summation.

Summing the terms in (28) and (29), we get

$$
\sum_{i=0}^{k-1} \mu(\xi D_\kappa^{\pi^*} P^{\pi^*})^i \bar{\delta}_i \leq \frac{1}{1-\gamma}\left((1-\kappa)C^{\pi^*(1)}(\mu,\nu) + (1-\gamma)\kappa c(0)\right)\delta = \frac{1-\kappa\gamma}{1-\gamma} C_\kappa^{\pi^*(1)}(\mu,\nu)\delta,
$$

where we identify $C_\kappa^{\pi^*(1)}(\mu,\nu)$ to be the same expression as in Definition 7.

For the second statement of the lemma, (26), we continue by using the identity

$$
(\xi D_\kappa^{\pi^*} P^{\pi^*})^i \leq (I - \xi D_\kappa^{\pi^*} P^{\pi^*})^{-1}.
$$

$$
\begin{aligned}
\sum_{i=0}^{k-1} \mu(\xi D_\kappa^{\pi^*} P^{\pi^*})^i \bar{\delta}_i &\leq \sum_{i=0}^{k-1} \mu(I - \xi D_\kappa^{\pi^*} P^{\pi^*})^{-1} \bar{\delta}_i \\
&\leq \sum_{i=0}^{k-1} \frac{C_\kappa^{\pi^*}(\mu,\nu)}{1-\xi}\nu\bar{\delta}_i \leq k\frac{C_\kappa^{\pi^*}(\mu,\nu)}{1-\xi}\delta = k\frac{1-\kappa\gamma}{1-\gamma} C_\kappa^{\pi^*}(\mu,\nu)\delta,
\end{aligned}
$$

where the second relation holds due to the definition of $C_\kappa^{\pi^*}(\mu,\nu)$. ■

To prove Theorem 9, we follow the arguments of (Scherrer, 2014, Appendix A), while using the operators T_κ^π instead of T^π and the approximate operator defined in Definition 6, and then use Lemma 15. We define the component-wise error at the i-th iteration, $\bar{\delta}_i$, which satisfies $\nu\bar{\delta}_i \leq \delta$. We have that for all k,

$$
\begin{aligned}
v^* - v^{\sigma_{\kappa,k}} &= T_\kappa^{\pi^*} v^* - T_\kappa^{\pi^*} v^{\sigma_{k-1}} + T_\kappa^{\pi^*} v^{\sigma_{k-1}} - T_\kappa^{\pi_k} v^{\sigma_{k-1}} \\
&\leq \xi D_\kappa^{\pi^*} P^{\pi^*}(v^* - v^{\sigma_{k-1}}) + \bar{\delta}_k.
\end{aligned}
$$

Thus, by induction on k, we obtain:

$$\begin{aligned} v^* - v^{\sigma_{\kappa,k}} &\leq \sum_{i=0}^{k-1} (\xi D_\kappa^{\pi^*} P^{\pi^*})^i \bar{\delta}_i + (\xi D_\kappa^{\pi^*} P^{\pi^*})^k (v^* - v^{\pi_0}) \\ &\leq \sum_{i=0}^{k-1} (\xi D_\kappa^{\pi^*} P^{\pi^*})^i \bar{\delta}_i + \xi^k \frac{R_{\max}}{1-\gamma} \end{aligned}$$

We can directly bound this term by applying Lemma 15. The two statements in that lemma lead to the two statements in Theorem 9. For the second statement, we carefully choose the iteration number k to make the last term smaller than δ:

$$k = \left\lceil \frac{\log \frac{R_{max}}{\delta(1-\gamma)}}{1-\xi} \right\rceil = \left\lceil \frac{(1-\kappa\gamma)\log \frac{R_{max}}{\delta(1-\gamma)}}{1-\gamma} \right\rceil . \tag{30}$$

By doing so we see that $\xi^{k^*} \frac{R_{\max}}{1-\gamma} < \delta$ and obtain the second statement of the result.