# Mirror descent actor-critic methods for KL-regularised MDPs in general spaces: stability and convergence

**Denis Zorba**
University of Edinburgh
e.zorba@sms.ed.ac.uk

**David Šiška**
University of Edinburgh
d.siska@ed.ac.uk

**Lukasz Szpruch**
University of Edinburgh
l.szpruch@ed.ac.uk

## Abstract

We provide theoretical guarantees for the convergence of discrete-time policy mirror descent with inexact advantage functions updated using temporal difference (TD) learning for KL-regularised MDPs in Polish state and action spaces. We rigorously derive sufficient conditions under which the single-loop actor-critic scheme is stable and convergent. To weaken these conditions, we introduce a variant that performs multiple TD steps per policy update and derive an explicit lower bound on the number of TD steps required to ensure stability. Finally, we establish sub-linear convergence when the number of TD steps grows logarithmically with the number of policy updates, and linear convergence when it grows linearly under a concentrability assumption.

## 1 Introduction

In reinforcement learning (RL), an agent seeks to find an optimal policy that minimises their expected cumulative cost by interacting with their environment. Such a framework has had numerous remarkable applications (Silver et al., 2018; Makoviychuk et al., 2021; Mnih et al., 2015). Theoretical understanding of key RL algorithms and their convergence properties has been growing over the last decades (Jaakkola et al., 1993; Tsitsiklis, 1994; Mei et al., 2020; Agarwal et al., 2021).

In practice, actor-critic algorithms are among the most widely used. The actor updates the policy using a policy gradient method (which requires the advantage function), while the critic, which provides an estimate of the advantage function, is updated based on costs and new states observed by interacting directly with the environment using, e.g., a temporal-difference (TD) loss.

Classical policy gradient methods require careful tuning of step sizes to achieve stability, which led to the development of trust-region approaches such as TRPO (Schulman et al., 2015) and its practical variant PPO (Schulman et al., 2017). These methods improve stability by controlling the change between successive policies via a KL-divergence constraint or policy-ratio clipping, respectively. Policy mirror descent (Ju and Lan, 2026; Lan, 2023; Kerimkulov et al., 2025) provides a closely related first-order viewpoint, replacing the hard trust-region constraint with a KL-based penalty and yielding algorithms for which it is possible to obtain theoretical guarantees (under the assumption of access to the exact advantage function). Adding entropic regularisation to the objective then leads to a number of theoretical and practical advantages (Ahmed et al., 2019; Cayci et al., 2024a; Leahy et al., 2022). Firstly, KL-regularised MDPs are guaranteed to have a unique optimal policy and can accelerate the convergence of policy gradient methods (Mei et al., 2020; Ju and Lan, 2026; Lan, 2023; Kerimkulov et al., 2025). Secondly, the entropic regularisation ensures persistent exploration (Rawlik et al., 2012; Haarnoja et al., 2017) and provides some (albeit minimal) robustness to changes in costs and environment (Ziebart, 2010; Eysenbach and Levine, 2021).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, while the theory of discrete time policy mirror descent is well understood for MDPs without KL regularisation (Tomar et al., 2022) and for tabular MDPs with KL regularisation (Lan, 2023), its stability and convergence remain open for KL-regularised MDPs in general state and action spaces. This paper fills this gap in the literature by rigorously proving the stability and convergence of policy mirror descent for KL-regularised MDPs in Polish state and action spaces, where the critic is updated using one step or multiple steps of temporal difference learning.

### 1.1 Related works

The theory of actor-critic algorithms for KL-regularised MDPs in the tabular setting has been studied extensively (Tomar et al., 2022; Qiu et al., 2021; Ju and Lan, 2026; Cayci et al., 2024a). However, many important applications involve continuous state and action spaces, naturally leading to Polish state/action spaces. We stress that the emphasis on Polish spaces is not merely technical. Polish spaces include Euclidean spaces, separable Hilbert spaces, and many common function spaces. This is relevant, for example, in control problems where the action itself is a function rather than a finite-dimensional vector, such as boundary control of PDEs (Lions, 1971). These methods have been applied to frontier problems in science, e.g., control of plasma in tokamaks (Degrave et al., 2022).

A key difficulty in such analyses is the coupling between the actor and critic updates. The actor step uses the current critic to form the policy improvement step, and hence any inaccuracy in the critic is fed directly into the next policy update. As a result, if the critic is not accurate enough, the actor update may produce a policy for which $\mathrm{KL}(\pi^n(\cdot \mid s) \mid \mu)$ becomes infinite. In the finite-action setting this issue does not arise: if the reference policy $\mu$ is uniform over $A$, then one always has $\mathrm{KL}(\pi^n(\cdot \mid s) \mid \mu) \leq \log |A|$ for every policy $\pi$, where $|A|$ denotes the cardinality of the action space. By contrast, in general action spaces, stability of the actor iterate depends on how accurately the critic is updated.

One attempt in answering this question is Zorba et al. (2025), which studies the Polish state and action space case through idealised continuous-time dynamics for policy mirror descent and TD learning. The continuous-time limit is mathematically convenient, but its connection to a given discrete algorithm is not automatic: many discrete schemes can have the same limiting dynamics as the step size goes to zero, while stability of the actual iterates as the number of steps goes to infinity is a separate question. Moreover, the convergence of the dynamics in Zorba et al. (2025) depends on an exponentially increasing timescale separation, whose algorithmic interpretation in discrete time is unclear.

It is precisely this interaction between critic accuracy and actor stability that our paper studies. To the best of our knowledge, this is the first work showing that for KL-regularised MDPs with general state/action spaces, one can guarantee stability of $\mathrm{KL}(\pi^n(\cdot \mid s) \mid \mu)$ by increasing the number of TD steps with the number of actor steps, and in particular that logarithmically many TD steps are sufficient for stability. The same coupling also explains the resulting convergence statements.

### 1.2 Outline of this paper and contributions

Our main contributions are as follows.

- We introduce the background material for KL-regularised MDPs and and policy mirror descent with a temporal difference (TD) critic in Section 1.4 and Section 2 respectively. Thereafter, we introduce and analyse a single-loop actor-critic algorithm in which the actor is updated using policy mirror descent and the critic is updated using one step of TD learning, per actor update. We derive explicit sufficient conditions under which the critic parameters and $\sup_s \mathrm{KL}(\pi^n(\cdot|s)|\mu)$ remain uniformly bounded along the iterations.
- Using the restrictive conditions required for the stability of the single-loop algorithm as motivation, we introduce a double-loop algorithm in Section 4 where one performs $M(n) \geq 1$ TD steps before performing the $n$th actor update. Under $Q^\pi_\tau$-realisability, we prove that performing logarithmically many inner TD steps is sufficient for stability.
- For the double-loop algorithm we derive explicit finite-time convergence rates in Section 4.2. More precisely, we demonstrate that when the number of TD steps, $M(n)$, grows logarithmically $M(n) \sim \log n$ we obtain a rate of $\mathcal{O}(1/n)$, and under an additional concentrability

assumption, $M(n) \sim n$ growth yields linear convergence. To the best of our knowledge, these results are the first of their kind in the literature.

### 1.3 Notation

Let $(E, d)$ denote a Polish space (i.e., a complete separable metric space). We always equip a Polish space with its Borel sigma-field $\mathcal{B}(E)$. Denote by $B_b(E)$ the space of bounded measurable functions $f : E \to \mathbb{R}$ endowed with the supremum norm $|f|_{B_b(E)} = \sup_{x \in E} |f(x)|$. Denote by $\mathcal{M}(E)$ the space of finite signed measures on $E$, and by $\mathcal{P}(E) \subset \mathcal{M}(E)$ the set of probability measures on $E$. Moreover, we denote the Euclidean norm on $\mathbb{R}^N$ by $|\cdot|$ with inner product $\langle \cdot, \cdot \rangle$. Given some $A, B \in \mathbb{R}^{N \times N}$, we denote by $\lambda_{\min}(A)$ the minimum eigenvalue of $A$.

### 1.4 KL-regularised Markov Decision Processes

Consider an infinite horizon Markov Decision Process $(S, A, P, c, \gamma)$, where the state space $S$ and action space $A$ are Polish, $P \in \mathcal{P}(S|S \times A)$ is the state transition probability kernel, $c$ is a bounded cost function and $\gamma \in (0, 1)$ is a discount factor. Let $\mu \in \mathcal{P}(A)$ denote a reference probability measure and $\tau > 0$ denote a regularisation parameter. To ease notation, for each $\pi \in \mathcal{P}(A|S)$, $s \in S$ and $a \in A$ we define $P_\pi(ds'|s) := \int_A P(ds'|s, a)\pi(da|s)$ and $P^\pi(ds', da'|s, a) := P(ds'|s, a)\pi(da'|s')$. For each stochastic policy $\pi \in \mathcal{P}(A|S)$ and $s \in S$, we define the regularised value function by

$$V_\tau^\pi(s) = \mathbb{E}_s^\pi \left[ \sum_{n=0}^{\infty} \gamma^n \Big( c(s_n, a_n) + \tau \, \mathrm{KL}(\pi(\cdot|s_n)|\mu) \Big) \right].$$

with $V_\tau^\pi(\rho) := \int_S V_\tau^\pi(s)\rho(ds)$ for any initial state distribution $\rho \in \mathcal{P}(S)$. Here $\mathrm{KL}(\pi(\cdot|s)|\mu)$ is the Kullback-Leibler (KL) divergence of $\pi(\cdot|s)$ with respect to $\mu$, defined as $\mathrm{KL}(\pi(\cdot|s)|\mu) := \int_A \log \frac{d\pi}{d\mu}(s, a)\pi(da|s)$ if $\pi(\cdot|s)$ is absolutely continuous with respect to $\mu$, and infinity otherwise. As a result, for any $s \in S$ we have that $V_\tau^\pi(s) \in \mathbb{R} \cup \{+\infty\}$. The optimal value function is defined as

$$V_\tau^*(\rho) = \inf_{\pi \in \mathcal{P}(A|S)} V_\tau^\pi(\rho).$$

We refer to $\pi^* \in \mathcal{P}(A|S)$ as the optimal policy if $V_\tau^*(\rho) = V_\tau^{\pi^*}(\rho)$. The Bellman Principle for KL-regularised MDPs, see Theorem A.1, ensures that without loss of generality, it is sufficient to consider policies from the class given by Definition 1.1 below.

**Definition 1.1** (Admissible Policies)**.** Let $\Pi_\mu$ denote the class of policies for which there exists $f \in B_b(S \times A)$ with

$$\pi(da|s) = \frac{\exp(f(s, a))}{\int_A \exp(f(s, a))\mu(da)}\mu(da).$$

Moreover, for each $\pi \in \Pi_\mu$, the state-action value function $Q_\tau^\pi \in B_b(S \times A)$ is defined as

$$Q_\tau^\pi(s, a) = c(s, a) + \gamma \int_S V_\tau^\pi(s')P(ds'|s, a).$$

and the soft advantage function $A_\tau^\pi \in B_b(S \times A)$ as

$$A_\tau^\pi(s, a) = Q_\tau^\pi(s, a) + \tau \log \frac{d\pi}{d\mu}(s, a) - V_\tau^\pi(s).$$

Finally we define the Bellman operator $\mathrm{T}_\tau^\pi : B_b(S \times A) \to B_b(S \times A)$ as

$$\mathrm{T}_\tau^\pi f(s, a) = c(s, a) + \gamma \int_{S \times A} f(s', a')P^\pi(ds', da'|s, a) + \tau\gamma \int_S \mathrm{KL}(\pi(\cdot|s')|\mu)P(ds'|s, a). \tag{1}$$

A direct calculation shows that for any $\pi \in \Pi_\mu$, $\mathrm{T}_\tau^\pi$ is a $\gamma$-contraction on $B_b(S \times A)$ with respect to $|\cdot|_{B_b(S \times A)}$ and hence by Banach's fixed point theorem $Q_\tau^\pi : S \times A \to \mathbb{R}$ is the unique fixed point of $\mathrm{T}_\tau^\pi$. It is worth noting that throughout this work, the reference measure $\mu \in \mathcal{P}(A)$ may also be a state-dependent, i.e $\mu \in \mathcal{P}(A|S)$ a probability kernel, as the existing principles of KL-regularised MDPs in general state and action spaces extend directly.

## 2 Mirror Descent and Temporal Difference

In this section we recap a classical actor-critic algorithm to find the optimal policy $\pi^* \in \Pi_\mu$ of the KL-regularised MDP. First consider the following policy mirror descent updates. For some $\lambda > 0$, define $\mathrm{MD} : \Pi_\mu \times \Pi_\mu \to \mathbb{R}$ as

$$\mathrm{MD}(\pi, \pi') = \int_S \left( \int_A A_\tau^{\pi'}(s,a)\pi(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi|\pi')(s) \right) d_\rho^{\pi'}(ds),$$

and consider the updates

$$\pi^{n+1} = \operatorname*{arg\,min}_{\pi \in \Pi_\mu} \mathrm{MD}(\pi, \pi^n). \tag{2}$$

In the setting where the advantage function is accessible for all $s \in S$ and $a \in A$, Lan (2023); Kerimkulov et al. (2025) demonstrate convergence of policy mirror descent for KL-regularised MDPs. To extend these works to the case of a parametrised critic, we consider the following finite-dimensional parametrisation of the advantage function. Given some feature mapping $\phi : S \times A \to \mathbb{R}^N$, we parametrise the state-action value function as $Q(s,a;\theta) := \langle \theta, \phi(s,a) \rangle$ and define the approximate soft advantage function as in Definition 2.1.

**Definition 2.1.** Let $Q(s,a;\theta) := \langle \theta, \phi(s,a) \rangle$, for some $\phi : S \times A \to \mathbb{R}^N$ and any $\tau > 0$. For all $s \in S$ and $a \in A$, the approximate advantage function is defined as

$$A(s,a;\theta,\pi) = Q(s,a;\theta) + \tau \log \frac{d\pi}{d\mu}(s,a) - \int_A \left( Q(s,a;\theta) + \tau \log \frac{d\pi}{d\mu}(s,a) \right) \pi(da|s). \tag{3}$$

To update the parameters $\theta \in \mathbb{R}^N$ of the critic, we perform semi-gradient descent on the mean-squared Bellman error, whose objective and semi-gradient are defined in Definition 2.2.

**Definition 2.2.** For some $\beta \in \mathcal{P}(S \times A)$, let $d_\beta^\pi \in \mathcal{P}(S \times A)$ be the state-action occupancy measure defined in Appendix A. The Mean Squared Bellman Error (MSBE) is defined as

$$\mathrm{MSBE}(\theta, \pi) := \frac{1}{2} \int_{S \times A} (Q(s,a;\theta) - \mathrm{T}_\tau^\pi Q(s,a;\theta))^2 d_\beta^\pi(da, ds).$$

Moreover, the semi-gradient $g : \mathbb{R}^N \times \mathcal{P}(A|S) \to \mathbb{R}^N$ of the MSBE with respect to $\theta$ is given by

$$g(\theta, \pi) := \int_{S \times A} (Q(s,a;\theta) - \mathrm{T}_\tau^\pi Q(s,a;\theta)) \phi(s,a)\, d_\beta^\pi(da, ds).$$

## 3 Single loop actor-critic

At each iteration of policy mirror descent (2), an ideal approach would be to compute $\theta^* \in \arg\min_{\theta \in \mathbb{R}^N} \mathrm{MSBE}(\theta, \pi^n)$ for each $n \in \mathbb{N}$ and use the parametrisation $A(s,a;\theta^*,\pi^n)$ for each $s \in S$ and $a \in A$. However, this requires solving an optimisation problem for each policy update, for example by performing many steps of TD learning until some small optimisation threshold is reached. Hence it is common in practice to perform a single step of temporal difference learning on the mean-squared Bellman error and use the updated parameters in the policy mirror descent update. More precisely, let

$$\widetilde{\mathrm{MD}}(\pi, \pi', \theta) = \int_S \left( \int_A A(s,a;\theta,\pi')\pi(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi|\pi')(s) \right) d_\rho^{\pi'}(ds),$$

then for some critic step size $h > 0$ and actor step size $\lambda > 0$, initial policy and parameters $\pi^0 = \pi_0 \in \Pi_\mu$, $\theta^0 = \theta_0 \in \mathbb{R}^N$, consider Algorithm 1.

**Algorithm 1** Single-loop actor-critic

**Require:** Critic step size $h > 0$, actor step size $\lambda > 0$, initial parameters $\theta^0 \in \mathbb{R}^N$, initial policy $\pi^0 \in \Pi_\mu$

1: **for** $n = 0, 1, 2, \ldots$ **do**
2: $\quad \theta^{n+1} \leftarrow \theta^n - h\, g(\theta^n, \pi^n)$
3: $\quad \pi^{n+1} \leftarrow \operatorname*{arg\,min}_{\pi \in \Pi_\mu} \widetilde{\mathrm{MD}}(\pi, \pi^n, \theta^{n+1})$
4: **end for**

As pointed out in Section 1.2, the stability of Algorithm 1 is still unresolved for KL-regularised MDPs in general spaces. We address this in the following section, under the following assumptions.

**Assumption 3.1.** Let $\beta \in \mathcal{P}(S \times A)$ be fixed. Then

$$\lambda_\beta := \lambda_{\min}\left(\int_{S\times A} \phi(s,a)\phi(s,a)^\top \beta(ds\,da)\right) > 0.$$

Note that unlike the analogous assumptions in the literature (Hong et al., 2023), Assumption 3.1 is independent of the policy and only depends on some $\beta \in \mathcal{P}(S \times A)$ and $\phi : S \times A \to \mathbb{R}^N$ which are both design choices. Assumption 3.1 is crucial for the stability analysis of Algorithm 1 and is equivalent to saying that no nonzero linear combination of the features vanishes $\beta$-almost everywhere.

**Assumption 3.2.** For all $(s,a) \in S \times A$ it holds that $|\phi(s,a)|_2 \leq 1$.

Assumption 3.2 is for convention and is without loss of generality in the finite-dimensional case.

### 3.1 Stability of Algorithm 1

Throughout this section, to ease notation we let $\Gamma := \lambda_\beta(1-\gamma)(1-\sqrt{\gamma})$. To establish conditions for stability of Algorithm 1, we first prove a coupled recursion for the critic parameters.

**Lemma 3.3.** *Let Assumption 3.1 and 3.2 hold. Let $0 < h \leq \frac{\Gamma}{6(1+\gamma)^2}$ and for some $\theta^0 = \theta_0 \in \mathbb{R}^N$ and $\pi^0 = \pi_0 \in \Pi_\mu$, let $\{\theta^n, \pi^n\}_{n\in\mathbb{N}}$ be the iterates for Algorithm 1. Then there exists $a_1, a_2 \geq 0$ such that for all $n \in \mathbb{N}$ it holds that*

$$|\theta^{n+1}|_2 \leq a_1 \sup_{0\leq r\leq n} \sup_{s\in S} \mathrm{KL}(\pi^r(\cdot|s)|\mu) + a_2.$$

See Appendix B.1 for a proof.

Lemma 3.3 demonstrates the coupling between the actor and the critic and its effect on stability. For example, suppose that we are dealing with an action space with finite cardinality, i.e., $|A| < \infty$. In this case, it holds that $\mathrm{KL}(\pi(\cdot|s)|\mu) \leq \log|A|$ for all $s \in S$ and $\pi \in \Pi_\mu$ given that $\min_{a\in A}\mu(a) > 0$. Thus, as an immediate corollary, given that $|A| < \infty$ one arrives at the stability of Algorithm 1 through the upper bound

$$|\theta^{n+1}|_2 \leq a_1 \log|A| + a_2.$$

However in the present setting of general action spaces, such a bound fails to hold. Thus to understand the scaling of Algorithm 1 to general spaces, one has to address exactly how the policy mirror descent updates couple with the temporal difference updates. The next result closes this loop by combining the critic recursion with a careful analysis of the log densities, Lemma E.1, arising from the policy mirror descent updates. Theorem 3.4 then shows that even if one just performs a single TD step for each policy update, Algorithm 1 can exhibit stable behaviour, but only under restrictive conditions on the effective horizon.

**Theorem 3.4.** *Let Assumptions 3.1 and 3.2 hold. Let*

$$0 < h \leq \frac{1}{2}\min\left\{\frac{\Gamma}{3(1+\gamma)^2}, \frac{\Gamma^2 - 32\gamma^2}{\Gamma\big(48\gamma^2 + 3(1+\gamma)^2\big)}\right\}$$

*and $0 < \tau\lambda < 1$. Moreover, let $\{\theta^n, \pi^n\}_{n\in\mathbb{N}}$ be the iterates of Algorithm 1. For $\frac{32\gamma^2}{\Gamma^2} < 1$, there exists $R \geq 0$ and for all $n \in \mathbb{N}$ and $s \in S$, it holds that*

$$\mathrm{KL}(\pi^n(\cdot|s)|\mu) \leq R.$$

See Appendix B.2 for a proof.

Theorem 3.4 suggests a single TD step may only be stable for a very restricted class of MDPs. This indicates that one needs more control over the critic accuracy in order to control the error propagation in the coupled updates. With this in mind, in Section 4 we analyse the setting where one performs $M(n) \geq 1$ steps of temporal difference for each policy update $n \in \mathbb{N}$, instead of $M(n) = 1$ for all $n \in \mathbb{N}$ as done in Algorithm 1, where the number of inner steps depends on the policy iterate $n \in \mathbb{N}$.

To establish the convergence of Algorithm 1, we work under $Q_\tau^\pi$-realisability.

**Assumption 3.5.** For all $\pi \in \Pi_\mu$ there exists $\theta_\pi \in \mathbb{R}^N$ such that $Q_\tau^\pi(s,a) = \langle \theta_\pi, \phi(s,a) \rangle$ for all $(s,a) \in S \times A$.

For the purposes of analytic tractability, the $Q_\tau^\pi$-realisability assumption is common in the literature (Devraj and Meyn, 2017; Meyn, 2024; Tian et al., 2024; Cayci et al., 2024a,b; Hong et al., 2023). One classical family satisfying this condition is the class of linear MDPs; see, e.g., Li et al. (2021); Yang and Wang (2019); Zanette et al. (2020). Moreover, $Q_\tau^\pi$-realisability holds in the limit $N \to \infty$ when $\phi_i$ are the basis functions of $L^2(\beta)$ for some $\beta \in \mathcal{P}(S \times A)$ (Brezis, 2011). With this perspective in mind, Ma et al. (2024); Ren et al. (2023) demonstrate empirical success of related finite-dimensional spectral and representation-based approximations in control problems. Our reason for adopting linear function approximation and realisability is to provide a baseline that can serve as a reference for future study of actor-critic algorithms for KL-regularised MDPs in general spaces under richer function approximation schemes, such as those studied in Wang et al. (2020); Li et al. (2021); Abbe et al. (2022); Chizat and Bach (2018); Liu et al. (2019). If exact realisability is omitted, the same proof strategy would have to track an additional approximation error measuring how well $Q_\tau^\pi$ can be represented in the span of $\phi$. This error would enter the critic recursion, affect the approximate advantage used by the actor, and ultimately appear as an additional residual term in the convergence guarantees, as in finite state-action analyses with approximation error (Qiu et al., 2021; Cayci et al., 2024a). We do not pursue this extension here because our goal is to isolate the stability mechanism caused by the coupling between TD critic error and KL mirror descent.

### 3.2 Convergence of Algorithm 1

In this section, we demonstrate convergence guarantees of Algorithm 1 in terms of the optimal value gap. These results hold in two cases. First, when the action space has finite cardinality, stability is guaranteed a priori through Lemma 3.3. Second, Theorem 3.6 also holds in the case of general action spaces under the conditions of Theorem 3.4.

**Theorem 3.6.** *Let Assumptions 3.1, 3.2 and 3.5 hold. Let $0 < \tau\lambda < 1$. For some $\theta^0 = \theta_0 \in \mathbb{R}^N$ and $\pi^0 = \pi_0 \in \Pi_\mu$ let $\{\theta^n, \pi^n\}_{n\in\mathbb{N}}$ be the iterates of Algorithm 1. Suppose that either $|A| < \infty$ with $\min_{a\in A} \mu(a) > 0$ or that the conditions of Theorem 3.4 hold. Let $0 < h \leq \frac{\Gamma}{6(1+\gamma)^2}$. Then there exists a constant $C \geq 0$ independent of $h, \lambda$ such that for any $\rho \in \mathcal{P}(S)$ and all $n \geq 1$, it holds that*

$$\min_{0\leq r\leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \leq C\left(\frac{1}{n\lambda} + \frac{1}{(nh)^{\frac{1}{2}}} + \frac{\lambda^{\frac{1}{2}}}{h(1-\tau\lambda)^{\frac{1}{2}}}\right).$$

See Appendix B.3 for a proof.

Theorem 3.6 clearly demonstrates the timescale separation required to guarantee the convergence of Algorithm 1. In particular, for a fixed admissible critic step size $h$ and sufficiently small $\varepsilon > 0$, choosing $\lambda$ proportional to $h^2\varepsilon^2$ with a sufficiently small proportionality constant gives an $\varepsilon$-optimal policy iterate within $\mathcal{O}\left(\frac{1}{h^2\varepsilon^3}\right)$ iterations.

## 4 Double loop actor-critic

We now formalise the double loop algorithm. Let $\{\theta^{n,k}\}_{k=0}^{M(n)}$ denote the total number of temporal difference steps for each $n \in \mathbb{N}$ and consider Algorithm 2.

**Algorithm 2** Double-loop actor-critic

**Require:** Critic step size $h > 0$, actor step size $\lambda > 0$, critic-step schedule $M : \mathbb{N} \to \mathbb{N}$, initial parameters $\theta^0 \in \mathbb{R}^N$, initial policy $\pi^0 \in \Pi_\mu$

**for** $n = 0, 1, 2, \ldots$ **do**
  $\theta^{n,0} \leftarrow \theta^n$
  **for** $k = 1, \ldots, M(n)$ **do**
    $\theta^{n,k} \leftarrow \theta^{n,k-1} - h\, g(\theta^{n,k-1}, \pi^n)$
  **end for**
  $\theta^{n+1} \leftarrow \theta^{n,M(n)}$
  $\pi^{n+1} \leftarrow \arg\min_{\pi\in\Pi_\mu} \widetilde{\mathrm{MD}}(\pi, \pi^n, \theta^{n+1})$
**end for**

We address the following question: For each $n \in \mathbb{N}$ and for all $\gamma \in (0,1)$, how many temporal-difference steps $M(n) \geq 1$ are sufficient to ensure stability of Algorithm 2 for KL-regularised MDPs in general spaces? It is worth noting that we only derive sufficient conditions for stability as necessary conditions for stability are problem dependent.

### 4.1 Stability of Algorithm 2

Intuitively, the number of TD steps $M(n)$ controls the accuracy of the critic with $M(n) \to \infty$ formally corresponding to obtaining the exact state-action value function. In this formal limit, properties from the study of policy mirror descent with exact advantage/ critic should follow, such as dissipation of the value function. With that in mind, Theorem 4.1 establishes a policy improvement property of the actor-critic updates, up to a critic error which in turn is controlled by $M(n)$.

**Theorem 4.1.** *Let Assumptions 3.1, 3.2 and 3.5 hold. Let* $0 < h < \min\left\{\frac{\Gamma}{2(1+\gamma)}, \frac{1}{\Gamma}\right\}$. *Then for all* $n \in \mathbb{N}$ *and* $s \in S$ *it holds that*

$$V_\tau^{\pi^*}(s) \leq V_\tau^{\pi^{n+1}}(s) \leq V_\tau^{\pi^n}(s) + \frac{2e^{-\frac{M(n)h\Gamma}{2}}}{1-\gamma}|\theta^n - \theta_{\pi^n}|_2.$$

See Appendix C.1 for a proof. Theorem 4.2 then gives a sufficient growth condition on the number of inner temporal-difference steps required to guarantee the stability of Algorithm 2 for KL-regularised MDPs in Polish action spaces.

**Theorem 4.2.** *Let Assumptions 3.1, 3.2 and 3.5 hold. Let* $0 < h < \min\left\{\frac{\Gamma}{2(1+\gamma)}, \frac{1}{\Gamma}\right\}$ *and* $0 < \tau\lambda < 1$. *Then there exists* $c_0 > 0$ *such that, for any* $c \geq c_0$, *if the number of inner temporal difference steps at mirror descent step* $n \in \mathbb{N}$ *satisfies*

$$M(n) \geq \frac{4}{h\Gamma}\log\left(c(n+1)\right),$$

*then there exists* $R \geq 0$ *such that for all* $s \in S$, $\gamma \in (0,1)$ *and* $n \in \mathbb{N}$ *it holds that*

$$\mathrm{KL}(\pi^n(\cdot|s)|\mu) \leq R.$$

See Appendix C.2 for a proof.

Theorem 4.2 demonstrates that logarithmic growth of the number of temporal difference steps is sufficient for stability of Algorithm 2 in general Polish action spaces. In the following section, we demonstrate the convergence of Algorithm 2 when one increases the number of TD steps logarithmically and linearly.

### 4.2 Convergence of Algorithm 2

Now that we have established that a logarithmically increasing schedule for the number of inner TD steps suffices for the stability of Algorithm 2, we prove that this schedule yields a sublinear convergence rate to the optimal policy.

**Theorem 4.3.** *Let Assumptions 3.1, 3.2 and 3.5 hold and let* $0 < h < \frac{\Gamma}{2(1+\gamma)}$. *Moreover for each* $n \in \mathbb{N}$, *let the number of inner critic steps satisfy* $M(n) \geq \frac{4}{h\Gamma}\log\left(c(n+1)\right)$ *with* $c \geq c_0$, *where* $c_0$ *is the stability constant from Theorem 4.2. Then, for any* $\rho \in \mathcal{P}(S)$, *there exists* $a \geq 0$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$\min_{0 \leq r \leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \leq \frac{a}{n}.$$

See Appendix D.1 for a proof.

Theorem 4.4 demonstrates that under a concentrability assumption, if we increase the number of temporal difference steps linearly with $n \in \mathbb{N}$, we obtain linear convergence of Algorithm 2.

**Theorem 4.4.** *Let Assumptions 3.1, 3.2 and 3.5 hold. Moreover suppose that for some* $2 \leq \xi < \infty$,

$$\frac{1}{1-\gamma}\Big|\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}\rho}\Big|_{B_b(S)} \leq \xi.$$

*Let* $0 < h < \frac{\Gamma}{2(1+\gamma)}$ *and* $\frac{1}{\xi} \leq \tau\lambda \leq \frac{\xi-1}{\xi}$. *For each policy update* $n \in \mathbb{N}$ *let the number of inner temporal difference steps satisfy* $M(n) \geq \frac{4c}{h\Gamma}(n+1)$ *with* $c \geq c_0$ *where* $c_0$ *is the constant from Theorem 4.2. Then there exists* $b \geq 0$ *such that for all* $n \in \mathbb{N}$ *it holds that*

$$V_\tau^{\pi^n}(\rho) - V_\tau^{\pi^*}(\rho) \leq be^{-\min\left\{\frac{1}{\xi}, c\right\}n}.$$

See Appendix D.2 for a proof.

Theorems 4.3 and 4.4 are in terms of the actor steps, while the actual critic work is the total number of TD steps used inside those updates. For the logarithmic schedule, reaching an $\varepsilon$-optimal iterate requires $\mathcal{O}(\frac{1}{\varepsilon})$ policy updates; thus a direct calculation shows that the cumulative number of TD steps up to that point is $\mathcal{O}(\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon}))$. Similarly, under the conditions of Theorem 4.4, the linear schedule needs only $\mathcal{O}(\log(\frac{1}{\varepsilon}))$ policy updates; thus the cumulative number of TD steps up to that point is $\mathcal{O}(\log^2(\frac{1}{\varepsilon}))$.

## 5 Conclusion, limitations, and future directions

In this work, we study the stability and convergence of a fundamental actor-critic algorithm for KL-regularised MDPs on general state-action spaces under $Q_\tau^\pi$-realisability with linear function approximation. In the algorithm the policy is updated through mirror descent and the critic parameters are updated through TD learning.

We provide sufficient, but restrictive, conditions guaranteeing stability and convergence for the single-step variant. For the multi-step variant we avoid the restrictive single-step stability condition by appropriately controlling the number of TD updates per policy update, and establish sublinear and, under an additional concentrability assumption, linear convergence rates.

To the best of our knowledge, these results address a significant theoretical gap for actor-critic methods in KL-regularised MDPs. To focus on the algorithm's behaviour in general spaces, we analyse the population setting in which all integrals are evaluated exactly. Extending the analysis to the sample-based setting with inexact gradients remains an important direction for future work.


## Acknowledgments and Disclosure of Funding

DZ was supported by the EPSRC Centre for Doctoral Training in Mathematical Modelling, Analysis and Computation (MAC-MIGS) funded by the UK Engineering and Physical Sciences Research Council (grant EP/S023291/1), Heriot-Watt University and the University of Edinburgh. We acknowledge funding from the UKRI Prosperity Partnerships grant APP43592: AI2 – Assurance and Insurance for Artificial Intelligence, which supported this work. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Bridging Stochastic Control And Reinforcement Learning, where work on this paper was undertaken. This work was supported by EPSRC grant EP/V521929/1.

The authors declare no competing interests.


## References


Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The Merged-Staircase Property: A Necessary and Nearly Sufficient Condition for SGD Learning of Sparse Functions on Two-Layer Neural Networks. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4782–4887. PMLR, 2022.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the Impact of Entropy on Policy Optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 09–15 Jun 2019.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror Descent with Relative Smoothness in Measure Spaces, with Application to Sinkhorn and EM. In *Advances in Neural Information Processing Systems*, volume 35, pages 17263–17275, 2022.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, NY, 1 edition, 2011. ISBN 978-0-387-70913-0. doi: 10.1007/978-0-387-70914-7. Published: 02 November 2010, eBook ISBN: 978-0-387-70914-7, Series ISSN: 0172-5939, Series E-ISSN: 2191-6675.

Semih Cayci, Niao He, and R. Srikant. Convergence of Entropy-Regularized Natural Policy Gradient with Linear Function Approximation. *SIAM Journal on Optimization*, 34(3):2729–2755, 2024a. doi: 10.1137/22M1540156.

Semih Cayci, Niao He, and R. Srikant. Finite-Time Analysis of Entropy-Regularized Neural Natural Actor-Critic Algorithm. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856.

Lénaïc Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models Using Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 31, pages 3040–3050, 2018.

J. Degrave, F. Felici, J. Buchli, et al. Magnetic Control of Tokamak Plasmas through Deep Reinforcement Learning. *Nature*, 602:414–419, 2022. doi: 10.1038/s41586-021-04301-9.

Adithya M Devraj and Sean Meyn. Zap Q-Learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Paul Dupuis and Richard S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1997. ISBN 9780471076728. doi: 10.1002/9781118165904.

Benjamin Eysenbach and Sergey Levine. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. *arXiv preprint arXiv:2103.06257*, 2021.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement Learning with Deep Energy-Based Policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A Two-Timescale Stochastic Algorithm Framework for Bilevel Optimization: Complexity Analysis and Application to Actor-Critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023. doi: 10.1137/20M1387341.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in Neural Information Processing Systems*, 6, 1993.

Caleb Ju and Guanghui Lan. Policy Optimization over General State and Action Spaces. *arXiv preprint arXiv:2211.16715*, 2026.

Sham M. Kakade and John Langford. Approximately Optimal Approximate Reinforcement Learning. In *International Conference on Machine Learning*, 2002.

Bekzhan Kerimkulov, James-Michael Leahy, David Siska, Lukasz Szpruch, and Yufei Zhang. A Fisher-Rao Gradient Flow for Entropy-Regularised Markov Decision Processes in Polish Spaces. *Foundations of Computational Mathematics*, 2025. doi: 10.1007/s10208-025-09729-3.

Guanghui Lan. Policy Mirror Descent for Reinforcement Learning: Linear Convergence, New Sampling Complexity, and Generalized Problem Classes. *Mathematical Programming*, 198(1): 1059–1106, 2023. doi: 10.1007/s10107-022-01816-5.

James-Michael Leahy, Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Convergence of Policy Gradient for Entropy Regularized MDPs with Neural Network Approximation in the Mean-Field Regime. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 12222–12252. PMLR, 2022.

Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, and Yuting Wei. Sample-Efficient Reinforcement Learning Is Feasible for Linearly Realizable MDPs with Limited Revisiting. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*, volume 170 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 1971.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural Trust Region/Proximal Policy Optimization Attains Globally Optimal Policy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

H. Ma, Z. Ren, B. Dai, and N. Li. Skill Transfer and Discovery for Sim-to-Real Learning: A Representation-Based Viewpoint. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8603–8609, Abu Dhabi, United Arab Emirates, 2024. doi: 10.1109/IROS58592.2024.10801637.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Mingzhe Lu, Kristina Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High Performance GPU Based Physics Simulation for Robot Learning. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

Sean Meyn. The Projected Bellman Equation in Reinforcement Learning. *IEEE Transactions on Automatic Control*, 69(12):8323–8337, December 2024. doi: 10.1109/TAC.2024.3409647.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, Feb 2015.

Shuang Qiu, Zhaoran Yang, Jianfeng Ye, and Zhuoran Wang. On Finite-Time Convergence of Actor-Critic Algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, June 2021. doi: 10.1109/JSAIT.2021.3078754.

Konrad Rawlik, Marc Toussaint, and Sethu Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Proceedings of Robotics: Science and Systems VIII*, 2012.

T. Ren, Z. Ren, N. Li, and B. Dai. Stochastic Nonlinear Control via Finite-dimensional Spectral Dynamic Embedding. In *Proceedings of the 2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 795–800, Singapore, Singapore, 2023. doi: 10.1109/CDC49753.2023.10383842.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust Region Policy Optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Tian Tian, Lin F. Yang, and Csaba Szepesvári. Confident Natural Policy Gradient for Local Planning in $q_\pi$-Realizable Constrained MDPs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 76139–76176. Curran Associates, Inc., 2024. doi: 10.52202/079017-2424.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror Descent Policy Optimization. In *International Conference on Learning Representations*, 2022.

John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3): 185–202, 1994.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural Policy Gradient Methods: Global Optimality and Rates of Convergence. In *International Conference on Learning Representations*, 2020.

Lin Yang and Mengdi Wang. Sample-Optimal Parametric Q-Learning Using Linearly Additive Features. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 6995–7004. PMLR, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist Regret Bounds for Randomized Least-Squares Value Iteration. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 1954–1964. PMLR, 2020.

Yufeng Zhang, Siyu Chen, Zhuoran Yang, Michael Jordan, and Zhaoran Wang. Wasserstein Flow Meets Replicator Dynamics: A Mean-Field Analysis of Representation Learning in Actor-Critic. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Brian D Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. Carnegie Mellon University, 2010.

Denis Zorba, David Šiška, and Lukasz Szpruch. Convergence of actor-critic for entropy regularised MDPs in general action spaces. *arXiv preprint arXiv:2510.14898*, 2025.

# A Technical Details

The state-occupancy kernel $d^\pi \in \mathcal{P}(S|S)$ is defined by

$$d^\pi(ds'|s) = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n P_\pi^n(ds'|s)\,, \tag{4}$$

where $P_\pi^n$ is the $n$-times product of the kernel $P_\pi$ with $P_\pi^0(ds'|s) \coloneqq \delta_s(ds')$. Moreover, for each $\pi \in \mathcal{P}(A|S)$ and $(s,a) \in S \times A$, we define the state-action occupancy kernel as

$$d^\pi(ds, da|s, a) = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n (P^\pi)^n(ds, da|s, a)$$

where $(P^\pi)^n$ is the $n$-times product of the kernel $P^\pi$ with $(P^\pi)^0(ds', da'|s, a) \coloneqq \delta_{(s,a)}(ds', da')$. Given some initial state-action distribution $\beta \in \mathcal{P}(S \times A)$ with initial state distribution $\rho(ds) = \int_A \beta(da, ds)$, we define the state-occupancy and state-action occupancy measures as

$$d_\rho^\pi(ds) = \int_S d^\pi(ds|s')\rho(ds'), \quad d_\beta^\pi(ds, da) = \int_{S\times A} d^\pi(ds, da|s', a')\beta(da', ds').$$

Note that for all $E \in \mathcal{B}(S \times A)$, by defining the linear operator $J_\pi : \mathcal{P}(S \times A) \to \mathcal{P}(S \times A)$ as

$$J_\pi\beta(E) = \int_{S\times A} P^\pi(E|s', a')\beta(ds', da'), \tag{5}$$

it directly holds that

$$d_\beta^\pi(da, ds) = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n J_\pi^n\beta(da, ds),$$

with $J_\pi^n$ the $n$-fold product of the operator $J_\pi$ with $J_\pi^0 = I$, the identity operator on $\mathcal{P}(S \times A)$.

A proof for the following dynamic programming principle for KL-regularised MDPs in general spaces can be found in Kerimkulov et al. (2025).

**Theorem A.1** (Dynamic Programming Principle). *Let $\tau > 0$. The optimal value function $V_\tau^*$ is the unique bounded solution of the following Bellman equation:*

$$V_\tau^*(s) = -\tau \log \int_A \exp\left(-\frac{1}{\tau} Q_\tau^*(s,a)\right) \mu(da),$$

*where $Q_\tau^* \in B_b(S \times A)$ is defined by*

$$Q_\tau^*(s,a) = c(s,a) + \gamma \int_S V_\tau^*(s') P(ds'|s,a)\,, \quad \forall (s,a) \in S \times A\,.$$

*Moreover, there is an optimal policy $\pi^* \in \mathcal{P}(A|S)$ given by*

$$\pi^*(da|s) = \exp\left(-\frac{1}{\tau}(Q_\tau^*(s,a) - V_\tau^*(s))\right) \mu(da)\,, \quad \forall s \in S.$$

*Finally, the value function $V_\tau^\pi$ is the unique bounded solution of the following Bellman equation for all $s \in S$*

$$V_\tau^\pi(s) = \int_A \left(Q_\tau^\pi(s,a) + \tau \log \frac{d\pi}{d\mu}(s,a)\right) \pi(da|s)\,.$$

The performance difference lemma, first introduced for tabular unregularised MDPs, has become fundamental in the analysis of MDPs as it acts a substitute for the strong convexity of the $\pi \mapsto V_\tau^\pi$ if the state-occupancy measure $d_\rho^\pi$ is ignored; see, e.g., Kakade and Langford (2002); Zhang et al. (2021); Ju and Lan (2026). By virtue of Kerimkulov et al. (2025), we have the following performance difference for KL-regularised MDPs in Polish state and action spaces.

**Lemma A.2** (Performance difference). *For all $\rho \in \mathcal{P}(S)$ and $\pi, \pi' \in \Pi_\mu$,*

$$\begin{aligned} &V_\tau^\pi(\rho) - V_\tau^{\pi'}(\rho) \\ &= \frac{1}{1-\gamma} \int_S \left[ \int_A \left(Q_\tau^{\pi'}(s,a) + \tau \log \frac{d\pi'}{d\mu}(s,a)\right)(\pi - \pi')(da|s) + \tau\, \mathrm{KL}(\pi(\cdot|s)|\pi'(\cdot|s)) \right] d_\rho^\pi(ds)\,. \end{aligned}$$

The following three lemmas become useful in the stability analysis of both Algorithm 1 and Algorithm 2 as well as the convergence analysis of Algorithm 2.

**Lemma A.3.** *For all $\pi \in \Pi_\mu$ and $s \in S$ it holds that*

$$\mathrm{KL}(\pi(\cdot|s)|\mu) \le 2 \left| \log \frac{d\pi}{d\mu}(s,\cdot) - \int_A \log \frac{d\pi}{d\mu}(s,a)\mu(da) \right|_{B_b(A)}.$$

*Proof.* By the definition and non-negativity of KL divergence it holds that

$$\begin{aligned} \mathrm{KL}(\pi(\cdot|s)|\mu) &\le \mathrm{KL}(\pi(\cdot|s)|\mu) + \mathrm{KL}(\mu|\pi(\cdot|s)) \\ &= \int_A \log \frac{d\pi}{d\mu}(s,a)\pi(da|s) + \int_A \log \frac{d\mu}{d\pi}(s,a)\mu(da) \\ &= \int_A \log \frac{d\pi}{d\mu}(s,a)(\pi(da|s) - \mu(da)) \\ &= \int_A \left(\log \frac{d\pi}{d\mu}(s,a) - \int_A \log \frac{d\pi}{d\mu}(s,a)\mu(da)\right)(\pi(da|s) - \mu(da)) \\ &\le \left| \log \frac{d\pi}{d\mu}(s,\cdot) - \int_A \log \frac{d\pi}{d\mu}(s,a)\mu(da) \right|_{B_b(A)} |\pi(\cdot|s) - \mu|_{\mathcal{M}(A)} \\ &\le 2 \left| \log \frac{d\pi}{d\mu}(s,\cdot) - \int_A \log \frac{d\pi}{d\mu}(s,a)\mu(da) \right|_{B_b(A)}. \end{aligned}$$

where we used that $|\pi(\cdot|s) - \mu|_{\mathcal{M}(A)} \le 2$ for all $s \in S$ in the final inequality. □

**Lemma A.4.** *For all $n \ge 0$, let $\{a_n\}_{n\ge 0}$ be a sequence of non-negative numbers such that*

$$a_{n+1} \le \kappa A_n + c$$

*with $c \ge 0$, $0 < \kappa < 1$ and $\mathrm{A}_n = \sup_{0\le r\le n} a_r$. Then for all $n \in \mathbb{N}$ it holds that*

$$\mathrm{A}_n \le a_0 + \frac{c}{1-\kappa}.$$

*Proof.* By definition, for any $n \in \mathbb{N}$ it holds that

$$A_{n+1} = \max\{A_n, a_{n+1}\} \leq \max\{A_n, \kappa A_n + c\}.$$

If $\max\{A_n, a_{n+1}\} = A_n$, then $A_{n+1} \leq A_0 + \frac{c}{1-\kappa} = a_0 + \frac{c}{1-\kappa}$ holds trivially for all $n \in \mathbb{N}$. On the other hand, we have

$$A_{n+1} \leq a_{n+1} \leq \kappa A_n + c.$$

Iterating this inequality also yields $A_{n+1} \leq A_0 + \frac{c}{1-\kappa} = a_0 + \frac{c}{1-\kappa}$. □

**Lemma A.5** (Zorba et al. (2025))**.** *Let $Q_\tau^\pi$-realisability hold. Then for all $\theta \in \mathbb{R}^N$ and $\pi \in \Pi_\mu$ it holds that*

$$-\langle g(\theta, \pi), \theta - \theta_\pi \rangle \leq -(1 - \sqrt{\gamma})(1 - \gamma)\langle \nabla_\theta L(\theta, \pi; \beta), \theta - \theta_\pi \rangle$$

*with*

$$\nabla_\theta L(\theta, \pi; \beta) = \int_{S \times A} (\langle \theta, \phi(s, a) \rangle - Q_\tau^\pi(s, a))\phi(s, a)\beta(da, ds).$$

**Lemma A.6.** *Suppose that $\mathrm{KL}(\pi^n(\cdot|s)|\mu) \leq R$ for all $n \in \mathbb{N}$ and $s \in S$. Then it holds that*

$$\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)} \leq \frac{1}{1-\gamma}\left(|c|_{B_b(S\times A)} + \tau\gamma R\right),$$
$$|\theta_{\pi^n}|_2 \leq \frac{1}{(1-\gamma)\lambda_\beta}\left(|c|_{B_b(S\times A)} + \tau\gamma R\right),$$

*for all $n \in \mathbb{N}$. Moreover suppose that the critic step size satisfies $0 < h \leq \frac{\Gamma}{6(1+\gamma)^2}$, then there exists $\tilde{R} \geq 0$ independent of $h, \lambda > 0$ such that for all $n \geq 0$ it holds that*

$$|\theta^{n+1}|_2^2 \leq \tilde{R}.$$

*Proof.* Recall that for any $\pi \in \Pi_\mu$, $Q_\tau^\pi \in B_b(S \times A)$ is a fixed point of the Bellman operator defined in (1). Therefore it holds that

$$Q_\tau^\pi(s, a) = c(s, a) + \gamma \int_{S\times A} Q_\tau^\pi(s', a')P^\pi(ds', da'|s, a) + \tau\gamma \int_S \mathrm{KL}(\pi(\cdot|s')|\mu)P(ds'|s, a).$$

Taking the $|\cdot|_{B_b(S\times A)}$ norm on both sides gives the result. Now let $\Sigma_\beta = \int_{S\times A} \phi(s, a)\phi(s, a)^\top \beta(da, ds)$. By $Q_\tau^\pi$-realisability and Assumptions 3.2 and 3.1 it holds that

$$\begin{aligned}|\theta_\pi|_2 &= \left|\Sigma_\beta \Sigma_\beta^{-1}\theta_\pi\right|_2 \\ &\leq \frac{1}{\lambda_\beta}\left|Q_\tau^\pi\right|_{B_b(S\times A)}.\end{aligned}$$

Using the first result then yields the desired bound.

Furthermore, suppose that $0 < h < \frac{\Gamma}{3(1+\gamma)^2}$. By Lemma 3.3, it holds that

$$\begin{aligned}|\theta^{n+1}|_2^2 &\leq |\theta_0|_2^2 + \frac{2\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}R^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2} \\ &= |\theta_0|_2^2 + 2\tau^2\gamma^2R^2\alpha(h) + |c|_{B_b(S\times A)}^2\alpha(h)\end{aligned}$$

where $\alpha(h) = \frac{3h+\frac{2}{\Gamma}}{\Gamma - 3h(1+\gamma)^2}$ and observe that $\alpha(h)$ is non-decreasing and positive for all $h \in \left(0, \frac{\Gamma}{3(1+\gamma)^2}\right)$. Hence we further restrict the critic step size to

$$0 < h \leq \bar{h} := \frac{\Gamma}{6(1+\gamma)^2}.$$

Thus it holds that $\alpha(h) \leq \alpha(\bar{h})$ for all $h \in \left(0, \frac{\Gamma}{6(1+\gamma)^2}\right]$. Substituting this we obtain that for all $n \in \mathbb{N}$ we have

$$|\theta^{n+1}|_2^2 \leq \bar{R}$$

for some $\bar{R} > 0$ independent of $h, \lambda > 0$. □

### A.1 Convergence proof ingredients

Note that throughout the remainder of this work, to ease notation for any $n \in \mathbb{N}$ we will denote the soft advantage function as $A(s,a;\theta^{n+1},\pi^n) := A(s,a;\theta^{n+1})$. One main ingredient of the proofs is the following Bregman proximal inequality. Let

$$M_\mu = \left\{ m \in \mathcal{P}(A) : \log \frac{dm}{d\mu} \in B_b(A) \right\},$$

and notice that this is a convex subset of $\mathcal{P}(A)$. A proof of the following classical three-point lemma can be found in Aubin-Frankowski et al. (2022).

**Lemma A.7** (Three point lemma/Bregman proximal inequality)**.** *Let $G : M_\mu \to \mathbb{R}$ be convex. For all $m' \in M_\mu$ let*

$$m^* = \operatorname*{arg\,min}_{m \in M_\mu} \left\{ G(m) + \mathrm{KL}(m|m') \right\}.$$

*Then for all $m \in M_\mu$ we have*

$$G(m) + \mathrm{KL}(m|m') \geq G(m^*) + \mathrm{KL}(m|m^*) + \mathrm{KL}(m^*|m')$$

**Lemma A.8.** *Let $F : S \to \mathbb{R}$ be such that $F \leq 0$. Then for any $\pi \in \Pi_\mu$ and $s \in S$ it holds that*

$$\frac{1}{1-\gamma} \int_S F(s') d_s^\pi(ds') \leq F(s)$$

*Proof.* By definition of the state occupancy measure $d^\pi(\cdot|s) \in \mathcal{P}(S)$ in (4) for any $s \in S$ and the fact that $P_\pi^0(ds'|s) = \delta_s(ds')$, we have for all $s \in S$ and $\pi \in \mathcal{P}(A|S)$ that

$$\begin{aligned} \frac{1}{1-\gamma} \int_S F(s') d_s^\pi(s') &= \int_S F(s') P_\pi^0(ds'|s) + \sum_{k=1}^\infty \int_S \gamma^k F(s') P_\pi^k(ds'|s) \\ &\leq \int_S F(s') \delta_s(ds') = F(s). \end{aligned}$$

This concludes the proof. □

**Theorem A.9.** *For some $\theta^0 = \theta_0$ and $\pi^0 = \pi_0$, let $\left\{\pi^k, \theta^k\right\}_{k \geq 0}$ be the iterates of Algorithm 1 and Algorithm 2. Suppose that $0 < \tau\lambda < 1$, then for all $\rho \in \mathcal{P}(S)$ and $n \in \mathbb{N}$ it holds that*

$$\begin{aligned} &\min_{0 \leq r \leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \qquad (6) \\ &\leq \frac{1}{\lambda(1-\gamma)n} \left( \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) + \lambda \left( V^0(\rho) - V^*(\rho) \right) + \lambda c(\gamma) \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2 \right). \end{aligned}$$

*Proof.* To ease notation let $V^k := V_\tau^{\pi^k}$ for $k \in \mathbb{N}$ and let $V^* := V_\tau^{\pi^*}$. Fix $s \in S$ and $\pi^k \in \Pi_\mu$. By the Three Point Lemma we have

$$\begin{aligned} &\lambda \int_A A(s,a;\theta^{k+1})(\pi - \pi^k)(da|s) + \mathrm{KL}(\pi|\pi^k)(s) \\ &\geq \lambda \int_A A(s,a;\theta^{k+1})(\pi^{k+1} - \pi^k)(da|s) + \mathrm{KL}(\pi|\pi^{k+1})(s) + \mathrm{KL}(\pi^{k+1}|\pi^k)(s). \end{aligned}$$

Rearranging, dividing through by $\lambda > 0$ and using that $0 < \tau\lambda < 1$ we have

$$\begin{aligned} &\frac{1}{\lambda} \left( \mathrm{KL}(\pi|\pi^{k+1})(s) - \mathrm{KL}(\pi|\pi^k)(s) \right) \qquad (7) \\ &\leq \int_A A(s,a;\theta^{k+1})(\pi - \pi^k)(da|s) - \int_A A(s,a;\theta^{k+1})(\pi^{k+1} - \pi^k)(da|s) - \tau\, \mathrm{KL}(\pi^{k+1}|\pi^k)(s). \end{aligned}$$

Now define

$$F(s) = \int_A A(s,a;\theta^{k+1})(\pi^{k+1} - \pi^k)(da|s) + \tau\, \mathrm{KL}(\pi^{k+1}|\pi^k)(s).$$

Recall that by the policy mirror descent updates (as clarified in the calculation in (14)), it holds that $F(s) \leq 0$ for all $s \in S$. Therefore, by the performance difference lemma and Lemma A.8 it holds that

$$\begin{aligned}
&(V^{k+1} - V^k)(s) \\
&= \frac{1}{1-\gamma} \int_S \left( \int_A A_\tau^{\pi^k}(s', a)(\pi^{k+1} - \pi^k)(da|s') + \tau \, \mathrm{KL}(\pi^{k+1}|\pi^k)(s') \right) d_s^{\pi^{k+1}}(ds') \\
&= \frac{1}{1-\gamma} \int_S F(s') d_s^{\pi^{k+1}}(ds') \\
&\quad + \frac{1}{1-\gamma} \int_S \int_A \left( A_\tau^{\pi^k}(s', a) - A(s', a; \theta^{k+1}) \right) (\pi^{k+1} - \pi^k)(da|s') d_s^{\pi^{k+1}}(ds') \\
&\leq \frac{1}{1-\gamma} \int_S F(s') d_s^{\pi^{k+1}}(ds') + \frac{2}{1-\gamma} |\theta^{k+1} - \theta_{\pi^k}|_2 \\
&\leq F(s) + \frac{2}{1-\gamma} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

where we added and subtracted the approximate advantage function in the second equality and used that for all $s \in S$ and $a \in A$, by Hölder's inequality, $Q_\tau^\pi$-realisability and Assumption 3.2, it holds that

$$A(s, a; \theta^{k+1}) - A_\tau^{\pi^k}(s, a) \leq \left| \theta^{k+1} - \theta_{\pi^k} \right|_2.$$

Substituting this into (7), for all $s \in S$ and $k \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\frac{1}{\lambda} \left( \mathrm{KL}(\pi|\pi^{k+1})(s) - \mathrm{KL}(\pi|\pi^k)(s) \right) \\
&\leq \int_A A(s, a; \theta^{k+1})(\pi - \pi^k)(da|s) - \int_A A(s, a; \theta^{k+1})(\pi^{k+1} - \pi^k)(da|s) - \tau \, \mathrm{KL}(\pi^{k+1}|\pi^k)(s) \\
&\leq \int_A A(s, a; \theta^{k+1})(\pi - \pi^k)(da|s) - F(s) \\
&\leq \int_A A(s, a; \theta^{k+1})(\pi - \pi^k)(da|s) - \left( V^{k+1}(s) - V^k(s) \right) + \frac{2}{1-\gamma} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

Now let $\pi = \pi^*$. Summing both sides over $k = 0, 1, \ldots, n-1$ for any $n \in \mathbb{N}$ we have that

$$\begin{aligned}
&\mathrm{KL}(\pi^*|\pi^n)(s) - \mathrm{KL}(\pi^*|\pi^0)(s) \\
&\leq \lambda \sum_{k=0}^{n-1} \int_A A(s, a; \theta^{k+1})(\pi^* - \pi^k)(da|s) - \lambda (V^n - V^0)(s) + \frac{2\lambda}{1-\gamma} \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

Adding and subtracting the true advantage function, we have

$$\begin{aligned}
&\mathrm{KL}(\pi^*|\pi^n)(s) - \mathrm{KL}(\pi^*|\pi^0)(s) \\
&\leq \lambda \sum_{k=0}^{n-1} \int_A A_\tau^{\pi^k}(s, a)(\pi^* - \pi^k)(da|s) - \lambda (V^n - V^0)(s) \\
&\quad + \lambda \sum_{k=0}^{n-1} \int_A \left( A(s, a; \theta^{k+1}) - A_\tau^{\pi^k}(s, a) \right) (\pi^* - \pi^k)(da|s) + \frac{2\lambda}{1-\gamma} \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

Let $c(\gamma) := \max\left\{1, \frac{2}{1-\gamma}\right\}$. Again using that $A(s, a; \theta^{k+1}) - A_\tau^{\pi^k}(s, a) \leq \left| \theta^{k+1} - \theta_{\pi^k} \right|_2$ for all $s \in S$, $a \in A$ and $k \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\mathrm{KL}(\pi^*|\pi^n)(s) - \mathrm{KL}(\pi^*|\pi^0)(s) \\
&\leq \lambda \sum_{k=0}^{n-1} \int_A A_\tau^{\pi^k}(s, a)(\pi^* - \pi^k)(da|s) - \lambda (V^n - V^0)(s) + \lambda c(\gamma) \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

Moreover, using that $(V^n(s) - V^0(s)) \geq (V^*(s) - V^0(s))$, which holds for all $s \in S$ by the definition of $V^*$, and integrating over $d_\rho^{\pi^*} \in \mathcal{P}(S)$ we obtain

$$\begin{aligned}
&\int_S \mathrm{KL}(\pi^*|\pi^n)(s) d_\rho^{\pi^*}(ds) - \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) \\
&\leq \lambda \sum_{k=0}^{n-1} \int_S \int_A A_\tau^{\pi^k}(s,a)(\pi^* - \pi^k)(da|s) d_\rho^{\pi^*}(ds) \\
&\qquad - \lambda(V^*(\rho) - V^0(\rho)) + \lambda c(\gamma) \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

Finally, applying the performance difference lemma on the first term on the right hand side, we obtain

$$\begin{aligned}
&\int_S \mathrm{KL}(\pi^*|\pi^n)(s) d_\rho^{\pi^*}(ds) - \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) \\
&\leq \lambda(1-\gamma) \sum_{k=0}^{n-1} \left( (V^* - V^k)(\rho) - \tau \int_S \mathrm{KL}(\pi^*|\pi^k)(s) d_\rho^{\pi^*}(ds) \right) - \lambda(V^*(\rho) - V^0(\rho)) \\
&\quad + \lambda c(\gamma) \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2.
\end{aligned}$$

Rearranging, dropping negative terms from the right hand side and using that the minimum of a sequence is less than or equal to the average, for all $\rho \in \mathcal{P}(S)$ and $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\min_{0 \leq r \leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \\
&\leq \frac{1}{\lambda(1-\gamma)n} \left( \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) + \lambda \left( V^0(\rho) - V^*(\rho) \right) + \lambda c(\gamma) \sum_{k=0}^{n-1} |\theta^{k+1} - \theta_{\pi^k}|_2 \right).
\end{aligned}$$

which is exactly (6) and thus concludes the proof. □

# B Proofs of Section 3

## B.1 Proof of Lemma 3.3

*Proof.* By definition of the critic updates in Algorithm 1, for any $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
|\theta^{n+1}|_2^2 &= |\theta^n - h g(\theta^n, \pi^n)|_2^2 \\
&= |\theta^n|_2^2 + h^2 |g(\theta^n, \pi^n)|_2^2 - 2h \langle \theta^n, g(\theta^n, \pi^n) \rangle \\
&= |\theta^n|_2^2 + h^2 I_1^n - 2h I_2^n.
\end{aligned} \tag{8}$$

where we let $I_1^n = |g(\theta^n, \pi^n)|_2^2$ and $I_2^n = \langle \theta^n, g(\theta^n, \pi^n) \rangle$. Now let $\Gamma := \lambda_\beta (1-\gamma)(1-\sqrt{\gamma})$ and $\mathrm{K}_n = \sup_{s \in S} \mathrm{KL}(\pi^n(\cdot|s)|\mu)$. Lemma 5.2 of Zorba et al. (2025) shows that for any $\theta \in \mathbb{R}^N$ and $\pi \in \Pi_\mu$, it holds that

$$-\langle g(\theta, \pi), \theta \rangle \leq -\frac{\Gamma}{2} |\theta|_2^2 + \frac{\tau^2 \gamma^2}{\Gamma} \sup_{s \in S} \mathrm{KL}(\pi(\cdot|s)|\mu)^2 + \frac{|c|_{B_b(S \times A)}^2}{\Gamma}.$$

Therefore, for all $n \in \mathbb{N}$ we can upper bound $I_2^n$ as

$$-2h I_2^n = -2h \langle g(\theta^n, \pi^n), \theta^n \rangle \leq 2h \left( -\frac{\Gamma}{2} |\theta^n|_2^2 + \frac{\tau^2 \gamma^2}{\Gamma} \mathrm{K}_n^2 + \frac{|c|_{B_b(S \times A)}^2}{\Gamma} \right).$$

Moreover, using the identity $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, a direct calculation shows that

$$I_1^n = |g(\theta^n, \pi^n)|_2^2 \leq 3(1+\gamma)^2 |\theta^n|_2^2 + 3|c|_{B_b(S \times A)}^2 + 3\tau^2 \gamma^2 \mathrm{K}_n^2.$$

Substituting these upper bounds into (8) it holds that

$$|\theta^{n+1}|_2^2 \leq \left(1 + 3h^2(1+\gamma)^2 - h\Gamma\right)|\theta^n|_2^2 + \left(3h^2\tau^2\gamma^2 + \frac{2h\tau^2\gamma^2}{\Gamma}\right)\mathrm{K}_n^2 \tag{9}$$
$$+ |c|_{B_b(S\times A)}^2\left(3h^2 + \frac{2h}{\Gamma}\right).$$

Now we choose $h > 0$ such that

$$0 < \left(1 + 3h^2(1+\gamma)^2 - h\Gamma\right) < 1,$$

which is achieved when $0 < h < \frac{\Gamma}{3(1+\gamma)^2}$. Now let $\mathrm{M}_n := \sup_{0\leq r\leq n}\mathrm{K}_r$, iterating (9) using that $\sum_{n=0}^{\infty} a^n = \frac{1}{1-a}$ for $0 < a < 1$, it holds that

$$|\theta^{n+1}|_2^2 \leq |\theta_0|_2^2 + \frac{\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\mathrm{M}_n^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}.$$

This concludes the proof. □

### B.2 Proof of Theorem 3.4

*Proof.* Let $l_n(s,a) = \log\frac{d\pi^n}{d\mu}(s,a) - \int_A \log\frac{d\pi^n}{d\mu}(s,a')\mu(da')$. By Lemma E.1, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}|l_{n+1}|_{B_b(S\times A)} &\leq (1-\tau\lambda)|l_n|_{B_b(S\times A)} + 2\lambda\left|Q(\cdot,\cdot;\theta^{n+1})\right|_{B_b(S\times A)}\\ &\leq (1-\tau\lambda)|l_n|_{B_b(S\times A)} + 2\lambda|\theta^{n+1}|_2\end{aligned}$$

where we used that $\left|Q(\cdot,\cdot;\theta^{n+1})\right|_{B_b(S\times A)} \leq |\theta^{n+1}|_2$ by Hölder's inequality and Assumption 3.2. Moreover, by assumption we have $0 < \tau\lambda < 1$, thus iterating this recursion we have

$$\begin{aligned}|l_{n+1}|_{B_b(S\times A)} &\leq (1-\tau\lambda)^{n+1}|l_0|_{B_b(S\times A)} + 2\lambda\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}|\theta^{k+1}|_2.\\ &\leq C_1 + 2\lambda\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}|\theta^{k+1}|_2,\end{aligned}$$

where $C_1 > 0$ is such that $|l_0|_{B_b(S\times A)} \leq C_1$, which holds without loss of generality since $\pi_0 \in \Pi_\mu$. Now squaring both sides, using the identity $(a+b)^2 \leq 2a^2 + 2b^2$ and Hölder's inequality we have

$$\begin{aligned}|l_{n+1}|_{B_b(S\times A)}^2 &\leq C_1^2 + 4\lambda^2\left(\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}|\theta^{k+1}|_2\right)^2\\ &\leq C_1^2 + \frac{4\lambda}{\tau}\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}|\theta^{k+1}|_2^2.\end{aligned} \tag{10}$$

Now we turn to Lemma 3.3, where given that $0 < h < \frac{\Gamma}{3(1+\gamma)^2}$ for all $n \in \mathbb{N}$ it holds that

$$|\theta^{n+1}|_2^2 \leq |\theta_0|_2^2 + \frac{\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\mathrm{M}_n^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}.$$

where $\mathrm{M}_n := \sup_{0\leq r\leq n}\mathrm{K}_r$ and $\mathrm{K}_n := \sup_{s\in S}\mathrm{KL}(\pi^n(\cdot|s)|\mu)$. By Lemma A.3, $\mathrm{KL}(\pi^n(\cdot|s)|\mu) \leq 2|l_n|_{B_b(S\times A)}$ for all $s \in S$ and $n \in \mathbb{N}$. Thus it also holds that $\mathrm{M}_n^2 \leq 4\mathrm{L}_n^2$ for all $n \in \mathbb{N}$ and hence we have

$$|\theta^{n+1}|_2^2 \leq |\theta_0|_2^2 + \frac{4\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\mathrm{L}_n^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}.$$

Substituting this into (10) and using that $\sum_{m=0}^{n}(1-\tau\lambda)^m \leq \frac{1}{\tau\lambda}$ for any $n \in \mathbb{N}$, a direct calculation shows that

$$|l_{n+1}|_{B_b(S\times A)}^2 \leq \kappa(h)\mathrm{L}_n^2 + C_2(h)$$

such that

$$\kappa(h) := \frac{16\gamma^2(3h + \frac{2}{\Gamma})}{\Gamma - 3h(1+\gamma)^2}, \quad C_2(h) := C_1^2 + \frac{4}{\tau^2}\,|\theta_0|_2^2 + \frac{4\,|c|^2_{B_b(S\times A)}\left(3h + \frac{2}{\Gamma}\right)}{\tau^2\left(\Gamma - 3h(1+\gamma)^2\right)}.$$

A direct calculation shows that given that

$$0 < h < \min\left\{\frac{\Gamma}{3(1+\gamma)^2}\,,\ \frac{\Gamma^2 - 32\gamma^2}{\Gamma\big(48\gamma^2 + 3(1+\gamma)^2\big)}\right\},$$

then the conditions of Lemma 3.3 and $0 < \kappa(h) < 1$ both hold, provided that $\frac{32\gamma^2}{\Gamma^2} < 1$. Therefore, by Lemma A.4 for all $n \geq 0$ it holds that

$$\mathrm{L}_n^2 \leq \mathrm{L}_0^2 + \frac{C_2(h)}{1-\kappa(h)}.$$

Furthermore, by Lemma A.3 it holds that $\mathrm{KL}(\pi^n(\cdot|s)|\mu) \leq \mathrm{K}_n \leq 2\mathrm{L}_n$ for all $s \in S$, therefore we have

$$\mathrm{KL}(\pi^n(\cdot|s)|\mu)^2 \leq 4\mathrm{L}_0^2 + \frac{4C_2(h)}{1-\kappa(h)}.$$

Finally, we further restrict the critic step size by defining

$$\bar{h} := \frac{1}{2}\min\left\{\frac{\Gamma}{3(1+\gamma)^2}\,,\frac{\Gamma^2 - 32\gamma^2}{\Gamma\big(48\gamma^2 + 3(1+\gamma)^2\big)}\right\},$$

and restricting $0 < h \leq \bar{h}$. Since $0 < \kappa(h) < 1$ on $(0, \bar{h}]$, we have $1 - \kappa(h) \geq 1 - \kappa(\bar{h}) > 0$. Moreover, both $\kappa(h)$ and $C_2(h)$ are non-decreasing in $h$ on $(0, \bar{h}]$, hence

$$\kappa(h) \leq \kappa(\bar{h}), \qquad C_2(h) \leq C_2(\bar{h}), \qquad 0 < h \leq \bar{h}.$$

Therefore, for all $n \geq 0$ and all $s \in S$,

$$\mathrm{KL}(\pi^n(\cdot|s)|\mu)^2 \leq 4\mathrm{L}_0^2 + \frac{4C_2(h)}{1-\kappa(h)} \leq 4\mathrm{L}_0^2 + \frac{4C_2(\bar{h})}{1-\kappa(\bar{h})} := R,$$

where $R$ depends only on $\gamma, \tau, \pi_0, |\theta_0|_2, |c|_{B_b(S\times A)}, \lambda_\beta$ and is independent of $h > 0$.

□

### B.3 Proof of Theorem 3.6

Before we address the proof of Theorem 3.6 we must first establish the following regularity property of the policies produced by Algorithm 1.

**Lemma B.1.** *For some $\theta^0 = \theta_0 \in \mathbb{R}^N$ and $\pi^0 = \pi_0 \in \Pi_\mu$ let $\{\theta^n, \pi^n\}_{n\in\mathbb{N}}$ be the iterates of Algorithm 1. Let $0 < \lambda\tau < 1$. Then for all $n \in \mathbb{N}$ and $s \in S$, it holds that*

$$\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s)) \leq \frac{\lambda}{1-\lambda\tau}\left(2|\theta^{n+1}|_2 + \tau\,\mathrm{KL}(\pi^n(\cdot|s)|\mu)\right).$$

*Proof.* By definition, for each $s \in S$ the mirror descent updates can be applied pointwise through

$$\pi^{n+1}(\cdot|s) = \operatorname*{arg\,min}_{m\in P(A)}\left\{\int_A A(s,a;\theta^{n+1})(m(da) - \pi^n(da|s)) + \frac{1}{\lambda}\,\mathrm{KL}(m|\pi^n(\cdot|s))\right\}.$$

Therefore for all $s \in S$ it holds that

$$\int_A A(s,a;\theta^{n+1})(\pi^{n+1} - \pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s)) \leq 0.$$

Rearranging and using the definition of the approximate advantage function, for each $s \in S$ it holds that

$$\begin{aligned}
\frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s)) &\leq -\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s)\\
&= -\int_A \left(Q(s,a;\theta^{n+1}) + \tau \log\frac{d\pi^n}{d\mu}(s,a)\right)(\pi^{n+1}-\pi^n)(da|s)\\
&= -\int_A Q(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s)\\
&\quad -\tau\int_A \log\frac{d\pi^n}{d\mu}(s,a)\pi^{n+1}(da|s) + \tau\int_A \log\frac{d\pi^n}{d\mu}(s,a)\pi^n(da|s)\\
&\leq 2|\theta^{n+1}|_2 - \tau\int_A \log\left(\frac{d\pi^n}{d\pi^{n+1}}(s,a)\frac{d\pi^{n+1}}{d\mu}(s,a)\right)\pi^{n+1}(da|s)\\
&\quad +\tau\,\mathrm{KL}(\pi^n(\cdot|s)|\mu)\\
&= 2|\theta^{n+1}|_2 - \tau\int_A \left(\log\frac{d\pi^n}{d\pi^{n+1}}(s,a) + \log\frac{d\pi^{n+1}}{d\mu}(s,a)\right)\pi^{n+1}(da|s)\\
&\quad +\tau\,\mathrm{KL}(\pi^n(\cdot|s)|\mu)\\
&= 2|\theta^{n+1}|_2 + \tau\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s)) - \tau\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\mu)\\
&\quad +\tau\,\mathrm{KL}(\pi^n(\cdot|s)|\mu)\\
&\leq 2|\theta^{n+1}|_2 + \tau\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s)) + \tau\,\mathrm{KL}(\pi^n(\cdot|s)|\mu).
\end{aligned}$$

where we used the non-negativity of KL divergence in the final inequality. After rearranging we conclude the proof. □

We are now ready to prove Theorem 3.6.

*Proof.* Theorem A.9 states that for all $n \in \mathbb{N}$ and $\rho \in \mathcal{P}(S)$, we have

$$\begin{aligned}
&\min_{0\leq r\leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho)\\
&\leq \frac{1}{\lambda(1-\gamma)n}\left(\int_S \mathrm{KL}(\pi^*|\pi^0)(s)d_\rho^{\pi^*}(ds) + \lambda\left(V^0(\rho) - V^*(\rho)\right) + \lambda c(\gamma)\sum_{k=0}^{n-1}|\theta^{k+1}-\theta_{\pi^k}|_2\right).
\end{aligned}$$

where $c(\gamma) = \max\left\{1, \frac{2}{1-\gamma}\right\}$. Then by definition of the temporal difference updates in Algorithm 1, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
\left|\theta^{n+1} - \theta_{\pi^n}\right|_2^2 &= |\theta^n - hg(\theta^n,\pi^n) - \theta_{\pi^n}|_2^2\\
&= |\theta^n - \theta_{\pi^n}|_2^2 + h^2|g(\theta^n,\pi^n)|_2^2 - 2h\langle\theta^n - \theta_{\pi^n}, g(\theta^n,\pi^n)\rangle\\
&\leq |\theta^n - \theta_{\pi^n}|_2^2 + 2h^2(1+\gamma)|\theta^n - \theta_{\pi^n}|_2^2 - 2h\langle\theta^n - \theta_{\pi^n}, g(\theta^n,\pi^n)\rangle\\
&\leq |\theta^n - \theta_{\pi^n}|_2^2 + 2h^2(1+\gamma)|\theta^n - \theta_{\pi^n}|_2^2\\
&\qquad - 2h(1-\gamma)(1-\sqrt{\gamma})\langle\nabla_\theta L(\theta^n,\pi^n), \theta^n - \theta_{\pi^n}\rangle\\
&\leq |\theta^n - \theta_{\pi^n}|_2^2 + 2h^2(1+\gamma)|\theta^n - \theta_{\pi^n}|_2^2 - 2h\Gamma|\theta^n - \theta_{\pi^n}|_2^2\\
&= (1-h\Gamma)|\theta^n - \theta_{\pi^n}|_2^2 + (2(1+\gamma)h^2 - h\Gamma)|\theta^n - \theta_{\pi^n}|_2^2.
\end{aligned}$$

where we used Lemma E.3 in the first inequality, Lemma A.5 in the second inequality and the strong convexity of $\theta \mapsto L(\theta,\pi;\beta)$ in the final inequality. Now choosing $0 < h < \frac{\Gamma}{2(1+\gamma)}$ it holds that $2(1+\gamma)h^2 - h\Gamma \leq 0$ and thus we arrive at

$$\left|\theta^{n+1} - \theta_{\pi^n}\right|_2^2 \leq (1-h\Gamma)\left|\theta^n - \theta_{\pi^n}\right|_2^2.$$

Now we add and subtract $\theta_{\pi^{n-1}}$ and use Young's inequality with some $\epsilon > 0$ to arrive at

$$\left|\theta^{n+1} - \theta_{\pi^n}\right|_2^2 \leq (1+\epsilon)(1-h\Gamma)|\theta^n - \theta_{\pi^{n-1}}|_2^2 + \left(1+\frac{1}{\epsilon}\right)(1-h\Gamma)|\theta_{\pi^n} - \theta_{\pi^{n-1}}|_2^2.$$

Choosing $\epsilon > 0$ such that $(1+\epsilon)(1-h\Gamma) = 1 - \frac{h\Gamma}{2}$, we obtain $\epsilon = \frac{h\Gamma}{2(1-h\Gamma)}$ which also implies that $\left(1 + \frac{1}{\epsilon}\right) = \frac{2}{h\Gamma} - 1 > 0$. Hence for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\left|\theta^{n+1} - \theta_{\pi^n}\right|_2^2 &\leq \left(1 - \frac{h\Gamma}{2}\right)|\theta^n - \theta_{\pi^{n-1}}|_2^2 + \left(\frac{2}{h\Gamma} - 1\right)(1-h\Gamma)|\theta_{\pi^n} - \theta_{\pi^{n-1}}|_2^2 \qquad (11)\\ &\leq \left(1 - \frac{h\Gamma}{2}\right)|\theta^n - \theta_{\pi^{n-1}}|_2^2 + \frac{2}{h\Gamma}|\theta_{\pi^n} - \theta_{\pi^{n-1}}|_2^2.\end{aligned}$$

Now for each $n \in \mathbb{N}$, we seek to control the errors coming from the moving target $|\theta_{\pi^n} - \theta_{\pi^{n-1}}|_2^2$. To that end, Theorem E.2 shows that for some $\alpha_1 > 0$ independent of $h, \lambda > 0$, for all $n \in \mathbb{N}$ it holds that

$$\left|Q_\tau^{\pi^n} - Q_\tau^{\pi^{n-1}}\right|_{B_b(S\times A)}^2 \leq \alpha_1 \sup_{s\in S} \mathrm{KL}(\pi^n \mid \pi^{n-1})(s).$$

Then Lemma B.1 states that for each $s \in S$ and $n \in \mathbb{N}$ it holds that

$$\mathrm{KL}(\pi^n|\pi^{n-1})(s) \leq \frac{\lambda}{1-\tau\lambda}\left(2|\theta^n|_2 + \tau\,\mathrm{KL}(\pi^{n-1}(\cdot|s)|\mu)\right) \leq \frac{\lambda}{1-\tau\lambda}R,$$

where we used that the critic parameters are uniformly bounded by a constant $R > 0$, not depending on $h, \lambda > 0$, by the theorem statement. Finally observe that for each $n \in \mathbb{N}$ it holds that $|\theta_{\pi^n} - \theta_{\pi^{n-1}}|_2 \leq \frac{1}{\lambda_\beta}\left|Q_\tau^{\pi^n} - Q_\tau^{\pi^{n-1}}\right|_{B_b(S\times A)}$ where $\lambda_\beta > 0$ by Assumption 3.1. Therefore, it also holds that

$$|\theta_{\pi^n} - \theta_{\pi^{n-1}}|_2^2 \leq \frac{\alpha_1 R}{\lambda_\beta^2}\frac{\lambda}{1-\tau\lambda}. \qquad (12)$$

Let $\alpha_3 := \frac{\alpha_1 R}{\lambda_\beta^2}$. Substituting (12) into (11) it holds that

$$\begin{aligned}\left|\theta^{n+1} - \theta_{\pi^n}\right|_2^2 &\leq \left(1 - \frac{h\Gamma}{2}\right)^n \left|\theta^1 - \theta_{\pi^0}\right|_2^2 + \frac{2\alpha_3}{h\Gamma}\frac{\lambda}{1-\tau\lambda}\sum_{k=0}^{n-1}\left(1 - \frac{h\Gamma}{2}\right)^k\\ &= \left(1 - \frac{h\Gamma}{2}\right)^n \left|\theta^1 - \theta_{\pi^0}\right|_2^2 + \frac{2\alpha_3}{h\Gamma}\frac{\lambda}{1-\tau\lambda}\frac{1 - \left(1 - \frac{h\Gamma}{2}\right)^n}{\frac{h\Gamma}{2}}\\ &\leq \left(1 - \frac{h\Gamma}{2}\right)^n \left|\theta^1 - \theta_{\pi^0}\right|_2^2 + \frac{4\alpha_3}{h^2\Gamma^2}\frac{\lambda}{1-\tau\lambda}.\end{aligned}$$

Taking the average over $0 \leq k \leq n-1$, a direct calculation shows that

$$\frac{1}{n}\sum_{k=0}^{n-1}\left|\theta^{k+1} - \theta_{\pi^k}\right|_2^2 \leq \frac{2}{nh\Gamma}\left|\theta^1 - \theta_{\pi^0}\right|_2^2 + \frac{4\alpha_3}{h^2\Gamma^2}\frac{\lambda}{1-\tau\lambda}.$$

By Hölder's inequality and the fact that $0 < \lambda\tau < 1$, it also holds that

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{n-1}|\theta^{k+1} - \theta_{\pi^k}|_2 &\leq \left(\frac{1}{n}\sum_{k=0}^{n-1}|\theta^{k+1} - \theta_{\pi^k}|_2^2\right)^{\frac{1}{2}}\\ &\leq \left(\frac{2}{nh\Gamma}\left|\theta^1 - \theta_{\pi^0}\right|_2^2 + \frac{4\alpha_3}{h^2\Gamma^2}\frac{\lambda}{1-\tau\lambda}\right)^{\frac{1}{2}}\\ &\leq C\left(\frac{1}{nh} + \frac{\lambda}{h^2(1-\tau\lambda)}\right)^{\frac{1}{2}}\\ &\leq C\left(\frac{1}{n^{\frac{1}{2}}h^{\frac{1}{2}}} + \frac{\lambda^{\frac{1}{2}}}{h(1-\tau\lambda)^{\frac{1}{2}}}\right).\end{aligned}$$

where $C^2 := \max\left\{\frac{2}{\Gamma}\left|\theta^0 - \theta_{\pi^0}\right|_2^2, \frac{4\alpha_3}{\Gamma^2}\right\}$. Substituting this into the bound from Theorem A.9, we conclude with

$$\begin{aligned}
&\min_{0\le r\le n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \\
&\le \frac{1}{\lambda(1-\gamma)n}\Bigg(\int_S \mathrm{KL}(\pi^*|\pi^0)(s)d_\rho^{\pi^*}(ds) \\
&\qquad + \lambda\left(V^0(\rho) - V^*(\rho)\right) + C\lambda c(\gamma)\left(\frac{n^{\frac{1}{2}}}{h^{\frac{1}{2}}} + \frac{n\lambda^{\frac{1}{2}}}{h(1-\tau\lambda)^{\frac{1}{2}}}\right)\Bigg),
\end{aligned}$$

which concludes the proof.

□

# C Proofs of Section 4

## C.1 Proof of Theorem 4.1

*Proof.* By the performance difference lemma, for all $s \in S$ and $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&V_\tau^{\pi^{n+1}}(s) - V_\tau^{\pi^n}(s) \\
&= \frac{1}{1-\gamma}\int_S\left[\int_A A_\tau^{\pi^n}(s',a)(\pi^{n+1}-\pi^n)(da|s') + \tau\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s')\right]d_s^{\pi^{n+1}}(ds') \\
&\le \frac{1}{1-\gamma}\int_S\left[\int_A A_\tau^{\pi^n}(s',a)(\pi^{n+1}-\pi^n)(da|s') + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s')\right]d_s^{\pi^{n+1}}(ds') \\
&= \frac{1}{1-\gamma}\int_S\left[\int_A A(s',a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s') + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s')\right]d_s^{\pi^{n+1}}(ds') \\
&\quad + \frac{1}{1-\gamma}\int_S\int_A\left(A_\tau^{\pi^n}(s',a) - A(s',a;\theta^{n+1})\right)(\pi^{n+1}-\pi^n)(da|s')d_s^{\pi^{n+1}}(ds') \\
&\le \frac{1}{1-\gamma}\int_S\left[\int_A A(s',a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s') + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s')\right]d_s^{\pi^{n+1}}(ds') \quad (13) \\
&\quad + \frac{2}{(1-\gamma)}|\theta^{n+1} - \theta_{\pi^n}|_2.
\end{aligned}$$

where we used that $\tau \le \frac{1}{\lambda}$ in the first inequality, added and subtracted the approximate advantage function in the second equality, and used Hölder's inequality with $Q_\tau^\pi$-realisability and Assumption 3.2. In particular, for all $s \in S$ and $a \in A$,

$$A(s,a;\theta^{n+1}) - A_\tau^{\pi^n}(s,a) \le \left|\theta^{n+1} - \theta_{\pi^n}\right|_2.$$

Now recall the policy mirror descent update in Algorithm 2

$$\pi^{n+1}(\cdot|s) = \operatorname*{arg\,min}_{m\in\mathcal{P}(A)}\left\{\int_A A(s,a;\theta^{n+1})(m(da) - \pi^n(da|s)) + \frac{1}{\lambda}\,\mathrm{KL}(m|\pi^n(\cdot|s))\right\}. \quad (14)$$

Since we obtain the minimum at $\pi^{n+1}$, for any $s \in S$ and $\pi(\cdot|s) \in \mathcal{P}(A)$ it holds that

$$\begin{aligned}
&\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \\
&\qquad \le \int_A A(s,a;\theta^{n+1})(\pi-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi|\pi^n)(s).
\end{aligned}$$

Choosing $\pi = \pi^n$ yields

$$\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \le 0.$$

Substituting this into (13) and using the definition of $\pi^* \in \Pi_\mu$, it hence holds that

$$V_\tau^{\pi^*}(s) \leq V_\tau^{\pi^{n+1}}(s) \leq V_\tau^{\pi^n}(s) + \frac{2}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|_2$$

for all $s \in S$. Finally, applying Lemma E.4 we arrive at

$$V_\tau^{\pi^*}(s) \leq V_\tau^{\pi^{n+1}}(s) \leq V_\tau^{\pi^n}(s) + \frac{2e^{-\frac{M(n)h\Gamma}{2}}}{1-\gamma}|\theta^{n} - \theta_{\pi^n}|_2.$$

This concludes the proof.

□

### C.2 Proof of Theorem 4.2

*Proof.* Recall that for each $n \in \mathbb{N}$ and $s \in S$, $a \in A$, the normalised log densities are defined as $l_n(s,a) = \log \frac{d\pi^n}{d\mu}(s,a) - \int_A \log \frac{d\pi^n}{d\mu}(s,a')\mu(da')$. Then by Lemma E.1, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
|l_{n+1}|_{B_b(S\times A)} &\leq (1-\tau\lambda)|l_n|_{B_b(S\times A)} + 2\lambda \left|Q(\cdot,\cdot;\theta^{n+1})\right|_{B_b(S\times A)} \\
&\leq (1-\tau\lambda)|l_n|_{B_b(S\times A)} + 2\lambda \left(\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)} + \left|Q(\cdot,\cdot;\theta^{n+1}) - Q_\tau^{\pi^n}\right|_{B_b(S\times A)}\right) \\
&\leq (1-\tau\lambda)|l_n|_{B_b(S\times A)} + 2\lambda \left(\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)} + |\theta^{n+1} - \theta_{\pi^n}|_2\right) \\
&\leq (1-\tau\lambda)\mathrm{L}_n + 2\lambda \left(\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)} + e^{-\frac{M(n)h\Gamma}{2}}|\theta^n - \theta_{\pi^n}|_2\right). \qquad (15)
\end{aligned}$$

where we added and subtracted $Q_\tau^{\pi^n}$ in the second inequality and used Hölder's inequality along with Assumption 3.2 in the third inequality. In the final inequality we used Theorem E.4 and used that by definition $|l_n|_{B_b(S\times A)} \leq \sup_{0\leq r\leq n}|l_r|_{B_b(S\times A)} := \mathrm{L}_n$. Now to upper bound $Q_\tau^{\pi^n} \in B_b(S\times A)$, by definition it holds that

$$Q_\tau^{\pi^n}(s,a) = c(s,a) + \gamma \int_S V_\tau^{\pi^n}(s')P(ds'|s,a),$$

for all $s \in S$ and $a \in A$. Taking the $|\cdot|_{B_b(S\times A)}$ and rearranging, it holds that

$$\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)} \leq |c|_{B_b(S\times A)} + \gamma \left|V_\tau^{\pi^n}\right|_{B_b(S)}. \qquad (16)$$

Now by Theorem 4.1, for all $n \in \mathbb{N}$ we have

$$V_\tau^{\pi^*}(s) \leq V_\tau^{\pi^{n+1}}(s) \leq V_\tau^{\pi^n}(s) + \frac{2e^{-\frac{M(n)h\Gamma}{2}}}{1-\gamma}|\theta^{n} - \theta_{\pi^n}|_2. \qquad (17)$$

Recursively applying the upper bound (17), for all $s \in S$ and $n \in \mathbb{N}$ it holds that

$$V_\tau^{\pi^*}(s) \leq V_\tau^{\pi^n}(s) \leq V_\tau^{\pi^0}(s) + \frac{2}{1-\gamma}\sum_{k=0}^{n-1} e^{-\frac{M(k)h\Gamma}{2}}|\theta^k - \theta_{\pi^k}|_2.$$

Taking the $|\cdot|_{B_b(S)}$ norm on both sides it also holds that

$$\begin{aligned}
\left|V_\tau^{\pi^n}\right|_{B_b(S)} &\leq \max\left\{\left|V_\tau^{\pi^0}\right|_{B_b(S)} + \frac{2}{1-\gamma}\sum_{k=0}^{n-1} e^{-\frac{M(k)h\Gamma}{2}}|\theta^k - \theta_{\pi^k}|_2, \left|V_\tau^{\pi^*}\right|_{B_b(S)}\right\} \\
&\leq \left|V_\tau^{\pi^0}\right|_{B_b(S)} + \left|V_\tau^{\pi^*}\right|_{B_b(S)} + \frac{2}{1-\gamma}\sum_{k=0}^{n-1} e^{-\frac{M(k)h\Gamma}{2}}|\theta^k - \theta_{\pi^k}|_2.
\end{aligned}$$

Substituting this into (16), for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)} &\leq |c|_{B_b(S\times A)} + \gamma \left|V_\tau^{\pi^n}\right|_{B_b(S)} \\ &\leq |c|_{B_b(S\times A)} + \gamma \left|V_\tau^{\pi^0}\right|_{B_b(S)} + \gamma \left|V_\tau^{\pi^*}\right|_{B_b(S)} + \frac{2\gamma}{1-\gamma}\sum_{k=0}^{n-1} e^{-\frac{M(k)h\Gamma}{2}}|\theta^k - \theta_{\pi^k}|_2.\end{aligned}$$

Let $\alpha_1 := |c|_{B_b(S\times A)} + \gamma \left|V_\tau^{\pi^0}\right|_{B_b(S)} + \gamma \left|V_\tau^{\pi^*}\right|_{B_b(S)}$ and $c(\gamma) = \max\left\{1, \frac{2\gamma}{1-\gamma}\right\}$. Substituting into (15), we have

$$|l_{n+1}|_{B_b(S\times A)} \leq (1-\tau\lambda)\mathrm{L}_n + 2\lambda\left(c(\gamma)\sum_{k=0}^{n} e^{-\frac{M(k)h\Gamma}{2}}|\theta^k - \theta_{\pi^k}|_2 + \alpha_1\right). \tag{18}$$

By the triangle inequality it holds that $|\theta^k - \theta_{\pi^k}|_2 \leq |\theta^k|_2 + |\theta_{\pi^k}|_2$ for each $k \in \mathbb{N}$. Then by Lemma 3.3 for all $k \in \mathbb{N}$ it holds that

$$\begin{aligned}|\theta^k|_2^2 &\leq |\theta_0|_2^2 + \frac{\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\sup_{0\leq r\leq k}\mathrm{K}_r^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h+\frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2} \\ &\leq |\theta_0|_2^2 + \frac{4\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\mathrm{L}_k^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h+\frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2} \\ &\leq |\theta_0|_2^2 + \frac{4\tau^2\gamma^2\left(3h + \frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\mathrm{L}_n^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h+\frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2},\end{aligned}$$

where we used Lemma A.3 in the second inequality and the fact that $\mathrm{L}_k \leq \mathrm{L}_n$ for all $k \leq n$. Similarly by Lemma A.6 it holds that

$$\begin{aligned}|\theta_{\pi^k}|_2 &\leq \frac{|c|_{B_b(S\times A)}}{(1-\gamma)\lambda_\beta} + \frac{2\tau\gamma}{(1-\gamma)\lambda_\beta}\mathrm{L}_k \\ &\leq \frac{|c|_{B_b(S\times A)}}{(1-\gamma)\lambda_\beta} + \frac{2\tau\gamma}{(1-\gamma)\lambda_\beta}\mathrm{L}_n,\end{aligned}$$

where we again used that $\mathrm{L}_k \leq \mathrm{L}_n$ for $k \leq n$. Therefore for all $k \in \{0, \ldots, n\}$ it holds that

$$|\theta^k - \theta_{\pi^k}|_2 \leq (\delta_1 + \delta_2\,\mathrm{L}_n), \tag{19}$$

with

$$\begin{aligned}\delta_1 &= \left(|\theta_0|_2^2 + \frac{|c|_{B_b(S\times A)}^2\left(3h+\frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\right)^{\frac{1}{2}} + \frac{|c|_{B_b(S\times A)}}{(1-\gamma)\lambda_\beta}, \\ \delta_2 &= 2\tau\gamma\left(\frac{\left(3h+\frac{2}{\Gamma}\right)}{\Gamma - 3h(1+\gamma)^2}\right)^{\frac{1}{2}} + \frac{2\tau\gamma}{(1-\gamma)\lambda_\beta}.\end{aligned}$$

Hence substituting (19) into (18), we arrive at

$$\begin{aligned}|l_{n+1}|_{B_b(S\times A)} &\leq (1-\tau\lambda)\mathrm{L}_n + 2\lambda\left(c(\gamma)\sum_{k=0}^{n} e^{-\frac{M(k)h\Gamma}{2}}|\theta^k - \theta_{\pi^k}|_2 + \alpha_1\right) \\ &\leq \beta_1(n)\mathrm{L}_n + \beta_2(n),\end{aligned} \tag{20}$$

where

$$\beta_1(n) := (1-\tau\lambda) + 2\lambda c(\gamma)\delta_2\sum_{k=0}^{n} e^{-\frac{M(k)h\Gamma}{2}},$$

$$\beta_2(n) := 2\lambda c(\gamma)\delta_1 \sum_{k=0}^{n} e^{-\frac{M(k)h\Gamma}{2}} + 2\lambda\alpha_1.$$

For all $n \in \mathbb{N}$ we choose $M(n)$ large enough such that

$$\beta_1(n) \leq 1 - \frac{\tau\lambda}{2}.$$

A direct calculation shows that this is satisfied when

$$\sum_{k=0}^{n} e^{-\frac{M(k)h\Gamma}{2}} \leq \frac{\tau}{4c(\gamma)\delta_2}.$$

Therefore let $c_0 > 0$ satisfy $c_0^2 \geq \frac{8c(\gamma)\delta_2}{\tau}$, and let $c \geq c_0$. If $M(n) \geq \frac{4}{h\Gamma}\log(c(n+1))$, then for all $n \in \mathbb{N}$

$$\sum_{k=0}^{n} e^{-\frac{M(k)h\Gamma}{2}} \leq \frac{1}{c^2}\sum_{k=0}^{n}\frac{1}{(k+1)^2} \leq \frac{2}{c^2} \leq \frac{\tau}{4c(\gamma)\delta_2}.$$

Therefore, with this choice of $M(n)$ for all $n \in \mathbb{N}$, (20) becomes

$$\begin{aligned} |l_{n+1}|_{B_b(S\times A)} &\leq \beta_1(n)\mathrm{L}_n + \beta_2(n) \\ &\leq \left(1 - \frac{\tau\lambda}{2}\right)\mathrm{L}_n + \alpha_3. \end{aligned}$$

A direct application of Lemma A.4 yields that for all $n \in \mathbb{N}$

$$\begin{aligned} \mathrm{L}_n &\leq \mathrm{L}_0 + \frac{\alpha_3}{1 - \left(1 - \frac{\tau\lambda}{2}\right)} \\ &= \mathrm{L}_0 + \frac{2\alpha_3}{\tau\lambda}. \end{aligned}$$

Finally, after applying Lemma A.3 and Lemma A.6, we conclude the proof. □

# D Proofs of Section 4.2

## D.1 Proof of Theorem 4.3

*Proof.* Theorem A.9 states that for all $n \in \mathbb{N}$ and $\rho \in \mathcal{P}(S)$, we have

$$\begin{aligned} &\min_{0\leq r\leq n-1} V_\tau^{\pi^r}(\rho) - V_\tau^{\pi^*}(\rho) \\ &\leq \frac{1}{\lambda(1-\gamma)n}\left(\int_S \mathrm{KL}(\pi^*|\pi^0)(s)d_\rho^{\pi^*}(ds) + \lambda\left(V^0(\rho) - V^*(\rho)\right) + \lambda c(\gamma)\sum_{k=0}^{n-1}|\theta^{k+1} - \theta_{\pi^k}|_2\right). \end{aligned}$$

with $c(\gamma) = \max\left\{1, \frac{2}{1-\gamma}\right\}$. We are thus left with controlling the average of the critic errors. To that end, by Theorem E.4, for any $n \in \mathbb{N}$ it holds that

$$\left|\theta^{n+1} - \theta_{\pi^n}\right|_2 \leq e^{-\frac{M(n)h\Gamma}{2}}\left|\theta^n - \theta_{\pi^n}\right|_2.$$

Now by assumption we have that for each $n \in \mathbb{N}$, $M := M(n) \geq \frac{4}{h\Gamma}\log\big(c(n+1)\big)$ with $c \geq c_0$ as in Theorem 4.2. Thus, by Theorem 4.2 and Lemma A.6, there exists $R > 0$ such that $\left|\theta^k - \theta_{\pi^k}\right|_2 \leq R$ for all $k \geq 0$. Therefore for any $n \in \mathbb{N}$ we have

$$\left|\theta^{n+1} - \theta_{\pi^n}\right|_2 \leq e^{-\frac{M(n)h\Gamma}{2}}\left|\theta^n - \theta_{\pi^n}\right|_2 \leq Re^{-\frac{M(n)h\Gamma}{2}}.$$

For any $n \in \mathbb{N}$, this implies that

$$\sum_{k=0}^{n-1}\left|\theta^{k+1} - \theta_{\pi^k}\right|_2 \leq R\sum_{k=0}^{n-1}\frac{1}{c^2(k+1)^2} = \frac{R}{c^2}\sum_{j=1}^{n}\frac{1}{j^2}.$$

Now using the classical result $\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{\pi^2}{6} \leq 2$, we obtain

$$\sum_{k=0}^{n-1}\left|\theta^{k+1} - \theta_{\pi^k}\right|_2 \leq \frac{2R}{c^2}.$$

Substituting this into the bound from Theorem A.9, we conclude the proof. □

### D.2 Proof of Theorem 4.4

*Proof.* To ease notation let $V^n := V_\tau^{\pi^n}$ for $n \in \mathbb{N}$ and let $V^* := V_\tau^{\pi^*}$. Note that $\frac{\mathrm{d}d_\rho^\pi}{\mathrm{d}\rho}(s) \geq (1-\gamma)$ for any $s \in S$ and $\pi \in \Pi_\mu$. Hence it holds that

$$\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}\rho}(s) = \frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^\pi}(s)\frac{\mathrm{d}d_\rho^\pi}{\mathrm{d}\rho}(s) \geq (1-\gamma)\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^\pi}(s).$$

for any $\pi \in \Pi_\mu$ and $s \in S$. Thus we have $\nu^n := \left|\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^{\pi^n}}\right|_{B_b(S)} \leq \xi$ for all $n \in \mathbb{N}$. Moreover, recall that by (14) for all $s \in S$ it holds that

$$\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s)) \leq 0.$$

Therefore, it holds that

$$\begin{aligned}
&\int_S \left(\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) d_\rho^{\pi^*}(ds)\\
&= \int_S \left(\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) \frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^{\pi^{n+1}}}(s) d_\rho^{\pi^{n+1}}(ds)\\
&\geq \left|\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^{\pi^{n+1}}}\right|_{B_b(S)} \int_S \left(\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \frac{1}{\lambda}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) d_\rho^{\pi^{n+1}}(ds)\\
&\geq \nu^{n+1} \int_S \left(\int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) + \tau\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) d_\rho^{\pi^{n+1}}(ds)\\
&= \nu^{n+1} \int_S \left(\int_A A_\tau^{\pi^n}(s,a)(\pi^{n+1}-\pi^n)(da|s) + \tau\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) d_\rho^{\pi^{n+1}}(ds)\\
&\quad + \nu^{n+1} \int_S \int_A \left(A(s,a;\theta^{n+1}) - A_\tau^{\pi^n}(s,a)\right)(\pi^{n+1}-\pi^n)(da|s) d_\rho^{\pi^{n+1}}(ds)\\
&= \nu^{n+1}(1-\gamma)\left(V^{n+1}(\rho) - V^n(\rho)\right)\\
&\quad + \nu^{n+1} \int_S \int_A \left(A(s,a;\theta^{n+1}) - A_\tau^{\pi^n}(s,a)\right)(\pi^{n+1}-\pi^n)(da|s) d_\rho^{\pi^{n+1}}(ds),
\end{aligned}$$

where the first inequality is due to the integrand in the $s \in S$ variable being non-positive, the second inequality is due to $0 \leq \tau\lambda \leq 1$ and the final equality is then an application of performance difference. To ease notation, let $\int_S \mathrm{KL}(\pi^*(\cdot|s)|\pi^n(\cdot|s))\, d_\rho^{\pi^*}(ds) := \mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right)$. Now identically to (7) from the proof of Theorem A.9, after an application of the Three Point Lemma, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\frac{1}{\lambda}\left(\mathrm{KL}(\pi^*|\pi^{n+1})\left(d_\rho^{\pi^*}\right) - \mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right)\right) \qquad (21)\\
&\leq \int_S \int_A A(s,a;\theta^{n+1})(\pi^*-\pi^n)(da|s) d_\rho^{\pi^*}(ds)\\
&\quad - \left(\int_S \int_A A(s,a;\theta^{n+1})(\pi^{n+1}-\pi^n)(da|s) d_\rho^{\pi^*}(ds) + \tau\,\mathrm{KL}(\pi^{n+1}|\pi^n)\left(d_\rho^{\pi^*}\right)\right).
\end{aligned}$$

To upper bound the first term on the right hand side, observe that

$$\begin{aligned}
&\int_{S\times A} A(s,a;\theta^{n+1})(\pi^*-\pi^n)(da|s) d_\rho^{\pi^*}(ds)\\
&= \int_{S\times A} A_\tau^{\pi^n}(s,a)(\pi^*-\pi^n)(da|s) d_\rho^{\pi^*}(ds)\\
&\quad + \int_{S\times A} \left(A(s,a;\theta^{n+1}) - A_\tau^{\pi^n}(s,a)\right)(\pi^*-\pi^n)(da|s) d_\rho^{\pi^*}(ds)\\
&= (1-\gamma)(V^*(\rho) - V^n(\rho)) - \tau\,\mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right)\\
&\quad + \int_{S\times A} \left(A(s,a;\theta^{n+1}) - A_\tau^{\pi^n}(s,a)\right)(\pi^*-\pi^n)(da|s) d_\rho^{\pi^*}(ds)
\end{aligned}$$

where we added and subtracted the true advantage function and applied the performance difference lemma on the first term in the second equality. Now observe that

$$\begin{aligned}
&\int_{S\times A} A(s,a;\theta^{n+1})(\pi^* - \pi^n)(da|s)d_\rho^{\pi^*}(ds) \\
&= (1-\gamma)(V^*(\rho) - V^n(\rho)) - \tau\,\mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right) \\
&\qquad + \int_{S\times A}\left(A(s,a;\theta^{n+1}) - A_\tau^{\pi^n}(s,a)\right)(\pi^* - \pi^n)(da|s)\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^{\pi^{n+1}}}(s)d_\rho^{\pi^{n+1}}(ds) \\
&\leq (1-\gamma)(V^*(\rho) - V^n(\rho)) - \tau\,\mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right) \\
&\qquad + 2\left|\frac{\mathrm{d}d_\rho^{\pi^*}}{\mathrm{d}d_\rho^{\pi^{n+1}}}\right|_{B_b(S)}|\theta^{n+1} - \theta_{\pi^n}|_2 \\
&= (1-\gamma)(V^*(\rho) - V^n(\rho)) - \tau\,\mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right) \\
&\qquad + 2\nu^{n+1}|\theta^{n+1} - \theta_{\pi^n}|_2,
\end{aligned}$$

where we performed a change of measure on the second term in the first equality and used Hölder's inequality along with Assumption 3.2 in the final inequality. Substituting this into (21), for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\frac{1}{\lambda}\left(\mathrm{KL}(\pi^*|\pi^{n+1})\left(d_\rho^{\pi^*}\right) - \mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right)\right) \\
&\leq (1-\gamma)(V^*(\rho) - V^n(\rho)) + (1-\gamma)\nu^{n+1}(V^{n+1}(\rho) - V^n(\rho)) - \tau\,\mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right) \\
&\qquad + 4\nu^{n+1}|\theta^{n+1} - \theta_{\pi^n}|_2.
\end{aligned}$$

Now let $y^n = \mathrm{KL}(\pi^*|\pi^n)\left(d_\rho^{\pi^*}\right)$ and $\delta^n = V^n(\rho) - V^*(\rho)$. Rearranging, a direct calculation shows that for any $n \in \mathbb{N}$

$$\delta^n - \nu^{n+1}\left(\delta^{n+1} - \delta^n - \frac{4}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|_2\right) \leq \frac{1}{\lambda(1-\gamma)}\left((1-\tau\lambda)y^n - y^{n+1}\right).$$

Now recall that by Theorem 4.1, for all $n \in \mathbb{N}$ it holds that

$$\delta^{n+1} - \delta^n = V^{n+1}(\rho) - V^n(\rho) \leq \frac{2}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|.$$

Therefore it also holds that $\delta^{n+1} - \delta^n - \frac{4}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|_2 \leq 0$. Hence using that $\nu^n \leq \xi$ for all $n \in \mathbb{N}$, we have

$$\delta^n + \xi\left(\delta^{n+1} - \delta^n - \frac{4}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|_2\right) \leq \frac{1}{\lambda(1-\gamma)}\left((1-\tau\lambda)y^n - y^{n+1}\right).$$

Thus dividing through by $\xi > 0$ and rearranging again, for all $n \in \mathbb{N}$ it holds that

$$\delta^{n+1} + \frac{1}{\lambda(1-\gamma)\xi}y^{n+1} \leq \frac{\xi-1}{\xi}\left(\delta^n + \frac{1-\tau\lambda}{\lambda(1-\gamma)(\xi-1)}y^n\right) + \frac{4}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|_2. \quad (22)$$

Moreover, by assumption we have that $\xi \geq 2$ and $\tau\lambda \geq \frac{1}{\xi}$, which imply that $\frac{1-\tau\lambda}{\lambda(1-\gamma)(\xi-1)} \leq \frac{1}{\lambda(1-\gamma)\xi}$. Therefore after simplifying (22), for all $n \in \mathbb{N}$ it holds that

$$\delta^{n+1} + \frac{1}{\lambda(1-\gamma)\xi}y^{n+1} \leq \frac{\xi-1}{\xi}\left(\delta^n + \frac{1}{\lambda(1-\gamma)\xi}y^n\right) + \frac{4}{1-\gamma}|\theta^{n+1} - \theta_{\pi^n}|_2.$$

Iterating this inequality, we have

$$\begin{aligned}
&\delta^{n+1} + \frac{1}{\lambda(1-\gamma)\xi}y^{n+1} \\
&\leq \left(\frac{\xi-1}{\xi}\right)^{n+1}\left(\delta^0 + \frac{1}{\lambda(1-\gamma)\xi}y^0\right) + \frac{4}{1-\gamma}\sum_{k=0}^{n}\left(\frac{\xi-1}{\xi}\right)^{n-k}|\theta^{k+1} - \theta_{\pi^k}|_2 \\
&\leq e^{-\frac{n+1}{\xi}}\left(\delta^0 + \frac{1}{\lambda(1-\gamma)\xi}y^0\right) + \frac{4}{1-\gamma}\sum_{k=0}^{n}\left(\frac{\xi-1}{\xi}\right)^{n-k}|\theta^{k+1} - \theta_{\pi^k}|_2, \qquad (23)
\end{aligned}$$

where we used the identity $(1-x)^M \leq e^{-xM}$ for any $x \in (0,1)$ in the final inequality. To ease notation, let $\kappa = \frac{\xi-1}{\xi} \in (0,1)$. Focusing on the error term on the right hand side, by assumption we have that for each policy update $n \in \mathbb{N}$, the number of temporal difference steps satisfies $M := M(n) \geq \frac{4c}{h\Gamma}(n+1)$, where $c > 0$ is chosen as in Theorem 4.4. Therefore by Theorem 4.2 and Lemma A.6, there exists $R > 0$ such that $|\theta^k - \theta_{\pi^k}|_2 \leq R$ for all $k \in \mathbb{N}$. Hence for all $n \in \mathbb{N}$ we have

$$\sum_{k=0}^{n} \kappa^{n-k} |\theta^{k+1} - \theta_{\pi^k}|_2 \leq R \sum_{k=0}^{n} \kappa^{n-k} e^{-\frac{M(k)h\Gamma}{2}} \leq R \sum_{k=0}^{n} \kappa^{n-k} e^{-2c(k+1)}.$$

Simplifying, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
\sum_{k=0}^{n} \kappa^{n-k} e^{-2c(k+1)} &= e^{-2c(n+1)} \sum_{k=0}^{n} \kappa^{n-k} e^{2c(n-k)} \\
&= e^{-2c(n+1)} \sum_{m=0}^{n} \left(\kappa e^{2c}\right)^m \\
&= e^{-2c(n+1)} \left( \frac{\left(\kappa e^{2c}\right)^{n+1} - 1}{\kappa e^{2c} - 1} \right),
\end{aligned}$$

where we used the property of the geometric series $\sum_{m=0}^{n} a^m = \frac{a^{n+1}-1}{a-1}$ for any positive real number $a \neq 1$ and for all $n \in \mathbb{N}$. Therefore we have

$$\begin{aligned}
\sum_{k=0}^{n} \kappa^{n-k} |\theta^{k+1} - \theta_{\pi^k}|_2 &\leq R \sum_{k=0}^{n} \kappa^{n-k} e^{-2c(k+1)} \\
&\leq R e^{-2c(n+1)} \left( \frac{\left(\kappa e^{2c}\right)^{n+1} - 1}{\kappa e^{2c} - 1} \right) \\
&\leq R e^{-2c(n+1)} \left( \frac{\left(\kappa e^{2c}\right)^{n+1} + 1}{|\kappa e^{2c} - 1|} \right) \\
&\leq R \frac{\kappa^{n+1} + e^{-2c(n+1)}}{|\kappa e^{2c} - 1|} \\
&\leq \frac{R}{|\kappa e^{2c} - 1|} \left( e^{-\frac{n+1}{\xi}} + e^{-2c(n+1)} \right) \\
&\leq \frac{2R}{|\kappa e^{2c} - 1|} e^{-\min\left\{\frac{1}{\xi}, 2c\right\}(n+1)},
\end{aligned}$$

where in the final inequality we used the standard identity $(1-x)^M \leq e^{-xM}$ for any $x \in (0,1)$. Substituting this into (23) we obtain

$$\begin{aligned}
\delta^{n+1} + \frac{1}{\lambda(1-\gamma)\xi} y^{n+1} &\leq e^{-\frac{n+1}{\xi}} \left( \delta^0 + \frac{1}{\lambda(1-\gamma)\xi} y^0 \right) + \frac{8R}{(1-\gamma)|\kappa e^{2c} - 1|} e^{-\min\left\{\frac{1}{\xi}, 2c\right\}(n+1)} \\
&\leq e^{-\min\left\{\frac{1}{\xi}, 2c\right\}(n+1)} \left( \delta^0 + \frac{1}{\lambda(1-\gamma)\xi} y^0 + \frac{8R}{(1-\gamma)|\kappa e^{2c} - 1|} \right).
\end{aligned}$$

This concludes the proof.

□

# E Additional Results

**Lemma E.1.** *Let Assumption 3.2 hold. Let* $l_n(s,a) = \log \frac{d\pi^n}{d\mu}(s,a) - \int_A \log \frac{d\pi^n}{d\mu}(s,a')\mu(da')$. *Then for all* $n \in \mathbb{N}$ *it holds that*

$$|l_{n+1}|_{B_b(S\times A)} \leq (1 - \tau\lambda)|l_n|_{B_b(S\times A)} + 2\lambda|\theta^{n+1}|_2.$$

*Proof.* To ease notation, throughout the proof we drop the dependence on the policy in the approximate advantage function, i.e., $A(s,a;\theta^{n+1},\pi^n) := A(s,a;\theta^{n+1})$. Then by the policy updates in Algorithm 1, it holds that

$$\pi^{n+1}(\cdot|s) = \underset{m\in\mathcal{P}(A)}{\arg\min}\left\{\int_A A(s,a;\theta^{n+1})(m(da)-\pi^n(da|s)) + \frac{1}{\lambda}\,\mathrm{KL}(m|\pi^n(\cdot|s))\right\}$$

for each $s \in S$. Lemma 1.4.3 of Dupuis and Ellis (1997) shows that for each $s \in S$ the minimiser is achieved at

$$\log\frac{d\pi^{n+1}}{d\mu}(s,a) = \log\frac{d\pi^n}{d\mu}(s,a) - \lambda A(s,a;\theta^{n+1}) - \log Z_n(s)$$

such that $Z_n(s) = \int_A \exp\left(-\lambda A(s,a';\theta^{n+1})\right)\pi^n(da'|s)$. Thus it holds that

$$\begin{aligned} l_{n+1}(s,a) &= \log\frac{d\pi^n}{d\mu}(s,a) - \lambda A(s,a;\theta^{n+1}) - \log Z_n(s) \\ &\quad - \left(\int_A\left(\log\frac{d\pi^n}{d\mu}(s,a') - \lambda A(s,a';\theta^{n+1}) - \log Z_n(s)\right)\mu(da')\right). \end{aligned}$$

Cancelling out the normalisation constants $Z_n(s)$ for each $s \in S$ (as they are independent of $a \in A$) and using the definition of the approximate advantage function (3), we have

$$l_{n+1}(s,a) = l_n(s,a) - \lambda\left(Q(s,a;\theta^{n+1}) - \int_A Q(s,a';\theta^{n+1})\mu(da') + \tau l_n(s,a)\right).$$

After grouping like terms we arrive at

$$l_{n+1}(s,a) = (1-\tau\lambda)l_n(s,a) - \lambda\left(Q(s,a;\theta^{n+1}) - \int_A Q(s,a';\theta^{n+1})\mu(da')\right).$$

Taking the $|\cdot|_{B_b(S\times A)}$ norm on both sides and using the triangle inequality on the right hand side yields

$$|l_{n+1}|_{B_b(S\times A)} \le (1-\tau\lambda)\,|l_n|_{B_b(S\times A)} + 2\lambda\left|Q(\cdot,\cdot;\theta^{n+1})\right|_{B_b(S\times A)}.$$

□

**Theorem E.2.** *Suppose that there exists $R \ge 0$ such that $\left|\log\frac{d\pi^n}{d\mu}\right|_{B_b(S\times A)} \le R$ for all $n \in \mathbb{N}$. Then there exists $\alpha_1 > 0$ such that for all $n \in \mathbb{N}$ and $s \in S$ it holds that*

$$\left|Q_\tau^{\pi^{n+1}} - Q_\tau^{\pi^n}\right|^2_{B_b(S\times A)} \le \alpha_1 \sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s).$$

*Proof.* Let $l_n(s,a) = \log\frac{d\pi^n}{d\mu}(s,a) - \int_A \log\frac{d\pi^n}{d\mu}(s,a')\mu(da')$. Recall that $Q_\tau^\pi \in B_b(S\times A)$ is a fixed point of the Bellman operator defined in (1) for all $\pi \in \Pi_\mu$. Hence it holds that

$$\begin{aligned} &Q_\tau^\pi(s,a) - Q_\tau^{\pi'}(s,a) \\ &= \gamma\int_{S\times A} Q_\tau^\pi(s',a')\pi(da'|s')P(ds'|s,a) - \gamma\int_{S\times A} Q_\tau^{\pi'}(s',a')\pi'(da'|s')P(ds'|s,a) \\ &\quad + \tau\gamma\left(\int_S\left(\mathrm{KL}(\pi|\mu)(s') - \mathrm{KL}(\pi'|\mu)(s')\right)P(ds'|s,a)\right). \end{aligned}$$

Now add and subtract $\gamma\int_{S\times A} Q_\tau^{\pi'}(s',a')\pi(da'|s')P(ds'|s,a)$ on the right hand side for each $s \in S$ and $a \in A$ to arrive at

$$\begin{aligned} &Q_\tau^\pi(s,a) - Q_\tau^{\pi'}(s,a) \\ &= \gamma\int_{S\times A}\left(Q_\tau^\pi(s',a') - Q_\tau^{\pi'}(s',a')\right)\pi(da'|s')P(ds'|s,a) \\ &\quad + \gamma\int_{S\times A} Q_\tau^{\pi'}(s',a')(\pi-\pi')(da'|s')P(ds'|s,a) \\ &\quad + \tau\gamma\left(\int_S\left(\mathrm{KL}(\pi|\mu)(s') - \mathrm{KL}(\pi'|\mu)(s')\right)P(ds'|s,a)\right). \end{aligned}$$

Rearranging and taking the $|\cdot|_{B_b(S\times A)}$ norm on both sides and using the triangle inequality we have

$$\begin{aligned}&\left|Q_\tau^\pi - Q_\tau^{\pi'}\right|_{B_b(S\times A)} \\ &\leq \frac{\gamma}{1-\gamma}\left|Q_\tau^{\pi'}\right|_{B_b(S\times A)} \sup_{s\in S}|\pi(\cdot|s)-\pi'(\cdot|s)|_{\mathcal{M}(A)} + \frac{\tau\gamma}{1-\gamma}\left|\mathrm{KL}(\pi|\mu)-\mathrm{KL}(\pi'|\mu)\right|_{B_b(S)}\end{aligned} \tag{24}$$

We shift our focus to the last term on the right hand side. By definition for any $\pi \in \Pi_\mu$ it holds that

$$\begin{aligned}&\left|\mathrm{KL}(\pi|\mu)-\mathrm{KL}(\pi'|\mu)\right|_{B_b(S)} \\ &= \left|\int_A \log\frac{d\pi}{d\mu}(\cdot,a)\pi(da|\cdot) - \int_A \log\frac{d\pi'}{d\mu}(\cdot,a)\pi'(da|\cdot)\right|_{B_b(S)} \\ &= \left|\int_A \log\left(\frac{d\pi}{d\pi'}(\cdot,a)\frac{d\pi'}{d\mu}(\cdot,a)\right)\pi(da|\cdot) - \int_A \log\frac{d\pi'}{d\mu}(\cdot,a)\pi'(da|\cdot)\right|_{B_b(S)} \\ &= \left|\int_A \log\frac{d\pi}{d\pi'}(\cdot,a)\pi(da|\cdot) + \int_A \log\frac{d\pi'}{d\mu}(\cdot,a)\pi(da|\cdot) - \int_A \log\frac{d\pi'}{d\mu}(\cdot,a)\pi'(da|\cdot)\right|_{B_b(S)} \\ &\leq \sup_{s\in S}\mathrm{KL}(\pi|\pi')(s) + \left|\int_A \log\frac{d\pi'}{d\mu}(\cdot,a)\left(\pi-\pi'\right)(da|\cdot)\right|_{B_b(S)} \\ &= \sup_{s\in S}\mathrm{KL}(\pi|\pi')(s) + \left|\int_A \left(\log\frac{d\pi'}{d\mu}(\cdot,a) - \int_A \log\frac{d\pi'}{d\mu}(\cdot,a)\mu(da)\right)\left(\pi-\pi'\right)(da|\cdot)\right|_{B_b(S)}.\end{aligned}$$

Now let $\pi = \pi^{n+1}$ and $\pi' = \pi^n$. By Pinsker's inequality, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}&\left|\mathrm{KL}(\pi^{n+1}|\mu)-\mathrm{KL}(\pi^n|\mu)\right|_{B_b(S)} \\ &\leq \sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s) + |l_n|_{B_b(S\times A)}\sup_{s\in S}\left|\pi^{n+1}(\cdot|s)-\pi^n(\cdot|s)\right|_{\mathcal{M}(A)} \\ &\leq \sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s) + \sqrt{2}\,|l_n|_{B_b(S\times A)}\sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s)^{\frac{1}{2}}.\end{aligned}$$

Substituting into (24) and using Pinsker's inequality once again we arrive at

$$\begin{aligned}&\left|Q_\tau^{\pi^{n+1}} - Q_\tau^{\pi^n}\right|_{B_b(S\times A)} \\ &\leq \sqrt{2}\frac{\gamma}{1-\gamma}\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)}\sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s)^{\frac{1}{2}} \\ &\quad + \frac{\tau\gamma}{1-\gamma}\sup_{s\in S}\left(\mathrm{KL}(\pi^{n+1}|\pi^n)(s) + \sqrt{2}\,|l_n|_{B_b(S\times A)}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)^{\frac{1}{2}}\right).\end{aligned}$$

After grouping terms we have

$$\begin{aligned}&\left|Q_\tau^{\pi^{n+1}} - Q_\tau^{\pi^n}\right|_{B_b(S\times A)} \\ &\leq \left(\frac{\sqrt{2}\tau\gamma\,|l_n|_{B_b(S\times A)}}{1-\gamma} + \sqrt{2}\frac{\gamma\left|Q_\tau^{\pi^n}\right|_{B_b(S\times A)}}{1-\gamma}\right)\sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s)^{\frac{1}{2}} \\ &\quad + \frac{\tau\gamma}{1-\gamma}\sup_{s\in S}\mathrm{KL}(\pi^{n+1}|\pi^n)(s).\end{aligned}$$

By assumption it holds that $|l_n|_{B_b(S\times A)} \leq 2R$ for all $n \in \mathbb{N}$, thus it also holds that $\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \leq 2R$ for each $s \in S$ and

$$\mathrm{KL}(\pi^{n+1}|\pi^n)(s) \leq (2R)^{\frac{1}{2}}\,\mathrm{KL}(\pi^{n+1}|\pi^n)(s)^{\frac{1}{2}}.$$

Substituting and squaring, together with Lemma A.6, gives the claimed bound. □

**Lemma E.3.** *Let Assumptions 3.1 and 3.2, and $Q_\tau^\pi$-realisability hold. Then for all $\theta \in \mathbb{R}^N$ and $\pi \in \Pi_\mu$ it holds that*

$$|g(\theta,\pi)|_2^2 \leq 2(1+\gamma)\,|\theta-\theta_\pi|_2^2$$

*Proof.* First observe that, by adding and subtracting $Q_\tau^\pi \in B_b(S \times A)$ and using that $\mathrm{T}_\tau^\pi Q_\tau^\pi = Q_\tau^\pi$ for all $\pi \in \Pi_\mu$ where $\mathrm{T}_\tau^\pi : B_b(S \times A) \to B_b(S \times A)$ is the Bellman operator defined in (1), we can express the semi-gradient as

$$\begin{aligned} g(\theta, \pi) &= \int_{S\times A} \left(Q(s,a;\theta) - \mathrm{T}_\tau^\pi Q(s,a;\theta)\right) \phi(s,a) d_\beta^\pi(ds, da) \\ &= \int_{S\times A} \left(Q(s,a;\theta) - Q_\tau^\pi(s,a) - (\mathrm{T}_\tau^\pi Q(s,a;\theta) - \mathrm{T}_\tau^\pi Q_\tau^\pi(s,a))\right) \phi(s,a) d_\beta^\pi(ds, da) \end{aligned}$$

for all $\theta \in \mathbb{R}^N$ and $\pi \in \Pi_\mu$. To ease notation, for each $s \in S$ and $a \in A$ we define $\varepsilon(s,a) = Q(s,a;\theta) - Q_\tau^\pi(s,a)$. Expanding the Bellman operator, a direct calculation shows that

$$g(\theta, \pi) = \int_{S\times A} \left(\varepsilon(s,a) - \gamma \int_{S\times A} \varepsilon(s',a') P^\pi(ds', da'|s,a)\right) \phi(s,a) d_\beta^\pi(ds, da).$$

Moreover, by using Young's and Hölder's inequalities along with Assumption 3.2, it holds that

$$\begin{aligned} |g(\theta,\pi)|_2^2 &\le \int_{S\times A} \left(\varepsilon(s,a) - \gamma \int_{S\times A} \varepsilon(s',a') P^\pi(ds', da'|s,a)\right)^2 d_\beta^\pi(ds, da) \\ &\le 2 \int_{S\times A} \varepsilon(s,a)^2 d_\beta^\pi(da, ds) + 2\gamma^2 \int_{S\times A\times S\times A} \varepsilon(s',a')^2 P^\pi(ds', da'|s,a) d_\beta^\pi(ds, da) \\ &= 2 \int_{S\times A} \varepsilon(s,a)^2 d_\beta^\pi(da, ds) + 2\gamma^2 \int_{S\times A} \varepsilon(s',a')^2 d_{J_\pi \beta}^\pi(ds', da'), \end{aligned}$$

where $J_\pi : \mathcal{P}(S \times A) \to \mathcal{P}(S \times A)$ is the one-step transition operator defined in (5). By Lemma 5.1 of Zorba et al. (2025), it holds that $d_{J_\pi\beta}^\pi(E) \le \frac{1}{\gamma} d_\beta^\pi(E)$ for all $E \in \mathcal{B}(S \times A)$. Therefore using the non-negativity of the integrand, for all $\theta \in \mathbb{R}^N$ and $\pi \in \Pi_\mu$ it holds that

$$\begin{aligned} |g(\theta,\pi)|_2^2 &\le 2 \int_{S\times A} \varepsilon(s,a)^2 d_\beta^\pi(da, ds) + 2\gamma \int_{S\times A} \varepsilon(s,a)^2 d_\beta^\pi(ds, da) \\ &= 2\,(1+\gamma) \int_{S\times A} \varepsilon(s,a)^2 d_\beta^\pi(da, ds) \\ &\le 2(1+\gamma) |\theta - \theta_\pi|_2^2, \end{aligned}$$

where the last inequality follows from the fact that for all $s \in S$ and $a \in A$ we have $\varepsilon^2(s,a) = \left(Q(s,a;\theta) - Q_\tau^\pi(s,a)\right)^2 = \left(\langle \theta - \theta_\pi, \phi(s,a)\rangle\right)^2 \le |\theta - \theta_\pi|_2^2 |\phi(s,a)|_2^2$ from Hölder's inequality, $Q_\tau^\pi$-realisability and Assumption 3.2.

□

**Theorem E.4.** *Let Assumptions 3.1 and 3.2, and $Q_\tau^\pi$-realisability hold. Let $0 < h < \min\left\{\frac{\Gamma}{2(1+\gamma)}, \frac{1}{\Gamma}\right\}$. Then for all $n \in \mathbb{N}$ it holds that*

$$|\theta^{n+1} - \theta_{\pi^n}|_2^2 \le e^{-M(n)h\Gamma} |\theta^n - \theta_{\pi^n}|_2^2.$$

*Proof.* Fix $n \in \mathbb{N}$ and let $M := M(n)$ to ease notation. By definition of the critic updates in Algorithm 2, for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} |\theta^{n+1} - \theta_{\pi^n}|_2^2 &= |\theta^{n,M} - \theta_{\pi^n}|_2^2 \\ &= |\theta^{n,M-1} - h g(\theta^{n,M-1}, \pi^n) - \theta_{\pi^n}|_2^2 \\ &= |\theta^{n,M-1} - \theta_{\pi^n}|_2^2 + h^2 \left|g(\theta^{n,M-1}, \pi^n)\right|_2^2 \\ &\quad - 2h\langle \theta^{n,M-1} - \theta_{\pi^n}, g(\theta^{n,M-1}, \pi^n)\rangle. \end{aligned}$$

By Lemma E.3, it holds that $|g(\theta,\pi)|_2^2 \le 2(1+\gamma)\,|\theta - \theta_\pi|_2^2$ for all $\pi \in \Pi_\mu$ and $\theta \in \mathbb{R}^N$. Hence it holds that

$$\begin{aligned} |\theta^{n+1} - \theta_{\pi^n}|_2^2 \le |\theta^{n,M-1} - \theta_{\pi^n}|_2^2 + 2h^2(1+\gamma)|\theta^{n,M-1} - \theta_{\pi^n}|_2^2 \\ - 2h\langle \theta^{n,M-1} - \theta_{\pi^n}, g(\theta^{n,M-1}, \pi^n)\rangle. \end{aligned}$$

Using Lemma A.5 on the final term we arrive at

$$|\theta^{n,M} - \theta_{\pi^n}|_2^2 \leq |\theta^{n,M-1} - \theta_{\pi^n}|_2^2 + 2h^2(1+\gamma)|\theta^{n,M-1} - \theta_{\pi^n}|_2^2 \\ - 2h(1-\gamma)(1-\sqrt{\gamma})\langle \nabla_\theta L(\theta^{n,M-1}, \pi^n; \beta), \theta^{n,M-1} - \theta_{\pi^n}\rangle.$$

We now proceed by using standard tools from convex analysis. That is, using the $\lambda_\beta$-strong convexity of $L(\cdot, \pi; \beta)$ and $Q_\tau^\pi$-realisability, we have

$$\begin{aligned} |\theta^{n,M} - \theta_{\pi^n}|_2^2 &\leq |\theta^{n,M-1} - \theta_{\pi^n}|_2^2 + 2h^2(1+\gamma)|\theta^{n,M-1} - \theta_{\pi^n}|_2^2 \\ &\quad - 2h(1-\gamma)(1-\sqrt{\gamma})\lambda_\beta|\theta^{n,M-1} - \theta_{\pi^n}|_2^2 \\ &= \left(1 - 2h\Gamma + 2(1+\gamma)h^2\right)|\theta^{n,M-1} - \theta_{\pi^n}|_2^2. \end{aligned}$$

with $\Gamma = \lambda_\beta(1-\gamma)(1-\sqrt{\gamma})$. Choosing the critic step size $h$ such that

$$0 < h < \min\left\{\frac{\Gamma}{2(1+\gamma)}, \frac{1}{\Gamma}\right\},$$

it holds that $2(1+\gamma)h^2 \leq h\Gamma$ and $h\Gamma \in (0,1)$, therefore

$$|\theta^{n,M} - \theta_{\pi^n}|_2^2 \leq (1 - h\Gamma)|\theta^{n,M-1} - \theta_{\pi^n}|_2^2.$$

For each fixed $n \in \mathbb{N}$, iterating over the $M > 0$ critic steps yields

$$|\theta^{n,M} - \theta_{\pi^n}|_2^2 \leq (1 - h\Gamma)^M|\theta^n - \theta_{\pi^n}|_2^2.$$

We conclude by using the standard identity $(1-x)^M \leq e^{-xM}$ for $x \in (0,1)$ and $M > 0$ to arrive at

$$|\theta^{n+1} - \theta_{\pi^n}|_2^2 \leq e^{-Mh\Gamma}|\theta^n - \theta_{\pi^n}|_2^2.$$

□

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract, introduction, and contribution list state that the paper provides population-level theory for KL-regularised actor-critic methods in general spaces. The assumptions and limitations are discussed in the assumptions discussion and conclusion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses realisability, feature covariance, population-level updates, exact actor optimisation, sampling, and function approximation error in the assumptions discussion and conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: Each theorem states its assumptions, and the proofs are provided in the appendix. Additional discussion explains the role and restrictiveness of the main assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A].

Justification: The paper is theory-only and does not include experiments or numerical results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A].

   Justification: The paper does not include experiments, datasets, or experimental code. The algorithms and theoretical schedules are specified in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A].

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments or empirical estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments. The computational cost of the theoretical TD-step schedules is discussed in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical, uses no human subjects or private data, and does not release models or datasets.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A].

    Justification: This is foundational theory with no direct deployment, dataset, or model release. No specific societal impact pathway is claimed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper releases no data, models, or high-risk assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A].

    Justification: The paper does not use existing datasets, code assets, or models; it only cites prior literature.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce or release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method and theoretical analysis do not use LLMs as an important, original, or non-standard component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.