AN ANALYSIS OF MIRROR-DESCENT SOFT ACTOR-CRITIC

Anonymous authors
Paper under double-blind review

ABSTRACT

Soft Actor-Critic (SAC) is widely used for entropy-regularised reinforcement learning with continuous action spaces, and practical implementations perform only a few actor steps towards an evolving target. In this work, we prove convergence guarantees when the target policy arises from policy mirror descent and compare it with the classical Gibbs target. We derive sufficient conditions for the strong convexity and smoothness of the actor objective, characterised by the curvature of the Q-function estimate through the Legendre differential operator, and establish an $\mathcal{O}\left(N^{-\frac{1}{5}}\right)$ best-iterate finite-time convergence rate up to actor and critic approximation errors. Moreover, the mirror-descent step size λ directly controls the target drift and hence actor tracking error, whereas the analogous Gibbs bound contains a non-vanishing tracking term.

1 INTRODUCTION

In reinforcement learning (RL), an agent seeks to learn an optimal policy that minimises its expected cumulative cost through interactions with an environment. This framework has led to remarkable successes across a range of challenging domains, including robotics, autonomous control and games (Levine et al., 2016; Lillicrap et al., 2016; Mnih et al., 2015). Broadly speaking, methods to tackle RL problems may be divided into two families: model-based approaches, which explicitly learn or exploit a model of the environment dynamics (Sutton & Barto, 2018; Levine & Koltun, 2013), and model-free approaches, which seek to learn a policy or value function directly from observed transitions (Sutton & Barto, 2018; Lillicrap et al., 2016; Haarnoja et al., 2018a).

In continuous action spaces, tackling model-free RL problems typically involves two learning problems. Firstly, equipped with function approximations, a critic (an estimate of the advantage or Q-function) is learned from observed transitions, commonly through some variant of temporal-difference learning (Sutton, 1988; Watkins & Dayan, 1992), which is then in turn used to update the actor (the policy). Such actor updates may be performed directly through policy-gradient methods (Williams, 1992; Sutton et al., 1999), or through proximal methods which improve stability by controlling the change between successive policies (Schulman et al., 2015; 2017). A closely related idea arises from policy mirror descent (Lan, 2023; Zhan et al., 2023; Ju & Lan, 2026), where a KL-divergence penalty is introduced directly into the policy-improvement problem, with the resulting update admitting a closed-form policy. However, in continuous action spaces, this target generally involves an intractable normalisation constant (Haarnoja et al., 2017; 2018a).

Entropy regularisation leads to a number of theoretical and practical benefits (Geist et al., 2019; Cayci et al., 2024a; Leahy et al., 2022). For example, entropy regularisation induces a unique statewise optimal policy under standard assumptions such as bounded costs (Geist et al., 2019; Kerimkulov et al., 2026) and can accelerate the convergence of policy gradient methods (Mei et al., 2020; Cen et al., 2022; Lan, 2023). It also encourages stochastic policies and persistent exploration (Haarnoja et al., 2017; 2018a).

The Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018b) was introduced as an efficient approach to tackling large-scale continuous model-free RL tasks with entropy regularisation, initially derived as a practical implementation of soft policy iteration (Haarnoja et al., 2018a). During SAC training, one performs gradient descent steps on a reverse Kullback–Leibler (KL) divergence with respect to some target policy, typically the exponential of the Q-function which in turn is motivated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by the Bellman principle. Crucially, the normalisation constant is independent of the actor parameters and therefore disappears when taking gradients of the actor objective. This gives a practical policy-improvement procedure without requiring the intractable normalisation constant.

Despite the practical success of SAC, its theoretical analysis remains sparse and challenging. In practice, the actor is commonly parametrised by a squashed Gaussian distribution with a parametrised mean, that is a Gaussian followed by a $\tanh$ transformation to enforce the bounds of the action space (after scaling). Moreover, the reverse-KL projection onto the target policy is not solved exactly, but rather a single or a few steps of gradient descent are performed (Haarnoja et al., 2018a). Such an objective is in general non-convex, and the $\tanh$ transformation produces policy log-densities which become unbounded near the boundary of the action space. Therefore regularity hypotheses used in policy-gradient analyses can fail (Bedi et al., 2022), while policy-space convergence results (Lan, 2023; Kerimkulov et al., 2026) do not directly cover such a framework.

In this work, we provide the first theoretical guarantees for a population-level formalisation of the SAC framework, with a fixed-covariance squashed Gaussian actor, linear mean and a critic oracle, where one treats a policy arising from mirror descent as the target, as well as the classical Gibbs target. We consider the setting where only a single projected gradient step is performed towards an evolving target policy. Moreover we identify conditions on the curvature of the critic under which the reverse-KL actor objective becomes strongly convex and smooth, allowing us to control the error incurred by the actor as it tracks the changing target.

2 Related works

There is a tremendous amount of research literature on RL theory, which underscores its importance. Our work relates to policy optimisation in entropy-regularised MDPs and to the analysis of actor-critic algorithms, which we summarise here.

Entropy regularised RL. Encouraging explorative policies during the training of RL policies have been demonstrated to be an important foundation of practical deep RL methods (Haarnoja et al., 2018a; Lan, 2023) which in turn establishes the importance of understanding entropy regularisation in MDPs. A particular line of work where the presence of entropy regularisation has been shown to have a meaningful contribution are policy mirror descent methods (Lan, 2023; Ju & Lan, 2026; Kerimkulov et al., 2026), where it was proven that the entropy regularisation can induce an exponential convergence to the optimal policy when the critic is known. More recently, it has been shown that one can maintain this exponential convergence when the critic is solved to sufficient accuracy in both the tabular case (Labbi et al., 2026) and the general case (Zorba et al., 2026).

Having said that, such theoretical analyses typically rely on the updates existing in the policy space and using measure theoretic techniques. Moving from policy-space updates to a practical parametrisation, a line of work (Sherman et al., 2025; Alfano et al., 2023; Yuan et al., 2023; Agarwal et al., 2021) approximates mirror-descent updates by solving the minimisation sub-optimally under various assumptions.

Actor-critic analysis. Early convergence analyses use multi-timescale stochastic approximation and ODE based approaches (Borkar & Konda, 1997; Konda & Tsitsiklis, 1999). Since then, the study of different regimes of actor critic algorithms can fall into a few categories. Firstly, a line of work studies the sample complexity of certain actor critic regimes in tabular and function-approximation settings under various assumptions in the unregularised case (Olshevsky & Gharesifard, 2023; Kumar et al., 2025; Gaur et al., 2024) and regularised case (Cayci et al., 2024b; Labbi et al., 2026). Another line of work studies the underlying convergence and stability behaviour of actor critic methods in general state and action spaces by focusing on a population level analysis (Zorba et al., 2026; Hong et al., 2023; Kerimkulov et al., 2026; Leahy et al., 2022).

3 Contributions

- We formalise the SAC policy class as the push-forward under $\tanh$ of a Gaussian distribution with parametrised mean, as standard in modern implementations of Soft Actor Critic (Haarnoja et al., 2018a;b). We study the reverse-KL objective for the actor parameters with target policies including the classical Gibbs measure (Haarnoja et al., 2018a) and a policy

arising from performing a step of policy mirror descent with step size/ penalty λ (Tomar et al., 2021; Lan, 2023). Note that in general, such a loss function is non-convex. In this work, we identify sufficient conditions under which the strong convexity of such an objective is guaranteed on any Euclidean ball of radius $\mathrm{R} > 0$. We demonstrate that the curvature of the critic, quantified through the Legendre differential operator, plays a particular role alongside the entropy regularisation parameter $\tau > 0$.

- We prove global finite-time convergence guarantees (up to approximation errors) under a $\varepsilon_{\mathrm{crit}}$-optimal critic oracle assumption. We derive bounds on the actor tracking error for both the classical Gibbs target and the mirror-descent target. We show that the mirror descent step size λ can directly control the tracking error and provides an explicit trade off between balancing policy improvement against the actor's ability to track the target. In contrast, the corresponding bound for the classical Gibbs target contains a non-vanishing contribution determined by the regularisation parameter τ.

4 Entropy Regularised MDPs

Consider an infinite horizon Markov Decision Process (S, A, P, c, γ), where S is the state space (which may be an arbitrary Polish space) and A is the action space, which we set as $A := (-1, 1)^M$ for a fixed action dimension $M \geq 1$. Moreover, $P \in \mathcal{P}(S|S \times A)$ is the state transition probability kernel, c is a bounded cost function and $\gamma \in (0, 1)$ is the discount factor. Let $\Lambda_{\mathbb{R}}$ denote Lebesgue measure on $\mathbb{R}$, let $\Lambda_{\mathbb{R}^M} := \Lambda_{\mathbb{R}}^{\otimes M}$, and let Λ denote normalised Lebesgue measure on A, that is $\Lambda(B) := 2^{-M}\Lambda_{\mathbb{R}^M}(B)$ for $B \in \mathcal{B}(A)$. Finally, let $\tau > 0$ denote a regularisation parameter. For each stochastic policy $\pi \in \mathcal{P}(A|S)$ and $s \in S$, we define the regularised value function by

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

where $\mathbb{E}_s^\pi$ denotes the expectation over the state-action trajectory $(s_0, a_0, s_1, a_1, \ldots)$ generated by policy π and kernel P such that $s_0 := s,\ a_n \sim \pi(\cdot \mid s_n), s_{n+1} \sim P(\cdot|s_n, a_n)$ for all $n \geq 0$ and for any $\nu \in \mathcal{P}(A)$, the entropy is defined as

$$\mathrm{H}(\nu) := -\int_A \log \frac{d\nu}{d\Lambda}(a)\nu(da).$$

if ν is absolutely continuous with respect to the Lebesgue measure and $-\infty$ otherwise. As a result, for any $s \in S$ we have that $V_\tau^\pi(s) \in \mathbb{R} \cup \{+\infty\}$. Moreover for any state distribution $\rho \in \mathcal{P}(S)$, we denote $V_\tau^\pi(\rho) := \int_S V_\tau^\pi(s)\rho(ds)$. The optimal value function is defined as

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

We refer to π^* as the optimal policy if $V_\tau^*(\rho) = V_\tau^{\pi^*}(\rho)$. For each π, the state-action value function Q_τ^π is defined as

$$Q_\tau^\pi(s, a) = c(s, a) + \gamma \int_S V_\tau^\pi(s')P(ds'|s, a). \tag{1}$$

Then for every $\pi \in \mathcal{P}(A|S)$ the advantage function is defined as $A_\tau^\pi = Q_\tau^\pi + \tau \log \frac{d\pi}{d\Lambda} - V_\tau^\pi$. The occupation measure for policy π with initial distribution ρ is denoted by d_ρ^π and is defined in Appendix B, alongside the Bellman principle for entropy-regularised MDPs.

5 Soft Actor Critic and Mirror Descent

In the seminal work of Haarnoja et al. (2018a), a framework for tackling off-policy entropy regularised MDPs was introduced, which has since been regarded as one of the state of the art approaches for entropy regularised MDPs with continuous action spaces. The idea is relatively simple and consists of learning a critic through some variant of temporal difference/ Q-learning and then learning the new actor by projecting the exponential of the critic onto your policy class using the Kullback-Leibler (KL) divergence. That is, given some policy class $\Pi \subset \mathcal{P}(A|S)$, for each $s \in S$ compute

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} \mathrm{KL}\left(\pi'(\cdot|s) \middle| \frac{1}{Z^{\pi_{\text{old}}}(s)} \exp\left(-\frac{1}{\tau} Q^{\pi_{\text{old}}}(s, \cdot) \right) \right). \tag{2}$$

One can then establish that (2) guarantees policy improvement when the minimisation is solved fully, at least in the tabular setting (Haarnoja et al., 2018a) and thus with exact critics, iterations of (2) results in a sequence of policies which converges to an optimal policy inside Π, although it is worth mentioning that in most cases the true optimal policy π^* need not belong in Π. Moreover, such guarantees only hold in the case where the optimisation in (2) is solved to full accuracy and when the action space is finite.

One issue that can arise from such a formulation is that the new policy π_{new} may be drastically different from the old policy π_{old}, resulting in a large relative entropy $\mathrm{KL}(\pi_{\text{new}}(\cdot|s)|\pi_{\text{old}}(\cdot|s))$. Policy mirror descent provides a solution to this by penalising such policies. Formally, for some $\lambda > 0$, the policies are updated by

$$\pi_{\mathrm{MD}}(\cdot|s) = \operatorname*{arg\,min}_{m\in\mathcal{P}(A)} \left\{ \int_A \left(Q_\tau^{\pi_{\text{old}}}(s,a) + \tau \log \frac{d\pi_{\text{old}}}{d\Lambda}(s,a) \right) m(da) + \frac{1}{\lambda} \mathrm{KL}\left(m \mid \pi_{\text{old}}(\cdot|s)\right) \right\}.$$

Suppose that the initial policy is uniform on A. By Dupuis & Ellis (1997), we have the following closed form solution to the optimisation problem

$$\frac{d\pi_{\mathrm{MD}}}{d\Lambda}(s,a) = \frac{1}{Z_{\mathrm{MD}}(s)} \left(\frac{d\pi_{\text{old}}}{d\Lambda}(s,a) \right)^{1-\tau\lambda} \exp\left(-\lambda Q_\tau^{\pi_{\text{old}}}(s,a)\right). \tag{3}$$

From here, it is clear to see that the policy mirror descent policy behaves as a geometric interpolation between the old policy and the full soft policy-improvement update, controlled by the proximal penalty/ step size $\lambda > 0$ and the entropy regularisation $\tau > 0$. In particular, they coincide when precisely we have $\lambda = \frac{1}{\tau}$, giving the soft policy-improvement update of (Haarnoja et al., 2018a).

Given that computing the normalisation constant in (3) is just as intractable as (2), one natural extension is to replace the target in (2) with (3), since the normalisation constant does not affect the gradient steps. This approach has demonstrated promising empirical results (Tomar et al., 2021).

6 Policy Class

Let us now introduce the policy class before presenting the algorithm. Recall that $A = (-1,1)^M$ and $\mathfrak{B}(\mathrm{R}) = \left\{w \in \mathbb{R}^d : |w|_2 \le \mathrm{R}\right\}$ is the Euclidean ball of radius R in $\mathbb{R}^d$. Let $x : S \to \mathbb{R}^{d\times M}$ be a bounded feature map satisfying $|x(s)|_{\text{op}} \le 1$ for all $s \in S$, where $|\cdot|_{\text{op}}$ is the Euclidean operator norm (see Appendix A for an overview of notation). For any $\varsigma > 0$, we denote the centred Gaussian with covariance $\varsigma^2 I_M$ by $q^\varsigma := \mathcal{N}(0, \varsigma^2 I_M)$. Let $\Sigma := \mathrm{diag}(\sigma_1^2, \ldots, \sigma_M^2)$ with $\sigma_i > 0$ for all i. For any $w \in \mathfrak{B}(\mathrm{R}) \subset \mathbb{R}^d$, we parametrise the mean by $m_w(s) = x(s)^\top w \in \mathbb{R}^M$ for each $s \in S$ and denote $q_w(\cdot|s) := \mathcal{N}(m_w(s), \Sigma)$. We denote this class of Gaussian kernels by $\mathcal{Q}_\Sigma := \{q_w : w \in \mathfrak{B}(\mathrm{R})\} \subset \mathcal{P}(\mathbb{R}^M|S)$. Throughout, $\tanh$ and its inverse act component-wise. To formalise the policy class, we must first introduce the notion of a push forward measure.

Definition 1 (Push-forward measure, Ambrosio et al., 2008, §5.2)**.** *Let $T : \mathbb{R}^M \to A$ be measurable and let $\nu \in \mathcal{P}(\mathbb{R}^M)$. The push-forward of ν by T is the probability measure $T_\#\nu \in \mathcal{P}(A)$ such that for any $B \in \mathcal{B}(A)$,*

$$T_\#\nu(B) := \nu(T^{-1}(B)).$$

Equivalently, for every bounded measurable function $f : A \to \mathbb{R}$,

$$\int_A f(a)\, T_\#\nu(da) = \int_{\mathbb{R}^M} f(T(u))\, \nu(du).$$

The policy class is then simply the push-forward under $\tanh$ of $\mathcal{Q}_\Sigma$.

Definition 2 (Squashed Gaussian Policies)**.** *The admissible policy class is given by*

$$\Pi_\Sigma := \left\{\pi_w = (\tanh)_\# q_w : q_w \in \mathcal{Q}_\Sigma\right\}.$$

The following lemma of Dupuis-Ellis becomes very useful in the proofs of the main results.

Lemma 1 (Dupuis & Ellis, 1997, Lemma E.2.1)**.** *Let $T : \mathbb{R}^M \to A$ be a measurable bijection with measurable inverse. Then for any $\nu, \eta \in \mathcal{P}(\mathbb{R}^M)$,*

$$\mathrm{KL}\left(T_\#\nu | T_\#\eta\right) = \mathrm{KL}(\nu|\eta).$$

Note that in general the log densities $\log \frac{d\pi_w}{d\Lambda}$ are unbounded near the boundary of the action space $A = (-1, 1)^M$. Therefore, one cannot directly apply classical tools such as the performance difference lemma for entropy regularised MDPs, which require bounded log densities, without careful treatment. However one can show that although the log densities are unbounded, the entropy $\mathrm{H}(\pi_w(\cdot|s))$ remains bounded for all $s \in S$ when $w \in \mathfrak{B}(\mathrm{R})$, which can serve as a remedy (see Lemma 5). Furthermore, while we adopt the $\tanh$ transformation due to its tractability and widespread use in practical implementations of soft actor-critic, much of the analysis extends to other sufficiently regular bijections from $\mathbb{R}^M$ to $(-1, 1)^M$.

7 Algorithm

For any iteration $n \in \mathbb{N}$, we denote the replay buffer state law by $\rho_n \in \mathcal{P}(S)$ with $\rho_0 = \rho$ the initial state distribution, which we model as an exponentially weighted mixture of all preceding state-occupancies up to the iteration $n > 0$ with parameter $\chi > 0$ (see Definition 5 in Appendix B). Then for any target policy $\pi_{\mathrm{target}} \in \mathcal{P}(A|S)$, $\pi_w \in \Pi_\Sigma$ and $n \in \mathbb{N}$ we define

$$J_n(w, \pi_{\mathrm{target}}) = \int_S \mathrm{KL}(\pi_w(\cdot|s)|\pi_{\mathrm{target}}(\cdot|s))\rho_n(ds),$$

Now suppose that one has access to an estimate of the Q-function for each encountered policy, and denote this sequence by $\{\widehat{Q}^n\}_{n\geq 0}$. Let $\pi^0(da|s) := \Lambda(da)$ for all $s \in S$. Firstly, we will denote the Gibbs target as

$$\frac{d\pi_{\mathrm{G}}^{n+1}}{d\Lambda}(s, a) = \frac{1}{Z_n(s)} \exp\left(-\frac{1}{\tau}\widehat{Q}^n(s, a)\right). \tag{4}$$

Furthermore, we denote the target arising from policy mirror descent as

$$\pi^{n+1}(\cdot|s) = \operatorname*{arg\,min}_{m\in\mathcal{P}(A)} \left\{ \int_A \left(\widehat{Q}^n(s, a) + \tau \log \frac{d\pi^n}{d\Lambda}(s, a)\right) m(da) + \frac{1}{\lambda}\,\mathrm{KL}(m|\pi^n(\cdot|s)) \right\}. \tag{5}$$

Throughout, $\{\pi^n\}_{n\geq 0}$ denotes the evolving target sequence, updated recursively from the previous target rather than from the current actor, whereas π_{w^n} denotes the actor tracking it. Moreover, throughout we take the initial actor satisfies $w^0 \in \mathfrak{B}(\mathrm{R})$, and the replay parameters satisfy $0 < \chi\lambda \leq 1$. We present the training procedure in Algorithm 1. To summarise, we study the case where one performs a single step of gradient descent of the objective with a mirror descent target for each critic estimate, then projects the new parameters onto $\mathfrak{B}(\mathrm{R})$.

Algorithm 1 Projected actor update

1: **for** $n = 0, 1, 2, \ldots$ **do**
2: Obtain critic estimate $\widehat{Q}^n$
3: Update replay law ρ_{n+1}
4: $w^{n+1} = \mathcal{P}_{\mathfrak{B}(\mathrm{R})}\left(w^n - h\nabla_w J_{n+1}(w^n, \pi^{n+1})\right)$
5: **end for**

We work under the following assumptions.

Assumption 1 (Q-function oracle)**.** *There exists $\varepsilon_{\mathrm{crit}} \geq 0$ such that, for all $n \geq 0$, the critic estimate $\widehat{Q}^n \in B_b(S \times A)$ satisfies $\widehat{Q}^n(s, \cdot) \in C^2(A)$ for each $s \in S$, $\left|\widehat{Q}^n\right|_{B_b(S\times A)} \leq \widehat{Q}_{\max}$ and*

$$\left|\widehat{Q}^n - Q_\tau^{\pi_{w^n}}\right|_{B_b(S\times A)} \leq \varepsilon_{\mathrm{crit}}.$$

Assumption 1 allows us to isolate dynamics of training the Gaussian policies π_{w^n}. While one could additionally introduce a parametrised critic and analyse its update jointly with the actor, we abstract away from the critic dynamics and capture the resulting approximation error through $\varepsilon_{\mathrm{crit}}$. This further allows our results to accommodate general twice-differentiable Q-function parametrisations, including standard smooth neural network architectures. Assumption 1 holds in many practical implementations when one freezes the policy and performs sufficiently many steps of temporal difference learning and when the parametrisation of the critic is sufficiently rich.

Assumption 2 (Optimal-occupancy concentrability)**.** *There exists $A_2 < \infty$ such that*

$$\left| \frac{dd_\rho^{\pi^*}}{d\rho} \right|_{L^2(\rho)} \leq A_2.$$

Assumption 2 is designed to control the distributional shift between the optimal discounted occupancy measure and the initial state distribution. Such concentrability or distribution-mismatch conditions are standard in analyses of approximate dynamic programming, off-policy learning and policy-gradient methods (Munos & Szepesvári, 2008; Chen & Jiang, 2019; Agarwal et al., 2021). Note that, we require that only the optimal discounted occupancy measure $d_\rho^{\pi^*}$ has a density with respect to the initial state distribution $\rho \in \mathcal{P}(S)$ and is square integrable with respect to ρ which is strictly weaker than imposing a uniform bound in the supremum norm (Agarwal et al., 2021) and imposing a similar assumption for every encountered policy during training (Uehara et al., 2020; Munos & Szepesvári, 2008). One case where Assumption 2 holds is when the transition kernel is square integrable with respect to the initial state distribution (see Lemma 7). This includes, under appropriate covariance and mean conditions, dynamics with Gaussian transition noise when the initial state distribution is also Gaussian.

Assumption 3 (Eigenvalues for actor parametrisation)**.** *Assume that*

$$\lambda_P := \lambda_{\min} \left(\int_S x(s)x(s)^\top \rho(ds) \right) > 0.$$

Assumption 3 is a standard non-degeneracy condition on the feature map and requires that the feature covariance matrix under the initial state distribution be positive definite, ensuring that every direction in the actor parameter space is sufficiently represented under only the initial state distribution ρ.

8 Main results

While we do not impose any assumptions regarding the realisability of the critic or policy in the main results, we inevitably must account for the bias that arises when the policy class of squashed Gaussians are not sufficiently rich enough to represent the target policies.

Definition 3 (Gaussian approximation error)**.** *For each $\{\widehat{Q}^n\}_{n\geq 0}$, the Gaussian approximation error to the mirror descent policies $\{\pi^n\}_{n\geq 0}$ is defined as*

$$\varepsilon_{\text{act}}^n = \min_{w \in \mathfrak{B}(\mathrm{R})} J_n(w, \pi^n).$$

Moreover, for any $n \geq 1$, we define $\varepsilon_{\text{approx}}(n) = \varepsilon_{\text{crit}} + \frac{1}{n} \sum_{k=0}^{n-1} \left(\varepsilon_{\text{act}}^k\right)^{\frac{1}{2}}$

In the following results, we denote $w_*^n \in \arg\min_{w \in \mathfrak{B}(\mathrm{R})} J_n(w, \pi^n)$, the optimal parameters $w \in \mathfrak{B}(\mathrm{R})$ for the target at iteration $n \in \mathbb{N}$. Theorem 1 establishes the first bound on the value function optimality gap for Algorithm 1.

Theorem 1. *Let Assumptions 1 and 2 hold, let $0 < \tau\lambda < 1$. Then there exists $\mathrm{C} < \infty$ such that for all $n \geq 1$,*

$$\begin{aligned} &\min_{0 \leq r \leq n-1} V_\tau^{\pi_{w^r}}(\rho) - V_\tau^{\pi^*}(\rho) \\ &\quad \leq \mathrm{C} \left(\lambda \left(\frac{1}{\lambda^2 n} + 1 \right) + \frac{1}{n} \sum_{k=0}^{n-1} \left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) \right)^{\frac{1}{2}} + \varepsilon_{\text{approx}}(n) \right). \end{aligned}$$

See Appendix C for a proof.

Theorem 1 shows that, up to approximation errors, the convergence of Algorithm 1 relies on two components. The first is the actor tracking error, that is how well can performing just one step of gradient descent on the objective track the evolving target. Secondly, one can see that the first term yields a constant which can be controlled by λ. Thus if one can also control the actor tracking error by a function of λ, after a careful fine tuning a convergence rate may be possible. To that end,

as standard in the analyses of coupled actor-critic and online gradient descent algorithms (Barakat et al., 2022; Olshevsky & Gharesifard, 2023; Mokhtari et al., 2016), we must establish sufficient regularity of the objective in order to control the tracking error.

To that end, since in general the reverse KL objective is non-convex, we must derive some sufficient conditions under which convexity holds. The first step towards doing so is by defining the following operator.

Definition 4. *Let $D : A \to \mathbb{S}^M$ be $D(a) := \operatorname{diag}(1 - a_1^2, \dots, 1 - a_M^2)$ and for any $f \in C^2(A)$ define $\mathcal{T} : C^2(A) \to C(A; \mathbb{S}^M)$ by*

$$(\mathcal{T}f)(a) := D(a)^{\frac{1}{2}} \nabla_a^2 f(a) D(a)^{\frac{1}{2}} - 2 \operatorname{diag}\big(a_1 \partial_{a_1} f(a), \dots, a_M \partial_{a_M} f(a)\big).$$

Firstly observe that since $f \in C^2(A)$, $\mathcal{T}f$ is continuous on A and thus $\mathcal{T}$ is well defined. Moreover for any $a \in A$, we can express $\mathcal{T}$ component-wise as

$$(\mathcal{T}f)_{ij}(a) = \begin{cases} (1 - a_i^2)\partial_{a_i a_i} f(a) - 2a_i \partial_{a_i} f(a), & i = j, \\ \sqrt{(1 - a_i^2)(1 - a_j^2)}\, \partial_{a_i a_j} f(a), & i \neq j. \end{cases}$$

For the special case of $M = 1$, $\mathcal{T}$ reduces to the classical Legendre differential operator $\frac{d}{da}((1 - a^2)f'(a))$ (DLMF, §18.8). In the present setting, $\mathcal{T}$ provides a convenient way of expressing the curvature of the actor objective and critic estimate and which in turn allows us to establish sufficient convexity conditions on the actor objective. Recall that, for each integer $k \geq 1$, $\mathbb{S}^k$ denotes the space of real symmetric matrices in $\mathbb{R}^{k \times k}$.

Theorem 2. *Let Assumptions 1 and 3 hold and define*

$$R_{\widehat{Q}} := \sup_{n \in \mathbb{N},\ s \in S,\ a \in A} \left| \mathcal{T}(\widehat{Q}^n(s, \cdot))(a) \right|_{\mathrm{op}},$$

Moreover let $0 < \tau\lambda < 1$ and $c_{\Sigma,\mathrm{R}} := \min_{1 \leq i \leq M} \left(1 - \tanh^2(\mathrm{R} + \sigma_i)\right)(2\Phi(1) - 1)$ with $\Phi : \mathbb{R} \to (0, 1)$ the standard Gaussian cumulative distribution function. If it holds that

$$\frac{R_{\widehat{Q}}}{2} < \tau,$$

then the functions $w \mapsto J_n(w, \pi^n)$ and $w \mapsto J_n(w, \pi_{\mathrm{G}}^n)$ are $(1 - \gamma)\left(2 - \frac{R_{\widehat{Q}}}{\tau}\right) c_{\Sigma,\mathrm{R}} \lambda_P$-strongly convex and $2 + \frac{R_{\widehat{Q}}}{\tau}$-smooth on $\mathfrak{B}(\mathrm{R})$ for all $n \in \mathbb{N}$.

See Appendix D for a proof.

Theorem 2 demonstrates sufficient conditions under which for any $n \in \mathbb{N}$, the actor objective with mirror descent target $w \mapsto J_n(w, \pi^n)$ is strongly convex on the Euclidean ball of radius $\mathrm{R} > 0$. Fundamentally speaking, this condition ensures that the Hessian of the actor objective is uniformly positive definite. Intuitively speaking, the curvature induced by the differential entropy relative to Lebesgue measure dominates the worst-case adverse curvature introduced by the critic, thereby preventing the actor objective from developing locally non-convex directions. See Remark 1 in Appendix D for an explicit example of $R_{\widehat{Q}}$ for Fourier critic parametrisations and see Figure 1 for a comparison of training dynamics for the MDP defined in Appendix K in the cases where the condition in Theorem 2 is satisfied and violated.

With Theorem 2 in hand, we now seek to control the actor tracking error.

Lemma 2 (Actor tracking)**.** *Let Assumptions 1 and 3 and the condition of Theorem 2 hold, and suppose that $0 < h < \frac{\tau}{2\tau + R_{\widehat{Q}}}$. Then there exists a non-negative constant $C_1 < \infty$ such that for every $n \geq 1$, it holds that*

$$\frac{1}{n} \sum_{k=0}^{n-1} \left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k)\right)^{\frac{1}{2}} \leq C_1 \left(\frac{1}{hn} + \frac{\lambda}{h^2}\right)^{\frac{1}{4}}.$$

See Appendix F for a proof.

Lemma 2 separates the actor optimisation error from the drift of the target and shows that, when the target policy arises from the mirror-descent update (5), the drift of the target can be directly controlled through the mirror-descent step size $\lambda > 0$. In particular, small λ makes consecutive target policies closer and therefore allows the actor, despite performing only a single gradient step per iteration, to track the evolving target more accurately.

However, of course choosing λ arbitrarily small does not come for free. Indeed, Theorem 1 contains the term $\frac{1}{\lambda n}$, reflecting the progress made by the underlying mirror-descent updates that we are approximating. This demonstrates an intrinsic trade off for Algorithm 1 between smaller values of λ, which produce a slowly moving target that is easier for the actor to track, and larger values of λ which allow for more aggressive policy improvement but increase the actor tracking error. Therefore balancing these effects is required to establish a result which demonstrates convergence up to approximation errors.

It is now worth addressing the behaviour of the actor tracking error when one has the classical Gibbs target (4) rather than the mirror descent target. For the following corollary and its proof, $\{w^n\}_{n\geq 0}$ denotes the iterates of Algorithm 1 with π^{n+1} replaced by π_{G}^{n+1} in the actor update in Algorithm 1. To that end, for $k \geq 1$ let us denote $w_{\mathrm{G},*}^k = \arg\min_{w\in\mathfrak{B}(\mathrm{R})} J_k(w, \pi_{\mathrm{G}}^k)$, which is unique under the conditions of Theorem 2.

Corollary 1. *Let Assumptions 1 and 3 and the condition of Theorem 2 hold and let* $0 < h < \frac{\tau}{2\tau + R_{\widehat{Q}}}$. *Then there exists* $C_{\mathrm{G}} < \infty$ *such that, for every* $n \geq 1$,

$$\frac{1}{n}\sum_{k=1}^{n-1}\left(J_k(w^k, \pi_{\mathrm{G}}^k) - J_k(w_{\mathrm{G},*}^k, \pi_{\mathrm{G}}^k)\right)^{\frac{1}{2}} \leq C_{\mathrm{G}}\left(\frac{1}{hn} + \frac{1}{h^2}\left(\frac{1}{\tau} + \chi\lambda\right)\right)^{\frac{1}{4}}.$$

The classical Gibbs target incurs the non-vanishing term $\frac{1}{\tau}$. Consequently, even when the replay distribution changes slowly, the bound in Corollary 1 does not guarantee that the actor can track the Gibbs target arbitrarily closely. In contrast, the mirror-descent formulation provides an explicit mechanism for slowing the evolution of the target to match the optimisation timescale of the actor. See Figure 2.

Theorem 3. *Let Assumptions 1, 2, and 3 hold and suppose that the conditions of Theorem 2 hold. For any integer* $N \geq 1$, *let* $\vartheta := \min\left\{\frac{1}{\tau}, \frac{1}{\chi}\right\}$ *and let* $\{\pi_{w^r}\}_{r=0}^{N-1}$ *be the iterates of Algorithm 1 with* $\lambda = \frac{\vartheta}{2N^{\frac{4}{5}}}$ *and* $0 < h < \frac{\tau}{2\tau + R_{\widehat{Q}}}$. *Then there exists* $C < \infty$ *such that*

$$\min_{0\leq r\leq N-1} V_\tau^{\pi_{w^r}}(\rho) - V_\tau^{\pi^*}(\rho) \leq C\left(N^{-\frac{1}{5}} + \varepsilon_{\mathrm{approx}}(N)\right).$$

Theorem 3 demonstrates that for a carefully tuned proximal penalty/step-size $\lambda > 0$ and step size $h > 0$ and any projecting radius $\mathrm{R} > 0$ one can establish a finite time convergence rate up to approximation errors. One promising route to achieving a faster convergence rate, at the cost of a more intricate analysis, is by also incorporating the critic dynamics and controlling the accuracy to which the critic is solved to at each iteration. Another one is to restrict $\mathrm{R} > \delta$ for some $\delta > 0$ sufficiently large in order to ensure that the minimisers w_*^n lie in the interior of $\mathfrak{B}(\mathrm{R})$. We do not pursue this here in order to retain generality for every $\mathrm{R} > 0$.

9 Conclusion and limitations

In this work, we prove convergence guarantees for a mathematical formalisation of Soft Actor Critic and demonstrate theoretical advantages of having a target policy arising from mirror descent rather than the classical Gibbs measure.

While the results in this work can be directly demonstrated in simplified settings (see Figures 1 and 2), the analysis is only performed in the population setting and thus the effect of sampling remains open. Moreover, we do not incorporate explicit critic dynamics in this work which remains a challenging direction for future work.

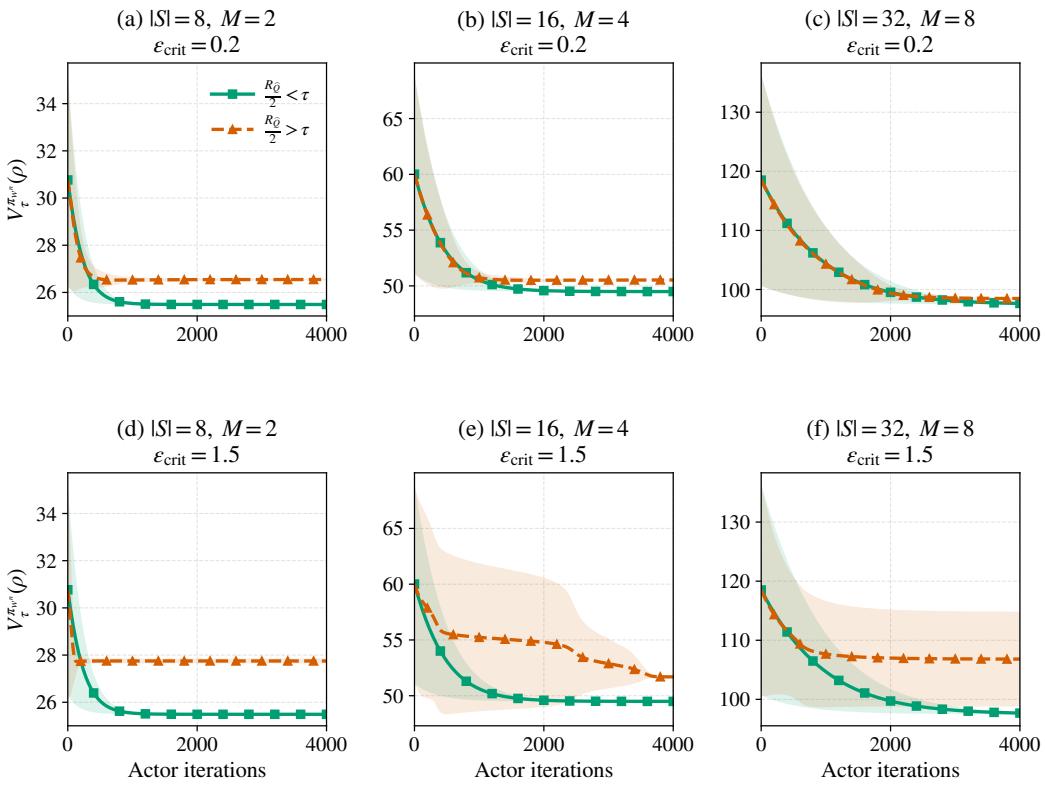

Figure 1: Training curves of Algorithm 1 for the MDP in Appendix K, with varying state-space size $|S|$ and action dimension M, $A = (-1, 1)^M$, and temperature $\tau = 1$. Within each panel, the two critics have equal error $\varepsilon_{\text{crit}}$ but different $R_{\widehat{Q}}$. The curvature condition of Theorem 2 is satisfied in green and violated in orange. Curves show means and shaded regions show one standard deviation across nine paired initialisations.

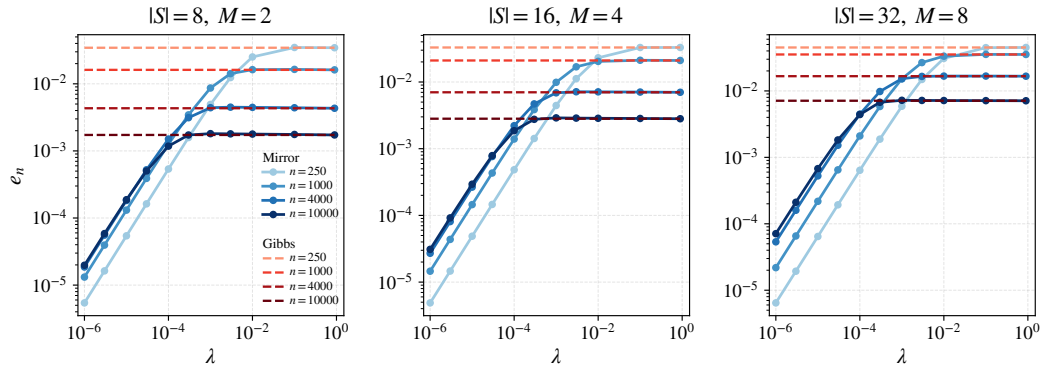

Figure 2: Averaged actor tracking error $e_n := \frac{1}{n}\sum_{k<n}\left(J_k(w^k, \pi^k) - J_k(w^k_*, \pi^k)\right)^{\frac{1}{2}}$ against the mirror-descent step size λ for the MDP defined in Appendix K with exact critics, evaluated at four actor budgets n (light to dark). Solid lines use the mirror-descent target (5) and dashed lines the Gibbs target (4) at the same budget. For the Gibbs target λ does not enter (4) itself, only the replay dynamics of Definition 5. See Appendix K for more details.

AI USE STATEMENT

In this work, we used generative AI tools for the purposes of aiding and polishing the contents of this work, including proofreading and the interpretation of results. Generative AI tools were also used for the purpose of assisting with the implementation of the experiment described in Appendix K.

We have not used generative AI to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, or propose or refine hypotheses. The use of generative AI for translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and generating synthetic datasets is not applicable to this work.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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. URL http://jmlr.org/papers/v22/19-736.html.

Carlo Alfano, Rui Yuan, and Patrick Rebeschini. A novel framework for policy mirror descent with general parameterization and linear convergence. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30681–30725. Curran Associates, Inc., 2023. doi: 10.52202/075280-1338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/61a9278dfef5f871b5e472389f8d6fa1-Abstract-Conference.html.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, second edition, 2008. doi: 10.1007/978-3-7643-8722-8.

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, pp. 17263–17275. Curran Associates, Inc., 2022. doi: 10.52202/068431-1255. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6e3daaeca6be8579573f69082b2dd58b-Abstract.html.

Anas Barakat, Pascal Bianchi, and Julien Lehmann. Analysis of a target-based actor-critic algorithm with linear function approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 991–1040. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/barakat22a.html.

Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M. Sadler, Pratap Tokekar, and Alec Koppel. On the hidden biases of policy mirror ascent in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1716–1731. PMLR, 2022. URL https://proceedings.mlr.press/v162/bedi22a.html.

Vivek S. Borkar and Vijaymohan R. Konda. The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22(4):525–543, 1997. doi: 10.1007/BF02745577.

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. URL https://doi.org/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. URL https://openreview.net/forum?id=BkEqk7pS1I.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578, 2022. doi: 10.1287/opre.2021.2151. URL https://arxiv.org/abs/2007.06558.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1042–1051. PMLR, 2019. URL https://proceedings.mlr.press/v97/chen19e.html.

DLMF. *NIST Digital Library of Mathematical Functions*. https://dlmf.nist.gov/, Release 1.2.8 of 2026-09-15, 2026. URL https://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

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. URL https://doi.org/10.1002/9781118165904.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (Last iterate) of actor-critic under Markovian sampling with neural network parametrization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15153–15179. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gaur24a.html.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 2017. URL https://proceedings.mlr.press/v70/haarnoja17a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018a. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications, 2018b. URL https://arxiv.org/abs/1812.05905.

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. URL https://doi.org/10.1137/20M1387341.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 1*. John Wiley & Sons, New York, second edition, 1994. ISBN 9780471584957.

Caleb Ju and Guanghui Lan. Policy optimization over general state and action spaces. *SIAM Journal on Optimization*, 36(3):1731–1772, 2026. doi: 10.1137/25M1791585. URL https://doi.org/10.1137/25M1791585.

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*, 26:2395–2469, 2026. doi: 10.1007/s10208-025-09729-3. URL https://doi.org/10.1007/s10208-025-09729-3.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12, pp. 1008–1014. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Navdeep Kumar, Priyank Agrawal, Giorgia Ramponi, Kfir Y. Levy, and Shie Mannor. On the convergence of single-timescale actor-critic. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 36406–36438. Curran Associates, Inc., 2025. doi: 10.52202/085713-1223. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/33ed8e42e696dfcb77379777c5036472-Paper-Conference.pdf.

Safwan Labbi, Paul Mangold, Daniil Tiapkin, and Eric Moulines. Refined analysis of entropy-regularized actor-critic. arXiv:2605.24357, 2026. URL https://arxiv.org/abs/2605.24357.

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12222–12252. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/leahy22a.html.

Sergey Levine and Vladlen Koltun. Guided policy search. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1–9. PMLR, 2013. URL https://proceedings.mlr.press/v28/levine13.html. Part 3.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. URL https://jmlr.org/papers/v17/15-522.html.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.02971.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829. PMLR, 2020. URL https://proceedings.mlr.press/v119/mei20b.html.

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, 2015. doi: 10.1038/nature14236.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7195–7201. IEEE, 2016. doi: 10.1109/CDC.2016.7799379.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008. URL https://jmlr.org/papers/v9/munos08a.html.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, 2004. doi: 10.1007/978-1-4419-8853-9.

Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023. doi: 10.1137/22M1483335.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Uri Sherman, Tomer Koren, and Yishay Mansour. Convergence of policy mirror descent beyond compatible function approximation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54825–54863. PMLR, 2025. URL https://proceedings.mlr.press/v267/sherman25a.html.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988. doi: 10.1007/BF00115009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, second edition, 2018. ISBN 0262039249.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063. MIT Press, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization, 2021. URL https://arxiv.org/abs/2005.09814.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In Hal Daumé, III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9659–9668. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/uehara20a.html.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Rui Yuan, Simon S. Du, Robert M. Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda, 2023. OpenReview.net. URL https://openreview.net/forum?id=-z9hdsyUwVQ.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D. Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023. doi: 10.1137/21M1456789. URL https://arxiv.org/abs/2105.11066.

Denis Zorba, David Šiška, and Lukasz Szpruch. Mirror descent actor-critic methods for entropy regularised MDPs in general spaces: stability and convergence, 2026. URL https://arxiv.org/abs/2602.10838.

A NOTATION

Let (E, d_E) denote a Polish space, that is, a complete separable metric space. We always equip a Polish space with its Borel σ-field $\mathcal{B}(E)$. We denote by $B_b(E)$ the space of bounded measurable functions $f : E \to \mathbb{R}$, endowed with the supremum norm

$$|f|_{B_b(E)} := \sup_{x \in E} |f(x)|.$$

We denote by $\mathcal{M}(E)$ the space of finite signed measures on E, endowed with the total variation norm

$$|\mu|_{\mathcal{M}(E)} := \sup_{|f|_{B_b(E)} \leq 1} \left| \int_E f(x)\mu(dx) \right|,$$

and by $\mathcal{P}(E) \subset \mathcal{M}(E)$ the set of probability measures on E. For Polish spaces E and F, $\mathcal{P}(E|F)$ denotes the set of probability kernels from F to E. For $\mu, \nu \in \mathcal{P}(E)$, their relative entropy is

$$\mathrm{KL}(\mu|\nu) := \int_E \log \frac{d\mu}{d\nu}(x)\mu(dx)$$

when μ is absolutely continuous with respect to ν, and $\mathrm{KL}(\mu|\nu) := +\infty$ otherwise. For $\mu \in \mathcal{P}(E)$ and a fixed σ-finite reference measure ν on E, the entropy is $\mathrm{H}(\mu) := -\int_E \log \frac{d\mu}{d\nu}(x)\mu(dx)$ when μ is absolutely continuous with respect to ν and the integral is well defined, and $-\infty$ when it is not absolutely continuous; the reference measure is Λ on A and $\Lambda_{\mathbb{R}^M}$ for the Gaussian laws on $\mathbb{R}^M$. Given $p \geq 1$, a measure $\mu \in \mathcal{P}(E)$, and a measurable function $f : E \to \mathbb{R}$, we write

$$|f|_{L^p(\mu)} := \left(\int_E |f(x)|^p \mu(dx) \right)^{\frac{1}{p}}.$$

We let $\mathbb{N} := \{0, 1, 2, \ldots\}$. For any $k \in \mathbb{N}$ with $k \geq 1$, we denote the Euclidean inner product on $\mathbb{R}^k$ by $\langle \cdot, \cdot \rangle$ with norm denoted by $| \cdot |_2$, and we write $|z|_1 := \sum_{i=1}^k |z_i|$ for the ℓ^1 norm. We denote the Euclidean ball of radius $R > 0$ in $\mathbb{R}^d$ by $\mathfrak{B}(\mathrm{R}) := \{w \in \mathbb{R}^d : |w|_2 \leq \mathrm{R}\}$. We denote by $\mathbb{S}^k$ the space of real symmetric $k \times k$ matrices and by I_k the $k \times k$ identity matrix. For a real matrix H, its Euclidean operator norm is

$$|H|_{\mathrm{op}} := \sup_{|v|_2 = 1} |Hv|_2.$$

For $H \in \mathbb{S}^k$, $\lambda_{\min}(H)$ denotes its minimum eigenvalue. Given $H, G \in \mathbb{S}^k$, we write $H \preceq G$ if $G - H$ is positive semidefinite and $H \prec G$ if $G - H$ is positive definite. For $z = (z_1, \ldots, z_k) \in \mathbb{R}^k$, $\mathrm{diag}(z_1, \ldots, z_k)$ denotes the $k \times k$ diagonal matrix with diagonal entries $z_1, \ldots, z_k$.

For an open set $U \subset \mathbb{R}^k$, $C^2(U)$ denotes the space of twice continuously differentiable real-valued functions on U. For a Polish space E, $C(E; \mathbb{S}^k)$ denotes the space of continuous functions from E to $\mathbb{S}^k$.

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 $P^\pi(ds', da'|s, a) := P(ds'|s, a)\pi(da'|s')$.

B BACKGROUND

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

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

where P_π^n is the n-times product of the kernel P_π with $P_\pi^0(ds'|s) := \delta_s(ds')$. Given any state distribution $\rho \in \mathcal{P}(S)$, we define the state-occupancy measure as

$$d_\rho^\pi(ds) = \int_S d^\pi(ds|s')\rho(ds'). \tag{6}$$

Definition 5. *Let $\rho \in \mathcal{P}(S)$ be an initial-state distribution, let $\chi > 0$, and suppose that $0 < \chi\lambda \leq 1$. Given a sequence of policies $\{\pi_{w^n}\}_{n\geq 0}$, define the sequence of replay distributions $\{\rho_n\}_{n\geq 0} \subset \mathcal{P}(S)$ recursively by*

$$\rho_0 := \rho, \qquad \rho_{n+1} := (1-\chi\lambda)\rho_n + \chi\lambda\, d_\rho^{\pi_{w^n}}, \qquad n \geq 0.$$

Equivalently, for every $n \geq 1$,

$$\rho_n = (1-\chi\lambda)^n \rho + \chi\lambda \sum_{j=0}^{n-1} (1-\chi\lambda)^{n-1-j} d_\rho^{\pi_{w^j}}.$$

Lemma 3. *For every $n \geq 0$, the replay measure satisfies*

$$\rho_n \geq (1-\gamma)\rho,$$

$$\lambda_{\min}\left(\int_S x(s)x(s)^\top \rho_n(ds)\right) \geq (1-\gamma)\lambda_P.$$

Proof of Lemma 3. Recall that $\rho_{n+1} := (1-\chi\lambda)\rho_n + \chi\lambda d_\rho^{\pi_{w^n}}$. Let $\beta := 1-\chi\lambda$. Unrolling this gives

$$\rho_n = \beta^n \rho + (1-\beta)\sum_{j=0}^{n-1} \beta^{n-1-j} d_\rho^{\pi_{w^j}}.$$

Since $d_\rho^\pi \geq (1-\gamma)\rho$ for every policy π,

$$\rho_n \geq \big(\beta^n + (1-\gamma)(1-\beta^n)\big)\rho \geq (1-\gamma)\rho.$$

Finally, for any $v \in \mathbb{R}^d$,

$$v^\top \left(\int_S x(s)x(s)^\top \rho_n(ds)\right) v \geq (1-\gamma) v^\top \left(\int_S x(s)x(s)^\top \rho(ds)\right) v,$$

which concludes the proof. □

We recall the dynamic programming principle from (Kerimkulov et al., 2026, Appendix B).

Theorem 4 (Dynamic Programming Principle)**.** *Let $\tau > 0$. The optimal value function V_τ^* is the unique bounded solution of the following Bellman equation:*

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

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

$$\pi^*(da|s) = \exp\left(-\frac{1}{\tau}(Q_\tau^*(s,a) - V_\tau^*(s))\right)\Lambda(da), \quad \forall s \in S. \tag{7}$$

Finally, for every $\pi \in \mathcal{P}(A|S)$ with $\mathrm{H}(\pi(\cdot|\cdot)) \in B_b(S)$, the value function V_τ^π is the unique bounded solution of the following Bellman equation for all $s \in S$

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

We now prove a performance difference lemma for policies with possibly unbounded log densities but with bounded entropy.

Lemma 4 (Performance difference)**.** *For every $\rho \in \mathcal{P}(S)$ and every $\pi \in \mathcal{P}(A|S)$ such that $V_\tau^\pi \in B_b(S)$ and $\mathrm{H}(\pi(\cdot|\cdot)) \in B_b(S)$, it holds that*

$$V_\tau^\pi(\rho) - V_\tau^{\pi^*}(\rho) = \frac{1}{1-\gamma}\int_S \left(\int_A Q_\tau^\pi(s,a)(\pi - \pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi(\cdot|s))\right)\right) d_\rho^{\pi^*}(ds).$$

Proof. Fix $\rho \in \mathcal{P}(S)$ and $\pi \in \mathcal{P}(A|S)$. By Theorem 4, for all $s \in S$ it holds that

$$\begin{aligned}
&\int_A Q_\tau^\pi(s,a)(\pi - \pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi(\cdot|s))\right) \\
&= \left(\int_A Q_\tau^\pi(s,a)\pi(da|s) - \tau\mathrm{H}(\pi(\cdot|s))\right) - \left(\int_A Q_\tau^\pi(s,a)\pi^*(da|s) - \tau\mathrm{H}(\pi^*(\cdot|s))\right) \\
&= V_\tau^\pi(s) - V_\tau^{\pi^*}(s) - \int_A \left(Q_\tau^\pi - Q_\tau^{\pi^*}\right)(s,a)\pi^*(da|s) \\
&= V_\tau^\pi(s) - V_\tau^{\pi^*}(s) - \gamma\int_S \left(V_\tau^\pi(s') - V_\tau^{\pi^*}(s')\right) P_{\pi^*}(ds'|s),
\end{aligned}$$

where in the second equality we used that $V_\tau^\pi(s) = \int_A Q_\tau^\pi(s,a)\pi(da|s) - \tau\mathrm{H}(\pi(\cdot|s))$. In the third equality, we used the definition of the state-action value function in (1). Now set $f := V_\tau^\pi - V_\tau^{\pi^*}$ and let g denote the left hand side. By assumption we have $V_\tau^\pi \in B_b(S)$ and by the Bellman principle, since $c \in B_b(S \times A)$, we have $V_\tau^{\pi^*} \in B_b(S)$ and thus $f \in B_b(S)$.

Similarly, since $V_\tau^\pi \in B_b(S)$ it holds that $Q_\tau^\pi \in B_b(S \times A)$ and $\mathrm{H}(\pi(\cdot|\cdot)) \in B_b(S)$. By the Bellman principle, it also holds that $\mathrm{H}(\pi^*(\cdot|\cdot)) \in B_b(S)$. Therefore, by Kerimkulov et al. (2026, Lemma 3.2), for all $s \in S$ it holds that

$$V_\tau^\pi(s) - V_\tau^{\pi^*}(s) = \frac{1}{1-\gamma}\int_S g(s')d^{\pi^*}(ds'|s).$$

Integrating both sides with respect to ρ and using the definition of the state occupancy measure in (6) concludes the proof. □

Define

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

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

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

Then for all $m \in M_\Lambda$ we have

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

C Proof of Theorem 1

Proof. For some $\lambda > 0$ with $1 - \tau\lambda \in (0,1)$ and for all $n \geq 0$ and $s \in S$, recall that

$$\pi^{n+1}(\cdot|s) = \operatorname*{arg\,min}_{m \in \mathcal{P}(A)}\left\{ \int_A \left(\widehat{Q}^n(s,a) + \tau\log\frac{d\pi^n}{d\Lambda}(s,a)\right) m(da) + \frac{1}{\lambda}\mathrm{KL}(m|\pi^n(\cdot|s))\right\}.$$

By (7) and (46), $\pi^*(\cdot|s), \pi^n(\cdot|s), \pi^{n+1}(\cdot|s) \in M_\Lambda$, and the minimiser in the preceding display belongs to M_Λ. The Three Point Lemma (Lemma 6 with $m = \pi^*(\cdot|s)$) yields

$$\begin{aligned}
&\int_A \left(\widehat{Q}^n(s,a) + \tau \log \frac{d\pi^n}{d\Lambda}(s,a)\right)(\pi^n - \pi^*)(da|s) \\
&\leq \frac{1}{\lambda}\left(\mathrm{KL}(\pi^*|\pi^n)(s) - \mathrm{KL}(\pi^*|\pi^{n+1})(s) - \mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) \\
&\quad + \int_A \left(\widehat{Q}^n(s,a) + \tau \log \frac{d\pi^n}{d\Lambda}(s,a)\right)(\pi^n - \pi^{n+1})(da|s).
\end{aligned} \tag{8}$$

We will now lower bound the left hand side. Firstly recall the standard identity

$$\int_A \log \frac{d\pi^n}{d\Lambda}(s,a)\pi^*(da|s) = -\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{KL}(\pi^*(\cdot|s)|\pi^n(\cdot|s)).$$

Consequently,

$$\begin{aligned}
&\int_A \widehat{Q}^n(s,a)(\pi^n - \pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\right) \\
&= \int_A \left(\widehat{Q}^n(s,a) + \tau \log \frac{d\pi^n}{d\Lambda}(s,a)\right)(\pi^n - \pi^*)(da|s) - \tau\,\mathrm{KL}(\pi^*|\pi^n)(s) \\
&\leq \int_A \left(\widehat{Q}^n(s,a) + \tau \log \frac{d\pi^n}{d\Lambda}(s,a)\right)(\pi^n - \pi^*)(da|s),
\end{aligned}$$

Therefore (8) becomes

$$\begin{aligned}
&\int_A \widehat{Q}^n(s,a)(\pi^n - \pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\right) \\
&\leq \frac{1}{\lambda}\left(\mathrm{KL}(\pi^*|\pi^n)(s) - \mathrm{KL}(\pi^*|\pi^{n+1})(s) - \mathrm{KL}(\pi^{n+1}|\pi^n)(s)\right) \\
&\quad + \int_A \left(\widehat{Q}^n(s,a) + \tau \log \frac{d\pi^n}{d\Lambda}(s,a)\right)(\pi^n - \pi^{n+1})(da|s).
\end{aligned} \tag{9}$$

Now we seek to upper bound the final term on the right hand side of (9). By Pinsker's inequality, it holds that

$$\begin{aligned}
&\int_A \left(\widehat{Q}^n(s,a) + \tau \log \frac{d\pi^n}{d\Lambda}(s,a)\right)(\pi^n - \pi^{n+1})(da|s) \\
&\leq \left(\widehat{Q}_{\max} + \tau \sup_{n\in\mathbb{N}} \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)}\right)\left|\pi^{n+1}(\cdot|s) - \pi^n(\cdot|s)\right|_{\mathcal{M}(A)} \\
&\leq \left(\widehat{Q}_{\max} + \tau C_{\log}\right)\sqrt{2\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}
\end{aligned}$$

where we used that $\sup_{n\in\mathbb{N}} \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)} \leq \frac{2\widehat{Q}_{\max}}{\tau} := C_{\log}$ by (47). Now, by Young's inequality in the form $ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$, with $a = \widehat{Q}_{\max} + \tau C_{\log}$ and $b = \sqrt{2\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}$, we obtain

$$\begin{aligned}
&\left(\widehat{Q}_{\max} + \tau C_{\log}\right)\sqrt{2\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))} - \frac{\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}{\lambda} \\
&\leq \frac{\lambda}{2}\left(\widehat{Q}_{\max} + \tau C_{\log}\right)^2 + \frac{1}{2\lambda}\left(\sqrt{2\,\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}\right)^2 - \frac{\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}{\lambda} \\
&= \frac{\lambda}{2}\left(\widehat{Q}_{\max} + \tau C_{\log}\right)^2 + \frac{\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}{\lambda} - \frac{\mathrm{KL}(\pi^{n+1}(\cdot|s)|\pi^n(\cdot|s))}{\lambda} \\
&= \frac{\lambda}{2}\left(\widehat{Q}_{\max} + \tau C_{\log}\right)^2.
\end{aligned}$$

Therefore (9) simplifies to

$$\begin{aligned}
&\int_A \widehat{Q}^n(s,a)(\pi^n - \pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\right) \\
&\leq \frac{1}{\lambda}\left(\mathrm{KL}(\pi^*|\pi^n)(s) - \mathrm{KL}(\pi^*|\pi^{n+1})(s)\right) + \frac{\lambda}{2}\left(\widehat{Q}_{\max} + \tau C_{\log}\right)^2.
\end{aligned}$$

Therefore after integrating over S with respect to $d_\rho^{\pi^*} \in \mathcal{P}(S)$ and summing over $k = 0, ..., n-1$ we arrive at

$$\sum_{k=0}^{n-1} \left(\int_S \left(\int_A \widehat{Q}^k(s,a)(\pi^k - \pi^*)(da|s) + \tau \left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^k(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds) \right) \tag{10}$$
$$\leq \frac{1}{\lambda} \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) + \frac{n\lambda}{2} \left(\widehat{Q}_{\max} + \tau C_{\log} \right)^2,$$

Now we turn to the performance difference lemma (Lemma 4) for between π_{w^n} and π^*, which states that

$$\left(V_\tau^{\pi_{w^n}}(\rho) - V_\tau^{\pi^*}(\rho) \right)$$
$$= \frac{1}{1-\gamma} \int_S \left(\int_A Q_\tau^{\pi_{w^n}}(s,a)(\pi_{w^n} - \pi^*)(da|s) + \tau \left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi_{w^n}(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds).$$

Adding and subtracting $\widehat{Q}^n$ and using Assumption 1, it holds that

$$\left(V_\tau^{\pi_{w^n}}(\rho) - V_\tau^{\pi^*}(\rho) \right)$$
$$\leq \frac{1}{1-\gamma} \int_S \left(\int_A \widehat{Q}^n(s,a)(\pi_{w^n} - \pi^*)(da|s) + \tau \left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi_{w^n}(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds)$$
$$+ \frac{2\varepsilon_{\mathrm{crit}}}{1-\gamma}.$$

Furthermore adding and subtracting π^n and $\mathrm{H}(\pi^n(\cdot|s))$ we arrive at

$$V_\tau^{\pi_{w^n}}(\rho) - V_\tau^{\pi^*}(\rho)$$
$$\leq \frac{1}{1-\gamma} \int_S \left(\int_A \widehat{Q}^n(s,a)(\pi^n - \pi^*)(da|s) + \tau \left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds)$$
$$+ \frac{1}{1-\gamma} \int_S \left(\int_A \widehat{Q}^n(s,a)(\pi_{w^n} - \pi^n)(da|s) + \tau \left(\mathrm{H}(\pi^n(\cdot|s)) - \mathrm{H}(\pi_{w^n}(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds)$$
$$+ \frac{2\varepsilon_{\mathrm{crit}}}{1-\gamma}. \tag{11}$$

To bound the second term on the right hand side, observe that by Theorem 5 and Pinsker's inequality, it holds that

$$\int_A \widehat{Q}^n(s,a)(\pi_{w^n} - \pi^n)(da|s) + \tau \left(\mathrm{H}(\pi^n(\cdot|s)) - \mathrm{H}(\pi_{w^n}(\cdot|s)) \right)$$
$$\leq \widehat{Q}_{\max} |\pi_{w^n}(\cdot|s) - \pi^n(\cdot|s)|_{\mathcal{M}(A)} + \tau C_{\mathrm{KL}} \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}}$$
$$\leq (\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}}) \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}}.$$

Therefore after integrating over S with respect to $d_\rho^{\pi^*}$ it also holds that

$$\int_S \left(\int_A \widehat{Q}^n(s,a)(\pi_{w^n} - \pi^n)(da|s) + \tau \left(\mathrm{H}(\pi^n(\cdot|s)) - \mathrm{H}(\pi_{w^n}(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds)$$
$$\leq (\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}}) \int_S \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}} d_\rho^{\pi^*}(ds).$$

By Holder's inequality and assumption 2, it holds that

$$
\begin{aligned}
&\int_S \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}} d_\rho^{\pi^*}(ds) \qquad (12)\\
&= \int_S \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}} \frac{dd_\rho^{\pi^*}}{d\rho}(s)\rho(ds)\\
&\leq \left(\int_S \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))\rho(ds)\right)^{\frac{1}{2}} \left(\int_S \left(\frac{dd_\rho^{\pi^*}}{d\rho}(s)\right)^2 \rho(ds)\right)^{\frac{1}{2}}\\
&= \left|\frac{dd_\rho^{\pi^*}}{d\rho}\right|_{L^2(\rho)} \left(\int_S \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))\rho(ds)\right)^{\frac{1}{2}}\\
&\leq \frac{1}{\sqrt{1-\gamma}} \left|\frac{dd_\rho^{\pi^*}}{d\rho}\right|_{L^2(\rho)} \left(\int_S \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))\rho_n(ds)\right)^{\frac{1}{2}}\\
&= \frac{1}{\sqrt{1-\gamma}} \left|\frac{dd_\rho^{\pi^*}}{d\rho}\right|_{L^2(\rho)} J_n(w^n,\pi^n)^{\frac{1}{2}}\\
&\leq \frac{A_2}{\sqrt{1-\gamma}} J_n(w^n,\pi^n)^{\frac{1}{2}}
\end{aligned}
$$

Hence substituting (12) into (11) it holds that

$$
\begin{aligned}
&V_\tau^{\pi_{w^n}}(\rho) - V_\tau^{\pi^*}(\rho)\\
&\quad\leq \frac{1}{1-\gamma}\int_S \left(\int_A \widehat{Q}^n(s,a)(\pi^n-\pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\right)\right) d_\rho^{\pi^*}(ds)\\
&\qquad + \frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}}} J_n(w^n,\pi^n)^{\frac{1}{2}} + \frac{2\varepsilon_{\mathrm{crit}}}{1-\gamma}
\end{aligned}
$$

Now let $w_*^n \in \arg\min_{w\in\mathfrak{B}(\mathrm{R})} J_n(w,\pi^n)$. Adding and subtracting $J_n(w_*^n,\pi^n)$ and using that $(a+b)^{\frac{1}{2}} \leq a^{\frac{1}{2}} + b^{\frac{1}{2}}$ for any non-negative a, b it holds that

$$
\begin{aligned}
&V_\tau^{\pi_{w^n}}(\rho) - V_\tau^{\pi^*}(\rho)\\
&\quad\leq \frac{1}{1-\gamma}\int_S \left(\int_A \widehat{Q}^n(s,a)(\pi^n-\pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\right)\right) d_\rho^{\pi^*}(ds)\\
&\qquad + \frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}}} \left(J_n(w^n,\pi^n) - J_n(w_*^n,\pi^n) + J_n(w_*^n,\pi^n)\right)^{\frac{1}{2}} + \frac{2\varepsilon_{\mathrm{crit}}}{1-\gamma}\\
&\leq \frac{1}{1-\gamma}\int_S \left(\int_A \widehat{Q}^n(s,a)(\pi^n-\pi^*)(da|s) + \tau\left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\right)\right) d_\rho^{\pi^*}(ds)\\
&\qquad + \frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}}} \left(\left(J_n(w^n,\pi^n) - J_n(w_*^n,\pi^n)\right)^{\frac{1}{2}} + J_n(w_*^n,\pi^n)^{\frac{1}{2}}\right) + \frac{2\varepsilon_{\mathrm{crit}}}{1-\gamma}.
\end{aligned}
$$

Summing over $k = 0, \dots, n-1$ for any $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
&\sum_{k=0}^{n-1} \left(V_\tau^{\pi_{w^k}}(\rho) - V_\tau^{\pi^*}(\rho) \right) \\
&\leq \frac{1}{1-\gamma} \sum_{k=0}^{n-1} \int_S \left(\int_A \widehat{Q}^k(s,a)(\pi^k - \pi^*)(da|s) + \tau \left(\mathrm{H}(\pi^*(\cdot|s)) - \mathrm{H}(\pi^k(\cdot|s)) \right) \right) d_\rho^{\pi^*}(ds) \\
&\quad + \frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}}} \sum_{k=0}^{n-1} \left(\left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) \right)^{\frac{1}{2}} + J_k(w_*^k, \pi^k)^{\frac{1}{2}} \right) + \frac{2\varepsilon_{\mathrm{crit}} n}{1-\gamma}.
\end{aligned}$$

Substituting in (10), we arrive at

$$\begin{aligned}
&\sum_{k=0}^{n-1} \left(V_\tau^{\pi_{w^k}}(\rho) - V_\tau^{\pi^*}(\rho) \right) \\
&\leq \frac{1}{\lambda(1-\gamma)} \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) + \frac{n\lambda}{2(1-\gamma)} \left(\widehat{Q}_{\max} + \tau C_{\log} \right)^2 \\
&\quad + \frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}}} \sum_{k=0}^{n-1} \left(\left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) \right)^{\frac{1}{2}} + J_k(w_*^k, \pi^k)^{\frac{1}{2}} \right) + \frac{2\varepsilon_{\mathrm{crit}} n}{1-\gamma}.
\end{aligned}$$

Dividing through by $n > 0$, using that the minimum is less than the average and $J_k(w_*^k, \pi^k) = \varepsilon_{\mathrm{act}}^k$, we arrive at

$$\begin{aligned}
&\min_{0 \leq r \leq n-1} V_\tau^{\pi_{w^r}}(\rho) - V_\tau^{\pi^*}(\rho) \\
&\leq \frac{1}{\lambda(1-\gamma)n} \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds) + \frac{\lambda}{2(1-\gamma)} \left(\widehat{Q}_{\max} + \tau C_{\log} \right)^2 \\
&\quad + \frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}} n} \sum_{k=0}^{n-1} \left(\left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) \right)^{\frac{1}{2}} + (\varepsilon_{\mathrm{act}}^k)^{\frac{1}{2}} \right) + \frac{2\varepsilon_{\mathrm{crit}}}{1-\gamma} \\
&\leq \mathrm{C} \left(\frac{1}{\lambda n} + \lambda + \frac{1}{n} \sum_{k=0}^{n-1} \left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) \right)^{\frac{1}{2}} + \varepsilon_{\mathrm{approx}}(n) \right),
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{C} := \max \Bigg\{ &\frac{1}{1-\gamma} \int_S \mathrm{KL}(\pi^*|\pi^0)(s) d_\rho^{\pi^*}(ds), \frac{(\widehat{Q}_{\max} + \tau C_{\log})^2}{2(1-\gamma)}, \\
&\frac{A_2(\sqrt{2}\widehat{Q}_{\max} + \tau C_{\mathrm{KL}})}{(1-\gamma)^{\frac{3}{2}}}, \frac{2}{1-\gamma} \Bigg\}.
\end{aligned}$$

□

D Proof of Theorem 2

Proof. For any $n \in \mathbb{N}$ and $w \in \mathfrak{B}(\mathrm{R})$, recall that

$$J_{n+1}(w, \pi^{n+1}) = \int_S \mathrm{KL}(\pi_w(\cdot|s)|\pi^{n+1}(\cdot|s)) \rho_{n+1}(ds).$$

Focusing on the integrand, by Lemma 1, for every $s \in S$ it holds that

$$\begin{aligned}
\mathrm{KL}(\pi_w(\cdot|s)|\pi^{n+1}(\cdot|s)) &= \mathrm{KL}\left((\tanh)_\# q_w(\cdot|s) | \pi^{n+1}(\cdot|s) \right) \\
&= \mathrm{KL}\left(q_w(\cdot|s) | \left(\tanh^{-1}\right)_\# \pi^{n+1}(\cdot|s) \right).
\end{aligned} \tag{13}$$

Let $\Lambda_{\mathbb{R}^M}$ denote Lebesgue measure on $\mathbb{R}^M$. Since $\tanh : \mathbb{R}^M \to A$ is a measurable bijection with measurable inverse, performing the change of variable $a = \tanh u$ with $u \in \mathbb{R}^M$, for every $B \in \mathcal{B}(\mathbb{R}^M)$ it holds that

$$\begin{aligned}\left((\tanh^{-1})_{\#}\pi^{n+1}\right)(B|s) &= \pi^{n+1}(\tanh(B)|s)\\ &= \int_{\tanh(B)} \frac{d\pi^{n+1}}{d\Lambda}(s,a)\Lambda(da)\\ &= \int_B \frac{d\pi^{n+1}}{d\Lambda}(s,\tanh u)2^{-M}\det D(\tanh u)\Lambda_{\mathbb{R}^M}(du).\end{aligned}$$

Thus it holds that

$$\begin{aligned}\frac{d\left(\tanh^{-1}\right)_{\#}\pi^{n+1}}{d\Lambda_{\mathbb{R}^M}}(s,u) &= \frac{\det D(\tanh u)}{2^M}\frac{d\pi^{n+1}}{d\Lambda}(s,\tanh u)\\ &= \frac{\det D(\tanh u)}{2^M\overline{Z}_{n+1}(s)}\exp\left(-\lambda\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}\widehat{Q}^k(s,\tanh u)\right),\end{aligned} \tag{14}$$

such that $\overline{Z}_{n+1}(s) := \int_A \exp\left(-\lambda\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}\widehat{Q}^k(s,a)\right)\Lambda(da)$. To ease notation, let

$$F_n(s,u) := \lambda\sum_{k=0}^{n}(1-\tau\lambda)^{n-k}\widehat{Q}^k(s,\tanh u).$$

Starting from (13) we therefore have that

$$\begin{aligned}&\mathrm{KL}(\pi_w(\cdot|s)|\pi^{n+1}(\cdot|s))\\ &= \mathrm{KL}\left(q_w(\cdot|s)|\left(\tanh^{-1}\right)_{\#}\pi^{n+1}(\cdot|s)\right)\\ &= \int_{\mathbb{R}^M}\log\frac{dq_w}{d\left(\tanh^{-1}\right)_{\#}\pi^{n+1}}(s,u)q_w(du|s)\\ &= \int_{\mathbb{R}^M}\log\left(\frac{dq_w}{d\Lambda_{\mathbb{R}^M}}(s,u)\frac{d\Lambda_{\mathbb{R}^M}}{d\left(\tanh^{-1}\right)_{\#}\pi^{n+1}}(s,u)\right)q_w(du|s)\\ &= \int_{\mathbb{R}^M}\log\frac{dq_w}{d\Lambda_{\mathbb{R}^M}}(s,u)q_w(du|s) - \int_{\mathbb{R}^M}\log\frac{d\left(\tanh^{-1}\right)_{\#}\pi^{n+1}}{d\Lambda_{\mathbb{R}^M}}(s,u)q_w(du|s)\\ &= -\mathrm{H}(q_w(\cdot|s)) + \int_{\mathbb{R}^M}\left(\log\left(2^M\overline{Z}_{n+1}(s)\right) + F_n(s,u) - \log\det D(\tanh u)\right)q_w(du|s)\\ &= -\mathrm{H}(q_w(\cdot|s)) + \log\left(2^M\overline{Z}_{n+1}(s)\right) + \int_{\mathbb{R}^M}\left(2\sum_{i=1}^{M}\log\cosh u_i + F_n(s,u)\right)q_w(du|s)\\ &= \int_{\mathbb{R}^M}\left(2\sum_{i=1}^{M}\log\cosh u_i + F_n(s,u)\right)q_w(du|s) + \log\frac{2^M\overline{Z}_{n+1}(s)}{(2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}} - \frac{M}{2},\end{aligned} \tag{15}$$

where in the fifth equality we used (14) and the definition of the entropy $\mathrm{H}(q_w(\cdot|s))$. In the sixth equality we used $\det D(\tanh u) = \prod_{i=1}^{M}\cosh^{-2}u_i$ and in the final equality we used that $\mathrm{H}(q_w(\cdot|s)) = \log((2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}) + \frac{M}{2}$. We now further ease notation through

$$\widetilde{F}_n(s,u) := 2\sum_{i=1}^{M}\log\cosh u_i + F_n(s,u), \qquad C_n := \int_S\left(\log\frac{2^M\overline{Z}_{n+1}(s)}{(2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}} - \frac{M}{2}\right)\rho_{n+1}(ds).$$

Therefore integrating (15) over S with respect to ρ_{n+1} it holds that

$$J_{n+1}(w,\pi^{n+1}) = \int_S\int_{\mathbb{R}^M}\widetilde{F}_n(s,u)q_w(du|s)\rho_{n+1}(ds) + C_n.$$

Now recall that for $f \in C^2(A)$ and $a = \tanh u$, the definition of $\mathcal{T}$ and the chain rule give, for every $i, j \in \{1, \dots, M\}$,

$$\partial_{u_i u_j}(f(\tanh u)) = (1-a_i^2)(1-a_j^2)\partial_{a_i a_j} f(a) - 2\mathbf{1}_{\{i=j\}} a_i (1-a_i^2)\partial_{a_i} f(a),$$
$$\nabla_u^2 (f(\tanh u)) = D(a)^{\frac{1}{2}} (\mathcal{T} f)(a) D(a)^{\frac{1}{2}}.$$

Moreover recall that $R_{\widehat{Q}} = \sup_{k\in\mathbb{N},\, s\in S,\, a\in A} |\mathcal{T}(\widehat{Q}^k(s,\cdot))(a)|_{\mathrm{op}}$. Since $0 \prec D(a) \preceq I_M$, it therefore holds that

$$\nabla_u^2 \widetilde{F}_n(s,u) = D(a)^{\frac{1}{2}} \left(2I_M + \lambda \sum_{k=0}^{n} (1-\tau\lambda)^{n-k} \mathcal{T}(\widehat{Q}^k(s,\cdot))(a) \right) D(a)^{\frac{1}{2}}$$
$$\succeq \left(2 - R_{\widehat{Q}} \lambda \sum_{k=0}^{n} (1-\tau\lambda)^{n-k} \right) D(a).$$

Analogously it also holds that

$$\nabla_u^2 \widetilde{F}_n(s,u) = D(a)^{\frac{1}{2}} \left(2I_M + \lambda \sum_{k=0}^{n} (1-\tau\lambda)^{n-k} \mathcal{T}(\widehat{Q}^k(s,\cdot))(a) \right) D(a)^{\frac{1}{2}}$$
$$\preceq \left(2 + R_{\widehat{Q}} \lambda \sum_{k=0}^{n} (1-\tau\lambda)^{n-k} \right) D(a).$$

and thus after using that $\lambda \sum_{k=0}^{n} (1-\tau\lambda)^{n-k} \leq \frac{1}{\tau}$, for all $s \in S$ and $u \in \mathbb{R}^M$ it holds that

$$\left(2 - \frac{R_{\widehat{Q}}}{\tau} \right) D(a) \preceq \nabla_u^2 \widetilde{F}_n(s,u) \preceq \left(2 + \frac{R_{\widehat{Q}}}{\tau} \right) D(a).$$

Since $a = \tanh u$, we may write these bounds as

$$\left(2 - \frac{R_{\widehat{Q}}}{\tau} \right) D(\tanh u) \preceq \nabla_u^2 \widetilde{F}_n(s,u) \preceq \left(2 + \frac{R_{\widehat{Q}}}{\tau} \right) D(\tanh u). \tag{16}$$

Now we have upper and lower bounds on the Hessian of $\widetilde{F}_n$, we translate these into upper and lower bounds on the Hessian of the objective $J_{n+1}(w, \pi^{n+1}) = \int_S \int_{\mathbb{R}^M} \widetilde{F}_n(s,u) q_w(du|s) \rho_{n+1}(ds) + C_n$. On that end, firstly observe that after performing the change of variable $u = m_w(s) + \Sigma^{\frac{1}{2}} z$ it holds that

$$\int_{\mathbb{R}^M} \widetilde{F}_n(s,u) q_w(du|s) = \int_{\mathbb{R}^M} \widetilde{F}_n(s, m_w(s) + \Sigma^{\frac{1}{2}} z) q^1(dz).$$

Applying the chain rule twice (after recalling that $m_w(s) = x(s)^\top w$) yields

$$\nabla_w^2 \left(\int_{\mathbb{R}^M} \widetilde{F}_n(s,u) q_w(du|s) \right) = x(s) \left(\int_{\mathbb{R}^M} \nabla_u^2 \widetilde{F}_n(s,u) \Big|_{u = m_w(s) + \Sigma^{\frac{1}{2}} z} q^1(dz) \right) x(s)^\top.$$

Consequently, for every $v \in \mathbb{R}^d$, it holds that

$$v^\top \nabla_w^2 \left(\int_{\mathbb{R}^M} \widetilde{F}_n(s,u) q_w(du|s) \right) v \tag{17}$$
$$= \int_{\mathbb{R}^M} (x(s)^\top v)^\top \, \nabla_u^2 \widetilde{F}_n(s,u) \Big|_{u = m_w(s) + \Sigma^{\frac{1}{2}} z} (x(s)^\top v) q^1(dz).$$

Now we apply the lower bound in (16) at $u = m_w(s) + \Sigma^{\frac{1}{2}} z$ to arrive at

$$v^\top \nabla_w^2 \left(\int_{\mathbb{R}^M} \widetilde{F}_n(s,u) q_w(du|s) \right) v \tag{18}$$
$$\geq \left(2 - \frac{R_{\widehat{Q}}}{\tau} \right) \sum_{i=1}^{M} (x(s)^\top v)_i^2 \int_{\mathbb{R}^M} \left(1 - \tanh^2(m_w(s)_i + \sigma_i z_i) \right) q^1(dz).$$

It thus remains to lower bound each integral $\int_{\mathbb{R}^M}(1-\tanh^2(m_w(s)_i+\sigma_i z_i))q^1(dz)$. To that end, since the integrand is non-negative and $r\mapsto 1-\tanh^2 r$ is even and decreasing for $r\geq 0$, for every $i\in\{1,\dots,M\}$, $w\in\mathfrak{B}(\mathrm{R})$ and $s\in S$, it holds that

$$\begin{aligned}\int_{\mathbb{R}^M}\left(1-\tanh^2(m_w(s)_i+\sigma_i z_i)\right)q^1(dz) &\geq \int_{\{z\in\mathbb{R}^M:|z_i|\leq 1\}}\left(1-\tanh^2(m_w(s)_i+\sigma_i z_i)\right)q^1(dz)\\ &\geq \left(1-\tanh^2(\mathrm{R}+\sigma_i)\right)\int_{\{z\in\mathbb{R}^M:|z_i|\leq 1\}}q^1(dz)\\ &= \left(1-\tanh^2(\mathrm{R}+\sigma_i)\right)\mathbb{P}(|Z_i|\leq 1)\\ &= \left(1-\tanh^2(\mathrm{R}+\sigma_i)\right)(2\Phi(1)-1).\end{aligned}$$

Here $Z\sim q^1$, Φ is the standard univariate Gaussian distribution function, and $|m_w(s)_i|\leq |x(s)^\top w|_2\leq \mathrm{R}$. Taking the minimum of these lower bounds over i gives

$$\begin{aligned}\int_{\mathbb{R}^M}\left(1-\tanh^2(m_w(s)_i+\sigma_i z_i)\right)q^1(dz) &\geq c_{\Sigma,\mathrm{R}}\\ &= \min_{1\leq j\leq M}\left(1-\tanh^2(\mathrm{R}+\sigma_j)\right)(2\Phi(1)-1)>0.\end{aligned}$$

Let $c_{\Sigma,\mathrm{R}}:=\min_{1\leq j\leq M}\left(1-\tanh^2(\mathrm{R}+\sigma_j)\right)(2\Phi(1)-1)$. (18) becomes

$$v^\top\nabla_w^2\left(\int_{\mathbb{R}^M}\widetilde{F}_n(s,u)q_w(du|s)\right)v\geq\left(2-\frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}|x(s)^\top v|_2^2. \tag{19}$$

Integrating (19) over S with respect to ρ_{n+1} gives

$$\begin{aligned}v^\top\nabla_w^2 J_{n+1}(w,\pi^{n+1})v &\geq \left(2-\frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}\int_S|x(s)^\top v|_2^2\rho_{n+1}(ds)\\ &= \left(2-\frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}v^\top\left(\int_S x(s)x(s)^\top\rho_{n+1}(ds)\right)v\\ &\geq (1-\gamma)\left(2-\frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}v^\top\left(\int_S x(s)x(s)^\top\rho(ds)\right)v\\ &\geq (1-\gamma)\left(2-\frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}\lambda_P|v|_2^2,\end{aligned} \tag{20}$$

where we used Lemma 3 in the second inequality and Assumption 3 in the final inequality. Thus (20) implies that the Hessian of $w\mapsto J_{n+1}(w,\pi^{n+1})$ has a minimum eigenvalue bounded from below by $(1-\gamma)\lambda_P(2-\frac{R_{\widehat{Q}}}{\tau})c_{\Sigma,\mathrm{R}}$ for all $n\in\mathbb{N}$, which in turn implies that $w\mapsto J_{n+1}(w,\pi^{n+1})$ is $(1-\gamma)\lambda_P(2-\frac{R_{\widehat{Q}}}{\tau})c_{\Sigma,\mathrm{R}}$-strongly convex for all $n\in\mathbb{N}$ and $w\in\mathfrak{B}(\mathrm{R})$.

Now we turn to analogously demonstrate smoothness. Using the upper bound in (16) in (17), we have

$$\begin{aligned}&v^\top\nabla_w^2\left(\int_{\mathbb{R}^M}\widetilde{F}_n(s,u)q_w(du|s)\right)v\\ &\leq\left(2+\frac{R_{\widehat{Q}}}{\tau}\right)\sum_{i=1}^M(x(s)^\top v)_i^2\int_{\mathbb{R}^M}\left(1-\tanh^2(m_w(s)_i+\sigma_i z_i)\right)q^1(dz).\end{aligned}$$

Since $1-\tanh^2 u_i\leq 1$ and q^1 is a probability measure, it holds that $\int_{\mathbb{R}^M}(1-\tanh^2(m_w(s)_i+\sigma_i z_i))q^1(dz)\leq 1$ for every i. Hence it holds that

$$v^\top\nabla_w^2\left(\int_{\mathbb{R}^M}\widetilde{F}_n(s,u)q_w(du|s)\right)v\leq\left(2+\frac{R_{\widehat{Q}}}{\tau}\right)|x(s)^\top v|_2^2. \tag{21}$$

Integrating (21) over S with respect to ρ_{n+1} gives

$$\begin{aligned}v^\top\nabla_w^2 J_{n+1}(w,\pi^{n+1})v &\leq \left(2+\frac{R_{\widehat{Q}}}{\tau}\right)\int_S|x(s)^\top v|_2^2\rho_{n+1}(ds)\\ &\leq\left(2+\frac{R_{\widehat{Q}}}{\tau}\right)|v|_2^2,\end{aligned}$$

where the final inequality follows from $|x(s)|_{\mathrm{op}} \leq 1$ for all $s \in S$. Thus, the maximum eigenvalue of the Hessian of $w \mapsto J_{n+1}(w, \pi^{n+1})$ is bounded from above by $2 + \frac{R_{\widehat{Q}}}{\tau}$ uniformly over $n \in \mathbb{N}$ and $w \in \mathfrak{B}(\mathrm{R})$. Hence, $w \mapsto J_{n+1}(w, \pi^{n+1})$ is $2 + \frac{R_{\widehat{Q}}}{\tau}$-smooth. For the Gibbs target (4), the same calculations hold with $F_n(s,u) = \frac{1}{\tau}\widehat{Q}^n(s, \tanh u)$. Since $|\mathcal{T}(\widehat{Q}^n(s,\cdot))(a)|_{\mathrm{op}} \leq R_{\widehat{Q}}$, the bound (16) still holds, and the remaining argument gives the same strong-convexity and smoothness constants for $w \mapsto J_{n+1}(w, \pi_G^{n+1})$. □

Remark 1 (Linear Fourier critic)**.** *Suppose that $\widehat{Q}^n(s,a) = (\theta^n)^\top \xi(s,a)$, where $|\theta^n|_2 \leq \mathrm{C}$ and $\xi = (\xi_1, \ldots, \xi_K)$ with $\xi_j(s,a) = \phi_j(s)\cos(\omega_j^\top a)$ with $\omega_j \in \mathbb{R}^M$ and $|\phi_j(s)| \leq 1$. A direct calculation gives that for each $s \in S$ and $a \in A$, $|\nabla_a \xi_j(s,a)|_2 \leq |\omega_j|_2$ and $|\nabla_a^2 \xi_j(s,a)|_{\mathrm{op}} \leq |\omega_j|_2^2$. Therefore it also holds that*

$$\begin{aligned}|\mathcal{T}\xi_j(s,\cdot)(a)|_{\mathrm{op}} &\leq |D(a)^{\frac{1}{2}}|_{\mathrm{op}}^2 |\nabla_a^2 \xi_j(s,a)|_{\mathrm{op}} + 2 \max_{1\leq i\leq M} |a_i \partial_{a_i} \xi_j(s,a)| \\ &\leq |\omega_j|_2^2 + 2|\omega_j|_2.\end{aligned}$$

Hence we have

$$|\mathcal{T}\widehat{Q}^n(s,\cdot)(a)|_{\mathrm{op}} \leq \sum_{j=1}^K |\theta_j^n| |\mathcal{T}\xi_j(s,\cdot)(a)|_{\mathrm{op}} \leq \mathrm{C}\left(\sum_{j=1}^K \left(|\omega_j|_2^2 + 2|\omega_j|_2\right)^2\right)^{\frac{1}{2}}.$$

Thus the conditions of Theorem 2 is satisfied whenever

$$\left(\sum_{j=1}^K (|\omega_j|_2^2 + 2|\omega_j|_2)^2\right)^{\frac{1}{2}} < \frac{2\tau}{\mathrm{C}}.$$

E Proof of Lemma 9

Lemma 9. *Let Assumptions 1 and 3 and the condition of Theorem 2 hold and let $0 < \chi\lambda \leq 1$. Then for all $n \in \mathbb{N}$ it holds that*

$$|w_*^{n+1} - w_*^n|_2^2 \leq C_{\mathit{drift}}\lambda$$

such that $C_{\mathit{drift}} = \frac{1}{\kappa}\left(16\widehat{Q}_{\max} + 4\chi\left(K_{\Sigma,\mathrm{R}} + \frac{2\widehat{Q}_{\max}}{\tau}\right)\right)$ and $\kappa := (1-\gamma)\left(2 - \frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}\lambda_P$.

Proof. Firstly define $D_n(w) := J_{n+1}(w, \pi^{n+1}) - J_n(w, \pi^n)$ for any $w \in \mathfrak{B}(\mathrm{R})$ and observe that by Theorem 2, $w \mapsto J_n(w, \pi^n)$ is κ-strongly convex for all $w \in \mathfrak{B}(R)$ and $n \in \mathbb{N}$, which implies that

$$\frac{\kappa}{2}|w_*^{n+1} - w_*^n|_2^2 \leq J_{n+1}(w_*^n, \pi^{n+1}) - J_{n+1}(w_*^{n+1}, \pi^{n+1}). \tag{22}$$

Moreover since $J_{n+1}(w_*^n, \pi^{n+1}) - J_{n+1}(w_*^{n+1}, \pi^{n+1}) \leq D_n(w_*^n) - D_n(w_*^{n+1})$ for any $n \in \mathbb{N}$, it suffices to bound $D_n(w_*^n) - D_n(w_*^{n+1})$. To that end define

$$l_n(s,a) := \log \frac{d\pi^n}{d\Lambda}(s,a) - \int_A \log \frac{d\pi^n}{d\Lambda}(s,a')\Lambda(da')$$

Observe that by (46), a direct calculation yields

$$l_{n+1}(s,a) = (1-\tau\lambda)l_n(s,a) - \lambda\left(\widehat{Q}^n(s,a) - \int_A \widehat{Q}^n(s,a')\pi^0(da'|s)\right).$$

Subtracting l_n from both sides for each $s \in S$ and $a \in A$ gives

$$l_{n+1}(s,a) - l_n(s,a) = -\tau\lambda l_n(s,a) - \lambda\left(\widehat{Q}^n(s,a) - \int_A \widehat{Q}^n(s,a')\pi^0(da'|s)\right).$$

Taking absolute values and using the triangle inequality, we obtain

$$|l_{n+1}(s,a) - l_n(s,a)| \leq \tau\lambda|l_n(s,a)| + \lambda\left|\widehat{Q}^n(s,a) - \int_A \widehat{Q}^n(s,a')\pi^0(da'|s)\right|. \tag{23}$$

Moreover observe that

$$\begin{aligned}\left|\widehat{Q}^n(s,a) - \int_A \widehat{Q}^n(s,a')\pi^0(da'|s)\right| &\leq |\widehat{Q}^n(s,a)| + \int_A |\widehat{Q}^n(s,a')|\pi^0(da'|s) \\ &\leq 2|\widehat{Q}^n|_{B_b(S\times A)} \\ &\leq 2\widehat{Q}_{\max},\end{aligned}$$

Thus, for all $n \in \mathbb{N}$, (23) becomes

$$|l_{n+1} - l_n|_{B_b(S\times A)} \leq \lambda\left(\tau|l_n|_{B_b(S\times A)} + 2\widehat{Q}_{\max}\right). \tag{24}$$

By (46), $\Lambda(A) = 1$ and the triangle inequality, it holds that

$$|l_n|_{B_b(S\times A)} \leq 2\widehat{Q}_{\max}\lambda\sum_{k=1}^{n}(1-\tau\lambda)^{n-k} \leq \frac{2\widehat{Q}_{\max}}{\tau}.$$

Substituting this into (24), we obtain

$$\begin{aligned}|l_{n+1} - l_n|_{B_b(S\times A)} &\leq \lambda\left(\tau\left(\frac{2\widehat{Q}_{\max}}{\tau}\right) + 2\widehat{Q}_{\max}\right) \\ &= 4\lambda\widehat{Q}_{\max}.\end{aligned} \tag{25}$$

Thus for every $w, w' \in \mathfrak{B}(\mathrm{R})$ and $n \in \mathbb{N}$, by (25) it holds that

$$\begin{aligned}&\left|\left(J_{n+1}(w,\pi^{n+1}) - J_{n+1}(w,\pi^n)\right) - \left(J_{n+1}(w',\pi^{n+1}) - J_{n+1}(w',\pi^n)\right)\right| \\ &= \left|\int_S\int_A \left(l_n(s,a) - l_{n+1}(s,a)\right)(\pi_w - \pi_{w'})(da|s)\rho_{n+1}(ds)\right| \\ &\leq 2\,|l_{n+1} - l_n|_{B_b(S\times A)} \\ &\leq 8\lambda\widehat{Q}_{\max}.\end{aligned} \tag{26}$$

Now recall that by the proof of Theorem 5, for all $n \in \mathbb{N}$ it holds that $\mathrm{KL}(\pi_w(\cdot|s)|\pi^n(\cdot|s)) \leq K_{\Sigma,\mathrm{R}} + \frac{2\widehat{Q}_{\max}}{\tau} =: M_J$ for all $s \in S$. Moreover, by definition of the replay distribution we have $\rho_{n+1} - \rho_n = \chi\lambda(d_\rho^{\pi_{w^n}} - \rho_n)$. Thus it follows that

$$\begin{aligned}&|(J_{n+1}(w,\pi^n) - J_n(w,\pi^n)) - (J_{n+1}(w',\pi^n) - J_n(w',\pi^n))| \\ &= \left|\int_S \mathrm{KL}(\pi_w(\cdot|s)|\pi^n(\cdot|s))(\rho_{n+1} - \rho_n)(ds) - \int_S \mathrm{KL}(\pi_{w'}(\cdot|s)|\pi^n(\cdot|s))(\rho_{n+1} - \rho_n)(ds)\right| \\ &= \chi\lambda\left|\int_S \left(\mathrm{KL}(\pi_w(\cdot|s)|\pi^n(\cdot|s)) - \mathrm{KL}(\pi_{w'}(\cdot|s)|\pi^n(\cdot|s))\right) d_\rho^{\pi_{w^n}}(ds)\right. \\ &\qquad \left. - \int_S \left(\mathrm{KL}(\pi_w(\cdot|s)|\pi^n(\cdot|s)) - \mathrm{KL}(\pi_{w'}(\cdot|s)|\pi^n(\cdot|s))\right)\rho_n(ds)\right| \\ &\leq 2\chi\lambda M_J.\end{aligned} \tag{27}$$

Now define $D_n(w) := J_{n+1}(w,\pi^{n+1}) - J_n(w,\pi^n)$. Combining (26) and (27) yields

$$|D_n(w) - D_n(w')| \leq \lambda\left(8\widehat{Q}_{\max} + 2\chi M_J\right). \tag{28}$$

Now observe that, since $w \mapsto J_{n+1}(w,\pi^{n+1})$ is κ-strongly convex on $\mathfrak{B}(\mathrm{R})$ and w_*^{n+1} minimises $w \mapsto J_{n+1}(w,\pi^{n+1})$ over $\mathfrak{B}(\mathrm{R})$, we have

$$\frac{\kappa}{2}|w_*^{n+1} - w_*^n|_2^2 \leq J_{n+1}(w_*^n,\pi^{n+1}) - J_{n+1}(w_*^{n+1},\pi^{n+1}).$$

Adding and subtracting $J_n(w_*^n,\pi^n)$ and $J_n(w_*^{n+1},\pi^n)$ and using (28) yields

$$J_{n+1}(w_*^n,\pi^{n+1}) - J_{n+1}(w_*^{n+1},\pi^{n+1}) \leq D_n(w_*^n) - D_n(w_*^{n+1}) \leq \lambda\left(8\widehat{Q}_{\max} + 2\chi M_J\right). \tag{29}$$

Combining (22) and (29), and rearranging, concludes the proof. □

F Proof of Lemma 2

Proof. Let $\kappa := (1-\gamma)\left(2 - \frac{R_{\widehat{Q}}}{\tau}\right) c_{\Sigma,\mathrm{R}} \lambda_P$ and $L := 2 + \frac{R_{\widehat{Q}}}{\tau}$. By assumption and Theorem 2, $w \mapsto J_n(w, \pi^n)$ is κ-strongly convex and L-smooth for any $n \in \mathbb{N}$. Now recall that by definition, for any $n \geq 1$,

$$w^n = \mathcal{P}_{\mathfrak{B}(\mathrm{R})}\left(w^{n-1} - h\nabla_w J_n(w^{n-1}, \pi^n)\right).$$

Since $w_*^n = \arg\min_{w \in \mathfrak{B}(\mathrm{R})} J_n(w, \pi^n)$, the first-order optimality condition gives that for all $w \in \mathfrak{B}(\mathrm{R})$

$$\left(\nabla_w J_n(w_*^n, \pi^n)\right)^\top (w - w_*^n) \geq 0. \tag{30}$$

Since $\mathfrak{B}(\mathrm{R})$ is closed and convex, the projection characterisation gives $x = \mathcal{P}_{\mathfrak{B}(\mathrm{R})}(y)$ if and only if $(y-x)^\top(w-x) \leq 0$ for all $w \in \mathfrak{B}(\mathrm{R})$. Applying this with $x = w_*^n$ and $y = w_*^n - h\nabla_w J_n(w_*^n, \pi^n)$, (30) implies that

$$w_*^n = \mathcal{P}_{\mathfrak{B}(\mathrm{R})}\left(w_*^n - h\nabla_w J_n(w_*^n, \pi^n)\right).$$

Therefore, it holds that

$$\begin{aligned}
|w^n - w_*^n|_2^2 &= \left|\mathcal{P}_{\mathfrak{B}(\mathrm{R})}\left(w^{n-1} - h\nabla_w J_n(w^{n-1}, \pi^n)\right) - w_*^n\right|_2^2 \qquad (31)\\
&= \left|\mathcal{P}_{\mathfrak{B}(\mathrm{R})}\left(w^{n-1} - h\nabla_w J_n(w^{n-1}, \pi^n)\right) - \mathcal{P}_{\mathfrak{B}(\mathrm{R})}\left(w_*^n - h\nabla_w J_n(w_*^n, \pi^n)\right)\right|_2^2 \\
&\leq \left|w^{n-1} - w_*^n - h\left(\nabla_w J_n(w^{n-1}, \pi^n) - \nabla_w J_n(w_*^n, \pi^n)\right)\right|_2^2 \\
&= |w^{n-1} - w_*^n|_2^2 - 2h\left(\nabla_w J_n(w^{n-1}, \pi^n) - \nabla_w J_n(w_*^n, \pi^n)\right)^\top (w^{n-1} - w_*^n) \\
&\quad + h^2\left|\nabla_w J_n(w^{n-1}, \pi^n) - \nabla_w J_n(w_*^n, \pi^n)\right|_2^2,
\end{aligned}$$

where in the first inequality we used the non-expansiveness of the projection operator. By Nesterov (2004, Theorem 2.1.12), it holds that

$$\begin{aligned}
&\left(\nabla_w J_n(w, \pi^n) - \nabla_w J_n(\bar{w}, \pi^n)\right)^\top (w - \bar{w}) \qquad (32)\\
&\quad \geq \frac{\kappa L}{\kappa + L}|w - \bar{w}|_2^2 + \frac{1}{\kappa + L}\left|\nabla_w J_n(w, \pi^n) - \nabla_w J_n(\bar{w}, \pi^n)\right|_2^2.
\end{aligned}$$

Applying (32) with $w = w^{n-1}$ and $\bar{w} = w_*^n$ and substituting into the right hand side of (31), we have

$$\begin{aligned}
|w^n - w_*^n|_2^2 &\leq |w^{n-1} - w_*^n|_2^2 - \frac{2h\kappa L}{\kappa + L}|w^{n-1} - w_*^n|_2^2 \qquad (33)\\
&\quad + \left(h^2 - \frac{2h}{\kappa + L}\right)\left|\nabla_w J_n(w^{n-1}, \pi^n) - \nabla_w J_n(w_*^n, \pi^n)\right|_2^2 \\
&= \left(1 - \frac{2h\kappa L}{\kappa + L}\right)|w^{n-1} - w_*^n|_2^2 \\
&\quad + \left(h^2 - \frac{2h}{\kappa + L}\right)\left|\nabla_w J_n(w^{n-1}, \pi^n) - \nabla_w J_n(w_*^n, \pi^n)\right|_2^2.
\end{aligned}$$

Since $hL < 1$ and $\kappa \leq L$, it holds that $\frac{2Lh\kappa}{\kappa+L} \geq h\kappa$ and

$$h^2 - \frac{2h}{\kappa + L} \leq \frac{h}{L} - \frac{2h}{\kappa + L} = h\frac{\kappa - L}{L(\kappa + L)} \leq 0.$$

Hence, after dropping the nonpositive gradient term in (33) and using that $1 - \frac{2Lh\kappa}{\kappa+L} \leq 1 - h\kappa$, we arrive at

$$\begin{aligned}
|w^n - w_*^n|_2^2 &\leq \left(1 - \frac{2h\kappa L}{\kappa + L}\right)|w^{n-1} - w_*^n|_2^2 \\
&\leq (1 - h\kappa)|w^{n-1} - w_*^n|_2^2.
\end{aligned}$$

Now adding and subtracting w_*^{n-1} on the right hand side and using Young's inequality for any $\delta > 0$, it holds that

$$|w^n - w_*^n|_2^2 \leq (1+\delta)(1 - h\kappa)|w^{n-1} - w_*^{n-1}|_2^2 + \left(1 + \frac{1}{\delta}\right)(1 - h\kappa)\,|w_*^n - w_*^{n-1}|_2^2.$$

By Lemma 9, for every $n \geq 1$ we have

$$|w^n - w_*^n|_2^2 \leq (1+\delta)(1-h\kappa)|w^{n-1} - w_*^{n-1}|_2^2 + \lambda C_{\text{drift}} \left(1 + \frac{1}{\delta}\right)(1-h\kappa)\,. \tag{34}$$

Now choose $\delta := \frac{h\kappa}{2(1-h\kappa)}$, which is well-defined because $0 < h\kappa \leq hL < 1$. Then (34) becomes

$$\begin{aligned} |w^n - w_*^n|_2^2 &\leq \left(1 + \frac{h\kappa}{2(1-h\kappa)}\right)(1-h\kappa)|w^{n-1} - w_*^{n-1}|_2^2 + \lambda C_{\text{drift}} \left(1 + \frac{2(1-h\kappa)}{h\kappa}\right)(1-h\kappa) \\ &\leq \left(1 - \frac{h\kappa}{2}\right)|w^{n-1} - w_*^{n-1}|_2^2 + \frac{2\lambda C_{\text{drift}}}{h\kappa}. \end{aligned} \tag{35}$$

where in the last inequality we used that $\left(1 + \frac{2(1-h\kappa)}{h\kappa}\right)(1-h\kappa) \leq \frac{2}{h\kappa}$. Iterating (35), for any $k \in \mathbb{N}$ gives

$$\begin{aligned} |w^k - w_*^k|_2^2 &\leq \left(1 - \frac{h\kappa}{2}\right)^k |w^0 - w_*^0|_2^2 + \frac{2\lambda C_{\text{drift}}}{h\kappa} \sum_{j=0}^{k-1} \left(1 - \frac{h\kappa}{2}\right)^j \\ &\leq \left(1 - \frac{h\kappa}{2}\right)^k |w^0 - w_*^0|_2^2 + \frac{4\lambda C_{\text{drift}}}{h^2\kappa^2}. \end{aligned} \tag{36}$$

To ease notation, let $C_{\text{init}} := \frac{2|w^0 - w_*^0|_2^2}{\kappa}$ and $C_{\text{track}} := \frac{4C_{\text{drift}}}{\kappa^2}$. Averaging (36) over $k = 0, \ldots, n-1$, we obtain

$$\begin{aligned} \frac{1}{n}\sum_{k=0}^{n-1} |w^k - w_*^k|_2^2 &\leq \frac{|w^0 - w_*^0|_2^2}{n} \sum_{k=0}^{n-1} \left(1 - \frac{h\kappa}{2}\right)^k + \frac{\lambda C_{\text{track}}}{h^2} \\ &\leq \frac{C_{\text{init}}}{hn} + \frac{\lambda C_{\text{track}}}{h^2}. \end{aligned} \tag{37}$$

Now fix $k \in \{0, \ldots, n-1\}$. Since $w \mapsto J_k(w, \pi^k)$ is L-smooth on $\mathfrak{B}(\mathrm{R})$ by Theorem 2 for any $k \in \mathbb{N}$ we have

$$J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) \leq \left(\nabla_w J_k(w_*^k, \pi^k)\right)^\top (w^k - w_*^k) + \frac{L}{2}|w^k - w_*^k|_2^2. \tag{38}$$

By Lemma 8 for every $k \in \mathbb{N}$ and every $w \in \mathfrak{B}(\mathrm{R})$ it holds that

$$\left|\nabla_w J_k(w, \pi^k)\right|_2 \leq 2\sqrt{M} + \sqrt{\frac{2}{\pi}}\frac{\widehat{Q}_{\max}}{\tau}\sum_{i=1}^{M}\frac{1}{\sigma_i} := C_{\text{grad}}$$

Moreover, since $w^k, w_*^k \in \mathfrak{B}(\mathrm{R})$, it holds that $|w^k - w_*^k|_2 \leq 2\mathrm{R}$, which in turn implies that $|w^k - w_*^k|_2^2 \leq 2\mathrm{R}|w^k - w_*^k|_2$. Using this bound in (38) we have

$$\begin{aligned} J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k) &\leq C_{\text{grad}}|w^k - w_*^k|_2 + L\mathrm{R}|w^k - w_*^k|_2 \\ &= (C_{\text{grad}} + L\mathrm{R})|w^k - w_*^k|_2. \end{aligned} \tag{39}$$

Since w_*^k minimises $w \mapsto J_k(w, \pi^k)$ over $\mathfrak{B}(\mathrm{R})$ and $w^k \in \mathfrak{B}(\mathrm{R})$, the left hand side of (39) is nonnegative. Define $C_3 := (C_{\text{grad}} + L\mathrm{R})^{\frac{1}{2}}$. Then taking the square root of (39), averaging and using Hölder's inequality, it holds that

$$\begin{aligned} \frac{1}{n}\sum_{k=0}^{n-1}\left(J_k(w^k, \pi^k) - J_k(w_*^k, \pi^k)\right)^{\frac{1}{2}} &\leq \frac{C_3}{n}\sum_{k=0}^{n-1}|w^k - w_*^k|_2^{\frac{1}{2}} \\ &\leq C_3\left(\frac{1}{n}\sum_{k=0}^{n-1}|w^k - w_*^k|_2^2\right)^{\frac{1}{4}}. \end{aligned}$$

Substituting (37), we obtain

$$\frac{1}{n}\sum_{k=0}^{n-1}\left(J_k(w^k,\pi^k)-J_k(w_*^k,\pi^k)\right)^{\frac{1}{2}} \le C_3\left(\frac{C_{\text{init}}}{hn}+\frac{\lambda C_{\text{track}}}{h^2}\right)^{\frac{1}{4}}$$
$$\le C_3\max\left\{C_{\text{init}},C_{\text{track}}\right\}^{\frac{1}{4}}\left(\frac{1}{hn}+\frac{\lambda}{h^2}\right)^{\frac{1}{4}}.$$

Thus the result follows with $C_1 := C_3\max\left\{C_{\text{init}},C_{\text{track}}\right\}^{\frac{1}{4}}$. □

G Proof of Theorem 3

Proof. The proof is a direct calculation. Let $\vartheta := \min\left\{\frac{1}{\tau},\frac{1}{\chi}\right\}$, $\lambda = \frac{\vartheta}{2N^{\frac{4}{5}}}$ and $0 < h < \frac{\tau}{2\tau+R_{\widehat{Q}}}$. Since $N \ge 1$ we have $0 < \tau\lambda \le \frac{1}{2}$ and $0 < \chi\lambda \le \frac{1}{2}$. Therefore by Theorem 1 and Lemma 2, there exists $\mathrm{C} < \infty$, independent of N, such that

$$\begin{aligned}
&\min_{0\le r\le N-1} V_\tau^{\pi_{w^r}}(\rho) - V_\tau^{\pi^*}(\rho)\\
&\le \mathrm{C}\left(\frac{1}{\lambda N}+\lambda+\left(\frac{1}{hN}+\frac{\lambda}{h^2}\right)^{\frac{1}{4}}+\varepsilon_{\text{approx}}(N)\right)\\
&= \mathrm{C}\left(\frac{2}{\vartheta}N^{-\frac{1}{5}}+\frac{\vartheta}{2}N^{-\frac{4}{5}}+N^{-\frac{1}{5}}\left(\frac{1}{hN^{\frac{1}{5}}}+\frac{\vartheta}{2h^2}\right)^{\frac{1}{4}}+\varepsilon_{\text{approx}}(N)\right)\\
&\le \mathrm{C}\left(\left(\frac{2}{\vartheta}+\frac{\vartheta}{2}+\left(\frac{1}{h}+\frac{\vartheta}{2h^2}\right)^{\frac{1}{4}}\right)N^{-\frac{1}{5}}+\varepsilon_{\text{approx}}(N)\right)\\
&\le C\left(N^{-\frac{1}{5}}+\varepsilon_{\text{approx}}(N)\right),
\end{aligned}$$

where

$$C := \mathrm{C}\max\left\{\frac{2}{\vartheta}+\frac{\vartheta}{2}+\left(\frac{1}{h}+\frac{\vartheta}{2h^2}\right)^{\frac{1}{4}},1\right\}.$$

□

H Proof of Corollary 1

Proof. Firstly recall that Theorem 2 holds in the case where the target policy is π_G^{n+1} for all $n \in \mathbb{N}$. Therefore $w \mapsto J_n(w,\pi_G^n)$ is $(1-\gamma)\left(2-\frac{R_{\widehat{Q}}}{\tau}\right)c_{\Sigma,\mathrm{R}}\lambda_P$-strongly convex and $2+\frac{R_{\widehat{Q}}}{\tau}$-smooth for all $w \in \mathfrak{B}(\mathrm{R})$ and $n \ge 1$. Moreover, it also directly holds that

$$\left|\log\frac{d\pi_{\mathrm{G}}^k}{d\Lambda}\right|_{B_b(S\times A)} \le \frac{2\widehat{Q}_{\max}}{\tau},$$

and consequently by (42) we have

$$\begin{aligned}
\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi_{\mathrm{G}}^n(\cdot|s)) &\le K_{\Sigma,\mathrm{R}} + \left|\log\frac{d\pi_G^n}{d\Lambda}\right|_{B_b(S\times A)}\\
&\le K_{\Sigma,\mathrm{R}} + \frac{2\widehat{Q}_{\max}}{\tau}\\
&:= M_{\mathrm{G}}
\end{aligned}$$

Therefore, analogously to the proof of Lemma 2, let

$$D_k(w) := J_{k+1}(w,\pi_{\mathrm{G}}^{k+1}) - J_k(w,\pi_{\mathrm{G}}^k),$$

then for all $k \geq 1$ it holds that

$$\begin{aligned}
&\left|\left(J_{k+1}(w, \pi_{\mathrm{G}}^{k+1}) - J_{k+1}(w, \pi_{\mathrm{G}}^{k})\right) - \left(J_{k+1}(w', \pi_{\mathrm{G}}^{k+1}) - J_{k+1}(w', \pi_{\mathrm{G}}^{k})\right)\right| \\
&= \frac{1}{\tau}\left|\int_S \int_A \left(\widehat{Q}^k(s,a) - \widehat{Q}^{k-1}(s,a)\right)(\pi_w - \pi_{w'})(da|s)\rho_{k+1}(ds)\right| \\
&\leq \frac{4\widehat{Q}_{\max}}{\tau}.
\end{aligned}$$

Moreover, using $\rho_{k+1} - \rho_k = \chi\lambda(d_\rho^{\pi_{w^k}} - \rho_k)$ exactly as in (27) gives

$$\left|\left(J_{k+1}(w, \pi_{\mathrm{G}}^{k}) - J_{k}(w, \pi_{\mathrm{G}}^{k})\right) - \left(J_{k+1}(w', \pi_{\mathrm{G}}^{k}) - J_{k}(w', \pi_{\mathrm{G}}^{k})\right)\right| \leq 2\chi\lambda M_{\mathrm{G}}.$$

Therefore it also holds that

$$|D_k(w) - D_k(w')| \leq \frac{4\widehat{Q}_{\max}}{\tau} + 2\chi\lambda M_{\mathrm{G}}.$$

Now by strong convexity and optimality, we have

$$\begin{aligned}
\frac{\kappa}{2}|w_{\mathrm{G},*}^{k+1} - w_{\mathrm{G},*}^{k}|_2^2 &\leq J_{k+1}(w_{\mathrm{G},*}^{k}, \pi_{\mathrm{G}}^{k+1}) - J_{k+1}(w_{\mathrm{G},*}^{k+1}, \pi_{\mathrm{G}}^{k+1}) \\
&\leq D_k(w_{\mathrm{G},*}^{k}) - D_k(w_{\mathrm{G},*}^{k+1}) \\
&\leq \frac{4\widehat{Q}_{\max}}{\tau} + 2\chi\lambda M_{\mathrm{G}}.
\end{aligned}$$

Hence after rearranging it holds that

$$|w_{\mathrm{G},*}^{k+1} - w_{\mathrm{G},*}^{k}|_2^2 \leq C_{\mathrm{drift}}\left(\frac{1}{\tau} + \chi\lambda\right),$$

where $C_{\mathrm{drift}} := \frac{1}{\kappa}\max\left\{8\widehat{Q}_{\max}, 4M_{\mathrm{G}}\right\}$. Now by (32) and (33), for all $k \geq 1$ it holds that

$$|w^{k+1} - w_{\mathrm{G},*}^{k+1}|_2^2 \leq \left(1 - \frac{h\kappa}{2}\right)|w^k - w_{\mathrm{G},*}^{k}|_2^2 + \frac{2C_{\mathrm{drift}}}{h\kappa}\left(\frac{1}{\tau} + \chi\lambda\right).$$

Since $|w^1 - w_{\mathrm{G},*}^1|_2 \leq 2\mathrm{R}$, iterating and averaging yields

$$\frac{1}{n}\sum_{k=1}^{n-1}|w^k - w_{\mathrm{G},*}^{k}|_2^2 \leq \frac{8\mathrm{R}^2}{h\kappa n} + \frac{4C_{\mathrm{drift}}}{h^2\kappa^2}\left(\frac{1}{\tau} + \chi\lambda\right).$$

Moreover, by Lemma 8, for every $w \in \mathfrak{B}(\mathrm{R})$ and $k \geq 1$ it holds that

$$|\nabla_w J_k(w, \pi_{\mathrm{G}}^{k})|_2 \leq C_{\text{grad}} := 2\sqrt{M} + \sqrt{\frac{2}{\pi}}\frac{\widehat{Q}_{\max}}{\tau}\sum_{i=1}^{M}\frac{1}{\sigma_i}.$$

Thus, by L-smoothness and Hölder's inequality,

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n-1}\left(J_k(w^k, \pi_{\mathrm{G}}^{k}) - J_k(w_{\mathrm{G},*}^{k}, \pi_{\mathrm{G}}^{k})\right)^{\frac{1}{2}} \\
&\leq (C_{\text{grad}} + L\mathrm{R})^{\frac{1}{2}}\left(\frac{1}{n}\sum_{k=1}^{n-1}|w^k - w_{\mathrm{G},*}^{k}|_2^2\right)^{\frac{1}{4}} \\
&\leq C\left(\frac{1}{hn} + \frac{1}{h^2}\left(\frac{1}{\tau} + \chi\lambda\right)\right)^{\frac{1}{4}},
\end{aligned}$$

where

$$C := (C_{\text{grad}} + L\mathrm{R})^{\frac{1}{2}}\max\left\{\frac{8\mathrm{R}^2}{\kappa}, \frac{4C_{\mathrm{drift}}}{\kappa^2}\right\}^{\frac{1}{4}}.$$

□

I AUXILIARY RESULTS

I.1 PROOF OF LEMMA 5

Lemma 5 (Entropy of a squashed Gaussian policy)**.** *For every $s \in S$ and every $w \in \mathfrak{B}(\mathrm{R})$, it holds that*

$$\begin{aligned}\mathrm{H}(\pi_w(\cdot|s)) &= \log\left(\frac{(2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}}{2^M}\right) + \frac{M}{2} - 2\int_{\mathbb{R}^M}\sum_{i=1}^{M}\log\cosh(u_i)q_w(du|s) \qquad (40)\\ &\geq -2\sqrt{M}\,\mathrm{R} - 2\sqrt{\frac{2}{\pi}}\sum_{i=1}^{M}\sigma_i - \left|\log\left(\frac{(2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}}{2^M}\right) + \frac{M}{2}\right|\\ &=: -K_{\Sigma,\mathrm{R}}.\end{aligned}$$

Proof. Since the Jacobian of $u \mapsto \tanh u$ is $D(\tanh u)$, the change-of-variables formula gives, for every $a \in A$ with $u = \tanh^{-1} a$,

$$\frac{d\pi_w}{d\Lambda}(s,a) = \frac{2^M}{(2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}}\frac{1}{\det D(a)}\exp\left(-\frac{1}{2}(u-m_w(s))^\top\Sigma^{-1}(u-m_w(s))\right).$$

Therefore, by Definition 1 it holds that

$$\begin{aligned}&\mathrm{H}(\pi_w(\cdot|s))\\ &= -\int_{\mathbb{R}^M}\left(M\log 2 - \log\big((2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}\big) - \frac{1}{2}(u-m_w(s))^\top\Sigma^{-1}(u-m_w(s))\right)q_w(du|s)\\ &\quad + \int_{\mathbb{R}^M}\sum_{i=1}^{M}\log(1-\tanh^2 u_i)q_w(du|s)\\ &= \log\left(\frac{(2\pi)^{\frac{M}{2}}\sqrt{\det\Sigma}}{2^M}\right) + \frac{M}{2} - 2\int_{\mathbb{R}^M}\sum_{i=1}^{M}\log\cosh(u_i)q_w(du|s),\end{aligned}$$

where the second equality uses $1-\tanh^2 u_i = \cosh^{-2} u_i$ and $\int_{\mathbb{R}^M}(u-m_w(s))^\top\Sigma^{-1}(u-m_w(s))q_w(du|s) = M$. Finally, $\log\cosh u_i \leq |u_i|$ and

$$\begin{aligned}\int_{\mathbb{R}^M}|u|_1 q_w(du|s) &\leq |m_w(s)|_1 + \int_{\mathbb{R}^M}|u-m_w(s)|_1 q_w(du|s)\\ &= |m_w(s)|_1 + \sqrt{\frac{2}{\pi}}\sum_{i=1}^{M}\sigma_i \leq \sqrt{M}\,\mathrm{R} + \sqrt{\frac{2}{\pi}}\sum_{i=1}^{M}\sigma_i,\end{aligned}$$

where $|m_w(s)|_1 \leq \sqrt{M}\,|x(s)^\top w|_2 \leq \sqrt{M}\,\mathrm{R}$. Substitution into (40) concludes the proof. □

I.2 PROOF OF THEOREM 5

Theorem 5. *Let Assumption 1 hold and suppose that $0 < \tau\lambda < 1$. Then there exists a non negative constant $C_{\mathrm{KL}} < \infty$ such that for all $n \in \mathbb{N}$ and $s \in S$ it holds that*

$$|\mathrm{H}(\pi_{w^n}(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))| \leq C_{\mathrm{KL}}\left(\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))\right)^{\frac{1}{2}}.$$

Proof. Firstly observe that, by definition of the entropy, for all $s \in S$ and for all $n \geq 0$ we have

$$\begin{aligned}&\mathrm{H}(\pi_{w^n}(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))\\ &= -\int_A \log\frac{d\pi_{w^n}}{d\Lambda}(s,a)\pi_{w^n}(da|s) + \int_A \log\frac{d\pi^n}{d\Lambda}(s,a)\pi^n(da|s)\\ &= -\int_A \log\left(\frac{d\pi_{w^n}}{d\pi^n}(s,a)\frac{d\pi^n}{d\Lambda}(s,a)\right)\pi_{w^n}(da|s) + \int_A \log\frac{d\pi^n}{d\Lambda}(s,a)\pi^n(da|s)\\ &= -\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)) - \int_A \log\frac{d\pi^n}{d\Lambda}(s,a)\left(\pi_{w^n}-\pi^n\right)(da|s).\end{aligned}$$

Therefore, by the triangle inequality, it holds that

$$\begin{aligned} &|\mathrm{H}(\pi_{w^n}(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))| \\ &\leq \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)) + \left|\int_A \log \frac{d\pi^n}{d\Lambda}(s,a)\left(\pi_{w^n} - \pi^n\right)(da|s)\right|. \end{aligned} \tag{41}$$

Furthermore recall the standard identity

$$\begin{aligned} \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)) &= \int_A \log \frac{d\pi_{w^n}}{d\pi^n}(s,a)\pi_{w^n}(da|s) \\ &= \int_A \log \frac{d\pi_{w^n}}{d\Lambda}(s,a)\pi_{w^n}(da|s) - \int_A \log \frac{d\pi^n}{d\Lambda}(s,a)\pi_{w^n}(da|s) \\ &\leq -\mathrm{H}(\pi_{w^n}(\cdot|s)) + \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)}. \end{aligned}$$

Since $\pi_{w^n} \in \Pi_\Sigma$ for all $n \geq 0$, Lemma 5 provides the following bound

$$-\mathrm{H}(\pi_{w^n}(\cdot|s)) \leq K_{\Sigma,\mathrm{R}}.$$

Therefore, it holds that

$$\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}} \leq \left(K_{\Sigma,\mathrm{R}} + \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)}\right)^{\frac{1}{2}}. \tag{42}$$

Multiplying both sides by $(\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)))^{\frac{1}{2}} \geq 0$ we have

$$\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)) \leq \left(K_{\Sigma,\mathrm{R}} + \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)}\right)^{\frac{1}{2}} \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}}. \tag{43}$$

Using Pinsker's inequality we bound the second term of (41) by

$$\begin{aligned} \left|\int_A \log \frac{d\pi^n}{d\Lambda}(s,a)\left(\pi_{w^n} - \pi^n\right)(da|s)\right| &\leq \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)} |\pi_{w^n}(\cdot|s) - \pi^n(\cdot|s)|_{\mathcal{M}(A)} \\ &\leq \sqrt{2}\left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)} (\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)))^{\frac{1}{2}}. \end{aligned} \tag{44}$$

Substituting (43) and (44) into (41), we arrive at

$$\begin{aligned} &|\mathrm{H}(\pi_{w^n}(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))| \\ &\leq \left(\left(K_{\Sigma,\mathrm{R}} + \left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)}\right)^{\frac{1}{2}} + \sqrt{2}\left|\log \frac{d\pi^n}{d\Lambda}\right|_{B_b(S\times A)}\right) (\mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s)))^{\frac{1}{2}}. \end{aligned} \tag{45}$$

We now prove that the log densities are uniformly bounded. To that end, by iterating (5) from $\pi^0 = \Lambda$ we have

$$\frac{d\pi^{n+1}}{d\Lambda}(s,a) = \frac{1}{\overline{Z}_{n+1}(s)} \exp\left(-\lambda \sum_{k=1}^{n+1} (1-\tau\lambda)^{n+1-k} \widehat{Q}^{k-1}(s,a)\right) \tag{46}$$

with $\overline{Z}_{n+1}(s) = \int_A \exp\left(-\lambda \sum_{k=1}^{n+1}(1-\tau\lambda)^{n+1-k}\widehat{Q}^{k-1}(s,a)\right)\Lambda(da)$. Therefore, by assumption 1 we have $\left|\widehat{Q}^n\right|_{B_b(S\times A)} \leq \widehat{Q}_{\max}$ thus it also holds that

$$\begin{aligned} \left|\log \frac{d\pi^n}{d\Lambda}(s,a)\right| &= \left|-\log \overline{Z}_n(s) - \lambda \sum_{k=1}^{n} (1-\tau\lambda)^{n-k}\widehat{Q}^{k-1}(s,a)\right| \\ &\leq \left|\log \overline{Z}_n(s)\right| + \left|\lambda \sum_{k=1}^{n}(1-\tau\lambda)^{n-k}\widehat{Q}^{k-1}(s,a)\right| \\ &\leq \frac{2\widehat{Q}_{\max}}{\tau} \end{aligned} \tag{47}$$

Therefore, substituting this bound into (45) gives

$$\begin{aligned}
&|\mathrm{H}(\pi_{w^n}(\cdot|s)) - \mathrm{H}(\pi^n(\cdot|s))| \\
&\leq \left(\left(K_{\Sigma,\mathrm{R}} + \frac{2\widehat{Q}_{\max}}{\tau}\right)^{\frac{1}{2}} + \sqrt{2}\left(\frac{2\widehat{Q}_{\max}}{\tau}\right)\right) \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}} \\
&= C_{\mathrm{KL}}\, \mathrm{KL}(\pi_{w^n}(\cdot|s)|\pi^n(\cdot|s))^{\frac{1}{2}},
\end{aligned}$$

with

$$C_{\mathrm{KL}} := \left(\left(K_{\Sigma,\mathrm{R}} + \frac{2\widehat{Q}_{\max}}{\tau}\right)^{\frac{1}{2}} + \sqrt{2}\left(\frac{2\widehat{Q}_{\max}}{\tau}\right)\right).$$

□

I.3 Proof of Lemma 7

Lemma 7. *Let $\rho \in \mathcal{P}(S)$ and suppose that there exists a constant $A_2 < \infty$ such that, for every $s \in S$ and $a \in A$,*

$$\left|\frac{dP}{d\rho}(s,a,\cdot)\right|_{L^2(\rho)} \leq A_2.$$

Then for any $\pi \in \mathcal{P}(A|S)$ it holds that

$$\left|\frac{dd^\pi_\rho}{d\rho}\right|_{L^2(\rho)} \leq (1-\gamma) + \gamma A_2$$

Proof. Firstly we seek to identify the density of P_π with respect to ρ for any $\pi \in \mathcal{P}(A|S)$. To that end, for any $B \in \mathcal{B}(S)$ and any $s \in S$ we have

$$\begin{aligned}
P_\pi(B|s) &= \int_A P(B|s,a)\pi(da|s) \\
&= \int_A \int_B \frac{dP}{d\rho}(s,a,s')\rho(ds')\pi(da|s) \\
&= \int_B \left(\int_A \frac{dP}{d\rho}(s,a,s')\pi(da|s)\right)\rho(ds').
\end{aligned}$$

Since this holds for all $B \in \mathcal{B}(S)$, we have

$$\frac{dP_\pi}{d\rho}(s,s') = \int_A \frac{dP}{d\rho}(s,a,s')\pi(da|s).$$

Therefore, by Minkowski's integral inequality it holds that for all $s \in S$,

$$\begin{aligned}
\left|\frac{dP_\pi}{d\rho}(s,\cdot)\right|_{L^2(\rho)} &= \left(\int_S \left|\int_A \frac{dP}{d\rho}(s,a,s')\pi(da|s)\right|^2 \rho(ds')\right)^{\frac{1}{2}} \\
&\leq \int_A \left(\int_S \left|\frac{dP}{d\rho}(s,a,s')\right|^2 \rho(ds')\right)^{\frac{1}{2}} \pi(da|s) \\
&\leq A_2.
\end{aligned} \tag{48}$$

We now prove a bound that holds for any initial state distribution $\rho \in \mathcal{P}(S)$. Fix $n \geq 1$. We wish to identify the density of $\int_S P^n_\pi(\cdot|s)\rho(ds)$ with respect to ρ. To that end, observe that for any

$B \in \mathcal{B}(S)$,

$$\begin{aligned}\int_S P_\pi^n(B|s)\rho(ds) &= \int_S \int_S P_\pi(B|s')P_\pi^{n-1}(ds'|s)\rho(ds) \\ &= \int_S P_\pi(B|s') \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right) \\ &= \int_S \int_B \frac{dP_\pi}{d\rho}(s', s'')\rho(ds'') \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right) \\ &= \int_B \left(\int_S \frac{dP_\pi}{d\rho}(s', s'') \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right) \right) \rho(ds'').\end{aligned}$$

Since this holds for all $B \in \mathcal{B}(S)$, it follows that

$$\frac{d\left(\int_S P_\pi^n(\cdot|s)\rho(ds)\right)}{d\rho}(s'') = \int_S \frac{dP_\pi}{d\rho}(s', s'') \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right).$$

Consequently, by Minkowski's integral inequality and (48), it holds that

$$\begin{aligned}\left| \frac{d\left(\int_S P_\pi^n(\cdot|s)\rho(ds)\right)}{d\rho} \right|_{L^2(\rho)} &= \left(\int_S \left| \int_S \frac{dP_\pi}{d\rho}(s', s'') \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right) \right|^2 \rho(ds'') \right)^{\frac{1}{2}} \\ &\leq \int_S \left| \frac{dP_\pi}{d\rho}(s', \cdot) \right|_{L^2(\rho)} \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right) \\ &\leq A_2 \int_S \left(\int_S P_\pi^{n-1}(ds'|s)\rho(ds) \right) \\ &= A_2,\end{aligned}$$

where in the final equality we used that $\int_S P_\pi^{n-1}(ds'|s)\rho(ds)$ is a probability measure on S. Now recall that by definition, we have

$$d_\rho^\pi(ds') = (1-\gamma) \sum_{n=0}^{\infty} \gamma^n \int_S P_\pi^n(ds'|s)\rho(ds),$$

with $P_\pi^0(ds'|s) = \delta_s(ds')$. Hence applying the triangle inequality in $L^2(\rho)$ to the partial sums and then taking the limit, we obtain

$$\begin{aligned}\left| \frac{dd_\rho^\pi}{d\rho} \right|_{L^2(\rho)} &\leq (1-\gamma) \left| \frac{d\rho}{d\rho} \right|_{L^2(\rho)} + (1-\gamma) \sum_{n=1}^{\infty} \gamma^n \left| \frac{d\left(\int_S P_\pi^n(\cdot|s)\rho(ds)\right)}{d\rho} \right|_{L^2(\rho)} \\ &\leq (1-\gamma) + (1-\gamma) \sum_{n=1}^{\infty} \gamma^n A_2 \\ &= (1-\gamma) + \gamma A_2.\end{aligned}$$

□

I.4 Proof of Lemma 8

Lemma 8 (Uniform boundedness of the actor gradient)**.** *Let Assumption 1 hold. Then for all $n \in \mathbb{N}$ and all $w \in \mathfrak{B}(\mathrm{R})$, it holds that*

$$\left| \nabla_w J_{n+1}(w, \pi^{n+1}) \right|_2 \leq 2\sqrt{M} + \sqrt{\frac{2}{\pi}} \frac{\widehat{Q}_{\max}}{\tau} \sum_{i=1}^{M} \frac{1}{\sigma_i}.$$

Moreover, for any $n \in \mathbb{N}$, the same bounds holds for $\left| \nabla_w J_{n+1}(w, \pi_G^{n+1}) \right|_2$.

Proof. Firstly observe that for every $n \in \mathbb{N}$ and every $w \in \mathfrak{B}(\mathrm{R})$, we can write

$$\begin{aligned}
J_{n+1}(w, \pi^{n+1}) &= \int_S \mathrm{KL}(\pi_w(\cdot|s)|\pi^{n+1}(\cdot|s))\rho_{n+1}(ds) \\
&= \int_S \int_A \log \frac{d\pi_w}{d\pi^{n+1}}(s,a)\pi_w(da|s)\rho_{n+1}(ds) \\
&= \int_S \int_A \log \left(\frac{d\pi_w}{d\Lambda}(s,a) \frac{d\Lambda}{d\pi^{n+1}}(s,a) \right) \pi_w(da|s)\rho_{n+1}(ds) \\
&= -\int_S \mathrm{H}(\pi_w(\cdot|s))\rho_{n+1}(ds) - \int_S \int_A \log \frac{d\pi^{n+1}}{d\Lambda}(s,a)\pi_w(da|s)\rho_{n+1}(ds).
\end{aligned}$$

We now bound the gradients of these two terms separately. For the first term, using Lemma 5, for every $s \in S$ we have

$$\begin{aligned}
-\mathrm{H}(\pi_w(\cdot|s)) &= 2\int_{\mathbb{R}^M} \sum_{i=1}^M \log\cosh(u_i) q_w(du|s) - \log\left(\frac{(2\pi)^{\frac{M}{2}}\sqrt{\det \Sigma}}{2^M} \right) - \frac{M}{2} \\
&= 2\int_{\mathbb{R}^M} \sum_{i=1}^M \log\cosh(m_w(s)_i + \sigma_i z_i) q^1(dz) - \log\left(\frac{(2\pi)^{\frac{M}{2}}\sqrt{\det \Sigma}}{2^M} \right) - \frac{M}{2}.
\end{aligned}$$

Therefore, differentiating under the integral, it follows that

$$\nabla_w \mathrm{H}(\pi_w(\cdot|s)) = -2\int_{\mathbb{R}^M} x(s)\tanh(m_w(s) + \Sigma^{\frac{1}{2}} z) q^1(dz).$$

Hence, using $|\tanh u|_2 \le \sqrt{M}$ and $|x(s)|_{\mathrm{op}} \le 1$ for all $s \in S$, it holds that

$$|\nabla_w \mathrm{H}(\pi_w(\cdot|s))|_2 \le 2\sqrt{M}. \tag{49}$$

Before proceeding with the second term, let us simplify the notation with

$$F_n(s,u) := \lambda \sum_{k=0}^n (1-\tau\lambda)^{n-k} \widehat{Q}^k(s, \tanh u),$$

and thus recall that

$$\frac{d\pi^{n+1}}{d\Lambda}(s, \tanh u) = \frac{1}{\overline{Z}_{n+1}(s)} \exp\left(-F_n(s,u)\right).$$

Let $\Lambda_{\mathbb{R}^M}$ denote Lebesgue measure on $\mathbb{R}^M$. By Definition 1 it holds that

$$\begin{aligned}
&\int_A \log \frac{d\pi^{n+1}}{d\Lambda}(s,a)\pi_w(da|s) \\
&= \int_A \log \frac{d\pi^{n+1}}{d\Lambda}(s,a)(\tanh)_{\#} q_w(da|s) \\
&= \int_{\mathbb{R}^M} \log \frac{d\pi^{n+1}}{d\Lambda}(s, \tanh u) q_w(du|s) \\
&= \int_{\mathbb{R}^M} \left(-F_n(s,u) - \log \overline{Z}_{n+1}(s) \right) q_w(du|s) \\
&= \frac{1}{(2\pi)^{\frac{M}{2}}\sqrt{\det \Sigma}} \int_{\mathbb{R}^M} \left(-F_n(s,u) - \log \overline{Z}_{n+1}(s) \right) \\
&\qquad \exp\left(-\frac{1}{2}(u - m_w(s))^\top \Sigma^{-1}(u - m_w(s)) \right) \Lambda_{\mathbb{R}^M}(du).
\end{aligned}$$

Therefore differentiation under the integral also gives

$$\nabla_w \int_A \log \frac{d\pi^{n+1}}{d\Lambda}(s,a)\pi_w(da|s) = -\int_{\mathbb{R}^M} F_n(s,u)x(s)\Sigma^{-1}(u - m_w(s))q_w(du|s).$$

Here the term containing $\log \overline{Z}_{n+1}(s)$ vanishes because $\int_{\mathbb{R}^M} q_w(du|s) = 1$ for every $w \in \mathfrak{B}(\mathrm{R})$. Hence for all $s \in S$ it holds that

$$\left|\nabla_w \int_A \log \frac{d\pi^{n+1}}{d\Lambda}(s,a)\pi_w(da|s)\right|_2 \leq |F_n|_{B_b(S\times\mathbb{R}^M)} \int_{\mathbb{R}^M} |\Sigma^{-1}(u - m_w(s))|_2 q_w(du|s)$$
$$\leq \frac{\widehat{Q}_{\max}}{\tau} \int_{\mathbb{R}^M} |\Sigma^{-1}(u - m_w(s))|_2 q_w(du|s),$$

Since $q_w(\cdot|s)$ is Gaussian with mean $m_w(s)$ and covariance Σ, the Gaussian first moment (Johnson et al., 1994, Ch. 13) yields

$$\int_{\mathbb{R}^M} |\Sigma^{-1}(u - m_w(s))|_2 q_w(du|s) \leq \sum_{i=1}^M \frac{1}{\sigma_i^2} \int_{\mathbb{R}^M} |u_i - m_w(s)_i| q_w(du|s)$$
$$= \sqrt{\frac{2}{\pi}} \sum_{i=1}^M \frac{1}{\sigma_i}.$$

Hence it holds that

$$\left|\nabla_w \int_A \log \frac{d\pi^{n+1}}{d\Lambda}(s,a)\pi_w(da|s)\right|_2 \leq \sqrt{\frac{2}{\pi}} \frac{\widehat{Q}_{\max}}{\tau} \sum_{i=1}^M \frac{1}{\sigma_i}. \tag{50}$$

Combining (49) and (50), and using that ρ_{n+1} is a probability measure, gives

$$\left|\nabla_w J_{n+1}(w, \pi^{n+1})\right|_2 \leq 2\sqrt{M} + \sqrt{\frac{2}{\pi}} \frac{\widehat{Q}_{\max}}{\tau} \sum_{i=1}^M \frac{1}{\sigma_i}.$$

To bound $\left|\nabla_w J_{n+1}(w, \pi_G^{n+1})\right|_2$, the same arguments hold with $F_n(s,u) = \frac{\widehat{Q}^n(s,\tanh u)}{\tau} \leq \frac{\widehat{Q}_{\max}}{\tau}$. □

J Implementation details

K Implementation details

In this section we give an overview on the environment used in Figures 1 and 2. In order to validate the theory, we restrict to an MDP in which the critics can be solved exactly. To that end, we require that for all $s \in S$, $c(s,\cdot) : A \mapsto \mathbb{R}$ is quadratic and the transition kernel is affine in the action variable. To ease notation, let $g : S \mapsto \mathbb{R}^M$ be $g(s) = f(s)v$ with $f(s) = -\frac{2}{5} + \frac{4}{5}\frac{s}{|S|-1}$ and $v = (1, -\frac{1}{2}, 1, -\frac{1}{2}, \ldots) \in \mathbb{R}^M$. Denoting $\bar{a} = \frac{1}{M}\sum_{i=1}^M a_i$ with $a = (a_1, a_2, \ldots, a_M) \in (-1,1)^M$ and $s^+ := \min(s+1, |S|-1)$, $s^- := \max(s-1, 0)$, the cost function and transition dynamics are defined as

$$c(s,a) = \frac{1}{10}|a - g(s)|_2^2 + \frac{3}{100(M-1)} \sum_{i<j} (a_i - a_j)^2 + \frac{3}{10}\frac{s}{|S|-1},$$

$$P(s'|s,a) = \frac{1}{10|S|} + \frac{9}{10}\left(\frac{1+\bar{a}}{2}\delta_{s^+}(s') + \frac{1-\bar{a}}{2}\delta_{s^-}(s')\right)$$

Given that the state space is tabular, we parametrise the actor using the canonical basis function on $\mathbb{R}^{|S|}$, that is $x(s) = e_s \otimes I_M$ so that $m_w(s) = x(s)^\top w$, $|x(s)|_{\mathrm{op}} = 1$ which in turn implies that $\lambda_P = \frac{1}{|S|}$. We take $\Sigma = 0.15^2 I_M$ and define the projection radius in Algorithm 1 as $\mathrm{R} = \sqrt{|S|M}$.

In Figure 1, to compare the effect on Algorithm 1 in the regimes where the condition of Theorem 2 is satisfied and not, we firstly learn the exact critic through a linear solve (Theorem 4), then add perturbations such that both resulting perturbed critics have the same approximation error $\varepsilon_{\mathrm{crit}}$ but different curvatures measures through $R_{\widehat{Q}}$. Concretely, we construct two sequences of critics such that

$$\widehat{Q}^n(s,a) = Q_\tau^{\pi_{w^n}}(s,a) + \varepsilon_{\mathrm{crit}} \sum_{j=1}^4 \alpha_j \cos\left(\omega_j^\top a\right)$$

with amplitudes $\alpha = (0.55, 0.20, 0.15, 0.10)$, $\sum_{j=1}^{4} \alpha_j = 1$, and such that $\omega_j = \frac{r_j}{\sqrt{M}} \mathbf{1}_M$. Then we construct two instances of $r \in \mathbb{R}^4$, with $r = (0.6, 0.8, 1, 1.2)$ and $r = (6, 7, 8, 9)$ giving $\frac{\mathrm{R}_{\widehat{Q}}}{2} < \tau$ and $\frac{\mathrm{R}_{\widehat{Q}}}{2} > \tau$ respectively, while $\left| \widehat{Q}^n - Q_\tau^{\pi_{w^n}} \right|_{B_b(S \times A)} = \varepsilon_{\mathrm{crit}}$ in both.

In Figure 2, we use the same MDP with varying dimensions but with exact critics ($\varepsilon_{\mathrm{crit}} = 0$) in order to isolate the effect of the proximal step-size/ penalty λ on the actor tracking error. Finally, in all instances, we set the initial state distribution to $\rho(s) = \frac{1}{|S|}$, the entropy temperature to $\tau = 1$, discount factor $\gamma = 0.9$, the actor step size to $h = 0.01$ and the exponential moving average weight in the replay buffer dynamics to $\chi = 1$.