

000 CONVERGENCE OF AN ACTOR-CRITIC GRADIENT 001 FLOW FOR ENTROPY REGULARISED MDPs IN GEN- 002 ERAL SPACES 003

004 **Anonymous authors**
 005
 006

007 Paper under double-blind review
 008
 009

010 ABSTRACT 011

012 We prove the stability and global convergence of a coupled actor-critic gradi-
 013 ent flow for infinite-horizon and entropy-regularised Markov decision processes
 014 (MDPs) in continuous state and action space with linear function approximation
 015 under Q-function realisability. We consider a version of the actor critic gradient
 016 flow where the critic is updated using temporal difference (TD) learning while the
 017 policy is updated using a policy mirror descent method on a separate timescale.
 018 For general action spaces, the relative entropy regularizer is unbounded and thus it
 019 is not clear a priori that the actor-critic flow does not suffer from finite-time blow-
 020 up. Therefore we first demonstrate stability which in turn enables us obtain a
 021 convergence rate of the actor critic flow to the optimal regularised value function.
 022 The arguments presented show that timescale separation is crucial for stability and
 023 convergence in this setting.
 024

025 1 INTRODUCTION 026

027 In reinforcement learning (RL) an agent aims to learn an optimal policy that maximizes the expected
 028 cumulative reward through repeated interactions with its environment. Such methods typically in-
 029 volve two key components: policy evaluation and policy improvement. During policy evaluation,
 030 the advantage function corresponding to a policy, or its function approximation, is updated using
 031 state, action and reward data generated under this policy. Policy improvement then uses this ap-
 032 proximate advantage function to update the policy, most commonly through some policy gradient
 033 method. Algorithms that explicitly combine these two components are known as actor-critic (AC)
 034 methods Konda & Tsitsiklis (1999), where the actor corresponds to policy improvement and the
 035 critic to policy evaluation.

036 There are many policy gradient methods to choose from. In the last decade trust region policy opti-
 037 mization (TRPO) methods Schulman et al. (2015) and methods inspired by these like PPO Schulman
 038 et al. (2017) have become increasingly well-established due to their impressive empirical perfor-
 039 mance. Largely, this is because they alleviate the difficulty in choosing appropriate step sizes for the
 040 policy gradient updates: for vanilla policy gradient even a small change in the parameter may result
 041 in large change in the policy, leading to instability, but TRPO prevents this by explicitly ensuring the
 042 KL divergence between successive updates is smaller than some tolerance. Mirror descent replaces
 043 the TRPO’s hard constraint with a penalty leading to a first order method which is also amenable to
 044 analysis. Indeed, at least for direct parametrization, it is known to converge with sub-linear and even
 045 linear rate for entropy regularised problems (depending on exact assumptions) Ju & Lan (2024); Lan
 046 (2023); Kerimkulov et al. (2025).

047 Due to the favourable analytical properties of mirror descent, in this paper we consider a version of
 048 the actor critic gradient flow where the policy is updated using a policy mirror descent method while
 049 the critic is updated using temporal difference (TD) on a separate timescale.

050 Entropy-regularised MDPs are widely used in practice since the entropic regularizer leads to a num-
 051 ber of desirable properties: it has a natural interpretation as something that drives exploration, it
 052 ensures that there is a unique optimal policy and it can accelerate convergence of mirror descent Ker-
 053 imkulov et al. (2025), as well as classical policy gradient Mei et al. (2020). However, analysing the
 054 stability and convergence of actor-critic methods in this entropy-regularised setting with general

054 state and action spaces remains highly non-trivial due to lack of a priori bounds on the value functions.
 055

056 To address the actor critic methods for entropy regularised MDPs in general action spaces, a careful
 057 treatment of tools from two timescale analysis, convex analysis over both Euclidean spaces and
 058 measure spaces must be deployed.
 059

060 In this paper, we address precisely this challenge. We study the stability and convergence of a
 061 widely used actor-critic algorithm in which the critic is updated using Temporal Difference (TD)
 062 learning Sutton (1988), and the policy is updated through Policy Mirror Descent Ju & Lan (2024).
 063 Our analysis employs a two-timescale update scheme Borkar & Konda (1997), where both the actor
 064 and critic are updated at each iteration with the critic updated on a faster timescale.
 065

066 1.1 RELATED WORKS

067 We focus on the subset of RL literature that address the convergence of coupled actor-critic al-
 068 gorithms. In the unregularised setting, actor-critic methods have been studied extensively. The
 069 first convergence results in the two-timescale regime established asymptotic convergence in the
 070 continuous-time limit of coupled updates (Borkar & Konda (1997); Konda & Tsitsiklis (1999)).
 071 Most modern research employs linear function approximation for the critic, where linear conver-
 072 gence rates have been obtained under various assumptions on the step-sizes of the actor and critic
 073 (Barakat et al. (2022); Zhang et al. (2020); Hong et al. (2023)).

074 Closely related to our work is Zhang et al. (2021), which considers the same two-timescale actor-
 075 critic scheme in the continuous-time limit for unregularised MDPs, with an overparameterized neu-
 076 ral network used for the critic. However, convergence to the optimal policy was not established, and
 077 a restarting mechanism was required to ensure the stability of the dynamics.
 078

079 In the entropy-regularised setting, Cayci et al. (2024a;b) address the convergence of a natural actor
 080 critic algorithm. However, the convergence and stability of these results rely on the finite cardinality
 081 of the action space in presence of entropy regularisation.
 082

083 1.2 OUR CONTRIBUTION

084 Under linear Q_τ^π -realisability assumption, we address the following question:

085 “*Is the actor-critic gradient flow for entropy-regularised MDPs in general action spaces stable and
 086 convergent, and if so, at what rate?*”
 087

088 There are two main technical challenges one has to overcome when working with entropy-
 089 regularised MDPs in general action spaces.
 090

- 091 • Even in mirror descent with exact advantage, the rate of convergence depends on a constant
 092 term $\int_S \text{KL}(\pi^*(\cdot|s) \|\pi_0(\cdot|s) d\pi_\rho^*(ds)$. See Lan (2023); Kerimkulov et al. (2025). In general
 093 action spaces, without entropy regularisation it is almost impossible to choose π_0 which
 094 would make this term finite, see Remark 2.1. Thus we need to include the regularisation in
 095 the analysis.
- 096 • Moreover, ensuring that the relative entropy does not blow up is difficult in general action
 097 spaces. In the finite action space setting, for any measure $\mu \in \mathcal{P}(A)$ such that $\mu(a_i) > 0$
 098 and for all $s \in S$ it holds that $\text{KL}(\pi(\cdot|s) \|\mu) \leq \log |A|$. In general action spaces the KL
 099 divergence has no upper bound (can be $+\infty$) even if μ has full support. Under mild assump-
 100 tions we show that the KL divergence does not blow up in finite time, see Corollary 5.1.

101 Our main contributions are as follows:
 102

- 103 • We study a common variant of actor-critic where the critic is updated using temporal dif-
 104 ference (TD) learning and the policy is updated using mirror descent. Similarly to Konda
 105 & Tsitsiklis (1999); Zhang et al. (2021), we analyse the coupled updates in the continuous-
 106 time limit, resulting in a dynamical system where the critic flow is captured by a *semi-*
 107 *gradient* flow and the actor flow corresponds to an approximate Fisher–Rao gradient flow
 over the space of probability kernels.

- By combining convex analysis over the space of probability measures with classical Euclidean convex analysis, we develop a Lyapunov-based stability framework that captures the interplay between entropy regularisation and timescale separation, and establish stability of the resulting dynamics.
- We prove convergence of the actor-critic dynamics for entropy-regularised MDPs with infinite action spaces.

1.3 NOTATION

Let (E, d) denote a Polish space (i.e., a complete separable metric space). We always equip a Polish space with its Borel sigma-field $\mathcal{B}(E)$. Denote by $B_b(E)$ the space of bounded measurable functions $f : E \rightarrow \mathbb{R}$ endowed with the supremum norm $|f|_{B_b(E)} = \sup_{x \in E} |f(x)|$. Denote by $\mathcal{M}(E)$ the Banach space of finite signed measures μ on E endowed with the total variation norm $|\mu|_{\mathcal{M}(E)} = |\mu|(E)$, where $|\mu|$ is the total variation measure. Recall that if $\mu = f d\rho$, where $\rho \in \mathcal{M}_+(E)$ is a nonnegative measure and $f \in L^1(E, \rho)$, then $|\mu|_{\mathcal{M}(E)} = |f|_{L^1(E, \rho)}$. Denote by $\mathcal{P}(E) \subset \mathcal{M}(E)$ the set of probability measures on E . Moreover, we denote the Euclidean norm on \mathbb{R}^N by $|\cdot|$ with inner product $\langle \cdot, \cdot \rangle$. Given some $A, B \in \mathbb{R}^{N \times N}$, we denote by $\lambda_{\min}(A)$ the minimum eigenvalue of A and denote $A \succeq B$ if and only if $A - B$ is positive semidefinite. Moreover, we denote by $|A|_{\text{op}}$ the operator norm of A induced by the Euclidean norm, $|A|_{\text{op}} := \sup_{|x| \neq 0} \frac{|Ax|}{|x|}$.

1.4 ENTROPY REGULARISED MARKOV DECISION PROCESSES

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

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

where $\text{KL}(\pi(\cdot|s) | \mu)$ is the Kullback-Leibler (KL) divergence of $\pi(\cdot|s)$ with respect to μ , define as $\text{KL}(\pi(\cdot|s) | \mu) := \int_A \ln \frac{d\pi}{d\mu}(a|s) \pi(da|s)$ if $\pi(\cdot|s)$ is absolutely continuous with respect to μ , and infinity otherwise.

For a given initial distribution $\rho \in \mathcal{P}(S)$, the optimal value function is defined as

$$V_\tau^*(\rho) = \min_{\pi \in \mathcal{P}(A|S)} V_\tau^\pi(\rho), \quad \text{with } V_\tau^\pi(\rho) := \int_S V_\tau^\pi(s) \rho(ds)$$

and we refer to $\pi^* \in \mathcal{P}(A|S)$ as the optimal policy if $V_\tau^*(\rho) = V_\tau^{\pi^*}(\rho)$. The Bellman Principle for entropy regularised MDPs, see Theorem A.1, tells us two important things: first, the optimal value function $V_\tau^* \in B_b(S)$ is the unique bounded solution of the following Bellman equation:

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

where for all $s \in S$ and $a \in A$, $Q_\tau^* \in B_b(S \times A)$ is defined by

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

Second, for all $s \in S$ there is a unique optimal policy $\pi^* \in \mathcal{P}(A|S)$ given by

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

Hence, without loss of generality, it is sufficient to consider policies from the class given by Definition 1.1 below.

162 **Definition 1.1** (Admissible Policies). *Let Π_μ denote the class of policies for which there exists
163 $f \in B_b(S \times A)$ with*

$$164 \quad 165 \quad \pi(da|s) = \frac{\exp(f(s, a))}{\int_A \exp(f(s, a)) \mu(da)} \mu(da).$$

166 For each $\pi \in \Pi_\mu$ the value function V_τ^π is the unique bounded solution of the on-policy Bellman
167 equation

$$168 \quad V_\tau^\pi(s) = \int_A \left(Q_\tau^\pi(s, a) + \tau \ln \frac{d\pi}{d\mu}(a, s) \right) \pi(da|s),$$

171 see e.g. (Kerimkulov et al., 2025, Lemma B.2).

172 For each $\pi \in \Pi_\mu$, we define the state-action value function $Q_\tau^\pi \in B_b(S \times A)$ by

$$173 \quad Q_\tau^\pi(s, a) = c(s, a) + \gamma \int_S V_\tau^\pi(s') P(ds'|s, a). \quad (1)$$

174 We see that $Q_\tau^\pi : S \times A \rightarrow \mathbb{R}$ is the unique fixed point of the operator $T^\pi : B_b(S \times A) \rightarrow B_b(S \times A)$,
175 defined as

$$176 \quad T^\pi f(s, a) = c(s, a) + \gamma \int_{S \times A} f(s', a') P^\pi(ds', da'|s, a) + \tau \gamma \int_S \text{KL}(\pi(\cdot|s')|\mu) P(ds'|s, a). \quad (2)$$

182 2 MIRROR-DESCENT AND THE FISHER-RAO GRADIENT FLOW

184 Defining the soft advantage function as

$$185 \quad A_\tau^\pi(s, a) := Q_\tau^\pi(s, a) + \tau \ln \frac{d\pi}{d\mu}(s, a) - V_\tau^\pi(s),$$

186 then for some $\lambda > 0$ and $\pi_0 \in \Pi_\mu$, the Policy Mirror Descent update rule reads as

$$187 \quad \pi^{n+1}(\cdot|s) = \arg \min_{m \in \mathcal{P}(A)} \left[\int_A A_\tau^{\pi^n}(s, a) (m(da) - \pi^n(da|s)) + \frac{1}{\lambda} \text{KL}(m|\pi^n(\cdot|s)) \right]$$

188 Dupuis & Ellis (1997) shows that the pointwise optimisation is achieved by

$$189 \quad \frac{d\pi^{n+1}}{d\pi^n}(a, s) = \frac{\exp(-\lambda A_\tau^{\pi^n}(s, a))}{\int_A \exp(-\lambda A_\tau^{\pi^n}(s, a)) \pi^n(da|s)}. \quad (3)$$

190 Observe that for any $\pi \in \mathcal{P}(A|S)$, it holds that $\int_A A_\tau^\pi(s, a) \pi(da|s) = 0$. Hence taking the logarithm
191 of (3) we have

$$192 \quad \ln \frac{d\pi^{n+1}}{d\mu}(s, a) - \ln \frac{d\pi^n}{d\mu}(s, a) = -\lambda A_\tau^{\pi^n}(s, a) - \ln \int_A e^{-\lambda A_\tau^{\pi^n}(s, a)} \pi^n(da|s).$$

193 Interpolating in the time variable and letting $\lambda \rightarrow 0$ we retrieve the Fisher-Rao gradient flow for the
194 policies

$$195 \quad \partial_t \ln \frac{d\pi_t}{d\mu}(s, a) = - \left(A_\tau^{\pi_t}(s, a) - \int_A A_\tau^{\pi_t}(s, a) \pi_t(da|s) \right) = -A_\tau^{\pi_t}(s, a). \quad (4)$$

196 Note that the soft advantage formally corresponds to the functional derivative of the value function
197 with respect to the policy π^n and thus (4) can be seen as a gradient flow of the value function over the
198 space of kernels $\mathcal{P}(A|S)$ (see Kerimkulov et al. (2025) for a detailed description of the functional
199 derivative).

200 **Remark 2.1.** *In the case where the advantage function is fully accessible for all $t \geq 0$, Kerimkulov
201 et al. (2025)[Theorem 2.8] shows that the entropy regularisation in the value function induces an
202 exponential convergence to the optimal policy. More specifically, their result shows that for all $t \geq 0$
203 we have*

$$204 \quad 0 \leq V_\tau^{\pi_t}(\rho) - V_\tau^{\pi^*}(\rho) \leq \frac{\tau}{(1 - \gamma)(e^{\tau t} - 1)} \left(\int_S \text{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi^*}(ds) \right).$$

216 If the action space has finite cardinality and π_0 is chosen to be uniform we see that
 217 $\text{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s)) \leq \log |A|$ for all $s \in S$, where $|A|$ represents the cardinality of the action
 218 space. One can then let $\tau \rightarrow 0$ in the above estimate to formally obtain convergence rate of order
 219 $1/t$ for the unregularised problem.

220 In the setting of general action spaces $\text{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s))$ is finite only if the density $\frac{d\pi^*}{d\pi_0}$ exists.
 221 However, by the dynamic programming principle for the unregularised problem (Hernández-Lerma
 222 & Lasserre, 1996, Theorem 4.2.3) shows that the optimal policies will have support on a mixture of
 223 Dirac distributions. Therefore, $\text{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s))$ will be finite only if π_0 is also a mixture of Dirac
 224 distributions with support which contains the support of $\pi^*(\cdot|s)$ for all $s \in S$. It is not realistic to
 225 assume that one can guess the initial policy π_0 which will have the above property. However, in the
 226 entropy regularised case, Theorem A.1 tells us $\pi^*(\cdot|s)$ has full support on A and so one simply has
 227 to choose $\pi_0(\cdot|s)$ to have full support on the action space A for all $s \in S$.

3 ACTOR CRITIC METHODS

231 Given some feature mapping $\phi : S \times A \rightarrow \mathbb{R}^N$, we parametrise the state-action value function as
 232 $Q(s, a; \theta) := \langle \theta, \phi(s, a) \rangle$. Moreover, we denote the approximate soft advantage function as

$$234 \quad A(s, a; \theta) = Q(s, a; \theta) + \tau \ln \frac{d\pi}{d\mu}(s, a) - \int_A \left(Q(s, a; \theta) + \tau \ln \frac{d\pi}{d\mu}(s, a) \right) \pi(da|s).$$

237 The Mean Squared Bellman Error (MSBE) is defined as

$$239 \quad \text{MSBE}(\theta, \pi) = \frac{1}{2} \int_{S \times A} (Q(s, a; \theta) - \mathbb{T}^\pi Q(s, a; \theta))^2 d_\beta^\pi(da, ds)$$

241 where for some fixed $\beta \in \mathcal{P}(S \times A)$, $d_\beta^\pi \in \mathcal{P}(S \times A)$ is the state-action occupancy measure defined
 242 in (9). Given that $\beta \in \mathcal{P}(S \times A)$ has full support, by (2) it holds that $\text{MSBE}(\theta, \pi) = 0$ if and
 243 only if $Q(s, a; \theta) = Q_\pi^\pi(s, a)$ for all $s \in S$ and $a \in A$. Hence one approach to implementing the
 244 policy mirror descent updates is to calculate the optimal parameters for $Q(s, a; \theta)$ by minimising
 245 the MSBE at each policy mirror descent iteration and then update the policy using variable steps
 246 $\{\lambda_n\}_{n \geq 0}$. This reads as

$$248 \quad \begin{cases} \theta^{n+1} = \arg \min_{\theta \in \mathbb{R}^N} \text{MSBE}(\theta, \pi^n), \\ 249 \quad \frac{d\pi^{n+1}}{d\pi^n}(a, s) = \frac{\exp(-\lambda_n A(s, a; \theta^{n+1}))}{\int_A \exp(-\lambda_n A(s, a; \theta^{n+1})) \pi^n(da|s)}. \end{cases} \quad (5)$$

253 To avoid fully solving the arg min in (5) for each policy update, one can update the critic using a
 254 semi-gradient descent on a different timescale to the policy update. Let $\{h_n\}_{n \geq 0}$ be the step-sizes
 255 of the critic at iteration $n \geq 0$. Let the semi-gradient $g : \mathbb{R}^N \times \mathcal{P}(A|S) \rightarrow \mathbb{R}^N$ of the MSBE with
 256 respect to θ be

$$258 \quad g(\theta, \pi) := \int_{S \times A} (Q(s, a; \theta) - \mathbb{T}^\pi Q(s, a; \theta)) \phi(s, a) d_\beta^\pi(da, ds). \quad (6)$$

260 The full arg min update in (5) is then replaced by

$$262 \quad \theta^{n+1} = \theta^n - h_n g(\theta^n, \pi^n),$$

263 where timescale separation $\eta_n := \frac{h_n}{\lambda_n} > 1$ ensures that the critic is updated on a much faster
 264 timescale than the policy to improve the local estimation of the policy updates. With general action
 265 spaces, which allow the KL term may be unbounded, one may need to go even further and choose
 266 a scheme which does several (and possibly increasing) number of updates of the critic before doing
 267 an actor update.

268 In this paper we focus on a continuous-time idealisation of the above which is presented in the next
 269 section.

270 **4 DYNAMICS**
 271

272 In this paper, we study the stability and convergence of the two-timescale actor-critic Mirror Descent
 273 scheme in the continuous-time limit. Let $Q_t(s, a) := Q(s, a; \theta_t)$ and $A_t(s, a) := A(s, a; \theta_t)$. Let
 274 $\eta : [0, \infty) \rightarrow [1, \infty)$ be a non-decreasing function representing the timescale separation, then for
 275 some $\theta_0 \in \mathbb{R}^N$, $\pi_0 \in \Pi_\mu$ and $\beta \in \mathcal{P}(S \times A)$, we have the following coupled dynamics

276
$$\frac{d\theta_t}{dt} = -\eta_t g(\theta_t, \pi_t), \quad \theta_0 = \theta^0 \in \mathbb{R}^N, \quad (7)$$

 277

278
$$\partial_t \pi_t(da|s) = -A_t(s, a) \pi_t(da|s), \quad t \geq 0, \quad \pi_0 = \pi^0 \in \Pi_\mu, \quad (8)$$

 279

280 where $g : \mathbb{R}^N \times \mathcal{P}(A|S)$ is the semi-gradient of the MSBE defined in (6). We refer to (8) as the
 281 approximate Fisher–Rao Gradient flow.

282 We perform our analysis under the following assumptions.

283 **Assumption 4.1** (Q_τ^π -realisability). *For all $\pi \in \Pi_\mu$ there exists $\theta_\pi \in \mathbb{R}^N$ such that $Q^\pi(s, a) = \langle \theta_\pi, \phi(s, a) \rangle$ for all $(s, a) \in S \times A$.*

286 A simple example of when this holds is in the tabular case, where one can choose ϕ to be a one-hot
 287 encoding of the state-action space. Moreover, all linear MDPs are Q^π -realisable. In a linear MDP
 288 there exists $\phi : S \times A \rightarrow \mathbb{R}^N$, $w \in \mathbb{R}^N$ and a sequence $\{\psi_i\}_{i=1}^N$ with $\psi_i \in \mathcal{M}(S)$ such that
 289 for all $(s, a) \in S \times A$,

290
$$c(s, a) = \langle w, \phi(s, a) \rangle, \quad P(ds' | s, a) = \sum_{i=1}^N \phi_i(s, a) \psi_i(ds').$$

 291

293 In this case it holds that $(\theta_\pi)_i = w_i + \int_S V^\pi(s') \psi_i(ds')$. Assumption 4.1 can be seen as a con-
 294 vention to omit function approximation errors in the final convergence results. This assumption, or
 295 the presence of approximation errors in convergence results, are widely present in the actor-critic
 296 literature (Cayci et al. (2024a), Zhang et al. (2020), Zanette et al. (2021), Hong et al. (2023), Qiu
 297 et al. (2021)).

298 More recently, Lin et al. (2025) derives some weaker ordering conditions in the bandit case (empty
 299 state space) which guarantee the convergence of soft-max policy gradient in the tabular setting be-
 300 yond realisability. However as of now it is unclear how this applies to MDPs and also fundamentally
 301 depends on the finite cardinality of the action space.

303 Since for all $\pi \in \Pi_\mu$ we know that $Q_\tau^\pi \in B_b(S \times A)$ we also have $Q_\tau^\pi \in L^2(S \times A; \beta)$, which
 304 is a Hilbert space. By (Brezis, 2011, Theorem 5.11), Assumption 4.1 holds in the limit $N \rightarrow \infty$
 305 when ϕ_i are the basis functions of $L^2(S \times A; \beta)$. However, analysis in such a Hilbert space becomes
 306 more involved and intricate and is the result of ongoing work. Combining this approach with careful
 307 truncation of the basis functions has demonstrated empirical success in Ma et al. (2024); Ren et al.
 308 (2023).

309 **Assumption 4.2.** *For all $(s, a) \in S \times A$ it holds that $|\phi(s, a)| \leq 1$.*

310 Assumption 4.2 is purely for convention and is without loss of generality in the finite-dimensional
 311 case.

312 **Assumption 4.3.** *Let $\beta \in \mathcal{P}(S \times A)$ be fixed. Then*

314
$$\lambda_\beta := \lambda_{\min} \left(\int_{S \times A} \phi(s, a) \phi(s, a)^\top \beta(ds da) \right) > 0.$$

 315

316 Note that unlike the analogous assumptions imposed in Hong et al. (2023), Assumption 4.3 is in-
 317 dependent of the policy. This property allows us to remove any dependence on the continuity of
 318 eigenvalues.

319 **Definition 4.1.** *For all $\pi \in \Pi_\mu$ and $\zeta \in \mathcal{P}(S \times A)$, the squared loss with respect to ζ is defined as*

321
$$L(\theta, \pi; \zeta) = \frac{1}{2} \int_{S \times A} (\langle \theta, \phi(s, a) \rangle - Q_\tau^\pi(s, a))^2 \zeta(da, ds)$$

 322

323 where Q_τ^π is defined in (1).

324 A straightforward calculation given in Lemma B.3 shows that due to Lemma A.1 and Assumption
 325 4.3, for any $\pi \in \Pi_\mu$ it holds that $L(\cdot, \pi; d_\beta^\pi)$ is $(1 - \gamma)\lambda_\beta$ -strongly convex.
 326

327 The following result then connects the geometry of the semi-gradient of the MSBE and the gradient
 328 of $L(\cdot, \pi; \beta)$, which can be seen as an extension of Lemma 3 of Bhandari et al. (2021) to the current
 329 entropy regularised setting and where the measure of integration in the MSBE is not necessarily
 330 stationary.

331 **Lemma 4.1.** *Let Assumption 4.1 hold. Then for all $\theta \in \mathbb{R}^N$ and $\pi \in \Pi_\mu$ it holds that*

$$332 -\langle g(\theta, \pi), \theta - \theta_\pi \rangle \leq -(1 - \sqrt{\gamma})(1 - \gamma) \langle \nabla_\theta L(\theta, \pi; \beta), \theta - \theta_\pi \rangle$$

333 with

$$334 \nabla_\theta L(\theta, \pi; \beta) = \int_{S \times A} (\langle \theta, \phi(s, a) \rangle - Q_\pi^\pi(s, a)) \phi(s, a) \beta(da, ds).$$

335 See Appendix B.1 for a proof.

336 5 STABILITY

341 In this section we analyse the stability of the coupled actor-critic flow. Under mild assumptions, see
 342 Corollary 5.1 shows that for all $s \in S$, $\text{KL}(\pi(\cdot|s)|\mu)$ does not blow up in finite time in the sense that
 343 there is no $T > 0$ and no $s \in S$ such that $\lim_{t \nearrow T} \text{KL}(\pi_t(\cdot|s)|\mu) = +\infty$. Existence of such a time
 344 $T > 0$ would result in a singularity in the actor-critic dynamics.

345 Throughout this section, to ease notation we let

$$347 \Gamma := \lambda_\beta(1 - \gamma)(1 - \sqrt{\gamma}),$$

$$348 K_t := \sup_{s \in S} \text{KL}(\pi_t(\cdot|s)|\mu),$$

350 with $\lambda_\beta > 0$ the constant from Assumption 4.3.

351 Using Lemma A.1, Lemma 5.1 then establishes the effect of the coupling and timescale separation
 352 in the actor-critic flow and its effect on the stability of the critic parameters.

353 **Lemma 5.1.** *Let Assumptions 4.2 and 4.3 hold. Then for all $t \geq 0$ it holds that*

$$355 \frac{1}{2\eta_t} \frac{d}{dt} |\theta_t|^2 \leq -\frac{\Gamma}{2} |\theta_t|^2 + \frac{\tau^2 \gamma^2 K_t^2}{\Gamma} + \frac{|c|_{B_b(S \times A)}^2}{\Gamma}$$

358 See Appendix C.1 for a proof. By connecting the result from Lemma 5.1 with the approximate
 359 Fisher–Rao gradient flow, we are able to establish a Grönwall-type inequality for the KL divergence
 360 of the policies with respect to the reference measure. Lemma 5.1 also illustrates that the coupled
 361 actor-critic flow is a forcing-damping system, where the damping comes from the strong convexity
 362 of the loss $\theta \mapsto L(\theta, \pi_t; d_\beta^{\pi_t})$ and the forcing coming from the policy updates manifesting as the K_t
 363 term on the right-hand-side of the estimate. Here we can see that in the finite action space setting,
 364 the forcing K_t term is upper bounded by a constant and thus we can arrive at stability straight
 365 away. In the current setting this is not possible and we must perform analysis over $\mathcal{P}(A|S)$ on the
 366 approximate Fisher–Rao gradient flow to arrive at stability.

367 **Theorem 5.1.** *Let Assumptions 4.2 and 4.3 hold. Let $\eta_0 > \frac{\tau}{\Gamma}$. Then there exists constants*

$$369 a_1 = a_1 \left(\tau, \eta_0, \gamma, \lambda_\beta, |c|_{B_b(S \times A)}, \left| \frac{d\pi_0}{d\mu} \right|_{B_b(S \times A)} \right) > 0$$

371 and $a_2 = a_2(\tau, \eta_0, \gamma, \lambda_\beta) > 0$ such that for all $\gamma \in (0, 1)$ and $t \geq 0$ it holds that

$$373 K_t^2 \leq a_1 + a_2 \int_0^t e^{-\tau(t-r)} K_r^2 dr.$$

376 See Appendix C.2 for a proof. Through applications of Grönwall’s Lemma (Lemma A.3), two direct
 377 corollaries of Theorem 5.1 show that the KL divergence of the policies with respect to the reference
 378 measure and the critic parameters do not blow up in finite time.

378 **Corollary 5.1** (Stability of π_t). *Under the same assumptions as Theorem 5.1, for all $\gamma \in (0, 1)$,
379 $s \in S$ and $t \geq 0$ it holds that*

$$380 \quad \text{KL}(\pi_t(\cdot|s)|\mu)^2 \leq a_1 e^{a_2 t}.$$

381 **Corollary 5.2** (Stability of θ_t). *Under the same assumptions as Theorem 5.1, suppose that there
382 exists $\alpha > 0$ such that $\frac{d}{dt}\eta_t \leq \alpha\eta_t$. Then for all $\gamma \in (0, 1)$ there exists $r_1, r_2 > 0$ such that for all
383 $t \geq 0$ it holds that*

$$384 \quad |\theta_t| \leq r_1 e^{r_2 t}.$$

385 See Appendix C.3 and C.4 for the proofs. Corollaries E.1 and E.2 in the appendix show that if the
386 MDP is sufficiently regularised through a sufficiently small discounting factor, the KL divergence
387 of the policies with respect to the reference measure remains uniformly bounded along the flow.
388

389 6 CONVERGENCE

390 In this section we will present three convergence results of the coupled actor-critic flow. Firstly,
391 we characterise the time derivative of the state-action value function along the approximate gradient
392 flow for the policies.

393 **Lemma 6.1.** *For all $t \geq 0$ and $(s, a) \in S \times A$, it holds that*

$$396 \quad \frac{d}{dt}Q_\tau^{\pi_t}(s, a) = \frac{\gamma}{1 - \gamma} \int_S \left(\int_{S \times A} A_\tau^{\pi_t}(s'', a'') \partial_t \pi_t(da''|s'') d^{\pi_t}(ds''|s') \right) P(ds'|s, a)$$

397 See Appendix D.1 for a proof. Observe that in the exact setting, (4), we obtain the dissipative
398 property of $\{Q_\tau^{\pi_t}\}_{t \geq 0}$ along the flow

$$401 \quad \frac{d}{dt}Q_\tau^{\pi_t}(s, a) = \frac{-\gamma}{1 - \gamma} \int_S \left(\int_{S \times A} A_\tau^{\pi_t}(s'', a'')^2 d^{\pi_t}(ds''|s') \right) P(ds'|s, a) \leq 0.$$

402 Furthermore, Theorem 6.1 shows that the actor-critic flow maintains the exponential convergence to
403 the optimal policy induced by the τ -regularisation up to a error term arising from not solving the
404 critic to full accuracy.

405 **Theorem 6.1.** *Let $\{\pi_t, \theta_t\}_{t \geq 0}$ be the trajectories of the actor critic flow. Let Assumptions 4.1 and
406 4.2 hold. Then for all $t > 0$ it holds that*

$$410 \quad \min_{r \in [0, t]} V_\tau^{\pi_r}(\rho) - V_\tau^{\pi^*}(\rho) \leq \frac{\tau}{2(1 - \gamma)(1 - e^{-\frac{\tau}{2}t})} \left(e^{-\frac{\tau}{2}t} \int_S \text{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi^*}(ds) \right. \\ 411 \quad \left. + \frac{1}{2\tau} \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \right)$$

412 See Appendix D.2 for a proof. Theorem 6.1 shows that the exponentially weighted error term
413 determines the rate of convergence of the actor-critic dynamics. On this note, Theorem 6.2 shows
414 that this error term decays exponentially up to an integral which now depends on the rate of change
415 of the true state-action value function and the timescale separation.

416 **Theorem 6.2.** *Let Assumptions 4.1, 4.2 and 4.3 hold. Let $\eta_0 > \frac{1}{\Gamma}$ and $0 < \tau < 1$. Then for all
417 $t \geq 0$ there exists constants $b_1, b_2 > 0$ such that*

$$418 \quad \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq b_1 e^{-\frac{\tau}{2}t} + b_2 \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \left| \frac{d\theta_{\pi_r}}{dt} \right|^2 dr.$$

419 See Appendix D.3 for a proof. By using Corollary 5.1 and by choosing η_t such that the critic
420 flows runs much faster than the actor, Theorem 6.3 demonstrates an exponential convergence to the
421 optimal policy for all $\gamma \in (0, 1)$.

422 **Theorem 6.3.** *Under the same assumptions as Theorem 6.2, there exists $k_1 > 0$ with $\eta_t = \eta_0 e^{k_1 t}$
423 and $k_2 > 0$ such that for all $\gamma \in (0, 1)$ and $t > 0$ it holds that*

$$424 \quad \min_{r \in [0, t]} V_\tau^{\pi_r}(\rho) - V_\tau^{\pi^*}(\rho) \leq \frac{\tau e^{-\frac{\tau}{2}t}}{2(1 - \gamma)(1 - e^{-\frac{\tau}{2}t})} \left(\int_S \text{KL}(\pi^*(\cdot|s)|\pi_0(\cdot|s)) d_\rho^{\pi^*}(ds) + \frac{k_2}{2\tau} \right)$$

432 See Appendix D.4 for a proof.
433
434 Corollary E.3 then shows that if the MDP is sufficiently regularised through a small discounting
435 factor, one can arrive at convergence for a much more general class of functions η_t .
436

437 7 LIMITATIONS

438

439 In this work, we only study the continuous-time dynamics of the actor-critic algorithm. Although
440 this formulation gives insights into the discrete counterpart, a rigorous treatment of the discrete-time
441 setting is more realistic for practical purposes and is left for future research.

442 Moreover, for the purposes of analysis our critic approximation is linear while in practice non-linear
443 neural networks are used to approximate the critic.
444

445 Finally, our work assumes all integrals are evaluated exactly, in particular the semi-gradient (6). In
446 practice these would need to be estimated from samples leading to additional Monte-Carlo errors.
447 To fully analyse this is left for future work.

448 8 ACKNOWLEDGEMENTS

449

450 The first author is supported by the xxx, funded by the xxx, xxx and xxx. The second and third
451 authors gratefully acknowledge funding from xxx.
452

453 **The authors would like to thank the anonymous reviewers whose insightful comments have helped
454 improve the paper.**
455

456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

486 REFERENCES
487

488 Anas Barakat, Pascal Bianchi, and Julien Lehmann. Analysis of a target-based actor-critic algorithm
489 with linear function approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel
490 Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and*
491 *Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 991–1040. PMLR, 28–
492 30 Mar 2022. URL <https://proceedings.mlr.press/v151/barakat22a.html>.

493 Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference
494 learning with linear function approximation. *Operations Research*, 69(3):950–973, 2021. doi:
495 10.1287/opre.2020.2024. URL <https://doi.org/10.1287/opre.2020.2024>.

496 V. S. Borkar and V. R. Konda. The actor-critic algorithm as multi-time-scale stochastic approxima-
497 tion. *Sadhana*, 22(5):525–543, 1997. doi: 10.1007/BF02745577.

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

502 Semih Cayci, Niao He, and R. Srikant. Convergence of entropy-regularized natural policy gradient
503 with linear function approximation. *SIAM Journal on Optimization*, 34(3):2729–2755, 2024a.
504 doi: 10.1137/22M1540156. URL <https://doi.org/10.1137/22M1540156>.

505 Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural
506 actor-critic algorithm. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856.
507 URL <https://openreview.net/forum?id=BkEqk7pS1I>.

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

511 Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Stochastic Modelling and Applied Probability. Springer, New York,
512 NY, 1996. ISBN 978-0-387-94579-8. doi: 10.1007/978-1-4612-0729-0. First edition.

513 Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm
514 framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM
515 Journal on Optimization*, 33(1):147–180, 2023. doi: 10.1137/20M1387341. URL <https://doi.org/10.1137/20M1387341>.

516 Caleb Ju and Guanghui Lan. Policy optimization over general state and action spaces, 2024. URL
517 <https://arxiv.org/abs/2211.16715>.

518 Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learn-
519 ing. In *International Conference on Machine Learning*, 2002. URL <https://api.semanticscholar.org/CorpusID:31442909>.

520 Bekzhan Kerimkulov, David ika, ukasz Szpruch, and Yufei Zhang. Mirror descent for stochastic
521 control problems with measure-valued controls, 2024. URL <https://arxiv.org/abs/2401.01198>.

522 Bekzhan Kerimkulov, James-Michael Leahy, David Siska, Lukasz Szpruch, and Yufei Zhang. A
523 fisher-rao gradient flow for entropy-regularised markov decision processes in polish spaces.
524 *Foundations of Computational Mathematics*, 2025. ISSN 1615-3383. doi: 10.1007/s10208-025-09729-3.
525 URL <https://doi.org/10.1007/s10208-025-09729-3>.

526 Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and
527 K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press,
528 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

540 Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling
 541 complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106,
 542 2023.

543 Max Qiushi Lin, Jincheng Mei, Matin Aghaei, Michael Lu, Bo Dai, Alekh Agarwal, Dale Schu-
 544 urmans, Csaba Szepesvari, and Sharan Vaswani. Rethinking the global convergence of softmax
 545 policy gradient with linear function approximation, 2025. URL <https://arxiv.org/abs/2505.03155>.

546 L. Liu, M. B. Majka, and . Szpruch. Polyak–ojasiewicz inequality on the space of measures
 547 and convergence of mean-field birth-death processes. *Applied Mathematics and Optimization*,
 548 87:48, 2023. doi: 10.1007/s00245-022-09962-0. URL <https://doi.org/10.1007/s00245-022-09962-0>.

549 H. Ma, Z. Ren, B. Dai, and N. Li. Skill transfer and discovery for sim-to-real learning: A
 550 representation-based viewpoint. In *Proceedings of the 2024 IEEE/RSJ International Conference
 551 on Intelligent Robots and Systems (IROS)*, pp. 8603–8609, Abu Dhabi, United Arab Emirates,
 552 2024. doi: 10.1109/IROS58592.2024.10801637.

553 Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence
 554 rates of softmax policy gradient methods. In *International conference on machine learning*, pp.
 555 6820–6829. PMLR, 2020.

556 Shuang Qiu, Zhaoran Yang, Jianfeng Ye, and Zhuoran Wang. On finite-time convergence of actor-
 557 critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, June
 558 2021. doi: 10.1109/JSAIT.2021.3078754.

559 T. Ren, Z. Ren, N. Li, and B. Dai. Stochastic nonlinear control via finite-dimensional spectral
 560 dynamic embedding. In *Proceedings of the 2023 62nd IEEE Conference on Decision and Control
 561 (CDC)*, pp. 795–800, Singapore, Singapore, 2023. doi: 10.1109/CDC49753.2023.10383842.

562 John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region
 563 policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR,
 564 2015.

565 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 566 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

567 Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*,
 568 3(1):9–44, 1988. doi: 10.1007/BF00115009.

569 Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable benefits of actor-critic meth-
 570 ods for offline reinforcement learning. In *Proceedings of the 35th International Conference on
 571 Neural Information Processing Systems*, NIPS ’21, Red Hook, NY, USA, 2021. Curran Associates
 572 Inc. ISBN 9781713845393.

573 Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-
 574 timescale off-policy actor-critic with function approximation. In Hal Daum III and Aarti Singh
 575 (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of
 576 *Proceedings of Machine Learning Research*, pp. 11204–11213. PMLR, 13–18 Jul 2020. URL
 577 <https://proceedings.mlr.press/v119/zhang20s.html>.

578 Yufeng Zhang, Siyu Chen, Zhuoran Yang, Michael Jordan, and Zhaoran Wang. Wasserstein flow
 579 meets replicator dynamics: A mean-field analysis of representation learning in actor-critic. In
 580 A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural
 581 Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=9IJLHPuLpvZ>.

582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593

594 **A KNOWN PROPERTIES MDPs AND OTHER USEFUL RESULTS**
 595

596 The state-occupancy kernel $d^\pi \in \mathcal{P}(S|S)$ is defined by
 597

598
$$d^\pi(ds'|s) = (1 - \gamma) \sum_{n=0}^{\infty} \gamma^n P_\pi^n(ds'|s),$$

 599

600 where P_π^n is the n -times product of the kernel P_π with $P_\pi^0(ds'|s) := \delta_s(ds')$. Moreover, for each
 601 $\pi \in \mathcal{P}(A|S)$ and $(s, a) \in S \times A$, we define the state-action occupancy kernel as
 602

603
$$d^\pi(ds, da|s, a) = (1 - \gamma) \sum_{n=0}^{\infty} \gamma^n (P^\pi)^n(ds, da|s, a)$$

 604

605 where $(P^\pi)^n$ is the n -times product of the kernel P^π with $(P^\pi)^0(ds', da'|s, a) := \delta_{(s, a)}(ds', da')$. Given some initial state-action distribution $\beta \in \mathcal{P}(S \times A)$ with initial state distribution given by
 606 $\rho(ds) = \int_A \beta(da, ds)$, we define the state-occupancy and state-action occupancy measures as
 607

608
$$d_\rho^\pi(ds) = \int_S d^\pi(ds|s') \rho(ds'), \quad d_\beta^\pi(ds, da) = \int_{S \times A} d^\pi(ds, da|s', a') \beta(da', ds'). \quad (9)$$

 609

610 Note that for all $E \in \mathcal{B}(S \times A)$, by defining the linear operator $J_\pi : \mathcal{P}(S \times A) \rightarrow \mathcal{P}(S \times A)$ as
 611

612
$$J_\pi \beta(E) = \int_{S \times A} P^\pi(E|s', a') \beta(ds', da'), \quad (10)$$

 613

614 it directly holds that
 615

616
$$d_\beta^\pi(da, ds) = (1 - \gamma) \sum_{n=0}^{\infty} \gamma^n J_\pi^n \beta(da, ds),$$

 617

618 with J_π^n the n -fold product of the operator J_π with $J_\pi^0 = I$, the identity operator on $\mathcal{P}(S \times A)$. The
 619 following lemma establishes properties of the state-action occupancy measure defined in (9) and
 620 which are useful in the proofs.
 621

622 **Lemma A.1.** *For all $\pi \in \mathcal{P}(A|S)$, $\beta \in \mathcal{P}(S \times A)$ and $E \in \mathcal{B}(S \times A)$ it holds that*

623
$$d_{J_\pi \beta}^\pi(E) = J_\pi^\pi d_\beta^\pi(E).$$

 624

625 Moreover, for all $\gamma \in (0, 1)$ we have

626
$$d_\beta^\pi(E) - \gamma d_{J_\pi \beta}^\pi(E) = (1 - \gamma) \beta(E). \quad (11)$$

 627

628 *Proof.* For any $\beta \in \mathcal{P}(S \times A)$, $\pi \in \mathcal{P}(A|S)$ and $E \in \mathcal{B}(S \times A)$, it holds that
 629

630
$$\begin{aligned} d_{J_\pi \beta}^\pi(E) &= (1 - \gamma) \sum_{n=0}^{\infty} \gamma^n (J_\pi^n J_\pi \beta)(E) \\ &= J_\pi d_\beta^\pi(E) \end{aligned}$$

 631

632 where we just used the associativity of the operator J_π . Furthermore by letting $m = n + 1$ it holds
 633 that
 634

635
$$\begin{aligned} d_{J_\pi \beta}^\pi(E) &= (1 - \gamma) \sum_{n=0}^{\infty} \gamma^n J_\pi^{n+1} \beta(E) \\ &= (1 - \gamma) \sum_{m=1}^{\infty} \gamma^{m-1} J_\pi^m \beta(E) \\ &= \frac{1 - \gamma}{\gamma} \sum_{m=1}^{\infty} \gamma^m J_\pi^m \beta(E) \\ &= \frac{1}{\gamma} (d_\beta^\pi(E) - (1 - \gamma) \beta(E)). \end{aligned}$$

 636

637 Rearranging concludes the proof. \square
 638

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

$$651 \quad V_\tau^*(s) = -\tau \ln \int_A \exp \left(-\frac{1}{\tau} Q_\tau^*(s, a) \right) \mu(da),$$

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

$$655 \quad Q_\tau^*(s, a) = c(s, a) + \gamma \int_S V_\tau^*(s') P(ds'|s, a), \quad \forall (s, a) \in S \times A.$$

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

$$659 \quad \pi^*(da|s) = \exp \left(-\frac{1}{\tau} (Q_\tau^*(s, a) - V_\tau^*(s)) \right) \mu(da), \quad \forall s \in S.$$

661 Finally, the value function V_τ^π is the unique bounded solution of the following Bellman equation for
662 all $s \in S$

$$663 \quad V_\tau^\pi(s) = \int_A \left(Q_\tau^\pi(s, a) + \tau \ln \frac{d\pi}{d\mu}(a, s) \right) \pi(da|s).$$

665 The performance difference lemma, first introduced for tabular unregularised MDPs, has become
666 fundamental in the analysis of MDPs as it acts a substitute for the strong convexity of the $\pi \mapsto V_\tau^\pi$ if
667 the state-occupancy measure d_ρ^π is ignored (e.g Kakade & Langford (2002), Zhang et al. (2021), Ju
668 & Lan (2024)). By virtue of Kerimkulov et al. (2025), we have the following performance difference
669 for entropy regularised MDPs in Polish state and action spaces.

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

$$672 \quad V_\tau^\pi(\rho) - V_\tau^{\pi'}(\rho) \\ 673 = \frac{1}{1-\gamma} \int_S \left[\int_A \left(Q_\tau^{\pi'}(s, a) + \tau \ln \frac{d\pi'}{d\mu}(a, s) \right) (\pi - \pi')(da|s) + \tau \text{KL}(\pi(\cdot|s) \|\pi'(\cdot|s)) \right] d_\rho^\pi(ds).$$

676 **Lemma A.3** (Grönwall). *Let $\lambda(s) \geq 0$, $a = a(s)$, $b = b(s)$ and $y = y(s)$ be locally integrable, real-
677 valued functions defined on $[0, T]$ such that y is also locally integrable and for almost all $s \in [0, T]$,*

$$678 \quad y(s) + a(s) \leq b(s) + \int_0^s \lambda(t) y(t) dt.$$

680 Then

$$681 \quad y(s) + a(s) \leq b(s) + \int_0^s \lambda(t) \left[\int_0^t \lambda(r) (b(r) - a(r)) dr \right] dt, \quad \forall s \in [0, T].$$

684 Furthermore, if b is monotone increasing and a is non-negative, then

$$685 \quad y(s) + a(s) \leq b(s) e^{\int_0^s \lambda(r) dr}, \quad \forall s \in [0, T].$$

688 B AUXILIARY RESULTS

690 **Lemma B.1.** *For some $\beta \in \mathcal{P}(S \times A)$, let $d_\beta^\pi \in \mathcal{P}(S \times A)$ be the state-action occupancy measure.
691 Moreover let $\kappa(ds, da, ds', da') := P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da)$. Then for any $\pi \in \Pi_\mu$ and any
692 integrable $f : S \times A \rightarrow \mathbb{R}$, it holds that*

$$693 \quad \int_{S \times A \times S \times A} f(s, a) f(s', a') \kappa(ds, da, ds', da') \leq \frac{1}{\sqrt{\gamma}} \int_{S \times A} f(s, a)^2 d_\beta^\pi(ds, da)$$

696 *Proof.* By Hölder's inequality, it holds that

$$698 \quad \int_{S \times A \times S \times A} f(s, a) f(s', a') \kappa(ds, da, ds', da') \quad (12) \\ 699 \\ 700 \leq \left(\int_{S \times A \times S \times A} f(s, a)^2 \kappa(ds, da, ds', da') \right)^{\frac{1}{2}} \left(\int_{S \times A \times S \times A} f(s', a')^2 \kappa(ds, da, ds', da') \right)^{\frac{1}{2}}.$$

702 Moreover, observe that
 703

$$\begin{aligned} 704 \int_{S \times A \times S \times A} f(s, a)^2 \kappa(ds, da, ds', da') &= \int_{S \times A} \left(\int_{S \times A} P^\pi(ds', da'|s, a) \right) f(s, a)^2 d_\beta^\pi(ds, da) \\ 705 \\ 706 &= \int_{S \times A} f(s, a)^2 d_\beta^\pi(ds, da), \\ 707 \end{aligned}$$

708 hence (12) becomes
 709

$$\begin{aligned} 710 \left(\int_{S \times A \times S \times A} f(s, a)^2 \kappa(ds, da, ds', da') \right)^{\frac{1}{2}} &\left(\int_{S \times A \times S \times A} f(s', a')^2 \kappa(ds, da, ds', da') \right)^{\frac{1}{2}} \\ 711 \\ 712 &\leq \left(\int_{S \times A} f(s, a)^2 d_\beta^\pi(ds, da) \right)^{\frac{1}{2}} \left(\int_{S \times A \times S \times A} f(s', a')^2 \kappa(ds, da, ds', da') \right)^{\frac{1}{2}}. \\ 713 \\ 714 \end{aligned}$$

715 Now by the first part of Lemma A.1, it holds that
 716

$$\begin{aligned} 717 \int_{S \times A \times S \times A} f(s', a')^2 \kappa(ds, da, ds', da') &= \int_{S \times A \times S \times A} f(s', a')^2 P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da) \\ 718 \\ 719 &= \int_{S \times A} f(s, a)^2 d_{J^\pi \beta}^\pi(ds, da), \\ 720 \end{aligned}$$

721 where $J^\pi : \mathcal{P}(S \times A) \rightarrow \mathcal{P}(S \times A)$ is defined in (10). Then by the second part of Lemma A.1 we
 722 have

$$\begin{aligned} 723 \left(\int_{S \times A} f(s, a)^2 d_\beta^\pi(ds, da) \right)^{\frac{1}{2}} &\left(\int_{S \times A \times S \times A} f(s', a')^2 \kappa(ds, da, ds', da') \right)^{\frac{1}{2}} \\ 724 \\ 725 &\leq \left(\int_{S \times A} f(s, a)^2 d_\beta^\pi(ds, da) \right)^{\frac{1}{2}} \left(\int_{S \times A} f(s, a)^2 d_{J^\pi \beta}^\pi(ds, da) \right)^{\frac{1}{2}} \\ 726 \\ 727 &\leq \frac{1}{\sqrt{\gamma}} \int_{S \times A} f(s, a)^2 d_\beta^\pi(ds, da), \\ 728 \\ 729 \end{aligned}$$

730 which concludes the proof.
 731

732 \square
 733

734 **Lemma B.2.** For some $\theta_0 \in \mathbb{R}^N$ and $\pi_0 \in \Pi_\mu$, let $\{\pi_t, \theta_t\}_{t \geq 0}$ be the trajectory of coupled actor-
 735 critic flow. Moreover let $K_t = \sup_{s \in S} \text{KL}(\pi_t(\cdot|s) \|\mu)$. There exists $C_1 > 0$ such that for all $t \geq 0$ it
 736 holds that

$$\begin{aligned} 737 \sup_{s \in S} |\partial_t \pi_t(\cdot|s)|_{\mathcal{M}(A)} &\leq |A_t|_{B_b(S \times A)}, \\ 738 \\ 739 |A_t|_{B_b(S \times A)} &\leq 2 |Q_t|_{B_b(S \times A)} + 2\tau \left| \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)}, \\ 740 \\ 741 |Q_\tau^{\pi_t}|_{B_b(S \times A)} &\leq \frac{1}{1-\gamma} \left(|c|_{B_b(S \times A)} + \tau \gamma K_t \right), \\ 742 \\ 743 \left| \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} &\leq C_1 + \frac{2}{\tau} \sup_{r \in [0, t]} |\theta_r| + \sup_{r \in [0, t]} K_r. \\ 744 \\ 745 \end{aligned}$$

746 *Proof.* The first claim $\sup_{s \in S} |\partial_t \pi_t(\cdot|s)|_{\mathcal{M}(A)} \leq |A_\tau^{\pi_t}|_{B_b(S \times A)}$ follows trivially from the definition
 747 of the approximate Fisher–Rao gradient flow defined in (8). Moreover, it holds that
 748

$$\begin{aligned} 749 |A_t|_{B_b(S \times A)} &= \left| Q_t + \tau \ln \frac{d\pi_t}{d\mu} - \int_A \left(Q_t(\cdot, a) + \tau \ln \frac{d\pi_t}{d\mu}(\cdot, a) \right) \pi_t(da|\cdot) \right|_{B_b(S \times A)} \\ 750 \\ 751 &\leq 2 \left| Q_t + \tau \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} \\ 752 \\ 753 &\leq 2 |Q_t|_{B_b(S \times A)} + 2\tau \left| \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} \\ 754 \\ 755 \end{aligned}$$

756 where we used the triangle inequality in the final inequality. Moreover, the state-action value function
 757 $Q_\tau^{\pi_t}$ is a fixed point of the Bellman operator defined in (2). Hence, for all $(s, a) \in S \times A$, we
 758 have

759

$$760 Q_\tau^{\pi_t}(s, a) = c(s, a) + \gamma \int_{S \times A} Q_\tau^{\pi_t}(s', a') P^{\pi_t}(ds', da'|s, a) + \tau \gamma \int_S \text{KL}(\pi_t(\cdot|s')\|\mu) P(ds'|s, a).$$

761

762 Taking absolute values and using the triangle inequality we have

763

$$764 |Q_\tau^{\pi_t}(s, a)| \leq |c|_{B_b(S \times A)} + \gamma |Q_\tau^{\pi_t}|_{B_b(S \times A)} + \tau \gamma \sup_{s' \in S} \text{KL}(\pi_t(\cdot|s')\|\mu)$$

765

$$766 = |c|_{B_b(S \times A)} + \gamma |Q_\tau^{\pi_t}|_{B_b(S \times A)} + \tau \gamma K_t.$$

767

768 Taking the supremum over $(s, a) \in S \times A$ on the left-hand side yields

769

$$770 |Q_\tau^{\pi_t}|_{B_b(S \times A)} \leq |c|_{B_b(S \times A)} + \gamma |Q_\tau^{\pi_t}|_{B_b(S \times A)} + \tau \gamma K_t.$$

771 Rearranging gives

772

$$(1 - \gamma) |Q_\tau^{\pi_t}|_{B_b(S \times A)} \leq |c|_{B_b(S \times A)} + \tau \gamma K_t,$$

773 which is the desired bound. Recall the approximate Fisher–Rao gradient flow for the policies
 774 $\{\pi_t\}_{t \geq 0}$, which for all $t \geq 0$ and for all $(s, a) \in S \times A$ is given by

775

$$776 \partial_t \ln \frac{d\pi_t}{d\mu}(s, a) = - \left(Q_t(s, a) + \tau \ln \frac{d\pi_t}{d\mu}(a, s) - \int_A \left(Q_t(s, a') + \tau \ln \frac{d\pi_t}{d\mu}(a', s) \right) \pi_t(da'|s) \right).$$

777

778 Duhamel’s principle yields for all $t \geq 0$ that

779

$$780 \ln \frac{d\pi_t}{d\mu}(s, a) = e^{-\tau t} \ln \frac{d\pi_0}{d\mu}(a, s) + \int_0^t e^{-\tau(t-r)} \left(\int_A Q_r(s, a') \pi_r(da'|s) - Q_r(s, a) \right) dr \quad (13)$$

781

$$782 + \tau \int_0^t e^{-\tau(t-r)} \text{KL}(\pi_r(\cdot|s)\|\mu) dr.$$

783

784 Since $\pi_0 \in \Pi_\mu$, there exists $C_1 \geq 1$ such that $\left| \ln \frac{d\pi_0}{d\mu} \right|_{B_b(S \times A)} \leq C_1$. Then by Assumption 4.2 we
 785 have that for all $t \geq 0$,

786

$$787 \left| \ln \frac{d\pi_t}{d\mu}(s, a) \right| \leq C_1 + \int_0^t e^{-\tau(t-r)} \left| \int_A Q_r(s, a') \pi_r(da'|s) - Q_r(s, a) \right| dr$$

788

$$789 + \tau \int_0^t e^{-\tau(t-r)} \text{KL}(\pi_r(\cdot|s)\|\mu) dr$$

790

$$791 \leq C_1 + 2 \int_0^t e^{-\tau(t-r)} |\theta_r| dr + \tau \int_0^t e^{-\tau(t-r)} K_r dr$$

792

$$793 \leq C_1 + \frac{2}{\tau} \sup_{r \in [0, t]} |\theta_r| + \sup_{r \in [0, t]} K_r,$$

794

795 where in the last inequality we used $\int_0^t e^{-\tau(t-r)} dr \leq \frac{1}{\tau}$. Taking the supremum over $(s, a) \in S \times A$
 796 yields

797

$$798 \left| \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} \leq C_1 + \frac{2}{\tau} \sup_{r \in [0, t]} |\theta_r| + \sup_{r \in [0, t]} K_r,$$

799 which is the desired bound. \square

800 **Lemma B.3.** *Let Assumption 4.3 hold. Then for all $\pi \in \Pi_\mu$, it holds that $L(\cdot, \pi; d_\beta^\pi)$ is $\lambda_\beta(1 - \gamma)$ -
 801 strongly convex.*

802 *Proof.* For any $\xi \in \mathcal{P}(S \times A)$, let $\Sigma_\xi := \int_{S \times A} \phi(s, a) \phi(s, a)^\top \xi(ds, da) \in \mathbb{R}^{N \times N}$. Then by
 803 Lemma A.1 and Assumption 4.3 it holds that $\Sigma_{d_\beta^\pi} \succeq (1 - \gamma) \Sigma_\beta \succeq (1 - \gamma) \lambda_\beta I$ and thus $L(\cdot, \pi; d_\beta^\pi)$
 804 is $\lambda_\beta(1 - \gamma)$ -strongly convex. \square

810 B.1 PROOF OF LEMMA 4.1
811

812 *Proof.* Recall that $Q(s, a) = \langle \theta, \phi(s, a) \rangle$ for some $\theta \in \mathbb{R}^N$ and that for all $\pi \in \Pi_\mu$, there exists
813 $\theta_\pi \in \mathbb{R}^N$ such that $Q^\pi(s, a) = \langle \theta_\pi, \phi(s, a) \rangle$ by Assumption 4.1. Then by definition of the semi-
814 gradient of the MSBE $g : \mathbb{R}^N \times \mathcal{P}(A|S) \rightarrow \mathbb{R}^N$ in (6), it holds that

$$\begin{aligned}
815 \langle g(\theta, \pi), \theta - \theta_\pi \rangle &= \left\langle \int_{S \times A} (Q(s, a) - \mathbf{T}^\pi Q(s, a)) \phi(s, a) d_\beta^\pi(da, ds), \theta - \theta_\pi \right\rangle \\
816 &= \left\langle \int_{S \times A} (Q(s, a) - Q_\tau^\pi(s, a)) \phi(s, a) d_\beta^\pi(da, ds), \theta - \theta_\pi \right\rangle \\
817 &\quad + \left\langle \int_{S \times A} (Q_\tau^\pi(s, a) - \mathbf{T}^\pi Q(s, a)) \phi(s, a) d_\beta^\pi(da, ds), \theta - \theta_\pi \right\rangle \\
818 &= \left\langle \int_{S \times A} (Q(s, a) - Q_\tau^\pi(s, a)) \phi(s, a) d_\beta^\pi(da, ds), \theta - \theta_\pi \right\rangle \\
819 &\quad - \gamma \left\langle \int_{S \times A \times S \times A} (Q(s', a') - Q_\tau^\pi(s', a')) \phi(s, a) P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da), \theta - \theta_\pi \right\rangle,
\end{aligned}$$

820 where we added and subtracted the true state-action value function $Q_\tau^\pi \in B_b(S \times A)$ in the second
821 equality and used the fact that it is a fixed point of the Bellman operator defined in (2). To ease
822 notation, let $\varepsilon(s, a) := Q(s, a) - Q_\tau^\pi(s, a)$. Multiplying both sides by -1 and using the associativity
823 of the inner product, we have

$$\begin{aligned}
824 - \langle g(\theta, \pi), \theta - \theta_\pi \rangle &= - \left\langle \int_{S \times A} \varepsilon(s, a) \phi(s, a) d_\beta^\pi(da, ds), \theta - \theta_\pi \right\rangle \\
825 &\quad + \gamma \left\langle \int_{S \times A} \varepsilon(s', a') \phi(s, a) P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da), \theta - \theta_\pi \right\rangle \\
826 &= - \int_{S \times A} \varepsilon(s, a) \langle \phi(s, a), \theta - \theta_\pi \rangle d_\beta^\pi(da, ds) \\
827 &\quad + \gamma \int_{S \times A} \varepsilon(s', a') \langle \phi(s, a), \theta - \theta_\pi \rangle P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da) \\
828 &= - \int_{S \times A} \varepsilon(s, a)^2 d_\beta^\pi(da, ds) \\
829 &\quad + \gamma \int_{S \times A \times S \times A} \varepsilon(s, a) \varepsilon(s', a') P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da) \\
830 &= I^{(1)} + \gamma I^{(2)}.
\end{aligned}$$

831 Now applying Lemma B.1 to $I^{(2)}$ we have
832

$$\begin{aligned}
833 I^{(2)} &:= \int_{S \times A \times S \times A} \varepsilon(s, a) \varepsilon(s', a') P^\pi(ds', da'|s, a) d_\beta^\pi(ds, da) \\
834 &\leq \frac{1}{\sqrt{\gamma}} \int_{S \times A} \varepsilon(s, a)^2 d_\beta^\pi(ds, da).
\end{aligned}$$

835 Thus it holds that
836

$$\begin{aligned}
837 - \langle g(\theta, \pi), \theta - \theta_\pi \rangle &\leq I^{(1)} + \gamma I^{(2)} \\
838 &\leq -(1 - \sqrt{\gamma}) \int_{S \times A} \varepsilon(s, a)^2 d_\beta^\pi(da, ds) \\
839 &= -(1 - \sqrt{\gamma}) \int_{S \times A} (Q(s, a) - Q_\tau^\pi(s, a))^2 d_\beta^\pi(da, ds) \\
840 &= -(1 - \sqrt{\gamma}) \langle \nabla_\theta L(\theta, \pi; d_\beta^\pi), \theta - \theta_\pi \rangle,
\end{aligned}$$

841 where the last inequality follows from the Assumption 4.1 and the definition of $Q(s, a) =$
842 $\langle \theta, \phi(s, a) \rangle$.
843

□

864 C PROOF OF STABILITY RESULTS

866 C.1 PROOF OF LEMMA 5.1

868 *Proof.* Consider the following equation

$$\begin{aligned}
 870 \quad & \frac{1}{2\eta_t} \frac{d}{dt} |\theta_t|^2 = \frac{1}{\eta_t} \left\langle \frac{d}{dt} \theta_t, \theta_t \right\rangle \\
 871 \quad & = -\langle g(\theta_t, \pi_t), \theta_t \rangle \\
 872 \quad & = -\left\langle \int_{S \times A} (Q_t(s, a) - T^{\pi_t} Q_t(s, a)) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 873 \quad & = -\left\langle \int_{S \times A} Q_t(s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 874 \quad & \quad + \left\langle \int_{S \times A} T^{\pi_t} Q_t(s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 875 \quad & := -J_t^{(1)} + J_t^{(2)}
 \end{aligned} \tag{14}$$

882 where we used the θ_t dynamics from (7) in the second equality and the definition of the semi-gradient
 883 in the third equality. For any $\pi \in \Pi_{\mu}$, let $\Sigma^{\pi} \in \mathbb{R}^{N \times N}$ be

$$\Sigma^{\pi} = \int_{S \times A} \phi(s, a) \phi(s, a)^{\top} d_{\beta}^{\pi}(da, ds).$$

887 Then by definition we have that $Q_t(s, a) = \langle \theta_t, \phi(s, a) \rangle$, hence for $J_t^{(1)}$ we have

$$\begin{aligned}
 889 \quad J_t^{(1)} &= \left\langle \int_{S \times A} Q_t(s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 890 \quad &= \left\langle \int_{S \times A} \langle \theta_t, \phi(s, a) \rangle \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 891 \quad &= \left\langle \theta_t, \left(\int_{S \times A} \phi(s, a) \phi(s, a)^{\top} d_{\beta}^{\pi_t}(da, ds) \right) \theta_t \right\rangle \\
 892 \quad &= \langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle
 \end{aligned} \tag{15}$$

897 Now dealing with $J_t^{(1)}$, expanding the Bellman operator defined in (2) we have

$$\begin{aligned}
 899 \quad J_t^{(2)} &= \left\langle \int_{S \times A} T^{\pi_t} Q_t(s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 900 \quad &= \left\langle \int_{S \times A} c(s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 901 \quad & \quad + \gamma \left\langle \int_{S \times A} \langle \theta_t, \phi(s', a') \rangle \phi(s, a) P^{\pi_t}(ds', da' | s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\
 902 \quad & \quad + \tau \gamma \left\langle \int_{S \times A} \left(\int_S \text{KL}(\pi_t(\cdot | s'), \mu) P(ds' | s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds) \right), \theta_t \right\rangle \\
 903 \quad & \leq |c|_{B_b(S \times A)} |\theta_t| + \gamma I_t^{(1)} + \tau \gamma I_t^{(2)}
 \end{aligned}$$

910 where we defined

$$\begin{aligned}
 911 \quad I_t^{(1)} &= \left\langle \int_{S \times A} \langle \theta_t, \phi(s', a') \rangle \phi(s, a) P^{\pi_t}(ds', da' | s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle, \\
 912 \quad I_t^{(2)} &= \left\langle \int_{S \times A} \left(\int_S \text{KL}(\pi_t(\cdot | s'), \mu) P(ds' | s, a) \phi(s, a) d_{\beta}^{\pi_t}(da, ds) \right), \theta_t \right\rangle.
 \end{aligned}$$

913 Moreover, to ease notation let

$$K_t := \sup_{s \in S} \text{KL}(\pi_t(\cdot | s) | \mu)$$

918 and temporarily let $\kappa_t(ds, da, ds', da') := P^{\pi_t}(ds', da'|s, a)d_{\beta}^{\pi_t}(da, ds)$. Now focusing on $I_t^{(1)}$, it
919 holds that
920

$$\begin{aligned} 921 \quad I_t^{(1)} &= \left\langle \int_{S \times A \times S \times A} \langle \theta_t, \phi(s', a') \rangle \phi(s, a) \kappa_t(da', ds', da, ds), \theta_t \right\rangle \\ 922 \\ 923 \quad &= \int_{S \times A \times S \times A} \langle \theta_t, \phi(s, a) \rangle \langle \theta_t, \phi(s', a') \rangle \kappa_t(ds', da', ds, da). \\ 924 \\ 925 \end{aligned}$$

926 Now using Lemma B.1 with $f = \langle \theta, \phi(\cdot, \cdot) \rangle$ we have

$$\begin{aligned} 927 \quad I_t^{(1)} &\leq \frac{1}{\sqrt{\gamma}} \left(\int_{S \times A} \langle \theta_t, \phi(s, a) \rangle^2 d_{\beta}^{\pi_t}(ds, da) \right)^{\frac{1}{2}} \left(\int_{S \times A} \langle \theta_t, \phi(s, a) \rangle^2 d_{\beta}^{\pi_t}(ds, da) \right)^{\frac{1}{2}} \\ 928 \\ 929 \quad &= \frac{1}{\sqrt{\gamma}} \int_{S \times A} \langle \theta_t, \phi(s, a) \rangle^2 d_{\beta}^{\pi_t}(ds, da) \\ 930 \\ 931 \quad &= \frac{1}{\sqrt{\gamma}} \langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle. \\ 932 \\ 933 \end{aligned}$$

934 Thus all together it holds that

$$\gamma I_t^{(1)} \leq \sqrt{\gamma} \langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle.$$

935 Now focusing on $I_t^{(2)}$, we have

$$\begin{aligned} 936 \quad I_t^{(2)} &= \left\langle \int_{S \times A} \left(\int_S \text{KL}(\pi_t(\cdot|s'), \mu) P(ds'|s, a) \right) \phi(s, a) d_{\beta}^{\pi_t}(da, ds), \theta_t \right\rangle \\ 937 \\ 938 \quad &\leq K_t \left| \int_{S \times A} \phi(s, a) d_{\beta}^{\pi_t}(ds, da) \right| |\theta_t| \\ 939 \\ 940 \quad &\leq K_t |\theta_t| \\ 941 \\ 942 \end{aligned}$$

943 where we used Assumption 4.2 in the final inequality. Hence along with (15), (14) becomes

$$\begin{aligned} 944 \quad \frac{1}{2\eta_t} \frac{d}{dt} |\theta_t|^2 &\leq -J_t^{(1)} + J_t^{(2)} \\ 945 \\ 946 \quad &\leq -\langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle + |c|_{B_b(S \times A)} |\theta_t| + \gamma I_t^{(1)} + \tau \gamma I_t^{(2)} \\ 947 \\ 948 \quad &\leq -\langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle + \sqrt{\gamma} \langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle + |c|_{B_b(S \times A)} |\theta_t| + \tau \gamma K_t |\theta_t| \\ 949 \\ 950 \quad &= -(1 - \sqrt{\gamma}) \langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle + (|c|_{B_b(S \times A)} + \tau \gamma K_t) |\theta_t|. \\ 951 \\ 952 \end{aligned} \tag{16}$$

953 Observe that by (11) and Assumption 4.2, $\Sigma^{\pi} \in \mathbb{R}^{N \times N}$ is positive definite for all $\pi \in \mathcal{P}(A|S)$,
954 hence it holds that

$$\langle \theta_t, \Sigma^{\pi_t} \theta_t \rangle \geq (1 - \gamma) \lambda_{\beta} |\theta_t|^2.$$

955 Therefore (16) becomes

$$\frac{1}{2\eta_t} \frac{d}{dt} |\theta_t|^2 \leq -(1 - \sqrt{\gamma})(1 - \gamma) \lambda_{\beta} |\theta_t|^2 + (|c|_{B_b(S \times A)} + \tau \gamma K_t) |\theta_t|$$

956 Let $\Gamma := \lambda_{\beta}(1 - \gamma)(1 - \sqrt{\gamma})$. By Young's inequality, there exists $\epsilon > 0$ such that

$$\begin{aligned} 957 \quad \frac{1}{2\eta_t} \frac{d}{dt} |\theta_t|^2 &\leq -\Gamma |\theta_t|^2 + \frac{\epsilon}{2} |\theta_t|^2 + \frac{(|c|_{B_b(S \times A)} + \tau \gamma K_t)^2}{2\epsilon} \\ 958 \\ 959 \quad &\leq -\Gamma |\theta_t|^2 + \frac{\epsilon}{2} |\theta_t|^2 + \frac{|c|_{B_b(S \times A)}^2 + \tau^2 \gamma^2 K_t^2}{\epsilon}, \\ 960 \\ 961 \end{aligned}$$

962 where we used the identity $(a + b)^2 \leq 2a^2 + 2b^2$. Choosing $\epsilon = \Gamma$ we arrive at

$$\frac{1}{2\eta_t} \frac{d}{dt} |\theta_t|^2 \leq -\frac{\Gamma}{2} |\theta_t|^2 + \frac{\tau^2 \gamma^2 K_t^2}{\Gamma} + \frac{|c|_{B_b(S \times A)}^2}{\Gamma}$$

963 which concludes the proof. \square

972 C.2 PROOF OF THEOREM 5.1
 973

974 *Proof.* By Lemma 5.1, we have that for all $r \geq 0$

$$975 \quad 976 \quad 977 \quad \frac{1}{2\eta_r} \frac{d}{dr} |\theta_r|^2 \leq -\frac{\Gamma}{2} |\theta_r|^2 + \frac{\tau^2 \gamma^2 K_r^2}{\Gamma} + \frac{|c|_{B_b(S \times A)}^2}{\Gamma}.$$

978 Rearranging, it holds that for all $t \geq 0$

$$979 \quad 980 \quad 981 \quad |\theta_r|^2 \leq -\frac{1}{\Gamma \eta_r} \frac{d}{dr} |\theta_r|^2 + \frac{2|c|_{B_b(S \times A)}^2 + 2\tau^2 \gamma^2 K_r^2}{\Gamma^2}.$$

982 Multiplying both sides by $e^{-\tau(t-r)}$ and integrating over r from 0 to t we have that for all $t \geq 0$

$$983 \quad 984 \quad 985 \quad \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr \leq -\frac{1}{\Gamma} \int_0^t e^{-\tau(t-r)} \frac{1}{\eta_r} \frac{d}{dr} |\theta_r|^2 dr + \frac{2|c|_{B_b(S \times A)}^2}{\Gamma^2} \int_0^t e^{-\tau(t-r)} dr \quad (17)$$

$$986 \quad 987 \quad + \frac{2\tau^2 \gamma^2}{\Gamma^2} \int_0^t e^{-\tau(t-r)} K_r^2 dr$$

$$988 \quad 989 \quad \leq -\frac{1}{\Gamma} \int_0^t e^{-\tau(t-r)} \frac{1}{\eta_r} \frac{d}{dr} |\theta_r|^2 dr + \frac{2|c|_{B_b(S \times A)}^2}{\Gamma^2 \tau} + \frac{2\tau^2 \gamma^2}{\Gamma^2} \int_0^t e^{-\tau(t-r)} K_r^2 dr,$$

990 where we used that $\int_0^t e^{-\tau(t-r)} dr \leq \frac{1}{\tau}$. Integrating the first term by parts, we have

$$991 \quad 992 \quad 993 \quad - \int_0^t e^{-\tau(t-r)} \frac{1}{\eta_r} \frac{d}{dr} |\theta_r|^2 dr = -\frac{|\theta_t|^2}{\eta_t} + e^{-\tau t} \frac{|\theta_0|^2}{\eta_0} + \tau \int_0^t |\theta_r|^2 \frac{e^{-\tau(t-r)}}{\eta_r} dr \quad (18)$$

$$994 \quad 995 \quad 996 \quad - \int_0^t |\theta_r|^2 \frac{e^{-\tau(t-r)} \frac{d}{dr} \eta_r}{\eta_r^2} dr.$$

997 Since by definition we have that for all $t \geq 0$, $\eta_t \geq 1$ and $\frac{d}{dt} \eta_t \geq 0$ it holds that

$$1000 \quad 1001 \quad 1002 \quad \int_0^t |\theta_r|^2 \frac{e^{-\tau(t-r)} \frac{d}{dr} \eta_r}{\eta_r^2} dr \geq 0.$$

1003 Hence dropping the negative terms on the right hand side of (18) and using that $\eta_t \geq \eta_0$ for all
 1004 $t \geq 0$, we have

$$1005 \quad 1006 \quad 1007 \quad -\frac{1}{\Gamma} \int_0^t e^{-\tau(t-r)} \frac{1}{\eta_r} \frac{d}{dr} |\theta_r|^2 dr \leq e^{-\tau t} \frac{|\theta_0|^2}{\Gamma \eta_0} + \frac{\tau}{\Gamma \eta_0} \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr.$$

1008 Substituting this back into (17), for all $t \geq 0$ we have that

$$1009 \quad 1010 \quad 1011 \quad \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr \leq e^{-\tau t} \frac{|\theta_0|^2}{\Gamma \eta_0} + \frac{\tau}{\Gamma \eta_0} \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr$$

$$1012 \quad 1013 \quad + \frac{2|c|_{B_b(S \times A)}^2}{\Gamma^2 \tau} + \frac{2\tau^2 \gamma^2}{\Gamma^2} \int_0^t e^{-\tau(t-r)} K_r^2 dr.$$

1014 Grouping like terms we have

$$1015 \quad \left(1 - \frac{\tau}{\Gamma \eta_0}\right) \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr \leq e^{-\tau t} \frac{|\theta_0|^2}{\Gamma \eta_0} + \frac{2|c|_{B_b(S \times A)}^2}{\Gamma^2 \tau} + \frac{2\tau^2 \gamma^2}{\Gamma^2} \int_0^t e^{-\tau(t-r)} K_r^2 dr.$$

1016 Recall that we have $\eta_0 > \frac{\tau}{\Gamma}$ to ensure that $1 - \frac{\tau}{\Gamma \eta_0} > 0$. Dividing through by $1 - \frac{\tau}{\Gamma \eta_0}$ gives for all
 1017 $t \geq 0$ that

$$1018 \quad 1019 \quad 1020 \quad \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr \leq \sigma_1 + \sigma_2 \int_0^t e^{-\tau(t-r)} K_r^2 dr \quad (19)$$

1021 where we've set

$$1022 \quad 1023 \quad \sigma_1 := \frac{|\theta_0|^2}{\Gamma \eta_0 \left(1 - \frac{\tau}{\Gamma \eta_0}\right)} + \frac{2|c|_{B_b(S \times A)}^2}{\Gamma^2 \tau \left(1 - \frac{\tau}{\Gamma \eta_0}\right)},$$

$$\sigma_2 := \frac{2\tau^2\gamma^2}{\Gamma^2 \left(1 - \frac{\tau}{\Gamma\eta_0}\right)}.$$

Recall the approximate Fisher–Rao gradient flow for the policies $\{\pi_t\}_{t \geq 0}$, which for all $t \geq 0$ and for all $s \in S, a \in A$ is

$$\partial_t \ln \frac{d\pi_t}{d\mu}(s, a) = - \left(Q_t(s, a) + \tau \ln \frac{d\pi_t}{d\mu}(a, s) - \int_A \left(Q_t(s, a) + \tau \ln \frac{d\pi_t}{d\mu}(a, s) \right) \pi_t(da|s) \right)$$

Duhamel’s principle yields for all $t \geq 0$ that

$$\begin{aligned} \ln \frac{d\pi_t}{d\mu}(s, a) &= e^{-\tau t} \ln \frac{d\pi_0}{d\mu}(a, s) + \int_0^t e^{-\tau(t-r)} \left(\int_A Q_r(s, a) \pi_r(da|s) - Q_r(s, a) \right) dr \quad (20) \\ &\quad + \tau \int_0^t e^{-\tau(t-r)} \text{KL}(\pi_r(\cdot|s) | \mu) dr \end{aligned}$$

Observe that since $\pi_0 \in \Pi_\mu$, there exists $C_1 \geq 1$ such that $\ln \left| \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} \leq C_1$. Using that $e^{-\tau t} \leq 1$ and assumption 4.2 gives that for all $t \geq 0$

$$\begin{aligned} \ln \frac{d\pi_t}{d\mu}(s, a) &\leq C_1 + 2 \int_0^t e^{-\tau(t-r)} |\theta_r| dr + \tau \int_0^t e^{-\tau(t-r)} \text{KL}(\pi_r(\cdot|s) | \mu) dr \\ &\leq C_1 + 2 \int_0^t e^{-\tau(t-r)} |\theta_r| dr + \tau \int_0^t e^{-\tau(t-r)} K_r dr \end{aligned}$$

Integrating over the actions with respect to $\pi_t(\cdot|s) \in \mathcal{P}(A)$ gives for all $t \geq 0$ that

$$\text{KL}(\pi_t(\cdot|s) | \mu) \leq C_1 + 2 \int_0^t e^{-\tau(t-r)} |\theta_r| dr + \tau \int_0^t e^{-\tau(t-r)} K_r dr$$

where we again use that $K_r = \sup_{s \in S} \text{KL}(\pi_r(\cdot|s) | \mu)$. Following from the techniques in Liu et al. (2023), observe that from (20) and Assumption 4.2 we similarly get for all $t \geq 0$ that

$$\ln \frac{d\mu}{d\pi_t}(a, s) = - \ln \frac{d\pi_t}{d\mu}(s, a) \leq C_1 + 2 \int_0^t e^{-\tau(t-r)} |\theta_r| dr - \tau \int_0^t e^{-\tau(t-r)} K_r dr.$$

Now integrating over the actions with respect to the reference measure $\mu \in \mathcal{P}(A)$ we have

$$\text{KL}(\mu | \pi_t(\cdot|s)) \leq C_1 + 2 \int_0^t e^{-\tau(t-r)} |\theta_r| dr - \tau \int_0^t e^{-\tau(t-r)} K_r dr$$

Moreover, using the non-negativity of the KL divergence, it holds for all $t \geq 0$ that

$$\text{KL}(\pi_t(\cdot|s) | \mu) \leq \text{KL}(\pi_t(\cdot|s) | \mu) + \text{KL}(\mu | \pi_t(\cdot|s)) \leq 2C_1 + 4 \int_0^t e^{-\tau(t-r)} |\theta_r| dr$$

Since this holds for any $s \in S$, it holds for all $t \geq 0$ that

$$K_t \leq 2C_1 + 4 \int_0^t e^{-\tau(t-r)} |\theta_r| dr$$

Now squaring both sides and using the Hölder’s inequality, we have

$$\begin{aligned} K_t^2 &\leq \left(2C_1 + 4 \int_0^t e^{-\tau(t-r)} |\theta_r| dr \right)^2 \\ &\leq 8(C_1)^2 + 32 \left(\int_0^t e^{-\tau(t-r)} |\theta_r| dr \right)^2 \\ &= 8(C_1)^2 + 32 \left(\int_0^t e^{-\frac{\tau}{2}(t-r)} e^{-\frac{\tau}{2}(t-r)} |\theta_r| dr \right)^2 \\ &\leq 8(C_1)^2 + 32 \left(\int_0^t e^{-\tau(t-r)} dr \right) \left(\int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr \right) \\ &\leq 8(C_1)^2 + \frac{32}{\tau} \int_0^t e^{-\tau(t-r)} |\theta_r|^2 dr, \end{aligned} \quad (21)$$

1080 where we again used $\int_0^t e^{-\tau(t-r)} dr \leq \frac{1}{\tau}$. We can now substitute (19) into (21) to arrive at
 1081

$$\begin{aligned} 1082 \quad K_t^2 &\leq 8(C_1)^2 + \frac{32}{\tau}\sigma_1 + \frac{32}{\tau}\sigma_2 \int_0^t e^{-\tau(t-r)} K_r^2 dr \\ 1083 \\ 1084 \quad &:= a_1 + a_2 \int_0^t e^{-\tau(t-r)} K_r^2 dr \\ 1085 \\ 1086 \end{aligned}$$

1087 with $a_1 = 8(C_1)^2 + \frac{32}{\tau}\sigma_1$ and $a_2 = \frac{32\sigma_2}{\tau}$. □
 1088

1089 **C.3 PROOF OF COROLLARY 5.1**

1090 *Proof.* By Theorem 5.1 it holds that
 1091

$$1092 \quad K_t^2 \leq a_1 + a_2 \int_0^t e^{-\tau(t-r)} K_r^2 dr. \\ 1093 \\ 1094$$

1095 Observe that by multiplying through by $e^{\tau t}$, we can rewrite this as
 1096

$$1097 \quad e^{\tau t} K_t^2 \leq e^{\tau t} a_1 + a_2 \int_0^t e^{\tau r} K_r^2 dr. \\ 1098$$

1099 Hence after defining $g(t) = e^{\tau t} K_t^2$ and applying Grönwall's inequality (Lemma A.3), for all $\gamma \in$
 1100 $(0, 1)$ it holds for all $t \geq 0$ that
 1101

$$K_t^2 \leq a_1 e^{a_2 t}.$$

1102 □

1103 **C.4 PROOF OF COROLLARY 5.2**

1104 *Proof.* By Corollary 5.1 and Lemma 5.1, for all $\gamma \in (0, 1)$ it holds that
 1105

$$1106 \quad \frac{1}{2} \frac{d}{dt} |\theta_t|^2 \leq -\frac{\Gamma}{2} \eta_t |\theta_t|^2 + b_t \eta_t$$

1107 such that
 1108

$$1109 \quad b_t = \left(\frac{2|c|_{B_b(S \times A)}^2 + 2\tau^2 \gamma^2 a_1 e^{a_2 t}}{\Gamma^2} \right).$$

1110 Recall that there exists $\alpha > 0$ such that $\frac{d}{dt} \eta_t \leq \alpha \eta_t$, then another application of Grönwall's Lemma
 1111 then concludes the proof. □
 1112

1113 **D PROOF OF CONVERGENCE RESULTS**

1114 **D.1 PROOF OF LEMMA 6.1**

1115 *Proof.* By the definition of the state-action value function (1) it holds that
 1116

$$\begin{aligned} 1117 \quad \frac{d}{dt} Q_\tau^{\pi_t}(s, a) &= \lim_{h \rightarrow 0} \frac{Q_\tau^{\pi_{t+h}}(s, a) - Q_\tau^{\pi_t}(s, a)}{h} \\ 1118 \\ 1119 \quad &= \gamma \int_S \frac{d}{dt} V_\tau^{\pi_t}(s') P(ds'|s, a). \end{aligned}$$

1120 Now observe that by Kerimkulov et al. (2025)[Proof of Proposition 2.6], we have
 1121

$$1122 \quad \frac{d}{dt} V_\tau^{\pi_t}(s) = \frac{1}{1-\gamma} \int_{S \times A} A_\tau^{\pi_t}(s, a) \partial_t \pi_t(da|s') d^{\pi_t}(ds'|s).$$

1123 Thus we have
 1124

$$1125 \quad \frac{d}{dt} Q_\tau^{\pi_t}(s, a) = \frac{\gamma}{1-\gamma} \int_S \left(\int_{S \times A} A_\tau^{\pi_t}(s'', a'') \partial_t \pi_t(da''|s'') d^{\pi_t}(ds''|s') \right) P(ds'|s, a).$$

1126 □

1134 D.2 PROOF OF THEOREM 6.1
11351136 *Proof.* Recall the performance difference Lemma (Lemma A.2): for all $\rho \in \mathcal{P}(S)$ and $\pi, \pi' \in \Pi_\mu$,

1137
1138
$$V_\tau^\pi(\rho) - V_\tau^{\pi'}(\rho) = \frac{1}{1-\gamma} \int_S \left[\int_A \left(Q_\tau^{\pi'}(s, a) + \tau \ln \frac{d\pi'}{d\mu}(a, s) \right) (\pi - \pi')(da|s) + \tau \text{KL}(\pi(\cdot|s) \|\pi'(\cdot|s)) \right] d_\rho^\pi(ds).$$

1139
1140
1141

1142 Now let $\pi = \pi^*$ and $\pi' = \pi_t$ and multiply both sides by -1 we have
1143
1144

1145
$$V_\tau^{\pi_t}(\rho) - V_\tau^{\pi^*}(\rho) = \frac{-1}{1-\gamma} \int_S \left(\int_A \left(Q^{\pi_t}(s, a) + \tau \ln \frac{d\pi_t}{d\mu}(a, s) \right) (\pi^* - \pi_t)(da|s) + \tau \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s)) \right) d_\rho^{\pi^*}(ds). \quad (22)$$

1146
1147
1148
1149
1150

1151 Recall the approximate Fisher–Rao dynamics, which we write as
1152

1153
$$\partial_t \ln \frac{d\pi_t}{d\mu}(s, a) + \left(Q_t(s, a) + \tau \ln \frac{d\pi_t}{d\mu}(a, s) - \int_A \left(Q_t(s, a') + \tau \ln \frac{d\pi_t}{d\mu}(a', s) \right) \pi_t(da'|s) \right) = 0. \quad (23)$$

1154
1155

1156 Observe that since the normalisation constant (enforcing the conservation of mass along the flow)
1157 $\int_A \left(Q_t(s, a) + \tau \ln \frac{d\pi_t}{d\mu}(a, s) \right) \pi_t(da|s)$ is independent of $a \in A$, it holds that
1158

1159
$$\int_A \left(\int_A \left(Q_t(s, a') + \tau \ln \frac{d\pi_t}{d\mu}(a', s) \right) \pi_t(da'|s) \right) (\pi^* - \pi_t)(da|s) = 0.$$

1160
1161

1162 Hence adding 0 in the form of (23) into (22) it holds that for all $t \geq 0$
1163

1164
$$V_\tau^{\pi_t}(\rho) - V_\tau^{\pi^*}(\rho) = \frac{1}{1-\gamma} \left(\int_{S \times A} \partial_t \ln \frac{d\pi_t}{d\mu}(a, s) (\pi^* - \pi_t)(da|s) d_\rho^{\pi^*}(ds) \right. \quad (24)$$

1165
1166
1167
$$\left. + \int_{S \times A} (Q_t(s, a) - Q^{\pi_t}(s, a)) (\pi^* - \pi_t)(da|s) d_\rho^{\pi^*}(ds) - \tau \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s)) d_\rho^{\pi^*}(ds) \right).$$

1168
1169

1170 By (Kerimkulov et al., 2024, Lemma 3.8) and Corollary 5.1, for any fixed $\nu \in \Pi_\mu$, the map $t \rightarrow$
1171 $\text{KL}(\nu \|\pi_t)$ is differentiable. Hence we have
1172

1173
$$\int_A \partial_t \ln \frac{d\pi_t}{d\mu}(s, a) (\pi^* - \pi_t)(da|s) = \int_A \partial_t \ln \frac{d\pi_t}{d\mu}(s, a) \pi^*(da|s) - \int_A \partial_t \ln \frac{d\pi_t}{d\mu}(s, a) \pi_t(da|s)$$

1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
27

1188 Focusing on the second term, we have
 1189

$$\begin{aligned}
 1190 \int_{S \times A} (Q_t(s, a) - Q^{\pi_t}(s, a))(\pi^* - \pi_t)(da|s)d_{\rho}^{\pi^*}(ds) \\
 1191 \leq |Q_t(s, a) - Q^{\pi_t}(s, a)|_{B_b(S \times A)} \int_S \text{TV}(\pi^*(\cdot|s), \pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) \\
 1192 \leq \frac{1}{\sqrt{2}} |\theta_t - \theta_{\pi_t}| \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))^{1/2} d_{\rho}^{\pi^*}(ds) \\
 1193 \leq \frac{1}{\sqrt{2}} |\theta_t - \theta_{\pi_t}| \left(\int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) \right)^{1/2},
 \end{aligned}$$

1194 where we used Pinsker's Inequality in the second inequality and Hlder's inequality in the final inequality.
 1195 Now applying Young's inequality, there exists $\epsilon > 0$ such that
 1196

$$1202 |\theta_t - \theta_{\pi_t}| \left(\int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) \right)^{1/2} \leq \frac{1}{2\epsilon} |\theta_t - \theta_{\pi_t}|^2 + \frac{\epsilon}{2} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds).$$

1203 Substituting this back into (25) and choosing $\epsilon = \sqrt{2}\tau$ we have
 1204

$$\begin{aligned}
 1205 V_{\tau}^{\pi_t}(\rho) - V_{\tau}^{\pi^*}(\rho) &= \frac{1}{1-\gamma} \left(-\frac{d}{dt} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) \right. \\
 1206 &\quad \left. - \frac{\tau}{2} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) + \frac{1}{4\tau} |\theta_t - \theta_{\pi_t}|^2 \right).
 \end{aligned}$$

1207 Rearranging, we arrive at
 1208

$$\begin{aligned}
 1209 \frac{d}{dt} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) &\leq -\frac{\tau}{2} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) \\
 1210 &\quad - (1-\gamma) \left(V_{\tau}^{\pi_t}(\rho) - V_{\tau}^{\pi^*}(\rho) \right) + \frac{1}{4\tau} |\theta_t - \theta_{\pi_t}|^2.
 \end{aligned}$$

1211 Applying Duhamel's principle yields
 1212

$$\begin{aligned}
 1213 \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) &\leq e^{-\frac{\tau}{2}t} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_0(\cdot|s))d_{\rho}^{\pi^*}(ds) \\
 1214 &\quad - (1-\gamma) \int_0^t e^{-\frac{\tau}{2}(t-r)} (V_{\tau}^{\pi_r}(\rho) - V_{\tau}^{\pi^*}(\rho)) dr + \frac{1}{2\tau} \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr.
 \end{aligned}$$

1215 Now using that $\int_0^t e^{-\frac{\tau}{2}(t-r)} dr = \frac{2(1-e^{-\frac{\tau}{2}})}{\tau}$, we have
 1216

$$\begin{aligned}
 1217 \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_t(\cdot|s))d_{\rho}^{\pi^*}(ds) &\leq e^{-\frac{\tau}{2}t} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_0(\cdot|s))d_{\rho}^{\pi^*}(ds) \\
 1218 &\quad - \frac{2(1-\gamma)(1-e^{-\frac{\tau}{2}})}{\tau} \min_{r \in [0,t]} (V_{\tau}^{\pi_r}(\rho) - V_{\tau}^{\pi^*}(\rho)) + \frac{1}{2\tau} \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr.
 \end{aligned}$$

1219 Rearranging, we have
 1220

$$\begin{aligned}
 1221 \min_{r \in [0,t]} V_{\tau}^{\pi_r}(\rho) - V_{\tau}^{\pi^*}(\rho) &\leq \frac{\tau}{2(1-\gamma)(1-e^{-\frac{\tau}{2}})} \left(e^{-\frac{\tau}{2}t} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_0(\cdot|s))d_{\rho}^{\pi^*}(ds) \right. \\
 1222 &\quad \left. + \frac{1}{2\tau} \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \right).
 \end{aligned}$$

1223 which concludes the proof.
 1224

□

1242 D.3 PROOF OF THEOREM 6.2
1243

1244 *Proof.* Using the chain rule and the critic dynamics in (7), we have that for all $r \geq 0$
1245

$$\begin{aligned} \frac{1}{2\eta_r} \frac{d}{dr} |\theta_r - \theta_{\pi_r}|^2 &= \frac{1}{\eta_r} \left(\left\langle \frac{d\theta_r}{dr}, \theta_r - \theta_{\pi_r} \right\rangle - \left\langle \frac{d\theta_{\pi_r}}{dr}, \theta_r - \theta_{\pi_r} \right\rangle \right) \\ &= -\langle g(\theta_r, \pi_r), \theta_r - \theta_{\pi_r} \rangle - \frac{1}{\eta_r} \left\langle \frac{d\theta_{\pi_r}}{dr}, \theta_r - \theta_{\pi_r} \right\rangle \end{aligned}$$

1250 Let $\Gamma = \lambda_\beta(1 - \gamma)(1 - \sqrt{\gamma})$. Using Lemma 4.1 and the λ_β -strong convexity of $L(\cdot, \pi; \beta)$ and
1251 recalling that $L(\theta_{\pi_r}, \pi_r) = 0$ for all $r \geq 0$, it holds for all $r \geq 0$ that
1252

$$\begin{aligned} \frac{1}{2\eta_t} \frac{d}{dt} |\theta_t - \theta_{\pi_t}|^2 &= -\langle g(\theta_t, \pi_t), \theta_t - \theta_{\pi_t} \rangle - \frac{1}{\eta_t} \left\langle \frac{d\theta_{\pi_t}}{dt}, \theta_t - \theta_{\pi_t} \right\rangle \\ &\leq -(1 - \gamma)(1 - \sqrt{\gamma}) \langle \nabla_\theta L(\theta_t, \pi_t; \beta), \theta_t - \theta_{\pi_t} \rangle - \frac{1}{\eta_t} \left\langle \frac{d\theta_{\pi_t}}{dt}, \theta_t - \theta_{\pi_t} \right\rangle \\ &\leq -(1 - \gamma)(1 - \sqrt{\gamma}) L(\theta_t, \pi_t; \beta) - \frac{\Gamma}{2} |\theta_t - \theta_{\pi_t}|^2 - \frac{1}{\eta_t} \left\langle \frac{d\theta_{\pi_t}}{dt}, \theta_t - \theta_{\pi_t} \right\rangle \\ &\leq -(1 - \gamma)(1 - \sqrt{\gamma}) L(\theta_t, \pi_t; \beta) - \frac{\Gamma}{2} |\theta_t - \theta_{\pi_t}|^2 + \frac{1}{2\eta_t} \left(\left| \frac{d\theta_{\pi_t}}{dt} \right|^2 + |\theta_t - \theta_{\pi_t}|^2 \right) \\ &= -(1 - \gamma)(1 - \sqrt{\gamma}) L(\theta_t, \pi_t; \beta) - \left(\frac{\Gamma}{2} - \frac{1}{2\eta_t} \right) |\theta_t - \theta_{\pi_t}|^2 + \frac{1}{2\eta_t} \left| \frac{d\theta_{\pi_t}}{dt} \right|^2, \end{aligned} \tag{26}$$

1266 where we used Hölder's and Young's inequalities in (26). Since $\eta_0 > \frac{1}{\Gamma}$ and η_t is a non-decreasing
1267 function, it holds that $\eta_t > \frac{1}{\Gamma}$ for all $t \geq 0$. Hence $\frac{\Gamma}{2} - \frac{1}{2\eta_t} > 0$ and thus we can drop the second term.
1268 Moreover the λ_β -strong convexity of $L(\cdot, \pi; \beta)$ along with $L(\theta_\pi, \pi; \beta) = 0$ and $\nabla_\theta L(\theta_\pi, \pi) = 0$
1269 for all $\pi \in \Pi_\mu$ gives that
1270

$$|\theta_t - \theta_{\pi_t}|^2 \leq \frac{2}{\lambda_\beta} L(\theta_t, \pi_t; \beta).$$

1273 Hence for all $r \geq 0$ we arrive at

$$\frac{1}{2\eta_r} \frac{d}{dr} |\theta_r - \theta_{\pi_r}|^2 \leq -\frac{\Gamma}{2} |\theta_r - \theta_{\pi_r}|^2 + \frac{1}{2\eta_r} \left| \frac{d\theta_{\pi_r}}{dr} \right|^2.$$

1277 Rearranging, multiplying by $e^{-\tau(t-r)}$ and integrating over r from 0 to t , it holds for all $t \geq 0$ that
1278

$$\int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq -\frac{1}{\Gamma} \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \frac{d}{dr} |\theta_r - \theta_{\pi_r}|^2 dr + \frac{1}{\Gamma} \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \left| \frac{d\theta_{\pi_r}}{dr} \right|^2 dr.$$

1282 Integrating the first term by parts (identically to (18) from the proof of Theorem 5.1), we have
1283

$$\begin{aligned} \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr &\leq \frac{1}{\Gamma} \left(-\frac{|\theta_t - \theta_{\pi_t}|^2}{\eta_t} + e^{-\frac{\tau}{2}t} \frac{|\theta_0 - \theta_{\pi_0}|^2}{\eta_0} \right. \\ &\quad \left. + \frac{\tau}{2} \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} |\theta_r - \theta_{\pi_r}|^2 dr - \int_0^t |\theta_r - \theta_{\pi_r}|^2 \frac{e^{-\frac{\tau}{2}(t-r)} \frac{d}{dr} \eta_r}{\eta_r^2} dr \right. \\ &\quad \left. + \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \left| \frac{d\theta_{\pi_r}}{dr} \right|^2 dr \right). \end{aligned}$$

1292 Since for all $t \geq 0$ it holds that $\eta_t \geq 1$ and $\frac{d}{dt} \eta_t \geq 0$, we have that
1293

$$\int_0^t |\theta_r - \theta_{\pi_r}|^2 \frac{e^{-\frac{\tau}{2}(t-r)} \frac{d}{dr} \eta_r}{\eta_r^2} dr \geq 0.$$

1296 Thus after dropping all negative terms and using that $\eta_t \geq \eta_0$ for all $t \geq 0$, we have
 1297

$$1298 \left(1 - \frac{\tau}{2\Gamma\eta_0}\right) \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq e^{-\frac{\tau}{2}} \frac{|\theta_0 - \theta_{\pi_0}|^2}{\Gamma\eta_0} + \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \left| \frac{d\theta_{\pi_r}}{dr} \right|^2 dr.$$

1300 Since $\eta_0 > \frac{1}{2\Gamma}$ and $\tau < 1$, it holds that $1 - \frac{\tau}{2\Gamma\eta_0} > 0$ and hence it holds that
 1301

$$1302 \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq e^{-\frac{\tau}{2}} \frac{|\theta_0 - \theta_{\pi_0}|^2}{\Gamma\eta_0 \left(1 - \frac{\tau}{2\Gamma\eta_0}\right)} + \frac{1}{\left(1 - \frac{\tau}{2\Gamma\eta_0}\right)} \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \left| \frac{d\theta_{\pi_r}}{dr} \right|^2 dr,$$

1305 which concludes the proof. \square
 1306

1307 D.4 PROOF OF THEOREM 6.3

1309 *Proof.* By Theorem 6.2, we have

$$1310 \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq e^{-\frac{\tau}{2}} \frac{|\theta_0 - \theta_{\pi_0}|^2}{\Gamma\eta_0 \left(1 - \frac{\tau}{2\Gamma\eta_0}\right)} + \frac{1}{\left(1 - \frac{\tau}{2\Gamma\eta_0}\right)} \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} \left| \frac{d\theta_{\pi_r}}{dr} \right|^2 dr.$$

1313 Hence it remains to characterise the growth of the final integral. Observe that for all $\pi \in \mathcal{P}(A|S)$,
 1314 $\theta_\pi \in \mathbb{R}^N$ satisfies the least-squares optimality condition given by
 1315

$$1316 \theta_\pi = \arg \min_{\theta} L(\theta, \pi; \beta) = \left(\int_{S \times A} \phi(s, a) \phi(s, a)^\top \beta(da, ds) \right)^{-1} \left(\int_{S \times A} \phi(s, a) Q_\tau^\pi(s, a) \beta(ds, da) \right).$$

1318 Setting $\pi = \pi_t$ and differentiating time we arrive at
 1319

$$1320 \frac{d\theta_{\pi_t}}{dt} = \left(\int_{S \times A} \phi(s, a) \phi(s, a)^\top \beta(da, ds) \right)^{-1} \left(\int_{S \times A} \phi(s, a) \frac{d}{dt} Q^{\pi_t}(s, a) \beta(ds, da) \right).$$

1322 Hence by Lemma 6.1, Assumption 4.2 and Assumption 4.3, for all $t \geq 0$ it holds that
 1323

$$1324 \left| \frac{d\theta_{\pi_t}}{dt} \right| = \left| \left(\int_{S \times A} \phi(s, a) \phi(s, a)^\top \beta(da, ds) \right)^{-1} \left(\int_{S \times A} \phi(s, a) \frac{d}{dt} Q^{\pi_t}(s, a) \beta(ds, da) \right) \right|$$

$$1325 \leq \left| \left(\int_{S \times A} \phi(s, a) \phi(s, a)^\top \beta(da, ds) \right)^{-1} \right|_{\text{op}} \left| \frac{d}{dt} Q^{\pi_t} \right|_{B_b(S \times A)}$$

$$1326 = \frac{1}{\lambda_\beta} \left| \frac{d}{dt} Q^{\pi_t} \right|_{B_b(S \times A)}$$

$$1327 = \frac{\gamma}{\lambda_\beta(1-\gamma)} \left| \int_S \left(\int_{S \times A} A_\tau^{\pi_t}(s'', a'') \partial_t \pi_t(da''|s'') d^{\pi_t}(ds''|s') \right) P(ds'|\cdot, \cdot) \right|_{B_b(S \times A)}$$

$$1328 \leq \frac{\gamma}{\lambda_\beta(1-\gamma)} |A_\tau^{\pi_t}|_{B_b(S \times A)} \sup_{s \in S} |\partial_t \pi_t(\cdot|s)|_{\mathcal{M}(A)}.$$

1336 Now using Lemma B.2, it holds that
 1337

$$1338 |A_\tau^{\pi_t}|_{B_b(S \times A)} \sup_{s \in S} |\partial_t \pi_t(\cdot|s)|_{\mathcal{M}(A)} \leq |A_\tau^{\pi_t}|_{B_b(S \times A)} |A_t|_{B_b(S \times A)}$$

$$1339 \leq \left(2 |Q_\tau^{\pi_t}|_{B_b(S \times A)} + 2\tau \left| \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} \right) \left(2 |Q_t|_{B_b(S \times A)} + 2\tau \left| \ln \frac{d\pi_t}{d\mu} \right|_{B_b(S \times A)} \right).$$

1343 Hence by Corollaries 5.1 and 5.2 and Lemma B.2, there exists $\alpha_1, \alpha_2 > 0$ such that
 1344

$$1345 \left| \frac{d\theta_{\pi_t}}{dt} \right|^2 \leq \alpha_1 e^{\alpha_2 t}.$$

1346 Thus Theorem 6.2 becomes
 1347

$$1348 \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq e^{-\frac{\tau}{2}} \frac{|\theta_0 - \theta_{\pi_0}|^2}{\Gamma\eta_0 \left(1 - \frac{\tau}{2\Gamma\eta_0}\right)} + \frac{\alpha_1}{\left(1 - \frac{\tau}{2\Gamma\eta_0}\right)} \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{e^{\alpha_2 r}}{\eta_r} dr.$$

1350 Let $\eta_t = \eta_0 e^{k_1 t}$ for any $k_1 > \frac{\tau}{2} + \alpha_2$. Then observe that

$$\begin{aligned} 1352 \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{e^{\alpha_2 r}}{\eta_r} dr &= \frac{1}{\eta_0} e^{-\frac{\tau}{2}t} \int_0^t e^{(\frac{\tau}{2} + \alpha_2 - k_1)r} dr \\ 1354 &\leq \frac{1}{\eta_0} e^{-\frac{\tau}{2}t} \left(\frac{e^{(\frac{\tau}{2} + \alpha_2 - k_1)t} - 1}{\frac{\tau}{2} + \alpha_2 - k_1} \right) \\ 1356 &\leq \frac{e^{-\frac{\tau}{2}t}}{\eta_0 (\frac{\tau}{2} + \alpha_2 - k_1)}, \end{aligned}$$

1359 hence all together it holds that

$$1361 \int_0^t e^{-\frac{\tau}{2}(t-r)} |\theta_r - \theta_{\pi_r}|^2 dr \leq e^{-\frac{\tau}{2}} \frac{|\theta_0 - \theta_{\pi_0}|^2}{\Gamma \eta_0 \left(1 - \frac{\tau}{2\Gamma \eta_0}\right)} + e^{-\frac{\tau}{2}t} \frac{\alpha_1}{(\eta_0 - \frac{\tau}{2\Gamma}) (\frac{\tau}{2} + \alpha_2 - k_1)}.$$

1363 Substituting this into the result from Theorem 6.2 concludes the proof. \square

E ADDITIONAL RESULTS

1367 **Corollary E.1** (Uniform boundedness). *Under the same assumptions as Theorem 5.1, for $\gamma \in (0, 1)$ such that $\frac{64\gamma^2}{\Gamma^2 - \frac{\Gamma\tau}{\eta_0}} < 1$ it holds that $a_2 < \tau$ and for all $t \geq 0$ it holds that*

$$1371 \text{KL}(\pi_t(\cdot|s)|\mu)^2 \leq \frac{a_1 \tau}{\tau - a_2}$$

E.1 PROOF OF COROLLARY E.1

1375 *Proof.* By Theorem 5.1 we have that

$$1377 K_t^2 \leq a_1 + a_2 \int_0^t e^{-\tau(t-r)} K_r^2 dr.$$

1379 Taking the supremum over $[0, t]$ on the right hand side, we have

$$1380 K_t^2 \leq a_1 + \frac{a_2}{\tau} \sup_{r \in [0, t]} K_r^2.$$

1382 Since this holds for all $t \geq 0$, we have

$$1383 \sup_{r \in [0, t]} K_r^2 \leq a_1 + \frac{a_2}{\tau} \sup_{r \in [0, t]} K_r^2.$$

1385 Now forcing $1 - \frac{a_2}{\tau} > 0$, which is equivalent to the condition

$$1387 \frac{64\gamma^2}{\Gamma^2 - \frac{\Gamma\tau}{\eta_0}} < 1.$$

1389 Hence after rearranging we have

$$1390 K_t^2 \leq \sup_{r \in [0, t]} K_r^2 \leq \frac{a_1 \tau}{\tau - a_2}$$

\square

1393 **Remark E.1.** *Observe that if one does not apply the loose upper bound $e^{-\tau t} \leq 1$ in (20) from the proof of Theorem 5.1, it holds that*

$$1396 a_1 = a_1(t) = 8e^{-2\tau t} (C_1)^2 + \frac{32}{\tau} \sigma_1$$

1398 with $\sigma_1 := \frac{|\theta_0|^2}{\Gamma \eta_0 \left(1 - \frac{\tau}{2\Gamma \eta_0}\right)} + \frac{2|c|_{B_b(S \times A)}^2}{\Gamma^2 \tau \left(1 - \frac{\tau}{2\Gamma \eta_0}\right)}$. Then choosing $\eta_0 = \tau + \epsilon$ for any $\epsilon > 0$ so that the 1400 conditions of Theorem 5.1 holds, formally sending $\tau \rightarrow \infty$ we obtain $\text{KL}(\pi_t(\cdot|s)|\mu) \rightarrow 0$ for all 1401 $s \in S$.

1402 **Corollary E.2.** *Under the conditions of Corollary E.1, there exists $R > 0$ such that for all $t \geq 0$ it 1403 holds that*

$$|\theta_t| \leq R$$

1404 E.2 PROOF OF COROLLARY E.2
14051406 *Proof.* By Corollary E.1, for sufficiently small $\gamma > 0$ it holds that for all $t \geq 0$,

1407
$$K_t^2 \leq \frac{a_1 \tau}{\tau - a_2}.$$

1408

1409 Hence by Lemma 5.1 we have
1410

1411
$$\frac{1}{2} \frac{d}{dt} |\theta_t|^2 \leq -\eta_t \frac{\Gamma}{2} |\theta_t|^2 + \eta_t \left(\frac{2|c|_{B_b(S \times A)}^2 + 2\tau^2 \gamma^2 \left(\frac{a_1 \tau}{\tau - a_2} \right)}{\Gamma^2} \right).$$

1412

1413 The uniform boundedness in time of $|\theta_t|$ then follows by Grönwall's Lemma (Lemma A.3). \square
14141415 **Corollary E.3.** *Under the same assumptions as Theorem 6.2, for $\gamma \in (0, 1)$ such that $\frac{2\sqrt{2}\gamma}{\sqrt{\Gamma^2 - \frac{\Gamma\tau}{\eta_0}}} < 1$
1416 there exists $d_1 > 0$ such that for all $t \geq 0$,*
1417

1418
$$\min_{r \in [0, t]} V_r^{\pi_r}(\rho) - V_r^{\pi^*}(\rho) \leq \frac{\tau}{2(1-\gamma)(1-e^{-\frac{\tau}{2}t})} \left(e^{-\frac{\tau}{2}t} \int_S \text{KL}(\pi^*(\cdot|s) \|\pi_0(\cdot|s)) d_r^{\pi^*}(ds) \right. \\ 1419 \left. + d_1 \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} dr \right).$$

1420

1421 E.3 PROOF OF THEOREM 6.3
14221423 *Proof.* Following completely identically to the proof of Theorem 6.3, we have
1424

1425
$$\left| \frac{d\theta_{\pi_t}}{dt} \right| \leq \frac{\gamma}{\lambda_\beta(1-\gamma)} |A_r^{\pi_t}|_{B_b(S \times A)} \sup_{s \in S} |\partial_t \pi_t(\cdot|s)|_{\mathcal{M}(A)} \\ 1426 \leq \frac{4}{(1-\gamma)^2} \left(|c|_{B_b(S \times A)} + K_t \right)^2 + 4\tau \left(C_1 + \frac{2}{\tau} \sup_{r \in [0, t]} |\theta_r| + \sup_{r \in [0, t]} K_r \right)^2.$$

1427

1428 Then by Corollaries E.1 and E.2, there exists $b_2 > 0$ such that $\left| \frac{d\theta_{\pi_t}}{dt} \right|^2 \leq d_1$. Hence by Theorem 6.2
1429 we have
1430

1431
$$\min_{r \in [0, t]} V_r^{\pi_r}(\rho) - V_r^{\pi^*}(\rho) \leq \frac{\tau}{2(1-\gamma)(1-e^{-\frac{\tau}{2}})} \left(e^{-\frac{\tau}{2}t} \left(\int_S \text{KL}(\pi^*(\cdot|s) \|\pi_0(\cdot|s)) d_r^{\pi^*}(ds) \right. \right. \\ 1432 \left. \left. + d_1 \int_0^t e^{-\frac{\tau}{2}(t-r)} \frac{1}{\eta_r} dr \right) \right).$$

1433

1434 \square
14351436 For example, suppose the small discounting factor condition is satisfied, choosing $\eta_t = t^{\frac{1}{2}} + \eta_0$ with
1437 $\eta_0 > \frac{1}{\Gamma}$ and $\tau = 0.5$, it can be shown that asymptotically
1438

1439
$$\min_{r \in [0, t]} V_r^{\pi_r}(\rho) - V_r^{\pi^*}(\rho) \sim \frac{1}{\sqrt{t}}.$$

1440

1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457