Journal of Machine Learning Research 27 (2026) 1-68 Submitted 5/25; Revised 4/26; Published 4/26

A Two-Timescale Primal-Dual Framework for Reinforcement Learning via Online Dual Variable Guidance

Axel Friedrich Wolter AXEL-FRIEDRICH.WOLTER@UNI-KONSTANZ.DE
Department of Computer Science
University of Konstanz, Germany
Universitätsstrasse 10, 78464 Konstanz

Tobias Sutter TOBIAS.SUTTER@UNISG.CH
Department of Economics
University of St.Gallen, Switzerland
Rosenbergstrasse 20-22, 9000 St. Gallen

Editor: Nan Jiang

Abstract

We study reinforcement learning by combining recent advances in regularized linear programming formulations with the classical theory of stochastic approximation. Motivated by the challenge of designing algorithms that leverage off-policy data while maintaining on-policy exploration, we propose PGDA-RL, a novel primal-dual projected gradient descent-ascent algorithm for solving regularized Markov decision processes (MDPs). PGDA-RL integrates experience replay-based gradient estimation with a two-timescale decomposition of the underlying nested optimization problem. The algorithm operates asynchronously, interacts with the environment through a single trajectory of correlated data, and updates its policy online in response to the dual variable associated with the occupancy measure of the underlying MDP. We prove that PGDA-RL converges almost surely to the optimal value function and policy of the regularized MDP. Our convergence analysis relies on tools from stochastic approximation theory and holds under weaker assumptions than those required by existing primal-dual RL approaches, notably removing the need for a simulator or a fixed behavioral policy. Under a strengthened ergodicity assumption on the underlying Markov chain, we establish a last-iterate finite-time guarantee with $\widetilde{\mathcal{O}}(k^{-2/3})$ mean-square convergence, aligning with the best-known rates for two-timescale stochastic approximation methods under Markovian sampling and biased gradient estimates.

Keywords: reinforcement learning, regularized Markov decision processes, stochastic approximation, primal-dual method, two-timescale optimization

1. Introduction

Recent years have seen many new promising developments in the theory and application of reinforcement learning (RL) as a framework for sequential decision making with notable successes of regularization-based algorithms, for example, the Soft Actor-Critic (Haarnoja et al., 2018) or Trust Region Policy Optimization (Schulman et al., 2015). The regularization typically takes the form of a convex penalty term in the policy, which is a setting that has been subsequently formalized by Geist et al. (2019); Neu et al. (2017). The adjustment of the objective aims at inducing desirable exploratory behavior of the learned policies and stability in their training (Schulman et al., 2017), and allows the application of dynamical systems analysis that takes advantage of the adjusted problem's strongly convex structure (Li et al., 2024).

As another development, the linear programming (LP) reformulation of Markov decision processes (MDPs), originally dating back to Manne (1960); Hernández-Lerma and Lasserre (1996); Borkar (2002), has recently received renewed interest. It has led to LP-based algorithms shown to be on par with broadly applied benchmarks, for example, the Dual Actor-Critic (Dai et al., 2018a). The LP approach to RL allows the reformulation of the Bellman optimality equation, which characterizes an optimal policy, as a min-max saddle-point problem that can be addressed with gradient methods. The analysis of the various algorithms proposed in this line of research often relies on simplifying assumptions such as access to a sampling generator (Chen and Wang, 2016), a known model (Li et al., 2024), or the uniqueness of the optimal policy (Lee and He, 2019). A further challenge in solving the min-max RL problem formulation lies in the nested loop structure that arises when the inner optimization is first solved approximately during each outer iteration, which complicates the convergence analysis (Gabbianelli et al., 2024; Dai et al., 2018a).

The asymptotic convergence under general model-access assumptions is well understood for (approximate) dynamic programming algorithms; (see Tsitsiklis (1994); Jaakkola et al. (1993)). However, finite-time guarantees for the *last iterate* remain scarce for LP-based primal-dual schemes driven by Markovian single-trajectory data and iterate-dependent (on-policy) exploration, particularly when gradients are formed from replay buffers and are therefore biased at finite times. This perspective is complementary to regret-minimization in online RL: rather than optimizing regret or minimax sample complexity via exploration bonuses, we study the stability and last-iterate behavior of a primalreplay-buffer biasd-ual stochastic approximation method for an LP saddle-point formulation in a realistically correlated-data regime.

In this work, we combine recent advances on regularized LP formulations of RL (Li et al., 2024) with the classical theory of stochastic approximation (Borkar, 2023) and modern finite-time analyses for two-timescale stochastic approximation (Zeng et al., 2024). Our goal is to design a provably convergent algorithm that maintains on-policy exploration while performing off-policy updates using previously observed transitions stored in an experience replay buffer, and to quantify its last-iterate finite-time rate. The main contributions of this paper are summarized as follows:

1. We introduce a novel primal-dual Projected Gradient Descent-Ascent (PGDA-RL) algorithm for regularized MDPs. The algorithm combines experience replay-based gradient estimation with a two-timescale decomposition of the nested optimization problem and leverages on-policy exploration. PGDA-RL operates asynchronously and interacts with the environment through a single trajectory of data, generated under a policy that evolves in response to the dual variable associated with the occupancy measure.

2. We establish that PGDA-RL converges almost surely to the optimal value function and policy of the regularized MDP. Our convergence analysis is based on tools from the stochastic approximation literature and requires significantly weaker assumptions than

existing primal-dual RL methods. Notably, we do not assume access to a simulator or a fixed behavior policy.

3. Under a stronger ergodicity assumption on the underlying Markov chain, we establish a *last-iterate* finite-time guarantee for the asynchronous PGDA-RL scheme despite replay-buffer biased gradient estimates. Concretely, on a high-probability visitation event $\mathcal{G}_\delta$ ensuring linear growth of the least-visited state-action count after a burn-in time, the dual iterate satisfies $\mathbb{E}[\|\rho_k - \rho^\star\|_2^2 \mid \mathcal{G}_\delta] = \widetilde{\mathcal{O}}(k^{-2/3})$. This yields corresponding $\widetilde{\mathcal{O}}(k^{-2/3})$ mean-square convergence guarantees for the regularized optimal value and policy.

1.1 Related Work

Recent research has explored several directions in reinforcement learning that build upon primal-dual methods and the regularized Markov decision processes framework. Table 1 provides a concise overview of the most closely related LP-based primal-dual RL methods and highlights the key differences compared to our approach regarding underlying model access, function approximation setting, and convergence analysis.

Dai et al. (2018a) introduced a dual actor-critic algorithm rooted in the LP formulation that incorporates multi-step Bellman equations, path regularization, and a nested stochastic dual-ascent update. While their approach shares similarities with ours in starting from an LP/Lagrangian viewpoint and employing convexity-inducing regularizers, we adopt a two-timescale stochastic approximation framework that yields an *incremental* update rule and enables an almost sure convergence analysis under Markovian sampling. Similarly, works by Chen and Wang (2016); Wang (2017); Chen et al. (2018) study LP formulations for discounted, average reward, and finite-horizon MDPs through projected stochastic primal-dual methods (including incremental coordinate updates). These works provide finite-sample or asymptotic convergence results under the crucial assumption of access to a sampling generator. In contrast, we focus on the *last iterate* and establish *almost sure convergence* assuming only access to a single trajectory of state-action pairs (Markovian model access). Lee and He (2019) extends Chen and Wang (2016) by incorporating a Q-function estimation step. Their analysis allows time-varying behavioral policies under conditions requiring sublinear convergence of the induced state-action distribution, which can be difficult to verify a priori and makes fully on-policy analysis nontrivial. Moreover, their algorithm relies on samples drawn from the stationary state-action distribution of the current policy. In Bas-Serrano et al. (2021) and the subsequent thesis Bas-Serrano (2022), the authors introduce the logistic Bellman error derived from an entropy-regularized LP formulation and establish finite-sample rates in the tabular setting under stationary-distribution generator access. In comparison, our stochastic approximation analysis further allows us to avoid stationary distribution sampling and to analyze iterate-dependent (on-policy) exploration under Markovian single-trajectory access. Under a strengthened ergodicity condition on the Markov chain induced by the iterate-dependent behavioral policy, we additionally establish a *last-iterate* finite-time mean-square convergence rate.

Two-timescale stochastic approximation has been applied to dynamic programming-based actor-critic algorithms (Konda and Borkar, 1999) and to constrained MDPs (Borkar, 2005), where safety constraints are added to the reward maximization objective. The La-

Paper	Model Access	Func. Appr.	Analysis	Guarantee	Rate
Chen and Wang (2016)	Generator	Tabular	PAC	Duality gap	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
Dai et al. (2018a)	Markovian	Non-linear	Sketch	–	–
Lee and He (2019)	Stationary Distr. Generator	Tabular	PAC	Duality gap	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
Bas-Serrano (2022)	Stationary Distr. Generator	Tabular	PAC	ϵ-optimal policy	$\widetilde{\mathcal{O}}(\epsilon^{-9})$
Gabbianelli et al. (2024)	Offline	Linear	PAC	ϵ-optimal policy	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
Li et al. (2024)	Full model	Tabular	Deterministic asymptotic	–	–
Our paper	Markovian	Tabular	Almost sure, Convergence rate	Dual iterate MSE after burn-in	$\widetilde{\mathcal{O}}(\epsilon^{-3/2})$

Table 2: Overview of LP-based RL methods and their guarantees. We summarize related approaches based on linear-programming or Lagrangian formulations of MDPs, organized by the model access assumption, the function-approximation regime, the type of analysis, and the corresponding guarantee and reported rate. The "Rate" column is stated with respect to the corresponding guarantee, and is therefore not directly comparable across rows when the guarantees or access models differ.

grangian relaxation of the constrained optimization problem motivates the application of various primal-dual methods (Chen et al., 2024; Hong and Tewari, 2024; Li et al., 2023).

The LP approach has also been applied to Offline RL problems, where near-optimal policies are learned from static data sets of transitions (Zhan et al., 2022; Ozdaglar et al., 2023; Gabbianelli et al., 2024). Even though the analysis objective differs, since we consider the online setting, their saddle-point problem resulting from the Lagrangian dual formulation of the LP is closely related to our optimization target.

Parallel investigations into regularized MDPs have also been influential. Li et al. (2024) explore a primal-dual formulation of entropy-regularized MDPs with additional convexity in the primal variable, providing asymptotic convergence analysis using natural gradient methods, albeit under the assumption of full model knowledge. Neu et al. (2017) and Geist et al. (2019) formalize the entropy-regularized MDP framework, outlining the conditional entropy regularizer and the contraction properties of the regularized Bellman evaluation operator. Ying and Zhu (2020) offers a comprehensive overview of various LP formulations in both regularized and unregularized contexts, emphasizing their equivalence to Bellman equations and policy gradient methods. Furthermore, convex Q-learning (Meyn, 2022; Mehta and Meyn, 2020; Lu and Meyn, 2023) is motivated by the idea of deriving a convex relaxation of the squared Bellman error based on the LP approach. The method involves solving a constrained convex optimization problem over the primal variable, the Q-function. It is primarily designed for analyzing the non-tabular setting, where the value function is learned via function approximation, which distinguishes this approach from our own.

Finally, our finite-time analysis builds on recent last-iterate convergence-rate results for two-timescale stochastic approximation Zeng et al. (2024). We face additional challenges in our setting, namely asynchronous updates and replay-buffer gradient estimates that introduce bias. Despite this, on the high-probability good visitation event $\mathcal{G}_\delta$ (after a burn-in period), we prove a last-iterate mean-square rate $\widetilde{\mathcal{O}}(k^{-2/3})$ for the dual iterates, imply-ing corresponding mean-square rates for the induced regularized value function and policy. Equivalently, achieving ϵ accuracy in mean-square error requires $k(\epsilon) = \widetilde{\mathcal{O}}(\epsilon^{-3/2})$ iterations to ensure $\mathbb{E}[\|\rho_k - \rho^\star\|_2^2 \mid \mathcal{G}_\delta] \leq \epsilon$ (up to logarithmic factors).

Structure. The structure of the paper is as follows. Section 2 formally introduces regularized MDPs and the associated saddle-point formulation that underlies our proposed algorithms. Section 3 presents a first algorithm in a synchronous setting, assuming access to a generative model. In Section 4, we extend this setting to the asynchronous case, where learning occurs from a single trajectory of data generated under a potentially changing policy. The section further contains the almost sure convergence result and our convergence rate analysis. Section 5 demonstrates the effectiveness of the proposed asynchronous algorithm on a standard reinforcement learning benchmark. Finally, Section 6 concludes the paper. The proofs details and technical results are deferred to the appendix to improve the readability of the main text.

Notation. For any state $s \in \mathcal{S}$, we denote by δ_s the Dirac measure in s. For a variable $v(s,a)$ depending on state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$, we denote the state marginal by $\tilde{v}(s) := \sum_{a \in \mathcal{A}} v(s,a)$. For a finite set $\mathcal{S}$ with cardinality n, the probability simplex over $\mathcal{S}$ is denoted by $\Delta_{\mathcal{S}} = \{x \in \mathbb{R}^n_+ : \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} x_i = 1\}$. We use the symbol $e_i \in \mathbb{R}^{|\mathcal{S}|}$ for the i-th standard basis vector on $\mathbb{R}^{|\mathcal{S}|}$. For any logical expression $\mathcal{E}$, the indicator function $\mathbf{1}\{\mathcal{E}\}$ evaluates to 1 if $\mathcal{E}$ is true and to 0 otherwise. To denote sequences, we use $\{\beta_k\}$ and $\{\beta(k)\}$ interchangeably for $k \in \mathbb{N}$. For $a, b \in \mathbb{R}$ we denote with $a \vee b$ the larger of a and b and with $\lfloor a \rfloor$ the integer part of a. For a compact set $H \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we denote the Euclidean projection operator as Π_H, where $\Pi_H : \mathbb{R}^d \to H$ is defined by $\Pi_H(x) = \arg\min_{y \in H} \|x - y\|_2$. We denote the total variation norm as $\|\cdot\|_{\mathrm{TV}}$. For kernels $\mathcal{P}$, $\mathcal{Q}$ on $\mathcal{X}$, we define $\|\mathcal{P} - \mathcal{Q}\|_\infty := \sup_{x \in \mathcal{X}} \|\mathcal{P}(\cdot|x) - \mathcal{Q}(\cdot|x)\|_{\mathrm{TV}}$.

2. Problem Formulation

We consider a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote finite state and action spaces, respectively. The transition kernel $\mathcal{P}(\cdot|s,a) \in \Delta_{\mathcal{S}}$ specifies the distribution over next states given the current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. Each state-action pair (s,a) is associated with a reward $r(s,a) \in \mathbb{R}$, and $\gamma \in (0,1)$ denotes the discount factor. Our objective is to characterize and subsequently compute optimal randomized stationary Markovian policies $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ that maximize the expected total discounted reward. We denote by $\Pi = (\Delta_{\mathcal{A}})^{|\mathcal{S}|}$ the set of all stationary Markovian policies. We assume a positive and bounded deterministic reward, that is, $r(s,a) \in [0, C_r]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and some positive constant $C_r < \infty$. The unregularized value function of a stationary Markovian policy $\pi \in \Pi$ is defined as

$$V_{ur}^\pi(s) = \mathbb{E}_s^\pi \left[\sum_{k=0}^\infty \gamma^k r(s_k, a_k) \right],$$

where the notation $\mathbb{E}_{\bar{s}}^\pi$ refers to the expectation under $s_0 = s, a_k \sim \pi(\cdot|s_k)$, $s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$. The optimal value of the unregularized MDP is $V_{ur}^\star := \max_{\pi \in \Pi} V_{ur}^\pi$, the maximum value of all stationary Markovian policies. The Bellman optimality equation states that the optimal value function satisfies

$$V_{ur}^{\star}(s) = \max_{\pi\in\Pi} \mathbb{E}_s^{\pi}\left[r(s,a) + \gamma \sum_{s'\in\mathcal{S}} V_{ur}^{\pi}\left(s'\right)\mathcal{P}(s'|s,a) \right], \qquad \forall s \in \mathcal{S},$$

which admits a unique solution. We approach the unregularized MDP by its formulation as a linear program, Puterman (1994, Section 6.9). Reformulating the Bellman optimality equations as inequalities yields that a bounded function $V : \mathcal{S} \to \mathbb{R}$ that satisfies

$$V(s) \geq r(s,a) + \gamma \sum_{s'\in\mathcal{S}} V\left(s'\right)\mathcal{P}(s'|s,a) \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},$$

is an upper bound to the value $V_{ur}^{\star}$ of the MDP.

Building on this finding, the primal formulation of the Bellman LP is then given by

$$P : \begin{cases} \min_{V\in\mathbb{R}^{|\mathcal{S}|}} & \mu^{\top} V \\ \text{s.t.} & 0 \geq -V(s) + r(s,a) + \gamma \sum_{s'\in\mathcal{S}} V\left(s'\right)\mathcal{P}(s'|s,a) \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}, \end{cases} \tag{1}$$

where $\mu \in \mathbb{R}_{+}^{|\mathcal{S}|}$ is any strictly positive vector such that $\sum_{s\in\mathcal{S}} \mu(s) = 1$. It is well-known (Puterman, 1994, Section 6.9) that the optimizer to the primal LP (1) is exactly the optimal value function $V_{ur}^{\star}$. The corresponding dual linear problem is defined as

$$D : \begin{cases} \max_{\rho\geq 0} & \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a) r(s,a) \\ \text{s.t.} & \sum_{a\in\mathcal{A}} \rho(s,a) - \gamma \sum_{s'\in\mathcal{S}, a'\in\mathcal{A}} \rho\left(s',a'\right)\mathcal{P}\left(s|s',a'\right) = \mu(s) \quad \forall s \in \mathcal{S}, \end{cases} \tag{2}$$

where μ is the vector specified in the primal Bellman LP (1). A dual feasible variable ρ corresponds to the *discounted occupancy measure* of a stationary randomized policy π_{ρ} defined for $s \in \mathcal{S}$ as

$$\pi_{\rho}(a|s) = \frac{\rho(s,a)}{\tilde{\rho}(s)}, \qquad \forall a \in \mathcal{A}, \tag{3}$$

where $\tilde{\rho}(s) := \sum_{a'\in\mathcal{A}} \rho(s,a')$ denotes the state marginal of the dual variable. Let $\rho_{ur}^{\star}$ denote a solution to the dual problem (2), then $\pi_{\rho_{ur}^{\star}}$ is an optimal policy, that is, $V_{ur}^{\pi_{\rho_{ur}^{\star}}} = V_{ur}^{\star}$, see (Puterman, 1994, Theorem 6.9.4). The Lagrangian formulation of the primal LP (1) is

$$\max_{\rho\geq 0} \min_{V\in\mathbb{R}^{|\mathcal{S}|}} L_{ur}(V,\rho) := \mu^{\top} V + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\Delta[V](s,a), \tag{4}$$

where

$$\Delta[V](s,a) = -V(s) + r(s,a) + \gamma \sum_{s'\in\mathcal{S}} V\left(s'\right)\mathcal{P}(s'|s,a)$$

denotes the Bellman error. It can be directly seen that the primal variable V can be assumed without loss of generality to belong to the compact set $\mathcal{V} := \{v \in \mathbb{R}_{+}^{|\mathcal{S}|} : v_i \leq \frac{C_r}{1-\gamma} \ \forall i = 1, \dots |\mathcal{S}|\}$.

2.1 Regularized MDP

Our approach is based on the regularized MDP formulation as formally described in Geist et al. (2019). Let $G : \mathbb{R}_+^{|\mathcal{A}|} \to \mathbb{R}$ be a strongly concave function and $\eta_\rho > 0$ a regularization parameter. For a stationary policy $\pi \in \Pi$, the regularized Bellman operator $T_{\pi,G} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}, V \mapsto T_{\pi,G}V$ is defined state-wise as

$$[T_{\pi,G}V](s) := \mathbb{E}_{\bar{s}}^{\pi}\left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} V\left(s'\right) \mathcal{P}(s'|s,a)\right] + \eta_\rho G(\pi(\cdot|s)). \tag{5}$$

The regularized Bellman optimality operator is defined accordingly as

$$T_{\star,G} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}, V \mapsto T_{\star,G}V := \max_{\pi \in \Pi} T_{\pi,G}V, \tag{6}$$

where the maximum is state-wise. Both operators are γ-contractions in the supremum norm, (Geist et al., 2019, Prop. 2). The regularized value function of a policy π, denoted V_r^π, is defined as the fixed point of the operator $T_{\pi,G}$. The regularized value function of the regularized MDP, denoted $V_r^\star$, is defined as the fixed point of the regularized optimality operator $T_{\star,G}$. Due to the strong concavity of G, the regularized Bellman optimality operator has a unique maximizing argument.

Proposition 1 *(Geist et al., 2019, Theorem 1) The policy* $\pi_r^\star := \arg\max_{\pi \in \Pi} T_{\pi,G}V_r^\star$ *is the unique optimal regularized policy in the sense that, state-wise,* $V_r^{\pi_r^\star} = V_r^\star \geq V_r^\pi$ *for all policies* $\pi \in \Pi$.

The optimal regularized policy $\pi_r^\star$ is a strictly positive and exploratory softmax policy for entropy-type G. The Lagrangian formulation of the regularized MDP (Ying and Zhu, 2020) includes the strongly concave function of the dual variable $g(\rho) := \sum_{s \in \mathcal{S}} G(\rho_s)$ with $\rho_s = \rho(s,\cdot) \in \mathbb{R}_+^{|\mathcal{A}|}$ and is given as

$$\max_{\rho \geq 0} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} L_r(V,\rho) := \mu^\top V + \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s,a)\Delta[V](s,a) + \eta_\rho g(\rho). \tag{7}$$

The following corollary is a consequence of Proposition 1 and characterizes the solution to the regularized Lagrangian.

Corollary 2 *Problem (7) has the unique saddle-point* $(V_r^\star, \rho_r^\star)$. *By the uniqueness of the optimal regularized policy, we have* $\pi_r^\star = \pi_{\rho_r^\star}$.

The proof is provided in Appendix D. The regularization affects the optimal value, that is, the regularized value function $V_r^\star$ differs from the unregularized value function $V_{ur}^\star$. However, the regularized policy $\pi_r^\star$ approximates the unregularized problem well. More precisely, the suboptimality of $\pi_r^\star$ for the unregularized MDP can be quantified and controlled.

Proposition 3 *(Geist et al., 2019, Theorem 2) Assume that constants* L_G *and* U_G *exist such that for all* $\pi \in \Pi$ *we have* $L_G \leq \max_{s \in \mathcal{S}} G(\pi(\cdot|s)) \leq U_G$. *Then,* $V_{ur}^\star(s) - \eta_\rho \frac{U_G - L_G}{1-\gamma} \leq V_{ur}^{\pi_r^\star}(s) \leq V_{ur}^\star(s)$ *for all* $s \in \mathcal{S}$.

For the Shannon entropy regularizer $G(\pi(\cdot|s)) := \sum_{a\in\mathcal{A}} \pi(a|s)\log(\pi(a|s))$, we have $0 \leq G(\pi(\cdot|s)) \leq \log|\mathcal{A}|$, for all $\pi \in \Pi$ and we can take $L_G = 0$, $U_G = \log|\mathcal{A}|$ in the above stated bounds. To ensure better computational properties, we seek to make the regularized problem (7), defined by the regularized Lagrangian L_r, strongly convex in V. Since L_r is originally linear in V, we propose an additional modification to problem (7). Specifically, we replace the linear term $\mu^\top V$ with a quadratic term and introduce a specific dual regularization function g of the form

$$g(\rho) = -\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\log\left(\frac{\rho(s,a)}{\tilde{\rho}(s)}\right). \tag{8}$$

With these modifications, we arrive at the main formulation of our study, a further regularized problem that takes the form

$$\max_{\rho\geq 0}\min_{V\in\mathbb{R}^{|\mathcal{S}|}} L(V,\rho) := \frac{\eta_V}{2}\|V\|_2^2 + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\left(\Delta[V](s,a) - \eta_\rho \log\left(\frac{\rho(s,a)}{\tilde{\rho}(s)}\right)\right). \tag{9}$$

The closely related recent work Li et al. (2024) also studies the convexified regularized MDP (9). They prove that the convex modification does not affect the optimal regularized value and only changes the optimal dual variable by a scaling factor that does not influence the optimal regularized policy.

Proposition 4 *(Li et al., 2024, Theorem 2.1) The unique solution $(V^\star, \rho^\star)$ to the double-regularized problem (9) is such that $V^\star = V_r^\star$ and the optimal policy induced by $\rho^\star$ coincides with $\pi_r^\star$, that is, $\pi_{\rho^\star} = \pi_{\rho_r^\star} = \pi_r^\star$.*

The optimal dual variable $\rho^\star$ of the double-regularized problem (9) depends on the convexity parameter η_V and does not form a state-action occupancy measure. The following technical lemma provides bounds for the optimal dual variable. Its proof is deferred to Appendix D.

Lemma 5 (Optimal dual variable) *The optimal dual variable of problem (9) with bounded reward is strictly positive and bounded from above, that is, there exist $C^L, C^U > 0$ such that $C^L < \rho^\star(s,a) < C^U$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.*

Let C^L and C^U be the explicit bounds derived in the constructive proof of Lemma 5. Define H as

$$H := \{\rho \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|} \mid C^L \leq \rho(s,a) \leq C^U,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}\}. \tag{10}$$

It holds that $\rho^\star \in \operatorname{int} H$. This compact set H will be used in the algorithms to specify the projection region of the dual variable iterates without impacting the optimal solution.

3. Synchronous Algorithm

We now propose our solution to the saddle-point problem (9) that is based on a projected stochastic gradient descent-ascent method summarized in Algorithm 1. The algorithm does not assume to know the transition probabilities of the MDP but has access to a generative model (Kakade, 2003; Chen and Wang, 2016), also known as a sampling oracle, which

Algorithm 1 Synchronous Projected Gradient Descent-Ascent RL

Require: H as in (10), $V_0 \in \mathbb{R}^{|\mathcal{S}|}, \rho_0 \in H, \eta_\rho, \eta_V > 0, \{\alpha_i\}_{i=1}^K, \{\beta_i\}_{i=1}^K \subset \mathbb{R}_{>0}$

for $k = 1, 2, \dots, K$ **do**
  Sample state-transitions $s' \sim \mathcal{P}(\cdot|s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$
  Update V with an SGD step $V_k = V_{k-1} - \alpha_k \widehat{\nabla}_V L(V_{k-1}, \rho_{k-1})$
  Update ρ with a projected SGA step $\rho_k = \Pi_H \left[\rho_{k-1} + \beta_k \widehat{\nabla}_\rho L(V_{k-1}, \rho_{k-1})\right]$
end for

return V_K, ρ_K

provides sampling access to any state in the environment. Given a state-action pair (s, a) as input, the generative model provides the reward $r(s, a)$ and a follow-up state s' with $s' \sim \mathcal{P}(\cdot|s, a)$. The almost sure convergence analysis is based on the two-timescale stochastic approximation framework developed in Borkar (1997), Borkar (2023). The dynamics of the projected recursion are based on results of Dupuis (1987) and Dupuis and Nagurney (1993).

Since the true gradients of the Lagrangian L depend on the unknown transition probabilities, we use stochastic gradient approximations, denoted by $\widehat{\nabla}_V L$ and $\widehat{\nabla}_\rho L$, to perform the updates. To ensure that the iterates remain bounded, we project the dual iterates onto the feasible set H using the bounds on the optimal dual variable established in Lemma 5. In the synchronous setting, we fully update the primal and dual variables in each iteration. We construct the stochastic gradients from generative model transition samples obtained for each state-action pair.

Specifically, let $\widetilde{\Xi} := \{\tilde{s}(s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ be a collection of random variables such that

$$\tilde{s}(s, a) \sim \mathcal{P}(\cdot|s, a), \quad \text{and } \{\tilde{s}(s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}} \text{ are independent.}$$

Recall the definition of the state-marginal $\tilde{\rho}(s) := \sum_{a\in\mathcal{A}} \rho(s, a)$. Given $\widetilde{\Xi}$, the stochastic gradients used in Algorithm 1 are

$$\widehat{\nabla}_V L(V, \rho)(s') = \eta_V V(s') - \tilde{\rho}(s') + \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s, a)\, \mathbf{1}\{\tilde{s}(s, a) = s'\}, \qquad s' \in \mathcal{S}, \tag{11a}$$

$$\widehat{\nabla}_\rho L(V, \rho)(s, a) = -V(s) + r(s, a) + \gamma V\big(\tilde{s}(s, a)\big) - \eta_\rho \log\Big(\frac{\rho(s, a)}{\tilde{\rho}(s)}\Big),\ (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{11b}$$

Algorithm 1 is a stochastic approximation algorithm based on the primal-dual gradient method for saddle-point estimation (Arrow et al., 1958). Chen and Wang (2016), Lee and He (2019), among others, study versions of a related approach for unregularized LP formulations, where the optimal policy's uniqueness must be assumed. They provide sample complexity bounds by applying a concentration bound to the average iterate. In contrast, we take the dynamical systems viewpoint (Borkar, 2023) on the stochastic approximation algorithm and analyze the asymptotic almost sure convergence. This change in perspective is reflected in the choice of stepsizes, where our approaches feature diminishing stepsizes operating on two timescales, such that the last iterates asymptotically approach the limiting ODEs. In contrast, they rely on a fixed stepsize scheme, which suffices for the average iterate analysis.

We first present this synchronous algorithm to introduce the convergence analysis and generalize the setting in Section 4 to allow asynchronous updates based on observations generated by a single trajectory of state-action pairs under a fixed or varying behavioral policy.

3.1 Almost Sure Convergence Analysis of Algorithm 1

Theorem 6 (Convergence of Algorithm 1) *Consider the stochastic gradients as defined in* (11) *and the projection region H defined in* (10)*. Let the stepsize sequences $\{\alpha_k\}, \{\beta_k\}$ satisfy*

$$\sum_{k=1}^{\infty} \alpha_k = \sum_{k=1}^{\infty} \beta_k = \infty, \quad \sum_{k=1}^{\infty} \left(\alpha_k^2 + \beta_k^2\right) < \infty, \quad \lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0.$$

Let $\{(V_k, \rho_k)\}$ be a sequence generated by Algorithm 1 with an arbitrary initialization $V_0 \in \mathbb{R}^{|\mathcal{S}|}$, $\rho_0 \in H$. Then the last iterates converge $\mathbb{P}$-a.s. to the regularized saddle-point of (9)*, that is,*

$$\lim_{K\to\infty} (V_K, \rho_K) = (V^\star, \rho^\star) \quad \mathbb{P}\text{-a.s.}$$

Examples of suitable stepsize sequences are $\alpha_k = \frac{1}{k^q}, q \in (\frac{1}{2}, 1)$ and $\beta_k = \frac{1}{k}$ or $\alpha_k = \frac{1}{k}$, $\beta_k = \frac{1}{1+k\log k}$. The proof of Theorem 6 builds on the two-timescale stochastic approximation framework of Borkar (1997). This framework takes the dynamical systems viewpoint on the stochastic iteration and applies to the regularized MDP problem addressed with Algorithm 1 as follows.

The recursion in Algorithm 1 corresponds to the following coupled iterations

$$V_k = V_{k-1} - \alpha_k(\nabla_V L(V_{k-1}, \rho_{k-1}) + M_k^{(1)}), \tag{12a}$$

$$\rho_k = \Pi_H \left[\rho_{k-1} + \beta_k(\nabla_\rho L(V_{k-1}, \rho_{k-1}) + M_k^{(2)})\right], \tag{12b}$$

where $\nabla_V L : \mathbb{R}^{|\mathcal{S}|} \times H \to \mathbb{R}^{|\mathcal{S}|}$, $\nabla_\rho L : \mathbb{R}^{|\mathcal{S}|} \times H \to \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ are the gradients of the regularized Lagrangian and $\{M_k^{(1)}\}$, $\{M_k^{(2)}\}$ are the sequences of differences between the true gradients and their noisy estimates given in (11), that is,

$$M_k^{(1)} = \widehat{\nabla}_V L(V_{k-1}, \rho_{k-1}) - \nabla_V L(V_{k-1}, \rho_{k-1}),\ M_k^{(2)} = \widehat{\nabla}_\rho L(V_{k-1}, \rho_{k-1}) - \nabla_\rho L(V_{k-1}, \rho_{k-1}).$$

The iterate sequences of Algorithm 1 fulfill the following properties regarding the gradients and noise.

Proposition 7 (Gradient and noise properties of Algorithm 1) *The following conditions are satisfied.*

(A1) *The regularized Lagrangian, defined in* (9)*, has Lipschitz continuous gradients.*

(A2) *The noise terms $\{M_k^{(i)}\}$ for $i = 1, 2$ arising from the iterates of Algorithm 1 are martingale difference sequences with respect to the increasing σ-fields*

$$\mathcal{F}_k := \sigma(V_\ell, \rho_\ell, M_\ell^{(1)}, M_\ell^{(2)}, \ell \le k), \quad k \ge 0,$$

and satisfy

$$\mathbb{E}\left[\|M_k^{(i)}\|^2|\mathcal{F}_{k-1}\right] \leq C(1 + \|V_{k-1}\|^2 + \|\rho_{k-1}\|^2), \quad i = 1, 2,$$

for a constant $C > 0$ for $k \geq 1$.

The proof is provided in Appendix A.

Together with the two-timescale stepsize sequences assumed in Theorem 6, these gradient and noise properties ensure that the coupled system (12) asymptotically approximates a dynamical system with a fast transient $V(t)$ and a slowly evolving dual variable $\rho(t)$. The separation of the timescales allows the fast transient to view the dual variable as quasi-constant, and in turn, the dual variable to assume the transient to be equilibrated. The convergence of the stochastic recursion is derived from the asymptotic properties of the limiting dynamical system. The ODE describing the dynamics of the primal variable V is the gradient flow

$$\dot{V}(t) = -\nabla_V L(V(t), \rho) \tag{13a}$$

for fixed $\rho \in H$. Since the mapping $V \mapsto L(V, \rho)$ is strictly convex, the set of equilibrated primal variables given a fixed $\rho \in H$ is a singleton. We denote this "optimal response" as

$$\lambda(\rho) := \arg\min_{V \in \mathbb{R}^{|\mathcal{S}|}} L(V, \rho) = \{V \in \mathbb{R}^{|\mathcal{S}|} : \nabla_V L(V, \rho) = 0\}.$$

The dual variable's dynamic is restricted to the compact set H. We follow the notation of (Kushner and Yin, 2003, Section 4.3) for the projected dynamics. For $\rho \in \partial H$, the boundary of H, let $C(\rho)$ be the infinite convex cone generated by the outer normals at ρ of the faces on which ρ lies. For ρ in the interior of H, let $C(\rho) = \{0\}$. The projected ODE describing the flow of the dual variable is then given by

$$\dot{\rho}(t) = \nabla_\rho L(\lambda(\rho(t)), \rho(t)) + \zeta(t), \quad \zeta(t) \in -C(\rho(t)), \tag{13b}$$

with the *projection term* ζ, the minimum force needed to keep ρ in H. To obtain almost sure convergence of the stochastic recursion, the limiting dynamical systems need to meet asymptotic stability criteria that are fulfilled by the recursion of Algorithm 1.

Proposition 8 (Dynamical systems properties of Algorithm 1) *The following conditions are satisfied.*

(A3) *For each $\rho \in H$, the primal ODE (13a) has a globally asymptotically stable equilibrium $\lambda(\rho)$, with $\lambda : H \to \mathbb{R}^{|\mathcal{S}|}$ a Lipschitz map.*

(A4) *The projected dual ODE (13b) has a globally asymptotically stable equilibrium $\rho^\star$.*

(A5) *Let $\{(V_k, \rho_k)\}$ be a sequence of iterates generated by (12) with a stepsize sequence fulfilling the assumption of Theorem 6. Then $\sup_k(\|V_k\| + \|\rho_k\|) < \infty$, a.s.*

Proof Sketch The full proof is stated in Appendix A. The main steps are as follows. To prove the convergence of the fast transient, $V(t) \to \lambda(\rho)$ for a fixed dual variable ρ stated in (A3), we provide a strict Lyapunov function argument. We show the convergence of the slow component given the equilibrated fast transient, (A4), by stating the well-posed projected dynamic system using results found in Dupuis and Nagurney (1993) and again constructing a strict Lyapunov function to establish global asymptotic stability with Dupuis (1987, Theorem 4.6). Lastly, to prove the iterate stability property, (A5), we apply the scaling limit or ODE@∞ method, established in Borkar and Meyn (2000). The dual estimators are bounded by projection to the compact hyperrectangle H with $\rho^\star \in H$. ■

Given the above stability properties, existing stochastic approximation theory by Borkar (2023) ensures the asymptotic convergence as follows.

Proof of Theorem 6 Given Propositions 7, and 8, and the two-timescale stepsize sequences assumed in Theorem 6, (Borkar, 2023, Theorem 8.1) applies and the coupled iterates (V_k, ρ_k) generated by the recursions (12), and hence by Algorithm 1, converge $\mathbb{P}$-a.s. to $(\lambda(\rho^\star), \rho^\star)$. ■

4. Asynchronous Algorithm

Assuming access to a generative model, as in Algorithm 1, can be restrictive in applications where samples can only be obtained by interacting with the environment. We therefore extend the approach to the standard online setting in which data are generated along a trajectory induced by a (fixed or time-varying) behavioral policy. This covers classical online methods such as Q-learning (Watkins and Dayan, 1992). The fixed behavioral policy setting is also related to offline RL where policies are learned from a fixed set of past experiences, see Levine et al. (2020) for an overview and Gabbianelli et al. (2024); Hong and Tewari (2024) for LP-based formulations.

Our algorithm retains the projected stochastic gradient descent-ascent structure, but performs *asynchronous* updates: only the state-action components visited by the behavioral policy are updated at each iteration. The almost-sure convergence analysis follows the asynchronous stochastic approximation framework of Konda and Borkar (1999); Perkins and Leslie (2013), while our finite-time rate bounds build on the two-timescale analysis developed in Zeng et al. (2024).

In addition to the related works mentioned in the discussion of Algorithm 1, the asynchronous algorithm shares features with the algorithms SBEED (Dai et al., 2018b) and Dual-AC (Dai et al., 2018a). They approximate different saddle-point formulations of the RL problem with behavioral policy access to the stochastic environment by sampling transitions from a growing experience replay buffer. In their convergence analysis, they assume a fixed behavioral policy. Our proposed algorithm combines the experience replay-based gradient estimation with the separation of the nested loop minimization by the two-timescale method and on-policy exploration. This combination allows us to apply the asynchronous stochastic approximation analysis in a way that has not been proposed for the regularized LP approach to reinforcement learning.

The asynchronous stochastic approximation setting features two sources of randomness: first, the index selection for the updates, which is based on the random transitions following the behavioral policy, and second, the stochastic gradient estimators that incorporate information from previous transitions.

We propose the following procedure for the index selection at a given update step in Algorithm 2. Let $\pi_k^b \in \Pi$ for $k \geq 0$ denote the behavioral policy at iteration k, and $\{X_k\} \subset \mathcal{X}$ the state-action process generated by the sequence of behavioral policies $\{\pi_k^b\}$ where

$$X_k := (s_{k-1}, a_{k-1}),\ s_k \sim \mathcal{P}(\cdot|s_{k-1}, a_{k-1}),\ a_k \sim \pi_{k-1}^b(\cdot|s_k).$$

At iteration k, the primal and dual variables are updated with projected stochastic gradient steps in the entries corresponding to the state s_k and state-action pair X_k, respectively.

Without the generative model, the trajectory $\{X_k\}$ is our only interaction with the environment. Since we cannot obtain new unbiased samples of every transition in each iteration under this Markovian observational model, we propose gradient estimators based on a version of experience replay to obtain asymptotically unbiased gradient directions. The experience replay buffer stores the observed transitions, which we use to estimate the transition probabilities featured in the true gradients as follows.

At each iteration, we add the current transition of the behavioral policy to the experience replay buffer, denoted as $\mathcal{D}_k$ in iteration k. To efficiently leverage the gathered experience, we opt for a structured buffer with a separate list $\mathcal{D}_k(s,a)$ for each state-action pair (s,a). Furthermore, for each state s', we maintain a set, denoted as $\mathcal{D}_{\text{inc}}(s')$, containing the "incoming" state-action pairs that have previously led to state s'.[1] Hence, if $X_k \notin \mathcal{D}_{\text{inc}}(s_k)$, we update $\mathcal{D}_{\text{inc}}(s_k) := \mathcal{D}_{\text{inc}}(s_k) \cup \{X_k\}$.

We estimate the transition probability vector $\mathcal{P}(\cdot|s_{k-1}, a_{k-1})$ in $\nabla_\rho L(V,\rho)(s_{k-1}, a_{k-1})$ with the unbiased new transitions s_k as the one-hot vector e_{s_k}. For the primal variable gradient entry $\nabla_V L(V,\rho)(s_k)$, we use buffer-based estimates of $\mathcal{P}(s_k|s,a)$ for state-action pairs with a previously observed transition to s_k. Concretely, for each $(s,a) \in \mathcal{D}_{\text{inc}}(s_k)$ we draw one next-state sample $\xi_k(s,a) \sim \mathcal{P}_{\mathcal{D}_k}(\cdot|s,a)$ from the corresponding buffer, where $\mathcal{P}_{\mathcal{D}_k}(s'|s,a) := \nu_k(s,a)^{-1} \sum_{\bar{s} \in \mathcal{D}_k(s,a)} \mathbf{1}\{\bar{s} = s'\}$, and set

$$\widehat{\mathcal{P}}_k(s_k|s,a) = \mathbf{1}\{\xi_k(s,a) = s_k\}.$$

For $(s,a) \notin \mathcal{D}_{\text{inc}}(s_k)$ no draw is made and we define $\widehat{\mathcal{P}}_k(s_k|s,a) = 0$. Denote the combined buffer-draw random variable as Ξ_k. The structured buffer draw can be interpreted as sampling from the empirical transition kernel. Under infinite visitation, this empirical kernel converges a.s. to the true kernel, hence the estimator is consistent.

Summarizing the above, the stochastic gradients for the asynchronous algorithm at iteration k are

$$\widehat{g}_k(\rho, V; X_k, \Xi_k) := e_{s_k}\left(\eta_V V(s_k) - \tilde{\rho}(s_k) + \gamma \sum_{(s,a)\in\mathcal{D}_{\text{inc}}(s_k)} \rho(s,a)\mathbf{1}\{\xi_k(s,a) = s_k\}\right) \tag{14a}$$

$$\widehat{h}_k(\rho, V; X_k) := e_{X_k}\left(-V(s_{k-1}) + r(X_k) + \gamma V(s_k) - \eta_\rho \log\left(\tfrac{\rho(X_k)}{\tilde{\rho}(s_{k-1})}\right)\right), \tag{14b}$$

where $\widehat{g}_k$ and $\widehat{h}_k$ are the estimates of $\nabla_V L$ and $\nabla_\rho L$, respectively.

1. The set of incoming state-action pairs does not contain duplicates. In contrast, the replay buffer consists of lists that allow duplicate entries.

Algorithm 2 Asynchronous PGDA-RL with Structured Experience Replay

Require: H as in (10), $V_0 \in \mathbb{R}^{|\mathcal{S}|}, \rho_0 \in H, \eta_\rho, \eta_V > 0, \{\alpha_k\}_{k\in\mathbb{N}}, \{\beta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_{>0}, \pi_0^b \in \Pi, \mu \in \Delta_{\mathcal{S}}, \{\epsilon_k\}_{k=1}^K \subset [0,1]$

Initialize visitation counters: $\tilde{\nu}_0 = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}|}, \nu_0 = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$

Initialize the structured experience replay buffer: For all $(s,a) \in \mathcal{S} \times \mathcal{A}$ set $\mathcal{D}_0(s,a) = \emptyset$

Initialize incoming states sets: For all $s \in \mathcal{S}$ set $\mathcal{D}_{\text{inc}}(s) = \emptyset$

Sample initial state-action pair: $s_0 \sim \mu, a_0 \sim \pi_0^b(\cdot|s_0)$

for $k = 1, 2, \ldots, K$ **do**

Set $X_k = (s_{k-1}, a_{k-1})$

Sample state-transition: $s_k \sim \mathcal{P}(\cdot|X_k)$

Sample action with behavioral policy: $a_k \sim \pi_{k-1}^b(\cdot|s_k)$

Update visitation counters: $\tilde{\nu}_k = \tilde{\nu}_{k-1} + e_{s_k}, \quad \nu_k = \nu_{k-1} + e_{X_k}$

Update experience replay buffers:

$$\mathcal{D}_k(X_k) = \mathcal{D}_{k-1}(X_k) \cup \{s_k\} \text{ (and } \mathcal{D}_k(s,a) = \mathcal{D}_{k-1}(s,a) \text{ for } (s,a) \neq X_k)$$

if $X_k \notin \mathcal{D}_{\text{inc}}(s_k)$ **then** $\mathcal{D}_{\text{inc}}(s_k) = \mathcal{D}_{\text{inc}}(s_k) \cup \{X_k\}$

For each $(s,a) \in \mathcal{D}_{\text{inc}}(s_k)$ draw a transition $\xi_k(s,a)$ uniformly from $\mathcal{D}_k(s,a)$, set

$$\widehat{\mathcal{P}}_k(s_k|s,a) = \mathbf{1}\{\xi_k(s,a) = s_k\}$$

Update V with an SGD step in the entry s_k with $\widehat{g}_k$ as in (14a):

$$V_k = V_{k-1} - \alpha\big(\tilde{\nu}_k(s_k)\big)\widehat{g}_k(\rho_{k-1}, V_{k-1}; X_k, \Xi_k)$$

Update ρ with a projected SGA step in the entry X_k with $\widehat{h}_k$ as in (14b):

$$\rho_k = \Pi_H\left[\rho_{k-1} + \beta\big(\nu_k(X_k)\big)\widehat{h}_k(\rho_{k-1}, V_{k-1}; X_k)\right]$$

(Optional) Update the behavioral policy: ▷ On-Policy Exploration

$$\pi_k^b(a|s) = (1-\epsilon_k)\frac{\rho_k(s,a)}{\sum_{a\in\mathcal{A}} \rho_k(s,a)} + \epsilon_k \mathcal{U}(\mathcal{A}), \qquad (s,a) \in \mathcal{S} \times \mathcal{A}$$

end for

return V_K, ρ_K

To ensure that each entry receives the same sequence of stepsizes when updates occur, we adjust the effective stepsizes based on visitation counts. We denote the number of visits to $(s,a) \in \mathcal{X}$ until iteration k as $\nu_k(s,a) := \sum_{\ell \leq k} \mathbf{1}\{X_\ell = (s,a)\}$ and correspondingly the number of visits to state s is the state marginal $\tilde{\nu}_k(s) := \sum_{a\in\mathcal{A}} \nu_k(s,a)$. Then, the asynchronous update at iteration k applies the random scalar stepsize $\bar{\alpha}_k := \alpha(\tilde{\nu}_k(s_k))$ and

$\bar{\beta}_k := \beta(\nu_k(X_k))$ for the primal and dual updates respectively.[2] The visitation counts act as "local clocks".

Rather than performing updates along the trajectory of a fixed behavior policy, it is often desirable to explore the state-action space using the current policy iterate in an online manner. The updating of the behavioral policy in Algorithm 2 allows for exploiting the learned policy during training while maintaining sufficient exploration through the projection of the dual iterate. While data collection follows an on-policy scheme, the updates are still performed off-policy using past experiences stored in the replay buffer. This approach ensures more stable off-policy learning with effective use of observed data.

4.1 Almost Sure Convergence Analysis of the Algorithm 2

We make the following assumptions to ensure sufficient exploration of the state-action space and obtain convergence results.

Assumption 9 (Ergodicity) *For any strictly exploratory policy π (that is, $\pi(a|s) > 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$), the induced Markov chain over the state space is ergodic.*

This is weaker than assuming ergodicity for all stationary policies. It further ensures that Markov chains on the state-action space induced by strictly exploratory policies are also ergodic.

Assumption 10 (Asynchronous Step-Size Sequences) *The stepsize sequences $\{\alpha_k\}$, $\{\beta_k\}$ satisfy*

1. $\sum_{k=1}^{\infty} \alpha_k = \sum_{k=1}^{\infty} \beta_k = \infty, \quad \sum_{k=1}^{\infty} \left(\alpha_k^2 + \beta_k^2\right) < \infty, \quad \lim_{k\to\infty} \frac{\beta_k}{\alpha_k} = 0.$
2. *For all k, for both $a_k = \alpha_k$ and $a_k = \beta_k$, $a_k \geq a_{k+1}$ and for all $x \in (0,1)$, there exists A_x with $\sup_k \frac{a_{[xk]}}{a_k} < A_x < \infty$.*

With the above assumptions, we obtain the following almost sure convergence result for both versions of Algorithm 2.

Theorem 11 (Almost Sure Convergence of Algorithm 2) *Consider the stochastic gradients defined in* (14)*. Let Assumption 9 hold and choose stepsize sequences $\{\alpha_k\}$,$\{\beta_k\}$ that satisfy Assumption 10. Let $\{(V_k, \rho_k)\}$ be a sequence generated by Algorithm 2 either with on-policy exploration or with a fixed strictly exploratory behavioral policy. Then, the last iterates converge $\mathbb{P}$-a.s. to the regularized saddle-point of* (9)*, that is,*

$$\lim_{K\to\infty} (V_K, \rho_K) = (V^\star, \rho^\star) \quad \mathbb{P}\text{-}a.s.$$

The proof of Theorem 11 is based on the asynchronous stochastic approximation theory by Borkar (1998), Konda and Borkar (1999), and the generalization in Perkins and Leslie (2013) that clarifies the conditions on the update frequencies.

2. Note that $\beta(k)$ and β_k both denote the k^{th} stepsize sequence element. Hence $\beta(\nu_k(s,a))$ is entry number $\nu_k(s,a) \in \mathbb{N}$ of the stepsize sequence $\{\beta(k)\}$.

To analyze the dynamics of the iterates generated by Algorithm 2, we define the filtrations

$$\mathcal{F}_k := \sigma(V_\ell, \rho_\ell, s_\ell, a_\ell, \Xi_\ell; \ell \leq k) \quad \text{and} \quad \mathcal{F}_k^- := \sigma(\mathcal{F}_{k-1}, s_k, a_k),\ k \geq 1.$$

The first filtration contains the complete history of the algorithm, while the latter filtration captures the information prior to sampling Ξ_k from the buffer $\mathcal{D}_k$ at iteration k. Indeed, $\mathcal{D}_k$ is a measurable function of $\{(s_\ell, a_\ell)\}_{\ell \leq k}$, hence $\mathcal{D}_k$ is $\mathcal{F}_k^-$-measurable. Clearly, $\mathcal{F}_{k-1} \subset \mathcal{F}_k^- \subset \mathcal{F}_k$.

We redefine the noise sequences $\{M_k^{(1)}\}$ and $\{M_k^{(2)}\}$ as

$$M_k^{(1)} := \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a) \left(\widehat{\mathcal{P}}_k(\cdot|s,a) - \mathbb{E}[\widehat{\mathcal{P}}_k(\cdot|s,a)|\mathcal{F}_k^-]\right), \tag{15a}$$

$$M_k^{(2)} := e_{X_k} \gamma \left(e_{s_k} - \mathcal{P}(\cdot|X_k)\right)^\top V_{k-1}, \tag{15b}$$

where $\widehat{\mathcal{P}}_k(\cdot|s,a) \equiv 0$ for pairs not drawn at time k. Note that V_{k-1} and $X_k = (s_{k-1}, a_{k-1})$ are $\mathcal{F}_{k-1}$-measurable and, since $s_k \sim \mathcal{P}(\cdot|X_k)$, we have $\mathbb{E}[e_{s_k}|\mathcal{F}_{k-1}] = \mathcal{P}(\cdot|X_k)$. Consequently, $\mathbb{E}[M_k^{(2)}|\mathcal{F}_{k-1}] = 0$. We further introduce the bias sequence

$$\mathcal{E}_k := \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a) \left(\mathbb{E}[\widehat{\mathcal{P}}_k(\cdot|s,a)|\mathcal{F}_k^-] - \mathcal{P}(\cdot|s,a)\right). \tag{16}$$

Note that the noise and bias sequences depend on the current iterates. When necessary, this dependence is made explicit. Denote the square diagonal matrices with the update index selectors as the diagonal entries as $\Lambda_k^{(1)} := \operatorname{diag}(e_{s_k}), \Lambda_k^{(2)} := \operatorname{diag}(e_{X_k})$. The asynchronous stochastic gradients (14) satisfy

$$\widehat{g}_k(\rho_{k-1}, V_{k-1}; X_k, \Xi_k) = \Lambda_k^{(1)}\big(\nabla_V L(V_{k-1}, \rho_{k-1}) + M_k^{(1)} + \mathcal{E}_k\big), \tag{17a}$$

$$\widehat{h}_k(\rho_{k-1}, V_{k-1}; X_k) = \Lambda_k^{(2)}\big(\nabla_\rho L(V_{k-1}, \rho_{k-1}) + M_k^{(2)}\big), \qquad \text{for all } k \in \mathbb{N}. \tag{17b}$$

Similarly to the synchronous case, the asymptotic analysis of the asynchronous stochastic iteration consists of analyzing the equilibrium properties of the dynamical systems that describe the limiting behavior of the iteration. In the asynchronous setting, the random selection of update components impact the limiting dynamics. Following the notation of Perkins and Leslie (2013), define for general $\epsilon > 0, d \in \mathbb{N}$, $\Omega_d^\epsilon := \{\operatorname{diag}(w_1, \ldots, w_d) | w_i \in [\epsilon, 1], i = 1, \ldots, d\}$. For the fast iterate V, we are concerned with the ODE

$$\dot{V}(t) = -\Lambda^{(1)}(t)\nabla_V L(V(t), \rho), \tag{18a}$$

where $\Lambda^{(1)} : \mathbb{R}_{\geq 0} \to \Omega_{|\mathcal{S}|}^\epsilon$ for some $\epsilon > 0$ is a measurable function. As in the synchronous case, we denote by $\lambda(\rho)$ the equilibrated value of V for a given ρ, that is, $\lambda(\rho) = \{V \in \mathbb{R}^{|\mathcal{S}|} : \nabla_V L(V, \rho) = 0\}$ and recall that since the mapping $V \mapsto L(V, \rho)$ is strictly convex, $\lambda(\rho)$ is unique. For the slower dual iterate, the ODE of interest is

$$\dot{\rho}(t) = \Lambda^{(2)}(t)\nabla_\rho L(\lambda(\rho(t)), \rho(t)) + \zeta(t), \quad \zeta(t) \in C(\rho(t)), \tag{18b}$$

with reflection term ζ and where $\Lambda^{(2)} : \mathbb{R}_{\geq 0} \to \Omega_{|\mathcal{S}|\cdot|\mathcal{A}|}^\epsilon$ for some $\epsilon > 0$ is a measurable function.

If we do not specify the diagonal matrix-valued functions $\Lambda^{(1)}(\cdot)$ and $\Lambda^{(2)}(\cdot)$ further, we can equivalently look at the properties of the class of dynamics captured by the differential inclusions

$$\dot{V}(t) \in -\Omega^{\epsilon}_{|\mathcal{S}|} \cdot \nabla_V L(V(t), \rho), \tag{19a}$$

$$\dot{\rho}(t) \in \Omega^{\epsilon}_{|\mathcal{S}|\cdot|\mathcal{A}|} \cdot \nabla_\rho L(\lambda(\rho(t)), \rho(t)) + \zeta(t), \quad \zeta(t) \in C(\rho(t)). \tag{19b}$$

Statements regarding the asymptotic properties of any solution to the differential inclusions defined in (19) also apply to the non-autonomous systems (18).

To relate the stochastic iterations to the dynamical systems and obtain convergence, we use the following properties of the (stochastic) gradients, the dynamical systems, and the update scheme.

Proposition 12 (Asynchronous SA) *Let $\{(V_k, \rho_k)\}$ be a sequence generated by Algorithm 2. The following properties hold.*

(B1) *For the noise terms in* (15) *and the bias in* (16) *it holds that*

- *(a)* $\{M_k^{(1)}\}$ *and* $\{M_k^{(2)}\}$ *are martingale-difference sequences with respect to* $\{\mathcal{F}_k\}$*, that is,* $\mathbb{E}[M_k^{(i)}|\mathcal{F}_{k-1}] = 0,\ i = 1, 2.$
- *(b) For a constant* $C > 0$ *and* $k \geq 1$

$$\mathbb{E}\left[\|M_k^{(1)}\|^2|\mathcal{F}_k^-\right] \leq C(1 + \|V_{k-1}\|^2 + \|\rho_{k-1}\|^2),$$
$$\mathbb{E}\left[\|M_k^{(2)}\|^2|\mathcal{F}_{k-1}\right] \leq C(1 + \|V_{k-1}\|^2 + \|\rho_{k-1}\|^2).$$

- *(c) The bias sequence* $\{\mathcal{E}_k\}$ *is a bounded random sequence satisfying* $\lim_{k\to\infty} \mathcal{E}_k = 0$ $\mathbb{P}$*-a.s.*

(B2) *For each* $\rho \in H$ *and every* $\epsilon > 0$*, for all* $\Omega^{\epsilon}_{|\mathcal{S}|}$*-valued measurable functions* $\Lambda^{(1)}(\cdot)$ *the ODE (18a) has a unique globally asymptotically stable equilibrium* $\lambda(\rho)$*. The function* $\lambda : H \to \mathbb{R}^{|\mathcal{S}|}$ *is Lipschitz continuous.*

(B3) *For every* $\epsilon > 0$*, for all* $\Omega^{\epsilon}_{|\mathcal{S}|\cdot|\mathcal{A}|}$*-valued measurable functions* $\Lambda^{(2)}(\cdot)$*, the projected ODE (18b) has a unique globally asymptotically stable equilibrium* $\rho^\star$*.*

(B4) $\sup_k(\|V_k\| + \|\rho_k\|) < \infty,\ \mathbb{P}$*-a.s.*

(B5) *For* $(s', a'), (s, a) \in \mathcal{X}$ *and any* $\rho \in H$*, let* $\mathcal{Q}_{(s,a),(s',a')}(\rho) := \mathcal{P}(s'|s, a)\ \pi^b_\rho(a'|s')$*. Then*

- *(a)* $\mathbb{P}((s_{k+1}, a_{k+1}) = (s', a')|\mathcal{F}_k) = \mathcal{Q}_{(s_k,a_k),(s',a')}(\rho_k)$*;*
- *(b) For each* $\rho \in H$*, the transition probabilities* $\mathcal{Q}_{(s,a),(s',a')}(\rho)$ *define an aperiodic, irreducible Markov chain over* $\mathcal{X}$*.*
- *(c) The mapping* $\rho \mapsto \mathcal{Q}_{(s,a),(s',a')}(\rho)$ *is Lipschitz continuous.*

The proof is stated in Appendix B.

The properties (B1)-a,b and (B4) do not change compared to the respective assumptions for the synchronous scheme. Under property (B1)-c, the bias term $\mathcal{E}_k$ vanishes and does not affect the limiting dynamics (see Remark 3 in (Borkar, 2023, Section 2.2)). Properties (B2) and (B4) are the asynchronous counterparts of (A4) and (A5. Property (B5) ensures that the state-action index process (s_k, a_k) is an ergodic Markov chain on $\mathcal{X}$, which implies that every state-action component of ρ and every state component of V is updated with positive asymptotic frequency (see also (Perkins and Leslie, 2013, Lemma 3.4)).

Proof of Theorem 11 Let Assumptions 9 and 10 hold, and let the stochastic gradients be as stated in (14). Given the properties listed in Proposition 12 and with the asynchronous step sizes defined in Assumption 10, the fast iterates' interpolated trajectory converges to the trajectory of a solution of the differential inclusion (19a) (Perkins and Leslie, 2013, Lemma 4.3), and the slow iterate asymptotically tracks a solution to (19b) almost surely (Perkins and Leslie, 2013, Theorem 4.7).

Finally, the almost sure convergence of the last iterates of Algorithm 2 follows from the global asymptotic equilibrium property of the projected ODE (18b), (B3), by applying (Perkins and Leslie, 2013, Corollary 4.8). ■

4.2 Convergence Rate Analysis of Algorithm 2

In this section, we derive the convergence rate for the two-timescale asynchronous PGDA algorithm. For the derivation we impose one change compared to the stated Algorithm 2, by projecting the primal variable iterates V_k to the compact set $\mathcal{V}_r \subset \mathbb{R}_+^{|\mathcal{S}|}$, where

$$\mathcal{V}_r := \{v \in \mathbb{R}_+^{|\mathcal{S}|} : v_i \leq \frac{C_r + \eta_\rho U_G}{1-\gamma} \ \forall i = 1, \dots |\mathcal{S}|\}. \tag{20}$$

As discussed in the proof of Corollary 2, this projection does not affect the optimal primal variable.

To establish the almost sure convergence result in Theorem 11, we assumed that for any strictly exploratory policy, the induced Markov chain is ergodic (Assumption 9). For the derivation of the convergence rate, we strengthen this assumption slightly as follows.

We denote Dobrushin's ergodic coefficient for a Markov kernel Q on $\mathcal{X}$ as

$$\delta(Q) := \sup_{x,y \in \mathcal{X}} \|\delta_x Q - \delta_y Q\|_{\mathrm{TV}}.$$

It measures the maximal contraction of the kernel in total variation distance; see Dobrushin (1956) and Brémaud (2020, Definition 4.3.11).

Assumption 13 (Uniform mixing of dual-induced chains) *Let $\{\mathcal{P}_\rho\}_{\rho \in H}$ denote the family of state-action transition kernels induced by the dual-induced policies $\{\pi_\rho\}_{\rho \in H}$.*

(i) There exist constants $C_\mathcal{X} > 0$ and $\varrho \in (0, 1)$ such that, for every $\rho \in H$, the Markov chain with kernel $\mathcal{P}_\rho$ has a unique stationary law μ_ρ and

$$\sup_{x \in \mathcal{X}} \|\delta_x \mathcal{P}_\rho^k - \mu_\rho\|_{\mathrm{TV}} \leq C_\mathcal{X} \varrho^k, \qquad k \geq 0.$$

(ii) *There exist constants* $m_\star \in \mathbb{N}$ *and* $\kappa \in (0,1)$ *such that, for every sequence* $\{\rho_t\}_{t\in\mathbb{N}} \subset H$,

$$\sup_{k\geq 0} \delta\left(\prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t}\right) \leq \kappa.$$

Remark 14 *A sufficient condition for Assumption 13 is the existence of a deterministic state-feedback policy* $\pi_d : \mathcal{S} \to \mathcal{A}$ *whose induced state chain is uniformly geometrically ergodic, that is, it has a unique stationary law and converges to it geometrically fast in total variation, uniformly over initial states; see Lemma 24.*

The convergence rate analysis presented here builds on the recent work of Zeng et al. (2024). The analysis applies with a specific choice of the two-timescale stepsize sequences. In the following let the stepsize sequences be

$$\alpha_k = \frac{\alpha_0}{(k+1)^{2/3}}, \qquad \beta_k = \frac{\beta_0}{k+1}, \qquad \alpha_0,\, \beta_0 > 0. \tag{21}$$

It can be easily verified that this choice satisfies Assumption 10.

We extend and adapt the convergence rate analysis of Zeng et al. (2024) for our LP-based algorithm framework and asynchronous updating scheme in two important nontrivial aspects.

First, while Zeng et al. (2024) considers only the global decaying stepsize setting, our focus is on the asynchronous regime with local stepsizes, which introduces randomness and non-monotonicity to the realized stepsize sequence. To handle the non-monotonicity, we switch the analysis to the currently least frequently visited state or state-action. This coordinate still receives updates linear in iteration counter k eventually, with a high-probability bound on the time when the linear growth rate applies. The proof is based on the uniform geometric ergodicity of the dual policy induced chains and Lipschitz properties of the dual-induced Markov kernels.

Second, unlike Zeng et al. (2024), the stochastic gradient estimator based on the replay buffer is biased in our setting. We find that the bias decays at a rate fast enough to not become dominating.

These properties of the asynchronous PGDA iterates as well as further regularity and boundedness properties form the basis of our convergence rate proof and are stated in the following proposition. Throughout, let $p_\star$ denote a uniform lower bound on the dual-variable induced stationary distributions over the state-action space,[3] valid for all $\rho \in H$. Denote the minimal number of visits to any state-action pair until iteration k as $\underline{\nu}_k := \min_{(s,a)\in\mathcal{X}} \nu_k(s,a)$. It holds that $\underline{\nu}_k \leq \nu_k(s,a) \leq k$ for all $(s,a) \in \mathcal{X}$. Denote the reduced Lagrangian objective as

$$f(\rho) = \min_V L(V,\rho). \tag{22}$$

Recall that $\bar{\alpha}_k = \alpha(\tilde{\nu}_k(s_k))$ and $\bar{\beta}_k = \beta(\nu_k(X_k))$.

Proposition 15 *There exist positive constants* μ_{opt}, L, B, D, *and for each* $\delta \in (0,1)$, *an event* $\mathcal{G}_\delta$ *with* $\mathbb{P}(\mathcal{G}_\delta) \geq 1-\delta$ *and an index* $K(\delta) \in \mathbb{N}$ *such that on* $\mathcal{G}_\delta$ *and for all* $k \geq K(\delta)$,

3. An explicit expression for $p_\star$ is derived in the proof of Lemma 25.

(i) *(Visitation floor) It holds that* $\underline{\nu}_k \geq (p_\star/2)\,k$*, and there exists a constant* c *such that* $c \cdot \alpha(\underline{\nu}_k) \leq \bar{\alpha}_k \leq \alpha(\underline{\nu}_k)$*, and* $c \cdot \beta(\underline{\nu}_k) \leq \bar{\beta}_k \leq \beta(\underline{\nu}_k)$ *with probability at least* $1-\delta$ *for* $k \geq K(\delta)$*.*

(ii) *(Replay-buffer bias) There exists a constant* $C_{\text{buf}}(\delta) > 0$*, such that for all* $\rho \in H$ *the buffer-induced bias* $\mathcal{E}_k(\rho)$*, see* (16)*, decays as*

$$\|\mathbb{E}_{s\sim\tilde{\mu}_\rho}[e_s\,\mathcal{E}_k(\rho)(s)]\|_\infty \leq C_{\text{buf}}(\delta)\sqrt{\tfrac{1}{k}}$$

with $C_{\text{buf}}(\delta) = O(\log(1/\delta))$*.*

Moreover, the following regularity and boundedness properties hold.

(iii) *(Regularity)* $V \mapsto L(V,\rho)$ *is* η_V*-strongly convex for all* $\rho \in H$*;* $\rho \mapsto \lambda(\rho)$ *is Lipschitz on* H*;* $\rho \mapsto f(\rho)$ *is* L*-smooth and* μ_{opt}*-strongly concave on* H*.*

(iv) *(Uniform bounds) The stochastic gradients are uniformly bounded over* $H \times \mathcal{V}_r$ *for all* $k \geq 1$*:* $\sup_{\rho\in H,\,V\in\mathcal{V}_r} \|\widehat{g}_k(\rho,V;\cdot,\cdot)\| \leq B$*,* $\sup_{\rho\in H,\,V\in\mathcal{V}_r} \|\widehat{h}_k(\rho,V;\cdot)\| \leq D$*.*

The proof of Proposition 15 is provided in Section C.1.

Before stating the convergence rate theorem, we provide the necessary additional notation. We focus the analysis on the least-visited state or state-action entry of the iterates. With the above stated visitation floor $\underline{\nu}_k \geq (p_\star/2)k$, we can bound the minimal visits to all state-action pairs and introduce the monotone decreasing upper envelope on the primal and dual stepsizes respectively as

$$\alpha_{\text{env},k} := \alpha(\lfloor p_\star k/2\rfloor), \qquad \beta_{\text{env},k} := \beta(\lfloor p_\star k/2\rfloor). \tag{23}$$

Note that $\frac{\alpha_{\text{env},k}}{\alpha_k} \leq \frac{4}{p_\star} =: C_{\text{env}}$. We define the mixing-time function as the worst-case mixing time across all admissible dual iterates. For $\zeta > 0$

$$\tau(\zeta) := \sup_{\rho\in H} \min\{\ell \in \mathbb{N} : \sup_{x\in\mathcal{X}} \|\delta_x \mathcal{P}_\rho^\ell - \mu_\rho\|_{\text{TV}} \leq \zeta\}, \quad \tau_k := \tau\big(\beta_{\text{env},k}\big).$$

The uniform ergodicity of the Markov chains corresponding to the strictly positive dual policies now implies that there exists a constant $C_\tau > 0$ such that

$$\tau_k \leq C_\tau \log(k+1)$$

and hence $\lim_{k\to\infty} \tau_k^2 \alpha_k = 0$. Consequently, there exists $\mathcal{K} \in \mathbb{N}$ such that for all $k \geq \mathcal{K}$

$$\tau_k^2 \alpha_{\text{env},k-\tau_k} \leq \frac{\tilde{\eta}_V}{18\,C_1\,C_{\text{env}}},\ \tau_k \alpha_{\text{env},k-\tau_k} \leq \min\{\tfrac{1}{\sqrt{3D}}, \tfrac{1}{\sqrt{3LB}}\},\ \alpha_k \leq \tfrac{1}{\tilde{\eta}_V},\ \tfrac{p_\star k}{2} \geq 1, \tag{24}$$

where

$$C_1 := 2(LDB + LB + LD + D^2 + \tfrac{\gamma|\mathcal{X}|C^U}{\beta_0}), \qquad \tilde{\eta}_V := \eta_V\, p_\star |\mathcal{A}|. \tag{25}$$

Lastly, there exists $c_\tau \in (0,1)$, such that for all $k \geq \tau_k$, $c_\tau(k+1) \leq k+1-\tau_k$.

With these definitions at hand, we can quantify the convergence rate of the asynchronous PGDA iterates as follows.

Theorem 16 (Convergence rate of Algorithm 2) *Let Assumption 13 hold. Let $\{\rho_k\}$, $\{V_k\}$ denote the iterates of Algorithm 2 with the projection on the primal iterates to $\mathcal{V}_r$. Fix $\delta \in (0,1)$ and let $\mathcal{G}_\delta$, and $K(\delta)$ be as in Proposition 15. Let the stepsize schedules be as in (21) with*

$$\beta_0 \geq \frac{2}{p_\star \mu_{\text{opt}}}, \quad \frac{\beta_0}{\alpha_0} \leq \frac{\tilde{\eta}_V}{8 p_\star \mu_{\text{opt}}}.$$

Then, on $\mathcal{G}_\delta$, for all $k \geq k_0 := \max\{K(\delta), \mathcal{K}\}$,

$$\mathbb{E}[f(\rho^\star) - f(\rho_k)|\mathcal{G}_\delta] \leq \left(\mathbb{E}[f(\rho^\star) - f(\rho_{k_0})|\mathcal{G}_\delta] + \frac{16L^2}{p_\star^3 \tilde{\eta}_V} \frac{\beta_{k_0}}{\alpha_{k_0}} \mathbb{E}\big[\|V_{k_0} - \lambda(\rho_{k_0})\|_2^2 |\mathcal{G}_\delta\big]\right) \frac{k_0+1}{k+1} + \frac{2C_2(\delta)\log^2(k+1)}{3(k+1)^{2/3}},$$

where L is a Lipschitz parameter and $C_2(\delta)$ is a constant depending only on model parameters and δ but not on k. Moreover,

$$\mathbb{E}\big[\|\rho_k - \rho^\star\|_2^2 |\mathcal{G}_\delta\big] \leq \tfrac{2}{\mu_{\text{opt}}} \mathbb{E}[f(\rho^\star) - f(\rho_k)|\mathcal{G}_\delta] = \tilde{\mathcal{O}}(k^{-2/3}).$$

The rate of convergence of the order $\widetilde{\mathcal{O}}(k^{-2/3})$ recovers the existing rate of Zeng et al. (2024) for two-timescale stochastic gradient schemes under Markovian model access, albeit under the more challenging setting of buffer-based biased gradient estimates and accounting for asynchronous updating with visitation-adjusted stepsizes.

As a consequence of the finite-time convergence rate of the dual iterates (Theorem 16), we obtain the following bound on the unregularized value suboptimality of the dual-induced policies; see Appendix C.3 for the proof.

Corollary 17 (Value suboptimality of the induced policy) *Under the assumptions of Theorem 16, let $L_\pi > 0$ satisfy*

$$\max_{s\in\mathcal{S}} \|\pi_\rho(\cdot|s) - \pi_{\hat{\rho}}(\cdot|s)\|_1 \leq L_\pi \|\rho - \hat{\rho}\|_2, \qquad \forall\, \rho, \hat{\rho} \in H,$$

and define

$$L_V := \frac{C_r}{(1-\gamma)^2} L_\pi, \qquad B_{\text{reg}} := \frac{\eta_\rho \log|\mathcal{A}|}{1-\gamma}.$$

Then, for every $s \in \mathcal{S}$ and all $k \geq k_0$,

$$\mathbb{E}\left[\left(V_{ur}^\star(s) - V_{ur}^{\pi_{\rho_k}}(s) - B_{\text{reg}}\right)_+^2 \,\middle|\, \mathcal{G}_\delta\right] \leq L_V^2\, \mathbb{E}\big[\|\rho_k - \rho^\star\|_2^2 \,\big|\, \mathcal{G}_\delta\big].$$

In particular, the excess unregularized value suboptimality beyond the deterministic regularization bias B_{reg} inherits the conditional mean-square rate $\widetilde{\mathcal{O}}(k^{-2/3})$ from Theorem 16.

The proof of Theorem 16 builds on two contraction inequalities, one each for the primal and the dual variable, which are then connected by the two-timescale lemma (Zeng et al., 2024, Lemma 4). Recall that we overload the Lipschitz constant notation and denote all Lipschitz constants by the envelope L.

Proposition 18 *For $\delta \in (0,1)$ let $\mathcal{G}_\delta$ denote the high-probability visitation event of Proposition 15. Let C_1 be as in (25). Then, for all $k \geq \max\{K(\delta), \mathcal{K}\}$, the primal variable tracking error $\mathbb{E}\left[\|V_k - \lambda(\rho_k)\|_2^2 | \mathcal{G}_\delta\right]$ satisfies*

$$\mathbb{E}\left[\|V_{k+1} - \lambda(\rho_{k+1})\|_2^2 | \mathcal{G}_\delta\right] \leq (1 - \tfrac{\tilde{\eta}_V}{4}\alpha_k)\mathbb{E}\left[\|V_k - \lambda(\rho_k)\|_2^2 | \mathcal{G}_\delta\right] + r_k(\delta),$$

where

$$r_k(\delta) := \tfrac{9}{2} C_1 \tau_k^2 \alpha_{\mathrm{env},k-\tau_k} \alpha_{\mathrm{env},k} + \tfrac{3}{2} \tfrac{C_{\mathrm{buf}}(\delta)^2 |\mathcal{S}|}{\tilde{\eta}_V} \tfrac{\alpha_{\mathrm{env},k}}{k} + \tfrac{3}{2} D^2 \alpha_{\mathrm{env},k}^2 + L^2 B^2 \beta_{\mathrm{env},k}^2 + \tfrac{2L^2B^2}{\tilde{\eta}_V} \tfrac{\beta_{\mathrm{env},k}^2}{\alpha_k}.$$

Proposition 19 *For $\delta \in (0,1)$ let $\mathcal{G}_\delta$ denote the high-probability visitation event of Proposition 15. Then for all $k \geq \max\{K(\delta), \mathcal{K}\}$, the reduced objective error $\mathbb{E}[f(\rho^\star) - f(\rho_k) | \mathcal{G}_\delta]$ satisfies*

$$\mathbb{E}[f(\rho^\star) - f(\rho_{k+1}) | \mathcal{G}_\delta] \leq (1 - p_\star \mu_{\mathrm{opt}}\, \beta_k)\mathbb{E}[f(\rho^\star) - f(\rho_k) | \mathcal{G}_\delta] + \frac{L^2 \beta_{\mathrm{env},k}}{2p_\star} \mathbb{E}[\|V_k - \lambda(\rho_k)\|_2^2 | \mathcal{G}_\delta] + \frac{25 L^2 B^3}{2} \tau_k^2 \beta_{\mathrm{env},k-\tau_k} \beta_{\mathrm{env},k},$$

where $\mu_{\mathrm{opt}} > 0$ is the strong concavity modulus of the reduced objective f.

The proofs of the two propositions are stated in Appendix C.2.

The proof of Theorem 16 now consists of combining the recursive contractions from Propositions 18 and 19 using the two-timescale lemma (Zeng et al., 2024, Lemma 4, Case 1), restated here for completeness.

Lemma 20 (Two-timescale lemma, Zeng et al. 2024) *Let $\{a_k, b_k, c_k, d_k, e_k, f_k\}$ be non-negative sequences satisfying $\frac{a_{k+1}}{d_{k+1}} \leq \frac{a_k}{d_k} < 1$ for all $k \geq 0$. Let $\{x_k\}, \{y_k\}$ be two nonnegative sequences. Suppose that x_k, y_k satisfy the following coupled inequalities:*

$$x_{k+1} \leq (1 - a_k)\, x_k + b_k y_k + c_k, \quad y_{k+1} \leq (1 - d_k)\, y_k + e_k x_k + f_k.$$

In addition, assume that there exists a constant $A \in \mathbb{R}$ such that

$$A a_k \big(1 - \frac{a_k}{d_k}\big) - b_k \geq 0 \quad \text{and} \quad \frac{A e_k}{d_k} \leq \frac{1}{2} \quad \forall k \geq 0.$$

Then we have for all $0 \leq \tau \leq k$

$$x_k \leq \left(x_\tau + \frac{A a_\tau}{d_\tau} y_\tau\right) \prod_{t=\tau}^{k-1} \left(1 - \frac{a_t}{2}\right) + \sum_{\ell=\tau}^{k-1} \left(c_\ell + \frac{A a_\ell f_\ell}{d_\ell}\right) \prod_{t=\ell+1}^{k-1} \left(1 - \frac{a_t}{2}\right).$$

Proof of Theorem 16 Fix $\delta \in (0,1)$ and work on $\mathcal{G}_\delta$ defined in Proposition 15. Let $k_0 := \max\{K(\delta), \mathcal{K}\}$ be the burn-in index such that the visitation floor bounds and the technical conditions in (24) hold for all $k \geq k_0$. Set

$$x_k := \mathbb{E}\left[f(\rho^\star) - f(\rho_k) | \mathcal{G}_\delta\right], \qquad y_k := \mathbb{E}\left[\|V_k - \lambda(\rho_k)\|_2^2 | \mathcal{G}_\delta\right].$$

The proof proceeds in four steps.

(I.) Following the two coupled inequalities of Proposition 18 and Proposition 19 it holds that for all $k \geq k_0$

$$x_{k+1} \leq (1 - a_k)x_k + b_k y_k + c_k, \tag{26}$$

where

$$a_k := p_\star \mu_{\text{opt}} \beta_k, \qquad b_k := \frac{L^2}{2p_\star}\beta_{\text{env},k}, \qquad c_k := \frac{25L^2B^3}{2}\,\tau_k^2\,\beta_{\text{env},k-\tau_k}\beta_{\text{env},k},$$

and

$$y_{k+1} \leq (1 - d_k)y_k + f_k, \tag{27}$$

where

$$d_k := \frac{\tilde{\eta}_V}{4}\alpha_k, \qquad e_k := 0, \qquad f_k := r_k(\delta).$$

Inserting the stepsize sequences into $r_k(\delta)$ (see Proposition 18) and adding the constant factors, we find that, with constant $C_R(\delta) := \frac{12}{p_\star \tilde{\eta}_V} C_{\text{buf}}(\delta)^2 |\mathcal{S}| \alpha_0 + \frac{12}{p_\star^2}\alpha_0^2(\frac{6C_1C_\tau^2}{c_\tau} + 2D^2) + \frac{16}{p_\star^2} L^2 B^2 \beta_0^2 (1 + \frac{2}{\alpha_0 \tilde{\eta}_V})$, for all $k \geq k_0$,

$$f_k \leq \frac{C_R(\delta)\,\log^2(k+1)}{(k+1)^{4/3}}. \tag{28}$$

(II.) We proceed to determine a feasible constant $A \in \mathbb{R}$ to apply Lemma 20. Let $k \geq k_0$. Inserting the chosen stepsize scheme (21) into the fraction a_k/d_k gives

$$\frac{a_k}{d_k} = \frac{p_\star \mu_{\text{opt}} \beta_k}{(\tilde{\eta}_V/4)\alpha_k} = \frac{4p_\star \mu_{\text{opt}}}{\tilde{\eta}_V}\frac{\beta_k}{\alpha_k} = \frac{4p_\star \mu_{\text{opt}}}{\tilde{\eta}_V}\frac{\beta_0}{\alpha_0 (k+1)^{1/3}}. \tag{29}$$

By the imposed conditions on the stepsize parameters, $\beta_0 \geq \frac{2}{p_\star \mu_{\text{opt}}}$ and α_0 is set such that $\frac{\beta_0}{\alpha_0} \leq \frac{\tilde{\eta}_V}{8p_\star \mu_{\text{opt}}}$. Since $\frac{\beta_k}{\alpha_k} = \frac{\beta_0}{\alpha_0 (k+1)^{1/3}}$ is non-increasing in k, this implies $\frac{a_k}{d_k} \leq \frac{1}{2}(k+1)^{-1/3} \leq \frac{1}{2}$. Moreover, by the stepsize envelope construction (see 23),

$$\beta_{\text{env},k} \leq \beta(\lfloor \tfrac{p_\star k}{2} \rfloor) = \frac{\beta_0}{\lfloor \frac{p_\star k}{2} \rfloor + 1} \leq \tfrac{4}{p_\star}\,\beta_k.$$

We set $A := \frac{4L^2}{p_\star^3 \mu_{\text{opt}}}$. Then, using $\frac{a_k}{d_k} \leq \frac{1}{2}$ and $b_k = \frac{L^2}{2p_\star}\beta_{\text{env},k} \leq \frac{2L^2}{p_\star^2}\beta_k$, we obtain

$$A a_k \Big(1 - \frac{a_k}{d_k}\Big) - b_k \geq \frac{A}{2} a_k - b_k \geq \frac{2L^2}{p_\star^2 \mu_{\text{opt}}}\mu_{\text{opt}}\beta_k - \frac{2L^2}{p_\star^2}\beta_k = 0,$$

and $A e_k / d_k = 0$ since $e_k = 0$. Hence, Lemma 20 applies for the coupled system (26)–(27) and gives

$$x_k \leq \Big(x_{k_0} + A\frac{a_{k_0}}{d_{k_0}} y_{k_0}\Big) \prod_{t=k_0}^{k-1}\Big(1 - \frac{a_t}{2}\Big) + \sum_{\ell=k_0}^{k-1}\Bigg[\prod_{t=\ell+1}^{k-1}\Big(1 - \frac{a_t}{2}\Big)\Big(c_\ell + A\frac{a_\ell}{d_\ell} f_\ell\Big)\Bigg]. \tag{30}$$

(III.) To derive the convergence rate result, it remains to bound the terms of the above inequality. First, using $\tau_k \leq C_\tau \log(k+1)$ and $\beta_{\text{env},k-\tau_k} \leq \frac{4}{c_\tau p_\star}\beta_k$,

$$c_k \leq \frac{25L^2B^3}{2c_\tau}\,(C_\tau^2 \log^2(k+1))\,(\tfrac{4}{p_\star}\beta_k)^2 = \frac{200L^2B^3C_\tau^2\beta_0^2}{c_\tau p_\star^2} \cdot \frac{\log^2(k+1)}{(k+1)^2}.$$

Second, combine (28) and (29) to get

$$A\frac{a_k}{d_k}f_k \le \frac{2L^2C_R(\delta)}{p_\star^3\mu_{\rm opt}}\cdot\frac{\log^2(k+1)}{(k+1)^{5/3}}.$$

Therefore, the multiplier in the second term of (30) satisfies

$$c_k + A\frac{a_k}{d_k}f_k \le C_2(\delta)\frac{\log^2(k+1)}{(k+1)^{5/3}}, \qquad C_2(\delta) := \frac{2L^2C_R(\delta)}{p_\star^3\mu_{\rm opt}} + \frac{200L^2B^3C_\tau^2\beta_0^2}{c_\tau p_\star^2}, \tag{31}$$

since $(k+1)^{-2} \le (k+1)^{-5/3}$ for all $k \ge 1$.

(IV.) Using $1-u \le e^{-u}$ and the harmonic-sum bound $\sum_{t=k_1}^{k_2}\frac{1}{t+1} \ge \log(\frac{k_2+2}{k_1+1})$, for all $k_0 \le \ell < k$,

$$\begin{aligned}\prod_{t=\ell+1}^{k-1}\Big(1-\frac{a_t}{2}\Big) &\le \exp\Big(-\frac{1}{2}\sum_{t=\ell+1}^{k-1}a_t\Big) = \exp\Big(-\frac{p_\star\mu_{\rm opt}\beta_0}{2}\sum_{t=\ell+1}^{k-1}\frac{1}{t+1}\Big)\\ &\le \Big(\frac{\ell+2}{k+1}\Big)^{\frac{p_\star\mu_{\rm opt}\beta_0}{2}} \le \frac{\ell+2}{k+1},\end{aligned}$$

where the last inequality is due to the condition $\beta_0 \ge \frac{2}{p_\star\mu_{\rm opt}}$. With the same reasoning $\prod_{t=k_0}^{k-1}\big(1-\frac{a_t}{2}\big) \le \frac{k_0+1}{k+1}$. Inserting the harmonic sum bounds and (31) in (30) and using $\frac{\ell+2}{k+1} \le 2\frac{\ell+1}{k+1}$ yields, for all $k \ge k_0$,

$$x_k \le \left(x_{k_0} + \frac{16L^2}{p_\star^3\tilde\eta_V}\frac{\beta_{k_0}}{\alpha_{k_0}}y_{k_0}\right)\frac{k_0+1}{k+1} + \frac{2C_2(\delta)\log^2(k+1)}{k+1}\sum_{\ell=k_0}^{k-1}\frac{1}{(\ell+1)^{2/3}},$$

where for the second summand, we use the monotonicity $\log^2(\ell+1) \le \log^2(k+1)$ for $\ell \le k$. We apply an integral bound to the sum. Since $y \mapsto y^{-2/3}$ is decreasing on $(0,\infty)$, we have $(\ell+1)^{-2/3} \le \int_\ell^{\ell+1} y^{-2/3}dy$. Hence, $\sum_{\ell=k_0}^{k-1}(\ell+1)^{-2/3} \le \int_{k_0}^k y^{-2/3}dy = 3\big(k^{1/3}-k_0^{1/3}\big) \le 3(k+1)^{1/3}$. Inserting the integral bound finally yields the claimed rate

$$\begin{aligned}\mathbb{E}\left[f(\rho^\star)-f(\rho_k)|\mathcal{G}_\delta\right] \le \Big(\mathbb{E}\left[f(\rho^\star)-f(\rho_{k_0})|\mathcal{G}_\delta\right] + \frac{16L^2}{p_\star^3\tilde\eta_V}\frac{\beta_{k_0}}{\alpha_{k_0}}\mathbb{E}\left[\|V_{k_0}-\lambda(\rho_{k_0})\|_2^2|\mathcal{G}_\delta\right]\Big)\frac{k_0+1}{k+1}\\ + \frac{6C_2(\delta)\log^2(k+1)}{(k+1)^{2/3}}.\end{aligned}$$

Since $\rho^\star \in {\rm int}H$, it holds that $\nabla f(\rho^\star) = 0$, hence with the strong concavity of f (see Proposition 15-(iv) and the proof in Lemma 29)

$$\begin{aligned}f(\rho_k) &\le f(\rho^\star) + \langle\nabla f(\rho^\star), \rho_k-\rho^\star\rangle - \tfrac{\mu_{\rm opt}}{2}\|\rho_k-\rho^\star\|_2^2\\ &\Rightarrow \|\rho_k-\rho^\star\|_2^2 \le \tfrac{2}{\mu_{\rm opt}}\left(f(\rho^\star)-f(\rho_k)\right),\end{aligned}$$

■

5. Numerical Results

We evaluate PGDA-RL on the FrozenLake-v1 environment (Towers et al., 2024), a standard reinforcement learning benchmark with discrete states and actions.[4] The environment consists of a 4×4 grid in which the agent must reach a goal state while avoiding holes. The action space contains four actions (up, down, left, right) and transitions are stochastic due to the "slippery" dynamics. To adapt this episodic task to our infinite-horizon setting, we treat reaching the goal or a hole as a transition back to the start state, allowing continuous learning. The reward for reaching the goal is 100, and all other transitions yield zero reward.

We set the model parameters to $\{\gamma, \eta_\rho, \eta_V\} = \{0.9, 0.1, 0.1\}$. The asynchronous stepsizes are $\alpha_k = (1+\tilde{k})^{-2/3}$ and $\beta_k = (1+\tilde{k})^{-1}$ with $\tilde{k} = 9+k/100$. We apply Algorithm 2 with on-policy exploration and an additional ϵ-greedy strategy with linearly decaying ϵ ($[\epsilon_0, \epsilon_K] = [1, 0.1]$, $K = 10^5$) to mitigate infrequent updates of rarely visited state-action pairs. Replay buffers are capped at size 10^3. In the numerical implementation, we additionally impose a lower bound of 10^{-12} on the entries of the dual iterate after projection to avoid numerical instabilities when projected components become extremely small.

We track value-function and policy errors throughout training. In particular, the reported relative root-MSE (rRMSE) is the relative ℓ_2 error $\|V - \bar{V}\|_2 / \|\bar{V}\|_2$ evaluated on non-terminal states, where $\bar{V}$ denotes the corresponding reference value.

Figure 1 summarizes the learning dynamics. Panel (a) shows that both the primal value iterate and the value induced by the current dual policy steadily reduce their relative value error on non-terminal states. The dual-policy value improves faster and reaches a lower error level over the shown horizon. This separation is consistent with the two-timescale design: while ρ_k evolves slowly toward $\rho^\star$, the value update primarily tracks the moving best response $\lambda(\rho_k)$, so the policy-induced value can stabilize earlier than the primal iterate over the finite training horizon. Panel (b) tracks the initial-state value under the dual-iterate policy: after a rapid increase during early training, the curve approaches the initial state value of the optimal regularized policy. The separation between the optimal unregularized value and the value of the optimal regularized policy shows the suboptimality induced by the entropy regularization. Panel (c) compares PGDA-RL against the generator-based benchmark SPD-dMDP (Chen and Wang, 2016) that builds on the LP approach to (unregularized) MDPs. SPD-dMDP improves more quickly early on, whereas PGDA-RL closes the gap over longer training and achieves comparable final accuracy over the shown horizon. A key difference is the sampling model: SPD-dMDP draws transitions from uniformly sampled state-action pairs, while our method relies on trajectory data generated by the current behavior policy. Moreover, SPD-dMDP reports the averaged iterate as output, which yields a visibly smoother curve than the last-iterate behavior shown for PGDA-RL. Finally, panel (d) reports the KL divergence between the learned dual iterate policy and the optimal regularized policy summed uniformly over non-terminal states, $\mathsf{D}(\pi_r^\star \parallel \pi_{\rho_k}) = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \pi_r^\star(a|s) \log\left(\frac{\pi_r^\star(a|s)}{\pi_{\rho_k}(a|s)}\right)$, which decreases consistently over training, indicating convergence in policy space alongside value improvement. In summary, the simulations support our theoretical findings by demonstrating convergence of the dual-induced policy π_{ρ_k} toward the unique optimal regularized policy $\pi_r^\star$ on all non-terminal states.

4. All experiments are implemented in Python, and the code for reproducing all numerical results is available from `https://github.com/AxelFW/tt-primal-dual-rl`.

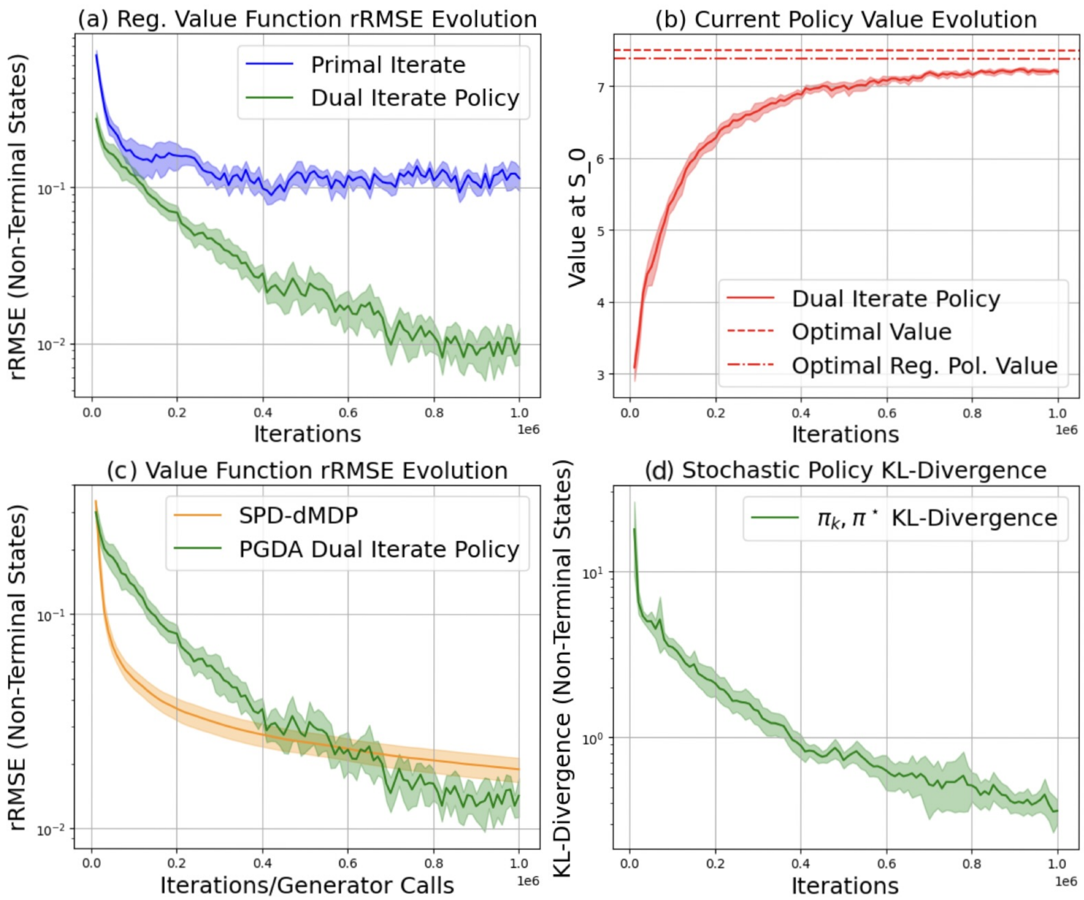

Figure 1: Value function and policy evaluation during training. The plots show the mean (solid lines) and shaded ± 2 standard errors over 10 independent training runs. (a) rRMSE to the optimal regularized value. (b) Initial-state value under the dual-iterate policy. (c) rRMSE to the optimal (unregularized) value and comparison to the generator-based benchmark *SPD-dMDP* (Chen and Wang, 2016). (d) KL-divergence between the learned policy π_{ρ_k} and the optimal regularized policy $\pi_r^\star$.

6. Concluding Reflections

We proposed a new LP-based primal-dual RL algorithm for entropy-regularized MDPs, suing a single on-policy trajectory generated by a behavioral policy that may evolve with the dual iterate. Using the dynamical-systems approach to stochastic approximation (Borkar, 2023), we establish almost sure last-iterate convergence of the primal and dual iterates to the optimal regularized value function and policy, respectively, under an asynchronous two-timescale scheme with diminishing stepsizes. To the best of our knowledge, such almost sure guarantees have not previously been established for LP-based methods in the entropy-regularized setting.

Under a strengthened ergodicity condition, we additionally derive a finite-time last-iterate guarantee, showing that the expected dual objective gap, and consequently the dual mean-squared error, decays at rate $\widetilde{\mathcal{O}}(k^{-2/3})$. This matches the best-known last-iterate behavior for two-timescale stochastic approximation under Markovian sampling (Zeng et al., 2024), despite replay-buffer bias and asynchronous, visitation-adjusted updates along a single trajectory.

Our finite-time analysis relies on a strengthened ergodicity condition, namely the existence of a deterministic ergodic policy, and a high-probability visitation event ensuring sufficient exploration along the trajectory. A natural future direction is to weaken these requirements and to derive last-iterate rates under milder mixing conditions, potentially by combining refined concentration tools for inhomogeneous Markov chains with adaptive exploration schedules. Moreover, while our guarantees are stated in terms of the dual objective gap (and the induced mean-square error), it would be valuable to connect primalreplay-buffer biasdual convergence more directly to control-centric criteria such as (regularized) return suboptimality and constraint-violation measures, thereby enabling sharper comparisons with alternative performance guarantees in online RL.

While this work focuses on the tabular setting, extending PGDA-RL to function approximation remains a pressing challenge. One promising route is to combine our primal-dual viewpoint and stochastic approximation analysis with LP-based frameworks under linear function approximation, as studied for offline settings by Gabbianelli et al. (2024), and to develop corresponding finite-time guarantees under Markovian single-trajectory data. Such extensions will likely require reformulating the asynchronous update scheme and the structured experience replay mechanism to accommodate approximation error and stability issues. Finally, it is an interesting open question to integrate acceleration ideas for convexified regularized MDP formulations, which have been shown to speed up convergence under full model knowledge (Li et al., 2024), into the present Markovian, replay-based setting.

Acknowledgments

The authors were supported by the German Research Foundation (DFG) under the grant "SU 1433/2-1".

Appendix A. Synchronous Almost Sure Convergence: Proof Details

This section contains the proofs of Propositions 7 and 8 that are at the heart of the almost sure convergence proof of Algorithm 1.

Proof of Proposition 7 The true gradients of the regularized Lagrangian (9) are given by

$$\nabla_V L(V,\rho)(s') = \eta_V V(s') + \tilde{\rho}(s') + \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\mathcal{P}(s'|s,a), \quad s' \in \mathcal{S} \tag{32a}$$

$$\nabla_\rho L(V,\rho)(s,a) = -V(s) + r(s,a) + \gamma \sum_{s'\in\mathcal{S}} V\left(s'\right)\mathcal{P}(s'|s,a) - \eta_\rho \log\left(\frac{\rho(s,a)}{\tilde{\rho}(s)}\right), \tag{32b}$$

where $(s,a) \in \mathcal{S} \times \mathcal{A}$.

The Lagrangian's gradient with respect to V, (32a), is linear in V and therefore Lipschitz continuous. Continuously differentiable functions are Lipschitz continuous on compact sets. Hence, the projection of the dual variable to the compact hyper-rectangle H ensures the Lipschitz continuity of the gradient (32b) and property (A1) The noise terms resulting from the stochastic gradients (11) with the iterates $\rho_{k-1},\ V_{k-1}$ inserted at iteration k are

$$M_k^{(1)}(s') = \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a)\left(\mathbf{1}\{\tilde{s}(s,a) = s'\} - \mathcal{P}(s'|s,a)\right), \quad s' \in \mathcal{S} \tag{33a}$$

$$M_k^{(2)}(s,a) = \gamma \left(V_{k-1}\left(\tilde{s}(s,a)\right) - \sum_{s'\in\mathcal{S}} V_{k-1}(s')\mathcal{P}(s'|s,a) \right),\ (s,a) \in \mathcal{S}\times\mathcal{A}. \tag{33b}$$

Recall the definition of the filtration $\mathcal{F}_k := \sigma(V_\ell, \rho_\ell, M_\ell^{(1)}, M_\ell^{(2)}, \ell \le k),\ k \ge 0$. The martingale difference properties of the noise terms with respect to the filtration $\{\mathcal{F}_k\}$ are now a direct consequence of the access to the generative model to obtain the state transitions $\tilde{s}(s,a) \sim \mathcal{P}(\cdot|s,a)$ for all state-action pairs in each update step.

We proceed with the bounds on the second moment and begin with $M_k^{(1)}$

$$\begin{aligned}
\mathbb{E}\left[\|M_k^{(1)}\|_2^2|\mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\sum_{s'\in\mathcal{S}} (M_k^{(1)}(s'))^2|\mathcal{F}_{k-1}\right] \\
&= \gamma^2 \sum_{s'\in\mathcal{S}} \mathbb{E}\left[\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a)\left(\mathbf{1}\{s' = \tilde{s}(s,a)\} - \mathcal{P}(s'|s,a)\right)\right)^2 |\mathcal{F}_{k-1}\right] \\
&\le \gamma^2|\mathcal{S}|\mathbb{E}\left[\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a)\right)^2 |\mathcal{F}_{k-1}\right] \\
&= \gamma^2|\mathcal{S}|\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a)\right)^2 \\
&= \gamma^2|\mathcal{S}|\|\rho_{k-1}\|_1^2 \ \le\ \gamma^2|\mathcal{S}|^2|\mathcal{A}|\|\rho_{k-1}\|_2^2.
\end{aligned}$$

The first inequality holds since $\mathbf{1}\{s' = \tilde{s}(s,a)\} - \mathcal{P}(s'|s,a) \in [-1,1]$. Now consider $M_k^{(2)}$

$$\begin{aligned}
\mathbb{E}\left[\|M_k^{(2)}\|_2^2|\mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} (M_k^{(2)}(s,a))^2|\mathcal{F}_{k-1}\right] \\
&= \gamma^2\, \mathbb{E}\left[\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left(V_{k-1}(\tilde{s}(s,a)) - \sum_{s'\in\mathcal{S}} V_{k-1}(s')\mathcal{P}(s'|s,a)\right)^2 |\mathcal{F}_{k-1}\right] \\
&\leq \gamma^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{E}\left[V_{k-1}(\tilde{s}(s,a))^2|\mathcal{F}_{k-1}\right] \\
&= \gamma^2 \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)V_{k-1}(s')^2 \\
&\leq \gamma^2|\mathcal{S}||\mathcal{A}|\|V_{k-1}\|_2^2,
\end{aligned}$$

which completes the proof of property (A2). ■

Proof of Proposition 8 We start with the proof of the global asymptotic stability property (A3) for the primal-dual gradient flow.

Let $\rho \in H$ be fixed. Recall that due to the strict convex-concave structure of the regularized Lagrangian, there exists a unique saddle-point $(V^\star, \rho^\star)$. With the first-order optimality condition based on (32a), the best response $\lambda(\rho)$ is

$$\nabla_V L(V(t),\rho)(s') = \eta_V V(t)(s') - \tilde{\rho}(s') + \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\mathcal{P}(s'|s,a) \overset{!}{=} 0$$

$$\Leftrightarrow \lambda(\rho)(s') = \frac{1}{\eta_V}\left(\tilde{\rho}(s') - \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\mathcal{P}(s'|s,a)\right).$$

Since $\tilde{\rho}(s') = \sum_{a\in\mathcal{A}} \rho(s',a)$, the mapping λ is linear in ρ, hence Lipschitz continuous.

Next, we establish the global asymptotic stability by providing a strict Lyapunov function, $\varphi_\rho : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}_+$. For fixed ρ set $\varphi_\rho(V(t)) = \frac{1}{2}\|V(t) - \lambda(\rho)\|_2^2$. Then,

$$\begin{aligned}
\dot{\varphi}_\rho(V(t)) &= \dot{V}(t)^\top(V(t) - \lambda(\rho)) \\
&= \nabla_V L(V(t),\rho)^\top(\lambda(\rho) - V(t)) \\
&\leqslant L(\lambda(\rho),\rho) - L(V(t),\rho) \leq 0 \quad \text{equality only for } V(t) = \lambda(\rho),
\end{aligned}$$

where the first inequality is due to the convexity and the last inequality is due to the optimality of $\lambda(\rho)$. Due to the strong convexity, the last inequality is strict for any $V(t) \neq \lambda(\rho)$.

We next address the property (A4). The projected dynamic system (13b) has to be well-posed to make claims about its stability. The Lipschitz continuity of the gradient $\nabla_\rho L(\lambda(\cdot),\cdot)$ follows from the linearity of the $\lambda(\rho)$ and the Lipschitz continuity of $\nabla_\rho L(V,\cdot)$ on the compact set H. To clarify the properties of the projection term in (13b), we present the derivation of the projected dynamic following Dupuis (1987). Define the projection of vector $v \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ at $\rho \in H$ as

$$\pi(\rho, v) := \lim_{\delta\searrow 0} \frac{\Pi_H[\rho + \delta v] - \rho}{\delta}.$$

The projected dynamics of ρ (13b) are then given by

$$\dot{\rho} = \pi(\rho, v), \quad \text{with} \quad v = \nabla_\rho L(\lambda(\rho), \rho). \tag{34}$$

Let $N(\rho)$ be the set-valued function that gives the outward normals at $\rho \in \partial H$, that is, $N(\rho) = \{\zeta \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|} \mid \forall \rho' \in H : \langle \zeta, \rho - \rho' \rangle \geq 0, \|\zeta\| = 1\}$. Then, by (Dupuis, 1987, Lemma 4.6), the system (34) equals

$$\pi(\rho, v) = v + (\langle v, -\zeta(\rho, v)\rangle \vee 0)\zeta(\rho, v), \tag{35}$$

with $\zeta(\rho, v) \in \arg\max_{\zeta \in -N(\rho)} \langle v, -\zeta \rangle$.

The existence and uniqueness of a solution to the projected dynamical system are guaranteed by Dupuis and Nagurney (1993, Theorem 2 (uniqueness), Theorem 3 (existence)). Note that due to the discontinuity of the dynamics, a solution is defined as an absolutely continuous function that fulfills the ODE save on a set of Lebesgue measure zero. To prove the global asymptotic stability of the projected ODE (13b), we next provide a Lyapunov argument based on (35).

Due to the first order optimality condition, $\rho^\star$ is an equilibrium of the well-defined solution to the projected dynamic system (13b). The function

$$\phi(\rho(t)) := L(V^\star, \rho^\star) - L(\lambda(\rho(t)), \rho(t)) \tag{36}$$

serves as a strict Lyapunov function. The inequality $\phi(\rho(t)) \geq 0$ holds since $L(\lambda(\rho), \rho) = \min_{V \in \mathbb{R}^{|\mathcal{S}|}} L(V, \rho) \leq \max_{\rho \in H} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} L(V, \rho) = L(V^\star, \rho^\star)$. The equality only holds for $\rho = \rho^\star$ due to the unique saddle-point as established in Corollary 2. Define $\Gamma(\rho(t)) := \nabla_\rho L(\lambda(\rho(t)), \rho(t))$ and with a slight abuse of notation $\zeta(\rho(t)) := \zeta(\rho(t), \Gamma(\rho(t)))$. Then,

$$\begin{aligned}
\dot{\phi}(\rho(t)) &= -\Gamma(\rho(t))^\top \dot{\rho}(t) \\
&= -\Gamma(\rho(t))^\top (\Gamma(\rho(t)) + (\langle \Gamma(\rho(t)), -\zeta(\rho(t))\rangle \vee 0)\zeta(\rho(t))) \\
&= -\|\Gamma(\rho(t))\|^2 + (\langle \Gamma(\rho(t)), -\zeta(\rho(t))\rangle \vee 0)\langle \Gamma(\rho(t)), \zeta(\rho(t))\rangle \\
&= -\|\Gamma(\rho(t))\|^2 - (\langle \Gamma(\rho(t)), -\zeta(\rho(t))\rangle \vee 0)\langle \Gamma(\rho(t)), -\zeta(\rho(t))\rangle \leq 0,
\end{aligned}$$

where the last inequality is strict except for the equilibrium point $\rho^\star$ that satisfies the first-order optimality condition. In the second equality, we apply the decomposition (35). Note that $\rho^\star \in H$. Hence, the first-order optimality condition also holds for the projected system.

Lastly, we address the stability of the iterate sequence, property (A5). The dual variable is constrained to the compact set H. We, therefore, only have to ensure that $\sup_k \|V_k\| < \infty$ almost surely. We apply the scaling limit method outlined in Borkar and Meyn (2000) and Borkar (2023, Theorem 4.1), which states that the global asymptotic stability of the origin for a certain scaling limit ODE is a sufficient condition for the almost sure boundedness of the iterates. The scaling limit for the value iterates for a fixed occupancy ρ is defined as

$$h_\infty(V(t)) := \lim_{c \to \infty} -\frac{1}{c} \nabla_V L(cV(t), \rho).$$

As can be seen from the gradient (32a), the scaling limit exists for all $V \in \mathbb{R}^{|\mathcal{S}|}$ and is given by

$$h_\infty(V(t)) = -\eta_V V(t),$$

with global asymptotic equilibrium at the origin for $\eta_V > 0$. This shows the almost sure boundedness of the value iterates. Note that the stability of the value iterates is achieved through the added convex regularization and does not hold with $\eta_V = 0$ due to the lack of curvature in the unregularized linear program and its Lagrangian formulation (4). ■

Appendix B. Asynchronous Almost Sure Convergence: Proof Details

This section contains the proofs of the asynchronous algorithm's properties listed in Proposition 12. The proofs are presented in a series of lemmas. For (B4) the proof of the synchronous convergence analysis still applies. The two options regarding the behavioral policy update only need to be addressed separately for property (B5).

Lemma 21 *The noise components and bias terms of the gradient estimators* (14) *satisfy property (B1). The conditional expectation of the buffer-based transition probability estimator is*

$$\mathbb{E}[\widehat{\mathcal{P}}_k(s'|s,a)|\mathcal{F}_k^-] = \frac{\nu_{k-1}(s,a)}{\nu_k(s,a)}\mathcal{P}_{\mathcal{D}_{k-1}}(s'|s,a) + \frac{\nu_k(s,a) - \nu_{k-1}(s,a)}{\nu_k(s,a)}\mathcal{P}(s'|s,a). \tag{37}$$

Proof We first derive the conditional expectation of the buffer-based transition estimator. Fix $(s,a) \in \mathcal{D}_{\text{inc}}(s_k)$, such that at iteration k there is a draw from the corresponding buffer (otherwise $\widehat{\mathcal{P}}_k(\cdot|s,a) \equiv 0$ by convention and the claims are trivial). Recall that $\xi_k(s,a)$ is drawn uniformly from the list $\mathcal{D}_k(s,a)$ and $\widehat{\mathcal{P}}_k(s'|s,a) = \mathbf{1}\{\xi_k(s,a) = s'\}$. If $(s,a) = X_k = (s_{k-1}, a_{k-1})$, then $\mathcal{D}_k(s,a) = \mathcal{D}_{k-1}(s,a) \cup \{s_k\}$ and $\nu_k(s,a) = \nu_{k-1}(s,a) + 1$. Conditionally on $\mathcal{F}_k^- = \sigma(\mathcal{F}_{k-1}, s_k, a_k)$, the multiset $\mathcal{D}_k(s,a)$ is fixed and the uniform draw satisfies

$$\mathbb{P}\big(\xi_k(s,a) \in \mathcal{D}_{k-1}(s,a)|\mathcal{F}_k^-\big) = \frac{\nu_{k-1}(s,a)}{\nu_k(s,a)}, \qquad \mathbb{P}\big(\xi_k(s,a) = s_k|\mathcal{F}_k^-\big) = \frac{1}{\nu_k(s,a)}.$$

Moreover, conditional on $\mathcal{F}_k^-$ and on the event $\{\xi_k(s,a) \in \mathcal{D}_{k-1}(s,a)\}$, the law of $\xi_k(s,a)$ is the empirical distribution $\mathcal{P}_{\mathcal{D}_{k-1}}(\cdot|s,a)$. Therefore,

$$\mathbb{E}[\widehat{\mathcal{P}}_k(s'|s,a)|\mathcal{F}_k^-] = \frac{\nu_{k-1}(s,a)}{\nu_k(s,a)}\,\mathcal{P}_{\mathcal{D}_{k-1}}(s'|s,a) + \frac{1}{\nu_k(s,a)}\,\mathcal{P}(s'|s,a).$$

If $(s,a) \neq X_k$, then the buffer cell is not updated, that is, $\mathcal{D}_k(s,a) = \mathcal{D}_{k-1}(s,a)$ and $\nu_k(s,a) = \nu_{k-1}(s,a)$, hence $\mathbb{E}[\widehat{\mathcal{P}}_k(s'|s,a)|\mathcal{F}_k^-] = \mathcal{P}_{\mathcal{D}_{k-1}}(s'|s,a)$. Combining both cases yields (37).

Based on this explicit conditional expectation, we turn to the noise and bias analysis. The noise and bias sequences of the gradient estimators are defined in equations (15), (16). By construction $\mathbb{E}[M_k^{(1)}|\mathcal{F}_k^-] = 0$. Since $\mathcal{F}_{k-1} \subset \mathcal{F}_k^-$, using the tower-property gives $\mathbb{E}[M_k^{(1)}|\mathcal{F}_{k-1}] = \mathbb{E}\Big[\mathbb{E}[M_k^{(1)}|\mathcal{F}_k^-]|\mathcal{F}_{k-1}\Big] = 0$. For $M_k^{(2)}$, we use $s_k \sim \mathcal{P}(\cdot|X_k)$ and $X_k, V_{k-1} \in \mathcal{F}_{k-1}$ to get $\mathbb{E}[M_k^{(2)}|\mathcal{F}_{k-1}] = 0$. This shows that the noise sequences $\{M_k^{(1)}\}, \{M_k^{(2)}\}$ are martingale difference sequences with respect to $\{\mathcal{F}_k\}$ and (B1)-a holds.

The bounds on the second moments of the martingale differences, (B1)-b, can be derived similarly to the proof of Proposition 7 upon noting that $\|\widehat{\mathcal{P}}_k(\cdot|s,a)\|_2 \le 1$ and $\|\mathbb{E}[\widehat{\mathcal{P}}_k(\cdot|s,a)|\mathcal{F}_k^-]\|_2 \le 1$. Recall

$$\mathcal{E}_k = \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho_{k-1}(s,a)\Big(\mathbb{E}\big[\widehat{\mathcal{P}}_k(\cdot|s,a)|\mathcal{F}_k^-\big] - \mathcal{P}(\cdot|s,a)\Big).$$

Fix $(s,a) \in \mathcal{X}$ and $s' \in \mathcal{S}$. In case that $(s,a) \in \mathcal{D}_{\text{inc}}(s_k)$, the estimator $\widehat{\mathcal{P}}_k(s'|s,a) = \mathbf{1}\{\xi_k(s,a) = s'\}$ is a one-hot vector of a sample $\xi_k(s,a)$ drawn uniformly from the list $\mathcal{D}_k(s,a)$. Hence,

$$\mathbb{E}\big[\widehat{\mathcal{P}}_k(s'|s,a)|\mathcal{F}_k^-\big] = \mathcal{P}_{\mathcal{D}_k}(s'|s,a).$$

By convention, if $D_k(s,a)$ is not drawn from at time k, then $\widehat{\mathcal{P}}_k(\cdot|s,a) \equiv 0$ and the same identity holds with $\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) \equiv 0$.

Now note that $\mathcal{D}_k(s,a)$ consists of the subsequence of next-states observed upon visits to (s,a), that is,

$$\mathcal{D}_k(s,a) = \{\, s_\ell \,:\, 1 \le \ell \le k,\ X_\ell = (s,a)\,\}, \qquad |\mathcal{D}_k(s,a)| = \nu_k(s,a).$$

Conditional on $\{X_\ell = (s,a)\}$, the next state satisfies $s_\ell \sim \mathcal{P}(\cdot|s,a)$. Under (B5) we have $\nu_k(s,a) \to \infty$ a.s. for all $(s,a) \in \mathcal{X}$, and thus by the strong law of large numbers,

$$\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) \xrightarrow[k\to\infty]{\text{a.s.}} \mathcal{P}(\cdot|s,a) \quad \text{for each fixed } (s,a) \in \mathcal{X}.$$

Since H is compact and $\rho_{k-1} \in H$ for all k, the weights $\rho_{k-1}(s,a)$ are uniformly bounded. Therefore, dominated convergence yields

$$\big\|\mathcal{E}_k\big\| \le \gamma \sum_{(s,a)\in\mathcal{X}} \big|\rho_{k-1}(s,a)\big|\, \big\|\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\big\| \xrightarrow[k\to\infty]{\text{a.s.}} 0.$$

Moreover, boundedness of ρ_{k-1} and $\|\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a)\|_2 \le 1$ imply that $\{\mathcal{E}_k\}$ is bounded uniformly. This proves (B1)-c.

■

Lemma 22 *The non-autonomous ODEs* (18) *satisfy properties (B2) and (B3).*

Proof The properties (B2) and (B3) regarding the equilibria of the dynamical systems, hold for the asynchronous projected stochastic ascent-descent scheme because we can relate the asymptotic behavior of any solution to the non-autonomous rescaled ODEs (18) to the limiting ODEs of the synchronous setting (13) following (Borkar, 2023, Section 6.4).

To see that (B2) and (B3) hold, we specify strict Lyapunov functions for the ODEs (13) that keep their Lyapunov function properties for the non-autonomous rescaled ODEs. We show the proof only for (B2) as (B3) follows with the same arguments.

Let $\rho \in H$ be a fixed arbitrary dual variable. Let $\epsilon > 0$ and $\Lambda^{(1)}(\cdot)$ be a measurable function with $\Lambda^{(1)}(t) \in \Omega^{\epsilon}_{|\mathcal{S}|}$ for all t. By definition, $\Lambda^{(1)}(t)$ is a diagonal matrix for each t with bounded diagonal entries $w_i(t) \in [\epsilon, 1]$. The measurability and boundedness of $t \mapsto \Lambda^{(1)}(t)$ together with the Lipschitz continuity of $V \mapsto \nabla_V L(V,\rho)$ imply that the vector field $-\Lambda^{(1)}(t)\nabla_V L(V,\rho)$ is measurable in t and locally Lipschitz in V. Hence, for every initial condition, the ODE (18a) admits a (unique) Carathéodory solution; see, e.g., standard results in differential equations (Filippov, 2013).

Define the Lyapunov function

$$\Phi_\rho(V) := L(V,\rho) - L(\lambda(\rho),\rho),$$

where $\lambda(\rho)$ is the unique minimizer of $V \mapsto L(V,\rho)$ (unique by strong convexity). Along any solution $V(\cdot)$ of (18a), using $\nabla_V \Phi_\rho(V) = \nabla_V L(V,\rho)$ we have

$$\begin{aligned}\frac{d}{dt}\Phi_\rho(V(t)) &= \nabla_V L(V(t),\rho)^\top \dot{V}(t) = -\nabla_V L(V(t),\rho)^\top \Lambda^{(1)}(t)\nabla_V L(V(t),\rho) \\ &= -\sum_{s\in\mathcal{S}} w_s(t)\big(\nabla_V L(V(t),\rho)(s)\big)^2 \le -\epsilon\|\nabla_V L(V(t),\rho)\|_2^2 \;\le\; 0.\end{aligned}$$

Moreover, equality holds if and only if $\nabla_V L(V(t),\rho) = 0$, that is, $V(t) = \lambda(\rho)$. By η_V-strong convexity of $V \mapsto L(V,\rho)$, the Polyak-Łojasiewicz inequality holds: $\|\nabla_V L(V,\rho)\|_2^2 \ge 2\eta_V\, \Phi_\rho(V)$ for all V. Therefore,

$$\frac{d}{dt}\Phi_\rho(V(t)) \;\le\; -2\epsilon\eta_V\, \Phi_\rho(V(t)),$$

which implies $\Phi_\rho(V(t)) \le e^{-2\epsilon\eta_V t}\Phi_\rho(V(0))$ and hence $V(t) \to \lambda(\rho)$ as $t \to \infty$. Thus $\lambda(\rho)$ is a globally asymptotically stable equilibrium for (18a), uniformly over measurable $\Lambda^{(1)}(\cdot) \in \Omega^\epsilon_{|\mathcal{S}|}$.

For (B3) the same arguments hold with the unchanged Lyapunov function used in the proof of condition (A4), see (36), since multiplying the vector field by a diagonal $\Lambda^{(2)}(t) \in \Omega^\epsilon_{|\mathcal{X}|}$ preserves negativity of the Lyapunov derivative. ∎

Lemma 23 *Under Assumption 9, the update component selections in Algorithm 2 fulfill property (B5).*

Proof Let Assumption 9 hold. In Algorithm 2, conditional on the history $\mathcal{F}_k$, the next state-action pair (s_{k+1}, a_{k+1}) is generated by first sampling $s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)$ and then sampling $a_{k+1} \sim \pi_k^b(\cdot|s_{k+1})$, where the behavioral policy π_k^b is either fixed or (in the on-policy exploration variant) a deterministic function of ρ_k via (3), optionally ϵ_k-mixed with the uniform policy. Therefore, for any $(s,a),(s',a') \in \mathcal{X}$,

$$\mathbb{P}\left((s_{k+1}, a_{k+1}) = (s',a')|\mathcal{F}_k\right) = \mathcal{P}(s'|s_k,a_k)\,\pi_k^b(a'|s').$$

In particular, in the on-policy exploration case where $\pi_k^b = \pi^{\rho_k}$ (or its ϵ_k-mixture), this can be written as

$$\mathbb{P}\left((s_{k+1}, a_{k+1}) = (s',a')|\mathcal{F}_k\right) = \mathcal{Q}_{(s_k,a_k),(s',a')}(\rho_k),$$

which shows (B5)-a. By Assumption 9, for any strictly exploratory policy, the induced Markov chain on $\mathcal{X}$ is irreducible and aperiodic. Since $\rho \in H$ implies a uniform lower bound on action probabilities (and ϵ_k-mixing preserves strict exploration), the policies π_ρ^b are strictly exploratory uniformly over $\rho \in H$. Hence, $\mathcal{Q}(\rho)$ defines an irreducible and aperiodic Markov chain on $\mathcal{X}$ for all $\rho \in H$ and (B5)-b holds.

Recall the definition of H with constants $0 < C^L \le C^U$ such that $C^L < \rho(s,a) < C^U$ for all (s,a) and all $\rho \in H$. Then $\sum_{b\in\mathcal{A}} \rho(s',b) \ge |\mathcal{A}|C^L$ for every $s' \in \mathcal{S}$, and for all $\rho_1, \rho_2 \in H$ and $a \in \mathcal{A}$,

$$\begin{aligned}
|\pi_{\rho_1}(a|s') - \pi_{\rho_2}(a|s')| &= \left| \frac{\rho_1(s',a)}{\sum_{a'\in\mathcal{A}} \rho_1(s',a')} - \frac{\rho_2(s',a)}{\sum_{a'\in\mathcal{A}} \rho_2(s',a')} \right| \\
&\le \left| \frac{\rho_1(s',a) - \rho_2(s',a)}{\sum_{a'\in\mathcal{A}} \rho_1(s',a')} \right| + \left| \rho_2(s',a) \frac{\sum_{a'\in\mathcal{A}}(\rho_2(s',a') - \rho_1(s',a'))}{\sum_{a'\in\mathcal{A}} \rho_1(s',a') \sum_{a'\in\mathcal{A}} \rho_2(s',a')} \right| \\
&\le \frac{1}{|\mathcal{A}|C^L} |\rho_1(s',a) - \rho_2(s',a)| + \frac{C^U}{|\mathcal{A}|^2 (C^L)^2} \sum_{a'\in\mathcal{A}} |\rho_1(s',a') - \rho_2(s',a')| \\
&\le C_{\mathcal{Q}} \sum_{a'\in\mathcal{A}} |\rho_1(s',a') - \rho_2(s',a')|,
\end{aligned}$$

with $C_{\mathcal{Q}} := \frac{1}{|\mathcal{A}|C^L} + \frac{C^U}{|\mathcal{A}|^2(C^L)^2}$, where the last inequality uses

$$|\rho_1(s',a) - \rho_2(s',a)| \le \sum_{a'} |\rho_1(s',a') - \rho_2(s',a')|.$$

Therefore, for any $(s,a), (s',a') \in \mathcal{X}$,

$$\begin{aligned}
\left| \mathcal{Q}_{(s,a),(s',a')}(\rho_1) - \mathcal{Q}_{(s,a),(s',a')}(\rho_2) \right| &= \mathcal{P}(s'|s,a) \left| \pi^{\rho_1}(a'|s') - \pi^{\rho_2}(a'|s') \right| \\
&\le C_{\mathcal{Q}} \sum_{b\in\mathcal{A}} \left| \rho_1(s',b) - \rho_2(s',b) \right|.
\end{aligned}$$

Since the bound holds uniformly for all $(s,a), (s',a')$, the kernel mapping $\rho \mapsto \mathcal{Q}(\rho)$ is Lipschitz (for example in the max-entry norm), and hence in any equivalent norm on the finite-dimensional space of kernels. This shows (B5)-c. ■

Appendix C. Asynchronous Convergence Rate: Proof Details

This section contains the proof details of the convergence rate theorem, Theorem 16. The first section addresses Remark 14 and Proposition 15, followed by the proofs of Propositions 18 and 19 and finally some additional technical lemmata collected in the third subsection.

C.1 Proofs Related to Assumption 13 and Proposition 15

We first address the sufficient condition for the uniform geometric ergodicity with common constants of the family of Markov chains induced by the dual iterate.

Lemma 24 (Deterministic mixing policy implies Assumption 13) *Assume that there exists a deterministic state-feedback policy $\pi_d : \mathcal{S} \to \mathcal{A}$ such that the resulting Markov chain on $\mathcal{S}$ with transition kernel*

$$\mathcal{P}_d(s'|s) := \mathcal{P}\big(s'|s, \pi_d(s)\big)$$

is uniformly geometrically ergodic. Then Assumption 13(i) and (ii) hold. Specifically, let $M \in \mathbb{N}$, $\epsilon_0 \in (0, 1]$, *and a probability measure* υ *on* $\mathcal{S}$ *be such that*

$$\mathcal{P}_d^M(\cdot|s) \geq \epsilon_0 \, \upsilon(\cdot), \qquad \forall s \in \mathcal{S}.$$

Further define

$$\pi_{\min} := \frac{C^L}{|\mathcal{A}| \, C^U}, \qquad \epsilon := \pi_{\min}^M \epsilon_0, \qquad \varrho := (1-\epsilon)^{1/M}.$$

Then the following hold.

(a) For every $\rho \in H$*, the state-action kernel* $\mathcal{P}_\rho$ *has a unique stationary distribution* μ_ρ*, and*

$$\sup_{x \in \mathcal{X}} \|\delta_x \mathcal{P}_\rho^k - \mu_\rho\|_{\mathrm{TV}} \leq C_{\mathcal{X}} \varrho^k, \qquad k \geq 0,$$

with the common constant $C_{\mathcal{X}} := \varrho^{-M}$.

(b) For every sequence $\{\rho_t\}_{t \in \mathbb{N}} \subset H$*,*

$$\sup_{k \geq 0} \delta\left(\prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t} \right) \leq \kappa, \qquad m_\star := M+1, \quad \kappa := 1-\epsilon.$$

Proof By the equivalence of uniform geometric ergodicity and the multi-step Doeblin condition, see Meyn and Tweedie (2012, Theorem 16.0.2), there exist $M \in \mathbb{N}$, $\epsilon_0 \in (0, 1]$, and a probability measure υ on $\mathcal{S}$ such that

$$\mathcal{P}_d^M(\cdot|s) \geq \epsilon_0 \, \upsilon(\cdot), \qquad \forall s \in \mathcal{S}.$$

For $\rho \in H$, let π_ρ denote the dual-induced policy (3) and define the corresponding state-transition kernel

$$\tilde{\mathcal{P}}_\rho(s'|s) := \sum_{a \in \mathcal{A}} \pi_\rho(a|s) \mathcal{P}(s'|s, a).$$

Since $C^L \leq \rho(s, a) \leq C^U$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have the uniform lower bound

$$\pi_\rho(a|s) = \frac{\rho(s, a)}{\sum_{b \in \mathcal{A}} \rho(s, b)} \geq \frac{C^L}{|\mathcal{A}| \, C^U} = \pi_{\min}, \qquad \forall \rho \in H, \ s \in \mathcal{S}, \ a \in \mathcal{A}.$$

Hence, for every $\rho \in H$ and $s, s' \in \mathcal{S}$,

$$\begin{aligned} \tilde{\mathcal{P}}_\rho(s'|s) &= \sum_{a \in \mathcal{A}} \pi_\rho(a|s) \mathcal{P}(s'|s, a) \\ &\geq \pi_\rho(\pi_d(s)|s) \, \mathcal{P}\big(s'|s, \pi_d(s)\big) \\ &\geq \pi_{\min} \, \mathcal{P}_d(s'|s). \end{aligned}$$

By monotonicity under composition, iterating the above inequality M times yields

$$\tilde{\mathcal{P}}_\rho^M(\cdot|s) \geq \pi_{\min}^M \, \mathcal{P}_d^M(\cdot|s) \geq \pi_{\min}^M \epsilon_0 \, \upsilon(\cdot) = \epsilon \, \upsilon(\cdot), \qquad \forall \rho \in H, \ s \in \mathcal{S}.$$

Therefore, for every $\rho \in H$, the kernel $\tilde{\mathcal{P}}_\rho^M$ satisfies the same Doeblin minorization, and therefore

$$\delta(\tilde{\mathcal{P}}_\rho^M) \le 1-\epsilon, \qquad \forall \rho \in H.$$

We first use this to obtain uniform geometric ergodicity of the frozen state kernels. Let $k = mM + r$ with $m \in \mathbb{N}_0$ and $0 \le r < M$. For any probability measures η, η' on $\mathcal{S}$, by submultiplicativity of Dobrushin's coefficient and the bound $\delta(Q) \le 1$ for any Markov kernel Q,

$$\|\eta\tilde{\mathcal{P}}_\rho^k - \eta'\tilde{\mathcal{P}}_\rho^k\|_{\mathrm{TV}} \le (1-\epsilon)^m \|\eta - \eta'\|_{\mathrm{TV}} \le \varrho^{-(M-1)}\varrho^k \|\eta - \eta'\|_{\mathrm{TV}}.$$

In particular, each $\tilde{\mathcal{P}}_\rho$ has a unique stationary law $\tilde{\mu}_\rho$ on $\mathcal{S}$, and

$$\sup_{s\in\mathcal{S}} \|\delta_s \tilde{\mathcal{P}}_\rho^k - \tilde{\mu}_\rho\|_{\mathrm{TV}} \le \varrho^{-(M-1)}\varrho^k, \qquad k \ge 0.$$

We now lift this to the state-action chain on $\mathcal{X} = \mathcal{S}\times\mathcal{A}$. Define the kernels

$$\mathcal{T}(s'|s,a) := \mathcal{P}(s'|s,a), \qquad \mathcal{B}_\rho\big(s,a \mid s'\big) := \mathbf{1}_{\{s=s'\}}\pi_\rho(a|s).$$

Then

$$\mathcal{P}_\rho = \mathcal{T}\mathcal{B}_\rho, \qquad \tilde{\mathcal{P}}_\rho = \mathcal{B}_\rho\mathcal{T}.$$

Therefore, if $\tilde{\mu}_\rho$ is stationary for $\tilde{\mathcal{P}}_\rho$, then

$$\mu_\rho := \tilde{\mu}_\rho \mathcal{B}_\rho$$

is stationary for $\mathcal{P}_\rho$, that is,

$$\mu_\rho(s,a) = \tilde{\mu}_\rho(s)\pi_\rho(a|s).$$

Moreover, for any $x \in \mathcal{X}$ and any $k \ge 1$,

$$\delta_x \mathcal{P}_\rho^k = (\delta_x\mathcal{T})\tilde{\mathcal{P}}_\rho^{k-1}\mathcal{B}_\rho.$$

Using non-expansiveness of total variation under Markov kernels, we obtain

$$\begin{aligned}\|\delta_x\mathcal{P}_\rho^k - \mu_\rho\|_{\mathrm{TV}} &= \|(\delta_x\mathcal{T})\tilde{\mathcal{P}}_\rho^{k-1}\mathcal{B}_\rho - \tilde{\mu}_\rho\mathcal{B}_\rho\|_{\mathrm{TV}}\\ &\le \|(\delta_x\mathcal{T})\tilde{\mathcal{P}}_\rho^{k-1} - \tilde{\mu}_\rho\|_{\mathrm{TV}}\\ &\le \varrho^{-(M-1)}\varrho^{k-1} = \varrho^{-M}\varrho^k.\end{aligned}$$

Since $\|\delta_x - \mu_\rho\|_{\mathrm{TV}} \le 1 \le \varrho^{-M}$ for $k=0$, this proves assertion (a) with $C_{\mathcal{X}} := \varrho^{-M}$.

It remains to prove assetion (b). Fix an arbitrary sequence $\{\rho_t\}_{t\in\mathbb{N}} \subset H$ and a block starting at time i. By the pointwise bound $\tilde{\mathcal{P}}_{\rho_t} \ge \pi_{\min}\mathcal{P}_d$ derived above, monotonicity under composition gives

$$\prod_{t=i}^{i+M-1} \tilde{\mathcal{P}}_{\rho_t}(\cdot|s) \ge \pi_{\min}^M \mathcal{P}_d^M(\cdot|s) \ge \epsilon\, \upsilon(\cdot), \qquad \forall s \in \mathcal{S}.$$

Hence,

$$\delta\left(\prod_{t=i}^{i+M-1} \tilde{\mathcal{P}}_{\rho_t}\right) \le 1-\epsilon.$$

Now using again the factorization of the state-action kernels, we have

$$\prod_{t=i}^{i+M} \mathcal{P}_{\rho_t} = \mathcal{T}\left(\prod_{t=i}^{i+M-1} \tilde{\mathcal{P}}_{\rho_t}\right)\mathcal{B}_{\rho_{i+M}}.$$

Therefore, for any probability measures μ, μ' on $\mathcal{X}$,

$$\begin{aligned}\left\|\mu \prod_{t=i}^{i+M} \mathcal{P}_{\rho_t} - \mu' \prod_{t=i}^{i+M} \mathcal{P}_{\rho_t}\right\|_{\mathrm{TV}} &\leq \left\|(\mu\mathcal{T}) \prod_{t=i}^{i+M-1} \tilde{\mathcal{P}}_{\rho_t} - (\mu'\mathcal{T}) \prod_{t=i}^{i+M-1} \tilde{\mathcal{P}}_{\rho_t}\right\|_{\mathrm{TV}} \\ &\leq (1-\epsilon)\|\mu\mathcal{T} - \mu'\mathcal{T}\|_{\mathrm{TV}} \\ &\leq (1-\epsilon)\|\mu - \mu'\|_{\mathrm{TV}}.\end{aligned}$$

Taking the supremum over μ, μ' yields

$$\delta\left(\prod_{t=i}^{i+M} \mathcal{P}_{\rho_t}\right) \leq 1-\epsilon.$$

Since i was arbitrary, Assumption 13(ii) holds with

$$m_\star = M+1, \qquad \kappa = 1-\epsilon.$$

This completes the proof. ■

Next, we provide the proofs of the convergence rate-related properties of the iterates generated by Algorithm 2 with primal variable projection to $\mathcal{V}_r$, summarized in Proposition 15. The visitation floor and replay-buffer bias decay results are of particular interest since they are the main differences of our asynchronous approach compared to Zeng et al. (2024).

Throughout, assume that the exploration of the state-action space in Algorithm 2 is performed based on the dual-induced strictly exploratory policies. The alternatives of the fixed exploratory behavioral policy and ϵ-exploration follow along the same arguments but with simpler constants and are left out.

We start with the visitation floor property, stated in Proposition 15(i).

Lemma 25 *The following hold*

(i) There exists $p_\star \in (0,1)$ such that for every fixed $\rho \in H$,

$$\mu_\rho(s,a) \geq p_\star, \qquad \forall (s,a) \in \mathcal{X}.$$

(ii) For every $\delta \in (0,1)$, there exists a deterministic burn-in time $K(\delta) < \infty$ such that

$$\mathbb{P}\left(\underline{\nu}_k \geq \frac{p_\star}{2}k \quad \text{for all } k \geq K(\delta)\right) \geq 1-\delta.$$

(iii) The envelopes $\alpha(\underline{\nu}_k)$, $\beta(\underline{\nu}_k)$ are non-increasing in k and for all $k \geq K(\delta)$,

$$\bar{\alpha}_k \leq \alpha(\underline{\nu}_k) \leq \alpha(\lfloor \tfrac{p_\star}{2}k \rfloor), \qquad \bar{\beta}_k \leq \beta(\underline{\nu}_k) \leq \beta(\lfloor \tfrac{p_\star}{2}k \rfloor)$$

with probability of at least $1-\delta$.

Proof (i) Since $\rho \in H$ implies $C^L \leq \rho(s,a) \leq C^U$, we have for each s, $\pi_\rho(a|s) = \rho(s,a)/\tilde{\rho}(s) \geq C^L/(|\mathcal{A}|C^U)$. Under Assumption 13, every state transition kernel corresponding to a policy π_ρ, induces a corresponding unique stationary law $\tilde{\mu}_\rho$ with all components strictly positive. The compactness of H and continuity of $\rho \mapsto \pi_\rho$ and $\pi_\rho \mapsto \tilde{\mu}_\rho$ yield the existence of a uniform lower bound $c_S := \inf_{\rho \in H} \min_{s \in \mathcal{S}} \tilde{\mu}_\rho(s) > 0$. Thus $\mu_\rho(s,a) = \tilde{\mu}_\rho(s)\pi_\rho(a|s) \geq c_S \cdot C^L/(|\mathcal{A}|C^U) =: p_\star$.

(ii) To show the high-probability linear growth bound for all state-action pairs, we first establish lower bounds for the time-varying expected visits, then apply a concentration bound for inhomogeneous Markov chains by Paulin (2015) to establish concentration for any time step k, and lastly apply a union bound to achieve the high-probability minimal growth bound for all steps $k \geq K(\delta)$.

We begin by deriving a block contraction of the time-inhomogeneous kernel sequence with Dobrushin's ergodic coefficient.

By (Brémaud, 2020, Corollary 4.3.17) the coefficient satisfies the contraction inequality $\|\mu Q - \nu Q\|_{\mathrm{TV}} \leq \delta(Q)\|\mu - \nu\|_{\mathrm{TV}}$ for any two probability measures μ, ν on $\mathcal{X}$.

Let $x_0 \in \mathcal{X}$ denote the initial state-action pair. Write $p_k := \delta_{x_0} \prod_{t=0}^{k-1} \mathcal{P}_{\rho_t}$ with the time-inhomogeneous kernel sequence $\{\mathcal{P}_{\rho_k}\}$, and let

$$d_k := \|p_k - \mu_{\rho_k}\|_{\mathrm{TV}}.$$

Under Assumption 13, we can choose constants $m_\star \in \mathbb{N}$ and $\kappa \in (0,1)$ such that

$$\sup_{k \geq 0} \delta\left(\prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t}\right) \leq \kappa.$$

Then, by the contraction inequality, for any probability measure q on $\mathcal{X}$,

$$\left\|(q - \mu_\rho) \prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t}\right\|_{\mathrm{TV}} \leq \delta\left(\prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t}\right) \|q - \mu_\rho\|_{\mathrm{TV}} \leq \kappa \|q - \mu_\rho\|_{\mathrm{TV}}.$$

Using this inhomogeneous Markov chain block contraction and the triangle inequality, for all $k \geq 0$,

$$\begin{aligned} d_{k+m_\star} &= \left\|p_k \prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t} - \mu_{\rho_{k+m_\star}}\right\|_{\mathrm{TV}} \\ &\leq \left\|(p_k - \mu_{\rho_k}) \prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t}\right\|_{\mathrm{TV}} + \left\|\mu_{\rho_k} \prod_{t=k}^{k+m_\star-1} \mathcal{P}_{\rho_t} - \mu_{\rho_{k+m_\star}}\right\|_{\mathrm{TV}} \\ &\leq \kappa d_k + \sum_{t=k}^{k+m_\star-1} \|\mu_{\rho_t} - \mu_{\rho_{t+1}}\|_{\mathrm{TV}}. \end{aligned} \tag{38}$$

The last inequality holds due to the following telescoping argument. Insert and subtract the intermediate measures and use the triangle inequality:

$$\begin{aligned}\left\|\mu_{\rho_k}\prod_{t=k}^{k+m_\star-1}\mathcal{P}_{\rho_t}-\mu_{\rho_{k+m_\star}}\right\|_{\mathrm{TV}} &\le \sum_{t=k}^{k+m_\star-1}\left\|\mu_{\rho_t}\prod_{j=t}^{k+m_\star-1}\mathcal{P}_{\rho_j}-\mu_{\rho_{t+1}}\prod_{j=t+1}^{k+m_\star-1}\mathcal{P}_{\rho_j}\right\|_{\mathrm{TV}}\\ &= \sum_{t=k}^{k+m_\star-1}\left\|(\mu_{\rho_t}\mathcal{P}_{\rho_t}-\mu_{\rho_{t+1}})\prod_{j=t+1}^{k+m_\star-1}\mathcal{P}_{\rho_j}\right\|_{\mathrm{TV}},\end{aligned}$$

where the empty product is taken as the identity. Using the stationary law property of $\mu_{\rho_t}\mathcal{P}_{\rho_t}=\mu_{\rho_t}$ and the non-expansivity of Markov kernels, each summand is bounded by the corresponding term $\|\mu_{\rho_t}-\mu_{\rho_{t+1}}\|$ and (38) follows.

The stationary distributions are Lipschitz continuous in the dual variable (see Lemma 32), so there exists $L_\mu>0$ such that $\|\mu_{\rho_t}-\mu_{\rho_{t+1}}\|_{\mathrm{TV}} \le L_\mu\|\rho_{t+1}-\rho_t\|_2$. Therefore,

$$\sum_{t=k}^{k+m_\star-1}\|\mu_{\rho_t}-\mu_{\rho_{t+1}}\|_{\mathrm{TV}}\le L_\mu\sum_{t=k}^{k+m_\star-1}\|\rho_{t+1}-\rho_t\|_2.$$

Since

$$\mathbb{E}[\nu_k(s,a)] \;=\; \sum_{t=0}^{k-1}p_t(s,a), \tag{39}$$

and by definition $p_k(s,a)\ge\mu_{\rho_k}(s,a)-d_k$, we next aim to bound $\sum_{t=0}^{k-1}d_t$ from above. For this, we partition the timeline into $m_\star$-residual classes. For each $\ell\in\{0,\dots,m_\star-1\}$, define

$$g_n^{(\ell)}:=d_{\ell+nm_\star},\qquad S_n^{(\ell)}:=\sum_{i=\ell+nm_\star}^{\ell+(n+1)m_\star-1}\|\rho_{i+1}-\rho_i\|_2,\qquad n=0,\dots,N_\ell,$$

where $N_\ell:=\lfloor(k-1-\ell)/m_\star\rfloor$. Note that the blocks form a partition of $\{0,\dots,k-2\}$, hence $\sum_{\ell=0}^{m^\star-1}\sum_{n=0}^{N_\ell}S_n^{(\ell)}=\sum_{t=0}^{k-2}\|\rho_{t+1}-\rho_t\|_2$ and $\sum_{\ell=0}^{m_\star-1}\sum_{n=0}^{N_\ell}g_n^{(\ell)}=\sum_{t=0}^{k-1}d_t$.

Then (38) implies

$$g_{n+1}^{(\ell)}\;\le\;\kappa\,g_n^{(\ell)}\;+\;L_\mu\,S_n^{(\ell)}.$$

Summing over $n=0,\dots,N_\ell-1$ gives

$$\sum_{n=0}^{N_\ell-1}g_{n+1}^{(\ell)}\;\le\;\kappa\sum_{n=0}^{N_\ell-1}g_n^{(\ell)}\;+\;L_\mu\sum_{n=0}^{N_\ell-1}S_n^{(\ell)}.$$

Since

$$\sum_{n=0}^{N_\ell-1}g_{n+1}^{(\ell)}-\kappa\sum_{n=0}^{N_\ell-1}g_n^{(\ell)}=(1-\kappa)\sum_{n=0}^{N_\ell-1}g_{n+1}^{(\ell)}\;+\;\kappa\big(g_{N_\ell}^{(\ell)}-g_0^{(\ell)}\big),$$

dropping the nonnegative term $\kappa g_{N_\ell}^{(\ell)}$ yields

$$(1-\kappa)\sum_{n=0}^{N_\ell-1}g_{n+1}^{(\ell)}\;\le\;\kappa\,g_0^{(\ell)}\;+\;L_\mu\sum_{n=0}^{N_\ell-1}S_n^{(\ell)}.$$

Dividing by $(1-\kappa)$ and adding $g_0^{(\ell)}$ to both sides,

$$\sum_{n=0}^{N_\ell} g_n^{(\ell)} \;\le\; \Big(1+\frac{\kappa}{1-\kappa}\Big) g_0^{(\ell)} \;+\; \frac{L_\mu}{1-\kappa}\sum_{n=0}^{N_\ell-1} S_n^{(\ell)} \;=\; \frac{1}{1-\kappa}\, g_0^{(\ell)} \;+\; \frac{L_\mu}{1-\kappa}\sum_{n=0}^{N_\ell-1} S_n^{(\ell)}.$$

Summing this bound over $\ell = 0, \dots, m_\star - 1$ yields

$$\sum_{t=0}^{k-1} d_t \;\le\; \frac{1}{1-\kappa}\sum_{\ell=0}^{m_\star-1} d_\ell \;+\; \frac{L_\mu}{1-\kappa}\sum_{t=0}^{k-2}\|\rho_{t+1}-\rho_t\|_2.$$

Since $d_\ell \le 1$ for all ℓ, $\sum_{\ell=0}^{m_\star-1} d_\ell \le m_\star$, which gives the final bound

$$\sum_{t=0}^{k-1} d_t \;\le\; \frac{m_\star}{1-\kappa} \;+\; \frac{L_\mu}{1-\kappa}\sum_{t=0}^{k-2}\|\rho_{t+1}-\rho_t\|_2.$$

Inserting this bound into (39), for any fixed $(s,a) \in \mathcal{X}$,

$$\mathbb{E}[\nu_k(s,a)] \;=\; \sum_{t=0}^{k-1} p_t(s,a) \;\ge\; \sum_{t=0}^{k-1}\big(\mu_{\rho_t}(s,a) - d_t\big) \;\ge\; k\,p_\star \;-\; \frac{m_\star}{1-\kappa} \;-\; \frac{L_\mu}{1-\kappa}\sum_{i=0}^{k-2}\|\rho_{i+1}-\rho_i\|_2.$$

By the 1-Lipschitz property of projection in ℓ_2, $\|\rho_{t+1}-\rho_t\|_2 \;\le\; \beta(\nu_t(s_t,a_t))\|\hat{h}_t\|$, where $\hat{h}_t$ is the single-coordinate gradient from (14b). Due to the projection step, the iterates V_t are bounded, and there is a constant $B < \infty$ with $\|\hat{h}_t\| \le B$ for all $t \ge 0$; hence

$$\|\rho_{t+1}-\rho_t\|_2 \;\le\; B\,\beta(\nu_t(s_t,a_t)).$$

Note that $\sum_{(s,a)\in\mathcal{X}} \nu_k(s,a) = k$. The specific choice of the base stepsize schedule $\beta_k = \beta_0/(k+1)$ gives the following bound on the total step length up to iteration k

$$\sum_{t=0}^{k-1}\beta(\nu_t(s_t,a_t)) = \sum_{(s,a)\in\mathcal{X}}\sum_{t=0}^{\nu_k(s,a)-1}\frac{\beta_0}{t+1} \;\le\; \beta_0\,|\mathcal{X}|\sum_{t=0}^{k-1}\frac{1}{t+1}.$$

Using $\int_t^{t+1}\frac{dx}{x} \ge \frac{1}{t+1}$ for $t \ge 1$, we get

$$\sum_{t=0}^{k-1}\frac{1}{t+1} = 1 + \sum_{t=1}^{k-1}\frac{1}{t+1} \le 1 + \int_1^k \frac{dx}{x} = 1 + \log(k),$$

and therefore,

$$\sum_{t=0}^{k-1}\beta(\nu_t(s_t,a_t)) \le \beta_0\,|\mathcal{X}|(1+\log k).$$

Hence, for any $\epsilon \in (0, p_\star)$ there exists $K_0(\epsilon)$ such that for all $k \ge K_0(\epsilon)$,

$$\mathbb{E}\Big[\frac{\nu_k(s,a)}{k}\Big] \;\ge\; p_\star - \frac{\epsilon}{2}. \tag{40}$$

Fix $(s, a) \in \mathcal{X}$. We next establish the concentration lower bound for any $k \in \mathbb{N}$. Denote the path of the process $\{X_k\}$ up to iteration k by $x_{0:k-1} \in \mathcal{X}^k$. Define the normalized count function for the fixed (s, a) pair by

$$\varphi : \mathcal{X}^k \to \mathbb{R},\ x_{0:k-1} \mapsto \varphi(x_{0:k-1}) := \frac{1}{k}\sum_{t=0}^{k-1} \mathbf{1}\{x_t = (s, a)\}.$$

Then, $\varphi(x_{0:k-1}) = \frac{\nu_k(s,a)}{k}$. Let $y_{0:k-1} \in \mathcal{X}^k$ denote a second path. It can be seen that

$$\varphi(x_{0:k-1}) - \varphi(y_{0:k-1}) \le \sum_{t=0}^{k-1} c_t \mathbf{1}\{x_t \neq y_t\}, \quad \text{with } c_t = \frac{1}{k},\ \sum_{t=0}^{k-1} c_t^2 = \frac{1}{k}$$

and we can apply the one-sided McDiarmid inequality for time-inhomogeneous Markov chains (Paulin, 2015, Corollary 2.10) to φ and obtain

$$\mathbb{P}\Big(\frac{\nu_k(s,a)}{k} - \mathbb{E}\big[\frac{\nu_k(s,a)}{k}\big] \le -\eta\Big) \ \le\ \exp\Big(-\frac{2k\,\eta^2}{\tau_{\min}}\Big), \qquad \tau_{\min} := \inf_{0\le\zeta<1} \tau_{\mathrm{inh}}(\zeta)\Big(\tfrac{2-\zeta}{1-\zeta}\Big)^2, \tag{41a}$$

where

$$\tau_{\mathrm{inh}}(\zeta) := \min\{\ell \in \mathbb{N} : \sup_{n\in\mathbb{N}} \sup_{x,y\in\mathcal{X}} \|\delta_x \prod_{t=n}^{\ell} \mathcal{P}_{\rho_t} - \delta_y \prod_{t=n}^{\ell} \mathcal{P}_{\rho_t}\|_{\mathrm{TV}} \le \zeta\}, \tag{41b}$$

see (Paulin, 2015, Definition 1.4), is the inhomogeneous mixing time for $\{\mathcal{P}_{\rho_k}\}$. It holds that $\tau_{\min} < \infty$ with a simple bound in our setting deferred to Lemma 31. From the mean lower bound in (40), there exists $K_0\big(\frac{p_\star}{2}\big)$ such that, for all $k \ge K_0\big(\frac{p_\star}{2}\big)$,

$$\mathbb{E}\left[\tfrac{\nu_k(s,a)}{k}\right] \ \ge\ \tfrac{3}{4}p_\star.$$

Hence, on the event $\{\frac{\nu_k(s,a)}{k} \le \frac{p_\star}{2}\}$ we have a downward deviation of at least $\mathbb{E}[\frac{\nu_k(s,a)}{k}] - \frac{p_\star}{2} \ \ge \frac{p_\star}{4}$, and applying (41a) yields

$$\mathbb{P}\left(\tfrac{\nu_k(s,a)}{k} \le \tfrac{p_\star}{2}\right) \ \le\ \exp\Big(-\frac{k\ p_\star^2}{8\,\tau_{\min}}\Big), \qquad k \ \ge\ K_0\Big(\tfrac{p_\star}{2}\Big). \tag{42}$$

Union bound over $(s, a) \in \mathcal{X}$ and sum the geometric tail over $k \ge K$:

$$\mathbb{P}\Big(\exists k \ge K : \min_{(s,a)\in\mathcal{X}} \tfrac{\nu_k(s,a)}{k} < p_\star - \epsilon\Big) \le |\mathcal{X}| \sum_{k=K}^{\infty} e^{-ck} = \frac{|\mathcal{X}|}{1 - e^{-c}}\, e^{-cK}, \quad c = \frac{p_\star^2}{8\,\tau_{\min}}.$$

Thus, for any $\delta \in (0, 1)$, taking

$$K(\delta)\ :=\ \max\left\{K_0\Big(\tfrac{p_\star}{2}\Big),\ \left\lceil \frac{1}{c}\,\log\frac{|\mathcal{X}|}{(1-e^{-c})\,\delta} \right\rceil\right\}, \tag{43}$$

ensures $\mathbb{P}\big(\forall k \ge K(\delta) :\ \underline{\nu}_k/k \ge p_\star/2\big)\ \ge\ 1 - \delta$. which completes the proof of (ii).

(iii) On the minimum-visitation event in (ii), for all $k \ge K(\delta)$ and all $(s, a) \in \mathcal{X}$ we have $\nu_k(s, a) \ge (p_\star/2)\,k$ and $\tilde{\nu}_k(s) \ge (p_\star/2)\,k$. Since $\alpha(\cdot)$ and $\beta(\cdot)$ are nonincreasing,

$$\alpha\big(\tilde{\nu}_k(s)\big) \;\le\; \alpha\left(\left\lfloor \tfrac{p_\star}{2}\, k \right\rfloor\right), \qquad \beta\big(\nu_k(s,a)\big) \;\le\; \beta\left(\left\lfloor \tfrac{p_\star}{2}\, k \right\rfloor\right).$$

■

Next, we prove the replay-buffer bias decay stated in Proposition 15(ii). In the following, we use the shorthand notation $\bar{\mathcal{E}}_k(\rho)$ for the stationary law expectation of the bias vector given a fixed $\rho \in H$, that is,

$$\bar{\mathcal{E}}_k(\rho) := \mathbb{E}_{s\sim\tilde{\mu}_\rho}[e_s\, \mathcal{E}_k(\rho)(s)]. \tag{44}$$

Lemma 26 (Replay-buffer bias under high-probability visitation) *Fix $\delta \in (0,1)$, let $\mathcal{G}_\delta$ denote the high-probability event of Lemma 25, and let $K(\delta)$ be the corresponding burn-in time. Define*

$$\Lambda(\delta) \;:=\; \sqrt{\log\Big(\frac{(2^{|\mathcal{S}|}-2)\,|\mathcal{X}|}{\delta}\Big)}.$$

Then on $\mathcal{G}_\delta$ and for all $k \ge K(\delta)$,

$$\big\|\bar{\mathcal{E}}_k(\rho)\big\|_\infty \;\le\; C_{\mathrm{buf}}(\delta)\frac{1}{\sqrt{k}}, \quad C_{\mathrm{buf}}(\delta) := \frac{3\gamma|\mathcal{X}|C^U\Lambda(\delta)}{\sqrt{p_\star}}. \tag{45}$$

In particular, $\|\bar{\mathcal{E}}_k(\rho)\|_\infty = O\big(\sqrt{1/k}\big)$ on $\mathcal{G}_\delta$, uniformly in ρ, V.

Proof For each fixed $(s,a) \in \mathcal{X}$ and corresponding buffer size $|\mathcal{D}_k(s,a)| = \nu_k(s,a)$, the L^1-bound for a categorical law on $|\mathcal{S}|$ outcomes (Weissman et al., 2003, Theorem 2.1) gives for all $\epsilon > 0$

$$\mathbb{P}\big(\big\|\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\big\|_1 \ge \epsilon\big) \;\le\; (2^{|\mathcal{S}|}-2)\,\exp\Big(-\tfrac{\nu_k(s,a)\epsilon^2}{2}\Big).$$

Condition on the σ-field generated by the realized counts $\{\nu_k(s,a)\}_{(s,a)\in\mathcal{X}}$. To union bound over the $|\mathcal{X}|$ pairs, we choose

$$\epsilon_k \;:=\; \sqrt{\frac{2}{\underline{\nu}_k}\;\log\Big((2^{|\mathcal{S}|}-2)\,|\mathcal{X}|\,/\delta\Big)},$$

Then

$$\mathbb{P}\Big(\bigcup_{(s,a)\in\mathcal{X}} \{\|\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\|_1 \ge \epsilon_k\} \;\Big|\; \sigma(\{\nu_k(s,a)\}_{(s,a)\in\mathcal{X}})\Big) \;\le\; \delta.$$

Therefore, with probability at least $1-\delta$, simultaneously for all $(s,a) \in \mathcal{X}$,

$$\big\|\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\big\|_1 \;\le\; \sqrt{\frac{2}{\underline{\nu}_k}\;\log\Big((2^{|\mathcal{S}|}-2)\,|\mathcal{X}|/\delta\Big)} \;\le\; \sqrt{\frac{2}{\underline{\nu}_k}}\,\Lambda(\delta).$$

On the high-probability visitation event $\mathcal{G}_\delta$ of Lemma 25, we have $\underline{\nu}_k \ge \frac{p_\star}{2}\,k$ for all $k \ge K(\delta)$, hence

$$\max_{(s,a)\in\mathcal{X}} \;\big\|\mathcal{P}_{\mathcal{D}_k}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\big\|_1 \;\le\; \frac{2\Lambda(\delta)}{\sqrt{p_\star\,k}}. \tag{46}$$

Now we pass from kernel error to the buffer-bias term. The stationary law expectations can be written as the Hadamard product between the buffer bias term defined in (16) and the strictly positive stationary state and state-action distributions induced by the dual policy, that is,

$$\bar{\mathcal{E}}_k(\rho)(s') = \tilde{\mu}_\rho(s')\,\gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\,\tfrac{\nu_{k-1}(s,a)}{\nu_k(s,a)}\left(\mathcal{P}_{\mathcal{D}_{k-1}}(s'|s,a) - \mathcal{P}(s'|s,a)\right).$$

With the trivial bound $\|\tilde{\mu}_\rho\|_\infty \leq 1$, and the increasing number of visits $\nu_k(s,a) \geq \nu_{k-1}(s,a)$, we can omit the preconditioning with the stationary distribution in the following norm bounds and find for $\bar{\mathcal{E}}_k(\rho)$,

$$\begin{aligned}\left\|\bar{\mathcal{E}}_k(\rho)\right\|_\infty \leq& \gamma \left\| \sum_{(s,a)\in\mathcal{X}} \rho(s,a)\big(\mathcal{P}_{\mathcal{D}_{k-1}}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\big)\right\|_\infty \\ \leq& \gamma\|\rho\|_1 \max_{(s,a)\in\mathcal{X}} \left\|\mathcal{P}_{\mathcal{D}_{k-1}}(\cdot|s,a) - \mathcal{P}(\cdot|s,a)\right\|_1 \leq \frac{3\gamma|\mathcal{X}|C^U\Lambda(\delta)}{\sqrt{p_\star}}\frac{1}{\sqrt{k}},\end{aligned}$$

since $\|\rho\|_1 \leq |\mathcal{X}|C^U$ by the restriction to $\rho \in H$ and $\frac{2}{\sqrt{k-1}} \leq \frac{2\sqrt{2}}{\sqrt{k}} \leq \frac{3}{\sqrt{k}}$ for all $k \geq 2$. ■

We next address the regularity properties stated in Proposition 15(iii). The η_V-strong convexity of the mapping $V \mapsto L(V,\rho)$ has been established already in equations (32) and the proof of Proposition 8. The explicit Lipschitz constant of the best response mapping λ is derived in the following.

Lemma 27 *The best response map $\lambda : H \to \mathbb{R}^{|\mathcal{S}|}$ is affine and hence globally Lipschitz on H. In particular,*

$$\lambda(\rho)_{s'} = \frac{1}{\eta_V}\big(\tilde{\rho}(s') - \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho(s,a)\mathcal{P}(s'|s,a)\big), \tag{47}$$

and for all ρ, $\hat{\rho} \in H$,

$$\|\lambda(\rho) - \lambda(\hat{\rho})\|_2 \leq \frac{\sqrt{|\mathcal{S}||\mathcal{A}|(1+\gamma^2)}}{\eta_V}\|\rho - \hat{\rho}\|_2.$$

Proof The displayed linear form (47) of the optimizer follows from the first-order optimality condition derived in the proof of Proposition 8. For the Lipschitz constant, we define the matrix $M \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}|}$ with rows

$$\mathrm{row}_{sa}(M) = e_s^\top - \gamma\mathcal{P}(\cdot|s,a)^\top, \text{ so that } \lambda(\rho) = \frac{1}{\eta_V}M^\top\rho.$$

Then $\|\lambda(\rho) - \lambda(\hat{\rho})\|_2 = \frac{1}{\eta_V}\|M^\top(\rho-\hat{\rho})\|_2 \leq \frac{\|M^\top\|_2}{\eta_V}\|\rho-\hat{\rho}\|_2$. Each row of M satisfies $\|\mathrm{row}_{sa}(M)\|_2^2 \leq \|e_s^\top\|_2^2 + \gamma^2\|\mathcal{P}(\cdot|s,a)\|_2^2 \leq (1+\gamma^2)$. Therefore, the claim follows by noting $\|M^\top\|_2^2 = \|M\|_2^2 \leq \|M\|_F^2 = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\mathrm{row}_{sa}(M)\|_2^2 \leq |\mathcal{S}||\mathcal{A}|(1+\gamma^2)$. ■

We prove the L-smoothness of the reduced objective by showing the Lipschitz property for $\nabla f(\rho)$. Upon inserting the best response formula (47) in the Lagrangian, we obtain

$$f(\rho) = L(\lambda(\rho),\rho) = -\frac{1}{2\eta_V}\|M^\top\rho\|_2^2 + \langle\rho, r\rangle + \eta_\rho g(\rho),$$

since $\Delta[V](s,a) = (-MV)(s,a) + r(s,a)$ and hence, $L(V,\rho) = \frac{\eta_V}{2}\|V\|_2^2 - \langle \rho, MV\rangle + \langle \rho, r\rangle + \eta_\rho g(\rho)$. Then $\nabla f(\rho) = -\frac{MM^\top}{\eta_V}\rho + r + \eta_\rho \nabla g(\rho)$ and we obtain the following smoothness for the reduced objective: For all $\rho, \hat\rho \in H$

$$\|\nabla f(\rho) - \nabla f(\hat\rho)\|_2 \;\leq \Big(\eta_\rho \frac{2}{C^L} + \frac{\|M\|_2^2}{\eta_V}\Big)\|\rho - \hat\rho\|_2 \leq \Big(\eta_\rho \frac{2}{C^L} + \frac{|\mathcal{S}||\mathcal{A}|(1+\gamma^2)}{\eta_V}\Big)\|\rho - \hat\rho\|_2,$$

where $\frac{2}{C^L}$ is the Lipschitz constant of $\nabla g(\rho)$ on H as derived in the following Lemma, and where we have used the upper bound for $\|M\|_2^2$ from the proof of Lemma 27.

Lemma 28 (Lipschitz gradient of g) *For the entropy regularization term $g(\rho)$, defined in* (8)*, it holds that, for all $\rho, \hat\rho \in H$,*

$$\|\nabla g(\rho) - \nabla g(\hat\rho)\|_2 \;\leq\; \frac{2}{C^L}\,\|\rho - \hat\rho\|_2.$$

Proof The coordinate-wise gradient is

$$\nabla g(\rho)(s,a) = -\log\Big(\frac{\rho(s,a)}{\tilde\rho(s)}\Big) = \log\tilde\rho(s) - \log\rho(s,a).$$

The Hessian is block-diagonal across states. For a fixed s and $a, a' \in \mathcal{A}$,

$$\big[\nabla^2 g(\rho)\big]_{(s,a),(s,a')} = \frac{1}{\tilde\rho(s)} - \frac{\mathbf{1}\{a = a'\}}{\rho(s,a)}.$$

Hence the $|\mathcal{A}| \times |\mathcal{A}|$ block at state s is

$$H_s(\rho) \;=\; -\operatorname{diag}\Big(\tfrac{1}{\rho(s,a)}\Big)_{a\in\mathcal{A}} \;+\; \frac{1}{\tilde\rho(s)}\,\mathbf{1}\mathbf{1}^\top.$$

Using the triangle inequality for operator norms and the facts $\|\operatorname{diag}(u)\|_2 = \|u\|_\infty$, $\|\mathbf{1}\mathbf{1}^\top\|_2 = |\mathcal{A}|$, and $\tilde\rho(s) \geq |\mathcal{A}|C^L$ for $\rho \in H$, we get

$$\|H_s(\rho)\|_2 \leq \max_a \frac{1}{\rho(s,a)} + \frac{1}{\tilde\rho(s)}|\mathcal{A}| \leq \frac{1}{C^L} + \frac{|\mathcal{A}|}{|\mathcal{A}|C^L} = \frac{2}{C^L}.$$

Since $\nabla^2 g(\rho)$ is block-diagonal with blocks $H_s(\rho)$, it follows that

$$\sup_\rho \|\nabla^2 g(\rho)\|_2 \;=\; \sup_{s,\rho}\|H_s(\rho)\|_2 \;\leq\; \frac{2}{C^L}.$$

Finally,

$$\|\nabla g(\rho) - \nabla g(\hat\rho)\|_2 \leq \Big(\sup_z \|\nabla^2 g(z)\|_2\Big)\,\|\rho - \hat\rho\|_2 \leq \frac{2}{C^L}\,\|\rho - \hat\rho\|_2.$$

■

The following Lemma states the strong concavity of the reduced objective on H.

Lemma 29 (Strong concavity of the reduced objective) *The reduced objective $f(\rho) = \min_V L(V, \rho)$ is μ-strongly concave on H with any modulus $0 < \mu \le \mu_{\text{opt}}$, where*

$$\mu_{\text{opt}} = \frac{1}{4}\Big[(A+B+C) - \sqrt{(A+B+C)^2 - 4AC}\Big] \; > \; 0, \tag{48}$$

with

$$A := \frac{\eta_\rho}{C^U}, \qquad B := \frac{|\mathcal{S}||\mathcal{A}|(1+\gamma^2)}{\eta_V}, \qquad C := \frac{(1-\gamma)^2\,|\mathcal{A}|\,{C^L}^2}{\eta_V\,{C^U}^2}.$$

Proof As shown previously, it holds that

$$\nabla f(\rho) = -\tfrac{1}{\eta_V} M M^\top \rho + r + \eta_\rho \nabla g(\rho), \qquad \nabla^2 f(\rho) = -\tfrac{1}{\eta_V} M M^\top + \eta_\rho \nabla^2 g(\rho).$$

Since g is concave, $\nabla^2 g(\rho) \preceq 0$, and together with the negative semidefinite term $-MM^\top$ this implies that f is concave. For any $h \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,

$$-h^\top \nabla^2 f(\rho) h = \frac{1}{\eta_V}\|M^\top h\|_2^2 - \eta_\rho\; h^\top \nabla^2 g(\rho) h. \tag{49}$$

For any $h \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, decompose $h = u + v$ per state, where for each $s \in \mathcal{S}$, $\sum_{a\in\mathcal{A}} u(s,a) = 0$ and $v(s,a) = q(s)\,\frac{\rho(s,a)}{\bar\rho(s)}$ for some $q(s) \in \mathbb{R}$. With this decomposition, u redistributes mass across actions while keeping the state marginal fixed, and v changes only the state marginals. In particular, $v^\top \nabla^2 g(\rho) v = 0$. For any u with $\sum_{a\in\mathcal{A}} u(s,a) = 0$, it holds that for all $s \in \mathcal{S}$ and corresponding block matrices $H_s(\rho)$ in the Hessian of g, specified in the proof of Lemma 28,

$$u(s,\cdot)^\top H_s(\rho) u(s,\cdot) = \sum_{a\in\mathcal{A}} \tfrac{u(s,a)^2}{\rho(s,a)} \ge \tfrac{1}{C^U} \sum_{a\in\mathcal{A}} u(s,a)^2.$$

Since the Hessian of g is block diagonal, it follows that uniformly on H,

$$u^\top\big(-\eta_\rho \nabla^2 g(\rho)\big) u = \sum_{s\in\mathcal{S}} \big(u(s,\cdot)^\top H_s(\rho) u(s,\cdot)\big) \ge \tfrac{\eta_\rho}{C^U}\|u\|_2^2 = A\|u\|_2^2. \tag{50}$$

It holds that $\|M^\top u\|_2 \le \|M\|_2 \|u\|_2$, and together with the operator bound $\|M\|_2^2 \le |\mathcal{S}||\mathcal{A}|(1+\gamma^2)$ derived in the proof of Lemma 27, we have

$$\|M^\top u\|_2^2 \le |\mathcal{S}||\mathcal{A}|(1+\gamma^2)\|u\|_2^2. \tag{51}$$

Let again $\tilde{\mathcal{P}}_\rho(s'|s) = \sum_{a\in\mathcal{A}} \pi_\rho(a|s)\mathcal{P}(s'|s,a)$ denote the state transition kernel induced by the dual policy. And with some abuse of notation let $\bar P_\rho \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the corresponding state transition matrix with $[\bar P_\rho]_{ss'} = \tilde{\mathcal{P}}_\rho(s'|s)$. It holds that

$$M^\top v = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} v(s,a)(e_s - \gamma \mathcal{P}(\cdot|s,a)) = \sum_{s\in\mathcal{S}} \left(\sum_{a\in\mathcal{A}} v(s,a)\right) e_s - \gamma \sum_{s\in\mathcal{S}, a\in\mathcal{A}} v(s,a)\mathcal{P}(\cdot|s,a)$$

Since $\sum_{a\in\mathcal{A}} v(s,a) = q(s)$ and

$$\sum_{a\in\mathcal{A}} v(s,a)\mathcal{P}(\cdot|s,a) = q(s)\sum_{a\in\mathcal{A}} \pi_\rho(a|s)\mathcal{P}(\cdot|s,a) = q(s)\bar{\mathcal{P}}_\rho(\cdot|s),$$

we have

$$M^\top v = (I - \gamma \bar{P}_\rho^\top)\ q.$$

Because $\bar{P}_\rho^\top$ is column-stochastic, $\|\bar{P}_\rho^\top q\|_1 \leq \|q\|_1$ and consequently, $\|M^\top v\|_1 \geq (1-\gamma)\|q\|_1 \geq (1-\gamma)\|q\|_2$. Using $\|x\|_2 \geq \|x\|_1/\sqrt{|\mathcal{S}|}$ for $x \in \mathbb{R}^{|\mathcal{S}|}$,

$$\|M^\top v\|_2^2 \geq \frac{1}{|\mathcal{S}|}\,\|M^\top v\|_1^2 \geq \frac{(1-\gamma)^2}{|\mathcal{S}|}\|q\|_2^2.$$

Moreover,

$$\|v\|_2^2 = \sum_{s\in\mathcal{S}} q(s)^2 \sum_{a\in\mathcal{A}} \Big(\frac{\rho(s,a)}{\tilde{\rho}(s)}\Big)^2 \leq \frac{C^{U^2}}{|\mathcal{A}|C^{L^2}}\|c\|_2{}^2,$$

because $\tilde{\rho}(s) \geq |\mathcal{A}|C^L$ and $\rho(s,a) \leq C^U$. Together,

$$\|M^\top v\|_2{}^2 \geq \frac{(1-\gamma)^2|\mathcal{A}|}{|\mathcal{S}|}\frac{C^{L^2}}{C^{U^2}}\|v\|_2{}^2. \tag{52}$$

From Equation (49) we have

$$-h^\top \nabla^2 f(\rho) h = \frac{1}{\eta_V}\|M^\top(u+v)\|_2^2 - \eta_\rho\, u^\top \nabla^2 g(\rho)\, u.$$

For any $\varepsilon \in (0,1)$, Young's inequality gives

$$\|a+b\|_2^2 \geq (1-\varepsilon)\|a\|_2^2 - \Big(\frac{1}{\varepsilon}-1\Big)\|b\|_2^2,$$

so with $a = M^\top v$, $b = M^\top u$,

$$-h^\top \nabla^2 f(\rho) h \ \geq\ \frac{1}{\eta_V}\Big[(1-\varepsilon)\|M^\top v\|_2^2 - \Big(\tfrac{1}{\varepsilon}-1\Big)\|M^\top u\|_2^2\Big] + \frac{\eta_\rho}{C^U}\|u\|_2^2.$$

Substituting the operator bounds above yields

$$-h^\top \nabla^2 f(\rho) h \ \geq\ \Big[A - B\Big(\tfrac{1}{\varepsilon}-1\Big)\Big]\|u\|_2^2 \ +\ C(1-\varepsilon)\,\|v\|_2^2.$$

Moreover, for all decompositions $h = u + v$ we have $\|h\|_2^2 \leq 2(\|u\|_2^2 + \|v\|_2^2)$, hence

$$-h^\top \nabla^2 f(\rho) h \ \geq\ 2\,\mu(\varepsilon)\,\|h\|_2^2,$$

where for every $\varepsilon \in (0,1)$,

$$\mu(\varepsilon) = \frac{1}{2}\min\Big\{A - B\Big(\tfrac{1}{\varepsilon}-1\Big),\ C(1-\varepsilon)\Big\}.$$

Maximizing over $\varepsilon \in (0,1)$ amounts to maximizing $\mu(\varepsilon)$, which is achieved by equating the two arguments of the minimum. Solving

$$A - B\Big(\tfrac{1}{\varepsilon}-1\Big) \ =\ C(1-\varepsilon)$$

for ε and inserting back gives the optimal modulus

$$\mu_{\text{opt}} = \max_{0<\varepsilon<1} \mu(\varepsilon) = \frac{1}{4}\Big[(A+B+C) - \sqrt{(A+B+C)^2 - 4AC}\Big].$$

Since $A > 0$ and $C > 0$ by construction, the discriminant $(A+B+C)^2 - 4AC$ is strictly smaller than $(A+B+C)^2$, so the bracket is positive and hence $\mu_{\text{opt}} > 0$. Thus $-\nabla^2 f(\rho) \succeq \mu I$ on H for every $0 < \mu \leq \mu_{\text{opt}}$, which proves strong concavity. ■

Lastly, the explicit definitions of the deterministic gradient bounds are

$$D := \sqrt{|\mathcal{S}|}(\eta_V \tfrac{C_r+\eta_\rho U_G}{1-\gamma} + (1+\gamma)|\mathcal{X}|C^U)$$

and

$$B := \sqrt{|\mathcal{S}||\mathcal{A}|}\big(C_r + (1+\gamma)\tfrac{C_r+\eta_\rho U_G}{1-\gamma} + \eta_\rho \log(\tfrac{1}{\pi_{\min}})\big).$$

Then for all attainable (ρ, V, X, Ξ) it holds that $\|\hat{g}(\rho, V, X, \Xi)\|_2 \leq D$, $\|\nabla_V L(V, \rho)\|_2 \leq D$, and $\|\nabla_\rho L(V, \rho)\|_2 \leq B$. This completes the proof of Proposition 15.

C.2 Proofs of the Propositions 18 and 19

Given Assumption 13, we can formulate the limiting update rules by taking expectations under the stationary distribution of the dual-induced Markov chains.

Under the stationary distribution and using asynchronous indicator selection,

$$\begin{aligned}\mathbb{E}_{X\sim\mu_\rho}[\widehat{g}_k(\rho, V; X, \Xi)] &= \operatorname{diag}(\tilde{\mu}_\rho)\,\nabla_V L(V, \rho) \ + \bar{\mathcal{E}}_k(\rho),\\ \mathbb{E}_{X\sim\mu_\rho}[\widehat{h}_k(\rho, V; X)] &= \operatorname{diag}(\mu_\rho)\,\nabla_\rho L(V, \rho).\end{aligned}$$

Proof of Proposition 18 Fix some $\delta \in (0,1)$ and let $k \geq \max\{K(\delta), \mathcal{K}\}$. We argue on the high-probability event $\mathcal{G}_\delta$ in the following. Recall that on $\mathcal{G}_\delta$ the random stepsizes $\bar{\alpha}_k$, $\bar{\beta}_k$ have the following deterministic bounds

$$\alpha_k \leq \bar{\alpha}_k \leq \alpha_{\text{env},k}, \quad \beta_k \leq \bar{\beta}_k \leq \beta_{\text{env},k}.$$

(I.) To separate the two sources of randomness, namely the update index selection at iteration k, $X_k = (s_{k-1}, a_{k-1})$ and s_k, and the buffer draw Ξ_k, we work with the pre-sampling filtration $\mathcal{F}_k^-$, defined as the σ-field containing the entire history up to and including the buffer update $\mathcal{D}_k$, but before sampling Ξ_k. In particular, Ξ_k is conditionally independent of the past given $\mathcal{F}_k^-$ and satisfies $\Xi_k \sim \mathcal{U}(\mathcal{D}_k(\cdot))$.

Conditionally on $\mathcal{F}_k^-$, we draw a buffer sample Ξ_k to form the estimator

$$\widehat{G}_k := \widehat{g}_k(\rho_{k-1}, V_{k-1}; X_k, \Xi_k).$$

The shorthand notation of the update is

$$V_{k+1} = \Pi_{\mathcal{V}_r}\big[V_k - \bar{\alpha}_{k+1}\,\widehat{G}_{k+1}\big],$$

Recall that $\lambda(\rho) := \arg\min_V L(V, \rho)$. By non-expansiveness of $\Pi_{\mathcal{V}_r}$ and expanding the square,

$$\|V_{k+1} - \lambda(\rho_k)\|_2^2 \leq \|V_k - \lambda(\rho_k)\|_2^2 - 2\bar{\alpha}_{k+1}\left\langle V_k - \lambda(\rho_k),\, \widehat{G}_{k+1}\right\rangle + \bar{\alpha}_{k+1}^2\,\|\widehat{G}_{k+1}\|_2^2. \quad (53)$$

Taking the conditional expectation to remove the buffer noise, we set

$$\widetilde{G}_k := \mathbb{E}\big[\widehat{G}_k|\mathcal{F}_k^-\big].$$

Correspondingly, in explicit notation

$$\tilde{g}_k(\rho, V; X) := \mathbb{E}\big[\widehat{g}_k(\rho, V; X, \Xi)|\mathcal{F}_k^-\big].$$

Further, we specify the associated martingale term

$$\bar{M}_k := \widehat{G}_k - \widetilde{G}_k,$$

so that $\mathbb{E}[\bar{M}_k|\mathcal{F}_k^-] = 0$ and $\mathbb{E}[\|\bar{M}_k\|_2^2|\mathcal{F}_k^-] \le \mathbb{E}[\|\widehat{G}_k\|_2^2|\mathcal{F}_k^-] \le D^2$. Decompose $\widetilde{G}_k$ into a target term, a Markovian mismatch (depending only on the law of X_k), and the stationary replay bias:

$$\widetilde{G}_k = \underbrace{\mathrm{diag}(\tilde{\mu}_{\rho_{k-1}})\,\nabla_V L(V_{k-1}, \rho_{k-1})}_{\text{target at } \rho_{k-1}} + \underbrace{\Delta_k^X}_{\text{mismatch of the index law}} + \underbrace{\bar{\mathcal{E}}_k}_{\text{buffer bias at stationarity}}, \tag{54}$$

where

$$\Delta_k^X := \widetilde{G}_k - \mathbb{E}_{X\sim\mu_{\rho_{k-1}}}[\tilde{g}_k(\rho_{k-1}, V_{k-1}; X)], \tag{55}$$

and we omit the dependence of the bias term under stationary-law (see 44) on ρ_{k-1} by setting

$$\bar{\mathcal{E}}_k := \bar{\mathcal{E}}_k(\rho_{k-1}) = \mathbb{E}_{X\sim\mu_{\rho_{k-1}}}[\tilde{g}_k(\rho_{k-1}, V_{k-1}; X)] - \mathrm{diag}(\tilde{\mu}_{\rho_{k-1}})\,\nabla_V L(V_{k-1}, \rho_{k-1}).$$

Therefore,

$$\widehat{G}_k = \mathrm{diag}(\tilde{\mu}_{\rho_{k-1}})\,\nabla_V L(V_{k-1}, \rho_{k-1}) + \Delta_k^X + \bar{\mathcal{E}}_k + \bar{M}_k.$$

Taking the conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_k^-]$ in the one-step bound (53) removes the martingale term and yields

$$\begin{aligned}\mathbb{E}\big[\|V_{k+1} - \lambda(\rho_k)\|_2^2|\mathcal{F}_k^-\big] \le \|V_k - \lambda(\rho_k)\|_2^2 &- 2\bar{\alpha}_{k+1}\Big\langle V_k - \lambda(\rho_k),\, \mathrm{diag}(\tilde{\mu}_{\rho_k})\,\nabla_V L(V_k, \rho_k)\Big\rangle \\ &- 2\bar{\alpha}_{k+1}\Big\langle V_k - \lambda(\rho_k),\, \Delta_{k+1}^X\Big\rangle - 2\bar{\alpha}_{k+1}\Big\langle V_k - \lambda(\rho_k),\, \bar{\mathcal{E}}_{k+1}\Big\rangle \\ &+ \bar{\alpha}_{k+1}^2\,\mathbb{E}\big[\|\widehat{G}_{k+1}\|_2^2|\mathcal{F}_k^-\big].\end{aligned}$$

The remaining terms are controlled as follows.

(II.) Due to the η_V-strong convexity of the mapping $V \mapsto L(V, \rho_k)$, the gradient $\nabla_V L(\cdot, \rho_k)$ is η_V-strongly monotone and $\nabla_V L(\lambda(\rho_k), \rho_k) = 0$. Hence with $\tilde{\eta}_V = \eta_V\, p_\star\, |\mathcal{A}|$, it holds that

$$\langle V_k - \lambda(\rho_k), \mathrm{diag}(\tilde{\mu}_{\rho_k})\nabla_V L(V_k, \rho_k)\rangle \;\ge\; \tilde{\eta}_V\,\|V_k - \lambda(\rho_k)\|_2^2.$$

Expanding, using the above, and the boundedness of the gradient and its stochastic single entry estimate, $\|\widehat{G}_k\|_2^2 \le D^2$, $\|\nabla_V L\|_2^2 \le D^2$, gives

$$\begin{aligned}\mathbb{E}\big[\|V_{k+1} - \lambda(\rho_k)\|_2^2|\mathcal{F}_k^-\big] \le \|V_k - \lambda(\rho_k)\|_2^2 &- 2\bar{\alpha}_{k+1}\,\tilde{\eta}_V\,\|V_k - \lambda(\rho_k)\|_2^2 \\ &- 2\bar{\alpha}_{k+1}\,\langle V_k - \lambda(\rho_k), \Delta_{k+1}^X\rangle - 2\bar{\alpha}_{k+1}\,\langle V_k - \lambda(\rho_k), \bar{\mathcal{E}}_{k+1}\rangle + \bar{\alpha}_{k+1}^2\,D^2.\end{aligned}$$

(III.) The technical proof of the following bound on $-2\bar{\alpha}_{k+1}\langle V_k - \lambda(\rho_k), \Delta^X_{k+1}\rangle$ applies a comparison of a frozen-chain to the mixing of the inhomogeneous chain and is provided in Lemma 33. We obtain the following

$$\begin{aligned} -2\mathbb{E}\big[\bar{\alpha}_{k+1}\langle V_k - \lambda(\rho_k), \Delta^X_{k+1}\rangle|\mathcal{G}_\delta\big] &\leq 2\mathbb{E}\big[\bar{\alpha}_{k+1}|\langle V_k - \lambda(\rho_k), \Delta^X_{k+1}\rangle||\mathcal{G}_\delta\big] \\ &\leq C_1\tau_k^2\alpha_{env,k}\alpha_{env,k-\tau_k}\mathbb{E}\big[\|V_{k-\tau_k} - \lambda(\rho_{k-\tau_k})\|_2^2 + 1|\mathcal{G}_\delta\big], \end{aligned}$$

where $C_1 = 2(LDB + LB + LD + D^2 + \frac{\gamma|\mathcal{X}|C^U}{\beta_0})$.

To address the replay-buffer bias, we apply the inequality $\langle 2\upsilon_1, \upsilon_2\rangle \leq c\|\upsilon_1\|_2^2 + \frac{1}{c}\|\upsilon_2\|_2^2$ that holds for any vectors υ_1, υ_2 and $c > 0$ with $\upsilon_1 = -(V_k - \lambda(\rho_k))$, $\upsilon_2 = \bar{\mathcal{E}}_{k+1}$ and $c = \tilde{\eta}_V$ to get

$$\begin{aligned} &-2\mathbb{E}\big[\bar{\alpha}_{k+1}\langle V_k - \lambda(\rho_k), \bar{\mathcal{E}}_{k+1}\rangle|\mathcal{F}_k^-, \mathcal{G}_\delta\big] \\ &\qquad \leq \tilde{\eta}_V\mathbb{E}\big[\bar{\alpha}_{k+1}\|V_k - \lambda(\rho_k)\|_2^2|\mathcal{F}_k^-, \mathcal{G}_\delta\big] + \tfrac{1}{\tilde{\eta}_V}\,\mathbb{E}\big[\bar{\alpha}_{k+1}\|\bar{\mathcal{E}}_{k+1}\|_2^2|\mathcal{F}_k^-, \mathcal{G}_\delta\big] \\ &\qquad \leq \tilde{\eta}_V\mathbb{E}\big[\bar{\alpha}_{k+1}\|V_k - \lambda(\rho_k)\|_2^2|\mathcal{F}_k^-, \mathcal{G}_\delta\big] + \tfrac{C_{\mathrm{buf}}(\delta)^2|\mathcal{S}|}{\tilde{\eta}_V}\tfrac{\alpha_{env,k}}{k}, \end{aligned}$$

where the last inequality uses the bound on the bias under stationary law derived in Lemma 26.

Recall the deterministic bounds on the random stepsize sequence $\bar{\alpha}_k$, $\alpha_k \leq \bar{\alpha}_k \leq \alpha_{\mathrm{env},k}$. Putting the bounds together and taking the conditional expectation gives us

$$\begin{aligned} \mathbb{E}\big[\|V_{k+1} - \lambda(\rho_k)\|_2^2|\mathcal{G}_\delta\big] \leq (1 - \tilde{\eta}_V\alpha_k)&\mathbb{E}[\|V_k - \lambda(\rho_k)\|_2^2|\mathcal{G}_\delta] \\ &+ C_1\tau_k^2\alpha_{\mathrm{env},k-\tau_k}\alpha_{\mathrm{env},k}(\mathbb{E}[\|V_{k-\tau_k} - \lambda(\rho_{k-\tau_k})\|_2^2|\mathcal{G}_\delta] + 1) \\ &+ \tfrac{C_{\mathrm{buf}}(\delta)^2|\mathcal{S}|}{\tilde{\eta}_V}\tfrac{\alpha_{\mathrm{env},k}}{k} + D^2\alpha_{\mathrm{env},k}^2 \end{aligned}$$

(IV.) To obtain the desired contraction property of $\|V_{k+1} - \lambda(\rho_{k+1})\|_2^2$ we use that for any $c > 0$

$$\|V_{k+1} - \lambda(\rho_{k+1})\|_2^2 \leq (1 + c)\,\|V_{k+1} - \lambda(\rho_k)\|_2^2 + (1 + \frac{1}{c})\,\|\lambda(\rho_{k+1}) - \lambda(\rho_k)\|_2^2.$$

With the Lipschitz continuity of the best response (see Lemma 27) and the bound on the dual update, we have

$$\|\lambda(\rho_{k+1}) - \lambda(\rho_k)\|_2^2 \leq L^2\|\rho_{k+1} - \rho_k\|_2^2 \leq L^2B^2\beta_{\mathrm{env,k}}^2.$$

With the adaptive choice of the inequality $c_k = \frac{\tilde{\eta}_V}{2}\alpha_k$ and the shorthand notation $y_k = \mathbb{E}\big[\|V_k - \lambda(\rho_k)\|_2^2|\mathcal{G}_\delta\big]$, we have the combined bound

$$\begin{aligned} y_{k+1} &\leq (1 + \tfrac{\tilde{\eta}_V}{2}\alpha_k)(1 - \tilde{\eta}_V\alpha_k)y_k + (1 + \tfrac{\tilde{\eta}_V}{2}\alpha_k)C_1\tau_k^2\alpha_{\mathrm{env},k-\tau_k}\alpha_{\mathrm{env},k}(y_{k-\tau_k} + 1) \\ &\quad + (1 + \tfrac{\tilde{\eta}_V}{2}\alpha_k)\tfrac{C_{\mathrm{buf}}(\delta)^2|\mathcal{S}|}{\tilde{\eta}_V}\tfrac{\alpha_{\mathrm{env},k}}{k} + (1 + \tfrac{\tilde{\eta}_V}{2}\alpha_k)D^2\alpha_{\mathrm{env},k}^2 \\ &\quad + L^2B^2\beta_{\mathrm{env},k}^2 + \frac{2L^2B^2\beta_{\mathrm{env},k}^2}{\tilde{\eta}_V\,\alpha_k} \end{aligned}$$

To simplify the above bound, we use $(1+\frac{\tilde{\eta}_V}{2}\alpha_k)(1-\tilde{\eta}_V\alpha_k) \le 1-\frac{\tilde{\eta}_V}{2}\alpha_k$ and the following bound on $y_{k-\tau_k}$ in the second summand

$$\begin{aligned} y_{k-\tau_k} &= \mathbb{E}\big[\|V_k-\lambda(\rho_k)-(V_k-V_{k-\tau_k})+\big(\lambda(\rho_k)-\lambda(\rho_{k-\tau_k})\big)\|_2^2|\mathcal{G}_\delta\big] \\ &\le 3(y_k+\mathbb{E}[\|V_k-V_{k-\tau_k}\|_2^2|\mathcal{G}_\delta]+\mathbb{E}[\|\lambda(\rho_k)-\lambda(\rho_{k-\tau_k})\|_2^2|\mathcal{G}_\delta]) \\ &\le 3y_k+3D^2\tau_k^2\alpha_{\mathrm{env},k-\tau_k}^2+3L^2B^2\tau_k^2\beta_{\mathrm{env},k-\tau_k}^2, \end{aligned}$$

where the second inequality is a consequence of the Cauchy-Schwarz inequality, and the last inequality follows from Lipschitz continuity of the best response and the bounded changes to the variables over the mixing interval. Using the condition $\frac{\tilde{\eta}_V}{2}\alpha_k \le \frac{1}{2}$ for $k \ge \max\{K(\delta),\mathcal{K}\}$ the inequality simplifies to

$$\begin{aligned} y_{k+1} \;&\le (1-\tfrac{\tilde{\eta}_V}{2}\alpha_k)y_k+\tfrac{9}{2}C_1\tau_k^2\alpha_{\mathrm{env},k-\tau_k}\alpha_{\mathrm{env},k}y_k \\ &+\tfrac{9}{2}C_1\tau_k^2\alpha_{\mathrm{env},k-\tau_k}\alpha_{\mathrm{env},k}(D^2\tau_k^2\alpha_{\mathrm{env},k-\tau_k}^2+L^2B^2\tau_k^2\beta_{\mathrm{env},k-\tau_k}^2+\tfrac{1}{3}) \\ &+\tfrac{3}{2}\tfrac{C_{\mathrm{buf}}(\delta)^2|\mathcal{S}|}{\tilde{\eta}_V}\tfrac{\alpha_{\mathrm{env},k}}{k}+\tfrac{3}{2}D^2\alpha_{\mathrm{env},k}^2+L^2B^2\beta_{\mathrm{env},k}^2+\tfrac{2L^2B^2\beta_{\mathrm{env},k}^2}{\tilde{\eta}_V\,\alpha_k} \end{aligned}$$

By the definition of $\mathcal{K}$ in (24), for all $k \ge \max\{K(\delta),\mathcal{K}\}$, $C_1C_{\mathrm{env}}\tau_k^2\alpha_{\mathrm{env},k-\tau_k} \le \frac{\tilde{\eta}_V}{18}$ and $\tau_k\alpha_{\mathrm{env},k-\tau_k} \le \min\{\frac{1}{\sqrt{3}D},\frac{1}{\sqrt{3}LB}\}$. Inserting the bounds simplifies the inequality to

$$\begin{aligned} y_{k+1} \;&\le (1-\tfrac{\tilde{\eta}_V}{4}\alpha_k)y_k+\tfrac{9}{2}C_1\tau_k^2\alpha_{\mathrm{env},k-\tau_k}\alpha_{\mathrm{env},k}+\tfrac{3}{2}\tfrac{C_{\mathrm{buf}}(\delta)^2|\mathcal{S}|}{\tilde{\eta}_V}\tfrac{\alpha_{\mathrm{env},k}}{k} \\ &+\tfrac{3}{2}D^2\alpha_{\mathrm{env},k}^2+L^2B^2\beta_{\mathrm{env},k}^2+\tfrac{2L^2B^2}{\tilde{\eta}_V}\tfrac{\beta_{\mathrm{env},k}^2}{\alpha_k}, \end{aligned} \tag{56}$$

which finishes the primal variable contraction proof. ■

Proof of Proposition 19 Fix $\delta \in (0,1)$ and let $k \ge \max\{K(\delta),\mathcal{K}\}$. We work on the high-probability event $\mathcal{G}_\delta$. Since $\tau_k \ge 1$, the burn-in condition $\tau_k\alpha_{\mathrm{env},k-\tau_k} \le \frac{1}{\sqrt{3}LB}$ implies $\alpha_{\mathrm{env},k-\tau_k} \le \frac{1}{\sqrt{3}LB}$. For the chosen two-timescale stepsizes, $\beta_t \le \alpha_t$ for all sufficiently large t, hence

$$\bar{\beta}_k \le \beta_{\mathrm{env},k} \le \beta_{\mathrm{env},k-\tau_k} \le \alpha_{\mathrm{env},k-\tau_k} \le \frac{1}{\sqrt{3}LB} \le \frac{1}{L},$$

where the last inequality follows from $B \ge 1$. Similarly to the proof of Proposition 18, we set

$$\widehat{H}_k := \widehat{h}_k(\rho_{k-1},V_{k-1};X_k), \qquad \widetilde{H}_k := \mathbb{E}[\widehat{H}_k \mid \mathcal{F}_{k-1}].$$

Then, by taking the expectation under the stationary distribution in (17b), we have

$$\mathbb{E}_{X\sim\mu_{\rho_{k-1}}}\Big[\widehat{h}_k(\rho_{k-1},V_{k-1};X)\Big] = \mathrm{diag}(\mu_{\rho_{k-1}})\nabla_\rho L(V_{k-1},\rho_{k-1}).$$

Using this stationary mean-field identity and $\nabla f(\rho) = \nabla_\rho L(\lambda(\rho),\rho)$, we decompose

$$\widetilde{H}_{k+1} = D(\rho_k)\nabla f(\rho_k)+\Delta_{k+1}^V+\bar{\Delta}_{k+1}^X,$$

where

$$D(\rho_k) := \mathrm{diag}(\mu_{\rho_k}), \quad \Delta_{k+1}^V := \mathbb{E}_{X\sim\mu_{\rho_k}}\Big[\widehat{h}_{k+1}(\rho_k,V_k;X)-\widehat{h}_{k+1}(\rho_k,\lambda(\rho_k);X)\Big],$$

and

$$\bar{\Delta}^X_{k+1} := \widetilde{H}_{k+1} - \mathbb{E}_{X\sim\mu_{\rho_k}}\big[\widehat{h}_{k+1}(\rho_k, \lambda(\rho_k); X)\big].$$

To isolate the core finite-time contraction mechanism, we carry out the following one-step estimate in the regime where the clipping is inactive after burn-in. By L-smoothness of f,

$$f(\rho_{k+1}) \geq f(\rho_k) + \bar{\beta}_{k+1}\langle \nabla f(\rho_k), \widehat{H}_{k+1}\rangle - \frac{L}{2}\bar{\beta}^2_{k+1}\|\widehat{H}_{k+1}\|_2^2.$$

By Proposition 15(iv), $\|\widehat{H}_{k+1}\|_2^2 \leq B^2$. Substituting the decomposition of $\widetilde{H}_{k+1}$, we obtain

$$\begin{aligned} f(\rho_{k+1}) \geq f(\rho_k) &+ \bar{\beta}_{k+1}\langle \nabla f(\rho_k), D(\rho_k)\nabla f(\rho_k)\rangle \\ &+ \bar{\beta}_{k+1}\langle \nabla f(\rho_k), \Delta^V_{k+1} + \bar{\Delta}^X_{k+1} + (\widehat{H}_{k+1} - \widetilde{H}_{k+1})\rangle + \frac{LB^2}{2}\bar{\beta}^2_{k+1}. \end{aligned}$$

Rearranging gives

$$\begin{aligned} f(\rho^\star) - f(\rho_{k+1}) \leq f(\rho^\star) - f(\rho_k) &- \bar{\beta}_{k+1}\langle \nabla f(\rho_k), D(\rho_k)\nabla f(\rho_k)\rangle \\ &- \bar{\beta}_{k+1}\langle \nabla f(\rho_k), \Delta^V_{k+1} + \bar{\Delta}^X_{k+1} + (\widehat{H}_{k+1} - \widetilde{H}_{k+1})\rangle + \frac{LB^2}{2}\bar{\beta}^2_{k+1}. \end{aligned} \tag{57}$$

Using $\mu_{\rho_k}(s,a) \geq p_\star$ for all $(s,a) \in \mathcal{X}$

$$-\langle \nabla f(\rho_k), D(\rho_k)\nabla f(\rho_k)\rangle \leq -p_\star\|\nabla f(\rho_k)\|_2^2 \leq -2p_\star\mu_{\text{opt}}\big(f(\rho^\star) - f(\rho_k)\big). \tag{58}$$

By Lipschitz continuity of $\widehat{h}_k(\rho, \cdot; X)$ in V, $\|\Delta^V_{k+1}\|_2 \leq L\|V_k - \lambda(\rho_k)\|_2$. Using Young's inequality with parameter $p_\star$,

$$-\langle \nabla f(\rho_k), \Delta^V_{k+1}\rangle \leq \frac{p_\star}{2}\|\nabla f(\rho_k)\|_2^2 + \frac{1}{2p_\star}\|\Delta^V_{k+1}\|_2^2,$$

and therefore

$$-\bar{\beta}_{k+1}\langle \nabla f(\rho_k), \Delta^V_{k+1}\rangle \leq \frac{p_\star}{2}\bar{\beta}_{k+1}\|\nabla f(\rho_k)\|_2^2 + \frac{L^2}{2p_\star}\bar{\beta}_{k+1}\|V_k - \lambda(\rho_k)\|_2^2. \tag{59}$$

By the Polyak–Łojasiewicz inequality for μ_{opt}-strongly concave f,

$$\tfrac{1}{2}\|\nabla f(\rho)\|_2^2 \geq \mu_{\text{opt}}\big(f(\rho^\star) - f(\rho)\big).$$

Combining (59) with (58) and the Polyak–Łojasiewicz inequality and inserting the bounds into Equation (57) gives

$$\begin{aligned} f(\rho^\star) - f(\rho_{k+1}) \leq (1 - p_\star\mu_{\text{opt}}\bar{\beta}_{k+1})\big(f(\rho^\star) - f(\rho_k)\big) &+ \frac{L^2}{2p_\star}\bar{\beta}_{k+1}\|V_k - \lambda(\rho_k)\|_2^2 \\ &- \bar{\beta}_{k+1}\langle \nabla f(\rho_k), \bar{\Delta}^X_{k+1} + (\widehat{H}_{k+1} - \widetilde{H}_{k+1})\rangle + \frac{LB^2}{2}\bar{\beta}^2_{k+1}. \end{aligned} \tag{60}$$

To address the Martingale difference term $\widehat{H}_{k+1} - \widetilde{H}_{k+1}$, note that

$$\mathbb{E}[\langle \nabla f(\rho_k), \widehat{H}_{k+1} - \widetilde{H}_{k+1}\rangle | \mathcal{F}_k] = 0$$

Taking expectation conditional on $\mathcal{G}_\delta$, using $\beta_k \le \bar{\beta}_{k+1} \le \beta_{\mathrm{env},k}$, and invoking the Markovian mismatch bound from Lemma 34,

$$-\mathbb{E}\big[\langle \nabla f(\rho_k), \bar{\Delta}^X_{k+1}\rangle | \mathcal{G}_\delta\big] \le 12L^2B^3\tau_k^2\beta_{\mathrm{env},k-\tau_k},$$

we arrive at

$$\mathbb{E}[f(\rho^\star) - f(\rho_{k+1})|\mathcal{G}_\delta] \le \big(1 - p_\star\mu_{\mathrm{opt}}\beta_k\big)\mathbb{E}[f(\rho^\star) - f(\rho_k)|\mathcal{G}_\delta] + \frac{L^2\beta_{\mathrm{env},k}}{2p_\star}\mathbb{E}[\|V_k - \lambda(\rho_k)\|_2^2|\mathcal{G}_\delta]$$
$$+ 12L^2B^3\tau_k^2\beta_{\mathrm{env},k-\tau_k}\beta_{\mathrm{env},k} + \frac{L}{2}B^2\beta^2_{\mathrm{env},k}.$$

The tower property with $\mathcal{G}_\delta$ eliminates the Martingale difference term. Finally, since $\tau_k \ge 1$ and $\beta_{\mathrm{env},k-\tau_k} \ge \beta_{\mathrm{env},k}$, $\beta^2_{\mathrm{env},k} \le \tau_k^2\beta_{\mathrm{env},k-\tau_k}\beta_{\mathrm{env},k}$, so the last term can be absorbed into the same remainder term. Thus,

$$\mathbb{E}[f(\rho^\star) - f(\rho_{k+1})|\mathcal{G}_\delta] \le \big(1 - p_\star\mu_{\mathrm{opt}}\beta_k\big)\mathbb{E}[f(\rho^\star) - f(\rho_k)|\mathcal{G}_\delta] + \frac{L^2}{2p_\star}\beta_{\mathrm{env},k}\mathbb{E}[\|V_k - \lambda(\rho_k)\|_2^2|\mathcal{G}_\delta]$$
$$+ \frac{25L^2B^3}{2}\tau_k^2\beta_{\mathrm{env},k-\tau_k}\beta_{\mathrm{env},k},$$

which is the claimed recursion. ■

C.3 Auxiliary Results and Proofs

This section contains additional auxiliary results required for the convergence rate proof as well as the proof of Corollary 17.

We first show the Lipschitz-continuity of the mapping $\pi \mapsto V^\pi_{ur}$.

Lemma 30 (Lipschitz continuity of the unregularized value in the policy) *For a stationary policy $\pi \in \Pi$, let V^π_{ur} denote the unregularized discounted value function. Then, for any two stationary policies $\pi, \hat{\pi} \in \Pi$,*

$$\|V^\pi_{ur} - V^{\hat{\pi}}_{ur}\|_\infty \le \frac{C_r}{(1-\gamma)^2}\max_{s\in\mathcal{S}}\|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1.$$

Proof For a stationary policy π, define

$$r_\pi(s) := \sum_{a\in\mathcal{A}}\pi(a|s)r(s,a), \qquad [\mathcal{P}_\pi]_{ss'} := \sum_{a\in\mathcal{A}}\pi(a|s)\mathcal{P}(s'|s,a).$$

Then, V^π_{ur} satisfies the Bellman equation

$$V^\pi_{ur} = r_\pi + \gamma\mathcal{P}_\pi V^\pi_{ur}.$$

Subtracting the Bellman equations for π and $\hat{\pi}$ yields

$$V^\pi_{ur} - V^{\hat{\pi}}_{ur} = \gamma\mathcal{P}_\pi\big(V^\pi_{ur} - V^{\hat{\pi}}_{ur}\big) + (r_\pi - r_{\hat{\pi}}) + \gamma(\mathcal{P}_\pi - \mathcal{P}_{\hat{\pi}})V^{\hat{\pi}}_{ur}.$$

Hence,

$$V^\pi_{ur} - V^{\hat{\pi}}_{ur} = (I - \gamma\mathcal{P}_\pi)^{-1}\Big[(r_\pi - r_{\hat{\pi}}) + \gamma(\mathcal{P}_\pi - \mathcal{P}_{\hat{\pi}})V^{\hat{\pi}}_{ur}\Big].$$

Since $\mathcal{P}_\pi$ is a stochastic matrix, we have $\|\mathcal{P}_\pi^t v\|_\infty \le \|v\|_\infty$ for all $t \ge 0$ and $v \in \mathbb{R}^{|\mathcal{S}|}$. Hence, using the Neumann-series expansion,

$$(I - \gamma \mathcal{P}_\pi)^{-1} v = \sum_{t=0}^{\infty} \gamma^t \mathcal{P}_\pi^t v,$$

it follows that

$$\|(I - \gamma \mathcal{P}_\pi)^{-1} v\|_\infty \le \sum_{t=0}^{\infty} \gamma^t \|\mathcal{P}_\pi^t v\|_\infty \le \sum_{t=0}^{\infty} \gamma^t \|v\|_\infty = \frac{1}{1-\gamma}\|v\|_\infty, \qquad \forall\, v \in \mathbb{R}^{|\mathcal{S}|}.$$

See also Puterman (1994, Theorem 6.1.1) for the corresponding discounted-value representation. Moreover, for every $s \in \mathcal{S}$,

$$|r_\pi(s) - r_{\hat{\pi}}(s)| = \left| \sum_{a \in \mathcal{A}} (\pi(a|s) - \hat{\pi}(a|s)) r(s,a) \right| \le C_r \|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1,$$

and therefore

$$\|r_\pi - r_{\hat{\pi}}\|_\infty \le C_r \max_{s \in \mathcal{S}} \|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1.$$

Likewise, for any $v \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$,

$$\begin{aligned} |((\mathcal{P}_\pi - \mathcal{P}_{\hat{\pi}})v)(s)| &= \left| \sum_{a \in \mathcal{A}} (\pi(a|s) - \hat{\pi}(a|s)) \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a) v(s') \right| \\ &\le \|v\|_\infty \|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1. \end{aligned}$$

Hence,

$$\|(\mathcal{P}_\pi - \mathcal{P}_{\hat{\pi}})v\|_\infty \le \|v\|_\infty \max_{s \in \mathcal{S}} \|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1.$$

Since $0 \le r(s,a) \le C_r$, we also have

$$\|V_{ur}^{\hat{\pi}}\|_\infty \le \frac{C_r}{1-\gamma}.$$

Combining the above bounds gives

$$\begin{aligned} \|V_{ur}^{\pi} - V_{ur}^{\hat{\pi}}\|_\infty &\le \frac{1}{1-\gamma} \left(\|r_\pi - r_{\hat{\pi}}\|_\infty + \gamma \|(\mathcal{P}_\pi - \mathcal{P}_{\hat{\pi}}) V_{ur}^{\hat{\pi}}\|_\infty \right) \\ &\le \frac{1}{1-\gamma} \left(C_r + \gamma \frac{C_r}{1-\gamma} \right) \max_{s \in \mathcal{S}} \|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1 \\ &= \frac{C_r}{(1-\gamma)^2} \max_{s \in \mathcal{S}} \|\pi(\cdot|s) - \hat{\pi}(\cdot|s)\|_1, \end{aligned}$$

which proves the claim. ■

Proof of Corollary 17. By Proposition 4, the optimal dual solution $\rho^\star$ induces the optimal regularized policy, that is,

$$\pi_{\rho^\star} = \pi_r^\star.$$

By the definition of L_π in Corollary 17, whose existence follows directly from the Lipschitz estimate derived in the proof of Lemma 23, we have

$$\max_{s\in\mathcal{S}} \|\pi_{\rho_k}(\cdot|s) - \pi_{\rho^\star}(\cdot|s)\|_1 \leq L_\pi \|\rho_k - \rho^\star\|_2.$$

Applying Lemma 30 with $\pi = \pi_{\rho_k}$ and $\hat{\pi} = \pi_{\rho^\star} = \pi_r^\star$, we obtain

$$\|V_{ur}^{\pi_{\rho_k}} - V_{ur}^{\pi_r^\star}\|_\infty \leq \frac{C_r}{(1-\gamma)^2} \max_{s\in\mathcal{S}} \|\pi_{\rho_k}(\cdot|s) - \pi_{\rho^\star}(\cdot|s)\|_1 \leq L_V \|\rho_k - \rho^\star\|_2.$$

Next, by Proposition 1 and the specialization $0 \leq G(\pi(\cdot|s)) \leq \log|\mathcal{A}|$ for the Shannon entropy regularizer, it holds for every $s \in \mathcal{S}$ that

$$0 \leq V_{ur}^\star(s) - V_{ur}^{\pi_r^\star}(s) \leq \frac{\eta_\rho \log|\mathcal{A}|}{1-\gamma} = B_{\text{reg}}.$$

Therefore,

$$\begin{aligned} V_{ur}^\star(s) - V_{ur}^{\pi_{\rho_k}}(s) &= \big(V_{ur}^\star(s) - V_{ur}^{\pi_r^\star}(s)\big) + \big(V_{ur}^{\pi_r^\star}(s) - V_{ur}^{\pi_{\rho_k}}(s)\big) \\ &\leq B_{\text{reg}} + \|V_{ur}^{\pi_r^\star} - V_{ur}^{\pi_{\rho_k}}\|_\infty. \end{aligned}$$

Since $V_{ur}^\star(s) - V_{ur}^{\pi_{\rho_k}}(s) \geq 0$, this implies

$$0 \leq \Big(V_{ur}^\star(s) - V_{ur}^{\pi_{\rho_k}}(s) - B_{\text{reg}}\Big)_+ \leq \|V_{ur}^{\pi_r^\star} - V_{ur}^{\pi_{\rho_k}}\|_\infty \leq L_V \|\rho_k - \rho^\star\|_2.$$

Squaring both sides and taking conditional expectations on $\mathcal{G}_\delta$ yields

$$\mathbb{E}\left[\Big(V_{ur}^\star(s) - V_{ur}^{\pi_{\rho_k}}(s) - B_{\text{reg}}\Big)_+^2 \,\Big|\, \mathcal{G}_\delta\right] \leq L_V^2 \mathbb{E}\big[\|\rho_k - \rho^\star\|_2^2 \mid \mathcal{G}_\delta\big].$$

The final rate statement follows immediately from Theorem 16. ■

Lemma 31 (Direct bound on $\tau_{\text{inh}}(\zeta)$ and $\tau_{\min}$) *Let Assumption 13(ii) hold. Let $\tau_{\min}$ be as defined in* (41a) *and let $\tau_{\text{inh}}(\zeta)$ be the inhomogeneous mixing time of $\{P_{\rho_k}\}$ as defined in* (41b)*. Then, for any $\zeta \in (0,1)$,*

$$\tau_{\text{inh}}(\zeta) \leq (m_\star - 1) + m_\star \left\lceil \frac{\log(1/\zeta)}{\log(1/\kappa)} \right\rceil,$$

and consequently

$$\tau_{\min} = \inf_{0\leq\zeta<1} \tau_{\text{inh}}(\zeta)\Big(\frac{2-\zeta}{1-\zeta}\Big)^2 \leq 9\left[(m_\star - 1) + m_\star \left\lceil \frac{\log 2}{\log(1/\kappa)} \right\rceil\right].$$

Proof By Assumption 13(ii), there exist constants $m_\star \in \mathbb{N}$ and $\kappa \in (0,1)$ such that for every sequence $\{\rho_t\}_{t\in\mathbb{N}} \subset H$,

$$\sup_{k\geq 0} \delta\left(\prod_{t=k}^{k+m_\star-1} P_{\rho_t}\right) \leq \kappa.$$

Fix an initial time index $i \geq 0$, and let $k = q\,m_\star + r$ with $q \in \mathbb{N}_0$ and $0 \leq r < m_\star$. Write the product of k consecutive kernels as

$$\prod_{t=i}^{i+k-1} P_{\rho_t} = \left(\prod_{b=0}^{q-1} \prod_{t=i+bm_\star}^{i+(b+1)m_\star-1} P_{\rho_t}\right)\left(\prod_{t=i+qm_\star}^{i+qm_\star+r-1} P_{\rho_t}\right).$$

Using the submultiplicativity of Dobrushin's ergodic coefficient,

$$\delta(Q_1Q_2) \leq \delta(Q_1)\delta(Q_2),$$

together with the bound $\delta(Q) \leq 1$ for every Markov kernel Q, we obtain

$$\begin{aligned}\delta\left(\prod_{t=i}^{i+k-1} P_{\rho_t}\right) &\leq \prod_{b=0}^{q-1} \delta\left(\prod_{t=i+bm_\star}^{i+(b+1)m_\star-1} P_{\rho_t}\right)\cdot \delta\left(\prod_{t=i+qm_\star}^{i+qm_\star+r-1} P_{\rho_t}\right)\\ &\leq \kappa^q.\end{aligned}$$

Hence, for all $x, y \in \mathcal{X}$,

$$\left\|\delta_x \prod_{t=i}^{i+k-1} P_{\rho_t} - \delta_y \prod_{t=i}^{i+k-1} P_{\rho_t}\right\|_{\mathrm{TV}} \leq \delta\left(\prod_{t=i}^{i+k-1} P_{\rho_t}\right) \leq \kappa^q.$$

Therefore, if $\kappa^q \leq \zeta$, then any product of length $k = q\,m_\star + r$ contracts total variation distance by at most ζ. Choosing

$$q = \left\lceil \frac{\log(1/\zeta)}{\log(1/\kappa)} \right\rceil$$

yields

$$\tau_{\mathrm{inh}}(\zeta) \leq (m_\star - 1) + m_\star \left\lceil \frac{\log(1/\zeta)}{\log(1/\kappa)} \right\rceil.$$

For $\tau_{\min}$, we proceed exactly as in the current argument and evaluate at $\zeta = \frac{1}{2}$:

$$\tau_{\min} = \inf_{0\leq\zeta<1} \tau_{\mathrm{inh}}(\zeta)\Big(\frac{2-\zeta}{1-\zeta}\Big)^2 \leq \tau_{\mathrm{inh}}(1/2)\Big(\frac{2-1/2}{1-1/2}\Big)^2 = 9\,\tau_{\mathrm{inh}}(1/2).$$

Combining this with the previous bound gives

$$\tau_{\min} \leq 9\left[(m_\star - 1) + m_\star \left\lceil \frac{\log 2}{\log(1/\kappa)} \right\rceil\right],$$

which proves the claim. ∎

The Lipschitz property of the dual-induced transition kernel has been established in the proof of Lemma 23. We next provide a proof of the Lipschitz continuity of the dual-induced stationary distribution which is applied in the proof of the visitation floor lemma.

Lemma 32 (Lipschitz stationary law under Assumption 13) *Let $d, \hat{d}$ be probability distributions on $\mathcal{X}$. There exists a constant $L_\mu < \infty$ (depending only on $C_\mathcal{X}$ and ϱ) such that for all $\rho, \hat{\rho} \in H$,*

$$\|\mu_\rho - \mu_{\hat{\rho}}\|_{\mathrm{TV}} \;\leq\; L_\mu \, \|\mathcal{P}_\rho - \mathcal{P}_{\hat{\rho}}\|_\infty \;\leq\; L_\mu \, C_\mathcal{Q} \max_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \big|\rho(s', a') - \hat{\rho}(s', a')\big|. \tag{61}$$

Moreover, for all $\rho, \hat{\rho} \in H$,

$$\|d\mathcal{P}_\rho - \hat{d}\mathcal{P}_{\hat{\rho}}\|_{\mathrm{TV}} \leq \|d - \hat{d}\|_{\mathrm{TV}} + \|\mathcal{P}_\rho - \mathcal{P}_{\hat{\rho}}\|_\infty. \tag{62}$$

Proof Under Assumption 13, for each $\rho \in H$ there is a unique stationary distribution μ_ρ and applying (Mitrophanov, 2005, Corollary 3.1) we have

$$\|\mu_\rho - \mu_{\hat{\rho}}\|_{\mathrm{TV}} \leq L_\mu \|\mathcal{P}_\rho - \mathcal{P}_{\hat{\rho}}\|_\infty,$$

where $L_\mu := \log_\varrho(C_\mathcal{X}^{-1}) + \frac{1}{1-\varrho}$. It remains to control $\|\mathcal{P}_\rho - \mathcal{P}_{\hat{\rho}}\|_\infty$ in terms of $\rho - \hat{\rho}$, which has already been shown in the proof of Lemma 23, where the Lipschitz property of the $\mathcal{P}_\rho$ has been established. This shows (61). The perturbation bound (62) follows from the triangle inequality applied as

$$\|d\mathcal{P}_\rho - \hat{d}\mathcal{P}_{\hat{\rho}}\|_{\mathrm{TV}} \leq \|d\mathcal{P}_\rho - \hat{d}\mathcal{P}_\rho\|_{\mathrm{TV}} + \|\hat{d}\mathcal{P}_\rho - \hat{d}\mathcal{P}_{\hat{\rho}}\|_{\mathrm{TV}}.$$

For the first term, total variation is non-expansive under any fixed Markov kernel K, that is, $\|dK - \hat{d}K\|_{\mathrm{TV}} \leq \|d - \hat{d}\|_{\mathrm{TV}}$, giving the first summand in (62). For the second term, $\|\hat{d}(\mathcal{P}_\rho - \mathcal{P}_{\hat{\rho}})\|_{\mathrm{TV}} \leq \sup_{x \in \mathcal{X}} \|\mathcal{P}_\rho(\cdot|x) - \mathcal{P}_{\hat{\rho}}(\cdot|x)\|_{\mathrm{TV}} = \|\mathcal{P}_\rho - \mathcal{P}_{\hat{\rho}}\|_\infty$, which is (62). Note that

$$\begin{aligned}\|\mathcal{P}_\rho(\cdot|x) - \mathcal{P}_{\hat{\rho}}(\cdot|x)\|_{\mathrm{TV}} &= \tfrac{1}{2} \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \mathcal{P}(s'|x) \big|\pi_\rho(a'|s') - \pi_{\hat{\rho}}(a'|s')\big| \\ &\leq \tfrac{1}{2} \sup_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} |\pi_\rho(a'|s') - \pi_{\hat{\rho}}(a'|s')|.\end{aligned}$$

Applying Lemma 23 gives the bound on the policy difference by $C_\mathcal{Q}$ times the ρ-difference per state. ■

The following lemma provides the bound on the term $\mathbb{E}[\langle V_k - \lambda(\rho_k), \Delta^X_{k+1}\rangle | \mathcal{G}_\delta]$ that appears in step (III.) of the proof of Proposition 18. Recall that Δ^X_{k+1}, defined in (55), denotes the mismatch between the law driving the update index selection and the ideal, mixed law.

Lemma 33 *Let $K(\delta)$, $\mathcal{G}_\delta$ be as in Prop. 15 and $\mathcal{K}$ as in (24). Moreover, let the Lipschitz envelope L and the gradient bounds B, D be as in Prop. 18. For any $k \geq \max\{K(\delta), \mathcal{K}\}$, it holds that*

$$\mathbb{E}[\langle V_k - \lambda(\rho_k), \Delta^X_{k+1}\rangle | \mathcal{G}_\delta] \leq C_1 \tau_k^2 \alpha_{\mathrm{env,k}-\tau_\mathrm{k}} \mathbb{E}[\|V_{k-\tau_k} - \lambda(\rho_{k-\tau_k})\|_2^2 + 1 | \mathcal{G}_\delta],$$

where $C_1 = 2(LDB + LB + LD + D^2 + \frac{\gamma|\mathcal{X}|C^U}{\beta_0})$.

Proof The lemma is an adjusted version of (Zeng et al., 2024, Lemma 2), with the difference lying in the boundedness of the gradient in our projected setting. The proof is also closely related to theirs, with changes occurring at the points where a uniform bound on the gradient can be applied.

The idea for handling the Markovian mismatch is to compare the inhomogeneous index process $\{X_t\}$ to an auxiliary chain frozen at parameter $\rho_{k-\tau_k}$; see also the proof of (Zeng et al., 2024, Lemma 1). Specifically, define

$$\{\widetilde{X}_{k-\tau_k+t}\}_{t=0,\dots,\tau_k} \quad \text{by} \quad \widetilde{X}_{k-\tau_k} = X_{k-\tau_k}, \qquad \widetilde{X}_{k-\tau_k+t+1} \sim \mathcal{P}_{\rho_{k-\tau_k}}(\,\cdot \mid \widetilde{X}_{k-\tau_k+t}).$$

Using $z_k := V_k - \lambda(\rho_k)$, we define the following shorthand notation

$$T_1 = \mathbb{E}[\langle V_k - V_{k-\tau_k}, \mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta]$$

$$T_2 = \mathbb{E}[\langle \lambda(\rho_{k-\tau_k}) - \lambda(\rho_k), \mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta]$$

$$T_3 = \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}[\widetilde{G}_{k+1}|\mathcal{F}_k^-] - \mathbb{E}[\tilde{g}_{k+1}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_k^-]\rangle|\mathcal{G}_\delta]$$

$$T_4 = \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}[\tilde{g}_{k+1}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_k^-] - \mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_{k-\tau_k}^-]\rangle|\mathcal{G}_\delta]$$

$$T_5 = \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_{k-\tau_k}^-] - \mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \widetilde{X}_k)|\mathcal{F}_{k-\tau_k}^-]\rangle|\mathcal{G}_\delta]$$

$$T_6 = \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \widetilde{X}_k)|\mathcal{F}_{k-\tau_k}^-] - \mathbb{E}_{\bar{X}\sim\mu_{\rho_{k-\tau_k}}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \bar{X})]\rangle|\mathcal{G}_\delta]$$

$$T_7 = \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}_{\bar{X}\sim\mu_{\rho_{k-\tau_k}}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \bar{X})] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X)]\rangle|\mathcal{G}_\delta]$$

$$T_8 = \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X)] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta]$$

With the above and the application of the tower property, we get the following decomposition

$$\begin{aligned}
\mathbb{E}[\langle z_k, \Delta_{k+1}^X\rangle|\mathcal{G}_\delta] &= \mathbb{E}[\langle z_k, \mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta] \\
&= \mathbb{E}[\langle V_k - V_{k-\tau_k}, \mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta] \\
&\quad + \mathbb{E}[\langle \lambda(\rho_{k-\tau_k}) - \lambda(\rho_k), \mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta] \\
&\quad + \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\rangle|\mathcal{G}_\delta] \\
&= T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8
\end{aligned}$$

We proceed by bounding the above terms $T_1, \dots, T_8$. First, using the boundedness of the stochastic gradients $\|\widehat{G}_k\|_2 \leq D$ and $\bar{\alpha}_k \leq \alpha_{\text{env},k}$ on $\mathcal{G}_\delta$, we have

$$\begin{aligned}
T_1 &\leq \mathbb{E}[\|V_k - V_{k-\tau_k}\|_2 \|\mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\|_2|\mathcal{G}_\delta] \\
&\leq \mathbb{E}[\underbrace{\|V_k - V_{k-\tau_k}\|_2}_{\leq \sum_{t=k+1-\tau_k}^{k} \bar{\alpha}_t \|\widehat{G}_t\|_2} \underbrace{\left(\|\mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big]\|_2 + \|\mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\|_2\right)}_{\leq 2D}|\mathcal{G}_\delta] \\
&\leq 2D^2 \tau_k \alpha_{\text{env},k-\tau_k}.
\end{aligned}$$

Similarly, using the boundedness of the dual update $\|\rho_{k+1} - \rho_k\|_2 \leq \bar{\beta}_k B$ and the Lipschitz continuity of the best response,

$$T_2 \leq \mathbb{E}[\| \underbrace{\lambda(\rho_{k-\tau_k}) - \lambda(\rho_k)\|_2}_{\leq L\|\rho_{k-\tau_k} - \rho_k\|_2} \underbrace{\|\mathbb{E}\big[\widetilde{G}_{k+1}|\mathcal{F}_k^-\big] - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1}(\rho_k, V_k; X)]\|_2}_{\leq 2D} |\mathcal{G}_\delta]$$
$$\leq 2DL\mathbb{E}[\sum_{t=k+1-\tau_k}^{k} \bar{\beta}_t \sup_{V,\rho} \|\nabla_\rho L(V,\rho)\|_2 |\mathcal{G}_\delta] \leq 2LDB\tau_k\beta_{\text{env},k-\tau_k}.$$

Next, use the Lipschitz continuity of $\nabla_V L$ with joint Lipschitz constant L to get

$$\begin{aligned}
T_3 &\leq \mathbb{E}[\|z_{k-\tau_k}\|_2 \, \|\mathbb{E}[\widetilde{G}_{k+1}|\mathcal{F}_k^-] - \mathbb{E}[\tilde{g}_{k+1}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_k^-]\|_2 |\mathcal{G}_\delta] \\
&\leq L\mathbb{E}[\|z_{k-\tau_k}\|_2 \big(\|V_k - V_{k-\tau_k}\|_2 + \|\rho_k - \rho_{k-\tau_k}\|_2\big)|\mathcal{G}_\delta] \\
&\leq L\tau_k(D\alpha_{\text{env},k-\tau_k} + B\beta_{\text{env},k-\tau_k})\mathbb{E}[\|z_{k-\tau_k}\|_2|\mathcal{G}_\delta] \\
&\leq L(D+B)\tau_k\alpha_{\text{env},k}\,\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta],
\end{aligned}$$

where the last inequality uses $\beta_k \leq \alpha_k$.

The term T_4 addresses the drift in the buffer over the mixing interval. Using the linearity of the gradient $\nabla_V L$ in the transition probabilities (see 32a), we have

$$\begin{aligned}
T_4 &= \mathbb{E}[\langle z_{k-\tau_k}, \mathbb{E}[\tilde{g}_{k+1}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_k^-] - \mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_{k-\tau_k}^-]\rangle|\mathcal{G}_\delta] \\
&\leq \gamma\mathbb{E}[\|z_{k-\tau_k}\|_2 \|\sum_{x\in\mathcal{X}} \rho_{k-\tau_k}(x)\big(\mathcal{P}_{\mathcal{D}_k}(\cdot|x) - \mathcal{P}_{\mathcal{D}_{k-\tau_k}}(\cdot|x)\big)\|_2|\mathcal{G}_\delta] \\
&\leq 2\gamma\mathbb{E}[\|z_{k-\tau_k}\|_2\|\rho_{k-\tau_k}\|_1\|\mathcal{P}_{\mathcal{D}_k} - \mathcal{P}_{\mathcal{D}_{k-\tau_k}}\|_\infty|\mathcal{G}_\delta] \\
&\leq 2\gamma|\mathcal{X}|C^U\mathbb{E}[\|z_{k-\tau_k}\|_2\|\mathcal{P}_{\mathcal{D}_k} - \mathcal{P}_{\mathcal{D}_{k-\tau_k}}\|_\infty|\mathcal{G}_\delta],
\end{aligned}$$

where for the last inequality we recall the bound $\|\rho\|_1 \leq |\mathcal{X}|C^U$ for all $\rho \in H$ from Lemma 26. Fix a state-action pair $x = (s,a) \in \mathcal{X}$, and set $m := \nu_{k-\tau_k}(x)$. Let $v_x \leq \tau_k$ be the number of new samples for x in the window, and let u_s be the number of transitions from x to s in the window. Let $c_s \in \mathbb{N}$ be such that $\mathcal{P}_{\text{old}}(s|x) := \frac{c_s}{m}$ and $\mathcal{P}_{\text{new}}(s|x) := \frac{c_s+u_s}{m+v_x}$ are the empirical next-state distributions, then

$$\|\mathcal{P}_{\text{new}}(\cdot|x) - \mathcal{P}_{\text{old}}(\cdot|x)\|_{\text{TV}} = \tfrac{1}{2}\sum_{j\in\mathcal{S}} \left|\tfrac{c_j+u_j}{m+v_x} - \tfrac{c_j}{m}\right| \;\leq\; \frac{v_x}{m}.$$

Therefore,

$$\|\mathcal{P}_{\mathcal{D}_k} - \mathcal{P}_{\mathcal{D}_{k-\tau_k}}\|_\infty \;\leq\; \max_{x\in\mathcal{X}} \frac{v_x}{\nu_{k-\tau_k}(x)} \;\leq\; \frac{\tau_k}{\min_{x\in\mathcal{X}} \nu_{k-\tau_k}(x)}.$$

On $\mathcal{G}_\delta$, it holds that $\nu_t(x) \geq \frac{p_\star}{2}t$ for $t \geq K(\delta)$. By the definition of $\beta_{\text{env},k-\tau_k} = \frac{\beta_0}{\lfloor p_\star/2\,(k-\tau_k)\rfloor+1}$, we have

$$\tfrac{1}{p_\star/2(k-\tau_k)} \leq (1 + \tfrac{2}{p_\star\mathcal{K}})\tfrac{1}{p_\star/2(k-\tau_k)+1} \leq \tfrac{2}{\beta_0}\beta_{\text{env},k-\tau_k}.$$

Consequently,

$$T_4 \leq \tfrac{2\gamma|\mathcal{X}|C^U}{\beta_0}\tau_k\beta_{\text{env},k-\tau_k}\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta].$$

The next term, T_5, compares the inhomogeneous chain $\{X_k\}$ over the mixing window of length τ_k with the frozen chain $\{\widetilde{X}_k\}$. Let us introduce the following notation for the conditional probability laws

$$p_t(\cdot) = \mathbb{P}(X_t = \cdot | \mathcal{F}_{t-1}^-) \qquad \text{and} \qquad \tilde{p}_t(\cdot) = \mathbb{P}(\widetilde{X}_t = \cdot | \mathcal{F}_{t-1}^-).$$

By construction of the frozen chain, $\widetilde{X}_{k-\tau_k} = X_{k-\tau_k}$, hence $\|p_{k-\tau_k} - \tilde{p}_{k-\tau_k}\|_{\mathrm{TV}} = 0$. Using the tower property for $\mathcal{F}_{k-\tau_k}^- \subset \mathcal{F}_{k-1}^-$ and the definition of the total variation distance, $\|p_t - \tilde{p}_t\|_{\mathrm{TV}} = \frac{1}{2}\sup_{h:\mathcal{X}\to[-1,1]}\left|\int_{\mathcal{X}} h(x)(p_t - \tilde{p}_t)(dx)\right|$,

$$\begin{aligned}
&\|\mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k)|\mathcal{F}_{k-\tau_k}^-] - \mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \widetilde{X}_k)|\mathcal{F}_{k-\tau_k}^-]\|_2 \\
&\quad = \|\mathbb{E}\big[\mathbb{E}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X_k) - \tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \widetilde{X}_k)|\mathcal{F}_{k-1}^-]|\mathcal{F}_{k-\tau_k}^-\big]\|_2 \\
&\quad = \left\|\mathbb{E}\left[\int_{\mathcal{X}} \tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; x)(p_k - \tilde{p}_k(dx)|\mathcal{F}_{k-\tau_k}^-\right]\right\|_2 \\
&\quad \le 2D\mathbb{E}[\|p_k - \tilde{p}_k\|_{\mathrm{TV}}|\mathcal{F}_{k-\tau_k}^-],
\end{aligned}$$

where we used Jensen's inequality and the bound $\|\int g\, d(p-\tilde{p})\|_2 \le 2\sup_x \|g(x)\|_2\, \|p-\tilde{p}\|_{\mathrm{TV}}$ together with $\sup_x \|\tilde{g}_{k+1-\tau_k}(\cdot; x)\|_2 \le D$. The one-step laws satisfy $p_t = p_{t-1}\mathcal{P}_{\rho_{t-1}}, \qquad \tilde{p}_t = \tilde{p}_{t-1}\mathcal{P}_{\rho_{k-\tau_k}}$, therefore, with (62) and the Lipschitz bound on the kernels it holds that

$$\|p_t - \tilde{p}_t\|_{\mathrm{TV}} \le \|p_{t-1} - \tilde{p}_{t-1}\|_{\mathrm{TV}} + \|\mathcal{P}_{\rho_{t-1}} - \mathcal{P}_{\rho_{k-\tau_k}}\|_\infty \le \|p_{t-1} - \tilde{p}_{t-1}\|_{\mathrm{TV}} + L\|\rho_{t-1} - \rho_{k-\tau_k}\|_2.$$

Iterating from $t = k$ to $k - \tau_k + 1$ and using $\|p_{k-\tau_k} - \tilde{p}_{k-\tau_k}\|_{\mathrm{TV}} = 0$ gives

$$\|p_k - \tilde{p}_k\|_{\mathrm{TV}} \le L \sum_{t=k-\tau_k}^{k-1} \|\rho_t - \rho_{k-\tau_k}\|_2.$$

Moreover, on $\mathcal{G}_\delta$ the dual increments for $t \ge k - \tau_k$ satisfy $\|\rho_{t+1} - \rho_t\|_2 \le B\bar{\beta}_{t+1} \le B\beta_{\mathrm{env},k-\tau_k}$, hence $\|\rho_t - \rho_{k-\tau_k}\|_2 \le B(t-(k-\tau_k))\beta_{\mathrm{env},k-\tau_k}$ and thus $\sum_{t=k-\tau_k}^{k-1} \|\rho_t - \rho_{k-\tau_k}\|_2 \le B\tau_k^2\beta_{\mathrm{env},k-\tau_k}$. Combining the bounds yields

$$2D\,\mathbb{E}[\|p_k - \tilde{p}_k\|_{\mathrm{TV}} \mid \mathcal{F}_{k-\tau_k}^-] \le 2LDB\,\tau_k^2\beta_{\mathrm{env},k-\tau_k}.$$

Consequently,

$$T_5 \le 2LDB\,\tau_k^2\beta_{\mathrm{env},k-\tau_k}\,\mathbb{E}[\|z_{k-\tau_k}\|_2 \mid \mathcal{G}_\delta] \le LDB\,\tau_k^2\beta_{\mathrm{env},k-\tau_k}\,\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1 \mid \mathcal{G}_\delta].$$

The term T_6 compares expectation under the fixed law chain to its stationary law expectation, and by definition of $\tau_k = \tau(\beta_{\mathrm{env},k})$ we have

$$\begin{aligned}
T_6 &\le 2D\mathbb{E}[\|z_{k-\tau_k}\|_2\mathbb{E}[\underbrace{\sup_{x\in\mathcal{X}} \|\delta_x(\mathcal{P}_{\rho_{k-\tau_k}})^{\tau_k} - \mu_{\rho_{k-\tau_k}}\|_{\mathrm{TV}}}_{\le\beta_{\mathrm{env},k}}|\mathcal{F}_{k-\tau_k}^-]|\mathcal{G}_\delta] \\
&\le D\beta_{\mathrm{env},k-\tau_k}\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta].
\end{aligned}$$

To bound the stationary law drift comparison term T_7, we use the Lipschitz continuity of the stationary law, derived in Lemma 32, and obtain

$$\begin{aligned} T_7 &\le \mathbb{E}[\|z_{k-\tau_k}\|_2 \,\|\mathbb{E}_{\bar{X}\sim\mu_{\rho_{k-\tau_k}}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; \bar{X})] \\ &\qquad\qquad\qquad\qquad - \mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X)]\|_2|\mathcal{G}_\delta] \\ &\le 2D\mathbb{E}[\|z_{k-\tau_k}\|_2 \,\|\mu_{\rho_{k-\tau_k}} - \mu_{\rho_k}\|_{\mathrm{TV}}|\mathcal{G}_\delta] \\ &\le LDB\tau_k\beta_{\mathrm{env},k-\tau_k}\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta], \end{aligned}$$

where the last inequality combines the Lipschitz bound with the bounded dual update over the mixing window, similar to the derivation in the bound to T_2.

Finally, to bound T_8, we have by the Lipschitz continuity of the primal gradient

$$\begin{aligned} T_8 &\le \mathbb{E}[\|z_{k-\tau_k}\|_2 \,\|\mathbb{E}_{X\sim\mu_{\rho_k}}[\tilde{g}_{k+1-\tau_k}(\rho_{k-\tau_k}, V_{k-\tau_k}; X) - \tilde{g}_{k+1}(\rho_k, V_k; X)]\|_2|\mathcal{G}_\delta] \\ &\le L\mathbb{E}[\|z_{k-\tau_k}\|_2 \left(\|V_k - V_{k-\tau_k}\|_2 + \|\rho_k - \rho_{k-\tau_k}\|_2\right)|\mathcal{G}_\delta] \\ &\le L(D+B)\tau_k\alpha_{\mathrm{env},k-\tau_k}\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta], \end{aligned}$$

where the final inequality follows by the same arguments as used for the bound in T_3. Putting it all together, we find with $\beta_k \le \alpha_k$ that

$$\begin{aligned} &\mathbb{E}[\langle z_k, \Delta_{k+1}^X\rangle|\mathcal{G}_\delta] \le \\ &\quad \big(2L(D+B) + LDB\tau_k + D + LDB + \tfrac{2\gamma|\mathcal{X}|C^U}{\beta_0}\big)\tau_k\alpha_{\mathrm{env},k}\,\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta] \\ &\qquad\qquad + 2D(LB+D)\tau_k\alpha_{\mathrm{env},k-\tau_k} \\ &\qquad\qquad \le \underbrace{2\big(LDB + LB + LD + D^2 + \tfrac{\gamma|\mathcal{X}|C^U}{\beta_0}\big)}_{C_1}\tau_k^2\alpha_{\mathrm{env},k-\tau_k}\mathbb{E}[\|z_{k-\tau_k}\|_2^2 + 1|\mathcal{G}_\delta], \end{aligned}$$

which finishes the proof. ■

The bound on the Markov mismatch term in Proposition 19 is given by (Zeng et al., 2024, Lemma 1) which is restated here for completeness, without proof. We adapted the statement to our notation using $\bar{\beta}_k \le \beta_{\mathrm{env},k}$ on $\mathcal{G}_\delta$.

Lemma 34 *Fix $\delta \in (0,1)$ and let $k \ge \max\{K(\delta), \mathcal{K}\}$. On the event $\mathcal{G}_\delta$*

$$-\mathbb{E}\big[\big\langle \nabla f(\rho_k),\, \Delta_{k+1}^X\big\rangle \,\big|\, \mathcal{G}_\delta\big] \le 12L^2B^3\,\tau_k^2\,\beta_{\mathrm{env},k-\tau_k}.$$

Appendix D. Additional Proofs

Proof of Corollary 2 The proof for the entropy-regularized MDP is provided in (Ying and Zhu, 2020, Section 3). We adapt the proof to apply to a general strongly convex regularizer. Although the proof closely follows Ying and Zhu (2020), we include it to ensure the paper remains self-contained. Throughout this proof we adopt the shorthand notation v_s or v_{sa} for $v(s)$ and $v(s,a)$, respectively for vectors in $\mathbb{R}^{|\mathcal{S}|}$ and $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ or corresponding mappings from $\mathcal{S}$ or $\mathcal{S}\times\mathcal{A}$ to $\mathbb{R}$.

We show the correspondence of the solution of the regularized Bellman optimality operator and the optimal primal variable in (7). We begin by formulating the corresponding primal linear program and then verify that its solution is the unique fixed point of the regularized Bellman optimality operator (6). Decompose $\rho_{sa} = \tilde{\rho}_s\pi_{sa}$ with $\tilde{\rho}_s = \sum_{a\in\mathcal{A}}\rho_{sa}$ and the policy $\pi = \pi_\rho \in \Pi$. For a homogeneous regularizer G we have $G(\rho_s) = G(\tilde{\rho}_s \pi_s) = \tilde{\rho}_s G(\pi_s)$ for all $s \in \mathcal{S}$. Due to the convex regularizing term, we can replace the condition $\rho \geq 0$ of the regularized primal-dual problem with the strict inequality $\rho > 0$.

When considering the problem (7), we recall the fact that the primal variable V can without loss of generality be restricted to a compact set $\mathcal{V}_r \subset \mathbb{R}_+^{|\mathcal{S}|}$, where

$$\mathcal{V}_r = \{v \in \mathbb{R}_+^{|\mathcal{S}|} : v_i \leq \frac{C_r + \eta_\rho U_G}{1-\gamma},\ \forall i = 1, \dots |\mathcal{S}|\}. \tag{63}$$

Inserting the decomposition above into (7) and using the convex-concave structure of the regularized Lagrangian, which allows us to use Sion's minimax theorem (Sion, 1958), we obtain

$$\begin{aligned}
&\min_{V \in \mathcal{V}_r} \max_{\pi \in \Pi, \tilde{\rho}_s > 0} \mu^\top V + \sum_{s \in \mathcal{S}} \tilde{\rho}_s \left(\sum_{a \in \mathcal{A}} \pi_{sa} \Delta[V]_{sa} + \eta_\rho G(\pi_s) \right) \\
&\Leftrightarrow \min_{V \in \mathcal{V}_r} \left(\mu^\top V + \max_{\pi \in \Pi, \tilde{\rho}_s > 0} \sum_{s \in \mathcal{S}} \tilde{\rho}_s \left(\sum_{a \in \mathcal{A}} \pi_{sa} (r_{sa} + \gamma \sum_{s' \in \mathcal{S}} V_{s'} \mathcal{P}(s'|s,a)) + \eta_\rho G(\pi_s) - V_s \right) \right) \\
&\Leftrightarrow \min_{V \in \mathcal{V}_r} \left(\mu^\top V + \sup_{\tilde{\rho}_s > 0} \sum_{s \in \mathcal{S}} \tilde{\rho}_s \max_{\pi_s \in \Delta_\mathcal{A}} \left(\sum_{a \in \mathcal{A}} \pi_{sa} (r_{sa} + \gamma \sum_{s' \in \mathcal{S}} V_{s'} \mathcal{P}(s'|s,a)) + \eta_\rho G(\pi_s) - V_s \right) \right).
\end{aligned}$$

Then the primal linear programming formulation of (7) is

$$P_r : \begin{cases} \min_{V \in \mathbb{R}^{|\mathcal{S}|}} \mu^\top V \\ \text{s.t.} \max_{\pi_s \in \Delta_\mathcal{A}} \sum_{a \in \mathcal{A}} \pi_{sa} (r_{sa} + \gamma \sum_{s' \in \mathcal{S}} V_{s'} \mathcal{P}(s'|s,a)) + \eta_\rho G(\pi_s) \leq V_s \quad \forall s \in \mathcal{S}. \end{cases}$$

We note that the constraint $V \in \mathcal{V}_r$ can be relaxed to $V \in \mathbb{R}^{|\mathcal{S}|}$ in P_r without loss of generality. Indeed, if we suppose that there is a feasible candidate V of P_r such that $V \notin \mathcal{V}_r$, we can construct a feasible candidate with a smaller objective value by projecting it onto $\mathcal{V}_r$.

The constraint of the primal LP is equivalently expressed as $T_{\star,G} V \leq V$ with the regularized Bellman optimality operator defined in (6). Let $\widehat{V}$ be a solution to P_r. Then by primal feasibility, the slack variable $S := \widehat{V} - T_{\star,G}\widehat{V} \geq 0$. Assume $S \neq 0$. Note that $T_{\star,G}$ is a monotone operator, (Geist et al., 2019, Prop. 2), that is, for $V_1 \geq V_2$, $T_{\star,G} V_1 \geq T_{\star,G} V_2$. Therefore $\widetilde{V} := T_{\star,G}\widehat{V}$ fulfills $T_{\star,G}\widetilde{V} \leq T_{\star,G}\widehat{V} = \widetilde{V}$, that is, $\widetilde{V}$ is primal feasible. For $S \neq 0$, there exists a state s with $\widetilde{V}_s < \widehat{V}_s$ and hence $\mu^\top \widetilde{V} < \mu^\top \widehat{V}$. This contradicts the optimality of $\widehat{V}$, therefore $S = 0$ holds. The solution to the primal problem is the unique fixed point of the regularized Bellman optimality operator.

The dual problem for the regularized MDP is

$$D_r : \begin{cases} \sup_{\rho > 0} \ \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} r_{sa} \rho_{sa} + \eta_\rho \sum_{s \in \mathcal{S}} \tilde{\rho}_s G(\pi_s) \\ \text{s.t. } \tilde{\rho}_{s'} - \gamma \sum_{(s,a) \in \in \mathcal{S} \times \mathcal{A}} \rho_{sa} \mathcal{P}(s'|s,a) = \mu_{s'} \quad \forall s' \in \mathcal{S}. \end{cases}$$

Inserting the decomposition $\rho_{sa} = \tilde{\rho}_s \pi_{sa}$ the dual constraint is

$$\tilde{\rho}_{s'} - \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}(s'|s,a)\pi_{sa}\tilde{\rho}_s = \mu_{s'}.$$

In vector notation with $\sum_{a\in\mathcal{A}} \mathcal{P}(s'|s,a)\pi_{sa} = \mathcal{P}^{\pi}_{ss'} = (\mathcal{P}^{\pi})^{\top}_{s's}$ the constraint can be rearranged to

$$\tilde{\rho} = \left(I - \gamma \left(\mathcal{P}^{\pi}\right)^{\top}\right)^{-1} \mu.$$

Denote this as $\tilde{\rho}^{\pi}$, highlighting the dependence on π. By inserting the optimal $\tilde{\rho}^{\pi}$ into the objective, we can reformulate the regularized dual problem as

$$\begin{aligned}
&\max_{\pi\in\Pi} \sum_{s\in\mathcal{S}} \tilde{\rho}^{\pi}_s \underbrace{\sum_{a\in\mathcal{A}} r_{sa}\pi_{sa}}_{=r^{\pi}_s} + \eta_\rho \sum_{s\in\mathcal{S}} \tilde{\rho}^{\pi}_s G\left(\pi_s\right) \\
&= \max_{\pi\in\Pi} \sum_{s\in\mathcal{S}} \tilde{\rho}^{\pi}_s \left(r^{\pi}_s + \eta_\rho G\left(\pi_s\right)\right) \\
&= \max_{\pi\in\Pi} (\tilde{\rho}^{\pi})^{\top} \left(r^{\pi} + \eta_\rho G\left(\pi\right)\right) \\
&= \max_{\pi\in\Pi} \mu^{\top} \left(I - \gamma\mathcal{P}^{\pi}\right)^{-1} \left(r^{\pi} + \eta_\rho G\left(\pi\right)\right)
\end{aligned}$$

The regularized Bellman operator states $V^{\pi} = r^{\pi} + \eta_\rho G(\pi) + \gamma\mathcal{P}^{\pi}V^{\pi}$ or equivalently $V^{\pi} = (I - \gamma\mathcal{P}^{\pi})^{-1}(r^{\pi} + \eta_\rho G(\pi))$. Hence, the regularized dual problem D_r is equivalent to

$$\max_{\pi\in\Pi} \mu^{\top} V \quad \text{s.t.} \quad V = T_{\pi,G}V = r^{\pi} + \eta_\rho G(\pi) + \gamma\mathcal{P}^{\pi}V.$$

Since μ is an arbitrary positive vector, this is the policy gradient problem with the unique solution $\pi^{\star}_r = \arg\max_{\pi\in\Pi} T_{\pi,G}V^{\pi}_r$. By the decomposition of ρ, we have $\pi^{\star}_{sa} = \frac{\rho^{\star}_{sa}}{\tilde{\rho}^{\pi^{\star}}_s}$, that is, $\pi_{\rho^{\star}_r} = \pi^{\star}_r$. ■

Proof of Lemma 5 We begin by establishing an upper bound for $\rho^{\star}$. The optimal primal and dual solutions are linked by the first-order optimality condition of the primal variable V. We then use the bound on $V^{\star}$ to obtain a bound on $\rho^{\star}$. Moreover, thanks to Theorem 4, we know that the tuple $(\pi^{\star}, V^{\star})$ is a fixed point of the regularized Bellman operator (6) and therefore fulfills the regularized Bellman equation

$$V^{\star} = r_{\pi^{\star}} + \gamma\mathcal{P}^{\pi^{\star}}V^{\star} + \eta_\rho G_{\pi^{\star}},$$

with $(r_\pi)_s := \sum_{a\in\mathcal{A}} r_{sa}\pi_{sa}$, $(\mathcal{P}^{\pi})_{ss'} = \sum_{a\in\mathcal{A}} \pi_{sa}\mathcal{P}(s'|s,a)$ and $(G_\pi)_s = G(\pi(\cdot|s))$. Therefore,

$$\begin{aligned}
\|V^{\star}\|_\infty &= \|r_{\pi^{\star}} + \gamma\mathcal{P}^{\pi^{\star}}V^{\star} + \eta_\rho G_{\pi^{\star}}\|_\infty \\
&\le \|r_{\pi^{\star}}\|_\infty + \gamma \underbrace{\|\mathcal{P}^{\pi^{\star}}\|_\infty}_{=1} \|V^{\star}\|_\infty + \eta_\rho\|G_{\pi^{\star}}\|_\infty \\
&\le C_r + \gamma\|V^{\star}\|_\infty + \eta_\rho U_G.
\end{aligned}$$

In the last inequality, we applied the upper bounds on the reward and regularizer. Rearranging terms we get

$$\|V^\star\|_\infty \leq \frac{C_r + \eta_\rho U_G}{1-\gamma}. \tag{64}$$

Based on the gradient (32a), the first-order optimality conditions for (9) state

$$\begin{aligned}
&\tilde{\rho}^\star_{s'} - \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho^\star_{sa}\mathcal{P}\left(s'|s,a\right) = \eta_V V^\star_{s'} \leqslant \frac{\eta_V(C_r+\eta_\rho U_G)}{1-\gamma} \\
\Rightarrow \quad & \sum_{s'\in\mathcal{S}} \tilde{\rho}^\star_{s'} \leqslant \frac{|\mathcal{S}|\eta_V(C_r+\eta_\rho U_G)}{1-\gamma} + \gamma \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho^\star_{sa} \underbrace{\sum_{s'\in\mathcal{S}} \mathcal{P}\left(s'|s,a\right)}_{=1} \\
\Leftrightarrow \quad & \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \rho^\star_{sa} \leqslant \frac{|\mathcal{S}|\eta_V(C_r+\eta_\rho U_G)}{(1-\gamma)^2}.
\end{aligned} \tag{65}$$

Combined with non-negativity, we find that for $C^{\max} := \frac{|\mathcal{S}|\eta_V(C_r+\eta_\rho U_G)}{(1-\gamma)^2}$ and $C^U := 2C^{\max}$ it holds that $\|\rho^\star\|_\infty \leq \|\rho^\star\|_1 \leq C^{\max} < C^U < \infty$. The lower bound for $\rho^\star$ follows from the normalized first-order optimality conditions for ρ based on the gradient (32b), stating that the optimal policy takes the form of a Boltzmann policy

$$\begin{aligned}
\pi^\star_{sa} = \frac{\rho^\star_{sa}}{\tilde{\rho}^\star_s} &= \frac{\exp(\frac{1}{\eta_\rho}\Delta[V^\star]_{sa})}{\sum_{a\in\mathcal{A}}\exp(\frac{1}{\eta_\rho}\Delta[V^\star]_{sa})} \\
&= \frac{\exp(\frac{1}{\eta_\rho}(-V_s + r_{sa} + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s,a)V^\star_{s'})}{\sum_{a\in\mathcal{A}}\exp(\frac{1}{\eta_\rho}(-V_s + r_{sa} + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s,a)V^\star_{s'})}.
\end{aligned}$$

As we derived in (64), the optimal regularized value function is bounded, and together with $r_{sa} \in [0, C_r]$ we get

$$\Delta[V^\star]_{sa} = -V^\star_s + r_{sa} + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s,a)V^\star_{s'} \geq -\|V^\star\|_\infty$$

and

$$\Delta[V^\star]_{sa} \leq C_r + \gamma\|V^\star\|_\infty.$$

Hence,

$$\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi^\star_{sa} \geq \frac{\exp(-\frac{1}{\eta_\rho}\|V^\star\|_\infty)}{|\mathcal{A}|\exp(\frac{1}{\eta_\rho}(C_r+\gamma\|V^\star\|_\infty))} = \frac{1}{|\mathcal{A}|}\exp(-\frac{1}{\eta_\rho}[C_r + (1+\gamma)\|V^\star\|_\infty]).$$

With $\|V^\star\|_\infty \leq \frac{C_r+\eta_\rho U_G}{1-\gamma}$ we obtain the lower bound for the optimal policy

$$\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi^\star_{sa} \geq \frac{1}{|\mathcal{A}|}\exp(-\frac{2\eta_\rho^{-1}C_r + (1+\gamma)U_G}{1-\gamma}) =: C^L_1.$$

Next, we establish a lower bound for $\tilde{\rho}$. Note that as a consequence of (32a),

$$\tilde{\rho}^\star = \eta_V (I - \gamma(\mathcal{P}^{\pi^\star})^\top)^{-1} V^\star$$

with $(P^\pi)_{ss'} := \sum_{a\in\mathcal{A}} \pi_{sa}\mathcal{P}(s'|s,a)$. With the Neumann series, we can expand the inverse and obtain

$$\tilde{\rho}^\star = \eta_V \left(\sum_{k=0}^{\infty} \gamma^k ((\mathcal{P}^{\pi^\star})^\top)^k\right) V^\star = \eta_V V^\star + \eta_V \left(\sum_{k=1}^{\infty} \gamma^k ((\mathcal{P}^{\pi^\star})^\top)^k\right) V^\star \geq \eta_V V^\star.$$

Therefore, for all $s' \in \mathcal{S}$ it holds that

$$\begin{aligned}
\tilde{\rho}^\star_{s'} &\geq \eta_V V^\star_{s'} \\
&= \eta_V \underbrace{(r_{\pi^\star} + \gamma P^{\pi^\star} V^\star)_{s'}}_{\geq 0} + \eta_V \eta_\rho \underbrace{(G_{\pi^\star})_{s'}}_{= -\sum_{a\in\mathcal{A}} \pi^\star_{s'a} \log \pi^\star_{s'a}} \\
&\geq -\eta_V \eta_\rho \sum_{a\in\mathcal{A}} \pi^\star_{s'a} \log \pi^\star_{s'a} \\
&\geq \min_{s\in\mathcal{S}} -\eta_V \eta_\rho \sum_{a\in\mathcal{A}} \pi^\star_{sa} \log \pi^\star_{sa}.
\end{aligned}$$

The equality in the second line holds since $V^\star$ is the fixed point of the regularized Bellman operator (5). The above derivation states that $\tilde{\rho}^\star_{s'}$ is bounded from below by the minimal entropy of the optimal policy, where the minimum is over the states $s \in \mathcal{S}$. Using the Boltzmann form of the optimal policy, we now turn to bound the minimal entropy.

Define $\Delta_{\max}(s) := \max_{a\in\mathcal{A}} \Delta[V^\star]_{sa}$ and $\Delta_{\min}(s) := \min_{a\in\mathcal{A}} \Delta[V^\star]_{sa}$. We then denote the maximum difference as $\Delta_s := \Delta_{\max}(s) - \Delta_{\min}(s)$. Set $K_1 := |\mathcal{A}| - 1$, $K_2(\Delta_s) := \exp(-\frac{1}{\eta_\rho}\Delta_s)$. The minimum entropy of the softmax policy is obtained when in state s one action gets the "high" weight, $\Delta_{\max}(s)$, and the other K_1 states get "low" weight $\Delta_{\min}$. In that case, the policy assigns the probabilities

$$\begin{aligned}
p_{\max} &= \frac{\exp(\frac{1}{\eta_\rho}\Delta_{\max}(s))}{\exp(\frac{1}{\eta_\rho}\Delta_{\max}(s)) + K_1 \exp(\frac{1}{\eta_\rho}\Delta_{\min}(s))} = \frac{1}{1 + K_1 \exp(-\frac{1}{\eta_\rho}\Delta_s)} = \frac{1}{1 + K_1 K_2(\Delta_s)} \\
p_{\min} &= \frac{K_2(\Delta_s)}{1 + K_1 K_2(\Delta_s)}.
\end{aligned}$$

The corresponding entropy is

$$(G_\pi)_s = -(p_{\max}\log p_{\max} + K_1 p_{\min} \log p_{\min}) = \log(1 + K_1 K_2(\Delta_s)) + \frac{K_1}{\eta_\rho} \frac{\Delta_s K_2(\Delta_s)}{1 + K_1 K_2(\Delta_s)}.$$

This term is monotonically decreasing in Δ_s. We know that $\Delta_s \leq C_r + (1+\gamma)\|V^\star\|_\infty \leq \frac{2C_r + (1+\gamma)\eta_\rho U_G}{1-\gamma} =: \bar{\Delta}$. Therefore, we can bound the minimum entropy of the optimal policy from below by inserting the upper bound for Δ_s and find

$$\tilde{\rho}^\star_{s'} \geq \min_{s\in\mathcal{S}} -\eta_V \eta_\rho \sum_{a\in\mathcal{A}} \pi^\star_{sa} \log \pi^\star_{sa} \geq \eta_V \eta_\rho \log(1 + K_1 K_2(\bar{\Delta})) + \eta_V \frac{K_1 \bar{\Delta} K_2(\bar{\Delta})}{1 + K_1 K_2(\bar{\Delta})} =: C_2^L.$$

Applying the decomposition $\rho^\star_{sa} = \pi^\star_{sa}\tilde{\rho}^\star_s$, we find that $\rho^\star_{sa} \geq C_1^L C_2^L =: C^{\min}$ and for $C^L := \frac{1}{2}C^{\min}$ the Lemma statement now follows.

■

References

Kenneth J. Arrow, Leonid Hurwicz, Hirofumi Uzawa, Hollis B. Chenery, Selmer Johnson, and Samuel Karlin. *Studies in Linear and Non-Linear Programming*, volume 2. Stanford University Press Stanford, 1958.

Joan Bas-Serrano. *Lagragian Duality for Efficient Large-Scale Reinforcement Learning*. Phd thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2022. Date of defense: 2022-06-28.

Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu. Logistic Q-learning. In *International conference on artificial intelligence and statistics*, pages 3610–3618. PMLR, 2021.

Vivek S. Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

Vivek S. Borkar. Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851, 1998.

Vivek S. Borkar. *Convex Analytic Methods in Markov Decision Processes*, pages 347–375. Springer US, Boston, MA, 2002. ISBN 978-1-4615-0805-2. doi: 10.1007/978-1-4615-0805-2_11.

Vivek S. Borkar. An actor-critic algorithm for constrained Markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer Nature Singapore, second edition, 2023.

Vivek S. Borkar and Sean P. Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2): 447–469, 2000.

Pierre Brémaud. *Markov Chains: Gibbs fields, Monte Carlo Simulation and Queues*. Springer, Cham, 2 edition, 2020. ISBN 978-3030459819.

Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained Markov decision processes. *Management Science*, 2024.

Yichen Chen and Mengdi Wang. Stochastic primal-dual methods and sample complexity of reinforcement learning. *arXiv preprint arXiv:1612.02516*, 2016.

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear pi-learning using state and action features. In *International Conference on Machine Learning*, pages 834–843, 2018.

Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. In *International Conference on Learning Representations*, 2018a.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134, 2018b.

Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

Paul Dupuis. Large deviations analysis of reflected diffusions and constrained stochastic approximation algorithms in convex sets. *Stochastics: An International Journal of Probability and Stochastic Processes*, 21(1):63–96, 1987.

Paul Dupuis and Anna Nagurney. Dynamical systems and variational inequalities. *Annals of Operations Research*, 44:7–42, 1993.

Aleksei F. Filippov. *Differential equations with discontinuous righthand sides: control systems*, volume 18. Springer Science & Business Media, 2013.

Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M. Okolo. Offline primal-dual reinforcement learning for linear MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 3169–3177, 2024.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

Onesimo Hernández-Lerma and Jean B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Applications of Mathematics Series. Springer, 1996. ISBN 9780387945798.

Kihyuk Hong and Ambuj Tewari. A primal-dual algorithm for offline constrained reinforcement learning with linear MDPs. In *International Conference on Machine Learning*, pages 18711–18737, 2024.

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

Sham M. Kakade. *On the Sample Complexity of Reinforcement Learning*. University of London, University College London, 2003.

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

Harold J. Kushner and George Yin. Convergence with probability one: Correlated noise. *Stochastic Approximation and Recursive Algorithms and Applications*, pages 161–212, 2003.

Donghwan Lee and Niao He. Stochastic primal-dual Q-learning algorithm for discounted MDPs. In *2019 American Control Conference*, pages 4897–4902, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Haoya Li, Hsiang-Fu Yu, Lexing Ying, and Inderjit S. Dhillon. Accelerating primal-dual methods for regularized Markov decision processes. *SIAM Journal on Optimization*, 34 (1):764–789, 2024.

Zihao Li, Boyi Liu, Zhuoran Yang, Zhaoran Wang, and Mengdi Wang. Double duality: Variational primal-dual policy optimization for constrained reinforcement learning. *Journal of Machine Learning Research*, 24(385):1–43, 2023.

Fan Lu and Sean P. Meyn. Convex Q-learning in a stochastic environment: Extended version. *arXiv preprint arXiv:2309.05105*, 2023.

Alan S. Manne. Linear programming and sequential decisions. *Management Science*, 6(3): 259–267, 1960. ISSN 00251909, 15265501.

Prashant G. Mehta and Sean P. Meyn. Convex Q-learning, part 1: Deterministic optimal control. *arXiv preprint*, 2008.03559, 2020.

Sean P. Meyn. *Control Systems and Reinforcement Learning*. Cambridge University Press, 2022.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

Alexander Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Asuman E. Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline RL with general function approximation. In *International Conference on Machine Learning*, pages 26769–26791, 2023.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 2015.

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

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1st edition, 1994. ISBN 0471619779.

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

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft Q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulao, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

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

Mengdi Wang. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Christopher Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, page 125, 2003.

Lexing Ying and Yuhua Zhu. A note on optimization formulations of Markov decision processes. *arXiv preprint arXiv:2012.09417*, 2020.

Sihan Zeng, Thinh T. Doan, and Justin Romberg. A two-time-scale stochastic optimization framework with applications in control and reinforcement learning. *SIAM Journal on Optimization*, 34(1):946–976, 2024.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775, 2022.