FINITE-TIME CONVERGENCE AND REGRET GUARANTEES FOR NEURAL REINFORCEMENT LEARNING IN POMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) in Partially Observable Markov Decision Processes (POMDPs) is computationally challenging due to the curse of long-range dependence on history. Fortunately, recent studies have shown that a range of POMDPs that satisfy the filter stability assumption can be well approximated by a superstate Markov Decision Process (MDP) based on a truncated history in a finite-memory window, with a value gap that decays with the window length. However, existing finite-time convergence analyses on such superstate MDP surrogates are either limited to over-simplified linear function approximation or rely on an i.i.d. sampling oracle that is violated by POMDP trajectories, both of which severely limit the applicability of these theoretical results in practice. This motivates us to bridge the gap by investigating RL in POMDPs with *nonlinear* neural value function approximation. Our main contributions are threefold: i) We first establish a new finite-state MDP *point-wise* approximation error bound, which lays the foundation for our RL algorithm design and analysis in POMDPs; ii) By applying Temporal-Difference (TD) learning to POMDP-generated superstate transitions and targeting the Bellman fixed point of the superstate MDP, we prove a finite-time critic error bound with nonlinear neural value function approximation; and iii) By combining the neural critic function approximation error bound with the policy mirror descent framework, we obtain a finite-time sublinear regret against the optimal POMDP value. To the best of our knowledge, such a finite-time sublinear regret based on neural function approximation is the first in the literature for finite-memory POMDPs. Our numerical experiments also verify our theoretical results.

1 INTRODUCTION

Reinforcement learning (RL) has achieved enormous empirical success in fully observable environments, yet many real-world decision problems suffer from partial observability of the underlying state, such as robots operating under noisy sensing (Kurniawati, 2021), autonomous vehicles having to reason about hidden intentions of surrounding agents (Sunberg & Kochenderfer, 2022), and clinical decision-making relying on indirect measurements of patient states (Hauskrecht & Fraser, 2000). More recently, agentic AI systems have also been modeled as partially observable problems (Zhang et al., 2025a). Thus, it is more appropriate to model such problems as partially observable Markov decision processes (POMDPs) (Kaelbling et al., 1998), and the corresponding reinforcement learning problem in POMDPs is known as partially observable reinforcement learning (PORL).

In POMDPs, the current observation is generally insufficient to infer the underlying current state. Therefore, the agent needs to rely on the full action-observation history. A standard approach for solving POMDPs is to maintain a *belief state* (Kaelbling et al., 1998), which is a sufficient statistic of the history. However, the belief state lies in a continuous simplex, and exact belief updates require known transition and observation models, making exact belief-based methods computationally intractable. Practical algorithms therefore approximate the belief state through either a recurrent representation of the history (Cayci & Eryilmaz, 2024) or a finite memory (Kara & Yüksel, 2023; Anjarlekar et al., 2025). Although recurrent representations are more expressive, training them over long time horizons could suffer from vanishing or exploding gradients (Pascanu et al., 2013). In contrast, finite-memory representations act directly on a fixed-length window, leading to simpler and

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

more parallelizable training. Empirically, finite-memory baselines such as simple frame stacking have been shown to be competitive memory models on partially observable benchmarks (Morad et al., 2023). Also, Hausknecht & Stone (2015) found no systematic advantage of recurrence over finite-memory on Atari-style benchmarks. Beyond these practical motivations, truncating the history into a finite-memory window yields a so-called *superstate*, which facilitates an explicit accuracy–memory trade-off through the window length and induces a tractable finite-state surrogate MDP.

Given the finite-state surrogate MDP, one can potentially solve RL problems in POMDPs by using RL algorithmic frameworks developed for conventional MDP (e.g., the Policy Mirror Descent (PMD) framework (Abbasi-Yadkori et al., 2019; Xiao, 2022; Lan, 2023; Zhan et al., 2023) , which further unifies a wide range of algorithms including natural policy gradient (NPG), trust-region policy optimization (TRPO), and proximal policy optimization (PPO) (Tomar et al., 2020; Alfano et al., 2023)). However, existing finite-time analyses for these RL algorithms in PORL have two limitations:

1) In finite-memory methods, most existing works adopt linear function approximation. For example, Cayci et al. (2024) analyzed NAC algorithm with a finite-state controller, and Anjarlekar et al. (2025) analyzed PMD algorithm on the superstate MDP. However, it is well-known that linear models' expressiveness is limited and do not offer universal approximation no matter how many linear layers are stacked. When a POMDP value function is not well approximated by linear models, the bias manifests as a persistent error term in the theoretical analysis.
2) In contrast, neural function approximation enjoys the powerful universal approximation property (i.e., neural networks can approximate broad classes of functions arbitrarily well by scaling up the network width and depth (Hornik et al., 1989; Shen et al., 2022)). However, existing works on neural function approximation for POMDPs assume i.i.d. sampling from the discounted visitation distribution (e.g., Cayci & Eryilmaz (2024)). Such an i.i.d. sampling assumption does not hold in on-policy POMDP trajectories, where samples are long-range temporal dependent.

In light of the growing importance of PORL in many emerging applications and the above limitations in its theoretical understanding, the following fundamental open question naturally arises:

(Q): Could we establish a rigorous theoretical understanding of solving POMDPs via reinforcement learning with *neural* value function approximation through a superstate surrogate MDP?

Answering the above question is highly non-trivial due to the following fundamental challenges:

- **Model Mismatch:** In PORL through superstate MDPs, the neural critic is updated with data sampled from POMDP trajectories, whereas the Bellman fixed point is defined on the superstate MDP. This discrepancy is absent in conventional temporal RL critic analyses, where the sampling distribution and the Bellman target are defined on the same MDP (Cai et al., 2019). Establishing a finite-time convergence guarantee for neural critics in PORL therefore requires a new analysis.
- **Non-i.i.d. Sampling within Iterations and Non-stationarity across Iterations.** Within each iteration, the neural critic is updated along a POMDP trajectory rather than i.i.d. samples. Standard Markovian temporal difference (TD) analyses (Bhandari et al., 2018; Cai et al., 2019) compare an MDP trajectory with the stationary distribution of the same underlying Markov chain. In PORL, however, the POMDP trajectory does *not* follow the superstate MDP kernel. Exacerbating the problem is the fact that, across iterations, the discounted visitation distribution is *non-stationary* and changes with the policy. Thus, the critic error established under the current visitation needs to be updated to that of the optimal policy, which introduces another layer of difficulty.

Our key contribution in this paper is that we address these two fundamental challenges by demonstrating that standard TD learning for critics with *neural* function approximation, when coupled with the PMD framework for superstate MDPs, achieves *sublinear regret* against the optimal POMDP value. Our main results are summarized as follows and organized in the roadmap as shown in Fig. 1:

- **Point-wise POMDP-to-Superstate MDP Approximation:** We first establish a *pointwise* value approximation bound between the POMDP and the induced superstate MDP under a filter-stability assumption (Step ① in Fig. 1). We also show that the reward-function and transition-kernel mismatches between the POMDP and superstate MDP models can be controlled by the same filter-stability error. This identifies the key mechanism behind the superstate MDP approximation and provides the foundation for our learning algorithm.

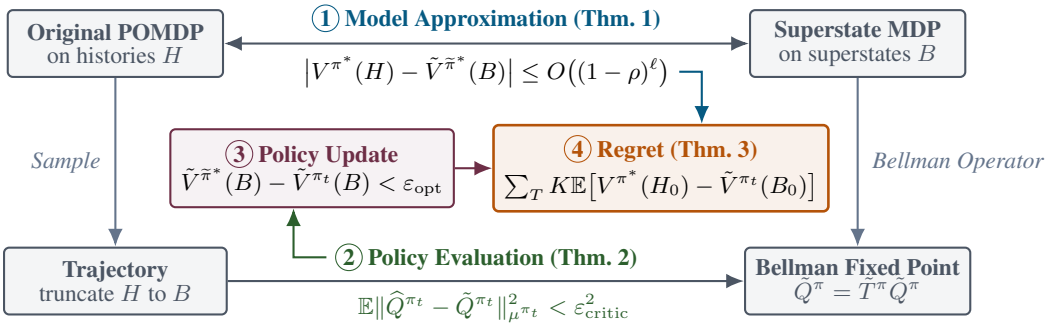


Figure 1: Roadmap for our theoretical finite-time analysis of reinforcement learning in POMDPs with neural value function approximation.

- **Finite-Time Neural TD Evaluation under POMDP Trajectory Sampling:** We establish a finite-time error bound showing that the neural critic trained from POMDP trajectories could well approximate the Bellman fixed point of the induced superstate MDP (Step ② in Fig. 1). To handle the model mismatch and non-i.i.d. trajectory sampling, we develop a *new* Lyapunov descent analysis that separates trajectory-sampling error, neural-model error, and partial-observability error.
- **Sublinear Regret for Neural PMD Algorithm:** We integrate the critic guarantee into the PMD analysis (Step ③ in Fig. 1) and prove sublinear regret for the overall algorithm (Step ④ in Fig. 1). To address the non-stationarity across policy iterations, we bridge this distribution mismatch through a concentrability assumption, and show that the regret bound explicitly separates the effects of neural critic error, finite-memory approximation error and policy optimization error.

2 Related Work

In this section, we provide an overview of three lines of research that are closely related to this work, thereby placing our contributions into a comparative perspective.

1) Finite-memory Learning for POMDPs: Since exact belief-state methods are intractable for POMDPs (Smallwood & Sondik, 1973; Lovejoy, 1991; Kaelbling et al., 1998), various practical surrogates have been proposed based on recurrent representations (Hausknecht & Stone, 2015; Cayci & Eryilmaz, 2024) or finite-memory controllers and sliding-window superstates (Poupart & Boutilier, 2003; Amato et al., 2012; Meuleau et al., 2013; Kara & Yuksel, 2022; Kara & Yüksel, 2023; Anjarlekar et al., 2025). A central tool for the latter is *filter stability*, which controls the gap between distinct histories that map to the same superstate. Exponential filter stability can be ensured under γ-observability (Golowich et al., 2023; Cai et al., 2024), or via Dobrushin coefficients of the transition and observation kernels (McDonald & Yüksel, 2020). The closest work to ours is Anjarlekar et al. (2025), which analyzes policy-based learning on a superstate MDP with linear function approximation. We adopt the same finite-memory perspective, and further investigate neural network approximation under milder assumptions compared to Anjarlekar et al. (2025).

2) Neural Function Approximation for RL in MDPs: For fully observable MDPs, finite-time guarantees have been developed for neural value-based and policy-gradient methods (Cai et al., 2019; Sun et al., 2022; Gaur et al., 2024; Zhang et al., 2025b). In POMDPs, neural methods typically aggregate histories through recurrent or latent-state representations (Wierstra et al., 2007; Hausknecht & Stone, 2015; Igl et al., 2018). Recent finite-time analyses have studied recurrent natural actor-critic methods under an i.i.d. sampling assumption (Cayci & Eryilmaz, 2024). However, finite-time analysis for neural approximation in POMDPs under standard on-policy trajectory sampling remains open.

3) Policy Mirror Descent (PMD) and Policy Gradient Methods: A growing line of work casts NPG, TRPO, and PPO as instances of policy mirror descent (Kakade, 2001; Schulman et al., 2015; 2017; Shani et al., 2020; Tomar et al., 2020; Alfano et al., 2023). For fully observable MDPs, global convergence and finite-time guarantees have been established for tabular softmax methods (Mei et al., 2020; Cen et al., 2022), regularized policy mirror descent (Lan, 2023; Zhan et al., 2023), and broader parametric policy classes, including log-linear or general parameterizations (Agarwal et al., 2021; Yuan et al., 2022; Alfano et al., 2023). However, the finite-time behavior of PMD with neural critic approximation remains unexplored in POMDP settings.

3 Problem Statement and Preliminaries

1) POMDP Model: We consider a POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, P, \Phi, r, \gamma)$ with finite state space $\mathcal{S}$, finite observation space $\mathcal{O}$, finite action space $\mathcal{A}$, transition probabilities $P : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$, observation probabilities $\Phi(o \mid s)$, reward function $r : \mathcal{S} \times \mathcal{A} \to [0, r_{\max}]$, and discount factor $\gamma \in (0, 1)$. The initial state $s_0 \in \mathcal{S}$ is sampled from a distribution $\mathcal{D}$. The dynamics satisfy:

$$\mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a) = P(s'|s, a), \qquad \forall t = 0, 1, \ldots,$$
$$\mathbb{P}(O_{t+1} = o | S_{t+1} = s') = \Phi(o|s'), \qquad \forall t = 0, 1, \ldots .$$

Since the latent state process is unobserved, the agent acts on the history $H_{t+1} = (H_t, A_t, O_{t+1}) \in \mathbb{H}$ with $H_0 = \varnothing$. The *belief state* $b_H(s) := \mathbb{P}(S_t = s \mid H_t = H)$ is a sufficient statistic of H.

2) Superstate MDP: To construct a superstate MDP, we fix a window length $\ell \in \mathbb{N}$ and let $\mathcal{G} : \mathbb{H} \to \mathsf{B}$ map each history to its most recent ℓ action–observation pairs: $B_t := \mathcal{G}(H_t) = \big(A_{\max\{0,t-\ell\}}, O_{\max\{1,t-\ell+1\}}, \ldots, A_{t-1}, O_t\big)$, with $\mathsf{B} := \{\mathcal{G}(H) : H \in \mathbb{H}\}$ being the resulting finite superstate space. Following Anjarlekar et al. (2025), given a belief $b \in \Delta(\mathcal{S})$, an action a, and an observation o, we define the belief-state update operator $\mathcal{U}_{a,o}$ using Bayes' rule as follows:

$$[\mathcal{U}_{a,o}(b)]\,(s') = \frac{\Phi(o \mid s') \sum_{s \in \mathcal{S}} P(s' \mid s, a) b(s)}{\sum_{\bar{s} \in \mathcal{S}} \Phi(o \mid \bar{s}) \sum_{s \in \mathcal{S}} P(\bar{s} \mid s, a) b(s)}.$$

Set $B_0 = \varnothing$ and $b_{B_0} = \mathcal{D}$. For $t \geq 1$, define $b_{B_t} = \mathcal{U}_{A_{t-1},O_t} \circ \cdots \circ \mathcal{U}_{A_{\max\{0,t-\ell\}},O_{\max\{1,t-\ell+1\}}}(\mathcal{D})$. We write b_B for the belief associated with superstate B. The superstate MDP $(\mathsf{B}, \mathcal{A}, \widetilde{\mathcal{P}}, \tilde{r}, \gamma)$ has reward and transition kernel defined as follows:

$$\tilde{r}(B, a) = \sum_s b_B(s)\, r(s, a), \;\; \widetilde{\mathcal{P}}(B'|B, a) = \sum_{o,s,s'} \mathbf{1}[\mathcal{G}(B \,\|\, \{a, o\}) = B']\, \Phi(o \mid s')\, P(s' \mid s, a)\, b_B(s),$$

where $\mathbf{1}[\cdot]$ is the indicator function, and $B \parallel \{a, o\}$ denotes the concatenation of the action-observation pair $\{a, o\}$ to the current superstate B. For any $B \in \mathsf{B}$ and any policy $\pi(\cdot|B)$ on the superstate MDP, the value function and action-value function are defined as

$$\tilde{V}^\pi(B) = \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t \tilde{r}(B_t, a_t) \;\big|\; B_0 = B\big], \;\; \tilde{Q}^\pi(B, a) = \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t \tilde{r}(B_t, a_t) \;\big|\; B_0 = B,\, a_0 = a\big],$$

and the Bellman operator can be written as:

$$(\tilde{T}^\pi \tilde{Q})(B, a) := \tilde{r}(B, a) + \gamma \sum_{B' \in \mathsf{B}} \sum_{a' \in \mathcal{A}} \widetilde{\mathcal{P}}(B' \mid B, a)\, \pi(a' \mid B')\, \tilde{Q}(B', a').$$

And the true Q-function $\tilde{Q}^\pi$ is the unique Bellman fixed point satisfying $\tilde{Q}^\pi = \tilde{T}^\pi(\tilde{Q}^\pi)$.

3) Neural Function Approximation: We parameterize the critic $\widehat{Q}(B, a; W)$ by a d-layer ReLU network of width m. For each superstate-action pair $(B, a) \in \mathsf{B} \times \mathcal{A}$, let $x = \psi(B, a) \in \mathbb{R}^p$ with $\|x\|_2 = 1$ be a one-to-one feature encoding. The neural network is parameterized as:

$$x^{(0)} = Ux, \quad x^{(h)} = \frac{1}{\sqrt{m}} \mathrm{ReLU}\Big(W^{(h)} x^{(h-1)}\Big) \text{ for each } h \in [d], \quad \widehat{Q}(x; W) = c^\top x^{(d)},$$

where $[d] := \{1, \ldots, d\}$, $U \in \mathbb{R}^{m \times p}$, $W^{(h)} \in \mathbb{R}^{m \times m}$, and $c \in \mathbb{R}^m$ are the network parameters, and $\mathrm{ReLU}(u) := \max\{u, 0\}$ is the activation function. For convenience, we write $W = \big(\mathrm{vec}(W^{(1)})^\top, \ldots, \mathrm{vec}(W^{(d)})^\top\big)^\top \in \mathbb{R}^{dm^2}$. At initialization, entries of U and $\{W^{(h)}\}_{h=1}^d$ are independently sampled from $\mathcal{N}(0, 2)$ and entries of c are independently sampled from $\mathcal{N}(0, 1)$. During training, we only update W, while keeping U and c fixed at their random initialization. It is worth noting that the bias-free notation is adopted for simplicity. Without loss of generality, affine ReLU layers with bias can be handled by augmenting the input and each hidden representation with a fixed coordinate, thereby absorbing the bias terms into the weight matrices. We also note that the use of ReLU activation here is only for illustrative purposes and our theoretical results extend to many common activations (e.g., Sigmoid and Tanh) with the same approximation order up to constant width and depth factors (Zhang et al., 2024).

4 Theory and Algorithm Design for Solving POMDPs via Superstate MDP

In this section, we present our theoretical results and algorithmic designs for solving POMDPs via a superstate MDP with neural value function approximation.

4.1 A New Point-wise POMDP-to-Superstate MDP Approximation Bound

We first quantify the error introduced by replacing the POMDP with the finite-memory superstate MDP by showing a *new* point-wise approximation bound between them. Toward this end, we first state two assumptions that are needed to establish the point-wise approximation bound.

Assumption 1 (Bounded Reward)**.** There exists a constant $r_{\max} > 0$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $|r(s, a)| \leq r_{\max}$.

Assumption 2 (Filter Stability)**.** There exists a constant $\rho \in (0, 1)$ such that for any two histories H and H' satisfying $\mathcal{G}(H) = \mathcal{G}(H')$, the corresponding belief states satisfy $\|b_H - b_{H'}\|_{\mathrm{TV}} \leq (1-\rho)^{\ell}$.

Assumption 2 is a point-wise form of exponential filter stability. Intuitively, the influence of observations and actions prior to the last ℓ steps decays at a geometric rate. This condition controls the mismatch between the true POMDP history process and superstate MDP. Noticing that the *Uniform Filter Stability Condition* in Anjarlekar et al. (2025) is a sufficient condition for Assumption 2. Hence, our assumption is *weaker.* Also, in the literature, the filter-stability assumption has been verified in numerous practical POMDP applications (e.g., software-reliability tracking (Debrabant & Stannat, 2013), robot localization (Lourenço et al., 2018), and neural spike-train models (Whiteley, 2021)).

We now state the resulting POMDP-to-Superstate MDP approximation guarantee, which bounds the value gap between the POMDP and the induced superstate MDP, thereby justifying the superstate MDP as an appropriate finite-memory surrogate used by our learning algorithm.

Theorem 1 (Finite-Memory Near-Optimality)**.** *Under Assumptions 1 and 2, for any history $H \in \mathbb{H}$, let $B = \mathcal{G}(H)$. Let π^* be an optimal policy of the original POMDP among all unrestricted history-dependent policies. Let $\widetilde{\pi}^*$ be an optimal policy of the induced superstate MDP. It then holds that $\|V^{\pi^*}(H) - \tilde{V}^{\widetilde{\pi}^*}(B)\| \leq \frac{2r_{\max}}{(1-\gamma)^2}(1-\rho)^{\ell}$.*

Theorem 1 says that the approximation bound is point-wise controlled by the window length ℓ for any history H. Intuitively, under filter stability, histories with the same superstate induce close beliefs. Hence, the corresponding reward-function and transition-kernel mismatches also decay at the rate of $O((1-\rho)^{\ell})$. This is the fundamental reason why the superstate MDP approximates the POMDP. See Appendix A.2 for proof details. Our result is in the same spirit as those in (Kara & Yüksel, 2023) and (Anjarlekar et al., 2025) and yet with key differences. Specifically, compared to our point-wise result, the approximation bound in (Kara & Yüksel, 2023) is in a weaker expectation sense. Meanwhile, with a stronger assumption, a similar point-wise bound with a better dependence on the $(1-\gamma)$-factor is proved in (Anjarlekar et al., 2025). Note that our bound is already sufficient for proving a sublinear regret, which is infeasible with linear function approximation adopted in (Anjarlekar et al., 2025).

4.2 Policy Optimization Algorithm Design through the Lens of Superstate MDP

Based on the superstate MDP approximation, our goal is to develop an RL algorithm with neural function approximation to learn a near-optimal policy in the sense of achieving sublinear regret. We consider a two-level neural policy iteration framework consisting of: (i) a neural TD critic (Algorithm 1) for estimating Bellman fixed point $\tilde{Q}^{\pi}$ of superstate MDP; and (ii) a policy update step (Algorithm 2) based on a closed-form exponential-weights update on the action simplex. Within each policy evaluation phase, training samples are collected from a single on-policy POMDP trajectory.

1) Policy Evaluation: We consider restart sampling: $(S, H, B) \leftarrow (S_0 \sim \mathcal{D}, \varnothing, \varnothing)$ with probability $1-\gamma$, otherwise the sampler follows the original kernel P. Then we adopt the standard semi-gradient TD-learning for policy evaluation in Algorithm 1. For a given policy π, our goal is to find the best neural network parameters $W \in \mathcal{B}$ (defined in Algorithm 1) such that $\widehat{Q}(\cdot, \cdot; W)$ approximates the Bellman fixed point $\tilde{Q}^{\pi}$ of the superstate MDP. At step k, if no reset occurs, then the agent observes the current superstate $B_k = \mathcal{G}(H_k)$, selects $a_k \sim \pi(\cdot \mid B_k)$, and the underlying POMDP returns reward r_k and the next observation o_{k+1}. The next superstate is obtained by concatenation and truncation: $H_{k+1} = H_k \parallel (a_k, o_{k+1})$ and $B_{k+1} = \mathcal{G}(H_{k+1})$, followed by $a_{k+1} \sim \pi(\cdot \mid B_{k+1})$. The critic is updated by TD learning on the transition $(B_k, a_k, r_k, B_{k+1}, a_{k+1})$, with the average parameter $\bar{W}_K$ taken as the output. If a reset occurs at step k, the critic parameters remain unchanged and the next action is sampled from $\pi(\cdot \mid B_0)$.

2) Policy Update: Algorithm 2 starts from the uniform policy and performs T outer policy updates. In each outer iteration t, we run Algorithm 1 under the current policy π_t for K critic steps, obtaining

Algorithm 1 Policy Evaluation.

Input: policy π, discount factor γ, step-size η, projection radius R, iterations K, superstate map $\mathcal{G}$.
Initialize: sample initialization $\widehat{W}_0$ as in Sec. 3; set $\bar{W}_0 = \widehat{W}_0$, define projection ball as $\mathcal{B} := \{W : \|W^{(h)} - \widehat{W}_0^{(h)}\|_F \le R, \forall h \in [d]\}$. Set $H_0 = \varnothing$, $B_0 = \mathcal{G}(H_0)$, sample $a_0 \sim \pi(\cdot \mid B_0)$.
for $k = 0, 1, \ldots, K-1$ **do**
  Sample $Z_{k+1} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\gamma)$.
  if $Z_{k+1} = 0$ **then**
    $(S_{k+1}, H_{k+1}, B_{k+1}) \leftarrow (S_0 \sim \mathcal{D}, \varnothing, \varnothing)$, sample $a_{k+1} \sim \pi(\cdot \mid B_0)$, $\widehat{W}_{k+1} \leftarrow \widehat{W}_k$.
  else
    Execute a_k, observe r_k, o_{k+1}.
    $H_{k+1} \leftarrow H_k \| (a_k, o_{k+1})$, $B_{k+1} \leftarrow \mathcal{G}(H_{k+1})$.
    Sample $a_{k+1} \sim \pi(\cdot \mid B_{k+1})$.
    $y_k \leftarrow r_k + \gamma \widehat{Q}(B_{k+1}, a_{k+1}; \widehat{W}_k)$.
    $\delta_k \leftarrow \widehat{Q}(B_k, a_k; \widehat{W}_k) - y_k$.
    $\widetilde{W}_{k+1} \leftarrow \widehat{W}_k - \eta \delta_k \nabla_W \widehat{Q}(B_k, a_k; \widehat{W}_k)$.
    $\widehat{W}_{k+1} \leftarrow \text{Proj}_{\mathcal{B}}(\widetilde{W}_{k+1})$.
  end if
  $\bar{W}_{k+1} \leftarrow \frac{k+1}{k+2} \bar{W}_k + \frac{1}{k+2} \widehat{W}_{k+1}$.
end for
Output: $\widehat{Q}^\pi := \widehat{Q}(\cdot, \cdot; \bar{W}_K)$.

Algorithm 2 Policy Update.

Input: discount factor γ, policy step-size κ, critic step-size η, projection radius R, number of policy updates T, critic iterations K, superstate map $\mathcal{G}$.
Initialize: $\pi_0(\cdot \mid B)$ uniform over $\mathcal{A}$ for every $B \in \mathsf{B}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Run Alg. 1 under π_t for K steps with an independent critic initialization; obtain
  $\widehat{Q}_t(B, a) := \widehat{Q}^{\pi_t}(B, a), \ \forall (B, a)$.
  for each $B \in \mathsf{B}$ **do**
    Update the policy as
    $$\pi_{t+1}(a \mid B) \propto \pi_t(a \mid B) e^{\kappa \widehat{Q}_t(B,a)}$$
    with
    $$\sum_{a \in \mathcal{A}} \pi_{t+1}(a \mid B) = 1.$$
  end for
end for
Output: last iterate π_T.

the averaged critic $\widehat{Q}_t(B, a)$. The policy is then updated independently in each superstate by the KL-regularized mirror-descent step as follows:

$$\pi_{t+1}(\cdot \mid B) := \underset{p \in \Delta(\mathcal{A})}{\text{argmin}} \Big\{ -\kappa \langle p, \widehat{Q}_t(B, \cdot) \rangle + D_{\text{KL}}\big(p \,\|\, \pi_t(\cdot \mid B)\big) \Big\},$$

which admits an exponential-weights closed-form update in Algorithm 2 (see Appendix D for details).

4.3 Finite-Time Convergence and Regret Analysis

We next analyze the finite-time convergence and regret performance of our proposed PMD-based policy optimization algorithm for POMDPs. The analysis first evaluates the neural TD critic error under POMDP-generated trajectory samples, and then combines this critic guarantee with PMD to obtain a sublinear regret bound against the optimal POMDP value.

1) Neural TD Critic: For a fixed policy π, let μ^π be the stationary distribution of the superstate-action Markov chain under π with the full-reset mechanism (μ^π is also the normalized discounted visitation distribution of the original SMDP, see Appendix B for proof). For critic evaluation, define the weighted L^2 norm of $f : \mathsf{B} \times \mathcal{A} \to \mathbb{R}$ by $\|f\|_{\mu^\pi} := \Big(\sum_{B,a} \mu^\pi(B, a) f(B, a)^2\Big)^{1/2}$. To analyze TD learning of our neural critic, we compare the POMDP-generated superstate process with the superstate-action chain whose stationary distribution is μ^π. Toward this end, we derive the following geometric mixing property:

Lemma 1 (Geometric Mixing). *For a given policy π, let $\{(\tilde{B}_k, \tilde{a}_k)\}_{k \ge 0}$ be the superstate-action Markov chain generated by the superstate MDP with full resets under π, which admits a unique stationary distribution μ^π. Moreover, the following bound holds for all $k \ge 0$,*

$$\sup_{(B,a) \in \mathsf{B} \times \mathcal{A}} \Big\| \mathbb{P}\big((\tilde{B}_k, \tilde{a}_k) \in \cdot \mid (\tilde{B}_0, \tilde{a}_0) = (B, a)\big) - \mu^\pi \Big\|_{\text{TV}} \le \gamma^k.$$

The geometric mixing property in Lemma 1 is commonly imposed as an assumption in the literature for finite-time TD analyses under Markovian sampling rather than i.i.d. sampling (see, e.g., Li et al.,

2020; Gaur et al., 2024). Notice that we no longer impose this property as an assumption. Rather, we show that the full-reset sampling yields the explicit mixing coefficients. Intuitively, it measures how fast the superstate-action pair mixes into the stationary distribution from any initial distribution.

We begin with the neural approximation component of the critic analysis. Lemma 2 shows that, when the projection radius R is sufficiently large, the neural critic can approximate $\tilde{Q}^\pi$ with a decreasing error as the neural network width and depth increase.

Lemma 2 (Universal Approximation). *Consider the policies $\{\pi_t\}_{t=0}^{T-1}$ generated by Algorithm 2, following from the universal approximation theorem in Shen et al. (2022), there exists a parameter $W_t^\dagger \in \mathbb{R}^{dm^2}$ for each $0 \le t < T$ such that $\big\|\widehat{Q}(\cdot,\cdot;W_t^\dagger) - \tilde{Q}^{\pi_t}\big\|_{\mu^{\pi_t}} \le \varepsilon_{\mathrm{app}}(m,d)$, where $\varepsilon_{\mathrm{app}}(m,d) := \widetilde{O}\big(m^{-2/p}d^{-2/p}\big)$. Let $\widehat{W}_{0,t}$ denote the critic initialization at iteration t. Define $R_t^\dagger := \max_{h\in[d]} \Big\|W_t^{\dagger,(h)} - \widehat{W}_{0,t}^{(h)}\Big\|_F$ and $R^\dagger := \max_{0\le t<T} R_t^\dagger$. Let $\mathcal{B}_t(R)$ be projection ball of radius R centered at $\widehat{W}_{0,t}$. Then, for any projection radius $R \ge R^\dagger$, we have $W_t^\dagger \in \mathcal{B}_t(R)$ for every $0 \le t < T$, and hence $\max_{0\le t<T} \inf_{W\in\mathcal{B}_t(R)} \Big\|\widehat{Q}(\cdot,\cdot;W) - \tilde{Q}^{\pi_t}\Big\|_{\mu^{\pi_t}} \le \varepsilon_{\mathrm{app}}(m,d)$.*

We now turn to the finite-time TD dynamics and establish the following critic error bound.

Theorem 2 (Critic Error Bound). *Under Assumptions 1 and 2, for a given policy $\pi := \pi(\cdot|B)$, let $\widehat{Q}(\cdot,\cdot;\bar{W}_K)$ be the Q-function learned by Algorithm 1. Suppose the projection radius satisfies $R^\dagger \le R = \widetilde{O}(m^{1/2}d^{-6})$, where $R^\dagger$ is defined in Lemma 2. Then, with probability at least $1 - e^{-\Omega(\log^2 m)}$,*

$$\mathbb{E}\|\widehat{Q}(\cdot,\cdot;\bar{W}_K) - \tilde{Q}^\pi\|_{\mu^\pi}^2 \le \underbrace{\widetilde{O}\Big(R^{\frac{8}{3}}m^{-\frac{1}{3}}d^6 + R^{\frac{4}{3}}m^{-\frac{1}{6}}d^3 + R^{\frac{7}{3}}m^{-\frac{1}{6}}d^{\frac{9}{2}}\Big)}_{\varepsilon_{\mathrm{NTK}}^2} + \underbrace{\widetilde{O}\Big(m^{-\frac{4}{p}}d^{-\frac{4}{p}}\Big)}_{\varepsilon_{\mathrm{app}}^2}$$
$$+ \underbrace{\widetilde{O}\Big((d^2 + d^{\frac{7}{2}}R + d^5R^2)/\sqrt{K}\Big)}_{\varepsilon_{\mathrm{TD}}^2} + \underbrace{\widetilde{O}\Big((d^{\frac{3}{2}}R + d^3R^2)(1-\rho)^\ell\Big)}_{\varepsilon_{\mathrm{hist}}^2} + \underbrace{\widetilde{O}\Big((d^{\frac{3}{2}}R + d^3R^2)/\sqrt{K}\Big)}_{\varepsilon_{\mathrm{mix}}^2}.$$

The critic error bound in Theorem 2 consists of three types of errors, which are interpreted as follows:

- **Architecture Error ($\varepsilon_{\mathrm{NTK}}^2$ and $\varepsilon_{\mathrm{app}}^2$):** These two terms come from the neural critic architecture. $\varepsilon_{\mathrm{NTK}}^2$ measures the gap between the nonlinear neural critic and its NTK linearization around the initialization $\widehat{W}_0$. This term comes from two uses of the NTK approximation in our proof: 1) We compare the learned critic $\widehat{Q}(\cdot,\cdot;\bar{W}_K)$ with its linearized version; and 2) During the TD analysis, we replace the nonlinear network values and gradients in the TD update by their linearized counterparts. $\varepsilon_{\mathrm{NTK}}^2$ is controlled jointly by the width m, depth d, and projection radius R. $\varepsilon_{\mathrm{app}}^2$ is the universal-approximation error, measuring how well the neural function class approximates $\tilde{Q}^\pi$ and decreasing as the width m and depth d of the neural network increase.
- **Sampling Error ($\varepsilon_{\mathrm{TD}}^2$ and $\varepsilon_{\mathrm{mix}}^2$).** $\varepsilon_{\mathrm{TD}}^2$ is the standard projected-TD optimization error and decays with the increase of the critic update horizon K. $\varepsilon_{\mathrm{mix}}^2$ is the price we pay for using a single on-policy trajectory instead of an i.i.d. sampling oracle. Under geometric mixing (Lemma 1), choosing the analysis lag $\tau = \max\{1, \lceil \log K/(2|\log\gamma|)\rceil\}$ gives $\gamma^\tau \le K^{-1/2}$. Hence, this term decays as $\widetilde{O}(K^{-1/2})$ with the other parameters fixed.
- **Partial-Observability Error ($\varepsilon_{\mathrm{hist}}^2$).** $\varepsilon_{\mathrm{hist}}^2$ is unique to PORL and reflects the *model mismatch*: the critic is updated by data from POMDP transitions $(H_k, a_k, r_k, H_{k+1}, a_{k+1})$, whereas its Bellman target is defined on the superstate MDP. Under filter stability (Assumption 2), the resulting bias decays as $(1-\rho)^\ell$. Thus increasing the memory length ℓ reduces this error exponentially.

We note that the critic bound is not monotone in either the width m or the depth d alone, because it also depends on the projection radius R. The NTK local-linearization argument requires $R = \widetilde{O}(m^{1/2}d^{-6})$, whereas Lemma 2 requires the projection ball to contain the approximating parameter, i.e., $R^\dagger \le R$. These conditions are compatible when $m \gtrsim (R^\dagger)^2 d^{12}$. Further, if $R^\dagger = O(1)$, then $m \gtrsim d^{12}$ suffices and $d = O(\log m)$ is admissible, yielding a *simplified* critic error bound as follows:

Corollary 1 (Simplified Critic Error Bound). *Suppose the conditions of Theorem 2 hold and assume $R^\dagger = O(1)$. Choose a projection radius $R = O(1)$ such that $R^\dagger \le R$, and let $d = O(\log m)$. Then,*

with probability at least $1 - e^{-\Omega(\log^2 m)}$*, it holds that* $\mathbb{E}\|\widehat{Q}^\pi - \tilde{Q}^\pi\|_{\mu^\pi}^2 \le \widetilde{O}\big(m^{-1/6} + m^{-4/p} + K^{-1/2} + (1-\rho)^\ell\big)$.

Corollary 1 implies that, for any $\varepsilon > 0$, the critic error can be made at most ε by taking $m \gtrsim \varepsilon^{-\max\{6,p/4\}}$, $K \gtrsim \varepsilon^{-2}$, and $\ell \gtrsim \log(1/\varepsilon)/|\log(1-\rho)|$. Thus, the neural TD critic can approximate the superstate Bellman fixed point to arbitrary accuracy.

2) Policy Update: We now combine the neural critic with the PMD policy update. Since the critic is based on the sampling distribution μ^{π_t} of the current policy, while the policy update analysis uses the optimal superstate policy $\widetilde{\pi}^*$ as its comparator, we adopt the following assumption to characterize the distribution mismatch.

Assumption 3 (Concentrability Coefficient)**.** There exists $\bar{C}_\infty < \infty$ such that, for all $t \ge 0$, $C_t := \mathbb{E}_{(B,a)\sim\mu^{\pi_t}}\left[\left(\frac{\mu^{\widetilde{\pi}^*}(B,a)}{\mu^{\pi_t}(B,a)}\right)^2 \middle| \pi_t\right]$ satisfies $\mathbb{E}[C_t] \le \bar{C}_\infty$.

The concentrability coefficient C_t quantifies how much the sampling distribution μ^{π_t} covers the distribution $\mu^{\widetilde{\pi}^*}$. Intuitively, C_t is small only if μ^{π_t} has full support over the regions visited by $\widetilde{\pi}^*$. Assumption 3 is the standard condition used in finite-time analyses of NPG and PMD methods (Yuan et al., 2022; Alfano et al., 2023), and has also been adopted in POMDPs by Cayci et al. (2024).

To evaluate the performance of Algorithm 2, we use a regret notion similar to Anjarlekar et al. (2025). Let H_0 be the initial history and $B_0 = \mathcal{G}(H_0)$ be the corresponding initial superstate. We define $R_N := \sum_{t=0}^{T-1}\sum_{k=0}^{K-1}\mathbb{E}\big[V^{\pi^*}(H_0) - \tilde{V}^{\pi_t}(B_0)\big] = \sum_{t=0}^{T-1} K\,\mathbb{E}\big[V^{\pi^*}(H_0) - \tilde{V}^{\pi_t}(B_0)\big]$, where $N = TK$ is the total interaction budget, and $V^{\pi^*}(H_0)$ and $\tilde{V}^{\pi_t}(B_0)$ denote the value functions of the optimal policy π^* in the original POMDP and the iterate π_t in the superstate MDP, respectively.

Theorem 3 (Regret Bound)**.** *Under Assumptions 1–3, let* $\{\pi_t\}_{t=0}^T$ *be the policies generated by Algorithm 2, and suppose that at each policy iteration the critic is trained by Algorithm 1. Let* $N = TK$ *denote the total interaction budget, and choose* $K = N^{2/3}$ *and* $T = N^{1/3}$*. Then, under the conditions of Theorem 2, with probability at least* $1 - N^{1/3}e^{-\Omega(\log^2 m)}$*, the total regret satisfies:*

$$R_N \le \widetilde{O}\Big(\underbrace{N(1-\rho)^\ell}_{\text{Model Approx.}} + \underbrace{N(\varepsilon_{\mathrm{NTK}} + \varepsilon_{\mathrm{app}})}_{\text{Architecture}} + \underbrace{N\varepsilon_{\mathrm{hist}}}_{\substack{\text{Critic History}\\ \text{Truncation}}} + \underbrace{N^{5/6}P_{\mathrm{mix}}}_{\substack{\text{Trajectory}\\ \text{Mixing}}} + \underbrace{N^{5/6}P_{\mathrm{opt}}}_{\substack{\text{PMD}\\ \text{Optimization}}} \Big), \tag{1}$$

where $\varepsilon_{\mathrm{NTK}}, \varepsilon_{\mathrm{app}}, \varepsilon_{\mathrm{hist}}$ *are the square roots of the corresponding terms in Theorem 2, and* $P_{\mathrm{mix}}(d,R) = d^{3/4}R^{1/2} + d^{3/2}R$, $P_{\mathrm{opt}}(d,R) = 1 + d + d^{3/2}R + d^{7/4}R^{1/2} + d^{5/2}R$.

The total regret bound in Theorem 3 consists of three parts, which can be interpreted as follows:

- **Model Approximation:** By Theorem 1, the term $N(1-\rho)^\ell$ comes from accumulating the value gap between the original POMDP and the surrogate superstate MDP over the total budget N, which is controlled by the memory length ℓ through filter stability and **decays exponentially fast**.
- **Critic-Induced Errors:** The middle three terms come from *neural* TD evaluation: the architecture term $N(\varepsilon_{\mathrm{NTK}} + \varepsilon_{\mathrm{app}})$, the critic-level history truncation term $N\varepsilon_{\mathrm{hist}}$, and the trajectory-mixing term $N^{5/6}P_{\mathrm{mix}}(d,R)$. The concentrability coefficient in Assumption 3 transfers these critic errors from the sampling distribution μ^{π_t} to the distribution associated with $\widetilde{\pi}^*$. As shown in Corollary 1, the underlying critic errors can be reduced through suitable parameter choices.
- **PMD Optimization.** The term $\widetilde{O}(N^{5/6}P_{\mathrm{opt}}(d,R))$ corresponds to the finite-sample optimization errors from both critic evaluation and policy update, which is **sublinear** in N.

The critic-induced terms have been characterized in Theorem 2 and Corollary 1. The following corollary offers a clean and simplified selection of hyperparameters under the assumption $R^\dagger = O(1)$:

Corollary 2 (Sublinear Regret)**.** *Under the conditions of Corollary 1, let* $K = N^{2/3}$ *and* $T = N^{1/3}$*. If the memory length* ℓ *and width* m *satisfy that* $\ell \ge \log N/(3|\log(1-\rho)|)$*, and* $m \gtrsim \max\{N^2, N^{p/12}\}$*, then, with probability at least* $1 - N^{1/3}e^{-\Omega(\log^2 m)}$*, the regret of Algorithm 2 satisfies* $R_N = \widetilde{O}(N^{5/6})$.

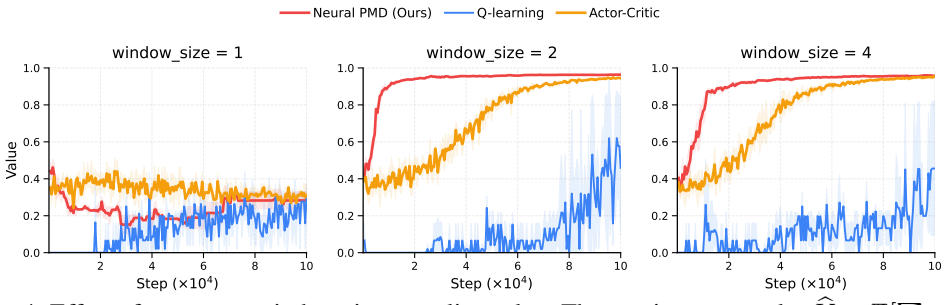


Figure 4: Effect of superstate window size on policy value. The y-axis reports value $\widehat{V} = \mathbb{E}[\sum_h \gamma^h r_h]$ estimated over 50 episodes every 500 training steps. Curves show mean $\pm$ one standard deviation.

Algorithm 2 achieves a sublinear regret $\widetilde{O}(N^{5/6})$. The universal approximation property of the neural critic, combined with a sufficiently large width m and logarithmic memory length ℓ, ensures the overall regret remains *sublinear* in N.

5 Experimental Results

We evaluate our policy optimization method on the *CheeseMaze* benchmark shown in Fig. 2, a partially observable navigation task introduced in McCallum (1993). The environment contains 11 latent states and 7 observations, where multiple states share the same observation. The action space consists of 4 directions. The agent should navigate to the goal state with observation 6, and receives a reward of $+1$ only upon reaching the goal and 0 otherwise. We compare against two standard PORL baselines: Q-learning and Actor-Critic. All methods use the same finite-memory superstate input and the same interaction budget.

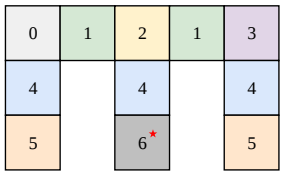


Figure 2: CheeseMaze.

Fig. 4 compares our method with Q-learning and Actor-Critic under different superstate window sizes $\ell \in \{1, 2, 4\}$. When $\ell = 1$, the agent only observes the most recent action–observation pair, giving a minimal-memory baseline. All methods perform poorly in this setting. As the window size increases to $\ell = 2$ and $\ell = 4$, our method becomes more stable than Q-learning and converges earlier than Actor-Critic. Q-learning is unstable because its greedy max-based update can amplify estimation errors. Actor-Critic improves more smoothly, but is slower because it also learns actor parameters. Increasing ℓ from 2 to 4 yields only marginal improvement, suggesting that a short finite-memory window already captures most useful information in this benchmark. Fig. 3 shows that the regret curve of our neural PMD-based method grows more slowly and stays below the baseline curves, which indicates a sublinear empirical regret consistent with our analysis. Sensitivity to policy stepsize κ and additional hyper-parameters are deferred to the Appendix E.

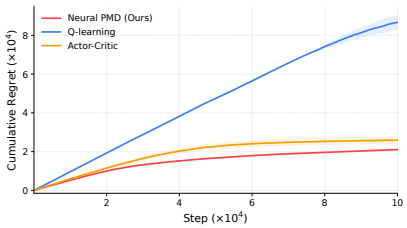


Figure 3: Empirical regret ($\ell = 2$): The y-axis reports the empirical cumulative regret $\widehat{R}_N := \sum_{t=0}^{T-1} K(V^* - \widehat{V}_t)$. Curves show mean ± 1 standard deviation.

6 Conclusion

In this paper, we studied RL in POMDPs with neural function approximation through a finite-memory superstate MDP. We first characterized the superstate MDP approximation, and then proved a finite-time critic error bound and a sublinear regret bound against the optimal POMDP value. We conducted experiments on the CheeseMaze benchmark to validate our theoretical results, showing greater stability than Q-learning and higher sample efficiency than Actor–Critic. A limitation of this work is that it focuses on finite state, observation, and action spaces, which allows truncated histories to enable a finite superstate MDP. Future directions include sharpening theoretical results, relaxing structural assumptions, extending the framework to more sophisticated architectures (e.g., Transformers), and generalizing it to continuous state and action spaces.

AI USE STATEMENT

Generative AI tools were used to polish the writing, improve the presentation of figures, and check portions of mathematical derivations and proofs. The authors independently verified all AI-assisted text, mathematical claims, proofs, citations, and code. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The supplementary material contains the complete implementation of our neural policy mirror descent method. The package also contains the Q-learning baseline and the Actor-Critic baseline. Algorithm 1 states the neural critic procedure and Algorithm 2 states the full policy optimization procedure. Assumption 1, Assumption 2 and Assumption 3 state every condition used in the analysis. Appendix A through Appendix D give complete proofs of every lemma, theorem and corollary in the main text. The CheeseMaze environment is implemented inside the package, so the experiments require no external dataset. Furthermore, the README states the package requirements, the installation steps and the exact command that reproduces each reported result, including the random seeds and the memory lengths. Appendix E.2 lists every hyperparameter setting and Appendix E.3 reports the compute resources used for the reported runs.

REFERENCES

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International conference on machine learning*, pp. 3692–3702. PMLR, 2019.

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

Carlo Alfano, Rui Yuan, and Patrick Rebeschini. A novel framework for policy mirror descent with general parameterization and linear convergence. *Advances in Neural Information Processing Systems*, 36:30681–30725, 2023.

Christopher Amato, Daniel S Bernstein, and Shlomo Zilberstein. Optimizing memory-bounded controllers for decentralized pomdps. *arXiv preprint arXiv:1206.5258*, 2012.

Ameya Anjarlekar, Rasoul Etesami, and R Srikant. Scalable policy-based rl algorithms for pomdps. *arXiv preprint arXiv:2510.06540*, 2025.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Q Cai, Zhuoran Yang, Jason Lee, and Zhaoran Wang. Neural temporal-difference and q-learning provably converge to global optima. arxiv. *Learning*, 2019.

Yang Cai, Xiangyu Liu, Argyris Oikonomou, and Kaiqing Zhang. Provable partially observable reinforcement learning with privileged information. *Advances in Neural Information Processing Systems*, 37:63790–63857, 2024.

Semih Cayci and Atilla Eryilmaz. Recurrent natural policy gradient for pomdps. *arXiv preprint arXiv:2405.18221*, 2024.

Semih Cayci, Niao He, and R Srikant. Finite-time analysis of natural actor-critic for pomdps. *SIAM Journal on Mathematics of Data Science*, 6(4):869–896, 2024.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Birgit Debrabant and Wilhelm Stannat. Stability of the optimal filter in a hidden markov model with multiplicative noise. *arXiv preprint arXiv:1301.4428*, 2013.

Mudit Gaur, Amrit Singh Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (last iterate) of actor-critic under markovian sampling with neural network parametrization. *arXiv preprint arXiv:2405.01843*, 2024.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning and learning in partially observable systems via filter stability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 349–362, 2023.

Matthew J Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. In *AAAI fall symposia*, volume 45, pp. 141, 2015.

Milos Hauskrecht and Hamish Fraser. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial intelligence in medicine*, 18(3):221–244, 2000.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and JoÃĢo GM AraÃšjo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International conference on machine learning*, pp. 2117–2126. PMLR, 2018.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Ali Kara and Serdar Yuksel. Near optimality of finite memory feedback policies in partially observed markov decision processes. *Journal of Machine Learning Research*, 23(11):1–46, 2022.

Ali Devran Kara and Serdar Yüksel. Convergence of finite memory q learning for pomdps and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 48(4): 2066–2093, 2023.

Hanna Kurniawati. Partially observable markov decision processes (pomdps) and robotics. *arXiv preprint arXiv:2107.07599*, 2021.

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

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.

Pedro Lourenço, Pedro Batista, Paulo Oliveira, and Carlos Silvestre. A globally exponentially stable filter for bearing-only simultaneous localization and mapping with monocular vision. *Robotics and Autonomous Systems*, 100:61–77, 2018.

William S Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65, 1991.

Andrew McCallum. Overcoming incomplete perception with util distinction memory. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, pp. 190–196, 1993.

Curtis McDonald and Serdar Yüksel. Exponential filter stability via dobrushin's coefficient. 2020.

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

Nicolas Meuleau, Leonid Peshkin, Kee-Eung Kim, and Leslie Pack Kaelbling. Learning finite-state controllers for partially observable environments. *arXiv preprint arXiv:1301.6721*, 2013.

Steven Morad, Ryan Kortvelesy, Matteo Bettini, Stephan Liwicki, and Amanda Prorok. Popgym: Benchmarking partially observable reinforcement learning. *arXiv preprint arXiv:2303.01859*, 2023.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. Pmlr, 2013.

Pascal Poupart and Craig Boutilier. Bounded finite state controllers. *Advances in neural information processing systems*, 16, 2003.

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

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

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5668–5675, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Tao Sun, Dongsheng Li, and Bao Wang. Finite-time analysis of adaptive temporal difference learning with deep neural networks. *Advances in Neural Information Processing Systems*, 35:19592–19604, 2022.

Zachary Sunberg and Mykel J Kochenderfer. Improving automated driving through pomdp planning with human internal states. *IEEE Transactions on Intelligent Transportation Systems*, 23(11): 20073–20083, 2022.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

Nick Whiteley. Dimension-free wasserstein contraction of nonlinear filters. *Stochastic Processes and their Applications*, 135:31–50, 2021.

Daan Wierstra, Alexander Foerster, Jan Peters, and Juergen Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *International conference on artificial neural networks*, pp. 697–706. Springer, 2007.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025a.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. Deep network approximation: Beyond relu to diverse activation functions. *Journal of Machine Learning Research*, 25(35):1–39, 2024.

Zhiyao Zhang, Myeung Suk Oh, FNU Hairi, Ziyue Luo, Alvaro Velasquez, and Jia Liu. Finite-time global optimality convergence in deep neural actor-critic methods for decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:2505.18433*, 2025b.

APPENDIX

A PROOF OF POMDP-TO-SMDP APPROXIMATION

A.1 PRELIMINARIES

Lemma 3 (Superstate-belief approximation)**.** *Under Assumption 2, for every history H and $B = \mathcal{G}(H)$,*

$$\|b_H - b_B\|_{\mathrm{TV}} \le (1-\rho)^{\ell},$$

where b_B is the fixed Bayes-filter belief defined in Section 3.

Proof. If $|B| < \ell$, no prefix has been discarded. Thus, $H = B$ and $b_H = b_B$. If $|B| = \ell$, the finite sequence B can itself be considered as an admissible history initialized from $\mathcal{D}$. Moreover, $\mathcal{G}(B) = B = \mathcal{G}(H)$. Applying Assumption 2 directly to the histories H and B gives

$$\|b_H - b_B\|_{\mathrm{TV}} \le (1-\rho)^{\ell}.$$

□

Lemma 4 (Reward and transition mismatch under filter stability)**.** *Under Assumptions 1 and 2, for every history-action pair (H, a), letting $B = \mathcal{G}(H)$, we have*

$$|\mathbb{E}[r_k \mid H_k = H, a_k = a] - \tilde{r}(B, a)| \le 2r_{\max}(1-\rho)^{\ell}.$$

Moreover, for any superstate policy π, if $a_{k+1} \sim \pi(\cdot \mid B_{k+1})$, then

$$\Big\|\mathbb{P}\big((B_{k+1}, a_{k+1}) \in \cdot \mid H_k = H, a_k = a\big) - \widetilde{\mathcal{P}}^{\pi}(\cdot \mid B, a)\Big\|_{\mathrm{TV}} \le (1-\rho)^{\ell}.$$

Proof. Fix any history-action pair (H, a) and let $B = \mathcal{G}(H)$. For the reward mismatch, the POMDP-generated reward satisfies $\mathbb{E}[r_k \mid H_k = H, a_k = a] = \sum_s b_H(s) r(s, a)$, whereas the superstate reward is $\tilde{r}(B, a) = \sum_s b_B(s) r(s, a)$. Therefore,

$$\begin{aligned}
|\mathbb{E}[r_k \mid H_k = H, a_k = a] - \tilde{r}(B, a)| &= \left|\sum_s \big(b_H(s) - b_B(s)\big) r(s, a)\right| \\
&\le r_{\max} \sum_s |b_H(s) - b_B(s)| \\
&= 2r_{\max}\|b_H - b_B\|_{\mathrm{TV}} \\
&\le 2r_{\max}(1-\rho)^{\ell},
\end{aligned}$$

where the last inequality follows from Lemma 3.

For a fixed policy π, define the superstate-action transition kernel $\widetilde{\mathcal{P}}^{\pi}(B', a' \mid B, a) := \widetilde{\mathcal{P}}(B' \mid B, a)\pi(a' \mid B')$. For the transition mismatch, recall that the POMDP-generated next-superstate transition and the superstate MDP transition are

$$\mathbb{P}(B_{k+1} = B' \mid H_k = H, a_k = a) = \sum_{o,s,s'} \mathbf{1}\{\mathcal{G}(H \parallel \{a, o\}) = B'\}\Phi(o \mid s')P(s' \mid s, a)b_H(s),$$

$$\widetilde{\mathcal{P}}(B' \mid B, a) = \sum_{o,s,s'} \mathbf{1}\{\mathcal{G}(B \parallel \{a, o\}) = B'\}\Phi(o \mid s')P(s' \mid s, a)b_B(s).$$

Since $B = \mathcal{G}(H)$, appending (a, o) and then truncating gives $\mathcal{G}(H \parallel \{a, o\}) = \mathcal{G}(B \parallel \{a, o\})$. Hence the two kernels differ only through the belief terms b_H and b_B. After sampling $a_{k+1} \sim \pi(\cdot \mid B_{k+1})$, for every (B', a') we have

$$\begin{aligned}
&\mathbb{P}(B_{k+1} = B', a_{k+1} = a' \mid H_k = H, a_k = a) - \widetilde{\mathcal{P}}^{\pi}(B', a' \mid B, a) \\
&\quad = \pi(a' \mid B') \sum_{o,s,s'} \mathbf{1}\{\mathcal{G}(B \parallel \{a, o\}) = B'\}\Phi(o \mid s')P(s' \mid s, a)\big(b_H(s) - b_B(s)\big).
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\Big\|\mathbb{P}\big((B_{k+1}, a_{k+1}) \in \cdot \mid H_k = H, a_k = a\big) - \widetilde{\mathcal{P}}^{\pi}(\cdot \mid B, a)\Big\|_{\mathrm{TV}} \\
&\quad = \frac{1}{2} \sum_{B', a'} \Big|\mathbb{P}(B_{k+1} = B', a_{k+1} = a' \mid H_k = H, a_k = a) - \widetilde{\mathcal{P}}^{\pi}(B', a' \mid B, a)\Big| \\
&\quad \le \frac{1}{2} \sum_{B', a'} \pi(a' \mid B') \sum_{o, s, s'} \mathbf{1}\{\mathcal{G}(B \parallel \{a, o\}) = B'\} \Phi(o \mid s') P(s' \mid s, a) |b_H(s) - b_B(s)| \\
&\quad = \frac{1}{2} \sum_{s} |b_H(s) - b_B(s)| \sum_{o, s'} \Phi(o \mid s') P(s' \mid s, a) \sum_{B'} \mathbf{1}\{\mathcal{G}(B \parallel \{a, o\}) = B'\} \sum_{a'} \pi(a' \mid B') \\
&\quad = \frac{1}{2} \sum_{s} |b_H(s) - b_B(s)| \\
&\quad = \|b_H - b_B\|_{\mathrm{TV}} \\
&\le (1 - \rho)^{\ell},
\end{aligned}
$$

where the last inequality follows from Lemma 3. □

A.2 Proof of Theorem 1

Proof. Let π^* be an unrestricted optimal policy of the original POMDP. Let $\widetilde{\pi}^*$ be an optimal policy of the induced superstate MDP. Define $\Delta := \sup_{H \in \mathbb{H}} |V^{\pi^*}(H) - \tilde{V}^{\widetilde{\pi}^*}(\mathcal{G}(H))|$. For any history H with $B = \mathcal{G}(H)$, the Bellman optimality equations give

$$
\begin{aligned}
\big|V^{\pi^*}(H) - \tilde{V}^{\widetilde{\pi}^*}(B)\big| \le \max_{a \in \mathcal{A}} \Big\{ &\big|\mathbb{E}[r_k \mid H_k = H, a_k = a] - \tilde{r}(B, a)\big| \\
&+ \gamma \big|\mathbb{E}[V^{\pi^*}(H_{k+1}) \mid H_k = H, a_k = a] - \mathbb{E}_{\widetilde{\mathcal{P}}(\cdot \mid B, a)}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1})]\big| \Big\}.
\end{aligned}
$$

The inequality uses $|\max_a f_a - \max_a g_a| \le \max_a |f_a - g_a|$ and the triangle inequality. For each action a, we add and subtract the conditional expectation $\mathbb{E}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1}) \mid H_k = H, a_k = a]$ under the original POMDP to obtain

$$
\begin{aligned}
&\mathbb{E}[V^{\pi^*}(H_{k+1}) \mid H_k = H, a_k = a] - \mathbb{E}_{\widetilde{\mathcal{P}}(\cdot \mid B, a)}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1})] \\
&= \mathbb{E}[V^{\pi^*}(H_{k+1}) \mid H_k = H, a_k = a] - \mathbb{E}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1}) \mid H_k = H, a_k = a] \\
&\quad + \mathbb{E}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1}) \mid H_k = H, a_k = a] - \mathbb{E}_{\widetilde{\mathcal{P}}(\cdot \mid B, a)}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1})].
\end{aligned}
$$

The triangle inequality and $|\mathbb{E}[X]| \le \mathbb{E}[|X|]$ therefore imply

$$
\begin{aligned}
\big|V^{\pi^*}(H) - \tilde{V}^{\widetilde{\pi}^*}(B)\big| \le \max_{a \in \mathcal{A}} \Big\{ &\big|\mathbb{E}[r_k \mid H_k = H, a_k = a] - \tilde{r}(B, a)\big| \\
&+ \gamma \mathbb{E}\Big[\big|V^{\pi^*}(H_{k+1}) - \tilde{V}^{\widetilde{\pi}^*}(B_{k+1})\big| \mid H_k = H, a_k = a\Big] \\
&+ \gamma \big|\mathbb{E}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1}) \mid H_k = H, a_k = a] - \mathbb{E}_{\widetilde{\mathcal{P}}(\cdot \mid B, a)}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1})]\big| \Big\}.
\end{aligned}
$$

The reward, continuation-value, and transition terms are bounded separately as follows:

$$
\begin{aligned}
&\big|\mathbb{E}[r_k \mid H_k = H, a_k = a] - \tilde{r}(B, a)\big| \le 2 r_{\max} (1 - \rho)^{\ell}, \\
&\mathbb{E}\Big[\big|V^{\pi^*}(H_{k+1}) - \tilde{V}^{\widetilde{\pi}^*}(B_{k+1})\big| \mid H_k = H, a_k = a\Big] \le \Delta, \\
&\big|\mathbb{E}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1}) \mid H_k = H, a_k = a] - \mathbb{E}_{\widetilde{\mathcal{P}}(\cdot \mid B, a)}[\tilde{V}^{\widetilde{\pi}^*}(B_{k+1})]\big| \\
&\qquad \le 2 \|\tilde{V}^{\widetilde{\pi}^*}\|_{\infty} \big\|\mathbb{P}(B_{k+1} \in \cdot \mid H_k = H, a_k = a) - \widetilde{\mathcal{P}}(\cdot \mid B, a)\big\|_{\mathrm{TV}} \\
&\qquad \le 2 \|\tilde{V}^{\widetilde{\pi}^*}\|_{\infty} (1 - \rho)^{\ell}.
\end{aligned}
$$

The reward bound follows from Lemma 4. The continuation-value bound follows from the definition of Δ. For the transition term, we use $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \le 2 \|f\|_{\infty} \|P - Q\|_{\mathrm{TV}}$ and marginalize the next action in Lemma 4. Substituting these bounds yields

$$
\big|V^{\pi^*}(H) - \tilde{V}^{\widetilde{\pi}^*}(B)\big| \le 2 r_{\max} (1 - \rho)^{\ell} + \gamma \Delta + 2 \gamma \|\tilde{V}^{\widetilde{\pi}^*}\|_{\infty} (1 - \rho)^{\ell}.
$$

Since $\|\tilde{V}^{\widetilde{\pi}^*}\|_\infty \le r_{\max}/(1-\gamma)$, taking the supremum over H gives

$$\Delta \le \gamma\Delta + 2r_{\max}(1-\rho)^\ell + \frac{2\gamma r_{\max}}{1-\gamma}(1-\rho)^\ell = \gamma\Delta + \frac{2r_{\max}}{1-\gamma}(1-\rho)^\ell.$$

Rearranging proves the desired bound for every H and $B = \mathcal{G}(H)$:

$$\left|V^{\pi^*}(H) - \tilde{V}^{\widetilde{\pi}^*}(B)\right| \le \Delta \le \frac{2r_{\max}}{(1-\gamma)^2}(1-\rho)^\ell. \qquad \square$$

B Full-Reset Occupancy and Stationarity

Lemma 5 (Exact occupancy-stationarity identity)**.** *Fix a superstate policy π. Let*

$$d^\pi(B) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(B_t = B \mid B_0 = \varnothing, \pi)$$

be the normalized discounted visitation distribution of the original superstate process. Then d^π is the unique stationary distribution of the full-reset process

$$\mathcal{P}^\pi(B' \mid B) = (1-\gamma)\delta_\varnothing(B') + \gamma\sum_{a\in\mathcal{A}}\pi(a \mid B)\widetilde{\mathcal{P}}(B' \mid B, a).$$

Consequently, $\mu^\pi(B,a) := d^\pi(B)\pi(a \mid B)$ is the stationary distribution of the superstate-action chain with full resets.

Proof. All time-indexed probabilities below refer to the original superstate process without resets. First, d^π is normalized because

$$\begin{aligned}\sum_{B\in\mathsf{B}} d^\pi(B) &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\sum_{B\in\mathsf{B}}\Pr(B_t = B \mid B_0 = \varnothing, \pi)\\ &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t = 1.\end{aligned}$$

For every $t \ge 0$, the Markov property of the original SMDP and the law of total probability give

$$\Pr(B_{t+1} = B' \mid B_0 = \varnothing, \pi) = \sum_{B\in\mathsf{B}}\Pr(B_t = B \mid B_0 = \varnothing, \pi)\sum_{a\in\mathcal{A}}\pi(a \mid B)\widetilde{\mathcal{P}}(B' \mid B, a).$$

This sums over the current superstate and the action selected there. Substituting the restart kernel separates the reset and non-reset terms. Substituting d^π into the latter term and applying the preceding identity then yields

$$\begin{aligned}\sum_{B\in\mathsf{B}} d^\pi(B)\mathcal{P}^\pi(B' \mid B) &= (1-\gamma)\delta_\varnothing(B')\sum_{B\in\mathsf{B}} d^\pi(B) + \gamma\sum_{B\in\mathsf{B}}\sum_{a\in\mathcal{A}} d^\pi(B)\pi(a \mid B)\widetilde{\mathcal{P}}(B' \mid B, a)\\ &= (1-\gamma)\delta_\varnothing(B') + \gamma(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\Pr(B_{t+1} = B' \mid B_0 = \varnothing, \pi)\\ &= (1-\gamma)\delta_\varnothing(B') + (1-\gamma)\sum_{t=1}^{\infty}\gamma^t\Pr(B_t = B' \mid B_0 = \varnothing, \pi)\\ &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\Pr(B_t = B' \mid B_0 = \varnothing, \pi)\\ &= d^\pi(B').\end{aligned}$$

Hence $d^\pi\mathcal{P}^\pi = d^\pi$. Since the next action is sampled from $\pi(\cdot \mid B')$, the same identity gives stationarity of $\mu^\pi(B,a) = d^\pi(B)\pi(a \mid B)$ for the state-action process.

It remains to prove uniqueness. Let λ and λ' be any two probability distributions on B. For this superstate chain, write $\widetilde{\mathcal{P}}^\pi(B' \mid B) := \sum_{a\in\mathcal{A}} \pi(a \mid B)\widetilde{\mathcal{P}}(B' \mid B, a)$.

$$\begin{aligned}\|\lambda\mathcal{P}^\pi - \lambda'\mathcal{P}^\pi\|_{\mathrm{TV}} &= \left\|(1-\gamma)\delta_\varnothing + \gamma\lambda\widetilde{\mathcal{P}}^\pi - (1-\gamma)\delta_\varnothing - \gamma\lambda'\widetilde{\mathcal{P}}^\pi\right\|_{\mathrm{TV}} \\ &= \gamma\left\|\lambda\widetilde{\mathcal{P}}^\pi - \lambda'\widetilde{\mathcal{P}}^\pi\right\|_{\mathrm{TV}} \\ &\le \gamma\|\lambda - \lambda'\|_{\mathrm{TV}}.\end{aligned}$$

The inequality follows from the nonexpansiveness of a Markov kernel in total variation. If d_1 and d_2 were stationary, then

$$\|d_1 - d_2\|_{\mathrm{TV}} = \|d_1\mathcal{P}^\pi - d_2\mathcal{P}^\pi\|_{\mathrm{TV}} \le \gamma\|d_1 - d_2\|_{\mathrm{TV}}.$$

Since $\gamma < 1$, this implies $d_1 = d_2$. □

B.1 Proof of Lemma 1

Proof. For the full-reset superstate-action chain, define the transition kernel

$$\mathcal{K}^\pi(B', a' \mid B, a) := \left[(1-\gamma)\delta_\varnothing(B') + \gamma\widetilde{\mathcal{P}}(B' \mid B, a)\right]\pi(a' \mid B').$$

By Lemma 5, $\mu^\pi(B, a) = d^\pi(B)\pi(a \mid B)$ is stationary for $\mathcal{K}^\pi$. Let λ and λ' be any two probability distributions on $\mathsf{B} \times \mathcal{A}$. Define

$$\nu^\pi(B', a') := \delta_\varnothing(B')\pi(a' \mid \varnothing), \qquad \widetilde{\mathcal{K}}^\pi(B', a' \mid B, a) := \widetilde{\mathcal{P}}(B' \mid B, a)\pi(a' \mid B').$$

Then $\mathcal{K}^\pi = (1-\gamma)\nu^\pi + \gamma\widetilde{\mathcal{K}}^\pi$, where the reset distribution ν^π is independent of the current state-action pair. Therefore,

$$\begin{aligned}\|\lambda\mathcal{K}^\pi - \lambda'\mathcal{K}^\pi\|_{\mathrm{TV}} &= \gamma\|\lambda\widetilde{\mathcal{K}}^\pi - \lambda'\widetilde{\mathcal{K}}^\pi\|_{\mathrm{TV}} \\ &\le \gamma\|\lambda - \lambda'\|_{\mathrm{TV}},\end{aligned}$$

where the inequality follows from the nonexpansiveness of a Markov kernel in total variation. This contraction also implies that μ^π is the unique stationary distribution of $\mathcal{K}^\pi$. Iterating this contraction and using stationarity gives, for every $(B, a) \in \mathsf{B} \times \mathcal{A}$ and $k \ge 0$,

$$\begin{aligned}\left\|\delta_{(B,a)}(\mathcal{K}^\pi)^k - \mu^\pi\right\|_{\mathrm{TV}} &= \left\|\delta_{(B,a)}(\mathcal{K}^\pi)^k - \mu^\pi(\mathcal{K}^\pi)^k\right\|_{\mathrm{TV}} \\ &\le \gamma^k\left\|\delta_{(B,a)} - \mu^\pi\right\|_{\mathrm{TV}} \le \gamma^k.\end{aligned}$$

Taking the supremum over (B, a) proves Lemma 1. The bound is uniform over policies and memory lengths. □

B.2 The actual full-reset POMDP sampling chain

Lemma 6 (Actual mixing and SMDP occupancy mismatch)**.** *Fix $\pi(\cdot \mid B)$ during a critic run. The actual process $X_k = (S_k, B_k, a_k)$ is a finite Markov chain under the joint reset protocol. Write ξ^π for its stationary distribution and μ^π_{P} for its (B, a) marginal. Then*

$$\sup_x \|\mathcal{L}(X_k \mid X_0 = x) - \xi^\pi\|_{\mathrm{TV}} \le \gamma^k.$$

Under Lemma 4,

$$\|\mu^\pi_{\mathrm{P}} - \mu^\pi\|_{\mathrm{TV}} \le \min\left\{1, \frac{\gamma}{1-\gamma}(1-\rho)^\ell\right\}.$$

Proof. Mixing of the actual sampling chain. On a reset, draw $S' \sim \mathcal{D}$ and set $B' = \varnothing$. Then draw $a' \sim \pi(\cdot \mid \varnothing)$. Without a reset, generate the next state and observation using P and Φ. Update the window and sample the next action from $\pi(\cdot \mid B')$. Thus, the transition kernel on (s, B, a) is

$$\mathcal{K}^\pi(s', B', a' \mid s, B, a) = (1-\gamma)\mathcal{D}(s')\delta_\varnothing(B')\pi(a' \mid \varnothing) + \gamma P(s' \mid s, a)\pi(a' \mid B')\sum_o \Phi(o \mid s')\mathbf{1}\{\mathcal{G}(B\|(a, o)) = B'\}.$$

The reset distribution $\mathcal{D}(s')\delta_{\varnothing}(B')\pi(a' \mid \varnothing)$ does not depend on the previous (s, B, a). For any two probability distributions λ and λ' on the augmented state space, their reset contributions satisfy

$$(1-\gamma)\mathcal{D}(s')\delta_{\varnothing}(B')\pi(a' \mid \varnothing)\underbrace{\left(\sum_x \lambda(x) - \sum_x \lambda'(x)\right)}_{1-1=0} = 0.$$

The remaining transition follows the original dynamics with probability γ. Applying the TV nonexpansiveness argument in Lemma 5 gives

$$\|\lambda\mathcal{K}^\pi - \lambda'\mathcal{K}^\pi\|_{\mathrm{TV}} \le \gamma\|\lambda - \lambda'\|_{\mathrm{TV}}.$$

A stationary distribution ξ^π exists because the augmented state space is finite. The contraction makes it unique. Iterating the bound gives

$$\|\mathcal{L}(X_k \mid X_0 = x) - \xi^\pi\|_{\mathrm{TV}} \le \gamma^k\|\delta_x - \xi^\pi\|_{\mathrm{TV}} \le \gamma^k.$$

Comparison of the original processes. For the following comparison, both processes run without resets. Write $\mathcal{L}_{\mathrm{P}}(B_j, a_j)$ and $\mathcal{L}_{\mathrm{SMDP}}(B_j, a_j)$ for their window-action distributions. Both start with $B_0 = \varnothing$ and $a_0 \sim \pi(\cdot \mid \varnothing)$. The POMDP additionally starts with $S_0 \sim \mathcal{D}$. The product with the SMDP kernel propagates the current POMDP distribution through one SMDP transition:

$$\left[\mathcal{L}_{\mathrm{P}}(B_j, a_j)\widetilde{\mathcal{P}}^\pi\right](B', a') = \sum_{B,a} \Pr_{\mathrm{P}}(B_j = B, a_j = a)\widetilde{\mathcal{P}}^\pi(B', a' \mid B, a).$$

By Lemma 4, the next-pair distribution conditional on each actual history and action differs from $\widetilde{\mathcal{P}}^\pi(\cdot \mid B_j, a_j)$ by at most $(1-\rho)^\ell$ in TV. Since $B_j = \mathcal{G}(H_j)$, averaging over those histories and actions gives

$$\begin{aligned}\left\|\mathcal{L}_{\mathrm{P}}(B_{j+1}, a_{j+1}) - \mathcal{L}_{\mathrm{P}}(B_j, a_j)\widetilde{\mathcal{P}}^\pi\right\|_{\mathrm{TV}} &= \left\|\mathbb{E}_{\mathrm{P}}\left[\mathcal{L}_{\mathrm{P}}(B_{j+1}, a_{j+1} \mid H_j, a_j) - \widetilde{\mathcal{P}}^\pi(\cdot \mid B_j, a_j)\right]\right\|_{\mathrm{TV}} \\ &\le \mathbb{E}_{\mathrm{P}}\left[\left\|\mathcal{L}_{\mathrm{P}}(B_{j+1}, a_{j+1} \mid H_j, a_j) - \widetilde{\mathcal{P}}^\pi(\cdot \mid B_j, a_j)\right\|_{\mathrm{TV}}\right] \\ &\le (1-\rho)^\ell.\end{aligned}$$

The equality uses the law of total probability. The first inequality uses convexity of TV and the second uses Lemma 4. The SMDP evolves exactly according to its Markov kernel:

$$\mathcal{L}_{\mathrm{SMDP}}(B_{j+1}, a_{j+1}) = \mathcal{L}_{\mathrm{SMDP}}(B_j, a_j)\widetilde{\mathcal{P}}^\pi.$$

Adding and subtracting $\mathcal{L}_{\mathrm{P}}(B_j, a_j)\widetilde{\mathcal{P}}^\pi$ yields the following bound by the triangle inequality and TV nonexpansiveness:

$$\|\mathcal{L}_{\mathrm{P}}(B_{j+1}, a_{j+1}) - \mathcal{L}_{\mathrm{SMDP}}(B_{j+1}, a_{j+1})\|_{\mathrm{TV}} \le (1-\rho)^\ell + \|\mathcal{L}_{\mathrm{P}}(B_j, a_j) - \mathcal{L}_{\mathrm{SMDP}}(B_j, a_j)\|_{\mathrm{TV}}.$$

The initial window-action distributions coincide. Induction therefore yields

$$\|\mathcal{L}_{\mathrm{P}}(B_j, a_j) - \mathcal{L}_{\mathrm{SMDP}}(B_j, a_j)\|_{\mathrm{TV}} \le j(1-\rho)^\ell.$$

Only the SMDP window-action process is treated as Markov in this comparison. The POMDP bound is obtained by conditioning on full histories.

From time marginals to stationary distributions. Apply the occupancy-stationarity calculation of Lemma 5 to $\mathcal{K}^\pi$ and then marginalize out the latent state. Together with the SMDP identity in that lemma, this gives

$$\mu^\pi_{\mathrm{P}} = (1-\gamma)\sum_{j=0}^{\infty}\gamma^j\mathcal{L}_{\mathrm{P}}(B_j, a_j),$$

$$\mu^\pi = (1-\gamma)\sum_{j=0}^{\infty}\gamma^j\mathcal{L}_{\mathrm{SMDP}}(B_j, a_j).$$

Hence,

$$\begin{aligned}\|\mu^{\pi}_{\mathrm{P}} - \mu^{\pi}\|_{\mathrm{TV}} &\le (1-\gamma)\sum_{j=0}^{\infty}\gamma^{j}\,\|\mathcal{L}_{\mathrm{P}}(B_j,a_j) - \mathcal{L}_{\mathrm{SMDP}}(B_j,a_j)\|_{\mathrm{TV}}\\ &\le (1-\gamma)\sum_{j=0}^{\infty}\gamma^{j} j(1-\rho)^{\ell}\\ &= \frac{\gamma}{1-\gamma}(1-\rho)^{\ell}.\end{aligned}$$

The TV distance between two probability distributions is also at most one. Combining the two bounds proves the stated minimum. □

C Proof of Critic Convergence

C.1 Preliminaries

Goal for Critic. Fix a policy π. Let $\widehat{Q}_k(\cdot,\cdot) = \widehat{Q}(\cdot,\cdot;\bar{W}_k)$ denote the critic estimate produced at iteration k, let $\tilde{Q}^{\pi}$ be the Bellman fixed point of the original superstate MDP under π, and let μ^{π} denote the corresponding stationary state-action distribution of the superstate MDP with full resets (which is also the discounted visitation distribution of original superstate MDP). Our goal is to bound the expected critic error under μ^{π}:

$$\mathbb{E}\big[\|\widehat{Q}_k - \tilde{Q}^{\pi}\|^2_{\mu^{\pi}}\big] = \mathbb{E}\Big[\mathbb{E}_{(B,a)\sim\mu^{\pi}}\big[(\widehat{Q}_k(B,a) - \tilde{Q}^{\pi}(B,a))^2\big]\Big]. \tag{2}$$

The main difficulty is that $\widehat{W}_k$ is not updated using samples drawn directly from the induced superstate MDP. Instead, the critic $\widehat{Q}_k(\cdot,\cdot)$ is trained using trajectories generated by the underlying POMDP. We evaluate the error under μ^{π}, because $\tilde{Q}^{\pi}$ is the Bellman fixed point of the superstate MDP and μ^{π} is the stationary distribution naturally associated with this target.

Data-generating model for $\widehat{Q}_k$. The estimate $\widehat{Q}_k$ is obtained from trajectory samples $(H_k, a_k, r_k, H_{k+1}, a_{k+1})$, which are mapped into the superstate representation $(B_k, a_k, r_k, B_{k+1}, a_{k+1})$ with $B_k := \mathcal{G}(H_k)$. These samples are generated by the underlying POMDP dynamics rather than by the superstate MDP kernel. The reward expectation and next-superstate transition induced by the POMDP are given in Lemma 4.

Locally linearized critic and projected Bellman fixed point. Following Cai et al. (2019), we introduce the locally linearized critic $\widehat{Q}_0$ around the initialization $\widehat{W}_0$:

$$\widehat{Q}_0(B,a;W) := \widehat{Q}(B,a;\widehat{W}_0) + \big\langle \nabla_W \widehat{Q}(B,a;\widehat{W}_0),\, W - \widehat{W}_0\big\rangle.$$

Let $\mathcal{B} := \Big\{W : \|W^{(h)} - \widehat{W}_0^{(h)}\|_F \le R, \forall h \in [d]\Big\}$ and $\mathcal{F} := \{\widehat{Q}_0(\cdot,\cdot;W) : W \in \mathcal{B}\}$, then $\|W - \widehat{W}_0\|_2^2 = \sum_{h=1}^{d}\|W^{(h)} - \widehat{W}_0^{(h)}\|_F^2 \le dR^2$. Let $\Pi_{\mathcal{F}}$ denote the projection operator onto $\mathcal{F}$, defined by $\Pi_{\mathcal{F}} f := \arg\min_{g\in\mathcal{F}}\|f-g\|^2_{\mu^{\pi}}$. Lemma 5 gives $\mu^{\pi}(B,a) = (1-\gamma)\delta_{\varnothing}(B)\pi(a \mid \varnothing) + \gamma(\mu^{\pi}\widetilde{\mathcal{P}}^{\pi})(B,a)$. Hence $\mu^{\pi}\widetilde{\mathcal{P}}^{\pi} \le \mu^{\pi}/\gamma$. Since $\mathcal{F}$ is closed and convex, projection nonexpansiveness and Jensen's inequality yield

$$\begin{aligned}\|\Pi_{\mathcal{F}}\tilde{T}^{\pi} f - \Pi_{\mathcal{F}}\tilde{T}^{\pi} g\|^2_{\mu^{\pi}} &\le \|\tilde{T}^{\pi} f - \tilde{T}^{\pi} g\|^2_{\mu^{\pi}}\\ &\le \gamma^2\|f-g\|^2_{\mu^{\pi}\widetilde{\mathcal{P}}^{\pi}}\\ &\le \gamma\|f-g\|^2_{\mu^{\pi}}.\end{aligned}$$

Taking square roots shows that $\Pi_{\mathcal{F}}\tilde{T}^{\pi}$ is a $\sqrt{\gamma}$-contraction in $L^2(\mu^{\pi})$. Therefore, $Q^{*,\pi}_{\mathcal{F}}$ is the unique projected Bellman fixed point in $\mathcal{F}$ which satisfies $Q^{*,\pi}_{\mathcal{F}} = \Pi_{\mathcal{F}}\tilde{T}^{\pi}Q^{*,\pi}_{\mathcal{F}}$, and there exists $W^* \in \mathcal{B}$ such that $\widehat{Q}_0(\cdot,\cdot;W^*) = Q^{*,\pi}_{\mathcal{F}}$.

Fact 1. *According to Cai et al. (2019)'s Lemmas F.1 and F.2, let* $m = \Omega\Big(\frac{p^{3/2}\log^{3/2}(m^{1/2}/R)}{Rd^{3/2}}\Big)$, $R = O\big(m^{1/2}d^{-6}\log^{-3} m\big)$, *with probability at least* $1 - e^{-\Omega(\log^2 m)}$ *over the random initialization, the following hold for every admissible superstate-action pair* (B, a) *satisfying* $\|\psi(B,a)\|_2 = 1$ *and every* $W \in \mathcal{B}$:

$$\begin{aligned}
\|\nabla_W \widehat{Q}(B,a;W) - \nabla_W \widehat{Q}_0(B,a;W)\|_2 &= O\left(R^{1/3}m^{-1/6}d^{5/2}\log^{1/2} m\right)\\
\|\nabla_W \widehat{Q}(B,a;W)\|_2 &= O(d)\\
|\widehat{Q}(B,a;\widehat{W}_0)| &= O(\log m)\\
|\widehat{Q}(B,a;W) - \widehat{Q}_0(B,a;W)| &= O\left(R^{4/3}m^{-1/6}d^{3}\log^{1/2} m\right)
\end{aligned}$$

Fact 1 depends only on the neural network architecture and the input feature $x = \psi(B,a)$, through the condition $\|x\|_2 = 1$, and is independent of the distribution from which (B,a) is sampled to input and calculate W.

Lemma 7 (Lipschitz and bounded critic)**.** *Under the same condition of Fact 1, with probability at least* $1-e^{-\Omega(\log^2 m)}$ *over the random initialization, there exist* $L_Q = O(d)$ *and* $Q_{\max} = O(\log m + d^{3/2}R)$ *such that for all* $W, W' \in \mathcal{B}$ *and* (B,a) *satisfying* $\|\psi(B,a)\|_2 = 1$*:*

$$\begin{aligned}
\|\nabla_W \widehat{Q}(B,a;W)\|_2 &\le L_Q,\\
|\widehat{Q}(B,a;W) - \widehat{Q}(B,a;W')| &\le L_Q\|W - W'\|_2,\\
|\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W')| &\le L_Q\|W - W'\|_2,\\
\max\{|\widehat{Q}(B,a;W)|, |\widehat{Q}_0(B,a;W)|\} &\le Q_{\max}.
\end{aligned}$$

Proof. By Fact 1, $\|\nabla_W \widehat{Q}(B,a;W)\|_2 = O(d)$, which allows us to set $L_Q = \sup_{W\in\mathcal{B},(B,a)} \|\nabla_W \widehat{Q}(B,a;W)\|_2$. Since $\mathcal{B}$ is convex, we have $|\widehat{Q}(B,a;W) - \widehat{Q}(B,a;W')| \le L_Q\|W - W'\|_2$. For the linearized critic $\widehat{Q}_0$, the difference can be expressed and bounded as follows:

$$\begin{aligned}
|\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W')| &= |\langle \nabla_W \widehat{Q}(B,a;\widehat{W}_0), W - W'\rangle|\\
&\le \|\nabla_W \widehat{Q}(B,a;\widehat{W}_0)\|_2\|W - W'\|_2\\
&\le L_Q\|W - W'\|_2,
\end{aligned}$$

For boundedness, Fact 1 implies $|\widehat{Q}(B,a;\widehat{W}_0)| = O(\log m)$. Given $W \in \mathcal{B}$, we have $\|W - \widehat{W}_0\|_2 \le \sqrt{d}R$. Hence, by the Lipschitz property,

$$\begin{aligned}
|\widehat{Q}(B,a;W)| &\le |\widehat{Q}(B,a;\widehat{W}_0)| + L_Q\|W - \widehat{W}_0\|_2\\
&\le O(\log m) + O(d)\cdot\sqrt{d}R\\
&= O(\log m + d^{3/2}R).
\end{aligned}$$

Similarly, by the linearized definition $\widehat{Q}_0(B,a;W) = \widehat{Q}(B,a;\widehat{W}_0) + \big\langle \nabla_W \widehat{Q}(B,a;\widehat{W}_0),\, W - \widehat{W}_0\big\rangle$, we have

$$\begin{aligned}
|\widehat{Q}_0(B,a;W)| &\le |\widehat{Q}(B,a;\widehat{W}_0)| + \|\nabla_W \widehat{Q}(B,a;\widehat{W}_0)\|_2\|W - \widehat{W}_0\|_2\\
&\le O(\log m) + O(d)\cdot\sqrt{d}R\\
&= O(\log m + d^{3/2}R).
\end{aligned}$$

which completes the proof by taking $Q_{\max} = O(\log m + d^{3/2}R)$. □

C.2 Proof of Theorem 2

We decompose the error in equation 2 by:

$$\begin{aligned}\mathbb{E}\big[\|\widehat{Q}(\cdot,\cdot;\bar{W}_K)-\tilde{Q}^{\pi}\|_{\mu^\pi}^2\big] \le\ & 3\mathbb{E}\underbrace{\big[\|\widehat{Q}(\cdot,\cdot;\bar{W}_K)-\widehat{Q}_0(\cdot,\cdot;\bar{W}_K)\|_{\mu^\pi}^2\big]}_{\text{NTK linearization error}}\\ &+3\mathbb{E}\underbrace{\big[\|\widehat{Q}_0(\cdot,\cdot;\bar{W}_K)-Q_{\mathcal{F}}^{*,\pi}\|_{\mu^\pi}^2\big]}_{\text{projected optimization error}}\\ &+3\mathbb{E}\big[\underbrace{\|Q_{\mathcal{F}}^{*,\pi}-\tilde{Q}^{\pi}\|_{\mu^\pi}^2}_{\text{approximation error}}\big].\end{aligned} \tag{3}$$

C.2.1 Bound NTK linearization error

Since each iterate $\widehat{W}_k \in \mathcal{B}$ and $\mathcal{B}$ is convex, we have $\bar{W}_K \in \mathcal{B}$. Using Fact 1 with $W = \bar{W}_K$, we obtain that, with probability at least $1 - e^{-\Omega(\log^2 m)}$,

$$\mathbb{E}\big[\|\widehat{Q}(\cdot,\cdot;\bar{W}_K)-\widehat{Q}_0(\cdot,\cdot;\bar{W}_K)\|_{\mu^\pi}^2\big] = O\Big(R^{8/3}m^{-1/3}d^6\log m\Big).$$

C.2.2 Bound projected optimization error

Step 1. Lyapunov Analysis. We use Lyapunov analysis to control $\|\widehat{Q}_0(\cdot,\cdot;\widehat{W}_k)-\widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi}^2$. Define $e_k := \widehat{W}_k - W^*$. The projection ball gives $\|e_k\|_2 \le 2\sqrt{d}\,R$. Due to $\widehat{W}_{k+1} = \mathrm{Proj}_{\mathcal{B}}(\widehat{W}_k - \eta g_k(\widehat{W}_k))$ and the non-expansion property of projection, we can get

$$\|e_{k+1}\|_2^2 \le \|e_k\|_2^2 - 2\eta\langle g_k(\widehat{W}_k), e_k\rangle + \eta^2\|g_k(\widehat{W}_k)\|_2^2. \tag{4}$$

where the true POMDP stochastic semigradient $g_k(\widehat{W}_k)$ is defined by

$$g_k(W) := Z_{k+1}\delta_k(W)\nabla_W\widehat{Q}(B_k,a_k;W),$$

$$\delta_k(W) := \widehat{Q}(B_k,a_k;W) - [r_k + \gamma\widehat{Q}(B_{k+1},a_{k+1};W)].$$

Here Z_{k+1} is the Bernoulli variable in Algorithm 1. The displayed TD residual is evaluated only when $Z_{k+1} = 1$ and the successor follows the original POMDP dynamics. On $Z_{k+1} = 0$, set both TD residuals $\delta_k(W)$ and $\delta_{0,k}(W)$ to zero for notation. No critic update is performed on that branch.

According to Lemma 7 and Assumption 1,

$$\|g_k(\widehat{W}_k)\|_2^2 \le \left(r_{\max} + (1+\gamma)Q_{\max}\right)^2 L_Q^2 =: \delta_{\max}^2 L_Q^2, \qquad \delta_{\max} = O(\log m + d^{3/2}R).$$

For the analysis, we further introduce two semigradients based on the locally linearized critic $\widehat{Q}_0$. The first is the linearized POMDP semigradient

$$g_{0,k}(W) = Z_{k+1}\delta_{0,k}(W)\nabla_W\widehat{Q}_0(B_k,a_k),$$

$$\delta_{0,k}(W) = \widehat{Q}_0(B_k,a_k;W) - [r_k + \gamma\widehat{Q}_0(B_{k+1},a_{k+1};W)].$$

The second is the linearized semigradient under the induced superstate MDP:

$$g_0^{\mathrm{SMDP}}(B,a;W) := \delta_0^{\mathrm{SMDP}}(B,a;W)\nabla_W\widehat{Q}_0(B,a),$$

$$\delta_0^{\mathrm{SMDP}}(B,a;W) := \widehat{Q}_0(B,a;W) - \left[\tilde{r}(B,a) + \gamma\sum_{B',a'}\widetilde{\mathcal{P}}^{\pi}(B',a' \mid B,a)\widehat{Q}_0(B',a';W)\right].$$

Define the corresponding population semigradient

$$\bar{g}_0^{\mathrm{SMDP}}(W) := \mathbb{E}_{(B,a)\sim\mu^\pi}\left[g_0^{\mathrm{SMDP}}(B,a;W)\right].$$

Let $\mathcal{I}_k$ contain the initialization and the augmented chain $X_j = (S_j, B_j, a_j)$ through time k. It also contains the past rewards and the critic iterates. This filtration is used only in the analysis. Define the actual stationary population direction by

$$\bar{g}_0^{\mathrm{POMDP}}(W) := \mathbb{E}_{X_k \sim \xi^\pi}\left[\mathbb{E}\Big[\delta_{0,k}(W)\nabla_W \widehat{Q}_0(B_k, a_k) \mid X_k,\ Z_{k+1} = 1\Big]\right].$$

Here ξ^π is the stationary law from Lemma 6. This expectation uses a stationary current state and one original POMDP transition. Set $G_{\max} := \delta_{\max} L_Q$. Independence of the reset coin gives

$$\mathbb{E}[\|g_k(\widehat{W}_k)\|_2^2 \mid \mathcal{I}_k] \le \gamma G_{\max}^2.$$

We compare the two population directions before controlling temporal dependence.

For the full-reset sampling scheme, we decompose

$$\begin{aligned} g_k(\widehat{W}_k) - \gamma \bar{g}_0^{\mathrm{SMDP}}(\widehat{W}_k) = \underbrace{\big(g_k(\widehat{W}_k) - g_{0,k}(\widehat{W}_k)\big)}_{\nu_k \text{ (linearization error)}} \quad + \underbrace{\gamma\big(\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_k) - \bar{g}_0^{\mathrm{SMDP}}(\widehat{W}_k)\big)}_{\omega_k \text{ (population mismatch)}} \\ + \underbrace{\big(g_{0,k}(\widehat{W}_k) - \gamma \bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_k)\big)}_{\zeta_k \text{ (Markovian sampling error)}}. \end{aligned} \tag{5}$$

Substituting (5) into (4) gives

$$\begin{aligned} \|e_{k+1}\|_2^2 \le \|e_k\|_2^2 - 2\eta\gamma\langle \bar{g}_0^{\mathrm{SMDP}}(\widehat{W}_k), e_k\rangle - 2\eta\langle \nu_k, e_k\rangle - 2\eta\langle \omega_k, e_k\rangle \\ - 2\eta\langle \zeta_k, e_k\rangle + \eta^2 \|g_k(\widehat{W}_k)\|_2^2. \end{aligned} \tag{6}$$

Step 2. Main Descent Term $\langle \bar{g}_0^{\mathrm{SMDP}}(\widehat{W}_k), e_k\rangle$.

Lemma 8 (Projected stationary point and one-point monotonicity)**.** *Let $W^* \in \mathcal{B}$ satisfy $\widehat{Q}_0(\cdot,\cdot;W^*) = Q_{\mathcal{F}}^{*,\pi}$. Then, for all $W \in \mathcal{B}$,*

$$\langle \bar{g}_0^{\mathrm{SMDP}}(W^*),\, W - W^*\rangle \ge 0,$$

and

$$\left\langle \bar{g}_0^{\mathrm{SMDP}}(W) - \bar{g}_0^{\mathrm{SMDP}}(W^*),\, W - W^*\right\rangle \ge (1 - \sqrt{\gamma})\|\widehat{Q}_0(\cdot,\cdot;W) - \widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi}^2.$$

Proof. We first prove the stationary-point inequality. For any $W \in \mathcal{B}$, starting from the definition of $\bar{g}_0^{\mathrm{SMDP}}(W^*)$, we have

$$\begin{aligned} \langle \bar{g}_0^{\mathrm{SMDP}}(W^*),\, W - W^*\rangle &= \left\langle \mathbb{E}_{(B,a)\sim\mu^\pi}\left[\delta_0^{\mathrm{SMDP}}(B,a;W^*)\nabla_W \widehat{Q}_0(B,a)\right],\, W - W^*\right\rangle \\ &= \mathbb{E}_{(B,a)\sim\mu^\pi}\left[\delta_0^{\mathrm{SMDP}}(B,a;W^*)\langle \nabla_W \widehat{Q}_0(B,a),\, W - W^*\rangle\right] \\ &= \mathbb{E}_{(B,a)\sim\mu^\pi}\left[\delta_0^{\mathrm{SMDP}}(B,a;W^*)\big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big)\right] \\ &\ge 0. \end{aligned}$$

where the last inequality holds since $Q_{\mathcal{F}}^{*,\pi} = \widehat{Q}_0(\cdot,\cdot;W^*)$ and $Q_{\mathcal{F}}^{*,\pi} = \Pi_{\mathcal{F}}\tilde{T}^\pi Q_{\mathcal{F}}^{*,\pi}$, the first-order optimality condition of projection onto the closed convex set $\mathcal{F}$ yields

$$\left\langle Q_{\mathcal{F}}^{*,\pi} - \tilde{T}^\pi Q_{\mathcal{F}}^{*,\pi},\, \widehat{Q}_0(\cdot,\cdot;W) - Q_{\mathcal{F}}^{*,\pi}\right\rangle_{\mu^\pi} \ge 0.$$

Using $Q_{\mathcal{F}}^{*,\pi}(B,a) - (\tilde{T}^\pi Q_{\mathcal{F}}^{*,\pi})(B,a) = \delta_0^{\mathrm{SMDP}}(B,a;W^*)$, then we are done.

We next prove the one-point monotonicity inequality. We have

$$\begin{aligned}
&\langle \bar{g}_0^{\mathrm{SMDP}}(W) - \bar{g}_0^{\mathrm{SMDP}}(W^*),\, W - W^* \rangle \\
&= \Big\langle \mathbb{E}_{(B,a)\sim\mu^\pi}\Big[\big(\delta_0^{\mathrm{SMDP}}(B,a;W) - \delta_0^{\mathrm{SMDP}}(B,a;W^*)\big) \nabla_W \widehat{Q}_0(B,a) \Big],\, W - W^* \Big\rangle \\
&= \mathbb{E}_{(B,a)\sim\mu^\pi}\Big[\big(\delta_0^{\mathrm{SMDP}}(B,a;W) - \delta_0^{\mathrm{SMDP}}(B,a;W^*)\big) \langle \nabla_W \widehat{Q}_0(B,a),\, W - W^* \rangle \Big] \\
&= \mathbb{E}_{(B,a)\sim\mu^\pi}\Big[\big(\delta_0^{\mathrm{SMDP}}(B,a;W) - \delta_0^{\mathrm{SMDP}}(B,a;W^*)\big) \big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big) \Big].
\end{aligned}$$

Next, by the definition of δ_0^{SMDP},

$$\begin{aligned}
&\delta_0^{\mathrm{SMDP}}(B,a;W) - \delta_0^{\mathrm{SMDP}}(B,a;W^*) \\
&= \widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*) - \gamma \sum_{B',a'} \widetilde{\mathcal{P}}^\pi(B',a'|B,a) \big(\widehat{Q}_0(B',a';W) - \widehat{Q}_0(B',a';W^*)\big).
\end{aligned}$$

Substituting this identity into the previous display yields

$$\begin{aligned}
&\langle \bar{g}_0^{\mathrm{SMDP}}(W) - \bar{g}_0^{\mathrm{SMDP}}(W^*),\, W - W^* \rangle \\
&= \mathbb{E}_{\mu^\pi}\Big[\big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big)^2 \Big] \\
&\quad - \gamma\, \mathbb{E}_{\mu^\pi}\Big[\big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big)\, \mathbb{E}\Big[\widehat{Q}_0(B',a';W) - \widehat{Q}_0(B',a';W^*) \,\Big|\, B,a \Big] \Big] \\
&= \mathbb{E}_{\mu^\pi}\Big[\big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big)^2 \Big] \\
&\quad - \gamma \mathbb{E}_{\mu^\pi}\Big[\big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big) \big(\widehat{Q}_0(B',a';W) - \widehat{Q}_0(B',a';W^*)\big) \Big]
\end{aligned}$$

where the last step comes from the tower property.

The successor follows the original SMDP kernel. Its marginal is $\mu^\pi \widetilde{\mathcal{P}}^\pi$. By Lemma 5,

$$\mu^\pi(B',a') = (1-\gamma)\delta_{\varnothing}(B')\pi(a' \mid \varnothing) + \gamma(\mu^\pi \widetilde{\mathcal{P}}^\pi)(B',a'), \qquad \mu^\pi \widetilde{\mathcal{P}}^\pi \le \mu^\pi/\gamma.$$

Therefore, by Cauchy–Schwarz,

$$\begin{aligned}
&\mathbb{E}_{\mu^\pi}\Big[\big(\widehat{Q}_0(B,a;W) - \widehat{Q}_0(B,a;W^*)\big) \big(\widehat{Q}_0(B',a';W) - \widehat{Q}_0(B',a';W^*)\big) \Big] \\
&\le \|\widehat{Q}_0(\cdot,\cdot;W) - \widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi} \|\widehat{Q}_0(\cdot,\cdot;W) - \widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi \widetilde{\mathcal{P}}^\pi} \\
&\le \gamma^{-1/2} \|\widehat{Q}_0(\cdot,\cdot;W) - \widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi}^2.
\end{aligned}$$

Combining the above displays, we obtain

$$\langle \bar{g}_0^{\mathrm{SMDP}}(W) - \bar{g}_0^{\mathrm{SMDP}}(W^*),\, W - W^* \rangle \ge (1-\sqrt{\gamma}) \|\widehat{Q}_0(\cdot,\cdot;W) - \widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi}^2.$$

This completes the proof. □

Applying Lemma 8 at $W = \widehat{W}_k$, we obtain

$$-2\eta\gamma \langle \bar{g}_0^{\mathrm{SMDP}}(\widehat{W}_k), e_k \rangle \le -2\eta\gamma(1-\sqrt{\gamma}) \|\widehat{Q}_0(\cdot,\cdot;\widehat{W}_k) - \widehat{Q}_0(\cdot,\cdot;W^*)\|_{\mu^\pi}^2.$$

This is crucial because it yields the descent term in the Lyapunov recursion.

Step 3. Bound on $\langle \nu_k, e_k \rangle$. Recall that $\nu_k = g_k(\widehat{W}_k) - g_{0,k}(\widehat{W}_k)$.

On a non-reset branch, by definition,

$$\begin{aligned}\|\nu_k\|_2 &= \left\|\delta_k(\widehat{W}_k)\nabla_W \widehat{Q}(B_k, a_k; \widehat{W}_k) - \delta_{0,k}(\widehat{W}_k)\nabla_W \widehat{Q}_0(B_k, a_k)\right\|_2 \\ &\le \left|\delta_k(\widehat{W}_k)\right| \left\|\nabla_W \widehat{Q}(B_k, a_k; \widehat{W}_k) - \nabla_W \widehat{Q}_0(B_k, a_k)\right\|_2 \\ &\quad + \left|\delta_k(\widehat{W}_k) - \delta_{0,k}(\widehat{W}_k)\right| \left\|\nabla_W \widehat{Q}_0(B_k, a_k)\right\|_2 .\end{aligned}$$

Moreover,

$$\left|\delta_k(\widehat{W}_k)\right| \le r_{\max} + (1+\gamma)Q_{\max} =: \delta_{\max},$$

and

$$\begin{aligned}\left|\delta_k(\widehat{W}_k) - \delta_{0,k}(\widehat{W}_k)\right| &\le \left|\widehat{Q}(B_k, a_k; \widehat{W}_k) - \widehat{Q}_0(B_k, a_k; \widehat{W}_k)\right| \\ &\quad + \gamma \left|\widehat{Q}(B_{k+1}, a_{k+1}; \widehat{W}_k) - \widehat{Q}_0(B_{k+1}, a_{k+1}; \widehat{W}_k)\right| .\end{aligned}$$

Thus, by Fact 1, with probability at least $1 - e^{-\Omega(\log^2 m)}$,

$$\|\nu_k\|_2^2 = O\Big(\delta_{\max}^2 R^{2/3} m^{-1/3} d^5 \log m + L_Q^2 R^{8/3} m^{-1/3} d^6 \log m\Big) .$$

On a reset branch $\nu_k = 0$. Applying Young's inequality on the non-reset branch and using the independence of Z_{k+1} gives, for any $\alpha > 0$,

$$\begin{aligned}-2\eta \mathbb{E}\langle \nu_k, e_k \rangle &\le \eta\gamma\alpha \mathbb{E}\|e_k\|_2^2 \\ &\quad + \frac{\eta\gamma}{\alpha} O\Big(\delta_{\max}^2 R^{2/3} m^{-1/3} d^5 \log m + R^{8/3} m^{-1/3} d^8 \log m\Big) .\end{aligned}$$

The first term is retained in the Lyapunov recursion and bounded after summation.

Step 4. Bound on $\langle \omega_k, e_k \rangle$. The term ω_k captures the stationary population mismatch between the POMDP and SMDP directions.

For fixed $W \in \mathcal{B}$, insert the SMDP direction averaged under the actual marginal μ_{P}^{π}:

$$\begin{aligned}\bar{g}_0^{\mathrm{POMDP}}(W) - \bar{g}_0^{\mathrm{SMDP}}(W) &= \underbrace{\Big(\bar{g}_0^{\mathrm{POMDP}}(W) - \mathbb{E}_{\mu_{\mathrm{P}}^{\pi}}[g_0^{\mathrm{SMDP}}(B, a; W)]\Big)}_{\text{one-step reward and transition mismatch}} \\ &\quad + \underbrace{\Big(\mathbb{E}_{\mu_{\mathrm{P}}^{\pi}}[g_0^{\mathrm{SMDP}}(B, a; W)] - \bar{g}_0^{\mathrm{SMDP}}(W)\Big)}_{\text{occupancy mismatch}} .\end{aligned}$$

By Lemma 4, averaging over the full histories of stationary POMDP episodes gives

$$\left\|\bar{g}_0^{\mathrm{POMDP}}(W) - \mathbb{E}_{\mu_{\mathrm{P}}^{\pi}}[g_0^{\mathrm{SMDP}}(B, a; W)]\right\|_2 \le 2L_Q(r_{\max} + \gamma Q_{\max})(1-\rho)^{\ell}.$$

Since $\|g_0^{\mathrm{SMDP}}(B, a; W)\|_2 \le G_{\max}$, Lemma 6 bounds the second term. Hence,

$$\begin{aligned}\|\bar{g}_0^{\mathrm{POMDP}}(W) - \bar{g}_0^{\mathrm{SMDP}}(W)\|_2 &\le 2L_Q(r_{\max} + \gamma Q_{\max})(1-\rho)^{\ell} + 2G_{\max}\|\mu_{\mathrm{P}}^{\pi} - \mu^{\pi}\|_{\mathrm{TV}} \\ &\le \frac{2G_{\max}}{1-\gamma}(1-\rho)^{\ell}.\end{aligned}$$

This bound is uniform over $W \in \mathcal{B}$. Substitute $W = \widehat{W}_k$ and apply Young's inequality separately to the one-step and occupancy differences above. For any $\alpha > 0$, this gives

$$-2\eta\mathbb{E}\langle\omega_k, e_k\rangle \le 2\eta\gamma\alpha\mathbb{E}\|e_k\|_2^2 + \frac{\eta\gamma}{\alpha}O\left(\frac{L_Q^2\delta_{\max}^2}{(1-\gamma)^2}(1-\rho)^{2\ell}\right).$$

The belief comparison above is for a fixed parameter. It does not condition on the adaptive critic or past rewards.

Step 5. Bound on $\langle\zeta_k, e_k\rangle$. To control the Markovian bias term, fix an integer $\tau \ge 1$. For $k \ge \tau$, define the older filtration by

$$\mathcal{I}_{k-\tau} := \sigma\Big(X_0, \dots, X_{k-\tau}, R_0, \dots, R_{k-\tau-1}, \widehat{W}_0, \dots, \widehat{W}_{k-\tau}\Big).$$

We use the actual augmented chain X_k from Step 1. Lemma 6 gives its mixing bound. For $k \ge \tau$, freeze the critic at $\widehat{W}_{k-\tau}$. Also recall that

$$\delta_{\max} := r_{\max} + (1+\gamma)Q_{\max}, \qquad G_{\max} := \delta_{\max}L_Q,$$

so that $\|g_k(\widehat{W}_k)\|_2 \le G_{\max}$.

For the sampled linearized direction, we have

$$\begin{aligned}\|g_{0,k}(W)\|_2 &\le Z_{k+1}\Big(|r_k| + |\widehat{Q}_0(B_k, a_k; W)| + \gamma|\widehat{Q}_0(B_{k+1}, a_{k+1}; W)|\Big)\|\nabla_W\widehat{Q}_0(B_k, a_k)\|_2 \\ &\le Z_{k+1}\big(r_{\max} + (1+\gamma)Q_{\max}\big)L_Q = Z_{k+1}G_{\max}.\end{aligned}$$

Taking the stationary expectation conditional on a non-reset transition gives

$$\|\bar{g}_0^{\mathrm{POMDP}}(W)\|_2 \le \delta_{\max}L_Q = G_{\max}.$$

By the definition of ζ_k, we have

$$\begin{aligned}\zeta_k = &\underbrace{\Big(g_{0,k}(\widehat{W}_k) - g_{0,k}(\widehat{W}_{k-\tau})\Big)}_{\text{iterate drift}} \\ &+ \underbrace{\Big(g_{0,k}(\widehat{W}_{k-\tau}) - \gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_{k-\tau})\Big)}_{\text{lagged bias at frozen iterate}} \\ &+ \underbrace{\Big(\gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_{k-\tau}) - \gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_k)\Big)}_{\text{population drift}}.\end{aligned} \tag{7}$$

STEP 5.1. LIPSCHITZ CONTINUITY OF $g_{0,k}$ AND $\bar{g}_0^{\mathrm{POMDP}}$. For any $W, W' \in \mathcal{B}$, the Temporal Difference (TD) error satisfies

$$\begin{aligned}|\delta_{0,k}(W) - \delta_{0,k}(W')| &\le |\widehat{Q}_0(B_k, a_k; W) - \widehat{Q}_0(B_k, a_k; W')| \\ &\quad + \gamma|\widehat{Q}_0(B_{k+1}, a_{k+1}; W) - \widehat{Q}_0(B_{k+1}, a_{k+1}; W')| \\ &\le (1+\gamma)L_Q\|W - W'\|_2.\end{aligned}$$

The bound holds on a non-reset transition. Both residuals are zero on a reset. Hence,

$$\begin{aligned}\|g_{0,k}(W) - g_{0,k}(W')\|_2 &= Z_{k+1}|\delta_{0,k}(W) - \delta_{0,k}(W')|\,\|\nabla_W\widehat{Q}_0(B_k, a_k)\|_2 \\ &\le Z_{k+1}(1+\gamma)L_Q^2\|W - W'\|_2.\end{aligned}$$

Setting $L_0 := (1+\gamma)L_Q^2$, the bounds above and conditional Jensen's inequality give

$$\begin{aligned}\|g_{0,k}(W) - g_{0,k}(W')\|_2 &\le Z_{k+1}L_0\|W - W'\|_2, \\ \|\bar{g}_0^{\mathrm{POMDP}}(W) - \bar{g}_0^{\mathrm{POMDP}}(W')\|_2 &\le L_0\|W - W'\|_2.\end{aligned}$$

Projection nonexpansiveness also gives $\|\widehat{W}_k - \widehat{W}_{k-\tau}\|_2 \le \eta\tau G_{\max}$.

By the triangle inequality and the update bound, we have

$$\mathbb{E}\left\|\Big(g_{0,k}(\widehat{W}_k) - g_{0,k}(\widehat{W}_{k-\tau})\Big) + \Big(\gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_{k-\tau}) - \gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_k)\Big)\right\|_2$$
$$\le \mathbb{E}[(Z_{k+1}+\gamma)L_0\|\widehat{W}_k - \widehat{W}_{k-\tau}\|_2] \le 2\gamma L_0\eta\tau G_{\max}.$$

For the drift terms, Young's inequality gives

Here $\mathbb{E}[Z_{k+1} \mid \mathcal{I}_k] = \gamma$. Both e_k and $\widehat{W}_k - \widehat{W}_{k-\tau}$ are $\mathcal{I}_k$-measurable. Young's inequality therefore gives

$$-2\eta\mathbb{E}\Big\langle\Big(g_{0,k}(\widehat{W}_k) - g_{0,k}(\widehat{W}_{k-\tau})\Big) + \gamma\Big(\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_{k-\tau}) - \bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_k)\Big), e_k\Big\rangle$$
$$\le 4\eta\gamma L_0\mathbb{E}\Big[\|\widehat{W}_k - \widehat{W}_{k-\tau}\|_2\|e_k\|_2\Big]$$
$$\le \eta\gamma\alpha\mathbb{E}\|e_k\|_2^2 + \frac{4\eta\gamma L_0^2}{\alpha}\mathbb{E}\|\widehat{W}_k - \widehat{W}_{k-\tau}\|_2^2 \le \eta\gamma\alpha\mathbb{E}\|e_k\|_2^2 + \frac{4\gamma L_0^2}{\alpha}\eta^3\tau^2 G_{\max}^2.$$

STEP 5.2. BOUND ON LAGGED BIAS. For the lagged bias term in (7), we freeze the iterate at $\widehat{W}_{k-\tau}$ and condition on $\mathcal{I}_{k-\tau}$.

Then

$$g_{0,k}(\widehat{W}_{k-\tau}) - \gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_{k-\tau})$$
$$= \underbrace{\Big(g_{0,k}(\widehat{W}_{k-\tau}) - \mathbb{E}[g_{0,k}(\widehat{W}_{k-\tau}) \mid \mathcal{I}_{k-\tau}]\Big)}_{\Xi_{k,\tau}\text{ (martingale difference)}}$$
$$+ \underbrace{\Big(\mathbb{E}[g_{0,k}(\widehat{W}_{k-\tau}) \mid \mathcal{I}_{k-\tau}] - \gamma\bar{g}_0^{\mathrm{POMDP}}(\widehat{W}_{k-\tau})\Big)}_{\Gamma_{k,\tau}^{\mathrm{mix}}\text{ (mixing bias)}}.$$

The population mismatch is already included in ω_k.

Since $\mathbb{E}[\Xi_{k,\tau} \mid \mathcal{I}_{k-\tau}] = 0$, the term simplifies to

$$\mathbb{E}\langle\Xi_{k,\tau}, e_k\rangle = \mathbb{E}\langle\Xi_{k,\tau}, e_k - e_{k-\tau}\rangle.$$

We have $\|\Xi_{k,\tau}\|_2 \le 2G_{\max}$ and $\|e_k - e_{k-\tau}\|_2 \le \eta\tau G_{\max}$. The reset coin also gives $\mathbb{E}\|\Xi_{k,\tau}\|_2 \le 2\gamma G_{\max}$. Therefore,

$$-2\eta\mathbb{E}\langle\Xi_{k,\tau}, e_k\rangle = -2\eta\mathbb{E}\langle\Xi_{k,\tau}, e_k - e_{k-\tau}\rangle$$
$$\le 2\eta\mathbb{E}[\|\Xi_{k,\tau}\|_2\|e_k - e_{k-\tau}\|_2] \le 4\eta^2\gamma\tau G_{\max}^2.$$

The expectation bound follows from $\|g_{0,k}(W)\|_2 \le Z_{k+1}G_{\max}$ and conditional Jensen's inequality.

Bound on $\Gamma_{k,\tau}^{\mathrm{mix}}$.

Given X_k, the mean of $g_{0,k}(W)$ is bounded by $\gamma G_{\max}$ for every fixed W.

For the actual augmented POMDP chain, Lemma 6 gives

$$\|\mathcal{L}(X_k \mid \mathcal{I}_{k-\tau}) - \xi^\pi\|_{\mathrm{TV}} \le \gamma^\tau.$$

Hence,

$$\|\Gamma_{k,\tau}^{\mathrm{mix}}\|_2 \le 2\gamma G_{\max}\|\mathcal{L}(X_k \mid \mathcal{I}_{k-\tau}) - \xi^\pi\|_{\mathrm{TV}} \le 2\gamma G_{\max}\gamma^\tau.$$

Applying Young's inequality for any $\alpha > 0$, the inner product is bounded by

$$-2\eta\mathbb{E}\langle\Gamma_{k,\tau}^{\mathrm{mix}}, e_k\rangle \le 2\eta\mathbb{E}[\|\Gamma_{k,\tau}^{\mathrm{mix}}\|_2\|e_k\|_2] \le \eta\gamma\alpha\mathbb{E}\|e_k\|_2^2 + \frac{4\eta\gamma G_{\max}^2\gamma^{2\tau}}{\alpha}.$$

Combining the drift, martingale, and mixing bounds gives

$$-2\eta\mathbb{E}\langle\zeta_k, e_k\rangle \le 2\eta\gamma\alpha\mathbb{E}\|e_k\|_2^2 + \frac{4\gamma L_0^2}{\alpha}\eta^3\tau^2 G_{\max}^2 + 4\eta^2\gamma\tau G_{\max}^2 + \frac{4\eta\gamma G_{\max}^2\gamma^{2\tau}}{\alpha}.$$

Step 6. Combine all. Taking expectation in (6) and combining the bounds from Steps 2–5, we obtain the following inequality for every $k \ge \tau$.

$$\begin{aligned}
2\eta\gamma(1-\sqrt{\gamma})\mathbb{E}\|\widehat{Q}_0(\cdot,\cdot;\widehat{W}_k) - Q_{\mathcal{F}}^{*,\pi}\|_{\mu^\pi}^2 &\le (1+5\eta\gamma\alpha)\mathbb{E}\|e_k\|_2^2 - \mathbb{E}\|e_{k+1}\|_2^2 \\
&+ \frac{\eta\gamma}{\alpha}O\left(\delta_{\max}^2 R^{2/3}m^{-1/3}d^5\log m + R^{8/3}m^{-1/3}d^8\log m + \frac{L_Q^2\delta_{\max}^2}{(1-\gamma)^2}(1-\rho)^{2\ell}\right) \\
&+ \frac{4\gamma L_0^2}{\alpha}\eta^3\tau^2 G_{\max}^2 + 4\eta^2\gamma\tau G_{\max}^2 + \frac{4\eta\gamma G_{\max}^2\gamma^{2\tau}}{\alpha} + \eta^2\gamma G_{\max}^2.
\end{aligned}$$

We sum the preceding inequality from $k = \tau$ to $K-1$. The Lyapunov terms telescope as

$$\sum_{k=\tau}^{K-1}\left(\mathbb{E}\|e_k\|_2^2 - \mathbb{E}\|e_{k+1}\|_2^2\right) = \mathbb{E}\|e_\tau\|_2^2 - \mathbb{E}\|e_K\|_2^2 \le 4dR^2.$$

Moreover, $\|e_k\|_2^2 \le 4dR^2$ implies

$$5\eta\gamma\alpha\sum_{k=\tau}^{K-1}\mathbb{E}\|e_k\|_2^2 \le 20\eta\gamma\alpha dR^2(K-\tau).$$

Algorithm 1 averages $\widehat{W}_0, \dots, \widehat{W}_K$. Since $\widehat{Q}_0$ is affine in W, Jensen's inequality gives

$$\mathbb{E}\|\widehat{Q}_0(\bar{W}_K) - Q_{\mathcal{F}}^{*,\pi}\|_{\mu^\pi}^2 \le \frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\widehat{Q}_0(\widehat{W}_k) - Q_{\mathcal{F}}^{*,\pi}\|_{\mu^\pi}^2.$$

The first τ iterates and the last iterate have squared projected function error at most $4Q_{\max}^2$. Therefore, for an absolute constant C,

$$\begin{aligned}
&\mathbb{E}\|\widehat{Q}_0(\bar{W}_K) - Q_{\mathcal{F}}^{*,\pi}\|_{\mu^\pi}^2 \\
&\le \frac{C}{1-\sqrt{\gamma}}\left(\frac{dR^2}{\gamma\eta K} + dR^2\alpha + \eta(\tau+1)G_{\max}^2\right) \\
&\quad + \frac{1}{\alpha(1-\sqrt{\gamma})}O\left(\delta_{\max}^2 R^{2/3}m^{-1/3}d^5\log m + R^{8/3}m^{-1/3}d^8\log m + \frac{L_Q^2\delta_{\max}^2}{(1-\gamma)^2}(1-\rho)^{2\ell}\right) \\
&\quad + \frac{CL_0^2\eta^2\tau^2 G_{\max}^2 + CG_{\max}^2\gamma^{2\tau}}{\alpha(1-\sqrt{\gamma})} + \frac{4(\tau+1)Q_{\max}^2}{K+1}.
\end{aligned}$$

Choose

$$\eta = K^{-1/2}, \qquad \tau = \max\left\{1, \left\lceil\frac{\log K}{2|\log\gamma|}\right\rceil\right\}, \qquad K > \tau.$$

Then $\gamma^\tau \le K^{-1/2}$. We use the same Young parameter throughout the preceding bounds and set

$$\alpha = \frac{1}{\sqrt{d}\,R}\Big(\delta_{\max}^2 R^{2/3} m^{-1/3} d^5 \log m + R^{8/3} m^{-1/3} d^8 \log m + \frac{L_Q^2 \delta_{\max}^2}{(1-\gamma)^2}(1-\rho)^{2\ell} + L_0^2 \eta^2 \tau^2 G_{\max}^2 + G_{\max}^2 \gamma^{2\tau}\Big)^{1/2}.$$

This choice balances the $dR^2\alpha$ term with the terms divided by α. Using $G_{\max} = L_Q \delta_{\max}$ and the subadditivity of the square root gives

$$\begin{aligned}
&\mathbb{E}\|\widehat{Q}_0(\bar{W}_K) - Q_{\mathcal{F}}^{*,\pi}\|_{\mu^\pi}^2 \\
&\le \frac{C}{1-\sqrt{\gamma}}\left(\frac{dR^2}{\gamma\eta K} + \eta(\tau+1)G_{\max}^2\right) \\
&\quad + \frac{O\Big(\delta_{\max} R^{4/3} m^{-1/6} d^3 \log^{1/2} m + R^{7/3} m^{-1/6} d^{9/2} \log^{1/2} m\Big)}{1-\sqrt{\gamma}} \\
&\quad + \frac{C\sqrt{d}\,R G_{\max}}{1-\sqrt{\gamma}}\left(\frac{(1-\rho)^\ell}{1-\gamma} + L_0\eta\tau + \gamma^\tau\right) + \frac{4(\tau+1)Q_{\max}^2}{K+1}.
\end{aligned}$$

Finally, $\delta_{\max} = r_{\max} + (1+\gamma)Q_{\max}$ and $Q_{\max} = O(\log m + d^{3/2}R)$. Thus, $\delta_{\max} = O(\log m + d^{3/2}R)$ for fixed $r_{\max}$. Substituting these relations together with $L_Q = O(d)$ and $L_0 = O(d^2)$ yields

$$\begin{aligned}
&\mathbb{E}\Big[\big\|\widehat{Q}_0(\cdot,\cdot;\bar{W}_K) - Q_{\mathcal{F}}^{*,\pi}\big\|_{\mu^\pi}^2\Big] \\
&\le \widetilde{O}\left(\frac{d^2 + d^{7/2}R + d^5R^2}{\sqrt{K}}\right) + \widetilde{O}\Big(R^{4/3} m^{-1/6} d^3 + R^{7/3} m^{-1/6} d^{9/2}\Big) \\
&\quad + \widetilde{O}\Big((d^{3/2}R + d^3R^2)(1-\rho)^\ell\Big) + \widetilde{O}\left(\frac{d^{3/2}R + d^3R^2}{\sqrt{K}}\right).
\end{aligned}$$

C.2.3 Bound Approximation Error

There exists $W^\dagger \in \mathcal{B}$ such that

$$\|\widehat{Q}(\cdot,\cdot;W^\dagger) - \tilde{Q}^\pi\|_{\mu^\pi} \le \widetilde{O}\Big(m^{-2/p} d^{-2/p}\Big).$$

We first decompose

$$\begin{aligned}
\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^\pi\|_{\mu^\pi} &\le \|Q_{\mathcal{F}}^{*,\pi} - \widehat{Q}_0(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} + \|\widehat{Q}_0(\cdot,\cdot;W^\dagger) - \widehat{Q}(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} \\
&\quad + \|\widehat{Q}(\cdot,\cdot;W^\dagger) - \tilde{Q}^\pi\|_{\mu^\pi}.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\|Q_{\mathcal{F}}^{*,\pi} - \widehat{Q}_0(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} &\le \|Q_{\mathcal{F}}^{*,\pi} - \Pi_{\mathcal{F}}\tilde{Q}^\pi\|_{\mu^\pi} + \|\Pi_{\mathcal{F}}\tilde{Q}^\pi - \widehat{Q}_0(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} \\
&\le \|\Pi_{\mathcal{F}}\tilde{T}^\pi Q_{\mathcal{F}}^{*,\pi} - \Pi_{\mathcal{F}}\tilde{T}^\pi\tilde{Q}^\pi\|_{\mu^\pi} + \|\widehat{Q}_0(\cdot,\cdot;W^\dagger) - \tilde{Q}^\pi\|_{\mu^\pi} \\
&\le \sqrt{\gamma}\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^\pi\|_{\mu^\pi} + \|\widehat{Q}_0(\cdot,\cdot;W^\dagger) - \widehat{Q}(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} + \|\widehat{Q}(\cdot,\cdot;W^\dagger) - \tilde{Q}^\pi\|_{\mu^\pi}.
\end{aligned}$$

Here, the first inequality is by the triangle inequality; the second uses $Q_{\mathcal{F}}^{*,\pi} = \Pi_{\mathcal{F}}\tilde{T}^\pi Q_{\mathcal{F}}^{*,\pi}$, $\tilde{Q}^\pi = \tilde{T}^\pi\tilde{Q}^\pi$, and the projection property $\|\Pi_{\mathcal{F}}\tilde{Q}^\pi - \widehat{Q}_0(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} \le \|\tilde{Q}^\pi - \widehat{Q}_0(\cdot,\cdot;W^\dagger)\|_{\mu^\pi}$ since $\widehat{Q}_0(\cdot,\cdot;W^\dagger) \in \mathcal{F}$; the third follows from the non-expansiveness of $\Pi_{\mathcal{F}}$, the $\sqrt{\gamma}$-contraction of $\tilde{T}^\pi$, and the triangle inequality.

Substituting this bound into the initial decomposition gives

$$\begin{aligned}
\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^\pi\|_{\mu^\pi} &\le \sqrt{\gamma}\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^\pi\|_{\mu^\pi} + 2\|\widehat{Q}_0(\cdot,\cdot;W^\dagger) - \widehat{Q}(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} \\
&\quad + 2\|\widehat{Q}(\cdot,\cdot;W^\dagger) - \tilde{Q}^\pi\|_{\mu^\pi}.
\end{aligned}$$

Rearranging yields

$$\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^{\pi}\|_{\mu^\pi} \leq \frac{2}{1-\sqrt{\gamma}} \left(\|\widehat{Q}_0(\cdot,\cdot;W^\dagger) - \widehat{Q}(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} + \|\widehat{Q}(\cdot,\cdot;W^\dagger) - \tilde{Q}^{\pi}\|_{\mu^\pi} \right).$$

Finally, by Fact 1

$$\|\widehat{Q}_0(\cdot,\cdot;W^\dagger) - \widehat{Q}(\cdot,\cdot;W^\dagger)\|_{\mu^\pi} \leq O\Big(R^{4/3} m^{-1/6} d^3 \log^{1/2} m\Big),$$

and

$$\|\widehat{Q}(\cdot,\cdot;W^\dagger) - \tilde{Q}^{\pi}\|_{\mu^\pi} \leq \widetilde{O}\Big(m^{-2/p} d^{-2/p}\Big).$$

Therefore,

$$\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^{\pi}\|_{\mu^\pi} \leq \widetilde{O}\Big(\frac{R^{4/3} m^{-1/6} d^3 + m^{-2/p} d^{-2/p}}{1-\sqrt{\gamma}}\Big).$$

Hence,

$$\|Q_{\mathcal{F}}^{*,\pi} - \tilde{Q}^{\pi}\|_{\mu^\pi}^2 \leq \widetilde{O}\left(\frac{\left(R^{4/3} m^{-1/6} d^3 + m^{-2/p} d^{-2/p}\right)^2}{(1-\sqrt{\gamma})^2}\right).$$

Combining this bound with the NTK linearization bound and the projected optimization bound, yields Theorem 2.

D Proof of Policy Update

D.1 Preliminaries

First, we give the equivalence of policy update forms for analysis, where we use Recursive Softmax Form in Algorithm 2.

Lemma 9 (Equivalence of Policy Update Forms)**.** *For a fixed superstate $B \in \mathsf{B}$ and a stepsize $\kappa > 0$, starting from a uniform initial policy $\pi_0(\cdot|B)$, the following three update rules are equivalent for all $t \geq 0$:*

1. ***Mirror Descent Form:***

$$\pi_{t+1}(\cdot|B) = \arg\min_{p \in \Delta(\mathcal{A})} \Big\{ -\kappa \langle p, \widehat{Q}_t(B,\cdot)\rangle + D_{\mathrm{KL}}(p \| \pi_t(\cdot|B)) \Big\}. \tag{8}$$

2. ***Recursive Softmax Form:***

$$\pi_{t+1}(a|B) = \frac{\pi_t(a|B) \exp\big(\kappa \widehat{Q}_t(B,a)\big)}{\sum_{a' \in \mathcal{A}} \pi_t(a'|B) \exp\big(\kappa \widehat{Q}_t(B,a')\big)}. \tag{9}$$

3. ***Cumulative Exponential-Weights Form:***

$$\pi_{t+1}(a|B) = \frac{\exp\big(\kappa \sum_{j=0}^{t} \widehat{Q}_j(B,a)\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\kappa \sum_{j=0}^{t} \widehat{Q}_j(B,a')\big)}. \tag{10}$$

Proof. Fix any superstate $B \in \mathsf{B}$. First, to show equivalence between (8) and (9), we consider the Lagrangian of the Mirror Descent objective with multiplier λ for the simplex constraint $\sum_a p(a) = 1$:

$$\mathcal{L}(p,\lambda) = -\kappa \sum_a p(a) \widehat{Q}_t(B,a) + \sum_a p(a) \log \frac{p(a)}{\pi_t(a|B)} + \lambda \Big(\sum_a p(a) - 1 \Big).$$

Setting the first-order optimality condition $\frac{\partial \mathcal{L}}{\partial p(a)} = -\kappa \widehat{Q}_t(B,a) + \log \frac{p(a)}{\pi_t(a|B)} + 1 + \lambda = 0$ yields the characterization:

$$\pi_{t+1}(a|B) = \pi_t(a|B) \exp(\kappa \widehat{Q}_t(B,a) - 1 - \lambda) \propto \pi_t(a|B) \exp(\kappa \widehat{Q}_t(B,a)). \tag{11}$$

Normalizing over $\mathcal{A}$ recovers the recursive softmax update (9).

Next, the equivalence between the recursive form (9) and cumulative form (10) follows by induction. For the base case $t = 0$, the uniform initialization $\pi_0(a|B) = 1/|\mathcal{A}|$ directly gives $\pi_1(a|B) \propto \exp(\kappa \widehat{Q}_0(B, a))$. Assuming the equivalence holds for π_t, then for π_{t+1} we have the chained relation:

$$\pi_{t+1}(a|B) \propto \pi_t(a|B) e^{\kappa \widehat{Q}_t(B,a)} \propto e^{\kappa \sum_{j=0}^{t-1} \widehat{Q}_j(B,a)} \cdot e^{\kappa \widehat{Q}_t(B,a)} = e^{\kappa \sum_{j=0}^{t} \widehat{Q}_j(B,a)}, \tag{12}$$

where the second proportionality follows from the induction hypothesis. Normalization recovers the cumulative exponential-weights form, completing the proof. □

The Goal. Let $N := TK$. Let π^* denote an unrestricted optimal policy of the original POMDP. Let $\widetilde{\pi}^*$ denote an optimal policy of the fixed superstate MDP. Following Anjarlekar et al. (2025), we compare the unrestricted POMDP optimum with the surrogate value achieved by each learned superstate policy. Starting from $H_0 = \varnothing$ and $B_0 = \mathcal{G}(H_0) = \varnothing$, define

$$R_N := K \sum_{t=0}^{T-1} \mathbb{E}\Big[V^{\pi^*}(H_0) - \tilde{V}^{\pi_t}(B_0)\Big].$$

We can decompose

$$R_N = \underbrace{K \sum_{t=0}^{T-1} \mathbb{E}\big[V^{\pi^*}(H_0) - \tilde{V}^{\widetilde{\pi}^*}(B_0)\big]}_{\text{model approximation error}} + \underbrace{K \sum_{t=0}^{T-1} \mathbb{E}\big[\tilde{V}^{\widetilde{\pi}^*}(B_0) - \tilde{V}^{\pi_t}(B_0)\big]}_{\text{policy update regret}}.$$

Bound Model Approximation Error. By Theorem 1, we have

$$\Big|V^{\pi^*}(H_0) - \tilde{V}^{\widetilde{\pi}^*}(B_0)\Big| \le \frac{2r_{\max}}{(1-\gamma)^2}(1-\rho)^{\ell}.$$

Therefore, the model approximation part of the regret satisfies

$$K \sum_{t=0}^{T-1} \mathbb{E}\big[V^{\pi^*}(H_0) - \tilde{V}^{\widetilde{\pi}^*}(B_0)\big] \le \frac{2Nr_{\max}}{(1-\gamma)^2}(1-\rho)^{\ell}.$$

D.2 Bound Superstate Policy Regret

We now bound the superstate policy-regret term $K \sum_{t=0}^{T-1} \mathbb{E}\big[\tilde{V}^{\widetilde{\pi}^*}(B_0) - \tilde{V}^{\pi_t}(B_0)\big]$.

Starting from $B_0 = \varnothing$, define the normalized discounted visitation distribution under policy π by

$$d^{\pi}(B) := (1-\gamma) \sum_{j=0}^{\infty} \gamma^j \Pr(B_j = B \mid B_0, \pi).$$

Under the full-reset SMDP construction, d^{π} coincides with the stationary superstate distribution by Lemma 5. Hence μ^{π} has marginal d^{π}. These are the discounted occupancies of the original SMDP.

Step 1. Performance-difference decomposition. By the performance difference lemma, for every t,

$$\tilde{V}^{\widetilde{\pi}^*}(B_0) - \tilde{V}^{\pi_t}(B_0) = \frac{1}{1-\gamma} \sum_{B' \in \mathsf{B}} d^{\widetilde{\pi}^*}(B') \sum_{a \in \mathcal{A}} \big(\widetilde{\pi}^*(a \mid B') - \pi_t(a \mid B')\big) \tilde{Q}^{\pi_t}(B', a).$$

Summing over t and multiplying by K, and then adding and subtracting $\widehat{Q}_t$, we obtain

$$\begin{aligned}
&K \sum_{t=0}^{T-1} \mathbb{E}\big[\tilde{V}^{\widetilde{\pi}^*}(B_0) - \tilde{V}^{\pi_t}(B_0)\big] \\
&= \frac{K}{1-\gamma} \sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{B,a} d^{\widetilde{\pi}^*}(B)\big(\widetilde{\pi}^*(a \mid B) - \pi_t(a \mid B)\big)\widehat{Q}_t(B,a)\right] \\
&\quad - \frac{K}{1-\gamma} \sum_{t=0}^{T-1} \mathbb{E}\left[\sum_{B,a} d^{\widetilde{\pi}^*}(B)\big(\widetilde{\pi}^*(a \mid B) - \pi_t(a \mid B)\big)\big(\widehat{Q}_t(B,a) - \tilde{Q}^{\pi_t}(B,a)\big)\right].
\end{aligned}$$

The first term is the PMD optimization error with respect to the estimated critic, and the second term is the critic-induced error.

Step 2. PMD optimization error. For each fixed superstate B, the recursive softmax update is the entropic mirror-descent update on the simplex:

$$\pi_{t+1}(\cdot \mid B) = \arg\min_{p \in \Delta(\mathcal{A})} \left\{ -\kappa \langle p, \widehat{Q}_t(B, \cdot) \rangle + D_{\mathrm{KL}}(p \| \pi_t(\cdot \mid B)) \right\}.$$

Therefore, the standard exponential-weights regret bound gives, for comparator $\widetilde{\pi}^*(\cdot \mid B)$,

$$\sum_{t=0}^{T-1} \langle \widetilde{\pi}^*(\cdot \mid B) - \pi_t(\cdot \mid B), \widehat{Q}_t(B, \cdot) \rangle \le \frac{D_{\mathrm{KL}}(\widetilde{\pi}^*(\cdot \mid B) \| \pi_0(\cdot \mid B))}{\kappa} + \frac{\kappa}{2} \sum_{t=0}^{T-1} \|\widehat{Q}_t(B, \cdot)\|_\infty^2.$$

Since $\pi_0(\cdot \mid B)$ is uniform and $\|\widehat{Q}_t(B, \cdot)\|_\infty \le Q_{\max}$, we have

$$\sum_{t=0}^{T-1} \langle \widetilde{\pi}^*(\cdot \mid B) - \pi_t(\cdot \mid B), \widehat{Q}_t(B, \cdot) \rangle \le \frac{\log |\mathcal{A}|}{\kappa} + \frac{\kappa T Q_{\max}^2}{2}.$$

Taking expectation over $B \sim d^{\widetilde{\pi}^*}$ gives

$$\sum_{t=0}^{T-1} \mathbb{E}_{B \sim d^{\widetilde{\pi}^*}} \left[\langle \widetilde{\pi}^*(\cdot \mid B) - \pi_t(\cdot \mid B), \widehat{Q}_t(B, \cdot) \rangle \right] \le \frac{\log |\mathcal{A}|}{\kappa} + \frac{\kappa T Q_{\max}^2}{2}.$$

Taking expectation over algorithmic randomness preserves this upper bound.

Step 3. Critic-induced error under concentrability. Conditioning on π_t, by the triangle inequality $|\widetilde{\pi}^*(a \mid B) - \pi_t(a \mid B)| \le \widetilde{\pi}^*(a \mid B) + \pi_t(a \mid B)$, we have

$$\begin{aligned} &\left| \sum_{B,a} d^{\widetilde{\pi}^*}(B) \big(\widetilde{\pi}^*(a \mid B) - \pi_t(a \mid B) \big) \big(\widehat{Q}_t(B, a) - \tilde{Q}^{\pi_t}(B, a) \big) \right| \\ &\le \sum_{B,a} \big(\mu^{\widetilde{\pi}^*}(B, a) + d^{\widetilde{\pi}^*}(B) \pi_t(a \mid B) \big) \big| \widehat{Q}_t(B, a) - \tilde{Q}^{\pi_t}(B, a) \big|. \end{aligned}$$

For the first term, Cauchy–Schwarz gives

$$\sum_{B,a} \mu^{\widetilde{\pi}^*}(B, a) \big| \widehat{Q}_t - \tilde{Q}^{\pi_t} \big| \le C_t^{1/2} \|\widehat{Q}_t - \tilde{Q}^{\pi_t}\|_{\mu^{\pi_t}}.$$

For the second term, since $d^{\widetilde{\pi}^*}(B) \pi_t(a \mid B) = \mu^{\pi_t}(B, a) d^{\widetilde{\pi}^*}(B) / d^{\pi_t}(B)$, another application of Cauchy–Schwarz yields

$$\sum_{B,a} d^{\widetilde{\pi}^*}(B) \pi_t(a \mid B) \big| \widehat{Q}_t - \tilde{Q}^{\pi_t} \big| \le \left(\sum_B \frac{(d^{\widetilde{\pi}^*}(B))^2}{d^{\pi_t}(B)} \right)^{1/2} \|\widehat{Q}_t - \tilde{Q}^{\pi_t}\|_{\mu^{\pi_t}}.$$

Moreover,

$$\sum_B \frac{(d^{\widetilde{\pi}^*}(B))^2}{d^{\pi_t}(B)} \le \sum_{B,a} \frac{(\mu^{\widetilde{\pi}^*}(B, a))^2}{\mu^{\pi_t}(B, a)} = C_t.$$

Thus,

$$\left| \sum_{B,a} d^{\widetilde{\pi}^*}(B) \big(\widetilde{\pi}^*(a \mid B) - \pi_t(a \mid B) \big) \big(\widehat{Q}_t(B, a) - \tilde{Q}^{\pi_t}(B, a) \big) \right| \le 2 C_t^{1/2} \|\widehat{Q}_t - \tilde{Q}^{\pi_t}\|_{\mu^{\pi_t}}.$$

Taking expectation and applying Cauchy–Schwarz, Assumption 3, and Theorem 2, we obtain

$$\begin{aligned} &\mathbb{E} \left[\left| \sum_{B,a} d^{\widetilde{\pi}^*}(B) \big(\widetilde{\pi}^*(a \mid B) - \pi_t(a \mid B) \big) \big(\widehat{Q}_t(B, a) - \tilde{Q}^{\pi_t}(B, a) \big) \right| \right] \\ &\le 2 \mathbb{E} \left[C_t^{1/2} \|\widehat{Q}_t - \tilde{Q}^{\pi_t}\|_{\mu^{\pi_t}} \right] \le 2 \sqrt{\mathbb{E}[C_t]} \sqrt{\mathbb{E} \left[\|\widehat{Q}_t - \tilde{Q}^{\pi_t}\|_{\mu^{\pi_t}}^2 \right]} \le 2 \sqrt{C_\infty \varepsilon_K}. \end{aligned}$$

where $\varepsilon_K = \mathbb{E} \|\widehat{Q}(\cdot, \cdot; \bar{W}_K) - \tilde{Q}^\pi\|_{\mu^\pi}^2$ denotes the squared critic-error upper bound from Theorem 2.

Step 4. Combine the two bounds. Combining Steps 1–3 gives

$$K\sum_{t=0}^{T-1}\mathbb{E}\big[\tilde{V}^{\tilde{\pi}^*}(B_0)-\tilde{V}^{\pi_t}(B_0)\big]\le\frac{K}{1-\gamma}\left(\frac{\log|\mathcal{A}|}{\kappa}+\frac{\kappa TQ_{\max}^2}{2}+2T\sqrt{\bar{C}_\infty\varepsilon_K}\right).$$

Together with the model-approximation bound,

$$K\sum_{t=0}^{T-1}\mathbb{E}\big[V^{\pi^*}(H_0)-\tilde{V}^{\tilde{\pi}^*}(B_0)\big]\le\frac{2Nr_{\max}}{(1-\gamma)^2}(1-\rho)^\ell,$$

we obtain

$$R_N\le\frac{2Nr_{\max}}{(1-\gamma)^2}(1-\rho)^\ell+\frac{K}{1-\gamma}\left(\frac{\log|\mathcal{A}|}{\kappa}+\frac{\kappa TQ_{\max}^2}{2}+2T\sqrt{\bar{C}_\infty\varepsilon_K}\right).$$

Choosing $\kappa=\frac{\sqrt{2\log|\mathcal{A}|}}{Q_{\max}\sqrt{T}}$ balances the two PMD terms and gives

$$\frac{\log|\mathcal{A}|}{\kappa}+\frac{\kappa TQ_{\max}^2}{2}=Q_{\max}\sqrt{2T\log|\mathcal{A}|}.$$

Since $N=TK$, this yields

$$R_N\le\frac{2Nr_{\max}}{(1-\gamma)^2}(1-\rho)^\ell+\frac{N}{1-\gamma}\left(Q_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}}+2\sqrt{\bar{C}_\infty\varepsilon_K}\right).$$

Step 5. Substitute the critic bound and choose K,T. Each critic run satisfies the conditions of Theorem 2 as required in Theorem 3. By Theorem 2 and $\sqrt{a_1+\cdots+a_s}\le\sqrt{a_1}+\cdots+\sqrt{a_s}$, we obtain

$$\begin{aligned}\sqrt{\varepsilon_K}&\le\widetilde{O}\Big(m^{-2/p}d^{-2/p}\Big)+\widetilde{O}\Big(R^{4/3}m^{-1/6}d^3+R^{2/3}m^{-1/12}d^{3/2}+R^{7/6}m^{-1/12}d^{9/4}\Big)\\&\quad+\widetilde{O}\left(\frac{d+d^{7/4}R^{1/2}+d^{5/2}R}{K^{1/4}}\right)+\widetilde{O}\Big((d^{3/4}R^{1/2}+d^{3/2}R)(1-\rho)^{\ell/2}\Big)\\&\quad+\widetilde{O}\Big((d^{3/4}R^{1/2}+d^{3/2}R)\gamma^{\tau/2}\Big).\end{aligned}$$

Choose $K=N^{2/3},T=N^{1/3}$.

Choose the same analysis lag as above, $\tau=\max\{1,\lceil\log K/(2|\log\gamma|)\rceil\}$, with $K>\tau$. Then

$$T^{-1/2}=K^{-1/4}=N^{-1/6},\qquad\gamma^{\tau/2}\le K^{-1/4}.$$

The mixing polynomial satisfies $d^{3/4}R^{1/2}+d^{3/2}R\le P_{\mathrm{opt}}(d,R)$ for $d\ge1$.

Using also $Q_{\max}=O(\log m+d^{3/2}R)$ and absorbing logarithmic factors into $\widetilde{O}(\cdot)$, we obtain the following bound:

$$\begin{aligned}R_N&\le N\Big[\widetilde{O}(m^{-2/p}d^{-2/p})+\widetilde{O}(R^{4/3}m^{-1/6}d^3+R^{2/3}m^{-1/12}d^{3/2}+R^{7/6}m^{-1/12}d^{9/4})\\&\quad+\widetilde{O}((1-\rho)^\ell)+\widetilde{O}((d^{3/4}R^{1/2}+d^{3/2}R)(1-\rho)^{\ell/2})\Big]\\&\quad+\widetilde{O}\Big(N^{5/6}(1+d+d^{3/2}R+d^{7/4}R^{1/2}+d^{5/2}R)\Big).\end{aligned}$$

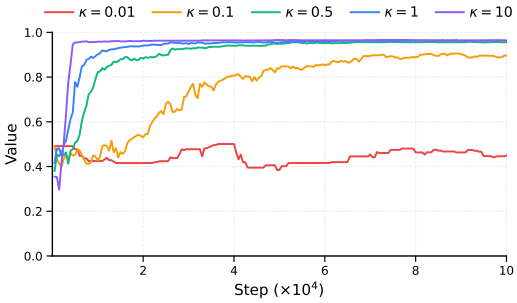


Figure 5: **Sensitivity to the mirror descent parameter κ in CheeseMaze.** The y-axis reports policy value, defined as the Monte Carlo estimate of the mean discounted return $\widehat{V} = \mathbb{E}[\sum_h \gamma^h r_h]$ over 50 evaluation episodes. Evaluation is performed every 500 training steps. Curves show the mean over 3 random seeds.

E Additional Experimental Results

E.1 Sensitivity to the Policy Stepsize

Figure 5 reports the policy value of our neural PMD on CheeseMaze ($\ell = 2$) for $\kappa \in \{0.01, 0.1, 0.5, 1, 10\}$. Performance is robust over $\kappa \in [0.1, 10]$; only the smallest setting $\kappa = 0.01$ leads to a slow update and noticeably degraded performance. We additionally observed (not shown) that significantly larger values of κ harm performance, as expected: the mirror-descent step becomes too aggressive and amplifies the inexact critic.

E.2 Hyperparameters

Table 1 summarizes the hyperparameters used in our experiments. All methods share the same superstate input encoding (concatenated one-hot observations and actions) and the same critic architecture; entries marked "–" indicate that the corresponding component is not part of the algorithm.

Table 1: Hyperparameters used in CheeseMaze experiments.

	Mirror Descent (ours)	**Q-learning**	**Actor-Critic**
Discount γ	0.99	0.99	0.99
Outer iterations T	200	200	200
Critic steps per iter K	500	500	500
Critic learning rate	10^{-3}	10^{-3}	10^{-3}
Critic hidden dim	64	64	64
Critic hidden layers	2	2	2
Actor learning rate	–	–	10^{-3}
Actor hidden dim	–	–	64
Actor hidden layers	–	–	1
Mirror-descent stepsize κ	0.1	–	–
ϵ-greedy schedule	–	$1.0 \to 0.05$ (linear)	–
Optimizer	Adam	Adam	Adam
Random seeds	$\{1, 2, 3\}$	$\{1, 2, 3\}$	$\{1, 2, 3\}$
Evaluation episodes	50	50	50
Evaluation interval (steps)	500	500	500

E.3 Compute Resources

All reported experiments were run on a single NVIDIA H200 GPU with 141GB HBM3e memory. The CheeseMaze experiments are lightweight, and each training run typically takes only a few minutes in our implementation. GPU acceleration was used mainly for convenience rather than necessity. The

experiments can be run on smaller GPUs and should also be feasible on CPU-only machines with longer runtime. No special storage resources are required, since the CheeseMaze environment is generated procedurally. The full set of reported experiments can be completed within a few hours.

E.4 Existing Assets and Licenses

We acknowledge the CleanRL project (Huang et al., 2022), which our implementation builds on for the experimental framework and implementation structure. CleanRL is publicly available and released under the MIT license. Our code is included in the supplemental material. No external datasets are used.

F Broader Impacts

This work is primarily foundational and aims to improve the understanding of reinforcement learning under partial observability. Potential positive impacts include practical applications such as autonomous driving, robotics, and resource allocation. Potential negative impacts are indirect: if such methods are deployed in autonomous decision-making systems without sufficient validation, they could lead to harmful decisions. In addition, scaling reinforcement learning methods to larger domains may increase computational and energy costs.