LAST-ITERATE GUARANTEES FOR ONLINE REINFORCEMENT LEARNING IN STRUCTURED CONSTRAINED MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

In safety-critical applications, deployment uses a single policy, whose performance and constraint satisfaction should hold directly rather than only for an average or mixture of training policies. This motivates last-iterate guarantees in constrained reinforcement learning. Recent progress has established such guarantees in exact-gradient or tabular online settings, yet scalable results for structured large-state problems remain open. We develop a general, statistically efficient framework for last-iterate convergence in structured Constrained MDPs (CMDPs). Our analysis separates contraction of the regularised primal-dual dynamics from actor approximation and statistical errors in policy evaluation under online exploration. This enables model-free on- and off-policy learning with structured function approximation: optimistic policy evaluation avoids explicit transition-model construction, while a compact parametric actor avoids maintaining mixtures or histories of past policies. We instantiate the framework for linear CMDPs and general function approximation, obtaining representation-dependent complexity and improved target-accuracy dependence over prior optimistic regularised primal-dual analyses. We further validate the stabilising effect predicted by our theory on a synthetic linear CMDP: the regularised method exhibits stable last-iterate behaviour, whereas its unregularised counterpart shows larger oscillations.

1 INTRODUCTION

Constrained Reinforcement Learning (RL) studies sequential decision making in which an agent maximizes reward while satisfying explicit safety, resource, or performance constraints. Constrained Markov decision processes (CMDPs) provide a standard framework for this problem (Altman, 1999; Efroni et al., 2020). Safe robotics offers a concrete motivation: reinforcement learning is increasingly used for locomotion (Margolis et al., 2024), manipulation (Kroemer et al., 2021), and navigation (Xu et al., 2025), where optimising reward alone is insufficient. A robot must also avoid unsafe behaviour, and deployment ultimately relies on a single learned policy rather than an average of the policies encountered during training. Other applications in which the deployed policy must satisfy constraints are discussed in Appendix A. Moreover, in these applications, state spaces are often large or effectively continuous, and practical methods must remain computationally manageable. These applications therefore make three general requirements especially clear: reliability of the final iterate, scalability beyond tabular models, and efficient computation.

Safety requirements are often encoded through handcrafted rules, such as barrier-style constraints (Ames et al., 2017; Wang et al., 2017; Cheng et al., 2019), or through fixed reward penalties (Lee et al., 2024; Fu et al., 2022; Xu et al., 2025). Handcrafted rules can be difficult to design for complex long-horizon tasks, while a fixed scalarized reward need not recover an optimal constrained policy (Ding et al., 2023). Lagrangian-based primal-dual methods instead adapt the reward-constraint tradeoff during learning. Yet they also face a fundamental challenge: their iterates can oscillate, so average reward and constraint satisfaction do not guarantee that the final policy is near-optimal and feasible (Ding et al., 2023). For constrained learning, this distinction is essential: one ultimately deploys a single policy, not an average of all policies produced during training. Last-iterate convergence is therefore the natural theoretical objective.

Recent work has made important progress in this direction. In the exact setting, where the CMDP is known, Ding et al. (2023) showed that regularised policy-gradient primal-dual methods can stabilise

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.

the learning dynamics and achieve last-iterate convergence. More recently, Montenegro et al. (2024); Lu et al. (2026) established model-free last-iterate convergence for general parameterised policies using stochastic policy gradients. This moves the theory beyond tabular policies, but relies on conditions ensuring that on-policy gradient estimates make global progress and does not explicitly address exploration. In a complementary direction, Müller et al. (2024); Zuo et al. (2026) used optimism under uncertainty to address exploration, but their online framework is tabular and model-based. Thus, a gap remains between scalable model-free policy optimisation and explicit online exploration. The tabular dependence is inadequate for large state spaces, while constructing an explicit optimistic transition model can be computationally demanding. Moreover, the implicit optimistic policy update accumulates past critic estimates; without a compact parametric representation, the policy representation and its statistical complexity can grow with the number of iterations (Lin et al., 2025).

A natural way to move beyond the tabular regime is to exploit structured function approximation, including linear CMDPs (Jin et al., 2020; Ghosh et al., 2024) and more general value-function classes. Under suitable structural assumptions, complexity can depend on the feature dimension or function-class complexity rather than the number of states. Moreover, when the state space is large, model-free methods can be computationally more efficient since they avoid explicitly constructing a transition model. Extending online last-iterate guarantees to these settings requires preserving contraction across optimisation steps while controlling both statistical policy-evaluation error and approximation error from a compact parametric actor. These challenges call for a unified, estimator-agnostic analysis.

Our central idea is a general framework that separates optimisation from statistical learning. The master analysis controls the contraction of the regularised primal-dual dynamics, while modular actor and critic conditions quantify the approximation and estimation errors arising from actor fitting and policy evaluation under online exploration. Consequently, the same convergence argument accommodates on- and off-policy data collection and different structured CMDP classes whenever these errors can be controlled. The model-free critic exploits problem structure without constructing a transition model, while the actor compresses the history-dependent policy update into a fixed-dimensional parameterization rather than storing all past critics.

Contributions. Motivated by these limitations, we ask whether *a model-free primal-dual method can achieve last-iterate convergence in structured CMDPs while explicitly addressing exploration from online interaction*. We answer this question affirmatively. Our main contributions are threefold:

First, we develop a general algorithmic and analytical framework for last-iterate convergence of online regularised policy-gradient primal-dual methods. It separates the contraction analysis from the approximation and statistical errors introduced by actor fitting, policy evaluation, and exploration, thereby accommodating both on- and off-policy data collection and different structured CMDP classes whenever these errors can be controlled. Its model-free optimistic critic and compact parametric actor avoid explicit transition-model construction and storage of the full critic history.

Second, we instantiate the framework for linear CMDPs and general function approximation, with statistical complexity governed by the feature dimension or function-class complexity rather than state-space size. Moreover, the linear-CMDP instantiation yields a computationally efficient algorithm.

Third, we sharpen the dependence on the target accuracy ϵ in the regularised primal-dual analysis of Müller et al. (2024). In the comparable online tabular setting without actor approximation error, this improves the last-iterate episode complexity from $\widetilde{O}(\epsilon^{-14})$ to $\widetilde{O}(\epsilon^{-6})$.

There is a large literature on constrained reinforcement learning and primal-dual methods for CMDPs; we defer a detailed discussion to Appendix A. Table A lists the sample complexities of CMDP methods, highlighting that our work achieves the tightest rate among existing methods for online CMDPs. Moreover, to the best of our knowledge, we establish the first non-asymptotic *last-iterate guarantees* for online RL in *structured CMDPs* while *explicitly addressing the exploration problem*.

2 Preliminaries

For $K \in \mathbb{N}_+$, let $[K] \triangleq \{1, \dots, K\}$. For a finite set $\mathcal{X}$, let $\Delta(\mathcal{X})$ denote the probability simplex over $\mathcal{X}$, and for $p, q \in \Delta(\mathcal{X})$, let $\mathrm{KL}(p\|q) \triangleq \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$. The notations $\widetilde{O}$ and $\widetilde{\Omega}$ suppress polylogarithmic factors. For any $x \in \mathbb{R}$, let $[x]_+ \triangleq \max\{0, x\}$. For a matrix A, let $A^\dagger$ denote its Moore-Penrose pseudoinverse.

Episodic CMDPs. We consider an episodic constrained Markov decision process (CMDP) $M = (\mathcal{S}, \mathcal{A}, H, P, r, u, s_1)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space , $P = \{P_h\}_{h=1}^H$ is the transition kernel, $r = \{r_h\}_{h=1}^H$ is the reward function, and $u = \{u_{i,h}\}_{i\in[I],h\in[H]}$ are the utilities associated with the safety constraints, where I is the number of constraints. The learner observes stochastic rewards and utilities supported on $[0, 1]$, with conditional means $r_h(s, a)$ and $u_{i,h}(s, a)$, respectively.

A policy $\pi = \{\pi_h\}_{h=1}^H$ is a sequence of stochastic decision rules $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$. Let Π denote the class of these policies. For any policy π, let $d_h^\pi(s, a) \triangleq \mathbb{P}_\pi(s_h = s, a_h = a \mid s_1)$, $d_h^\pi(s) \triangleq \sum_{a\in\mathcal{A}} d_h^\pi(s, a)$ denote the occupancy measure and state marginal induced by π. For each (h, s, a), and every function $g : \mathcal{S} \to \mathbb{R}$, define $(P_h g)(s, a) \triangleq \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[g(s')]$. For any $r' = \{r'_h\}_{h=1}^H$ defined on state-action pairs, define $Q^\pi_{r',h}(s, a) \triangleq \mathbb{E}_\pi\left[\sum_{t=h}^H r'_t(s_t, a_t) \,\middle|\, s_h = s, a_h = a\right]$. For any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $h \in [H]$, define the π-Bellman operator by

$$(\mathbb{B}^\pi_{r',h} f)(s, a) \triangleq \begin{cases} r'_h(s, a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a),\, a'\sim\pi_{h+1}(\cdot|s')}[f(s', a')], & h < H, \\ r'_H(s, a), & h = H. \end{cases}$$

Then $Q^\pi_{r',h}$ satisfies the Bellman recursion

$$Q^\pi_{r',h}(s, a) = (\mathbb{B}^\pi_{r',h} Q^\pi_{r',h+1})(s, a), \qquad V^\pi_{r',h}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[Q^\pi_{r',h}(s, a)\right].$$

Here $Q^\pi_{r',H+1} \equiv 0$. In particular, V^π_r denotes the expected cumulative reward, and $V^\pi_u \triangleq \left(V^\pi_{u_1}, \ldots, V^\pi_{u_I}\right) \in \mathbb{R}^I$ collects the expected cumulative utilities. When $h = 1$ and the initial state is s_1, we write $V^\pi_{r'} \triangleq V^\pi_{r',1}(s_1)$. In this work, our goal is to solve the constrained control problem

$$\max_{\pi\in\Pi} V^\pi_r \qquad \text{s.t.} \qquad V^\pi_{u_i} \geq c_i, \quad \forall i \in [I], \tag{1}$$

where $c \in [0, H]^I$ is the vector of safety thresholds. Let $\pi^\star$ denote an optimal feasible policy. We also define the log-linear policy class Π_{lin}: given an actor feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_a}$ satisfying $\|\varphi(s, a)\|_2 \leq 1$ and parameters $\omega = \{\omega_h\}_{h=1}^H$, define

$$\pi_h(a \mid s, \omega) = \frac{\exp\big(\langle\varphi(s, a), \omega_h\rangle\big)}{\sum_{a'\in\mathcal{A}} \exp\big(\langle\varphi(s, a'), \omega_h\rangle\big)}. \tag{2}$$

We denote the corresponding policy by $\pi(\omega)$. The real-valued scores supplied to a softmax are called logits; here, for a log-linear policy, the logit for action a at state s and stage h is $\langle\varphi(s, a), \omega_h\rangle$.

Regularised primal-dual formulation. To obtain a bounded dual domain, we impose the standard Slater condition.

Assumption 2.1 (Slater condition)**.** There exist a policy $\bar\pi \in \Pi$ and a constant $\xi > 0$ such that $V^{\bar\pi}_{u_i} \geq c_i + \xi$ for every $i \in [I]$.

It implies $\|\lambda^\star\|_1 \leq H/\xi$ for an optimal dual solution $\lambda^\star$ (Ying et al., 2022); hence, we work on $\Lambda \triangleq [0, \lambda_{\max}]^I$ with $\lambda_{\max} \geq H/\xi$. By strong duality, the constrained problem (1) is equivalent to the Lagrangian saddle-point problem (Altman, 1999; Paternain et al., 2019)

$$\max_{\pi\in\Pi} \min_{\lambda\in\Lambda} L(\pi, \lambda), \qquad L(\pi, \lambda) \triangleq V^\pi_r + \lambda^\top (V^\pi_u - c). \tag{3}$$

Although (3) involves the scalarized reward $r + \lambda^\top u$, it is important not to confuse the CMDP with a single unconstrained MDP obtained by fixing λ. optimising a fixed scalarization is generally not equivalent to solving the constrained problem, since the appropriate tradeoff must be adjusted through the dual dynamics (Altman, 1999; Ding et al., 2023). Because vanilla primal-dual dynamics can oscillate, we add entropy regularisation to the primal variable and quadratic regularisation to the dual variable to stabilise the last iterate (Ding et al., 2023).

For any policy π, define the entropy cost $\psi^\pi_h(s, a) \triangleq -\log \pi_h(a \mid s)$, and the entropy-value function $V^\pi_{\psi^\pi} \triangleq \mathbb{E}_\pi\left[\sum_{h=1}^H \psi^\pi_h(s_h, a_h) \,\middle|\, s_1\right]$. For a regularisation parameter $\tau > 0$, define the regularised Lagrangian

$$L_\tau(\pi, \lambda) \triangleq V^\pi_r + \lambda^\top (V^\pi_u - c) + \tau V^\pi_{\psi^\pi} + \frac{\tau}{2}\|\lambda\|_2^2, \qquad (\pi, \lambda) \in \Pi \times \Lambda. \tag{4}$$

Algorithm 1 Online RPGPD

1: **Input**: $K, N, \eta, \tau, \Lambda$, a sampling scheme (on-policy or off-policy), ActorFit, and OPE.
2: **Initialize**: set $\omega_h^1 = 0$ for every h, $\lambda_1 = 0$, and $\mathcal{D}^0 = \varnothing$; define $\pi^1 = \pi(\omega^1)$.
3: **for** $k = 1, \dots, K-1$ **do**
4: Collect data: $\mathcal{D}^k \leftarrow \begin{cases} \text{on-policy}: & \{N \text{ fresh traj. } \sim \pi^k\}, \\ \text{off-policy}: & \mathcal{D}^{k-1} \cup \{1 \text{ fresh traj. } \sim \pi^k\}. \end{cases}$
5: Update critic: $\{Q^k_{z_k,h}\}_{h=1}^H, V_u^k \leftarrow \mathrm{OPE}(\pi^k, \lambda_k, \mathcal{D}^k)$.
6: Update actor: $\omega^{k+1} \leftarrow \mathrm{ActorFit}(\omega^k, Q^k_{z_k}, \eta)$ and define $\pi^{k+1} = \pi(\omega^{k+1})$.
7: Update dual variable: $\lambda_{k+1} \leftarrow \mathrm{Proj}_\Lambda[(1-\eta\tau)\lambda_k - \eta(V_u^k - c)]$.
8: **end for**
9: **Output**: π^K.

Let $(\pi^\star_\tau, \lambda^\star_\tau)$ denote a saddle point of L_τ over $\Pi \times \Lambda$. Given the current iterate (π^k, λ_k), define $\psi_{k,h}(s,a) \triangleq -\log \pi_h^k(a \mid s)$, $z_k \triangleq r + \lambda_k^\top u + \tau\psi_k$. Following (Müller et al., 2024), the exact primal-dual update admit closed-form expressions

$$\pi_h^{k+1}(\cdot \mid s) \propto \pi_h^k(\cdot \mid s) \exp\big(\eta Q^{\pi^k}_{z_k,h}(s,\cdot)\big), \quad \lambda_{k+1} = \mathrm{proj}_\Lambda\big((1-\eta\tau)\lambda_k - \eta(V_u^{\pi^k} - c)\big). \tag{5}$$

Moreover, uniform initialization and exponential primal update ensure that $\pi_h^k(a \mid s) > 0$ for every (k,h,s,a), thus $\psi_{k,h}(s,a)$ is finite.

3 Master Algorithm and Analysis

So far, we have presented the regularised primal-dual formulation in an idealised setting. We now move to the online setting, where the CMDP is unknown, and policy gradients must be estimated from data collected through interaction with the environment. To separate the convergence analysis from model-specific estimation and optimisation details, we formulate an oracle-based algorithm with two abstract components: an optimistic policy-evaluation oracle (critic) and an actor-fitting oracle. We state sufficient conditions on the critic and actor oracles. Under these conditions, we derive a master contraction recursion that makes explicit how the two sources of error affect the last iterate. Concrete policy-evaluation constructions are developed in Sections 4 and 5.

Master algorithm. Let $\mathrm{KL}_{k,h}(s) \triangleq \mathrm{KL}\big(\pi^\star_{\tau,h}(\cdot \mid s) \,\|\, \pi_h^k(\cdot \mid s)\big)$, and define the potential $\Phi_k \triangleq \sum_{h=1}^H \mathbb{E}_{s\sim d_h^{\pi^\star_\tau}}\Big[\mathrm{KL}_{k,h}(s)\Big] + \frac{1}{2}\|\lambda_k - \lambda^\star_\tau\|_2^2$. Let $\omega^k = (\omega_h^k)_{h=1}^H$ and $Q^k_{z_k} = (Q^k_{z_k,h})_{h=1}^H$ denote the collections of stagewise actor parameters and critic estimates, respectively. At iteration k, the Online Policy-Evaluation oracle (OPE) returns the optimistic critic $Q^k_{z_k}$, together with the utility estimates V_u^k. The actor-fitting oracle ActorFit then returns ω^{k+1} by fitting the H stagewise logit targets $\{\langle \varphi, \omega_h^k\rangle + \eta Q^k_{z_k,h}\}_{h=1}^H$. The on- and off-policy variants differ only in the dataset supplied to the OPE: in the on-policy mode, $\mathcal{D}^k$ consists of a fresh batch of N independent trajectories generated by π^k; in the off-policy mode, $\mathcal{D}^k$ augments $\mathcal{D}^{k-1}$ with a single trajectory generated by π^k. The resulting Regularsied Policy-Gradient Primal Dual (RPGPD) procedure is presented in Algorithm 1.

Policy-evaluation oracle. We first state the critic requirements used by the master analysis.

Assumption 3.1 (Policy-evaluation oracle)**.** With probability at least $1-\delta$, simultaneously for every $k \in [K-1]$, the oracle returns $Q^k_{z_k}$ and V_u^k such that, with $Q^k_{z_k,H+1} \equiv 0$ and $V^k_{z_k,h}(s) \triangleq \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}[Q^k_{z_k,h}(s,a)]$ and $V^k_{z_k} \triangleq V^k_{z_k,1}(s_1)$:

(i) **Optimism:** $Q^k_{z_k,h} \geq \mathbb{B}^{\pi^k}_{z_k,h} Q^k_{z_k,h+1}$ pointwise for every $h \in [H]$;

(ii) **Consistency:** there is $\beta_k \geq 0$ for which $V^k_{z_k} - V^{\pi^k}_{z_k} \leq \beta_k$ and $\max_{i\in[I]} |V^k_{u_i} - V^{\pi^k}_{u_i}| \leq \beta_k$;

(iii) **Boundedness:** $0 \leq V^k_{u_i} \leq H$ and, for every (h,s),

$$\sum_{a\in\mathcal{A}} \pi_h^k(a \mid s) \exp\big(\eta Q^k_{z_k,h}(s,a)\big)\big(Q^k_{z_k,h}(s,a)\big)^2 \leq C_{\eta,\tau,\Lambda,k}. \tag{6}$$

The three conditions capture standard requirements for optimistic online policy evaluation. Optimism makes $Q^k_{z_k}$ an upper bound for $Q^{\pi^k}_{z_k}$, a common mechanism in online RL for handling uncertainty and encouraging exploration. Consistency keeps the estimated regularised return and utilities close to their true values, with error controlled by β_k, while boundedness controls the dual update and the second-order term of the exponential actor update. Importantly, these are modular oracle-level conditions: they do not prescribe a particular estimator or sampling scheme and can be verified by suitable constructions for both on- and off-policy data collection and for structured CMDP classes, including linear CMDPs.

Actor oracle. Having specified the critic requirements, we next characterize how accurately the actor oracle must approximate the ideal policy update. Given $Q^k_{z_k,h}$, the ideal entropy-regularised mirror-ascent step is

$$\widetilde{\pi}^{k+1}_h(a \mid s) \propto \pi^k_h(a \mid s) \exp\big(\eta Q^k_{z_k,h}(s,a)\big),$$

whereas ActorFit returns a policy π^{k+1} in the log-linear policy class. The master analysis requires only that, for some $\epsilon_{\mathrm{act}} \geq 0$ and every $(k,h,s) \in [K-1] \times [H] \times \mathcal{S}$, the fitted update satisfies

$$\mathrm{KL}\big(\pi^\star_{\tau,h}(\cdot \mid s) \,\|\, \pi^{k+1}_h(\cdot \mid s)\big) - \mathrm{KL}\big(\pi^\star_{\tau,h}(\cdot \mid s) \,\|\, \widetilde{\pi}^{k+1}_h(\cdot \mid s)\big) \leq \eta \epsilon_{\mathrm{act}}. \tag{7}$$

This condition measures the KL progress lost by projecting the ideal update onto the log-linear class Π_{lin} and contributes $\eta H \epsilon_{\mathrm{act}}$ to the master recursion. Although explicit, the ideal update accumulates past critics and need not remain in the log-linear class. Following Lin et al. (2025), ActorFit approximates each target increment $\eta Q^k_{z_k,h}$ using a linear combination of the actor features φ. Moreover, we can relax this approximation requirement by observing that, for a state-dependent baseline $b : \mathcal{S} \to \mathbb{R}$, subtracting $b(s)$ multiplies every unnormalized action weight at state s by the same factor $e^{-\eta b(s)}$, which cancels upon normalization:

$$\frac{\pi^k_h(a \mid s) \exp\big(\eta(Q^k_{z_k,h}(s,a) - b(s))\big)}{\sum_{a' \in \mathcal{A}} \pi^k_h(a' \mid s) \exp\big(\eta(Q^k_{z_k,h}(s,a') - b(s))\big)} = \widetilde{\pi}^{k+1}_h(a \mid s).$$

In words, a log-linear policy may represent the action differences of $Q^k_{z_k,h}$ without representing a state-dependent term $b(s)$ common to all actions. Thus, choosing b to remove such a term can reduce the actor approximation error compared with fitting $Q^k_{z_k,h}$ directly as in Lin et al. (2025).

Furthermore, fitting the actor over every triplet in $\mathcal{S} \times \mathcal{A}^2$ can be infeasible when the state space is large. Adapting the coreset approach of Lin et al. (2025) to action differences, we instead evaluate the actor loss on a fixed, finite set of triplets $\mathcal{D}_{\mathrm{exp}}$.

To state the actor approximation condition, for any triplet (s,a,a'), define the action differences $\Delta\varphi(s,a,a') \triangleq \varphi(s,a) - \varphi(s,a')$ and $\Delta Q^k_{z_k,h}(s,a,a') \triangleq Q^k_{z_k,h}(s,a) - Q^k_{z_k,h}(s,a')$.

Assumption 3.2 (Log-linear actor approximation)**.** There exists a fixed finite weighted coreset $(\mathcal{D}_{\mathrm{exp}}, \rho_{\mathrm{exp}})$, where $\mathcal{D}_{\mathrm{exp}} \subset \mathcal{S} \times \mathcal{A} \times \mathcal{A}$ and $\rho_{\mathrm{exp}} \in \Delta(\mathcal{D}_{\mathrm{exp}})$, with

$$G_{\mathrm{exp}} \triangleq \sum_{(s,a,a') \in \mathcal{D}_{\mathrm{exp}}} \rho_{\mathrm{exp}}(s,a,a') \Delta\varphi(s,a,a') \Delta\varphi(s,a,a')^\top, \quad \kappa_G \triangleq \sup_{s,a,a'} \|\Delta\varphi(s,a,a')\|_{G^\dagger_{\mathrm{exp}}} < \infty.$$

We require $\Delta\varphi(s,a,a') \in \mathrm{range}(G_{\mathrm{exp}})$ for every (s,a,a'). Moreover, there is an $\epsilon_{\mathrm{bias}} \geq 0$ such that

$$\max_{\substack{k \in [K-1] \\ h \in [H]}} \inf_{\substack{\omega \in \mathbb{R}^{d_{\mathrm{a}}} \\ b : \mathcal{S} \to \mathbb{R}}} \sup_{s,a} \big|\langle \varphi(s,a), \omega - \omega^k_h \rangle - \eta\big(Q^k_{z_k,h}(s,a) - b(s)\big)\big| \leq \eta \epsilon_{\mathrm{bias}}. \tag{8}$$

A detailed comparison explaining why the Assumption 3.2 is weaker than that of Lin et al. (2025) is deferred to Appendix B.1. The coreset $\mathcal{D}_{\mathrm{exp}}$ and its corresponding weight distribution ρ_{exp} can be constructed using G-optimal design, yielding $|\mathcal{D}_{\mathrm{exp}}| = \widetilde{O}(d_{\mathrm{a}})$ and $\kappa_G = O(\sqrt{d_{\mathrm{a}}})$ (Lattimore et al., 2020; Lin et al., 2025); see Appendix B.1 for details. Since this construction is addressed in the literature, we assume for simplicity that $(\mathcal{D}_{\mathrm{exp}}, \rho_{\mathrm{exp}})$ is given.

In implementation, we fit differences in Q-values between actions, which eliminate b without estimating it. In particular, for class Π_{lin}, ActorFit computes $\omega^{k+1} = (\omega^{k+1}_h)_{h=1}^H$ by solving

$$\omega^{k+1}_h \in \arg\min_{w_h \in \mathbb{R}^{d_{\mathrm{a}}}} \sum_{(s,a,a') \in \mathcal{D}_{\mathrm{exp}}} \rho_{\mathrm{exp}}(s,a,a') \left(\langle \Delta\varphi(s,a,a'), w_h - \omega^k_h \rangle - \eta \Delta Q^k_{z_k,h}(s,a,a')\right)^2. \tag{9}$$

Thus, the loss fits the ideal policy update $\widetilde{\pi}^{k+1}$ within the log-linear class Π_{lin}, without estimating the baseline function b. Here, ϵ_{bias} measures uniform approximation error, while κ_G controls its amplification under coreset fitting.

Proposition 3.3 (Log-linear actor certificate)**.** *Under Assumption 3.2, the* ActorFit *update in* (9) *satisfies* (7) *with* $\epsilon_{\text{act}} = 2(\kappa_G + 1)\epsilon_{\text{bias}}$.

The actor error ϵ_{act} measures how accurately the log-linear actor represents and fits the ideal update. It can decrease with a richer actor class and vanishes under exact representability and fitting, as for tabular actors with independent state-action logits, where $\epsilon_{\text{act}} = 0$.

Last-iterate convergence of Algorithm 1. Combining the abstract critic condition with the actor projection bound yields a geometric recursion for the master potential. Under the additional Slater and fixed-domain conditions, the same result converts this potential bound into last-iterate reward suboptimality and constraint-violation guarantees for the original CMDP.

Theorem 3.4 (General bound for the master algorithm)**.** *Fix* $K \geq 1$. *Let* $G_\tau \triangleq \sqrt{I}(H + \tau\lambda_{\max})$. *Under Assumption 3.1, suppose that* ActorFit *satisfies* (7) *and* $0 < \eta\tau \leq 1$. *Then, with probability at least* $1 - \delta$, *Algorithm 1 satisfies*

$$\Phi_K \leq e^{-\eta\tau(K-1)}\Phi_1 + \sum_{k=1}^{K-1}(1-\eta\tau)^{K-1-k}\left[\frac{\eta^2}{2}\left(HC_{\eta,\tau,\Lambda,k} + G_\tau^2\right) + \eta(1 + I\lambda_{\max})\beta_k + \eta H\epsilon_{\text{act}}\right]. \tag{10}$$

Moreover, if Assumption 2.1 holds, $0 < \tau \leq 1$, *and* $\lambda_{\max} = 2H\log(e|\mathcal{A}|)/\xi$, *then, for every* $i \in [I]$,

$$V_r^{\pi^\star} - V_r^{\pi^K} \leq H^{3/2}\sqrt{2\Phi_K} + \tau H\log|\mathcal{A}|, \quad c_i - V_{u_i}^{\pi^K} \leq H^{3/2}\sqrt{2\Phi_K} + \frac{\tau H(1 + \tau\log|\mathcal{A}|)}{\xi}. \tag{11}$$

The key distinction from Müller et al. (2024, Lemma 4.2) is that our choice $\lambda_{\max} = 2H\log(e|\mathcal{A}|)/\xi$, independent of the target accuracy ϵ, gives the strict inequality $\lambda^\star_{\tau,i} < \lambda_{\max}$ for every $i \in [I]$. Therefore, dual optimality yields $[c_i - V_{u_i}^{\pi^\star_\tau}]_+ = \tau\lambda^\star_{\tau,i} = O(\tau)$. Thus, our conversion keeps $\lambda_{\max}$ fixed and permits $\tau = \Theta(\epsilon)$, which is the primary source of the sharper ϵ dependence later in Section 4 and 5. By contrast, Müller et al. (2024)'s radius-dependent bound balances terms of order $1/\lambda_{\max}$ and $\tau\lambda_{\max}$, requiring $\lambda_{\max} = \Theta(\epsilon^{-1})$ and $\tau = \Theta(\epsilon^2)$ to attain $O(\epsilon)$ error. Remark C.1 gives the detailed comparison, and Appendix B.2 records its sample-complexity consequences.

Environment structure and sampling enter the recursion only through the oracle certificates β_k, $C_{\eta,\tau,\Lambda,k}$, and ϵ_{act}. The conversion bounds in Theorem 3.4 then translate a bound on Φ_K into last-iterate reward and constraint guarantees for the original CMDP. Unlike the tabular model-based framework of Müller et al. (2024), this analysis supports model-free evaluation with linear or general function approximation and either on- or off-policy sampling, provided the oracle assumptions hold.

4 Last-Iterate Convergence for Linear CMDPs

This section instantiates the policy-evaluation interface of Section 3 under linear-CMDP realizability. We first give a linear policy-evaluation routine shared by both sampling modes, followed by joint oracle and last-iterate guarantees for fresh on-policy batches and cumulative off-policy data. Throughout this section, we work under the following linear-CMDP realizability assumption.

Assumption 4.1 (Linear CMDP)**.** There exist a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_c}$, signed measures $\{\mu_h\}_{h=1}^H$ with $\mu_h(\cdot) \in \mathbb{R}^{d_c}$, reward parameters $\{\upsilon_{r,h}\}_{h=1}^H \subset \mathbb{R}^{d_c}$, and utility parameters $\{\upsilon_{u_i,h}\}_{i\in[I],\,h\in[H]} \subset \mathbb{R}^{d_c}$ such that for all (s,a,h), $P_h(\cdot \mid s,a) = \langle\phi(s,a), \mu_h(\cdot)\rangle$, $r_h(s,a) = \langle\phi(s,a), \upsilon_{r,h}\rangle$, $u_{i,h}(s,a) = \langle\phi(s,a), \upsilon_{u_i,h}\rangle$. Moreover, for all (s,a,h), $\|\phi(s,a)\|_2 \leq 1$, $\|\upsilon_{r,h}\|_2 \leq \sqrt{d_c}$, $\|\upsilon_{u_i,h}\|_2 \leq \sqrt{d_c}$, $\left\|\left(|\mu_h^{(1)}|(\mathcal{S}), \ldots, |\mu_h^{(d_c)}|(\mathcal{S})\right)\right\|_2 \leq \sqrt{d_c}$.

Linear policy-evaluation oracle. The policy-evaluation computation is shared by both sampling modes. Given a trajectory dataset $\mathcal{D}$, let $\mathcal{D}_h$ denote its stage-h transitions and let Σ_h^k be their regularised empirical covariance, as specified in Algorithm 2. Initialize $V_{r,H+1}^k = V_{u_i,H+1}^k = V_{\psi,H+1}^k = 0$. Index the samples by ℓ and set $y_{r,h}^\ell = r_h^\ell$, $y_{u_i,h}^\ell = u_{i,h}^\ell$, and $y_{\psi,h}^\ell = 0$. Then, for $j \in \{r, u_1, \ldots, u_I, \psi\}$, the ridge-regression coefficients are

Algorithm 2 Linear OPE$(\pi^k, \lambda_k, \mathcal{D})$

1: **Input**: π^k, λ_k, a trajectory dataset $\mathcal{D}$, and confidence radii $\alpha_r, \alpha_u, \alpha_\psi$.
2: Set $V^k_{r,H+1} = V^k_{u_i,H+1} = V^k_{\psi,H+1} = 0$ for every $i \in [I]$.
3: **for** $h = H, H-1, \dots, 1$ **do**
4: Set $\Sigma^k_h \leftarrow I_{d_c} + \sum_{(s_h,a_h,r_h,u_h,s_{h+1})\in\mathcal{D}_h} \phi(s_h,a_h)\phi(s_h,a_h)^\top$.
5: Compute the regression coefficients in (12).
6: Set $\alpha_{u_i} \equiv \alpha_u$ for $i \in [I]$ and $b^k_{j,h}(s,a) = \alpha_j \|\phi(s,a)\|_{(\Sigma^k_h)^{-1}}$ for $j \in \{r, u_1, \dots, u_I, \psi\}$. For $j \in \{r, u_1, \dots, u_I\}$, define

$$\begin{aligned} Q^k_{j,h} &= \mathrm{Truncate}_{[0,H+1-h]}\big(\langle \phi, \widehat{\theta}^k_{j,h}\rangle + b^k_{j,h}\big), \\ Q^k_{\psi,h}(s,a) &= \psi_{k,h}(s,a) + \mathrm{Truncate}_{[0,(H-h)\log|\mathcal{A}|]}\big(\langle \phi(s,a), \widehat{\theta}^k_{\psi,h}\rangle + b^k_{\psi,h}(s,a)\big). \end{aligned} \tag{13}$$

7: Set $V^k_{j,h}(s) = \mathbb{E}_{a\sim\pi^k_h(\cdot|s)}[Q^k_{j,h}(s,a)]$ for $j \in \{r, u_1, \dots, u_I, \psi\}$.
8: **end for**
9: Return $Q^k_{z_k,h} = Q^k_{r,h} + \lambda_k^\top Q^k_{u,h} + \tau Q^k_{\psi,h}$ and $V^k_u = (V^k_{u_1,1}(s_1), \dots, V^k_{u_I,1}(s_1))$.

$$\widehat{\theta}^k_{j,h} = (\Sigma^k_h)^{-1} \sum_{\ell=1}^{|\mathcal{D}_h|} \phi(s^\ell_h, a^\ell_h)\big(y^\ell_{j,h} + V^k_{j,h+1}(s^\ell_{h+1})\big). \tag{12}$$

Algorithm 2 summarizes the resulting backward recursion. At each stage, it evaluates the coefficients in (12), adds elliptical bonuses, truncates the component critics to their natural ranges, and propagates their policy averages. The reward, utility, and entropy components are estimated separately before being combined using λ_k and τ. In particular, the immediate entropy cost $\psi_{k,h}$ is known from π^k, so only its value function $V^k_{\psi,h}$ is estimated.

On- and off-policy guarantees. The same linear oracle supports both modes. At iteration k, on-policy evaluation uses N fresh trajectories, whereas off-policy evaluation appends one trajectory to the cumulative dataset. The following lemma gives the corresponding certificates.

Lemma 4.2 (Linear policy-evaluation certificates). *Under Assumptions 3.2 and 4.1, suppose* $0 < \tau \le 1$, $0 < \eta\tau \le 1/4$, *and* $\eta H(1 + I\lambda_{\max} + \tau \log|\mathcal{A}|) \le 1/4$. *For each sampling mode, there exist mode-specific confidence radii such that Algorithm 2 satisfies Assumption 3.1 with probability at least* $1-\delta$. *For the on-policy mode,*

$$\sup_{k<K} \beta_k = \widetilde{O}\left(\big(1 + I\lambda_{\max} + \tau\log|\mathcal{A}|\big)\frac{d_c^{3/2}H^2}{\sqrt{N}}\right). \tag{14}$$

For the off-policy mode,

$$\sum_{k=1}^{K-1} \beta_k = \widetilde{O}\left(\big(1 + I\lambda_{\max} + \tau\log|\mathcal{A}|\big)H^2\sqrt{d_c(d_c^2 + d_a)K}\right). \tag{15}$$

Moreover, in both modes, $\sup_{k<K} C_{\eta,\tau,\Lambda,k} = \widetilde{O}\left(H^2\big(1 + I\lambda_{\max} + \tau\log|\mathcal{A}|\big)^2\right)$.

Theorem 4.3 (On- and off-policy linear-CMDP guarantees). *Fix* $\epsilon, \delta \in (0,1)$. *Under Assumptions 2.1, 4.1, and 3.2, run Algorithm 1 with Algorithm 2 in either sampling mode and use the corresponding parameter choices in the appendix. Then, with probability at least* $1-\delta$,

$$\max\left\{\big[V_r^{\pi^\star} - V_r^{\pi^K}\big]_+, \max_{i\in[I]}\big[c_i - V_{u_i}^{\pi^K}\big]_+\right\} \le \widetilde{O}\left(\epsilon + \frac{H^{5/3}}{\xi^{1/3}}\epsilon_{\mathrm{act}}^{1/3}\right). \tag{16}$$

For the on-policy mode, use the parameter choices in (69) of Appendix D.3. The total number of on-policy episode rollouts satisfies

$$NK = \widetilde{O}\left(\frac{d_c^3 I^6 H^{22}}{\xi^6\epsilon^6}\max\left\{1, \min\left\{\frac{H^4}{\xi^4\epsilon^4}, \frac{1}{H^{8/3}\xi^{8/3}\epsilon_{\mathrm{act}}^{4/3}}\right\}\right\}\right). \tag{17}$$

For the off-policy mode, use the parameter choices in (104) *of Appendix D.5. The number of off-policy episode rollouts satisfies*

$$K = \widetilde{O}\left(\frac{d_{\mathrm{c}}(d_{\mathrm{c}}^2 + d_{\mathrm{a}})I^4H^{14}}{\xi^4\epsilon^4}\max\left\{1, \min\left\{\frac{H^2}{\xi^2\epsilon^2}, \frac{1}{H^{4/3}\xi^{4/3}\epsilon_{\mathrm{act}}^{2/3}}\right\}\right\}\right). \tag{18}$$

Beyond any dependence encoded by the representation dimensions d_{c} and d_{a}, both guarantees depend on $|\mathcal{A}|$ only through logarithmic factors and contain no separate polynomial factor in $|\mathcal{A}|$. This is a substantial improvement in action-space dependence over the online tabular bound of Müller et al. (2024, Theorem 5.1), which is polynomial in $|\mathcal{A}|$. We refer the reader to Appendix B.3 for a detailed explanation of this improvement.

Moreover, the sharper dependence on ϵ follows from the tighter conversion in Theorem 3.4 and the contraction analysis. Unlike the radius-dependent conversion of Müller et al. (2024, Lemma 4.2), our argument avoids the inverse-radius term and preserves geometric contraction with less conservative choices of τ and η. In particular, under a direct tabular specialization, our off-policy bound scales as $\widetilde{O}(\epsilon^{-6})$, improving upon the $\widetilde{O}(\epsilon^{-14})$ accuracy dependence implied by the model-based tabular analysis (Müller et al., 2024). Appendix B.2 provides the full sample-complexity and tabular comparison.

Although our off-policy sample complexity does not attain the $\Omega(\epsilon^{-2})$ lower bound (Domingues et al., 2021), its accuracy dependence matches the $O(\epsilon^{-6})$ iteration complexity of RPGPD (Ding et al., 2023). Thus, sampling adds only an additional multiplicative factor depending on the representation complexity.

5 General Function Approximation

This section extends the policy-evaluation construction of Section 4 to general critic classes while keeping the master algorithm and log-linear actor of Section 3. Here, we state only the function-class condition, the oracle idea, and the final guarantees; precise definitions, algorithms, oracle certificates, and proofs are deferred to Appendix E.

Function classes and complexity. Let $\mathcal{X} \triangleq \mathcal{S}\times\mathcal{A}$ and $\mathcal{J} = \{r, u_1, \dots, u_I, \psi\}$. For every $h \in [H]$, let $\mathcal{F}_{r,h}$ and $\mathcal{F}_{u_i,h}$ take values in $[0, H+1-h]$, and let $\mathcal{F}_{\psi,h}$ take values in $[0, (H-h)\log|\mathcal{A}|]$. The entropy class models only $P_hV^k_{\psi,h+1}$, since the known immediate cost $\psi_{k,h}(s,a)$ is included explicitly in the action-value estimate.

Assumption 5.1 (Value closeness)**.** For every $h \in [H]$, every function $V : \mathcal{S} \to [0, H-h]$, and every function $W : \mathcal{S} \to [0, (H-h)\log|\mathcal{A}|]$, the following inclusions hold:

$$r_h + P_hV \in \mathcal{F}_{r,h}, \qquad u_{i,h} + P_hV \in \mathcal{F}_{u_i,h},\ i \in [I], \qquad P_hW \in \mathcal{F}_{\psi,h}.$$

This is a standard componentwise value-closeness condition, similar to that of Liu et al. (2023). We summarize the critic complexity by the scale-dependent eluder dimension d_{E}, which controls sequential confidence widths, and the domain metric entropy $\mathfrak{h}_{\mathcal{X}}$, which controls the discretization of data-dependent bonuses. Their precise definitions and evaluation scales are given in Appendix E.1.

General optimistic policy evaluation. The general oracle fits the reward, utility, and entropy-value critics backward, adds a nonnegative stable bonus to each component, adds $-\log\pi_h^k$ exactly to the entropy critic, and scalarizes only after componentwise fitting. The complete construction is Algorithm 4 in Appendix E.2; its on- and off-policy certificates are proved in Appendices E.5 and E.6.

Last-iterate guarantees. We obtain the sample complexity stated in the theorem below.

Theorem 5.2 (General last-iterate guarantees)**.** *Fix $\epsilon, \delta \in (0,1)$. Under Assumptions 2.1, 3.2, and 5.1, Algorithm 1, instantiated with either the on-policy or off-policy general oracle and the corresponding parameter choices specified below, guarantees, with probability at least $1-\delta$,*

$$\max\left\{\left[V_r^{\pi^\star} - V_r^{\pi^K}\right]_+, \max_{i\in[I]}\left[c_i - V_{u_i}^{\pi^K}\right]_+\right\} \le \widetilde{O}\left(\epsilon + \frac{H^{5/3}}{\xi^{1/3}}\epsilon_{\mathrm{act}}^{1/3}\right). \tag{19}$$

For the on-policy oracle, use the parameter choices in (143) of Appendix E.5. The total number of on-policy episode rollouts satisfies

$$NK = \widetilde{O}\left(\frac{d_{\mathrm{E}}^3 \mathfrak{h}_{\mathcal{X}} I^6 H^{22}}{\xi^6 \epsilon^6} \max\left\{1, \min\left\{\frac{H^4}{\xi^4 \epsilon^4}, \frac{1}{H^{8/3}\xi^{8/3}\epsilon_{\mathrm{act}}^{4/3}}\right\}\right\}\right). \tag{20}$$

For the off-policy oracle, use the parameter choices in (149) of Appendix E.6. The number of off-policy episodes satisfies

$$K = \widetilde{O}\left(\frac{d_{\mathrm{E}}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}}) I^4 H^{14}}{\xi^4 \epsilon^4} \max\left\{1, \min\left\{\frac{H^2}{\xi^2 \epsilon^2}, \frac{1}{H^{4/3}\xi^{4/3}\epsilon_{\mathrm{act}}^{2/3}}\right\}\right\}\right). \tag{21}$$

The theorem extends last-iterate convergence beyond linear realizability: its statistical complexity is governed by the critic-class quantities d_{E} and $\mathfrak{h}_{\mathcal{X}}$, rather than the number of states.

6 Experiment

We conduct a synthetic experiment to examine the mechanism and empirical effectiveness of regularisation. We consider an episodic linear CMDP with unknown dynamics under fresh on-policy and cumulative off-policy sampling. The learner uses sample-based optimistic policy evaluation, while the hidden model is used only for post-hoc evaluation. The log-linear actor is fitted using the action-difference objective in (9). We present the setup and results here and defer all details to Appendix F.

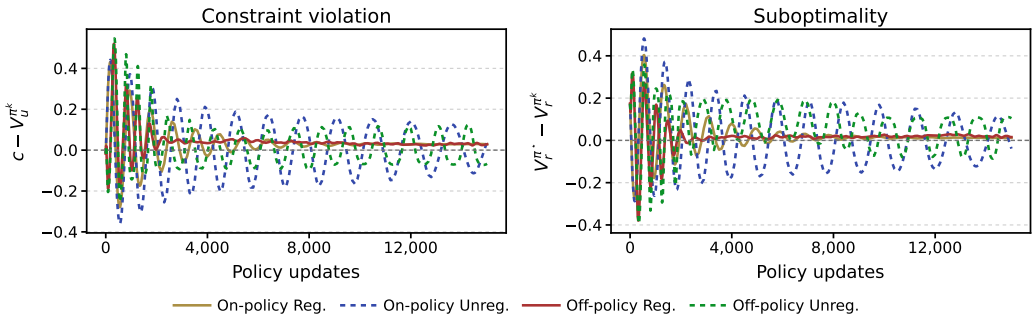

Figure 1: Online last-iterate performance in the synthetic linear CMDP over 1.5×10^4 policy updates.

We examine the last-iterate behaviour of the primal-dual dynamics on a small episodic linear CMDP whose optimal constrained policy and value functions can be computed exactly. We compare the regularised primal-dual method with a vanilla baseline using $\tau = 0$, both trained at the same constraint threshold c. Both use sample-based optimistic policy evaluation; exact evaluation is used only to assess the resulting policies. Figure 1 reports the signed constraint gap $c - V_u^{\pi^k}$ and signed reward gap $V_r^{\pi^\star} - V_r^{\pi^k}$. Since pointwise averaging across seeds can mask fluctuations, Figure 1 shows one seed in each sampling mode. The vanilla baseline (dashed lines) exhibits larger oscillations, whereas the regularised configuration (solid lines) damps its initial oscillations more quickly and yields substantially flatter trajectories in both sampling modes. We further compare within-run fluctuations of the unregularised and regularised methods using the 30-seed statistic summarised in Table B of Appendix F. In summary, regularised configurations substantially reduce last-iterate oscillations in both settings.

7 Conclusion

We developed a model-free framework for last-iterate convergence in structured CMDPs, with a computationally efficient linear-CMDP instantiation. By separating regularised primal-dual dynamics from modular actor and critic conditions, it accommodates on- and off-policy sampling and yields guarantees for linear CMDPs and general function approximation. Empirically, our synthetic experiment illustrates stable last-iterate behaviour. Extending the framework to continuous actions and neural actor-critic models is a promising direction toward practical application.

AI USE STATEMENT

We used AI tools to assist with the design and implementation of the experimental code and to improve the manuscript's grammar, word choice, and clarity. We take responsibility for the final content of this work, including the text, claims, code, results, produced with the aid of generative AI.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 2017.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022.

Archana Bura, Aria Hasanzadezonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *Advances in Neural Information Processing Systems*, 2020.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained MDPs. In *Advances in Neural Information Processing Systems*, 2023.

Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Başar, and Mihailo R. Jovanović. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs. *Journal of Machine Learning Research*, 2025.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, 2021.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps, 2020.

Zipeng Fu, Ashish Kumar, Ananye Agarwal, Haozhi Qi, Jitendra Malik, and Deepak Pathak. Coupling vision and proprioception for navigation of legged robots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained RL with linear function approximation. In *Advances in Neural Information Processing Systems*, 2022.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free RL. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Egor Gladin, Maksim Lavrik-Karmazin, Karina Zainullina, Varvara Rudenko, Alexander Gasnikov, and Martin Takac. Algorithm for constrained markov decision process with linear convergence. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Steve Hanneke. The star number and eluder dimension: Elementary observations about the dimensions of disagreement. In *Proceedings of the Thirty-Seventh Conference on Learning Theory*, 2024.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Conference on Learning Theory*, 2020.

Toshinori Kitamura, Tadashi Kozuno, Masahiro Kato, Yuki Ichihara, Soichiro Nishimori, Akiyoshi Sannai, Sho Sonoda, Wataru Kumagai, and Yutaka Matsuo. A policy gradient primal-dual algorithm for constrained mdps with uniform pac guarantees, 2024. URL https://arxiv.org/abs/2401.17780.

Toshinori Kitamura, Arnob Ghosh, Tadashi Kozuno, Wataru Kumagai, Kazumi Kasaura, Kenta Hoshino, Yohei Hosoe, and Yutaka Matsuo. Provably efficient RL under episode-wise safety in constrained MDPs with linear function approximation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of Machine Learning Research*, 2021.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Joonho Lee, Marko Bjelonic, Alexander Reske, Lorenz Wellhausen, Takahiro Miki, and Marco Hutter. Learning robust autonomous navigation and locomotion for wheeled-legged robots. *Science Robotics*, 2024.

Max Qiushi Lin, Reza Asad, Kevin Tan, Haque Ishfaq, Csaba Szepesvari, and Sharan Vaswani. Optimistic actor-critic with parametric policies: Unifying sample efficiency and practicality. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025.

Chang Liu, Yunfan Li, and Lin F. Yang. Near-optimal sample complexity for online constrained MDPs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvari. Optimistic natural policy gradient: a simple efficient policy optimization framework for online RL. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Policy optimization for constrained MDPs with provable fast global convergence. *arXiv preprint arXiv:2111.00552*, 2021.

Michael Lu, Max Qiushi Lin, Mo Chen, and Sharan Vaswani. Augmented lagrangian method for last-iterate convergence for constrained mdps, 2026. URL https://arxiv.org/abs/2605.11694.

Gabriel B. Margolis, Ge Yang, Kartik Paigwar, Tao Chen, and Pulkit Agrawal. Rapid locomotion via reinforcement learning. *The International Journal of Robotics Research*, 2024.

Washim Uddin Mondal and Vaneet Aggarwal. Last-iterate convergence of general parameterized policies in constrained MDPs. *Transactions on Machine Learning Research*, 2026.

Alessandro Montenegro, Marco Mussi, Matteo Papini, and Alberto Maria Metelli. Last-iterate global convergence of policy gradients for constrained reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Santiago Paternain, Luiz F. O. Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems*, 2019.

Santiago Paternain, Miguel Calvo-Fullana, Luiz F. O. Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2023.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for CMDP with adversarial loss. In *Advances in Neural Information Processing Systems*, 2020.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, 2013.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained MDPs via policy optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

Li Wang, Aaron D. Ames, and Magnus Egerstedt. Safety barrier certificates for collisions-free multirobot systems. *IEEE Transactions on Robotics*, 2017.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In *Advances in Neural Information Processing Systems*, 2020.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Zhefan Xu, Xinming Han, Haoyu Shen, Hanyu Jin, and Kenji Shimada. Navrl: Learning safe flight in dynamic environments. *IEEE Robotics and Automation Letters*, 2025.

Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained markov decision processes with entropy regularization. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Qian Zuo, Zhiyong Wang, and Fengxiang He. Near-constant strong violation and last-iterate convergence for online cmdps via decaying safety margins, 2026. URL https://arxiv.org/abs/2602.10917.

CONTENTS OF APPENDIX

A Comprehensive Related Works **14**

B Discussion of the Statistical Improvements **16**

B.1 On actor assumption 16

B.2 Origin of the sharper accuracy dependence 17

B.3 Actor- and action-space dependence 18

C Proofs for Section 3 **19**

C.1 Actor projection and the master recursion 19

C.2 Conversion from the regularised saddle point 21

D Proofs for Section 4 **23**

D.1 Technical lemmas for uniform confidence 23

D.2 Analysis of the on-policy oracle 25

D.3 Proof of the on-policy final rate 27

D.4 Analysis of the off-policy oracle 29

D.5 Proof of the off-policy final rate 32

E Proofs for Section 5 **33**

E.1 General function classes and complexity measures 33

E.2 Stable optimistic policy evaluation 34

E.3 Technical lemmas for uniform confidence 36

E.4 Eluder-width bounds 41

E.5 On-policy oracle certificate and tuning 43

E.6 Off-policy oracle certificate and tuning 44

E.7 Proofs of the full last-iterate guarantees 46

F Additional Experimental Details **49**

A Comprehensive Related Works

Table A: Comparison of online CMDP results with last-iterate guarantees for the final policy. We include only works that explicitly account for exploration through online interaction with an unknown CMDP. For a fair comparison with these prior tabular, model-based results, we report sample complexity in the tabular setting using one-hot state-action features for the actor and critic, with $\epsilon_{\text{act}} = 0$. Sample complexity counts the complete episodes required to obtain an $O(\epsilon)$-optimal, $O(\epsilon)$-feasible final policy. Rates suppress logarithmic factors and dependence on problem parameters other than ϵ. The *Structured CMDPs* column identifies guarantees based on features or function classes, while *Model free* indicates that the method does not explicitly estimate the transition kernel.

	Sample complexity	Model free	Structured CMDPs
Müller et al. (2024)	$\widetilde{O}(\epsilon^{-14})$	✗	✗
Kitamura et al. (2024)	$\widetilde{O}(\epsilon^{-7})$	✗	✗
Zuo et al. (2026)	$\widetilde{O}(\epsilon^{-6})$	✗	✗
This work (off-policy)	$\widetilde{O}(\epsilon^{-6})$	✓	✓
This work (on-policy)	$\widetilde{O}(\epsilon^{-10})$	✓	✓
Lower bound Domingues et al. (2021)	$\widetilde{O}(\epsilon^{-2})$	—	—

Importance of last-iterate convergence. Last-iterate guarantees are particularly valuable for a parameterised actor. For the log-linear class defined above, the pointwise average $K^{-1}\sum_{k=1}^{K}\pi_h(\cdot \mid s, \omega^k)$ generally cannot be represented as $\pi_h(\cdot \mid s, \omega)$ by one parameter set ω. Sampling a past actor at the start of each episode implements a mixture, but requires storing those actors; averaging their parameters produces a log-linear policy without inheriting the mixture's reward or feasibility guarantee. This distinction is relevant to constrained applications such as language-model fine-tuning that optimises helpfulness subject to a harmful-response cost (Dai et al., 2024) and wireless communication, where a deployed control policy must respect average transmission-power limits (Bura et al., 2022). In these applications, the single paramaterised policy deployed after training must satisfy the constraints, making a last-iterate guarantee particularly desirable.

Last-iterate convergence *without* online exploration A related line of work studies last-iterate convergence under specified gradient, rollout, or coverage conditions. Ding et al. (2023) establish nonasymptotic guarantees for regularised and optimistic policy-gradient primal-dual updates; their sample-based extension assumes access to policy rollouts from a simulator. Moskovitz et al. (2023) use optimistic ascent-descent to stabilise the primal-dual iterates. Their "optimism" is a correction to the saddle-point update, distinct from confidence bounds for an unknown CMDP. For general parameterised policies, Montenegro et al. (2024) analyse last-iterate convergence under gradient-domination and gradient-estimator assumptions, while Mondal and Aggarwal (2026) use compatible-approximation and Fisher non-degeneracy conditions together with sampled rollouts. Recent augmented-Lagrangian work by Lu et al. (2026) also establishes last-iterate guarantees under coverage and gradient-evaluation assumptions. These results do not jointly establish confidence-guided exploration from adaptively collected trajectories in an unknown CMDP and a final-policy guarantee that accounts for the resulting estimation error.

By contrast, our method explicitly addresses exploration in an unknown CMDP: it constructs optimistic Bellman estimates from the collected trajectories, whose high-probability certificates drive exploration and interface with the last-iterate primal-dual contraction under both fresh on-policy sampling and cumulative off-policy data. Our general-function result therefore broadens the critic and statistical-exploration axes under structural Bellman assumptions, fixed-coreset leverage control, and an explicit policy-space actor-projection error.

Last-iterate convergence *with* online exploration. The closest works to ours in Table A are Müller et al. (2024), Kitamura et al. (2024), and Zuo et al. (2026). All three address exploration in an unknown tabular CMDP using optimistic, model-based methods. Müller et al. (2024) are closest in their use of regularised primal-dual updates and a policy-dual contraction; their online tuning yields $\widetilde{O}(\epsilon^{-14})$ episodes for an $O(\epsilon)$-optimal, $O(\epsilon)$-feasible final policy. Kitamura et al. (2024) obtain a

$\widetilde{O}(\epsilon^{-7})$ uniform-PAC bound, which controls the number of episodes with inaccurate or infeasible policies. Zuo et al. (2026) use decaying safety margins to control online strong regret and constraint violation. Their online analysis yields $\widetilde{O}(\epsilon^{-6})$ episode scaling, matching the accuracy exponent of our off-policy bound, although their algorithmic approach and analysis differ from ours.

Our work retains the optimistic, regularised primal-dual structure while estimating Bellman quantities directly from trajectories, without explicitly constructing a transition model. High-probability critic certificates account for exploration error in the last-iterate analysis under both fresh on-policy sampling and cumulative off-policy data. This gives *model-free guarantees* for linear CMDPs and general function classes under the stated structural assumptions, with a compact log-linear actor. In the tabular specialisation with state-action features and $\epsilon_{\text{act}} = 0$, our episode bounds are $\widetilde{O}(\epsilon^{-6})$ off-policy and $\widetilde{O}(\epsilon^{-10})$ on-policy. The off-policy rate matches the state-of-the-art rate of Zuo et al. (2026).

Guarantee criteria for online CMDPs. Several notions of safety and convergence appear in the online CMDP literature and are not interchangeable. Standard weak regret and cumulative violation control signed sums (Efroni et al., 2020) and therefore allow cancellation: an unsafe episode may be offset by an overly conservative one. Strong regret and hard constraint violation instead sum positive-part errors, while zero cumulative violation and episode-wise safety impose still different requirements on the learning trajectory (Wei et al., 2022; Ghosh et al., 2024; Stradi et al., 2025; Kitamura et al., 2025). Last-iterate convergence concerns the reward and feasibility of the policy returned for deployment, rather than the complete path followed during learning (Ding et al., 2023; Müller et al., 2024; Montenegro et al., 2024; Lu et al., 2026; Zuo et al., 2026; Kitamura et al., 2024; Liu et al., 2025). A uniform tail guarantee for all sufficiently late iterates would imply a strong regret. However, sublinear strong regret provides a guarantee for the uniform mixture of policies, $\frac{1}{K}\sum_{k=1}^{K}\pi_k$, but does not directly guarantee the final policy π_K.

Primal-dual methods. CMDPs provide a standard framework for reinforcement learning with explicit safety constraints, with roots in stochastic control and broad use in safe RL and robotics (Altman, 1999; Achiam et al., 2017; Tessler et al., 2019; Brunke et al., 2022; Gu et al., 2024). Classical algorithmic approaches include LP-based, dual, and primal-dual methods (Efroni et al., 2020; Paternain et al., 2019; 2023). LP-based methods often provide strong theoretical guarantees (Efroni et al., 2020), but require planning over occupancy measures and therefore scale poorly with large state spaces. Dual methods optimise over the Lagrange multipliers, but evaluating each dual iterate typically requires solving the associated unconstrained MDP or policy-optimisation problem, leading to an inner planning or policy-optimisation loop (Paternain et al., 2019; Ying et al., 2022). Entropy-regularised dual formulations and cutting-plane methods can improve the optimisation dependence when sufficiently accurate planning or policy-optimisation subroutines are available (Ying et al., 2022; Gladin et al., 2023). Primal-dual methods instead update the policy and multipliers together in a single loop, making them especially compatible with policy-gradient and actor-critic implementations in large-scale RL (Ding et al., 2020; 2025; Liu et al., 2021; Bai et al., 2022; Qiu et al., 2020; Ghosh et al., 2022). This computational convenience, however, comes with a known limitation: vanilla primal-dual dynamics may oscillate, and most existing guarantees control averaged values, mixture policies, or weak regret rather than the final policy iterate (Ding et al., 2023). Our work builds on the efficiency of primal-dual methods while targeting a last-iterate guarantee for the final deployed policy.

Technical novelty compared with (Müller et al., 2024) The online tabular analysis of Müller et al. (2024) is the closest regularised optimistic predecessor. Both methods combine entropy and quadratic dual regularisation and are built around an exponentiated ideal policy update and a policy-dual Lyapunov argument. Müller et al. (2024) estimate a tabular transition model and perform optimistic dynamic programming, whereas our critic is obtained directly from trajectory data by Bellman regression. The master theorem isolates the initial-state evaluation error β_k, the local exponential moment $C_{\eta,\tau,\Lambda,k}$, and the actor error ϵ_{act}, thereby separating the last-iterate argument from a particular statistical estimator. This permits representation-dependent linear and general-function instantiations and a fixed-dimensional actor instead of an exact tabular policy. A further analytical difference is the fixed dual domain $\lambda_{\max} = 2H\log(e|\mathcal{A}|)/\xi$, which yields an $O(\tau)$ regularisation bias without making the dual radius accuracy-dependent. Holding the dimensions, horizon, constraint count, and Slater gap fixed, the exact-actor tabular specialisation has target-accuracy dependence $\widetilde{O}(\epsilon^{-10})$ on-policy and $\widetilde{O}(\epsilon^{-6})$ with off-policy data, compared with the $\widetilde{O}(\epsilon^{-14})$ terminal-accuracy

dependence implied by their online analysis. This comparison concerns the returned policy; their strong-regret theorem additionally controls cumulative positive-part errors. Appendix B.2 and the following comparison with the prior tabular analysis give the detailed rate accounting.

B Discussion of the Statistical Improvements

B.1 On actor assumption

Comparison with uniform raw-critic approximation. The comparison is with Assumption 4.1 of (Lin et al., 2025), which uniformly approximates the raw critic increment. For the same actor features and critic targets, choosing $b \equiv 0$ gives, for every (k, h),

$$\begin{aligned}\inf_{\omega,b} \sup_{s,a} \left|\langle \varphi(s,a), \omega - \omega_h^k \rangle - \eta\big(Q^k_{z_k,h}(s,a) - b(s)\big)\right| \\ \le \inf_{\omega} \sup_{s,a} \left|\langle \varphi(s,a), \omega - \omega_h^k \rangle - \eta Q^k_{z_k,h}(s,a)\right|.\end{aligned} \tag{22}$$

Here $\omega \in \mathbb{R}^{d_a}$ and $b : \mathcal{S} \to \mathbb{R}$. Their bias tolerance corresponds to $\eta\epsilon_{\text{bias}}$ in our normalization. Their printed assumption places the infimum before the supremum over iterations and stages, and therefore also implies the per-update raw bound on the right of (22). Consequently, their uniform raw-critic requirement implies (8). The example below shows that the converse fails. Thus, our approximation condition is strictly weaker than the uniform raw-critic requirement of Lin et al. (2025).

Why action differences suffice. Fix (k, h, ω) and write $e(s,a) = \langle \varphi(s,a), \omega - \omega_h^k \rangle - \eta Q^k_{z_k,h}(s,a)$. Since $\mathcal{A}$ is finite,

$$\begin{aligned}\inf_b \sup_{s,a} |e(s,a) + \eta b(s)| &= \frac{1}{2} \sup_{s,a,a'} |e(s,a) - e(s,a')| \\ &= \frac{1}{2} \sup_{s,a,a'} \left|\langle \Delta\varphi(s,a,a'), \omega - \omega_h^k \rangle - \eta \Delta Q^k_{z_k,h}(s,a,a')\right|.\end{aligned} \tag{23}$$

Indeed, $|e(s,a) - e(s,a')| \le 2\sup_a |e(s,a) + \eta b(s)|$ for every b, and equality in the first line is attained by $b(s) = -(\max_a e(s,a) + \min_a e(s,a))/(2\eta)$. Thus (8) is equivalent to a uniform action-difference approximation bound with tolerance $2\eta\epsilon_{\text{bias}}$. This identity explains both the baseline-free loss (9) and the factor 2 in the certificate $\epsilon_{\text{act}} = 2(\kappa_G + 1)\epsilon_{\text{bias}}$. The proof of Proposition 3.3 below establishes the coreset transfer and the KL bound.

A strict counterexample. Consider $H = 2$, $I = 1$, $\mathcal{A} = \{-, +\}$, and four states: an initial state s_{init} (so $s_1 = s_{\text{init}}$), two second-stage states s_{zero} and s_{sign}, and an absorbing state s_{abs}. Only at s_{sign} does the feature depend on the action. Specifically, let $x(s_{\text{sign}}, +) = 1$, $x(s_{\text{sign}}, -) = -1$, and $x(s,a) = 0$ elsewhere. Take $\phi(s,a) = \varphi(s,a) = (1, x(s,a))^\top/\sqrt{2}$, so $d_c = d_a = 2$. Set

$$\begin{aligned} P_1(\cdot \mid s,a) &= \tfrac{1}{2}\delta_{s_{\text{zero}}} + \tfrac{1}{2}\delta_{s_{\text{sign}}}, \qquad P_2(\cdot \mid s,a) = \delta_{s_{\text{abs}}}, \\ r_1 &= 0, \qquad r_2(s,a) = \tfrac{1+x(s,a)}{2}, \qquad u_{1,h} = 1, \qquad c = 0. \end{aligned}$$

Both second-stage states are reachable. This is a linear CMDP satisfying Assumption 4.1: take $v_{r,1} = 0$, $v_{r,2} = (1,1)^\top/\sqrt{2}$, $v_{u_1,h} = (\sqrt{2}, 0)^\top$, and $\mu_h = (\sqrt{2}P_h, 0)^\top$. All stated norm bounds hold, and Slater's condition holds with $\xi = 1$.

Fix $0 < \eta\tau \le 1$, initialize $\omega_1^1 = 0$ and $\omega_2^1 = (0, \sqrt{2}v_1)^\top$ with $v_1 > 0$, and use exact evaluation. At iteration k, write $\omega_2^k = (0, \sqrt{2}v_k)^\top$, so $\pi_2^k(a \mid s) \propto e^{v_k x(s,a)}$. The terminal shaped critic is

$$Q^k_{z_k,2}(s,a) = \underbrace{\frac{1}{2} + \lambda_k + \tau \log \sum_{a' \in \mathcal{A}} e^{v_k x(s,a')}}_{b_k(s)} + \left(\frac{1}{2} - \tau v_k\right) x(s,a). \tag{24}$$

Consequently, choosing the baseline b_k and $\omega = \omega_2^k + \eta\sqrt{2}(0, \tfrac{1}{2} - \tau v_k)^\top$ makes the baseline-adjusted error zero. At the first stage, the uniform policy is preserved and the shaped critic is constant, so its

error is also zero. In contrast,

$$\frac{\varphi(s_{\text{sign}},+)+\varphi(s_{\text{sign}},-)}{2}=\varphi(s_{\text{zero}},+),$$
$$\frac{Q^k_{z_k,2}(s_{\text{sign}},+)+Q^k_{z_k,2}(s_{\text{sign}},-)}{2}-Q^k_{z_k,2}(s_{\text{zero}},+)=\tau\log\cosh v_k.$$

The triangle inequality therefore gives

$$\inf_{\theta\in\mathbb{R}^2}\sup_{s,a}\left|\langle\varphi(s,a),\theta\rangle-Q^k_{z_k,2}(s,a)\right|\geq\frac{\tau}{2}\log\cosh v_k>0. \tag{25}$$

For the increment $\eta Q^k_{z_k,2}$, this lower bound is multiplied by η. Thus our approximation condition holds with $\epsilon_{\text{bias}}=0$, while the uniform raw-critic condition cannot hold with zero bias.

A single coreset triplet $(s_{\text{sign}},+,-)$ of weight one gives $G_{\text{exp}}=\operatorname{diag}(0,2)$ and $\kappa_G=1$. The fitted policy equals the ideal update, with slope

$$v_{k+1}=(1-\eta\tau)v_k+\eta/2,\qquad v_k\longrightarrow\frac{1}{2\tau}\quad\text{as } k\to\infty.$$

In particular, $v_k>0$ for every k. Moreover, $\frac{\tau}{2}\log\cosh(1/(2\tau))\longrightarrow 1/4$ as $\tau\downarrow 0$. Hence the actual projection error is zero even though the raw-critic bias certificate need not vanish. This does not assert that linear-CMDP realizability alone makes ϵ_{bias} small: nonlinear bonuses and clipping may still require richer actor features.

Coreset coverage. Only the span of $\{\Delta\varphi(s,a,a')\}$ affects the policy; its orthogonal complement consists of action-independent feature directions. The range condition in Assumption 3.2 ensures that coreset fitting controls every direction in this span, while $G^{\dagger}_{\text{exp}}$ allows irrelevant directions to remain singular. For a finite feature image of rank $m\geq 2$, applying Lattimore et al. (2020, Theorem 4.4) in this span gives $|\mathcal{D}_{\text{exp}}|=\widetilde{O}(m)$ and $\kappa_G^2\leq 2m$. For a bounded, possibly infinite image, the same asymptotic claim follows by a finite-net argument. In span coordinates, choose image vectors $y_1,\dots,y_m$ forming a basis and set $A=m^{-1}\sum_j y_j y_j^{\top}\succ 0$. Take a finite ζ-net containing this basis with $\zeta\leq\sqrt{\lambda_{\min}(A)}$, and let G be its design matrix. Then, for every image vector y and a net point y' within ζ,

$$\operatorname{tr}(G^{-1}A)\leq 2m,\qquad \|G^{-1}\|_{\text{op}}\leq\frac{2m}{\lambda_{\min}(A)},\qquad \|y\|_{G^{-1}}\leq\sqrt{2m}+\zeta\sqrt{\frac{2m}{\lambda_{\min}(A)}}\leq\sqrt{8m}.$$

For $m=1$, a singleton whose feature norm is at least half the supremum gives $\kappa_G\leq 2$; for $m=0$, any singleton gives $G_{\text{exp}}=0$ and $\kappa_G=0$. Since $m\leq d_{\text{a}}$ and $\|\Delta\varphi\|_2\leq 2$, this proves the stated support and leverage orders without a compactness assumption. It is an existence statement; the coreset is treated as given, not as obtainable from an arbitrary infinite feature image at no computational cost.

B.2 Origin of the sharper accuracy dependence

The improvement begins with the conversion from the regularised saddle point to the original CMDP. The conversion argument in Theorem 3.4 bounds the multiplier $\lambda^{\star}_{\tau}$ using Slater's condition, so the box radius $\lambda_{\max}=2H\log(e|\mathcal{A}|)/\xi$ need not grow as the target accuracy decreases. The regularisation bias is consequently $O(\tau)$, allowing $\tau=\Theta(\epsilon)$ when the other problem parameters are fixed. The radius-dependent conversion of Müller et al. (2024, Lemma 4.2) instead balances $1/\lambda_{\max}$ against $\tau\lambda_{\max}$, leading to $\lambda_{\max}=\Theta(\epsilon^{-1})$ and $\tau=\Theta(\epsilon^2)$. Remark C.1 gives the two conversion inequalities.

To isolate this effect, take $\epsilon_{\text{act}}=0$, hold all remaining problem parameters fixed, and use the same oracle certificates in both analyses. In the on-policy case, the master recursion requires

$$\eta=\widetilde{\Theta}\left(\frac{\tau\epsilon^2}{\lambda_{\max}^2}\right),\qquad K=\widetilde{O}\left(\frac{1}{\eta\tau}\right),\qquad N=\widetilde{O}\left(\frac{\lambda_{\max}^4}{\tau^2\epsilon^4}\right).$$

Here the stepsize controls the local-norm term, geometric contraction controls the initial error, and the batch size controls critic error. For off-policy sampling, the cumulative certificate gives

$K = \widetilde{O}(\lambda_{\max}^4/(\tau^2\epsilon^4))$ with $\eta = \widetilde{\Theta}(1/(\tau K))$. Substitution yields the following accuracy exponents; entries suppress universal constants and logarithmic factors.

Quantity	Radius-dependent conversion	Fixed-domain conversion
Dual radius $\lambda_{\max}$	ϵ^{-1}	1
Regularisation τ	ϵ^{2}	ϵ
On-policy iterations K	ϵ^{-8}	ϵ^{-4}
On-policy batch size N	ϵ^{-12}	ϵ^{-6}
On-policy episodes NK	ϵ^{-20}	ϵ^{-10}
Off-policy episodes K	ϵ^{-12}	ϵ^{-6}

These are comparisons within our master analysis, not a restatement of the published rates. Both columns retain geometric contraction, isolating the effect of the conversion. For our fixed-domain guarantees with actor approximation, $\epsilon_{\text{act}} \lesssim \xi\epsilon^3/H^5$ keeps the approximation floor in Theorem 4.3 at the target order.

Comparison with the prior tabular analysis. The online model-based result of Müller et al. (2024, Theorem 5.1) uses one episode per iteration and proves $\widetilde{O}(K^{13/14})$ strong regret. Their Remark 5.1 also identifies a last-iterate guarantee: the slowest terminal error term under that tuning is $\widetilde{O}(K^{-1/14})$, giving $\widetilde{O}(\epsilon^{-14})$ episodes. Their exact-value Theorem 4.1 counts optimisation iterations and is therefore a different comparison.

For a direct tabular specialisation of our theorem, take one-hot critic and actor features, so $d_{\text{c}} = d_{\text{a}} = |\mathcal{S}||\mathcal{A}|$ and $\epsilon_{\text{bias}} = \epsilon_{\text{act}} = 0$. The off-policy guarantee then gives

$$K = \widetilde{O}\left(\frac{(|\mathcal{S}||\mathcal{A}|)^3 I^4 H^{16}}{\xi^6\epsilon^6}\right).$$

The sharper accuracy dependence comes from the fixed-domain conversion, geometric contraction, and terminal-error calibration. It is an improvement in the analysis, not an inherent statistical advantage of model-free evaluation: a suitable model-based oracle could use the same master recursion. The guarantees also have different scopes. The cited online theorem controls cumulative strong regret during learning, whereas our theorem controls the returned policy and, in this specialisation, pays cubic dependence on $|\mathcal{S}||\mathcal{A}|$.

B.3 Actor- and action-space dependence

Remark B.1 (Statistical roles of the actor class and action space)**.** For the fresh-batch on-policy oracle, conditionally on the pre-iteration history, π^k is fixed and $\mathcal{D}^k$ contains N independent trajectories generated by this single policy. The uniform confidence argument therefore covers only the d_{c}-dimensional critic coefficient and the elliptical-bonus geometry, whose covering dimension is of order $d_{\text{c}}^2 + d_{\text{c}}$. No cover of the actor class is needed, which yields the $d_{\text{c}}^{3/2}/\sqrt{N}$ consistency term in (14). With off-policy data, by contrast, the fitted value function depends on policies generated from the accumulated data. Uniform control over the reachable actor class is then required and contributes the d_{a} term in $\sqrt{d_{\text{c}}(d_{\text{c}}^2 + d_{\text{a}})K}$ in (15). Thus, d_{a} is absent from the on-policy statistical rate but may enter the off-policy rate; in either setting it may also affect the actor approximation error and the computational cost of ActorFit. The general-function guarantees exhibit the same split: (20) has no d_{a} term, whereas (21) contains one in the off-policy setting.

In both sampling regimes, the remaining explicit action-space dependence arises from entropy and normalization terms through powers of $\log(e|\mathcal{A}|)$. In particular, retaining the escort moment $|\mathcal{A}|^{\eta\tau}$ and imposing the stated stepsize condition keep this factor bounded by a universal constant. Hence, for fixed representation dimensions, neither guarantee contains a polynomial factor in $|\mathcal{A}|$, whereas the online tabular guarantee of Müller et al. (2024, Theorem 5.1 and Appendix F) is polynomial in $|\mathcal{A}|$ (its leading $K^{13/14}$ coefficient includes $|\mathcal{A}|^{1/4}$). This comparison concerns the explicit dependence on action cardinality: with one-hot tabular features, $d_{\text{c}} = d_{\text{a}} = |\mathcal{S}||\mathcal{A}|$, so the representation dimensions themselves recover polynomial dependence on $|\mathcal{A}|$.

C Proofs for Section 3

We first transfer the coreset fitting error to the actor certificate, then derive the primal and dual inequalities that give the master recursion. Finally, we convert distance to the regularised saddle point into reward and constraint bounds. The constants in these three steps are retained because they specify the oracle interface and the fixed dual radius.

C.1 Actor projection and the master recursion

Proof of Proposition 3.3. The proof controls the fitted logit differences uniformly, then transfers that error through softmax. Fix (k, h) and $\nu > 0$. By Assumption 3.2, choose $\omega_h^{k,0} \in \mathbb{R}^{d_a}$ and $b : \mathcal{S} \to \mathbb{R}$ such that, with

$$e_0(s,a) = \langle \varphi(s,a), \omega_h^{k,0} - \omega_h^k \rangle - \eta\big(Q^k_{z_k,h}(s,a) - b(s)\big),$$

we have $\|e_0\|_\infty \le \eta\epsilon_{\text{bias}} + \nu$. The baseline cancels in $e_0(s,a) - e_0(s,a')$. The normal equation for the exact minimizer in (9) gives

$$G_{\exp}(\omega_h^{k+1} - \omega_h^{k,0}) = -\sum_{(s,a,a')\in\mathcal{D}_{\exp}} \rho_{\exp}(s,a,a')\Delta\varphi(s,a,a')\big(e_0(s,a) - e_0(s,a')\big). \tag{26}$$

Taking the inner product of (26) with $\omega_h^{k+1} - \omega_h^{k,0}$ and applying Cauchy-Schwarz gives

$$\|\omega_h^{k+1} - \omega_h^{k,0}\|_{G_{\exp}} \le \left(\sum_{(s,a,a')\in\mathcal{D}_{\exp}} \rho_{\exp}(s,a,a')\big(e_0(s,a) - e_0(s,a')\big)^2\right)^{1/2} \le 2(\eta\epsilon_{\text{bias}} + \nu).$$

Since every $\Delta\varphi(s,a,a')$ belongs to $\operatorname{range}(G_{\exp})$, the pseudoinverse Cauchy-Schwarz inequality and the leverage bound give

$$\begin{aligned}\left|\left\langle \Delta\varphi(s,a,a'), \omega_h^{k+1} - \omega_h^{k,0}\right\rangle\right| &\le \|\Delta\varphi(s,a,a')\|_{G_{\exp}^\dagger}\|\omega_h^{k+1} - \omega_h^{k,0}\|_{G_{\exp}} \\ &\le 2\kappa_G(\eta\epsilon_{\text{bias}} + \nu).\end{aligned} \tag{27}$$

Consequently, for

$$\Delta_{k,h}(s,a) = \langle \varphi(s,a), \omega_h^{k+1} - \omega_h^k \rangle - \eta Q^k_{z_k,h}(s,a),$$

the triangle inequality gives $\sup_{s,a,a'} |\Delta_{k,h}(s,a) - \Delta_{k,h}(s,a')| \le 2(\kappa_G + 1)(\eta\epsilon_{\text{bias}} + \nu)$. Letting $\nu \downarrow 0$ yields

$$\sup_{s,a,a'} |\Delta_{k,h}(s,a) - \Delta_{k,h}(s,a')| \le 2\eta(\kappa_G + 1)\epsilon_{\text{bias}} = \eta\epsilon_{\text{act}}. \tag{28}$$

No bound on the action-independent part of $\Delta_{k,h}$ is needed, since

$$\pi_h^{k+1}(a \mid s) = \frac{\widetilde{\pi}_h^{k+1}(a \mid s)e^{\Delta_{k,h}(s,a)}}{\mathbb{E}_{a'\sim\widetilde{\pi}_h^{k+1}(\cdot|s)}[e^{\Delta_{k,h}(s,a')}]}.$$

Therefore, for every state s and $p \in \Delta(\mathcal{A})$, suppressing s in the next display,

$$\begin{aligned}\mathrm{KL}(p\|\pi_h^{k+1}) - \mathrm{KL}(p\|\widetilde{\pi}_h^{k+1}) &= -\mathbb{E}_p[\Delta_{k,h}] + \log \mathbb{E}_{\widetilde{\pi}_h^{k+1}}[e^{\Delta_{k,h}}] \\ &\le \max_a \Delta_{k,h}(s,a) - \min_a \Delta_{k,h}(s,a) \le \eta\epsilon_{\text{act}}.\end{aligned}$$

Taking $p = \pi^\star_{\tau,h}(\cdot \mid s)$ proves (7) with $\epsilon_{\text{act}} = 2(\kappa_G + 1)\epsilon_{\text{bias}}$. □

Proof of the recursion bound in Theorem 3.4. Work on the joint event in Assumption 3.1 and fix $k \in [K-1]$. We first derive a one-step inequality.

Primal progress. The entropy identity $V^{\pi^\star_\tau}_{\psi_k} - V^{\pi^\star_\tau}_{\psi^{\pi^\star_\tau}} = \sum_h \mathbb{E}_{d_h^{\pi^\star_\tau}}[\mathrm{KL}_{k,h}]$ separates the regularisation term from the shaped reward z_k. Define the Bellman residual

$$\mathcal{E}_{k,h}(s,a) \triangleq Q^k_{z_k,h}(s,a) - (\mathbb{B}^{\pi^k}_{z_k,h}Q^k_{z_k,h+1})(s,a) \ge 0.$$

Bellman telescoping gives

$$V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{k} = \sum_{h=1}^{H} \mathbb{E}_{s\sim d_h^{\pi_\tau^\star}} \left[\left\langle \pi_{\tau,h}^\star(\cdot \mid s) - \pi_h^k(\cdot \mid s), Q_{z_k,h}^k(s,\cdot) \right\rangle \right] - \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d_h^{\pi_\tau^\star}} [\mathcal{E}_{k,h}(s,a)].$$

Dropping the nonnegative residual term and using $V_{z_k}^{k} - V_{z_k}^{\pi^k} \le \beta_k$ yields

$$L_\tau(\pi_\tau^\star, \lambda_k) - L_\tau(\pi^k, \lambda_k) = V_{z_k}^{\pi_\tau^\star} - V_{z_k}^{\pi^k} - \tau \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_\tau^\star}}[\mathrm{KL}_{k,h}] \le \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_\tau^\star}} \left[\langle \pi_{\tau,h}^\star - \pi_h^k, Q_{z_k,h}^k \rangle - \tau \mathrm{KL}_{k,h} \right] + \beta_k.$$

For the ideal exponential update, the log-normaliser identity gives

$$\mathrm{KL}_{k,h} - \mathrm{KL}(\pi_{\tau,h}^\star \| \widetilde{\pi}_h^{k+1}) = \eta \langle \pi_{\tau,h}^\star, Q_{z_k,h}^k \rangle - \log \mathbb{E}_{\pi_h^k}[e^{\eta Q_{z_k,h}^k}].$$

Since $z_k \ge 0$ and $Q_{z_k,H+1}^k = 0$, optimism implies $Q_{z_k,h}^k \ge 0$ by backward induction. Using $e^x \le 1 + x + x^2 e^x/2$ for $x \ge 0$, $\log(1+x) \le x$, and the local-norm bound (6), we obtain

$$\log \mathbb{E}_{\pi_h^k}[e^{\eta Q_{z_k,h}^k}] \le \eta \langle \pi_h^k, Q_{z_k,h}^k \rangle + \frac{\eta^2}{2} C_{\eta,\tau,\Lambda,k}.$$

Combining the preceding three displays gives

$$L_\tau(\pi_\tau^\star, \lambda_k) - L_\tau(\pi^k, \lambda_k) \le \frac{1}{\eta} \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_\tau^\star}} \left[\mathrm{KL}(\pi_{\tau,h}^\star \| \pi_h^k) - \mathrm{KL}(\pi_{\tau,h}^\star \| \widetilde{\pi}_h^{k+1}) \right] - \tau \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_\tau^\star}} \left[\mathrm{KL}(\pi_{\tau,h}^\star \| \pi_h^k) \right] + \frac{\eta H}{2} C_{\eta,\tau,\Lambda,k} + \beta_k. \tag{29}$$

The state argument of each KL divergence is implicit. By (7), the first line on the right-hand side of (29) is at most

$$\frac{1}{\eta} \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_\tau^\star}} \left[\mathrm{KL}(\pi_{\tau,h}^\star \| \pi_h^k) - \mathrm{KL}(\pi_{\tau,h}^\star \| \pi_h^{k+1}) \right] + H\epsilon_{\mathrm{act}}. \tag{30}$$

Dual progress. Set $g_k = V_u^k - c + \tau\lambda_k$. The dual update is $\lambda_{k+1} = \mathrm{Proj}_\Lambda(\lambda_k - \eta g_k)$. Because $V_u^k, c \in [0,H]^I$ and $\lambda_k \in \Lambda$, $\|g_k\|_2 \le G_\tau$. Nonexpansiveness of projection therefore gives

$$\langle V_u^k - c + \tau\lambda_k, \lambda_k - \lambda_\tau^\star \rangle \le \frac{\|\lambda_k - \lambda_\tau^\star\|_2^2 - \|\lambda_{k+1} - \lambda_\tau^\star\|_2^2}{2\eta} + \frac{\eta G_\tau^2}{2}.$$

Replacing V_u^k by $V_u^{\pi^k}$ costs at most $\|\lambda_k - \lambda_\tau^\star\|_1 \beta_k \le I\lambda_{\max}\beta_k$. Moreover,

$$L_\tau(\pi^k, \lambda_k) - L_\tau(\pi^k, \lambda_\tau^\star) + \frac{\tau}{2}\|\lambda_k - \lambda_\tau^\star\|_2^2 = \left\langle V_u^{\pi^k} - c + \tau\lambda_k, \lambda_k - \lambda_\tau^\star \right\rangle.$$

Combining this identity with the projection inequality yields

$$L_\tau(\pi^k, \lambda_k) - L_\tau(\pi^k, \lambda_\tau^\star) + \frac{\tau}{2}\|\lambda_k - \lambda_\tau^\star\|_2^2 \le \frac{\|\lambda_k - \lambda_\tau^\star\|_2^2 - \|\lambda_{k+1} - \lambda_\tau^\star\|_2^2}{2\eta} + \frac{\eta G_\tau^2}{2} + I\lambda_{\max}\beta_k. \tag{31}$$

Contraction. Substituting (30) into (29), adding (31), using the saddle inequality $L_\tau(\pi_\tau^\star, \lambda_k) \ge L_\tau(\pi_\tau^\star, \lambda_\tau^\star) \ge L_\tau(\pi^k, \lambda_\tau^\star)$, and multiplying by η gives

$$\Phi_{k+1} \le (1 - \eta\tau)\Phi_k + \frac{\eta^2}{2}(HC_{\eta,\tau,\Lambda,k} + G_\tau^2) + \eta(1 + I\lambda_{\max})\beta_k + \eta H\epsilon_{\text{act}}. \tag{32}$$

Iterating (32) gives

$$\Phi_K \le (1 - \eta\tau)^{K-1}\Phi_1 + \sum_{k=1}^{K-1}(1 - \eta\tau)^{K-1-k}\left[\frac{\eta^2}{2}(HC_{\eta,\tau,\Lambda,k} + G_\tau^2) + \eta(1 + I\lambda_{\max})\beta_k + \eta H\epsilon_{\text{act}}\right].$$

The inequality $(1 - \eta\tau)^{K-1} \le e^{-\eta\tau(K-1)}$ proves (10). □

C.2 Conversion from the regularised saddle point

Proof of the conversion bounds in Theorem 3.4. We first bound the bias at $\pi_\tau^\star$, then control the value change from $\pi_\tau^\star$ to π^k using the KL part of Φ_k.

The regularised multiplier $\lambda_{\tau,i}^\star$ is strictly smaller than $\lambda_{\max}$. Primal optimality, Slater's condition, and nonnegativity of reward and entropy give

$$L_\tau(\pi_\tau^\star, \lambda_\tau^\star) \ge L_\tau(\bar{\pi}, \lambda_\tau^\star) \ge \xi\|\lambda_\tau^\star\|_1. \tag{33}$$

Dual optimality allows comparison with the zero multiplier:

$$L_\tau(\pi_\tau^\star, \lambda_\tau^\star) \le L_\tau(\pi_\tau^\star, 0) \le H + \tau H \log|\mathcal{A}|. \tag{34}$$

Therefore

$$\|\lambda_\tau^\star\|_1 \le \frac{H(1 + \tau\log|\mathcal{A}|)}{\xi} \le \frac{H(1 + \log|\mathcal{A}|)}{\xi} < \lambda_{\max}. \tag{35}$$

Bias at the saddle point. By (35), $\lambda_{\tau,i}^\star < \lambda_{\max}$ (i.e., $\lambda_{\tau,i}^\star$ is *strictly* smaller than $\lambda_{\max}$) for every $i \in [I]$, so the upper box constraint $\lambda_i \le \lambda_{\max}$ does not bind. For fixed $\pi_\tau^\star$, the minimization over the i-th multiplier coordinate is

$$\lambda_{\tau,i}^\star \in \operatorname*{arg\,min}_{0 \le \lambda_i \le \lambda_{\max}} \left\{\lambda_i\big(V_{u_i}^{\pi_\tau^\star} - c_i\big) + \frac{\tau}{2}\lambda_i^2\right\}.$$

The solution is therefore

$$\lambda_{\tau,i}^\star = \min\left\{\lambda_{\max}, \left[\frac{c_i - V_{u_i}^{\pi_\tau^\star}}{\tau}\right]_+\right\} = \left[\frac{c_i - V_{u_i}^{\pi_\tau^\star}}{\tau}\right]_+, \tag{36}$$

where the final equality follows from $\lambda_{\tau,i}^\star < \lambda_{\max}$. Equivalently, if $\lambda_{\tau,i}^\star = 0$, lower-bound optimality gives $V_{u_i}^{\pi_\tau^\star} - c_i \ge 0$; otherwise, $0 < \lambda_{\tau,i}^\star < \lambda_{\max}$, and stationarity gives $V_{u_i}^{\pi_\tau^\star} - c_i = -\tau\lambda_{\tau,i}^\star$. Hence,

$$\big[c_i - V_{u_i}^{\pi_\tau^\star}\big]_+ = \tau\lambda_{\tau,i}^\star \le \frac{\tau H(1 + \tau\log|\mathcal{A}|)}{\xi}. \tag{37}$$

For reward, primal optimality at $\lambda_\tau^\star$ gives

$$V_r^{\pi^\star} \le V_r^{\pi_\tau^\star} + \lambda_\tau^{\star\top}(V_u^{\pi_\tau^\star} - c) + \tau\big(V_{\psi^{\pi_\tau^\star}}^{\pi_\tau^\star} - V_{\psi^{\pi^\star}}^{\pi^\star}\big) - \lambda_\tau^{\star\top}(V_u^{\pi^\star} - c).$$

The last term is nonpositive because $\pi^\star$ is feasible. The first multiplier term equals $-\tau\|\lambda_\tau^\star\|_2^2$ by (36), and policy entropy lies in $[0, H\log|\mathcal{A}|]$. Consequently,

$$V_r^{\pi^\star} - V_r^{\pi_\tau^\star} \le \tau H \log|\mathcal{A}|. \tag{38}$$

Transfer to the current iterate. For any per-stage function $f_h \in [0,1]$, the performance-difference identity uses the comparator occupancy already present in Φ_k:

$$V_f^{\pi_\tau^\star} - V_f^{\pi^k} = \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi_\tau^\star}}\left[\langle \pi_{\tau,h}^\star - \pi_h^k, Q_{f,h}^{\pi^k}\rangle\right].$$

Since $0 \le Q_{f,h}^{\pi^k} \le H$, Pinsker's inequality bounds the absolute value of each integrand by $H\sqrt{2\mathrm{KL}_{k,h}}$. Cauchy-Schwarz over the stages and their state distributions gives

$$|V_f^{\pi^k} - V_f^{\pi_\tau^\star}| \le H^{3/2}\left(2\sum_{h=1}^{H}\mathbb{E}_{d_h^{\pi_\tau^\star}}[\mathrm{KL}(\pi_{\tau,h}^\star\|\pi_h^k)]\right)^{1/2} \le H^{3/2}\sqrt{2\Phi_k}. \tag{39}$$

Applying (39) to r and each u_i, then using (37) and (38), gives the two inequalities for every k. Setting $k = K$ proves Theorem 3.4. Since both right-hand sides are nonnegative, the same bounds hold for the positive parts used below. □

Remark C.1 (Comparison with the radius-dependent conversion)**.** The prior analysis does not exclude the possibility that a regularised multiplier reaches the boundary of Λ. In the extreme case, $\lambda_\tau^\star = \lambda_{\max} e_i$, so only its i-th coordinate reaches $\lambda_{\max}$. The clipped optimality condition then implies only $\left[c_i - V_{u_i}^{\pi_\tau^\star}\right]_+ \ge \tau\lambda_{\max}$, and therefore provides no upper bound on the constraint violation.

To cover this case, Müller et al. (2024) compare the regularised saddle point with the feasible multiplier $\lambda = \lambda_{\max} e_i$. After dropping the nonnegative regularisation terms at $\pi^\star$, the regularised saddle inequality gives

$$V_r^{\pi^\star} + \lambda_\tau^{\star\top}(V_u^{\pi^\star} - c) \le V_r^{\pi_\tau^\star} + \lambda_{\max}\left(V_{u_i}^{\pi_\tau^\star} - c_i\right) + \frac{\tau}{2}\lambda_{\max}^2 + \tau V_{\psi^{\pi_\tau^\star}}^{\pi_\tau^\star}.$$

Let $\lambda^\star$ be an optimal multiplier for the unregularised CMDP. Optimality of $(\pi^\star, \lambda^\star)$ gives

$$V_r^{\pi_\tau^\star} - V_r^{\pi^\star} \le \lambda^{\star\top}\left(V_u^{\pi^\star} - V_u^{\pi_\tau^\star}\right).$$

Adding these two inequalities eliminates the reward terms. Moreover, feasibility of $\pi^\star$ and nonnegativity of $\lambda_\tau^\star$ imply $\lambda_\tau^{\star\top}(V_u^{\pi^\star} - c) \ge 0$. It follows that

$$\lambda_{\max}\left(c_i - V_{u_i}^{\pi_\tau^\star}\right) \le \frac{\tau}{2}\lambda_{\max}^2 + \lambda^{\star\top}\left(V_u^{\pi^\star} - V_u^{\pi_\tau^\star}\right) + \tau V_{\psi^{\pi_\tau^\star}}^{\pi_\tau^\star}.$$

Since $\|V_u^{\pi^\star} - V_u^{\pi_\tau^\star}\|_\infty \le H$ and $\|\lambda^\star\|_1 \le H/\xi$, the inner product is at most H^2/ξ. Also, $V_{\psi^{\pi_\tau^\star}}^{\pi_\tau^\star} \le H\log|\mathcal{A}|$. The resulting right-hand side is nonnegative; hence, taking positive parts and dividing by $\lambda_{\max}$ gives

$$\left[c_i - V_{u_i}^{\pi_\tau^\star}\right]_+ \le \frac{\tau}{2}\lambda_{\max} + \frac{1}{\lambda_{\max}}\left(\frac{H^2}{\xi} + \tau H\log|\mathcal{A}|\right).$$

Adding the policy-transfer error and relaxing $\tau\lambda_{\max}/2 \le \tau\lambda_{\max}$ gives

$$\left[c_i - V_{u_i}^{\pi^k}\right]_+ \le H^{3/2}\sqrt{2\Phi_k} + \tau\lambda_{\max} + \frac{1}{\lambda_{\max}}\left(\frac{H^2}{\xi} + \tau H\log|\mathcal{A}|\right). \tag{40}$$

In contrast, our analysis rules out the saturated case by proving $\lambda_{\tau,i}^\star < \lambda_{\max}$. The coordinatewise optimality condition (36) therefore gives directly

$$\left[c_i - V_{u_i}^{\pi_\tau^\star}\right]_+ = \tau\lambda_{\tau,i}^\star \le \frac{\tau H(1 + \tau\log|\mathcal{A}|)}{\xi}.$$

Thus, our conversion avoids both the $\tau\lambda_{\max}$ and $1/\lambda_{\max}$ terms. Consequently, $\lambda_{\max}$ remains independent of ϵ, and $\tau = \Theta(\epsilon)$ suffices, whereas the radius-dependent conversion of Müller et al. (2024) requires $\lambda_{\max} = \Theta(\epsilon^{-1})$ and $\tau = \Theta(\epsilon^2)$. Combined with the geometric contraction in Theorem 3.4, this yields the sharper accuracy dependence detailed in Appendix B.2.

Conventions for the complexity proofs. The remaining proofs use the global notation of the main paper. Universal numerical constants are denoted by C or c only locally; their values may change between inequalities. The notation $\widetilde{O}$ suppresses logarithmic factors in the horizon, representation complexities, action and constraint counts, inverse confidence and accuracy levels, and the parameter radii appearing in covering arguments. Polynomial dependence on the displayed problem parameters is retained.

For the simplified rates, choose a Slater witness with $0 < \xi \le 1$. A larger certified margin can always be replaced by $\min\{\xi, 1\}$; the resulting complexity is evaluated at that chosen margin. The accuracy-dependent comparison is informative when $\epsilon_{\text{act}} \le \xi/H^2$: otherwise the displayed approximation floor $H^{5/3}\xi^{-1/3}\epsilon_{\text{act}}^{1/3}$ is at least H, which already bounds every reward gap and constraint violation. The sample bounds retain the outer maximum induced by the cap $\tau \le 1$; this maximum can be removed when the cap is inactive.

D Proofs for Section 4

We first verify the three oracle conditions, then substitute their bounds into the master recursion. Fresh batches require uniform confidence only over the critic class; off-policy data also require uniformity over the actor class. The local-norm calculation is common to both modes and is proved once below. Throughout this section, $C, c > 0$ denote universal constants.

D.1 Technical lemmas for uniform confidence

For any bounded V, define

$$\theta_{V,h} \triangleq \int_{\mathcal{S}} V(s')\,\mu_h(ds'). \tag{41}$$

Assumption 4.1 gives

$$P_hV(s,a) = \langle \phi(s,a), \theta_{V,h}\rangle, \qquad \|\theta_{V,h}\|_2 \le \sqrt{d_{\text{c}}}\|V\|_\infty. \tag{42}$$

For $j \in \{r, u_1, \dots, u_I, \psi\}$, set

$$\theta_{j,h}(V) \triangleq \begin{cases} \upsilon_{r,h} + \theta_{V,h}, & j = r, \\ \upsilon_{u_i,h} + \theta_{V,h}, & j = u_i,\ i \in [I], \\ \theta_{V,h}, & j = \psi, \end{cases} \qquad R_{j,h} \triangleq \begin{cases} H + 1 - h, & j \in \{r, u_1, \dots, u_I\}, \\ (H-h)\log|\mathcal{A}|, & j = \psi. \end{cases} \tag{43}$$

For the backward recursion, write $\theta_{j,h}^k \triangleq \theta_{j,h}(V_{j,h+1}^k)$. Given a policy π, define $\ell_{j,h}^\pi(s,a) = 0$ for $j \in \{r, u_1, \dots, u_I\}$ and $\ell_{\psi,h}^\pi(s,a) = -\log \pi_h(a \mid s)$. For $B > 0$, let

$$v_{j,h}^{\pi,w,M}(s) \triangleq \mathbb{E}_{a\sim\pi_h(\cdot|s)}\Big[\ell_{j,h}^\pi(s,a) + \text{Truncate}_{[0,R_{j,h}]}\left(\langle\phi(s,a), w\rangle + \alpha_j\sqrt{\phi(s,a)^\top M\phi(s,a)}\right)\Big], \tag{44}$$

$$\mathcal{V}_{j,h}(\pi, B) \triangleq \left\{ v_{j,h}^{\pi,w,M} \;\middle|\; \|w\|_2 \le B,\ 0 \preceq M \preceq I_{d_{\text{c}}} \right\}.$$

Let $\mathcal{V}_h(\pi,B) \triangleq \bigcup_{j\in\{r,u_1,\dots,u_I,\psi\}} \mathcal{V}_{j,h}(\pi,B)$, and at the terminal stage set $\mathcal{V}_{j,H+1}(\pi,B) = \mathcal{V}_{H+1}(\pi,B) = \{0\}$.

Lemma D.1 (Covering the linear OPE value classes). *Let $\bar\alpha = \max\{\alpha_r, \alpha_u, \alpha_\psi\}$. For every fixed full-support policy π, $B > 0$, and $\varepsilon \in (0,1]$,*

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{V}_h(\pi,B)) \le C(d_{\text{c}}^2 + d_{\text{c}})\log\left(\frac{C(1+B+\bar\alpha+d_{\text{c}}+I)}{\varepsilon}\right). \tag{45}$$

If a class $\mathcal{F}_h$ induces policies $\pi(\omega)$ and corresponding class elements satisfy $\|v_{j,h}^{\pi(\omega),w,M} - v_{j,h}^{\pi(\omega'),w,M}\|_\infty \le L_h\|f_{\omega,h} - f_{\omega',h}\|_\infty$, then an $\varepsilon/(2L_h)$-net of $\mathcal{F}_h$, together with $\varepsilon/2$-nets of the corresponding fixed-policy classes, is an ε-net of their union.

Proof. Truncation and policy averaging are nonexpansive. If $\|w - w'\|_2 \le \varepsilon/2$, then $|\langle \phi, w - w'\rangle| \le \varepsilon/2$. Moreover,

$$\left|\sqrt{\phi^\top M\phi} - \sqrt{\phi^\top M'\phi}\right| \le \sqrt{\|M - M'\|_{\mathrm{op}}} \le \sqrt{\|M - M'\|_{\mathrm{F}}}.$$

Hence an $\varepsilon/2$-cover of the radius-B Euclidean ball and an $\varepsilon^2/(4\bar{\alpha}^2)$-cover of the positive-semidefinite contractions in Frobenius norm give an ε-cover of each component class. These parameter sets have dimensions d_{c} and $d_{\mathrm{c}}(d_{\mathrm{c}}+1)/2$ and radii B and $\sqrt{d_{\mathrm{c}}}$, respectively. Their volumetric covering bounds, followed by a union over the $I+2$ components, give

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{V}_h(\pi, B)) \le d_{\mathrm{c}} \log\left(1 + \frac{4B}{\varepsilon}\right) + \frac{d_{\mathrm{c}}(d_{\mathrm{c}}+1)}{2} \log\left(1 + \frac{8\sqrt{d_{\mathrm{c}}}\,\bar{\alpha}^2}{\varepsilon^2}\right) + \log(I+2),$$

which implies (45).

For the second claim, take an $\varepsilon/(2L_h)$-net of $\mathcal{F}_h$. At every actor in this net, take an $\varepsilon/2$-net of the corresponding fixed-policy value class. The assumed transfer inequality and the triangle inequality show that the resulting product net has radius ε, which proves the second assertion. □

For an adapted stage-h sample prefix and any value function V, write

$$\Sigma_{h,m} = I_{d_{\mathrm{c}}} + \sum_{\ell=1}^{m} \phi_h^\ell (\phi_h^\ell)^\top,$$
$$\widehat{\theta}_{j,h,m}(V) = \Sigma_{h,m}^{-1} \sum_{\ell=1}^{m} \phi_h^\ell \big(y_{j,h}^\ell + V(s_{h+1}^\ell)\big),$$
$$\zeta_{j,h}^\ell(V) = y_{j,h}^\ell + V(s_{h+1}^\ell) - \langle \phi_h^\ell, \theta_{j,h}(V)\rangle.$$

Here $\mathcal{G}_{\ell,h}$ denotes the information available through (s_h^ℓ, a_h^ℓ), before observing $(y_{j,h}^\ell, s_{h+1}^\ell)$.

Lemma D.2 (Uniform confidence over a value function class)**.** *Fix (j,h) and a class $\mathcal{W}_{h+1}$ that is deterministic, or fixed before a fresh batch is collected. Suppose that, uniformly over $V \in \mathcal{W}_{h+1}$ and ℓ, $\mathbb{E}[\zeta_{j,h}^\ell(V) \mid \mathcal{G}_{\ell,h}] = 0$, $|y_{j,h}^\ell + V(s_{h+1}^\ell)| \le Y_j$, and $\|\theta_{j,h}(V)\|_2 \le S_j$. If $\log \mathcal{N}_\infty(\varepsilon_0, \mathcal{W}_{h+1}) \le \Gamma$ and $(\sqrt{m_{\max}} + \sqrt{d_{\mathrm{c}}})\varepsilon_0 \le 1$, then, with probability at least $1 - \delta_0$, uniformly over $m \le m_{\max}$, $V \in \mathcal{W}_{h+1}$, and (s,a),*

$$\left|\langle \phi(s,a), \widehat{\theta}_{j,h,m}(V) - \theta_{j,h}(V)\rangle\right| \le \alpha_j \|\phi(s,a)\|_{\Sigma_{h,m}^{-1}}, \tag{46}$$

for $\alpha_j = C\{Y_j\sqrt{d_{\mathrm{c}}\log(1 + m_{\max}/d_{\mathrm{c}}) + \Gamma + \log(m_{\max}/\delta_0)} + S_j + 1\}$.

Proof. Let $\mathcal{C}$ be an ε_0-net of $\mathcal{W}_{h+1}$. For every $V \in \mathcal{C}$, the ridge normal equations give

$$\widehat{\theta}_{j,h,m}(V) - \theta_{j,h}(V) = \Sigma_{h,m}^{-1}\left(\sum_{\ell=1}^{m} \phi_h^\ell \zeta_{j,h}^\ell(V) - \theta_{j,h}(V)\right). \tag{47}$$

Self-normalized concentration and a union bound over $\mathcal{C}$ and $m \le m_{\max}$ yield

$$\left\|\sum_{\ell=1}^{m} \phi_h^\ell \zeta_{j,h}^\ell(V)\right\|_{\Sigma_{h,m}^{-1}} \le CY_j\sqrt{d_{\mathrm{c}}\log(1 + m_{\max}/d_{\mathrm{c}}) + \Gamma + \log(m_{\max}/\delta_0)}$$

simultaneously for all net elements and prefixes. The ridge term is at most $\|\theta_{j,h}(V)\|_2$. For any V' with $\|V - V'\|_\infty \le \varepsilon_0$,

$$\left\|\sum_{\ell=1}^{m} \phi_h^\ell (V - V')(s_{h+1}^\ell)\right\|_{\Sigma_{h,m}^{-1}} \le \sqrt{m}\,\varepsilon_0, \qquad \|\theta_{V,h} - \theta_{V',h}\|_2 \le \sqrt{d_c}\,\varepsilon_0.$$

These terms are absorbed by the displayed choice of α_j. Combining the preceding bounds with (47) and Cauchy-Schwarz proves (46). If $\mathcal{W}_{h+1}$ is conditionally fixed, the argument is applied after conditioning on the preceding history. □

D.2 Analysis of the on-policy oracle

Proof of Lemma 4.2, on-policy branch. Condition on the history $\mathcal{F}_{k-1}$ before iteration k. Then π^k and λ_k are fixed, and the N trajectories collected at iteration k are conditionally independent. Define

$$\iota_{\mathrm{on}} \triangleq \log\left(\frac{CKNH(I+1)|\mathcal{A}|d_c(1+\lambda_{\max})}{\delta \min\{1,\tau\}}\right). \tag{48}$$

Choose, for a sufficiently large universal constant C,

$$\alpha_r = \alpha_u = CH\sqrt{(d_c^2 + d_c)\iota_{\mathrm{on}}}, \qquad \alpha_\psi = CH\log(e|\mathcal{A}|)\sqrt{(d_c^2 + d_c)\iota_{\mathrm{on}}}. \tag{49}$$

We prove the oracle properties on a common event of probability at least $1 - \delta$.

Value class and uniform confidence. The reward and utility regression targets are bounded by H, while the entropy value target is bounded by $H\log(e|\mathcal{A}|)$. Since $\|\phi\|_2 \le 1$ and $0 \preceq (\Sigma_h^k)^{-1} \preceq I_{d_c}$, the normal equations give

$$\|\widehat{\theta}_{j,h}^k\|_2 \le \sum_{n=1}^{N} \|\phi_h^{k,n}\|_2 \left|y_{j,h}^{k,n} + V_{j,h+1}^k(s_{h+1}^{k,n})\right| \le NH\log(e|\mathcal{A}|) \triangleq B_N. \tag{50}$$

Define

$$\mathcal{V}_{k,j,h}^{\mathrm{on}} \triangleq \mathcal{V}_{j,h}(\pi^k, B_N), \qquad \mathcal{V}_{k,h}^{\mathrm{on}} \triangleq \mathcal{V}_h(\pi^k, B_N). \tag{51}$$

The backward recursion satisfies

$$V_{j,h}^k = v_{j,h}^{\pi^k, \widehat{\theta}_{j,h}^k, (\Sigma_h^k)^{-1}} \in \mathcal{V}_{k,j,h}^{\mathrm{on}} \subseteq \mathcal{V}_{k,h}^{\mathrm{on}}. \tag{52}$$

Conditionally on $\mathcal{F}_{k-1}$, this class is fixed. Applying Lemma D.1 with $B = B_N$ gives

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{V}_{k,h}^{\mathrm{on}}) \le C(d_c^2 + d_c)\left[\iota_{\mathrm{on}} + \log(1 + 1/\varepsilon)\right]. \tag{53}$$

Use the bounds $Y_r = Y_u = H$, $Y_\psi = H\log(e|\mathcal{A}|)$, $S_r, S_u \le 2H\sqrt{d_c}$, and $S_\psi \le H\sqrt{d_c}\log(e|\mathcal{A}|)$. Applying Lemma D.2 conditionally with $\mathcal{W}_{h+1} = \mathcal{V}_{k,j,h+1}^{\mathrm{on}}$, $m_{\max} = N$, $\varepsilon_0 = (2KNHd_c)^{-2}$, and $\delta_0 = \delta/[2KH(I+2)]$, followed by a union bound over (k,h,j), gives, with probability at least $1 - \delta/2$,

$$\left|\left\langle \phi(s,a), \widehat{\theta}_{j,h}^k - \theta_{j,h}^k \right\rangle\right| \le \alpha_j \|\phi(s,a)\|_{(\Sigma_h^k)^{-1}} = b_{j,h}^k(s,a) \tag{54}$$

simultaneously for all (k,h,j,s,a). The resulting radii are precisely (49). No actor cover is required because π^k is fixed before the fresh batch is collected.

Optimism and consistency. The population targets defined by (43) lie in their corresponding truncation intervals. Therefore, (54) and (13) imply

$$Q_{r,h}^k \ge r_h + P_h V_{r,h+1}^k, \quad Q_{u_i,h}^k \ge u_{i,h} + P_h V_{u_i,h+1}^k, \quad Q_{\psi,h}^k \ge \psi_{k,h} + P_h V_{\psi,h+1}^k. \tag{55}$$

Since $\lambda_k \ge 0$ and $\tau > 0$, summing the component inequalities gives

$$Q_{z_k,h}^k \ge \mathbb{B}_{z_k,h}^{\pi^k} Q_{z_k,h+1}^k, \qquad h \in [H]. \tag{56}$$

For $j = r, u_i, \psi$, let $(Q_{j,h}^{\pi^k}, V_{j,h}^{\pi^k}, V_j^{\pi^k})$ denote the corresponding action value, stagewise value, and total value; for $j = \psi$, this means $(Q_{\psi_k,h}^{\pi^k}, V_{\psi_k,h}^{\pi^k}, V_{\psi_k}^{\pi^k})$. Also set $V_j^k \triangleq V_{j,1}^k(s_1)$. Backward induction

in (55) gives $Q^k_{j,h} \geq Q^{\pi^k}_{j,h}$. The upper side of the confidence interval gives

$$0 \leq Q^k_{j,h}(s,a) - Q^{\pi^k}_{j,h}(s,a) \leq P_h\big(V^k_{j,h+1} - V^{\pi^k}_{j,h+1}\big)(s,a) + 2b^k_{j,h}(s,a). \tag{57}$$

The known immediate entropy term cancels from the difference when $j = \psi$. Taking the π^k-expectation of (57) and telescoping over h gives

$$0 \leq V^k_j - V^{\pi^k}_j \leq 2\sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\pi^k}_h}[b^k_{j,h}(s,a)]. \tag{58}$$

To bound the expected bonuses, define the population feature covariance

$$\Gamma_{k,h} \triangleq \mathbb{E}_{(s,a)\sim d^{\pi^k}_h}[\phi(s,a)\phi(s,a)^\top]. \tag{59}$$

and write $\phi^{k,n}_h = \phi(s^{k,n}_h, a^{k,n}_h)$. We now derive the required empirical covariance bound. Set

$$a \triangleq \frac{C\iota_{\mathrm{on}}}{N}, \qquad A_{k,h} \triangleq \Gamma_{k,h} + aI_{d_{\mathrm{c}}},$$

and define the normalized population and empirical covariances

$$G_{k,h} \triangleq A^{-1/2}_{k,h}\Gamma_{k,h}A^{-1/2}_{k,h}, \qquad \widehat{G}_{k,h} \triangleq \frac{1}{N}\sum_{n=1}^{N} A^{-1/2}_{k,h}\phi^{k,n}_h(\phi^{k,n}_h)^\top A^{-1/2}_{k,h}.$$

Because $A_{k,h} \succeq aI_{d_{\mathrm{c}}}$ and $\|\phi^{k,n}_h\|_2 \leq 1$, matrix Bernstein gives

$$\mathbb{P}\left(\|\widehat{G}_{k,h} - G_{k,h}\|_{\mathrm{op}} > \frac{1}{2}\,\middle|\,\mathcal{F}_{k-1}\right) \leq 2d_{\mathrm{c}}e^{-cNa} \leq \frac{\delta}{2KH}, \tag{60}$$

where the final inequality follows by choosing the universal constant in a sufficiently large. On the complementary event, $G_{k,h} \preceq \widehat{G}_{k,h} + \frac{1}{2}I_{d_{\mathrm{c}}}$. Multiplying this inequality on both sides by $A^{1/2}_{k,h}$ gives

$$\Gamma_{k,h} \preceq \frac{1}{N}\sum_{n=1}^{N}\phi^{k,n}_h(\phi^{k,n}_h)^\top + \frac{1}{2}\big(\Gamma_{k,h} + aI_{d_{\mathrm{c}}}\big).$$

Moving $\Gamma_{k,h}/2$ to the left and substituting $Na = C\iota_{\mathrm{on}}$ yields

$$N\Gamma_{k,h} \preceq 2\sum_{n=1}^{N}\phi^{k,n}_h(\phi^{k,n}_h)^\top + C\iota_{\mathrm{on}}I_{d_{\mathrm{c}}}. \tag{61}$$

A union bound over (k,h) and the tower property show that (61) holds jointly with (54) with probability at least $1-\delta$. Since $\Sigma^k_h = I_{d_{\mathrm{c}}} + \sum_{n=1}^{N}\phi^{k,n}_h(\phi^{k,n}_h)^\top$ and $0 \prec (\Sigma^k_h)^{-1} \preceq I_{d_{\mathrm{c}}}$, multiplying (61) by $(\Sigma^k_h)^{-1}$ and taking the trace gives

$$\begin{aligned}\mathrm{tr}((\Sigma^k_h)^{-1}\Gamma_{k,h}) &\leq \frac{2}{N}\,\mathrm{tr}\left((\Sigma^k_h)^{-1}(\Sigma^k_h - I_{d_{\mathrm{c}}})\right) + \frac{C\iota_{\mathrm{on}}}{N}\,\mathrm{tr}((\Sigma^k_h)^{-1}) \\ &\leq C\frac{d_{\mathrm{c}}\iota_{\mathrm{on}}}{N}.\end{aligned} \tag{62}$$

Moreover,

$$\mathbb{E}_{d^{\pi^k}_h}[\|\phi(s,a)\|^2_{(\Sigma^k_h)^{-1}}] = \mathrm{tr}((\Sigma^k_h)^{-1}\Gamma_{k,h}).$$

Jensen's inequality and (62) therefore yield

$$\begin{aligned}\mathbb{E}_{d^{\pi^k}_h}[\|\phi(s,a)\|_{(\Sigma^k_h)^{-1}}] &\leq \sqrt{\mathrm{tr}((\Sigma^k_h)^{-1}\Gamma_{k,h})} \\ &\leq C\sqrt{\frac{d_{\mathrm{c}}\iota_{\mathrm{on}}}{N}}.\end{aligned} \tag{63}$$

By (58),

$$0 \le V_{z_k}^k - V_{z_k}^{\pi^k} \le 2\big(\alpha_r + I\lambda_{\max}\alpha_u + \tau\alpha_\psi\big) \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^k}}[\|\phi(s,a)\|_{(\Sigma_h^k)^{-1}}],$$

$$\max_{i\in[I]} |V_{u_i}^k - V_{u_i}^{\pi^k}| \le 2\alpha_u \sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^k}}[\|\phi(s,a)\|_{(\Sigma_h^k)^{-1}}].$$

Substituting (49) and (63), the common choice

$$\beta_k \le C\big(1 + I\lambda_{\max} + \tau\log(e|\mathcal{A}|)\big)H^2\sqrt{\frac{d_c^3}{N}}\,\iota_{\mathrm{on}}. \tag{64}$$

controls both quantities required by Assumption 3.1(ii). The utility truncation also gives $0 \le V_{u_i}^k \le H$.

Local norm. The truncation levels give, for $p_a = \pi_h^k(a \mid s)$,

$$0 \le Q_{z_k,h}^k(s,a) \le H(1 + I\lambda_{\max} + \tau\log|\mathcal{A}|) + \tau\log(1/p_a). \tag{65}$$

For this calculation only, set $B = H(1 + I\lambda_{\max} + \tau\log|\mathcal{A}|)$ and $\alpha = \eta\tau$. Equation (65) gives $e^{\eta Q_{z_k,h}^k(s,a)} \le e^{\eta B} p_a^{-\alpha}$ and $(Q_{z_k,h}^k(s,a))^2 \le 2B^2 + 2\tau^2\log^2(1/p_a)$. Hence

$$\begin{aligned}\sum_a p_a e^{\eta Q_{z_k,h}^k(s,a)}(Q_{z_k,h}^k(s,a))^2 &\le 2e^{\eta B}\left[B^2\sum_a p_a^{1-\alpha} + \tau^2\sum_a p_a^{1-\alpha}\log^2(1/p_a)\right] \\ &\le C\left[B^2 + \tau^2\log^2(e|\mathcal{A}|)\right]. \end{aligned}\tag{66}$$

The first sum satisfies $\sum_a p_a^{1-\alpha} \le |\mathcal{A}|^\alpha$ by concavity. If $|\mathcal{A}| = 1$, the second sum is zero. Otherwise, set $t = 1/\log(e|\mathcal{A}|)$. The inequality $x^2 \le 4e^{-2}t^{-2}e^{tx}$ for $x \ge 0$ gives

$$\sum_a p_a^{1-\alpha}\log^2(1/p_a) \le C\log^2(e|\mathcal{A}|)\sum_a p_a^{1-\alpha-t} \le C\log^2(e|\mathcal{A}|)|\mathcal{A}|^{\alpha+t}. \tag{67}$$

The stepsize conditions imply $\eta B \le 1/4$ and $\alpha\log|\mathcal{A}| \le 1/4$, while $|\mathcal{A}|^t \le e$. Substituting these three bounds proves the second line of (66). Therefore,

$$C_{\eta,\tau,\Lambda,k} \le CH^2\big(1 + I\lambda_{\max} + \tau\log(e|\mathcal{A}|)\big)^2. \tag{68}$$

Since $0 < \tau \le 1$,

$$1 + I\lambda_{\max} + \tau\log(e|\mathcal{A}|) \le 2\big(1 + I\lambda_{\max} + \tau\log|\mathcal{A}|\big),$$

so (68) has the form stated in the lemma. Equations (56), (64), and (68) prove Assumption 3.1; specifically, they establish the on-policy consistency bound in (14) and the common local-norm bound in Lemma 4.2. □

D.3 Proof of the on-policy final rate

Fix $\epsilon, \delta \in (0,1)$ and choose the parameters at the orders

$$\begin{aligned}\lambda_{\max} &= \frac{2H\log(e|\mathcal{A}|)}{\xi}, \quad \tau = \Theta\left(\min\left\{1, \max\left\{\frac{\xi\epsilon}{H\log(e|\mathcal{A}|)}, \frac{H^{2/3}\xi^{2/3}\epsilon_{\mathrm{act}}^{1/3}}{\log^{2/3}(e|\mathcal{A}|)}\right\}\right\}\right), \\ \eta &= \widetilde{\Theta}\left(\frac{\tau\xi^2\epsilon^2}{I^2H^8}\right), \qquad K = \widetilde{\Theta}\left(\frac{I^2H^8}{\tau^2\xi^2\epsilon^2}\right), \qquad N = \widetilde{\Theta}\left(\frac{d_c^3I^4H^{14}}{\tau^2\xi^4\epsilon^4}\right).\end{aligned}\tag{69}$$

Proof of Theorem 4.3, on-policy branch. We first choose the iteration and batch budgets to control the potential, and then tune τ to convert that control into a guarantee for the original CMDP. We use the normalized Slater witness specified in the complexity-proof conventions above. For the logarithmic factors only, write $L = \log(e|\mathcal{A}|)$ and take $L_\epsilon = \log(e + H^4L/\epsilon^2 + H^5L^2/(\xi^2\epsilon^2))$.

Potential control. Uniform policy initialization and $\lambda_1 = 0$ give

$$\Phi_1 \le H \log |\mathcal{A}| + \frac{H^2 L^2}{2\xi^2}. \tag{70}$$

Thus L_ϵ bounds $\log(e + H^3\Phi_1/\epsilon^2)$ using only known problem parameters. The geometric sum in Theorem 3.4 is at most $1/(\eta\tau)$. Thus the uniform on-policy certificate gives

$$\begin{aligned}\Phi_K \le e^{-\eta\tau(K-1)}\Phi_1 + \frac{\eta}{2\tau}\sup_{k<K}\left(HC_{\eta,\tau,\Lambda,k} + G_\tau^2\right) \\ + \frac{1 + I\lambda_{\max}}{\tau}\sup_{k<K}\beta_k + \frac{H\epsilon_{\text{act}}}{\tau}.\end{aligned} \tag{71}$$

Using $\lambda_{\max} = 2HL/\xi$, the oracle bounds reduce to

$$\begin{aligned}\sup_{k<K}\left(HC_{\eta,\tau,\Lambda,k} + G_\tau^2\right) &\le C\frac{I^2H^5L^2}{\xi^2}, \\ (1 + I\lambda_{\max})\sup_{k<K}\beta_k &\le C\frac{I^2H^4L^2}{\xi^2}\sqrt{\frac{d_c^3}{N}}\,\iota_{\text{on}},\end{aligned} \tag{72}$$

where ι_{on} is defined in (48). Choose the stepsize at the following upper-bound scale, and take the smallest integer batch and iteration budgets satisfying

$$\begin{aligned}\eta \le \frac{c\tau\xi^2\epsilon^2}{I^2H^8L^2}, \qquad K = 1 + \left\lceil \frac{CL_\epsilon}{\eta\tau} \right\rceil, \\ N \ge \frac{Cd_c^3I^4H^{14}L^4}{\tau^2\xi^4\epsilon^4}\,\iota_{\text{on}}^2.\end{aligned} \tag{73}$$

The dependence of ι_{on} on N, K is logarithmic, so these choices can be satisfied without changing their displayed polynomial orders. Taking c sufficiently small also ensures the two oracle stepsize conditions. Each of the first three terms in (71) is then $O(\epsilon^2/H^3)$, giving

$$\Phi_K \le C\frac{\epsilon^2}{H^3} + \frac{H\epsilon_{\text{act}}}{\tau}. \tag{74}$$

In particular, the iteration and batch budgets are

$$K = O\left(\frac{I^2H^8L^2L_\epsilon}{\tau^2\xi^2\epsilon^2}\right), \qquad N = O\left(\frac{d_c^3I^4H^{14}L^4\iota_{\text{on}}^2}{\tau^2\xi^4\epsilon^4}\right), \tag{75}$$

which verifies the parameter orders in (69).

Temperature and actor approximation. The conversion bounds in Theorem 3.4 and (74) imply

$$\max\left\{[V_r^{\pi^\star} - V_r^{\pi^K}]_+, \max_{i\in[I]}[c_i - V_{u_i}^{\pi^K}]_+\right\} \le C\left[\epsilon + \frac{HL}{\xi}\tau + H^2\sqrt{\frac{\epsilon_{\text{act}}}{\tau}}\right]. \tag{76}$$

The last two terms express the only temperature tradeoff: a larger τ increases the regularisation error but decreases the effect of actor approximation. They balance at $\tau = (H^2\xi^2\epsilon_{\text{act}}/L^2)^{1/3}$. The target-accuracy scale is $\xi\epsilon/(HL)$. Taking the larger of these two scales, capped at one, gives (69) and the error bound (16). Indeed, below the cap, the regularisation and actor terms are both bounded by $O(\epsilon + H^{5/3}L^{1/3}\epsilon_{\text{act}}^{1/3}/\xi^{1/3})$. If the cap is active, the actor-error term in this bound is already at least a constant multiple of H, and the same conclusion follows from the range $[0, H]$ of reward and utility values. A genuine $\widetilde{O}(\epsilon)$ guarantee requires $\epsilon_{\text{act}} = \widetilde{O}(\xi\epsilon^3/H^5)$.

Sample count. Multiplying the two budgets in (75) gives

$$NK = O\left(\frac{d_c^3I^6H^{22}L^6}{\xi^6\epsilon^6\tau^4}\,\iota_{\text{on}}^2L_\epsilon\right). \tag{77}$$

The capped temperature satisfies

$$\frac{1}{\tau^4} = O\left(\max\left\{1,\ \min\left\{\frac{H^4L^4}{\xi^4\epsilon^4},\ \frac{L^{8/3}}{H^{8/3}\xi^{8/3}\epsilon_{\text{act}}^{4/3}}\right\}\right\}\right). \tag{78}$$

For $\epsilon_{\text{act}} = 0$, the second entry of the minimum is interpreted as $+\infty$. Substitution into (77) gives (17) after suppressing polylogarithmic factors. In the accuracy-sensitive regime $\epsilon_{\text{act}} \le \xi/H^2$, the cap is inactive and the outer maximum can be removed. Each episode contains H transitions, so the transition count is HNK. □

D.4 Analysis of the off-policy oracle

Proof of Lemma 4.2, off-policy branch. In the off-policy mode, $\mathcal{D}^k$ contains the entire data prefix, including samples that influenced the current actor. We therefore first construct a confidence event uniform over both actors and critics, and then bound the cumulative evaluation error through predictable elliptical widths. Set

$$\iota_K \triangleq \log\left(\frac{CKH(I+1)|\mathcal{A}|(d_{\text{c}}+d_{\text{a}})(1+\lambda_{\max})(1+\kappa_G+\epsilon_{\text{bias}})}{\delta\min\{1,\tau\}}\right), \tag{79}$$

and take

$$\begin{aligned}\alpha_r = \alpha_u &= CH\sqrt{(d_{\text{c}}^2+d_{\text{c}}+d_{\text{a}})\iota_K},\\ \alpha_\psi &= CH\log(e|\mathcal{A}|)\sqrt{(d_{\text{c}}^2+d_{\text{c}}+d_{\text{a}})\iota_K}.\end{aligned} \tag{80}$$

Uniform confidence. For every k, the off-policy dataset satisfies $|\mathcal{D}_h^k| \le K$. The normal-equation calculation in (50) therefore gives

$$\|\widehat{\theta}_{j,h}^k\|_2 \le KH\log(e|\mathcal{A}|) \triangleq B_K. \tag{81}$$

For a fixed actor, the critic still varies only through a regression coefficient $w \in \mathbb{R}^{d_{\text{c}}}$ and a symmetric matrix $0 \preceq M \preceq I_{d_{\text{c}}}$, as in (44). The additional step is to cover the actors that can be computed from $\mathcal{D}^k$.

Define the actor-fitting residual

$$\Delta_{k,h}(s,a) \triangleq \langle\varphi(s,a), \omega_h^{k+1} - \omega_h^k\rangle - \eta Q_{z_k,h}^k(s,a).$$

Equation (28) gives $\sup_{s,a,a'}|\Delta_{k,h}(s,a) - \Delta_{k,h}(s,a')| \le \eta\epsilon_{\text{act}}$. By the component truncation bounds,

$$Q_{z_k,h}^k(s,a) = \overline{Q}_{k,h}(s,a) - \tau\log\pi_h^k(a\mid s), \qquad 0 \le \overline{Q}_{k,h}(s,a) \le H(1+I\lambda_{\max}+\tau\log|\mathcal{A}|).$$

Subtracting the fitted-logit identities for actions a and a' therefore gives

$$\begin{aligned}\log\frac{\pi_h^{k+1}(a\mid s)}{\pi_h^{k+1}(a'\mid s)} &= (1-\eta\tau)\log\frac{\pi_h^k(a\mid s)}{\pi_h^k(a'\mid s)}\\ &\quad + \eta\big(\overline{Q}_{k,h}(s,a) - \overline{Q}_{k,h}(s,a')\big) + \Delta_{k,h}(s,a) - \Delta_{k,h}(s,a').\end{aligned}$$

Taking absolute values yields

$$\begin{aligned}\left|\log\frac{\pi_h^{k+1}(a\mid s)}{\pi_h^{k+1}(a'\mid s)}\right| &\le (1-\eta\tau)\left|\log\frac{\pi_h^k(a\mid s)}{\pi_h^k(a'\mid s)}\right|\\ &\quad + \eta\left[H(1+I\lambda_{\max}+\tau\log|\mathcal{A}|)+\epsilon_{\text{act}}\right].\end{aligned} \tag{82}$$

Since $\omega_h^1 = 0$, iterating (82) gives

$$\sup_{k,s,a,a'}\left|\log\frac{\pi_h^k(a\mid s)}{\pi_h^k(a'\mid s)}\right| \le \frac{H(1+I\lambda_{\max}+\tau\log|\mathcal{A}|)+\epsilon_{\text{act}}}{\tau} \triangleq L_{\text{act}}. \tag{83}$$

For a stage-h actor parameter $\omega_h \in \mathbb{R}^{d_a}$, define

$$f_{\omega,h}(s,a) \triangleq \langle \varphi(s,a), \omega_h \rangle - \frac{1}{|\mathcal{A}|} \sum_{b \in \mathcal{A}} \langle \varphi(s,b), \omega_h \rangle. \tag{84}$$

The subtracted term is independent of a, so (2) gives

$$\pi_h(a \mid s, \omega) = \frac{\exp(f_{\omega,h}(s,a))}{\sum_{b \in \mathcal{A}} \exp(f_{\omega,h}(s,b))}, \qquad f_{\omega,h}(s,a) = \frac{1}{|\mathcal{A}|} \sum_{b \in \mathcal{A}} \log \frac{\pi_h(a \mid s, \omega)}{\pi_h(b \mid s, \omega)}. \tag{85}$$

Equations (83) and (85) imply $\|f_{\omega^k,h}\|_\infty \le L_{\mathrm{act}}$. Define

$$\mathcal{F}_h^{\mathrm{off}} \triangleq \left\{ f_{\omega,h} : \omega_h \in \mathbb{R}^{d_a},\ \|f_{\omega,h}\|_\infty \le L_{\mathrm{act}} \right\}. \tag{86}$$

Thus $f_{\omega^k,h} \in \mathcal{F}_h^{\mathrm{off}}$ for every k and every realization of $\mathcal{D}^k$. Since $\omega_h \mapsto f_{\omega,h}$ is linear and its image has dimension at most d_a, a volumetric bound gives

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{F}_h^{\mathrm{off}}) \le d_a \log\left(1 + \frac{2L_{\mathrm{act}}}{\varepsilon}\right), \tag{87}$$

for every $\varepsilon > 0$.

The fixed-policy classes in (44) already contain every value function that can be produced by the backward recursion. Define

$$\mathcal{V}_{j,h}^{\mathrm{off}}(\omega) \triangleq \mathcal{V}_{j,h}(\pi(\omega), B_K), \qquad \mathcal{V}_{j,h}^{\mathrm{off}} \triangleq \bigcup_{\omega:\, f_{\omega,h} \in \mathcal{F}_h^{\mathrm{off}}} \mathcal{V}_{j,h}^{\mathrm{off}}(\omega), \qquad \mathcal{V}_h^{\mathrm{off}} \triangleq \bigcup_{j \in \{r, u_1, \dots, u_I, \psi\}} \mathcal{V}_{j,h}^{\mathrm{off}}. \tag{88}$$

At the terminal stage, set $\mathcal{V}_{j,H+1}^{\mathrm{off}} = \mathcal{V}_{H+1}^{\mathrm{off}} = \{0\}$. The bounds defining (86) and (88) depend only on the problem parameters and algorithmic budgets; hence both classes are fixed independently of the realized data. The oracle definition and (81) imply

$$V_{j,h}^k = v_{j,h}^{\pi^k, \widehat{\theta}_{j,h}^k, (\Sigma_h^k)^{-1}} \in \mathcal{V}_{j,h}^{\mathrm{off}} \subseteq \mathcal{V}_h^{\mathrm{off}} \tag{89}$$

for every (k,h,j).

Suppose $\|f_{\omega,h} - f_{\omega',h}\|_\infty \le \nu \le 1$. Then

$$\left| \log \frac{\pi_h(a \mid s, \omega)}{\pi_h(a \mid s, \omega')} \right| \le 2\nu, \qquad \|\pi_h(\cdot \mid s, \omega) - \pi_h(\cdot \mid s, \omega')\|_1 \le e^{2\nu} - 1 \le C\nu. \tag{90}$$

Using (85) and $\|f_{\omega,h}\|_\infty, \|f_{\omega',h}\|_\infty \le L_{\mathrm{act}}$ further gives

$$\begin{aligned} &\Big| \mathbb{E}_{a \sim \pi_h(\cdot \mid s, \omega)}[-\log \pi_h(a \mid s, \omega)] \\ &\qquad - \mathbb{E}_{a \sim \pi_h(\cdot \mid s, \omega')}[-\log \pi_h(a \mid s, \omega')] \Big| \le C(1 + L_{\mathrm{act}})\nu, \\ &\left| \mathbb{E}_{\pi_h(\cdot \mid s, \omega)}[Q] - \mathbb{E}_{\pi_h(\cdot \mid s, \omega')}[Q] \right| \le C\|Q\|_\infty \nu. \end{aligned} \tag{91}$$

Consequently, for the same (j, w, M),

$$\left\| v_{j,h}^{\pi(\omega), w, M} - v_{j,h}^{\pi(\omega'), w, M} \right\|_\infty \le C\big(H \log(e|\mathcal{A}|) + L_{\mathrm{act}}\big)\nu. \tag{92}$$

Applying Lemma D.1 with $B = B_K$, $L_h = C(H \log(e|\mathcal{A}|) + L_{\mathrm{act}})$, and $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{off}}$, and then using (87), gives

$$\log \mathcal{N}_\infty(\varepsilon, \mathcal{V}_h^{\mathrm{off}}) \le C(d_c^2 + d_c + d_a)\left[\iota_K + \log(1 + 1/\varepsilon)\right]. \tag{93}$$

The $d_c^2 + d_c$ terms come from the critic class, while d_a comes from covering the actor class.

Since each $\mathcal{V}_{j,h+1}^{\mathrm{off}}$ is deterministic and contains every corresponding $V_{j,h+1}^k$, Lemma D.2, applied with $\mathcal{W}_{h+1} = \mathcal{V}_{j,h+1}^{\mathrm{off}}$, $m_{\max} = K$, $\varepsilon_0 = (2KHd_c)^{-2}$, $\delta_0 = \delta/[2H(I+2)]$, and the same (Y_j, S_j)

bounds as in the on-policy case, gives, with probability at least $1-\delta/2$,

$$\left|\langle \phi(s,a), \widehat{\theta}^k_{j,h} - \theta^k_{j,h}\rangle\right| \le \alpha_j \|\phi(s,a)\|_{(\Sigma^k_h)^{-1}}, \qquad j \in \{r, u_1, \dots, u_I, \psi\}, \tag{94}$$

simultaneously for every iterate and (s,a,h). Uniformity over $\mathcal{V}^{\mathrm{off}}_{j,h+1}$ and (89) permits the $V^k_{j,h+1}$ to depend on the full dataset $\mathcal{D}^k$ used in the regression; the actor π^k depends on the preceding prefix $\mathcal{D}^{k-1} \subseteq \mathcal{D}^k$.

Optimism and evaluation error. Applying the componentwise derivation (55)- (58) on the event (94) gives, for every component j,

$$0 \le V^k_j - V^{\pi^k}_j \le 2\sum_{h=1}^{H} \mathbb{E}_{d^{\pi^k}_h}[b^k_{j,h}], \tag{95}$$

where $V^{\pi^k}_j = V^{\pi^k}_{\psi_k}$ for $j = \psi$. Since $\Sigma^k_h \succeq \Sigma^{k-1}_h$, $\|\phi(s,a)\|_{(\Sigma^k_h)^{-1}} \le \|\phi(s,a)\|_{(\Sigma^{k-1}_h)^{-1}}$. Define

$$W_k \triangleq \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim d^{\pi^k}_h}\left[\|\phi(s,a)\|_{(\Sigma^{k-1}_h)^{-1}}\right]. \tag{96}$$

Therefore, (95) and (80) imply

$$\begin{aligned} 0 \le V^k_{z_k} - V^{\pi^k}_{z_k} &\le 2\big(\alpha_r + I\lambda_{\max}\alpha_u + \tau\alpha_\psi\big)W_k, \\ \max_{i\in[I]} |V^k_{u_i} - V^{\pi^k}_{u_i}| &\le 2\alpha_u W_k. \end{aligned}$$

The first right-hand side dominates the second and satisfies

$$\begin{aligned} 2\big(\alpha_r + I\lambda_{\max}\alpha_u + \tau\alpha_\psi\big)W_k \le CH\sqrt{(d_{\mathrm{c}}^2 + d_{\mathrm{c}} + d_{\mathrm{a}})\iota_K} \\ \times \big(1 + I\lambda_{\max} + \tau\log(e|\mathcal{A}|)\big)W_k. \end{aligned} \tag{97}$$

We take the right-hand side of (97) as β_k. Utility truncation gives $0 \le V^k_{u_i} \le H$.

Cumulative width. Let $X_{k,h} = \|\phi(s^k_h, a^k_h)\|_{(\Sigma^{k-1}_h)^{-1}}$. Both π^k and Σ^{k-1}_h are measurable before trajectory k, so $W_k = \mathbb{E}[\sum_h X_{k,h} \mid \mathcal{F}_{k-1}]$. Since $\sum_h X_{k,h} \in [0,H]$, Azuma-Hoeffding gives, with probability at least $1-\delta/2$,

$$\sum_{k=1}^{K-1} W_k \le \sum_{k=1}^{K-1}\sum_{h=1}^{H} X_{k,h} + H\sqrt{2(K-1)\log(2/\delta)}. \tag{98}$$

For each stage, the elliptical-potential identity and $\mathrm{tr}(\Sigma^{K-1}_h) \le d_{\mathrm{c}} + K - 1$ imply

$$\sum_{k=1}^{K-1} X^2_{k,h} \le 2\log\det(\Sigma^{K-1}_h) \le 2d_{\mathrm{c}}\log\left(1 + \frac{K-1}{d_{\mathrm{c}}}\right). \tag{99}$$

Cauchy-Schwarz over (k,h) gives

$$\sum_{k=1}^{K-1}\sum_{h=1}^{H} X_{k,h} \le H\sqrt{2d_{\mathrm{c}}(K-1)\log\left(1 + \frac{K-1}{d_{\mathrm{c}}}\right)}. \tag{100}$$

Combining (98) and (100) yields

$$\sum_{k=1}^{K-1} W_k \le CH\sqrt{d_{\mathrm{c}}K\log(eK/\delta)}. \tag{101}$$

Combining (97) and (101) proves the cumulative consistency bound in (15), since $d_{\mathrm{c}}^2 + d_{\mathrm{c}} + d_{\mathrm{a}} = \Theta(d_{\mathrm{c}}^2 + d_{\mathrm{a}})$. The confidence and width events hold jointly with probability at least $1-\delta$.

Local norm. This calculation is independent of the sampling mode. For $p_a = \pi_h^k(a \mid s)$, clipping gives

$$0 \le Q^k_{z_k,h}(s,a) \le H(1 + I\lambda_{\max} + \tau \log|\mathcal{A}|) + \tau \log(1/p_a). \tag{102}$$

Applying the entropy-moment argument (66)-(67) under the two stated stepsize conditions therefore yields

$$\sum_a p_a e^{\eta Q^k_{z_k,h}(s,a)} \big(Q^k_{z_k,h}(s,a)\big)^2 \le CH^2\big(1 + I\lambda_{\max} + \tau \log(e|\mathcal{A}|)\big)^2. \tag{103}$$

This verifies the remaining oracle condition and completes the off-policy branch. □

D.5 Proof of the off-policy final rate

Fix $\epsilon, \delta \in (0,1)$ and choose the parameters as

$$\begin{aligned} \lambda_{\max} &= \frac{2H\log(e|\mathcal{A}|)}{\xi}, \quad \tau = \Theta\left(\min\left\{1, \max\left\{\frac{\xi\epsilon}{H\log(e|\mathcal{A}|)}, \frac{H^{2/3}\xi^{2/3}\epsilon_{\text{act}}^{1/3}}{\log^{2/3}(e|\mathcal{A}|)}\right\}\right\}\right), \\ \eta &= \widetilde{\Theta}\left(\frac{\tau\xi^4\epsilon^4}{d_{\text{c}}(d_{\text{c}}^2 + d_{\text{c}} + d_{\text{a}})I^4H^{14}}\right), \qquad K = \widetilde{\Theta}\left(\frac{1}{\eta\tau}\right). \end{aligned} \tag{104}$$

Proof of Theorem 4.3, off-policy branch. Off-policy sampling supplies a cumulative $O(\sqrt{K})$ oracle certificate rather than a uniform $O(N^{-1/2})$ certificate. Set $L = \log(e|\mathcal{A}|)$ and $L_\epsilon = \log(e + H^4L/\epsilon^2 + H^5L^2/(\xi^2\epsilon^2))$. Uniform policy initialization, $\lambda_1 = 0$, and (35) give

$$\Phi_1 \le H\log|\mathcal{A}| + \frac{H^2L^2}{2\xi^2}. \tag{105}$$

Potential control. For the local-norm and actor terms, sum the geometric weights in Theorem 3.4. For the nonnegative oracle errors, bound each weight by one. This gives

$$\begin{aligned} \Phi_K \le e^{-\eta\tau(K-1)}\Phi_1 + \frac{\eta}{2\tau}\sup_{k<K}\big(HC_{\eta,\tau,\Lambda,k} + G_\tau^2\big) \\ + \eta(1 + I\lambda_{\max})\sum_{k=1}^{K-1}\beta_k + \frac{H\epsilon_{\text{act}}}{\tau}. \end{aligned} \tag{106}$$

Substituting $\lambda_{\max} = 2HL/\xi$ into the off-policy certificate gives

$$\begin{aligned} \sup_{k<K}\big(HC_{\eta,\tau,\Lambda,k} + G_\tau^2\big) &\le C\frac{I^2H^5L^2}{\xi^2}, \\ (1 + I\lambda_{\max})\sum_{k=1}^{K-1}\beta_k &\le C\frac{I^2H^4L^2}{\xi^2}\sqrt{d_{\text{c}}(d_{\text{c}}^2 + d_{\text{a}})K}\,\iota_K. \end{aligned} \tag{107}$$

Here ι_K is defined in (79); it also absorbs the logarithm in the cumulative-width bound. Take

$$\eta = \frac{c\tau\xi^4\epsilon^4}{d_{\text{c}}(d_{\text{c}}^2 + d_{\text{a}})I^4H^{14}L^4\iota_K^2L_\epsilon}, \qquad K = 1 + \left\lceil\frac{CL_\epsilon}{\eta\tau}\right\rceil.$$

These implicit logarithmic choices have the polynomial orders in (104). The normalized parameter ranges imply that this stepsize is no larger than a constant multiple of $\tau\xi^2\epsilon^2/(I^2H^8L^2)$, which controls the local-norm contribution, and is small enough for the oracle stability conditions. Moreover, $K = O(L_\epsilon/(\eta\tau))$ gives

$$\eta(1 + I\lambda_{\max})\sum_{k=1}^{K-1}\beta_k \le C\frac{I^2H^4L^2}{\xi^2}\sqrt{\frac{d_{\text{c}}(d_{\text{c}}^2 + d_{\text{a}})\eta L_\epsilon}{\tau}}\,\iota_K = O(\epsilon^2/H^3).$$

The contraction term has the same order by the choice of K. Consequently,

$$\Phi_K \le C\frac{\epsilon^2}{H^3} + \frac{H\epsilon_{\text{act}}}{\tau}. \tag{108}$$

Temperature and sample count. Applying (11) to (108) and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ yields

$$\max\left\{[V_r^{\pi^\star} - V_r^{\pi^K}]_+, \max_{i\in[I]}[c_i - V_{u_i}^{\pi^K}]_+\right\} \le C\left[\epsilon + \frac{HL}{\xi}\tau + H^2\sqrt{\frac{\epsilon_{\mathrm{act}}}{\tau}}\right]. \tag{109}$$

The two temperature-dependent terms balance at $\tau = (H^2\xi^2\epsilon_{\mathrm{act}}/L^2)^{1/3}$, while the target-accuracy scale is $\xi\epsilon/(HL)$. Taking the larger scale and capping it at one gives (104) and (16).

Each off-policy iteration adds one episode to $\mathcal{D}^k$, and each episode contains H transitions. Before substituting the temperature, the total transition count is

$$HK = O\left(\frac{d_{\mathrm{c}}(d_{\mathrm{c}}^2 + d_{\mathrm{a}})I^4H^{15}L^4}{\xi^4\epsilon^4\tau^2}\,\iota_K^2 L_\epsilon^2\right). \tag{110}$$

This count already dominates the local-norm and stability budgets, as shown by the stepsize comparisons above. The temperature contributes

$$\frac{1}{\tau^2} = O\left(\max\left\{1, \min\left\{\frac{H^2L^2}{\xi^2\epsilon^2}, \frac{L^{4/3}}{H^{4/3}\xi^{4/3}\epsilon_{\mathrm{act}}^{2/3}}\right\}\right\}\right). \tag{111}$$

For $\epsilon_{\mathrm{act}} = 0$, the second entry of the minimum is interpreted as $+\infty$. Substituting (111) into (110), dividing by H, and suppressing polylogarithmic factors gives (18). When $\epsilon_{\mathrm{act}} \le \xi/H^2$, the cap is inactive and the outer maximum can be removed. □

E Proofs for Section 5

We prove the general oracle certificates in three steps: control the fitted Bellman targets uniformly over the value function used by the oracle, bound the resulting confidence widths, and substitute these bounds into the master contraction. Throughout, $\mathcal{X}$, $\mathcal{J}$, $\mathcal{F}_{j,h}$, and the actor notation are those of the main text.

E.1 General function classes and complexity measures

We first specify the two complexity measures used in the main theorem. Following Russo and Van Roy (2013, Definitions 3-4), a point $x \in \mathcal{X}$ is ρ-dependent on $x_1, \dots, x_m$ with respect to $\mathcal{F}$ if every $f, f' \in \mathcal{F}$ satisfying $\left(\sum_{q=1}^m (f(x_q) - f'(x_q))^2\right)^{1/2} \le \rho$ also satisfies $|f(x) - f'(x)| \le \rho$; otherwise it is ρ-independent.

Definition E.1 (Scale-dependent eluder dimension)**.** The eluder dimension $d_{\mathrm{E}}(\mathcal{F}, \rho)$ is the supremum of the integers $L \ge 0$ for which there exist $x_1, \dots, x_L$ and a common $\rho' \ge \rho$ such that every x_ℓ is ρ'-independent of its predecessors. For the component critic classes, define

$$d_{\mathrm{E}}(\rho) \triangleq 1 \vee \max_{j\in\mathcal{J},\, h\in[H]} d_{\mathrm{E}}(\mathcal{F}_{j,h}, \rho). \tag{112}$$

Definition E.2 (Domain metric entropy)**.** For a critic class $\mathcal{F}$, let $d_{\mathcal{F}}(x, x') = \sup_{f\in\mathcal{F}} |f(x) - f(x')|$. The aggregate domain metric entropy is

$$\mathfrak{h}_{\mathcal{X}}(\rho) \triangleq 1 \vee \max_{j\in\mathcal{J},\, h\in[H]} \log \mathcal{N}\big(\rho, \mathcal{X}, d_{\mathcal{F}_{j,h}}\big), \tag{113}$$

where $\mathcal{N}(\rho, \mathcal{X}, d_{\mathcal{F}})$ is the smallest cardinality of a ρ-cover of $\mathcal{X}$ in the class-induced pseudometric $d_{\mathcal{F}}$.

These measures have different roles. The eluder dimension controls how often confidence widths can remain large along a sequence of observations. The domain entropy bounds the number of possible rounded representatives of state-action pairs $(s, a) \in \mathcal{X}$, which is used below to bound the number of possible data-dependent bonuses. This is the domain-cover term in Wang et al. (2020, Assumption 2). The sup-norm critic cover $\mathcal{N}_\infty(\rho, \mathcal{F}_{j,h})$ need not be imposed separately because Hanneke (2024, Theorem 14) bounds it by the scale-dependent eluder dimension; see Lemma E.4. For a finite state-action domain, $\mathfrak{h}_{\mathcal{X}}(\rho) \le 1 \vee \log(|\mathcal{S}||\mathcal{A}|)$.

Accuracy-dependent resolutions. Fix the target accuracy ϵ and choose τ as in (143). Write $L = \log(e|\mathcal{A}|)$, $B = HL$, and $L_\epsilon = \log(e + H^4L/\epsilon^2 + H^5L^2/(\xi^2\epsilon^2))$. For a sufficiently small fixed universal constant $c_0 > 0$, define, in this order,

$$r_\epsilon = \min\left\{1, \frac{c_0\tau\xi^2\epsilon^2}{L_\epsilon I^2 H^6 L^2}\right\}, \quad d_0 = d_{\mathrm{E}}(r_\epsilon), \qquad q_\epsilon = \frac{r_\epsilon}{\sqrt{d_0}}, \quad \nu_\epsilon = \frac{c_0 q_\epsilon^2}{B}. \tag{114}$$

The quantities in Theorem 5.2 are evaluated at

$$d_{\mathrm{E}} = d_{\mathrm{E}}(\nu_\epsilon/24), \qquad \mathfrak{h}_{\mathcal{X}} = \mathfrak{h}_{\mathcal{X}}(\nu_\epsilon). \tag{115}$$

We consider classes for which these positive-resolution quantities are finite; no restriction on their growth as the resolution decreases is imposed. All scales are fixed before choosing the sample cap $M = N$ or $M = K$, and $d_0 \le d_{\mathrm{E}}$. The confidence level and sample cap enter only logarithmic factors, not the arguments in (115). In particular, the stated rates retain any accuracy dependence of d_{E} and $\mathfrak{h}_{\mathcal{X}}$; these quantities are not treated as constants for an arbitrary class.

The construction below retains a squared-norm error of order Mq_ϵ^2 instead of forcing that error to vanish with M. Eluder counting at resolution r_ϵ converts it into a width contribution of order $Mq_\epsilon\sqrt{d_0} = Mr_\epsilon$. The choice of r_ϵ makes its contribution to the master potential $O(\epsilon^2/H^3)$.

Remark E.3 (Linear specialization)**.** Consider the bounded linear class $\mathcal{F}_{j,h}^{\mathrm{lin}} = \{x \mapsto \langle\phi(x), \theta\rangle : \|\theta\|_2 \le R,\ 0 \le \langle\phi(x), \theta\rangle \le B_{j,h} \text{ for all } x\}$, where $B_{r,h} = B_{u_i,h} = H + 1 - h$, $B_{\psi,h} = (H - h)\log|\mathcal{A}|$, and $\|\phi(x)\|_2 \le 1$. Then

$$d_{\mathcal{F}_{j,h}^{\mathrm{lin}}}(x, x') \le R\|\phi(x) - \phi(x')\|_2.$$

A Euclidean cover of the feature image therefore gives

$$\log\mathcal{N}(\rho, \mathcal{X}, d_{\mathcal{F}_{j,h}^{\mathrm{lin}}}) \le d_{\mathrm{c}}\log\left(1 + \frac{2R}{\rho}\right).$$

The linear-CMDP Bellman targets of Section 4 belong to these classes with $R \le H\sqrt{d_{\mathrm{c}}}\log(e|\mathcal{A}|)$. Consequently,

$$\mathfrak{h}_{\mathcal{X}}(\rho) \le 1 \vee \left[d_{\mathrm{c}}\log\left(1 + \frac{2H\sqrt{d_{\mathrm{c}}}\log(e|\mathcal{A}|)}{\rho}\right)\right] = \widetilde{O}(d_{\mathrm{c}}). \tag{116}$$

The cover depends only on the feature image, so $\mathcal{S}$ need not be finite. Since the classes consist of bounded linear functions, the linear eluder-dimension bound also gives $d_{\mathrm{E}}(\rho) = \widetilde{O}(d_{\mathrm{c}})$ (Russo and Van Roy, 2013). The oracle clips its optimistic estimates separately; clipping is not part of the critic-class definition used here.

E.2 Stable optimistic policy evaluation

For a finite multiset $Z = (x_\ell)_{\ell=1}^n$ of points in $\mathcal{X}$, write $\|f\|_Z^2 = \sum_{\ell=1}^n f(x_\ell)^2$, so repeated points contribute separately (Z may contain repeated points in $\mathcal{X}$). If occurrence x_ℓ has weight $w_\ell \ge 0$, use $\|f\|_Z^2 = \sum_{\ell=1}^n w_\ell f(x_\ell)^2$. We adapt the sensitivity sampling and rounding of Wang et al. (2020, Algorithm 3) to the fixed resolutions in (114). Unlike their budget-dependent construction, we fix the error Mq_ϵ^2 in the confidence radius. Lemma E.5 states the two properties needed below: domination of the full-data confidence width and a deterministic bound on the number of possible bonuses.

Algorithm 3 $\mathrm{StableBonus}(\mathcal{F}, \overline{f}, Z, \alpha, M, \delta_0)$

1: Write $Z = (x_\ell)_{\ell=1}^n$, where $n = |Z| \le M$ and $\alpha \ge Mq_\epsilon^2$. Set
$$R_{\mathcal{F}} = \sup_{x\in\mathcal{X}} \sup_{f,g\in\mathcal{F}} |f(x) - g(x)|.$$
2: If $R_{\mathcal{F}} = 0$, return $b \equiv 0$. If $n = 0$, return $b(x) = \sup_{f,g\in\mathcal{F}} |f(x) - g(x)|$.
 # Use sensitivity sampling to decide which x_ℓ enter Z'.
3: **for** $\ell = 1, \dots, n$ **do**
4: Compute sens_ℓ by (121) and p_ℓ by (123).
5: Independently draw $\zeta_\ell \sim \mathrm{Bernoulli}(p_\ell)$.
6: **end for**
7: Let Z' contain x_ℓ with weight $1/p_\ell$ whenever $\zeta_\ell = 1$.
8: Set J_M by (124).
9: **if** $\sum_{\ell=1}^n \zeta_\ell > J_M$ or $\sum_{\ell=1}^n \zeta_\ell/p_\ell > 2n$ **then**
10: **return** $b \equiv R_{\mathcal{F}}$.
11: **end if**
 # Replace each selected x_ℓ by a nearby cover point
12: Initialize $Z^\sharp$ as an empty weighted multiset.
13: **for** $\ell = 1, \dots, n$ **do**
14: **if** $\zeta_\ell = 1$ **then**
15: Choose $x_\ell^\sharp$ from a fixed ν_ϵ-cover of $(\mathcal{X}, d_{\mathcal{F}})$ such that $d_{\mathcal{F}}(x_\ell, x_\ell^\sharp) \le \nu_\epsilon$.
16: Add $x_\ell^\sharp$ to $Z^\sharp$ with weight $1/p_\ell$.
17: **end if**
18: **end for**
19: Choose $f^\sharp$ from a fixed internal sup-norm ν_ϵ-cover of $\mathcal{F}$ such that $\|\overline{f} - f^\sharp\|_\infty \le \nu_\epsilon$.
20: Set $\widehat{\mathcal{F}} = \{f \in \mathcal{F} : \|f - f^\sharp\|_{Z^\sharp}^2 \le 7\alpha\}$.
21: **return** $b(x) = \sup_{f,g\in\widehat{\mathcal{F}}} |f(x) - g(x)|$ for $x \in \mathcal{X}$.

Algorithm 3 takes a critic class $\mathcal{F} \subseteq (\mathcal{X} \to [0, B])$, a fitted center $\overline{f} \in \mathcal{F}$, data $Z = (x_\ell)_{\ell=1}^n$, a squared empirical-norm radius α, a sample cap M, and a compression failure budget δ_0. It returns a bonus $b : \mathcal{X} \to [0, B]$.

Line 1 specifies $n \le M$, $\alpha \ge Mq_\epsilon^2$, and the class diameter $R_{\mathcal{F}}$; line 2 handles a zero-range class or empty data. Lines 4–7 compute the sensitivity sens_ℓ, select x_ℓ with probability p_ℓ, and assign selected points weight $1/p_\ell$. Larger sensitivity increases the selection probability, while inverse-probability weighting compensates for subsampling. The budget δ_0 enters p_ℓ and J_M through their definitions. Lines 8–10 return $R_{\mathcal{F}}$ if the selected count exceeds J_M or the total weight exceeds $2n$. Lines 15–16 round selected points into $Z^\sharp$ without changing their weights; line 19 rounds $\overline{f}$ to $f^\sharp$ which is its nearest point in the cover of $\mathcal{F}$. Lines 20–21 form $\widehat{\mathcal{F}}$ and return its pointwise width as b. The checks and fixed covers give a deterministic finite family of possible bonuses. Regression still uses all supplied samples.

Given the stable bonus computation in Algorithm 3, Algorithm 4 follows the backward stagewise structure of the linear oracle in Algorithm 2: it fits each component critic, adds an optimism bonus, truncates the estimate, and computes its policy average. Here $\widehat{f}_{j,h}^k$ is fitted over $\mathcal{F}_{j,h}$, while Algorithm 3 supplies its bonus $b_{j,h}^k$.

Write $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_H)$, where $\mathcal{D}_h$ contains the stage-h transition from *every trajectory* in $\mathcal{D}$. Importantly, we note that it does not split $\mathcal{D}$ into separate training subsets and the stage-h regression uses every transition in $\mathcal{D}_h$. Set $y_{r,h}^\ell = r_h^\ell$, $y_{u_i,h}^\ell = u_{i,h}^\ell$, and $y_{\psi,h}^\ell = 0$. Given $V_{j,h+1}^k$, define

$$\widehat{f}_{j,h}^k \in \arg\min_{f\in\mathcal{F}_{j,h}} \sum_{\ell\in\mathcal{D}_h} \left(f(s_h^\ell, a_h^\ell) - y_{j,h}^\ell - V_{j,h+1}^k(s_{h+1}^\ell)\right)^2. \tag{117}$$

We assume measurable exact minimizers. An additive least-squares optimisation tolerance can instead be included in the confidence radius.

Algorithm 4 Stable General $\mathrm{OPE}_{\mathrm{gen}}(\pi^k, \lambda_k, \mathcal{D})$

Input: π^k, λ_k, the stage-indexed dataset $\mathcal{D}$, critic classes, radii $\{\alpha_{j,h}\}$, sample cap M, and failure budgets $\{\delta_{j,h}\}$.
Set $V^k_{j,H+1} = 0$ for every $j \in \mathcal{J}$.
for $h = H, H-1, \dots, 1$ **do**
  for $j \in \mathcal{J}$ **do**
    Compute $\widehat{f}^k_{j,h}$ by (117) using all of $\mathcal{D}_h$.
    $b^k_{j,h} \leftarrow \mathrm{StableBonus}(\mathcal{F}_{j,h}, \widehat{f}^k_{j,h}, \{(s^\ell_h, a^\ell_h) : \ell \in \mathcal{D}_h\}, \alpha_{j,h}, M, \delta_{j,h})$.
    Set $\overline{Q}^k_{j,h} = \mathrm{Truncate}_{[0,B_{j,h}]}(\widehat{f}^k_{j,h} + b^k_{j,h})$, where $B_{r,h} = B_{u_i,h} = H + 1 - h$ and $B_{\psi,h} = (H-h)\log|\mathcal{A}|$.
  end for
  Set $Q^k_{r,h} = \overline{Q}^k_{r,h}$, $Q^k_{u_i,h} = \overline{Q}^k_{u_i,h}$, and $Q^k_{\psi,h}(s,a) = -\log \pi^k_h(a \mid s) + \overline{Q}^k_{\psi,h}(s,a)$.
  Set $V^k_{j,h}(s) = \mathbb{E}_{a\sim\pi^k_h(\cdot|s)}[Q^k_{j,h}(s,a)]$ for every $j \in \mathcal{J}$.
end for
Return $Q^k_{z_k,h} = Q^k_{r,h} + \lambda_k^\top Q^k_{u,h} + \tau Q^k_{\psi,h}$ and $V^k_u = (V^k_{u_1,1}(s_1), \dots, V^k_{u_I,1}(s_1))$.

E.3 Technical lemmas for uniform confidence

The stage-h regression uses $V^k_{j,h+1}$, which is computed from the same trajectories. We therefore need a confidence bound that holds for every value function the oracle might produce, rather than for one fixed value function. This subsection shows that the stable bonus $b^k_{j,h}$ controls the empirical critic width and belongs to a deterministic finite family. Together with covers of the critics and actors, this gives a class containing $V^k_{j,h+1}$ over which the regression bound holds uniformly.

For $\mathcal{F} \subseteq (\mathcal{X} \to [0,B])$, a finite multiset $Z \subset \mathcal{X}$, and $\gamma \geq 0$, define the empirical confidence set and pointwise width as

$$\mathcal{C}_Z(\overline{f}, \gamma) = \{f \in \mathcal{F} : \|f - \overline{f}\|^2_Z \leq \gamma\}, \qquad w(\mathcal{C}, x) = \sup_{f,f' \in \mathcal{C}} |f(x) - f'(x)|.$$

Recall the pseudometric $d_{\mathcal{F}}(x,x') = \sup_{f\in\mathcal{F}} |f(x) - f(x')|$.

Lemma E.4 (Critic entropy from eluder dimension). *Let $\mathcal{F} \subseteq (\mathcal{X} \to [0,B])$ and $0 < \rho \leq B$. Then*

$$\log \mathcal{N}_\infty(\rho, \mathcal{F}) \leq C\left[1 + d_{\mathrm{E}}\Big(\mathcal{F}, \frac{\rho}{6}\Big)\right] \log\left(\frac{eB\,[1 + d_{\mathrm{E}}(\mathcal{F}, \rho/6)]}{\rho}\right). \tag{118}$$

Here $\mathcal{N}_\infty$ is the internal sup-norm covering number used by the stable-bonus construction.

Proof. Rescale to $\mathcal{F}/B \subseteq [0,1]$ and let $m = \log \widetilde{\mathcal{N}}_\infty(\rho/2, \mathcal{F})$, where the tilde denotes an external cover whose centers need not lie in $\mathcal{F}$. If $m \leq 1$, the right-hand side of (118) is at least a universal constant, so the result holds. Otherwise, Hanneke (2024, Theorem 14), at scale $\rho/(6B)$, gives

$$\left\lfloor \frac{2m}{\log(144B^2 m/\rho^2)} \right\rfloor \leq d_{\mathrm{E}}(\mathcal{F}, \rho/6).$$

The dimension used in that theorem restricts one function in each witnessing pair to a fixed center. Removing this restriction can only increase the longest admissible sequence, so it is bounded by Definition E.1. Inverting the preceding inequality gives

$$m \leq C\,[1 + d_{\mathrm{E}}(\mathcal{F}, \rho/6)] \log\left(\frac{eB[1 + d_{\mathrm{E}}(\mathcal{F}, \rho/6)]}{\rho}\right).$$

Choose one function of $\mathcal{F}$ from each nonempty external cover cell. These functions form an internal ρ-cover of the same cardinality, proving the result. □

Lemma E.5 (Stable-width compression). *For a component class $\mathcal{F}$, $|Z| \leq M$, $\overline{f} \in \mathcal{F}$, and $\delta_0 \in (0,1)$, set $\alpha_{\mathcal{F}}(M, \delta_0) = Mq^2_\epsilon$. For every fixed $\alpha \geq \alpha_{\mathcal{F}}(M, \delta_0)$, Algorithm 3 satisfies, with*

probability at least $1-\delta_0/(16M)$*,*

$$w\big(\mathcal{C}_Z(\overline{f},\alpha),x\big)\le b(x)\le w\big(\mathcal{C}_Z(\overline{f},34\alpha),x\big),\qquad \forall x\in\mathcal{X}. \tag{119}$$

For each such radius, every bonus b *returned by Algorithm 3 belongs to a deterministic finite family* $\mathcal{W}(\mathcal{F},M,\delta_0;\alpha)$ *satisfying*

$$\sup_{\alpha\ge\alpha_{\mathcal{F}}(M,\delta_0)}\log|\mathcal{W}(\mathcal{F},M,\delta_0;\alpha)|=\widetilde{O}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}}). \tag{120}$$

The complexity scales are those of (115)*; only logarithmic factors depend on* M *and* δ_0*.*

Proof. We follow the sensitivity-sampling strategy of Wang et al. (2020, Appendix A.1), but use the fixed denominator floor $|Z|q_\epsilon^2$ and prove the resulting additive-error bounds explicitly. The empty dataset and zero-range class have exact widths, so suppose $Z=(x_1,\dots,x_n)$ with $1\le n\le M$. The component range is at most $B=H\log(e|\mathcal{A}|)$.

Sensitivity and retained support. For each index $\ell\in[n]$, define

$$\mathrm{sens}_\ell=\sup_{f,g\in\mathcal{F}}\frac{(f(x_\ell)-g(x_\ell))^2}{\max\{\|f-g\|_Z^2,nq_\epsilon^2\}}. \tag{121}$$

Each sensitivity is at most one. For every positive sensitivity, choose a witness $v_\ell=f_\ell-g_\ell$ whose quotient is at least $\mathrm{sens}_\ell/2$. If $|v_\ell(x_\ell)|\le q_\epsilon$, then $\mathrm{sens}_\ell\le 2/n$. Partition the remaining indices by $a<|v_\ell(x_\ell)|\le 2a$, where $a=2^jq_\epsilon$. Process the points in this group in their original order. Append each x_ℓ to the first existing subsequence for which it is a-independent of the points already assigned to that subsequence. If there is no such subsequence, start a new one with x_ℓ. Every subsequence has length at most $d_{\mathrm{E}}(\mathcal{F},a)$. If x_ℓ enters subsequence $m\ge 2$, it is a-dependent on each of the $m-1$ preceding subsequences. Since $|v_\ell(x_\ell)|>a$, their disjointness gives

$$\|v_\ell\|_Z^2>(m-1)a^2,\qquad \mathrm{sens}_\ell\le\min\{1,8/(m-1)\}.$$

Summing over the at most $d_{\mathrm{E}}(\mathcal{F},a)$ entries of each subsequence and then the dyadic groups yields

$$\sum_{\ell=1}^n\mathrm{sens}_\ell\le C\left[1+d_{\mathrm{E}}(\mathcal{F},q_\epsilon)\log(eB/q_\epsilon)\log(en)\right]. \tag{122}$$

Let $\ell_M=\log(64M\mathcal{N}_\infty(\nu_\epsilon,\mathcal{F})^2/\delta_0)$. For a sufficiently large universal constant C_{s}, take

$$a_\ell=\min\{1,C_{\mathrm{s}}\ell_M(\mathrm{sens}_\ell+1/n)\},\qquad p_\ell=\frac{1}{\lfloor 1/a_\ell\rfloor},\qquad \zeta_\ell\sim\mathrm{Bernoulli}(p_\ell)\quad\text{independently.} \tag{123}$$

Then $a_\ell\le p_\ell\le 2a_\ell$, and each retained weight $1/p_\ell$ is an integer in $[n]$. By (122) and Bernstein's inequality, the number of retained indices is at most

$$J_M=\lceil C\ell_M\left[1+d_{\mathrm{E}}(\mathcal{F},q_\epsilon)\log(eB/q_\epsilon)\log(eM)\right]\rceil \tag{124}$$

except on an event of probability at most $\delta_0/(64M)$. Moreover, for indices with $p_\ell<1$, $1/p_\ell\le n/(C_{\mathrm{s}}\ell_M)$. Thus

$$\sum_{\ell=1}^n\mathrm{Var}(\zeta_\ell/p_\ell)\le\frac{n^2}{C_{\mathrm{s}}\ell_M},\qquad \Pr\left\{\sum_{\ell=1}^n\zeta_\ell/p_\ell>2n\right\}\le\frac{\delta_0}{64M}.$$

Uniform preservation of empirical distances. Let Z' contain each retained point with weight $1/p_\ell$. For a fixed difference $v=f-g$, put $T_v=\max\{\|v\|_Z^2,nq_\epsilon^2\}$ and $Y_\ell=(\zeta_\ell/p_\ell-1)v(x_\ell)^2$. The indices with $p_\ell=1$ have $Y_\ell=0$, and the others satisfy

$$|Y_\ell|\le\frac{T_v}{C_{\mathrm{s}}\ell_M},\qquad \sum_{\ell=1}^n\mathbb{E}[Y_\ell^2]\le\frac{T_v\|v\|_Z^2}{C_{\mathrm{s}}\ell_M}\le\frac{T_v^2}{C_{\mathrm{s}}\ell_M}.$$

Bernstein's inequality and a union bound over pairs from a fixed internal ν_ϵ-net give $|\|v\|_{Z'}^2-\|v\|_Z^2|\le T_v/32$ for all net pairs, with failure probability at most $\delta_0/(64M)$. For any pair f,g, choose net

approximants $\widetilde{f}, \widetilde{g}$ and write $\widetilde{v} = \widetilde{f} - \widetilde{g}$. On the total-weight event,

$$\|v - \widetilde{v}\|_Z \le 2\nu_\epsilon\sqrt{n}, \qquad \|v - \widetilde{v}\|_{Z'} \le 2\nu_\epsilon\sqrt{2n}.$$

Applying the triangle inequality in these two seminorms and $2ab \le ta^2 + b^2/t$, with a fixed small $t > 0$, extends the net bound to

$$\frac{1}{2}\|f-g\|_Z^2 - nq_\epsilon^2 \le \|f-g\|_{Z'}^2 \le \frac{3}{2}\|f-g\|_Z^2 + nq_\epsilon^2, \qquad f, g \in \mathcal{F}. \tag{125}$$

Here $\nu_\epsilon \le c_0 q_\epsilon$, so the net-extension errors are absorbed by nq_ϵ^2 when c_0 is sufficiently small.

Rounding and confidence sets. Round each retained point in $d_{\mathcal{F}}$ and the center $\overline{f}$ in sup norm, both at resolution ν_ϵ. The total retained weight is at most $2n$, so, for $U = \|f - \overline{f}\|_Z$ and $R = \|f - f^\sharp\|_{Z^\sharp}$,

$$\left|R - \|f - \overline{f}\|_{Z'}\right| \le 3\nu_\epsilon\sqrt{2n} \le q_\epsilon\sqrt{n}.$$

Equation (125) therefore implies

$$R^2 \le 3U^2 + 4nq_\epsilon^2, \qquad U^2 \le 4R^2 + 6nq_\epsilon^2.$$

Since $\alpha \ge Mq_\epsilon^2 \ge nq_\epsilon^2$, these inequalities give

$$\mathcal{C}_Z(\overline{f}, \alpha) \subseteq \mathcal{C}_{Z^\sharp}(f^\sharp, 7\alpha) \subseteq \mathcal{C}_Z(\overline{f}, 34\alpha). \tag{126}$$

Taking pairwise widths proves (119). The preceding failure probabilities sum to at most $\delta_0/(16M)$.

Number of possible bonuses. For fixed α, a bonus constructed from $Z^\sharp$ is determined by the rounded center $f^\sharp$, at most J_M rounded state-action pairs $x_\ell^\sharp$, and their integer weights $1/p_\ell \le M$. Including the constant failure output and the empty-data width gives

$$|\mathcal{W}(\mathcal{F}, M, \delta_0; \alpha)| \le 2 + \mathcal{N}_\infty(\nu_\epsilon, \mathcal{F}) \sum_{m=0}^{J_M} \left[M\mathcal{N}(\nu_\epsilon, \mathcal{X}, d_{\mathcal{F}})\right]^m. \tag{127}$$

Lemma E.4 gives $\log \mathcal{N}_\infty(\nu_\epsilon, \mathcal{F}) = \widetilde{O}(d_{\mathrm{E}})$. Also $d_{\mathrm{E}}(\mathcal{F}, q_\epsilon) \le d_{\mathrm{E}}$, so $J_M = \widetilde{O}(d_{\mathrm{E}}^2)$. Taking logarithms in (127) proves (120). Its right-hand side is independent of the numerical value of the fixed radius. The deterministic support and total-weight checks ensure that the same family contains every output, including on the failure event. □

Using $B \vee 1$ accommodates component ranges smaller than one without changing the stated rates. A zero-range component, in particular the terminal entropy-value function, is known exactly and uses a zero bonus.

For a value function V, define the component Bellman target

$$f_{j,h}[V](s,a) \triangleq \begin{cases} r_h(s,a) + P_h V(s,a), & j = r, \\ u_{i,h}(s,a) + P_h V(s,a), & j = u_i,\ i \in [I], \\ P_h V(s,a), & j = \psi. \end{cases} \tag{128}$$

For $j = r, u_i, \psi$, respectively, let $V_j^{\pi^k}$ denote $V_r^{\pi^k}$, $V_{u_i}^{\pi^k}$, and $V_{\psi_k}^{\pi^k}$. Thus, for $f_{j,k,h}^\star \triangleq f_{j,h}[V_{j,h+1}^k]$, Assumption 5.1 gives $f_{j,k,h}^\star \in \mathcal{F}_{j,h}$, and the corresponding regression target lies in $[0, B_{j,h}]$.

We now establish uniform least-squares confidence. Uniformity over the value function is what allows the backward recursion to reuse the same trajectories at every stage.

Lemma E.6 (Uniform adaptive-value function confidence)**.** *Let $\mathcal{V}$ be a deterministic class of functions $\mathcal{S} \to [0, B]$, and suppose that for every $V \in \mathcal{V}$ the Bellman target $f_V(x) = y(x) + PV(x)$ belongs to $\mathcal{F} \subseteq (\mathcal{X} \to [0, B])$. Let $(X_t, Y_t, S_t')_{t=1}^M$ be a sequential dataset. Conditionally on the history $\mathscr{F}_{t-1}$ and the possibly adaptive X_t, assume $\mathbb{E}[Y_t \mid X_t, \mathscr{F}_{t-1}] = y(X_t)$ and $S_t' \sim P(\cdot \mid X_t)$. Assume also that $Y_t + V(S_t') \in [0, B]$ for every $V \in \mathcal{V}$. For each $m \le M$ and V, let*

$$\widehat{f}_{V,m} \in \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{m} \big(f(X_t) - Y_t - V(S_t')\big)^2.$$

Then, for every $\nu \in (0,1)$, with probability at least $1-\delta_0$, simultaneously for all $m \le M$ and $V \in \mathcal{V}$,

$$\|\widehat{f}_{V,m} - f_V\|_{Z_m}^2 \le CB^2 \log\left(\frac{M\mathcal{N}_\infty(\nu,\mathcal{F})\mathcal{N}_\infty(\nu,\mathcal{V})}{\delta_0}\right) + CMB\nu, \qquad Z_m = \{X_1,\dots,X_m\}. \tag{129}$$

The same conclusion holds conditionally when the class $\mathcal{V}$ is fixed by a sigma-field preceding the dataset.

Proof. For fixed $f \in \mathcal{F}$ and $V \in \mathcal{V}$, write $\varepsilon_t(V) = Y_t + V(S'_t) - f_V(X_t)$. Conditionally on $\mathscr{F}_{t-1}$ and X_t, this is a mean-zero random variable bounded in magnitude by B. The exponential-supermartingale bound for $\varepsilon_t(V)(f(X_t) - f_V(X_t))$, followed by a union bound over sup-norm ν-nets of $\mathcal{F}$ and $\mathcal{V}$ and over $m \le M$, gives

$$2\sum_{t=1}^{m} \varepsilon_t(V)(f(X_t) - f_V(X_t)) \le \frac{1}{2}\|f - f_V\|_{Z_m}^2 + CB^2 \log\left(\frac{M\mathcal{N}_\infty(\nu,\mathcal{F})\mathcal{N}_\infty(\nu,\mathcal{V})}{\delta_0}\right).$$

This event holds simultaneously for all net pairs. Since $\|f_V - f_{\widetilde{V}}\|_\infty \le \|V - \widetilde{V}\|_\infty$, replacing either net center by a function within distance ν changes the displayed loss comparison by at most $CMB\nu$. The resulting inequality therefore holds for every (f, V).

By realizability, f_V is an admissible competitor for the least-squares fit, so

$$\|\widehat{f}_{V,m} - f_V\|_{Z_m}^2 \le 2\sum_{t=1}^{m} \varepsilon_t(V)\big(\widehat{f}_{V,m}(X_t) - f_V(X_t)\big).$$

Substitute $f = \widehat{f}_{V,m}$ into the uniform inequality and move the half squared-norm term to the left to obtain (129). The event is uniform in V, so it also covers a value function chosen after observing the data. If $\mathcal{V}$ is measurable with respect to a sigma-field preceding the dataset, condition on that sigma-field, apply the uniform bound, and then integrate the conditional probability. □

E.3.1 Value function and actor covers

For a policy π_h, critic $f \in \mathcal{F}_{j,h}$, and stable bonus $b \in \mathcal{W}(\mathcal{F}_{j,h}, M, \delta_0; \alpha)$ for a fixed radius α, the oracle forms

$$V_{j,h}^{\pi,f,b}(s) = \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\mathrm{Truncate}_{[0,B_{j,h}]}\big(f(s,a) + b(s,a)\big), \qquad j \ne \psi,$$

$$V_{\psi,h}^{\pi,f,b}(s) = \mathrm{Ent}\big(\pi_h(\cdot \mid s)\big) + \sum_{a\in\mathcal{A}} \pi_h(a \mid s)\mathrm{Truncate}_{[0,B_{\psi,h}]}\big(f(s,a) + b(s,a)\big).$$

Here $\mathrm{Ent}(p) = -\sum_a p(a)\log p(a)$. Only the value function part of the entropy critic is fitted; its immediate entropy contribution is evaluated exactly. For later use, write

$$\mathcal{W}_{j,h}^{(M)}(\alpha) \triangleq \mathcal{W}(\mathcal{F}_{j,h}, M, \delta_0; \alpha), \qquad \mathcal{V}_{j,h}(\pi_h; M, \delta_0, \alpha) \triangleq \left\{V_{j,h}^{\pi,f,b} : f \in \mathcal{F}_{j,h},\ b \in \mathcal{W}_{j,h}^{(M)}(\alpha)\right\}. \tag{130}$$

At the terminal stage, set $\mathcal{V}_{j,H+1}(\cdot; M, \delta_0, \alpha) = \{0\}$.

Lemma E.7 (Stable value function cover)**.** *Fix $\alpha \ge \alpha_{\mathcal{F}_{j,h}}(M, \delta_0)$. For a fixed policy π_h, the class $\mathcal{V}_{j,h}(\pi_h; M, \delta_0, \alpha)$ has a ν-cover whose logarithmic size is at most*

$$\log \mathcal{N}_\infty(\nu, \mathcal{F}_{j,h}) + \log|\mathcal{W}(\mathcal{F}_{j,h}, M, \delta_0; \alpha)|. \tag{131}$$

For a varying log-linear policy, combine a $\nu/2$-cover of the critic with a cover of its logits at sup-norm resolution $\nu' = \nu/[4(B_{j,h} + \log(e|\mathcal{A}|))]$, allowing state-dependent additive shifts. The logarithmic size of this logit cover is added to (131), with the critic resolution replaced by $\nu/2$. This construction covers both ordinary and entropy value function.

Proof. Clipping and averaging under a fixed policy are nonexpansive in the sup norm. Thus a critic cover, paired with every member of the finite bonus family, proves the fixed-policy statement. For

varying policies, define $\|\pi - \pi'\|_{1,\infty} = \sup_s \sum_a |\pi(a \mid s) - \pi'(a \mid s)|$. For $Q, Q' \in [0, B_{j,h}]$,

$$\left| \sum_a \pi(a \mid s) Q(s,a) - \sum_a \pi'(a \mid s) Q'(s,a) \right| \le \|Q - Q'\|_\infty + B_{j,h} \|\pi - \pi'\|_{1,\infty}.$$

If two logits differ by at most ν' after a state-dependent shift, interpolation between them gives $\|\pi - \pi'\|_{1,\infty} \le 2\nu'$. Along this interpolation, with logit increment Δg and induced policy π_t,

$$\frac{d}{dt} \operatorname{Ent}(\pi_t) = -\operatorname{Cov}_{\pi_t}(\Delta g, \log \pi_t).$$

The derivative is bounded in magnitude by $2\|\Delta g\|_\infty \operatorname{Ent}(\pi_t) \le 2 \log |\mathcal{A}| \, \|\Delta g\|_\infty$. Integrating and adding the critic-cover error gives a total error of at most $\nu/2 + 2(B_{j,h} + \log |\mathcal{A}|)\nu' \le \nu$. Thus a logit cover controls the entropy term without requiring a lower bound on the action probabilities. □

For off-policy data, the actor depends on the preceding data prefix. Recall the centered actor function $f_{\omega,h}$, radius L_{act}, and deterministic class $\mathcal{F}_h^{\mathrm{off}}$ from (84)-(86), and define

$$\Pi_h^{\mathrm{off}} \triangleq \left\{ \pi_h(\cdot \mid \cdot, \omega) : f_{\omega,h} \in \mathcal{F}_h^{\mathrm{off}} \right\}. \tag{132}$$

Lemma E.8 (Off-policy log-linear actor cover)**.** *Run Algorithm 1 with Algorithm 4, Assumption 3.2, and $0 < \eta\tau \le 1$. For every stage h, the iterates initialized at $\omega_h^1 = 0$ satisfy $\pi_h^k \in \Pi_h^{\mathrm{off}}$ and*

$$\sup_s \max_{a,a'} \left| \langle \varphi(s,a) - \varphi(s,a'), \omega_h^k \rangle \right| \le L_{\mathrm{act}} = \frac{H(1 + I\lambda_{\max} + \tau \log |\mathcal{A}|) + \epsilon_{\mathrm{act}}}{\tau}. \tag{133}$$

Moreover, for every $\nu > 0$,

$$\log \mathcal{N}\left(\nu, \Pi_h^{\mathrm{off}}, \|\cdot\|_{1,\infty}\right) \le d_{\mathrm{a}} \log\left(1 + \frac{C L_{\mathrm{act}}}{\nu}\right). \tag{134}$$

The cover is obtained from logits modulo state-dependent shifts, so the same covering bound also supplies the logit cover required by Lemma E.7.

Proof. The actor and fitting rule are unchanged from Section 4. Hence the residual bound in (28) and the log-odds recursion (82) give (133). Equations (85)-(87) then give $\pi_h^k \in \Pi_h^{\mathrm{off}}$ and the stated cover. The centered representation removes exactly the state-dependent shifts that leave softmax policies unchanged. □

Lemma E.9 (Value-function-class complexity)**.** *At the fixed resolution $\nu_\epsilon/2$, the logarithmic covering numbers of the fixed-policy and off-policy value function classes are bounded, respectively, by*

$$\Gamma_{\mathrm{on}} = \widetilde{O}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}}), \qquad \Gamma_{\mathrm{off}} = \widetilde{O}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}}). \tag{135}$$

These bounds hold uniformly over admissible deterministic radii. Choosing the squared-loss radius $C(B_{j,h} \vee 1)^2 \Gamma_{\mathrm{on}} + CNq_\epsilon^2$ or $C(B_{j,h} \vee 1)^2 \Gamma_{\mathrm{off}} + CKq_\epsilon^2$, respectively, satisfies both the stable-bonus threshold and the uniform regression bound in (129).

Proof. For $M = N, K$, let $\overline{w}^{(M)} = \max_{j,h} \sup_{\alpha \ge \alpha_{\mathcal{F}_{j,h}}(M,\delta_0)} \log |\mathcal{W}(\mathcal{F}_{j,h}, M, \delta_0; \alpha)|$. A zero-range component contributes zero to this maximum. Define

$$\begin{aligned} \Gamma_{\mathrm{on}} &= 1 + \log \frac{N}{\delta_0} + 2 \max_{j,h} \log \mathcal{N}_\infty(\nu_\epsilon/4, \mathcal{F}_{j,h}) + \overline{w}^{(N)}, \\ \Gamma_{\mathrm{off}} &= 1 + \log \frac{K}{\delta_0} + 2 \max_{j,h} \log \mathcal{N}_\infty(\nu_\epsilon/4, \mathcal{F}_{j,h}) + \overline{w}^{(K)} \\ &\quad + d_{\mathrm{a}} \log\left(1 + \frac{C L_{\mathrm{act}}(B + \log(e|\mathcal{A}|))}{\nu_\epsilon}\right). \end{aligned} \tag{136}$$

For a fixed admissible radius and policy, Lemma E.7 gives

$$\log \mathcal{N}_\infty(\nu_\epsilon/2, \mathcal{V}_{j,h}(\pi_h; N, \delta_0, \alpha)) \le \log \mathcal{N}_\infty(\nu_\epsilon/2, \mathcal{F}_{j,h}) + \overline{w}^{(N)} \le \Gamma_{\mathrm{on}}.$$

For the union over $\pi_h \in \Pi_h^{\mathrm{off}}$, use critic resolution $\nu_\epsilon/4$ and logit resolution $\nu_\epsilon/[8(B_{j,h} + \log(e|\mathcal{A}|))]$ in that lemma. Equation (134) bounds the added logit-cover term by the last term of Γ_{off}. The separate maxima over (j, h) in (136) also cover a stage-h critic paired with a stage-$(h+1)$ value function.

Apply Lemma E.6 at resolution $\nu = \nu_\epsilon/2$. Its logarithmic term is bounded by the corresponding Γ, and the definition of ν_ϵ gives

$$CMB_{j,h}\nu_\epsilon \le CMB\nu_\epsilon \le CMq_\epsilon^2.$$

The additive radius CMq_ϵ^2 therefore covers both this approximation term and $\alpha_{\mathcal{F}_{j,h}}(M, \delta_0) = Mq_\epsilon^2$. For a zero-range component, the target, fit, and bonus are identically zero and no confidence radius is needed. Finally, Lemma E.4, at resolution $\nu_\epsilon/4$, bounds the critic entropy by $\widetilde{O}(d_{\mathrm{E}})$; Lemma E.5 gives $\overline{w}^{(M)} = \widetilde{O}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}})$. The actor term contributes $\widetilde{O}(d_{\mathrm{a}})$. This proves (135); the budget enters only through logarithms, not through the scales at which the complexity measures are evaluated. □

Lemma E.10 (Common boundedness certificate). *Let $S \triangleq 1 + I\lambda_{\max} + \tau \log(e|\mathcal{A}|)$. If $0 < \eta\tau \le 1/4$ and $\eta HS \le 1/4$, every output of Algorithm 4 satisfies*

$$0 \le V_{u_i}^k \le H, \qquad \sum_{a \in \mathcal{A}} \pi_h^k(a \mid s) e^{\eta Q^k_{z_k,h}(s,a)} \left(Q^k_{z_k,h}(s,a)\right)^2 \le CH^2S^2. \tag{137}$$

Proof. Utility boundedness follows from truncation. For $p_a = \pi_h^k(a \mid s)$, the component ranges give

$$Q^k_{z_k,h}(s,a) = q_{k,h}(s,a) + \tau \log(1/p_a), \qquad 0 \le q_{k,h}(s,a) \le HS.$$

Set $\alpha = \eta\tau$ and $L = \log(e|\mathcal{A}|)$. Then

$$\sum_a p_a e^{\eta Q^k_{z_k,h}(s,a)} \left(Q^k_{z_k,h}(s,a)\right)^2 \le 2e^{\eta HS} \left[H^2S^2 \sum_a p_a^{1-\alpha} + \tau^2 \sum_a p_a^{1-\alpha} \log^2(1/p_a) \right].$$

Concavity gives $\sum_a p_a^{1-\alpha} \le |\mathcal{A}|^\alpha$. Moreover, $x^2 \le CL^2 e^{x/(2L)}$ for $x \ge 0$ implies

$$\sum_a p_a^{1-\alpha} \log^2(1/p_a) \le CL^2 |\mathcal{A}|^{\alpha + 1/(2L)}.$$

The two stepsize conditions give $\eta HS \le 1/4$ and $\alpha \log|\mathcal{A}| \le 1/4$, while $|\mathcal{A}|^{1/(2L)} \le e^{1/2}$ and $\tau L \le S$. Substitution proves (137). □

E.4 Eluder-width bounds

We now bound the size of the bonus under the current policy, which determines the cost of optimism in the oracle guarantee. For a fresh on-policy batch, we control $\mathbb{E}_{x \sim d_h^{\pi^k}}[b_{j,h}^k(x)]$ using an eluder-width bound. With off-policy data, the bonus depends on the newly observed trajectory; we upper-bound it by a width determined from the preceding data Z_{k-1}, then control the sum of its conditional expectations.

Lemma E.11 (Population eluder width). *Let $\mathcal{F} \subseteq (\mathcal{X} \to [0, B])$, let $X_1, \dots, X_N$ be i.i.d. from ν, fix $\gamma \ge 0$, and set*

$$w_N(x) = \sup \left\{ |f(x) - f'(x)| : f, f' \in \mathcal{F}, \ \sum_{n=1}^N (f(X_n) - f'(X_n))^2 \le \gamma \right\}.$$

For every fixed $\rho > 0$, with probability at least $1 - \delta_0$,

$$\mathbb{E}_{X\sim\nu}[w_N(X)] \le \rho + \frac{Bd}{N} + 4\sqrt{\frac{\gamma d}{N}} + B\sqrt{\frac{2\log(1/\delta_0)}{N}}, \qquad d = d_{\mathrm{E}}(\mathcal{F}, \rho). \tag{138}$$

Proof. Set $Z_{n-1} = \{X_1, \dots, X_{n-1}\}$ and define w_{n-1} by the same pairwise constraint on Z_{n-1}. We adapt the counting argument of Russo and Van Roy (2013, Proposition 3 and Lemma 2) to the fixed cutoff ρ. For $t \ge \rho$, greedily partition the points with $w_{n-1}(X_n) > t$ into t-independent subsequences. At each such point there is a pair (f, f') with $|f(X_n) - f'(X_n)| > t$ and $\|f - f'\|^2_{Z_{n-1}} \le \gamma$. If this point were t-dependent on m disjoint preceding subsequences, the same pair would satisfy

$$\gamma \ge \|f - f'\|^2_{Z_{n-1}} > mt^2.$$

Thus $\lfloor \gamma/t^2 \rfloor + 1$ subsequences suffice, and each contains at most $d_{\mathrm{E}}(\mathcal{F}, t) \le d$ points. Consequently,

$$\sum_{n=1}^{N} \mathbf{1}\{w_{n-1}(X_n) > t\} \le d\left(1 + \frac{\gamma}{t^2}\right), \qquad t \ge \rho. \tag{139}$$

If $\rho \ge B$, the desired width bound is immediate. Otherwise, integrating the count gives

$$\begin{aligned}\sum_{n=1}^{N} w_{n-1}(X_n) &\le N\rho + \int_{\rho}^{B} \min\left\{N, d + \frac{\gamma d}{t^2}\right\} dt \\ &\le N\rho + Bd + \int_{0}^{\infty} \min\left\{N, \frac{\gamma d}{t^2}\right\} dt \le N\rho + Bd + 2\sqrt{\gamma d N}.\end{aligned}$$

Adding observations only shrinks the feasible pair set, so $w_N \le w_{n-1}$ pointwise. Independence therefore gives, for every realized batch,

$$N\mathbb{E}_{X\sim\nu}[w_N(X)] \le \sum_{n=1}^{N} \mathbb{E}[w_{n-1}(X_n) \mid X_1, \dots, X_{n-1}].$$

Each conditional expectation differs from $w_{n-1}(X_n)$ by a martingale difference bounded in magnitude by B. Azuma-Hoeffding and the preceding width sum prove (138). □

Lemma E.12 (Sequential predictable width)**.** *Let $\mathcal{F} \subseteq (\mathcal{X} \to [0, B])$, let X_k be generated sequentially, and let $\mathscr{F}_{k-1}$ contain the history before X_k is drawn. Suppose b_k is $\mathscr{F}_{k-1}$-measurable, $0 \le b_k \le B$, and*

$$b_k(x) \le \sup\left\{|f(x) - f'(x)| : f, f' \in \mathcal{F}, \quad \sum_{\ell<k}(f(X_\ell) - f'(X_\ell))^2 \le \gamma\right\}.$$

For every fixed $\rho > 0$, with probability at least $1 - \delta_0$,

$$\sum_{k=1}^{K-1} \mathbb{E}[b_k(X_k) \mid \mathscr{F}_{k-1}] \le C\left[K\rho + Bd + \sqrt{\gamma d K} + B\sqrt{K\log(1/\delta_0)}\right], \quad d = d_{\mathrm{E}}(\mathcal{F}, \rho). \tag{140}$$

Proof. The deterministic count (139) applies to the pairwise widths dominating b_k, regardless of how the points X_k are selected. Integrating it as in the preceding proof gives

$$\sum_{k=1}^{K-1} b_k(X_k) \le K\rho + Bd + 2\sqrt{\gamma d K}.$$

By predictability, $\mathbb{E}[b_k(X_k) \mid \mathscr{F}_{k-1}] - b_k(X_k)$ is a martingale difference bounded by B. Azuma-Hoeffding proves (140). This conditional-expectation bound is the form needed by the off-policy oracle. □

Lemma E.13 (Predictable envelope for a post-data width)**.** *Let $Z_k = Z_{k-1} \cup \{X_k\}$ be multisets in $\mathcal{X}$, let $\overline{f}_k \in \mathcal{F}$, and let $\gamma \geq 0$. Then, for every $x \in \mathcal{X}$,*

$$w\big(\mathcal{C}_{Z_k}(\overline{f}_k, \gamma), x\big) \leq \sup_{\substack{f, f' \in \mathcal{F} \\ \|f - f'\|_{Z_{k-1}}^2 \leq 4\gamma}} |f(x) - f'(x)|. \tag{141}$$

In particular, if Z_{k-1} is known before X_k is drawn and γ is deterministic, the right-hand side is predictable even when $\overline{f}_k$ depends on X_k.

Proof. For any $f, f' \in \mathcal{C}_{Z_k}(\overline{f}_k, \gamma)$, the triangle inequality for the empirical seminorm gives

$$\|f - f'\|_{Z_k} \leq \|f - \overline{f}_k\|_{Z_k} + \|f' - \overline{f}_k\|_{Z_k} \leq 2\sqrt{\gamma}.$$

Restricting the empirical norm from Z_k to Z_{k-1} can only decrease it, so the pair (f, f') is feasible on the right-hand side of (141). Taking the supremum proves the claim. □

E.5 On-policy oracle certificate and tuning

Throughout these certificates, take $\lambda_{\max} \geq 1$, as satisfied by the prescribed choice $2H \log(e|\mathcal{A}|)/\xi$. Write $S = 1 + I\lambda_{\max} + \tau \log(e|\mathcal{A}|)$ and use the fixed scales $r_\epsilon, q_\epsilon, \nu_\epsilon$ of (114).

Proposition E.14 (Non-split general on-policy oracle)**.** *Under Assumption 5.1, choose the radii and stable-bonus budgets as below. If $N \geq d_0 d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}}$, $0 < \tau \leq 1$, $0 < \eta\tau \leq 1/4$, and $\eta H(1 + I\lambda_{\max} + \tau \log(e|\mathcal{A}|)) \leq 1/4$, then, with probability at least $1 - \delta$, Algorithm 4 satisfies Assumption 3.1 and*

$$\begin{aligned} \sup_{k<K} \beta_k &\leq \widetilde{O}\left(SH^2 \sqrt{\frac{d_{\mathrm{E}}^3 \mathfrak{h}_{\mathcal{X}}}{N}} \right) + CSHr_\epsilon, \\ \sup_{k<K} C_{\eta,\tau,\Lambda,k} &= \widetilde{O}\left(H^2 \big(1 + I\lambda_{\max} + \tau \log|\mathcal{A}|\big)^2 \right). \end{aligned} \tag{142}$$

Here and below, $\widetilde{O}$ hides universal powers of the confidence, contraction, and cover-scale logarithms. In particular, all polynomial dependence on d_{E}, $\mathfrak{h}_{\mathcal{X}}$, and d_{a} is displayed. Fix $\epsilon, \delta \in (0, 1)$ and choose

$$\begin{aligned} \lambda_{\max} &= \frac{2H \log(e|\mathcal{A}|)}{\xi}, \\ \tau &= \Theta\left(\min\left\{ 1, \max\left\{ \frac{\xi\epsilon}{H \log(e|\mathcal{A}|)}, \frac{H^{2/3}\xi^{2/3}\epsilon_{\mathrm{act}}^{1/3}}{\log^{2/3}(e|\mathcal{A}|)} \right\} \right\} \right), \\ \eta &= \widetilde{\Theta}\left(\frac{\tau\xi^2\epsilon^2}{I^2H^8} \right), \qquad K = \widetilde{\Theta}\left(\frac{I^2H^8}{\tau^2\xi^2\epsilon^2} \right), \\ N &= \widetilde{\Theta}\left(\frac{d_{\mathrm{E}}^3 \mathfrak{h}_{\mathcal{X}} I^4 H^{14}}{\tau^2\xi^4\epsilon^4} \right). \end{aligned} \tag{143}$$

These choices yield the on-policy part of Theorem 5.2.

Proof of Proposition E.14. We verify optimism, consistency, and boundedness in that order. Allocate $\delta_0 = \delta/[CKH(I+2)]$ to each component, stage, and iteration, set $\delta_{j,h} = \delta_0$, take $M = N$, and choose

$$\alpha_{j,h} = C\left[(B_{j,h} \vee 1)^2 \Gamma_{\mathrm{on}} + N q_\epsilon^2 \right] \qquad B_{j,h} > 0, \tag{144}$$

where Γ_{on} is the value-function-cover complexity in Lemma E.9. A sufficiently large universal C makes these radii satisfy both the stable-bonus and regression-confidence requirements. For $B_{j,h} = 0$, set $\alpha_{j,h} = b_{j,h}^k = 0$.

For $h \in [H]$, define the conditional value function class

$$\mathcal{V}_{k,j,h}^{\mathrm{on}} \triangleq \mathcal{V}_{j,h}(\pi_h^k; N, \delta_0, \alpha_{j,h}), \qquad \mathcal{V}_{k,j,H+1}^{\mathrm{on}} \triangleq \{0\}.$$

By construction of Algorithm 4, $V^k_{j,h} \in \mathcal{V}^{\mathrm{on}}_{k,j,h}$ for every $h \in [H+1]$.

Fix k and condition on the history before its batch is collected. The policy π^k is fixed and the N trajectories are independent, but the fitted value function $V^k_{j,h+1}$ still depends on this batch. Let $Z^k_h \triangleq \{(s^{k,n}_h, a^{k,n}_h) : n \in [N]\}$. Applying Lemma E.6 conditionally to $\mathcal{V}^{\mathrm{on}}_{k,j,h+1}$ at resolution $\nu_\epsilon/2$ and using Lemma E.9, backward induction gives, simultaneously for every component and stage,

$$\left\| \widehat{f}^k_{j,h} - f^\star_{j,k,h} \right\|^2_{Z^k_h} \le \alpha_{j,h}. \tag{145}$$

The uniform value function cover is what permits every stage to use the same batch without sample splitting.

On the joint confidence and stable-bonus event, both $\widehat{f}^k_{j,h}$ and $f^\star_{j,k,h}$ belong to the inner full-data confidence ball. The lower width sandwich gives $|\widehat{f}^k_{j,h} - f^\star_{j,k,h}| \le b^k_{j,h}$ pointwise. Truncation preserves this optimism because the target lies in the truncation interval:

$$0 \le \overline{Q}^k_{j,h}(s,a) - f^\star_{j,k,h}(s,a) \le 2b^k_{j,h}(s,a). \tag{146}$$

For the entropy component, the known immediate cost $\psi_{k,h}$ is included exactly. Scalarizing with the nonnegative weights $(1, \lambda_k, \tau)$ therefore proves Assumption 3.1(i).

For consistency, the upper width sandwich and the empirical-norm triangle inequality give

$$b^k_{j,h}(x) \le \sup_{\substack{f,f' \in \mathcal{F}_{j,h} \\ \|f-f'\|^2_{Z^k_h} \le 136\alpha_{j,h}}} |f(x) - f'(x)|. \tag{147}$$

Applying Lemma E.11 at cutoff r_ϵ conditionally under the occupancy $d^{\pi^k}_h$ yields, for reward and utility components,

$$\mathbb{E}_{d^{\pi^k}_h}[b^k_{j,h}] \le CH \left[\sqrt{\frac{d_0 \Gamma_{\mathrm{on}}}{N}} + \frac{d_0}{N} + \sqrt{\frac{\log(1/\delta_0)}{N}} \right] + Cr_\epsilon.$$

The additional radius contributes $Cq_\epsilon\sqrt{d_0} = Cr_\epsilon$; also $d_0 \le d_{\mathrm{E}}$. The entropy-value function bound can be enlarged by $\log(e|\mathcal{A}|)$; at $h = H$ this value function and its bonus are zero. Telescoping the Bellman residuals in (146) now gives

$$0 \le V^k_{j,1}(s_1) - V^{\pi^k}_j \le 2 \sum_{h=1}^{H} \mathbb{E}_{d^{\pi^k}_h}[b^k_{j,h}],$$

where $V^{\pi^k}_j$ denotes $V^{\pi^k}_{\psi_k}$ for $j = \psi$. Consequently, a common tolerance for both consistency requirements is

$$\beta_k = 2 \sum_{h=1}^{H} \left[\mathbb{E}_{d^{\pi^k}_h} b^k_{r,h} + \max_i \mathbb{E}_{d^{\pi^k}_h} b^k_{u_i,h} + \sum_{i=1}^{I} \lambda_{k,i} \mathbb{E}_{d^{\pi^k}_h} b^k_{u_i,h} + \tau \mathbb{E}_{d^{\pi^k}_h} b^k_{\psi,h} \right].$$

Since (135) gives $\Gamma_{\mathrm{on}} = \widetilde{O}(d^2_{\mathrm{E}} \mathfrak{h}_{\mathcal{X}})$ and $N \ge d_0 d^2_{\mathrm{E}} \mathfrak{h}_{\mathcal{X}}$,

$$\beta_k \le \widetilde{O}\left(SH^2 \sqrt{\frac{d_0 d^2_{\mathrm{E}} \mathfrak{h}_{\mathcal{X}}}{N}} \right) + CSHr_\epsilon.$$

Finally, Lemma E.10 gives the utility and local-moment bounds. Since $d_0 \le d_{\mathrm{E}}$ and $0 < \tau \le 1$, $1 + I\lambda_{\max} + \tau \log(e|\mathcal{A}|) \le 2(1 + I\lambda_{\max} + \tau \log|\mathcal{A}|)$, so this bound has the form stated in (142). A union bound over the allocated events completes the oracle certificate. □

E.6 Off-policy oracle certificate and tuning

Proposition E.15 (Non-split general off-policy oracle). *Under Assumptions 3.2 and 5.1, choose the radii and stable-bonus budgets as below. If* $0 < \tau \le 1$, $0 < \eta\tau \le 1/4$, *and* $\eta H(1 + I\lambda_{\max} +$

$\tau \log(e|\mathcal{A}|)) \le 1/4$, *then, with probability at least* $1-\delta$, *Algorithm 4 on the cumulative off-policy dataset* $\mathcal{D}^k$ *satisfies Assumption 3.1 and*

$$\sum_{k=1}^{K-1} \beta_k \le \widetilde{O}\left(SH^2\left[\sqrt{d_{\mathrm{E}}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})K} + d_{\mathrm{E}}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})\right]\right) + CSHKr_\epsilon,$$
$$\sup_{k<K} C_{\eta,\tau,\Lambda,k} = \widetilde{O}\left(H^2\left(1 + I\lambda_{\max} + \tau \log|\mathcal{A}|\right)^2\right). \tag{148}$$

Choose $\lambda_{\max}$ and τ as in (143), and set

$$\eta = \widetilde{\Theta}\left(\frac{\tau \xi^4 \epsilon^4}{d_{\mathrm{E}}(d_{\mathrm{E}}^2 \mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}}) I^4 H^{14}}\right), \qquad K = \widetilde{\Theta}\left(\frac{1}{\eta\tau}\right). \tag{149}$$

These choices yield the off-policy part of Theorem 5.2. The proof below dominates the additive complexity term in (148) and bounds the contribution of Kr_ϵ separately.

Proof of Proposition E.15. The confidence argument is uniform over both value function and the off-policy actor class; the width argument must additionally account for the current sample in the fitted bonus. Take $M = K$, allocate $\delta_0 = \delta/[CKH(I+2)]$, set $\delta_{j,h} = \delta_0$, and choose

$$\alpha_{j,h} = C\left[(B_{j,h} \vee 1)^2 \Gamma_{\mathrm{off}} + Kq_\epsilon^2\right] \qquad B_{j,h} > 0,$$

where Γ_{off} is defined in Lemma E.9; set $\alpha_{j,h} = b_{j,h}^k = 0$ when $B_{j,h} = 0$. For $h \in [H]$, define the deterministic class

$$\mathcal{V}_{j,h}^{\mathrm{off}} \triangleq \bigcup_{\pi_h \in \Pi_h^{\mathrm{off}}} \mathcal{V}_{j,h}(\pi_h; K, \delta_0, \alpha_{j,h}), \qquad \mathcal{V}_{j,H+1}^{\mathrm{off}} \triangleq \{0\}.$$

Algorithm 4 and Lemma E.8 give $V_{j,h}^k \in \mathcal{V}_{j,h}^{\mathrm{off}}$ for every $h \in [H+1]$. For $Z_h^k \triangleq \{(s_h^\ell, a_h^\ell) : \ell \in [k]\}$, applying Lemma E.6 at resolution $\nu_\epsilon/2$ over this deterministic class and all dataset prefixes gives, simultaneously for every (k, h, j),

$$\|\widehat{f}_{j,h}^k - f_{j,k,h}^\star\|_{Z_h^k}^2 \le \alpha_{j,h}. \tag{150}$$

Uniformity permits substituting the value function fitted on $\mathcal{D}^k$, including the current trajectory; the actor π^k depends only on $\mathcal{D}^{k-1}$. On the joint confidence and stable-bonus event, the lower width sandwich and truncation give

$$0 \le \overline{Q}_{j,h}^k - f_{j,k,h}^\star \le 2b_{j,h}^k.$$

After adding the immediate entropy term exactly, scalarization proves optimism. Bellman telescoping also gives

$$0 \le V_{j,1}^k(s_1) - V_j^{\pi^k} \le 2\sum_{h=1}^H \mathbb{E}_{d_h^{\pi^k}}[b_{j,h}^k].$$

Accordingly, use the consistency certificate

$$\beta_k = 2\sum_{h=1}^H \left[\mathbb{E}_{d_h^{\pi^k}} b_{r,h}^k + \max_i \mathbb{E}_{d_h^{\pi^k}} b_{u_i,h}^k + \sum_{i=1}^I \lambda_{k,i} \mathbb{E}_{d_h^{\pi^k}} b_{u_i,h}^k + \tau \mathbb{E}_{d_h^{\pi^k}} b_{\psi,h}^k\right]. \tag{151}$$

Let $\mathscr{F}_{k-1}$ contain the history before trajectory k is collected. The stable-width sandwich gives

$$b_{j,h}^k(x) \le w\big(\mathcal{C}_{Z_h^k}(\widehat{f}_{j,h}^k, 34\alpha_{j,h}), x\big).$$

Removing the current sample from the empirical norm and applying Lemma E.13 yields the predictable envelope

$$b_{j,h}^k(x) \le \widetilde{b}_{j,h}^k(x) \triangleq \sup_{\substack{f,f' \in \mathcal{F}_{j,h} \\ \|f-f'\|_{Z_h^{k-1}}^2 \le 136\alpha_{j,h}}} |f(x) - f'(x)|. \tag{152}$$

Both $\widetilde{b}_{j,h}^k$ and π^k are $\mathscr{F}_{k-1}$-measurable. Consequently,

$$\mathbb{E}_{d_h^{\pi^k}}[b_{j,h}^k] \le \mathbb{E}_{d_h^{\pi^k}}[\widetilde{b}_{j,h}^k] = \mathbb{E}\left[\widetilde{b}_{j,h}^k(s_h^k, a_h^k) \mid \mathscr{F}_{k-1}\right].$$

The first inequality holds for each realized fitted bonus. Only the predictable envelope is evaluated on the actual current trajectory in the last equality, which avoids conditioning incorrectly on a bonus fitted using that trajectory.

The envelope radius is $O(H^2\Gamma_{\mathrm{off}} + Kq_\epsilon^2)$ for reward and utility components. Lemma E.12 at cutoff r_ϵ, summed over stages, therefore gives

$$\sum_{k=1}^{K-1}\sum_{h=1}^{H} \mathbb{E}_{d_h^{\pi^k}}[b_{j,h}^k] \le \widetilde{O}\left(H^2\left[\sqrt{d_0\Gamma_{\mathrm{off}}K} + d_0\right]\right) + CHKr_\epsilon$$

for reward and utility components, with one additional factor $\log(e|\mathcal{A}|)$ for entropy. Here the additional radius contributes $CHKq_\epsilon\sqrt{d_0} = CHKr_\epsilon$. The additive complexity term retains the initial unresolved widths; it will be dominated only after choosing K. Substituting these bounds into (151), and using $\Gamma_{\mathrm{off}} = \widetilde{O}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})$ from (135) proves the first line of (148). Lemma E.10, together with $0 < \tau \le 1$, establishes boundedness in the stated form. A union bound over the confidence, stable-bonus, and sequential-width events completes the proof. □

E.7 Proofs of the full last-iterate guarantees

We make the logarithmic tuning explicit to avoid a sample-budget fixed point. After choosing τ and the scales in (114), set

$$\iota_\epsilon = \log\left(e + \frac{|\mathcal{A}|H(I+2)(1+d_{\mathrm{E}}+\mathfrak{h}_{\mathcal{X}}+d_{\mathrm{a}})(1+L_{\mathrm{act}})}{\delta\xi\tau\epsilon\nu_\epsilon}\right).$$

All quantities in this logarithm are independent of N and K. For the choices below, $\log(NK) = O(\iota_\epsilon)$ and $\log(1/\delta_0) = O(\iota_\epsilon)$, with constants depending only on the universal integer p. The explicit covering and confidence bounds above involve only fixed powers of these logarithms. Choose a universal integer p larger than their total degree plus six. Then powers of ι_ϵ suffice for all suppressed logarithmic factors, without evaluating a complexity measure at a budget-dependent scale.

Proof of the on-policy part of Theorem 5.2. We first reduce the potential to its actor-approximation floor, then choose the temperature and count trajectories. Uniform initialization, $\lambda_1 = 0$, and (35) give

$$\Phi_1 \le H\log|\mathcal{A}| + \frac{H^2(1+\log|\mathcal{A}|)^2}{2\xi^2}.$$

The geometric weights in Theorem 3.4 sum to at most $1/(\eta\tau)$. Substituting Proposition E.14 and $\lambda_{\max} = 2H\log(e|\mathcal{A}|)/\xi$ therefore yields

$$\begin{aligned}\Phi_K \le{}& e^{-\eta\tau(K-1)}\Phi_1 + \frac{H\epsilon_{\mathrm{act}}}{\tau} \\ &+ \widetilde{O}\left(\frac{\eta I^2H^5}{\tau\xi^2} + \frac{I^2H^4}{\tau\xi^2}\sqrt{\frac{d_{\mathrm{E}}^3\mathfrak{h}_{\mathcal{X}}}{N}}\right) \\ &+ C\frac{I^2H^3L^2r_\epsilon}{\tau\xi^2}.\end{aligned} \tag{153}$$

The two terms in $\widetilde{O}$ come from the local second moment (including G_τ^2) and the statistical critic error. An explicit realization of (143) is

$$\eta = \frac{c\tau\xi^2\epsilon^2}{I^2H^8\iota_\epsilon^p}, \qquad K = 1 + \left\lceil \frac{CL_\epsilon}{\eta\tau} \right\rceil,$$
$$N = \left\lceil \frac{Cd_{\mathrm{E}}^3\mathfrak{h}_{\mathcal{X}}I^4H^{14}\iota_\epsilon^p}{\tau^2\xi^4\epsilon^4} \right\rceil.$$

Here c is sufficiently small and C sufficiently large. These choices satisfy the stepsize conditions and make both statistical terms $O(\epsilon^2/H^3)$. Since $L_\epsilon \geq \log(1 + H^3\Phi_1/\epsilon^2)$, they also control the initial term. The retained approximation error satisfies

$$C\frac{I^2H^3L^2r_\epsilon}{\tau\xi^2} \leq Cc_0\frac{\epsilon^2}{H^3L_\epsilon} \leq Cc_0\frac{\epsilon^2}{H^3}$$

by the choice of r_ϵ. Consequently,

$$\Phi_K \leq \frac{C\epsilon^2}{H^3} + \frac{H\epsilon_{\mathrm{act}}}{\tau}. \tag{154}$$

Since the right-hand sides of the conversion bounds in Theorem 3.4 are nonnegative, applying those bounds and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ gives, in the normalized Slater regime specified above,

$$\max\left\{[V_r^{\pi^\star} - V_r^{\pi^K}]_+, \max_{i\in[I]}[c_i - V_{u_i}^{\pi^K}]_+\right\} \leq C\left[\epsilon + \frac{H\log(e|\mathcal{A}|)}{\xi}\tau + H^2\epsilon_{\mathrm{act}}^{1/2}\tau^{-1/2}\right].$$

The regularisation term is $O(\epsilon)$ at temperature $\xi\epsilon/[H\log(e|\mathcal{A}|)]$. Balancing it with the actor term instead gives temperature $H^{2/3}\xi^{2/3}\epsilon_{\mathrm{act}}^{1/3}/\log^{2/3}(e|\mathcal{A}|)$. Taking the larger of these two temperatures, capped at one, gives (143). When the cap is inactive, the resulting bound is (19). In particular, the cap is inactive throughout the nontrivial regime $\epsilon_{\mathrm{act}} \leq \xi/H^2$ specified above. Outside this regime the stated approximation floor is already at least H, the largest possible reward or constraint gap.

Each oracle call uses N trajectories shared by all stages. Multiplying the batch size by the number of calls gives

$$NK = \widetilde{O}\left(\frac{d_{\mathrm{E}}^3\mathfrak{h}_{\mathcal{X}}I^6H^{22}}{\xi^6\epsilon^6\tau^4}\right).$$

To keep the temperature cap explicit, its reciprocal satisfies

$$\frac{1}{\tau^4} = \widetilde{O}\left(\max\left\{1, \min\left\{\frac{H^4}{\xi^4\epsilon^4}, \frac{1}{H^{8/3}\xi^{8/3}\epsilon_{\mathrm{act}}^{4/3}}\right\}\right\}\right),$$

where the second entry of the minimum is interpreted as $+\infty$ when $\epsilon_{\mathrm{act}} = 0$. Substitution into the preceding display yields (20). When the cap is inactive, the outer maximum can be removed. Each trajectory contains H transitions, so the transition count is HNK. □

Proof of the off-policy part of Theorem 5.2. Use $\lambda_{\max} = 2H\log(e|\mathcal{A}|)/\xi$. Uniform policy initialization, $\lambda_1 = 0$, and (35) give

$$\Phi_1 \leq H\log|\mathcal{A}| + \frac{H^2(1+\log|\mathcal{A}|)^2}{2\xi^2}.$$

For the critic term, bound the geometric weights by one and apply the cumulative certificate of Proposition E.15. For the second-moment and actor terms, retain the geometric sum $1/(\eta\tau)$. The

resulting master bound is

$$\begin{aligned}\Phi_K \le e^{-\eta\tau(K-1)}\Phi_1 + \frac{H\epsilon_{\mathrm{act}}}{\tau} + \widetilde{O}\left(\frac{\eta I^2 H^5}{\tau\xi^2}\right) \\ + \widetilde{O}\left(\frac{\eta I^2 H^4}{\xi^2}\sqrt{d_{\mathrm{E}}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})K}\right) \\ + \widetilde{O}\left(\frac{\eta I^2 H^4}{\xi^2} d_{\mathrm{E}}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})\right) \\ + C\frac{\eta K I^2 H^3 L^2}{\xi^2} r_\epsilon.\end{aligned} \tag{155}$$

An explicit realization of (149) is

$$\eta = \frac{c\tau\xi^4\epsilon^4}{d_{\mathrm{E}}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})I^4 H^{14}\iota_\epsilon^p}, \qquad K = 1 + \left\lceil \frac{CL_\epsilon}{\eta\tau} \right\rceil.$$

These choices ensure $K \ge d_{\mathrm{E}}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})$, so the additive width term is dominated by the square-root term. Substituting $K = \widetilde{O}(1/(\eta\tau))$, the latter contributes

$$\widetilde{O}\left(\frac{I^2 H^4}{\xi^2}\sqrt{\frac{\eta d_{\mathrm{E}}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})}{\tau}}\right) = O(\epsilon^2/H^3),$$

with the logarithmic factors included in the stepsize choice. The second-moment term is also $O(\epsilon^2/H^3)$ under this stepsize. The last term in (155) satisfies

$$C\frac{\eta K I^2 H^3 L^2}{\xi^2} r_\epsilon \le C'\frac{L_\epsilon I^2 H^3 L^2}{\tau\xi^2} r_\epsilon \le C' c_0 \frac{\epsilon^2}{H^3}.$$

Thus retaining the compression and discretization error does not change the required potential accuracy, and

$$\Phi_K \le \frac{C\epsilon^2}{H^3} + \frac{H\epsilon_{\mathrm{act}}}{\tau}.$$

Applying (11) and $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ gives

$$\max\left\{[V_r^{\pi^\star} - V_r^{\pi^K}]_+, \max_{i\in[I]}[c_i - V_{u_i}^{\pi^K}]_+\right\} \le C\left[\epsilon + \frac{H\log(e|\mathcal{A}|)}{\xi}\tau + H^2\sqrt{\frac{\epsilon_{\mathrm{act}}}{\tau}}\right].$$

Balancing the last two terms and imposing the target-accuracy scale gives the temperature in (143), hence (19).

The off-policy branch collects one trajectory per iteration and, before substituting τ,

$$K = \widetilde{O}\left(\frac{d_{\mathrm{E}}(d_{\mathrm{E}}^2\mathfrak{h}_{\mathcal{X}} + d_{\mathrm{a}})I^4 H^{14}}{\xi^4\epsilon^4\tau^2}\right).$$

Keeping the temperature cap gives

$$\frac{1}{\tau^2} = \widetilde{O}\left(\max\left\{1, \min\left\{\frac{H^2}{\xi^2\epsilon^2}, \frac{1}{H^{4/3}\xi^{4/3}\epsilon_{\mathrm{act}}^{2/3}}\right\}\right\}\right)$$

where the second entry of the minimum is interpreted as $+\infty$ when $\epsilon_{\mathrm{act}} = 0$. Substitution proves (21); when the cap is inactive, the outer maximum can be removed. The transition count is HK. □

Remark E.16 (Relation to the linear oracle)**.** Both oracle constructions fit the components backward, add optimistic bonuses, include $\psi_{k,h}$ exactly, and scalarize after fitting. They also share the same sampling modes: a fresh batch or the cumulative off-policy dataset. The distinction lies in the bonus geometry. The linear oracle uses a covariance matrix and an elliptical-potential argument (Jin et al., 2020); the general oracle uses sensitivity sampling and rounding to obtain a controlled family of widths, followed by eluder counting (Wang et al., 2020).

This general construction sacrifices some dimension dependence. For bounded linear critics, $d_{\mathrm{E}}, \mathfrak{h}_{\mathcal{X}} = \widetilde{O}(d_{\mathrm{c}})$, so the on-policy complexity factor $d_{\mathrm{E}}^3 \mathfrak{h}_{\mathcal{X}}$ becomes $\widetilde{O}(d_{\mathrm{c}}^4)$, compared with $\widetilde{O}(d_{\mathrm{c}}^3)$ for the tailored linear analysis. The analogous off-policy factors are $\widetilde{O}(d_{\mathrm{c}}^4 + d_{\mathrm{c}} d_{\mathrm{a}})$ and $\widetilde{O}(d_{\mathrm{c}}^3 + d_{\mathrm{c}} d_{\mathrm{a}})$, respectively. The same loss under linear specialization is discussed by Wang et al. (2020, Theorem 1 and Remarks 2-3). Only the bonus representation is compressed; every regression continues to use all supplied data.

F Additional Experimental Details

The synthetic experiment evaluates the online linear CMDP algorithm under both fresh on-policy and cumulative off-policy optimistic evaluation.

Linear CMDP. We use a horizon-4 CMDP with one utility constraint, six actions, and $d_{\mathrm{c}} = d_{\mathrm{a}} = 6$. There are ten nonterminal states: one initial state and three observable clones of each of three outcome groups $\{\mathrm{bad}, \mathrm{neutral}, \mathrm{good}\}$; an additional terminal state is used only for bookkeeping. The known critic and actor features coincide, $\phi = \varphi$, but the transition, reward, and utility parameters are hidden from the learner.

For completeness, let $\mathbf{e}_j$ be the jth coordinate vector, let $g(s) \in \{0, 1, 2\}$ index the group of a noninitial state, and let $\oplus$ denote cyclic addition over the six feature coordinates. With $(\rho, \gamma) = (6.5, 1.5) \times 10^{-1}$, the dense simplex features are

$$\phi(s_1, a) = \frac{1-\rho}{6}\mathbf{1} + \rho \mathbf{e}_a, \qquad \phi(s, a) = \frac{1-\rho-\gamma}{6}\mathbf{1} + \rho \mathbf{e}_{a \oplus 2g(s)} + \gamma \mathbf{e}_{2g(s)+1}.$$

The six transition anchors distribute the next state over the three groups according to

$$M = 10^{-1} \begin{pmatrix} 9.0 & 0.8 & 0.2 \\ 0.2 & 0.8 & 9.0 \\ 1.2 & 8.0 & 0.8 \\ 0.8 & 2.5 & 6.7 \\ 6.7 & 2.5 & 0.8 \\ 1.8 & 2.0 & 6.2 \end{pmatrix}.$$

We use M and its two cyclic column shifts at the first three stages and sample uniformly among the three clones of the selected group; the fourth stage transitions to the terminal state. Reward and utility means are linear in ϕ with parameters

$$\upsilon_r = 10^{-1}(9.5, 0.5, 8.0, 2.5, 6.5, 1.0)^\top, \qquad \upsilon_u = 10^{-1}(0.5, 9.5, 2.5, 8.5, 4.5, 9.0)^\top.$$

Conditional rewards and utilities are independent Bernoulli observations with these means. Thus, learning is genuinely online and stochastic even though the finite hidden model permits exact post-hoc evaluation.

Benchmark and plotted quantities. The lower-bound constraint is $c = 2.14$. Solving the occupancy-measure linear programme gives $V_r^{\pi^\star} \approx 2.2$ and $V_u^{\pi^\star} \approx 2.1$, with the utility constraint active. For every current policy, we use the hidden model only to report

$$\Delta_c^k = c - V_u^{\pi^k}, \qquad \Delta_r^k = V_r^{\pi^\star} - V_r^{\pi^k}.$$

Hence $\Delta_c^k \le 0$ denotes feasibility. The reward gap is signed: an infeasible policy may have $\Delta_r^k < 0$ because it can exceed the reward of the constrained optimum.

Compared methods. Both methods use the online primal-dual loop of Algorithm 1 and the componentwise optimistic ridge evaluation of Algorithm 2. The saved runs use a fixed-dimensional log-linear actor fitted by the weighted action-difference objective in (9) on a balanced 15-triplet coreset. The fit uses the minimum-norm solution on the rank-five feature-difference span and no actor ridge. The critic ridge parameter is 1.0, the confidence level is $\delta = 5.0 \times 10^{-2}$, and $\lambda_{\max} \approx 22.5$. Each run performs $K = 1.5 \times 10^4$ policy updates, with stepsize $\eta = 2.0 \times 10^{-2}$ on-policy and $\eta = 5.0 \times 10^{-2}$ off-policy. The regularised method uses $\tau = 2.0 \times 10^{-2}$, whereas the baseline uses $\tau = 0$. Both methods train and are evaluated at $c = 2.14$.

On- and off-policy data. In the on-policy condition, the oracle discards past data and receives a fresh batch of $N = 3.2 \times 10^4$ trajectories at every update. In the off-policy condition, it appends one

new trajectory per update to cumulative ridge sufficient statistics. This gives 4.8×10^8 trajectories per on-policy seed and 1.5×10^4 trajectories per off-policy seed. The large fresh batches isolate last-iterate behaviour rather than establish an on-policy sample-efficiency advantage.

The figures show one seed in each sampling mode.

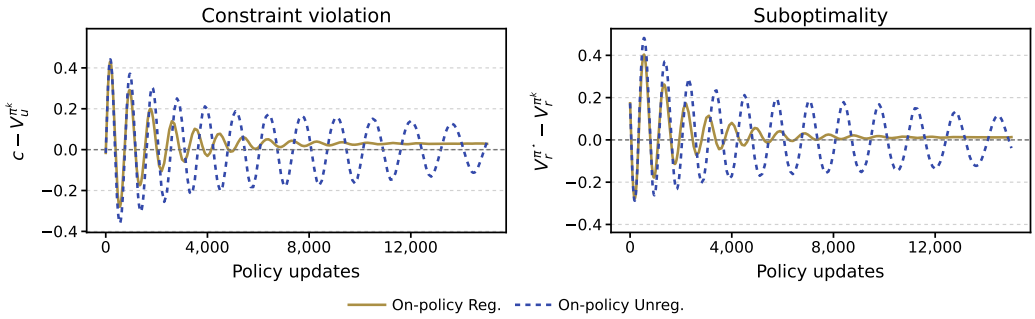

Figure 2: On-policy last-iterate performance for one seed. Solid and dashed curves denote the regularised and unregularised methods, both trained at c.

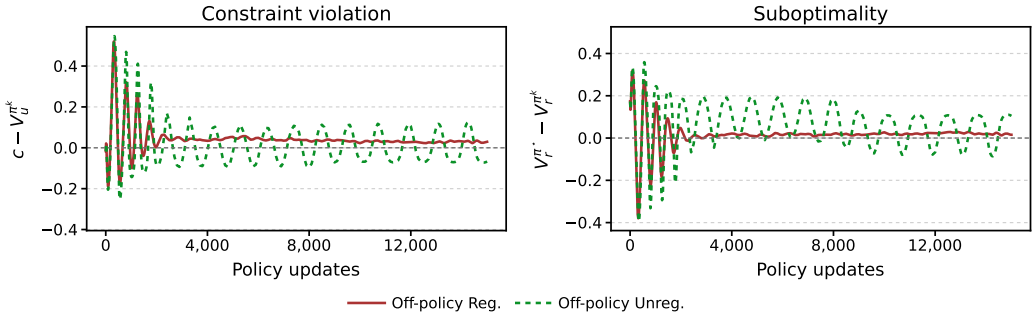

Figure 3: Off-policy last-iterate performance for one seed, using the same panels, visual encoding, and axis limits as Figure 2.

Table B: Late-training fluctuations of the signed constraint gap. For each run, we sort the recorded gaps during updates 13,000–15,000, discard the lowest and highest 10%, and measure the width of the remaining values. Each entry gives the median width across 30 seeds, with the [25th, 75th] percentiles in brackets.

Sampling mode	Regularised	Unregularised
On-policy	0.0027 [0.0023, 0.0037]	0.2234 [0.2199, 0.2247]
Off-policy	0.0068 [0.0052, 0.0084]	0.1790 [0.0794, 0.3122]

Figures 2 and 3 show the two sampling modes separately. Over 1.5×10^4 updates, the regularised curves damp their initial oscillations more rapidly and remain flatter than the unregularised curves in both modes.

We further complement the training-curve plot with the following statistic, computed over 30 seed runs, to highlight the difference in fluctuations between the vanilla primal-dual method and its regularised counterpart. For seed j, let $\Delta_{c,j}^k$ and $\Delta_{r,j}^k$ denote the signed constraint and reward gaps at update k. Over the recorded checkpoints $\mathcal{K}_{\text{tail}} = \{13{,}010, 13{,}020, \ldots, 15{,}000\}$, the fluctuation score in Table B is

$$W_j = \text{Quant}_{0.9}\big(\Delta_{c,j}^k : k \in \mathcal{K}_{\text{tail}}\big) - \text{Quant}_{0.1}\big(\Delta_{c,j}^k : k \in \mathcal{K}_{\text{tail}}\big),$$

where Quant_p is the empirical p-quantile. Intuitively, a larger W_j indicates greater oscillation in the late-training curve.