When Does Trajectory-Level Supervision Permit Efficient Offline Reinforcement Learning?

Anonymous authors
Paper under double-blind review

Abstract

Offline reinforcement learning is typically analyzed under process-level reward supervision, yet many sequential decision datasets record only trajectory-level outcomes, such as task success, a final score, or a human preference between two trajectories. We develop a statistical theory of offline policy optimization from such feedback. For the cumulative-reward objective with one scalar outcome label per trajectory, we propose OPAC, a pessimistic actor-critic algorithm that learns a latent per-step reward together with the critic. OPAC finds an ε-optimal policy from $\widetilde{O}(H^4 C_{sa}/\varepsilon^2)$ trajectories under single-policy concentrability C_{sa}, and we prove a matching lower bound, characterizing the statistical cost of replacing process-level rewards with one trajectory-level label. The approach extends to pairwise trajectory preferences with comparable guarantees. We then study nonlinear aggregations of per-step rewards that serve as both the supervision and the objective. Here efficiency is not guaranteed: all-success outcomes can require exponentially many trajectories even with deterministic transitions and constant coverage. For structured aggregations affine in the continuation value, which admit Bellman-style dynamic programming, we identify two coefficients that quantify the information lost in outcome aggregation and Bellman updates, and show that generalized OPAC is sample-efficient when both are polynomially bounded.

1 Introduction

Offline reinforcement learning (RL) asks whether a good policy can be learned from a fixed dataset of past interactions (Lange et al., 2012; Levine et al., 2020; Jiang & Xie, 2025), and has produced practical deep RL algorithms (e.g., Fujimoto et al., 2019; Kumar et al., 2020; Fujimoto & Gu, 2021; Kostrikov et al., 2022). Most of its theory assumes process-level rewards (Agarwal et al., 2022; Foster & Rakhlin, 2023): each logged state–action pair comes with its reward. This gives direct access to Bellman targets, but long-horizon applications rarely provide it. A dataset may record only whether a task eventually succeeded (Andrychowicz et al., 2017; Lightman et al., 2024), a final score, whether a final answer was correct (Uesato et al., 2022), whether a patient recovered (Komorowski et al., 2018; Gottesman et al., 2019), or which of two trajectories a human preferred (Christiano et al., 2017). The feedback is trajectory-level; the decisions remain sequential (Jia et al., 2025; Chen et al., 2025).

We develop a statistical theory of offline policy optimization from trajectory-level supervision. Beyond whether it suffices, we ask where the statistical difficulty comes from: distribution shift in offline control (Xie et al., 2021a; Cheng et al., 2022; Jiang & Xie, 2025), recovering latent per-step rewards from aggregated labels (Xu et al., 2024; Jia et al., 2025; Chen et al., 2025), or the trajectory-level objective itself (Zhang et al., 2020). They lead to qualitatively different regimes.

We begin with the regime closest to classical offline RL. The objective is still the expected cumulative reward $J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{h=1}^{H} r_h^\star(s_h, a_h)\right]$, but each trajectory in the data carries only a bounded scalar outcome whose conditional mean is its cumulative reward. We ask:

What sample complexity is sufficient and necessary under this outcome model?

We answer this question up to logarithmic factors. Our algorithm, OPAC, is a pessimistic actor-critic method (Xie et al., 2021a; Cheng et al., 2022) that learns a latent per-step reward jointly with the

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.

Table 1: Summary of results: number of trajectories needed for ε-optimality, with $C_{sa} = C_{sa}(\pi^\star)$. Preference returns are normalized to $[0, 1]$. The tabular row treats the numbers of states and actions and the coverage as constants; see the referenced results for precise conditions.

Supervision	Upper bound	Lower bound	Reference
Cumulative objective $J(\pi) = \mathbb{E}_\pi\left[\sum_{h=1}^H r_h^\star(s_h, a_h)\right]$			
Per-step rewards, tabular	$\widetilde{O}(H^3/\varepsilon^2)$	$\Omega(H^3/\varepsilon^2)$	Xie et al., 2021b
Scalar outcome, $Y \in [0, H]$	$\widetilde{O}(H^4 C_{sa}/\varepsilon^2)$	$\Omega(H^4 C_{sa}/\varepsilon^2)$	Thms. 1, 2
Pairwise preference	$\widetilde{O}\big((1+\alpha_{\mathrm{C}}^{-1})(H^2 + c_{\mathrm{C}} H) C_{sa}/\varepsilon^2\big)$	—	Thm. 3
Generalized objective $J(\pi) = \mathbb{E}_\pi\left[\sigma(r_1^\star, \ldots, r_H^\star)\right]$			
All-success, $\sigma = \prod_h r_h$	—	$\Omega(2^H)$	Thm. 5
Structured σ	$\widetilde{O}\big(V_{\max}^4 L^2 (\kappa_\mu H^2 C_{sa} + \chi_\mu H^4)/\varepsilon^2\big)$	—	Thm. 9

critic: the reward is fit by trajectory-level regression, the critic by a Bellman error under that reward, and pessimism guards against distribution shift. Under realizability and Bellman completeness, OPAC finds an ε-optimal policy from $\widetilde{O}(H^4 C_{sa}/\varepsilon^2)$ trajectories, where C_{sa} is the single-policy concentrability (Theorem 1). Conversely, $\Omega(H^4 C_{sa}/\varepsilon^2)$ trajectories are necessary for every $C_{sa} \geq 2$ and large H, even with deterministic transitions (Theorem 2). The key point is that OPAC never identifies the per-step rewards: for the cumulative objective, learning their trajectory sums suffices.

We then turn to pairwise trajectory preferences generated by a Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Christiano et al., 2017), with the same cumulative objective. OPAC applies after replacing squared return regression by logistic preference regression, and for returns normalized to $[0, 1]$ its guarantee matches the scalar-outcome one up to constants of the BTL model (Theorem 3). Calibrated scalar outcomes are thus not essential: binary comparisons support the same pessimistic mechanism, even though they reveal returns only up to a common additive shift.

The cumulative reward is only one way to evaluate a trajectory. Often the recorded outcome is the performance criterion itself, such as whether every stage of a task succeeded, and it may combine per-step rewards nonlinearly, as in RL with general utilities (Zhang et al., 2020; Barakat et al., 2023). We therefore fix a known aggregation map σ and consider the objective $J(\pi) = \mathbb{E}_{\tau\sim\pi}[\sigma(r_1^\star, \ldots, r_H^\star)]$, where $r_h^\star$ is evaluated at (s_h, a_h). The data record outcomes Y with $\mathbb{E}[Y \mid \tau] = \sigma(r_1^\star, \ldots, r_H^\star)$, so the outcome is both the supervision signal and the optimization target. We ask:

Is efficient learning still possible when both the supervision and the objective are generalized?

In general it is not. For the all-success objective $\sigma(r_1, \ldots, r_H) = \prod_{h=1}^H r_h$ with binary rewards, any offline learner may need $\Omega(2^H)$ trajectories to reach suboptimality below $1/8$, even with deterministic transitions and constant state–action concentrability (Theorem 5), because the outcome is informative only on an exponentially small set of trajectories, so coverage is not the bottleneck.

We then identify a learnable structured subclass. When σ is affine in its continuation value, the objective admits a generalized Bellman recursion and a performance-difference identity with prefix weights (Proposition 6), so OPAC extends with a reweighted policy loss. Two sources of information loss then govern the sample complexity. The reward process coefficient $\kappa_\mu(\sigma)$ (Definition 1) measures how much aggregation hides differences between per-step rewards, and it governs how fast least squares on outcomes recovers them (Proposition 7). The Bellman inverse coefficient $\chi_\mu(\sigma)$ (Definition 2) measures how much generalized Bellman targets compress the same differences. Neither controls the other: on simple chains, each stays bounded while the other grows exponentially in H (Proposition 8, Table 2). Generalized OPAC is polynomial in both coefficients and the other parameters (Theorem 9). These conditions are sufficient; we do not claim that they are necessary.

Contributions (i) For the cumulative objective with scalar outcomes, we propose OPAC and prove matching upper and lower bounds of order $H^4 C_{sa}/\varepsilon^2$, which characterize the statistical cost of replacing process-level rewards with one trajectory-level label (Theorems 1 and 2). (ii) We extend OPAC to BTL preferences with a guarantee that matches the scalar-outcome rate up to BTL constants (Theorem 3). (iii) For generalized objectives, we prove an exponential lower bound (Theorem 5),

develop generalized dynamic programming and two incomparable information-loss coefficients (Propositions 6 to 8), and give a polynomial upper bound for generalized OPAC (Theorem 9). Table 1 summarizes the rates, and Appendix A discusses related work.

2 PRELIMINARIES

Episodic MDPs and policies An episodic MDP $M = (\mathcal{S}, \mathcal{A}, P, r^\star, H, s_1)$ has layered states $\mathcal{S} = \bigsqcup_{h=1}^{H} \mathcal{S}_h$, a finite action set $\mathcal{A}$ (written $\mathcal{A}_h$ at stage h), transitions $P_h : \mathcal{S}_h \times \mathcal{A}_h \to \Delta(\mathcal{S}_{h+1})$, a fixed initial state $s_1 \in \mathcal{S}_1$, and an unknown mean reward $r_h^\star : \mathcal{S}_h \times \mathcal{A}_h \to [0,1]$ at each stage $h \in [H]$. A trajectory is $\tau = (s_1, a_1, \dots, s_H, a_H)$. The reward $r^\star$ defines the objective, but the offline data never reveal it step by step. A Markov policy $\pi = (\pi_h)_{h=1}^H$ draws $a_h \sim \pi_h(\cdot \mid s_h)$; we write $\mathbb{E}_\pi$ for expectations over the trajectories it generates from s_1 and d_h^π for the law of (s_h, a_h). We fix a policy class Π that contains the behavior policy μ used to collect the data.

Rewards, values, and function classes For a candidate reward $r = (r_h)_{h=1}^H$, the cumulative reward of τ is $R(\tau; r) := \sum_{h=1}^H r_h(s_h, a_h)$. The action value $Q_h^{\pi,r}(s,a)$ is the expected reward collected from stage h onward when π is run from $(s_h, a_h) = (s,a)$, and the state value $V_h^{\pi,r}(s)$ is its average over $a \sim \pi_h(\cdot \mid s)$. We drop r when $r = r^\star$ and write $R^\star(\tau) := R(\tau; r^\star)$. The goal is to maximize $J(\pi) := \mathbb{E}_\pi[R^\star(\tau)] = V_1^\pi(s_1)$, and $\pi^\star \in \operatorname{argmax}_{\pi \in \Pi} J(\pi)$ is the in-class comparator. The learner uses a critic class $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ with $\mathcal{F}_h \subseteq \{f : \mathcal{S}_h \times \mathcal{A}_h \to [0, V_{\max}]\}$ and a class $\mathcal{R}$ of candidate rewards $r_h : \mathcal{S}_h \times \mathcal{A}_h \to [0,1]$. For simplicity, Π, $\mathcal{F}$, and $\mathcal{R}$ are finite and fixed before the data are drawn, and Π contains every actor iterate generated by the algorithm; for linear critics, these iterates are log-linear policies, whose log-covering numbers grow only logarithmically in K.

Bellman operator For a reward r and a policy π, the Bellman operator $\mathcal{T}_r^\pi$ acts on $f = (f_h)_{h=1}^H$ by

$$(\mathcal{T}_r^\pi f)_h(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot \mid s,a)}\big[f_{h+1}(s', \pi)\big], \qquad (s,a) \in \mathcal{S}_h \times \mathcal{A}_h, \tag{1}$$

where $f_h(s, \pi) := \mathbb{E}_{a \sim \pi_h(\cdot \mid s)}[f_h(s,a)]$, $f_{H+1} \equiv 0$, and $\mathcal{T}^\pi := \mathcal{T}_{r^\star}^\pi$. Thus $Q^{\pi,r}$ is the unique fixed point of $\mathcal{T}_r^\pi$, and a critic f with small $f - \mathcal{T}_r^\pi f$ approximately evaluates π under r.

Offline data and coverage Unless stated otherwise, the dataset $\mathcal{D} = \{(\tau^{(i)}, Y^{(i)})\}_{i=1}^n$ consists of n i.i.d. trajectory–outcome pairs with $\tau^{(i)} \sim \mu$ and $\mathbb{E}[Y^{(i)} \mid \tau^{(i)}] = R^\star(\tau^{(i)})$, and $\mathbb{E}_\mathcal{D}$ denotes the empirical average over $\mathcal{D}$. We measure coverage by the all-step state–action concentrability (Chen & Jiang, 2019; Xie et al., 2021a) of π relative to μ,

$$C_{sa}(\pi) := \max_{h \in [H]} \sup_{(s,a) \in \mathcal{S}_h \times \mathcal{A}_h} \frac{d_h^\pi(s,a)}{d_h^\mu(s,a)}, \qquad \text{with } 0/0 = 0 \text{ and } x/0 = +\infty \text{ for } x > 0. \tag{2}$$

It is small when the data visit every state–action pair that π visits with comparable frequency, and we require it only for $\pi^\star$. We use this classical density-ratio notion (Munos, 2007; Farahmand et al., 2010; Chen & Jiang, 2019) rather than function-class-dependent coverage (Duan et al., 2020; Xie et al., 2021a; Song et al., 2023), as we focus on the granularity of reward supervision.

Notation For a distribution ν on $\mathcal{S}_h \times \mathcal{A}_h$, let $\|q\|_{2,\nu}^2 := \mathbb{E}_\nu[q^2]$; for $q = (q_h)_{h=1}^H$, the summed norm is $\|q\|_{2,\mu}^2 := \sum_{h=1}^H \mathbb{E}_{d_h^\mu}[q_h^2]$. The symbols $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, and $\lesssim$ hide universal constants; their tilde versions also hide factors polylogarithmic in $n, H, |\mathcal{A}|, |\Pi|, |\mathcal{F}|, |\mathcal{R}|$, and $1/\delta$.

3 OFFLINE RL FROM SCALAR OUTCOMES

In this section each offline trajectory carries a single scalar label: the learner observes $Y \in [0, H]$ with $\mathbb{E}[Y \mid \tau] = R^\star(\tau)$ but none of the per-step rewards $r_h^\star(s_h, a_h)$, and the target is still J; hence $V_{\max} = H$. The usual Bellman target $r_h^\star(s_h, a_h) + V_{h+1}(s_{h+1})$ therefore cannot be formed from the data. OPAC instead treats the reward as a latent function and learns it jointly with the critic: the reward must explain the outcome labels, and the critic must be Bellman-consistent under that reward.

Algorithm 1 Outcome-based Pessimistic Actor-Critic (OPAC)

Input: dataset $\mathcal{D}$, classes $\mathcal{F}, \mathcal{R}$, pessimism $\beta > 0$, temperature $\eta > 0$, iterations K

1: $\pi_{1,h}(\cdot \mid s) \leftarrow \mathrm{Unif}(\mathcal{A}_h)$ for all $h \in [H]$ and $s \in \mathcal{S}_h$
2: **for** $k = 1, \ldots, K$ **do**
3: $\quad (f_k, r_k) \in \operatorname{argmin}_{(f,r)\in\mathcal{F}\times\mathcal{R}} \mathcal{L}_{\mathcal{D}}(\pi_k, f) + \beta\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r, f) + \beta\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r)$ ▷ evaluation
4: $\quad \pi_{k+1,h}(\cdot \mid s) \propto \pi_{k,h}(\cdot \mid s)\exp\big(f_{k,h}(s,\cdot)/\eta\big)$ for all h and s ▷ improvement
5: **end for**
6: **return** $\bar{\pi} = \mathrm{Unif}(\pi_1, \ldots, \pi_K)$ ▷ draw k once, then run π_k for the episode

3.1 Algorithm

Empirical criteria OPAC scores a policy π, a critic f, and a candidate reward r by three empirical criteria built on the one-step target $y_h^{r,f,\pi} := r_h(s_h, a_h) + f_{h+1}(s_{h+1}, \pi)$, where $f_{H+1} \equiv 0$:

$$\mathcal{L}_{\mathcal{D}}(\pi, f) := \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}\big[f_h(s_h, \pi) - f_h(s_h, a_h)\big], \quad (3a) \qquad \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r) := \mathbb{E}_{\mathcal{D}}\big[(Y - R(\tau; r))^2\big], \quad (3b)$$

$$\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi, r, f) := \sum_{h=1}^{H} \Big\{\mathbb{E}_{\mathcal{D}}\big[(f_h(s_h, a_h) - y_h^{r,f,\pi})^2\big] - \min_{f' \in \mathcal{F}_h} \mathbb{E}_{\mathcal{D}}\big[(f'(s_h, a_h) - y_h^{r,f,\pi})^2\big]\Big\}. \quad (3c)$$

By the performance difference lemma, the population version of the policy loss $\mathcal{L}_{\mathcal{D}}(\pi, f)$ equals $J(\pi) - J(\mu)$ when $f = Q^\pi$, so it estimates how much π improves on the behavior policy according to f. The Bellman error $\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}$ measures how far f is from a fixed point of $\mathcal{T}_r^\pi$: its first term alone is inflated by the conditional variance of the target $y_h^{r,f,\pi}$, and subtracting the best fit within $\mathcal{F}_h$ removes this bias under completeness (Lemma 14). Finally, $\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}$ regresses each outcome on the reward sum of its trajectory. Because r enters both $\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}$ and $\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}$, it ties the critic to the outcome labels.

Outcome-based Pessimistic Actor-Critic (OPAC) Algorithm 1 alternates two steps. The evaluation step (line 3) is pessimistic: the penalties keep (f, r) consistent with the data, and among such pairs the first term selects the critic under which π_k looks least favorable relative to μ, so π_k receives no credit for actions whose value the data cannot confirm. The improvement step (line 4) runs exponential weights on this pessimistic critic at every state. In practice, OPAC can be implemented with the actor-critic algorithm of Cheng et al. (2022, Algorithm 2) and one additional reward network.

3.2 Upper Bound

Our analysis uses three assumptions. The only outcome-specific one is reward realizability; the other two are approximate value realizability and Bellman completeness, standard in offline RL with function approximation. Completeness is imposed uniformly over candidate rewards $r \in \mathcal{R}$, since OPAC evaluates Bellman backups under rewards that are themselves estimated from outcome labels.

Assumption 1 (Reward realizability)**.** *The ground-truth mean process reward satisfies $r^\star \in \mathcal{R}$.*

Assumption 2 (Value realizability)**.** *For every $\pi \in \Pi$, $\min_{f\in\mathcal{F}} \max_{h,\nu} \|f_h - (\mathcal{T}^\pi f)_h\|^2_{2,\nu} \le \varepsilon_{\mathcal{F}}$, where h ranges over $[H]$ and ν over the state–action distributions $\{d_h^\rho : \rho \in \Pi\}$ induced by Π.*

Assumption 3 (Completeness)**.** *For every $(\pi, r, f) \in \Pi \times \mathcal{R} \times \mathcal{F}$, $\min_{f'\in\mathcal{F}} \|f' - \mathcal{T}_r^\pi f\|^2_{2,\mu} \le \varepsilon_{\mathcal{F},\mathcal{F}}$.*

Theorem 1 (Upper bound for outcome-based offline RL)**.** *Under Assumptions 1 to 3 and $C_{sa}(\pi^\star) < \infty$, OPAC with suitable η, β and K iterations returns a policy $\bar{\pi}$ such that*

$$J(\pi^\star) - J(\bar{\pi}) = \widetilde{O}\left(H^2\sqrt{\frac{C_{sa}(\pi^\star)}{n}} + \frac{H^2}{\sqrt{K}} + H\sqrt{C_{sa}(\pi^\star)\varepsilon_{\mathcal{F}}} + \sqrt{HC_{sa}(\pi^\star)\varepsilon_{\mathcal{F},\mathcal{F}}} + H\sqrt{\varepsilon_{\mathcal{F}}}\right) \quad (4)$$

with probability at least $1 - \delta$. In particular, if $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$ and $K \ge n$, the suboptimality is $\widetilde{O}\big(H^2\sqrt{C_{sa}(\pi^\star)/n}\big)$, so $n = \widetilde{O}\big(H^4 C_{sa}(\pi^\star)/\varepsilon^2\big)$ trajectories suffice for ε-optimality.

The proof in Appendix C specifies η and β; the latter depends on $C_{sa}(\pi^\star)$. The terms of (4) are the statistical error under coverage $C_{sa}(\pi^\star)$, the optimization error of K iterations, and the approximation

errors $\varepsilon_{\mathcal{F}}$ and $\varepsilon_{\mathcal{F},\mathcal{F}}$. The statistical error has two sources. Bellman errors are per-step quantities, so transferring them from $d_h^{\pi^\star}$ to d_h^{μ} costs $C_{sa}(\pi^\star)$. Reward errors are observed only through trajectory sums, and a change of trajectory measure (Yuan et al., 2025) bounds $\big(\mathbb{E}_{\pi^\star}[R^\star(\tau) - R(\tau;r)]\big)^2$ by $HC_{sa}(\pi^\star)$ times the outcome-regression error $\mathbb{E}_\mu\big[(R^\star(\tau) - R(\tau;r))^2\big]$. OPAC thus never needs to identify $r^\star$ stage by stage: shifting reward between stages without changing trajectory sums alters neither the labels nor J. At fixed class sizes, the leading term stems from Bellman errors, which accumulate over H stages with targets of scale H.

3.3 Lower Bound

The next theorem shows that, up to class-complexity logarithms, the dependence of (4) on H, n, and $C_{sa}(\pi^\star)$ cannot be improved: ε-optimality can require $\Omega(H^4 C_{sa}(\pi^\star)/\varepsilon^2)$ trajectories.

Theorem 2 (Lower bound for outcome-based offline RL)**.** *There exist universal constants $c > 0$ and H_0 such that, for every $C \geq 2$, $H \geq H_0$, and $n \geq 32CH^2$, there is a class of outcome-based offline RL instances with deterministic transitions, at most three actions, and $C_{sa}(\pi^\star) = C$ on which every algorithm that uses n trajectories and outputs a policy $\widehat{\pi}$ satisfies*

$$\sup_M \mathbb{E}_M\big[J_M(\pi_M^\star) - J_M(\widehat{\pi})\big] \geq cH^2\sqrt{\frac{C_{sa}(\pi^\star)}{n}}.$$

The proof is in Appendix E.1. In the hard instances, each stage of a deterministic chain has a good and a bad action whose rewards differ by Δ, the behavior policy tries each with probability $1/C$, and the label $Y \in \{0, H\}$ has noise of order H. Since one stage shifts the mean label by at most Δ, each trajectory carries $O(\Delta^2/(CH^2))$ information about that stage. Identifying the good action at a constant fraction of stages thus needs $n \gtrsim CH^2/\Delta^2$ trajectories, and failing to do so costs regret of order $H\Delta$; setting $H\Delta \asymp \varepsilon$ gives $n \gtrsim H^4 C/\varepsilon^2$, matching the explicit dependence in Theorem 1.

Remark 1 (Comparison with process supervision)**.** *Under exact realizability and completeness and $K \geq n$, Theorems 1 and 2 determine the explicit trajectory dependence $H^4 C_{sa}(\pi^\star)/\varepsilon^2$ of expected ε-optimality, up to the class-complexity logarithms of the upper bound. With process supervision, the tabular rate is $\widetilde{\Theta}(H^3/\varepsilon^2)$ at fixed per-stage state- and action-space sizes and coverage, for $\varepsilon \leq H^{-5/2}$ (Xie et al., 2021b). The displayed horizon dependence thus differs by one factor of H in sample complexity, equivalently $\sqrt{H}$ in suboptimality, under the bounded label model $Y \in [0, H]$, whose conditional variance can be of order H^2. For linear MDPs with trajectory-sum labels, Xu et al. (2024) report an $\widetilde{O}(dH^3/\sqrt{n})$ upper bound under a well-explored-dataset condition, one factor of H in suboptimality above the per-step-reward bound of pessimistic value iteration (Jin et al., 2021b); that gap is not shown to be necessary, whereas ours is tight in its explicit dependence.*

4 Offline RL from Pairwise Preferences

We next replace scalar outcomes by pairwise comparisons: the learner observes only which of two trajectories is preferred. This is the offline analogue of the preference-based outcome-feedback model of Chen et al. (2025) and a stylized abstraction of reward modeling in RLHF. The target remains the cumulative return, normalized in this section so that $r_h^\star(s,a) \in [0,1]$ and $R^\star(\tau) \in [0,1]$.

Preference observation model The dataset $\mathcal{D}_{\mathrm{Pref}}$ consists of n i.i.d. triples $(\tau^{(i),+}, \tau^{(i),-}, y^{(i)})$, in which the two trajectories are drawn independently from μ and $y^{(i)} \in \{0,1\}$ indicates whether the first is preferred. Labels follow the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Christiano et al., 2017): given the pair, $y \sim \mathrm{Bern}(\mathrm{C}_{r^\star}(\tau^+, \tau^-))$, where for a reward r and a fixed inverse temperature $\gamma > 0$, which controls the sensitivity to return differences,

$$\mathrm{C}_r(\tau^+, \tau^-) := \frac{\exp\big(\gamma R(\tau^+; r)\big)}{\exp\big(\gamma R(\tau^+; r)\big) + \exp\big(\gamma R(\tau^-; r)\big)}. \tag{5}$$

Preference loss and algorithm Preference-based OPAC fits the latent reward by maximum likelihood: it replaces the squared loss $\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}$ in (3b) by the negative BTL log-likelihood

$$\mathcal{L}_{\mathcal{D}_{\mathrm{Pref}}}^{\mathrm{Pref}}(r) := -\frac{1}{n}\sum_{i=1}^{n}\Big[y^{(i)} \log \mathrm{C}_r(\tau^{(i),+}, \tau^{(i),-}) + (1 - y^{(i)}) \log\big(1 - \mathrm{C}_r(\tau^{(i),+}, \tau^{(i),-})\big)\Big]. \tag{6}$$

Both losses fit the latent per-step reward to trajectory feedback, but preferences identify rewards only through return differences; the common shift of all returns that they cannot identify is irrelevant for policy optimization. Preference-based OPAC runs Algorithm 1 on the $2n$ trajectories $\mathcal{D}_{\text{traj}}$ of the preference pairs, with $\mathcal{L}^{\text{Pref}}_{\mathcal{D}_{\text{Pref}}}$ in place of $\mathcal{L}^{\text{RM}}_{\mathcal{D}}$ in line 3; all other steps are unchanged.

Sample complexity Let every $r \in \mathcal{R}$ satisfy $0 \le r_h \le 1$ and $R(\tau; r) \in [0, 1]$, and set $V_{\max} = 1$. The constants α_{C} and c_{C} of (61) depend only on γ: α_{C} is half the minimal curvature of the negative BTL log-likelihood in the return difference on $[-1, 1]$, small when comparisons are nearly random (γ small) or saturated (γ large), and c_{C} controls the range and variance of log-likelihood ratios.

Theorem 3 (Upper bound for preference-based offline RL)**.** *Suppose Assumptions 1 to 3 hold and $C_{sa}(\pi^\star) < \infty$. With suitable parameters η and β, preference-based OPAC run for K iterations returns a policy $\bar{\pi}$ such that, with probability at least $1 - \delta$ over the data,*

$$J(\pi^\star) - J(\bar{\pi}) = \widetilde{O}\Bigg(\sqrt{1 + \frac{1}{\alpha_{\text{C}}}}\left(H + \sqrt{c_{\text{C}} H}\right)\sqrt{\frac{C_{sa}(\pi^\star)}{n}} + \frac{H}{\sqrt{K}} + \sqrt{1 + \frac{1}{\alpha_{\text{C}}}}\left(H\sqrt{C_{sa}(\pi^\star)\varepsilon_{\mathcal{F}}} + \sqrt{H C_{sa}(\pi^\star)\varepsilon_{\mathcal{F},\mathcal{F}}}\right) + H\sqrt{\varepsilon_{\mathcal{F}}}\Bigg). \tag{7}$$

In particular, if $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$ and $K \ge n$, the bound is dominated by its first term.

The proof is in Appendix D. Rescaling Theorem 1 from returns in $[0, H]$ to $[0, 1]$ divides its rate by H and gives $H\sqrt{C_{sa}(\pi^\star)/n}$, the leading term of (7). For fixed γ, binary comparisons thus match scalar outcome labels up to α_{C} and c_{C}, even though they reveal returns only up to a common shift.

5 Generalized Trajectory Objectives

We now let the outcome be both the signal and the objective, possibly nonlinear in per-step rewards via a known aggregation map $\sigma : [0, 1]^H \to [0, V_{\max}]$; the trajectory return and policy value are

$$R(\tau; r) \coloneqq \sigma\big(r_1(s_1, a_1), \ldots, r_H(s_H, a_H)\big), \qquad J(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi}\big[R(\tau; r^\star)\big], \tag{8}$$

and $\pi^\star \in \operatorname{argmax}_{\pi\in\Pi} J(\pi)$. In this section and its appendices, R, V, Q, J, $\mathcal{T}$, and all losses are defined via σ. The sum gives the cumulative setting; the all-success outcome below is nonlinear.

Example 4 (All-success outcome)**.** *Let $r_h(s, a) \in \{0, 1\}$ indicate success at stage h, and let $\sigma(u_1, \ldots, u_H) = \prod_{h=1}^{H} u_h$. Then $R(\tau; r) = 1$ if every stage of τ succeeds and $R(\tau; r) = 0$ otherwise, so $J(\pi)$ is the probability that π completes a trajectory without a failed stage.*

The data $\mathcal{D}$ are as in Section 2, except that $Y^{(i)} \in [0, V_{\max}]$ and $\mathbb{E}[Y^{(i)} \mid \tau^{(i)}] = R(\tau^{(i)}; r^\star)$.

5.1 An Exponential Lower Bound

Without further structure, the sample complexity can be exponential in H: under the all-success outcome, the label is informative only on an exponentially small set of trajectories.

Theorem 5 (Exponential lower bound under all-success aggregation)**.** *For every $H \ge 1$, there is a family of deterministic MDPs indexed by $\boldsymbol{\theta} \in \{0, 1\}^H$, with the all-success outcome and $C_{sa}(\pi^\star) \le 2$, such that every learner $\widehat{\pi}$ using n i.i.d. outcome-labeled trajectories satisfies $\sup_{\boldsymbol{\theta}\in\{0,1\}^H} \mathbb{E}[J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi})] \ge \frac{1}{4}(1 - n \cdot 2^{1-H})$. Hence expected suboptimality below $1/8$ requires $\Omega(2^H)$ trajectories.*

The proof is in Appendix E.2. We therefore restrict attention to structured aggregations and bound the sample complexity of generalized OPAC by two coefficients quantifying the information lost in outcome regression and in Bellman backups; we claim sufficiency, not necessity.

5.2 Generalized Dynamic Programming

For the upper bound, suppose σ is computed by a backward recursion with known maps $\sigma_h : \mathbb{R}^2 \to \mathbb{R}$ and a terminal value g: for each suffix $\tau_h = (s_h, a_h, \ldots, s_H, a_H)$ of a trajectory,

$$R_{H+1} \coloneqq g, \qquad R_h(\tau_h; r) \coloneqq \sigma_h\big(r_h(s_h, a_h), R_{h+1}(\tau_{h+1}; r)\big), \quad h = H, \ldots, 1, \tag{9}$$

with $R_1(\tau_1; r) = R(\tau; r)$ for every $r \in \mathcal{R}$. All candidate continuation returns, including $R_{H+1} = g$, and all critics f_h take values in $[0, V_{\max}]$. Cumulative returns use $\sigma_h(u, v) = u + v$ and $g = 0$; all-success (Example 4) uses $\sigma_h(u, v) = uv$ and $g = 1$, so that $R_h(\tau_h; r) = \prod_{j \geq h} r_j(s_j, a_j)$.

Values and Bellman operator Let $Q_h^{\pi,r}(s, a) := \mathbb{E}_\pi[R_h(\tau_h; r) \mid s_h = s, a_h = a]$, $V_h^{\pi,r}(s) := Q_h^{\pi,r}(s, \pi)$, and $J_r(\pi) := V_1^{\pi,r}(s_1)$, so $J = J_{r^\star}$. The Bellman operator aggregates through σ_h:

$$(\mathcal{T}_r^\pi f)_h(s, a) := \sigma_h\Big(r_h(s, a),\, \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\big[f_{h+1}(s', \pi)\big]\Big), \qquad (s, a) \in \mathcal{S}_h \times \mathcal{A}_h, \tag{10}$$

with $f_{H+1}(\cdot, \pi) \equiv g$ and $\mathcal{T}^\pi := \mathcal{T}_{r^\star}^\pi$. When σ_h is affine in v, dynamic programming carries over.

Proposition 6 (Generalized dynamic programming). *Let $r \in \mathcal{R}$ and each $\sigma_h(u, v) = a_h(u)v + b_h(u)$.*

(i) Evaluation. *For every policy π, $Q^{\pi,r}$ is the unique fixed point of $\mathcal{T}_r^\pi$.*

(ii) Optimality. *If $a_h \geq 0$, the optimality equation*

$$f_h(s, a) = \sigma_h\Big(r_h(s, a), \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\big[\max_{a'} f_{h+1}(s', a')\big]\Big), \qquad f_{H+1} \equiv g,$$

has the unique solution $Q_h^{\star,r} = \sup_\pi Q_h^{\pi,r}$, and every policy greedy with respect to $Q^{\star,r}$ maximizes J_r over Markov policies.

(iii) Performance difference. *For all policies π, π', with the advantage $A_h^{\pi',r} := Q_h^{\pi',r} - V_h^{\pi',r}$,*

$$J_r(\pi) - J_r(\pi') = \sum_{h=1}^{H} \mathbb{E}_{\tau \sim \pi}\big[W_h^r(\tau)\, A_h^{\pi',r}(s_h, a_h)\big], \qquad W_h^r(\tau) := \prod_{j=1}^{h-1} a_j\big(r_j(s_j, a_j)\big).$$

The proofs are in Appendix F. Affinity lets conditional expectations pass through σ_h, giving (i) and (iii); monotonicity gives (ii), so a deterministic Markov policy is optimal. In (iii), the prefix weight W_h^r reweights the stage-h advantage. For all-success, $W_h^r(\tau) = \prod_{j<h} r_j(s_j, a_j)$ indicates that the first $h - 1$ stages succeeded, so a mistake at stage h matters only on trajectories still alive. When $0 \leq a_h \leq 1$, these weights lie in $[0, 1]$ and errors at later stages are not amplified; this is the role of the no-expansion condition in Assumption 4, and generalized OPAC uses W_h^r in its policy loss.

5.3 Two Coefficients of Information Loss

The upper bound depends on two coefficients, one for outcome regression and one for Bellman backups. The first involves only whole-trajectory returns and not the recursion (9).

Definition 1 (Reward process coefficient). *For the trajectory distribution induced by μ, define*

$$\kappa_\mu(\sigma) := \sup_{r, r' \in \mathcal{R},\, r \neq r'} \frac{\|r - r'\|_{2,\mu}^2}{\mathbb{E}_{\tau \sim \mu}\big[(R(\tau; r) - R(\tau; r'))^2\big]}.$$

Here $\|r - r'\|_{2,\mu}^2 = \sum_h \mathbb{E}_\mu[(r_h - r'_h)^2(s_h, a_h)]$ is the summed norm of Section 2. A large $\kappa_\mu(\sigma)$ means that distinct rewards become nearly indistinguishable once aggregated by σ. For cumulative returns, $\kappa_\mu(\sigma)$ can be infinite for every behavior policy, since moving reward between stages leaves every trajectory sum unchanged, whereas it equals one for a nontrivial class whose rewards share the same behavior average at every state (Appendix H); Theorem 1 is unaffected, as it works with trajectory sums. We next show that $\kappa_\mu(\sigma)$ is the price of least-squares reward recovery from outcomes.

Proposition 7 (Reward recovery, informal). *Let $\widehat{r}$ minimize $\sum_i (Y^{(i)} - R(\tau^{(i)}; r))^2$ over $\mathcal{R} \ni r^\star$.*

(i) With probability at least $1 - \delta$, $\|\widehat{r} - r^\star\|_{2,\mu}^2 \lesssim \kappa_\mu(\sigma)\, V_{\max}^2 \log(|\mathcal{R}|/\delta)/n$.

(ii) Conversely, fix distinct $r, r' \in \mathcal{R}$ with $\|r - r'\|_{2,\mu}^2 = \Delta^2$, and let $\kappa_{r,r'}(\sigma)$ be the ratio in Definition 1 at this pair. Under Gaussian outcome noise of variance $V_{\max}^2$, every estimator has $\max_{r^\star \in \{r, r'\}} \mathbb{E}\|\widehat{r} - r^\star\|_{2,\mu}^2 \geq \Delta^2/16$ whenever $n \leq V_{\max}^2 \kappa_{r,r'}(\sigma)/\Delta^2$.

Table 2: Coefficients on the reward classes of Appendix H (deterministic binary-action chains, independent behavior). The biased row uses two rewards differing at one stage, and its behavior takes the rewarded action with probability $1 - 1/H$. For χ_μ, Π consists of the uniform policy and $\mathcal{F}$ of its true critics; in the biased row they contain them. Larger $\mathcal{R}$, Π, or $\mathcal{F}$ only increase the coefficients. For cumulative returns, κ_μ is set by the reward class: one for nontrivial classes with a common behavior-averaged reward, infinite when stage shifts cancel on every trajectory.

Objective	$\sigma_h(u, v)$	Behavior μ	$\kappa_\mu(\sigma)$	$\chi_\mu(\sigma)$
Cumulative return	$u + v$	any	may be ∞	1
Discounted return, $c \in (0, 1)$	$u + cv$	uniform	$c^{-2(H-1)}$	1
Hybrid, k final gates	$u + v$, then uv	uniform	$k2^{k-1}$	4^{k-1}
All-success	uv	uniform	$H2^{H-1}$	4^{H-1}
All-success	uv	biased	$\leq e$	$\geq 4^{H-1}$

Applying (ii) to a pair that nearly attains the supremum in Definition 1 shows that the linear dependence on $\kappa_\mu(\sigma)$ in (i) is not an artifact of the analysis. The two parts use different noise models, so together they do not give a matching minimax rate for every class; the formal statements are Theorems 34 and 35 in Appendix I. The second coefficient concerns the one-step targets of (10).

Definition 2 (Bellman inverse coefficient)**.** *For the trajectory distribution induced by μ, define*

$$\chi_\mu(\sigma) := \sup_{\pi \in \Pi,\ f \in \mathcal{F},\ r, r' \in \mathcal{R},\ r \neq r'} \frac{\|r - r'\|_{2,\mu}^2}{\|\mathcal{T}_r^\pi f - \mathcal{T}_{r'}^\pi f\|_{2,\mu}^2}.$$

In both definitions, a positive numerator over zero is $+\infty$, $0/0 = 0$, and an empty supremum is zero. Thus $\chi_\mu(\sigma)$ measures how much Bellman targets compress reward differences: the smaller it is, the more tightly close Bellman targets force the rewards to be close. With additive returns the targets differ exactly by $r_h - r'_h$, so $\chi_\mu(\sigma) = 1$ for every behavior μ and all $\Pi, \mathcal{F}$, for a nontrivial class on the data support. For all-success they differ by $(r_h - r'_h)\,\mathbb{E}_{s'}[f_{h+1}(s', \pi)]$: reward differences shrink with the continuation value and vanish where the critic predicts failure. Generalized OPAC uses $\kappa_\mu(\sigma)$ to recover rewards from outcomes and $\chi_\mu(\sigma)$ to recover them from Bellman errors.

The discounted and biased rows of Table 2 give the following separation (proof in Appendix H).

Proposition 8 (The two coefficients are not comparable)**.** *For every $H \geq 2$ and $c \in (0, 1)$, there is a discounted-return instance with $\chi_\mu(\sigma) = 1$ and $\kappa_\mu(\sigma) = c^{-2(H-1)}$, and an all-success instance with $\kappa_\mu(\sigma) \leq e$ and $\chi_\mu(\sigma) \geq 4^{H-1}$. Hence a bound on either coefficient does not bound the other.*

In the hybrid row both coefficients grow with the number k of multiplicative gates, not with H, and $\chi_\mu(\sigma)$ stays bounded for stochastic policies: if Π consists of the policies choosing every gate action with probability at least $p \leq 1/2$ and $\mathcal{F}$ of their true critics, then $\chi_\mu(\sigma) = p^{-2(k-1)}$, whereas for $k \geq 2$ a deterministic policy can make it infinite by zeroing a continuation. Outcomes hide per-step differences on rare events, which inflates κ_μ; Bellman targets shrink them by the continuation value, which inflates χ_μ. Informative behavior removes the first effect but not the second, so both enter Theorem 9. On the hard class of Theorem 5 with $\mathcal{R} = \{r_{\boldsymbol{\theta}}\}$, $\kappa_\mu(\sigma) = H2^{H-1}$, consistent with its $\Omega(2^H)$ lower bound, but a large $\kappa_\mu(\sigma)$ alone does not imply a policy lower bound: under max-reward ($\sigma_h(u, v) = \max\{u, v\}$), it can be infinite while all policies have the same value (Appendix H).

5.4 Generalized OPAC

Generalized OPAC is analyzed under the following assumption.

Assumption 4 (σ-regularity and reward-class closure)**.** *For every stage h and $u \in [0, 1]$: (1)* affinity*: $\sigma_h(u, v) = a_h(u)v + b_h(u)$; (2)* no expansion*: $a_h(u) \in [0, 1]$; (3)* smoothness*: a_h and b_h are L-Lipschitz in u; (4)* closure*: for every $(\pi, f) \in \Pi \times \mathcal{F}$, some reward $\tilde{r}^{\pi,f} \in \mathcal{R}$ satisfies $\mathcal{T}^\pi_{\tilde{r}^{\pi,f}} f = f$.*

We use Assumptions 2 and 3 with the generalized operator (10). Items (1)–(2) are the conditions of Proposition 6, and item (3) gives $|(\mathcal{T}_r^\pi f)_h - (\mathcal{T}_{r'}^\pi f)_h| \leq L(V_{\max} + 1)|r_h - r'_h|$, converting reward errors into target errors. Item (4) says every critic is the exact value of π under some reward in $\mathcal{R}$;

it is used only for pessimism. Since both coefficients are suprema over $\mathcal{R}$, enlarging $\mathcal{R}$ to satisfy item (4) can increase them. The objectives in Table 2 satisfy items (1)–(3) with $L = 1$, as does the exponential utility $\sigma_h(u, v) = 1 - e^{-\lambda u}(1 - v)$ of risk-sensitive RL (Howard & Matheson, 1972) with $g = 0$ and $\lambda \in (0, 1]$, where $J(\pi) = 1 - \mathbb{E}_\pi[e^{-\lambda \sum_h r^\star_h}]$. The game-over objective (Appendix H) has a non-Lipschitz gate and max-reward is not affine, so Theorem 9 covers neither.

Generalized criteria Generalized OPAC changes three ingredients of OPAC: the target becomes $y_h^{r,f,\pi} := \sigma_h\big(r_h(s_h, a_h), f_{h+1}(s_{h+1}, \pi)\big)$ with $f_{H+1} \equiv g$, the reward loss $\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r) := \mathbb{E}_{\mathcal{D}}[(Y - R(\tau; r))^2]$ uses the return $R(\tau; r)$, and the policy loss is weighted by W_h^r from Proposition 6(iii):

$$\mathcal{L}^{W^r}_{\mathcal{D}}(\pi, r, f) := \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}\big[W_h^r(\tau)\,(f_h(s_h, \pi) - f_h(s_h, a_h))\big], \tag{11a}$$

$$\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi, r, f) := \sum_{h=1}^{H} \Big\{ \mathbb{E}_{\mathcal{D}}\big[(f_h(s_h, a_h) - y_h^{r,f,\pi})^2\big] - \min_{f' \in \mathcal{F}_h} \mathbb{E}_{\mathcal{D}}\big[(f'(s_h, a_h) - y_h^{r,f,\pi})^2\big] \Big\}. \tag{11b}$$

By Proposition 6(iii) with the pair (μ, π), the population version of $\mathcal{L}^{W^r}_{\mathcal{D}}$ at $f = Q^{\pi,r}$ equals $J_r(\pi) - J_r(\mu)$, so the weighted loss again measures improvement over μ; for cumulative returns $W_h^r \equiv 1$ and it reduces to $\mathcal{L}_{\mathcal{D}}(\pi, f)$. Generalized OPAC runs Algorithm 1 with these losses in line 3.

Theorem 9 (Sample complexity of generalized OPAC). *Suppose Assumptions 1 to 4 hold with the generalized operator, $V_{\max}, L \geq 1$, $C_{sa}(\pi^\star) < \infty$, and $1 \leq \kappa_\mu(\sigma), \chi_\mu(\sigma) < \infty$. With suitable η and β, generalized OPAC run for K iterations returns $\bar{\pi}$ such that, with probability at least $1 - \delta$,*

$$\begin{aligned} J(\pi^\star) - J(\bar{\pi}) = \widetilde{O}\Bigg(& V_{\max}^2 L \sqrt{\frac{\kappa_\mu(\sigma) H^2 C_{sa}(\pi^\star)}{n}} + V_{\max}^2 L \sqrt{\frac{\chi_\mu(\sigma) H^4}{n}} \\ & + V_{\max} H \sqrt{\frac{\log|\mathcal{A}|}{K}} + \Phi \sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}} + \big(\sqrt{H}\,\Phi + H\big)\sqrt{\varepsilon_{\mathcal{F}}} \Bigg), \end{aligned} \tag{12}$$

where $\Phi = \widetilde{O}\big(V_{\max} L \sqrt{\kappa_\mu(\sigma) H C_{sa}(\pi^\star)} + V_{\max} L \sqrt{\chi_\mu(\sigma) H^3} + \sqrt{H C_{sa}(\pi^\star)}\big)$.

The proof, which also defines Φ, is in Appendix G. With exact realizability and completeness and $K \geq n$, only the two statistical terms remain, so $n = \widetilde{O}\big(V_{\max}^4 L^2 (\kappa_\mu(\sigma) H^2 C_{sa}(\pi^\star) + \chi_\mu(\sigma) H^4)/\varepsilon^2\big)$ trajectories suffice. The first arises when the learned reward is used along $\pi^\star$: the regret contains per-step reward errors under $d_h^{\pi^\star}$, which $\kappa_\mu(\sigma)$ converts into outcome-regression error and $C_{sa}(\pi^\star)$ transfers to the data. The second arises from pessimism: by item (4), the critic is the exact value of the policy under some $\tilde{r} \in \mathcal{R}$, whereas the policy-loss weights use the candidate reward r. Since $\mathcal{T}^\pi_{\tilde{r}} f = f$, Definition 2 bounds $\|\tilde{r} - r\|^2_{2,\mu}$ by $\chi_\mu(\sigma)$ times the Bellman error of (f, r), which controls the weight mismatch. This comparison is on the data distribution, so the second term has no $C_{sa}(\pi^\star)$.

6 Conclusion

We developed a sample-complexity theory of offline RL with trajectory-level supervision. For the cumulative-reward objective, OPAC attains the rate $\widetilde{O}(H^2 \sqrt{C_{sa}(\pi^\star)/n})$ from bounded, unbiased trajectory-level labels, and a lower bound matches its explicit horizon and coverage dependence. The same template extends to trajectory preferences. For nonlinear aggregations, we showed an exponential barrier in general and polynomial sample complexity for structured aggregations controlled by $\kappa_\mu(\sigma)$ and $\chi_\mu(\sigma)$. Together, these results delineate when trajectory-level supervision enables efficient offline policy optimization and when missing process-level rewards are a fundamental obstacle.

Limitations Our generalized framework assumes a known aggregation rule with Bellman-style structure, and it does not fully characterize which trajectory-level objectives are both statistically learnable and practically meaningful. Broader sufficient conditions for efficient outcome-based learning, and lower bounds in terms of $\kappa_\mu(\sigma)$ and $\chi_\mu(\sigma)$, remain to be identified.

AI USE STATEMENT

The theoretical ideas and original proofs were developed by the human authors. Generative AI tools assisted with drafting and polishing the manuscript, checking mathematical derivations, and revising proof exposition. The authors take responsibility for all mathematical claims and the final manuscript.

REPRODUCIBILITY STATEMENT

The problem definitions, assumptions, and algorithms are specified in Sections 2 to 5. The appendices provide the proofs, lower-bound constructions, and coefficient calculations for the stated results.

REFERENCES

Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. Monograph, 2022. URL https://rltheorybook.github.io/.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in Neural Information Processing Systems*, 30, 2017.

Marco Bagatella, Andreas Krause, and Georg Martius. Directed exploration in reinforcement learning from linear temporal logic. *Transactions on Machine Learning Research*, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, pp. 1753–1800. PMLR, 2023.

Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36259–36269, 2022.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36:66845–66859, 2023.

Alper Kamil Bozkurt, Yu Wang, Michael M Zavlanos, and Miroslav Pajic. Control synthesis from linear temporal logic specifications using model-free reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10349–10355. IEEE, 2020.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Asaf Cassel, Haipeng Luo, Aviv Rosenberg, and Dmitry Sotnikov. Near-optimal regret in linear MDPs with aggregate bandit feedback. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5757–5791, 2024.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. *Advances in Neural Information Processing Systems*, 34: 3401–3412, 2021.

Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Outcome-based online reinforcement learning: Algorithms and fundamental limits. In *Advances in Neural Information Processing Systems*, 2025.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pp. 3773–3793. PMLR, 2022.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Wei Cui and Wei Yu. Reinforcement learning with non-cumulative objective. *IEEE Transactions on Machine Learning in Communications and Networking*, 1:124–137, 2023.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7288–7295, 2021.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in Neural Information Processing Systems*, 23, 2010.

Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3489, 2022.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Noah Golowich and Ankur Moitra. The role of inherent Bellman error in offline reinforcement learning with linear function approximation. *Reinforcement Learning Journal*, 1:302–341, 2024.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

Sai Krishna Gottipati, Yashaswi Pathak, Rohan Nuttall, Sahir, Raviteja Chunduru, Ahmed Touati, Sriram Ganapathi Subramanian, Matthew E Taylor, and Sarath Chandar. Maximum reward formulation in reinforcement learning. *arXiv preprint arXiv:2010.03744*, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Shanyu Han, Yang Liu, and Xiang Yu. Risk-sensitive reinforcement learning based on convex scoring functions. *arXiv preprint arXiv:2505.04553*, 2025.

Ronald A Howard and James E Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Xiang Ji, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. Self-play with adversarial critic: Provable and scalable offline alignment for language models. *arXiv preprint arXiv:2406.04274*, 2024.

Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? Rethinking process supervision from a theoretical perspective. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27373–27398, 2025.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 40(4):570–596, 2025.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *Mathematics of Operations Research*, 48(3):1496–1521, 2023.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pp. 267–274, 2002.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Tal Lancewicki and Yishay Mansour. Near-optimal regret using policy optimization in online MDPs with aggregate bandit feedback. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32467–32491, 2025.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pp. 45–73. Springer, 2012.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Rémi Munos. Performance bounds in L_p-norm for approximate value iteration. *SIAM Journal on Control and Optimization*, 46(2):541–561, 2007.

Thanh Nguyen-Tang, Ming Yin, Sunil Gupta, Svetha Venkatesh, and Raman Arora. On instance-dependent bounds for offline reinforcement learning with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9310–9318, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Kian Hong Quah and Chai Quek. Maximum reward reinforcement learning: A non-cumulative reward criterion. *Expert Systems with Applications*, 31(2):351–359, 2006.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*, 2025.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pp. 19967–20025. PMLR, 2022.

Yuda Song, Yifei Zhou, Ayush Sekhari, J Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *International Conference on Learning Representations*, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Grigorii Veviurko, Wendelin Böhmer, and Mathijs de Weerdt. To the max: Reinventing reward in reinforcement learning. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49455–49470, 2024.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *International Conference on Machine Learning*, pp. 35864–35907. PMLR, 2023.

Kaiwen Wang, Dawen Liang, Nathan Kallus, and Wen Sun. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. *arXiv preprint arXiv:2403.06323*, 2024a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b.

Ruosong Wang, Hanrui Zhang, Devendra Singh Chaplot, Denis Garagić, and Ruslan Salakhutdinov. Planning with submodular objective functions. *arXiv preprint arXiv:2010.11863*, 2020a.

Ruosong Wang, Peilin Zhong, Simon S Du, Russ R Salakhutdinov, and Lin Yang. Planning with general objective functions: Going beyond total rewards. *Advances in Neural Information Processing Systems*, 33:14486–14497, 2020b.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021a.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 34:27395–27407, 2021b.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *International Conference on Machine Learning*, pp. 54715–54754. PMLR, 2024.

Tengyu Xu, Yue Wang, Shaofeng Zou, and Yingbin Liang. Provably efficient offline reinforcement learning with trajectory-wise reward. *IEEE Transactions on Information Theory*, 70(9):6481–6518, 2024.

Cambridge Yang, Michael Littman, and Michael Carbin. Computably continuous reinforcement-learning objectives are PAC-learnable. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10729–10736, 2023.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, 34:4065–4078, 2021.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

Yurun Yuan, Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Trajectory Bellman residual minimization: A simple value-based method for LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 13626–13640, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *International Conference on Learning Representations*, 2024.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvári, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

Appendix

A Related Work

Offline RL Theory Our work builds on theoretical results in offline RL; see Jiang & Xie (2025) for a recent survey. Pessimism in the face of uncertainty counteracts distribution shift by penalizing value estimates in regions poorly covered by data (Yin & Wang, 2021; Jin et al., 2021b; Rashidinejad et al., 2021; Xie et al., 2021a; Cheng et al., 2022; Blanchet et al., 2023). Closely related to our algorithmic approach are pessimistic and adversarially trained actor-critic methods for offline RL (Xie et al., 2021a; Zanette et al., 2021; Cheng et al., 2022; Ji et al., 2024). On the data coverage side, recent works have progressively weakened the required assumptions from all-policy concentrability to single-policy concentrability and partial coverage (Xie et al., 2021a; Zhan et al., 2022; Shi et al., 2022), including minimax sample-complexity results for tabular model-based offline RL (Rashidinejad et al., 2021; Xie et al., 2021b; Li et al., 2024). On the function approximation side, a key question is what structural conditions on the function class enable efficient learning, spanning realizability-only guarantees, unifying complexity measures such as the Bellman eluder dimension, fundamental impossibility results, instance-dependent bounds, and characterizations of inherent Bellman error (Xie & Jiang, 2021; Jin et al., 2021a; Foster et al., 2022; Nguyen-Tang et al., 2023; Jin et al., 2023; Golowich & Moitra, 2024).

Preference Learning A complementary line of theoretical work studies RL from preference feedback rather than scalar rewards. Empirically, reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) is used to train generative models from human preferences, and direct preference optimization (Rafailov et al., 2023) reformulates the RLHF objective as a supervised loss on preference data. On the theoretical side, Zhu et al. (2023) provide guarantees for RLHF under both pairwise and K-wise comparison models, Xiong et al. (2024) bridge theory and practice for iterative preference learning under KL regularization, and Zhan et al. (2024) establish guarantees for offline preference-based RL. Our preference setting in Section 4 adopts the standard Bradley–Terry–Luce comparison model (Bradley & Terry, 1952; Christiano et al., 2017) and operates entirely in the offline regime. Compared with these prior analyses, we maintain dependence on a single-policy state–action concentrability $C_{sa}(\pi^\star)$ rather than the trajectory-level concentrability that often appears in preference-based analyses, and our guarantee depends on the BTL model through constants α_{C} and c_{C} that capture preference identifiability and concentration on bounded return ranges.

Outcome-based RL Outcome-based RL—where the learner observes only trajectory-level signals rather than per-step rewards—has been studied under several names. A theoretical line on aggregate feedback and trajectory preferences (Efroni et al., 2021; Chatterji et al., 2021; Chen et al., 2022; Cassel et al., 2024; Lancewicki & Mansour, 2025) considers learners that receive episode-level rewards or comparisons, primarily in online RL. A second, more recent line studies outcome supervision under general function approximation, motivated by reasoning datasets and human-feedback alignment in large language models (Stiennon et al., 2020; Cobbe et al., 2021; Ouyang et al., 2022; Bai et al., 2022; Jaech et al., 2024; Guo et al., 2025): Jia et al. (2025) show that offline outcome data can be transformed into per-step pseudo-rewards via trajectory-level regression at a cost controlled by the single-policy state–action concentrability rather than the much larger trajectory-level concentrability; Yuan et al. (2025) propose trajectory Bellman residual minimization for LLM reasoning and sharpen the change-of-trajectory-measure inequality introduced by Jia et al. (2025); and Chen et al. (2025) establish an online optimism-based algorithm with $\widetilde{O}(C_{\mathrm{cov}}H^3/\varepsilon^2)$ sample complexity together with an exponential outcome-versus-process separation. Closest to our scalar-outcome setting, Xu et al. (2024) study offline RL in which each trajectory is labeled by the sum of its per-step rewards. Their PARTED algorithm redistributes trajectory returns into per-step proxy rewards by least squares and runs pessimistic value iteration with elliptical bonuses, giving $\widetilde{O}(dH^3/\sqrt{n})$ suboptimality for linear MDPs with feature dimension d and an analogous neural-tangent-kernel bound. Their guarantee assumes a well-explored dataset whose trajectory-level feature covariance is well conditioned, so that per-step rewards are identifiable from trajectory sums. This condition fails whenever constant functions lie in the feature span, because shifting reward mass between two steps leaves every trajectory sum unchanged. No lower bound is given. OPAC never identifies per-step

rewards: it controls only trajectory-level reward error through a change of trajectory measure, uses single-policy state–action coverage under general function approximation, and is complemented by a matching lower bound. Outcome-level feedback is also implicit in goal-reaching with sparse rewards (Andrychowicz et al., 2017), in clinical and healthcare RL where labels are revealed only at trajectory boundaries (Komorowski et al., 2018; Gottesman et al., 2019), and in trajectory-level human preferences (discussed above). Compared to these works, we focus on the *offline* regime under partial (single-policy) coverage and additionally study *generalized* (nonlinear) trajectory-level objectives, for which we give an exponential lower bound and a sufficient regime described by two coefficients.

Compared with PRM Methods An empirical literature trains reward models that score either intermediate steps (process reward models, PRMs) or completed trajectories (outcome reward models, ORMs), particularly for verifying language-model reasoning (Lightman et al., 2024; Uesato et al., 2022). Uesato et al. (2022) compare process- and outcome-based feedback at scale on math word problems, while Lightman et al. (2024) report that step-by-step process supervision can outperform outcome supervision on challenging math benchmarks. Jia et al. (2025) subsequently provide a theoretical perspective showing that, under suitable coverage and realizability conditions, outcome supervision can match process supervision up to polynomial-in-horizon factors. Our upper and lower bounds give the explicit $H^4 C_{sa}/\varepsilon^2$ trajectory dependence for bounded unbiased outcomes, up to class-complexity logarithms. In the comparison regime of Remark 1, the displayed horizon dependence is one factor of H above the process-supervised tabular benchmark. We further extend this analysis to BTL preferences and to nonlinear trajectory-level objectives, where the picture changes qualitatively: our structural and function-class assumptions yield an upper bound in terms of κ_μ and χ_μ, while all-success outcomes admit an exponential lower bound even under constant state–action coverage.

Generalized RL Objectives Different optimization objectives correspond to different ways of aggregating stage-wise signals into trajectory-level criteria. Theoretical lines in this direction include risk-sensitive RL (Howard & Matheson, 1972; Bastani et al., 2022; Wang et al., 2023; 2024a; Han et al., 2025), planning with submodular or other general (e.g., permutation-symmetric) trajectory objectives (Wang et al., 2020a;b), max-reward RL that replaces the cumulative return by the maximum per-step reward (Quah & Quek, 2006; Gottipati et al., 2020; Veviurko et al., 2024), dynamic-programming foundations for non-cumulative objectives that introduce parallel notions of generalized reward aggregation, value functions, and Bellman operators (Cui & Yu, 2023), optimization of state–action occupancy utilities beyond cumulative rewards under hidden convexity or general utilities where Bellman dynamic programming breaks down (Zhang et al., 2020; Barakat et al., 2023), control synthesis from linear temporal logic specifications (Bozkurt et al., 2020; Bagatella et al., 2025), and PAC learnability for computably continuous RL objectives (Yang et al., 2023). Most of these works analyze convergence of dynamic-programming or online policy-gradient procedures rather than offline finite-sample complexity.

Process- and outcome-supervised reward modeling for mathematical reasoning (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024b; Setlur et al., 2025) motivates studying different ways to aggregate stepwise feedback. Our all-success objective is one specified aggregation; its lower bound concerns the observation model and instance family in Theorem 5. We study offline finite-sample guarantees for known trajectory aggregations and give sufficient conditions in terms of $\kappa_\mu(\sigma)$ and $\chi_\mu(\sigma)$.

B Technical Lemmas for Outcome-Based Offline RL

Conventions Since the stages are disjoint, we identify $f \in \mathcal{F}$ with $(f_1, \ldots, f_H)$ and write $f(s,a)$ for $f_h(s,a)$ when $s \in \mathcal{S}_h$. For general state spaces, the ratio in Eq. (2) denotes the Radon–Nikodym derivative and the supremum is essential with respect to d_h^μ; $C_{sa}(\pi) = +\infty$ unless $d_h^\pi \ll d_h^\mu$ at every stage.

Population losses This appendix uses the following population loss notation throughout. For a policy $\pi \in \Pi$, a candidate reward $r \in \mathcal{R}$, and a critic $f \in \mathcal{F}$, define

$$\mathcal{L}_\mu(\pi, f) := \sum_{h=1}^{H} \mathbb{E}_\mu \left[f_h(s_h, \pi) - f_h(s_h, a_h) \right], \tag{13a}$$

$$\mathcal{L}_\mu^{\mathrm{BE}}(\pi, r, f) := \sum_{h=1}^{H} \mathbb{E}_\mu \left[\left((f - \mathcal{T}_r^\pi f)(s_h, a_h) \right)^2 \right], \tag{13b}$$

$$\mathcal{L}_\mu^{\mathrm{RM}}(r) := \mathbb{E}_\mu \left[\left(\sum_{h=1}^{H} (r_h - r_h^\star)(s_h, a_h) \right)^2 \right] = \mathbb{E}_\mu \left[(R(\tau; r) - R^\star(\tau))^2 \right]. \tag{13c}$$

B.1 Change-of-Measure and Optimization Tools

This subsection collects the two auxiliary tools used to control the change-of-measure terms and the policy-optimization term in the proof of the upper bound.

The following statement follows from Lemma 1 or Corollary 6 of Yuan et al. (2025).

Lemma 10 (Change of trajectory measure)**.** *For any finite-horizon MDP, Markov policies μ, π with $C_{sa}(\pi) < \infty$, and bounded measurable $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,*

$$\left(\mathbb{E}_\pi \left[\sum_{h=1}^{H} g(s_h, a_h) \right] \right)^2 \le H C_{sa}(\pi)\, \mathbb{E}_\mu \left[\left(\sum_{h=1}^{H} g(s_h, a_h) \right)^2 \right].$$

The next lemma is the standard exponential-weights no-regret bound (Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006), specialized to our layer-wise policy update; we include the proof for completeness.

Lemma 11 (No-regret bound)**.** *Let $\mathcal{A}$ be finite, with $|\mathcal{A}| \ge 2$. Let $\{f_k\}_{k=1}^K \subset \mathcal{F}$ be an arbitrary (possibly data-adaptive) sequence of critics such that, for every $k \in [K]$ and $h \in [H]$,*

$$f_{k,h} : \mathcal{S}_h \times \mathcal{A} \to [0, V_{\max}].$$

Set $\eta = V_{\max} \sqrt{K/(8 \log |\mathcal{A}|)}$ in the update below. Initialize

$$\pi_{1,h}(\cdot \mid s) = \mathrm{Unif}(\mathcal{A}), \qquad \forall h \in [H],\ s \in \mathcal{S}_h,$$

and for each $k = 1, \dots, K$ and $h \in [H]$, update

$$\pi_{k+1,h}(a \mid s) = \frac{\pi_{k,h}(a \mid s) \exp\big(f_{k,h}(s,a)/\eta\big)}{\sum_{a' \in \mathcal{A}} \pi_{k,h}(a' \mid s) \exp\big(f_{k,h}(s,a')/\eta\big)}. \tag{14}$$

For any comparator $\pi \in \Pi$, let $d_{h,S}^\pi(s) := \sum_{a \in \mathcal{A}_h} d_h^\pi(s,a)$ denote the step-h state marginal, and define

$$\mathrm{Reg}_k(\pi) := \sum_{h=1}^{H} \mathbb{E}_{s \sim d_{h,S}^\pi} \Big[\mathbb{E}_{a \sim \pi_h(\cdot|s)}[f_{k,h}(s,a)] - \mathbb{E}_{a \sim \pi_{k,h}(\cdot|s)}[f_{k,h}(s,a)] \Big].$$

Then, for every comparator $\pi \in \Pi$,

$$\frac{1}{K} \sum_{k=1}^{K} \mathrm{Reg}_k(\pi) \le H V_{\max} \sqrt{\frac{\log |\mathcal{A}|}{2K}}. \tag{15}$$

Proof. We prove a pointwise regret bound for each fixed pair (s,h), and then average over $s \sim d_{h,S}^\pi$ and sum over h.

Fix any $h \in [H]$ and $s \in \mathcal{S}_h$. Define

$$q_k(\cdot) := \pi_{k,h}(\cdot \mid s), \qquad g_k(\cdot) := f_{k,h}(s, \cdot) \in [0, V_{\max}]^{|\mathcal{A}|}.$$

Then Eq. (14) becomes the exponential-weights update

$$q_{k+1}(a) = \frac{q_k(a)\exp(g_k(a)/\eta)}{\sum_{a'\in\mathcal{A}} q_k(a')\exp(g_k(a')/\eta)}.$$

Introduce weights

$$w_1(a) = q_1(a) = \frac{1}{|\mathcal{A}|}, \qquad w_{k+1}(a) = w_k(a)\exp(g_k(a)/\eta),$$

and let $W_k := \sum_{a\in\mathcal{A}} w_k(a)$. Since $W_1 = 1$ and $q_k(a) = w_k(a)/W_k$,

$$\frac{W_{k+1}}{W_k} = \sum_{a\in\mathcal{A}} q_k(a)\exp(g_k(a)/\eta) = \mathbb{E}_{a\sim q_k}\left[\exp(g_k(a)/\eta)\right].$$

Upper bound on $\log W_{K+1}$ By Hoeffding's lemma, since $g_k(a)/\eta \in [0, V_{\max}/\eta]$ for all $a \in \mathcal{A}$,

$$\log \mathbb{E}_{a\sim q_k}\left[\exp(g_k(a)/\eta)\right] \le \frac{1}{\eta}\mathbb{E}_{a\sim q_k}[g_k(a)] + \frac{V_{\max}^2}{8\eta^2}.$$

Summing over $k = 1, \dots, K$ and using $W_1 = 1$,

$$\log W_{K+1} \le \frac{1}{\eta}\sum_{k=1}^{K}\mathbb{E}_{a\sim q_k}[g_k(a)] + \frac{V_{\max}^2 K}{8\eta^2}. \tag{16}$$

Lower bound on $\log W_{K+1}$

$$W_{K+1} = \sum_{a\in\mathcal{A}} q_1(a)\exp\left(\frac{1}{\eta}\sum_{k=1}^{K} g_k(a)\right).$$

Since $q_1 = \mathrm{Unif}(\mathcal{A})$ is strictly positive, we can write, for any $p \in \Delta(\mathcal{A})$,

$$W_{K+1} \ge \sum_{a:p(a)>0} p(a)\cdot\frac{q_1(a)}{p(a)}\exp\left(\frac{1}{\eta}\sum_{k=1}^{K} g_k(a)\right) = \mathbb{E}_{a\sim p}\left[\frac{q_1(a)}{p(a)}\exp\left(\frac{1}{\eta}\sum_{k=1}^{K} g_k(a)\right)\right].$$

Applying Jensen's inequality to the concave function $\log(\cdot)$ yields

$$\log W_{K+1} \ge \mathbb{E}_{a\sim p}\left[\log\frac{q_1(a)}{p(a)} + \frac{1}{\eta}\sum_{k=1}^{K} g_k(a)\right] = \frac{1}{\eta}\sum_{k=1}^{K}\mathbb{E}_{a\sim p}[g_k(a)] - \mathrm{KL}(p\|q_1).$$

Combining with Eq. (16) gives

$$\sum_{k=1}^{K}\left(\mathbb{E}_{a\sim p}[g_k(a)] - \mathbb{E}_{a\sim q_k}[g_k(a)]\right) \le \eta\mathrm{KL}(p\|q_1) + \frac{V_{\max}^2 K}{8\eta}.$$

Since $q_1 = \mathrm{Unif}(\mathcal{A})$,

$$\mathrm{KL}(p\|q_1) = \sum_{a\in\mathcal{A}} p(a)\log\left(p(a)|\mathcal{A}|\right) = \log|\mathcal{A}| + \sum_{a\in\mathcal{A}} p(a)\log p(a) \le \log|\mathcal{A}|,$$

where the last inequality uses $\sum_a p(a)\log p(a) \le 0$ for any probability distribution p. Hence, for every $p \in \Delta(\mathcal{A})$,

$$\sum_{k=1}^{K}\left(\mathbb{E}_{a\sim p}[g_k(a)] - \mathbb{E}_{a\sim q_k}[g_k(a)]\right) \le \eta\log|\mathcal{A}| + \frac{V_{\max}^2 K}{8\eta}. \tag{17}$$

Setting $p(\cdot) = \pi_h(\cdot \mid s)$, recalling $g_k(a) = f_{k,h}(s,a)$ and $q_k(\cdot) = \pi_{k,h}(\cdot \mid s)$, Eq. (17) becomes a deterministic (pointwise-in-s) inequality. Taking $\mathbb{E}_{s\sim d_{h,S}^{\pi}}[\cdot]$ preserves it, and summing over $h = 1, \dots, H$ yields the intermediate bound

$$\sum_{k=1}^{K}\mathrm{Reg}_k(\pi) \le H\left(\eta\log|\mathcal{A}| + \frac{V_{\max}^2 K}{8\eta}\right). \tag{18}$$

Finally, when $|\mathcal{A}| \geq 2$, AM–GM minimizes the right-hand side of Eq. (18) over $\eta > 0$ at

$$\eta^\star = V_{\max}\sqrt{\frac{K}{8\log|\mathcal{A}|}}, \qquad \eta^\star \log|\mathcal{A}| + \frac{V_{\max}^2 K}{8\eta^\star} = V_{\max}\sqrt{\frac{K\log|\mathcal{A}|}{2}},$$

and dividing by K gives Eq. (15). The case $|\mathcal{A}| = 1$ is trivial. □

B.2 Concentration Lemmas

This subsection collects the high-probability estimates used to transfer the population losses in the analysis to their empirical counterparts. When combining concentration bounds, we divide the total failure probability δ among their events; this changes only logarithmic constants.

Lemma 12 (Uniform concentration of the policy loss)**.** *Assume every $f = (f_1, \ldots, f_H) \in \mathcal{F}$ satisfies $f_h(s,a) \in [0, V_{\max}]$ for all $h \in [H]$ and $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, and that $\mathcal{D} = \{\tau_i\}_{i=1}^n$ consists of n i.i.d. trajectories drawn from μ. Fix $\delta \in (0,1)$ and define*

$$\varepsilon_{\mathrm{Perf}} := 2H^2 V_{\max}^2 \frac{\log(2|\Pi||\mathcal{F}|/\delta)}{n}. \tag{19}$$

Then with probability at least $1-\delta$, simultaneously for every $\pi \in \Pi$ and every $f \in \mathcal{F}$,

$$\big|\mathcal{L}_\mu(\pi, f) - \mathcal{L}_{\mathcal{D}}(\pi, f)\big| \leq \sqrt{\varepsilon_{\mathrm{Perf}}}. \tag{20}$$

Proof. Fix $(\pi, f) \in \Pi \times \mathcal{F}$. For each trajectory $\tau_i = (s_{i,1}, a_{i,1}, \ldots, s_{i,H}, a_{i,H})$, set

$$Z_i(\pi, f) := \sum_{h=1}^{H} \big[f_h(s_{i,h}, \pi(s_{i,h})) - f_h(s_{i,h}, a_{i,h})\big].$$

Since the τ_i are i.i.d. under μ, the variables $\{Z_i(\pi,f)\}_{i=1}^n$ are i.i.d., and by definition of $\mathcal{L}_\mu(\pi,f)$ and $\mathcal{L}_{\mathcal{D}}(\pi,f)$ in Eqs. (3a) and (13),

$$\mathbb{E}_{\tau_i \sim \mu}\big[Z_i(\pi,f)\big] = \mathcal{L}_\mu(\pi,f), \qquad \frac{1}{n}\sum_{i=1}^n Z_i(\pi,f) = \mathcal{L}_{\mathcal{D}}(\pi,f).$$

Moreover, since f_h takes values in $[0, V_{\max}]$, each summand satisfies

$$f_h(s_{i,h}, \pi(s_{i,h})) - f_h(s_{i,h}, a_{i,h}) \in [-V_{\max}, V_{\max}],$$

and therefore $|Z_i(\pi,f)| \leq HV_{\max}$ almost surely. Hoeffding's inequality applied to the i.i.d. bounded random variables $\{Z_i(\pi,f)\}_{i=1}^n$ gives, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\big|\mathcal{L}_\mu(\pi,f) - \mathcal{L}_{\mathcal{D}}(\pi,f)\big| \leq HV_{\max}\sqrt{\frac{2\log(2/\delta')}{n}}.$$

Taking a union bound over all pairs $(\pi, f) \in \Pi \times \mathcal{F}$ with $\delta' = \delta/(|\Pi||\mathcal{F}|)$, and noting that $HV_{\max}\sqrt{2\log(2|\Pi||\mathcal{F}|/\delta)/n} = \sqrt{\varepsilon_{\mathrm{Perf}}}$ by definition Eq. (19), yields Eq. (20) simultaneously for every $(\pi, f) \in \Pi \times \mathcal{F}$. □

Lemma 13 (Reward-model concentration via Bernstein)**.** *Let $\mathcal{D} = \{(\tau_i, Y_i)\}_{i=1}^n$ be n i.i.d. samples collected under μ, where*

$$Y_i = \sum_{h=1}^{H} r_h^\star(s_{i,h}, a_{i,h}) + \xi_i, \qquad \mathbb{E}[\xi_i \mid \tau_i] = 0, \qquad |\xi_i| \leq H \text{ a.s.},$$

Assume $r^\star \in \mathcal{R}$ and every $r \in \mathcal{R}$ satisfies $0 \leq r_h(s,a) \leq 1$. Define the population and empirical reward-model losses

$$\mathcal{L}_\mu^{\mathrm{RM}}(r) := \mathbb{E}_\mu\left[\Big(\textstyle\sum_{h=1}^H (r - r^\star)(s_h, a_h)\Big)^2\right],$$

$$\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}(r) := \frac{1}{n}\sum_{i=1}^n \Big(Y_i - \textstyle\sum_{h=1}^H r_h(s_{i,h}, a_{i,h})\Big)^2.$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $r \in \mathcal{R}$,

$$\mathcal{L}_\mu^{\mathrm{RM}}(r) \le 2\big[\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r) - \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star)\big] + \varepsilon_{\mathrm{RM}}, \tag{21}$$

where

$$\varepsilon_{\mathrm{RM}} = \frac{128}{3} \cdot \frac{H^2 \log(|\mathcal{R}|/\delta)}{n},$$

Proof. Fix any $r \in \mathcal{R}$ and write $\Delta_i := \sum_{h=1}^H (r - r^\star)(s_{i,h}, a_{i,h})$. Under the assumption $0 \le r_h, r_h^\star \le 1$ we have $|\Delta_i| \le H$. A direct expansion gives

$$\big(Y_i - \textstyle\sum_h r\big)^2 - \big(Y_i - \sum_h r^\star\big)^2 = (\Delta_i - \xi_i)^2 - \xi_i^2 = \Delta_i^2 - 2\Delta_i\xi_i,$$

so, letting $U_i(r) := \Delta_i^2 - 2\Delta_i\xi_i$,

$$\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r) - \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star) = \frac{1}{n}\sum_{i=1}^n U_i(r). \tag{22}$$

Mean. Since $\mathbb{E}[\xi_i \mid \tau_i] = 0$,

$$\mathbb{E}[U_i(r)] = \mathbb{E}[\Delta_i^2] - 2\mathbb{E}[\Delta_i \mathbb{E}[\xi_i \mid \tau_i]] = \mathbb{E}[\Delta_i^2] = \mathcal{L}_\mu^{\mathrm{RM}}(r). \tag{23}$$

Variance (self-bounding). Using $\mathbb{E}[\xi_i \mid \tau_i] = 0$, we have $\mathrm{Cov}(\Delta_i^2, \Delta_i\xi_i) = \mathbb{E}[\Delta_i^3 \mathbb{E}[\xi_i \mid \tau_i]] = 0$, so

$$\mathrm{Var}(U_i(r)) = \mathrm{Var}(\Delta_i^2) + 4\mathrm{Var}(\Delta_i\xi_i).$$

Each term is self-bounded by $\mathcal{L}_\mu^{\mathrm{RM}}(r)$:

$$\mathrm{Var}(\Delta_i^2) \le \mathbb{E}[\Delta_i^4] \le H^2\mathbb{E}[\Delta_i^2] = H^2\mathcal{L}_\mu^{\mathrm{RM}}(r),$$

and, using the tower property,

$$\mathrm{Var}(\Delta_i\xi_i) = \mathbb{E}[\Delta_i^2\xi_i^2] = \mathbb{E}[\Delta_i^2\mathbb{E}[\xi_i^2 \mid \tau_i]] \le H^2\mathcal{L}_\mu^{\mathrm{RM}}(r).$$

Combining,

$$\mathrm{Var}(U_i(r)) \le 5H^2\mathcal{L}_\mu^{\mathrm{RM}}(r). \tag{24}$$

Range. Since $|\xi_i| \le H$, $|U_i(r)| \le H^2 + 2H^2 = 3H^2$. We use the looser bound $|U_i(r)| \le M := 8H^2$, so that $|U_i(r) - \mathbb{E}U_i(r)| \le 2M$.

Bernstein and self-bounding. Apply one-sided Bernstein to the i.i.d. variables $\{U_i(r)\}_{i=1}^n$ with confidence $\eta \in (0,1)$, using Eq. (23)–Eq. (24) and the centered range bound $2M$: with probability at least $1-\eta$,

$$\mathcal{L}_\mu^{\mathrm{RM}}(r) - \big[\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r) - \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star)\big] \le \sqrt{\frac{10H^2\mathcal{L}_\mu^{\mathrm{RM}}(r)\log(1/\eta)}{n}} + \frac{2M\log(1/\eta)}{3n}.$$

Taking $\eta = \delta/|\mathcal{R}|$ and a union bound over $r \in \mathcal{R}$, set

$$x := \mathcal{L}_\mu^{\mathrm{RM}}(r), \qquad T := \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r) - \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star), \qquad \beta' := 4H^2\frac{\log(|\mathcal{R}|/\delta)}{n}.$$

Since $10H^2\log(|\mathcal{R}|/\delta)/n \le 8\beta'$ and $2M\log(|\mathcal{R}|/\delta)/(3n) = 4\beta'/3$, the Bernstein inequality implies

$$x \le T + \sqrt{8\beta' x} + \frac{4}{3}\beta'. \tag{25}$$

Finally, $\sqrt{8\beta' x} \le x/2 + 4\beta'$ by AM–GM. Substituting this into Eq. (25) and moving $x/2$ to the left gives

$$x \le 2T + \frac{32}{3}\beta' = 2T + \frac{128}{3} \cdot \frac{H^2\log(|\mathcal{R}|/\delta)}{n},$$

which is Eq. (21). □

For a critic f and policy π, write $V^{\pi}_{f,h}(s) := f_h(s, \pi_h)$ and set $V^{\pi}_{f,H+1} \equiv 0$. In the outcome-based setting, the Bellman error in Eq. (3c) is computed by plugging in the candidate reward r; no observed per-step reward is used.

Lemma 14 (Bellman-error concentration)**.** *Assume $\Pi, \mathcal{R}, \mathcal{F}$ are finite, every $r \in \mathcal{R}$ satisfies $0 \le r_h(s,a) \le 1$, every $f \in \mathcal{F}$ satisfies $0 \le f_h(s,a) \le V_{\max}$ with $V_{\max} \ge 1$, and uniform approximate Bellman completeness holds with error $\varepsilon_{\mathcal{F},\mathcal{F}}$ (Assumption 3). Then with probability at least $1-\delta$, the following holds simultaneously for all $(\pi, r, f) \in \Pi \times \mathcal{R} \times \mathcal{F}$:*

$$\mathcal{L}^{\mathrm{BE}}_{\mu}(\pi, r, f) \le 2\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi, r, f) + \varepsilon_{\mathrm{BE}} + 4\varepsilon_{\mathcal{F},\mathcal{F}}, \qquad \varepsilon_{\mathrm{BE}} := \frac{256}{3} \cdot \frac{H V_{\max}^2 \log\big(2|\Pi||\mathcal{R}||\mathcal{F}|/\delta\big)}{n}.$$

Proof. Fix any $(\pi, r, f) \in \Pi \times \mathcal{R} \times \mathcal{F}$. For any bounded g, define the population and empirical unminimized Bellman-regression residuals

$$L_{\mu}(g) := \sum_{h=1}^{H} \mathbb{E}_{\mu}\left[\big(g_h(s_h,a_h) - r_h(s_h,a_h) - f_{h+1}(s_{h+1},\pi)\big)^2\right],$$
$$L_{\mathcal{D}}(g) := \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{H}\big(g_h(s_{i,h},a_{i,h}) - r_h(s_{i,h},a_{i,h}) - f_{h+1}(s_{i,h+1},\pi)\big)^2.$$

The key observation is that $r_h(s_h,a_h)$ is deterministic given (s_h,a_h), so conditioning on (s_h,a_h),

$$\begin{aligned}\mathbb{E}\left[r_h(s_h,a_h) + V^{\pi}_{f,h+1}(s_{h+1})\middle| s_h,a_h\right] &= r_h(s_h,a_h) + \mathbb{E}[V^{\pi}_{f,h+1}(s_{h+1}) \mid s_h,a_h] \\ &= (\mathcal{T}^{\pi}_r f)_h(s_h,a_h),\end{aligned}$$

and moreover $\mathrm{Var}(r_h(s_h,a_h)+V^{\pi}_{f,h+1}(s_{h+1}) \mid s_h,a_h) = \mathrm{Var}(V^{\pi}_{f,h+1}(s_{h+1}) \mid s_h,a_h)$, since adding a (s_h,a_h)-measurable constant leaves the conditional variance unchanged. Hence the bias–variance decomposition gives

$$\begin{aligned}&\mathbb{E}\left[\big(g_h(s_h,a_h) - r_h(s_h,a_h) - V^{\pi}_{f,h+1}(s_{h+1})\big)^2\middle| s_h,a_h\right] \\ &\quad= \big(g_h(s_h,a_h) - (\mathcal{T}^{\pi}_r f)_h(s_h,a_h)\big)^2 + \mathrm{Var}\left(V^{\pi}_{f,h+1}(s_{h+1})\middle| s_h,a_h\right).\end{aligned}$$

Taking expectation over $(s_h,a_h) \sim d^{\mu}_h$ and summing over $h = 1,\dots,H$, we obtain

$$L_{\mu}(g) = \|g - \mathcal{T}^{\pi}_r f\|^2_{2,\mu} + \sum_{h=1}^{H} \mathbb{E}_{d^{\mu}_h}\left[\mathrm{Var}\left(V^{\pi}_{f,h+1}(s_{h+1})\middle| s_h,a_h\right)\right]. \tag{26}$$

The second term does not depend on g. Hence, setting $g = f$ and $g = \mathcal{T}^{\pi}_r f$ in Eq. (26), we obtain

$$L_{\mu}(f) - L_{\mu}(\mathcal{T}^{\pi}_r f) = \|f - \mathcal{T}^{\pi}_r f\|^2_{2,\mu} = \mathcal{L}^{\mathrm{BE}}_{\mu}(\pi, r, f), \tag{27}$$

For each trajectory i, define

$$Z_i(\pi, r, f) := \sum_{h=1}^{H}\Big[\ell_{i,h}(f; \pi, r, f) - \ell_{i,h}(\mathcal{T}^{\pi}_r f; \pi, r, f)\Big],$$

where

$$\ell_{i,h}(g; \pi, r, f) := \big(g_h(s_{i,h},a_{i,h}) - r_h(s_{i,h},a_{i,h}) - V^{\pi}_{f,h+1}(s_{i,h+1})\big)^2.$$

Since the trajectories are i.i.d., the variables $\{Z_i(\pi, r, f)\}_{i=1}^{n}$ are i.i.d. Moreover, by Eq. (27),

$$\mathbb{E}[Z_i(\pi, r, f)] = L_{\mu}(f) - L_{\mu}(\mathcal{T}^{\pi}_r f) = \mathcal{L}^{\mathrm{BE}}_{\mu}(\pi, r, f). \tag{28}$$

We now establish a self-bounding variance inequality. For each (i,h), define

$$A_{i,h} := f_h(s_{i,h},a_{i,h}) - r_h(s_{i,h},a_{i,h}) - V^{\pi}_{f,h+1}(s_{i,h+1}),$$
$$B_{i,h} := (\mathcal{T}^{\pi}_r f)_h(s_{i,h},a_{i,h}) - r_h(s_{i,h},a_{i,h}) - V^{\pi}_{f,h+1}(s_{i,h+1}).$$

Then

$$\ell_{i,h}(f;\pi,r,f) - \ell_{i,h}(\mathcal{T}_r^\pi f;\pi,r,f) = A_{i,h}^2 - B_{i,h}^2.$$

Since

$$0 \le f_h(s,a) \le V_{\max}, \qquad 0 \le (\mathcal{T}_r^\pi f)_h(s,a) \le V_{\max}+1 \le 2V_{\max},$$
$$0 \le r_h(s,a) \le 1, \qquad 0 \le V_{f,h+1}^\pi(s') \le V_{\max},$$

we have

$$|A_{i,h}|,\ |B_{i,h}| \le 2V_{\max}.$$

Therefore

$$\begin{aligned}|A_{i,h}^2 - B_{i,h}^2| &= |A_{i,h} - B_{i,h}| \cdot |A_{i,h} + B_{i,h}| \\ &\le 4V_{\max}|A_{i,h} - B_{i,h}|.\end{aligned}$$

Since

$$A_{i,h} - B_{i,h} = f_h(s_{i,h},a_{i,h}) - (\mathcal{T}_r^\pi f)_h(s_{i,h},a_{i,h}),$$

it follows that

$$\big(A_{i,h}^2 - B_{i,h}^2\big)^2 \le 16V_{\max}^2\big(f_h(s_{i,h},a_{i,h}) - (\mathcal{T}_r^\pi f)_h(s_{i,h},a_{i,h})\big)^2. \tag{29}$$

Using Cauchy–Schwarz $(\sum_{h=1}^H a_h)^2 \le H\sum_{h=1}^H a_h^2$,

$$\begin{aligned}Z_i(\pi,r,f)^2 &= \left(\sum_{h=1}^H \big[\ell_{i,h}(f;\pi,r,f) - \ell_{i,h}(\mathcal{T}_r^\pi f;\pi,r,f)\big]\right)^2 \\ &\le H\sum_{h=1}^H \big[\ell_{i,h}(f;\pi,r,f) - \ell_{i,h}(\mathcal{T}_r^\pi f;\pi,r,f)\big]^2 \\ &\le 16HV_{\max}^2\sum_{h=1}^H \big(f_h(s_{i,h},a_{i,h}) - (\mathcal{T}_r^\pi f)_h(s_{i,h},a_{i,h})\big)^2,\end{aligned}$$

where the last step used Eq. (29). Taking expectation,

$$\begin{aligned}\mathrm{Var}(Z_i(\pi,r,f)) &\le \mathbb{E}[Z_i(\pi,r,f)^2] \\ &\le 16HV_{\max}^2\sum_{h=1}^H \mathbb{E}_{d_h^\mu}\Big[\big(f_h(s_h,a_h) - (\mathcal{T}_r^\pi f)_h(s_h,a_h)\big)^2\Big] \\ &= 16HV_{\max}^2\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f).\end{aligned} \tag{30}$$

Also, since $|A_{i,h}|, |B_{i,h}| \le 2V_{\max}$, each per-step gap satisfies

$$\big|\ell_{i,h}(f;\pi,r,f) - \ell_{i,h}(\mathcal{T}_r^\pi f;\pi,r,f)\big| \le 8V_{\max}^2, \tag{31}$$

and summing over $h = 1,\dots,H$ gives

$$|Z_i(\pi,r,f)| \le 8HV_{\max}^2. \tag{32}$$

Applying Bernstein's inequality to the i.i.d. variables $\{Z_i(\pi,r,f)\}_{i=1}^n$, we obtain that for any fixed triple (π,r,f) and any $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f) \le \frac{1}{n}\sum_{i=1}^n Z_i(\pi,r,f) + \sqrt{\frac{32HV_{\max}^2\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f)\log(1/\delta')}{n}} + \frac{16HV_{\max}^2\log(1/\delta')}{3n}. \tag{33}$$

Setting $\delta' = \delta/(2|\Pi||\mathcal{R}||\mathcal{F}|)$ and taking a union bound over all triples $(\pi,r,f) \in \Pi\times\mathcal{R}\times\mathcal{F}$, with probability at least $1-\delta/2$, simultaneously for all such triples,

$$\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f) \le \frac{1}{n}\sum_{i=1}^n Z_i(\pi,r,f) + \sqrt{\frac{32HV_{\max}^2\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f)\log(2|\Pi||\mathcal{R}||\mathcal{F}|/\delta)}{n}}$$

$$+\frac{16HV_{\max}^2\log(2|\Pi||\mathcal{R}||\mathcal{F}|/\delta)}{3n}. \tag{34}$$

We now convert $\frac{1}{n}\sum_i Z_i(\pi,r,f) = L_{\mathcal{D}}(f) - L_{\mathcal{D}}(\mathcal{T}_r^\pi f)$ into an empirical quantity computable from $\mathcal{F}$. Let

$$g^\star(\pi,r,f) \in \operatorname*{argmin}_{g\in\mathcal{F}} \|g - \mathcal{T}_r^\pi f\|_{2,\mu}^2,$$

so that by completeness, $\|g^\star - \mathcal{T}_r^\pi f\|_{2,\mu}^2 \le \varepsilon_{\mathcal{F},\mathcal{F}}$. Since $g^\star \in \mathcal{F}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n Z_i(\pi,r,f) &= \big[L_{\mathcal{D}}(f) - L_{\mathcal{D}}(g^\star)\big] + \big[L_{\mathcal{D}}(g^\star) - L_{\mathcal{D}}(\mathcal{T}_r^\pi f)\big] \\ &\le \underbrace{\big[L_{\mathcal{D}}(f) - \min_{g\in\mathcal{F}} L_{\mathcal{D}}(g)\big]}_{=\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi,r,f)} + \underbrace{\big[L_{\mathcal{D}}(g^\star) - L_{\mathcal{D}}(\mathcal{T}_r^\pi f)\big]}_{=:T(\pi,r,f)},\end{aligned} \tag{35}$$

where we suppress the arguments (π,r,f) in $L_{\mathcal{D}}(\cdot)$ for brevity. The bias term $T(\pi,r,f)$ is handled by the same Bernstein self-bounding argument applied to the pair $(g^\star, \mathcal{T}_r^\pi f)$ in place of $(f, \mathcal{T}_r^\pi f)$: replaying Eq. (29)–Eq. (32) with $g^\star$ instead of f gives

$$\mathbb{E}[T(\pi,r,f)] = \|g^\star - \mathcal{T}_r^\pi f\|_{2,\mu}^2 \le \varepsilon_{\mathcal{F},\mathcal{F}},$$
$$\mathrm{Var}\big(T(\pi,r,f) \text{ per sample}\big) \le 16HV_{\max}^2\|g^\star - \mathcal{T}_r^\pi f\|_{2,\mu}^2 \le 16HV_{\max}^2\varepsilon_{\mathcal{F},\mathcal{F}}.$$

Writing $b := 4HV_{\max}^2\log(2|\Pi||\mathcal{R}||\mathcal{F}|/\delta)/n$, Bernstein's inequality (union-bounded over $(\pi,r,f) \in \Pi\times\mathcal{R}\times\mathcal{F}$, which also fixes $g^\star$) yields, with probability at least $1-\delta/2$,

$$T(\pi,r,f) \le \varepsilon_{\mathcal{F},\mathcal{F}} + \sqrt{8b\varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{4}{3}b \le 2\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{16}{3}b, \tag{36}$$

where the last step uses AM–GM $\sqrt{8b\varepsilon_{\mathcal{F},\mathcal{F}}} \le \varepsilon_{\mathcal{F},\mathcal{F}} + 2b$ and then loosens the constant.

Combining Eq. (34), Eq. (35) and Eq. (36), and writing

$$x := \mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f), \qquad M := \mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi,r,f),$$

we obtain

$$x \le M + 2\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{16}{3}b + \sqrt{8bx} + \frac{4}{3}b.$$

Using AM–GM, $\sqrt{8bx} \le x/2 + 4b$, so

$$x \le M + 2\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{x}{2} + 4b + \frac{16}{3}b + \frac{4}{3}b = M + 2\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{x}{2} + \frac{32}{3}b,$$

and rearranging yields $x \le 2M + 4\varepsilon_{\mathcal{F},\mathcal{F}} + \frac{64}{3}b$. Equivalently,

$$\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f) \le 2\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi,r,f) + \varepsilon_{\mathrm{BE}} + 4\varepsilon_{\mathcal{F},\mathcal{F}},$$

as claimed. □

B.3 Approximation and Comparator Lemmas

This subsection collects the approximation statements associated with the comparator pair $(f_\pi, r^\star)$ used in the pessimism step. We use the Bellman operator $\mathcal{T}_{r^\star}^\pi$ from Eq. (1). Here $V_{\max}$ is the critic bound of Section 2, and $\varepsilon_{\mathcal{F}}$ is the critic-realizability error in Assumption 2.

Lemma 15 (Adapted from Theorem 8 in Cheng et al. (2022))**.** *For any $\pi \in \Pi$, let f_π be defined as follows,*

$$f_\pi := \operatorname*{argmin}_{f\in\mathcal{F}} \sup_{h,\nu} \|f_h - (\mathcal{T}_{r^\star}^\pi f)_h\|_{2,\nu}^2.$$

Under Assumption 2, the following inequality holds with probability at least $1-\delta$ for all $\pi \in \Pi$:

$$\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi,r^\star,f_\pi) \le \varepsilon_{\mathrm{apx}}, \qquad \varepsilon_{\mathrm{apx}} := O\left(\frac{HV_{\max}^2\log(|\mathcal{F}||\Pi|/\delta)}{n} + H\varepsilon_{\mathcal{F}}\right).$$

Proof. Fix π and write $f = f_\pi$, $m = \mathcal{T}^\pi_{r^\star} f$, and $e(g) = \|g - m\|^2_{2,\mu}$. Let $L_\mathcal{D}(g)$ be the unminimized Bellman squared loss with target $r^\star + f_{h+1}(\cdot, \pi)$. The squared-loss differences have mean $e(g)$ and second moment $O(HV^2_{\max} e(g))$, as in Eq. (29)–Eq. (30). Bernstein's inequality and a union bound over $(\pi, g) \in \Pi \times \mathcal{F}$ therefore give, simultaneously,

$$\tfrac{1}{2} e(g) - t \le L_\mathcal{D}(g) - L_\mathcal{D}(m) \le 2e(g) + t, \qquad t = O\left(\frac{HV^2_{\max} \log(2|\Pi||\mathcal{F}|/\delta)}{n} \right).$$

Since $e(f_\pi) \le H\varepsilon_\mathcal{F}$ and $e(g) \ge 0$,

$$\mathcal{L}^{\mathrm{BE}}_\mathcal{D}(\pi, r^\star, f_\pi) = L_\mathcal{D}(f_\pi) - \min_{g \in \mathcal{F}} L_\mathcal{D}(g) \le 2H\varepsilon_\mathcal{F} + 2t.$$

□

Lemma 16 (Finite-horizon performance difference lemma; Kakade & Langford, 2002)**.** *For any fixed per-step reward function r and any policies $\pi, \pi' \in \Pi$, let $J_r(\pi) := \mathbb{E}_{\tau\sim\pi}[R(\tau; r)]$. Then*

$$J_r(\pi) - J_r(\pi') = \sum_{h=1}^{H} \mathbb{E}_{s_h, a_h \sim d^\pi_h} \left[Q^{\pi', r}_h(s_h, a_h) - V^{\pi', r}_h(s_h) \right].$$

In particular, when $r = r^\star$, this reduces to the same identity under the shorthand notation introduced in Section 2.

Lemma 17 (Finite-horizon analogue of Eq. (16) in Cheng et al. (2022))**.** *Under Assumption 2, for every $\pi_k \in \Pi$,*

$$\left| \mathcal{L}_\mu(\pi_k, Q^{\pi_k}) - \mathcal{L}_\mu(\pi_k, f_{\pi_k}) \right| \le 2H\sqrt{\varepsilon_\mathcal{F}} = O(H\sqrt{\varepsilon_\mathcal{F}}).$$

Proof. Write $f := f_{\pi_k}$, $\pi := \pi_k$ and $\mathcal{T}^\pi := \mathcal{T}^\pi_{r^\star}$ throughout the proof. Define the *fake per-step reward induced by* (f, π),

$$\widetilde{r}_h(s_h, a_h) := f_h(s_h, a_h) - \mathbb{E}_{s_{h+1}\sim P(\cdot|s_h,a_h)} \left[f_{h+1}(s_{h+1}, \pi(s_{h+1})) \right], \qquad h \in [H],$$

so that $R(\tau; \widetilde{r}) = \sum_{h=1}^H \widetilde{r}_h(s_h, a_h)$, with $f_{H+1} \equiv 0$. By construction f satisfies the Bellman equation under reward $\widetilde{r}$, so $f_h = Q^{\pi, \widetilde{r}}_h$ and, letting $J_{\widetilde{r}}(\cdot)$ denote the return in the MDP with reward $\widetilde{r}$,

$$J_{\widetilde{r}}(\pi) = f_1(s_1, \pi), \qquad J_{\widetilde{r}}(\mu) = \mathbb{E}_{\tau\sim\mu}[R(\tau; \widetilde{r})]. \tag{37}$$

Step 1: two applications of the performance difference lemma. By Lemma 16 applied in the true MDP,

$$\mathcal{L}_\mu(\pi, Q^\pi) = \sum_{h=1}^{H} \mathbb{E}_{d^\mu_h} \left[V^\pi_h(s_h) - Q^\pi_h(s_h, a_h) \right] = J(\pi) - J(\mu). \tag{38}$$

By Lemma 16 applied in the fake MDP (with reward $\widetilde{r}$, same dynamics) and using Eq. (37),

$$\mathcal{L}_\mu(\pi, f) = \sum_{h=1}^{H} \mathbb{E}_{d^\mu_h} \left[f_h(s_h, \pi) - f_h(s_h, a_h) \right] = J_{\widetilde{r}}(\pi) - J_{\widetilde{r}}(\mu) = f_1(s_1, \pi) - \mathbb{E}_\mu[R(\tau; \widetilde{r})]. \tag{39}$$

Subtracting Eq. (39) from Eq. (38),

$$\mathcal{L}_\mu(\pi, Q^\pi) - \mathcal{L}_\mu(\pi, f) = \underbrace{\left[J(\pi) - f_1(s_1, \pi) \right]}_{=:A} + \underbrace{\left[\mathbb{E}_\mu[R(\tau; \widetilde{r})] - J(\mu) \right]}_{=:B}. \tag{40}$$

Step 2: express A and B as Bellman residuals of f. By the definition of $\widetilde{r}$ and $\mathcal{T}^\pi f_h(s_h, a_h) = r^\star(s_h, a_h) + \mathbb{E}_{s_{h+1}}[f_{h+1}(s_{h+1}, \pi)]$,

$$r^\star(s_h, a_h) - \widetilde{r}_h(s_h, a_h) = \mathcal{T}^\pi f_h(s_h, a_h) - f_h(s_h, a_h).$$

Taking expectations under d^π_h and d^μ_h respectively and summing over h,

$$A = \mathbb{E}_{\tau\sim\pi}[R^\star(\tau) - R(\tau; \widetilde{r})] = \sum_{h=1}^{H} \mathbb{E}_{d^\pi_h} \left[(\mathcal{T}^\pi f_h - f_h)(s_h, a_h) \right], \tag{41}$$

$$B = \mathbb{E}_{\tau\sim\mu}[R(\tau;\widetilde{r}) - R^\star(\tau)] = \sum_{h=1}^{H} \mathbb{E}_{d_h^\mu}\big[(f_h - \mathcal{T}^\pi f_h)(s_h, a_h)\big]. \tag{42}$$

Step 3: bound via Assumption 2. Combining Eq. (40)–Eq. (42) and applying Jensen's inequality ($\|\cdot\|_{1,\nu} \le \|\cdot\|_{2,\nu}$) term-by-term,

$$\big|\mathcal{L}_\mu(\pi, Q^\pi) - \mathcal{L}_\mu(\pi, f)\big| \le \sum_{h=1}^{H} \big\|f_h - \mathcal{T}^\pi f_h\big\|_{2,d_h^\pi} + \sum_{h=1}^{H} \big\|f_h - \mathcal{T}^\pi f_h\big\|_{2,d_h^\mu}.$$

Since $f = f_\pi$ attains the argmin in the definition of f_π, Assumption 2 gives

$$\sup_{h,\nu} \|f_h - (\mathcal{T}^\pi f)_h\|_{2,\nu}^2 \le \varepsilon_{\mathcal{F}}.$$

Taking $\nu = d_h^\pi$ and $\nu = d_h^\mu$ gives $\|f_h - \mathcal{T}^\pi f_h\|_{2,\nu}^2 \le \varepsilon_{\mathcal{F}}$ for each $h \in [H]$ and each $\nu \in \{d_h^\pi, d_h^\mu\}$. Summing $2H$ such terms gives

$$\big|\mathcal{L}_\mu(\pi, Q^\pi) - \mathcal{L}_\mu(\pi, f)\big| \le 2H\sqrt{\varepsilon_{\mathcal{F}}},$$

which is the claim. □

We recall the policy-loss notation used in the decomposition. For the behavior policy μ,

$$\mathcal{L}_\mu(\pi, f) \coloneqq \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\big[f_h(s_h, \pi(s_h)) - f_h(s_h, a_h)\big].$$

Lemma 18 (Regret decomposition)**.** *Let π be an arbitrary competitor policy, $\widehat{\pi} \in \Pi$ be some learned policy, $\widehat{r} \in \mathcal{R}$ be some learned reward function, and $f \in \mathcal{F}$ be an arbitrary function over $\mathcal{S} \times \mathcal{A}$. Then we have the following decomposition:*

$$\begin{aligned}
J(\pi) - J(\widehat{\pi}) = &\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\big[\mathcal{T}_{\widehat{r}}^{\widehat{\pi}} f_h(s_h, a_h) - f_h(s_h, a_h)\big] \\
&+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\big[f_h(s_h, a_h) - \mathcal{T}_{\widehat{r}}^{\widehat{\pi}} f_h(s_h, a_h)\big] \\
&+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\big[f_h(s_h, a_h) - f_h(s_h, \widehat{\pi}(s_h))\big] - \mathcal{L}_\mu(\widehat{\pi}, Q^{\widehat{\pi}}) + \mathcal{L}_\mu(\widehat{\pi}, f) \\
&+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\big[r^\star(s_h, a_h) - \widehat{r}(s_h, a_h)\big] + \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\big[\widehat{r}(s_h, a_h) - r^\star(s_h, a_h)\big].
\end{aligned}$$

Proof. Define $r_h^{f,\widehat{\pi}}(s_h, a_h) \coloneqq f_h(s_h, a_h) - \mathbb{E}_{s_{h+1}}[f_{h+1}(s_{h+1}, \widehat{\pi})]$. Then $r^{f,\widehat{\pi}} = (r_1^{f,\widehat{\pi}}, \ldots, r_H^{f,\widehat{\pi}})$ is a fake reward function given f and $\widehat{\pi}$, with trajectory return $R(\tau; r^{f,\widehat{\pi}})$. We use the subscript $(\cdot)_{r^{f,\widehat{\pi}}}$ to denote functions or operators under the auxiliary MDP $\mathcal{M}_{r^{f,\widehat{\pi}}}$, which shares the dynamics of $\mathcal{M}$ but uses reward $r^{f,\widehat{\pi}}$. For any fixed policy π, these two MDPs induce the same occupancies d^π, since they have the same transition kernel.

Since f satisfies the Bellman equation with respect to $\mathcal{M}_{r^{f,\widehat{\pi}}}$, i.e., $f_h(s_h, a_h) = r_h^{f,\widehat{\pi}}(s_h, a_h) + \mathbb{E}_{s_{h+1}}[f_{h+1}(s_{h+1}, \widehat{\pi})]$, we know $f_h = Q_h^{\widehat{\pi}, r^{f,\widehat{\pi}}}$, for $h \in [H]$.

We perform a performance decomposition:

$$J(\pi) - J(\widehat{\pi}) = J(\pi) - J(\mu) - (J(\widehat{\pi}) - J(\mu)). \tag{43}$$

By the performance difference lemma (Lemma 16),

$$\begin{aligned}
J(\widehat{\pi}) - J(\mu) &= -\big(J(\mu) - J(\widehat{\pi})\big) \\
&= -\left(\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\big[Q_h^{\widehat{\pi}}(s_h, a_h) - V_h^{\widehat{\pi}}(s_h)\big]\right)
\end{aligned}$$

$$= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[V_h^{\widehat{\pi}}(s_h) - Q_h^{\widehat{\pi}}(s_h,a_h)\right]$$
$$= \mathcal{L}_\mu(\widehat{\pi}, Q^{\widehat{\pi}}).$$

Define $\Delta(\widehat{\pi}) \coloneqq \mathcal{L}_\mu(\widehat{\pi}, Q^{\widehat{\pi}}) - \mathcal{L}_\mu(\widehat{\pi}, f)$, then we can rewrite the second term of Eq. (43) as

$$\begin{aligned}
J(\widehat{\pi}) - J(\mu) &= \mathcal{L}_\mu(\widehat{\pi}, Q^{\widehat{\pi}}) \\
&= \Delta(\widehat{\pi}) + \mathcal{L}_\mu(\widehat{\pi}, f) \\
&= \Delta(\widehat{\pi}) + \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[f_h(s_h,\widehat{\pi}(s_h)) - f_h(s_h,a_h)\right] \\
&= \Delta(\widehat{\pi}) + \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[V_h^{\widehat{\pi},r^{f,\widehat{\pi}}}(s_h) - Q_h^{\widehat{\pi},r^{f,\widehat{\pi}}}(s_h,a_h)\right] \\
&= \Delta(\widehat{\pi}) + J_{r^{f,\widehat{\pi}}}(\widehat{\pi}) - J_{r^{f,\widehat{\pi}}}(\mu) && \text{(Lemma 16)} \\
&= \Delta(\widehat{\pi}) + V_1^{\widehat{\pi},r^{f,\widehat{\pi}}}(s_1) - \mathbb{E}_{\tau\sim\mu}[R(\tau; r^{f,\widehat{\pi}})] \\
&= \Delta(\widehat{\pi}) + f_1(s_1,\widehat{\pi}) - \mathbb{E}_{\tau\sim\mu}[R(\tau; r^{f,\widehat{\pi}})].
\end{aligned}$$

Substituting this identity into Eq. (43) gives

$$J(\pi) - J(\widehat{\pi}) = \big(J(\pi) - f_1(s_1,\widehat{\pi})\big) + \big(\mathbb{E}_{\tau\sim\mu}[R(\tau; r^{f,\widehat{\pi}})] - J(\mu)\big) - \Delta(\widehat{\pi}). \tag{44}$$

By the definition of the expected return $J(\mu)$,

$$\begin{aligned}
\mathbb{E}_{\tau\sim\mu}[R(\tau; r^{f,\widehat{\pi}})] - J(\mu) &= \mathbb{E}_{\tau\sim\mu}\left[R(\tau; r^{f,\widehat{\pi}}) - R^\star(\tau)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[r^{f,\widehat{\pi}}(s_h,a_h) - r^\star(s_h,a_h)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}[f_h(s_h,a_h) - \mathbb{E}_{s_{h+1}}[f_{h+1}(s_{h+1},\widehat{\pi})] - r^\star(s_h,a_h)] \\
&\qquad \text{(Definition of } r^{f,\widehat{\pi}}\text{)} \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[f_h(s_h,a_h) - (\mathcal{T}^{\widehat{\pi}} f)_h(s_h,a_h)\right]. && (45)
\end{aligned}$$

Since $J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{d_h^\pi}[r^\star(s_h,a_h)] = \mathbb{E}_{\tau\sim\pi}[R^\star(\tau)]$,

$$\begin{aligned}
J(\pi) - f_1(s_1,\widehat{\pi}) &= (J(\pi) - \mathbb{E}_{\tau\sim\pi}[R(\tau; r^{f,\widehat{\pi}})]) + (\mathbb{E}_{\tau\sim\pi}[R(\tau; r^{f,\widehat{\pi}})] - f_1(s_1,\widehat{\pi})) \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}[r^\star(s_h,a_h) - r^{f,\widehat{\pi}}(s_h,a_h)] + J_{r^{f,\widehat{\pi}}}(\pi) - J_{r^{f,\widehat{\pi}}}(\widehat{\pi}) \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\left[\mathcal{T}^{\widehat{\pi}} f(s_h,a_h) - f(s_h,a_h)\right] \\
&\quad + \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\left[Q_h^{\widehat{\pi},r^{f,\widehat{\pi}}}(s_h,a_h) - V_h^{\widehat{\pi},r^{f,\widehat{\pi}}}(s_h)\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\left[\mathcal{T}^{\widehat{\pi}} f(s_h,a_h) - f(s_h,a_h)\right] \\
&\quad + \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\left[f_h(s_h,a_h) - f_h(s_h,\widehat{\pi}(s_h))\right]. && (46)
\end{aligned}$$

Putting Eq. (45) and Eq. (46) into Eq. (44), we get

$$\begin{aligned} J(\pi) - J(\widehat{\pi}) = &\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[\mathcal{T}^{\widehat{\pi}} f_h(s_h,a_h) - f_h(s_h,a_h)\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[f_h(s_h,a_h) - f_h(s_h,\widehat{\pi}(s_h))\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\mu}}\left[f_h(s_h,a_h) - \mathcal{T}^{\widehat{\pi}} f_h(s_h,a_h)\right] - \mathcal{L}_{\mu}(\widehat{\pi}, Q^{\widehat{\pi}}) + \mathcal{L}_{\mu}(\widehat{\pi}, f). \end{aligned} \tag{47}$$

For a given learned reward function $\widehat{r}$, we can write

$$\begin{aligned} \mathcal{T}^{\widehat{\pi}} f_h(s_h,a_h) &= r^{\star}(s_h,a_h) + \mathbb{E}_{s_{h+1}\sim P(\cdot|s_h,a_h)}\left[f_{h+1}(s_{h+1},\widehat{\pi}(s_{h+1}))\right] \\ &= \widehat{r}(s_h,a_h) + \mathbb{E}_{s_{h+1}\sim P(\cdot|s_h,a_h)}\left[f_{h+1}(s_{h+1},\widehat{\pi}(s_{h+1}))\right] + r^{\star}(s_h,a_h) - \widehat{r}(s_h,a_h) \\ &= \mathcal{T}^{\widehat{\pi}}_{\widehat{r}} f_h(s_h,a_h) + r^{\star}(s_h,a_h) - \widehat{r}(s_h,a_h). \end{aligned}$$

Plugging this into Eq. (47), we get

$$\begin{aligned} J(\pi) - J(\widehat{\pi}) = &\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[\mathcal{T}^{\widehat{\pi}}_{\widehat{r}} f_h(s_h,a_h) - f_h(s_h,a_h)\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\mu}}\left[f_h(s_h,a_h) - \mathcal{T}^{\widehat{\pi}}_{\widehat{r}} f_h(s_h,a_h)\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[f_h(s_h,a_h) - f_h(s_h,\widehat{\pi}(s_h))\right] - \mathcal{L}_{\mu}(\widehat{\pi}, Q^{\widehat{\pi}}) + \mathcal{L}_{\mu}(\widehat{\pi}, f) \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[r^{\star}(s_h,a_h) - \widehat{r}(s_h,a_h)\right] + \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\mu}}\left[\widehat{r}(s_h,a_h) - r^{\star}(s_h,a_h)\right]. \end{aligned}$$

□

C Proof of the Outcome-Based Offline RL Upper Bound

In this section we prove Theorem 1, using the population losses defined in Eq. (13).

C.1 Regret Decomposition

Our objective is to find an upper bound for

$$J(\pi) - J(\bar{\pi}) = \frac{1}{K}\sum_{k=1}^{K}\big(J(\pi) - J(\pi_k)\big).$$

By Lemma 18, for any $\pi_k \in \Pi$, $f_k \in \mathcal{F}$, and $r_k \in \mathcal{R}$ learned in the algorithm, we have

$$\begin{aligned} J(\pi) - J(\pi_k) = &\underbrace{\begin{aligned} &\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[\mathcal{T}^{\pi_k}_{r_k} f_{k,h}(s_h,a_h) - f_{k,h}(s_h,a_h)\right] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\mu}}\left[f_{k,h}(s_h,a_h) - \mathcal{T}^{\pi_k}_{r_k} f_{k,h}(s_h,a_h)\right] \end{aligned}}_{(\mathrm{I})_k} \\ &+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^{\pi}}\left[f_{k,h}(s_h,a_h) - f_{k,h}(s_h,\pi_k(s_h))\right]}_{(\mathrm{II})_k} \end{aligned}$$

$$+\underbrace{\mathcal{L}_\mu(\pi_k, f_k) - \mathcal{L}_\mu(\pi_k, Q^{\pi_k})}_{(\mathrm{III})_k}$$
$$+\underbrace{\begin{aligned}&\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\pi}\left[r^\star(s_h,a_h) - r_k(s_h,a_h)\right]\\ &+\sum_{h=1}^{H} \mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[r_k(s_h,a_h) - r^\star(s_h,a_h)\right]\end{aligned}}_{(\mathrm{IV})_k}. \tag{48}$$

C.2 Bounding the Policy-Optimization Term

Term $(\mathrm{II})_k = \sum_{h=1}^H \mathbb{E}_{s_h,a_h\sim d_h^\pi}\left[f_{k,h}(s_h,a_h) - f_{k,h}(s_h,\pi_k(s_h))\right]$ is controlled by the no-regret property of the layer-wise softmax update Algorithm 1, line 4. Writing $d_{h,S}^\pi(s) := \sum_a d_h^\pi(s,a)$ for the state marginal, we have

$$(\mathrm{II})_k = \sum_{h=1}^{H} \mathbb{E}_{s\sim d_{h,S}^\pi}\Big[\mathbb{E}_{a\sim\pi_h(\cdot|s)}[f_{k,h}(s,a)] - \mathbb{E}_{a\sim\pi_{k,h}(\cdot|s)}[f_{k,h}(s,a)]\Big] = \mathrm{Reg}_k(\pi),$$

because $d_h^\pi(s,a) = d_{h,S}^\pi(s)\pi_h(a \mid s)$. Since every $f_k \in \mathcal{F}$ satisfies $f_{k,h}(s,a) \in [0, V_{\max}]$, Lemma 11 applies with the comparator $\pi \in \Pi$. Choosing $\eta = \eta^\star = V_{\max}\sqrt{K/(8\log|\mathcal{A}|)}$, Eq. (15) yields

$$\frac{1}{K}\sum_{k=1}^{K}(\mathrm{II})_k \le HV_{\max}\sqrt{\frac{\log|\mathcal{A}|}{2K}} = \widetilde{O}\left(\frac{HV_{\max}}{\sqrt{K}}\right). \tag{49}$$

C.3 Bounding the Bellman and Reward-Mismatch Terms

Recall the state–action concentrability $C_{sa}(\pi)$ from Eq. (2). For any $g : \mathcal{S}_h \times \mathcal{A} \to \mathbb{R}$, the pointwise bound $d_h^\pi(s,a) \le C_{sa}(\pi) d_h^\mu(s,a)$ gives

$$\mathbb{E}_{d_h^\pi}[g(s_h,a_h)^2] \le C_{sa}(\pi)\mathbb{E}_{d_h^\mu}[g(s_h,a_h)^2].$$

Applying this inequality, Cauchy–Schwarz, and AM–GM with free parameter $\beta/2 > 0$ (so that, after the factor-2 Bellman-error concentration Eq. (53a) below, the pessimism coefficient matches the algorithmic one, β, in Algorithm 1, line 3), we have

$$\mathbb{E}_{s_h,a_h\sim d_h^\pi}[(\mathcal{T}_{r_k}^{\pi_k} f_k - f_k)(s_h,a_h)] \le \frac{C_{sa}(\pi)}{\beta} + \frac{\beta}{4}\mathbb{E}_{s_h,a_h\sim d_h^\mu}[((\mathcal{T}_{r_k}^{\pi_k} f_k - f_k)(s_h,a_h))^2],$$
$$\mathbb{E}_{s_h,a_h\sim d_h^\mu}[(f_k - \mathcal{T}_{r_k}^{\pi_k} f_k)(s_h,a_h)] \le \frac{1}{\beta} + \frac{\beta}{4}\mathbb{E}_{s_h,a_h\sim d_h^\mu}[((\mathcal{T}_{r_k}^{\pi_k} f_k - f_k)(s_h,a_h))^2].$$

Summing over $h \in [H]$,

$$(\mathrm{I})_k \le \frac{H(C_{sa}(\pi)+1)}{\beta} + \frac{\beta}{2}\sum_{h=1}^{H}\mathbb{E}_{s_h,a_h\sim d_h^\mu}\left[((\mathcal{T}_{r_k}^{\pi_k} f_k - f_k)(s_h,a_h))^2\right]. \tag{50}$$

Similarly, by the change of trajectory measure (Lemma 10)

$$\sup_f \frac{\left(\mathbb{E}_{\tau\sim\pi}\left[\sum_h f(s_h,a_h)\right]\right)^2}{\mathbb{E}_{\tau\sim\mu}\left[\left(\sum_h f(s_h,a_h)\right)^2\right]} \le HC_{sa}(\pi),$$

selecting $g_k = r^\star - r_k$, we have

$$\sum_{h=1}^{H} g_k(s_h,a_h) = \sum_{h=1}^{H}\left(r^\star(s_h,a_h) - r_k(s_h,a_h)\right) = R^\star(\tau) - R(\tau; r_k).$$

Therefore, applying Young's inequality,

$$\mathbb{E}_{\tau\sim\pi}\Big[\sum_{h=1}^{H}\big(r^\star(s_h,a_h)-r_k(s_h,a_h)\big)\Big] \le \frac{HC_{sa}(\pi)}{\beta}+\frac{\beta}{4}\mathbb{E}_{\tau\sim\mu}[(R^\star(\tau)-R(\tau;r_k))^2],$$

$$\mathbb{E}_{\tau\sim\mu}\Big[\sum_{h=1}^{H}\big(r_k(s_h,a_h)-r^\star(s_h,a_h)\big)\Big] \le \frac{1}{\beta}+\frac{\beta}{4}\mathbb{E}_{\tau\sim\mu}[(R^\star(\tau)-R(\tau;r_k))^2].$$

$$(\mathrm{IV})_k \le \frac{HC_{sa}(\pi)+1}{\beta}+\frac{\beta}{2}\mathbb{E}_{\tau\sim\mu}[(R^\star(\tau)-R(\tau;r_k))^2]. \tag{51}$$

Then by combining Eq. (50) and Eq. (51) we get

$$\begin{aligned}(\mathrm{I})_k+(\mathrm{III})_k+(\mathrm{IV})_k &\le \frac{H(C_{sa}(\pi)+1)}{\beta}+\frac{\beta}{2}\sum_{h=1}^{H}\mathbb{E}_{s_h,a_h\sim d_h^\mu}\big[((\mathcal{T}_{r_k}^{\pi_k}f_k-f_k)(s_h,a_h))^2\big]\\ &\quad+\frac{HC_{sa}(\pi)+1}{\beta}+\frac{\beta}{2}\mathbb{E}_{\tau\sim\mu}[(R^\star(\tau)-R(\tau;r_k))^2]\\ &\quad+\mathcal{L}_\mu(\pi_k,f_k)-\mathcal{L}_\mu(\pi_k,Q^{\pi_k})\\ &\le \frac{2H(C_{sa}(\pi)+1)}{\beta}+\frac{\beta}{2}\mathcal{L}_\mu^{\mathrm{BE}}(\pi_k,r_k,f_k)+\frac{\beta}{2}\mathcal{L}_\mu^{\mathrm{RM}}(r_k)\\ &\quad+\mathcal{L}_\mu(\pi_k,f_k)-\mathcal{L}_\mu(\pi_k,Q^{\pi_k}),\end{aligned} \tag{52}$$

where in the last inequality we used $HC_{sa}(\pi)+1\le H(C_{sa}(\pi)+1)$ and the definitions Eq. (13).

C.4 Pessimism and Empirical Transfer

Now fix the high-probability event on which the concentration and approximation lemmas from Appendices B.2 and B.3 hold simultaneously for every $\pi\in\Pi$, $r\in\mathcal{R}$, and $f\in\mathcal{F}$. Because the change-of-measure step Eq. (52) produces the coefficient $\beta/2$ on the population BE/RM losses, and the Bellman-error (Lemma 14) and reward-model (Lemma 13) concentration inequalities introduce a factor 2 in front of the empirical losses, the two factors cancel so that the pessimistic evaluation step Algorithm 1, line 3 is invoked with the algorithmic coefficient β. On this event we have, for every $k\in[K]$,

$$\mathcal{L}_\mu^{\mathrm{BE}}(\pi_k,r_k,f_k)\le 2\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi_k,r_k,f_k)+\varepsilon_{\mathrm{BE}}+4\varepsilon_{\mathcal{F},\mathcal{F}}, \tag{53a}$$

$$\mathcal{L}_\mu^{\mathrm{RM}}(r_k)\le 2\big[\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}(r_k)-\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}(r^\star)\big]+\varepsilon_{\mathrm{RM}}, \tag{53b}$$

$$\sup_{\pi\in\Pi,f\in\mathcal{F}}\big|\mathcal{L}_\mu(\pi,f)-\mathcal{L}_{\mathcal{D}}(\pi,f)\big|\le\sqrt{\varepsilon_{\mathrm{Perf}}}, \tag{53c}$$

$$|\mathcal{L}_\mu(\pi_k,Q^{\pi_k})-\mathcal{L}_\mu(\pi_k,f_{\pi_k})|\le O(H\sqrt{\varepsilon_{\mathcal{F}}})\quad\text{(Lemma 17)}. \tag{53d}$$

We now bound the quantity

$$\Delta_k:=\frac{\beta}{2}\mathcal{L}_\mu^{\mathrm{BE}}(\pi_k,r_k,f_k)+\frac{\beta}{2}\mathcal{L}_\mu^{\mathrm{RM}}(r_k)+\mathcal{L}_\mu(\pi_k,f_k)-\mathcal{L}_\mu(\pi_k,Q^{\pi_k})$$

that appears on the right-hand side of Eq. (52), in four explicit steps.

Step (A): transfer BE/RM losses to the empirical data. Using Eq. (53a)–Eq. (53b),

$$\begin{aligned}\frac{\beta}{2}\mathcal{L}_\mu^{\mathrm{BE}}(\pi_k,r_k,f_k)+\frac{\beta}{2}\mathcal{L}_\mu^{\mathrm{RM}}(r_k)&\le\beta\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi_k,r_k,f_k)+\beta\big[\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}(r_k)-\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}(r^\star)\big]\\&\quad+\frac{\beta}{2}\big(\varepsilon_{\mathrm{BE}}+\varepsilon_{\mathrm{RM}}+4\varepsilon_{\mathcal{F},\mathcal{F}}\big).\end{aligned} \tag{54}$$

The $r^\star$-centered form is essential under noisy observations, but it cancels exactly when the same quantity is subtracted from the pessimism RHS in Step (D) below.

Step (B): transfer the performance-loss gap $\mathcal{L}_\mu(\pi_k,f_k)-\mathcal{L}_\mu(\pi_k,Q^{\pi_k})$ to the empirical data. Insert f_{π_k} as a pivot and use Eq. (53c) together with Eq. (53d):

$$\mathcal{L}_\mu(\pi_k,f_k)-\mathcal{L}_\mu(\pi_k,Q^{\pi_k})=\underbrace{\big[\mathcal{L}_\mu(\pi_k,f_k)-\mathcal{L}_{\mathcal{D}}(\pi_k,f_k)\big]}_{\le\sqrt{\varepsilon_{\mathrm{Perf}}}\text{ by Eq. (53c)}}$$

$$
\begin{aligned}
&+ \underbrace{\big[\mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) - \mathcal{L}_{\mu}(\pi_k, f_{\pi_k})\big]}_{\leq \sqrt{\varepsilon_{\mathrm{Perf}}} \text{ by } Eq.\ (53c)} \\
&+ \underbrace{\big[\mathcal{L}_{\mu}(\pi_k, f_{\pi_k}) - \mathcal{L}_{\mu}(\pi_k, Q^{\pi_k})\big]}_{\leq O(H\sqrt{\varepsilon_{\mathcal{F}}}) \text{ by } Eq.\ (53d)} \\
&+ \mathcal{L}_{\mathcal{D}}(\pi_k, f_k) - \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) \\
&\leq \mathcal{L}_{\mathcal{D}}(\pi_k, f_k) - \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) + 2\sqrt{\varepsilon_{\mathrm{Perf}}} + O(H\sqrt{\varepsilon_{\mathcal{F}}}), \qquad (55)
\end{aligned}
$$

where we applied the uniform concentration Eq. (53c) twice (once to $(\pi_k, f_k) \in \Pi \times \mathcal{F}$ and once to $(\pi_k, f_{\pi_k}) \in \Pi \times \mathcal{F}$; the latter uses $f_{\pi_k} \in \mathcal{F}$ from Lemma 15).

Step (C): invoke pessimistic optimality at coefficient β. By the definition of the pessimistic estimator Algorithm 1, line 3, $(f_k, r_k) \in \mathcal{F} \times \mathcal{R}$ minimizes

$$
(f, r) \mapsto \mathcal{L}_{\mathcal{D}}(\pi_k, f) + \beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r, f) + \beta \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r).
$$

Plugging in the comparator $(f_{\pi_k}, r^\star) \in \mathcal{F} \times \mathcal{R}$ (using $f_{\pi_k} \in \mathcal{F}$ from Assumption 2 and $r^\star \in \mathcal{R}$ from Assumption 1) yields

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\pi_k, f_k) + \beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r_k, f_k) + \beta \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r_k) \\
&\quad \leq \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) + \beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r^\star, f_{\pi_k}) + \beta \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r^\star). \qquad (56)
\end{aligned}
$$

Step (D): bound the comparator Bellman-error loss, and cancel $\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r^\star)$. Subtracting $\beta \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r^\star)$ from both sides of Eq. (56) gives the $r^\star$-centered pessimism inequality

$$
\begin{aligned}
&\mathcal{L}_{\mathcal{D}}(\pi_k, f_k) + \beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r_k, f_k) + \beta \big[\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r_k) - \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r^\star)\big] \\
&\quad \leq \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) + \beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r^\star, f_{\pi_k}). \qquad (57)
\end{aligned}
$$

By Lemma 15, the remaining comparator Bellman-error term on the RHS is controlled by

$$
\beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r^\star, f_{\pi_k}) \leq \beta \varepsilon_{\mathrm{apx}}. \qquad (58)
$$

Combining (A)–(D). Adding Eq. (54) and Eq. (55) gives

$$
\begin{aligned}
\Delta_k \leq\ & \big[\beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r_k, f_k) + \beta\big(\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r_k) - \mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r^\star)\big) + \mathcal{L}_{\mathcal{D}}(\pi_k, f_k) - \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k})\big] \\
&+ \frac{\beta}{2}\big(\varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{RM}} + 4\varepsilon_{\mathcal{F},\mathcal{F}}\big) + 2\sqrt{\varepsilon_{\mathrm{Perf}}} + O(H\sqrt{\varepsilon_{\mathcal{F}}}).
\end{aligned}
$$

By the $r^\star$-centered pessimism inequality Eq. (57), the bracketed empirical quantity is bounded by $\beta \mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r^\star, f_{\pi_k})$, which is in turn controlled by Eq. (58). Note that the potentially constant-order term $\mathcal{L}^{\mathrm{RM}}_{\mathcal{D}}(r^\star)$ has cancelled on both sides and does not appear in the final bound. Therefore

$$
\Delta_k \leq \frac{\beta}{2}\big(\varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{RM}} + 4\varepsilon_{\mathcal{F},\mathcal{F}} + 2\varepsilon_{\mathrm{apx}}\big) + 2\sqrt{\varepsilon_{\mathrm{Perf}}} + O(H\sqrt{\varepsilon_{\mathcal{F}}}). \qquad (59)
$$

C.5 Final Rate and Parameter Choice

Define

$$
\varepsilon := \varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{RM}} + 4\varepsilon_{\mathcal{F},\mathcal{F}} + 2\varepsilon_{\mathrm{apx}}, \qquad \xi := 2\sqrt{\varepsilon_{\mathrm{Perf}}} + O(H\sqrt{\varepsilon_{\mathcal{F}}}).
$$

Substituting the bound on Δ_k from Eq. (59) into Eq. (52) gives

$$
(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \leq \frac{2H(C_{sa}(\pi) + 1)}{\beta} + \Delta_k \leq \frac{2H(C_{sa}(\pi) + 1)}{\beta} + \frac{\beta}{2}\varepsilon + \xi. \qquad (60)
$$

Since n denotes the number of trajectories, the concentration terms above all scale with n^{-1}. In this cumulative-return setting, the boundedness convention before Eq. (3a) sets $V_{\max} = H$, so under the sum-over-h convention of Eq. (13), we may upper bound every concentration term at the trajectory scale H:

$$
\varepsilon_{\mathrm{BE}} = \widetilde{O}\left(\frac{H^3}{n}\right) \quad (\text{Lemma 14}, V_{\max} = H), \qquad \varepsilon_{\mathrm{RM}} = \widetilde{O}\left(\frac{H^2}{n}\right) \quad (\text{Lemma 13}),
$$

$$\varepsilon_{\mathrm{apx}} = \widetilde{O}\left(\frac{H^3}{n} + H\varepsilon_{\mathcal{F}}\right) \quad \text{(Lemma 15)},$$
$$\varepsilon_{\mathrm{Perf}} = \widetilde{O}\left(\frac{H^4}{n}\right) \quad \text{(Lemma 12)}.$$

The factor H in $\varepsilon_{\mathrm{BE}}$ and $\varepsilon_{\mathrm{apx}}$ relative to $\varepsilon_{\mathrm{RM}}$ comes from the sum over $h \in [H]$ in $\mathcal{L}^{\mathrm{BE}}$. Hence

$$\varepsilon = \widetilde{O}\left(\frac{H^3}{n} + H\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}\right), \qquad \xi = \widetilde{O}\left(\frac{H^2}{\sqrt{n}} + H\sqrt{\varepsilon_{\mathcal{F}}}\right).$$

Optimizing the bound $\frac{2H(C_{sa}(\pi)+1)}{\beta} + \frac{\beta}{2}\varepsilon$ over $\beta > 0$ yields the minimizer

$$\beta^\star = 2\sqrt{\frac{H(C_{sa}(\pi)+1)}{\varepsilon}},$$

at which

$$(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \le 2\sqrt{H(C_{sa}(\pi)+1)\varepsilon} + \xi.$$

Substituting the bound on ε and using $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$,

$$(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \le \widetilde{O}\Bigg(H\sqrt{\frac{H^2(C_{sa}(\pi)+1)}{n}} + H\sqrt{(C_{sa}(\pi)+1)\varepsilon_{\mathcal{F}}} + \sqrt{H(C_{sa}(\pi)+1)\varepsilon_{\mathcal{F},\mathcal{F}}}\Bigg) + \xi.$$

Finally, applying the no-regret bound Eq. (49) for the $(\mathrm{II})_k$ terms and substituting $\xi = \widetilde{O}(H^2/\sqrt{n} + H\sqrt{\varepsilon_{\mathcal{F}}})$,

$$\begin{aligned}
J(\pi) - J(\bar{\pi}) &= \frac{1}{K}\sum_{k=1}^{K}\big(J(\pi) - J(\pi_k)\big) \\
&\le \frac{1}{K}\sum_{k=1}^{K}(\mathrm{II})_k + \frac{1}{K}\sum_{k=1}^{K}\big((\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k\big) \\
&\le \frac{1}{K}\sum_{k=1}^{K}(\mathrm{II})_k + 2\sqrt{H(C_{sa}(\pi)+1)\varepsilon} + \xi \\
&\le H^2\sqrt{\frac{\log|\mathcal{A}|}{2K}} + 2\sqrt{H(C_{sa}(\pi)+1)\varepsilon} + \xi \\
&\le \widetilde{O}\left(H^2\sqrt{\frac{C_{sa}(\pi)}{n}} + \frac{H^2}{\sqrt{K}} + H\sqrt{C_{sa}(\pi)\varepsilon_{\mathcal{F}}} + \sqrt{HC_{sa}(\pi)\varepsilon_{\mathcal{F},\mathcal{F}}} + H\sqrt{\varepsilon_{\mathcal{F}}}\right),
\end{aligned}$$

where the performance-loss contribution $H^2/\sqrt{n}$ has been absorbed into the leading statistical rate $H^2\sqrt{C_{sa}(\pi)/n}$.

In particular, taking $K \ge n$ and assuming exact realizability ($\varepsilon_{\mathcal{F}} = 0$) and exact Bellman completeness ($\varepsilon_{\mathcal{F},\mathcal{F}} = 0$),

$$J(\pi) - J(\bar{\pi}) \le \widetilde{O}\left(H^2\sqrt{\frac{C_{sa}(\pi)}{n}}\right).$$

D Proof of the Preference-Based Upper Bound

We prove Theorem 3. In this cumulative-return setting, $V_{\max} = 1$ and every candidate trajectory return lies in $[0, 1]$. The argument follows the proof of Theorem 1. The policy and Bellman empirical losses use the $2n$ unlabeled trajectories contained in the preference pairs, with $\mathcal{D}$ abbreviating $\mathcal{D}_{\mathrm{traj}}$;

since these trajectories are i.i.d. from μ, the corresponding concentration bounds change only by universal constants. Relative to the scalar-outcome proof, the main changes are that the squared reward-model loss is replaced by the logistic preference loss, and the reward-mismatch term is controlled through a centered change-of-measure step, since preferences identify rewards only up to uniform shifts of cumulative returns.

We use the following constants for the fixed BTL model on the return range $[0, 1]$:

$$\begin{aligned}
\alpha_{\mathrm{C}} &:= \frac{\gamma^2}{2} \inf_{|w|\le 1} \frac{\exp(\gamma w)}{(1+\exp(\gamma w))^2}, \\
a_{\mathrm{C}} &:= \frac{1}{1+e^{\gamma}}, \qquad B_{\mathrm{C}} := \log \frac{1-a_{\mathrm{C}}}{a_{\mathrm{C}}}, \\
b_{\mathrm{C}} &:= \sup_{p,q\in[a_{\mathrm{C}},1-a_{\mathrm{C}}],p\neq q} \frac{p\log^2(p/q) + (1-p)\log^2((1-p)/(1-q))}{\mathrm{KL}(\mathrm{Bern}(p)\|\mathrm{Bern}(q))}, \qquad c_{\mathrm{C}} := 2b_{\mathrm{C}} + \frac{8}{3}B_{\mathrm{C}}.
\end{aligned} \tag{61}$$

These constants depend only on γ.

D.1 Preference Identifiability and Concentration

For a pair (τ^+, τ^-), define the return-difference notation

$$\Delta_r(\tau^+, \tau^-) := R(\tau^+; r) - R(\tau^-; r), \qquad \Delta^\star(\tau^+, \tau^-) := \Delta_{r^\star}(\tau^+, \tau^-).$$

Let

$$\ell_r(\tau^+, \tau^-, y) := -y \log \mathrm{C}_r(\tau^+, \tau^-) - (1-y)\log(1 - \mathrm{C}_r(\tau^+, \tau^-))$$

be the BTL log loss induced by r. The population preference loss is

$$\mathcal{L}_\mu^{\mathrm{Pref}}(r) := \mathbb{E}_{\tau^+,\tau^-\sim\mu, y|\tau^\pm}\left[\ell_r(\tau^+, \tau^-, y)\right].$$

Lemma 19 (BTL preference identifiability)**.** *Under the BTL model Eq. (5), with α_{C} defined in Eq. (61), for every $r \in \mathcal{R}$,*

$$\mathcal{L}_\mu^{\mathrm{Pref}}(r) - \mathcal{L}_\mu^{\mathrm{Pref}}(r^\star) \ge \alpha_{\mathrm{C}} \mathbb{E}_{\mu\otimes\mu}\left[(\Delta_r - \Delta^\star)^2\right] = 2\alpha_{\mathrm{C}} \mathrm{Var}_\mu\left(R(\cdot; r) - R^\star(\cdot)\right).$$

Equivalently,

$$\mathrm{Var}_\mu\left(R(\cdot; r) - R^\star(\cdot)\right) \le \frac{1}{2\alpha_{\mathrm{C}}}\big(\mathcal{L}_\mu^{\mathrm{Pref}}(r) - \mathcal{L}_\mu^{\mathrm{Pref}}(r^\star)\big).$$

Proof. Fix a trajectory pair (τ^+, τ^-). Write

$$d_r = \Delta_r(\tau^+, \tau^-), \qquad d_\star = \Delta^\star(\tau^+, \tau^-).$$

Since $R(\tau; r) \in [0, 1]$, both d_r and $d_\star$ lie in $[-1, 1]$. Define the one-dimensional logistic link

$$s_\gamma(w) = \frac{\exp(\gamma w)}{1+\exp(\gamma w)}.$$

For this fixed pair, the true label distribution is

$$y \sim \mathrm{Bern}(s_\gamma(d_\star)).$$

Let $p_\star = s_\gamma(d_\star)$ and define the conditional log-risk

$$\phi(w) = -p_\star \log s_\gamma(w) - (1-p_\star)\log(1 - s_\gamma(w)).$$

Then

$$\mathbb{E}_y[\ell_r(\tau^+, \tau^-, y)] = \phi(d_r), \qquad \mathbb{E}_y[\ell_{r^\star}(\tau^+, \tau^-, y)] = \phi(d_\star).$$

A direct differentiation gives

$$\phi'(w) = \gamma\big(s_\gamma(w) - p_\star\big), \qquad \phi''(w) = \gamma^2 s_\gamma(w)(1 - s_\gamma(w)).$$

Since $p_\star = s_\gamma(d_\star)$, we have $\phi'(d_\star) = 0$. Also,

$$m_{\mathrm{C}} := \inf_{|w|\le 1} \gamma^2 s_\gamma(w)(1 - s_\gamma(w)) > 0,$$

because $s_\gamma(w)(1-s_\gamma(w))$ is continuous and strictly positive on the compact interval $[-1,1]$. Taylor's theorem with remainder gives, for some point $\bar{w}$ between d_r and $d_\star$,

$$\phi(d_r)-\phi(d_\star)=\phi'(d_\star)(d_r-d_\star)+\frac{1}{2}\phi''(\bar{w})(d_r-d_\star)^2\geq\frac{m_{\mathrm{C}}}{2}(d_r-d_\star)^2.$$

By the definition of α_{C} in Eq. (61), $m_{\mathrm{C}}/2=\alpha_{\mathrm{C}}$. Since the trajectory pair was arbitrary, taking expectation over $\tau^+,\tau^-\sim\mu$ gives

$$\mathcal{L}^{\mathrm{Pref}}_\mu(r)-\mathcal{L}^{\mathrm{Pref}}_\mu(r^\star)\geq\alpha_{\mathrm{C}}\mathbb{E}_{\mu\otimes\mu}\left[(\Delta_r-\Delta^\star)^2\right].$$

It remains to rewrite the pairwise error as a variance. Let

$$g(\tau)=R(\tau;r)-R^\star(\tau).$$

Then

$$\Delta_r(\tau^+,\tau^-)-\Delta^\star(\tau^+,\tau^-)=g(\tau^+)-g(\tau^-).$$

Since τ^+ and τ^- are independent draws from μ,

$$\begin{aligned}\mathbb{E}_{\mu\otimes\mu}[(g(\tau^+)-g(\tau^-))^2]&=\mathbb{E}_\mu[g(\tau)^2]+\mathbb{E}_\mu[g(\tau)^2]-2\mathbb{E}_\mu[g(\tau)]^2\\&=2\mathrm{Var}_\mu(g),\end{aligned}$$

as desired. □

Lemma 20 (Preference log loss concentration)**.** *Let a_{C}, B_{C}, b_{C}, and c_{C} be defined in Eq. (61). With probability at least $1-\delta$, simultaneously for all $r\in\mathcal{R}$,*

$$\mathcal{L}^{\mathrm{Pref}}_\mu(r)-\mathcal{L}^{\mathrm{Pref}}_\mu(r^\star)\leq 2\Big[\mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r)-\mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r^\star)\Big]+c_{\mathrm{C}}\frac{\log(|\mathcal{R}|/\delta)}{n}.$$

Proof. The structure is the same as the $r^\star$-centered reward-model concentration in Lemma 13: we control the population excess loss by the empirical loss difference relative to $r^\star$. The difference is that the random variable is now a log-likelihood ratio, whose mean is a KL divergence and whose variance is self-bounded by that same KL divergence.

Fix $r\in\mathcal{R}$. Let $Z=(\tau^+,\tau^-,y)$ and let $P_\star$ denote the true law of Z: the pair (τ^+,τ^-) is drawn from $\mu\otimes\mu$, and then

$$y\mid\tau^+,\tau^-\sim\mathrm{Bern}\big(\mathrm{C}_{r^\star}(\tau^+,\tau^-)\big).$$

Let P_r be the same law except that the conditional Bernoulli success probability is $\mathrm{C}_r(\tau^+,\tau^-)$. Since the pair marginal is the same under $P_\star$ and P_r, only the conditional Bernoulli law changes.

Write $p_\star=\mathrm{C}_{r^\star}(\tau^+,\tau^-)$ and $p_r=\mathrm{C}_r(\tau^+,\tau^-)$. Since $P_\star$ and P_r have the same pair distribution, their likelihood ratio is just the Bernoulli label likelihood ratio. Thus

$$\begin{aligned}\ell_r(\tau^+,\tau^-,y)-\ell_{r^\star}(\tau^+,\tau^-,y)&=[-y\log p_r-(1-y)\log(1-p_r)]\\&\quad-[-y\log p_\star-(1-y)\log(1-p_\star)]\\&=y\log\frac{p_\star}{p_r}+(1-y)\log\frac{1-p_\star}{1-p_r}\\&=\log\frac{p_\star^y(1-p_\star)^{1-y}}{p_r^y(1-p_r)^{1-y}}.\end{aligned}$$

Thus the single-sample loss difference is exactly the log-likelihood ratio:

$$W_r(Z)\coloneqq\ell_r(\tau^+,\tau^-,y)-\ell_{r^\star}(\tau^+,\tau^-,y)=\log\frac{P_\star(y\mid\tau^+,\tau^-)}{P_r(y\mid\tau^+,\tau^-)}.$$

For this fixed trajectory pair, taking expectation only over the label $y\mid\tau^+,\tau^-$ gives

$$\begin{aligned}\mathbb{E}_{y|\tau^+,\tau^-}[W_r(Z)]&=p_\star\log\frac{p_\star}{p_r}+(1-p_\star)\log\frac{1-p_\star}{1-p_r}\\&=\mathrm{KL}\left(\mathrm{Bern}(p_\star)\|\mathrm{Bern}(p_r)\right).\end{aligned}$$

Now taking the outer expectation over the trajectory pair $(\tau^+, \tau^-) \sim \mu \otimes \mu$,

$$\begin{aligned}\mathcal{L}_\mu^{\mathrm{Pref}}(r) - \mathcal{L}_\mu^{\mathrm{Pref}}(r^\star) &= \mathbb{E}_{(\tau^+,\tau^-)\sim\mu\otimes\mu}\left[\mathbb{E}_{y|\tau^+,\tau^-}[W_r(Z)]\right] \\ &= \mathbb{E}_{Z\sim P_\star}[W_r(Z)] = \mathrm{KL}(P_\star \| P_r).\end{aligned}$$

Also, for the observed dataset $Z_1, \ldots, Z_n$,

$$\frac{1}{n}\sum_{i=1}^n W_r(Z_i) = \mathcal{L}_{\mathcal{D}_{\mathrm{Pref}}}^{\mathrm{Pref}}(r) - \mathcal{L}_{\mathcal{D}_{\mathrm{Pref}}}^{\mathrm{Pref}}(r^\star).$$

Write

$$L_r \coloneqq \mathrm{KL}(P_\star \| P_r), \qquad \widehat{L}_r \coloneqq \frac{1}{n}\sum_{i=1}^n W_r(Z_i).$$

We now prove concentration of $\widehat{L}_r$ around L_r. The BTL probabilities are uniformly bounded away from 0 and 1 on the return range $[0, 1]$. Indeed, for all $r \in \mathcal{R}$ and all trajectory pairs,

$$a_{\mathrm{C}} \le \mathrm{C}_r(\tau^+, \tau^-) \le 1 - a_{\mathrm{C}}.$$

Hence the log-likelihood ratio is bounded:

$$|W_r(Z)| \le B_{\mathrm{C}}.$$

We also need the standard self-bounding property of bounded Bernoulli log likelihoods. For $p, q \in [a_{\mathrm{C}}, 1 - a_{\mathrm{C}}]$, define

$$k(p, q) = \mathrm{KL}(\mathrm{Bern}(p) \| \mathrm{Bern}(q))$$

and

$$m(p, q) = p \log^2 \frac{p}{q} + (1 - p) \log^2 \frac{1 - p}{1 - q}.$$

Since the square $[a_{\mathrm{C}}, 1 - a_{\mathrm{C}}]^2$ is compact and $m(p, q)/k(p, q)$ has a finite continuous extension at $p = q$, the constant b_{C} in Eq. (61) is finite and satisfies $m(p, q) \le b_{\mathrm{C}} k(p, q)$ for all p, q in this interval.

Applying this pointwise with $p = \mathrm{C}_{r^\star}(\tau^+, \tau^-)$ and $q = \mathrm{C}_r(\tau^+, \tau^-)$, the conditional second moment over the label satisfies

$$\mathbb{E}_{y|\tau^+,\tau^-}[W_r(Z)^2] \le b_{\mathrm{C}} \mathbb{E}_{y|\tau^+,\tau^-}[W_r(Z)].$$

Taking the outer expectation over $(\tau^+, \tau^-) \sim \mu \otimes \mu$ gives

$$\mathbb{E}_{Z\sim P_\star}[W_r(Z)^2] \le b_{\mathrm{C}} \mathbb{E}_{Z\sim P_\star}[W_r(Z)] = b_{\mathrm{C}} L_r.$$

Therefore $\mathrm{Var}_{Z\sim P_\star}(W_r(Z)) \le b_{\mathrm{C}} L_r$.

Now apply one-sided Bernstein to the i.i.d. variables $W_r(Z_1), \ldots, W_r(Z_n)$. Since $|W_r(Z)| \le B_{\mathrm{C}}$ and $L_r = \mathbb{E}_{Z\sim P_\star}[W_r(Z)] \le \mathbb{E}_{Z\sim P_\star}[|W_r(Z)|] \le B_{\mathrm{C}}$, the lower-tail centered variables satisfy $L_r - W_r(Z) \le 2B_{\mathrm{C}}$. Thus, for any fixed r and any $\eta \in (0, 1)$, with probability at least $1 - \eta$,

$$L_r - \widehat{L}_r \le \sqrt{\frac{2 b_{\mathrm{C}} L_r \log(1/\eta)}{n}} + \frac{4 B_{\mathrm{C}} \log(1/\eta)}{3n}.$$

Take $\eta = \delta/|\mathcal{R}|$ and take a union bound over $r \in \mathcal{R}$. On the resulting event, set

$$a_n = \frac{\log(|\mathcal{R}|/\delta)}{n}.$$

Then every $r \in \mathcal{R}$ satisfies

$$L_r \le \widehat{L}_r + \sqrt{2 b_{\mathrm{C}} L_r a_n} + \frac{4 B_{\mathrm{C}}}{3} a_n.$$

Using $\sqrt{2 b_{\mathrm{C}} L_r a_n} \le L_r/2 + b_{\mathrm{C}} a_n$ and rearranging,

$$L_r \le 2\widehat{L}_r + c_{\mathrm{C}} a_n.$$

Substituting back the definitions of L_r and $\widehat{L}_r$ proves the lemma. Here B_{C}, b_{C}, and c_{C} are constants determined by the BTL comparison model on the bounded return range. More explicitly, they depend on $a_{\mathrm{C}} = 1/(1 + e^\gamma)$, and hence on γ, but not on n, δ, or the size of the reward class except through the logarithmic factor above. □

Lemma 21 (Centered change of trajectory measure)**.** *For any Markov policy π with $C_{sa}(\pi) < \infty$ and any bounded measurable per-step function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, write $\bar{g}(\tau) = \sum_{h=1}^{H} g(s_h, a_h)$. Then*

$$\left(\mathbb{E}_\pi[\bar{g}] - \mathbb{E}_\mu[\bar{g}]\right)^2 \le HC_{sa}(\pi)\mathrm{Var}_\mu(\bar{g}).$$

Proof. Let $\bar{g}_\mu = \mathbb{E}_\mu[\bar{g}]$ and define $\tilde{g}(s_h, a_h) = g(s_h, a_h) - \bar{g}_\mu/H$. Then $\sum_h \tilde{g}(s_h, a_h) = \bar{g}(\tau) - \bar{g}_\mu$. Applying Lemma 10 to $\tilde{g}$ gives

$$\left(\mathbb{E}_\pi[\bar{g} - \bar{g}_\mu]\right)^2 \le HC_{sa}(\pi)\mathbb{E}_\mu[(\bar{g} - \bar{g}_\mu)^2],$$

which is the desired inequality. □

D.2 Proof of the Preference-Based Upper Bound

Fix a competitor $\pi \in \Pi$. The regret decomposition Eq. (48) is unchanged because the objective is still the cumulative return $J(\pi) = \mathbb{E}_\pi[R^\star(\tau)]$. Terms $(\mathrm{I})_k$, $(\mathrm{II})_k$, and $(\mathrm{III})_k$ are controlled exactly as in the scalar-outcome proof. The only population-level change is the reward-mismatch term

$$(\mathrm{IV})_k = \mathbb{E}_\pi[R^\star(\tau) - R(\tau; r_k)] + \mathbb{E}_\mu[R(\tau; r_k) - R^\star(\tau)].$$

Let $g_{k,h}(s, a) \coloneqq r_{k,h}(s, a) - r^\star_h(s, a)$ and $\bar{g}_k(\tau) \coloneqq \sum_{h=1}^{H} g_{k,h}(s_h, a_h) = R(\tau; r_k) - R^\star(\tau)$. Then

$$(\mathrm{IV})_k = -\big(\mathbb{E}_\pi[\bar{g}_k] - \mathbb{E}_\mu[\bar{g}_k]\big).$$

By Lemma 21, Lemma 19, and AM–GM with parameter $\beta > 0$,

$$\begin{aligned}
|(\mathrm{IV})_k| &\le \sqrt{HC_{sa}(\pi)\mathrm{Var}_\mu(\bar{g}_k)} \\
&\le \sqrt{\frac{HC_{sa}(\pi)}{2\alpha_{\mathrm{C}}}\big(\mathcal{L}^{\mathrm{Pref}}_\mu(r_k) - \mathcal{L}^{\mathrm{Pref}}_\mu(r^\star)\big)} \\
&\le \frac{HC_{sa}(\pi)}{4\alpha_{\mathrm{C}}\beta} + \frac{\beta}{2}\big(\mathcal{L}^{\mathrm{Pref}}_\mu(r_k) - \mathcal{L}^{\mathrm{Pref}}_\mu(r^\star)\big).
\end{aligned} \tag{62}$$

Combining this with the bound on term (I) from Eq. (50) and adding term (III) gives

$$\begin{aligned}
(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \le{}& \frac{H(C_{sa}(\pi) + 1)}{\beta} + \frac{HC_{sa}(\pi)}{4\alpha_{\mathrm{C}}\beta} + \frac{\beta}{2}\mathcal{L}^{\mathrm{BE}}_\mu(\pi_k, r_k, f_k) \\
&+ \frac{\beta}{2}\big(\mathcal{L}^{\mathrm{Pref}}_\mu(r_k) - \mathcal{L}^{\mathrm{Pref}}_\mu(r^\star)\big) + \mathcal{L}_\mu(\pi_k, f_k) - \mathcal{L}_\mu(\pi_k, Q^{\pi_k}).
\end{aligned} \tag{63}$$

We now transfer the population losses to empirical losses. On the same high-probability event used in the proof of Theorem 1, Eq. (53a), Eq. (53c), and Eq. (53d) hold. In addition, Lemma 20 gives the centered preference concentration bound with

$$\varepsilon_{\mathrm{Pref}} \coloneqq c_{\mathrm{C}}\frac{\log(|\mathcal{R}|/\delta)}{n}.$$

Therefore,

$$\begin{aligned}
&\frac{\beta}{2}\mathcal{L}^{\mathrm{BE}}_\mu(\pi_k, r_k, f_k) + \frac{\beta}{2}\big(\mathcal{L}^{\mathrm{Pref}}_\mu(r_k) - \mathcal{L}^{\mathrm{Pref}}_\mu(r^\star)\big) \\
&\quad\le \beta\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r_k, f_k) + \beta\big[\mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r_k) - \mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r^\star)\big] + \frac{\beta}{2}\big(\varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{Pref}} + 4\varepsilon_{\mathcal{F},\mathcal{F}}\big).
\end{aligned} \tag{64}$$

The performance-loss transfer is exactly Eq. (55):

$$\mathcal{L}_\mu(\pi_k, f_k) - \mathcal{L}_\mu(\pi_k, Q^{\pi_k}) \le \mathcal{L}_{\mathcal{D}}(\pi_k, f_k) - \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) + 2\sqrt{\varepsilon_{\mathrm{Perf}}} + O(H\sqrt{\varepsilon_{\mathcal{F}}}).$$

By the optimality of (f_k, r_k) in Algorithm 1, line 3 with the preference loss, and subtracting the common term $\beta\mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r^\star)$ from both sides,

$$\mathcal{L}_{\mathcal{D}}(\pi_k, f_k) + \beta\mathcal{L}^{\mathrm{BE}}_{\mathcal{D}}(\pi_k, r_k, f_k) + \beta\big[\mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r_k) - \mathcal{L}^{\mathrm{Pref}}_{\mathcal{D}_{\mathrm{Pref}}}(r^\star)\big]$$

$$\leq \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) + \beta \mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi_k, r^\star, f_{\pi_k}). \tag{65}$$

Combining Eq. (63), Eq. (64), the performance-loss transfer above, and Eq. (65), the empirical terms are bounded as

$$\begin{aligned}\mathcal{L}_{\mathcal{D}}(\pi_k, f_k) + \beta \mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi_k, r_k, f_k) + \beta\big[\mathcal{L}_{\mathcal{D}_{\mathrm{Pref}}}^{\mathrm{Pref}}(r_k) - \mathcal{L}_{\mathcal{D}_{\mathrm{Pref}}}^{\mathrm{Pref}}(r^\star)\big] - \mathcal{L}_{\mathcal{D}}(\pi_k, f_{\pi_k}) \\ \leq \beta \mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi_k, r^\star, f_{\pi_k}).\end{aligned}$$

By Lemma 15,

$$\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi_k, r^\star, f_{\pi_k}) \leq \varepsilon_{\mathrm{apx}},$$

so the left-hand side of the previous display is at most $\beta\varepsilon_{\mathrm{apx}}$, which is written below as $(\beta/2)\cdot 2\varepsilon_{\mathrm{apx}}$ to match the other slack terms. Therefore,

$$\begin{aligned}(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \leq \frac{H(C_{sa}(\pi)+1)}{\beta} + \frac{HC_{sa}(\pi)}{4\alpha_{\mathrm{C}}\beta} \\ + \frac{\beta}{2}\big(\varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{Pref}} + 4\varepsilon_{\mathcal{F},\mathcal{F}} + 2\varepsilon_{\mathrm{apx}}\big) + 2\sqrt{\varepsilon_{\mathrm{Perf}}} + O(H\sqrt{\varepsilon_{\mathcal{F}}}).\end{aligned} \tag{66}$$

It remains to optimize β. By Lemma 20, $\varepsilon_{\mathrm{Pref}} = \widetilde{O}(c_{\mathrm{C}}/n)$. Since $0 \leq r_h, f_h \leq 1$, Lemmas 12, 14 and 15 apply with $V_{\max} = 1$ and give

$$\varepsilon_{\mathrm{BE}} = \widetilde{O}(H/n), \qquad \varepsilon_{\mathrm{apx}} = \widetilde{O}(H/n + H\varepsilon_{\mathcal{F}}), \qquad \varepsilon_{\mathrm{Perf}} = \widetilde{O}(H^2/n),$$

and the β-dependent terms in Eq. (66) are minimized by choosing

$$\beta = \sqrt{\frac{2\big(H(C_{sa}(\pi)+1) + HC_{sa}(\pi)/(4\alpha_{\mathrm{C}})\big)}{\varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{Pref}} + 4\varepsilon_{\mathcal{F},\mathcal{F}} + 2\varepsilon_{\mathrm{apx}}}}.$$

With this choice, Eq. (66) gives the explicit intermediate bound

$$\begin{aligned}(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \leq \widetilde{O}\Bigg(\sqrt{\left[H(C_{sa}(\pi)+1) + \frac{HC_{sa}(\pi)}{4\alpha_{\mathrm{C}}}\right]\left[\frac{H + c_{\mathrm{C}}}{n} + H\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}\right]} \\ + \frac{H}{\sqrt{n}} + H\sqrt{\varepsilon_{\mathcal{F}}}\Bigg).\end{aligned}$$

Consequently, using $C_{sa}(\pi) \geq 1$ and expanding the square-root product,

$$\begin{aligned}(\mathrm{I})_k + (\mathrm{III})_k + (\mathrm{IV})_k \leq \widetilde{O}\Bigg(\sqrt{1 + \frac{1}{\alpha_{\mathrm{C}}}}\Bigg[H\sqrt{\frac{C_{sa}(\pi)}{n}} + \sqrt{\frac{c_{\mathrm{C}}HC_{sa}(\pi)}{n}} \\ + H\sqrt{C_{sa}(\pi)\varepsilon_{\mathcal{F}}} + \sqrt{HC_{sa}(\pi)\varepsilon_{\mathcal{F},\mathcal{F}}}\Bigg] + H\sqrt{\varepsilon_{\mathcal{F}}}\Bigg).\end{aligned}$$

Adding the no-regret bound Eq. (49) for term (II) and using $V_{\max} = 1$ in this cumulative-return setting, averaging over $k = 1, \ldots, K$ gives

$$\begin{aligned}J(\pi) - J(\bar{\pi}) \leq \widetilde{O}\Bigg(\sqrt{1 + \frac{1}{\alpha_{\mathrm{C}}}}\Bigg[H\sqrt{\frac{C_{sa}(\pi)}{n}} + \sqrt{\frac{c_{\mathrm{C}}HC_{sa}(\pi)}{n}} \\ + H\sqrt{C_{sa}(\pi)\varepsilon_{\mathcal{F}}} + \sqrt{HC_{sa}(\pi)\varepsilon_{\mathcal{F},\mathcal{F}}}\Bigg] + \frac{H}{\sqrt{K}} + H\sqrt{\varepsilon_{\mathcal{F}}}\Bigg).\end{aligned}$$

Here α_{C} and c_{C} are the BTL preference constants defined in Eq. (61). They depend on γ alone under this normalization, but not on n, δ, or $|\mathcal{R}|$ except through the explicit factor $\log(|\mathcal{R}|/\delta)$ already included in the concentration term. Taking $\pi = \pi^\star$ in the explicit display above proves Eq. (7).

E Lower Bound for Outcome-Based Learning

E.1 Proof of Theorem 2: Lower Bound for Sum-Reward Outcomes

Hard instances and coverage Fix $C \geq 2$, $H \geq H_0$, and $n \geq 32CH^2$ as in Theorem 2. Consider a deterministic chain with one state s_h at each stage $h \in [H]$, action set $\mathcal{A} = \{0, 1, \perp\}$, and transitions $P_h(s_{h+1} \mid s_h, a) = 1$ for every $h < H$ and $a \in \mathcal{A}$. For $\boldsymbol{\theta} \in \{0,1\}^H$ and a gap $\Delta \in (0, 1/2]$ to be chosen, define

$$r_{h,\boldsymbol{\theta}}(s_h, a) = \frac{1}{2} + \frac{\Delta}{2} z_h(a, \boldsymbol{\theta}), \qquad z_h(a, \boldsymbol{\theta}) = \begin{cases} 1, & a = \theta_h, \\ -1, & a = 1 - \theta_h, \\ 0, & a = \perp. \end{cases}$$

All rewards lie in $[0, 1]$. The optimal policy $\pi^\star_{\boldsymbol{\theta}}$ selects θ_h at each stage, so $J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) = H(1 + \Delta)/2$. The behavior policy chooses actions independently across stages with

$$\mu_h(0) = \mu_h(1) = \frac{1}{C}, \qquad \mu_h(\perp) = 1 - \frac{2}{C}.$$

The occupancy ratio on the optimal action is exactly C, and all other target occupancies are zero; hence $C_{sa}(\pi^\star_{\boldsymbol{\theta}}) = C$. When $C = 2$, omit the unused action $\perp$ to obtain a two-action instance.

Outcome observations and information bound The learner observes n independent trajectories and their labels, with

$$Y \mid a_{1:H} \sim H \operatorname{Bern}(p_{\boldsymbol{\theta}}(a_{1:H})), \qquad p_{\boldsymbol{\theta}}(a_{1:H}) = \frac{1}{2} + \frac{\Delta}{2H} \sum_{h=1}^{H} z_h(a_h, \boldsymbol{\theta}).$$

Thus $Y \in \{0, H\}$ and $\mathbb{E}[Y \mid a_{1:H}] = \sum_h r_{h,\boldsymbol{\theta}}(s_h, a_h)$, as required by the bounded unbiased outcome model. Since $\Delta \leq 1/2$, all success probabilities lie in $[1/4, 3/4]$. Let $P_{\boldsymbol{\theta}}$ denote the law of one labeled trajectory. For two parameters $\boldsymbol{\theta}, \boldsymbol{\theta}'$, let $I = \{h : \theta_h \neq \theta'_h\}$ and $m = |I|$. The trajectory marginal is the same under both parameters, and

$$p_{\boldsymbol{\theta}} - p_{\boldsymbol{\theta}'} = \frac{\Delta}{H} \sum_{h \in I} X_h, \qquad X_h = \frac{z_h(a_h, \boldsymbol{\theta}) - z_h(a_h, \boldsymbol{\theta}')}{2}.$$

Under μ, these variables are independent, with $\mathbb{E}_\mu[X_h] = 0$ and $\mathbb{E}_\mu[X_h^2] = 2/C$. Therefore

$$\mathbb{E}_\mu[(p_{\boldsymbol{\theta}} - p_{\boldsymbol{\theta}'})^2] = \frac{2\Delta^2 m}{CH^2}.$$

Using $\mathrm{KL}(\operatorname{Bern}(p) \| \operatorname{Bern}(q)) \leq 8(p - q)^2$ for $p, q \in [1/4, 3/4]$, and independence of the samples, we obtain

$$\mathrm{KL}(P_{\boldsymbol{\theta}}^{\otimes n} \| P_{\boldsymbol{\theta}'}^{\otimes n}) \leq 8n\, \mathbb{E}_\mu[(p_{\boldsymbol{\theta}} - p_{\boldsymbol{\theta}'})^2] = \frac{16n\Delta^2 m}{CH^2}.$$

Packing and estimation error By the Varshamov–Gilbert lemma, there exists $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\} \subseteq \{0,1\}^H$ with $\log M \geq (H/8) \log 2$ and pairwise Hamming distance at least $H/8$. Choose

$$\Delta = \frac{H}{32} \sqrt{\frac{C}{n}}.$$

The condition $n \geq 32CH^2$ ensures $\Delta \leq 1/2$. Since $m \leq H$, the pairwise dataset KL is at most $H/64$ and hence at most $(\log M)/4$. Under the metric $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \Delta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1$, the packing separation is at least $\delta = \Delta H/8$. Choose the universal constant $H_0 = 32$. Then $\log 2 \leq (\log M)/4$ for $H \geq H_0$, and Fano's inequality (Yu, 1997) yields

$$\begin{aligned} \mathfrak{R}^* &:= \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\boldsymbol{\theta}}[\Delta \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1] \\ &\geq \frac{\delta}{2} \left(1 - \frac{M^{-2} \sum_{j,k} \mathrm{KL}(P_{\boldsymbol{\theta}_j}^{\otimes n} \| P_{\boldsymbol{\theta}_k}^{\otimes n}) + \log 2}{\log M} \right) \geq \frac{\Delta H}{32}. \end{aligned}$$

Reduction from policy learning For an arbitrary output policy, including a mixture or a policy depending on past states and actions, let $q_h(a)$ be its marginal action probability on this deterministic chain. These marginals depend only on the output policy, since transitions are independent of $\boldsymbol{\theta}$. Decode $\widehat{\theta}_h$ as the more probable of actions 0 and 1, ignoring the dummy and breaking ties arbitrarily. If this decoder is wrong, $q_h(\theta_h) \le q_h(1-\theta_h)$. The expected stage-wise regret is

$$\Delta q_h(1-\theta_h) + \frac{\Delta}{2} q_h(\perp) = \frac{\Delta}{2}\big(1 - q_h(\theta_h) + q_h(1-\theta_h)\big) \ge \frac{\Delta}{2}.$$

Here $q_h(\perp) = 0$ for the two-action construction at $C = 2$. Summing over incorrectly decoded coordinates gives

$$J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi}) \ge \frac{\Delta}{2}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1.$$

Consequently,

$$\inf_{\widehat{\pi}} \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{\theta}}[J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi})] \ge \frac{1}{2}\mathfrak{R}^* \ge \frac{\Delta H}{64} = \frac{H^2}{2048}\sqrt{\frac{C}{n}}.$$

Taking the instance class indexed by Θ proves Theorem 2 with $c = 1/2048$.

Sample-complexity interpretation For sufficiently small ε, achieving worst-case expected suboptimality at most ε therefore requires

$$n \ge \frac{H^4 C}{2048^2 \varepsilon^2}.$$

The cutoff $n \ge 32CH^2$ does not affect this small-error conclusion: at $n_0 = \lceil 32CH^2 \rceil$ the lower bound is a positive universal constant times H, and a learner with more samples can always ignore the extra samples. To compare with the high-probability upper bound in Theorem 1, set $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$, take $K \ge n$, and choose confidence level $\delta_{\mathrm{conf}} = \varepsilon/(2H)$ and enough trajectories that its high-probability error bound is at most $\varepsilon/2$. Since policy regret lies in $[0, H]$, the expected regret is then at most $\varepsilon/2 + H\delta_{\mathrm{conf}} = \varepsilon$. This changes only logarithmic factors and yields the $\widetilde{\Theta}(H^4C/\varepsilon^2)$ expected-error comparison in Remark 1.

E.2 Proof of Theorem 5: Exponential Lower Bound for All-Success Aggregation

We prove the lower bound for the all-success outcome described in Example 4. We construct a hard family of deterministic finite-horizon MDPs with binary rewards. Let the horizon be H, the state space be

$$\mathcal{S} = \{s_1, \dots, s_H\},$$

and the action space be $\mathcal{A} = \{0, 1\}$. The transition kernel is deterministic:

$$P_h(s_{h+1} \mid s_h, a) = 1, \qquad \forall h \in [H-1],\ a \in \mathcal{A}.$$

Thus the state is again only the time index, and the learner chooses one binary action at each step.

The environments differ only in their reward functions. For a parameter

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_H) \in \{0,1\}^H,$$

define the binary reward at step h by

$$r_{h,\boldsymbol{\theta}}(s_h, a) = \mathbf{1}\{a = \theta_h\} \in \{0, 1\}, \tag{67}$$

and define the observed outcome using the all-success aggregation of Example 4:

$$Y(\tau) = R_{\boldsymbol{\theta}}(\tau) = \prod_{h=1}^{H} r_{h,\boldsymbol{\theta}}(s_h, a_h) = \mathbf{1}\{a_h = \theta_h\ \forall h\}.$$

The optimal policy $\pi^\star_{\boldsymbol{\theta}}$ selects action θ_h at every step, and therefore

$$J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) = 1.$$

We assume the behavior policy μ chooses the two actions uniformly at every step. Identifying a trajectory with its action vector

$$\mathbf{x} = (a_1, \dots, a_H) \in \{0,1\}^H,$$

we have $\mathbf{x} \sim \mathrm{Unif}(\{0,1\}^H)$ and the observed label is

$$y = \mathbf{1}\{\mathbf{x} = \boldsymbol{\theta}\}.$$

Coverage Since μ is uniform,

$$d_h^{\mu}(s_h, a) = \frac{1}{2}.$$

For the optimal policy, $d_h^{\pi^\star_{\boldsymbol{\theta}}}(s_h, \theta_h) = 1$, and hence

$$C_{sa}(\pi^\star_{\boldsymbol{\theta}}) = \max_{h,s,a} \frac{d_h^{\pi^\star_{\boldsymbol{\theta}}}(s,a)}{d_h^{\mu}(s,a)} = 2.$$

At the trajectory level, however, $\mu(\tau) = 2^{-H}$ for every trajectory, while $\pi^\star_{\boldsymbol{\theta}}$ puts all its mass on the single trajectory $\mathbf{x} = \boldsymbol{\theta}$. Thus

$$C_\tau(\pi^\star_{\boldsymbol{\theta}}) \coloneqq \max_{\tau} \frac{\mathbb{P}_{\pi^\star_{\boldsymbol{\theta}}}[\tau]}{\mathbb{P}_{\mu}[\tau]} = 2^H.$$

The construction therefore has constant state–action concentrability but exponential trajectory concentrability.

Reduction to parameter estimation For any fixed output policy $\widehat{\pi}$, including a mixture or a history-dependent policy, let $q_h(a) = \mathbb{P}_{\widehat{\pi}}(a_h = a)$ on the deterministic chain. These marginals are defined without reference to $\boldsymbol{\theta}$, and

$$J_{\boldsymbol{\theta}}(\widehat{\pi}) = \mathbb{P}_{\widehat{\pi}}(a_{1:H} = \boldsymbol{\theta}) \le \min_{h\in[H]} q_h(\theta_h), \qquad J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) = 1. \tag{68}$$

Given $\widehat{\pi}$, define an induced estimator $\widehat{\boldsymbol{\theta}}(\widehat{\pi})$ by choosing the most likely action at every step, breaking ties arbitrarily:

$$\widehat{\theta}_h(\widehat{\pi}) \coloneqq \arg\max_{a\in\{0,1\}} q_h(a) \in \{0,1\}, \qquad \widehat{\boldsymbol{\theta}}(\widehat{\pi}) \coloneqq (\widehat{\theta}_1, \dots, \widehat{\theta}_H). \tag{69}$$

The following lemma converts value suboptimality into zero-one parameter-estimation loss.

Lemma 22 (Reduction from value gap to zero-one loss)**.** *For every policy $\widehat{\pi}$ and every $\boldsymbol{\theta} \in \{0,1\}^H$,*

$$J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi}) \ge \frac{1}{2}\mathbf{1}\{\widehat{\boldsymbol{\theta}}(\widehat{\pi}) \ne \boldsymbol{\theta}\}.$$

Proof. If $\widehat{\boldsymbol{\theta}}(\widehat{\pi}) = \boldsymbol{\theta}$, the right-hand side is zero and there is nothing to prove. Otherwise, there exists $h^\star \in [H]$ such that $\widehat{\theta}_{h^\star} \ne \theta_{h^\star}$. By Eq. (69),

$$q_{h^\star}(\theta_{h^\star}) \le \frac{1}{2}.$$

By Eq. (68),

$$J_{\boldsymbol{\theta}}(\widehat{\pi}) \le q_{h^\star}(\theta_{h^\star}) \le \frac{1}{2},$$

so $J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi}) \ge 1 - 1/2 = 1/2$. □

Indistinguishability via total variation Let $P_{\boldsymbol{\theta}}$ be the law of one sample $(\mathbf{x}, y)$, where $\mathbf{x} \sim \mathrm{Unif}(\{0,1\}^H)$ and $y = \mathbf{1}\{\mathbf{x} = \boldsymbol{\theta}\}$. Let $P_{\boldsymbol{\theta}}^{\otimes n}$ denote the law of n i.i.d. samples. We now bound the total variation distance between the observation laws corresponding to two distinct parameters.

Lemma 23 (Trajectory-coverage TV bound)**.** *For any distinct $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \{0,1\}^H$,*

$$\mathrm{TV}(P_{\boldsymbol{\theta}_0}, P_{\boldsymbol{\theta}_1}) = 2^{1-H}, \qquad \mathrm{TV}(P_{\boldsymbol{\theta}_0}^{\otimes n}, P_{\boldsymbol{\theta}_1}^{\otimes n}) \le n \cdot 2^{1-H}.$$

Proof. Under both $P_{\boldsymbol{\theta}_0}$ and $P_{\boldsymbol{\theta}_1}$, the marginal distribution of $\mathbf{x}$ is uniform. If $\mathbf{x} \notin \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$, then $y = 0$ under both laws. If $\mathbf{x} = \boldsymbol{\theta}_0$, then $y = 1$ under $P_{\boldsymbol{\theta}_0}$ and $y = 0$ under $P_{\boldsymbol{\theta}_1}$; the case $\mathbf{x} = \boldsymbol{\theta}_1$ is symmetric. Therefore

$$\mathrm{TV}(P_{\boldsymbol{\theta}_0}, P_{\boldsymbol{\theta}_1}) = \frac{1}{2}\sum_{\mathbf{x},y} \left|P_{\boldsymbol{\theta}_0}(\mathbf{x}, y) - P_{\boldsymbol{\theta}_1}(\mathbf{x}, y)\right| = \frac{1}{2} \cdot 2 \cdot \left(2^{-H} + 2^{-H}\right) = 2^{1-H}.$$

For the n-sample bound, couple the two experiments by drawing

$$\mathbf{x}_1, \dots, \mathbf{x}_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{0,1\}^H), \qquad y_i^{(b)} \coloneqq \mathbf{1}\{\mathbf{x}_i = \boldsymbol{\theta}_b\}, \quad b \in \{0,1\}.$$

Under this coupling,

$$X^{(b)} \coloneqq (\mathbf{x}_1, y_1^{(b)}, \dots, \mathbf{x}_n, y_n^{(b)})$$

has marginal law $P_{\boldsymbol{\theta}_b}^{\otimes n}$ for each $b \in \{0,1\}$. Define

$$E \coloneqq \Big\{\mathbf{x}_i \notin \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\} \text{ for all } i \in [n]\Big\}.$$

On E, all labels satisfy $y_i^{(0)} = y_i^{(1)} = 0$, so $X^{(0)} = X^{(1)}$. Therefore, for any event A in the observation space,

$$\mathbf{1}\{X^{(0)} \in A\} - \mathbf{1}\{X^{(1)} \in A\} = \big(\mathbf{1}\{X^{(0)} \in A\} - \mathbf{1}\{X^{(1)} \in A\}\big)\mathbf{1}\{E^c\} \le \mathbf{1}\{E^c\},$$

where the first equality uses $X^{(0)} = X^{(1)}$ on E. Taking expectations gives

$$P_{\boldsymbol{\theta}_0}^{\otimes n}(A) - P_{\boldsymbol{\theta}_1}^{\otimes n}(A) = \mathbb{P}[X^{(0)} \in A] - \mathbb{P}[X^{(1)} \in A] \le \mathbb{P}[E^c].$$

Taking the supremum over A,

$$\mathrm{TV}\big(P_{\boldsymbol{\theta}_0}^{\otimes n}, P_{\boldsymbol{\theta}_1}^{\otimes n}\big) = \sup_A \big[P_{\boldsymbol{\theta}_0}^{\otimes n}(A) - P_{\boldsymbol{\theta}_1}^{\otimes n}(A)\big] \le \mathbb{P}[E^c].$$

By a union bound,

$$\mathbb{P}[E^c] \le \sum_{i=1}^{n} \mathbb{P}\big[\mathbf{x}_i \in \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}\big] = n \cdot 2 \cdot 2^{-H} = n \cdot 2^{1-H},$$

which completes the proof. □

Combining the lemmas We now combine the reduction and the TV bound by the two-point Le Cam method. Fix any two distinct parameters $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \{0,1\}^H$. By Lemma 22, for every policy $\widehat{\pi}$ and $i \in \{0,1\}$,

$$J_{\boldsymbol{\theta}_i}(\pi^\star_{\boldsymbol{\theta}_i}) - J_{\boldsymbol{\theta}_i}(\widehat{\pi}) \ge \frac{1}{2}\mathbf{1}\big\{\widehat{\boldsymbol{\theta}}(\widehat{\pi}) \ne \boldsymbol{\theta}_i\big\} = \frac{1}{2}\ell_{01}\big(\widehat{\boldsymbol{\theta}}(\widehat{\pi}), \boldsymbol{\theta}_i\big),$$

where

$$\ell_{01}(\boldsymbol{\theta}, \boldsymbol{\theta}') \coloneqq \mathbf{1}\{\boldsymbol{\theta} \ne \boldsymbol{\theta}'\}.$$

Taking expectations under $P_{\boldsymbol{\theta}_i}^{\otimes n}$ and then taking the supremum over $\boldsymbol{\theta} \in \{0,1\}^H$,

$$\inf_{\widehat{\pi}} \sup_{\boldsymbol{\theta} \in \{0,1\}^H} \mathbb{E}\big[J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi})\big] \ge \frac{1}{2} \inf_{\widehat{\boldsymbol{\theta}}} \max_{i \in \{0,1\}} \mathbb{E}_{P_{\boldsymbol{\theta}_i}^{\otimes n}}\big[\ell_{01}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_i)\big], \tag{70}$$

where the infimum on the right ranges over all possibly randomized estimators $\widehat{\boldsymbol{\theta}}$ based on the data. Since ℓ_{01} is a metric and $\ell_{01}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = 1$, the two-point Le Cam inequality (see Yu, 1997) gives

$$\inf_{\widehat{\boldsymbol{\theta}}} \max_{i \in \{0,1\}} \mathbb{E}_{P_{\boldsymbol{\theta}_i}^{\otimes n}}\big[\ell_{01}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}_i)\big] \ge \frac{1}{2}\big(1 - \mathrm{TV}(P_{\boldsymbol{\theta}_0}^{\otimes n}, P_{\boldsymbol{\theta}_1}^{\otimes n})\big). \tag{71}$$

Combining Eq. (70), Eq. (71), and Lemma 23,

$$\inf_{\widehat{\pi}} \sup_{\boldsymbol{\theta} \in \{0,1\}^H} \mathbb{E}\big[J_{\boldsymbol{\theta}}(\pi^\star_{\boldsymbol{\theta}}) - J_{\boldsymbol{\theta}}(\widehat{\pi})\big] \ge \frac{1}{4}\big(1 - \mathrm{TV}(P_{\boldsymbol{\theta}_0}^{\otimes n}, P_{\boldsymbol{\theta}_1}^{\otimes n})\big) \ge \frac{1}{4}\big(1 - n \cdot 2^{1-H}\big),$$

as desired. In particular, if $n \le 2^{H-2}$, then the right-hand side is at least $1/8$. Therefore achieving expected suboptimality below any constant $\varepsilon < 1/8$ requires

$$n = \Omega(2^H) = \Omega\big(C_\tau(\pi^\star)\big).$$

This completes the proof of Theorem 5.

F GENERALIZED REINFORCEMENT LEARNING THEORY

This section develops the RL theory underlying the generalized objective introduced in Section 5. The MDP, policy, and occupancy notation is otherwise the same as in Section 2. We restrict attention to Markov policies $\pi = (\pi_h)_{h=1}^H$, where each $\pi_h : \mathcal{S}_h \to \Delta(\mathcal{A}_h)$ depends only on the current state. All expectations $\mathbb{E}_{\tau\sim\pi}$ and occupancies d_h^π are taken under the trajectory law induced by such a Markov policy and the fixed transition kernel. As in the main text, we work with the stage-wise affine-in-continuation aggregation

$$\sigma_h(u, v) = a_h(u)v + b_h(u),$$

with terminal constant $g \in \mathbb{R}$. For any candidate reward $r = (r_h)_{h=1}^H$, define

$$\begin{aligned} R_{H+1}(\cdot; r) &\coloneqq g, \\ R_h(\tau_h; r) &\coloneqq \sigma_h\big(r_h(s_h, a_h), R_{h+1}(\tau_{h+1}; r)\big), \qquad h = H, H-1, \dots, 1, \end{aligned}$$

and set $R(\tau; r) \coloneqq R_1(\tau_1; r)$ and $J_r(\pi) \coloneqq \mathbb{E}_{\tau\sim\pi}[R(\tau; r)]$. For the true environment reward $r^\star$, we recover the objective of the main text as $J(\pi) = J_{r^\star}(\pi)$. We establish Bellman evaluation, Bellman optimality, greedy optimality, and the generalized performance-difference lemma with the reward argument kept explicit, as these results are used throughout Appendix G.

F.1 GENERALIZED VALUE FUNCTIONS AND BELLMAN RECURSION

Since $\sigma_h(u, v)$ is affine in its continuation argument v, conditional expectations commute with the aggregation in that coordinate. Thus the generalized action-value and state-value functions for a policy π and reward r,

$$\begin{aligned} Q_h^{\pi,r}(s, a) &\coloneqq \mathbb{E}_\pi[R_h(\tau_h; r) \mid s_h = s, a_h = a], \\ V_h^{\pi,r}(s) &\coloneqq \mathbb{E}_\pi[R_h(\tau_h; r) \mid s_h = s], \end{aligned}$$

satisfy the backward Bellman recursion

$$\begin{aligned} Q_H^{\pi,r}(s, a) &= \sigma_H\big(r_H(s, a), g\big), \\ Q_h^{\pi,r}(s, a) &= \sigma_h\big(r_h(s, a), \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,r}(s')]\big), \qquad h \in [H-1], \\ V_h^{\pi,r}(s) &= \mathbb{E}_{a\sim\pi_h(\cdot|s)}[Q_h^{\pi,r}(s, a)], \end{aligned}$$

with $V^{\pi,r}(s) \coloneqq V_1^{\pi,r}(s)$.

For any collection of stage-wise functions

$$f = (f_1, \dots, f_H), \qquad f_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R},$$

define the associated state-value surrogates by

$$V_{H+1}^{\pi,f} \coloneqq g, \tag{72a}$$

$$V_h^{\pi,f}(s) \coloneqq \mathbb{E}_{a\sim\pi_h(\cdot|s)}[f_h(s, a)], \qquad h \in [H], \tag{72b}$$

$$V_{H+1}^{f} \coloneqq g, \tag{72c}$$

$$V_h^{f}(s) \coloneqq \max_{a\in\mathcal{A}} f_h(s, a), \qquad h \in [H]. \tag{72d}$$

We also use the convention $V_{H+1}^{\pi,r} \coloneqq g$.

F.2 GENERALIZED BELLMAN EQUATIONS AND OPTIMALITY

Definition 3 (Generalized Bellman equation)**.** *Fix a policy π and reward r. We say that $f = (f_1, \dots, f_H)$ satisfies the Bellman equation for (π, r) if, for every $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$,*

$$f_H(s, a) = \sigma_H\big(r_H(s, a), g\big), \qquad f_h(s, a) = \sigma_h\big(r_h(s, a), \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,f}(s')]\big), \quad h \in [H-1]. \tag{73}$$

Definition 4 (Generalized Bellman optimality equation)**.** *Fix a reward r. We say that $f = (f_1, \dots, f_H)$ satisfies the Bellman optimality equation for r if, for every $h \in [H]$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$,*

$$f_H(s,a) = \sigma_H\big(r_H(s,a), g\big), \qquad f_h(s,a) = \sigma_h\big(r_h(s,a), \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^f(s')]\big), \quad h \in [H-1]. \tag{74}$$

Policy evaluation only uses affinity in the continuation value. For optimal control, we use $a_h(u) \geq 0$, so $v \mapsto \sigma_h(u,v)$ is nondecreasing. The statistical upper bound additionally assumes $a_h(u) \leq 1$.

Proposition 24 (Characterization of $Q^{\pi,r}$)**.** *Fix a policy π and reward r. A collection $f = (f_1, \dots, f_H)$ satisfies the Bellman equation Eq. (73) for (π, r) if and only if*

$$f_h(s,a) = Q_h^{\pi,r}(s,a), \qquad \forall h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}.$$

Equivalently, the Bellman equation for (π, r) has a unique solution, namely $Q^{\pi,r}$.

Proof. The "if" direction follows immediately from the recursive definitions of $Q^{\pi,r}$ and $V^{\pi,r}$.

For the "only if" direction, suppose f satisfies Eq. (73). We prove by backward induction on h that

$$f_h(s,a) = Q_h^{\pi,r}(s,a), \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

At stage H, we have

$$f_H(s,a) = \sigma_H\big(r_H(s,a), g\big) = Q_H^{\pi,r}(s,a).$$

Now assume that for some $h < H$,

$$f_{h+1}(s,a) = Q_{h+1}^{\pi,r}(s,a), \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A}.$$

Then for every $s' \in \mathcal{S}$,

$$V_{h+1}^{\pi,f}(s') = \mathbb{E}_{a' \sim \pi_{h+1}(\cdot|s')}[f_{h+1}(s',a')] = \mathbb{E}_{a' \sim \pi_{h+1}(\cdot|s')}[Q_{h+1}^{\pi,r}(s',a')] = V_{h+1}^{\pi,r}(s').$$

Substituting this into Eq. (73) yields

$$f_h(s,a) = \sigma_h\big(r_h(s,a), \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,r}(s')]\big) = Q_h^{\pi,r}(s,a).$$

This completes the induction. □

In this subsection, optimization ranges over all Markov policies. We now define the optimal action-value function for reward r by

$$Q_h^{\star,r}(s,a) := \sup_{\pi} Q_h^{\pi,r}(s,a), \qquad h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A},$$

and the optimal state-value function by

$$V_h^{\star,r}(s) := \max_{a \in \mathcal{A}} Q_h^{\star,r}(s,a), \qquad h \in [H],$$

and set $V_{H+1}^{\star,r} := g$.

Theorem 25 (Bellman optimality equation and greedy optimality)**.** *Fix a reward r, and suppose $a_h(u) \geq 0$ holds. Then:*

1. *The Bellman optimality equation Eq. (74) has a unique solution $\bar{Q} = (\bar{Q}_1, \dots, \bar{Q}_H)$.*
2. *This solution coincides with the optimal action-value function:*

   $$\bar{Q}_h(s,a) = Q_h^{\star,r}(s,a), \qquad \forall h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}.$$

   Hence a collection $f = (f_1, \dots, f_H)$ satisfies Eq. (74) if and only if $f = Q^{\star,r}$.
3. *Any greedy policy with respect to $\bar{Q}$, namely any policy π^{gr} satisfying*

   $$\pi_h^{\mathrm{gr}}(a \mid s) = 0, \qquad \forall a \notin \arg\max_{a' \in \mathcal{A}} \bar{Q}_h(s,a'),$$

   for every $h \in [H]$ and $s \in \mathcal{S}$, is optimal for J_r. Equivalently, any policy greedy with respect to $Q^{\star,r}$ is optimal for J_r.

Proof. We divide the proof into three steps.

Step 1: Existence and uniqueness of a solution to the Bellman optimality equation The equation Eq. (74) can be solved uniquely by backward recursion.

At stage H, the equation uniquely determines

$$\bar{Q}_H(s,a) = \sigma_H\big(r_H(s,a), g\big).$$

Once $\bar{Q}_{h+1}, \dots, \bar{Q}_H$ are known, Eq. (74) uniquely determines $\bar{Q}_h$ through

$$\bar{Q}_h(s,a) = \sigma_h\big(r_h(s,a), \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[\bar{V}_{h+1}(s')]\big), \qquad \bar{V}_{h+1}(s') \coloneqq \max_{a'} \bar{Q}_{h+1}(s',a').$$

Thus the solution exists and is unique.

Step 2: Every policy is dominated by $\bar{Q}$ We claim that for every policy π,

$$Q_h^{\pi,r}(s,a) \le \bar{Q}_h(s,a), \qquad \forall h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}. \tag{75}$$

We prove this by backward induction on h.

At stage H, for every policy π,

$$Q_H^{\pi,r}(s,a) = \sigma_H\big(r_H(s,a), g\big) = \bar{Q}_H(s,a),$$

so Eq. (75) holds.

Now assume Eq. (75) holds at stage $h+1$. Then for every $s' \in \mathcal{S}$,

$$V_{h+1}^{\pi,r}(s') = \mathbb{E}_{a'\sim\pi_{h+1}(\cdot|s')}[Q_{h+1}^{\pi,r}(s',a')] \le \max_{a'} \bar{Q}_{h+1}(s',a') = \bar{V}_{h+1}(s').$$

Taking expectation with respect to $P_h(\cdot \mid s,a)$ in the preceding display and using $a_h(u) \ge 0$ gives

$$\begin{aligned} Q_h^{\pi,r}(s,a) &= \sigma_h\big(r_h(s,a), \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,r}(s')]\big) \\ &\le \sigma_h\big(r_h(s,a), \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[\bar{V}_{h+1}(s')]\big) = \bar{Q}_h(s,a). \end{aligned}$$

This proves Eq. (75).

Step 3: A greedy policy attains $\bar{Q}$ Let π^{gr} be any greedy policy with respect to $\bar{Q}$, and define

$$\bar{V}_h(s) \coloneqq \max_a \bar{Q}_h(s,a).$$

We claim that

$$Q_h^{\pi^{\mathrm{gr}},r}(s,a) = \bar{Q}_h(s,a), \qquad \forall h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}, \tag{76}$$

and therefore

$$V_h^{\pi^{\mathrm{gr}},r}(s) = \bar{V}_h(s), \qquad \forall h \in [H],\ s \in \mathcal{S}. \tag{77}$$

Again we use backward induction.

At stage H, the equality is immediate since both sides equal

$$\sigma_H\big(r_H(s,a), g\big).$$

Assume Eq. (76) holds at stage $h+1$. Then for every $s' \in \mathcal{S}$,

$$V_{h+1}^{\pi^{\mathrm{gr}},r}(s') = \mathbb{E}_{a'\sim\pi_{h+1}^{\mathrm{gr}}(\cdot|s')}[Q_{h+1}^{\pi^{\mathrm{gr}},r}(s',a')].$$

Since π_{h+1}^{gr} places mass only on maximizers of $\bar{Q}_{h+1}(s',\cdot)$ and $Q_{h+1}^{\pi^{\mathrm{gr}},r} = \bar{Q}_{h+1}$ by the induction hypothesis,

$$V_{h+1}^{\pi^{\mathrm{gr}},r}(s') = \max_{a'} \bar{Q}_{h+1}(s',a') = \bar{V}_{h+1}(s').$$

Using the recursion that defines $Q_h^{\pi^{\mathrm{gr}},r}$, we obtain

$$Q_h^{\pi^{\mathrm{gr}},r}(s,a) = \sigma_h\big(r_h(s,a), \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[\bar{V}_{h+1}(s')]\big) = \bar{Q}_h(s,a),$$

which proves Eq. (76). Then

$$V_h^{\pi^{\mathrm{gr}},r}(s) = \mathbb{E}_{a\sim\pi_h^{\mathrm{gr}}(\cdot|s)}[Q_h^{\pi^{\mathrm{gr}},r}(s,a)] = \max_a \bar{Q}_h(s,a) = \bar{V}_h(s),$$

proving Eq. (77).

Combining Steps 2 and 3, for every policy π,

$$Q_h^{\pi,r}(s,a) \le \bar{Q}_h(s,a) = Q_h^{\pi^{\mathrm{gr}},r}(s,a), \qquad \forall h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}.$$

Hence

$$\bar{Q}_h(s,a) = \sup_\pi Q_h^{\pi,r}(s,a) = Q_h^{\star,r}(s,a),$$

so $\bar{Q} = Q^{\star,r}$. This proves item 2. Item 3 follows from

$$V_h^{\pi^{\mathrm{gr}},r}(s) = \bar{V}_h(s) = \max_a Q_h^{\star,r}(s,a) = V_h^{\star,r}(s), \qquad \forall h \in [H],\ s \in \mathcal{S}.$$

□

Corollary 26. *Fix a reward r. Under $a_h(u) \ge 0$, a collection $f = (f_1, \dots, f_H)$ satisfies the Bellman optimality equation Eq. (74) for r if and only if*

$$f_h(s,a) = Q_h^{\star,r}(s,a), \qquad \forall h \in [H],\ s \in \mathcal{S},\ a \in \mathcal{A}.$$

Moreover, any policy greedy with respect to f is optimal for J_r.

Proof. If f satisfies Eq. (74), then Theorem 25 implies $f = Q^{\star,r}$. Conversely, $Q^{\star,r}$ satisfies Eq. (74) because it is the unique solution established in Theorem 25. The greedy-optimality statement follows from the same theorem. □

F.3 Generalized Performance Difference Lemma

The following lemma extends the standard performance difference lemma (Lemma 16) to the generalized objective. For a fixed reward r, the prefix scaling factors $\prod_{j<h} a_j(r_j(s_j,a_j))$ appear naturally as weights on the per-step advantage terms.

Lemma 27 (Generalized performance difference lemma)**.** *Under the generalized objective with $\sigma_h(u,v) = a_h(u)v + b_h(u)$, for any reward r and any two policies π and π',*

$$J_r(\pi) - J_r(\pi') = \sum_{h=1}^{H} \mathbb{E}_{\tau\sim\pi}\left[\left(\prod_{j=1}^{h-1} a_j\big(r_j(s_j,a_j)\big)\right) A_h^{\pi',r}(s_h,a_h)\right],$$

where $A_h^{\pi',r}(s,a) := Q_h^{\pi',r}(s,a) - V_h^{\pi',r}(s)$ and we use the convention $\prod_{j=1}^{0}(\cdot) = 1$.

Proof. We prove by backward induction that for every $h \in [H]$ and $s \in \mathcal{S}$,

$$V_h^{\pi,r}(s) - V_h^{\pi',r}(s) = \sum_{h'=h}^{H} \mathbb{E}_{\tau\sim\pi}\left[\left(\prod_{j=h}^{h'-1} a_j(r_j(s_j,a_j))\right) A_{h'}^{\pi',r}(s_{h'},a_{h'}) \,\middle|\, s_h = s\right]. \tag{78}$$

The lemma then follows by setting $h = 1$ and $s = s_1$.

Base case ($h = H$) At $h = H$ the only term in the sum is $h' = H$ with an empty product:

$$V_H^{\pi,r}(s) - V_H^{\pi',r}(s) = \mathbb{E}_{a\sim\pi_H}[Q_H^{\pi',r}(s,a) - V_H^{\pi',r}(s)] = \mathbb{E}_{a\sim\pi_H}[A_H^{\pi',r}(s,a)].$$

It remains to verify this equality. Since $V_{H+1}^{\pi,r} = V_{H+1}^{\pi',r} = g$, both $Q_H^{\pi,r}(s,a)$ and $Q_H^{\pi',r}(s,a)$ equal $\sigma_H(r_H(s,a),g)$, so $Q_H^{\pi,r} = Q_H^{\pi',r}$. Therefore

$$\begin{aligned} V_H^{\pi,r}(s) - V_H^{\pi',r}(s) &= \mathbb{E}_{a\sim\pi_H}[Q_H^{\pi,r}(s,a)] - \mathbb{E}_{a\sim\pi'_H}[Q_H^{\pi',r}(s,a)] \\ &= \mathbb{E}_{a\sim\pi_H}[Q_H^{\pi',r}(s,a)] - V_H^{\pi',r}(s) = \mathbb{E}_{a\sim\pi_H}[A_H^{\pi',r}(s,a)]. \end{aligned}$$

Inductive step Assume Eq. (78) holds at $h+1$. At stage h, using the Bellman equation

$$Q_h^{\pi,r}(s,a) = a_h(r_h(s,a))\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\pi,r}(s')] + b_h(r_h(s,a)),$$

(and analogously for π'), we get

$$Q_h^{\pi,r}(s,a) - Q_h^{\pi',r}(s,a) = a_h(r_h(s,a))\mathbb{E}_{s'\sim P_h}[V_{h+1}^{\pi,r}(s') - V_{h+1}^{\pi',r}(s')].$$

Therefore

$$\begin{aligned}
V_h^{\pi,r}(s) - V_h^{\pi',r}(s) &= \mathbb{E}_{a\sim\pi_h}[Q_h^{\pi,r}(s,a)] - V_h^{\pi',r}(s) \\
&= \mathbb{E}_{a\sim\pi_h}[Q_h^{\pi',r}(s,a) - V_h^{\pi',r}(s)] + \mathbb{E}_{a\sim\pi_h}[Q_h^{\pi,r}(s,a) - Q_h^{\pi',r}(s,a)] \\
&= \mathbb{E}_{a\sim\pi_h}[A_h^{\pi',r}(s,a)] + \mathbb{E}_{a\sim\pi_h}\left[a_h(r_h(s,a))\mathbb{E}_{s'\sim P_h}[V_{h+1}^{\pi,r}(s') - V_{h+1}^{\pi',r}(s')]\right].
\end{aligned}$$

Substituting the induction hypothesis Eq. (78) at $h+1$ into the second term gives

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi_h}\left[a_h(r_h(s,a))\mathbb{E}_{s'}\sum_{h'=h+1}^{H}\mathbb{E}_{\tau\sim\pi}\left[\prod_{j=h+1}^{h'-1}a_j(r_j(s_j,a_j))\cdot A_{h'}^{\pi',r}(s_{h'},a_{h'})\middle| s_{h+1}=s'\right]\right] \\
&= \sum_{h'=h+1}^{H}\mathbb{E}_{\tau\sim\pi}\left[a_h(r_h(s_h,a_h))\prod_{j=h+1}^{h'-1}a_j(r_j(s_j,a_j))\cdot A_{h'}^{\pi',r}(s_{h'},a_{h'})\middle| s_h=s\right] \\
&= \sum_{h'=h+1}^{H}\mathbb{E}_{\tau\sim\pi}\left[\prod_{j=h}^{h'-1}a_j(r_j(s_j,a_j))\cdot A_{h'}^{\pi',r}(s_{h'},a_{h'})\middle| s_h=s\right].
\end{aligned}$$

Combining gives

$$V_h^{\pi,r}(s) - V_h^{\pi',r}(s) = \sum_{h'=h}^{H}\mathbb{E}_{\tau\sim\pi}\left[\prod_{j=h}^{h'-1}a_j(r_j(s_j,a_j))\cdot A_{h'}^{\pi',r}(s_{h'},a_{h'})\middle| s_h=s\right],$$

completing the induction. □

G Proof of the Generalized Outcome-Based Upper Bound

This appendix proves Theorem 9. The argument follows the same blueprint as the proof of Theorem 1 in Appendix C: a regret decomposition produces three population losses, each is transferred to the data via concentration, and the resulting AM–GM bound is balanced by tuning β. The two new ingredients relative to the additive-return analysis are:

- the trajectory return is no longer the per-step sum, so the per-step gap $r_h^\star - r_{k,h}$ is replaced by the Bellman-operator gap $\Delta_h^{r_k} := (\mathcal{T}_{r^\star}^{\widehat{\pi}} f)_h - (\mathcal{T}_{r_k}^{\widehat{\pi}} f)_h$, and trajectory return differences no longer decompose into per-step increments;
- policy occupancies are weighted by the trajectory factor $W_h^r(\tau) = \prod_{j<h} a_j(r_j(s_j,a_j))$ from Proposition 6, which is in general policy- and reward-dependent.

Both are instances of the same principle: the generalized PDL (Lemma 27) replaces the additive performance-difference identity by a multiplicative one whose weights and per-step gaps couple to the unknown reward $r^\star$.

We adopt the notation of Section 5 and Appendix F: the affine aggregation $\sigma_h(u,v) = a_h(u)v + b_h(u)$; the offline dataset $\mathcal{D} = \{(\tau_i, Y_i)\}_{i=1}^n$ with $\mathbb{E}[Y_i \mid \tau_i] = R(\tau_i; r^\star)$ and $Y_i \in [0, V_{\max}]$, so the centered noise $\xi_i := Y_i - R(\tau_i; r^\star)$ satisfies $\mathbb{E}[\xi_i \mid \tau_i] = 0$ and $|\xi_i| \le V_{\max}$ a.s.; the policy class Π, reward class $\mathcal{R}$, and value class $\mathcal{F}$; the generalized Bellman operator $\mathcal{T}_r^\pi$ from Eq. (10), with $\mathcal{T}^\pi := \mathcal{T}_{r^\star}^\pi$ (by Proposition 24, Q^π is the unique fixed point of $\mathcal{T}^\pi$); the trajectory return $R(\tau; r)$ from Eq. (9); the empirical losses $\mathcal{L}_{\mathcal{D}}^{W^r}, \mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}$ in Eq. (11) and $\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}$; the trajectory weights W_h^r; the state–action concentrability $C_{sa}(\pi^\star)$ from Eq. (2); the reward-process coefficient $\kappa := \kappa_\mu(\sigma)$ from Definition 1; and the Bellman inverse coefficient $\chi := \chi_\mu(\sigma)$ from Definition 2. Throughout $r^\star = (r_h^\star)_{h=1}^H$ is the true per-step reward and $r = (r_h)_{h=1}^H \in \mathcal{R}$ a candidate.

G.1 Setup

Population losses For $(\pi, r, f) \in \Pi \times \mathcal{R} \times \mathcal{F}$, the population analogues of Eq. (11) and of $\mathcal{L}_{\mathcal{D}}^{\mathrm{RM}}$ are

$$\begin{aligned}
\mathcal{L}_\mu(\pi, f) &:= \sum_{h=1}^{H} \mathbb{E}_\mu\big[f_h(s_h, \pi) - f_h(s_h, a_h)\big], \\
\mathcal{L}_\mu^{W^r}(\pi, r, f) &:= \sum_{h=1}^{H} \mathbb{E}_\mu\big[W_h^r(\tau)\big(f_h(s_h, \pi) - f_h(s_h, a_h)\big)\big], \\
\mathcal{L}_\mu^{\mathrm{BE}}(\pi, r, f) &:= \sum_{h=1}^{H} \mathbb{E}_\mu\big[\big((f_h - \mathcal{T}_r^\pi f)(s_h, a_h)\big)^2\big], \\
\mathcal{L}_\mu^{\mathrm{RM}}(r) &:= \mathbb{E}_{\tau\sim\mu}\big[\big(R(\tau; r) - R(\tau; r^\star)\big)^2\big].
\end{aligned} \tag{79}$$

These extend Eq. (13) from Appendix C in two ways: the policy loss carries the trajectory weight W_h^r, and the Bellman-error and reward-model losses use the generalized Bellman operator $\mathcal{T}_r^\pi$ and trajectory return $R(\tau; \cdot)$ respectively. Because σ_h is affine in its continuation, the one-step target satisfies $\mathbb{E}[y_h^{r,f,\pi} \mid s_h, a_h] = (\mathcal{T}_r^\pi f)_h(s_h, a_h)$. At the population level, subtracting the squared loss of the conditional-mean target cancels the conditional variance. The empirical minimum in Eq. (11b) need not be unbiased; its deviation and the generalized completeness error are controlled by the concentration argument in Step 3(c) below.

Per-step gaps and the r-free identity Fix $\widehat{\pi} \in \Pi$, $f \in \mathcal{F}$, $r \in \mathcal{R}$, and define

$$\begin{aligned}
\delta_h(s, a) &:= f_h(s, a) - (\mathcal{T}_r^{\widehat{\pi}} f)_h(s, a), \\
\Delta_h^r(s, a) &:= (\mathcal{T}_{r^\star}^{\widehat{\pi}} f)_h(s, a) - (\mathcal{T}_r^{\widehat{\pi}} f)_h(s, a) \\
&= [a_h(r_h^\star) - a_h(r_h)]\, \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\widehat{\pi}, f}(s')] + [b_h(r_h^\star) - b_h(r_h)].
\end{aligned}$$

Both are per-step functions of (s, a). Their difference satisfies the *r-free identity*

$$(\mathcal{T}_{r^\star}^{\widehat{\pi}} f)_h(s, a) - f_h(s, a) = \Delta_h^r(s, a) - \delta_h(s, a), \tag{80}$$

whose right-hand side does not depend on r even though each summand does. Identity Eq. (80) is the key replacement, in the generalized setting, of the additive identity $\mathcal{T}^\pi f - f = r^\star - \tilde{r}$ used in Lemma 17.

The next lemma propagates this one-step identity through the generalized Bellman recursion.

Lemma 28 (Generalized simulation lemma)**.** *Let* $e_h(s) := V_h^{\widehat{\pi}}(s) - V_h^{\widehat{\pi}, f}(s) = \mathbb{E}_{a\sim\widehat{\pi}_h(\cdot|s)}[Q_h^{\widehat{\pi}}(s, a) - f_h(s, a)]$ *with the convention* $e_{H+1} \equiv 0$*. Then for every* $h \in [H]$*,* $(s, a) \in \mathcal{S} \times \mathcal{A}$*,*

$$Q_h^{\widehat{\pi}}(s, a) - f_h(s, a) = \Delta_h^r(s, a) - \delta_h(s, a) + a_h(r_h^\star(s, a))\, \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[e_{h+1}(s')]. \tag{81}$$

Proof. Since $Q^{\widehat{\pi}}$ is the fixed point of $\mathcal{T}_{r^\star}^{\widehat{\pi}}$, expanding via $\sigma_h(u, v) = a_h(u)v + b_h(u)$ gives

$$Q_h^{\widehat{\pi}}(s, a) - (\mathcal{T}_{r^\star}^{\widehat{\pi}} f)_h(s, a) = a_h(r_h^\star(s, a))\, \mathbb{E}_{s'}[V_{h+1}^{\widehat{\pi}}(s') - V_{h+1}^{\widehat{\pi}, f}(s')] = a_h(r_h^\star(s, a))\, \mathbb{E}_{s'}[e_{h+1}(s')],$$

since the $b_h(r_h^\star)$ terms cancel. Adding $\big[(\mathcal{T}_{r^\star}^{\widehat{\pi}} f)_h - f_h\big] = \Delta_h^r - \delta_h$ from Eq. (80) yields Eq. (81). □

G.2 Generalized Regret Decomposition

Lemma 29 (Generalized regret decomposition)**.** *For any* $\pi, \widehat{\pi} \in \Pi$*,* $r \in \mathcal{R}$*, and* $f \in \mathcal{F}$*, with* δ_h*,* Δ_h^r *from Eq. (80) and* $W_h^{r^\star}$ *from Proposition 6,*

$$J(\pi) - J(\widehat{\pi}) = \underbrace{\sum_{h=1}^{H} \mathbb{E}_\pi\big[W_h^{r^\star}(\tau)\big(\Delta_h^r - \delta_h\big)(s_h, a_h)\big]}_{(\mathrm{I}^\sharp)}$$

$$+\underbrace{\sum_{h=1}^{H}\mathbb{E}_\pi\big[W_h^{r^\star}(\tau)\big(f_h(s_h,a_h)-f_h(s_h,\widehat{\pi})\big)\big]}_{(\mathrm{II}^\sharp)}+\underbrace{f_1(s_1,\widehat{\pi})-J(\widehat{\pi})}_{(\mathrm{III}^\sharp)}. \tag{82}$$

The right-hand side does not depend on the choice of $r\in\mathcal{R}$, by Eq. (80).

Proof. By the generalized PDL (Lemma 27) at reward $r^\star$, $J(\pi)-J(\widehat{\pi})=\sum_h\mathbb{E}_\pi\big[W_h^{r^\star}(\tau)\,A_h^{\widehat{\pi}}(s_h,a_h)\big]$ with $A_h^{\widehat{\pi}}(s,a)=Q_h^{\widehat{\pi}}(s,a)-V_h^{\widehat{\pi}}(s)$. Using $V_h^{\widehat{\pi},f}(s)=f_h(s,\widehat{\pi})$, decompose

$$A_h^{\widehat{\pi}}(s,a)=\big[Q_h^{\widehat{\pi}}(s,a)-f_h(s,a)\big]+\big[f_h(s,a)-f_h(s,\widehat{\pi})\big]-e_h(s).$$

The simulation lemma Eq. (81) gives $Q_h^{\widehat{\pi}}-f_h=(\Delta_h^r-\delta_h)+a_h(r_h^\star)\,\mathbb{E}_{s'}[e_{h+1}]$, and the identity $W_h^{r^\star}(\tau)\cdot a_h(r_h^\star(s_h,a_h))=W_{h+1}^{r^\star}(\tau)$ together with the tower rule yields

$$\sum_{h=1}^{H}\mathbb{E}_\pi\big[W_h^{r^\star}\cdot a_h(r_h^\star)\,\mathbb{E}_{s'}[e_{h+1}(s')]\big]=\sum_{h=2}^{H}\mathbb{E}_\pi[W_h^{r^\star}\,e_h(s_h)],$$

using $e_{H+1}\equiv 0$. Combining the displays and using $W_1^{r^\star}=1$, $e_1(s_1)=J(\widehat{\pi})-f_1(s_1,\widehat{\pi})$, the only surviving e_h term is $-\mathbb{E}_\pi[W_1^{r^\star}e_1(s_1)]=f_1(s_1,\widehat{\pi})-J(\widehat{\pi})$, which is $(\mathrm{III}^\sharp)$. The remaining two terms are $(\mathrm{I}^\sharp)$ and $(\mathrm{II}^\sharp)$. □

G.3 Auxiliary Lemmas

We collect the structural lemmas used to bound the three terms of Eq. (82). Throughout, δ_h and Δ_h^r are as in Eq. (80).

Fake reward Given a critic $f\in\mathcal{F}$ and a policy $\widehat{\pi}\in\Pi$, select the *fake reward* $\tilde{r}=\tilde{r}^{\widehat{\pi},f}\in\mathcal{R}$ supplied by Assumption 4(4), so that f is Bellman-consistent with $\widehat{\pi}$:

$$\sigma_h\big(\tilde{r}_h(s,a),\,\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^{\widehat{\pi},f}(s')]\big)=f_h(s,a),\qquad h\in[H]. \tag{83}$$

Since $\tilde{r}\in\mathcal{R}$, the same continuation-return bounds and coefficient definitions apply. Thus f is exactly the value function of $\widehat{\pi}$ in the auxiliary MDP with reward $\tilde{r}$: $f=Q^{\widehat{\pi},\tilde{r}}$ and $J_{\tilde{r}}(\widehat{\pi})=f_1(s_1,\widehat{\pi})$. The fake reward is not estimated by the algorithm; it is a proof device that converts value-function inconsistency into a reward-level discrepancy. We write

$$\tilde{W}_h(\tau)\coloneqq\prod_{j=1}^{h-1}a_j(\tilde{r}_j(s_j,a_j)),\qquad G(\tau)\coloneqq R(\tau;r^\star)-R(\tau;\tilde{r}).$$

The proof below uses the Bellman inverse coefficient only through the consequence

$$\sum_{h=1}^{H}\mathbb{E}_\mu\big[(\tilde{r}_h-r_h)^2\big]\le\chi\,\mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi},r,f),\qquad\forall(\widehat{\pi},r,f)\in\Pi\times\mathcal{R}\times\mathcal{F}. \tag{84}$$

This follows from Definition 2 because $(r,\tilde{r})\in\mathcal{R}\times\mathcal{R}$, $(\mathcal{T}_{\tilde{r}}^{\widehat{\pi}}f)_h=f_h$, and Theorem 9 assumes $\chi<\infty$.

Trajectory return telescoping The next lemma gives an exact expression for the difference between two generalized trajectory returns, by interpolating one reward at a time.

Lemma 30 (Trajectory return telescoping)**.** *For any rewards r,r' and trajectory $\tau=(s_1,a_1,\dots,s_H,a_H)$,*

$$\begin{aligned}&R(\tau;r')-R(\tau;r)\\&\quad=\sum_{k=1}^{H}\Big(\prod_{j=1}^{k-1}a_j(r_j'(s_j,a_j))\Big)\big\{[a_k(r_k')-a_k(r_k)]\,R_{k+1}(\tau_{k+1};r)+[b_k(r_k')-b_k(r_k)]\big\},\end{aligned} \tag{85}$$

where $R_{k+1}(\tau_{k+1};r)$ is the generalized return from step $k+1$ onward under reward r.

Proof. Define interpolated rewards $r_h^{(k)} = r'_h$ for $h \le k$ and $r_h^{(k)} = r_h$ for $h > k$, so that $r^{(0)} = r$ and $r^{(H)} = r'$. Then $R(\tau; r') - R(\tau; r) = \sum_k [R(\tau; r^{(k)}) - R(\tau; r^{(k-1)})]$. Since $r^{(k)}$ and $r^{(k-1)}$ agree on $h > k$, $R_{k+1}(\tau_{k+1}; r^{(k)}) = R_{k+1}(\tau_{k+1}; r^{(k-1)}) = R_{k+1}(\tau_{k+1}; r)$, so at step k,

$$R_k(\tau_k; r^{(k)}) - R_k(\tau_k; r^{(k-1)}) = [a_k(r'_k) - a_k(r_k)]\, R_{k+1}(\tau_{k+1}; r) + [b_k(r'_k) - b_k(r_k)].$$

For $h < k$ both use r'_h, so the difference propagates by the multiplicative factor $a_h(r'_h)$. Unrolling from $h = 1$ to $h = k - 1$ gives Eq. (85) for the k-th summand; sum over k. □

Lemma 31 (Telescoping upper bound, contractive case)**.** *Suppose $a_h(u) \in [0, 1]$, a_h, b_h are L-Lipschitz, and $|R_h(\tau_h; r)| \le V_{\max}$ for all h, r, τ. Then, for any reward functions r, r' and any trajectory τ,*

$$|R(\tau; r') - R(\tau; r)| \le L(V_{\max} + 1) \sum_{h=1}^{H} |r'_h - r_h|(s_h, a_h).$$

Proof. Apply Lemma 30. Since $a_h(u) \in [0, 1]$, $\prod_{j<k} a_j(r'_j) \in [0, 1]$; since a_h, b_h are L-Lipschitz, $|a_k(r'_k) - a_k(r_k)| \le L|r'_k - r_k|$ and similarly for b_k; and by boundedness, $|R_{k+1}(\tau_{k+1}; r)| \le V_{\max}$. □

Lemma 32 (Weight Lipschitz bound)**.** *Suppose $a_h(u) \in [0, 1]$ and a_h is L-Lipschitz. Then, for any reward functions r, r' and any trajectory τ,*

$$|W_h^{r'}(\tau) - W_h^{r}(\tau)| \le L \sum_{j=1}^{h-1} |r'_j - r_j|(s_j, a_j).$$

Proof. Write $p_j := a_j(r_j(s_j, a_j))$ and $p'_j := a_j(r'_j(s_j, a_j))$. The standard telescoping identity gives

$$\prod_{j=1}^{h-1} p'_j - \prod_{j=1}^{h-1} p_j = \sum_{k=1}^{h-1} \Big(\prod_{j<k} p'_j \Big)(p'_k - p_k) \Big(\prod_{k<j\le h-1} p_j \Big).$$

Since $a_h(u) \in [0, 1]$, $p_j, p'_j \in [0, 1]$, so both surrounding products are bounded by 1. Since a_h is L-Lipschitz, $|p'_k - p_k| \le L|r'_k - r_k|(s_k, a_k)$. Summing over k yields the claim. □

Lemma 33 (Trajectory L^2 bound on G)**.** *Suppose $a_h(u) \in [0, 1]$, a_h, b_h are L-Lipschitz, and $|R_h(\tau_h; r)| \le V_{\max}$ for all h, r, τ. Then, for any $r \in \mathcal{R}$, $f \in \mathcal{F}$, $\widehat{\pi} \in \Pi$,*

$$\mathbb{E}_\mu[G(\tau)^2] \le 2\mathcal{L}_\mu^{\mathrm{RM}}(r) + 2L^2(V_{\max} + 1)^2 \chi H \mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi}, r, f).$$

Proof. Write $G = (R(\tau; r^\star) - R(\tau; r)) + (R(\tau; r) - R(\tau; \tilde{r}))$ and use $(x + y)^2 \le 2x^2 + 2y^2$. The first part is $\mathcal{L}_\mu^{\mathrm{RM}}(r)$. For the second, Lemma 31 with $r' = \tilde{r}$ gives $|R(\tau; r) - R(\tau; \tilde{r})| \le L(V_{\max} + 1) \sum_h |r_h - \tilde{r}_h|$; squaring and applying Cauchy–Schwarz,

$$(R(\tau; r) - R(\tau; \tilde{r}))^2 \le L^2(V_{\max} + 1)^2 H \sum_h (r_h - \tilde{r}_h)^2.$$

Taking $\mathbb{E}_\mu$ and applying Eq. (84) yields $\mathbb{E}_\mu[(R(\tau; r) - R(\tau; \tilde{r}))^2] \le L^2(V_{\max} + 1)^2 \chi H \mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi}, r, f)$. □

G.4 Proof of Theorem 9

We use the temperature parameter $\eta = V_{\max}\sqrt{K/(8 \log |\mathcal{A}|)}$, as in Lemma 11.

Proof of Theorem 9. Apply Lemma 29 with comparator $\pi = \pi^\star$, current policy $\widehat{\pi} = \pi_k$, and the pair $(r, f) = (r_k, f_k)$ chosen by Generalized OPAC. For each $k \in [K]$,

$$J(\pi^\star) - J(\pi_k) = \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[W_h^{r^\star}(\tau)(\Delta_h^{r_k} - \delta_{k,h})(s_h, a_h)\right]}_{(\mathrm{I}^\sharp)_k}$$

$$+\underbrace{\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[W_h^{r^\star}(\tau)\big(f_{k,h}(s_h,a_h)-f_{k,h}(s_h,\pi_k)\big)\right]}_{(\mathrm{II}^\sharp)_k}$$
$$+\underbrace{f_{k,1}(s_1,\pi_k)-J(\pi_k)}_{(\mathrm{III}^\sharp)_k}, \tag{86}$$

where

$$\delta_{k,h}\coloneqq f_{k,h}-(\mathcal{T}^{\pi_k}_{r_k}f_k)_h,\qquad \Delta_h^{r_k}\coloneqq(\mathcal{T}^{\pi_k}_{r^\star}f_k)_h-(\mathcal{T}^{\pi_k}_{r_k}f_k)_h.$$

We average the three terms in Eq. (86) over $k\in[K]$ on a single high-probability event on which the concentration bounds below hold uniformly over $\Pi\times\mathcal{R}\times\mathcal{F}$.

Step 1: bound $(\mathrm{I}^\sharp)$**.** Since $a_h(u)\in[0,1]$, $W_h^{r^\star}\in[0,1]$. Thus

$$|(\mathrm{I}^\sharp)_k|\le\sum_{h=1}^{H}\mathbb{E}_{d_h^{\pi^\star}}\left[|\delta_{k,h}|+|\Delta_h^{r_k}|\right].$$

The pointwise change of measure $d_h^{\pi^\star}(s,a)\le C_{sa}(\pi^\star)d_h^{\mu}(s,a)$ and Cauchy–Schwarz give

$$\sum_h\mathbb{E}_{d_h^{\pi^\star}}\left[|\delta_{k,h}|\right]\le\sqrt{HC_{sa}(\pi^\star)\mathcal{L}_\mu^{\mathrm{BE}}(\pi_k,r_k,f_k)}.$$

Since a_h,b_h are L-Lipschitz and $|V_{h+1}^{\pi_k,f_k}|\le V_{\max}$,

$$|\Delta_h^{r_k}(s,a)|\le L(V_{\max}+1)|r_h^\star-r_{k,h}|(s,a).$$

Applying the same change of measure and using the definition of $\kappa=\kappa_\mu(\sigma)$ from Definition 1,

$$\sum_h\mathbb{E}_{d_h^{\pi^\star}}\left[|\Delta_h^{r_k}|\right]\le L(V_{\max}+1)\sqrt{\kappa HC_{sa}(\pi^\star)\mathcal{L}_\mu^{\mathrm{RM}}(r_k)}. \tag{87}$$

Combining,

$$|(\mathrm{I}^\sharp)_k|\le\phi^{\mathrm{I}}_{\mathrm{BE}}\sqrt{\mathcal{L}_\mu^{\mathrm{BE}}(\pi_k,r_k,f_k)}+\phi^{\mathrm{I}}_{\mathrm{RM}}\sqrt{\mathcal{L}_\mu^{\mathrm{RM}}(r_k)}, \tag{88}$$

with $\phi^{\mathrm{I}}_{\mathrm{BE}}\coloneqq\sqrt{HC_{sa}(\pi^\star)}$ and $\phi^{\mathrm{I}}_{\mathrm{RM}}\coloneqq L(V_{\max}+1)\sqrt{\kappa HC_{sa}(\pi^\star)}$.

Remark 2. *Eq. (87) is the only place where $\kappa_\mu(\sigma)$ enters the final bound. In the cumulative-return case, the separate additive analysis instead applies change of trajectory measure (Lemma 10) directly to the expected signed reward sum. For this generalized analysis, one may alternatively use*

$$\sum_h\mathbb{E}_{d_h^{\pi^\star}}\left[|\Delta_h^{r_k}|\right]\le L(V_{\max}+1)\sqrt{H\kappa_{\pi^\star}C_\tau(\pi^\star)\mathcal{L}_\mu^{\mathrm{RM}}(r_k)},$$
$$\kappa_{\pi^\star}\coloneqq\sup_{r\neq r'}\frac{\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[(r_h-r'_h)^2(s_h,a_h)\right]}{\mathbb{E}_{\pi^\star}\left[(R(\tau;r)-R(\tau;r'))^2\right]},$$

where the supremum ranges over distinct $r,r'\in\mathcal{R}$ and $C_\tau(\pi^\star)\coloneqq\max_\tau\mathbb{P}_{\pi^\star}[\tau]/\mathbb{P}_\mu[\tau]$. This inserts $\kappa_{\pi^\star}C_\tau(\pi^\star)$ in place of $\kappa_\mu C_{sa}(\pi^\star)$. This trajectory-level form can be tighter when μ places sufficient mass on trajectories visited under $\pi^\star$. For consistency with the rest of the paper, which is stated under state–action concentrability $C_{sa}(\pi^\star)$ rather than trajectory concentrability $C_\tau(\pi^\star)$, we retain the present analysis.

Step 2: bound $(\mathrm{II}^\sharp)$**.** Let $\mathfrak{H}_{h-1}$ denote the trajectory prefix before action a_h is drawn. Conditional on $\mathfrak{H}_{h-1}$, the state s_h and the weight

$$W_h^{r^\star}(\tau)=\prod_{j<h}a_j(r_j^\star(s_j,a_j))$$

are fixed, and $W_h^{r^\star}(\tau)\in[0,1]$. Therefore

$$\mathbb{E}_{\pi^\star}\Big[W_h^{r^\star}(\tau)\big(f_{k,h}(s_h,a_h)-f_{k,h}(s_h,\pi_k)\big)\Big|\mathfrak{H}_{h-1}\Big]=W_h^{r^\star}(\tau)\big(f_{k,h}(s_h,\pi^\star)-f_{k,h}(s_h,\pi_k)\big).$$

For every fixed h and state s, the pointwise bound Eq. (17) in the proof of Lemma 11, with $p = \pi_h^\star(\cdot \mid s)$ and $q_k = \pi_{k,h}(\cdot \mid s)$, gives

$$\frac{1}{K}\sum_{k=1}^{K}\big(f_{k,h}(s,\pi^\star) - f_{k,h}(s,\pi_k)\big) \le V_{\max}\sqrt{\frac{\log|\mathcal{A}|}{2K}},$$

since $f_{k,h}(s,\cdot) \in [0, V_{\max}]$. Multiplying this pointwise inequality by the nonnegative prefix weight $W_h^{r^\star}(\tau) \le 1$, taking expectation over the prefix, and summing over h yields

$$\frac{1}{K}\sum_{k=1}^{K}(\mathrm{II}^\sharp)_k \le HV_{\max}\sqrt{\frac{\log|\mathcal{A}|}{2K}} = \widetilde{O}\big(HV_{\max}/\sqrt{K}\big). \tag{89}$$

Step 3: bound $(\mathrm{III}^\sharp)$**.** Fix k and write $\widehat{\pi} = \pi_k$, $r = r_k$, and $f = f_k$.

(3a) Express $(\mathrm{III}^\sharp)$ *via population losses.* Apply the generalized PDL (Lemma 27) with the first policy μ and the comparator policy $\widehat{\pi}$. Under reward $\tilde{r}$, using $f = Q^{\widehat{\pi},\tilde{r}}$ and hence $J_{\tilde{r}}(\widehat{\pi}) = f_1(s_1,\widehat{\pi})$,

$$f_1(s_1,\widehat{\pi}) - J_{\tilde{r}}(\mu) = \sum_h \mathbb{E}_\mu\big[\tilde{W}_h(\tau)(f_h(s_h,\widehat{\pi}) - f_h(s_h,a_h))\big] = \mathcal{L}_\mu^{W^{\tilde{r}}}(\widehat{\pi},\tilde{r},f),$$

and under reward $r^\star$,

$$J(\widehat{\pi}) - J(\mu) = \sum_h \mathbb{E}_\mu\big[W_h^{r^\star}(\tau)(V_h^{\widehat{\pi}}(s_h) - Q_h^{\widehat{\pi}}(s_h,a_h))\big] = \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi},r^\star,Q^{\widehat{\pi}}),$$

where both last equalities use the definition of the weighted population loss in Eq. (79). Subtracting and using $J_{\tilde{r}}(\mu) - J(\mu) = -\mathbb{E}_\mu[G(\tau)]$ with $G(\tau) := R(\tau;r^\star) - R(\tau;\tilde{r})$,

$$(\mathrm{III}^\sharp) = \mathcal{L}_\mu^{W^{\tilde{r}}}(\widehat{\pi},\tilde{r},f) - \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi},r^\star,Q^{\widehat{\pi}}) - \mathbb{E}_\mu[G]. \tag{90}$$

(3b) Pivot via the in-class realizer. Let $f_{\widehat{\pi}} \in \mathcal{F}$ satisfy

$$\sup_{h,\nu}\|f_{\widehat{\pi},h} - (\mathcal{T}_{r^\star}^{\widehat{\pi}} f_{\widehat{\pi}})_h\|_{2,\nu}^2 \le \varepsilon_{\mathcal{F}},$$

as guaranteed by Assumption 2 for the generalized operator. Insert $\pm\mathcal{L}_\mu^{W^r}(\widehat{\pi},r,f)$ and $\pm\mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi},r^\star,f_{\widehat{\pi}})$:

$$\begin{aligned}(\mathrm{III}^\sharp) = {} & \underbrace{\mathcal{L}_\mu^{W^{\tilde{r}}}(\widehat{\pi},\tilde{r},f) - \mathcal{L}_\mu^{W^r}(\widehat{\pi},r,f)}_{\text{(weight replacement)}} + \underbrace{\mathcal{L}_\mu^{W^r}(\widehat{\pi},r,f) - \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi},r^\star,f_{\widehat{\pi}})}_{\text{(algorithm gap)}} \\ & + \underbrace{\mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi},r^\star,f_{\widehat{\pi}}) - \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi},r^\star,Q^{\widehat{\pi}})}_{\text{(realizability gap)}} - \mathbb{E}_\mu[G].\end{aligned} \tag{91}$$

(3c) Concentration estimates. On the high-probability event, the following generalized concentration bounds hold uniformly over $(\pi,r,f) \in \Pi \times \mathcal{R} \times \mathcal{F}$. They use the same Bernstein/Hoeffding arguments as in Appendix B.2, but with three generalized ingredients: the Bellman target is $y_h^{r,f,\pi} = \sigma_h(r_h, f_{h+1}(\cdot,\pi))$, the reward-model residual is the trajectory return $R(\tau;r) - R(\tau;r^\star)$, and the policy-loss summand carries the trajectory weight W_h^r.

- *BE concentration*. Conditional on (s_h,a_h),

  $$\mathbb{E}[y_h^{r,f,\pi} \mid s_h,a_h] = (\mathcal{T}_r^\pi f)_h(s_h,a_h),$$

  so the same bias–variance cancellation behind Lemma 14 applies with $\mathcal{T}_r^\pi$ in place of the additive Bellman operator. Since $y_h^{r,f,\pi} - R_h(\tau_h;r) = a_h(r_h)(f_{h+1}(s_{h+1},\pi) - R_{h+1}(\tau_{h+1};r))$, the continuation bounds imply $|y_h^{r,f,\pi}| \le 2V_{\max}$. Thus per-step squared residuals are $O(V_{\max}^2)$, giving

  $$\mathcal{L}_\mu^{\mathrm{BE}}(\pi,r,f) \le 2\mathcal{L}_{\mathcal{D}}^{\mathrm{BE}}(\pi,r,f) + \varepsilon_{\mathrm{BE}} + 4\varepsilon_{\mathcal{F},\mathcal{F}}, \qquad \varepsilon_{\mathrm{BE}} = \widetilde{O}(HV_{\max}^2/n).$$

  The $4\varepsilon_{\mathcal{F},\mathcal{F}}$ term is the slack from Assumption 3 for the generalized operator, uniformly over the single class $\mathcal{R}$ containing both candidate and fake rewards.

- *RM concentration, $r^\star$-centered.* Apply the Bernstein self-bounding proof of Lemma 13 to
$$U_i(r) := (R(\tau_i; r) - Y_i)^2 - (R(\tau_i; r^\star) - Y_i)^2.$$
Since $\mathbb{E}[Y_i \mid \tau_i] = R(\tau_i; r^\star)$, $\mathbb{E}[U_i(r)] = \mathcal{L}_\mu^{\mathrm{RM}}(r)$; since $|R(\tau; r)| \le V_{\max}$ and $|Y_i - R(\tau_i; r^\star)| \le V_{\max}$, the range and variance terms are of order $V_{\max}^2$:
$$\mathcal{L}_\mu^{\mathrm{RM}}(r) \le 2[\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r) - \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star)] + \varepsilon_{\mathrm{RM}}, \qquad \varepsilon_{\mathrm{RM}} = \widetilde{O}(V_{\max}^2/n).$$

- *Weighted policy-loss concentration.* This is Hoeffding's inequality as in Lemma 12, but union-bounded over (π, r, f) rather than only (π, f). The extra reward argument only changes the summand through the measurable weight $W_h^r(\tau)$; contractivity gives $W_h^r \in [0, 1]$, so the per-trajectory summand remains bounded by $O(HV_{\max})$:
$$\sup_{(\pi,r,f)} \left|\mathcal{L}_\mu^{W^r}(\pi, r, f) - \mathcal{L}_\mathcal{D}^{W^r}(\pi, r, f)\right| \le \sqrt{\varepsilon_{\mathrm{Perf}}}, \qquad \sqrt{\varepsilon_{\mathrm{Perf}}} = \widetilde{O}(HV_{\max}/\sqrt{n}).$$

(3d) Algorithm gap. Since $r^\star \in \mathcal{R}$, the pessimism step (Algorithm 1, line 3, with the losses Eq. (11) and $\mathcal{L}_\mathcal{D}^{\mathrm{RM}}$) can be compared with $(f_{\widehat{\pi}}, r^\star)$. After subtracting $\beta\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star)$ from both sides,
$$\mathcal{L}_\mathcal{D}^{W^r}(\widehat{\pi}, r, f) + \beta\mathcal{L}_\mathcal{D}^{\mathrm{BE}}(\widehat{\pi}, r, f) + \beta\big[\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r) - \mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star)\big] \le \mathcal{L}_\mathcal{D}^{W^{r^\star}}(\widehat{\pi}, r^\star, f_{\widehat{\pi}}) + \beta\varepsilon_{\mathrm{apx}}, \tag{92}$$
The per-stage realizability bound gives $\mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi}, r^\star, f_{\widehat{\pi}}) \le H\varepsilon_\mathcal{F}$; the squared-loss concentration argument yields
$$\mathcal{L}_\mathcal{D}^{\mathrm{BE}}(\widehat{\pi}, r^\star, f_{\widehat{\pi}}) \le \varepsilon_{\mathrm{apx}}, \qquad \varepsilon_{\mathrm{apx}} = \widetilde{O}(HV_{\max}^2/n + H\varepsilon_\mathcal{F}).$$
Combining Eq. (92) with the three concentration bounds, the centered noisy term $\mathcal{L}_\mathcal{D}^{\mathrm{RM}}(r^\star)$ cancels exactly, and
$$\mathcal{L}_\mu^{W^r}(\widehat{\pi}, r, f) - \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi}, r^\star, f_{\widehat{\pi}}) \le -\frac{\beta}{2}(\mathcal{L}_\mu^{\mathrm{BE}} + \mathcal{L}_\mu^{\mathrm{RM}}) + \frac{\beta}{2}(\varepsilon_{\mathrm{BE}} + \varepsilon_{\mathrm{RM}} + 4\varepsilon_{\mathcal{F},\mathcal{F}}) + \beta\varepsilon_{\mathrm{apx}} + 2\sqrt{\varepsilon_{\mathrm{Perf}}}. \tag{93}$$

(3e) Realizability gap (third bracket). Write $W_h = W_h^{r^\star}$, $\phi_h = f_{\widehat{\pi},h} - Q_h^{\widehat{\pi}}$, and $\delta_h = f_{\widehat{\pi},h} - (\mathcal{T}_{r^\star}^{\widehat{\pi}} f_{\widehat{\pi}})_h$. Bellman evaluation and affine dependence on the continuation give
$$\delta_h(s, a) = \phi_h(s, a) - a_h(r_h^\star(s, a))\mathbb{E}_{s' \sim P_h(\cdot|s,a)}[\phi_{h+1}(s', \widehat{\pi})], \qquad \phi_{H+1} \equiv 0.$$
Since $W_{h+1} = W_h a_h(r_h^\star)$ and $W_1 = 1$, weighted telescoping under any rollout policy ρ gives
$$\sum_{h=1}^H \mathbb{E}_\rho[W_h\delta_h] = \phi_1(s_1, \widehat{\pi}) - \sum_{h=1}^H \mathbb{E}_\rho[W_h(\phi_h(s_h, \widehat{\pi}) - \phi_h(s_h, a_h))].$$
The last sum vanishes for $\rho = \widehat{\pi}$. Subtracting the identity for $\rho = \mu$ from that for $\rho = \widehat{\pi}$ therefore yields
$$\mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi}, r^\star, f_{\widehat{\pi}}) - \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi}, r^\star, Q^{\widehat{\pi}}) = \sum_{h=1}^H \mathbb{E}_{\widehat{\pi}}[W_h\delta_h] - \sum_{h=1}^H \mathbb{E}_\mu[W_h\delta_h].$$
Using $0 \le W_h \le 1$, Cauchy–Schwarz, and Assumption 2 for $\mu, \widehat{\pi} \in \Pi$ gives
$$\left|\mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi}, r^\star, f_{\widehat{\pi}}) - \mathcal{L}_\mu^{W^{r^\star}}(\widehat{\pi}, r^\star, Q^{\widehat{\pi}})\right| \le 2H\sqrt{\varepsilon_\mathcal{F}}. \tag{94}$$

(3f) Weight-replacement bracket. Since $|f_h| \le V_{\max}$,
$$\left|\mathcal{L}_\mu^{W^{\tilde{r}}}(\widehat{\pi}, \tilde{r}, f) - \mathcal{L}_\mu^{W^r}(\widehat{\pi}, r, f)\right| \le 2V_{\max} \sum_{h=1}^H \mathbb{E}_\mu[|\tilde{W}_h - W_h^r|].$$
By the weight Lipschitz bound (Lemma 32) and the trivial estimate $\sum_{h=1}^H \sum_{j<h} \le H \sum_{j=1}^H$, then Cauchy–Schwarz over j followed by Eq. (84),
$$\sum_h \mathbb{E}_\mu[|\tilde{W}_h - W_h^r|] \le LH \sum_j \mathbb{E}_\mu[|\tilde{r}_j - r_j|] \le L\sqrt{\chi}\, H^{3/2} \sqrt{\mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi}, r, f)},$$

hence

$$\left|\mathcal{L}_\mu^{W^{\tilde{r}}}(\widehat{\pi},\tilde{r},f)-\mathcal{L}_\mu^{W^{r}}(\widehat{\pi},r,f)\right|\le 2V_{\max}L\sqrt{\chi}\,H^{3/2}\sqrt{\mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi},r,f)}. \tag{95}$$

(3g) The $-\mathbb{E}_\mu[G]$ piece. By Cauchy–Schwarz, $|\mathbb{E}_\mu[G]|\le\sqrt{\mathbb{E}_\mu[G^2]}$. Substituting the bound on $\mathbb{E}_\mu[G^2]$ from Lemma 33 and using $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$,

$$|\mathbb{E}_\mu[G]|\le\sqrt{2}\sqrt{\mathcal{L}_\mu^{\mathrm{RM}}(r)}+L(V_{\max}+1)\sqrt{2\chi H\,\mathcal{L}_\mu^{\mathrm{BE}}(\widehat{\pi},r,f)}. \tag{96}$$

(3h) Combine. Adding Eq. (93)–Eq. (96) (the sign of $-\mathbb{E}_\mu[G]$ is absorbed into $|\mathbb{E}_\mu[G]|$) yields

$$\begin{aligned}(\mathrm{III}^\sharp)\le&-\frac{\beta}{2}(\mathcal{L}_\mu^{\mathrm{BE}}+\mathcal{L}_\mu^{\mathrm{RM}})+\frac{\beta}{2}(\varepsilon_{\mathrm{BE}}+\varepsilon_{\mathrm{RM}}+4\varepsilon_{\mathcal{F},\mathcal{F}})\\&+\beta\varepsilon_{\mathrm{apx}}+2\sqrt{\varepsilon_{\mathrm{Perf}}}+2H\sqrt{\varepsilon_{\mathcal{F}}}+\phi_{\mathrm{BE}}^{\mathrm{III}}\sqrt{\mathcal{L}_\mu^{\mathrm{BE}}}+\phi_{\mathrm{RM}}^{\mathrm{III}}\sqrt{\mathcal{L}_\mu^{\mathrm{RM}}},\end{aligned} \tag{97}$$

with $\phi_{\mathrm{BE}}^{\mathrm{III}}:=2V_{\max}L\sqrt{\chi}\,H^{3/2}+L(V_{\max}+1)\sqrt{2\chi H}$ and $\phi_{\mathrm{RM}}^{\mathrm{III}}:=\sqrt{2}$. All BE/RM losses are evaluated at $(\widehat{\pi},r,f)$.

Step 4: combine and tune β. Adding Eq. (88) and Eq. (97), with all BE/RM losses evaluated at $(\widehat{\pi},r,f)$,

$$(\mathrm{I}^\sharp)+(\mathrm{III}^\sharp)\le-\frac{\beta}{2}(\mathcal{L}_\mu^{\mathrm{BE}}+\mathcal{L}_\mu^{\mathrm{RM}})+\frac{\beta}{2}\bar{\varepsilon}+2\sqrt{\varepsilon_{\mathrm{Perf}}}+2H\sqrt{\varepsilon_{\mathcal{F}}}+\Phi_{\mathrm{BE}}\sqrt{\mathcal{L}_\mu^{\mathrm{BE}}}+\Phi_{\mathrm{RM}}\sqrt{\mathcal{L}_\mu^{\mathrm{RM}}}, \tag{98}$$

where $\bar{\varepsilon}:=\varepsilon_{\mathrm{BE}}+\varepsilon_{\mathrm{RM}}+4\varepsilon_{\mathcal{F},\mathcal{F}}+2\varepsilon_{\mathrm{apx}}$ aggregates the slacks (mirroring ε in Eq. (59)) and

$$\Phi_{\mathrm{BE}}:=\phi_{\mathrm{BE}}^{\mathrm{I}}+\phi_{\mathrm{BE}}^{\mathrm{III}},\qquad\Phi_{\mathrm{RM}}:=\phi_{\mathrm{RM}}^{\mathrm{I}}+\phi_{\mathrm{RM}}^{\mathrm{III}},\qquad\Phi^2:=\Phi_{\mathrm{BE}}^2+\Phi_{\mathrm{RM}}^2.$$

With the coefficients from Eq. (88) and Eq. (97),

$$\Phi=\widetilde{O}\big(L(V_{\max}+1)\sqrt{\kappa HC_{sa}(\pi^\star)}+V_{\max}L\sqrt{\chi H^3}+L(V_{\max}+1)\sqrt{\chi H}+\sqrt{HC_{sa}(\pi^\star)}+1\big).$$

Under the bounds $V_{\max},L,\kappa,\chi\ge1$ in Theorem 9, with $H,C_{sa}(\pi^\star)\ge1$, the $L(V_{\max}+1)\sqrt{\chi H}$ term is dominated by $V_{\max}L\sqrt{\chi H^3}$ and the constant term is dominated by $\sqrt{HC_{sa}(\pi^\star)}$, giving the simplified order for Φ stated in Theorem 9. Young's inequality $X\sqrt{Y}\le X^2/(2\beta)+\beta Y/2$ cancels the negative BE/RM terms and leaves

$$(\mathrm{I}^\sharp)+(\mathrm{III}^\sharp)\le\frac{\Phi^2}{2\beta}+\frac{\beta\bar{\varepsilon}}{2}+2\sqrt{\varepsilon_{\mathrm{Perf}}}+2H\sqrt{\varepsilon_{\mathcal{F}}}.$$

Choosing

$$\beta=\frac{\Phi}{\sqrt{\bar{\varepsilon}}},\qquad\bar{\varepsilon}=\widetilde{O}\big(HV_{\max}^2/n+H\varepsilon_{\mathcal{F}}+\varepsilon_{\mathcal{F},\mathcal{F}}\big).$$

Hence, using $\sqrt{x+y+z}\le\sqrt{x}+\sqrt{y}+\sqrt{z}$,

$$\begin{aligned}\frac{\Phi^2}{2\beta}+\frac{\beta\bar{\varepsilon}}{2}&=\Phi\sqrt{\bar{\varepsilon}}\\&=\widetilde{O}\Big(\Phi V_{\max}\sqrt{\frac{H}{n}}+\sqrt{H}\,\Phi\sqrt{\varepsilon_{\mathcal{F}}}+\Phi\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}}\Big).\end{aligned} \tag{99}$$

Expanding the statistical part with the above bound on Φ gives

$$\Phi V_{\max}\sqrt{\frac{H}{n}}=\widetilde{O}\Big(V_{\max}^2L\sqrt{\frac{\kappa H^2C_{sa}(\pi^\star)}{n}}+V_{\max}^2L\sqrt{\frac{\chi H^4}{n}}+V_{\max}\sqrt{\frac{H^2C_{sa}(\pi^\star)}{n}}\Big). \tag{100}$$

The concentration remainder satisfies

$$2\sqrt{\varepsilon_{\mathrm{Perf}}}=\widetilde{O}\big(V_{\max}H/\sqrt{n}\big).$$

Under the same normalization, the last term in Eq. (100) and the concentration remainder are bounded by the displayed statistical terms in Eq. (12). Therefore,

$$(\mathrm{I}^\sharp)+(\mathrm{III}^\sharp)\le\widetilde{O}\left(V_{\max}^2L\sqrt{\frac{\kappa H^2C_{sa}(\pi^\star)}{n}}+V_{\max}^2L\sqrt{\frac{\chi H^4}{n}}+\Phi\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}}+(\sqrt{H}\,\Phi+H)\sqrt{\varepsilon_{\mathcal{F}}}\right).$$

Averaging over k and adding the no-regret contribution Eq. (89) gives Eq. (12). □

H Examples of Generalized Objectives

This section illustrates the generalized objective framework of Section 5. For a policy π and per-step reward tuple $r = (r_1, \dots, r_H)$, write the trajectory return as

$$R(\tau; r) = \sigma\big(r_1(s_1, a_1), \dots, r_H(s_H, a_H)\big), \qquad J(\pi) = \mathbb{E}_{\tau\sim\pi}[R(\tau; r^\star)].$$

Theorem 5 shows that without structure on σ, outcome-based learning can require exponentially many trajectories. We give sufficient conditions for the upper bound using two coefficients. First, because the trajectory outcome aggregates per-step rewards, we use the *reward-process coefficient* $\kappa_\mu(\sigma)$ (Definition 1) to measure information loss from observing only a scalar return:

$$\kappa_\mu(\sigma) = \sup_{r\neq r'\in\mathcal{R}} \frac{\sum_{h=1}^H \mathbb{E}_\mu\big[(r_h - r'_h)^2(s_h, a_h)\big]}{\mathbb{E}_{\tau\sim\mu}\big[(R(\tau; r) - R(\tau; r'))^2\big]}.$$

Large $\kappa_\mu(\sigma)$ means distinct reward profiles are nearly indistinguishable from scalar outcomes alone. Second, generalized objectives need not admit a standard Bellman dynamic-programming decomposition (Zhang et al., 2020; Barakat et al., 2023). We therefore restrict attention to affine-in-continuation aggregations $\sigma_h(u, v) = a_h(u)v + b_h(u)$ and use the *Bellman inverse coefficient* $\chi_\mu(\sigma)$ (Definition 2) to measure how much one-step generalized Bellman targets can compress per-step reward differences:

$$\chi_\mu(\sigma) = \sup_{\pi\in\Pi,\, f\in\mathcal{F},\, r\neq r'\in\mathcal{R}} \frac{\sum_{h=1}^H \mathbb{E}_\mu\big[(r_h - r'_h)^2(s_h, a_h)\big]}{\sum_{h=1}^H \mathbb{E}_\mu\big[\big((\mathcal{T}_r^\pi f)_h - (\mathcal{T}_{r'}^\pi f)_h\big)^2(s_h, a_h)\big]}.$$

Under all the hypotheses of Theorem 9, the statistical terms in the generalized OPAC bound are

$$\widetilde{O}\left(V_{\max}^2 L\sqrt{\frac{\kappa_\mu(\sigma)H^2 C_{sa}(\pi^\star)}{n}} + V_{\max}^2 L\sqrt{\frac{\chi_\mu(\sigma)H^4}{n}}\right), \tag{101}$$

up to the no-regret term in K and approximation errors.

The following examples describe aggregation structure and coefficient calculations. A Bellman recursion by itself does not verify the joint reward-class closure, generalized realizability/completeness, or finite coefficients required by Theorem 9.

Cumulative return Take $a_h(u) = 1$, $b_h(u) = u$, and $g = 0$, then $\sigma_h(u, v) = u + v$, $R(\tau) = \sum_{h=1}^H r_h(s_h, a_h)$. Here $J(\pi) = \mathbb{E}_\pi[\sum_h r_h]$ is the standard cumulative return. The generalized Bellman operator Eq. (10) reduces to the usual Bellman backup, and items (1)–(3) of Assumption 4 hold with $L = 1$. Item (4) additionally requires $(f_h - \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[f_{h+1}(s', \pi)])_{h=1}^H \in \mathcal{R}$ for every (π, f), which is not automatic for arbitrary bounded critics. Moreover, for every π, f and every $h, (s, a)$,

$$(\mathcal{T}_r^\pi f)_h(s, a) - (\mathcal{T}_{r'}^\pi f)_h(s, a) = r_h(s, a) - r'_h(s, a),$$

since the continuation term $\mathbb{E}_{s'}[f_{h+1}(s', \pi)]$ cancels in the difference. Thus $\chi_\mu(\sigma) = 1$ for every behavior μ and all classes $\Pi, \mathcal{F}$ whenever some pair has positive squared reward distance under μ (and is zero for a class trivial on the data support).

The reward-process coefficient is instead determined by how the reward class distributes differences across stages. For $r, r' \in \mathcal{R}$ write $\Delta_h := r_h - r'_h$. The numerator of Definition 1 is $\sum_h \mathbb{E}_\mu[\Delta_h^2(s_h, a_h)]$, whereas the denominator is

$$\mathbb{E}_\mu\Big[\Big(\sum_{h=1}^H \Delta_h(s_h, a_h)\Big)^2\Big] = \sum_{h=1}^H \mathbb{E}_\mu\big[\Delta_h^2(s_h, a_h)\big] + 2\sum_{h<h'} \mathbb{E}_\mu\big[\Delta_h(s_h, a_h)\Delta_{h'}(s_{h'}, a_{h'})\big].$$

The ratio therefore equals one when the cross-stage terms sum to zero, and it is large when they are negative and nearly cancel the diagonal terms. A pair that differs at a single stage has no cross terms, so $\kappa_\mu(\sigma) \geq 1$ for any class containing such a pair. Three regimes follow.

Infinite for every behavior policy. Let $H \geq 2$, and suppose that $\mathcal{R}$ contains r and r' with $r'_1 = r_1 + c$ and $r'_2 = r_2 - c$ for a constant $c \neq 0$, and $r'_h = r_h$ for $h \geq 3$. Then $\sum_h \Delta_h = 0$ on every trajectory,

while the numerator equals $2c^2$. Hence $\kappa_\mu(\sigma) = +\infty$ for every μ, uniform behavior included: the two rewards induce the same return on every trajectory, so no data distribution separates them. More generally, $\kappa_\mu(\sigma) = +\infty$ whenever some pair has $\sum_h \Delta_h(s_h, a_h) = 0$ μ-almost surely but $\sum_h \mathbb{E}_\mu[\Delta_h^2(s_h, a_h)] > 0$.

Exactly one for μ-centered classes. Suppose that all members of $\mathcal{R}$ have the same behavior-averaged reward at every state reached by μ: for every h, every such $s \in \mathcal{S}_h$, and all $r, r' \in \mathcal{R}$, $\sum_a \mu_h(a \mid s)\Delta_h(s, a) = 0$. For $h < h'$, $\Delta_h(s_h, a_h)$ is determined by the trajectory prefix up to $s_{h'}$, and, because μ is Markov, the conditional mean of $\Delta_{h'}(s_{h'}, a_{h'})$ given this prefix is $\sum_a \mu_{h'}(a \mid s_{h'})\Delta_{h'}(s_{h'}, a) = 0$. Every cross term therefore vanishes, the two squared distances coincide, and $\kappa_\mu(\sigma) = 1$ for any such class that is nontrivial on the data support, in any MDP. Such a class fixes the behavior average of the reward and varies only the contrasts between actions; the discounted-return calculation below with $c = 1$ is an instance.

Intermediate values: stage offsets versus action contrasts. Consider, as in Table 2, a deterministic chain on which behavior draws actions independently across stages, and rewards $r_h(a_h)$ that depend only on the stage and the action. Decompose $\Delta_h(a) = m_h + \xi_h(a)$ into the offset $m_h := \mathbb{E}_{a\sim\mu_h}[\Delta_h(a)]$ and the centered contrast ξ_h, and let $v_h := \mathbb{E}_{a\sim\mu_h}[\xi_h(a)^2]$. Independence across stages gives, whenever $\sum_h v_h > 0$,

$$\frac{\sum_h \mathbb{E}_\mu[\Delta_h^2(a_h)]}{\mathbb{E}_\mu[(\sum_h \Delta_h(a_h))^2]} = \frac{\sum_h m_h^2 + \sum_h v_h}{(\sum_h m_h)^2 + \sum_h v_h} \le 1 + \frac{\sum_h m_h^2}{\sum_h v_h},$$

with equality when $\sum_h m_h = 0$. Under uniform binary behavior, $m_h = (\Delta_h(0) + \Delta_h(1))/2$ and $v_h = (\Delta_h(1) - \Delta_h(0))^2/4$. Consequently, $\kappa_\mu(\sigma) = 1$ when every offset vanishes, recovering the centered regime; $\kappa_\mu(\sigma) \le 1 + \lambda$ when $\sum_h m_h^2 \le \lambda \sum_h v_h$ for all pairs, that is, when offsets that can cancel across stages are small relative to action contrasts; and $\kappa_\mu(\sigma)$ is unbounded when $\mathcal{R}$ contains pairs with $\sum_h m_h = 0$ and $\sum_h v_h / \sum_h m_h^2 \to 0$, the zero-contrast limit being the infinite regime above. Aligned offsets can make the ratio smaller than one; equal offsets with no contrast give $1/H$. Unlike the discounted and multiplicative cases below, additive aggregation introduces no H-dependent factor: the scale is set by the offset-to-contrast ratio of the class, and behavior enters only through the definition of offsets and contrasts.

The additive upper bound does not use this coefficient. Its proof applies change of trajectory measure (Lemma 10) to the signed sum $\sum_h (r_h^\star - r_{k,h})(s_h, a_h)$, whose second moment under μ is the outcome regression error, so only trajectory sums, not per-stage rewards, need to be identified; see Remark 2. The entry "may be ∞" in Table 2 therefore records a failure of process-reward identification, not of policy learning, and under its own assumptions the rate of Theorem 1 does not depend on which regime applies.

All-success aggregation For the all-success objective in Example 4, take $\sigma_h(u, v) = uv$ and $g = 1$, so that $R(\tau; r) = \prod_{h=1}^H r_h(s_h, a_h)$. This objective is Bellman-recursive and affine in the continuation value, but it can still be statistically hard because the multiplicative aggregation strongly attenuates per-step reward differences. To see this concretely, consider the deterministic-chain construction underlying Theorem 5: under the uniform behavior policy μ, each action sequence has probability 2^{-H}, and for each $\boldsymbol{\theta} \in \{0, 1\}^H$ define $r_{\boldsymbol{\theta},h}(s_h, a) = \mathbf{1}\{a = \theta_h\}$. For any $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, let $S(\boldsymbol{\theta}, \boldsymbol{\theta}') = \{h : \theta_h \neq \theta'_h\}$ and $k = |S(\boldsymbol{\theta}, \boldsymbol{\theta}')|$. Then the per-step reward profiles differ exactly on these k stages, so

$$\sum_{h=1}^H \mathbb{E}_\mu\big[(r_{\boldsymbol{\theta},h} - r_{\boldsymbol{\theta}',h})^2(s_h, a_h)\big] = k.$$

On the other hand, $R(\tau; r_{\boldsymbol{\theta}}) = 1$ iff $a_h = \theta_h$ for all h, an event of probability 2^{-H} under μ; the analogous event for $\boldsymbol{\theta}'$ is disjoint and also has probability 2^{-H}. Hence

$$\mathbb{E}_\mu\big[(R(\tau; r_{\boldsymbol{\theta}}) - R(\tau; r_{\boldsymbol{\theta}'}))^2\big] = 2 \cdot 2^{-H} = 2^{-(H-1)}.$$

Therefore

$$\frac{\sum_{h=1}^H \mathbb{E}_\mu[(r_{\boldsymbol{\theta},h} - r_{\boldsymbol{\theta}',h})^2(s_h, a_h)]}{\mathbb{E}_\mu[(R(\tau; r_{\boldsymbol{\theta}}) - R(\tau; r_{\boldsymbol{\theta}'}))^2]} = k\, 2^{H-1},$$

which is maximized at $k = H$, giving $\kappa_\mu(\sigma) = H2^{H-1}$ on this class. Thus κ_μ grows exponentially on this class; Theorem 5 separately proves a policy-learning lower bound. A similar intuition applies to the Bellman inverse coefficient $\chi_\mu(\sigma)$. For $\sigma_h(u, v) = uv$,

$$(\mathcal{T}_r^\pi f)_h(s,a) - (\mathcal{T}_{r'}^\pi f)_h(s,a) = (r_h - r'_h)(s,a)\,\mathbb{E}_{s'\sim P_h(\cdot|s,a)}[f_{h+1}(s',\pi)].$$

Thus a reward difference at stage h is multiplied by the continuation all-success value after (s, a). If all candidate continuations are at least $v_{\min} > 0$, then $\chi_\mu(\sigma) \le v_{\min}^{-2}$. On the binary chain, a true critic for a policy choosing each remaining correct action independently with probability p_0 has continuation p_0^{H-h}. Consequently $\chi_\mu(\sigma)$ can grow exponentially, and is infinite if an admissible pair differs on the data support only where a candidate continuation is zero. All-success aggregation therefore admits Bellman recursion while either coefficient can grow exponentially; finite coefficients alone do not verify the other assumptions of Theorem 9.

Concrete coefficient calculations The coefficients depend on the reward class and data distribution, and χ_μ also depends on $\Pi, \mathcal{F}$. The following calculations illustrate their magnitudes; satisfying all assumptions of Theorem 9, including reward-class closure, is a separate requirement.

For cumulative returns, the target difference is exactly $r_h - r'_h$, so $\chi_\mu = 1$ for any nontrivial class on the data support. The scale of κ_μ is set by the class, as derived in the cumulative-return paragraph above: it equals one for μ-centered classes, is at most one plus the offset-to-contrast ratio for action-dependent rewards under independent behavior, and is infinite when the class contains cross-stage shifts that cancel on every trajectory. The additive analysis in Theorem 1 bypasses process-reward identification and therefore needs no finite-κ_μ assumption.

For a hybrid objective with $k \ge 1$ multiplicative stages, consider

$$R(\tau; r) = \sum_{h=1}^{H-k} r_h(s_h, a_h) + \prod_{j=1}^{k} r_{H-k+j}(s_{H-k+j}, a_{H-k+j}).$$

This has additive recursion on the prefix, multiplicative recursion on the suffix, and terminal value $g = 1$. On a deterministic chain with two actions and independent uniform behavior, fix the prefix rewards throughout the class and take suffix rewards $r_{\theta,H-k+j}(a) = \mathbf{1}\{a = \theta_j\}$, $\theta \in \{0, 1\}^k$. For two parameters at Hamming distance m, the numerator of κ_μ is m, while the outcome difference is nonzero on two disjoint events of probability 2^{-k} each. Hence $\kappa_\mu = k2^{k-1}$ exactly. For χ_μ, the prefix stages contribute no target difference, and at gate j the target difference is the reward difference times the continuation $\mathbb{E}_{s'}[f_{H-k+j+1}(s', \pi)]$, with $f_{H+1} = g = 1$. Hence $\chi_\mu \le v_{\min}^{-2}$ whenever every candidate continuation at the gates is at least $v_{\min} > 0$. For $0 < p \le 1/2$, let Π consist of the policies that choose each action at every gate with probability at least p, and let $\mathcal{F}$ consist of their true critics under $\mathcal{R}$. The continuation at gate j is then the probability that the remaining $k - j$ gates succeed, which is at least p^{k-j} and, at the first gate, equals p^{k-1} for a suitable policy. A pair differing only at the first gate attains this bound, so $\chi_\mu = p^{-2(k-1)}$. The uniform policy ($p = 1/2$) gives $\chi_\mu = 4^{k-1}$, which is a lower bound whenever Π and $\mathcal{F}$ contain the uniform policy and its true critic. Both coefficients thus grow with k but not with H.

Objective and stated class	κ_μ	χ_μ
Cumulative return, μ-centered class	1	1
Cumulative return, cancelling stage shifts	Infinite	1
Hybrid, k binary gates	$k2^{k-1}$	4^{k-1}
All-success, $k = H$	$H2^{H-1}$	4^{H-1}

The χ_μ entries take Π to be the uniform policy and $\mathcal{F}$ its true critics, and are lower bounds for larger classes; a zero candidate continuation, for instance from a deterministic policy, makes χ_μ infinite for $k \ge 2$. These are coefficient calculations on the stated classes, not assertions that their reward-class closure or completeness holds. They explain two different losses of information: outcomes hide gate differences on rare events, whereas Bellman targets shrink their amplitude by the continuation value.

The two coefficients can separate The preceding hybrid calculation shows simultaneous growth. Two further class-specific calculations isolate the sources of growth. As before, these illustrate coefficients and do not assert all hypotheses of the generalized OPAC theorem.

First, take the discounted sum $R(\tau; r) = \sum_{h=1}^{H} c^{h-1} r_h(a_h)$ on a deterministic chain, with $0 < c < 1$ and independent uniform binary actions. Fix a baseline $\bar{r}_h = 1/2$, and let $r^{(j)}$ differ from it only at stage j, by $\delta(2\mathbf{1}\{a_j = 1\} - 1)$, where $0 < \delta \leq 1/2$. Set $\mathcal{R} = \{\bar{r}, r^{(1)}, \ldots, r^{(H)}\}$. The recursion $\sigma_h(u, v) = u + cv$ has target difference $r_h - r'_h$, so $\chi_\mu = 1$ independently of the critic class. For the pair $(\bar{r}, r^{(j)})$, the reward squared distance is δ^2 and the outcome squared distance is $c^{2(j-1)}\delta^2$. For $(r^{(j)}, r^{(\ell)})$ with $j \neq \ell$, these distances are $2\delta^2$ and $\delta^2(c^{2(j-1)} + c^{2(\ell-1)})$, respectively, because cross terms have mean zero. It follows exactly that

$$\kappa_\mu = c^{-2(H-1)}, \qquad \chi_\mu = 1.$$

Thus κ_μ can grow exponentially while $\chi_\mu = 1$.

Second, take all-success rewards on a binary chain with $H \geq 2$, and the two-point class $\{r_\theta, r_{\theta'}\}$ where θ, θ' differ only at the first stage. Let behavior choose the action θ_h with probability $p = 1 - 1/H$ independently at each stage. The squared reward distance is one. The two distinct success events have combined probability $p^H + (1-p)p^{H-1} = p^{H-1}$, giving

$$\kappa_\mu = p^{-(H-1)} \leq e.$$

If Π includes the uniform policy and $\mathcal{F}$ includes its true critic, the first-stage continuation for that policy is $2^{-(H-1)}$. The two rewards differ only at that stage, and their target squared distance is $4^{-(H-1)}$, irrespective of the behavior bias. Hence $\chi_\mu \geq 4^{H-1}$. Informative behavior can therefore control outcome identification while candidate Bellman continuations still suppress reward differences.

Conditional continuation (game-over) Take $a_h(u) = \mathbf{1}\{u > 0\}$, $b_h(u) = u$, and $g = 0$: $\sigma_h(u, v) = u + \mathbf{1}\{u > 0\}v$, $R(\tau) = \sum_{h=1}^{H_{\text{stop}}} r_h(s_h, a_h)$, where $H_{\text{stop}} = \min\{h : r_h(s_h, a_h) \leq 0\} \wedge H$, with $\min \varnothing = +\infty$. Once a non-positive reward is received, all subsequent rewards are discarded. Per-step feedback reveals the stopping time. The scalar outcome need not identify it, and the gate $a_h(u) = \mathbf{1}\{u > 0\}$ is not Lipschitz. For example, when $H = 10$, both reward sequences

$$(1, 0, \star, \ldots, \star) \qquad \text{and} \qquad (0.1, 0.1, \ldots, 0.1)$$

produce the same scalar outcome $R(\tau) = 1$. The first trajectory stops at stage 2, after which all downstream rewards are discarded, while the second makes small positive progress at every stage. Thus per-step feedback distinguishes these cases immediately, but the scalar outcome collapses them into the same observation. Although the recursion is affine in the continuation value and contractive, the exact game-over objective lies outside Assumption 4 because of the hard, non-Lipschitz gate. A smoothed continuation gate can satisfy the Lipschitz condition, but class closure and finite coefficients must still be verified.

Exponential utility Fix a risk parameter $\lambda \in (0, 1]$ and take $a_h(u) = e^{-\lambda u}$, $b_h(u) = 1 - e^{-\lambda u}$, and $g = 0$, so that $\sigma_h(u, v) = 1 - e^{-\lambda u}(1 - v)$. Since $1 - R_h = e^{-\lambda r_h}(1 - R_{h+1})$ and $R_{H+1} = 0$,

$$R(\tau) = 1 - e^{-\lambda \sum_{h=1}^{H} r_h(s_h, a_h)} \in [0, 1), \qquad J(\pi) = 1 - \mathbb{E}_{\tau \sim \pi}\left[e^{-\lambda \sum_{h=1}^{H} r_h}\right].$$

As $t \mapsto -\lambda^{-1} \log(1 - t)$ is increasing, maximizing J maximizes the risk-averse entropic criterion $-\lambda^{-1} \log \mathbb{E}_\pi[e^{-\lambda \sum_h r_h}]$ of risk-sensitive RL (Howard & Matheson, 1972; Wang et al., 2024a; Han et al., 2025). The recursion is affine and increasing in the continuation value, $a_h(u) \in [e^{-\lambda}, 1]$, and a_h, b_h are λ-Lipschitz, so items (1)–(3) of Assumption 4 hold with $L = 1$, and all values lie in $[0, 1]$. For a risk-seeking parameter $\lambda > 0$, the normalized recursion $\sigma_h(u, v) = e^{\lambda(u-1)}v$ with $g = 1$ gives $R(\tau) = e^{\lambda(\sum_h r_h - H)}$, so maximizing J maximizes $\lambda^{-1} \log \mathbb{E}_\pi[e^{\lambda \sum_h r_h}]$; here $a_h(u) \in [e^{-\lambda}, 1]$ and items (1)–(3) hold with $L = \max\{1, \lambda\}$. In both cases, reward-class closure and finite coefficients remain additional requirements.

Max-reward RL Quah & Quek (2006); Gottipati et al. (2020); Cui & Yu (2023); Veviurko et al. (2024) propose a paradigm in which an agent optimizes the maximum (or minimum) reward rather than the cumulative reward. Consider $R^{\max}(\tau; r) = \max_{h \in [H]} r_h(s_h, a_h)$. This objective can be represented recursively as $\sigma_h(u, v) = \max\{u, v\}$, but the recursion is not affine in v, so our generalized Bellman theory in Appendix F and Theorem 9 does not apply. We can also illustrate a

separate information-theoretic obstruction arising from reward-process aggregation. Take a simple example: suppose every candidate reward satisfies $r_H(s,a) = 1$ at every terminal state–action pair and $r_h(s_h, a_h) < 1$ for all $h < H$ throughout the μ-support. Then $R^{\max}(\tau; r) = 1$ holds μ-almost surely for every $r \in \mathcal{R}$, so $\mathbb{E}_\mu[(R^{\max}(\tau; r) - R^{\max}(\tau; r'))^2] = 0$ for all pairs r, r', while earlier-stage rewards may still differ. If $\mathcal{R}$ contains such distinct pairs with $\sum_{h=1}^{H} \mathbb{E}_\mu[(r_h - r'_h)^2(s_h, a_h)] > 0$, then the denominator in Definition 1 vanishes and $\kappa_\mu(\sigma) = +\infty$. This prevents reward-process identification, but is not a policy-learning lower bound: in this particular example every policy has the same return one. Thus infinite κ_μ does not by itself imply hard policy optimization.

I The Underlying Statistical Problem: σ-Composition Regression

This appendix records a self-contained regression abstraction: i.i.d. draws $x_i = (x_{i,1}, \dots, x_{i,H}) \sim \mu$, scalar outcomes Y_i, and the reward-process coefficient $\kappa_\mu(\sigma)$ from Definition 1, read with $R(x; r)$ below in place of $R(\tau; r)$ and with r_h evaluated at the coordinate x_h. The coefficient controls reward-process error through prediction error; the lower bound below concerns a specified pair under Gaussian noise.

I.1 Regression Formulation

Fix $V_{\max} > 0$, measurable spaces $\mathcal{X}_1, \dots, \mathcal{X}_H$, a law μ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_H$, a known $\sigma : \mathbb{R}^H \to \mathbb{R}$, and classes $\mathcal{R}_h$ of measurable maps $r_h : \mathcal{X}_h \to \mathbb{R}$ with $\mathcal{R} \subseteq \mathcal{R}_1 \times \cdots \times \mathcal{R}_H$. For $x = (x_1, \dots, x_H)$ and $r = (r_1, \dots, r_H) \in \mathcal{R}$, define the composed outcome $R(x; r) := \sigma\big(r_1(x_1), \dots, r_H(x_H)\big)$. Assume $|R(x; r)| \le V_{\max}$ for every $r \in \mathcal{R}$ and μ-a.e. x. Let $r^\star \in \mathcal{R}$ and suppose we observe

$$(x_i, Y_i) \overset{\text{i.i.d.}}{\sim} P, \qquad P_x = \mu, \quad i \in [n],$$

with $\mathbb{E}[Y_i \mid x_i] = R(x_i; r^\star)$ and $|Y_i| \le V_{\max}$ almost surely. Write

$$\|r - r'\|^2 := \sum_{h=1}^{H} \mathbb{E}_\mu\big[(r_h(x_h) - r'_h(x_h))^2\big].$$

The coefficient $\kappa_\mu(\sigma)$ is as in Definition 1; equivalently, with $\kappa = \kappa_\mu(\sigma)$, it is the smallest $\kappa \ge 0$ such that

$$\|r - r'\|^2 \le \kappa\, \mathbb{E}_{x\sim\mu}\big[(R(x; r) - R(x; r'))^2\big] \quad \text{for all distinct } r, r' \in \mathcal{R},$$

adopting the ratio conventions $a/0 = +\infty$ for $a > 0$ and $0/0 = 0$. For distinct $r, r' \in \mathcal{R}$ with $\mathbb{E}_{x\sim\mu}\big[(R(x; r) - R(x; r'))^2\big] > 0$, define the pair-specific ratio

$$\kappa_{r,r'}(\sigma) := \frac{\|r - r'\|^2}{\mathbb{E}_{x\sim\mu}\big[(R(x; r) - R(x; r'))^2\big]}, \tag{102}$$

so $\kappa_\mu(\sigma) = \sup_{r \ne r'} \kappa_{r,r'}(\sigma)$ under the same conventions. The ERM estimator is

$$\hat{r} \in \operatorname*{argmin}_{r\in\mathcal{R}} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - R(x_i; r)\big)^2. \tag{103}$$

I.2 Bounds for σ-Composition Regression

The next theorem records a high-probability upper bound for the empirical risk minimizer from Eq. (103).

Theorem 34 (Upper bound for σ-composition regression)**.** *Suppose $\mathcal{R}$ is finite, $\kappa_\mu(\sigma) < \infty$, $r^\star \in \mathcal{R}$, and $(x_i, Y_i)_{i=1}^n$ are i.i.d. with $x_i \sim \mu$ and*

$$\mathbb{E}[Y_i \mid x_i] = R(x_i; r^\star).$$

Assume moreover that

$$|R(x; r)| \le V_{\max} \quad \textit{for all } r \in \mathcal{R} \textit{ and } \mu\textit{-a.e. } x, \qquad |Y_i| \le V_{\max} \quad \textit{a.s.}$$

Then there exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$, any empirical risk minimizer $\hat{r}$ from Eq. (103) satisfies

$$\|\hat{r} - r^\star\|^2 \le C\, \kappa_\mu(\sigma)\, \frac{V_{\max}^2 \log(|\mathcal{R}|/\delta)}{n}.$$

The proof of Theorem 34 is given in Appendix I.3. For a minimax lower bound, we isolate the hardness coming from compressing per-step structure into a single trajectory outcome: we take two candidates r, r' at $\|\cdot\|$-separation. Under i.i.d. Gaussian outcome noise at scale $V_{\max}$, Theorem 35 shows that $\Omega\big(V_{\max}^2 \kappa_{r,r'}(\sigma)/\Delta^2\big)$ samples can be necessary even for distinguishing this pair.

Theorem 35 (Lower bound for σ-composition regression)**.** *Suppose* $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^n$ *with* $x_i \overset{i.i.d.}{\sim} \mu$ *and*

$$Y_i = R(x_i; r^\star) + \xi_i, \qquad \xi_1, \dots, \xi_n \overset{i.i.d.}{\sim} \mathcal{N}(0, V_{\max}^2),$$

with $(\xi_i)_{i=1}^n$ *independent of* $(x_i)_{i=1}^n$*. Fix distinct* $r, r' \in \mathcal{R}$ *with* $\|r - r'\|^2 = \Delta^2 > 0$ *and finite* $\kappa_{r,r'}(\sigma)$ *as in Eq. (102). Then whenever* $n \le V_{\max}^2 \kappa_{r,r'}(\sigma)/\Delta^2$*,*

$$\inf_{\hat{r}} \max_{r^\star \in \{r, r'\}} \mathbb{E}\big[\|\hat{r} - r^\star\|^2\big] \ge \frac{\Delta^2}{16}.$$

In particular, for any $\varepsilon < \Delta^2/16$*, every estimator* $\hat{r}$ *with* $\max_{r^\star \in \{r,r'\}} \mathbb{E}[\|\hat{r} - r^\star\|^2] \le \varepsilon$ *requires* $n = \Omega\big(V_{\max}^2 \kappa_{r,r'}(\sigma)/\Delta^2\big)$.

The proof of Theorem 35 is given in Appendix I.4. If $\kappa_{r,r'}(\sigma) = \infty$, the two data laws coincide, and the same bound holds for every n. The upper bound transfers prediction error to reward-process error using κ_μ. The lower bound relates pairwise distinguishability to $\kappa_{r,r'}$ under Gaussian noise, which differs from the bounded-label model of the upper bound; these results do not assert a matching minimax rate for every class.

I.3 Proof of Theorem 34

For $r \in \mathcal{R}$ set $g_r(x) := R(x; r)$ and $g_\star(x) := R(x; r^\star)$. Let (x, Y) be an independent copy from the same distribution as each sample pair, so $x \sim \mu$ and $\mathbb{E}[Y \mid x] = g_\star(x)$. Let $\mathcal{D} = \{(x_i, Y_i)\}_{i=1}^n$ and write

$$\mathbb{E}_{\mathcal{D}}[\phi] := \frac{1}{n} \sum_{i=1}^n \phi(x_i, Y_i)$$

for measurable ϕ. Define the squared loss $\ell_r(x, Y) := (Y - g_r(x))^2$ and the excess loss

$$Z_r(x, Y) := \ell_r(x, Y) - \ell_{r^\star}(x, Y).$$

Then

$$\begin{aligned} \mathbb{E}[Z_r(x, Y)] &= \mathbb{E}\big[(Y - g_r(x))^2 - (Y - g_\star(x))^2\big] \\ &= \mathbb{E}_{x \sim \mu}\big[(g_r(x) - g_\star(x))^2\big] =: L_r. \end{aligned}$$

In particular, L_r is the population squared prediction error of g_r under μ.

Fix $r \in \mathcal{R}$. Expanding the square gives

$$Z_r(x, Y) = (Y - g_r(x))^2 - (Y - g_\star(x))^2 = (g_\star(x) - g_r(x))\big(2Y - g_r(x) - g_\star(x)\big).$$

Since $|Y|, |g_r(x)|, |g_\star(x)| \le V_{\max}$ almost surely,

$$|Z_r(x, Y)| \le 4V_{\max}^2$$

and

$$Z_r(x, Y)^2 \le 16 V_{\max}^2 (g_r(x) - g_\star(x))^2.$$

Therefore

$$\mathrm{Var}(Z_r(x, Y)) \le \mathbb{E}[Z_r(x, Y)^2] \le 16 V_{\max}^2 L_r.$$

Also $L_r = \mathbb{E}_{x \sim \mu}[(g_r(x) - g_\star(x))^2] \le 4V_{\max}^2$, hence $|Z_r(x, Y) - L_r| \le 8V_{\max}^2$ almost surely. Apply one-sided Bernstein to the i.i.d. variables $Z_r(x_1, Y_1), \dots, Z_r(x_n, Y_n)$: for any fixed $r \in \mathcal{R}$ and any $\eta \in (0, 1)$, with probability at least $1 - \eta$,

$$\mathbb{E}_{\mathcal{D}}[Z_r] - L_r \ge -\sqrt{\frac{32 V_{\max}^2 L_r \log(1/\eta)}{n}} - \frac{16 V_{\max}^2 \log(1/\eta)}{3n}.$$

Set $\eta = \delta/|\mathcal{R}|$ and take a union bound over $r \in \mathcal{R}$. With probability at least $1-\delta$, the preceding display holds simultaneously for every $r \in \mathcal{R}$. On this event, if $\mathbb{E}_{\mathcal{D}}[Z_r] \le 0$ then

$$L_r \le \sqrt{\frac{32V_{\max}^2 L_r \log(|\mathcal{R}|/\delta)}{n}} + \frac{16V_{\max}^2 \log(|\mathcal{R}|/\delta)}{3n}.$$

Using $\sqrt{32V_{\max}^2 L_r \log(|\mathcal{R}|/\delta)/n} \le L_r/2 + 16V_{\max}^2 \log(|\mathcal{R}|/\delta)/n$ and rearranging yields

$$L_r \le C\frac{V_{\max}^2 \log(|\mathcal{R}|/\delta)}{n}$$

for a universal constant $C > 0$.

If $\hat{r}$ is an empirical risk minimizer and $r^\star \in \mathcal{R}$, then $\mathbb{E}_{\mathcal{D}}[Z_{\hat{r}}] \le 0$.

On this same event,

$$\mathbb{E}_{x\sim\mu}\big[(R(x;\hat{r}) - R(x;r^\star))^2\big] = L_{\hat{r}} \le C\frac{V_{\max}^2 \log(|\mathcal{R}|/\delta)}{n}.$$

Finally, by Definition 1,

$$\|\hat{r} - r^\star\|^2 \le \kappa_\mu(\sigma)\, \mathbb{E}_{x\sim\mu}\big[(R(x;\hat{r}) - R(x;r^\star))^2\big].$$

Combining the last two displays completes the proof.

I.4 Proof of Theorem 35

For $r^\star \in \{r, r'\}$, let $\mathbb{P}_{r^\star}$ be the law of one draw (x, Y) under that truth, and let $\mathbb{E}_{r^\star}$ denote expectation when $\mathcal{D}$ consists of n i.i.d. draws from $\mathbb{P}_{r^\star}$. Write $\delta_R^2 := \mathbb{E}_{x\sim\mu}\big[(R(x;r) - R(x;r'))^2\big] = \Delta^2/\kappa_{r,r'}(\sigma)$. Conditional on x, the two models differ only in the mean of a Gaussian with variance $V_{\max}^2$, so

$$\mathrm{KL}(\mathbb{P}_r \| \mathbb{P}_{r'}) = \frac{\delta_R^2}{2V_{\max}^2}.$$

By tensorization, $\mathrm{KL}(\mathbb{P}_r^{\otimes n} \| \mathbb{P}_{r'}^{\otimes n}) = n\,\mathrm{KL}(\mathbb{P}_r \| \mathbb{P}_{r'})$, and Pinsker's inequality gives

$$\mathrm{TV}\big(\mathbb{P}_r^{\otimes n}, \mathbb{P}_{r'}^{\otimes n}\big) \le \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathbb{P}_r^{\otimes n} \| \mathbb{P}_{r'}^{\otimes n})} = \sqrt{\frac{n\delta_R^2}{4V_{\max}^2}} \le \frac{1}{2}$$

when $n \le V_{\max}^2/\delta_R^2$.

The pair (r, r') is Δ-separated in $\|\cdot\|$ because $\|r - r'\| = \Delta$. By the two-point Le Cam method (Yu, 1997), when $n \le V_{\max}^2/\delta_R^2$,

$$\inf_{\hat{r}} \max_{r^\star \in \{r,r'\}} \mathbb{E}_{r^\star}\left[\|\hat{r} - r^\star\|^2\right] \ge \frac{\Delta^2}{8}\Big(1 - \mathrm{TV}\big(\mathbb{P}_r^{\otimes n}, \mathbb{P}_{r'}^{\otimes n}\big)\Big) \ge \frac{\Delta^2}{16}.$$