TOWARDS $\sqrt{K}$ REGRET FOR PRIVATE RL WITH GENERAL FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on private reinforcement learning with general function approximation establishes $\widetilde{O}(K^{3/5})$ regret over K episodes, but whether privacy permits the canonical $\widetilde{O}(\sqrt{K})$ dependence remains open. We resolve this question for a finite Bellman-complete value-function class with bounded Bellman–Eluder dimension. Our main algorithm, `Priv-LazyGOLF`, privatizes rare policy switching through a twofold use of `AboveThreshold`: it privately detects when the current hypothesis has accumulated enough Bellman inconsistency and privately selects an optimistic feasible replacement. We show that `Priv-LazyGOLF` is (ε, δ)-jointly differentially private and, with high probability, achieves $\widetilde{O}(\sqrt{K})$ regret with only logarithmically many policy switches in K. Thus, privacy preserves the $\sqrt{K}$ regret of non-private learning. We also establish that the same private rare-switching template extends, through suitable decomposable losses, to a broader class of reinforcement learning problems while retaining $\widetilde{O}(\sqrt{K})$ regret and privacy.

1 INTRODUCTION

Reinforcement learning (RL) increasingly personalizes sequential decisions in healthcare and online services. A treatment trajectory may contain medical history, measurements, treatments, and outcomes, while a recommendation trajectory may reveal contexts, choices, and feedback. Such data improve future policies but can expose sensitive user information through later recommendations (Vietri et al., 2020; Qiao & Wang, 2023; Zhou, 2022).

This concern has motivated a growing theoretical literature on private online RL, whose goal is to achieve sublinear regret while providing formal privacy guarantees. The standard privacy notion here is *joint differential privacy* (JDP), which protects one user's trajectory from the outputs shown to all other users while allowing that user's own action to depend on their data (Kearns et al., 2014; Vietri et al., 2020). Before the recent work of He & Zhou (2026), online regret guarantees under JDP were restricted to tabular MDPs (Vietri et al., 2020; Chowdhury & Zhou, 2022; Qiao & Wang, 2023) or linear function approximation (Zhou, 2022; Luyo et al., 2021). Those algorithms rely heavily on counts or low-dimensional sufficient statistics provided by the tabular or linear structure.

He & Zhou (2026) took the first step towards private RL with *general function approximation*. By combining fixed-size batching with the exponential mechanism, they obtained $\widetilde{O}(K^{3/5})$ regret under JDP, where K is the number of episodes. Without privacy, however, RL with general function approximation can achieve the canonical $\widetilde{O}(\sqrt{K})$ regret under structural conditions such as Bellman–Eluder dimension or coverability (Jin et al., 2021; Xie et al., 2023). The same bound can even be attained with only logarithmically many policy updates via the rare-switching method (Xiong et al., 2024). The result of He & Zhou (2026) therefore leaves open the following question:

Can private RL with general function approximation attain $\widetilde{O}(\sqrt{K})$ regret?

We answer this question affirmatively, taking the first step towards closing the gap between private and non-private RL with general function approximation. Our contributions are summarized below.

1. Privatizing `GOLF` through `AboveThreshold`. Our algorithm, `Priv-LazyGOLF` (Algorithm 2), builds on `GOLF` (Jin et al., 2021) and incorporates the rare-switching idea of Xiong et al. (2024). Rare switching requires two data-dependent decisions: when the current hypothesis becomes inconsistent

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.

and which plausible replacement is most optimistic. Because both depend on an evolving confidence set, the underlying constrained optimization is difficult to privatize directly (He & Zhou, 2026). Our key observation is that both decisions admit the same one-positive threshold interface (Dwork & Roth, 2014): an `AboveThreshold` trigger determines when to update, and an `AboveThreshold` selector chooses the optimistic replacement.

2. $\sqrt{K}$ regret under JDP. For finite general function classes satisfying Bellman completeness and bounded Bellman–Eluder dimension, Theorem 5.1 proves that `Priv-LazyGOLF` is (ε, δ)-JDP while attaining $\widetilde{O}(\sqrt{K})$ regret. The key to our result is that `AboveThreshold` turns rare switching into a privacy advantage: negative answers incur no additional privacy cost within an invocation, so the total privacy cost is composed over only logarithmically many switches. This avoids the polynomial composition of fixed batching (He & Zhou, 2026) and enables the canonical $\sqrt{K}$ regret dependence.

3. Extension to the ℓ_2-EC framework. Finally, we extend both our algorithm and analysis to the general ℓ_2-Eluder Condition (EC) framework of Xiong et al. (2024). Theorem 6.1 in Section 6 establishes JDP and $\widetilde{O}(\sqrt{K})$ regret under this broader framework. The EC framework subsumes our Bellman setting: its decomposable-loss structure generalizes the Bellman regression induced by Bellman completeness, while its ℓ_2-Eluder condition generalizes the exploration control provided by Bellman–Eluder dimension.

1.1 Related Work

Private RL. Regret guarantees for private RL were first established for tabular MDPs (Vietri et al., 2020; Chowdhury & Zhou, 2022; Qiao & Wang, 2023) and later for linear function approximation (Zhou, 2022; Luyo et al., 2021). Existing private rare-switching analyses for linear MDPs (Ngo et al., 2022; Sahu, 2025) contain gaps identified by He & Zhou (2026). Moving beyond linear structure, He & Zhou (2026) obtained the first JDP regret guarantee for general function approximation, achieving $\widetilde{O}(K^{3/5})$ regret under coverability via fixed batching and the exponential mechanism. We instead achieve $\widetilde{O}(\sqrt{K})$ regret through private rare switching under Bellman–Eluder dimension. This condition and coverability are incomparable, see Appendix A for detailed discussions.

General function approximation and adaptivity. Exploration in non-private RL beyond linear models has been studied through the Eluder and Bellman–Eluder dimensions, Bellman and witness ranks, bilinear classes, the decision–estimation coefficient, coverability, the sequential extrapolation coefficient, and the generalized Eluder coefficient (Wang et al., 2020; Jin et al., 2021; Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Foster et al., 2021; Xie et al., 2023; Zhong et al., 2022). Rare-switching methods aim to retain $\sqrt{K}$ regret with logarithmically many updates (Abbasi-Yadkori et al., 2011; Bai et al., 2019; Kong et al., 2021; Qiao et al., 2023; Xiong et al., 2024; Zhao et al., 2024). We bring rare switching to private RL with general function approximation by using `AboveThreshold` (Dwork & Roth, 2014) to privatize both when to switch and which hypothesis to select.

2 Preliminaries

2.1 Online RL with General Function Approximation

Finite-horizon MDPs. We consider an unknown episodic MDP $M^\star = (\mathcal{S}, \mathcal{A}, H, P, R, s_1)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, horizon H, stage-h transition kernel $P_h(\cdot \mid s, a)$, conditional mean reward $R_h(s, a) \in [0, 1]$, and fixed public initial state s_1. A policy $\pi = (\pi_1, \ldots, \pi_H)$ maps states to action distributions and generates trajectories by $a_h \sim \pi_h(\cdot \mid s_h)$, $\mathbb{E}[r_h \mid s_h, a_h] = R_h(s_h, a_h)$, $s_{h+1} \sim P_h(\cdot \mid s_h, a_h)$. We assume $r_h \in [0, 1]$ a.s. and normalize so that $\sum_{h=1}^{H} r_h \in [0, 1]$ a.s. For a policy π, its value and action-value functions at any $h \in [H]$ are

$$V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=h}^{H} r_t \,\middle|\, s_h = s\right], \qquad Q_h^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{t=h}^{H} r_t \,\middle|\, s_h = s, a_h = a\right].$$

We define $J(\pi) := V_1^\pi(s_1)$. Let $\pi^\star$ be an optimal policy and write $V_h^\star = V_h^{\pi^\star}$ and $Q_h^\star = Q_h^{\pi^\star}$. These functions obey the following Bellman optimality equations:

$$V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s, a), \qquad Q_h^\star(s, a) = R_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot \mid s, a)} V_{h+1}^\star(s').$$

For any $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define the optimal Bellman operator by $[\mathcal{T}_h f](s,a) := R_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)} \max_{a' \in \mathcal{A}} f(s', a')$.

Online interaction and performance. The learner interacts with $M^\star$ for K episodes. Before episode k, it selects a policy π^k using only the preceding interaction history and then observes the resulting trajectory $\{(s_h^k, a_h^k, r_h^k, s_{h+1}^k)\}_{h=1}^H$ for episode k. We evaluate the learner's performance by the cumulative regret:

$$\mathsf{Reg}(K) = \sum_{k=1}^{K} \big(J(\pi^\star) - J(\pi^k)\big).$$

General function approximation. We adopt value-function approximation so that the regret guarantee need not scale with the cardinality of $\mathcal{S}$. In particular, the learner is given a class of product action-value functions $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_H$ where $\mathcal{F}_h \subseteq \{f_h : \mathcal{S} \times \mathcal{A} \to [0,1]\}$. Each $f = (f_1, \ldots, f_H) \in \mathcal{F}$ induces the value function and greedy policy for each $h \in [H]$ as follows.

$$V_{f,h}(s) := \max_{a \in \mathcal{A}} f_h(s,a), \qquad \pi_{f,h}(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} f_h(s,a).$$

Sublinear regret with a general function class in RL typically requires two types of structural conditions: *representation conditions*, which ensure that $\mathcal{F}$ contains the value functions required by the learner, and *exploration conditions*, which limit the exploration complexity. In the main paper, we use Bellman completeness for the former and low Bellman–Eluder dimension for the latter.

Assumption 2.1 (Bellman completeness)**.** $\mathcal{T}_h \mathcal{F}_{h+1} \subseteq \mathcal{F}_h$ for every $h \in [H]$ and $\mathcal{F}_{H+1} = \{0\}$.

Here $\mathcal{T}_h \mathcal{F}_{h+1} := \{\mathcal{T}_h q : q \in \mathcal{F}_{h+1}\}$. Bellman completeness, a common assumption in RL, implies realizability: $Q_h^\star = \mathcal{T}_h Q_{h+1}^\star \in \mathcal{F}_h$, which holds by backward induction.

For $f \in \mathcal{F}$, define its stage-h Bellman residual and the induced Bellman residual class by

$$e_h(f; s, a) := f_h(s,a) - [\mathcal{T}_h f_{h+1}](s,a), \qquad \mathcal{E}_h(\mathcal{F}) := \{e_h(f; \cdot) : f \in \mathcal{F}\}.$$

We measure the exploration complexity of $\mathcal{F}$ through its Bellman residual classes via the point-mass Bellman–Eluder dimension (Jin et al., 2021).

Definition 2.2 (Point-mass Bellman–Eluder dimension)**.** Let Φ be a real-valued function class on a domain $\mathcal{X}$, and let D be a family of probability measures on $\mathcal{X}$. A measure ν is η-independent of $\mu_1, \ldots, \mu_n$ with respect to Φ if there exists $\phi \in \Phi$ such that $\sum_{i=1}^n (\mathbb{E}_{x \sim \mu_i} \phi(x))^2 \le \eta^2$ and $|\mathbb{E}_{x \sim \nu} \phi(x)| > \eta$. Define $\dim_{\mathsf{DE}}(\Phi, D, \eta)$ as the largest n for which there exist $\mu_1, \ldots, \mu_n \in D$ and a common scale $\eta' \ge \eta$ such that each μ_j is η'-independent of $\mu_1, \ldots, \mu_{j-1}$. For each $h \in [H]$, let $D_{\Delta,h} = \{\delta_x : x = (s,a) \in \mathcal{S} \times \mathcal{A}\}$ be the family of point masses at stage h, and define $D_\Delta := \{D_{\Delta,h}\}_{h=1}^H$. The *point-mass Bellman–Eluder dimension* of $\mathcal{F}$ at scale η is

$$\dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, \eta) := \max_{h \in [H]} \dim_{\mathsf{DE}}\big(\mathcal{E}_h(\mathcal{F}), D_{\Delta,h}, \eta\big).$$

Because the measures in $D_{\Delta,h}$ are Dirac, Definition 2.2 has a direct pointwise interpretation: a sequence is long only if, at every new point x_j, some Bellman residual is small in squared norm on $x_1, \ldots, x_{j-1}$ but remains larger than the common scale at x_j. Hence, a small BE-dimension rules out too many such "surprising" residuals and controls the exploration complexity.

Remark 2.3 (Choice of distribution family in BE dimension)**.** Jin et al. (2021) consider two canonical distribution families: the greedy-policy-induced roll-in distributions $D_\mathcal{F}$ and the point-mass family D_Δ. They define the Bellman–Eluder dimension as the minimum of the two induced dimensions, which are incomparable in general. We consider only the D_Δ-Bellman–Eluder dimension, following the low-switching analysis of Xiong et al. (2024).

2.2 Joint Differential Privacy in RL

Each episode is associated with one unique user. Let $U = (u_1, \ldots, u_K) \in \mathcal{U}^K$ be the user sequence, where a user record contains all private information governing that episode's interaction. Two sequences U, U' are j-neighbors if they agree except at user j. The complete output of an online learner contains the actions delivered during all episodes, and its final policy. Requiring standard DP (Dwork et al., 2006) for this entire output is too restrictive: replacing user j may legitimately

change the actions personalized to that same user. Indeed, even in contextual bandits, protecting the complete action sequence can preclude sublinear regret (Shariff & Sheffet, 2018). Joint differential privacy (Kearns et al., 2014) resolves this issue by excluding user j's own actions from the view protected against changes to u_j and is the standard privacy notion adopted in prior private-RL work (Vietri et al., 2020; Chowdhury & Zhou, 2022; Qiao & Wang, 2023; Zhou, 2022).

Definition 2.4 (JDP in RL (Vietri et al., 2020))**.** Let $\mathcal{M}_{-j}(U)$ denote every output visible outside episode j, including all other users' actions and the final policy, but excluding $(a_1^j, \dots, a_H^j)$. An online RL algorithm is (ε, δ)-jointly differentially private if, for every $j \in [K]$, every pair of j-neighboring user sequences U, U', and every measurable event E,

$$\mathbb{P}\big(\mathcal{M}_{-j}(U) \in E\big) \le e^{\varepsilon}\mathbb{P}\big(\mathcal{M}_{-j}(U') \in E\big) + \delta.$$

JDP still protects user j against an arbitrary coalition of all other users, provided that user j does not disclose the actions delivered privately to them. We use the standard unique-user interface: replacing u_j changes only episode j's trajectory row, while another user's actions depend only on the public transcript and that user's own record. Under this interface, public-transcript standard DP implies JDP by the Billboard Lemma (Hsu et al., 2016; Vietri et al., 2020).

3 The Non-Private LazyGOLF Algorithm

In this section, we review the Bellman-regression specialization of Algorithm 1 in Xiong et al. (2024), which extends GOLF (Jin et al., 2021) by updating its policy only when the current hypothesis accumulates sufficient Bellman error. We call this specialization LazyGOLF and use it as the non-private template for our main private algorithm. Section 6 treats the broader ℓ_2-EC formulation. We begin by defining the empirical Bellman-loss quantities used by LazyGOLF. For $q \in \mathcal{F}_{h+1}$, recall that $V_q(s) = \max_{a\in\mathcal{A}} q(s,a)$ and that the value at stage $H+1$ is zero. Given $u \in \mathcal{F}_h$ and $q \in \mathcal{F}_{h+1}$,

$$\widehat{L}_{h,t}(u;q) := \sum_{i=1}^{t}\big(u(s_h^i,a_h^i) - r_h^i - V_q(s_{h+1}^i)\big)^2, \tag{1}$$

$$\widehat{\mathcal{E}}_{h,t}(f) := \widehat{L}_{h,t}(f_h; f_{h+1}) - \min_{u\in\mathcal{F}_h}\widehat{L}_{h,t}(u; f_{h+1}),$$

$$\widehat{\mathcal{E}}_t(f) := \max_{h\in[H]}\widehat{\mathcal{E}}_{h,t}(f), \qquad \widehat{\mathcal{E}}_0(f) := 0. \tag{2}$$

Here, $\widehat{L}_{h,t}$ is the empirical squared Bellman loss, $\widehat{\mathcal{E}}_{h,t}(f)$ is the corresponding excess loss relative to the best fit in $\mathcal{F}_h$, and $\widehat{\mathcal{E}}_t(f)$ is the maximum excess loss over all stages.

LazyGOLF algorithm. GOLF (Jin et al., 2021) recomputes an optimistic hypothesis every episode; LazyGOLF (Xiong et al., 2024) retains it until the excess Bellman loss exceeds a switching threshold. Let $\beta_{\mathsf{fit}} < \beta_{\mathsf{switch}}$ be suitable thresholds and initialize $f^1 \in \operatorname{argmax}_{f\in\mathcal{F}} V_{f,1}(s_1)$ with fixed public tie-breaking. LazyGOLF executes the greedy policy π_{f^k} in episode k and updates:

$$f^{k+1} \in \begin{cases} \{f^k\}, & \widehat{\mathcal{E}}_k(f^k) < \beta_{\mathsf{switch}}, \\ \operatorname*{argmax}_{f\in\mathcal{F}:\,\widehat{\mathcal{E}}_k(f)\le\beta_{\mathsf{fit}}} V_{f,1}(s_1), & \widehat{\mathcal{E}}_k(f^k) \ge \beta_{\mathsf{switch}}. \end{cases} \tag{3}$$

Thus, LazyGOLF performs the same global optimistic update as GOLF, but only when the current hypothesis becomes empirically inconsistent with the accumulated data, as represented by $\widehat{\mathcal{E}}_k(f^k)$. The separation between β_{fit} and β_{switch} ensures that a newly selected hypothesis must accumulate sufficient additional Bellman error before another switch occurs.

Why laziness matters for privacy. LazyGOLF retains the standard $\widetilde{O}(\sqrt{K})$ regret bound with only $O(\log K)$ adaptive, data-dependent policy updates (Xiong et al., 2024). Because privacy loss composes across updates, rare switching limits its accumulation, unlike the fixed-size, *data-independent* batching of He & Zhou (2026), which requires polynomially many updates.[1]

[1] Adaptive data-dependent rare switching for private linear MDPs has been explored (Ngo et al., 2022; Sahu, 2025), though He & Zhou (2026) identified gaps in those analyses.

4 OUR APPROACH: PRIVATIZING LAZYGOLF

Motivated by the preceding discussion, our algorithm privatizes LazyGOLF while preserving its lazy-update structure. The key idea centers on the following two data-dependent operations in (3):

$$\underbrace{\widehat{\mathcal{E}}_k(f^k) \geq \beta_{\text{switch}}}_{\text{trigger}} \qquad \text{and} \qquad \underbrace{\operatorname*{argmax}_{f:\,\widehat{\mathcal{E}}_k(f)\leq\beta_{\text{fit}}} V_{f,1}(s_1)}_{\text{selector}}. \tag{4}$$

The trigger determines when the current hypothesis is replaced, whereas the selector determines which hypothesis is used thereafter. Although the numerical excess losses remain internal to the learner, both decisions are revealed through its subsequent policies and must therefore be privatized.

Remark 4.1 (Finite-class setting)**.** As in He & Zhou (2026), we consider a finite function class $\mathcal{F}$ with $N := |\mathcal{F}|$, so that the selector can examine the hypotheses in a fixed public order.

Privatizing control flow rather than scores. A direct approach would privately release a perturbed score $\widehat{\mathcal{E}}_k(f^k)$ after every episode, incurring up to K numerical releases even though the trigger requires only a binary threshold test: retain the current hypothesis until the first positive answer initiates an update. The selector admits the same interface. Because $\mathcal{F}$ and s_1 are public, we order candidates by decreasing $V_{f,1}(s_1)$, reject each candidate receiving a negative answer, and select the first candidate receiving a positive answer. Both procedures therefore produce a sequence of negative answers followed by at most one positive answer, exactly the structure handled by AboveThreshold.

4.1 THE ABOVETHRESHOLD PRIMITIVE

AboveThreshold is a standard sparse-vector primitive (Dwork & Roth, 2014) for processing adaptively chosen scalar queries against a public threshold without releasing their numerical values. It returns $\perp$ until a noisy query first crosses the noisy threshold, then returns $\top$ and halts.

Algorithm 1 One-positive AboveThreshold

Input: database D, query stream $q_1, q_2, \ldots$, threshold T, sensitivity Δ, and privacy level ε_0.
Draw $Z_0 \sim \mathrm{Lap}(2\Delta/\varepsilon_0)$ and set $\widetilde{T} \leftarrow T + Z_0$.
for each submitted query q_i **do**
  Draw $Z_i \sim \mathrm{Lap}(4\Delta/\varepsilon_0)$.
  if $q_i(D) + Z_i \geq \widetilde{T}$ **then**
    Output $\top$ and halt. // Positive answer.
  else
    Output $\perp$. // Negative answer.
  end if
end for

We have the following privacy and utility guarantees of AboveThreshold. For privacy, one complete invocation of Algorithm 1 is $(\varepsilon_0, 0)$-DP and its privacy cost is *independent* of the number of preceding $\perp$ answers. For utility, its accuracy holds with high probability because Laplace noise is unbounded. Suppose that an invocation processes at most n queries. For any $\alpha \in (0, 1)$, a union bound on the Laplace tails gives

$$\mathbb{P}\left(|Z_0| \leq \omega_{\text{single}} \text{ and } \max_{i\in[n]} |Z_i| \leq \omega_{\text{single}}\right) \geq 1 - \alpha, \qquad \omega_{\text{single}} := \frac{4\Delta}{\varepsilon_0} \log\left(\frac{n+1}{\alpha}\right).$$

On this event, simultaneously for every processed query,

$$\top \Rightarrow q_i(D) \geq T - 2\omega_{\text{single}}, \qquad \perp \Rightarrow q_i(D) < T + 2\omega_{\text{single}}, \qquad q_i(D) \geq T + 2\omega_{\text{single}} \Rightarrow \top. \tag{5}$$

Thus, AboveThreshold incurs an ambiguity margin of order $\Delta \log(n/\alpha)/\varepsilon_0$, which grows only logarithmically with the number of submitted queries. This logarithmic utility dependence—together with a privacy cost that does not depend on n—is the key advantage over privately releasing every numerical score. Appendix C gives its formal privacy and accuracy proofs for completeness.

Algorithm 2 `Priv-LazyGOLF`

Input: $\mathcal{F}, K, H$ and $\widetilde{\beta}_{\mathsf{fit}}, \widetilde{\beta}_{\mathsf{switch}}, M, \Delta, \varepsilon_0$, where $M < K$.
Publicly order $\mathcal{F} = \{f^{(1)}, \dots, f^{(N)}\}$ so that $V_{f^{(1)},1}(s_1) \geq \cdots \geq V_{f^{(N)},1}(s_1)$, with fixed tie-breaking.
Set $f \leftarrow f^{(1)}$ and $m \leftarrow 0$. `// Initialization.`
Start a trigger `AboveThreshold` call with threshold $\widetilde{\beta}_{\mathsf{switch}}$, sensitivity Δ, and privacy level ε_0.
for $k = 1, \dots, K$ **do**
  Execute π_f in episode k and add the observed trajectory to the data.
  if $m < M$ **then** `// Make no further private calls after` M `positive triggers.`
    Submit $\widehat{\mathcal{E}}_k(f)$ to the current trigger call.
    if the trigger returns $\top$ **then**
      Set $m \leftarrow m + 1$.
      if $m < M$ and $k < K$ **then**
        Start selector `AboveThreshold` with $(-\widetilde{\beta}_{\mathsf{fit}}, \Delta, \varepsilon_0)$.
        Submit $-\widehat{\mathcal{E}}_k(f^{(1)}), \dots, -\widehat{\mathcal{E}}_k(f^{(N)})$ in order until the first $\top$.
        If the first $\top$ occurs at index j, set $f \leftarrow f^{(j)}$; otherwise set $f \leftarrow f^{(1)}$.
        Start fresh trigger `AboveThreshold` for the new f with $(\widetilde{\beta}_{\mathsf{switch}}, \Delta, \varepsilon_0)$.
      end if
    end if
  end if
end for

4.2 Private Trigger and Selector via AboveThreshold

We now privatize the trigger and selector using two separate instances of `AboveThreshold`. To account for the ambiguity margin in (5), we replace the non-private thresholds β_{switch} and β_{fit} with their privacy-inflated counterparts $\widetilde{\beta}_{\mathsf{switch}}$ and $\widetilde{\beta}_{\mathsf{fit}}$, respectively.

Private trigger. Each epoch begins when the learner fixes a hypothesis f and starts a trigger instance of `AboveThreshold` with threshold $\widetilde{\beta}_{\mathsf{switch}}$. After every episode k in that epoch, the learner submits $\widehat{\mathcal{E}}_k(f)$. An answer $\perp$ retains f and continues the epoch, whereas the first $\top$ ends the epoch and invokes the private selector below. Thus, a single `AboveThreshold` invocation privatizes all trigger decisions in an epoch, regardless of how many episodes the epoch contains.

Private selector. Recall that the selector in (4) chooses the highest-value hypothesis whose excess loss is below the fitting threshold. To privatize this constrained maximization, after a positive trigger at episode k, the learner orders $\mathcal{F}$ by decreasing $V_{f,1}(s_1)$ and submits the queries $-\widehat{\mathcal{E}}_k(f)$ to `AboveThreshold` with threshold $-\widetilde{\beta}_{\mathsf{fit}}$. A response $\top$ means that the current candidate passes this noisy threshold test. The learner stops the scan at the first $\top$ and selects that candidate, which is the highest-value candidate accepted by the test. The ordering is public and incurs no privacy loss. Moreover, on the high-probability good event, $Q^\star$ receives $\top$, so the selector stops no later than $Q^\star$ in the ordering and hence returns an *optimistic* hypothesis, which is the key for our analysis.

4.3 The Private LazyGOLF Algorithm

Algorithm 2 privatizes both the trigger and selector of `LazyGOLF` using `AboveThreshold`. It publicly orders $\mathcal{F}$, sets $f = f^{(1)}$, and starts a trigger. In episode k, the learner executes π_f, records the trajectory, and submits $\widehat{\mathcal{E}}_k(f)$. An answer $\perp$ retains f. After an answer $\top$, if $m < M$ and $k < K$, the selector chooses the first accepted hypothesis in the public ordering (defaulting to $f^{(1)}$) and starts a fresh trigger. No private calls follow the M-th positive trigger. For the analysis, we write f^k for the hypothesis f executed in episode k and $\pi^k := \pi_{f^k}$.

Remark 4.2 (Why a fixed cap M is necessary)**.** The logarithmic switching bound holds only with high probability, so it cannot by itself justify privacy composition on every realization path. We therefore fix $M < K$ before the interaction and make no further private calls after the M-th positive

trigger. The algorithm makes at most $L := 2M - 1$ `AboveThreshold` calls.[2] This deterministic bound makes privacy composition valid unconditionally. Meanwhile, with our choice of M, the cap is inactive with high probability and does not affect the regret bound.

Remark 4.3 (Information-theoretic algorithm)**.** Our goal is statistical: can JDP preserve the non-private $\widetilde{O}(\sqrt{K})$ dependence? Following the finite-class setting of He & Zhou (2026), we permit exhaustive enumeration and set computation aside, because failure in this special finite case would rule out a general result. Fortunately, Theorem 5.1 below answers it affirmatively.

5 Theoretical Guarantees

We present the regret and privacy guarantee of Algorithm 2, whose parameter choices are given below. Fix $0 < \varepsilon \leq 1$ and $0 < \alpha, \delta < 1/2$, and recall that $N = |\mathcal{F}|$. Define $d := 1 \vee \dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, K^{-1/2})$, choose a public, data-independent certified $d_{\mathsf{max}} \geq d$, and let $\Lambda_K := \lceil \log_2(2K) \rceil$. Uniform concentration over stages, episode intervals, and pairs of consecutive-stage functions gives the fitting threshold

$$\beta_{\mathsf{fit}} = c_0 \left[\log\left(\frac{24HK^2 \left(\max_{h\in[H]} |\mathcal{F}_h|\right)^2}{\alpha} \right) + 1 \right], \tag{6}$$

where c_0 is a sufficiently large universal constant. The score $\widehat{\mathcal{E}}_t(f)$ has trajectory-replacement sensitivity $\Delta = 8$. Set $M := \lceil 64 d_{\mathsf{max}} H \Lambda_K \rceil$, $L := 2M - 1$, and $\varepsilon_0 := \varepsilon / [4\sqrt{2L \log(2/\delta)}]$. Each `AboveThreshold` invocation contains at most K trigger queries or N selector queries. A union bound over all L invocations gives the uniform margin $\omega := (4\Delta/\varepsilon_0) \log(3L(K + N + 1)/\alpha)$. Finally, set $\widetilde{\beta}_{\mathsf{fit}} := \beta_{\mathsf{fit}} + 2\omega$ and $\widetilde{\beta}_{\mathsf{switch}} := 16(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)$.

Theorem 5.1. *Under Assumption 2.1, with the stated parameter choices above, Algorithm 2 is (ε, δ)-JDP. Moreover, with probability at least $1 - \alpha$,*

$$\mathsf{Reg}(K) = \widetilde{O}\left(H \sqrt{dK \left(1 + \log \max_{h\in[H]} |\mathcal{F}_h|\right)} + d^{1/2} d_{\mathsf{max}}^{1/4} H^{5/4} \sqrt{\frac{K(1 + \log N)}{\varepsilon}} \right). \tag{7}$$

Here, $\widetilde{O}$ suppresses logarithmic factors in $K, H, d_{\mathsf{max}}, 1/\alpha$, and $1/\delta$.

Theorem 5.1 closes the statistical-rate gap between private and non-private RL: `Priv-LazyGOLF` is JDP and achieves the canonical $\widetilde{O}(\sqrt{K})$ regret. Thus, for fixed privacy parameters, privacy adds a separate cost but does not worsen the dependence on the number of episodes.

Comparison with prior work. The closest general-class result, He & Zhou (2026), combines fixed-size batching with the exponential mechanism, achieving $\widetilde{O}(K^{3/5})$ regret with polynomially many private updates. Our adaptive, data-dependent rare-switching rule uses only $O(\log K)$ updates, reducing privacy composition and hence enabling $\widetilde{O}(\sqrt{K})$ regret. The two results rely on distinct, incomparable exploration conditions—coverability and point-mass Bellman–Eluder dimension; Section 7 discusses whether the $\sqrt{K}$ regret is attainable under coverability.

Remark 5.2 (Role of d_{max})**.** The public certified bound $d_{\mathsf{max}} \geq d$ sets the deterministic cap M required for unconditional privacy composition. Any valid upper bound suffices; $d_{\mathsf{max}} = d$ is sharpest, while conservative choices cost only a $d_{\mathsf{max}}^{1/4}$ factor up to logarithms.

5.1 Proof Sketch

Privacy. Each trigger and selector query has trajectory-replacement sensitivity $\Delta = 8$, and the deterministic cap permits at most $L = 2M - 1$ `AboveThreshold` calls in total. Each call is $(\varepsilon_0, 0)$-DP regardless of the number of preceding negative answers. Advanced adaptive composition therefore makes the public transcript (ε, δ)-DP, and the Billboard Lemma (Hsu et al., 2016) yields JDP.

[2] The algorithm starts one trigger call. Each of the first $M - 1$ positive triggers can start one selector and one fresh trigger, while the M-th positive trigger starts neither, giving $1 + 2(M - 1) = 2M - 1$ calls.

Regret. For an episode interval $I \subseteq [K]$, let $\widehat{\mathcal{E}}_{h,I}(f)$ denote the excess loss in (1)–(2) computed only on episodes in I, so that $\widehat{\mathcal{E}}_{h,t} = \widehat{\mathcal{E}}_{h,[t]}$. Uniformly over every stage, interval I, and $f \in \mathcal{F}$, the concentration bound gives, for $S_{h,I}(f) := \sum_{i \in I} e_h(f; s_h^i, a_h^i)^2$,

$$S_{h,I}(f) \lesssim \widehat{\mathcal{E}}_{h,I}(f) + \beta_{\mathsf{fit}}, \qquad \widehat{\mathcal{E}}_{h,I}(f) \lesssim S_{h,I}(f) + \beta_{\mathsf{fit}}. \tag{$\star$}$$

Let $\mathcal{G}$ be the event on which ($\star$) and all `AboveThreshold` accuracy bounds hold, and condition on $\mathcal{G}$.

Step 1: Optimism and the global residual budget. Let $\Gamma := \widetilde{\beta}_{\mathsf{fit}} + 2\omega = \beta_{\mathsf{fit}} + 4\omega$, so $\widetilde{\beta}_{\mathsf{switch}} = 16\Gamma$. Because $Q^\star$ has zero Bellman residual, `AboveThreshold` accuracy forces the ordered selector to stop no later than $Q^\star$. Thus every epoch starting at episode s begins with an optimistic hypothesis f satisfying $V_{f,1}(s_1) \geq V_1^\star(s_1)$ and $\widehat{\mathcal{E}}_{s-1}(f) \leq \Gamma$. At any later episode, the current hypothesis was either just selected or retained after a negative trigger, so its excess loss on all previous data is $O(\Gamma)$. The first inequality in ($\star$) therefore gives, for every stage, $\sum_{i<k} e_h(f^k; s_h^i, a_h^i)^2 \lesssim \Gamma$. By the bounded point-mass Bellman–Eluder dimension, this bound over past state–action pairs implies a cumulative residual bound along the on-policy trajectory. That is, for every prefix before the cap,

$$\sum_{k=1}^{n} e_h(f^k; s_h^k, a_h^k)^2 \lesssim d\Gamma\Lambda_K, \qquad h \in [H]. \tag{8}$$

Step 2: Positive triggers require large epochwise residuals. Suppose an epoch $[s,t]$ ends with a positive trigger. The hypothesis had score at most Γ when the epoch began, while the terminal score crosses a constant-factor larger threshold. By subadditivity, some stage h has large epoch excess loss. The second inequality in ($\star$) then implies a large sum of squared Bellman residuals over that epoch:

$$\widehat{\mathcal{E}}_{h,[s,t]}(f) \gtrsim \Gamma \quad \Longrightarrow \quad \sum_{i=s}^{t} e_h(f; s_h^i, a_h^i)^2 \gtrsim \Gamma.$$

Epochs assigned to a fixed stage are disjoint. Since each trigger contributes $\gtrsim \Gamma$ over its epoch, (8) permits only $O(d\Lambda_K)$ positive triggers per stage and $O(dH\Lambda_K) < M$ overall. If the cap were reached, the prefix ending at the M-th trigger would contradict this bound. Hence the cap is nonbinding on $\mathcal{G}$. Crucially, the privacy-inflated radius Γ cancels between the two bounds.

Step 3: From residual control to regret. Optimism and Bellman telescoping give

$$V_1^\star(s_1) - V_1^{\pi^k}(s_1) \leq \sum_{h=1}^{H} \mathbb{E}^{\pi^k}[e_h(f^k; s_h, a_h)].$$

A martingale argument replaces these conditional expectations by the observed signed residuals. Cauchy–Schwarz and (8) then imply

$$\mathsf{Reg}(K) \lesssim \sum_{h=1}^{H} \sqrt{K \sum_{k=1}^{K} e_h(f^k; s_h^k, a_h^k)^2} + \widetilde{O}(H\sqrt{K}) = \widetilde{O}\Big(H\sqrt{d\Gamma K}\Big).$$

The concentration, `AboveThreshold` accuracy, and martingale events hold jointly with probability at least $1-\alpha$. Substituting the definitions of $\Gamma, \beta_{\mathsf{fit}}, \omega$, and L gives (7). See Appendix D for full proof.

6 Extension to the ℓ_2-EC Framework

Bellman completeness with bounded point-mass BE dimension is one instance of the ℓ_2-Eluder Condition (EC) framework of Xiong et al. (2024), which replaces Bellman residuals with conditional means of a general vector-valued loss and abstracts the square-potential transfer (Step 1 in Section 5.1) and Bellman telescoping (Step 3) as two class-level conditions. Our private rare-switching template extends to *finite* EC classes with the same trigger–selector control flow; Appendix E gives formal definitions, parameters, and proofs; finiteness is essential because the private selector must enumerate $\mathcal{F}$.

Decomposable losses. Each $f \in \mathcal{F}$ specifies a policy π_f and a predicted value $V_{f,1}(s_1)$, both computable without private data, and $\mathcal{G} \supseteq \mathcal{F}$ is a finite auxiliary class. At stage h of episode i, write

$\eta_h^i = (s_h^i, a_h^i)$ and $\zeta_h^i = (r_h^i, s_{h+1}^i)$. The framework uses a vector-valued loss $\ell_{h,f'}(\zeta, \eta, f, g)$ with $\|\ell_{h,f'}\|_2 \le R$. The loss is *decomposable* (Definition E.1) if a map $\mathfrak{T} : \mathcal{F} \to \mathcal{G}$ separates signal from noise: the loss evaluated at $g = \mathfrak{T}(f)$ is the noise term (zero conditional mean), while the conditional mean $\mu_{i,h}(f)$ of the loss at $g = f$ plays the role of the Bellman residual $e_h(f; s_h^i, a_h^i)$. An anchor $f^\star \in \mathcal{F}$ with $\mathfrak{T}(f^\star) = f^\star$ and $V_{f^\star,1}(s_1) = V_1^\star(s_1)$ is required.

EC class. An ℓ_2-type EC class (Definition E.2) has parameters $d, \kappa \ge 1$ and satisfies two conditions: (i) the ℓ_2-*Eluder condition* (cf. Step 1)—if $\sum_{i<k} \|\mu_{i,h}(f^k)\|_2^2 \le \Gamma$ for all $k \le n$, then $\sum_{i \le n} \|\mu_{i,h}(f^i)\|_2^2 \le d\Gamma \log(en)$; (ii) κ-*dominance* (cf. Step 3), bounding the optimism gap $V_{f,1}(s_1) - V_1^{\pi_f}(s_1)$ by κ times the expected stagewise loss norms.

Private algorithm and guarantee. `Priv-EC-RS` (Algorithm 3 in Appendix E) follows the same trigger–selector structure as Algorithm 2, replacing the Bellman excess $\widehat{\mathcal{E}}_t(f)$ by the empirical EC excess loss

$$\widehat{\mathcal{E}}_t^{\mathrm{EC}}(f) = \max_{h \in [H]} \Big[\sum_{i=1}^{t} \big\|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f, f)\big\|_2^2 - \min_{g \in \mathcal{G}} \sum_{i=1}^{t} \big\|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f, g)\big\|_2^2 \Big],$$

whose trajectory-replacement sensitivity is $2R^2$. As in Section 5, the switching cap is set from a public certified bound $d_{\mathsf{max}}^{\mathrm{EC}} \ge d$.

Theorem 6.1. *Under Definitions E.1 and E.2, with the parameter choices in Appendix E, Algorithm 3 is (ε, δ)-JDP. Moreover, with probability at least $1-\alpha$, it makes at most $\frac{9}{2} dH \log(eK)$ policy switches and*

$$\mathsf{Reg}(K) = \widetilde{O}\left(\kappa RH \sqrt{dK(1 + \log(NG))} + \kappa R d^{1/2} \big(d_{\mathsf{max}}^{\mathrm{EC}}\big)^{1/4} H^{5/4} \sqrt{\frac{K(1+\log N)}{\varepsilon}} \right), \quad (9)$$

where $N = |\mathcal{F}|$, $G = |\mathcal{G}|$, and $\widetilde{O}$ suppresses logarithmic factors in $K, H, d_{\mathsf{max}}^{\mathrm{EC}}, 1/\alpha$, and $1/\delta$.

Theorem 6.1 retains both JDP and the $\sqrt{K}$ dependence. For product classes, its first term exceeds that of Theorem 5.1 by $\sqrt{H}$ because the EC analysis applies a union bound over all stages and intervals jointly, whereas the Bellman analysis exploits the stagewise product structure of $\mathcal{F}$. Appendix E also shows how Theorem 5.1 arises as the Bellman-regression instance of the same proof.

7 Concluding Remarks

Multiplicative vs. additive privacy cost. In Theorems 5.1 and 6.1, privacy enters multiplicatively: the `AboveThreshold` noise adds $\widetilde{O}(\sqrt{L}/\varepsilon)$ to the squared-error radius Γ (Section 5.1), and the square root in the regret bound converts this to the $\sqrt{K/\varepsilon}$ term. Additive privacy costs are known in the tabular setting (Qiao & Wang, 2023), but every existing JDP guarantee under function approximation—linear (Zhou, 2022; Luyo et al., 2021) and general (He & Zhou, 2026)—incurs a multiplicative dependence. Whether an additive cost is achievable beyond the tabular setting remains open.

Towards oracle efficiency and infinite classes. The selector's linear scan over $\mathcal{F}$ is the sole computational bottleneck. A natural idea is to replace it with a noisy binary search on a public value-ordered tree, where each node issues one private scalar query to locate a nearly feasible, optimistic hypothesis, reducing oracle calls per switch to $O(\log N)$. Formalizing this requires a constrained range-feasibility oracle that is stronger than standard ERM; whether low Bellman–Eluder dimension alone implies such an oracle remains open, as does extending the approach to infinite classes via implicit covers.

Can coverability support private rare switching? We conjecture that $O(\log K)$ deterministic policy switching and $\sqrt{K}$ regret are incompatible under finite coverability. This would show that the gap between our results (which assume low Bellman–Eluder dimension) and those relying on coverability (He & Zhou, 2026) is structural rather than an artifact of our analysis (Appendix A). The reason is that coverability bounds visit frequencies but not the number of deterministic policies needed to discover rewarding actions: behind a rare entrance, arbitrarily many candidate branches coexist at constant coverability, yet each deterministic policy reaches only one. Stochastic policies escape this obstruction and offer a promising route to private $\sqrt{K}$ regret under coverability.

AI USE STATEMENT

Generative AI tools were used to help formulate and check mathematical claims, assist with proof writing and verification, organize the manuscript, draft and edit prose, and format LaTeX. No datasets, experiments, or empirical results were generated using AI. The authors independently verified all AI-assisted mathematical arguments, proofs, citations, and text and take responsibility for the final content.

ETHICS STATEMENT

This is a theoretical privacy paper and uses no human-subject data. The proposed guarantees depend on the stated unique-user neighboring relation and causal interface; deploying the method outside that interface may not protect users. Differential privacy limits information leakage but does not by itself address fairness, safety, or the quality of learned actions.

REPRODUCIBILITY STATEMENT

All assumptions, algorithmic parameters, sensitivity calculations, and theorem statements appear in the main paper. The appendices provide complete proofs of the statistical, privacy, switching, and regret claims, including constants sufficient to reproduce the stated rates.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Outcome-based online reinforcement learning: Algorithms and fundamental limits. *arXiv preprint arXiv:2505.20268*, 2025. URL https://arxiv.org/abs/2505.20268.

Sayak Ray Chowdhury and Xingyu Zhou. Differentially private regret minimization in episodic Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6375–6383, 2022.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in RL. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2826–2836, 2021.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9 of *Foundations and Trends in Theoretical Computer Science*. Now Publishers, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284, 2006.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 51–60, 2010.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Yi He and Xingyu Zhou. Towards differentially private reinforcement learning with general function approximation. *arXiv preprint arXiv:2605.07049*, 2026. URL https://arxiv.org/abs/2605.07049.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1704–1713, 2017.

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

Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, pp. 403–410, 2014.

Dingwen Kong, Ruslan Salakhutdinov, Ruosong Wang, and Lin F. Yang. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

Paul Luyo, Evrard Garcelon, Alessandro Lazaric, and Matteo Pirotta. Differentially private exploration in reinforcement learning with linear representation. *arXiv preprint arXiv:2112.01585*, 2021.

Dung Daniel T. Ngo, Giuseppe Vietri, and Zhiwei Steven Wu. Improved regret for differentially private exploration in linear MDP. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16529–16552, 2022.

Dan Qiao and Yu-Xiang Wang. Near-optimal differentially private reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9914–9940, 2023.

Dan Qiao, Ming Yin, and Yu-Xiang Wang. Logarithmic switching cost in reinforcement learning beyond linear MDPs. *arXiv preprint arXiv:2302.12456*, 2023.

Sharan Sahu. Towards optimal differentially private regret bounds in linear MDPs. *arXiv preprint arXiv:2504.09339*, 2025. URL https://arxiv.org/abs/2504.09339.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. In *Advances in Neural Information Processing Systems*, pp. 4296–4306, 2018.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 2898–2933, 2019.

Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, and Zhiwei Steven Wu. Private reinforcement learning with PAC and regret guarantees. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9754–9764, 2020.

Ruosong Wang, Russ R. Salakhutdinov, and Lin F. Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded Eluder dimension. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6123–6135, 2020.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Nuoya Xiong, Zhaoran Wang, and Zhuoran Yang. A general framework for sequential decision-making under adaptivity constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54792–54830, 2024. URL https://proceedings.mlr.press/v235/xiong24d.html.

Heyang Zhao, Jiafan He, and Quanquan Gu. A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. GEC: A unified framework for interactive decision making in MDP, POMDP, and beyond. *arXiv preprint arXiv:2211.01962*, 2022.

Xingyu Zhou. Differentially private reinforcement learning with linear function approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–27, 2022.

A Bellman–Eluder Dimension versus Coverability

We compare the point-mass Bellman–Eluder dimension in Definition 2.2 with the state–action coverability coefficient C_{cov} of Xie et al. (2023), evaluated on the greedy-policy class induced by $\mathcal{F}$. Although Xie et al. (2023, Proposition 4) give a related separation for Q-type Bellman–Eluder dimension, the following one-step constructions establish both directions directly for the point-mass notion used in this paper.

Proposition A.1 (Incomparability with coverability)**.** *Even for one-state, one-step MDPs and finite Bellman-complete classes, neither point-mass Bellman–Eluder dimension nor coverability bounds the other. More precisely:*

1. *for every n, there is an instance with $C_{\mathsf{cov}} = 1$ but $\dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, \eta) \geq n$ for every $\eta \leq 1/4$;*
2. *for every N, there is an instance with $C_{\mathsf{cov}} = N$ but $\dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, \eta) = O(\log(1+\eta^{-2}))$, independently of N.*

Proof. Fix a public deterministic rule for breaking ties. For the first direction, let $H = 1$, let the single state have actions $\{a_0, a_1, \dots, a_n\}$, and set the reward identically to zero. Let $\mathcal{F}_1 = \{0, f^1, \dots, f^n\}$, where

$$f^j(a_0) = 1, \qquad f^j(a_j) = \tfrac{1}{2}, \qquad f^j(a) = 0 \quad \text{otherwise.}$$

The class is Bellman complete because $\mathcal{T}_1 0 = 0 \in \mathcal{F}_1$. Every induced greedy policy selects a_0 (with the tie-breaking rule choosing a_0 for the zero function), so $C_{\mathsf{cov}} = 1$. On the other hand, the Bellman residuals are the functions themselves. The sequence $a_1, \dots, a_n$ is $1/4$-independent: at step j, the witness f^j vanishes on all preceding actions and equals $1/2$ at a_j. Hence $\dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, \eta) \geq n$ for every $\eta \leq 1/4$.

For the reverse direction, let the actions be $\{a_1, \dots, a_N\}$ and choose distinct unit vectors $u_1, \dots, u_N \in \mathbb{R}^2$. Set the reward to $1/2$ and let $\mathcal{F}_1 = \{f^0, f^1, \dots, f^N\}$, where

$$f^0(a_i) = \tfrac{1}{2}, \qquad f^j(a_i) = \tfrac{1}{2} + \tfrac{1}{2}\langle u_j, u_i \rangle.$$

Again the class is Bellman complete because $\mathcal{T}_1 0 = f^0$. Moreover, a_j is the unique greedy action of f^j, so the induced policy class contains all N point-mass action policies. Therefore

$$C_{\mathsf{cov}} = \min_{\mu \in \Delta(\mathcal{A})} \max_{j \in [N]} \frac{1}{\mu(a_j)} = N.$$

The residual class is contained in the two-dimensional linear class $a_i \mapsto \langle \theta, u_i \rangle$ with $\|\theta\|_2 \leq 1/2$. For completeness, consider an independent sequence $b_1, \dots, b_m$ at common scale $\eta' \geq \eta$, write $x_t = u_j$ when $b_t = a_j$, and set $V_0 = 4(\eta')^2 I_2$ and $V_t = V_0 + \sum_{i=1}^t x_i x_i^\top$. Each witness θ_t satisfies

$$\theta_t^\top V_{t-1} \theta_t \leq 2(\eta')^2, \qquad \left|\theta_t^\top x_t\right| > \eta',$$

where the first bound uses $\|\theta_t\|_2 \leq 1/2$ and $\sum_{i<t} (\theta_t^\top x_i)^2 \leq (\eta')^2$. By Cauchy–Schwarz in the V_{t-1}-norm,

$$(\eta')^2 < (\theta_t^\top x_t)^2 \leq (\theta_t^\top V_{t-1} \theta_t)(x_t^\top V_{t-1}^{-1} x_t) \leq 2(\eta')^2\, x_t^\top V_{t-1}^{-1} x_t,$$

so $x_t^\top V_{t-1}^{-1} x_t > 1/2$. The matrix determinant lemma and the trace–determinant inequality then give

$$(3/2)^m < \frac{\det(V_m)}{\det(V_0)} \leq \left(1 + \frac{m}{8(\eta')^2}\right)^2$$

for a sequence of length m. If $\eta' \geq 1/2$, no witness exists because $|\langle \theta, u_i \rangle| \leq 1/2$ for every admissible θ and i, so $m = 0$. Otherwise, using $1 + m/(8(\eta')^2) \leq (1+m)(1+(\eta')^{-2})$,

$$m \log(3/2) < 2\log(1+m) + 2\log\bigl(1 + (\eta')^{-2}\bigr).$$

Since $2\log(1+m) \leq \frac{1}{2} m \log(3/2) + c$ for a universal constant c, and $\log(1+(\eta')^{-2}) \geq \log 5$, this yields $m = O(\log(1+(\eta')^{-2}))$, which is $O(\log(1+\eta^{-2}))$ because $\eta' \geq \eta$. This is the standard linear Eluder-dimension bound (Jin et al., 2021), with no dependence on N. □

B Non-Private Statistical Ingredients

B.1 Bellman Completeness Implies Realizability

Lemma B.1. *Under Assumption 2.1, $Q^\star \in \mathcal{F}$ and $|e_h(f; s, a)| \le 1$ for every f, h, s, a.*

Proof. Set $Q^\star_{H+1} = 0 \in \mathcal{F}_{H+1}$. If $Q^\star_{h+1} \in \mathcal{F}_{h+1}$, Bellman completeness gives $\mathcal{T}_h Q^\star_{h+1} \in \mathcal{F}_h$, while Bellman optimality gives $Q^\star_h = \mathcal{T}_h Q^\star_{h+1}$. Backward induction proves $Q^\star \in \prod_h \mathcal{F}_h$. Both f_h and $\mathcal{T}_h f_{h+1}$ lie in $[0, 1]$: the former by definition of $\mathcal{F}_h$ and the latter by completeness. Hence the residual is in $[-1, 1]$. □

B.2 Uniform Convergence

For the proofs, we extend the prefix notation to an interval $I \subseteq [K]$ by setting

$$\widehat{L}_{h,I}(u; q) = \sum_{i \in I} \big(u(s^i_h, a^i_h) - r^i_h - V_q(s^i_{h+1})\big)^2,$$
$$\widehat{\mathcal{E}}_{h,I}(f) = \widehat{L}_{h,I}(f_h; f_{h+1}) - \min_{u \in \mathcal{F}_h} \widehat{L}_{h,I}(u; f_{h+1}).$$

Lemma B.2 (Uniform excess-loss concentration)**.** *Under Assumption 2.1, for β_{fit} in (6), a sufficiently large universal constant c_0 gives an event $\mathcal{G}_{\mathsf{stat}}$ of probability at least $1-\alpha/3$ on which, simultaneously for every $h \in [H]$, interval $I \subseteq [K]$, and $f \in \mathcal{F}$,*

$$\frac{3}{4} \sum_{i \in I} e_h(f; s^i_h, a^i_h)^2 - \beta_{\mathsf{fit}} \le \widehat{\mathcal{E}}_{h,I}(f) \le \frac{5}{4} \sum_{i \in I} e_h(f; s^i_h, a^i_h)^2 + 2\beta_{\mathsf{fit}}. \tag{10}$$

On the same event, $\widehat{\mathcal{E}}_t(Q^\star) \le \beta_{\mathsf{fit}}$ for every $t \in [K]$.

Proof. Fix h, an interval I, $q \in \mathcal{F}_{h+1}$, and $u \in \mathcal{F}_h$. Let

$$d_i = u(s^i_h, a^i_h) - [\mathcal{T}_h q](s^i_h, a^i_h), \qquad \eta_i = [\mathcal{T}_h q](s^i_h, a^i_h) - r^i_h - V_q(s^i_{h+1}).$$

Let $\mathcal{H}_{i,h}$ be the sigma-field immediately before (r^i_h, s^i_{h+1}) is observed. It contains the preceding trajectories, the public transcript, all algorithmic privacy randomness drawn by that time, and the current state–action pair. Thus d_i is $\mathcal{H}_{i,h}$-measurable and $\mathbb{E}[\eta_i \mid \mathcal{H}_{i,h}] = 0$. Expanding squares gives

$$\widehat{L}_{h,I}(u; q) - \widehat{L}_{h,I}(\mathcal{T}_h q; q) = \sum_{i \in I} d_i^2 + 2 \sum_{i \in I} d_i \eta_i.$$

Set $X_i := 2 d_i \eta_i$. The reward and value bounds give $|X_i| \le 4$ and $\mathbb{E}[X_i^2 \mid \mathcal{H}_{i,h}] \le 16 d_i^2$. Applying Freedman's inequality (cf. Lemma 5 of Chen et al. (2025)) to the masked martingale differences $\{\mathbf{1}\{i \in I\} X_i\}_{i=1}^K$ and their negatives, as in their proof of Lemma 6, gives, with probability at least $1 - 2\rho$,

$$\left| \sum_{i \in I} X_i \right| \le \lambda \sum_{i \in I} \mathbb{E}[X_i^2 \mid \mathcal{H}_{i,h}] + \frac{\log(1/\rho)}{\lambda} \le 16\lambda \sum_{i \in I} d_i^2 + \frac{\log(1/\rho)}{\lambda}.$$

Taking $\lambda = 1/64$ and $\rho = \alpha/[6HK^2(\max_h |\mathcal{F}_h|)^2]$, we obtain

$$\left| \widehat{L}_{h,I}(u; q) - \widehat{L}_{h,I}(\mathcal{T}_h q; q) - \sum_{i \in I} d_i^2 \right| \le \frac{1}{4} \sum_{i \in I} d_i^2 + \beta_{\mathsf{fit}} \tag{11}$$

simultaneously over all tuples (h, I, u, q) with probability at least $1 - \alpha/3$, since there are at most $HK^2(\max_h |\mathcal{F}_h|)^2$ such tuples and $c_0 \ge 64$ in (6).

Now fix h, I, f. Bellman completeness ensures that $\mathcal{T}_h f_{h+1} \in \mathcal{F}_h$. Define

$$X := \widehat{L}_{h,I}(f_h; f_{h+1}) - \widehat{L}_{h,I}(\mathcal{T}_h f_{h+1}; f_{h+1}), \qquad G := \widehat{L}_{h,I}(\mathcal{T}_h f_{h+1}; f_{h+1}) - \min_{u \in \mathcal{F}_h} \widehat{L}_{h,I}(u; f_{h+1}),$$

and let $S := \sum_{i \in I} e_h(f; s_h^i, a_h^i)^2$. Then $G \geq 0$, $\widehat{\mathcal{E}}_{h,I}(f) = X + G$, and (11) with $u = f_h, q = f_{h+1}$ gives

$$\frac{3}{4}S - \beta_{\mathsf{fit}} \leq X \leq \frac{5}{4}S + \beta_{\mathsf{fit}}.$$

Let $\widehat{u} \in \operatorname{argmin}_{u \in \mathcal{F}_h} \widehat{L}_{h,I}(u; f_{h+1})$ be the empirical minimizer. Applying the lower bound in (11) to $\widehat{u}$ gives

$$\widehat{L}_{h,I}(\widehat{u}; f_{h+1}) - \widehat{L}_{h,I}(\mathcal{T}_h f_{h+1}; f_{h+1}) \geq -\beta_{\mathsf{fit}},$$

and hence $G \leq \beta_{\mathsf{fit}}$. Combining these inequalities proves (10). For $f = Q^\star$, the Bellman equation gives $S = X = 0$, so $\widehat{\mathcal{E}}_{h,I}(Q^\star) = G \leq \beta_{\mathsf{fit}}$. Taking $I = [t]$ and the maximum over h proves the final claim. □

B.3 Bellman–Eluder Square Potential

For completeness, we record the standard squared-residual argument based on the Bellman–Eluder dimension, cf. Jin et al. (2021).

Lemma B.3 (Squared-residual potential)**.** *Recall that $d = 1 \vee \dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, K^{-1/2})$. Fix a stage h and $n \leq K$. Suppose $f^1, \ldots, f^n \in \mathcal{F}$ and state–action pairs $(s_h^1, a_h^1), \ldots, (s_h^n, a_h^n)$ satisfy, for every $k \in [n]$ and some $\Gamma \geq 1$,*

$$\sum_{i<k} e_h(f^k; s_h^i, a_h^i)^2 \leq \Gamma.$$

Then

$$\sum_{k=1}^{n} e_h(f^k; s_h^k, a_h^k)^2 \leq 12 d \Gamma \Lambda_K.$$

Proof. Fix $u \geq K^{-1/2}$ and consider the indices k for which $\left|e_h(f^k; s_h^k, a_h^k)\right| > u$. Place these indices sequentially into buckets. If the current state–action pair is u-dependent on every existing bucket, then the witness $e_h(f^k; \cdot)$ contributes more than u^2 squared residual to each bucket; otherwise that pair can be added to a bucket on which it is u-independent. Since its total squared residual on all preceding pairs is at most Γ, at most Γ/u^2 buckets can reject the current pair. Thus, at most $\Gamma/u^2 + 1$ buckets are needed, and each bucket contains at most $\dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, u) \leq d$ indices. Therefore,

$$\sum_{k=1}^{n} \mathbf{1}\{\left|e_h(f^k; s_h^k, a_h^k)\right| > u\} \leq \left(\frac{\Gamma}{u^2} + 1\right) \dim_{\mathsf{BE}}(\mathcal{F}, D_\Delta, u). \tag{12}$$

Sort $v_k := \left|e_h(f^k; s_h^k, a_h^k)\right|$ in decreasing order. Values at most $K^{-1/2}$ contribute at most 1, and the first d larger values contribute at most d because Bellman residuals are bounded by 1. For every later rank $t > d$, apply (12) with a threshold approaching $v_{(t)}$ to obtain

$$v_{(t)}^2 \leq \frac{\Gamma d}{t - d}.$$

Summing over the ranks gives

$$\sum_{k=1}^{n} v_k^2 \leq 1 + d + \Gamma d \sum_{j=1}^{K} \frac{1}{j} \leq 12 d \Gamma \Lambda_K,$$

where the final inequality uses $d, \Gamma \geq 1$ and the definition of Λ_K. □

C AboveThreshold Privacy and Accuracy

For completeness, we provide privacy and accuracy proofs for the one-positive version of `AboveThreshold` used in Algorithm 2; see Dwork & Roth (2014) for the standard `AboveThreshold` mechanism and its connection to the Sparse Vector technique.

C.1 PRIVACY

Theorem C.1. *A complete invocation of Algorithm 1 is $(\varepsilon_0, 0)$-DP, even when queries are chosen adaptively from earlier released bits.*

Proof. Fix neighboring databases D, D' and a transcript $y_k = (\perp^{k-1}, \top)$. Along this common public prefix, the adaptive analyst selects the same query maps $q_1, \ldots, q_k$ in both worlds, although their evaluated values differ. Fix the analyst's data-independent randomness. Write $u = (z_1, \ldots, z_{k-1})$ for the earlier query-noise coordinates and define

$$g_D(u) = \max_{i<k}\{q_i(D) + z_i\}, \qquad g_{D'}(u) = \max_{i<k}\{q_i(D') + z_i\}.$$

For a fixed dummy variable u, the transcript event under D is

$$A_D(u) = \{(z_0, z_k) : g_D(u) < T + z_0 \le q_k(D) + z_k\}.$$

This use of the same u compares two integrands; it does not assert that two independent executions realize identical earlier noises.

Set

$$s(u) = g_{D'}(u) - g_D(u), \qquad c(u) = s(u) + q_k(D) - q_k(D').$$

Sensitivity and the Lipschitzness of the maximum give $|s(u)| \le \Delta$ and $|c(u)| \le 2\Delta$. The translation

$$(z_0, z_k) \mapsto (z_0 + s(u), z_k + c(u))$$

maps $A_D(u)$ bijectively to $A_{D'}(u)$. A Laplace density p_b satisfies $p_b(x) \le e^{|c|/b} p_b(x+c)$. Therefore the threshold density ratio is at most $e^{\varepsilon_0/2}$ and the final-query density ratio is at most $e^{\varepsilon_0/2}$. The earlier-noise density is unchanged and is integrated over u. Hence

$$\mathbb{P}_D(Y = y_k) \le e^{\varepsilon_0} \mathbb{P}_{D'}(Y = y_k).$$

The first-query-positive transcript is the same calculation without u. A finite all-negative transcript shifts only the threshold coordinate by $g_{D'}(u) - g_D(u)$ and costs at most $e^{\varepsilon_0/2}$. Swapping D, D' gives the reverse comparison. Summing singleton bounds over an arbitrary transcript event proves pure DP. □

C.2 ACCURACY

Lemma C.2 (Accuracy of `AboveThreshold`)**.** *An invocation of Algorithm 1 that processes at most n queries satisfies (5) simultaneously for all processed queries with probability at least $1 - \alpha$, where*

$$\omega_{\mathsf{single}} = \frac{4\Delta}{\varepsilon_0} \log\left(\frac{n+1}{\alpha}\right).$$

Proof. Let $\mathcal{E}$ be the event that $|Z_0| \le \omega_{\mathsf{single}}$ and $|Z_i| \le \omega_{\mathsf{single}}$ for every processed query. The Laplace tails and a union bound give

$$\mathbb{P}(\mathcal{E}^c) \le e^{-\omega_{\mathsf{single}}\varepsilon_0/(2\Delta)} + n e^{-\omega_{\mathsf{single}}\varepsilon_0/(4\Delta)} \le (n+1) e^{-\omega_{\mathsf{single}}\varepsilon_0/(4\Delta)} = \alpha.$$

On $\mathcal{E}$, an answer $\top$ implies $q_i(D) \ge T + Z_0 - Z_i \ge T - 2\omega_{\mathsf{single}}$, while $\perp$ implies $q_i(D) < T + Z_0 - Z_i \le T + 2\omega_{\mathsf{single}}$. Finally, $q_i(D) \ge T + 2\omega_{\mathsf{single}}$ implies $q_i(D) + Z_i \ge T + \omega_{\mathsf{single}} \ge T + Z_0$, forcing $\top$. □

D PROOF OF THEOREM 5.1

We prove the two conclusions of Theorem 5.1 separately. Section D.1 establishes JDP without conditioning on the utility events. For regret, Sections D.2–D.4 work on the statistical concentration and `AboveThreshold` accuracy events: we first establish optimism and historical residual control, then bound the number of positive triggers and remove the deterministic cap from the analysis, and finally convert the resulting residual bound into regret. Section D.5 combines these arguments and substitutes the parameter choices from Section 5.

D.1 PRIVACY

Lemma D.1 (Trajectory-replacement sensitivity). *Conditional on any public transcript prefix, $\widehat{\mathcal{E}}_t(f)$ and $-\widehat{\mathcal{E}}_t(f)$ have sensitivity at most 8 for every fixed query time and candidate.*

Proof. For $u \in \mathcal{F}_h$ and $q \in \mathcal{F}_{h+1}$, we have $u(s,a), V_q(s') \in [0,1]$, and the stagewise reward assumption gives $r_h \in [0,1]$ almost surely. Hence $u(s,a) - r_h - V_q(s') \in [-2,1]$, so every squared summand in (1) lies in $[0,4]$. Replacing one trajectory therefore changes $\widehat{L}_{h,t}(u;q)$ by at most 4 for every fixed h,t,u,q. The minimum over $u \in \mathcal{F}_h$ changes by at most the largest change of any candidate loss, which is also 4. Hence $\widehat{\mathcal{E}}_{h,t}(f)$ has sensitivity at most 8. Finally, $|\max_h A_h - \max_h B_h| \le \max_h |A_h - B_h|$, so taking the maximum over stages adds no factor of H; negating the query does not change its sensitivity. □

Proposition D.2 (Joint differential privacy). *Algorithm 2 is (ε,δ)-JDP.*

Proof. Fix an arbitrary public transcript prefix. The prefix determines the current hypothesis, query time, and next candidate examined by the selector. Lemma D.1 and Theorem C.1 imply that, conditional on this prefix, the next complete trigger or selector invocation is $(\varepsilon_0, 0)$-DP.

The cap permits at most $L = 2M - 1$ physical invocations on every execution path. We may pad any shorter execution with data-independent invocations. Advanced adaptive composition (Dwork et al., 2010) therefore gives $(\varepsilon_{\mathsf{comp}}, \delta)$-DP, where

$$\varepsilon_{\mathsf{comp}} = \sqrt{2L\log(1/\delta)}\,\varepsilon_0 + L\varepsilon_0(e^{\varepsilon_0} - 1).$$

The calibration of ε_0 uses the stronger $\log(2/\delta) \ge \log(1/\delta)$, so the first term is at most $\varepsilon/4$. Moreover, $\varepsilon_0 \le 1$, so $e^{\varepsilon_0} - 1 \le 2\varepsilon_0$ and

$$L\varepsilon_0(e^{\varepsilon_0} - 1) \le 2L\varepsilon_0^2 = \frac{\varepsilon^2}{16\log(2/\delta)} < \frac{\varepsilon}{16}.$$

Thus $\varepsilon_{\mathsf{comp}} < \varepsilon$, proving (ε,δ)-DP for the public transcript. The selected hypotheses, switching times, and final policy are post-processing of this transcript. Under the causal unique-user interface, every other user's actions depend only on the public transcript and that user's unchanged local state. Hence, the Billboard Lemma (Hsu et al., 2016) yields that Algorithm 2 is (ε,δ)-JDP. This argument holds on every execution path and does not condition on any utility event. □

D.2 THE ACCURACY EVENT AND PRIVATE OPTIMISM

Let $\mathcal{G}_{\mathsf{AT}}$ be the event that every threshold and query noise drawn by Algorithm 2 has magnitude at most ω. The algorithm draws at most $L(K+N+1)$ such variables, each with Laplace scale at most $4\Delta/\varepsilon_0$. The Laplace tail bound and the definition of ω give

$$\mathbb{P}(\mathcal{G}_{\mathsf{AT}}) \ge 1 - \alpha/3.$$

Lemma D.3 (Exact private optimism). *On $\mathcal{G}_{\mathsf{stat}} \cap \mathcal{G}_{\mathsf{AT}}$, every selector invoked after episode t returns $\widehat{f}$ satisfying*

$$V_{\widehat{f},1}(s_1) \ge V_1^\star(s_1), \qquad \widehat{\mathcal{E}}_t(\widehat{f}) \le \widetilde{\beta}_{\mathsf{fit}} + 2\omega.$$

The initial hypothesis $f^{(1)}$ satisfies the same value inequality and $\widehat{\mathcal{E}}_0(f^{(1)}) = 0$.

Proof. Lemma B.2 and the definition of $\widetilde{\beta}_{\mathsf{fit}}$ give

$$-\widehat{\mathcal{E}}_t(Q^\star) \ge -\beta_{\mathsf{fit}} = -\widetilde{\beta}_{\mathsf{fit}} + 2\omega.$$

For the selector query $q(f) = -\widehat{\mathcal{E}}_t(f)$ and threshold $T = -\widetilde{\beta}_{\mathsf{fit}}$, Lemma C.2 therefore forces a positive answer no later than the point at which the public scan reaches $Q^\star$. The selected hypothesis $\widehat{f}$ precedes $Q^\star$ in the decreasing-value order, so $V_{\widehat{f},1}(s_1) \ge V_1^\star(s_1)$. A positive answer for $\widehat{f}$ also implies

$$-\widehat{\mathcal{E}}_t(\widehat{f}) \ge -\widetilde{\beta}_{\mathsf{fit}} - 2\omega,$$

which is equivalent to the claimed excess-loss bound. Finally, $f^{(1)}$ maximizes the public initial-state value over a class containing $Q^\star$, and $\widehat{\mathcal{E}}_0(f^{(1)}) = 0$ by definition. □

D.3 Residual Energy and the Deterministic Cap

Let

$$\tau_M := \inf\left\{ t \in [K] : \sum_{k=1}^{t} \mathbf{1}\{\text{the trigger returns } \top \text{ after episode } k\} \geq M \right\},$$

with $\tau_M = K + 1$ if this event never occurs. Thus τ_M is the episode at whose completion the M-th positive trigger fires.

Lemma D.4 (Historical residual budget before the cap)**.** *On $\mathcal{G}_{\mathsf{stat}} \cap \mathcal{G}_{\mathsf{AT}}$, for every $k \leq \min\{K, \tau_M\}$ and stage h,*

$$\sum_{i<k} e_h(f^k; s_h^i, a_h^i)^2 \leq 24\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big).$$

Proof. If $k = 1$, then $\widehat{\mathcal{E}}_0(f^1) = 0$. If episode k is the first episode after a selector call, Lemma D.3 gives

$$\widehat{\mathcal{E}}_{k-1}(f^k) \leq \widetilde{\beta}_{\mathsf{fit}} + 2\omega.$$

Otherwise, the hypothesis was retained after a negative trigger at episode $k-1$. Lemma C.2 and the definitions of the inflated thresholds give

$$\widehat{\mathcal{E}}_{k-1}(f^k) < \widetilde{\beta}_{\mathsf{switch}} + 2\omega = 16\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big) + 2\omega \leq \frac{33}{2}\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big),$$

where the last inequality uses $2\omega \leq (\widetilde{\beta}_{\mathsf{fit}} + 2\omega)/2$. Thus, in all three cases, $\widehat{\mathcal{E}}_{k-1}(f^k) \leq \frac{33}{2}(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)$. Since $\widehat{\mathcal{E}}_{h,k-1}(f^k) \leq \widehat{\mathcal{E}}_{k-1}(f^k)$, the lower bound in (10) implies

$$\frac{3}{4}\sum_{i<k} e_h(f^k; s_h^i, a_h^i)^2 - \beta_{\mathsf{fit}} \leq \frac{33}{2}\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big).$$

Because $\beta_{\mathsf{fit}} \leq \widetilde{\beta}_{\mathsf{fit}} + 2\omega$, rearranging gives the result. □

Applying Lemma B.3 with $\Gamma = 24(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)$ gives, for every $n \leq \min\{K, \tau_M\}$ and $h \in [H]$,

$$\sum_{k=1}^{n} e_h(f^k; s_h^k, a_h^k)^2 \leq 288d\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big)\Lambda_K. \tag{13}$$

Lemma D.5 (Subadditivity of empirical excess loss)**.** *For disjoint episode intervals I and J and every fixed h, f,*

$$\widehat{\mathcal{E}}_{h,I\cup J}(f) \leq \widehat{\mathcal{E}}_{h,I}(f) + \widehat{\mathcal{E}}_{h,J}(f).$$

Proof. The empirical loss is additive over disjoint intervals. Moreover,

$$\min_{u\in\mathcal{F}_h}\big\{\widehat{L}_{h,I}(u; f_{h+1}) + \widehat{L}_{h,J}(u; f_{h+1})\big\} \geq \min_{u\in\mathcal{F}_h}\widehat{L}_{h,I}(u; f_{h+1}) + \min_{u\in\mathcal{F}_h}\widehat{L}_{h,J}(u; f_{h+1}).$$

Subtracting this inequality from the additive loss of f_h proves the claim. □

Lemma D.6 (Fresh energy at a positive trigger)**.** *Suppose an epoch $[s, t]$ ends in a positive trigger with $t \leq \tau_M$. For some stage h,*

$$\sum_{i=s}^{t} e_h(f; s_h^i, a_h^i)^2 \geq 9\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big),$$

where f is the epoch hypothesis.

Proof. The positive trigger and Lemma C.2 give

$$\widehat{\mathcal{E}}_t(f) \geq \widetilde{\beta}_{\mathsf{switch}} - 2\omega = 16\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big) - 2\omega.$$

Choose a stage h attaining the maximum in $\widehat{\mathcal{E}}_t(f)$. At the beginning of the epoch, Lemma D.3 gives $\widehat{\mathcal{E}}_{h,s-1}(f) \le \widetilde{\beta}_{\mathsf{fit}} + 2\omega$; for the initial epoch, this quantity is zero. Lemma D.5 therefore implies

$$\begin{aligned}\widehat{\mathcal{E}}_{h,[s,t]}(f) &\ge \widehat{\mathcal{E}}_{h,t}(f) - \widehat{\mathcal{E}}_{h,s-1}(f) \\ &\ge 15\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big) - 2\omega \\ &\ge 14\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big),\end{aligned}$$

where the last step uses $2\omega \le (\widetilde{\beta}_{\mathsf{fit}} + 2\omega)/2$. Applying the upper bound in (10) to the epoch gives

$$14\big(\widetilde{\beta}_{\mathsf{fit}} + 2\omega\big) \le \frac{5}{4}\sum_{i=s}^{t} e_h(f; s_h^i, a_h^i)^2 + 2\beta_{\mathsf{fit}}.$$

Since $\beta_{\mathsf{fit}} \le \widetilde{\beta}_{\mathsf{fit}} + 2\omega$, rearranging yields a lower bound of $48(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)/5$, which is larger than the claimed bound. □

Theorem D.7 (The cap does not bind)**.** *On $\mathcal{G}_{\mathsf{stat}} \cap \mathcal{G}_{\mathsf{AT}}$, Algorithm 2 has at most $32dH\Lambda_K < M$ positive triggers, and hence at most $32dH\Lambda_K$ policy switches. In particular, the cap is never reached, i.e., $\tau_M = K + 1$, and (13) holds with $n = K$.*

Proof. Consider any prefix ending at $n \le \min\{K, \tau_M\}$. Assign each positive trigger in this prefix to a stage supplied by Lemma D.6. The epoch intervals assigned to a fixed stage are disjoint, and the hypothesis remains fixed within each epoch. Thus, (13) bounds the number of triggers assigned to that stage by

$$\frac{288d(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)\Lambda_K}{9(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)} = 32d\Lambda_K.$$

Summing over the H stages gives at most $32dH\Lambda_K$ positive triggers in any such prefix. If $\tau_M \le K$, the prefix ending at τ_M contains exactly M positive triggers, while the preceding bound gives

$$M \le 32dH\Lambda_K < 64d_{\mathsf{max}}H\Lambda_K \le M,$$

a contradiction. Hence $\tau_M = K + 1$: the cap is not reached on $\mathcal{G}_{\mathsf{stat}} \cap \mathcal{G}_{\mathsf{AT}}$. Taking $n = K$ in (13) proves the full-horizon energy bound. Finally, a policy can change only after a positive trigger, so the same upper bound applies to the number of policy switches. □

D.4 Regret on the Good Event

Proposition D.8 (Regret on the good event)**.** *There is an event $\mathcal{G}_{\mathsf{Az}}$ with probability at least $1 - \alpha/3$ such that, on $\mathcal{G}_{\mathsf{stat}} \cap \mathcal{G}_{\mathsf{AT}} \cap \mathcal{G}_{\mathsf{Az}}$,*

$$\mathsf{Reg}(K) \le 12\sqrt{2}\,H\sqrt{d(\widetilde{\beta}_{\mathsf{fit}} + 2\omega)K\Lambda_K} + H\sqrt{2K\log(3H/\alpha)}. \tag{14}$$

Proof. For every f, Bellman telescoping gives

$$V_{f,1}(s_1) - V_1^{\pi_f}(s_1) = \sum_{h=1}^{H} \mathbb{E}^{\pi_f}[e_h(f; s_h, a_h)]. \tag{15}$$

Because f^k is fixed before episode k, the conditional expectation of $e_h(f^k; s_h^k, a_h^k)$ equals $\mathbb{E}^{\pi^k}[e_h(f^k; s_h, a_h)]$. The residual lies in $[-1, 1]$, so subtracting its conditional mean gives a martingale difference whose conditional range has length at most 2. The conditional-range form of Azuma–Hoeffding and a union bound over h give an event $\mathcal{G}_{\mathsf{Az}}$ of probability at least $1 - \alpha/3$ on which

$$\sum_{k=1}^{K} \mathbb{E}^{\pi^k}[e_h(f^k; s_h, a_h)] \le \sum_{k=1}^{K} e_h(f^k; s_h^k, a_h^k) + \sqrt{2K\log(3H/\alpha)}$$

simultaneously for all h. Combining this bound with exact optimism from Lemma D.3 and (15) gives

$$\mathsf{Reg}(K) \le \sum_{h=1}^{H}\left(\sum_{k=1}^{K} e_h(f^k; s_h^k, a_h^k) + \sqrt{2K\log(3H/\alpha)}\right).$$

Cauchy–Schwarz and Theorem D.7 yield

$$\sum_{k=1}^{K} e_h(f^k; s_h^k, a_h^k) \le \sqrt{K \sum_{k=1}^{K} e_h(f^k; s_h^k, a_h^k)^2} \le 12\sqrt{2}\sqrt{d(\widetilde{\beta}_{\mathsf{fit}} + 2\omega) K \Lambda_K}.$$

Summing over h proves (14). □

D.5 Completing the Proof

Proof of Theorem 5.1. The privacy conclusion follows from Proposition D.2 without conditioning on any of the events used below. For the regret conclusion, Lemma B.2, the definition of $\mathcal{G}_{\mathsf{AT}}$, and Proposition D.8 give

$$\mathbb{P}(\mathcal{G}_{\mathsf{stat}}) \ge 1 - \alpha/3, \qquad \mathbb{P}(\mathcal{G}_{\mathsf{AT}}) \ge 1 - \alpha/3, \qquad \mathbb{P}(\mathcal{G}_{\mathsf{Az}}) \ge 1 - \alpha/3.$$

Consequently, their intersection has probability at least $1 - \alpha$. On this intersection, Theorem D.7 ensures that the deterministic cap never binds, and Proposition D.8 gives (14).

It remains only to substitute the parameters. By definition, $\widetilde{\beta}_{\mathsf{fit}} + 2\omega = \beta_{\mathsf{fit}} + 4\omega$. Substituting the choices of ε_0 and ω yields

$$\widetilde{\beta}_{\mathsf{fit}} + 2\omega = \beta_{\mathsf{fit}} + \frac{64\Delta}{\varepsilon}\sqrt{2L\log(2/\delta)}\log\left(\frac{3L(K+N+1)}{\alpha}\right). \tag{16}$$

Since $L = 2M - 1 = O(d_{\mathsf{max}} H \Lambda_K)$, substituting (6) and (16) into (14) proves (7) and completes the proof. □

E Extension to the ℓ_2-EC Class

This section extends the preceding construction to the single-MDP ℓ_2-Eluder Condition (EC) class of Xiong et al. (2024). We keep the same online interaction, public initial state, trajectory-level neighboring relation, and one-positive `AboveThreshold` primitive. The only new ingredient is a bounded decomposable loss whose conditional mean plays the role of the Bellman residual. We state the extension for finite classes because the private selector requires a fixed public candidate order.

E.1 Single-MDP EC Setup

Each $f \in \mathcal{F}$ specifies a policy π_f and a predicted initial value $V_{f,1}(s_1) \in [0,1]$, both computable without private trajectories. Let $\mathcal{G}$ be a finite auxiliary class with $\mathcal{F} \subseteq \mathcal{G}$. In episode i and stage h, write

$$\eta_h^i = (s_h^i, a_h^i), \qquad \zeta_h^i = (r_h^i, s_{h+1}^i).$$

Thus η_h^i is available before the one-step outcome ζ_h^i. Let $\mathcal{H}_{i-1}$ be the episode filtration immediately before f^i is chosen. It is generated by all trajectories through episode $i-1$, the public transcript, and every internal random variable drawn by the trigger and selector mechanisms up to that time. The active hypothesis f^i is $\mathcal{H}_{i-1}$-measurable. Let $\mathcal{H}_{i,h}$ augment this filtration with the episode-i observations through η_h^i and any algorithmic randomness drawn before ζ_h^i is observed. We write $\mathbb{E}_{\zeta_h^i}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}_{i,h}]$.

For $h \in [H]$, $f', f \in \mathcal{F}$, and $g \in \mathcal{G}$, let

$$\ell_{h,f'}(\zeta, \eta, f, g) \in \mathbb{R}^p$$

be a loss with $\|\ell_{h,f'}(\zeta,\eta,f,g)\|_2 \le R$ almost surely. The dimension p need not enter the guarantee.

Definition E.1 (Decomposable loss)**.** The loss ℓ is decomposable if there is a map $\mathfrak{T} : \mathcal{F} \to \mathcal{G}$ such that, for every admissible argument,

$$\ell_{h,f'}(\zeta_h^i, \eta_h^i, f, g) - \mathbb{E}_{\zeta_h^i}\left[\ell_{h,f'}(\zeta_h^i, \eta_h^i, f, g)\right] = \ell_{h,f'}(\zeta_h^i, \eta_h^i, f, \mathfrak{T}(f)). \tag{17}$$

We additionally assume that there is $f^\star \in \mathcal{F}$ satisfying

$$V_{f^\star,1}(s_1) = V_1^\star(s_1), \qquad \mathfrak{T}(f^\star) = f^\star. \tag{18}$$

For a hypothesis f tested on episode i, define its conditional mean loss at stage h by

$$\mu_{i,h}(f) := \mathbb{E}_{\zeta_h^i}\left[\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f, f)\right].$$

Jensen's inequality gives $\|\mu_{i,h}(f)\|_2 \le R$. Following Definition 3.1 of Xiong et al. (2024), we use the following single-MDP, explicit-constant specialization of the ℓ_2-type EC framework.

Definition E.2 (Single-MDP ℓ_2-type EC class)**.** Fix the target failure probability α. The tuple $(M^\star, \mathcal{F}, \mathcal{G}, \ell, d, \kappa)$ is a single-MDP ℓ_2-type EC class if $d, \kappa \ge 1$ and the following hold.

1. *ℓ_2-type Eluder condition.* For every stage h, every realized interaction and algorithmic privacy randomness, every sequence $f^1, \dots, f^K \in \mathcal{F}$ with f^i measurable with respect to $\mathcal{H}_{i-1}$, every $n \le K$, and every $\Gamma \ge R^2$,

$$\sum_{i<k} \left\|\mu_{i,h}(f^k)\right\|_2^2 \le \Gamma \quad \text{for every } k \in [n], \tag{19}$$

   implies

$$\sum_{i=1}^{n} \left\|\mu_{i,h}(f^i)\right\|_2^2 \le d\Gamma \log(en). \tag{20}$$

2. *κ-dominance.* For every sequence $f^1, \dots, f^K \in \mathcal{F}$ with f^i measurable with respect to $\mathcal{H}_{i-1}$, with probability at least $1 - \alpha/4$, simultaneously for every $n \le K$,

$$\sum_{i=1}^{n} \left(V_{f^i,1}(s_1) - V_1^{\pi_{f^i}}(s_1)\right) \le \kappa \sum_{h=1}^{H} \sum_{i=1}^{n} \mathbb{E}\left[\left\|\mu_{i,h}(f^i)\right\|_2 \mid \mathcal{H}_{i-1}\right].$$

The first condition is pathwise and therefore applies to every realized prefix, including one selected by a stopping rule. We absorb the universal constant in the original EC definition into d.

Remark E.3 (Examples and finite-class scope)**.** The general ℓ_2-EC framework contains several useful model classes, including low D_Δ-Bellman–Eluder dimension, linear-mixture MDPs, and kernelized nonlinear regulators, under suitable decomposable losses (Xiong et al., 2024). Theorem 6.1 has a narrower scope because its private ordered selector assumes that both $\mathcal{F}$ and $\mathcal{G}$ are finite. It therefore applies directly to finite EC instances and to a finite parameter subclass of one of the preceding examples only when that subclass still satisfies Definition E.1, contains an exact realizability anchor $f^\star$, and obeys Definition E.2.

E.2 Empirical EC Score and Private Algorithm

For an episode interval $I \subseteq [K]$, define

$$\widehat{L}_{h,I}^{\mathrm{EC}}(f, g) := \sum_{i \in I} \left\|\ell_{h,f^i}(\zeta_h^i, \eta_h^i, f, g)\right\|_2^2, \tag{21}$$

$$\widehat{\mathcal{E}}_{h,I}^{\mathrm{EC}}(f) := \widehat{L}_{h,I}^{\mathrm{EC}}(f, f) - \min_{g \in \mathcal{G}} \widehat{L}_{h,I}^{\mathrm{EC}}(f, g), \tag{22}$$

$$\widehat{\mathcal{E}}_t^{\mathrm{EC}}(f) := \max_{h \in [H]} \widehat{\mathcal{E}}_{h,[t]}^{\mathrm{EC}}(f), \qquad \widehat{\mathcal{E}}_0^{\mathrm{EC}}(f) := 0.$$

These are the direct EC analogues of (1)–(2). The minimization class changes from a stagewise Bellman-regression class to $\mathcal{G}$, but the score remains a scalar empirical excess loss.

Fix $0 < \varepsilon \le 1$ and $0 < \alpha, \delta < 1/2$, let $N = |\mathcal{F}|$ and $G = |\mathcal{G}|$, and choose a public, data-independent certified $d_{\mathsf{max}}^{\mathrm{EC}} \ge d$. Set

$$\beta_{\mathrm{EC}} := 8R^2 \log\left(\frac{8HK^2NG}{\alpha}\right), \qquad \Delta_{\mathrm{EC}} := 2R^2,$$

$$M_{\mathrm{EC}} := \left\lceil 9 d_{\mathsf{max}}^{\mathrm{EC}} H \log(eK) \right\rceil, \qquad L_{\mathrm{EC}} := 2M_{\mathrm{EC}} - 1,$$

$$\varepsilon_{0,\mathrm{EC}} := \frac{\varepsilon}{4\sqrt{2L_{\mathrm{EC}} \log(2/\delta)}}, \qquad \omega_{\mathrm{EC}} := \frac{4\Delta_{\mathrm{EC}}}{\varepsilon_{0,\mathrm{EC}}} \log\left(\frac{4L_{\mathrm{EC}}(K+N+1)}{\alpha}\right),$$

$$B_{\mathrm{EC}} := \beta_{\mathrm{EC}} + 4\omega_{\mathrm{EC}}, \qquad \widetilde{\beta}_{\mathsf{fit}}^{\mathrm{EC}} := \beta_{\mathrm{EC}} + 2\omega_{\mathrm{EC}}, \qquad \widetilde{\beta}_{\mathsf{switch}}^{\mathrm{EC}} := 16B_{\mathrm{EC}}.$$

Algorithm 3 `Priv-EC-RS`

Input: $\mathcal{F}, \mathcal{G}, K, H$ and $\widetilde{\beta}^{\mathrm{EC}}_{\mathsf{fit}}, \widetilde{\beta}^{\mathrm{EC}}_{\mathsf{switch}}, M_{\mathrm{EC}}, \Delta_{\mathrm{EC}}, \varepsilon_{0,\mathrm{EC}}$.
Publicly order $\mathcal{F} = \{f^{(1)}, \dots, f^{(N)}\}$ so that $V_{f^{(1)},1}(s_1) \geq \cdots \geq V_{f^{(N)},1}(s_1)$, with fixed tie-breaking.
Set $f \leftarrow f^{(1)}$ and $m \leftarrow 0$. `// Initialization.`
Start a trigger `AboveThreshold` call with threshold $\widetilde{\beta}^{\mathrm{EC}}_{\mathsf{switch}}$, sensitivity Δ_{EC}, and privacy level $\varepsilon_{0,\mathrm{EC}}$.
for $k = 1, \dots, K$ **do**
  Execute π_f in episode k and add the observed trajectory to the data.
  if $m < M_{\mathrm{EC}}$ **then** `// Make no further private calls after` M_{EC} `positive triggers.`
    Submit $\widehat{\mathcal{E}}^{\mathrm{EC}}_k(f)$ to the current trigger call.
    if the trigger returns $\top$ **then**
      Set $m \leftarrow m + 1$.
      if $m < M_{\mathrm{EC}}$ and $k < K$ **then**
        Start selector `AboveThreshold` with $(-\widetilde{\beta}^{\mathrm{EC}}_{\mathsf{fit}}, \Delta_{\mathrm{EC}}, \varepsilon_{0,\mathrm{EC}})$.
        Submit $-\widehat{\mathcal{E}}^{\mathrm{EC}}_k(f^{(1)}), \dots, -\widehat{\mathcal{E}}^{\mathrm{EC}}_k(f^{(N)})$ in order until the first $\top$.
        If the first $\top$ occurs at index j, set $f \leftarrow f^{(j)}$; otherwise set $f \leftarrow f^{(1)}$.
        Start fresh trigger `AboveThreshold` for the new f with $(\widetilde{\beta}^{\mathrm{EC}}_{\mathsf{switch}}, \Delta_{\mathrm{EC}}, \varepsilon_{0,\mathrm{EC}})$.
      end if
    end if
  end if
end for

Algorithm 3 has exactly the same private control flow as Algorithm 2. It therefore makes at most $L_{\mathrm{EC}} = 2M_{\mathrm{EC}} - 1$ physical `AboveThreshold` invocations. The utility proof below shows that the cap does not bind on the good event.

E.3 Guarantee

With the parameter choices above, Theorem 6.1 in Section 6 states the privacy, switching, and regret guarantees of Algorithm 3; we prove it below.

For a stagewise product specialization with $N = \prod_h |\mathcal{F}_h|$ and an analogous product auxiliary class, the global cardinality term satisfies $\log(NG) = \Theta(H)$ times a homogeneous stagewise logarithm. The first term of the EC bound then incurs an extra $\sqrt{H}$ relative to Theorem 5.1. This loss comes from the general EC concentration union bound over global tuples (h, I, f, g), whereas Lemma B.2 indexes the Bellman regression locally by $(u, q) \in \mathcal{F}_h \times \mathcal{F}_{h+1}$. The sharper main theorem exploits that stagewise structure; the monolithic EC theorem does not.

E.4 Proof of Theorem 6.1

We follow the proof flow of Theorem 5.1: first establish the statistical comparison and a privacy guarantee independent of the utility events, then prove optimism and historical control, charge fresh EC energy to every positive trigger, and finally apply dominance.

Step 1: decomposability gives the statistical sandwich. Fix h, an interval $I = [k_1, k_2]$, and $f \in \mathcal{F}, g \in \mathcal{G}$. Write

$$Y_i(f,g) := \ell_{h,f^i}(\zeta^i_h, \eta^i_h, f, g),$$
$$\nu_i(f,g) := \mathbb{E}_{\zeta^i_h}[Y_i(f,g)],$$
$$\xi_i(f) := \ell_{h,f^i}(\zeta^i_h, \eta^i_h, f, \mathfrak{T}(f)).$$

By (17), $Y_i(f,g) = \nu_i(f,g) + \xi_i(f)$ and $\mathbb{E}_{\zeta^i_h}[\xi_i(f)] = 0$. Therefore

$$\widehat{L}^{\mathrm{EC}}_{h,I}(f,g) - \widehat{L}^{\mathrm{EC}}_{h,I}(f, \mathfrak{T}(f)) = S_{h,I}(f,g) + Z_{h,I}(f,g), \tag{23}$$

where

$$S_{h,I}(f,g) := \sum_{i\in I} \|\nu_i(f,g)\|_2^2, \qquad Z_{h,I}(f,g) := 2\sum_{i\in I}\langle \nu_i(f,g), \xi_i(f)\rangle.$$

For each summand $X_i = 2\langle \nu_i(f,g), \xi_i(f)\rangle$, conditional mean zero, boundedness, and Jensen's inequality give

$$|X_i| \le 2R^2, \qquad \mathbb{E}_{\zeta_h^i}[X_i^2] \le 4R^2 \|\nu_i(f,g)\|_2^2.$$

Applying Lemma 5 of Chen et al. (2025) with $C = 2R^2$ to the masked martingale differences $\{\mathbf{1}\{i \in I\}X_i\}_{i=1}^K$ and their negatives, as in their proof of Lemma 6, gives, with probability at least $1-2\rho$,

$$|Z_{h,I}(f,g)| \le \lambda \sum_{i\in I} \mathbb{E}_{\zeta_h^i}[X_i^2] + \frac{\log(1/\rho)}{\lambda} \le 4R^2\lambda S_{h,I}(f,g) + \frac{\log(1/\rho)}{\lambda}.$$

Taking $\lambda = 1/(8R^2)$ and $\rho = \alpha/(8HK^2NG)$, and union bounding over at most HK^2NG tuples (h, I, f, g), shows that, with probability at least $1-\alpha/4$, simultaneously for all such tuples,

$$|Z_{h,I}(f,g)| \le \tfrac{1}{2} S_{h,I}(f,g) + \beta_{\mathrm{EC}}. \tag{24}$$

Call this event $\mathcal{G}_{\mathrm{EC}}$. Adaptivity of the behavior hypotheses f^i causes no additional union bound because f^i is predictable before episode i, and Freedman's inequality is applied conditionally on that history. The event is uniform over deterministic intervals, so it also holds for the random epoch intervals selected by the algorithm.

For $g = f$, $\nu_i(f,f) = \mu_{i,h}(f)$. Define $S_{h,I}(f) := \sum_{i\in I} \|\mu_{i,h}(f)\|_2^2$. Since $\mathfrak{T}(f) \in \mathcal{G}$, minimizing (23) over $g \in \mathcal{G}$ and using (24) gives

$$\frac{1}{2} S_{h,I}(f) - \beta_{\mathrm{EC}} \le \widehat{\mathcal{E}}_{h,I}^{\mathrm{EC}}(f) \le \frac{3}{2} S_{h,I}(f) + 2\beta_{\mathrm{EC}}. \tag{25}$$

To verify the upper bound carefully, (24) implies

$$0 \le \widehat{L}_{h,I}^{\mathrm{EC}}(f, \mathfrak{T}(f)) - \min_{g\in\mathcal{G}} \widehat{L}_{h,I}^{\mathrm{EC}}(f,g) \le \beta_{\mathrm{EC}},$$

while the $g = f$ instance of (23) lies between $\frac{1}{2}S_{h,I}(f) - \beta_{\mathrm{EC}}$ and $\frac{3}{2}S_{h,I}(f) + \beta_{\mathrm{EC}}$. For $f = f^\star$, (18) gives $\mathfrak{T}(f^\star) = f^\star$, so the $g = f$ instance of (23) vanishes and the preceding display yields

$$\widehat{\mathcal{E}}_t^{\mathrm{EC}}(f^\star) \le \beta_{\mathrm{EC}} \quad \text{for every } t. \tag{26}$$

Step 2: sensitivity and privacy. Every summand of (21) lies in $[0, R^2]$. Replacing one trajectory changes the first term in (22) by at most R^2 and changes its minimum by at most R^2. Hence $\widehat{\mathcal{E}}_t^{\mathrm{EC}}(f)$ and $-\widehat{\mathcal{E}}_t^{\mathrm{EC}}(f)$ have sensitivity at most $\Delta_{\mathrm{EC}} = 2R^2$ conditional on any public transcript prefix. That prefix fixes the behavior/reference hypotheses f^i appearing in (21); the neighboring rows need not share any realized private history. Taking a maximum over stages adds no factor of H.

Theorem C.1 makes each complete trigger or selector call $(\varepsilon_{0,\mathrm{EC}}, 0)$-DP. The private control flow permits at most $L_{\mathrm{EC}} = 2M_{\mathrm{EC}} - 1$ physical calls on every sample path; padding shorter executions to L_{EC} calls is data independent. The advanced-composition calculation in Proposition D.2, with L, ε_0 replaced by $L_{\mathrm{EC}}, \varepsilon_{0,\mathrm{EC}}$, proves that the complete public transcript is (ε, δ)-DP. The Billboard Lemma then gives (ε, δ)-JDP. This conclusion does not use $\mathcal{G}_{\mathrm{EC}}$ or any bounded-noise event.

Step 3: private optimism and historical EC error. Let $\mathcal{G}_{\mathrm{AT}}^{\mathrm{EC}}$ be the event that every threshold and query noise drawn in Algorithm 3 has magnitude at most ω_{EC}. There are at most $L_{\mathrm{EC}}(K+N+1)$ such variables, each with scale at most $4\Delta_{\mathrm{EC}}/\varepsilon_{0,\mathrm{EC}}$, so a Laplace tail union bound gives $\mathbb{P}(\mathcal{G}_{\mathrm{AT}}^{\mathrm{EC}}) \ge 1 - \alpha/4$.

On $\mathcal{G}_{\mathrm{EC}} \cap \mathcal{G}_{\mathrm{AT}}^{\mathrm{EC}}$, (26) and the last implication in (5) force every selector to stop no later than $f^\star$ in the public order. Thus every selected f satisfies

$$V_{f,1}(s_1) \ge V_1^\star(s_1), \qquad \widehat{\mathcal{E}}_t^{\mathrm{EC}}(f) \le \widetilde{\beta}_{\mathsf{fit}}^{\mathrm{EC}} + 2\omega_{\mathrm{EC}} = B_{\mathrm{EC}}. \tag{27}$$

The initial hypothesis $f^{(1)}$ also satisfies $V_{f^{(1)},1}(s_1) \ge V_1^\star(s_1)$, since it maximizes $V_{f,1}(s_1)$ over a class containing $f^\star$.

The EC stopping episode is

$$\tau_{M_{\mathrm{EC}}} := \inf\left\{ t \in [K] : \sum_{k=1}^{t} \mathbf{1}\{\text{the EC trigger returns } \top \text{ after episode } k\} \geq M_{\mathrm{EC}} \right\},$$

with value $K+1$ if the event never occurs.

For every $k \leq \min\{K, \tau_{M_{\mathrm{EC}}}\}$, either episode k begins an epoch and (27) applies at time $k-1$, or the preceding trigger answer was negative. In the latter case, (5) gives

$$\widehat{\mathcal{E}}^{\mathrm{EC}}_{k-1}(f^k) < 16B_{\mathrm{EC}} + 2\omega_{\mathrm{EC}} \leq 17B_{\mathrm{EC}}.$$

The same bound is trivial for the initial hypothesis at time zero. Applying the lower half of (25) to $[k-1]$ gives, for every h,

$$\sum_{i<k} \left\|\mu_{i,h}(f^k)\right\|_2^2 \leq 36B_{\mathrm{EC}}.$$

Since $36B_{\mathrm{EC}} \geq R^2$, the pathwise EC implication (19)–(20) yields, for every $n \leq \min\{K, \tau_{M_{\mathrm{EC}}}\}$,

$$\sum_{i=1}^{n} \left\|\mu_{i,h}(f^i)\right\|_2^2 \leq 36dB_{\mathrm{EC}} \log(en). \tag{28}$$

Evaluating the pathwise statement at this random prefix requires no stopping-time extension.

Step 4: every positive trigger consumes fresh EC energy. The generalized excess loss is subadditive over disjoint intervals by the same argument as Lemma D.5: the empirical loss is additive and the minimum of a sum is at least the sum of the separate minima. Consider an epoch $[s,t]$ ending in a positive trigger with $t \leq \tau_{M_{\mathrm{EC}}}$. The positive answer and (5) imply

$$\widehat{\mathcal{E}}^{\mathrm{EC}}_{t}(f) \geq 16B_{\mathrm{EC}} - 2\omega_{\mathrm{EC}} \geq 15B_{\mathrm{EC}}.$$

Choose a stage h attaining the maximum. At the epoch start, (27) gives $\widehat{\mathcal{E}}^{\mathrm{EC}}_{h,[s-1]}(f) \leq B_{\mathrm{EC}}$; it is zero for the initial epoch. Subadditivity therefore gives

$$\widehat{\mathcal{E}}^{\mathrm{EC}}_{h,[s,t]}(f) \geq 14B_{\mathrm{EC}}.$$

The upper half of (25), together with $\beta_{\mathrm{EC}} \leq B_{\mathrm{EC}}$, now implies

$$\sum_{i=s}^{t} \left\|\mu_{i,h}(f)\right\|_2^2 \geq 8B_{\mathrm{EC}}. \tag{29}$$

Assign each positive trigger to one stage satisfying (29). Epochs assigned to a fixed stage are disjoint, and $f^i = f$ within each epoch. Combining (28) and (29) shows that the number assigned to stage h is at most $\frac{9}{2}d\log(eK)$. Summing over stages proves the stated switching bound. If the cap were reached, the prefix ending at $\tau_{M_{\mathrm{EC}}}$ would contain M_{EC} positive triggers, contradicting

$$M_{\mathrm{EC}} \leq \frac{9}{2}dH\log(eK) < 9d^{\mathrm{EC}}_{\mathsf{max}}H\log(eK) \leq M_{\mathrm{EC}}.$$

Thus the cap does not bind, and (28) holds with $n = K$.

Step 5: dominance converts EC energy into regret. Let

$$Y_{i,h} := \left\|\mu_{i,h}(f^i)\right\|_2 \in [0,R], \qquad \overline{Y}_{i,h} := \mathbb{E}[Y_{i,h} \mid \mathcal{H}_{i-1}].$$

For each fixed h, $\overline{Y}_{i,h} - Y_{i,h}$ is a martingale difference bounded in absolute value by R. Azuma–Hoeffding and a union bound over h give an event of probability at least $1-\alpha/4$ on which

$$\sum_{i=1}^{K} \overline{Y}_{i,h} \leq \sum_{i=1}^{K} Y_{i,h} + R\sqrt{2K\log(4H/\alpha)} \quad \text{for every } h.$$

On the intersection of this event, the EC dominance event, and the two good events above, exact optimism (27) gives

$$\begin{aligned}\mathsf{Reg}(K) &\le \sum_{i=1}^{K}\left(V_{f^i,1}(s_1) - V_1^{\pi_{f^i}}(s_1)\right)\\ &\le \kappa \sum_{h=1}^{H}\sum_{i=1}^{K} \overline{Y}_{i,h}\\ &\le \kappa \sum_{h=1}^{H}\left(\sqrt{K\sum_{i=1}^{K} Y_{i,h}^2} + R\sqrt{2K\log(4H/\alpha)}\right)\\ &\le 6\kappa H\sqrt{dB_{\mathrm{EC}}K\log(eK)} + \kappa HR\sqrt{2K\log(4H/\alpha)},\end{aligned}$$

which gives (9) after substituting the parameter choices. The statistical, `AboveThreshold`, dominance, and martingale events each fail with probability at most $\alpha/4$, so their intersection has probability at least $1-\alpha$. Together with the privacy argument in Step 2, this completes the proof.

E.5 Recovery of the Bellman–Eluder Specialization

For the model-free value-function setting in the main paper, take $\eta_h = (s_h, a_h)$, $\zeta_h = (r_h, s_{h+1})$, let $\mathcal{G}$ contain the Bellman regression targets, and use the scalar temporal-difference loss

$$\ell_{h,f'}(\zeta_h, \eta_h, f, g) = g_h(s_h, a_h) - r_h - V_{f,h+1}(s_{h+1}).$$

Then the conditional mean at $g = f$ is the Bellman residual $e_h(f; s_h, a_h)$, decomposability follows from Bellman completeness, and the point-mass Bellman–Eluder square-potential argument of Lemma B.3 supplies (20) with d replaced by a constant multiple of the Bellman–Eluder dimension. Hence Theorem 5.1 is the concrete Bellman-regression instance of the EC proof pattern, with sharper constants and class-size dependence obtained by exploiting its stagewise structure.