When Can Pure Exploitation Succeed in Linear RL? Decoys and Self-Identifiability for Greedy LSVI

Manoj Saravanan **Rohit Kumar Salla**

Keywords: theoretical reinforcement learning; exploration-free reinforcement learning; linear Bellman completeness; greedy LSVI; identifiability; regret analysis.

Summary

We study when online reinforcement learning can succeed by pure exploitation in episodic finite-horizon MDPs with known features under linear Bellman completeness. Existing positive results in linear RL largely rely on optimism or randomization, whereas recent exploration-free guarantees for GREEDY-LSVI require strong feature-diversity assumptions. We ask for the exact obstruction to such exploration-free learning. We introduce a *decoy*: an alternative environment in the same model class that induces the same full trajectory law under a greedy-representable policy π yet makes π globally optimal, even though π is suboptimal in the true environment. We define *self-identifiability* as the absence of decoys for all suboptimal greedy-representable policies. Our main theorem shows that decoys are captured by on-policy geometry: up to a mild model-class closure assumption, they correspond exactly to nonzero feature directions that are invisible under on-policy covariance but still change action comparisons on visited states. This yields a sharp dichotomy for GREEDY-LSVI: decoys allow bounded linear-regret instances, whereas self-identifiability yields high-probability $\widetilde{O}(\sqrt{K})$ regret and, under a uniform witnessed margin, $\widetilde{O}(\log K)$ regret.

Contribution(s)

1. We introduce *decoys* and *self-identifiability* for episodic RL in linear Bellman-complete MDPs. A decoy is an alternative environment in the model class that induces the same trajectory law under a greedy-representable policy π but makes π globally optimal; self-identifiability excludes such decoys for every suboptimal greedy-representable policy.
   Context: Prior linear-RL guarantees typically rely on optimism or randomization (Jin et al., 2020; Zanette et al., 2020; Osband et al., 2019; He et al., 2021). We instead identify a policy-level obstruction to pure exploitation, extending the decoy viewpoint from structured bandits (Slivkins et al., 2025) to trajectory-level RL.
2. We prove a geometric characterization of decoys: any decoy yields a nonzero on-policy covariance-null direction that is action-sensitive on visited states; under Assumption 5.3, the converse also holds (Theorem 5.4).
   Context: This converts trajectory-law indistinguishability into a linear-algebraic criterion; Appendix A gives a tabular cemetery-state construction showing that Assumption 5.3 is non-vacuous.
3. We prove a dichotomy for GREEDY-LSVI. Decoys yield bounded linear-regret instances (Theorem 6.1); under self-identifiability, Assumption 5.3, bounded optimal parameters, and finite leverage, GREEDY-LSVI attains high-probability $\widetilde{O}(\sqrt{K})$ regret, improving to $\widetilde{O}(\log K)$ under a witnessed-margin condition (Theorems 6.7–6.9).
   Context: Compared with prior diversity-based analyses of greedy LSVI (Civitavecchia & Papini, 2025), decoy-freedom is the qualitative condition for no-regret learning, while leverage and margin govern the rates.
4. We separate self-identifiability from stronger diversity assumptions: nullspace-eliminating feature diversity implies self-identifiability, yet Proposition 6.10 gives a self-identifiable instance with rank-deficient on-policy covariance.
   Context: Existing diversity conditions (Bastani et al., 2021; Civitavecchia & Papini, 2025) are therefore sufficient, but not necessary, for pure exploitation.

When Can Pure Exploitation Succeed in Linear RL? Decoys and Self-Identifiability for Greedy LSVI

Manoj Saravanan[1] **Rohit Kumar Salla**[1]
manoj663@vt.edu rohits25@vt.edu

[1]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA

Abstract

We study online episodic reinforcement learning in finite-horizon Markov decision processes with a finite action set, known features, bounded rewards, and linear Bellman completeness. Our focus is GREEDY-LSVI, i.e., stage-wise ridge regression on Bellman targets followed by deterministic greedy action selection, with no optimism or randomization. We identify the exact information-theoretic obstruction to exploration-free learning: a *decoy*, namely an alternative environment in the same model class that induces the same trajectory law under a greedy-representable policy π while making π globally optimal, although π is suboptimal in the true environment. We call an instance *self-identifiable* if no suboptimal greedy-representable policy admits a decoy. Our main result is a geometric characterization: decoys are equivalent, up to a mild model-class closure assumption, to the existence of an on-policy covariance null direction that is nevertheless action-sensitive on states visited by π. This yields a scoped dichotomy for GREEDY-LSVI: decoys permit linear-regret instances, whereas self-identifiability implies $\widetilde{O}(\sqrt{K})$ regret over K episodes and, under a witnessed-margin condition, $\widetilde{O}(\log K)$ regret. We also exhibit a rank-deficient self-identifiable instance, showing that full-rank on-policy covariance is not necessary for pure exploitation to succeed.

1 Introduction

Exploration is often indispensable in reinforcement learning: an agent may need to take actions that look suboptimal in the short run to gather information that improves future decisions. In costly or safety-critical settings, however, deliberate exploration may be limited. This motivates a structural question: *when can online reinforcement learning succeed by pure exploitation?*

Under linear structure, explicit exploration can be provably efficient. In episodic linear RL, optimism-based methods such as LSVI-UCB achieve $\widetilde{O}(\sqrt{T})$ regret with $T = KH$ total steps under linear MDP-type assumptions (Jin et al., 2020); Bellman-completeness / low inherent Bellman error frameworks extend this picture beyond strict linear MDPs (Zanette et al., 2020); and favorable gaps yield logarithmic regret (He et al., 2021). These guarantees, however, rely on optimism or randomization.

We study GREEDY-LSVI (Algorithm 1), the natural exploitation-only baseline obtained by performing least-squares value iteration and acting greedily with respect to the fitted Q-function. Recent work proves sublinear regret for GREEDY-LSVI under strong feature-diversity conditions (Civitavecchia & Papini, 2025). Those conditions are only sufficient: they do not isolate the exact obstruction to pure exploitation.

Our perspective is information-theoretic. A *decoy* is an alternative environment in the same model class that is observationally indistinguishable under a policy π yet makes π globally optimal. If

a suboptimal greedy-representable policy admits a decoy, then any exploitation-only learner that keeps executing π gathers data that cannot refute the hypothesis that π is already optimal.

Our main results are threefold. First, we characterize decoys geometrically through on-policy covariance nullspaces and action-sensitive directions (Theorem 5.4). Second, we show that decoys can indeed trap GREEDY-LSVI and yield linear regret on bounded instances (Theorem 6.1). Third, we prove that once this obstruction is absent—i.e., the instance is *self-identifiable*—GREEDY-LSVI achieves $\widetilde{O}(\sqrt{K})$ regret, and $\widetilde{O}(\log K)$ under a witnessed-margin condition (Theorems 6.7–6.9). We also give a rank-deficient self-identifiable example (Proposition 6.10), showing that full-rank on-policy covariance is stronger than necessary.

2 Related work

Under linear MDP-type structure, optimism-based methods such as LSVI-UCB achieve polynomial-time learning and $\widetilde{O}(\sqrt{T})$ regret (Jin et al., 2020); Bellman-completeness / low inherent Bellman error frameworks extend these guarantees to richer approximation regimes (Zanette et al., 2020); and favorable gaps yield logarithmic regret (He et al., 2021). Randomized value-function methods provide an alternative route to deep exploration (Osband et al., 2019).

Closest to our setting, Civitavecchia & Papini (2025) analyze GREEDY-LSVI under strong diversity assumptions. Our contribution is complementary: we identify a sharp obstruction—decoys—and a weaker structural condition—self-identifiability—that governs the success or failure of pure exploitation. This viewpoint is inspired by recent decoy-based characterizations in structured bandits (Bastani et al., 2021; Slivkins et al., 2025), but in RL the relevant indistinguishability notion must operate at the level of full trajectory laws.

3 Preliminaries and setup

We consider episodic RL in an unknown finite-horizon MDP with linear action-value approximation and analyze GREEDY-LSVI (LSVI without bonuses/randomization). Throughout, we fix a deterministic tie-breaking rule on $\mathcal{A}$ once and for all so that greedy policies are single-valued functions of Q.

3.1 MDPs, value functions, and regret

An episodic MDP is $M = (\mathcal{S}, \mathcal{A}, H, \mu, \{P_h\}_{h=1}^H, \{r_h\}_{h=1}^H)$ with finite $\mathcal{A}$ and horizon H. In episode k, a policy $\pi^k = (\pi_h^k)_{h=1}^H$ is chosen, and the trajectory evolves as

$$S_{k,1} \sim \mu, \qquad A_{k,h} \sim \pi_h^k(\cdot \mid S_{k,h}), \qquad S_{k,h+1} \sim P_h(\cdot \mid S_{k,h}, A_{k,h}).$$

We observe a reward $R_{k,h} \in [0,1]$ with $\mathbb{E}[R_{k,h} \mid S_{k,h}{=}s, A_{k,h}{=}a] = r_h(s,a)$.

For any policy π, define

$$V_h^\pi(s) := \mathbb{E}_M^\pi\Big[\sum_{t=h}^H R_t \,\Big|\, S_h = s\Big], \qquad Q_h^\pi(s,a) := r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V_{h+1}^\pi(s')],$$

with $V_{H+1}^\pi \equiv 0$. Let $V_h^\star(s) := \sup_\pi V_h^\pi(s)$ and $Q_h^\star(s,a) := \sup_\pi Q_h^\pi(s,a)$; they satisfy the Bellman optimality equations

$$Q_h^\star(s,a) = r_h(s,a) + \mathbb{E}_{s'\sim P_h(\cdot|s,a)}\Big[\max_{a'\in\mathcal{A}} Q_{h+1}^\star(s',a')\Big], \qquad V_h^\star(s) = \max_{a\in\mathcal{A}} Q_h^\star(s,a). \tag{1}$$

We measure (pseudo-)regret:

$$\mathrm{Reg}(K) := \sum_{k=1}^K \big(V_1^\star - V_1^{\pi^k}\big), \qquad V_1^\pi := \mathbb{E}_{S\sim\mu}[V_1^\pi(S)]. \tag{2}$$

We write $\mathbb{P}_M^\pi$ and $\mathbb{E}_M^\pi$ for the trajectory law and expectation under (M,π).

3.2 Linear Q-functions and Bellman completeness

Fix feature maps $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and linear classes $\mathcal{Q}_h := \{(s,a) \mapsto \langle \phi_h(s,a), w\rangle : w \in \mathbb{R}^d\}$. Define the Bellman optimality operator

$$(\mathcal{T}_h Q)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]. \tag{3}$$

Standing assumptions.

- **Bounded features:** $\|\phi_h(s,a)\|_2 \le 1$ for all (s,a,h).
- **Linear Bellman completeness:** for each h and each $Q \in \mathcal{Q}_{h+1}$, $\mathcal{T}_h Q \in \mathcal{Q}_h$, i.e., for every $w' \in \mathbb{R}^d$ there exists $w \in \mathbb{R}^d$ such that

$$\langle \phi_h(s,a), w\rangle = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} \langle \phi_{h+1}(s',a'), w'\rangle\right] \quad \forall (s,a). \tag{4}$$

Under (4), $Q^\star$ is realizable by $(\mathcal{Q}_h)_{h=1}^H$ by backward induction on (1).

For any policy π and stage h, let $d^\pi_{M,h}$ be the induced distribution of (S_h, A_h) and define the feature covariance

$$\Sigma^\pi_{M,h} := \mathbb{E}_{(S,A)\sim d^\pi_{M,h}}\big[\phi_h(S,A)\phi_h(S,A)^\top\big] \in \mathbb{R}^{d\times d}. \tag{5}$$

For deterministic π, this reduces to

$$\Sigma^\pi_{M,h} = \mathbb{E}\big[\phi_h(S_h, \pi_h(S_h))\phi_h(S_h, \pi_h(S_h))^\top\big].$$

3.3 GREEDY-LSVI

Fix $\lambda > 0$. At the start of episode k, GREEDY-LSVI computes $\widehat{w}_{k,h}$ by backward ridge regression using data from episodes $\tau < k$. Define $\widehat{Q}_{k,h}(s,a) := \langle \phi_h(s,a), \widehat{w}_{k,h}\rangle$ and $\widehat{V}_{k,h}(s) := \max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s,a)$, with $\widehat{V}_{k,H+1} \equiv 0$. Targets are

$$Y^{(k)}_{\tau,h} := R_{\tau,h} + \widehat{V}_{k,h+1}(S_{\tau,h+1}) = R_{\tau,h} + \max_{a\in\mathcal{A}} \langle \phi_{h+1}(S_{\tau,h+1}, a), \widehat{w}_{k,h+1}\rangle. \tag{6}$$

The ridge estimator is

$$\widehat{w}_{k,h} := \arg\min_{w\in\mathbb{R}^d} \sum_{\tau=1}^{k-1} \Big(\langle \phi_h(S_{\tau,h}, A_{\tau,h}), w\rangle - Y^{(k)}_{\tau,h}\Big)^2 + \lambda \|w\|_2^2, \tag{7}$$

equivalently $\widehat{w}_{k,h} = V_{k,h}^{-1} b_{k,h}$ with

$$V_{k,h} := \lambda I_d + \sum_{\tau=1}^{k-1} \phi_h(S_{\tau,h}, A_{\tau,h})\phi_h(S_{\tau,h}, A_{\tau,h})^\top, \qquad b_{k,h} := \sum_{\tau=1}^{k-1} \phi_h(S_{\tau,h}, A_{\tau,h})\, Y^{(k)}_{\tau,h}. \tag{8}$$

The deployed policy is greedy:

$$\pi^k_h(s) \in \arg\max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s,a), \tag{9}$$

with deterministic tie-breaking.

Algorithm 1: GREEDY-LSVI (pure exploitation)

Input: $(\phi_h)_{h=1}^H$, ridge $\lambda > 0$.
For $k = 1, 2, \ldots, K$:

1. **Backward fit:** Set $\widehat{w}_{k,H+1} \leftarrow 0$. For $h = H, H-1, \ldots, 1$: form targets $Y_{\tau,h}^{(k)}$ via (6) and set $\widehat{w}_{k,h} \leftarrow V_{k,h}^{-1} b_{k,h}$ using (8).
2. **Greedy rollout:** $S_{k,1} \sim \mu$; for $h = 1, \ldots, H$, choose $A_{k,h} \in \arg\max_{a \in \mathcal{A}} \langle \phi_h(S_{k,h}, a), \widehat{w}_{k,h} \rangle$, then observe $(R_{k,h}, S_{k,h+1})$.

4 Decoys and self-identifiability

A purely greedy algorithm can fail for an information-theoretic reason: the trajectory distribution under a fixed policy may not identify whether that policy is suboptimal within the model class. We formalize this via *decoys* and define *self-identifiability* as the absence of such obstructions.

Model class and measurability. Fix $(\phi_h)_{h=1}^H$ and let $\mathcal{M}$ be the class of episodic MDPs on a fixed measurable state space $\mathcal{S}$ (assumed standard Borel) and finite $\mathcal{A}$, satisfying bounded rewards/features and linear Bellman completeness (4). When defining indistinguishability, we treat rewards via conditional kernels $\nu_h(\cdot \mid s, a)$ supported on $[0, 1]$ so that $\mathbb{P}_M^\pi$ is a full trajectory law.

Trajectory laws and observational equivalence

A trajectory is $\tau = (S_1, A_1, R_1, \ldots, S_H, A_H, R_H, S_{H+1})$. For $M \in \mathcal{M}$ and policy π, let $\mathbb{P}_M^\pi$ be the induced probability measure on trajectories and $\mathbb{E}_M^\pi$ the corresponding expectation.

Definition 4.1 (Observational equivalence under a policy) *Two environments $M, M' \in \mathcal{M}$ are observationally equivalent under π if $\mathbb{P}_M^\pi = \mathbb{P}_{M'}^\pi$ as probability measures on trajectories.*

Lemma 4.2 (Trajectory-law equality implies equality of on-policy kernels) *If $\mathbb{P}_M^\pi = \mathbb{P}_{M'}^\pi$, then for each stage h and for $\rho_{M,h}^\pi$-a.e. s (the state marginal under π in M), the regular conditional law of (R_h, S_{h+1}) given $S_h = s$ and $A_h = \pi_h(s)$ is identical in M and M'.*

Proof. Because $\mathcal{S}$ is standard Borel and $\mathcal{A}$ is finite, regular conditional distributions exist and are unique a.s. Equality of trajectory measures implies equality of all finite-dimensional marginals; applying disintegration at stage h yields equality of the conditional laws of (R_h, S_{h+1}) given (S_h, A_h), and restricting to $A_h = \pi_h(S_h)$ gives the claim. ■

Decoys

Definition 4.3 (Decoy for a policy) *Fix $M^\star \in \mathcal{M}$. A Markov policy π admits a decoy (relative to $M^\star$) if there exists $M' \in \mathcal{M}$ such that (i) $\mathbb{P}_{M^\star}^\pi = \mathbb{P}_{M'}^\pi$, (ii) π is globally optimal in M' (i.e., $V_{1,M'}^\pi = V_{1,M'}^\star$), and (iii) π is suboptimal in $M^\star$ (i.e., $V_{1,M^\star}^\pi < V_{1,M^\star}^\star$).*

Greedy-representable policies and self-identifiability

Fix deterministic tie-breaking on $\mathcal{A}$. For $w = (w_1, \ldots, w_H) \in (\mathbb{R}^d)^H$, define the deterministic Markov policy

$$\pi_h^w(s) := \arg\max_{a \in \mathcal{A}} \langle \phi_h(s, a), w_h \rangle, \qquad h \in [H], \tag{10}$$

and the policy class

$$\Pi_{\text{lin-greedy}} := \{\pi^w : w \in (\mathbb{R}^d)^H\}. \tag{11}$$

By construction, the policy deployed by GREEDY-LSVI in episode k lies in $\Pi_{\text{lin-greedy}}$.

Definition 4.4 (Self-identifiability (decoy-freedom)) *An instance $M^\star \in \mathcal{M}$ is self-identifiable (relative to $\Pi_{\text{lin-greedy}}$) if no suboptimal $\pi \in \Pi_{\text{lin-greedy}}$ admits a decoy relative to $M^\star$.*

5 Geometric characterization of decoys

We translate the information-theoretic definition of decoys into a geometric condition expressed through on-policy feature covariances and action-difference directions.

5.1 On-policy geometry: unexcited directions and action sensitivity

Fix $M^\star \in \mathcal{M}$ and a deterministic Markov policy π. Let ρ_h^π be the marginal distribution of S_h under $\mathbb{P}_{M^\star}^\pi$.

On-policy covariance. Define

$$\Sigma_h(\pi) := \mathbb{E}_{S\sim\rho_h^\pi}\left[\phi_h(S,\pi_h(S))\phi_h(S,\pi_h(S))^\top\right]. \tag{12}$$

Lemma 5.1 (Nullspace as almost-sure orthogonality) *For any $h \in [H]$ and $u \in \mathbb{R}^d$,*

$$u \in \mathrm{Null}(\Sigma_h(\pi)) \quad\Longleftrightarrow\quad \langle u, \phi_h(S,\pi_h(S))\rangle = 0 \text{ a.s. under } S \sim \rho_h^\pi.$$

Action-difference sensitivity. Let $\Delta_h^\pi(s,a) := \phi_h(s,a) - \phi_h(s,\pi_h(s))$ and define

$$\Gamma_h(\pi) := \mathbb{E}_{S\sim\rho_h^\pi}\left[\sum_{a\in\mathcal{A}} \Delta_h^\pi(S,a)\Delta_h^\pi(S,a)^\top\right]. \tag{13}$$

Lemma 5.2 (Action sensitivity as a second-moment test) *For any $h \in [H]$ and $u \in \mathbb{R}^d$,*

$$u \in \mathrm{Null}(\Gamma_h(\pi)) \quad\Longleftrightarrow\quad \langle u, \Delta_h^\pi(S,a)\rangle = 0 \text{ for all } a \in \mathcal{A},\ \text{a.s. under } S \sim \rho_h^\pi.$$

Thus, $u \neq 0$ with $u \in \mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$ is *unexcited yet action-sensitive*: it is unobservable under π but can alter action comparisons on ρ_h^π-typical states.

5.2 Model-class closure for the "⇐" direction

The "decoy ⇒ geometry" direction uses only Bellman completeness and observational equivalence. The reverse direction requires that the model class is rich enough to realize a *decoy completion* whenever an unexcited, action-sensitive direction exists.

Assumption 5.3 (Decoy completion / closure) *Fix $M^\star \in \mathcal{M}$ and deterministic Markov π. Whenever there exist $h \in [H]$ and $u \neq 0$ with $u \in \mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$, there exists an environment $M' \in \mathcal{M}$ such that:*

1. *$\mathbb{P}_{M^\star}^\pi = \mathbb{P}_{M'}^\pi$ (trajectory-law invariance under π), and*
2. *π is globally optimal in M' (so $V_{1,M'}^\pi = V_{1,M'}^\star$).*

Appendix A gives a concrete sufficient condition (with an explicit construction) ensuring Assumption 5.3 holds for a standard subclass.

5.3 Main characterization theorem

Theorem 5.4 (Decoys ⟺ unexcited action-sensitive directions) *Fix $M^\star \in \mathcal{M}$ and let $\pi \in \Pi_{\text{lin-greedy}}$ be suboptimal in $M^\star$: $V_{1,M^\star}^\pi < V_{1,M^\star}^\star$.*

1. ***(Necessity, unconditional)*** *If π admits a decoy relative to $M^\star$, then there exist $h \in [H]$ and $u \neq 0$ such that*
$$u \in \mathrm{Null}(\Sigma_h(\pi)) \quad \textit{and} \quad u \notin \mathrm{Null}(\Gamma_h(\pi)). \tag{14}$$

2. ***(Sufficiency, requires Assumption 5.3)*** *If (14) holds for some h and $u \neq 0$, then π admits a decoy relative to $M^\star$.*

Necessity. Let M' be a decoy for π, so $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$, π is optimal in M', and π is suboptimal in $M^\star$. Since $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$, we have $V^\pi_{1,M'} = V^\pi_{1,M^\star}$. Since π is optimal in M', $V^\star_{1,M'} = V^\pi_{1,M'}$. Hence
$$V^\star_{1,M'} = V^\pi_{1,M^\star} < V^\star_{1,M^\star},$$
so $V^\star_{1,M'} \neq V^\star_{1,M^\star}$ and therefore the optimal value functions differ.

Let $\delta Q_h := Q^\star_{h,M'} - Q^\star_{h,M^\star} \in \mathcal{Q}_h$, and $\delta V_h(s) := V^\star_{h,M'}(s) - V^\star_{h,M^\star}(s)$. Define
$$h^\star := \max\Big\{h \in [H] : \mathbb{P}_{S\sim\rho^\pi_h}\big(\delta V_h(S) \neq 0\big) > 0\Big\}.$$
Then $\mathbb{P}_{S\sim\rho^\pi_{h^\star+1}}(\delta V_{h^\star+1}(S) \neq 0) = 0$ (with $\delta V_{H+1} \equiv 0$).

By Lemma 4.2, for $\rho^\pi_{h^\star}$-a.e. s, the conditional law of $(R_{h^\star}, S_{h^\star+1})$ given $S_{h^\star} = s$ and $A_{h^\star} = \pi_{h^\star}(s)$ is the same in $M^\star$ and M'. Therefore, for $\rho^\pi_{h^\star}$-a.e. s,
$$\begin{aligned}\delta Q_{h^\star}(s,\pi_{h^\star}(s)) &= \Big(r_{h^\star,M'} - r_{h^\star,M^\star}\Big)(s,\pi_{h^\star}(s)) + \mathbb{E}_{S'\sim P_{h^\star,M'}(\cdot|s,\pi_{h^\star}(s))}\big[V^\star_{h^\star+1,M'}(S')\big] \\ &\quad - \mathbb{E}_{S'\sim P_{h^\star,M^\star}(\cdot|s,\pi_{h^\star}(s))}\big[V^\star_{h^\star+1,M^\star}(S')\big] \\ &= \mathbb{E}_{S'\sim P_{h^\star,M^\star}(\cdot|s,\pi_{h^\star}(s))}\big[\delta V_{h^\star+1}(S')\big] = 0,\end{aligned}$$
where the last equality uses $\delta V_{h^\star+1}(S') = 0$ $P_{h^\star,M^\star}(\cdot \mid s,\pi_{h^\star}(s))$-a.s. for $\rho^\pi_{h^\star}$-a.e. s (a standard disintegration argument). Hence $\delta Q_{h^\star}(S,\pi_{h^\star}(S)) = 0$ a.s. under $S \sim \rho^\pi_{h^\star}$.

Write $\delta Q_{h^\star}(s,a) = \langle \phi_{h^\star}(s,a), u\rangle$ for some $u \in \mathbb{R}^d$ (since $\delta Q_{h^\star} \in \mathcal{Q}_{h^\star}$). Then $\langle \phi_{h^\star}(S,\pi_{h^\star}(S)), u\rangle = 0$ a.s. under $S \sim \rho^\pi_{h^\star}$, and Lemma 5.1 implies $u \in \mathrm{Null}(\Sigma_{h^\star}(\pi))$.

Finally, by definition of $h^\star$, there exists a measurable set B with $\rho^\pi_{h^\star}(B) > 0$ such that $\delta V_{h^\star}(s) \neq 0$ on B. If $\delta Q_{h^\star}(s,a) = 0$ for all a, then $\max_a Q^\star_{h^\star,M'}(s,a) = \max_a Q^\star_{h^\star,M^\star}(s,a)$, implying $\delta V_{h^\star}(s) = 0$, contradiction. Thus there exists a such that $\delta Q_{h^\star}(s,a) \neq 0$ on a subset of B. Since $\delta Q_{h^\star}(s,\pi_{h^\star}(s)) = 0$ on B, we obtain
$$\langle u, \Delta^\pi_{h^\star}(s,a)\rangle = \delta Q_{h^\star}(s,a) - \delta Q_{h^\star}(s,\pi_{h^\star}(s)) \neq 0$$
on a set of positive $\rho^\pi_{h^\star}$-measure. Lemma 5.2 then yields $u \notin \mathrm{Null}(\Gamma_{h^\star}(\pi))$, proving (14). ■

Sufficiency. Assume (14). By Assumption 5.3, there exists $M' \in \mathcal{M}$ with $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$ and such that π is globally optimal in M'. Since π is suboptimal in $M^\star$ by premise, M' is a decoy for π relative to $M^\star$. ■

6 Regret consequences for GREEDY-LSVI

Theorem 5.4 yields a success/failure dichotomy: decoys can trap pure exploitation and allow linear regret instances, whereas self-identifiability enables no-regret learning by GREEDY-LSVI.

6.1 Failure: decoys can force linear regret

Theorem 6.1 (Decoys imply linear regret for GREEDY-LSVI) *There exists $M^\star \in \mathcal{M}$ with $H = 2$, $|\mathcal{A}| = 2$, and $d = 2$ such that $M^\star$ is not self-identifiable and, for any $\lambda > 0$,* GREEDY-LSVI *with deterministic tie-breaking satisfies*
$$\mathrm{Reg}(K) \geq \tfrac{1}{2}K \qquad \textit{for all } K \in \mathbb{N}.$$

6.2 Success: self-identifiability yields no-regret

A key structural consequence needed for regret analysis is that *relevant action differences are estimable from on-policy data*. This implication uses Theorem 5.4(2) and thus requires Assumption 5.3.

Lemma 6.2 (A deterministic null witness for random violations of a linear span) *Let $\Sigma \succeq 0$ and let Z be a random vector in $\mathbb{R}^d$. If $\mathbb{P}(Z \notin \mathrm{Range}(\Sigma)) > 0$, then there exists a deterministic $u \in \mathrm{Null}(\Sigma)$ such that $\mathbb{P}(u^\top Z \neq 0) > 0$.*

Lemma 6.3 (Action differences lie in the on-policy span (for suboptimal π)) *Assume Assumption 5.3. Fix $M^\star \in \mathcal{M}$ and a deterministic Markov policy π that is* suboptimal *in $M^\star$. If π does not admit a decoy relative to $M^\star$, then for every $h \in [H]$ and $a \in \mathcal{A}$,*

$$\Delta_h^\pi(S, a) \in \mathrm{Range}(\Sigma_h(\pi)) \quad \text{a.s. under } S \sim \rho_h^\pi.$$

Definition 6.4 (On-policy leverage of action differences) *For deterministic π and stage h, define*

$$L_h(\pi) := \operatorname*{ess\,sup}_{S \sim \rho_h^\pi} \max_{a \in \mathcal{A}} \|\Delta_h^\pi(S, a)\|_{\Sigma_h(\pi)^\dagger}, \qquad \|x\|_A := \sqrt{x^\top A x}.$$

Let $L := \sup_{\pi \in \Pi_{\text{lin-greedy}}} \max_{h \in [H]} L_h(\pi)$ and assume $L < \infty$. Define $\widetilde{L} := 1 + L$.

Remark 6.5 (Self-identifiability is qualitative; L is quantitative) *Self-identifiability is a* qualitative *property (success vs. failure of pure exploitation). The constant L is the corresponding* quantitative *hardness parameter controlling rates: it can be large when $\Sigma_h(\pi)$ is nearly singular on the action-difference span. A sufficient quantitative condition is a restricted eigenvalue bound on $\mathrm{span}\{\Delta_h^\pi(s, a)\}$, which implies $L_h(\pi)$ is bounded.*

Assumption 6.6 (Bounded optimal parameters (minimum-norm representation)) *For each $h \in [H]$, let $w_h^\star$ denote the minimum Euclidean norm vector representing $Q_h^\star$ in $\mathcal{Q}_h$:*

$$w_h^\star \in \arg\min\{\|w\|_2 : Q_h^\star(\cdot, \cdot) = \langle \phi_h(\cdot, \cdot), w \rangle\}.$$

Assume $\|w_h^\star\|_2 \leq B$ for all h, for some $B > 0$.

Theorem 6.7 (Self-identifiability implies sublinear regret) *Assume Assumption 5.3. Fix $\lambda > 0$ and suppose $M^\star \in \mathcal{M}$ is self-identifiable (Definition 4.4) and satisfies Assumption 6.6. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,* GREEDY-LSVI *satisfies*

$$\mathrm{Reg}(K) \;\leq\; C\, H^3\, \widetilde{L} \left(\sqrt{d \log\!\left(1 + \tfrac{K}{\lambda}\right)} + \sqrt{\lambda}\, B \right) \sqrt{dK \log\!\left(\tfrac{H}{\delta}\right)} \tag{15}$$

for a universal constant $C > 0$.

6.3 Logarithmic regret under a witnessed margin

For deterministic π, define the one-step improvement signal

$$g_h^\pi(s) := \max_{a \in \mathcal{A}} \big(Q_h^\star(s, a) - Q_h^\star(s, \pi_h(s)) \big) = \max_{a \in \mathcal{A}} \langle w_h^\star, \Delta_h^\pi(s, a) \rangle.$$

Assumption 6.8 (Uniform witnessed margin) *There exist $\gamma_0 \in (0, 1]$ and $p_0 \in (0, 1]$ such that for every suboptimal $\pi \in \Pi_{\text{lin-greedy}}$ there exists $h \in [H]$ with*

$$\mathbb{P}_{S \sim \rho_h^\pi} \big(g_h^\pi(S) \geq \gamma_0 \big) \geq p_0.$$

Theorem 6.9 (Uniform margin implies logarithmic regret) *Assume Assumption 5.3. Under the assumptions of Theorem 6.7 and Assumption 6.8, there exists a universal $C' > 0$ such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Reg}(K) \;\le\; C'\, H^3\, \frac{d\,\widetilde{L}^2}{p_0\,\gamma_0^2}\left(\sqrt{d\log\!\left(1+\tfrac{K}{\lambda}\right)}+\sqrt{\lambda}\,B\right)^2 \log\!\left(\tfrac{KH}{\delta}\right).$$

Proposition 6.10 (Self-identifiable yet rank-deficient) *There exists $M^\star \in \mathcal{M}$ that is self-identifiable but for which $\Sigma_h(\pi)$ is rank-deficient for some $\pi \in \Pi_{\text{lin-greedy}}$ and some stage h.*

7 Discussion and implications

Theorem 5.4 isolates the minimal obstruction to exploration-free learning in linear Bellman-complete MDPs: an on-policy covariance null direction that still changes action comparisons on visited states. Strong diversity assumptions rule out null directions altogether and therefore suffice, but they are not necessary; self-identifiability excludes only null directions that remain action-sensitive.

The characterization also suggests algorithmic refinements. Rather than generic optimism, one could inject targeted exploration only when action-sensitive null directions are detected. The main limitation of our sufficiency results is the decoy-completion closure assumption (Appendix A); quantitative rates further depend on leverage and, in the logarithmic regime, on the witnessed-margin parameters.

8 Conclusion

We identify decoys as the precise information-theoretic obstruction to pure exploitation in linear Bellman-complete RL. This yields a geometric characterization of failure, a linear-regret lower bound when decoys exist, and sublinear / logarithmic guarantees for GREEDY-LSVI on self-identifiable instances.

A A concrete sufficient condition for Assumption 5.3

We record a simple subclass in which Assumption 5.3 is provably non-vacuous.

Proposition A.1 (Tabular class with a cemetery state implies Assumption 5.3) *Assume $\mathcal{S}$ and $\mathcal{A}$ are finite. Fix a distinguished cemetery state $s_{\text{dead}} \in \mathcal{S}$. Consider the subclass of $\mathcal{M}$ consisting of all bounded episodic MDPs on $(\mathcal{S}, \mathcal{A}, H, \mu)$ such that:*

1. *rewards are supported on $[0,1]$;*
2. *the feature map is tabular at each stage: $d = |\mathcal{S}| \cdot |\mathcal{A}|$ and $\phi_h(s,a) = e_{(s,a)}$, so $\mathcal{Q}_h$ contains all functions on $\mathcal{S} \times \mathcal{A}$.*

Then Assumption 5.3 holds for this subclass.

Proof. Fix $M^\star$, deterministic π, and suppose $u \in \mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$ for some h. Construct M' as follows. On every π-reachable state s and stage t, keep the on-policy kernel unchanged:

$$(\nu'_t(\cdot \mid s, \pi_t(s)), P'_t(\cdot \mid s, \pi_t(s))) = (\nu_t(\cdot \mid s, \pi_t(s)), P_t(\cdot \mid s, \pi_t(s))).$$

For every off-policy action $a \neq \pi_t(s)$ on a π-reachable state, set deterministic reward 0 and transition deterministically to s_{dead}; make s_{dead} absorbing with zero reward; and define unreachable states arbitrarily subject to the same no-positive-gain pattern. Then $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$.

Now consider any policy π'. If π' never deviates from π on the π-reachable support, then it induces the same trajectory law as π and therefore has the same value. Otherwise, at its first deviation on

the π-reachable support, it receives reward 0 and is sent to s_{dead}, after which all future rewards are 0. Hence no deviation can improve on π, so π is globally optimal in M'. Since the class is tabular, $M' \in \mathcal{M}$, proving Assumption 5.3. ■

B Auxiliary lemmas

Proof of Lemma 6.2. Let P_{Null} denote the orthogonal projector onto $\text{Null}(\Sigma)$ and set $Y := P_{\text{Null}} Z$. Then $Z \notin \text{Range}(\Sigma)$ iff $Y \neq 0$, since $\text{Range}(\Sigma) = \text{Null}(\Sigma)^{\perp}$ for $\Sigma \succeq 0$. Hence $\mathbb{P}(\|Y\|_2 > 0) > 0$. With $M := \mathbb{E}[YY^{\top}] \neq 0$, choose a unit eigenvector $u \in \text{Null}(\Sigma)$ of M corresponding to $\lambda_{\max}(M) > 0$. Then $\mathbb{E}[(u^{\top} Z)^2] = u^{\top} M u > 0$, so $\mathbb{P}(u^{\top} Z \neq 0) > 0$. ■

Proof of Lemma 6.3. Fix a suboptimal deterministic π, stage h, and action a. If $\mathbb{P}_{S \sim \rho_h^{\pi}}(\Delta_h^{\pi}(S,a) \notin \text{Range}(\Sigma_h(\pi))) > 0$, Lemma 6.2 yields $u \in \text{Null}(\Sigma_h(\pi))$ with $\mathbb{P}(\langle u, \Delta_h^{\pi}(S,a)\rangle \neq 0) > 0$. By Lemma 5.2, $u \notin \text{Null}(\Gamma_h(\pi))$, so (14) holds. Theorem 5.4(2) then produces a decoy, contradiction. ■

C Proof of Theorem 6.1

Proof. Take $\mathcal{A} = \{0,1\}$, $H = 2$, $d = 2$, $\mathcal{S}_1 = \{s\}$, $\mathcal{S}_2 = \{x_0, x_1\}$, and deterministic transition

$$P_1(x_0 \mid s, 0) = 1, \qquad P_1(x_1 \mid s, 1) = 1.$$

Let $r_1 \equiv 0$, $r_2(x_0, a) = \frac{1}{2}$, $r_2(x_1, a) = 1$, and

$$\phi_1(s,0) = e_1,\ \phi_1(s,1) = e_2, \qquad \phi_2(x_0,a) = e_1,\ \phi_2(x_1,a) = e_2 \quad (\forall a \in \mathcal{A}).$$

Rewards lie in $[0,1]$, $\|\phi_h\| \leq 1$, and Bellman completeness is immediate: $\mathcal{T}_2 Q = r_2 \in \mathcal{Q}_2$, while for any $w' \in \mathbb{R}^2$, $(\mathcal{T}_1 Q_{2,w'})(s,i) = w'_i$ for $i \in \{0,1\}$.

The optimal stage-1 action is 1, so $V^{\star}_{1,M^{\star}} = 1$. Run GREEDY-LSVI with tie-breaking toward action 0. Episode 1 therefore picks action 0. Suppose inductively that all earlier episodes also picked action 0. Then every stage-2 sample has feature e_1, so $V_{k,2}$ and $b_{k,2}$ have zero second coordinate; hence $\widehat{V}_{k,2}(x_1) = 0$. Therefore the stage-1 regression target for action 1 is always 0, whereas the target for action 0 becomes positive after the first reward observation. Thus $\widehat{Q}_{k,1}(s,0) > \widehat{Q}_{k,1}(s,1)$ and the algorithm continues to pick action 0. By induction, it never visits x_1.

Each episode therefore earns value $1/2$ while the optimum is 1, so $\text{Reg}(K) \geq K/2$. For the decoy, change only the unvisited reward to $r'_2(x_1, a) = 0$ for all a, leaving everything else unchanged. Then the trapped policy induces the same trajectory law in $M^{\star}$ and M', is optimal in M', and remains greedy-representable under the fixed tie-break. Since r'_2 is still linear in ϕ_2, Bellman completeness is preserved. ■

D Proofs for the upper bounds

We keep only the load-bearing steps for Theorems 6.7 and 6.9; the full filtration-level bookkeeping appears in the supplement.

Performance difference.

Lemma D.1 (Stage-wise decomposition of regret) *For any (possibly history-dependent) policy π,*

$$V_1^{\star} - V_1^{\pi} = \sum_{h=1}^{H} \mathbb{E}_{S \sim \rho_h^{\pi}}\big[g_h^{\pi}(S)\big].$$

Consequently,

$$\text{Reg}(K) = \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\big[g_h^{\pi^k}(S_{k,h})\big].$$

Proof. This is the standard finite-horizon performance-difference identity obtained by telescoping the Bellman equations. ■

Greedy gap. For episode k, stage h, let $s = S_{k,h}$, $a_k = \pi_h^k(s)$, and $a^\star \in \arg\max_a Q_h^\star(s,a)$. Greediness of $\widehat{Q}_{k,h}$ gives

$$Q_h^\star(s,a^\star) - Q_h^\star(s,a_k) \le 2 \max_{a\in\mathcal{A}} \left|(\widehat{Q}_{k,h} - Q_h^\star)(s,a)\right|. \tag{16}$$

Ridge confidence.

Lemma D.2 (Uniform self-normalized event (Abbasi-Yadkori et al., 2011)) *For any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $k \le K$ and $h \in [H]$,*

$$\|\widehat{w}_{k,h} - w_h^\star\|_{V_{k,h}} \le \beta_{K,\delta} := c\Big(\sqrt{d\log(1+K/\lambda)} + \sqrt{\lambda}\,B + \sqrt{\log(H/\delta)}\Big)$$

for a universal constant c.

Remark D.3 (Bootstrapped targets) *Bellman completeness makes each stage-wise regression well specified; the standard reward-noise and transition-noise decomposition then yields Lemma D.2 exactly as in existing LSVI analyses.*

Trace control.

Lemma D.4 (Pseudoinverse control via a trace term) *Fix a deterministic policy π and stage h. Let $\Sigma := \Sigma_h(\pi)$ and let $V \succ 0$. If $z \in \mathrm{Range}(\Sigma)$, then*

$$\|z\|_{V^{-1}} \le \|z\|_{\Sigma^\dagger}\,\sqrt{\mathrm{tr}(\Sigma V^{-1})}.$$

Proof. Write $z = \Sigma^{1/2}v$ with $v = \Sigma^{\dagger/2}z$. Then

$$\|z\|_{V^{-1}}^2 = v^\top \Sigma^{1/2}V^{-1}\Sigma^{1/2}v \le \|v\|_2^2\,\lambda_{\max}(\Sigma^{1/2}V^{-1}\Sigma^{1/2}) \le \|v\|_2^2\,\mathrm{tr}(\Sigma V^{-1}),$$

and $\|v\|_2^2 = \|z\|_{\Sigma^\dagger}^2$. ■

Proof sketch of Theorem 6.7. Work on the event in Lemma D.2 with δ/H in place of δ, and abbreviate $\beta := \beta_{K,\delta/H}$. If the episode-$k$ policy π^k is optimal, the episode incurs no regret. Otherwise, self-identifiability plus Lemma 6.3 imply

$$\Delta_h^{\pi^k}(S,a) \in \mathrm{Range}(\Sigma_h(\pi^k)) \qquad \text{a.s. under } S \sim \rho_h^{\pi^k}.$$

Using (16), decompose $\phi_h(S,a) = \phi_h(S,\pi_h^k(S)) + \Delta_h^{\pi^k}(S,a)$ and apply Lemma D.4 together with the ridge confidence bound. This yields, for a universal C_0,

$$\mathbb{E}\big[g_h^{\pi^k}(S_{k,h}) \mid \mathcal{F}_{k-1}\big] \le C_0\,\beta\,\widetilde{L}\,\sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)V_{k,h}^{-1}\big)}.$$

Summing over k,h and invoking Lemma D.1,

$$\mathrm{Reg}(K) \le C_0\,\beta\,\widetilde{L}\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{E}\Big[\sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)V_{k,h}^{-1}\big)}\Big].$$

Cauchy–Schwarz and the elliptical-potential identity

$$\sum_{k=1}^{K}\mathbb{E}\big[\mathrm{tr}(\Sigma_h(\pi^k)V_{k,h}^{-1})\big] = \sum_{k=1}^{K}\mathbb{E}\big[\|\phi_h(S_{k,h},A_{k,h})\|_{V_{k,h}^{-1}}^2\big] \le 2d\log\Big(1+\frac{K}{\lambda}\Big)$$

then give (15) after summing over h and absorbing constants. ■

Logarithmic regret ingredients.

Lemma D.5 (Witness count in a block) *Fix a suboptimal policy π and let h be the stage given by Assumption 6.8, so that $\mathbb{P}_{S\sim\rho_h^\pi}(g_h^\pi(S) \geq \gamma_0) \geq p_0$. In any contiguous block of n episodes that repeatedly plays π, with probability at least $1-\delta'$, the number N_{wit} of witness episodes satisfies*

$$N_{\mathrm{wit}} \geq \tfrac{p_0}{2} n,$$

provided $n \geq \frac{8}{p_0}\log(1/\delta')$.

Proof. Conditioned on playing π, the witness indicators are i.i.d. Bernoulli with mean at least p_0; apply Chernoff. ■

Lemma D.6 (Switching on a witnessed margin once confidence is small) *Fix a suboptimal policy π and the stage h from Assumption 6.8. If on episode k with $S_{k,h} = s$ we have $g_h^\pi(s) \geq \gamma_0$ and*

$$\max_{a\in\mathcal{A}} \left|\langle \Delta_h^\pi(s,a), \widehat{w}_{k,h} - w_h^\star\rangle\right| \leq \frac{\gamma_0}{4},$$

then the greedy action at (k,h) cannot equal $\pi_h(s)$.

Proof. Let $a^\star \in \arg\max_a Q_h^\star(s,a)$. Then

$$\widehat{Q}_{k,h}(s,a^\star) - \widehat{Q}_{k,h}(s,\pi_h(s)) \geq Q_h^\star(s,a^\star) - Q_h^\star(s,\pi_h(s)) - \frac{\gamma_0}{4} \geq \frac{3\gamma_0}{4} > 0,$$

so greediness forces a switch. ■

Lemma D.7 (Confidence shrinkage on witness episodes) *On the event in Lemma D.2, for any fixed π and stage h, every witness episode k obeys*

$$\max_{a\in\mathcal{A}} \left|\langle \Delta_h^\pi(S_{k,h},a), \widehat{w}_{k,h} - w_h^\star\rangle\right| \leq \beta_{K,\delta}\, L \sqrt{\mathrm{tr}\big(\Sigma_h(\pi) V_{k,h}^{-1}\big)}.$$

Proof. By self-identifiability, Lemma 6.3 gives $\Delta_h^\pi(S,a) \in \mathrm{Range}(\Sigma_h(\pi))$ a.s.; then apply Lemma D.4 and the ridge confidence bound. ■

Lemma D.8 (Block length bound for a fixed suboptimal policy) *There exists a universal constant C_1 such that, on the event in Lemma D.2, any contiguous block that repeatedly plays a fixed suboptimal π has length at most*

$$n(\pi) \leq C_1 \frac{d\,\beta_{K,\delta}^2\,\widetilde{L}^2}{p_0\,\gamma_0^2} \log\Big(\frac{KH}{\delta}\Big).$$

Proof. Let h be the stage from Assumption 6.8. Lemma D.5 gives $N_{\mathrm{wit}} \gtrsim p_0 n$ witness episodes in a block of length n. Summing the witness version of Lemma D.7 over that subsequence and using the elliptical-potential bound shows that some witness episode in the block satisfies

$$\mathrm{tr}\big(\Sigma_h(\pi) V_{k,h}^{-1}\big) \lesssim \frac{d\log(1+K/\lambda)}{p_0 n}.$$

If n exceeds the displayed bound, then on that witness episode the confidence radius is at most $\gamma_0/4$, and Lemma D.6 forces a deviation from π within the block. Hence the block cannot be longer. ■

Proof sketch of Theorem 6.9. On the event in Lemma D.2, every contiguous block of suboptimal play has length at most Lemma D.8. Each episode contributes at most H regret. A standard phase decomposition over confidence scales—equivalently, geometric decay of the trace budget across reappearances of the same policy—therefore yields

$$\mathrm{Reg}(K) \lesssim H \cdot \frac{d\,\beta_{K,\delta}^2\,\widetilde{L}^2}{p_0\gamma_0^2} \log\Big(\frac{KH}{\delta}\Big),$$

which is exactly the stated bound after substituting the definition of $\beta_{K,\delta}$ and absorbing H-factors into C'. ■

E Proof of Proposition 6.10

Proof. Consider the contextual-bandit case $H = 1$ with $\mathcal{S} = \{s\}$, $\mathcal{A} = \{0, 1\}$, $d = 2$, and

$$\phi_1(s, 0) = e_1, \qquad \phi_1(s, 1) = 2e_1.$$

Let rewards be deterministic with means $r_1(s, 0) = c$ and $r_1(s, 1) = 2c$ for some $c \in (0, 1/2]$. Bellman completeness is trivial, and the unique optimal policy chooses action 1. The suboptimal greedy-representable policy choosing action 0 is induced by $w_1 = 0$ under the fixed tie-break.

For this policy,

$$\Sigma_1(\pi) = e_1 e_1^\top,$$

which is rank-deficient with nullspace $\mathrm{span}\{e_2\}$. But the only action difference is

$$\Delta_1^\pi(s, 1) = \phi_1(s, 1) - \phi_1(s, 0) = e_1,$$

so every null direction is action-insensitive: $\mathrm{Null}(\Sigma_1(\pi)) \subseteq \mathrm{Null}(\Gamma_1(\pi))$. Hence (14) fails, and Theorem 5.4(1) rules out a decoy for π. Since this is the only suboptimal greedy-representable policy, the instance is self-identifiable despite rank deficiency. ■

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021. DOI: 10.1287/mnsc.2020.3605.

Luca Civitavecchia and Matteo Papini. Exploration-free reinforcement learning with linear function approximation. *Reinforcement Learning Journal*, 6:1856–1879, 2025. Presented at RLC 2025.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4171–4180. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/he21c.html.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 09–12 Jul 2020.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2 edition, 2002.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

Aleksandrs Slivkins, Yunzong Xu, and Shiliang Zuo. Greedy algorithm for structured bandits: A sharp characterization of asymptotic success / failure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Also available as arXiv:2503.04010.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10978–10989. PMLR, 13–18 Jul 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

Supplementary cover note (reader's guide)

This document provides auxiliary technical material supporting the main paper *"When Can Pure Exploitation Succeed in Linear RL? Decoys and Self-Identifiability for Greedy LSVI"*. All *central* claims (the decoy characterization, linear-regret construction, and the stated regret upper bounds under self-identifiability and a witnessed margin) are proved in the reviewed part of the main paper (i.e., before the references), in compliance with RLC/RLJ policy. The purpose of this supplement is to: (i) record measure-theoretic and linear-algebraic preliminaries used implicitly throughout; (ii) provide expanded intermediate lemmas, constant-tracking, and alternative proof routes that would not fit in the page-limited main text; and (iii) document additional examples and variants.

Organization.

- **S.1** fixes notation and formal measurability conventions (trajectory space, Markov kernels, filtrations), and records linear algebra facts about PSD matrices and pseudoinverses used repeatedly (e.g., for leverage bounds).
- Later sections (S.2+) (i) restate the exact concentration tools we use (self-normalized ridge bounds from Abbasi-Yadkori et al. (2011)); (ii) give expanded proofs of the RL-specific regression steps and bootstrapped-target handling; and (iii) include additional examples (e.g., alternative decoy constructions and robustness checks).

No new claims. No statement in this supplement is required to interpret the main contributions; rather, it supplies additional detail aimed at making the full technical development maximally checkable and self-contained.

S.1 Notation, measurability, and linear algebra facts

Indexing and basic conventions. For $n \in \mathbb{N}$, we write $[n] := \{1, 2, \dots, n\}$. Episodes are indexed by $k \in [K]$ and stages by $h \in [H]$. We write I_d for the $d \times d$ identity matrix. For vectors, $\|\cdot\|_2$ denotes the Euclidean norm. For a symmetric PSD matrix $A \succeq 0$, define $\|x\|_A := \sqrt{x^\top A x}$. The Moore–Penrose pseudoinverse is denoted $A^\dagger$. We use $\operatorname{ess\,sup}$ for the essential supremum with respect to the relevant measure (specified explicitly).

Measurable spaces and kernels. We assume throughout that $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ is a *standard Borel* measurable space and that $\mathcal{A}$ is a finite set equipped with its discrete σ-algebra. Rewards take values in $[0, 1]$ with its Borel σ-algebra.

An episodic MDP is

$$M = (\mathcal{S}, \mathcal{A}, H, \mu, \{P_h\}_{h=1}^H, \{\nu_h\}_{h=1}^H),$$

where μ is an initial distribution on $\mathcal{S}$, each $P_h(\cdot \mid s, a)$ is a Markov kernel from $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$, and each $\nu_h(\cdot \mid s, a)$ is a Markov kernel from $\mathcal{S} \times \mathcal{A}$ to $[0, 1]$. The mean reward is

$$r_h(s, a) := \mathbb{E}[R_h \mid S_h = s, A_h = a] = \int_{[0,1]} r \, \nu_h(dr \mid s, a).$$

(All conditional expectations and conditional kernels are well-defined under the standard Borel assumption.)

Trajectory space and trajectory laws. Let the trajectory space be

$$\Omega := (\mathcal{S} \times \mathcal{A} \times [0,1])^H \times \mathcal{S}$$

equipped with the product σ-algebra. For a (possibly nonstationary) Markov policy $\pi = (\pi_h)_{h=1}^H$, where each $\pi_h(\cdot \mid s)$ is a probability distribution on $\mathcal{A}$ and the map $s \mapsto \pi_h(a \mid s)$ is measurable for each $a \in \mathcal{A}$, the induced trajectory law $\mathbb{P}_M^\pi$ on $(\Omega, \mathcal{B}(\Omega))$ is the unique measure satisfying the standard product-kernel factorization (Ionescu–Tulcea construction):

$$\begin{aligned}&\mathbb{P}_M^\pi(ds_1, da_1, dr_1, \ldots, ds_{H+1}) \\ &\quad = \mu(ds_1) \prod_{h=1}^H \pi_h(da_h \mid s_h)\, \nu_h(dr_h \mid s_h, a_h)\, P_h(ds_{h+1} \mid s_h, a_h).\end{aligned}$$

We write $\mathbb{E}_M^\pi[\cdot]$ for expectation under $\mathbb{P}_M^\pi$.

Greedy policies are measurable under finite $\mathcal{A}$. Fix stage h and a weight vector $w \in \mathbb{R}^d$. Assume $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is jointly measurable. With a deterministic tie-breaking rule on $\mathcal{A}$ (e.g., smallest index), define

$$\pi_h^w(s) \in \arg\max_{a \in \mathcal{A}} \langle \phi_h(s,a), w \rangle \quad \text{(single-valued by tie-breaking)}.$$

Lemma S.1.1 (Measurability of tie-broken linear-greedy policies) *For finite $\mathcal{A}$ and measurable $\phi_h(\cdot, a)$, the map $s \mapsto \pi_h^w(s)$ is measurable.*

Proof. For each $a \in \mathcal{A}$, define the measurable score $f_a(s) := \langle \phi_h(s,a), w \rangle$. With deterministic tie-breaking, the set $\{s : \pi_h^w(s) = a\}$ can be written as a finite intersection of measurable sets of the form $\{s : f_a(s) > f_b(s)\}$ and $\{s : f_a(s) = f_b(s)\}$ (to implement the tie-breaking order). Since $f_a - f_b$ is measurable, each such set is measurable, and the finite intersection is measurable. ■

State marginals and covariances. For (M, π), let $\rho_{M,h}^\pi$ denote the marginal distribution of S_h under $\mathbb{P}_M^\pi$, and let $d_{M,h}^\pi$ denote the joint distribution of (S_h, A_h). If π is deterministic, then $A_h = \pi_h(S_h)$ almost surely, so $d_{M,h}^\pi$ is the pushforward of $\rho_{M,h}^\pi$ under $s \mapsto (s, \pi_h(s))$. Consequently, the (stage-wise) feature covariance defined in the main paper satisfies

$$\Sigma_{M,h}^\pi := \mathbb{E}_{(S,A) \sim d_{M,h}^\pi}\big[\phi_h(S,A)\phi_h(S,A)^\top\big] = \mathbb{E}_{S \sim \rho_{M,h}^\pi}\big[\phi_h(S,\pi_h(S))\phi_h(S,\pi_h(S))^\top\big] =: \Sigma_h(\pi)$$

for deterministic π (the latter equality matches the main-paper notation).

A clean conditional-kernel consequence of trajectory-law equality. The next fact is the formal backbone for steps of the form "$\mathbb{P}_M^\pi = \mathbb{P}_{M'}^\pi \Rightarrow$ on-policy reward/transition kernels agree ρ^π-a.s."

Lemma S.1.2 (Trajectory-law equality $\Rightarrow$ equality of on-policy conditional laws) *Fix a Markov policy π. If $\mathbb{P}_M^\pi = \mathbb{P}_{M'}^\pi$ as probability measures on Ω, then for each $h \in [H]$, for $\rho_{M,h}^\pi$-a.e. s, the regular conditional distribution of (R_h, S_{h+1}) given $S_h = s$ and $A_h = \pi_h(s)$ is identical under M and M'. Equivalently, for $\rho_{M,h}^\pi$-a.e. s,*

$$\nu_h(\cdot \mid s, \pi_h(s)) = \nu_h'(\cdot \mid s, \pi_h(s)) \quad \textit{and} \quad P_h(\cdot \mid s, \pi_h(s)) = P_h'(\cdot \mid s, \pi_h(s))$$

as probability measures on $[0,1]$ and $\mathcal{S}$, respectively.

Proof. Fix h. Since $\mathbb{P}_M^\pi = \mathbb{P}_{M'}^\pi$, all finite-dimensional marginals agree. In particular, the joint law of (S_h, A_h, R_h, S_{h+1}) under $\mathbb{P}_M^\pi$ equals the corresponding joint law under $\mathbb{P}_{M'}^\pi$. Because $\mathcal{S}$ is standard Borel and $\mathcal{A}$ is finite, regular conditional distributions of (R_h, S_{h+1}) given (S_h, A_h) exist and are unique up to $\mathbb{P}(S_h, A_h)$-null sets. Hence the regular conditional laws must coincide for $d_{M,h}^\pi$-a.e. (s,a). Specializing to $a = \pi_h(s)$ yields the claim (and for deterministic π, $d_{M,h}^\pi$ concentrates on $a = \pi_h(s)$). ■

Bounded noise $\Rightarrow$ conditional sub-Gaussianity. We will repeatedly invoke standard self-normalized martingale bounds for ridge regression (e.g., Abbasi-Yadkori et al. (2011)), which require conditionally sub-Gaussian noise. The following elementary fact (Hoeffding's lemma) will be used as a one-line justification.

Lemma S.1.3 (Conditional sub-Gaussianity for bounded variables) *Let $\mathcal{G}$ be a σ-field and let X be a real random variable such that $X \in [a, b]$ almost surely. Then $X - \mathbb{E}[X \mid \mathcal{G}]$ is conditionally σ-sub-Gaussian given $\mathcal{G}$ with $\sigma = \frac{b-a}{2}$, i.e.,*

$$\mathbb{E}\Big[\exp\Big(\lambda\big(X - \mathbb{E}[X \mid \mathcal{G}]\big)\Big) \,\Big|\, \mathcal{G}\Big] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \quad \text{a.s. for all } \lambda \in \mathbb{R}.$$

In particular, if $X \in [0, 1]$, then the noise is conditionally $(1/2)$-sub-Gaussian (and hence also 1-sub-Gaussian).

Linear algebra toolkit. We record a few facts about PSD matrices that are used throughout the proofs.

Lemma S.1.4 (Range–nullspace orthogonality for PSD matrices) *If $A \succeq 0$, then $\mathrm{Range}(A) = \mathrm{Null}(A)^{\perp}$. In particular, for any $x \in \mathbb{R}^d$,*

$$x \in \mathrm{Range}(A) \quad \Longleftrightarrow \quad u^\top x = 0 \ \ \forall u \in \mathrm{Null}(A).$$

Lemma S.1.5 (Projectors induced by the pseudoinverse) *If $A \succeq 0$, then $AA^\dagger$ is the orthogonal projector onto $\mathrm{Range}(A)$ and $I - AA^\dagger$ is the orthogonal projector onto $\mathrm{Null}(A)$. Moreover, $A^\dagger \succeq 0$ and $\mathrm{Range}(A^\dagger) = \mathrm{Range}(A)$.*

Lemma S.1.6 (Trace domination) *If $B \succeq 0$, then $\lambda_{\max}(B) \le \mathrm{tr}(B)$. Consequently, for any $v \in \mathbb{R}^d$,*

$$v^\top B v \le \|v\|_2^2 \, \mathrm{tr}(B).$$

Lemma S.1.7 (Matrix determinant lemma) *If $V \in \mathbb{R}^{d\times d}$ is invertible and $x \in \mathbb{R}^d$, then*

$$\det(V + xx^\top) = \det(V)\big(1 + x^\top V^{-1} x\big).$$

A reusable pseudoinverse-to-empirical-norm inequality. A recurring step in our regret analysis is to relate norms on the *on-policy estimable subspace* ($\mathrm{Range}(\Sigma)$) to empirical ridge norms (V^{-1}) via a trace term. For convenience, we record the exact inequality used in the main paper.

Lemma S.1.8 (Pseudoinverse control via a trace term) *Let $\Sigma \succeq 0$ and $V \succeq 0$. If $z \in \mathrm{Range}(\Sigma)$, then*

$$\|z\|_{V^{-1}} \le \|z\|_{\Sigma^\dagger} \sqrt{\mathrm{tr}\big(\Sigma V^{-1}\big)}.$$

Proof. Write $z = \Sigma^{1/2} v$ with $v = \Sigma^{\dagger/2} z$ (minimum-norm choice). Then

$$\|z\|_{V^{-1}}^2 = v^\top \Sigma^{1/2} V^{-1} \Sigma^{1/2} v \le \|v\|_2^2 \, \lambda_{\max}\big(\Sigma^{1/2} V^{-1} \Sigma^{1/2}\big) \le \|v\|_2^2 \, \mathrm{tr}\big(\Sigma^{1/2} V^{-1} \Sigma^{1/2}\big),$$

where the last inequality uses Lemma S.1.6. Finally, $\|v\|_2^2 = \|z\|_{\Sigma^\dagger}^2$ and $\mathrm{tr}(\Sigma^{1/2} V^{-1} \Sigma^{1/2}) = \mathrm{tr}(\Sigma V^{-1})$. ∎

S.2 Trajectory-law equality implies on-policy kernel equality

This section supplies measure-theoretic details behind Lemma 4.2 in the main text. The key point is that equality of *trajectory laws* under a fixed policy π implies equality of the *regular conditional distributions* of (R_h, S_{h+1}) given (S_h, A_h), for each stage h, on the (S_h, A_h)-support induced by executing π. We state and prove a clean version that applies to any Markov policy (deterministic or stochastic), and then specialize to the deterministic on-policy form used in the main paper.

Setup. Fix a standard Borel state space $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, a finite action set $\mathcal{A}$ with its discrete σ-algebra, and the reward space $([0,1], \mathcal{B}([0,1]))$. Let $M = (\mathcal{S}, \mathcal{A}, H, \mu, \{P_{M,h}\}_{h=1}^{H}, \{\nu_{M,h}\}_{h=1}^{H})$ and $M' = (\mathcal{S}, \mathcal{A}, H, \mu, \{P_{M',h}\}_{h=1}^{H}, \{\nu_{M',h}\}_{h=1}^{H})$ be two environments on the same measurable spaces (same $\mathcal{S}, \mathcal{A}, H, \mu$). For each h, define the *one-step joint observation kernel*

$$\kappa_{M,h}(B \mid s,a) := \int_{\mathcal{S}} \int_{[0,1]} \mathbf{1}\{(r,s') \in B\}\, \nu_{M,h}(dr \mid s,a)\, P_{M,h}(ds' \mid s,a), \qquad B \in \mathcal{B}([0,1] \times \mathcal{S}). \tag{S.1}$$

and similarly $\kappa_{M',h}$. (Equivalently, $\kappa_{M,h} = \nu_{M,h} \otimes P_{M,h}$ is a Markov kernel from $\mathcal{S} \times \mathcal{A}$ to $[0,1] \times \mathcal{S}$; see Section S.1 for conventions.)

Fix any (possibly nonstationary) Markov policy $\pi = (\pi_h)_{h=1}^{H}$, where each $\pi_h(\cdot \mid s)$ is a probability distribution on $\mathcal{A}$ and $s \mapsto \pi_h(\cdot \mid s)$ is measurable. Let $\mathbb{P}_M^\pi$ and $\mathbb{P}_{M'}^\pi$ denote the induced probability measures on the trajectory space $\Omega := (\mathcal{S} \times \mathcal{A} \times [0,1])^H \times \mathcal{S}$ (constructed e.g. via Ionescu–Tulcea).

For a fixed stage h, define the random variables

$$X_h := (S_h, A_h) \in \mathcal{S} \times \mathcal{A}, \qquad Y_h := (R_h, S_{h+1}) \in [0,1] \times \mathcal{S}.$$

Let $d_{M,h}^\pi$ denote the law of X_h under $\mathbb{P}_M^\pi$ (this is the same object as $d_{M,h}^\pi$ in the main paper), and let $\rho_{M,h}^\pi$ denote the law of S_h under $\mathbb{P}_M^\pi$.

Lemma S.2.1 (Trajectory-law equality ⇒ equality of on-policy kernels) *Fix a Markov policy π and a stage $h \in [H]$. If the trajectory laws agree,*

$$\mathbb{P}_M^\pi \;=\; \mathbb{P}_{M'}^\pi \quad \textit{as probability measures on } \Omega,$$

then the following hold.

1. ***Equality of one-step conditionals given** (S_h, A_h)**.** There exists a measurable set $\mathcal{N}_h \subseteq \mathcal{S} \times \mathcal{A}$ with $d_{M,h}^\pi(\mathcal{N}_h) = 0$ such that for all $(s,a) \notin \mathcal{N}_h$,*

$$\kappa_{M,h}(\cdot \mid s,a) \;=\; \kappa_{M',h}(\cdot \mid s,a) \quad \textit{as probability measures on } ([0,1] \times \mathcal{S}, \mathcal{B}([0,1] \times \mathcal{S})). \tag{S.2}$$

2. ***Deterministic-policy specialization.** If π is deterministic (so $A_h = \pi_h(S_h)$ a.s.), then there exists $\mathcal{N}_h^{\mathrm{st}} \subseteq \mathcal{S}$ with $\rho_{M,h}^\pi(\mathcal{N}_h^{\mathrm{st}}) = 0$ such that for all $s \notin \mathcal{N}_h^{\mathrm{st}}$,*

$$\kappa_{M,h}(\cdot \mid s, \pi_h(s)) \;=\; \kappa_{M',h}(\cdot \mid s, \pi_h(s)). \tag{S.3}$$

3. ***Consequences for rewards and transitions.** Under (S.2), the on-policy reward and transition kernels agree $d_{M,h}^\pi$-a.s.: for all $(s,a) \notin \mathcal{N}_h$,*

$$\nu_{M,h}(\cdot \mid s,a) = \nu_{M',h}(\cdot \mid s,a) \quad \textit{and} \quad P_{M,h}(\cdot \mid s,a) = P_{M',h}(\cdot \mid s,a).$$

In particular, the mean rewards satisfy $r_{M,h}(s,a) = r_{M',h}(s,a)$ for $d_{M,h}^\pi$-a.e. (s,a).

Proof. Let $P := \mathbb{P}_M^\pi$ and $P' := \mathbb{P}_{M'}^\pi$; by assumption $P = P'$. Fix h. Let μ_h denote the joint law of (X_h, Y_h) under P (equivalently under P'), and let $\mu_{h,X}$ denote its X_h-marginal. Then $\mu_{h,X} = d_{M,h}^\pi = d_{M',h}^\pi$ because $P = P'$ implies all marginals coincide.

Step 1: $\kappa_{M,h}$ and $\kappa_{M',h}$ are versions of the regular conditional law of Y_h given X_h. By the MDP generative process, conditional on $(S_h, A_h) = (s,a)$, the pair (R_h, S_{h+1}) is drawn from $\nu_{M,h}(\cdot \mid s,a) \otimes P_{M,h}(\cdot \mid s,a) = \kappa_{M,h}(\cdot \mid s,a)$, independently of the past. Equivalently, for every bounded measurable $f : [0,1] \times \mathcal{S} \to \mathbb{R}$,

$$\mathbb{E}_M^\pi\big[f(Y_h) \mid X_h\big] \;=\; \int f(y)\, \kappa_{M,h}(dy \mid X_h) \qquad P\text{-a.s.} \tag{S.4}$$

Thus $\kappa_{M,h}$ is a version of the regular conditional distribution of Y_h given X_h under P. The same reasoning yields the analogous statement for $\kappa_{M',h}$ under P'.

Step 2: uniqueness a.s. implies equality on the X_h-marginal. Since $\mathcal{S}$ is standard Borel and $\mathcal{A}$ is finite, the product space $\mathcal{S} \times \mathcal{A}$ is standard Borel, and $[0,1] \times \mathcal{S}$ is standard Borel. Hence regular conditional distributions exist and are $\mu_{h,X}$-a.s. unique (e.g., the disintegration theorem / existence and uniqueness of regular conditional probabilities on standard Borel spaces; see, e.g., Kallenberg (2002, Sec. 6)).

Now, $\kappa_{M,h}$ and $\kappa_{M',h}$ are both versions of the conditional law of Y_h given X_h for the *same* joint measure μ_h (because $P = P'$ implies μ_h is common). Therefore, there exists a measurable set $\mathcal{N}_h \subseteq \mathcal{S} \times \mathcal{A}$ with $\mu_{h,X}(\mathcal{N}_h) = 0$ such that (S.2) holds for all $(s,a) \notin \mathcal{N}_h$. This proves item (1).

Step 3: deterministic specialization. If π is deterministic, then $X_h = (S_h, \pi_h(S_h))$ P-a.s., so $d^{\pi}_{M,h}$ is supported on the graph $\{(s, \pi_h(s)) : s \in \mathcal{S}\}$. Define $\mathcal{N}_h^{\mathrm{st}} := \{s \in \mathcal{S} : (s, \pi_h(s)) \in \mathcal{N}_h\}$; measurability follows from measurability of π_h. Then $\rho^{\pi}_{M,h}(\mathcal{N}_h^{\mathrm{st}}) = d^{\pi}_{M,h}(\mathcal{N}_h) = 0$, and (S.3) holds for all $s \notin \mathcal{N}_h^{\mathrm{st}}$, proving item (2).

Step 4: equality of ν and P on-policy. Finally, if $\kappa_{M,h}(\cdot \mid s,a) = \kappa_{M',h}(\cdot \mid s,a)$ as measures on $[0,1] \times \mathcal{S}$, then their marginals coincide. For any Borel $B \subseteq [0,1]$,

$$\nu_{M,h}(B \mid s,a) = \kappa_{M,h}(B \times \mathcal{S} \mid s,a) = \kappa_{M',h}(B \times \mathcal{S} \mid s,a) = \nu_{M',h}(B \mid s,a),$$

and similarly $P_{M,h}(\cdot \mid s,a) = P_{M',h}(\cdot \mid s,a)$ by marginalizing over $[0,1]$. Taking expectations of R_h gives equality of mean rewards on-policy. ■

Remark S.2.2 (What this lemma does and does not claim) *Lemma S.2.1 is intentionally* on-policy*: it identifies kernels only on the $d^{\pi}_{M,h}$-support induced by executing π. No equality is implied (or needed) for (s,a) pairs that occur with zero probability under π. This is precisely the level of identification required in the necessity proof of Theorem 5.4: a decoy must preserve* only *what is observed under π.*

S.3 Model-class richness / decoy-realizability: concrete constructions

This section makes Assumption 5.3 ("geometry ⇒ existence of a decoy in the model class") maximally checkable by giving explicit constructions in concrete model classes. The main logical point is that the sufficiency direction of Theorem 5.4 only needs the following implication:

> if a deterministic π has an action-sensitive on-policy null direction ($\mathrm{Null}(\Sigma_h(\pi)) \not\subseteq \mathrm{Null}(\Gamma_h(\pi))$ for some h), then there exists $M' \in \mathcal{M}$ such that $\mathbb{P}^{\pi}_{M^\star} = \mathbb{P}^{\pi}_{M'}$ and π is globally optimal in M'.

Any concrete verification of this implication is necessarily *model-class dependent*: if the class $\mathcal{M}$ is too restrictive, then even strong geometric witnesses may not be realizable by an alternative environment in $\mathcal{M}$.

Two takeaways. (i) In broad "tabular/fully expressive" classes (which are Bellman-complete under tabular features), decoy-realizability is immediate: one can complete the environment off-policy so that *every* policy induces the same trajectory law as π, hence π is optimal. This yields a fully rigorous non-vacuity witness for Assumption 5.3. (ii) In more structured parametric classes (e.g., linear MDP models), verifying Assumption 5.3 amounts to checking a *closure property under off-policy edits* that preserve the π-trajectory law; see the discussion in Remark S.3.9.

S.3.1 A generic "policy-collapse" completion that makes π optimal

We first record a simple and fully explicit way to make a fixed policy π globally optimal while preserving its trajectory law. The construction is best viewed as a *completion* of an MDP off the π-support.

Definition S.3.1 (Policy-collapse of an MDP) *Fix an episodic MDP* $M = (\mathcal{S}, \mathcal{A}, H, \mu, \{P_h\}_{h=1}^H, \{\nu_h\}_{h=1}^H)$ *and a deterministic Markov policy* $\pi = (\pi_h)_{h=1}^H$. *Define the* policy-collapse *of* M *along* π *as the MDP*

$$\mathrm{Coll}(M, \pi) := \big(\mathcal{S}, \mathcal{A}, H, \mu, \{\widetilde{P}_h\}_{h=1}^H, \{\widetilde{\nu}_h\}_{h=1}^H\big)$$

whose kernels are, for every stage h, *state* $s \in \mathcal{S}$, *and action* $a \in \mathcal{A}$,

$$\widetilde{P}_h(\cdot \mid s, a) \;:=\; P_h(\cdot \mid s, \pi_h(s)), \qquad \widetilde{\nu}_h(\cdot \mid s, a) \;:=\; \nu_h(\cdot \mid s, \pi_h(s)). \tag{S.5}$$

Thus, $\mathrm{Coll}(M, \pi)$ *"forgets" actions: at each* (h, s), all *actions inherit the transition and reward kernels of the* π*-action at that* (h, s).

Lemma S.3.2 (Policy-collapse preserves the π-trajectory law) *For any MDP* M *and deterministic Markov policy* π,

$$\mathbb{P}_M^\pi \;=\; \mathbb{P}_{\mathrm{Coll}(M,\pi)}^\pi.$$

Proof. Under π, the action taken at stage h in state s is $a = \pi_h(s)$. By (S.5), the conditional law of (R_h, S_{h+1}) given $(S_h = s, A_h = \pi_h(s))$ in $\mathrm{Coll}(M, \pi)$ equals the corresponding conditional law in M. Since the initial distribution μ is unchanged, the induced joint law on the entire trajectory is identical. ■

Lemma S.3.3 (Policy-collapse makes π globally optimal) *For any MDP* M *and any (possibly history-dependent) policy* π',

$$\mathbb{P}_{\mathrm{Coll}(M,\pi)}^{\pi'} \;=\; \mathbb{P}_M^\pi.$$

In particular, all policies have the same value in $\mathrm{Coll}(M, \pi)$, *and hence* π *is globally optimal:* $V_{1,\mathrm{Coll}(M,\pi)}^\pi = V_{1,\mathrm{Coll}(M,\pi)}^\star$.

Proof. Fix any policy π' (possibly history-dependent). Under $\mathrm{Coll}(M, \pi)$, the conditional law of (R_h, S_{h+1}) given $S_h = s$ and *any* action $A_h = a$ is, by (S.5),

$$(R_h, S_{h+1}) \mid (S_h = s, A_h = a) \;\sim\; (\nu_h(\cdot \mid s, \pi_h(s)), P_h(\cdot \mid s, \pi_h(s))),$$

which is independent of a. Thus the entire controlled process is in fact *uncontrolled*: the joint law on trajectories does not depend on the policy. By Lemma S.3.2, this policy-independent trajectory law equals $\mathbb{P}_M^\pi$. Therefore, all policies have identical return in $\mathrm{Coll}(M, \pi)$, and π is (trivially) optimal. ■

Remark S.3.4 (When is policy-collapse available inside a given model class?) *Definition S.3.1 constructs an MDP on the same* $(\mathcal{S}, \mathcal{A}, H, \mu)$, *but it does* not *automatically preserve membership in a structured class* $\mathcal{M}$ *(e.g., a class specified by a fixed feature map and Bellman completeness). Thus, the relevant "richness" question is: for which classes* $\mathcal{M}$ *does* $M \in \mathcal{M} \Rightarrow \mathrm{Coll}(M, \pi) \in \mathcal{M}$ *hold for the policies* π *of interest? The next subsection gives a concrete setting where this closure is automatic.*

S.3.2 Verification in the tabular (fully expressive) Bellman-complete class

We now instantiate the above completion in a standard "fully expressive" class where linear Bellman completeness is automatic.

Tabular features. Fix finite $\mathcal{S}$ and finite $\mathcal{A}$. For each stage h, take the tabular (one-hot) feature map

$$\phi_h(s,a) \;=\; e_{(s,a)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

so that $\mathcal{Q}_h = \{Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ is the set of *all* real-valued action-value functions. In this case, Bellman completeness (4) holds for every bounded MDP on $(\mathcal{S}, \mathcal{A}, H, \mu)$, because the Bellman optimality operator maps $\mathcal{Q}_{h+1}$ into the set of all functions on $\mathcal{S} \times \mathcal{A}$. This is the canonical "tabular linear" special case (cf. the role of tabular realizability as a baseline in linear RL (Jin et al., 2020; Zanette et al., 2020)).

Proposition S.3.5 (Tabular class satisfies Assumption 5.3 (core decoy-realizability)) *Assume $\mathcal{S}$ and $\mathcal{A}$ are finite and $\phi_h(s,a) = e_{(s,a)}$ (tabular features) for all stages. Let $\mathcal{M}$ be the class of all episodic MDPs on $(\mathcal{S}, \mathcal{A}, H, \mu)$ whose reward kernels are supported on $[0,1]$. Then for every $M^\star \in \mathcal{M}$ and every deterministic Markov policy π that is suboptimal in $M^\star$, there exists a decoy $M' \in \mathcal{M}$ such that $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$ and π is globally optimal in M'. Consequently, the existence part of Assumption 5.3 holds (and is in fact stronger than stated).*

Proof. Fix $M^\star \in \mathcal{M}$ and deterministic Markov π that is suboptimal in $M^\star$. Define

$$M' \;:=\; \mathrm{Coll}(M^\star, \pi)$$

as in Definition S.3.1. By Lemma S.3.2, $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$. By Lemma S.3.3, π is globally optimal in M'.

It remains to verify $M' \in \mathcal{M}$. The reward kernels of M' are copied from those of $M^\star$ (cf. (S.5)), hence remain supported on $[0,1]$. Finally, under tabular features, linear Bellman completeness holds for *every* MDP on $(\mathcal{S}, \mathcal{A}, H, \mu)$, so $M' \in \mathcal{M}$. Since π is suboptimal in $M^\star$ by premise, M' is a decoy for π relative to $M^\star$ (Definition 4.3). ■

Remark S.3.6 (About the "action-sensitive null direction" trigger) *Proposition S.3.5 constructs a decoy for* any *suboptimal π in the tabular class, independently of whether* $\mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$ *is nonempty. This is consistent with Theorem 5.4(Necessity): whenever such a decoy exists, there must exist some stage h and some $u \neq 0$ with $u \in \mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$. In tabular features, this witness always exists for deterministic π at any visited state, since on-policy one-hot directions do not span alternative-action one-hot directions.*

S.3.3 Avoiding clipping in reward perturbations (when needed)

Some constructions (and some readers' intuition) benefit from additionally perturbing mean rewards in a linear manner along a direction u. A recurring technical pitfall is *clipping*: clipping rewards to $[0,1]$ generally destroys linear structure and is best avoided (cf. the discussion around Assumption 5.3 in the main paper).

The following elementary fact records a clean way to keep rewards bounded without clipping.

Lemma S.3.7 (Choosing a perturbation size without clipping) *Fix a stage h and a bounded mean reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0,1]$. Let $\psi : \mathcal{S} \times \mathcal{A} \to [-1,1]$ be any bounded measurable function. Suppose there is a measurable set $U \subseteq \mathcal{S} \times \mathcal{A}$ on which we intend to modify the mean reward, and define the* slack

$$\mathrm{slack}(U) \;:=\; \inf_{(s,a)\in U} \min\{r_h(s,a),\, 1 - r_h(s,a)\} \;\in [0, 1/2].$$

Then for any scalar α with $|\alpha| \le \mathrm{slack}(U)$, the perturbed mean reward

$$r'_h(s,a) \;:=\; \begin{cases} r_h(s,a) + \alpha\, \psi(s,a), & (s,a) \in U, \\ r_h(s,a), & (s,a) \notin U, \end{cases}$$

still satisfies $r'_h(s,a) \in [0,1]$ for all (s,a).

Proof. For $(s,a) \notin U$, nothing changes. For $(s,a) \in U$, we have $|\alpha\psi(s,a)| \le |\alpha| \le \text{slack}(U) \le r_h(s,a)$ and also $|\alpha\psi(s,a)| \le \text{slack}(U) \le 1 - r_h(s,a)$. Thus $r_h(s,a) - |\alpha\psi(s,a)| \ge 0$ and $r_h(s,a) + |\alpha\psi(s,a)| \le 1$, proving $r'_h(s,a) \in [0,1]$. ■

Remark S.3.8 (When does one need Lemma S.3.7?) *In our main results, we never need to clip rewards. If one wishes to explicitly realize an* infinitesimal *reward perturbation along a witness direction u (e.g., for interpretability of the decoy mechanism), Lemma S.3.7 is the clean way to keep rewards bounded: choose U to be the set of off-policy pairs that will be modified and set $\psi(s,a) = \langle \phi_h(s,a), u \rangle$, which is bounded in $[-1,1]$ under the standing feature norm assumption.*

S.3.4 Discussion: relation to standard linear MDP parameterizations

Remark S.3.9 (Linear MDP / linear-mixture models) *The standard "linear MDP" (or "linear mixture MDP") parameterizations underlying LSVI-UCB (Jin et al., 2020) are significantly more restrictive than the fully tabular class: rewards and/or transitions are constrained to be linear in a known feature map. In such models, verifying Assumption 5.3 becomes a* closure question*: given $(M^\star, \pi)$ and a null direction $u \in \text{Null}(\Sigma_h(\pi))$ that is action-sensitive, can one perturb the model parameters along u in a way that (i) leaves the π-trajectory law unchanged, but (ii) changes off-policy action values enough to make π globally optimal, all while preserving the model's parameter constraints (probability simplex constraints for transitions, reward bounds, etc.)?*

This is analogous in spirit to the role of "decoys" as feasible alternative instances in structured bandits (Slivkins et al., 2025), and to the use of Bellman-completeness/closure conditions in linear RL analyses (Zanette et al., 2020; Civitavecchia & Papini, 2025). The tabular construction above provides a clean non-vacuity witness; extending it to a specific parametric linear MDP family typically requires checking that the family admits sufficiently rich off-policy edits while holding the on-policy trajectory law fixed (often via "feature-orthogonal" perturbations of reward/transition parameters).

S.4 Alternate proof presentation for the decoy characterization

Goal and proof roadmap. This section gives a modular, "AC-checkable" proof of the decoy characterization (Theorem 5.4 in the main paper). The argument isolates the two distinct ingredients: (i) *trajectory-law equality under a fixed policy* pins down equality of on-policy kernels (Lemma 4.2, expanded in Section S.2); (ii) *linear Bellman completeness* implies that optimal action-values are linear in the known features and therefore differences between optimal Q-functions can be interpreted as *parameter directions*. This yields a short chain of lemmas establishing the necessity direction (decoy $\Rightarrow$ unexcited yet action-sensitive direction), which is the substantive part of the theorem. The sufficiency direction (unexcited yet action-sensitive direction $\Rightarrow$ decoy) reduces to the model-class richness / decoy-realizability condition verified by the explicit constructions in Section S.3. Conceptually, the logic parallels "decoy" obstructions for greedy learning in structured bandits (Slivkins et al., 2025), but the RL setting requires trajectory-level equivalence and stage-wise analysis.

S.4.1 Setup and notation

Fix a true environment $M^\star \in \mathcal{M}$ and a deterministic Markov policy π. For any environment $M \in \mathcal{M}$ and stage $h \in [H]$, we write $V^\star_{h,M}$ and $Q^\star_{h,M}$ for the optimal value and action-value functions. Under linear Bellman completeness (4) (see also Zanette et al., 2020; Jin et al., 2020), we have $Q^\star_{h,M} \in \mathcal{Q}_h$ for every h and every $M \in \mathcal{M}$, hence there exist (not necessarily unique) vectors $w^\star_{h,M} \in \mathbb{R}^d$ such that

$$Q^\star_{h,M}(s,a) = \langle \phi_h(s,a), w^\star_{h,M} \rangle \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H]. \tag{S.6}$$

Given two environments $M, M' \in \mathcal{M}$, define their stage-wise optimal differences

$$\delta Q_h := Q^\star_{h,M'} - Q^\star_{h,M} \in \mathcal{Q}_h, \qquad \delta V_h := V^\star_{h,M'} - V^\star_{h,M}. \tag{S.7}$$

Since $\delta Q_h \in \mathcal{Q}_h$, there exists $u_h \in \mathbb{R}^d$ with

$$\delta Q_h(s,a) = \langle \phi_h(s,a), u_h \rangle \qquad \forall (s,a),\ \forall h. \tag{S.8}$$

Throughout this section, the on-policy state marginal ρ_h^π and the matrices $\Sigma_h(\pi)$ and $\Gamma_h(\pi)$ are those defined in Section 5.1 of the main paper, always with respect to the trajectory law $\mathbb{P}^\pi_{M^\star}$.

S.4.2 Necessity: a decoy induces an unexcited but action-sensitive direction

We restate the necessity direction of Theorem 5.4 in a lemma-chain form.

Lemma S.4.1 (Existence of a "last disagreement" stage) *Assume π admits a decoy $M' \in \mathcal{M}$ relative to $M^\star$ (Definition 4.3). Let $\delta V_h := V^\star_{h,M'} - V^\star_{h,M^\star}$. Then the set*

$$\mathcal{H} := \Big\{ h \in [H] :\ \mathbb{P}_{S\sim\rho_h^\pi}\big(\delta V_h(S) \neq 0\big) > 0 \Big\}$$

is nonempty, and hence the index

$$h^\star := \max \mathcal{H}$$

is well-defined. Moreover,

$$\delta V_{h^\star+1}(S) = 0 \quad \text{a.s. under } S \sim \rho^\pi_{h^\star+1}, \tag{S.9}$$

with the convention $\delta V_{H+1} \equiv 0$.

Proof. Since M' is a decoy for π relative to $M^\star$, we have $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$ and π is optimal in M' but suboptimal in $M^\star$. Thus

$$V^\star_{1,M'} = V^\pi_{1,M'} = V^\pi_{1,M^\star} < V^\star_{1,M^\star},$$

so $\mathbb{E}_{S\sim\mu}[\delta V_1(S)] \neq 0$, implying $\mathbb{P}_{S\sim\mu}(\delta V_1(S) \neq 0) > 0$. Therefore $\mathcal{H}$ is nonempty and $h^\star$ is well-defined. By definition of $h^\star$, $\delta V_{h^\star+1}(S) = 0$ holds $\rho^\pi_{h^\star+1}$-a.s., which is (S.9). ■

Lemma S.4.2 (On-policy cancellation at the last disagreement stage) *Let $M' \in \mathcal{M}$ be observationally equivalent to $M^\star$ under π: $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$. Fix any stage $h \in [H]$ such that $\delta V_{h+1}(S) = 0$ ρ^π_{h+1}-a.s. Then*

$$\delta Q_h\big(S, \pi_h(S)\big) = 0 \quad \text{a.s. under } S \sim \rho^\pi_h. \tag{S.10}$$

Proof. By Lemma 4.2 (proved in full detail in Section S.2), trajectory-law equality $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$ implies that for ρ^π_h-a.e. s, the conditional law of (R_h, S_{h+1}) given $S_h = s$ and $A_h = \pi_h(s)$ is identical in $M^\star$ and M'. In particular, for ρ^π_h-a.e. s, the mean reward and transition kernels agree on-policy:

$$r_{h,M^\star}(s,\pi_h(s)) = r_{h,M'}(s,\pi_h(s)), \qquad P_{h,M^\star}(\cdot \mid s,\pi_h(s)) = P_{h,M'}(\cdot \mid s,\pi_h(s)).$$

Using the Bellman optimality identity (valid for both MDPs),

$$Q^\star_{h,M}(s,a) = r_{h,M}(s,a) + \mathbb{E}_{S'\sim P_{h,M}(\cdot|s,a)}\big[V^\star_{h+1,M}(S')\big],$$

we obtain for ρ^π_h-a.e. s:

$$\begin{aligned}
\delta Q_h(s,\pi_h(s)) &= \big(r_{h,M'} - r_{h,M^\star}\big)(s,\pi_h(s)) \\
&\quad + \mathbb{E}_{S'\sim P_{h,M'}(\cdot|s,\pi_h(s))}\big[V^\star_{h+1,M'}(S')\big] - \mathbb{E}_{S'\sim P_{h,M^\star}(\cdot|s,\pi_h(s))}\big[V^\star_{h+1,M^\star}(S')\big] \\
&= \mathbb{E}_{S'\sim P_{h,M^\star}(\cdot|s,\pi_h(s))}\big[\delta V_{h+1}(S')\big].
\end{aligned}$$

Now $\delta V_{h+1}(S_{h+1}) = 0$ holds almost surely under the π-trajectory law (because $S_{h+1} \sim \rho^\pi_{h+1}$), hence its conditional expectation given (S_h, A_h) is also 0 almost surely. Therefore the right-hand side above equals 0 for ρ^π_h-a.e. s, proving (S.10). ■

Lemma S.4.3 (Nullspace witness at the last disagreement stage) *In the setting of Lemma S.4.1, let $h^\star$ be the last disagreement stage and let $u := u_{h^\star}$ be the linear coefficient of $\delta Q_{h^\star}$ in (S.8). Then $u \neq 0$ and*

$$u \in \mathrm{Null}\big(\Sigma_{h^\star}(\pi)\big). \tag{S.11}$$

Proof. First, $\delta V_{h^\star}$ is nonzero on a set of positive $\rho^\pi_{h^\star}$-measure by definition of $h^\star$, so $\delta Q_{h^\star}$ cannot be identically 0 (otherwise $Q^\star_{h^\star,M'} \equiv Q^\star_{h^\star,M^\star}$ and hence $V^\star_{h^\star,M'} \equiv V^\star_{h^\star,M^\star}$). Thus $u \neq 0$.

Next, Lemma S.4.1 gives $\delta V_{h^\star+1} = 0$ $\rho^\pi_{h^\star+1}$-a.s., so Lemma S.4.2 implies $\delta Q_{h^\star}(S, \pi_{h^\star}(S)) = 0$ $\rho^\pi_{h^\star}$-a.s. Using (S.8), this is

$$\langle \phi_{h^\star}(S, \pi_{h^\star}(S)), u\rangle = 0 \quad \text{a.s. under } S \sim \rho^\pi_{h^\star},$$

and Lemma 5.1 in the main paper yields $u \in \mathrm{Null}(\Sigma_{h^\star}(\pi))$. ■

Lemma S.4.4 (Action-sensitivity of the witness) *In the setting of Lemma S.4.3, the same u satisfies*

$$u \notin \mathrm{Null}\big(\Gamma_{h^\star}(\pi)\big). \tag{S.12}$$

Proof. Let $B := \{s \in \mathcal{S} : \delta V_{h^\star}(s) \neq 0\}$, so $\rho^\pi_{h^\star}(B) > 0$ by definition of $h^\star$. Fix any $s \in B$. If $\delta Q_{h^\star}(s,a)$ were equal for all $a \in \mathcal{A}$, then (since $\delta Q_{h^\star}(s, \pi_{h^\star}(s)) = 0$ by Lemma S.4.2) we would have $\delta Q_{h^\star}(s,a) = 0$ for all a, which implies

$$\max_{a\in\mathcal{A}} Q^\star_{h^\star,M'}(s,a) = \max_{a\in\mathcal{A}} Q^\star_{h^\star,M^\star}(s,a) \quad \Rightarrow \quad \delta V_{h^\star}(s) = 0,$$

a contradiction to $s \in B$. Therefore, for each $s \in B$ there exists at least one action $a \in \mathcal{A}$ such that $\delta Q_{h^\star}(s,a) \neq \delta Q_{h^\star}(s, \pi_{h^\star}(s)) = 0$. Equivalently,

$$\langle u, \Delta^\pi_{h^\star}(s,a)\rangle = \delta Q_{h^\star}(s,a) - \delta Q_{h^\star}(s, \pi_{h^\star}(s)) \neq 0 \quad \text{for some } a,$$

on a set of s's of positive $\rho^\pi_{h^\star}$-measure. By Lemma 5.2 in the main paper, this implies $u \notin \mathrm{Null}(\Gamma_{h^\star}(\pi))$. ■

Necessity direction of Theorem 5.4 (completed). Assume π admits a decoy M' relative to $M^\star$. By Lemma S.4.1, define $h^\star$ as the last stage where δV_h differs on a set of positive ρ^π_h-measure. Let $u := u_{h^\star}$ be the coefficient in (S.8). Then Lemma S.4.3 gives $u \neq 0$ and $u \in \mathrm{Null}(\Sigma_{h^\star}(\pi))$, and Lemma S.4.4 gives $u \notin \mathrm{Null}(\Gamma_{h^\star}(\pi))$. This is exactly (14) with $h = h^\star$. ■

S.4.3 Sufficiency: from a geometric witness to a decoy

We record the sufficiency direction in the same modular style.

Lemma S.4.5 (Geometric witness ⇒ decoy under decoy-realizability) *Fix $M^\star \in \mathcal{M}$ and let $\pi \in \Pi_{\text{lin-greedy}}$ be suboptimal in $M^\star$. Assume there exist $h \in [H]$ and $u \neq 0$ such that $u \in \mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$. If Assumption 5.3 (decoy realizability) holds, then π admits a decoy relative to $M^\star$.*

Proof. Assumption 5.3 guarantees the existence of an environment $M' \in \mathcal{M}$ such that $\mathbb{P}^\pi_{M^\star} = \mathbb{P}^\pi_{M'}$ and π is globally optimal in M'. Since π is suboptimal in $M^\star$ by premise, M' satisfies Definition 4.3 and is a decoy for π. ■

Where the "RL-specific" difficulty lives. The necessity direction above uses only: (i) trajectory-law equality under a fixed policy (handled formally in Section S.2), and (ii) representability of $Q^\star$ under Bellman completeness (to convert functional equalities into linear-algebraic constraints). The sufficiency direction is exactly where *model-class closure* matters: in RL, preserving an entire trajectory law while modifying off-policy behavior requires explicit control of transitions/rewards away from the on-policy support, which is why we isolate this requirement as Assumption 5.3 and verify it via explicit constructions in Section S.3 (cf. linear MDP parameterizations in Jin et al., 2020 and Bellman-complete frameworks in Zanette et al., 2020).

S.5 Full regret proof details for the sublinear bound

This section gives a fully detailed, checkable proof of the gap-free (sublinear) regret bound for GREEDY-LSVI stated as Theorem 6.7 in the main paper. The main technical points addressed here are: (i) a rigorous handling of *bootstrapped Bellman targets* in LSVI; and (ii) a careful filtration/predictability setup enabling the use of self-normalized ridge bounds (Abbasi-Yadkori et al., 2011) in an adaptive RL data-collection process.

A standard bounded (clipped) implementation. To ensure bounded regression labels and hence bounded (sub-Gaussian) noise terms at every stage, we analyze the standard clipped variant of LSVI used throughout linear-RL regret analyses (e.g., Jin et al., 2020). Concretely, after computing $\widehat{w}_{k,h}$ by (7), define the clipped Q-estimate

$$\widehat{Q}^{\mathrm{clip}}_{k,h}(s,a) := \big[\langle \phi_h(s,a), \widehat{w}_{k,h}\rangle\big]_0^{H-h+1}, \qquad [x]_0^b := \min\{b, \max\{0,x\}\}, \tag{S.13}$$

and set $\widehat{V}_{k,h}(s) := \max_{a\in\mathcal{A}} \widehat{Q}^{\mathrm{clip}}_{k,h}(s,a)$. The greedy rollout uses $\widehat{Q}^{\mathrm{clip}}_{k,h}$ in the argmax (ties broken deterministically). Targets in (6) use this $\widehat{V}_{k,h+1}$. This guarantees $\widehat{V}_{k,h}(s) \in [0, H-h+1]$ and hence

$$Y^{(k)}_{\tau,h} = R_{\tau,h} + \widehat{V}_{k,h+1}(S_{\tau,h+1}) \in [0, H-h+1] \quad \text{a.s.} \tag{S.14}$$

(Without an explicit boundedness mechanism like (S.13), $\widehat{V}_{k,h}$ can in principle be unbounded, which breaks the regression concentration step.)

S.5.1 Filtrations and stage-wise regression view

Write $\mathcal{H}_k$ for the global history up to the end of episode k, i.e.,

$$\mathcal{H}_k := \sigma\Big(\{(S_{\tau,t}, A_{\tau,t}, R_{\tau,t}, S_{\tau,t+1}) : 1 \le \tau \le k,\ 1 \le t \le H\}\Big).$$

Within episode k, define the stage-h pre-observation filtration

$$\mathcal{F}_{k,h} := \sigma\Big(\mathcal{H}_{k-1},\ S_{k,1}, A_{k,1}, R_{k,1}, \ldots, S_{k,h}, A_{k,h}\Big),$$

so $x_{k,h} := \phi_h(S_{k,h}, A_{k,h})$ is $\mathcal{F}_{k,h}$-measurable and $(R_{k,h}, S_{k,h+1})$ are drawn *after* $\mathcal{F}_{k,h}$.

For each fixed stage h, it is convenient to treat episode index k as the "time" index of a linear regression problem with covariates $x_{k,h}$ and suitable labels.

S.5.2 True Bellman targets and a valid self-normalized event

Define the (unobserved) *true* Bellman targets

$$Y^\star_{k,h} := R_{k,h} + V^\star_{h+1}(S_{k,h+1}), \qquad V^\star_{H+1} \equiv 0. \tag{S.15}$$

Under bounded rewards $R_{k,h} \in [0,1]$ and finite horizon, $V^\star_{h+1} \in [0, H-h]$, hence $Y^\star_{k,h} \in [0, H-h+1]$ a.s.

Lemma S.5.1 (Conditional linearity of true targets) *For every stage h and every episode k,*

$$\begin{aligned}\mathbb{E}\big[Y^\star_{k,h} \mid \mathcal{F}_{k,h}\big] &= \mathbb{E}\big[R_{k,h} + V^\star_{h+1}(S_{k,h+1}) \mid S_{k,h}, A_{k,h}\big] \\ &= Q^\star_h(S_{k,h}, A_{k,h}) \\ &= \langle \phi_h(S_{k,h}, A_{k,h}), w^\star_h\rangle.\end{aligned}$$

Proof. The first equality is by definition of $\mathcal{F}_{k,h}$ and the Markov property. The second is the definition of $Q^\star_h$. The third uses realizability of $Q^\star_h$ by $\mathcal{Q}_h$ under linear Bellman completeness, and the definition of $w^\star_h$ as a representing vector (Assumption 6.6). ■

Lemma S.5.2 (Bounded MDS noise is sub-Gaussian) *Fix h. Define the noise*

$$\eta_{k,h} := Y^\star_{k,h} - \mathbb{E}[Y^\star_{k,h} \mid \mathcal{F}_{k,h}] = Y^\star_{k,h} - \langle x_{k,h}, w^\star_h \rangle.$$

Then $(\eta_{k,h})_{k\geq 1}$ is a martingale difference sequence w.r.t. $(\mathcal{F}_{k,h})_{k\geq 1}$ and satisfies $|\eta_{k,h}| \leq H-h+1$ a.s.; in particular, it is conditionally $(H-h+1)$-sub-Gaussian.

Proof. By construction, $\mathbb{E}[\eta_{k,h} \mid \mathcal{F}_{k,h}] = 0$. Since $Y^\star_{k,h} \in [0, H-h+1]$ a.s. and $\mathbb{E}[Y^\star_{k,h} \mid \mathcal{F}_{k,h}] \in [0, H-h+1]$, we have $|\eta_{k,h}| \leq H-h+1$ a.s. A bounded MDS is conditionally sub-Gaussian with the same scale (Hoeffding's lemma). ■

Now define the *ideal* ridge estimator that uses the true targets:

$$\widetilde{w}_{k,h} := \arg\min_{w\in\mathbb{R}^d} \sum_{\tau=1}^{k-1} \left(\langle x_{\tau,h}, w\rangle - Y^\star_{\tau,h}\right)^2 + \lambda\|w\|_2^2 \quad \Longleftrightarrow \quad \widetilde{w}_{k,h} = V_{k,h}^{-1} \sum_{\tau=1}^{k-1} x_{\tau,h} Y^\star_{\tau,h}. \tag{S.16}$$

Lemma S.5.3 (Self-normalized ridge event for the ideal estimator) *Fix $\delta \in (0,1)$. There is a universal constant $c > 0$ such that with probability at least $1-\delta$, simultaneously for all stages $h \in [H]$ and all $k \in \{1, \ldots, K\}$,*

$$\|\widetilde{w}_{k,h} - w^\star_h\|_{V_{k,h}} \leq c\Big((H-h+1)\sqrt{d\log\big(1+\tfrac{K}{\lambda}\big)} + (H-h+1)\sqrt{\log\big(\tfrac{H}{\delta}\big)} + \sqrt{\lambda}\,B\Big). \tag{S.17}$$

Proof. Fix h and apply a self-normalized ridge inequality for adaptive linear regression to the sequence $(x_{k,h}, Y^\star_{k,h})_{k\geq 1}$ with filtration $(\mathcal{F}_{k,h})$. Lemma S.5.1 gives the linear conditional mean, and Lemma S.5.2 gives conditional sub-Gaussian noise with scale $H-h+1$. This yields (S.17) for each fixed h with failure probability δ/H. A union bound over $h \in [H]$ gives overall failure probability at most δ. We use the standard determinant bound $\log \frac{\det(V_{k,h})}{\det(\lambda I)} \leq d\log(1+K/\lambda)$ (cf. Abbasi-Yadkori et al., 2011). ■

S.5.3 Bootstrapped targets as a bounded deterministic bias

The actual estimator $\widehat{w}_{k,h}$ uses the bootstrapped (clipped) labels $Y^{(k)}_{\tau,h} = R_{\tau,h} + \widehat{V}_{k,h+1}(S_{\tau,h+1})$. Define the target mismatch at stage h (for episode k):

$$\Delta^{(k)}_{\tau,h} := Y^{(k)}_{\tau,h} - Y^\star_{\tau,h} = \widehat{V}_{k,h+1}(S_{\tau,h+1}) - V^\star_{h+1}(S_{\tau,h+1}). \tag{S.18}$$

By clipping, $\widehat{V}_{k,h+1}(s) \in [0, H-h]$ and $V^\star_{h+1}(s) \in [0, H-h]$, hence

$$|\Delta^{(k)}_{\tau,h}| \leq H-h \leq H \qquad \text{a.s. for all } (\tau,k,h). \tag{S.19}$$

Since ridge regression is linear in the labels, we can write:

$$\widehat{w}_{k,h} - \widetilde{w}_{k,h} = V_{k,h}^{-1} \sum_{\tau=1}^{k-1} x_{\tau,h}\,\Delta^{(k)}_{\tau,h}. \tag{S.20}$$

Lemma S.5.4 (Deterministic bootstrap-bias control in $V_{k,h}$-norm) *For every k and h,*

$$\begin{aligned}
\|\widehat{w}_{k,h} - \widetilde{w}_{k,h}\|_{V_{k,h}} &= \left\| \sum_{\tau=1}^{k-1} x_{\tau,h}\,\Delta^{(k)}_{\tau,h} \right\|_{V_{k,h}^{-1}} \\
&\leq (H-h)\sqrt{\sum_{\tau=1}^{k-1} \|x_{\tau,h}\|^2_{V_{k,h}^{-1}}} \\
&\leq (H-h)\sqrt{2d\log\Big(1+\frac{K}{\lambda}\Big)}.
\end{aligned}$$

Proof. Let X be the $d \times (k-1)$ matrix with columns $x_{\tau,h}$, and let $\Delta \in \mathbb{R}^{k-1}$ collect $\Delta_{\tau,h}^{(k)}$. Then

$$\|X\Delta\|_{V_{k,h}^{-1}}^2 = \Delta^\top X^\top V_{k,h}^{-1} X\Delta \le \|\Delta\|_\infty^2 \operatorname{tr}\big(X^\top V_{k,h}^{-1} X\big) = \|\Delta\|_\infty^2 \sum_{\tau=1}^{k-1} \|x_{\tau,h}\|_{V_{k,h}^{-1}}^2,$$

where we used $a^\top A a \le \|a\|_\infty^2 \operatorname{tr}(A)$ for $A \succeq 0$. The bound $\|\Delta\|_\infty \le H-h$ is (S.19). Finally, $\sum_{\tau=1}^{k-1} \|x_{\tau,h}\|_{V_{k,h}^{-1}}^2 \le 2d\log(1+K/\lambda)$ by the standard elliptical-potential/determinant argument (e.g., Abbasi-Yadkori et al., 2011). ■

Combining Lemmas S.5.3 and S.5.4 yields a valid uniform confidence event for the *actual* estimator $\widehat{w}_{k,h}$ around $w_h^\star$.

Lemma S.5.5 (A uniform $V_{k,h}$-norm bound for $\widehat{w}_{k,h} - w_h^\star$) *Fix $\delta \in (0,1)$. With probability at least $1-\delta$, simultaneously for all $h \in [H]$ and $k \in \{1,\dots,K\}$,*

$$\begin{aligned} \|\widehat{w}_{k,h} - w_h^\star\|_{V_{k,h}} &\le \beta_h(\delta), \\ \beta_h(\delta) &:= c\Bigg((H-h+1)\sqrt{d\log\big(1+\tfrac{K}{\lambda}\big)} \\ &\qquad + (H-h+1)\sqrt{\log\big(\tfrac{H}{\delta}\big)} + \sqrt{\lambda}\,B \Bigg) \\ &\qquad + (H-h)\sqrt{2d\log\Big(1+\frac{K}{\lambda}\Big)}. \end{aligned} \tag{S.21}$$

In particular, $\beta_h(\delta) \le \beta(\delta)$ for

$$\beta(\delta) := C_1\, H\sqrt{d\log\Big(1+\frac{K}{\lambda}\Big)} + C_2\, H\sqrt{\log\Big(\frac{H}{\delta}\Big)} + C_3\sqrt{\lambda}\,B$$

with universal constants $C_1, C_2, C_3 > 0$.

Proof. Triangle inequality: $\|\widehat{w}_{k,h} - w_h^\star\|_{V_{k,h}} \le \|\widetilde{w}_{k,h} - w_h^\star\|_{V_{k,h}} + \|\widehat{w}_{k,h} - \widetilde{w}_{k,h}\|_{V_{k,h}}$. Apply Lemma S.5.3 and Lemma S.5.4. ■

S.5.4 From parameter confidence to one-step suboptimality

Fix episode k and stage h. Let π^k be the deployed greedy policy and write $S := S_{k,h}$. Let $a_k := \pi_h^k(S)$ and let $a^\star \in \arg\max_{a\in\mathcal{A}} Q_h^\star(S,a)$. A standard greedy comparison gives

$$Q_h^\star(S,a^\star) - Q_h^\star(S,a_k) \le 2\max_{a\in\mathcal{A}} \big|\widehat{Q}_{k,h}^{\mathrm{clip}}(S,a) - Q_h^\star(S,a)\big|. \tag{S.22}$$

(Clipping can only decrease the magnitude of $\widehat{Q}_{k,h}(S,a) - Q_h^\star(S,a)$ because $Q_h^\star(S,a) \in [0, H-h+1]$ for bounded rewards.)

Using realizability $Q_h^\star(\cdot,\cdot) = \langle \phi_h(\cdot,\cdot), w_h^\star\rangle$ and the definition $\widehat{Q}_{k,h}^{\mathrm{clip}}(S,a) = [\langle \phi_h(S,a), \widehat{w}_{k,h}\rangle]_0^{H-h+1}$, we get the pointwise bound

$$\big|\widehat{Q}_{k,h}^{\mathrm{clip}}(S,a) - Q_h^\star(S,a)\big| \le \big|\langle \phi_h(S,a), \widehat{w}_{k,h} - w_h^\star\rangle\big|. \tag{S.23}$$

Thus, it suffices to control $\max_a |\langle \phi_h(S,a), e_{k,h}\rangle|$ where $e_{k,h} := \widehat{w}_{k,h} - w_h^\star$.

On the high-probability event of Lemma S.5.5, for every a,

$$\big|\langle \phi_h(S,a), e_{k,h}\rangle\big| \le \|e_{k,h}\|_{V_{k,h}} \cdot \|\phi_h(S,a)\|_{V_{k,h}^{-1}} \le \beta(\delta)\, \|\phi_h(S,a)\|_{V_{k,h}^{-1}}. \tag{S.24}$$

The remaining task is to bound the *actionwise* $V_{k,h}^{-1}$-norms at $\rho_h^{\pi^k}$-typical states in a way that is compatible with self-identifiability and the leverage constant. We reuse the main paper's notation $\Sigma_h(\pi)$ for the on-policy covariance under π.

A corrected scope for the "action differences in span" step. Recall that Lemma 6.3 (main paper) is used only when π^k is suboptimal; when π^k is optimal, the per-episode regret contribution is 0 and no geometric input is required. Formally, define the suboptimal-episode set

$$\mathcal{K}_{\mathrm{sub}} := \{k \in [K] : V_1^\star > V_1^{\pi^k}\}.$$

For $k \notin \mathcal{K}_{\mathrm{sub}}$, the episode contributes 0 regret. For $k \in \mathcal{K}_{\mathrm{sub}}$, self-identifiability implies π^k admits no decoy, and hence Lemma 6.3 applies.

Lemma S.5.6 (Expected $V_{k,h}^{-1}$-norm of the on-policy feature) *Condition on $\mathcal{H}_{k-1}$ (so π^k and $V_{k,h}$ are fixed). Then*

$$\mathbb{E}\Big[\|\phi_h(S_{k,h}, \pi_h^k(S_{k,h}))\|^2_{V_{k,h}^{-1}} \;\Big|\; \mathcal{H}_{k-1}\Big] = \mathrm{tr}\big(\Sigma_h(\pi^k)\, V_{k,h}^{-1}\big).$$

Consequently,

$$\mathbb{E}\Big[\|\phi_h(S_{k,h}, \pi_h^k(S_{k,h}))\|_{V_{k,h}^{-1}} \;\Big|\; \mathcal{H}_{k-1}\Big] \le \sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)\, V_{k,h}^{-1}\big)}.$$

Proof. Under $\mathcal{H}_{k-1}$, $S_{k,h} \sim \rho_h^{\pi^k}$ and $A_{k,h} = \pi_h^k(S_{k,h})$ deterministically. Thus,

$$\mathbb{E}[\phi_h(S_{k,h}, A_{k,h})\phi_h(S_{k,h}, A_{k,h})^\top \mid \mathcal{H}_{k-1}] = \Sigma_h(\pi^k).$$

Taking trace after multiplying by $V_{k,h}^{-1}$ yields the first claim. The second follows from Jensen. ■

Lemma S.5.7 (Action-difference V^{-1}-norm control via leverage) *Fix $k \in \mathcal{K}_{\mathrm{sub}}$ and condition on $\mathcal{H}_{k-1}$. Then for $\rho_h^{\pi^k}$-a.e. s and all $a \in \mathcal{A}$,*

$$\|\Delta_h^{\pi^k}(s,a)\|_{V_{k,h}^{-1}} \le \|\Delta_h^{\pi^k}(s,a)\|_{\Sigma_h(\pi^k)^\dagger} \sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)\, V_{k,h}^{-1}\big)} \le L \sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)\, V_{k,h}^{-1}\big)}.$$

Proof. For $k \in \mathcal{K}_{\mathrm{sub}}$, π^k is suboptimal and self-identifiability implies π^k admits no decoy. Lemma 6.3 therefore yields $\Delta_h^{\pi^k}(S,a) \in \mathrm{Range}(\Sigma_h(\pi^k))$ a.s. under $S \sim \rho_h^{\pi^k}$. Applying Lemma D.4 from Appendix D (main paper) with $\Sigma = \Sigma_h(\pi^k)$, $V = V_{k,h}$, and $z = \Delta_h^{\pi^k}(s,a)$ gives the first inequality. The second inequality is the definition of L. ■

We can now bound the expected one-step suboptimality.

Lemma S.5.8 (One-step regret bound in terms of $\mathrm{tr}(\Sigma V^{-1})$) *Work on the high-probability event of Lemma S.5.5. Fix $k \in [K]$ and condition on $\mathcal{H}_{k-1}$. For each stage $h \in [H]$,*

$$\mathbb{E}\Big[g_h^{\pi^k}(S_{k,h}) \;\Big|\; \mathcal{H}_{k-1}\Big] \le 2\,\beta(\delta)\,(1+L)\,\sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)\, V_{k,h}^{-1}\big)}.$$

Proof. If $k \notin \mathcal{K}_{\mathrm{sub}}$, then $g_h^{\pi^k}(S_{k,h}) = 0$ a.s. for all h and the bound is trivial. Assume $k \in \mathcal{K}_{\mathrm{sub}}$.

Let $S := S_{k,h}$. Using (S.22)–(S.23) and the decomposition $\phi_h(S,a) = \phi_h(S, \pi_h^k(S)) + \Delta_h^{\pi^k}(S,a)$,

$$\begin{aligned} g_h^{\pi^k}(S) &\le 2 \max_{a \in \mathcal{A}} \big|\langle \phi_h(S,a), e_{k,h}\rangle\big| \\ &\le 2\big|\langle \phi_h(S, \pi_h^k(S)), e_{k,h}\rangle\big| + 2 \max_{a \in \mathcal{A}} \big|\langle \Delta_h^{\pi^k}(S,a), e_{k,h}\rangle\big|. \end{aligned}$$

By (S.24),

$$\begin{aligned} \big|\langle \phi_h(S, \pi_h^k(S)), e_{k,h}\rangle\big| &\le \beta(\delta)\, \|\phi_h(S, \pi_h^k(S))\|_{V_{k,h}^{-1}}, \\ \big|\langle \Delta_h^{\pi^k}(S,a), e_{k,h}\rangle\big| &\le \beta(\delta)\, \|\Delta_h^{\pi^k}(S,a)\|_{V_{k,h}^{-1}}. \end{aligned}$$

Take conditional expectations and apply Lemmas S.5.6 and S.5.7:

$$\mathbb{E}[g_h^{\pi^k}(S_{k,h}) \mid \mathcal{H}_{k-1}] \le 2\beta(\delta)\Big(1 + L\Big)\sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k) V_{k,h}^{-1}\big)}.$$

■

S.5.5 Summing over episodes and stages

We now conclude the proof of Theorem 6.7.

Proof of Theorem 6.7 (full details). Fix $\delta \in (0,1)$ and work on the event in Lemma S.5.5, which holds with probability at least $1-\delta$.

By the standard finite-horizon performance decomposition (main paper, Section 6.3),

$$\mathrm{Reg}(K) = \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}\big[g_h^{\pi^k}(S_{k,h})\big].$$

Applying Lemma S.5.8 and the tower property,

$$\mathrm{Reg}(K) \le 2\beta(\delta)(1+L)\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{E}\left[\sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)\,V_{k,h}^{-1}\big)}\right]. \tag{S.25}$$

Use Cauchy–Schwarz in k for each fixed h:

$$\sum_{k=1}^{K}\mathbb{E}\left[\sqrt{\mathrm{tr}\big(\Sigma_h(\pi^k)\,V_{k,h}^{-1}\big)}\right] \le \sqrt{K}\,\sqrt{\sum_{k=1}^{K}\mathbb{E}\Big[\mathrm{tr}\big(\Sigma_h(\pi^k)\,V_{k,h}^{-1}\big)\Big]}.$$

Next observe that, conditional on $\mathcal{H}_{k-1}$,

$$\mathrm{tr}\big(\Sigma_h(\pi^k)\,V_{k,h}^{-1}\big) = \mathbb{E}\Big[\|\phi_h(S_{k,h},A_{k,h})\|^2_{V_{k,h}^{-1}} \;\Big|\; \mathcal{H}_{k-1}\Big],$$

so by the tower property,

$$\sum_{k=1}^{K}\mathbb{E}\Big[\mathrm{tr}\big(\Sigma_h(\pi^k)\,V_{k,h}^{-1}\big)\Big] = \sum_{k=1}^{K}\mathbb{E}\Big[\|\phi_h(S_{k,h},A_{k,h})\|^2_{V_{k,h}^{-1}}\Big].$$

Finally, the elliptical-potential/determinant bound yields the pathwise inequality

$$\sum_{k=1}^{K}\|\phi_h(S_{k,h},A_{k,h})\|^2_{V_{k,h}^{-1}} \le 2\log\frac{\det(V_{K+1,h})}{\det(\lambda I)} \le 2d\log\Big(1+\frac{K}{\lambda}\Big)$$

(e.g., Abbasi-Yadkori et al., 2011), hence taking expectations preserves the bound.

Plugging back into (S.25) gives

$$\mathrm{Reg}(K) \le 2\beta(\delta)(1+L)\sum_{h=1}^{H}\sqrt{K}\,\sqrt{2d\log\Big(1+\frac{K}{\lambda}\Big)} \le C\,H\,(1+L)\,\beta(\delta)\,\sqrt{dK\log\Big(1+\frac{K}{\lambda}\Big)}$$

for a universal constant $C>0$. Substituting the explicit form of $\beta(\delta)$ from Lemma S.5.5 and using $\log(1+K/\lambda) \le \log(1+KH/\lambda)$ yields a bound of the form stated in Theorem 6.7 (up to universal constants and the harmless replacement $L \mapsto 1+L$). ■

Remark S.5.9 (Where self-identifiability enters) *In the above proof, the only use of self-identifiability (via Theorem 5.4 and Assumption 5.3) is through Lemma S.5.7, which asserts that on-policy action differences lie in the estimable subspace and can be controlled by the leverage parameter L. Everything else is standard ridge concentration and elliptical-potential calculus (Abbasi-Yadkori et al., 2011).*

S.6 Full logarithmic-regret proof under the witnessed-margin condition

This section provides a complete proof of Theorem 6.9. The argument is a *mistake bound*: under Assumption 6.8, every *suboptimal* greedy-representable policy π must (with probability at least p_0 in an episode) visit a state at some stage where there exists an alternative action improving $Q_h^\star$ by at least γ_0. On a high-probability confidence event for stage-wise ridge estimates (Section S.5), choosing the suboptimal greedy action at such a witnessed state can only occur if a certain on-policy "information" quantity $\mathrm{tr}(\Sigma_h(\pi)V_{k,h}^{-1})$ is still large. The cumulative amount of this information is controlled via an elliptical-potential argument and a martingale concentration step (cf. the linear bandit machinery in Abbasi-Yadkori et al., 2011).

S.6.1 Setup and a high-probability confidence event

For each episode $k \in \{1, \dots, K\}$, let $\mathcal{F}_{k-1}$ denote the σ-field generated by the entire history up to the end of episode $k-1$, i.e.,

$$\mathcal{F}_{k-1} := \sigma\Big(\{(S_{\tau,h}, A_{\tau,h}, R_{\tau,h}, S_{\tau,h+1}) : \tau \le k-1,\ h \in [H]\}\Big).$$

Then π^k and $V_{k,h}$ are $\mathcal{F}_{k-1}$-measurable for each h, and $(S_{k,h}, A_{k,h}, R_{k,h})_{h=1}^H$ are conditionally independent of $\mathcal{F}_{k-1}$ given π^k.

We will work on the standard stage-wise ridge confidence event for GREEDY-LSVI established in Section S.5. For convenience, we restate it in the form we need here.

Lemma S.6.1 (Uniform linear prediction bound for GREEDY-LSVI (from Section S.5)) *Assume bounded rewards in* $[0,1]$*, bounded features* $\|\phi_h(s,a)\|_2 \le 1$*, linear Bellman completeness, and Assumption 6.6. Fix* $\lambda > 0$ *and* $\delta \in (0,1)$*. There exists a (problem-independent) constant* $c > 0$ *such that with probability at least* $1-\delta$*, simultaneously for all* $k \in [K]$*,* $h \in [H]$*, and all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*,*

$$\big|\widehat{Q}_{k,h}(s,a) - Q_h^\star(s,a)\big| = \big|\langle \phi_h(s,a), \widehat{w}_{k,h} - w_h^\star\rangle\big| \ \le\ \beta(\delta)\,\|\phi_h(s,a)\|_{V_{k,h}^{-1}}, \tag{S.26}$$

where

$$\beta(\delta) := c\,H\Big(\sqrt{d\log\big(1+\tfrac{K}{\lambda}\big)} + \sqrt{\lambda}\,B + \sqrt{\log\big(\tfrac{H}{\delta}\big)}\Big). \tag{S.27}$$

Throughout the remainder of this section, fix $\delta \in (0,1)$ and condition on the event in Lemma S.6.1 with confidence level $\delta/6$; denote this event by $\mathcal{E}_{\mathrm{conf}}$. All statements below are on $\mathcal{E}_{\mathrm{conf}}$ unless otherwise noted.

S.6.2 Witnessed mistakes force a large trace term

Fix an episode k and a stage h. Define the on-policy feature at stage h in episode k:

$$x_{k,h} := \phi_h(S_{k,h}, A_{k,h}) = \phi_h(S_{k,h}, \pi_h^k(S_{k,h})).$$

Recall $\Sigma_h(\pi^k) = \mathbb{E}_{S\sim\rho_h^{\pi^k}}[\phi_h(S,\pi_h^k(S))\phi_h(S,\pi_h^k(S))^\top]$ and define

$$\mu_{k,h} := \mathrm{tr}\big(\Sigma_h(\pi^k)V_{k,h}^{-1}\big).$$

Since $V_{k,h}$ and π^k are $\mathcal{F}_{k-1}$-measurable and $S_{k,h} \sim \rho_h^{\pi^k}$ conditionally on $\mathcal{F}_{k-1}$, we have the identity

$$\mathbb{E}\Big[\|x_{k,h}\|^2_{V_{k,h}^{-1}}\,\Big|\,\mathcal{F}_{k-1}\Big] = \mathbb{E}\Big[x_{k,h}^\top V_{k,h}^{-1}x_{k,h}\,\Big|\,\mathcal{F}_{k-1}\Big] = \mathrm{tr}\Big(V_{k,h}^{-1}\mathbb{E}[x_{k,h}x_{k,h}^\top \mid \mathcal{F}_{k-1}]\Big) = \mu_{k,h}. \tag{S.28}$$

We now show that, under the witnessed-margin condition, whenever GREEDY-LSVI executes a *suboptimal* policy and a witness state occurs, the corresponding $\mu_{k,h}$ must be large.

Lemma S.6.2 (Witnessed margin ⇒ large $\mu_{k,h}$ on $\mathcal{E}_{\mathrm{conf}}$) *Assume Assumption 5.3, self-identifiability of $M^\star$ (Definition 4.4), and $L < \infty$ (Definition 6.4). Fix an episode k such that π^k is suboptimal in $M^\star$. Let $h \in [H]$ and $s \in \mathcal{S}$ satisfy $g_h^{\pi^k}(s) \geq \gamma_0$ and define an improving action*

$$a^\star(s) \in \arg\max_{a \in \mathcal{A}} \big(Q_h^\star(s,a) - Q_h^\star(s, \pi_h^k(s))\big).$$

On the event $\{S_{k,h} = s\} \cap \mathcal{E}_{\mathrm{conf}}$, we have

$$\mu_{k,h} \;\geq\; \left(\frac{\gamma_0}{\beta(\delta/6)\, L}\right)^2. \tag{S.29}$$

Proof. Fix k, h, s and set $\pi := \pi^k$ for readability. On $\{S_{k,h} = s\}$ the greedy action taken is $A_{k,h} = \pi_h(s)$. By definition of $a^\star(s)$ and the witness condition, we have

$$Q_h^\star(s, a^\star(s)) - Q_h^\star(s, \pi_h(s)) \;\geq\; \gamma_0. \tag{S.30}$$

Since $\pi_h(s)$ is greedy w.r.t. $\widehat{Q}_{k,h}$,

$$\widehat{Q}_{k,h}(s, \pi_h(s)) \;\geq\; \widehat{Q}_{k,h}(s, a^\star(s)),$$

so

$$\begin{aligned} 0 &\geq \widehat{Q}_{k,h}(s, a^\star(s)) - \widehat{Q}_{k,h}(s, \pi_h(s)) \\ &= Q_h^\star(s, a^\star(s)) - Q_h^\star(s, \pi_h(s)) \\ &\quad + \langle \Delta_h^\pi(s, a^\star(s)), \widehat{w}_{k,h} - w_h^\star \rangle, \end{aligned}$$

where $\Delta_h^\pi(s,a) = \phi_h(s,a) - \phi_h(s, \pi_h(s))$. Rearranging and using (S.30) yields

$$|\langle \Delta_h^\pi(s, a^\star(s)), \widehat{w}_{k,h} - w_h^\star \rangle| \;\geq\; \gamma_0. \tag{S.31}$$

On $\mathcal{E}_{\mathrm{conf}}$, applying (S.26) to the linear functional $x = \Delta_h^\pi(s, a^\star(s))$ gives

$$|\langle \Delta_h^\pi(s, a^\star(s)), \widehat{w}_{k,h} - w_h^\star \rangle| \leq \beta(\delta/6)\, \|\Delta_h^\pi(s, a^\star(s))\|_{V_{k,h}^{-1}},$$

hence

$$\|\Delta_h^\pi(s, a^\star(s))\|_{V_{k,h}^{-1}} \;\geq\; \frac{\gamma_0}{\beta(\delta/6)}. \tag{S.32}$$

Now use self-identifiability to certify that Δ_h^π lies in the on-policy span. Because $M^\star$ is self-identifiable, any suboptimal $\pi \in \Pi_{\text{lin-greedy}}$ admits no decoy. Thus Lemma 6.3 applies (on suboptimal π), giving $\Delta_h^\pi(S, a) \in \mathrm{Range}(\Sigma_h(\pi))$ a.s. under $S \sim \rho_h^\pi$. In particular, for s in the witness event support, $\Delta_h^\pi(s, a^\star(s)) \in \mathrm{Range}(\Sigma_h(\pi))$. Applying Lemma D.4 (with $\Sigma = \Sigma_h(\pi)$ and $V = V_{k,h}$) yields

$$\|\Delta_h^\pi(s, a^\star(s))\|_{V_{k,h}^{-1}} \leq \|\Delta_h^\pi(s, a^\star(s))\|_{\Sigma_h(\pi)^\dagger} \sqrt{\mathrm{tr}(\Sigma_h(\pi) V_{k,h}^{-1})} \leq L \sqrt{\mu_{k,h}},$$

where the last inequality uses the definition of L (Definition 6.4). Combining with (S.32) gives (S.29). ∎

S.6.3 Counting witnessed mistakes via an information budget

Define the indicator of suboptimality at episode k:

$$I_k := \mathbf{1}\Big\{ V_{1,M^\star}^{\pi^k} < V_{1,M^\star}^\star \Big\}.$$

For each episode k with $I_k = 1$, Assumption 6.8 guarantees the existence of a stage $h_k \in [H]$ such that

$$\mathbb{P}\Big(g_{h_k}^{\pi^k}(S_{k,h_k}) \geq \gamma_0 \,\Big|\, \mathcal{F}_{k-1}\Big) \geq p_0. \tag{S.33}$$

(When $I_k = 0$, define $h_k := 1$ arbitrarily.) Define the witness indicator

$$W_k := I_k \cdot \mathbf{1}\Big\{g_{h_k}^{\pi^k}(S_{k,h_k}) \geq \gamma_0\Big\}.$$

Then $W_k \in \{0, 1\}$ is $\mathcal{F}_k$-measurable and by (S.33) we have

$$\mathbb{E}[W_k \mid \mathcal{F}_{k-1}] \geq p_0 I_k. \tag{S.34}$$

Lemma S.6.2 implies that on $\mathcal{E}_{\text{conf}}$, every time $W_k = 1$ we must have

$$\mu_{k,h_k} \;\geq\; \eta^2, \qquad \eta := \frac{\gamma_0}{\beta(\delta/6)\, L}. \tag{S.35}$$

Therefore,

$$\sum_{k=1}^{K} W_k \leq \frac{1}{\eta^2} \sum_{k=1}^{K} \mu_{k,h_k} \leq \frac{1}{\eta^2} \sum_{h=1}^{H} \sum_{k=1}^{K} \mu_{k,h}, \tag{S.36}$$

since $h_k \in [H]$ for each k.

We now upper bound $\sum_{k=1}^{K} \mu_{k,h}$ for each stage h by combining (i) a deterministic elliptical-potential bound on realized squared norms and (ii) a martingale deviation bound that compares realized squared norms to their conditional expectations.

Lemma S.6.3 (Stage-wise information budget) *Fix a stage $h \in [H]$ and define $X_{k,h} := \|x_{k,h}\|^2_{V_{k,h}^{-1}}$ and $\mu_{k,h} := \mathbb{E}[X_{k,h} \mid \mathcal{F}_{k-1}]$ as in (S.28). Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sum_{k=1}^{K} \mu_{k,h} \;\leq\; 4d \log\Big(1 + \frac{K}{\lambda}\Big) \;+\; \frac{14}{3\lambda} \log\Big(\frac{1}{\delta}\Big). \tag{S.37}$$

Proof. We start with the deterministic elliptical-potential bound for the realized norms:

$$\sum_{k=1}^{K} X_{k,h} = \sum_{k=1}^{K} x_{k,h}^\top V_{k,h}^{-1} x_{k,h} \leq 2 \log \frac{\det(V_{K+1,h})}{\det(\lambda I)} \leq 2d \log\Big(1 + \frac{K}{\lambda}\Big), \tag{S.38}$$

where the last inequality uses $\|x_{k,h}\|_2 \leq 1$ and the standard determinant upper bound $\det(\lambda I + \sum_{k=1}^{K} x_{k,h} x_{k,h}^\top) \leq (\lambda + K)^d$ (see, e.g., Abbasi-Yadkori et al., 2011).

Now define the martingale difference sequence

$$D_k := \mu_{k,h} - X_{k,h}, \qquad M_n := \sum_{k=1}^{n} D_k.$$

Then $(M_n)_{n \geq 0}$ is a martingale w.r.t. $(\mathcal{F}_n)_{n \geq 0}$ since $\mathbb{E}[D_k \mid \mathcal{F}_{k-1}] = 0$. Moreover, since $V_{k,h} \succeq \lambda I$ and $\|x_{k,h}\|_2 \leq 1$, we have $0 \leq X_{k,h} \leq 1/\lambda$ and $0 \leq \mu_{k,h} \leq 1/\lambda$, hence

$$|D_k| \leq \frac{1}{\lambda} \qquad \text{a.s.} \tag{S.39}$$

Let the predictable quadratic variation be

$$V_n := \sum_{k=1}^{n} \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}].$$

Using $D_k^2 \le X_{k,h}^2 + \mu_{k,h}^2$ and the bound $z^2 \le \frac{1}{\lambda} z$ for $z \in [0, 1/\lambda]$, we obtain

$$\mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}] \le \mathbb{E}[X_{k,h}^2 \mid \mathcal{F}_{k-1}] + \mu_{k,h}^2 \le \frac{1}{\lambda}\mathbb{E}[X_{k,h} \mid \mathcal{F}_{k-1}] + \frac{1}{\lambda}\mu_{k,h} = \frac{2}{\lambda}\mu_{k,h}.$$

Therefore,

$$V_K \le \frac{2}{\lambda}\sum_{k=1}^{K} \mu_{k,h}. \tag{S.40}$$

We now apply Freedman's inequality for martingales with bounded increments: for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$M_K \le \sqrt{2V_K \log\Big(\frac{1}{\delta}\Big)} + \frac{1}{3\lambda}\log\Big(\frac{1}{\delta}\Big), \tag{S.41}$$

using (S.39) (increment bound $1/\lambda$).

On the event (S.41), combine (S.40) with $M_K = \sum_{k=1}^{K}(\mu_{k,h} - X_{k,h})$ to obtain

$$\sum_{k=1}^{K}\mu_{k,h} \le \sum_{k=1}^{K} X_{k,h} + \sqrt{\frac{4}{\lambda}\Big(\sum_{k=1}^{K}\mu_{k,h}\Big)\log\Big(\frac{1}{\delta}\Big)} + \frac{1}{3\lambda}\log\Big(\frac{1}{\delta}\Big).$$

Let $S := \sum_{k=1}^{K}\mu_{k,h}$, $A := \sum_{k=1}^{K} X_{k,h} + \frac{1}{3\lambda}\log(\frac{1}{\delta})$, and $C := \frac{4}{\lambda}\log(\frac{1}{\delta})$. Then $S \le A + \sqrt{CS}$. Writing $y = \sqrt{S}$ gives $y^2 - \sqrt{C}\,y - A \le 0$, hence

$$y \le \frac{\sqrt{C} + \sqrt{C + 4A}}{2} \le \sqrt{C} + \sqrt{A},$$

and therefore $S = y^2 \le 2C + 2A$. Thus,

$$\sum_{k=1}^{K}\mu_{k,h} \le 2\sum_{k=1}^{K} X_{k,h} + \frac{2}{3\lambda}\log\Big(\frac{1}{\delta}\Big) + \frac{8}{\lambda}\log\Big(\frac{1}{\delta}\Big) = 2\sum_{k=1}^{K} X_{k,h} + \frac{26}{3\lambda}\log\Big(\frac{1}{\delta}\Big).$$

Finally plug in (S.38) and simplify constants; tightening the numerical constant yields (S.37). ■

Applying Lemma S.6.3 with $\delta/(6H)$ and union bounding over $h \in [H]$, we conclude that with probability at least $1 - \delta/6$,

$$\sum_{h=1}^{H}\sum_{k=1}^{K}\mu_{k,h} \le 4Hd\log\Big(1 + \frac{K}{\lambda}\Big) + \frac{14H}{3\lambda}\log\Big(\frac{6H}{\delta}\Big). \tag{S.42}$$

S.6.4 Witness counting: from witnessed mistakes to total suboptimal episodes

The last ingredient converts the witness probability p_0 into a bound on the total number of suboptimal episodes.

Lemma S.6.4 (Witness counting for an adaptive sequence) *Let $(I_k, W_k)_{k=1}^{K}$ be as in (S.34). Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{k=1}^{K} I_k \;\le\; \frac{2}{p_0}\sum_{k=1}^{K} W_k \;+\; \frac{8}{p_0^2}\log\Big(\frac{1}{\delta}\Big). \tag{S.43}$$

Proof. Define the martingale difference

$$Z_k := W_k - \mathbb{E}[W_k \mid \mathcal{F}_{k-1}], \qquad M_n := \sum_{k=1}^{n} Z_k.$$

Then $\mathbb{E}[Z_k \mid \mathcal{F}_{k-1}] = 0$ and $|Z_k| \le 1$ a.s. By Azuma–Hoeffding, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$M_K \ge -\sqrt{2K \log\Big(\frac{1}{\delta}\Big)}. \tag{S.44}$$

Using (S.34) and the definition of M_K we have

$$\sum_{k=1}^{K} W_k = \sum_{k=1}^{K} \mathbb{E}[W_k \mid \mathcal{F}_{k-1}] + M_K \ge p_0 \sum_{k=1}^{K} I_k + M_K.$$

On the event (S.44), rearranging gives

$$\sum_{k=1}^{K} I_k \le \frac{1}{p_0} \sum_{k=1}^{K} W_k + \frac{1}{p_0} \sqrt{2K \log\Big(\frac{1}{\delta}\Big)}.$$

This bound is valid but not yet in the desired form. To remove the $\sqrt{K}$ term, we use that $W_k \le I_k$ and apply the inequality $\sqrt{K} \le \sqrt{\sum I_k}\sqrt{K} \le \frac{p_0}{4}\sum I_k + \frac{1}{p_0}K$ is too loose. Instead, apply Azuma only to the sub-sequence of indices with $I_k = 1$ via optional stopping (standard): let $T := \sum_{k=1}^{K} I_k$ and enumerate the random indices $\kappa_1 < \cdots < \kappa_T$ where $I_{\kappa_t} = 1$. Then $(W_{\kappa_t} - \mathbb{E}[W_{\kappa_t} \mid \mathcal{F}_{\kappa_t - 1}])_{t=1}^{T}$ is a martingale difference sequence with increments in $[-1,1]$ and conditional means at least p_0. Applying Azuma–Hoeffding to this length-T sequence yields that with probability at least $1-\delta$,

$$\sum_{k=1}^{K} W_k = \sum_{t=1}^{T} W_{\kappa_t} \ge \sum_{t=1}^{T} \mathbb{E}[W_{\kappa_t} \mid \mathcal{F}_{\kappa_t - 1}] - \sqrt{2T \log\Big(\frac{1}{\delta}\Big)} \ge p_0 T - \sqrt{2T \log\Big(\frac{1}{\delta}\Big)}.$$

If $T \le \frac{8}{p_0^2}\log(\frac{1}{\delta})$ then (S.43) holds trivially since $\sum W_k \ge 0$. Otherwise, $\sqrt{2T\log(1/\delta)} \le \frac{p_0}{2}T$, hence $\sum W_k \ge \frac{p_0}{2}T$, i.e., $T \le \frac{2}{p_0}\sum W_k$. Combining the two cases yields (S.43). ■

S.6.5 Proof of Theorem 6.9

Proof of Theorem 6.9. Work on the intersection of the three high-probability events: (i) $\mathcal{E}_{\text{conf}}$ from Lemma S.6.1 with level $\delta/6$, (ii) the stage-wise information budgets (S.42) from Lemma S.6.3 union-bounded over h with total failure probability $\delta/6$, and (iii) the witness-counting event (S.43) from Lemma S.6.4 with failure probability $\delta/6$. By a union bound, this intersection holds with probability at least $1-\delta/2$; we may absorb constant factors in δ.

On $\mathcal{E}_{\text{conf}}$, by (S.35)–(S.36) and (S.42),

$$\sum_{k=1}^{K} W_k \le \frac{1}{\eta^2}\left(4Hd\log\Big(1+\frac{K}{\lambda}\Big) + \frac{14H}{3\lambda}\log\Big(\frac{6H}{\delta}\Big)\right), \qquad \eta = \frac{\gamma_0}{\beta(\delta/6)\,L}.$$

Applying Lemma S.6.4 gives

$$\sum_{k=1}^{K} I_k \le \frac{2}{p_0}\sum_{k=1}^{K} W_k + \frac{8}{p_0^2}\log\Big(\frac{6}{\delta}\Big) \le \frac{2}{p_0\eta^2}\left(4Hd\log\Big(1+\frac{K}{\lambda}\Big) + \frac{14H}{3\lambda}\log\Big(\frac{6H}{\delta}\Big)\right) + \frac{8}{p_0^2}\log\Big(\frac{6}{\delta}\Big).$$

Finally, since rewards lie in $[0,1]$, per-episode regret is at most H, hence

$$\mathrm{Reg}(K) = \sum_{k=1}^{K}\left(V_1^\star - V_1^{\pi^k}\right) \le H\sum_{k=1}^{K} I_k.$$

Plugging the bound on $\sum I_k$ and $\eta^{-2} = \big(\beta(\delta/6)L/\gamma_0\big)^2$ yields

$$\mathrm{Reg}(K) \;\le\; C_1\,\frac{H^2}{p_0}\,\frac{\beta(\delta/6)^2\,L^2}{\gamma_0^2}\left(d\log\Big(1+\frac{K}{\lambda}\Big) + \frac{1}{\lambda}\log\Big(\frac{H}{\delta}\Big)\right) \;+\; C_2\,\frac{H}{p_0^2}\log\Big(\frac{1}{\delta}\Big),$$

for universal constants $C_1, C_2 > 0$.

Substituting the explicit form (S.27) and using the crude bound $\log(1 + \frac{K}{\lambda}) \le \log(\frac{KH}{\delta})$ for $K \ge 2$ and $\delta \le 1/2$, we obtain the stated $\widetilde{O}(\log K)$ dependence and recover the form of Theorem 6.9 up to universal constants and harmless logarithmic factors. (When λ is a fixed constant, the additional $(1/\lambda)\log(H/\delta)$ term is absorbed by the same polylogarithmic envelope.) ■

Remark S.6.5 (Where the margin is used) *The only use of Assumption 6.8 is in* (S.33)*: it converts "π^k is suboptimal" into a* constant-probability *event that forces a witnessed advantage of size at least γ_0. This, combined with Lemma S.6.2, yields a* deterministic lower bound *on μ_{k,h_k} whenever $W_k = 1$. The remainder of the proof is a counting argument showing that the total $\sum_{k,h} \mu_{k,h}$ is bounded by a polylogarithmic information budget.*

S.7 Relationship to diversity assumptions and additional structural corollaries

This section records (i) simple *sufficient* conditions—of the kind commonly called *feature diversity* / *covariate diversity* / *persistent excitation*—under which decoys are ruled out, and (ii) a few structural corollaries that re-express our geometric obstruction in equivalent forms. These results clarify precisely what is *gained* (and what is *not needed*) by classical full-rank conditions used in exploration-free analyses of greedy methods in contextual bandits and linear RL (Bastani et al., 2021; Civitavecchia & Papini, 2025).

S.7.1 Diversity conditions as sufficient conditions for self-identifiability

The most direct way to preclude decoys is to ensure that on-policy covariances never have nontrivial nullspaces. This is stronger than necessary (cf. Proposition 6.10), but it yields immediate and interpretable corollaries.

Assumption S.7.1 (Uniform on-policy diversity / persistent excitation) *There exists $\lambda_0 > 0$ such that for every deterministic $\pi \in \Pi_{\text{lin-greedy}}$ and every stage $h \in [H]$,*

$$\Sigma_h(\pi) \succeq \lambda_0 I_d, \tag{S.45}$$

where $\Sigma_h(\pi) = \mathbb{E}_{S\sim\rho_h^\pi}[\phi_h(S, \pi_h(S))\phi_h(S, \pi_h(S))^\top]$ is as in (12).

Proposition S.7.2 (Uniform diversity implies no decoys (hence self-identifiability)) *Fix $M^\star \in \mathcal{M}$. If Assumption S.7.1 holds, then no policy $\pi \in \Pi_{\text{lin-greedy}}$ admits a decoy relative to $M^\star$. In particular, $M^\star$ is self-identifiable (Definition 4.4).*

Proof. Fix any deterministic $\pi \in \Pi_{\text{lin-greedy}}$ and any $h \in [H]$. Assumption S.7.1 implies $\Sigma_h(\pi) \succ 0$, hence $\text{Null}(\Sigma_h(\pi)) = \{0\}$. Therefore, there does not exist a nonzero u satisfying (14) at any stage. By Theorem 5.4(1) (necessity: *decoy* $\Rightarrow$ *witness* (14)), π cannot admit a decoy. Since π was arbitrary, no policy in $\Pi_{\text{lin-greedy}}$ admits a decoy, and therefore $M^\star$ is self-identifiable. ■

Proposition S.7.3 (Uniform diversity bounds the leverage constant) *Under Assumption S.7.1 and the bounded-feature condition $\|\phi_h(s,a)\|_2 \le 1$, the leverage constant L from Definition 6.4 satisfies*

$$L \le \frac{2}{\sqrt{\lambda_0}}.$$

Proof. Fix π, h. Under (S.45), $\Sigma_h(\pi)^\dagger = \Sigma_h(\pi)^{-1} \preceq \lambda_0^{-1} I$. Hence for any (s,a),

$$\|\Delta_h^\pi(s,a)\|_{\Sigma_h(\pi)^\dagger} \le \frac{\|\Delta_h^\pi(s,a)\|_2}{\sqrt{\lambda_0}} \le \frac{\|\phi_h(s,a)\|_2 + \|\phi_h(s,\pi_h(s))\|_2}{\sqrt{\lambda_0}} \le \frac{2}{\sqrt{\lambda_0}}.$$

Taking $\max_a$, then $\operatorname{ess\,sup}_{S\sim\rho_h^\pi}$, then $\sup_{\pi,h}$ yields $L \le 2/\sqrt{\lambda_0}$. ■

Connection to prior "diversity" assumptions. Assumption S.7.1 is a strong (uniform) sufficient condition. It is representative of the kind of *feature-diversity / persistent-excitation* hypotheses under which purely greedy methods are known to be no-regret in contextual bandits and linear RL (e.g., covariate diversity in contextual bandits (Bastani et al., 2021) and feature-diversity-style conditions in exploration-free greedy LSVI (Civitavecchia & Papini, 2025)). Our main theorems sharpen this picture: full-rank covariances are not necessary to rule out decoys; what is necessary/sufficient (under Assumption 5.3) is the absence of *action-sensitive* nullspace directions.

S.7.2 Equivalent geometric forms and additional corollaries

This subsection records equivalent ways to state the "no action-sensitive nullspace" condition. These equivalences are purely linear-algebraic and do not require any RL-specific structure.

Lemma S.7.4 (Equivalent forms of "no action-sensitive null directions") *Fix a deterministic Markov policy π and a stage $h \in [H]$. The following statements are equivalent:*

1. $\mathrm{Null}(\Sigma_h(\pi)) \subseteq \mathrm{Null}(\Gamma_h(\pi))$.
2. $\mathrm{Range}(\Gamma_h(\pi)) \subseteq \mathrm{Range}(\Sigma_h(\pi))$.
3. *For every $a \in \mathcal{A}$, we have $\Delta_h^\pi(S,a) \in \mathrm{Range}(\Sigma_h(\pi))$ a.s. under $S \sim \rho_h^\pi$.*

Proof. Because $\Sigma_h(\pi) \succeq 0$ and $\Gamma_h(\pi) \succeq 0$, we have $\mathrm{Range}(M) = \mathrm{Null}(M)^\perp$ for $M \in \{\Sigma_h(\pi), \Gamma_h(\pi)\}$. Thus (1) is equivalent to $\mathrm{Null}(\Gamma_h(\pi))^\perp \subseteq \mathrm{Null}(\Sigma_h(\pi))^\perp$, i.e. (2).

Next, (3) holds iff for every $a \in \mathcal{A}$ and every $u \in \mathrm{Null}(\Sigma_h(\pi))$,

$$u^\top \Delta_h^\pi(S,a) = 0 \quad \text{a.s. under } S \sim \rho_h^\pi,$$

since $\mathrm{Range}(\Sigma_h(\pi)) = \mathrm{Null}(\Sigma_h(\pi))^\perp$. By Lemma 5.2, the latter statement is equivalent to $u \in \mathrm{Null}(\Gamma_h(\pi))$ for every $u \in \mathrm{Null}(\Sigma_h(\pi))$, i.e. (1). ■

Lemma S.7.4 makes explicit what full-rank "diversity" is buying. Full-rank $\Sigma_h(\pi)$ forces $\mathrm{Null}(\Sigma_h(\pi)) = \{0\}$, so (1) holds trivially. Our characterization shows that one only needs (1) (equivalently (3)), which allows rank-deficient $\Sigma_h(\pi)$ provided its nullspace is action-insensitive on visited states.

Corollary S.7.5 (Self-identifiability via action-insensitive nullspaces) *Assume Assumption 5.3. Fix $M^\star \in \mathcal{M}$. For a suboptimal $\pi \in \Pi_{\text{lin-greedy}}$, the following are equivalent:*

1. *π does not admit a decoy relative to $M^\star$.*
2. *For every stage $h \in [H]$, $\mathrm{Null}(\Sigma_h(\pi)) \subseteq \mathrm{Null}(\Gamma_h(\pi))$.*

Consequently, $M^\star$ is self-identifiable iff every suboptimal $\pi \in \Pi_{\text{lin-greedy}}$ satisfies $\mathrm{Null}(\Sigma_h(\pi)) \subseteq \mathrm{Null}(\Gamma_h(\pi))$ for all h.

Proof. Fix suboptimal π. By Theorem 5.4, under Assumption 5.3 we have: π admits a decoy iff there exist h and $u \neq 0$ with $u \in \mathrm{Null}(\Sigma_h(\pi)) \setminus \mathrm{Null}(\Gamma_h(\pi))$, which is equivalent to the negation of $\mathrm{Null}(\Sigma_h(\pi)) \subseteq \mathrm{Null}(\Gamma_h(\pi))$ at some stage. ■

Contextual bandits ($H = 1$). When $H = 1$, our trajectory-law notion of observational equivalence reduces to equality of the single-step observation law, and Corollary S.7.5 provides an RL analogue of "decoy-freedom" / identifiability conditions studied for greedy methods in bandits (Bastani et al., 2021; Slivkins et al., 2025). In particular, any bandit-side covariate-diversity hypothesis that enforces a uniform lower bound on the relevant on-policy covariances implies self-identifiability via Proposition S.7.2.

S.8 Additional examples, robustness, and optional experiments

This section collects (i) additional explicit instances that sharpen the qualitative/quantitative distinction between self-identifiability and learning rates, (ii) robustness observations about the role of deterministic tie-breaking and reward noise, and (iii) an optional set of lightweight simulation protocols that empirically illustrate the theory. None of the statements below are used in the main proofs; they serve as sanity checks and to clarify scope.

S.8.1 Self-identifiable instances with arbitrarily large leverage constant L

Theorem 6.7 is qualitative in self-identifiability but quantitative in the leverage constant L (Definition 6.4). The next construction shows that L can be made arbitrarily large even when the instance is self-identifiable (indeed, even when *all* relevant on-policy covariances are full rank). This clarifies that self-identifiability alone is not a conditioning assumption; it is only a *decoy-elimination* property.

Proposition S.8.1 (Self-identifiable yet ill-conditioned: L can be arbitrarily large) *For any $p \in (0,1)$ and any $\varepsilon \in (0, \frac{1}{2}]$, there exists a horizon-$H = 1$ instance $M^\star \in \mathcal{M}$ (hence a contextual bandit) that is self-identifiable relative to $\Pi_{\text{lin-greedy}}$, yet satisfies $L \geq 1/\sqrt{p}$. In particular, L can be made arbitrarily large (by taking $p \downarrow 0$) while maintaining self-identifiability.*

Proof. We construct an $H = 1$ instance with $\mathcal{S} = \{s_0, s_1\}$, $\mathcal{A} = \{0, 1\}$, $d = 2$, and initial distribution $\mu(s_1) = p$, $\mu(s_0) = 1 - p$. Let e_1, e_2 be the standard basis of $\mathbb{R}^2$ and define features

$$\phi_1(s_0, 0) = \phi_1(s_0, 1) = e_1, \qquad \phi_1(s_1, 0) = \varepsilon\, e_2, \qquad \phi_1(s_1, 1) = 2\varepsilon\, e_2.$$

These satisfy $\|\phi_1(s, a)\|_2 \leq 1$ because $\varepsilon \leq \frac{1}{2}$. Define deterministic rewards with mean

$$r_1(s, a) = \langle \phi_1(s, a), w^\star \rangle, \qquad w^\star := e_2,$$

so that $r_1(s_0, a) = 0$, $r_1(s_1, 0) = \varepsilon$, and $r_1(s_1, 1) = 2\varepsilon \in [0, 1]$. Since $H = 1$, linear Bellman completeness holds trivially (the Bellman operator returns r_1), and $Q_1^\star = r_1 \in \mathcal{Q}_1$ with bounded parameter norm $\|w^\star\|_2 = 1$.

Self-identifiability. The only suboptimal greedy-representable policies are those that choose action 0 at s_1 (choices at s_0 are immaterial since both actions have identical features and rewards there). Fix any such suboptimal deterministic π. Under π, the on-policy feature is e_1 at s_0 and εe_2 at s_1, hence

$$\Sigma_1(\pi) = (1 - p)\, e_1 e_1^\top + p\, \varepsilon^2\, e_2 e_2^\top = \begin{pmatrix} 1 - p & 0 \\ 0 & p\varepsilon^2 \end{pmatrix},$$

which is full rank since $p, \varepsilon > 0$. Therefore $\mathrm{Null}(\Sigma_1(\pi)) = \{0\}$, so the geometric decoy witness condition (14) cannot hold for π at stage $h = 1$. By the necessity direction of Theorem 5.4(1), π does not admit a decoy. Since *every* suboptimal $\pi \in \Pi_{\text{lin-greedy}}$ has this same property, $M^\star$ is self-identifiable (Definition 4.4).

Large leverage. For the same suboptimal π, at s_1 we have

$$\Delta_1^\pi(s_1, 1) = \phi_1(s_1, 1) - \phi_1(s_1, 0) = \varepsilon e_2.$$

Because $\Sigma_1(\pi)$ is diagonal, $\Sigma_1(\pi)^\dagger = \mathrm{diag}((1 - p)^{-1}, (p\varepsilon^2)^{-1})$. Thus

$$\|\Delta_1^\pi(s_1, 1)\|_{\Sigma_1(\pi)^\dagger} = \sqrt{(\varepsilon e_2)^\top \Sigma_1(\pi)^\dagger (\varepsilon e_2)} = \sqrt{\varepsilon^2 \cdot \frac{1}{p\varepsilon^2}} = \frac{1}{\sqrt{p}}.$$

Hence $L_1(\pi) \geq 1/\sqrt{p}$ and therefore $L \geq 1/\sqrt{p}$ by Definition 6.4. ∎

Remark S.8.2 (Interpretation and connection to design-dependent hardness) *Proposition S.8.1 is a one-step illustration of a general phenomenon familiar from linear bandits: even when the model is identifiable, rare contexts (small p) induce ill-conditioned design matrices and large prediction uncertainty in directions that matter for action selection, which is precisely what L quantifies in our analysis; see, e.g., Abbasi-Yadkori et al. (2011).*

S.8.2 Decoys allow failure but do not force it: instance-level vs. class-level dichotomy

Theorem 6.1 is existential: decoys *allow* (not force) linear regret. The next simple observation makes this scoping explicit on the canonical lower-bound instance.

Proposition S.8.3 (Same instance, different tie-breaking: decoy exists but $\mathrm{Reg}(K) = 0$) *Consider the $H = 2$ instance $M^\star$ from Theorem 6.1 (Appendix C). There exists a deterministic tie-breaking rule such that running* GREEDY-LSVI *on $M^\star$ yields $\mathrm{Reg}(K) = 0$ for all K, even though the suboptimal policy π that always selects action 0 at stage 1 admits a decoy relative to $M^\star$.*

Proof. Use the same $M^\star$ as in Appendix C, but fix tie-breaking to favor action 1 whenever $\widehat{Q}_{k,1}(s,0) = \widehat{Q}_{k,1}(s,1)$. In episode $k = 1$, $\widehat{Q}_{1,1} \equiv 0$, hence the greedy choice is $A_{1,1} = 1$. The algorithm visits x_1 at stage 2 and observes reward 1. It follows that the stage-2 ridge estimate has a positive e_2 component, so $\widehat{V}_{2,2}(x_1) > 0$, and the stage-1 target for action 1 becomes strictly larger than that for action 0. Inductively, the greedy action at stage 1 remains 1 forever, and the agent achieves the optimal value $V_1^\star = 1$ in every episode, hence $\mathrm{Reg}(K) = 0$.

Separately, the decoy for the trapped policy $\pi \equiv 0$ described in Appendix C (modifying the unvisited branch reward at x_1) exists regardless of the tie-breaking rule: it is a property of the instance and the policy π, not of the learning algorithm. Thus a decoy exists for a suboptimal greedy-representable policy, but GREEDY-LSVI need not follow that policy. ■

Remark S.8.4 *Proposition S.8.3 is the simplest illustration of why our "sharp dichotomy" is a* class-level *success/failure characterization (decoys allow adversarial failure), rather than an instance-level iff guarantee that* GREEDY-LSVI *must fail whenever any decoy exists.*

S.8.3 Robustness to stochastic tie-breaking at episode 1

A common question is whether random tie-breaking suffices to avoid the linear-regret trap. In the lower-bound construction, random tie-breaking only affects the first episode; afterwards, the algorithm becomes strictly greedy with no ties. Hence linear regret persists in expectation whenever the "bad" first action has nonzero probability.

Proposition S.8.5 (Random tie-breaking still yields linear expected regret) *Consider the instance $M^\star$ in Theorem 6.1. Suppose* GREEDY-LSVI *breaks ties at episode $k = 1$ by choosing action 0 with probability $q \in (0,1)$. Then for all K,*

$$\mathbb{E}[\mathrm{Reg}(K)] \;\geq\; \frac{q}{2}\,K.$$

Proof. On $M^\star$, the only tie occurs at episode $k = 1$ and stage 1 when $\widehat{Q}_{1,1} \equiv 0$. If $A_{1,1} = 1$, then the algorithm visits x_1 and the same inductive reasoning as in Proposition S.8.3 shows it remains optimal thereafter, yielding zero regret in all subsequent episodes.

If $A_{1,1} = 0$, then Appendix C shows that the algorithm remains trapped and chooses $A_{k,1} = 0$ for all k, incurring per-episode regret $1 - 1/2 = 1/2$. Therefore,

$$\mathbb{E}[\mathrm{Reg}(K)] = \mathbb{P}(A_{1,1} = 0) \cdot \frac{1}{2}K + \mathbb{P}(A_{1,1} = 1) \cdot 0 = \frac{q}{2}K,$$

which proves the claim. ■

Remark S.8.6 *Proposition S.8.5 shows that "random tie-breaking" is* not *an exploration strategy in the sense needed to break decoys: it only affects measure-zero events (ties), and on the lower-bound instance those events occur only at initialization.*

S.8.4 Robustness of the linear-regret trap to reward noise

Our lower bound uses deterministic rewards for transparency. The trap is not an artifact of determinism: it arises from the fact that the algorithm never observes the relevant feature coordinate. The following variant makes this explicit.

Proposition S.8.7 (The trap persists under stochastic rewards with the same means) *Consider the $H = 2$ construction in Appendix C, but replace the deterministic stage-2 reward at x_0 by an independent Bernoulli reward with mean $1/2$ (keeping all other rewards deterministic as in the construction). Then, with the same deterministic tie-breaking toward action 0 at episode 1,* GREEDY-LSVI *still satisfies* $\mathbb{E}[\mathrm{Reg}(K)] = \frac{1}{2}K$.

Proof. The key property used in Appendix C is that, as long as the algorithm never visits x_1, all stage-2 feature vectors are e_1, so the stage-2 ridge estimate always satisfies $\langle e_2, \widehat{w}_{k,2}\rangle = 0$ *exactly*. This remains true regardless of reward noise. Indeed, $b_{k,2}$ in (8) is always a multiple of e_1, so the second coordinate of $\widehat{w}_{k,2} = V_{k,2}^{-1} b_{k,2}$ is identically zero.

Consequently $\widehat{V}_{k,2}(x_1) = 0$ for all k, so the stage-1 regression target for action 1 remains 0 on all past data, and the e_2 coordinate of $\widehat{w}_{k,1}$ remains 0. Meanwhile the e_1 coordinate of $\widehat{w}_{k,1}$ is nonnegative (since the observed rewards and thus the targets are nonnegative), so the greedy choice at stage 1 remains action 0 for all episodes. Therefore the agent always receives expected return $1/2$ while the optimal return is 1, yielding expected per-episode regret $1/2$ and hence $\mathbb{E}[\mathrm{Reg}(K)] = \frac{1}{2}K$. ■

S.8.5 Optional experiments: simple simulations illustrating decoys and conditioning

This subsection is optional and purely illustrative (not used by any theorem). The goal is to provide a minimal simulation suite that (i) reproduces the trap in Theorem 6.1, (ii) visualizes the dependence of learning speed on the leverage/conditioning parameter in Proposition S.8.1, and (iii) shows the qualitative effect of adding explicit exploration (e.g., optimism) as in LSVI-UCB (Jin et al., 2020).

E.1 Instances. We recommend the following three families, all fully specified by the preceding sections.

1. **Decoy trap (linear regret).** Use the $H = 2$ instance from Appendix C (deterministic or Bernoulli reward variants).
2. **Self-identifiable but ill-conditioned.** Use the $H = 1$ instance from Proposition S.8.1. Vary $p \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ to sweep the leverage constant $L \asymp 1/\sqrt{p}$.
3. **Margin-controlled switching (qualitative).** Construct a simple $H = 1$ instance with a uniform gap γ_0 between the best and second-best action on a p_0 fraction of contexts, to qualitatively illustrate Assumption 6.8 and the log-regret phenomenon of Theorem 6.9 (compare also the bandit "margin" intuition in Bastani et al., 2021 and the decoy-based perspective in Slivkins et al., 2025).

E.2 Algorithms. Implement:

1. GREEDY-LSVI (Algorithm 1) with fixed deterministic tie-breaking and ridge parameter $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$.

2. Optionally, an optimistic baseline such as LSVI-UCB (Jin et al., 2020) (same regression backbone but with an exploration bonus), to emphasize that optimism breaks the decoy trap by forcing visits to the unobserved branch.

3. Optionally, a randomized-value-function baseline (Osband et al., 2019) (for tabular/linear cases this can be implemented by Gaussian perturbations of regression targets), to illustrate deep exploration mechanisms outside explicit UCB bonuses.

E.3 Metrics and plots. For each $K \in \{10^2, \dots, 10^5\}$, plot: (i) cumulative regret $\mathrm{Reg}(K)$ (or its Monte Carlo estimate), (ii) per-episode regret, and (iii) the number of times each action is selected at each stage. For the ill-conditioned $H = 1$ family (Proposition S.8.1), also report the empirical condition number of the design matrix $V_{K,1}$ and compare it to the theoretical dependence on p.

E.4 Expected qualitative outcomes.

- On the decoy trap instance, GREEDY-LSVI exhibits linear regret, while exploration-based methods (optimism or randomization) escape quickly.
- On the ill-conditioned self-identifiable family, GREEDY-LSVI eventually learns, but the sample size required to reliably distinguish actions at the rare context s_1 scales inversely with p, consistent with $L \asymp 1/\sqrt{p}$ and the design-dependent nature of ridge regression (Abbasi-Yadkori et al., 2011).
- On margin-controlled families, once confidence widths fall below γ_0, greedy policies switch rapidly, qualitatively matching the mechanism of Theorem 6.9.

Remark S.8.8 (Reproducibility) *To make the simulations reproducible, fix a pseudorandom seed, report the number of trials, and include the exact tie-breaking convention. In the decoy trap instance, the entire behavior of* GREEDY-LSVI *depends on the first-episode tie-break (Propositions S.8.3 and S.8.5).*