ALMOST EXOGENOUS: REGRET GUARANTEES FOR REINFORCEMENT LEARNING UNDER BOUNDED ACTION INFLUENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Exogenous Markov Decision Processes (Exo-MDPs) model settings where uncertainty is driven by external factors independent of the learner's actions. In many real systems, this separation is only approximate: actions weakly influence variables otherwise treated as exogenous. We study reinforcement learning under bounded violations of exogeneity, where each action perturbs the exogenous transition law. The perturbation is controlled by an impact parameter ε. We introduce two optimistic algorithms for this perturbation: a tabular method and Exogenous Value-Targeted Regression (Exo-VTR), which exploits linear-mixture structure in large exogenous spaces. We prove high-probability regret of the form $\tilde{O}(\sqrt{K}) + \tilde{O}(\varepsilon K)$, where K is the number of episodes, retaining the statistical benefit of exogenous structure while paying an explicit perturbation penalty. In the tabular setting, no-bonus predict-then-optimize (PTO) attains the same regret order. We also provide lower bounds on first-order loss for specified exogenous planner classes. A separate two-stage construction establishes first-order regret for Exo-UCBVI-ε, PTO, and Exo-VTR-ε with a one-dimensional reference representation. Experiments validate the execution-family separation and probe the roles of clipping and optimism.

1 INTRODUCTION

Many sequential decision problems contain an asymmetry: the learner controls part of the system, while much of the uncertainty originates outside its control. Inventory decisions respond to demand, resource-allocation policies respond to arrivals, and execution decisions respond to market conditions. Exogenous Markov decision processes (Exo-MDPs) formalize this structure by separating the state into an endogenous component that responds to the learner's decisions and an exogenous component whose evolution is independent of them (Dietterich et al., 2018; Efroni et al., 2022; Sinclair et al., 2023). This separation can substantially simplify learning, since observations of the exogenous process can be reused across policies rather than explored separately throughout the endogenous state-action space.

Recent theory has sharpened the statistical value of this structure. Wan et al. (2024) connect discrete Exo-MDPs to linear-mixture MDPs and obtain regret guarantees governed by the exogenous structure rather than by the potentially much larger endogenous state and action spaces. More recently, Liang et al. (2026) show that exact exogeneity can even remove the need for deliberate exploration: because the exogenous process is unaffected by the learner, observed trajectories provide policy-independent information that supports pure-exploitation learning in both tabular and linear-function-approximation settings. These results show that exogeneity is more than a modeling convenience; it fundamentally changes what must be learned.

Yet the same assumption is brittle. Exact Exo-MDPs require the learner to have no influence on the future exogenous process, while many real systems violate this idealization without becoming fully endogenous. Large orders can exert market impact, resource-allocation decisions can weakly alter future congestion, and interventions in physical or economic systems can feed back into variables otherwise treated as external. In such settings, discarding exogenous structure entirely can be unnecessarily pessimistic, whereas treating the system as exactly exogenous introduces a systematic

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

modeling error that need not disappear with more data. Existing Exo-MDP guarantees most directly comparable to ours therefore leave an important intermediate regime uncharacterized: the environment remains largely driven by external uncertainty, but the learner can weakly influence it. This leads to our central question:

How much action influence can an Exo-MDP tolerate before its statistical advantage is lost, and what is the price of treating an approximately exogenous process as exogenous?

We study this intermediate regime through *bounded violations of exogeneity*. Informally, the exogenous transition may depend weakly on the learner's state and action, with the strength of this departure controlled by an impact parameter ε and the magnitude of the action taken. Exact exogeneity is recovered when this influence vanishes, while positive ε allows limited action-induced feedback. Rather than learning an unrestricted action-dependent transition model, we ask how much of the statistical advantage of the exogenous representation can be retained under such controlled violations. We summarize our contributions as follows.

Bounded violations of exogeneity. We introduce a near-exogenous model that interpolates between exact Exo-MDPs and action-dependent environments. The exogenous process may respond weakly to the learner while remaining within a controlled ℓ_1 neighborhood of an action-independent reference process, with ε quantifying the strength of this violation. This formulation allows us to study how the statistical advantages of exogenous structure degrade as action influence increases.

Upper and lower regret guarantees. We develop Exo-UCBVI-ε for tabular settings and Exo-VTR-ε for linear-mixture exogenous kernels. Both retain the leading statistical rate of the exogenous model while adding impact-dependent terms that are linear in ε and the number of episodes, up to logarithmic factors. For Exo-VTR-ε, our bound makes the dependence on the action cap and trajectory budget explicit, building on the robust VTR misspecification framework of Li & Yang (2024). We further establish first-order lower bounds for consistent action-independent planners on a hard family, and separate algorithm-specific lower bounds for Exo-UCBVI-ε, its cap-free variant, no-bonus PTO, and Exo-VTR-ε.

Mechanism-level empirical validation. Controlled experiments reproduce the separation predicted by the theory and examine how bounded action influence affects optimistic planning. In the tabular case, impact widening changes the maximizing action only through truncation, while Exo-VTR-ε can produce action-dependent bonuses even without active truncation. Ablations and an impact-aware parametric reference further isolate these effects.

2 Related Work

Learning with exogenous structure. A growing line of work exploits exogenous information to reduce the statistical difficulty of reinforcement learning (RL) (Dietterich et al., 2018; Efroni et al., 2022; Sinclair et al., 2023). Wan et al. (2024) show that discrete Exo-MDPs induce a linear representation through the exogenous distribution, yielding regret whose leading dependence can avoid the endogenous state and action space sizes; Liang et al. (2026) show that exact action independence can make deliberate exploration unnecessary; and Maran et al. (2026) obtain matching upper and lower bounds for action-independent exogenous dynamics. We instead allow action independence to fail in a bounded, action-linked manner.

Decision-dependent dynamics, misspecification, and robustness. Performative RL allows the deployed policy to modify rewards and transitions (Mandal et al., 2023; Mandal & Radanovic, 2025), whereas we retain a reference exogenous process and bound the deviation induced by individual actions. Misspecification analyses degrade gracefully with the misspecification level (Vial et al., 2022; Zhang et al., 2024; Du et al., 2020; Lattimore et al., 2020), while corruption-robust and distributionally robust methods address adversarial or off-dynamics deviations (Lykouris et al., 2025; Wei et al., 2022; Ye et al., 2024; He et al., 2025). Our deviation model instead ties the departure from exogeneity to action magnitude and a pathwise budget. Theorem 4 isolates a first-order loss for the specified consistent exogenous planners on the hard family.

Linear mixture MDPs and value targeted regression. Our linear mixture analysis builds on value-targeted model estimation (Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021; Huang et al.,

2024), which Wan et al. (2024) connect to Exo-MDPs through the exogenous distribution. Li & Yang (2024) analyze Robust-UCRL-VTR under locally bounded model misspecification, including linearly parameterized transitions. Our Exo-VTR analysis specializes this setting to an exogenous reference kernel and gives explicit dependence on the action cap and trajectory budget.

A comprehensive comparison with each of these lines of work is given in Appendix B.

3 SETTING

We study episodic RL over K episodes of horizon H. A state decomposes as $s = (x, \xi)$, where $x \in \mathcal{X}$ is *endogenous* and $\xi \in \Xi$ is *exogenous*, with $|\Xi| = S_\bullet \geq 2$ in the tabular case. Actions are scalar, $a \in \mathcal{A} \subset \mathbb{R}$, and the reward $r_h(x, a, \xi) \in [0, 1]$ is deterministic and known. The initial state s_1 is fixed. The endogenous variable is determined by a known map of the current endogenous state, the action, and the next exogenous state,

$$x_{h+1} = f_h(x_h, a_h, \xi_{h+1}), \tag{1}$$

so all transition uncertainty is generated by the exogenous process, which may affect both successor components. Assuming known, deterministic f_h and r_h places us in the setting of Liang et al. (2026). We relax only the independence of the exogenous process from the learner's actions.

Reference and actual exogenous kernels. The exogeneity idealization asserts that the exogenous coordinate evolves without reference to the agent,

$$\xi_{h+1} \sim P_0(\cdot \mid \xi_h), \tag{2}$$

and it is this action-independent kernel that an exogenous-MDP method estimates and plans against. We take Eq. (2) as a *reference* model rather than as a description of the environment. The reference exogenous kernel P_0 is shared across all stages. This permits tabular transition counts to be pooled across stages for each exogenous state and, under Assumption A1L, VTR regression samples to be pooled across stages under one common transition parameter. The rewards r_h and endogenous map f_h may still depend on h. The learner instead interacts with a kernel that may respond to what it does,

$$\xi_{h+1} \sim P_\varepsilon(\cdot \mid \xi_h, x_h, a_h), \tag{3}$$

and the content of our setting is a bound on how far Eq. (3) can travel from Eq. (2).

Conditionally on the current state (x, ξ) and action a, the joint law of the successor (x', ξ') is $\delta_{f_h(x,a,\xi')}(dx')\, P_\varepsilon(d\xi' \mid \xi, x, a)$ in the actual model, and $\delta_{f_h(x,a,\xi')}(dx')\, P_0(d\xi' \mid \xi)$ in the reference model. The Dirac mass depends on the same successor ξ' that is integrated against the exogenous kernel; these laws need not factor into independent successor coordinates.

Assumption A2 (Bounded action influence)**.** *For every* (ξ, x, a)*,*

$$\left\| P_\varepsilon(\cdot \mid \xi, x, a) - P_0(\cdot \mid \xi) \right\|_1 \;\leq\; \varepsilon\, |a|. \tag{4}$$

Write M_0 and M_ε for the reference and actual models. They share the initial state, f_h and r_h, and differ only in the exogenous kernel. At $\varepsilon = 0$ they coincide; for $\varepsilon > 0$, Eq. (4) bounds the action-induced deviation from P_0. The model is illustrated in Appendix C.

Assumption A3 (Action budget)**.** *Pathwise,* $\sum_{h=1}^{H} |a_h| \leq \bar{X}$*, and per-step* $|a_h| \leq a_{\max}$*.*

The budget $\bar{X}$ is what makes the perturbation in Eq. (4) summable across a trajectory, and it is the quantity that appears in our rates. The per-step cap $a_{\max}$ yields the sharper capped guarantee of Theorem 1; Theorem 2 removes it and uses only the pathwise budget, at the cost of an additional logarithmic factor and a K-independent remainder.

Remark 1 (Optimal execution)**.** In optimal trade execution, x_h is the remaining inventory, ξ_h is the price or market state, a_h is the quantity traded at step h, and the endogenous map is $x_{h+1} = x_h - a_h$, so the budget $\bar{X}$ is the parent order. The exogeneity idealization is the standard modeling assumption that prices are unaffected by the trader; Eq. (4) replaces it with the assumption that a trade of size $|a_h|$ changes the price law by at most $\varepsilon|a_h|$, capturing bounded market impact. This execution update is the special case $f_h(x, a, \xi') = x - a$, which ignores the successor argument. Further details are in Appendix C.

Admissibility. The remaining budget is carried in the endogenous state. Let $\mathcal{A}_h(x, \xi)$ be the known set of actions respecting that budget and the instance constraints, admitting a feasible continuation through the horizon for every candidate exogenous successor. In capped statements it also imposes $|a| \le a_{\max}$; in cap-free statements learner and comparator both omit this separate cap. All policies and maximizations below use these feasible sets. Appendix C gives the formal state augmentation and construction.

Policies, values, and regret. A policy $\pi = (\pi_h)_{h=1}^{H}$ is a possibly randomized, non-stationary Markov policy, with π_h mapping the current state to a distribution supported on $\mathcal{A}_h(x, \xi)$. For a model $M \in \{M_0, M_\varepsilon\}$ we write $V^{\pi}_{M,h}(s)$ for the expected sum of rewards collected from step h onward under π in M, and $V^{\star}_{M,h}(s) = \sup_\pi V^{\pi}_{M,h}(s)$, where the supremum is over admissible Markov policies; since each of the $H-h+1$ rewards lies in $[0, 1]$, we have $0 \le V^{\pi}_{M,h}(s) \le H-h+1$ throughout. We abbreviate $V^{\pi}_{M_\varepsilon,h}$ and $V^{\pi}_{M_0,h}$ as $V^{\pi}_{\varepsilon,h}$ and $V^{\pi}_{0,h}$, respectively, and use the analogous notation for optimal values. The Bellman equations evaluate continuation at $(f_h(x, a, \xi'), \xi')$ for the same successor ξ' (see full forms in Appendix C). An algorithm commits to a policy π_k at the start of episode k, and interacts with M_ε, which leads to regret

$$\mathrm{Regret}_{M_\varepsilon}(K) \;=\; \sum_{k=1}^{K} \Big[V^{\star}_{\varepsilon,1}(s_1) - V^{\pi_k}_{\varepsilon,1}(s_1) \Big]. \tag{5}$$

Regret is measured against the optimal policy in the environment actually encountered by the learner. Because the data are generated by M_ε while the learner estimates the reference kernel P_0, action influence introduces a systematic bias in addition to the usual transition noise. Our analysis separates and controls these two effects.

Linear-mixture exogenous kernels. The tabular results of Section 4 require no additional structural assumption beyond Assumptions A2 and A3. For Section 6 we replace the tabular structure on the reference kernel, and only on the reference kernel, by a linear-mixture representation.

Assumption A1L (Linear-mixture reference kernel; used only in Section 6)**.** *There exist a known feature map $\phi : \Xi \times \Xi \to \mathbb{R}^d$, a dominating measure ν on Ξ, and an unknown parameter $\theta^\star \in \mathbb{R}^d$ with $\|\theta^\star\|_2 \le B_\theta$, such that*

$$P_0(d\xi' \mid \xi) = \big\langle \phi(\xi', \xi), \theta^\star \big\rangle \, \nu(d\xi'), \tag{6}$$

which reduces to $P_0(\xi' \mid \xi) = \langle \phi(\xi', \xi), \theta^\star \rangle$ for finite Ξ. For a bounded measurable $g : \Xi \to \mathbb{R}$ set

$$\psi_g(\xi) := \int_{\Xi} \phi(\xi', \xi)\, g(\xi')\, \nu(d\xi'), \tag{7}$$

and assume the normalization $\|\psi_g(\xi)\|_2 \le 1$ whenever g takes values in $[0, 1]$.

Combining Eqs. (6) and (7) gives the identity the linear algorithm is built on: for every bounded g and every ξ,

$$\big\langle P_0(\cdot \mid \xi), g \big\rangle = \psi_g(\xi)^\top \theta^\star. \tag{8}$$

A single parameter serves every test function, so each transition can be paired with the value function current when it was collected; since every value function below lies in $[0, H]$, linearity gives $\|\psi\|_2 \le H$ for all features used. Assumption A1L constrains P_0 alone: P_ε need not admit such a representation and is controlled solely through Eq. (4), which lets the same ε-analysis serve both geometries. For general spaces, we assume the measurability conditions and exact integration and planning oracles specified in Appendix F.

4 Tabular near-exogenous learning

Throughout this section we assume Ξ is finite and no structure beyond Assumptions A2 and A3 is imposed. Since f_h is known, the algorithm pools exogenous transitions across actions to estimate the shared reference kernel P_0. By Eq. (4), the mixture of laws producing a state's data differs from P_0 by at most ε times its average action magnitude, hence by at most $\varepsilon a_{\max}$ (Lemma 8, Appendix D).

Estimator and confidence radius. Let $n_k(\xi)$ count the transition-generating visits to ξ in episodes $1, \dots, k-1$, and let $\widehat{P}_k(\cdot \mid \xi)$ be the empirical successor distribution formed from exactly those transitions, taken uniform when $n_k(\xi) = 0$. Two errors separate $\widehat{P}_k$ from P_0. The sampling error is controlled, simultaneously over all states and all counts, by

$$c_\beta := \sqrt{2\big(S_\bullet \ln 2 + \ln(S_\bullet KH/\delta)\big)}, \qquad \beta_0 := 2, \qquad \beta_n := \frac{c_\beta}{\sqrt{n}} \ (n \geq 1), \tag{9}$$

where $\delta \in (0,1)$ is the failure probability. The second error is systematic and is exactly the displacement just described. The bonus admits both,

$$b_{k,h}(\xi) := \tfrac{1}{2}(H-h)\big(\beta_{n_k(\xi)} + \varepsilon a_{\max}\big), \tag{10}$$

the factor $\frac{1}{2}(H-h)$ arising from the change of measure bound applied to a value function of span at most $H-h$. The first term of Eq. (10) shrinks as data accumulate and the second does not. That asymmetry produces the two regret terms below, and it is the reason the worst-case impact component does not vanish solely as more data are collected.

Algorithm. Exo-UCBVI-ε performs optimistic value iteration against $\widehat{P}_k$ with the bonus Eq. (10), truncating each backup from above at $H-h+1$ so that the span bound the bonus relies on is maintained. Counts and estimates are frozen at episode boundaries, which makes the value function used at step h measurable at the start of the episode, as the regret recursion requires. Algorithm 1 in Appendix D states the procedure in full. Lemma 12 shows that freezing adds only $S_\bullet H$ to the count bound and preserves the leading $\sqrt{K}$ rate.

Theorem 1 (Regret of Exo-UCBVI-ε)**.** *Under Assumptions A2 and A3, with probability at least $1-2\delta$, we have*

$$\mathrm{Regret}_{M_\varepsilon}(K) \ \leq \ c_\beta H\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) \ + \ \frac{\varepsilon a_{\max} H^2 K}{2} \ + \ \varepsilon \bar{X} HK \ + \ 2H\sqrt{2KH\ln(1/\delta)}.$$

Writing the first term as a rate, $\widetilde{O}\big(S_\bullet\sqrt{H^3K}\big) + \widetilde{O}\big(S_\bullet^{3/2}H^2\big)$, the second summand being the frozen count burn-in carried by $S_\bullet H$ in Lemma 12; the impact contribution is $O\big(\varepsilon a_{\max}H^2K + \varepsilon\bar{X}HK\big)$. The statistical term carries the $\sqrt{K}$ dependence of the exogenous model and involves neither $|\mathcal{X}|$ nor $|\mathcal{A}|$, so the benefit identified by Wan et al. (2024) survives the relaxation. The impact terms are linear in K and vanish at $\varepsilon = 0$, recovering an exact exogeneity guarantee. When the per-step cap satisfies $a_{\max} \leq c\bar{X}/H$, the two impact terms are bounded by $(1+\frac{c}{2})\varepsilon\bar{X}HK$.

Pure exploitation. No-bonus predict-then-optimize (PTO), which plans greedily in the same frozen pooled empirical model, satisfies the bound of Theorem 1 with $2\varepsilon\bar{X}HK$ in place of $\varepsilon\bar{X}HK$, with probability at least $1-2\delta$ (Theorem PTO, Appendix D.5). Thus optimism is unnecessary for this tabular regret order under bounded action influence.

Removing the per-step cap. The cap-free variant replaces the uniform $\varepsilon a_{\max}$ widening in Eq. (10) by the observed average action magnitude at each exogenous state. Appendix D.6 defines its computable bonus; learner and comparator both use the cap-free feasible sets. The resulting guarantee is as follows.

Theorem 2 (Cap-free regret)**.** *Under Assumption A2 and the pathwise budget of Assumption A3 alone, Exo-UCBVI-ε run with the bonus $b'_{k,h}$ of Appendix D.6 satisfies, with probability at least $1-2\delta$,*

$$\begin{aligned}\mathrm{Regret}_{M_\varepsilon}(K) \leq c_\beta H\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) + \varepsilon H\big(1+\ln(KH)\big)(K + S_\bullet H)\,\bar{X} \\ + \varepsilon\bar{X}HK + 2H\sqrt{2KH\ln(1/\delta)}.\end{aligned}$$

Proof sketch. Appendix D gives the full argument. Contaminated concentration and optimism lead to a regret recursion under the actual transition law; the frozen-count bound of Lemma 12 controls its statistical term. The volume weighted simulation gap of Lemma 7 is paid once per episode, and a martingale bound handles the remaining noise.

Remark 2 (How the widening reaches the policy)**.** The bonus Eq. (10) depends on ξ and h but not on the action. Before upper truncation it shifts all action scores equally, so its widening can change the maximizer only through truncation. In Section 6, the linear-mixture feature and elliptical bonus can instead vary across actions.

5 LOWER BOUNDS

Under the hypotheses of Proposition 16, the unique reference-optimal policy incurs $o(\varepsilon)$ loss per episode. The execution bounds below instead use a tied reference problem, in which every full selling schedule has the same reference value. Theorem 4 is the principal result for the specified reward-aware exogenous planners; Theorem 3 covers a more restricted class. These are bounds for those planner classes, not a minimax bound over all learners.

A tied instance family. This family uses the successor-independent execution map of Remark 1, and hence is a special case of Eq. (1). Fix $g \in (0,1)$ and take $H = 3$, $\bar{X} = 1$, and $a_{\max} = \frac{1}{2}$, so that the feasible schedules form the simplex $\mathcal{S} = \{a \in [0, \frac{1}{2}]^3 : \sum_h a_h = 1\}$. Let $\Xi = \{G, B\}$ with $\xi_1 = G$ and B absorbing, and let the reward be $r_h = a_h\, p(\xi_h)$ for a price p with $p(G) = 1$. The family contains two variants, indexed by $v \in \{\mathrm{UP}, \mathrm{DOWN}\}$, which differ only in the price attached to B: $p(B) = 1 + g$ under UP and $p(B) = 1 - g$ under DOWN. Since $a_h \le \frac{1}{2}$ and $p \le 1 + g < 2$, the rewards lie in $[0,1]$ as Section 3 requires. The kernels are common to both variants,

$$P_0(B \mid G) = 0, \qquad P_\varepsilon(B \mid G, x, a) = \tfrac{1}{2}\varepsilon a, \tag{11}$$

so that the bound in Eq. (4) is attained on the G row, the two kernels coinciding at B. Under the reference kernel, the exogenous state never leaves G, every schedule in $\mathcal{S}$ collects the same total, and the idealized problem is completely tied. Under P_ε the tie is broken at first-order by

$$V_\varepsilon^{\mathrm{UP}}(a) = 1 + \tfrac{1}{2} g\varepsilon\, e_2(a) + O(\varepsilon^2), \qquad V_\varepsilon^{\mathrm{DOWN}}(a) = 1 - \tfrac{1}{2} g\varepsilon\, e_2(a) + O(\varepsilon^2), \tag{12}$$

where $e_2(a) = \sum_{i<j} a_i a_j$ and the remainders are uniform on $\mathcal{S}$. Appendix E shows that spreading maximizes e_2 and is favored under UP, while concentration minimizes it and is favored under DOWN. The two first-order suboptimality gaps sum to $\frac{1}{24} g\varepsilon$; hence, for sufficiently small ε, every fixed schedule incurs at least $\frac{1}{48} g\varepsilon(1 - C_g\varepsilon)$ in one variant, for some finite C_g depending only on g.

Planners blind to the variant. Because the two variants share the initial state, endogenous map and kernels and differ only in rewards, any algorithm whose action selection ignores the reward function induces the same interaction law in both. Call an algorithm *instance oblivious and statistic blind* if its randomized policy sequence is the same function of the pooled transition counts in each variant; Remark 7 explains why this class must exclude use of the reward function.

Theorem 3. *Every instance oblivious and statistic blind algorithm satisfies*

$$\max_{v \in \{\mathrm{UP}, \mathrm{DOWN}\}} \mathbb{E}\big[\mathrm{Regret}_v(K)\big] \;\ge\; \frac{g\varepsilon}{48}\, K\,(1 - C_g\varepsilon).$$

Appendix E proves the bound using the common interaction law of the two variants and their opposing optimal schedules.

Planners that estimate an action-independent kernel. Theorem 3 constrains algorithms that ignore rewards, which Exo-UCBVI-ε does not. We next consider *consistent exogenous planners*. An algorithm belongs to this class if at each episode it plays a policy optimal in the exogenous model determined by some action-independent kernel $\widehat{Q}_k(\cdot \mid \xi)$, where $\widehat{Q}_k$ is either the pooled maximum likelihood estimate or any kernel within the ℓ_1 radius $\beta_{n_k(\xi)}$ of it. Rewards may be used freely in planning, and ties among models are broken adversarially. This class contains the pooled point estimate and confidence set exogenous planners analyzed below.

Under UP, the pooled point-estimate planner eventually treats an action-induced $G \to B$ transition as part of an action-independent kernel and selects the concentrated schedule. A confidence-set planner can retain a temporary partial hedge. Appendix E gives the two pinning arguments and the hedge geometry.

Theorem 4. *Consider variant UP. (I) For the point estimate planner, with T its random pinning time,*

$$\mathbb{E}\big[\mathrm{Regret}_{\mathrm{UP}}(K)\big] \;\ge\; \frac{g\varepsilon}{24}\,(1 - C_g\varepsilon)\,\mathbb{E}\big[(K - T)^+\big].$$

(II) For the confidence set planner, with T_{cs} its deterministic burn-in threshold, there is an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \ge 1 - 2\delta$, on which every episode $k \ge T_{\mathrm{cs}}$ is pinned, such that

$$\mathrm{Regret}_{\mathrm{UP}}(K) \;\ge\; \frac{g\varepsilon}{96}\,(K - T_{\mathrm{cs}})^+\,(1 - C_g\varepsilon) \qquad \text{on } \mathcal{E}.$$

The weaker constant in Theorem 4 *(II)* reflects a partial hedge available to the confidence-set planner. Using that hedge also generates observations that eventually exclude it. Proposition 21 bounds its aggregate contribution by $O(g \log(S_\bullet KH/\delta))$, so both planner types have the aggregate post-burn-in first-order coefficient $\frac{1}{24} g\varepsilon$, subject to the theorem's confidence qualification.

What the separation isolates. A variant-aware comparator, permitted to read the variant from the known reward function and play the corresponding first-order optimal schedule, incurs $O(g\varepsilon^2 K)$ in both variants. Thus, for the blind and consistent exogenous planner classes covered by these lower bounds, a first-order εK dependence is unavoidable on this family, and it matches in shape the impact terms of Theorems 1 and 2. In UP, the loss arises because the action-independent model attributes action-induced transitions to the exogenous kernel and, after pinning, selects a concentrated schedule although the true first-order optimum is the spread schedule.

Algorithm-specific lower bounds. A separate construction in Appendix E.8 applies directly to Algorithm 1. For every $0 < \varepsilon \leq 1$, it has $H = 2$, $\bar{X} = a_{\max} = 1$, and two endogenous states, exogenous states and actions. For every $K \geq 1$ and $\delta \in (0, 1)$, regardless of optimistic tie-breaking, Algorithm 1 satisfies almost surely

$$\mathrm{Regret}_{M_\varepsilon}(K) \geq \frac{\varepsilon}{16} \left(K - \left\lceil \frac{c_\beta^2}{4} \right\rceil \right)_+, \qquad c_\beta^2 = 2 \ln \frac{16K}{\delta}.$$

The cap-free variant obeys the same bound; PTO has regret exactly $\varepsilon K/16$. With the one-dimensional representation $\phi(\xi', \xi) = P_0(\xi' \mid \xi)$, Algorithm 2 and its cap-free variant also have regret exactly $\varepsilon K/16$ from the first episode, as proved in Appendix E.8. This representation determines the reference kernel. The known reward is chosen separately for each ε, so these results do not contradict the fixed-reference conclusion of Proposition 16. They are algorithm-specific results, not minimax lower bounds.

6 Linear mixture near-exogenous learning

We now replace the tabular structure on the reference kernel by Assumption A1L, so that Ξ may be large or infinite and the statistical dependence is governed by the mixture dimension d in place of $S_\bullet$. The proof keeps the tabular structure (confidence control, optimism and a regret recursion under the actual law), the volume weighted simulation bound, and the separation of contamination bias from sampling noise, with least squares replacing tabular estimation. Theorem PTO removes optimism only in the tabular setting; no no-bonus guarantee is established here. A tabular indicator embedding gives a weaker dimension-dependent guarantee than Theorem 1 (Appendix F).

Frozen value targeted regression. Let D_{k-1} be the set of transition-generating steps of episodes $1, \ldots, k-1$, indexed by a global time t. For a step t occurring at stage h_t of episode j, with current state and action (ξ_t, x_t, a_t), define

$$V_t(\xi') := \overline{V}_{j,h_t+1}\big(f_{h_t}(x_t, a_t, \xi'), \xi'\big).$$

This uses the value function in force when the step was taken. The function V_t is determined before the successor draw. After observing ξ_{t+1}, define the regression pair

$$\psi_t := \psi_{V_t}(\xi_t), \qquad y_t := V_t(\xi_{t+1}). \tag{13}$$

With ridge parameter $\lambda := H^2$, set

$$\Sigma_k := \lambda I + \sum_{t \in D_{k-1}} \psi_t \psi_t^\top, \qquad \widehat{\theta}_k := \Sigma_k^{-1} \sum_{t \in D_{k-1}} \psi_t y_t, \tag{14}$$

where Σ_k is the regularized design matrix and $\widehat{\theta}_k$ the ridge estimate of $\theta^\star$. Each regression target separates into a reference-kernel prediction, action-induced bias, and conditionally mean-zero noise (Appendix F). The resulting confidence radius is

$$\beta_k := HB_\theta + \frac{H}{2} \sqrt{\ln \frac{\det \Sigma_k}{\lambda^d} + 2 \ln \frac{1}{\delta}} + \frac{\varepsilon H}{2} \sqrt{a_{\max} \bar{X}\,(k-1)}, \tag{15}$$

whose three terms pay for regularization, for martingale noise, and for contamination. The last is the analog of the tabular impact widening. The contamination contribution is proportional to $\sqrt{k-1}$; Assumption A3 controls its growth through the squared action volumes, whose sum over the first $k-1$ episodes is at most $a_{\max}\bar{X}(k-1)$ (Lemma 24).

Optimistic backup. With $\psi_{k,h}(x,\xi,a) := \psi_{\overline{V}_{k,h+1}(f_h(x,a,\cdot),\cdot)}(\xi)$ and $\overline{V}_{k,H+1} \equiv 0$, Exo-VTR-$\varepsilon$ computes, for $h = H, \dots, 1$,

$$\overline{Q}_{k,h}(x,\xi,a) = \min\Big\{H-h+1,\ \max\big\{0,\ r_h(x,a,\xi) + \psi_{k,h}^{\top}\widehat{\theta}_k + \beta_k\|\psi_{k,h}\|_{\Sigma_k^{-1}}\big\}\Big\}, \tag{16}$$

sets $\overline{V}_{k,h}(x,\xi) = \max_{a\in\mathcal{A}_h(x,\xi)} \overline{Q}_{k,h}(x,\xi,a)$, and plays the greedy policy for the whole episode. Appendix F states the procedure as Algorithm 2. The clipping is two-sided, unlike the single upper truncation of Algorithm 1, because $\psi_{k,h}^{\top}\widehat{\theta}_k$ need not be an expectation under a probability kernel; the lower clip keeps every stored target bounded on every sample path. Separately, $\lambda = H^2$ gives $\|\psi\|_{\Sigma_k^{-1}} \le 1$, which removes norm clipping cases from the potential argument.

Alignment and the frozen potential. The feature in the bonus at the played action is exactly the feature later stored for that transition, since both are built from the same frozen value function before the successor is drawn (Remark 12). This alignment allows an elliptical potential argument; with $\Gamma := \ln\big(\det\Sigma_{K+1}/\lambda^d\big) \le d\ln(1+KH/d)$, episode-wise freezing costs only an additive burn-in and leaves the leading $\sqrt{K}$ rate intact.

Theorem 5 (Regret of Exo-VTR-ε)**.** *Under Assumptions A1L, A2 and A3, with probability at least $1-2\delta$, Exo-VTR-ε satisfies*

$$\mathrm{Regret}_{M_\varepsilon}(K) \ \le\ 2\beta_K\Big(2\sqrt{KH\Gamma} + \frac{H\Gamma}{\ln 2}\Big) \ +\ \varepsilon\bar{X}HK \ +\ 2H\sqrt{2KH\ln(1/\delta)}.$$

Expanding β_K and Γ, the statistical contribution is $\widetilde{O}\big((d + B_\theta\sqrt{d})\sqrt{H^3K}\big)$ and the leading impact contribution is $O(\varepsilon\bar{X}HK) + \widetilde{O}\big(\varepsilon K\sqrt{d\,a_{\max}\bar{X}H^3}\big)$, with frozen update burn-in terms that Appendix F displays in full. As in Theorem 1, the statistical part retains the $\sqrt{K}$ rate of the exogenous model, now with d in place of $S_\bullet$, and every impact term vanishes at $\varepsilon = 0$. The per-step cap can again be removed; Corollary 30 gives the cap-free radius and resulting rates, with the comparison to the capped guarantee.

Action dependence of the bonus. Unlike the tabular bonus of Remark 2, the bonus $\beta_k\|\psi_{k,h}(x,\xi,a)\|_{\Sigma_k^{-1}}$ depends on the action, because f_h carries it into the value function whose aggregation defines the feature. It can therefore change the optimistic maximizer with neither clip active, the mechanism examined in Section 7.

Scope. Theorem 5 is of Hoeffding type; Appendix F records the Bernstein question as a conjecture. The statement assumes the exact integration and measurable planning oracles specified in Appendix F; it makes no general computational efficiency claim.

7 Experiments

The experiments below examine the mechanisms underlying the preceding results. Full protocols, seeds and the remaining families are in Appendix G.

7.1 Controlled instances

Setup. Four pre-declared environment families use $H = 20$, $S_\bullet = 5$, $K = 3000$ episodes and six matched seeds over $\varepsilon \in \{0, 0.02, 0.05, 0.10, 0.20\}$, comparing Exo-UCBVI-$\varepsilon$ with the low-bonus ablation, TWAP and an Almgren–Chriss schedule. The hard family of Section 5 is run in both

variants with $K = 1500$ and six seeds over $\varepsilon \in \{0, 0.01, 0.02, 0.05, 0.10\}$, comparing PTO as an exact exogeneity stress test, the pooled exogenous maximum likelihood planner of Theorem 4, Exo-UCBVI-ε and an impact-aware parametric maximum likelihood reference.

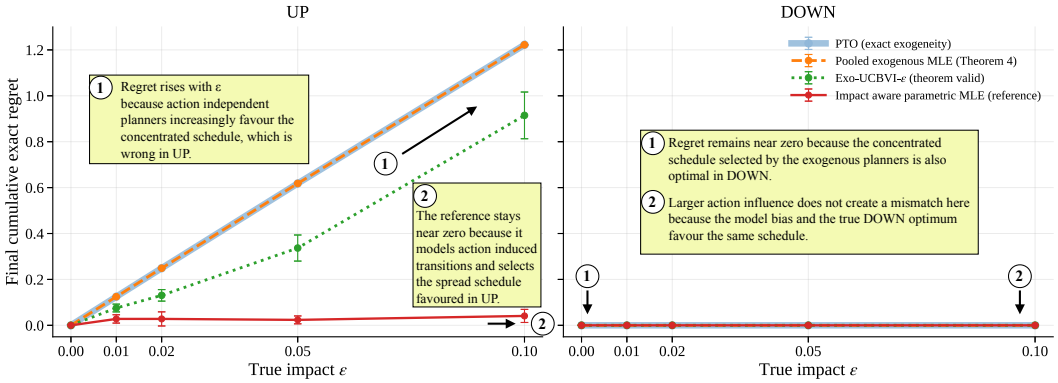


Figure 1: **Execution hard family of Theorems 3 and 4.** Final cumulative exact regret against the true action influence ε, for the UP variant (left) and DOWN variant (right). Error bars show 95% confidence intervals over six seeds.

Figure 1 reproduces the separation established in Section 5: in UP, the pooled exogenous MLE exhibits the first-order growth in ε predicted by Theorem 4, while the impact-aware reference stays at the $O(\varepsilon^2 K)$ level of Proposition 22. The same model bias, growing with ε, is present in both variants yet costs nothing in DOWN: the first-order loss arises not from an action-independent abstraction itself but from one whose induced schedule is wrong for the instance, which is why Theorem 4 is stated on UP.

On the persistent impact family, final regret is approximately linear in ε, as Theorem 1 suggests; the other families are non-monotone (Appendix G). For Exo-VTR, the elliptical bonus remains action-dependent and can change the greedy action before clipping (Appendix G), unlike the tabular bonus of Remark 2.

7.2 Historical market replay

Replay summary. On one year of Binance spot BTCUSDT one-minute prices, we calibrate a five-state simulator on the chronological calibration split, train Exo-UCBVI for 4000 simulated episodes, and evaluate the frozen policy on 2000 held-out blocks with $H = 20$. Public prices do not identify causal impact, so the kernel is action-independent ($\varepsilon = 0$) and η is an assumed temporary execution cost. The policy reproduces TWAP for $\eta \geq 0.0016$ and becomes state-dependent below it, with crossover $0.0008 < \eta_{\text{cross}} \leq 0.0016$, yet never beats TWAP on the tested grid; an information matched oracle fitted on held-out data also does not, suggesting that the state description rather than estimation error may be the binding limitation under this protocol (Appendix G.5).

8 Conclusion and limitations

We studied RL when actions weakly perturb an otherwise exogenous process. The tabular and linear-mixture guarantees retain statistical dependence on exogenous structure while accounting for action influence through action magnitude and the trajectory budget; in the tabular setting, PTO attains the same order without optimism. The execution-family lower bounds establish first-order loss for the specified exogenous planner classes, and a two-stage construction does so for Exo-UCBVI-ε, PTO and Exo-VTR-ε with a one-dimensional reference representation; neither is a minimax lower bound. Controlled experiments reproduce the separation and show how clipping and elliptical bonuses affect decisions. The analysis assumes known influence bounds, known deterministic endogenous maps and rewards, and the stated oracles; learning the influence bound, Bernstein upper bounds, richer state representations and observed order book impact remain open.

AI USE STATEMENT

AI tools were used to assist with code generation, experimental implementation, and manuscript editing. ChatGPT also assisted with proof checking of the algorithm-specific lower-bound arguments. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the manuscript's content, proofs, experiments, and conclusions.

ETHICS STATEMENT

This work is primarily theoretical and studies RL in approximately exogenous Markov decision processes under bounded action influence. It does not involve human subjects, personal data, or direct deployment in real-world systems, and therefore raises no immediate ethical concerns. However, the framework may be applied in domains such as inventory management, execution, and resource allocation, where automated decisions can affect individuals, organizations, or markets. In such applications, practitioners should account for domain-specific risks, including distributional effects, fairness, operational constraints, and unintended consequences. The theoretical guarantees in this work should not be interpreted as providing causal identification or as validating deployment in a particular application domain.

REPRODUCIBILITY STATEMENT

All assumptions used in the analysis are stated explicitly in the main text, and proofs of the theoretical results are provided in Appendices D and F, with the supporting lower-bound arguments in the corresponding appendix sections. The tabular and linear-mixture procedures are given in algorithmic form in the manuscript. The synthetic experiments and historical market replay are described in the experimental section and Appendix G.5. The anonymized code and experiment archive are provided in the supplementary material, with the dataset manifest and hash files and the configuration JSONs recording the settings used for each experimental campaign.

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.

Akram S. Awad, Shihab Ahmed, Yue Wang, and George K. Atia. Robust transfer learning with side information. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 255–286, 2026.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. 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. 463–474. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/ayoub20a.html.

Woojin Chae, Kihyuk Hong, Yufan Zhang, Ambuj Tewari, and Dabeen Lee. Learning infinite-horizon average-reward linear mixture MDPs of bounded span. In *International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2737–2745, 2025.

Thomas Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1262–1270. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/dietterich18a.html.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1genAVKPB.

Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5062–5127. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/efroni22a.html.

Yihong Guo, Yu Yang, Pan Xu, and Anqi Liu. MOBODY: Model-based off-dynamics offline reinforcement learning. In *International Conference on Learning Representations*, 2026.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=pJdMOKqdSV.

Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Horizon-free and instance-dependent regret bounds for reinforcement learning with general function approximation. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3673–3681. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/huang24b.html.

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger (eds.), *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 666–686. PMLR, 10–11 Jun 2020. URL https://proceedings.mlr.press/v120/jia20a.html.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. 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. 5662–5670. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lattimore20a.html.

Alexander Levine, Peter Stone, and Amy Zhang. Learning a fast mixing exogenous block MDP using a single trajectory. In *International Conference on Learning Representations*, 2025a.

Alexander Levine, Peter Stone, and Amy Zhang. Offline action-free learning of Ex-BMDPs by comparing diverse datasets. *Reinforcement Learning Journal*, 6:2452–2484, 2025b.

Yunfan Li and Lin Yang. On the model-misspecification in reinforcement learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2764–2772. PMLR, 2024. URL https://proceedings.mlr.press/v238/li24m.html.

Hao Liang, Jiayu Cheng, Sean Sinclair, and Yali Du. Is pure exploitation sufficient in exogenous mdps with linear function approximation? In *International Conference on Learning Representations*, volume 2026, pp. 15304–15349, 2026.

Zhishuai Liu and Pan Xu. Linear mixture distributionally robust Markov decision processes. *Advances in Neural Information Processing Systems*, 38:138326–138361, 2025.

Thodoris Lykouris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. *Mathematics of Operations Research*, 50(2):1277–1304, 2025.

Debmalya Mandal and Goran Radanovic. Performative reinforcement learning with linear Markov decision process. In *International Conference on Artificial Intelligence and Statistics*, 2025.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23642–23680. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/mandal23a.html.

Davide Maran, Davide Salaorni, and Marcello Restelli. Learning in markov decision processes with exogenous dynamics. *arXiv preprint arXiv:2603.02862*, 2026.

Vasilis Pollatos, Debmalya Mandal, and Goran Radanovic. On corruption-robustness in performative reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19939–19947, 2025.

Manoj Saravanan and Rohit Kumar Salla. When can pure exploitation succeed in linear RL? Decoys and self-identifiability for greedy LSVI. *Reinforcement Learning Journal*, 7, 2026.

Sean R. Sinclair, Felipe Vieira Frujeri, Ching-An Cheng, Luke Marshall, Hugo De Oliveira Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, and Adith Swaminathan. Hindsight learning for MDPs with exogenous inputs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31877–31914. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/sinclair23a.html.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Improved algorithms for misspecified linear markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 4723–4746. PMLR, 2022.

Jia Wan, Sean R. Sinclair, Devavrat Shah, and Martin J Wainwright. Exploiting exogenous structure for sample-efficient reinforcement learning. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024. URL https://openreview.net/forum?id=z55DUMR4bE.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Chenlu Ye, Jiafan He, Quanquan Gu, and Tong Zhang. Towards robust model-based reinforcement learning against adversarial corruption. *arXiv preprint arXiv:2402.08991*, 2024.

Weitong Zhang, Zhiyuan Fan, Jiafan He, and Quanquan Gu. Achieving constant regret in linear markov decision processes. *Advances in Neural Information Processing Systems*, 37:130694–130738, 2024.

Zhongjun Zhang and Sean R. Sinclair. Offline-online reinforcement learning for linear mixture MDPs. *arXiv preprint arXiv:2604.11994*, 2026.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4532–4576. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/zhou21a.html.

A TABLE OF NOTATION

Table 1: List of common notation used throughout the paper.

Symbol	Definition
Exo-MDP specification and bounded action influence	
$\mathcal{X}$	Endogenous state space.
Ξ	Exogenous state space.
$S_\bullet = \lvert\Xi\rvert$	Number of exogenous states in the tabular setting.
$\mathcal{A} \subset \mathbb{R}$	Scalar action space.
H	Episode horizon.
K	Number of episodes.

Continued on next page

Table 1 – continued

Symbol	Definition
$s_h = (x_h, \xi_h)$	Full state at stage h, consisting of endogenous and exogenous coordinates.
$x_h \in \mathcal{X}$	Endogenous state at stage h.
$\xi_h \in \Xi$	Exogenous state at stage h.
$a_h \in \mathcal{A}$	Action selected at stage h; $\lvert a_h\rvert$ denotes its scalar magnitude.
$r_h(x, a, \xi) \in [0, 1]$	Known deterministic planning reward at stage h.
$f_h(x, a, \xi')$	Known endogenous update map; the next endogenous state satisfies $x_{h+1} = f_h(x_h, a_h, \xi_{h+1})$.
s_1	Fixed initial state.
$P_0(\cdot \mid \xi)$	Reference exogenous transition kernel, shared across stages and independent of the action.
$P_\varepsilon(\cdot \mid \xi, x, a)$	Actual exogenous transition kernel encountered by the learner.
ε	Action-influence parameter satisfying $\lVert P_\varepsilon(\cdot \mid \xi, x, a) - P_0(\cdot \mid \xi)\rVert_1 \le \varepsilon\lvert a\rvert$.
$M_0,\ M_\varepsilon$	Reference and actual MDPs, respectively; they differ only in the exogenous transition kernel.
$\bar{X}$	Pathwise action budget: $\sum_{h=1}^{H} \lvert a_h\rvert \le \bar{X}$.
$a_{\max}$	Per-stage action cap, $\lvert a_h\rvert \le a_{\max}$, when the capped setting is used.
$\mathsf{b}(x)$	Remaining action budget encoded in the endogenous state when it is not already determined by x.
$\mathcal{A}_h(x, \xi)$	Known feasible action set at stage h, incorporating the remaining-budget and instance-specific constraints.
$\pi = (\pi_h)_{h=1}^{H}$	Admissible policy; in the main formulation, a possibly randomized non-stationary Markov policy.
π_k	Policy selected at the beginning of episode k and held fixed throughout that episode.
$V^{\pi}_{M,h}(s)$	Expected cumulative reward from stage h onward under policy π in model M.
$V^{\star}_{M,h}(s)$	Optimal value from stage h onward in model M, over the relevant admissible policy class.
$V^{\pi}_{\varepsilon,h},\ V^{\pi}_{0,h}$	Abbreviations for $V^{\pi}_{M_\varepsilon,h}$ and $V^{\pi}_{M_0,h}$, respectively.
$Q^{\star}_{0,h}(x, \xi, a)$	Optimal reference-model state-action value at stage h.
$\mathrm{Regret}_{M_\varepsilon}(K)$	Cumulative regret over K episodes against the optimal policy in the actual model M_ε.
δ	Failure-probability parameter used in the high-probability guarantees.
Tabular estimation, Exo-UCBVI, and PTO	
$n_k(\xi)$	Number of transition-generating visits to exogenous state ξ before episode k, pooled across stages.
$\widehat{P}_k(\cdot \mid \xi)$	Empirical exogenous successor distribution formed from the pooled observations available before episode k.
c_β	Tabular concentration constant $c_\beta = \sqrt{2(S_\bullet \ln 2 + \ln(S_\bullet KH/\delta))}$.
β_n	Tabular ℓ_1 statistical radius: $\beta_0 = 2$ and $\beta_n = c_\beta/\sqrt{n}$ for $n \ge 1$.
$b_{k,h}(\xi)$	Tabular Exo-UCBVI bonus $\frac{1}{2}(H - h)\big(\beta_{n_k(\xi)} + \varepsilon a_{\max}\big)$.
$\overline{Q}_{k,h},\ \overline{V}_{k,h}$	Optimistic action-value and value functions used by Exo-UCBVI-ε.
$\bar{a}_k(\xi)$	Empirical average action magnitude over the visits to ξ available before episode k; used in the cap-free tabular construction.
$b'_{k,h}(\xi)$	Cap-free tabular bonus obtained by replacing $\varepsilon a_{\max}$ with $\varepsilon \bar{a}_k(\xi)$.
$\widehat{M}_k$	Empirical exogenous MDP with kernel $\widehat{P}_k$, used by the no-bonus PTO planner.
$\widetilde{Q}_{k,h},\ \widehat{V}_{k,h}$	PTO action-value and value functions computed in $\widehat{M}_k$ without an optimism bonus.
$\widehat{\pi}_k$	PTO maximizing policy selected in $\widehat{M}_k$ for episode k.

Continued on next page

Table 1 – continued

Symbol	Definition
$E_{k,h}(\xi)$	Analysis-only PTO error envelope controlling the discrepancy between reference-model and empirical-model policy values.
Hard-family lower-bound notation	
$G,\ B$	Good and bad exogenous states in the two-state hard construction; B is absorbing.
$g \in (0,1)$	Price-gap parameter of the hard family.
$p(\xi)$	Price attached to exogenous state ξ in the hard construction, with $p(G) = 1$ and $p(B) = 1 \pm g$.
$v \in \{\mathrm{UP}, \mathrm{DOWN}\}$	Index of the two hard-family variants.
$\mathcal{S}$	Feasible full-liquidation schedule simplex in the hard construction: $\{a \in [0, \frac{1}{2}]^3 : \sum_h a_h = 1\}$.
$e_2(a)$	Second elementary symmetric polynomial of the schedule, $e_2(a) = \sum_{i<j} a_i a_j$, measuring its first-order concentration effect in the hard family.
C_g	Finite constant depending only on g, used to control the uniform $O(\varepsilon^2)$ remainder in the hard-family bounds.
$\widehat{Q}_k(\cdot \mid \xi)$	Action-independent kernel selected by a consistent exogenous planner in episode k.
q_k	Selected-model probability $\widehat{Q}_k(B \mid G)$ in episode k.
$q_{B,k}$	Selected-model probability $\widehat{Q}_k(B \mid B)$ in episode k.
T	Random pinning time of the pooled point-estimate planner.
T_{cs}	Deterministic burn-in threshold after which the confidence-set planner is pinned on the high-probability event.
N_B	Number of transitions with source state B required to exclude the hedging kernel in the confidence-set analysis.
$\mathcal{E}$	High-probability pinning event used in the confidence-set lower bound.
$\hat{p}$	Empirical Bernoulli transition probability used in the entrywise Bernstein admissible set.
r_n	Entrywise Bernstein radius $\sqrt{4\hat{p}(1-\hat{p})\Lambda/n} + 5\Lambda/n$ for $n \geq 1$.
J	Dyadic peeling factor $1 + \lceil \log_2(KH) \rceil$ in the Bernstein analysis.
Λ	Bernstein logarithmic factor $\ln(4S_\bullet^2 KHJ/\delta)$.
T_B	Burn-in threshold for the Bernstein-set planner.
Linear-mixture model and Exo-VTR	
d	Linear-mixture feature dimension.
$\phi(\xi', \xi) \in \mathbb{R}^d$	Known linear-mixture feature map for the reference exogenous kernel.
ν	Dominating measure on the exogenous space Ξ.
$\theta^\star \in \mathbb{R}^d$	Unknown parameter shared by the reference kernel across stages, satisfying $\|\theta^\star\|_2 \leq B_\theta$.
B_θ	Known upper bound on $\|\theta^\star\|_2$.
$\psi_g(\xi)$	Integrated feature $\int_\Xi \phi(\xi', \xi) g(\xi')\, \nu(d\xi')$.
D_{k-1}	Set of transition-generating steps from episodes $1, \ldots, k-1$ used by Exo-VTR before episode k.
$t,\ h_t$	Global transition index and the stage associated with that transition.
$V_t(\xi')$	Frozen continuation function associated with sample t: $\overline{V}_{j,h_t+1}(f_{h_t}(x_t, a_t, \xi'), \xi')$.
ψ_t	Regression feature $\psi_{V_t}(\xi_t)$ computed before observing ξ_{t+1}.
y_t	Regression target $V_t(\xi_{t+1})$ observed after the successor draw.
λ	Ridge regularization parameter, fixed to $\lambda = H^2$.
Σ_k	Episode-k regularized design matrix $\lambda I + \sum_{t \in D_{k-1}} \psi_t \psi_t^\top$.
$\widehat{\theta}_k$	Ridge estimate $\Sigma_k^{-1} \sum_{t \in D_{k-1}} \psi_t y_t$ of $\theta^\star$.

Continued on next page

Table 1 – continued

Symbol	Definition
ζ_t	Contamination-bias term $\langle P_\varepsilon(\cdot \mid \xi_t, x_t, a_t) - P_0(\cdot \mid \xi_t), V_t\rangle$.
ω_t	Conditionally mean-zero regression noise $y_t - \langle P_\varepsilon(\cdot \mid \xi_t, x_t, a_t), V_t\rangle$.
β_k	Exo-VTR confidence radius containing regularization, martingale-noise, and contamination terms.
$\psi_{k,h}(x, \xi, a)$	Planning feature $\psi_{\overline{V}_{k,h+1}(f_h(x,a,\cdot),\cdot)}(\xi)$.
$\overline{Q}_{k,h}$, $\overline{V}_{k,h}$	Two-sided clipped optimistic action-value and value functions used by Exo-VTR-ε.
Γ	Log-determinant potential $\ln(\det \Sigma_{K+1}/\lambda^d)$, bounded by $d\ln(1 + KH/d)$.
$\mathcal{H}_t$	Pre-successor filtration at global step t in the Exo-VTR analysis.
$\mathcal{F}_t$	Post-successor filtration $\sigma(\mathcal{H}_t, \xi_{t+1})$.
$\mathcal{G}_{\text{lin}}$	High-probability confidence-ellipsoid event $\{\|\widehat{\theta}_k - \theta^\star\|_{\Sigma_k} \le \beta_k \text{ for all } k\}$.
Historical market replay	
p_h	Observed relative market price at stage h of a held-out replay block.
η	Assumed temporary execution-impact coefficient appearing in the replay payoff; it is distinct from the transition-influence parameter ε.
τ	Proportional execution fee.
r_h^{replay}	Unnormalized held-out execution payoff $a_h(p_h - \eta a_h) - \tau a_h$.
$\mathcal{R}(\pi)$	Total realized execution revenue of policy π on a held-out block.
$c(\pi)$	Paired execution cost relative to TWAP, reported in basis points.
TWAP	Time-weighted average-price execution benchmark.

B Detailed related work

Learning with exogenous structure. A growing line of work exploits exogenous information to reduce the statistical difficulty of reinforcement learning (RL). Dietterich et al. (2018) study the identification and removal of exogenous state variables and rewards, while Efroni et al. (2022) consider observations that factor into a small controllable component and a large exogenous component. Sinclair et al. (2023) formulate MDPs with exogenous inputs and use observed exogenous traces for hindsight learning, while Wan et al. (2024) show that discrete Exo-MDPs induce a linear representation through the exogenous distribution, yielding regret guarantees whose leading dependence can avoid the sizes of the endogenous state and action spaces. More recently, Liang et al. (2026) show that exact action independence can make deliberate exploration unnecessary, with pure exploitation attaining finite sample regret guarantees in tabular and linear function approximation settings. Maran et al. (2026) study MDPs with action-independent exogenous dynamics and obtain regret bounds whose leading terms depend on the exogenous state space with a matching lower bound for their setting. Our work instead allows action independence to fail in a bounded, action linked manner and asks how much of the exogenous advantage survives and what first-order cost is unavoidable on the hard family for the planner classes of Theorems 3 and 4.

A parallel line of work studies exogenous structure that is latent rather than observed. In the Exogenous Block MDP formulation, observations decompose into a controllable factor evolving deterministically under the agent's actions and a temporally correlated exogenous noise factor, and the learning problem is to recover an encoder onto the controllable latent space. Levine et al. (2025a) give STEEL, a sample efficient algorithm for learning the controllable dynamics from a single trajectory under function approximation, with sample complexity depending on the controllable latent space and the encoder class and, at worst linearly, on the mixing time of the exogenous factor. Levine et al. (2025b) show that action free offline data drawn from several agents with differing policies makes the controllable representation recoverable, using differences in controllable feature dynamics across datasets. These works recover exogenous structure the learner cannot see, whereas we take the decomposition as given and ask what happens when the independence it asserts holds

only approximately. From the opposite direction, Saravanan & Salla (2026) ask when exploration free learning can succeed in linear RL without appealing to exogenous structure at all, characterizing through decoy constructions and self identifiability when exploration free learning can avoid linear regret; under additional regularity conditions, their Greedy-LSVI algorithm attains sublinear regret. Their route to pure exploitation is an identifiability property of the instance; the route in Liang et al. (2026) is action independence of the exogenous process, and it is the latter that our ε relaxation weakens.

Decision dependent dynamics, misspecification, and robustness. Our model is also related to settings in which the transition law is imperfect or changes with the learner's decisions. Performative RL allows the deployed policy to modify rewards and transition dynamics (Mandal et al., 2023; Mandal & Radanovic, 2025); in contrast, we retain a reference exogenous process and study bounded deviations induced by individual actions. In a related performative setting, Pollatos et al. (2025) study performative RL under Huber ε contamination, in which an ε fraction of data points is corrupted by arbitrary adversarial noise, and obtain last iterate convergence to an approximately stable policy with approximation error linear in $\sqrt{\varepsilon}$. Their contamination corrupts the data from which a policy dependent environment is estimated; ours instead constrains the transition law itself by an amount tied to the magnitude of the action played and is not modeled as adversarial sample corruption. For model misspecification, Vial et al. (2022) derive regret guarantees that degrade with the misspecification level, while Zhang et al. (2024) obtain instance dependent regret that is constant in the number of episodes when misspecification is sufficiently small relative to the suboptimality gap. Du et al. (2020) show that an approximately accurate representation alone need not suffice for sample efficient learning, while Lattimore et al. (2020) characterize the tradeoff between approximation quality and learning complexity in misspecified linear bandits and RL with a generative model. Theorem 4 isolates a different obstruction: a first-order loss for the specified consistent exogenous planners on the hard family under bounded action linked violations. Corruption robust RL permits adversarial changes to rewards or transitions (Lykouris et al., 2025; Wei et al., 2022; Ye et al., 2024), while distributionally robust off dynamics RL addresses discrepancies between nominal and deployment models (He et al., 2025). We impose a specific action magnitude bound and pathwise budget on deviations from an exogenous kernel, and our goal is to quantify the price of preserving that abstraction rather than to optimize against general transition uncertainty.

Recent work on transition mismatch has moved in two complementary directions, both of which sharpen that contrast. Guo et al. (2026) address the conservatism of reward penalization and data filtering in off dynamics offline RL by learning the target domain dynamics from source data and limited target data and optimizing against rollouts from the learned model, rather than restricting attention to source transitions in low shift regions. Awad et al. (2026) retain the robust optimization perspective while tightening the uncertainty model, constructing estimate-centered uncertainty sets from limited target samples and side information, including bounds on feature moments, distributional distances, and density ratios, and deriving finite-sample guarantees for the resulting robust policy. Both treat the mismatch as a property of two environments, to be learned around or hedged against; in our setting the deviation is generated by the learner within a single environment, is indexed by the action it plays, and is absorbed into an optimistic radius rather than resolved by estimating the deviating kernel or by playing a worst-case within a set.

Linear mixture MDPs and value targeted regression. Our linear mixture analysis builds on value targeted model estimation, which regresses transition models against downstream value functions (Jia et al., 2020; Ayoub et al., 2020). Zhou et al. (2021) develop a computationally efficient, nearly minimax optimal algorithm for linear mixture MDPs, while subsequent work develops weighted value targeted regression with sharper horizon and instance dependent guarantees (Huang et al., 2024). Wan et al. (2024) connect this structure directly to Exo-MDPs through the exogenous distribution. Relative to the realizable models just discussed, Exo-VTR-ε operates in a different regime: the reference exogenous kernel retains linear mixture structure, while the observed dynamics contain systematic action linked deviations. Its analysis must therefore control both statistical estimation error and contamination bias while retaining the structural benefits of value targeted estimation.

For misspecified models, Li & Yang (2024) already analyze Robust-UCRL-VTR; their published Theorem 5.6 and Remark 5.8 include linearly parameterized transitions. Stored value targets and

confidence widening for bias are also present in their analysis. Their assumption controls policy occupancy moments through order four, so it can also benefit from a trajectory budget. Here Lemmas 24 and 7 give explicit pathwise dependence on the action cap and total volume. The contribution is this budget dependence and its exogenous interpretation, rather than a new generic statistical plus linear misspecification rate.

The linear mixture model has since been extended in several directions relevant here. Chae et al. (2025) treat the infinite horizon average reward case under the Bellman optimality condition, combining a span clipped value iteration argument with weighted ridge regression to attain a nearly minimax optimal regret bound of order $\widetilde{O}(d\sqrt{\mathrm{sp}(v^\star)T})$, where $\mathrm{sp}(v^\star)$ is the span of the optimal bias function. Liu & Xu (2025) introduce linear mixture distributionally robust MDPs, placing the uncertainty on the mixture weighting parameters rather than directly on the transition kernel; their main theoretical procedure uses transition-targeted ridge regression, and they also study a value-targeted robust variant empirically. That construction is the natural point of contrast for Exo-VTR-ε: they optimize against a worst-case within a parameter ball, whereas we retain a single reference parameter $\theta^\star$ and absorb the action-linked deviation into the confidence radius Eq. (15), which is what preserves the exogenous estimation structure rather than replacing it. Closer to the data side, Zhang & Sinclair (2026) study offline online learning in linear mixture MDPs under environment shift, where offline data collected by an unknown behavior policy may come from a mismatched environment, and give an algorithm that provably improves over purely online learning when that data is informative and matches online only performance when it is not, with regret upper bounds and nearly matching lower bounds characterizing the boundary. Their shift is between the environment that produced the data and the one the learner faces; ours is between the reference kernel the learner models and the kernel its own actions induce, within a single online interaction.

C Model details and illustration

The three modeling regimes discussed in Section 3 are illustrated below.

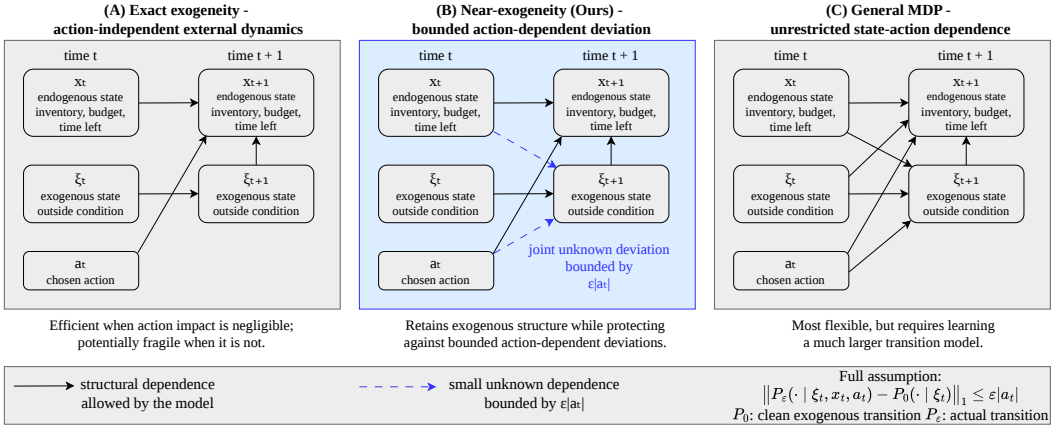


Figure 2: One transition step under three modeling assumptions. In each panel x_t is the endogenous coordinate, ξ_t the exogenous coordinate, and a_t the action. Solid arrows mark dependence the model permits outright; the dashed arrows in (B) mark the additional dependence admitted by Assumption A2, whose ℓ_1 distance from the reference kernel is at most $\varepsilon|a_t|$. Panel (A) is the reference model M_0 of Eq. (2) and panel (B) the model M_ε of Eq. (3) actually encountered by the learner. In panels (A) and (B), the known update is $x_{t+1} = f_t(x_t, a_t, \xi_{t+1})$, so a solid arrow $\xi_{t+1} \to x_{t+1}$ is also permitted under exact exogeneity.

C.1 Control and budget conventions

In the scalar-action model of Section 3, $|a|$ denotes the ordinary absolute magnitude of the control. Vector or categorical controls would require a known non-negative magnitude or cost map in both Assumptions A2 and A3; we retain the scalar formulation.

If $0 \in \mathcal{A}$ represents a no-action control, then wherever it is feasible, Assumption A2 implies $P_\varepsilon(\cdot \mid \xi, x, 0) = P_0(\cdot \mid \xi)$. This is a restriction on the exogenous transition law; the state may still change under that control. Availability of a zero control is not required by the results. Note that Eq. (4) constrains only the exogenous kernel; the endogenous dynamics remain exactly known through Eq. (1). The resulting model lies between exact exogeneity and unrestricted action dependence, as Fig. 2 illustrates. Exact exogeneity allows the action to reach the endogenous coordinate alone; an unrestricted MDP allows it to reach both without qualification; Eq. (4) admits the second channel but caps how much can pass through it. The relaxation is thus quantitative rather than structural, and it closes continuously as $\varepsilon \to 0$.

The cap alone gives total volume at most $Ha_{\max}$, so the trajectory budget gives a smaller bound when $\bar{X} < Ha_{\max}$. Conversely, the trajectory budget already implies $|a_h| \le \bar{X}$. Thus cap-free means that no separate $a_{\max}$ restriction is imposed, while the remaining-budget constraint is retained.

C.2 Successor-dependent inventory example

Inventory control, also considered by Liang et al. (2026), illustrates successor dependence. Let x_h denote net inventory (negative for backorders), $a_h \ge 0$ the order quantity, $d_h \ge 0$ known deterministic consumption, and $\xi_h \in \{1/2, 1\}$ an observed supplier condition specifying the delivery fraction, reported even without an order. Then

$$x_{h+1} = x_h + a_h \xi_{h+1} - d_h.$$

The same order yields different next inventories under full and partial delivery. Assumption A2 can model weak order-induced supplier congestion, while Assumption A3 bounds total ordered volume. Known, bounded costs based on current inventory and orders give rewards in the stated range after normalization.

C.3 Admissibility

Write $\mathsf{b}(x)$ for the remaining action budget. If it is not already determined by x, append it to the endogenous coordinate and continue to write x and f_h for the augmented state and known map, with

$$\mathsf{b}(x_1) = \bar{X}, \qquad \mathsf{b}\big(f_h(x, a, \xi')\big) = \mathsf{b}(x) - |a|.$$

The rewards and exogenous kernels ignore this added coordinate. The motivating inventory equations display only the physical coordinate when this augmentation is needed.

Let $\mathcal{A}_h(x, \xi) \subseteq \mathcal{A}$ denote the known set of actions satisfying $|a| \le \mathsf{b}(x)$ and the instance's existing constraints, and admitting a feasible continuation through the horizon for every candidate exogenous successor. In capped statements also impose $|a| \le a_{\max}$; in cap-free statements omit this separate restriction for both learner and comparator. The initial state is feasible. These sets are nonempty on feasible states, are the same in the actual, reference and planning models, and are subject to the existing measurability and planning-oracle requirements. A policy is admissible if it selects from these sets on every feasible history. Throughout, all policy quantifiers, including those for history dependent policies, concern admissible policies of the specified class; all Bellman equations and action maximizations concern feasible states and actions.

C.4 Bellman equations

The reference optimal Bellman equations use the same successor in both coordinates:

$$Q^\star_{0,h}(x, \xi, a) = r_h(x, a, \xi) + \int_\Xi V^\star_{0,h+1}\big(f_h(x, a, \xi'), \xi'\big) P_0(d\xi' \mid \xi),$$
$$V^\star_{0,h}(x, \xi) = \sup_{a \in \mathcal{A}_h(x,\xi)} Q^\star_{0,h}(x, \xi, a), \qquad V^\star_{0,H+1} \equiv 0.$$

Policy evaluation averages the corresponding action values under $\pi_h(da \mid x, \xi)$. The actual and empirical Bellman equations use $P_\varepsilon(\cdot \mid \xi, x, a)$ and $\widehat{P}_k(\cdot \mid \xi)$, respectively, with the same successor-dependent map.

Algorithm 1 Exo-UCBVI-ε (frozen counts)

input: H, K, ε, $a_{\max}$, δ
for $k = 1, \dots, K$ **do**
  form $n_k(\xi)$ and $\widehat{P}_k(\cdot \mid \xi)$ from episodes $1, \dots, k-1$
  $\overline{V}_{k,H+1} \leftarrow 0$
  for $h = H, \dots, 1$ **do**
    $\overline{Q}_{k,h}(x,\xi,a) \leftarrow \min\{H-h+1,\ r_h(x,a,\xi) + \sum_{\xi'} \widehat{P}_k(\xi' \mid \xi)\, \overline{V}_{k,h+1}(f_h(x,a,\xi'),\xi') + b_{k,h}(\xi)\}$
    $\overline{V}_{k,h}(x,\xi) \leftarrow \max_{a \in \mathcal{A}_h(x,\xi)} \overline{Q}_{k,h}(x,\xi,a)$, $\pi_{k,h}(x,\xi) \in \arg\max_{a \in \mathcal{A}_h(x,\xi)} \overline{Q}_{k,h}(x,\xi,a)$
  end for
  play $\pi_k = (\pi_{k,h})_{h=1}^H$ for the whole episode in M_ε
end for

D Proofs for the tabular setting

Throughout this appendix Ξ is finite, $|\Xi| = S_\bullet \geq 2$, and Assumptions A2 and A3 are in force. Observed exogenous transitions within an episode are the pairs (ξ_h, ξ_{h+1}) for $h = 1, \dots, H-1$, so there are at most KH transitions in total. For reference, the procedure analyzed below is stated in full as Algorithm 1.

D.1 Change of measure and the simulation gap

Lemma 6 (Change of measure)**.** *For distributions p, q on Ξ and any g of finite span, $\left|\sum_{\xi'}(p-q)(\xi')\, g(\xi')\right| \leq \frac{1}{2}\|p-q\|_1 \operatorname{span}(g)$.*

Proof. Let $c = \frac{1}{2}(\max g + \min g)$. Since p and q are probability distributions, $\sum_{\xi'}(p-q)(\xi')\, c = 0$, so the sum equals $\sum_{\xi'}(p-q)(\xi')(g(\xi')-c)$. The midpoint of an interval lies within half its length of every point of the interval, so $|g-c| \leq \frac{1}{2}\operatorname{span}(g)$ pointwise, and Hölder's inequality gives the claim. □

In the applications below, each continuation is evaluated at $(f_h(x,a,\xi'),\xi')$. Its span as a function of ξ' is bounded using the value bounds over all full states. A span bound for a single fixed-endogenous-state slice is not used.

Lemma 7 (Volume weighted simulation)**.** *Under Assumptions A2 and A3, for every policy π, possibly history dependent and randomized,*

$$\left|V^\pi_{\varepsilon,1}(s_1) - V^\pi_{0,1}(s_1)\right| \ \leq\ \frac{\varepsilon H \bar{X}}{2}. \tag{17}$$

Proof. For $j = 0, \dots, H$ let $M^{(j)}$ be the hybrid model that uses P_ε at steps $1, \dots, j$ and P_0 afterwards, with f_h throughout, and set $W_j := V^\pi_{M^{(j)},1}(s_1)$, so that $V^\pi_{\varepsilon,1} - V^\pi_{0,1} = \sum_{j=1}^H (W_j - W_{j-1})$.

Fix j. The hybrids $M^{(j)}$ and $M^{(j-1)}$ share their dynamics through step $j-1$ and share π, so the law μ_j of the prefix $\tau_j = (s_1, a_1, \dots, s_j, a_j)$ is common to both. Because π may be history dependent, define the conditional idealized continuation

$$G_{j+1}(\tau_j, s') := \mathbb{E}_{M_0,\pi}\Big[\sum_{h=j+1}^{H} r_h \ \Big|\ \tau_j,\ s_{j+1} = s'\Big] \in [0, H-j],$$

which is well defined because both hybrids run P_0 after step j, and the future actions of π are functions of the history, whose law is determined by (τ_j, s') and P_0; the range is the value bound of Section 3 applied conditionally.

The two hybrids differ only in the step j kernel, and the map $\xi' \mapsto f_j(x_j, a_j, \xi')$ is common to both, so the reward prefixes cancel in law and

$$W_j - W_{j-1} = \mathbb{E}_{\mu_j}\Big[\sum_{\xi'}\big(P_\varepsilon(\xi' \mid \xi_j, x_j, a_j) - P_0(\xi' \mid \xi_j)\big)\, G_{j+1}\big(\tau_j, (f_j(x_j, a_j, \xi'), \xi')\big)\Big].$$

Applying Lemma 6 at each fixed τ_j, to the composed continuation, whose range is $[0, H-j]$, and with Assumption A2,

$$|W_j - W_{j-1}| \le \tfrac{1}{2}(H-j)\,\varepsilon\,\mathbb{E}_{\mu_j}|a_j| \le \tfrac{H}{2}\,\varepsilon\,\mathbb{E}_{M_\varepsilon,\pi}|a_j|.$$

Summing over j and applying the pathwise budget of Assumption A3 gives Eq. (17). □

Remark 3 (Why the assembly is not routed through a regret transfer)**.** Writing $\Delta := \sup_\pi |V^\pi_{\varepsilon,1} - V^\pi_{0,1}| \le \varepsilon H\bar{X}/2$, a direct transfer argument gives $\sum_k [V^\star_{\varepsilon,1} - V^{\pi_k}_{\varepsilon,1}] \le \sum_k [V^\star_{0,1} - V^{\pi_k}_{0,1}] + 2\Delta K$, since the per episode decomposition $(V^\star_\varepsilon - V^\star_0) + (V^\star_0 - V^{\pi_k}_0) + (V^{\pi_k}_0 - V^{\pi_k}_\varepsilon)$ pays Δ twice. The assembly below avoids this for two reasons. Base guarantees stated for M_0 presume M_0 generated data, whereas the data are generated by M_ε, and the pigeonhole argument requires the states whose bonuses are summed to be the states whose counts grow, which happens along M_ε trajectories. The assembly therefore works under the true measure and pays Δ once per episode.

D.2 Contaminated estimation

Fix ξ and index its transition-generating visits chronologically by i, writing (x_i, a_i) for the state and action at visit i and ξ'_i for the successor. Define the retrospective mixture

$$\bar{P}_n(\cdot \mid \xi) := \frac{1}{n}\sum_{i=1}^{n} P_\varepsilon(\cdot \mid \xi, x_i, a_i), \tag{18}$$

and decompose $\widehat{P}_k - P_0 = (\widehat{P}_k - \bar{P}_{n_k(\xi)}) + (\bar{P}_{n_k(\xi)} - P_0)$.

Lemma 8 (Bias)**.** $\big\|\bar{P}_n(\cdot \mid \xi) - P_0(\cdot \mid \xi)\big\|_1 \le \varepsilon\,\bar{a}(\xi) \le \varepsilon a_{\max}$*, where* $\bar{a}(\xi)$ *is the average volume over the first* n *visits to* ξ*.*

Proof. $P_0(\cdot \mid \xi)$ is the average of n copies of itself, so by the triangle inequality and Assumption A2, $\|\bar{P}_n - P_0\|_1 \le \frac{1}{n}\sum_i \|P_\varepsilon(\cdot \mid \xi, x_i, a_i) - P_0(\cdot \mid \xi)\|_1 \le \frac{\varepsilon}{n}\sum_i |a_i|$, which is $\varepsilon\bar{a}(\xi)$, and at most $\varepsilon a_{\max}$ by Assumption A3. □

Lemma 9 (Noise)**.** *With probability at least* $1-\delta$*, simultaneously for every* ξ *and every* $n \in \{1, \dots, KH\}$ *for which the first* n *visits to* ξ *occur, the estimator formed from those visits satisfies* $\|\widehat{P}_n - \bar{P}_n\|_1 \le \beta_n$ *with* β_n *as in Eq.* (9)*.*

Proof. Step 1: sign vector representation. For every real vector v on Ξ,

$$\|v\|_1 = \max_{f \in \{-1,+1\}^{S_\bullet}} \langle f, v\rangle,$$

the maximum being attained coordinatewise at $f(\xi') = \mathrm{sign}(v(\xi'))$. Applying this with $v = \widehat{P}_n - \bar{P}_n$, and writing $P_i := P_\varepsilon(\cdot \mid \xi, x_i, a_i)$ for the law governing the ith successor draw,

$$\|\widehat{P}_n - \bar{P}_n\|_1 = \max_{f \in \{-1,+1\}^{S_\bullet}} \frac{1}{n}\sum_{i=1}^{n}\big(f(\xi'_i) - \langle f, P_i\rangle\big). \tag{19}$$

No separate lower tail statement is needed, since $-f$ also lies in the sign vector class.

Step 2: a stopped global time martingale. Index transition opportunities by a global time $t \le T_{\max} \le KH$. Let $\mathcal{H}_t$ contain the interaction and internal randomness through the current state and the chosen action, but not the successor draw, and let $\mathcal{F}_t := \sigma(\mathcal{H}_t, \xi_{t+1})$, so that $\mathcal{F}_{t-1} \subseteq \mathcal{H}_t \subseteq \mathcal{F}_t$. The update Eq. (1) makes the next endogenous coordinate measurable once ξ_{t+1} is observed, so the post-transition sigma-field already contains the full successor state. Fix ξ and a sign vector f, put

$e_t := \mathbf{1}\{\xi_t = \xi\}$, which is $\mathcal{F}_{t-1}$ measurable because the current state is known after the preceding transition, and set

$$D_t^f := e_t\big(f(\xi_{t+1}) - \langle f, P_\varepsilon(\cdot \mid \xi, x_t, a_t)\rangle\big), \qquad M_u^f := \sum_{t \le u} D_t^f.$$

Conditionally on $\mathcal{H}_t$ the successor is drawn from $P_\varepsilon(\cdot \mid \xi, x_t, a_t)$, so $\mathbb{E}[D_t^f \mid \mathcal{H}_t] = 0$, and by the tower rule the same holds given $\mathcal{F}_{t-1}$. When $e_t = 1$ the increment lies in an interval of length at most 2, and when $e_t = 0$ it vanishes, so for every $\alpha \in \mathbb{R}$ Hoeffding's lemma gives $\mathbb{E}[e^{\alpha D_t^f} \mid \mathcal{H}_t] \le \exp(\alpha^2 e_t/2)$, and again the same bound conditionally on $\mathcal{F}_{t-1}$.

Let τ_n be the time of the nth selected visit, with $\tau_n = \infty$ if it never occurs, and let $\sigma_n := \tau_n \wedge T_{\max}$. Since $\sum_{t \le \sigma_n} e_t \le n$, the stopped exponential supermartingale argument yields, for every $\beta > 0$,

$$\mathbb{P}\big(\tau_n \le T_{\max},\ \langle f, \widehat{P}_n - \bar{P}_n\rangle \ge \beta\big) \le \exp\Big(-\frac{n\beta^2}{2}\Big), \tag{20}$$

because on $\{\tau_n \le T_{\max}\}$ the stopped sum equals $n\langle f, \widehat{P}_n - \bar{P}_n\rangle$, and the statement is vacuous otherwise. This stopped construction is what makes the random visit times harmless.

Step 3: union bound. By Eq. (19), if $\|\widehat{P}_n - \bar{P}_n\|_1 > \beta$ then at least one of the $2^{S_\bullet}$ sign vectors satisfies $\langle f, \widehat{P}_n - \bar{P}_n\rangle > \beta$. A fixed state generates at most one transition per global opportunity, so n ranges over at most KH values. Assigning failure probability $\delta/(2^{S_\bullet} S_\bullet KH)$ to each of the at most $2^{S_\bullet} S_\bullet KH$ statements and solving $\exp(-n\beta_n^2/2) \le \delta/(2^{S_\bullet} S_\bullet KH)$ gives $\frac{1}{2} n\beta_n^2 \ge S_\bullet \ln 2 + \ln(S_\bullet KH/\delta)$, which is exactly $\beta_n = c_\beta/\sqrt{n}$ as in Eq. (9). □

Remark 4 (The constant)**.** The sharpened constant 2 in $n\beta_n^2$ comes from the sign vector representation. A subset representation writes $\|\widehat{P} - \bar{P}\|_1 = 2\max_B(\widehat{P}(B) - \bar{P}(B))$, converting an ℓ_1 deviation of size β into a subset deviation of only $\beta/2$. For a fixed subset B, the centered indicator increment has conditional range length at most 1. Applying conditional Hoeffding with the same stopped global time construction as Lemma 9 therefore gives exponent $-2n(\beta/2)^2 = -n\beta^2/2$. The sign vector representation gives the same exponent using conditional range length at most 2 and deviation threshold β. Thus the subset representation does not inherently incur a weaker exponent; both representations yield the same constant when their conditional ranges are used correctly.

Remark 5 (Good event)**.** Let $\mathcal{G} := \{\forall \xi, k \text{ with } n_k(\xi) \ge 1 : \|\widehat{P}_k(\cdot \mid \xi) - \bar{P}_{n_k(\xi)}(\cdot \mid \xi)\|_1 \le \beta_{n_k(\xi)}\}$. Lemma 9 holds for all positive counts simultaneously, so freezing which count is read changes nothing and $\mathbb{P}(\mathcal{G}) \ge 1-\delta$. For $n_k(\xi) \ge 1$, Lemma 8 gives $\|\widehat{P}_k(\cdot \mid \xi) - P_0(\cdot \mid \xi)\|_1 \le \beta_{n_k(\xi)} + \varepsilon a_{\max}$, and for $n_k(\xi) = 0$ the same inequality holds directly because the ℓ_1 distance between any two probability distributions is at most $2 = \beta_0$. Hence on $\mathcal{G}$,

$$\big\|\widehat{P}_k(\cdot \mid \xi) - P_0(\cdot \mid \xi)\big\|_1 \le \beta_{n_k(\xi)} + \varepsilon a_{\max} \qquad \text{for all } \xi, k. \tag{21}$$

D.3 Optimism and the regret recursion

Lemma 10 (Optimism)**.** *On $\mathcal{G}$, $\overline{V}_{k,h} \ge V^\star_{0,h}$ pointwise for all k and h.*

Proof. Backward induction on h; the base case is trivial. For a candidate action a at state (x, ξ), evaluate continuations at $(f_h(x, a, \xi'), \xi')$. If the truncation binds then $\overline{Q}_{k,h} = H - h + 1 \ge Q^\star_{0,h}$ by the value bound of Section 3. Otherwise the rewards cancel, and adding and subtracting $\sum_{\xi'} \widehat{P}_k(\xi' \mid \xi) V^\star_{0,h+1}(f_h(x, a, \xi'), \xi')$ splits $\overline{Q}_{k,h} - Q^\star_{0,h}$ into three parts: a term $\sum_{\xi'} \widehat{P}_k(\xi' \mid \xi)\big(\overline{V}_{k,h+1} - V^\star_{0,h+1}\big)(f_h(x, a, \xi'), \xi')$ that is nonnegative by the induction hypothesis and the nonnegativity of $\widehat{P}_k$; a term at least $-\frac{1}{2}(H-h)(\beta_{n_k(\xi)} + \varepsilon a_{\max})$ by Lemma 6 with Eq. (21) and the span bound $\operatorname{span}\big(V^\star_{0,h+1}(f_h(x, a, \cdot), \cdot)\big) \le H - h$; and $+b_{k,h}(\xi)$, which cancels the worst-case by Eq. (10). Taking maxima over a preserves the inequality. □

Lemma 11 (Regret recursion)**.** *On $\mathcal{G}$, along the realized M_ε trajectory of π_k,*

$$\overline{V}_{k,1}(s_1) - V^{\pi_k}_{\varepsilon,1}(s_1) \le \sum_{h=1}^{H} \Big[2b_{k,h}(\xi_h) + \tfrac{1}{2}(H-h)\,\varepsilon\,|a_h| + m_{k,h}\Big], \qquad |m_{k,h}| \le 2(H-h),$$

where $(m_{k,h})$ is a martingale difference sequence with respect to the within episode filtration.

Proof. Set $Y_{k,h} := \overline{V}_{k,h}(x_h,\xi_h) - V^{\pi_k}_{\varepsilon,h}(x_h,\xi_h)$ and $Y_{k,H+1} := 0$. Greediness and discarding the truncation give $\overline{V}_{k,h}(x_h,\xi_h) \le r_h + \sum_{\xi'} \widehat{P}_k(\xi' \mid \xi_h)\overline{V}_{k,h+1}(f_h(x_h,a_h,\xi'),\xi') + b_{k,h}(\xi_h)$. Subtracting the M_ε Bellman equation and adding and subtracting $\sum_{\xi'} P_\varepsilon(\xi' \mid \xi_h,x_h,a_h)\overline{V}_{k,h+1}(f_h(x_h,a_h,\xi'),\xi')$ yields

$$Y_{k,h} \le \underbrace{\sum_{\xi'} (\widehat{P}_k - P_\varepsilon)(\xi')\,\overline{V}_{k,h+1}(f_h(x_h,a_h,\xi'),\xi')}_{\text{(I)}} + \underbrace{\sum_{\xi'} P_\varepsilon(\xi')\left(\overline{V}_{k,h+1} - V^{\pi_k}_{\varepsilon,h+1}\right)(f_h(x_h,a_h,\xi'),\xi')}_{\text{(II)}} + b_{k,h}(\xi_h).$$

For (I), the truncation gives $\operatorname{span}\big(\overline{V}_{k,h+1}(f_h(x_h,a_h,\cdot),\cdot)\big) \le H-h$, and by Eq. (21), Assumption A2 and the triangle inequality $\|\widehat{P}_k - P_\varepsilon\|_1 \le \beta_{n_k(\xi_h)} + \varepsilon a_{\max} + \varepsilon|a_h|$, so Lemma 6 gives (I) $\le b_{k,h}(\xi_h) + \frac{1}{2}(H-h)\varepsilon|a_h|$. The on trajectory error is therefore automatically weighted by the action actually taken. For (II), let $\mathcal{F}_{k,h}$ be the history through the current state (x_h,ξ_h) and chosen action a_h, before the successor draw. The conditional law of (x_{h+1},ξ_{h+1}) is $\delta_{f_h(x_h,a_h,\xi')}(dx')\,P_\varepsilon(d\xi' \mid \xi_h,x_h,a_h)$, and both value functions are $\mathcal{F}_{k,h}$ measurable because counts are frozen, so (II) $= \mathbb{E}[Y_{k,h+1} \mid \mathcal{F}_{k,h}]$. Put $m_{k,h} := \mathbb{E}[Y_{k,h+1} \mid \mathcal{F}_{k,h}] - Y_{k,h+1}$. Both value functions lie in $[0,H-h]$, so $Y_{k,h+1} \in [-(H-h),H-h]$ and $|m_{k,h}| \le 2(H-h)$. Unrolling gives the claim. □

D.4 Frozen pigeonhole and proof of Theorem 1

Lemma 12 (Frozen pigeonhole)**.** $\sum_{k=1}^{K}\sum_{h=1}^{H-1}\big(\max\{n_k(\xi_{k,h}),1\}\big)^{-1/2} \le 2\sqrt{S_\bullet KH} + S_\bullet H.$

Proof. Fix ξ with $N(\xi)$ total transition-generating visits and index them chronologically by i, writing $k(i)$ for the containing episode. At most H of the first i visits lie in episode $k(i)$, so $n_{k(i)}(\xi) \ge i-H$ and hence $\max(n_{k(i)}(\xi),1) \ge \max(i-H,1)$. The contribution of ξ is therefore at most $H + \sum_{j=1}^{\max\{N(\xi)-H,0\}} j^{-1/2} \le H + 2\sqrt{N(\xi)}$ by integral comparison. Summing over states and applying Cauchy–Schwarz with $\sum_\xi N(\xi) \le K(H-1) \le KH$ gives the claim. □

Proof of Theorem 1. Work on $\mathcal{G}$. Per episode, by Lemma 7 and then Lemma 10,

$$V^\star_{\varepsilon,1} - V^{\pi_k}_{\varepsilon,1} \le \left(V^\star_{0,1} + \tfrac{\varepsilon H\bar{X}}{2}\right) - V^{\pi_k}_{\varepsilon,1} \le \overline{V}_{k,1} - V^{\pi_k}_{\varepsilon,1} + \tfrac{\varepsilon H\bar{X}}{2},$$

so that Δ is paid once per episode as recorded in Remark 3. Summing over k and inserting Lemma 11 gives three pieces.

Piece 1. Since $S_\bullet \ge 2$, the definition of c_β in Eq. (9) gives $c_\beta > 2 = \beta_0$, hence $\beta_n \le c_\beta/\sqrt{\max(n,1)}$ for every $n \ge 0$. Since $b_{k,H} = 0$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} 2b_{k,h}(\xi_{k,h}) = \sum_{k=1}^{K}\sum_{h=1}^{H-1} 2b_{k,h}(\xi_{k,h}).$$

Using $\sum_h (H-h) \le H^2/2$ and Lemma 12,

$$\sum_{k,h} 2b_{k,h}(\xi_{k,h}) \le c_\beta H\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) + \frac{\varepsilon a_{\max} H^2 K}{2}.$$

Piece 2. By Assumption A3, $\sum_h \frac{1}{2}(H-h)\varepsilon|a_h| \le \varepsilon H\bar{X}/2$ per episode; combined with the standing per-episode term this totals $\varepsilon\bar{X}HK$.

Piece 3. The concatenated $(m_{k,h})$ is a martingale difference sequence of length KH with increments at most $2H$, so Azuma's inequality gives $\sum_{k,h} m_{k,h} \le 2H\sqrt{2KH\ln(1/\delta)}$ with probability at least $1-\delta$.

A union bound over $\mathcal{G}$ and the Azuma event gives total failure probability at most 2δ. □

Corollary 13. *If $a_{\max} \le c\bar{X}/H$ then the two impact terms of Theorem 1 are bounded by $(1+\frac{c}{2})\varepsilon\bar{X}HK$.*

Crossover in the upper bound. The combined impact term is no larger than the sum of the statistical, frozen-count and martingale terms when $\bar{X} + a_{\max}H/2 > 0$ and

$$\varepsilon \le \frac{c_\beta\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) + 2\sqrt{2KH\ln(1/\delta)}}{K\big(\bar{X} + a_{\max}H/2\big)}.$$

If $\bar{X} + a_{\max}H/2 = 0$, the impact term is zero. For a positive denominator and $K \ge S_\bullet H$, the crossover scale, up to logarithmic factors, is $S_\bullet\sqrt{H}/[(\bar{X}+a_{\max}H)\sqrt{K}]$. Under $a_{\max} = O(\bar{X}/H)$, this becomes $S_\bullet\sqrt{H}/(\bar{X}\sqrt{K})$. This is a crossover between terms in the upper bound, not a sharp threshold for loss of the exogenous statistical advantage.

D.5 Pure exploitation under bounded action influence

At the start of episode k, let $\widehat{M}_k$ be the exogenous MDP with kernel $\widehat{P}_k$, formed from exactly the same pooled transition observations and with the same zero-count convention as Algorithm 1. Define

$$\begin{aligned}
&\widehat{V}_{k,H+1}(x,\xi) = 0,\\
&\widetilde{Q}_{k,h}(x,\xi,a) = r_h(x,a,\xi) + \sum_{\xi'} \widehat{P}_k(\xi' \mid \xi)\widehat{V}_{k,h+1}(f_h(x,a,\xi'),\xi'),\\
&\widehat{V}_{k,h}(x,\xi) = \max_{a\in\mathcal{A}_h(x,\xi)} \widetilde{Q}_{k,h}(x,\xi,a).
\end{aligned} \tag{PTO.1}$$

PTO selects a measurable maximizing policy $\widehat{\pi}_k$ and plays it for the whole episode in M_ε. The estimates and policy are frozen within the episode. No bonus or truncation is added to Eq. (PTO.1).

The policy may be a deterministic measurable selector, or a randomized Markov policy supported on the maximizing actions. Its selection may depend on all preceding interaction and on internal randomness available at the episode boundary.

Theorem PTO (Regret of no-bonus PTO)**.** *Under Assumptions A2 and A3, the PTO algorithm Eq.* (PTO.1) *satisfies, with probability at least $1-2\delta$,*

$$\begin{aligned}
\mathrm{Regret}_{M_\varepsilon}(K) \le{}& c_\beta H\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) + \frac{\varepsilon a_{\max}H^2K}{2}\\
&+ 2\varepsilon\bar{X}HK + 2H\sqrt{2KH\ln(1/\delta)},
\end{aligned} \tag{PTO.2}$$

where c_β is exactly the constant in Eq. (9)*. No coverage, mixing, ergodicity, or minimum visitation assumption is required.*

Proof. For a policy π, write $\widehat{V}^\pi_{k,h}$ for its value in $\widehat{M}_k$. Because $\widehat{P}_k$ is a probability kernel and rewards lie in $[0,1]$,

$$0 \le \widehat{V}^\pi_{k,h}(x,\xi) \le H-h+1,$$
$$0 \le \widehat{V}_{k,h}(x,\xi) \le H-h+1.$$

Backward induction in Eq. (PTO.1) gives

$$\widehat{V}^{\widehat{\pi}_k}_{k,h} = \widehat{V}_{k,h} = \sup_\pi \widehat{V}^\pi_{k,h}.$$

For a randomized maximizing policy, this equality holds because averaging over maximizing actions preserves the maximizing value at every backward step.

Step 1: a bounded error envelope. Use the event $\mathcal{G}$ of Remark 5, under the actual interaction law of PTO in M_ε. The stopped global time proof of Lemma 9 applies to this adaptive policy without modification. In particular, $\mathbb{P}(\mathcal{G}) \geq 1 - \delta$, and on $\mathcal{G}$,

$$\|\widehat{P}_k(\cdot \mid \xi) - P_0(\cdot \mid \xi)\|_1 \leq \beta_{n_k(\xi)} + \varepsilon a_{\max} \quad \text{for every } k, \xi.$$

This includes $n_k(\xi) = 0$, since the distance between two probability distributions is at most $2 = \beta_0$.

For analysis only, define the error-envelope functions $E_{k,h} : \Xi \to \mathbb{R}$ by $E_{k,H+1} = 0$ and, for $h = H, \ldots, 1$,

$$E_{k,h}(\xi) := \min \left\{ H - h,\, b_{k,h}(\xi) + \sum_{\xi'} P_0(\xi' \mid \xi) E_{k,h+1}(\xi') \right\}. \tag{PTO.3}$$

The quantity $b_{k,h}$ here is the existing expression Eq. (10); it is used only to bound model error. Neither $b_{k,h}$ nor $E_{k,h}$ is used by PTO to select actions. The envelope may depend on the unknown P_0 because it is solely a proof device.

For every outcome of the interaction, including outcomes outside $\mathcal{G}$,

$$0 \leq E_{k,h}(\xi) \leq H - h, \qquad E_{k,H} = 0. \tag{PTO.4}$$

It is measurable at the start of episode k: its only random inputs are the frozen counts.

Step 2: uniform control over policies. On $\mathcal{G}$, we claim that

$$\begin{gathered} \left| V^\pi_{0,h}(x, \xi) - \widehat{V}^\pi_{k,h}(x, \xi) \right| \leq E_{k,h}(\xi) \\ \text{for every policy } \pi,\ x, \xi, h. \end{gathered} \tag{PTO.5}$$

This is a pointwise consequence of the row bounds, so it requires no union bound over policies.

At $h = H$, the two values equal the expected immediate reward under the same policy, so their difference is zero. For the induction step, fix (x, ξ) and an action a. Adding and subtracting $\sum_{\xi'} P_0(\xi' \mid \xi) \widehat{V}^\pi_{k,h+1}(f_h(x, a, \xi'), \xi')$ gives

$$\begin{aligned} &\sum_{\xi'} P_0(\xi' \mid \xi) V^\pi_{0,h+1}(f_h(x, a, \xi'), \xi') \\ &\quad - \sum_{\xi'} \widehat{P}_k(\xi' \mid \xi) \widehat{V}^\pi_{k,h+1}(f_h(x, a, \xi'), \xi') \\ &\quad = \sum_{\xi'} (P_0 - \widehat{P}_k)(\xi' \mid \xi) \widehat{V}^\pi_{k,h+1}(f_h(x, a, \xi'), \xi') \\ &\qquad + \sum_{\xi'} P_0(\xi' \mid \xi) \big(V^\pi_{0,h+1} - \widehat{V}^\pi_{k,h+1} \big)(f_h(x, a, \xi'), \xi'). \end{aligned}$$

The absolute value of the first sum is at most $b_{k,h}(\xi)$ by Lemma 6, Eq. (21), and $0 \leq \widehat{V}^\pi_{k,h+1} \leq H - h$. The absolute value of the second sum is at most $\sum_{\xi'} P_0(\xi' \mid \xi) E_{k,h+1}(\xi')$ by the induction hypothesis, which holds at every endogenous state, including $f_h(x, a, \xi')$ for each candidate successor. The first bound likewise uses the global value interval, so the composed continuation has span at most $H - h$.

The immediate rewards cancel when the two policy Bellman equations are subtracted. Averaging the preceding bound over $\pi_h(da \mid x, \xi)$ therefore gives

$$\begin{aligned} &|V^\pi_{0,h}(x, \xi) - \widehat{V}^\pi_{k,h}(x, \xi)| \\ &\qquad \leq b_{k,h}(\xi) + \sum_{\xi'} P_0(\xi' \mid \xi) E_{k,h+1}(\xi'). \end{aligned}$$

There is also the bound $H - h$: after cancellation of the identical expected immediate reward, each remaining expected return lies in $[0, H - h]$. Taking the minimum of these two bounds proves Eq. (PTO.5).

Taking suprema over the same admissible policies yields

$$V^{\star}_{0,h}(x,\xi) \le \widehat{V}_{k,h}(x,\xi) + E_{k,h}(\xi).$$

Applying Eq. (PTO.5) to $\widehat{\pi}_k$ gives

$$V^{\widehat{\pi}_k}_{0,h}(x,\xi) \ge \widehat{V}_{k,h}(x,\xi) - E_{k,h}(\xi).$$

Consequently,

$$V^{\star}_{0,1}(s_1) - V^{\widehat{\pi}_k}_{0,1}(s_1) \le 2E_{k,1}(\xi_1). \tag{PTO.6}$$

No existence assumption for a maximizing reference policy is needed; the argument uses suprema.

Lemma 7 also implies

$$V^{\star}_{\varepsilon,1}(s_1) \le V^{\star}_{0,1}(s_1) + \frac{\varepsilon H \bar{X}}{2},$$
$$V^{\widehat{\pi}_k}_{\varepsilon,1}(s_1) \ge V^{\widehat{\pi}_k}_{0,1}(s_1) - \frac{\varepsilon H \bar{X}}{2}.$$

Combining these inequalities with Eq. (PTO.6),

$$V^{\star}_{\varepsilon,1}(s_1) - V^{\widehat{\pi}_k}_{\varepsilon,1}(s_1) \le 2E_{k,1}(\xi_1) + \varepsilon H \bar{X}. \tag{PTO.7}$$

Both simulation terms have been included explicitly. This is a fixed-data value comparison, not an invocation of a learning guarantee whose data were generated by M_0. The remaining task is to control $E_{k,1}$ using the actual M_ε observations.

Step 3: transport along the actual exogenous trace. Fix an actual transition in episode k, and abbreviate its generating row by

$$p^{\varepsilon}_{k,h}(\cdot) := P_\varepsilon(\cdot \mid \xi_{k,h}, x_{k,h}, a_{k,h}).$$

By Assumption A2, Lemma 6, and Eq. (PTO.4),

$$\begin{aligned} &\left| \sum_{\xi'} \big(P_0(\xi' \mid \xi_{k,h}) - p^{\varepsilon}_{k,h}(\xi')\big) E_{k,h+1}(\xi') \right| \\ &\qquad \le \frac{\varepsilon |a_{k,h}|}{2} \operatorname{span}(E_{k,h+1}) \le \frac{\varepsilon |a_{k,h}|}{2}(H-h). \end{aligned} \tag{PTO.8}$$

For $h \le H-1$, discarding the minimum in Eq. (PTO.3), then applying Eq. (PTO.8), gives

$$\begin{aligned} 2E_{k,h}(\xi_{k,h}) \le\ & 2b_{k,h}(\xi_{k,h}) + \varepsilon(H-h)|a_{k,h}| \\ & + 2\sum_{\xi'} p^{\varepsilon}_{k,h}(\xi') E_{k,h+1}(\xi'). \end{aligned} \tag{PTO.9}$$

This inequality holds for each realized action, including actions drawn from a randomized maximizing policy. There is no change to a comparator's trajectory law.

The bounded envelope in Eq. (PTO.3) is important here: its continuation span is bounded by the remaining horizon, independently of the confidence widths and counts. Thus the drift in Eq. (PTO.8) introduces neither an additional statistical factor nor a coverage coefficient.

Step 4: a martingale under the actual interaction law. Let $\mathcal{H}_{k,h}$ contain the full interaction history through the current state and chosen action, including internal randomness already used, but not the successor draw. In particular, $E_{k,h+1}$ is $\mathcal{H}_{k,h}$ measurable and

$$\mathbb{P}(\xi_{k,h+1} \in \cdot \mid \mathcal{H}_{k,h}) = p^{\varepsilon}_{k,h}(\cdot).$$

Define, for $h = 1, \dots, H-1$,

$$m^{\mathrm{PTO}}_{k,h} := 2\left[\sum_{\xi'} p^{\varepsilon}_{k,h}(\xi') E_{k,h+1}(\xi') - E_{k,h+1}(\xi_{k,h+1}) \right]. \tag{PTO.10}$$

These increments are defined on the entire probability space, not only on $\mathcal{G}$. They satisfy

$$\mathbb{E}[m_{k,h}^{\mathrm{PTO}} \mid \mathcal{H}_{k,h}] = 0,$$
$$|m_{k,h}^{\mathrm{PTO}}| \le 2(H-h) \le 2H.$$

Each increment is measurable after its successor is observed. The sigma-field after the preceding observed transition is contained in $\mathcal{H}_{k,h}$, including across episode boundaries. The tower rule therefore makes the chronologically concatenated increments a martingale difference sequence with respect to the successive post-transition sigma-fields.

Substituting Eq. (PTO.10) into Eq. (PTO.9) and telescoping to $E_{k,H} = 0$ gives

$$2E_{k,1}(\xi_1) \le \sum_{h=1}^{H-1} \left[2b_{k,h}(\xi_{k,h}) + \varepsilon(H-h)|a_{k,h}| + m_{k,h}^{\mathrm{PTO}}\right]. \tag{PTO.11}$$

Every count appearing on the right is evaluated at an actual transition-generating visit.

There are $K(H-1) \le KH$ increments. Azuma's inequality, applied under the original interaction law without conditioning on $\mathcal{G}$, gives an event $\mathcal{A}_{\mathrm{PTO}}$ of probability at least $1-\delta$ on which

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1} m_{k,h}^{\mathrm{PTO}} \le 2H\sqrt{2KH\ln(1/\delta)}. \tag{PTO.12}$$

Step 5: assembly. On $\mathcal{G} \cap \mathcal{A}_{\mathrm{PTO}}$, sum Eq. (PTO.7) and insert Eq. (PTO.11). First, $\beta_n \le c_\beta/\sqrt{\max\{n,1\}}$ for every $n \ge 0$, as in the proof of Theorem 1. Hence Lemma 12 gives

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1} 2b_{k,h}(\xi_{k,h}) \\
&\quad= \sum_{k=1}^{K}\sum_{h=1}^{H-1}(H-h)\big(\beta_{n_k(\xi_{k,h})} + \varepsilon a_{\max}\big) \\
&\quad\le c_\beta H\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) + \frac{\varepsilon a_{\max} KH(H-1)}{2} \\
&\quad\le c_\beta H\big(2\sqrt{S_\bullet KH} + S_\bullet H\big) + \frac{\varepsilon a_{\max} H^2 K}{2}.
\end{aligned}$$

Second, the action-dependent drift is bounded pathwise by

$$\begin{aligned}
&\varepsilon\sum_{k=1}^{K}\sum_{h=1}^{H-1}(H-h)|a_{k,h}| \\
&\quad\le \varepsilon H\sum_{k=1}^{K}\sum_{h=1}^{H}|a_{k,h}| \le \varepsilon H\bar{X}K.
\end{aligned}$$

With the $\varepsilon H\bar{X}$ per episode already present in Eq. (PTO.7), this gives $2\varepsilon\bar{X}HK$. Finally, Eq. (PTO.12) supplies the martingale term. The union bound $\mathbb{P}(\mathcal{G} \cap \mathcal{A}_{\mathrm{PTO}}) \ge 1-2\delta$ proves Eq. (PTO.2).

For $H = 1$, PTO maximizes the known immediate reward and regret is zero; all transition sums above are empty. The argument also covers $K = 1$ and unvisited states: $\beta_0 = 2$ supplies the row bound, and Eq. (PTO.4) holds regardless of counts. □

The counterfactual identity and its correction. Appendix F.3.1 of Liang et al. (2026) controls the two policy-evaluation errors through weighted reciprocal square roots of exogenous-state counts. Appendix F.3.2 uses a shared exogenous realization to make the counterfactual simulation-lemma increments conditionally mean zero. The conditional-mean calculation is the step that uses exact exogeneity.

To see precisely what changes here, fix a deterministic comparator π and define its replay on the original interaction probability space by

$$\widetilde{x}_{k,1} = x_1, \qquad \widetilde{a}_{k,h} = \pi_h(\widetilde{x}_{k,h}, \xi_{k,h}),$$
$$\widetilde{x}_{k,h+1} = f_h(\widetilde{x}_{k,h}, \widetilde{a}_{k,h}, \xi_{k,h+1}).$$

This replay is a measurable function of the actual exogenous trace. It need not have the law of π in either M_0 or M_ε.

For example, let

$$g_{k,h}(\xi') := \Big| V^{\pi}_{0,h+1}\big(f_h(\widetilde{x}_{k,h}, \widetilde{a}_{k,h}, \xi'), \xi'\big) - \widehat{V}^{\pi}_{k,h+1}\big(f_h(\widetilde{x}_{k,h}, \widetilde{a}_{k,h}, \xi'), \xi'\big)\Big|.$$

The function $g_{k,h}$ is known given $\mathcal{H}_{k,h}$ and lies in $[0, H-h]$: the candidate-successor map is determined before the draw, although the realized $\widetilde{x}_{k,h+1}$ is only known afterwards. Under the actual interaction law,

$$\mathbb{E}\left[\langle P_0(\cdot \mid \xi_{k,h}), g_{k,h}\rangle - g_{k,h}(\xi_{k,h+1}) \mid \mathcal{H}_{k,h}\right] = \left\langle P_0(\cdot \mid \xi_{k,h}) - p^{\varepsilon}_{k,h}, g_{k,h}\right\rangle.$$

The right side is not generally zero. Its absolute value is at most $\varepsilon|a_{k,h}|(H-h)/2$ by Assumption A2 and Lemma 6. Thus replay remains a valid measurable construction, but exact conditional centering must be replaced by centering under the actual kernel and an explicit drift.

The proof above implements this correction through $E_{k,h}$, which bounds the model error uniformly over policies. Eq. (PTO.8) controls the discrepancy, and Eq. (PTO.10) centers under the actual generating law. The pigeonhole inequality is consequently applied only to actual visits. Neither a comparator occupancy bound nor a positive lower bound on visitation probabilities is used.

Comparison with Theorem 1. The statistical term, frozen-count remainder, $\varepsilon a_{\max} H^2 K/2$ term, and displayed martingale term in Eq. (PTO.2) are identical to those in Theorem 1. The coefficient of $\varepsilon \bar{X} HK$ increases from 1 to 2. One copy in the PTO proof comes from the two fixed-policy simulation comparisons, and the other comes from transporting the error envelope along the actual trace.

Thus PTO attains

$$\widetilde{O}\big(S_\bullet \sqrt{H^3 K}\big) + \widetilde{O}\big(S_\bullet^{3/2} H^2\big) + O\big(\varepsilon a_{\max} H^2 K + \varepsilon \bar{X} HK\big),$$

with no dependence on $|\mathcal{X}|$ or $|\mathcal{A}|$. If $a_{\max} \le c\bar{X}/H$, its two impact terms are at most $(2 + c/2)\varepsilon \bar{X} HK$. Optimism is therefore unnecessary for attaining this tabular upper-bound order under Assumptions A2 and A3.

At $\varepsilon = 0$, both impact terms vanish. This is consistent with the exact-exogeneity PTO guarantee of Liang et al. (2026), whose Theorem 1 states $\widetilde{O}(H^2 S_\bullet \sqrt{K})$ for stage-dependent exogenous kernels. The present setting has a common kernel P_0 and pools transition counts across stages, yielding the displayed $\widetilde{O}(S_\bullet \sqrt{H^3 K})$ leading term.

D.6 The cap-free variant

Throughout this subsection only the pathwise budget of Assumption A3 is imposed. Planning and comparison use the cap-free feasible action sets, and $a_{\max}$ is not required as an input. For each exogenous state ξ, let

$$\bar{a}_k(\xi) := \frac{1}{n_k(\xi)} \sum_{i=1}^{n_k(\xi)} \big|a_i(\xi)\big|, \qquad \bar{a}_k(\xi) := 0 \text{ if } n_k(\xi) = 0, \tag{22}$$

where $a_i(\xi)$ is the action played at the ith transition-generating visit to ξ, indexed chronologically over episodes $1, \dots, k-1$, so that the sum runs over exactly the visits that define $\widehat{P}_k(\cdot \mid \xi)$. Run Exo-UCBVI-$\varepsilon$ with the bonus

$$b'_{k,h}(\xi) := \tfrac{1}{2}(H-h)\big(\beta_{n_k(\xi)} + \varepsilon\, \bar{a}_k(\xi)\big). \tag{23}$$

The algorithm played those volumes, so $b'_{k,h}$ is computable, and the empirical widening upper bounds the contamination bias of the retrospective transition mixture relative to P_0 at each state, the sampling noise remaining the responsibility of $\beta_{n_k(\xi)}$.

By Lemma 8 the widening $\varepsilon\bar{a}_k(\xi)$ dominates the bias at ξ exactly, so Lemmas 10 and 11 hold verbatim with b' in place of b, inequality Eq. (21) holding in the sharpened form $\|\widehat{P}_k - P_0\|_1 \le \beta_{n_k(\xi)} + \varepsilon\bar{a}_k(\xi)$, which is vacuous at $n_k(\xi) = 0$ through β_0 alone.

Lemma 14 (Weighted pigeonhole)**.** *For the transition-generating stages,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1} \bar{a}_k(\xi_{k,h}) \;\le\; \big(1+\ln(KH)\big)(K + S_\bullet H)\,\bar{X}.$$

Proof. Fix ξ and let its transition-generating visits, in chronological order $i = 1, \dots, N = N(\xi)$, carry volumes $v_i := |a_i|$. Let $k(i)$ denote the episode containing visit i, and let $m_i := n_{k(i)}(\xi)$ be the frozen count read at that visit. Then $\bar{a}_{k(i)}(\xi) = \frac{1}{m_i}\sum_{j\le m_i} v_j$ when $m_i \ge 1$, and the term is 0 when $m_i = 0$. As in Lemma 12, $\max(i-H, 0) \le m_i \le i-1$. Exchanging the order of summation,

$$\sum_{i=1}^{N} \bar{a}_{k(i)}(\xi) = \sum_{j=1}^{N} v_j \sum_{i:\, m_i \ge j} \frac{1}{m_i}.$$

Fix j and put $I_j := \{i : m_i \ge j\}$. Since $m_i \le i-1$, every $i \in I_j$ satisfies $i \ge j+1$. Split I_j into the window $j+1 \le i \le j+H$ and the tail $i \ge j+H+1$. In the window, membership in I_j gives $m_i \ge j$, so each included summand is at most $1/j$ and there are at most H such indices. In the tail, $m_i \ge i - H \ge j+1 > j$, so every tail index belongs to I_j and $1/m_i \le 1/(i-H)$. Hence

$$\sum_{i:\, m_i \ge j} \frac{1}{m_i} \le \frac{H}{j} + \sum_{i=j+H+1}^{N} \frac{1}{i-H} \le \frac{H}{j} + 1 + \ln N,$$

using the harmonic bound. The pathwise budget gives $v_j \le \bar{X}$, so $\sum_j H v_j/j \le H\bar{X}(1+\ln N)$, and with $A(\xi) := \sum_j v_j$ the remaining part is $(1+\ln N)A(\xi)$. Summing over states with $\sum_\xi A(\xi) \le K\bar{X}$ and $N(\xi) \le KH$ gives the claim. □

Proof of Theorem 2. Identical to the proof of Theorem 1 with two substitutions. Lemmas 10 and 11 hold with b', and term (I) of Lemma 11 uses $\|\widehat{P}_k - P_\varepsilon\|_1 \le \beta_{n_k(\xi_h)} + \varepsilon\bar{a}_k(\xi_h) + \varepsilon|a_h|$. In Piece 1 the terminal stage term vanishes, so

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(H-h)\,\varepsilon\,\bar{a}_k(\xi_{k,h}) = \sum_{k=1}^{K}\sum_{h=1}^{H-1}(H-h)\,\varepsilon\,\bar{a}_k(\xi_{k,h}) \le \varepsilon H \sum_{k=1}^{K}\sum_{h=1}^{H-1}\bar{a}_k(\xi_{k,h}),$$

which is bounded by Lemma 14. The β part of Piece 1, Piece 2 and Piece 3 are unchanged. □

E Proofs for the lower bounds

E.1 Generic instances

The first-order results of this subsection use a refinement of Assumption A2. Throughout this subsection, write $V_\varepsilon(\pi) := V^\pi_{\varepsilon,1}(s_1)$ and $V_0(\pi) := V^\pi_{0,1}(s_1)$, set $V^\star_\varepsilon := \sup_{\pi\in\Pi} V_\varepsilon(\pi)$ for the policy class Π fixed in the statement at hand, and use τ_j and G_{j+1} as defined in the proof of Lemma 7.

Definition 1 (A2′)**.** $P_\varepsilon = P_0 + \varepsilon D + R_\varepsilon$ with $D(\Xi \mid \cdot) = 0$, $\|D(\cdot \mid \xi, x, a)\|_1 \le |a|$, and $\rho(\varepsilon) := \sup_{\xi,x,a\neq 0} \|R_\varepsilon(\cdot \mid \xi, x, a)\|_1/|a| = o(\varepsilon)$.

Lemma 15 (Uniform first-order expansion)**.** *Let Assumptions A2, A3 and Definition 1 hold, let* Π *be any set of possibly history dependent randomized policies, and set*

$$L(\pi) := \sum_{j=1}^{H} \mathbb{E}_{M_0,\pi}\Big[\sum_{\xi'} D(\xi' \mid \xi_j, x_j, a_j)\, G_{j+1}\big(\tau_j, (f_j(x_j, a_j, \xi'), \xi')\big)\Big].$$

Then $\sup_{\pi\in\Pi}\big|V_\varepsilon(\pi) - V_0(\pi) - \varepsilon L(\pi)\big| \le \frac{H\bar{X}}{2}\rho(\varepsilon) + \frac{1}{2}\varepsilon^2\bar{X}H^2 a_{\max} = o(\varepsilon)$*, with constants independent of* π.

Proof. Step 1. By the hybrid construction of Lemma 7, $V_\varepsilon(\pi) - V_0(\pi) = \sum_{j=1}^{H} \mathbb{E}_{\mu_j^\varepsilon}[\langle P_\varepsilon - P_0,\, G_{j+1}(\tau_j, (f_j(x_j, a_j, \cdot), \cdot))\rangle]$, where μ_j^ε is the prefix law under P_ε dynamics, common to the two neighbouring hybrids, and write μ_j^0 for the corresponding prefix law under P_0. *Step 2.* Substituting Definition 1, the remainder contributes at most $\frac{1}{2}\rho(\varepsilon)\mathbb{E}|a_j|(H-j)$ per-step, since Lemma 6 applies verbatim to R_ε, whose total mass is zero. Summing and using Assumption A3 bounds the total remainder by $\frac{H\bar{X}}{2}\rho(\varepsilon) = o(\varepsilon)$ uniformly. The leading term per-step is $\varepsilon\,\mathbb{E}_{\mu_j^\varepsilon}[\Phi_j]$ with $\Phi_j(\tau_j) := \langle D(\cdot \mid \xi_j, x_j, a_j), G_{j+1}(\tau_j, (f_j(x_j, a_j, \cdot), \cdot))\rangle$, and Lemma 6 with $\|D\|_1 \le |a_j|$ gives $\|\Phi_j\|_\infty \le \frac{1}{2}a_{\max}(H-j)$. Note that Φ_j is signed, since D is a signed kernel.

Step 3. For any bounded prefix functional φ of τ_j,

$$\left|\mathbb{E}_{\mu_j^\varepsilon}\varphi - \mathbb{E}_{\mu_j^0}\varphi\right| \le \|\varphi\|_\infty \sum_{h=1}^{j-1} \mathbb{E}[\varepsilon|a_h|] \le \|\varphi\|_\infty\, \varepsilon\bar{X}. \tag{24}$$

To see this, telescope the prefix process through its own hybrids. Let μ_h denote the common prefix law of the two neighbouring hybrids at the swap occurring at step h. Swapping the kernel at step $h < j$ changes $\mathbb{E}[\varphi(\tau_j)]$ by $\mathbb{E}_{\mu_h}[\langle P_\varepsilon - P_0, \chi_{h+1}(\tau_h, a_h, (f_h(x_h, a_h, \cdot), \cdot))\rangle]$ with $\chi_{h+1}(\tau_h, a_h, s') := \mathbb{E}[\varphi(\tau_j) \mid \tau_h, a_h, s_{h+1} = s']$ under the common continuation dynamics. Since φ may be signed, χ_{h+1} ranges in an interval of length $2\|\varphi\|_\infty$, so Lemma 6 bounds the per swap change by $\|\varphi\|_\infty\varepsilon|a_h|$; summing over $h < j$ and applying Assumption A3 gives Eq. (24).

Applying Eq. (24) with $\varphi = \Phi_j$ and summing over j with $\sum_j (H-j) \le H^2/2$ bounds the total measure change error by $\frac{1}{4}\varepsilon^2\bar{X}a_{\max}H^2$, and $\sum_j \varepsilon\,\mathbb{E}_{\mu_j^0}\Phi_j = \varepsilon L(\pi)$ by definition. □

Proposition 16. *Let Assumptions A2, A3 and Definition 1 hold, let Π be a compact metric space on which V_0 and L are continuous, and let V_0 have a unique maximiser $\pi_0^\star$. Then $V_\varepsilon^\star - V_\varepsilon(\pi_0^\star) = o(\varepsilon)$.*

Proof. By Lemma 15, $V_\varepsilon = V_0 + \varepsilon L + E_\varepsilon$ with $\|E_\varepsilon\|_\infty =: \kappa(\varepsilon) = o(\varepsilon)$, so $V_\varepsilon \to V_0$ uniformly. Take $\pi_\varepsilon^\star$ with $V_\varepsilon(\pi_\varepsilon^\star) \ge V_\varepsilon^\star - \varepsilon^2$. Every subsequence of $(\pi_\varepsilon^\star)$ has a convergent sub-subsequence by compactness, whose limit maximised V_0 by uniform convergence and therefore equals $\pi_0^\star$ by uniqueness; hence $\pi_\varepsilon^\star \to \pi_0^\star$ and $L(\pi_\varepsilon^\star) \to L(\pi_0^\star)$ by continuity. Then

$$V_\varepsilon^\star - V_\varepsilon(\pi_0^\star) \le \underbrace{V_0(\pi_\varepsilon^\star) - V_0(\pi_0^\star)}_{\le 0} + \varepsilon\big(L(\pi_\varepsilon^\star) - L(\pi_0^\star)\big) + 2\kappa(\varepsilon) + \varepsilon^2 = o(\varepsilon). \qquad \square$$

Remark 6. Under the hypotheses of Proposition 16, the loss $V_\varepsilon^\star - V_\varepsilon(\pi_0^\star)$ is $o(\varepsilon)$. The uniform $o(\varepsilon)$ remainder in Lemma 15 does not supply an $O(\varepsilon)$ optimizer displacement rate or an $O(\varepsilon^2)$ loss bound. The tied execution instance below exhibits worst-case first-order loss for the planner classes of Theorems 3 and 4; such loss is not required in every near-exogenous problem.

E.2 The tie construction

Recall the family of Section 5. In M_ε, writing $\mathbb{P}(\xi_2 = B) = \frac{1}{2}\varepsilon a_1$ and $\mathbb{P}(\xi_3 = B) = \frac{1}{2}\varepsilon(a_1 + a_2) - \frac{1}{4}\varepsilon^2 a_1 a_2$, the values of full selling open-loop schedules are given by Eq. (12). On $\mathcal{S}$ we have $e_2(a) = \frac{1}{2}(1 - \|a\|_2^2)$, so the convexity of $\|a\|_2^2$ places its maximum at the vertices, the permutations of $(\frac{1}{2}, \frac{1}{2}, 0)$ where $e_2 = \frac{1}{4}$, and its minimum at the center where $e_2 = \frac{1}{3}$. Damage therefore depends on concentration only, and not on order. The suboptimality gaps are $\mathrm{gap}^{\mathrm{DOWN}}(a) = \frac{1}{2}g\varepsilon(e_2(a) - \frac{1}{4})$ and $\mathrm{gap}^{\mathrm{UP}}(a) = \frac{1}{2}g\varepsilon(\frac{1}{3} - e_2(a))$, which sum identically to $\frac{1}{24}g\varepsilon$.

E.3 Planners blind to the variant

Lemma 17 (No leakage through the process law)**.** *For any algorithm whose action selection is a possibly randomized function of the observed state, action and transition history only, and not of the reward function, the joint law of the entire interaction is identical under UP and DOWN.*

Proof. The variants share the initial state, f and P_ε, and differ only in rewards, which the action selection does not reference. Induct over the KH steps: the conditional law of the next action is a reward free map of the history and the conditional law of the next state is given by the shared kernel, so both coincide across variants. □

Remark 7 (Statistic blindness alone is vacuous)**.** The class of algorithms whose policy sequence factors through pooled transition counts admits the algorithm that ignores all data and hard codes the variant optimal schedule read off the known reward function, which is a constant function of the counts and achieves $O(\varepsilon^2 K)$ in its variant. No $\Theta(\varepsilon K)$ bound holds against that class, so the class must exclude use of the reward function, as Section 5 requires.

Lemma 18 (Closed loop value formulas)**.** *Let π be any closed loop randomized policy in the family. Denote by $a^G = (a_1^G, a_2^G, a_3^G)$ its random actions along the all G path, put $S := \mathbb{E}[\sum_h a_h^G]$ and $u := 1 - S \geq 0$, let w_1 and w_2 be the volumes sold at B prices following a transition to B at step 2 and step 3 respectively, and set $\delta_1 :=$ (total sold $\mid$ B at 2) $- \sum_h a_h^G$ and $\delta_2 :=$ (total sold $\mid$ B at 3) $- \sum_h a_h^G$.*

(i) *If π sells its full budget on every path then $V_\varepsilon^{\mathrm{DOWN}}(\pi) = 1 - \frac{1}{2} g\varepsilon\, \mathbb{E}[e_2(a^G)] + O(\varepsilon^2)$, and symmetrically for UP.*

(ii) *Without any selling assumption, $V_\varepsilon^{\mathrm{DOWN}}(\pi) + V_\varepsilon^{\mathrm{UP}}(\pi) = 2S + \varepsilon\, \mathbb{E}[\Delta] + O(\varepsilon^2)$ with $\Delta := a_1^G \delta_1 + a_2^G \delta_2$ and $\mathbb{E}[\Delta] \leq u$, the $O(\varepsilon^2)$ being uniform over policies.*

Proof. Write $V_\varepsilon^v(\pi) = \mathbb{E}[\sum_h A_h] \mp g\, \mathbb{E}[\sum_h A_h \mathbf{1}\{\xi_h = B\}]$ with A_h the realized actions. The B events have probability $O(\varepsilon)$: conditional on the step 1 action, taken at G surely, $\mathbb{P}(\xi_2 = B) = \frac{\varepsilon}{2}\mathbb{E}[a_1^G]$, and $\mathbb{P}(\xi_2 = G, \xi_3 = B) = \frac{\varepsilon}{2}\mathbb{E}[a_2^G](1 + O(\varepsilon))$. Hence

$$\mathbb{E}\Big[\sum_h A_h\Big] = S + \tfrac{\varepsilon}{2}\mathbb{E}\big[a_1^G\delta_1 + a_2^G\delta_2\big] + O(\varepsilon^2), \quad \mathbb{E}\Big[\sum_h A_h \mathbf{1}\{\xi_h = B\}\Big] = \tfrac{\varepsilon}{2}\mathbb{E}\big[a_1^G w_1 + a_2^G w_2\big] + O(\varepsilon^2).$$

For (ii), summing the variants cancels the $\mp g$ terms, so the damage functional is never evaluated, leaving $2S + \varepsilon\mathbb{E}[\Delta] + O(\varepsilon^2)$; and $\delta_i \leq 1 - \sum_h a_h^G$ pathwise gives $\Delta \leq (a_1^G + a_2^G)(1 - \sum_h a_h^G) \leq 1 - \sum_h a_h^G$, whence $\mathbb{E}[\Delta] \leq u$. For (i), full selling forces $\delta_i = 0$, $w_1 = 1 - a_1^G$, feasible over two capped steps, and $w_2 = a_3^G$, feasible since $a_3^G \leq \frac{1}{2}$, so the damage term is $\frac{\varepsilon}{2}\mathbb{E}[a_1^G(1 - a_1^G) + a_2^G a_3^G] = \frac{\varepsilon}{2}\mathbb{E}[e_2(a^G)]$, using $a_1(1 - a_1) = a_1(a_2 + a_3)$ on the full selling simplex. Transitions to B occur only after the damage relevant volumes are committed, which is why no closed loop policy evades e_2 at first-order. All statements hold conditionally on the history and therefore for randomized policies. □

Definition 2 (Instance oblivious and statistic blind)**.** An algorithm is instance oblivious and statistic blind if its randomized policy sequence is a function of pooled transition counts only, the same function in both variants.

Proof of Theorem 3. Playing the concentrated full selling schedule in DOWN and the uniform one in UP gives $V_{\mathrm{DOWN}}^\star + V_{\mathrm{UP}}^\star \geq (1 - \frac{1}{8}g\varepsilon) + (1 + \frac{1}{6}g\varepsilon) + O(\varepsilon^2) = 2 + \frac{1}{24}g\varepsilon + O(\varepsilon^2)$. By Lemma 17 the algorithm's episode k policy and interaction law are identical across variants, so with expectations under the common law, Lemma 18(ii) gives, per episode,

$$\mathbb{E}[\mathrm{gap}_k^{\mathrm{DOWN}}] + \mathbb{E}[\mathrm{gap}_k^{\mathrm{UP}}] \geq 2 + \tfrac{1}{24}g\varepsilon - \big(2S_k + \varepsilon\mathbb{E}[\Delta_k]\big) - C\varepsilon^2 = \tfrac{1}{24}g\varepsilon + 2u_k - \varepsilon\mathbb{E}[\Delta_k] - C\varepsilon^2 \geq \tfrac{1}{24}g\varepsilon - C\varepsilon^2,$$

since $\mathbb{E}[\Delta_k] \leq u_k$ and $2u_k - \varepsilon u_k \geq 0$, so that under selling only enlarges the gap sum, its zeroth order cost dominating its first-order credit. Summing over k and halving gives the claim for any finite C_g depending only on g with $C_g \geq 48C/g$. □

E.4 Consistent exogenous planners

Definition 3 (Consistent exogenous planner)**.** At each episode k, the algorithm plays a policy optimal in the exogenous model with some action-independent kernel $\widehat{Q}_k(\cdot \mid \xi)$, where either $\widehat{Q}_k$ is the pooled count maximum likelihood estimate with the uniform convention at zero counts, or $\|\widehat{Q}_k(\cdot \mid \xi) - \widehat{P}_k^{\mathrm{MLE}}(\cdot \mid \xi)\|_1 \leq \beta_{n_k(\xi)}$ for all ξ. Model ties are broken adversarially and rewards may be used freely in planning.

Lemma 19 (Model backward induction, general B row)**.** *In variant UP, suppose the planner's kernel satisfies $q := \widehat{Q}_k(B \mid G) > 0$, and write $q_B := \widehat{Q}_k(B \mid B) \in [0, 1]$ for its arbitrary B*

row. Then every model optimal closed loop policy sells the full budget on every path, and its G path schedule lies in

$$\mathcal{C}(q_B) := \begin{cases} \{(0, \frac{1}{2}, \frac{1}{2})\}, & q_B > 0, \\ \{(a_1,\ \frac{1}{2} - a_1,\ \frac{1}{2}) : a_1 \in [0, \frac{1}{2}]\}, & q_B = 0. \end{cases}$$

Consequently $e_2 = \frac{1}{4}$ when $q_B > 0$, while for $q_B = 0$, $e_2(a_1, \frac{1}{2} - a_1, \frac{1}{2}) = \frac{1}{4} + \frac{a_1}{2} - a_1^2 \in [\frac{1}{4}, \frac{5}{16}]$, the maximum $\frac{5}{16}$ being attained at $a_1 = \frac{1}{4}$ and the value $\frac{1}{4}$ at both endpoints.

Proof. Backward induction in the planner's model, with prices $p(G) = 1$ and $p(B) = 1 + g$, kernel $G \to B$ with probability q and $B \to B$ with probability q_B at every step, action-independent.

Step 3. Sell $\min(x, \frac{1}{2})$, which is strictly optimal since prices are positive.

Step 2 at $\xi = B$ with inventory y. The objective is $(1 + g)s + (1 + gq_B)\min(y - s, \frac{1}{2})$, because the model's expected step 3 price from B is $q_B(1 + g) + (1 - q_B) = 1 + gq_B$. Since $1 + g \geq 1 + gq_B$, selling now weakly dominates deferring, strictly if $q_B < 1$; at $q_B = 1$ the split is indifferent but every optimal split sells the full sellable amount, and the functionals of Lemma 18 depend only on the B volumes, which full selling pins. For $y \in [\frac{1}{2}, 1]$, $W_2(y, B) = \frac{1+g}{2} + (1 + gq_B)(y - \frac{1}{2})$ and $\partial_y W_2(y, B) = 1 + gq_B$.

Step 2 at $\xi = G$ with inventory y. The objective is $s + (1 + gq)\min(y - s, \frac{1}{2})$, the expected step 3 price from G being $1 + gq$, independent of the B row. For $y \in (\frac{1}{2}, 1]$ the objective increases on $s \leq y - \frac{1}{2}$ with slope 1 and strictly decreases on $s \geq y - \frac{1}{2}$ with slope $-gq < 0$, using $q > 0$, so $s^\star = y - \frac{1}{2}$ uniquely, $W_2(y, G) = (y - \frac{1}{2}) + \frac{1+gq}{2}$ and $\partial_y W_2(y, G) = 1$. The marginal unit is sold at step 2 at price 1 because the step 3 cap is already saturated at the optimum.

Step 1 at $\xi = G$ with $x = 1$. Writing $F(a_1) = a_1 + q\, W_2(1 - a_1, B) + (1 - q)\, W_2(1 - a_1, G)$ for $a_1 \in [0, \frac{1}{2}]$,

$$F'(a_1) = 1 - q(1 + gq_B) - (1 - q) = -\,g\,q\,q_B.$$

If $q_B > 0$ then F strictly decreases, $a_1^\star = 0$, step 2 at G forces $s^\star = \frac{1}{2}$ and step 3 sells $\frac{1}{2}$, giving $(0, \frac{1}{2}, \frac{1}{2})$. If $q_B = 0$ then F is constant, every $a_1 \in [0, \frac{1}{2}]$ is model optimal, after which step 2 at G uniquely forces $s^\star = \frac{1}{2} - a_1$ and step 3 sells $\frac{1}{2}$, giving $\mathcal{C}(0)$. Finally $e_2(a_1, \frac{1}{2} - a_1, \frac{1}{2}) = a_1(\frac{1}{2} - a_1) + \frac{a_1}{2} + \frac{1}{2}(\frac{1}{2} - a_1) = \frac{1}{4} + \frac{a_1}{2} - a_1^2$, maximised at $a_1 = \frac{1}{4}$ with value $\frac{5}{16}$ and equal to $\frac{1}{4}$ at both endpoints. Full selling on every path follows from positive model prices at every stage. □

Remark 8 (Adjudication of the step 1 coefficient)**.** Three derivations of the step 1 coefficient are possible: the absorbing B special case $-gq$, the proposal $-g[qq_B + (1-q)q]$, and the present $-gqq_B$. The last is correct, on three independent checks. First, at the step 2 optimum from $y \in (\frac{1}{2}, 1]$ the step 3 allocation equals the cap $\frac{1}{2}$ exactly, so the marginal unit cannot be deferred and is sold at step 2 at price 1, giving $\partial_y W_2(y, G) = 1$ and hence $F'(a_1) = -gqq_B$. Second, at $(q, q_B) = (\frac{1}{2}, 0)$ one computes $W_2(y, B) = y + \frac{g}{2}$ and $W_2(y, G) = y + \frac{g}{4}$, so $F(a_1) = 1 + \frac{3g}{8}$ for every a_1, exactly constant as $-gqq_B = 0$ predicts, whereas the alternative predicts slope $-\frac{g}{4}$. Third, at $q_B = 1$ the present formula gives $-gq$, matching the absorbing B case, whereas the alternative gives $-gq(2-q)$. The alternative arises from assigning $\partial_y W_2(y, G) = 1 + gq$, which is correct in DOWN, where the cap stays slack, and incorrect in UP, where back loading saturates it.

Lemma 20 (Pinning and persistence)**.** *In variant UP, write $q_k := \widehat{Q}_k(B \mid G)$. Then:*

(a) *For the point estimate planner, the first observed $G \to B$ transition makes $q_k > 0$ permanently, and the pinning time satisfies $\mathbb{E}[T] \leq \frac{4}{\varepsilon} + O(1)$ with a geometric tail. Moreover the point estimate B row always satisfies $q_B \geq \frac{1}{2}$: the uniform convention at $n_k(B) = 0$ gives $q_B = \frac{1}{2}$, and once any B row transition is observed the absorbing truth makes every observation $B \to B$, so the estimate gives $q_B = 1$.*

(b) *For the confidence set planner, let $k_0 := \lceil 32 \ln(K/\delta) \rceil$ and $n^\star := \left\lfloor \frac{1024 c_\beta^2}{\varepsilon^2} \right\rfloor + 1 = O\big(\varepsilon^{-2}\big[S_\bullet + \log(S_\bullet K H/\delta)\big]\big)$. With probability at least $1 - 2\delta$, for every episode $k \geq T_{\mathrm{cs}} := n^\star + k_0 + 1$ the value 0 lies outside every admissible confidence set for $\widehat{Q}_k(B \mid G)$, so $q_k > 0$ for all such k simultaneously.*

(c) *The B row need never be pinned. A B row observation requires $\xi_2 = B$, which has probability $\frac{1}{2}\varepsilon a_1$ per episode, so a planner whose realized a_1 is 0 generates no B row data and $q_B = 0$ remains admissible for all K.*

Proof. (a) Counts are cumulative. Full selling holds for every model optimal planner, since positive model prices make stranding inventory strictly suboptimal, and every full selling schedule has $a_1 + a_2 \geq \frac{1}{2}$ because $a_3 \leq \frac{1}{2}$. Steps 1 and 2 occur at G with probability $1 - O(\varepsilon)$, and each step at G with action a yields a $G \to B$ observation with probability $\frac{1}{2}\varepsilon a$, so the per episode discovery probability is at least $\frac{\varepsilon}{2} \cdot \frac{1}{2} \cdot (1 - O(\varepsilon)) \geq \frac{\varepsilon}{4}(1 - O(\varepsilon))$ and geometric domination applies. The B row statements are immediate from the conventions and the absorbing truth.

(b) The planner selects $q_k = 0$ only while $\widehat{P}_k^{\mathrm{MLE}}(B \mid G) \leq \beta_{n_k(G)}$. On the good event $\mathcal{G}$, $\widehat{P}^{\mathrm{MLE}}(B \mid G) \geq \bar{P}(B \mid G) - \beta_n$ with $\bar{P}(B \mid G) = \frac{1}{2}\varepsilon\bar{a}(G)$ exactly. Each episode contributes two G visits with volumes summing to at least $\frac{1}{2}$ if $\xi_2 = G$, or one G visit otherwise, so with $\#B_k$ the number of B episodes, $\bar{a}(G) \geq \frac{1}{4}(1 - \#B_k/k)$. Each episode is a B episode with conditional probability at most $\frac{1}{2}\varepsilon a_1 \leq \frac{\varepsilon}{4}$, and Azuma applied to the compensated count gives, with probability at least $1 - \delta$ simultaneously for all $k \geq k_0$, $\#B_k/k \leq \frac{\varepsilon}{4} + \sqrt{2\ln(K/\delta)/k} \leq \frac{1}{2}$, since for $k \geq k_0$ the root term is at most $\frac{1}{4}$ and $\varepsilon \leq 1$. Hence $\bar{a}(G) \geq \frac{1}{8}$ and $\bar{P}(B \mid G) \geq \frac{\varepsilon}{16}$ from episode k_0 on. Zero exits every admissible set once $\frac{\varepsilon}{16} - \beta_n > \beta_n$, that is once $\beta_n < \frac{\varepsilon}{32}$, that is once $n \geq n^\star$. Since $\xi_1 = G$ every episode, $n_k(G) \geq k - 1$, so both conditions hold from episode T_{cs} on, and the simultaneity over counts in Lemma 9 makes the exclusion permanent. A union bound over the two failure events gives 2δ.

(c) Immediate from the observation structure. □

Proof of Theorem 4. After pinning, each episode's realized policy is model optimal for some admissible kernel, so by Lemma 19 it sells fully with G path schedule in $\mathcal{C}(q_B)$, and by Lemma 18(i) its UP gap is $\frac{1}{2}g\varepsilon(\frac{1}{3} - e_2) + O(\varepsilon^2)$ against $V_{\mathrm{UP}}^\star \geq 1 + \frac{1}{6}g\varepsilon - O(\varepsilon^2)$.

For (i), Lemma 20(a) gives $q_B \geq \frac{1}{2} > 0$ always, so $e_2 = \frac{1}{4}$ and the per-episode gap is $\frac{1}{24}g\varepsilon(1 - O(\varepsilon))$; summing over the pinned episodes and taking expectations gives the stated bound with $\mathbb{E}[(K - T)^+]$.

For (ii), the adversarial choice $q_B = 0$ with tie break $a_1 = \frac{1}{4}$ attains $e_2 = \frac{5}{16}$, and no admissible choice exceeds it, since the bound $e_2 \leq \frac{5}{16}$ holds pointwise over $\mathcal{C}(q_B)$ for every admissible q_B by Lemma 19, hence also in expectation under any adversarial randomization. The per-episode gap is therefore at least $\frac{1}{2}g\varepsilon(\frac{1}{3} - \frac{5}{16})(1 - O(\varepsilon)) = \frac{1}{96}g\varepsilon(1 - O(\varepsilon))$ on the pinning event of Lemma 20(b), which has probability at least $1 - 2\delta$. □

E.5 The hedge budget

The confidence set hedge lowers the per episode constant only temporarily, because its use is limited by the number of B row observations available before the hedging kernel is excluded. The following proposition makes the resulting aggregate correction explicit.

Proposition 21. *Let N_B denote the number of transitions with source state B after which the hedging kernel of Theorem 4*(ii) *ceases to be admissible. For the confidence set planner of Definition 3,*

$$N_B = \left\lfloor \tfrac{1}{4} c_\beta^2 \right\rfloor + 1 \;=\; O\big(S_\bullet + \log(S_\bullet K H/\delta)\big),$$

with c_β as in Eq. (9)*. Let $\varepsilon_0(g) > 0$ be chosen so that $C_g\varepsilon_0(g) \leq 1$. Assume $0 < \delta < \frac{1}{2}$ and $0 < \varepsilon \leq \varepsilon_0(g)$. Then the confidence set planner satisfies, unconditionally,*

$$\mathbb{E}\big[\mathrm{Regret}_{\mathrm{UP}}(K)\big] \;\geq\; \frac{g\varepsilon}{24}\,(K - T_{\mathrm{cs}})^+\,(1 - C_g\varepsilon)(1 - 2\delta) \;-\; \frac{g}{2}N_B.$$

The leading post burn-in first-order coefficient is therefore $\frac{1}{24}g\varepsilon$ for both planner types, up to the displayed confidence factor and the additive hedge budget correction, the latter being logarithmic in K rather than a constant fraction of the regret. The weaker constant of Theorem 4*(ii)* remains the

correct uniform per episode guarantee, since the adversary may alternate between hedging and not, but alternation cannot exceed the aggregate budget.

Proof of Proposition 21. The factor $(1 - 2\delta)$ multiplies only the positive post pinning term; the hedge budget correction is unconditional. We first prove

$$\mathbb{E}\Big[\sum_{k=1}^{K} v_k\Big] \le \frac{g}{2} N_B, \tag{25}$$

where v_k is the leading order amount by which a genuine live hedge reduces the baseline post pinning gap.

Filtration. Let $\mathcal{F}_{k-1}$ be the history through episode $k-1$. The planner may use fresh internal randomness at the start of episode k, so define the pre draw sigma field $\widetilde{\mathcal{F}}_k := \sigma(\mathcal{F}_{k-1}, \text{episode } k \text{ internal randomness}, \widehat{Q}_k, q_k, q_{B,k}, a_1^{(k)})$, where $q_k := \widehat{Q}_k(B \mid G)$ and $q_{B,k} := \widehat{Q}_k(B \mid B)$, so that the selected model and the step 1 tie break are fixed before ξ_2 is drawn. With $Z_k := \mathbf{1}\{\xi_2 = B \text{ in episode } k\}$ and $p_k := \frac{1}{2}\varepsilon a_1^{(k)}$, the true kernel gives

$$\mathbb{E}[Z_k \mid \widetilde{\mathcal{F}}_k] = p_k. \tag{26}$$

Exclusion threshold. At zero B row observations the maximum likelihood convention is uniform, and the distance from the row corresponding to $q_B = 0$ is $\|(1,0) - (\frac{1}{2}, \frac{1}{2})\|_1 = 1 \le \beta_0 = 2$. After at least one B row observation the absorbing truth gives $\widehat{P}_n^{\mathrm{MLE}}(\cdot \mid B) = (0,1)$ for $n \ge 1$, so $q_B = 0$ is admissible if and only if $2 \le \beta_n$, that is $n \le c_\beta^2/4$. The confidence ball being closed, equality is included, so $q_B = 0$ is excluded permanently once the cumulative count reaches $N_B = \lfloor c_\beta^2/4 \rfloor + 1$.

Live hedge indicator. The condition $a_1^{(k)} > 0$ alone does not identify a hedge before the G row is pinned, so define

$$J_k := \mathbf{1}\big\{k > T_{\mathrm{cs}},\ q_k > 0,\ q_{B,k} = 0,\ 0 < a_1^{(k)} < \tfrac{1}{2}\big\},$$

which is $\widetilde{\mathcal{F}}_k$ measurable. When $J_k = 1$, Lemma 19 gives the all G schedule $(a_1^{(k)}, \frac{1}{2} - a_1^{(k)}, \frac{1}{2})$ and $e_{2,k} = \frac{1}{4} + \frac{a_1^{(k)}}{2} - (a_1^{(k)})^2$. Define

$$v_k := J_k \tfrac{1}{2} g\varepsilon\big(e_{2,k} - \tfrac{1}{4}\big) = J_k \tfrac{1}{2} g\varepsilon\, a_1^{(k)}\big(\tfrac{1}{2} - a_1^{(k)}\big) = J_k\, g\, p_k\big(\tfrac{1}{2} - a_1^{(k)}\big) \le \tfrac{g}{2} J_k p_k.$$

Budget. Whenever $J_k Z_k = 1$, one new B row observation is generated while the selected model still uses the admissible choice $q_{B,k} = 0$, and there can be at most N_B such observations before permanent exclusion. Hence pathwise $\sum_{k=1}^{K} J_k Z_k \le N_B$. By Eq. (26) and the $\widetilde{\mathcal{F}}_k$ measurability of J_k, $\mathbb{E}[J_k Z_k \mid \widetilde{\mathcal{F}}_k] = J_k p_k$, so $\mathbb{E}[\sum_k J_k p_k] = \mathbb{E}[\sum_k J_k Z_k] \le N_B$ and Eq. (25) follows. This argument is unconditional and permits fully adaptive randomized model selection and tie-breaking, provided those choices are fixed before ξ_2 is drawn.

Assembly. On the pinning event $\mathcal{E}$, every episode $k > T_{\mathrm{cs}}$ has $q_k > 0$, and Lemma 19 leaves two possibilities. If $q_{B,k} > 0$ the schedule is $(0, \frac{1}{2}, \frac{1}{2})$ with $e_{2,k} = \frac{1}{4}$. If $q_{B,k} = 0$ the schedule lies in $\mathcal{C}(0)$, whose two endpoints again give $e_{2,k} = \frac{1}{4}$, while every interior excess is charged exactly by v_k. Hence on $\mathcal{E}$ and for every $k > T_{\mathrm{cs}}$, $e_{2,k} - \frac{1}{4} = J_k a_1^{(k)}(\frac{1}{2} - a_1^{(k)})$.

By Lemma 18(i) there are finite constants C_1, C_2, independent of the post pinning policy, with $V_{\mathrm{UP}}^{\star} \ge 1 + \frac{1}{6} g\varepsilon - C_1\varepsilon^2$ and $V_{\mathrm{UP}}^{\pi_k} \le 1 + \frac{1}{2} g\varepsilon\, e_{2,k} + C_2\varepsilon^2$, so with $C := C_1 + C_2$, $\mathrm{gap}_k \ge \frac{1}{2} g\varepsilon(\frac{1}{3} - e_{2,k}) - C\varepsilon^2 \ge \frac{1}{24} g\varepsilon - v_k - C\varepsilon^2$. Using the same C_g, enlarged if necessary so that $C_g \ge 24C/g$, the assumption $C_g\varepsilon \le 1$ makes $\frac{1}{24} g\varepsilon(1 - C_g\varepsilon)$ nonnegative.

Per-episode gaps are nonnegative, so $\mathrm{Regret}_{\mathrm{UP}}(K) \ge \sum_{k=T_{\mathrm{cs}}+1}^{K} \mathrm{gap}_k$, and pathwise

$$\mathrm{Regret}_{\mathrm{UP}}(K) \ge \mathbf{1}_{\mathcal{E}}\Big[\tfrac{1}{24} g\varepsilon (K - T_{\mathrm{cs}})^{+}(1 - C_g\varepsilon) - \sum_{k=1}^{K} v_k\Big].$$

Taking expectations, using $v_k \ge 0$ with Eq. (25) and $\mathbb{P}(\mathcal{E}) \ge 1 - 2\delta$, and noting that the nonnegative coefficient permits substituting the probability lower bound, gives the claim. □

Remark 9. The constant $\frac{1}{96}g\varepsilon$ of Theorem 4(ii) is the correct uniform per episode floor, while the aggregate rate is $\frac{1}{24}g\varepsilon$ for both planner types. The hedged phase contributes an expected correction of $O(g\log(S_\bullet KH/\delta))$, an additive term logarithmic in K, and under any fixed hedge a_1 it self terminates after an expected $O(\varepsilon^{-1}\log(S_\bullet KH/\delta))$ hedged episodes, which is shorter than the confidence set pinning burn-in itself.

E.6 Completeness of the separation

Proposition 22. *The algorithm that reads the variant from the known reward function and plays the leading-order variant optimal schedule achieves $\mathbb{E}[\mathrm{Regret}_v(K)] = O(g\varepsilon^2 K)$ in both variants, by the uniform remainder bounds of Lemma 15. Eq.* (11) *is exactly linear in ε, so $R_\varepsilon \equiv 0$ in Definition 1, and that lemma supplies an $O(\varepsilon^2)$ remainder. Theorems 3 and 4 give the first-order lower bounds for their specified planner classes, with the burn-in qualifications of Theorem 4. Theorem 1 separately bounds the regret of Exo-UCBVI-ε, with impact contribution $O(\varepsilon a_{\max}H^2K + \varepsilon\bar{X}HK)$ in addition to its non-impact terms.*

Remark 10 (Why the bound lives on UP)**.** Under DOWN, pinning drives the planner toward front loaded concentration, which is DOWN optimal, so the blind planner is accidentally correct. This holds for arbitrary admissible B rows. Redoing the backward induction with $p(B) = 1-g$, the planner defers at B, since the step 2 objective $(1-g)s + (1-gq_B)\min(y-s, \frac{1}{2})$ decreases on the capped branch because $1 - gq_B \geq 1-g$, giving $s^\star = y - \frac{1}{2}$ and $\partial_y W_2^{\mathrm{DOWN}}(y, B) = 1-g$; and it sells the cap immediately at G, since $s + (1-gq)\min(y-s, \frac{1}{2})$ increases on both branches, giving $s^\star = \min(y, \frac{1}{2})$ and $\partial_y W_2^{\mathrm{DOWN}}(y, G) = 1 - gq$, the step 3 cap staying slack so that the marginal unit genuinely defers. The step 1 derivative is $1 - q(1-g) - (1-q)(1-gq) = gq(2-q) > 0$ for every $q \in (0,1]$ and every $q_B \in [0,1]$, so front loading is forced, the G path schedule is $(\frac{1}{2}, \frac{1}{2}, 0)$ and $e_2 = \frac{1}{4}$, which is DOWN optimal.

E.7 Bernstein admissible sets

Definition 4 (Bernstein set planner)**.** As Definition 3, except that the admissible set is entrywise Bernstein: for $n \geq 1$, $Q(\cdot \mid \xi)$ is admissible if and only if for every ξ', $|Q(\xi' \mid \xi) - \hat{p}| \leq r_n(\xi' \mid \xi)$ with $r_n := \sqrt{4\hat{p}(1-\hat{p})\Lambda/n} + 5\Lambda/n$, where $\hat{p} = \widehat{P}_n^{\mathrm{MLE}}(\xi' \mid \xi)$, $J := 1 + \lceil \log_2(KH) \rceil$ and $\Lambda := \ln(4S_\bullet^2 KHJ/\delta)$. When $n = 0$, every probability row is admissible. The factor J pays for the dyadic variance peeling below, and since $\ln J = O(\ln\ln(KH))$ it changes only logarithmic factors.

Proposition 23. *In variant UP there is an absolute constant C, for which $C = 2^{14}$ suffices, such that with probability at least $1 - 2\delta$:*

(i) *the Bernstein set planner has $q_k > 0$ for every $k \geq T_B := \lceil C\Lambda/\varepsilon \rceil + k_0 + 1$, so its pinning burn-in is $\widetilde{O}(\varepsilon^{-1})$, matching the point estimate rate; the ε^{-2} of Lemma 20(b) is an artifact of ℓ_1 ball radii rather than of the problem.*

(ii) *its B row hedge budget satisfies $N_B \leq \lfloor 5\Lambda \rfloor + 1 = O(\Lambda)$, so Proposition 21 holds for it with the improved additive constant $\frac{g}{2}(\lfloor 5\Lambda \rfloor + 1)$.*

Proof. (ii) The truth is B absorbing, so after $m \geq 1$ B row observations $\hat{p}(B \mid B) = 1$ exactly, the variance factor $\hat{p}(1-\hat{p})$ vanishes and $r_m = 5\Lambda/m$. Admissibility of $q_B = 0$ requires $|0 - 1| \leq r_m$, hence $m \leq 5\Lambda$. Counts are cumulative, so exclusion is permanent and the budget argument of Proposition 21 applies verbatim with $N_B = \lfloor 5\Lambda \rfloor + 1$.

(i) Consider the entry $(B \mid G)$. At its ith observed visit let $p_i := P_\varepsilon(B \mid G, x_i, a_i) \in [0, \frac{\varepsilon}{4}]$, $\bar{p}_n := \frac{1}{n}\sum_{i \leq n} p_i$ and $\hat{p}_n := \frac{1}{n}\sum_{i \leq n} \mathbf{1}_i$ with $\mathbf{1}_i$ indicating a $G \to B$ transition. Then $M_n := \sum_{i \leq n}(\mathbf{1}_i - p_i)$ is a martingale and $V_n := \sum_{i \leq n} p_i(1-p_i) \leq n\bar{p}_n$.

For each entry, count, sign and dyadic level $\ell \in \{0, \ldots, J-1\}$, consider the event $2^\ell \leq n\bar{p}_n < 2^{\ell+1}$, on which $V_n \leq 2^{\ell+1}$. Freedman's inequality gives $\mathbb{P}(\pm M_n \geq \sqrt{2 \cdot 2^{\ell+1}\Lambda} + \frac{2}{3}\Lambda,\ 2^\ell \leq n\bar{p}_n < 2^{\ell+1}) \leq e^{-\Lambda}$, and since $2^{\ell+1} \leq 2n\bar{p}_n$ on the same event the threshold is at most $\sqrt{4n\bar{p}_n\Lambda} + \frac{2}{3}\Lambda$. There are at most two signs, $S_\bullet^2$ entries, KH counts and J levels, and by the definition of Λ we have $2S_\bullet^2 KHJ\, e^{-\Lambda} \leq \delta$. Hence with probability at least $1 - \delta$, simultaneously for all entries and

all counts with $n\bar{p}_n \geq 1$,

$$|\hat{p}_n - \bar{p}_n| \leq \sqrt{\frac{4\bar{p}_n\Lambda}{n}} + \frac{2\Lambda}{3n}. \tag{27}$$

Intersect this with the Azuma scaffold of Lemma 20(b), which has probability at least $1-\delta$ and gives $\bar{p}_n \geq \frac{\varepsilon}{16}$ from episode k_0 on. Take $n \geq C\Lambda/\varepsilon$ with $C = 2^{14}$. Then $n\bar{p}_n \geq C\Lambda/16 \geq 1$, so Eq. (27) applies, and $\Lambda/n \leq \varepsilon/C \leq \frac{16}{C}\bar{p}_n = 2^{-10}\bar{p}_n$. Therefore $\sqrt{4\bar{p}_n\Lambda/n} \leq 2^{-4}\bar{p}_n$ and $\frac{2\Lambda}{3n} \leq \frac{2}{3}2^{-10}\bar{p}_n$, and since $2^{-4} + \frac{2}{3}2^{-10} < \frac{1}{10}$ we obtain $\frac{9}{10}\bar{p}_n \leq \hat{p}_n \leq \frac{11}{10}\bar{p}_n$. The Bernstein radius then satisfies $r_n \leq \big(\sqrt{\frac{44}{10}2^{-10}} + 5 \cdot 2^{-10}\big)\bar{p}_n < \frac{1}{5}\bar{p}_n$, so $\hat{p}_n - r_n \geq (\frac{9}{10} - \frac{1}{5})\bar{p}_n > 0$ and the value 0 lies outside the admissible interval, giving $q_k > 0$. Since $n_k(G) \geq k-1$ this holds for every $k \geq T_B$, and cumulative counts make the exclusion permanent. The two events fail with total probability at most 2δ. □

E.8 An algorithm-specific two-stage lower bound

Proposition (First-order loss of the tabular algorithms). For every $0 < \varepsilon \leq 1$, there is a finite instance satisfying Assumptions A2 and A3, with $H = 2$, $\bar{X} = a_{\max} = 1$ and $\mathcal{X} = \Xi = \mathcal{A} = \{0,1\}$, such that, for every integer $K \geq 1$, every $\delta \in (0,1)$ and every permitted tie-breaking rule, Algorithm 1 satisfies almost surely

$$\mathrm{Regret}_{M_\varepsilon}(K) \geq \frac{\varepsilon}{16}\left(K - \left\lceil \frac{c_\beta^2}{4} \right\rceil\right)_+, \qquad c_\beta^2 = 2\ln\frac{16K}{\delta}.$$

The subscript $+$ denotes the positive part. The same bound holds for the cap-free variant of Appendix D.6. No-bonus PTO of Appendix D.5 has regret exactly $\varepsilon K/16$. All three statements use the optimal admissible policy in M_ε as comparator, as in Eq. (5). The instance depends on ε, but not on K, δ or the tie-breaking rule.

Proof. Take $s_1 = (1,0)$, let the remaining budget be $\mathsf{b}(x) = x$, and set, for $h = 1, 2$,

$$\mathcal{A}_h(x,\xi) = \{a \in \{0,1\} : a \leq x\}, \qquad f_h(x,a,\xi') = \max\{x-a, 0\}.$$

For feasible actions the update is $x - a$, so every path satisfies $\sum_{h=1}^{2}|a_h| \leq 1$ and $|a_h| \leq 1$. Action zero provides a feasible continuation from every state. Removing the separate cap does not change these feasible sets. Define the time-homogeneous exogenous kernels by

$$P_0(1 \mid 0) = 0, \qquad P_\varepsilon(1 \mid 0, x, a) = \frac{\varepsilon a}{2},$$
$$P_0(1 \mid 1) = P_\varepsilon(1 \mid 1, x, a) = 1.$$

Complementary probabilities define the remaining entries. These are probability kernels for $0 < \varepsilon \leq 1$. On the row $\xi = 0$, their ℓ_1 distance is $2(\varepsilon a/2) = \varepsilon|a|$; on the row $\xi = 1$ it is zero. Thus Assumption A2 holds, including equality of the actual and reference rows at action zero. Give both models the same known rewards

$$r_1(x,a,\xi) = \frac{\varepsilon}{16}\mathbf{1}\{a = 0\}, \qquad r_2(x,a,\xi) = \frac{1}{4}\mathbf{1}\{\xi = 1\}.$$

These rewards are deterministic and lie in $[0,1]$; there is no terminal reward. The second-stage action does not affect any remaining reward. At s_1, choosing $a_1 = 0$ therefore has value $\varepsilon/16$, whereas choosing $a_1 = 1$ has value $(\varepsilon/2)(1/4) = \varepsilon/8$ in M_ε. An admissible optimal policy chooses $a_1 = 1$ and then $a_2 = 0$. For a possibly randomized episode policy π_k, write $\pi_{k,1}(\{0\} \mid 1, 0)$ for its probability of selecting action zero at the initial state. Then

$$V^\star_{\varepsilon,1}(s_1) = \frac{\varepsilon}{8},$$
$$V^\star_{\varepsilon,1}(s_1) - V^{\pi_k}_{\varepsilon,1}(s_1) = \frac{\varepsilon}{16}\pi_{k,1}(\{0\} \mid 1, 0).$$

In Algorithm 1, the final-stage bonus is zero and the upper truncation at 1 is inactive, so

$$\overline{V}_{k,2}(x,\xi) = \frac{1}{4}\mathbf{1}\{\xi = 1\}.$$

This value is independent of x. Composing it with $f_1(x, a, \xi')$ therefore gives the same continuation for both initial actions. Before upper truncation at 2, their backups are

$$\frac{\varepsilon}{16}\mathbf{1}\{a = 0\} + \frac{1}{4}\widehat{P}_k(1 \mid 0) + b_{k,1}(0).$$

Only $h = 1$ generates an observed transition when $H = 2$, and every episode starts at $\xi_1 = 0$. Hence $n_k(0) = k - 1$ for every preceding interaction history. With $S_\bullet = H = 2$, Eq. (9) gives $c_\beta^2 = 2\ln(16K/\delta)$. We use its untruncated radius: $\beta_0 = 2$ and $\beta_n = c_\beta/\sqrt{n}$ for $n \geq 1$, which may exceed 2. For $k \geq 1 + \lceil c_\beta^2/4 \rceil$, however, $\beta_{n_k(0)} \leq 2$, and Eq. (10) gives $b_{k,1}(0) \leq 1 + \varepsilon/2$. Since $0 \leq \widehat{P}_k(1 \mid 0) \leq 1$, both untruncated backups are at most

$$\frac{\varepsilon}{16} + \frac{1}{4} + 1 + \frac{\varepsilon}{2} \leq \frac{29}{16} < 2.$$

Upper truncation is therefore inactive, and action zero is the unique maximizer, with margin $\varepsilon/16$, for every empirical row and history. Each of these episodes incurs regret $\varepsilon/16$, regardless of earlier tie-breaking, including randomized choices among maximizers. There are $(K - \lceil c_\beta^2/4 \rceil)_+$ such episodes, and all earlier episode regrets are nonnegative. Summing Eq. (5) proves the asserted almost-sure bound without a concentration event.

For the cap-free variant, $0 \leq \bar{a}_k(0) \leq 1$ on every history, including the zero-count convention. Its bonus Eq. (23) is no larger than the bonus just bounded and remains common to both actions. Its final-stage value is unchanged, so the same count, clipping and strict-maximizer argument proves the same lower bound.

For PTO, Eq. (PTO.1) has neither a bonus nor truncation. Its final-stage value is again $\widehat{V}_{k,2}(x, \xi) = \frac{1}{4}\mathbf{1}\{\xi = 1\}$, and its initial backup difference is exactly $\varepsilon/16$ in favor of action zero in every episode. Thus it incurs exactly $\varepsilon/16$ per episode and $\varepsilon K/16$ in total. The same conclusion holds for any policy optimal under any action-independent planning kernel: changing that kernel changes only the common continuation term. □

The same two-stage instance also gives a lower bound for Algorithm 2 under a one-dimensional representation of its reference kernel.

Proposition (First-order loss of Exo-VTR). For every $0 < \varepsilon \leq 1$, equip the preceding instance with counting measure ν on $\Xi = \{0, 1\}$ and the representation

$$d = 1, \qquad \phi(\xi', \xi) = P_0(\xi' \mid \xi) = \mathbf{1}\{\xi' = \xi\},$$
$$\theta^\star = 1, \qquad B_\theta = 1.$$

This instance satisfies Assumptions A1L, A2 and A3, with $H = 2$ and $\bar{X} = a_{\max} = 1$. For every integer $K \geq 1$, every $\delta \in (0, 1)$ and every permitted tie-breaking rule, Algorithm 2 and its cap-free variant in Corollary 30 each satisfy, almost surely,

$$\mathrm{Regret}_{M_\varepsilon}(K) = \frac{\varepsilon K}{16}.$$

The comparator is the optimal admissible policy in M_ε, as in Eq. (5). The instance, including its known reward and feature map, is fixed before K, δ and the tie-breaking rule are chosen. The reward depends on ε, whereas the feature map does not.

Proof. For every bounded function g on Ξ, the integrated feature is $\psi_g(\xi) = g(\xi)$. Thus $\|\psi_g(\xi)\|_2 \leq 1$ whenever g takes values in $[0, 1]$, and the representation identity holds with $\|\theta^\star\|_2 = B_\theta = 1$. All spaces are finite, and the integration and feasible-action maximization are finite operations. This verifies Assumption A1L and the oracle requirements. The rewards, maps, feasible sets and kernels are unchanged from the preceding construction. In particular, the row distance at $\xi = 0$ is $2(\varepsilon a/2) = \varepsilon|a|$ and is zero at $\xi = 1$, including at action zero. Every successor branch satisfies $\sum_{h=1}^{2} |a_h| \leq 1$, and removing the separate cap leaves the feasible sets unchanged.

Fix an episode k and any preceding history with the prescribed initial state and transition-recording convention. Since $\overline{V}_{k,3} = 0$, the stage-2 feature is zero for every feasible state and action. Both the prediction and bonus in Eq. (16) vanish there. Clipping to $[0, 1]$ therefore gives

$$\overline{Q}_{k,2}(x, \xi, a) = \overline{V}_{k,2}(x, \xi) = \frac{1}{4}\mathbf{1}\{\xi = 1\}.$$

Consequently, composing the value with the known successor map gives

$$\psi_{k,1}(x,\xi,a) = \sum_{\xi' \in \{0,1\}} \phi(\xi',\xi)\overline{V}_{k,2}(f_1(x,a,\xi'),\xi') = \frac{1}{4}\mathbf{1}\{\xi = 1\}.$$

In particular, both feasible actions at $(x,\xi) = (1,0)$ have feature zero. Every recorded transition has $h_t = 1$ and $\xi_t = 0$, so the stored pairs in Eq. (13) satisfy

$$V_t(\xi') = \frac{1}{4}\mathbf{1}\{\xi' = 1\}, \qquad \psi_t = 0, \qquad y_t = \frac{1}{4}\mathbf{1}\{\xi_{t+1} = 1\}.$$

This calculation does not assume that earlier actions were zero. Even a recorded transition to state 1 gives a nonzero target multiplied by a zero feature. Since $\lambda = H^2 = 4$, Eq. (14) therefore gives, for every such history,

$$\Sigma_k = 4I, \qquad \widehat{\theta}_k = 0, \qquad \frac{\det \Sigma_k}{\lambda^d} = 1.$$

The full radius in Eq. (15), including its contamination term, is

$$\beta_k = 2 + \sqrt{2\ln(1/\delta)} + \varepsilon\sqrt{k-1}.$$

It is finite for every stated k and δ and multiplies a zero feature norm at both initial actions. The two-sided clipped backups are thus

$$\overline{Q}_{k,1}(1,0,0) = \frac{\varepsilon}{16}, \qquad \overline{Q}_{k,1}(1,0,1) = 0.$$

The upper clip at 2 changes neither value, and the lower clip preserves the zero value of action 1. Action zero is the unique maximizer from episode 1, regardless of the history or any randomized tie-breaking at other states. Its actual value is $\varepsilon/16$, while the admissible policy choosing $a_1 = 1$ and $a_2 = 0$ attains the actual optimum $\varepsilon/8$. Every selected policy therefore incurs exactly $\varepsilon/16$ regret, including when its second-stage action is randomized. Summing Eq. (5) gives the asserted identity.

For the cap-free variant, the contamination term is $\frac{1}{2}\varepsilon H\bar{X}\sqrt{k-1} = \varepsilon\sqrt{k-1}$, the same as above. Its feasible sets are also unchanged, so every preceding calculation applies. □

For the tabular optimistic algorithms, when $K \geq 2\lceil c_\beta^2/4\rceil$, the preceding lower bound is at least $\varepsilon K/32$. These results establish algorithm-specific worst-case first-order loss, not a minimax lower bound over learning algorithms or matching dependence on the horizon or model dimensions. The one-dimensional feature map above determines P_0; normalization of its probability rows forces $\theta^\star = 1$. Assumption A1L does not exclude this case. The Exo-VTR result concerns its specified backup under action-dependent misspecification, rather than the difficulty of estimating an unknown reference kernel, and does not assert the same behavior for other feature representations of P_0. The known reward is selected separately for each ε and is identical within each reference/actual model pair, so these results do not contradict the fixed-reference conclusion of Proposition 16. An action-aware policy has zero regret on this instance.

F Proofs for the linear mixture setting

The linear mixture bound does not subsume Theorem 1. A direct pair indicator embedding of the tabular model uses $d = S_\bullet^2$, and satisfying the normalization in Assumption A1L requires rescaling ϕ by $S_\bullet^{-1/2}$, which inflates $\|\theta^\star\|_2$ to order $S_\bullet$. The resulting guarantee is weaker than Theorem 1 by at least a factor of order $S_\bullet$, so neither theorem contains the other.

F.1 Measurability and computational oracles

For the large or infinite space statement, Ξ, $\mathcal{X}$ and $\mathcal{A}$ are standard Borel spaces, and the rewards, deterministic maps, kernels and feature map are jointly measurable (in particular, f_h is measurable in (x,a,ξ')), with $\int \|\phi(\xi',\xi)\|_2\,\nu(d\xi') < \infty$ for every relevant ξ, so ψ_g is well defined for every

Algorithm 2 Exo-VTR-ε (frozen updates)

input: H, K, ε, $a_{\max}$, $\bar{X}$, B_θ, δ, $\lambda = H^2$
for $k = 1, \dots, K$ **do**
  form Σ_k and $\widehat{\theta}_k$ by Eq. (14) from the pairs Eq. (13), and β_k by Eq. (15)
  $\overline{V}_{k,H+1} \leftarrow 0$
  for $h = H, \dots, 1$ **do**
    $\psi_{k,h}(x, \xi, a) \leftarrow \psi_{\overline{V}_{k,h+1}(f_h(x,a,\cdot),\cdot)}(\xi)$
    $\overline{Q}_{k,h}(x, \xi, a) \leftarrow$ the two sided clipped backup Eq. (16)
    $\overline{V}_{k,h}(x, \xi) \leftarrow \max_{a \in \mathcal{A}_h(x,\xi)} \overline{Q}_{k,h}(x, \xi, a)$, $\pi_{k,h}(x, \xi) \in \arg\max_{a \in \mathcal{A}_h(x,\xi)} \overline{Q}_{k,h}(x, \xi, a)$
  end for
  play $\pi_k = (\pi_{k,h})_{h=1}^{H}$ for the whole episode in M_ε
end for

bounded measurable g. The learner has access to an integration oracle returning $\psi_g(\xi)$ for every bounded measurable successor-composed continuation function generated by the algorithm, and to a measurable planning oracle returning an action in $\mathcal{A}_h(x, \xi)$ that maximises Eq. (16) over that set. Both oracle operations are exact for every input generated by the algorithm. Their outputs, including tie-breaking choices, are jointly measurable in the episode-start information and the queried states and actions. At backward step h, the continuation $\overline{V}_{k,h+1}$ is already fixed. The returned selector defines $\pi_{k,h}$, and the corresponding maximum defines $\overline{V}_{k,h}$. These functions are available for exact evaluation at feasible states, and $\overline{V}_{k,h}$ is a valid continuation input for the next backward step. Starting from $\overline{V}_{k,H+1} \equiv 0$, the recursion fixes all value functions and the policy before the episode begins. Theorem 5 is therefore a statistical guarantee; computational efficiency additionally requires these oracle operations to be tractable for the chosen feature and action classes. For example, a finite inventory grid, including any budget coordinate, that is closed under f_h, with finite Ξ and enumerable finite feasible action sets, permits finite sums and backward enumeration when the known rewards, maps and features can be evaluated. The cost depends on these set sizes and evaluations, not only on d. The change of measure argument of Lemma 6 and the simulation bound of Lemma 7 pass to integrals without change, since the centering step uses only that the difference of two probability kernels has total mass zero.

Algorithm 2 specifies what is computed; Fig. 3 shows when each operation occurs. This ordering is required by the arguments that follow. Because the estimate and the policy are fixed before the episode begins and are not revised within it, ψ_t is measurable with respect to the pre-transition field $\mathcal{H}_t$ in Lemma 25, and $\overline{V}_{k,h+1}$ is measurable at episode start in Lemma 28. Because the feature stored for a step is the one evaluated in the bonus at the played action, the vectors summed in Lemma 29 are precisely those incorporated into Σ_{k+1}, which is the invariant recorded in Remark 12.

F.2 Contaminated least squares

Writing

$$\zeta_t := \big\langle P_\varepsilon(\cdot \mid \xi_t, x_t, a_t) - P_0(\cdot \mid \xi_t),\, V_t \big\rangle, \qquad \omega_t := y_t - \big\langle P_\varepsilon(\cdot \mid \xi_t, x_t, a_t),\, V_t \big\rangle, \tag{28}$$

identity Eq. (8) gives $y_t = \psi_t^\top \theta^\star + \zeta_t + \omega_t$, in which ζ_t is the contamination bias created by Eq. (4) and ω_t is conditionally mean zero noise. This is the least squares counterpart of the decomposition behind Eq. (10). These components are controlled separately in the analysis of Eq. (15).

Lemma 24 (Bias). *$|\zeta_t| \le \frac{1}{2}\varepsilon |a_t| (H - h_t)$, and consequently*

$$\sum_{t \in D_{k-1}} \zeta_t^2 \le \Big(\frac{\varepsilon H}{2}\Big)^2 \sum_{t \in D_{k-1}} a_t^2 \le \Big(\frac{\varepsilon H}{2}\Big)^2 a_{\max} \bar{X}\, (k-1).$$

Proof. The signed measure $P_\varepsilon - P_0$ has total mass zero, so Lemma 6 gives $|\zeta_t| \le \frac{1}{2}\|P_\varepsilon - P_0\|_1 \operatorname{span}(V_t) \le \frac{1}{2}\varepsilon |a_t| (H - h_t)$, the span bound holding because the two-sided clipping in Eq.

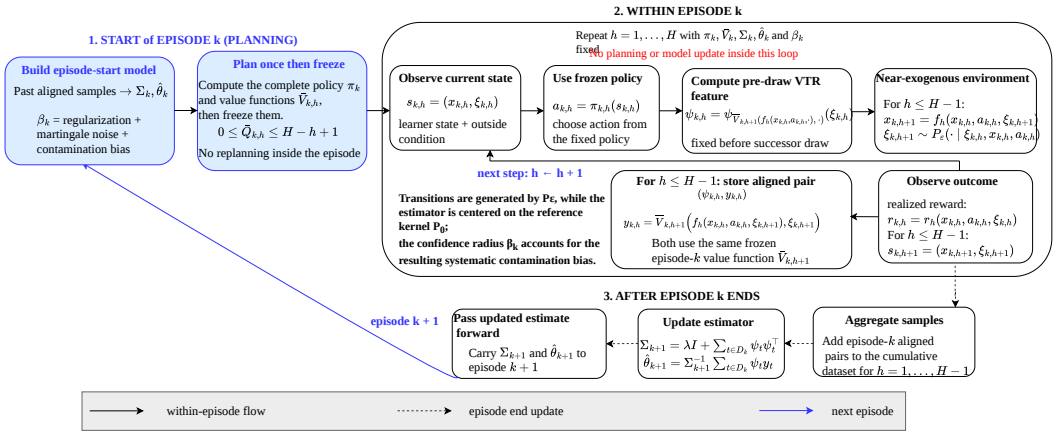


Figure 3: **One episode of Exo-VTR-ε.** Planning occurs once at the start of episode k, using the design matrix and ridge estimate constructed from previous episodes, and the resulting policy and value functions are then held fixed throughout the episode. Transitions are generated by the contaminated kernel P_ε while the estimator is centered on the reference kernel P_0, and the confidence radius β_k accounts for the resulting contamination bias. Each transition-generating step $h \le H-1$ contributes one aligned pair (ψ_t, y_t), and these pairs are incorporated into the design matrix only after the episode ends. The feature ψ_t is computed from the function $\xi' \mapsto \overline{V}_{k,h+1}(f_h(x_t, a_t, \xi'), \xi')$ before the draw; the target y_t is its value at the observed ξ_{t+1}.

(16) confines $\overline{V}_{j,h_t+1}$ to $[0, H-h_t]$ at every full state, and hence $V_t(\xi') \in [0, H-h_t]$ for every candidate successor ξ'. For the sum of squares, within one episode $\sum_h a_h^2 \le a_{\max} \sum_h |a_h| \le a_{\max} \bar{X}$ by Assumption A3, and there are at most $k-1$ completed episodes. □

Remark 11. The bias enters through squared volumes, which is the structural distinction from the tabular proof: the least squares geometry aggregates the bias directly in a quadratic norm and the episode budget controls that quadratic sum. Moreover $\sum_h a_h^2 \le (\sum_h |a_h|)^2 \le \bar{X}^2$ holds without any per-step cap, which is the source of the cap-free variant recorded in Corollary 30.

The radius Eq. (15) uses this deterministic bound on the squared action volumes; it does not adapt to their realized sum. Using the cap alone would give $H a_{\max}^2 (k-1)$ in place of $a_{\max} \bar{X}(k-1)$, so the trajectory budget can tighten the contamination widening when $\bar{X} < H a_{\max}$.

Lemma 25 (Noise)**.** *Let $\mathcal{H}_t$ be generated by the interaction and internal randomness through (ξ_t, x_t, a_t), before ξ_{t+1} is observed, and let $\mathcal{F}_t := \sigma(\mathcal{H}_t, \xi_{t+1})$. Then ψ_t is $\mathcal{H}_t$ measurable, ω_t is $\mathcal{F}_t$ measurable with $\mathbb{E}[\omega_t \mid \mathcal{H}_t] = 0$, and ω_t is conditionally sub-Gaussian with parameter $H/2$.*

Proof. The frozen specification fixes $\overline{V}_{j,h_t+1}$ at the start of episode j, and (ξ_t, x_t, a_t) belong to $\mathcal{H}_t$. Because f_{h_t} is known, the function $\xi' \mapsto V_t(\xi') = \overline{V}_{j,h_t+1}(f_{h_t}(x_t, a_t, \xi'), \xi')$ is determined before the draw, so ψ_t is predictable. The realized $x_{t+1} = f_{h_t}(x_t, a_t, \xi_{t+1})$ is $\mathcal{F}_t$ measurable. Conditionally on $\mathcal{H}_t$ the law of ξ_{t+1} is $P_\varepsilon(\cdot \mid \xi_t, x_t, a_t)$, giving $\mathbb{E}[\omega_t \mid \mathcal{H}_t] = 0$. By the two sided clipping, $V_t(\xi_{t+1}) \in [0, H-h_t]$ on every sample path, so it lies in an interval of length at most H and Hoeffding's lemma makes its centered version conditionally sub-Gaussian with parameter at most $H/2$. □

Lemma 26 (Confidence ellipsoid)**.** *Let $\mathcal{G}_{\text{lin}} := \{\forall k \ge 1 : \|\widehat{\theta}_k - \theta^\star\|_{\Sigma_k} \le \beta_k\}$ with β_k as in Eq. (15). Then $\mathbb{P}(\mathcal{G}_{\text{lin}}) \ge 1 - \delta$.*

Proof. Ridge algebra gives $\sum_t \psi_t y_t = (\Sigma_k - \lambda I)\theta^\star + \sum_t \psi_t \zeta_t + \sum_t \psi_t \omega_t$, so $\widehat{\theta}_k - \theta^\star = \Sigma_k^{-1}\big(\sum_t \psi_t \omega_t + \sum_t \psi_t \zeta_t - \lambda\theta^\star\big)$ and, by the triangle inequality in the Σ_k norm,

$$\|\widehat{\theta}_k - \theta^\star\|_{\Sigma_k} \le \Big\|\sum_t \psi_t \omega_t\Big\|_{\Sigma_k^{-1}} + \Big\|\sum_t \psi_t \zeta_t\Big\|_{\Sigma_k^{-1}} + \lambda\|\theta^\star\|_{\Sigma_k^{-1}}.$$

For the third term, $\Sigma_k \succeq \lambda I$ gives $\lambda\|\theta^\star\|_{\Sigma_k^{-1}} \le \sqrt{\lambda}\|\theta^\star\|_2 \le H B_\theta$.

For the second, take any v with $\|v\|_{\Sigma_k} \le 1$. Cauchy–Schwarz over t gives $\langle \sum_t \psi_t \zeta_t, v\rangle = \sum_t \zeta_t(\psi_t^\top v) \le \sqrt{\sum_t \zeta_t^2}\sqrt{\sum_t (\psi_t^\top v)^2}$, and $\sum_t (\psi_t^\top v)^2 = v^\top(\sum_t \psi_t\psi_t^\top)v \le v^\top \Sigma_k v \le 1$ because Σ_k dominates the sample Gram matrix. Taking the supremum over such v and applying Lemma 24 bounds this term by $\frac{1}{2}\varepsilon H\sqrt{a_{\max}\overline{X}(k-1)}$. This step is where the volume weighting of the bias occurs.

For the first, Lemma 25 places us in the setting of the self normalized concentration inequality for vector valued martingales of Abbasi-Yadkori et al. (2011): ψ_t is measurable with respect to the pre noise sigma field $\mathcal{H}_t$, ω_t is conditionally $H/2$ sub-Gaussian given $\mathcal{H}_t$, and the result is simultaneous over all times, so no union bound over k is required. It gives, with probability at least $1-\delta$ simultaneously for all k, $\|\sum_t \psi_t\omega_t\|_{\Sigma_k^{-1}} \le \frac{H}{2}\sqrt{\ln(\det\Sigma_k/\lambda^d) + 2\ln(1/\delta)}$. Summing the three bounds matches Eq. (15). □

F.3 Optimism and the regret recursion

Lemma 27 (Optimism). *On $\mathcal{G}_{\mathrm{lin}}$, $\overline{V}_{k,h} \ge V^\star_{0,h}$ pointwise for all k, h.*

Proof. Backward induction on h; the base case is trivial. For a candidate action a at state (x,ξ), abbreviate $\psi_{k,h} := \psi_{k,h}(x,\xi,a)$. Let $U_{k,h}(x,\xi,a) := r_h(x,a,\xi) + \psi_{k,h}^\top\widehat{\theta}_k + \beta_k\|\psi_{k,h}\|_{\Sigma_k^{-1}}$ be the untruncated optimistic backup. Adding and subtracting $\psi_{k,h}^\top\theta^\star = \langle P_0(\cdot\mid\xi), \overline{V}_{k,h+1}(f_h(x,a,\cdot),\cdot)\rangle$, which holds by Eq. (8), gives

$$U_{k,h} - Q^\star_{0,h} = \underbrace{\psi_{k,h}^\top(\widehat{\theta}_k - \theta^\star)}_{\ge -\beta_k\|\psi_{k,h}\|_{\Sigma_k^{-1}}} + \underbrace{\big\langle P_0(\cdot\mid\xi), (\overline{V}_{k,h+1} - V^\star_{0,h+1})(f_h(x,a,\cdot),\cdot)\big\rangle}_{\ge 0} + \beta_k\|\psi_{k,h}\|_{\Sigma_k^{-1}} \ \ge\ 0.$$

Thus $U_{k,h} \ge Q^\star_{0,h} \ge 0$, so the lower clip is inactive on $\mathcal{G}_{\mathrm{lin}}$. If the upper clip binds then $\overline{Q}_{k,h} = H-h+1 \ge Q^\star_{0,h}$; otherwise $\overline{Q}_{k,h} = U_{k,h} \ge Q^\star_{0,h}$. The first bound is Cauchy–Schwarz in the Σ_k geometry and the second uses the induction hypothesis against the genuine probability kernel P_0; nothing requires $\psi^\top\widehat{\theta}_k$ itself to be a probability expectation. Maxima over a preserve the inequality. □

Lemma 28 (Regret recursion). *On $\mathcal{G}_{\mathrm{lin}}$, along the realized M_ε trajectory of π_k, with $b_{k,h} := \beta_k\|\psi_{k,h}(x_h,\xi_h,a_h)\|_{\Sigma_k^{-1}}$ evaluated at the played action,*

$$\overline{V}_{k,1}(s_1) - V^{\pi_k}_{\varepsilon,1}(s_1) \le \sum_{h=1}^{H}\Big[2b_{k,h} + \tfrac{1}{2}(H-h)\varepsilon|a_h| + m_{k,h}\Big], \qquad |m_{k,h}| \le 2(H-h),$$

with $(m_{k,h})$ a martingale difference sequence.

Proof. Set $Y_{k,h} := \overline{V}_{k,h}(x_h,\xi_h) - V^{\pi_k}_{\varepsilon,h}(x_h,\xi_h)$ and $Y_{k,H+1} := 0$, and let $\mathcal{F}_{k,h}$ denote the history through the current state (x_h,ξ_h) and chosen action a_h, before the successor draw. Identical in structure to Lemma 11; only term (I) changes. On $\mathcal{G}_{\mathrm{lin}}$ the lower clip is inactive at every action by Lemma 27, so greediness and discarding only the upper truncation give $\overline{V}_{k,h} \le r_h + \psi_{k,h}^\top\widehat{\theta}_k + b_{k,h}$. Subtracting the M_ε Bellman equation and adding and subtracting $\langle P_\varepsilon, \overline{V}_{k,h+1}(f_h(x_h,a_h,\cdot),\cdot)\rangle$ produces

$$Y_{k,h} \le \underbrace{\psi_{k,h}^\top\widehat{\theta}_k - \big\langle P_\varepsilon, \overline{V}_{k,h+1}(f_h(x_h,a_h,\cdot),\cdot)\big\rangle}_{\text{(I)}} + \underbrace{\big\langle P_\varepsilon, (\overline{V}_{k,h+1} - V^{\pi_k}_{\varepsilon,h+1})(f_h(x_h,a_h,\cdot),\cdot)\big\rangle}_{\text{(II)}} + b_{k,h}.$$

For (I), adding and subtracting $\psi_{k,h}^\top \theta^\star$ and applying Eq. (8),

$$(\mathrm{I}) = \psi_{k,h}^\top(\widehat{\theta}_k - \theta^\star) + \big\langle P_0 - P_\varepsilon, \overline{V}_{k,h+1}(f_h(x_h, a_h, \cdot), \cdot)\big\rangle \le \beta_k \|\psi_{k,h}\|_{\Sigma_k^{-1}} + \tfrac{1}{2}(H-h)\varepsilon|a_h|,$$

the second bound by Lemma 6 with Assumption A2 at the action actually taken, so that the on trajectory model error is automatically volume weighted while only the historical bias inside β_k carries the aggregate volume. For (II), the frozen design makes $\overline{V}_{k,h+1}$ measurable at episode start and the conditional law of (x_{h+1}, ξ_{h+1}) is $\delta_{f_h(x_h,a_h,\xi')}(dx')\, P_\varepsilon(d\xi' \mid \xi_h, x_h, a_h)$, so $(\mathrm{II}) = \mathbb{E}[Y_{k,h+1} \mid \mathcal{F}_{k,h}]$; the resulting increment is bounded by $2(H-h)$ since both value functions lie in $[0, H-h]$. □

Remark 12 (Sample to bonus alignment)**.** The elliptical potential of Lemma 29 requires the feature vectors appearing in the bonus sums to be the vectors added to the design matrix. This holds exactly: the bonus at step h of episode k is evaluated at $(x_{k,h}, \xi_{k,h}, a_{k,h})$ using the function

$$\xi' \mapsto \overline{V}_{k,h+1}\big(f_h(x_{k,h}, a_{k,h}, \xi'), \xi'\big).$$

For the corresponding global step t, this is exactly V_t, so the stored feature satisfies

$$\psi_t = \psi_{k,h}(x_{k,h}, \xi_{k,h}, a_{k,h}).$$

Both features are determined before the successor draw. The regression target is $y_t = V_t(\xi_{k,h+1})$. The frozen design pairs each sample with its own episode's estimate. Had the assembly been routed through Remark 3's idealized regret comparison, the features in the bonus summation would be distributed under M_0 while the design matrix grows along M_ε trajectories, and the potential argument would fail. In an implementation this invariant should be asserted at every stored step, since it breaks silently if the regression pair is recomputed with a value function updated after the episode ends.

F.4 The frozen elliptical potential

Lemma 29. *With* $\Gamma := \ln(\det \Sigma_{K+1}/\lambda^d) \le d\ln(1 + KH/d)$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \|\psi_{k,h}\|_{\Sigma_k^{-1}} \le 2\sqrt{KH\Gamma} + \frac{H\Gamma}{\ln 2}.$$

Proof. Introduce, as an analysis device only, the per-step running matrix $\Sigma_t := \lambda I + \sum_{s<t} \psi_s \psi_s^\top$ over global time; for t inside episode k we have $\Sigma_k \preceq \Sigma_t \preceq \Sigma_{k+1}$ in the Loewner order, because the episode start matrix omits the current episode's samples and the episode end matrix includes them. Call episode k loud if $\det \Sigma_{k+1} > 2 \det \Sigma_k$ and quiet otherwise. Determinant ratios telescope to at most e^Γ, so there are at most $\Gamma/\ln 2$ loud episodes. Within a loud episode each of the at most H step terms is bounded by 1, which is valid because $\lambda = H^2$ forces $\|\psi\|_{\Sigma_k^{-1}} \le \|\psi\|_2/H \le 1$, so loud episodes contribute at most $H\Gamma/\ln 2$.

Within a quiet episode we use

$$x^\top A^{-1} x \le \big(\det B/\det A\big)\, x^\top B^{-1} x \qquad \text{for positive definite } A \preceq B. \tag{29}$$

This follows from the direct form result of Abbasi-Yadkori et al. (2011), which states that for positive semi-definite A', B', C' with $A' = B' + C'$ one has $\sup_{x \ne 0} x^\top A' x / x^\top B' x \le \det A'/\det B'$. Since $A \preceq B$ with both positive definite, $B^{-1} \preceq A^{-1}$, so $A^{-1} = B^{-1} + (A^{-1} - B^{-1})$ with the bracket positive semi-definite; applying the direct form with $A' = A^{-1}$, $B' = B^{-1}$ and $C' = A^{-1} - B^{-1}$ gives $x^\top A^{-1} x / x^\top B^{-1} x \le \det(A^{-1})/\det(B^{-1}) = \det B/\det A$, which is Eq. (29).

With $A = \Sigma_k$, $B = \Sigma_t$ and $\det \Sigma_t/\det \Sigma_k \le \det \Sigma_{k+1}/\det \Sigma_k \le 2$ in a quiet episode, Eq. (29) gives $\|\psi_t\|_{\Sigma_k^{-1}} \le \sqrt{2}\|\psi_t\|_{\Sigma_t^{-1}}$. The matrix determinant lemma gives $\det \Sigma_{t+1} = \det \Sigma_t(1 + \|\psi_t\|^2_{\Sigma_t^{-1}})$, and since each $\|\psi_t\|^2_{\Sigma_t^{-1}} \le 1$ the elementary inequality $u \le 2\ln(1+u)$ on $[0,1]$ yields $\sum_t \|\psi_t\|^2_{\Sigma_t^{-1}} \le 2\Gamma$; Cauchy–Schwarz over the at most KH steps then gives $\sum_t \|\psi_t\|_{\Sigma_t^{-1}} \le \sqrt{2KH\Gamma}$. Combining, quiet episodes contribute at most $\sqrt{2} \cdot \sqrt{2KH\Gamma} = 2\sqrt{KH\Gamma}$. □

F.5 Proof of Theorem 5 and the cap-free variant

Proof of Theorem 5. Work on the intersection of $\mathcal{G}_{\mathrm{lin}}$ and the Azuma event below. Per episode, by Lemma 7 and then Lemma 27,

$$V^{\star}_{\varepsilon,1}(s_1) - V^{\pi_k}_{\varepsilon,1}(s_1) \le \left(V^{\star}_{0,1}(s_1) + \tfrac{\varepsilon H \bar{X}}{2}\right) - V^{\pi_k}_{\varepsilon,1}(s_1) \le \overline{V}_{k,1}(s_1) - V^{\pi_k}_{\varepsilon,1}(s_1) + \tfrac{\varepsilon H \bar{X}}{2},$$

Lemma 7 being a statement about the two true MDPs with no reference to the algorithm or to tabular structure. The gap is paid once per episode, for the reason recorded in Remarks 3 and 12. Summing over k and inserting Lemma 28 gives three pieces.

Piece 1. $\sum_{k,h} 2b_{k,h} = 2\sum_{k,h} \beta_k \|\psi_{k,h}\|_{\Sigma_k^{-1}} \le 2\beta_K \sum_{k,h} \|\psi_{k,h}\|_{\Sigma_k^{-1}}$, since β_k is nondecreasing in k, the determinant and the contamination sum both growing; Lemma 29 bounds the remaining sum.

Piece 2. Per episode $\sum_h \frac{1}{2}(H-h)\varepsilon|a_h| \le \varepsilon H\bar{X}/2$ by Assumption A3; combined with the standing per-episode term the total is $\varepsilon \bar{X} HK$.

Piece 3. The concatenated $(m_{k,h})$ is a martingale difference sequence of length KH with increments at most $2H$, so Azuma gives $2H\sqrt{2KH\ln(1/\delta)}$ with probability at least $1-\delta$. A union bound over the two failure events gives 2δ.

For the rate summary, the noise part of β_K is of order $H\sqrt{\Gamma}$, and its product with the leading potential term $2\sqrt{KH\Gamma}$ is $H\Gamma\sqrt{KH} = \widetilde{O}(d\sqrt{H^3K})$; the regularization part contributes $\widetilde{O}(B_\theta\sqrt{d}\sqrt{H^3K})$; and the contamination part contributes $\widetilde{O}(\varepsilon K\sqrt{d\, a_{\max}\bar{X}H^3})$. Multiplication by the frozen update term $H\Gamma/\ln 2$ contributes $\widetilde{O}(B_\theta dH^2)$, $\widetilde{O}(d^{3/2}H^2)$ and $\widetilde{O}(\varepsilon dH^2\sqrt{a_{\max}\bar{X}K})$ respectively. The last grows as $\sqrt{K}$, so it is of lower order than the linear contamination term asymptotically but is not polylogarithmic, and is therefore displayed in Section 6. □

Corollary 30 (Cap-free impact price)**.** *Assume Assumptions A1L, A2 and only the pathwise budget of Assumption A3. Use the cap-free feasible action sets for both learner and comparator; $a_{\max}$ is not required as an input. Replacing the contamination term of Eq.* (15) *by $\frac{1}{2}\varepsilon H\bar{X}\sqrt{k-1}$, which is valid by the cap-free bound $\sum_h a_h^2 \le \bar{X}^2$ per episode in Lemma 24, Theorem 5 holds with impact terms $\widetilde{O}(\varepsilon\bar{X}\sqrt{dH^3K}) + O(\varepsilon\bar{X}HK) + \widetilde{O}(\varepsilon\bar{X}dH^2\sqrt{K})$ and no per-step cap anywhere.*

Under the proportional cap $a_{\max} \le c\bar{X}/H$ of Corollary 13, the capped radius gives the main linear impact term $\widetilde{O}(\varepsilon\bar{X}HK\sqrt{cd})$, since $\sqrt{d\, a_{\max}\bar{X}H^3} \le \bar{X}H\sqrt{cd}$, with the lower order burn-in $\widetilde{O}(\varepsilon\bar{X}dH^{3/2}\sqrt{cK})$, whereas the cap-free radius gives $\widetilde{O}(\varepsilon\bar{X}HK\sqrt{dH})$ and $\widetilde{O}(\varepsilon\bar{X}dH^2\sqrt{K})$ whether or not a cap holds. The capped contamination term is no larger whenever $a_{\max} \le \bar{X}$; the value of the cap-free form is that it requires no separate per-step cap.

Conjecture 31. *The impact widening composes additively with Bernstein-type base analyzes. The claim is plausible, since the widening only enlarges a radius, but it is unproven because variance aware radii are estimated from the same contaminated samples whose bias the widening exists to absorb. Theorem 5 settles the Hoeffding case for the linear mixture class, and Proposition 23 settles the corresponding question on the lower bound side; the upper bound composition remains open.*

Remark 13 (Relation to the exact exogeneity literature)**.** The linear structure of Liang et al. (2026) and the present one are orthogonal axes. Their post decision function approximation linearises the endogenous side, with features on the post-decision endogenous state and a weight vector for each exogenous state and stage, while their exogenous process remains finite and is estimated tabularly, so $|\Xi|$ appears multiplicatively in their guarantees. Assumption A1L does the reverse: it linearises the exogenous kernel, permits Ξ to be large or infinite under the oracle assumptions above, and handles the endogenous side exactly through the known deterministic map. Neither model class contains the other. Their impossibility result concerns the removal of known endogenous transitions or known rewards rather than an action-dependent exogenous kernel, so the two works study distinct failure modes, and whether their pure exploitation framework admits a robust ε impact analogue remains open.

G EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

G.1 CONTROLLED FAMILIES

The controlled campaign evaluates four pre-declared environment families, denoted *default sanity*, *predictive mean reversion*, *persistent impact* and *high state impact*, with $H = 20$, $S_\bullet = 5$, $K = 3000$ episodes, six matched seeds $0, \ldots, 5$, and $\varepsilon \in \{0, 0.02, 0.05, 0.10, 0.20\}$. All methods within a family and ε level share seeds, so comparisons are paired. Regret is computed exactly, by dynamic programming against the optimal value of the true environment, rather than by Monte Carlo estimation of realized reward. The theorem valid planner uses the confidence radius and bonus of Section 4 with a pre-declared lexicographic tie break among exact optimistic maximizers; the low bonus ablation scales the bonus by 0.05 and carries no theorem claim.

G.1.1 ALL CONTROLLED FAMILIES

Fig. 4 shows the persistent impact family reported in Section 7.1, while Fig. 5 shows the remaining three controlled families. Persistent impact exhibits an approximately linear increase in final regret across the tested ε values, consistent with the linear impact dependence appearing in Theorem 1.

The other three families are non-monotone. Default sanity rises sharply through $\varepsilon = 0.10$ and then changes little by $\varepsilon = 0.20$. Predictive mean reversion falls initially and then rises substantially as ε increases. High state impact reaches its largest regret at $\varepsilon = 0.10$ and falls again at $\varepsilon = 0.20$. These patterns do not contradict Theorem 1, which is an upper bound and does not require regret to increase monotonically with ε on every instance. The linear fits in Fig. 5 are therefore descriptive summaries rather than tests of the theorem's rate.

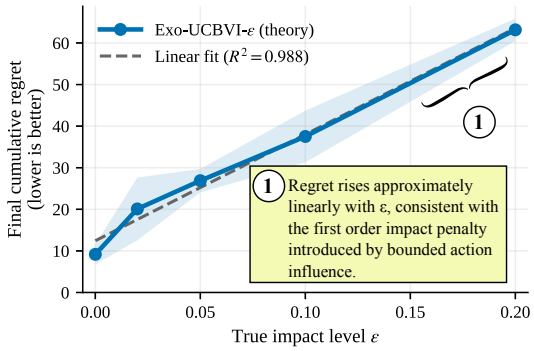


Figure 4: **Persistent impact family.** Final cumulative regret against the true action influence ε, with a linear fit. Shading shows the spread across six matched seeds.

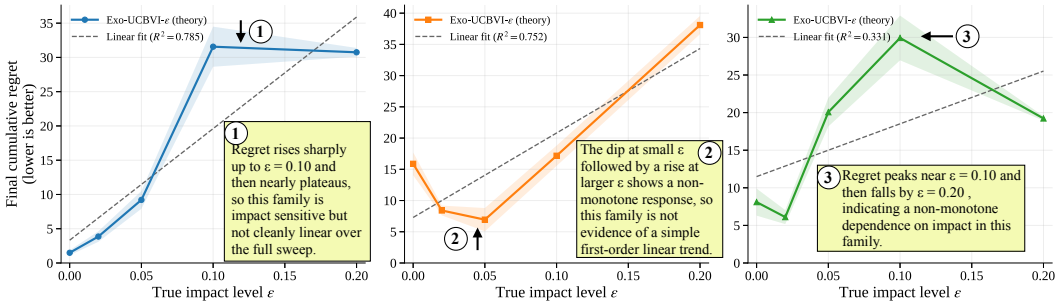


Figure 5: **The three non-monotone controlled families.** Final cumulative regret against ε for default sanity (left), predictive mean reversion (center) and high state impact (right), each with a linear fit and its R^2. Shading shows the spread across six matched seeds. Axis scales differ across panels.

G.1.2 LOW-BONUS EXO-UCBVI ABLATION

The low bonus ablation scales the theorem-valid bonus by 0.05 and is reported separately, since it carries no theorem claim below unit scale. Fig. 6 summarises its effect on the four controlled families through the ratio of final regret under the low-bonus variant to final regret under the theorem-valid method. A ratio near 1 means that the two methods finish with similar regret, whereas a smaller ratio means that reducing the bonus lowers final regret. Default sanity, predictive mean reversion and persistent impact all move toward the equal-regret line as ε increases, so on those families the low-bonus variant approaches the theorem-valid method at larger impact levels. High state impact is the exception: its ratio remains well below 1 throughout and is still close to 0.5 at $\varepsilon = 0.20$. Thus the low-bonus variant generally tracks the theorem-valid method, but the effect of reducing the bonus is clearly instance-dependent.

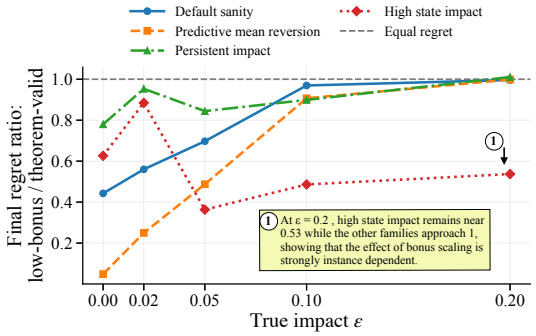


Figure 6: **Low-bonus Exo-UCBVI ablation.** Final-regret ratio (low-bonus)/(theorem-valid) against the true action influence ε for default sanity, predictive mean reversion, persistent impact and high state impact. The dashed horizontal line marks equal regret. Ratios below 1 indicate that the low-bonus variant attains lower final regret than the theorem-valid method.

G.2 HARD FAMILY

The hard family of Section 5 is run in both variants with $K = 1500$, six seeds and $\varepsilon \in \{0, 0.01, 0.02, 0.05, 0.10\}$, giving 240 runs. Four methods are compared: PTO as an exact exogeneity stress test, the pooled exogenous maximum likelihood planner, the theorem valid Exo-UCBVI-ε, and an impact aware parametric maximum likelihood reference. Predictions were registered before the campaign was run. The tail pinning fraction, the realized e_2 and the final exact regret are recorded per run.

G.3 EXO-VTR BONUS DIAGNOSTICS

The Exo-VTR impact sweep uses $H = 20$, $S_\bullet = 20$, $\bar{X} = 20$, $a_{\max} = 2$, $d = 4$, $K = 600$ and $\delta = 0.05$, with seeds $0, \ldots, 5$. The dimension sweep uses $K = 400$ with the same horizon, exogenous state space, budget, cap, confidence level and seeds. The bonus diagnostic campaign records, at fixed (ξ, h) and at each checkpoint, the mean, ninetieth percentile and maximum spread of $\beta_k \|\psi_{k,h}\|_{\Sigma_k^{-1}}$ across feasible actions; the fraction of feasible states at which the bonus changes the greedy action with neither clip active; the fraction at which it changes the argmax set; the fraction of action cells at which the upper clip binds; and the fraction of policy cells that change under a full replan with the bonus removed. The main text reports the slice with mixture dimension $d = 4$ at the final checkpoint, sweeping $\varepsilon_{\text{true}} = \varepsilon_{\text{assumed}} \in \{0, 0.02, 0.05, 0.10\}$.

Figure 7 reports two additional diagnostics. The left panel tests whether the action dependence identified in Section 6 persists as data accumulate. The mean bonus spread across feasible actions generally falls over learning, but remains strictly positive at late checkpoints for every tested ε. Moreover, larger assumed impact is associated with a larger late-stage spread, consistent with the ε-dependent contamination widening in Eq. (15). Thus the elliptical bonus does not collapse to the action-neutral geometry of the tabular bonus.

The right panel measures whether this action dependence affects decisions. At the final checkpoint, the fraction of states at which the bonus varies across actions decreases with ε, while the fraction of

upper-clipped action cells increases strongly over the same sweep; the two patterns are consistent with greater upper clipping coinciding with less observable action-level bonus variation at larger impact levels. Crucially, the fraction of states at which the greedy action changes with neither clip active remains small but nonzero throughout. The full-replan policy-change fraction is substantially larger and remains roughly stable across ε, consistent with local bonus differences propagating through the backward value recursion.

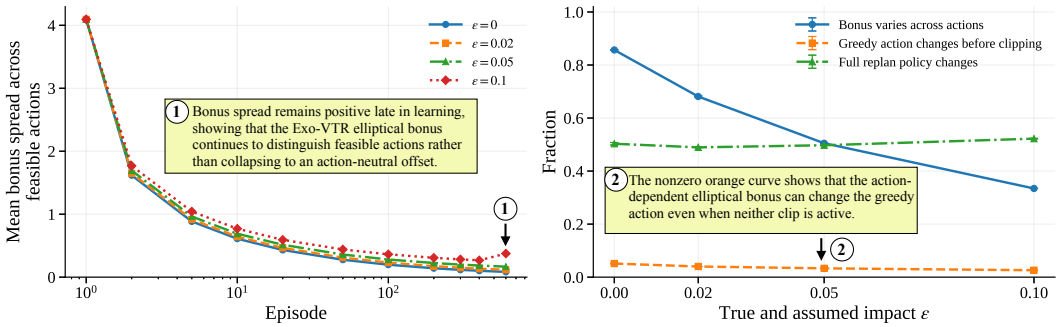


Figure 7: **Exo-VTR bonus diagnostics at $d = 4$.** Left: mean spread of the elliptical bonus across feasible actions at fixed (ξ, h), against episode on a logarithmic scale, for four values of $\varepsilon_{\text{true}} = \varepsilon_{\text{assumed}}$. Right: at the final checkpoint, the fraction of feasible states at which the bonus varies across actions, at which the greedy action changes with neither clip active, and at which a full replan with the bonus removed changes the policy cell. Right-panel error bars show 95% confidence intervals.

G.3.1 No-bonus Exo-VTR ablation

The diagnostics above establish that the bonus is action-dependent and that it alters greedy choices, but not how much of the observed regret gap is associated with the bonus. Fig. 8 isolates this contribution by comparing four variants on the same instances: the theorem-valid method, a no-bonus variant identical except that the optimistic term is removed, an impact-blind variant that assumes $\varepsilon = 0$, and an action-conditioned plug-in reference that models the action dependence directly. The plug-in reference has the lowest final regret throughout the sweep. The right panel therefore reports how much of the empirical gap from either the theorem-valid or impact-blind method to this reference is closed by removing the bonus.

Removing the bonus lowers final regret at every ε on the grid and closes approximately 45–49 per cent of the theorem-valid-to-plug-in gap across the sweep. The corresponding fraction for the impact-blind method declines as ε increases, because the no-bonus regret rises substantially while the impact-blind and plug-in regrets change much less. Neither fraction approaches 100 per cent, so removing the bonus explains only part of the observed gap on these instances. The no-bonus variant carries no theorem claim and is reported only as an empirical ablation.

Figure 9 adds the temporal view of this ablation. Across all four tested impact levels, the two variants remain close through the early and intermediate checkpoints, and the separation emerges later in training, when the cumulative regret of the theorem-valid Exo-VTR-ε variant grows more rapidly. Thus the final-regret difference reported above is associated with a late-stage divergence in these configurations. This empirical pattern does not establish that the elliptical optimism bonus is unnecessary in the linear-mixture setting in general.

G.4 Exo-VTR robustness and dimension scaling

The diagnostics of the previous subsection fix the mixture dimension at $d = 4$. This subsection varies it over $d \in \{2, 4, 8, 12\}$ and asks two questions: whether the implementation invariants the analysis relies on continue to hold, and whether the action dependence they enable survives as the feature space grows.

Figure 10 answers both. The left panel records four run-level validation checks, all of which pass at every dimension tested. These are binary checks rather than performance estimates: no run in the

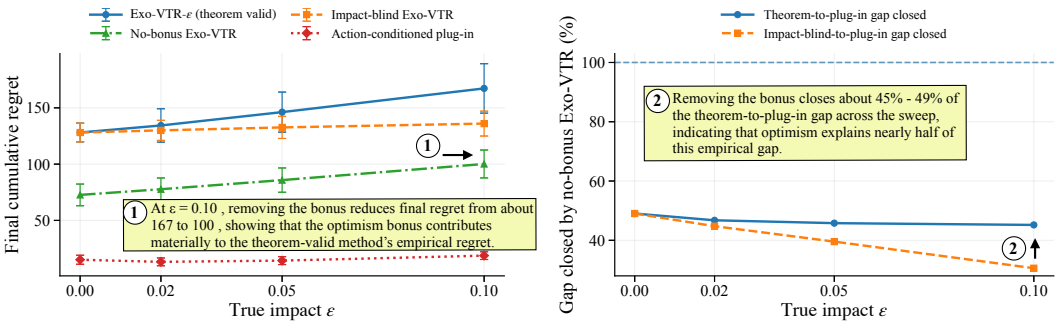


Figure 8: **No-bonus ablation at $d = 4$.** Left: final cumulative regret against the true action influence ε for the theorem-valid Exo-VTR-ε, the no-bonus variant, the impact-blind variant run with $\varepsilon = 0$, and the action-conditioned plug-in reference. Right: the percentage of the gap to the plug-in reference that the no-bonus variant closes, relative to the theorem-valid method and to the impact-blind variant. Left-panel error bars show 95% confidence intervals.

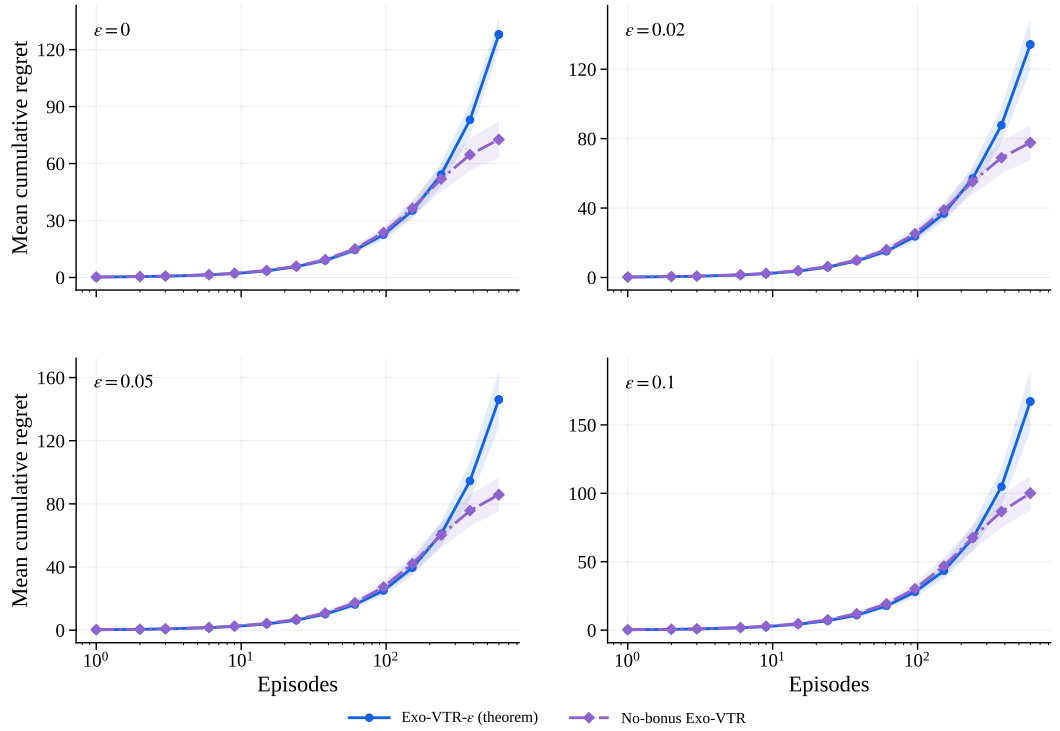


Figure 9: **Exo-VTR versus no-bonus Exo-VTR.** Cumulative regret over learning for the theorem-valid Exo-VTR-ε method and the otherwise identical no-bonus ablation across the tested impact levels. Curves show mean cumulative regret across six matched seeds, with 95% confidence intervals.

sweep violates the replay, confidence, feature-alignment, or primary-maximizer conditions. In particular, the feature-alignment check verifies the sample-to-bonus alignment described in Remark 12.

The right panel is graded. The fraction of states at which the bonus varies across actions declines with d, whereas the fraction of policy cells changed by a full replan rises slightly; the two curves cross between $d = 4$ and $d = 8$. Their divergence is consistent with local bonus differences propagating through the backward replanning recursion. The fraction of states at which the greedy action changes with neither clip active is small at every dimension but remains strictly positive throughout the tested range, so the mechanism identified in Section 6 is not specific to the $d = 4$ slice.

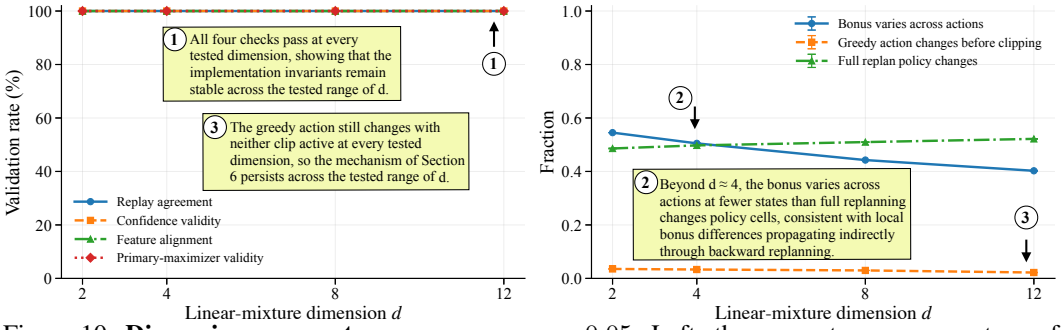


Figure 10: **Dimension sweep at $\varepsilon_{\text{true}} = \varepsilon_{\text{assumed}} = 0.05$.** Left: the pass rate, as a percentage of runs, of the replay agreement, confidence validity, feature alignment and primary-maximizer validity checks, against the linear-mixture dimension d. Right: the fraction of feasible states at which the bonus varies across actions, at which the greedy action changes with neither clip active, and the fraction of policy cells changed by a full replan with the bonus removed. Right-panel error bars show 95% confidence intervals.

G.5 Historical market replay

G.5.1 Market data and protocol

The market data are Binance spot BTCUSDT one minute klines for the twelve calendar months from August 2025 to July 2026 inclusive, merged chronologically into a single series of $525{,}600$ observations. Integrity was verified before use: the twelve monthly files were checked against an exact expected manifest by symbol, interval and month; each file's timestamps and row count were checked against its declared calendar month; the merged series was checked for duplicate minutes, for gaps in the minute grid, for internal consistency of each bar, and for agreement between the row count and the calendar span. The merged file is identified by its SHA-256 digest `5b7edb05bd01c9bd1f13ce1b895d14d647e9f1be0e7989cbe7561bcb38ed005e`, and the merge reproduces byte identically across machines.

The chronological split places the first $367{,}920$ observations in the calibration period and the remaining $157{,}680$ in the hold out period. The exogenous state is the quintile of the trailing twenty minute log price residual, with quantile boundaries fitted on the calibration period only and the trailing mean excluding the current observation, so there is no look ahead. Calibration occupancy is exactly 20% in each of the five states. The representative price levels are the centered empirical medians of $\exp$ of the residual within each state, on the scale the data exhibit. Evaluation draws 2000 contiguous blocks of length $H = 20$ from the hold out period with replacement, normalising each block by its first price, and every method is executed on the identical blocks. Four protocol checks are asserted at run time: that all methods use identical blocks, that the fixed baselines reproduce their reference values exactly, that every method completes liquidation, and that the paired TWAP cost is identically zero.

The calibration prefix fixes the exogenous state boundaries, representative state levels and simulator transition matrix. Each Exo-UCBVI learner is then initialized with zero transition counts and trained for 4000 simulated episodes under this fixed simulator. The learner's empirical kernel, confidence widths and policy are recomputed between simulated episodes; the calibrated simulator kernel itself is not refitted. After training, the final policy is frozen for the full 2000 block held-out replay. Dur-

ing held-out replay, the previously trained policy and its learned estimates remain fixed throughout evaluation. The impact sweep uses the base Exo-UCBVI implementation. The benchmark table uses the lexicographic tie breaking implementation, which selects an empirical value maximizer among the exact optimistic maximizers.

held-out execution is scored using the per stage replay payoff

$$r_h^{\mathrm{replay}} = a_h\big(p_h - \eta\, a_h\big) - \tau\, a_h, \tag{30}$$

where p_h is the observed relative price, η is a temporary impact coefficient and τ a proportional fee. During simulated training, Bellman planning uses the same economic payoff with the fixed representative calibrated price level of the current exogenous state in place of p_h. The resulting finite reward table is transformed by one common affine normalization over feasible model rewards to $[0, 1]$, so the planning reward is known and deterministic once the calibrated model is fixed. The learned policy is indexed by (h, x, ξ); the continuous observed price is not an additional policy input, although it is used to construct the discrete residual state and to score held-out execution. The calibrated kernel is action-independent, so Exo-UCBVI is run here with $\varepsilon = 0$; η affects the execution reward and not the transition law. Writing $\mathcal{R}(\pi)$ for the revenue realized by a schedule on a held-out block, we report the paired cost against TWAP in basis points,

$$c(\pi) := \frac{\mathcal{R}(\mathrm{TWAP}) - \mathcal{R}(\pi)}{\bar{X}} \times 10^4, \tag{31}$$

so that positive values indicate higher execution cost than TWAP and c vanishes for TWAP by construction. Every method is evaluated on identical blocks.

G.5.2 Calibration sensitivity

The representative levels of Section 7.2 are centered empirical medians. An alternative normalization rescales the most deviant state to $\pm 10\%$, which on this data amplifies the cross state spread from 32.19 basis points to 1984.59 basis points, a factor of 61.7. Under that normalization the learned policy deviates from TWAP and incurs $+74.59$ basis points at $\eta = 0.02$, whereas under the empirical levels it reproduces TWAP exactly at the same η. The amplified variant is reported here for completeness and is not used for any claim in the main text, since the agent would then be trained on an exogenous signal substantially larger than the hold out data contain.

G.5.3 Kernel diagnostics

The calibrated five state kernel has row maxima $(0.807, 0.596, 0.567, 0.597, 0.806)$ and mean self transition probability 0.675 against a uniform baseline of 0.200. No row is near deterministic, so the concentration of the estimate at $367{,}920$ calibration observations does not degenerate the exogenous process. State occupancy is exactly uniform by construction of the quantile bins.

G.5.4 Reference policies

Two references are computed on the same blocks. The information matched oracle is the optimal policy of the form $\pi(h, x, \xi)$ under an (h, ξ) Markov fit of the hold out dynamics and payoffs, obtained by backward induction; it is deliberately fitted on the evaluation data, so it serves as a reference rather than a method. The clairvoyant policy is the per block dynamic programming optimum given the realized price path in advance. Neither is attainable. The oracle maximizes expected value under a Markov fit rather than directly over the empirical path distribution, so it is a reference point and not a proven upper bound on the policy class.

Table 2 reports the paired costs at the two impact levels highlighted in the main text. The information matched oracle does not beat TWAP at either value, which suggests that learning error is not the dominant limitation under this state description. The gap between the oracle and the clairvoyant reference also reflects future uncertainty and the Markov approximation, and should not be read as a pure representation gap.

G.5.5 Impact sweep

The full eight point impact sweep records, at each η, the mean paired cost with bootstrap confidence interval, the fraction of executed actions differing from TWAP and the fraction of paths

Table 2: Paired cost against TWAP in basis points, held-out blocks. Lower is better. The information matched oracle is fitted on the evaluation data and is a reference, not an attainable method.

	$\eta = 0.0008$	$\eta = 0.0002$
TWAP	0.0000	0.0000
Exo-UCBVI ($\varepsilon = 0$)	$+0.7413$	$+1.0972$
Information matched oracle	$+0.0000$	$+0.8577$
Clairvoyant	-1.8906	-4.5838

containing any deviation, both measured by replaying the learned policy and TWAP along the same held-out state paths with the feasibility projection applied to each. Because η is not identified by public price data, it is treated as an assumed impact regime and swept over the pre-declared grid $\eta \in \{0.02, 0.01, 0.005, 0.0032, 0.0016, 0.0008, 0.0004, 0.0002\}$, fixed before the sweep was run. Equating the cost of the smallest deviation from TWAP with its largest possible gain gives a first-order two-period threshold $\eta^\star \approx 0.00161$, obtained as half the 32.19 basis point spread of the calibrated state levels. Fig. 11 visualizes these quantities. The left panel shows that the paired cost is exactly zero throughout the upper part of the grid and becomes positive only once η falls below the first-order threshold near $\eta^\star \approx 0.0016$. Among the three smallest values of η, the paired cost is largest at $\eta = 0.0004$ and then falls at $\eta = 0.0002$, so the cost response is non-monotone after deviations from TWAP begin. The right panel shows the corresponding behavioral change. Both deviation measures are exactly zero for $\eta \geq 0.0016$ and then rise as η decreases, with the path level deviation increasing faster than the executed action deviation because changing a small subset of actions is already enough for a trajectory to count as deviating. The information-matched oracle executes TWAP on every path at $\eta = 0.0008$ and deviates on 87.6% of paths at $\eta = 0.0002$ without beating TWAP at either level. At $\eta = 0.0002$ the learned policy is 0.24 basis points above the information- matched oracle. The action-dependent execution cost in Eq. (30) is imposed by the model rather than measured from the order-book, so neither the ground-truth optimum nor the causal impact of a hypothetical order is identified by this replay.

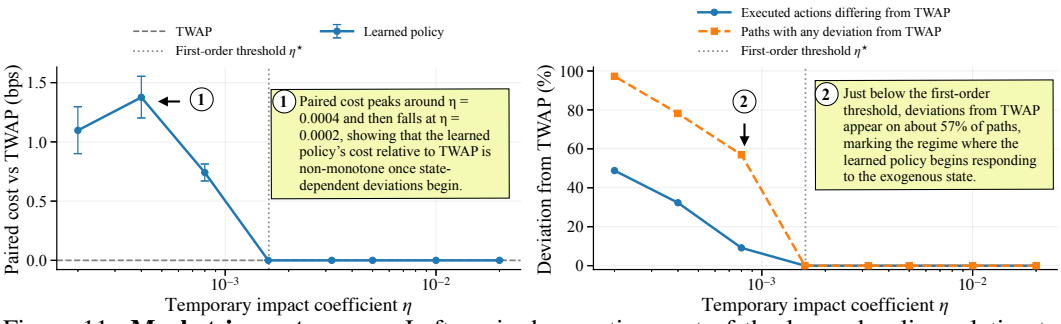


Figure 11: **Market impact sweep.** Left: paired execution cost of the learned policy relative to TWAP, in basis points, against the temporary impact coefficient η on a logarithmic scale; error bars show bootstrap 95% confidence intervals. Right: the percentage of executed actions that differ from TWAP and the percentage of held-out paths with any deviation from TWAP. In both panels, the vertical dotted line marks the first order threshold $\eta^\star \approx 0.00161$.