REGRET MINIMIZATION IN FACTORED MDPS WITH ENDOGENOUS AND EXOGENOUS FACTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study regret minimization in non-episodic factored MDPs whose state has two parts. The exogenous part, such as the demand arriving at a server, has transitions that do not depend on the action. The endogenous part has transitions that do. The reward may depend jointly on both parts and the action, and the endogenous dynamics are unknown and stochastic. In the factored-MDP framework an exogenous factor is a factor whose parent scope omits the action. Its visits are then counted across actions, and the existing analysis gives it a regret term free of the number of actions $|\mathcal{A}|$. We show that this is all the partition gives in the rate, and that it can give more in the prefactor. We prove the first regret lower bound for this class. Its endogenous term scales with $\sqrt{|\mathcal{A}|}$ and its exogenous term does not, so it matches the upper rate in T and in $|\mathcal{A}|$, and the exogenous term persists under a joint reward, which separates this setting from the exogenous-information line, where a restriction on the reward removes the exogenous part from learning. The prefactor of the existing bounds is the diameter, which counts the time to reach the rarest exogenous configuration, a time no policy controls. When the exogenous transitions are i.i.d., a plug-in variant without optimism on the exogenous part carries the endogenous part of the diameter in place of the diameter, and plans on a model smaller by a factor exponential in the number of exogenous factors. In experiments pooling has lower regret than every baseline, on serving instances driven by a real request trace it removes almost all of the regret of the baseline that ignores the partition, and the plug-in variant matches its regret on instances whose diameter is large while its endogenous part is small, at a fraction of the planning time.

1 INTRODUCTION

Provably efficient tabular RL, from (Kearns & Singh, 1998; Brafman & Tennenholtz, 2002; Auer & Ortner, 2007) to UCRL2 (Jaksch et al., 2010) and its bandit-style optimism (Auer et al., 2002; Lattimore & Szepesvári, 2020), has regret polynomial in the number of states, which is uninformative when the state is a product of many components. **Factored MDPs** (Osband & Van Roy, 2014) represent the transition kernel as a sparse dynamic Bayesian network over the factors and replace that dependence by a polynomial one in the number of factors.

Factorization leaves a second kind of structure unused. In the standard model every factor's transition depends on the action. In many systems, however, a large part of the state evolves independently of the action, for example order arrivals and demand in a warehouse, prices and downstream demand in inventory control, user mobility and traffic load in a cellular network (Dietterich et al., 2018), and the request stream at an LLM serving system. We call a factor *endogenous* if the action enters its transition and *exogenous* if it does not.

In the factored-MDP framework of Osband & Van Roy (2014) the action is one coordinate of a factor's parent scope, and an exogenous factor is a factor whose scope omits it. Its transition is then estimated at parent values instead of parent–action pairs, each cell collects the visits of every action, and the analysis of Xu & Tewari (2020); Rosenberg & Mansour (2021) gives it a regret term that does not grow with the number of actions, whereas the term of an endogenous factor grows with its square root. Declaring the factor endogenous, as one must when the action is not a scope coordinate

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.

(Rosenberg & Mansour, 2021, NFA-DORL), pays that square root on a component no action can explore. We write the pooled algorithm in that form and call it **Hybrid-DORL**.

The partition gives nothing more in the dependence on T and on the number of actions. We prove the first regret lower bound for factored MDPs with a mixed endogenous–exogenous partition. Its endogenous term keeps the square root of the number of actions, so the pooling cannot be extended to the endogenous factors, and its exogenous term does not vanish, so the exogenous part must be learned although no action affects it. The second point separates our setting from the exogenous-information line (Dietterich et al., 2018; Efroni et al., 2022a;b; Sinclair et al., 2023; Wan et al., 2024), which removes the exogenous part from learning under a restriction on the reward or on the endogenous dynamics (Appendix I). Without that restriction the exogenous part must be learned, and the learner saves only that square root.

The partition can give more in the prefactor. The upper bounds of Xu & Tewari (2020); Rosenberg & Mansour (2021) are multiplied by the diameter of the MDP, the longest expected time to steer between two states. With exogenous factors this includes the time to reach the rarest exogenous configuration, which no policy controls, so the diameter can be arbitrarily large while the endogenous part of the MDP stays small. When the exogenous transitions are i.i.d., a plug-in variant, **Hybrid-DORL-PI**, gives the planner the estimated exogenous distribution with no optimism on the exogenous factors, and its bound carries the endogenous part of the diameter, the time to steer between two endogenous states, in place of the diameter. A memoryless exogenous state contributes at most one step to the bias span of the planning model, whereas the standard analysis charges its diameter. The planning model also shrinks by a factor exponential in the number of exogenous factors.

Contributions. (1) The first regret lower bound for factored MDPs with a mixed endogenous–exogenous partition, whose endogenous term grows with the number of actions and whose exogenous term does not, so that pooling is all the partition gives in the dependence on T and on the number of actions, and the exogenous part must be learned under a joint reward. It is matched against the upper bound that the existing analysis gives to NFA-DORL with pooled exogenous counts, which we call Hybrid-DORL. (2) Hybrid-DORL-PI, a plug-in variant for i.i.d. exogenous transitions whose bound carries the endogenous part of the diameter in place of the diameter and whose planning model shrinks exponentially in the number of exogenous factors. (3) Experiments on random instances and on a real LLM-serving trace, in which Hybrid-DORL has lower regret than every baseline, on the trace it removes almost all of the regret of the baseline that ignores the partition, and the plug-in variant matches it on instances whose diameter is large while the endogenous part stays small, at a fraction of the planning cost.

2 PRELIMINARIES

Notation. For a finite set X, Δ_X denotes the probability simplex over X and $\mathrm{Unif}(X) \in \Delta_X$ the uniform distribution on it. We write $\log$ for the natural logarithm. $\widetilde{O}(\cdot)$ hides polylogarithmic factors and $\Omega(\cdot)$ universal constants.

We study a non-episodic, average-reward MDP over an infinite horizon (Puterman, 1994; Sutton & Barto, 2018). It is a tuple $M = (\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite sets of states and actions, $P(\cdot \mid s, a) \in \Delta_{\mathcal{S}}$ is the distribution of the next state given the current state-action pair, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a deterministic reward.[1] Interaction proceeds in discrete rounds $t = 1, 2, \ldots$ from an arbitrary starting state s^1. At round t the learner observes s^t, selects $a^t \in \mathcal{A}$, receives reward $r^t = r(s^t, a^t)$, and moves to $s^{t+1} \sim P(\cdot \mid s^t, a^t)$.

A (stationary, deterministic) policy is a map $\pi : \mathcal{S} \to \mathcal{A}$, and its *gain* is its long-run average reward $\lambda(M, \pi, s) := \lim_{T\to\infty} \frac{1}{T}\,\mathbb{E}\big[\sum_{t=1}^{T} r(s^t, \pi(s^t)) \mid s^1 = s\big]$, where $s^{t+1} \sim P(\cdot \mid s^t, \pi(s^t))$. Let $T(s' \mid M, \pi, s)$ be the random first-passage time to s' when π is run from s in M. The *diameter* of M is $D(M) := \max_{s \neq s'} \min_{\pi:\mathcal{S}\to\mathcal{A}} \mathbb{E}[T(s' \mid M, \pi, s)]$, and we assume throughout that $D(M) < \infty$, the standard *communicating* assumption. Under it the optimal gain is achieved by a stationary

[1] Known deterministic rewards are standard in this line (Jaksch et al., 2010), and stochastic rewards add only lower-order terms.

policy independent of the initial state, so we write $\pi^\star \in \arg\max_\pi \lambda(M,\pi,s^1)$ and $\lambda^\star(M) := \lambda(M,\pi^\star,s^1)$. After T rounds the *regret* of the learner is $\mathrm{Reg}_T(M) := \sum_{t=1}^{T}\big(\lambda^\star(M) - r^t\big)$.

The *bias function* $h^\star : \mathcal{S} \to \mathbb{R}$ measures the relative value of the starting state, and with $\lambda^\star(M)$ it satisfies the Bellman optimality equation

$$\lambda^\star(M) + h^\star(s) \;=\; \max_{a\in\mathcal{A}}\Big[r(s,a) + \sum_{s'\in\mathcal{S}} P(s' \mid s,a)\, h^\star(s')\Big] \tag{1}$$

at every $s \in \mathcal{S}$, where $\pi^\star(s)$ attains the maximum. For a function f on $\mathcal{S}$ we write $\mathrm{sp}(f) := \max_s f(s) - \min_s f(s)$. The span $\mathrm{sp}(h^\star)$ is at most $D(M)$ (Bartlett & Tewari, 2009).

Factored MDPs. Suppose $\mathcal{S} = \bigotimes_{i\in[d]} \mathcal{S}_i$ factorizes into a product of d component sets, where $[d] = \{1,\dots,d\}$. For any subset $Z \subseteq [d]$, write $\mathcal{S}[Z] := \bigotimes_{i\in Z}\mathcal{S}_i$ and $s[Z] := (s[i])_{i\in Z}$ for the scope set and scope variable, and $s[i]$ for $s[\{i\}]$. Given parent scopes $\{Z_i^P \subseteq [d]\}_{i\in[d]}$, the MDP is *factored* if its transition kernel decomposes as $P(s' \mid s,a) = \prod_{i\in[d]} P_i(s'[i] \mid s[Z_i^P], a)$, forming a sparse dynamic Bayesian network (DBN) (Osband & Van Roy, 2014). This is the non-factored-action form of Rosenberg & Mansour (2021). In the form of Osband & Van Roy (2014) the action is a coordinate of the scope, and the two coincide when every scope contains it.

Planning oracle. Exact planning on a factored MDP can take time exponential in d (Goldsmith et al., 1997; Littman, 1997). Following Xu & Tewari (2020) and Rosenberg & Mansour (2021), we assume an FMDP planning oracle. Given a *communicating* factored MDP, the oracle returns a policy, a gain and a bias that satisfy (1) on it. An algorithm is *oracle-efficient* if it makes polynomially many oracle calls and does polynomial work outside them.

3 Problem Setting

We consider factored MDPs $M = (\mathcal{S},\mathcal{A},P,r)$ of §2 whose factor index set is partitioned as $[d] = I^{\mathrm{endo}} \cup I^{\mathrm{exo}}$, so that $\mathcal{S} = \mathcal{S}_{\mathrm{endo}} \times \mathcal{S}_{\mathrm{exo}}$ with $\mathcal{S}_{\mathrm{endo}} = \bigotimes_{i\in I^{\mathrm{endo}}}\mathcal{S}_i$ and $\mathcal{S}_{\mathrm{exo}} = \bigotimes_{i\in I^{\mathrm{exo}}}\mathcal{S}_i$, and we write $s = (s_{\mathrm{endo}}, s_{\mathrm{exo}})$. The partition and the parent scopes are known to the learner, and the kernel P is unknown and learned from a single non-episodic interaction. The reward r may depend jointly on s_{endo}, s_{exo} and the action, which separates our setting from the exogenous-information line of Appendix I. We write $d_{\mathrm{endo}} := |I^{\mathrm{endo}}|$ and $d_{\mathrm{exo}} := |I^{\mathrm{exo}}|$, order the indices so that $I^{\mathrm{endo}} = \{1,\dots,d_{\mathrm{endo}}\}$, and let $m := \max_i |Z_i^P|$, $W := \max_i |\mathcal{S}_i|$ and $\Lambda := \max_i |\mathcal{S}[Z_i^P]| \le W^m$.

Definition 1 (Factored MDP with exogenous structure)**.** *A factored MDP with the partition $[d] = I^{\mathrm{endo}} \cup I^{\mathrm{exo}}$ has* exogenous structure *if every exogenous factor $i \in I^{\mathrm{exo}}$ satisfies: (i) the kernel P_i is action-independent, i.e. $P_i(s'[i] \mid s[Z_i^P], a) = P_i(s'[i] \mid s[Z_i^P])$ for all $a \in \mathcal{A}$, and (ii) the parent scope is confined to the exogenous index set, $Z_i^P \subseteq I^{\mathrm{exo}}$.*

Under Definition 1 the factored kernel of §2 specializes to

$$P(s' \mid s,a) = \prod_{i\in I^{\mathrm{endo}}} P_i\big(s'[i] \,\big|\, s[Z_i^P],\, a\big) \;\cdot\; \prod_{i\in I^{\mathrm{exo}}} P_i\big(s'[i] \,\big|\, s[Z_i^P]\big). \tag{2}$$

Grouping the two products gives $P(s' \mid s,a) = P_{\mathrm{endo}}\big(s'_{\mathrm{endo}} \,\big|\, s_{\mathrm{endo}}, s_{\mathrm{exo}}, a\big) \cdot P_{\mathrm{exo}}\big(s'_{\mathrm{exo}} \,\big|\, s_{\mathrm{exo}}\big)$, so the exogenous component evolves as an autonomous Markov chain on $\mathcal{S}_{\mathrm{exo}}$, and its marginal is independent of the agent's policy. Figure 1(a) shows an instance with $d = 5$, $d_{\mathrm{endo}} = 2$ and $d_{\mathrm{exo}} = 3$. Its scopes are $Z_1^P = \{1,3\}$, $Z_2^P = \{1,2,4\}$, $Z_3^P = \{3\}$, $Z_4^P = \{3,4\}$ and $Z_5^P = \{4,5\}$, and the reward reads $s[1]$, $s[3]$ and a.

4 Algorithm

4.1 Overview

Hybrid-DORL is the DORL of Rosenberg & Mansour (2021) in its non-factored-action form, with one change: the visit counts of an exogenous factor are pooled across actions. In the factored-action

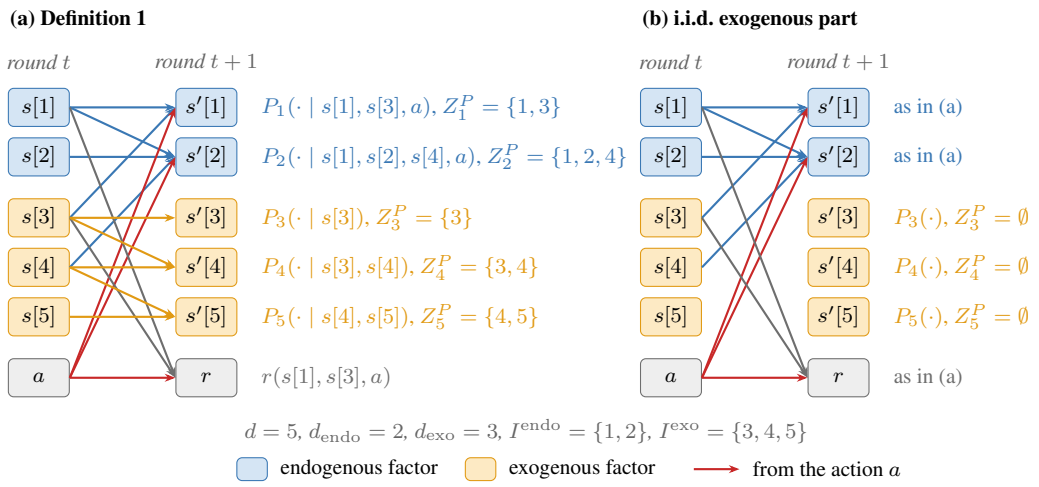


Figure 1: One round of the instance of §3. An arrow into a round-$(t+1)$ node marks an argument of that node's kernel or of the reward, listed on the right. (a) Under Definition 1, no arrow leaves a toward an exogenous factor and no exogenous factor has an endogenous parent. (b) The same instance with $Z_i^P = \emptyset$ on the exogenous factors, the i.i.d. case of §4.4.

form of the same paper it is DORL with the action left out of every exogenous scope. It runs in episodes in the optimism-in-the-face-of-uncertainty pattern of UCRL2 (Jaksch et al., 2010). At the start of episode k it computes an optimistic policy π^k and plays it until the episode ends (Algorithm 1, with full pseudocode in Appendix A). It has three components: **(i) asymmetric estimation** builds confidence sets for every factor, differing only in how visits are counted; **(ii) lifting** embeds the factor-wise choice of optimistic kernel into an augmented factored MDP (§4.3); **(iii) planning** calls a standard FMDP oracle on it. The lift and the planning step are those of Xu & Tewari (2020); Rosenberg & Mansour (2021), so the algorithm inherits their oracle-efficiency, and the exogenous structure enters in the estimation step.

4.2 Asymmetric Estimation

Take an endogenous factor $i \in I^{\text{endo}}$. Let $N^k_{Z_i^P}(v, a)$ be the number of rounds up to the end of episode k at which its parent scope took the value $v \in \mathcal{S}[Z_i^P]$, its parent cell, and the action was a. Let $N^k_{i,Z_i^P}(v, a, w)$ be the number of those rounds at which factor i's next value was $w \in \mathcal{S}_i$. At the start of episode k only the counts through episode $k-1$ are available, so the empirical kernel is $\hat{P}_i^k(w \mid v, a) := N^{k-1}_{i,Z_i^P}(v, a, w)/N^{k-1}_{Z_i^P}(v, a)$, or $1/|\mathcal{S}_i|$ at a pair not yet visited. Its confidence radius is Bernstein-type,

$$\varepsilon_i^k(w \mid v, a) := \sqrt{\frac{2\,\hat{P}_i^k(w \mid v, a)\,\tau_i^k}{\max\{N^{k-1}_{Z_i^P}(v, a),\, 1\}}} + \frac{7\,\tau_i^k}{3\,\max\{N^{k-1}_{Z_i^P}(v, a) - 1,\, 1\}}, \tag{3}$$

where $\tau_i^k := \log(40\, d_{\text{endo}}\, W \Lambda\, |\mathcal{A}|\, t_k^3/\delta)$ is the log factor, t_k the round at which episode k begins and δ the confidence parameter.

An exogenous factor $i \in I^{\text{exo}}$ drops the action argument and pools the same two counts over it, $N^k_{Z_i^P}(v) := \sum_a N^k_{Z_i^P}(v, a)$ and $N^k_{i,Z_i^P}(v, w) := \sum_a N^k_{i,Z_i^P}(v, a, w)$. Then $\hat{P}_i^k(w \mid v)$ and $\varepsilon_i^k(w \mid v)$ are the empirical kernel and (3) with three replacements: $N^{k-1}_{Z_i^P}(v)$ for $N^{k-1}_{Z_i^P}(v, a)$, then $N^{k-1}_{i,Z_i^P}(v, w)$ for $N^{k-1}_{i,Z_i^P}(v, a, w)$, then d_{exo} for $d_{\text{endo}}|\mathcal{A}|$ in τ_i^k. Pooling is legitimate because P_i is action-free (Definition 1(i)), and it leaves an exogenous factor with $|\mathcal{S}[Z_i^P]|$ cells instead of $|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|$.

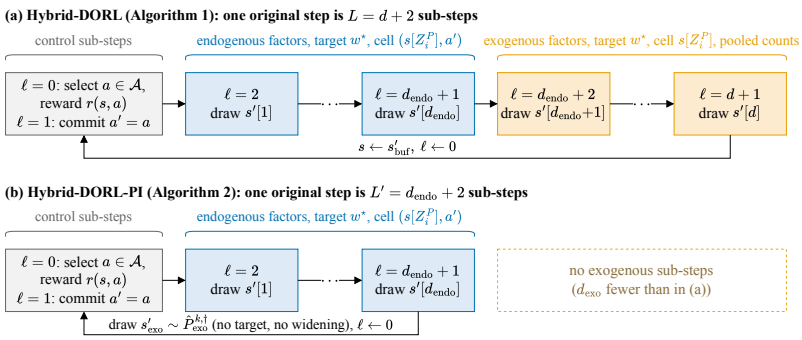


Figure 2: The two lifts. (a) Hybrid-DORL unfolds one original step into $L = d+2$ sub-steps, one per factor after the action is selected and committed, and draws each factor from the optimistic row (4) at the target chosen by the planner. (b) Hybrid-DORL-PI unfolds the endogenous factors only and draws the exogenous coordinate from the smoothed exogenous estimate at the last transition, with no target and no widening.

An episode ends as soon as the visit count to some endogenous (v, a) pair or to some exogenous parent value v doubles. A standard doubling argument bounds the number of episodes by $\widetilde{O}(\sum_{i \in I^{\text{endo}}} |\mathcal{S}[Z_i^P]|\,|\mathcal{A}| + \sum_{i \in I^{\text{exo}}} |\mathcal{S}[Z_i^P]|)$ (Appendix B).

4.3 Lifting Optimism into a Factored MDP

At each episode k the algorithm must pick a policy and a transition kernel together, where the kernel may lie anywhere within the radii ε_i^k of §4.2, which is intractable in general (Osband & Van Roy, 2014). We use the lift of Xu & Tewari (2020); Rosenberg & Mansour (2021) unchanged. It turns the choice of kernel into a choice of actions, one factor at a time: the planner names one next value per factor, a *target*, and an ordinary FMDP planner chooses the policy and the kernel together. Figure 2(a) shows the lift.

One original step of M is unfolded into $L := d+2$ sub-steps of an augmented factored MDP $\widetilde{M}^k_{\mathcal{B}} = (\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}, \widetilde{P}^k_{\mathcal{B}}, \tilde{r})$. The tilde marks the sub-step space, and $\mathcal{B}$ denotes the confidence sets of episode k. A state $\tilde{s} = (s, \ell, a', s'_{\text{buf}})$ holds the original state, the sub-step counter, the action committed at sub-step 0 and a buffer of the factor values drawn so far, with $\perp$ for a slot not yet filled (Appendix A.1). An action is either an original action or a target value for one factor, $\widetilde{\mathcal{A}} = \mathcal{A} \cup \bigcup_{i \in [d]} \mathcal{S}_i$. Sub-step 0 selects the action and collects $\tilde{r}(\tilde{s}, \tilde{a}) = r(s, \tilde{a})$, the only reward of the L sub-steps, and sub-step 1 records it in a'. Sub-step $i+1$ draws factor i at the parent cell $v = s[Z_i^P]$ from the row that moves the whole slack onto the target $w^\star \in \mathcal{S}_i$ named by the action,

$$P^k_{\mathcal{B},i}(w \mid v, a') \;=\; \hat{P}^k_i(w \mid v, a') - \mathcal{W}^k_i(w \mid v, a') + \mathbb{1}\{w = w^\star\} \sum_{w' \in \mathcal{S}_i} \mathcal{W}^k_i(w' \mid v, a'), \qquad (4)$$

where $\mathcal{W}^k_i := \min\{\varepsilon^k_i, \hat{P}^k_i\}$ is the slack by which the row is widened, truncated so that the row stays nonnegative, and exogenous factors drop the action argument. The $|\mathcal{S}_i|$ targets are the vertices of factor i's confidence polytope (Lemma 4), so a planner over augmented policies searches over the kernels in the confidence sets. $\widetilde{M}^k_{\mathcal{B}}$ has $2d+2$ components with parent scopes of size at most $m+3$, so a standard oracle plans on it (Appendix A.2). The oracle returns $\tilde{\pi}^k$, and we play $\pi^k(s) := \tilde{\pi}^k\big((s, 0, \perp, \perp)\big)$.

4.4 Hybrid-DORL-PI: No Lift on an i.i.d. Exogenous Part

Exogenous inputs are often modeled as i.i.d. across time, as arrivals and channel states in network control (Tassiulas & Ephremides, 1992; Georgiadis et al., 2006; Neely, 2010) and demands in inventory control (Bertsekas, 2017, §1.1). We model this by setting $Z_i^P = \emptyset$ for every $i \in I^{\text{exo}}$, as in Figure 1(b), and then the exogenous part needs neither optimism nor the lift. Optimism explores, but on an exogenous factor there is nothing to explore. By Definition 1 the samples arrive at every step

Algorithm 1 Hybrid-DORL (sketch)

1: **Input:** confidence parameter δ, parent scopes $\{Z_i^P\}_{i\in[d]}$, partition $[d] = I^{\mathrm{endo}} \cup I^{\mathrm{exo}}$, reward r.
2: **for** episode $k = 1, 2, \ldots$ **do**
3: *(1) Asymmetric estimation.* For every factor $i \in [d]$ form $\hat{P}_i^k$ and $\mathcal{W}_i^k := \min\{\varepsilon_i^k, \hat{P}_i^k\}$ from (3), with visits counted per (v, a) for $i \in I^{\mathrm{endo}}$ and per v for $i \in I^{\mathrm{exo}}$.
4: *(2) Lift.* Build the augmented factored MDP $\widetilde{M}_{\mathcal{B}}^k$ of §4.3.
5: *(3) Plan and execute.* Call the planning oracle on $\widetilde{M}_{\mathcal{B}}^k$ and play $\pi^k(s) := \tilde{\pi}^k((s, 0, \perp, \perp))$ until some visit count of §4.2 doubles.
6: **end for**

Algorithm 2 Hybrid-DORL-PI: Hybrid-DORL with the exogenous part plugged in

1: **Requires:** i.i.d. exogenous transitions, that is, $Z_i^P = \emptyset$ for every $i \in I^{\mathrm{exo}}$.
2: As Algorithm 1, with two changes at $i \in I^{\mathrm{exo}}$ in every episode k:
3: *(1) No widening.* $\mathcal{W}_i^k := 0$, and the planner is given the smoothed estimate $\hat{P}_i^{k,\dagger}$ of (5).
4: *(2) Endogenous lift.* Only the endogenous factors are unfolded, into $L' := d_{\mathrm{endo}} + 2$ sub-steps, and the exogenous coordinate is drawn from $\prod_{i\in I^{\mathrm{exo}}} \hat{P}_i^{k,\dagger}$ at the last transition (Appendix A.3).
5: Plan and play on the resulting augmented MDP as before.

under every policy, so the plug-in estimate converges at the rate $1/\sqrt{t}$ on its own, and the planned gain is off by at most the bias span of the planning model times that error. With an i.i.d. exogenous part that span does not include the time to reach a rare exogenous value, since the exogenous state is redrawn at every step. The lift is not needed because no target has to be chosen for a factor that is not widened. Algorithm 2 drops both on the exogenous factors, as Figure 2(b) shows. It sets $\mathcal{W}_i^k := 0$, plans with the smoothed estimate

$$\hat{P}_i^{k,\dagger} \;:=\; \Big(1 - \tfrac{1}{t_k}\Big)\hat{P}_i^k \;+\; \tfrac{1}{t_k}\,\mathrm{Unif}(\mathcal{S}_i), \tag{5}$$

which keeps the planning model communicating, and does not unfold the exogenous factors. The augmented state space then loses one buffer coordinate per exogenous factor and shrinks by a factor of about $W^{d_{\mathrm{exo}}}$, and the cost of one sweep of value iteration, which is proportional to the number of augmented states, shrinks by at least as much.

5 Regret Analysis

Proposition 1 (Regret upper bound)**.** *For any communicating factored MDP with endogenous–exogenous structure* (2)*, Hybrid-DORL satisfies, with probability at least* $1 - \delta$*,*

$$\mathrm{Reg}_T(M) \;=\; \widetilde{O}\Big(D \sum_{i\in I^{\mathrm{endo}}} \sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T} \;+\; D \sum_{i\in I^{\mathrm{exo}}} \sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,T}\Big), \qquad D := D(M). \tag{6}$$

Each factor contributes one term, with the number of cells at which it is estimated under the root: parent–action pairs for an endogenous factor, parent values alone for an exogenous one. The prefactor D bounds the bias span of the planning model, which multiplies every one-step estimation error. Proposition 1 is what the analysis of Xu & Tewari (2020); Rosenberg & Mansour (2021) gives for DORL in its factored-action form when the action is left out of every exogenous scope, written in the non-factored-action notation, and with no exogenous factors it is the known-reward case of Rosenberg & Mansour (2021, Theorem 2). We prove it under Definition 1 in Appendix C because the proof of the plug-in bound below reuses its parts.

For Theorem 1, write $T_{\mathrm{endo}}(s'_{\mathrm{endo}} \mid M, \pi, s)$ for the first-passage time to the *endogenous* coordinate s'_{endo}, and

$$D_{\mathrm{endo}}(M) \;:=\; \max_{s_{\mathrm{endo}} \neq s'_{\mathrm{endo}}} \; \max_{s_{\mathrm{exo}}} \; \min_{\pi} \; \mathbb{E}\big[T_{\mathrm{endo}}(s'_{\mathrm{endo}} \mid M, \pi, (s_{\mathrm{endo}}, s_{\mathrm{exo}}))\big] \;\le\; D(M) \tag{7}$$

for the endogenous part of the diameter, which requires only that the endogenous coordinates be reached.

Theorem 1 (Plug-in bound under i.i.d. exogenous factors)**.** *Suppose $Z_i^P = \emptyset$ for every $i \in I^{\mathrm{exo}}$, so that each exogenous P_i is a single distribution on $\mathcal{S}_i$. Then Hybrid-DORL-PI (Algorithm 2) satisfies, with probability at least $1-\delta$,*

$$\mathrm{Reg}_T(M) \;=\; \widetilde{O}\Big(\big(1+D_{\mathrm{endo}}\big)\Big[\sum_{i\in I^{\mathrm{endo}}}\sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T}\;+\;\sum_{i\in I^{\mathrm{exo}}}\sqrt{|\mathcal{S}_i|\,T}\Big]\Big), \tag{8}$$

with $D_{\mathrm{endo}} := D_{\mathrm{endo}}(M) \le D(M)$ of (7)*.*

Bound (8) is Proposition 1 at $|\mathcal{S}[Z_i^P]| = 1$ on the exogenous factors with D replaced by $1 + D_{\mathrm{endo}}$. Its proof (Appendix D) differs at two points. The exogenous coordinate of the plug-in model is redrawn from one fixed distribution at every step, so the Bellman equation reads the bias only through its average over that distribution, a function of the endogenous state alone, and the span is at most one plus the endogenous diameter of the plug-in model. Without optimism on the exogenous part the planned gain can fall short of $\lambda^\star(M)$ by the span times the exogenous estimation error per step, and that error sums to the exogenous term of (8).

6 Lower Bound

We now show that no learning algorithm can remove the factor $\sqrt{|\mathcal{A}|}$ from the endogenous term of Proposition 1, and that none can remove the exogenous term.

Theorem 2 (Regret lower bound)**.** *Fix $d_{\mathrm{endo}} \ge 1$, $d_{\mathrm{exo}} \ge 2m$, $m \ge 1$, $W \ge 2$ and $|\mathcal{A}| \ge 2$, and write $d = d_{\mathrm{endo}} + d_{\mathrm{exo}}$. There is a universal constant c such that for any learning algorithm and any $T \ge c\,|\mathcal{A}|\,d\,W^m$ there exists a factored MDP M satisfying Definition 1 and the model of §3, with $2d + \lceil \log_2 d \rceil + 3$ factors of size at most $W+1$, of which $d_{\mathrm{endo}}+1$ are endogenous, and with parent scopes of size at most $1 + \max\{m, \lceil \log_2 d \rceil + 2\}$, such that*

$$\mathbb{E}[\mathrm{Reg}_T(M)] \;=\; \Omega\Big(\sqrt{d_{\mathrm{endo}}\cdot W^m\cdot|\mathcal{A}|\cdot T}\;+\;\sqrt{d_{\mathrm{exo}}\cdot W^m\cdot T}\Big). \tag{9}$$

The endogenous term scales with $\sqrt{|\mathcal{A}|}$, and the exogenous term does not.

The rates match in T and in $|\mathcal{A}|$, which is present on the endogenous side and absent on the exogenous side. The exogenous term is the price of a joint reward: the learner has to learn transitions it cannot influence, and the partition saves it only the factor $\sqrt{|\mathcal{A}|}$. The remaining gap is the diameter, a $\sqrt{d\,W}$ Cauchy–Schwarz slack, and the hard instance having more factors and larger scopes than the parameters d and m. We claim matching rates, not minimax optimality (Remark 1). The two terms come from two hypothesis tests: at an endogenous parent–action pair the learner must find which of $|\mathcal{A}|$ actions is biased, and at an exogenous parent value it must find which way a coin is biased, which no action can speed up. Appendix E gives the construction and the proof.

7 Evaluation

We evaluate on synthetic instances generated at random (Appendix F) and on instances whose exogenous factors are fitted to a real LLM-serving trace (Appendix G). The main instance has $d_{\mathrm{endo}} = 1$, $d_{\mathrm{exo}} = 3$, $W = 4$ and $|\mathcal{A}| = 8$. Its kernel rows and reward vectors are sampled once, from a Dirichlet distribution and uniformly, and then fixed, and the reward depends jointly on one endogenous and one exogenous coordinate. The baselines are UCRL2 (Jaksch et al., 2010), FACTORED-UCRL (Osband & Van Roy, 2014) and NFA-DORL (Rosenberg & Mansour, 2021), and Appendix H specifies all four algorithms.

7.1 Cumulative regret

Figure 3 plots cumulative regret on the main instance. Hybrid-DORL has the lowest regret, 22% below FACTORED-UCRL and 38% below NFA-DORL. The comparison with NFA-DORL isolates the effect of the partition, since NFA-DORL is Hybrid-DORL with every confidence set indexed by the action. With the partition, each exogenous cell collects $|\mathcal{A}|$ times more samples and its radius is $\sqrt{|\mathcal{A}|}$ times smaller, the factor that separates the two sums of Proposition 1.

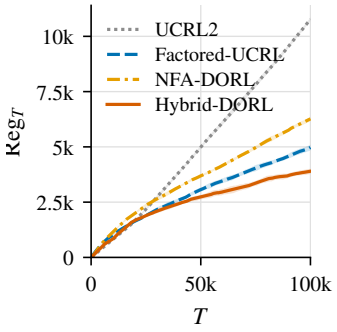


Figure 3: Cumulative regret.

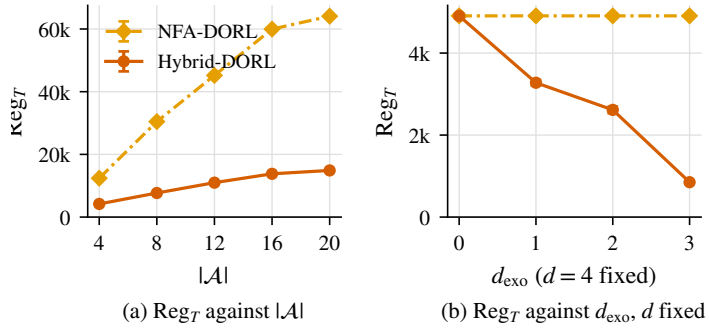
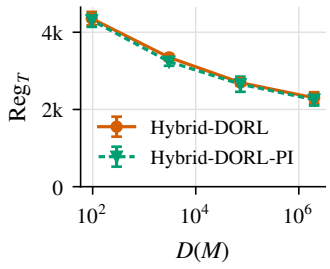
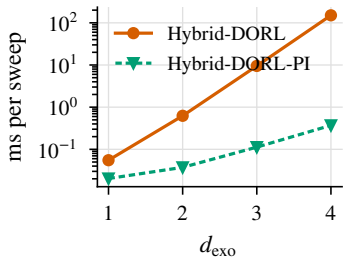


Figure 4: Regret against actions (a) and against d_{exo}, d fixed (b).



(a) Reg$_T$ against $D(M)$, $D_{\text{endo}}(M) = 2.1$



(b) sweep time, $d = 1 + d_{\text{exo}}$

Figure 5: Hybrid-DORL-PI against Hybrid-DORL: regret (a) and time per value-iteration sweep (b).

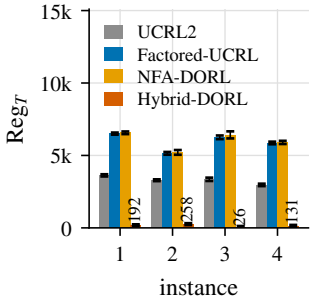


Figure 6: Regret on the trace.

7.2 Dependence on the number of actions

Figure 4(a) runs Hybrid-DORL and NFA-DORL, which ignores the endogenous–exogenous partition, at $|\mathcal{A}| = 4, 8, 12, 16, 20$ and $T = 10^6$. The action sets are nested, so a larger $|\mathcal{A}|$ adds actions with new kernel rows and changes nothing else. The regret of both grows with $|\mathcal{A}|$, the gap between them grows from 8.2×10^3 at $|\mathcal{A}| = 4$ to 4.9×10^4 at $|\mathcal{A}| = 20$, and their ratio grows from 3.0 to 4.3. Hence, with more actions, a larger share of NFA-DORL's regret is the exogenous exploration that pooling removes.

7.3 Dependence on the number of exogenous factors

Figure 4(b) runs Hybrid-DORL on a variant of the main instance with 0, 1, 2 or 3 of its three exogenous factors declared exogenous and the rest declared endogenous, next to NFA-DORL, which does not use the partition. The exogenous scopes of this instance form a chain, so each of these partitions satisfies Definition 1, and its reward depends on all three exogenous factors (Appendix F). The regret falls with every factor pooled, from 4907 at zero, where the two algorithms coincide, to 3276, 2615 and 851. Pooling all three removes 83% of NFA-DORL's regret.

7.4 Regret and planning cost of Hybrid-DORL-PI

Figure 5(a) runs an i.i.d. instance whose exogenous distributions are made more peaked from left to right. The diameter counts the time to reach the rarest joint exogenous value, whose probability falls to 7×10^{-7}, so $D(M)$ grows from 10^2 to 2×10^6, while $D_{\text{endo}}(M)$ counts only the time to reach an endogenous value and stays at 2.1. The regret of Hybrid-DORL-PI does not grow with D, as the prefactor $1 + D_{\text{endo}}$ of (8) predicts, and it even falls, from 4.3×10^3 to 2.3×10^3, since the peaked distributions leave fewer exogenous values to visit often. Hybrid-DORL matches it within 4% at every point, so optimism on an i.i.d. exogenous part costs no regret, and the factor D in Proposition 1 is not paid on these instances either. The plug-in reduces the planning cost instead.

In Figure 5(b) a sweep of value iteration over its lift takes one fixed exogenous expectation, while the lift of §4.3 chooses a target at every exogenous sub-step. The measured ratio reaches 407 at $d_{\mathrm{exo}} = 4$, above the $W^{d_{\mathrm{exo}}} = 256$ of the state count alone, because the full lift also spends d_{exo} more sub-steps per round.

7.5 A REAL EXOGENOUS PROCESS

Figure 6 uses four instances built from the BurstGPT request trace (Appendix G). The trace records the request side only, arrival times and token counts, and nothing about the server, so the exogenous factors come from the data and the server is modeled. Each instance is a server that decides once a minute. The exogenous factors are the arrival level of the minute (idle, light, busy or burst), the mean response length and the share of GPT-4 requests, each with four levels. Their transition kernels are estimated from one quarter of the sixty-one-day trace each, so the four instances are four consecutive periods. The endogenous factor is the backlog, with levels of 0, 6, 25 and 100 waiting requests. The action opens one of eight service capacities, from 0 to 40 slots. The reward is the number of served requests minus a holding cost on the backlog and a compute cost on every open slot. For example, at a burst minute with a backlog of 25, opening 40 slots serves the backlog and the arrivals but is charged for 40 slots, while opening 10 slots is cheaper and lets the backlog grow. No capacity choice changes the next arrival level. The fitted chains are far from i.i.d. (Appendix G), which is the case Proposition 1 covers.

The regret of Hybrid-DORL is at most 258 on every instance, against 5.1×10^3 to 6.6×10^3 for FACTORED-UCRL and NFA-DORL and 3.0×10^3 to 3.6×10^3 for UCRL2. NFA-DORL indexes its confidence set on the arrival kernel by the capacity, so at a rarely played capacity the optimistic model lets that capacity lead to a burst, and the planner keeps the capacity about three levels above the optimal one in idle minutes. Hybrid-DORL pools the arrival counts over the capacities, so no capacity is linked to a burst, and it plays the optimal capacity 98% of the time in idle minutes.

8 CONCLUSION

We studied regret in average-reward factored MDPs with a known endogenous–exogenous partition and a reward that depends on both parts. In the factored-MDP framework the partition is a choice of scope, and pooling the visits of an exogenous factor across actions removes $\sqrt{|\mathcal{A}|}$ from its regret term. Our lower bound shows that this is all the partition gives in the dependence on T and on $|\mathcal{A}|$: the endogenous term keeps $\sqrt{|\mathcal{A}|}$, and under a joint reward the exogenous term cannot be removed. The partition gives more in the prefactor. When the exogenous part is i.i.d., the plug-in variant pays the endogenous part of the diameter in place of the diameter, because a memoryless exogenous state adds at most one step to the bias span, and it plans on a model smaller by a factor of $W^{d_{\mathrm{exo}}}$. Whether a Markov exogenous part allows the same is open, and a bound finer than the diameter on the time the exogenous chain takes to forget its state is what is missing.

AI USE STATEMENT

We used generative AI tools for assisting in the revision of the proofs in the appendix, for checking and editing the experiment code under the authors' instructions, and for editing the manuscript for readability, creating figures and formatting references. We have not used generative AI tools to formulate the mathematical claims, to propose the hypotheses or to generate synthetic datasets, and the remaining tasks with required disclosure are not applicable to this work. All AI-assisted work was reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal*, 19(3):357–367, 1967. doi: 10.2748/tmj/1178243286.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. I*. Athena Scientific, 4th edition, 2017.

Dimitri P. Bertsekas and John N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

Ronen I. Brafman and Moshe Tennenholtz. R-max — a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Thomas G Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning. In *International Conference on Machine Learning*, 2018.

Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, pp. 5062–5127. PMLR, 2022a.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provable RL with exogenous distractors via multistep inverse dynamics. In *International Conference on Learning Representations*, 2022b.

Eyal Even-Dar, Sham M. Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Leonidas Georgiadis, Michael J. Neely, and Leandros Tassiulas. Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, 1(1):1–144, 2006. doi: 10.1561/1300000001.

Judy Goldsmith, Michael L. Littman, and Martin Mundhenk. The complexity of plan existence and evaluation in probabilistic domains. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pp. 182–189, 1997.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. In *International Conference on Machine Learning (ICML)*, 1998.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. doi: 10.1017/9781108571401.

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

Michael L. Littman. Probabilistic propositional planning: Representations and complexity. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 748–754, 1997.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

Michael J. Neely. *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Synthesis Lectures on Communication Networks. Morgan & Claypool, 2010. doi: 10.2200/S00271ED1V01Y201006CNT007.

Michael J. Neely. Opportunistic learning for Markov decision systems with application to smart robots. In *60th Annual Allerton Conference on Communication, Control, and Computing*, 2024. doi: 10.1109/Allerton63246.2024.10735289.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, 2014.

Raymond E. A. C. Paley and Antoni Zygmund. On some series of functions, (3). *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(2):190–205, 1932. doi: 10.1017/S0305004100010860.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored mdps with unknown structure. *Advances in Neural Information Processing Systems*, 34:11148–11159, 2021.

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 *International Conference on Machine Learning*, pp. 31877–31914. PMLR, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Mohammad Sadegh Talebi, Anders Jonsson, and Odalric-Ambrym Maillard. Improved exploration in factored average-reward MDPs. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Leandros Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936–1948, 1992. doi: 10.1109/9.182479.

Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored Markov decision processes. In *Advances in Neural Information Processing Systems*, 2020.

George Trimponias and Thomas G. Dietterich. Reinforcement learning with exogenous states and rewards. *arXiv preprint arXiv:2303.12957*, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Jia Wan, Sean R Sinclair, Devavrat Shah, and Martin J Wainwright. Exploiting exogenous structure for sample-efficient reinforcement learning. *arXiv preprint arXiv:2409.14557*, 2024.

Yuxin Wang, Yuhan Chen, Zeyu Li, et al. BurstGPT: A real-world workload dataset to optimize LLM serving systems. `https://github.com/HPMLL/BurstGPT`, 2024.

Xiaohan Wei, Hao Yu, and Michael J. Neely. Online learning in weakly coupled Markov decision processes: A convergence time study. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1), 2018.

Ziping Xu and Ambuj Tewari. Reinforcement learning in factored MDPs: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. In *Advances in Neural Information Processing Systems*, 2020.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 222–227, 1977. doi: 10.1109/SFCS.1977.24.

Peng Zhao, Long-Fei Li, and Zhi-Hua Zhou. Dynamic regret of online Markov decision processes. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

A Full Algorithm

This appendix gives the full pseudocode of Hybrid-DORL (Algorithm 3) and the optimistic-planning subroutine it calls (Algorithm 4). Algorithm 3 instantiates the three-step template of Algorithm 1 with concrete estimators, the empirical Bernstein radius of (3), the doubling episode-termination rule of §4.2, and per-factor visit counters. Algorithm 4 constructs the augmented MDP $\widetilde{M}^k_{\mathcal{B}}$ described in §4.3.

A.1 Detailed lift construction

The lift draws the factors in index order, so that by the convention of §3 the d_{endo} endogenous factors are drawn first and the d_{exo} exogenous ones after. The state space is

$$\widetilde{\mathcal{S}} = \underbrace{\mathcal{S}}_{s} \times \underbrace{\{0, 1, \ldots, L-1\}}_{\ell} \times \underbrace{(\mathcal{A} \cup \{\perp\})}_{a'} \times \underbrace{\textstyle\prod_{i\in[d]} (\mathcal{S}_i \cup \{\perp\})}_{s'_{\mathrm{buf}}}, \tag{10}$$

and the *consistent* part of (10), the part $\widetilde{M}^k_{\mathcal{B}}$ actually runs on, is the set

$$\big\{(s, \ell, a', s'_{\mathrm{buf}}) \;:\; s'_{\mathrm{buf}}[j] \neq \perp \iff j < \ell - 1,\;\; a' = \perp \iff \ell = 0\big\}.$$

It is closed under $\widetilde{P}^k_{\mathcal{B}}$ and under $\widetilde{P}$, since each sub-step fills exactly the next buffer slot and the reset clears all of them, and it contains every state the trajectory mapping of §C.1 produces. An exogenous factor's $\mathcal{W}^k_i$ in (4) is built from the pooled counts of §4.2. Write

$$\omega^k_i(v, a') \;:=\; \sum_{w' \in \mathcal{S}_i} \mathcal{W}^k_i(w' \mid v, a') \tag{11}$$

for the mass that (4) moves onto the target. By Lemma 4, ranging $w^\star$ over $\mathcal{S}_i$ in (4) enumerates every vertex of the one-sided confidence polytope around $\hat{P}^k_i$, so an optimal policy of $\widetilde{M}^k_{\mathcal{B}}$ is at least as optimistic as a direct bounded-FMDP solve. In addition, the lift lets a target depend on the coordinates already drawn.

Algorithm 3 Hybrid-DORL (full version)

Input: confidence parameter δ, parent scopes $\{Z^P_i\}_{i\in[d]}$, endogenous/exogenous partition $[d] = I^{\mathrm{endo}} \cup I^{\mathrm{exo}}$, state factor sizes $\{|\mathcal{S}_i|\}_{i\in[d]}$, action space $\mathcal{A}$, (known) reward function r.
Initialization: Set visit counters $N^0 \leftarrow 0$, and set $t \leftarrow 1$, and observe initial state s^1.
for episode $k = 1, 2, \ldots$ **do**
  # Start New Episode
  $t_k \leftarrow t$.
  # Update empirical estimates
  for each factor $i \in [d]$ **do**
    $\hat{P}^k_i \leftarrow$ the empirical kernel of §4.2, at every $v \in \mathcal{S}[Z^P_i]$, $w \in \mathcal{S}_i$ and $a \in \mathcal{A}$, where for $i \in I^{\mathrm{exo}}$ the action argument is dropped and the counts are aggregated over actions.
  end for
  # Set confidence radii (with the log factor τ^k_i of §4.2)
  for $i \in I^{\mathrm{endo}}$ **do**
    For all $v \in \mathcal{S}[Z^P_i]$, $a \in \mathcal{A}$, $w \in \mathcal{S}_i$:

$$\varepsilon^k_i(w \mid v, a) \;\leftarrow\; \sqrt{\frac{2\,\hat{P}^k_i(w \mid v, a)\,\tau^k_i}{\max\{N^{k-1}_{Z^P_i}(v, a),\, 1\}}} + \frac{7\,\tau^k_i}{3\,\max\{N^{k-1}_{Z^P_i}(v, a) - 1,\, 1\}}.$$

    Set transition confidence set $\mathcal{W}^k_i(w \mid v, a) := \min\big\{\varepsilon^k_i(w \mid v, a),\, \hat{P}^k_i(w \mid v, a)\big\}$.
  end for
  for $i \in I^{\mathrm{exo}}$ **do**
    For all $v \in \mathcal{S}[Z^P_i]$, $w \in \mathcal{S}_i$, with $N^{k-1}_{Z^P_i}(v) := \sum_a N^{k-1}_{Z^P_i}(v, a)$:

$$\varepsilon^k_i(w \mid v) \;\leftarrow\; \sqrt{\frac{2\,\hat{P}^k_i(w \mid v)\,\tau^k_i}{\max\{N^{k-1}_{Z^P_i}(v),\, 1\}}} + \frac{7\,\tau^k_i}{3\,\max\{N^{k-1}_{Z^P_i}(v) - 1,\, 1\}}.$$

Set $\mathcal{W}_i^k(w \mid v) := \min\{\varepsilon_i^k(w \mid v),\ \hat{P}_i^k(w \mid v)\}$.
end for
 # Compute optimistic policy
Construct optimistic augmented MDP $\widetilde{M}_{\mathcal{B}}^k$ using $\{\mathcal{W}_i^k\}_{i\in[d]}$ for every factor (see Algorithm 4).
Compute $\pi^k \leftarrow$ optimal policy of $\widetilde{M}_{\mathcal{B}}^k$ via FMDP planning oracle.
 # Execute Policy
while $\nu_Z^k(s^t[Z], \pi^k(s^t)) < \max\{N_Z^{k-1}(s^t[Z], \pi^k(s^t)),\ 1\}$ for every scope $Z \in \{Z_i^P\}_{i\in[d]}$ (endogenous cells indexed by (v, a), exogenous cells by v alone) **do**
    Play action $a^t \leftarrow \pi^k(s^t)$.
    Observe s^{t+1} and earn reward $r^t \leftarrow r(s^t, a^t)$.
    # Update in-episode visit counters
    for $i \in I^{\text{endo}}$ **do** ▷ count per (v, a)
        $\nu^k_{Z_i^P}(s^t[Z_i^P], a^t) \mathrel{+}= 1, \quad \nu^k_{i,Z_i^P}(s^t[Z_i^P], a^t, s^{t+1}[i]) \mathrel{+}= 1.$
    end for
    for $i \in I^{\text{exo}}$ **do** ▷ count per v, aggregated over action
        $\nu^k_{Z_i^P}(s^t[Z_i^P]) \mathrel{+}= 1, \quad \nu^k_{i,Z_i^P}(s^t[Z_i^P], s^{t+1}[i]) \mathrel{+}= 1.$
    end for
    $t \leftarrow t + 1$.
end while
 # End-of-episode counter update
$N^k \leftarrow N^{k-1} + \nu^k$ (for all counter variables above).
end for

Algorithm 4 Hybrid-DORL: Compute Optimistic Policy π^k

Input: confidence sets $\{\mathcal{W}_i^k\}_{i\in[d]}$ for every factor, the exogenous ones built from action-pooled counts, reward function r.
Construct augmented MDP $\widetilde{M}_{\mathcal{B}}^k = (\widetilde{\mathcal{S}}, \widetilde{\mathcal{A}}, \widetilde{P}_{\mathcal{B}}^k, \tilde{r})$ with the following components (matching §4.3):
State space $\widetilde{\mathcal{S}}$ as in (10), i.e. the *consistent* part of $\mathcal{S} \times \{0, 1, \ldots, L-1\} \times (\mathcal{A} \cup \{\perp\}) \times \prod_{i\in[d]}(\mathcal{S}_i \cup \{\perp\})$ with $L := d + 2$, and a state $\tilde{s} = (s, \ell, a', s'_{\text{buf}})$ stores the original state, sub-step counter, committed action and the partially filled successor buffer.
Action set $\widetilde{\mathcal{A}}(\tilde{s})$, depending on the state only through the counter: $\widetilde{\mathcal{A}}(\tilde{s}) := \mathcal{A}$ at $\ell = 0$, the targets $\mathcal{S}_{\ell-1}$ at $\ell \in \{2, \ldots, L-1\}$, and the pooled set $\widetilde{\mathcal{A}} := \mathcal{A} \cup \left(\bigcup_{i\in[d]} \mathcal{S}_i\right)$ at $\ell = 1$, where every action has the same effect.
Augmented reward $\tilde{r}(\tilde{s}, \tilde{a}) := r(s, \tilde{a})$ at $\ell = 0$ and 0 otherwise.
The transition $\widetilde{P}_{\mathcal{B}}^k$ implements a L-sub-step lift of one original step:
- **Sub-step** $\ell = 0$ **(action selection):** the state advances to $(s, 1, \tilde{a}, \perp)$.
- **Sub-step** $\ell = 1$ **(action commit):** deterministic advance to $(s, 2, a', \perp)$, and no learner choice.
- **Sub-step** $\ell = i+1$**,** $i \in [d]$ **(optimistic sampling):** the planner picks $w^\star \in \mathcal{S}_i$ and factor i's successor w is drawn from the optimistic row (4) at the parent cell $s[Z_i^P]$, with the action argument a' present if $i \in I^{\text{endo}}$ and absent if $i \in I^{\text{exo}}$. The buffer $s'_{\text{buf}}[i] \leftarrow w$ and the counter advances.
- **Commit at** $\ell = L - 1$**:** after factor d is sampled the buffer is complete and the next augmented state is $(s'_{\text{buf}}, 0, \perp, \perp)$.

Compute optimal policy $\tilde{\pi}^k$ of $\widetilde{M}_{\mathcal{B}}^k$ via a factored MDP planning oracle.
Extract: $\pi^k(s) \leftarrow \tilde{\pi}^k\big((s, 0, \perp, \perp)\big)$.

A.2 Planning on the lifted model

The planner is called on $\widetilde{M}_{\mathcal{B}}^k$, and the oracle of §2 requires two things of its input: one action set at every state, and a communicating model. Giving it the pooled $\widetilde{\mathcal{A}}$ everywhere, where each symbol

outside the current action set acts as one fixed element of it, only duplicates actions, so the maximum in (1) and its solutions are unchanged.

$\widetilde{M}^k_{\mathcal{B}}$ is communicating. The second term of (3) makes $\varepsilon^k_i > 0$, hence the slack mass ω^k_i of (11) positive at every cell, and by (4) the target is hit with probability at least ω^k_i. So under any action a' followed by the targets $w^\star_j = s''[j]$, the probability of moving from $(s, 0, \perp, \perp)$ to $(s'', 0, \perp, \perp)$ within L sub-steps is at least

$$\prod_{j\in[d]} \omega^k_j\big(s[Z^P_j], a'\big) \;>\; 0$$

for every s, s'', the intermediate states are visited during these sub-steps, and every state returns to some $(s, 0, \perp, \perp)$ within L sub-steps. Its gain is therefore the unique constant solving (1), which certifies λ^k.

The model is periodic with period L, since the counter advances deterministically under every action. Therefore extended value iteration does not converge on it without an aperiodicity transformation (Jaksch et al., 2010, Remark 9).

A.3 The plug-in lift

Hybrid-DORL-PI (Algorithm 2) changes Hybrid-DORL at the exogenous factors only, each of which has a single parent cell under the hypothesis of Theorem 1. Their widening is switched off, $\mathcal{W}^k_i := 0$ for $i \in I^{\text{exo}}$, and the planner uses the smoothed estimate $\hat{P}^{k,\dagger}_i$ of (5), so the exogenous block of the planning model is the single distribution $\hat{P}^{k,\dagger}_{\text{exo}} := \prod_{i\in I^{\text{exo}}} \hat{P}^{k,\dagger}_i$ and no target is chosen for it. The lift therefore unfolds the endogenous factors only. An original step becomes $L' := d_{\text{endo}} + 2$ sub-steps: $\ell = 0$ selects the action and $\ell = 1$ commits it as in §A.1, $\ell = i + 1$ draws the i-th endogenous factor from the optimistic row (4) at a target chosen by the planner, and the transition out of $\ell = L' - 1$ draws the whole exogenous coordinate from $\hat{P}^{k,\dagger}_{\text{exo}}$ and resets to $\ell = 0$. The action set is $\mathcal{A}$ at $\ell = 0$ and the targets of the $(\ell - 1)$-th endogenous factor at $\ell \geq 2$, as in §A.2.

Writing $y \in \mathcal{S}_{\text{endo}}$ and $x \in \mathcal{S}_{\text{exo}}$ for the two parts of a state s, the model on the original state space is M^k_{ce}. Its action set is $\mathcal{A} \times \mathcal{F}_{\text{endo}}$, where $\mathcal{F}_{\text{endo}}$ is the set of decision rules $f = (f_i)_{i\in I^{\text{endo}}}$, and f_i chooses the target of factor i from the endogenous values already drawn. Its endogenous kernel is $\widetilde{P}^k_{\text{endo}}(y' \mid y, x, a, f) := \prod_{i\in I^{\text{endo}}} P^k_{\mathcal{B},i}(y'[i] \mid s[Z^P_i], a)$ with the targets chosen by f, and its exogenous successor is drawn from $\hat{P}^{k,\dagger}_{\text{exo}}$. Its lift is $\widetilde{M}^k_{\text{ce}}$, on which the oracle is called, and $\pi^k(s) := \tilde{\pi}^k((s, 0, \perp, \perp))$ as before.

The smoothing keeps $\widetilde{M}^k_{\text{ce}}$ communicating: $\hat{P}^{k,\dagger}_i(w) \geq 1/(t_k|\mathcal{S}_i|) > 0$ for every w, so every exogenous value is reached in one original step, and the endogenous targets reach every endogenous value as in §A.2. The oracle of §2 therefore applies, and its gain is the constant λ^k.

B Auxiliary Lemmas

Four facts are used throughout. Lemma 1 is the concentration inequality behind the good event of §C.3. Lemmas 2 and 3 are the two consequences of the doubling rule: the number of episodes, which bounds the boundary term in §C.6, and the growth of the visit counts, which the sums of §C.5 need. Lemma 4 says that the optimistic rows (4) are exactly the vertices of the confidence sets, and Lemmas 5 and 8 use this fact to write the true kernel as a mixture of them.

B.1 Per-coordinate concentration

Lemma 1 (Per-coordinate concentration)**.** *Fix a factor i, a parent value v, an action a (omitted for exogenous factors), and a successor w. With the empirical Bernstein radius ε^k_i of* (3)*, at each fixed sample size n,*

$$\Pr\big[\,|\hat{P}^k_i(w \mid v, a) - P_i(w \mid v, a)| > \varepsilon^k_i(w \mid v, a)\,\big] \;\leq\; 4\,e^{-\tau^k_i},$$

with τ^k_i of §4.2. Consequently the truncated confidence set $\mathcal{W}^k_i$ contains P_i coordinatewise: $\max\{\hat{P}^k_i(w \mid v, a) - \varepsilon^k_i(w \mid v, a), 0\} \leq P_i(w \mid v, a)$ *whenever* $|\hat{P}^k_i(w \mid v, a) - P_i(w \mid v, a)| \leq \varepsilon^k_i(w \mid v, a)$.

Proof. Fix n. For $n \le 1$ the claim is vacuous, because the second term of (3) alone is $\frac{7}{3}\tau_i^k > 1$. So assume $n \ge 2$ and consider the first n visits to (v, a). Over those visits the indicators $X_p := \mathbb{1}\{s'[i] = w\}$ are i.i.d. $\mathrm{Ber}(P_i(w \mid v, a))$, because the factored kernel makes $s'[i]$ depend on the past only through (v, a).

By the empirical Bernstein inequality (Maurer & Pontil, 2009, Theorem 4), applied to these $n \ge 2$ i.i.d. indicators with unbiased sample variance $V_n = \frac{1}{n(n-1)} \sum_{p<q} (X_p - X_q)^2$, with probability at least $1 - \delta'$,

$$\mathbb{E}[X] - \bar{X} \;\le\; \sqrt{\tfrac{2V_n \ln(2/\delta')}{n}} \;+\; \tfrac{7\ln(2/\delta')}{3(n-1)}. \tag{12}$$

Set $\delta' := 2e^{-\tau_i^k}$, so that $\ln(2/\delta') = \tau_i^k$.

The sample variance is bounded by the estimate itself,

$$V_n \;\overset{(a)}{=}\; \tfrac{n}{n-1}\, \hat{P}_i^k(1 - \hat{P}_i^k) \;\overset{(b)}{\le}\; \hat{P}_i^k, \tag{13}$$

where step (a) is the unbiased sample variance of Bernoulli data; step (b) uses $\hat{P}_i^k \ge 1/n$ whenever $\hat{P}_i^k > 0$, so $1 - \hat{P}_i^k \le \frac{n-1}{n}$ and $\frac{n}{n-1}(1 - \hat{P}_i^k) \le 1$.

Applying (12) to $\{X_p\}$ and to $\{1 - X_p\}$ and taking a union over the two sides,

$$\left|\hat{P}_i^k(w \mid v, a) - P_i(w \mid v, a)\right| \;\le\; \sqrt{\tfrac{2V_n\tau_i^k}{n}} + \tfrac{7\tau_i^k}{3(n-1)} \;\le\; \sqrt{\tfrac{2\hat{P}_i^k\tau_i^k}{n}} + \tfrac{7\tau_i^k}{3(n-1)} \;=\; \varepsilon_i^k(w \mid v, a) \tag{14}$$

with probability at least $1 - 2\delta' = 1 - 4e^{-\tau_i^k}$, where the first inequality is (12) at $\ln(2/\delta') = \tau_i^k$, the second substitutes (13), and the equality is the definition (3) of ε_i^k at $n = \max\{N_{Z_i^P}^{k-1}(v, a), 1\}$.

Finally, coordinatewise containment is $\max\{\hat{P}_i^k - \varepsilon_i^k, 0\} \le P_i$, where the first branch holds by (14) on that event and the second by $P_i \ge 0$.

□

B.2 Episode-count bound

Write

$$\Sigma_{\mathrm{cell}} \;:=\; \sum_{i \in I^{\mathrm{endo}}} |\mathcal{S}[Z_i^P]|\,|\mathcal{A}| \;+\; \sum_{i \in I^{\mathrm{exo}}} |\mathcal{S}[Z_i^P]| \tag{15}$$

for the total number of confidence cells. An endogenous factor contributes one cell per parent value and action, and an exogenous factor contributes one per parent value.

Lemma 2 (Episode-count bound)**.** *Under the doubling rule of §4.2, applied to the endogenous (v, a) counts and to the pooled exogenous v counts, the number of episodes over horizon T satisfies*

$$K \;\le\; 1 + \Sigma_{\mathrm{cell}}\big(1 + \log_2 T\big) \;=\; \widetilde{O}\big(\Sigma_{\mathrm{cell}}\big).$$

Proof. Call a *cell* an (i, v, a) triple with $i \in I^{\mathrm{endo}}$ or an (i, v) pair with $i \in I^{\mathrm{exo}}$, so the number of cells is Σ_{cell}. Write D_c for the number of doubling events cell c triggers over the horizon. Then

$$K \;\overset{(a)}{\le}\; 1 + \sum_c D_c \;\overset{(b)}{\le}\; 1 + \Sigma_{\mathrm{cell}}\big(1 + \lfloor \log_2 T \rfloor\big) \;\le\; 1 + \Sigma_{\mathrm{cell}}\big(1 + \log_2 T\big),$$

where step (a) counts the initial episode and observes that every later episode change is triggered by at least one doubling event; step (b) bounds D_c by noting that cell c's pre-episode counts run through $0, 1, 2, 4, \ldots$ up to at most T, since each trigger at least doubles the count or crosses from 0 to 1, which is $1 + \lfloor \log_2 T \rfloor$ values. □

B.3 Episode-length bound

Lemma 3 (Episode length)**.** *Under the doubling rule of §4.2, applied to the endogenous (v, a) counts and to the pooled exogenous v counts, every episode satisfies*

$$t_{k+1} - t_k \;\le\; t_k + \Sigma_{\mathrm{cell}},$$

with Σ_{cell} *as in* (15). *In particular* $t_{k+1} \le 2t_k + \Sigma_{\text{cell}}$*, so episode start times grow geometrically up to the additive constant* Σ_{cell}*. Moreover the visit counts of a single episode never exceed the counts it started with: for every cell,*

$$\nu^k_{Z_i^P}(v,a) \;\le\; \max\{N^{k-1}_{Z_i^P}(v,a),\,1\} \qquad (i \in I^{\text{endo}}), \tag{16}$$

and likewise $\nu^k_{Z_i^P}(v) \le \max\{N^{k-1}_{Z_i^P}(v),1\}$ *for* $i \in I^{\text{exo}}$.

Proof. Fix a factor $i \in [d]$ and write $Z := Z_i^P$. When $i \in I^{\text{exo}}$, the pair (v,a) stands for v alone, since the doubling rule applies to the pooled exogenous counts too. Then

$$\begin{aligned} t_{k+1} - t_k \;&\overset{(a)}{=}\; \sum_{(v,a)} \nu^k_Z(v,a) \;\overset{(b)}{\le}\; \sum_{(v,a)} \max\{N^{k-1}_Z(v,a),\,1\} \\ &\le\; \sum_{(v,a)} N^{k-1}_Z(v,a) \;+\; |\mathcal{S}[Z]|\,|\mathcal{A}| \;\overset{(c)}{\le}\; t_k + \Sigma_{\text{cell}}, \end{aligned} \tag{17}$$

where step (a) holds because each time step of episode k increments ν^k_Z at exactly one pair; step (b) is the while-condition of Algorithm 3, checked before each increment and at the current cell of every scope, so no cell exceeds its pre-episode count during the episode, which also gives (16); step (c) uses that the pre-episode counts of a single scope sum to at most t_k, and $|\mathcal{S}[Z]|\,|\mathcal{A}| \le \Sigma_{\text{cell}}$. □

B.4 Validity of the lifted optimistic transitions

At a cell (v,a') of factor i, where the action argument is absent when $i \in I^{\text{exo}}$, the *one-sided confidence polytope* is

$$\mathcal{B}^k_i(v,a') \;:=\; \Big\{ Q \in \mathbb{R}^{|\mathcal{S}_i|} \;:\; Q(w') \ge \hat{P}^k_i(w' \mid v,a') - \mathcal{W}^k_i(w' \mid v,a')\;\forall w',\; \textstyle\sum_{w'} Q(w') = 1 \Big\}, \tag{18}$$

the kernels that give every successor at least the estimate less its slack.

Lemma 4 (Validity of the lifted optimistic transitions)**.** *Fix a factor* $i \in [d]$*, a parent value* $v = s[Z_i^P]$*, and an action* a'*. For every learner target* $w \in \mathcal{S}_i$*, the distribution* $P^k_{\mathcal{B},i}(\cdot \mid v,a')$ *defined by Eq.* (4) *satisfies (i)* $P^k_{\mathcal{B},i}(w' \mid v,a') \ge 0$ *for all* $w' \in \mathcal{S}_i$*; (ii)* $\sum_{w'} P^k_{\mathcal{B},i}(w' \mid v,a') = 1$*; and (iii)* $P^k_{\mathcal{B},i}(\cdot \mid v,a') \in \mathcal{B}^k_i(v,a')$*. Moreover, ranging* w *over* $\mathcal{S}_i$ *realizes every vertex of* $\mathcal{B}^k_i(v,a')$ *.*

Proof. The action argument is absent when $i \in I^{\text{exo}}$, with $N^{k-1}_{Z_i^P}(v)$ in place of $N^{k-1}_{Z_i^P}(v,a)$. By (4),

$$P^k_{\mathcal{B},i}(w' \mid v,a') \;=\; \begin{cases} \hat{P}^k_i(w' \mid v,a') - \mathcal{W}^k_i(w' \mid v,a'), & w' \ne w, \\ \hat{P}^k_i(w \mid v,a') - \mathcal{W}^k_i(w \mid v,a') + \omega^k_i(v,a'), & w' = w. \end{cases} \tag{19}$$

Part (i). For every w',

$$\hat{P}^k_i(w') - \mathcal{W}^k_i(w') \;=\; \max\{\hat{P}^k_i(w') - \varepsilon^k_i(w'),\,0\} \;\ge\; 0,$$

by the truncation $\mathcal{W}^k_i = \min\{\varepsilon^k_i, \hat{P}^k_i\}$. Both branches of (19) are therefore non-negative, and the branch $w' = w$ adds $\omega^k_i(v,a') \ge 0$.

Part (ii). Summing (19) over w',

$$\begin{aligned} \sum_{w'} P^k_{\mathcal{B},i}(w' \mid v,a') \;&=\; \sum_{w'} \hat{P}^k_i(w' \mid v,a') \;-\; \sum_{w'} \mathcal{W}^k_i(w' \mid v,a') \;+\; \omega^k_i(v,a') \\ &=\; \sum_{w'} \hat{P}^k_i(w' \mid v,a') \;=\; 1, \end{aligned} \tag{20}$$

where the middle sum cancels against $\omega^k_i(v,a')$ by its definition. The last equality holds at a visited pair because the counts sum to $N^{k-1}_{Z_i^P}(v,a)$, and at an unvisited pair by the uniform convention of §4.2. Without that convention the ratio-of-counts formula returns the all-zero vector at $N^{k-1}_{Z_i^P}(v,a) = 0$, and parts (ii) and (iii) fail.

Part (iii). Both branches of (19) satisfy

$$P^k_{\mathcal{B},i}(w' \mid v, a') \;\geq\; \hat{P}^k_i(w' \mid v, a') - \mathcal{W}^k_i(w' \mid v, a'),$$

with equality on the branch $w' \neq w$ and $\omega^k_i(v, a') \geq 0$ added on the branch $w' = w$. With (20) this is membership in $\mathcal{B}^k_i(v, a')$ as defined in (18).

Vertices. Writing $q := \hat{P}^k_i(\cdot \mid v, a') - \mathcal{W}^k_i(\cdot \mid v, a')$, which is nonnegative with total mass $1 - \omega^k_i(v, a')$, membership in (18) says exactly that $Q - q$ is nonnegative with total mass $\omega^k_i(v, a')$, that is, $Q = q + \omega^k_i(v, a')\,\alpha$ for a probability vector α on $\mathcal{S}_i$. So $\mathcal{B}^k_i(v, a')$ is the affine image of the probability simplex. Its vertices are the images $q + \omega^k_i(v, a')\,e_w$ of the unit vectors, which are the rows (19) as w ranges over $\mathcal{S}_i$, and they coincide when $\omega^k_i(v, a') = 0$. Hence $\mathcal{B}^k_i(v, a')$ is their convex hull. □

C Proof of Proposition 1 (Upper Bound)

The proofs use the models below. A tilde marks the sub-step state space of §4.3; a subscript $\mathcal{B}$ marks kernels taken from the confidence sets $\mathcal{B}^k_i$ of episode k, the optimistic rows (4); ce marks the plug-in variant of Theorem 1, whose exogenous block is the estimate itself; and L' marks the lift over the endogenous factors only. A prime is reserved for successor states.

	kernels	state space	defined
M	true, P	$\mathcal{S}$	§3
$\widetilde{M}$	true, $\widetilde{P}$	sub-step, $L = d + 2$	§C.1
$\widetilde{M}^k_{\mathcal{B}}$	rows (4), the target chosen at each sub-step, $\widetilde{P}^k_{\mathcal{B}}$	sub-step, L	§4.3, App. A
M^k_{ce}	endogenous rows (4), exogenous $\hat{P}^{k,\dagger}_{\mathrm{exo}}$ of (5)	$\mathcal{S}$	App. A.3
$\widetilde{M}^k_{\mathrm{ce}}$	the same, $\widetilde{P}^k_{\mathrm{ce}}$	sub-step, $L' = d_{\mathrm{endo}} + 2$	App. A.3
$\widetilde{M}_{L'}$	true	sub-step, L'	Appendix D

The proof has four parts. §C.1–§C.3 set up the tools: the true-kernel lift $\widetilde{M}$ against which the realized trajectory is measured, the bias span of the planned model (Lemma 5), and the good event of probability at least $1 - \delta$. §C.4 splits the regret into an optimism term, non-positive by Lemma 6, and a deviation term, and reduces the latter to per-factor sums of confidence radii. §C.5 evaluates those sums and §C.6 combines the terms.

C.1 Trajectory Mapping

The proof measures the realized trajectory against a reference MDP $\widetilde{M}$, which shares the augmented state and action spaces of $\widetilde{M}^k_{\mathcal{B}}$ and its sub-step timeline (§A.1), and differs in exactly one respect: every factor, endogenous or exogenous, moves under its *true* kernel $P_j(\cdot \mid s[Z^P_j], a')$ rather than under the optimistic row (4), so the augmented action at $\ell \geq 1$ has no effect, and its reward is the $\tilde{r}$ of §A.1. In the sub-step space the planned kernel and the true kernel differ at one factor per sub-step, so the deviation between them is counted one factor at a time, and the bound is therefore a sum over factors.

A trajectory of M maps into $\widetilde{M}$ sub-step by sub-step: given $(s^t, a^t)_{t \leq T}$ in M, the image is

$$s^{t,\ell} \;=\; \big(s^t,\; \ell,\; a',\; \beta^{(\ell)}\big), \qquad a' = \begin{cases} a^t, & \ell \geq 1, \\ \perp, & \ell = 0, \end{cases} \qquad \beta^{(\ell)}[j] = \begin{cases} s^{t+1}[j], & j \leq \ell - 2, \\ \perp, & j \geq \ell - 1, \end{cases}$$

with $a^{t,0} = a^t$ and $a^{t,\ell}$ arbitrary for $\ell \geq 1$. The draw of j is realized on the transition *out of* $\ell = j + 1$, so the state there cannot already contain it.

For a policy π in M we write π also for its embedding in $\widetilde{M}$, so $\pi(s^{t,0}) = \pi(s^t)$ and $\pi(s^{t,\ell})$ is arbitrary for $\ell \geq 1$. Its $\ell = 0$ states follow M under π and it collects $r(s^t, \pi(s^t))$ once per L sub-steps, so

$$\lambda\big(\widetilde{M}, \pi, (s, 0, \perp, \perp)\big) \;=\; \lambda(M, \pi, s)/L. \tag{21}$$

C.2 BIAS SPAN OF THE OPTIMISTIC MODEL

Optimism only adds transitions, so no state of the optimistic model is harder to reach than in M, and the lift multiplies hitting times by L while dividing the gain by L, so the bias, at most diameter times gain, stays at $D(M)$. The bound is stated for every solution of the optimality equation, since the oracle of §2 does not say which one it returns.

Let $\mathcal{E}^k$ be the event that $|\hat{P}_i^k - P_i| \le \varepsilon_i^k$ holds at every factor, cell and successor of episode k, written out in §C.3; on it $P_i \in \mathcal{B}_i^k$ for every i by Lemma 1.

Lemma 5 (Bias span of the optimistic model)**.** *On $\mathcal{E}^k$, every (π, λ, h) satisfying* (1) *on $\widetilde{M}_{\mathcal{B}}^k$ has*

$$\operatorname{sp}(h) \;\le\; D(M) + 1. \tag{22}$$

Proof. Fix $s, s'' \in \mathcal{S}$ and let π be a policy of M with $\mathbb{E}[T(s'' \mid M, \pi, s)] \le D(M)$. On $\mathcal{E}^k$ every true row $P_i(\cdot \mid v, a)$ lies in $\mathcal{B}_i^k(v, a)$ (Lemma 1), the convex hull of the target rows (Lemma 4), so it is a convex combination of them. Let σ be the policy of $\widetilde{M}_{\mathcal{B}}^k$ that plays $\pi(s)$ at $(s, 0, \bot, \bot)$ and at sub-step $i+1$ draws the target of factor i from those weights at the cell $(s[Z_i^P], \pi(s))$. Factor i's draw then has law $P_i(\cdot \mid s[Z_i^P], \pi(s))$ and the d draws are conditionally independent, so the $\ell = 0$ states under σ follow M under π, and the first-passage time T from $(s, 0, \bot, \bot)$ to $(s'', 0, \bot, \bot)$ satisfies $\mathbb{E}[T] = L\,\mathbb{E}[T(s'' \mid M, \pi, s)] \le L\,D(M)$. The gain satisfies $\lambda \le 1/L$, because reward is earned at $\ell = 0$ only and is at most 1.

These two facts bound the bias at $\ell = 0$ by diameter times gain. Dropping the maximum in (1) gives, for every $\tilde{a} \in \widetilde{\mathcal{A}}(\tilde{s})$,

$$h(\tilde{s}) \;\ge\; \tilde{r}(\tilde{s}, \tilde{a}) - \lambda + \sum_{\tilde{s}'} \widetilde{P}_{\mathcal{B}}^k(\tilde{s}' \mid \tilde{s}, \tilde{a})\, h(\tilde{s}') \;\ge\; -\lambda + \mathbb{E}_{\tilde{a}}\big[h(\tilde{s}')\big],$$

using $\tilde{r} \ge 0$, and averaging over the randomized action of σ keeps it. Iterating along σ and stopping at T, legitimate because h is bounded and $\mathbb{E}[T] < \infty$,

$$h(s, 0, \bot, \bot) - h(s'', 0, \bot, \bot) \;\ge\; -\lambda\,\mathbb{E}[T] \;\ge\; -D(M),$$

and exchanging s with s'' gives $\operatorname{sp}(h|_{\ell=0}) \le D(M)$.

At $\ell \ge 1$ the reward vanishes and the counter advances deterministically, so (1) reads $h(\cdot, \ell, \cdot) = -\lambda + \max_{\tilde{a}} \mathbb{E}_{\tilde{a}}[h(\cdot, \ell+1, \cdot)]$ for $\ell \le L-2$ and $h(\cdot, L-1, \cdot) = -\lambda + \max_{\tilde{a}} \mathbb{E}_{\tilde{a}}[h|_{\ell=0}]$. A maximum of expectations of $h|_{\ell=0}$ lies in $[\min h|_{\ell=0}, \max h|_{\ell=0}]$, so downward induction on ℓ gives, for every $\ell \ge 1$,

$$\min h|_{\ell=0} - (L-\ell)\lambda \;\le\; h(\cdot, \ell, \cdot) \;\le\; \max h|_{\ell=0} - (L-\ell)\lambda,$$

so

$$\operatorname{sp}(h) \;\le\; \operatorname{sp}(h|_{\ell=0}) + (L-1)\lambda \;\le\; D(M) + \tfrac{L-1}{L} \;<\; D(M) + 1.$$

□

We write

$$H \;:=\; D(M) + 1 \tag{23}$$

for the span bound of (22), so $\operatorname{sp}(h^k) \le H$ on $\mathcal{E}^k$ for the oracle's returned bias, and $H \le 2D(M)$ since $D(M) \ge 1$.

C.3 FAILURE EVENTS

The analysis runs on the event that every confidence set contains the truth and that the martingale part of the deviation term concentrates. This subsection defines the three failure events and bounds each by $\delta/5$. The reward r is deterministic and known (§3), so no reward event is needed. We use the Bernstein radii $\varepsilon_i^k(w \mid v, a)$ and $\varepsilon_i^k(w \mid v)$ of (3) and the log factors τ_i^k of §4.2.

- F_{endo}^P is the event that some empirical estimate of an endogenous transition kernel is far from the true kernel: there exist k, $i \in I^{\text{endo}}$, $v \in \mathcal{S}[Z_i^P]$, $a \in \mathcal{A}$, and $w \in \mathcal{S}_i$ such that

$$|\hat{P}_i^k(w \mid v, a) - P_i(w \mid v, a)| > \varepsilon_i^k(w \mid v, a).$$

- F^P_{exo} is the event that some empirical estimate of an exogenous transition kernel is far from the true kernel: there exist k, $i \in I^{\text{exo}}$, $v \in \mathcal{S}[Z_i^P]$, and $w \in \mathcal{S}_i$ such that

$$|\hat{P}_i^k(w \mid v) - P_i(w \mid v)| > \varepsilon_i^k(w \mid v).$$

- F_{Az} is the event that the augmented-trajectory martingale concentration is violated (two-sided):

$$\left| \sum_{k,t,\ell} \left[\sum_{\tilde{s}'} \widetilde{P}(\tilde{s}' \mid s^{t,\ell}, \tilde{\pi}^k(s^{t,\ell}))\, h^k(\tilde{s}') - h^k(\tilde{s}^{t,\ell+1}) \right] \right| > H\sqrt{2\,TL\log(10/\delta)},$$

where h^k is the oracle's returned bias and $\tilde{s}^{t,\ell+1}$ the next augmented state. By Lemma 5, $\mathrm{sp}(h^k) \le D(M) + 1 = H$ whichever solution the oracle returns.

We show $\Pr[F^P_{\text{endo}}], \Pr[F^P_{\text{exo}}], \Pr[F_{Az}] \le \delta/5$, whose union is at most $3\delta/5 < \delta$, and write $\mathcal{E} := (F^P_{\text{endo}} \cup F^P_{\text{exo}})^c$.

Write $\mathcal{T}^{\text{endo}} := \{(i, v, a, w) : i \in I^{\text{endo}},\, v \in \mathcal{S}[Z_i^P],\, a \in \mathcal{A},\, w \in \mathcal{S}_i\}$ for the endogenous tuple set. Then

$$\begin{aligned}
\Pr[F^P_{\text{endo}}] &= \Pr\Big[\, \exists\, k,\, (i,v,a,w) \in \mathcal{T}^{\text{endo}} :\ |\hat{P}_i^k(w \mid v, a) - P_i(w \mid v, a)| > \varepsilon_i^k(w \mid v, a) \,\Big] \\
&\le \sum_{t_k \in [T]} \sum_{(i,v,a,w) \in \mathcal{T}^{\text{endo}}} \Pr\big[\, |\hat{P}_i^k - P_i| > \varepsilon_i^k \,\big] \\
&\overset{(a)}{\le} \sum_{k=1}^{K} \sum_{n=1}^{t_k} \sum_{(i,v,a,w) \in \mathcal{T}^{\text{endo}}} 4\, e^{-\tau_i^k} \\
&\overset{(b)}{\le} \sum_{k=1}^{K} t_k \cdot d_{\text{endo}}\, \Lambda\, |\mathcal{A}|\, W \cdot \frac{4\,\delta}{40\, d_{\text{endo}}\, W\, \Lambda\, |\mathcal{A}|\, t_k^3} \;=\; \sum_{k=1}^{K} \frac{\delta}{10\, t_k^2} \;\overset{(c)}{\le}\; \frac{\delta}{10} \sum_{t=1}^{\infty} \frac{1}{t^2} \;\le\; \frac{\delta}{5},
\end{aligned} \tag{24}$$

where step (a) is Lemma 1, applied in each direction of $|\hat{P}_i^k - P_i|$ for $4e^{-\tau_i^k}$ per tuple, where the visit count is random, so the union also runs over the sample size $n = 1, \ldots, t_k$; step (b) substitutes τ_i^k from §4.2 and $|\mathcal{T}^{\text{endo}}| \le d_{\text{endo}}\Lambda|\mathcal{A}|W$, and the cube t_k^3 absorbs the extra t_k from that union; step (c) uses that the $\{t_k\}$ are distinct positive integers, so $\sum_k 1/t_k^2 \le \zeta(2) < 2$.

The exogenous chain is parallel, with $\mathcal{T}^{\text{exo}} := \{(i, v, w) : i \in I^{\text{exo}},\, v \in \mathcal{S}[Z_i^P],\, w \in \mathcal{S}_i\}$ (no action, since exogenous transitions are action-free):

$$\begin{aligned}
\Pr[F^P_{\text{exo}}] &= \Pr\Big[\, \exists\, k,\, (i,v,w) \in \mathcal{T}^{\text{exo}} :\ |\hat{P}_i^k(w \mid v) - P_i(w \mid v)| > \varepsilon_i^k(w \mid v) \,\Big] \\
&\le \sum_{k=1}^{K} \sum_{n=1}^{t_k} \sum_{(i,v,w) \in \mathcal{T}^{\text{exo}}} 4\, e^{-\tau_i^k} \\
&\le \sum_{k=1}^{K} t_k \cdot d_{\text{exo}}\, \Lambda\, W \cdot \frac{4\,\delta}{40\, d_{\text{exo}}\, W\, \Lambda\, t_k^3} \;=\; \sum_{k=1}^{K} \frac{\delta}{10\, t_k^2} \;\le\; \frac{\delta}{10} \sum_{t \ge 1} \frac{1}{t^2} \;\le\; \frac{\delta}{5},
\end{aligned} \tag{25}$$

with the justifications of (24) under three substitutions: no action argument, the visit count aggregated as $N^{k-1}_{Z_i^P}(v) := \sum_a N^{k-1}_{Z_i^P}(v, a)$, and d_{exo} for $d_{\text{endo}}|\mathcal{A}|$ in τ_i^k.

For $\Pr[F_{Az}]$, write $\mathcal{F}_{t,\ell}$ for the augmented trajectory through sub-step (t, ℓ), and define

$$\Xi_{t,\ell} \;:=\; \mathbb{E}[h^k(\tilde{s}^{t,\ell+1}) \mid \mathcal{F}_{t,\ell}] - h^k(\tilde{s}^{t,\ell+1}),$$

the planned bias at the next sub-step minus its conditional mean given the trajectory so far, so that $\mathbb{E}[\Xi_{t,\ell} \mid \mathcal{F}_{t,\ell}] = 0$. The bound $|\Xi_{t,\ell}| \le \mathrm{sp}(h^k) \le H$ of (23) holds only on $\mathcal{E}$, so we stop the sum at

the first sub-step where it can fail. Written out, the per-episode event $\mathcal{E}^k$ of §C.2 is

$$\mathcal{E}^k \;:=\; \big\{\forall(i,v,a,w)\in\mathcal{T}^{\mathrm{endo}}:\ |\hat{P}_i^k(w\mid v,a)-P_i(w\mid v,a)|\le\varepsilon_i^k(w\mid v,a)\big\}$$
$$\cap\ \big\{\forall(i,v,w)\in\mathcal{T}^{\mathrm{exo}}:\ |\hat{P}_i^k(w\mid v)-P_i(w\mid v)|\le\varepsilon_i^k(w\mid v)\big\},$$

which is decided by the data collected before episode k starts, since $\hat{P}^k$ and ε^k are, and satisfies $\mathcal{E}=\bigcap_{k\ge1}\mathcal{E}^k$. Order sub-steps lexicographically by (t,ℓ) and let τ be the first at which $\mathcal{E}^{k(t)}$ has failed, where $k(t)$ is the episode containing t, and $\tau=\infty$ if it never fails. Whether $\tau>(t,\ell)$ is decided by the events $\mathcal{E}^{k(t')}$ with $t'\le t$, each of them by data collected before $t_{k(t')}\le t'$, so it is a function of $\mathcal{F}_{t,\ell}$.

Put $\widetilde{\Xi}_{t,\ell}:=\Xi_{t,\ell}\cdot\mathbb{1}\{(t,\ell)<\tau\}$. The indicator is a function of $\mathcal{F}_{t,\ell}$ and comes out of the conditional expectation, so $\mathbb{E}[\widetilde{\Xi}_{t,\ell}\mid\mathcal{F}_{t,\ell}]=\mathbb{1}\{(t,\ell)<\tau\}\,\mathbb{E}[\Xi_{t,\ell}\mid\mathcal{F}_{t,\ell}]=0$, and $|\widetilde{\Xi}_{t,\ell}|\le H$ holds unconditionally: on $\{(t,\ell)<\tau\}$ Lemma 5 gives $|\Xi_{t,\ell}|\le\mathrm{sp}(h^k)\le H$, and on $\{(t,\ell)\ge\tau\}$ the increment is zero. The Azuma–Hoeffding inequality (Azuma, 1967) for a sum of $n:=TL$ terms, each bounded by H and of conditional mean zero given the trajectory before it, gives

$$\Pr\Big[\,\big|\textstyle\sum_{k,t,\ell}\widetilde{\Xi}_{t,\ell}\big|>H\sqrt{2n\log(10/\delta)}\,\Big]\;\le\;2\exp\!\Big(-\tfrac{2nH^2\log(10/\delta)}{2nH^2}\Big)\;=\;\tfrac{\delta}{5}. \tag{26}$$

On $\mathcal{E}$, $\tau=\infty$ and $\widetilde{\Xi}\equiv\Xi$, so (26) bounds $\Pr[F_{Az}\cap\mathcal{E}]$, the quantity the union below uses: $\Pr[F^P_{\mathrm{endo}}\cup F^P_{\mathrm{exo}}\cup(F_{Az}\cap\mathcal{E})]\le\frac{3\delta}{5}<\delta$, with $\mathcal{E}^c=F^P_{\mathrm{endo}}\cup F^P_{\mathrm{exo}}$ already inside.

C.4 Regret Decomposition

The regret splits into an optimism term, which the planned gain makes non-positive, and a deviation term, whose endogenous and exogenous parts give the two summands of Proposition 1. The first needs the planned gain to dominate the true one.

Lemma 6 (Optimism). *On $\mathcal{E}^k$,*

$$\lambda^\star(\widetilde{M}^k_{\mathcal{B}})\;\ge\;\lambda^\star(M)/L. \tag{27}$$

Proof. Let (λ^k,h) solve (1) on $\widetilde{M}^k_{\mathcal{B}}$ (Appendix A.2). We show that at every $\tilde{s}$ and every action $\tilde{a}$ available there,

$$\lambda^k+h(\tilde{s})\;\ge\;\tilde{r}(\tilde{s},\tilde{a})+\sum_{\tilde{s}'}\widetilde{P}(\tilde{s}'\mid\tilde{s},\tilde{a})\,h(\tilde{s}'), \tag{28}$$

where the right-hand side is that of the optimality equation of $\widetilde{M}$, which has the same state space. At $\ell\in\{0,1\}$ the two kernels coincide, so (28) is (1) on $\widetilde{M}^k_{\mathcal{B}}$ with the maximum dropped. At $\ell=i+1$ write $\tilde{s}'_w$ for the successor of $\tilde{s}$ with $s'_{\mathrm{buf}}[i]=w$, and omit the cell $(s[Z_i^P],a')$, with the action absent for an exogenous factor, from the arguments of P_i, $\hat{P}_i^k$, $\mathcal{W}_i^k$ and ω_i^k. On $\mathcal{E}^k$, $P_i\ge\hat{P}_i^k-\mathcal{W}_i^k$ coordinatewise (Lemma 1), and $\sum_w\big(P_i(w)-\hat{P}_i^k(w)+\mathcal{W}_i^k(w)\big)=\omega_i^k$, so

$$\begin{aligned}\lambda^k+h(\tilde{s})\;&\overset{(a)}{=}\;\max_{w^\star\in\mathcal{S}_i}\Big[\sum_w\big(\hat{P}_i^k(w)-\mathcal{W}_i^k(w)\big)\,h(\tilde{s}'_w)+\omega_i^k\,h(\tilde{s}'_{w^\star})\Big]\\&\overset{(b)}{\ge}\;\sum_w\big(\hat{P}_i^k(w)-\mathcal{W}_i^k(w)\big)\,h(\tilde{s}'_w)+\sum_w\big(P_i(w)-\hat{P}_i^k(w)+\mathcal{W}_i^k(w)\big)\,h(\tilde{s}'_w)\\&=\;\sum_w P_i(w)\,h(\tilde{s}'_w),\end{aligned} \tag{29}$$

where step (a) is (1) with the row (4) split into its two terms and zero reward at $\ell\ge1$; step (b) takes $w^\star$ at a maximum of $w\mapsto h(\tilde{s}'_w)$ and spreads the same mass ω_i^k over coordinates where h is no larger. The last expression is the right-hand side of (28), because $\widetilde{M}$ ignores $\tilde{a}$ at $\ell\ge1$.

Fix any policy π of $\widetilde{M}$ and any initial state $\tilde{s}^1$. Summing (28) at $(\tilde{s}^u,\tilde{a}^u)$ over $u\le n$ and taking expectations,

$$n\lambda^k+h(\tilde{s}^1)\;\ge\;\mathbb{E}_\pi\Big[\sum_{u\le n}\tilde{r}(\tilde{s}^u,\tilde{a}^u)\Big]+\mathbb{E}_\pi\big[h(\tilde{s}^{n+1})\big]\;\Longrightarrow\;\lambda^k\;\ge\;\frac{1}{n}\,\mathbb{E}_\pi\Big[\sum_{u\le n}\tilde{r}(\tilde{s}^u,\tilde{a}^u)\Big]-\frac{\mathrm{sp}(h)}{n},$$

and $n \to \infty$ gives $\lambda^k \geq \lambda(\widetilde{M}, \pi, \tilde{s}^1)$ (Puterman, 1994, Thm. 8.4.1). Taking π the embedding of $\pi^\star(M)$ and $\tilde{s}^1 = (s, 0, \perp, \perp)$, whose gain is $\lambda^\star(M)/L$ by (21), gives (27), since $\lambda^k = \lambda^\star(\widetilde{M}^k_{\mathcal{B}})$ by Appendix A.2. □

Using $r^t = r(s^t, a^t)$, the regret decomposes as

$$
\begin{aligned}
\mathrm{Reg}_T(M) &= \sum_{t=1}^{T} \big(\lambda^\star(M) - r(s^t, a^t)\big) \\
&= \sum_{t=1}^{T} \left(\frac{\lambda^\star(M)}{L} - r(s^t, a^t)\right) + \sum_{t=1}^{T} \sum_{\ell=1}^{L-1} \left(\frac{\lambda^\star(M)}{L} - 0\right) \\
&\overset{(a)}{=} \sum_{t=1}^{T} \sum_{\ell=0}^{L-1} \big(\lambda^\star(M)/L - \tilde{r}(s^{t,\ell}, a^{t,\ell})\big) \\
&\overset{(b)}{=} \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{\ell=0}^{L-1} \big(\lambda^\star(M)/L - \tilde{r}(s^{t,\ell}, \pi^k(s^{t,\ell}))\big) \\
&\overset{(c)}{=} \underbrace{\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{\ell=0}^{L-1} \big(\lambda^\star(M)/L - \lambda^k\big)}_{=:\text{ Term (I): optimism}} + \underbrace{\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{\ell=0}^{L-1} \big(\lambda^k - \tilde{r}(s^{t,\ell}, \pi^k(s^{t,\ell}))\big)}_{=:\text{ Term (II): deviation}},
\end{aligned} \tag{30}
$$

where step (a) is the definition of $\tilde{r}$; step (b) splits the sum into episodes, where the algorithm plays $a^t = \pi^k(s^t)$; step (c) adds and subtracts the planned gain $\lambda^k = \lambda^\star(\widetilde{M}^k_{\mathcal{B}})$ (Appendix A.2). Term (I) is non-positive on $\mathcal{E}^k$ by Lemma 6, at every k, so on $\mathcal{E} = \bigcap_k \mathcal{E}^k$

$$
\mathrm{Reg}_T(M) \;\leq\; \text{Term (II)}. \tag{31}
$$

It remains to bound Term (II). Let $\tilde{\pi}^k$ be the policy returned by the planning oracle on $\widetilde{M}^k_{\mathcal{B}}$ and h^k the bias returned alongside it. Term (II) is evaluated at $\tilde{\pi}^k$, whose augmented action at $\ell \geq 2$ is a target $w^\star \in \mathcal{S}_{\ell-1}$ that moves the state, and not at the embedding of π^k, which moves nothing there. The gain of $\widetilde{M}^k_{\mathcal{B}}$ is the constant λ^k since the model is communicating (Appendix A.2). Hence (1) holds there with $\tilde{\pi}^k$ attaining the maximum, and rearranges, at every $\tilde{s}$, to

$$
\lambda^k - \tilde{r}\big(\tilde{s}, \tilde{\pi}^k(\tilde{s})\big) \;=\; \sum_{\tilde{s}'} \widetilde{P}^k_{\mathcal{B}}\big(\tilde{s}' \mid \tilde{s}, \tilde{\pi}^k(\tilde{s})\big) h^k(\tilde{s}') - h^k(\tilde{s}).
$$

Substituting this at each sub-step of each episode gives the first display below. Write $\rho := 3 \log \frac{40\, d_{\mathrm{endo}}\, W\, \Lambda\, |\mathcal{A}|\, T}{\delta}$ and $\rho_{\mathrm{exo}} := 3 \log \frac{40\, d_{\mathrm{exo}}\, W\, \Lambda\, T}{\delta}$ for the logarithmic factors of §4.2. The coefficient 3 comes from the exponent t_k^3: with $C := 40\, d_{\mathrm{endo}} W \Lambda |\mathcal{A}|/\delta \geq 1$, $\tau_i^k = \log(C t_k^3) \leq \log(C T^3) \leq 3 \log(CT) = \rho$ at $i \in I^{\mathrm{endo}}$ whenever $t_k \leq T$, and likewise ρ_{exo} bounds τ_i^k at

$i \in I^{\mathrm{exo}}$. Then

$$
\begin{aligned}
\text{Term (II)} &= \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{\ell=0}^{L-1} \left[\sum_{\tilde{s}' \in \widetilde{\mathcal{S}}} \widetilde{P}^k_{\mathcal{B}}(\tilde{s}' \mid s^{t,\ell}, \tilde{\pi}^k(s^{t,\ell})) h^k(\tilde{s}') - h^k(s^{t,\ell}) \right] \\
&\overset{(a)}{=} \underbrace{\sum_{k,t,\ell} \sum_{\tilde{s}'} \big[\widetilde{P}^k_{\mathcal{B}} - \widetilde{P}\big](\tilde{s}' \mid s^{t,\ell}, \tilde{\pi}^k(s^{t,\ell}))\, h^k(\tilde{s}')}_{=:\text{Term (II.A): kernel gap}} \\
&\quad + \underbrace{\sum_{k,t,\ell} \left[\sum_{\tilde{s}'} \widetilde{P}(\tilde{s}' \mid s^{t,\ell}, \tilde{\pi}^k(s^{t,\ell})) h^k(\tilde{s}') - h^k(s^{t,\ell}) \right]}_{=:\text{Term (II.B): martingale term}} \\
&\overset{(b)}{\le} \tfrac{3}{2} H \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{i \in I^{\mathrm{endo}}} \sum_{w \in \mathcal{S}_i} \varepsilon_i^k\big(w \mid s^t[Z_i^P],\, a^t\big) \\
&\quad + \tfrac{3}{2} H \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{i \in I^{\mathrm{exo}}} \sum_{w \in \mathcal{S}_i} \varepsilon_i^k\big(w \mid s^t[Z_i^P]\big) \\
&\quad + \text{Term (II.B)} \\
&\overset{(c)}{\le} \tfrac{3}{2} H \sum_{k=1}^{K} \sum_{i \in I^{\mathrm{endo}}} \sum_{v \in \mathcal{S}[Z_i^P]} \sum_{a \in \mathcal{A}} \nu^k_{Z_i^P}(v,a) \\
&\qquad \cdot \left(\sqrt{\frac{2\,|\mathcal{S}_i|\,\rho}{\max\{N^{k-1}_{Z_i^P}(v,a), 1\}}} + \frac{7\,|\mathcal{S}_i|\,\rho}{3\max\{N^{k-1}_{Z_i^P}(v,a) - 1, 1\}} \right) \\
&\qquad + \tfrac{3}{2} H \sum_{k=1}^{K} \sum_{i \in I^{\mathrm{exo}}} \sum_{v \in \mathcal{S}[Z_i^P]} \nu^k_{Z_i^P}(v) \\
&\qquad \cdot \left(\sqrt{\frac{2\,|\mathcal{S}_i|\,\rho_{\mathrm{exo}}}{\max\{N^{k-1}_{Z_i^P}(v), 1\}}} + \frac{7\,|\mathcal{S}_i|\,\rho_{\mathrm{exo}}}{3\max\{N^{k-1}_{Z_i^P}(v) - 1, 1\}} \right) \\
&\qquad + KH + H\sqrt{2\,TL\log(10/\delta)} \\
&\overset{(d)}{\le} \widetilde{O}\Big(\sum_{i \in I^{\mathrm{endo}}} D\sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T} \Big) \\
&\quad + \widetilde{O}\Big(\sum_{i \in I^{\mathrm{exo}}} D\sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,T} \Big),
\end{aligned}
\tag{32}
$$

where step (a) adds and subtracts $\sum_{\tilde{s}'} \widetilde{P}(\tilde{s}' \mid s^{t,\ell}, \tilde{\pi}^k(s^{t,\ell}))h^k(\tilde{s}')$; step (b) uses that at $\ell = j+1$ exactly one factor, j, is drawn, so $[\widetilde{P}^k_{\mathcal{B}} - \widetilde{P}](\cdot \mid s^{t,\ell}, \cdot)$ is supported on that coordinate, while at $\ell \in \{0,1\}$ the two kernels agree, giving one term per factor per *original* step rather than per sub-step; step (c) substitutes the Bernstein radius (3) and the bound on Term (II.B) derived below; step (d) is the doubling and Cauchy–Schwarz argument of Appendix C.5.

The factor $\frac{3}{2}H$ in step (b) comes from two applications of the span. For a vector q whose entries sum to zero, $|\langle q, h^k\rangle| = |\langle q, h^k - c\rangle| \le \frac{1}{2}\mathrm{sp}(h^k)\|q\|_1$ with c the midpoint of the range of h^k, and on $\mathcal{E}$ we have $\|P^k_{\mathcal{B},i} - P_i\|_1 \le 2\omega_i^k + \sum_w \varepsilon_i^k \le 3\sum_w \varepsilon_i^k$ together with $\mathrm{sp}(h^k) \le H$ by (23).

Term (II.B) is the martingale of §C.3 plus a telescoping sum. Adding and subtracting $h^k(\tilde{s}^{t,\ell+1})$, the realized next state,

$$\text{Term (II.B)} \;=\; \sum_{k,t,\ell} \Xi_{t,\ell} \;+\; \sum_{k=1}^{K} \sum_{t,\ell} \big[h^k(\tilde{s}^{t,\ell+1}) - h^k(s^{t,\ell})\big] \;\le\; H\sqrt{2TL\log(10/\delta)} \;+\; KH,$$

where the first sum is at most $H\sqrt{2TL\log(10/\delta)}$ outside F_{Az}, and the second telescopes within each episode to at most $\mathrm{sp}(h^k) \le H$.

C.5 Bounding Term (II)

We continue from the endogenous part of (32),

$$\text{Term (II.A, endo)} \;\le\; \tfrac{3}{2} \sum_{i \in I^{\mathrm{endo}}} H\, S_i,$$

$$S_i \;:=\; \sum_{k=1}^{K} \sum_{v,a} \nu^k_{Z_i^P}(v,a) \Big(\sqrt{\tfrac{2\,|\mathcal{S}_i|\,\rho}{\max\{N^{k-1}_{Z_i^P}(v,a),1\}}} + \tfrac{7\,|\mathcal{S}_i|\,\rho}{3\max\{N^{k-1}_{Z_i^P}(v,a)-1,\,1\}} \Big),$$

using $\sum_w \sqrt{\hat{P}^k_i(w \mid v,a)} \le \sqrt{|\mathcal{S}_i|}$ by Cauchy–Schwarz over the $|\mathcal{S}_i|$ successors.

The doubling rule controls S_i. By Lemma 3, $\nu^k_{Z_i^P}(v,a) \le \max\{N^{k-1}_{Z_i^P}(v,a),1\}$, which is the hypothesis of Jaksch et al. (2010, Lemma 19): for $z_1,\dots,z_n \ge 0$ with $z_k \le Z_{k-1} := \max\{1, \sum_{j<k} z_j\}$, $\sum_k z_k/\sqrt{Z_{k-1}} \le (\sqrt{2}+1)\sqrt{Z_n}$. Applied at each cell with $z_k = \nu^k_{Z_i^P}(v,a)$, and with $z_k/Z_{k-1} \le 2z_k/Z_k \le 2\log(Z_k/Z_{k-1})$ from $Z_k \le 2Z_{k-1}$ for the second sum,

$$\sum_{k=1}^{K} \frac{\nu^k_{Z_i^P}(v,a)}{\sqrt{\max\{N^{k-1}_{Z_i^P}(v,a),1\}}} \;\le\; (\sqrt{2}+1)\sqrt{N^K_{Z_i^P}(v,a)},$$

$$\sum_{k=1}^{K} \frac{\nu^k_{Z_i^P}(v,a)}{\max\{N^{k-1}_{Z_i^P}(v,a),1\}} \;\le\; 2\log N^K_{Z_i^P}(v,a) + 1,$$

where the second denominator of S_i is $\max\{N^{k-1}_{Z_i^P}(v,a)-1,1\} \ge \frac{1}{2}\max\{N^{k-1}_{Z_i^P}(v,a),1\}$. Summing the first over cells introduces the action count, which appears nowhere else:

$$\sum_{v \in \mathcal{S}[Z_i^P]} \sum_{a \in \mathcal{A}} \sqrt{N^K_{Z_i^P}(v,a)} \;\le\; \sqrt{|\mathcal{S}[Z_i^P]|\,|\mathcal{A}| \sum_{v,a} N^K_{Z_i^P}(v,a)} \;=\; \sqrt{|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T},$$

by Cauchy–Schwarz over the $|\mathcal{S}[Z_i^P]||\mathcal{A}|$ cells and because each original step increments exactly one of them.

The exogenous piece of (32) runs the same two telescoping bounds on the pooled counts, whose hypothesis $\nu^k_{Z_i^P}(v) \le \max\{N^{k-1}_{Z_i^P}(v),1\}$ is the second statement of Lemma 3, with ρ_{exo} for ρ, and the cell sum runs over $|\mathcal{S}[Z_i^P]|$ cells,

$$\sum_{v \in \mathcal{S}[Z_i^P]} \sqrt{N^K_{Z_i^P}(v)} \;\le\; \sqrt{|\mathcal{S}[Z_i^P]| \sum_{v} N^K_{Z_i^P}(v)} \;=\; \sqrt{|\mathcal{S}[Z_i^P]|\,T},$$

giving Term (II.A, exo) $= \widetilde{O}\big(H \sum_{i \in I^{\mathrm{exo}}} \sqrt{|\mathcal{S}_i||\mathcal{S}[Z_i^P]|T}\big)$, with no factor $|\mathcal{A}|$.

C.6 Assembling Proposition 1

Let $F := F^P_{\mathrm{endo}} \cup F^P_{\mathrm{exo}} \cup (F_{Az} \cap \mathcal{E})$. Each of the three has probability at most $\delta/5$ by (24), (25) and (26), so $\Pr[F] \le \frac{3\delta}{5} < \delta$, and $F \supseteq \mathcal{E}^c$ together with $F \supseteq F_{Az} \cap \mathcal{E}$ gives $F \supseteq F_{Az}$.

Work on F^c. There $\mathcal{E}$ holds, so Lemma 1 places the true P_i in $\mathcal{B}_i^k$ for every $i \in [d]$ and every episode, Lemma 6 gives $\lambda^\star(M)/L \le \lambda^\star(\widetilde{M}_{\mathcal{B}}^k) = \lambda^k$, so Term (I) of (30) is non-positive and (31) holds. Chaining (32) with the telescoping bounds of Appendix C.5,

$$\mathrm{Reg}_T(M) \;\le\; \widetilde{O}\Big(H \sum_{i\in I^{\mathrm{endo}}} \sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T}\Big) \;+\; \widetilde{O}\Big(H \sum_{i\in I^{\mathrm{exo}}} \sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,T}\Big) \\ +\; 7H \sum_{i\in[d]} |\mathcal{S}_i|\,\rho \sum_{v,a} \big(2\log N_{Z_i^P}^K(v,a) + 1\big) \;+\; KH \;+\; H\sqrt{2TL\log(10/\delta)},$$

with ρ_{exo} and v alone at exogenous factors. The last three terms are absorbed into the $\widetilde{O}$ of (6). The first two are polylogarithmic in T by $N_{Z_i^P}^K \le T$ and Lemma 2, and the third is at most $\sqrt{6\log(10/\delta)}\,H\sqrt{d\,T}$ by $L \le 3d$. Finally $H = D(M) + 1 \le 2D(M)$ gives the prefactor D. The bound holds with probability at least $1-\delta$. □

D Proof of Theorem 1

Throughout, $T_k := t_{k+1} - t_k$ is the length of episode k, everything is stated on $\mathcal{E}^k$, and y, x are the two parts of a state as in §A.3. The oracle returns $(\tilde\pi^k, \lambda^k, h^k)$ satisfying (1) on $\widetilde{M}_{\mathrm{ce}}^k$. The reference model is $\widetilde{M}_{L'}$, the lift of M under its true kernels with the timeline of $\widetilde{M}_{\mathrm{ce}}^k$: L' sub-steps, the endogenous factors drawn in index order and the exogenous coordinate at the transition out of $\ell = L' - 1$. The embedding of a policy π of M in it has gain $\lambda(M,\pi,s)/L'$, as in (21). Write $B_{\mathrm{exo}} := \sum_{j\in I^{\mathrm{exo}}} \sqrt{|\mathcal{S}_j|}$ and $H_{\mathrm{ce}} := 2 + 8\,D_{\mathrm{endo}}(M) \le 10H$, the bound on $\mathrm{sp}(h^k)$ past the threshold that Lemmas 7 and 8 below establish, and let k_0 be the first episode past the threshold of Lemma 8 below.

The smoothing costs $\|\hat P_i^{k,\dagger} - \hat P_i^k\|_1 \le 2/t_k$ by (5), and the exogenous estimation error is $\Delta_{\mathrm{exo}}^k := \sum_{i\in I^{\mathrm{exo}}} \|\hat P_i^{k,\dagger} - P_i\|_1$. We write $\hat P_i^k$ for $\hat P_i^{k,\dagger}$ except where the two are compared.

The rounds before t_{k_0} cost at most one unit each, and past them the regret splits into two terms,

$$\mathrm{Reg}_T(M) \;\le\; t_{k_0} \;+\; \sum_{k\ge k_0} T_k\big(\lambda^\star(M) - L'\lambda^k\big) \;+\; \text{Term (II)}, \tag{33}$$

by (30) with L' for L, from t_{k_0} on. The second term is Term (I) of (30), the benchmark defect, no longer non-positive without optimism on the exogenous part, and the third is Term (II), the deviation term, at the span of h^k. Lemmas 7 and 8 bound that span by H_{ce}, Lemma 9 bounds the defect, and the three terms are added at the end.

Lemmas 7 and 8 below both need a policy that reaches a given endogenous value quickly, and it comes from the following fact. Reaching $\{s : s_{\mathrm{endo}} = y''\}$ for a fixed y'', in M or in $\widetilde{M}_{\mathrm{ce}}^k$, is a stochastic shortest-path problem with unit costs in which some stationary policy reaches the target with probability one from every state, because both models are communicating. So by Bertsekas & Tsitsiklis (1991) one stationary deterministic policy attains the minimum expected hitting time from every start at once, and no randomized policy does better. This is needed because the minimum in (7) is taken inside the maxima, so a policy attaining D_{endo} could otherwise depend on the start.

The span of h^k is bounded in two steps: Lemma 7 by the endogenous diameter of the planned model, Lemma 8 by that of M.

Lemma 7 (Bias span of the plug-in planned model)**.** *Let* (π,λ,h) *satisfy* (1) *on* $\widetilde{M}_{\mathrm{ce}}^k$*. Then*

$$\mathrm{sp}(h) \;\le\; 1 + D_{\mathrm{endo}}\big(\widetilde{M}_{\mathrm{ce}}^k\big).$$

Proof. Write $u(y) := \sum_x \hat P_{\mathrm{exo}}^k(x)\,h(y,x)$. Since the exogenous successor is drawn from $\hat P_{\mathrm{exo}}^k$ regardless of (y,x) and of the augmented action, the optimality equation reads

$$\lambda + h(y,x) \;=\; \max_{(a,f)} \Big[\, r(y,x,a) \;+\; \sum_{y'} \widetilde P_{\mathrm{endo}}^k(y' \mid y,x,a,f)\,u(y') \,\Big], \tag{34}$$

The bracket is a reward in $[0,1]$ plus an expectation of u, so it lies in $[\min u, 1+\max u]$ for every (y,x), and $\mathrm{sp}(h) \le 1 + \mathrm{sp}(u)$.

It remains to bound $\mathrm{sp}(u)$, by the argument of Lemma 5. Fix y, y'' and let π be the shortest-path policy above for M^k_{ce}, so that $\mathbb{E}[T_{\mathrm{endo}}(y'' \mid M^k_{\mathrm{ce}}, \pi, (y,x))] \le D_{\mathrm{endo}}(M^k_{\mathrm{ce}})$ for every x. Dropping the maximum in (34) at $\pi(y,x)$ and using $r \ge 0$ gives $h(y,x) \ge -\lambda + \mathbb{E}[u(y') \mid y, x]$, and averaging over $x \sim \hat{P}^k_{\mathrm{exo}}$, independent of y because the exogenous coordinate is redrawn at every step, gives $u(y) \ge -\lambda + \mathbb{E}[u(y') \mid y]$ along the endogenous chain under π. Iterating to T_{endo} and exchanging y with y'',

$$|u(y) - u(y'')| \;\le\; \lambda\, \mathbb{E}[T_{\mathrm{endo}}] \;\le\; D_{\mathrm{endo}}(M^k_{\mathrm{ce}}),$$

since the gain of M^k_{ce} is at most 1. □

Unrolling (1) on $\widetilde{M}^k_{\mathrm{ce}}$ over one original step composes the per-factor maxima, each taken after the earlier draws, into one maximum over $\mathcal{A} \times \mathcal{F}_{\mathrm{endo}}$, so $(L'\lambda^k, h^k|_{\ell=0})$ solves (1) on M^k_{ce} and Lemma 7 applies to it; the sub-steps $\ell \ge 1$ add at most $(L'-1)\lambda^k \le 1$ as in Lemma 5, so $\mathrm{sp}(h^k) \le 2 + D_{\mathrm{endo}}\big(M^k_{\mathrm{ce}}\big)$.

Lemma 8 (Endogenous diameter of the plug-in model)**.** *On $\mathcal{E}^k$, if*

$$t_k \;\ge\; 512\, \tau^k_i\, B^2_{\mathrm{exo}}\, D_{\mathrm{endo}}(M)^2 \tag{35}$$

for $i \in I^{\mathrm{exo}}$, all of which share one τ^k_i, then $\Delta^k_{\mathrm{exo}} \le 1/(4D_{\mathrm{endo}}(M))$ and $D_{\mathrm{endo}}(M^k_{\mathrm{ce}}) \le 8\, D_{\mathrm{endo}}(M)$.

Proof. Fix a target y'', let π be the shortest-path policy above for M, and let σ be π run in M^k_{ce} with the endogenous targets randomized to reproduce P_{endo}. This randomization exists on $\mathcal{E}^k$, because $P_i \in \mathcal{B}^k_i$ for $i \in I^{\mathrm{endo}}$ and $\mathcal{B}^k_i$ is the convex hull of the target rows by Lemma 4. With $n := 2D_{\mathrm{endo}}(M)$, from every state s,

$$\Pr_{M^k_{\mathrm{ce}},\sigma}\big(T_{\mathrm{endo}} \le n \,\big|\, s\big) \overset{(a)}{\ge} \Pr_{M,\pi}\big(T_{\mathrm{endo}} \le n \,\big|\, s\big) - \frac{n}{2}\,\big\|\hat{P}^k_{\mathrm{exo}} - P_{\mathrm{exo}}\big\|_1 \overset{(b)}{\ge} \frac{1}{2} - \frac{n}{2}\,\Delta^k_{\mathrm{exo}} \overset{(c)}{\ge} \frac{1}{4}, \tag{36}$$

where step (a) holds because the two path laws over n steps differ only in the exogenous draw, whose total variation per step is $\frac{1}{2}\|\hat{P}^k_{\mathrm{exo}} - P_{\mathrm{exo}}\|_1$; step (b) is Markov's inequality from $\mathbb{E}_{M,\pi}[T_{\mathrm{endo}} \mid s] \le D_{\mathrm{endo}}(M)$; step (c) is $\Delta^k_{\mathrm{exo}} \le 1/(4D_{\mathrm{endo}}(M))$, shown last. Hence

$$D_{\mathrm{endo}}(M^k_{\mathrm{ce}}) \overset{(d)}{\le} \mathbb{E}_{M^k_{\mathrm{ce}},\sigma}[T_{\mathrm{endo}}] \overset{(e)}{\le} n \sum_{j \ge 1} \big(\tfrac{3}{4}\big)^{j-1} \;=\; 4n \;=\; 8\, D_{\mathrm{endo}}(M), \tag{37}$$

where step (d) is the shortest-path fact above, and step (e) restarts the bound at every multiple of n: each block reaches the target with probability at least $\frac{1}{4}$ from every starting state, so T_{endo}/n is dominated by a geometric variable of success probability $\frac{1}{4}$. It remains to check that (35) gives $\Delta^k_{\mathrm{exo}} \le 1/(4D_{\mathrm{endo}}(M))$: for $t_k \ge 2$,

$$\begin{aligned}
\Delta^k_{\mathrm{exo}} \;&=\; \sum_{i \in I^{\mathrm{exo}}} \big\|\hat{P}^{k,\dagger}_i - P_i\big\|_1 \;\le\; \sum_{i \in I^{\mathrm{exo}}} \Big(\big\|\hat{P}^{k,\dagger}_i - \hat{P}^k_i\big\|_1 + \sum_w \big|\hat{P}^k_i(w) - P_i(w)\big|\Big) \\
&\overset{(a)}{\le} \sum_{i \in I^{\mathrm{exo}}} \Big(\frac{2}{t_k} + \sum_w \sqrt{\frac{2\hat{P}^k_i(w)\,\tau^k_i}{t_k}} + \frac{7|\mathcal{S}_i|\,\tau^k_i}{3(t_k-1)}\Big) \;\le\; \sum_{i \in I^{\mathrm{exo}}} \Big(\frac{2}{t_k} + \sqrt{\frac{2\tau^k_i|\mathcal{S}_i|}{t_k}} + \frac{7|\mathcal{S}_i|\,\tau^k_i}{3(t_k-1)}\Big) \\
&\overset{(b)}{\le} 2 \sum_{i \in I^{\mathrm{exo}}} \Big(\sqrt{\frac{2\tau^k_i|\mathcal{S}_i|}{t_k}} + \frac{7|\mathcal{S}_i|\,\tau^k_i}{3(t_k-1)}\Big) \\
&\le\; 2\sqrt{\frac{2\tau^k_i}{t_k}}\, B_{\mathrm{exo}} + \frac{28\,\tau^k_i}{3\,t_k}\, B^2_{\mathrm{exo}} \\
&\overset{(c)}{\le} \frac{1}{8D_{\mathrm{endo}}(M)} + \frac{1}{8D_{\mathrm{endo}}(M)} \;=\; \frac{1}{4D_{\mathrm{endo}}(M)},
\end{aligned} \tag{38}$$

where step (a) is the smoothing cost $2/t_k$ and Lemma 1 at the single parent cell of factor i, whose count is t_k; step (b) uses $2/t_k \le \sqrt{2\tau_i^k|\mathcal{S}_i|/t_k}$, since $\tau_i^k \ge \log 40 > 2$; step (c) is $t_k \ge 512\,\tau_i^k B_{\text{exo}}^2 D_{\text{endo}}(M)^2$ with $D_{\text{endo}}(M) \ge 1$. □

The first term of (33) is free of T: by $t_{k_0} \le 2t_{k_0-1} + \Sigma_{\text{cell}}$ from Lemma 3 with $\tau_i^k \le \rho_{\text{exo}}$, $t_{k_0} \le 1024\,\rho_{\text{exo}} B_{\text{exo}}^2 D_{\text{endo}}(M)^2 + \Sigma_{\text{cell}}$, which is absorbed into the $\widetilde{O}$ of (8). Past the threshold, Lemma 7 transferred to the lift, $\text{sp}(h^k) \le 2 + D_{\text{endo}}(M_{\text{ce}}^k)$, and Lemma 8 give

$$\text{sp}(h^k) \;\le\; 2 + 8\,D_{\text{endo}}(M) \;=\; H_{\text{ce}} \qquad \text{for } k \ge k_0. \tag{39}$$

Lemma 9 (Benchmark defect)**.** *On $\mathcal{E}^k$,*

$$\lambda^k \;\ge\; \frac{\lambda^\star(M)}{L'} \;-\; \frac{\text{sp}(h^k)}{2L'}\,\Delta_{\text{exo}}^k. \tag{40}$$

Proof. Run the proof of Lemma 6 with $\widetilde{M}_{L'}$ in place of $\widetilde{M}$ and the oracle's (λ^k, h^k) on $\widetilde{M}_{\text{ce}}^k$. At $\ell \in \{0,1\}$ and at the endogenous sub-steps (28) holds unchanged, since $P_i \in \mathcal{B}_i^k$ for $i \in I^{\text{endo}}$ on $\mathcal{E}^k$. At the transition out of $\ell = L'-1$ the exogenous coordinate is drawn from $\hat{P}_{\text{exo}}^k$ instead of P_{exo}, so there

$$\begin{aligned} \lambda^k + h^k(\tilde{s}) \;&\overset{(a)}{\ge}\; \sum_{\tilde{s}'} \widetilde{P}(\tilde{s}' \mid \tilde{s}, \tilde{a})\, h^k(\tilde{s}') - \frac{\text{sp}(h^k)}{2}\,\big\|\hat{P}_{\text{exo}}^k - P_{\text{exo}}\big\|_1 \\ &\overset{(b)}{\ge}\; \sum_{\tilde{s}'} \widetilde{P}(\tilde{s}' \mid \tilde{s}, \tilde{a})\, h^k(\tilde{s}') - \frac{\text{sp}(h^k)}{2}\,\Delta_{\text{exo}}^k, \end{aligned} \tag{41}$$

where step (a) is (28) with the exogenous draw exchanged, in which only the span of h^k matters because the entries of $\hat{P}_{\text{exo}}^k - P_{\text{exo}}$ sum to zero, and step (b) is $\|\prod_i \hat{P}_i^k - \prod_i P_i\|_1 \le \sum_i \|\hat{P}_i^k - P_i\|_1$. Summing along n sub-steps of any policy π of $\widetilde{M}_{L'}$ from $\tilde{s}^1$, where the defect enters once per L' sub-steps,

$$\lambda^k \;\ge\; \frac{1}{n}\,\mathbb{E}_\pi\Big[\sum_{u \le n} \tilde{r}(\tilde{s}^u, \tilde{a}^u)\Big] - \frac{\text{sp}(h^k)}{n} - \Big\lceil \frac{n}{L'} \Big\rceil \frac{\text{sp}(h^k)}{2n}\,\Delta_{\text{exo}}^k \;\xrightarrow[n\to\infty]{}\; \lambda(\widetilde{M}_{L'}, \pi, \tilde{s}^1) - \frac{\text{sp}(h^k)}{2L'}\,\Delta_{\text{exo}}^k, \tag{42}$$

and π the embedding of $\pi^\star(M)$, whose gain is $\lambda^\star(M)/L'$, gives (40). □

The three terms of (33) add up as follows:

$$\begin{aligned} \text{Reg}_T(M) &\overset{(a)}{\le} t_{k_0} + \tfrac{H_{\text{ce}}}{2} \sum_{k \ge k_0} T_k\,\Delta_{\text{exo}}^k \;+\; \text{Term (II.A)} \;+\; \text{Term (II.B)} \\ &\overset{(b)}{\le} t_{k_0} + H_{\text{ce}} \sum_{k \ge k_0} T_k\,\Delta_{\text{exo}}^k \;+\; \tfrac{3}{2} H_{\text{ce}} \sum_{k \ge k_0} \sum_{t=t_k}^{t_{k+1}-1} \sum_{i \in I^{\text{endo}}} \sum_{w} \varepsilon_i^k\big(w \mid s^t[Z_i^P], a^t\big) \\ &\qquad +\; \text{Term (II.B)} \\ &\overset{(c)}{\le} t_{k_0} + \widetilde{O}\Big(H_{\text{ce}} \sum_{i \in I^{\text{exo}}} \sum_{k \ge k_0} T_k \sqrt{|\mathcal{S}_i|/t_k}\Big) \;+\; \widetilde{O}\Big(H_{\text{ce}} \sum_{i \in I^{\text{endo}}} \sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T}\Big) \\ &\qquad +\; \widetilde{O}\big(H_{\text{ce}} \sqrt{d\,T}\big) \\ &\overset{(d)}{=} \widetilde{O}\Big(H_{\text{ce}} \sum_{i \in I^{\text{endo}}} \sqrt{|\mathcal{S}_i|\,|\mathcal{S}[Z_i^P]|\,|\mathcal{A}|\,T} \;+\; H_{\text{ce}} \sum_{i \in I^{\text{exo}}} \sqrt{|\mathcal{S}_i|\,T}\Big), \end{aligned}$$

which is (8). Step (a) is (33) followed by Lemma 9 with $\text{sp}(h^k) \le H_{\text{ce}}$ from (39), in which the L' cancels, and the split of Term (II) at the true kernel as in (32)(a); step (b) is (32)(b), where the exogenous part costs $\frac{H_{\text{ce}}}{2}\Delta_{\text{exo}}^k$ per original step since the entries of $\hat{P}_{\text{exo}}^k - P_{\text{exo}}$ sum to zero; step (c) bounds the endogenous part and Term (II.B) as in (32)(c)–(d), with TL' sub-steps and the sum of Term (II.B) started at $(t_{k_0}, 0)$, which is a sub-step decided by the data before it, so that the truncation

of §C.3 applies unchanged, and bounds the exogenous part by Lemma 1 at the single parent cell of an i.i.d. factor, whose count is t_k; step (d) uses Lemma 3: the episodes with $t_k < \Sigma_{\text{cell}}$ contribute at most $\sum_k T_k \le 3\Sigma_{\text{cell}}$, every later one has $t_{k+1} \le 3t_k$, so $T_k t_k^{-1/2} \le \sqrt{3}\int_{t_k}^{t_{k+1}} u^{-1/2}\,\mathrm{d}u$ and the rest sums to at most $2\sqrt{3T}$, and the lower-order terms are absorbed as in §C.6, where t_{k_0} does not depend on T.

E Proof of Theorem 2 (Lower Bound)

E.1 Matching upper and lower rates

Remark 1 (Matching upper and lower rates)**.** *Substituting the paper-wide bounds $|\mathcal{S}_i| \le W$ and $|\mathcal{S}[Z_i^P]| \le \Lambda \le W^m$ (§3) into* (6) *gives the factored form of Proposition 1,*

$$\mathrm{Reg}_T \;=\; \widetilde{O}\Big(D\, d_{\text{endo}}\sqrt{W^{m+1}\,|\mathcal{A}|\,T} \;+\; D\, d_{\text{exo}}\sqrt{W^{m+1}\,T}\Big), \tag{43}$$

which is compared term by term with the lower bound (9)*:*

	Endogenous	*Exogenous*
Upper (43)	$D\, d_{\text{endo}}\sqrt{W^{m+1}\,\lvert\mathcal{A}\rvert\,T}$	$D\, d_{\text{exo}}\sqrt{W^{m+1}\,T}$
Lower (9)	$\sqrt{d_{\text{endo}}\,W^m\,\lvert\mathcal{A}\rvert\,T}$	$\sqrt{d_{\text{exo}}\,W^m\,T}$

The rate matches in T and in the $|\mathcal{A}|$-dependence, present on the endogenous side and absent on the exogenous side, and no constant beyond the diameter and the factor cardinalities enters either side. The residual gap in the symbols is the diameter D and the $\sqrt{d\,W}$ Cauchy–Schwarz slack, and the instance realizes those symbols at larger values, as below.

This is not minimax optimality. The table compares the bounds symbol for symbol, whereas the hard instance realizes those symbols at larger values: factor cardinality $W+1$ rather than W, maximum parent scope $1+\max\{m, \lceil\log_2 d\rceil+2\}$ rather than m, and $d+d_{\text{exo}}+\lceil\log_2 d\rceil+2$ exogenous factors rather than d_{exo}. On the endogenous side this leaves a further $W^{(1+\max\{m,\,\lceil\log_2 d\rceil+2\}-m)/2}$, which is $\sqrt{W}$ when $m \ge \lceil\log_2 d\rceil + 2$ and larger otherwise. On the exogenous side it leaves a further factor $\Theta\big((d + d_{\text{exo}} + \lceil\log_2 d\rceil + 2)/d_{\text{exo}}\big)$, unbounded when $d_{\text{endo}} \gg d_{\text{exo}}$. It also carries no diameter dependence. The construction below puts the instance in the model class by redrawing each coordinate at a rate $\kappa = 1/(dT^2)$, and at that rate $D(M) = \kappa^{-\Theta(d)} = (dT^2)^{\Theta(d)}$ while $\mathrm{sp}(h^\star) = O(m)$, so the instance does not determine a D-prefactor. We therefore claim matching rates *in T and in $|\mathcal{A}|$, not matching bounds.*

E.2 Full proof

Proof. We suppress the time index and write s, s' for the current and next states of a single transition. Throughout, M denotes a member of the family of hard instances constructed below, indexed by its hidden parameters.

The proof has four parts: the instance, rules (i)–(vi), which makes every second step test one hidden parameter; the decomposition (53), which writes the regret, up to an additive 3, as the sum over hidden parameters of the regret incurred on the blocks that test it; one bound per parameter, (64) for an endogenous one (a bandit problem) and (68) for an exogenous one (a two-point test); and their sum (70).

The instance has six kinds of factor.

	values	class	$\lvert Z^P\rvert$	job
c	$\{0,1\}$	exo	1	the clock, alternates $0,1,0,1,\ldots$
$b_0,\ldots,b_{\lceil\log_2 d\rceil}$	$\{0,1\}$	exo	0	fair coins encoding a random index $J \in [d]$
$v_1,\ldots,v_d$	$\{0,\ldots,W\}$	exo	$\lceil\log_2 d\rceil + 3$	nonzero on the window $[J, J+m)$ only
$s_1^{\text{endo}},\ldots,s_{d_{\text{endo}}}^{\text{endo}}$	$\{0,1\}$	endo	$m+1$	biased only under one hidden action
$s_1^{\text{exo}},\ldots,s_{d_{\text{exo}}}^{\text{exo}}$	$\{0,1\}$	exo	$m+1$	a coin with a hidden bias no action steers
β	$\{+,-\}$	endo	2	records the learner's guess at that direction

A *block* is a $c = 1$ step together with the $c = 0$ step that follows it. At the $c = 1$ step the learner acts, β records the action, and the binary factors are drawn using the window currently stored in v. At the $c = 0$ step the reward is collected, the next window is opened, and the location bits redraw J. Call factor j *active* when the m value factors it depends on are all nonzero, $v_j, \dots, v_{j+m-1} \neq 0$. A binary factor is biased only when it is active, and its window and the opened one both have length m, so

$$\begin{aligned} &J \le d-m+1: \quad v_j,\dots,v_{j+m-1}\neq 0 \iff [j,j{+}m)\subseteq[J,J{+}m) \iff j=J,\\ &J \ge d-m+2: \quad \{\, j : v_j,\dots,v_{j+m-1}\neq 0\,\}=\emptyset. \end{aligned} \tag{44}$$

In the second case the window reaches past v_d, and $j = J$ is one of the $m - 1$ boundary factors of rule (v) below, which are never active.

The two classes differ in one way. An endogenous parameter is one of the $|\mathcal{A}|$ actions, and evidence about it is split across the actions, so the gap can be $\sqrt{|\mathcal{A}|}$ larger before the learner can detect it. An exogenous parameter is the direction of a bias, every action produces the same evidence, and no $\sqrt{|\mathcal{A}|}$ factor arises. Hence the two terms of (9) differ by $\sqrt{|\mathcal{A}|}$.

Rules (i)–(vi) give the transition kernel and the reward.

(i) *Counter.* $c' = 1 - c$, started at $c = 0$, with parent scope $\{c\}$.

(ii) *Location bits.* Each b_q is an independent fair coin regardless of the action, with empty parent scope. Here b_0 selects the class, endogenous or exogenous, with probability exactly $\frac{1}{2}$ each, and the remaining $\lceil \log_2 d \rceil$ bits select the index within it.

(iii) *Value factors.* With 0 the reserved inactive marker and $\kappa := 1/(dT^2)$, whose purpose appears below, v_i transitions independently at each coordinate as

$$v_i' = \begin{cases} \mathrm{Unif}(\{0,\dots,W\}) & \text{with probability } \kappa,\\ \mathrm{Unif}(\{1,\dots,W\}) & \text{otherwise, if } c=0 \text{ and } J \le i < J+m,\\ v_i & \text{otherwise, if } c=1,\\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

Its parent scope is $\{c, b_0, b_1, \dots, b_{\lceil \log_2 d \rceil}, v_i\}$. Window indices $i > d$, which occur when $J \ge d - m + 2$, correspond to no factor and are ignored.

(iv) *Bet.* For two fixed actions $a_+, a_- \in \mathcal{A}$, the endogenous bit β has parent scope $\{c, \beta\}$ and

$$\beta' \;=\; \begin{cases} - & \text{if } c=1 \text{ and } a=a_-,\\ + & \text{if } c=1 \text{ and } a\neq a_-,\\ \beta & \text{if } c=0. \end{cases}$$

The bet and the binary coins of rule (v) are written by the same transition, so the guess is committed before the coin it bets on is revealed.

(v) *Binary factors.* Endogenous $s_1^{\mathrm{endo}}, \dots, s_{d_{\mathrm{endo}}}^{\mathrm{endo}} \in \{0,1\}$ with one hidden parameter $a^\star_{j,v} \in \mathcal{A}$ per window value $v \in \{1,\dots,W\}^m$, gap Δ_{endo} of (48) below and parent scope $\{c, v_j, \dots, v_{j+m-1}\}$ of size $m + 1$,

$$\Pr\big(s_j'^{\mathrm{endo}} = 1 \;\big|\; v, a\big) = \begin{cases} \frac{1}{2} + \Delta_{\mathrm{endo}} & \text{if } c=1,\ v_j,\dots,v_{j+m-1}\neq 0,\ a = a^\star_{j,v},\\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{46}$$

and exogenous binary factors $s_1^{\mathrm{exo}}, \dots, s_{d_{\mathrm{exo}}}^{\mathrm{exo}} \in \{0,1\}$, with $j' := d_{\mathrm{endo}} + j$, split by whether the natural parent scope fits inside $\{1,\dots,d\}$. *Interior* factors, those with $j \le d_{\mathrm{exo}} - m + 1$, carry one hidden parameter $\theta_{j,v} \in \{-1,+1\}$ per window value v, gap Δ_{exo} of (48) and parent scope $\{c, v_{j'}, \dots, v_{j'+m-1}\}$ of size $m + 1$, and transition independently of a as

$$\Pr\big(s_j'^{\mathrm{exo}} = 1 \;\big|\; v\big) = \begin{cases} \frac{1}{2} + \theta_{j,v}\cdot\Delta_{\mathrm{exo}} & \text{if } c=1,\ v_{j'},\dots,v_{j'+m-1}\neq 0,\\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{47}$$

The remaining $m - 1$ *boundary* factors would reach past v_d, and we set $s_j'^{\mathrm{exo}} \sim \mathrm{Ber}(\frac{1}{2})$ independently of (s, a), with empty parent scope and never active, so that they contribute reward 0 under every policy.

Two gap parameters, one per class, are set at the Le Cam thresholds of the per-pair sub-problems solved below,

$$\Delta_{\mathrm{endo}} = c_1\sqrt{\frac{|\mathcal{A}|\cdot d_{\mathrm{endo}}W^m}{T}}, \qquad \Delta_{\mathrm{exo}} = c_0\sqrt{\frac{d_{\mathrm{exo}}W^m}{T}}, \tag{48}$$

with $c_0 = c_1 = 1/32$.

(vi) *Reward.* The deterministic known reward function is

$$r(s,a) = \mathbb{1}\{c=0\}\cdot\min\Big\{1,\ \sum_{j=1}^{d_{\mathrm{endo}}} r_j^{\mathrm{endo}}(v_j, s_j^{\mathrm{endo}}) + \sum_{j=1}^{d_{\mathrm{exo}}} r_j^{\mathrm{exo}}(v_{d_{\mathrm{endo}}+j}, s_j^{\mathrm{exo}}, \beta)\Big\}, \tag{49}$$

with per-factor pieces

$$r_j^{\mathrm{endo}}(v_j, s_j^{\mathrm{endo}}) = s_j^{\mathrm{endo}}\cdot\mathbb{1}\big\{v_j,\dots,v_{j+m-1}\neq 0\big\}, \tag{50}$$

$$r_j^{\mathrm{exo}}(v_{j'}, s_j^{\mathrm{exo}}, \beta) = \begin{cases} s_j^{\mathrm{exo}} & v_{j'},\dots,v_{j'+m-1}\neq 0,\ \beta=+,\\ 1-s_j^{\mathrm{exo}} & v_{j'},\dots,v_{j'+m-1}\neq 0,\ \beta=-,\\ 0 & v_{j'},\dots,v_{j'+m-1}\text{ not all nonzero, or } j'+m-1>d. \end{cases} \tag{51}$$

The instance is started at $c=0$ with $v\equiv 0$, $s^{\mathrm{endo}}\equiv s^{\mathrm{exo}}\equiv 0$, $\beta=+$ and the location bits drawn as fair coins, so that its first transition opens a window by rule (iii) and the $c=1$ step that follows begins block 1.

The instance is admissible for Theorem 2. It has $1+1+(\lceil\log_2 d\rceil+1)+d+d_{\mathrm{endo}}+d_{\mathrm{exo}} = 2d+\lceil\log_2 d\rceil+3$ factors, of which $d_{\mathrm{endo}}+1$ are endogenous, its largest factor is a value factor of cardinality $W+1$, and the table gives $\max_i|Z_i^P| = 1+\max\{m,\lceil\log_2 d\rceil+2\}$. Every exogenous factor takes no action argument and depends only on exogenous indices, by rules (i)–(iii) and (v), which is Definition 1. The endogenous factors, the coins s_j^{endo} and the bet β, depend on the action, and the definition places no condition on them. The reward (49) is deterministic, known and, by its truncation, in $[0,1]$. The first branch of (45) ensures reachability. With it every configuration is reached with positive probability within three steps, because the counter alternates and the bet is rewritten only on leaving a $c=1$ state. So $D(M)<\infty$, and the branch is action-free and has an empty scope.

That branch is also the only way (44) can fail, so we track it. Let $\xi_i^t\in\{0,1\}$ indicate that coordinate i took the first branch of (45) at the transition out of step t; it is recorded separately because that branch may return the current value and leave v unchanged. Write

$$\mathcal{G} \;:=\; \big(c^t,\ b_0^t,\dots,b_{\lceil\log_2 d\rceil}^t,\ v_1^t,\dots,v_d^t,\ \xi_1^t,\dots,\xi_d^t\big)_{t\le T}$$

for the trajectories of the action-free, parameter-free factors; the coins $s^{\mathrm{endo}}, s^{\mathrm{exo}}$ are not in it, and the per-pair arguments below condition on it. Block k occupies steps $2k$ and $2k+1$, and its window is opened by the transition out of step $2k-1$. Block k is *mixed* if

$$\max_{i\in[d]}\max\big\{\xi_i^{2k-1},\ \xi_i^{2k}\big\} \;=\; 1.$$

These $2d$ indicators are disjoint from those of every other block, so block types are independent. A block that is not mixed carries the window opened at $2k-1$ throughout, so (44) applies to it. Whether a block is mixed, and hence its type below, is a function of $\mathcal{G}$.

A block's regret belongs to its one active pair. We index pairs (j,v) by the *global* index $j\in[d]$: $j\le d_{\mathrm{endo}}$ refers to the j-th endogenous factor s_j^{endo}, and $j>d_{\mathrm{endo}}$ to the $(j-d_{\mathrm{endo}})$-th exogenous factor, written s_j^{exo}. Only pairs with $v\in\{1,\dots,W\}^m$ are tracked, since a tuple with a zero coordinate leaves its factor inactive. Call factor j *inactive* at the current step when its parent window $v_j,\dots,v_{j+m-1}$ contains at least one zero coordinate. By (44), every global index other than the active $j=J$ is inactive throughout a block that is not mixed.

Claim 1. *For any inactive factor j and any action a, at every step,*

$$r_j^{\mathrm{endo}} = r_j^{\mathrm{exo}} = 0 \qquad \textit{and} \qquad \Pr\big(s_j'=1\mid s,a\big)=\tfrac{1}{2},$$

and neither quantity depends on any hidden parameter of the instance.

Proof. Both reward pieces (50)–(51) are gated by the indicator that the full window $v_j, \dots, v_{j+m-1}$ is nonzero, which fails by hypothesis, so both vanish identically. The biased branches of (46) and (47) carry the same guard, so the fallback branch applies and $\Pr(s'_j = 1) = \frac{1}{2}$ regardless of $a^\star_{j,v}$ and $\theta_{j,v}$, and at steps with $c \neq 1$ it applies to every factor. □

We now decompose the regret. Each block ends in one reward-collecting step, and by (44) it has one active pair or, at a boundary J, none. Write R_k for the reward block k collects and ϑ_k for its type, where the type is (j, v) for the active pair, ∂ for a boundary block and $\varnothing$ for a mixed block, and set

$$R_k^{\max} := \begin{cases} \frac{1}{2} + \Delta_{\text{endo}}, & \vartheta_k = (j,v),\ j \le d_{\text{endo}}, \\ \frac{1}{2} + \Delta_{\text{exo}}, & \vartheta_k = (j,v),\ j > d_{\text{endo}}, \\ 0, & \vartheta_k = \partial, \\ 1, & \vartheta_k = \varnothing, \end{cases} \qquad \Delta_k := R_k^{\max} - R_k, \qquad \bar{R} := \mathbb{E}\big[R_k^{\max}\big],$$

together with $\mathrm{Reg}(j,v) := \sum_{k\,:\,\vartheta_k=(j,v)} \Delta_k$. The first three rows are the largest expected reward the block can carry, attained by playing $a^\star_{j,v}$ at an endogenous block and betting $\theta_{j,v}$ at an exogenous one, and equal to 0 at a boundary block by Claim 1. On a mixed block no pair is named and the fourth row is only the trivial bound $r \le 1$, which makes $\Delta_k = 1 - R_k \ge 0$ there. On a block of type (j,v) exactly one pair is active and $R_k^{\max}$ is the largest conditional expected block reward given $\mathcal{G}$ and the history $\mathcal{H}_k$ preceding the block, so $\mathbb{E}[\Delta_k \mid \mathcal{G}, \mathcal{H}_k] \ge 0$ there.

No action changes which pair is active or how often, since the exogenous coordinates evolve autonomously, and the block reward depends only on the action at the block's $c = 1$ step. Every policy therefore collects at most $R_k^{\max}$ in expectation at block k, and the policy just named attains it except on a mixed block, where it loses at most 1, so with blocks of length two

$$\frac{\bar{R}}{2} - d\kappa \;\le\; \lambda^\star(M) \;\le\; \frac{\bar{R}}{2}, \qquad \Pr\big(\vartheta_k = \varnothing\big) \;=\; 1 - (1-\kappa)^{2d} \;\le\; 2d\kappa \;=\; \frac{2}{T^2}. \tag{52}$$

Let Π be the uniform prior on the hidden parameters, independent across pairs, $a^\star_{j,v} \sim \mathrm{Unif}(\mathcal{A})$ at an endogenous pair and $\theta_{j,v} \sim \mathrm{Unif}(\{\pm 1\})$ at an exogenous one, and write $\mathbb{E}_\Pi$ for expectation over Π together with the trajectory it induces. With $K_T := \lfloor (T-1)/2 \rfloor$ the number of blocks, which is deterministic and satisfies $T/3 \le K_T \le T/2$ for $T \ge 6$, and $\rho_T \in [0,2]$ the reward of the at most two steps left at the ends of the horizon,

$$\begin{aligned}
\sup_M \mathbb{E}\big[\mathrm{Reg}_T(M)\big] &\overset{(a)}{\ge} \mathbb{E}_\Pi\big[\mathrm{Reg}_T(M)\big] \\
&\overset{(b)}{\ge} \mathbb{E}_\Pi\Big[\sum_{k \le K_T} \big(2\lambda^\star(M) - R_k\big)\Big] \;-\; \mathbb{E}_\Pi[\rho_T] \\
&\overset{(c)}{=} \underbrace{K_T\, \mathbb{E}_\Pi\big[2\lambda^\star(M) - \bar{R}\big]}_{=:\text{ Term (A): type average}} + \underbrace{\sum_{k \le K_T} \mathbb{E}_\Pi\big[\Delta_k\big]}_{=:\text{ Term (B): block sub-optimality}} - \underbrace{\mathbb{E}_\Pi\big[\rho_T\big]}_{=:\text{ Term (C): horizon ends}} \\
&\overset{(d)}{=} K_T\, \mathbb{E}_\Pi\big[2\lambda^\star(M) - \bar{R}\big] \;+\; \sum_{(j,v)} \mathbb{E}_\Pi\big[\mathrm{Reg}(j,v)\big] \\
&\qquad + \sum_{k \le K_T} \mathbb{E}_\Pi\big[\Delta_k \mathbb{1}\{\vartheta_k = \varnothing\}\big] \;-\; \mathbb{E}_\Pi\big[\rho_T\big] \\
&\overset{(e)}{\ge} \sum_{(j,v)} \mathbb{E}_\Pi\big[\mathrm{Reg}(j,v)\big] \;-\; \frac{1}{T} \;-\; 2,
\end{aligned} \tag{53}$$

where step (a) is Yao's principle (Yao, 1977), since the supremum over the family is at least the average under Π; step (b) uses that reward is collected at the $c = 0$ step of a block and nowhere else, where the K_T blocks partition the horizon up to the at most two steps at its ends, and drops $(T - 2K_T)\lambda^\star(M) \ge 0$; step (c) adds and subtracts $R_k^{\max}$ and takes $\mathbb{E}_\Pi[R_k^{\max}] = \bar{R}$ for every k,

since the types ϑ_k are i.i.d. across blocks with a law free of the hidden parameters; step (d) splits Term (B) along $1 = \mathbb{1}\{\vartheta_k = (j,v)\} + \mathbb{1}\{\vartheta_k = \partial\} + \mathbb{1}\{\vartheta_k = \varnothing\}$, the boundary type contributing nothing because $R_k = R_k^{\max} = 0$ there by Claim 1; and step (e) bounds the three remaining terms, Term (A) by $-2d\kappa \le 2\lambda^\star(M) - \bar{R} \le 0$ from (52) for every member of the family, with $K_T \le T/2$, the mixed sum by $\Delta_k = 1 - R_k \ge 0$, which lets it be discarded, and Term (C) by $r \in [0,1]$.

Each summand of (53) is a problem in isolation: a pair is observed only on the blocks in which it is active, and its parameter must be identified from those. We first count the blocks assigned to a pair, then bound one summand of each class. The two classes differ only in the number of candidates, $|\mathcal{A}|$ versus two.

The location bits put J almost uniformly on each class. The $\lceil \log_2 d \rceil$ bits below b_0 take $A_{\mathrm{loc}} := 2^{\lceil \log_2 d \rceil} \ge d$ equally likely values, split among the d_{endo} endogenous indices into groups of $\lfloor A_{\mathrm{loc}}/d_{\mathrm{endo}} \rfloor$ or $\lceil A_{\mathrm{loc}}/d_{\mathrm{endo}} \rceil$ values, so for $1 \le j \le d_{\mathrm{endo}}$

$$\frac{1}{4d_{\mathrm{endo}}} \overset{(a)}{\le} \frac{1}{2} \cdot \frac{\lfloor A_{\mathrm{loc}}/d_{\mathrm{endo}} \rfloor}{A_{\mathrm{loc}}} \le \Pr(J = j) \le \frac{1}{2} \cdot \frac{\lceil A_{\mathrm{loc}}/d_{\mathrm{endo}} \rceil}{A_{\mathrm{loc}}} \overset{(b)}{\le} \frac{1}{d_{\mathrm{endo}}}, \tag{54}$$

where the factor $\frac{1}{2}$ is b_0, step (a) is $\lfloor x \rfloor \ge x/2$ and step (b) is $\lceil x \rceil \le 2x$, both at $x = A_{\mathrm{loc}}/d_{\mathrm{endo}} \ge 1$; the same holds with d_{exo} for $d_{\mathrm{endo}} < j \le d$.

A block has type (j,v) when $J = j$ at its opening, the value tuple on the window, uniform on $\{1,\dots,W\}^m$ and independent of J, equals v, and the first branch is taken at none of its $2d$ coordinate draws, so by (54) and $(1-\kappa)^{2d} \ge 1 - 2/T^2 \ge \frac{7}{8}$,

$$p_{j,v} := \Pr\big(\vartheta_k = (j,v)\big) = \Pr(J = j)\, W^{-m}(1-\kappa)^{2d} \in \Big[\frac{7}{32\, d_{\mathrm{endo}} W^m},\ \frac{1}{d_{\mathrm{endo}} W^m}\Big]$$

for $j \le d_{\mathrm{endo}}$, and the same with d_{exo} for $j > d_{\mathrm{endo}}$. Let $n_{j,v}$ be the number of blocks of type (j,v), so $n_{j,v} = \sum_{k \le K_T} X_k$ with $X_k \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p_{j,v})$, since the types are independent across blocks. We fix the two *reference* activation counts

$$\mu_{\mathrm{endo}} := \frac{T}{4 d_{\mathrm{endo}} W^m}, \qquad j \le d_{\mathrm{endo}}, \tag{55}$$

$$\mu_{\mathrm{exo}} := \frac{T}{4 d_{\mathrm{exo}} W^m}, \qquad j > d_{\mathrm{endo}}. \tag{56}$$

These bracket the true expectations $\mathbb{E}[n_{j,v}] = K_T\, p_{j,v}$ within a factor 4 below and 2 above,

$$\tfrac{1}{4}\,\mu_{\mathrm{endo}} \le \mathbb{E}[n_{j,v}] \le 2\,\mu_{\mathrm{endo}} \quad (j \le d_{\mathrm{endo}}), \qquad \tfrac{1}{4}\,\mu_{\mathrm{exo}} \le \mathbb{E}[n_{j,v}] \le 2\,\mu_{\mathrm{exo}} \quad (j > d_{\mathrm{endo}}). \tag{57}$$

The per-pair bounds need only that the count is of the right order with constant probability.

Claim 2. *For every tracked pair (j,v), the event $\mathcal{V}_{j,v} := \{\frac{\mu}{8} \le n_{j,v} \le 128\,\mu\} \cap \{\vartheta_k \ne \varnothing \ \forall k \le K_T\}$ satisfies*

$$\Pr\big(\mathcal{V}_{j,v}\big) \ge \tfrac{3}{32}, \tag{58}$$

where μ is μ_{endo} for an endogenous pair and μ_{exo} for an exogenous one.

Proof. Write $\bar{n} := \mathbb{E}[n_{j,v}] = K_T\, p_{j,v}$, so $\mu/4 \le \bar{n} \le 2\mu$ by (57). Upper tail: by Markov's inequality, $\Pr(n_{j,v} > 128\mu) \le \bar{n}/(128\mu) \le 1/64$. Lower tail: $n_{j,v} = \sum_{k \le K_T} X_k$ with $X_k \sim \mathrm{Ber}(p_{j,v})$ i.i.d., so $\mathrm{Var}(n_{j,v}) \le \bar{n}$ and $\mathbb{E}[n_{j,v}^2] \le \bar{n} + \bar{n}^2$, and the Paley–Zygmund inequality (Paley & Zygmund, 1932) at level $\frac{1}{2}$ gives

$$\Pr\big(n_{j,v} \ge \tfrac{\bar{n}}{2}\big) \ge \frac{1}{4} \cdot \frac{\bar{n}^2}{\bar{n} + \bar{n}^2} = \frac{1}{4} \cdot \frac{\bar{n}}{1 + \bar{n}} \ge \frac{1}{8},$$

using $\bar{n} \ge \mu/4 = T/(16\, d_{\mathrm{endo}} W^m) \ge 2d/d_{\mathrm{endo}} > 1$ by the horizon condition with $|\mathcal{A}| \ge 2$, likewise for an exogenous pair, and $\bar{n}/2 \ge \mu/8$. Mixed blocks: by (52) and a union bound over the $K_T \le T/2$ blocks, $\Pr(\exists\, k \le K_T : \vartheta_k = \varnothing) \le 1/T$. Hence

$$\begin{aligned}\Pr\big(\mathcal{V}_{j,v}\big) &\ge \Pr\big(n_{j,v} \ge \tfrac{\bar{n}}{2}\big) - \Pr(n_{j,v} > 128\mu) - \Pr(\exists\, k \le K_T : \vartheta_k = \varnothing) \\ &\overset{(a)}{\ge} \tfrac{1}{8} - \tfrac{1}{64} - \tfrac{1}{T} \overset{(b)}{\ge} \tfrac{3}{32},\end{aligned}$$

where step (a) collects the three bounds above and step (b) is $T \geq 192$, which the horizon condition gives from $|\mathcal{A}| \geq 2$, $d \geq 3$ and $W^m \geq 2$. □

Fix an endogenous pair (j, v) with $j \leq d_{\text{endo}}$. Given $\mathcal{G}$ with $n_{j,v} = n$ and no block mixed, the n activated blocks are the only steps whose law depends on $a^\star_{j,v}$, by Claim 1 and (46), and each is one draw of $\text{Ber}(\frac{1}{2} + \Delta_{\text{endo}}\mathbb{1}\{a = a^\star_{j,v}\})$, so the algorithm faces an $|\mathcal{A}|$-armed Bernoulli bandit with n pulls in which arm $a^\star_{j,v}$ has mean $\frac{1}{2} + \Delta_{\text{endo}}$ and all other arms have mean $\frac{1}{2}$. Write H_i for the member of the family whose biased arm is arm i, H_0 for the auxiliary instance outside the family in which every arm has mean $\frac{1}{2}$, and $\mathbb{P}_{H_0}$, $\mathbb{P}_{H_i}$ for the law of the whole interaction given $\mathcal{G}$, on which $n_{j,v} = n$, under H_0 and H_i. With N_i the number of activated blocks in which arm i was played, where the same arm is biased and scored,

$$\textstyle\sum_{i \in \mathcal{A}} N_i \;=\; n \tag{59}$$

pathwise. Every display below is on the event (58). The divergence between the null and the i-th instance is

$$\text{KL}(\mathbb{P}_{H_0} \,\|\, \mathbb{P}_{H_i}) \overset{(a)}{=} \mathbb{E}_{H_0}[N_i] \cdot \text{KL}\big(\text{Ber}(\tfrac{1}{2}) \,\big\|\, \text{Ber}(\tfrac{1}{2} + \Delta_{\text{endo}})\big) \overset{(b)}{\leq} 4\Delta^2_{\text{endo}}\, \mathbb{E}_{H_0}[N_i], \tag{60}$$

where step (a) is the chain rule for the divergence of the two trajectory laws on the event (58): H_0 and H_i agree on every parameter but $a^\star_{j,v}$, so their kernels differ only at factor j in an activated block where arm i is played, and each such step contributes the Bernoulli divergence (Lattimore & Szepesvári, 2020, Lemma 15.1); step (b) is $\text{KL}(\text{Ber}(\frac{1}{2})\|\text{Ber}(\frac{1}{2} + \Delta)) \leq 4\Delta^2$ for $\Delta \leq 1/4$, whose precondition holds because $\Delta^2_{\text{endo}} = c_1^2 |\mathcal{A}| d_{\text{endo}} W^m / T \leq c_1^2/4 \leq 1/16$ since $c_1 = 1/32$ and $T \geq 16|\mathcal{A}| d W^m$. Summing over arms,

$$\begin{aligned}
\sum_{i \in \mathcal{A}} \text{TV}(\mathbb{P}_{H_0}, \mathbb{P}_{H_i}) &\overset{(a)}{\leq} \sum_i \sqrt{\tfrac{1}{2}\text{KL}(\mathbb{P}_{H_0} \,\|\, \mathbb{P}_{H_i})} \;\leq\; \sum_i \Delta_{\text{endo}} \sqrt{2\, \mathbb{E}_{H_0}[N_i]} \\
&\overset{(b)}{\leq} \Delta_{\text{endo}} \sqrt{2|\mathcal{A}| \textstyle\sum_i \mathbb{E}_{H_0}[N_i]} \;=\; \Delta_{\text{endo}} \sqrt{2|\mathcal{A}| n} \\
&= \tfrac{c_1}{2} |\mathcal{A}| \sqrt{2n/\mu_{\text{endo}}} \overset{(c)}{\leq} 8c_1 |\mathcal{A}|,
\end{aligned} \tag{61}$$

where step (a) is Pinsker's inequality (Lattimore & Szepesvári, 2020, Eq. 14.12) followed by (60); step (b) is Cauchy–Schwarz over the $|\mathcal{A}|$ arms, after which (59) enters as a constant determined by $\mathcal{G}$; step (c) uses $n \leq 128\mu_{\text{endo}}$ on the event (58). Averaging the scored counts over the uniform prior,

$$\begin{aligned}
\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbb{E}_{H_i}[N_i] &\overset{(a)}{\leq} \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \Big(\mathbb{E}_{H_0}[N_i] + n\, \text{TV}(\mathbb{P}_{H_0}, \mathbb{P}_{H_i}) \Big) \\
&\leq \frac{n}{|\mathcal{A}|} \;+\; \frac{n}{|\mathcal{A}|} \cdot 8c_1 |\mathcal{A}| \;=\; n\Big(\frac{1}{|\mathcal{A}|} + 8c_1 \Big) \overset{(b)}{\leq} \frac{3n}{4},
\end{aligned} \tag{62}$$

where step (a) is $\mathbb{E}_{H_i}[N_i] \leq \mathbb{E}_{H_0}[N_i] + n\, \text{TV}(\mathbb{P}_{H_0}, \mathbb{P}_{H_i})$, since $0 \leq N_i \leq n$, and the next line follows from (59) and (61); step (b) uses $c_1 = 1/32$ and $|\mathcal{A}| \geq 2$. Hence

$$\begin{aligned}
\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbb{E}_{H_i}\big[\text{Reg}(j, v) \,\big|\, n_{j,v} = n\big] \;&=\; \Delta_{\text{endo}} \Big(n - \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \mathbb{E}_{H_i}[N_i] \Big) \\
&\geq\; \Delta_{\text{endo}} \cdot \frac{n}{4},
\end{aligned} \tag{63}$$

where the equality counts Δ_{endo} for each of the $n - N_{a^\star}$ scored pulls of a non-optimal arm, where the biased arm is arm i under H_i. Then (62) leaves at least $n/4$ of them in Bayes average. Combining with the two-sided event,

$$\begin{aligned}
\mathbb{E}_\Pi[\text{Reg}(j, v)] \;&\geq\; \Pr\big(\mathcal{V}_{j,v}\big) \cdot \Delta_{\text{endo}} \cdot \frac{\mu_{\text{endo}}}{32} \;\geq\; \frac{3}{32} \cdot \frac{1}{32} \cdot \frac{1}{32} \sqrt{\frac{|\mathcal{A}|\, d_{\text{endo}} W^m}{T}} \cdot \frac{T}{4\, d_{\text{endo}} W^m} \\
&= \frac{3}{2^{17}} \sqrt{\frac{|\mathcal{A}|\, T}{d_{\text{endo}} W^m}},
\end{aligned} \tag{64}$$

where the first inequality applies (63) on the event (58) at $n \geq \mu_{\text{endo}}/8$ and discards the contribution off the event, which is non-negative by $\mathbb{E}[\Delta_k \mid \mathcal{G}, \mathcal{H}_k] \geq 0$, and the second substitutes (48) and (55). The first step is legitimate because $\mathcal{G}$ determines $n_{j,v}$ and its law does not depend on $a^\star_{j,v}$, since $\mathcal{G}$ is generated by action-free, parameter-free factors.

For an exogenous pair the action does not affect the exogenous transitions, and hence it does not alter the rate at which information about $\theta_{j,v}$ accumulates. We therefore use Le Cam's two-point method (Tsybakov, 2009, Chapter 2) with $\mathbb{P}_\pm$ the law of the whole interaction under $\theta_{j,v} = \pm 1$, which produces observations $\text{Ber}(\frac{1}{2} \pm \Delta_{\text{exo}})$. Per draw,

$$\text{KL}\big(\text{Ber}(\tfrac{1}{2} - \Delta_{\text{exo}}) \,\big\|\, \text{Ber}(\tfrac{1}{2} + \Delta_{\text{exo}})\big) \;=\; 2\Delta_{\text{exo}} \cdot \log\frac{1 + 2\Delta_{\text{exo}}}{1 - 2\Delta_{\text{exo}}} \;\overset{(a)}{\leq}\; \frac{8\Delta_{\text{exo}}^2}{1 - 2\Delta_{\text{exo}}} \;\overset{(b)}{\leq}\; 16\Delta_{\text{exo}}^2, \tag{65}$$

where step (a) is $\log((1+u)/(1-u)) \leq 2u/(1-u)$ at $u = 2\Delta_{\text{exo}}$; step (b) is $\Delta_{\text{exo}} \leq 1/4$, since $c_0 = 1/32$ gives $\Delta_{\text{exo}}^2 \leq c_0^2/4 < 1/16$. On the event (58), writing TV_n for the total variation between $\mathbb{P}_+$ and $\mathbb{P}_-$ given $\mathcal{G}$ with $n_{j,v} = n$ on it,

$$\text{TV}_n \;\overset{(a)}{\leq}\; \Delta_{\text{exo}}\sqrt{8n} \;\leq\; \Delta_{\text{exo}}\sqrt{8 \cdot 128\,\mu_{\text{exo}}} \;=\; 16c_0 \;\overset{(b)}{\leq}\; \tfrac{1}{2},$$

where step (a) is Pinsker's inequality followed by (65) at each of the n activated blocks, which on the event (58) are the only steps at which the parent window of factor j equals v, conditioned on $\mathcal{G}$, whose law does not depend on $\theta_{j,v}$ since none of its generating factors has s_j^{exo} among its parents; the middle bound is the event (58) with $\mu_{\text{exo}}\Delta_{\text{exo}}^2 = c_0^2/4$ by (48) and (56); and step (b) is $c_0 = 1/32$.

Let N_+ and N_- count the *scored* decisions at which the latched bet is $+$ and $-$, so that $N_+ + N_- = n$ pathwise and the bet is wrong at $N_{-\theta_{j,v}}$ of them, since $+$ is optimal under $\theta_{j,v} = +1$ and $-$ under $\theta_{j,v} = -1$. Then

$$\begin{aligned}
\frac{1}{2}\sum_{\theta \in \{\pm 1\}} \mathbb{E}_\theta\big[\text{Reg}(j,v) \,\big|\, n_{j,v} = n\big] \;&\overset{(a)}{=}\; \Delta_{\text{exo}}\big(\mathbb{E}_+[N_-] + \mathbb{E}_-[N_+]\big) \\
&\overset{(b)}{\geq}\; \Delta_{\text{exo}} \cdot n\big(1 - \text{TV}_n\big) \qquad (66) \\
&\geq\; \Delta_{\text{exo}} \cdot \frac{n}{2}, \qquad (67)
\end{aligned}$$

where step (a) counts $2\Delta_{\text{exo}}$ for each wrong scored decision, since $\mathbb{E}[r_j^{\text{exo}} \mid \beta = +] - \mathbb{E}[r_j^{\text{exo}} \mid \beta = -] = 2\theta_{j,v}\Delta_{\text{exo}}$ on activation, because the coin is drawn *after* the bet was latched, and averages over the uniform prior on $\theta_{j,v}$; step (b) is $\mathbb{E}_+[N_-] + \mathbb{E}_-[N_+] = n - \big(\mathbb{E}_-[N_-] - \mathbb{E}_+[N_-]\big) \geq n\big(1 - \text{TV}_n\big)$ by $N_+ + N_- = n$ and $N_- \in [0, n]$, and the last line uses $\text{TV}_n \leq \frac{1}{2}$. Combining with the visit concentration (58),

$$\begin{aligned}
\mathbb{E}_\Pi[\text{Reg}(j,v)] \;&\geq\; \Pr\big(\mathcal{V}_{j,v}\big) \cdot \Delta_{\text{exo}} \cdot \frac{\mu_{\text{exo}}}{16} \;\geq\; \frac{3}{32} \cdot \frac{1}{16} \cdot \frac{1}{32}\sqrt{\frac{d_{\text{exo}}W^m}{T}} \cdot \frac{T}{4\,d_{\text{exo}}W^m} \\
&=\; \frac{3}{2^{16}}\sqrt{\frac{T}{d_{\text{exo}}W^m}}, \qquad (68)
\end{aligned}$$

where the first inequality applies (67) on the event (58) at $n \geq \mu_{\text{exo}}/8$, where the discarded off-event contribution is non-negative as at (64), and the second substitutes (48) and (56).

Adding the per-pair bounds, the sum of step (e) of (53) splits by class. Write $\text{Reg}_{\text{endo}} := \sum_{j=1}^{d_{\text{endo}}} \sum_{v \in [W]^m} \mathbb{E}_\Pi[\text{Reg}(j,v)]$ and $\text{Reg}_{\text{exo}} := \sum_{j=d_{\text{endo}}+1}^{d_{\text{endo}}+d_{\text{exo}}-m+1} \sum_{v \in [W]^m} \mathbb{E}_\Pi[\text{Reg}(j,v)]$. Summing the endogenous contributions,

$$\text{Reg}_{\text{endo}} \;\geq\; d_{\text{endo}}W^m \cdot \frac{3}{2^{17}}\sqrt{\frac{|\mathcal{A}|\,T}{d_{\text{endo}}W^m}} \;=\; \frac{3}{2^{17}}\sqrt{d_{\text{endo}}W^m|\mathcal{A}|\,T}, \tag{69}$$

by (64) at each summand. The exogenous sum is the same with (68) at each summand, where only the $d_{\text{exo}} - m + 1$ interior factors contribute,

$$\text{Reg}_{\text{exo}} \;\geq\; (d_{\text{exo}} - m + 1)W^m \cdot \frac{3}{2^{16}}\sqrt{\frac{T}{d_{\text{exo}}W^m}} \;\geq\; \frac{3}{2^{17}}\sqrt{d_{\text{exo}}W^m T},$$

where the last inequality uses $d_{\text{exo}} \geq 2m$, giving $d_{\text{exo}} - m + 1 \geq d_{\text{exo}}/2$.

Step (e) of (53) subtracts $1/T + 2 \leq 3$ once rather than per pair. By (69) and the horizon condition $T \geq 2^{36}|\mathcal{A}|dW^m$, $\text{Reg}_{\text{endo}} \geq 3 \cdot 2^{-17}\sqrt{d_{\text{endo}}W^m|\mathcal{A}|T} \geq 3 \cdot 2^{-17}\sqrt{2^{36}} = 6$, so

$$\begin{aligned}\sup_M \mathbb{E}[\text{Reg}_T(M)] \;&\geq\; \text{Reg}_{\text{endo}} + \text{Reg}_{\text{exo}} - 3 \;\geq\; \tfrac{1}{2}\big(\text{Reg}_{\text{endo}} + \text{Reg}_{\text{exo}}\big) \\ &\geq\; \frac{3}{2^{18}}\Big(\sqrt{d_{\text{endo}}W^m|\mathcal{A}|\,T} \;+\; \sqrt{d_{\text{exo}}W^m T}\Big), \end{aligned} \tag{70}$$

which is (9) with the constant $3 \cdot 2^{-18}$ and completes the proof. □

F Synthetic Instance: Full Specification

The instance of §7 is sampled, not designed. Fix $d_{\text{endo}}, d_{\text{exo}} \geq 1$, an alphabet size $W \geq 2$, an action count $|\mathcal{A}|$, a concentration $\alpha > 0$ and an instance seed. Every factor takes values in $\{1, \ldots, W\}$, so $|\mathcal{S}| = W^{d_{\text{endo}}+d_{\text{exo}}}$.

Parent scopes. Writing $[d] = I^{\text{endo}} \cup I^{\text{exo}}$ with $I^{\text{endo}} = \{1, \ldots, d_{\text{endo}}\}$, factor $i \in I^{\text{endo}}$ has $Z_i^P = \{i\} \cup \{i \bmod d_{\text{endo}} + 1\} \cup \{d_{\text{endo}} + ((i-1) \bmod d_{\text{exo}}) + 1\}$, and factor $d_{\text{endo}} + j \in I^{\text{exo}}$ has $Z^P = \{d_{\text{endo}} + j\} \cup \{d_{\text{endo}} + (j \bmod d_{\text{exo}}) + 1\}$, which is a singleton when $d_{\text{exo}} = 1$. So $m \leq 3$, with equality when $d_{\text{endo}} \geq 2$, every exogenous scope lies inside I^{exo}, and every endogenous factor depends on one exogenous coordinate. Both clauses of Definition 1 hold, and neither holds vacuously.

Kernels. Each row is an independent $\text{Dir}(\alpha \mathbf{1}_W)$ sample drawn once from the instance seed: for $i \in I^{\text{endo}}$ one row per (parent value, action) pair, and for $i \in I^{\text{exo}}$ one row per parent value with no action argument. All entries are positive with probability one, so $D(M) < \infty$ and the instance is communicating; small α makes the rows peaked, which keeps the chain slow to mix and the problem non-trivial.

Reward. Two vectors $u, w \in [0,1]^W$ are drawn uniformly once, and

$$r(s) \;=\; \frac{u\big(s[1]\big)\, w\big(s[d_{\text{endo}}+1]\big)}{\max_v u(v) \cdot \max_v w(v)} \;\in\; [0,1].$$

It is a product of one endogenous and one exogenous coordinate, so it is joint in the two blocks and is not additively separable. It carries no action argument, which the setting of §3 permits, so the only role of the action is to steer the endogenous block. The partition instance of Figure 4(b) uses instead a product over all exogenous coordinates, given below.

The instance we run. At $d_{\text{endo}} = 1$, $d_{\text{exo}} = 3$, $W = 4$, $|\mathcal{A}| = 8$, $\alpha = 0.3$ and instance seed 0 we obtain $|\mathcal{S}| = 256$, parent scopes $\{1,2\}, \{2,3\}, \{3,4\}, \{2,4\}$, so $m = 2$ and $\Lambda = 16$, and $\lambda^\star = 0.353$ with $\text{sp}(h^\star) = 2.54$, $D(M) = 646$ and $D_{\text{endo}}(M) = 2.16$. Of the $\Sigma_{\text{cell}} = 176$ confidence cells, 48 are exogenous. A uniformly random policy earns 0.219 and the best constant action 0.290, so 17.8% of the optimal gain has to be learned, and 7 of the 8 actions appear in the optimal policy. The endogenous rows are drawn as 32 columns per parent cell, of which the first $|\mathcal{A}|$ are used, so the action sets of Figure 4(a) are nested and the instance at $|\mathcal{A}| = 8$ is the one above. That sweep runs to $T = 10^6$, ten times the horizon of the other panels, because at $T = 10^5$ the exploration of each additional action is not yet complete for either algorithm. The exogenous factors are Markov rather than i.i.d., so the instance exercises Proposition 1 rather than Theorem 1.

The partition instance. For Figure 4(b) the exogenous scopes form a chain, $Z^P_{d_{\text{endo}}+1} = \{d_{\text{endo}} + 1\}$ and $Z^P_{d_{\text{endo}}+j} = \{d_{\text{endo}} + j - 1,\, d_{\text{endo}} + j\}$ for $j \geq 2$, in place of the cycle above, so that the first n exogenous factors are closed under the scope relation for every n and declaring exactly those exogenous satisfies Definition 1(ii); with the cycle no proper subset is closed. Everything else is drawn as above except the reward, which depends on every exogenous factor, $r(s) = u(s[1]) \prod_{j=1}^{d_{\text{exo}}} w_j(s[d_{\text{endo}} + j])$ normalized to $[0, 1]$, so that each declared factor matters for the decision. This gives $\lambda^\star = 0.224$ and $\text{sp}(h^\star) = 1.04$.

The i.i.d. exogenous instance. For Figure 5(a) the exogenous parent scopes are emptied, $Z^P_i = \emptyset$ for $i \in I^{\text{exo}}$, so each exogenous factor is drawn i.i.d. from one $\text{Dir}(\alpha \mathbf{1}_W)$ row, which is the hypothesis of Theorem 1; the endogenous kernels are those of the main instance and the reward vectors are fresh draws from the same seed. Each exogenous distribution is then mixed with the uniform one, $p \leftarrow (1 - \beta)p + \beta/W$, at $\beta \in \{1, 0.3, 0.1, 0.03\}$. The endogenous kernels do not depend on β, so $D_{\text{endo}}(M)$ stays at 2.1, while the rarest joint exogenous value has probability 1.6×10^{-2} at $\beta = 1$ and 6.7×10^{-7} at $\beta = 0.03$, which puts $D(M)$ at 96, 3.0×10^3, 7.4×10^4 and 2.0×10^6; at $d_{\text{exo}} = 4$ and $\beta = 0.1$ the same construction gives $|\mathcal{S}| = 1024$, $D(M) = 2.9 \times 10^6$ and $D_{\text{endo}}(M) = 2.1$; both diameters are computed exactly by policy iteration on the hitting-time problem. The plug-in variant is Hybrid-DORL with $\mathcal{W}^k_i := 0$ on the exogenous factors and the planner given (5). The timings of Figure 5(b) are one sweep of tabular value iteration over each lifted model at $d_{\text{endo}} = 1$, $W = 4$ and $|\mathcal{A}| = 8$, with the layers stored per (s, a', s'_{buf}), at true rows widened by a constant radius 0.05, taking the median of ten sweeps. At radius 0 both lifts return the gain of M, which checks the implementation.

G EXOGENOUS STRUCTURE IN LLM-SERVING TRACES

Whether any real system satisfies Definition 1 is an empirical question. We measure it on the public BurstGPT trace (Wang et al., 2024), which records arrival times and token counts and nothing about the server, so the arrival stream is the one part of such a system that no scheduling decision alters. Bucketed into per-tick counts, it is action-independent and depends only on itself, which is the definition, and it is *not* i.i.d. Its autocorrelations do not decay over five lags (Table 1), which is the signature of a slowly drifting rate rather than of short-range dependence, and its hourly rate moves by $11.7\times$ across a day. This is the relevant case. An exogenous block is worth exploiting only when it has structure of its own. Proposition 1 covers this case, and Theorem 1 does not.

BurstGPT covers sixty-one days of ChatGPT and GPT-4 requests. Each row is an arrival time with the prompt and response token counts. The trace does not record the serving system, its batching, or its scheduler, so the endogenous dynamics of §7 cannot be extracted from it.

Table 1: Per-tick arrival counts of BurstGPT. A Poisson stream would have $\text{var}/\text{mean} = 1$ and autocorrelation 0 at every lag.

tick	rounds	mean	var/mean	$\rho(1), \ldots, \rho(5)$
1 s	5,269,974	0.27	3.84	.60 .61 .61 .60 .61
60 s	87,833	16.3	138	.94 .90 .87 .84 .82

The serving instances. The instances of Figure 6 take three exogenous factors from BurstGPT at a one-minute tick, each bucketed into four levels: the arrival count, cut at the 50th, 80th and 95th percentiles of the segment so that the levels are idle, light, busy and burst; the mean response length; and the share of GPT-4 requests. Their kernels are the empirical transition counts of one quarter of the trace, one instance per quarter, with parent scopes fixed after inspecting the conditional entropies of the levels, so that the response length and the model mix each depend on the arrival level. The endogenous factor is the backlog with levels 0, 6, 25 and 100 requests. The action opens one of eight capacities, equally spaced from 0 to 40 slots. A slot serves 900 tokens per minute divided by the mean response length of the current level, and requests arriving during the minute can be served in it. The next backlog level is the bucket of the current backlog plus the arrivals minus the served requests, with the arrivals drawn from the minutes of the segment at the current arrival level (4000 Monte Carlo draws per cell). The reward is the expected number of served requests, where a

GPT-4 request counts twice, normalized by its maximum, minus 0.6 times the normalized backlog and 0.2 times the normalized capacity, rescaled to $[0, 1]$. A behaviour trace on the first segment (one seed, 3×10^4 steps) shows the source of the regret. At the three lower arrival levels NFA-DORL plays the optimal capacity 14 to 18% of the time and keeps it about three levels too high, and its per-quarter regret does not fall, while Hybrid-DORL plays the optimal capacity 98% of the time in idle minutes. UCRL2 is less affected than the factored baselines because its joint rows are sparse on these instances, and the truncation $\min\{\varepsilon, \hat{P}\}$ applied to a joint row vanishes on successors never observed.

H Algorithms Compared

This appendix specifies the four algorithms of §7 and Hybrid-DORL-PI in full. All runs use $\delta = 0.1$, five seeds and $T = 10^5$ unless stated otherwise, every algorithm plans exactly on the joint state space, and every band and error bar in the figures is one standard deviation over the seeds.

H.1 Algorithm descriptions

The three algorithms that use the confidence radius of (3) take the union-bound constants of §4.2 reduced from 40 and t_k^3 to 6 and t_k, and FACTORED-UCRL takes the same ρ inside its ℓ_1 radius. The cube covers a union over the random sample size that is far too conservative at simulation horizons. The reduction is applied identically to all four, so it cannot favor any of them.

Algorithm	Factored	Confidence set	Exo. counts pooled	Description
UCRL2	no	Bernstein	no	Learns the joint transition $P(s' \mid s, a)$ with a single $(\lvert\mathcal{S}\rvert, \lvert\mathcal{A}\rvert, \lvert\mathcal{S}\rvert)$ box.
FACTORED-UCRL	yes	ℓ_1 ball	no	One ℓ_1 ball of radius $\sqrt{2\lvert\mathcal{S}_i\rvert\rho / N^{k-1}_{Z_i^P}(v, a)}$ per factor, indexed by parent scope and action.
NFA-DORL	yes	Bernstein	no	Per-factor learning treating all factors as endogenous. One Bernstein box per factor, indexed by parent scope and action.
HYBRID-DORL	yes	Bernstein	yes	Every factor, endogenous or exogenous, is learned per factor with Bernstein widening. The one difference is that the visit counts of an exogenous factor are pooled across actions.
Hybrid-DORL-PI (Alg. 2)	yes	Bernstein	yes	Hybrid-DORL with no widening on the exogenous factors, which the planner takes from the smoothed estimate (5). Figure 5 only.

Table 2: The algorithms compared. All share the doubling-trick episode schedule of §4. The three factored algorithms plan by value iteration on the lift folded into one original step, where each factor's row is chosen within its confidence set given the factors drawn before it, and FACTORED-UCRL uses the exact inner maximum over its ℓ_1 ball. UCRL2 plans by extended value iteration on its joint box. Each row changes one component relative to the row above it.

I Related Work in Detail

Factored MDPs. Regret minimization in factored MDPs with known structure was initiated by Osband & Van Roy (2014) and improved by Xu & Tewari (2020) (non-episodic, relaxed oracle),

Tian et al. (2020) (Bernstein rates and lower bounds), Talebi et al. (2021) (factored diameter) and Rosenberg & Mansour (2021) (unknown structure, oracle-efficient). All of them allow the action among the coordinates of a parent scope, so the counts of a factor whose scope omits the action are pooled over actions, and in the per-factor bounds its term carries no $|\mathcal{A}|$. None singles out such factors, and none gives a lower bound with a mixed partition. The lower bound of Tian et al. (2020, Theorem 4) covers the episodic case in which no factor depends on the action. All of these bounds, including ours, contain a diameter. A bias-span bound would be smaller, but it needs a planner restricted to bounded-span solutions (Bartlett & Tewari, 2009; Fruit et al., 2018), and no oracle-efficient factored version of such a planner is known.

MDPs with exogenous information. A second group of papers uses the split under additional restrictions. Dietterich et al. (2018); Trimponias & Dietterich (2023) assume an additive reward and use the split for variance reduction rather than for a rate; Efroni et al. (2022a) lets the reward depend on the endogenous part only; Efroni et al. (2022b); Levine et al. (2025) drop the reward and learn the endogenous latent state; Sinclair et al. (2023); Wan et al. (2024) allow joint reward but need known deterministic endogenous dynamics. Closest is Neely (2024): a controlled component next to an i.i.d. vector of unknown distribution, with a guarantee free of that vector's cardinality, but for time-average cost under constraints, on an unfactored state, and as a convergence time rather than a regret bound. Joint reward with unknown stochastic endogenous dynamics, as a regret problem, has to our knowledge not been addressed. Online MDPs with adversarial rewards (Even-Dar et al., 2009; Wei et al., 2018; Zhao et al., 2022) put the uncontrolled part in the reward instead and are incomparable.