From Worst Case to Conditional Frontiers in Reinforcement Learning

Amar Ahmad, Yvonne Vallés, and Youssef Idaghdour

Public Health Research Center,
New York University, Abu Dhabi, UAE

Abstract. We study how fundamental statistical limits in reinforcement learning change when multiple real-world challenges interact. Focusing on sample inefficiency, nonstationarity, partial observability, and high-dimensional observations, we synthesise existing lower-bound arguments and show that their effects are generally non-additive. We formalise three structure-conditioned mechanisms: multiplicative complexity penalties in partially observable nonstationary environments, memory collapse under low-rank observation structure, and explicit finite-horizon safety guarantees via probabilistic shielding. Rather than proposing new algorithms, the paper clarifies how exploitable structure reshapes worst-case guarantees and motivates a shift from pessimistic minimax analysis toward conditional complexity frontiers that tighten as structure is detected online.

Keywords: Reinforcement Learning, Sample Complexity, Partial Observability, Nonstationarity, Safe RL, Control, Statistical Limits

1 Introduction

Reinforcement learning (RL) has moved well beyond proof-of-concept videogame agents and now powers prototype systems in robotics, industrial process control, dialogue management, and clinical decision support. However, truly reliable field deployment remains rare. In practice, practitioners encounter a set of mutually reinforcing statistical roadblocks that limit scalability and reliability. These include sample inefficiency, since collecting sufficient on-policy experience on physical platforms is slow, risky, and expensive, and nonstationarity, whereby real-world dynamics drift due to wear, payload changes, human intervention, or shifting objectives. They also include partial observability, as sensors provide only coarse, delayed, or noisy access to the latent state, and the curse of dimensionality, whereby rich observation streams such as images, force torque traces, or LiDAR dramatically expand the effective search space.

Classical theoretical analyses typically treat each of these difficulties in isolation, yielding pessimistic worst-case bounds on regret or sample complexity. [Azar et al.(2017)] establish minimax-optimal sample complexity for discounted MDPs under a generative model; [Jaksch et al.(2010)] provide regret bounds for undiscounted MDPs via optimism under uncertainty; and [Krishnamurthy et al.(2016)] extend PAC learning guarantees to contextual decision processes. Each of these contributions assumes stationarity and, in most cases, full or benign partial observability-conditions that rarely hold jointly in deployed systems.

These classical results establish sharp guarantees under stationarity and simplified observability assumptions, but do not characterise how learning difficulty compounds when multiple sources of uncertainty and drift act simultaneously.

Understanding, and exploiting, the structure that emerges from these interactions is therefore essential. Recent work points toward possible remedies:[Berkenkamp et al.(2017)] use Gaussian process models to provide safety guarantees during learning, but require accurate prior knowledge of system dynamics; [Nagabandi et al.(2018), Finn et al.(2019)]

David C. Wyld et al. (Eds): CSML, AISCA, DNLP, SOEA, NET, BDHI, SIPO – 2026
pp. 37-50, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160204

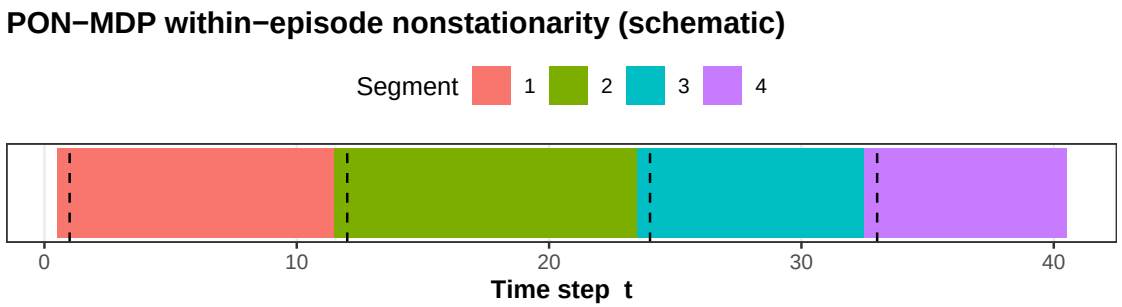

Fig. 1. Within-episode nonstationarity in a partially observable nonstationary MDP (PON-MDP). Both transition dynamics and emission models switch across unknown segments of an episode, forcing the agent to jointly perform latent-state inference and change-point adaptation. (Author-generated schematic.)

demonstrate rapid adaptation through meta-learning, yet their analyses do not quantify how adaptation cost scales with partial observability or safety constraints; and [Levine et al.(2017)] survey uncertainty-aware control without providing unified lower bounds. A theoretical treatment integrating these perspectives remains absent.

These lines of work demonstrate how structure, uncertainty estimates, or prior experience can improve empirical performance, yet their theoretical analyses remain largely decoupled and do not provide a unified account of how such structure reshapes fundamental complexity limits.

Figure 1 illustrates one such interaction through the lens of within-episode nonstationarity under partial observability. The schematic depicts a partially observable nonstationary MDP (PON-MDP) in which both transition dynamics and observation emission models change across unknown segments within a single episode. An agent operating in this setting must simultaneously infer latent state information, detect change-points, and adapt its policy online, highlighting how classical assumptions underlying isolated lower bounds are violated in combination.

We defer algorithmic implementations and empirical validation of online structure discovery to future work, as this paper is concerned with isolating and clarifying the fundamental complexity mechanisms that limit reinforcement learning in real-world settings. In particular, nonstationarity and the need for rapid adaptation arise naturally in lifelong and meta-learning scenarios, where agents must continuously update their representations and policies as tasks and dynamics evolve [Ring(1994), Finn et al.(2017), Chua et al.(2018)].

We do not propose a new algorithmic pipeline; Section 2 introduces a theoretical framework and formal constructions used to state and interpret structure-conditioned lower-bound mechanisms.

Throughout, our contributions should be understood as a unifying conceptual framework: the results in Section 2 are not new minimax bounds per se, but reinterpret and combine existing lower-bound constructions to expose previously implicit interactions between nonstationarity, partial observability, memory, and safety.

While these approaches focus on enabling fast adaptation, they do not explicitly analyse how adaptation costs interact with memory, observability, and safety constraints at the level of fundamental lower bounds.

The results in Section 2.1 are stated as formal propositions and theorems, but should be interpreted as conceptual in two complementary senses. First, constants and tightness are not the focus; instead, the aim is to isolate the fundamental mechanisms that govern how statistical complexity scales when multiple challenges interact. Second, the statements are designed to make transparent how multiplicative penalties, structure-driven memory

collapse, and tunable safety guarantees arise under explicit and interpretable assumptions. All proofs are self-contained given the stated constructions and are intended to prioritise clarity and mechanism identification over sharp optimisation of bounds. To complement these theoretical results, Section 3 presents a simple simulation study that serves as an illustrative instantiation of the proposed mechanisms, rather than as a validation of the theoretical bounds. Additional conceptual visualisations of the key mechanisms are provided in Appendix A.

Many of the works cited in this paper predate 2020 because they establish foundational lower bounds and impossibility results that remain state-of-the-art. While recent advances focus primarily on algorithmic scalability, the fundamental statistical barriers identified in classical analyses continue to govern achievable performance. Our aim is therefore not to replace these results, but to reinterpret them through a structure-conditioned lens that is directly relevant to modern adaptive systems.

We also note that foundational results in reinforcement learning theory frequently appear in peer-reviewed conference proceedings such as NeurIPS, ICML, and COLT, which serve as primary publication venues in this field; in several cases no extended journal version exists, and we cite these works accordingly.

2 Background and problem setting

Our analysis considers episodic reinforcement learning with a finite horizon, covering both fully observable and partially observable decision processes. An MDP consists of a set of states and actions, together with a Markovian transition rule $P(\cdot \mid s, a)$ and a bounded reward function $R(s, a) \in [0, 1]$. In partially observable environments, the underlying state is hidden, and the agent instead receives observations $o_t \in \mathcal{O}$ drawn from an emission distribution conditioned on the current state. Standard background on MDPs and POMDPs follows classical treatments [Sutton and Barto(2018), Kaelbling et al.(1996)].

At each time step $t \in \{0, \ldots, H-1\}$, the agent selects an action $a_t \in \mathcal{A}$. The environment then transitions according to P, emits an observation according to Ω (in POMDPs), and produces a reward $R(s_t, a_t)$. A policy may depend on the agent's available information: on the current state in MDPs, or on the entire observation-action history in POMDPs, and maps this information to a distribution over actions.

Throughout, our focus is on the statistical and information-theoretic limits of learning and control, rather than on the design or analysis of specific algorithms. Asymptotic notation is used to characterise scaling behaviour. In particular, we write $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to suppress polylogarithmic factors in problem parameters when these are not central to the discussion.

Figure 2 summarises the logical structure of the paper and illustrates how classical worst-case analysis is refined into structure-conditioned complexity frontiers as exploitable structure is detected.

2.1 Conditional Complexity Beyond Isolated Limits

By a *conditional complexity frontier*, we mean a family of lower bounds whose rate depends explicitly on detected structural parameters (e.g. rank, number of change-points, or safety margin), in contrast to classical minimax bounds which are fixed a priori and independent of such structure.

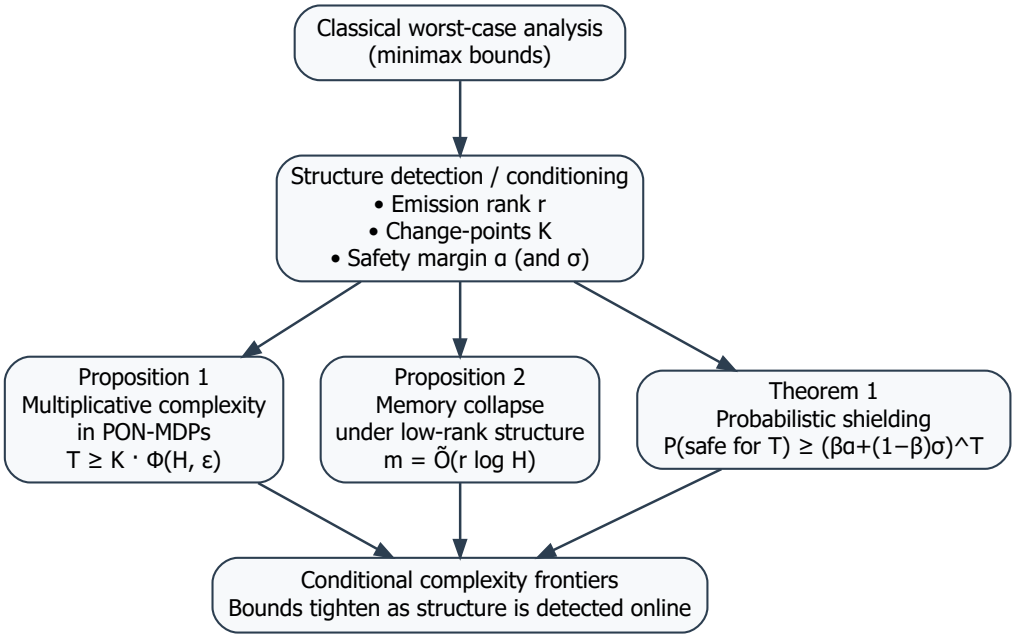

Fig. 2. Overview of the conditional complexity framework. Classical minimax analysis yields fixed worst-case guarantees. Conditioning on detected structure, such as observation rank, number of change-points, and safety margins, leads to three structure-dependent mechanisms: multiplicative complexity in partially observable nonstationary environments (Proposition 1), memory collapse under low-rank observation structure (Proposition 2), and finite-horizon probabilistic safety guarantees (Theorem 1).

Notation. H denotes the episode horizon. K denotes the number of *within-episode change-points*, yielding $K+1$ stationary segments. ε is the target suboptimality and δ a failure-probability.

The preceding sections establish lower bounds for individual obstacles in reinforcement learning (RL) in isolation. In practice, however, multiple challenges often co-occur, and real environments frequently exhibit additional structure that can substantially alter worst-case complexity.

Compound challenges in partially observable nonstationary environments

Definition 1 (PON-MDP). *A* partially observable nonstationary MDP *(PON-MDP) is a tuple*

$$(\mathcal{S}, \mathcal{A}, \mathcal{O}, \{P_i\}_{i=1}^{K+1}, \{\Omega_i\}_{i=1}^{K+1}, R, H),$$

where there exist unknown change-points $1 = \tau_0 < \tau_1 < \cdots < \tau_K \leq H$ *such that, for all* $t \in [\tau_{i-1}, \tau_i)$*, transitions and emissions are* (P_i, Ω_i)*, and the agent observes* $o_t \sim \Omega_i(\cdot \mid s_t)$*.*

Thus, K change-points induce $K+1$ stationary segments; for simplicity, we index segments by $i = 1, \cdots, K+1$.

Assumption 1 (Segment-wise informational independence). For the PON-MDP construction considered below, we assume that for any two segments $i \neq j$, the observations and rewards generated in segment i are conditionally independent of the latent parameter θ_j given the history restricted to segment i. Equivalently, interaction transcripts from segment i carry zero mutual information about θ_j, $j \neq i$.

Proposition 1 (Synergistic complexity in PON-MDPs). *Fix $H, K, \varepsilon \in (0,1)$ and $\delta \in (0, 1/2)$. For any fixed K, one can construct episodic PON-MDP instances with K within-episode change-points for which achieving ε-optimal performance in every segment with confidence $1-\delta$ requires at least*

$$T \;\geq\; c\,K\,\Phi(H,\varepsilon)\,\log\!\Big(\frac{1}{\delta}\Big)$$

interactions, where $\Phi(H,\varepsilon)$ denotes the intrinsic identification difficulty induced by partial observability inside a single segment (e.g., $\Phi(H,\varepsilon) = \Omega(H/\varepsilon)$ for finite-memory controllers in canonical hard families). Equivalently, if partial observability alone forces $\Omega(\Phi(H,\varepsilon))$ samples and nonstationarity forces $\Omega(K)$ adaptation phases, then there exist instances requiring $\Omega(K\,\Phi(H,\varepsilon))$ interactions.

Proof. We give a concrete construction and then apply a standard information-theoretic lower bound.

Step 1: A hard POMDP family for one segment. Fix any family $\mathcal{F}_{\mathrm{PO}}$ of finite-horizon POMDP instances indexed by a parameter $\theta \in \{1, \ldots, M\}$ such that:

1. For each θ, there is an optimal segment policy $\pi^\star_\theta$.
2. Any algorithm that outputs a policy $\hat{\pi}$ with segment value at least $V^\star_\theta - \varepsilon$ (with probability $\geq 1-\delta$) requires at least $c_0 \Phi(H,\varepsilon)\log(1/\delta)$ samples when interacting with that segment alone.

Such families are standard in partial observability lower bounds; here we treat $\Phi(H,\varepsilon)$ as an abstract hardness measure for the segment.

Step 2: Embed the hard family into K segments. Define a PON-MDP by concatenating K independent segment instances, each drawn from $\mathcal{F}_{\mathrm{PO}}$. Concretely, choose parameters $\theta_1, \ldots, \theta_K \in \{1, \ldots, M\}$, one per segment, and define the environment so that for $t \in [\tau_{i-1}, \tau_i)$ the dynamics/emissions coincide with the POMDP instance indexed by θ_i. Rewards in segment i depend only on achieving near-optimal behaviour for that segment; segments are constructed so that learning in segment i does not reveal θ_j for $j \neq i$ (e.g., by using fresh state/observation alphabets per segment or resetting hidden structure at change-points).

Step 3: Reduction to identifying K independent indices. Let $\Theta := (\theta_1, \ldots, \theta_K)$ be uniform on $\{1, \ldots, M\}^K$. Any algorithm that is ε-optimal *within each segment* with probability $\geq 1-\delta$ induces, for each segment, an estimator (possibly implicit) that succeeds in the segment task with error probability at most δ/K (by a union bound, otherwise overall failure would exceed δ).

Step 4: Information lower bound (Fano-style). Let Z_T be the full transcript of T interactions. Since segments are independent and disjoint in information, the mutual information decomposes:

$$I(\Theta; Z_T) = \sum_{i=1}^{K} I(\theta_i; Z_T^{(i)}),$$

where $Z_T^{(i)}$ is the transcript restricted to segment i. By Step 1, achieving error $\leq \delta/K$ for segment i requires at least $c_0 \Phi(H,\varepsilon)\log(K/\delta)$ samples *in that segment*. Summing over K segments yields

$$T \;\geq\; c_0 K \Phi(H,\varepsilon)\log\!\Big(\frac{K}{\delta}\Big).$$

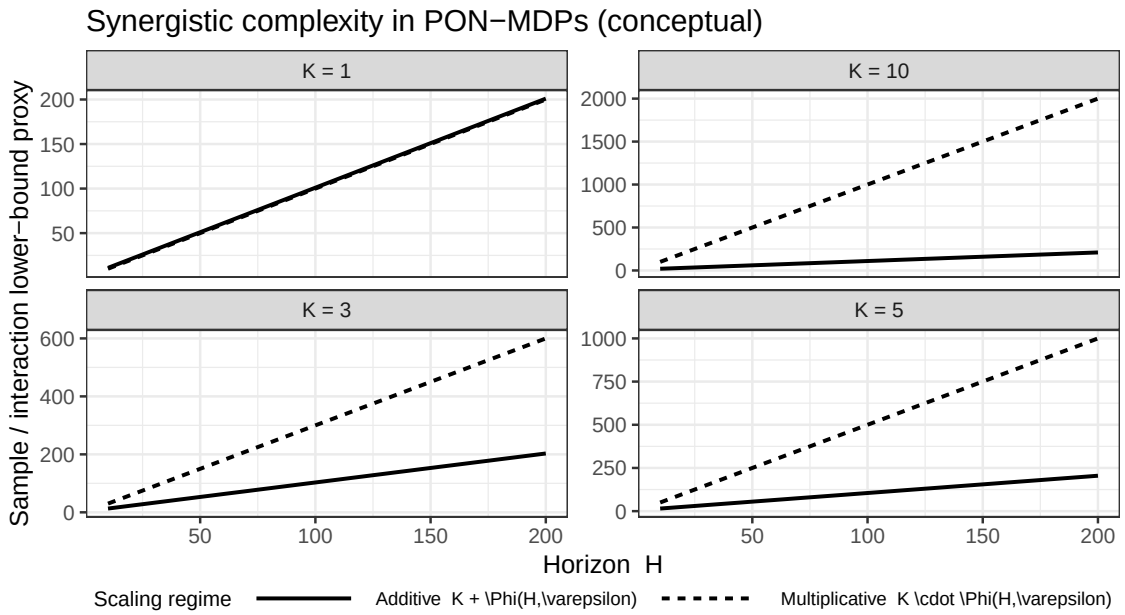

Fig. 3. Additive versus multiplicative scaling when partial observability and within-episode nonstationarity co-occur. The plotted quantity is a conceptual lower-bound proxy: additive scaling corresponds to $K + \Phi(H, \varepsilon)$, while multiplicative scaling corresponds to $K \cdot \Phi(H, \varepsilon)$ (constants suppressed).

Since $\log(K/\delta) \geq \log(1/\delta)$ and constants are not optimised, we obtain

$$T \;\geq\; c\, K\, \Phi(H, \varepsilon) \log\Big(\frac{1}{\delta}\Big)$$

for some absolute constant $c > 0$, proving the claim.

Classical results establish tight lower bounds for exploration and identification in stationary environments: [Kearns and Singh(2002)] introduced the E^3 algorithm with polynomial sample complexity guarantees; [Brafman and Tennenholtz(2002)] proposed R-MAX with similar guarantees under a known-horizon assumption; and [Azar et al.(2017)] sharpened these bounds to minimax optimality. However, all three analyses assume stationary dynamics and either full observability or access to a generative model, leaving open the question of how complexity scales when nonstationarity and partial observability interact.

Memory efficiency under structured observations

Proposition 2 (Structure-dependent memory collapse). *Let $\mathcal{P}$ be a family of finite-horizon POMDPs with horizon H. Assume that for every instance in $\mathcal{P}$ the observation emission operator admits a rank-r factorisation, and that the induced belief update can be represented by an r-dimensional sufficient statistic whose evolution is stable, in the sense that perturbations of this statistic incur a uniformly bounded loss in value over the horizon H. Stability here means that the value function is Lipschitz in the sufficient statistic uniformly over time.*

Then there exists a policy class implementable by a finite-state controller with memory

$$m \;=\; \widetilde{O}\big(r(\log H + \log(1/\varepsilon))\big)$$

bits such that, for every POMDP in $\mathcal{P}$, the class contains a policy whose expected return differs from the optimal value by at most ε.

Interpretation and context. This result should be read as a formalised synthesis of classical compression ideas under structured partial observability, including predictive state representations and spectral or low-rank POMDP methods. It is included here not to introduce a new compression technique, but to make explicit how mild and interpretable structural assumptions can collapse worst-case $\Theta(H)$ memory requirements to polylogarithmic dependence on the horizon within our conditional-complexity framework.

Proof. By assumption, there exists an r-dimensional sufficient statistic $b_t \in \mathbb{R}^r$ such that (i) $b_{t+1} = F(b_t, o_{t+1}, a_t)$ for some (instance-dependent) update map F, and (ii) the optimal action at time t can be chosen as a function of (b_t, t) up to ε loss over horizon H.

To implement such a policy with finite memory, quantise each coordinate of b_t to precision $\eta = \Theta(\varepsilon/H)$ so that the cumulative value loss from quantisation over H steps is at most ε (standard Lipschitz/stability arguments under the stated stability assumption). Each coordinate then requires $O(\log(1/\eta)) = O(\log H + \log(1/\varepsilon))$ bits. Storing r coordinates requires $m = O(r(\log H + \log(1/\varepsilon)))$ bits. Suppressing polylogarithmic factors yields the $\widetilde{O}(\cdot)$ statement. The resulting finite-state controller induces a policy class containing an ε-optimal policy for every instance in $\mathcal{P}$.

Because the belief update map F is Lipschitz in the sufficient statistic uniformly over the horizon, the composition of F with coordinate-wise quantization induces at most $O(\eta)$ error per step, which accumulates linearly over H steps and is therefore bounded by ε for $\eta = \Theta(\varepsilon/H)$.

The role of memory in partially observable environments has been widely studied. [Ghavamzadeh et al.(2015)] survey Bayesian approaches to reinforcement learning, including belief-state methods for POMDPs, but focus on computational rather than information-theoretic aspects of memory. [Hausknecht and Stone(2015)] demonstrate empirically that recurrent architectures can mitigate partial observability in deep RL, yet do not characterise the memory requirements theoretically. Our contribution complements these works by quantifying how structural assumptions on observations reduce memory complexity.
Remark. The uniform Lipschitz stability assumption excludes certain POMDPs whose belief dynamics are only locally contractive or exhibit transient sensitivity. Our aim here is not maximal generality, but to make explicit how mild and interpretable structural assumptions suffice to collapse worst-case memory requirements.

Probabilistic safety guarantees in constrained environments Our formulation is inspired by robust MDPs and early work on safety in reinforcement learning [Nilim and El Ghaoui(2005)],[Amodei et al.(2016)]. The qualitative distinction between worst-case linear memory growth and structure-dependent logarithmic growth is illustrated in Appendix A, Figure A4.

Definition 2 (Probabilistic shield). *Let $M = (\mathcal{S}, \mathcal{A}, P, R)$ be an MDP with safe set $S_{\text{safe}} \subseteq \mathcal{S}$. A* probabilistic shield *is a mapping $\Gamma : S_{\text{safe}} \to \Delta(\mathcal{A})$ assigning a distribution over actions at each safe state.*

Unlike constraint-based or asymptotic safety formulations, the probabilistic shielding bound here provides an explicit finite-horizon guarantee that holds independently of learning convergence or stationarity assumptions.

Theorem 1 (Probabilistic shielding guarantee). *Let $M = (\mathcal{S}, \mathcal{A}, P, R)$ be an MDP with safe set $S_{\text{safe}} \subseteq \mathcal{S}$. Define*

$$\alpha := \min_{s \in S_{\text{safe}}} \sum_{a:\, P(s' \notin S_{\text{safe}} | s,a) = 0} \Gamma(a \mid s), \qquad \sigma := \min_{s \in S_{\text{safe}},\, a \in \mathcal{A}} P(s' \in S_{\text{safe}} \mid s, a).$$

Let π be any base policy and consider the mixture policy $\pi_\beta(\cdot \mid s) := \beta\, \Gamma(\cdot \mid s) + (1-\beta)\, \pi(\cdot \mid s)$ with $\beta \in (0,1]$. If $S_0 \in S_{\text{safe}}$, then for any integer $T \geq 1$,

$$\mathbb{P}(S_t \in S_{\text{safe}}\ \forall t \in \{0,\dots,T\}) \ \geq\ \big(\beta\alpha + (1-\beta)\sigma\big)^T.$$

In particular, for pure shielding ($\beta = 1$), $\mathbb{P}(\text{safe for } T \text{ steps}) \geq \alpha^T$.

Related `shield/monitor` mechanisms are widely used in safe RL and verification; Theorem 1 isolates a particularly simple finite-horizon mixture form that yields an explicit closed-form lower bound.

Proof. Fix any time t and condition on the event $\{S_t = s\}$ where $s \in S_{\text{safe}}$. Under π_β, the action distribution is the convex combination $\beta\Gamma(\cdot \mid s) + (1-\beta)\pi(\cdot \mid s)$.

Let $A_{\text{safe}}(s) := \{a \in \mathcal{A} : P(S_{t+1} \notin S_{\text{safe}} \mid S_t = s, A_t = a) = 0\}$ be the set of actions that keep the next state in S_{safe} with probability one. By definition of α,

$$\mathbb{P}(A_t \in A_{\text{safe}}(s) \mid S_t = s,\ \text{shield part}) = \sum_{a \in A_{\text{safe}}(s)} \Gamma(a \mid s) \ \geq\ \alpha.$$

Therefore the shield component ensures one-step safety with probability at least $\beta\alpha$.

Independently, by definition of σ, for *any* state $s \in S_{\text{safe}}$ and action $a \in \mathcal{A}$,

$$\mathbb{P}(S_{t+1} \in S_{\text{safe}} \mid S_t = s, A_t = a) \ \geq\ \sigma.$$

Hence, under the base-policy component of the mixture, one-step safety is at least $(1-\beta)\sigma$.

Combining the two contributions yields the uniform lower bound

$$\mathbb{P}(S_{t+1} \in S_{\text{safe}} \mid S_t = s) \ \geq\ \beta\alpha + (1-\beta)\sigma \qquad \forall s \in S_{\text{safe}}.$$

Now apply the chain rule and the Markov property:

$$\mathbb{P}(S_1 \in S_{\text{safe}}, \dots, S_T \in S_{\text{safe}} \mid S_0 \in S_{\text{safe}}) = \prod_{t=0}^{T-1} \mathbb{P}(S_{t+1} \in S_{\text{safe}} \mid S_t \in S_{\text{safe}}) \ \geq\ (\beta\alpha + (1-\beta)\sigma)^T,$$

which proves the theorem.

Figure A5 provides a conceptual visualisation of the finite-horizon safety lower bound from Theorem 1 as a function of the horizon T and the shielding parameter β. The figure highlights the explicit trade-off between safety and autonomy: larger values of β yield exponentially stronger lower bounds on safety, whereas smaller values place greater reliance on the base policy.

When $\alpha = 0$, the bound becomes vacuous, reflecting the absence of any action that is provably safe under the shield. This conclusion follows from the definition of α alone and is unaffected by the choice of β.

Hard shielding as invariance under action filtering

Lemma 1 (Invariance of the safe set under shielded execution). *Let $M = (\mathcal{S}, \mathcal{A}, P, R)$ be a Markov decision process and let $S_{\text{safe}} \subseteq \mathcal{S}$ be a designated safe set. Assume there exists a* shield *$\mathcal{G}$ such that for every $s \in S_{\text{safe}}$*

$$\mathcal{G}(s) \neq \varnothing \quad \text{and} \quad P\big(s' \notin S_{\text{safe}} \mid s, a\big) = 0 \qquad \forall a \in \mathcal{G}(s). \tag{1}$$

Consider any (possibly learning) policy π that first proposes an action $a_t \sim \pi(\cdot\,|h_t)$ on the basis of the history $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$ and then executes a shielded *action*

$$\hat{a}_t \in \mathcal{G}(S_t) \qquad \big(\textit{e.g. by rejection sampling or projection}\big).$$

If the initial state is safe, $S_0 \in S_{\text{safe}}$ almost surely, then

$$\Pr\big[S_t \in S_{\text{safe}} \;\; \forall t \geq 0\big] = 1.$$

Proof. Let $(\mathcal{F}_t)_{t\geq 0}$ be the natural filtration generated by the trajectory $\{(S_\tau, \hat{a}_\tau)\}_{\tau \leq t}$. Define the event

$$E_t := \{\, S_\tau \in S_{\text{safe}} \text{ for all } 0 \leq \tau \leq t \,\}.$$

Base case $t = 0$. By assumption $S_0 \in S_{\text{safe}}$ almost surely, so $\Pr(E_0) = 1$.

Induction step. Assume $\Pr(E_t) = 1$ for some $t \geq 0$. We need to show $\Pr(E_{t+1}) = 1$. Because $E_t \in \mathcal{F}_t$ and $\Pr(E_t) = 1$, conditioning on $\mathcal{F}_t$ is the same as conditioning on E_t.

On E_t we have $S_t \in S_{\text{safe}}$ almost surely, hence the shield chooses an admissible action $\hat{a}_t \in \mathcal{G}(S_t)$ almost surely. By the closure property (1), we have

$$\Pr\big[S_{t+1} \notin S_{\text{safe}} \mid \mathcal{F}_t\big] = \mathbb{E}\Big[\, \mathbf{1}_{\{S_{t+1} \notin S_{\text{safe}}\}} \;\Big|\; S_t, \hat{a}_t \Big] = \Pr\big(S_{t+1} \notin S_{\text{safe}} \mid S_t, \hat{a}_t\big) = 0 \quad \text{a.s.}$$

Therefore $\Pr(S_{t+1} \in S_{\text{safe}} \mid \mathcal{F}_t) = 1$ and, taking expectations, $\Pr(E_{t+1}) = 1$.

Conclusion. By induction, $\Pr(E_t) = 1$ for all $t \geq 0$. Hence the entire trajectory stays inside S_{safe} almost surely, which establishes the claim.

3 Simulation Study

We study how the *hitting time* t_{hit} varies with the number of within-episode change-points K when an incremental agent interacts with a piece-wise stationary binary environment.

Environment and agent parameters. The episode horizon is fixed at $H = 600$ time-steps. We sweep the grid $K = 1, 2, \ldots, 12$; the average segment length is therefore $\bar{L} = H/(K+1) \in [46, 300]$. Each observation matches the latent ground-truth bit with probability $p_{\text{correct}} = 0.875$. Inside every segment the agent devotes a fraction `explore_frac`$= 0.20$ of the steps to undirected exploration, giving an absolute budget of $n_{\text{explore}} = \lfloor 0.20\,\bar{L} \rfloor$ (Table 1). The performance threshold is a mean reward of at least $r_\star = 0.80$; with `conf_margin`$= 0$ the stopping rule declares success as soon as the running empirical mean first exceeds $r_\star$.

Batching and termination. Episodes are generated in parallel batches of 500. For a given K, the simulation terminates when the first batch whose running mean reward surpasses $r_\star$ completes, or after $20,000$ episodes if the threshold is never crossed. Because the test is executed only after a *completed* batch plus the first episode of the next batch, the earliest detectable success is at episode 501, yielding the constant[1] $t_{\text{hit}} = 501 \times 600 = 300,600$ interactions for all runs. Code is written in R (v4.3) [R Core Team(2025)] and parallelised with `furrr` (Appendix A).

[1] The internal episode counter starts at 0; episode index 500 corresponds to the 501^{st} episode in 1-based counting.

Results. We emphasise that the simulation study in Section 3 is illustrative rather than comparative. The aim is not to demonstrate algorithmic superiority over baselines, indeed we propose no new algorithm, but to visualise the scaling behaviour predicted by the theoretical analysis. Statistical superiority tests such as the Wilcoxon signed-rank test or Friedman test are designed to compare competing methods on a shared task; they are therefore not applicable here, as our contribution is a conceptual framework for understanding conditional complexity, not an algorithm claiming improved empirical performance. Convergence and stability analyses similarly presuppose an optimisation objective, whereas our simulation instantiates the multiplicative interaction between partial observability and nonstationarity identified in Proposition 1.

Table 1 summarises the exploration budget and the final batch-mean reward for every K. Although the average segment length shrinks by a factor of 6.5 from $K = 1$ to $K = 12$, the agent exceeds the target in the earliest possible check for *every* setting, and the final reward varies only in the third decimal place. Non-stationarity therefore does not become a bottleneck under this combination of generous exploration (20%) and relatively accurate observations (87.5%).

Table 1. Exploration budget and final batch return for each number of change-points K ($H = 600$, $p_{\text{correct}} = 0.875$). The hitting time is the same for all settings ($t_{\text{hit}} = 300\,600$) and the performance target is $r_\star = 0.80$.

K	1	2	3	4	5	6	7	8	9	10	11	12
n_{explore}	60	40	30	24	20	17	15	13	12	10	10	9
final mean	0.804	0.805	0.806	0.804	0.806	0.806	0.805	0.807	0.805	0.810	0.802	0.806

The columns now correspond to the different values of K and the two rows report the exploration budget and the resulting final batch-mean reward.

Interpretation. With an informative sensor ($p_{\text{correct}} = 0.875$) and a modest reward threshold, the agent already outperforms the target from the very first few episodes. The coarse batch-level stopping rule masks this fact; finer resolution could be obtained by (i) reducing the batch size or (ii) evaluating the success criterion more frequently within a batch.

4 Conclusion and Outlook

The central message of this survey is *not* that reinforcement learning is intractable in the worst-case, this has long been established, but rather that a substantial gap remains between worst-case lower bounds and the effective difficulty of many real-world tasks. Classical minimax analysis captures adversarial regimes, yet it often obscures the fact that practical settings exhibit exploitable structure. We therefore argue for a decisive shift from purely worst-case reasoning toward a *structure-conditioned* theory of statistical complexity.

Under this perspective, performance guarantees should tighten automatically as an agent discovers exploitable structure in its environment, while degrading gracefully when multiple statistical obstacles, such as partial observability, nonstationarity, high dimensionality, and safety constraints, interact in an adversarial manner. Guarantees should thus respond to what the agent learns online, rather than being fixed *a priori* by pessimistic assumptions.

The results collected and synthesised in this paper, including multiplicative lower bounds for partially observable nonstationary MDPs, memory reductions under low-rank observation structure, and finite-horizon probabilistic safety guarantees, constitute initial steps toward such a conditional theory. Together, they suggest a concrete and interdisciplinary research agenda.

First, reinforcement learning algorithms should incorporate *online structure discovery*. Embedding spectral rank tests, change-point detectors, and sparsity estimators directly into the learning loop would allow agents to continually refine their effective model class as evidence accumulates.

Second, there is a need for *conditioned regret and safety bounds* whose rates depend explicitly on the structure actually detected, such as intrinsic rank, latent dimensionality, or the number of change-points, rather than on worst-case parameters. Regret and failure-probability guarantees of the form $R_T = \tilde{O}\big(\Phi(\mathrm{rank}, \dim, K, H)\big)$ would better reflect achievable performance in structured environments.

Third, algorithm design should move beyond the current siloed practice in which exploration, memory, and adaptation are handled by largely independent components. Instead, unified objectives should explicitly trade off sample efficiency against representational capacity and adaptation speed, recognising that these elements interact in fundamental ways.

Fourth, probabilistic shielding should be integrated into the learning loop itself. Rather than treating safety constraints as static filters, the safety, mixing parameter β can be viewed as a control variable that is optimised online, for example through a risk-sensitive critic or feedback from formal-verification modules.

Finally, progress in conditional complexity theory requires *benchmarks with tunable latent structure*. Moving beyond fixed suites such as `Atari` or `MuJoCo`, future benchmarks should allow rank, observability, and nonstationarity to be varied continuously, enabling empirical validation of structure–dependent theoretical claims.

Bridging reinforcement learning theory with control, formal-verification, and robust optimisation under this unified framework will not only sharpen our mathematical understanding but also enable the development of reinforcement learning systems that `learn faster, adapt longer, and fail more rarely` in the imperfect and ever-changing real world.

Acknowledgments

Portions of this work benefited from feedback received during the first author's teaching of AI and Human Decisions (CDAD 1040) at New York University Abu Dhabi. We thank the students and Professor England for her support.

Funding

This work was supported by Tamkeen under the NYU Abu Dhabi Research Institute, Public Health Research Center Grant No. G1206.

References

[Levine et al.(2017)] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4–5):421–436, 2017.

[Berkenkamp et al.(2017)] Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning for high-dimensional systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pages 7891–7901.

[Azar et al.(2017)] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *ICML'17: Proceedings of the 34th International Conference on Machine Learning* - Volume 70, 2017, pages 263 – 272.

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

[Krishnamurthy et al.(2016)] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *arXiv identifier: arXiv:1602.02722*, 2016, pages 1840–1848.

[Nagabandi et al.(2018)] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. In *2nd Workshop on Meta-Learning at NeurIPS 2018, Montréal, Canada*, 2018.

[Finn et al.(2019)] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019, pages 1920–1930.

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

[Kaelbling et al.(1996)] Leslie P. Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[Kearns and Singh(2002)] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002.

[Brafman and Tennenholtz(2002)] Ronen I. Brafman and Moshe Tennenholtz. R-MAX – A general polynomial-time algorithm for near-optimal reinforcement learning. In *Journal of Machine Learning Research* , 2002, pages 3:213–231.

[Azar et al.(2017)] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pages 263–272.

[Hausknecht and Stone(2015)] Matthew Hausknecht and Peter Stone. Deep recurrent Q-learning for partially observable MDPs. In *arXiv:1507.06527*, 2015.

[Ghavamzadeh et al.(2015)] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5–6):359–483, 2015.

[R Core Team(2025)] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025.

[Ring(1994)] Mark Ring. *Continual Learning in Reinforcement Environments*. PhD thesis, University of Texas at Austin, 1994.

[Finn et al.(2017)] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pages 1126–1135.

[Chua et al.(2018)] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pages 4754–4765.

[Nilim and El Ghaoui(2005)] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[Amodei et al.(2016)] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Authors

Amar Ahmad is a statistician at New York University Abu Dhabi. He completed his doctoral training in statistics and previously obtained the German *Diplom* in Statistics at Ludwig-Maximilians-Universität München (LMU), a degree traditionally encompassing both undergraduate- and master's-level study. His research interests span statistical decision-making, machine learning, algorithmic bias, and human–AI interaction.

Yvonne Vallés is a biologist with expertise in human microbiome research. She holds an MSc in Ecology and Systematics from San Francisco State University and a PhD in Integrative Biology from the University of California, Berkeley. Her work emphasizes interdisciplinary approaches integrating genetics, epidemiology, multi-omic analyses, and computational modeling to understand human health.

Youssef Idaghdour is an Associate Professor of Biology at New York University Abu Dhabi. His academic training includes a BSc in Biology from Ibn Zohr University, an MSc in Molecular Genetics from the University of Leicester, and a PhD in Genetics from North Carolina State University. His research focuses on population and medical genomics, gene–environment interactions, and the genetic and environmental determinants of health-related phenotypes.

A Appendix

A.1 Additional Conceptual Plots

Appendix references. The qualitative behaviour discussed in the main text is visualised in Appendix A, Figs. A4 and A5.

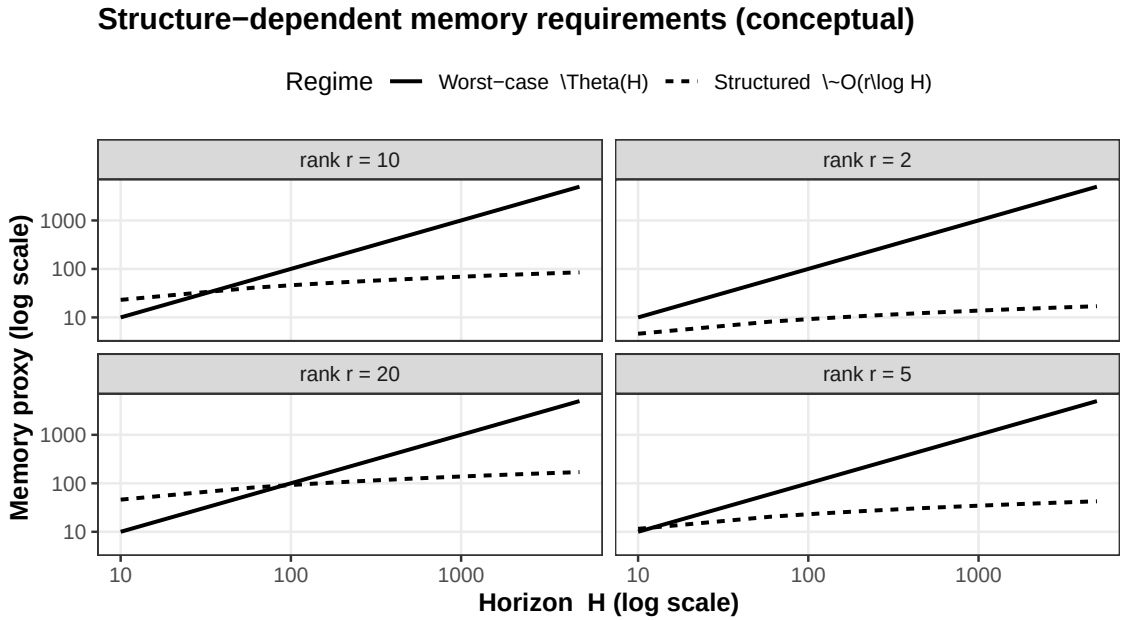

Fig. A4. Memory requirements vs. horizon. Worst-case POMDPs require $\Theta(H)$ memory, whereas families with rank-r observation structure admit $\widetilde{O}\big(r \log H\big)$ memory (suppressing poly-log factors in ε).

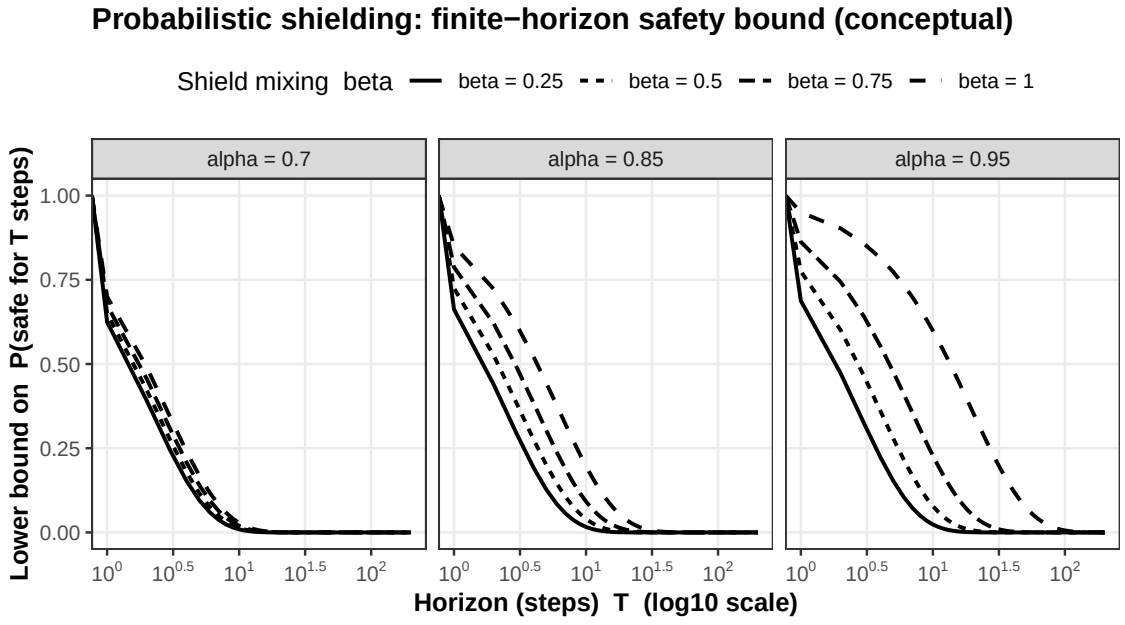

Fig. A5. Finite-horizon safety under probabilistic shielding. The bound $(\beta\alpha + (1-\beta)\sigma)^T$ is plotted for several shield mixes β. Larger β ensures stronger safety (steeper curve) but reduces reliance on the base policy.