Anytime Posterior Sampling under Nested CVaR Constraints

Abstract

We develop an anytime posterior-sampling algorithm for finite episodic Markov decision processes with unknown transitions and a separate nested upper-tail CVaR constraint. The method extends the bounded-violation posterior-sampling framework of Kalagarla et al. [18] beyond additive expected-cost constraints to time-consistent tail risk. This extension faces a structural obstruction: when action randomization lies inside CVaR, generic threshold tightening can cause a constant value loss and a strict scalar duality gap. Our solution is a risk bank that sparsely executes minimum-risk policies and uses their sampled-model slack to support untightened constrained planning. With known rewards and costs, model-wise strict feasibility on the support of an arbitrary prior, and exact posterior-sampling and constrained-planning oracles, the learner needs neither a common feasible baseline, the induced feasibility gap, nor the learning horizon. For fixed problem parameters, we prove a $\widetilde{O}(\sqrt{K})$ upper bound on signed Bayesian reward shortfall after K episodes, a K-independent upper bound on expected signed risk-value debt, and a sublinear number of calibration episodes. The comparator is a randomized physical-state Markov policy, and risk control is a signed aggregate guarantee rather than per-episode safety. A support-size-independent CVaR–Hellinger inequality and distorted-occupancy analysis quantify the dependence on tail mass and horizon. We also establish a fixed-K high-probability certificate, a two-model lower bound attaining the worst-case rare-event risk-debt exponent, and NP-completeness of a one-state planning subclass at every fixed rational tail mass in $(0, 1)$.

Note. This paper was generated entirely by AI, including MiMo, using an automated research pipeline developed by Chenghua Liu and Hanyu Li.

1 Introduction

This paper asks whether posterior sampling can learn under a separate, time-consistent tail-risk constraint without being given a common feasible policy, a numerical feasibility gap, or the number of episodes. The challenge is structural: nested CVaR is nonlinear in randomized actions and evaluates model error under a tail-distorted path law rather than the nominal visitation law that produces data.

Posterior sampling is well understood for unconstrained MDPs [26, 27]. For additive expected-cost constraints, Kalagarla et al. [18] obtain near-square-root Bayesian reward regret and bounded signed violation without requiring a supplied safe policy. Their linear mixing, vanishing-tightening, and additive simulation arguments do not transfer to the nested-risk setting. We therefore work in an explicit oracle model: exact draws from the conditional posterior and exact solutions of the known-model constrained problem are available.

The comparator is a randomized Markov policy on the physical state, with its action randomization inside each CVaR operator. This choice is deliberate. A policy that is Markov in an augmented risk-budget state is generally history-dependent relative to the physical state, while drawing one deterministic policy per episode places the randomization outside CVaR. Both alternatives change the feasible set.

Our contributions are:

- For every $\alpha \in (0,1)$, an interior one-stage construction gives a constant value jump under arbitrarily small positive tightening and a strict scalar duality gap. Thus tightening cannot serve as a generic every-episode device in this policy class.
- An anytime risk bank uses computed sampled-model slack to decide when to run a polynomial-time minimum-risk calibration policy. Under model-wise strict feasibility, it gives explicit one-sided bounds: signed Bayesian reward shortfall is near $\sqrt{K}$, expected signed risk-value debt has a K-independent upper bound, and $N_K^{\mathrm{cal}} \le (T_K + \tau)/\Delta_*$ almost surely. For a fixed problem, $T_K = \widetilde{O}(W_{\alpha,H} S\sqrt{AK})$, so calibration is sparse.
- We prove a support-size-independent CVaR–Hellinger inequality and combine it with distorted occupancy and posterior symmetry. Within this proof framework, Hellinger confidence saves one half-power of the tail mass relative to total variation and yields product weights with the correct worst-case rare-event exponent.
- For each fixed rational $\alpha \in (0,1)$, rational one-state planning is NP-complete; the rational general problem lies in $\exists\mathbb{R}$. A two-model rare-chain family matches the risk-debt exponent at a tailored episode count. With a common fallback, it yields a Pareto lower bound rather than simultaneous lower bounds on both coefficients.

Relation to prior work. Recursive risk Bellman equations and risk-budget augmentation are classical [3, 10, 11, 32]. Online work on iterated CVaR or recursive optimized certainty equivalents optimizes a risk-sensitive objective rather than expected reward under a separate nested-risk constraint [8, 13, 15, 22, 35]. Separate-constraint work instead studies static trajectory CVaR, entropic utility, one-step optimized certainty equivalents, or known-model policy optimization, with different protocols and guarantees [2, 9, 17, 19, 21, 25, 36].

For additive CMDPs, optimistic, primal–dual, model-free, and posterior-sampling methods give sublinear reward and constraint regret under a variety of episodic or average-reward assumptions [1, 5, 6, 14, 16, 18, 24, 28, 33]. Risk-sensitive Bayes-adaptive objectives and risk over model uncertainty form another distinct interface [23, 29]. A recent forecast-driven framework combines posterior sampling with a static return-CVaR objective or chance constraint in a belief-augmented planner, not the nested constraint and signed-debt theorem studied here [20]. To our knowledge, prior work does not simultaneously study a separate nested-CVaR constraint, this physical-state Markov comparator, posterior-sampling interaction, near-$\sqrt{K}$ expected-reward shortfall, and a K-independent upper bound on expected signed risk-value debt. This is a scoped distinction, not an absolute priority claim; Table 2 compares the relevant axes explicitly.

2 Problem formulation and performance criteria

We first specify the MDP, the physical-state Markov comparator, and the nested CVaR functional. We then state the model-wise feasibility assumption and the two signed performance criteria controlled by the algorithm.

Let $\mathcal{S}$ and $\mathcal{A}$ be finite sets of cardinalities S and A. An episode has stages $h = 1, \dots, H$ and starts at a fixed state s_1. The reward $r_h(s,a)$ and cost $c_h(s,a)$ are known and belong to $[0,1]$. We allow $0 < \tau \le H$. Because $0 \le \rho_1^{\pi,p}(s_1) \le H$, the endpoint $\tau = H$ makes the original constraint vacuous; strict feasibility below already forces $\tau > 0$. The time-homogeneous transition kernel

$p(\cdot \mid s, a)$ is drawn once from a prior μ on the finite-dimensional kernel space and is then fixed. Write $\mathcal{Q} := \mathrm{supp}(\mu)$. The prior may be arbitrary: it need not be conjugate, product-form, or row-factorized.

The policy class is

$$\Pi_{\mathrm{M}} := \prod_{h=1}^{H} \prod_{s \in \mathcal{S}} \Delta(\mathcal{A}),$$

the randomized Markov policies on the physical state. For $\pi \in \Pi_{\mathrm{M}}$, let $V_h^{\pi,p}(s)$ be its expected reward-to-go, with

$$V_{H+1}^{\pi,p} = 0, \qquad V_h^{\pi,p}(s) = \sum_a \pi_h(a \mid s) \left[r_h(s,a) + \sum_{s'} p(s' \mid s, a) V_{h+1}^{\pi,p}(s') \right]. \tag{2.1}$$

We use upper-tail CVaR for losses, with $\alpha \in (0, 1]$ denoting tail mass. For a probability law P on a finite outcome space Ω,

$$\mathrm{CVaR}_{\alpha,P}^{\mathrm{up}}(X) = \inf_{z \in \mathbb{R}} \left\{ z + \frac{1}{\alpha} \mathbb{E}_P[(X - z)_+] \right\} = \max_{\substack{\xi : \Omega \to [0, 1/\alpha] \\ \mathbb{E}_P \xi = 1}} \mathbb{E}_P[\xi X]. \tag{2.2}$$

The first identity is the Rockafellar–Uryasev representation and the second is its finite-dimensional risk-envelope dual [30, 31]; the second equality also follows directly from finite-dimensional linear-programming duality. Define nested risk by

$$\rho_{H+1}^{\pi,p} = 0, \qquad \rho_h^{\pi,p}(s) = \mathrm{CVaR}_{\alpha}^{\mathrm{up}}\big(c_h(s, A_h) + \rho_{h+1}^{\pi,p}(S_{h+1})\big), \tag{2.3}$$

where, inside every CVaR operator, $A_h \sim \pi_h(\cdot \mid s)$ and $S_{h+1} \sim p(\cdot \mid s, A_h)$. Thus action randomization is itself part of the tail-risk distribution; it is not applied outside the risk operator.

We assume only model-wise strict feasibility on the prior support. Define $g(q) := \min_{\pi \in \Pi_{\mathrm{M}}} \rho_1^{\pi,q}(s_1)$ and suppose $g(q) < \tau$ for every $q \in \mathcal{Q}$.

Lemma 2.1 (Automatic uniform margin)**.** *Model-wise strict feasibility induces the positive uniform gap*

$$\Delta_* := \tau - \max_{q \in \mathcal{Q}} g(q) > 0. \tag{2.4}$$

Proof sketch. Backward induction and the primal CVaR representation make nested risk continuous in (π, q). Berge's theorem then makes g continuous; it attains its maximum on compact $\mathcal{Q}$, where model-wise strict feasibility makes the gap positive. The complete argument is in Appendix A.1.

The minimizing policy may depend on q. No single policy is assumed safe for every model, no safe policy is supplied, and the algorithm is not given Δ_* or any lower bound on it. The quantity Δ_* appears only in performance bounds.

For any kernel q with a nonempty feasible set (in particular, every $q \in \mathcal{Q}$), write

$$\mathcal{V}_\tau(q) := \max_{\pi \in \Pi_{\mathrm{M}} :\, \rho_1^{\pi,q}(s_1) \le \tau} V_1^{\pi,q}(s_1). \tag{2.5}$$

After K episodes, an algorithm executing policies $\pi_1, \ldots, \pi_K$ is evaluated by the signed Bayesian reward shortfall and signed risk-value debt

$$\mathrm{BR}_r(K) := \mathbb{E} \sum_{k=1}^{K} \left[\mathcal{V}_\tau(p) - V_1^{\pi_k,p}(s_1) \right], \tag{2.6}$$

$$\mathrm{BR}_\rho(K) := \mathbb{E} \sum_{k=1}^{K} \left[\rho_1^{\pi_k,p}(s_1) - \tau \right]. \tag{2.7}$$

The expectation covers the prior draw, transition observations, posterior samples, and policy randomization. Neither quantity is necessarily nonnegative, and all results below control only their upper sides. In particular, the risk criterion is not the sum of positive violations.

Target guarantees. The goal is a near-$\sqrt{K}$ upper bound on (2.6) and a K-independent upper bound on (2.7), without supplying a common feasible policy, a known uniform gap, or a terminal episode count. We also ask for the computational status of the known-model constrained planner.

Remark 2.2 (Meaning of the risk guarantee). The quantity in (2.7) is a Bayesian expectation of a *signed* sum. Negative margins in some episodes may cancel positive margins in others. A K-independent upper bound therefore does not imply per-episode feasibility, a bound on $\sum_k (\rho_1^{\pi_k,p} - \tau)_+$, and is not a safety guarantee. We later give a high-probability certificate for the same cumulative *risk values*; that certificate likewise does not control realized trajectory costs or individual episodes. This distinction is essential to all of the results.

3 An obstruction to generic direct tightening

The additive Safe-PSRL template plans against a slightly tightened threshold. For nested CVaR, constrained value need not be continuous from below in that threshold, even when the original constraint is nonvacuous.

Proposition 3.1 (Value jump and scalar duality gap)**.** *Fix $\alpha \in (0,1)$. There is a one-stage known model satisfying* (2.4) *with a nonvacuous original budget $\tau = t \in (0,1)$, $g(q) = 0$, and $\Delta_* = t$. Every positive tightening has a reward loss that tends to $1-\alpha$ as the tightened budget approaches t from below. At every budget $b \in (0,t)$, the primal value is $\alpha b/t$ and the scalar Lagrangian dual value is b/t.*

Proof. There are three actions: safe has $(c,r) = (0,0)$, high has $(c,r) = (t,1)$, and bad has $(c,r) = (1,0)$. At the original budget t, high is feasible and optimal, while bad is infeasible, so the constraint is nonvacuous. For any budget $b < t$, replacing probability on bad by safe preserves reward and weakly decreases risk. It therefore suffices to let x be the probability of high, in which case

$$\rho(x) = t\ \mathrm{CVaR}_\alpha^{\mathrm{up}}(\mathrm{Bernoulli}(x)) = t \min\{x/\alpha, 1\}. \tag{3.1}$$

At budget $b \in (0,t)$, the constrained optimum is $x = \alpha b/t$. The bad action is dominated in the scalar inner problem as well, and

$$\inf_{\lambda \geq 0} \left\{ \lambda b + \max_{x \in [0,1]} [x - \lambda \rho(x)] \right\} = \inf_{\lambda \geq 0} \{\lambda b + \max(0, 1 - \lambda t)\} = \frac{b}{t}.$$

The inner maximizer uses only $x = 0$ or $x = 1$. As $b \uparrow t$, the tightened optimum tends to α, whereas the original value is one. The construction and conclusion require $\alpha < 1$. □

Corollary 3.2 (Linear regret from strict tightening)**.** *In the model of Proposition 3.1, any policy feasible for a strictly tightened budget $t - \varepsilon_k$, with $\varepsilon_k \in (0,t)$, loses at least $1-\alpha$ reward relative to the original budget-t optimum. Thus an algorithm that uses a positive tightening in every episode has reward regret at least $(1-\alpha)K$, even though the model is known.*

Proof. Feasibility gives $x_k \leq \alpha(t - \varepsilon_k)/t$, whereas the original constrained value is one. Hence $1 - x_k \geq 1 - \alpha + \alpha\varepsilon_k/t \geq 1 - \alpha$; summing proves the claim. □

4 Anytime sparse-calibration posterior sampling

This section defines the risk bank and its minimum-risk calibration policy, then states the main performance theorem. The bank turns computed slack in a sparse set of episodes into cancellation of sampled-model debt without using the unknown uniform gap.

Define the rare-event weights

$$W_{\alpha,H}^2 := \sum_{h=1}^{H-1} (H-h+1)^2 \alpha^{-h}, \qquad w_{\alpha,H} := \max_{1\le h\le H-1} (H-h)\alpha^{-h}. \tag{4.1}$$

Empty sums and maxima are zero. The first weight comes from Hellinger stability and the second pays for transition rows with very few observations. For analysis at an arbitrary horizon K, set

$$L_K := S\log 2 + \log(8SAHK^2), \tag{4.2}$$

$$\Lambda_K := S\log(KH+1) + \log(8SAHK^2),$$

$$C_K := 4\sqrt{2}\, W_{\alpha,H}\sqrt{SA\Lambda_K \log(eKH)}. \tag{4.3}$$

Neither confidence quantity is used by the algorithm. Its anytime bank target is defined from

$$\nu := S+2, \qquad \lambda_0 := S\log(H+1) + \log(8SAH), \qquad m_0 := \log(eH),$$

$$A_0 := \lambda_0 + 2\nu, \qquad M_0 := m_0 + 2, \qquad \mathfrak{c} := 4\sqrt{2}\, W_{\alpha,H}\sqrt{SA}, \tag{4.4}$$

and, for integers $k \ge 1$, by

$$T_0 := 0, \qquad T_k := \mathfrak{c}\sqrt{k(A_0 + \nu\log k)(M_0 + \log k)}. \tag{4.5}$$

For $H = 1$, the convention $W_{\alpha,H} = 0$ makes $T_k = 0$.

Let $\mathcal{H}_k$ be the sigma-field generated by all trajectories, posterior draws, selected policies, bank values, and internal randomization before episode k, but not by the latent kernel p or the current draw $\widehat{p}_k$. A trajectory includes $(S_1, A_1, \dots, S_H, A_H, S_{H+1})$, and $N_k(s,a)$ counts before episode k all visits to (s,a) for which this successor is observed. Conditional on $\mathcal{H}_k$, the latent p and the fresh draw $\widehat{p}_k$ are independent copies from the regular conditional posterior $\mu_k(\cdot) := \mathbb{P}(p \in \cdot \mid \mathcal{H}_k)$; the selected policy is measurable with respect to $\sigma(\mathcal{H}_k, \widehat{p}_k)$.

Continuity, compactness, and the measurable maximum theorem allow us to fix Borel selectors for every optimization below. Standard-Borel history and kernel spaces also provide regular conditional posteriors and conditionally independent draws supported on $\mathcal{Q}$ almost surely; details appear in Appendix A.2.

The online result assumes exact access to both the posterior-sampling oracle and the constrained-planning oracle. The arbitrary-prior claim is statistical: it does not assert that exact inference or planning is tractable for every prior.

Algorithm 1 (Anytime sparse-calibration posterior sampling)**.** Initialize the sampled-model risk bank at $\mathsf{B}_0 = 0$. For episode $k = 1, 2, \dots$, draw $\widehat{p}_k \sim \mu_k$ independently conditional on $\mathcal{H}_k$, and then:

1. If $\mathsf{B}_{k-1} < T_k$, execute a minimum-risk policy in the sampled model:

$$\pi_k \in \operatorname*{arg\,min}_{\pi\in\Pi_{\mathrm{M}}} \rho_1^{\pi,\widehat{p}_k}(s_1).$$

   Deposit its exact sampled-model slack

$$\delta_k := \tau - g(\widehat{p}_k) > 0, \qquad \mathsf{B}_k := \mathsf{B}_{k-1} + \delta_k.$$

2. If $\mathsf{B}_{k-1} \geq T_k$, execute an exact, un-tightened constrained optimum:

$$\pi_k \in \arg\max_{\pi \in \Pi_{\mathrm{M}}} \left\{ V_1^{\pi,\widehat{p}_k}(s_1) : \rho_1^{\pi,\widehat{p}_k}(s_1) \leq \tau \right\}.$$

Set $\mathsf{B}_k := \mathsf{B}_{k-1}$.

3. Observe $(S_1, A_1, \ldots, S_H, A_H, S_{H+1})$ and update the posterior.

The bank contains only calibration deposits; unused slack from exploitation episodes is deliberately ignored. Every deposit is computed by the same minimum-risk dynamic program that produces the calibration policy. Thus the algorithm knows neither the terminal episode K nor the analysis parameter Δ_*.

Calibration differs sharply from exploitation: the former admits an ordinary backward recursion even though the latter is computationally hard.

Proposition 4.1 (Minimum-risk calibration)**.** *For every kernel q, define $g_{H+1}^q(s) = 0$ and*

$$g_h^q(s) = \min_{a \in \mathcal{A}} \mathrm{CVaR}^{\mathrm{up}}_{\alpha,\, S' \sim q(\cdot|s,a)} \left(c_h(s,a) + g_{h+1}^q(S') \right). \tag{4.6}$$

Then $g_h^q(s)$ is the minimum nested risk from (h,s) over randomized physical-state Markov tail policies. A deterministic Markov minimizer exists, and all its actions can be computed in $O(H[S \log S + AS^2]) = O(HAS^2)$ arithmetic operations.

Proof. Proceed backward. The terminal risk is zero. Suppose the assertion holds at $h+1$, and fix a randomized action row $\lambda \in \Delta(\mathcal{A})$ at (h,s). Every continuation policy has a risk vector R_{h+1} satisfying $R_{h+1}(s') \geq g_{h+1}^q(s')$ pointwise. Monotonicity of CVaR therefore makes its risk at least

$$\mathrm{CVaR}^{\mathrm{up}}_{\alpha}\big(c_h(s,A) + g_{h+1}^q(S')\big), \quad A \sim \lambda, \quad S' \sim q(\cdot \mid s, A).$$

For fixed g_{h+1}^q, the primal formula (2.2) expresses this quantity as a pointwise infimum of functions affine in λ, hence as a concave function ϕ. If $\lambda = \sum_a \lambda_a e_a$, then $\phi(\lambda) \geq \sum_a \lambda_a \phi(e_a) \geq \min_a \phi(e_a)$, so a pure action minimizes current risk. Combining the minimizing actions at all successor states into the same deterministic Markov tail policy attains (4.6).

At a fixed stage, $c_h(s,a)$ does not change the ordering of the successor values. Sort these S values once in $O(S \log S)$ time. Each of the SA rows then needs an $O(S)$ scan to accumulate the upper α-tail probability. Summing over stages gives $O(H[S \log S + AS^2]) = O(HAS^2)$ arithmetic operations. □

The complexity result in Proposition 6.2 explains why the exact exploitation qualification is substantive; its proof appears in Appendix E, while posterior computation remains a separate sampling-oracle issue.

For a fully explicit K-independent risk bound, define

$$G_0 := \sqrt{A_0 M_0}, \qquad \gamma_0 := \frac{A_0 + \nu M_0}{2G_0}, \tag{4.7}$$

and

$$R_* := \frac{\mathfrak{c}^2}{2\Delta_*} \left[G_0 + 2\gamma_0 \log\left(1 + \frac{4\gamma_0 \mathfrak{c}}{\Delta_*}\right) \right]^2, \tag{4.8}$$

$$\mathcal{G}_* := 2SAH w_{\alpha,H} + \frac{H}{2} + R_*. \tag{4.9}$$

Here R_* is an envelope for the largest trailing-block bank deficit, and $\mathcal{G}_*$ adds the low-count and confidence-failure terms to obtain the total risk-debt envelope.

Theorem 4.2 (Anytime posterior sampling under nested CVaR)**.** *Fix $\alpha \in (0,1]$ and assume model-wise strict feasibility and the two exact oracles stated above. Then Algorithm 1 uses neither K nor Δ_* and, simultaneously for every evaluation horizon $K \geq 1$, satisfies the one-sided bounds*

$$\mathrm{BR}_r(K) \leq 8H^{3/2}\sqrt{SAKL_K} + \frac{H}{2} + \frac{H(T_K+\tau)}{\Delta_*}, \tag{4.10}$$

$$\mathrm{BR}_\rho(K) \leq \mathcal{G}_* \qquad (H \geq 2). \tag{4.11}$$

The number N_K^{cal} of calibration episodes obeys, almost surely and simultaneously for all K, the pathwise bound

$$N_K^{\mathrm{cal}} \leq \min\left\{K, \frac{T_K+\tau}{\Delta_*}\right\}. \tag{4.12}$$

For $H = 1$, no transition is learned, $T_K = N_K^{\mathrm{cal}} = 0$, $\mathrm{BR}_r(K) = 0$, and $\mathrm{BR}_\rho(K) \leq 0$. For every fixed instance with $\Delta_* > 0$, (4.12) implies $N_K^{\mathrm{cal}}/K \to 0$ almost surely. This is the precise sense in which calibration is sparse; it is not a parameter-uniform small constant. The right-hand side of (4.10) scales as $\widetilde{O}(S\sqrt{A}[H^{3/2} + HW_{\alpha,H}/\Delta_*]\sqrt{K})$, while $\mathcal{G}_*$ scales as $\widetilde{O}(SAHw_{\alpha,H} + S^2AW_{\alpha,H}^2/\Delta_* + H)$. The logarithms in the latter expression depend only on fixed problem parameters, not on K; neither statement is a two-sided bound on the signed criteria.

Extensions and endpoints. Pathwise certified exploitation errors add their cumulative feasibility and optimization errors to the two right-hand sides; uncontrolled approximate inference is not covered. The theorem does not require identifiability and remains valid at $\alpha = 1$, although the obstruction and hardness constructions use $\alpha < 1$. Stage-dependent tail masses replace powers of α by the corresponding prefix products. Precise statements and proofs are in Corollaries B.1 and B.3 and Remark B.2.

5 Analysis

This section states the central simulation results and proves Theorem 4.2. The concentration, measurable-selection, and local perturbation details are collected in the appendix.

Write $p_{sa} := p(\cdot \mid s,a)$ and define the nominal occupancies

$$d_h^{\pi,p}(s,a) := \mathbb{P}^{\pi,p}(S_h = s, A_h = a), \qquad d_h^{\pi,p}(s) := \sum_a d_h^{\pi,p}(s,a).$$

For a policy π, kernel q, continuation f, and stage h, put

$$T_{q,h}^{\pi} f(s) := \mathrm{CVaR}_\alpha^{\mathrm{up}}(c_h(s,A) + f(S')), \quad A \sim \pi_h(\cdot \mid s), \quad S' \sim q(\cdot \mid s, A),$$

and use

$$\mathrm{TV}(P,Q) = \tfrac{1}{2}\|P - Q\|_1, \qquad d_{\mathrm{H}}^2(P,Q) := \frac{1}{2}\sum_\omega (\sqrt{P(\omega)} - \sqrt{Q(\omega)})^2.$$

The first result replaces the usual total-variation factor $1/\alpha$ by a support-size-independent Hellinger factor $1/\sqrt{\alpha}$.

Lemma 5.1 (CVaR–Hellinger stability)**.** *Let P, Q be finite laws and let X take values in an interval of length R. For every $\alpha \in (0,1]$,*

$$\left|\mathrm{CVaR}_{\alpha,P}^{\mathrm{up}}(X) - \mathrm{CVaR}_{\alpha,Q}^{\mathrm{up}}(X)\right| \leq \frac{2\sqrt{2}\,R}{\sqrt{\alpha}}\, d_{\mathrm{H}}(P,Q). \tag{5.1}$$

Proof. Translation reduces to $0 \le X \le R$. The quantile formula and layer-cake identity give

$$\mathrm{CVaR}^{\mathrm{up}}_{\alpha,P}(X) = \frac{1}{\alpha}\int_0^R \min\{\alpha, P(X > t)\}\, dt. \tag{5.2}$$

For an event E, put $p = P(E)$ and $q = Q(E)$ and suppose without loss that $p \le q$. Hellinger distance contracts under the map $\omega \mapsto \mathbf{1}_E$, so $\left|\sqrt{p} - \sqrt{q}\right| \le \sqrt{2}\, d_{\mathrm{H}}(P, Q)$. If $q \le \alpha$, then $q - p \le 2\sqrt{2\alpha}\, d_{\mathrm{H}}(P, Q)$. If $p \le \alpha \le q$, the same bound holds for $\alpha - p$; if $\alpha \le p$, the capped probabilities coincide. Applying this event bound to $E = \{X > t\}$ in (5.2) and integrating proves the claim. □

The proof uses a hybrid local estimate that applies total variation only to low-count rows and Hellinger distance to the remaining rows.

Lemma 5.2 (Hybrid local perturbation)**.** *Let $\mathcal{I} \subseteq \mathcal{A}$ be any set of action rows and $0 \le f \le H - h$. Then*

$$\begin{aligned}\left|T^{\pi}_{p,h} f(s) - T^{\pi}_{q,h} f(s)\right| \le{}& \frac{H-h}{\alpha}\sum_{a \notin \mathcal{I}} \pi_h(a \mid s)\, \mathrm{TV}(p_{sa}, q_{sa}) \\ &+ \frac{2\sqrt{2}(H-h+1)}{\sqrt{\alpha}}\left(\sum_{a \in \mathcal{I}} \pi_h(a \mid s) d^2_{\mathrm{H}}(p_{sa}, q_{sa})\right)^{1/2}.\end{aligned} \tag{5.3}$$

The next lemma propagates these local errors through an adverse path law. Its measure-domination conclusion is what permits learning from nominal visits.

Lemma 5.3 (Distorted simulation)**.** *For every p, q and π, there are probability distributions $\widetilde{d}_h$ on states, with $\widetilde{d}_1(s_1) = 1$, such that, for $u_h(s, a) := \widetilde{d}_h(s)\pi_h(a \mid s)$,*

$$\begin{aligned}\rho^{\pi,p}_1(s_1) - \rho^{\pi,q}_1(s_1) &\le \sum_{h=1}^{H-1}\sum_{s} \widetilde{d}_h(s)\varepsilon_h(s), \\ \varepsilon_h(s) &:= \left|T^{\pi}_{p,h}\rho^{\pi,q}_{h+1}(s) - T^{\pi}_{q,h}\rho^{\pi,q}_{h+1}(s)\right|.\end{aligned} \tag{5.4}$$

In particular,

$$\rho^{\pi,p}_1(s_1) - \rho^{\pi,q}_1(s_1) \le \sum_{h=1}^{H-1} \frac{H-h}{\alpha}\sum_{s,a} u_h(s, a)\, \mathrm{TV}(p_{sa}, q_{sa}), \tag{5.5}$$

and, pointwise,

$$u_h(s, a) \le \alpha^{-(h-1)} d^{\pi,p}_h(s, a). \tag{5.6}$$

No ratio is asserted when nominal occupancy is zero.

Corollary 5.4 (Nominal-occupancy simulation)**.** *For every p, q, π,*

$$\rho^{\pi,p}_1(s_1) - \rho^{\pi,q}_1(s_1) \le \sum_{h=1}^{H-1}(H-h)\alpha^{-h}\mathbb{E}_{(S_h,A_h)\sim d^{\pi,p}_h}[\mathrm{TV}(p_{S_hA_h}, q_{S_hA_h})]. \tag{5.7}$$

Reverse-KL row-prefix confidence sets, posterior symmetry, and frozen-count charging now give the expected weighted simulation bound. Full proofs of Lemmas 5.2 and 5.3 and Corollary 5.4 and the next result appear in Appendix C.

Lemma 5.5 (Weighted posterior simulation)**.** *For every sequence π_k measurable with respect to $\sigma(\mathcal{H}_k, \widehat{p}_k)$,*

$$\sum_{k=1}^{K} \mathbb{E}\left[\rho_1^{\pi_k,p}(s_1) - \rho_1^{\pi_k,\widehat{p}_k}(s_1)\right] \le C_K\sqrt{K} + D, \qquad D := 2SAHw_{\alpha,H} + \frac{H}{2}. \tag{5.8}$$

This holds for every prior μ.

For a pathwise statement, predictable occupancy sums must also be converted to realized visits. The resulting certificate is pointwise in a fixed evaluation horizon and a fixed admissible policy-selection rule.

Theorem 5.6 (Fixed-K risk-bank certificate)**.** *Fix $K \ge 1$ and $\delta \in (0,1)$, and define*

$$\Lambda_{K,\delta} := S\log(KH+1) + \log\frac{4SAHK^2}{\delta}, \qquad \ell_\delta := \log\frac{4}{\delta}, \tag{5.9}$$

$$\mathcal{E}_{K,\delta} := w_{\alpha,H}\big(4SAH + 2H\ell_\delta\big) + 4W_{\alpha,H}\sqrt{\Lambda_{K,\delta}K\big(4SA\log(eKH) + 2\ell_\delta\big)}. \tag{5.10}$$

For each fixed admissible policy-selection rule producing π_k measurable with respect to $\sigma(\mathcal{H}_k, \widehat{p}_k)$, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}\left[\rho_1^{\pi_k,p}(s_1) - \rho_1^{\pi_k,\widehat{p}_k}(s_1)\right] \le \mathcal{E}_{K,\delta}. \tag{5.11}$$

The probability covers the prior draw, posterior samples, and trajectories. Consequently, for anytime sparse-calibration posterior sampling,

$$\sum_{k=1}^{K}[\rho_1^{\pi_k,p}(s_1) - \tau] \le \mathcal{E}_{K,\delta} - \mathsf{B}_K \tag{5.12}$$

with the same probability. The right side is computable from the chosen (K,δ), the problem parameters, and the observed bank.

Remark 5.7 (Scope of the certificate)*.* The guarantee in (5.12) concerns the signed sum of true-model nested-risk values. It controls neither realized trajectory costs nor per-episode or positive-part violation. It is stated for a fixed pair (K,δ); a simultaneous confidence sequence would require an additional time-uniform construction.

The reward argument uses the ordinary, undistorted analogue.

Lemma 5.8 (Policy-uniform reward simulation)**.** *For every policy sequence measurable with respect to $\sigma(\mathcal{H}_k, \widehat{p}_k)$,*

$$\sum_{k=1}^{K}\mathbb{E}\left[V_1^{\pi_k,\widehat{p}_k}(s_1) - V_1^{\pi_k,p}(s_1)\right] \le 8H^{3/2}\sqrt{SAKL_K} + \frac{H}{2}. \tag{5.13}$$

The remaining deterministic fact about the bank target isolates the calculus needed to turn an anytime target into a horizon-independent residual.

Lemma 5.9 (Deterministic target properties)**.** *For $H \geq 2$ and all integers $k \geq 1$ and $j, n \geq 0$,*

$$C_k\sqrt{k} \leq T_k, \qquad T_{j+n} \leq T_j + T_n, \qquad \sup_{m\geq 0}(T_m - m\Delta_*)_+ \leq R_*. \tag{5.14}$$

The proofs of Lemmas 5.8 and 5.9 and Theorem 5.6 are in the appendix. We can now give the core argument for the main theorem without hiding the bank cancellation that drives it.

Proof of Theorem 4.2. If $H = 1$, then $W_{\alpha,H} = T_k = 0$ and the bank rule always selects an exact constrained optimum. Transitions affect neither reward nor risk, so $N_K^{\mathrm{cal}} = \mathrm{BR}_r(K) = 0$ and $\mathrm{BR}_\rho(K) \leq 0$. Hence assume $H \geq 2$. The measurable selectors fixed in Appendix A.2 make every executed policy admissible for the simulation lemmas. Every calibration deposit satisfies, almost surely,

$$\Delta_* \leq \delta_k = \tau - g(\widehat{p}_k) \leq \tau. \tag{5.15}$$

In an exploitation episode, sampled-model debt is nonpositive; in a calibration episode it equals $-\delta_k$. Thus

$$\mathrm{BR}_\rho(K) \leq C_K\sqrt{K} + D - \mathbb{E}\mathsf{B}_K \leq D + \mathbb{E}[T_K - \mathsf{B}_K], \tag{5.16}$$

where the first inequality is Lemma 5.5 and the second is Lemma 5.9.

We bound the deficit pathwise. If episode K is exploitation, the rule gives $\mathsf{B}_K \geq T_K$. Otherwise let $r, \ldots, K$ be the maximal trailing block of calibration episodes and put $n = K - r + 1$. If $r = 1$, then $\mathsf{B}_K \geq n\Delta_*$. If $r > 1$, episode $r - 1$ is exploitation, so $\mathsf{B}_{r-1} \geq T_{r-1}$ and $\mathsf{B}_K \geq T_{r-1} + n\Delta_*$. Subadditivity therefore gives

$$T_K - \mathsf{B}_K \leq T_n - n\Delta_* \leq \sup_{m\geq 0}(T_m - m\Delta_*)_+ \leq R_*. \tag{5.17}$$

Together with (5.16) and $\mathcal{G}_* = D + R_*$, this proves (4.11).

If no episode calibrates, the count bound is immediate. Immediately before the last calibration, the bank is below its then-current target and hence below T_K; the last deposit is at most τ, and the bank does not subsequently change. Since every deposit is at least Δ_*,

$$N_K^{\mathrm{cal}}\Delta_* \leq \mathsf{B}_K < T_K + \tau.$$

Combining this inequality with $N_K^{\mathrm{cal}} \leq K$ proves (4.12).

For reward regret, posterior symmetry and measurability of the constrained value give

$$\mathbb{E}[\mathcal{V}_\tau(p) - \mathcal{V}_\tau(\widehat{p}_k) \mid \mathcal{H}_k] = 0. \tag{5.18}$$

Decompose

$$\begin{aligned}\mathcal{V}_\tau(p) - V_1^{\pi_k,p} = {}& [\mathcal{V}_\tau(p) - \mathcal{V}_\tau(\widehat{p}_k)] \\ & + [\mathcal{V}_\tau(\widehat{p}_k) - V_1^{\pi_k,\widehat{p}_k}] + [V_1^{\pi_k,\widehat{p}_k} - V_1^{\pi_k,p}].\end{aligned}$$

The middle term is zero in exploitation episodes and at most H in calibration episodes. Summing, using (5.18), Lemma 5.8, and (4.12) proves (4.10). The scaling statements follow by substituting the definitions of T_K and R_*. □

6 Computational complexity

Exact calibration is polynomial by Proposition 4.1, but exact exploitation can be hard even for a known model. We make the oracle qualification precise for rationally encoded instances.

Definition 6.1 (Rational dynamic-CVaR Markov planning)**.** The input is a binary encoding of finite $H, \mathcal{S}, \mathcal{A}$, a rational transition kernel, rational rewards and costs, a rational tail mass $\alpha \in (0, 1]$, a budget τ, and a reward target v_{tar}. The question is whether some $\pi \in \Pi_{\mathrm{M}}$ satisfies $\rho_1^{\pi,p}(s_1) \le \tau$ and $V_1^{\pi,p}(s_1) \ge v_{\mathrm{tar}}$.

Proposition 6.2 (Planning complexity)**.** *For every fixed rational $\alpha \in (0, 1)$, Definition 6.1 is NP-complete on the subclass with one state, two actions, and a deterministic self-loop, even when a strictly feasible all-skip policy is known. With unrestricted rational input data, the decision problem belongs to $\exists\mathbb{R}$ and hence to PSPACE.*

The lower bound is a Subset-Sum reduction in which equality between expected reward and nested risk forces every action probability to be binary. For the upper classification, policy, reward-value, risk-epigraph, CVaR-threshold, and positive-part variables form a polynomial-size degree-two semialgebraic system. The full reduction and equivalence proof are in Appendix E. The exact complexity of the unrestricted problem remains open: the result gives an $\exists\mathbb{R}$/PSPACE upper classification, not $\exists\mathbb{R}$-completeness. It also does not implement exact inference or planning for arbitrary real-valued posterior draws.

7 Information-theoretic lower bounds

The exponential tail dependence in Theorem 4.2 is not only an artifact of the upper-bound proof. An admissible learner does not observe the latent model index θ and, in episode k, uses past complete trajectories and independent internal randomness to choose a policy $\pi_k \in \Pi_{\mathrm{M}}$. Every expectation below covers the uniform prior on θ, transition randomness, action randomization, and the learner's internal randomness.

Theorem 7.1 (Risk debt with zero reward regret)**.** *Fix $0 < \alpha \le 1/2$, $H \ge 3$, and $\tau = 1/4$. There is an MDP with $S = 2H$, $A = 2$, and a uniform prior on two time-homogeneous kernels, with known rewards and costs and with $g(q) = 0$ in each model, such that every admissible learner has $\mathrm{BR}_r(K) = 0$ for all K. Moreover, its exact finite-horizon Bayes optimum is*

$$\inf_{\mathsf{Alg}} \mathrm{BR}_\rho^{\mathsf{Alg}}(K) = \frac{1 - (1-\beta)^K}{2\beta} - \frac{K}{4}, \qquad \beta := \alpha^{H-1}, \tag{7.1}$$

where the infimum ranges over all admissible learners. In particular, every learner, at $K_0 = \lfloor 1/(2\beta) \rfloor$, satisfies

$$\mathrm{BR}_\rho(K_0) \ge \frac{1}{32\beta} = \frac{1}{32}\alpha^{-(H-1)}. \tag{7.2}$$

The construction contains two branches that reveal which model is latent only after $H - 1$ consecutive transitions of probability α. Before this geometric revelation, the posterior stays uniform and the two model risks are $\rho_{\theta=0} = x_1$ and $\rho_{\theta=1} = x_0$ for initial branch probabilities (x_0, x_1). Thus every learner incurs expected debt $1/4$ until revelation, while all policies have the same reward. Summing the resulting geometric waiting time proves (7.1); the full construction and constant calculation are in Appendix F.

Adding a known common fallback changes the conclusion from a pure risk lower bound to a risk–reward Pareto tradeoff.

Corollary 7.2 (Known common safe fallback)**.** *Modify the construction of Theorem 7.1 by enlarging the common action set to* $\{a_0, a_1, \ell\}$*, so* $A = 3$*. At* s_1*, the known action* ℓ *has reward zero and moves to the zero-cost absorbing state* f*; at every other state it duplicates* a_0*. Then* ℓ *is safe in both models, and every admissible learner satisfies*

$$\mathrm{BR}_\rho(K) + \frac{1}{2}\,\mathrm{BR}_r(K) \geq \frac{1-(1-\beta)^K}{2\beta} - \frac{K}{4}, \qquad \beta = \alpha^{H-1}. \tag{7.3}$$

Consequently, if $F, G \geq 0$ *and* $\mathrm{BR}_r(K) \leq F\sqrt{K}$ *and* $\mathrm{BR}_\rho(K) \leq G$ *for every* K*, then*

$$G \geq \frac{1}{64\beta} \qquad \text{or} \qquad F \geq \frac{\sqrt{2}}{32\sqrt{\beta}}. \tag{7.4}$$

The no-fallback theorem is stronger for risk debt alone: its safe policies are model-dependent, exactly as allowed by (2.4), and it forces $G = \Omega(\alpha^{-(H-1)})$ even when reward regret is zero. Under the stronger common-fallback assumption, Corollary 7.2 gives only a Pareto conclusion and does not lower-bound both coefficients simultaneously.

8 Limitations and conclusion

The theorem assumes a correctly specified prior, known bounded rewards and costs, a finite time-homogeneous tabular MDP, and model-wise strict feasibility on the prior support. It also assumes exact posterior sampling and exact constrained planning; Corollary B.1 covers only pathwise certified planning errors. The comparator is randomized Markov on the physical state, with action randomness inside CVaR. Changing to an augmented-state or episode-level-mixture comparator changes the feasible set.

The controlled risk quantity is an expected signed aggregate, useful when surplus and deficit may be balanced across episodes. It is not per-episode feasibility, a sum of positive parts, a guarantee for realized costs, or a safety certificate. It is the nonlinear-risk analogue of the signed constraint debt used in additive constrained learning. Within this oracle model, sparse calibration gives a near-$\sqrt{K}$ one-sided upper bound on signed Bayesian reward shortfall and a K-independent upper bound on expected signed risk-value debt without a common feasible baseline or knowledge of the induced uniform gap. The rare-chain construction shows that the worst-case risk-debt exponent is necessary at a tailored episode count.

References

[1] M. Agarwal, Q. Bai, and V. Aggarwal. Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 22–31, 2022. PMLR 180.

[2] M. Ahmadi, U. Rosolia, M. D. Ingham, R. M. Murray, and A. D. Ames. Constrained risk-averse Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11718–11725, 2021. doi:10.1609/aaai.v35i13.17393.

[3] N. Bäuerle and A. Glauner. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 296(3):953–966, 2022. doi:10.1016/j.ejor.2021.04.030.

[4] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in Real Algebraic Geometry*. Springer, second edition, 2006. doi:10.1007/3-540-33099-2.

[5] K. Brantley, M. Dudik, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems 33*, pages 16315–16326, 2020. Proceedings version; corrected extended version: arXiv:2006.05051.

[6] A. Bura, A. HasanzadeZonuzy, D. Kalathil, S. Shakkottai, and J.-F. Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *Advances in Neural Information Processing Systems 35*, pages 1047–1059, 2022. doi:10.52202/068431-0077; arXiv:2112.00885.

[7] J. Canny. Some algebraic and geometric computations in PSPACE. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 460–467, 1988. doi:10.1145/62212.62257.

[8] Y. Chen, Y. Du, P. Hu, S. Wang, D. Wu, and L. Huang. Provably efficient iterated CVaR reinforcement learning with function approximation and human feedback. In *International Conference on Learning Representations*, pages 30478–30515, 2024. Proceedings version.

[9] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018. JMLR version.

[10] Y.-L. Chow and M. Pavone. Stochastic optimal control with dynamic, time-consistent risk constraints. In *2013 American Control Conference*, pages 390–395, 2013. doi:10.1109/ACC.2013.6579868.

[11] S. Chu and Y. Zhang. Markov decision processes with iterated coherent risk measures. *International Journal of Control*, 87(11):2286–2293, 2014. doi:10.1080/00207179.2014.909947.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2006. doi:10.1002/047174882X.

[13] Z. Deng, A. Velasquez, and S. Zou. Minimax optimal sample complexity for iterated CVaR reinforcement learning with a generative model. *IEEE Transactions on Information Theory*, 72(7):5077–5103, 2026. doi:10.1109/TIT.2026.3692816.

[14] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. R. Jovanović. Provably efficient safe exploration via primal-dual policy optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3304–3312, 2021. PMLR 130; extended version: arXiv:2003.00534.

[15] Y. Du, S. Wang, and L. Huang. Provably efficient risk-sensitive reinforcement learning: Iterated CVaR and worst path. In *International Conference on Learning Representations*, 2023. arXiv:2206.02678.

[16] Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained MDPs. arXiv:2003.02189, 2020.

[17] A. Ghosh and M. Moharrami. Online learning in risk sensitive constrained MDP. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19406–19425, 2025. PMLR 267.

[18] K. C. Kalagarla, R. Jain, and P. Nuzzo. A safe Bayesian learning algorithm for constrained MDPs with bounded constraint violation. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2782–2790, 2025. PMLR 258; extended version: arXiv:2301.11547.

[19] D. Kim, T. Cho, S. Han, H. Chung, K. Lee, and S. Oh. Spectral-risk safe reinforcement learning with convergence guarantees. In *Advances in Neural Information Processing Systems 37*, pages 92465–92498, 2024. Proceedings version. doi:10.52202/079017-2936.

[20] M. Koren, O. Peretz, T. Dinh, and P. S. Yu. UAMDP: Uncertainty-aware Markov decision process for risk-constrained reinforcement learning from probabilistic forecasts. arXiv:2510.08226v2, 2025.

[21] J. Lee, B. Saglam, S. Pougkakiotis, A. Karbasi, and D. Kalogerias. Risk-averse constrained reinforcement learning with optimized certainty equivalents. In *Advances in Neural Information Processing Systems 38*, 2025. Proceedings version; doi:10.52202/085713-1250.

[22] H. Liang and Z. Luo. Regret bounds for risk-sensitive reinforcement learning with Lipschitz dynamic risk measures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1774–1782, 2024. PMLR 238.

[23] Y. Lin, Y. Ren, and E. Zhou. Bayesian risk Markov decision processes. In *Advances in Neural Information Processing Systems 35*, pages 17430–17442, 2022. doi:10.52202/068431-1267; Proceedings version.

[24] T. Liu, R. Zhou, D. Kalathil, P. R. Kumar, and C. Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Advances in Neural Information Processing Systems 34*, pages 17183–17193, 2021. Proceedings version; arXiv:2106.02684.

[25] Y. Luo and E. Delage. Actor-critic algorithm for dynamic expectile and CVaR. arXiv:2605.07857, 2026.

[26] I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems 26*, pages 3003–3011, 2013. Proceedings version; arXiv:1306.0940.

[27] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems 30*, pages 1333–1342, 2017. Proceedings version; arXiv:1709.04570.

[28] D. Provodin, M. C. Kaptein, and M. Pechenizkiy. Efficient exploration in average-reward constrained reinforcement learning: Achieving near-optimal regret with posterior sampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41144–41162, 2024. PMLR 235.

[29] M. Rigter, B. Lacerda, and N. Hawes. Risk-averse Bayes-adaptive reinforcement learning. In *Advances in Neural Information Processing Systems 34*, pages 1142–1154, 2021. Proceedings version; arXiv:2102.05762.

[30] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000. doi:10.21314/JOR.2000.038.

[31] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi:10.1016/S0378-4266(02)00271-6.

[32] A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010. doi:10.1007/s10107-010-0393-3.

[33] H. Wei, X. Liu, and L. Ying. Triple-Q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3274–3307, 2022. PMLR 151; arXiv:2106.01577.

[34] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003.

[35] W. Xu, X. Gao, and X. He. Regret bounds for Markov decision processes with recursive optimized certainty equivalents. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38400–38427, 2023. PMLR 202.

[36] Z. Yang, H. Jin, Y. Tang, and G. Fan. Risk-aware constrained reinforcement learning with non-stationary policies. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2029–2037, 2024. Proceedings version. doi:10.65109/ZRHO4945.

A Regularity and measurable selection

This appendix supplies the compactness and measurability details used in the setup and algorithm.

A.1 Proof of the automatic uniform gap

Proof of Lemma 2.1. The policy space Π_{M} and the kernel simplex are compact. We verify that $(\pi, q) \mapsto \rho_h^{\pi,q}(s)$ is continuous by backward induction. At $H+1$ both sides are zero. If the continuation is continuous and lies in $[0, H-h]$, the primal representation in (2.2) writes the stage-h risk as the minimum over $z \in [0, H-h+1]$ of a function continuous in (π, q, z); the loss range ensures that this restriction loses no minimizer. Berge's maximum theorem preserves continuity at stage h. Applying it once more to the minimum over compact Π_{M} shows that g is continuous. Finally, $\mathcal{Q}$ is closed in the compact kernel simplex, so g attains its maximum at some $q_* \in \mathcal{Q}$. Model-wise strict feasibility gives $g(q_*) < \tau$, proving (2.4). □

A.2 Measurable selectors

The set $\mathcal{Q}$ is closed in the compact kernel simplex, and

$$\Gamma(q) := \{\pi \in \Pi_{\mathrm{M}} : \rho_1^{\pi,q}(s_1) \le \tau\}$$

has a Borel graph and nonempty compact sections because reward and nested risk are continuous and (2.4) holds. The measurable maximum theorem therefore gives Borel optimizer correspondences. Repeatedly minimizing policy coordinates over each nonempty compact optimizer set selects its coordinatewise lexicographically least point and gives a Borel selector. The kernel and history spaces

are standard Borel, so regular conditional posteriors, and hence the conditional independent draws used below, exist. On realized histories the posterior and its independent draw are supported on $\mathcal{Q}$ almost surely.

B Extensions and endpoint cases

This appendix records approximate-oracle robustness, endpoint behavior, and stage-dependent tail masses.

Corollary B.1 (Certified approximate exploitation)**.** *Suppose the exploitation oracle in episode k returns a policy satisfying*

$$\rho_1^{\pi_k,\widehat{p}_k}(s_1) \le \tau + \varepsilon_k, \qquad V_1^{\pi_k,\widehat{p}_k}(s_1) \ge \mathcal{V}_\tau(\widehat{p}_k) - \zeta_k$$

for pathwise certified deterministic errors $\varepsilon_k, \zeta_k \ge 0$ at every realized call, while calibration remains exact. For $H \ge 2$, the right-hand sides of (4.11) and (4.10) increase by at most $\sum_{k\le K}\varepsilon_k$ and $\sum_{k\le K}\zeta_k$, respectively. For $H = 1$, $\mathrm{BR}_\rho(K) \le \sum_{k\le K}\varepsilon_k$ and $\mathrm{BR}_r(K) \le \sum_{k\le K}\zeta_k$.

Proof. In the risk decomposition, exploitation episode k contributes at most ε_k sampled-model debt. In the reward decomposition, its sampled suboptimality is at most ζ_k. All model-error and bank terms are unchanged. At $H = 1$ there is no model error, giving the displayed endpoint bounds directly. □

Remark B.2 (Endpoints and identifiability). The bounds hold for every $K \ge 1$ and do not assume posterior concentration or model identifiability. At $\alpha = 1$, upper-tail CVaR is expectation, the risk distortion is one, and Theorem 4.2 remains valid with polynomial weights. The value-jump, duality-gap, and integrality constructions in Propositions 3.1 and 6.2 use $\alpha < 1$.

For $H \ge 2$,

$$4\alpha^{-(H-1)} \le W_{\alpha,H}^2 \le H^3\alpha^{-(H-1)}, \qquad w_{\alpha,H} \le H\alpha^{-(H-1)}. \tag{B.1}$$

Thus $W_{\alpha,H}^2$ has the same α-exponent as $\alpha^{-(H-1)}$ up to polynomial factors in H; no soft-O notation here hides those factors.

Corollary B.3 (Stage-dependent tail masses)**.** *Suppose the stage-h risk operator is upper-tail CVaR with known mass $\alpha_h \in (0, 1]$. Put*

$$P_0 := 1, \qquad P_h := \prod_{j=1}^{h}\alpha_j, \qquad W_{\boldsymbol{\alpha},H}^2 := \sum_{h=1}^{H-1}\frac{(H-h+1)^2}{P_h}, \qquad w_{\boldsymbol{\alpha},H} := \max_{1\le h\le H-1}\frac{H-h}{P_h}.$$

Replace $W_{\alpha,H}$ and $w_{\alpha,H}$ by these quantities in the bank target and all performance bounds, and use α_h in the stage-h calibration recursion. Then Theorems 4.2 and 5.6 and Corollary B.1 remain valid.

Proof. At stage h, the local TV and Hellinger coefficients become respectively $(H-h)/\alpha_h$ and $2\sqrt{2}(H-h+1)/\sqrt{\alpha_h}$. The distorted state occupancy satisfies

$$\widetilde{d}_h(s) \le P_{h-1}^{-1} d_h^{\pi,p}(s),$$

because each preceding envelope density is at most $1/\alpha_j$. Combining these facts gives $(H-h)/P_h$ for low-count rows and $(H-h+1)^2/P_h$ after squaring the high-count coefficient. The remainder of the proof is unchanged. The mass α_H does not enter transition-model error because no transition follows stage H. □

C Simulation under nested CVaR

This section proves the model-error and bank results stated in Section 5. It develops the local perturbation, confidence-set, visit-counting, and predictable-to-realized tools used by those proofs, using the notation introduced in the main text.

Lemma C.1 (Local perturbation)**.** *If $0 \le f \le H - h$ pointwise, then*

$$\left|T^{\pi}_{p,h} f(s) - T^{\pi}_{q,h} f(s)\right| \le \frac{H-h}{\alpha} \sum_{a} \pi_h(a \mid s)\, \mathrm{TV}(p_{sa}, q_{sa}). \tag{C.1}$$

Moreover, $T^{\pi}_{q,h}$ is 1-Lipschitz in continuation sup norm.

Proof. Fix z in the primal representation (2.2). For each action a, the function

$$s' \longmapsto (c_h(s,a) + f(s') - z)_+$$

has span at most $H - h$. Because $p_{sa} - q_{sa}$ has total mass zero, the difference of its expectations under the two rows is at most that span times total variation. Average over a, divide by α, and use $|\inf_z F(z) - \inf_z G(z)| \le \sup_z |F(z) - G(z)|$. The current cost does not enlarge the span because it is constant in s' after fixing a. Monotonicity and translation equivariance of CVaR give the Lipschitz claim. □

Proof of Lemma 5.2. Introduce an intermediate kernel r whose rows at state s equal q_{sa} for $a \notin \mathcal{I}$ and p_{sa} for $a \in \mathcal{I}$. The difference between p and r involves only rows outside $\mathcal{I}$, so the row-wise span argument in Lemma C.1 gives the first term. The joint action–successor laws induced by r and q have squared Hellinger distance

$$\sum_{a \in \mathcal{I}} \pi_h(a \mid s) d^2_{\mathrm{H}}(p_{sa}, q_{sa}).$$

Under either law, $c_h(s, A) + f(S')$ lies in an interval of length at most $H - h + 1$. Apply Lemma 5.1 to obtain the second term and use the triangle inequality. □

Corollary C.2 (Uniform fixed-policy perturbation)**.** *For every fixed π, every pair of kernels p, q, and every stage h,*

$$\left\|\rho^{\pi,p}_h - \rho^{\pi,q}_h\right\|_\infty \le \frac{(H-h)(H-h+1)}{2\alpha} \max_{s,a} \mathrm{TV}(p_{sa}, q_{sa}). \tag{C.2}$$

Proof. Split the Bellman difference by inserting $T^{\pi}_{p,j} \rho^{\pi,q}_{j+1}$. The continuation Lipschitz property and Lemma C.1 give, for $j = h, \dots, H-1$,

$$\left\|\rho^{\pi,p}_j - \rho^{\pi,q}_j\right\|_\infty \le \left\|\rho^{\pi,p}_{j+1} - \rho^{\pi,q}_{j+1}\right\|_\infty + \frac{H-j}{\alpha} \max_{s,a} \mathrm{TV}(p_{sa}, q_{sa}).$$

Backward summation and $\sum_{j=h}^{H-1} (H-j) = (H-h)(H-h+1)/2$ prove the claim. □

Thus fixed-policy nested CVaR is uniformly Lipschitz in the transition model. The exponential weights in the upper bound below arise because data are collected under nominal paths while nested CVaR amplifies nominally rare paths. This is also distinct from the discontinuity of the *optimized constrained value in the budget* in Proposition 3.1.

We next prove the one-sided distorted-simulation result used for signed risk debt.

Proof of Lemma 5.3. The proof selects an adverse one-step law, unrolls the resulting recursion, and then compares that distorted law with nominal occupancy. Put $\delta_h(s) = \rho_h^{\pi,p}(s) - \rho_h^{\pi,q}(s)$. At each (h, s), the envelope correspondence $\Xi(P) = \{\xi \in [0, 1/\alpha]^{\mathcal{A}\times\mathcal{S}} : \mathbb{E}_P\xi = 1\}$ has nonempty compact sections and a measurable closed graph in the finite-dimensional law P. The measurable maximum theorem therefore gives a measurable dual maximizer $\xi_h(a, s' \mid s)$ in (2.2) for the p-model random variable with continuation $\rho_{h+1}^{\pi,p}$; fix one such selector. The distorted joint transition

$$M_h(a, s' \mid s) := \xi_h(a, s' \mid s)\pi_h(a \mid s)p(s' \mid s, a)$$

is a probability kernel and satisfies $M_h \le \alpha^{-1}\pi_h p$ pointwise. More explicitly, split

$$T_{p,h}^{\pi}\rho_{h+1}^{\pi,p} - T_{q,h}^{\pi}\rho_{h+1}^{\pi,q} = [T_{p,h}^{\pi}\rho_{h+1}^{\pi,p} - T_{p,h}^{\pi}\rho_{h+1}^{\pi,q}] + [T_{p,h}^{\pi}\rho_{h+1}^{\pi,q} - T_{q,h}^{\pi}\rho_{h+1}^{\pi,q}].$$

For random variables X, Y under the same law, the risk-envelope representation and a maximizer ξ_X for X imply

$$\mathrm{CVaR}_{\alpha}^{\mathrm{up}}(X) - \mathrm{CVaR}_{\alpha}^{\mathrm{up}}(Y) \le \mathbb{E}[\xi_X(X - Y)].$$

Apply this inequality to the first bracket and bound the second bracket above by its absolute value. This yields

$$\delta_h(s) \le \mathbb{E}_{(A,S')\sim M_h(\cdot|s)}[\delta_{h+1}(S')] + \varepsilon_h(s). \tag{C.3}$$

Unroll (C.3), taking $\widetilde{d}_h$ to be the state marginal reached through $M_1, \ldots, M_{h-1}$. This gives (5.4); Lemma C.1 then gives (5.5). To verify domination without treating a zero-probability ratio, sum M_h over actions:

$$\sum_a M_h(a, s' \mid s) \le \alpha^{-1}\sum_a \pi_h(a \mid s)p(s' \mid s, a).$$

Induction on h therefore gives $\widetilde{d}_h(s) \le \alpha^{-(h-1)}d_h^{\pi,p}(s)$. Multiplication by the same, undistorted current policy row $\pi_h(a \mid s)$ proves (5.6). Notice that M_h may distort the action in the preceding transition; the claim needed here is domination of its *state marginal*, followed by the nominal current action row. □

Proof of Corollary 5.4. Substitute the measure domination $u_h \le \alpha^{-(h-1)}d_h^{\pi,p}$ into (5.5). □

For episode k, abbreviate $d_{k,h} := d_h^{\pi_k,p}$, and use $\widetilde{d}_{k,h}$ and $u_{k,h}$ for the distorted measures from Lemma 5.3 with $q = \widehat{p}_k$.

We next prove the weighted posterior simulation bound.

Proof of Lemma 5.5. The proof couples adaptive data to fixed row tapes, transfers every history-measurable confidence event to the posterior draw, and charges low- and high-count rows separately. The latter charge is controlled by a frozen-count logarithmic potential.

For each row $x = (s, a)$, pre-generate an infinite successor sequence with law p_x; adaptive visits reveal prefixes of these row tapes. Let $\overline{p}_{x,n}$ be the empirical law of the first n successors. The method-of-types bound [12] gives

$$\mathbb{P}\left(\mathrm{KL}(\overline{p}_{x,n}\|p_x) > \frac{\Lambda_K}{n}\right) \le (n+1)^S e^{-\Lambda_K} \le \frac{1}{8SAHK^2}$$

for $1 \le n \le KH$. A union bound over rows and prefixes shows that the true kernel lies in every reverse-KL prefix ball except on an event $\mathcal{E}_0^c$ of probability at most $1/(8K)$. Formally, put

$$\mathcal{C}_{x,n} := \left\{v \in \Delta(\mathcal{S}) : \mathrm{KL}(\overline{p}_{x,n}\|v) \le \frac{\Lambda_K}{n}\right\} \quad (n \ge 1), \qquad \mathcal{C}_{x,0} := \Delta(\mathcal{S}),$$

and let $\mathcal{C}_k := \prod_x \mathcal{C}_{x,N_k(x)}$, which is $\mathcal{H}_k$-measurable.

More generally, conditional independence gives, for every $\mathcal{H}_k$-measurable random Borel set C_k, $\mathbb{P}(\widehat{p}_k \in C_k \mid \mathcal{H}_k) = \mathbb{P}(p \in C_k \mid \mathcal{H}_k)$. Applying this identity to $C_k = \mathcal{C}_k^c$ gives

$$\mathbb{P}(\widehat{p}_k \notin \mathcal{C}_k \mid \mathcal{H}_k) = \mu_k(\mathcal{C}_k^c) = \mathbb{P}(p \notin \mathcal{C}_k \mid \mathcal{H}_k). \tag{C.4}$$

Consequently, $\mathbb{P}(\widehat{p}_k \notin \mathcal{C}_k) = \mathbb{P}(p \notin \mathcal{C}_k) \le 1/(8K)$. Since a one-sided risk difference is at most H, the single true-prefix failure contributes at most $KH\mathbb{P}(\mathcal{E}_0^c) \le H/8$, while the episode-wise sampled-kernel failures contribute at most $H\sum_k \mathbb{P}(\widehat{p}_k \notin \mathcal{C}_k) \le H/8$. Their total is at most $H/4$; we retain the convenient allowance $H/2$ in D. On $\mathcal{E}_0 \cap \{\widehat{p}_k \in \mathcal{C}_k\}$ and for $N_k(x) \ge 1$,

$$d_{\mathrm{H}}^2(p_x, \widehat{p}_{k,x}) \le \frac{2\Lambda_K}{N_k(x)}. \tag{C.5}$$

Indeed, $2d_{\mathrm{H}}^2(r, v) \le \mathrm{KL}(r\|v)$ and the Hellinger triangle inequality apply with $r = \overline{p}_{x,N_k(x)}$.

Apply the local-error form of Lemma 5.3 with $q = \widehat{p}_k$, and in Lemma 5.2 take $\mathcal{I}_{k,s} := \{a : N_k(s,a) \ge H\}$. For a low-count row $x = (s,a)$, measure domination gives

$$\frac{H-h}{\alpha} u_{k,h}(x) \le w_{\alpha,H} d_{k,h}(x).$$

Counts are frozen within an episode. Before a row's episode-start count reaches H, fewer than $2H$ realized visits can be charged to it: the last such episode begins below H and adds at most H visits. Taking expectations, all low-count terms are at most $2SAHw_{\alpha,H}$.

For the remaining terms define, with the high-count restriction inside the sum,

$$Y_K := \sum_{k=1}^{K} \sum_{h=1}^{H-1} \sum_{x: N_k(x) \ge H} \frac{d_{k,h}(x)}{N_k(x)}. \tag{C.6}$$

For each fixed (k,h), Jensen's inequality, measure domination, and (C.5) give

$$\begin{aligned}
&\sum_s \widetilde{d}_{k,h}(s) \left(\sum_{a \in \mathcal{I}_{k,s}} \pi_{k,h}(a \mid s) d_{\mathrm{H}}^2(p_{sa}, \widehat{p}_{k,sa}) \right)^{1/2} \\
&\quad \le \left(\sum_s \sum_{a: N_k(s,a) \ge H} u_{k,h}(s,a) d_{\mathrm{H}}^2(p_{sa}, \widehat{p}_{k,sa}) \right)^{1/2} \\
&\quad \le \alpha^{-(h-1)/2} \left(2\Lambda_K \sum_s \sum_{a: N_k(s,a) \ge H} \frac{d_{k,h}(s,a)}{N_k(s,a)} \right)^{1/2}.
\end{aligned}$$

Multiplying by the Hellinger coefficient in (5.3), summing, and applying Cauchy–Schwarz over (k,h) yields the high-count bound

$$4W_{\alpha,H}\sqrt{\Lambda_K K Y_K}. \tag{C.7}$$

Let $n_k(x)$ be the realized visits to row x in episode k, and condition on $\mathcal{F}_k^+ := \sigma(\mathcal{H}_k, p, \widehat{p}_k)$; the selected π_k is then fixed. The definition of nominal occupancy first gives the conditional identity

$$\mathbb{E}[n_k(x) \mid \mathcal{F}_k^+] = \sum_{h=1}^{H} d_{k,h}(x).$$

The tower property therefore gives, for every nonnegative $\mathcal{H}_k$-measurable function f,

$$\mathbb{E}\left[\sum_{h=1}^{H} d_{k,h}(x) f(N_k(x))\right] = \mathbb{E}[n_k(x) f(N_k(x))]. \tag{C.8}$$

Thus the expectation of (C.6) is at most the corresponding sum with the full-episode realized visits $n_k(x)$. If $N_k(x) \geq H \geq n_k(x)$, put $t = n_k(x)/N_k(x) \in [0,1]$. Since $\log(1+t) \geq t/2$,

$$\frac{n_k(x)}{N_k(x)} \leq 2 \log \frac{N_{k+1}(x)}{N_k(x)}.$$

Telescoping these logarithms separately for every row, and using $N_{K+1}(x) \leq KH$, yields $\mathbb{E}Y_K \leq 2SA \log(eKH)$. Jensen's inequality gives $\mathbb{E}\sqrt{Y_K} \leq \sqrt{\mathbb{E}Y_K}$ in (C.7), and hence the high-count contribution is at most

$$4\sqrt{2}\, W_{\alpha,H} \sqrt{SA\Lambda_K \log(eKH)} \sqrt{K} = C_K \sqrt{K}.$$

Combining high counts, low counts, and failures proves (5.8). The argument used only row-prefix concentration and conditional posterior symmetry, so neither prior factorization nor a policy fixed before $\widehat{p}_k$ is required. □

The proof of Theorem 5.6 needs a pathwise conversion from predictable occupancy sums to realized visits.

Lemma C.3 (Predictable-to-realized conversion)**.** *Let $(\mathcal{G}_k)_{k=0}^K$ be a filtration, let Z_k be $\mathcal{G}_k$-measurable, let $m_k := \mathbb{E}[Z_k \mid \mathcal{G}_{k-1}]$, and suppose $0 \leq Z_k \leq b$ almost surely for $b \geq 0$. For every $\ell > 0$, with probability at least $1 - e^{-\ell}$,*

$$\sum_{k=1}^{K} m_k \leq 2 \sum_{k=1}^{K} Z_k + 2b\ell. \tag{C.9}$$

Proof. If $b = 0$, then $Z_k = m_k = 0$ almost surely and the claim is immediate. Assume $b > 0$. Put $c := 1 - e^{-1}$. Convexity on $[0, b]$ gives $e^{-z/b} \leq 1 - cz/b$, and hence

$$\mathbb{E}[e^{-Z_k/b} \mid \mathcal{G}_{k-1}] \leq 1 - cm_k/b \leq e^{-cm_k/b}.$$

Therefore $\exp\{c \sum_{j \leq k} m_j/b - \sum_{j \leq k} Z_j/b\}$ is a nonnegative supermartingale starting from one. Markov's inequality bounds its terminal value by e^{ℓ} except on an event of probability $e^{-\ell}$. Rearranging and using $c^{-1} < 2$ proves (C.9). □

Proof of Theorem 5.6. We intersect true- and sampled-kernel prefix confidence events, apply the same hybrid simulation decomposition as in expectation, and then use Lemma C.3 separately for low- and high-count visits. For $n \geq 1$, define

$$\mathcal{C}_{x,n}^{\delta} := \left\{ v \in \Delta(\mathcal{S}) : \mathrm{KL}(\overline{p}_{x,n} \| v) \leq \frac{\Lambda_{K,\delta}}{n} \right\}, \qquad \mathcal{C}_{x,0}^{\delta} := \Delta(\mathcal{S}),$$

and put $\mathcal{C}_k^{\delta} := \prod_x \mathcal{C}_{x,N_k(x)}^{\delta}$. The method-of-types failure probability for one row and one prefix is at most $\delta/(4SAHK^2)$. Union bounding over SA rows and at most KH prefixes shows that some true prefix fails with probability at most $\delta/(4K)$. For the current balls, conditional posterior symmetry then gives

$$\mathbb{P}(\widehat{p}_k \notin \mathcal{C}_k^{\delta}) = \mathbb{P}(p \notin \mathcal{C}_k^{\delta}) \leq \frac{\delta}{4K}.$$

A second union bound over episodes shows that all sampled rows also lie in their balls except with probability $\delta/4$.

On this joint confidence event, repeat the hybrid local-error argument leading to (C.7). With $x = (s, a)$, it yields the pathwise inequality

$$\sum_{k=1}^{K} [\rho_1^{\pi_k, p}(s_1) - \rho_1^{\pi_k, \widehat{p}_k}(s_1)] \le w_{\alpha,H} Y_{\mathrm{lo}} + 4W_{\alpha,H}\sqrt{\Lambda_{K,\delta} K Y_{\mathrm{hi}}}, \tag{C.10}$$

where

$$Y_{\mathrm{lo}} := \sum_{k=1}^{K}\sum_{h=1}^{H-1}\sum_{x} d_{k,h}(x)\mathbf{1}\{N_k(x) < H\},$$
$$Y_{\mathrm{hi}} := \sum_{k=1}^{K}\sum_{h=1}^{H-1}\sum_{x:N_k(x)\ge H} \frac{d_{k,h}(x)}{N_k(x)}.$$

We control these predictable sums without taking expectations. Let $n_k(x)$ be the number of visits to row x in episode k and put

$$Z_k^{\mathrm{lo}} := \sum_{x} n_k(x)\mathbf{1}\{N_k(x) < H\}, \qquad Z_k^{\mathrm{hi}} := \sum_{x:N_k(x)\ge H} \frac{n_k(x)}{N_k(x)}.$$

Let $\mathcal{G}_{k-1} := \sigma(\mathcal{H}_k, p, \widehat{p}_k, \pi_k)$ be the analytic information immediately before the episode-k trajectory. These sigma-fields are nested because $\mathcal{H}_{k+1}$ contains the current draw, policy, and trajectory. Conditional nominal occupancy gives

$$\mathbb{E}[Z_k^{\mathrm{lo}} \mid \mathcal{G}_{k-1}] = \sum_{h=1}^{H}\sum_{x} d_{k,h}(x)\mathbf{1}\{N_k(x) < H\},$$
$$\mathbb{E}[Z_k^{\mathrm{hi}} \mid \mathcal{G}_{k-1}] = \sum_{h=1}^{H}\sum_{x:N_k(x)\ge H} \frac{d_{k,h}(x)}{N_k(x)}.$$

These means dominate the episode contributions to Y_{lo} and Y_{hi}. Moreover, $0 \le Z_k^{\mathrm{lo}} \le H$, and

$$0 \le Z_k^{\mathrm{hi}} \le H^{-1}\sum_{x} n_k(x) \le 1.$$

Frozen-count charging gives $\sum_k Z_k^{\mathrm{lo}} < 2SAH$: for each row, fewer than H visits occur before the last episode starting below H, and that episode contributes at most H more. Applying Lemma C.3 with $b = H$ and $\ell = \ell_\delta$ therefore gives, except with probability $\delta/4$,

$$Y_{\mathrm{lo}} \le 4SAH + 2H\ell_\delta. \tag{C.11}$$

For a high-count row, $N_k(x) \ge H \ge n_k(x)$ and hence

$$\frac{n_k(x)}{N_k(x)} \le 2\log\frac{N_{k+1}(x)}{N_k(x)}.$$

Telescoping over each row gives $\sum_k Z_k^{\mathrm{hi}} \le 2SA\log(eKH)$. A second application of Lemma C.3, now with $b = 1$, gives, except with probability $\delta/4$,

$$Y_{\mathrm{hi}} \le 4SA\log(eKH) + 2\ell_\delta. \tag{C.12}$$

The total exceptional probability is at most $\delta/(4K) + 3\delta/4 \le \delta$. Substituting (C.11) and (C.12) into (C.10) proves (5.11).

Finally, exploitation has sampled-model debt at most zero, while calibration has sampled-model debt exactly $-\delta_k$. Hence

$$\sum_{k=1}^{K} [\rho_1^{\pi_k, \widehat{p}_k}(s_1) - \tau] \le -\mathsf{B}_K.$$

Adding this inequality to (5.11) proves (5.12). □

We next prove the ordinary, undistorted reward analogue.

Proof of Lemma 5.8. Ordinary Bellman telescoping has the fixed-horizon identity

$$V_1^{\pi,q}(s_1) - V_1^{\pi,p}(s_1) = \sum_{h=1}^{H-1} \sum_{s,a} d_h^{\pi,p}(s,a) \sum_{s'} (q-p)(s' \mid s,a) V_{h+1}^{\pi,q}(s').$$

Because $0 \le V_{h+1}^{\pi,q} \le H - h$, setting $q = \widehat{p}_k$ gives

$$V_1^{\pi_k, \widehat{p}_k} - V_1^{\pi_k, p} \le \sum_{h=1}^{H-1} (H-h) \sum_{x} d_{k,h}(x) \, \mathrm{TV}(p_x, \widehat{p}_{k,x}). \tag{C.13}$$

The true-prefix and episode-wise posterior-sample events needed here follow from a separate total-variation confidence set. Indeed, the multinomial inequality of Weissman et al. [34] gives

$$\mathbb{P}\left(\mathrm{TV}(\overline{p}_{x,n}, p_x) > \sqrt{L_K/(2n)} \right) \le (2^S - 2) e^{-L_K} \le \frac{1}{8SAHK^2}.$$

The same union bound and conditional posterior identity (C.4), now applied to these TV balls, show that all true and sampled rows satisfy

$$\mathrm{TV}(p_x, \widehat{p}_{k,x}) \le r_k(x), \qquad r_k(x) := \begin{cases} 1, & N_k(x) = 0, \\ \min\{1, \sqrt{2L_K/N_k(x)}\}, & N_k(x) \ge 1. \end{cases} \tag{C.14}$$

outside failure events contributing at most $H/2$ in total. The events are uniform over rows and continuation values, hence remain valid for sample-dependent π_k. On the joint good event, low counts contribute at most $2SAH^2$ by the same frozen-count charging argument.

For high counts, use $H - h \le H$ and (C.8) to replace occupancy by realized visits after taking expectations. Since $N_k(x) \ge H \ge n_k(x)$,

$$\frac{n_k(x)}{\sqrt{N_k(x)}} \le 3 \left(\sqrt{N_{k+1}(x)} - \sqrt{N_k(x)} \right).$$

If $n_k(x) = 0$, both sides vanish. If $n_k(x) > 0$, rationalizing the square-root difference and cancelling $n_k(x)$ shows that the inequality is equivalent to $\sqrt{N_{k+1}(x)} + \sqrt{N_k(x)} \le 3\sqrt{N_k(x)}$, which follows from $n_k(x) \le N_k(x)$ and $1 + \sqrt{2} < 3$. After telescoping and applying Cauchy–Schwarz over the SA rows,

$$\sum_{x} \sum_{k : N_k(x) \ge H} \frac{n_k(x)}{\sqrt{N_k(x)}} \le 3\sqrt{SAKH}.$$

Together with (C.14), the high-count term is at most $3\sqrt{2} H^{3/2} \sqrt{SAKL_K}$.

If $K \le SAH$, the deterministic bound HK is already at most the right side of (5.13). If $K > SAH$, the low-count term is absorbed by the same right side, since

$$2SAH^2 \le 2H^{3/2}\sqrt{SAK} \le \frac{2}{\sqrt{L_K}} H^{3/2}\sqrt{SAKL_K}.$$

The numerical constant 8 dominates these contributions in both cases. This proves the claim without any restriction on how π_k depends on the posterior sample. □

D Deterministic target calculations

Proof of Lemma 5.9. For $k \ge 1$,

$$\Lambda_k \le \lambda_0 + \nu \log k \le A_0 + \nu \log k, \qquad \log(ekH) = m_0 + \log k \le M_0 + \log k,$$

where the first inequality uses $kH + 1 \le k(H + 1)$. Hence

$$C_k\sqrt{k} \le T_k. \tag{D.1}$$

To establish subadditivity, for $t \ge 1$ write

$$G(x) := \sqrt{(A_0 + \nu x)(M_0 + x)}, \qquad T(t) := \mathfrak{c}\sqrt{t}\, G(\log t).$$

Direct differentiation gives

$$G''(x) = -\frac{(A_0 - \nu M_0)^2}{4G(x)^3} \le 0, \qquad T''(t) = \mathfrak{c}t^{-3/2}\left[G''(\log t) - \frac{1}{4}G(\log t)\right] < 0.$$

Extend T linearly on $[0, 1]$ with $T(0) = 0$. Because $A_0 \ge 2\nu$ and $M_0 \ge 2$,

$$\frac{T'_+(1)}{T(1)} = \frac{1}{2} + \frac{\nu}{2A_0} + \frac{1}{2M_0} \le 1,$$

so the extension is nondecreasing and concave. A nonnegative concave function vanishing at zero is subadditive; therefore

$$T_{j+n} \le T_j + T_n \qquad (j, n \in \mathbb{Z}_{\ge 0}). \tag{D.2}$$

It remains to verify the explicit bound on the residual supremum. Concavity of G on $[0, \infty)$ gives

$$G(x) \le G_0 + \gamma_0 x.$$

For an integer $j \ge 1$, put $y = \sqrt{j}$ and use

$$\log y \le \frac{\Delta_* y}{4\gamma_0 \mathfrak{c}} + \log\left(1 + \frac{4\gamma_0\mathfrak{c}}{\Delta_*}\right).$$

It follows that

$$\begin{aligned} T_j - j\Delta_* &\le \mathfrak{c}y\left[G_0 + 2\gamma_0 \log\left(1 + \frac{4\gamma_0\mathfrak{c}}{\Delta_*}\right)\right] - \frac{\Delta_*}{2}y^2 \\ &\le R_*. \end{aligned}$$

The final display is bounded above by R_* after maximizing the concave quadratic in y. This proves the third inequality in (5.14); the first two were proved above. □

E Proof of the planning-complexity result

Proof of Proposition 6.2. The proof has two independent parts. A Subset-Sum reduction establishes the restricted lower bound, while a polynomial-size CVaR epigraph formulation gives the general existential-real upper classification.

Reduce from Subset-Sum. Let the positive integer weights be $w_1, \dots, w_n$ and let the target satisfy $B > 0$. Append a dummy weight $B + 1$ if necessary so that $C := \sum_i w_i > B$, without changing whether a subset sums to B, and reindex the resulting list as $w_1, \dots, w_n$. Set $H = n$, $\tau = B/C$, and at stage i define

$$c_i(\text{take}) = r_i(\text{take}) = w_i/C, \qquad c_i(\text{skip}) = r_i(\text{skip}) = 0.$$

If x_i is the take probability, translation equivariance of CVaR and the deterministic self-loop give

$$V_1^\pi = \sum_i \frac{w_i}{C} x_i, \qquad \rho_1^\pi = \sum_i \frac{w_i}{C} \min\{x_i/\alpha, 1\}. \tag{E.1}$$

For $\alpha \in (0, 1)$, $\min\{x/\alpha, 1\} \geq x$, with equality only at $x \in \{0, 1\}$. Thus risk at most B/C and reward at least B/C force every x_i to be binary and the selected weights to sum exactly to B. The converse is immediate. All-skip has risk zero, so strict feasibility is known.

We next show that the restricted subclass belongs to NP. In an arbitrary one-state, two-action, deterministic-self-loop instance, let x_h be the probability of action one at stage h, and denote the two stage costs by $d_{h,0}$ and $d_{h,1}$. Translation equivariance and the deterministic continuation imply

$$\rho_1^\pi = \sum_{h=1}^{H} \phi_h(x_h), \qquad \phi_h(x) = \mathrm{CVaR}_\alpha^{\mathrm{up}}\big(d_{h,A}\big), \quad A \sim (1 - x, x).$$

If $d_{h,1} \geq d_{h,0}$, then

$$\phi_h(x) = d_{h,0} + (d_{h,1} - d_{h,0}) \min\{x/\alpha, 1\};$$

if $d_{h,1} < d_{h,0}$, the analogous formula is $d_{h,1} + (d_{h,0} - d_{h,1}) \min\{(1 - x)/\alpha, 1\}$. Thus every ϕ_h is rational piecewise affine with two pieces, while the expected reward is affine in $(x_1, \dots, x_H)$. A nondeterministic certificate guesses the active piece at each stage. The corresponding piece-domain inequalities, the risk constraint, the reward-target constraint, and $0 \leq x_h \leq 1$ form a rational linear program. Whenever it is feasible, it has a rational basic feasible solution of bit length polynomial in the input encoding, which supplies a polynomially verifiable certificate. At a piece boundary either adjacent piece may be included because the formulas agree there. Therefore this subclass is in NP and is NP-complete. The displayed reduction is weak and does not establish strong NP-hardness.

For the upper classification, introduce variables for a policy π, reward values $v_h(s)$, risk epigraphs $R_h(s)$, CVaR thresholds $z_h(s)$, and positive parts $t_h(s, a, s')$. Impose the policy simplices, terminal values $v_{H+1} = R_{H+1} = 0$, the reward equations

$$v_h(s) = \sum_a \pi_h(a \mid s) \left[r_h(s, a) + \sum_{s'} p(s' \mid s, a) v_{h+1}(s') \right], \tag{E.2}$$

and

$$t_h(s, a, s') \geq c_h(s, a) + R_{h+1}(s') - z_h(s), \qquad t_h(s, a, s') \geq 0, \tag{E.3}$$

$$R_h(s) \geq z_h(s) + \frac{1}{\alpha} \sum_{a, s'} \pi_h(a \mid s) p(s' \mid s, a) t_h(s, a, s'). \tag{E.4}$$

Add

$$R_1(s_1) \le \tau, \qquad v_1(s_1) \ge v_{\mathrm{tar}}. \tag{E.5}$$

This is a polynomial-size degree-two semialgebraic system. A feasible policy supplies a witness by taking exact reward values, exact nested risks, optimizing CVaR thresholds, and exact positive parts. Conversely, (E.3) implies

$$t_h(s,a,s') \ge [c_h(s,a) + R_{h+1}(s') - z_h(s)]_+.$$

Substitution in (E.4), followed by minimization over a hypothetical threshold, gives $R_h(s) \ge \mathrm{CVaR}_\alpha^{\mathrm{up}}(c_h(s,A) + R_{h+1}(S'))$. Starting from $R_{H+1} = 0$, backward induction and monotonicity of CVaR yield $R_h(s) \ge \rho_h^{\pi,p}(s)$ at every stage. Hence no witness can certify a falsely feasible policy.

The decision formulation is therefore exact and lies in $\exists\mathbb{R}$, which is contained in PSPACE [7]. When the feasible policy set is nonempty, compactness of Π_{M}, continuity of risk, and continuity of reward imply attainment. Standard real-algebraic methods, such as cylindrical algebraic decomposition, decide each target query and can represent an exact algebraic optimizer in finite time [4]. The restricted NP-completeness result rules out a generic polynomial-time exact planner unless $P = NP$. □

F A rare-chain lower bound

This appendix gives the construction and exact calculations underlying Theorem 7.1 and Corollary 7.2.

Proof of Theorem 7.1. We construct two indistinguishable rare branches, compute their nested risks exactly, and then reduce optimal learning to the geometric waiting time until an endpoint reveals the latent model.

The states are the initial state s_1, a failure state f, two public endpoint states g, b, and branch-depth states $z_{i,j}$ for $i \in \{0,1\}$ and $1 \le j \le H-2$. This is $2H$ states. At s_1, action a_i has reward one and cost zero, and moves to $z_{i,1}$ with probability α and to f otherwise. For $1 \le j \le H-3$, both actions at $z_{i,j}$ have zero reward and cost and move to $z_{i,j+1}$ with probability α and to f otherwise. At $z_{i,H-2}$, both actions move with probability α to a model-dependent endpoint and to f otherwise. Kernel $\theta \in \{0,1\}$ makes that endpoint g when $i = \theta$ and b when $i = 1-\theta$. Both actions are identical away from s_1. For every action a, set $p(\cdot \mid f, a) = \delta_f$, $p(\cdot \mid g, a) = \delta_g$, and $p(\cdot \mid b, a) = \delta_b$.

All unspecified rewards and costs are zero, except that $c_H(b,a) = 1$ for both actions; in particular, $c_H(g,a) = c_H(f,a) = 0$. The rewards and costs are common and known, and the two kernels differ only in the last branch transition. Because branch and depth are encoded in the state, the kernels are time-homogeneous.

We next compute nested risk. At stage H, the risk at b is one and that at f, g is zero. On a bad branch, backward induction gives risk one at every surviving branch-depth state: at the last depth and at every earlier depth, the continuation is one with probability α and zero after failure, whose upper-tail CVaR is $\mathrm{CVaR}_\alpha^{\mathrm{up}}(\mathrm{Bernoulli}(\alpha)) = 1$. Let (x_0, x_1) be the learner's initial action distribution. Under model θ, the stage-one random variable is one exactly when the learner chooses the bad branch $1-\theta$ and survives the first rare transition, an event of probability $\alpha x_{1-\theta}$. Therefore

$$\rho_\theta = \mathrm{CVaR}_\alpha^{\mathrm{up}}(\mathrm{Bernoulli}(\alpha x_{1-\theta})) = x_{1-\theta},$$

or equivalently

$$\rho_{\theta=0} = x_1, \qquad \rho_{\theta=1} = x_0. \tag{F.1}$$

Each model has a risk-zero, reward-one policy, so $g(q) = 0$ and $\Delta_* = \tau$. Every policy also receives reward one; therefore reward regret is identically zero.

An episode reaches g or b with probability $\beta = \alpha^{H-1}$, independently of the chosen branch. The endpoint and the learner's known branch choice identify θ. Before any endpoint has been reached, all observed transitions have identical likelihood under the two models, so the posterior remains uniform. Let U_k be the event of no revelation before episode k. Conditional on any history in U_k, (x_0, x_1) may depend arbitrarily on that history, but (F.1) and $x_0 + x_1 = 1$ still give

$$\mathbb{E}_{\theta\sim\mathrm{Unif}\{0,1\}}[\rho - \tau \mid U_k] = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \tag{F.2}$$

After revelation, risk is nonnegative, so per-episode signed debt is at least $-\tau = -1/4$. Moreover, $\mathbb{P}(U_k) = (1-\beta)^{k-1}$. Conditioning on U_k and its complement, the episode-k expected debt is at least

$$\frac{1}{4}\mathbb{P}(U_k) - \frac{1}{4}\mathbb{P}(U_k^c) = \frac{1}{2}(1-\beta)^{k-1} - \frac{1}{4}.$$

Summing this geometric series proves the lower bound in (7.1). It is attained by a learner that chooses either branch before revelation and, once an endpoint identifies θ, always chooses the good branch. Such a learner has debt exactly $1/4$ before revelation and exactly $-1/4$ afterward, while revelation has the constant hazard β. This proves the asserted Bayes-optimal equality.

Here $\beta \le 1/4$, so $K_0 \ge 1/(4\beta)$ and $K_0\beta \le 1/2$. The first two Bonferroni terms give

$$1 - (1-\beta)^{K_0} \ge K_0\beta - \binom{K_0}{2}\beta^2.$$

Substitution into (7.1) yields

$$\mathrm{BR}_\rho(K_0) \ge \frac{K_0}{4}[1 - (K_0 - 1)\beta] \ge \frac{K_0}{8} \ge \frac{1}{32\beta}.$$

This completes the proof. □

Remark F.1 (Heterogeneous rare chain). The same construction matches Corollary B.3. Let the transition after stage h survive with probability α_h and use $\mathrm{CVaR}^{\mathrm{up}}_{\alpha_h}$ at that stage. Since $\mathrm{CVaR}^{\mathrm{up}}_{\alpha_h}(\mathrm{Bernoulli}(\alpha_h)) = 1$, the risk calculation is unchanged, while $\beta = \prod_{h=1}^{H-1}\alpha_h$. Equation (7.1) remains an exact Bayes identity with this β; when $\beta \le 1/4$, (7.2) remains valid as well.

Proof of Corollary 7.2. Let U_k again denote no revelation before episode k. Conditional on a history in U_k, write (x_0, x_1, x_ℓ) for the initial action probabilities. In either model the constrained comparator earns one by choosing its good branch. Hence the episode reward regret is x_ℓ, while (F.1) gives

$$\mathbb{E}_{\theta\sim\mathrm{Unif}\{0,1\}}\left[(\rho - \tau) + \frac{1}{2}(\text{reward regret})\right] = \left[\frac{x_0 + x_1}{2} - \frac{1}{4}\right] + \frac{x_\ell}{2} = \frac{1}{4}.$$

The conditional probability of identification in any unrevealed episode is $\beta(x_0 + x_1) \le \beta$, so $\mathbb{P}(U_k) \ge (1-\beta)^{k-1}$. After identification, risk and reward regret are nonnegative before subtracting τ, and the displayed combined quantity is therefore at least $-1/4$. Conditioning as in the theorem gives a lower bound $\frac{1}{2}(1-\beta)^{k-1} - \frac{1}{4}$ for episode k. Summation proves (7.3).

At $K_0 = \lfloor 1/(2\beta)\rfloor$, the right side is at least $1/(32\beta)$ by the calculation in Theorem 7.1. If $G < 1/(64\beta)$, then

$$\frac{1}{2}F\sqrt{K_0} > \frac{1}{64\beta}.$$

Since $K_0 \le 1/(2\beta)$, this implies $F > \sqrt{2}/(32\sqrt{\beta})$ and proves (7.4). □

G Notation, quantifiers, and comparison axes

The following tables collect the scopes that are easiest to conflate in the main text and compare the nearest methodological lines along the axes relevant to the scoped novelty statement in the introduction.

Table 1: Notation and quantifier scope.

Symbol or phrase	Meaning	Scope and quantifier
$p \sim \mu$, $\mathcal{Q}$	Latent transition kernel and prior support	p is drawn once before interaction and then fixed; model-wise strict feasibility holds for every $q \in \mathcal{Q}$.
$\mathcal{H}_k$, $\widehat{p}_k$	Episode-start history and sampled kernel	Conditional on $\mathcal{H}_k$, p and $\widehat{p}_k$ are independent copies from the posterior μ_k.
Π_{M}, π_k	Randomized physical-state Markov policies	Action randomization is inside every CVaR operator, and π_k is measurable with respect to $\sigma(\mathcal{H}_k, \widehat{p}_k)$.
$g(q)$, Δ_*	Minimum nested risk and induced uniform gap	$\Delta_* = \tau - \sup_{q \in \mathcal{Q}} g(q) > 0$ exists by compactness but is not given to the learner.
B_k, T_k, N_K^{cal}	Risk bank, anytime target, and calibration count	The algorithm uses T_k and observed sampled-model slack, but neither the terminal horizon K nor Δ_*.
$\mathrm{BR}_r(K)$, $\mathrm{BR}_\rho(K)$	Signed Bayesian reward shortfall and signed risk-value debt	Expectations cover the prior, observations, posterior draws, and policy randomization; only one-sided upper bounds are claimed.
"simultaneously for all K"	Anytime expectation and pathwise count bounds	The inequalities in Theorem 4.2 hold for every evaluation horizon along the same algorithmic run.
"fixed K" certificate	High-probability statement in Theorem 5.6	It is pointwise in a fixed admissible measurable policy-selection rule and is not a confidence sequence.
Rational planning result	Decision problem in Definition 6.1	It classifies rationally encoded known-model inputs; it does not implement the oracle for arbitrary real posterior samples.

Table 2: Axis-by-axis comparison with the nearest methodological lines. "Static" means one risk functional applied to an episode return; "nested" means a stagewise time-consistent recursion.

Work	Risk functional and task	Online protocol and comparator	Guarantee or boundary relevant here
This work	Nested upper-tail CVaR constraint with expected-reward objective	Bayesian posterior sampling; randomized physical-state Markov policies with action mixing inside CVaR	Near-$\sqrt{K}$ one-sided signed reward shortfall, K-independent upper bound on expected signed debt, and sparse calibration; exact sampling and planning oracles.
[8, 15]	Iterated-CVaR objective	Episodic online learning; risk-sensitive objective rather than a separate constraint	Objective-regret or sample-efficiency results, not a separate signed-debt guarantee.
[13, 22, 35]	Recursive OCE or dynamic-risk objective	Episodic interaction or a generative-model protocol	Regret or sample complexity for optimizing the risk-sensitive objective.
[9]	Static percentile or trajectory-CVaR constraint	Risk-constrained policy optimization	Static trajectory risk and optimization guarantees rather than nested-risk Bayesian regret.
[17]	Entropic-risk constraint with expected-reward objective	Online finite-horizon learning with optimism and primal–dual control	Sublinear reward and constraint performance for entropic risk, not nested CVaR or K-independent signed debt.
[18, 28]	Additive expected-cost CMDPs	Posterior sampling in episodic Bayesian or communicating average-reward models	Reward and additive-constraint guarantees; the risk recursion and comparator studied here are absent.
[21]	OCE-based risk-aware constraints	Policy optimization under constraint qualifications	Convergence of an OCE-constrained optimizer, not finite-horizon posterior-sampling regret.
[20]	Static return-CVaR objective or chance constraint	Probabilistic forecasts, posterior sampling, and a belief-augmented planner	Different risk functional and comparator; the cited preprint does not state the signed-debt theorem considered here.
[23, 29]	Risk applied to Bayesian model uncertainty	Bayes-adaptive or Bayesian-risk MDP formulations	Risk over epistemic uncertainty rather than a separate nested constraint under a fixed latent kernel.