Learning Chance-Constrained MDPs with Bellman Distributional Certificates

Chenbei Lu
Cornell University AI for Science Institute
Cornell University
Ithaca, NY, USA
chenbei.lu@cornell.edu

Hongyu Yi*
Department of Electrical & Computer Engineering
University of Washington
Seattle, WA, USA
hyyi@uw.edu

Abstract

Safe reinforcement learning (RL) commonly enforces expected-cost constraints, but such expectation-form safety may fail to control the probability of rare high-cost trajectories. Chance-constrained MDPs (CCMDPs) impose a stronger probability-level requirement, but are widely viewed as harder because the chance constraint itself is nonconvex and depends on the full trajectory rather than a Bellman-linear expectation. In this paper, we reveal that this computational difficulty does not necessarily imply a higher statistical price. For tabular discounted CCMDPs with bounded successor support and access to a certified planning oracle, we establish a model-based upper bound, with a matching lower bound up to logarithmic terms. Technically, our key idea is the *Bellman distributional certificate*, which constructs a Bellman recursion for constraint violation probabilities before policy selection. The certificate reduces trajectory-level safety analysis to local transition uncertainty, avoiding a union bound over the potentially exponential number of candidate policies. For stochastic policies, we give a model-free variance-reduced policy-gradient algorithm with a finite-sample expected KKT-residual guarantee and independent validation of every accepted policy. Numerical experiments on synthetic CCMDPs and an IEEE 14-bus energy storage control benchmark illustrate the safety and mechanism behavior of the proposed algorithms.

1 Introduction

Reinforcement learning (RL) is a central framework for online decision-making in uncertain environments [43]. In safety-critical applications, however, maximizing expected reward is not sufficient: the agent must also satisfy certain safety or reliability requirements, the central concern in safe reinforcement learning [1, 24, 49]. A common formulation for safe reinforcement learning is constrained Markov decision processes (CMDPs), which maximize the cumulative reward subject to constraints on expected cumulative costs [4, 9, 19]. This expectation-form constraint is mathematically convenient: expected-cost constraints are linear in the occupancy measure, so a CMDP with known transitions can be formulated as a linear program [4, 23, 47]. However, CMDPs control safety only through expectations. They do not directly control the probability distribution of rare high-cost events such as overloads or failures [10, 14].

Chance-constrained Markov decision processes (CCMDPs) provide a stronger probability-level formulation of safety [33, 37]. Instead of constraining only the average cumulative cost, a CCMDP requires the cumulative trajectory cost to remain below a prescribed threshold with high probability [12]. In the discounted setting, this takes the form

$$\mathbb{P}_\pi\left(\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\le d\right)\ge 1-\delta.$$

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Unlike an expected-cost constraint, this requirement directly limits the probability of high-cost trajectories and therefore better captures reliability requirements in applications such as energy systems, transportation, and financial markets [8, 17, 53]. However, this stronger safety object appears much harder to learn. A chance constraint is generally nonconvex and depends on the full trajectory distribution rather than an expectation of additive costs, so it breaks the Bellman-linear and occupancy-measure structure used in CMDP analyses, which leads to a generally nonconvex feasible set [4, 40, 47].

Learning CCMDPs is challenging for two closely related reasons. First, chance constraints restrict the distribution of cumulative trajectory costs rather than their expectations. This breaks the Bellman-linear and occupancy-measure structure that makes CMDPs tractable, and the resulting feasible set is generally nonconvex [4, 40, 47]. Second, without a reusable certificate, one must estimate violation probabilities for each policy separately, leading to inefficient sample reuse [50]. Addressing these challenges is essential for developing sample-efficient RL algorithms with reliable safety guarantees. See Appendix A for a detailed literature review.

To overcome these challenges, our contributions can be summarized as follows.

Bellman distributional certificates for chance constraints. We introduce Bellman distributional certificates, which track a remaining safety budget to express the chance-constraint violation probability through a Bellman recursion. This replaces separate policy-level safety checks with reusable one-step Bellman updates for violation probabilities, which is the key to enable sample-efficient constraint feasibility certification across candidate policies.

Model-based learning with Bellman certificates. We develop a model-based algorithm for learning a near-optimal deterministic policy with empirical Bellman distributional certificates in CCMDPs. Under Assumption 1, which requires bounded successor support and a certified planning oracle, we prove that the returned policy is chance constraint-feasible and near-optimal with sample complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|[(1-\gamma)^{-3}\varepsilon^{-2}+(1-\gamma)^{-1}\rho^{-2}])$, which significantly improves the best-known deterministic CCMDP learning result [50]. We also prove a matching lower bound up to logarithmic terms, showing that the sample requirements are unavoidable. The proof uses shared row-wise reverse-KL confidence sets and a trajectory-level KL transfer argument, which control the safety of data-dependent policies without a union bound over exponentially many policies.

Model-free policy-gradient algorithm with trajectory certificates. We further develop a model-free algorithm for stochastic policy learning using variance-reduced policy gradients and trajectory certificates. Under standard regularity assumptions, the optimization phase attains an expected KKT-residual guarantee with $\widetilde{O}(\varepsilon_{\mathrm{kkt}}^{-3})$ samples. To our knowledge, this is the first result for finite sample stochastic policy learning with original chance constraints [12, 14, 51, 52].

2 Chance-Constrained Markov Decision Processes

Consider a standard infinite-horizon discounted chance-constrained MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma, \mathcal{C}, \mu \rangle$. Here $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the unknown transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\gamma \in (0, 1)$ is the discount factor, and μ is the initial distribution of the state. The set $\mathcal{C} = \{(c_i, d_i, \delta_i)\}_{i=1}^m$ specifies the following chance constraints:

$$\mathbb{P}_{P,\pi,\mu}\Big(\sum\nolimits_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) \le d_i\Big) \ge 1 - \delta_i, \qquad \forall i \in [m]. \tag{1}$$

where $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the instantaneous safety cost, d_i is the safety budget, and δ_i is the allowed violation probability.

Since the chance constraints depend on the full trajectory, we allow the policy π to also depend on trajectory parameters in addition to the current state. Specifically, a policy $\pi : \mathcal{S} \times \Gamma \to \Delta(\mathcal{A})$ maps current state s and trajectory parameters $g \in \Gamma$ to the distribution simplex over actions. Ordinary Markov policies correspond to the special case where Γ is a singleton.

For any policy π, let $\mathbb{P}_{P,\pi,\mu}$ and $\mathbb{E}_{P,\pi,\mu}$ denote probability and expectation for trajectories generated by P, policy π, and initial distribution μ, including updates of trajectory parameters when present.

We define

$$q_{P,i}(\pi) = \mathbb{P}_{P,\pi,\mu}\Big(\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) > d_i\Big), \qquad J_P(\pi) = \mathbb{E}_{P,\pi,\mu}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big]. \quad (2)$$

For a policy class Π, the feasible set and optimal value are

$$\Pi^\star = \{\pi \in \Pi : q_{P,i}(\pi) \le \delta_i,\ i \in [m]\}, \qquad V_{\mathrm{cc}}^\star = \sup_{\pi \in \Pi^\star} J_P(\pi). \quad (3)$$

3 Bellman Distributional Certificates

We construct upper confidence bounds on chance-constraint violation probabilities. We first truncate the discounted tail and round the remaining budget to obtain a finite violation event. We then bound its probability using a robust Bellman recursion.

3.1 Violation-Probability Bellman Table

Estimating violation probabilities policy by policy is statistically inefficient: samples used to validate one completed policy do not directly certify another, while the deterministic policy class can be exponentially large. We instead evaluate each policy with a Bellman table whose entries are violation probabilities from a state, a rounded remaining budget, and a time index.

For a tail allowance α_{tail} and grid widths $\eta_i > 0$, define

$$H_\alpha = \left\lceil \frac{\log(1/((1-\gamma)\alpha_{\mathrm{tail}}))}{1-\gamma} \right\rceil, \qquad b_i^0 = \left\lfloor \frac{d_i - \alpha_{\mathrm{tail}}}{\eta_i} \right\rfloor, \qquad w_{i,h}(s,a) = \left\lceil \frac{\gamma^h c_i(s,a)}{\eta_i} \right\rceil. \quad (4)$$

Here H_α makes the discounted tail after the certificate horizon at most α_{tail}, b_i^0 is the initial integer budget, and $w_{i,h}(s,a)$ is the integer charge of taking action a in state s at time index h. The initial budget is rounded down and each one-step charge is rounded up, so the finite violation event is conservative for the original chance constraint. The finite budget state is $b = (b_1, \dots, b_m)$, where each coordinate takes integer values $0, \dots, b_i^0$ together with the failure value -1; if $b_i^0 < 0$, the coordinate starts at -1. We write $\mathcal{B}$ for the product of these coordinate sets. The deterministic update used below, denoted $B_h(s,a,b)$, subtracts the charges $w_{i,h}(s,a)$ coordinatewise and sets any exhausted coordinate to -1. Below we write $H = H_\alpha$ when the tail allowance is fixed.

Definition 1 (Bellman violation-probability table)**.** *For a fixed certificate horizon H, policy π, constraint i, and transition kernel P, define $q_{P,i}^\pi(s,b,h)$ as the conditional probability that the i-th rounded budget coordinate becomes negative by the certificate horizon, when the rounded process starts from original state s, rounded remaining-budget vector b, and time index h.*

Lemma 1 (Markov sufficiency of rounded budgets)**.** *Fix a certificate horizon H, budget grid η, and a deterministic policy π whose actions are selected as $a_h = \pi_h(s_h, b_h)$. Then, for each constraint i, the conditional probability of rounded finite-prefix violation after time h, given the trajectory history $\mathcal{F}_h$, depends on $\mathcal{F}_h$ only through (s_h, b_h, h). In particular, there exists a function $q_{P,i}^\pi$ such that*

$$\mathbb{P}_{P,\pi}(\exists t \in \{h, \dots, H\} : b_{i,t} < 0 \mid \mathcal{F}_h) = q_{P,i}^\pi(s_h, b_h, h). \quad (5)$$

Moreover, $q_{P,i}^\pi$ satisfies the Bellman recursion

$$q_{P,i}^\pi(s,b,h) = \sum\nolimits_{s'} P(s' \mid s, \pi_h(s,b)) q_{P,i}^\pi(s', B_h(s, \pi_h(s,b), b), h+1), \quad (6)$$

with boundary condition $q_{P,i}^\pi(s,b,H) = \mathbb{I}\{b_i < 0\}$.

The proof is given in Appendix B. For a fixed policy $\pi_h(s,b)$, the lemma shows that (s,b,h) suffices to evaluate the finite violation probability. For fixed H and grid eta, let $\Pi_{H,\eta}$ denote the deterministic policies that use $\pi_h(s,b)$ for $h = 0, ..., H-1$. All policies in this class follow the same deterministic continuation, fixed in advance, from time H onward.

3.2 Pessimistic Bellman Certificates

Lemma 1 gives the exact Bellman recursion for the finite violation probability when the transition kernel is known. In learning, P is unknown, so the certificate maximizes each Bellman backup over a row-wise confidence set. Since the table represents violation probabilities, this robust backup overestimates the true violation probability whenever the true transition row belongs to the confidence set.

Definition 2 (Pessimistic Bellman distributional certificate)**.** *Fix a policy $\pi \in \Pi_{H,\eta}$, an empirical transition kernel $\widehat{P}$, and row-wise confidence sets $\{\mathcal{C}_{s,a}\}_{s,a}$. For each constraint i, define the empirical violation table $\overline{q}_i^{\pi}(s,b,h)$ backward by $\overline{q}_i^{\pi}(s,b,H) = \mathbb{I}\{b_i < 0\}$, and for $h = H-1,\dots,0$ by*

$$\overline{q}_i^{\pi}(s,b,h) = \sup_{p \in \mathcal{C}_{s,a}} p^{\top} \overline{q}_i^{\pi}(\cdot, B_h(s,a,b), h+1), \qquad a = \pi_h(s,b), \tag{7}$$

The pessimistic Bellman distributional certificate of π is $\overline{q}_{i,H,\eta}(\pi) = \sum_s \mu(s)\overline{q}_i^{\pi}(s,b^0,0)$ for all $i \in [m]$.

When the confidence sets contain all true transition rows, conservative rounding and the robust recursion give $q_{P,i}(\pi) \le \mathbb{P}_{P,\pi,\mu}(b_{i,H} < 0) \le \overline{q}_{i,H,\eta}(\pi)$.

Proposition 1 (Certified safety, Proof in Appendix D)**.** *With high probability over the shared row samples, simultaneously for every $\pi \in \Pi_{H,\eta}$, $\overline{q}_{i,H,\eta}(\pi) \le \delta_i$ for every $i \in [m]$ implies that π is feasible for the original infinite-horizon chance-constrained MDP.*

Why the certificate is reusable. The key statistical point is that one confidence set is constructed for each original transition row and reused at every time and budget state. On the simultaneous row-coverage event, the robust recursion is valid for every policy, including a policy selected after observing the empirical model. Thus the statistical analysis needs neither fresh samples for different Bellman continuations nor a union bound over complete policies.

4 Model-Based Learning with Bellman Certificates

This section develops a model-based learning algorithm built on Bellman distributional certificates. We first describe the learning procedure and then present its finite-sample guarantee together with a separate statistical lower bound.

4.1 Empirical Bellman-Certified Learning

Bellman-certified model-based learning has three steps: environment model estimation, policy-wise Bellman value and certificate computation, and optimization over certified policies.

Model estimation. We assume access to a standard generative model. At each original state-action pair (s,a), the learner draws n independent next-state samples from $P(\cdot \mid s,a)$ and forms one empirical row $\widehat{P}_{s,a}$. The same empirical kernel is reused across all time indices, budget states, constraints, and reward comparisons. The row reward $r(s,a)$ and the costs are known bounded functions. With the known support bound d_0 from Assumption 1, define

$$\kappa_n := \frac{(d_0-1)\log(n+1) + \log(|\mathcal{S}||\mathcal{A}|/\zeta)}{n}, \qquad \mathcal{C}_{s,a} := \{p \in \Delta(\mathcal{S}) : \mathrm{KL}(\widehat{P}_{s,a} \| p) \le \kappa_n\}. \tag{8}$$

We use KL from the empirical distribution to the candidate distribution so that the confidence set can include models that assign positive probability to unobserved successors.

Policy evaluation. For a deterministic policy π, we compute its safety certificates using the KL sets in (8) and the robust recursion in (7).

We estimate its return under the shared empirical kernel. Set $\widehat{V}_H^{\pi}(s,H,b) = 0$ and, for $h = H-1,\dots,0$, compute

$$\widehat{V}_H^{\pi}(s,h,b) = r(s,\pi_h(s,b)) + \gamma\, \widehat{P}(\cdot \mid s,\pi_h(s,b))^{\top} \widehat{V}_H^{\pi}(\cdot, h+1, B_h(s,\pi_h(s,b),b)), \tag{9}$$

and define $\widehat{J}(\pi) = \sum_s \mu(s)\widehat{V}_H^{\pi}(s,0,b^0)$. Thus the robust Bellman distributional certificate checks feasibility, while ordinary empirical Bellman evaluation ranks the feasible candidates.

Finite certified problem. We maximize estimated return over the solver's finite policy class $\Pi_{\mathcal{O}}$, subject to the certified chance constraints:

$$\max_{\pi \in \Pi_{\mathcal{O}}} \widehat{J}(\pi) \qquad \text{s.t.} \quad \overline{q}_{i,H_\alpha,\eta}(\pi) \le \delta_i, \quad i \in [m]. \tag{10}$$

Algorithm 1 Bellman-certified model-based learning

Require: Generative model, finite certified solver, known support bound d_0, samples per row n, tail allowance α_{tail}, grids η_i, margin ρ, confidence ζ, solver gap ξ

Compute H_α, b^0, $w_{i,h}$, and B_h as in (4) and Section 3.
for each original row (s, a) **do**
    Draw n next-state samples, form $\widehat{P}_{s,a}$, and construct $\mathcal{C}_{s,a}$ by (8).
end for
Evaluate robust safety backups by (7) and empirical reward backups by (9), reusing $\widehat{P}$.
Solve (10) over $\Pi_{\mathcal{O}}$ with δ_i replaced by $\delta_i - 3\rho/4$, to gap ξ; denote the output by $\widehat{\pi}$.
if the tightened certified set is empty **then**
    return UNRESOLVED
else
    return $\widehat{\pi}$
end if

Definition 3 (ξ-optimal finite certified solver)**.** *For $\xi \geq 0$, whenever the certified feasible set is nonempty, a finite certified solver returns a policy $\widehat{\pi} \in \Pi_{\mathcal{O}}$ and its Bellman reward and safety tables such that $\overline{q}_{i,H_\alpha,\eta}(\widehat{\pi}) \leq \delta_i$ for all $i \in [m]$ and*

$$\widehat{J}(\widehat{\pi}) \geq \max_{\pi \in \Pi_{\mathcal{O}}:\, \overline{q}_{i,H_\alpha,\eta}(\pi) \leq \delta_i,\ i \in [m]} \widehat{J}(\pi) - \xi. \tag{11}$$

When $\xi = 0$, the solver is exact.

Assumption 1 (Bounded successor support and certified planning)**.** *There is a known integer $d_0 \geq 2$, independent of $|\mathcal{S}|$, $|\mathcal{A}|$, γ, ε, and ρ, such that*

$$|\operatorname{supp} P(\cdot \mid s, a)| \leq d_0, \qquad (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{12}$$

The finite policy class $\Pi_{\mathcal{O}} \subseteq \Pi_{H,\eta}$ is fixed before sampling. Whenever the tightened problem is feasible, the certified planner returns an ξ-optimal policy in the sense of Definition 3.

The bounded-support condition restricts the number of possible successors. The planning requirement is an oracle condition: it controls statistical accuracy but does not assert that the constrained deterministic optimization is polynomial-time solvable. Algorithm 1 summarizes the procedure.

4.2 Sample Complexity Guarantees

As is standard in finite-sample constrained RL, we state the guarantee relative to a comparator with a safety margin. Set $\bar{\eta}_\varepsilon := \min\{\varepsilon/8, (1-\gamma)/128\}$, use $\alpha_{\mathrm{tail}} = \eta_i = \bar{\eta}_\varepsilon$ in the rounded certificate, and let $q^{\mathrm{rnd}}_{P,i}(\pi)$ denote the resulting true rounded violation probability. For $\rho > 0$, define the rounded ρ-interior optimal value

$$V^\star_\rho = \max_{\pi \in \Pi_{\mathcal{O}}} \{J_P(\pi) : q^{\mathrm{rnd}}_{P,i}(\pi) \leq \delta_i - \rho,\ i \in [m]\}, \tag{13}$$

and assume this set is nonempty. Let $\pi^\star_\rho$ be a maximizer. The margin is used only for statistical certification; conservative rounding separately guarantees safety for the original chance constraint. Such interior comparators are standard in finite-sample constrained RL [11, 47].

The output must satisfy the original constraints, but its return is compared with the best policy in $\Pi_{\mathcal{O}}$ that has margin ρ under the rounded test. A policy can satisfy the original constraint but fail the rounded test if its cumulative cost often falls just below d_i. This can happen even with known transitions. Appendix E explains when a policy with margin 2ρ under the original constraints also has margin ρ under the rounded test.

We provide the upper and lower bounds below, with proofs in Appendix E and Appendix F.

Theorem 1 (Bellman-certified upper bound)**.** *There are universal constants $C, c > 0$ such that the following holds. Fix $\zeta \in (0,1)$, $\rho \in (0,1]$, $\varepsilon \in (0,1)$, and $\xi \geq 0$, and assume the set defining $V^\star_\rho$ is nonempty. Suppose Algorithm 1 uses a solver satisfying Assumption 1, with fixed support bound d_0. Run it with $\alpha_{\mathrm{tail}} = \eta_i = \bar{\eta}_\varepsilon$ for every constraint and $n = \lfloor D/(|\mathcal{S}||\mathcal{A}|) \rfloor$ samples per row. If the total generative-model transition-sample budget D satisfies*

$$D \geq C d_0 |\mathcal{S}||\mathcal{A}| \left(\frac{1}{(1-\gamma)^3 \varepsilon^2} + \frac{1}{(1-\gamma)\rho^2} \right) \log^c \left(\frac{8 d_0 |\mathcal{S}||\mathcal{A}|}{\zeta \varepsilon \rho (1-\gamma)} \right). \tag{14}$$

Then with probability at least $1-\zeta$, Algorithm 1 returns a policy $\widehat{\pi}$ satisfying $q_{P,i}(\widehat{\pi}) \le \delta_i$ for every $i \in [m]$ and $J_P(\widehat{\pi}) \ge V_\rho^\star - \varepsilon - \xi$.

Theorem 2 (Model-based CCMDP lower bound)**.** *There are universal constants $c, c_0 > 0$ such that the following holds for every fixed $d_0 \ge 2$, $\gamma \in [1/2, 1)$, $\varepsilon \in (0, c_0/(1-\gamma))$, $\rho \in (0, c_0)$, $|\mathcal{S}| \ge 4$, and $|\mathcal{A}| \ge 2$. Use $\alpha_{\mathrm{tail}} = \eta = \bar{\eta}_\varepsilon$ and take $\Pi_{\mathcal{O}} = \Pi_{H,\eta}$ in (13). Let Π_{det} be the full deterministic policy class represented by the trajectory parameters in Section 2. Any generative-model learner that, on every one-constraint CCMDP satisfying $|\operatorname{supp} P(\cdot \mid s,a)| \le d_0$ and having a nonempty rounded ρ-interior feasible set, returns with probability at least $2/3$ a policy $\widehat{\pi} \in \Pi_{\mathrm{det}}$ satisfying jointly $q_P(\widehat{\pi}) \le \delta$ and $J_P(\widehat{\pi}) \ge V_\rho^\star - \varepsilon$ must, in the worst case, use a total number D of transition samples satisfying*

$$D \ge c|\mathcal{S}||\mathcal{A}| \left(\frac{1}{(1-\gamma)^3 \varepsilon^2} + \frac{1}{(1-\gamma)\rho^2} \right). \tag{15}$$

Sample complexity and comparison to prior CCMDP learning. In Theorem 1, the ε^{-2} term is the usual value-estimation cost, while the ρ^{-2} term is the cost of certifying exact chance safety near the constraint boundary. The proof uses the shared KL sets to control the trajectory law without adding a confidence buffer at each Bellman step. A bound on total variation directly bounds the error in a violation probability. For discounted return, the same bound is multiplied by the return range, at most $1/(1-\gamma)$. The same transition samples must meet both accuracy requirements, so the larger of the two sample sizes suffices, or equivalently their sum up to a constant factor.

For fixed d_0, the upper bound improves the displayed state and horizon dependences of the deterministic CCMDP bound of [50]. Theorem 2 shows that both the reward-learning term and the chance-boundary term are unavoidable, matching the upper bound in these parameters up to logarithmic factors.

Comparison with MDPs and CMDPs. The reward term has the same dependence as in discounted MDPs and relaxed-feasibility CMDPs [2, 6, 47]. Strict-feasibility CMDP bounds use a different feasibility margin, so their safety terms are not directly comparable.

5 Model-Free Stochastic Policy Learning

We use sampled trajectories to optimize expected return subject to the rounded chance constraints, without estimating the transition kernel.

5.1 Trajectory Violation-Probability Policy Gradient

Compared with standard policy gradient for unconstrained MDPs, the chance-constrained setting requires three additional designs. First, the rollout must keep enough trajectory information to evaluate the chance event. Fix the certificate length H and budget grid from Section 3. During each rollout, the algorithm maintains the rounded remaining budget b_t and the clipped time index $\bar{h}_t = \min\{t, H\}$, and uses the stochastic policy class $\Pi_\Theta = \{\pi_\theta(a \mid s, \bar{h}, b) : \theta \in \Theta \subset \mathbb{R}^p\}$, where Θ is an open parameter domain and $(\bar{h}, b)$ are trajectory parameters rather than environment states. From time H onward, the policy uses $(\bar{h}_t, b_t) = (H, b_H)$. Only the violation event is truncated; the reward objective still uses the infinite-horizon return.

Second, we maximize expected return subject to constraints on the rounded violation probabilities. For constraint i, this probability is $q_{P,i}^H(\theta) = \mathbb{P}_{P,\pi_\theta,\mu}(b_{i,H} < 0)$. This is the probability of the same finite event used by the Bellman distributional certificate.

Third, the optimization step must handle probability constraints estimated from trajectories. Algorithm 2 uses variance-reduced policy gradient [51, 52] for the tightened problem. We estimate reward gradients from discounted rollouts, and chance-violation gradients from length-H violation indicators and likelihood-ratio scores.

Let $\bar{J}_P(\theta) = (1-\gamma)J_P(\pi_\theta)$, $f_H(\theta) = -\bar{J}_P(\theta)$, and $g_i^H(\theta) = q_{P,i}^H(\theta) + 2\rho - \delta_i$. The 2ρ tightening leaves room for optimization and validation error. We write $g^H(\theta) \le 0$ as $g^H(\theta) + z = 0$ with $z \ge 0$. The algorithm uses the smooth penalty $\Phi_\beta(\theta, z) = f_H(\theta) + (\beta/2)\|g^H(\theta) + z\|_2^2$ with fixed β. The proximal term h is zero for $z \ge 0$ and $+\infty$ otherwise, so the proximal step keeps the slack nonnegative.

We estimate the candidate's rounded violation probabilities on fresh trajectories. We accept it if each upper confidence bound is at most the allowed violation probability, and return UNRESOLVED otherwise. When the bounds hold, conservative rounding guarantees that the accepted policy satisfies the original constraints.

Algorithm 2 Model-free trajectory-sampling violation-probability policy gradient

Require: Rollout access, budget-aware policy class $\pi_\theta(a \mid s, \bar{h}, b)$, certificate length H, grids η_i, margin ρ, confidence ζ, KKT tolerance $\varepsilon_{\mathrm{kkt}}$

1: Initialize $x_0 = (\theta_0, z_0)$ using the feasible-initialization and merit-gap clause of Assumption 3.
2: **for** $k = 0, \dots, T-1$ **do**
3: Draw reward rollouts and length-H violation-indicator rollouts, maintaining $(\bar{h}_t, b_t)$ and sharing the full-budget rollout across constraints.
4: Update the recursive estimator v_k of the smooth penalty gradient $\nabla\Phi_\beta(x_k)$, with periodic independent refresh batches.
5: Take the analyzed proximal penalty step $x_{k+1} = \mathrm{prox}_{\eta_{\mathrm{opt}} h}(x_k - \eta_{\mathrm{opt}} v_k)$, using independent samples for Jacobian–constraint products.
6: **end for**
7: Draw R uniformly from $\{0, \dots, T-1\}$ and take the policy component θ_{cand} of the already-computed x_{R+1}. Validate it with fresh length-H trajectories; accept only if $\widehat{q}^{H}_{i,\mathrm{val}}(\theta_{\mathrm{cand}}) + \rho/2 \le \delta_i$ for all i, and otherwise return UNRESOLVED.

5.2 Sample Complexity Guarantees

Assumption 2 (Local regularity)**.** *The objective and constraints are locally smooth. The policy scores, oracle variances, and multipliers are bounded. The parameter iterates stay in a compact subset of the open domain, and the initial policy is feasible for the tightened problem [29, 51, 52]. Assumption 3 gives the full conditions, including the additional residual-domination condition* (35)*.*

Define the KKT residual:

$$\mathcal{R}_H(\theta) = \min_{\lambda \in [0,\Lambda]^m} \left\{ \left\| \nabla f_H(\theta) + \sum\nolimits_i \lambda_i \nabla g_i^H(\theta) \right\|_2 + \|[g^H(\theta)]_+\|_1 + \sum\nolimits_i |\lambda_i g_i^H(\theta)| \right\}, \quad (16)$$

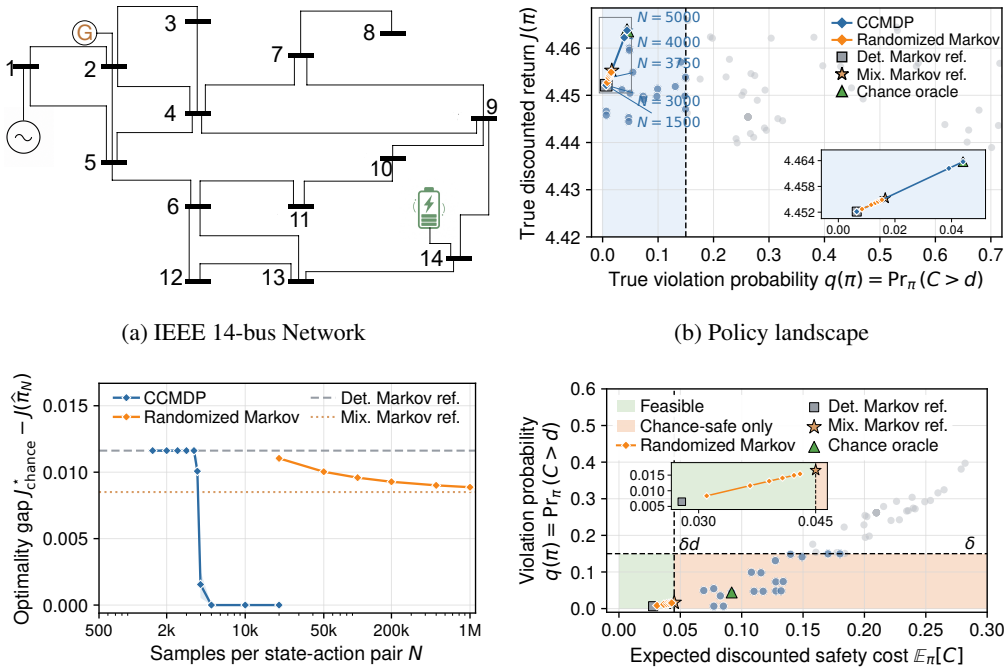
where Λ is the local multiplier bound from Assumption 2.

Theorem 3 (Model-free trajectory certificate)**.** *There is a universal constant $C_{\mathrm{pg}} > 0$ such that the following holds. Fix $\alpha_{\mathrm{tail}} \in (0, (1-\gamma)^{-1})$, grid widths $\eta_i > 0$, margin $\rho \in (0, 1]$, KKT accuracy $\varepsilon_{\mathrm{kkt}} \in (0, 1)$, and confidence $\zeta \in (0, 1)$. Run Algorithm 2 with $H = \lceil \log(1/((1-\gamma)\alpha_{\mathrm{tail}}))/(1-\gamma) \rceil$, and use $M_{\mathrm{val}} = \left\lceil 2\log(2m/\zeta)/\rho^2 \right\rceil$ fresh validation trajectories per constraint. Let K_{opt} denote the resulting problem-dependent local optimization constant. Under Assumptions 2 and 3, it suffices that the expected total number $D := \mathbb{E}[N_{\mathrm{trans}}]$ of environment transitions for optimization and validation satisfies*

$$\begin{aligned} D \ge & \frac{C_{\mathrm{pg}} K_{\mathrm{opt}} G^2 m}{\varepsilon_{\mathrm{kkt}}^3 (1-\gamma)^3} \log^3\left(\frac{2}{(1-\gamma)\alpha_{\mathrm{tail}}}\right) \log^4\left(\frac{8 K_{\mathrm{opt}} G^2 m}{\varepsilon_{\mathrm{kkt}} \rho \zeta (1-\gamma)^3 \alpha_{\mathrm{tail}}}\right) \\ & + \frac{C_{\mathrm{pg}} m}{\rho^2 (1-\gamma)} \log\left(\frac{2}{(1-\gamma)\alpha_{\mathrm{tail}}}\right) \log\left(\frac{2m}{\zeta}\right). \end{aligned} \quad (17)$$

Then the optimization phase returns a candidate satisfying $\mathbb{E}[\mathcal{R}_H(\theta_{\mathrm{cand}})] \le \varepsilon_{\mathrm{kkt}}$. Moreover, with probability at least $1 - \zeta$ over the independent validation trajectories, Algorithm 2 either returns validation unresolved or accepts a policy feasible for the original infinite-horizon CCMDP. On the same event, if the returned candidate satisfies $q^H_{P,i}(\theta_{\mathrm{cand}}) \le \delta_i - \rho$ for all i, then it is accepted.

Discussion. The first term in (17) is the optimization cost, with local constants collected in K_{opt} (Appendix H). The second is the cost of estimating the rounded violation probabilities on fresh trajectories. Bernoulli mean estimation with length-H rollouts gives the validation cost $\widetilde{O}(m/((1-\gamma)\rho^2))$ in transition samples. The optimization guarantee is about stationarity: it neither bounds the gap to the global optimum nor guarantees that the candidate passes validation.



(a) IEEE 14-bus Network

(b) Policy landscape

(c) Optimality gap

(d) Expected-cost comparison

Figure 1: IEEE 14-bus storage control. **(a)** DC network with the battery at bus 14. **(b)** Return–violation landscape of the candidate policy class. Blue and gray dots denote chance-feasible and chance-infeasible policies. Blue and orange diamonds show mean performance of CCMDP and randomized Markov-CMDP, respectively. The indicated N values denote transition samples per state-action pair. The gray square, light-orange star, and green triangle mark the true-model deterministic Markov reference, mixture Markov reference, and deterministic chance oracle, respectively. **(c)** Optimality gap $J^{\star}_{\text{chance}} - J(\widehat{\pi}_N)$ versus N. The two curves and shaded bands show the mean and ± 1.96 empirical standard errors from 60 independent empirical-model trials per sample size, conditional on a policy being returned. The gray dashed and light-orange dotted lines show the deterministic and mixture Markov reference gaps, respectively. **(d)** Violation probability versus expected discounted safety cost. The horizontal and vertical dashed lines mark $q_P(\pi) = \delta = 0.15$ and $\mathbb{E}[C_\pi] = \delta d = 0.045$, respectively. The green region satisfies the Markov expected-cost condition; the orange region contains chance-feasible policies excluded by that condition. Orange diamonds trace the learned mixtures. The reference markers are as in (b).

6 Numerical Studies

We test Algorithm 1 on a finite-state CCMDP built from a DC power-flow simulator of the IEEE 14-bus system [46]. The state records storage state of charge, time block, and load regime; actions charge, discharge, or idle a storage device placed at bus 14. Rewards measure normalized operating benefit, while the safety cost measures normalized transmission-line overload severity. The chance constraint limits the probability that the discounted cumulative overload severity exceeds a prescribed threshold. Appendix I.2 gives the complete experimental specification. Additional synthetic diagnostics and model-free variance-reduced policy-gradient (VR-PG) experiments, including optimization trajectories and independent policy validation, are demonstrated in Appendix I.1.

We compare the Bellman-certified CCMDP selector in Algorithm 1 with a Markov-CMDP baseline using initial policy mixtures and a statistical constraint margin [31, 47]. The CCMDP selector reuses one empirical transition kernel for reverse-KL robust safety backups and finite-horizon reward evaluation. Markov-CMDP maximizes the same empirical return over mixing weights whose weighted robust expected-cost upper bound is at most $\delta d = 0.045$, using the same KL confidence sets.

One component policy is drawn at the start of each trajectory and followed throughout. The chance-constrained oracle maximizes true return subject to $q_P(\pi) \leq \delta$ within the original deterministic class.

The curves in Figure 1(c) use only trials that return a policy. At $N = 1500$, CCMDP returns a policy in 59 of 60 trials. The Markov-CMDP baseline returns no policy at any tested $N \leq 10{,}000$.

At smaller sample sizes, the policies returned by the algorithm coincide with the deterministic Markov-CMDP reference, with return 4.45213. As the confidence sets shrink, the selected policies approach the oracle (Figure 1). At $N = 5000$ samples per state-action pair, all 60 trials select the oracle, with return 4.46376. Markov-CMDP's mean return increases from 4.45271 at $N = 20{,}000$ to 4.45489 at $N = 10^6$, approaching the true-model mixture reference 4.45525. This reference retains a gap of 0.00851 to the deterministic chance oracle because of the expected-cost constraint. The experiment thus separates finite-sample conservatism of the robust certificate from the conservatism of the expected-cost surrogate.

7 Conclusion

We introduced Bellman distributional certificates for learning infinite-horizon CCMDPs. The certificate turns chance safety into local violation-probability Bellman updates. We also proved a lower bound separating the reward-learning term from the chance-boundary certification term. Under the fixed bounded-successor-support and certified-planning conditions in Assumption 1, the model-based upper bound achieves exact safety and matches these primary dependences up to logarithmic factors. We developed a model-free policy-gradient method with independent safety validation. Our experiments use a synthetic CCMDP and an IEEE 14-bus storage benchmark.

Limitations and Future Directions. The model-based guarantee is limited to deterministic policies, assumes a fixed bound on the number of successors of every state–action pair, and relies on the certified planning oracle in Assumption 1. The model-free guarantee applies to stochastic policies, but provides only an approximate KKT-residual guarantee and may return UNRESOLVED after validation. Extending the model-based analysis to broader policy classes, transition structures, and concrete solvers, and obtaining global model-free guarantees, are left for future work.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 22–31. PMLR, 2017. Sydney, Australia, August 6–11, 2017.

[2] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020. Graz, Austria (virtual), July 9–12, 2020.

[3] Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, 2020. doi: 10.1109/TIT.2020.2996134. URL `https://doi.org/10.1109/TIT.2020.2996134`.

[4] Eitan Altman. *Constrained Markov Decision Processes*. Routledge, London, UK, 1st edition, 1999. doi: 10.1201/9781315140223. URL `https://doi.org/10.1201/9781315140223`.

[5] Rashid Alyassi and Majid Khonji. Dual formulation for chance constrained stochastic shortest path with application to autonomous vehicle behavior planning, 2023. URL `https://arxiv.org/abs/2302.13115`.

[6] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013. doi: 10.1007/s10994-013-5368-1. URL `https://hal.science/hal-00831875`.

[7] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*,

volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

[8] Patrizia Beraldi and Maria Elena Bruni. Enhanced indexation via chance constraints. *Operational Research*, 22:1553–1573, 2022. doi: 10.1007/s12351-020-00594-2. URL https://doi.org/10.1007/s12351-020-00594-2.

[9] Frederick J. Beutler and Keith W. Ross. Optimal policies for controlled markov chains with a constraint. *Journal of Mathematical Analysis and Applications*, 112(1):236–252, 1985. doi: 10.1016/0022-247X(85)90288-4. URL https://doi.org/10.1016/0022-247X(85)90288-4.

[10] Vivek S. Borkar and Rahul Jain. Risk-constrained markov decision processes. *IEEE Transactions on Automatic Control*, 59(9):2574–2579, 2014. doi: 10.1109/TAC.2014.2309262.

[11] Max Buckley, Konstantinos Papathanasiou, and Andreas Spanopoulos. Primal-dual sample complexity bounds for constrained markov decision processes with multiple constraints. *arXiv preprint arXiv:2503.06751*, 2025. URL https://arxiv.org/abs/2503.06751.

[12] Weiqin Chen, Dharmashankar Subramanian, and Santiago Paternain. Probabilistic constraint for safety-critical reinforcement learning. *IEEE Transactions on Automatic Control*, 69(10):6789–6804, 2024. doi: 10.1109/TAC.2024.3379246.

[13] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

[14] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18 (167):1–51, 2018. URL https://jmlr.org/papers/v18/15-636.html.

[15] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8103–8112, Red Hook, NY, USA, 2018. Curran Associates Inc.

[16] Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, pages 2892–2901, 2018. doi: 10.1609/aaai.v32i1.11791.

[17] Emiliano Dall'Anese, Kyri Baker, and Tyler Summers. Chance-constrained AC optimal power flow for distribution systems with renewables. *IEEE Transactions on Power Systems*, 32(5):3427–3438, 2017. doi: 10.1109/TPWRS.2017.2656080. URL https://doi.org/10.1109/TPWRS.2017.2656080.

[18] Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

[19] Cyrus Derman and Arthur F. Veinott. Constrained markov decision chains. *Management Science*, 19(4-part-1):389–390, 1972. doi: 10.1287/mnsc.19.4.389.

[20] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R. Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.neurips.cc/paper/2020/hash/5f7695debd8cde8db5abcb9f161b49ea-Abstract.html.

[21] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo R. Jovanovic. 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. PMLR, 2021. URL https://proceedings.mlr.press/v130/ding21d.html.

[22] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020. URL https://arxiv.org/abs/2003.02189.

[23] Eugene A. Feinberg and Adam Shwartz. Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21(4):922–945, 1996. doi: 10.1287/moor.21.4.922. URL https://doi.org/10.1287/moor.21.4.922.

[24] Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. URL https://jmlr.org/papers/v16/garcia15a.html.

[25] Alessandro Giuseppi and Antonio Pietrabissa. Chance-constrained control with lexicographic deep reinforcement learning. *IEEE Control Systems Letters*, 4(3):755–760, 2020. doi: 10.1109/LCSYS.2020.2979635.

[26] William B. Haskell and Rahul Jain. Stochastic dominance-constrained markov decision processes. *SIAM Journal on Control and Optimization*, 51(1):273–303, 2013. doi: 10.1137/120874679.

[27] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49:193–208, 2002.

[28] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 33:12861–12872, 2020.

[29] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications*, 87(1), 2024. doi: 10.1007/s10589-023-00521-z.

[30] Hao Liang and Zhi-Quan Luo. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *Journal of Machine Learning Research*, 25(221):1–56, 2024. URL http://jmlr.org/papers/v25/22-1253.html.

[31] Xingtu Liu, Lin F. Yang, and Sharan Vaswani. Sample complexity bounds for linear constrained MDPs with a generative model. *arXiv preprint arXiv:2507.02089*, 2025. URL https://arxiv.org/abs/2507.02089.

[32] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007. doi: 10.1137/050622328. URL https://doi.org/10.1137/050622328.

[33] Masahiro Ono, Marco Pavone, Yoshiaki Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Auton. Robots*, 39 (4):555–571, December 2015. ISSN 0929-5593. doi: 10.1007/s10514-015-9467-7. URL https://doi.org/10.1007/s10514-015-9467-7.

[34] Santiago Paternain, Miguel Calvo-Fullana, Luiz F. O. Chamon, and Alejandro Ribeiro. Learning safe policies via primal-dual methods. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6491–6497, 2019. doi: 10.1109/CDC40024.2019.9029423. Nice, France, December 11–13, 2019.

[35] Baiyu Peng, Yao Mu, Yang Guan, Shengbo Eben Li, Yuming Yin, and Jianyu Chen. Model-based actor-critic with chance constraint for stochastic system. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4694–4700, 2021. doi: 10.1109/CDC45484.2021.9683748. Austin, TX, USA, December 13–17, 2021.

[36] Baiyu Peng, Jingliang Duan, Jianyu Chen, Shengbo Eben Li, Genjin Xie, Congsheng Zhang, Yang Guan, Yao Mu, and Enxin Sun. Model-based chance-constrained reinforcement learning via separated proportional-integral lagrangian. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):466–478, 2024. doi: 10.1109/TNNLS.2022.3175595.

[37] Samuel Pfrommer, Tanmay Gautam, Alec Zhou, and Somayeh Sojoudi. Safe reinforcement learning with chance-constrained model predictive control. In Roya Firoozi, Negar Mehr, Esen Yel, Rika Antonova, Jeannette Bohg, Mac Schwager, and Mykel Kochenderfer, editors, *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pages 291–303. PMLR, 23–24 Jun 2022. URL https://proceedings.mlr.press/v168/pfrommer22a.html.

[38] Pedro Santana, Sylvie Thiébaux, and Brian Williams. RAO*: An algorithm for chance-constrained POMDP's. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3308–3314, 2016. doi: 10.1609/aaai.v30i1.10423.

[39] Harsh Satija, Philip Amortila, and Joelle Pineau. Constrained markov decision processes via backward value functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8502–8511. PMLR, 2020. URL https://proceedings.mlr.press/v119/satija20a.html.

[40] Xun Shen, Shuo Jiang, Akifumi Wachi, Kazumune Hashimoto, and Sebastien Gros. Flipping-based policy for chance-constrained markov decision processes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 64377–64411. Curran Associates, Inc., 2024. doi: 10.52202/079017-2053. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/760b5def8dcb1156aac454e9c0f5f406-Paper-Conference.pdf.

[41] Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5729–5738. PMLR, 2019. URL https://proceedings.mlr.press/v97/shen19d.html.

[42] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

[43] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

[44] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[45] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkfrvsA9FX.

[46] University of Washington. Power systems test case archive: 14 bus power flow test case. https://labs.ece.uw.edu/pstca/pf14/pg_tca14bus.htm. Power Systems Test Case Archive.

[47] Sharan Vaswani, Lin Yang, and Csaba Szepesvari. Near-optimal sample complexity bounds for constrained mdps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3110–3122. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/14a5ebc9cd2e507cd811df78c15bf5d7-Paper-Conference.pdf.

[48] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9797–9806. PMLR, 2020. URL https://proceedings.mlr.press/v119/wachi20a.html.

[49] Akifumi Wachi, Xun Shen, and Yanan Sui. A survey of constraint formulations in safe reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/913. URL https://doi.org/10.24963/ijcai.2024/913. Jeju, South Korea, August 3–9, 2024.

[50] Hongyu Yi, Chenbei Lu, and Chenye Wu. On the optimal deterministic policy learning in chance-constrained markov decision processes. *IEEE Control Systems Letters*, 9:2217–2222, 2025. doi: 10.1109/LCSYS.2025.3610666.

[51] Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods, 2020. URL https://arxiv.org/abs/2003.04302.

[52] Junyu Zhang, Chengzhuo Ni, Zheng Yu, Csaba Szepesvári, and Mengdi Wang. On the convergence and sample efficiency of variance-reduced policy gradient method. In *Advances in Neural Information Processing Systems*, volume 34, pages 2228–2240, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/11c484ea9305ea4c7bb6b2e6d570d466-Abstract.html.

[53] Yi Zhao, Ronghui Liu, Xi Zhang, and Anthony Whiteing. A chance-constrained stochastic approach to intermodal container routing problems. *PLOS ONE*, 13(2):e0192275, 2018. doi: 10.1371/journal.pone.0192275. URL https://doi.org/10.1371/journal.pone.0192275.

Roadmap for the appendix

The Appendix is organized as follows:

- Appendix A provides additional related work on CMDPs, CCMDPs, risk-sensitive RL, distributional RL, and post-selection certification.
- Appendix B proves the Markov sufficiency of the rounded remaining-budget state and the Bellman recursion for finite-prefix violation probabilities.
- Appendix C proves the conservative rounding lemma that connects the finite rounded violation event to the original infinite-horizon chance constraint.
- Appendix D establishes simultaneous KL row coverage and proves the certified-safety proposition.
- Appendix E proves the sample-complexity upper bound for Bellman-certified model-based learning.
- Appendix F proves the model-based CCMDP lower bound.
- Appendix G illustrates the computational hardness of exact deterministic planning for CCMDPs, showing that exact reward maximization over deterministic policies under a discounted chance constraint is NP-hard.
- Appendix H presents the model-free trajectory-certificate analysis, including the policy-gradient oracle identities and the proof of the model-free sample-complexity guarantee.
- Appendix I provides concise experimental settings for the synthetic diagnostics and the IEEE 14-bus storage-control experiment.

A Related Work

Our work contributes to a few key literatures within the safe RL and theoretical RL community. We discuss each in turn below.

Safe RL and Constrained MDPs. Safe reinforcement learning studies sequential decision-making problems in which the learner must optimize reward while satisfying safety requirements [24, 49]. A large part of this literature builds on constrained Markov decision processes, where safety is modeled through constraints on expected cumulative costs [4, 9, 19]. This formulation has enabled linear-programming and dynamic-programming characterizations in known models, as well as Lagrangian, primal-dual, policy-gradient, and Lyapunov-based methods for learning and policy optimization [1, 4, 15, 20, 34, 45]. It has also supported finite-sample analyses for CMDP learning under exploration, generative-model, or primal-dual settings [11, 21, 22, 47]. However, expected-cost constraints do not directly certify trajectory-level reliability. A policy may have small expected cost while still assigning non-negligible probability to unsafe trajectories, which motivates risk-sensitive, percentile, and chance-constrained formulations [10, 14, 26]. In contrast, our work focuses on chance constraints, where the safety requirement is imposed on the probability of violating a cumulative-cost threshold. This distinction changes both the algorithmic and statistical nature of the problem: the learner must control a tail-probability event rather than a first moment.

Chance-Constrained MDPs. Chance-constrained MDPs provide a direct way to model trajectory-level safety by requiring the probability of cumulative cost violation to stay below a prescribed threshold [12, 14, 33, 40], which are widely adopted in various practical scenarios [8, 17, 53]. This formulation is substantially harder than expected-cost CMDPs: the constraint depends on the distribution of the cumulative cost around a threshold, the feasible set is generally nonconvex, and small model errors can change feasibility by moving probability mass across the threshold [10, 14, 18, 26, 40, 50]. Known-model approaches address these difficulties through budget augmentation, dynamic programming over cost distributions, or search-based planning in fully or partially observable models [5, 33, 38]. Learning and control approaches often use chance-constraint approximations, model-predictive control, actor-critic or Lagrangian updates, probabilistic-constraint gradients, or risk/percentile surrogates such as CVaR [12–14, 25, 32, 35–37, 44]. These methods are useful computationally or empirically, but they typically do not give finite-sample guarantees for the original chance event after a policy has been selected from data. The closest deterministic CCMDP learning

result estimates policy-level violation probabilities for stationary deterministic policies [50]. Our work keeps the chance constraint itself as the certified object. On the model-based side, we use Bellman distributional certificates to certify the policy returned by a finite planner; on the model-free side, we optimize and validate the rounded violation probability directly from trajectories.

Finite-Sample Analysis for Model-Based Algorithms. Our model-based algorithm follows the standard model-based RL approach, where the learner estimates the transition model from a generative model and then plans in the empirical model. Model-based methods for discounted MDPs have been extensively studied [2, 6, 27, 28, 42], achieving the minimax-optimal sample complexity $\widetilde{O}(|\mathcal{S}||\mathcal{A}|\varepsilon^{-2}(1-\gamma)^{-3})$ in the generative-model setting. Finite-sample guarantees have also been developed for CMDPs with expected cumulative cost constraints [11, 21, 22, 47]; in this setting, the leading sample complexity has the same $|\mathcal{S}||\mathcal{A}|\varepsilon^{-2}(1-\gamma)^{-3}$ dependence, while exact constraint satisfaction typically requires a strict-feasibility margin. Chance-constrained MDPs are harder because feasibility is determined by a tail-probability event rather than an expected cumulative cost. The closest deterministic CCMDP learning result estimates policy-level violation probabilities after model learning with sample complexity $\widetilde{O}(|\mathcal{S}|^2|\mathcal{A}|(1-\gamma)^{-4}(\varepsilon^{-2}+\rho^{-2}))$ when translated to our notation [50]. In contrast, under fixed bounded successor support and access to a certified planning oracle, our upper bound has primary dependence $\widetilde{O}(|\mathcal{S}||\mathcal{A}|[(1-\gamma)^{-3}\varepsilon^{-2}+(1-\gamma)^{-1}\rho^{-2}])$. The separate deterministic-policy lower bound has the same primary parameter dependences over the same bounded-successor problem class and uses the same rounded interior comparator.

Finite-Sample Analysis for Model-Free Algorithms. Our model-free algorithm follows the policy-optimization view of RL, where the learner directly updates a policy from sampled trajectories without estimating the transition model [43]. Model-free methods for safe RL and CMDPs have been widely studied through Lagrangian, primal-dual, natural policy-gradient, reward-constrained, Lyapunov-based, and backward-value-function approaches [1, 15, 20, 21, 39, 45, 48]. These works provide important algorithmic and finite-sample guarantees for expected-cost constraints, but they do not directly control the probability of a trajectory-level violation event. A parallel line of work studies risk-sensitive or chance-constrained policy learning through CVaR objectives, chance-constraint approximations, model-predictive control, actor-critic updates, or probabilistic-constraint gradients [10, 12, 13, 25, 35–37, 44]. These methods move closer to event-level safety, but their guarantees typically concern surrogate risks, approximations, asymptotic convergence, or local stationarity rather than finite-sample certification of the original chance constraint after learning. To our knowledge, our result is the first to combine an explicit finite-sample approximate-KKT bound for the direct rounded chance objective with an independent post-learning certificate for the original chance event. Technically, our algorithm combines a trajectory-level violation-probability gradient estimator with stochastic augmented-Lagrangian and variance-reduced policy-gradient ideas [29, 41, 51, 52], and then certifies the final data-dependent policy using an independent validation batch.

Distributional RL and Post-Selection Certification. Distributional RL learns the distribution of cumulative return and is often used to improve value estimation or optimize risk-sensitive criteria [7, 16, 30]. Our use of "distributional" is different: we do not learn the full return distribution or replace the chance constraint with a surrogate risk measure. Instead, the certificate upper bounds the violation probability of the original threshold event. The main statistical issue is post-selection: the final policy is chosen after inspecting data, so fixed-policy validation is not sufficient. Our model-based algorithm addresses this by certifying Bellman violation-probability entries before policy selection, while our model-free algorithm uses an independent validation batch for the final learned policy.

B Proof of Lemma 1

Proof. Fix a constraint i. Let $\mathcal{F}_h$ be the history up to time h, including s_h and b_h. Define the future failure event

$$E_{i,h} = \{\exists t \in \{h, \dots, H\} : b_{i,t} < 0\}.$$

We show by backward induction that its conditional probability depends only on (s_h, b_h, h).

At $h = H$, no transition remains. Thus

$$\mathbb{P}_{P,\pi}(E_{i,H} \mid \mathcal{F}_H) = \mathbb{I}\{b_{i,H} < 0\}.$$

This is the terminal value of $q^{\pi}_{P,i}$. Now let $h < H$ and assume the result holds at time $h+1$. Given $\mathcal{F}_h$, the action is $a_h = \pi_h(s_h, b_h)$. The next state follows P, and the next budget is

$$b_{h+1} = B_h(s_h, a_h, b_h).$$

The event from time h is the event from time $h+1$ after this update. The induction hypothesis and the tower property give

$$\mathbb{P}_{P,\pi}(E_{i,h} \mid \mathcal{F}_h) = \sum\nolimits_{s'} P(s' \mid s_h, a_h)\, q^{\pi}_{P,i}(s', B_h(s_h, a_h, b_h), h+1) = q^{\pi}_{P,i}(s_h, b_h, h),$$

where the last equality is the Bellman recursion. The last expression uses the history only through (s_h, b_h, h). This proves the induction step and the lemma. □

C Conservative Rounding

Lemma 2 (Conservative rounding)**.** *For every policy $\pi \in \Pi_{H,\eta}$ and constraint $i \in [m]$, the original infinite-horizon violation probability is bounded by the rounded finite-prefix violation probability:*

$$q_{P,i}(\pi) \le q^{\mathrm{rnd}}_{P,i}(\pi). \tag{18}$$

Proof. Set $H = H_\alpha$. Fix a trajectory and a constraint i. Suppose the rounded test does not fail by time H. Then

$$\sum\nolimits_{h=0}^{H-1} w_{i,h}(s_h, a_h) \le b^0_i.$$

Each rounded charge is an upper bound, and b^0_i is rounded down. Hence

$$\sum\nolimits_{h=0}^{H-1} \gamma^h c_i(s_h, a_h) \le \eta_i \sum\nolimits_{h=0}^{H-1} w_{i,h}(s_h, a_h) \le \eta_i b^0_i \le d_i - \alpha_{\mathrm{tail}}.$$

The cost after time H is at most

$$\sum\nolimits_{h=H}^{\infty} \gamma^h c_i(s_h, a_h) \le \sum\nolimits_{h=H}^{\infty} \gamma^h = \frac{\gamma^H}{1-\gamma} \le \alpha_{\mathrm{tail}}.$$

The two bounds give $\sum_{h=0}^{\infty} \gamma^h c_i(s_h, a_h) \le d_i$. Thus, if the original constraint fails, the rounded test must also fail. Taking probabilities gives $q_{P,i}(\pi) \le q^{\mathrm{rnd}}_{P,i}(\pi)$. □

D Proof of Proposition 1

We first show that every true transition row belongs to its KL confidence set and then apply Bellman monotonicity.

Lemma 3 (Simultaneous KL row coverage)**.** *Under Assumption 1, the confidence sets in* (8) *satisfy*

$$\mathbb{P}(P(\cdot \mid s, a) \in \mathcal{C}_{s,a} \text{ for every } (s, a)) \ge 1 - \zeta. \tag{19}$$

For one row whose true support has size at most d_0, the method-of-types bound gives $\mathbb{P}(\mathrm{KL}(\widehat{P}_{s,a} \| P_{s,a}) > x) \le (n+1)^{d_0 - 1} e^{-nx}$; see, e.g., [3]. Substituting $x = \kappa_n$ and taking a union bound over the $|\mathcal{S}||\mathcal{A}|$ original rows proves the lemma. Notice that the unknown identities of the successors introduce no additional factor because every empirical type is supported on the true row support.

Fix any realized dataset in the event of Lemma 3, any policy $\pi \in \Pi_{H,\eta}$, and any constraint i. We show that the robust Bellman table dominates the true rounded table.

We use backward induction. At $h = H$, both tables equal $\mathbb{I}\{b_i < 0\}$. Suppose the bound holds at $h+1$. For $a = \pi_h(s, b)$,

$$\begin{aligned} P(\cdot \mid s, a)^{\top} q^{\pi}_{P,i}(\cdot, B_h(s, a, b), h+1) &\le P(\cdot \mid s, a)^{\top} \overline{q}^{\pi}_i(\cdot, B_h(s, a, b), h+1) \\ &\le \sup_{p \in \mathcal{C}_{s,a}} p^{\top} \overline{q}^{\pi}_i(\cdot, B_h(s, a, b), h+1). \end{aligned}$$

The first inequality uses the induction hypothesis, and the second uses $P_{s,a} \in \mathcal{C}_{s,a}$. Induction and averaging over $s_0 \sim \mu$ give $q^{\mathrm{rnd}}_{P,i}(\pi) \le \overline{q}_{i,H,\eta}(\pi)$. If the certificate accepts π, then

$$q_{P,i}(\pi) \le q^{\mathrm{rnd}}_{P,i}(\pi) \le \overline{q}_{i,H,\eta}(\pi) \le \delta_i.$$

The argument is deterministic on the row-coverage event and therefore holds simultaneously for every policy and constraint. This proves the proposition. □

E Proof of Theorem 1

The proof has four steps. We record the finite rounded problem, establish a uniform trajectory-transfer bound, prove safety, and then bound reward loss and sample size.

Let

$$H = H_\alpha = \left\lceil \frac{\log(1/((1-\gamma)\alpha_{\mathrm{tail}}))}{1-\gamma} \right\rceil, \qquad R_H = \sum_{h=0}^{H-1} \gamma^h, \qquad |\mathcal{B}| = \prod_{i=1}^{m} N_i(\alpha_{\mathrm{tail}}, \eta_i),$$

Here $N_i(\alpha_{\mathrm{tail}}, \eta_i)$ is the number of values in budget coordinate i, including failure. Fix a maximizer $\pi_\rho^\star$ of $V_\rho^\star$. Let $J_{P,H}(\pi) := \mathbb{E}_{P,\pi,\mu}[\sum_{h=0}^{H-1} \gamma^h r(S_h, A_h)]$ and set $r_n := \sqrt{H\kappa_n/2}$. Under the theorem's choice $\alpha_{\mathrm{tail}} = \eta_i = \bar{\eta}_\varepsilon$, we have $H = \widetilde{O}((1-\gamma)^{-1})$, $R_H \le (1-\gamma)^{-1}$, and $\gamma^H/(1-\gamma) \le \bar{\eta}_\varepsilon \le \varepsilon/8$.

Step 1: Record the rounded state space.

Coordinate i has at most $\max\{1, \lfloor (d_i - \bar{\eta}_\varepsilon)/\bar{\eta}_\varepsilon \rfloor + 2\}$ values. This finiteness makes the certified planning problem well defined, but $|\mathcal{B}|$ does not enter the confidence event because the same original-row KL set is valid for all time and budget states. The planner may still take time proportional to $|\mathcal{B}|$.

Step 2: Transfer the shared empirical model uniformly.

Let $\widehat{q}_i^{\,\pi}$ be the rounded violation probability under the shared empirical kernel $\widehat{P}$. On the row-coverage event of Lemma 3, the following bound holds for every policy in the class, including the policy selected from the data.

Lemma 4 (Uniform trajectory transfer)**.** *On the event of Lemma 3, every $\pi \in \Pi_{H,\eta}$ and $i \in [m]$ satisfy*

$$\begin{aligned} |\widehat{q}_i^{\,\pi} - q_{P,i}^{\mathrm{rnd}}(\pi)| \le r_n, \qquad 0 \le \overline{q}_{i,H,\eta}(\pi) - q_{P,i}^{\mathrm{rnd}}(\pi) \le 2r_n, \\ |\widehat{J}(\pi) - J_{P,H}(\pi)| \le R_H r_n. \end{aligned} \tag{20}$$

The sample condition in the theorem permits an integer $n = D/(|\mathcal{S}||\mathcal{A}|)$, up to rounding, such that $r_n \le \rho/8$ and $R_H r_n \le \varepsilon/8$. Applying Lemma 4 gives, simultaneously for $\pi_\rho^\star$ and the data-dependent output $\widehat{\pi}$,

$$\left|\widehat{J}(\pi) - J_{P,H}(\pi)\right| \le \frac{\varepsilon}{8}, \qquad 0 \le \overline{q}_{i,H,\eta}(\pi) - q_{P,i}^{\mathrm{rnd}}(\pi) \le \frac{\rho}{4}.$$

Step 3: Prove feasibility and safety.

The rounded comparator satisfies $q_{P,i}^{\mathrm{rnd}}(\pi_\rho^\star) \le \delta_i - \rho$. The uniform transfer bound therefore gives

$$\overline{q}_{i,H,\eta}(\pi_\rho^\star) \le \delta_i - \rho + \frac{\rho}{4} \le \delta_i - \frac{3\rho}{4}, \qquad i \in [m].$$

Thus $\pi_\rho^\star$ is feasible for the tightened empirical problem, so Algorithm 1 does not return UNRESOLVED on the row-coverage event. Its robust feasibility and Lemma 4 give, for each i,

$$q_{P,i}^{\mathrm{rnd}}(\widehat{\pi}) \le \overline{q}_{i,H,\eta}(\widehat{\pi}) \le \delta_i - \frac{3\rho}{4} \le \delta_i.$$

Lemma 2 now gives $q_{P,i}(\widehat{\pi}) \le \delta_i$.

Step 4: Bound reward loss and count samples.

Now compare rewards. Nonnegative rewards give $J_P(\widehat{\pi}) \ge J_{P,H}(\widehat{\pi})$. The reward after time H is at most $\gamma^H/(1-\gamma) \le \varepsilon/8$. Also, $\pi_\rho^\star$ is feasible for the tightened problem, so empirical ξ-optimality applies. Hence

$$\begin{aligned} J_P(\widehat{\pi}) \ge J_{P,H}(\widehat{\pi}) \ge \widehat{J}(\widehat{\pi}) - \frac{\varepsilon}{8} \ge \widehat{J}(\pi_\rho^\star) - \xi - \frac{\varepsilon}{8} \ge J_{P,H}(\pi_\rho^\star) - \xi - \frac{\varepsilon}{4} \\ \ge J_P(\pi_\rho^\star) - \xi - \frac{3\varepsilon}{8} \ge V_\rho^\star - \xi - \varepsilon. \end{aligned}$$

The next-to-last inequality uses the reward-tail bound. The last one uses $J_P(\pi^\star_\rho) = V^\star_\rho$ and $3\varepsilon/8 \le \varepsilon$.

Finally, $r_n \le \min\{\rho/8, \varepsilon/(8R_H)\}$ follows from $n = \widetilde{O}(d_0 H[\rho^{-2} + R_H^2 \varepsilon^{-2}])$. The algorithm draws this batch once at each original state–action row, so $D = |\mathcal{S}||\mathcal{A}|n$. Substituting $H = \widetilde{O}((1-\gamma)^{-1})$ and $R_H \le (1-\gamma)^{-1}$ gives (14). □

E.1 Proof of Lemma 4

Fix a realized dataset in the event of Lemma 3. For a policy π, let $\mathbb{P}^{0:H}_{M,\pi}$ denote the law of the augmented trajectory through time H under transition kernel M. The KL chain rule and (8) give

$$\mathrm{KL}\Big(\mathbb{P}^{0:H}_{\widehat{P},\pi} \,\Big\|\, \mathbb{P}^{0:H}_{P,\pi}\Big) = \sum_{h=0}^{H-1} \mathbb{E}_{\widehat{P},\pi}\Big[\mathrm{KL}(\widehat{P}_{S_h,A_h} \| P_{S_h,A_h})\Big] \le H\kappa_n.$$

Pinsker's inequality bounds the total variation distance between these trajectory laws by r_n. Applying it to the rounded failure indicator gives $|\widehat{q}^{\,\pi}_i - q^{\mathrm{rnd}}_{P,i}(\pi)| \le r_n$. Applying it to the truncated discounted reward, whose range is contained in $[0, R_H]$, gives $|\widehat{J}(\pi) - J_{P,H}(\pi)| \le R_H r_n$.

It remains to control the conservatism of the robust certificate. Rectangularity of the row-wise sets implies that the backward suprema in (7) are attained by a possibly nonstationary augmented-state kernel Q^π_i satisfying $Q^\pi_{i,h}(\cdot \mid s, b) \in \mathcal{C}_{s,\pi_h(s,b)}$ at every triple (s, b, h). Its rounded failure probability equals $\overline{q}_{i,H,\eta}(\pi)$. A second application of the KL chain rule yields

$$\mathrm{KL}\Big(\mathbb{P}^{0:H}_{\widehat{P},\pi} \,\Big\|\, \mathbb{P}^{0:H}_{Q^\pi_i,\pi}\Big) \le H\kappa_n.$$

Because $\widehat{P}_{s,a} \in \mathcal{C}_{s,a}$, robust maximization and Pinsker give $0 \le \overline{q}_{i,H,\eta}(\pi) - \widehat{q}^{\,\pi}_i \le r_n$. Because $P_{s,a} \in \mathcal{C}_{s,a}$, Proposition 1 gives $q^{\mathrm{rnd}}_{P,i}(\pi) \le \overline{q}_{i,H,\eta}(\pi)$. Combining these inequalities with $|\widehat{q}^{\,\pi}_i - q^{\mathrm{rnd}}_{P,i}(\pi)| \le r_n$ proves the safety claim.

All arguments are deterministic after conditioning on the row event and use no policy union bound. Hence they hold simultaneously for every $\pi \in \Pi_{H,\eta}$, including a data-dependent output policy. □

Relation to the original CCMDP optimum. The rounded finite event is a conservative certificate, not a different safety notion. Lemma 2 gives $q_{P,i}(\pi) \le q^{\mathrm{rnd}}_{P,i}(\pi)$, so any policy certified by the rounded event is safe for the original infinite-horizon chance constraint. If $C^\pi_i := \sum_{t\ge 0} \gamma^t c_i(S_t, A_t)$ and $\Delta_i := \alpha_{\mathrm{tail}} + H_\alpha \eta_i$, the same rounding inequalities imply

$$q_{P,i}(\pi) \le q^{\mathrm{rnd}}_{P,i}(\pi) \le \mathbb{P}_{P,\pi,\mu}(C^\pi_i > d_i - \Delta_i)\,.$$

Thus a true 2ρ-interior comparator also belongs to the rounded ρ-interior set whenever its probability mass on $(d_i - \Delta_i, d_i]$ is at most ρ for every constraint. Consequently, if $V^{\star,\mathrm{true}}_{2\rho} := \max\{J_P(\pi) : q_{P,i}(\pi) \le \delta_i - 2\rho,\ i \in [m]\}$ has such a maximizer, Theorem 1 also guarantees reward at least $V^{\star,\mathrm{true}}_{2\rho} - \varepsilon - \xi$. Without this boundary condition, the rounded and original comparators need not agree, even when the transition kernel is deterministic.

F Proof of Theorem 2

The proof has three steps. We prove the reward lower bound, prove the safety lower bound, and add the two bounds. Every transition row in both hard families has at most two successors, so the constructions satisfy Assumption 1 for every fixed $d_0 \ge 2$.

Step 1: Prove the reward lower bound.

Set $H_0 = \lceil (1-\gamma)^{-1} \rceil$, $M = |\mathcal{S}| - 1$, and $B = |\mathcal{A}|$. Build M decision states $x_1, \dots, x_M$ and one absorbing state z. The initial state is uniform over the decision states. The reward is one at each x_j and zero at z. Set $c \equiv 0$, $d = 1$, and $\delta = 1/2$. Then every policy is safe under both the original and rounded events. For $c_0 < 1/2$, the rounded ρ-interior optimum is the usual reward optimum.

Let $\vartheta = (\vartheta_1, \ldots, \vartheta_M) \in [B]^M$ be hidden. At x_j, action ϑ_j is good. The process returns to x_j with probability

$$p_s = 1 - \frac{1}{4H_0} \quad \text{if } a = \vartheta_j, \qquad p_u = p_s - \Delta \quad \text{otherwise}, \qquad \Delta = \frac{2048\varepsilon}{H_0^2};$$

and moves to z otherwise. Rewards and costs are known. Thus only transition samples contain information about ϑ. Choose c_0 small enough that $\Delta \le 1/(4H_0)$. Then $p_u \ge 1 - 1/(2H_0)$, so all probabilities are valid.

Fix any deterministic output policy π; it may depend on history. Consider the history that starts at x_j and stays there for t transitions. Let $a_{j,t}$ be the action of π on this history. Let $S_j(t)$ be the probability of staying at x_j until time t, with $S_j(0) = 1$. Let $W_j = \sum_{t=0}^{H_0-1} \mathbb{I}\{a_{j,t} \neq \vartheta_j\}$. Also let $V^\pi(x_j)$ be the value from x_j.

The all-good policy has value $V^\star = (1 - \gamma p_s)^{-1}$. Since $V^\star = 1 + \gamma p_s V^\star$, each bad action loses $\gamma \Delta V^\star$. Thus the performance-difference identity gives

$$V^\star - V^\pi(x_j) = \gamma \Delta V^\star \sum_{t \ge 0} \gamma^t S_j(t) \mathbb{I}\{a_{j,t} \neq \vartheta_j\}.$$

For $t < H_0$, we have $V^\star \ge H_0/2$, $\gamma^t \ge 1/4$, and $S_j(t) \ge p_u^t \ge 1/2$. Average over the initial state to get

$$V^\star - J_P(\pi) \ge \frac{\Delta H_0}{32M} \sum_{j=1}^{M} W_j.$$

If π is ε-optimal, the choice of Δ gives $\sum_j W_j \le MH_0/64$. Decode ϑ_j as the most common action in $(a_{j,0}, \ldots, a_{j,H_0-1})$. Use fixed tie-breaking. A wrong answer requires $W_j \ge H_0/2$. Thus the decoder is correct on at least $31M/32$ coordinates.

A sample from a decision row is Bernoulli with mean p_s or p_u. The Bernoulli KL bound gives

$$\kappa := \max\{\mathrm{kl}(\mathrm{Bern}(p_s), \mathrm{Bern}(p_u)), \mathrm{kl}(\mathrm{Bern}(p_u), \mathrm{Bern}(p_s))\} \le CH_0\Delta^2 \le C\frac{\varepsilon^2}{H_0^3}.$$

Choose c_0 small enough for Lemma 6. Draw ϑ uniformly from $[B]^M$, and draw a hidden index J uniformly from $[M]$. The learner succeeds with probability at least $2/3$. The decoder above then identifies ϑ_J with probability at least $(2/3)(31/32) > 3/5$.

Now draw the hidden actions for all contexts except J from the prior and simulate their samples internally. Send only queries at J to the one-good-row experiment. This gives the same product-prior law. The expected number of sent queries is at most D/M. Lemma 6 gives $D/M \ge cB/\kappa$. Since $M \ge |\mathcal{S}|/2$ and $H_0 \ge (1-\gamma)^{-1}$,

$$D \ge c\frac{MB}{\kappa} \ge c'\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2},$$

This is the reward lower bound.

Step 2: Prove the safety lower bound.

The next lemma gives the safety lower bound. We prove it in Appendix F.1.

Lemma 5 (Chance-boundary testing lower bound)**.** *There are universal constants $c, c_0 > 0$ such that, for every $\gamma \in [1/2, 1)$, $\rho \in (0, c_0)$, $|\mathcal{S}| \ge 4$, and $|\mathcal{A}| \ge 2$, there exists a zero-reward, one-constraint CCMDP family with two successors per row and a nonempty rounded ρ-interior feasible set for which any generative-model learner that returns an exactly safe policy in Π_{det} with probability at least $2/3$ must use, in the worst case,*

$$D \ge c\,\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)\rho^2} \tag{21}$$

transition samples.

In the family from Lemma 5, all rewards are zero and the rounded ρ-interior set is nonempty. Thus exact safety alone requires $D \geq c|\mathcal{S}||\mathcal{A}|/((1-\gamma)\rho^2)$.

Step 3: Add the two bounds.

Take the union of the two hard families. Its worst-case complexity is at least the larger bound. Since $\max\{x, y\} \geq (x+y)/2$, this gives the sum up to a universal constant. □

F.1 Proof of Lemma 5

The proof has four steps. We build the hard family, decode its safe actions, bound the KL of one sample, and apply a testing lower bound.

Step 1: Build the hard family.

Set $H_0 = \lceil (1-\gamma)^{-1} \rceil$, $M = |\mathcal{S}| - 2$, and $B = |\mathcal{A}|$. Build decision states $x_1, \ldots, x_M$, a failure state f, and a safe absorbing state z. The initial state is uniform over $x_1, \ldots, x_M$. Let $\vartheta \in [B]^M$ be hidden. Action ϑ_j is the safe action at x_j. Set

$$p_s = 1 - \frac{1}{4H_0}, \qquad \Delta = \frac{128\rho}{H_0}, \qquad p_u = p_s - \Delta.$$

At x_j, the process stays at x_j with probability p_s under action ϑ_j and with probability p_u under any other action. It moves to f otherwise. From f, it moves to z in one step. State z is absorbing.

All rewards are zero. The cost is one at f and zero elsewhere. Set

$$d = \frac{\gamma^{H_0} + \gamma^{H_0+1}}{2}, \qquad q_s = 1 - p_s^{H_0}, \qquad \delta = q_s + \rho.$$

Bernoulli's inequality gives $p_s^{H_0} \geq 3/4$. Choose c_0 small enough that $0 < \delta < 1$.

The state f is visited for one step. If it is reached at time τ, the total cost is γ^τ. Since $\gamma^{H_0+1} < d < \gamma^{H_0}$, violation means reaching f by time H_0. The all-safe policy has violation probability $q_s = \delta - \rho$. Thus the ρ-interior set is nonempty. Its value is zero because all rewards are zero.

We verify that this is also the rounded event used in the theorem. Here each trajectory pays at most one nonzero cost. Moreover,

$$\frac{\gamma^{H_0} - \gamma^{H_0+1}}{2} = \frac{\gamma^{H_0}(1-\gamma)}{2} \geq \frac{1-\gamma}{16}, \qquad \alpha_{\mathrm{tail}} + 2\eta \leq \frac{3(1-\gamma)}{128}.$$

Also, $\alpha_{\mathrm{tail}} \leq (1-\gamma)/128$ implies $H_\alpha \geq \log(128/(1-\gamma)^2)/(1-\gamma) > H_0 + 1$. Therefore a payment at time $\tau \leq H_0$ crosses the rounded threshold, while a payment at time $\tau \geq H_0 + 1$ does not: the latter follows from $\gamma^{H_0+1} + \alpha_{\mathrm{tail}} + 2\eta < d$. Hence the original and rounded violation indicators agree trajectory by trajectory throughout this hard family, and the all-safe policy is a rounded ρ-interior comparator.

Step 2: Recover the safe actions.

Fix a deterministic output policy π; it may depend on history. At x_j, consider the history that stays at x_j for the first t transitions. Let $a_{j,t}^\pi$ be the action on this history. This defines an action sequence for $t = 0, \ldots, H_0 - 1$. Let $W_j = \sum_{t=0}^{H_0-1} \mathbb{I}\{a_{j,t}^\pi \neq \vartheta_j\}$. Starting from x_j, the probability of violation is

$$q_j(\pi) = 1 - \prod_{t=0}^{H_0-1} \left(p_s - \Delta \mathbb{I}\{a_{j,t}^\pi \neq \vartheta_j\}\right).$$

Changing one factor from p_s to p_u lowers survival by at least $\Delta p_u^{H_0-1}$. A telescoping expansion gives

$$q_j(\pi) - q_s \geq \Delta p_u^{H_0-1} W_j, \qquad q_P(\pi) - q_s \geq \frac{\Delta p_u^{H_0-1}}{M} \sum_{j=1}^{M} W_j.$$

Choose c_0 so that $128\rho \leq 1/4$. Then $p_u \geq 1 - 1/(2H_0)$, $p_u^{H_0-1} \geq 1/2$, and $H_0 \Delta p_u^{H_0-1} \geq 64\rho$. If π is safe, $q_P(\pi) \leq q_s + \rho$. Hence $\sum_j W_j \leq M H_0/64$.

Decode ϑ_j as the most common action in $(a^\pi_{j,0}, \ldots, a^\pi_{j,H_0-1})$. Use fixed tie-breaking. A wrong answer needs $W_j \geq H_0/2$. The decoder can therefore be wrong on at most $M/32$ states. Every safe policy recovers at least $31M/32$ entries of ϑ.

Step 3: Bound the KL of one sample.

A decision-row sample is Bernoulli with mean p_s or p_u. The mean gap is $\Delta = O(\rho/H_0)$, and the failure probability is of order H_0^{-1}. The Bernoulli KL bound gives

$$\kappa := \max\{\mathrm{kl}(\mathrm{Bern}(p_s), \mathrm{Bern}(p_u)), \mathrm{kl}(\mathrm{Bern}(p_u), \mathrm{Bern}(p_s))\} \leq CH_0\Delta^2 \leq C\frac{\rho^2}{H_0}.$$

Thus one sample has only $O(\rho^2/H_0)$ KL. Choose c_0 small enough for the lemma used next.

Step 4: Reduce to one random state.

We use the following one-context testing bound. Its proof is in Appendix F.2.

Lemma 6 (Adaptive one-good-row identification)**.** *Let $B \geq 2$, let Θ be uniform on $[B]$, and suppose that a query to arm a returns an independent Bernoulli observation with mean $p_s \in (0,1)$ when $a = \Theta$ and mean $p_u \in (0,1)$ otherwise. Define $\kappa = \max\{\mathrm{kl}(\mathrm{Bern}(p_s), \mathrm{Bern}(p_u)), \mathrm{kl}(\mathrm{Bern}(p_u), \mathrm{Bern}(p_s))\}$. For a sufficiently small universal $\kappa_0 > 0$, every adaptive randomized procedure with $0 < \kappa \leq \kappa_0$ that identifies Θ with probability at least $3/5$, averaged over the uniform prior, uses an expected number N of queries satisfying*

$$\mathbb{E}[N] \geq c\,\frac{B}{\kappa} \tag{22}$$

for a universal constant $c > 0$. The conclusion is unchanged in the presence of auxiliary observations whose conditional law given the past is the same for every value of Θ.

Draw ϑ uniformly from $[B]^M$. Also draw a hidden index J uniformly from $[M]$. A safe output is correct on at least a $31/32$ fraction of states. The learner returns such an output with probability at least $2/3$. Hence $\mathbb{P}(\widehat{\vartheta}_J = \vartheta_J) \geq (2/3)(31/32) = 31/48 > 3/5$.

Simulate every state except J internally. Send only queries at J to the one-good-row oracle. This produces the same law as the original hard family, and the learner does not know J. If N_j is the number of samples from state j, then

$$\mathbb{E}[N_J] = \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}[N_j \mid J = j] = \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_{\mathrm{prior}}[N_j] \leq \frac{D}{M}.$$

Lemma 6 gives $D/M \geq cB/\kappa \geq cBH_0/\rho^2$. Use $M \geq |\mathcal{S}|/2$, $B = |\mathcal{A}|$, and $H_0 \geq (1-\gamma)^{-1}$ to get the claimed bound. □

F.2 Proof of Lemma 6

The proof has three steps. We truncate the run, compare it with a null model, and apply Pinsker's inequality.

Step 1: Stop the procedure.

Let $n = \mathbb{E}[N]$ under the uniform prior. Stop after $\lceil 32n \rceil$ queries. By Markov's inequality, this changes the output with probability at most $1/32$. The stopped procedure therefore succeeds with probability at least $3/5 - 1/32$.

Step 2: Define a null model.

Let $\mathbb{Q}_1$ be the law of the hidden arm and the stopped transcript. The transcript includes the algorithm's random bits and output. Define a null law $\mathbb{Q}_0$ in which every arm has mean p_u and Θ is still uniform. Under $\mathbb{Q}_0$, the transcript is independent of Θ. Thus any estimate is correct with probability $1/B$. If N_a is the number of queries to arm a, the KL chain rule gives

$$\mathrm{KL}(\mathbb{Q}_0, \mathbb{Q}_1) = \frac{1}{B}\sum_{a=1}^{B}\mathbb{E}_0[N_a]\,\mathrm{kl}(\mathrm{Bern}(p_u), \mathrm{Bern}(p_s)) \leq \frac{\kappa\lceil 32n \rceil}{B}.$$

Given $\Theta = a$, only samples from arm a differ under the two laws. Auxiliary observations have the same law, so they add zero KL.

Step 3: Apply Pinsker's inequality.

Since $1/B \le 1/2$, the two success probabilities differ by at least $(3/5 - 1/32) - 1/2 = 11/160$. Pinsker's inequality gives $\kappa\lceil 32n\rceil/B \ge c_1$. For small enough κ_0, this implies $n \ge cB/\kappa$. □

G Computational Hardness of Exact Deterministic Planning

Proposition 2 (Hardness of exact deterministic planning)**.** *Even when the transition model is known and deterministic, exact reward maximization over deterministic policies subject to a discounted chance constraint is* NP − hard.

Proof. Reduce from 0–1 knapsack. Let the item values be v_j, the weights be w_j, and the capacity be B. Fix a rational $\gamma \in (0, 1)$. Build a deterministic chain through states $1, \dots, n$. State $n + 1$ is absorbing. At state j, action *choose* gives reward $v_j/(M_v\gamma^{j-1})$ and cost $w_j/(M_w\gamma^{j-1})$. Action *skip* gives zero reward and cost. Both actions move to state $j + 1$. Set

$$M_v = \gamma^{-(n-1)} \max_j v_j, \qquad M_w = \gamma^{-(n-1)} \max_j w_j,$$

Replace a zero maximum by one. Then all one-step rewards and costs lie in $[0, 1]$.

A deterministic policy chooses a vector $x \in \{0, 1\}^n$. The factors γ^{j-1} cancel, so

$$J(x) = \frac{1}{M_v} \sum_{j=1}^{n} v_j x_j, \qquad C(x) = \frac{1}{M_w} \sum_{j=1}^{n} w_j x_j.$$

Set $d = B/M_w$ and take any $\delta \in (0, 1)$. The trajectory is deterministic, so its violation probability is zero or one. The chance constraint holds exactly when $C(x) \le d$, or $\sum_j w_j x_j \le B$.

Thus feasible policies are exactly feasible knapsack solutions. Their discounted reward is proportional to total item value. An exact planner would solve 0–1 knapsack. The construction has polynomial size for fixed rational γ, so the problem is NP-hard. □

H Proof of Theorem 3

The proof has three steps. We define the rounded problem, find a KKT candidate, and test this candidate on fresh samples.

Step 1: Define the rounded problem and its oracles.

Set H as in Theorem 3. A rollout starts from $b_0 = b^0$ and updates the rounded budget until time H. After time H, set $b_t = b_H$ and $\bar{h}_t = H$, so $a_t \sim \pi_\theta(\cdot \mid s_t, \bar{h}_t, b_t)$ with $\bar{h}_t = \min\{t, H\}$. For constraint i, let $q^H_{P,i}(\theta) = \mathbb{P}_{P,\pi_\theta,\mu}(b_{i,H} < 0)$ and define $f_H(\theta) = -\bar{J}_P(\theta)$ and $g^H_i(\theta) = q^H_{P,i}(\theta) + 2\rho - \delta_i$. The rounded problem is

$$\min_{\theta\in\Theta} f_H(\theta) \qquad \text{s.t.} \quad g^H_i(\theta) \le 0, \quad i \in [m].$$

All iterates stay inside $\mathcal{K} \Subset \Theta$. Thus the KKT residual has no boundary term for θ. Assumption 2 gives $\|\nabla_\theta \log \pi_\theta(a \mid s, \bar{h}, b)\|_2 \le G$. The next two lemmas give the stochastic gradients used by the optimizer.

Lemma 7 (Trajectory first-order oracle)**.** *Under Assumption 2, the rollouts defined below give unbiased estimates of $f_H(\theta)$, $\nabla f_H(\theta)$, $g^H_i(\theta)$, and $\nabla g^H_i(\theta)$. Their gradient second moments satisfy*

$$\mathbb{E}\|\widehat{\nabla f_H}(\theta)\|_2^2 \le \frac{2G^2}{(1-\gamma)^2}, \qquad \mathbb{E}\sum\nolimits_{i=1}^{m} \|\widehat{\nabla g^H_i}(\theta)\|_2^2 \le mG^2H^2, \tag{23}$$

and their value samples satisfy

$$|\widehat{f}_H(\theta)| \le 1, \qquad |\widehat{g}^H_i(\theta)| \le 3. \tag{24}$$

One call uses at most $(1-\gamma)^{-1}+H$ transitions in expectation. Its transition-weighted variance is therefore

$$\mathsf{W}_H = O\left(G^2\left[\frac{1}{(1-\gamma)^3}+mH^3\right]\right). \tag{25}$$

The recursive update also needs an unbiased gradient difference.

Lemma 8 (Trajectory gradient differences)**.** *For each $F \in \{g_1^H,\dots,g_m^H\}$, sample τ under π_θ and define*

$$L_{\theta'|\theta}(\tau) = \frac{p_{\theta'}(\tau)}{p_\theta(\tau)} \tag{26}$$

and $\Delta_F(\theta,\theta';\tau) = \widehat{\nabla}F(\theta;\tau) - L_{\theta'|\theta}(\tau)\widehat{\nabla}F(\theta';\tau)$. Then

$$\mathbb{E}_{\tau\sim p_\theta}\Delta_F(\theta,\theta';\tau) = \nabla F(\theta)-\nabla F(\theta'), \qquad \mathbb{E}_{\tau\sim p_\theta}\|\Delta_F(\theta,\theta';\tau)\|_2^2 \le L_{\mathrm{pg}}^2\|\theta-\theta'\|_2^2. \tag{27}$$

For the exact normalized reward with geometric stopping, the same statements hold under the mean-square condition for likelihood-ratio differences in item 3 of Assumption 3. They also hold for a truncated reward rollout, although Theorem 3 does not use that approximation.

Lemmas 7 and 8 give the required refresh and difference estimators, with transition-weighted variance $\mathsf{W}_H = O\big(G^2[(1-\gamma)^{-3}+mH^3]\big)$.

Step 2: Find a KKT candidate.

For $\min_\theta f(\theta)$ subject to $g_i(\theta)\le 0$, define

$$\mathcal{R}(\theta) = \min_{\lambda\in[0,\Lambda]^m}\left\{\left\|\nabla f(\theta)+\sum_i \lambda_i\nabla g_i(\theta)\right\|_2 + \|[g(\theta)]_+\|_1 + \sum_i|\lambda_i g_i(\theta)|\right\}.$$

The next lemma bounds this residual.

Lemma 9 (Finite violation-probability KKT theorem)**.** *Under Assumption 3, the recursive variance-reduced method returns $\widehat{\theta}$ with $\mathbb{E}[\mathcal{R}(\widehat{\theta})]\le\varepsilon$ after*

$$\widetilde{O}\big(K_{\mathrm{opt}}\mathsf{V}\varepsilon^{-3}\big) \tag{28}$$

stochastic first-order oracle calls, where V and K_{opt} are defined in that assumption. If oracle components have different rollout lengths, the expected number of transitions is $\widetilde{O}(K_{\mathrm{opt}}\mathsf{W}\varepsilon^{-3})$.

Apply Lemma 9 to f_H and g_i^H. Assumptions 2 and 3 give all needed conditions. Their constants are in K_{opt} and do not depend on $\varepsilon_{\mathrm{kkt}}$ or ζ. The candidate satisfies $\mathbb{E}[\mathcal{R}_H(\theta_{\mathrm{cand}})]\le\varepsilon_{\mathrm{kkt}}$. Let $L_\alpha = \log(2/((1-\gamma)\alpha_{\mathrm{tail}}))$. Then

$$H = \left\lceil\frac{\log(1/((1-\gamma)\alpha_{\mathrm{tail}}))}{1-\gamma}\right\rceil \le \frac{CL_\alpha}{1-\gamma},$$

for a universal constant C. Insert this bound and W_H into Lemma 9. The optimization cost is

$$\frac{C_{\mathrm{pg}}K_{\mathrm{opt}}G^2m}{\varepsilon_{\mathrm{kkt}}^3(1-\gamma)^3}L_\alpha^3\log^4\left(\frac{8K_{\mathrm{opt}}G^2m}{\varepsilon_{\mathrm{kkt}}\rho\zeta(1-\gamma)^3\alpha_{\mathrm{tail}}}\right)$$

expected transitions. This is the first term of (17).

Step 3: Validate safety.

Now validate the candidate. Condition on the optimization data. Then θ_{cand} is fixed, the validation trajectories are i.i.d., and $\widehat{q}_{i,\mathrm{val}}^H(\theta_{\mathrm{cand}}) = M_{\mathrm{val}}^{-1}\sum_{j=1}^{M_{\mathrm{val}}}\mathbb{I}\{b_{i,H}^{(j)}<0\}$ is a Bernoulli average with mean $q_{P,i}^H(\theta_{\mathrm{cand}})$. Hoeffding's inequality gives

$$\mathbb{P}\left(\left|\widehat{q}_{i,\mathrm{val}}^H(\theta_{\mathrm{cand}})-q_{P,i}^H(\theta_{\mathrm{cand}})\right| > \sqrt{\frac{\log(2m/\zeta)}{2M_{\mathrm{val}}}}\;\middle|\;\theta_{\mathrm{cand}}\right)\le\frac{\zeta}{m}.$$

A union bound over i and the choice of M_{val} give, with probability at least $1-\zeta$,

$$q^H_{P,i}(\theta_{\mathrm{cand}}) \le \widehat{q}^H_{i,\mathrm{val}}(\theta_{\mathrm{cand}}) + \sqrt{\frac{\log(2m/\zeta)}{2M_{\mathrm{val}}}} \le \widehat{q}^H_{i,\mathrm{val}}(\theta_{\mathrm{cand}}) + \rho/2$$

for every i. The next lemma links the rounded event to the original event.

Lemma 10 (Rounded finite tail event certifies infinite-horizon safety)**.** *Fix $H = \lceil \log(1/((1-\gamma)\alpha_{\mathrm{tail}}))/(1-\gamma) \rceil$ and the conservative rounded budget process used by Algorithm 2. For any budget-aware stochastic policy π_θ,*

$$\left\{ \sum\nolimits_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) > d_i \right\} \subseteq \{ b_{i,H} < 0 \}. \tag{29}$$

Consequently,

$$q_{P,i}(\theta) \le q^H_{P,i}(\theta). \tag{30}$$

Suppose validation passes and set $\widehat{\theta} = \theta_{\mathrm{cand}}$. Then $q^H_{P,i}(\widehat{\theta}) \le \delta_i$. Lemma 10 gives $q_{P,i}(\widehat{\theta}) \le \delta_i$. Thus every accepted policy is safe. If the test fails, the algorithm returns UNRESOLVED.

Validation uses $mM_{\mathrm{val}}H$ transitions. The bound on H gives the second term of (17).

It remains to prove acceptance under a margin. Suppose $q^H_{P,i}(\theta_{\mathrm{cand}}) \le \delta_i - \rho$ for every i. On the same event,

$$\widehat{q}^H_{i,\mathrm{val}}(\theta_{\mathrm{cand}}) \le q^H_{P,i}(\theta_{\mathrm{cand}}) + \rho/2 \le \delta_i - \rho/2.$$

Thus $\widehat{q}^H_{i,\mathrm{val}}(\theta_{\mathrm{cand}}) + \rho/2 \le \delta_i$ for all i. The test accepts the candidate. This proves Theorem 3.

H.1 Proof of Lemma 7

Proof. We handle reward and constraints separately.

Step 1: Construct the reward oracle.

Let $\bar{J}_P = (1-\gamma)J_P$. Draw τ with $\mathbb{P}(\tau = t) = (1-\gamma)\gamma^t$. Run the policy through time τ and set

$$\widehat{\bar{J}}(\theta) = r(s_\tau, a_\tau), \qquad \widehat{\nabla \bar{J}}(\theta) = r(s_\tau, a_\tau) \sum\nolimits_{u=0}^{\tau} \nabla_\theta \log \pi_\theta(a_u \mid s_u, \bar{h}_u, b_u).$$

Then $\mathbb{E}[\widehat{\bar{J}}(\theta)] = (1-\gamma)\sum_{t\ge 0}\gamma^t \mathbb{E}[r(s_t,a_t)] = \bar{J}_P(\theta)$. The score identity gives

$$\nabla \bar{J}_P(\theta) = (1-\gamma)\sum\nolimits_{t\ge 0}\gamma^t \nabla_\theta \mathbb{E}[r(s_t,a_t)] = \mathbb{E}[\widehat{\nabla \bar{J}}(\theta)],$$

Rewards and scores are bounded, so the derivative may pass through the sum. Moreover, $\|\widehat{\nabla \bar{J}}(\theta)\|_2^2 \le G^2(\tau+1)^2$. Using $\mathbb{E}[(\tau+1)^2] \le 2(1-\gamma)^{-2}$ and $\mathbb{E}[\tau+1] = (1-\gamma)^{-1}$ proves the reward bounds. For $f_H = -\bar{J}_P$, only the sign changes, and $|\widehat{f}_H| \le 1$.

Step 2: Construct the constraint oracle.

One length-H rollout gives all constraint samples. Define

$$\chi_i = \mathbb{I}\{b_{i,H} < 0\}, \qquad \widehat{g}^H_i(\theta) = \chi_i + 2\rho - \delta_i, \qquad \widehat{\nabla g^H_i}(\theta) = \chi_i \sum\nolimits_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t, t, b_t).$$

Since $\mathbb{E}\chi_i = q^H_{P,i}(\theta)$, the value sample is unbiased. For the gradient, fix a path $\omega = (s_0, a_0, \ldots, s_H)$. Its budgets and $\chi_i(\omega)$ are fixed. Only the policy terms depend on θ:

$$\mathbb{P}_\theta(\omega) = \mu(s_0)\prod\nolimits_{t=0}^{H-1} \pi_\theta(a_t \mid s_t, t, b_t) P(s_{t+1} \mid s_t, a_t),$$
$$\nabla_\theta \mathbb{P}_\theta(\omega) = \mathbb{P}_\theta(\omega) \sum\nolimits_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t, t, b_t).$$

Sum over all paths. This gives $\mathbb{E}[\widehat{\nabla g^H_i}(\theta)] = \nabla q^H_{P,i}(\theta) = \nabla g^H_i(\theta)$.

Finally, $\|\widehat{\nabla g^H_i}(\theta)\|_2^2 \le G^2H^2$ and $|\widehat{g}^H_i(\theta)| \le 3$. A constraint rollout therefore gives $O(mG^2H^3)$ transition-weighted variance. The reward rollout gives $O(G^2(1-\gamma)^{-3})$. Adding them gives W_H. □

H.2 Proof of Lemma 8

Proof. The proof has two steps: unbiasedness and mean-square smoothness.

Step 1: Prove unbiasedness.

The path fixes the rounded budgets. Thus its law depends on θ only through the policy, with $p_\theta(\tau) = \mu(s_0)\prod_{t=0}^{H-1}\pi_\theta(a_t \mid s_t, t, b_t)P(s_{t+1} \mid s_t, a_t)$. Policy positivity makes $L_{\theta'|\theta} = p_{\theta'}/p_\theta$ well defined. Hence

$$\mathbb{E}_{\tau\sim p_\theta}\Big[L_{\theta'|\theta}(\tau)\widehat{\nabla}F(\theta';\tau)\Big] = \mathbb{E}_{\tau\sim p_{\theta'}}[\widehat{\nabla}F(\theta';\tau)] = \nabla F(\theta').$$

Lemma 7 gives the other term. Thus $\Delta_F(\theta,\theta';\tau)$ is unbiased for $\nabla F(\theta) - \nabla F(\theta')$.

Step 2: Establish mean-square smoothness.

Let $R_F(u;\tau) = L_{u|\theta}(\tau)\widehat{\nabla}F(u;\tau)$. The mean-value theorem gives

$$\|\Delta_F(\theta,\theta';\tau)\|_2 \le \sup_{u\in[\theta,\theta']}\|\nabla_u R_F(u;\tau)\|_{\mathrm{op}}\|\theta-\theta'\|_2.$$

For a violation component, $\widehat{\nabla}F(u;\tau) = \chi(\tau)\sum_{t=0}^{H-1}\nabla\log\pi_u(a_t \mid s_t, t, b_t)$. Thus $\nabla_u R_F$ contains likelihood ratios, scores, and score derivatives. Their second moments are bounded on the compact set. Call this bound L_{pg}^2. This proves the second-moment claim.

For the reward oracle, the change of measure gives unbiasedness conditional on the geometric stopping time, which has the same law under both policies. We use item 3 of Assumption 3 for the mean-square bound on this difference estimator. For truncated reward rollouts, the finite-horizon argument above applies. □

H.3 Local KKT conditions

The fixed-penalty problem uses slack variables. The next lemma links its residuals to the KKT residual.

Lemma 11 (Slack residual implies inequality KKT residual)**.** *Consider* $\min_\theta f(\theta)$ *subject to* $g_i(\theta) \le 0$*,* $i \in [m]$*. For a slack variable* $z \in [0, Z]^m$ *and multiplier* $y \in \mathbb{R}^m$*, define*

$$r_\theta = \big\|\nabla f(\theta) + J_g(\theta)^\top y\big\|_2, \quad r_c = \|g(\theta)+z\|_1, \quad r_z = \operatorname{dist}\Big(0, y + N_{\mathbb{R}^m_+}(z)\Big). \tag{31}$$

Here $N_{\mathbb{R}^m_+}(z)$ *is the normal cone. If* $\|J_g(\theta)\|_{\mathrm{op}} \le B_J$ *and* $\|[y]_+\|_1 \le \Lambda$*, then* $\lambda = [y]_+$ *satisfies*

$$\mathcal{R}(\theta) \le r_\theta + B_J r_z + r_c + \Lambda r_c + Z\sqrt{m}\, r_z. \tag{32}$$

Assumption 3 (Finite violation-probability local regularity)**.** *Together with Assumption 2, assume the following local conditions. In particular, item 4 is an additional residual-domination condition. Consider* $\min_\theta f(\theta)$ *subject to* $g_i(\theta) \le 0$*,* $i \in [m]$*, on an open domain* Θ*. Every parameter iterate and every segment between consecutive iterates lies in a compact convex set* $\mathcal{K} \Subset \Theta$*. Set* $x = (\theta, z)$*,* $z \in \mathbb{R}^m_+$*,* $h(x) = \mathbb{I}_{\mathbb{R}^m_+}(z)$*, and* $c(x) = g(\theta) + z$*. Thus only the slack variable is constrained by* h*. Fix* $\beta \ge 1$*, independently of the target accuracy, and let* $\Phi_\beta(x) = f(\theta) + (\beta/2)\|c(x)\|_2^2$ *and* $F_\beta = \Phi_\beta + h$*. All constants below are finite and independent of the target accuracy:*

1. *The initialization* $x_0 = (\theta_0, z_0)$ *has* $g(\theta_0) \le 0$*,* $z_0 \ge 0$*, and* $F_\beta(x_0) - \inf_x F_\beta(x) \le \Delta$*.*
2. *The slack error bound and the following bounds hold:*

$$\begin{aligned} \nu\|c(x)\|_2 &\le \operatorname{dist}\big(-J_c(x)^\top c(x), \partial h(x)\big), \\ \|\nabla f(\theta)\|_2 \le B_f, \qquad \|J_g(\theta)\|_{\mathrm{op}} &\le B_J, \qquad \|z\|_\infty \le Z. \end{aligned} \tag{33}$$

   Here $\partial h(x) = \{0\} \times N_{\mathbb{R}^m_+}(z)$*.*

3. *The functions* f *and* g_i *have Lipschitz gradients, and* $y = \beta c(x)$ *satisfies* $\|[y]_+\|_1 \le \Lambda$ *along the visited iterates. The stochastic samples are unbiased; refresh estimators have bounded variance; and recursive differences are mean-square smooth. For fixed-distribution oracles,*

use common-random-number differences. For trajectory gradients, use the likelihood-ratio difference estimator from Lemma 8. With $C(x,\xi) = \widehat{g}(\theta,\xi) + z$ and $J_C(x,\xi) = [\,\widehat{J}_g(\theta,\xi)\;\; I_m\,]$, there are constants K_c, K_J, K_L, K_V such that

$$\begin{aligned}
&\mathbb{E}\|C(x,\xi)\|_2^2 \le K_c^2, \qquad \mathbb{E}\|J_C(x,\xi)\|_{\mathrm{op}}^2 \le K_J^2,\\
&\mathbb{E}\|C(x,\xi) - C(x',\xi)\|_2^2 \le K_L^2\|x-x'\|_2^2,\\
&\mathbb{E}\|J_C(x,\xi) - J_C(x',\xi)\|_{\mathrm{op}}^2 \le K_L^2\|x-x'\|_2^2,\\
&\mathbb{E}\|\widehat{\nabla} f(\theta;\xi) - \widehat{\nabla} f(\theta';\xi)\|_2^2 \le K_L^2\|x-x'\|_2^2,\\
&\mathbb{E}\|\widehat{\nabla} f(\theta;\xi) - \nabla f(\theta)\|_2^2 \le K_V \mathsf{V},\\
&\mathbb{E}\|C(x,\xi) - c(x)\|_2^2 + \mathbb{E}\|J_C(x,\xi) - J_c(x)\|_{\mathrm{op}}^2 \le K_V\mathsf{V}, \quad \mathsf{V} \ge 1.
\end{aligned} \tag{34}$$

4. *For every visited x, let $y = \beta c(x)$ and define r_θ, r_z, r_c as in Lemma 11. Assume*

$$r_\theta(x) + r_z(x) + r_c(x) \le K_{\mathrm{res}}\ \mathrm{dist}(0, \nabla\Phi_\beta(x) + \partial h(x)) \tag{35}$$

for a finite constant K_{res}.

We write

$$K_{\mathrm{opt}} = \mathrm{poly}\big(L_f, L_g, B_J, \Lambda, Z, \nu^{-1}, \Delta, K_c, K_J, K_L, K_V, L_{\mathrm{pg}}, K_{\mathrm{res}}, \beta\big) \tag{36}$$

Here L_f, L_g are the Lipschitz-gradient constants and L_{pg} is the trajectory-difference constant. These quantities may depend on the policy class, horizon, budget grid, tightening margin, and visited set, but not on ε or the confidence level. We absorb universal polynomial products of them into K_{opt}.

This assumption is used only for the KKT bound. The safety proof uses fresh validation data.

H.4 Proof of Lemma 9

Let $x^+ = (\theta^+, z^+)$ be the computed iterate x_{R+1} selected in Algorithm 2. We first bound the KKT residual by the stationarity residual of $F_\beta = \Phi_\beta + h$. Assumption 3 gives

$$r_\theta(x^+) + r_z(x^+) + r_c(x^+) \le K_{\mathrm{opt}}\ \mathrm{dist}\big(0, \nabla\Phi_\beta(x^+) + \partial h(x^+)\big)\,.$$

Lemma 11 shows that it is enough to make

$$\mathrm{dist}\big(0, \nabla\Phi_\beta(x^+) + \partial h(x^+)\big) \le c\,\varepsilon/K_{\mathrm{opt}}$$

for a small universal constant c. The next lemma gives an oracle for F_β.

Lemma 12 (Penalty-gradient oracle calibration)**.** *Under Assumption 3, for the fixed $\beta \ge 1$, the smooth part*

$$\Phi_\beta(x) = f(\theta) + \frac{\beta}{2}\|c(x)\|_2^2$$

has $K_{\mathrm{sm}}\beta$-Lipschitz gradient, an unbiased refresh-gradient oracle with variance at most $K_\sigma^2\beta^2\mathsf{V}$, and a recursive gradient-difference oracle with mean-square constant $K_{\mathrm{ms}}\beta$. Moreover, $F_\beta = \Phi_\beta + h$ satisfies $F_\beta(x_0) - \inf_x F_\beta(x) \le \Delta$.

Thus, after increasing K_{opt}, we have $L_\beta \le K_{\mathrm{opt}}$, $\sigma_\beta^2 \le K_{\mathrm{opt}}\mathsf{V}$, and $L_{\mathrm{ms},\beta} \le K_{\mathrm{opt}}$. Let $\bar\sigma_\beta = \max\{1, \sigma_\beta\}$. We use the next proximal bound.

Lemma 13 (Self-contained variance-reduced proximal subproblem)**.** *Let $F(x) = \Phi(x) + h(x)$, where Φ is L-smooth, h is closed convex with a tractable proximal map, and $F(x_0) - \inf_x F(x) \le \Delta$. Suppose the refresh estimator is unbiased with variance at most σ^2 and the gradient-difference estimator is unbiased with second moment at most $L_{\mathrm{ms}}^2\|x-x'\|_2^2$. Set $\bar\sigma = \max\{1,\sigma\}$ and run*

$$x_{t+1} = \mathrm{prox}_{\eta h}(x_t - \eta v_t), \tag{37}$$

using fresh independent samples, refreshing v_t every q steps and otherwise using the recursive mini-batch update. If

$$\eta \le (8L)^{-1}, \qquad B \ge \frac{256\bar\sigma^2}{\varepsilon^2}, \qquad b \ge 512q\eta^2 L_{\mathrm{ms}}^2, \qquad T \ge \frac{1024\Delta}{\eta\varepsilon^2}, \tag{38}$$

then, for R uniform on $\{0, \dots, T-1\}$, the computed iterate x_{R+1} satisfies

$$\mathbb{E}[\operatorname{dist}(0, \nabla\Phi(x_{R+1}) + \partial h(x_{R+1}))] \le \varepsilon. \tag{39}$$

With $q = \lceil \bar{\sigma}/\varepsilon \rceil$ and $L_{\mathrm{ms}} = O(L)$, the number of stochastic gradient calls is

$$\widetilde{O}\left(\frac{\bar{\sigma} L \Delta}{\varepsilon^3} + \frac{\bar{\sigma}^2}{\varepsilon^2} \right). \tag{40}$$

Apply the lemma to F_β with $\varepsilon_{\mathrm{int}} = c\varepsilon/K_{\mathrm{opt}}$. Choose

$$\eta \asymp L_\beta^{-1}, \qquad q \asymp \bar{\sigma}_\beta/\varepsilon_{\mathrm{int}}, \qquad B \asymp \bar{\sigma}_\beta^2/\varepsilon_{\mathrm{int}}^2, \qquad b \asymp q\eta^2 L_{\mathrm{ms},\beta}^2, \qquad T \asymp \Delta L_\beta/\varepsilon_{\mathrm{int}}^2.$$

The refresh steps use $\lceil T/q \rceil B$ samples. The other steps use Tb samples. Thus

$$\left\lceil \frac{T}{q} \right\rceil B + Tb \le \widetilde{O}\left(\frac{\bar{\sigma}_\beta L_\beta \Delta}{\varepsilon_{\mathrm{int}}^3} + \frac{\bar{\sigma}_\beta^2}{\varepsilon_{\mathrm{int}}^2} \right) \le \widetilde{O}\big(K_{\mathrm{opt}} \mathsf{V} \varepsilon^{-3} \big).$$

The last inequality uses $\varepsilon \in (0, 1)$ and the definition of K_{opt}. The stationarity residual is at most $c\varepsilon/K_{\mathrm{opt}}$ in expectation. Set $\widehat{\theta} = \theta^+$. The residual bound and Lemma 11 give $\mathbb{E}[\mathcal{R}(\widehat{\theta})] \le \varepsilon$. If each sample is weighted by its rollout length, V becomes W. □

H.5 Proof of Lemma 13

The proof has two steps. We first establish descent for the inexact proximal update and then control the recursive gradient error.

Step 1: Establish descent.

Define the exact gradient mapping G_t, the computed step d_t, and the gradient error e_t by

$$G_t = \frac{1}{\eta}\left(x_t - \operatorname{prox}_{\eta h}(x_t - \eta \nabla\Phi(x_t)) \right), \qquad d_t = \frac{1}{\eta}(x_t - x_{t+1}), \qquad e_t = v_t - \nabla\Phi(x_t).$$

The proximal map is nonexpansive, so $\|d_t - G_t\|_2 \le \|e_t\|_2$ and $\|G_t\|_2^2 \le 2\|d_t\|_2^2 + 2\|e_t\|_2^2$. Proximal optimality gives $d_t - v_t \in \partial h(x_{t+1})$, while smoothness gives

$$h(x_{t+1}) - h(x_t) \le \eta \langle v_t, d_t \rangle - \frac{\eta}{2}\|d_t\|_2^2,$$
$$\Phi(x_{t+1}) - \Phi(x_t) \le -\eta \langle \nabla\Phi(x_t), d_t \rangle + \frac{L\eta^2}{2}\|d_t\|_2^2.$$

Add the last two bounds and use $v_t = \nabla\Phi(x_t) + e_t$:

$$F(x_{t+1}) \le F(x_t) - \eta\left(\frac{1}{2} - \frac{L\eta}{2} \right)\|d_t\|_2^2 + \eta \langle e_t, d_t \rangle.$$

Use $\eta \le (8L)^{-1}$ and $\langle e_t, d_t \rangle \le \frac{1}{8}\|d_t\|_2^2 + 2\|e_t\|_2^2$. Then

$$\mathbb{E}F(x_{t+1}) \le \mathbb{E}F(x_t) - \frac{\eta}{4}\mathbb{E}\|d_t\|_2^2 + 2\eta\mathbb{E}\|e_t\|_2^2.$$

Sum over t and use $\|G_t\|^2 \le 2\|d_t\|^2 + 2\|e_t\|^2$:

$$\sum_{t=0}^{T-1} \mathbb{E}\|G_t\|_2^2 \le \frac{8\Delta}{\eta} + 18 \sum_{t=0}^{T-1} \mathbb{E}\|e_t\|_2^2.$$

Step 2: Control the recursive gradient error.

We next bound e_t. We use the following SARAH/STORM recursion [51, 52].

Lemma 14 (Variance-reduced gradient recursion)**.** *Let $G(\theta;\xi)$ be an unbiased refresh estimator and let $D(\theta,\theta';\xi)$ be a stochastic difference estimator satisfying*

$$\begin{aligned} \mathbb{E}[G(\theta;\xi)] &= \nabla\Phi(\theta), & \mathbb{E}\|G(\theta;\xi)-\nabla\Phi(\theta)\|_2^2 &\le \sigma^2, \\ \mathbb{E}[D(\theta,\theta';\xi)] &= \nabla\Phi(\theta)-\nabla\Phi(\theta'), & \mathbb{E}\|D(\theta,\theta';\xi)\|_2^2 &\le L_{\mathrm{ms}}^2\|\theta-\theta'\|_2^2. \end{aligned} \tag{41}$$

At the start s of an epoch and, respectively, at a recursive step $t>s$, set

$$v_s = \frac{1}{B}\sum\nolimits_{j=1}^{B} G(\theta_s;\xi_j), \qquad v_t = v_{t-1} + \frac{1}{b}\sum\nolimits_{j=1}^{b} D(\theta_t,\theta_{t-1};\xi_{t,j}), \tag{42}$$

using fresh samples in every mini-batch. Then

$$\mathbb{E}\|v_t-\nabla\Phi(\theta_t)\|_2^2 \le \frac{\sigma^2}{B} + \frac{L_{\mathrm{ms}}^2}{b}\sum\nolimits_{r=s+1}^{t}\mathbb{E}\|\theta_r-\theta_{r-1}\|_2^2. \tag{43}$$

Apply Lemma 14 to each epoch. If an epoch starts at s, let $\ell_s = \min\{q, T-s\}$. Then

$$\sum\nolimits_{t=s}^{s+\ell_s-1}\mathbb{E}\|e_t\|_2^2 \le \frac{\ell_s\sigma^2}{B} + \frac{qL_{\mathrm{ms}}^2}{b}\sum\nolimits_{r=s+1}^{s+\ell_s-1}\mathbb{E}\|x_r-x_{r-1}\|_2^2.$$

Sum over all epochs. Also use $\|x_r-x_{r-1}\|_2^2 = \eta^2\|d_{r-1}\|_2^2 \le 2\eta^2\|G_{r-1}\|_2^2 + 2\eta^2\|e_{r-1}\|_2^2$. We get

$$\sum\nolimits_{t=0}^{T-1}\mathbb{E}\|e_t\|_2^2 \le \frac{T\sigma^2}{B} + \frac{2q\eta^2L_{\mathrm{ms}}^2}{b}\sum\nolimits_{t=0}^{T-1}\mathbb{E}\|G_t\|_2^2 + \frac{2q\eta^2L_{\mathrm{ms}}^2}{b}\sum\nolimits_{t=0}^{T-1}\mathbb{E}\|e_t\|_2^2.$$

Since $b \ge 512q\eta^2L_{\mathrm{ms}}^2$, move the last term to the left:

$$\sum\nolimits_{t=0}^{T-1}\mathbb{E}\|e_t\|_2^2 \le \frac{2T\sigma^2}{B} + \frac{1}{128}\sum\nolimits_{t=0}^{T-1}\mathbb{E}\|G_t\|_2^2.$$

Insert this into the descent bound:

$$\sum\nolimits_{t=0}^{T-1}\mathbb{E}\|G_t\|_2^2 \le \frac{10\Delta}{\eta} + \frac{42T\sigma^2}{B}.$$

The choices of B and T give

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|G_t\|_2^2 \le \left(\frac{10}{1024}+\frac{42}{256}\right)\varepsilon^2, \qquad \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|e_t\|_2^2 \le \frac{2}{256}\varepsilon^2 + \frac{1}{128T}\sum_{t=0}^{T-1}\mathbb{E}\|G_t\|_2^2.$$

Also, $\|d_t-G_t\|_2 \le \|e_t\|_2$, so $T^{-1}\sum_t\mathbb{E}\|d_t\|_2^2 \le 2T^{-1}\sum_t\mathbb{E}(\|G_t\|_2^2+\|e_t\|_2^2)$. Proximal optimality gives

$$\operatorname{dist}(0,\nabla\Phi(x_{t+1})+\partial h(x_{t+1})) \le (1+L\eta)\|d_t\|_2 + \|e_t\|_2.$$

Square this bound and average over uniform R. Use $L\eta \le 1/8$ and the two bounds above. The result is at most ε^2. Jensen's inequality gives the stated expected residual. Finally, refresh steps use $\lceil T/q\rceil B$ samples and recursive steps use Tb samples. Substitute $q = \lceil\bar\sigma/\varepsilon\rceil$ and $L_{\mathrm{ms}} = O(L)$ to get $\widetilde{O}(\bar\sigma L\Delta\varepsilon^{-3} + \bar\sigma^2\varepsilon^{-2})$. □

H.6 Proof of Lemma 14

At a refresh step, independence gives $\mathbb{E}\|v_s-\nabla\Phi(\theta_s)\|_2^2 \le \sigma^2/B$. For $t>s$, define

$$\Delta_t = \frac{1}{b}\sum\nolimits_{j=1}^{b} D(\theta_t,\theta_{t-1};\xi_{t,j}) - \left(\nabla\Phi(\theta_t)-\nabla\Phi(\theta_{t-1})\right).$$

Thus $\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\|\Delta_t\|_2^2 \mid \mathcal{F}_{t-1}] \le (L_{\mathrm{ms}}^2/b)\|\theta_t-\theta_{t-1}\|_2^2$. Also, $v_t-\nabla\Phi(\theta_t) = v_{t-1}-\nabla\Phi(\theta_{t-1})+\Delta_t$. The cross term has conditional mean zero. Apply this identity from $s+1$ to t to get the stated bound. □

H.7 Proof of Lemma 11

Proof. Since $z_i \geq 0$, we have $[g_i(\theta)]_+ \leq |g_i(\theta) + z_i|$, and hence $\|[g(\theta)]_+\|_1 \leq r_c$. The normal cone is

$$N_{\mathbb{R}_+}(z_i) = \begin{cases} (-\infty, 0], & z_i = 0, \\ \{0\}, & z_i > 0. \end{cases}$$

Let $e_i = \operatorname{dist}(0, y_i + N_{\mathbb{R}_+}(z_i))$. If $z_i = 0$, then $|[y_i]_-| \leq e_i$ and $y_i z_i = 0$. If $z_i > 0$, then $|y_i| \leq e_i$. Hence

$$\|y - [y]_+\|_2 \leq \|e\|_2 \leq r_z, \qquad \sum\nolimits_i [y_i]_+ z_i \leq \sum\nolimits_i e_i z_i \leq Z\sqrt{m}\, r_z.$$

Set $\lambda = [y]_+$. Then

$$\left\|\nabla f(\theta) + J_g(\theta)^\top \lambda\right\|_2 \leq r_\theta + \|J_g(\theta)\|_{\mathrm{op}} \|y - \lambda\|_2 \leq r_\theta + B_J r_z.$$

Let $e_i^c = g_i(\theta) + z_i$. Since $g_i(\theta) = e_i^c - z_i$,

$$\sum\nolimits_i |\lambda_i g_i(\theta)| \leq \sum\nolimits_i \lambda_i |e_i^c| + \sum\nolimits_i \lambda_i z_i \leq \Lambda r_c + Z\sqrt{m}\, r_z.$$

These are the three terms in $\mathcal{R}(\theta)$. Adding them proves the lemma. □

H.8 Proof of Lemma 12

Proof. We prove the refresh bound, the recursive bound, and the smoothness bound in this order.

Step 1: Construct the refresh estimator.

Let $C(x, \xi) = \widehat{g}(\theta, \xi) + z$, whose sampled Jacobian is $J_C(x, \xi) = [\, \widehat{J}_g(\theta, \xi) \;\; I_m \,]$. Use independent samples for the two factors and define

$$\widehat{\nabla}\Phi_\beta(x; \xi_0, \xi_1, \xi_2) = \widehat{\nabla} f(\theta; \xi_0) + \beta J_C(x, \xi_1)^\top C(x, \xi_2)$$

Independence gives $\mathbb{E}[J_C(x, \xi_1)^\top C(x, \xi_2)] = \mathbb{E}[J_C(x, \xi_1)]^\top \mathbb{E}[C(x, \xi_2)] = J_c(x)^\top c(x)$. Thus this is an unbiased estimator of $\nabla\Phi_\beta(x) = \nabla f(\theta) + \beta J_c(x)^\top c(x)$.

For the variance, use $J_C^\top C - J_c^\top c = J_C^\top (C - c) + (J_C - J_c)^\top c$. The moment bounds then give $\mathbb{E}\|J_C^\top C - J_c^\top c\|_2^2 \leq 2K_J^2 K_V \mathsf{V} + 2K_c^2 K_V \mathsf{V}$. Add the variance of $\widehat{\nabla} f$ and use $\beta \geq 1$. This gives $K_\sigma^2 \beta^2 \mathsf{V}$.

Step 2: Construct the recursive difference estimator.

Now construct the recursive difference. Draw ξ_1 and ξ_2 independently at x. Let $L_{\xi_j}(x' \mid x)$ be the likelihood ratio from x to x'. It equals one for fixed-distribution samples. Define

$$\begin{aligned} D_\beta(x, x'; \xi_0, \xi_1, \xi_2) = {} & \widehat{\nabla} f(x; \xi_0) - L_{\xi_0}(x' \mid x) \widehat{\nabla} f(x'; \xi_0) \\ & + \beta \Big[J_C(x, \xi_1)^\top C(x, \xi_2) - L_{\xi_1}(x' \mid x) L_{\xi_2}(x' \mid x) J_C(x', \xi_1)^\top C(x', \xi_2) \Big]. \end{aligned}$$

The two ratios change measure for the two samples. Independence gives $\mathbb{E}[L_{\xi_1}(x' \mid x) L_{\xi_2}(x' \mid x) J_C(x', \xi_1)^\top C(x', \xi_2)] = J_c(x')^\top c(x')$. The same identity holds for the objective-gradient term, so $\mathbb{E}[D_\beta(x, x'; \xi_0, \xi_1, \xi_2)] = \nabla\Phi_\beta(x) - \nabla\Phi_\beta(x')$.

For the second moment, define $\mathcal{P}(u; \xi_1, \xi_2) = L_{\xi_1}(u \mid x) L_{\xi_2}(u \mid x) J_C(u, \xi_1)^\top C(u, \xi_2)$. The mean-value theorem gives $\|\mathcal{P}(x; \xi_1, \xi_2) - \mathcal{P}(x'; \xi_1, \xi_2)\|_2 \leq \sup_{u \in [x, x']} \|\nabla_u \mathcal{P}(u; \xi_1, \xi_2)\|_{\mathrm{op}} \|x - x'\|_2$. Each derivative acts on one likelihood ratio, J_C, or C. Assumption 3 and Lemma 8 bound these terms. Hence

$$\mathbb{E}\|\mathcal{P}(x; \xi_1, \xi_2) - \mathcal{P}(x'; \xi_1, \xi_2)\|_2^2 \leq K^2 \|x - x'\|_2^2$$

for a finite constant K. The same argument applies to the objective-gradient term. The factor β gives the bound $K_{\mathrm{ms}}^2 \beta^2 \|x - x'\|_2^2$.

Step 3: Verify smoothness and the initial gap.

The same product expansion gives the smoothness bound from $J_C(x)^\top C(x) - J_C(x')^\top C(x') = J_C(x)^\top (C(x) - C(x')) + (J_C(x) - J_C(x'))^\top C(x')$.

Finally, item 1 of Assumption 3 gives $F_\beta(x_0) - \inf_x F_\beta(x) \leq \Delta$. This proves the lemma. □

H.9 Proof of Lemma 10

Proof. If $b_i^0 < 0$, the result is immediate. Now suppose $b_{i,H} \geq 0$. The budget update gives $\sum_{t=0}^{H-1} w_{i,t}(s_t, a_t) \leq b_i^0 = \lfloor (d_i - \alpha_{\text{tail}})/\eta_i \rfloor$. Also, $\gamma^t c_i(s_t, a_t) \leq \eta_i w_{i,t}(s_t, a_t)$. Thus

$$\sum_{t=0}^{H-1} \gamma^t c_i(s_t, a_t) \leq \eta_i \left\lfloor \frac{d_i - \alpha_{\text{tail}}}{\eta_i} \right\rfloor \leq d_i - \alpha_{\text{tail}}.$$

The tail after time H is at most $\gamma^H/(1-\gamma) \leq \alpha_{\text{tail}}$. Hence $b_{i,H} \geq 0$ implies $\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t) \leq d_i$. The contrapositive gives (29). Taking probabilities gives (30). □

I Experiment Details

I.1 Synthetic CCMDP diagnostics

Synthetic instance and protocol. The synthetic diagnostic is a finite-state absorbing discounted MDP with random absorption time. It has eight decision states $s_0, \ldots, s_7$, one bad state b, and one absorbing terminal state s_{term}, with initial state s_0 and discount factor $\gamma = 0.95$. At each decision state there are two actions, safe $(a = 0)$ and risky $(a = 1)$. Both actions have self-loop probability 0.06, which makes the process genuinely infinite-horizon rather than fixed-horizon. The safe action has bad-transition probability 0.002 and moves forward with probability 0.938. The risky action has bad-transition probability p_i^r and moves forward with probability $0.94 - p_i^r$; the values of p_i^r and risky rewards are listed in Table A.1. The bad state moves to s_{term}, and s_{term} is absorbing.

The chance constraint is

$$\mathbb{P}\Big(\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) > 0.50\Big) \leq 0.13.$$

Only the bad state has nonzero cost, $c(b, a) = 1$, while all decision and terminal states have zero cost. The safe action reward is 0.45 at every decision state; the bad and terminal states have zero reward. We enumerate all $2^8 = 256$ stationary deterministic policies and use the same class for both learned selectors and the oracle. We use the sample sizes

$$\begin{aligned}\mathcal{N}_{\text{syn}} = \{&500, 1000, 2000, 5000, 10000, 20000, 50000, 62500, 125000, 250000, 500000\} \\ &\cup \{10^6, 2 \times 10^6, 4 \times 10^6, 8 \times 10^6\}.\end{aligned}$$

For each $N \in \mathcal{N}_{\text{syn}}$, we draw N next-state samples from each of the 16 decision-state/action rows, giving a total budget $D = 16N$. The bad and terminal transition rows are known. Within each repetition, both methods share one empirical kernel $\widehat{P}$, reused across time, policies, safety evaluation, and reward evaluation. We use 150 independent empirical-model draws at each budget, with random seed 20260503, and evaluate selected policies exactly under the true kernel. The datasets at different budgets are drawn separately and are not nested prefixes of one sample stream.

Robust Bellman selector and comparator. The CCMDP selector applies Algorithm 1 with $\rho = 0.0035$, $\zeta = 0.05$, $d_0 = 3$, and $\alpha_{\text{tail}} = \eta = 0.000390625$, giving $H = 217$ and $b^0 = 1279$. Each sampled row uses the reverse-KL confidence set in (8), over all ten successor states; the known bad and terminal rows have singleton confidence sets. Since the bad state is followed by the absorbing terminal state, every admissible trajectory pays the nonzero cost at most once. The rounded violation event is exactly a bad-state visit at time $t \leq 13$, so the robust recursion in (7) reduces to a 14-layer reachability calculation. We enumerate all 256 policies and maximize the H-step empirical Bellman return in (9) subject to

$$\overline{q}_{H,\eta}(\pi) \leq \delta - \frac{3\rho}{4} = 0.127375.$$

An empty certified set returns UNRESOLVED.

The Markov-CMDP comparator maximizes the same empirical return subject to $U_N^\pi(s_0) \leq \delta d = 0.065$, where the robust expected-cost value uses the same row-wise KL confidence sets and $\zeta = 0.05$:

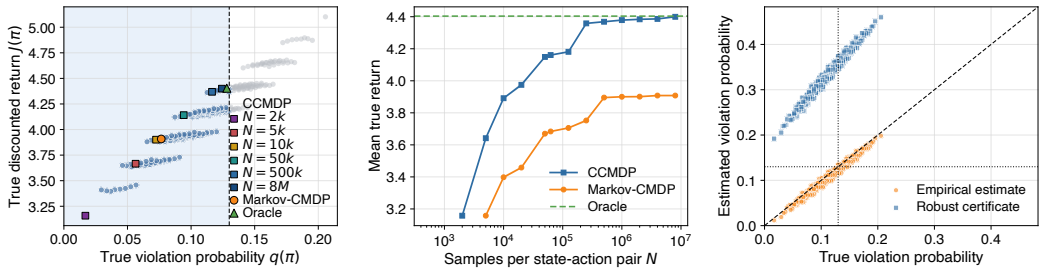
$$U_N^\pi(s) = c(s, \pi(s)) + \gamma \sup_{p \in \mathcal{C}_{s,\pi(s)}} p^\top U_N^\pi.$$

Its margin above $\mathbb{E}_{\widehat{P}}[C^\pi]$, with $C^\pi = \sum_{t \geq 0} \gamma^t c(s_t, a_t)$, is $O(\sqrt{\log N/N})$ for fixed γ. The oracle maximizes true discounted return subject to true chance feasibility within the same 256-policy class.

Table A.1: State-dependent risky-action parameters in the synthetic CCMDP. The safe action uses reward 0.45 and bad-transition probability 0.002 at every decision state.

Decision state s_i	s_0	s_1	s_2	s_3	s_4	s_5	s_6	s_7
Risky reward $r(s_i, 1)$	0.72	0.73	0.74	0.77	0.79	0.81	0.84	0.87
Risky bad probability p_i^r	0.014	0.018	0.021	0.024	0.028	0.032	0.036	0.040
Risky next-state probability $0.94 - p_i^r$	0.926	0.922	0.919	0.916	0.912	0.908	0.904	0.900

Synthetic results. Figure A.1 shows how the robust certificate's conservatism decreases with additional samples. CCMDP returns no policy at $N = 500$ or 1000, returns a policy in 30 of 150 trials at $N = 2000$, and returns a policy in all 150 trials at each larger sample size. Every returned CCMDP policy is truly feasible. At $N = 8 \times 10^6$, its mean true return is 4.39954, compared with 3.90793 for Markov-CMDP and the same-class oracle value 4.40376, a relative gap of approximately 0.096%. Markov-CMDP returns no policy for $N \leq 2000$, returns a policy in 111 of 150 trials at $N = 5000$, and returns a policy in all 150 trials at each larger sample size. Every returned Markov-CMDP policy satisfies the true expected-cost and chance constraints. Reported return means and standard errors are conditional on a policy being returned.



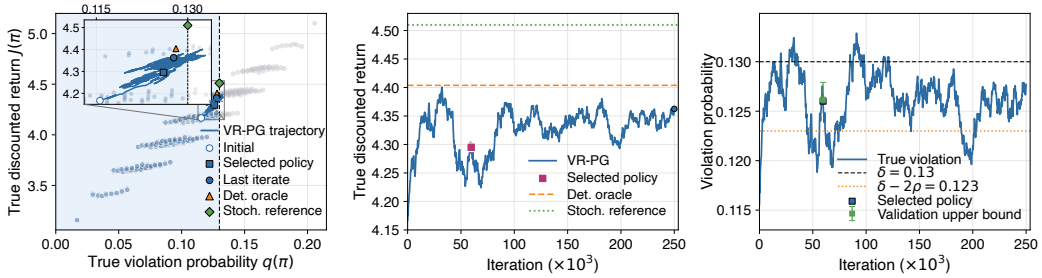
(a) Policy landscape (b) Return vs. samples (c) Certificate calibration

Figure A.1: Synthetic CCMDP diagnostics with shared transition samples. **(a)** True return–violation landscape of all 256 stationary deterministic policies, with CCMDP selections from repetition 0 at the indicated sample sizes, the Markov-CMDP selection at $N = 8 \times 10^6$, and the same-class oracle. **(b)** Mean true return among returned policies from 150 independent empirical-model draws per sample size; shaded bands show ± 1.96 empirical standard errors. Each learned transition row uses N samples, giving total budget $D = 16N$. **(c)** Empirical violation estimates and robust certificates versus exact violation probabilities for all 256 policies at $N = 1000$, repetition 0. The diagonal indicates equality, and dotted lines mark $\delta = 0.13$.

Model-free experiment. We apply the variance-reduced policy-gradient method (VR-PG) with the recursive proximal updates of Algorithm 2 to the same synthetic instance, using a stationary Bernoulli policy with one logit per decision state. The initial risky-action probability is 0.5 at every state, and the initial nonnegative slack is zero. We optimize the normalized negative return with the quadratic slack penalty, using $\rho = 0.0035$, fixed $\beta = 80$, numerical step size 0.01, and 250,000 updates. Thus the tightened training threshold is $\delta - 2\rho = 0.123$. The gradient estimator is refreshed every 20 updates using 2,048 independent trajectory triples; intervening updates use 128 triples with likelihood-ratio corrections. Each triple supplies independent reward-gradient, risk-gradient, and risk-value samples. The proximal step projects the slack onto the nonnegative half-line. Training uses 168 million trajectories. True-model evaluations are used to display the trajectory of the learned policies.

Figure A.2 shows the VR-PG optimization trajectory for training seed 11. The return rises from 4.16658 initially to 4.36245 at the last iterate, whose true violation probability is 0.12769. The best enumerated deterministic feasible return is 4.40376, and the multi-start numerical stochastic reference has return 4.50970 at violation probability 0.13.

After training, a separate random draw selects uniformly from the 250,000 post-update policies; the selected candidate is at iteration 59,611. Its true return is 4.29506 and its true violation probability is 0.12601. An independent batch of 602,267 trajectories gives a violation estimate of 0.126168; adding the validation margin $\rho/2 = 0.00175$ yields 0.127918. Since this upper bound is below $\delta = 0.13$, the candidate is accepted. The selected candidate and the last iterate are marked separately in the figure.



(a) Policy trajectory (b) Return trajectory (c) Risk and validation

Figure A.2: VR-PG on the synthetic CCMDP with $\rho = 0.0035$ and $\beta = 80$. **(a)** Optimization trajectory in the true return–violation plane, with the initial policy, selected policy, and last iterate marked against the deterministic policy landscape. The inset enlarges the trajectory and includes the deterministic oracle and numerical stochastic reference. **(b)** True discounted return over training. The magenta square marks the selected policy; horizontal lines show the deterministic oracle and stochastic reference returns. **(c)** True violation probability over training, with chance threshold $\delta = 0.13$ and tightened training threshold $\delta - 2\rho = 0.123$. The green marker and upward error bar show the independent validation estimate and its upper bound for the selected policy; the upper bound lies below δ.

I.2 IEEE 14-bus storage-control setting

The storage-control experiment in Section 6 is a finite-state CCMDP built from an IEEE 14-bus DC power-flow simulator. The state is $s_t = (e_t, \tau_t, \ell_t)$, where $e_t \in \{0, 0.1, \ldots, 1.0\}$ is storage state of charge, τ_t is one of four time blocks, and ℓ_t is one of four load regimes with multipliers $(0.82, 1.00, 1.18, 1.42)$. Thus $|\mathcal{S}| = 176$. The storage device is placed at bus 14, and the action set is $\{-12, -6, 0, 6, 12\}$ MW, where negative actions charge and positive actions discharge. Storage capacity is 48 MWh, with charging and discharging efficiencies both equal to 0.95; infeasible commands are clipped to the SOC range and rounded to the nearest SOC grid point.

The time block advances deterministically, while the load regime follows a time-dependent Markov chain. With rows and columns ordered as low, normal, high, and extreme, the transition row is $P_{\text{load}}(\cdot \mid \tau, \ell) = 0.75 P_{\text{base}}(\ell, \cdot) + 0.25 P_{\text{time}}(\tau, \cdot)$, where

$$P_{\text{base}} = \begin{bmatrix} 0.68 & 0.25 & 0.06 & 0.01 \\ 0.15 & 0.62 & 0.19 & 0.04 \\ 0.04 & 0.20 & 0.58 & 0.18 \\ 0.02 & 0.08 & 0.30 & 0.60 \end{bmatrix}, \qquad P_{\text{time}} = \begin{bmatrix} 0.55 & 0.35 & 0.09 & 0.01 \\ 0.18 & 0.55 & 0.22 & 0.05 \\ 0.05 & 0.25 & 0.45 & 0.25 \\ 0.10 & 0.35 & 0.40 & 0.15 \end{bmatrix}.$$

The initial distribution puts SOC at 0.5, time at night, and load-regime probabilities $(0.15, 0.70, 0.13, 0.02)$.

Rewards measure normalized operating benefit. The time-block prices are $(0.45, 0.75, 1.35, 1.05)$, and the reward is an affine normalization of negative operating cost to $[0, 1]$. The safety cost is normalized line-overload severity from the DC power-flow solution: if $\rho_{\max}(s, a)$ is the maximum line-loading ratio after applying action a, then $c(s, a) = \min\{1, [\rho_{\max}(s, a) - 1]_+ / 0.5\}$. The chance constraint used in the reported IEEE 14-bus experiment is

$$\mathbb{P}_{P,\pi,\mu}\Big(\sum\nolimits_{t=0}^{\infty} \gamma^t c(s_t, a_t) > 0.30\Big) \le 0.15, \qquad \gamma = 0.85. \tag{44}$$

For the finite Bellman violation table, we use tail allowance $\alpha_{\text{tail}} = 0.005$ and budget-grid width $\eta = 0.0015$. By (4), these choices give

$$H = 48, \qquad b^0 = \left\lfloor \frac{d - \alpha_{\text{tail}}}{\eta} \right\rfloor = 196.$$

The structured policy class contains 144 threshold storage policies together with the always-idle policy, for a total of 145 policies. A threshold policy discharges when the load regime and SOC exceed prescribed thresholds, charges during low-price periods when the load and SOC are below prescribed thresholds, and otherwise idles. CCMDP and the chance oracle use these deterministic candidates; Markov-CMDP also allows initial randomization over them.

For the sample-budget experiment, we use

$$\mathcal{N}_{\text{IEEE}} = \{25, 50, 100, 200, 500, 1000, 1500, 2000, 2500, 3000\} \cup \{3500, 3750, 4000, 5000, 10000, 20000\}.$$

For Markov-CMDP, we extend this grid by $\{50{,}000, 100{,}000, 200{,}000, 500{,}000, 1{,}000{,}000\}$, with 60 independent empirical models at each added sample size. For each $N \in \mathcal{N}_{\text{IEEE}}$, we obtain N independent successor samples per original state-action row, with 60 independent empirical models per sample size. Each model uses $D = 880N$ transitions. Its empirical kernel is shared across all time stages, rounded-budget nodes, candidate policies, and reward evaluations.

The main comparison is between the robust-backward CCMDP selector and a Markov-CMDP expected-cost baseline over initial mixtures of these candidates. We write $C_\pi = \sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)$. By Markov's inequality, $\mathbb{P}_{P,\pi,\mu}(C_\pi > d) \le \frac{\mathbb{E}_{P,\pi,\mu}[C_\pi]}{d}$, so the condition

$$\mathbb{E}_{P,\pi,\mu}[C_\pi] \le \delta d = 0.045 \tag{45}$$

is sufficient for the original chance constraint. For each deterministic candidate π_i, let $u_{i,N}$ be its robust infinite-horizon expected-cost upper bound under the same row-wise KL confidence sets ($\zeta = 0.05$). Markov-CMDP solves the linear program

$$\max_{w \ge 0,\, \sum_i w_i = 1} \sum_i w_i \widehat{J}_H(\pi_i) \quad \text{subject to} \quad \sum_i w_i u_{i,N} \le 0.045,$$

where $\widehat{J}_H$ is the $H = 48$-step empirical return. An optimal solution needs at most two components; one is drawn per trajectory and followed throughout. An empty certified set returns UNRESOLVED. The true-model deterministic and mixture Markov references maximize true return subject to (45) over the original candidates and their initial mixtures, respectively. The chance-constrained oracle is the highest-return policy in the original deterministic class satisfying (44).

True-model discounted returns and expected discounted safety costs are computed from the known finite transition kernel for evaluation only. Violation probabilities displayed in Figure 1 are estimated under the true transition kernel using 50,000 Monte Carlo trajectories of length 100. These true-model quantities are not available to the learner and are used only for evaluation and construction of the oracle references. Mixture statistics are the corresponding weighted averages of component-policy statistics.

Robust backward certificate. With $d_0 = 4$ and $\zeta = 0.05$, each row uses the reverse-KL confidence radius

$$\kappa_N = \frac{3\log(N+1) + \log(880/0.05)}{N}.$$

The robust backups maximize over the full 176-state simplex. We set $\rho = 0.0035$ and solve the finite policy search exactly, maximizing the empirical H-step return subject to

$$\overline{q}_{H,\eta}(\pi) \le \delta - \frac{3\rho}{4} = 0.147375.$$

An empty certified set returns UNRESOLVED. All trials with $N \le 1000$ are unresolved; 59 of 60 trials return a policy at $N = 1500$, and all 60 return a policy at each larger sample size. The returned policies through $N = 3500$ coincide with the deterministic Markov-CMDP reference. At $N = 3750$, eight of the 60 trials select the oracle, increasing to 52 at $N = 4000$ and all 60 at $N = 5000$. Markov-CMDP returns no policy for $N \le 10{,}000$ and returns a policy in all 60 trials at each tested $N \ge 20{,}000$. Reported means and standard errors condition on a policy being returned. All 599 returned CCMDP policies satisfy the exact rounded-risk constraint under the true model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction distinguish the conditional model-based upper bound, the deterministic-policy lower bound, and the local model-free guarantee with accept-or-unresolved validation. The corresponding interface and local-regularity assumptions are stated explicitly.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 7 discusses limitations and future directions, including computation for finite certified planning, the strict-feasibility margin, and the local nature of the model-free stochastic guarantee. The deterministic certificate does not assume density, anti-concentration, or atom separation, and the experiments include both a synthetic diagnostic instance and an IEEE 14-bus storage-control benchmark.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The appendix proves the implication from the stated certified-planning interface, gives the complete deterministic-policy lower-bound construction, and proves the model-free result under the explicit local residual-domination assumption.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The storage-control benchmark and both synthetic experiments include their implementations and numerical results. The appendix specifies the shared-sample deterministic protocol and the recursive variance-reduced model-free updates, numerical parameters, candidate selection, and independent validation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The experimental materials include implementations and saved results for the IEEE 14-bus benchmark, the deterministic synthetic comparison, and the model-free synthetic experiment, with reproduction instructions. Public access to these materials has not been specified.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The appendix specifies the storage benchmark and synthetic MDP, together with the model-free step size, penalty coefficient, refresh schedule, batch sizes, iteration count, and independent validation procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: For the repeated deterministic synthetic experiment, the appendix reports mean-return bands of ± 1.96 empirical standard errors across independent model draws. For the IEEE 14-bus experiment, Panel C of Figure 1 reports the mean optimality gap together with ± 1.96 empirical standard errors across 60 independent empirical-model trials. The model-free synthetic figure shows a single training run and reports a one-sided validation upper bound for the uniformly selected candidate.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: The available implementation is CPU-only, but exact hardware, peak memory, and wall-clock totals were not recorded for every reported run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is theoretical and algorithmic, uses no human-subject data, and does not involve restricted datasets, deployed systems, or prohibited applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper discusses practical relevance and societal impacts by illustrating a real-world storage-control case.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release models, datasets, or other assets that pose high-risk misuse concerns.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A].

Justification: The current draft does not use external datasets, pretrained models, or software assets as part of the research contribution.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: The paper does not introduce or release new datasets, code packages, models, or other research assets in the current draft.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects, so IRB approval or an equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLM tools assisted with experimental setup, code implementation and debugging, figure preparation, and manuscript editing. The paper is grounded in theoretical analysis of CCMDPs, and LLMs are not components of the proposed learning or certification methods.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.