CANTELLI CONSTRAINED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *Canary*, a risk-averse method designed to optimize Value-at-Risk (VaR) constrained reinforcement learning (RL) problems. We employ Cantelli's inequality to obtain a tractable, conservative and smooth bound on the VaR constraint based on the first two moments of the cost return. This yields a constraint estimator that remains stable with tight violation thresholds in dense cost regimes. Extending the trust-region framework of the Constrained Policy Optimization (CPO) method, we further provide worst-case bounds for both policy improvement and constraint violation during the training process. Empirically during training, Canary is the only method that reliably satisfies the VaR constraint in every environment tested.

1 INTRODUCTION

Reinforcement Learning (RL) has demonstrated remarkable utility in optimizing complex decision-making simulations such as autonomous driving (Kiran et al., 2021), robotics (Schulman et al., 2018; 2017b;a) and finance (Hambly et al., 2023; Briola et al., 2023). The standard RL objective typically seeks to maximize the expected return. While effective in simulations, this formulation is often insufficient for real-life scenarios, especially when bounding the probability of a costly catastrophic failure is a prerequisite for deployment. In these settings, designing a risk-aware reward function via a soft constraint can be impractical. The cost of tuning may be too high, while choosing a setting a priori might lead to intractable learning or unsafe policies. This motivates the use of hard constraints on tail risks, which are often a more natural formulation to facilitate safety.

A natural hard constraint to achieve safety is a *Value-at-Risk (VaR)* constraint, where the objective is to *first* satisfy a bound on the probability that the cumulative cost exceeds a limit, and only then to maximize reward (Stellato et al., 2017; Tagawa, 2017). Here, cost is a distinct penalty for *unsafe* states that might be distinct from the reward signal of the RL agent's objective. While the term VaR is chiefly used in finance, the way we employ it is general and extends naturally to probabilistic safety constraints in control and robotics. For example, in finance we may wish to avoid a margin call, while in robotics we would want to avoid actions that destroy physical components.

Existing methods struggle precisely in the practically important regime of strict (small) violation thresholds. Indicator based methods (Geibel & Wysotzki, 2005; Chow et al., 2017) suffer from sample inefficiency through cost sparsity (Andrychowicz et al., 2018) and instability as the violation probability shrinks (Section 5). In finance, *Conditional Value-at-Risk (CVaR)* is often favored as a tractable conservative bound on VaR (Rockafellar & Uryasev, 2000; 2002). However, empirical CVaR estimators are also unstable in tight VaR violation threshold regimes (Section 5).

We introduce *Canary*,[1] a novel approach with four primary contributions. We formulate a smooth and tractable safe approximation (Nemirovski, 2012) of VaR constraints in RL using the Cantelli inequality, the tightest possible distribution-free bound relying only on the first two moments of the cost return distribution (Bertsimas & Popescu, 2005) (Section 4). We provide a worst-case analysis of the constraint violation during training, extending the guarantees of the trust-region based CPO method (Achiam et al., 2017) to the VaR setting (Section 4). We characterize the relative variance of VaR constraint estimators, showing indicator and CVaR estimators collapse in low-violation regimes while Cantelli remains stable (Section 5). Finally, we empirically demonstrate that Canary settles at the deepest safety level, achieves the least converged violation-probability jitter, and maintains safety most reliably on all three benchmark environments (Section 6).

[1] Like a canary in a coal mine, the Cantelli constraint senses the approaching probabilistic tail risk boundary and alerts the policy.

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

2 Related Work

Trust region methods. Safe RL typically constrains expected cost returns with Lagrangian methods (Tessler et al., 2018), which offer no bounds on violation during training. Constrained Policy Optimization (CPO) (Achiam et al., 2017) instead uses trust-region theory (Schulman et al., 2017a) to bound worst-case training-time violations. Canary inherits these guarantees and extends them from expected cost to tail risk.

Value-at-Risk (VaR) optimization. Constraining the probability of violation casts the problem as a chance-constrained MDP, where dynamic programming or search bounds the violation probability under a known model (Ono et al., 2015; Santana et al., 2016); these methods do not extend to the model-free, continuous-control setting we address. Within RL, a Bernoulli indicator that triggers upon violation is often used to estimate the VaR bound directly, introducing an extremely sparse, sample-inefficient cost signal (Geibel & Wysotzki, 2005; Andrychowicz et al., 2018), and the Lagrangian policy gradients of Chow et al. (2017) lack training-time safety guarantees. Canary resolves both: state augmentation (Xu & Mannor, 2011; Sootla et al., 2022) applied to the second cost moment gives dense Markovian cost assignment, and the Cantelli formulation provides a smooth safe approximation compatible with worst-case trust-region bounds.

Conditional Value-at-Risk (CVaR). Most RL work on tail risk uses CVaR as a constraint (Zhang et al., 2025; Ying et al., 2022) or objective (Lim & Malik, 2022), typically estimated by distributional RL from a worst fraction of trajectories (Zhang et al., 2025; Lim & Malik, 2022), which yields fragile scaling of the constraint gradients (Section 5). Canary instead projects the boundary through the first two moments of the full batch, giving stable gradients that constrain the tail. WCSAC (Yang et al., 2021) also reasons about tail risk through moments, but assumes a Gaussian cost distribution within an off-policy framework; we exclude it from our on-policy comparison and note that the Gaussian tail underestimates risk in our experiments (Remark 6.1).

3 Preliminaries

3.1 Constrained Markov Decision Process

We model the agent's interaction with the environment as a *Constrained Markov Decision Process (CMDP)* (Altman, 1999), a tuple $(\mathcal{S}, \mathcal{A}, T, \mathcal{R}, \mathcal{C}, d_0)$ of state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward kernel $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, cost kernel $\mathcal{C} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ and initial state distribution $d_0 \in \Delta(\mathcal{S})$, where $\Delta(\mathcal{X})$ denotes a set of probability distributions over $\mathcal{X}$. The reward measures task performance and the cost measures safety. A *policy* $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ generates *trajectories* $\tau = (s_t, a_t, r_t, c_t)_{t \in \mathcal{N}}$ with *actions* $a_t \sim \pi(\cdot|s_t)$, *states* $s_{t+1} \sim T(\cdot|s_t, a_t)$, *rewards* $r_t \sim \mathcal{R}(\cdot|s_t, a_t)$ and *costs* $c_t \sim \mathcal{C}(\cdot|s_t, a_t)$, where in theoretical analysis $\mathcal{N} = \mathbb{N}_0$. The standard reinforcement learning objective is the *expected reward return*

$$J(\pi) = \mathbb{E}_{\tau \sim \pi}\Big[\sum_{t \in \mathcal{N}} \gamma^t r_t\Big], \tag{1}$$

where $\gamma \in [0, 1)$ is the *reward discount factor*. In a safety-critical setting, we are also concerned with the discounted *cost return* $C(\tau) = \sum_{t \in \mathcal{N}} \gamma_c^t c_t$, with *cost discount factor* $\gamma_c \in [0, 1)$, and its expectation

$$\mu(\pi) = \mathbb{E}_{\tau \sim \pi}[C(\tau)] = \mathbb{E}_{\tau \sim \pi}\Big[\sum_{t \in \mathcal{N}} \gamma_c^t c_t\Big]. \tag{2}$$

The standard CMDP objective maximizes $J(\pi)$ subject to a *cost limit* $\ell \in \mathbb{R}$ on $\mu(\pi)$, the Optimization Problem (*OP*):

$$\pi^* = \arg\max_{\pi} \; J(\pi) \quad \text{s.t.} \quad \mu(\pi) \leq \ell. \tag{3}$$

3.2 Constrained Policy Optimization

Constrained Policy Optimization (CPO) solves CMDPs iteratively, updating π_k to π_{k+1} within a trust region (Schulman et al., 2017a) and bounding the worst-case constraint violation and performance

degradation of each step (Achiam et al., 2017). Since $J(\pi)$ and $\mu(\pi)$ cannot be evaluated for a candidate $\pi \neq \pi_k$ from trajectories of π_k alone, CPO optimizes the *trust-region surrogates*

$$L(\pi) = J(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{\substack{s\sim d_{\pi_k} \\ a\sim\pi}}\left[A_{\pi_k}(s,a)\right] \tag{4}$$

$$L_\mu(\pi) = \mu(\pi_k) + \frac{1}{1-\gamma_c}\mathbb{E}_{\substack{s\sim d^c_{\pi_k} \\ a\sim\pi}}\left[A^C_{\pi_k}(s,a)\right], \tag{5}$$

where $d_\pi(s) = (1-\gamma)\sum_{t\in\mathcal{N}}\gamma^t\mathbb{P}(s_t = s \mid \pi)$ and $d^c_\pi(s) = (1-\gamma_c)\sum_{t\in\mathcal{N}}\gamma_c^t\mathbb{P}(s_t = s \mid \pi)$ are the reward and cost *discounted state visitation frequencies* of π, and A_{π_k} and $A^C_{\pi_k}$ are the reward and cost *advantage functions* (Schulman et al., 2018) of π_k. Crucially, $L(\pi_k) = J(\pi_k)$ and $\nabla L(\pi_k) = \nabla J(\pi_k)$. Therefore, although not necessarilly linear functions, these surrogates are often referred to as valid first-order approximations (Schulman et al., 2017a; Achiam et al., 2017). At each iteration k, CPO solves the local optimization problem

$$\pi_{k+1} = \arg\max_\pi \; L(\pi) \quad \text{s.t.} \quad L_\mu(\pi) \leq \ell, \quad \bar{D}_{KL}(\pi, \pi_k) \leq \delta, \tag{6}$$

where $\bar{D}_{KL}(\pi, \pi_k) = \mathbb{E}_{s\sim d^c_{\pi_k}}[D_{KL}(\pi(a \mid s), \pi_k(a \mid s))]$ is the expected KL divergence between π and π_k, and the radius hyperparameter $\delta > 0$ defines the *trust region* (Schulman et al., 2017a).

3.3 Value-at-Risk Objective

The *Value-at-Risk (VaR)* of the cost return at a *violation threshold* $\varepsilon \in (0,1)$ is

$$\mathrm{VaR}_\varepsilon(\pi) = \inf\left\{\, \eta \in \mathbb{R} : \mathbb{P}\big(C(\tau) > \eta\big) \leq \varepsilon \,\right\}. \tag{7}$$

We write $\eta^* := \mathrm{VaR}_\varepsilon(\pi)$; by right-continuity of $\eta \mapsto \mathbb{P}(C(\tau) > \eta)$, $\mathbb{P}(C(\tau) > \eta^*) \leq \varepsilon$. In contrast to the standard constraint on expected cost return in the CMDP objective (3), we are interested in bounding the tail risk: given the cost limit ℓ (Section 3.1), we impose the probabilistic constraint $\mathbb{P}(C(\tau) > \ell) \leq \varepsilon$, which is equivalent to $\mathrm{VaR}_\varepsilon(\pi) \leq \ell$. We refer to this as the *VaR constraint*. The resulting objective becomes

$$\boxed{\pi^* = \arg\max_\pi \; J(\pi) \quad \text{s.t.} \quad \mathbb{P}(C(\tau) > \ell) \leq \varepsilon.} \tag{8}$$

Note that we have not specified the policy space being maximized over. Typically, this will be given by some parametric class of policies, reducing the objective to a constrained parameter OP.

3.4 Indicator Based Methods

In principle, OP 8 could be solved as a special case of the standard CMDP problem in OP 3 by introducing an indicator-based cost return $I(\tau)$ (Geibel & Wysotzki, 2005; Chow et al., 2017):

$$I(\tau) = \mathbf{1}(C(\tau) > \ell) = \mathbf{1}\Big(\sum_{t\in\mathcal{N}}\gamma_c^t c_t > \ell\Big), \tag{9}$$

whose expectation is the violation probability; with $I(\tau)$ as the cost return, the constraint in OP 3 becomes:

$$\mu_I(\pi) = \mathbb{E}_{\tau\sim\pi}[I(\tau)] = \mathbb{P}(C(\tau) > \ell) \leq \varepsilon. \tag{10}$$

We will refer to this as the *indicator constraint*.

3.5 Conditional Value-at-Risk Methods

An alternative for the VaR objective is to constrain the Conditional Value-at-Risk (CVaR). This is a *safe approximation* of the VaR constraint (Nemirovski, 2012): any policy satisfying the CVaR constraint at target violation threshold ε also satisfies the VaR constraint. CVaR can be tractably formulated using the Rockafellar-Uryasev dual representation (Rockafellar & Uryasev, 2000; 2002):

$$\mathrm{CVaR}_\varepsilon(\pi) = \min_\eta \left(\eta + \frac{1}{\varepsilon}\mathbb{E}_\tau\left[(C(\tau) - \eta)^+\right]\right) \tag{11}$$

where $(x)^+ = \max(0, x)$; the $(1-\varepsilon)$-quantile η^* of Equation (7) is a minimizer. The CVaR constrained OP is then:

$$\pi^* = \arg\max_\pi \; J(\pi) \quad \text{s.t.} \quad \mathrm{CVaR}_\varepsilon(\pi) \leq \ell. \tag{12}$$

3.6 CANTELLI'S INEQUALITY

Cantelli's inequality (also known as the one-sided Chebyshev inequality) provides a tighter version of the Chebyshev inequality for one-sided tail bounds. Given the random cost return $C(\tau)$ with finite mean $\mu(\pi)$ (Equation (2)) and variance $\sigma^2(\pi) = \text{Var}_{\tau\sim\pi}[C(\tau)]$, Cantelli's inequality states that for any $\lambda > 0$:

$$\mathbb{P}(C(\tau) - \mu(\pi) \geq \lambda) \leq \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \lambda^2}. \tag{13}$$

This holds for any distribution with finite first and second moments, which makes it more conservative than a bound that exploits further knowledge of the distribution of $C(\tau)$. Crucially, however, no distribution-free bound using only the first two moments can be tighter on $\mathbb{R}$ (Bertsimas & Popescu, 2005).

4 METHOD

In this section, we present Canary, a risk-averse, stable (Section 5) algorithm for enforcing Value-at-Risk constraints in an online RL framework using Cantelli's inequality. Further proofs and derivations can be found in Appendix A.

4.1 CANTELLI APPROXIMATED VALUE-AT-RISK

We first employ Cantelli's inequality (13) to construct a smooth safe approximation of the original VaR constraint in OP 8. Substituting $\lambda = \ell - \mu(\pi)$ into (13) can be shown to be equivalent to the *Cantelli constraint* (refer to Appendix A.1):

$$B(\pi) := \left(\frac{1}{\varepsilon} - 1\right)\sigma^2(\pi) - [\ell - \mu(\pi)]^2 \leq 0. \tag{14}$$

Cantelli's inequality requires $\lambda > 0$, so $\mu(\pi) < \ell$. However, the boundary case $\mu(\pi) = \ell$ also satisfies the VaR constraint under $B(\pi) \leq 0$ (Appendix A.1). The non-strict *mean constraint* $\mu(\pi) \leq \ell$ can therefore be imposed, and OP 8 is transformed into the *Cantelli-VaR* form:

$$\pi^* = \arg\max_{\pi} \; J(\pi) \quad \text{s.t.} \quad B(\pi) \leq 0, \quad \mu(\pi) \leq \ell. \tag{15}$$

We call (ℓ, ε) a *feasible setting* if some policy π satisfies OP 15. The Cantelli constraint is equivalent to a worst-case VaR bound over all cost distributions sharing the first two moments of $C(\tau)$ (El Ghaoui et al., 2003), which equals the worst-case CVaR over that family (Zymler et al., 2013). Since CVaR bounds VaR (Rockafellar & Uryasev, 2000; 2002), the Cantelli-feasible set of policies for OP 15 is a subset of the CVaR-feasible set of OP 12, which is in turn a subset of the VaR-feasible set of OP 8. As such, any policy satisfying the Cantelli-VaR constraint is guaranteed to satisfy the original VaR constraint, but not vice versa. The Cantelli-VaR constraint is thus also a safe approximation of the VaR constraint. This conservatism means a policy that appears feasible only through estimation error can still satisfy the VaR constraint, at the price of a smaller feasible set of policies; that price is quantified in Section 6.3.

4.2 AUGMENTED COST FORMULATION

We wish to facilitate the derivation of worst-case training time constraint violations for the Cantelli-VaR constraint (14) in lieu of those provided by CPO. This requires the cost variance $\sigma^2(\pi) = \mathbb{E}[C(\tau)^2] - \mathbb{E}[C(\tau)]^2$ and expectation $\mu(\pi)$ to be estimable using trust-region surrogates. In turn, these make use of advantage functions, which have a form given by a discounted sum of local per-step terms. While $\mu(\pi)$ is already of this form, the second moment $\mathbb{E}[C(\tau)^2]$ is not; we therefore rewrite it as a discounted sum of local per-step terms.

Let y_t be the discounted accumulated cost up to time t with

$$y_{t+1} = y_t + \gamma_c^t c_t, \quad y_0 = 0. \tag{16}$$

The square of the cost return can be decomposed as follows (Appendix A.2):

$$C(\tau)^2 = \sum_{t \in \mathcal{N}} \gamma_c^t (\gamma_c^t c_t^2 + 2 y_t c_t). \tag{17}$$

Therefore, given $\beta = \frac{1}{\varepsilon} - 1$, the cost second moment can be written as the expected discounted return of the *augmented cost* $\tilde{c}_t$:

$$\tilde{c}_t = \gamma_c^t c_t^2 + 2 y_t c_t \tag{18}$$

$$\tilde{\mu}(\pi) = \mathbb{E}_{\tau \sim \pi}\Big[\sum_{t=0}^{\infty} \gamma_c^t \tilde{c}_t \Big] = \mathbb{E}_{\tau \sim \pi}[C(\tau)^2]. \tag{19}$$

However, this augmented cost is not Markovian; that is, the state-action tuple (s_t, a_t) alone is insufficient to determine the distribution of $\tilde{c}_t$. Akin to (Xu & Mannor, 2011; Sootla et al., 2022), we augment the state space with y_t and γ_c^t, to give the augmented state x_t:

$$x_t = (s_t, y_t, \gamma_c^t). \tag{20}$$

Letting $\rho(\pi) = \frac{1}{\varepsilon}\mu(\pi)^2 + \ell^2$ be the *policy-dependent bound*, the Cantelli constraint can thus be written as (Appendix A.3)

$$B(\pi) = \beta \tilde{\mu}(\pi) + 2\ell\mu(\pi) - \rho(\pi) \leq 0. \tag{21}$$

4.3 Update Step

We will update the policy with an iterative framework akin to the one used in CPO. Again, we cannot evaluate $J(\pi)$ or $B(\pi)$ over any candidate policy π, because we only sample trajectories from an initial policy $\pi_k \neq \pi$. We circumvent this by creating a policy trust region (Schulman et al., 2017a) and introducing trust-region surrogates $L_{\tilde{\mu}}(\pi)$ and $L_{\rho}(\pi)$ at the current policy π_k for augmented cost return $\tilde{\mu}(\pi)$, and policy-dependent bound $\rho(\pi)$ respectively. Defining $Z = \mathbb{E}_{\substack{x \sim d^c_{\pi_k} \\ a \sim \pi}}[A^C_{\pi_k}(x, a)]$, then as per Appendix A.4

$$L_{\tilde{\mu}}(\pi) = \tilde{\mu}(\pi_k) + \frac{1}{1-\gamma_c} \mathbb{E}_{\substack{x \sim d^c_{\pi_k} \\ a \sim \pi}} \left[A^{\tilde{C}}_{\pi_k}(x, a) \right], \tag{22}$$

$$L_{\rho}(\pi) = \rho(\pi_k) + \frac{1}{\varepsilon} \left(\frac{2\mu(\pi_k)}{(1-\gamma_c)} Z + \frac{1}{(1-\gamma_c)^2} Z^2 \right). \tag{23}$$

Taking the predefined trust-region surrogates for the reward return in Equation (4) and the cost return in Equation (5), a trust-region surrogate for $B(\pi)$ at π_k is given

$$L_B(\pi) = \beta L_{\tilde{\mu}}(\pi) + 2\ell L_{\mu}(\pi) - L_{\rho}(\pi). \tag{24}$$

This yields the final practical OP, including the trust-region surrogate of the mean constraint:

$$\boxed{\pi_{k+1} = \arg\max_{\pi} \; L(\pi) \quad \text{s.t.} \quad L_B(\pi) \leq 0, \quad L_{\mu}(\pi) \leq \ell, \quad \bar{D}_{KL}(\pi, \pi_k) \leq \delta.} \tag{25}$$

4.4 Worst-Case Violation Bounds

Extending the theoretical analysis of CPO, we derive a bound on the worst-case constraint violation introduced by the trust-region surrogates.

Theorem 4.1 (Worst-Case Cantelli Violation)**.** *Let solution policy π_{k+1} satisfy OP 25. Then π_{k+1} also satisfies:*

$$B(\pi_{k+1}) \leq K\Big(\beta \alpha^{\tilde{C}}_{\pi_{k+1}} + 2|\ell| \alpha^{C}_{\pi_{k+1}} + \frac{2\alpha^{C}_{\pi_{k+1}}}{\varepsilon} M \Big) \quad \textit{and} \quad \mu(\pi_{k+1}) \leq \ell + K \alpha^{C}_{\pi_{k+1}}, \tag{26}$$

where $\alpha^{\tilde{C}}_{\pi} = \max_x |\mathbb{E}_{a \sim \pi}[A^{\tilde{C}}_{\pi_k}(x, a)]|$ and $\alpha^{C}_{\pi} = \max_x |\mathbb{E}_{a \sim \pi}[A^{C}_{\pi_k}(x, a)]|$ represent the maximum expected augmented and standard cost advantages respectively over augmented states x (20), $K = \frac{\sqrt{2\delta}\gamma_c}{(1-\gamma_c)^2}$, and $M = |\mu(\pi_k)| + \frac{\alpha^{C}_{\pi_{k+1}}}{1-\gamma_c}$. See Appendix A.5 for a proof.

Theorem 4.1 gives constraint-violation magnitude bounds for the Cantelli and mean constraints in OP 15 during training under almost surely bounded costs. However, as in CPO, the $(1-\gamma_c)^{-2}$ dependence means the bound loosens as $\gamma_c \to 1$, unless the trust region δ shrinks correspondingly.

Since the trust-region constrained $L(\pi)$ objective in OP 25 is identical to CPO, it also inherits the worst case reward performance bound under almost surely bounded rewards (Achiam et al., 2017). Moreover, as in the CPO paper, our bounds omit accounting for any error due to the practical necessity of estimating the advantage functions or expected cost returns from policy roll-outs.

4.5 Recovery Mechanism

OP 25 may admit no solution given the trust region size. Extending CPO's treatment of infeasible updates, we replace an infeasible update with a recovery step. Writing $\Pi_\delta = \{\pi : \bar{D}_{KL}(\pi, \pi_k) \le \delta\}$ and $\Pi_\delta^\ell = \Pi_\delta \cap \{\pi : L_\mu(\pi) \le \ell\}$, define the reachable minima

$$m_\mu = \min_{\pi \in \Pi_\delta} L_\mu(\pi), \qquad m_B = \min_{\pi \in \Pi_\delta^\ell} L_B(\pi), \tag{27}$$

both of which admit closed forms (Appendix C). The update selection has three tiers: the nominal update (a), Cantelli restoration (b), and mean restoration (c),

$$\pi_{k+1} = \begin{cases} \text{(a) solution of OP 25} & m_\mu \le \ell,\ m_B \le 0, \\ \text{(b) } \arg\min_{\pi \in \Pi_\delta^\ell} L_B(\pi) & m_\mu \le \ell,\ m_B > 0, \\ \text{(c) } \arg\min_{\pi \in \Pi_\delta} L_\mu(\pi) & m_\mu > \ell, \end{cases} \tag{28}$$

equivalently evaluated as: if $m_\mu > \ell$ apply (c); else if $m_B > 0$ apply (b); else apply (a). Prioritizing safety, Tier c precedes the others because the mean constraint is a precondition for the Cantelli constraint.

No solution policy of any update tier guarantees satisfaction of the VaR constraint. Under almost surely bounded costs, Tier a alone admits the worst-case violation magnitude bounds of Theorem 4.1, while policies of Tiers a–b do not provably violate the VaR constraint if their cost standard deviation is large enough (Corollary 4.2):

Corollary 4.2 (Conditional VaR Violation)**.** *Let $V = \{\pi : \mu(\pi) > \ell + \sigma(\pi)/\sqrt{\beta}\}$ be the region where the VaR constraint is provably violated and π_{k+1} be produced by the selection rule 28, with K and $\alpha^C_{\pi_{k+1}}$ as in Theorem 4.1. For Tiers a–b, $\mu(\pi_{k+1}) \le \ell + K\alpha^C_{\pi_{k+1}}$, and consequently $\pi_{k+1} \notin V$ whenever $\sigma(\pi_{k+1}) \ge \sqrt{\beta}\, K\alpha^C_{\pi_{k+1}}$. For Tier c, $\mu(\pi_{k+1}) \le \mu(\pi_k) + K\alpha^C_{\pi_{k+1}}$. See Appendix A.6 for a proof.*

4.6 Practical Algorithm

For a policy parameterized by θ, given the Cantelli-VaR constrained OP 25 might be a non-linear problem. Therefore, in analogy to CPO, it is made computationally tractable using Taylor expansions. The reward objective and cost constraint are approximated to first order, while the KL divergence constraint is approximated to second order. To this end, let

$$c_B = L_B(\theta_k) = \beta\tilde{\mu}(\theta_k) + 2\ell\mu(\theta_k) - \rho(\theta_k) \tag{29}$$

$$b_B = \nabla L_B(\theta_k) = \beta\nabla L_{\tilde{\mu}}(\theta_k) + \left(2\ell - \frac{2\mu(\theta_k)}{\varepsilon}\right)\nabla L_\mu(\theta_k), \tag{30}$$

$$g = \nabla L(\theta_k), \quad c_\mu = \mu(\theta_k) - \ell, \quad b_\mu = \nabla L_\mu(\theta_k). \tag{31}$$

Then the problem is mapped to the following quadratically constrained linear program

$$\theta_{k+1} = \arg\max_\theta\ g^\top\Delta \quad \text{s.t.} \quad c_B + b_B^\top\Delta \le 0, \quad c_\mu + b_\mu^\top\Delta \le 0, \quad \tfrac{1}{2}\Delta^\top H\Delta \le \delta, \tag{32}$$

where $\Delta = \theta - \theta_k$ and H is the Hessian of the KL Divergence. The solution to OP 32 is closed-form (Appendix C). However, to ensure satisfaction of OP 25, we implement a backtracking line search as in CPO.

5 RELATIVE VARIANCE SCALING

We have described three formulations for the original VaR constraint in OP 8: the indicator constraint in Equation 10, and two safe approximations, the CVaR constraint in OP 12 and the Cantelli-VaR constraint in OP 15, which Canary is centered on. Each requires estimating a *constraint value* and its *policy gradient*, for Canary c_B, c_μ and b_B, b_μ. For now, we assume unbiased Monte Carlo estimators without a critic.

Given a metric m and its estimator $\hat{m}$, analyzing the total variance $\text{Var}[\hat{m}] = \mathbb{E}[||\hat{m} - \mathbb{E}[\hat{m}]||_2^2]$ in isolation is misleading because it might shrink as the signal m itself shrinks (Rubinstein & Kroese, 2016). Therefore, the *relative variance* (inverse signal-to-noise ratio) (Rubinstein & Kroese, 2016) defined by

$$\kappa^2[m] = \frac{\text{Var}[\hat{m}]}{||m||_2^2} \tag{33}$$

is often used instead. In deep RL, the explosion of this relative variance often manifests as a collapse in the cosine similarity between the empirical and true values, a phenomenon empirically identified as a driver of optimization failure in policy gradient methods (Ilyas et al., 2020). We call a method *stable* if, for a fixed policy, the relative variances of its constraint value and policy gradient remain bounded as the violation threshold becomes strict, $\varepsilon \to 0$.

Throughout this section and Appendix B, we assume costs are non-negative, $c_t \geq 0$ almost surely, so $C(\tau) \geq 0$. This permits an atom at zero, $\mathbb{P}(C(\tau) = 0) = 1 - f$, and defines cost regimes by $f = \mathbb{P}(C(\tau) > 0)$: the cost signal is *dense* if $f > \varepsilon$ and *sparse* otherwise. The results extend to signed costs. Replacing $\mathbb{P}(C(\tau) > 0)$ by $\mathbb{P}(C(\tau) \neq 0)$ throughout, the sparse regime is empty, absent an atom at zero. The scaling results further rely on four assumptions, stated here for reference and discussed in Appendix B.

Assumption 5.1 (Positive thresholds and regular quantile)**.** The cost limit, ℓ, and the $(1-\varepsilon)$-quantile of the cost return, η^* (7), are positive: $\ell > 0$ and $\eta^* > 0$. Consequently, $p \leq f$ and $\varepsilon < f$, where $p = \mathbb{P}(C(\tau) > \ell)$ is the violation probability. Moreover, the cost return has no atom at η^*, $\mathbb{P}(C(\tau) = \eta^*) = 0$, so $\mathbb{P}(C(\tau) > \eta^*) = \varepsilon$.

Assumption 5.2 (Bounded conditional relative second moment)**.** Let $X \in \{h(\tau), s(\tau)\, h(\tau)\}$, where $s(\tau) = \sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)$ is the score of the trajectory, and $h(\tau) = \mathbb{I}(C(\tau) > \ell)$ for the indicator, $h(\tau) = (C(\tau) - \eta^*)^+$ for CVaR, and $h(\tau) \in \{C(\tau), C(\tau)^2\}$ for the Cantelli moments. For $\mathcal{E} = \{h(\tau) \neq 0\}$: $\mathbb{E}[\|X\|^2 \mid \mathcal{E}] \,/\, \|\mathbb{E}[X \mid \mathcal{E}]\|^2 \leq K$ for a constant K independent of p, ε, and f.

Assumption 5.3 (No catastrophic cancellation)**.** Given $c_1 = 2(\ell - \mu(\theta)/\varepsilon)$ and $c_2 = \beta$. There exists a $\nu > 0$, independent of θ and (p, ε, f), such that $\|c_1 x_1 + c_2 x_2\|^2 \geq \nu(c_1^2 \|x_1\|^2 + c_2^2 \|x_2\|^2)$ for both the moment pair $(x_1, x_2) = (\mu(\theta), \tilde{\mu}(\theta))$ and the moment-gradient pair $(x_1, x_2) = (\nabla_\theta \mu(\theta), \nabla_\theta \tilde{\mu}(\theta))$.

Assumption 5.4 (Proportional slack)**.** Each constraint value decomposes as $J_C(\theta) = \mathbb{E}[x(\tau)] + b$, a *sampled part* $\mathbb{E}[x(\tau)]$, where the sampled statistic $x(\tau)$ vanishes off the active event $\mathcal{E}$, plus a batch-independent offset b. The *signal share* $r = |\mathbb{E}[x(\tau)]|/|J_C(\theta)|$ satisfies $r_{\min} \leq r \leq r_{\max}$ for constants $0 < r_{\min} \leq r_{\max} < \infty$ independent of (p, ε, f). For the Cantelli and mean constraints, for $X \in \{x(\tau), s(\tau)\, x(\tau)\}$, $\mathbb{E}[\|X\|^2 \mid \mathcal{E}] \,/\, \|\mathbb{E}[X \mid \mathcal{E}]\|^2 \geq 1 + \omega$ for a constant $\omega > 0$ independent of (p, ε, f).

Proposition 5.5 (Relative variance of constraints)**.** *Under Assumptions 5.1 to 5.4, for unbiased Monte Carlo estimators, the relative variance of the constraint value and gradient over a fixed batch of N trajectories satisfies the following bounds (Appendix B):* indicator*,* $\Theta(1/(pN))$ *(Rubinstein & Kroese, 2016);* CVaR*,* $\Theta(1/(\varepsilon N))$ *(Tamar et al., 2014); and* Cantelli*,* $\Theta(1/(fN))$*, where for signed costs* $f = \mathbb{P}(C(\tau) \neq 0)$*, which is* 1 *absent an atom at zero.*

The unbiased Cantelli estimators analyzed are not computable, so the relative variances of the biased *plug-in estimators* of c_B and b_B that Canary uses, whose bias shrinks as $1/N$, are also analyzed in Appendix B under one additional assumption (98). The constraint value keeps the $\Theta(1/(fN))$ rate to leading order and the gradient an $\mathcal{O}(1/(fN))$ upper bound. Therefore, only the Cantelli estimator is stable given dense cost regimes. The next section shows how this translates to improved safety for low ε regimes.

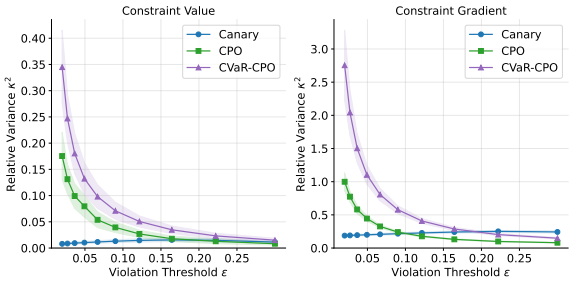


Figure 1: **Relative Variance Scaling:** IcyLake relative variance scaling for the constraint value (left) and constraint gradient (right) for Canary (Cantelli constraint) (blue), CPO (indicator) (green), and CVaR-CPO (purple). Each method is using a critic-dependent GAE advantage estimate for the policy gradient.

6 Results

We evaluate on three benchmarks run end-to-end on the GPU with JAX (Bradbury et al., 2018): **EcoAnt**, **PointGoal** (Ji et al., 2023), and **IcyLake** (Appendix D.1, and further experimental setup details in Appendix D). In practice, trajectories terminate or are truncated after a finite time T_τ, so $\mathcal{N} = \{0, \ldots, T_\tau - 1\}$. All results use 30 seeds; in every plot, lines are means and shaded areas one standard deviation over seeds.

We compare **Canary** against four baselines: the unconstrained **Proximal Policy Optimization (PPO)** (Schulman et al., 2017b); **Constrained Policy Optimization (CPO)** (Achiam et al., 2017), a trust-region CMDP method applied with the indicator constraint (Section 3.4) to enforce the VaR constraint (OP 8); **CVaR Proximal Policy Optimization (CPPO)** (Ying et al., 2022), Lagrangian PPO with a CVaR constraint (OP 12); and **CVaR Constrained Policy Optimization (CVaR-CPO)** (Zhang et al., 2025), a trust-region method with a CVaR constraint and a distributional cost critic.

6.1 Empirical Relative Variance Scaling

Figure 1 is consistent with the scaling behavior of Proposition 5.5, even with trust-region surrogates using critic-dependent GAE advantage estimates for policy gradients. The relative variances of the indicator and CVaR constraints grow as $\varepsilon \to 0$ while the Cantelli constraint remains flat.

6.2 Training Performance

We select violation threshold $\varepsilon = 0.02$ and cost limit ℓ for feasible settings in each environment (Table 1) with undiscounted costs $\gamma_c = 1$. Table 1 shows that Canary is the most *reliable* method (% safe) in every environment, and two measured properties contribute to this result: it converges furthest below the violation threshold ($\hat{P}$) and its violation probability varies least at convergence ($\hat{\sigma}$), so its margin per unit jitter, $(\varepsilon - \hat{P})/\hat{\sigma}$, is the largest of all methods. The theory supports these results. The Cantelli-VaR constraint is the most conservative surrogate (Section 4.1), which lowers the converged violation probability, and it is stable at low violation thresholds (Proposition 5.5), which lowers the jitter. A sweep over ε on IcyLake (Appendix D.5) shows that the convergence rate of the violation probability is also independent of ε, with converged rates settling 2–4$\times$ below the thresholds set.

6.3 Anatomy of the Conservatism

The converged violation probabilities in Table 1 settle well below the violation threshold ε, at some cost in reward. This conservatism could stem from *optimizer slack* (converging to the interior of the Cantelli-feasible set, forgoing reward the Cantelli constraint would permit) or *inequality slack* (the conservatism of the distribution-free Cantelli constraint itself). Figure 3 separates the two regimes: in EcoAnt and PointGoal every seed converges with $c_B > 0$, so the margin there is inequality slack, the regime whose training-time violation Theorem 4.1 bounds. On IcyLake most seeds (22/30) instead converge inside the boundary, so its conservatism stacks optimizer slack on top of inequality slack.

Remark 6.1 (Gaussian tails underestimate risk). It is common to model the cost return as Gaussian (Yang et al., 2021). At the converged cost return moments (mean μ and standard deviation σ), median $z = (\ell - \mu)/\sigma$ is 6.1 for EcoAnt and 5.0 for PointGoal, so a Gaussian tail predicts violation probabilities on the order of 10^{-9} and 10^{-7}, which are $2{\times}10^6$ times (EcoAnt) and $3{\times}10^4$ times (PointGoal) below the realized rates, an understatement of the exceedance probability by up to six orders of magnitude.

Table 1: Safety adherence across environments and methods (ε=0.02, 30 seeds, mean±standard error over seeds). *Converged violation probability* $\hat{P}$: per-seed average violation probability over the second half of training; *violation-probability jitter* $\hat{\sigma}$: per-seed standard deviation of the violation probability over the second half of training ($\times 10^{-3}$; Appendix D.4); *reliability* % safe: percent of seeds s whose converged violation probability satisfies $\hat{P}^{(s)} \leq \varepsilon$; *converged reward return* $\hat{J}$: per-seed average undiscounted reward return over the second half of training. Note, Canary settles at the lowest $\hat{P}$ and $\hat{\sigma}$ in every environment.

	Method	$\hat{P}\downarrow$	$\hat{\sigma}\downarrow$	% safe↑	$\hat{J}\uparrow$
EcoAnt ℓ=150	Canary	**.002** ±.000	**1.4** ±0.2	**100**	81±3
	PPO	.694 ±.003	36.6 ±2.1	0	1321±7
	CPO	.030 ±.000	4.2 ±0.4	0	203±6
	CPPO	.048 ±.003	20.9 ±2.2	3	48±5
	CVaR-CPO	.038 ±.002	16.8 ±3.6	0	123±5
PointGoal ℓ=100	Canary	**.011** ±.001	**6.1** ±1.4	**87**	4.0±0.6
	PPO	.512 ±.006	23.1 ±4.2	0	69.0±0.0
	CPO	.060 ±.002	31.0 ±2.6	0	12.1±1.2
	CPPO	.045 ±.010	22.2 ±4.0	47	7.3±1.6
	CVaR-CPO	.015 ±.001	11.8 ±1.5	**87**	6.4±2.0
IcyLake ℓ=5.5	Canary	**.016** ±.007	**7.1** ±2.7	**87**	.67±.06
	PPO	.308 ±.002	19.7 ±1.2	0	1.0±.00
	CPO	.041 ±.003	9.9 ±0.9	7	.96±.00
	CPPO	.354 ±.002	27.1 ±1.5	0	1.0±.00
	CVaR-CPO	.020 ±.002	7.5 ±1.3	53	.93±.01

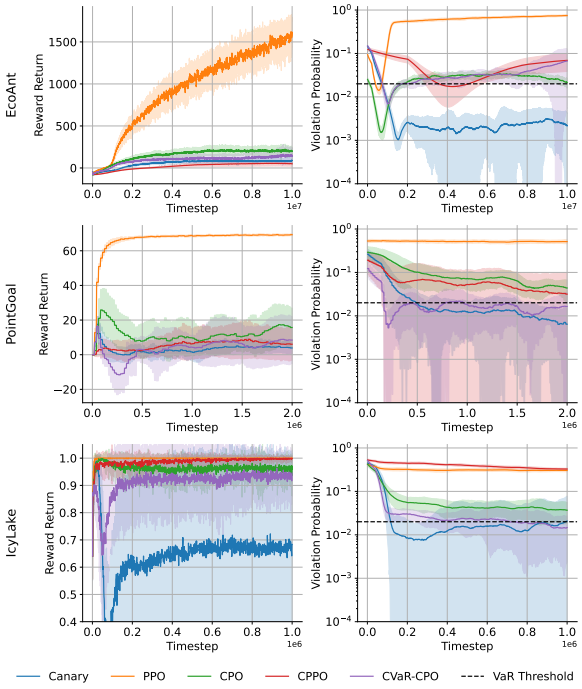


Figure 2: Reward return (left) and violation probability (right; log scale) during training on EcoAnt (ℓ=150, top), PointGoal (ℓ=100, middle) and IcyLake (ℓ=5.5, bottom) under the VaR constraint $\mathbb{P}(C(\tau) > \ell) \leq 0.02$.

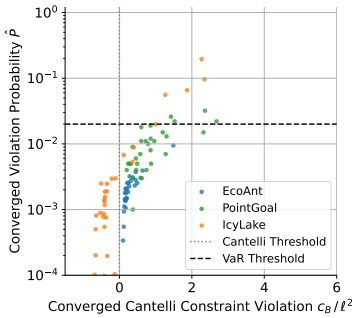


Figure 3: **Anatomy of the Conservatism**: Converged Cantelli constraint violation c_B/ℓ^2 (rescaled by ℓ^2 for plotting purposes, per-seed mean over the second half of training) against the converged violation probability $\hat{P}$ for EcoAnt (blue), PointGoal (green) and IcyLake (orange) at $\varepsilon = 0.02$ (log scale). Seeds right of the Cantelli threshold ($c_B > 0$) carry inequality slack only; seeds strictly inside the boundary add optimizer slack (Section 6.3). 82 of 90 seeds have converged violation probabilities below the VaR threshold.

7 Conclusion

We introduced Canary, an on-policy RL method that satisfies VaR constraints. We formulated a smooth and tractable safe approximation of VaR constraints from the first two moments of the cost return via Cantelli's inequality, extended CPO's worst-case violation bounds to the Canary update step, and showed that Canary remains stable at tight violation thresholds in dense cost regimes. Empirically, Canary attains the lowest converged violation probability and the least violation-probability jitter of the methods tested, and is therefore the most reliable across seeds.

Limitations. The Cantelli-VaR constraint can reject higher reward policies that satisfy the VaR constraint (Section 6.3), and settings infeasible for it may remain VaR-feasible. Stability requires a dense cost signal (Section 5). The bound of Theorem 4.1 omits advantage and moment estimation errors and loosens unboundedly as $\gamma_c \to 1$ at fixed δ.

AI USE STATEMENT

In this work, we used generative AI tools (large language model assistants) for three purposes. First, to aid and polish the writing: editing prose for clarity and concision, checking LaTeX formatting, and condensing sections to the page limit. Second, to draft sections of the paper from the authors' notes, derivations and results; every drafted passage was reviewed, revised and, where necessary, rewritten by the authors, and all claims in the text were checked by the authors against the results and proofs they describe. Third, for research execution: assistants were used to write and debug parts of the implementation of Canary, the baselines and the environments, to run and log experiments, and to discuss the design of the method and analysis. All code produced with assistance was reviewed by the authors and validated against reference implementations and the reported baseline behavior, and every reported number was produced by the authors' own runs of that code. The theoretical results and their proofs were derived and checked line by line by the authors. We have not used generative AI tools for literature retrieval, for generating data, or for proving mathematical claims. We take responsibility for the final content of this work, including any text, code or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Proofs of the Cantelli-VaR constraint (14), the augmented-cost decomposition (17), the trust-region surrogates (23), Theorem 4.1 and Corollary 4.2 are given in Appendix A, and the relative-variance results of Section 5, with the assumptions they rely on, are derived in Appendix B. The Canary algorithm and its closed-form solver are stated in Appendix C. The environments, the Canary and baseline hyperparameters, the tuning budget of each baseline, the seed protocol, and the definition of the reported statistics are given in Appendix D. Source code for Canary, the baselines and the JAX environments is provided as supplementary material.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization, 2017. URL https://arxiv.org/abs/1705.10528.

Erik Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1999.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay, 2018. URL https://arxiv.org/abs/1707.01495.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Antonio Briola, Jeremy Turiel, Riccardo Marcaccioli, Alvaro Cauderan, and Tomaso Aste. Deep reinforcement learning for active high frequency trading, 2023. URL https://arxiv.org/abs/2101.07107.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *J. Mach. Learn. Res.*, 18(1):6070–6120, January 2017. ISSN 1532-4435.

Laurent El Ghaoui, Maksim Oks, and François Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, 2003.

Peter Geibel and Fritz Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *Journal of Artificial Intelligence Research*, 24:81–108, 2005.

Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance, 2023. URL https://arxiv.org/abs/2112.04553.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients, 2020. URL https://arxiv.org/abs/1811.02553.

Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=WZmlxIuIGR.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey, 2021. URL https://arxiv.org/abs/2002.00444.

Shiau Hong Lim and Ilyas Malik. Distributional reinforcement learning for risk-sensitive policies. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30977–30989. Curran Associates, Inc., 2022.

Arkadi Nemirovski. On safe tractable approximations of chance constraints. *European Journal of Operational Research*, 219(3):707–718, 2012.

Masahiro Ono, Marco Pavone, Yoshiaki Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4): 555–571, 2015.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at risk. *Journal of Risk*, 3:21–41, 2000. URL https://api.semanticscholar.org/CorpusID:854622.

R.Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. ISSN 0378-4266. doi: https://doi.org/10.1016/S0378-4266(02)00271-6. URL https://www.sciencedirect.com/science/article/pii/S0378426602002716.

Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016.

Pedro Santana, Sylvie Thiébaux, and Brian Williams. RAO*: An algorithm for chance-constrained POMDP's. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 3308–3314, 2016.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2017a. URL https://arxiv.org/abs/1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017b. URL https://arxiv.org/abs/1707.06347.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018. URL https://arxiv.org/abs/1506.02438.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, 2009.

Aivar Sootla, Alexander I. Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H. Mguni, Jun Wang, and Haitham Ammar. Saute RL: Almost surely safe reinforcement learning using state augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20423–20443. PMLR, 2022.

Bartolomeo Stellato, Bart P. G. Van Parys, and Paul J. Goulart. Multivariate chebyshev inequality with estimated mean and variance. *The American Statistician*, 71(2):123–127, April 2017. ISSN 1537-2731. doi: 10.1080/00031305.2016.1186559. URL http://dx.doi.org/10.1080/00031305.2016.1186559.

Koyoharu Tagawa. Chebyshev inequality based approach to chance constrained portfolio optimization, 2017. URL https://www.iaras.org/iaras/filedownloads/ijmcm/2017/001-0009(2017).pdf.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling, 2014. URL https://arxiv.org/abs/1404.3862.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization, 2018. URL https://arxiv.org/abs/1805.11074.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Huan Xu and Shie Mannor. Probabilistic goal Markov decision processes. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 2046–2052, 2011.

Qisong Yang, Thiago D. Simão, Simon H. Tindemans, and Matthijs T. J. Spaan. WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10639–10646, 2021.

Chengyang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk, 2022. URL https://arxiv.org/abs/2206.04436.

Junzi Zhang, Jongho Kim, Brendan O'Donoghue, and Stephen Boyd. Sample efficient reinforcement learning with reinforce, 2020. URL https://arxiv.org/abs/2010.11364.

Qiyuan Zhang, Shu Leng, Xiaoteng Ma, Qihan Liu, Xueqian Wang, Bin Liang, Yu Liu, and Jun Yang. Cvar-constrained policy optimization for safe reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):830–841, 2025. doi: 10.1109/TNNLS.2023.3331304.

Steve Zymler, Daniel Kuhn, and Berç Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1–2):167–198, 2013.

A Canary Method Proofs

A.1 Derivation of Equation 14

Assume first that $\mu(\pi) < \ell$, so that $\lambda = \ell - \mu(\pi) > 0$. By Cantelli's inequality:

$$\mathbb{P}(C(\tau) - \mu(\pi) \geq \lambda) \leq \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \lambda^2},$$

which is equivalent to:

$$\mathbb{P}(C(\tau) > \ell) \leq \mathbb{P}(C(\tau) \geq \ell) \leq \frac{\sigma^2(\pi)}{\sigma^2(\pi) + [\ell - \mu(\pi)]^2},$$

where the first inequality holds because $\{C(\tau) > \ell\} \subseteq \{C(\tau) \geq \ell\}$.

To guarantee that this probability is bounded by some $\varepsilon \in (0, 1]$, it is sufficient to require that the upper bound itself is less than or equal to ε:

$$\frac{\sigma^2(\pi)}{\sigma^2(\pi) + [\ell - \mu(\pi)]^2} \leq \varepsilon.$$

We can now simplify this condition. Since $\lambda > 0$ and $\varepsilon > 0$, the following steps are logically equivalent:

$$\begin{aligned}
&\frac{\sigma^2(\pi)}{\sigma^2(\pi) + [\ell - \mu(\pi)]^2} \leq \varepsilon \\
\iff \quad &\sigma^2(\pi) \leq \varepsilon[\sigma^2(\pi) + [\ell - \mu(\pi)]^2] \\
\iff \quad &\sigma^2(\pi) \leq \varepsilon\sigma^2(\pi) + \varepsilon[\ell - \mu(\pi)]^2 \\
\iff \quad &(1 - \varepsilon)\sigma^2(\pi) \leq \varepsilon[\ell - \mu(\pi)]^2 \\
\iff \quad &\frac{1 - \varepsilon}{\varepsilon}\sigma^2(\pi) \leq [\ell - \mu(\pi)]^2 \\
\iff \quad &\left(\frac{1}{\varepsilon} - 1\right)\sigma^2(\pi) - [\ell - \mu(\pi)]^2 \leq 0.
\end{aligned}$$

In the boundary case $\mu(\pi) = \ell$, Cantelli's inequality no longer applies ($\lambda = 0$), so we verify the VaR constraint directly. Substituting $\mu(\pi) = \ell$ into the Cantelli constraint (14) eliminates its second term:

$$B(\pi) = \left(\frac{1}{\varepsilon} - 1\right)\sigma^2(\pi) - [\ell - \ell]^2 = \left(\frac{1}{\varepsilon} - 1\right)\sigma^2(\pi) \leq 0.$$

Since $\varepsilon \in (0, 1)$, the factor $\frac{1}{\varepsilon} - 1$ is strictly positive, and $\sigma^2(\pi) \geq 0$ by definition; their product can only be non-positive if $\sigma^2(\pi) = 0$. A cost return with zero variance equals its mean with probability one, so $C(\tau) = \mu(\pi) = \ell$ almost surely and $\mathbb{P}(C(\tau) > \ell) = 0 \leq \varepsilon$. Combining the two cases, $\mu(\pi) < \ell$ (via Cantelli's inequality) and $\mu(\pi) = \ell$ (via the zero-variance argument above): under the mean constraint $\mu(\pi) \leq \ell$, the Cantelli constraint implies the VaR constraint $\mathbb{P}(C(\tau) > \ell) \leq \varepsilon$.

A.2 Derivation of Equation 17

It is possible to break down the square cost return into the discounted sum of local per-step terms:

$$
\begin{aligned}
C(\tau)^2 &= \left(\sum_{t=0}^{\infty} \gamma_c^t c_t\right)^2 \\
&= \sum_{t=0}^{\infty} \gamma_c^t c_t \sum_{k=0}^{\infty} \gamma_c^k c_k \\
&= \sum_{t=0}^{\infty} \gamma_c^t c_t \left(\sum_{k=0}^{t-1} \gamma_c^k c_k + \gamma_c^t c_t + \sum_{k=t+1}^{\infty} \gamma_c^k c_k\right) \\
&= \sum_{t=0}^{\infty} \gamma_c^t c_t \left(y_t + \gamma_c^t c_t + \sum_{k=t+1}^{\infty} \gamma_c^k c_k\right) \\
&= \sum_{t=0}^{\infty} \gamma_c^{2t} c_t^2 + \sum_{t=0}^{\infty} \gamma_c^t c_t y_t + \sum_{t=0}^{\infty} \gamma_c^t c_t \sum_{k=t+1}^{\infty} \gamma_c^k c_k \\
&= \sum_{t=0}^{\infty} \gamma_c^{2t} c_t^2 + 2 \sum_{t=0}^{\infty} \gamma_c^t c_t y_t \\
&= \sum_{t=0}^{\infty} \gamma_c^t (\gamma_c^t c_t^2 + 2 c_t y_t).
\end{aligned}
$$

Writing the square as a product of two copies of the sum, indexed by t and k, lets us hold the outer index t fixed and split only the inner sum, into the terms strictly before t, the term at $k = t$, and the terms strictly after t. The first of these is by definition the accumulated cost $y_t = \sum_{k=0}^{t-1} \gamma_c^k c_k$, which is where the recursion in Equation 16 enters.

The only remaining step is the last cross term. Relabeling its two indices, $(t, k) \mapsto (k, t)$, maps the region $k > t$ onto the region $k < t$ and leaves the summand $\gamma_c^{k+t} c_k c_t$ unchanged, so

$$
\sum_{t=0}^{\infty} \gamma_c^t c_t \sum_{k=t+1}^{\infty} \gamma_c^k c_k = \sum_{t=0}^{\infty} \gamma_c^t c_t \sum_{k=0}^{t-1} \gamma_c^k c_k = \sum_{t=0}^{\infty} \gamma_c^t c_t y_t,
$$

which is what makes the two cross terms combine into the factor of 2. Each unordered pair $\{k, t\}$ with $k \neq t$ is counted once looking forward from t and once looking back from k; y_t collects only the backward half, and the factor of 2 accounts for the other.

A.3 Derivation of Equation 21

We need to show that we can reconstruct the Cantelli-VaR constraint in a CMDP form. First, we take the moments of the cost return to define the mean and variance:

$$
\begin{aligned}
\mu(\pi) &= \mathbb{E}_{\tau \sim \pi}\left[C(\tau)\right] \\
\tilde{\mu}(\pi) &= \mathbb{E}_{\tau \sim \pi}\left[C(\tau)^2\right] \\
\sigma^2(\pi) &= \tilde{\mu}(\pi) - \mu(\pi)^2.
\end{aligned}
$$

These can be combined with the square cost return in Equation 17 to break down the global episode-level Cantelli-VaR constraint into a form containing a local per-step discounted term with the augmented state space (20):

$$
\begin{aligned}
B(\pi) &= \left(\frac{1}{\varepsilon} - 1\right) \sigma^2(\pi) - [\ell - \mu(\pi)]^2 \\
&= \left(\frac{1}{\varepsilon} - 1\right) [\tilde{\mu}(\pi) - \mu(\pi)^2] - \ell^2 + 2\ell\mu(\pi) - \mu(\pi)^2 \\
&= \left(\frac{1}{\varepsilon} - 1\right) \tilde{\mu}(\pi) - \frac{1}{\varepsilon}\mu(\pi)^2 + 2\ell\mu(\pi) - \ell^2 \\
&= \left(\frac{1}{\varepsilon} - 1\right) \tilde{\mu}(\pi) + 2\ell\mu(\pi) - \left[\frac{1}{\varepsilon}\mu(\pi)^2 + \ell^2\right]
\end{aligned}
$$

This maps to the final constraint form in Equation 21 setting the policy-dependent bound $\rho(\pi) = \frac{1}{\varepsilon}\mu(\pi)^2 + \ell^2$:

$$
B(\pi) = \left(\frac{1}{\varepsilon} - 1\right) \tilde{\mu}(\pi) + 2\ell\mu(\pi) - \rho(\pi) \leq 0
$$

The bundling of $\mu(\pi)^2$ into $\rho(\pi)$ is not merely notational. The terms $\tilde{\mu}(\pi)$ and $\mu(\pi)$ are expected discounted sums of per-step costs on the augmented state space, so they admit advantage-based trust-region surrogates. In contrast, $\mu(\pi)^2$ is the square of an expectation: it equals $\mathbb{E}_{\tau,\tau'\sim\pi}[C(\tau)C(\tau')]$ for independent trajectories τ and τ', an expectation over a pair of trajectories rather than one, so it cannot be written as the expected return of any per-step cost, even after state augmentation. We therefore move $\mu(\pi)^2$, together with the constant ℓ^2, to the constraint-limit side as the policy-dependent bound $\rho(\pi)$. This term is approximated by substituting the trust-region surrogate $L_\mu(\pi)$ (Equation (5)) into the square rather than through a trust-region surrogate of its own, which is the origin of the quadratic term in Equation 23 (Appendix A.4), and its approximation error is bounded separately via a difference-of-squares argument in Section A.5.

A.4 Derivation of Equation 23

First, we can expand the difference between $\rho(\pi)$ and $\rho(\pi_k)$:

$$
\begin{aligned}
\rho(\pi) - \rho(\pi_k) &= \frac{1}{\varepsilon}[\mu(\pi)^2 - \mu(\pi_k)^2] \\
\rho(\pi) &= \rho(\pi_k) + \frac{1}{\varepsilon}[\mu(\pi)^2 - \mu(\pi_k)^2].
\end{aligned}
$$

Then, we can take the trust-region surrogate of $\mu(\pi)$ around some other policy π_k:

$$
\begin{aligned}
\mu(\pi) \approx L_\mu(\pi) &= \mu(\pi_k) + \frac{1}{1-\gamma_c}\mathbb{E}_{\substack{x\sim d^c_{\pi_k} \\ a\sim\pi}}[A^C_{\pi_k}(x,a)] \\
L_\mu(\pi)^2 &= \mu(\pi_k)^2 + \frac{2\mu(\pi_k)}{1-\gamma_c}\mathbb{E}_{\substack{x\sim d^c_{\pi_k} \\ a\sim\pi}}[A^C_{\pi_k}(x,a)] + \frac{1}{(1-\gamma_c)^2}\mathbb{E}_{\substack{x\sim d^c_{\pi_k} \\ a\sim\pi}}[A^C_{\pi_k}(x,a)]^2.
\end{aligned}
$$

Substituting this in gets the trust-region surrogate $L_\rho(\pi)$:

$$
\begin{aligned}
L_\rho(\pi) &= \rho(\pi_k) + \frac{1}{\varepsilon}[L_\mu(\pi)^2 - \mu(\pi_k)^2] \\
&= \rho(\pi_k) + \frac{1}{\varepsilon(1-\gamma_c)}\left(2\mu(\pi_k)\mathbb{E}_{\substack{x\sim d^c_{\pi_k} \\ a\sim\pi}}[A^C_{\pi_k}(x,a)] + \frac{1}{1-\gamma_c}\mathbb{E}_{\substack{x\sim d^c_{\pi_k} \\ a\sim\pi}}[A^C_{\pi_k}(x,a)]^2\right).
\end{aligned}
$$

A.5 Derivation of Theorem 4.1

Given a current policy π_k, we aim to show the worst-case constraint violation of a policy π_{k+1} following the Canary update rule (25). We start by considering the factored Cantelli-VaR constraint

$B(\pi)$:

$$B(\pi_{k+1}) = \left(\frac{1}{\varepsilon} - 1\right) \tilde{\mu}(\pi_{k+1}) + 2\ell\mu(\pi_{k+1}) - \rho(\pi_{k+1}).$$

Using the trust-region surrogates $L_{\tilde{\mu}}$, L_μ, and L_ρ, the algorithm update rule enforces the linearized constraint $\left(\frac{1}{\varepsilon} - 1\right) L_{\tilde{\mu}}(\pi_{k+1}) + 2\ell L_\mu(\pi_{k+1}) - L_\rho(\pi_{k+1}) \leq 0$. Therefore, the true constraint violation is bounded by the approximation error:

$$\begin{aligned} B(\pi_{k+1}) &\leq B(\pi_{k+1}) - \left[\left(\frac{1}{\varepsilon} - 1\right) L_{\tilde{\mu}}(\pi_{k+1}) + 2\ell L_\mu(\pi_{k+1}) - L_\rho(\pi_{k+1})\right] \\ &= \left(\frac{1}{\varepsilon} - 1\right) \left[\tilde{\mu}(\pi_{k+1}) - L_{\tilde{\mu}}(\pi_{k+1})\right] + 2\ell \left[\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})\right] + \left[L_\rho(\pi_{k+1}) - \rho(\pi_{k+1})\right]. \end{aligned}$$

Applying the triangle inequality gives:

$$B(\pi_{k+1}) \leq \left(\frac{1}{\varepsilon} - 1\right) |\tilde{\mu}(\pi_{k+1}) - L_{\tilde{\mu}}(\pi_{k+1})| + 2|\ell| \, |\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})| + |L_\rho(\pi_{k+1}) - \rho(\pi_{k+1})|.$$

From standard CPO theory, the worst-case approximation errors for the linear expected returns are bounded by:

$$\begin{aligned} |\tilde{\mu}(\pi_{k+1}) - L_{\tilde{\mu}}(\pi_{k+1})| &\leq \frac{\sqrt{2\delta}\gamma_c}{(1-\gamma_c)^2} \alpha^{\tilde{C}}_{\pi_{k+1}} \\ |\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})| &\leq \frac{\sqrt{2\delta}\gamma_c}{(1-\gamma_c)^2} \alpha^{C}_{\pi_{k+1}}, \end{aligned}$$

where $\alpha^{\tilde{C}}_{\pi_{k+1}} = \max_x |\mathbb{E}_{a\sim\pi_{k+1}}[A^{\tilde{C}}_{\pi_k}(x,a)]|$ and $\alpha^{C}_{\pi_{k+1}} = \max_x |\mathbb{E}_{a\sim\pi_{k+1}}[A^{C}_{\pi_k}(x,a)]|$.

We then need to deal with the approximation error in the constraint limit $\rho(\pi_{k+1})$ itself. First we define variables X and Y for notational brevity:

$$\begin{aligned} X &= \mathbb{E}_{\substack{x\sim d^c_{\pi_{k+1}} \\ a\sim\pi_{k+1}}}[A^C_{\pi_k}(x,a)] = \sum_x d^c_{\pi_{k+1}}(x) \sum_a \pi_{k+1}(a \mid x) A^C_{\pi_k}(x,a) \\ Y &= \mathbb{E}_{\substack{x\sim d^c_{\pi_k} \\ a\sim\pi_{k+1}}}[A^C_{\pi_k}(x,a)] = \sum_x d^c_{\pi_k}(x) \sum_a \pi_{k+1}(a \mid x) A^C_{\pi_k}(x,a). \end{aligned}$$

This allows us to write the error in the constraint limit itself as:

$$|L_\rho(\pi_{k+1}) - \rho(\pi_{k+1})| \leq \frac{1}{\varepsilon(1-\gamma_c)} \left[2|\mu(\pi_k)| \, |X - Y| + \frac{1}{1-\gamma_c}|X^2 - Y^2|\right].$$

To bound this, we now need to find the limits for $|X-Y|$ and $|X^2-Y^2|$. Defining the total variational divergence of the policy update in discrete action space as $D_{TV}(\pi_{k+1}, \pi_k) = \frac{1}{2}\sum_a |\pi_{k+1}(a \mid x) - \pi_k(a \mid x)|$, we can use the existing result of the worst case state visitation frequency difference:

$$||d^c_{\pi_{k+1}}(x) - d^c_{\pi_k}(x)||_1 \leq \frac{2\gamma_c}{1-\gamma_c} \mathbb{E}_{x\sim d^c_{\pi_k}}[D_{TV}(\pi_{k+1}, \pi_k)],$$

and applying the same trust region bound theory in CPO:

$$|X - Y| \leq \frac{2\gamma_c \alpha^C_{\pi_{k+1}}}{1-\gamma_c} \mathbb{E}_{x\sim d^c_{\pi_k}}[D_{TV}(\pi_{k+1}, \pi_k)].$$

To handle $|X^2 - Y^2|$, we can first decompose it:

$$|X^2 - Y^2| = |X - Y||X + Y|.$$

This is useful since we previously defined the bound for $|X - Y|$ already, so we are now left with $|X + Y|$:

$$|X + Y| \leq |X| + |Y|.$$

First, we can define an inner term of X and Y as a function of the augmented state x alone:

$$f(x) = \mathbb{E}_{a\sim\pi_{k+1}}[A^C_{\pi_k}(x,a)]$$
$$\alpha^C_{\pi_{k+1}} = \max_x |f(x)|.$$

Then we can derive the bound for $|X|$:

$$\begin{aligned}|X| &\leq \sum_x d^c_{\pi_{k+1}}(x)\cdot|f(x)| \\ &\leq \sum_x d^c_{\pi_{k+1}}(x)\cdot\alpha^C_{\pi_{k+1}} \\ &\leq \alpha^C_{\pi_{k+1}}\left(\sum_x d^c_{\pi_{k+1}}(x)\right) \\ &\leq \alpha^C_{\pi_{k+1}}.\end{aligned}$$

The same steps apply for $|Y|$:

$$|Y| \leq \alpha^C_{\pi_{k+1}}$$

such that the bound for $|X+Y|$ can be derived:

$$|X+Y| \leq 2\alpha^C_{\pi_{k+1}}$$

and $|X^2 - Y^2|$:

$$|X^2 - Y^2| \leq \frac{4\gamma_c(\alpha^C_{\pi_{k+1}})^2}{1-\gamma_c}\mathbb{E}_{x\sim d^c_{\pi_k}}[D_{TV}(\pi_{k+1},\pi_k)].$$

We can also use Pinsker's inequality to bound the total variational distance with the KL divergence, which in turn is bounded in the update step by δ:

$$D_{TV}(P,Q) \leq \sqrt{\frac{1}{2}D_{KL}(P,Q)},$$

and putting everything together, we can create the final worst-case constraint violation by summing the three separated bound components:

$$B(\pi_{k+1}) \leq \frac{\sqrt{2\delta}\gamma_c}{(1-\gamma_c)^2}\left(\left[\frac{1}{\varepsilon}-1\right]\alpha^{\tilde{C}}_{\pi_{k+1}} + 2|\ell|\alpha^C_{\pi_{k+1}} + \frac{2\alpha^C_{\pi_{k+1}}}{\varepsilon}\left[|\mu(\pi_k)| + \frac{\alpha^C_{\pi_{k+1}}}{1-\gamma_c}\right]\right).$$

The mean overshoot bound (26) follows directly from the linearized mean constraint of OP 25 together with the approximation error bound above: $L_\mu(\pi_{k+1}) \leq \ell$, and $|\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})| \leq \frac{\sqrt{2\delta}\gamma_c}{(1-\gamma_c)^2}\alpha^C_{\pi_{k+1}}$, so

$$\mu(\pi_{k+1}) \leq L_\mu(\pi_{k+1}) + |\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})| \leq \ell + \frac{\sqrt{2\delta}\gamma_c}{(1-\gamma_c)^2}\alpha^C_{\pi_{k+1}}.$$

A.6 Proof of Corollary 4.2

We first show that the VaR constraint is provably violated on V. Define the reflected cost return $X = -C(\tau)$, with

$$\mu_X(\pi) = -\mu(\pi), \qquad \sigma^2_X(\pi) = \sigma^2(\pi), \qquad \ell_X = -\ell.$$

Reflection exchanges the two tails,

$$\{X \geq \ell_X\} = \{-C(\tau) \geq -\ell\} = \{C(\tau) \leq \ell\},$$

so an upper bound on the upper tail of X is an upper bound on the lower tail of $C(\tau)$. Mirroring the construction of Appendix A.1, which sets $\lambda = \ell - \mu(\pi)$, we take

$$\lambda_X = \ell_X - \mu_X(\pi) = \mu(\pi) - \ell,$$

which is positive exactly when $\mu(\pi) > \ell$. Cantelli's inequality (13) applied to X then gives

$$\mathbb{P}\big(X - \mu_X(\pi) \geq \lambda_X\big) \leq \frac{\sigma_X^2(\pi)}{\sigma_X^2(\pi) + \lambda_X^2},$$

and substituting the definitions back,

$$\mathbb{P}\big(C(\tau) \leq \ell\big) \leq \frac{\sigma^2(\pi)}{\sigma^2(\pi) + [\mu(\pi) - \ell]^2}.$$

Since $\mathbb{P}(C(\tau) > \ell) = 1 - \mathbb{P}(C(\tau) \leq \ell)$, taking complements gives

$$\mathbb{P}\big(C(\tau) > \ell\big) \geq \frac{[\mu(\pi) - \ell]^2}{\sigma^2(\pi) + [\mu(\pi) - \ell]^2}.$$

Requiring this lower bound to exceed ε, and using $\varepsilon \in (0, 1)$ and $\lambda_X > 0$ throughout,

$$\begin{aligned} \frac{\lambda_X^2}{\sigma^2(\pi) + \lambda_X^2} > \varepsilon \quad &\iff \quad \lambda_X^2(1 - \varepsilon) > \varepsilon\sigma^2(\pi) \\ &\iff \quad \lambda_X > \sigma(\pi)\sqrt{\frac{\varepsilon}{1 - \varepsilon}} = \frac{\sigma(\pi)}{\sqrt{\frac{1}{\varepsilon} - 1}}, \end{aligned}$$

that is, $\mu(\pi) > \ell + \sigma(\pi)/\sqrt{\frac{1}{\varepsilon} - 1}$, which defines the provably violating region V.

We now bound the expected cost return of each tier. All three tiers of the selection rule in Section 4.5 constrain π_{k+1} to the trust region $\bar{D}_{KL}(\pi_{k+1}, \pi_k) \leq \delta$, so the CPO approximation-error bound used in Section A.5, $|\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})| \leq K\alpha^C_{\pi_{k+1}}$, applies to every tier. In particular,

$$\mu(\pi_{k+1}) \leq L_\mu(\pi_{k+1}) + |\mu(\pi_{k+1}) - L_\mu(\pi_{k+1})| \leq L_\mu(\pi_{k+1}) + K\alpha^C_{\pi_{k+1}}, \tag{34}$$

and it suffices to bound the surrogate $L_\mu(\pi_{k+1})$ in each tier.

Tiers a–b ($m_\mu \leq \ell$). Both OP 25 and Tier b of (28) contain the constraint $L_\mu(\pi_{k+1}) \leq \ell$, so (34) gives $\mu(\pi_{k+1}) \leq \ell + K\alpha^C_{\pi_{k+1}}$ directly. For Tier a, this is the expected cost return bound (26) of Theorem 4.1. Membership of V requires $\mu(\pi_{k+1}) > \ell + \sigma(\pi_{k+1})/\sqrt{\frac{1}{\varepsilon} - 1}$. If $\sigma(\pi_{k+1}) \geq \sqrt{\frac{1}{\varepsilon} - 1}\, K\alpha^C_{\pi_{k+1}}$, then

$$\ell + \frac{\sigma(\pi_{k+1})}{\sqrt{\frac{1}{\varepsilon} - 1}} \geq \ell + K\alpha^C_{\pi_{k+1}} \geq \mu(\pi_{k+1}),$$

so $\pi_{k+1} \notin V$.

Tier c ($m_\mu > \ell$). The current policy is always admissible, with $L_\mu(\pi_k) = \mu(\pi_k)$ by the first-order validity of the surrogate (Section 3.2), so the Tier c minimizer in (28) satisfies $L_\mu(\pi_{k+1}) = m_\mu \leq \mu(\pi_k)$. Combining with (34),

$$\mu(\pi_{k+1}) \leq \mu(\pi_k) + K\alpha^C_{\pi_{k+1}}.$$

B Relative Variance of VaR Constraint Estimators

As established in Section 5, a constrained policy update depends on estimates of both the constraint value and its policy gradient. We study the statistical quality of these estimates through their relative variance (inverse signal-to-noise ratio), defined in Equation (33). The purpose of this appendix is to characterize how the relative variance of the indicator, CVaR, and Cantelli estimators depends on the fraction of a trajectory batch that is informative for each estimator.

Throughout, as in Section 5, costs are non-negative, $c_t \geq 0$ almost surely, so $C(\tau) \geq 0$, and we consider a batch of N independent trajectories

$$\tau_1, \ldots, \tau_N \overset{\text{i.i.d.}}{\sim} \pi_\theta.$$

Independence is assumed only across trajectories; states and actions within an individual trajectory retain the temporal dependence induced by the MDP.

We write each method's constraint in the canonical form

$$J_C(\theta) \leq 0,$$

where $J_C(\theta)$ is the population constraint value and $\hat{J}_C(\theta)$ its Monte Carlo estimator. Similarly, $\nabla_\theta J_C(\theta)$ denotes the population constraint gradient and $\hat{g}$ its Monte Carlo estimator. These are the quantities consumed by a constrained optimizer, for example in a Lagrangian multiplier update or a trust-region feasibility test.

For a quantity m with estimator $\hat{m}$, recall from Equation (33) that its relative variance is

$$\kappa^2[m] = \frac{\mathrm{Var}[\hat{m}]}{\|m\|_2^2}.$$

Relative variance is therefore undefined when the underlying signal vanishes. In particular, in our case, when the quantity in question is $J_C(\theta)$, the relative variance of a constraint-value estimator diverges mechanically at the constraint boundary $J_C(\theta) = 0$, independently of the estimator being considered. This is an artifact of the metric rather than of any particular estimator, and Assumption 5.4 quantifies the distance from the boundary at which the constraint values are compared.

Effective sample size. The three estimators differ fundamentally in the fraction of a trajectory batch that enters their trajectory-level cost statistic. To make this precise, let $h(\tau)$ denote the trajectory-level statistic associated with an estimator, and define its *active event* by

$$\mathcal{E}_h := \{\tau : h(\tau) \neq 0\}.$$

For a realized batch, define the *effective sample size* associated with h as

$$N_{\mathrm{eff}}(h) := \sum_{i=1}^{N} \mathbf{1}\{\tau_i \in \mathcal{E}_h\} = \sum_{i=1}^{N} \mathbf{1}\{h(\tau_i) \neq 0\}.$$

Thus, $N_{\mathrm{eff}}(h)$ is a random variable. Since the trajectories are identically distributed,

$$\mathbb{E}[N_{\mathrm{eff}}(h)] = N\,\mathbb{P}(\mathcal{E}_h).$$

We use this expected effective sample size to characterize the sampling regime of each estimator.

Let

$$p := \mathbb{P}\big(C(\tau) > \ell\big)$$

denote the policy's violation probability, and let

$$f := \mathbb{P}\big(C(\tau) > 0\big)$$

denote the probability that a trajectory incurs any positive cost. Since Assumption 5.1 requires $\ell > 0$,

$$\{C(\tau) > \ell\} \subseteq \{C(\tau) > 0\},$$

and hence

$$p \leq f.$$

For the *indicator estimator*, the trajectory-level statistic is

$$h_{\mathrm{Ind}}(\tau) := \mathbf{1}\{C(\tau) > \ell\}.$$

Its active event is therefore

$$\mathcal{E}_{\mathrm{Ind}} = \{C(\tau) > \ell\},$$

which occurs with probability p. Consequently,

$$\mathbb{E}\left[N_{\mathrm{eff}}(h_{\mathrm{Ind}})\right] = pN.$$

Since $p \leq f$, the indicator estimator's expected effective sample size pN is at most fN, and shrinks as the threshold tightens while f does not.

For the *CVaR estimator*, the trajectory-level statistic is

$$h_{\mathrm{CVaR}}(\tau) := \big(C(\tau) - \eta^*\big)^+.$$

Its active event is

$$\mathcal{E}_{\mathrm{CVaR}} = \{C(\tau) > \eta^*\}.$$

Under Assumption 5.1, which requires that the cost return has no atom at η^*,

$$\mathbb{P}\big(C(\tau) > \eta^*\big) = \varepsilon,$$

and hence

$$\mathbb{E}\left[N_{\mathrm{eff}}(h_{\mathrm{CVaR}})\right] = \varepsilon N.$$

Akin to the indicator, this is thresholded by ε rather than f: since Assumption 5.1 requires $\eta^* > 0$,

$$\{C(\tau) > \eta^*\} \subseteq \{C(\tau) > 0\},$$

so $\varepsilon \leq f$. Furthermore, the inequality is strict, because $\eta^* > 0$ is the infimum in Equation (7), so $\eta = 0$ does not satisfy $\mathbb{P}(C(\tau) > \eta) \leq \varepsilon$, that is, $\varepsilon < f$. The CVaR estimator's expected effective sample size εN is therefore strictly below fN, and shrinks as the threshold tightens while f does not.

The *Cantelli estimator* does not threshold trajectories at either ℓ or η^*. Instead, it estimates the first two cost moments using the trajectory-level statistics

$$h_1(\tau) := C(\tau), \qquad h_2(\tau) := C(\tau)^2.$$

Since costs are non-negative, both statistics have the same active event,

$$\mathcal{E}_{\mathrm{Cantelli}} = \{C(\tau) > 0\},$$

which occurs with probability f. Therefore,

$$\mathbb{E}\left[N_{\mathrm{eff}}(h_1)\right] = \mathbb{E}\left[N_{\mathrm{eff}}(h_2)\right] = fN.$$

Thus, the expected effective sample sizes associated with the three methods are

Estimator	Active event	$\mathbb{E}[N_{\mathrm{eff}}]$
Indicator	$\{C(\tau) > \ell\}$	pN
CVaR	$\{C(\tau) > \eta^*\}$	εN
Cantelli	$\{C(\tau) > 0\}$	fN

.

These expected effective sample sizes provide the intuition for the relative-variance scaling derived below. Under the corresponding non-degeneracy assumptions, the derivations show that the relative variance scales, up to constant factors, as the inverse expected effective sample size,

$$\kappa^2 = \Theta\left(\frac{1}{\mathbb{E}[N_{\mathrm{eff}}]}\right).$$

This gives the scaling classes

$$\Theta\left(\frac{1}{pN}\right), \qquad \Theta\left(\frac{1}{\varepsilon N}\right), \qquad \Theta\left(\frac{1}{fN}\right)$$

for the indicator, CVaR, and Cantelli estimators, respectively. The following subsections establish the corresponding results and make explicit the conditions under which these scaling classes hold.

Dense and sparse cost regimes. The distinction between the three expected effective sample sizes is most important in the dense-cost, low-violation regime. If f remains bounded away from zero while $p, \varepsilon \to 0$, then the expected fractions of the batch active for the indicator and CVaR estimators satisfy

$$\frac{\mathbb{E}[N_{\mathrm{eff}}(h_{\mathrm{Ind}})]}{N} = p \to 0,$$

and

$$\frac{\mathbb{E}[N_{\mathrm{eff}}(h_{\mathrm{CVaR}})]}{N} = \varepsilon \to 0,$$

whereas for either Cantelli moment,

$$\frac{\mathbb{E}[N_{\mathrm{eff}}(h_i)]}{N} = f, \qquad i \in \{1, 2\},$$

remains bounded away from zero. Thus, indicator and CVaR estimation become increasingly dominated by rare-event sampling, while the Cantelli moment estimators continue to extract information from a non-vanishing fraction of the batch.

This distinction disappears when the environmental cost signal itself is sparse. If $f \leq \varepsilon$, the CVaR quantile satisfies $\eta^* = 0$. In this regime the CVaR active event coincides with the positive-cost support, and its expected effective sample size becomes fN, matching that of the Cantelli moment estimators. The effective sample size advantage considered in this appendix is therefore specific to dense-cost regimes, which is the requirement ($f > \varepsilon$) noted in Section 5 and stated as a limitation in Section 7.

From trajectory statistics to constraint values. The expected effective sample size alone does not determine relative variance: the distribution of the active contributions may itself change with p, ε, or f, and the constraint value is not the statistic $h(\tau)$ itself but a quantity built from it. The object a constrained optimizer consumes is the constraint value $J_C(\theta)$ and its policy gradient, so we now state the two results the per-method derivations instantiate in terms of $J_C(\theta)$ directly, and make explicit where each assumption of Section 5 enters.

Every constraint value below decomposes as

$$J_C(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[x(\tau)] + b, \tag{35}$$

where $x(\tau)$ is a *sampled statistic* of a single trajectory and b is an *offset* that does not depend on the sampled trajectories $\tau_1, \ldots, \tau_N$, such as the threshold $-\varepsilon$ in the indicator constraint or the cost limit $-\ell$ in the mean constraint. The sampled statistic is built from the method's trajectory-level statistic $h(\tau)$:

$$x_{\mathrm{Ind}}(\tau) = h_{\mathrm{Ind}}(\tau) \tag{36}$$

$$x_{\mathrm{CVaR}}(\tau) = \frac{h_{\mathrm{CVaR}}(\tau)}{\varepsilon} \tag{37}$$

$$x_{\mathrm{Cantelli}}(\tau) = c_1 h_1(\tau) + c_2 h_2(\tau) \quad \text{(Appendix B.3).} \tag{38}$$

In every case, $x(\tau) = 0$ whenever $h(\tau) = 0$, so $x(\tau)$ vanishes outside the active event $\mathcal{E}_h$ and has the same active probability

$$q := \mathbb{P}(\mathcal{E}_h) \in \{p, \varepsilon, f\}. \tag{39}$$

Let

$$m_a := \mathbb{E}[x(\tau) \mid \mathcal{E}_h], \qquad S_a := \mathbb{E}[x(\tau)^2 \mid \mathcal{E}_h], \qquad R_x := \frac{S_a}{m_a^2} \tag{40}$$

denote the conditional first and second moments of the sampled statistic and their *conditional ratio*, the relative second moment of a single active contribution. By the Law of Total Expectation, since $x(\tau)$ vanishes off $\mathcal{E}_h$,

$$J_C(\theta) = q m_a + b. \tag{41}$$

Define the *signal share*

$$r := \frac{q|m_a|}{|J_C(\theta)|}. \tag{42}$$

Assumption 5.4 bounds r above and below.

For the policy gradient, define the trajectory score

$$s(\tau) := \sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t). \tag{43}$$

The score-function identity (Zhang et al., 2020) applied to Equation (35) gives,

$$\nabla_\theta J_C(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)x(\tau)], \tag{44}$$

so the gradient is the expectation of the sampled statistic $s(\tau)x(\tau)$, which vanishes off the same active event $\mathcal{E}_h$ (whenever $x(\tau) = 0$, the gradient contribution is necessarily zero; we define the active event through $x(\tau)$ rather than $s(\tau)x(\tau)$ to isolate the sparsity induced by the cost statistic from incidental zeros of the score). Let

$$m_g := \mathbb{E}[s(\tau)x(\tau) \mid \mathcal{E}_h], \qquad S_g := \mathbb{E}[\|s(\tau)x(\tau)\|^2 \mid \mathcal{E}_h], \qquad R_g := \frac{S_g}{\|m_g\|^2} \tag{45}$$

denote the corresponding conditional moments and ratio for the gradient, so that $\nabla_\theta J_C(\theta) = qm_g$. The gradient has no offset, and therefore no signal share.

The conditional ratios R_x and R_g are upper bounded by Assumption 5.2 for the indicator and CVaR constraints since $x(\tau)$ is a scalar multiple of $h(\tau)$. For the Cantelli constraint, $x(\tau)$ is a linear combination of two statistics of Assumption 5.2, and bounding R_x and R_g additionally requires Assumption 5.3 (Appendix B.3).

Lemma B.1 (Relative variance of a constraint value)**.** *Let $J_C(\theta)$ decompose as in Equation* (35)*, with the Monte Carlo estimator $\hat{J}_C = \frac{1}{N}\sum_{i=1}^{N} x(\tau_i) + b$ over N independent trajectories. Then*

$$\kappa^2[J_C(\theta)] = r^2\left(\frac{R_x}{qN} - \frac{1}{N}\right), \tag{46}$$

and

$$\frac{(1-q)\,r^2}{qN} \;\le\; \kappa^2[J_C(\theta)] \;\le\; \frac{r^2 R_x}{qN}. \tag{47}$$

If $R_x \le K_x$ for a constant K_x independent of (p, ε, f), and Assumption 5.4 holds, then for q bounded away from one,

$$\kappa^2[J_C(\theta)] = \Theta\left(\frac{1}{qN}\right). \tag{48}$$

Proof. The estimator is unbiased: $\mathbb{E}[\hat{J}_C] = \mathbb{E}[x(\tau)] + b = qm_a + b = J_C(\theta)$ by Equation (41). Starting from the definition of relative variance (33),

$$\begin{aligned}
\kappa^2[J_C(\theta)] &= \frac{\mathrm{Var}(\hat{J}_C)}{J_C(\theta)^2} \\
&= \frac{1}{J_C(\theta)^2}\,\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} x(\tau_i)\right) \qquad (b \text{ does not depend on the batch}) \\
&= \frac{\mathrm{Var}(x(\tau))}{N\,J_C(\theta)^2} \qquad (\text{independence across trajectories}) \\
&= \frac{\mathbb{E}[x(\tau)^2] - \mathbb{E}[x(\tau)]^2}{N\,J_C(\theta)^2} \\
&= \frac{qS_a - q^2 m_a^2}{N\,J_C(\theta)^2} \qquad (\text{Law of Total Expectation; } x(\tau) = 0 \text{ off } \mathcal{E}_h) \\
&= \frac{(qm_a)^2}{J_C(\theta)^2} \cdot \frac{qS_a - q^2 m_a^2}{N(qm_a)^2} \\
&= r^2\left(\frac{R_x}{qN} - \frac{1}{N}\right) \qquad (\text{Equation (42); } R_x = S_a/m_a^2),
\end{aligned} \tag{49}$$

which is Equation (46). From it, the two sides of Equation (47) follow as

$$\begin{aligned}
&\kappa^2[J_C(\theta)] \ge r^2\left(\frac{1}{qN} - \frac{1}{N}\right) = \frac{(1-q)\,r^2}{qN} \qquad (\text{Jensen: } S_a \ge m_a^2, \text{ so } R_x \ge 1), \\
&\kappa^2[J_C(\theta)] \le \frac{r^2 R_x}{qN} \qquad (\text{drop the non-negative } r^2/N) \\
&\qquad\qquad \le \frac{r^2 K_x}{qN} \qquad (R_x \le K_x; \text{ Assumption 5.2 and Assumption 5.3}).
\end{aligned} \tag{50}$$

The lower bound uses no assumption. Under Assumption 5.4, $r_{\min} \le r \le r_{\max}$, so for q bounded away from one,

$$\frac{(1-q)\,r_{\min}^2}{qN} \le \kappa^2[J_C(\theta)] \le \frac{r_{\max}^2 K_x}{qN} \qquad (\text{Assumption 5.4}), \tag{51}$$

which is Equation (48). □

The two sides of Equation (47) have different standing. The lower bound holds for every policy with $r > 0$: whenever the sampled part of the constraint is comparable to the constraint value, an expected effective sample size qN below one puts the estimator below unit signal-to-noise ratio. The upper bound is where the ratio bound $R_x \le K_x$ enters, as the statement that the relative second moment of a single active contribution stays bounded while the active fraction shrinks. For the indicator and CVaR constraints, $x(\tau)$ is a scalar multiple of $h(\tau)$ and Assumption 5.2 supplies it directly; for the Cantelli constraint, $x(\tau)$ is a linear combination of two statistics of Assumption 5.2, and bounding R_x and R_g additionally requires Assumption 5.3 (Appendix B.3). Assumption 5.4 is needed on both sides, and for a reason specific to constraint values: the upper bound $\mathcal{O}(1/(qN))$ can be loose. If the sampled part qm_a vanishes while the offset b holds the signal fixed, then $r \to 0$ and Equation (46) shows the relative variance vanishes rather than diverges; the CVaR constraint value exhibits exactly this behavior for almost surely bounded cost returns (Appendix B.2). Assumption 5.4 also subsumes the boundary margin: $r \le r_{\max}$ implies $|J_C(\theta)| \ge q|m_a|/r_{\max} > 0$.

Lemma B.2 (Relative variance of a constraint gradient)**.** *Let $J_C(\theta)$ decompose as in Equation* (35) *and satisfy the gradient identity Equation* (44)*, $\nabla_\theta J_C(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)x(\tau)]$, which holds whenever the offset b is independent of θ, with the Monte Carlo gradient estimator $\hat{g} = \frac{1}{N}\sum_{i=1}^{N} s(\tau_i)x(\tau_i)$ over N independent trajectories. Then*

$$\kappa^2[\nabla_\theta J_C(\theta)] = \frac{R_g}{qN} - \frac{1}{N}, \tag{52}$$

and

$$\frac{1-q}{qN} \;\le\; \kappa^2[\nabla_\theta J_C(\theta)] \;\le\; \frac{R_g}{qN}. \tag{53}$$

If $R_g \le K_g$ for a constant K_g independent of (p, ε, f), then for q bounded away from one,

$$\kappa^2[\nabla_\theta J_C(\theta)] = \Theta\left(\frac{1}{qN}\right), \tag{54}$$

with constants $1 - q$ and K_g. The lower bound in Equation (53) *holds for every policy.*

Proof. By the gradient identity, $\hat{g}$ is the empirical mean of $Y(\tau) := s(\tau)x(\tau)$ and is unbiased for $\nabla_\theta J_C(\theta) = qm_g$. Starting from the definition of relative variance (33),

$$\begin{aligned}
\kappa^2[\nabla_\theta J_C(\theta)] &= \frac{\mathrm{Var}(\hat{g})}{\|\nabla_\theta J_C(\theta)\|^2} \\
&= \frac{\mathrm{Var}(Y(\tau))}{N\,\|\nabla_\theta J_C(\theta)\|^2} \qquad \text{(independence across trajectories)} \\
&= \frac{\mathbb{E}[\|Y(\tau)\|^2] - \|\mathbb{E}[Y(\tau)]\|^2}{N\,\|\nabla_\theta J_C(\theta)\|^2} \\
&= \frac{qS_g - q^2\|m_g\|^2}{N\,q^2\|m_g\|^2} \qquad \text{(Law of Total Expectation; } Y(\tau) = 0 \text{ off } \mathcal{E}_h\text{; } \nabla_\theta J_C(\theta) = qm_g\text{)} \\
&= \frac{R_g}{qN} - \frac{1}{N} \qquad (R_g = S_g/\|m_g\|^2),
\end{aligned} \tag{55}$$

which is Equation (52). There is no offset and hence no signal share. The two sides of Equation (53) follow as

$$\begin{aligned}
\kappa^2[\nabla_\theta J_C(\theta)] &\ge \frac{1}{qN} - \frac{1}{N} = \frac{1-q}{qN} \qquad \text{(Jensen for } y \mapsto \|y\|^2\text{: } S_g \ge \|m_g\|^2\text{, so } R_g \ge 1\text{)}, \\
\kappa^2[\nabla_\theta J_C(\theta)] &\le \frac{R_g}{qN} \qquad \text{(drop the non-negative } 1/N\text{)} \\
&\le \frac{K_g}{qN} \qquad (R_g \le K_g\text{; Assumption 5.2 and Assumption 5.3}),
\end{aligned} \tag{56}$$

which for q bounded away from one is Equation (54). The lower bound uses no assumption. A scalar factor multiplying both $x(\tau)$ and its estimator, such as the $1/\varepsilon$ of the CVaR constraint, cancels in the relative variance. □

The per-method derivations that follow therefore reduce to three steps: identify the sampled statistic $x(\tau)$, its active probability q, and its offset b; bound the conditional ratios R_x and R_g, by Assumption 5.2 directly or, for the Cantelli constraint, by Assumptions 5.2 and 5.3 together; and, for the constraint value, read off the signal share r and state what Assumption 5.4 requires of the policy. Lemmas B.1 and B.2 then supply the rates. We now derive the relative-variance scaling explicitly, beginning with the indicator estimator.

B.1 Indicator VaR: $N_{\text{eff}} = pN$

The indicator constraint (10) bounds the expectation of the Bernoulli indicator, $\mu_I(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[I(\tau)] = \mathbb{P}(C(\tau) > \ell) = p$, by the violation threshold. The VaR constraint is therefore:

$$J_C(\theta) = \mu_I(\theta) - \varepsilon \leq 0, \tag{57}$$

with Monte Carlo estimator

$$\hat{J}_C = \frac{1}{N}\sum_{i=1}^{N} I(\tau_i) - \varepsilon. \tag{58}$$

This is of the form Equation (35) with sampled statistic $x(\tau) = I(\tau)$, active probability $q = p$, and offset $b = -\varepsilon$, so Lemma B.1 applies. Since the indicator is idempotent ($I(\tau)^2 = I(\tau)$), its conditional moments are $m_a = S_a = 1$, so $R_x = 1$ and the ratio bound holds with $K_x = 1$. The signal is $J_C(\theta) = p - \varepsilon$ (41). Any feasible policy has $p \leq \varepsilon$, so we write $p = c\varepsilon$ with $c \in (0, 1)$ parameterizing the fraction of the threshold the policy uses; the slack is then $\varepsilon - p = \frac{1-c}{c}p$ and the signal share (42) is $r = p/(\varepsilon - p) = c/(1 - c)$. Holding c fixed is Assumption 5.4 for the indicator constraint. Substituting these into Lemma B.1, the relative variance is exactly

$$\kappa^2[J_C(\theta)] = \left(\frac{c}{1-c}\right)^2 \left(\frac{1}{pN} - \frac{1}{N}\right) = \left(\frac{c}{1-c}\right)^2 \frac{1-p}{pN}, \tag{59}$$

which, since $R_x = 1$, sits on the lower edge of the bounds of Lemma B.1,

$$\frac{(1-p)}{pN}\left(\frac{c}{1-c}\right)^2 \leq \kappa^2[J_C(\theta)] \leq \frac{1}{pN}\left(\frac{c}{1-c}\right)^2, \tag{60}$$

and hence, for fixed c and with $1 - \varepsilon \leq 1 - p \leq 1$,

$$\kappa^2[J_C(\theta)] = \Theta\left(\frac{1}{pN}\right). \tag{61}$$

Because $p = c\varepsilon$ ties the violation probability to the threshold, this rate diverges as the violation threshold becomes strict ($\varepsilon \to 0$): the constraint value cannot be evaluated with bounded relative error, and, the expression being exact, no choice of estimator constant can prevent it.

For the gradient, Lemma B.2 applies with the same $x(\tau)$ and $q = p$: the conditional gradient moments are $m_g = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau) \mid I(\tau) = 1]$ and, using $I(\tau)^2 = I(\tau)$, $S_g = \mathbb{E}_{\tau\sim\pi_\theta}[\|s(\tau)\|^2 \mid I(\tau) = 1]$, and Assumption 5.2 with $X = s(\tau)I(\tau)$ supplies the ratio bound $R_g = S_g/\|m_g\|^2 \leq K$. Substituting, the relative variance is exactly

$$\kappa^2[\nabla_\theta J_C(\theta)] = \frac{1}{pN}\left(\frac{S_g}{\|m_g\|^2}\right) - \frac{1}{N}, \tag{62}$$

bounded by

$$\frac{1-p}{pN} \leq \kappa^2[\nabla_\theta J_C(\theta)] \leq \frac{K}{pN}, \tag{63}$$

and hence

$$\kappa^2[\nabla_\theta J_C(\theta)] = \Theta\left(\frac{1}{pN}\right), \tag{64}$$

where the lower bound holds for every policy.

B.2 CVaR: $N_{\text{eff}} = \varepsilon N$

CVaR relies on the Rockafellar-Uryasev dual formulation, where η converges to the $(1-\varepsilon)$-quantile η^*, and the constraint of OP 12 evaluates the expected tail risk

$$F(\theta) = \eta^* + \frac{1}{\varepsilon}\mathbb{E}_{\tau\sim\pi_\theta}\left[(C(\tau)-\eta^*)^+\right] \tag{65}$$

against the cost limit. The CVaR constraint is therefore:

$$J_C(\theta) = F(\theta) - \ell \le 0, \tag{66}$$

with Monte Carlo estimator

$$\hat{J}_C = \eta^* + \frac{1}{\varepsilon N}\sum_{i=1}^{N}(C(\tau_i)-\eta^*)^+ - \ell. \tag{67}$$

This estimator holds η fixed at the population quantile η^*; the practical estimator instead minimizes over η, or equivalently plugs the empirical $(1-\varepsilon)$-quantile $\hat{\eta}$ of the same batch into Equation (11). Because η^* minimizes the Rockafellar-Uryasev objective (11), the objective is stationary in η at η^* (differentiable there under Assumption 5.1, which excludes an atom at η^*), so the perturbation $\hat{\eta} - \eta^* = \mathcal{O}_p(N^{-1/2})$ enters the estimator only at second order: the empirical-quantile estimator is first-order equivalent to the fixed-quantile estimator above, with the same asymptotic variance (Shapiro et al., 2009). The scaling derived below therefore transfers to the practical estimator to leading order in $1/N$.

In the form of Equation (35), the sampled statistic is the scaled tail penalty $x(\tau) = \frac{1}{\varepsilon}(C(\tau)-\eta^*)^+$ and the offset is $b = \eta^* - \ell$. The active event is the tail event $\{C(\tau) > \eta^*\}$. Throughout, we use $q = \mathbb{P}(C(\tau) > \eta^*) = \varepsilon$, which holds under Assumption 5.1: the cost return places no probability mass at η^*, and its zero-cost atom $\mathbb{P}(C(\tau) = 0) = 1 - f$ lies below $\eta^* > 0$. Let $m_{\text{tail}} = \mathbb{E}[(C(\tau)-\eta^*) \mid C(\tau) > \eta^*]$ and $S_{\text{tail}} = \mathbb{E}[(C(\tau)-\eta^*)^2 \mid C(\tau) > \eta^*]$ denote the conditional moments of the unscaled tail penalty, so that $m_a = m_{\text{tail}}/\varepsilon$, $S_a = S_{\text{tail}}/\varepsilon^2$, and the scaling cancels in the conditional ratio $R_x = S_{\text{tail}}/m_{\text{tail}}^2$, which Assumption 5.2 bounds. By Equation (41), $F(\theta) = \eta^* + qm_a = \eta^* + m_{\text{tail}}$: the tail mean m_{tail} is the excess of CVaR over VaR, and the signal is

$$J_C(\theta) = \eta^* + m_{\text{tail}} - \ell, \tag{68}$$

with signal share (42) $r = m_{\text{tail}}/|J_C(\theta)|$. Substituting into Lemma B.1 with $q = \varepsilon$, the relative variance is exactly

$$\kappa^2[J_C(\theta)] = \frac{m_{\text{tail}}^2}{J_C(\theta)^2}\left[\frac{1}{\varepsilon N}\left(\frac{S_{\text{tail}}}{m_{\text{tail}}^2}\right) - \frac{1}{N}\right], \tag{69}$$

bounded by

$$\frac{1-\varepsilon}{\varepsilon N}\cdot\frac{m_{\text{tail}}^2}{J_C(\theta)^2} \le \kappa^2[J_C(\theta)] \le \frac{K}{\varepsilon N}\cdot\frac{m_{\text{tail}}^2}{J_C(\theta)^2}, \tag{70}$$

and hence, under Assumption 5.4,

$$\kappa^2[J_C(\theta)] = \Theta\left(\frac{1}{\varepsilon N}\right). \tag{71}$$

For the CVaR constraint, Assumption 5.4 reads: the slack between the cost limit and the CVaR, $|\ell - F(\theta)|$, is proportional to the excess of CVaR over VaR, m_{tail}. This is the CVaR analogue of the indicator's $p = c\varepsilon$, and it is not vacuous. For a cost return with bounded support, η^* converges to the maximum cost return $\operatorname{ess\,sup} C(\tau)$ and $m_{\text{tail}} \to 0$ as $\varepsilon \to 0$ (for $C(\tau)$ uniform on $[0,1]$, $m_{\text{tail}} = \varepsilon/2$ and $S_{\text{tail}} = \varepsilon^2/3$), so at a fixed slack $|J_C(\theta)| \ge \zeta$ the exact expression (69) gives $\kappa^2[J_C(\theta)] \approx \varepsilon/(3\zeta^2 N) \to 0$: the upper bound $\mathcal{O}(1/(\varepsilon N))$ holds but is loose, and the constraint value is estimated with vanishing relative error because the sampled part of the constraint has itself vanished. The gradient, which has no offset, has no such escape.

For the gradient, the constant cost limit vanishes under differentiation, $\nabla_\theta J_C(\theta) = \nabla_\theta F(\theta)$, and the Envelope Theorem allows η^* to be held fixed while differentiating with respect to the policy parameters θ, so Equation (44) applies:

$$\nabla_\theta J_C(\theta) = \frac{1}{\varepsilon}\mathbb{E}_{\tau\sim\pi_\theta}\left[s(\tau)(C(\tau)-\eta^*)^+\right] = \frac{1}{\varepsilon}\cdot\varepsilon\cdot\mathbb{E}_{\tau\sim\pi_\theta}\left[s(\tau)(C(\tau)-\eta^*) \mid C(\tau) > \eta^*\right] = m, \tag{72}$$

where $m = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)(C(\tau) - \eta^*) \mid C(\tau) > \eta^*]$ is the conditional gradient of the unscaled tail penalty: the $1/\varepsilon$ of the estimator cancels the probability ε of the tail event, and the true gradient has squared magnitude $\|m\|^2$ independent of ε. The Monte Carlo estimator is $\hat{g} = \frac{1}{\varepsilon N}\sum_{i=1}^{N} s(\tau_i)(C(\tau_i) - \eta^*)^+$; because the $(x)^+$ operator evaluates to zero outside the tail event, only an expected εN trajectories contribute to the sum, defining the effective sample size $N_{\text{eff}} = \varepsilon N$. With $S = \mathbb{E}_{\tau\sim\pi_\theta}[\|s(\tau)(C(\tau) - \eta^*)\|^2 \mid C(\tau) > \eta^*]$, the scaling cancels in $R_g = S/\|m\|^2$, which Assumption 5.2 with $X = s(\tau)(C(\tau) - \eta^*)^+$ bounds, so substituting into Lemma B.2 with $q = \varepsilon$, the relative variance is exactly

$$\kappa^2[\nabla_\theta J_C(\theta)] = \frac{1}{\varepsilon N}\left(\frac{S}{\|m\|^2}\right) - \frac{1}{N}, \tag{73}$$

bounded by

$$\frac{1-\varepsilon}{\varepsilon N} \le \kappa^2[\nabla_\theta J_C(\theta)] \le \frac{K}{\varepsilon N}, \tag{74}$$

and hence

$$\kappa^2[\nabla_\theta J_C(\theta)] = \Theta\left(\frac{1}{\varepsilon N}\right), \tag{75}$$

where the lower bound holds for every policy, in particular for the bounded-support cost returns under which the constraint value escapes the rate above.

B.3 CANTELLI: $N_{\text{eff}} = fN$

Canary enforces the constraint pair of OP 15: the Cantelli constraint $B(\theta) \le 0$ and the mean constraint $\mu(\theta) - \ell \le 0$. We derive the relative variance of each in turn, in both cases treating the constraint value first and the policy gradient second.

B.3.1 CANTELLI CONSTRAINT

The Canary formulation avoids algorithmic thresholding operators, defining the Cantelli constraint $B(\theta)$ as in Equation 21, which is already in canonical form, $J_C(\theta) \le 0$, with

$$J_C(\theta) = B(\theta) = \left(\frac{1}{\varepsilon} - 1\right)\tilde{\mu}(\theta) + 2\ell\mu(\theta) - \frac{1}{\varepsilon}\mu(\theta)^2 - \ell^2. \tag{76}$$

Write $\mu_i(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[C(\tau)^i]$ for the i-th moment of the cost return, $i \in \{1, 2\}$, so that $\mu_1(\theta) = \mu(\theta)$ and $\mu_2(\theta) = \tilde{\mu}(\theta)$. The Monte Carlo estimators for these moments over a batch of N trajectories are the empirical means

$$\hat{\mu}_i = \frac{1}{N}\sum_{j=1}^{N} C(\tau_j)^i. \tag{77}$$

The constraint value is a smooth function of the two moments, with sensitivities

$$c_1 = \frac{\partial B}{\partial \mu_1} = 2\left(\ell - \frac{\mu_1(\theta)}{\varepsilon}\right), \qquad c_2 = \frac{\partial B}{\partial \mu_2} = \frac{1}{\varepsilon} - 1, \tag{78}$$

evaluated at the current θ, in terms of which it can be written exactly as

$$J_C(\theta) = c_1\mu_1(\theta) + c_2\mu_2(\theta) + b, \qquad b = J_C(\theta) - c_1\mu_1(\theta) - c_2\mu_2(\theta) = \frac{\mu(\theta)^2}{\varepsilon} - \ell^2. \tag{79}$$

Replacing the moments by their empirical means gives the unbiased Monte Carlo estimator of the constraint value,

$$\hat{J}_C = c_1\hat{\mu}_1 + c_2\hat{\mu}_2 + b. \tag{80}$$

$\hat{J}_C$ is unbiased but not computable, since c_1 and b involve the true $\mu(\theta)$; the computable, biased plug-in estimator that the algorithm evaluates is analyzed once the rate for $\hat{J}_C$ is established.

From Equation (79), the sampled statistic for $J_C(\theta)$ is given by

$$x(\tau) = c_1 C(\tau) + c_2 C(\tau)^2. \tag{81}$$

The offset b of Equation (79), which is built from the true moments at the current θ rather than from the sampled trajectories, is an offset in the sense of Equation (35): it depends on θ, but not on the batch. The statistic $x(\tau)$ vanishes off the active event $\{C(\tau) > 0\}$, which occurs with probability $q = f$, so the Cantelli constraint is in canonical form and can be substituted into Lemma B.1.

Let $\mu_{k,a} = \mathbb{E}_{\tau\sim\pi_\theta}[C(\tau)^k \mid C(\tau) > 0]$ for $k \in \{1, 2, 3, 4\}$ denote the cost moments conditional on the active trajectories, so that $\mu_i(\theta) = f\mu_{i,a}$ (moments up to order four appear because $x(\tau)^2$ involves $C(\tau)^4$). The conditional moments of the sampled statistic are

$$m_a = c_1\mu_{1,a} + c_2\mu_{2,a}, \qquad S_a = c_1^2\mu_{2,a} + 2c_1c_2\mu_{3,a} + c_2^2\mu_{4,a}, \tag{82}$$

so the sampled part of the constraint value is $\mathbb{E}[x(\tau)] = fm_a = c_1\mu_1(\theta) + c_2\mu_2(\theta)$ and the signal share (42) is

$$r = \frac{|c_1\mu_1(\theta) + c_2\mu_2(\theta)|}{|J_C(\theta)|}. \tag{83}$$

Lemma B.1 requires the conditional ratio $R_x = S_a/m_a^2$, which for the linear combination $x(\tau)$ is not bounded by Assumption 5.2 alone: the numerator contains the cross moment $\mu_{3,a}$, and the denominator is a squared linear combination that cancellation between its two terms can drive toward zero. Bounding the numerator from above and the denominator from below,

$$\begin{aligned}
R_x &= \frac{c_1^2\mu_{2,a} + 2c_1c_2\mu_{3,a} + c_2^2\mu_{4,a}}{\left(c_1\mu_{1,a} + c_2\mu_{2,a}\right)^2} \\
&\le \frac{\left(|c_1|\sqrt{\mu_{2,a}} + |c_2|\sqrt{\mu_{4,a}}\right)^2}{\left(c_1\mu_{1,a} + c_2\mu_{2,a}\right)^2} \qquad \text{(Cauchy–Schwarz: } \mu_{3,a} \le \sqrt{\mu_{2,a}\,\mu_{4,a}}\text{)} \\
&\le \frac{2\left(c_1^2\mu_{2,a} + c_2^2\mu_{4,a}\right)}{\left(c_1\mu_{1,a} + c_2\mu_{2,a}\right)^2} \qquad ((a+b)^2 \le 2(a^2+b^2)) \\
&\le \frac{2K\left(c_1^2\mu_{1,a}^2 + c_2^2\mu_{2,a}^2\right)}{\left(c_1\mu_{1,a} + c_2\mu_{2,a}\right)^2} \qquad \text{(Assumption 5.2: } \mu_{2,a} \le K\mu_{1,a}^2,\ \mu_{4,a} \le K\mu_{2,a}^2\text{)} \\
&\le \frac{2K\left(c_1^2\mu_{1,a}^2 + c_2^2\mu_{2,a}^2\right)}{\nu\left(c_1^2\mu_{1,a}^2 + c_2^2\mu_{2,a}^2\right)} \qquad \text{(Assumption 5.3, moment pair)} \\
&= \frac{2K}{\nu} =: K_x.
\end{aligned} \tag{84}$$

In the last inequality, Assumption 5.3 is stated for the unconditioned moments $\mu_i(\theta) = f\mu_{i,a}$, and the factor f^2 cancels on both sides. For this scalar moment pair, cancellation requires $c_1 < 0$, since $c_2 = \frac{1}{\varepsilon} - 1 > 0$ and both conditional moments are positive; that is, it requires $\mu(\theta) > \varepsilon\ell$, and then additionally $|c_1|\mu_{1,a} \approx c_2\mu_{2,a}$. Substituting into Lemma B.1 with $q = f$, signal share r of Equation (83), and $R_x \le K_x$, the relative variance of the constraint evaluation is bounded by

$$\frac{\omega\, r^2}{fN} \le \kappa^2[J_C(\theta)] \le \frac{r^2}{fN} \cdot \frac{2K}{\nu}, \tag{85}$$

where the lower bound follows from Equation (46) and the non-degeneracy clause $R_x \ge 1 + \omega$ of Assumption 5.4, since $R_x - f \ge \omega$ for every $f \in (0, 1]$, and hence, with Assumption 5.4 bounding r above and below,

$$\kappa^2[J_C(\theta)] = \Theta\left(\frac{1}{fN}\right). \tag{86}$$

The bound above is uniform in f, but Assumption 5.4 requires the signal share r of Equation (83) to be uniform in ε as well. Expanding the sensitivities, $c_1\mu_1(\theta) + c_2\mu_2(\theta) = J_C(\theta) + \ell^2 - \mu(\theta)^2/\varepsilon$, so

$$r = \left|1 + \frac{\ell^2 - \mu(\theta)^2/\varepsilon}{J_C(\theta)}\right|. \tag{87}$$

By the identity $J_C(\theta) = \left(\frac{1}{\varepsilon} - 1\right)\sigma^2(\theta) - (\mu(\theta) - \ell)^2$, for policies whose cost variance is bounded below, $\sigma^2(\theta) \ge \sigma^2_{\min} > 0$, both $J_C(\theta)$ and the correction $\ell^2 - \mu(\theta)^2/\varepsilon$ grow at the rate $1/\varepsilon$, and $r \to |\sigma^2(\theta) - \mu(\theta)^2|/\sigma^2(\theta)$ as $\varepsilon \to 0$. This limit is bounded above by $1 + \mu(\theta)^2/\sigma^2_{\min}$, and bounded away from zero unless the cost variance equals the squared mean cost, $\tilde{\mu}(\theta) = 2\mu(\theta)^2$, a knife-edge identity between the two moments. The gradient rate derived later needs no such condition.

Constraint value plug-in estimator. The algorithm does not evaluate $\hat{J}_C$ itself, but the biased plug-in estimator in which $\hat{\mu}_1$ also replaces $\mu(\theta)$ inside c_1 and b:

$$\hat{B} = \beta\hat{\mu}_2 + 2\ell\hat{\mu}_1 - \frac{\hat{\mu}_1^2}{\varepsilon} - \ell^2. \tag{88}$$

$\hat{J}_C$ is the first-order surrogate (the Delta Method) of $\hat{B}$, and the two differ by the exact remainder

$$\hat{B} - \hat{J}_C = -\frac{(\hat{\mu}_1 - \mu_1(\theta))^2}{\varepsilon}, \tag{89}$$

whose mean is $-\operatorname{Var}(\hat{\mu}_1)/\varepsilon = -\sigma^2(\theta)/(\varepsilon N)$: the absolute order is $\mathcal{O}(1/(\varepsilon N))$, not $\mathcal{O}(1/N)$, and the factor $1/\varepsilon$ cannot be discarded in the regime $\varepsilon \to 0$ this appendix concerns. The remainder is nevertheless second order relative to the signal, which carries the same factor: Assumption 5.4 gives $|J_C(\theta)| \geq |c_1\mu_1(\theta) + c_2\mu_2(\theta)|/r_{\max}$, Assumption 5.3 gives $|c_1\mu_1(\theta) + c_2\mu_2(\theta)| \geq \sqrt{\nu}\, c_2\mu_2(\theta)$, and $\sigma^2(\theta) \leq \mu_2(\theta)$, so with $c_2 = (1-\varepsilon)/\varepsilon$,

$$\frac{\left|\mathbb{E}[\hat{B} - \hat{J}_C]\right|}{|J_C(\theta)|} \leq \frac{\sigma^2(\theta)}{\varepsilon N} \cdot \frac{r_{\max}\,\varepsilon}{(1-\varepsilon)\sqrt{\nu}\,\mu_2(\theta)} \leq \frac{r_{\max}}{(1-\varepsilon)\sqrt{\nu}\,N}, \tag{90}$$

uniformly in (p, ε, f) for ε bounded away from one. For the variance of the remainder, $\operatorname{Var}((\hat{\mu}_1 - \mu_1(\theta))^2)/\varepsilon^2 \leq \mathbb{E}[(\hat{\mu}_1 - \mu_1(\theta))^4]/\varepsilon^2$, and for an empirical mean $\mathbb{E}[(\hat{\mu}_1 - \mu_1(\theta))^4] \leq 3\sigma^4(\theta)/N^2 + \mathbb{E}[(C(\tau) - \mu_1(\theta))^4]/N^3$, where the fourth moment is bounded through Assumption 5.2 ($\mu_{4,a} \leq K\mu_{2,a}^2$, so $\mathbb{E}[C(\tau)^4] \leq K\mu_2(\theta)^2/f$). Normalizing by $J_C(\theta)^2 \geq \nu c_2^2\mu_2(\theta)^2/r_{\max}^2$ as above and using $\sigma^4(\theta) \leq \mu_2(\theta)^2$, the remainder contributes to the relative variance at order $1/N^2 + 1/(fN^3)$, smaller than the $\Theta(1/(fN))$ rate above by at least a factor of $1/N$, uniformly in (p, ε, f). The cross term between the sampling error of $\hat{J}_C$ and the remainder is bounded by Cauchy–Schwarz at the geometric mean of the two contributions, of relative order $1/(\sqrt{f}N^{3/2})$, and the squared bias is of relative order $1/N^2$ by the mean bound above; both are likewise subleading. The constraint-value rate (85) therefore holds for $\hat{B}$ to leading order.

The Cantelli constraint policy gradient is

$$\nabla_\theta J_C(\theta) = c_1\nabla_\theta\mu_1(\theta) + c_2\nabla_\theta\mu_2(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[s(\tau)\left(c_1C(\tau) + c_2C(\tau)^2\right)\right] = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)x(\tau)], \tag{91}$$

where $\nabla_\theta\mu_i(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)C(\tau)^i] =: m_i(\theta)$, and $x(\tau)$ is the sampled statistic of Equation (81). The offset b of Equation (79) depends on θ, so the sufficient condition of Lemma B.2 fails, but the gradient identity Equation (44) it is used for holds regardless: differentiating the decomposition $J_C(\theta) = c_1\mu_1(\theta) + c_2\mu_2(\theta) + b$ produces, beyond $c_1\nabla_\theta\mu_1(\theta) + c_2\nabla_\theta\mu_2(\theta)$, the two terms $\mu_1(\theta)\nabla_\theta c_1 = -\frac{2\mu_1(\theta)}{\varepsilon}\nabla_\theta\mu_1(\theta)$ and $\nabla_\theta b = +\frac{2\mu_1(\theta)}{\varepsilon}\nabla_\theta\mu_1(\theta)$, which cancel exactly, recovering Equation (91). Lemma B.2 therefore applies exactly.

Let $m_{i,a} = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)C(\tau)^i \mid C(\tau) > 0]$ and $S_{i,a} = \mathbb{E}_{\tau\sim\pi_\theta}[\|s(\tau)C(\tau)^i\|^2 \mid C(\tau) > 0]$ denote the moment gradients and their second moments conditional on the active trajectories, so that $m_i(\theta) = fm_{i,a}$, with empirical moment gradients $\hat{m}_i = \frac{1}{N}\sum_{j=1}^N s(\tau_j)C(\tau_j)^i$ averaging over all N trajectories including the exact zeros.

The Monte Carlo estimator of the constraint gradient is the corresponding linear combination of the empirical moment gradients, $\hat{g} = c_1\hat{m}_1 + c_2\hat{m}_2$, with the sensitivity c_1 evaluated at the true moment $\mu_1(\theta)$. Like $\hat{J}_C$ (80), this fixed-coefficient estimator is not computable; the practical estimator plugs $\hat{\mu}_1$ into c_1. We analyze the fixed-coefficient estimator first and bound the plug-in perturbation, which unlike the value remainder (89) is not higher order, at the end of this derivation. By linearity, $\hat{g} = \frac{1}{N}\sum_{j=1}^N s(\tau_j)x(\tau_j)$ exactly, the gradient estimator of Lemma B.2 for the sampled statistic $x(\tau)$ of Equation (81), with conditional gradient moments $m_g = c_1m_{1,a} + c_2m_{2,a}$ and $S_g = c_1^2S_{1,a} + 2c_1c_2S_{12,a} + c_2^2S_{2,a}$, where the conditional cross moment is $S_{12,a} = \mathbb{E}_{\tau\sim\pi_\theta}[\langle s(\tau)C(\tau), s(\tau)C(\tau)^2\rangle \mid C(\tau) > 0] = \mathbb{E}_{\tau\sim\pi_\theta}[\|s(\tau)\|^2C(\tau)^3 \mid C(\tau) > 0]$. Lemma B.2 requires the conditional ratio $R_g = S_g/\|m_g\|^2$, bounded exactly as R_x in Equation (84), with Assumption 5.3 now applied to the

moment-gradient pair:

$$\begin{aligned} R_g &= \frac{c_1^2 S_{1,a} + 2c_1c_2 S_{12,a} + c_2^2 S_{2,a}}{\|c_1 m_{1,a} + c_2 m_{2,a}\|^2} \\ &\le \frac{\left(|c_1|\sqrt{S_{1,a}} + |c_2|\sqrt{S_{2,a}}\right)^2}{\|c_1 m_{1,a} + c_2 m_{2,a}\|^2} \qquad \text{(Cauchy–Schwarz: } S_{12,a} \le \sqrt{S_{1,a}\, S_{2,a}}\text{)} \\ &\le \frac{2\left(c_1^2 S_{1,a} + c_2^2 S_{2,a}\right)}{\|c_1 m_{1,a} + c_2 m_{2,a}\|^2} \qquad ((a+b)^2 \le 2(a^2+b^2)) \\ &\le \frac{2K\left(c_1^2\|m_{1,a}\|^2 + c_2^2\|m_{2,a}\|^2\right)}{\|c_1 m_{1,a} + c_2 m_{2,a}\|^2} \qquad \text{(Assumption 5.2: } S_{i,a} \le K\|m_{i,a}\|^2\text{)} \\ &\le \frac{2K\left(c_1^2\|m_{1,a}\|^2 + c_2^2\|m_{2,a}\|^2\right)}{\nu\left(c_1^2\|m_{1,a}\|^2 + c_2^2\|m_{2,a}\|^2\right)} \qquad \text{(Assumption 5.3, moment-gradient pair)} \\ &= \frac{2K}{\nu} =: K_g. \end{aligned} \tag{92}$$

As before, Assumption 5.3 is stated for the unconditioned gradients $m_i(\theta) = f m_{i,a}$ and the factor f^2 cancels on both sides. Substituting into Lemma B.2 with $q = f$ and $R_g \le K_g$, the relative variance of the gradient is exactly bounded by

$$\frac{\omega}{fN} \le \kappa^2[\nabla_\theta J_C(\theta)] \le \frac{1}{fN} \cdot \frac{2K}{\nu}, \tag{93}$$

where the lower bound follows from the exact form in Lemma B.2 and $R_g \ge 1 + \omega$ (Assumption 5.4), and hence

$$\kappa^2[\nabla_\theta J_C(\theta)] = \Theta\left(\frac{1}{fN}\right). \tag{94}$$

Policy gradient plug-in estimator. The practical gradient estimator plugs $\hat{\mu}_1$ into the sensitivity,

$$\hat{g}_B = \hat{c}_1 \hat{m}_1 + c_2 \hat{m}_2, \qquad \hat{c}_1 = 2\left(\ell - \frac{\hat{\mu}_1}{\varepsilon}\right), \tag{95}$$

and differs from the fixed-coefficient estimator $\hat{g}$ analyzed above by the exact perturbation

$$\hat{g}_B - \hat{g} = -\frac{2}{\varepsilon}\left(\hat{\mu}_1 - \mu_1(\theta)\right)\hat{m}_1. \tag{96}$$

Unlike the value remainder (89), which is quadratic in $\hat{\mu}_1 - \mu_1(\theta)$, this perturbation is linear in it, of order $\mathcal{O}_p(1/(\varepsilon\sqrt{N}))$ — the same order as the Monte Carlo error of $\hat{g}$ itself, whose coefficients also carry the factor $1/\varepsilon$. It is therefore a first-order term and must be bounded against the signal. Decomposing $\hat{m}_1 = m_1(\theta) + (\hat{m}_1 - m_1(\theta))$,

$$\mathbb{E}\left[\|\hat{g}_B - \hat{g}\|^2\right] \le \frac{8}{\varepsilon^2}\operatorname{Var}(\hat{\mu}_1)\,\|m_1(\theta)\|^2 + \frac{8}{\varepsilon^2}\,\mathbb{E}\left[\left(\hat{\mu}_1 - \mu_1(\theta)\right)^2 \|\hat{m}_1 - m_1(\theta)\|^2\right], \tag{97}$$

where the second term is of relative order $1/N^2$ under fourth-moment analogues of Assumption 5.2 for $C(\tau)$ and $s(\tau)C(\tau)$. For the first term, $\operatorname{Var}(\hat{\mu}_1) \le f\mu_{2,a}/N$ and $\|m_1(\theta)\|^2 = f^2\|m_{1,a}\|^2$, while Assumption 5.3 bounds the signal from below through the c_2 component alone, $\|\nabla_\theta J_C(\theta)\|^2 \ge \nu c_2^2 \|m_2(\theta)\|^2 = \nu c_2^2 f^2 \|m_{2,a}\|^2$ with $c_2 = (1-\varepsilon)/\varepsilon$; the c_1 component cannot serve here, since c_1 passes through zero at $\mu(\theta) = \varepsilon\ell$ while the perturbation does not vanish. Provided the conditional cross-moment condition

$$\mu_{2,a}\|m_{1,a}\|^2 \le K_2\|m_{2,a}\|^2 \tag{98}$$

holds for a constant K_2 independent of (p, ε, f) — the second-moment gradient carries at least the conditional cost scale of the first, the analogue for gradients of $\mu_{2,a} \ge \mu_{1,a}^2$ — the relative contribution of the first term is

$$\frac{8f^3\mu_{2,a}\|m_{1,a}\|^2/(\varepsilon^2 N)}{\nu(1-\varepsilon)^2 f^2\|m_{2,a}\|^2/\varepsilon^2} \le \frac{8K_2}{\nu(1-\varepsilon)^2} \cdot \frac{f}{N}, \tag{99}$$

of order $f/N \leq 1/(fN)$: the perturbation does not raise the $\mathcal{O}(1/(fN))$ upper bound. The plug-in estimator is biased, $\mathbb{E}[\hat{g}_B - \hat{g}] = -\frac{2}{\varepsilon}\operatorname{Cov}(\hat{\mu}_1, \hat{m}_1)$, but only at relative order $1/N$ by Cauchy–Schwarz together with Assumption 5.2 and Equation (98), so the squared bias is subleading and relative variance remains the fidelity measure. Since $\operatorname{Var}(\hat{g}_B) \leq 2\operatorname{Var}(\hat{g}) + 2\,\mathbb{E}[\|\hat{g}_B - \hat{g}\|^2]$,

$$\frac{\operatorname{Var}(\hat{g}_B)}{\|\nabla_\theta J_C(\theta)\|^2} \leq 2\,\kappa^2[\nabla_\theta J_C(\theta)] + \frac{2\,\mathbb{E}\left[\|\hat{g}_B - \hat{g}\|^2\right]}{\|\nabla_\theta J_C(\theta)\|^2} = \mathcal{O}\left(\frac{1}{fN}\right): \tag{100}$$

the relative variance of the practical gradient estimator obeys the same $\mathcal{O}(1/(fN))$ upper bound. The perturbation (96) is correlated with the sampling error of $\hat{g}$, so it can in principle also lower the variance; the two-sided $\Theta(1/(fN))$ statement of Proposition 5.5 is for the unbiased fixed-coefficient estimators $\hat{J}_C$ and $\hat{g}$, and the plug-in estimators $\hat{B}$ and $\hat{g}_B$ satisfy the $\mathcal{O}(1/(fN))$ upper bound in relative variance.

When the environmental cost signal is dense (e.g., energy consumption, continuous proximity penalties), essentially every trajectory incurs some cost, yielding $f \approx 1$ and $N_{\text{eff}} \approx N$. Here, the relative variances of both the constraint value and its gradient remain bounded by $\mathcal{O}(1/N)$ with constants independent of the violation threshold ε. This resolves the divergence of the baseline methods in feasible, dense-cost regimes.

B.3.2 Mean Constraint

The mean constraint is already in canonical form,

$$J_C(\theta) = \mu(\theta) - \ell \leq 0, \tag{101}$$

with Monte Carlo estimator $\hat{J}_C = \hat{\mu}_1 - \ell$ and $\hat{\mu}_1$ as in Equation (77).

This is of the form Equation (35) with sampled statistic $x(\tau) = C(\tau)$, $q = f$, and offset $b = -\ell$, so Lemma B.1 applies. The sampled part is $\mu(\theta) = f\mu_{1,a}$ (41) and the signal is $J_C(\theta) = \mu(\theta) - \ell$, so the signal share (42) is $r = \mu(\theta)/|\mu(\theta) - \ell|$, and the conditional ratio is $R_x = \mu_{2,a}/\mu_{1,a}^2$, bounded by Assumption 5.2. Substituting into Lemma B.1 with $q = f$, the relative variance is exactly

$$\kappa^2[J_C(\theta)] = r^2\left[\frac{1}{fN}\left(\frac{\mu_{2,a}}{\mu_{1,a}^2}\right) - \frac{1}{N}\right], \tag{102}$$

bounded by

$$\frac{\omega\, r^2}{fN} \leq \kappa^2[J_C(\theta)] \leq \frac{r^2 K}{fN}, \tag{103}$$

using $R_x \geq 1 + \omega$ (Assumption 5.4) for the lower bound, and hence, under Assumption 5.4,

$$\kappa^2[J_C(\theta)] = \Theta\left(\frac{1}{fN}\right). \tag{104}$$

The structure mirrors the Cantelli constraint value (85) with the single moment in place of the linear combination $x(\tau)$, so no analogue of Assumption 5.3 is needed. For the mean constraint, Assumption 5.4 reads: the policy uses a fixed fraction of the cost limit, $\mu(\theta) = c\ell$ with $c \in (0, 1)$ fixed, giving $r = c/(1 - c)$ exactly as for the indicator constraint.

For the policy gradient, $\nabla_\theta J_C(\theta) = \nabla_\theta \mu(\theta) = m_1(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[s(\tau)C(\tau)]$, with no sensitivities and no composition. The offset $b = -\ell$ is constant in θ, so Lemma B.2 applies with sampled statistic $x(\tau) = C(\tau)$, $q = f$, conditional gradient moments $m_g = m_{1,a}$ and $S_g = S_{1,a}$, and estimator $\hat{m}_1$: the relative variance is exactly

$$\kappa^2[\nabla_\theta J_C(\theta)] = \frac{1}{fN}\left(\frac{S_{1,a}}{\|m_{1,a}\|^2}\right) - \frac{1}{N}, \tag{105}$$

bounded by

$$\frac{\omega}{fN} \leq \kappa^2[\nabla_\theta J_C(\theta)] \leq \frac{K}{fN}, \tag{106}$$

and hence

$$\kappa^2[\nabla_\theta J_C(\theta)] = \Theta\left(\frac{1}{fN}\right), \tag{107}$$

which is Lemma B.2 with $R_g = S_{1,a}/\|m_{1,a}\|^2$ bounded above by Assumption 5.2 and below by Assumption 5.4.

C CANARY ALGORITHM

Algorithm 1 Canary

Input: Initial policy π_{θ_0}, value function estimates $V_\phi, V_\psi^C, V_\chi^{\tilde{C}}$, cost limit ℓ, violation threshold ε, KL-divergence limit δ.
Initialize: $\beta \leftarrow \frac{1}{\varepsilon} - 1$
for $k = 0, 1, 2, \ldots$ **do**
1. Data Collection:
Sample trajectories $\mathcal{D} = \{\tau_i\}$ using policy π_{θ_k}.
Compute augmented state $x(s_t, y_t, \gamma_c^t)$ (20).
2. Advantage & Return Estimation:
Value function estimates $V_\phi, V_\psi^C, V_\chi^{\tilde{C}}$ are used to estimate advantages $A_{\theta_k}, A_{\theta_k}^C, A_{\theta_k}^{\tilde{C}}$ respectively, employing the Generalized Advantage Estimation (GAE) (Schulman et al., 2018).
Estimate expected cost returns $\mu(\theta_k)$ and augmented cost returns $\tilde{\mu}(\theta_k)$ using a Monte Carlo (MC) form with $\mathcal{D}$.
3. Constraint Construction:
Calculate policy-dependent bound $\rho(\theta_k)$.
Set Cantelli constraint (c_B, b_B) via (29), (30).
Set mean constraint (c_μ, b_μ) via (31).
4. Policy Update:
Compute objective gradient $g \leftarrow \nabla_\theta L(\theta)|_{\theta_k}$.
Compute reachable minima m_μ, m_B (27) in closed form (Appendix C.1).
Select the tier of (28) by reachability and solve it by active-set enumeration (Appendix C.1).
5. Critic Update:
Update V_ϕ, V_ψ^C and $V_\chi^{\tilde{C}}$ by minimizing MSE against return targets.
end for

The value function estimates here are parameterized models; however, the Canary algorithm is agnostic to the estimate used.

C.1 CLOSED-FORM SOLVER

In step coordinates $s = \theta - \theta_k$, the Tier a update (32) is

$$\max_s \; g^\top s \quad \text{s.t.} \;\; c_B + b_B^\top s \le 0, \;\; c_\mu + b_\mu^\top s \le 0, \;\; \tfrac{1}{2} s^\top H s \le \delta.$$

Its solution lies in the span of $H^{-1}g$, $H^{-1}b_B$, and $H^{-1}b_\mu$, so after three conjugate-gradient solves the problem reduces to scalar algebra on

$$q = g^\top H^{-1} g, \qquad r = \begin{pmatrix} b_B^\top H^{-1} g \\ b_\mu^\top H^{-1} g \end{pmatrix}, \qquad S = \begin{pmatrix} b_B^\top H^{-1} b_B & b_B^\top H^{-1} b_\mu \\ b_\mu^\top H^{-1} b_B & b_\mu^\top H^{-1} b_\mu \end{pmatrix}.$$

Reachability. The minimum of a linear function $c' + b'^\top s$ over the trust region alone is $c' - \sqrt{2\delta\, b'^\top H^{-1} b'}$, which gives m_μ directly from (c_μ, b_μ). m_B is the minimum of $c_B + b_B^\top s$ over the trust region intersected with the half-space $c_\mu + b_\mu^\top s \le 0$; this is a linear objective under one linear constraint and the trust region, solved in closed form by the same one-active-constraint case analysis as below.

Active-set enumeration. With a linear objective the trust region is active at any optimum. Writing $\nu = (\nu_1, \nu_2) \ge 0$ for the multipliers of the linearized Cantelli and mean constraints and $\lambda > 0$ for the trust region, stationarity gives $s^* = \frac{1}{\lambda} H^{-1}(g - \nu_1 b_B - \nu_2 b_\mu)$, and the exact solution is found by enumerating which of the two bounds is active:

- *Neither*: $s^* = \sqrt{2\delta/q}\, H^{-1} g$, the natural gradient step.
- *One*: the analytic CPO dual with the corresponding pair (c', b') (Achiam et al., 2017).

- *Both*: tightness of the two bounds is linear in the multipliers, $\nu(\lambda) = S^{-1}(r + \lambda \mathbf{c})$ with $\mathbf{c} = (c_B, c_\mu)^\top$, and substituting into $\frac{1}{2} s^{*\top} H s^* = \delta$ the cross terms cancel, leaving

$$\lambda^* = \sqrt{\frac{q - r^\top S^{-1} r}{2\delta - \mathbf{c}^\top S^{-1} \mathbf{c}}}, \qquad \nu^* = S^{-1}(r + \lambda^* \mathbf{c}).$$

  Here $q - r^\top S^{-1} r \geq 0$ is the squared H^{-1}-norm of the component of g orthogonal to the two constraint gradients, and $2\delta > \mathbf{c}^\top S^{-1} \mathbf{c}$ is precisely the condition that both tightened constraints intersect inside the trust region (otherwise the case is skipped).

A candidate is retained only if it is primal feasible with non-negative multipliers; among retained candidates the one with the largest objective value is the exact solution. If S is near-singular (b_B and b_μ close to parallel), the mean constraint is directionally redundant and is dropped. The recovery tiers of (28) belong to the same family: Tier b minimizes the linear objective $b_B^\top s$ under the mean constraint and the trust region (the one-active-constraint case), and Tier c is the unconstrained trust-region minimizer $s^* = -\sqrt{2\delta / S_{22}}\, H^{-1} b_\mu$.

D Experimental Details

D.1 Environments

Although the Safety Gymnasium package provides an excellent set of benchmark environments for CMDPs (Ji et al., 2023), we leverage the benefits of running Just-in-Time (JIT) compiled JAX code end-to-end on the GPU for accelerated experiments (Bradbury et al., 2018). This required translating well-known benchmarks: **EcoAnt**, a reimplementation of the Ant environment (Schulman et al., 2018) in MJX, where the agent must maximize forward velocity while managing a limited battery budget ($c_t = 0.5||a_t||_2^2$); **PointGoal**, a reimplementation of SafetyPointGoal1 from Safety Gymnasium (Ji et al., 2023) built on top of MJX, where the agent must navigate toward a randomized target location while avoiding static hazardous areas distributed across the arena; and **IcyLake**, a reimplementation of the gymnasium FrozenLake environment (Towers et al., 2024) in JAX, where the agent must traverse a 4×4 frozen lake grid to reach the goal state while avoiding icy tiles which accumulate cost $+1$ per visit.

D.2 Canary Hyperparameters

Table 2: Canary Hyperparameter Settings

Hyperparameter	**Value**
Hidden Layers	2
Hidden Units	256
Activation	tanh
Learning Rate	3×10^{-4}
Optimizer	Adam
Grad Norm Clip	0.5
GAE λ	0.95
Reward Discount γ	0.99
Cost Discount γ_c	1.0
Trust Region δ	0.01
Critic epochs	80
Base Seed	0

To assess performance, trajectories are drawn from distributions that depend on random number generator (RNG) seeds, so for statistical trustworthiness we measure results over policies trained on multiple seeds. The base seed initializes a single PRNG key, from which the parallel multi-seed runs reported in Section 6 each receive an independent key by splitting.

D.3 Baseline Hyperparameters and Tuning Budget

All methods share the same actor-critic trunk (two hidden layers of 256 units, $\tanh$), optimizer (Adam at 3×10^{-4}, linearly annealed to zero), gradient-norm clip (0.5), reward discount ($\gamma = 0.99$), cost discount ($\gamma_c = 1.0$), and GAE parameter ($\lambda = 0.95$), and are run for 30 seeds. CVaR-CPO additionally carries a distributional cost critic (two hidden layers of 64 units, $\tanh$) that the other methods do not require. Environment-level settings, shared across methods, are given in Table 3.

Table 3: Environment-level settings, identical across all methods.

Setting	**EcoAnt**	**PointGoal**	**IcyLake**
Cost limit ℓ	150	100	5.5
Violation threshold ε	0.02	0.02	0.02
Parallel environments	20	20	5
Rollout length	100	125	200
Total environment steps	10M	2M	1M

The trust-region methods (Canary, CPO, CVaR-CPO) share CPO's update machinery and its settings: trust region $\delta = 0.01$, backtracking coefficient 0.8 with at most 10 backtracking iterations, conjugate-gradient damping 0.1, and no entropy bonus. They differ only in how the constraint pair (c, b) is constructed. CPO is applied to the breach-indicator variant of each environment with cost limit ε and an undiscounted indicator return, so that it constrains $\mathbb{P}(C(\tau) > \ell) \leq \varepsilon$ directly (Section 3.4). Per-method settings for the first-order methods are given in Table 4.

Table 4: Per-method hyperparameters for the first-order baselines. Ratio clip and value-loss coefficients follow standard PPO practice.

Method	**Hyperparameter**	**EcoAnt**	**PointGoal**	**IcyLake**
PPO	Ratio clip	0.2	0.2	0.2
	Epochs per update	4	4	4
	Minibatch size	500	500	40
	Entropy coeff.	10^{-5}	0.0075	0.01
CPPO	λ init	50	50	1
	λ step size	0.03	0.01	0.03
	PID gain k_p	600	60	15
	CVaR clip ratio	1000	300	300
	Epochs per update	10	4	4
	Entropy coeff.	0.0075	0.0075	0.01

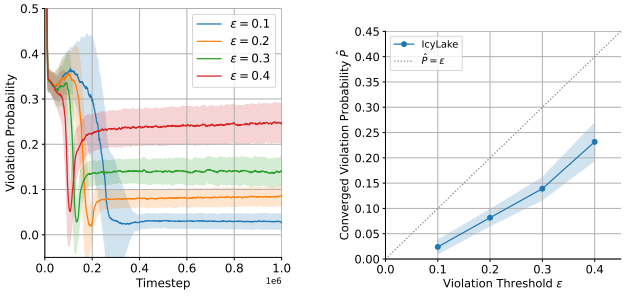
Tuning budget. CPPO is the only method whose hyperparameters were selected by search, because its Lagrangian gains have no established default in this setting. For each environment we swept a 12-point grid: on EcoAnt and PointGoal, $k_p \in \{60, 100, 150\}$ crossed with λ step size $\in \{0.01, 0.03\}$ and λ initialization $\in \{1, 50\}$ over 2 seeds per configuration, extended along the gain axis (up to $k_p = 2400$ with λ cap 4000 on EcoAnt, $k_p = 600$ on PointGoal); on IcyLake, $k_p \in \{0, 15, 60, 150\}$ crossed with λ step size $\in \{10^{-3}, 3 \times 10^{-2}, 3 \times 10^{-1}\}$ over 5 seeds per configuration. Configurations were ranked by steady-state adherence, leading candidates were re-run at 30 seeds before selection, and reward broke adherence ties. Canary, CPO and CVaR-CPO received no search: all three inherit CPO's default trust region $\delta = 0.01$, and PPO uses standard settings. The tuning budget therefore favors the constrained baseline rather than the proposed method.

The sweeps also show that CPPO's failures in Table 1 are not tuning artifacts. On IcyLake the sweep is flat: all 12 configurations converge to $\hat{P} \in [0.32, 0.37]$ at $\hat{J} \approx 1$, and driving λ to its cap changes neither, so the CVaR penalty supplies no usable adherence gradient at $\varepsilon = 0.02$ and CPPO behaves as unconstrained PPO. On EcoAnt, adherence improves monotonically with gain up to $k_p = 600$, bottoming out at roughly twice the threshold, then reverses as larger gains ($k_p \geq 1200$) destabilize the Lagrangian and destroy reward; no configuration in the $40\times$ gain range reaches ε.

D.4 Violation-Probability Jitter

Per seed, we partition the second half of training into 12 equal non-overlapping blocks and compute block violation probabilities p_b (exceedances over episodes) with episode counts n_b. A frozen policy would still show block-to-block variance from finite sampling; we subtract this binomial counting floor, $\mathrm{mean}_b[\bar{p}(1-\bar{p})/n_b]$ with $\bar{p}$ the seed's pooled second-half rate, from the observed block variance $\mathrm{Var}_b[p_b]$ and report the square root of the excess: the standard deviation with which the converged policy's violation probability wanders. Values are seed means with standard errors. The floor subtraction assumes episodes are independent given the policy, which holds in these environments.

D.5 Canary Violation Threshold (ε) Sensitivity



(a) Training Dynamics (b) Converged Calibration

Figure 4: **Violation Threshold Sensitivity:** (a) Probability of a cost return exceeding $\ell = 15$ in IcyLake with varying violation thresholds ε. Lines are mean and shaded areas represent one standard deviation over 500 seeds. (b) The same sweep with violation threshold vs the converged violation probability $\hat{P}$.

For Canary, we conduct a sensitivity analysis on the violation threshold ε in Figure 4a. Excluding the plateau around violation probability 0.35, the rate of convergence is independent of ε. The collapse from violation probability 0.3 to 0.1 takes a median of roughly 20k steps at every ε. Figure 4b further highlights the conservative nature of Canary established in Section 4.1, with empirical expected violation rates settling 2–4$\times$ below the thresholds set.