QUANTILE-CONSTRAINED REINFORCEMENT LEARNING VIA DISTRIBUTIONAL CREDIT ASSIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

In many sequential decision-making applications, such as healthcare and inventory control, the goal is not only to achieve high average returns but also to avoid poor lower-tail outcomes. We study expected-return maximization under a lower-quantile constraint and focus on the key challenge of assigning trajectory-level risk to individual decisions. Using an equivalent violation-probability formulation, we propose *distributional credit assignment* with two key components. (1) At each step, we use accumulated rewards to determine how much future return is needed to meet the constraint. We then define a transition-based risk advantage from the change in violation probability between consecutive augmented states. (2) We estimate these violation probabilities from conditional future-return distributions learned through distributional Bellman updates. This uses intermediate reward magnitudes to estimate risk, rather than relying solely on terminal binary outcomes. We combine the resulting risk advantages with reward advantages in proximal policy optimization (PPO). Theoretically, we derive a density-free risk-gradient representation, provide error bounds for risk-gradient estimation, and establish convergence under a three-time-scale stochastic-approximation formulation. Experiments in synthetic mean-risk trade-off and SimpleButton environments show that our method provides favorable performance in balancing expected return and lower-tail constraints compared with existing quantile-constrained baselines.

1 INTRODUCTION

In many sequential decision-making problems, maximizing expected cumulative reward alone is insufficient. Decision-makers must also limit how often outcomes fall below a minimum acceptable level. For example, in healthcare, one may seek to improve average health outcomes while ensuring that a sufficiently large proportion of patients achieve a clinically meaningful outcome (Fang et al., 2023). Quantile-based criteria have also been studied for sequential treatment decisions, including when to initiate HIV therapy (Li et al., 2022). In inventory management, high demand uncertainty makes it important to guard against poor lower-tail profit outcomes (Jiang et al., 2023). The goal may therefore be to maximize expected cumulative profit while limiting the probability of falling below a minimum profit target. Related concerns arise in dynamic portfolio management, where investment objectives are pursued subject to value-at-risk constraints on future wealth (Basak and Shapiro, 2001). These applications share a common goal: to optimize average performance while maintaining a minimum level of lower-tail performance. We study this problem in reinforcement learning (RL). A policy π specifies how actions are selected based on the information available at each decision. Let G^π denote the discounted cumulative reward obtained under this policy. We seek a policy that solves

$$\max_{\pi} \mathbb{E}[G^\pi] \qquad \text{subject to} \qquad Q_\alpha(G^\pi) \geq q,$$

where $Q_\alpha(G^\pi)$ denotes the lower α-quantile of G^π for $\alpha \in (0, 1)$ and q is a prescribed threshold.

This problem presents three related challenges:

1. Differentiating the return quantile with respect to policy parameters introduces the reciprocal of the return density at the current quantile. This density can be difficult to estimate accurately, especially when it is small, and taking its reciprocal can further amplify estimation error. Tilted quantile policy optimization (TQPO) (Li et al., 2025) avoids explicit density estimation by estimating

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.

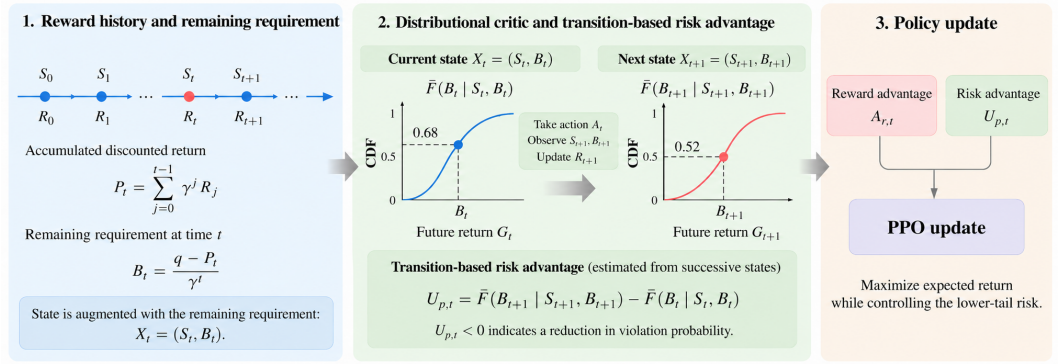

Figure 1: Overview of distributional credit assignment. Accumulated rewards determine the remaining return requirement, while successive state-level violation probabilities yield a transition-based risk advantage that is combined with the reward advantage in PPO.

quantile-gradient directions from samples. However, it still directly tracks the policy-dependent return quantile, whereas we use the equivalent violation probability at a fixed return threshold.

2. Unlike expectations, quantiles are not additive: the quantile of a sum is generally not the sum of its quantiles. As a result, return quantiles do not satisfy the standard Bellman recursion, and conventional temporal-difference updates cannot be applied directly. One way to recover a dynamic programming structure is through state augmentation (Li et al., 2022). Quantile-constrained policy optimization (QCPO) (Jung et al., 2022) takes a different approach by approximating the cumulative-cost quantile with an expected discounted sum of modified stage costs. This enables stepwise policy updates but introduces approximation error into the quantile representation.
3. The quantile constraint is equivalent to a bound on the probability that the total return falls at or below q. However, assigning this trajectory-level risk to individual actions remains challenging. In episodic tasks, a direct Monte Carlo gradient estimator weights each action's policy score by the same terminal indicator $\mathbf{1}\{G^\pi \leq q\}$. For early decisions, this indicator is also affected by later actions and environmental randomness, resulting in *delayed and noisy risk signals*. The key challenge is to construct informative per-step risk advantages from a trajectory-level constraint.

Building on the violation-probability formulation, we propose *distributional credit assignment* to convert trajectory-level risk into per-step risk advantages. Our proposal uses accumulated rewards and learned future-return distributions to evaluate how risk changes along the trajectory. Our contributions are fourfold.

- We reformulate the trajectory-level constraint using a remaining return requirement determined by the rewards observed so far. This yields an augmented Markov state and allows the original lower-quantile constraint to be evaluated through conditional violation probabilities.
- We learn these probabilities from conditional future-return distributions using distributional Bellman updates (Bellemare et al., 2017; Dabney et al., 2018b). Based on the change in violation probability across consecutive augmented states, we construct a transition-based risk advantage and combine it with the reward advantage in PPO.
- We establish a density-free representation of the risk gradient, derive error bounds for its estimation, and prove convergence of the proposed algorithm.
- We evaluate our method in synthetic mean–risk trade-off environments with different horizons and in the SimpleButton environment. Across these settings, our method shows favorable performance relative to existing quantile-constrained baselines in balancing expected return and constraints.

In summary, our approach uses observed rewards in two complementary ways: accumulated rewards determine the remaining return requirement, while intermediate reward magnitudes are used to learn the future-return distribution (see Figure 1). Together, these components enable the construction of per-step risk advantages for the original trajectory-level constraint.

2 RELATED WORK

Our work is related to three branches of research: quantile-based decision making, chance-constrained RL and distributional RL. We review related works in each branch in detail below.

Quantile-based decision making. One line of work maximizes the outcome quantile rather than the mean. Gilbert et al. (2017) study quantile-optimal planning in Markov decision processes. Li et al. (2022) develop dynamic programming methods that include the quantile level in the state. For model-free learning, Jiang et al. (2023) propose two-timescale policy-gradient algorithms, while Hau et al. (2025) develop a Q-learning method with convergence guarantees. Quantile maximization has also been studied for static and dynamic treatment regimes (Wang et al., 2018).

A different line of work retains the mean as the objective and imposes a quantile constraint on lower-tail performance. Fang et al. (2023) adopt this criterion for individualized treatment-rule learning and extend it to dynamic treatment regimes. Our work follows this mean-quantile formulation but focuses on temporal credit assignment.

Chance-constrained RL. Chance-constrained RL limits the probability of undesirable trajectory outcomes. The following approaches differ in how they represent the constraint and estimate risk for policy optimization.

- One approach expresses a chance constraint as a restriction on a cost quantile. QCPO (Jung et al., 2022) approximates the cost quantile as an expected sum of modified costs. Its policy updates rely on this additive approximation and tail-density estimation. TQPO (Li et al., 2025) instead estimates quantile-gradient directions from samples, avoiding both the additive approximation and explicit density estimation. Focusing on more extreme outcomes, EVO (Gao et al., 2025) uses extreme-value models to estimate extreme cost quantiles from limited tail observations.
- Another approach replaces the chance constraint with a conservative sufficient condition. Canary (Tangri and Calliess, 2026) uses Cantelli's inequality to bound violation probability through the mean and variance of cumulative cost. This avoids explicit tail modeling, but the resulting condition can exclude policies that satisfy the original chance constraint.
- A third approach works directly with violation probabilities to construct policy updates. For cumulative-cost constraints, Chow et al. (2018) augment the state with the remaining cost budget. Their actor-critic learns a scalar value function that includes a terminal violation penalty. Temporal-difference updates propagate this penalty to earlier decisions. Chen et al. (2024) instead study the probability of staying in a safe set throughout an episode. Their gradient combines the observed safety of the trajectory prefix with the conditional probability that the remaining trajectory is safe. In their implementation, this probability is fitted to binary outcomes indicating whether all subsequent states remain safe.

Our work also uses violation probabilities for policy updates, but learns the required risk feedback through future-return distributions. We fit these distributions using distributional Bellman updates that incorporate intermediate reward magnitudes. Accumulated rewards determine how much future return is needed to meet the original threshold. We then use the learned distributions to estimate each action's probability of falling below this requirement. Thus, the risk critic learns from numerical rewards rather than only from binary event labels, while policy updates target the original trajectory-level probability constraint.

Distributional RL. Distributional RL learns the distribution of cumulative reward, rather than only its mean (Bellemare et al., 2017). Quantile regression and implicit quantile networks provide practical methods for estimating this distribution (Dabney et al., 2018a;b). Beyond return estimation, Quantile Credit Assignment (Mesnard et al., 2023) uses learned distributions to construct baselines that reduce variance in policy-gradient estimation.

Distributional RL also provides a basis for evaluating and constraining different measures of risk. WCSAC (Yang et al., 2023) uses a distributional safety critic to estimate conditional value-at-risk (CVaR), which measures average cost in the upper tail. SRCPO (Kim et al., 2024) considers spectral risk measures, defined through weighted averages of cost quantiles. Zhang and Weng (2022) develop a broader distributional framework that accommodates constraints on CVaR, variance, and

the probability of reaching unsafe states. These studies establish distribution learning as a tool for risk-constrained policy optimization. Our focus is on using this tool to learn per-step risk feedback for the original trajectory-level violation-probability constraint.

Evaluating individual actions under a cumulative-return constraint also requires accounting for rewards already accumulated. RAC (Yang et al., 2024) handles trajectory-level risk constraints by introducing surrogate constraints at individual states. It learns distributions of past and future costs and combines their quantiles to construct risk upper bounds. The surrogate constraints require these bounds to remain below prescribed limits at each state. In contrast, we use the rewards already observed to determine how much future return is needed to meet the original threshold. We then estimate each action's violation probability from its conditional future-return distribution.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $\mathcal{R}$ specifies the reward distribution, P is the transition kernel, and $\gamma \in (0,1)$ is the discount factor. At each time t, the agent observes state S_t, selects action A_t under policy π, receives reward R_t, and transitions to S_{t+1}. We assume uniformly bounded rewards. The infinite-horizon discounted return under policy π is $G^\pi = \sum_{t=0}^{\infty} \gamma^t R_t$.

For a fixed $\alpha \in (0,1)$, define the lower quantile $Q_\alpha(G^\pi) = \inf\{z : \Pr(G^\pi \le z) \ge \alpha\}$. We seek to maximize expected return subject to $Q_\alpha(G^\pi) \ge q$, where q is a prescribed return threshold. Assume that the cumulative distribution function (CDF) of return is continuous and strictly increasing near its α-quantile for every policy considered. Then the quantile constraint is equivalent to $\Pr(G^\pi \le q) \le \alpha$. For a policy π_θ parameterized by θ, let $J(\theta) = \mathbb{E}[G^{\pi_\theta}]$ and $p(\theta) = \Pr(G^{\pi_\theta} \le q)$ denote its expected return and violation probability, respectively. Our problem becomes

$$\max_\theta \; J(\theta) \qquad \text{subject to} \qquad p(\theta) \le \alpha.$$

We use the Lagrangian

$$L(\theta, \lambda) = J(\theta) - \lambda\{p(\theta) - \alpha\}, \qquad \lambda \ge 0, \tag{1}$$

where λ is the Lagrange multiplier. The policy parameters are updated by gradient ascent and the multiplier by projected gradient descent. While the expected-return gradient can be estimated using standard methods, we focus next on constructing per-step risk advantages for estimating $\nabla_\theta p(\theta)$.

3.2 DISTRIBUTIONAL CREDIT ASSIGNMENT

To optimize the Lagrangian in (1), we need the gradient of the violation probability $p(\theta)$. Let H_t denote the information available before selecting action A_t. Since $p(\theta) = \mathbb{E}[\mathbf{1}\{G^{\pi_\theta} \le q\}]$, likelihood-ratio differentiation gives

$$\nabla_\theta p(\theta) = \sum_{t=0}^{\infty} \mathbb{E}[\mathbf{1}\{G^{\pi_\theta} \le q\} \nabla_\theta \log \pi_\theta(A_t \mid H_t)]\,, \tag{2}$$

under the conditions in Proposition 1. Each policy score is therefore weighted by the same trajectory-level violation indicator. For early decisions, this indicator also reflects later actions and environmental randomness. This motivates constructing per-step risk advantages that preserve the gradient of the violation probability.

We first use accumulated rewards to determine the remaining return required to satisfy the original threshold. Let $P_t = \sum_{j=0}^{t-1} \gamma^j R_j$ denote the accumulated discounted reward, with $P_0 = 0$, and let $G_t^\pi = \sum_{j=0}^{\infty} \gamma^j R_{t+j}$ denote the future return from time t. Before selecting A_t, P_t is known whereas G_t^π remains uncertain. Since $G^\pi = P_t + \gamma^t G_t^\pi$,

$$\{G^\pi \le q\} = \{G_t^\pi \le B_t\}, \qquad B_t = \frac{q - P_t}{\gamma^t}. \tag{3}$$

Thus, B_t is the remaining return requirement at time t. It satisfies $B_0 = q$ and $B_{t+1} = (B_t - R_t)/\gamma$. This recursion, together with the environment dynamics, makes $X_t = (S_t, B_t)$ an augmented Markov state. We henceforth consider stationary policies on this augmented state, so $\pi_\theta(\cdot \mid H_t) = \pi_\theta(\cdot \mid S_t, B_t)$.

The conditional future-return distribution allows us to evaluate the violation probability at each augmented state and action. For a fixed policy π, define

$$F^\pi(z \mid s, b, a) = \Pr(G_t^\pi \le z \mid S_t = s, B_t = b, A_t = a)\,,$$

where subsequent actions follow π. Evaluating the CDF at the remaining requirement b gives the conditional probability of violating the original return threshold. In particular, (3) implies

$$\mathbb{E}[\mathbf{1}\{G^{\pi_\theta} \le q\} \mid S_t, B_t, A_t] = F^{\pi_\theta}(B_t \mid S_t, B_t, A_t). \tag{4}$$

Thus, by the tower property, the trajectory-level indicator in (2) can be replaced by the corresponding conditional violation probability. Define the policy-average violation probability at (s, b) as

$$\bar{F}^\pi(s, b) = \mathbb{E}_{a\sim\pi(\cdot|s,b)}\left[F^\pi(b \mid s, b, a)\right]. \tag{5}$$

We then define the *risk advantage*

$$A_p^\pi(s, b, a) = F^\pi(b \mid s, b, a) - \bar{F}^\pi(s, b). \tag{6}$$

It measures how the violation probability of action a differs from the policy average at the same augmented state. Since $\bar{F}^\pi(s, b)$ does not depend on the selected action, its expected product with the policy score is zero. Together with (4), this yields the following result.

Proposition 1 (Exact risk policy gradient)**.** *Assume that π_θ has differentiable densities with parameter-independent support and that, for each augmented state x, differentiation can be interchanged with integration over actions. Let $Y = \mathbf{1}\{G^{\pi_\theta} \le q\}$ and $s_t = \nabla_\theta \log \pi_\theta(A_t \mid X_t)$. Suppose that $\mathbb{E}\|s_t\|_2 < \infty$ for every $t \ge 0$, that $p(\theta)$ is differentiable, and that $\nabla_\theta p(\theta) = \sum_{t=0}^\infty \mathbb{E}[Y s_t]$, $\sum_{t=0}^\infty \|\mathbb{E}[Y s_t]\|_2 < \infty$, where all expectations are taken under π_θ. Then*

$$\nabla_\theta p(\theta) = \sum_{t=0}^{\infty} \mathbb{E}\left[A_p^{\pi_\theta}(S_t, B_t, A_t)\nabla_\theta \log \pi_\theta(A_t \mid S_t, B_t)\right]. \tag{7}$$

Transition-based risk advantage. The risk advantage in (6) compares the violation probability of the selected action with the policy average at the same augmented state. For policy optimization, we use an equivalent transition-based representation that assigns risk credit through the realized reward and successor state. Specifically, the selected action induces a transition from (S_t, B_t) to (S_{t+1}, B_{t+1}), where the observed reward enters through $B_{t+1} = (B_t - R_t)/\gamma$. Recall that $\bar{F}^\pi(s, b)$ denotes the policy-average violation probability at augmented state (s, b). For a nonterminal transition, define

$$U_{p,t}^\pi = \bar{F}^\pi(S_{t+1}, B_{t+1}) - \bar{F}^\pi(S_t, B_t), \tag{8}$$

while at genuine termination, $U_{p,t}^\pi = Y - \bar{F}^\pi(S_t, B_t)$ and $Y = \mathbf{1}\{G^\pi \le q\}$. The observed reward enters through the update $B_{t+1} = (B_t - R_t)/\gamma$. Under the augmented Markov formulation,

$$\mathbb{E}\left[U_{p,t}^\pi \mid S_t, B_t, A_t\right] = A_p^\pi(S_t, B_t, A_t). \tag{9}$$

Thus, $U_{p,t}^\pi$ is a transition-based risk advantage whose conditional expectation equals the exact risk advantage. The derivation of (8) is given in Appendix A. Equations (6)-(8) motivate our distributional credit assignment.

3.3 Risk estimation and policy optimization

The transition-based identity in (8) requires conditional violation probabilities. We estimate them from the conditional distribution of future returns. This allows the critic to learn from reward observations through distributional Bellman updates and to evaluate the violation probability at the current remaining-return requirement.

Step 1: Estimate risk advantages. We represent the conditional future-return distribution by N predicted quantiles $z_{\psi,i}(s, b, a)$ at levels $\tau_i = (i - \frac{1}{2})/N$, $i = 1, \ldots, N$. With π_k fixed, we learn these

quantiles using distributional Bellman updates (Bellemare et al., 2017; Dabney et al., 2018a). The critic therefore learns from numerical reward observations rather than only from terminal violation indicators. Given the fitted critic, let $\widehat{F}_\psi(b \mid s, b, a)$ denote the estimated CDF evaluated at the remaining requirement b. We approximate the policy-average violation probability by

$$\widehat{\bar{F}}_t = \frac{1}{M}\sum_{m=1}^{M} \widehat{F}_\psi(B_t \mid S_t, B_t, \widetilde{A}_{t,m}), \qquad \widetilde{A}_{t,m} \sim \pi_k(\cdot \mid S_t, B_t). \tag{10}$$

For a nonterminal transition, an independent set of actions is sampled at (S_{t+1}, B_{t+1}) to obtain $\widehat{\bar{F}}_{t+1}$. Following the transition-based risk advantage introduced in Section 3.2, we use

$$\widehat{U}_{p,t} = \begin{cases} \widehat{\bar{F}}_{t+1} - \widehat{\bar{F}}_t, & \text{if the transition is nonterminal,} \\ Y - \widehat{\bar{F}}_t, & \text{at genuine termination,} \end{cases} \tag{11}$$

where $Y = \mathbf{1}\{G \le q\}$. For comparison, the corresponding direct risk-advantage estimator is

$$\widehat{A}_{p,t} = \widehat{F}_\psi(B_t \mid S_t, B_t, A_t) - \widehat{\bar{F}}_t. \tag{12}$$

This direct estimator is the object analyzed in Section 4, whereas the experiments use $\widehat{U}_{p,t}$.

Step 2: Update the policy. A separate scalar value critic provides the reward advantage $\widehat{A}_{r,t}$. Let $\rho_t(\theta) = \pi_\theta(A_t \mid S_t, B_t)/\pi_k(A_t \mid S_t, B_t)$ and let $\bar{\rho}_t(\theta) = \mathrm{clip}(\rho_t(\theta), 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}})$. We apply PPO clipping separately to the reward and risk terms. Up to a common positive normalization factor, the actor maximizes

$$\widehat{L}_k(\theta) = \widehat{\mathbb{E}}_k\Big[\gamma^t \min\{\rho_t(\theta)\widehat{A}_{r,t}, \bar{\rho}_t(\theta)\widehat{A}_{r,t}\} - \lambda_k \max\{\rho_t(\theta)\widehat{U}_{p,t}, \bar{\rho}_t(\theta)\widehat{U}_{p,t}\}\Big]. \tag{13}$$

Here $\widehat{\mathbb{E}}_k$ averages over valid rollout transitions. The factor γ^t aligns the reward term with the time-zero expected-return objective, while the risk term requires no additional discount.

Step 3: Update the multiplier. For the theoretical formulation, we use the standard projected dual update

$$\lambda_{k+1} = \mathrm{clip}\big(\lambda_k + \eta_\lambda(\widehat{p}_k - \alpha), 0, \lambda_{\max}\big),$$

where $\widehat{p}_k$ is the empirical violation probability computed from recent rollout returns, $\eta_\lambda > 0$ is the multiplier step size, and $\lambda_{\max} > 0$ is an upper bound used to keep the multiplier bounded. The multiplier increases when the observed violation rate exceeds α and decreases otherwise. Further implementation details are provided in Appendix D.

4 Theoretical Analysis

In this section, we first quantify the mean square error (MSE) of the full-trajectory risk-gradient estimator. We then analyze the joint convergence of the critic, actor, and multiplier, and characterize how critic error affects the limiting policy. Proofs are provided in Appendix B.

4.1 Error bounds for risk-gradient estimation

We study risk-gradient estimation for a fixed policy $\pi = \pi_\theta$. Write $s_t = \nabla_\theta \log \pi_\theta(A_t \mid X_t)$ for the policy score. Using L independent trajectories and M alternative actions per decision, we form

$$\widehat{g}_{p,\infty} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty} \widehat{A}_{p,\ell,t} s_{\ell,t}, \tag{14}$$

where $\widehat{A}_{p,\ell,t}$ is the risk-advantage estimate from Section 3.3. The indices ℓ and t identify the trajectory and decision, respectively. All expectations below include randomness from trajectories, critic fitting, and alternative-action sampling.

We assume $|R_t| \le r_{\max}$ and write $R_\star = r_{\max}/(1-\gamma)$ for the corresponding bound on future returns. Policies have differentiable densities with parameter-independent support and admit continuously differentiable finite-prefix likelihood-ratio formulas. Their scores satisfy

$$\mathbb{E}[s_t \mid X_t] = 0, \qquad \mathbb{E}[\|s_t\|_2^2 \mid X_t] \le C_s^2. \tag{15}$$

These conditions hold uniformly on a neighborhood of the parameters considered.

Assumption 1 (Probability mass near the return threshold)**.** There exist $t_0 \geq 1$, $C_q > 0$, and $\nu \in (0, 1]$ such that, $\Pr(|P_t - q| \leq u) \leq C_q u^\nu$, for all $t \geq t_0$ and $u > 0$.

This condition controls the contribution of late decisions. Since the remaining discounted return has magnitude at most $\gamma^t R_\star$, the risk advantage vanishes outside $E_t = \{|P_t - q| \leq \gamma^t R_\star\}$. Assumption 1 makes these probabilities decrease geometrically, allowing the risk-gradient series to converge without an additional discount factor.

Assumption 2 (Accuracy of risk estimation)**.** The estimated return distributions are supported on $[-R_\star, R_\star]$. Let $e_\psi(s, b, a) = \widehat{F}_\psi(b \mid s, b, a) - F^\pi(b \mid s, b, a)$. For a deterministic $\delta \geq 0$, assume that, almost surely at every visited augmented state, $\sup_{a,a' \in \operatorname{supp} \pi(\cdot|s,b)} |e_\psi(s, b, a) - e_\psi(s, b, a')| \leq \delta$. Here, $\operatorname{supp} \pi(\cdot \mid s, b)$ denotes the support of the action distribution under policy π at (s, b).

Here, δ bounds the error in estimating differences between actions' violation probabilities. This is the relevant error because the risk advantage subtracts the policy-action average. The support condition can be enforced by projecting predicted quantiles onto the interval $[-R_\star, R_\star]$. For the transition-based estimator $\widehat{U}_{p,t}$ used in practice, the analogous approximation error is the difference between state-risk estimation errors at successive augmented states. Writing $e_\psi(x) = \widehat{F}_\psi(x) - \bar{F}^\pi(x)$, the relevant quantity is $e_\psi(X_{t+1}) - e_\psi(X_t)$ for nonterminal transitions, with the terminal risk observed exactly. The present analysis focuses on the direct risk-advantage estimator; the transition-based case admits an analogous error decomposition.

Theorem 1 (MSE of risk-gradient estimation)**.** *Suppose Assumptions 1–2 hold. Freeze the fitted critic before drawing the actions. Then the series in (14) converges in mean square and*

$$\mathbb{E} \left\| \widehat{g}_{p,\infty} - \nabla_\theta p(\theta) \right\|_2^2 \leq \frac{2\sigma_p^2}{L} + 2C_s^2 H_\nu^2 \delta^2 + \frac{C_s^2 K_\nu}{4LM}, \tag{16}$$

$$\left\| \mathbb{E} \widehat{g}_{p,\infty} - \nabla_\theta p(\theta) \right\|_2 \leq C_s H_\nu \delta, \tag{17}$$

where $\sigma_p^2 = \mathbb{E} \left\| \sum_{t=0}^\infty A_{p,t}^\pi s_t - \nabla_\theta p(\theta) \right\|_2^2 < \infty$ *with* $A_{p,t}^\pi = A_p^\pi(S_t, B_t, A_t)$*, and* $H_\nu = t_0 + \sqrt{C_q}\, R_\star^{\nu/2} \gamma^{\nu t_0/2} / (1 - \gamma^{\nu/2})$*,* $K_\nu = t_0 + C_q R_\star^\nu \gamma^{\nu t_0} / (1 - \gamma^\nu)$*. In particular, the estimator is unbiased when* $\delta = 0$*. As* $L \to \infty$ *and* $\delta \to 0$*, it converges to* $\nabla_\theta p(\theta)$ *in mean square.*

The three terms in (16) describe different sources of error. The first, $2\sigma_p^2/L$, bounds trajectory-sampling error. Here, σ_p^2 is the total variance of one trajectory's gradient contribution using true risk advantages, including all cross-time covariances. The second, $2C_s^2 H_\nu^2 \delta^2$, bounds the contribution of critic estimation error. The third, $C_s^2 K_\nu / (4LM)$, accounts for approximating the policy-action average with M alternative actions. This sampling introduces no additional bias relative to exact averaging under the fitted critic. Its variance contribution decreases as $1/(LM)$, so fixed M suffices for mean-square consistency when $L \to \infty$ and $\delta \to 0$. The preceding bounds concern the infinite-horizon estimator, whereas practical updates use finite rollouts. Retaining only the first T decisions omits a portion of the true risk gradient whose norm is at most $C_s C_q R_\star^\nu \gamma^{\nu T} / (1 - \gamma^\nu)$ for $T \geq t_0$. This bound concerns truncating the gradient sum while retaining the infinite-horizon return target. Appendix B provides the full finite-rollout error bound. An analogous error bound for the transition-based estimator is provided in Appendix B.

4.2 Convergence and constraint satisfaction

The critic, actor and multiplier are updated jointly. For the convergence analysis, we let the critic adapt fastest, the actor update on a slower scale, and the multiplier change most slowly. This separation allows the critic to track the current policy and the actor to adapt before substantial changes in the multiplier. We refer to this coupled learning scheme as three-time-scale stochastic approximation. Let ψ_k, θ_k, and λ_k denote the distributional-critic, policy, and multiplier parameters at iteration k, respectively. Their projected updates are

$$\psi_{k+1} = \Pi_\Psi[\psi_k + \beta_k \widehat{h}_k],\ \theta_{k+1} = \Pi_\Theta[\theta_k + a_k\{\widehat{g}_{r,k} - \lambda_k \widehat{g}_{p,k}\}],\ \lambda_{k+1} = \Pi_\Lambda[\lambda_k + b_k(\widehat{p}_k - \alpha)],$$

where Π_C denotes Euclidean projection onto C. Here, $\widehat{h}_k$ is the stochastic critic update, $\widehat{g}_{r,k}$ and $\widehat{g}_{p,k}$ are the estimated reward and risk gradients, and $\widehat{p}_k$ estimates the current violation probability.

For a fixed policy parameter θ, let $h_\psi(\psi, \theta)$ denote the mean critic update direction. We define $\psi^*(\theta)$ as its stable equilibrium, satisfying $h_\psi(\psi^*(\theta), \theta) = 0$. Given critic parameter ψ, let $g_p(\theta, \psi)$ denote the mean risk-gradient direction. After the critic has equilibrated, the actor mean update directionis $h_\theta(\theta, \lambda) = \nabla_\theta J(\theta) - \lambda g_p(\theta, \psi^*(\theta))$. For a fixed multiplier λ, let $\theta^*(\lambda)$ denote a stable equilibrium of the actor update, satisfying $h_\theta(\theta^*(\lambda), \lambda) = 0$. We then define $r(\lambda) = p(\theta^*(\lambda)) - \alpha$, which measures the constraint residual at this equilibrium. Here, $\psi^*(\theta)$ and $\theta^*(\lambda)$ denote equilibria of the corresponding mean updates, not necessarily optima of the original constrained problem.

We assume that the critic and actor mean updates admit stable equilibria when the slower variables are held fixed, and that the equilibrium violation probability responds monotonically to the multiplier. The precise regularity and stability conditions are given in Appendix B.

Theorem 2. *Suppose Assumptions 1–3 hold together with the stochastic-approximation conditions specified in Appendix B.3. Then*

$$(\psi_k, \theta_k, \lambda_k) \xrightarrow{\text{a.s.}} (\psi^*, \theta^*, \lambda^*),$$

where $\theta^* = \theta^*(\lambda^*)$ and $\psi^* = \psi^*(\theta^*)$. Moreover, $p(\theta^*) \le \alpha$, $\lambda^*\{p(\theta^*) - \alpha\} = 0$. If δ^* denotes the equilibrium error in estimating conditional violation-probability differences, then

$$\|\nabla_\theta L(\theta^*, \lambda^*)\|_2 \le \lambda^* C_s H_\nu \delta^*.$$

When $\delta^* = 0$, the limit satisfies the KKT conditions of the original probability-constrained problem.

The theorem shows that the coupled critic-actor-multiplier updates converge to a feasible limiting policy. Critic estimation error determines the remaining error in the first-order optimality condition. When this error vanishes, the limiting solution satisfies the KKT conditions.

5 Experiments

We first evaluate *distributional credit assignment* (DCA) in a synthetic mean–risk trade-off environment and compare it with existing quantile-constrained baselines. The same framework can be extended to expected-reward maximization under a cumulative-cost quantile constraint. We evaluate this extension on SimpleButton in Appendix E.

Experimental setting. We first consider a synthetic downside-risk control problem designed to exhibit a clear trade-off between mean return and lower-tail risk. Each episode has $H = 20$ or $H = 40$ decision steps with discount factor $\gamma = 0.99$. At each step, the agent chooses a continuous action $A_t \in [0, 1]$ and receives $R_t = 1 + 2A_t - 50A_tD_t + 0.02\xi_t$, where $D_t \sim \text{Bernoulli}(0.025)$ and $\xi_t \sim \text{Uniform}[-1, 1]$ are sampled independently after the action is selected. Larger actions therefore yield higher expected rewards but expose the agent to larger losses when a downside shock occurs. We maximize the expected discounted return subject to $Q_{0.2}(G) \ge 16$ for $H = 20$ and $Q_{0.2}(G) \ge 29.09$ for $H = 40$. We compare our method with QCPO (Jung et al., 2022) and TQPO (Li et al., 2025). Each method is trained with ten independent seeds under the same interaction budget of 2.048 million transitions per seed. Final policies are evaluated using 10^5 independent episodes. We report the mean discounted return, the 20% return quantile, the violation probability, and the number of empirically feasible runs. Environment and evaluation details are given in Appendix C.

Implementation details. All methods use the same sigmoid-Gaussian policy architecture and the same on-policy rollout batches. Our method represents the conditional future-return distribution using $N = 64$ quantiles and learns it with one-step distributional Bellman targets and the pinball loss. Conditional violation probabilities are obtained by interpolating the predicted quantiles at the current remaining-return requirement. We estimate the policy-average risk at both the current and observed next augmented states using $M = 4$ policy actions, and construct the risk credit from their difference. The reward and risk terms are then combined using separately clipped PPO surrogates. The multiplier is updated from the observed episode-level violation rate using a proportional–integral (PI) controller (Stooke et al., 2020; Sohrabi et al., 2024), including during the critic warm-up period. We use three critic epochs and at most four PPO epochs per iteration. Further implementation details are provided in Appendix D.

Results. Figure 2 compares DCA with QCPO and TQPO under two horizons. For $H = 20$, DCA achieves the highest mean return, 27.565 ± 0.531. Its average violation probability is $(18.193 \pm$

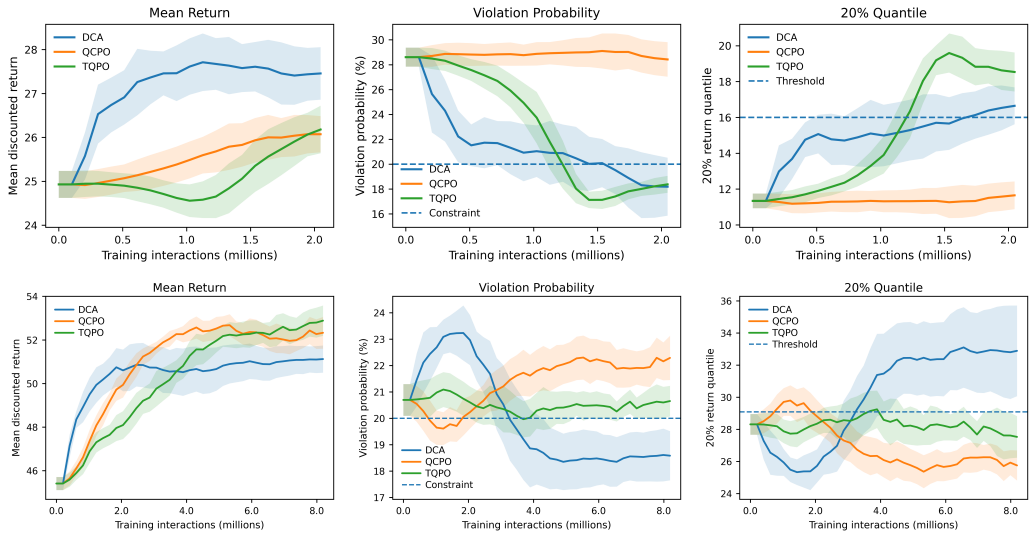

Figure 2: Learning performance for $H = 20$ (top) and $H = 40$ (bottom). Columns show mean discounted return, violation probability, and the 20% return quantile.

$2.221)\%$, very close to TQPO at $(18.288 \pm 0.619)\%$. DCA therefore improves the mean return by 4.98% relative to TQPO at a similar average risk level. TQPO, however, achieves a higher 20% return quantile, 18.606 ± 0.979, and more stable constraint satisfaction across seeds. The empirical constraint is satisfied by 8 of the 10 DCA policies and all 10 TQPO policies, whereas QCPO remains above the prescribed violation level on average. Thus, at $H = 20$, DCA mainly improves expected return, while TQPO provides stronger lower-tail performance.

For $H = 40$, DCA has a lower mean return than QCPO and TQPO, but provides much stronger constraint control. Its average violation probability is $(18.599 \pm 1.131)\%$, below the prescribed 20% level, compared with $(22.208 \pm 1.131)\%$ for QCPO and $(20.724 \pm 0.625)\%$ for TQPO. All 10 DCA policies satisfy the empirical constraint, compared with none for QCPO and one for TQPO. DCA also achieves the highest 20% return quantile, 32.618 ± 3.178, which exceeds the corresponding return threshold, while the average quantiles of both baselines remain below it. Hence, although the two baselines obtain higher mean returns, their policies generally violate the lower-tail requirement. The $H = 40$ results therefore highlight DCA's stronger constraint satisfaction and lower-tail performance.

The weak constraint control of QCPO may be related to how it represents the quantile constraint. In our problem, the lower-tail constraint is imposed directly on the same cumulative reward that appears in the objective. QCPO instead converts the problem to a cumulative-cost formulation and uses an approximate quantile representation for stepwise policy updates. This additional approximation may weaken the constraint signal in the present environment. We interpret this result as specific to the evaluated configuration, rather than as a general limitation of QCPO.

6 Conclusion

We propose distributional credit assignment for RL under a lower-quantile constraint. By combining accumulated rewards with conditional future-return distributions, our method converts a trajectory-level risk requirement into per-step risk advantages for policy optimization. Our analysis establishes a density-free risk-gradient representation, provides error bounds for risk-gradient estimation, and gives convergence guarantees. Experiments in synthetic mean–risk trade-off and SimpleButton environments demonstrate favorable performance compared with existing quantile-constrained baselines. An interesting direction for future work is to consider constraints with time-varying thresholds. Since DCA learns the conditional future-return distribution, it can evaluate violation probabilities at different remaining-return requirements, providing a natural basis for such extensions.

Ethics statement

The study uses simulated control tasks and no human or private data.

Reproducibility statement

Section 3 describes the proposed method, and Section 4 states the assumptions and theoretical results. Appendix A derives the transition-based risk-advantage identity, and Appendix B provides the theoretical proofs and extensions. Appendices C and D describe the synthetic environment, evaluation protocol, and implementation details. Appendix E describes the cumulative-cost extension and the SimpleButton experiments, including training configurations, evaluation procedures, and the aggregation of results across random seeds. Anonymous code for the H = 20 synthetic experiments is provided in the supplementary materials, together with dependency specifications and instructions for training, evaluation, and plotting.

AI use statement

We used generative AI tools solely to polish the language and improve the clarity and readability of the manuscript. The authors take full responsibility for the final manuscript and all scientific claims.

References

Suleyman Basak and Alexander Shapiro. Value-at-risk-based risk management: Optimal policies and asset prices. *The Review of Financial Studies*, 14(2):371–405, 2001. doi:10.1093/rfs/14.2.371.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458, 2017.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, 2008.

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

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

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

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1096–1105, 2018.

Ethan X. Fang, Zhaoran Wang, and Lan Wang. Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, 118(543):1733–1746, 2023. doi:10.1080/01621459.2021.2008402.

Shiqing Gao, Yihang Zhou, Shuai Shao, Haoyu Luo, Yiheng Bing, Jiaxin Ding, Luoyi Fu, and Xinbing Wang. Extreme value policy optimization for safe reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18772–18793, 2025.

Hugo Gilbert, Paul Weng, and Yan Xu. Optimizing quantiles in preference-based Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 3569–3575, 2017. doi:10.1609/aaai.v31i1.11026.

Jia Lin Hau, Erick Delage, Esther Derman, Mohammad Ghavamzadeh, and Marek Petrik. Q-learning for quantile MDPs: A decomposition, performance, and convergence analysis. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2665–2673, 2025.

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 *Advances in Neural Information Processing Systems*, volume 36, pages 18964–18993, 2023.

Jinyang Jiang, Jiaqiao Hu, and Yijie Peng. Quantile-based deep reinforcement learning using two-timescale policy gradient algorithms. *arXiv preprint arXiv:2305.07248*, 2023.

Whiyoung Jung, Myungsik Cho, Jongeui Park, and Youngchul Sung. Quantile constrained reinforcement learning: A reinforcement learning framework constraining outage probability. In *Advances in Neural Information Processing Systems*, volume 35, pages 6437–6449, 2022.

Dohyeong Kim, Taehyun Cho, Seungyub Han, Hojun Chung, Kyungjae Lee, and Songhwai Oh. Spectral-risk safe reinforcement learning with convergence guarantees. In *Advances in Neural Information Processing Systems*, volume 37, pages 92465–92498, 2024.

Xiaocheng Li, Huaiyang Zhong, and Margaret L. Brandeau. Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447, 2022. doi:10.1287/opre.2021.2123.

Chenglin Li, Guangchun Ruan, and Hua Geng. Tilted quantile gradient updates for quantile-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18236–18243, 2025. doi:10.1609/aaai.v39i17.34006.

Thomas Mesnard, Wenqi Chen, Alaa Saade, Yunhao Tang, Mark Rowland, Théophane Weber, Clare Lyle, Audrunas Gruslys, Michal Valko, Will Dabney, Georg Ostrovski, Eric Moulines, and Rémi Munos. Quantile credit assignment. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24517–24531, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Motahareh Sohrabi, Juan Ramirez, Tianyue H. Zhang, Simon Lacoste-Julien, and Jose Gallego-Posada. On PI controllers for updating Lagrange multipliers in constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45922–45954, 2024.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9133–9143, 2020.

Rohan Tangri and Jan-Peter Calliess. Cantelli constrained policy optimization. *arXiv preprint arXiv:2601.22993*, 2026.

Lan Wang, Yu Zhou, Rui Song, and Ben Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018. doi:10.1080/01621459.2017.1330204.

Qisong Yang, Thiago D. Simão, Simon H. Tindemans, and Matthijs T. J. Spaan. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3):859–887, 2023. doi:10.1007/s10994-022-06187-8.

Zhaoxing Yang, Haiming Jin, Yao Tang, and Guiyun Fan. Risk-aware constrained reinforcement learning with non-stationary policies. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2029–2037, 2024.

Jianyi Zhang and Paul Weng. Safe distributional reinforcement learning. In *Distributed Artificial Intelligence (DAI 2021)*, volume 13170 of *Lecture Notes in Computer Science*, pages 107–128. Springer, 2022. doi:10.1007/978-3-030-94662-3_8.

A Derivation of the transition-based risk-advantage identity

We show that the transition-based construction in (9) has the same conditional mean as the risk advantage defined in (6).

Let

$$\bar{F}^\pi(X_t) = \mathbb{E}_{a\sim\pi(\cdot|X_t)}\left[F^\pi(B_t \mid S_t, B_t, a)\right] = \Pr_\pi(G^\pi \le q \mid X_t)$$

denote the policy-average violation probability at the augmented state $X_t = (S_t, B_t)$. For a nonterminal transition, define

$$U^\pi_{p,t} = \bar{F}^\pi(X_{t+1}) - \bar{F}^\pi(X_t),$$

while at genuine termination let

$$U^\pi_{p,t} = Y - \bar{F}^\pi(X_t), \qquad Y = \mathbf{1}\{G^\pi \le q\}.$$

Consider first a nonterminal transition. Conditioning on (X_t, A_t) and using the tower property,

$$\mathbb{E}\big[\bar{F}^\pi(X_{t+1}) \mid X_t, A_t\big] = \Pr_\pi(G^\pi \le q \mid X_t, A_t).$$

Indeed, after observing the transition to X_{t+1}, all subsequent actions follow π, and therefore $\bar{F}^\pi(X_{t+1})$ is the conditional probability of eventual violation from the successor state. By the definition of the action-conditional violation probability,

$$\Pr_\pi(G^\pi \le q \mid X_t, A_t) = F^\pi(B_t \mid S_t, B_t, A_t).$$

Hence

$$\begin{aligned}\mathbb{E}\big[U^\pi_{p,t} \mid X_t, A_t\big] &= F^\pi(B_t \mid S_t, B_t, A_t) - \bar{F}^\pi(X_t) \\ &= A^\pi_p(S_t, B_t, A_t). \end{aligned} \tag{18}$$

At genuine termination, Y is the realized violation indicator. Thus

$$\mathbb{E}[Y \mid X_t, A_t] = \Pr_\pi(G^\pi \le q \mid X_t, A_t) = F^\pi(B_t \mid S_t, B_t, A_t),$$

and therefore

$$\begin{aligned}\mathbb{E}\big[U^\pi_{p,t} \mid X_t, A_t\big] &= F^\pi(B_t \mid S_t, B_t, A_t) - \bar{F}^\pi(X_t) \\ &= A^\pi_p(S_t, B_t, A_t). \end{aligned} \tag{19}$$

Thus, in both cases,

$$\mathbb{E}\big[U^\pi_{p,t} \mid X_t, A_t\big] = A^\pi_p(S_t, B_t, A_t).$$

Consequently, replacing the direct risk advantage by $U^\pi_{p,t}$ preserves the expected policy-gradient contribution:

$$\mathbb{E}\big[U^\pi_{p,t}\nabla_\theta \log \pi_\theta(A_t \mid X_t)\big] = \mathbb{E}\big[A^\pi_p(S_t, B_t, A_t)\nabla_\theta \log \pi_\theta(A_t \mid X_t)\big].$$

B Proofs and extensions

This appendix provides proofs of the main theoretical results and related extensions. We first prove the risk-gradient representation and compare the variance of the per-decision gradient contributions. We then bound risk-gradient estimation error, relate the critic contribution to distributional accuracy, and account for finite rollouts. Finally, we establish joint convergence of the critic, actor, and multiplier and bound the stationarity error at their limit. The main-text results are restated before their proofs.

Unless stated otherwise, we use the notation and regularity conditions of Section 4. The policy is defined on the augmented state $X_t = (S_t, B_t)$. All objectives and return distributions are defined in terms of infinite-horizon discounted returns. For finite rollouts, we truncate the gradient sum but retain the infinite-horizon return target.

B.1 Proof of Proposition 1

Proposition 1 expresses the exact violation-probability gradient through conditional risk advantages, assigning each action its own contribution without changing the objective.

Proposition 1 (Exact risk policy gradient)**.** *Assume that π_θ has differentiable densities with parameter-independent support and that, for each augmented state x, differentiation can be interchanged with integration over actions. Let $Y = \mathbf{1}\{G^{\pi_\theta} \le q\}$ and $s_t = \nabla_\theta \log \pi_\theta(A_t \mid X_t)$. Suppose that $\mathbb{E}\|s_t\|_2 < \infty$ for every $t \ge 0$, that $p(\theta)$ is differentiable, and that $\nabla_\theta p(\theta) = \sum_{t=0}^{\infty} \mathbb{E}[Y s_t]$, $\sum_{t=0}^{\infty} \|\mathbb{E}[Y s_t]\|_2 < \infty$, where all expectations are taken under π_θ. Then*

$$\nabla_\theta p(\theta) = \sum_{t=0}^{\infty} \mathbb{E}\big[A_p^{\pi_\theta}(S_t, B_t, A_t) \nabla_\theta \log \pi_\theta(A_t \mid S_t, B_t)\big]. \tag{20}$$

Proof. Fix θ and write $\pi = \pi_\theta$ throughout the proof. All probabilities and expectations are taken under π. Let $Y = \mathbf{1}\{G^\pi \le q\}$ and write $A^\pi_{p,t} = A^\pi_p(S_t, B_t, A_t)$.

Fix $t \ge 0$. Since $G^\pi = P_t + \gamma^t G^\pi_t$ and $B_t = (q - P_t)/\gamma^t$, we have $Y = \mathbf{1}\{G^\pi_t \le B_t\}$. By the definition of F^π,

$$\mathbb{E}[Y \mid X_t, A_t] = \Pr(G^\pi_t \le B_t \mid X_t, A_t) = F^\pi(B_t \mid S_t, B_t, A_t).$$

The assumption $\mathbb{E}\|s_t\|_2 < \infty$, together with $0 \le Y \le 1$, ensures that $Y s_t$ is integrable. Since s_t is a function of (X_t, A_t), the law of iterated expectations gives

$$\mathbb{E}[Y s_t] = \mathbb{E}[\mathbb{E}[Y \mid X_t, A_t] s_t] = \mathbb{E}[F^\pi(B_t \mid S_t, B_t, A_t) s_t].$$

The assumed interchange of differentiation and integration imply

$$\mathbb{E}[s_t \mid X_t = x] = \int \nabla_\theta \log \pi_\theta(a \mid x)\, \pi_\theta(a \mid x)\, da = \nabla_\theta \int \pi_\theta(a \mid x)\, da = 0.$$

Define

$$\overline{F}^\pi(X_t) = \mathbb{E}_{a \sim \pi(\cdot \mid X_t)}[F^\pi(B_t \mid S_t, B_t, a)].$$

For fixed θ, this is a function of X_t taking values in $[0, 1]$. Its product with s_t is therefore integrable, and

$$\mathbb{E}[\overline{F}^\pi(X_t) s_t] = \mathbb{E}\Big[\overline{F}^\pi(X_t) \mathbb{E}[s_t \mid X_t]\Big] = 0.$$

Subtracting this zero-mean term and using the definition of the risk advantage yields

$$\mathbb{E}[Y s_t] = \mathbb{E}\Big[\Big\{F^\pi(B_t \mid S_t, B_t, A_t) - \overline{F}^\pi(X_t)\Big\} s_t\Big] = \mathbb{E}[A^\pi_{p,t} s_t].$$

This identity gives

$$\nabla_\theta p(\theta) = \sum_{t=0}^{\infty} \mathbb{E}[Y s_t] = \sum_{t=0}^{\infty} \mathbb{E}[A^\pi_{p,t} s_t],$$

which proves the result. □

For a fixed decision t and an action-independent baseline $c(X_t)$, we replace Y in $(Y - c(X_t))s_t$ by $\mathbb{E}[Y \mid X_t, A_t]$. The following proposition shows that this replacement preserves the expected gradient contribution and does not increase its covariance matrix in the positive semidefinite order.

Proposition 2 (Variance reduction by conditional risk credit)**.** *Fix a decision t and an action-independent baseline $c(X_t)$ such that $(Y - c(X_t))s_t$ is square-integrable, where $Y = \mathbf{1}\{G^\pi \le q\}$. Define $V_t = (Y - c(X_t))s_t$ and $\overline{V}_t = (F^\pi(B_t \mid S_t, B_t, A_t) - c(X_t))s_t$. Then $\overline{V}_t = \mathbb{E}[V_t \mid X_t, A_t]$, $\mathbb{E}\overline{V}_t = \mathbb{E}V_t$, and*

$$\mathrm{Cov}(V_t) - \mathrm{Cov}(\overline{V}_t) = \mathbb{E}[\mathrm{Var}(Y \mid X_t, A_t) s_t s_t^\top] \succeq 0. \tag{21}$$

In particular, $c(X_t) = \overline{F}^\pi(X_t)$ gives $\overline{V}_t = A^\pi_{p,t} s_t$.

Proof. Since $G^\pi = P_t + \gamma^t G_t^\pi$ and $B_t = (q - P_t)/\gamma^t$, we have $Y = \mathbf{1}\{G_t^\pi \le B_t\}$. The definition of F^π therefore gives

$$\mathbb{E}[Y \mid X_t, A_t] = \Pr(G_t^\pi \le B_t \mid X_t, A_t) = F^\pi(B_t \mid S_t, B_t, A_t).$$

For fixed θ, the score s_t is a function of (X_t, A_t), while the baseline $c(X_t)$ depends only on X_t. Both can therefore be taken outside the conditional expectation:

$$\begin{aligned}\mathbb{E}[V_t \mid X_t, A_t] &= \big(\mathbb{E}[Y \mid X_t, A_t] - c(X_t)\big)s_t \\ &= \big(F^\pi(B_t \mid S_t, B_t, A_t) - c(X_t)\big)s_t = \overline{V}_t.\end{aligned} \tag{22}$$

Taking expectations gives $\mathbb{E}[\overline{V}_t] = \mathbb{E}[V_t]$.

Next, subtracting the conditional mean yields

$$V_t - \overline{V}_t = \big(Y - F^\pi(B_t \mid S_t, B_t, A_t)\big)s_t.$$

Since s_t is determined by (X_t, A_t), the conditional covariance satisfies

$$\begin{aligned}\mathrm{Cov}(V_t \mid X_t, A_t) &= \mathbb{E}\big[(V_t - \overline{V}_t)(V_t - \overline{V}_t)^\top \mid X_t, A_t\big] \\ &= \mathbb{E}\Big[\big(Y - F^\pi(B_t \mid S_t, B_t, A_t)\big)^2 \mid X_t, A_t\Big]\, s_t s_t^\top \\ &= \mathrm{Var}(Y \mid X_t, A_t) s_t s_t^\top.\end{aligned}$$

The law of total covariance now gives

$$\begin{aligned}\mathrm{Cov}(V_t) &= \mathrm{Cov}(\mathbb{E}[V_t \mid X_t, A_t]) + \mathbb{E}[\mathrm{Cov}(V_t \mid X_t, A_t)] \\ &= \mathrm{Cov}(\overline{V}_t) + \mathbb{E}\big[\mathrm{Var}(Y \mid X_t, A_t) s_t s_t^\top\big].\end{aligned} \tag{23}$$

Because $\mathrm{Var}(Y \mid X_t, A_t) \ge 0$ and $s_t s_t^\top$ is positive semidefinite, the last term is positive semidefinite. This proves (21).

Finally, setting $c(X_t) = \overline{F}^\pi(X_t)$ gives $\overline{V}_t = A_{p,t}^\pi s_t$ by the definition of the risk advantage. □

B.2 Proof of Theorem 1 and its extensions

Theorem 1 bounds the mean-square error of the infinite-horizon risk-gradient estimator, separating trajectory sampling, critic error, and alternative-action sampling. Proposition 3 relates the action-contrast error δ to a distributional Bellman residual. Proposition 4 then accounts for finite rollouts by bounding the omitted gradient tail and the error of the retained estimator. We also relate this risk-gradient bound to the complete Lagrangian direction.

Theorem 1 (MSE of risk-gradient estimation)**.** *Suppose Assumptions 1–2 hold, and the fitted critic is frozen before the alternative actions are sampled. Then the infinite-horizon estimator*

$$\widehat{g}_{p,\infty} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty} \widehat{A}_{p,\ell,t} s_{\ell,t}$$

converges in mean square and satisfies

$$\mathbb{E}\,\|\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2^2 \le \frac{2\sigma_p^2}{L} + 2C_s^2 H_\nu^2 \delta^2 + \frac{C_s^2 K_\nu}{4LM}, \tag{24}$$

$$\|\mathbb{E}\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2 \le C_s H_\nu \delta, \tag{25}$$

where $\sigma_p^2 := \mathbb{E}\,\big\|\sum_{t=0}^{\infty} A_{p,t}^\pi s_t - \nabla_\theta p(\theta)\big\|_2^2$ *and*

$$H_\nu := t_0 + \frac{\sqrt{C_q}\, R_\star^{\nu/2}\gamma^{\nu t_0/2}}{1-\gamma^{\nu/2}}, \qquad K_\nu := t_0 + \frac{C_q R_\star^\nu \gamma^{\nu t_0}}{1-\gamma^\nu}.$$

In particular, $\widehat{g}_{p,\infty}$ *is unbiased when* $\delta = 0$*. If* $L \to \infty$ *and* $\delta \to 0$ *with fixed* $M \ge 1$*, then*

$$\widehat{g}_{p,\infty} \xrightarrow{L^2} \nabla_\theta p(\theta).$$

Proof. The proof proceeds in three steps. We first show that the risk-gradient term at time t can be nonzero only when P_t lies within a shrinking $O(\gamma^t)$ neighborhood of the threshold q, namely, when $|P_t - q| \le \gamma^t R_\star$. Under Assumption 1, the probability of this event decays sufficiently fast with t, which provides the summability needed to define the infinite-horizon gradient series. We then decompose the gradient-estimation error into trajectory-sampling, critic-approximation, and Monte Carlo action-sampling errors and bound the three components separately. Finally, we combine these bounds, using the conditional zero-mean property of the Monte Carlo action-sampling error, to obtain the stated MSE and bias guarantees.

Step 1: The existence of infinite risk gradient and its finite variance. Recall that

$$G^\pi = P_t + \gamma^t G_t^\pi, \qquad |G_t^\pi| \le R_\star := \frac{r_{\max}}{1-\gamma}.$$

Thus $|\gamma^t G_t^\pi| \le \gamma^t R_\star$. Define

$$E_t := \{|P_t - q| \le \gamma^t R_\star\} = \{|B_t| \le R_\star\}, \qquad B_t = \frac{q - P_t}{\gamma^t}.$$

On E_t^c, B_t lies outside the support of G_t^π, so $F^\pi(B_t \mid S_t, B_t, a)$ is action-independent and hence $A_{p,t}^\pi = 0$. Since $|A_{p,t}^\pi| \le 1$, we have

$$|A_{p,t}^\pi| \le \mathbf{1}_{E_t}. \tag{26}$$

We next quantify how frequently E_t can occur. By Assumption 1, for every $t \ge t_0$,

$$\Pr(E_t) = \Pr\big(|P_t - q| \le \gamma^t R_\star\big) \le C_q(\gamma^t R_\star)^\nu = C_q R_\star^\nu \gamma^{\nu t}. \tag{27}$$

For $t < t_0$, we use only $\Pr(E_t) \le 1$. Therefore,

$$\sum_{t=0}^{\infty} \sqrt{\Pr(E_t)} \le t_0 + \sqrt{C_q}\, R_\star^{\nu/2} \sum_{t=t_0}^{\infty} \gamma^{\nu t/2} = t_0 + \frac{\sqrt{C_q}\, R_\star^{\nu/2} \gamma^{\nu t_0/2}}{1 - \gamma^{\nu/2}} = H_\nu, \tag{28}$$

$$\sum_{t=0}^{\infty} \Pr(E_t) \le t_0 + C_q R_\star^\nu \sum_{t=t_0}^{\infty} \gamma^{\nu t} = t_0 + \frac{C_q R_\star^\nu \gamma^{\nu t_0}}{1 - \gamma^\nu} = K_\nu. \tag{29}$$

These two summability relations are the key infinite-horizon estimates used throughout the remainder of the proof.

Before analyzing the estimated gradient, we use (28) to verify that the corresponding true infinite-horizon gradient contribution is well defined. Let

$$U_t := A_{p,t}^\pi s_t, \qquad \|Y\|_{L^2} := \big(\mathbb{E}\|Y\|_2^2\big)^{1/2}.$$

Because $E_t = \{|B_t| \le R_\star\}$ is measurable with respect to $X_t = (S_t, B_t)$, (26) and (15) imply

$$\begin{aligned} \|U_t\|_{L^2}^2 &= \mathbb{E}\big[|A_{p,t}^\pi|^2 \|s_t\|_2^2\big] \le \mathbb{E}\big[\mathbf{1}_{E_t}\|s_t\|_2^2\big] \\ &= \mathbb{E}\big[\mathbf{1}_{E_t}\mathbb{E}\big[\|s_t\|_2^2 \mid X_t\big]\big] \le C_s^2 \Pr(E_t). \end{aligned} \tag{30}$$

Hence

$$\|U_t\|_{L^2} \le C_s \sqrt{\Pr(E_t)}. \tag{31}$$

For integers $n > m$,

$$\left\| \sum_{t=m+1}^{n} U_t \right\|_{L^2} \le \sum_{t=m+1}^{n} \|U_t\|_{L^2} \le C_s \sum_{t=m+1}^{\infty} \sqrt{\Pr(E_t)} \xrightarrow{m\to\infty} 0, \tag{32}$$

where the last limit follows from (28). Therefore the partial sums $Z_N := \sum_{t=0}^{N} A_{p,t}^\pi s_t$ form a Cauchy sequence in L^2. Since L^2 is complete, there exists a random vector $Z \in L^2$ such that

$$Z_N \xrightarrow{L^2} Z, \qquad Z := \sum_{t=0}^{\infty} A_{p,t}^\pi s_t. \tag{33}$$

Furthermore, by Minkowski's inequality and (28),

$$\|Z\|_{L^2} \le \sum_{t=0}^{\infty} \|U_t\|_{L^2} \le C_s \sum_{t=0}^{\infty} \sqrt{\Pr(E_t)} \le C_s H_\nu. \tag{34}$$

Since L^2 convergence implies L^1 convergence,

$$\mathbb{E}\|Z_N - Z\|_2 \le \left(\mathbb{E}\|Z_N - Z\|_2^2\right)^{1/2} \longrightarrow 0.$$

Hence

$$\mathbb{E}Z = \lim_{N\to\infty} \mathbb{E}Z_N = \sum_{t=0}^{\infty} \mathbb{E}[A_{p,t}^{\pi} s_t] = \nabla_\theta p(\theta),$$

where the last equality follows from Proposition 1. Consequently,

$$\sigma_p^2 = \mathbb{E}\,\|Z - \nabla_\theta p(\theta)\|_2^2 = \mathbb{E}\|Z - \mathbb{E}Z\|_2^2 \le \mathbb{E}\|Z\|_2^2 \le C_s^2 H_\nu^2 < \infty. \tag{35}$$

Step 2: Decompose the estimation error and bound its three components. We introduce two intermediate estimators in order to separate the three sources of error. For trajectory ℓ, let

$$A_{p,\ell,t}^{\pi} := A_p^{\pi}(S_{\ell,t}, B_{\ell,t}, A_{\ell,t}).$$

Using the estimated CDF $\widehat{F}_\psi$, define

$$A_{p,\ell,t}^{\psi} := \widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, A_{\ell,t}) - \mathbb{E}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[\widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right]. \tag{36}$$

Here the expectation over actions is retained rather than approximated by a sample average. The corresponding infinite-horizon estimators are

$$\widehat{g}_{p,\infty}^{\mathrm{true}} := \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty} A_{p,\ell,t}^{\pi} s_{\ell,t}, \qquad \widehat{g}_{p,\infty}^{\psi} := \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty} A_{p,\ell,t}^{\psi} s_{\ell,t}. \tag{37}$$

The first uses the true conditional return distribution. The second replaces the true distribution by the fitted critic but still computes the action average exactly. The actual estimator further replaces this action expectation by the M-sample Monte Carlo average.

Accordingly,

$$\widehat{g}_{p,\infty} - \nabla_\theta p(\theta) = \underbrace{\left(\widehat{g}_{p,\infty}^{\mathrm{true}} - \nabla_\theta p(\theta)\right)}_{\Delta_{\mathrm{traj}}} + \underbrace{\left(\widehat{g}_{p,\infty}^{\psi} - \widehat{g}_{p,\infty}^{\mathrm{true}}\right)}_{\Delta_{\mathrm{crit}}} + \underbrace{\left(\widehat{g}_{p,\infty} - \widehat{g}_{p,\infty}^{\psi}\right)}_{\Delta_{\mathrm{MC}}}. \tag{38}$$

We now bound these three terms separately.

Trajectory-sampling error. For trajectory ℓ, define $Z_\ell := \sum_{t=0}^{\infty} A_{p,\ell,t}^{\pi} s_{\ell,t}$. By Step 1, Z_ℓ is well defined in L^2, and

$$\mathbb{E}Z_\ell = \nabla_\theta p(\theta), \qquad \mathbb{E}\,\|Z_\ell - \nabla_\theta p(\theta)\|_2^2 = \sigma_p^2.$$

Since the L rollout trajectories are independent, $Z_1, \dots, Z_L$ are i.i.d. and

$$\widehat{g}_{p,\infty}^{\mathrm{true}} = \frac{1}{L}\sum_{\ell=1}^{L} Z_\ell.$$

Therefore,

$$\begin{aligned}\mathbb{E}\|\Delta_{\mathrm{traj}}\|_2^2 &= \mathbb{E}\left\|\frac{1}{L}\sum_{\ell=1}^{L}\left(Z_\ell - \nabla_\theta p(\theta)\right)\right\|_2^2 \\ &= \frac{1}{L^2}\sum_{\ell=1}^{L}\mathbb{E}\,\|Z_\ell - \nabla_\theta p(\theta)\|_2^2 = \frac{\sigma_p^2}{L}.\end{aligned} \tag{39}$$

Here all cross terms vanish because different trajectories are independent and each centered variable has mean zero. In particular,

$$\mathbb{E}\Delta_{\text{traj}} = 0. \tag{40}$$

Critic-approximation error. Recall the fitted-CDF error

$$e_\psi(s, b, a) := \widehat{F}_\psi(b \mid s, b, a) - F^\pi(b \mid s, b, a).$$

By the definitions of $A^\psi_{p,\ell,t}$ and $A^\pi_{p,\ell,t}$,

$$\begin{aligned} A^\psi_{p,\ell,t} - A^\pi_{p,\ell,t} &= e_\psi(S_{\ell,t}, B_{\ell,t}, A_{\ell,t}) - \mathbb{E}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[e_\psi(S_{\ell,t}, B_{\ell,t}, a)\right] \\ &= \mathbb{E}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[e_\psi(S_{\ell,t}, B_{\ell,t}, A_{\ell,t}) - e_\psi(S_{\ell,t}, B_{\ell,t}, a)\right]. \end{aligned} \tag{41}$$

Assumption 2 therefore gives

$$|A^\psi_{p,\ell,t} - A^\pi_{p,\ell,t}| \le \delta.$$

Define $E_{\ell,t} := \{|P_{\ell,t} - q| \le \gamma^t R_\star\}$, on $E^c_{\ell,t}$, we have $|B_{\ell,t}| > R_\star$. Because both the true and fitted return distributions are supported on $[-R_\star, R_\star]$, both CDFs are constant in the action on this event, and hence

$$A^\pi_{p,\ell,t} = A^\psi_{p,\ell,t} = 0 \qquad \text{on } E^c_{\ell,t}.$$

Thus the previous bound sharpens to

$$|A^\psi_{p,\ell,t} - A^\pi_{p,\ell,t}| \le \delta \mathbf{1}_{E_{\ell,t}}. \tag{42}$$

Using (15),

$$\left\|(A^\psi_{p,\ell,t} - A^\pi_{p,\ell,t})s_{\ell,t}\right\|^2_{L^2} \le \delta^2 \mathbb{E}\big[\mathbf{1}_{E_{\ell,t}}\|s_{\ell,t}\|_2^2\big] \le C_s^2\delta^2 \Pr(E_t). \tag{43}$$

Hence, by Minkowski's inequality,

$$\begin{aligned} \|\Delta_{\text{crit}}\|_{L^2} &\le \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\left\|(A^\psi_{p,\ell,t} - A^\pi_{p,\ell,t})s_{\ell,t}\right\|_{L^2} \\ &\le C_s\delta\sum_{t=0}^{\infty}\sqrt{\Pr(E_t)} \le C_s H_\nu \delta. \end{aligned} \tag{44}$$

Therefore,

$$\mathbb{E}\|\Delta_{\text{crit}}\|_2^2 \le C_s^2 H_\nu^2 \delta^2. \tag{45}$$

The same argument applied to the tail $t = m+1, \dots, n$ yields

$$\left\|\frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=m+1}^{n}(A^\psi_{p,\ell,t} - A^\pi_{p,\ell,t})s_{\ell,t}\right\|_{L^2} \le C_s\delta\sum_{t=m+1}^{\infty}\sqrt{\Pr(E_t)} \longrightarrow 0.$$

Thus the fitted-critic series also converges in L^2.

Monte Carlo action-sampling error. Let

$$\mathcal{F} := \sigma(\{S_{\ell,t}, A_{\ell,t}, R_{\ell,t} : \ell = 1, \dots, L,\ t \ge 0\}, \psi) \tag{46}$$

be the σ-field generated by the complete rollout batch and the fitted critic parameter. Since

$$P_{\ell,t} = \sum_{j=0}^{t-1}\gamma^j R_{\ell,j}, \qquad B_{\ell,t} = \frac{q - P_{\ell,t}}{\gamma^t},$$

the quantities $P_{\ell,t}$, $B_{\ell,t}$, $X_{\ell,t} = (S_{\ell,t}, B_{\ell,t})$, $s_{\ell,t}$, and $\widehat{F}_\psi$ are all $\mathcal{F}$-measurable.

Conditional on $\mathcal{F}$, the fitted critic and all rollout quantities are therefore fixed, while the alternative actions satisfy

$$\widetilde{A}_{\ell,t,m} \sim \pi(\cdot \mid S_{\ell,t}, B_{\ell,t}), \qquad m = 1, \dots, M,$$

and are sampled independently. Define

$$\overline{F}^{\psi}_{\ell,t} := \mathbb{E}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[\widehat{F}_{\psi}(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right]$$

and

$$\xi_{\ell,t,m} := \widehat{F}_{\psi}(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, \widetilde{A}_{\ell,t,m}) - \overline{F}^{\psi}_{\ell,t}.$$

By construction,

$$\mathbb{E}[\xi_{\ell,t,m} \mid \mathcal{F}] = 0. \tag{47}$$

Moreover, from the definition of $\widehat{A}_{p,\ell,t}$,

$$\widehat{A}_{p,\ell,t} - A^{\psi}_{p,\ell,t} = -\frac{1}{M}\sum_{m=1}^{M} \xi_{\ell,t,m}. \tag{48}$$

Thus

$$\Delta_{\mathrm{MC}} = -\frac{1}{LM}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\sum_{m=1}^{M} s_{\ell,t}\xi_{\ell,t,m}. \tag{49}$$

Because $\widehat{F}_{\psi}$ is a CDF,

$$0 \le \widehat{F}_{\psi}(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a) \le 1.$$

Hence its variance under the policy action is at most $1/4$. Furthermore, on $E^{c}_{\ell,t}$ the support condition makes the fitted CDF constant in a. Therefore

$$\mathrm{Var}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[\widehat{F}_{\psi}(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right] \le \frac{1}{4}\mathbf{1}_{E_{\ell,t}}. \tag{50}$$

Conditional on $\mathcal{F}$, the variables $\xi_{\ell,t,m}$ are centered and the alternative draws are independent. Therefore the cross terms vanish, and

$$\begin{aligned}\mathbb{E}\big[\|\Delta_{\mathrm{MC}}\|_2^2 \mid \mathcal{F}\big] &= \frac{1}{L^2M}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\|s_{\ell,t}\|_2^2\,\mathrm{Var}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[\widehat{F}_{\psi}(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right] \\ &\le \frac{1}{4L^2M}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\mathbf{1}_{E_{\ell,t}}\|s_{\ell,t}\|_2^2.\end{aligned} \tag{51}$$

Taking expectation and using (15),

$$\begin{aligned}\mathbb{E}\|\Delta_{\mathrm{MC}}\|_2^2 &\le \frac{1}{4L^2M}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\mathbb{E}\big[\mathbf{1}_{E_{\ell,t}}\|s_{\ell,t}\|_2^2\big] \\ &\le \frac{C_s^2}{4L^2M}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\Pr(E_t) \le \frac{C_s^2K_{\nu}}{4LM}.\end{aligned} \tag{52}$$

The same calculation restricted to the tail $t = m+1, \dots, n$ gives

$$\mathbb{E}\left\|\frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=m+1}^{n}(\widehat{A}_{p,\ell,t} - A^{\psi}_{p,\ell,t})s_{\ell,t}\right\|_2^2 \le \frac{C_s^2}{4LM}\sum_{t=m+1}^{\infty}\Pr(E_t) \longrightarrow 0.$$

Hence the Monte Carlo error series is Cauchy in L^2. Together with the convergence of the true and fitted-critic series, this establishes the mean-square existence of $\widehat{g}_{p,\infty}$.

Finally, (47) implies

$$\mathbb{E}[\Delta_{\mathrm{MC}} \mid \mathcal{F}] = 0, \qquad \mathbb{E}\Delta_{\mathrm{MC}} = 0. \tag{53}$$

Step 3: Combine the componentwise bounds. From (38),

$$\widehat{g}_{p,\infty} - \nabla_\theta p(\theta) = \Delta_{\mathrm{traj}} + \Delta_{\mathrm{crit}} + \Delta_{\mathrm{MC}}.$$

Both Δ_{traj} and Δ_{crit} are $\mathcal{F}$-measurable, while (53) gives

$$\mathbb{E}[\Delta_{\mathrm{MC}} \mid \mathcal{F}] = 0.$$

Since $\mathbb{E}[\Delta_{\mathrm{MC}} \mid \mathcal{F}] = 0$, conditioning on $\mathcal{F}$ and applying the tower property gives

$$\mathbb{E}\big[(\Delta_{\mathrm{traj}} + \Delta_{\mathrm{crit}})^\top \Delta_{\mathrm{MC}}\big] = 0.$$

and therefore

$$\mathbb{E}\,\|\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2^2 = \mathbb{E}\|\Delta_{\mathrm{MC}}\|_2^2 + \mathbb{E}\,\|\Delta_{\mathrm{traj}} + \Delta_{\mathrm{crit}}\|_2^2\,. \tag{54}$$

We do not assume that Δ_{traj} and Δ_{crit} are independent, since the fitted critic may be trained using the same rollout batch used by the actor. Hence we use

$$\|x + y\|_2^2 \le 2\|x\|_2^2 + 2\|y\|_2^2.$$

Combining (39), (45), and (52),

$$\begin{aligned}\mathbb{E}\,\|\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2^2 &\le \mathbb{E}\|\Delta_{\mathrm{MC}}\|_2^2 + 2\mathbb{E}\|\Delta_{\mathrm{traj}}\|_2^2 + 2\mathbb{E}\|\Delta_{\mathrm{crit}}\|_2^2 \\ &\le \frac{C_s^2 K_\nu}{4LM} + \frac{2\sigma_p^2}{L} + 2C_s^2 H_\nu^2 \delta^2. \end{aligned} \tag{55}$$

Therefore

$$\mathbb{E}\,\|\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2^2 \le \frac{2\sigma_p^2}{L} + 2C_s^2 H_\nu^2 \delta^2 + \frac{C_s^2 K_\nu}{4LM}. \tag{56}$$

For the bias, taking expectations in (38) and using (40) and (53) gives

$$\mathbb{E}\widehat{g}_{p,\infty} - \nabla_\theta p(\theta) = \mathbb{E}\Delta_{\mathrm{crit}}.$$

Jensen's inequality and Cauchy–Schwarz inequality, together with (45) therefore yields

$$\begin{aligned}\|\mathbb{E}\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2 &= \|\mathbb{E}\Delta_{\mathrm{crit}}\|_2 \le \mathbb{E}\|\Delta_{\mathrm{crit}}\|_2 \\ &\le \big(\mathbb{E}\|\Delta_{\mathrm{crit}}\|_2^2\big)^{1/2} \le C_s H_\nu \delta. \end{aligned} \tag{57}$$

In particular, if $\delta = 0$, then

$$\mathbb{E}\widehat{g}_{p,\infty} = \nabla_\theta p(\theta).$$

Finally, for fixed $M \ge 1$,

$$\frac{2\sigma_p^2}{L} \longrightarrow 0, \qquad 2C_s^2 H_\nu^2 \delta^2 \longrightarrow 0, \qquad \frac{C_s^2 K_\nu}{4LM} \longrightarrow 0$$

as $L \to \infty$ and $\delta \to 0$. It follows from (56) that

$$\mathbb{E}\,\|\widehat{g}_{p,\infty} - \nabla_\theta p(\theta)\|_2^2 \longrightarrow 0,$$

that is,

$$\widehat{g}_{p,\infty} \xrightarrow{L^2} \nabla_\theta p(\theta).$$

□

Theorem 1 expresses the critic contribution in terms of the action-contrast error δ. The next proposition bounds this error using a distributional Bellman residual.

Transition-based risk-gradient estimation. We next extend the estimation guarantee to the transition-based risk credit in (11). To distinguish critic error from Monte Carlo action-sampling error, define the exact policy average of the fitted critic by

$$\bar{F}_\psi(x) := \mathbb{E}_{a\sim\pi(\cdot|x)}\big[\widehat{F}_\psi(b \mid s,b,a)\big], \qquad x=(s,b),$$

and let

$$e_\psi(x) := \bar{F}_\psi(x) - \bar{F}^\pi(x).$$

Here $\bar{F}_\psi$ uses exact action averaging, whereas $\widehat{\bar{F}}_t$ in (10) is its M-sample approximation.

Corollary 1 (Bias and MSE of transition-based risk-gradient estimation)**.** *Fix $\pi=\pi_\theta$ and retain the bounded-reward, policy-regularity, and score conditions of Section 4.1, together with Assumption 1. In place of Assumption 2, suppose that the fitted return distributions are supported on $[-R_\star, R_\star]$ and that, for a deterministic $\delta_U \geq 0$, almost surely over trajectories and critic fitting, every sampled transition satisfies*

$$|\Delta e_{\psi,t}| \leq \delta_U,$$
$$\Delta e_{\psi,t} := \begin{cases} e_\psi(X_{t+1}) - e_\psi(X_t), & \text{if nonterminal,} \\ -e_\psi(X_t), & \text{at genuine termination.} \end{cases} \tag{58}$$

At genuine termination, use the observed indicator $Y=\mathbf{1}\{G^\pi \leq q\}$ exactly and set all subsequent gradient contributions to zero.

Draw L independent trajectories under π and freeze the fitted critic before drawing the auxiliary actions. Conditional on these trajectories and all critic-fitting randomness, use M i.i.d. actions from $\pi(\cdot \mid x)$ for each queried state x, independently across trajectories, transitions, and the two endpoints of each transition. Each query uses fresh draws, including when a state also occurs in an adjacent transition. The fitted critic may depend on the trajectory batch.

For $\widehat{U}_{p,\ell,t}$ defined in (11), let

$$\widehat{g}^U_{p,\infty} := \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{\infty}\widehat{U}_{p,\ell,t}s_{\ell,t}. \tag{59}$$

Then this series converges in mean square and satisfies

$$\mathbb{E}\left\|\widehat{g}^U_{p,\infty} - \nabla_\theta p(\theta)\right\|_2^2 \leq \frac{2\sigma_U^2}{L} + 2C_s^2H_\nu^2\delta_U^2 + \frac{C_s^2K_\nu}{2LM}, \tag{60}$$

$$\left\|\mathbb{E}\widehat{g}^U_{p,\infty} - \nabla_\theta p(\theta)\right\|_2 \leq C_sH_\nu\delta_U, \tag{61}$$

where H_ν and K_ν are the constants in Theorem 1, and

$$\sigma_U^2 := \mathbb{E}\left\|\sum_{t=0}^{\infty}U^\pi_{p,t}s_t - \nabla_\theta p(\theta)\right\|_2^2 \leq C_s^2H_\nu^2 < \infty.$$

In particular, $\widehat{g}^U_{p,\infty}$ is unbiased when $\delta_U=0$. As $L\to\infty$ and $\delta_U\to 0$, it converges to $\nabla_\theta p(\theta)$ in mean square with any fixed $M\geq 1$.

The three terms in (60) account for trajectory sampling with the true transition credit, successive-state critic error, and finite action averaging, respectively. The variance σ_U^2 includes randomness in the observed transitions and all cross-time covariances. The action-sampling term accounts for two independently estimated state probabilities at each nonterminal transition; genuine termination requires only the current-state estimate. Thus, the transition-based estimator has the same $O(L^{-1})+O(\delta_U^2)+O((LM)^{-1})$ error structure as the direct estimator in Theorem 1. A uniform state-risk error bound $|e_\psi(x)|\leq\varepsilon$ at all queried states is sufficient with $\delta_U=2\varepsilon$.

For return-law kernels, write $\overline{W}_1(Z,Z') = \sup_{s,b,a} W_1(Z(s,b,a), Z'(s,b,a))$, where the supremum is over an augmented state–action domain closed under continuation dynamics and containing the queried inputs. The fixed-policy distributional Bellman operator is

$$(\mathcal{T}^\pi Z)(s,b,a) \overset{D}{=} R + \gamma Z(S', (b-R)/\gamma, A'), \qquad A'\sim\pi(\cdot \mid S', (b-R)/\gamma).$$

The joint environment kernel supplies (R, S'), with zero continuation at genuine termination. We use its standard γ-contraction in $\overline{W}_1$ and the midpoint-quantile projection bound $\overline{W}_1(\mathcal{Q}_N Z, Z) \le R_\star/N$ for laws on $[-R_\star, R_\star]$; $\mathcal{Q}_N$ places equal mass on the quantiles $(i - \frac{1}{2})/N$, $i = 1, \dots, N$ (Bellemare et al., 2017; Dabney et al., 2018a). These operator properties are used only to obtain the certificate below.

Proposition 3 (Distributional error certificate for risk credit)**.** *Fix $\pi = \pi_\theta$. Suppose Z_ψ and Z^π are supported on $[-R_\star, R_\star]$ and $\overline{W}_1(Z_\psi, \mathcal{T}^\pi Z_\psi) \le \varepsilon_B$ almost surely. Assume that, uniformly over queried (s, b, a), for every $u > 0$ and some $C_F > 0$, $\kappa \in (0, 1]$,*

$$\max\{F^\pi(b + u \mid s, b, a) - F^\pi(b \mid s, b, a), F^\pi(b \mid s, b, a) - F^\pi(b - u \mid s, b, a)\} \le C_F u^\kappa.$$

Let $e_\psi(s, b, a) = \widehat{F}_\psi(b \mid s, b, a) - F^\pi(b \mid s, b, a)$. Then

$$\begin{aligned} \overline{W}_1(Z_\psi, Z^\pi) &\le \frac{\varepsilon_B}{1 - \gamma}, \\ \sup_{s,b,a} |e_\psi(s, b, a)| &\le 2C_F^{1/(\kappa+1)} \left(\frac{\varepsilon_B}{1 - \gamma}\right)^{\kappa/(\kappa+1)}, \\ \sup_{s,b} \sup_{a,a' \in \operatorname{supp} \pi(\cdot|s,b)} |e_\psi(s, b, a) - e_\psi(s, b, a')| &\le 4C_F^{1/(\kappa+1)} \left(\frac{\varepsilon_B}{1 - \gamma}\right)^{\kappa/(\kappa+1)}. \end{aligned} \tag{62}$$

Thus the last right-hand side is a valid choice of δ in Theorem 1. If instead $\overline{W}_1(Z_\psi, \mathcal{Q}_N \mathcal{T}^\pi Z_\psi) \le \varepsilon_{\mathrm{proj}}$, the same conclusions hold with ε_B replaced by $\varepsilon_{\mathrm{proj}} + R_\star/N$. In particular, a valid choice is

$$\delta = 4C_F^{1/(\kappa+1)} \left(\frac{\varepsilon_{\mathrm{proj}} + R_\star/N}{1 - \gamma}\right)^{\kappa/(\kappa+1)}. \tag{63}$$

Proof. We first turn the Bellman residual into a distributional error, then bound the queried CDF error by a coupling argument and take action differences.

The Bellman residual and contraction imply

$$\overline{W}_1(Z_\psi, Z^\pi) \le \varepsilon_B + \gamma \overline{W}_1(Z_\psi, Z^\pi), \qquad d := \overline{W}_1(Z_\psi, Z^\pi) \le \varepsilon_B/(1 - \gamma).$$

For one queried input, couple $U \sim Z^\pi$ and $\widehat{U} \sim Z_\psi$ so that $\mathbb{E}|U - \widehat{U}| = W_1(Z^\pi, Z_\psi) \le d$. For every $u > 0$,

$$\Pr(U \le b - u) - d/u \le \Pr(\widehat{U} \le b) \le \Pr(U \le b + u) + d/u.$$

The CDF error is therefore at most $C_F u^\kappa + d/u$. Choosing $u = (d/C_F)^{1/(\kappa+1)}$ for $d > 0$ gives $2C_F^{1/(\kappa+1)} d^{\kappa/(\kappa+1)}$; for $d = 0$ the laws coincide. Taking the supremum over the queried inputs yields

$$\sup_{s,b,a} |\widehat{F}_\psi(b \mid s, b, a) - F^\pi(b \mid s, b, a)| \le 2C_F^{1/(\kappa+1)} d^{\kappa/(\kappa+1)}. \tag{64}$$

An action contrast is a difference of two CDF errors, so a valid bound is

$$\delta = 4C_F^{1/(\kappa+1)} \left(\frac{\varepsilon_B}{1 - \gamma}\right)^{\kappa/(\kappa+1)}. \tag{65}$$

Thus (65) is a valid choice of δ, with

$$C = 4C_F^{1/(\kappa+1)} (1 - \gamma)^{-\kappa/(\kappa+1)}. \tag{66}$$

This proves the unprojected-residual certificate. Under the conditions of Theorem 1, substitution into its bias and mean-square bounds then gives

$$\begin{aligned} \|\mathbb{E}\widehat{g}_{p,\infty} - \nabla p(\theta)\|_2 &\le C_s H_\nu C\, \varepsilon_B^{\kappa/(\kappa+1)}, \\ \mathbb{E}\|\widehat{g}_{p,\infty} - \nabla p(\theta)\|_2^2 &\le \frac{2\sigma_p^2}{L} + 2C_s^2 H_\nu^2 C^2 \varepsilon_B^{2\kappa/(\kappa+1)} + \frac{C_s^2 K_\nu}{4LM}. \end{aligned}$$

The residual and CDF bounds apply to each fitted critic satisfying the almost-sure premise; the estimator bounds additionally require all sampling conditions of Theorem 1.

Projected-residual case. For the projected residual in Proposition 3, the standard quantile-projection bound gives

$$\overline{W}_1(Z_\psi, \mathcal{T}^\pi Z_\psi) \le \varepsilon_{\mathrm{proj}} + R_\star/N.$$

The same contraction argument gives

$$\overline{W}_1(Z_\psi, Z^\pi) \le \frac{\varepsilon_{\mathrm{proj}} + R_\star/N}{1-\gamma}. \tag{67}$$

Combining this estimate with (64) bounds CDF and action-contrast errors at the learned critic equilibrium.

The projected-residual statement follows by the same substitution. □

Finite-rollout estimation. We now retain only the first T decisions, while keeping $\nabla_\theta p(\theta)$ as the target for the original infinite-horizon problem. The following proposition combines the omitted gradient tail with critic and sampling errors in the retained estimator.

For L independent trajectories, the retained estimator is

$$\widehat{g}_{p,T} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1} \widehat{A}_{p,\ell,t}\nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}). \tag{68}$$

For the estimation bounds, allow deterministic time-dependent bounds $\delta_t \ge 0$ and general fitted return supports. The action-contrast condition is that, almost surely over the rollout batch and critic fitting, at every stored (s,b) at time $t < T$,

$$\sup_{a,a' \in \mathrm{supp}\,\pi(\cdot|s,b)} \left| \begin{array}{l} \widehat{F}_\psi(b \mid s,b,a) - F^\pi(b \mid s,b,a) \\ \quad - \widehat{F}_\psi(b \mid s,b,a') + F^\pi(b \mid s,b,a') \end{array} \right| \le \delta_t. \tag{69}$$

Proposition 4 (Finite-rollout risk-gradient estimation)**.** *Under Proposition 1, the standing score regularity, and Assumption 1, for every integer $T \ge t_0$,*

$$\left\| \nabla_\theta p(\theta) - \sum_{t=0}^{T-1} \mathbb{E}[A^\pi_{p,t} s_t] \right\|_2 \le \frac{C_s C_q R_\star^\nu \gamma^{\nu T}}{1-\gamma^\nu}, \tag{70}$$

$$\left(\mathbb{E}\left\| \sum_{t=0}^{T-1} A^\pi_{p,t} s_t \right\|_2^2 \right)^{1/2} \le C_s H_\nu. \tag{71}$$

Moreover, truncating the return event itself gives

$$\Pr(\mathbf{1}\{P_T \le q\} \ne \mathbf{1}\{G^\pi \le q\}) \le C_q(\gamma^T R_\star)^\nu. \tag{72}$$

For the estimator in (68), suppose in addition that the standing regularity conditions and (69) hold. Freeze the fitted critic before drawing the M alternative actions per decision. Conditional on the batch and critic-fitting randomness, draw these actions independently across trajectories, decisions, and replicates from $\pi(\cdot \mid S_{\ell,t}, B_{\ell,t})$. Then

$$\begin{aligned} \mathbb{E}\|\widehat{g}_{p,T} - \nabla_\theta p(\theta)\|_2^2 \le \frac{2}{L}\mathbb{E}\left\| \sum_{t=0}^{T-1} \left(\begin{array}{l} A^\pi_{p,t}\nabla_\theta \log \pi_\theta(A_t \mid S_t, B_t) \\ \quad - \mathbb{E}\big[A^\pi_{p,t}\nabla_\theta \log \pi_\theta(A_t \mid S_t, B_t)\big] \end{array} \right) \right\|_2^2 \\ + 2C_s^2 \left(\sum_{t=0}^{T-1} \delta_t \right)^2 + \frac{C_s^2 T}{4LM} + \frac{2C_s^2 C_q^2 \gamma^{2\nu T}}{(1-\gamma^\nu)^2} R_\star^{2\nu}. \end{aligned} \tag{73}$$

The variance $\mathbb{E}\|\widehat{g}_{p,T} - \mathbb{E}\widehat{g}_{p,T}\|_2^2$ is bounded by the same expression with its final, truncation term removed. The bias satisfies

$$\|\mathbb{E}\widehat{g}_{p,T} - \nabla_\theta p(\theta)\|_2 \le C_s \sum_{t=0}^{T-1} \delta_t + \frac{C_s C_q \gamma^{\nu T}}{1-\gamma^\nu} R_\star^\nu. \tag{74}$$

Proof. We first bound the omitted gradient tail and then analyze the retained estimator.

Truncation error. We use the localization established in Theorem 1, sum the omitted first moments, and then apply the deterministic return-tail bound. Recall $E_t = \{|P_t - q| \le \gamma^t R_\star\}$ and $|A^\pi_{p,t}| \le \mathbf{1}_{E_t}$. Because E_t is determined by X_t, conditional Cauchy–Schwarz gives, for $t \ge t_0$,

$$\mathbb{E}\|A^\pi_{p,t} s_t\|_2 \le \mathbb{E}[\mathbf{1}_{E_t}\mathbb{E}(\|s_t\|_2 \mid X_t)] \le C_s \Pr(E_t) \le C_s C_q R_\star^\nu \gamma^{\nu t},$$
$$\mathbb{E}\|A^\pi_{p,t} s_t\|_2^2 \le \mathbb{E}[\mathbf{1}_{E_t}\mathbb{E}(\|s_t\|_2^2 \mid X_t)] \le C_s^2 \Pr(E_t).$$

The corresponding bounds for $t < t_0$ are C_s and C_s^2. The gradient identity in Proposition 1 therefore implies

$$\left\| \nabla_\theta p(\theta) - \sum_{t<T} \mathbb{E}[A^\pi_{p,t} s_t] \right\|_2 \le \sum_{t=T}^{\infty} \mathbb{E}\|A^\pi_{p,t} s_t\|_2 \le \frac{C_s C_q R_\star^\nu \gamma^{\nu T}}{1-\gamma^\nu},$$
$$\left(\mathbb{E} \left\| \sum_{t<T} A^\pi_{p,t} s_t \right\|_2^2 \right)^{1/2} \le \sum_{t<T} (\mathbb{E}\|A^\pi_{p,t} s_t\|_2^2)^{1/2} \le C_s H_\nu.$$

Finally, $|G^\pi - P_T| \le \gamma^T R_\star$, so disagreement of the two indicators requires $|P_T - q| \le \gamma^T R_\star$. Assumption 1 gives (72).

Finite-rollout estimation error. We next bound the fitted-critic error, isolate independent alternative-action noise, and combine these errors with the gradient-tail bound.

To isolate the two operations being approximated, introduce two intermediate estimators. The superscript true means that the true risk advantage is used; the superscript ψ means that the fitted CDF is used but its policy-action average is integrated exactly:

$$\widehat{g}^{\mathrm{true}}_{p,T} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1} A^\pi_{p,\ell,t} \nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}), \tag{75}$$

$$\widehat{g}^{\psi}_{p,T} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}) \begin{pmatrix} \widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, A_{\ell,t}) \\ \quad - \mathbb{E}_{a\sim\pi}\left[\widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right] \end{pmatrix}. \tag{76}$$

These intermediate estimators separate critic approximation from alternative-action sampling in the proof.

The fitted critic can depend on the actor batch. Replacing the true CDF by its fitted counterpart in both terms of (6) changes the advantage by

$$\widehat{F}_\psi(b \mid s,b,a) - F^\pi(b \mid s,b,a) - \mathbb{E}_{a'\sim\pi(\cdot|s,b)}\left[\widehat{F}_\psi(b \mid s,b,a') - F^\pi(b \mid s,b,a')\right]. \tag{77}$$

The absolute value of this centered difference is at most δ_t under (69): integrate the pairwise error difference against $\pi(\cdot \mid s,b)$. Action centering and (69) give, for every realized batch and fitted critic,

$$\|\widehat{g}^{\psi}_{p,T} - \widehat{g}^{\mathrm{true}}_{p,T}\|_2 \le \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1} \delta_t \|\nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t})\|_2.$$

Because the bounds δ_t are deterministic, Minkowski's inequality and (15) imply

$$\left(\mathbb{E}\|\widehat{g}^{\psi}_{p,T} - \widehat{g}^{\mathrm{true}}_{p,T}\|_2^2\right)^{1/2} \le C_s \sum_{t=0}^{T-1} \delta_t. \tag{78}$$

This step uses no independence between the fitted critic and the executed actions.

Mean-zero alternative-action noise. Condition on the complete rollout batch and all critic-fitting randomness. Each alternative is drawn after this conditioning, so

$$\mathbb{E}[\widehat{g}_{p,T} - \widehat{g}^{\psi}_{p,T} \mid \text{batch and fitted critic}] = 0.$$

Conditional independence across all alternative draws gives the exact conditional second moment

$$\frac{1}{L^2 M}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1}\|\nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t})\|_2^2 \cdot \mathrm{Var}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[\widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right]. \quad (79)$$

A scalar in $[0,1]$ has variance at most $1/4$. Averaging over the batch therefore yields

$$\mathbb{E}\|\widehat{g}_{p,T} - \widehat{g}^{\psi}_{p,T}\|_2^2 \le \frac{C_s^2 T}{4LM}. \quad (80)$$

The conditionally mean-zero difference is orthogonal in mean square to every square-integrable function of the batch and fitted critic. In particular, it is orthogonal to both the oracle sampling error and the critic approximation error.

Combining the three estimators. The exact decomposition is

$$\widehat{g}_{p,T} - \nabla_\theta p(\theta) = (\widehat{g}_{p,T} - \widehat{g}^{\psi}_{p,T}) + (\widehat{g}^{\psi}_{p,T} - \widehat{g}^{\mathrm{true}}_{p,T}) + (\widehat{g}^{\mathrm{true}}_{p,T} - \nabla_\theta p(\theta)). \quad (81)$$

The last estimator is an average of L independent oracle trajectory contributions. Its second moment about the infinite-horizon gradient is its trajectory variance divided by L, plus the squared bias in (70). Apply $\|x+y\|_2^2 \le 2\|x\|_2^2 + 2\|y\|_2^2$ to the last two terms in (81), and add the orthogonal alternative-action variance. Equations (78) and (80) give (73). For the variance statement, center the two remaining random vectors before using the same inequality; centering removes the deterministic tail and subtracts the squared mean from the critic-error second moment. For the bias statement, take expectations in (81), use the conditional mean-zero property, and apply (78) and (70). The bound in (71) controls the trajectory term uniformly in T.

A data-dependent action-contrast refinement. The proof actually controls the fitted, centered CDF error on the realized batch. Specifically,

$$\begin{aligned}
&\mathbb{E}\|\widehat{g}^{\psi}_{p,T} - \widehat{g}^{\mathrm{true}}_{p,T}\|_2^2 \\
&= \mathbb{E}\Bigg\| \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}) \\
&\qquad \cdot \begin{bmatrix} \widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, A_{\ell,t}) - F^\pi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, A_{\ell,t}) \\ -\,\mathbb{E}_{a\sim\pi(\cdot|S_{\ell,t},B_{\ell,t})}\left[\widehat{F}_\psi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a) - F^\pi(B_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}, a)\right] \end{bmatrix} \Bigg\|_2^2.
\end{aligned} \quad (82)$$

Twice this explicit quantity can replace $2C_s^2(\sum_{t<T}\delta_t)^2$ in both the mean-square and variance bounds. Its square root can replace the critic term in the bias bound. This retains nonuniform, data-dependent errors and cancellations without imposing a uniform approximation-error envelope.

If the CDF error is constant across actions at a stored state, its executed-action value and the average of its values on any M alternatives cancel for every realization. This cancellation is pathwise for every sampled set of alternatives. The mean-zero action-sampling term separately quantifies the approximation of the integrated policy average.

□

The gradient-tail bound retains the infinite-horizon CDF in earlier decisions. Changing the critic target to a return stopped at an artificial cutoff is a separate approximation and must also be included in its action-contrast error.

To connect this risk estimate to the actor update, use the corresponding reward estimate with the time-zero discount in $\nabla_\theta J(\theta)$:

$$\widehat{g}_{r,T} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{t=0}^{T-1}\gamma^t \widehat{A}_{r,\ell,t}\nabla_\theta \log \pi_\theta(A_{\ell,t} \mid S_{\ell,t}, B_{\ell,t}). \quad (83)$$

Under the conditions of Proposition 4, for fixed $\lambda \ge 0$ and a square-integrable reward estimate, the L^2 triangle inequality gives the Lagrangian error bound without requiring independence between

the reward and risk estimates:

$$\begin{aligned}&\left(\mathbb{E}\left\|\widehat{g}_{r,T}-\lambda\widehat{g}_{p,T}-\nabla_\theta L(\theta,\lambda)\right\|_2^2\right)^{1/2}\\&\leq\left(\mathbb{E}\left\|\widehat{g}_{r,T}-\nabla_\theta J(\theta)\right\|_2^2\right)^{1/2}+\lambda\left(\mathbb{E}\left\|\widehat{g}_{p,T}-\nabla_\theta p(\theta)\right\|_2^2\right)^{1/2}.\end{aligned}\tag{84}$$

B.3 Proof of Theorem 2

The preceding results quantify risk-gradient estimation error for a fixed policy. We now analyze the coupled critic, actor, and multiplier updates. Under the stability and stochastic-approximation conditions below, Theorem 2 establishes joint convergence and constraint satisfaction, with the remaining stationarity error controlled by the equilibrium action-contrast error δ^*.

Assumption 3 (Stable critic, actor, and risk response)**.** (i) For each fixed θ, the mean critic update has a unique interior equilibrium $\psi^*(\theta)$. Starting from any $\psi \in \Psi$, the corresponding mean dynamics converge to $\psi^*(\theta)$, uniformly over θ. The map $\psi^*(\theta)$ is Lipschitz in θ. (ii) For each fixed λ, the mean actor update has a unique interior equilibrium $\theta^*(\lambda)$. Starting from any $\theta \in \Theta$, the corresponding mean dynamics converge to $\theta^*(\lambda)$, uniformly over λ. The mean actor update is continuously differentiable in θ, and its Jacobian with respect to θ is nonsingular at $\theta^*(\lambda)$. (iii) $r(\lambda)$ is nonincreasing. Either $r(0) < 0$, or there exists $\lambda^* \in (0, \lambda_{\max})$ such that $r(\lambda^*) = 0$ and $r'(\lambda^*) < 0$.

The assumption requires stable critic and actor updates when the slower variables are held fixed. The final condition ensures that increasing the multiplier does not increase the equilibrium violation probability and that the original constraint can be reached.

Stochastic-approximation conditions. In addition to Assumptions 1–3, we impose the following conditions.

(i) **Bounded parameter sets.** The critic and policy parameter sets Ψ and Θ are compact and convex with nonempty interiors. The multiplier set is $\Lambda = [0, \lambda_{\max}]$, where $\lambda_{\max} > 0$. All projections are Euclidean.

Projection keeps the parameter sequences bounded. Convexity ensures that each projected update is uniquely defined.

(ii) **Step sizes.** The critic, actor, and multiplier step sizes β_k, a_k, and b_k, respectively, are positive, deterministic, and satisfy

$$\begin{gathered}\sum_{k=0}^{\infty}\beta_k=\sum_{k=0}^{\infty}a_k=\sum_{k=0}^{\infty}b_k=\infty,\qquad\sum_{k=0}^{\infty}(\beta_k^2+a_k^2+b_k^2)<\infty,\\\frac{a_k}{\beta_k}\longrightarrow 0,\qquad\frac{b_k}{a_k}\longrightarrow 0.\end{gathered}\tag{85}$$

The divergent sums give each update an unbounded total step size, while square summability controls accumulated sampling noise. The two ratios make the critic adapt faster than the actor, and the actor adapt faster than the multiplier.

(iii) **Regularity and stability of the mean updates.** The maps h_ψ, g_p, $\nabla_\theta J$, and p are Lipschitz on their respective parameter domains. The actor mean update after critic equilibration,

$$h_\theta(\theta,\lambda) := \nabla_\theta J(\theta) - \lambda g_p(\theta, \psi^*(\theta)),$$

is continuously differentiable on a neighborhood of $\Theta \times \Lambda$.

For a nonempty closed convex set C, let Π_C denote Euclidean projection onto C, and define

$$\mathcal{P}_C(x,v) := \lim_{\eta\downarrow 0}\frac{\Pi_C(x+\eta v)-x}{\eta},$$

where $x \in C$ and v is an update direction. The stability requirements in Assumption 3 concern the projected mean dynamics

$$\dot{\psi} = \mathcal{P}_\Psi(\psi, h_\psi(\psi,\theta)),\qquad \dot{\theta} = \mathcal{P}_\Theta(\theta, h_\theta(\theta,\lambda)),\tag{86}$$

with θ and λ held fixed, respectively. Their specified equilibria are globally asymptotically stable, uniformly over the fixed parameters.

Lipschitz continuity controls how much the mean updates change when the parameters change. The stability condition requires the projected dynamics to converge to the specified equilibria from every allowed initial parameter, including boundary points. Uniformity means that the stability neighborhoods and convergence-time bounds can be chosen independently of the fixed policy or multiplier. Since the equilibria lie in the interiors of their parameter sets, they satisfy the original equations

$$h_\psi(\psi^*(\theta), \theta) = 0, \qquad h_\theta(\theta^*(\lambda), \lambda) = 0.$$

(iv) **Mean errors and sampling noise.** Let $(\mathcal{F}_k)_{k\geq 0}$ be the filtration containing all information available before the update at iteration k. The parameters $(\psi_k, \theta_k, \lambda_k)$ are $\mathcal{F}_k$-measurable. The update estimates admit the decomposition

$$\begin{aligned} \widehat{h}_k &= h_\psi(\psi_k, \theta_k) + u_k^\psi + M_{k+1}^\psi, \\ \widehat{g}_{r,k} - \lambda_k \widehat{g}_{p,k} &= \nabla_\theta J(\theta_k) - \lambda_k g_p(\theta_k, \psi_k) + u_k^\theta + M_{k+1}^\theta, \\ \widehat{p}_k - \alpha &= p(\theta_k) - \alpha + u_k^\lambda + M_{k+1}^\lambda. \end{aligned} \tag{87}$$

Here u_k^ψ, u_k^θ, and u_k^λ are $\mathcal{F}_k$-measurable conditional mean errors satisfying

$$\|u_k^\psi\|_2 + \|u_k^\theta\|_2 + |u_k^\lambda| \xrightarrow{\text{a.s.}} 0.$$

The sampling noises M_{k+1}^ψ, M_{k+1}^θ, and M_{k+1}^λ are $\mathcal{F}_{k+1}$-measurable and satisfy

$$\mathbb{E}[M_{k+1}^\psi \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[M_{k+1}^\theta \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[M_{k+1}^\lambda \mid \mathcal{F}_k] = 0.$$

For a finite constant C_M, assume

$$\mathbb{E}\Big[\|M_{k+1}^\psi\|_2^2 + \|M_{k+1}^\theta\|_2^2 + |M_{k+1}^\lambda|^2 \,\Big|\, \mathcal{F}_k\Big] \leq C_M \qquad \text{a.s.}$$

The conditional mean errors must vanish so that they do not change the limiting mean updates. The sampling noise may remain nonzero at each iteration. Its zero conditional mean and bounded second moments, together with the square-summable step sizes, ensure convergence of the accumulated step-size-weighted noise.

(v) **Risk-gradient estimation.** The direction $g_p(\theta, \psi)$ is the mean of the risk-gradient estimator with policy π_θ and critic parameter ψ held fixed. The conditions of Theorem 1 hold uniformly over the parameters considered.

This allows the risk-gradient bias bound to be applied at the limiting policy and critic. The critic-induced bias $g_p(\theta, \psi) - \nabla_\theta p(\theta)$ is retained in the actor mean update. It is not part of u_k^θ and need not vanish.

Theorem 2 (Joint convergence and constraint satisfaction)**.** *Suppose Assumptions 1–3 and the stochastic-approximation conditions above hold. Then*

$$(\psi_k, \theta_k, \lambda_k) \xrightarrow{\text{a.s.}} (\psi^*, \theta^*, \lambda^*),$$

where

$$\theta^* := \theta^*(\lambda^*), \qquad \psi^* := \psi^*(\theta^*).$$

Moreover,

$$p(\theta^*) \leq \alpha, \qquad \lambda^* \geq 0, \qquad \lambda^*\{p(\theta^*) - \alpha\} = 0. \tag{88}$$

Let $\pi^* := \pi_{\theta^*}$ denote the limiting policy. For an augmented state (s, b) and an action a, define the equilibrium conditional-CDF error by

$$e^*(s, b, a) := \widehat{F}_{\psi^*}(b \mid s, b, a) - F^{\pi^*}(b \mid s, b, a).$$

Let $\delta^* \geq 0$ satisfy

$$\sup_{a, a' \in \operatorname{supp} \pi^*(\cdot|s,b)} |e^*(s, b, a) - e^*(s, b, a')| \leq \delta^*$$

almost surely at every visited augmented state. Then

$$\|\nabla_\theta L(\theta^*, \lambda^*)\|_2 \leq \lambda^* C_s H_\nu \delta^*, \tag{89}$$

where $L(\theta, \lambda) = J(\theta) - \lambda\{p(\theta) - \alpha\}$, and C_s and H_ν are the constants in Theorem 1.

In particular, when $\delta^* = 0$, the limit satisfies the KKT conditions of the original probability-constrained problem.

Proof. The proof proceeds in three steps. We first establish that the critic tracks the current policy and the actor tracks the current multiplier. We then prove convergence of the multiplier, which implies joint convergence of all three parameter sequences. Finally, we derive feasibility and complementarity, and apply Theorem 1 to bound the Lagrangian stationarity error.

Step 1: Critic and actor tracking. Recall that $\psi^*(\theta)$ is the critic equilibrium for a fixed policy parameter θ. Since $a_k/\beta_k \to 0$, the policy is asymptotically frozen on the critic time scale. Compactness gives bounded iterates, and the stated step-size and noise conditions allow the projected stochastic-approximation ODE method to be applied (Borkar, 2008). Together with the uniform stability of the critic dynamics in (86), this gives

$$\|\psi_k - \psi^*(\theta_k)\|_2 \xrightarrow{\text{a.s.}} 0. \tag{90}$$

We next replace the current critic in the actor mean update by its equilibrium. Define

$$\varepsilon_k^\theta := u_k^\theta + \lambda_k \left\{ g_p(\theta_k, \psi^*(\theta_k)) - g_p(\theta_k, \psi_k) \right\}.$$

Let L_g be a Lipschitz constant of g_p with respect to its critic argument. Since $0 \le \lambda_k \le \lambda_{\max}$,

$$\|\varepsilon_k^\theta\|_2 \le \|u_k^\theta\|_2 + \lambda_{\max} L_g \|\psi_k - \psi^*(\theta_k)\|_2 \xrightarrow{\text{a.s.}} 0. \tag{91}$$

Thus the actor recursion becomes

$$\theta_{k+1} = \Pi_\Theta \left[\theta_k + a_k \left\{ h_\theta(\theta_k, \lambda_k) + \varepsilon_k^\theta + M_{k+1}^\theta \right\} \right].$$

For each fixed λ, the actor mean update has the unique interior equilibrium $\theta^*(\lambda)$. The continuous differentiability of h_θ and its nonsingular Jacobian at this equilibrium allow the implicit-function theorem to be applied. Uniqueness joins the local equilibrium branches, so $\theta^*(\lambda)$ is continuously differentiable and Lipschitz on the compact interval Λ.

Since $b_k/a_k \to 0$, the multiplier is asymptotically frozen on the actor time scale. Applying the same tracking result to the uniformly stable projected actor dynamics yields

$$\|\theta_k - \theta^*(\lambda_k)\|_2 \xrightarrow{\text{a.s.}} 0. \tag{92}$$

Step 2: Multiplier and joint convergence. Recall the equilibrium constraint residual

$$r(\lambda) := p(\theta^*(\lambda)) - \alpha.$$

Define

$$\varepsilon_k^\lambda := u_k^\lambda + p(\theta_k) - p(\theta^*(\lambda_k)).$$

Let L_p be a Lipschitz constant of p. By (92),

$$|\varepsilon_k^\lambda| \le |u_k^\lambda| + L_p \|\theta_k - \theta^*(\lambda_k)\|_2 \xrightarrow{\text{a.s.}} 0.$$

The multiplier recursion therefore becomes

$$\lambda_{k+1} = \Pi_\Lambda \left[\lambda_k + b_k \left\{ r(\lambda_k) + \varepsilon_k^\lambda + M_{k+1}^\lambda \right\} \right].$$

Its limiting projected mean dynamics are

$$\dot\lambda = \mathcal{P}_\Lambda(\lambda, r(\lambda)) = \begin{cases} \max\{r(0), 0\}, & \lambda = 0, \\ r(\lambda), & 0 < \lambda < \lambda_{\max}, \\ \min\{r(\lambda_{\max}), 0\}, & \lambda = \lambda_{\max}. \end{cases} \tag{93}$$

We consider the two cases in Assumption 3(iii).

Case 1: $r(0) < 0$. Set $\lambda^* := 0$. Since r is nonincreasing,

$$r(\lambda) \le r(0) < 0, \qquad \lambda \in \Lambda.$$

Hence the drift is negative whenever $\lambda > 0$, and the projection makes $\lambda = 0$ an equilibrium. For the Lyapunov function $V(\lambda) := \lambda^2/2$,

$$\dot V(\lambda) = \lambda r(\lambda) < 0, \qquad \lambda > 0.$$

Therefore $\lambda^* = 0$ is globally asymptotically stable.

Case 2: $r(\lambda^*) = 0$ *and* $r'(\lambda^*) < 0$ *for some* $\lambda^* \in (0, \lambda_{\max})$. By the derivative condition, there exists $\eta > 0$ such that

$$\frac{r(\lambda) - r(\lambda^*)}{\lambda - \lambda^*} < 0 \qquad \text{whenever } 0 < |\lambda - \lambda^*| < \eta,$$

where η is chosen so that $(\lambda^* - \eta, \lambda^* + \eta) \subset \Lambda$. Since $r(\lambda^*) = 0$, the residual is positive immediately to the left of λ^* and negative immediately to its right. Monotonicity extends these signs to the entire interval:

$$r(\lambda) > 0 \quad \text{for } \lambda < \lambda^*, \qquad r(\lambda) < 0 \quad \text{for } \lambda > \lambda^*.$$

Thus the root is unique and the drift points inward at both endpoints of Λ. For $V(\lambda) := (\lambda - \lambda^*)^2/2$,

$$\dot{V}(\lambda) = (\lambda - \lambda^*) r(\lambda) < 0, \qquad \lambda \neq \lambda^*.$$

Therefore λ^* is globally asymptotically stable in this case as well.

In both cases, the multiplier recursion has a globally asymptotically stable projected mean equilibrium. The stochastic-approximation convergence result gives

$$\lambda_k \xrightarrow{\text{a.s.}} \lambda^*. \tag{94}$$

Using (92), (94), and continuity of $\theta^*(\cdot)$,

$$\|\theta_k - \theta^*\|_2 \leq \|\theta_k - \theta^*(\lambda_k)\|_2 + \|\theta^*(\lambda_k) - \theta^*(\lambda^*)\|_2 \xrightarrow{\text{a.s.}} 0. \tag{95}$$

Similarly, (90) and continuity of $\psi^*(\cdot)$ imply

$$\|\psi_k - \psi^*\|_2 \leq \|\psi_k - \psi^*(\theta_k)\|_2 + \|\psi^*(\theta_k) - \psi^*(\theta^*)\|_2 \xrightarrow{\text{a.s.}} 0. \tag{96}$$

Consequently,

$$(\psi_k, \theta_k, \lambda_k) \xrightarrow{\text{a.s.}} (\psi^*, \theta^*, \lambda^*). \tag{97}$$

Step 3: Feasibility and the stationarity bound. In Case 1,

$$\lambda^* = 0, \qquad p(\theta^*) - \alpha = r(0) < 0.$$

In Case 2,

$$\lambda^* > 0, \qquad p(\theta^*) - \alpha = r(\lambda^*) = 0.$$

Therefore,

$$p(\theta^*) \leq \alpha, \qquad \lambda^* \geq 0, \qquad \lambda^* \{p(\theta^*) - \alpha\} = 0. \tag{98}$$

Since θ^* is an interior actor equilibrium, projection does not alter its equilibrium equation:

$$0 = h_\theta(\theta^*, \lambda^*) = \nabla_\theta J(\theta^*) - \lambda^* g_p(\theta^*, \psi^*).$$

Hence the true Lagrangian gradient satisfies

$$\begin{aligned} \nabla_\theta L(\theta^*, \lambda^*) &= \nabla_\theta J(\theta^*) - \lambda^* \nabla_\theta p(\theta^*) \\ &= \lambda^* g_p(\theta^*, \psi^*) - \lambda^* \nabla_\theta p(\theta^*) \\ &= \lambda^* \left\{ g_p(\theta^*, \psi^*) - \nabla_\theta p(\theta^*) \right\}. \end{aligned} \tag{99}$$

Applying Theorem 1 to the equilibrium policy and critic gives

$$\|g_p(\theta^*, \psi^*) - \nabla_\theta p(\theta^*)\|_2 \leq C_s H_\nu \delta^*.$$

Taking norms in (99),

$$\begin{aligned} \|\nabla_\theta L(\theta^*, \lambda^*)\|_2 &= \lambda^* \left\| g_p(\theta^*, \psi^*) - \nabla_\theta p(\theta^*) \right\|_2 \\ &\leq \lambda^* C_s H_\nu \delta^*. \end{aligned} \tag{100}$$

Thus

$$\|\nabla_\theta L(\theta^*, \lambda^*)\|_2 \leq \lambda^* C_s H_\nu \delta^*. \tag{101}$$

If $\delta^* = 0$, then $\nabla_\theta L(\theta^*, \lambda^*) = 0$. Together with (98), this establishes stationarity, primal feasibility, dual feasibility, and complementary slackness. These are the KKT conditions of the original probability-constrained problem. □

C DETAILED EXPERIMENTAL SETTING

Downside-risk allocation environment. We construct a continuous-control environment with horizon $H = 20$ and $\gamma = 0.99$. At decision time t, the action $A_t \in [0, 1]$ represents the level of risk exposure. After the action is selected, the environment independently samples

$$D_t \sim \text{Bernoulli}(0.025), \qquad \xi_t \sim \text{Uniform}[-1, 1],$$

and generates the raw reward

$$R_t = 1 + 2A_t - 50A_tD_t + 0.02\xi_t. \tag{102}$$

The random variables (D_t, ξ_t) are independent across time. The shock variable D_t affects only the current reward and is not a persistent hidden state.

Conditional on $A_t = a$, the reward distribution is the mixture

$$0.975\,\text{Uniform}[0.98 + 2a,\ 1.02 + 2a] + 0.025\,\text{Uniform}[0.98 - 48a,\ 1.02 - 48a].$$

Hence

$$\mathbb{E}[R_t \mid A_t = a] = 1 + 0.75a,$$

so a larger action always improves the conditional mean reward, while simultaneously increasing exposure to the downside component.

The discounted episode return is

$$G = \sum_{t=0}^{H-1} \gamma^t R_t,$$

and the optimization problem is

$$\max_{\pi}\ \mathbb{E}_\pi[G] \qquad \text{subject to} \qquad \Pr_\pi(G \le q) \le \alpha, \tag{103}$$

with $q = 16$ and $\alpha = 0.20$.

Let

$$P_t = \sum_{j=0}^{t-1} \gamma^j R_j$$

denote the return accumulated before action A_t. The remaining return requirement is

$$B_t = \frac{q - P_t}{\gamma^t}, \qquad B_{t+1} = \frac{B_t - R_t}{\gamma}.$$

The policy therefore observes the augmented state

$$X_t = (t/H, B_t).$$

The episode terminates only after the 20th decision; reaching or missing the threshold earlier does not terminate the trajectory, and no additional terminal reward or penalty is introduced.

The environment has an analytically transparent mean–risk trade-off. Writing

$$C_H = \sum_{t=0}^{H-1} \gamma^t = 18.2093,$$

the expected return of any non-anticipating policy satisfies

$$\mathbb{E}_\pi[G] = C_H + 0.75 \sum_{t=0}^{H-1} \gamma^t \mathbb{E}_\pi[A_t].$$

Thus the unconstrained mean-optimal policy is $A_t \equiv 1$, with expected return 31.8663. For this policy, any downside shock causes the final return to fall below $q = 16$, giving

$$\Pr(G \le 16) = 1 - 0.975^{20} \approx 0.3973,$$

which substantially violates the prescribed 20% limit. In contrast, the conservative policy $A_t \equiv 0$ has zero violation probability. The environment therefore admits feasible policies while ensuring that mean maximization alone does not solve the constrained problem.

Numerical scaling. For neural-network training, we divide rewards, return thresholds, and remaining-return requirements by 20:

$$r_t = R_t/20, \qquad b_t = B_t/20, \qquad q_{\text{int}} = 0.8.$$

The policy input is therefore

$$x_t = (t/H, b_t).$$

This rescaling changes only numerical units:

$$\{G \le 16\} = \left\{ \sum_{t=0}^{H-1} \gamma^t r_t \le 0.8 \right\},$$

and hence does not change either the feasible policy set or the violation-probability constraint. All reported returns and quantiles are converted back to the original reward units.

Baseline cost representation. QCPO and TQPO are formulated using upper-tail cumulative costs. For both methods, we use the nonnegative stage cost

$$c_t = \frac{3.02 - R_t}{20}.$$

Because the horizon is fixed and the same discount factor is used, the cumulative discounted cost satisfies

$$C = \sum_{t=0}^{H-1} \gamma^t c_t = \frac{3.02 C_H - G}{20}.$$

Consequently,

$$G \le 16 \quad \Longleftrightarrow \quad C \ge d, \qquad d = \frac{3.02 C_H - 16}{20} \approx 1.949605.$$

The original lower-20% return constraint is therefore represented as an upper-80% cost-quantile constraint for QCPO and TQPO.

Training and evaluation protocol. For each method, we train ten independent runs using seeds $601, \ldots, 610$. Each run consists of 400 training iterations. At each iteration, 256 complete trajectories of length 20 are collected, giving

$$400 \times 256 \times 20 = 2{,}048{,}000$$

environment transitions per run. The first 20 iterations are used as critic warm-up iterations and do not update the actor.

For monitoring, we evaluate the policy on $3,000$ episodes at initialization and every 20 training iterations. Evaluation uses random streams separate from training. For each training seed, the same validation random seeds are reused across checkpoints. Final testing uses a separate random seed and $100,000$ episodes per trained policy. These validation rollouts are not used for multiplier updates, early stopping, hyperparameter selection, or checkpoint selection. The final checkpoint at iteration 400 is always reported. Each final policy is evaluated on $100{,}000$ independent episodes.

For each training seed, we compute the mean discounted return, the empirical 20% return quantile, and

$$\widehat{p} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbf{1}\{G_i \le 16\}.$$

A run is called empirically feasible when $\widehat{p} \le 0.20$. All summary statistics are first computed separately for each trained policy and then averaged over the ten training seeds. Reported "$\pm$" values denote the sample standard deviation across training seeds rather than Monte Carlo standard errors from pooled test episodes. Learning curves similarly show the mean across ten training seeds with one sample standard deviation band.

Details for $H = 40$. For the longer-horizon experiment, we retain the same per-step reward distribution and discount factor $\gamma = 0.99$, but set $H = 40$. Let $C_H = \sum_{t=0}^{H-1} \gamma^t$. To preserve the threshold per unit of discounted horizon mass, we set

$$q_{40} = 16 \, \frac{C_{40}}{C_{20}} = 29.086511.$$

Each method is trained on ten independent seeds (721–730) for 800 updates, with 256 complete length-40 trajectories per update, yielding $800 \times 256 \times 40 = 8.192$ million training transitions per run. We always report the final update, without checkpoint selection, and evaluate every learned policy on 100,000 independent episodes.

D IMPLEMENTATION DETAILS

Common policy and optimization setup. All three methods use the same policy parameterization. The actor is a two-layer MLP with 32 hidden units per layer and $\tanh$ activations. It outputs the mean of a latent Gaussian variable,

$$Z_t \sim \mathcal{N}(\mu_\theta(x_t), \sigma_\theta^2),$$

and the environment action is

$$A_t = \mathrm{sigmoid}(Z_t).$$

The log standard deviation is a learned state-independent parameter, initialized at -0.5 and clipped to $[-3, 0.5]$ when used. PPO ratios are computed from the Gaussian densities of the stored latent variables; the parameter-independent sigmoid Jacobian cancels from the new/old policy ratio.

A separate reward-value network consists of two 64-unit $\tanh$ hidden layers and a scalar output. The reward advantage is

$$\widehat{A}_t^r = g_t - V_\phi(x_t),$$

where

$$g_t = \sum_{j=t}^{H-1} \gamma^{j-t} r_j$$

is the observed numerical suffix return and the value prediction is taken before fitting the value network on the current batch. The time-zero reward contribution used by the actor is therefore $\gamma^t \widehat{A}_t^r$.

Each iteration uses three critic epochs with minibatches of size 1024, and at most four actor epochs with minibatches of size 2560. The PPO clipping parameter is 0.2. After each actor epoch, we compute the sample approximation

$$\widehat{\mathrm{KL}} = \widehat{\mathbb{E}}[\exp(\Delta_t) - 1 - \Delta_t], \qquad \Delta_t = \log \pi_\theta(A_t \mid x_t) - \log \pi_k(A_t \mid x_t),$$

and stop the remaining actor epochs if this quantity exceeds 0.02. Actor and critic gradient norms are clipped at 1 and 10, respectively.

Distributional critic. Our return-distribution critic is a two-layer MLP with 64 hidden units per layer and $N = 64$ outputs. The quantile levels are

$$\tau_i = \frac{i - \frac{1}{2}}{N}, \qquad i = 1, \ldots, N.$$

The critic receives (x_t, A_t) as input and predicts $z_{\psi,i}(x_t, A_t)$.

For each nonterminal transition, we sample

$$A_t' \sim \pi_k(\cdot \mid x_{t+1})$$

and construct the distributional target atoms

$$Y_{t,j} = r_t + \gamma z_{\bar{\psi},j}(x_{t+1}, A_t'), \qquad j = 1, \ldots, N.$$

At true termination, $Y_{t,j} = r_t$. The target network is held fixed within each critic epoch and is hard-updated from the online critic at the end of the epoch.

We minimize the full pairwise quantile-regression loss

$$\mathcal{L}_{\mathrm{dist}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{\tau_i}\left(Y_{t,j} - z_{\psi,i}(x_t, A_t)\right),$$

where ρ_τ is the pinball loss. Thus the target atoms represent the complete one-step Bellman target distribution; the implementation does not assume that corresponding quantiles add under Bellman recursion.

CDF evaluation and policy-average risk. After critic fitting, the predicted quantiles are sorted and the CDF is evaluated at the current remaining-return requirement using linear interpolation between adjacent predicted quantiles. Values below the smallest predicted quantile are assigned probability zero, and values above the largest predicted quantile are assigned probability one.

For a state $x = (s, b)$, define the fitted policy-average violation probability

$$\widehat{\bar{F}}_\psi(x) = \frac{1}{M}\sum_{m=1}^{M} \widehat{F}_\psi(b \mid x, \widetilde{A}_m), \qquad \widetilde{A}_m \overset{\text{i.i.d.}}{\sim} \pi_k(\cdot \mid x),$$

with $M = 4$. Fresh auxiliary actions are drawn at both x_t and, for nonterminal transitions, the observed next augmented state x_{t+1}. These actions are used only to query the fitted critic; no counterfactual environment transition is generated.

Observed-transition risk credit. The risk signal used in the current implementation is

$$\widehat{U}_t^p = \begin{cases} \widehat{\bar{F}}_\psi(x_{t+1}) - \widehat{\bar{F}}_\psi(x_t), & d_t = 0, \\ \mathbf{1}\{b_{t+1} \geq 0\} - \widehat{\bar{F}}_\psi(x_t), & d_t = 1, \end{cases} \tag{104}$$

where d_t denotes true episode termination. No additional factor γ is applied to the probability continuation; discounting is already incorporated through the recursion for b_{t+1}.

With the true conditional risk function, the conditional expectation of this sampled transition credit equals the action risk advantage. With a fitted distributional critic and finitely many auxiliary actions, the practical estimator can contain both critic error and Monte Carlo noise.

Actor and multiplier updates. Before updating the actor, we freeze the reward advantages, risk credits, multiplier λ_k, and the common scale

$$c_k = \max\left\{\mathrm{SD}\left(\gamma^t \widehat{A}_t^r - \lambda_k \widehat{U}_t^p\right), 0.01\right\}.$$

The actor maximizes

$$\frac{1}{c_k}\widehat{\mathbb{E}}\left[\gamma^t \min\{\rho_t \widehat{A}_t^r, \bar{\rho}_t \widehat{A}_t^r\} - \lambda_k \max\{\rho_t \widehat{U}_t^p, \bar{\rho}_t \widehat{U}_t^p\}\right],$$

where

$$\bar{\rho}_t = \mathrm{clip}(\rho_t, 0.8, 1.2).$$

The reward and risk terms are clipped separately and are not separately standardized.

The risk multiplier is updated once per rollout batch from the observed episode-level violation fraction,

$$e_k = \widehat{p}_k - 0.20,$$

using

$$m_{k+1} = 0.8 m_k + 0.2 e_k,$$

and

$$\lambda_{k+1} = \mathrm{clip}\left[\lambda_k + 0.2 e_k + 5(m_{k+1} - m_k), 0, 20\right].$$

We initialize $\lambda_0 = m_0 = 0$. The PI recursion is already active during the first 20 critic warm-up iterations, although the actor itself is held fixed during this period.

The actor learning rate is 10^{-3} and the reward-value and distributional-critic learning rates are both 2×10^{-3}. We use Adam throughout and do not decay the learning rates.

QCPO implementation. Our QCPO baseline follows the core quantile-constrained update of Jung et al. (2022) within the common PPO rollout scaffold. The cost critic predicts 25 positive cost quantiles and the parameters of a Weibull tail model. The quantiles are trained using one-step distributional targets and the full pairwise quantile-Huber loss with $\kappa = 1$. The upper-tail quantiles are additionally used to fit the Weibull tail, whose fitted density enters the quantile-gradient correction. Per-step quantile advantages are propagated using a trace parameter 0.97. With 25 midpoint quantiles, the implementation uses the stored risk index corresponding to $\tau = 0.78$ as the discrete approximation to the target 0.8 level.

The multiplier is updated from the empirical 80% quantile of the most recent discounted episode costs:

$$\lambda \leftarrow \max\{0, \lambda + 10(\widehat{Q}_{0.8}(C) - d)\}.$$

The actor objective retains separately clipped reward and risk terms and uses $1 + \lambda$ as its normalization. Both actor and critic use learning rate 10^{-4}.

TQPO implementation. Our TQPO baseline follows the sampled tilted-quantile update of Li et al. (2025). A scalar cost-value network is trained with an MSE objective, and cost advantages are computed using GAE with trace parameter 0.97. The resulting cost-return samples are stored in a step-level replay buffer and used to estimate the empirical upper-80% threshold $\widehat{q}_{0.8}^{\text{step}}$. The per-step risk signal is

$$\widehat{U}_t^{\text{TQPO}} = \frac{0.1}{0.2}\mathbf{1}\{\widehat{C}_t \geq \widehat{q}_{0.8}^{\text{step}}\}.$$

A separate buffer of recent complete-episode discounted costs is used to estimate the episode-level cost quantile for the multiplier update. Following the public implementation, the update is tilted according to the fraction of recent quantile estimates falling below the prescribed cost threshold.

The initial multiplier is 0.5 and the multiplier step size is 10. Actor and critic learning rates are both 10^{-4}.

Baseline implementation scope. QCPO and TQPO are source-derived core implementations embedded in the same lightweight PPO rollout scaffold used for this synthetic study, rather than verbatim executions of the authors' full recurrent `rlpyt` training pipelines. We preserve their baseline-specific risk estimators, multiplier rules, and normalization schemes. Hyperparameters were selected in a separate pilot run before the ten reported evaluation seeds and were not retuned on the reported runs.

E Supplementary experiments: cost-quantile constraints

The formulation in the main text constrains the lower tail of the same return whose expectation is maximized. Distributional credit assignment also applies when performance and risk are measured by separate reward and cost signals. We first describe this extension and then compare DCA with QCPO and TQPO on SimpleButton, a robot-navigation task used by Jung et al. (2022) to study cumulative-cost constraints.

E.1 Extension to cumulative-cost constraints

We retain the infinite-horizon discounted formulation of Section 3.1 and introduce a separate, uniformly bounded nonnegative stage cost C_t. With the same discount factor $\gamma \in (0, 1)$, define

$$G^\pi = \sum_{t=0}^{\infty} \gamma^t R_t, \qquad Z^\pi = \sum_{t=0}^{\infty} \gamma^t C_t.$$

The objective remains expected-return maximization, while the constraint now limits the upper tail of cumulative cost:

$$\max_\pi \ \mathbb{E}[G^\pi] \qquad \text{subject to} \qquad Q_{1-\alpha}(Z^\pi) \leq d. \tag{105}$$

Here d is a prescribed cost budget and α is the allowed violation probability. With the lower-quantile convention used in the main text, this constraint is equivalent to

$$p_c(\theta) := \Pr(Z^{\pi_\theta} > d) \leq \alpha. \tag{106}$$

The strict inequality excludes trajectories whose cost is exactly d; the equivalence therefore also holds for discrete costs. Writing $J(\theta) = \mathbb{E}[G^{\pi_\theta}]$ as before, we use $L_c(\theta, \lambda) = J(\theta) - \lambda\{p_c(\theta) - \alpha\}$. Thus, the mean objective and primal–dual structure are unchanged.

Let $P_t^c = \sum_{j=0}^{t-1} \gamma^j C_j$ denote the accumulated discounted cost and define the remaining cost budget

$$B_t^c = \frac{d - P_t^c}{\gamma^t}, \qquad B_{t+1}^c = \frac{B_t^c - C_t}{\gamma}, \qquad B_0^c = d. \tag{107}$$

For the future cost $Z_t^\pi = \sum_{j=0}^{\infty} \gamma^j C_{t+j}$,

$$Z^\pi = P_t^c + \gamma^t Z_t^\pi, \qquad \{Z^\pi > d\} = \{Z_t^\pi > B_t^c\}.$$

As in the main text, the budget recursion defines an augmented Markov state $X_t^c = (S_t, B_t^c)$, on which the policy is stationary. Let $F_c^\pi(z \mid s, b, a) = \Pr_\pi(Z_t^\pi \le z \mid S_t = s, B_t^c = b, A_t = a)$ be the conditional future-cost CDF. The action's violation probability is its upper-tail probability,

$$u_c^\pi(s, b, a) = 1 - F_c^\pi(b \mid s, b, a).$$

Subtracting the policy average gives the cost-risk advantage

$$A_{p_c}^\pi(s, b, a) = u_c^\pi(s, b, a) - \mathbb{E}_{\widetilde{a}\sim\pi(\cdot|s,b)}[u_c^\pi(s, b, \widetilde{a})]. \tag{108}$$

Under the differentiability and absolute-summability conditions of Proposition 1, applied to $Y_c = \mathbf{1}\{Z^{\pi_\theta} > d\}$, the same conditional-expectation argument yields

$$\nabla_\theta p_c(\theta) = \sum_{t=0}^{\infty} \mathbb{E}\big[A_{p_c}^{\pi_\theta}(S_t, B_t^c, A_t) \nabla_\theta \log \pi_\theta(A_t \mid S_t, B_t^c)\big]. \tag{109}$$

Hence the density-free credit-assignment construction carries over: we learn conditional future-cost distributions through distributional Bellman updates and evaluate their upper tails at the remaining budget. The reward advantage continues to optimize expected reward, while the cost-risk advantage controls the separate trajectory-level constraint.

The transition-based construction also follows the argument in Appendix A. Let $\bar{u}_c^\pi(x) = \mathbb{E}_{a\sim\pi(\cdot|x)}[u_c^\pi(x, a)]$. Its fitted counterpart gives the sampled risk credit

$$\widehat{U}_t^c = \widehat{\bar{u}}_c(X_{t+1}^c) - \widehat{\bar{u}}_c(X_t^c). \tag{110}$$

With the true conditional probabilities, its conditional expectation given (X_t^c, A_t) equals (108). There is no additional discount factor on the probability continuation: discounting is already incorporated in (107).

E.2 SimpleButton setting and implementation

Environment and constraint. We consider SimpleButtonEnv from the QCPO study (Jung et al., 2022), referred to as SimpleButton in this appendix. The task was originally implemented using Safety Gym; we use a Safety-Gymnasium implementation (Ji et al., 2023) with MuJoCo simulation. A point robot must repeatedly reach designated target buttons while avoiding hazard regions and contacts with the non-target button. The layout contains two fixed buttons and three fixed hazard regions. When the robot performs an action at time t, it receives a reward $\big\{\|p_{t+1} - p_{\text{goal}}\|_2 - \|p_t - p_{\text{goal}}\|_2\big\} + \mathbf{1}_{\text{goal reached}}$, where p_t is the position (x, y) of the robot at time t and p_{goal} is the current goal position at time t. It also receives a cost $+1$ if the robot touches a non-goal object (the non-goal button, a hazard, or a gremlin) and 0 otherwise.

Each episode lasts $H = 1000$ steps. The task therefore separates goal-directed performance from safety: reward encourages reaching targets, while cumulative cost records unsafe interactions over the episode. For this episodic experiment, we use the finite-horizon, undiscounted version of the objective, with episode return and cost

$$G = \sum_{t=0}^{999} R_t, \qquad Z = \sum_{t=0}^{999} C_t,$$

and impose

$$Q_{0.8}(Z) \le 15 \qquad \Longleftrightarrow \qquad \Pr(Z > 15) \le 0.20.$$

Unlike the synthetic experiment, the cost signal here is supplied by the environment and is not an affine transformation of reward. Mean cost is reported as a descriptive statistic; it does not determine whether the quantile constraint is satisfied.

The remaining budget is therefore $B_t^c = d - \sum_{j=0}^{t-1} C_j$, with update $B_{t+1}^c = B_t^c - C_t$. Time is included in the state for this finite-horizon task. At the episode boundary, the continuation in (110) is replaced by the observed terminal violation indicator, giving

$$\widehat{U}_{H-1}^c = \mathbf{1}\{B_H^c < 0\} - \widehat{\bar{u}}_c(X_{H-1}^c).$$

DCA implementation. We implement the cost-based extension using a recurrent Gaussian policy, a reward-value network, and a quantile critic for the future-cost distribution. The cost critic conditions on the observation, recurrent policy context, remaining time, and current action, and is trained with one-step distributional Bellman targets and the pairwise pinball loss. We evaluate its CDF by interpolating sorted quantiles at the remaining cost budget and construct the transition-based risk credit in (110) using sampled policy actions. The critic reuses recent trajectories, whereas actor and multiplier updates use fresh rollouts. Table 1 summarizes the implementation settings.

Reward advantages are Monte Carlo suffix returns minus the pre-update value prediction. The actor maximizes

$$\frac{1}{1+\lambda_k}\,\widehat{\mathbb{E}}\left[\frac{1}{100}\min\{\rho_t\widehat{A}^r_t,\bar{\rho}_t\widehat{A}^r_t\}-\lambda_k\max\{\rho_t\widehat{U}^c_t,\bar{\rho}_t\widehat{U}^c_t\}\right], \tag{111}$$

where $\bar{\rho}_t = \mathrm{clip}(\rho_t, 0.9, 1.1)$. The reward advantage is divided by 100 for numerical scaling. Following the official QCPO implementation (Jung et al., 2022), we normalize the combined surrogate by $1+\lambda_k$ to control its scale as the multiplier increases while preserving the relative weighting of reward and risk. Advantages and λ_k are frozen throughout the PPO epochs, so this common positive normalization preserves the surrogate's gradient direction for a fixed multiplier.

The multiplier uses an incremental PI update driven by the violation fraction in the fresh rollout, including during critic warm-up:

$$e_k = \widehat{p}_{c,k} - \alpha, \qquad m_{k+1} = \beta m_k + (1-\beta)e_k,$$
$$\lambda_{k+1} = \mathrm{clip}[\lambda_k + \eta_I e_k + \eta_P(m_{k+1}-m_k),\ 0, \lambda_{\max}]\,.$$

Table 1: DCA implementation settings for SimpleButton. Reward and cost targets are undiscounted.

Setting	Value
Policy network	Two 512-unit hidden layers; 512-unit LSTM
Critic networks	Two 512-unit hidden layers each
Distribution and risk estimation	128 midpoint quantiles; 4 actions per state
Optimizer and learning rates	Adam; actor initially 2×10^{-4}, critics 3×10^{-4}
Fresh rollout / critic warm-up	25 complete episodes / 20 rollouts
Critic replay / minibatch	Latest 100 episodes / 1024 transitions, 50% fresh
Critic updates per rollout	100 during warm-up; 400 thereafter
Target updates	Three fitting stages per rollout; hard update after each stage
PPO updates / clipping	Up to 8 full-batch epochs / 0.1
Policy KL control	Total budget 0.01; backtracking for rejected updates
PI coefficients	$\eta_I = 0.2$, $\eta_P = 5$, $\beta = 0.8$
Multiplier initialization / cap	$\lambda_0 = m_0 = 0$ / $\lambda_{\max} = 50$

Baselines. We compare with QCPO (Jung et al., 2022) and TQPO (Li et al., 2025) using recurrent implementations adapted to the same simulator. QCPO uses 25 cost quantiles, a Weibull tail model, and an empirical episode-cost quantile for its multiplier update. For this comparison, its reward and cost targets are undiscounted to match the episode-level criterion; this is a modification of its default discounted-target configuration. TQPO retains its source implementation's discounted reward and cost training targets with discount factor 0.99, while its multiplier is updated using undiscounted complete-episode costs. Both baselines use learning rate 10^{-4} and PPO clipping parameter 0.1.

Training and evaluation protocol. DCA and QCPO each use 10 million training transitions per seed, collected in batches of 25 complete episodes. DCA's budget includes the 20 critic-only warm-up rollouts. TQPO uses batches of 12 complete episodes and finishes at 10.008 million transitions, the first complete batch at or above the same nominal budget. We report all available runs of the final configurations: five seeds (0–4) for DCA and TQPO and four seeds (0–3) for QCPO. We use the final configurations selected in the SimpleButton tuning study.

Table 2 uses the final saved checkpoint of each run, evaluated with stochastic actions on independent episodes: 2048 episodes per seed for DCA and TQPO and 512 for QCPO. For each seed, we

compute mean return, mean cost, the empirical 80% cost quantile using linear interpolation, and the fraction of episodes with $Z > 15$. A seed is empirically feasible when this fraction is at most 0.20. Metrics are computed separately for each seed before averaging, and "$\pm$" denotes the sample standard deviation across training seeds.

For the learning curves, we first compute each seed's metrics from the latest 100 completed training episodes (or all available episodes initially), retaining the first rollout endpoint at or after each $50{,}000$-transition interval. We then apply a trailing 0.5-million-transition moving average within each seed before computing the across-seed mean and sample standard deviation. All plotted returns and costs are undiscounted.

E.3 Results

The objective in (105) is to maximize mean return while satisfying the cost-quantile constraint. Table 2 shows that DCA achieves both: it satisfies the empirical constraint in all five seeds and obtains a mean return of 18.19 ± 1.96, compared with 16.73 ± 1.97 for QCPO. This is an 8.71% increase in average return, together with a lower violation probability (11.39% versus 14.45%). QCPO satisfies the constraint in only three of four seeds, whereas DCA maintains feasibility across all reported runs.

Table 2: Final-checkpoint results on SimpleButton (mean $\pm$ sample SD). The constraint is $\Pr(Z > 15) \leq 20\%$; the last column counts feasible seeds.

Method	Mean return	Mean cost	$Q_{0.8}(Z)$	Violation (%)	Feasible
DCA (ours)	$\mathbf{18.19 \pm 1.96}$	5.32 ± 1.39	$\mathbf{5.20 \pm 1.30}$	$\mathbf{11.39 \pm 2.72}$	$\mathbf{5/5}$
QCPO	16.73 ± 1.97	5.08 ± 1.03	7.50 ± 6.03	14.45 ± 4.24	$3/4$
TQPO	22.11 ± 3.86	7.40 ± 1.66	16.92 ± 4.30	22.02 ± 3.68	$1/5$

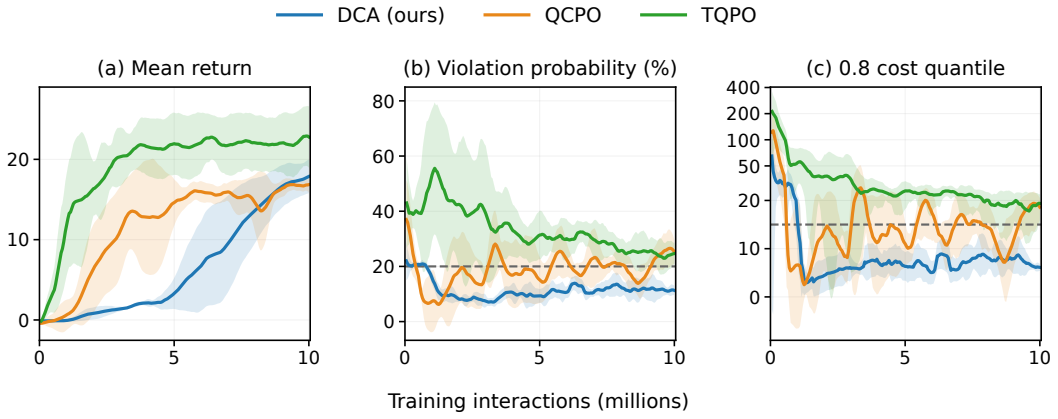

Figure 3: Learning performance on SimpleButton: mean return, violation probability, and 0.8 cost quantile (mean $\pm$ sample SD across seeds). Dashed lines mark the 20% violation limit and cost threshold 15.All panels show the full range; panel (c) uses a base-10 logarithmic scale, with a linear region $[0, 20]$. Curves follow the training-window and smoothing protocol above; Table 2 reports independent final-policy evaluations.

The comparison with TQPO further highlights the role of feasibility. TQPO violates the constraint in four of five seeds, with an average violation probability of 22.02% and an average 80% cost quantile of 16.92, exceeding the limits of 20% and 15, respectively. Its higher mean return is therefore obtained without consistently meeting the required constraint and does not constitute better performance on the stated constrained objective. DCA instead combines high reward with consistent empirical feasibility.

The cost statistics also distinguish tail control from mean-cost reduction. DCA and QCPO have similar mean costs (5.32 and 5.08), yet DCA achieves a lower average 80% cost quantile (5.20 versus 7.50) with substantially smaller variation across seeds (1.30 versus 6.03). Thus, DCA's constraint satisfaction is reflected directly in the cost tail, which is the quantity the problem requires controlling.

Figure 3 shows how this performance develops during training. After the initial phase, DCA keeps the smoothed violation rate below the prescribed limit while steadily improving mean return. QCPO exhibits larger fluctuations in its violation rate, while TQPO remains above the limit for much of training. The final-checkpoint evaluations confirm the resulting advantage: DCA is the only method that satisfies the empirical constraint in every reported seed, while also achieving a higher mean return than QCPO. Together, these results demonstrate that distributional credit assignment extends beyond lower-tail return constraints to expected-return maximization under a separate cumulative-cost quantile constraint.