ON THE RELIABILITY OF VALUE ESTIMATION FOR SAFE EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe exploration requires reinforcement learning (RL) agents to effectively trade off task performance against the risk of constraint violations during learning. Constrained RL algorithms use reward and cost value estimates to guide this trade-off, while their theoretical guarantees are stated in terms of the true value functions that these estimates approximate. Yet the reliability of the cost critic is rarely examined directly. We show that cost critics exhibit substantially poorer agreement with Monte Carlo reference values than reward critics. The regression problem for the cost critic spans a spectrum, from heavily bootstrapped targets that are stable but inherit the critic's own errors, to long-horizon targets that contain more information about future cost but are high variance and difficult to generalize from. Neither end yields a critic that correlates well with the reference cost value, though for different reasons. Because cost-critic errors do not exhibit a systematic bias, existing bias-control methods do not recover value fidelity and instead primarily make policy updates more conservative; regularization likewise fails to improve the overall safety–performance trade-off. We therefore explore successor representations as an alternative approach to learning cost value functions and show that they provide a promising direction for safer and more sample-efficient exploration.

1 INTRODUCTION

Reinforcement learning (RL) has demonstrated strong performance across a wide range of sequential decision-making problems, yet learning through direct interaction remains challenging in safety-critical domains. An agent must acquire sufficient experience to learn useful behavior while simultaneously limiting constraint violations incurred during learning (Achiam et al., 2017; Stooke et al., 2020; Mani et al., 2025). Safe exploration therefore requires not only learning a policy that eventually satisfies safety constraints, but doing so while making effective use of both samples and violations encountered along the way.

In constrained actor–critic methods (Stooke et al., 2020; Liu et al., 2022), this trade-off is mediated by estimates of future reward and cost. The monotonic improvement and constraint-feasibility bounds underlying many safe policy optimization methods (Achiam et al., 2017; Xu et al., 2021; Tamar et al., 2015) are expressed in terms of the true reward and cost values; however, in practice, these quantities are estimated by learned critics, making the reliability of the resulting policy updates dependent on their approximation quality. An inaccurate cost critic can underestimate dangerous behavior and lead to constraint violations, but it can also overestimate risk in otherwise useful regions of the state space, resulting in unnecessarily conservative exploration and sample inefficiency.

While value estimation has been studied extensively in standard RL, largely in the context of reward optimization and sample efficiency (Ilyas et al., 2018; Fujimoto et al., 2018a; Moskovitz et al., 2021), considerably less attention has been paid to the critic's quality in safe exploration, where errors in the cost critic directly affect the safety–performance trade-off. Existing analyses of cost critics have primarily focused on aggregate estimation error (Gao et al., 2025b), which can obscure state-dependent prediction errors that determine how individual policy updates trade reward against risk. In this work, we study value estimation as a fundamental component of safe exploration. We ask a basic empirical question: *how accurately do learned critics recover the underlying true value functions, and what kinds of estimation errors do they exhibit?* We evaluate learned reward and cost values on states sampled from the current on-policy distribution, using Monte Carlo (MC)

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.

reference estimates obtained by averaging repeated rollouts from each state under the current policy. Aggregate error metrics provide only a partial view of critic quality. Mean signed error can obscure compensating over- and under-estimation across states, whereas squared error captures error magnitude but discards its direction. In safe exploration, this distinction is important because underestimating high-cost states and overestimating low-cost states have qualitatively different consequences for policy improvement. We therefore complement aggregate estimation error with correlation, estimation error conditioned on the true cost-value regime, and *AUROC* for identifying high-cost states. Together, these diagnostics capture whether the critic preserves variation across states, where systematic over- or under-estimation occurs, and whether safety-critical states can be reliably distinguished.

Our experiments reveal a pronounced asymmetry between reward- and cost-value estimation. While the reward critic develops substantial agreement with its Monte Carlo (MC) reference early in training, the cost critic remains poorly correlated with the corresponding reference value, even when the policy satisfies the prescribed constraint. In many safe-control problems (Ray et al., 2019), cost feedback is sparse, so informative cost signals arise only along a small fraction of transitions. Estimating long-horizon cost therefore requires assigning these sparse signals to earlier states, while the realized cost-to-go from a given state can vary substantially across trajectories. We study this difficulty through the spectrum of TD(λ) targets. Short-horizon targets provide lower-variance supervision but rely heavily on bootstrapping from an inaccurate critic, whereas longer-horizon targets improve credit assignment and reduce bootstrap dependence at the cost of substantially higher variance. At the Monte Carlo limit, the target is better aligned with the underlying cost value, yet becomes considerably harder for the critic to learn reliably from finite on-policy data.

We next ask whether these two failure modes can be addressed with existing techniques. For strongly bootstrapped targets, we study representative bias-control methods (Gao et al., 2025b; Fujimoto et al., 2018a) and examine whether correcting over- or under-estimation also improves the state-wise fidelity of the cost critic; for high-variance targets, we evaluate regularization methods (Gal & Ghahramani, 2016) aimed at improving generalization from noisy supervision. Our results suggest that neither class of intervention fully resolves the estimation problem: bias-control methods can shift aggregate error without recovering the underlying value structure, while regularization provides limited improvement under high-variance targets. This motivates us to consider whether the prediction problem itself should be restructured. We therefore study successor representations (*SR*s) (Dayan, 1993b), which factorizes long-horizon cost prediction through future state visitation, and find that they improve early cost-value prediction and lead to a better safety–performance trade-off during exploration.

Our analysis leads to three key findings:

- *Constraint satisfaction does not imply an accurate cost critic.* Agents can satisfy the constraint even when their learned cost values remain poorly aligned with Monte Carlo references.
- *Aggregate estimation error can hide substantial structural error.* Cost critics may have near-zero average error while exhibiting poor correlation and systematic errors across different cost regimes, and existing bias-control methods do not necessarily recover this missing structure.
- *Predictive representations (SRs) improve cost-value estimation.* Successor representations factorize long-horizon cost prediction through future state visitation, improving early critic fidelity yielding a better safety–performance trade-off during exploration.

2 BACKGROUND

A *Constrained Markov decision process*(CMDP) (Altman, 2021) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, c, \gamma, \mu_0)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P(\cdot \mid s, a)$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, non-negative cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$, discount factor $\gamma \in [0, 1)$, and initial state distribution μ_0. Let Π denote the set of stationary policies. A policy $\pi \in \Pi$, together with μ_0 and P, induces a distribution over trajectories $\tau = (s_0, a_0, s_1, a_1, \dots)$, under which we define the discounted return and cost

$$J_r(\pi) = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t)\Big], \qquad J_c(\pi) = \mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty} \gamma^t\, c(s_t, a_t)\Big]. \tag{1}$$

The CMDP problem is

$$\pi^* \in \operatorname*{argmax}_{\pi \in \Pi_{\mathcal{C}}} J_r(\pi), \qquad \Pi_{\mathcal{C}} = \big\{\pi \in \Pi : J_c(\pi) \le \beta\big\}, \tag{2}$$

for a constraint threshold $\beta \ge 0$.

Value functions. In constrained RL, for every policy π, we have two value functions, one for cost (c) and one for reward (r), defined similarly by:

$$Q^\pi_\bullet(s,a) = \mathbb{E}_{\tau\sim\pi}\big[G_\bullet(\tau) \,\big|\, s_0 = s,\, a_0 = a\big], \qquad V^\pi_\bullet(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\big[Q^\pi_\bullet(s,a)\big], \tag{3}$$

where $\bullet$ stands for r or c and $G_r(\tau) = \sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)$ (resp. $G_c(\tau) = \sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)$) represents the discounted return (resp. cost) on trajectory τ. Note that $J_\bullet(\pi) = \mathbb{E}_{s_0\sim\mu_0} V^\pi_\bullet(s_0)$.

Policy gradient. For a parametrized policy π_θ, write $J_\bullet(\theta) = J_\bullet(\pi_\theta)$. The policy gradient theorem (Sutton et al., 1999) gives

$$\nabla_\theta J_\bullet(\theta) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim d^{\pi_\theta},\, a\sim\pi_\theta(\cdot|s)}\Big[\nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi_\theta}_\bullet(s,a)\Big], \tag{4}$$

which holds for any per-step signal and hence for reward and cost alike; it is the basis for solving the unconstrained problem $\max_\theta J_r(\theta)$ by stochastic gradient ascent, as well as for handling the constraint through the gradient of $J_c(\theta)$.

Lagrangian relaxation. Constrained Policy Optimization (***CPO*** (Achiam et al., 2017)) approximately maximize $J_r(\pi)$ under constraint and resort to minimizing $J_c(\pi)$ when the local problem is infeasible. Lagrangian methods instead introduce

$$\mathcal{L}_\eta(\theta) = J_r(\theta) - \eta\big(J_c(\theta) - \beta\big), \tag{5}$$

with multiplier $\eta \ge 0$. For fixed η, maximizing equation 5 over θ is an unconstrained MDP problem with scalar reward $r - \eta c$. Writing $G_\eta(\tau) = G_r(\tau) - \eta G_c(\tau)$, we have $J_\eta(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[G_\eta(\tau)] + \eta\beta$, where $\eta\beta$ is a constant independent of θ. Defining the corresponding value functions $Q^\pi_\eta = Q^\pi_r - \eta Q^\pi_c$ and $V^\pi_\eta = V^\pi_r - \eta V^\pi_c$, solutions to this unconstrained MDP can similarly be approximated using $\nabla_\theta J_\eta(\theta) = \nabla_\theta J_r(\theta) - \eta\nabla_\theta J_c(\theta)$ (equation 4) for a fixed η. However, in practice (TRPO-Lagrangian with a PID-controlled multiplier, abbreviated ***PIDLag*** (Stooke et al., 2020)) η is dynamically adapted to enforce the constraint. From here on $\bullet$ stands for r, c or η, with corresponding per-step signal $x_\bullet$ given by $r(s,a)$, $c(s,a)$ and $r(s,a) - \eta\, c(s,a)$ respectively.

2.1 Safe Exploration

On-policy deep RL algorithms for CMDPs learn a parametrized policy by iteratively collecting experience and updating the policy parameters, producing a sequence $\pi_1, \pi_2, \ldots, \pi_{\text{last}}$. The standard CMDP objective concerns the final policy,

$$\max_{\pi_{\text{last}}} J_r(\pi_{\text{last}}) \qquad \text{s.t.} \qquad J_c(\pi_{\text{last}}) \le \beta.$$

Safe exploration additionally concerns the cost incurred by the intermediate policies encountered while learning this final policy.

Suppose training generates N episodes, where episode j consists of a trajectory $s_0, a_0, \ldots, s_{T_j}$ with associated costs $c(s_t,a_t)$. Its undiscounted episodic cost is $C_j = \sum_{t=0}^{T_j-1} c(s_t,a_t)$

We summarize the cost incurred throughout learning using the *average episodic cost* (Avg EpCost),

$$\overline{C} = \frac{1}{N}\sum_{j=1}^{N} C_j. \tag{6}$$

Unlike the constraint on π_{last}, Avg EpCost captures the safety of the learning process itself. Among methods achieving comparable task performance, a lower Avg EpCost (also written as Avg episode cost in the paper) indicates that less cost was incurred while learning indicating safer exploration.

The policy updates of ***PIDLag*** and ***CPO*** are defined in terms of reward and cost advantages, which in practice must be estimated from finite on-policy data. We therefore ask how accurately these quantities are estimated in practice, and what aspects of estimation quality are important for safe exploration.

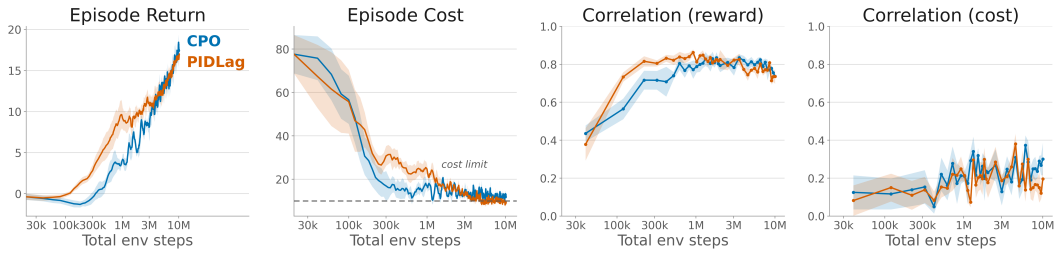


Figure 1: *Constraint satisfaction does not imply accurate cost critic*. Two constrained RL algorithms *CPO* and *PIDLag* on *PointGoal1* task from safety gym Ray et al. (2019). Despite having very poor correlation for the cost-critic both methods are able to satisfy the constraint ($\beta = 10$).

2.2 Value estimation from sampled trajectories

We estimate the reward and cost value functions from finite trajectory segments using identical procedures. Let x_t denote either $r(s_t, a_t)$ or $c(s_t, a_t)$, and let $\bar{V}$ be the corresponding critic frozen when the batch was collected. For a segment ending at T, define $\bar{V}(s_T) = 0$ at a true terminal state and retain its predicted value at a truncation. The m-step return is

$$\widehat{G}_t^{(m)} = \sum_{k=0}^{m-1} \gamma^k x_{t+k} + \gamma^m \bar{V}(s_{t+m}), \qquad 1 \le m \le T - t. \tag{7}$$

A geometrically weighted mixture gives the TD(λ) target

$$y_t(\lambda) = (1-\lambda) \sum_{m=1}^{T-t-1} \lambda^{m-1} \widehat{G}_t^{(m)} + \lambda^{T-t-1} \widehat{G}_t^{(T-t)}, \qquad \lambda \in [0, 1]. \tag{8}$$

Thus $\lambda = 0$ gives the one-step TD target, while $\lambda = 1$ gives the sampled return with a plug-in tail at truncation; intermediate values trade critic dependence against sampling variance.

Each critic is fitted to targets computed from its own frozen predictions:

$$\min_{\phi} \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \big(V_\phi(s_t) - y_t\big)^2. \tag{9}$$

Reward and cost critics therefore use the same targets, regression objective, and optimization budget. The m-step bias, mixture weights, endpoint conventions, and related derivations are given in Appendix A.

3 Safe agents with inaccurate cost critics

Constraint satisfaction does not imply an accurate cost critic. We first ask whether standard constrained RL methods learn cost critics that accurately represent future cost. Figure 1 evaluates *CPO* and *PIDLag* on the *PointGoal1* task from Safety Gym (Ray et al., 2019). We sample probe states from the current policy's on-policy state distribution and compare the critic predictions at these states with reference values obtained by resetting the simulator to each probe state and averaging independent rollouts of the current policy (Appendix B). Both methods substantially reduce episodic cost, from roughly 78 initially to below 25 within the first million environment steps, and continue toward the prescribed constraint ($\beta = 10$) over training. Their reward critics improve alongside the policy, reaching a correlation of approximately 0.8 with their reference values. In contrast, the cost critics remain poorly correlated, settling near 0.2 despite continued improvement in constraint satisfaction. This low correlation is not explained by noise in the Monte Carlo reference: on the same rollouts, the attainable correlation is approximately 0.98 (Appendix C). Thus, *standard constrained RL methods can learn policies that satisfy the cost constraint without learning an accurate cost value function.*

The errors are state dependent rather than a uniform bias. Figure 2 examines what this poor estimation looks like. The predicted and reference cost values form a pronounced "L"-shaped pattern: states with near-zero reference cost receive predictions spanning a wide range, while states with high reference cost are systematically underpredicted. The error therefore cannot be described as a simple global tendency to over- or underestimate cost.

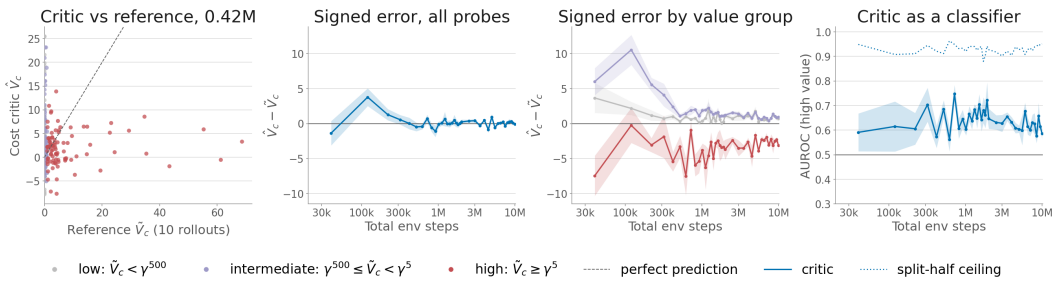


Figure 2: *A small aggregate error hides errors that cancel in the worst way.* CPO on *PointGoal1*, 5 seeds; lines are the seed mean and bands ± 1 standard error. From left to right: cost critic prediction against the Monte Carlo reference $\tilde{V}_c$, coloured by value group (low, $\tilde{V}_c < \gamma^{500}$; intermediate; high, $\tilde{V}_c \geq \gamma^5$); the signed error $\hat{V}_c - \tilde{V}_c$ over all probes across training; the same error split by value group; and the AUROC for identifying high value states, with the split-half ceiling attainable on these rollouts (dotted).

To quantify this structure, we group probe states according to their reference value $\tilde{V}_c$: *low* ($\tilde{V}_c < \gamma^{500}$), *intermediate* ($\gamma^5 \geq \tilde{V}_c \geq \gamma^{500}$), and *high* ($\tilde{V}_c \geq \gamma^5$). The high-value group corresponds to states from which cost is imminent: since a cost event contributes 1, incurring cost within t steps implies $V_c^\pi(s) \geq \gamma^t$. Further details on the thresholds are given in Appendix D.1.

This grouping reveals errors that are hidden by the aggregate statistic. The overall signed error $\hat{V}_c - \tilde{V}_c$ remains close to zero for most of training—$+0.02$ at 0.42M steps and $+0.09$ on average over the final 8M steps. Yet high-value states are systematically *underestimated*, by -5.5 at 0.42M steps and -2.9 over the final 8M, while intermediate-value states are simultaneously *overestimated*, by $+2.4$ and $+1.0$, respectively. These opposing errors cancel in aggregate. More importantly, they cancel in an unfavorable way for safe exploration: states where cost is imminent are underestimated, while comparatively safer states are assigned excessive cost.

The critic still preserves coarse information about high-risk states. This raises the question of how constrained RL methods succeed in satisfying constraints despite such inaccurate cost values. Figure 2 (col 4) suggests that the critic retains a weaker form of useful information. When asked only to distinguish high-value states from the rest, its *AUROC* remains between 0.55 and 0.70 throughout training and averages 0.62 over the final 8M steps, above the chance level of 0.5. Because the reference itself is estimated from finitely many stochastic rollouts, the relevant ceiling is below 1: splitting the ten rollouts for each probe into two halves and using one half to rank labels defined by the other gives an average *AUROC* of approximately 0.93. Thus, although the cost critic fails to recover the underlying cost value function accurately, it still preserves partial information about which states are associated with imminent cost. This coarse discrimination helps reconcile the two observations above: constrained RL can make progress toward satisfying the constraint even when its cost critic has poor value fidelity. Appendix D.2 further discusses this metric and its sensitivity to the threshold defining the high-value group.

Fitting the regression target does not imply learning the true cost value. The critic target controls a familiar bias–variance trade-off (Section 2.2). At $\lambda_c = 0$, the target relies almost entirely on bootstrapping from the critic's prediction at the next state. This produces low-variance targets, but also propagates errors already present in the critic. At $\lambda_c = 1$, the target approaches the sampled Monte Carlo return and depends little on the critic itself. It is therefore more faithful to the realized future cost, but is high variance: trajectories starting from the same state can accumulate very different costs, while the critic observes only a limited number of such realizations.

Figure 3 separates two questions: whether the cost critic learns the target it is given, and whether the resulting critic estimates the underlying cost value accurately. For the first question, strongly bootstrapped targets appear highly effective. At small λ_c, prediction–target correlation exceeds 0.95 on the training batch and remains similarly high on held-out states sampled from the same on-policy distribution. As λ_c approaches 1, this agreement deteriorates sharply; at $\lambda_c = 1$, test correlation with the MC target falls to approximately 0.2. The critic therefore generalizes well to low-variance bootstrapped targets, but struggles to predict high-variance Monte Carlo returns from individual states.

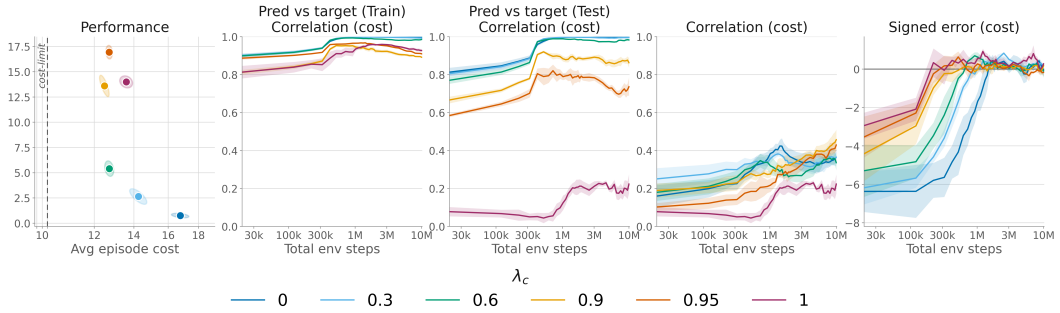


Figure 3: *Cost-value targets exhibit a trade-off between target fidelity and learnability.* Using CPO on *Point-Goal1*, we compare cost critics trained with TD(λ_c) targets ranging from low-variance, strongly bootstrapped TD targets to high-variance Monte Carlo (MC) returns. We evaluate train/test correlation with the regression target, together with correlation and estimation error against the MC reference value. Strongly bootstrapped targets are easy to fit and generalize to, but yield critics with substantially poorer agreement with the reference value. In contrast, MC targets reduce bootstrap dependence but their high variance leads to poor generalization and consequently weak cost-value correlation.

Evaluation against the MC reference gives a different picture. Correlation with the reference remains below approximately 0.45 across the entire sweep, including for critics that fit their bootstrapped targets almost perfectly. The estimation error tells the same story: highly bootstrapped targets exhibit substantially larger error with respect to the reference value, and at $\lambda_c = 0$ the critic requires more than 1M environment steps to reach the correct value scale, compared with only a few hundred thousand steps for $\lambda_c = 1$. Thus, low-λ_c critics are not merely poorly ordered relative to the reference; their predictions are also quantitatively farther from it. Strong target fit therefore reflects the learnability of the bootstrapped target rather than the fidelity of the learned cost value function.

The two ends of the spectrum therefore fail for different reasons. Low-λ_c targets are easy to learn but inherit errors through bootstrapping; high-λ_c targets better reflect realized future cost but are difficult to generalize from because of their variance. Policy performance follows the same trade-off: episodic cost decreases and return improves as λ_c increases up to approximately 0.95, while both degrade at $\lambda_c = 1$. Intermediate targets provide a better empirical compromise, but no value of λ_c yields a critic with strong agreement with the reference value.

Changing the target therefore shifts the source of the estimation problem rather thaneliminating it. This motivates two complementary directions: controlling the error introduced through bootstrapping at low λ_c, and improving generalization to high-variance targets at large λ_c.

4 Bias-Correction and Regularization

Bias correction changes value scale more than value structure. The previous analysis shows that strongly bootstrapped targets can be easy to fit while remaining poorly aligned with the reference cost value. We therefore ask whether methods designed to correct value-estimation bias can improve the learned cost critic. We compare *CPO* with *MICE* (Gao et al., 2025b) and clipped double-Q learning (CDQ) (Fujimoto et al., 2018a) (max of two cost value network), both of which introduce pessimistic corrections intended to reduce harmful cost underestimation. As shown in Figure 4, these methods primarily shift cost predictions upward. The shift is visible across low-, intermediate-, and high-value states, while correlation with the Monte Carlo reference remains poor. Thus, correcting the overall direction of the estimates does not recover the state-wise structure of the cost value function. This is consistent with the error pattern in Figure 2: cost-critic error is not a uniform bias that can be corrected by globally increasing or decreasing predictions. Different regions of the state distribution exhibit errors of different signs, so a pessimistic shift can reduce underestimation in some states while increasing overestimation in others.

MICE can improve behavior without improving the critic. *MICE* nevertheless changes policy performance because its intrinsic cost affects more than the critic target. In *CPO*, the empirical policy cost is used to select between reward-oriented and constraint-recovery updates; increasing this quantity therefore biases the algorithm toward conservative updates. To isolate this effect, we introduce a *biased-cost* baseline that leaves the critic unchanged and adds a positive, exponentially decaying bias only to the empirical cost used in this case-selection step. Figure 4 shows that this simple modification reproduces the early conservative behavior of *MICE* and achieves comparable or better performance. This ablation indicates that the benefit of *MICE* in this setting can largely

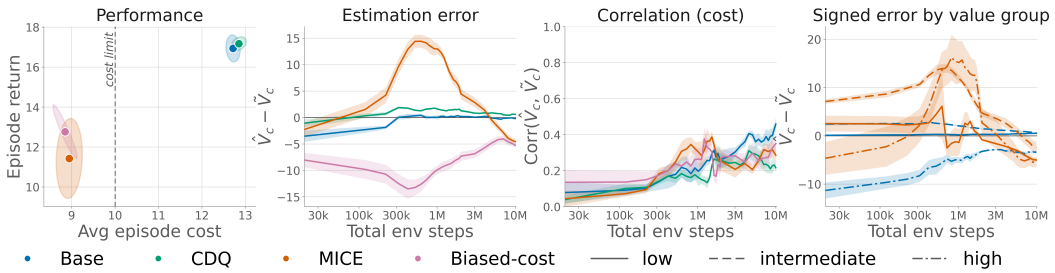


Figure 4: *Systematic bias correction does not recover cost-value structure.* Bias control methods like MICE (Gao et al., 2025b) and CDQ (Fujimoto et al., 2018a) primarily inflate cost predictions across low-, intermediate-, and high-cost states, without improving correlation with the Monte Carlo reference. A simple biased-cost baseline, which artificially increases the empirical constraint cost $J_c(\pi)$, achieves similar performance to MICE by inducing more conservative policy updates early in training.

be attributed to its effect on *CPO*'s update-selection mechanism, rather than to improved cost-value estimation.

Regularization improves estimation but not safe exploration. At the opposite end of the target spectrum, high-λ_c targets provide more informative supervision but are difficult to generalize from. We therefore test standard regularization through L_2 weight decay and dropout. Figure 5 shows that both can improve cost-value correlation relative to the base CPO critic. However, this improvement does not translate into a better safety–efficiency trade-off: episodic cost decreases, but so does episodic return. Regularization therefore produces more conservative policies rather than simultaneously improving safety and task performance.

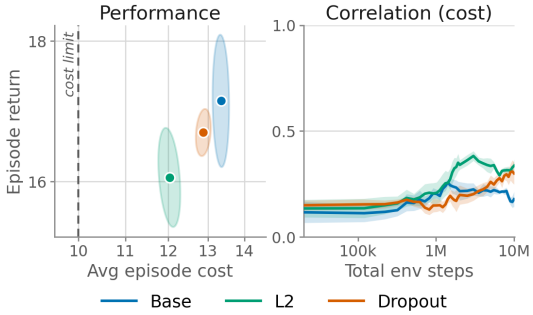


Figure 5: *Regularization improves safety but not the overall safety–efficiency trade-off.* Applying L_2 regularization or dropout to the cost critic reduces training cost, but at the expense of episodic return. While correlation improves later in training, early cost-value estimation remains poor, limiting its benefit for safe exploration.

Taken together, these results show that neither correcting global estimation bias nor regularizing the critic resolves the cost-value learning problem identified above. Bias correction can improve behavior through the optimization mechanism without recovering accurate state-wise values, while regularization can improve critic correlation without improving the overall safety–efficiency trade-off.

5 Predictive State Representations as Inductive Bias

The preceding analysis points to a limitation of directly learning the cost value function. Changing the return target trades bootstrap error against variance, while bias correction and regularization do not consistently recover an accurate cost-value function or improve the safety–performance trade-off. We therefore consider a different approach: rather than estimating long-horizon cost directly, can the critic first learn a predictive representation of future state visitation and use it to construct the value function?

Given a feature $\phi : \mathcal{S} \to \mathbb{R}^d$ for the state space, the successor representation of a policy π is defined as the vector valued function[1]:

$$\psi^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \,\middle|\, s_0 = s \right]. \tag{10}$$

If reward and cost are linear in these features, $r(s) = \phi(s)^\top w_r$ and $c(s) = \phi(s)^\top w_c$, their value functions become

$$V_r^\pi(s) = \psi^\pi(s)^\top w_r, \qquad V_c^\pi(s) = \psi^\pi(s)^\top w_c. \tag{11}$$

[1]We use contrastive learning for ϕ using InfoNCE (Oord et al., 2018)

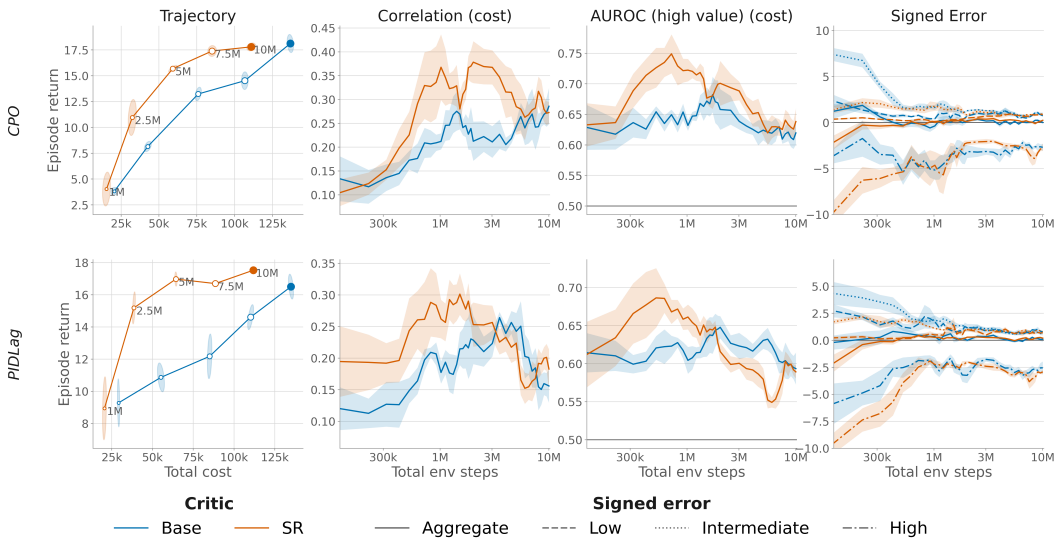


Figure 6: *Successor representations improve both cost-value estimation and safe exploration.* Results for *CPO* (top) and *PIDLag* (bottom) compare the base critic with an *SR*-based critic. From left to right: reward–total cost learning trajectories, cost-critic correlation with Monte Carlo reference values, *AUROC* for distinguishing low- and high-cost states, signed error aggregate and by group. The SR critic improves the reward–cost trade-off, achieves higher early-training correlation, and better separates low- from high-cost states.

We use the same successor representation ψ^π for both reward and cost, with separate reward and cost heads w_r and w_c. This factorization separates long-horizon prediction from instantaneous signal estimation: ψ^π learns the discounted future state occupancy from all visited states, while w_r and w_c are learned from per-step reward and cost. For the cost critic in particular, sparse cost therefore affects only the instantaneous cost regression rather than the entire long-horizon prediction problem.

***SR* improves the safety–performance trade-off.** For both *CPO* and *PIDLag*, the *SR* critic reduces the total cost incurred during training while also improving return. Average episodic training cost decreases from 13.9 to 11.3 for CPO and from 13.6 to 11.4 for PIDLag. The reward–cost trajectories in Figure 6 (col 1) show that this is not achieved by sacrificing task performance: beyond roughly 25k cumulative cost, the SR runs obtain higher return for the same cost budget, with the gap becoming significant from approximately 50k onward. At a budget of 75k, every SR run reaches a return of at least 16.4, while no base run exceeds 15.3. Conversely, a return of 12.5 is reached at roughly half the training cost: 36.9k versus 72.4k for *CPO*, and 29.4k versus 77.1k for *PIDLag*.

***SR* requires fewer conservative updates.** The improvement is accompanied by less reliance on cost-driven optimization (Figure 7). For *CPO*, the fraction of updates that fall back to pure cost recovery decreases from 82% to 59%. For *PIDLag*, the average Lagrange multiplier decreases from 17.0 to 7.5. Because both changes occur alongside lower training cost and higher return, they indicate that SR does not obtain safety simply by making the policy more conservative. Instead, the agent requires fewer or weaker cost-driven corrections while achieving better constraint optimization.

Fig. 6 (col 3) shows that the *SR* critic separates high-value states from the rest significantly better than the base critic (*AUROC* 0.72 against 0.66 for *CPO*, 0.65 against 0.63 for *PIDLag*), and in the first it over-predicts far less on low- and intermediate-value states ($+0.4$ against $+1.7$, and $+1.5$ against $+5.7$, for *CPO*), while under-predicting more strongly on high-value states (-8.5 against -3.6). The correlation with the reference values, by contrast, never separates significantly over the first two windows, and only does so for *CPO* beyond 2M steps: the evidence for an improvement is clearer for the ordering of states than for the accuracy of the estimates themselves.

The advantage in ordering as measured by *AUROC* is limited to the phase in which the cost is driven under limit. Beyond 2M steps the two critics are indistinguishable for *CPO* and the *SR* critic falls behind for *PIDLag*, but the two agents are then at very different points in learning — at 2M the SR agent has reached a return of 14, which the base agent only attains around 6M — so they are being trained and scored on very different state distributions.

Taken together, these results support the central diagnosis of this work: improving safe exploration requires more than correcting the aggregate bias of the cost critic. The structure of the prediction problem itself matters. By separating long-horizon predictive information from instantaneous cost estimation, successor representations provide one example of how richer predictive representations can improve cost-value estimation and translate into safer, more sample-efficient policy learning. We view this result as a promising direction rather than a complete solution, motivating further work on representations and learning objectives specifically designed for reliable safety prediction.

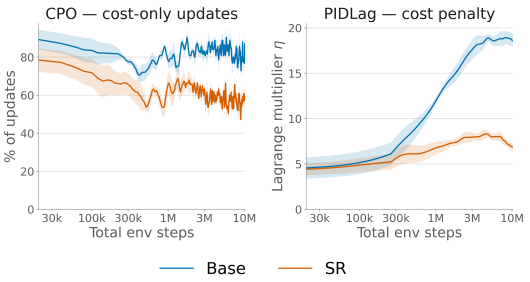


Figure 7: *SR enables safer exploration with less conservative updates. SR* reduces cost-only updates in CPO and lowers the Lagrange multiplier in *PIDLag*, while still improving constraint satisfaction.

6 Related Work

Safe Exploration: Safe exploration aims to improve policies while limiting constraint violations during learning (Achiam et al., 2017; Liu et al., 2022; Sootla et al., 2022; Stooke et al., 2020; Gu et al., 2024; Mani et al., 2026). Existing methods approach this through uncertainty-aware model-based planning (Berkenkamp, 2019; Huang et al., 2023), safety critics and action-filtering mechanisms (Fisac et al., 2019; Srinivasan et al., 2020; Kang et al., 2022; Bharadhwaj et al., 2020), or augmented state representations that encode accumulated cost or safety information (Sootla et al., 2022; Jiang et al., 2023). Across these approaches, effective policy improvement depends on accurately estimating future safety consequences from limited on-policy data. Our work studies this estimation problem directly, focusing on the reliability of learned cost-value functions and how their errors affect safe exploration.

Value Estimation and Policy Learning: Value estimation errors can substantially affect policy learning, particularly when bootstrapping and function approximation amplify prediction errors (van Hasselt, 2010; Fujimoto et al., 2018b). Prior analyses show that learned critics may poorly approximate reference values and produce inaccurate policy-gradient estimates (Ilyas et al., 2020; Wang et al., 2025). In constrained RL, recent work has focused on correcting systematic cost underestimation (Gao et al., 2025a). In contrast, we show that cost-critic errors can be structurally heterogeneous, such that small aggregate bias does not imply accurate state-wise cost estimation.

Successor Representations: Successor representations factor value prediction into expected future state occupancy and task-specific rewards or costs (Dayan, 1993a). They have primarily been used for representation learning, transfer, and temporal abstraction (Kulkarni et al., 2016; Barreto et al., 2017; Machado et al., 2018). We instead use this factorization for cost critics, showing that richer predictive structure improves cost-value fidelity and leads to safer, less conservative policy optimization.

7 Discussion and Conclusion

Our results identify reliable cost-value estimation as an important and under-examined component of safe exploration. We find that aggregate over- or under-estimation alone does not adequately characterize critic quality: cost critics can exhibit low average error while failing to preserve the state-wise structure of future cost, and existing bias-correction methods may shift predictions without improving this structure. Our target analysis further reveals a trade-off between stable but strongly bootstrapped targets and informative but high-variance long-horizon targets. Together with the gains obtained using successor representations, these findings suggest that future work should focus on richer predictive representations, uncertainty-aware targets, and adaptive bias–variance trade-offs, while evaluating safe RL algorithms not only through final return and constraint satisfaction but also through the fidelity of the safety information used to guide policy updates.

AI USE STATEMENT

In this work, we used generative AI tools for aid with polishing the manuscript fixing grammatical errors etc. along with coding assistance in forms of suggestions. We have not used generative AI tools for ideation, structuring the manuscript or to perform any analysis used in the paper. We have reviewed all AI-assisted work. All LLM generated code was reviewed by multiple authors and rigorously stress tested to verify its accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

In this work, we propose an analysis and a method for improving safe exploration in deep reinforcement learning algorithms. Since the method is application-agnostic, it does not introduce application-specific ethical concerns on its own. However, its use in real-world domains may raise ethical considerations depending on the application. We therefore encourage practitioners to carefully assess the potential risks and societal implications when deploying the method in areas such as healthcare, robotics, communications, and finance.

REPRODUCIBILITY STATEMENT

All code, configuration files, and exact hyperparameters required to reproduce our results are publicly available at `https://anonymous.4open.science/r/value-fn-reliablity`. Experiments were run in standard, publicly accessible environments (OpenAI Gym, Safety Gym), with version details documented in the repository.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31. PMLR, 2017.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, David Silver, and Hado van Hasselt. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4055–4065, 2017.

Felix Berkenkamp. *Safe exploration in reinforcement learning: Theory and applications in robotics*. PhD thesis, ETH Zurich, 2019.

Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. *arXiv preprint arXiv:2010.14497*, 2020.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4):613–624, 1993a. doi: 10.1162/neco.1993.5.4.613.

Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993b.

Jaime F Fisac, Neil F Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J Tomlin. Bridging hamilton-Jacobi safety analysis and reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8550–8556. IEEE, 2019.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pp. 1587–1596. Pmlr, 2018a.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 2018b.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016.

Shiqing Gao, Jiaxin Ding, Luoyi Fu, and Xinbing Wang. Controlling underestimation bias in constrained reinforcement learning for safe exploration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18290–18316. PMLR, 2025a.

Shiqing Gao, Jiaxin Ding, Luoyi Fu, and Xinbing Wang. Controlling underestimation bias in constrained reinforcement learning for safe exploration. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18290–18316. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/gao25b.html.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Weidong Huang, Jiaming Ji, Borong Zhang, Chunhe Xia, and Yaodong Yang. Safe dreamerv3: Safe reinforcement learning with world models. *arXiv preprint arXiv:2307.07176*, 2023.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Are deep policy gradient algorithms truly policy gradient algorithms. *arXiv preprint arXiv:1811.02553*, 2018.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020.

Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. OmniSafe: An infrastructure for accelerating safe reinforcement learning research. *Journal of Machine Learning Research*, 25(285):1–6, 2024.

Hao Jiang, Tien Mai, Pradeep Varakantham, and Minh Huy Hoang. Solving richly constrained reinforcement learning through state augmentation and reward penalties. *arXiv preprint arXiv:2301.11592*, 2023.

Katie Kang, Paula Gradu, Jason J Choi, Michael Janner, Claire Tomlin, and Sergey Levine. Lyapunov density models: Constraining distribution shift in learning-based control. In *International Conference on Machine Learning*, pp. 10708–10733. PMLR, 2022.

Tejas D. Kulkarni, Ardavan Saeedi, Simanta Gautam, and Samuel J. Gershman. Deep successor reinforcement learning. *arXiv preprint arXiv:1606.02396*, 2016.

Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pp. 13644–13668. PMLR, 2022.

Marlos C. Machado, Clemens Rosenbaum, Xiaoxiao Guo, Miao Liu, Gerald Tesauro, and Murray Campbell. Eigenoption discovery through the deep successor representation. In *International Conference on Learning Representations*, 2018.

Kaustubh Mani, Vincent Mai, Charlie Gauthier, Annie Chen, Samer Nashed, and Liam Paull. Safety representations for safer policy learning. *arXiv preprint arXiv:2502.20341*, 2025.

Kaustubh Mani, Yann Pequignot, Vincent Mai, and Liam Paull. Shapo: Sharpness-aware policy optimization for safe exploration. *arXiv preprint arXiv:2606.10228*, 2026.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.

Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano, Michael Arbel, and Michael Jordan. Tactical optimism and pessimism for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12849–12863, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.

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

Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. Learning to be safe: Deep RL with a safety critic. *arXiv preprint arXiv:2010.14603*, 2020.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *International Conference on Machine Learning*, pp. 9133–9143. PMLR, 2020.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. *Advances in neural information processing systems*, 28, 2015.

Hado van Hasselt. Double q-learning. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2613–2621, 2010.

Tao Wang, Ruipeng Zhang, and Sicun Gao. Improving value estimation critically enhances vanilla policy gradient. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65449–65463. PMLR, 2025.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pp. 11480–11491. PMLR, 2021.

A VALUE ESTIMATION FROM SAMPLED TRAJECTORIES

The objects of Subsection 2 are expectations over infinite trajectories; in practice one only has access to the finite batches of sampled ones introduced in Subsection 2.1. We record here several estimators that stand in for $V^\pi_\bullet$.

The frozen critic. The estimators below apply to both streams, so we write x_t for either $r(s_t, a_t)$ or $c(s_t, a_t)$ and V for $V_\bullet$ when no distinction is necessary. Throughout, $\bar{V}$ denotes the *frozen* critic, i.e. the parameters in force when the batch was collected, so that all targets below are computed from a fixed function and, conditional on $\bar{V}$, the regression that follows is an ordinary supervised problem.

m-step returns. Let $n = T - t$ be the number of steps left in the segment, and set $\bar{V}(s_T) = 0$ if s_T is a true terminal state (at a truncation $\bar{V}(s_T)$ is the critic's value). For $1 \le m \le n$, the m-step return

$$G^{(t,m)} \;=\; \sum_{k<m} \gamma^k x_{t+k} \;+\; \gamma^m \bar{V}(s_{t+m}) \tag{12}$$

accumulates m observed steps and lets the frozen critic supply the tail. If $\bar{V} = V^\pi$ it is an unbiased estimate of $V^\pi(s_t)$; otherwise its bias is $\gamma^m \,\mathbb{E}\big[\bar{V}(s_{t+m}) - V^\pi(s_{t+m}) \mid s_t\big]$. Averaging m-step returns with geometric weights gives a family of targets,

$$y_t(\lambda) \;=\; (1-\lambda) \sum_{m=1}^{T-t-1} \lambda^{m-1}\, G^{(t,m)} \;+\; \lambda^{T-t-1}\, G^{(t,T-t)}, \qquad \lambda \in [0,1], \tag{13}$$

whose weights sum to one (with $0^0 = 1$). Its endpoints are

$$y_t(0) = x_t + \gamma \bar{V}(s_{t+1}) \;\text{ (TD(0))}, \qquad y_t(1) = G^{(t,T-t)} \;\text{ (MC, with a plug-in tail at truncation)},$$

and intermediate values (TD(λ)) trade dependence on $\bar{V}$ against variance: the frozen critic's weight decreases from TD(0) to MC.

Critic regression. Each critic is then fit by minimizing squared error against targets obtained using its own frozen outputs,

$$\min_\phi \; \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \big(V_\phi(s_t) - y_t\big)^2, \tag{14}$$

over the batch $\mathcal{B}$, for a fixed number of gradient passes per batch. Nothing in this construction distinguishes reward from cost, which is what makes the comparison in the next section pointed: under the same targets, the same regression and the same budget, one of the two critics ends up far more faithful than the other.

Advantage estimation. In practice $Q^\pi_\bullet$ in equation 4 is replaced by the advantage $A^\pi_\bullet(s,a) = Q^\pi_\bullet(s,a) - V^\pi_\bullet(s)$, which leaves the gradient unchanged and reduces its variance, and the value functions themselves by a learned approximation $\hat{V}_\bullet$, the *critic*. Throughout we use the standard one-step temporal-difference estimate: for an action a taken in state s, producing the per-step signal $x_\bullet$ and the next state s',

$$A_\bullet(s,a) \;=\; x_\bullet + \gamma\, \hat{V}_\bullet(s') - \hat{V}_\bullet(s), \tag{15}$$

so that $A_\eta = A_r - \eta A_c$ whenever the two are combined as in *PIDLag*. Equation 15 describes how the cost critic $\hat{V}_c$ influences the policy update and provides the natural setting for the question we study.

B HOW MANY MONTE CARLO ROLLOUTS PER PROBE STATE?

When the critic is evaluated, and on whose states. The critic is evaluated at the point in the epoch where its predictions enter the update, on states visited by the policy that collected the batch. So the critic under evaluation has been trained on data from all previous policies except the current one, which we roll out to sample states for the evaluation. This way, the reported correlation (or any other metric) measures the out-of-sample quality the policy update actually relies on.

Value function reference We estimate the value of a probe state s and policy π by averaging K independent rollouts from π starting at state s:

$$\tilde{V}(s) \;=\; \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{t=0}^{T}\gamma^{\,t}\,x_t^{(k)}\Big), \qquad s_0^{(k)} = s, \tag{16}$$

With $T = \infty$, it is an unbiased but noisy estimate of $V^\pi(s)$. In our experiments, we use $T = 1000$ which has a small negative bias since the missing tail is set to zero. More rollouts (larger K) give a more precise estimate but require more simulations, hence more computation time. Since probe states are drawn from the current policy, the difficulty of estimating their values changes during training, so K must suffice at every checkpoint. We chose K from a dedicated run that collects 40 rollouts per probe state.

Setup. A CPO run on *PointGoal1* (one seed) is evaluated at 9 checkpoints between 2×10^4 and 10^7 steps, each with 150 probe states and 40 rollouts per state. We measure the quality of a K-rollout estimate in three ways.

- *Error.* We draw K of the 40 rollouts at random (200 draws) and compute the root-mean-square difference, across probe states, between the K-rollout and the 40-rollout estimates, divided by the standard deviation of the 40-rollout estimates across states. The K rollouts are a subset of the 40, so this understates the error against the true value, by a factor $\sqrt{1 - K/40}$ in expectation.
- *Reliability.* The correlation across probe states between a K-rollout estimate and the true value V,

$$\rho_{\max}(K) \;=\; \mathrm{Corr}\big(\tilde{V}^{(K)}, V\big) \;=\; \sqrt{\frac{\mathrm{Var}[V]}{\mathrm{Var}[V] + \mathbb{E}[\sigma^2]/K}}, \tag{17}$$

  where $\mathrm{Var}[V]$ is the variance of the true values across states and $\mathbb{E}[\sigma^2]$ the average variance of a single rollout, both estimated by a one-way analysis of variance on the 40 rollouts.
- *Split-half correlation.* The correlation across probe states between two estimates built from disjoint random sets of K rollouts each ($K \le 20$, 200 random splits). The two estimates are independent given the state, so this correlation equals $\rho_{\max}(K)^2$; it measures how well two independent K-rollout estimates agree, without relying on the variance model above.

Estimating the cost value is harder than the reward value. For the same number of rollouts, cost estimates are noisier relative to the spread of values across states. With a single rollout, the error is 33% of that spread for reward and 49% for cost (median over checkpoints), the reliability $\rho_{\max}(1)$ has median 0.95 for reward against 0.88 for cost, and two independent single-rollout estimates correlate at 0.90 for reward against 0.78 for cost, dropping to 0.55 for cost at some checkpoints.

Ten rollouts balance compute and precision. At $K = 10$ the error is 9% of the spread of values for reward and 14% for cost (at most 21% and 24% over checkpoints), and $\rho_{\max}$ has median 0.994 for reward and 0.985 for cost, never below 0.972 and 0.955. Two independent 10-rollout estimates correlate at 0.99 for reward and 0.97 for cost (never below 0.946 and 0.935), and splitting all 40 rollouts into halves of 20 gives 0.99 for both. The split-half correlations agree with the analysis-of-variance prediction $\rho_{\max}(K)^2$ to within 0.001 for reward and 0.023 for cost at $K = 10$. Quadrupling the simulation to $K = 40$ raises the reliability by a median 0.004 (reward) and 0.011 (cost), and by at most 0.021 and 0.033, while halving it to $K = 5$ raises the cost error from 14% to 21% and lowers its reliability by up to 0.039. We therefore use $K = 10$ rollouts per probe state throughout.

C Reliability of the Monte Carlo reference values

The correlations reported in Figure 1 compare each critic $\hat{V}_\bullet$ with the reference $\tilde{V}^\bullet$ of equation equation 16, an average of only $K = 10$ rollouts for a sample of states (probe set) collected with the current policy. Given how difficult the value estimation problem is, it is essential to assess that our reference estimate is good enough to guarantee the trustworthiness of the metrics reported. If the reference is too noisy, in principle, it could make any critic look poorly correlated, for example. We

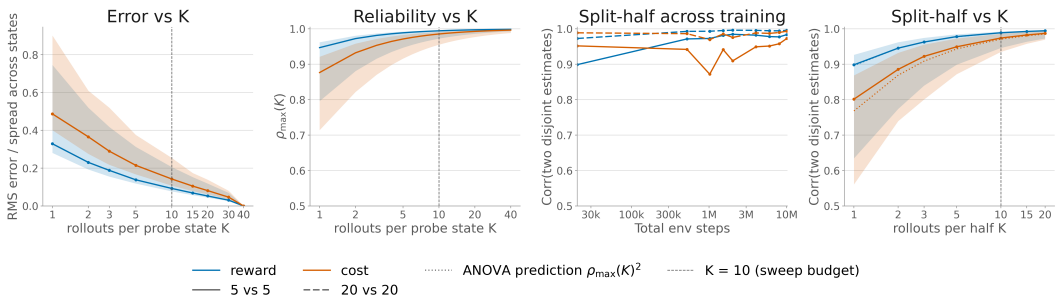


Figure 8: *The cost value is harder to estimate than the reward value, and $K = 10$ rollouts per probe state give a precise estimate of both.* CPO on *PointGoal1*, one seed, 9 checkpoints, 150 probe states with 40 rollouts each; reward in blue, cost in red; lines are medians over checkpoints and bands their range. *Error of a K-rollout estimate*: root-mean-square difference from the 40-rollout estimate, relative to the spread of values across states (median over 200 random draws of K rollouts). *Reliability of a K-rollout estimate*: correlation with the true value, from the analysis of variance. *Split-half correlation*: correlation between two estimates built from disjoint random sets of K rollouts each (solid), and its analysis-of-variance prediction $\rho_{\max}(K)^2$ (dotted). The dashed vertical line marks $K = 10$.

show that it does not: over the probe set of both algorithms, the reference is reliable enough that a perfect critic would reach a correlation of about 0.9 or higher at every checkpoint, for both reward and cost.

Attenuation by the reference. For a probe state s, write $\tilde{V}^\bullet(s) = V^\pi_\bullet(s) + \varepsilon(s)$. The K rollouts are independent given s and do not depend on the critic, so $\mathbb{E}[\varepsilon(s) \mid s] = 0$ and $\mathrm{Var}[\varepsilon(s) \mid s] = \sigma^2_\bullet(s)/K$, where $\sigma^2_\bullet(s)$ is the variance of a single discounted rollout return from s. Over the distribution of probe states this gives $\mathrm{Cov}(\hat{V}_\bullet, \tilde{V}^\bullet) = \mathrm{Cov}(\hat{V}_\bullet, V^\pi_\bullet)$ and $\mathrm{Var}[\tilde{V}^\bullet] = \mathrm{Var}[V^\pi_\bullet] + \mathbb{E}[\sigma^2_\bullet]/K$, hence

$$\mathrm{Corr}\big(\hat{V}_\bullet, \tilde{V}^\bullet\big) \;=\; \mathrm{Corr}\big(\hat{V}_\bullet, V^\pi_\bullet\big)\, \rho_{\max}(K), \qquad \rho_{\max}(K) \;=\; \sqrt{\frac{\mathrm{Var}[V^\pi_\bullet]}{\mathrm{Var}[V^\pi_\bullet] + \mathbb{E}[\sigma^2_\bullet]/K}}. \tag{18}$$

The factor $\rho_{\max}(K) \le 1$ is the correlation a perfect critic, $\hat{V}_\bullet = V^\pi_\bullet$, would obtain against the reference. It is the same for every critic evaluated on the same probe states, so it bounds what the measurement can show, not what a critic has learned.

Estimating the ceiling. Given returns G_{ik} from K rollouts at each of n probe states, we estimate $\rho_{\max}(K)$ by a one-way analysis of variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2, \qquad \widehat{\mathrm{Var}}[V] = \max\Big(S_b^2 - \frac{\hat{\sigma}^2}{K},\, 0\Big), \qquad \hat{\rho}_{\max}(K) = \sqrt{\widehat{\mathrm{Var}}[V]/S_b^2}, \tag{19}$$

where s_i^2 is the sample variance of the K returns from state i and S_b^2 is the sample variance of the reference values $\tilde{V}_i = \frac{1}{K}\sum_k G_{ik}$ across states.

A split-half check. The estimate above relies on the within-state variances. As an independent check, we split the $K = 10$ rollouts of every probe state at random into two halves of five and compute the correlation r between the two half-averages across states. The halves are independent given the state, so by the same argument as equation 18, $r = \rho_{\max}(5)^2$, and the Spearman–Brown relation gives $\rho_{\max}(10)^2 = 2r/(1 + r)$. This uses only the independence of rollouts; we average it over 200 random splits.

Results. Figure 10 reports both quantities for the configurations of Figure 1, over the same 5 seeds and 102 probe states.

- *Reward.* The ceiling has median 0.994 for both algorithms and is never below 0.969 (seed mean). The two halves of the reference agree with correlation $r \approx 0.98$, matching the analysis-of-variance prediction $\hat{\rho}_{\max}(5)^2$ to within 0.003.

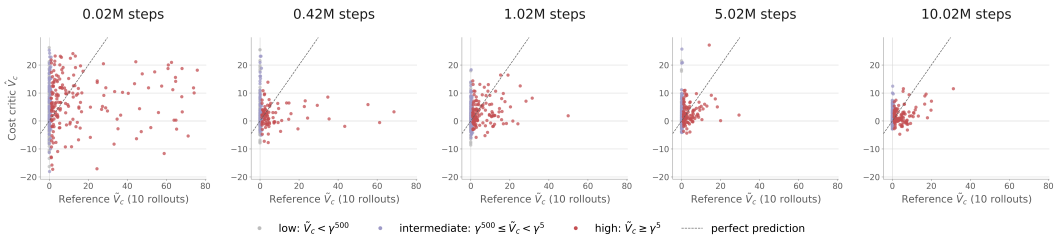


Figure 9: *Cost Critic prediction vs reference across training:* The figure shows the scatter plot for prediction vs true reference cost value function over different training steps for *CPO* on *PointGoal1* task. Its clear that even at later stages of learning cost critic overestimates cost values for intermediate states and underestimates cost values for high risk or highly unsafe states.

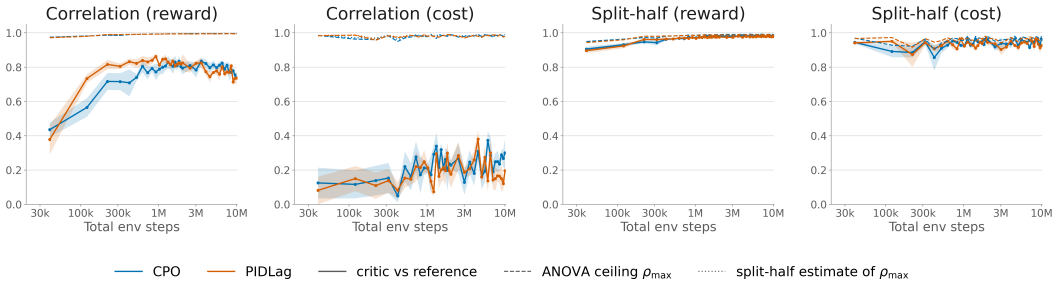


Figure 10: **The $K = 10$ Monte Carlo reference is reliable enough to support Figure 1.** CPO (blue) and PIDLag (orange) on *PointGoal1*, same configurations, 5 seeds and 102 probe states as Figure 1; solid lines and bands are the seed mean ± 1 standard deviation. *Correlation (reward)* and *Correlation (cost)*: correlation between the critic and the reference (as in Figure 1), with the ceiling $\rho_{\max}(10)$ attainable by a perfect critic, estimated by analysis of variance (dashed) and from split halves of the rollouts (dotted). *Split-half agreement*: correlation across probe states between the averages of two random halves of five rollouts (solid), and its analysis-of-variance prediction $\rho_{\max}(5)^2$ (dashed). The ceiling stays near or above 0.9 throughout, far above the cost critics' correlations.

- *Cost.* The cost signal is sparser, and the ceiling is slightly lower and noisier: median 0.983 (CPO) and 0.975 (PIDLag), never below 0.919 and 0.898 (seed mean); for a single seed at a single checkpoint it drops as low as 0.72 (CPO) and 0.60 (PIDLag). The split-half correlation has median 0.94 (CPO) and 0.93 (PIDLag), within 0.05 of its prediction, and the split-half ceiling stays within 0.025 of the analysis-of-variance one.

Consequence for Figure 1. Against these ceilings, the cost critics' correlations remain low: at most 0.49 (CPO) and 0.36 (PIDLag) over training, and 0.47 and 0.27 at the end, where the ceilings are 0.98 and 0.96. Dividing by the ceiling changes the final values by at most 0.01. A critic equal to the true cost value would score about 0.9 or higher on these very rollouts, so the gap between reward and cost correlations in Figure 1 reflects the critics, not the noise in the reference.

D METRICS FOR CRITIC QUALITY

D.1 VALUE GROUPS.

Several of our diagnostics are reported per group of probe states, split by the reference value $\tilde{V}_c$ into *low* ($< \gamma^{500}$), *intermediate*, and *high* ($\geq \gamma^5$); with $\gamma = 0.99$ these thresholds are 0.0066 and 0.951. They are stated as powers of γ because that makes them interpretable in units of time-to-cost. With a per-step cost of 1, a cost event occurring $k \leq t$ steps after s contributes $\gamma^k \geq \gamma^t$ to the discounted sum, and the remaining terms are nonnegative, so

$$V_c^\pi(s) \;\geq\; \gamma^t \,\Pr\big(\text{a cost event within } t \text{ steps} \mid s\big). \tag{20}$$

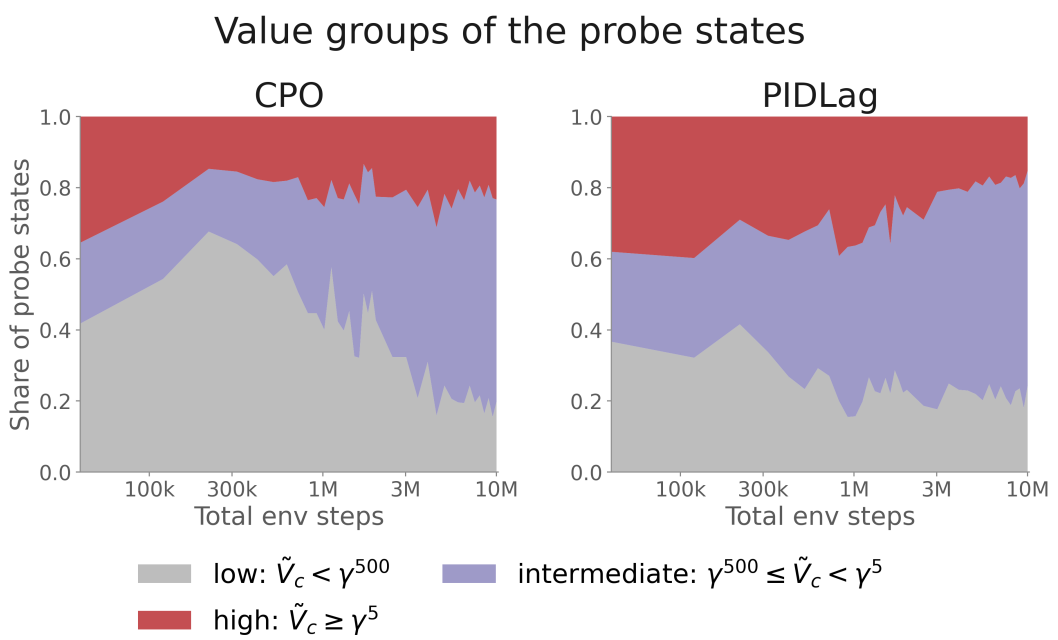


Figure 11: Evolution of share in each group across training, averaged over 5 seeds, complement to 1.

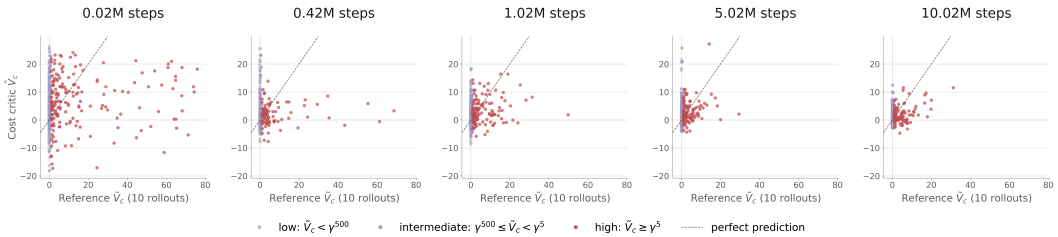


Figure 12: *Cost Critic prediction vs reference across training:* The figure shows the scatter plot for prediction vs true reference cost value function over different training steps for *CPO* on *PointGoal1* task. Its clear that even at later stages of learning cost critic overestimates cost values for intermediate states and underestimates cost values for high risk or highly unsafe states.

Read forward, equation 20 says that if cost is incurred within t steps with certainty then $V_c^\pi(s) \geq \gamma^t$: a state from which cost is imminent, within five steps, is a high value state. Read backward, it bounds how soon cost can arrive, $\Pr(\text{a cost event within } t \text{ steps} \mid s) \leq \gamma^{-t} V_c^\pi(s)$, so a low value state has at most a 1.8% chance of incurring any cost over the next hundred steps (γ^{400}), and at most 4.9% over the next two hundred. The backward reading becomes vacuous as t approaches 500, and it says nothing about what happens afterwards: a state that incurs no cost for five hundred steps may still have a value as large as $\gamma^{501}/(1-\gamma) \approx 0.65$ if cost accumulates beyond that point. The groups are therefore defined by the value itself and not by a time-to-cost.

The thresholds are fixed rather than quantiles of the current value distribution, so that the same state falls in the same group at every checkpoint and for every algorithm; the price is that the group sizes move over training, from a majority of low value probes while the cost is being brought down to a majority of intermediate ones once it has been. Figure 13 reports how the classification results vary with the threshold defining the high group, from γ^{500} up to values well above 1.

D.2 A WEAKER QUESTION: TELLING HIGH FROM LOW VALUE.

Correlation and squared error both ask the critic for the right *number*. We also ask a weaker question, and one closer to what the policy update needs: can the critic solve the binary task of separating states whose future cost is high from the rest? Call a probe state high value when its reference

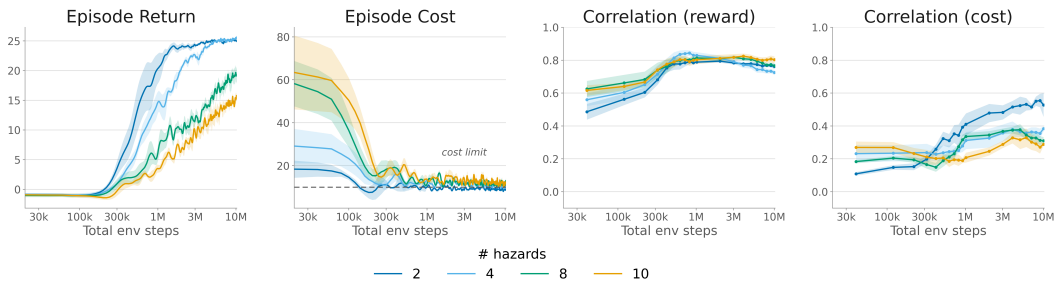


Figure 14: *Learning accurate cost value function becomes harder as the task complexity increase:* In this experiment we create an environment in SafetyGym modeled after PointGoal1, where the number of hazards can be controlled. We see that as we go from 2 hazards to 10 hazards, the correlation of cost value function with reference values substantially decreases but the reward value function doesn't share the same fate.

exceeds a threshold, $\tilde{V}^{c}(s) \geq v$, and let n_1 and n_0 be the numbers of probes above and below it. A critic addresses this task by scoring states rather than labelling them, and the standard measure of how well a score supports it is the area under the ROC curve,

$$\text{AUROC} = \frac{1}{n_1 n_0} \sum_{i\,:\,\text{high}} \sum_{j\,:\,\text{not high}} \Big[\mathbf{1}\{\hat{V}_c(s_i) > \hat{V}_c(s_j)\} + \tfrac{1}{2}\,\mathbf{1}\{\hat{V}_c(s_i) = \hat{V}_c(s_j)\} \Big], \tag{21}$$

which is the Mann–Whitney U statistic divided by $n_1 n_0$ (Mann & Whitney, 1947; Hanley & McNeil, 1982). It has a straightforward interpretation: among the $n_1 n_0$ pairs formed by taking one high-value state and one state below the threshold, it is the fraction the critic orders correctly, with ties counted as half a success. Equivalently, it is the probability that a randomly drawn high-value state receives a higher prediction than a randomly drawn state below the threshold. A critic that is constant, or that assigns predictions unrelated to the reference, scores 0.5; perfect separation scores 1.

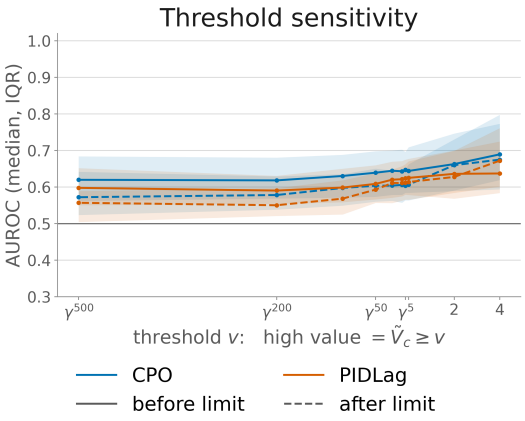


Figure 13: AUROC is quite robust to a wide range of thresholds for defining high value states.

Because equation 21 depends on the predictions only through their ranks, it is invariant to any strictly increasing transformation of the critic's output. A compressed scale, a shifted level, or predictions that are negative where the true cost-to-go cannot be, all of these leave it unchanged. Because the labels come from a finite-sample reference, we also report a ceiling (split-half ceiling) obtained by splitting the 10 rollouts of each probe into halves and using one half's estimate to rank the labels defined by the other: no critic can do better than this on these rollouts, so it allows to separate the critic's limitations from the reference's.

E Implementation Details

E.1 Hyperparameter Optimization and Selection

For the base algorithms (*CPO* and *PIDLag*) we use the optimized hyperparameters available in open-source implementations like Omnisafe (Ji et al., 2024) for the actor and do a hyperparameter search for the critic training. In particular for base methods, we search over two parameters *update-iters* or the number of updates to the critic over the whole on-policy samples and the $critic - lr$ or the learning rate for the critic.

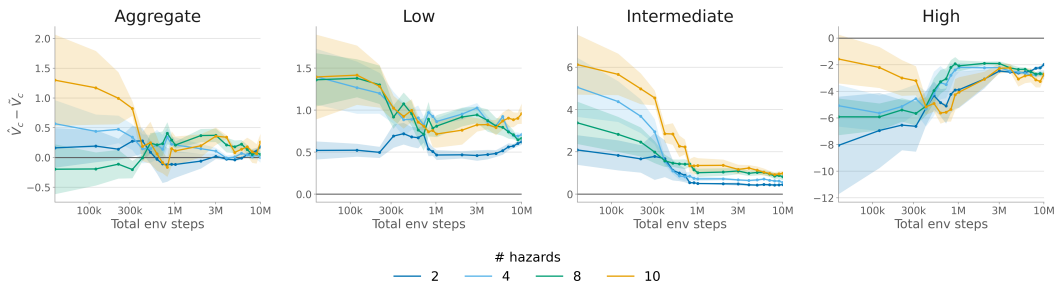


Figure 15: For very low hazard number the critic is able to estimate cost value well in the low and intermediate groups but it struggles in high value group because they are under-represented in the on-policy rollouts. On the other hand as the number of hazards increase overestimation on low and intermediate groups emerge and the estimation error on high group is quickly minimized.

For *SR* critics we optimized additionally a $\kappa \in 0.1, 0.01$ parameter which controls the ridge regression for the least squares solution to the cost and reward readouts. The parameters for InfoNCE Oord et al. (2018) that are optimized are learning rate $\phi_{lr} \in 0.0003, 0.001$ and contrastive temperature $\tau \in 0.001, 0.01$.