# BEYOND LAGRANGIAN METHODS FOR ROBUST PEAK-COST CONSTRAINED REINFORCEMENT LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study robust peak-cost constrained reinforcement learning, where the goal is to maximize reward while bounding the largest cost incurred along a trajectory – unlike the constrained MDP (CMDP), which only keeps the *expected cumulative* cost below a threshold. We first show that the problem is exactly equivalent to a CMDP on a state augmented with the running maximum cost, which restores zero duality gap but requires history-dependent policies, yields a cost signal that is active only at a new peak, and makes worst-case evaluation conservative. We therefore keep the reachability-constrained (RCRL) formulation on the original state space, for which strong duality can fail, and optimize a surrogate rather than a Lagrangian, estimating worst-case values through an integral probability metric. Our main contribution is the first convergence analysis for this surrogate. We prove that every iterate stays $\epsilon$-safe from a safe start which the primal-dual-based approach can not guarantee, and that a stationary point of the exact surrogate is reached. Under a nondegeneracy condition, we further obtain a globally $\epsilon$-optimal and $\epsilon$-safe policy in $\mathcal{O}(\epsilon^{-3})$ iterations from a safe start, and $\mathcal{O}(\epsilon^{-4})$ otherwise. Experiments on continuous-control benchmarks, including RCRL, augmented CMDP and robust CMDP baselines, show that the method sustains reward while enforcing peak-cost safety under dynamics perturbations.

## 1 INTRODUCTION

Safe decision making in cyber-physical systems requires avoiding hazardous behaviour while maximizing reward. The standard framework for safe reinforcement learning is the constrained Markov decision process (CMDP) (Altman, 2021; Efroni et al., 2020; Ghosh et al., 2024; 2023; Tessler et al., 2018), which bounds the expected *cumulative* cost of a policy. This is the right model when violations are additive, so that a large cost in one step can be offset by low costs elsewhere. It is the wrong model when a single large cost determines safety: two trajectories with the same cumulative cost differ sharply if only one of them contains a severe violation.

Many sequential tasks are of the latter kind, and safety depends on the largest cost encountered along a trajectory. In the grid-world navigation problem of Fig. 1, an agent must reach a goal while avoiding hazardous cells. A policy with low cumulative cost may still pass through a highly unsafe cell once, which is catastrophic if that cell represents a collision, hardware damage, or entry into a forbidden region. Conversely, a policy that merely avoids high instantaneous cost may take an apparently safe action that makes a later violation unavoidable. Safety is therefore trajectory-dependent, and the quantity to be constrained is the peak cost along the trajectory. A further difficulty is that policies are trained in a simulator and deployed in an environment whose dynamics differ, so the guarantee must hold under model mismatch. This motivates constraining the peak cost under the worst case over an uncertainty set of transition models.

A peak-cost constraint can be recast exactly as a CMDP by augmenting the state with the running maximum of the cost and charging the increment by which each step raises it. The augmented problem is additive and so inherits zero duality gap, but it requires policies on the enlarged state, yields a cost signal that fires only when a new peak is reached, and, in the robust setting, admits an adversary that reacts to the memory component. These drawbacks motivate the reachability-constrained formulation (RCRL) Yu et al. (2022); Ganai et al. (2023), which keeps Markov policies on the original state space and captures the peak through a non-additive value function. Existing RCRL methods are Lagrangian-based, but Mukhopadhyay et al. (2026) recently showed that peak-cost constrained MDPs need not admit strong duality, since the maximum-cost objective destroys



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.

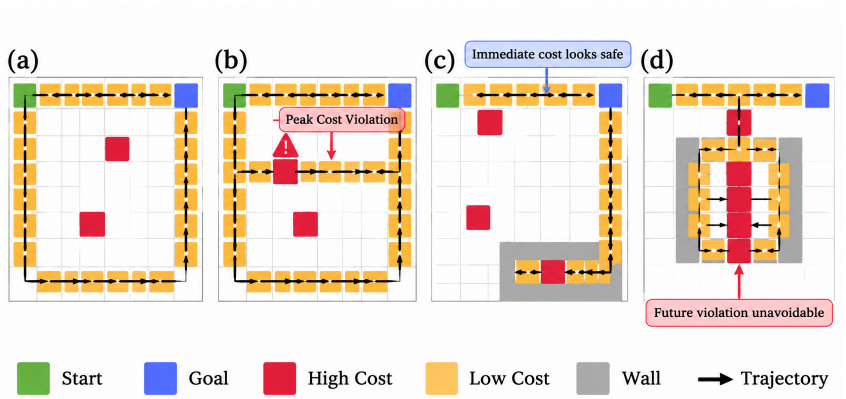



Figure 1: Grid-world motivation: cumulative-cost constraints can miss peak violations, while instantaneous-cost safety can ignore future unavoidable violations.

the convexity of the achievable values in the occupancy measure, so a primal–dual method need not recover the constrained optimum. A second limitation is temporal: the multiplier can grow only *after* a violation has been observed, so feasibility is not preserved during training even from a feasible start, and the resulting guarantee holds only on average. Building on the surrogate of Mukhopadhyay et al. (2026), which avoids dualizing the constraint, we ask what can be guaranteed for it: *that work provided no convergence analysis, no safety guarantee during training, and limited empirical validation*. The analysis is challenging, since the peak-cost value is nonsmooth and nonlinear in the occupancy measure and the worst case over the uncertainty set adds a further maximum.

**Our contributions.**

- We show that the expected peak-cost problem is *exactly* equivalent to a CMDP on the state augmented with the running peak-cost memory, recovering strong duality, and identify the three costs of this route: an enlarged state space with a history-dependent policy class, a sparse incremental cost, and a conservative robust evaluation caused by the lifted rectangular uncertainty set.
- We adopt the RCRL peak-cost value of Yu et al. (2022); Mukhopadhyay et al. (2026), which coincides with the expected peak for deterministic dynamics and policies, and, since strong duality fails for it, optimize the robust surrogate of Mukhopadhyay et al. (2026) rather than a Lagrangian. Robust evaluation uses an integral-probability-metric uncertainty set, for which the peak-cost Bellman update admits a closed form.
- We give the *first* convergence guarantees for this surrogate. The key observation is that our surrogate is the largest among a family of smooth values. This makes the surrogate weakly convex and permits a proximal analysis of the exact objective, with no smoothing. Away from the points where reward and constraint are equally active, we obtain a globally $\epsilon$-optimal and $\epsilon$-safe policy in $\mathcal{O}(\epsilon^{-3})$ iterations from a safe start and $\mathcal{O}(\epsilon^{-4})$ otherwise.
- Our surrogate keeps the constraint inside the objective being descended, rather than attaching it through a multiplier that must first be learned. We prove that the worst-case violation never rises above its initial level, so every iterate is $\epsilon$-safe when training begins from a safe policy – a guarantee no primal–dual method with the standard multiplier update can provide, and one that holds unconditionally, without the regularity conditions our optimality result requires.
- Since the proximal step is impractical at scale, we implement RP-CRL as a robust actor–critic: the critics estimate the robust reward and peak-cost values via the IPM update, and the actor takes a PPO step on whichever term of the surrogate is active. On continuous-control benchmarks with single and multiple constraints, against state-of-the-art baselines and the augmented CMDP and robust CMDP formulations, RP-CRL attains competitive reward while satisfying the peak-cost constraint under perturbations, where others either violate it or are markedly more conservative. It enforces multiple peak constraints without a multiplier per constraint; it satisfies the feasibility if starts from a feasible policy rather than oscillating across it unlike the primal-dual ones.

## 2 Related Work

**Constrained MDPs and State Augmentation.** CMDPs are the standard framework for constrained RL, bounding the expected cumulative cost Altman (2021). Linearity of the objective and the constraint in the occupancy measure yields convexity and zero duality gap Paternain et al. (2019), which underpins Lagrangian and primal–dual methods Paternain et al. (2022); Stooke et al. (2020); Tessler

et al. (2018); Zheng & Ratliff (2020); Ghosh et al. (2023); Ding et al. (2020); Ghosh et al. (2024); primal alternatives avoid multipliers altogether Xu et al. (2021); Achiam et al. (2017); Chow et al. (2018), and FAC Ma et al. (2021) learns state-dependent multipliers that diverge on infeasible states. Trajectory-level requirements are often made Markovian by augmenting the state: Sauté RL Sootla et al. (2022) carries the remaining safety budget, while anytime-constrained RL McMahan & Zhu (2024) carries the accumulated cost, showing that Markovian policies no longer suffice. Instantaneous constraints are also considered Bai et al. (2023); Amani et al. (2021). As we mentioned our peak-cost constraint formulation differs from the above. While complexity guarantees are provided for CMDP Ding et al. (2020); Xu et al. (2021), the peak-cost is not linear and gradient dominance condition cannot be established.

**Robust Constrained MDPs.** Robust MDPs optimize against the worst case over an uncertainty set, with rectangularity keeping robust dynamic programming tractable Iyengar (2005); Wiesemann et al. (2013) Their constrained counterparts lose convexity in the occupancy measure Wang et al. (2022), and existing methods either assume strong duality without full iteration guarantees Zhang et al. (2024) or use epigraph and binary-search formulations that are costly and sensitive to value-estimation noise Kitamura et al. (2025). Ma et al. (2025) proposed a robust CRPO, however, did not provide iteration complexity guarantee for any arbitrary $\epsilon > 0$. A surrogate objective with iteration-complexity guarantees removes the binary search Ganguly et al. (2026); we adapt it to the peak-cost setting, where the constraint is non-additive and its value function nonsmooth in the policy, hence, those results do not extend here.

**Feasible Sets and Reachability Analysis.** Feasible-set characterization is central to safe control and RL Brunke et al. (2022). Control barrier functions give certificates but are conservative and model-dependent Choi et al. (2021); Luo & Ma (2021); Ma et al. (2022), and Hamilton–Jacobi reachability is rigorous yet limited by PDE cost Bansal et al. (2017). RL-based formulations learn safety-preserving policies via a discounted safety Bellman recursion Fisac et al. (2018; 2019); So et al. (2024), and RCRL methods optimize reward and peak-cost safety jointly Yu et al. (2022); Ganai et al. (2023), with RESPO Ganai et al. (2023) extending this to stochastic settings. All are Lagrangian-based, and did not provide any convergence rate guarantee. Mukhopadhyay et al. (2026) showed that the peak-cost problem may fail to admit zero duality gap and proposed the robust surrogate and IPM-based value estimation on which we build; convergence analysis and large-scale numerical evaluations with continuous control were left open.

## 3 Preliminaries and Problem Formulation

**Constrained Markov Decision Processes** A CMDP is given by $M = \langle \mathcal{S}, \mathcal{A}, P, r, c, \gamma, d_0 \rangle$, where $\mathcal{S}, \mathcal{A}$ are state and action spaces, $P$ is the transition kernel, $r$ and $c$ are reward and cost functions, $\gamma \in (0, 1)$ is the discount factor, and $d_0$ is the initial distribution. A policy is denoted by $\pi(\cdot \mid s)$.

For policy $\pi$ and model $P$, we define the discounted negative-reward and cost values as

$$V_r^{P,\pi}(s) = -\mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) | s_0 = s, \pi, P\big], \qquad V_c^{P,\pi}(s) = \mathbb{E}\big[\textstyle\sum_{t=0}^{\infty} \gamma^t c(s_t,a_t) | s_0 = s, \pi, P\big].$$

The corresponding objectives are $J_r^P(\pi) = \mathbb{E}_{s\sim d_0}[V_r^{P,\pi}(s)]$ and $J_c^P(\pi) = \mathbb{E}_{s\sim d_0}[V_c^{P,\pi}(s)]$. A standard CMDP solves $\min_\pi \; J_r^P(\pi) \quad \text{s.t.} \quad J_c^P(\pi) \le b$. Under Slater's condition, CMDPs admit zero duality gap Paternain et al. (2019), making Lagrangian methods effective. Without loss of generality, we assume $-r \in [0, 1]$.

**Peak-Cost Constraints:** CMDPs constrain cumulative cost, but many safety-critical tasks require controlling the largest violation along a trajectory. For example, the safety requires that the peak cost along $\tau = (s_0, s_1, \ldots)$ remains below $b$ where the cost at a state might be inversely proportional to the distance from some obstacles. This motivates the peak-cost constrained problem

$$\min_\pi \; J_r^P(\pi) \qquad \text{s.t.} \qquad \mathbb{E}_{s_0\sim d_0}\big[\max_{t\le T-1} c(s_t, a_t) | s_0\big] \le b, \tag{1}$$

where $s_t$ is the state at time $t$. In finite-horizon settings, $t$ is restricted to $t \le T$. We assume at least one feasible policy exists.

**CMDP Reformulation of the Peak-Cost Constraint** The peak-cost constraint in Problem equation 1 is non-additive, since it depends on the maximum cost along a trajectory rather than a sum of per-step costs. Nevertheless, in the following we show that how the constraint can be recovered as a CMDP in the augmented state.

**Proposition 1.** *[Exact CMDP Reformulation of Peak Cost] Assume $c(s,a) \geq 0$. Define the augmented state $\bar{s}_t = (s_t, m_t)$ with $m_0 = 0$ and $m_{t+1} = \max\{m_t, c(s_t,a_t)\}$, and the incremental cost $\Delta(\bar{s}_t, \bar{a}_t) = (c(s_t,a_t) - m_t)_+$. Then for every trajectory, $\Delta(\bar{s}_t, \bar{a}_t) = m_{t+1} - m_t$, and hence $\sum_{t=0}^{T-1} \Delta(\bar{s}_t, \bar{a}_t) = m_T = \max_{0\leq t\leq T-1} c(s_t,a_t)$. Consequently, Problem equation 1 over history-dependent policies is equivalent to the additive CMDP on $\bar{S}$ with cost $\Delta$ over policies $\bar{\pi}(a \mid s, m)$, with identical optimal values.*

Thus, the augmented CMDP converts the trajectory-level peak constraint into an additive one and can recover the standard zero-duality-gap structure under usual CMDP conditions such as Slater's condition Paternain et al. (2019).

**Limitations of the augmentation:**

***Issue 1: Increased and potentially continuous state space.*** State augmentation enlarges the state space from $S$ to $\bar{S} = S \times \mathcal{M}$, where $\mathcal{M}$ is the range of the running maximum. Even when $S$ is finite, $\mathcal{M}$ can be continuous. For example, if $c_t \in [0, c_{\max}]$, then $\mathcal{M} = [0, c_{\max}]$.

and the augmented state space is uncountable. Consequently, tabular CMDP methods whose guarantees depend on the cardinality of the state space no longer apply directly. Discretizing the running maximum introduces an additional approximation error, while continuous augmentation requires function approximation over the additional state dimension. Further, the policy class now has the form $\bar{\pi}(a \mid s, m)$ instead of only depending on $s$.

***Issue 2:***: Although the reformulation is exact, the incremental cost $\Delta_t = (c_t - m_t)_+$ can be sparse because it is nonzero only when a new trajectory peak is reached. This can lead to high-variance cost estimates in model-free learning, particularly for long-horizon problems (See Figure 9 in Appendix C.2). Auxiliary dense signals, such as

$$\tilde{c}_t = \eta c_t + \alpha_t \Delta_t,$$

may provide additional learning feedback, but such shaping does not by itself preserve the original peak-cost constraint unless $\eta$ is 0 and $\alpha \approx 1$.

***Issue 3: Robustness under augmentation.*** A central motivation of this work is robustness to model mismatch, which we formalize in Section 4 through an uncertainty set over the transition kernel. The uncertainty concerns only the physical dynamics $P(s' \mid s, a)$, while the running maximum evolves deterministically. However, a worst-case model in the augmented space may select different kernels for the same $(s, a)$ depending on $m$, an adversary strictly more powerful than the physical uncertainty warrants. The resulting robust evaluation is therefore conservative in general. Moreover, even without augmentation, strong duality can fail for robust CMDPs , so the augmented reformulation does not by itself restore a zero duality gap in the robust setting.

## 4 RCRL Formulation

**Reachability-Constrained Formulation:** The augmented reformulation is exact, but it requires policies over the enlarged space $\bar{S}$ and suffers from the issues discussed above. An alternative, following Yu et al. (2022), keeps Markov policies $\pi \in \Pi_S$ on the original state space and captures the peak cost through a non-additive value function. For a fixed policy $\pi$, define

$$Q^{\pi,P}_{c,\text{peak}}(s,a) = (1-\gamma)c(s,a) + \gamma \max\left\{c(s,a),\, \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V^{\pi,P}_{c,\text{peak}}(s')\right]\right\}, \tag{2}$$

$$V^{\pi,P}_{c,\text{peak}}(s) = \left\langle \pi(\cdot \mid s),\, Q^{\pi,P}_{c,\text{peak}}(s,\cdot)\right\rangle. \tag{3}$$

Since $|\max\{c,x\} - \max\{c,y\}| \leq |x-y|$, the induced operator is a $\gamma$-contraction in the sup-norm for $\gamma < 1$, and $V^{\pi,P}_{c,\text{peak}}$ is its unique fixed point Fisac et al. (2019). We then consider

$$\min_{\pi\in\Pi_S} J^P_r(\pi) \qquad \text{s.t.} \qquad \max_{s_0\in\mathcal{S}_0} V^{\pi,P}_{c,\text{peak}}(s_0) \leq b, \tag{4}$$

where $\mathcal{S}_0$ is the set of initial states.

**Remark 1.** *$V^{\pi,P}_{c,\text{peak}}$ is not, in general, the expected peak cost. Because the expectation over actions and next states is taken inside the maximum, Jensen's inequality gives, for $\gamma = 1$, $V^{\pi,P}_{c,\text{peak}}(s) \leq$*

$\mathbb{E}[\max_t c(s_t, a_t) \mid s_0 = s]$, *with equality when both* $P$ *and* $\pi$ *are deterministic. Problem equation 4 is therefore a distinct formulation rather than a reformulation of Problem equation 1. However, the formulations match when* $P$ *and* $\pi$ *are deterministic for* $\gamma \to 1$.

Yu et al. (2022) solve Problem equation 4 through the Lagrangian

$$\min_{\pi \in \Pi_S} \max_{\lambda \geq 0} \; J_r^P(\pi) + \textstyle\sum_{s_0 \in \mathcal{S}_0} \lambda(s_0) \big( V_{c,\text{peak}}^{\pi,P}(s_0) - b \big), \tag{5}$$

with state-dependent multipliers $\lambda(s_0)$. Unlike the augmented CMDP, the constraint in equation 4 is not linear in the occupancy measure of $\pi$, strong duality can fail, so saddle points of equation 5 need not recover solutions of equation 4.

Unlike standard CMDPs, the formulation in equation 5 does not come with a strong-duality guarantee. This was formalized in Mukhopadhyay et al. (2026); we restate the result.

**Theorem 1** (Mukhopadhyay et al. (2026))**.** *There exists an MDP with a single initial state,* $|\mathcal{S}_0| = 1$*, for which Problem equation 4 satisfies Slater's condition but has a strictly positive duality gap: the optimal value of equation 4 differs from that of its Lagrangian dual. Since* $|\mathcal{S}_0| = 1$*, the state-dependent multiplier in equation 5 reduces to a scalar, so the gap applies to equation 5 directly. See Section A.2 for the proof.*

**Remark 2** (Implication for primal-dual methods)**.** *For standard CMDPs, zero duality gap follows from the convexity of the set of achievable objective–constraint pairs, which in turn follows from the linearity of both in the occupancy measure. Theorem 1 shows that this convexity fails for the peak-cost value* $V_{c,\text{peak}}^{\pi,P}$*. The Lagrangian dual instead equals the value of the convexified problem, which is attained by a policy that randomizes between Markov policies at the start of each episode. In the example of Theorem 1, the dual value is achieved by running a policy with* $V_{c,\text{peak}}^{\pi,P}(s_0) > b$ *in a fixed fraction of episodes and a conservative policy in the rest. Such a mixture satisfies the constraint only on average across episodes, which defeats the purpose of a peak-cost constraint. Primal-dual methods whose guarantees rely on averaged iterates converge toward this mixture rather than toward a solution of equation 4. This motivates our approach: a primal method that optimizes directly over* $\Pi_S$ *without dualizing the peak-cost constraint.*

**Remark 3** (Contrast with the augmented CMDP)**.** *The gap is specific to equation 4, which uses Markov policies on* $\mathcal{S}$ *and the non-additive constraint* $V_{c,\text{peak}}^{\pi,P}$*. The expected peak-cost problem equation 1 over the augmented state* $\bar{s} = (s, m)$ *is instead an additive CMDP, linear in the augmented occupancy measure, and thus has zero duality gap under Slater's condition Paternain et al. (2019). Strong duality is recovered by changing both the policy class, to* $\bar{\pi}(a \mid s, m)$*, and the constraint, to the expected peak (these coincide only in the deterministic case), at the cost of the issues in Section 3 .*

**Robust Formulation:** Problem equation 4 assumes that training and deployment dynamics coincide. Policies trained in simulation may face different real-world transitions and violate safety constraints after deployment. We therefore consider the robust counterpart

$$\min_{\pi \in \Pi_S} \max_{P \in \mathcal{P}} \; J_r^P(\pi) \qquad \text{s.t.} \qquad \max_{P \in \mathcal{P}} \max_{s_0 \in \mathcal{S}_0} V_{c,\text{peak}}^{\pi,P}(s_0) \leq b, \tag{6}$$

which optimizes worst-case reward subject to worst-case peak-cost safety, with the two worst cases taken separately. We use $(s,a)$-rectangular sets $\mathcal{P} = \bigotimes_{(s,a)} \mathcal{P}_{(s,a)}$ with $\mathcal{P}_{(s,a)} = \{P \in \Delta(\mathcal{S}) : D(P, P_0(\cdot \mid s, a)) \leq \rho\}$, where $P_0$ is the nominal simulator model. Rectangularity yields tractable robust dynamic programming Iyengar (2005), whereas robust evaluation over general sets is NP-hard Wiesemann et al. (2013). Since $x \mapsto \max\{c, x\}$ is monotone, the robust peak-cost operator

$$(\mathcal{T}_{\mathcal{P}}^{\pi} Q)(s,a) = (1-\gamma) c(s,a) + \gamma \max \Big\{ c(s,a), \sup_{P \in \mathcal{P}_{(s,a)}} \mathbb{E}_P \big[ \langle \pi(\cdot \mid s'), Q(s', \cdot) \rangle \big] \Big\}$$

is a $\gamma$-contraction whose fixed point gives $\max_{P \in \mathcal{P}} V_{c,\text{peak}}^{\pi,P}$. Note that an $(s,a)$-rectangular set on $\mathcal{S}$ lifts to an $(s,m,a)$-rectangular set on the augmented space, strictly enlarging the adversary (cf. Issue 3) for the state-augmented CMDP.

***IPM uncertainty sets.*** For large or continuous state spaces, we use an integral probability metric Zhou et al. (2023), $d_F(p,q) = \sup_{f \in F} \{\mathbb{E}_p[f] - \mathbb{E}_q[f]\}$, with $F = \{\Psi(\cdot)^\top \xi : \|\xi_{2:d}\|_2 \leq 1\}$ and

$\Psi_1 \equiv 1$. For a linear peak-cost critic $V_w(s) = \Psi(s)^\top w$, the robust update admits an upper bound

$$(\mathcal{T}^\pi_{\mathcal{P}} Q_w)(s,a) \le (1-\gamma)c(s,a) + \gamma \max\Big\{c(s,a),\, \mathbb{E}_{P_0(\cdot|s,a)}\big[V_w(s')\big] + \rho\|w_{2:d}\|_2\Big\}, \tag{7}$$

where $w_{2:d}$ excludes the weight of the constant feature. The derivation is in Appendix B.1. Since this is an upper bound, we use this for robust peak cost evaluation.

**Remark 4** (Robust evaluation under augmentation)**.** *Since the augmented cost $\Delta(\bar{s},a) = (c(s,a) - m)_+$ is additive, the robust operator is the standard one, $\bar{Q}_h(\bar{s},a) = \Delta(\bar{s},a) + \mathbb{E}_{P_0(\cdot|s,a)}[\bar{V}_{h+1}(s',m')] + \rho\|\bar{V}_{h+1}(\cdot,m')\|_F$, with $m' = \max\{m, c(s,a)\}$ fixed before $s'$ is drawn. The cost is one critic per value of the running maximum and an $(s,m,a)$-rectangular adversary, hence a conservative worst case (Issue 3), in exchange for dropping the non-additive maximum. See Appendix B.2.*

**Robust Peak-Cost Surrogate Formulation** Since equation 6 may have a nonzero duality gap (Theorem 1), we avoid dualizing the constraint. Let $J(\pi) = \max_{P\in\mathcal{P}} J_r^{\pi,P}$ and $W(\pi) = \max_{P\in\mathcal{P},\, s_0\in\mathcal{S}_0} V_{c,\text{peak}}^{\pi,P}(s_0) - b$. Following Ganguly et al. (2026), we solve the robust surrogate

$$\min_{\pi\in\Pi_{\mathcal{S}}} F_\beta(\pi) := \max\big\{\tfrac{J(\pi)}{\beta},\, W(\pi)\big\}. \tag{8}$$

For large $\beta$, the violation term dominates while the constraint is violated, and the reward term dominates once feasibility is achieved.

**Proposition 2.** *Let $\pi^*$ solve equation 6, and let $\hat{\pi}$ satisfy $F_\beta(\hat{\pi}) \le \min_\pi F_\beta(\pi) + \delta$ for some $\delta \ge 0$. Then*

$$J(\hat{\pi}) \le J(\pi^*) + \beta\delta, \qquad W(\hat{\pi})_+ \le \tfrac{1}{(1-\gamma)\beta} + \delta.$$

*In particular, if $\beta \ge 1/((1-\gamma)\epsilon)$, then $W(\hat{\pi})_+ \le \epsilon + \delta$, and for an exact minimizer ($\delta = 0$), $J(\hat{\pi}) \le J(\pi^*)$ and $W(\hat{\pi})_+ \le \epsilon$.*

The Proof is in Appendix A.1. Proposition 2 highlights the strength of the surrogate problem. Also, $\beta$ is one-sided (i.e., it only needs to be above $O(1/\epsilon)$) unlike the CMDP, where one needs to properly tune the dual variable.

**Theoretical Results: Convergence Guarantees** We optimize $F_\beta$ in equation 8 over the direct parametrization $\Pi = \prod_{s\in\mathcal{S}} \Delta(\mathcal{A}) \subset \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by the proximal point iteration

$$\pi^{t+1} \in \arg\min_{\pi\in\Pi} \Big\{F_\beta(\pi) + \tfrac{1}{2\mu}\|\pi - \pi^t\|_2^2\Big\}, \qquad t = 0, 1, \dots \tag{9}$$

Throughout, $H = (1-\gamma)^{-1}$, and $\pi^\star$ denote a solution of (6). $F_\beta(\pi)$ $\pi$, but it is a pointwise maximum of smooth *branch values* which preserves weak convexity. Consequently $F_\beta$ is $\omega$-weakly convex with $\omega = 2\gamma|\mathcal{A}|H^3$ for $\beta \ge 1$, uniformly over the uncertainty set, and with $\mu = 1/(2\omega)$ each subproblem in equation 9 is $\omega$-strongly convex (Lemma 4).

**Assumption 1.** *$\mathcal{S}$, and $\mathcal{A}$ are finite, $c, -r \in [0,1]$, and $\mathcal{P}$ is a compact set of transition kernels.*

Let $\Delta_0 := F_\beta(\pi^0)$ measure the initialization: $\Delta_0 \le \max\{H/\beta, W(\pi_0) - b\}$ in general, and $\Delta_0 = J(\pi^0)/\beta \le H/\beta$ if $\pi^0$ is feasible for equation 6. We use the above for theoretical guarantees, our empircal evaluations consider both infinite $\mathcal{S}$ and $\mathcal{A}$.

**Theorem 2** (Monotone safety and stationarity)**.** *Let Assumption 1 hold, $\beta \ge 1$, $\rho = 2\gamma|\mathcal{A}|H^3$, $\mu = 1/(2\omega)$. Then $(W(\pi^t) - b)_+ \le \Delta_0$ and $J(\pi^t) \le \beta\Delta_0$ for all $t$, and*

$$\min_{0\le t<T} \|\pi^{t+1} - \pi^t\|_2 \le 2\mu\sqrt{\omega\Delta_0/T}. \tag{10}$$

*In particular, from a feasible $\pi^0$ every iterate satisfies $(W(\pi^t) - b)_+ \le J(\pi^0)/\beta$ and $J(\pi^t) \le J(\pi^0)$. Since $\beta = O(\epsilon^{-1})$, thus, $(W(\pi_t) - b)_+ \le O(\epsilon)$, and $\min_t ||\pi_{t+1} - \pi_t||_2 \le O(\epsilon)$ if $T = O(\epsilon^{-1})$ when $\pi_0$ is feasible. If $\pi^0$ is infeasible then $\min_t ||\pi_{t+1} - \pi_t|| \le \epsilon$ if $T = O(\epsilon^{-2})$.*

**Theorem 3** (Global optimality at nondegenerate points)**.** *Let $\bar{t}$ attain the minimum in equation 10, and suppose all branches of $F_\beta$ active at $\pi^{\bar{t}+1}$ share a common gradient and the mismatch coefficient $C$ of Definition 2 is finite. Then, with $E_T = 2C\sqrt{2|\mathcal{S}|\omega\Delta_0/T}$, every $\pi^t$ with $t \ge \bar{t}+1$ – in particular $\pi^T$, which minimizes $F_\beta$ over the iterates – satisfies*

$$J(\pi^t) \le J(\pi^\star) + \beta E_T, \qquad \big(W(\pi^t) - b\big)_+ \le \min\{\Delta_0,\, J(\pi^\star)/\beta + E_T\}. \tag{11}$$

*Consequently, for $\epsilon \in (0, H]$ and $\beta = H/\epsilon$, the output is $\epsilon$-optimal and $\mathcal{O}(\epsilon)$-feasible after $T = \mathcal{O}\big(\gamma C^2|\mathcal{S}||\mathcal{A}|H^5\epsilon^{-3}\big)$ iterations from a feasible $\pi^0$ (since $\Delta_0 = J(\pi^0)/\beta$), and after $T = \mathcal{O}\big(\gamma C^2|\mathcal{S}||\mathcal{A}|H^5\epsilon^{-4}\big)$ from an arbitrary one.*

This is the first convergence result for (6) and did not use any primal-dual approach which can be of independent interest for the community. Theorem 3 states that the iteration complexity is $T = \mathcal{O}(\epsilon^{-3})$ when the initial policy is feasible; otherwise it is $\mathcal{O}(\epsilon^{-4})$. The extra $1/\epsilon$ relative to the $\epsilon^{-2}$ of primal-dual methods for standard CMDPs is the price of a penalty of a lack of a strong duality. A feasible initialization, e.g. a known conservative controller, is worth a further factor of $1/\epsilon$, since it makes $\Delta_0 \leq \epsilon$. Theorem 3 additionally requires $C$ and nondegeneracy in addition to Assumption 1. In particular, if both the branches are active, then, the gradients need to match. Relaxing this assumption is left for the future.

Theorem 2 needs no condition beyond Assumption 1, Also, Theorem 2 states that if the initial policy $\pi^0$ is feasible, the policy would be $\epsilon$-safe throughout the training, a guarantee which the primal-dual approaches cannot provide. Unlike the primal-dual, the convergence is on the best-iterate, thus, one does not need to store all the iterates and average it over. Further, the convergence to stationary point is $\mathcal{O}(\epsilon^{-1})$ starting from a feasible policy; otherwise, it is $\mathcal{O}(\epsilon^{-2})$.

**Practical instantiation.** The proximal step equation 9 is stated with an exact subproblem solve and a Euclidean penalty, which is convenient for analysis but impractical at scale. We therefore implement it with a clipped policy update in the style of PPO Schulman et al. (2017), using the robust actor–critic where critic is trained using IPM to estimate $J$ and $W$: the clipping (equivalently, a KL penalty) plays the role of the proximal term $\|\pi - \pi^t\|^2/2\mu$ in keeping $\pi^{t+1}$ close to $\pi^t$, while the objective remains the surrogate $F_\beta$. Details of our implementation are provided in Appendix C.1.

## 5 Experiments

**Environments, Baselines and Protocol:** We evaluate RP-CRL on Safe Mujoco(AppendixC.4), Safe Control Gym, and Safety Gym, covering low- to high-dimensional observations up to **376D**; state–action dimensions are in Appendix C.5. In Safe Mujoco, costs restrict joint torques to prevent aggressive or destabilizing motions. In Safe Control Gym, a two-rotor quadrotor is constrained by lower and upper thrust limits on each rotor, yielding four constraints. In Safety Gym, we use the circle task, where the agent must move quickly along a circular path without leaving its perimeter. Cost functions, safety thresholds, uncertainty radii, hyperparameters, policy parameterization, and network details are provided in Appendices Section C.3.1,C.3.2,C.3.3,C.4,C.5 and C.6.

We compare with Primal–Dual (PD), FAC Ma et al. (2021), RCRL Yu et al. (2022), RESPO Ganai et al. (2023), and the Surrogate Objective (SO), which trains equation 8 without robustness in the nominal environment. RP-CRL uses the same objective but with the robust IPM critics from Section 4. Results are averaged over N seeds, with shaded regions showing one standard deviation. Inference is performed in perturbed environments; perturbation levels and designs are listed in Appendix C.3.3 and C.6. Safety is evaluated using empirical trajectory peak cost averaged over evaluation episodes, rather than the learned critic, to avoid critic approximation error.

**Single-Constraint Results:** Figs. 2 and 3 show training and inference results. In nominal training, all methods learn to satisfy the constraint while improving return, so differences appear mainly under perturbation. At inference, RP-CRL gives the best overall return–safety tradeoff. PD attains the highest return on Ant and Swimmer but violates the peak-cost threshold in both, making its returns incomparable at matched safety. RP-CRL outperforms RCRL in HalfCheetah and Humanoid and remains competitive in Swimmer maintaining satisfying safety in each environment, while RCRL becomes unsafe under perturbation on Ant. SO follows RP-CRL in the nominal setting but loses both return and safety under perturbation, isolating the benefit of robust training since SO and RP-CRL differ only in that component. RESPO is more conservative and does poorly with respect to reward, consistent with its feasibility objective weighting constraints by predicted safety probability.

**Multi-constraint results** are in Appendix C.2 which also show a similar trend.

**Comparison with the Augmented CMDP Formulation:** Fig. 4 compares RP-CRL with the augmented CMDP and robust CMDP (RCMDP) (Appendix B.2) reformulations of Section 3. Both augmented variants exceed the inference peak-cost threshold, consistent with Section 3; it is likely due to the added per-step cost which makes the constraint relaxed. RCMDP improves CMDP re-

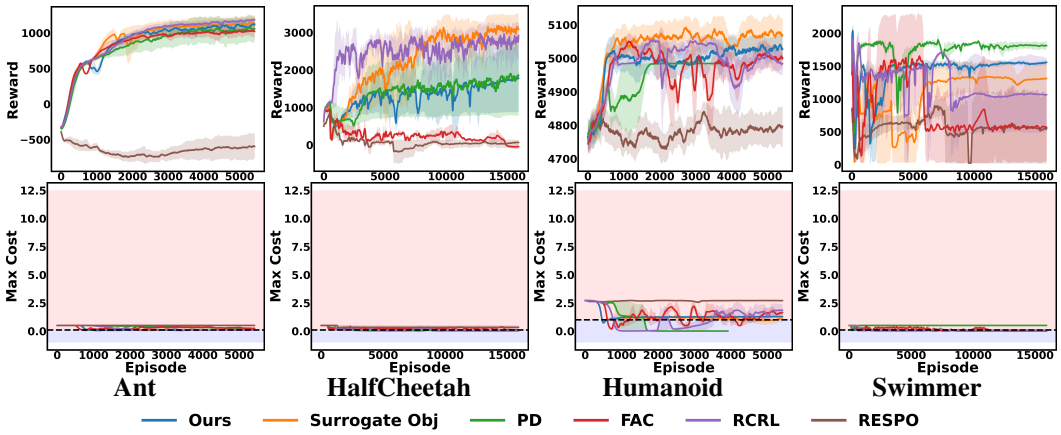



Figure 2: Training performance under a single peak-cost constraint across four environments. Top row: cumulative reward; bottom row: maximum trajectory cost.

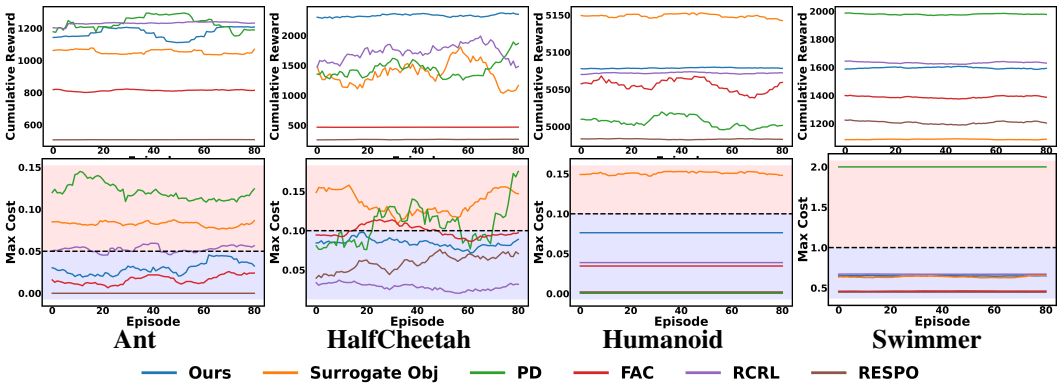



Figure 3: Inference performance under perturbed dynamics for single-constraint tasks. RP-CRL achieves the best overall trade-off between return and persistent safety, while high-reward PD baseline violates peak-cost constraints.

turns through IPM-based robust training and attains higher return than RP-CRL on Humanoid and Safety Gym, with similar returns on HalfCheetah, but violates the peak constraint at every environment. While RCRL achieves slightly higher return on Safety Gym, its violation rate is worse than RP-CRL (23% vs. 16%).

**Effect of the Surrogate Parameter $\beta$:** Proposition 2 bounds the violation by $1/((1-\gamma)\beta) + \delta$, so, as $\beta$ increases the violation decreases. Fig. 5 matches this on Ant, HalfCheetah and Swimmer ($b = 0.1$): $\beta \geq 3000$ RP-CRLconverges below the threshold. As the Proposition 2 indicates return is not sacrificed: $\beta = 300$ attains the *lowest* return on Humanoid and Swimmer, as expected since an exact minimizer incurs no return penalty from larger $\beta$ and unsafe behaviour is itself reward-costly here. On Humanoid, whose peak costs are higher, we need $\beta = 30000$ to make violation near 0.

**Effect of Feasible Policy Initialization:** Theorem 2 guarantees that a feasible start keeps the violation at its initial level and improves the complexity from $\mathcal{O}(\epsilon^{-4})$ to $\mathcal{O}(\epsilon^{-3})$. In Fig. 6, RP-CRL started feasible holds the peak cost at the threshold throughout each environment and has lower across-seed variance. The contrast with RCRL is sharper: infeasible RCRL on Humanoid drives the peak cost to zero and back to $\approx 1.9$, well above $b = 1.2$, the oscillation expected of a multiplier that reacts only after a violation. This trend is observed in HalfCheetah and Swimmer as well. Even for Ant, the RCRL takes a longer time to stabilize and make the violation near 0. Feasible RCRL on Swimmer collapses in return near episode 3800, whereas RP-CRL remains stable.

## 6 Conclusion and Future Work

We studied robust peak-cost constrained RL, where safety depends on the largest violation along a trajectory rather than the cumulative cost. The problem is exactly equivalent to a CMDP on a state augmented with the running maximum, which restores strong duality but is conservative under

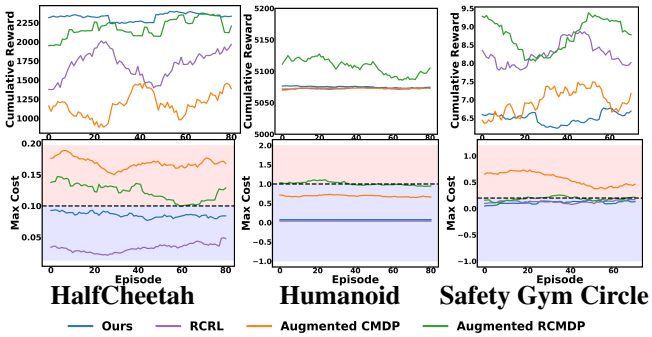



Figure 4: Inference comparison of RP-CRL, RCRL, Augmented CMDP, and Augmented RCMDP reformulations. RP-CRL maintains lower persistent-safety violations. CMDP-style approaches frequently exceed the safety threshold. RCRL violates 23% vs. RP-CRL 16% in Safety Gym.

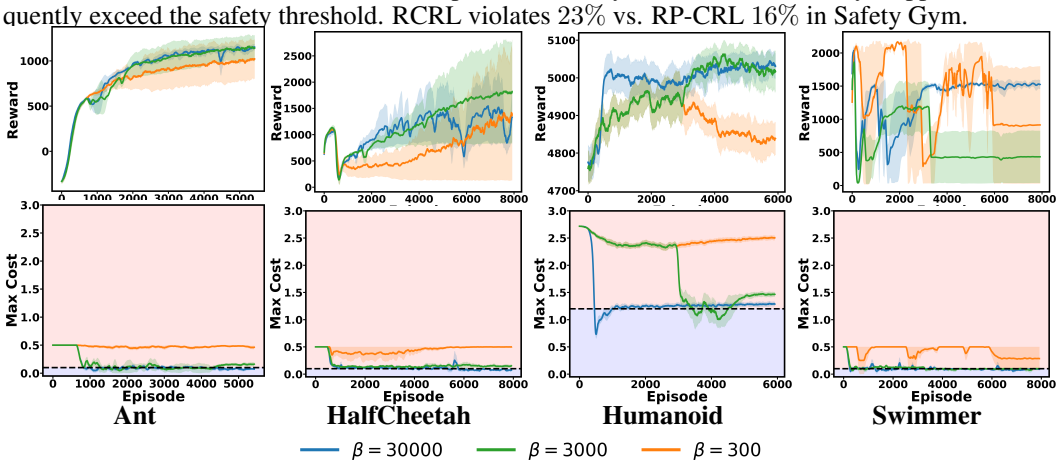



Figure 5: Ablation study on the parameter $\beta$ during training. The results illustrate that with increasing $\beta$, safety learning gets better.

model uncertainty; we therefore analyzed the surrogate on the original state space, giving its first convergence guarantees. The iterates never raise the worst-case violation above its initial level, so safety holds throughout training from a safe start, and reach a stationary point of the exact objective; under nondegeneracy and coverage conditions the output is also near-optimal. Empirically, RP-CRL attains competitive reward while maintaining peak-cost safety under perturbations.

The global guarantee requires a single active branch at the returned policy; whether gradient domination extends to points where the reward and constraint terms are equally active, or spurious stationary points arise there, is the main open question, and would also settle whether the $\mathcal{O}(\epsilon^{-3})$ complexity can be improved. Evaluating on physical systems, where the sim-to-real mismatch is present by construction, are natural next steps.

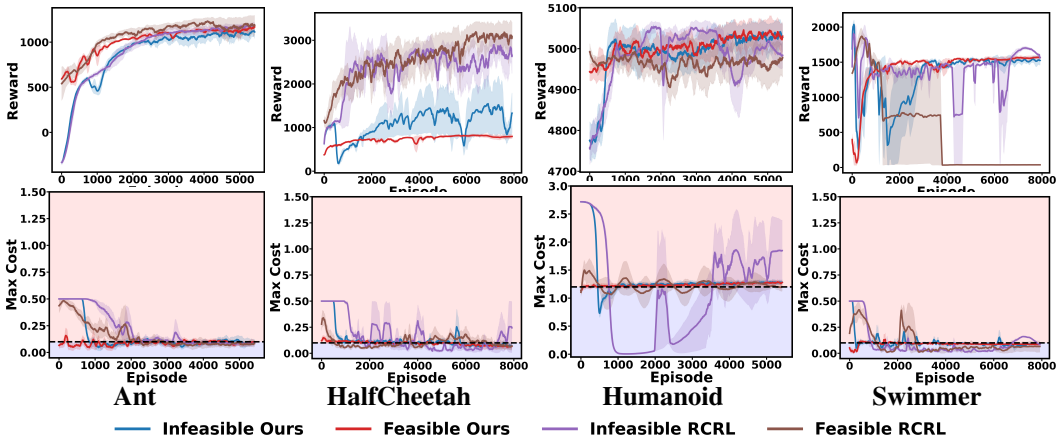



Figure 6: Training comparison between feasible and infeasible policy initialization. RP-CRL exhibits stable learning compared to RCRL which needs dual variable tuning.

### AI USE STATEMENT

In this work, we used generative AI tools for polishing the write-up and sometimes while polishing code. We have not used generative AI tools for research ideation or brainstorming, problem formulation, problem design, solutions, experimental designs, story development. We have reviewed all AI-assisted work. We checked LLM-generated text for grammatical errors, hallucinations if any, ensuring no modifications in our original concept of writing is done while polishing through manual readings by all co-authors multiple times. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

Our work is responsible for safety in AI agents. We develop the theoretical aspects of safety and validate them through simulations. We have not directly involved humans in the experiments. Hence, we have no ethical concerns for this work.

### REPRODUCIBILITY STATEMENT

We will make all the code public with detailed instructions to reproduce them for future researchers.

## REFERENCES


Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. Pmlr, 2017.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on learning theory*, pp. 64–66. PMLR, 2020.

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

Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 243–253. PMLR, 2021.

Qinbo Bai, Vaneet Aggarwal, and Ather Gattami. Provably sample-efficient model-free algorithm for mdps with peak constraints. *Journal of Machine Learning Research*, 24(60):1–25, 2023.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances. In *2017 IEEE 56th annual conference on decision and control (CDC)*, pp. 2242–2253. IEEE, 2017.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

Jason J Choi, Donggun Lee, Koushil Sreenath, Claire J Tomlin, and Sylvia L Herbert. Robust control barrier–value functions for safety-critical control. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 6814–6821. IEEE, 2021.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

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.

Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. Iterative reachability estimation for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36:69764–69797, 2023.

Sourav Ganguly, Kishan Panaganti, Arnob Ghosh, and Adam Wierman. Efficient policy optimization in robust constrained mdps with iteration complexity guarantees. *Advances in Neural Information Processing Systems*, 38:9398–9437, 2026.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pp. 1054–1062. PMLR, 2024.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Toshinori Kitamura, Tadashi Kozuno, Wataru Kumagai, Kenta Hoshino, Yohei Hosoe, Kazumi Kasaura, Masashi Hamaya, Paavo Parmas, and Yutaka Matsuo. Near-optimal policy identification in robust constrained markov decision processes via epigraph form. In *International Conference on Learning Representations*, volume 2025, pp. 56945–56981, 2025.

Yuping Luo and Tengyu Ma. Learning barrier certificates: Towards safe reinforcement learning with zero training-time violations. *Advances in Neural Information Processing Systems*, 34:25621–25632, 2021.

Haitong Ma, Yang Guan, Shegnbo Eben Li, Xiangteng Zhang, Sifa Zheng, and Jianyu Chen. Feasible actor-critic: Constrained reinforcement learning for ensuring statewise safety. *arXiv preprint arXiv:2105.10682*, 2021.

Haitong Ma, Changliu Liu, Shengbo Eben Li, Sifa Zheng, and Jianyu Chen. Joint synthesis of safety certificate and safe control policy using constrained reinforcement learning. In *Learning for Dynamics and Control Conference*, pp. 97–109. PMLR, 2022.

Shaocong Ma, Ziyi Chen, Yi Zhou, and Heng Huang. Rectified robust policy optimization for model-uncertain constrained reinforcement learning without strong duality. *arXiv preprint arXiv:2508.17448*, 2025.

Jeremy McMahan and Xiaojin Zhu. Anytime-constrained reinforcement learning. In *International conference on artificial intelligence and statistics*, pp. 4321–4329. PMLR, 2024.

Shilpa Mukhopadhyay, Sourav Ganguly, Santosh Mohan Rajkumar, Honghao Wei, Debdipta Goswami, and Arnob Ghosh. Robust peak-cost constrained reinforcement learning. *arXiv preprint arXiv:2607.15457*, 2026.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 68(3):1321–1336, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Oswin So, Cheng Ge, and Chuchu Fan. Solving minimum-cost reach avoid using reinforcement learning. *Advances in Neural Information Processing Systems*, 37:30951–30984, 2024.

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.

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.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2018.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

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.

Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. Reachability constrained reinforcement learning. In *International conference on machine learning*, pp. 25636–25655. PMLR, 2022.

Zhengfei Zhang, Kishan Panaganti, Laixi Shi, Yanan Sui, Adam Wierman, and Yisong Yue. Distributionally robust constrained reinforcement learning under strong duality. *arXiv preprint arXiv:2406.15788*, 2024.

Liyuan Zheng and Lillian Ratliff. Constrained upper confidence reinforcement learning. In *Learning for Dynamics and Control*, pp. 620–629. PMLR, 2020.

Ruida Zhou, Tao Liu, Min Cheng, Dileep Kalathil, PR Kumar, and Chao Tian. Natural actor-critic for robust reinforcement learning with function approximation. *Advances in neural information processing systems*, 36:97–133, 2023.

# APPENDIX CONTENTS


**A Theorems and Proofs** **14**

A.1 Proof of Proposition 2 . . . . . 14

A.2 Proof of Theorem 1 . . . . . 14

**B Derivations** **16**

B.1 Derivation of the IPM Robust Peak-Cost Critic . . . . . 16

B.2 IPM Robust Critic for the Augmented CMDP Reformulation . . . . . 17

**C Algorithm, Additional Experiments and Training Details** **19**

C.1 Proposed Algorithm . . . . . 19

C.1.1 Robust Critics . . . . . 19

C.1.2 Branch Selection and Actor Update . . . . . 20

C.2 Additional Experiments . . . . . 20

C.3 Training Details . . . . . 21

C.3.1 Policy Parameterization . . . . . 21

C.3.2 Network Architecture and Rollout Buffer . . . . . 25

C.3.3 Training Hyperparameters . . . . . 25

C.4 Cost Designs . . . . . 26

C.5 Environment Details . . . . . 27

C.6 Perturbation Design for Robustness Training and Evaluation . . . . . 27

**D Proofs of the Convergence Results** **27**

D.1 A unified template and notation . . . . . 27

D.2 Structure of the peak-cost value . . . . . 28

D.3 Proximal point dynamics and gradient domination . . . . . 30

D.4 Proofs of Theorems 2 and 3 . . . . . 31

# A THEOREMS AND PROOFS

## A.1 PROOF OF PROPOSITION 2

*Proof.* Since $-r \in [0,1]$, we have $0 \le J(\pi) \le (1-\gamma)^{-1}$. Feasibility of $\pi^*$ gives $W(\pi^*) \le 0 \le J(\pi^*)/\beta$, so $F_\beta(\pi^*) = J(\pi^*)/\beta$. Hence $\max\{J(\hat{\pi})/\beta,\, W(\hat{\pi})\} = F_\beta(\hat{\pi}) \le J(\pi^*)/\beta + \delta$. The first term yields $J(\hat{\pi}) \le J(\pi^*) + \beta\delta$, and the second yields $W(\hat{\pi}) \le J(\pi^*)/\beta + \delta \le ((1-\gamma)\beta)^{-1} + \delta$. □

## A.2 PROOF OF THEOREM 1

This result is motivated from Mukhopadhyay et al. (2026).

*Formal Statement of Theorem 1*: Consider a discounted MDP with discount factor $\gamma = 0.9$, one nonterminal state $s$, and one absorbing terminal state $\perp$. At state $s$, there are two actions:

- action $a_0$: reward $r(s,a_0) = 4$, cost $c(s,a_0) = 0$, and transition
$$P(\perp \mid s, a_0) = 1;$$
- action $a_1$: reward $r(s,a_1) = 8$, cost $c(s,a_1) = \frac{1}{2}$, and transition
$$P(s \mid s, a_1) = 1.$$

Let $\pi_p$ denote the stationary randomized policy that chooses $a_0$ with probability $p \in [0,1]$ and $a_1$ with probability $1-p$. Define the hard-max modified constraint value

$$V_c^{\pi_p}(s) = \sum_a \pi_p(a \mid s)\Big((1-\gamma)c(s,a) + \gamma \max\{c(s,a), \mathbb{E}[V_c^{\pi_p}(s') \mid s, a]\}\Big), \tag{12}$$

with $V_c^{\pi_p}(\perp) = 0$.

Consider the constrained optimization problem

$$P^\star := \min_{p\in[0,1]} J_r(\pi_p) \qquad \text{s.t.} \qquad V_c^{\pi_p}(s) \le b,$$

with threshold $b = 0.4$, where $J_r(\pi_p)$ is the standard negative discounted reward value starting from $s$. Here, note that peak cost-constraint across the trajectory is the same as here because of the two states with one state being terminal.

Then the associated Lagrangian dual

$$D^\star := \sup_{\nu\ge 0} \inf_{p\in[0,1]} \Big\{ J_r(\pi_p) + \nu\big(V_c^{\pi_p}(s) - b\big)\Big\}$$

satisfies

$$D^\star < P^\star.$$

More precisely,

$$P^\star = -\tfrac{180}{7} \approx -25.7143, \qquad D^\star = -64.8,$$

so that

$$D^\star - P^\star = -\tfrac{1368}{35} \approx -39.0857 < 0.$$

Hence the peak-cost-constraint problem admits a strict duality gap.

*Proof.* Let $J_r(\pi_p)$ denote the discounted negative reward value starting from $s$. By the Bellman equation,

$$J_r(\pi_p) = p\cdot(-4) + (1-p)\big((-8) + \gamma J_r(\pi_p)\big).$$

Since $\gamma = 0.9$, this becomes

$$J_r(\pi_p) = -4p + (1-p)(-8 + 0.9 J_r(\pi_p)).$$

Rearranging gives

$$J_r(\pi_p)\big(1 - 0.9(1-p)\big) = -(8 - 4p),$$

hence

$$J_r(\pi_p) = -\tfrac{8-4p}{0.1+0.9p}.$$

Next, let $V(p) := V_c^{\pi_p}(s)$. By definition of the modified hard-max constraint,

$$V(p) = p\Big((1-\gamma)\cdot 0 + \gamma \max\{0, 0\}\Big) + (1-p)\Big((1-\gamma)\tfrac{1}{2} + \gamma \max\{\tfrac{1}{2}, V(p)\}\Big). \tag{13}$$

The first term is zero, so

$$V(p) = (1-p)\Big(0.1 \cdot \tfrac{1}{2} + 0.9 \max\{\tfrac{1}{2}, V(p)\}\Big).$$

We now solve this fixed-point equation. Suppose first that $V(p) \le \frac{1}{2}$. Then

$$\max\{\tfrac{1}{2}, V(p)\} = \tfrac{1}{2},$$

and thus

$$V(p) = (1-p)\Big(0.05 + 0.9 \cdot \tfrac{1}{2}\Big) = (1-p)\cdot\tfrac{1}{2} = \tfrac{1-p}{2}.$$

Since $\frac{1-p}{2} \le \frac{1}{2}$ for all $p \in [0,1]$, this is self-consistent. Therefore,

$$V_c^{\pi_p}(s) = \tfrac{1-p}{2}, \qquad \forall p \in [0,1].$$

We now solve the primal problem

$$\min_{p\in[0,1]} J_r(\pi_p) \qquad \text{s.t.} \qquad \tfrac{1-p}{2} \le 0.4.$$

The constraint is equivalent to

$$1 - p \le 0.8 \iff p \ge 0.2.$$

Moreover,

$$J_r(\pi_p) = -\tfrac{8-4p}{0.1+0.9p},$$

and differentiating yields

$$\tfrac{d}{dp} J_r(\pi_p) = \tfrac{4(0.1+0.9p)+(8-4p)(0.9)}{(0.1+0.9p)^2} = \tfrac{7.6}{(0.1+0.9p)^2} > 0. \tag{14}$$

Hence $J_r(\pi_p)$ is strictly increasing in $p$, so the primal optimum is attained at the smallest feasible $p$, namely $p^\star = 0.2$. Therefore,

$$P^\star = J_r(\pi_{0.2}) = -\tfrac{8-4(0.2)}{0.1+0.9(0.2)} = -\tfrac{7.2}{0.28} = -\tfrac{180}{7}.$$

Now consider the Lagrangian

$$L(p,\nu) = J_r(\pi_p) + \nu\big(V_c^{\pi_p}(s) - 0.4\big), \qquad \nu \ge 0.$$

Substituting the formulas for $J_r(\pi_p)$ and $V_c^{\pi_p}(s)$ gives

$$L(p,\nu) = -\tfrac{8-4p}{0.1+0.9p} + \nu\left(\tfrac{1-p}{2} - 0.4\right) = -\tfrac{8-4p}{0.1+0.9p} + \nu(0.1 - 0.5p). \tag{15}$$

Fix $\nu \ge 0$. The derivative with respect to $p$ is

$$\tfrac{\partial}{\partial p} L(p,\nu) = \tfrac{7.6}{(0.1+0.9p)^2} - \tfrac{\nu}{2},$$

and the second derivative is

$$\tfrac{\partial^2}{\partial p^2} L(p,\nu) = -\tfrac{13.68}{(0.1+0.9p)^3} < 0.$$

Thus for every fixed $\nu$, the function $p \mapsto L(p, \nu)$ is concave on $[0, 1]$. Therefore its minimum over $[0, 1]$ is attained at one of the endpoints:

$$g(\nu) := \inf_{p \in [0,1]} L(p, \nu) = \min\{L(0, \nu), L(1, \nu)\}.$$

Evaluating the endpoints,

$$L(0, \nu) = -\tfrac{8}{0.1} + 0.1\nu = -80 + 0.1\nu,$$

and

$$L(1, \nu) = -\tfrac{4}{1} - 0.4\nu = -4 - 0.4\nu.$$

Hence

$$g(\nu) = \min\{-80 + 0.1\nu,\ -4 - 0.4\nu\}.$$

The maximum of this minimum of two affine functions occurs when they are equal:

$$-80 + 0.1\nu = -4 - 0.4\nu \iff 76 = 0.5\nu \iff \nu^\star = 152.$$

Substituting back,

$$D^\star = g(\nu^\star) = -80 + 0.1(-152) = -64.8.$$

Finally,

$$\begin{aligned} D^\star - P^\star = -64.8 + \tfrac{180}{7} = -\tfrac{324}{5} + \tfrac{180}{7} &= \tfrac{-2268+900}{35} \\ &= -\tfrac{1368}{35} < 0. \end{aligned}$$

Thus the duality gap is strict. □

# B Derivations

## B.1 Derivation of the IPM Robust Peak-Cost Critic

We derive the robust Bellman update used for the peak-cost critic under an integral probability metric (IPM) uncertainty set. For two transition distributions $p, q \in \Delta(\mathcal{S})$, define the IPM distance

$$d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} \left\{ p^\top f - q^\top f \right\},$$

where the function class is chosen as

$$\mathcal{F} = \left\{ f = \Psi\epsilon : \epsilon \in \mathbb{R}^d,\ \|\epsilon\| \le 1 \right\}.$$

Here $\Psi \in \mathbb{R}^{|\mathcal{S}| \times d}$ is the feature matrix. We approximate the peak-cost critic by a linear function

$$V^\pi_{c,\mathrm{peak},w}(s) = \psi(s)^\top w, \qquad \text{i.e.} \qquad V^\pi_{c,\mathrm{peak},w} = \Psi w.$$

For a nominal transition distribution $p^0_{s,a} = P_0(\cdot \mid s, a)$, the IPM uncertainty set is

$$\mathcal{P}_{s,a} = \left\{ q \in \Delta(\mathcal{S}) : d_{\mathcal{F}}(q, p^0_{s,a}) \le \rho \right\}.$$

Under transition uncertainty, the robust peak-cost Bellman operator is given by

$$\left(\mathcal{L}_{\mathcal{P}} V^\pi_{c,\mathrm{peak},w}\right)(s, a) = (1 - \gamma)c(s, a) + \gamma \max \left\{ c(s, a),\ \sup_{q \in \mathcal{P}_{s,a}} q^\top V^\pi_{c,\mathrm{peak},w} \right\}. \tag{16}$$

The above robust peak-cost Bellman update contains the term

$$\sup_{q \in \mathcal{P}_{s,a}} q^\top V^\pi_{c,\mathrm{peak},w}.$$

We now simplify this worst-case expectation. Let

$$u = q - p^0_{s,a}, \qquad q = u + p^0_{s,a}.$$

Then the objective becomes

$$q^\top V^\pi_{c,\mathrm{peak},w} = (u + p^0_{s,a})^\top \Psi w = u^\top \Psi w + (p^0_{s,a})^\top \Psi w.$$

The IPM constraint becomes

$$d_{\mathcal{F}}(q, p^0_{s,a}) = \sup_{\|\epsilon\|\le 1} (q - p^0_{s,a})^\top \Psi \epsilon = \sup_{\|\epsilon\|\le 1} u^\top \Psi \epsilon \le \rho.$$

Using the dual norm identity, this is equivalent to

$$\|\Psi^\top u\|_* \le \rho.$$

Because both $q$ and $p^0_{s,a}$ are probability distributions, we also have

$$\mathbf{1}^\top u = 0.$$

Assume that the first feature is the constant feature, i.e., $\psi_1(s) = 1$ for all $s$. Let

$$y = \Psi^\top u.$$

Then $\mathbf{1}^\top u = 0$ implies $y_1 = 0$. Therefore, the robust expectation can be written as

$$\sup_{q\in\mathcal{P}_{s,a}} q^\top V^\pi_{c,\mathrm{peak},w} \le (p^0_{s,a})^\top \Psi w + \sup_{\substack{\|y\|_*\le\rho \\ y_1=0}} y^\top w.$$

The inequality is due to the fact that we must have $q \ge 0$. Since $y_1 = 0$, only the non-constant coordinates of $w$ contribute:

$$y^\top w = y^\top_{2:d} w_{2:d}.$$

Thus,

$$\sup_{\substack{\|y\|_*\le\rho \\ y_1=0}} y^\top w = \sup_{\|y_{2:d}\|_*\le\rho} y^\top_{2:d} w_{2:d} = \rho\|w_{2:d}\|,$$

where the last equality again follows from duality of norms. Hence,

$$\sup_{q\in\mathcal{P}_{s,a}} q^\top V^\pi_{c,\mathrm{peak},w} \le (p^0_{s,a})^\top V^\pi_{c,\mathrm{peak},w} + \rho\|w_{2:d}\|.$$

Equivalently,

$$\sup_{q\in\mathcal{P}_{s,a}} \mathbb{E}_{s'\sim q}\left[V^\pi_{c,\mathrm{peak},w}(s')\right] = \mathbb{E}_{s'\sim P_0(\cdot|s,a)}\left[V^\pi_{c,\mathrm{peak},w}(s')\right] + \rho\|w_{2:d}\|.$$

Substituting this expression into the peak-cost Bellman recursion gives the IPM robust peak-cost Bellman operator

$$\left(\mathcal{L}_{\mathcal{P}} V^\pi_{c,\mathrm{peak},w}\right)(s,a) \le (1-\gamma)c(s,a) + \gamma \max\left\{c(s,a), \mathbb{E}_{s'\sim P_0(\cdot|s,a)}\left[V^\pi_{c,\mathrm{peak},w}(s')\right] + \rho\|w_{2:d}\|\right\}.$$

If the Euclidean norm is used in the IPM feature ball, then the dual norm is also Euclidean and the penalty becomes $\rho\|w_{2:d}\|_2$. Therefore, the final operator used in our implementation is

$$\left(\mathcal{L}_{\mathcal{P}} V^\pi_{c,\mathrm{peak},w}\right)(s,a) \le (1-\gamma)c(s,a) + \gamma \max\left\{c(s,a), \mathbb{E}_{s'\sim P_0(\cdot|s,a)}\left[V^\pi_{c,\mathrm{peak},w}(s')\right] + \rho\|w_{2:d}\|_2\right\}.$$

The first coordinate is excluded from the robustness penalty because it corresponds to the constant feature. Since $q$ and $p^0_{s,a}$ both have total mass one, perturbations $u = q - p^0_{s,a}$ satisfy $\mathbf{1}^\top u = 0$, and therefore the constant feature is unaffected by the IPM perturbation.

### B.2 IPM Robust Critic for the Augmented CMDP Reformulation

We now derive the robust critic used for the augmented CMDP baseline. The essential difference from Appendix B.1 is that the inner maximum of the peak recursion disappears: in the augmented state space the constraint is additive, so the operator is the standard one, and the only question is how the memory coordinate affects the robustness penalty. We show that it does so only through the dependence of the cost on the next state, and not at all when the cost is a function of $(s, a)$ alone.

**Augmented dynamics and additive cost.** Let $\bar{s}_h = (s_h, m_h)$ with $m_0 = 0$ and, writing $c_h$ for the cost incurred at step $h$,

$$m_{h+1} = \max\{m_h,\, c_h\}, \qquad \Delta_h = m_{h+1} - m_h = (c_h - m_h)_+. \tag{17}$$

By Proposition 1, $\sum_{h<H} \Delta_h = \max_{h<H} c_h$, so the peak cost is recovered as an *undiscounted, finite-horizon* sum of $\Delta_h$. We therefore write the augmented critic with a step index $h$ and no discount factor.

**The induced augmented uncertainty set.** Uncertainty concerns the physical dynamics only: the adversary selects $q \in \mathcal{P}_{s,a}$ for the next physical state $s'$, and the memory coordinate then follows deterministically through

$$\phi_{s,a,m}(s') \;=\; \max\{m,\, c(s,a,s')\}, \tag{18}$$

which reduces to the constant $\phi \equiv \max\{m, c(s,a)\}$ when the cost does not depend on the next state. The augmented uncertainty set is therefore not a free set of distributions over $\bar{s}'$, but the pushforward

$$\bar{\mathcal{P}}_{(s,m),a} = \Big\{\, q \circ \big(\mathrm{id}, \phi_{s,a,m}\big)^{-1} \;:\; q \in \mathcal{P}_{s,a} \Big\}. \tag{19}$$

Consequently, for any augmented critic $\bar{V}_{h+1}$,

$$\sup_{\bar{q} \in \bar{\mathcal{P}}_{(s,m),a}} \mathbb{E}_{\bar{s}' \sim \bar{q}}\big[\bar{V}_{h+1}(\bar{s}')\big] \;=\; \sup_{q \in \mathcal{P}_{s,a}} \mathbb{E}_{s' \sim q}\big[g(s')\big], \qquad g(s') := \bar{V}_{h+1}\big(s', \phi_{s,a,m}(s')\big). \tag{20}$$

Equation equation 20 is the key step: the worst case is taken over a single function $g$ of the *physical* next state, so the same IPM machinery as in Appendix B.1 applies, with $g$ in place of the critic. In particular, the adversary cannot perturb $s'$ and $m'$ independently.

**Linear critic, state–action cost.** Suppose $c = c(s,a)$, so that $m' = \max\{m, c(s,a)\}$ is determined before $s'$ is drawn, and let the critic be linear in the physical features for each value of the memory,

$$\bar{V}_h(s,m) = \psi(s)^\top w_h(m), \tag{21}$$

with $\psi_1 \equiv 1$ as before. Then $g = \Psi w_{h+1}(m')$ with $m'$ fixed, and the derivation of Appendix B.1 applies verbatim:

$$\sup_{q \in \mathcal{P}_{s,a}} \mathbb{E}_{s' \sim q}\big[g(s')\big] \le \mathbb{E}_{s' \sim P_0(\cdot|s,a)}\big[\bar{V}_{h+1}(s', m')\big] + \rho \big\| w_{h+1,2:d}(m') \big\|. \tag{22}$$

**General critic, next-state-dependent cost.** When $c = c(s,a,s')$, the map $\phi_{s,a,m}$ is non-constant and the incremental cost is itself random, so the relevant function is $g(s') = \Delta(\bar{s}, a, s') + \bar{V}_{h+1}(s', \phi_{s,a,m}(s'))$. Take $\mathcal{F}$ to be the unit ball of 1-Lipschitz functions, so that $d_{\mathcal{F}}$ is the Kantorovich–Rubinstein distance and

$$\sup_{q \in \mathcal{P}_{s,a}} \mathbb{E}_q[g] - \mathbb{E}_{P_0}[g] \le \rho\, \mathrm{Lip}(g). \tag{23}$$

Write

$$L_s = \sup_{s,m} \big\| \nabla_s \bar{V}_h(s,m) \big\|_2, \qquad L_m = \sup_{s,m} \big| \partial_m \bar{V}_h(s,m) \big|, \qquad L_c = \sup_{s,a} \mathrm{Lip}\big(c(s,a,\cdot)\big), \tag{24}$$

where $\partial_m$ is a scalar derivative (or any Lipschitz modulus in $m$, since $\bar{V}$ need not be differentiable there).

**Lemma 1** (Augmented sensitivity). *For all $s'_1, s'_2 \in \mathcal{S}$, $|g(s'_1) - g(s'_2)| \le \big(L_s + (1 + L_m)L_c\big)\|s'_1 - s'_2\|_2$, and hence $\mathrm{Lip}(g) \le L_s + (1 + L_m)L_c$. If $c = c(s,a)$ then $L_c = 0$ and $\mathrm{Lip}(g) \le L_s$.*

*Proof.* Since $x \mapsto \max\{m, x\}$ and $x \mapsto (x - m)_+$ are 1-Lipschitz, $|\phi_{s,a,m}(s'_1) - \phi_{s,a,m}(s'_2)| \le |c(s,a,s'_1) - c(s,a,s'_2)| \le L_c \|s'_1 - s'_2\|_2$, and the same bound holds for $\Delta$. Therefore

$$|g(s'_1) - g(s'_2)| \le \underbrace{L_c \|s'_1 - s'_2\|_2}_{\Delta} + \underbrace{L_s \|s'_1 - s'_2\|_2 + L_m\, |\phi(s'_1) - \phi(s'_2)|}_{\bar{V}_{h+1}},$$

and substituting the bound on $|\phi(s'_1) - \phi(s'_2)|$ gives the claim. □

**The robust augmented operator.** Combining equation 20 with equation 22 or Lemma 1,

$$\big(\mathcal{L}_{\mathcal{P}}\bar{Q}_h\big)(\bar{s},a) \le \Delta(\bar{s},a) + \mathbb{E}_{s'\sim P_0(\cdot|s,a)}\big[\bar{V}_{h+1}(s',m')\big] + \rho\,\Lambda_h(s,m,a), \tag{25}$$

with $\bar{V}_h(\bar{s}) = \langle \pi(\cdot \mid \bar{s}), \bar{Q}_h(\bar{s},\cdot)\rangle$, $\bar{V}_H \equiv 0$, and

$$\Lambda_h(s,m,a) = \begin{cases} \big\|w_{h+1,2:d}(m')\big\|, & \text{linear critic, } c = c(s,a), \\ L_s + (1+L_m)L_c, & \text{general critic, } c = c(s,a,s'). \end{cases} \tag{26}$$

In the first case the penalty is exact, by the same argument as in Appendix B.1; in the second it is an upper bound, so the operator is conservative.

**Remark 5** (Lifted rectangularity)**.** *The penalty in equation 25 is applied separately at each $(s,m,a)$, which grants the adversary an $(s,m,a)$-rectangular set: it may select a different physical kernel for the same $(s,a)$ depending on the value of the running maximum. The physical uncertainty is only $(s,a)$-rectangular, and the lifted set strictly contains the set it induces. The robust evaluation is therefore conservative even before any approximation, which is the quantitative form of Issue 3 in Section 3.*

**Remark 6** (Uniform versus state-dependent pessimism)**.** *When $\Lambda_h$ does not depend on $(s,m,a)$ – as in the second case of equation 26, where $L_s, L_m, L_c$ are global suprema – the penalty is a constant and the fixed point of equation 25 is the nominal value shifted by $\rho\Lambda(H-h)$. Robustness then acts only as a tightened constraint threshold rather than as state-dependent pessimism. The linear-critic form is preferable in this respect, since $\|w_{h+1,2:d}(m')\|$ varies with the running maximum. By contrast, in the direct peak-cost operator, the penalty sits inside the maximum, where it changes which branch is active and so is never a uniform shift.*

**Practical estimation.** For the linear critic, $\|w_{h+1,2:d}(m')\|$ is available in closed form. For a differentiable critic $\bar{V}_\theta$, the constants in equation 24 are suprema and are estimated from a batch $\{(s_i,m_i)\}_{i=1}^N$ by the empirical maxima

$$\widehat{L}_s = \max_{i\le N}\big\|\nabla_s \bar{V}_\theta(s_i,m_i)\big\|_2, \qquad \widehat{L}_m = \max_{i\le N}\big|\partial_m \bar{V}_\theta(s_i,m_i)\big|, \tag{27}$$

and $\widehat{\Lambda} = \widehat{L}_s + (1+\widehat{L}_m)L_c$, with $L_c$ known from the cost definition. A batch mean in place of the maximum would systematically under-estimate the supremum and hence under-robustify the critic; where a maximum over the batch is too noisy we use a high empirical quantile instead.

# C Algorithm, Additional Experiments and Training Details

## C.1 Proposed Algorithm

We now present our approach for solving the robust peak-constrained MDP. The algorithm is an actor–critic scheme with two critics – one for the reward and one for the peak cost – both robustified through the IPM penalty of equation 7, and an actor that follows whichever branch of the surrogate equation 8 is currently active.

### C.1.1 Robust Critics

Both critics are linear in the features produced by the penultimate network layer, $V(s) = \psi(s)^\top w$ with $\psi_1 \equiv 1$ the bias unit, so that the closed form of Appendix B.1 applies with $w$ the final-layer weights and $w_{2:d}$ those weights excluding the bias. Applying equation 7 to each critic gives the temporal-difference errors

$$\delta_{t,r} = -r_t + \gamma\Big(V_r(s_{t+1}) + \rho_r\|w_{r,2:d}\|_2\Big) - V_r(s_t), \tag{28}$$

$$\delta_{t,c} = (1-\gamma)c_t + \gamma\max\Big\{c_t,\ V_{c,\mathrm{peak}}(s_{t+1}) + \rho_c\|w_{c,2:d}\|_2\Big\} - V_{c,\mathrm{peak}}(s_t). \tag{29}$$

Here $V_r$ is the negative-reward value, hence the term $-r_t$. In equation 29 the robustness penalty appears *inside* the maximum, as in the operator equation 7: placed outside it would act as a uniform shift of the critic and could not change which branch of the peak recursion is active.

Because the reward objective is additive, we use generalized advantage estimation; the peak recursion is non-additive, so multi-step returns do not telescope and we use the one-step estimate:

$$A_{t,r} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l,r}, \qquad A_{t,c} = \delta_{t,c}, \tag{30}$$

with $\lambda$ the GAE decay, and value targets $v_{\text{target},j}(s_t) = A_{t,j} + V_j(s_t)$ for $j \in \{r, c\}$. The critics are fit by regression, $L_{V_j} = \big(V_j(s_t) - v_{\text{target},j}(s_t)\big)^2$.

### C.1.2 Branch Selection and Actor Update

At each iteration we evaluate the two terms of the surrogate equation 8 on the sampled batch $\mathcal{D}_t$,

$$A \;=\; \frac{1}{\beta}\,\widehat{\mathbb{E}}_{s\sim\mathcal{S}_0}\big[V_r(s)\big], \qquad B \;=\; \max_{s\in\mathcal{D}_t} V_{c,\text{peak}}(s) \;-\; b, \tag{31}$$

and update the actor on the reward advantage $A_{t,r}$ if $A > B$ and on the cost advantage $A_{t,c}$ otherwise, so that the update always follows the active term of $\max\{J/\beta,\, W - b\}$. No multiplier is maintained. The selected advantage is used in the PPO-style clipped objective Schulman et al. (2017),

$$L_{\text{actor}} = \mathbb{E}_t\Big[\min\big(\varrho_t(\theta)A_{t,j},\ \text{clip}\big(\varrho_t(\theta), 1-\epsilon, 1+\epsilon\big)A_{t,j}\big)\Big]\,, \tag{32}$$

where $\varrho_t(\theta)$ is the policy ratio and $j$ is the selected branch; since both $V_r$ and $V_{c,\text{peak}}$ are minimized, the advantages enter with the sign that makes equation 32 a descent step on the active term. Algorithm 1 summarizes the procedure.

**Algorithm 1:** RP-CRL

**Input:** Initial policy parameters $\theta_0$, initial value function parameters $\phi_{r,0}$ and $\phi_{c,0}$
**Parameters:** Discount factor $\gamma$, Peak-Cost budget $b$, GAE decay $\lambda$, Lagrangian multiplier $\beta$, PPO clipping parameter $\epsilon$, learning rates $\alpha_{\text{actor}}$, $\alpha_{\text{critic,r}}$, $\alpha_{\text{critic,c}}$ Reward/Cost Selector $\psi(.)$
**Initialize:** Policy $\pi_\theta$ with parameters $\theta_0$, robust value functions $V_r$ and $V_{c,\text{peak}}$ with parameters $\phi_{r,0}$ and $\phi_{c,0}$;
**for** $t = 0, 1, \ldots, T-1$ **do**
  Collect a set of trajectories $\mathcal{D}_t = \{\tau_i\}_{i=1}^N$ by running policy $\pi_\theta$ in the environment;
  $A = \frac{1}{|D_t|}\sum_{s\in D_t} \frac{V_r(s)}{\beta}$;
  $B = \max_{s\in S_0}\left(V_{c,\text{peak}}(s) - b\right)$;
  $F = \arg\max\{A, B\}$;
  Estimate Advantage & Update Critic: $\psi(F)$;
  Update Policy using Eqn 32 and update $\theta_t$ using SGD with learning rate $\alpha_{\text{actor}}$;
**Output:** Optimized policy $\pi_\theta$

## C.2 Additional Experiments

**Multi-Constraint Results**

Figs. 7 and 8 report the multi-constraint setting. The quadrotor imposes four constraints, one for the lower and one for the upper thrust threshold of each rotor; the HalfCheetah variant imposes two, on pitch and joint speed (Design details in Appendix C.4). During training, FAC attains the highest return but violates the thresholds, while the remaining methods are safe. At inference, all methods exhibit persistent safety. RP-CRL gives the best return among methods that keep every peak cost below its threshold, indicating that the surrogate handles several simultaneous peak constraints without a multiplier per constraint.

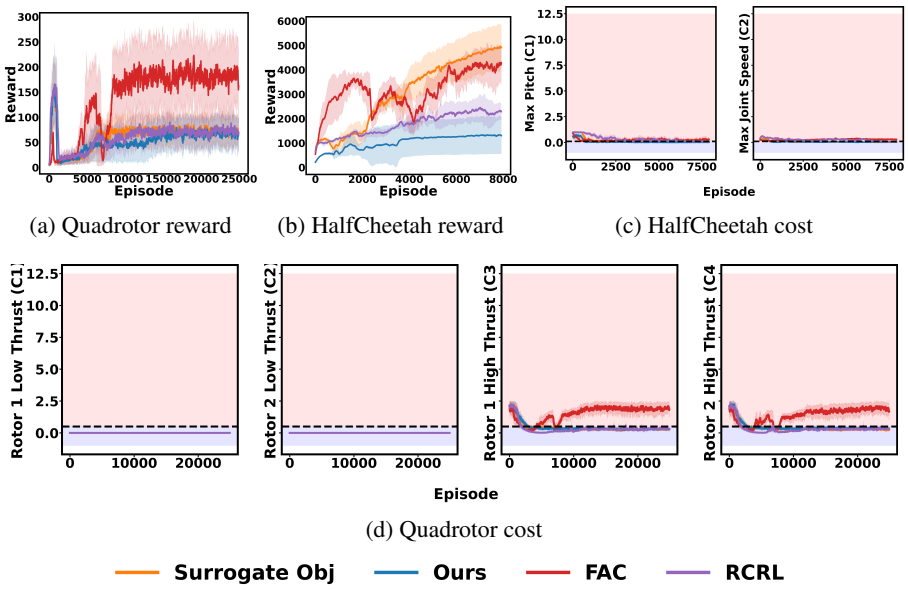
**Sparsity of Cost Signals in Augmented CMDP Reformulation as per Proposition 1**

Fig. 9 illustrates the sparsity of the cost signal in the HalfCheetah environment. The episode length is typically close to 1000 steps, but only about 0–25 of these steps contain nonzero cost. Consequently, the fraction of timesteps with zero cost is close to one for most episodes. This shows that the agent

**Algorithm 2:** Reward/Cost Selector $\psi$

**Input:** Choice function F, Discount factor $\gamma$, GAE decay $\lambda$,PPO clipping parameter $\epsilon$, learning rates $\alpha_{\text{actor}}$, $\alpha_{\text{critic,r}}$, $\alpha_{\text{critic,c}}$

**if** $F = 0$ **then**
  **Update Reward Critic:**;
  Compute TD errors $\delta_{t,r}$ using Eqn 28;
  Estimate advantages $A_{t,r}$ using Eqn 30;
**else if** $F = 1$ **then**
  **Update Cost Critic:**;
  Compute TD errors $\delta_{t,c}$ using Eqn 29;
  Estimate advantages $A_{t,c}$ using Eqn 30;
Estimate targets $v_{target,r}$ and reward critic loss $L_{V_r}$;
Update critic parameters $\phi_r$ using SGD with learning rate $\alpha_{\text{critic,r}}$;
Estimate targets $v_{target,c}$ and cost critic loss $L_{V_{c,\text{peak}}}$ ;
Update critic parameters $\phi_c$ using SGD with learning rate $\alpha_{\text{critic,c}}$;




(a) Quadrotor reward (b) HalfCheetah reward (c) HalfCheetah cost

(d) Quadrotor cost

Figure 7: Multi-constraint training results. FAC has highest performance but behaves unsafely while training. Rest methods remain persistently safe while learning their task performances.

receives very sparse cost feedback during training. Due to this sparsity, learning an accurate cost critic becomes extremely slow, and the reward learning can collapse before the safety signal is learned reliably.

### C.3 Training Details

This section summarizes the policy parameterization, neural network architecture, rollout buffer, and training hyperparameters used in our experiments. Unless otherwise stated, all continuous-control experiments use a Squashed Gaussian stochastic policy trained with a PPO-style clipped policy update. The same actor–critic backbone is used across environments, with environment-specific hyperparameters listed below.

#### C.3.1 Policy Parameterization

For continuous-action environments, we use a tanh-squashed Gaussian policy. This choice is motivated by the fact that the control inputs in our environments are bounded, whereas a standard Gaussian policy has unbounded support. The squashed Gaussian policy samples an unconstrained latent action and then maps it smoothly to the valid action range using a tanh transformation.

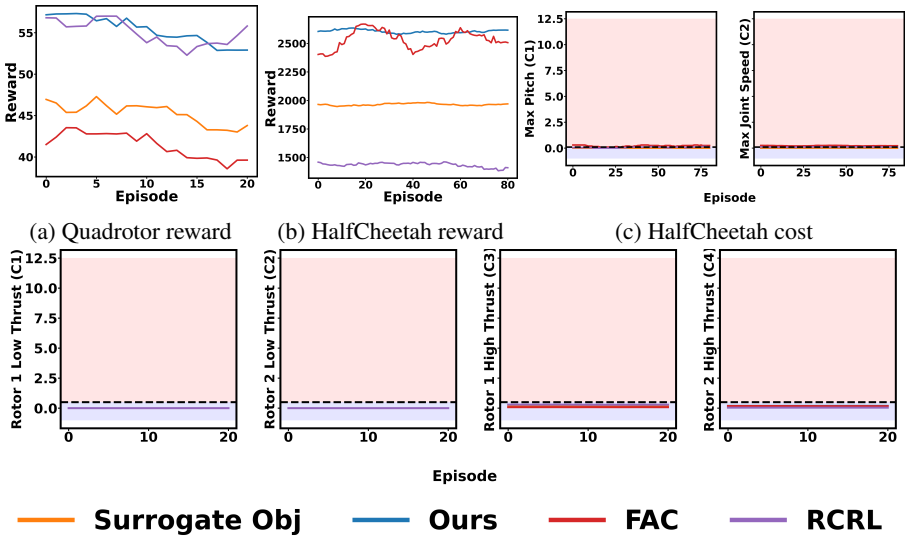



Figure 8: Multi-constraint inference results.While all methods are persistently safe, RP-CRL performs best in terms of returns.

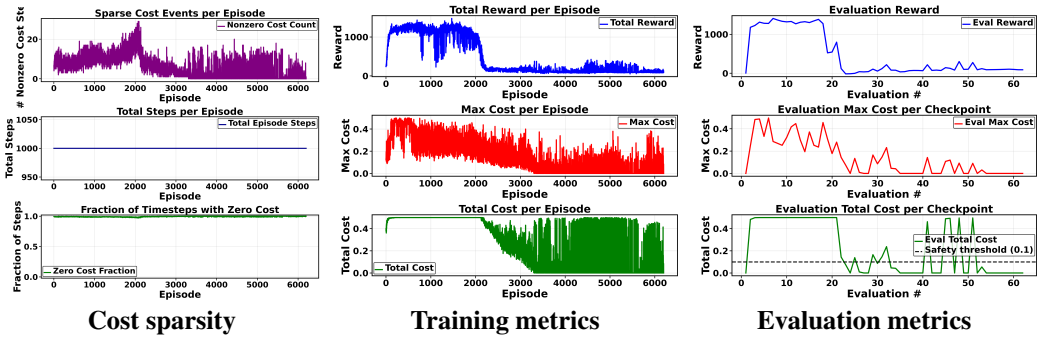



Figure 9: Sparse cost signal in HalfCheetah. Although episodes contain nearly 1000 timesteps, only a small number of timesteps, typically around 0–25, produce nonzero cost. Hence, the zero-cost fraction remains close to one for most episodes. This sparse safety signal slows cost learning substantially and can cause reward learning to collapse before the cost critic becomes informative.

**Comparison with a standard Gaussian policy.** A standard Gaussian policy directly samples the environment action from a Gaussian distribution. Given a state s, the policy outputs a mean action $\mu_\theta(s)$, and the action is sampled as $a \sim \mathcal{N}(\mu_\theta(s), \sigma_\theta^2)$. However, in continuous-control environments, the action must satisfy $a \in [-a_{\max}, a_{\max}]$. Since a Gaussian distribution has support over the whole real line, it can produce invalid actions outside this interval. A common workaround is to clip the sampled action: $\tilde{a} = \mathrm{clip}(a, -a_{\max}, a_{\max})$. This clipping operation changes the effective action distribution. In particular, many different Gaussian samples outside the valid action range are mapped to the same boundary action. For example, all samples larger than $a_{\max}$ are mapped to $a_{\max}$. Therefore, the clipped action no longer follows the Gaussian density that is used to compute the policy log-probability.

This creates a mismatch in PPO-style updates. PPO uses the likelihood ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$. If the environment receives a clipped action but the likelihood ratio is computed using the unclipped Gaussian density, then the update no longer corresponds to the true action distribution. This issue is especially important near the action boundaries, where many Gaussian samples collapse to the same clipped action.

To avoid this problem, we use a squashed Gaussian policy. Instead of clipping the sampled action, we sample a latent action from an unconstrained Gaussian and pass it through a smooth bounded function.

**Squashed Gaussian policy.** Given a state s, the actor network outputs the mean of a latent Gaussian distribution: $\mu_\theta(s) = \mathrm{MLP}_\theta(s)$. A latent pre-tanh action u is sampled from $u \sim p_\theta(u|s) = \mathcal{N}(\mu_\theta(s), \sigma_\theta^2)$. The actual environment action is then obtained by applying a tanh transformation: $a = f(u) = a_{\max}\tanh(u)$. This guarantees that the final action is always bounded: $a \in [-a_{\max}, a_{\max}]$.

Thus, the policy separates the sampled latent action u from the actual environment action a. The latent action u follows a Gaussian distribution and is used to compute log-probabilities, while the transformed action a is applied to the environment.

The log-standard deviation is represented as a state-independent trainable parameter.

**Change-of-variables correction.** Because the action applied to the environment is a nonlinear transformation of the latent Gaussian sample, the policy density over the environment action must include a change-of-variables correction.

For a one-dimensional action, the transformation is $a = a_{\max}\tanh(u)$. The derivative of this transformation is $\frac{da}{du} = a_{\max}\left(1 - \tanh^2(u)\right)$. The probability mass must be preserved under the transformation. For a small interval, this gives

$$p_\theta(u|s)\,du = \pi_\theta(a|s)\,da.$$

Therefore,

$$\pi_\theta(a|s) = p_\theta(u|s)\left|\frac{du}{da}\right|.$$

Since

$$\left|\frac{du}{da}\right| = \frac{1}{a_{\max}\left(1 - \tanh^2(u)\right)},$$

the transformed action density is

$$\pi_\theta(a|s) = \frac{p_\theta(u|s)}{a_{\max}\left(1 - \tanh^2(u)\right)}.$$

Taking the logarithm gives

$$\log \pi_\theta(a|s) = \log p_\theta(u|s) - \log\left[a_{\max}\left(1 - \tanh^2(u)\right)\right].$$

Equivalently,

$$\log \pi_\theta(a|s) = \log p_\theta(u|s) - \log a_{\max} - \log\left(1 - \tanh^2(u)\right).$$

The latent Gaussian density is

$$p_\theta(u|s) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}}\exp\left(-\frac{1}{2}\frac{(u - \mu_\theta(s))^2}{\sigma_\theta^2}\right).$$

Thus, its log-density is

$$\log p_\theta(u|s) = -\frac{1}{2}\log(2\pi) - \log\sigma_\theta - \frac{1}{2}\frac{(u - \mu_\theta(s))^2}{\sigma_\theta^2}.$$

Substituting this into the transformed density gives

$$\log \pi_\theta(a|s) = -\frac{1}{2}\log(2\pi) - \log\sigma_\theta - \frac{1}{2}\frac{(u - \mu_\theta(s))^2}{\sigma_\theta^2} - \log a_{\max} - \log\left(1 - \tanh^2(u)\right).$$

**Numerically stable form of the correction.** Directly computing the term $log(1 - tanh^2(u))$ can be unstable when the magnitude of u is large, because tanh(u) becomes very close to either 1 or -1. Therefore, we use an equivalent stable identity.

First, note that

$$1 - \tanh^2(u) = \mathrm{sech}^2(u) = \frac{1}{\cosh^2(u)}.$$

Taking logarithms,

$$\log\left(1-\tanh^2(u)\right) = -2\log(\cosh(u)).$$

Using

$$\cosh(u) = \frac{e^u + e^{-u}}{2},$$

we have

$$\log(\cosh(u)) = \log(e^u + e^{-u}) - \log 2.$$

Factoring out $e^u$,

$$\log(\cosh(u)) = \log\left(e^u\left(1+e^{-2u}\right)\right) - \log 2.$$

Therefore,

$$\log(\cosh(u)) = u + \log\left(1+e^{-2u}\right) - \log 2.$$

Using the softplus function,

$$\mathrm{softplus}(x) = \log(1+e^x),$$

we obtain

$$\log(\cosh(u)) = u + \mathrm{softplus}(-2u) - \log 2.$$

Hence,

$$\log\left(1-\tanh^2(u)\right) = -2\left[u + \mathrm{softplus}(-2u) - \log 2\right].$$

Equivalently,

$$\log\left(1-\tanh^2(u)\right) = 2\left[\log 2 - u - \mathrm{softplus}(-2u)\right].$$

This stable expression is used in our implementation.

**Multi-dimensional actions.** For a d-dimensional action, the latent action is $u = [u_1, u_2, \ldots, u_d]$. Each dimension is transformed independently:

$$a_i = a_{\max}\tanh(u_i), \qquad i = 1, \ldots, d.$$

The Jacobian matrix is diagonal because each action dimension depends only on its corresponding latent variable. Therefore,

$$\left|\det\frac{\partial a}{\partial u}\right| = \prod_{i=1}^{d} a_{\max}\left(1-\tanh^2(u_i)\right).$$

The transformed policy density is

$$\pi_\theta(a|s) = p_\theta(u|s)\left|\det\frac{\partial u}{\partial a}\right|.$$

Taking logarithms gives

$$\log\pi_\theta(a|s) = \log p_\theta(u|s) - \sum_{i=1}^{d}\log\left[a_{\max}\left(1-\tanh^2(u_i)\right)\right].$$

Equivalently,

$$\log\pi_\theta(a|s) = \sum_{i=1}^{d}\log\mathcal{N}\left(u_i;\mu_{\theta,i}(s),\sigma^2_{\theta,i}\right) - \sum_{i=1}^{d}\left[\log a_{\max} + \log\left(1-\tanh^2(u_i)\right)\right].$$

**Sampling and PPO update.** During rollout collection, we first sample the latent action: $u_t \sim \mathcal{N}(\mu_\theta(s_t), \sigma^2_\theta)$. Then we transform it into the bounded environment action: $a_t = a_{\max}\tanh(u_t)$. The environment receives $a_t$, while the PPO update uses the corrected log-probability of the squashed action. This ensures that the likelihood ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is computed using the actual transformed action distribution rather than an incorrectly clipped Gaussian distribution.

**Evaluation policy.** At evaluation time, we use the deterministic latent mean: $u_{\text{eval}} = \mu_\theta(s)$. The final deterministic action is $a_{\text{eval}} = a_{\max}\tanh(\mu_\theta(s))$.

### C.3.2 Network Architecture and Rollout Buffer

All actor and critic networks use fully connected multilayer perceptrons with three hidden layers. The hidden width is $64$ in all experiments unless otherwise specified.

**Actor and Critic networks.** The Squashed Gaussian actor consists of three hidden layers 64 neurons each. We use tanh activation. Orthogonal initialization is applied to all actor layers, with a smaller gain of $0.01$ for the final mean layer. For multi-constraint experiments, the peak-cost critic predicts cost information for each constraint dimension.

**Rollout buffer.** The rollout buffer stores on-policy trajectories. The default rollout batch size is $2048$ for MuJoCo-style tasks, while the Quadrotor experiments use a batch size of $1024$.

### C.3.3 Training Hyperparameters

**Common PPO and optimization settings.** The following settings are used throughout unless modified in the environment-specific tables:

| | |
|---|---|
| Policy distribution | Squashed Gaussian |
| Hidden width | $64$ |
| Discount factor $\gamma$ | $0.99$ |
| GAE parameter $\lambda$ | $0.95$ |
| PPO clip parameter $\epsilon$ | $0.2$ |
| PPO epochs $K$ | $5$ |
| Evaluation frequency | $100$ |
| Save frequency | $20$ |
| Advantage normalization | Enabled |
| Learning-rate decay | Enabled |
| Gradient clipping | Enabled |
| Orthogonal initialization | Enabled |
| Tanh activation | Enabled |
| Adaptive entropy coefficient | Disabled |

**Experiment specific parameters:**

**Single-constraint experiments.** For the single-constraint MuJoCo experiments, we used four environments: HalfCheetah, Swimmer, Ant, and Humanoid. **HalfCheetah** was trained for 16000 episodes with a warm-start period of 500 episodes, a safety baseline of 0.1, 5 policy update epochs, penalty coefficient 30000, and entropy coefficient 0.001. The nominal surrogate-objective run used HalfCheetah Cost without robustness regularization, while the robust version used HalfCheetah Cost Perturbed with robustness regularization 0.001 and gravity perturbation scale 2.0. **Swimmer** was trained for 16000 episodes with a warm-start period of 150 episodes, a safety baseline of 0.1, penalty coefficient 30000, 5 policy update epochs, and entropy coefficient 0.001. The robust version used Swimmer Cost with viscosity perturbation scale 0.7, while nominal runs used no viscosity perturbation. **Ant** was trained for 6000 episodes with a warm-start period of 700 episodes, a safety baseline of 0.05, penalty coefficient 30000, actor learning rate 1e-3, reward-critic learning rate 1e-3, cost-critic learning rate 1e-3, and entropy coefficient 0.001. The robust Ant run used Ant Cost Perturbed with gravity perturbation scale 0.7 and robustness regularization 0.001. While inference, a perturbation scale of 2.0 is used. **Humanoid** was trained for 6000 episodes with a warm-start period of 300 episodes, a safety baseline of 0.1, penalty coefficient 300000, actor learning rate 1e-3, reward-critic learning rate 5e-4, cost-critic learning rate 5e-4, and entropy coefficient 0.01. The robust Humanoid run used Humanoid Cost Perturbed with gravity perturbation scale 0.7 and robustness regularization 0.001.Inference is done on a perturbation of $\sigma = 3.0$.

**Multi-constraint experiments.** For the multi-constraint experiments, we used HalfCheetah and Quadrotor settings. The **HalfCheetah multi-constraint** environment used HalfCheetah MultiConstraint Cost and HalfCheetah MultiConstraint Cost Perturbed with the multi-constraint pitch-and-joint-speed cost version. This setting included two constraints, corresponding to pitch and joint-speed limits. The surrogate-objective and robust runs were trained for 10000 episodes, evaluated

Table 1: Single-constraint cost designs. Here, $a_t$ is the action, $\kappa = 0.5$ is the torque threshold, $d_{\text{wall}}(s_t)$ is the distance to the closest wall, and $d_{\text{safe}}$ is the safety-clearance distance.

| Environment | Cost |
|---|---|
| Ant | $c(s_t, a_t) = (\|a_t\|_\infty - 0.5)_+$ |
| HalfCheetah | $c(s_t, a_t) = (\|a_t\|_\infty - 0.5)_+$ |
| Humanoid | $c(s_t, a_t) = \|a_t\|_2^2$ |
| Swimmer | $c(s_t, a_t) = (\|a_t\|_\infty - 0.5)_+$ |
| Safety Gym Circle | $c(s_t) = 0, \quad d_{\text{wall}}(s_t) > d_{\text{safe}}$<br>$c(s_t) = \frac{d_{\text{safe}} - d_{\text{wall}}(s_t)}{d_{\text{safe}}}, \quad 0 < d_{\text{wall}}(s_t) \leq d_{\text{safe}}$<br>$c(s_t) = 1 + 10\,(-d_{\text{wall}}(s_t)), \quad d_{\text{wall}}(s_t) \leq 0$ |

Table 2: Multi-constraint cost designs. For Quadrotor, $u_{t,i}$ is the thrust of rotor $i$. For HalfCheetah, $\theta_t$ is the torso pitch and $\dot{q}_{t,j}$ is the velocity of joint $j$.

| Environment | Constraint costs |
|---|---|
| Quadrotor | $c_i(s_t, a_t) = (u_{t,i} - 0.12)_+ + (0 - u_{t,i})_+$<br>$i = 1, \ldots, 4$ |
| HalfCheetah | $c_{\text{pitch}}(s_t) = \left(\frac{\|\theta_t\| - \pi/2}{\pi/2}\right)_+$<br>$c_{joint}(s_t) = \max_j \left(\frac{\|\dot{q}_{t,j}\| - \nu_j}{\nu_j}\right)_+, \quad j = 1, \ldots, 6$ |

every 100 episodes, used random seed 5, and performed 5 policy update epochs per update. The robust version used the perturbed environment with robustness regularization 0.001, while the nominal surrogate-objective run used no robustness regularization. The RCRL and FAC multi-constraint HalfCheetah baselines were trained for 8000 steps, evaluated every 100 episodes, used random seed 5, and used no robustness regularization. For **Quadrotor**, the multi-constraint setting used four constraints with a safety baseline of 0.5, penalty coefficient 30000, 5 policy update epochs, a warm-start period of 1000 episodes, batch size 1024, mini-batch size 256, actor learning rate 5e-4, reward-critic learning rate 5e-4, cost-critic learning rate 5e-4, and cost scaling factor 100.0. The nominal Quadrotor multi-constraint run was trained for 35000 episodes without robustness regularization, the robust run was trained for 35000 episodes with robustness regularization 0.001, and the FAC baselines were trained for 35000 episodes as well.

**CMDP reformulation experiments.** For the CMDP reformulation baselines, we used CMDP versions of HalfCheetah, Humanoid, and Safety Gym. **HalfCheetah** CMDP experiments used HalfCheetah CMDP and HalfCheetah CMDP Perturbed. These runs were trained for 16000 episodes in the dense CMDP setting , with a warm-start period of 500 episodes, 5 policy update epochs, a safety baseline of 0.1, and cost-critic learning rates typically set to 5e-4. **Humanoid** CMDP experiments used Humanoid CMDP trained for 6000 episodes with a warm-start period of 300 episodes, a safety baseline of 0.1, actor learning rate 1e-3, reward-critic learning rate 5e-4, cost-critic learning rate 5e-4, and entropy coefficient 0.01. The perturbed sparse CMDP run used gravity perturbation scale 0.7. **Safety Gym** CMDP experiment runs were trained for 16000 episodes in the dense CMDP setting , with a warm-start period of 1300 episodes, 5 policy update epochs, a safety baseline of 0.2, and critic learning rates typically set to 3e-4 and actor learning rate 3e-3. The training used a minibatch of 256.

### C.4 Cost Designs

We summarize the cost functions used in our experiments. For single-constraint tasks, the costs are based on action magnitude, action energy, or distance to the closest safety boundary. For multi constraint tasks, the costs are based on action. Detailed cost designs are provided in Tables 1 and 2.

Table 3: Environment details. Here, $S_{dim}$ denotes the state dimension, $A_{dim}$ denotes the action dimension, and $a_{\max}$ denotes the maximum action magnitude.

| Environment | $S_{dim}$ | $A_{dim}$ | $a_{\max}$ |
|---|---|---|---|
| Ant | 113 | 8 | 1 |
| HalfCheetah | 18 | 6 | 1 |
| Humanoid | 376 | 17 | 0.4 |
| Swimmer | 10 | 2 | 1 |
| Safety Gym Circle | 40 | 20 | 1 |
| Quadrotor | 6 | 2 | 1 |

Table 4: Perturbation design used for robustness training and evaluation. Training perturbations are sampled stepwise unless otherwise stated, while inference perturbations are sampled once per episode and then held fixed.

| Environment | Perturbation parameter | Perturbation | Values |
|---|---|---|---|
| HalfCheetah, single constraint | Gravity | $g_t = g_0 + \xi_t,\ \xi_t \sim \mathcal{N}(0, \sigma^2_{\text{grav}})$ | Train: $\sigma_{\text{grav}} = 2.0$; Test: $\sigma_{\text{grav}} = 1.5$ |
| Ant, single constraint | Gravity | $g_t = g_0 + \xi_t,\ \xi_t \sim \mathcal{N}(0, \sigma^2_{\text{grav}})$ | Train: $\sigma_{\text{grav}} = 0.7$; Test: $\sigma_{\text{grav}} = 2.0$ |
| Humanoid, single constraint | Gravity | $g_t = g_0 + \xi_t,\ \xi_t \sim \mathcal{N}(0, \sigma^2_{\text{grav}})$ | Train: $\sigma_{\text{grav}} = 0.7$; Test: $\sigma_{\text{grav}} = 2.0$ |
| Swimmer, single constraint | Viscosity | $\nu_t = \nu_0 + \xi_t,\ \xi_t \sim \mathcal{N}(0, \sigma^2_{\nu})$ | Train: $\sigma_{\nu} = 0.7$; Test: $\sigma_{\nu} = 2.0$ |
| Safety Gym Circle, single constraint | Gravity | $g_t = g_0 + \xi_t,\ \xi_t \sim \mathcal{N}(0, \sigma^2_{\text{grav}})$ | Train: $\sigma_{\text{grav}} = 0.07$; Test: $\sigma_{\text{grav}} = 0.25$ |
| HalfCheetah, multi-constraint | Gravity | $g_t = g_0 + \xi_t,\ \xi_t \sim \mathcal{N}(0, \sigma^2_{\text{grav}})$ | Train: $\sigma_{\text{grav}} = 0.7$; Test: $\sigma_{\text{grav}} = 2.0$ |
| Quadrotor, multi-constraint | Wind | $w_t \sim \text{Unif}[-0.005, 0.005]$ | Train: 0.005; Test: 0.05 |

### C.5 Environment Details

Table 3 summarizes the state dimension, action dimension, and maximum action magnitude for each environment.

### C.6 Perturbation Design for Robustness Training and Evaluation

We use stochastic dynamics perturbations during robustness training and fixed episode-level perturbations during evaluation. For gravity-based perturbations, the nominal vertical gravity value is first stored as $g_0$. During training, at each environment step we sample a fresh perturbation $\xi_t \sim \mathcal{N}(0, \sigma^2)$ and set $g_t = g_0 + \xi_t$. At reset, gravity is restored to $g_0$. During inference, we instead sample a single perturbation per evaluation episode, $\xi \sim \mathcal{N}(0, \sigma^2)$, and keep the gravity fixed throughout the episode as $g = g_0 + \xi$. This evaluates robustness under persistent dynamics shifts rather than rapidly varying stepwise noise. Swimmer uses the same protocol but perturbs viscosity instead of gravity. For Quadrotor, we perturb wind uniformly. Details are in Table 4

## D Proofs of the Convergence Results

### D.1 A unified template and notation

We identify a policy with a vector $\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and use the Euclidean norm. Since each simplex has diameter $\sqrt{2}$, $\operatorname{diam}(\Pi) = \sqrt{2|\mathcal{S}|}$. Let $\mathcal{U} = \{\pi : \sum_a |\pi(a \mid s)| < \gamma^{-1}\ \forall s\}$, an open set containing $\Pi$.

**Template.** All quantities in the analysis are instances of a single object. A *killed MDP* $\mathfrak{M} = (g, K, \nu)$ consists of a cost $g : \mathcal{S} \times \mathcal{A} \to [0, 1]$, a nonnegative kernel $K$ with $\sum_{s'} K(s' \mid s, a) \le \gamma$, and an initial weight $\nu \in \Delta(\mathcal{S})$. For $\pi \in \mathcal{U}$ let $g_\pi(s) = \sum_a \pi(a \mid s) g(s, a)$ and $K_\pi(s, s') = \sum_a \pi(a \mid s) K(s' \mid s, a)$. Since $\|K_\pi\|_\infty < 1$ on $\mathcal{U}$ and $\|K_\pi\|_\infty \le \gamma$ on $\Pi$, the following are well defined, with $\|(I - K_\pi)^{-1}\|_\infty \le H$ on $\Pi$:

$$v_\pi = (I - K_\pi)^{-1} g_\pi, \qquad Q_\pi = g + K v_\pi, \qquad U(\pi) = \nu^\top v_\pi, \qquad d_\pi = \nu^\top (I - K_\pi)^{-1}.$$

Here $(Kv)(s, a) = \sum_{s'} K(s' \mid s, a) v(s')$, and $d_\pi$ is an unnormalized occupancy with total mass at most $H$. We write $\iota_\Pi$ for the indicator of $\Pi$ in the sense of convex analysis, $\iota_\Pi(\pi) = 0$ for $\pi \in \Pi$

and $+\infty$ otherwise, $N_\Pi(\pi)$ for the normal cone to $\Pi$ at $\pi$, $\partial$ for the Fréchet subdifferential, and $(x)_+ = \max\{x, 0\}$.

**Instances.** For $P \in \mathcal{P}$, the *reward branch* is $\mathfrak{M}_P = (-r, \gamma P, d_0)$, for which $U = J_r^{\pi,P}$. For $\sigma \in \Sigma = \{0,1\}^{\mathcal{S}\times\mathcal{A}}$, $P \in \mathcal{P}$ and $s_0 \in \mathcal{S}_0$, the *peak branch* $k = (P, \sigma, s_0) \in \mathcal{K}$ is

$$\mathfrak{M}_k = \big(\tilde{c}_\sigma,\ \gamma\sigma P,\ \delta_{s_0}\big), \qquad \tilde{c}_\sigma(s,a) = (1-\gamma\sigma(s,a))\, c(s,a) \in [0,1],$$

with value $V^k(\pi) := U_k(\pi)$. Let $\mathcal{J} = \mathcal{P} \sqcup \mathcal{K}$ and

$$\psi_P(\pi) = J_r^{\pi,P}/\beta, \qquad \psi_k(\pi) = V^k(\pi) - b.$$

We write $d_{\pi,j}$ and $Q_{\pi,j}$ for the occupancy and action-value of branch $j$. The active set at $\pi$ is $\mathcal{J}^\star(\pi) = \{j \in \mathcal{J} : \psi_j(\pi) = F_\beta(\pi)\}$.

**Definition 1** (Nondegeneracy; Frank–Wolfe gap)**.** *A point $\pi \in \Pi$ is* nondegenerate *if $\nabla\psi_j(\pi)$ is the same for all $j \in \mathcal{J}^\star(\pi)$. We then write $\nabla F_\beta(\pi)$ for this common vector and $\mathcal{G}(\pi) = \max_{p\in\Pi}\langle \pi - p, \nabla F_\beta(\pi)\rangle \geq 0$.*

**Definition 2** (Coverage)**.** *For $\pi \in \Pi$ and $j \in \mathcal{J}$, let $C_j(\pi) = \|d_{\pi^\star,j}/d_{\pi,j}\|_\infty$, where the ratio is taken on the support of $d_{\pi^\star,j}$ and equals $+\infty$ if $d_{\pi,j}$ vanishes there. Define $C = \sup_\pi \min_{j\in\mathcal{J}^\star(\pi)} C_j(\pi)$, the supremum being over nondegenerate $\pi \in \Pi$.*

For a reward branch, $d_{\pi,P} \geq d_0$ and $d_{\pi^\star,P} \leq H$ entrywise, so $C_P(\pi) \leq H/\min_s d_0(s)$ whenever $d_0$ has full support. No such bound holds for peak branches, whose occupancy starts from a single state and is killed where $\sigma = 0$.

### D.2 Structure of the peak-cost value

**Lemma 2** (Branch representation)**.** *For all $\pi \in \Pi$, $P \in \mathcal{P}$ and $s \in \mathcal{S}$, $V_{c,\mathrm{peak}}^{\pi,P}(s) = \max_{\sigma\in\Sigma} v_{\pi,P}^\sigma(s)$, where $v_{\pi,P}^\sigma$ is the value vector of $\mathfrak{M}_{(P,\sigma,\cdot)}$, and the maximum is attained by a single $\sigma^\star$ simultaneously for all $s$. Consequently $W = \max_{k\in\mathcal{K}} V^k$ and $F_\beta = \max_{j\in\mathcal{J}} \psi_j$ on $\Pi$.*

*Proof.* Fix $\pi \in \Pi$ and $P \in \mathcal{P}$, and abbreviate $(P\langle\pi, Q\rangle)(s,a) = \sum_{s'} P(s' \mid s,a) \sum_{a'} \pi(a' \mid s') Q(s',a')$. Define on $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ the operators

$$(\mathcal{T}Q)(s,a) = (1-\gamma)c + \gamma \max\{c,\ P\langle\pi,Q\rangle\}, \qquad (\mathcal{T}^\sigma Q)(s,a) = (1-\gamma)c + \gamma\big[(1-\sigma)c + \sigma\, P\langle\pi,Q\rangle\big],$$

all terms evaluated at $(s,a)$. We proceed in four steps.

*Step 1: both operators are monotone $\gamma$-contractions.* If $Q \leq Q'$ entrywise then $P\langle\pi,Q\rangle \leq P\langle\pi,Q'\rangle$, because $P \geq 0$ and $\pi \geq 0$ on $\Pi$; since $x \mapsto \max\{c,x\}$ is nondecreasing and $\sigma \geq 0$, monotonicity of $\mathcal{T}$ and $\mathcal{T}^\sigma$ follows. For contraction, $|\max\{c,x\} - \max\{c,y\}| \leq |x-y|$ and $\|P\langle\pi,Q\rangle - P\langle\pi,Q'\rangle\|_\infty \leq \|Q-Q'\|_\infty$ give $\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma\|Q-Q'\|_\infty$, and likewise for $\mathcal{T}^\sigma$ with modulus $\gamma\|\sigma\|_\infty \leq \gamma$. Each therefore has a unique fixed point, reached from any starting point.

*Step 2: the fixed point of $\mathcal{T}^\sigma$ is the branch action-value.* Let $v^\sigma = v_{\pi,P}^\sigma$ be the value vector of $\mathfrak{M}_{(P,\sigma,\cdot)}$, i.e. $v^\sigma = \tilde{c}_{\sigma,\pi} + K_{\pi,P}^\sigma v^\sigma$, and set $Q^\sigma = \tilde{c}_\sigma + \gamma\sigma P v^\sigma$. Averaging $Q^\sigma$ over $\pi(\cdot \mid s)$ gives $\langle\pi(\cdot \mid s), Q^\sigma(s,\cdot)\rangle = \tilde{c}_{\sigma,\pi}(s) + (K_{\pi,P}^\sigma v^\sigma)(s) = v^\sigma(s)$. Using $(1-\gamma) + \gamma(1-\sigma) = 1 - \gamma\sigma$,

$$(\mathcal{T}^\sigma Q^\sigma)(s,a) = (1-\gamma\sigma(s,a))c(s,a) + \gamma\sigma(s,a)(Pv^\sigma)(s,a) = \tilde{c}_\sigma(s,a) + \gamma\sigma(s,a)(Pv^\sigma)(s,a) = Q^\sigma(s,a),$$

so $Q^\sigma$ is the fixed point of $\mathcal{T}^\sigma$.

*Step 3: every branch is dominated.* Let $Q$ be the fixed point of $\mathcal{T}$, so $Q = Q_{c,\mathrm{peak}}^{\pi,P}$. Since $(1-\sigma)c + \sigma x \leq \max\{c,x\}$ for every $x$ and every $\sigma \in \{0,1\}$, we have $\mathcal{T}^\sigma Q' \leq \mathcal{T}Q'$ for all $Q'$; in particular $\mathcal{T}^\sigma Q \leq \mathcal{T}Q = Q$. Suppose $(\mathcal{T}^\sigma)^n Q \leq Q$ for some $n \geq 1$. Applying the monotone operator $\mathcal{T}^\sigma$ to both sides gives $(\mathcal{T}^\sigma)^{n+1}Q \leq \mathcal{T}^\sigma Q \leq Q$, so the inequality holds for all $n$ by induction. Letting $n \to \infty$ and using Step 1, $(\mathcal{T}^\sigma)^n Q \to Q^\sigma$, whence $Q^\sigma \leq Q$.

*Step 4: one branch is exact.* Define $\sigma^\star(s,a) = \mathbf{1}\{(P\langle\pi,Q\rangle)(s,a) \geq c(s,a)\}$. At every $(s,a)$, the bracket in $\mathcal{T}^{\sigma^\star}$ selects whichever of $c(s,a)$ and $(P\langle\pi,Q\rangle)(s,a)$ is larger, so $\mathcal{T}^{\sigma^\star}Q = \mathcal{T}Q = Q$. By uniqueness of the fixed point of $\mathcal{T}^{\sigma^\star}$, $Q = Q^{\sigma^\star}$.

Combining Steps 3 and 4, $Q = \max_\sigma Q^\sigma$ with the maximum attained at $\sigma^\star$ simultaneously at every $(s, a)$. Averaging over $\pi(\cdot \mid s) \geq 0$ preserves both the inequality $Q^\sigma \leq Q$ and the equality at $\sigma^\star$, so $v^\sigma \leq V^{\pi,P}_{c,\text{peak}} = v^{\sigma^\star}$, which is the first claim. Taking the maximum over $(P, s_0) \in \mathcal{P} \times \mathcal{S}_0$ gives $W = \max_{k\in\mathcal{K}} V^k$, and combining with $J/\beta = \max_{P\in\mathcal{P}} \psi_P$ gives $F_\beta = \max\{J/\beta, W - b\} = \max_{j\in\mathcal{J}} \psi_j$. □

**Lemma 3** (Performance difference and gradient)**.** *For any killed MDP and all $\pi, \pi' \in \Pi$,*

$$U(\pi') - U(\pi) = \sum_s d_{\pi'}(s) \left\langle \pi'(\cdot \mid s) - \pi(\cdot \mid s),\, Q_\pi(s, \cdot) \right\rangle,$$

*and for $\pi \in \mathcal{U}$, $\partial U(\pi)/\partial\pi(a \mid s) = d_\pi(s)\, Q_\pi(s, a)$.*

*Proof. Invertibility.* For $\pi \in \mathcal{U}$ the matrix $K_\pi$ has absolute row sums $\sum_{s'} |K_\pi(s, s')| \leq \gamma \sum_a |\pi(a \mid s)| < 1$, so $I - K_\pi$ is invertible with $(I - K_\pi)^{-1} = \sum_{n\geq 0} K_\pi^n$; on $\Pi$ the row sums are at most $\gamma$, giving $\|(I - K_\pi)^{-1}\|_\infty \leq H$.

*Performance difference.* Write $M' = I - K_{\pi'}$, so that $v_{\pi'} = M'^{-1} g_{\pi'}$. Then

$$v_{\pi'} - v_\pi = M'^{-1} g_{\pi'} - v_\pi = M'^{-1}\big(g_{\pi'} - M' v_\pi\big) = M'^{-1}\big(g_{\pi'} + K_{\pi'} v_\pi - v_\pi\big),$$

where the second equality inserts $v_\pi = M'^{-1} M' v_\pi$. We identify the bracket. By the definitions of $g_{\pi'}$, $K_{\pi'}$ and $Q_\pi = g + K v_\pi$,

$$\big(g_{\pi'} + K_{\pi'} v_\pi\big)(s) = \sum_a \pi'(a \mid s) \Big[ g(s, a) + \sum_{s'} K(s' \mid s, a) v_\pi(s') \Big] = \big\langle \pi'(\cdot \mid s), Q_\pi(s, \cdot) \big\rangle,$$

and the same computation with $\pi$ in place of $\pi'$ gives $v_\pi(s) = \langle \pi(\cdot \mid s), Q_\pi(s, \cdot)\rangle$, since $v_\pi = g_\pi + K_\pi v_\pi$. Hence the bracket has entries $\langle \pi'(\cdot \mid s) - \pi(\cdot \mid s), Q_\pi(s, \cdot)\rangle$. Multiplying on the left by $\nu^\top$ and using $d_{\pi'} = \nu^\top M'^{-1}$ gives the stated identity.

*Gradient.* On $\mathcal{U}$ the map $\pi \mapsto v_\pi$ is differentiable, being a composition of the linear maps $\pi \mapsto (g_\pi, K_\pi)$ with matrix inversion. Differentiating $v_\pi = g_\pi + K_\pi v_\pi$ in a coordinate direction gives $\partial v_\pi = \partial g_\pi + (\partial K_\pi) v_\pi + K_\pi \partial v_\pi$, that is, $\partial v_\pi = (I - K_\pi)^{-1}\big(\partial g_\pi + (\partial K_\pi) v_\pi\big)$. For the coordinate $\pi(a \mid s)$, both $g_\pi$ and $K_\pi$ depend on it only through row $s$, so the vector $\partial g_\pi + (\partial K_\pi) v_\pi$ equals $g(s, a) + \sum_{s'} K(s' \mid s, a) v_\pi(s') = Q_\pi(s, a)$ in entry $s$ and zero elsewhere. Multiplying by $\nu^\top$ and reading off the $s$-th column of $(I - K_\pi)^{-1}$ gives $\partial U(\pi)/\partial\pi(a \mid s) = d_\pi(s) Q_\pi(s, a)$.

Both statements apply verbatim to the reward branch $\mathfrak{M}_P$ and to each peak branch $\mathfrak{M}_k$, since these are instances of the template. □

**Lemma 4** (Smoothness and weak convexity)**.** *For every killed MDP and $\pi, \pi' \in \Pi$, the map $\tau \mapsto U(\pi + \tau(\pi' - \pi))$ on $[0, 1]$ has second derivative bounded in absolute value by $L\|\pi' - \pi\|^2$ with $L = 2\gamma|\mathcal{A}|H^3$. Consequently, with $\omega = L \max\{1, 1/\beta\}$, the function $F_\beta + \frac{\omega}{2}\|\cdot\|^2$ is convex on $\Pi$. For $\mu = 1/(2\omega)$, the objective in equation 9 is $\rho$-strongly convex on $\Pi$ and has a unique minimizer.*

*Proof. Smoothness.* Fix $\pi, \pi' \in \Pi$, let $u = \pi' - \pi$ and $\pi_\tau = \pi + \tau u$. Along this segment $\pi_\tau \in \Pi$, and $U(\pi_\tau) = \nu^\top (I - K_{\pi_\tau})^{-1} g_{\pi_\tau}$ with both $g_{\pi_\tau}$ and $K_{\pi_\tau}$ affine in $\tau$. Differentiating twice and using $\frac{d}{d\tau}(I - K_{\pi_\tau})^{-1} = (I - K_{\pi_\tau})^{-1}\big(\frac{d}{d\tau} K_{\pi_\tau}\big)(I - K_{\pi_\tau})^{-1}$ expresses $\frac{d^2}{d\tau^2} U(\pi_\tau)$ as a sum of terms, each a product of at most three resolvents, at most two derivative matrices, and one cost vector. This is precisely the computation in (Agarwal et al., 2020, Lemma D.3), whose bound $2\gamma|\mathcal{A}|/(1-\gamma)^3$ uses only that the per-step cost lies in $[0, 1]$, that the kernel has nonnegative entries with row sums at most $\gamma$ so that $\|(I - K_{\pi_\tau})^{-1}\|_\infty \leq H$, and that the kernel is linear in $\pi$. All three hold for every killed MDP: $\tilde{c}_\sigma \in [0, 1]$ and $-r \in [0, 1]$, and $\gamma\sigma P$ has row sums at most $\gamma$. Hence $|\frac{d^2}{d\tau^2} U(\pi_\tau)| \leq L\|u\|^2$ with $L = 2\gamma|\mathcal{A}|H^3$, so $\psi_k$ is $L$-smooth and $\psi_P$ is $(L/\beta)$-smooth along segments of $\Pi$; the constants are uniform in $P \in \mathcal{P}$, $\sigma \in \Sigma$ and $s_0 \in \mathcal{S}_0$, since they do not involve these quantities.

*Weak convexity.* Let $\omega = L \max\{1, 1/\beta\}$, so $L_j \leq \omega$ for every $j \in \mathcal{J}$. Fix $j$ and consider $\phi(\tau) = \psi_j(\pi_\tau) + \frac{\omega}{2}\|\pi_\tau\|^2$. Then $\phi''(\tau) = \frac{d^2}{d\tau^2}\psi_j(\pi_\tau) + \omega\|u\|^2 \geq -L_j\|u\|^2 + \omega\|u\|^2 \geq 0$, so $\phi$ is convex on $[0, 1]$. As $\pi, \pi'$ were arbitrary points of $\Pi$ and $\Pi$ is convex, $\psi_j + \frac{\omega}{2}\|\cdot\|^2$ is convex on $\Pi$.

By Lemma 2, $F_\beta + \frac{\omega}{2}\|\cdot\|^2 = \sup_{j\in\mathcal{J}} \left(\psi_j + \frac{\omega}{2}\|\cdot\|^2\right)$, a pointwise supremum of convex functions, hence convex on $\Pi$.

*Strong convexity of the subproblem.* With $\mu = 1/(2\omega)$, expand

$$F_\beta(\pi) + \tfrac{1}{2\mu}\|\pi - \pi^t\|^2 = \underbrace{\left[F_\beta(\pi) + \tfrac{\omega}{2}\|\pi\|^2\right]}_{\text{convex on } \Pi} + \underbrace{\left[\tfrac{\omega}{2}\|\pi\|^2 - 2\omega\langle\pi, \pi^t\rangle + \omega\|\pi^t\|^2\right]}_{\omega\text{-strongly convex}},$$

using $\frac{1}{2\mu} = \omega$. A convex function plus a $\omega$-strongly convex one is $\omega$-strongly convex, and a strongly convex function on a compact convex set has a unique minimizer. □

**Lemma 5** (Directional derivatives)**.** *Under Assumption 1(a), $F_\beta$ is continuous on $\Pi$. For $\pi \in \Pi$ and $u$ with $\pi + \tau u \in \Pi$ for small $\tau > 0$, $F'_\beta(\pi; u) = \max_{j\in\mathcal{J}^\star(\pi)} \langle \nabla\psi_j(\pi), u\rangle$. In particular, $F'_\beta(\pi; u) = \langle \nabla F_\beta(\pi), u\rangle$ at nondegenerate $\pi$.*

*Proof. Compactness of the index set.* $\mathcal{K} = \mathcal{P} \times \Sigma \times \mathcal{S}_0$ is compact, being a product of a compact set with two finite sets, and $\mathcal{J} = \mathcal{P} \sqcup \mathcal{K}$ is a disjoint union of two compact sets, hence compact in the topology that makes each summand open and closed.

*Joint continuity.* Fix $\pi \in \mathcal{U}$. For a branch $j$ with kernel $K^{(j)}$ and cost $g^{(j)}$, both $K^{(j)}_\pi$ and $g^{(j)}_\pi$ are bilinear in $(\pi, P)$ for the reward branch and in $(\pi, P)$ for fixed $\sigma$ for a peak branch, hence jointly continuous in $(j, \pi)$; note that $\sigma$ and $s_0$ range over finite sets, on which every function is continuous. Matrix inversion is continuous on the open set of invertible matrices, and $I - K^{(j)}_\pi$ is invertible throughout $\mathcal{J} \times \mathcal{U}$ by the Neumann series argument of Lemma 3. Therefore $\psi_j(\pi)$, and by the closed-form expression $d_{\pi,j}(s)Q_{\pi,j}(s, a)$ of Lemma 3 also $\nabla\psi_j(\pi)$, are jointly continuous on $\mathcal{J} \times \mathcal{U}$.

*Continuity of $F_\beta$.* A pointwise maximum of a family of functions that is jointly continuous in the index and the argument, over a compact index set, is continuous; this is Berge's maximum theorem, or directly: for $\pi$ in a compact neighbourhood the family is uniformly equicontinuous, and a uniform limit of continuous functions is continuous.

*Directional derivative.* The hypotheses of Danskin's theorem (Clarke, 1990) are exactly those verified above: $\mathcal{J}$ compact, $\psi_j(\pi)$ continuous in $(j, \pi)$, and $\psi_j$ continuously differentiable in $\pi$ with $\nabla_\pi\psi_j$ continuous in $(j, \pi)$. The theorem gives, for every direction $u$,

$$F'_\beta(\pi; u) = \lim_{\tau\downarrow 0} \frac{F_\beta(\pi + \tau u) - F_\beta(\pi)}{\tau} = \max_{j\in\mathcal{J}^\star(\pi)} \langle \nabla\psi_j(\pi), u\rangle.$$

We use it only for directions $u$ with $\pi + \tau u \in \Pi$ for small $\tau > 0$, which is legitimate because $\psi_j$ is defined and differentiable on the open set $\mathcal{U} \supset \Pi$. If $\pi$ is nondegenerate, all gradients in the maximum coincide with $\nabla F_\beta(\pi)$ and the maximum evaluates to $\langle \nabla F_\beta(\pi), u\rangle$. □

### D.3 Proximal point dynamics and gradient domination

**Lemma 6** (Proximal point steps)**.** *Under Assumption 1(a) and for any $\mu > 0$, the iteration equation 9 is well defined, and for all $t \geq 0$ and $T \geq 1$:*

*(a)* $F_\beta(\pi^{t+1}) + \frac{1}{2\mu}\|\pi^{t+1} - \pi^t\|^2 \leq F_\beta(\pi^t)$;

*(b)* $\sum_{t<T} \|\pi^{t+1} - \pi^t\|^2 \leq 2\mu\big(F_\beta(\pi^0) - \min_\Pi F_\beta\big)$;

*(c) if $\pi^{t+1}$ is nondegenerate, then* $\mathcal{G}(\pi^{t+1}) \leq \operatorname{diam}(\Pi)\, \|\pi^{t+1} - \pi^t\|/\mu$;

*Proof.* A minimizer exists because $F_\beta$ is continuous (Lemma 5) and $\Pi$ is compact. Part (a) compares the objective at $\pi^{t+1}$ with its value at $\pi^t$, and (b) telescopes (a). For (c), fix $p \in \Pi$ and let $u = p - \pi^{t+1}$. The function $\phi(\tau) = F_\beta(\pi^{t+1} + \tau u) + \frac{1}{2\mu}\|\pi^{t+1} + \tau u - \pi^t\|^2$ is minimized over $[0, 1]$ at $\tau = 0$, so $\phi'(0^+) \geq 0$. By Lemma 5, this gives $\langle \nabla F_\beta(\pi^{t+1}), u\rangle + \frac{1}{\mu}\langle \pi^{t+1} - \pi^t, u\rangle \geq 0$, that is, $\langle \pi^{t+1} - p, \nabla F_\beta(\pi^{t+1})\rangle \leq \frac{1}{\mu}\|\pi^{t+1} - \pi^t\|\, \|p - \pi^{t+1}\|$. Maximizing over $p$ completes the proof. □

**Lemma 7** (Gradient domination at nondegenerate points). *Let $\pi \in \Pi$ be nondegenerate and $\pi' \in \Pi$. For every $j \in \mathcal{J}^\star(\pi)$,*

$$F_\beta(\pi) - F_\beta(\pi') \le C'_j(\pi)\, \mathcal{G}(\pi), \qquad C'_j(\pi) = \|d_{\pi',j}/d_{\pi,j}\|_\infty.$$

*Proof.* Since $j$ is active and $F_\beta \ge \psi_j$, $F_\beta(\pi) - F_\beta(\pi') \le \psi_j(\pi) - \psi_j(\pi')$. Let $U_j$ be the value of branch $j$ and $\kappa_j \in \{1, 1/\beta\}$ its scaling, so that $\psi_j = \kappa_j U_j + \text{const}$ and $\nabla F_\beta(\pi) = \kappa_j \nabla U_j(\pi)$. By Lemma 3,

$$U_j(\pi) - U_j(\pi') = \sum_s d_{\pi',j}(s)\, \langle \pi(\cdot \mid s) - \pi'(\cdot \mid s), Q_{\pi,j}(s,\cdot)\rangle \le \sum_s d_{\pi',j}(s)\, h(s),$$

where $h(s) = \max_{q \in \Delta(\mathcal{A})} \langle \pi(\cdot \mid s) - q, Q_{\pi,j}(s,\cdot)\rangle \ge 0$. Since $h \ge 0$, we have $\sum_s d_{\pi',j}(s)h(s) \le C'_j(\pi) \sum_s d_{\pi,j}(s)h(s)$. The maximum over $\Pi$ separates across states, so by Lemma 3, $\sum_s d_{\pi,j}(s)h(s) = \max_{p \in \Pi} \langle \pi - p, \nabla U_j(\pi)\rangle$. Multiplying by $\kappa_j$ gives the claim. □

### D.4 Proofs of Theorems 2 and 3

*Proof of Theorem 2.* By definition $F_\beta(\pi^0) = \Delta_0$, and by Lemma 6(a) $F_\beta(\pi^t) \le \Delta_0$ for all $t$. Since $F_\beta \ge W - b$ and $F_\beta \ge J/\beta$; if $\pi^0$ is feasible then $W(\pi^0) - b \le 0 \le J(\pi^0)/\beta$, so $\Delta_0 = J(\pi^0)/\beta$. Next, $F_\beta \ge J/\beta \ge 0$ gives $\min_\Pi F_\beta \ge 0$, so Lemma 6(b) yields $\min_{t<T} \|\pi^{t+1} - \pi^t\|^2 \le 2\mu\Delta_0/T$. Dividing by $\mu^2$ and using $\mu = 1/(2\rho)$ gives equation 10. □

*Proof of Theorem 3.* By Lemma 6(c), equation 10 and $\operatorname{diam}(\Pi) = \sqrt{2|\mathcal{S}|}$,

$$\mathcal{G}(\pi^{\bar t+1}) \le \sqrt{2|\mathcal{S}|}\, \frac{\|\pi^{\bar t+1} - \pi^{\bar t}\|}{\mu} \le 2\sqrt{\frac{2|\mathcal{S}|\rho\Delta_0}{T}}, \qquad \text{so} \qquad C\,\mathcal{G}(\pi^{\bar t+1}) \le E_T.$$

Apply Lemma 7 at $\pi^{\bar t+1}$ with $\pi' = \pi^\star$ and an active branch attaining the minimum in Definition 2: $F_\beta(\pi^{\bar t+1}) \le F_\beta(\pi^\star) + C\mathcal{G}(\pi^{\bar t+1}) = J(\pi^\star)/\beta + C\mathcal{G}(\pi^{\bar t+1})$, where the equality uses feasibility of $\pi^\star$. By Lemma 6(a), $F_\beta(\pi^t) \le F_\beta(\pi^{\bar t+1}) \le J(\pi^\star)/\beta + E_T$ for every $t \ge \bar t + 1$; monotonicity also shows that $\pi^T$ minimizes $F_\beta$ over the iterates. Using $J(\pi^t)/\beta \le F_\beta(\pi^t)$ and $W(\pi^t) - b \le F_\beta(\pi^t)$, which gives equation 11.

Since $\epsilon \le H$ we have $\beta = H/\epsilon \ge 1$ and $\rho = 2\gamma|\mathcal{A}|H^3$. When $\pi^0$ is feasible, $\Delta_0 \le H/\beta = \epsilon$, so

$$\beta E_T = 2C\beta\sqrt{\frac{2|\mathcal{S}|\rho\Delta_0}{T}} \le 2C\sqrt{\frac{2|\mathcal{S}|\rho\beta H}{T}} = 4C\sqrt{\frac{\gamma|\mathcal{S}||\mathcal{A}|H^5}{\epsilon T}} \le \epsilon$$

for $T \ge 16\gamma C^2|\mathcal{S}||\mathcal{A}|H^5\epsilon^{-3}$, and the violation is at most $\Delta_0 \le \epsilon$ at every iterate. When $\pi^0$ is infeasible, $\Delta_0 \le \max\{H/\beta, 1-b\} = \bar\Delta$ and $\beta E_T = 4C(H/\epsilon)\sqrt{\gamma|\mathcal{S}||\mathcal{A}|H^3\bar\Delta/T} \le \epsilon$ for the stated $T$; then $E_T \le \epsilon/\beta \le \epsilon$ and the violation is at most $J(\pi^\star)/\beta + E_T \le \epsilon + \epsilon$. □

**Remark 7** (Inexact subproblems). *If $\pi^{t+1}$ is only $\epsilon_{\text{in}}$-optimal for equation 9, Lemma 6(a) holds with an additional $\epsilon_{\text{in}}$ term on the right-hand side. Hence $F_\beta(\pi^t) \le J(\pi^0) + t\epsilon_{\text{in}}$ holds with violation at most $J(\pi^0)/\beta + t\epsilon_{\text{in}}$ when $\pi^0$ is feasible, and Lemma 5(b) holds with $2\mu(F_\beta(\pi^0) - \min F_\beta + T\epsilon_{\text{in}})$ in the right-hand side. Transferring (c) requires relating $\pi^{t+1}$ to the exact proximal point, which is possible through $\rho$-strong convexity: $\|\pi^{t+1} - \hat\pi^{t+1}\| \le \sqrt{2\epsilon_{\text{in}}/\rho}$. We do not pursue this here and has been left for the future.*

**Remark 8** (Scope of Theorem 3). *The nondegeneracy condition excludes the kink $\{J/\beta = W - b\}$ and ties in the branch selection. When the peak-cost constraint binds, minimizers of $F_\beta$ typically lie on this kink, so Theorem 3 certifies global optimality for outputs away from it, while Theorem 2 holds unconditionally. At a kink the relevant gradient combines occupancies of two different chains – the full chain for $J$ and a killed chain for $W$ – to which Lemma 7 does not apply; extending gradient domination there is left open.*