Constraint-Conditioned Policy Optimization for Versatile Safe Reinforcement Learning

Yihang Yao*[1], Zuxin Liu*[1], Zhepeng Cen[1], Jiacheng Zhu[1,3], Wenhao Yu[2], Tingnan Zhang[2], Ding Zhao[1]

[1] Carnegie Mellon University
[2] Google DeepMind
[3] Massachusetts Institute of Technology

January 30, 2026

Foundation of Safe RL

Standard reinforcement learning (RL) aims to learn policies that maximize the task reward return.

Safe reinforcement learning (RL) aims not only maximize reward return, but also satisfy certain constraints (limit the constraint violation rate to a certain level) before deploying to safety-critical applications.

The "Versatile" needs vs current limitation:

- Traditional safe RL policies are typically trained for a single, fixed constraint threshold and cannot adapt to new safety requirements without extensive retraining.
- Real-world applications require agents that can adapt their conservativeness.
- E.g., an autonomous vehicle adapt safety thresholds for driving on an empty highway vs. crowded urban area to maximize transportation efficiency

Objective

Primary challenges:

- **Training efficiency:** train multiple policies under different constraint threshold is sampling inefficient
- **Zero-shot adaptation capability:** adapting the learned policy to accommodate unseen safety thresholds.

Conditioned Constrained Policy Optimization (CCPO): a sampling-efficient algorithm for versatile safe reinforcement learning that achieves zero-shot generalization to unseen cost thresholds during deployment.

Setup: CMDP

A Constrained Markov Decision Process (CMDP) $\mathcal{M}$ is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, c, \mu_0)$, that augments MDP with an additional element $c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ to characterize the cost of violating the constraint.

The Safe RL Objective

Find a policy π that maximizes reward return while limiting cost return under a threshold ϵ:

$$\pi^* = \arg\max_{\pi} V_r^{\pi}(\mu_0), \quad \text{s.t. } V_c^{\pi}(\mu_0) \le \epsilon \tag{1}$$

where $V_f^{\pi}(\mu_0) = \mathbb{E}_{\tau\sim\pi, s_0\sim\mu_0}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{f}_t\right], \mathbf{f} \in \{r, c\}$.

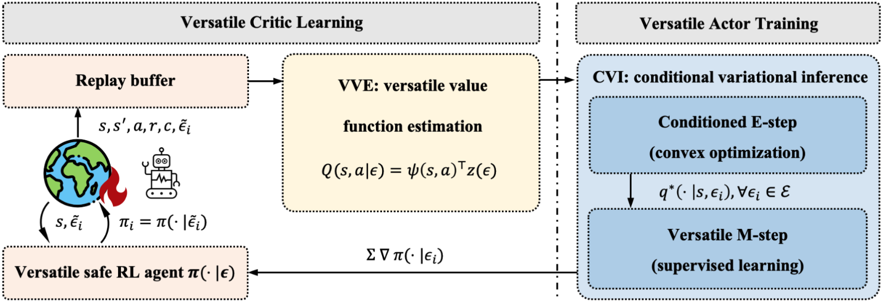
Versatile Safe RL Objective

Find the optimal versatile policy $\pi^*(\cdot|\epsilon)$ in constrained-conditioned policy space:

$$\pi^*(\cdot|\epsilon) = \arg\max_{\pi} V_r^{\pi}(\mu_0), \quad \text{s.t. } V_c^{\pi}(\mu_0) \le \epsilon, \quad \forall \epsilon \in \mathcal{E} \tag{2}$$

Proposed Method

- **Versatile Value Estimation (VVE)**: Uses representation learning to estimate value functions for unseen thresholds.
- **Conditioned Variational Inference (CVI)**: Uses "RL as inference" framework to encode arbitrary thresholds into the policy optimization with convergence guarantee.



VVE

Motivation: To adapt to unseen thresholds, the agent must estimate value functions (Q_r, Q_c) for conditions not present in the training data.

Assumption 1: (Linear decomposition)

The optimal versatile Q-functions $Q_f^*(s, a|\epsilon)$ with respect to the optimal versatile policy π^* can be represented as:

$$Q_f^*(s, a|\epsilon) = \psi_f(s, a)^\top z_f^*(\epsilon), \quad f \in \{r, c\} \tag{3}$$

where $\psi_f(s, a)$ represents state-action features representing environment dynamics, and $z_f^*(\epsilon)$ is the task features representing the specific constraint threshold.

Result: VVE disentangles environmental dynamics and target thresholds within a latent space, effectively encodes the threshold information ϵ into the Q functions and achieve accurate estimations for unseen thresholds.

Bounded Estimation Error of VVE

Assumption 2: (Polynomial feature space)

The optimal constraint-conditioned policy feature can be approximated by $z_f^*(\epsilon) = \text{Poly}(\epsilon, p) + e$.

Theorem 1: Bounded estimation error

With confidence level $1 - \alpha$, the error for an arbitrary threshold $\epsilon \in [\epsilon_L, \epsilon_H]$ is bounded by:

$$\|\hat{Q}_f(s, a|\epsilon) - Q_f^*(s, a|\epsilon)\| \leq \frac{z_{\alpha/2} B(p)}{N^{\beta(p)}} \sqrt{\sigma^2 K_f^2 M}, \quad (4)$$

- p is the polynomial degree corresponds to the $z(\epsilon)$ representation capability;
- N is the number of selected thresholds for behavior policies: $\{\tilde{\epsilon}_i\}_{i=1,2,\ldots,N}$;
- K_f is the norm constraints on feature function: $\|\psi_f(s, a)\|_\infty \leq K_f$;
- M is the dimension of $\psi_f(s, a)$ and $z_f^*(\epsilon)$.

CVI

CVI is a **constraint-conditioned extension** build on *safe RL as inference* framework (CVPO, 2022), which decomposes safe RL to convex optimization followed by supervised learning.

Objective function for safe RL:

$$\pi^* = \arg\max_{\pi} V_r^{\pi}(\mu_0), \quad s.t. \quad V_c^{\pi}(\mu_0) \leq \epsilon.$$

Standard primal-dual style approaches:

- Transform the primal objective into dual by introducing the Lagrange multiplier λ, and solve the min-max problem iteratively:

$$(\pi^*, \lambda^*) = \arg\min_{\lambda \geq 0} \max_{\pi} J_r(\pi) - \lambda(J_c(\pi) - \epsilon_1). \quad (5)$$

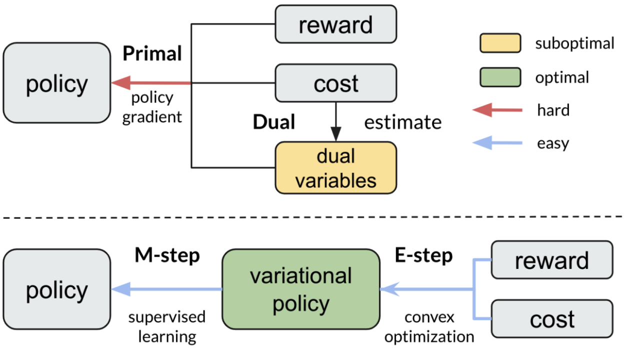
- Suffer from instability issue and lack optimality guarantees.

Safe RL as inference framework:

- Introducing a variational distribution and solving constrained optimization by EM algorithm.
- Provide sample efficiency, stable performance, and optimality guarantees.

Primal-dual view vs. inference view

- **Primal-dual formulation:** what are the actions that could maximize task rewards while satisfying the constraints?
- **RL as inference:** given future success in maximizing task rewards, what are the feasible actions most likely to have been taken?



ELBO Objective

For a given trajectory τ, the likelihood of being optimal is:

$$p(O = 1|\tau) \propto \exp\left(\sum_t \gamma^t r_t/\alpha\right)$$

The probability of getting a trajectory τ under the conditioned policy $\pi(\cdot|\epsilon_i)$ is:

$$p_{\pi(\cdot|\epsilon_i)}(\tau) = p(s_0) \prod_{t\geq 0} p(s_{t+1}|s_t, a_t)\pi(a_t|s_t, \epsilon_i)$$

The ELBO for the log-likelihood of trajectory-level optimality:

$$\begin{aligned} \log p_{\pi(\cdot|\epsilon_i)}(O = 1) &= \log \mathbb{E}_{\tau\sim q(\cdot|\epsilon_i)} \frac{p(O = 1|\tau)p_\pi(\tau|\epsilon_i)}{q(\tau|\epsilon_i)} \\ &\geq \mathbb{E}_{\tau\sim q(\cdot|\epsilon_i)} \log \frac{p(O = 1|\tau)p_{\pi(\cdot|\epsilon_i)}(\tau)}{q(\tau|\epsilon_i)} \\ &\propto \mathbb{E}_{\tau\sim q(\cdot|\epsilon_i)} \left[\sum_{t=0}^{\infty} \gamma^t r_t\right] - \alpha D_{\mathrm{KL}}(q(\tau|\epsilon_i)\|p_{\pi(\cdot|\epsilon_i)}(\tau)) := \mathcal{J}(q, \pi|\epsilon_i), \end{aligned} \tag{6}$$

where $q(\tau|\epsilon_i)$ is an auxiliary trajectory-wise variational distribution.

ELBO Objective

Given that $q(\tau|\epsilon_i) = p(s_0)\prod_{t\geq 0} p(s_{t+1}|s_t, a_t)q(a_t|s_t, \epsilon_i)$, reformat the ELBO over the state and conditioned action distribution:

$$\mathcal{J}(q, \theta|\epsilon_i) = \mathbb{E}_{\rho_q(\cdot|\epsilon_i)}\left[\sum_{t=0}^{\infty} \gamma^t r_t - \alpha D_{\mathrm{KL}}(q(\cdot|\epsilon_i)\|\pi_\theta(\cdot|\epsilon_i))\right] + \log p(\theta)$$

The objective function is solved following an EM-style:
E-step find the optimal variational distribution q^* to:

- maximize the return of task reward;
- satisfy the safety constraints;
- stay within trust region of old policy.

M-step minimize the KL divergence between $p_{\pi(\cdot|\epsilon_i)}(\tau)$ and $q(\cdot|\epsilon_i)$.

Constraint-Conditioned E-step

The ELBO objective w.r.t q as a constrained optimization problem:

$$\begin{aligned} \max_{q(a|s,\epsilon_i)} \ & \mathbb{E}_{\rho_q}\left[\int q(a|s,\epsilon_i)\hat{Q}_r^{\pi_{\theta_j}}(s,a|\epsilon_i)da\right] \\ \text{s.t. } & \mathbb{E}_{\rho_q}\left[\int q(a|s,\epsilon_i)\hat{Q}_c^{\pi_{\theta_j}}(s,a|\epsilon_i)da\right] \leq \epsilon_i, \\ & \mathbb{E}_{\rho_q}\left[D_{\mathrm{KL}}\left(q(a|s,\epsilon_i)\|\pi_{\theta_j}(\cdot|\epsilon_i)\right)\right] \leq \kappa; \end{aligned} \tag{7}$$

The solution of the optimal $q_i^* = q_i^*(a|s,\epsilon_i)$ has the closed form:

$$q_i^* = \frac{\pi_{\theta_j}(\cdot|\epsilon_i)}{Z(s,\epsilon_i)} \exp\left(\frac{\hat{Q}_r^{\pi_{\theta_j}}(\cdot|\epsilon_i) - \lambda_i^* \hat{Q}_c^{\pi_{\theta_j}}(\cdot|\epsilon_i)}{\eta_i^*}\right), \tag{8}$$

and the dual variables η_i^* and λ_i^* are solved by **convex optimization**:

$$\min_{\lambda_i,\eta_i\geq 0} g(\eta_i,\lambda_i) = \lambda_i\epsilon_i + \eta_i\kappa\mathbb{E}_{\rho_q}\left[\log \mathbb{E}_{\pi(\cdot|\epsilon_i)}\left[\exp\left(\frac{\hat{Q}_r(\cdot|\epsilon_i) - \lambda_i\hat{Q}_c(\cdot|\epsilon_i)}{\eta_i}\right)\right]\right]. \tag{10}$$

Versatile M-step

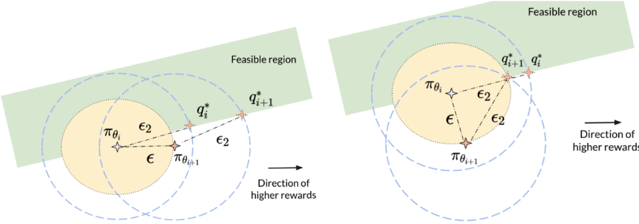
Improve the ELBO w.r.t. the policy parameter θ:

$$\mathcal{J}(\theta|\epsilon_i) = \mathbb{E}_{\rho_q}\left[\alpha \mathbb{E}_{q_i^*}[\log \pi_\theta(a|s,\epsilon_i)]\right] + \log(p|\epsilon_i)$$

Converts to supervised-learning problem with KL-divergence constraints:

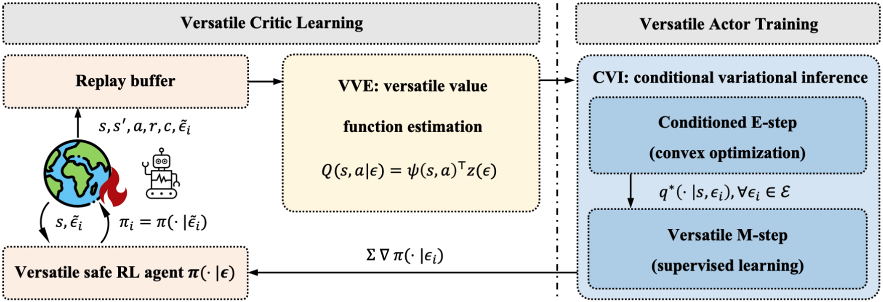
$$\max_\theta \mathbb{E}_{\rho_q}\left[\sum_{i=1}^{|\mathcal{E}|} \mathbb{E}_{q_i^*}[\log \pi_\theta(a|s,\epsilon_i)]/|\mathcal{E}|\right] s.t. \mathbb{E}_{\rho_q}[D_{\mathrm{KL}}(\pi_{\theta_j}(a|s,\epsilon_i)\|\pi_\theta(a|s,\epsilon_i))] \leq \gamma \forall i,$$

where $\mathcal{E}$ is the set of all the sampled versatile policy conditions $\{\epsilon_i\}$ in the fine-tuning stage of training.



(a) Regular policy updating (b) Worst-case policy updating

Overview of CCPO



Experiment

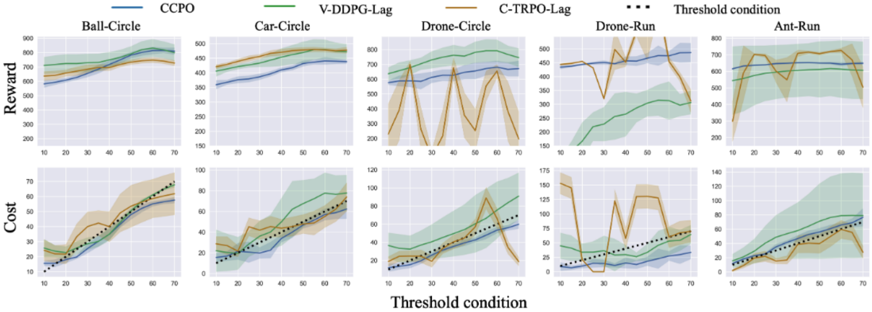
- **Tasks:** Run and Circle with four robots (Ball, Car, Drone, and Ant).
- **Reward:** running fast between two boundaries or in a circle
- **Constraint:** run across boundaries or exceed an agent-specific velocity threshold.
- **Baselines:**
  - **Constraint-conditioned baselines**:
    Directly integrating the threshold as part of the state and adopting commonly used safe RL to optimize with behavior policy conditions only.
  - **Policy linear combination baselines**:

$$\pi(\cdot|\epsilon) = w_1\pi(\cdot|\epsilon_1) + w_2\pi(\cdot|\epsilon_2); \quad w_1 = (\epsilon_2 - \epsilon)/(\epsilon_2 - \epsilon_1),\ w_2 = (\epsilon - \epsilon_1)/(\epsilon_2 - \epsilon_1)$$

Main Results

Task	Stats	CCPO (ours)	Constraint-conditioned		Linear combination	
			V-SAC-Lag	V-DDPG-Lag	C-PPO-Lag	C-TRPO-Lag
Ball-Circle	Avg. R ↑	710.86±20.47	774.16±20.34	762.61±58.65	637.85±14.03	699.38±1.94
	Avg. CV ↓	0.59±0.31	5.32±5.00	2.81±1.12	3.11±1.64	4.50±0.08
	Avg. R-G ↑	699.04±20.48	766.52±22.59	756.67±58.48	667.89±12.17	699.14±2.05
	Avg. CV-G ↓	0.83±0.42	6.29±5.72	3.53±1.26	3.40±1.75	5.59±0.25
Car-Circle	Avg. R ↑	406.06±6.30	331.80±11.57	448.82±18.65	440.01±2.59	461.14±1.39
	Avg. CV ↓	1.60±0.91	12.18±4.65	14.48±8.14	9.09±1.52	7.84±1.71
	Avg. R-G ↑	401.53±5.59	331.19±11.00	445.32±17.42	438.31±3.03	460.72±1.15
	Avg. CV-G ↓	1.49±0.38	12.74±4.32	14.63±8.69	11.07±1.58	9.14±2.01
Drone-Circle	Avg. R ↑	630.55±40.03	693.69±22.37	734.58±49.69	392.64±23.13	380.77±18.62
	Avg. CV ↓	0.32±0.38	13.24±8.80	19.62±11.15	0.45±0.38	6.55±1.95
	Avg. R-G ↑	625.51±40.12	699.14±24.88	730.29±48.43	342.77±19.06	291.87±19.88
	Avg. CV-G ↓	0.47±0.55	14.97±10.10	19.44±10.36	0.21±0.09	7.23±2.03
Drone-Run	Avg. R ↑	458.69±12.98	355.61±35.44	244.60±48.29	398.88±21.53	461.70±4.91
	Avg. CV ↓	0.23±0.25	8.66±4.30	11.33±9.63	9.46±5.63	47.97±3.49
	Avg. R-G ↑	455.64±11.83	354.61±33.34	236.61±43.49	386.77±30.09	464.07±6.61
	Avg. CV-G ↓	0.33±0.37	9.96±4.54	12.72±9.91	11.18±7.46	60.39±4.32
Ant-Run	Avg. R ↑	660.88±4.82	615.73±91.99	594.75±172.35	636.06±6.78	629.83±7.84
	Avg. CV ↓	3.13±1.67	8.47±3.55	23.69±30.42	5.16±1.59	0.22±0.17
	Avg. R-G ↑	660.07±5.26	626.27±84.61	592.50±173.01	620.46±9.99	605.07±10.63
	Avg. CV-G ↓	3.25±1.48	7.76±11.83	22.90±9.39	6.73±2.32	0.03±0.06

Main Results



Evaluation of ϵ-sampling efficiency

Training on behavior policy set $\tilde{\mathcal{E}} = \{20, 40, 60\}$ vs. $\tilde{\mathcal{E}}' = \{20, 30, 40, 50, 60, 70\}$, and evaluating on threshold conditions $\mathcal{E} = \{10, 15, ..., 70\}$

Table 2: ϵ-sampling efficiency evaluation. ↑: the higher reward, the better. ↓: the lower constraint violation (minimal 0), the better. The models are evaluated on a series of threshold conditions and we report the averaged reward and constraint violation values on all evaluation thresholds and generalized thresholds. Each value is reported as mean ± standard deviation for 50 episodes and 5 seeds. We shade the safest agent with the lowest averaged cost violation value.

Algorithm	Stats	Ball-Circle	Car-Circle	Drone-Circle	Drone-Run	Averaged Score
CCPO with $\tilde{\mathcal{E}}$	Avg. R ↑	710.86±20.47	406.06±6.30	630.55±40.03	458.69±12.98	551.54
	Avg. CV ↓	0.59±0.31	1.60±0.91	0.32±0.38	0.23±0.25	0.69
C-TRPO with $\tilde{\mathcal{E}}$	Avg. R ↑	699.38±1.94	461.14±1.39	380.77±18.62	461.70±4.91	500.75
	Avg. CV ↓	4.50±0.08	7.84±1.71	6.55±1.95	47.97±3.49	16.72
C-TRPO with $\tilde{\mathcal{E}}'$	Avg. R ↑	682.94±8.08	458.13±2.22	411.91±8.95	472.89±2.65	506.47
	Avg. CV ↓	2.66±0.37	11.90±2.12	5.20±0.81	30.20±2.47	12.49

Ablation Study

Table 3: Ablation study of removing the versatile value function estimation (VVE), and the conditioned variational inference (CVI). ↑: the higher reward, the better. ↓: the lower constraint violation (minimal 0), the better. Each value is reported as mean ± standard deviation for 50 episodes and 5 seeds. Each value is reported as mean ± standard deviation.

Algorithm	Stats	Ball-Circle	Car-Circle	Drone-Circle	Drone-Run	Ant-Run
CCPO (Full)	Avg. R ↑	710.86±20.47	406.06±6.30	630.55±40.03	458.69±12.98	660.88±4.82
	Avg. CV ↓	0.59±0.31	1.60±0.91	0.32±0.38	0.23±0.25	3.13±1.67
	Avg. R-G ↑	699.04±20.48	401.53±5.59	625.51±40.12	455.64±11.83	660.07±5.26
	Avg. CV-G ↓	0.83±0.42	1.49±0.38	0.47±0.55	0.33±0.37	3.25±1.48
CCPO w/o VVE	Avg. R ↑	674.55±17.81	370.42±14.38	426.47±49.30	417.84±8.24	428.59±88.39
	Avg. CV ↓	0.60±0.41	6.42±0.85	8.67±1.45	3.28±2.86	10.66±11.81
	Avg. R-G ↑	670.61±14.18	364.5±14.51	416.83±47.46	413.28±9.04	434.59±83.89
	Avg. CV-G ↓	0.73±0.36	5.64±0.92	7.74±1.36	3.33±3.16	12.01±10.08
CCPO w/o CVI	Avg. R ↑	641.33±40.14	387.31±5.76	520.70±42.18	386.81±39.44	465.80±31.78
	Avg. CV ↓	1.44±0.72	1.66±0.79	2.36±2.67	0.81±0.76	3.51±0.93
	Avg. R-G ↑	623.17±41.42	383.24±6.30	519.05±36.31	388.69±35.35	465.36±32.20
	Avg. CV-G ↓	1.78±0.70	2.17±1.09	2.73±3.03	1.15±1.08	3.96±1.01

Conclusion

- **Versatile Safe RL Framework:** Frames safe Reinforcement Learning as a generalized problem beyond fixed thresholds, addressing the limitations of traditional constrained optimization.
- **Zero-Shot Adaptation:** Introduces **CCPO**, a novel approach based on conditional variational inference that generalizes to unseen constraint thresholds without requiring policy retraining.
- **Core Technical Innovations:**
  - Developed two key techniques: **Value Variation Estimation (VVE)** and **Conditional Variational Inference (CVI)**.
  - Provides theoretical guarantees regarding data efficiency and safety.
- **Empirical Superiority:** Outperforms baselines in safety and task performance, particularly in **high-dimensional** state and action spaces where traditional methods fail to adapt.