SafeTAC: Safe Tsallis Actor-Critic Reinforcement Learning for Safer Exploration

Dohyeong Kim*, Jaeseok Heo*, and Songhwai Oh

***Abstract*— Satisfying safety constraints is the top priority in safe reinforcement learning (RL). However, without proper exploration, an overly conservative policy such as freezing at the same position can be generated. To this end, we utilize maximum entropy RL methods for exploration. In particular, an RL method with Tsallis entropy maximization, called Tsallis actor-critic (TAC), is used to synthesize policies which can explore with more promising actions. In this paper, we propose a Tsallis entropy-regularized safe RL method for safer exploration, called SafeTAC. For more expressiveness, we extend the TAC to use a Gaussian mixture model policy, which improves the safety performance. To stabilize the training process, the retrace estimators for safety critics are formulated, and a safe policy update rule using a trust region method is proposed.**

I. INTRODUCTION

As the number of successful cases of controlling complex robots using reinforcement learning (RL) increases [1]–[3], safety-aware RL, denoted as *safe RL*, is actively investigated for more diverse robotic applications. Safety can be dealt with in a primal-dual approach by explicitly modeling constraints [4]–[7] or can be handled implicitly by adding a penalty to the reward [8]. In particular, safe RL methods based on the primal-dual approach with chance constraints or risk measure constraints can produce a policy with near-zero constraint violations [5]–[7]. Furthermore, not only research for theory but also applications to robots such as legged robots [9] and manipulators [10] are being emerged.

Since safe RL has safety constraints, unlike traditional RL, there are two major issues to be resolved: 1) how to compute the constraints and 2) how to update policies to satisfy the constraints. First, since the exact computation of constraints is computationally prohibited, safety critics are widely used to estimate the constraints, and the Bellman operator can calculate targets for the critics [11]. However, as the constraint consists of constant values, a bias in the estimation can confuse the direction of the policy update. Thus, it is more preferable to use the TD-Lambda for the targets rather than TD to reduce the bias [12]. Second, there are several methods for safe policy updates, which update policies while satisfying the constraints. The most popular method is the Lagrangian method, which solves a Lagrange dual problem and updates the policy and Lagrange multipliers simultaneously. However, the multipliers can make the training process fluctuate or destabilize [13]. Stooke *et al.* [13] have proposed a novel PID-based Lagrangian method to alleviate the fluctuation but it is sensitive to hyperparameters [8]. In contrast, trust-region methods can achieve great performance without instability in training because they do not need to update Lagrange multipliers [4], [7]. However, the existing trust region-based methods only apply to Gaussian policies, so there is a need for a trust-region approach for non-Gaussian policies.

It is also important to develop a proper exploration strategy to generate a high-return policy while satisfying safety. For example, if focusing only on safety without proper exploration, excessively conservative policies such as freezing at a single location can be generated. For this reason, we focus on the entropy-regularized safe RL problem, since it is known that maximum entropy RL methods, such as soft-actor critic (SAC) [14], can create high-return policies with its outstanding exploration strategies. Yang *et al.* [6] have proposed a SAC-based safe RL method, called worst-case soft actor-critic (WCSAC). By constraining conditional value at risk (CVaR) of safety signals, WCSAC enables to synthesize a safer policy compared to methods using expectation-based constraints, and at the same time effectively solves the maximum entropy RL problem using the SAC method. However, Shannon-Gibbs (SG) entropy-based regularization, used in SAC and WCSAC, is not qualified for safe RL. Since the optimal policy of SAC is a stochastic policy which can sample unnecessary actions [15], SAC can cause agents to reach an irreversible situation due to such actions.

In this paper, we propose a safe RL method with entropy maximization. We use the CVaR constraints to better reflect safety, but use the Tsallis entropy for regularization. Lee *et al.* [15] have proposed the maximum Tsallis entropy RL method, called Tsallis actor-critic (TAC). As they have shown that the optimal policy can efficiently explore without sampling unnecessary actions, the Tsallis entropy-based regularization is highly suitable for safe exploration. However, since the optimal policy of TAC is not usually represented by a single Gaussian distribution, we extend the TAC by modeling the policy as a mixture of Gaussians. Next, we have developed a trust region method for TAC to update policies reflecting the safety constraints. As there is no closed term of the KL divergence between Gaussian mixture models (GMMs), we build a trust region constraint using an approximation proposed by Hershey *et al.* [16]. Approximating the objective function of TAC and the CVaR constraints within the trust region, search directions are calculated, and policies are safely updated using a line search method as done in [4], [7]. Finally, to train safety critics with low bias, we formulate

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01190, [SW Star Lab] Robot Learning: Efficient, Safe, and Socially-Acceptable Machine Learning, 50%) and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01309, Development of AI Technology for Guidance of a Mobile Robot to its Goal with Uncertain Maps in Indoor/Outdoor Environments, 50%). *(Corresponding authors: Songhwai Oh.)*

*Equal contribution.

D. Kim, J. Heo, and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: dohyeong.kim@rllab.snu.ac.kr, jaeseok.heo@rllab.snu.ac.kr, songhwai@snu.ac.kr).

targets for the critics using retrace estimators [17], which can efficiently estimate TD-Lambda using off-policy data. By accurately estimating constraints through the safety critics trained with low bias, we can avoid synthesizing policies that are overly conservative or violate safety.

The main contributions are threefold. First, we propose a novel Tsallis entropy-regularized safe RL method, called *SafeTAC*. By adjusting the entropic index of the Tsallis entropy, SafeTAC can generate various exploration strategies for safe RL. Second, we efficiently stabilize the training process using the proposed trust region method. Last, SafeTAC is evaluated through safe navigation tasks in both simulation and real environments and achieves the near-zero number of constraint violations in all experiments.

II. RELATED WORK

A survey done by García *et al.* [18] on safe RL has classified safe RL into two main groups, *optimization criterion* and *exploration process*. Methods that focus on the *exploration process* [5], [10] correct actions locally by changing action on whether the action is safe or not. Bharadhwaj *et al.* [5] and Thananjeyan *et al.* [10] have used a pre-trained safety-critic to determine if a sampled action is safe, and used rejection sampling [5] or a recovery policy [10] to guide the agent to a safe region. *Optimization criterion* based methods [4], [6]–[8], [13], [19], [20] focus on optimizing a policy within a given constraint. Among the methods, Lagrangian based methods [6], [13], [19] are widely used. These methods are easy to implement and have shown to work well with real robots [19]. Due to the nature of the Lagrangian approach, which updates the policy and multipliers concurrently, the multipliers can oscillate as training progresses. Stooke *et al.* introduced a PID Lagrangian method [13] to overcome this oscillation issue. On the other hand, Achiam *et al.* [4] introduced a method that searches for the optimal policy within a trust region [21]. This method theoretically shows a monotonic improvement of the objective functions, but compared to the Lagrangian based methods, it shows high number of constraint violations, as shown in [22]. Kim and Oh [7] extended the works of [4] by using a distributional safety critic with CVaR constraints.

Recently, off-policy safe RL methods have gained attention as the importance of sample efficiency is increasing. Bharadhwaj *et al.* [5] have extended the conservative Q-learning [23] to bound the expected chance of failures. Ha *et al.* [19] have extended SAC [14] with a Lagrange multiplier to achieve advanced performance and efficient exploration while satisfying constraints. Yang *et al.* used a CVaR-based constraint with an entropy reward term to enhance safety even in the worst case scenarios while exploring efficiently.

III. BACKGROUND

A. Constrained Markov Decision Process

Safe RL problems can be modeled by a constrained Markov decision process (CMDP), an MDP with cost functions. A CMDP can be defined as a set of: a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition model $\mathcal{P}$, an initial state distribution ρ, a discount factor γ, a reward function R, and a cost function C. The transition model, reward, and cost are functions that maps $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$ to $\mathbb{R}$. In a CMDP setting, a policy $\pi(\cdot|s)$ can be defined as a probability distribution of an action given a state s, and a cumulative cost sum is defined as $C_\pi(s,a) := \sum_{t=0}^{\infty} C(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a$, where s_t, a_t, s_{t+1} are sampled by π and $\mathcal{P}$. Then, a safe RL problem can be formulated as follows:

$$\begin{aligned} &\underset{\pi}{\text{maximize}} \ \mathop{\mathbb{E}}_{\rho,\pi,\mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \right] \\ &\text{s.t. } \mathcal{M} \left(\sum_{t=0}^{\infty} \gamma^t C(s_t, a_t, s_{t+1}) \right) \leq d, \end{aligned} \tag{1}$$

where d is a limit value, and $\mathcal{M}$ is a safety measure such as an expectation or conditional value at risk (CVaR). Since the safe RL problem aims to maximize the return while satisfying the constraint as shown in (1), optimization techniques such as Lagrangian methods or trust-region methods are required to solve the problem.

B. Maximum Tsallis Entropy Reinforcement Learning

Lee *et al.* [15] have proposed an actor-critic RL method maximizing Tsallis entropy, called TAC. Tsallis entropy is a generalized entropy and is defined as:

$$S_q(P) := \mathop{\mathbb{E}}_{X \sim P} \left[-\ln_q(P(X)) \right], \tag{2}$$

where q is an entropic index, P is a distribution, and $\ln_q$ is q-logarithm [15]. Tsallis entropy becomes identical to SG entropy when $q = 1$. The objective of TAC is written as:

$$\underset{\pi}{\text{maximize}} \ \mathop{\mathbb{E}}_{\rho,\pi,\mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t, s_{t+1}) + \beta S_q(\pi(\cdot|s_t)) \right) \right], \tag{3}$$

where β is an entropy weight. To handle (3), new variants of reward-critic are introduced by Lee *et al.* [15].

$$\begin{aligned} V_q^\pi(s) &:= \mathop{\mathbb{E}}_{\pi,\mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t \left(r_t + \beta S_q(\pi(\cdot|s_t)) \right) \middle| s_0 = s \right], \\ Q_q^\pi(s,a) &:= \mathbb{E} \left[R(s,a,s') + \gamma V_q^\pi(s') | s' \sim \mathcal{P}(\cdot|s,a) \right], \end{aligned} \tag{4}$$

where $r_t = R(s_t, a_t, s_{t+1})$. By adjusting the entropic index, TAC makes different exploration-exploitation strategies [15], which can be seen that TAC is suitable for safe RL.

C. Worst-Case Soft Actor-Critic

To solve a safe RL problem with entropy maximization, Yang *et al.* [6] have proposed a SAC-based safe RL method, called worst-case soft actor-critic (WCSAC). For the safety measure in (1), WCSAC uses the conditional value at risk (CVaR), which is written as:

$$\mathrm{CVaR}_\alpha(X) := \mathbb{E} \left[X | X > \mathrm{CDF}^{-1}(1-\alpha) \right], \tag{5}$$

where α is a risk level, and CDF is the cumulative density function of the random variable X. CVaR is calculated as an expectation above a certain level. Therefore, it can focus on the worst case, and WCSAC obtains a safe policy by constraining the CVaR. WCSAC improved the expectation-constrained SAC method proposed by Ha *et al.* [19], as CVaR constraints can effectively prevent incurring excessive costs [7]. To deal with the CVaR, safety-critic and safety-variance functions are defined in [6] as below.

$$\begin{aligned} Q_C^\pi(s,a) &:= \mathop{\mathbb{E}}_{\pi,\mathcal{P}} \left[C_\pi(s_0,a_0) | s_0 = s, a_0 = a \right], \\ V_C^\pi(s,a) &:= \mathop{\mathbb{E}}_{\pi,\mathcal{P}} \left[C_\pi(s_0,a_0)^2 | s_0 = s, a_0 = a \right] - Q_C^\pi(s,a)^2. \end{aligned} \tag{6}$$

With an assumption that the cumulative cost sum $C_\pi(s,a)$ follows a Gaussian distribution, the CVaR of $C_\pi(s,a)$ can be approximated as [6]:

$$\begin{aligned}\mathrm{CVaR}_\alpha(C_\pi(s,a)) &\approx Q_C^\pi(s,a) + \alpha^{-1}\phi(\Phi^{-1}(\alpha))\sqrt{V_C^\pi(s,a)} \\ &=: \Gamma_\pi(s,a,\alpha),\end{aligned} \tag{7}$$

where ϕ and Φ are probability and cumulative density functions for the standard normal distribution, respectively, and the approximated CVaR is denoted by Γ_π. Then, the safety constraint can be defined as follows:

$$\mathop{\mathbb{E}}_{\substack{s\sim D \\ a\sim\pi(\cdot|s)}} [\Gamma_\pi(s,a,\alpha)] \leq d, \tag{8}$$

where d is a limit value, and D is a replay buffer. Finally, WCSAC addresses the maximum entropy RL problem with the constraint (8) by solving the following dual problem.

$$\max_{\kappa\geq 0}\min_{\pi} \mathop{\mathbb{E}}_{\pi,D}\left[-Q_{q=1}^\pi(s,a) + \kappa\left(\Gamma_\pi(s,a,\alpha) - d\right)\right], \tag{9}$$

where κ is a Lagrange multiplier.

IV. PROPOSED METHOD

The maximum Tsallis entropy RL method can effectively prevent sampling an undesirable action (see [15] for more details), which means it is more suitable for safe RL than the SAC method. Thus, we aim to maximize the objective of TAC while satisfying the constraint (8), and the constrained problem for policy improvements can be expressed as:

$$\begin{aligned}\underset{\pi'}{\text{maximize}} \mathop{\mathbb{E}}_{\pi',D}\left[Q_q^\pi(s,a) - \beta\ln_q\left(\pi'(a|s)\right)\right] \\ \mathbf{s.t.} \mathop{\mathbb{E}}_{\pi',D}\left[\Gamma_\pi(s,a,\alpha)\right] \leq d.\end{aligned} \tag{10}$$

To solve constrained optimization problems, Lagrangian methods are widely used in safe RL methods including WCSAC, but Lagrange multipliers can oscillate, so other techniques, such as PID Lagrangian, are required [13]. Therefore, for the policy improvement, we adopt trust region methods, which do not require updating the multipliers. In the remainder of this section, we first introduce how to model a policy using a Gaussian mixture model (GMM). Next, the safe policy update rule using the trust region is described, and finally, we formulate retrace estimators [17] for targets of the safety-critic and variance to estimate Γ_π with low bias.

A. Policy Modeling Using Gaussian Mixture Model

We divide the action sampling process into two steps to ensure that sampled actions are bounded as in SAC [14]. First, an unbounded action is sampled from an unbounded policy: $u \sim \eta(\cdot|s)$, and the action is transformed using the *tanh* function to obtain a bounded action: $a = \tanh(u)$. Then we model the unbounded policy $\eta(\cdot|s)$ using a Gaussian mixture model (GMM) as follows:

$$\begin{aligned}\eta(u|s) &= \sum_{i=1}^K w_i(s)\mathcal{N}(u;\mu_i(s),\Sigma_i(s)), \\ \pi(a|s) &= \eta(u|s)\Big(\prod_{i=1}^{D_\mathcal{A}}(1-\tanh^2(u_i))\Big)^{-1},\end{aligned} \tag{11}$$

where $a = \tanh(u)$, K is the number of the Gaussians, w_i is the ith weight calculated by the *softmax* function, μ_i and Σ_i are the mean and variance of the ith Gaussian, and $D_\mathcal{A}$ is the dimension of action space. However, reparameterization tricks for action sampling are required to calculate a policy gradient of the TAC objective. To sample the weights w, we use the Gumbel-Softmax (GS) [24] as: $\overline{w} = \mathrm{GS}(\epsilon_G|g(s))$, ϵ_G is a noise vector sampled from the standard Gumbel distribution, and g is a logit function. Then, actions are sampled as follows:

$$\begin{aligned}a &= \tanh\left(\sum_{i=1}^K \overline{w}_i\left(\mu_i(s) + \epsilon_N^T\Sigma_i(s)\right)\right) \\ &=: f(\epsilon_G,\epsilon_N|s),\end{aligned} \tag{12}$$

where ϵ_N is sampled from the standard normal distribution, and the sampling function is denoted as f. Through the function f, the policy gradient can be calculated by sampling the noise vectors ϵ_N and ϵ_G.

B. Safe Policy Update Using Trust Region

We use the trust region method to solve the constrained problem (10). By adding a trust region constraint, the subproblem for the safe policy update can be rewritten as follows:

$$\begin{aligned}\underset{\pi'}{\text{maximize}} \mathop{\mathbb{E}}_{\pi',D}\left[Q_q^\pi(s,a) - \beta\ln_q\left(\pi'(a|s)\right)\right] \\ \mathbf{s.t.} \mathop{\mathbb{E}}_{\pi',D}\left[\Gamma_\pi(s,a,\alpha)\right] \leq d, \\ \mathbb{E}_D\left[D_{\mathrm{KL}}(\eta(\cdot|s)||\eta'(\cdot|s))\right] \leq \delta,\end{aligned} \tag{13}$$

where D_{KL} is the Kullback-Leibler (KL) divergence, δ is a size of the trust region, and η and η' are unbounded policies before and after being updated, respectively. Since the objective and constraints in (13) are nonlinear, we approximate (13) to find a search direction and update the policy using a line search method as in [4], [7]. We first parameterize the policy, reward-critic, safety-critic, and safety-variance with ϕ, θ, ψ_C, and ψ_V, respectively. Then, the gradients of the objective and safety constraint can be expressed as follows:

$$\begin{aligned}g = \mathop{\mathbb{E}}_{s\sim D}\Big[&\left(\nabla_a Q_q^\pi(s,a;\theta) - \nabla_a\ln_q\pi(a|s;\phi)\right)\nabla_\phi f(\epsilon_G,\epsilon_N|s;\phi) \\ &-\nabla_\phi\ln_q\pi(a|s;\phi) \mid a = f(\epsilon_G,\epsilon_N|s;\phi), \epsilon_N\sim\mathcal{N}, \epsilon_G\sim\mathcal{G}\Big], \\ b = \mathop{\mathbb{E}}_{s\sim D}\Big[&\nabla_a\Gamma(s,a,\alpha;\psi_C,\psi_V)^T\nabla_\phi f(\epsilon_G,\epsilon_N|s;\phi) \\ &|a = f(\epsilon_G,\epsilon_N|s;\phi), \epsilon_N\sim\mathcal{N}, \epsilon_G\sim\mathcal{G}\Big],\end{aligned} \tag{14}$$

where g and b are gradients of the objective and safety constraint, respectively, $\mathcal{N}$ is the standard normal distribution, and $\mathcal{G}$ is the standard Gumbel distribution. To approximate the trust region constraint, Hessian of the KL divergence is required, but there is no closed form of the KL divergence between GMMs. Thus, we use the following approximation proposed by Hershey *et al.* [16].

$$\begin{aligned}&\overline{\mu}(s;\phi) := \sum_{i=1}^K w_i(s;\phi)\mu_i(s;\phi), \\ &\overline{\Sigma}(s;\phi) := \sum_{i=1}^K w_i(s;\phi)\big(\Sigma(s;\phi) \\ &\qquad + (\mu_i(s;\phi) - \overline{\mu}(s;\phi))(\mu_i(s;\phi) - \overline{\mu}(s;\phi))^T\big), \\ &D_{\mathrm{KL}}^{\mathrm{approx}}(\eta(\cdot|s,\phi_{\mathrm{old}})||\eta(\cdot|s,\phi)) := \\ &\qquad D_{\mathrm{KL}}(\mathcal{N}(\overline{\mu}(s;\phi_{\mathrm{old}}),\overline{\Sigma}(s;\phi_{\mathrm{old}}))||\mathcal{N}(\overline{\mu}(s;\phi),\overline{\Sigma}(s;\phi))).\end{aligned} \tag{15}$$

In [16], when the true distance between two GMMs is close, the approximation is shown to be accurate. Hence, the Hessian matrix can be estimated as follows:

$$H = \frac{\partial^2}{\partial \phi^2} \mathop{\mathbb{E}}_{s\sim D} [D_{\mathrm{KL}}^{\mathrm{approx}}(\eta(\cdot|s;\phi_{\mathrm{old}})||\eta(\cdot|s;\phi))] \big|_{\phi=\phi_{\mathrm{old}}}. \quad (16)$$

Finally, the nonlinear problem (13) can be relaxed by linear approximation of the objective and safety constraint and quadratic approximation of the trust region constraint as follows:

$$\begin{aligned} \phi_{\mathrm{new}} = \mathop{\mathrm{argmax}}_{\phi}\ & g^T(\phi - \phi_{\mathrm{old}}) \\ \mathbf{s.t.}\ & \frac{1}{2}(\phi - \phi_{\mathrm{old}})^T H(\phi - \phi_{\mathrm{old}}) \le \delta, \\ & \mathop{\mathbb{E}}_{\substack{s\sim D \\ a\sim\pi(\cdot|s;\phi_{\mathrm{old}})}} [\Gamma_\pi(s,a,\alpha;\psi_C,\psi_V)] + b^T(\phi - \phi_{\mathrm{old}}) \le d. \end{aligned} \quad (17)$$

(17) can be solved using a linear and quadratic constrained linear programming (LQCLP) solver. However, since (17) is obtained through approximation, the solution of (17) is not used directly for policy updates but as a search direction. After the search direction is calculated, the policy is updated using a backtracking line search method. If the constrained problem (17) is infeasible, recovery steps are performed by minimizing the safety constraint within the trust region as described in Section 4.2 of [4].

C. Critic Update Using Retrace Estimators

SAC-based methods generally use temporal difference (TD) targets for critic updates, but TD targets induce huge biases [12]. As Γ_π is constrained by a constant d in (13), Γ_π should be estimated with low bias. Thus, we decide to use the retrace estimator [17], which can trade off bias and variance of target values by adjusting a trace-decay value. We extend the result of [25] from state value functions to action value functions as follows:

$$\begin{aligned} Q_{C,t}^{\mathrm{ret}} = {} & c_t + \gamma Q_C^\pi(s_{t+1}, a'_{t+1}) \\ & + \gamma\lambda\overline{\rho}_{t+1}(Q_{C,t+1}^{\mathrm{ret}} - Q_C^\pi(s_{t+1}, a_{t+1})), \\ V_{C,t}^{\mathrm{ret}} = {} & \left(c_t + \gamma Q_C^\pi(s_{t+1}, a'_{t+1})\right)^2 \\ & + \gamma^2 V_C^\pi(s_{t+1}, a'_{t+1}) - Q_C^\pi(s_t, a_t)^2 \\ & + \gamma^2\lambda\overline{\rho}_{t+1}(V_{C,t+1}^{\mathrm{ret}} - V_C^\pi(s_{t+1}, a_{t+1})), \end{aligned} \quad (18)$$

where $\overline{\rho}_t = \min(1, \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)})$, μ is a behavioral policy, $c_t = C(s_t, a_t, s_{t+1})$, $a'_t \sim \pi(\cdot|s_t)$, and λ is a trace-decay value. The safety-critic and variance are updated by minimizing differences between the target values from (18). Finally, the proposed method is summarized in Algorithm 1.

Algorithm 1 SafeTAC

Input: Policy $\pi(a|s;\phi)$, reward-critic $Q_q^\pi(s,a;\theta)$, safety-critic $Q_C^\pi(s,a;\psi_C)$, safety-variance $V_C^\pi(s,a;\psi_V)$, and replay buffer D.

Initialize the parameters ϕ, θ, ψ_C, and ψ_V.
Initialize the replay buffer D and the environment.
for epochs=1, P **do**
  for t=1, N **do**
    Sample an action $a_t \sim \pi(\cdot|s_t;\theta)$.
    Get a next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$, a reward $r_t = R(s_t, a_t, s_{t+1})$, and a cost $c_t = C(s_t, a_t, s_{t+1})$.
    Store $(s_t, a_t, \pi(a_t|s_t;\phi), r_t, c_t, s_{t+1})$ in D.
  end for
  Sample trajectories T from D.
  Calculate g, b, and H in (14) with T.
  Calculate the search direction by solving (17) and update ϕ using a line search method.
  Calculate target values for safety-critic and variance in (18) using T and update ψ_C and ψ_V.
  Calculate reward targets from (4) using T and update θ.
end for

V. Experiments

In this section, we evaluate SafeTAC, using several experiments, including sim-to-real experiments. In addition, subsequent ablation studies are conducted to show the effectiveness of the proposed method: **1)** experiments with different numbers of Gaussian mixtures, **2)** comparison between the Lagrangian and trust-region methods for the policy update, and **3)** comparison between the TD targets and retrace targets for safety critics. In the remainder of this section, the experimental setup for the simulation and sim-to-real are presented, and training details and baseline methods are introduced. The results of the experiments are discussed in Section VI.

A. Simulation Experimental Setup

We conduct experiments in the Safety Gym [22] which is a commonly used benchmark for safe RL methods. We have experimented with the *goal* task where robots need to reach a point goal while avoiding hazardous regions. Three different types of robots have been used for the experiments: *Point*, *Car*, and *Doggo*. The LiDAR sensor, velocities, and relative goal location are part of the 16-dimensional state space, and the action space has two-dimensional. The action of the Point and Car tasks are for turning and moving the agent forward/backward. Since the Doggo task is challenging to train directly using safe RL algorithms, we initially trained a low-level controller that moves the agent towards a target without taking dangerous areas into account. Then, we train RL agents in a reformed environment where the action is a target for the low-level controller, as done in [7]. The reward and cost functions are defined as follows for all tasks:

$$\begin{aligned} R(s,a,s') &= d_g(s) - d_g(s') + I_{\{x|x\le\Delta\}}(d_g(s)) \\ C(s,a,s') &= \mathrm{sigmoid}(k\cdot(r_h - d_h(s))), \end{aligned} \quad (19)$$

where d_g is the distance between the goal and the robot, I is an indicator, d_h is the minimum distance between the robot and the hazardous regions, Δ is a threshold ($= 0.3\ m$), r_h is the hazard region size ($= 0.2\ m$), and k is a constant ($= 10$).

B. Sim-to-Real Experimental Setup

The goal of the sim-to-real robot experiment is to navigate the Jackal robot [26], shown in Figure 1a, to a given goal position while avoiding obstacles. We restrict the robot from moving backward because the LiDAR of the robot does not cover the full range. We first train agents in a simulator based on the MuJoCo physics engine [27]. The simulator terminates the episode and give additional costs equivalent to the remaining steps of the episode once the agent collides with obstacles. The state space is the same as the simulation setup except for the linear acceleration. The action space consists of linear velocity and angular velocity, and the reward and cost functions are the same as the simulation. Once trained in a simulator, the agent is evaluated in the real world shown in Figure 1b.

(a) Jackal robot (b) Real-world environment

Fig. 1: The jackal robot and the real-world environment are presented. Using the jackal robot, agents are evaluated to navigate a given goal, avoiding obstacles in the environment. Goals are spawned in sequence (marked in red).

C. Baselines

SafeTAC is compared with three different safe RL methods. For trust-region based safe RL methods, constrained policy optimization (CPO) [4], and trust-region conditional value at risk (TRC) [7] are used. CPO uses expectations while TRC uses CVaR for the safety measure. For comparison with SAC based methods, WCSAC [6] is used.

To measure performance, the number of entering hazardous regions are counted, called constraint violation (CV). we use the same metric defined in [7] for the score, which is expressed as: $\sum_{t=0}^{T-1} R_t/(1+\sum_{t=0}^{T-1} CV_t)$. The reward sum is typically large for safe RL algorithms that ignore constraints, so dividing the sum by CV can produce fair results.

VI. Results

A. Safety Gym and Jackal Simulation

The average final scores, average final CVs, and the total number of CVs are presented in Table I. SafeTAC shows the best results in the Doggo and Car tasks and the second-best results in the Point and Jackal tasks. Also, SafeTAC is an off-policy RL method, which means it can learn from past experiences and enhance sample efficiency for better training. Thus, SafeTAC is able to find a more optimal policy compared to the baseline methods. However, due to the entropy regularization, SafeTAC can have a slightly dangerous policy with high numbers of CVs compared to methods without the entropy regularization term. This is shown in the Point and Jackal tasks, where SafeTAC has the slightly higher average number of CVs than TRC.

TRC performs worse than SafeTAC in the Car and Doggo tasks and slightly better in the Point and Jackal Tasks. Since TRC does not apply entropy regularization, it can achieve comparable performance while maintaining a safe policy and has shown the low average number of CVs. However, compared to SafeTAC, TRC tends to show lower performance as it is an on-policy RL method and has inefficient exploration compared to SafeTAC with an entropy regularization. WCSAC showed the lowest total CVs in all tasks but lower scores than SafeTAC. WCSAC appears to be overly conservative due to the overestimated Lagrange multiplier. Because of this, the policy seems to maintain safety constraints rather than maximizing reward sums. CPO shows lower scores and higher CVs than SafeTAC. Unlike SafeTAC, which considers the tail of the cost distribution, CPO only considers whether the expected discounted cost sum is within a certain threshold. Therefore, it is difficult for CPO to prevent failures.

In the Car and Jackal tasks, q value of 1.2 has shown the best performance, while in the Doggo task, q value of 1.4 has shown the best performance. Also, the q value of 1.2 has shown the lowest average number of CVs in the Point, Car, and Jackal tasks, while q value of 1.4 has shown the lowest average number of CVs in the Doggo task. It can be interpreted that when a new task is given, using the Tsallis entropy regularization with an appropriate q value can lower the chance of sampling unnecessary actions and enhance performance compared to the SG entropy regularization.

B. Real-World Evaluation

Results of the sim-to-real experiments are shown in Table II, and the demonstration video can be found in the attached material. SafeTAC shows the highest reward sum with low numbers of CVs with no failure. Due to the improved exploration by Tsallis entropy regularization, SafeTAC can find the policy with the highest return compared to other safe RL methods. The agent trained by TRC moves slowly near obstacles, but the agent trained by SafeTAC moves consistently fast, even near hazardous areas, as seen in the video. WCSAC shows the lowest value not only in the number of CVs but also in the reward sum, which means that WCSAC generates an overly conservative policy as in the simulation tasks. Finally, CPO shows the highest number of failures, which can be seen as the expectation-based constraint cannot effectively prevent failures.

C. Ablation Study

1) Gaussian Mixture Models: To show the effect of the GMM policy, we experiment with different numbers of Gaussian mixtures in the Point task with $q = 1.2$. Table III shows that higher the number of Gaussian mixtures K results in higher scores and lower numbers of CVs. From the results, it can be seen that the GMM further improves safety performance because it is more expressive than a single Gaussian and is more suitable for the Tsallis entropy maximization problem.

2) Safety Critic Targets: The second row of Table III shows the results of training SafeTAC with the TD targets ($\lambda = 0$) and the retrace targets ($\lambda = 0.97$). From the result, using the proposed retrace estimators for the targets is exceptionally effective in improving scores and lowering the number of CVs. Also, the result implies that training the safety critic and variance with low bias is critical because the safety constraints consist of constants.

3) Policy Update Rule: Finally, we compare two policy update rules: the Lagrangian method and the trust-region method. To implement the Lagrangian method, the multiplier and the policy are updated using a gradient descent optimizer, and the other settings, such as the retrace targets, are the same as SafeTAC. The result is shown in the last row of Table III. The Lagrangian method shows lower numbers of CVs than the trust-region method but also shows near zero scores. From the result, we can conclude that it is difficult for the Lagrangian method to balance constraint satisfaction and return maximization, but the trust-region method can be used to overcome the issue.

Through the ablation experiments, it is shown that the proposed method is highly effective for safe RL problems. In particular, performance degradation occurs without the

	q	Score ↑				CV ↓				Total CVs [10^5] ↓			
		Point	Car	Doggo	Jackal	Point	Car	Doggo	Jackal	Point	Car	Doggo	Jackal
SafeTAC (Ours)	1.4	8.4	14.5	**13.9**	8.02	0.009	0.005	0.001	0.040	1.31	0.80	0.25	2.47
	1.2	10.4	18.2	13.6	8.22	0.007	**0.002**	0.002	0.022	1.16	0.61	0.30	2.23
	1.0	10.1	**18.7**	13.1	4.49	0.008	0.002	0.002	0.066	1.19	0.59	0.35	3.85
TRC	-	**13.0**	15.9	11.0	**9.24**	**0.003**	0.002	0.004	0.007	0.83	0.65	0.45	2.01
WCSAC	-	-0.2	5.4	10.5	2.86	0.015	0.003	**0.001**	**0.004**	**0.77**	**0.33**	**0.10**	**0.96**
CPO	-	8.0	12.5	6.0	6.81	0.021	0.011	0.020	0.072	2.67	2.0	1.74	5.23

TABLE I: The final score, final numbers of CVs, and the total number of CVs in the simulated environments. Each method is trained five times with different seeds, and note that the unit used in the column of the total CVs is 10^5.

	Reward Sum ↑	CV ↓	# of failures ↓
SafeTAC (Ours)	**21.403**	20.7	0/10
TRC	10.217	10.0	0/10
WCSAC	1.900	**0.0**	0/10
CPO	7.381	616.1	8/10

TABLE II: Evaluation results of the real-world Jackal environments. The results are obtained by running 10 episodes for each method. It is considered a failure if the episode ends due to a collision with the robot and obstacles.

		Score ↑	CV ↓	Total CVs [10^5] ↓
GMM	K = 1	10.45	0.007	1.16
	K = 3	10.57	**0.002**	0.89
	K = 5	**11.78**	0.005	**0.83**
Retrace	λ = 0.97	**10.11**	**0.008**	**1.19**
	$\lambda = 0$	-0.02	0.026	1.80
Policy Update	Trust-Region	**10.11**	0.008	1.19
	Lagrangian	-0.15	**0.007**	**1.18**

TABLE III: Ablation results. All experiments are performed in the Point task. For the GMM, we use three different numbers of Gaussian mixtures and set the entropic index value q to 1.2. For the retrace targets and policy update ablations, q is set to 1.

proposed retrace estimators for the critic targets and the trust-region method for policy update. Also, SafeTAC with GMM policy increases the score while reducing the total number of CVs, which means that the Tsallis entropy regularization fits the GMM policy well and enables safer exploration.

VII. CONCLUSIONS

For safer exploration, we have proposed a safe RL method with Tsallis entropy maximization, SafeTAC. To express the optimal policy of the TAC, we extend the TAC to use a GMM policy through reparameterization tricks. By approximating the KL divergence between GMMs, the safe policy update rule based on the trust region method has been derived. SafeTAC shows higher or similar performance than other state-of-the-art safe RL methods in the simulation and sim-to-real experiments. Finally, we shows that the proposed retrace estimators and trust-region method improves the stability of the training process.

REFERENCES

[1] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science Robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[2] J. Siekmann, K. Green, J. Warila, A. Fern, and J. Hurst, "Blind bipedal stair traversal via sim-to-real reinforcement learning," in *Proc. Robotics: Science and Systems*, 2021.

[3] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[4] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. International Conference on Machine Learning*, 2017, pp. 22–31.

[5] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, "Conservative safety critics for exploration," in *Proc. International Conference on Learning Representations*, 2021.

[6] Q. Yang, T. D. Simão, S. H. Tindemans, and M. T. Spaan, "WCSAC: Worst-case soft actor critic for safety-constrained reinforcement learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2021, pp. 10 639–10 646.

[7] D. Kim and S. Oh, "TRC: Trust region conditional value at risk for safe reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2621–2628, 2022.

[8] T. Xu, Y. Liang, and G. Lan, "CRPO: A new approach for safe reinforcement learning with convergence guarantee," in *Proc. International Conference on Machine Learning*, 2021, pp. 11 480–11 491.

[9] S. Gangapurwala, A. Mitchell, and I. Havoutis, "Guided constrained policy optimization for dynamic quadrupedal robot locomotion," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3642–3649, 2020.

[10] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery rl: Safe reinforcement learning with learned recovery zones," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915–4922, 2021.

[11] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[12] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[13] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by PID lagrangian methods," in *Proc. International Conference on Machine Learning*, 2020, pp. 9133–9143.

[14] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. International Conference on Machine Learning*, 2018, pp. 1861–1870.

[15] K. Lee, S. Kim, S. Lim, S. Choi, M. Hong, J. Kim, Y.-L. Park, and S. Oh, "Generalized Tsallis Entropy Reinforcement Learning and Its Application to Soft Mobile Robots," in *Proc. Robotics: Science and Systems*, 2020.

[16] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. IV–317–IV–320.

[17] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," in *Proc. International Conference on Learning Representations*, 2017.

[18] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[19] S. Ha, P. Xu, Z. Tan, S. Levine, and J. Tan, "Learning to walk in the real world with minimal human effort," in *Proc. Conference on Robot Learning*, 2021, pp. 1110–1120.

[20] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *Journal of Machine Learning Research*, vol. 18, 12 2015.

[21] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. International Conference on Machine Learning*, 2015, pp. 1889–1897.

[22] A. Ray, J. Achiam, and D. Amodei, "Benchmarking Safe Exploration in Deep Reinforcement Learning," 2019.

[23] Y. J. Ma, D. Jayaraman, and O. Bastani, "Conservative offline distributional reinforcement learning," in *Advances in Neural Information Processing Systems*, 2021.

[24] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *Proc. International Conference on Learning Representations*, 2017.

[25] D. Kim and S. Oh, "Efficient off-policy safe reinforcement learning using trust region conditional value at risk," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7644–7651, 2022.

[26] "Jackal ugv - small weatherproof robot - clearpath," Dec 2020. [Online]. Available: https://clearpathrobotics.com/jackal-small-unmanned-ground-vehicle/

[27] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.