Learning diverse skills for safe reinforcement learning

Haoyuan Cai
Princeton University

Chenxiao Tian
Princeton University

1 Introduction

Reinforcement learning has been studied widely in the industrial settings, ranging from safe robotics to autonomous driving. In order to deal with the uncertainty in the uncontrolled environments, we need to face the challenges of robustness and scalability. Adversarial reinforcement learning is one of the most common way to train a robust agent. We want the agent (controller) to keep away from the failure conditions under a bounded adversarial disturbance which represents the uncertainty of environments.

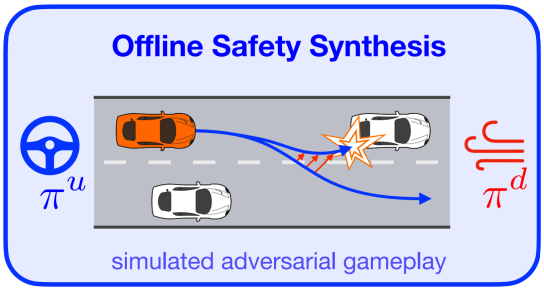

Figure 1: In adversarial reinforcement learning, we use game-theoretic methods to train a safe policy π^u and an adversary π^d iteratively. Figure taken from [Kai-Chieh Hsu]

However, during the training process, it's a common issue that the trained adversary fluctuates and overfits to the current controller. As a result, the training process becomes unstable, and the controller does not converge even the adversary has good exploration in the whole state space. For example, if there are two obstacles around the current position, it's possible that in the first 1000 episodes the adversary is trained to push the controller towards obstacle 1, and in the next 1000 episodes the adversary is to push the controller towards obstacle 2. This leads to the controller's overfitting to one of the adversary and having bad performance in the worst case. To overcome this issue, we hope that the adversary can frequently switching between pushing the controller between the two obstacles.

Our method is to train a latent-conditioned adversary. In each iteration, we randomly selects a latent, and feed the latent to the adversary. In this way, the adversary becomes diverse and not utilizable. This can reduce the possibilities of the controller's overfitting.

Related Work In this project, we mainly focus on safe reinforcement learning. Hamilton-Jacobi-Isaacs (HJI) theory models the safety problem in general nonlinear dyanmics as a two-play zero-sum game between the controller and an adversarial disturbance [**?**Mitchell et al., 2005]. The drawback of HJI methods is that they become computationally prohibitive when the state dimension is larger.

Recent works use neural networks to represent the control policy and the adversarial disturbance. Our algorithm is mainly based on ISAACS (iterative soft adversarial actor-critic for safety) [Hsu et al., 2023]. Similarly with other works in adversarial reinforcement learning, we discover that the trained controller tends to overfit the disturbance and have bad performances under some unseen

Preprint. Under review.

adversaries beyond the training process. From the simulation and comparision result, our method seems performing better than ISAAC method (Figure 2):

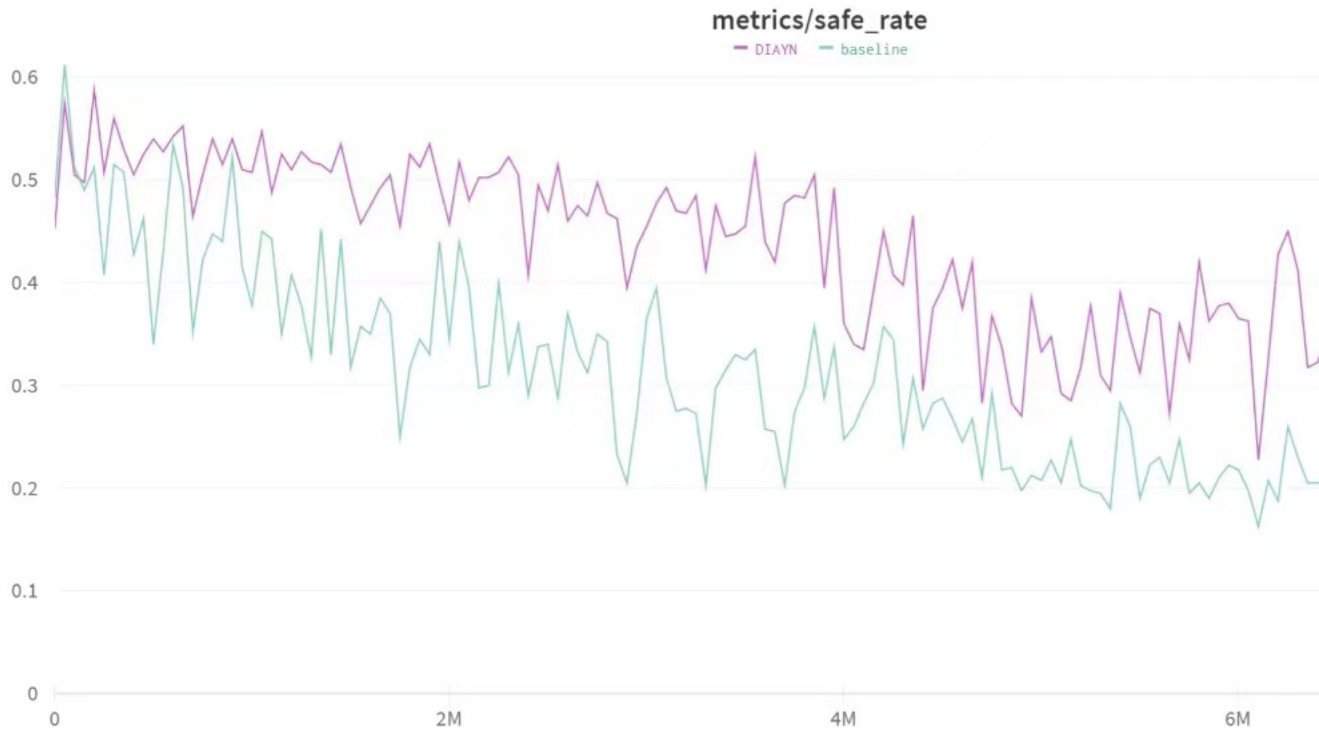

Figure 2: Our Method DIAYN vs ISAAC Method, where the metric of Y-axis refers to the success rate of the system or controller in maintaining safety standards during testing or simulation. For instance, in the context of autonomous driving, this could mean the proportion of successfully avoiding collisions or other safety violations.X-axis refers to lasting testing time minutes.It seems that DIAYN performs safer than the lower line method

In order to deal with this problem, we need to increase the diversity of the disturbance during the training process. One of the method is to train a latent-based disturbance, so that the disturbance can have multiple skills and become hard to utilize. Previous works on unsupervised discovery of skills introduce information-inspired critics to help the agent explore [Eysenbach et al., 2018, Sharma et al., 2019].

2 Preliminaries

We consider the discrete-time dynamics with a controller input u_t and a disturbance input d_t:

$$s_{t+1} = f(s_t, u_t, d_t).$$

Here t is the time step, and s_t is the state.

We define the failure set $\mathcal{F}$ as all the states that we need to prevent the system from entering.

Hamilton-Jacobi-Isaacs (HJI) reachability analysis models the safety problem as a two-player zero-sum game between the agent and the adversary. In HJI analysis, we introduce a Lipschitz continuous safety margin $g(s)$, where $g(s) > 0 \Leftrightarrow s \in \mathcal{F}$.

For any controller policy $\pi^u(u|s)$ and adversary $\pi^d(d|s)$, the value function of this game is characterized by a two-player Isaacs equation:

$$V^{safe}(s) = (1-\gamma)g(s) + \gamma \mathop{\mathbb{E}}_{\substack{u\sim\pi^u(\cdot|s),\\ d\sim\pi^d(\cdot|s)}} \max\{g(s), Q^{safe}(s,u,d)\}, \quad Q^{safe}(s,u,d) := V(f(s,u,d)).$$

It is similar with the Bellman equation, except that in the value function we have an additional max operator.

The objective of the controller is to minimize $V^{safe}(s)$, while the goal of the adversary is to maximize $V^{safe}(s)$.

Our method is to train a latent conditioned controller $\pi^u(u|s,z)$ and adversary $\pi^d(d|s,z)$. The latent z is sampled in the latent space $\mathcal{Z}$. To ensure that the disturbance depends on the latent z and avoid colliding into one single policy, we need to add a diversity regularizer in the loss function of the adversary. To realize this, we use previous works of unsupervised discovery of skills in reinforcement learning [Eysenbach et al., 2018].

3 Main Algorithm

Introducing Information-Inspired Critics Similarly with the algorithm DIAYN (Diversity Is All You Need) [Eysenbach et al., 2018], we denote the latent variable as $z \sim p(z)$, on which we condition our policy. The policy conditioned on a fixed variable z is a "skill". And we denote $I(\cdot,\cdot)$ and $\mathcal{H}(\cdot)$ as the mutual information and Shannon entropy, respectively.

The objective of the algorithm DIAYN is to maximize $I(S;Z) + \mathcal{H}(A|S) - I(A;Z|S)$. In intuition, this objective ensures that the skill depends on the state instead of the action. As we cannot integrate over all states and skills to compute this objective precisely, the algorithm DIAYN introduces a learned discriminator $q_\phi(z|s)$. And in each step, the reward function is defined as

$$r^{diayn}(s,z) = \log q_\phi(z|s) - \log p(z).$$

The objective of DIAYN is to maximize the cumulative reward of r^{diayn}. Similarly with DIAYN, we introduce an information-inspired critic $Q^{diayn}(s,u,d,z)$ to represent the cumulative reward of $r^{diayn}(s,z)$. Then, we can let train an adversary π^d which maximizes $Q^{safe} + \beta Q^{diayn}$ (β is some constant), so that the adversary contains multiple skills while maintaining its performance.

Now we introduce our main algorithm, Figure 3 is an intuitive illustration of ISAAC equation.

Figure 3: illustration of ISAAC equation

Now we focus on the loss functions of two actors π^u and π^d. Here the terms $\log \pi^u(u|s,z)$ and $\log \pi^d(d|s,z)$ are the entropy regularization in the soft actor-critic algorithm, so we mainly focus on the objective $Q^{safe} + \beta Q^{diayn}$.

While the disturbance maximizes the diversity Q^{diayn}, the controller minimizes the diversity Q^{diayn}. In the next section, we will prove that by adding the information-inspired critic, the training objective will be an evidence lower bound for the control policy.

4 Method and Theoretical Results: Lower Bound for Value Function $\mathbf{V}^{\pi^u,z^d}(s)$

Remark on the Method

As we know, adversarial Reinforcement learning has wide applications in learning safe control policies, while these policies lack safety guarantees and exhibit little robustness under new adversaries. The key reason is that when we train a controller and an adversary together to play a two-player zero-sum Markov game, the controller tends to overfit the adversary. So we develop this new approach to train a latent-based adversary by maximizing a mutual information inspired critic and also let the controller minimize this critic. Theoretically, we hope to prove when the adversary plays equilibrium policies, the training objective is a variational lower bound, closely related to the

Algorithm 1: Latent-conditioned ISAACS

Initialize: ctrl $\pi^u(\cdot|s_t, \tilde{z}_t)$, dstb $\pi^d(\cdot|s_t, z)$, encoder $q(z|s_t, u_{t-1}, s_{t-1}, \cdots)$, safety critic $Q^{safe}(s, u, d, z)$, DIAYN critic $Q^{diayn}(s, u, d, z)$, replay buffer $\mathcal{B} = \emptyset$.
for each episode **do**
  Sample $z \sim p(z)$, $s_0 \sim \mu(s_0)$.
  for $t = 0, 1, 2, \cdots$ **do**
    Sample $\tilde{z}_t \sim q(\cdot|s_t, u_{t-1}, s_{t-1}, \cdots)$.
    Ctrl takes action $u_t \sim \pi^u(\cdot|s_t, \tilde{z}_t)$.
    Dstb takes action $d_t \sim \pi^d(\cdot|s_t, z)$.
    Set $r_t^{diayn} = \log q(z|s_t, u_{t-1}, s_{t-1}, \cdots) - \log p(z)$.
    Get next state $s_{t+1} = f(s_t, u_t, d_t)$.
    Add $(z, s_t, \tilde{z}_t, u_t, d_t)$ to $\mathcal{B}$.
  end for
  Update $\pi^u, \pi^d, q, Q^{safe}, Q^{diayn}$:

$$\pi_u : \min_{\pi_u} \mathop{\mathbb{E}}_{\substack{(z,s,\tilde{z},d)\sim\mathcal{B},\\ u\sim\pi^u(\cdot|s,\tilde{z})}} [Q^{safe}(s, u, d, z) + \beta Q^{diayn}(s, u, d, z) + \alpha_u \log \pi^u(u|s, z)].$$

$$\pi_d : \max_{\pi_d} \mathop{\mathbb{E}}_{\substack{(z,s,u)\sim\mathcal{B},\\ d\sim\pi^d(\cdot|s,z)}} [Q^{safe}(s, u, d, z) + \beta Q^{diayn}(s, u, d, z) - \alpha_d \log \pi^d(d|s, z)].$$

$$q : \max \mathop{\mathbb{E}}_{(z,s_t,u_{t-1},s_{t-1},\cdots)\sim\mathcal{B}} [\log q(z|s_t, u_{t-1}, s_{t-1}, \cdots)].$$

$$Q^{safe}(s_t, u_t, d_t, z) : (1-\gamma)g(s_{t+1}) + \gamma \max\{g(s_{t+1}), Q^{safe}(s_{t+1}, u_{t+1}, d_{t+1}, z)\}.$$

$$Q^{diayn}(s_t, u_t, d_t, z) : r_{t+1}^{diayn} + \gamma Q^{diayn}(s_{t+1}, u_{t+1}, d_{t+1}, z).$$

end for

variational autoencoder. This result can explain why the controller overfits if the adversary is not diverse.To ensure that the disturbance depends on the latent z and avoid colliding into one single policy which may result in overfit problem, we need to add a diversity regularizer in the loss function of the adversary.So in order to realize this, we need to use previous works of unsupervised discovery of skills in reinforcement learning.

Why does using skills improving stability? Because,by training the controller to recognize and respond to different random generating skills (i.e.disturbances or challenges guided by latent variables), our method learns how to react in various scenarios. This diversity of challenges forces the controller to continuously adapt to new situations, thereby improving its stability and robustness in changing environments. Also,we think that through training with diverse skills, the controller and agent actually not only learns how to specifically handle the current challenges but also it could better generalizes to unseen situations. This means the controller remains more stable when facing new, unknown challenges, avoiding drastic performance drops, thereby we could improve its stability and robustness in changing environments in this way.

Theory Results Now we derive the evidence lower bound for the control policy. Define the failure set as $\mathcal{F}$.

Here we fix the initial state s_0 and the adversary π^d. For convenience, we define the safety margin $g(s) > 0$ such that $g(s) < \epsilon \iff s \in \mathcal{F}$. We note that in Section 2, we defined $g(s) > 0 \iff s \in \mathcal{F}$, so we need to inverse the sign of $g(x)$ in Section 2 to satisfy the definition here.

We fix the adversary π^d. The objective of the controller is to maximize

$$\max_{\pi^u} \mathop{\mathbb{E}}_{z^d\sim p(z)} [V^{\pi^u, z^d}(s_0)],$$

where we define the value function

$$V^{\pi^u, z^d}(s) \triangleq \mathbb{E}[(1-\gamma)g(s) + \gamma \min\{g(s), V^{\pi^u, z^d}(s')\} \mid z \sim p(z), u \sim \pi^u(\cdot|s, z), d \sim \pi^d(\cdot|s, z^d), s' = f(s, u, d)].$$

Here the disturbance fixes its skill z^d, which is sampled in the beginning. And the controller randomly guesses the disturbance skill z at each step.

In intuition, if we introduce a learning discriminator $q(z|s)$ to help the controller π^u obtain the information of z^d, the performance of π^u will increase. But this may result in overfitting. To avoid overfitting, we add the KL regularization $\log p(z) - \log q(z|s)$ in the training objective, so that we are optimizing a lower bound of $V^{\pi^u, z^d}(s)$.

Theorem 1

We consider arbitrarily another value function as the following form:

$$\widetilde{V}^{\pi^u, z^d}(s) \triangleq \mathop{\mathbb{E}}_{\substack{z\sim q(z|s),\\ u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)\log g(s) + \log p(z) - \log q(z|s) + \gamma \min\{\log g(s), \widetilde{V}^{\pi^u, z^d}(s')\}].$$

Then, we can obtain the following evidence inequality for lower bound, in fact for any state s,

$$\log V^{\pi^u, z^d}(s) - \widetilde{V}^{\pi^u, z^d}(s) \geq 0.$$

In fact, we note that in Algorithm 1, the controller minimizes $Q^{safe} + \beta Q^{diayn}$. In each step, the diayn reward is $\log q(z|s) - \log p(z)$. Therefore, by minimizing Q^{diayn}, we are actually maximizing $\log p(z) - \log q(z|s)$, which concurs with our theoretical analyses.

5 Limitations

Our method DIAYN at this stage,it may have some potential limitations which may lead to some future research and exploration :

Computational Complexity: The proposed methods, particularly involving Hamilton-Jacobi-Isaacs (HJI) theory and neural networks, may have high computational demands, limiting their applicability in resource-constrained environments.

Generalization Issues: While the approach aims to reduce overfitting by increasing disturbance diversity, it's unclear how well the method generalizes to completely unseen environments or adversaries beyond the training set.

Dependency on Accurate Models: The effectiveness of the proposed adversarial reinforcement learning algorithm relies heavily on the accuracy of the dynamic models used. Inaccurate modeling of the environment or the adversarial disturbance can significantly impair the performance.

Scalability Concerns: As the state dimension increases, the methods like HJI become computationally prohibitive, suggesting scalability issues for complex, high-dimensional environments.

Lack of Empirical Validation: Our method at this stage primarily focuses on theoretical aspects and somehow it lacks extensive empirical evidence to support the practical effectiveness of the proposed methods in real-world scenarios.

Risk of Over-Engineering: The introduction of latent-conditioned controllers and adversaries, while innovative, might lead to over-engineered solutions that are difficult to implement and maintain in less controlled, real-world settings.

Appendix A: Derivation for Theorem 1

Theorem 1 For any given state s,we have the following inequality for value functions given in section 5:

$$\log V^{\pi^u, z^d}(s) - \widetilde{V}^{\pi^u, z^d}(s) \geq 0.$$

Proof of Theorem 1:

$$V^{\pi^u,z^d}(s) = \mathop{\mathbb{E}}_{\substack{z\sim p(z),\\ u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)g(s) + \gamma\min\{g(s), V^{\pi^u,z^d}(s')\}].$$

$$\begin{aligned}
\log V^{\pi^u,z^d}(s) &= \log \mathop{\mathbb{E}}_{\substack{z\sim p(z),\\ u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)g(s) + \gamma\min\{g(s), V^{\pi^u,z^d}(s')\}] \\
&= \log \mathop{\mathbb{E}}_{z\sim p(z)} \left[\frac{q(z|s)}{q(z|s)} \mathop{\mathbb{E}}_{\substack{u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)g(s) + \gamma\min\{g(s), V^{\pi^u,z^d}(s')\}]\right]. \\
&\geq \mathop{\mathbb{E}}_{z\sim q(z|s)} \left[\log p(z) - \log q(z|s) + \log \mathop{\mathbb{E}}_{\substack{u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)g(s) + \gamma\min\{g(s), V^{\pi^u,z^d}(s')\}]\right] \\
&\geq \mathop{\mathbb{E}}_{z\sim q(z|s)} \left[\log p(z) - \log q(z|s) + \mathop{\mathbb{E}}_{\substack{u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} \log[(1-\gamma)g(s) + \gamma\min\{g(s), V^{\pi^u,z^d}(s')\}]\right] \\
&\geq \mathop{\mathbb{E}}_{z\sim q(z|s)} \left[\log p(z) - \log q(z|s) + \mathop{\mathbb{E}}_{\substack{u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)\log g(s) + \gamma\log\min\{g(s), V^{\pi^u,z^d}(s')\}]\right] \\
&\geq \mathop{\mathbb{E}}_{z\sim q(z|s)} \left[(1-\gamma)\log g(s) + \log p(z) - \log q(z|s) + \mathop{\mathbb{E}}_{\substack{u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [\gamma\min\{\log g(s), \log V^{\pi^u,z^d}(s')\}]\right].
\end{aligned}$$

Now that we define another value function as following:

$$\widetilde{V}^{\pi^u,z^d}(s) \triangleq \mathop{\mathbb{E}}_{\substack{z\sim q(z|s),\\ u\sim\pi^u(\cdot|s,z),\\ d\sim\pi^d(\cdot|s,z^d),\\ s'=f(s,u,d).}} [(1-\gamma)\log g(s) + \log p(z) - \log q(z|s) + \gamma\min\{\log g(s), \widetilde{V}^{\pi^u,z^d}(s')\}].$$

We can obtain the evidence lower bound: For any state s,

$$\log V^{\pi^u,z^d}(s) - \widetilde{V}^{\pi^u,z^d}(s) \geq 0.$$

References

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Kai-Chieh Hsu, Duy Phuong Nguyen, and Jaime Fernàndez Fisac. Isaacs: Iterative soft adversarial actor-critic for safety. In *Learning for Dynamics and Control Conference*, pages 90–103. PMLR, 2023.

Jaime F. Fisac Kai-Chieh Hsu, Duy P. Nguyen. Isaacs: Iterative soft adversarial actor-critic for safety. *Proceedings of Machine Learning Research vol XX:1–14, 2023*.

Ian M Mitchell, Alexandre M Bayen, and Claire J Tomlin. A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on automatic control*, 50(7):947–957, 2005.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.