Adversarial Distributional Reinforcement Learning against Extrapolated Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Distributional reinforcement learning (DiRL) accounts for stochasticity in the environment by learning the full distribution of return and has hugely improved performance due to better differentiating between states and training-phase policy evaluation. However, even if the environment is not relied upon for being deterministic, the agent still only gets to traverse a single possible path and therefore observe a single return backup feedback during online learning. Effectively, DiRL is learning the whole distribution using only one sample from it, relying substantially on inductive bias. This work aims to alleviate catastrophically generalizing from a similar-looking state whose behavioural consequence (under the current policy) is actually disparate, i.e., an attack, with adversarial training. To do this, we first identify the set of attacks in which the agent's behavioural consequences are sufficiently dissimilar to the current state, then pick the strongest which incurs the largest model distinguishability error: the smallest distance between predicted return distributions. Finally, we update the return distribution model by ascending the gradient of this minimal distance, effectively solving a minimax problem. In defining attacks, we use bisimulation metric to measure behavioural similarity. To decide the distance between predicted return distributions, which needs to be differentiable with respect to the return distribution model, we train a value discriminator recognizing true Bellman backups from fake ones, and use the contrastive score as a proxy. Experiments on MuJoCo environments suggest that the proposed method is able to improve DiRL performance however the return distribution is modelled.

1 Introduction

Rather than estimating the expected return or value function in reinforcement learning, distributional reinforcement learning (DiRL) Bellemare et al. (2017) models the full distribution of the return, viewing it as random variable whose stochasticity stems from the intrinsic randomness of the environment and potentially also from the agent itself. Thus said, the stochasticity of the environment is not nevertheless providing extra learning signals just because it is assumed to exist – During online training, the agent still only gets to traverse a single possible path and therefore observe a single return backup feedback.

Admittedly, some DiRL approaches approximate the return distribution as a parametric miniature subset of its examples Bellemare et al. (2017); Dabney et al. (2018a,b); Rowland et al. (2019); Martin et al. (2020); Barth-Maron et al. (2018); Singh et al. (2020); Kuznetsov et al. (2020), and therefore carry a multi-sample prediction for any step in computing the Bellman backup target; others Doan et al. (2018); Freirich et al. (2019); Choi et al. (2019); Li & Faisal (2021) represent the return distribution as a generative model, able to generate as many return samples as desired. However, these diversities all come from the model itself (i.e. multiple guesses), bearing no additional

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

learning information, whereas the feedback from the environment which contains the ground truths for improving the model, i.e., the trace of rewards and subsequent states, is only possible to be observed in *one* example[1]. Effectively, vanilla DiRL methods, which are usually parametric, are learning the return distributions with only one backup signal per state / state-action pair, relying heavily on generalization from nearby data and their observed backup signals.

As a consequence, it is possible that two states appear similar and therefore have similar predicted return distributions, whereas the same policy (e.g. as defined by the same function and set of parameters) has disparate behaviour consequences in them in reality. In this scenario, the return distributions in the two states should be different.

To this end, we propose to alleviate generalizing from similar-looking but behaviourally distinct states in return distribution learning with adversarial training Goodfellow et al. (2015). Specifically, for each state, we find the worst-case scenario which incurs the lowest distinguishability of the return distribution model within a vicinity of the state consisting of states behaviourally dissimilar to it, and update the return distribution model to increase this minimal gap as a regularizer to the original modelling objective.

Robust adversarial reinforcement learning is not a new topic, which frames the RL problem as a zero-sum Markov game to account for unavoidable and uncontrollable difference between training and testing environments. This can be a two-player game, with an additional policy that is either a destabilizing adversary applying disturbance forces to the system Morimoto & Doya (2005); Pinto et al. (2017), or a risk-seeking adversary Pan et al. (2019); Ren et al. (2020a,b). Alternatively, when the adversarial examples can be explicitly defined in the state space, the problem is akin to the perturbation issue in supervised learning Goodfellow et al. (2015); Madry et al. (2018); Cai et al. (2018), and can therefore be recast as a minimax optimization problem solved with adversarial training Zhang et al. (2020); Oikarinen et al. (2021).

In this work, we borrow the idea of adversarial training, not to improve robustness against perturbation though, but to prevent extrapolated generalization in learning return distributions due to having to rely excessively on inductive bias. Please note, our intuition works the other way round compared to the usual sense of adversarial training (which encourages generalization around states visually slightly different), *discouraging* generalization from behaviourally dissimilar states.

Experiments on MuJoCo tasks Todorov et al. (2012) suggest that the proposed method is able to improve DiRL performance however the return distribution is modelled. We believe we have proposed the first algorithm to discourage erroneous generalization in DiRL.

2 Methods

We consider a Markov decision process $\left(\mathcal{S}, \mathcal{A}, R, P, \gamma\right)$ Puterman (1994), where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces respectively, $R : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ a potentially stochastic reward function, $P : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{P}(\mathcal{S})$ a transition probability density function, and $\gamma \in (0, 1)$ a temporal discount factor. An RL agent has a policy that maps states to a probability distribution over actions $\pi : \mathcal{S} \mapsto \mathcal{P}(\mathcal{A})$. The return $G^\pi(s) := \sum_{t=0}^{\infty} \gamma^t r_t, s_0 = s$ is a random variable which quantifies the accumulated future rewards, its distribution being denoted as $\omega^\pi(s) \in \mathcal{P}(\mathbb{R})$. We consider the state-dependent return in this work, whilst the idea applies also to action-dependent return. The distributional Bellman operator Bellemare et al. (2017) allows the return distribution to be estimated with temporal difference as in scalar RL, its action-marginalized version being Li & Faisal (2021)

$$\mathcal{T}^\pi G^\pi(s) :\overset{D}{=} R(s) + \gamma G^\pi(s'),$$

where the distribution equation $\overset{D}{=}$ specifies that the random variables on both sides of the equation are distributed by the same law, and s' is the next state.

We learn a return distribution model $\hat{\omega}^\pi(s)$ along with a policy $\pi(a|s)$. The dependency on π is dropped in notations if no confusion is to be induced. In DiRL, one instance of Bellman backup target $\mathcal{T}^\pi G^\pi(s)$ is being observed for each s to update $\hat{\omega}$ at s, implicitly relying heavily on generalization from Bellman backups at neighbouring states, some of which might have little behavioural similarity to s. We leverage adversarial training to reduce generalization when state similarity is misconstrued, which contrasts model prediction with ground-truth reality.

[1]Unless the environment can be rewound.

2.1 Behavioural similarity

The adversarial examples, or attacks, are engendered by the properties of the environment and would mislead the model if no special treatment is applied. We define the attack set of s as a subset of the state space in which an attack s^* is so behaviourally dissimilar to s that generalization from $\hat{\omega}(s^*)$ to $\hat{\omega}(s)$ should be discouraged *and* that s^* and s appear similar enough to confuse $\hat{\omega}$:

$$\mathcal{AT}(s) := \big\{s^* \in \mathcal{S} : \ ||s - s^*||_1 < \varepsilon_1, \ \mathcal{M}(s, s^*) > \varepsilon_2\big\}. \tag{1}$$

State vicinity is defined under l_1 norm. $\mathcal{M}(\cdot,\cdot)$ is a functional that measures how well the consequences of two given states can be distinguished, for which we use the bisimulation metric Givan et al. (2003); Ferns et al. (2011); Zhang et al. (2021)

$$\mathcal{M}(s_i, s_j) := |r_i - r_j| + W\big(P(\cdot|s_i, a_i), P(\cdot|s_j, a_j)\big), \forall\, s_i, s_j \in \mathcal{S}.$$

Bisimulation metric quantifies behavioural similarity with a combination of the difference between rewards and the Wasserstein metric W Villani (2008) between the next state distributions, for which we are learning a hybrid transition function $P(s'|\phi(s), a)$ predicting the next state s' from the embedding of the current state $\phi(s)$. To simplify computation, we model $P(s'|\phi(s), a)$ as a Gaussian and use W_2 distance which has a closed-form expression.

Note that the attack set $\mathcal{AT}(s)$ depicts the oracle determining whether two states s and s^* *should* have similar return distributions ω in reality. This oracle will be referred to to identify the differentiating error of the model $\hat{\omega}$, and thus the model $\hat{\omega}$ cannot be involved in defining $\mathcal{AT}$.

2.2 Model distinguishability

To measure the mistake the model $\hat{\omega}$ is making in differentiating between states, we train a binary value discriminator Goodfellow et al. (2014) which when at its optimum is describing the relative probability of a given return value G being drawn according to the return distribution predicted for a given state s

$$\Phi(G|s) := \frac{p\big(G \sim \hat{\omega}(s)\big)}{p\big(G \sim \hat{\omega}(s)\big) + p\big(G \sim \hat{\omega}(\mathcal{S} \backslash s)\big)}, \ \forall\, G \in \mathbb{R}, \ s \in \mathcal{S}. \tag{2}$$

We use $\sim$ to denote the preceding sample being distributed according to the succeeding probability function, and $\hat{\omega}(\mathcal{S} \backslash s)$ to denote the marginal predicted return distribution under the current policy aggregated over the whole of the state space except s:

$$\hat{\omega}(\mathcal{S} \backslash s) := \frac{\int_{\mathcal{S} \backslash s} d^{\pi}(s')\hat{\omega}(s')\mathrm{d}s'}{\int_{\mathcal{S} \backslash s} d^{\pi}(s')\mathrm{d}s'}, \tag{3}$$

with d^{π} representing the stationary state distribution under policy π. The denominator is to normalize d^{π} to make it a proper distribution when s is excluded.

Specifically, Eq. (2) does not contrast $\hat{\omega}(s)$ with a particular state return distribution, but the rest of possible state return distributions as a whole.

Note that Eq. (2) is the equivalent *definition* of $\Phi(G|s)$ ended up with when trained against the cross-entropy loss given samples $\sim \hat{\omega}(s)$ and samples $\sim \hat{\omega}(\mathcal{S} \backslash s)$, rather than its computation formula, as we only assume the return distribution model $\hat{\omega}$ to be able to be sampled from without having access to its analytical form.

For each s, we use its observed Bellman backup target (computed from any backup method) as the "true" sample $\sim \hat{\omega}(s)$, and the Bellman targets at other states belonging to the same training batch as the "fake" samples $\sim \hat{\omega}(\mathcal{S} \backslash s)$.

With this relative probability estimator, we define the distance between the predicted return distributions for a state s and an attack s^* as a contrastive score

$$\begin{aligned} D\big(\hat{\omega}(s), \hat{\omega}(s^*)\big) :&= -\mathbb{E}_{G \sim \hat{\omega}(s)}\left[\log \frac{p\big(G \sim \hat{\omega}(s^*)\big)}{p\big(G \sim \hat{\omega}(\mathcal{S} \backslash s^*)\big)}\right] \\ &= -\mathbb{E}_{G \sim \hat{\omega}(s)}\big[\log \Phi(G|s^*) - \log\big(1 - \Phi(G|s^*)\big)\big], \end{aligned} \tag{4}$$

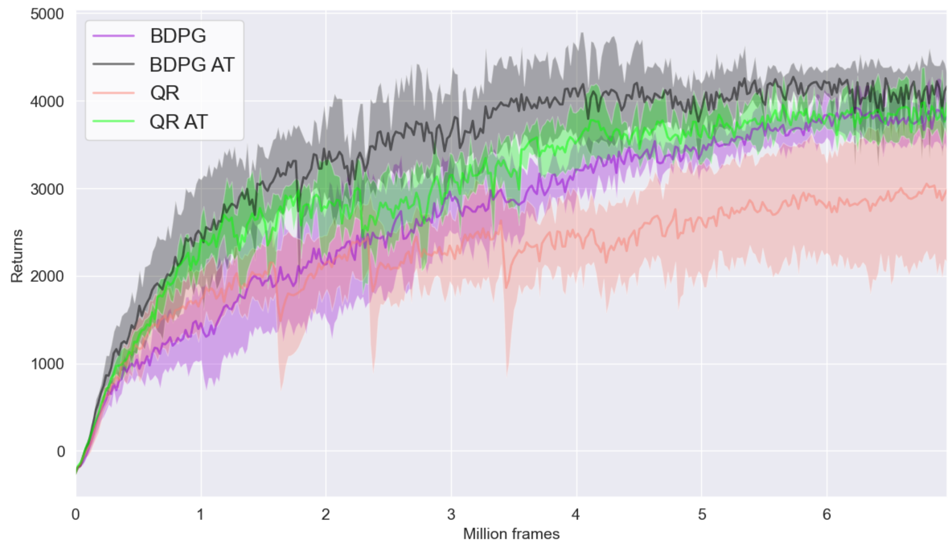

Figure 1: Training performance on HalfCheetah-v3. Solid lines and shaded areas represent mean and standard deviation over 5 runs respectively.

which can be directly computed from the value discriminator Φ. Specifically, Eq. (4) does not directly contrast $\hat{\omega}(s)$ with $\hat{\omega}(s^*)$, but describes how likely a sample drawn from $\hat{\omega}(s)$ is not deemed a sample from $\hat{\omega}$ predicted at s^* relative to at any other state than s^*, i.e., a one-versus-many binary decision. This setting can be thought of as a regularization (inside regularization) on diversity to prevent all return distributions being overly far away from each other due to the adversarial training.

We use Monte Carlo estimate of the expectation over $\hat{\omega}(s)$, by sampling multiple times i.i.d. from $\hat{\omega}(s)$ with reparameterization Kingma & Welling (2014). This is so that D remains differentiable with respect to the parameters of $\hat{\omega}$.

2.3 Adversarial loss

Finally, we define the adversarial objective to alleviate erroneous generalization as

$$\max_{\hat{\omega}} \min_{s^* \in \mathcal{AT}(s)} D\big(\hat{\omega}(s), \bar{\hat{\omega}}(s^*)\big), \tag{5}$$

in which the overhead bar $\bar{\cdot}$ indicates that the model in question is deemed fixed in the current context.

Basically, for each state s, we first find the worst-case scenario where an attack s^* is sufficiently visually similar and behaviourally dissimilar to s, yet however has a learned return distribution closest to its own according to the D in Eq. (4). Then we increase this minimal D in updating $\hat{\omega}$ at s, appended as a regularizer to the original return distribution modelling objective such that the model can still be updated with conventional stochastic gradient descent.

3 Results

This is an ongoing project, we have only partial results and implementation details are not final at this stage.

We use the scalar RL method PPO Schulman et al. (2017) as implementation backbone, substituting the value function with return distribution modelled from QR-DQN Dabney et al. (2018b) (adapted for state return distribution) and BDPG Li & Faisal (2021) respectively, as our two distributional RL baselines. QR represents the return distribution $\hat{\omega}$ as a set of return samples corresponding to a set of evenly distributed quantile levels. BDPG represents $\hat{\omega}$ as a variant of variational auto-encoder. We incorporate the proposed adversarial training on both baselines, denoted as "QR AT", "BDPG AT" respectively.

Experiments were conducted on Mujoco environments Todorov et al. (2012). From the results on $\mathrm{HalfCheetah}$ as shown in Fig. 1, we can see that the proposed adversarial training can improve upon both QR and BDPG.

4 Discussion & Conclusions

In this work, we propose to leverage minimax adversarial training to prevent extrapolated generalization in modelling parametric return distributions. For each state s, we first search for the attacks $s^* \in \mathcal{AT}(s)$ that are both visually similar to s so that there may be generalization between them in learning the return distribution model $\hat{\omega}$, and behaviourally dissimilar to s (as measured by bisimulation metric) so that generalization from $\hat{\omega}(s^*)$ to $\hat{\omega}(s)$ should be discouraged. Then the largest distinguishing error of the model among the attacks is being regularized during model update. To estimate the distance between two predicted return distributions D, we train a value discriminator Φ depicting whether a given return value is distributed according to the return distribution predicted by the model at the given state or not. The model distinguishability $D\big(\hat{\omega}(s), \hat{\omega}(s^*)\big)$ is therefore computed as how $\hat{\omega}(s)$ is farther away from $\hat{\omega}$ predicted at s^* than at any other state $\mathcal{S}\backslash s^*$.

Proof-of-concept results on $\mathrm{HalfCheetah}$ suggest that the proposed idea can considerably improve learning speed as well as asymptotic performance regardless whether the return distribution is approximated as a particle set or a generative model, two most prevalently adopted return distribution modelling methods.

Admittedly, the experiment is too scarce to be statistically remarkable at the moment. We are conducting further investigations into e.g. the inner work of adversarial training in DiRL, the degree and consequence of misgeneralization, as well as performing on more environments. Hopefully we will have a thorough experimental understanding of our idea soon. In the meantime, we do hope to get and would appreciate any feedback on the work.

References

Barth-Maron, G., Hoffman, M. W., Budden, D., Dabney, W., Horgan, D., TB, D., Muldal, A., Heess, N., and Lillicrap, T. Distributed distributional deterministic policy gradients. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 449–458, Sydney, Australia, Aug 2017.

Cai, Q.-Z., Liu, C., and Song, D. Curriculum adversarial training. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3740–3747, July 2018.

Choi, Y., Lee, K., and Oh, S. Distributional deep reinforcement learning with a mixture of Gaussians. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 9791–9797, 2019.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1096–1105, Stockholmsmässan, Stockholm Sweden, 2018a.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018b.

Doan, T., Mazoure, B., and Lyle, C. GAN q-learning, 2018.

Ferns, N., Panangaden, P., and Precup, D. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

Freirich, D., Shimkin, T., Meir, R., and Tamar, A. Distributional multivariate policy evaluation and exploration with the bellman gan. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1983–1992, Long Beach, California, USA, June 2019.

Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1):163–223, 2003.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, May 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Kuznetsov, A., Shvechikov, P., Grishin, A., and Vetrov, D. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Li, L. and Faisal, A. A. Bayesian distributional policy gradients. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35(10), pp. 8429–8437, 2021.

Madry, A., Makelov, A., Schmid, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, April 2018.

Martin, J., Lyskawinski, M., Li, X., and Englot, B. Stochastically dominant distributional reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, Virtual, July 2020.

Morimoto, J. and Doya, K. Robust reinforcement learning. *Neural Computation*, 17(2):335–359, Feb 2005.

Oikarinen, T., Zhang, W., Megretski, A., Daniel, L., and Weng, T.-W. Robust deep reinforcement learning through adversarial loss. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26156–26167, 2021.

Pan, X., Seita, D., Gao, Y., and Canny, J. Risk averse robust adversarial reinforcement learning, 2019.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 2817–2826, 2017.

Puterman, M. L. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Ren, K., Zheng, T., Qin, Z., and Liu, X. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020a.

Ren, Y., Duan, J., Li, S. E., Guan, Y., and Sun, Q. Improving generalization of reinforcement learning with minimax distributional soft actor-critic. In *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, 2020b.

Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M. G., and Dabney, W. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5528–5536, Long Beach, California, USA, June 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

Singh, R., Lee, K., and Chen, Y. Sample-based distributional policy gradient, 2020.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012.

Villani, C. *Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften.* Springer Berlin Heidelberg, 2008.

Zhang, A., McAllister, R. T., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21024–21037, 2020.