CURIOSITY AUGMENTED EXPLORATION FOR OFF-POLICY REINFORCEMENT LEARNING

Ashish Malik
Department of Mechanical Engineering
Punjab Engineering College
Chandigarh, 160012, India
ashishmalik.bemech14@pec.edu.in

ABSTRACT

Text.

1 INTRODUCTION

Reinforcement learning algorithms have gained tremendous success in challenging high-dimensional continuous control problems in recent years. State of the art algorithms such as Soft Actor Critic (SAC, Haarnoja et al. (2018)), Twin Delayed Deterministic Policy Gradient (TD3, Fujimoto et al. (2018)) and Proximal Policy Optimization (PPO, Schulman et al. (2017)) are actor-critic methods that employ Boltzmann exploration for generating new trajectories for learning. The off-policy learning algorthim SAC, achieves the greatesr sample efficiecy among these algorithms. SAC maximizes the trade-off between the expected extrinsic returns and entropy of the learned soft-policy. However, SAC still requires millions of environment interactions to learn a viable policy and value-function estimates in high-dimensional settings. Poor sample efficiency is a major obstacle in widespread adoptation of deep reinforcement learning (DRL) in the real world. A few limitations in existing algorithms and common exploration strategies that contribute to poor sample efficiency are:

1. *Directionally uninformed exploration:* Sampling actions from gaussian distribution of action probabilities is a standard practice in DRL. Actions from the opposite side of the mean are have equal probabilities of selections with this method. It causes high frequency perturbations during exploration which act as a low-pass filter and tend to cancel each other, leading to poor exploration Stulp & Sigaud (2013); Kober & Peters (2008). Also, current policies are obtained after incremental updates over past policies using temporal difference methods. Therefore, it is highly likely that action-spaces where the past-policies had high probability density have already been explored. These facts reveal the disadvantages of Directionally uninformed exploration.
2. *Pessimistic underexploration (Ciosek et al., 2019):* SAC and TD3 stabilize learning by avoiding overestimation (Van Hasselt et al., 2015) using greedy maximization of a lower bound of action-value estimates. However, exploration in a state-action space with spuriously high estimation of this lower bound will cause the policies covariance to collapse in the region. This will discourage exploration of new actions and prevent the improvement of critic's estimation. Entropy regularization in SAC prevents the policy's covariance from collapsing to zero, but it does not directly address the problem.
3. *Unimodal policy distribution:* Unimodal action-probability distribution (ex. Gaussian distribution) limits the expressivity of exploration policies. Multi-modal distributions for exploration policies increase exploration and accelerate discovery of good policies (Mazoure et al., 2020).
4. *Goal-focused exploration:* Goal-less exploration based on intrinsic rewards (curiosity) is an active research area in DRL. It is inspired by natural learning in babies and is shown to learn useful behaviors without any external rewards (Burda et al., 2018). Goal-focused exploration methods (with external reward signal) explores in an on-policy manner that narrows the active area for exploration. Off-policy exploration with external reward signal considerably slows down learning and tend to destabilize it (the deadly triad, (Sutton & Barto, 2018)).

The continuous version of popular mountain-car challenge is ideal for observing the effect of these shortcomings. In this version, the agent receives a large positive reward for reaching the goal and is penalized for every action it takes. This reward dynamic makes it a difficult exploration focused challenge as the agent may falsely determine that it is better to not take any actions at all if the target is not reached soon enough during exploration. States visited during exploration by various algorithms are shown in fig. FIG. DETAILS ABOUT THE FIGURE GENERATION AND OBSERVATIONS FROM THE FIG.

In this work, we present a novel method of off-policy learning that addresses the aforementioned shortcomings using only an additional hyperparameter. Our method disentangles the exploration and exploitation trade-off in off-policy learning methods and we name it NAME. NAME learns a separate exploration focused policy using intrinsic rewards. Both exploration focused policy and target polices are used for selecting actions for explorations. WE TEST, AND OTHER EXPERIMENTATION RESULTS. We also perform ablations to isolate the effect of the new hyperparameter introduced and demonstrate the stability of NAME.

2 Curiousity augmentation

C-SAC trains a separate exploration focused policy for aiding exploration and learning. From now on we will call this exploration policy and the target policy as exploitation policy. The exploration policy maximizes a tradeoff between the expected intrinsic returns and KL divergence from the exploition policy. Let $z(s_t, a_t, s_{t+1})$ be the intrinsic reward signal recieved when the agent takes action a_t in state s_t and makes the transition to s_{t+1}. The exploration policy with parameters ϕ maximizes the following objective:

$$J(\pi_\phi) = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim\rho_{\pi_\phi}}\Big[z(s_t,a_t,s_{t+1}) + \alpha\mathcal{H}\big(\pi_\phi(\cdot|s_t)\big) - \beta D_{KL}\Big(\pi_\phi(\cdot|s_t)\big|\big|\pi_\theta(\cdot|s_t)\Big)\Big] \tag{1}$$

π_θ is the exploitation policy parameterized with θ. The temperature parameter β determines the relative importance of the KL-divergence term against the intrinsic returns. β determines how far the exploration policy can stray from exploitation policy. A two component gaussian mixture model $\mathcal{M}(\pi_\theta, \pi_\phi)$ is used for selecting actions for generating exploration trajectories. The probability of selecting action a_t in in state s_t is given as:

$$p(a_t, s_t) = \eta\pi_\theta(s_t) + (1-\eta)\pi_\phi(s_t) \tag{2}$$

Where, η is the relative weightage of the two policies. This formulation has a number of conceptual and practical advantages. Use of a specialized exploration policy which uses an intrisic reward signal results in better exploration. The KL divergence penalty ensures that the exploration policy explores regions in the vicinity of the exploitation policy, which increases stability. KL constraint between exploration and exploitation policy has been previously employed for improving exploration (Ciosek et al., 2019). Bi-modal action probability distribution better express the high-priority action-space for learning and is more natural. Humans take actions that either results in best exploitation of our knowledge or actions that have the potential to result in greatest learning for future exploitations. Therefore, a bi-modal distribution is a natural choice.

2.1 Policy iteration for exploration policy

We calculate the value of the policy according to objective 1 in the policy evaluation step. The Q-value for the exploration policy can be iteratively computed by applying a modified Bellman backup operator $\mathcal{T}^{\pi_\phi}$ as:

$$\mathcal{T}^{\pi_\phi}Q_{\pi_\phi}(s_t,a_t) \triangleq z^*(s_t,a_t,s_{t+1}) + \gamma\mathbb{E}_{s_{t+1}\sim\rho_{\pi_\phi},a_{t+1}\sim\pi_\phi}\Big[Q(s_{t+1},a_{t+1})\Big] \tag{3}$$

$$z^*(s_t,a_t,s_{t+1}) = z(s_t,a_t,s_{t+1}) + \mathbb{E}_{s_{t+1}\sim\rho_{\pi_\phi}}\Big[\alpha\mathcal{H}\big(\pi_\phi(\cdot|s_t)\big) - \beta D_{KL}\big(\pi_\phi(\cdot|s_{t+1})\big|\big|\pi_\theta(\cdot|s_{t+1})\big)\Big] \tag{4}$$

Lemma 1. *Consider stationary soft-policies π_ϕ and π_θ, the Bellman backup operator in 3, and a mapping $Q^0 : \mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}$ with $|\mathcal{A}| < \infty$. Let $Q^{k+1} = \mathcal{T}^{\pi_\phi}Q^k$. The sequence Q^k will converge to the soft Q-value of π_ϕ as $k\rightarrow\infty$.*

Proof. Text. □

The policy is updated towards the exponential of the Q-function in the policy improvement step.

$$(\pi_\phi)_{\text{new}} = \arg\min_{\pi' \in \Pi} D_{\text{KL}}\left(\pi_\phi(\cdot|s_t)\Bigg|\Bigg|\frac{\exp Q^{\pi_\phi}(s_t,\cdot)}{Z^{\pi_\phi}(s_t)}\right) \tag{5}$$

where, Π is the parameterized distribution class in which we wish to project the new policy $(\pi_\phi)_{\text{new}}$. The partition function $Z^{\pi_\phi}(s_t)$ normalizes the exponential of Q-function distribution. It does not contribute to the gradient of the new policy and thus can be convienently ignored.

Lemma 2. *Let $\pi_\phi \in \Pi$ and $(\pi_\phi)_{new}$ be the solution of the minimization objective 5. Then, $Q^{(\pi_\phi)_{new}}(s_t,a_t) \geq Q^{\pi_\phi}(s_t,a_t)$ for all $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ with $|\mathcal{A}| < \infty$.*

Proof. Text. □

Theorem 1. *Repeated applications of policy evaluation and policy improvement on any $\pi_\phi \in \Pi$ will converge it to π^* such that $Q^{\pi^*}(s_t,a_t) \geq Q^{\pi}(s_t,a_t)$ for any $\pi \in \Pi$, $s_t \in \mathcal{S}$ and $a_t \in \mathcal{A}$ assuming $|\mathcal{A}| < \infty$.*

Proof. Text. □

2.2 Curious soft actor critic algorithm

Theorem 1 can provably learn the optimal exploration policy. But it is tractable only in the tabular setting. In this sub-section we derive a practical algorithm for augmenting off-policy learning with curious exploration. We consider a parameterized exploration policy $\pi_\phi(a_t|s_t)$, exploitation policy π_θ, action-value functions $Q_{i=1,2}(s_t,a_t)$ and target action-value functions $\hat{Q}_{i=1,2}(s_t,a_t)$ parameterized with ϕ, θ, ψ and, $\hat{\psi}$ respectively. The Q-functions are trained by minimizing the following Bellman residual:

$$J_{Q_i}(\psi) = \mathbb{E}_{(s_t,a_t,s_{t+1})\sim\mathcal{D}}\left[\frac{1}{2}\big(Q_i(s_t,a_t) - \bar{Q}(s_t,a_t)\big)^2\right] \tag{6}$$

Here, $\mathcal{D}$ is the replay buffer and $\bar{Q}(s_t,a_t)$ is given as:

$$\bar{Q}(s_t,a_t) = z(s_t,a_t,s_{t+1}) + \gamma \min\big(Q_{i=1,2}(s_{t+1},\bar{a}_{t+1})\big) + \alpha \tag{7}$$

$$-\beta D_{KL}\Big(\pi_\phi(\cdot|s_{t+1})\Big|\Big|\pi_\theta(\cdot|s_{t+1})\Big) \tag{8}$$

$\bar{a}$ are actions sampled using the current exploration policy. The policy is trained by minimizing the following objective:

$$J_{\pi_\phi}(\phi) = \mathbb{E}_{s_t\sim\mathcal{D},\epsilon\sim\mathcal{N}}\Big[\log \pi_\phi\big(f_\phi(\epsilon_t,s_t)|s_t\big) - \min\big(Q_{i=1,2}(s_t, f_\phi(\epsilon_t,s_t))\big)\Big] \tag{9}$$

where,

$$a_t = f_\phi(\epsilon_t, s_t) \tag{10}$$

signifies the reparameterization trick.

References

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. In *Advances in Neural Information Processing Systems*, pp. 1787–1798, 2019.

Scott Fujimoto, Herke Van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Jens Kober and Jan Peters. Policy search for motor primitives in robotics. *Advances in neural information processing systems*, 21:849–856, 2008.

Bogdan Mazoure, Thang Doan, Audrey Durand, Joelle Pineau, and R Devon Hjelm. Leveraging exploration in off-policy algorithms via normalizing flows. In *Conference on Robot Learning*, pp. 430–444, 2020.

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

Freek Stulp and Olivier Sigaud. Robot skill learning: From reinforcement learning to evolution strategies. *Paladyn, Journal of Behavioral Robotics*, 4(1):49–61, 2013.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *arXiv preprint arXiv:1509.06461*, 2015.