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

Actor-critic methods with greedy explorations have become a standard for achieving state of the art performance in model-free reinforcement learning. Prior works have identified and tackled a few shortcomings of greedy exploration. In this work, we avoid the common pitfalls of greedy exploration using a novel method that decouples exploration-exploitation objectives. The exploration objective is trained with intrinsic rewards (curiosity) to maximize learning and exploitation objective is trained with extrinsic rewards to maximize returns. We show that this decoupling addresses the limitations of greedy exploration. We show that our method can be combined with any model-free actor critic method and its performance is tested on several challenging continuous control tasks. Empirical results show that it achieves state of the art performance on continuous control benchmarks.

1 INTRODUCTION

Reinforcement learning algorithms have gained tremendous success in challenging high-dimensional continuous control problems. 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 rely on greedy exploration for generating trajectories for learning. SAC maximizes the trade-off between expected extrinsic returns and entropy of the learned soft-policy. The entropy augmented learning objective ensures that policy's covariance does not collapse to zero and thus maintains high exploration rate and avoids local minimas. Still, SAC (and other state of the art algorithms) require millions of interactions to learn viable policies in high-dimensional settings. This poor-sample efficiency is a direct result of limitations of greedy exploration, some of which are:

1. *Directionally uninformed exploration*: Adding uncorrelated noise in actions either by sampling from a normal distribution or by adding random noise is a standard practice in deep reinforcement learning. These practices lead to actions from opposite side of the mean to be selected with equal probabilities. High frequency perturbations around mean actions acts as a low pass filter and tend cancel each other out, leading to poor exploration (Stulp & Sigaud, 2013; Kober & Peters, 2008). Additionally, policies are obtained after incremental updates over previous policies using temporal difference methods. So, action-spaces where the past-policies had high densities are likely explored extensively. Therefore exploration efforts should be focused towards novel action-spaces.
2. *Pessimistic under-exploration* (Ciosek et al., 2019): SAC and TD3 avoid overestimation of Q-estimates and stabilize learning (Van Hasselt et al., 2015) using greedy maximization of a lower bound of action-value estimates. But exploration in state-action space with spuriously high estimation of this lower bound causes the policy's covariance to collapse in that region. It discourages exploration of new actions and prevents improvement of Q-estimate. SAC's entropy regularlization prevents policy's covariance from reducing to zero, but it also does not directly address the problem efficiently.

Similarly, few limitation of standard practices in deep reinforcement learning (DRL) are:

1. *Goal focused exploration*: "Goal-less" exploration based on intrinsic rewards (curiosity) is an active research area in (DRL). It is inspired by the natural learning in infants and is shown to bring about useful behaviors without any external rewards (Burda et al., 2018). Goal-focused exploration methods (using extrinsic reward signals) explores in an on-policy manner that narrows the active area of exploration. Off-policy exploration on the other hand considerably slows down and tend to destabilize learning (Sutton & Barto, 2018).
2. *Unimodal policy distribution*: Standard implementations of RL algorithms employ unimodal action-probability distributions (ex. Gaussians). These distributions limits the expressivity of policies to only a single area of interest. Multi-modal action-probability distributions increase exploration efficiency and thus accelerates discovery of good policies (Mazoure et al., 2020).

The continuous version of the popular mountain car challenge is ideal for observing the effect of fore-mentioned limitations. In this task, the agent receives a large positive reward for reaching the goal and is penalized for every action it takes. The challenge necessitates efficient exploration as the agent may spuriously learn that it is better not to take any action at all if it does not reach the goal position soon enough during exploration. This reward dynamic makes it a difficult exploration focused challenge. States visited by various algorithms on this challenge are shown in Fig. DETAILS.

In this work, we present a novel method of off-policy exploration that can be used for any temporal difference model-free RL algorithms that address the above-mentioned shortcomings. The method is based upon disentangling exploration and exploitation trade-off. Our method learns a separate exploration policy in addition to a target policy. The exploration policy is trained with intrinsic rewards and aims at maximizing learning, whereas the target policy trained by the base-algorithms focuses on extrinsic returns. Both exploration policy and target policies are used for active exploration.

2 Curiosity augmentation

Our method trains a separate exploration focused policy for aiding and expedite learning. For clearity we refer to the two policies as exploration and target policy. The exploration policy maximizes the expected intrinsic returns and while minimizing KL divergence from the target policy. Let z_t be the intrinsic reward signal recieved by when the agent at time-step t. The exploration policy, pi_ϕ, maximizes the following objective:

$$J(\pi_\phi) = \sum_{t=0}^{T} \mathbb{E}_{\tau\sim\pi_\phi}\left[z_t + \alpha\mathcal{H}\big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t)\big) - \frac{1}{\beta}D_{KL}\Big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t)\big|\big|\pi_{\boldsymbol{\theta}}(\cdot|\mathbf{s}_t)\Big)\right] \tag{1}$$

Here, π_θ is the exploration policy. The parameters α and β determines the relative importance of intrinsic returns, entropy of the policy and its KL-divergence from the target policy. The state value function for policy π_ϕ is obtained as:

$$V_{\pi_\phi}(s) = \mathbb{E}_{\tau\sim\pi_\phi}\left[\sum_{t=0}^{T}\gamma^t\Big(z_t + \alpha\mathcal{H}\big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t)\big) - \frac{1}{\beta}D_{KL}\Big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t)\big|\big|\pi_{\boldsymbol{\theta}}(\cdot|\mathbf{s}_t)\Big)\Big|s_0 = s\right] \tag{2}$$

$$\begin{aligned} Q_{\pi_\phi}(s,a) = \mathbb{E}_{\tau\sim\pi_\phi}\Bigg[\sum_{t=0}^{T}\gamma^t z_t + \sum_{t=1}^{T}\gamma^t\Bigg(&\alpha\mathcal{H}\big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t)\big) \\ &- \frac{1}{\beta}D_{KL}\Big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t)\big|\big|\pi_{\boldsymbol{\theta}}(\cdot|\mathbf{s}_t)\Big)\Bigg)\Bigg|s_0 = s, a_0 = a\Bigg] \end{aligned} \tag{3}$$

Parameter β puts a soft penalty on straying too far from the target policy. Therefore, the exploration policy explores regions in the vicinity of the target policy, which increases the stability of learning in the target policy. Such KL constraint between exploration and target policies have previously been employed to improve exploration Ciosek et al. (2019).

2.1 Policy iteration for exploration policy

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

$$\begin{aligned}\mathcal{T}^{\pi_\phi} Q_{\boldsymbol{\phi}}(\mathbf{s}_t, \mathbf{a}_t) \triangleq z_t + \gamma \mathbb{E}_{\tau\sim\pi_\phi}\Big[&Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) + \alpha\mathcal{H}\big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_{t+1})\big) \\ &- \frac{1}{\beta} D_{KL}\Big(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_{t+1})\Big|\Big|\pi_{\boldsymbol{\theta}}(\cdot|\mathbf{s}_{t+1})\Big)\Big] \qquad (4)\end{aligned}$$

Lemma 1. *(Exploration-policy evaluation) Consider the Bellman backup operator in 4 , stationary soft-policies π_ϕ and π_θ, and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 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 $k \to \infty$.*

Proof. Using the definitions of entropy and Kullback–Leibler divergence, we can write 4 as:

$$\begin{aligned}\mathcal{T}^{\pi_\phi} Q_{\boldsymbol{\phi}}(\mathbf{s}_t, \mathbf{a}_t) \triangleq z_t + \gamma \mathbb{E}_{\tau\sim\pi_\phi}\Big[&Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - (\alpha + \frac{1}{\beta}) \log(\pi_{\boldsymbol{\phi}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})) \\ &+ \frac{1}{\beta} \log(\pi_{\boldsymbol{\theta}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}))\Big] \qquad (5)\end{aligned}$$

Let,

$$z_t^* = z_t + \frac{\gamma}{\beta}\mathbb{E}_{\tau\sim\pi_\phi}\Big[\log(\pi_{\boldsymbol{\theta}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})) - (\alpha + \beta)\log(\pi_{\boldsymbol{\phi}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}))\Big] \qquad (6)$$

be the modified intrinsic reward signal. The update rule 4 can then be written as:

$$Q(\mathbf{s}_t, \mathbf{a}_t) \leftarrow z^*(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) + \gamma\mathbb{E}_{(s_{t+1}\sim\rho_{\pi_{\boldsymbol{\phi}}}, a_{t+1}\sim\pi_{\boldsymbol{\phi}})}\Big[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})\Big] \qquad (7)$$

Finally, apply the standard policy evaluation convergence results (Sutton & Barto, 2018). The assumption $|\mathcal{A}| < \infty$ ensures that the modified intrinsic reward signal z^* is bounded. □

The exploration policy is updated by moving closer towards the modified exponential of the exploration state-action function in the policy improvement step as:

$$\pi_{\boldsymbol{\phi}}^{\text{new}} = \arg\min_{\pi'_{\boldsymbol{\phi}}\in\boldsymbol{\Pi}} D_{KL}\left(\pi'_\phi(\cdot|\mathbf{s}_t)\middle|\middle|\frac{\exp\left(Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, \cdot) - (\frac{\alpha+\beta-1}{\beta})\log\pi_\phi(\cdot|\mathbf{s}_t) + \frac{1}{\beta}\log\pi_\theta(\cdot|\mathbf{s}_t)\right)}{Z_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t)}\right) \qquad (8)$$

where, $\boldsymbol{\Pi}$ is the distribution class for projecting the new policy $\pi_{\boldsymbol{\phi}}^{\text{news}}$. Partition function $Z_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t)$ normalizes the exponential of exploration Q-function distribution. It does not contribute to the gradient of the new policy and can be ignored for gradient based policy updates.

Lemma 2. *(Exploration-policy improvement) Let $\pi'_{\boldsymbol{\phi}} \in \boldsymbol{\Pi}$ and $\pi_{\boldsymbol{\phi}}^{new}$ be the solution of the minimization objective 8. Then, $Q_{\pi'_{\boldsymbol{\phi}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, \mathbf{a}_t)$ assuming $|\mathcal{A}| < \infty$, $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t) > 0$ and $\pi_\phi(\mathbf{a}_t|\mathbf{s}_t) > 0$, for all $\mathbf{s}_t \in \mathcal{S}$ and $\mathbf{a}_t \in \mathcal{A}$.*

Proof. Equation 8 can be rewritten as:

$$\begin{aligned}\pi_{\boldsymbol{\phi}}^{\text{new}}(\cdot|s_t) &= \arg\min_{\pi'_{\boldsymbol{\phi}}\in\boldsymbol{\Pi}} D_{KL}\Big(\pi'_\phi(\cdot|\mathbf{s}_t)\Big|\Big|\exp\Big(Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, \cdot) - \Big(\frac{\alpha+\beta-1}{\beta}\Big)\log\pi_\phi(\cdot|\mathbf{s}_t) \\ &\qquad + \frac{1}{\beta}\log\pi_\theta(\cdot|\mathbf{s}_t)) - \log Z_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t)\Big)\Big) \qquad (9) \\ &= \arg\min_{\pi'_{\boldsymbol{\phi}}\in\boldsymbol{\Pi}} J_{\pi_{\boldsymbol{\phi}}}(\pi'_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t))\end{aligned}$$

We have, $J_{\pi_{\boldsymbol{\phi}}}(\pi_{\boldsymbol{\phi}}^{\text{new}}(\cdot|\mathbf{s}_t)) \leq J_{\pi_{\boldsymbol{\phi}}}(\pi_{\boldsymbol{\phi}}(\cdot|\mathbf{s}_t))$ as we can always select $\pi_\phi^{new} = \pi_\phi \in \boldsymbol{\Pi}$. So,

$$\begin{aligned}&\mathbb{E}_{a_t\sim\pi_{\boldsymbol{\phi}}^{\text{new}}}\Big[(\alpha + \frac{1}{\beta})\log\pi_{\boldsymbol{\phi}}^{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) - Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta}\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t) + \log Z_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t)\Big] \\ &\quad\leq \quad \mathbb{E}_{\mathbf{a}_t\sim\pi_{\boldsymbol{\phi}}}\Big[(\alpha + \frac{1}{\beta})\log\pi_{\boldsymbol{\phi}}(\mathbf{a}_t|\mathbf{s}_t) - Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\beta}\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t) + \log Z_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t)\Big] \quad (10)\end{aligned}$$

As partition function Z_{π_ϕ} does not depend on $\mathbf{a}_t$,

$$\begin{aligned}
&\mathbb{E}_{a_t\sim\pi_{\boldsymbol{\phi}}^{\text{new}}}\Big[(\alpha+\frac{1}{\beta})\log\pi_{\boldsymbol{\phi}}^{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) - Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t,\mathbf{a}_t) - \frac{1}{\beta}\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\Big] \\
&\qquad\leq\quad \mathbb{E}_{\mathbf{a}_t\sim\pi_{\boldsymbol{\phi}}}\Big[(\alpha+\frac{1}{\beta})\log\pi_{\boldsymbol{\phi}}(\mathbf{a}_t|\mathbf{s}_t) - Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t,\mathbf{a}_t) - \frac{1}{\beta}\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\Big] \qquad (11)
\end{aligned}$$

The inequality is reduced to:

$$\mathbb{E}_{a_t\sim\pi_{\boldsymbol{\phi}}^{\text{new}}}\Big[Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t,\mathbf{a}_t) - (\alpha+\frac{1}{\beta})\log\pi_{\boldsymbol{\phi}}^{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) + \frac{1}{\beta}\log\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\Big] \geq V_{\pi_\phi}(\mathbf{s}_t) \qquad (12)$$

Next, considering the Bellman equation with intrinsic reward signal for value functions given in 2 and 3 :

$$\begin{aligned}
Q_{\pi_{\boldsymbol{\phi}}}(s_t,a_t) &= z(\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t+1}) + \gamma\mathbb{E}_{s_{t+1}\sim\rho}\big[V_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_{t+1})\big] \\
&\leq z(\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t+1}) + \gamma\mathbb{E}_{s_{t+1}\sim\rho}\Big[\mathbb{E}_{a_{t+1}\sim\pi_{\boldsymbol{\phi}}^{\text{new}}}\Big[Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_{t+1},\mathbf{a}_{t+1}) \\
&\qquad\qquad - (\alpha+\frac{1}{\beta})\log\pi_{\boldsymbol{\phi}}^{\text{new}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}) + \frac{1}{\beta}\log\pi_\theta(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})\Big]\Big] \\
&\ \ \vdots \\
&\leq Q_{\pi_{\phi}^{\text{new}}}(\mathbf{s}_t,\mathbf{a}_t)
\end{aligned}$$

By repeatedly expanding $Q_{\pi_{\boldsymbol{\phi}}}$ on RHS, using bound in 12 and convergence result in lemma 1. □

Theorem 1. *(Exploration-policy iteration) Repeated applications of policy evaluation (as given in lemma 1) and policy improvement (as given in lemma 2) on any $\pi_\phi \in \Pi$ will converge it to π_ϕ^* such that $Q^{\pi_\phi^*}(s_t,a_t) \geq Q^{\pi_\phi}(s_t,a_t)$ assuming $|\mathcal{A}| < \infty$, $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t) > 0$ and $\pi_\phi(\mathbf{a}_t|\mathbf{s}_t) > 0$, for all $\mathbf{s}_t \in \mathcal{S}$ and $\mathbf{a}_t \in \mathcal{A}$.*

Proof. Refer Sutton & Barto (2018) for proof of policy iteration. The assumption $|\mathcal{A}|$ bounds the modified intrinsic reward. □

2.2 Practical curiosity augmented exploration

The above algorithm should be practically approximated to be applicable to large continuous control domains. For that, we will be using function approximators for Q-function and the policy. Also, we will alternate between policy evaluation step and policy improvement instead of running them till convergence. We also use two separate Q-functions and use the minimum of the two to mitigate the positive bias in the policy improvement step. Consider function approximators for state-action value functions $Q_{\boldsymbol{\psi}_1}$, $Q_{\boldsymbol{\psi}_2}$ parameterized using $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$, exploration-policy $\pi_{\boldsymbol{\phi}}$ parameterized using $\boldsymbol{\phi}$ and target-policy $\pi_{\boldsymbol{\theta}}$ parameterized using $\boldsymbol{\theta}$. The soft-Q functions are trained to minimize the following error:

$$J_Q(\boldsymbol{\psi}_i) = \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\mathcal{D}}\left[\frac{1}{2}\Big(Q_{\boldsymbol{\psi}_i}(\mathbf{s}_t,\mathbf{a}_t) - \hat{Q}(\mathbf{s}_t,\mathbf{a}_t)\Big)^2\right], \qquad \text{For } i=1,2 \qquad (13)$$

with,

$$\begin{aligned}
\hat{Q}(\mathbf{s}_t,\mathbf{a}_t) = z_t + \gamma\mathbb{E}_{\mathbf{s}_{t+1}\sim\mathcal{D}}\Big[&\min\Big(Q_{\bar{\boldsymbol{\psi}}_i}(\mathbf{s}_{t+1},\mathbf{a}_{t+1})\Big) - (\alpha+\frac{1}{\beta})\log\pi_\phi(\mathbf{a}_{t+1}|\mathbf{s}_{t+1}) \\
&+ \frac{1}{\beta}\pi_\theta(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})\Big], \qquad \text{For } i=1,2 \text{ and } \mathbf{a}_{t+1}\sim\pi_\phi(\mathbf{s}_{t+1}) \qquad (14)
\end{aligned}$$

where, $\mathcal{D}$ is the replay buffer and $Q_{\bar{\boldsymbol{\psi}}_i}$ are the target networks that are updated using a moving average of the Q-function parameters and are shown to improve stability of the learning process (Mnih et al., 2015).

The exploration-policy is learned by minimizing the expected KL-divergence given in 8.

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[D_{KL} \left(\pi'_\phi(\cdot|\mathbf{s}_t) \middle\| \frac{\exp\left(Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, \cdot) - (\frac{\alpha+\beta-1}{\beta}) \log \pi_\phi(\cdot|\mathbf{s}_t) + \frac{1}{\beta} \log \pi_\theta(\cdot|\mathbf{s}_t)\right)}{Z_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t)} \right) \right]$$

We can use the reparameterization to minimize this objective as each component of the target density function is differentiable. Using reparameterization trick the objective can be written as:

$$\mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}, \epsilon \sim \mathcal{N}} \Big[(\alpha + \frac{1}{\beta}) \log \pi_{\boldsymbol{\phi}}(f_\phi(\epsilon_t, \mathbf{s}_t)|\mathbf{s}_t) - Q_{\pi_{\boldsymbol{\phi}}}(\mathbf{s}_t, f_\phi(\epsilon_t, \mathbf{s}_t)) - \frac{1}{\beta} \log \pi_\theta(f_\phi(\epsilon_t, \mathbf{s}_t)|\mathbf{s}_t) \Big] \quad (15)$$

with,

$$\mathbf{a}_t = f_\phi(\epsilon_t, \mathbf{s_t})$$

A two component Gaussian-mixture model $\mathcal{M}_{(\pi_{\boldsymbol{\phi}}, \pi_{\boldsymbol{\theta}})}$ is used for sampling actions for generating exploration trajectories. The probability distribution for sampling actions is given as:

$$\mathcal{M}_{(\pi_{\boldsymbol{\phi}}, \pi_{\boldsymbol{\theta}})}(\mathbf{a}_t|\mathbf{s}_t) = \eta \pi_{\boldsymbol{\phi}}(\mathbf{a}_t|\mathbf{s}_t) + (1-\eta)\pi_{\boldsymbol{\theta}}(\mathbf{a}_t|\mathbf{s}_t) \quad (16)$$

This Bi-modal action probability distribution better expresses the high-priority action-spaces for exploration in a natural manner. Humans tend to take actions that either lead greatest utilization of learned knowledge or actions that can potentially result in greatest learning of knowledge. The complete algorithm is described in Algorithm 1.

Algorithm 1: Curiosity augmented off-policy learning

```
Initialize $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}_{i=1,2}$, $\bar{\boldsymbol{\psi}}_{i=1,2}$ and $\mathcal{D}$ such that $\boldsymbol{\psi}_{i=1,2} = \bar{\boldsymbol{\psi}}_{i=1,2}$ and $\boldsymbol{\phi} = \boldsymbol{\theta}$;
Initialize off-policy learning algorithm $\mathcal{O}$;
for each iteration do
    Initialize $\mathcal{M}$ using $\eta$, $\pi_{\boldsymbol{\phi}}$ and $\pi_{\boldsymbol{\theta}}$;
    for each environment step do
        $\mathbf{a}_t \sim \mathcal{M}_{(\pi_{\boldsymbol{\phi}}, \pi_{\boldsymbol{\theta}})}(\mathbf{a}_t|\mathbf{s}_t)$;
        $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$;
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1}))\}$
    end
end
```

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.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

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.