参赛队员姓名：郭培扬

中学：上海市世界外国语中学国际部

省份：上海市

国家/地区：中华人民共和国

指导老师姓名：张伟楠

论文题目：基于能量模型的强化学习探索算法

Energy-Based Exploration for Reinforcement Learning

Peiyang Guo
Shanghai World Foreign Language Middle School
hellobrian18@163.com

Abstract

We aim to tackle the exploration problem in reinforcement learning (RL) tasks, which is the key problem to improve the sample efficiency, especially in environments with sparse rewards. The key idea for most previous works is to encourage the agent to visit a novel state that is rarely reached before. Motivated by recent progresses for applications of energy-based models (EBMs) in RL, in this paper, we inherit such an intuitive idea and propose a novel exploration method called energy-based exploration (EBEx) with an intrinsic reward computed via an energy model, which estimates the density of the past trajectories. In our experiments, we integrated our exploration framework with soft actor-critic algorithm and EBEx achieves superior performance on a sparse-reward 2D navigation task.

1 Introduction

Reinforcement learning (RL) is the branch of machine learning methods that learn the optimal policy of an agent so as to maximize its cumulative reward during the interaction with the environment (Sutton and Barto, 2018). Although lots of success have been achieved in a wide range of application domains like video games (Mnih *et al.*, 2015), Go (Silver *et al.*, 2017), and visual navigation (Silver *et al.*, 2017), they requires dense reward designing which may cost much efforts and may be not always accessible. An important solution to this problem is exploration, which is concerned about encouraging the agent to try unfamiliar states and thus to learn whether these states are of high value. However, naive exploration strategies adopted by many existing RL algorithms, such as ϵ-greedy (Mnih *et al.*, 2015) or random Gaussian noise (Lillicrap *et al.*, 2016), fail to work well in hard exploration environments, especially the ones with sparse reward signals.

To that end, researchers have proposed a range of intrinsic reward (or called exploration bonus) to encourage exploration, inspired by the concepts of curiosity and surprise (Schmidhuber, 1991; Itti and Baldi, 2006). For example, curiosity-driven intrinsic reward signals based on count (Kolter and Ng, 2009; Bellemare *et al.*, 2016; Machado *et al.*, 2018), information theoretic surprise (Mohamed and Rezende, 2015; Houthooft *et al.*, 2016) and prediction error (Pathak *et al.*, 2017; Burda *et al.*, 2019) all show promising results.

Motivated by recent progress in energy-based model (EBM), in this paper, we formulate the intrinsic reward from a energy view, and propose a novel Energy-Based Exploration (EBEx) framework. To be specific, we find that one can use the energy of the past experience as the intrinsic reward by constructing the cross entropy maximization objective, where agents aim to get away from historical states. Theoretically, we can apply any existing energy-modeling methods in EBEx but for generalization in high-dimensional and image-based state data, in this work, we use Energy-Based Generative Adversarial Networks (EBGAN) (Zhao *et al.*, 2016) as the backbone function to provide the value of the energy, where rarely seen states will be set with high energy and states in the replay buffer will be assigned with low energy. In addition, EBEx is an efficient RL framework with good modularity which can be easily extended to incorporate a variety of existing RL methods.

Research work for application of S.-T. Yau High School Science Award 2020

To illustrate the effectiveness of EBEx we conduct several quantitative and qualitative evaluation experiments in a sparse-reward 2D navigation environment, which is hard to explore for the agent and it can only learn a sub-optimal solution with a normal RL algorithm. By contrast, our experiments show that with EBEx, the agent finally learns to reach the optimal solution. To understand how EBEx framework provides good guidance on unseen states, we further visualize the changes of the sampled trajectories and the intrinsic reward during the training procedure.

The contributions of this paper are listed below: 1) we propose a novel framework named Energy-Based Exploration (EBEx) as a solution for the exploration of RL algorithms, especially in sparse-reward tasks; 2) we derive our algorithm by maximizing the distance between the current rollouts and the samples in the replay buffer, which provides theoretical guarantees for the effectiveness of our algorithm; 3) we evaluate our method in several sparse-reward tasks and show that EBEx can help the agent to explore by both quantitative and qualitative results.

2 Preliminaries

2.1 Markov Decision Process

The dynamics environment of Reinforcement Learning (RL) is normally formulated as a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \rho_0, r, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ represents the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the probability distribution of the state transition, $\rho_0 : \mathcal{S} \to [0, 1]$ is the initial state distribution, and $\gamma \in [0, 1]$ is the discounted factor. At timestep t, the agent utilizes its policy $\pi(a|s) : \mathcal{S} \times \mathcal{A} \to [0, 1]$ to make a decision of action a_t given the state s_t, and receives a reward signal r_t. Without ambiguity of the notation, we define the reward function as $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. For an arbitrary function $f : \langle s, a \rangle \to \mathbb{R}$, define $\mathbb{E}_\pi[f(s, a)] = \mathbb{E}_{s\sim P, a\sim\pi}[f(s, a)] \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right]$ as the expectation of f w.r.t the policy π, where $s_0 \sim \rho_0$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$. Therefore, the objective of the agent is to find a policy that maximizes the expectation of discounted accumulative reward as:

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[r(s, a)\right] \ . \tag{1}$$

Occupancy Measure The occupancy measure (Schulman *et al.*, 2015) can be seen as an unnormalized discounted density of the state-action pairs that are sampled by the agent:

$$\begin{aligned} \rho_\pi(s, a) &= \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = s|\pi) \\ &= \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi) = \pi(a|s)\rho_\pi(s) \ . \end{aligned} \tag{2}$$

With this definition, we are ready to write that $\mathbb{E}_\pi[\cdot] = \sum_{s,a} \rho_\pi(s, a)[\cdot] = \mathbb{E}_{(s,a)\sim\rho_\pi}[\cdot]$. We will denote $\rho_\pi^{s,a}$ as ρ_π in the following sections for simplicity, and we have $\rho_\pi \in \mathcal{D} \triangleq \{\rho_\pi : \pi \in \Pi\}$.

2.2 Exploration via Intrinsic Reward

Exploration in RL tasks requires agents to explore novel states as much as possible in order to achieve a better policy that can obtain more accumulated reward, especially when the raw reward signal is sparse and thus hard to learn. A common solution is to recognize the visiting frequency for certain states and to encourage the agent to enter such states via an *intrinsic* reward. Different from the extrinsic reward which is provided by the environment or the mechanism of the task, intrinsic reward is the one that agent provides for itself. Formally, denote the extrinsic reward r^e and the intrinsic reward r^i, the objective of the problem augments the additional exploration bonus term:

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[r^e(s, a) + \alpha r^i(s))\right] \ , \tag{3}$$

where α is the hyperparameter. It is worth noting that the intrinsic reward becomes less as the visiting frequency of some state becomes larger, thus the agent will not be encouraged by the exploration bonus to a state once it has encountered that state sufficient times.

2.3 Energy-Based Models

An energy-based model (EBM) (LeCun *et al.*, 2006) of a random variable $X \sim p(x)$ builds the density of data by estimating the energy function $E(x)$ using sample x in a Boltzmann distribution:

$$p(x) = \frac{1}{Z} \exp(-E(x)) \,, \tag{4}$$

where $Z = \int \exp(-E(x))dx$ is the partition function. Therefore, the energy function E can be seen as the unnormalized log-density of data.

3 Exploration via Estimating the Past Energy

In this section, we provide an energy-based perspective for encouraging agents to explore novel states that the agent seldom reached in the past trajectories.

3.1 Energy-Based Exploration

We start to present our energy-based exploration (EBEx) framework by introducing the basic idea of existing exploration works. As discussed before, exploration encourages agent to reach novel states. The novelty of states can be evaluated by the visiting count, as many previous works did (Strehl and Littman, 2005, 2008; Kolter and Ng, 2009). However, keeping exact visit counts is always impractical for most of problems, especially for problems with continuous or high-dimensional state space. To that end, Tang *et al.* (2017) utilized an auto-encoder with a hash function to reduce the dimension of state space. Bellemare *et al.* (2016) and Ostrovski *et al.* (2018) proposed to compute the pseudo-counts using a density model. Besides, information-theoretical methods can also be applied into modeling the information gain compared with the past experience (Mohamed and Rezende, 2015; Houthooft *et al.*, 2016). Recently, researchers has already recognize the ability of deep neural networks for modeling the distribution of the experience in order to judge the novelty by prediction error (Pathak *et al.*, 2017; Burda *et al.*, 2019). Motivated by these intuitive ideas, in this paper, we propose to take the advantage of the energy-based model (EBM) to model the density of the past experience.

Denote a replay buffer $\mathcal{B}$ that stores the experience samples, we model the state distribution $P_{\mathcal{B}}(a)$ as a Boltzmann distribution using an energy function $E(s)$:

$$P_{\mathcal{B}}(s) = \frac{1}{Z} \exp\left(-E_{\mathcal{B}}(s)\right) . \tag{5}$$

Remember our goal is to encourage the agent to reach novel states, in other words, trying to keep away from the past experience. This can be formulated as an optimization problem which maximizes the Cross Entropy (CE) between the past state distribution and the current state distribution $d_\pi(s)$ induced by the current policy π:

$$\max_\pi H(d_\pi(s), P_{\mathcal{B}}(s)) \,, \tag{6}$$

where $d_\pi(s) = (1-\gamma)\rho_\pi(s)$ is the normalized probabilistic distribution of the occupancy measure. An intriguing observation is that Eq. (6) can be further written as a discounted reward maximization problem that takes the energy as the intrinsic reward:

$$\begin{aligned} H(d_\pi(s), P_{\mathcal{B}}(s)) &= \sum_{s,a} d_\pi(s) \log \frac{1}{P_{\mathcal{B}}(s)} \\ &= (1-\gamma) \sum_{s,a} \rho_\pi \left(- \log \frac{(1-\gamma)e^{-E_{\mathcal{B}}(s)}}{Z'} \right) \\ &= (1-\gamma) \sum_{s,a} \rho_\pi \left(E_{\mathcal{B}}(s) + \log \frac{Z'}{(1-\gamma)} \right) \\ &= (1-\gamma) \mathbb{E}_\pi \left[E_{\mathcal{B}}(s) \right] + \text{const} \,. \end{aligned} \tag{7}$$

The additional constant can still be removed in the optimization problem, hence Eq. (6) is equivalent to the following problem:

$$\max_\pi \mathbb{E}_\pi \left[E_{\mathcal{B}}(s) \right] \,. \tag{8}$$

Algorithm 1 Energy-Based Exploration

Input: the state-action value network Q, policy network π, enerator network G, discriminator auto-encoder D, horizon T and replay buffer $\mathbf{M}$
Initialize Q,π, G and D with random weights, $\mathbf{M} \leftarrow$
for episode = $1, 2, \ldots, K$ **do**
  for $t = 1, 2, \ldots, T$ **do**
    Sample trajectories $\tau = \{s_0, a_0, \cdots, s_T, a_T\}$ and store into replay buffer $\mathbf{M}$.
    For every s_t in $\mathbf{M}$, compute $r_t^i(s_t)$ according to Eq. (12).
    if t is EBGAN's training step **then**
      Update generator G and discriminator D according to Eq. (10).
    end if
    Update policy π and value Q using any reinforcement algorithm, such as SAC.
  end for
end for

This fact remind us that if we want to keep away from the distributions of the past experience, we can instead tries to maximize the expected accumulated energy of the experience buffer w.r.t policy π.

Recall that exploration maximizes the intrinsic reward as exploration bonus, we find it naturally setting the state energy of buffer $E_{\mathcal{B}}(s)$ as the intrinsic reward. Therefore, the augmented objective Eq. (3) can be rewritten as

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[r^e(s,a) + \alpha E_{\mathcal{B}}(s)) \right] . \tag{9}$$

3.2 Past Energy Estimation from Historical Trajectories

As described above, our intrinsic reward function $r^i(s,a)$ is determined by $E_{\pi_{\mathcal{B}}}(s)$, a learned state energy function of the past experience. Therefore, in this section, we will elaborate on how to estimate $E_{\pi_{\mathcal{B}}}(s,a)$ from the historical experience.

Typically, any methods to training EBMs (LeCun *et al.*, 2006) can be used here to estimate the state energy. However, most of methods are limited in low state dimension (Hyvärinen, 2005; Kingma and Cun, 2010; Du and Mordatch, 2019), and thus are hard to be extended to high-dimensional environments such as image-based games.

As a particular solution to such tasks, in this paper, we consider to leverage Energy-Based Generative Adversarial Networks (EBGAN) (Zhao *et al.*, 2016), a scalable and efficient algorithm that takes the advantage of GAN to model the distribution of image-based data. EBGAN follows the basic idea of EBMs which is able to assign low values for data in the training set and higher values for data out of the training set via an auto-encoder based discriminator. As such, we can take the reconstruction error from the auto-encoder-based discriminator as the estimated energy and construct the intrinsic reward to help the agent explore in sparse reward environments.

Formally, let us denote the random variable $X = s \sim P_{\mathcal{B}}(s))$. Then, given a sample x from $\mathcal{B}$, a generated sample $G(z)$ and a margin m, similar as a normal GAN, we can construct an EBGAN model where the discriminator loss $\mathcal{L}_D$ and the generator loss $\mathcal{L}_G$ are defined as

$$\begin{aligned} \mathcal{L}_D &= D(x) + [m - D(G(z))]^+ \\ \mathcal{L}_D &= D(G(z)) \, , \end{aligned} \tag{10}$$

where $[\cdot]^+ = \max(0, \cdot)$. In detail, the discriminator D is structured as an auto-encoder:

$$D(x) = \|Dec(Enc(x)) - x\| \, . \tag{11}$$

Therefore, for a data instance sampled from the buffer $\mathcal{B}$, the discriminator is required to reconstruct the sample via the auto-encoder; while the unseen data is encouraged to reconstructed with an error upper-bounded by m, which can be further used to compose the intrinsic reward function. With such an energy estimation model, we build the overall Energy-Based Exploration (EBEx) architecture shown in Fig. 1 and list the step-by-step algorithm in Algo. 1.

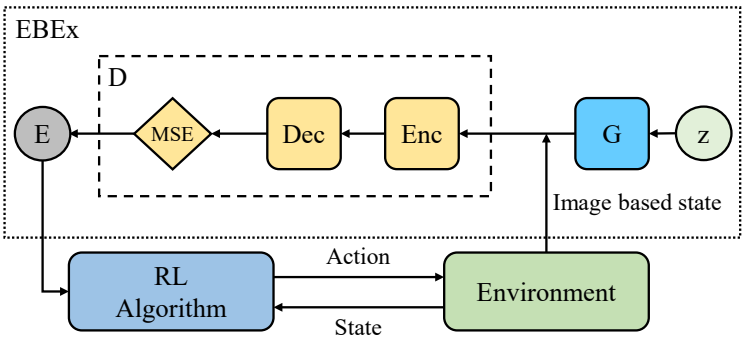


Figure 1: The Energy-Based Exploration (EBEx) framework where we illustrate EBGAN to provide the energy for example. EBEx is free to combine with any reinforcement learning algorithm and is able to handle high-dimensional data like images with a powerful energy-based model.

3.3 Further Implementation Details

So far we have presented our energy-based exploration framework where we utilize EBGAN as the estimator for high-dimensional image-based state data and construct the intrinsic reward function with the reconstruction error of the discriminator in EBGAN. Nonetheless, there are also many details to be carefully considered for implementing an effective exploration algorithm, which are described as follows.

Multiple Value Heads for Q function. Since the agent now receives two kinds of rewards, i.e., the extrinsic reward that is acquired from the environment and the intrinsic reward that is generated by the energy estimator to encourage exploration. One can simply use one value network for predicting the sum of the accumulative rewards $\mathbb{E}[\sum_{t=0}^{T}\gamma^t(r_t^e + \alpha r_t^i)]$, however, we refer to Burda *et al.* (2019) and Hong *et al.* (2019) that a more effective training design is to separately estimate the accumulative extrinsic reward $V^e = \mathbb{E}[\sum_{t=0}^{T}\gamma_1^t(r_t^e)]$ and the accumulative intrinsic reward $V^i = \mathbb{E}[\sum_{t=0}^{T}\gamma_2^t(r_t^i)]$, where T is the horizon. The rationale here is that the extrinsic reward is stationary since the environment stays along the whole training procedure, while the intrinsic reward actually varies as the agent's exploration proceeds.

Decaying Intrinsic Weights. Since our EBM is trained based on the experience data from the replay buffer, which only stores limited number of samples, a problem rises that the EBM may provide a high energy from an observed historical state which the agent has encountered a long time ago. A simple but effective solution is to decrease the weight of the intrinsic reward α every time the extrinsic reward is non-zero. The decreasing size is proportional to the extrinsic reward, i.e., $\alpha = \alpha * \mathbb{E}r_i$ Therefore, once the agent has reached the place and find a near-optimal policy, the intrinsic part will be released.

Surrogate Intrinsic Reward Function. Another important concern is about the intrinsic reward function. Instead of directly using the energy value $E(x)$ as the intrinsic reward function, we have more choices by selecting different surrogate reward function such that $r^i(s) = h(E(s))$ to adapt to different environments, where h is a mapping function from the energy of the state to the intrinsic reward value. Notice that if h is a monotonically increasing linear function, we only make translation or scaling transformation on the energy outputs without changing the optimal solution of the optimization problem Eq. (8).

As discussed in Kostrikov *et al.* (2018), the property of the task highly affects the reward design. For example, a positive reward encourages the agent to "survive" as long as possible in a given environment, while a negative one can lead to better performance in a "per-step-penalty" style environment. In our work, as most of our testing environments belong to the "surviving" tasks and the energy value $E(x)$ out of the discriminator is actually bounded in $[0, m]$, we choose to formulate h as

$$h(x) = \frac{1}{m}x \ , \tag{12}$$

since we try to normalize the intrinsic reward into $[0, 1]$. Such a formulation shows good results in our qualitative experiments as will be shown in Section 5.

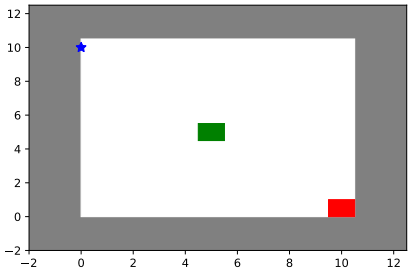


Figure 2: Illustration of the 2D navigation environment, where the agent starts from the top left corner (blue point) and sample in the world. There are small rewards (1) in the middle (green area) and large rewards (10) in the bottom right corner (red area).

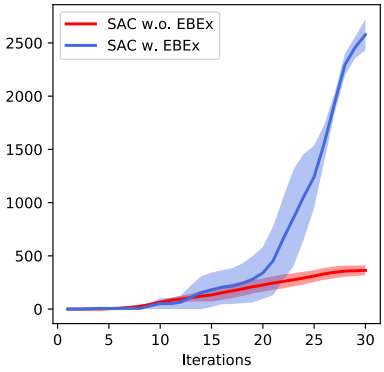


Figure 3: Learning curves in the 2D navigation environment, where the x-axis represent the training iteration and the y-axis denote the cumulative extrinsic reward gained from the environment. It is obvious that SAC with EBEx help agents reach the area with high rewards while SAC get stuck in the sub-optimal area.

4 Related Work

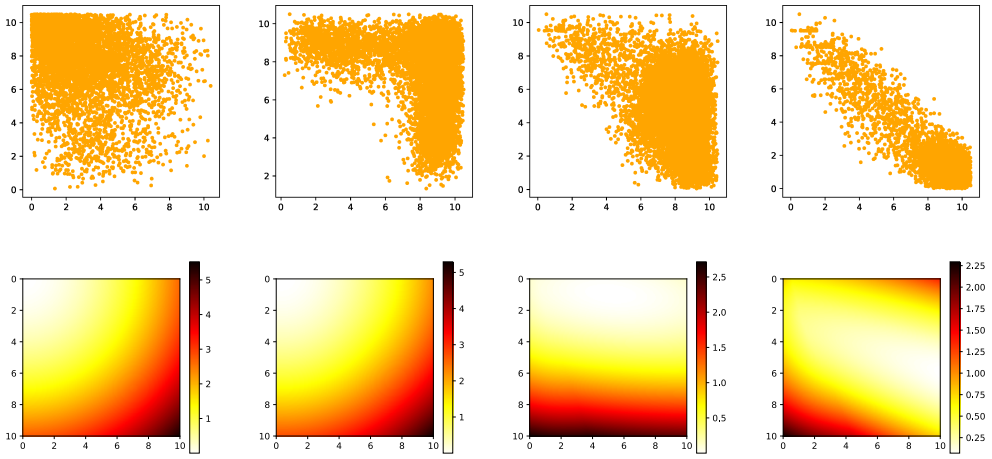
The problem of exploration in Reinforcement Learning (RL) has been an active research topic for decades and there are various solutions that have been investigated for encouraging the agent to explore novel states. Naive exploration strategies explore with simple heuristics, such as ϵ-greedy, Boltzmann exploration and random noise methods. However, such random-walk like exploration strategies usually fail to work well in sparse reward tasks, i.e., the agent will not get any reward signals in the most of times. To that end, researchers have been investigated on intrinsic reward (or called exploration bonus) as a general solution to help explore novel states so that the agent can more possibly achieve the optimal solution (Oudeyer and Kaplan, 2009; Schmidhuber, 2010).

Specifically, a straightforward way of designing intrinsic reward is to count the state/state-action visited frequency to calculate the bonus, and they practically work well in discrete and simple environments (Strehl and Littman, 2005, 2008; Kolter and Ng, 2009). To alleviate the challenge in continuous state spaces, Tang *et al.* (2017) utilized a hash function combined with an auto-encoder, while Bellemare *et al.* (2016) and Ostrovski *et al.* (2018) proposed to introduce the pseudo-counts with a density model.

Besides, since the novelty of unseen states can also be measured by their uncertainty, a lot of works, therefore, leverage information theory metrics to compute the intrinsic reward, such as information gain (Houthooft *et al.*, 2016), empowerment and mutual information (Mohamed and Rezende, 2015; Still and Precup, 2012) etc.

Recently, there are works based on prediction-error further developing the advances of curiosity-driven exploration, where the basic assumption is that the neural network model can to-some-extent memorize the visited state with certain generalization, so that the visited or similar state will be predicted (or scored) more accurately. Pathak *et al.* (2017) compared the difference between a predicted state feature predicted and the real state, while Burda *et al.* (2019) used the prediction error between a trained model and a fixed randomly initialized neural network.

Different from above heuristically designed intrinsic rewards, our EBEx begins with an optimization problem on maximizing the KL divergence between the current occupancy and the historical one, then naturally derives an energy-form intrinsic reward. Although we show similar properties that provide high rewards on novel states and low rewards on familiar ones, we interpret such a problem from a statistics point of view with mathematical guarantees, which builds the key technical contribution of this paper.



(a) 0 Iterations (b) 1000 Iterations (c) 2000 Iterations (d) 3000 Iterations

Figure 4: Trajectory points in past 10,000 timesteps sampled by agents (top) and heat maps of the intrinsic reward provided by EBGAN (bottom) at different training iterations in the 2D navigation environment. The agent starts from the top-left corner and sample in the grid world, which can obtain a small reward from the center region and a larger reward from the bottom right region. The horizontal and the vertical axis denote the *state space*. It is obvious that EBGAN learns a well intrinsic reward for unseen states.

5 Experiments

In this section, we conduct several experiments from different dimensions to answer the following research questions:

RQ1: Does EBEx framework indeed provide good guidance on unseen states?

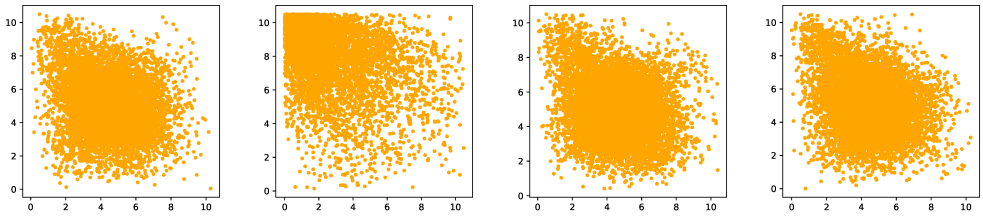
RQ2: Can EBEx work better than normal RL algorithms?

To answer these two research questions, we conduct qualitative and quantitative experiments in a sparse-reward 2D navigation task, where the environment is a 2-dimensional continuous grid world as shown in Fig. 2. The agent begins at the top-left corner and tries to get as much reward as it can. There are small rewards (1) around the center and a larger reward (10) at the bottom right corner. We assume that without appropriate exploration strategy, the agents will stuck in the center cannot reach the high-reward area. We compare the performance of EBEx and a regular RL algorithm, i.e., soft actor-critic (Haarnoja *et al.*, 2018).

Training Details. We apply soft actor-critic (SAC) (Haarnoja *et al.*, 2018) RL algorithm for 2D navigation task. The learning rates for the agent and the EBGAN are 2.5×10^{-4} and 5×10^{-6}, respectively. We train EBGAN every 500 policy training times using all data upon the replay buffer.

Sparse-Reward 2D Navigation We test EBEx in a sparse-reward 2D navigation task to evaluate whether EBEx can provide good guidance on unseen states and is able to help the agent learn a better policy. We first conduct quantitative experiments to evaluate the performance of the agent equipped with EBEx compared with a normal SAC algorithm. We show the learning curves of the cumulative extrinsic rewards in Fig. 3, which indicates that EBEx is capable of helping agent find a better policy in a sparse reward environment.

Besides, we also conduct qualitative experiments, where we visualize the historical trajectories in the past 10,000 timesteps and show how EBGAN guides the reward. As shown in Fig. 4, the agents starts from the top-left corner and sample in the world, and EBGAN provides much more energy in less-visited states compared with the more frequently-encountered ones. As the training proceeds, the agent finally learns to reach the bottom right corner which is associated with a higher reward. However, without EBEx, the agent loses the optimal policy and stucks in the middle region of the environment, as illustrated in Fig. 5.



(a) 0 Iterations (b) 1000 Iterations (c) 2000 Iterations (d) 3000 Iterations

Figure 5: Trajectory points in past 10000 timesteps sampled by agents **without** EBGAN in different training iterations. The agent starts from the top right corner, gets stuck in the middle of the environment and fails to get higher rewards in the bottom right corner.

6 Conclusion

In this paper, we proposed a novel exploration framework named Energy-Based Exploration (EBEx), which has shown its potential to address the exploration challenge in deep reinforcement learning by estimating the energy of the past trajectories. We implemented EBEx with soft-actor critic (SAC) and test our algorithm on a sparse-reward 2D navigation task, where through qualitative and quantitative experiments we show that EBEx is capable of providing reasonable intrinsic reward signals and help the agent achieve the optimality in a more sample-efficient manner.

For future work, we plan to deploy EBEx on more complex reinforcement learning environments to further investigate how to design good energy models to better guide the agent to efficiently explore the environment.

References

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *NeurIPS*, pages 1471–1479, 2016.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *ICLR*, 2019.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems*, pages 3603–3613, 2019.

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.

Weijun Hong, Menghui Zhu, Minghuan Liu, Weinan Zhang, Ming Zhou, Yong Yu, and Peng Sun. Generative adversarial exploration for reinforcement learning. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*, pages 1–10, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *NeurIPS*, pages 1109–1117, 2016.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Laurent Itti and Pierre F Baldi. Bayesian surprise attracts human attention. In *NeurIPS*, pages 547–554, 2006.

Durk P Kingma and Yann L Cun. Regularized estimation of image statistics by score matching. In *Advances in neural information processing systems*, pages 1126–1134, 2010.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *ICML*, pages 513–520. ACM, 2009.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *ICLR*, 2016.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. Count-based exploration with the successor representation. *arXiv preprint arXiv:1807.11622*, 2018.

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, 2015.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *NeurIPS*, pages 2125–2133, 2015.

Georg Ostrovski, Marc G Bellemare, Aaron van den Oord, and Rémi Munos. Count-based exploration with neural density models. *ICML*, 2018.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *FRONT NEUROROBOTICS*, 1:6, 2009.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *CVPRW*, pages 16–17, 2017.

Jürgen Schmidhuber. Curious model-building control systems. In *IJCNN*, pages 1458–1463. IEEE, 1991.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Autom. Control*, 2(3):230–247, 2010.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, volume 37, pages 1889–1897, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *THEOR BIOSCI*, 131(3):139–148, 2012.

Alexander L Strehl and Michael L Littman. A theoretical analysis of model-based interval estimation. In *ICML*, pages 856–863. ACM, 2005.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *J. Comput. Syst. Sci*, 74(8):1309–1331, 2008.

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

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *NeurIPS*, pages 2753–2762, 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

Acknowledgements

I have been deeply interested in the field of reinforcement learning since I started to learn machine learning in 9thgrade. I acquired the knowledge of baseline algorithms of RL by myself， including Markov Decision Processes, Actor-Critic, Policy Gradient, etc. In this competition, with the help of my tutor Dr. Zhang Weinan, I systematically learned the advanced RL algorithms, and carefully studied the previous researches of the algorithms that use the intrinsic reward to encourage exploration. I found that the traditional count-based methods could not be used in various situations of continuous action and/or state spaces and form my idea of this research.

In this study, I spent one month to study the advanced algorithm of reinforcement learning, select a topic for my research, and complete the derivation of the theoretical part. Then, I spent two months to do the coding for my experiment and finish the paper.

I am sincerely thankful to Apex Lab of Shanghai jiaotong university for the technical support.

I would like to express my special thanks to gratitude to my tutor Dr. Zhang Weinan for his great help and guidance during the whole research program, especially in the mathematical deduction.

I would like to convey my respectable thanks to Mr. Chen Yong, who is the vice-president of World Foreign Language Academy, for inspiring me to explore the computer science world.

I would also like to thank my family for their support and encouragement during the most difficult period of my research.

Finally, I greatly appreciate Yau-Science Award for providing me, a high school student who loves computer science and machine learning, with the opportunity to do the research. It has benefited me a lot.

Thanks again to Dr. Zhang Weinan, Mr. Chen Yong, and my family for their support and help!

致谢信

我从初中三年级了解机器学习以来，便一直对机器学习中强化学习这一领域有着浓厚的兴趣。然而在自己学习的过程中，更多接触的都是强化学习的基线算法，包括经典的马尔科夫决策过程、Actor-Critic 算法、策略梯度算法等等。在这次比赛中，张伟楠导师带我研究了学术界前沿的各类强化学习改进算法，仔细了解了前人为强化学习设计的十分重要的内在奖励机制的结构与相关算法。由于研究中，我发现传统的计数式内在奖励无法使用在各类连续样本空间、动作空间的环境中，故提出了本论文中的方法来解决这一问题。

本次研究中，我使用 1 个月了解了强化学习的前沿算法，确定了选题，并完成了理论部分的推导，使用 2 个月完成了实验代码的编写，与论文的书写。首先尤其感谢上海交通大学 Apex 实验室提供的服务器资源，使我研究中的实验部分可以进行，也非常感谢张伟楠导师在我的研究过程中对我持续的帮助。张伟楠导师为上海交通大学副教授，他在本课题的数学推导部分给予我指导、在实验部分遇到一些很棘手的漏洞时与我共同解决。同时，张教授对我研究有着非常严格的要求，使我的学术能力得到充分锻炼。感谢世外中学陈勇校长从我进入初中以来对我的培育，您让我有幸踏入计算机科学的殿堂并致力于将其发展为我的终生事业。同样感谢我的家人在我研究最困难的时候给予我的支持与鼓励。你们使我得以坚持完成本次研究。最后感谢丘成桐科学奖为我这个热爱计算机与机器学习的高中学生提供的机会，本次研究的过程使我受益匪浅。

再次感谢张伟楠导师、陈勇校长与我的家人对我的支持与帮助！

丘成桐中学科学奖-学术诚信声明

本参赛团队郑重声明：

1. 参赛团队提交的参赛队员和指导老师信息完整且属实无误。
2. 所提交的研究报告是在指导老师指导下进行的研究工作和取得的研究成果。
3. 尽本团队所知，除文中加以标注和致谢中所罗列的内容外，研究中不包含其他人已经发表或撰写过的研究成果，不存在代写或其他违规行为。

以上，若有不实之处，本人愿意承担一切相关责任，并服从丘成桐中学科学奖组织委员会的裁决。

参赛学生（签字）：

本校指导老师（签字）：

学校名称（加盖学校或教务处公章）：

外校指导老师（签字）：

单位名称（加盖单位公章）：

2020 年 9 月 15 日