Exploration via State Influence Modeling

Yongxin Kang[1,2*], **Enmin Zhao**[2,1*], **Kai Li**[2], **Junliang Xing**[2†]
[1] School of artificial intelligence, University of Chinese Academy of Sciences
[2] Institute of Automation, Chinese Academy of Sciences
{kangyongxin2018, zhaoenmin2018, kai.li}@ia.ac.cn, jlxing@nlpr.ia.ac.cn

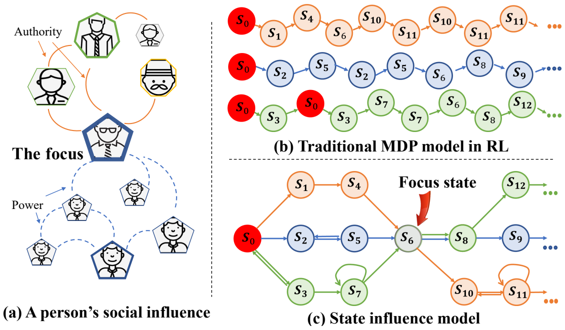
Abstract

This paper studies the challenging problem of reinforcement learning (RL) in hard exploration tasks with sparse rewards. It focuses on the exploration stage before the agent gets the first positive reward, in which case, traditional RL algorithms with simple exploration strategies often work poorly. Unlike previous methods using some attribute of a single state as the intrinsic reward to encourage exploration, this work leverages the social influence between different states to permit more efficient exploration. It introduces a general intrinsic reward construction method to evaluate the social influence of states dynamically. Three kinds of social influence are introduced for a state: *conformity*, *power*, and *authority*. By measuring the state influence, agents quickly find the focus state during the exploration process. The proposed RL framework with state influence evaluation works well in hard exploration task. Extensive experimental analyses and comparisons in *Grid Maze* and many hard exploration Atari 2600 games demonstrate its high exploration efficiency.

Introduction

Reinforcement learning (RL) in hard exploration tasks with sparse rewards is an essential problem in artificial intelligence. Unlike typical RL problems, hard exploration tasks with sparse rewards often consist of two stages. First, there is a long period of exploration before the agent obtains a new reward, which we term the no-reward exploration stage. Second, after the agent obtains some local rewards, it follows a process of experience utilization and continuing to explore, which we term the local-reward exploitation stage. In this paper, we focus on the first and more difficult no-reward exploration stage.

During the no-reward exploration stage, traditional RL algorithms based on value function (Mnih et al. 2015; Van Hasselt, Guez, and Silver 2016) or policy gradient (Schulman et al. 2015, 2017) often get trapped in some confusing states because the state value they used is only evaluated by the reward. Since only a few states contain rewards, the agent cannot distinguish between those no-reward states, even if it have experienced them many times.


Authority
The focus
Power
s_0 s_1 s_4 s_6 s_{10} s_{11} s_{10} s_{11} s_{11}
s_0 s_2 s_5 s_2 s_5 s_6 s_8 s_9
s_0 s_3 s_0 s_3 s_7 s_7 s_6 s_8 s_{12}
Focus state
s_0 s_1 s_4 s_2 s_5 s_6 s_8 s_9 s_{12} s_3 s_7 s_{10} s_{11}

(a) A person's social influence

(b) Traditional MDP model in RL

(c) State influence model

Figure 1: (a) The social influence of one person in a social network. Three characteristics are often used to represent a person's social influence: 1) the size of the node indicates his conformity, 2) the blue dotted line indicates his power which is usually related to the number of connections within his group, and 3) the solid orange line indicates the authority, that is, how many followers a person has. The most influential person in a social network is called the *focus*. Similarly, we regard the states in (b) the traditional MDP model in RL as the nodes in (c) social network and obtain the focus state by measuring the states' social influence. The focus state will be related to more states, and the exploration of it can accelerate the agent's cognition of the environment.

To deal with the hard exploration tasks with sparse rewards, many previous works try to imitate expert demonstrations (DQFD) (Hester et al. 2018) or their own successful experience (SIL) (Oh et al. 2018). In practice, however, expert demonstrations are often unavailable, and SIL focuses on the second stage, where the agent has already received local rewards. In the no-reward exploration stage, HER (Andrychowicz et al. 2017) randomly sets virtual goals from the experience replay buffer, regardless of which experience might be the most valuable. So it suffers from low sampling efficiency. To explore more meaningful directions, some researchers design intrinsic rewards based on the curiosity of states, where curiosity is measured by prediction errors (Pathak et al. 2017), reachability (Savinov et al. 2019) or pseudo-count (Bellemare et al. 2016). Although have achieved some success, they only consider the attributes of the state itself but ignore the relationships between states.

[*] Equal contribution, † corresponding author.

“But the human essence is no abstraction inherent in each single individual. In its reality it is the ensemble of the social relation” (Marx and Engels 1969). Social influence refers to the change of a person’s behavior after an interaction with other people or organizations. It consists of the process by which the individual opinions can be changed by the influence of another individual or other individuals (Friedkin 2006). Based on these considerations, when estimating the value of a state in RL, it’s better to consider the relationship with other states besides its own attributes.

Inspired by the concept of social influence in the social networks analysis community (Friedkin 2006), we regard each state as an individual, the relationship between states as a link in social networks, and the exploration process as a series of accesses to the opinion leader states (Figure 1). A general intrinsic reward construction method is thus introduced to measure the social influence of states dynamically, which is termed as Social Influence (SI) based intrinsic reward function. It comprises three kinds of social attributes: conformity, power, and authority. In particular, the conformity measures how often a state is visited, the power measures the relations with its former states in the MDP process, and the authority measures the relations it might have with its followers. By evaluating these social attributes of states, the agent can find the focus state and exploit this information to accelerate the exploration process. We form a general RL framework using this SI-based intrinsic reward function. The new RL framework applies to both value based RL algorithms like DQN (Mnih et al. 2015), Dueling DQN (Wang et al. 2016), policy gradient based RL algorithms like PPO (Schulman et al. 2017), TRPO (Schulman et al. 2015), and the hybrid one like A3C (Mnih et al. 2016).

For a series of hard exploration tasks like *Grid Maze* and *Montezuma’s Revenge*, we develop corresponding learning algorithms based on the proposed RL framework. The results demonstrate that, with the introduction of social influence, all the evaluated algorithms significantly improve their learning efficiency and quickly accomplish the goal of each task. In specific, on the *Grid Maze* game, the proposed method distinctly reduces the total exploration steps compared with the classical Q-learning based method. On the *Montezuma’s Revenge*, compared with existing algorithms, the proposed method converges faster and obtains higher scores.

To summarize, the main contributions of this work are listed as follows in threefold:

- We point out that in the no-reward exploration stage, it is not enough to define intrinsic rewards based on the attributes of the state alone. The relationship between different states is introduced as a new part of intrinsic rewards.
- According to the measurement of individual social influence in social network analysis, a generalized intrinsic reward function is defined for each state, including the attributes of the state itself and the relationships between the state and others.
- A new RL framework of SI-based intrinsic reward function is proposed and applied to Q-learning and A2C to improve the performance in some hard exploration games.

The source code, trained models, and all the experimental results will be released to facilitate further studies on reinforcement learning in hard exploration tasks.

Background

To expatiate the proposed intrinsic motivation model, we first introduce some background knowledge on basic reinforcement learning, intrinsic reward, and social influence.

Basic RL. The standard RL formulation involves an agent interacting with an environment. An MDP is a tuple $M = \langle S, A, R, T, \gamma\rangle$, consisting of a set of states S, a set of actions A, a reward function $R : S \times A \to \mathbb{R}$, a transition probability model $T(s_{t+1}, r_{t+1}|s_t, a_t)$, and a discount factor $\gamma \in [0, 1]$. A policy π maps a state to an action, $\pi : S \to A$. An episode starts with an initial state s_0, and at each timestamp t, the agent chooses an action $a_t = \pi(a|s_t)$ based on the current state s_t. The environment produces a reward r_{t+1} to the agent, which reaches to next state s_{t+1} sampled from the distribution $T(s_{t+1}, r_{t+1}|s_t, a_t)$. The reward might be discounted by a factor γ at each timestamp, and the goal of the agent is to maximize the accumulated reward,

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \tag{1}$$

Intrinsic reward. Intrinsic rewards become critical when extrinsic rewards are sparse (Pathak et al. 2017). They guide the agent based on the change in prediction error or learning progress (Bellemare et al. 2016; Schmidhuber 1991; Oudeyer, Kaplan, and Hafner 2007). If $e_n(A)$ is the error made by the agent at time n over some event A, and $e_{n+1}(A)$ the same one after observing a new piece of information, then the learning progress is $e_n(A) - e_{n+1}(A)$. To further quantify the learning process, researchers provide an information gain related method to explain the intrinsic reward (Bellemare et al. 2016). At each timestamp, the agent is trained with the reward $r_t = e_t + \beta i_t$, where e_t is the extrinsic reward provided by the environment, i_t is the intrinsic reward generated by the agent, and $\beta > 0$ is a scalar balancing between the intrinsic and extrinsic rewards (Taiga et al. 2020). The overall optimization problem solves the following Bellman equation,

$$V(s) = \max_{a \in A} [e_t + \beta i_t + \gamma E_\pi [V(s')]]. \tag{2}$$

Social influence. In social network analysis, social influence of a node is often characterized by three main features (Friedkin 2006): 1) *conformity*, that occurs when an individual expresses a particular opinion in order to meet the expectations of a given other, though he/she does not necessarily hold the belief that the opinion is appropriate; 2) *power*, that is the ability to force someone to behave in a particular way by controlling his/her outcomes; and 3) *authority*, that is the power that believed to be legitimated by those who are subjected to it.

The Model

There are two main obstacles in the process of exploring a sparse reward environment. The first obstacle comes from the vast state space. The other one comes from the uncertainty across states. Too many non-reward states and unknown transitions between states cause great confusion to agents and slow down the exploration process. Figure 1(b) is a simple example that regards state as the node and transition as the edge in a dynamic directed graph. When the number of states increases with the exploration process, the number of nodes in the graph increases and the number of possible edges also increases. This brings great trouble to the exploration task. Social networks can also be modeled by dynamic graphs (Figure 1(a)). However, as the population size increases, it can quickly find an effective way to spread information without confusion from the uncertainty. This benefits from the efficient utilization of the focus person and the modeling of the relationship between persons. Based on these observations, we introduce the concepts from the social network into the process of exploring to narrow the exploration space and reduce the uncertainty.

Intuitively, as the exploration task shown in Figure 1(c), if you want to explore state S_8, you must have a corresponding state S_6 appeared. Whether the arrival and re-exploration of a state are instructive to the policy improvement not only relates to the current state itself but also how to reach the state and how much potential the state can get in the future. The exploration process thus can not just be summarized as "explore what surprises the agent" as most previous methods did. It should take "exploit what influences the environment" into account. Just like the focus person in the social network in Figure 1(a), his importance to information broadcasting not only relates to his conformity (occurrence frequency), but also his power and authority (connections with other people). Therefore, we give a more reasonable formulation of intrinsic reward according to the concepts in social influence analysis, which consists of the state's characteristics and the relationship between states.

State Influence

Definition 1 (Conformity function) *We define the conformity function on the state space S, $f_C : S \rightarrow R$ mapping the state to a conformity level. $\forall s_i, s_j \in S$, if $f_C(s_i) < f_C(s_j)$, we say state s_j is visited more often than state s_i.*

In the study of social network, conformity indicates that the opinion of an individual is the same as that of the majority of people, or whether the opinion of the individual is expected by the public. In the exploration task of RL, we regard each state encountered as an individual and its visited characteristics as opinions. The conformity function f_C measures how often a state is visited. In social networks, the focus person will not blindly follow others, which means, less conformity. Meanwhile, in our exploration problems, we should avoid accessing the already familiar states.

In the episodic RL, we formalize the conformity as

$$f_C(s_i) := p(s_i). \tag{3}$$

Definition 2 (Power function) *We define the power function f_P on the state space S, $f_P : S \rightarrow R$, and it maps the state to a power level. $\forall s_i, s_j \in S$, if $f_P(s_i) < f_P(s_j)$, we say that there are more states which can lead the agent to state s_j than to s_i.*

Powerful person influences society by controlling other people's labor or information output and the power is the embodiment of compulsion. In this paper, we define the power of a state as how many states must achieve the ultimate goal through information exchange with the current state. The power function measures the relationship among states, which can be regarded as a structure of the partially explored environment. In episodic RL, we formalize the power function as

$$f_P(s_i) := \int_{s_k \in S_p} p(s_i|s_k) ds_k, \tag{4}$$

where the S_p is the states' set which appear before s_i.

Definition 3 (Authority function) *We define the authority function on the state space S, $f_A : S \rightarrow R$ mapping the state to an authority level. $\forall s_i, s_j \in S$, if $f_A(s_i) < f_A(s_j)$, we say that state s_j can lead to a more diverse state space, i.e., having more states followed.*

A person should influence more communicators through his "authority". It is necessary for individuals to shape their own information and then influence others. We define the authority of state as the influence of the absence of one state on other states The authority function measures another kind of relationship among states, which indicates how many future states can be reached from a specific state. In episodic RL, we formalize the authority function as

$$f_A(s_i) := \int_{s_k \in S_a} p(s_k|s_i) ds_k, \tag{5}$$

where the set S_a contains states after state s_i in the currently known trajectory.

Combined the three functions, the Social Influence (SI) based intrinsic reward function can be represented as:

$$i^{SI}(s) \triangleq \Psi(f_C(s), f_P(s), f_A(s)), \tag{6}$$

where Ψ can be a function with parameters or a task related deterministic function.

With these characteristics to describe the social influence of states, we combine the SI-based intrinsic reward function with the extrinsic reward $e_t(s)$ to evaluate the reward, $r_t = e_t + \beta i_t^{SI}$, where β is a balancing factor between the extrinsic environment value e_t and the proposed instinct reward $i^{SI}(s)$. In all the evaluations conducted in experiment part, β is experimentally set to 1. And then, in substitution of traditional formulation G_t in Equation 1, we evaluate the SI-based accumulated reward by

$$G^{SI}(S_t) = \sum_{k=0}^{\infty} \gamma^k (\beta * i^{SI}(S_{t+k+1}) + e(S_{t+k+1})), \tag{7}$$

where $0 < \gamma < 1$ is the discount factor.

According to $G^{SI}(S_t)$, we propose SI-based value function, which evaluates the expectation of the total value of a state S_t following current policy π.

$$\begin{aligned} V_\pi^{SI}(S_t) &= E_\pi[G^{SI}(S_t)|S_t = s] \\ &= E_\pi[\sum_{k=0}^{\infty} \gamma^k (\beta * i^{SI}(S_{t+k+1}) + e(S_{t+k+1}))]. \end{aligned} \tag{8}$$

Algorithm 1 SI-based RL framework

```
1: Initialize policy $\pi_0$, exploration steps $k_0$, and expansion interval $\delta$.
2: while not done do
3:     Generate $M$ trajectories in local environment by $\pi_i$.
4:     Evaluate $i^{SI}(s)$ of the known state by $i^{SI}(s) = \Psi(f_C(s), f_P(s), f_A(s))$.
5:     while $\pi_i$ not convergence do
6:         Update $\pi_i$ using reward $r_t = e_t + \beta i_t^{(SI)}$ with an RL algorithm, e.g., Eqn. (9) or Eqn. (10).
7:     end while
8:     Expand the scope of exploration by increasing $k_{i+1} = k_i + \delta$.
9:     if Agent get local rewards then
10:        Any local-reward exploitation stage algorithms will be used.
11:    end if
12: end while
```

Compared with the definition of the traditional $V(s)$ in Section , $V^{SI}(s)$ preserves extrinsic reward $e(s)$ and introduces intrinsic rewards $i^{SI}(s)$. This form of $V(s)$ can give general guidance without arriving at the reward state, which makes it an effective attempt to overcome the hard situation with no reward. Moreover, it inherits the characteristics of $i^{SI}(s)$, measures the structural information of the environment through the relationships between states, and its evaluation of state value is more instructive to the exploitation in the exploration process. It provides a generalized intrinsic rewarding mechanism for hard-exploration tasks in sparse reward environments and can be integrated with different RL algorithms. We will describe its iterative nature in the following and incorporate it into the traditional Q-learning and A2C frameworks.

SI-based RL Framework

The social influence based value function can be embedded into any RL algorithm that involves value iteration. The key idea of the SI-based RL framework is that an agent utilizes the current $i^{SI}(s)$ and π to explore a desired part of the environment, progressively broaden the exploration scope, and update the social influence of states in turn with newly acquired information about the environment. This process is iteratively advanced until the agent finds the target or receives extrinsic reward signals. Algorithm 1 shows the detail of SI-based RL framework.

The main advantages are in twofold: 1) RL framework with $i^{SI}(s)$ expands the agent's horizon incrementally, shown in Step 8 in Algorithm 1; and 2) the evaluation function $V^{SI}(s)$ contains not only the reward but also the structure information among states, shown in Eqn. (8). With the first advantage, agents can gradually explore the unknown environment and get sub-optimal solutions in each local context. With the second one, the agent can exploit the intrinsic state relationships even though there are no reward signals. In this way, the agent obtains a reasonable explore direction before getting a reward in the hard exploration environment. We incorporate it into the traditional RL algorithms to construct different SI-based RL algorithms, such as SI-based Q-learning and A2C.

Applying $i^{SI}(s)$ to traditional value based methods, such as Q-learning, we get the update formula:

$$
\begin{aligned}
Q^{SI}_{\pi} &= E_{\pi}[G^{SI}(S_t)|S_t = s, A_t = a] \\
&= \sum_{s',r} T(s', r|s, a)[\beta i^{SI}(s') + e(s') + \gamma V^{SI}_{\pi}(s')].
\end{aligned} \tag{9}
$$

State influence can also accelerate the exploration process of policy gradient based methods, such as AC. When apply to the One-step Actor-Critic, the parameters update rule is as follows:

$$
\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha(G^{SI}_{t:t+1}(S_t) - \hat{V}^{SI}(S_t, \eta)) \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)} \\
&= \theta_t + \alpha \delta_t \frac{\nabla \pi(A_t|S_t, \theta_t)}{\pi(A_t|S_t, \theta_t)},
\end{aligned} \tag{10}
$$

where $\hat{V}^{SI}(S_t, \eta)$ is the evaluation of SI-based value function and

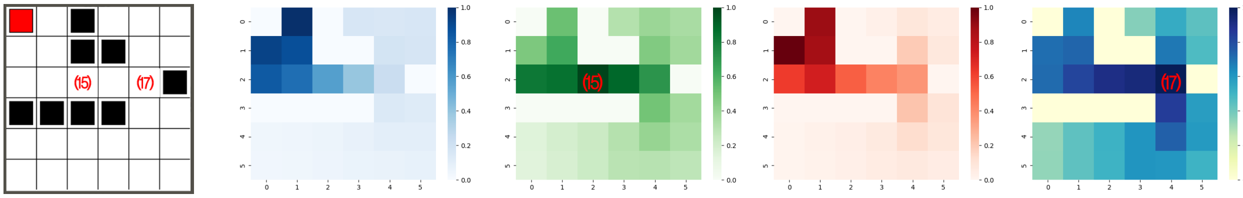
$$
\begin{aligned}
\delta_t &= G^{SI}_{t:t+1}(S_t) - \hat{V}^{SI}(S_t, \eta) \\
&= \beta i^{SI}(S_{t+1}) + e(S_{t+1}) + \gamma V^{SI}_{\pi}(S_{t+1}) - \hat{V}^{SI}(S_t, \eta).
\end{aligned} \tag{11}
$$

Connections to Other Intrinsic Motivation Methods

The state influence modeling is a generalization of the existing intrinsic reward construction methods. For example, the $\hat{N}_t$ in pseudo-count (PSC) (Bellemare et al. 2016) is a kind of conformity, and the core idea of the curiosity algorithm is to avoid large conformity. They formalize the intrinsic reward as $i^{PSC}(s_t) \triangleq (\hat{N}_t(s_t))^{-1/2}$. If we ignore power and authority, assume that all states are independent, and define our conformity as $f_C(s_i) \triangleq (\hat{N}_t(s_t))^{-1/2}$, $i^{SI}(s)$ will degenerate to pseudo-count in (Bellemare et al. 2016).

Similarly, both $i^{ICM}(s)$ (Pathak et al. 2017) and $i^{RND}(s)$ (Burda et al. 2019) are constructing curiosity by modeling the relationship between the current state s_i and the next state s_{i+1}, where they formulize the intrinsic reward as $i^{ICM}(s_t) \triangleq ||\hat{\Phi}(s_{t+1}) - \Phi(s_{t+1})||_2^2$ and $i^{RND}(s_t) \triangleq ||\hat{f}(s_t;\theta) - f(s_t)||_2^2$ (the $\hat{\Phi}(\cdot)$ and $\hat{f}(\cdot)$ are estimation of states by a trained model). Comparing with Definition 6, ICM or RND is just the special case of the State Influence.

In general, to achieve the sparse goal, agents should fight against the *uncertainty* of both environment and policy.



(a)MiniMaze **(b) Conformity** **(c) Power** **(d) Authority** **(e) State influence**

Figure 2: (a) MiniMaze. Visiting the focus states 15 and 17 can accelerate the agent's awareness of the environment because they contain more structure information. From the results calculated by $1,000$ random explorations with 20 steps at each episode, only the conformity in (b) might not well represent the structure of the environment. Combining with the power (c) and the authority (d), the focus state found in (e) is more consistent with the actual situation.

Since each exploration of a state will reduce its uncertainty, almost all intrinsic motivation methods are committed to improving the information gain of each encountered state. Only in this way can the uncertainty space be reduced faster and the goal be achieved. Most of the previous methods (Bellemare et al. 2016; Burda et al. 2019; Pathak et al. 2017) only pay attention to the novelty of a single state, while our state influence modeling reduces the uncertainty through both the state's attributes and the relationships between states. We will show the significant performance improvement of our methods in the following.

Experiments

To verify the efficiency of SI-based intrinsic reward $i^{SI}(s)$ and the proposed SI-based RL algorithms, we conduct three sets of experiments with corresponding analyses.

- State influence in the Mini Grid Maze: shows the necessity and the different roles of the various intrinsic potential signals of $i^{SI}(s)$
- SI-based Q-learning in Grid Maze: applies the SI-based Q-learning algorithm directly to solve the hard exploration problem in the Grid Maze.
- SI-based A2C in Atari games: extends $i^{SI}(s)$ and SI-based A2C to the Atari games, and shows the efficiency of proposed framework compared with baseline algorithms.

State Influence in Mini Grid Maze

We use a toy example shown in Figure 2 to illustrate the motivation for introducing social influence into the intrinsic rewards. The MiniMaze environment is a small maze with 6×6 states, encoded as 1 to 36 from left to right, top to bottom. The structure of the environment is constructed by several impassable black holes. Assuming that all states in the environment have no reward, its state space be regarded as a limited discrete one without reward. An agent starts from the beginning state in the top left corner shown in the Figure 2a. The action space is {"left", "right", "up", "down"}, and the transition limited by the walls and holes.

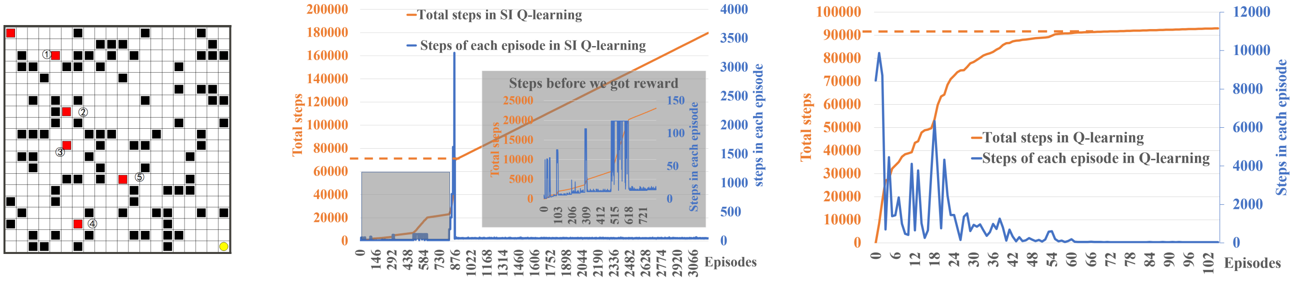
The set of trajectories is noted as T, and the i-th trajectory as $T_i = \{(s_{ij}, a_{ij})\}$, where the $s_{ij} \in S, a_{ij} \in A, i \in \{1, 2, \cdots, M\}$. In frequency statistics, the ith trajectory have W_i states, the first visit of state S_i is noted as $u_{iw_{it}}$, the w_{it} is the order of state in this trajectory, $w_{it} \in N^+, 0 < w_{it} < W_i$. The sort of states that first visit in this trajectory is $U_i = \{u_{i1}, \cdots, u_{iw_{it}}, \cdots, u_{iW_i}\}$, where $u_{iw_{it}} \in S$. Meanwhile, the number of times each state appears in T_i is $C_i = \{c_{i1}, \cdots, c_{iw_{it}}, \cdots, c_{iW_i}\}$, where, $c_{iw_{it}} \in N^+$. The appearing order of state S_k in trajectory T_i is w_{ik}, and we denotes the number of all the states currently known as N.

According to each component of social influence mentioned before, we use the total number of visits to states to indicate the conformity function (3) of discrete states $f_C(S_k) \triangleq \sum_{i=1}^{M} c_{iw_{ik}}/N$. And we use the number of states appear before S_k in each trajectory to define the power function (4), $f_P(S_k) \triangleq \sum_{i=1}^{M} \sum_{j=1}^{w_{ik}-1} c_{ij}/N$. In terms of the expressions of social influence from and to other states, this definition can be more complicated. Though this is the simplest, it will be helpful to the conduction of the following work. Similarly, we define the number of visited states S_k in each trajectory after S_k as the measurement of authority function (5) of state, $f_A(S_k) \triangleq \sum_{i=1}^{M} \sum_{j=w_{ik}+1}^{W_i} c_{ij}/N$.

Heatmaps of conformity, power, authority, and their combination are shown respectively in Figure 2. The blue heatmap (b) is the number of states that occur, as the counts in the count-based method, which is the simplest form of conformity. It is obviously not enough to evaluate the properties of states. We can see that the green (c) and orange (d) implementation in the figure can bring us more information. From the perspective of statistical information in Figure 2c, state 15 has strong power, and more information on state must be transmitted through it, which can be used as the focus state in the local area. To illustrate the effectiveness of social influence, we also combine the three attributes in the form $(f_P(s) + f_A(s))/\sqrt{f_C(s)}$. The results are shown in Figure 2 (e), from which we can say that state 17 is the most influential state. It can also be seen from Figure 2 (a) that node 15 and 17 are the focus state of the environment, which could improve exploration efficiency. This is an environment without any reward. Now we apply $i^{IS}(s)$ to the sparse reward environment.

SI-based Q-learning in Grid Maze

In Grid Maze, a frequency-based form is detailed to show the feasibility of the Social Influence based intrinsic reward function $i^{SI}(s)$. We define the $i^{SI}(s)$ used in this discrete



(a) Grid Maze (b) SI Q-learning (c) Q-learning

Figure 3: (a) Grid Maze is a 20 × 20 maze with random holes and a reward in the bottom right corner. Examples of some focus states during the local exploration process are represented by the numbers within the circles. (b) Steps in each episode and total steps by SI Q-learning. The enlarged gray area shows the process of gradual exploration. (c) Steps in each episode and total steps by Q-learning.

environment as:

$$s_* = \text{argmax}_s \frac{f_P(s) + f_A(s)}{\sqrt{f_C(s)}}, \tag{12}$$

$$i^{SI}(s) = \begin{cases} 1 & , \quad if\ s = s_*, \\ 0 & , \quad else. \end{cases} \tag{13}$$

The basic principle of constructing SI value is proportional to power and authority, and inversely proportional to conformity. As for the dominator of Eq. (12), we follow the count-based method (Bellemare et al. 2016; Pathak et al. 2017) for intrinsic reward construction. As for the numerator, we set the two terms have equal importance. Although much better results can be obtained by searching different configurations of these three terms, we found Eq. (2) already works well in the experiments. To verify the generality of the algorithm in a direct and straightforward way, we do not set the scaling factor in this experiment. By introducing the generalized intrinsic reward signals $i^{SI}(s)$, we apply SI Q-learning to a sparse reward Grid Maze. We also compare SI Q-learning with the traditional Q-learning to illustrate the feasibility and the efficiency of our framework.

To illustrate the progressive exploration, we show several representative states in Figure 3(a). They are focus states of local stabilization policy before the agent gets any extrinsic reward. And we can see the staged exploration in Figure 3(b). In Figure 3(b), we zoom in on the curve before we get the reward, which is the gray part of the graph. In the gray part, the steps of each episode grow up along with the expanding of exploration scope in each state of Figure 3(a). Combining Figure 3(a) with Figure 3(b), it can be seen that since we integrate the social influence into intrinsic reward, our SI-based Q-learning can give a reasonable exploration direction before getting rewards. Therefore, we can also reduce the total exploration steps.

A comparison between Figure 3(b) and Figure 3(c) illustrates this advantage. Figure 3(b) and Figure 3(c) respectively show the total training steps of SI Q-learning and Q-learning in the orange curve. The SI Q-learning gradually expands the scope of exploration and converges in about 75,000 steps, compared to the Q-learning, which converges in about 90,000 steps. These results show that the introduction of the SI based intrinsic reward function $i^{SI}(s)$ can effectively deal with the hard exploration tasks.

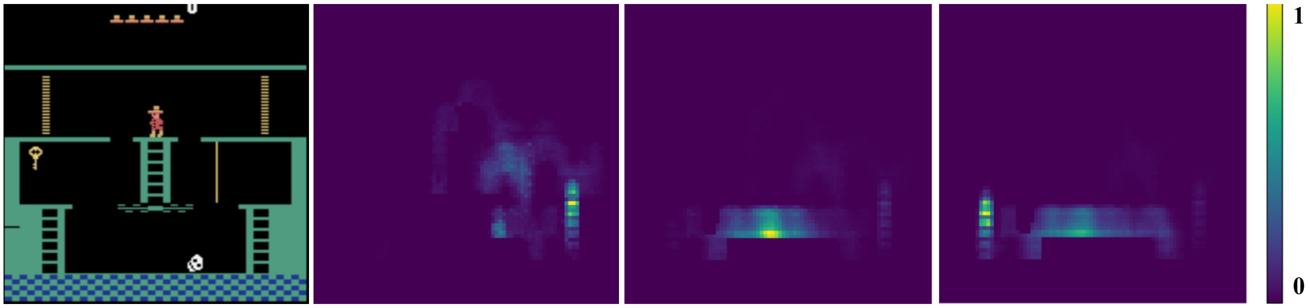
SI-based RL in Atari Games

To verify the efficiency and generalization ability of the proposed method, we apply $i^{SI}(s)$ to some hard exploration Atari games like *Montezuma's Revenge*, *Gravitar*, *Freeway* and so on. These games are generally considered as a notoriously difficult games in Atari games (Van Hasselt, Guez, and Silver 2016). For example, *Montezuma's Revenge* has three levels, each level has 24 different rooms, the agent navigate through different rooms to collect treasures. Here, we treat the coding of original frame as the state, and the state space is gradually increased along with the exploration. The action consists of 17 movements, such as up, down, jump, etc.

In the Atari experiments, we convert the 210 × 160 input RGB frames to grayscale images and resize them to 42 × 42 images following the practice in (Tang et al. 2017; Bellemare et al. 2016). The position of the agent is then discretized into a state in 42 × 42, and we set the $i^{SI}(s)$ with the same settings in Eqn. (12) and Eqn. (13).

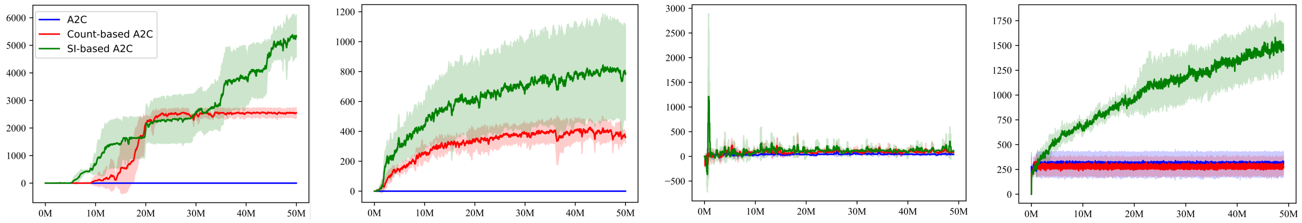
To compare the proposed SI-based A2C with the count-based A2C (Bellemare et al. 2016) on the first no-reward exploration stage of the sparse reward problem, we use the first room of *Montezuma's Revenge* as the experimental environment in Figure 4. We can see that the point with the highest heat in each stage are the points that the agent must explore to pass the first room. They can assist the agent to "exploit what influences the environment". According to the results shown in Figure 5(a), SI-A2C converges to the local goal in fewer training steps. This proves that the integration of relationship evaluation in social influence helps find the focus states. It is more effective to explore the focus states than only to explore the state with few visits, and SI-based A2C can get the reward in the environment faster as the green curve shown in Figure 5(a).

For the *Montezuma's Revenge* is a game with many rooms, it is a good example of multi-phase exploration. As shown in the Figure 5(a), after completing the first stage of exploration (about 2000 points average rewards have been obtained), our SI-based method also experiences a period of rest period (at about $20M$ to $28M$ steps). With the continuous increase of exploration scope (the $8th$ step in the Algorithm 1), our algorithm can quickly form the exploration advantage in the next stage. In contrast, the Count-based A2C will stay in the rest period for a long time. This shows that



(a) Montezuma's Revenge **(b) 1st stage** **(c) 2nd stage** **(d) 3rd stage**

Figure 4: *Heatmap of state influence in Montezuma's Revenge*. The state influence changes with the expand of the horizon of exploration. (b), (c) and (d) are heatmaps sampled from the first stage (max 500 steps per trajectory), the second stage (max 1000 steps per trajectory) and the third stage (max 1500 steps per trajectory) respectively.



(a) *Montezuma's Revenge* (b) *Venture* (c) *Private Eye* (d) *Gravitar*

Figure 5: Learning steps are compared with baseline algorithms in *Montezuma's Revenge*, *Venture*, *Private Eye* and *Gravitar*. The x axis is the number of steps. For the convenience of comparison, we limit the exploration steps of all games to 50 million steps. The y axis is the mean episodic rewards in each game. The results are averaged by 5 runs. These games are selected because they are all sparse reward games. The green curve represents SI-based algorithm's performance, in contrast to the red curve that only uses the Count-based method, and the blue cure represents the basic A2C which cannot get much performance.

our algorithm can effectively guide the agent to complete more valuable exploration without the extrinsic reward. In the performance of the other games, such as *Venture* in the Figure 5(b), *Private Eye* in the Figure 5(c) and *Gravitar* in the Figure 5(d), it also proves that our algorithm can help agents get better results faster than the baselines.

We further compare our proposed method with other baseline algorithms: SimHash (Tang et al. 2017), Curiosity Driven Exploration (ICM) (Pathak et al. 2017), Exploration with Mutual Information (EMI) (Kim et al. 2019), Count-based method(A3C+) (Bellemare et al. 2016) and A2C (Mnih et al. 2016) on four exploration games *Montezuma's Revenge*(MR), *Gravitar*(Gvt), *Freeway*(Fw) and *Venture*(Vt), and four games which are not sparse reward environment, *Berzerk*(Bz), *Jamesbond*(Jb), *Enduro*(Ed), and *Zaxxon*(Zx). Table 1 shows performance comparison of the proposed SI-based A2C and baseline methods. Our method achieves better results on most of the Atari games. The experimental results show that $i^{SI}(s)$ makes the learning process faster, and enable agents to explore further in the hard exploration tasks.

	MR	Gvt	Fw	Vt	Bz	Jb	Ed	Zx
SimHash	75	482	33	445	-	-	-	-
ICM	1011	424	**34**	418	-	-	-	-
EMI	387	558	**34**	646	-	-	-	-
RND	3442	1348	**34**	**1258**	-	-	-	-
A3C+	2551	284	30	361	-	-	-	-
A2C	6	329	0	0	1203	399	0	124
SI-A2C	**5342**	**1451**	**34**	783	**1559**	**1996**	**60**	**8602**

Table 1: Performance in Atari games of the SI-based A2C and baseline methods from the original papers. − indicates the non-reported corresponding data. The results are averaged by 5 runs with diverse random seeds in 50M steps.

Conclusion and Future Work

In this work, we introduce a social influence based intrinsic reward function to reinforcement learning in hard exploration tasks with sparse rewards. This definition effectively complements existing essential reinforcement learning solutions. We use $i^{SI}(s)$ on the value function-based and policy gradient based RL algorithms. In the Grid Maze with many obstacles and only one reward signal, our experimental results show that owing to $i^{SI}(s)$, the agent has a reasonable exploration direction before it gets a reward. Based on the SI based RL framework we proposed, the experience is collected in a gradually expanding way. Thus, the agent can explore and recognize the environment incrementally. Finally, by applying the proposed framework to some hard exploration Atari games, the results show that the algorithms combining with $i^{SI}(s)$ achieve the task objectives in fewer exploration steps than the baseline algorithms. We just focus the no-reward exploration stage in this paper, the local-reward exploitation stage will be considered in the future.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant No. 62076238 and 61902402, in part by the National Key Research and Development Program of China under Grant No. 2019AAA010340X, in part by the CCF-Tencent Open Research Fund under Grant No. RAGR20200104, and in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27000000.

References

Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Abbeel, O. P.; and Zaremba, W. 2017. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, 5048–5058.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 1471–1479.

Burda, Y.; Edwards, H.; Storkey, A.; and Klimov, O. 2019. Exploration by Random Network Distillation. In *International Conference on Learning Representations*, 1–17.

Friedkin, N. E. 2006. *A structural theory of social influence*, volume 13. Cambridge University Press.

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep Q-learning from demonstrations. In *AAAI Conference on Artificial Intelligence*, 3223–3230.

Kim, H.; Kim, J.; Jeong, Y.; Levine, S.; and Song, H. O. 2019. EMI: Exploration with Mutual Information. In *International Conference on Machine Learning*, 3360–3369.

Marx, K.; and Engels, F. 1969. Theses on Feuerbach, 1888. *It is available at http://www. marxists. org/archive/marx/works/1845/theses/theses. htm* .

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *International Conference on Machine Learning*, 1928–1937.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540): 529.

Oh, J.; Guo, Y.; Singh, S.; and Lee, H. 2018. Self-imitation learning. In *International Conference on Machine Learning*, 3875–3884.

Oudeyer, P.; Kaplan, F.; and Hafner, V. V. 2007. Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation* 11(2): 265–286.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2778–2787.

Savinov, N.; Raichuk, A.; Vincent, D.; Marinier, R.; Pollefeys, M.; Lillicrap, T.; and Gelly, S. 2019. Episodic Curiosity through Reachability. In *International Conference on Learning Representations*, 1–20.

Schmidhuber, J. 1991. A possibility for implementing curiosity and boredom in model-building neural controllers. In *International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, 222–227.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International Conference on Machine Learning*, 1889–1897.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .

Taiga, A. A.; Fedus, W.; Machado, M. C.; Courville, A.; and Bellemare, M. G. 2020. On Bonus Based Exploration Methods In The Arcade Learning Environment. In *International Conference on Learning Representations*, 1–20.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Chen, O. X.; Duan, Y.; Schulman, J.; DeTurck, F.; and Abbeel, P. 2017. # Exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2753–2762.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double Q-learning. In *AAAI Conference on Artificial Intelligence*, 2094–2100.

Wang, Z.; Schaul, T.; Hessel, M.; Van Hasselt, H.; Lanctot, M.; and De Freitas, N. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *International Conference on Machine Learning*, 1995–2003.