Curiosity Increases Equality in Competitive Resource Allocation

Bernadette Bucher*, Siddharth Singh*, Clélia de Mutalier, Kostas Daniilidis & Vijay Balasubramanian
University of Pennsylvania
Philadelphia, PA 19104, USA
{bucherb@seas, sidsingh@seas, clelia@sas, kostas@seas, vijay@physics}.upenn.edu

Abstract

We consider multiple agents using different strategies to compete for resources with a diverse distribution of rewards. Statistical theory shows that two kinds of equilibria are possible: (1) where some agents “settle” on a fixed resource while others visit diverse sites, and (2) where all agents pursue a similar strategy of visiting diverse sites. The first equilibrium shows a highly skewed reward distribution; in the second equilibrium most agents are similarly successful. We show that a population of agents can learn these equilibrium strategies through reinforcement learning. In conventional Q-learning, the population of agents learns the equilibrium strategy with skewed rewards. If we add curiosity, an intrinsic motivation to explore, Q-learning converges faster, and produces the second equilibrium in which most agents get similar average rewards. Thus, curiosity increases equality.

1 Introduction

We consider multiple agents competing for resources that provide diverse rewards. In such systems, the incentives provided to agents and constraints on their behavior can dramatically change population dynamics and outcomes. For example, constraints on agent demand can force the population to a Pareto optimal equilibrium (Matsuda et al., 2010). Domansky & Kreps (2002) show that private preferences of individual agents can significantly impact the population resource distribution. Mulatier et al. (2020) demonstrate that in over-crowded conditions, competition can lead to significant inequality in average rewards that individuals receive at equilibrium. Here, we implement a reinforcement learning approach to competitive resource allocation. We show that a population of such agents can learn an ensemble of strategies leading to the equilibrium in Mulatier et al. (2020), and that, additionally, greater equality in reward results when agents are “curious”, i.e. when they have an additional intrinsic motivation to explore.

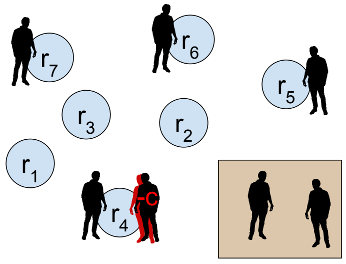


Figure 1: Agents can be either “at home” (brown box) or “out” occupying one of the resource locations (blue spots). Each location m provides an extrinsic reward $r_m > 0$. An agent trying an occupied location receives a negative reward $-c < 0$ (see red highlighted agent).

Curious behavior can be quantified in different ways (Schmidhuber, 2010; 1990; 1991; Jaegle et al., 2019), and is typically induced by incentivizing novelty seeking or learning progress, where the latter can be measured in terms of information gain. *Information gain* is approximated by Houthooft et al. (2016) by using a Bayesian neural network and by Osband et al. (2016) and Pathak et al. (2019) by measuring disagreements in an ensemble of models. Novelty seeking is induced by Pathak et al. (2017); Burda et al. (2018) by giving agents an intrinsic reward for visiting states yielding high *model error*. Bucher et al. (2019); Achiam & Sastry (2017) propose the use of *Bayesian surprise* measured by the negative log-likelihood of observed data conditioned on the model as another approach to novelty seeking curious behavior. Bellemare et al. (2016); Ostrovski et al. (2017); Kearns & Singh

* Denotes equal author contribution.

(2002) all present variations of a *count-based exploration bonus* in which a bonus is given for visiting less visited states. We consider count-based curiosity, the simplest form of introducing exploration incentives to the competitive resource allocation problem. Thus, we use the discrete version of count-based curiosity in the system we consider here in order to most directly analyze the impact of curiosity on dynamics.

We demonstrate that agents using a count-based curiosity reward achieve faster convergence to a more equal distribution of rewards across agents than agents following a greedy strategy in a competitive resource allocation task.

2 Resource Allocation Problem Formulation

2.1 The system

Consider a system with N agents competing for N discrete resource locations, which provide different payoffs and can be occupied by one agent at a time only. Agents can be either "home" or "out" occupying one of the locations (see Fig. 1). The ratio of the rate at which agents go out and the rate at which they go back home is the control parameter η, and the probability to find an agent out at any time of the simulation is therefore $p^{\text{out}} = \eta/(1+\eta)$. When an agent goes out it chooses a location m to visit based on its own strategy, which it learns during the simulation. If the chosen location is available, then the agent receives the positive extrinsic rewards r_m provided by the location. If instead the location turns out to be already occupied, then the agent receives a negative extrinsic reward denoted $-c$ and must try another location. Agents thus incur a cost for visiting occupied spots.

Agent behavior in this system is defined by the strategy they use to decide which location to go to. In prior work, equilibrium strategies of the agents were studied in the case where agents aim to maximize their expectation of extrinsic rewards. In this work, we compare the strategies learned by such greedy agents to the strategies learned by curious agents who choose the locations to visit based on both the extrinsic reward provided by the resources and an intrinsic reward they receive for exploring new resources. We compare the dynamics of the two systems during the learning process and study how it impacts the equilibrium strategies of the agents and their final expected rewards.

2.2 Curious Versus Greedy Agents

An agent's learning process can be described as a batched version of Q-learning. An agent's value function $V(k)$ is updated at the end of each episode, where an episode starts when the agent leaves home and ends when it finds a first available location (see Fig. 2). At the end of the k-th episode agent A_i updates the function V using:

$$V_m^i(k) = \frac{(k-1)\, V_m^i(k-1) + R_m^i(k)}{k}\,, \tag{1}$$

where $R_m^i(k)$ is the total extrinsic rewards accumulated by A_i during a trajectory after trying location m until the episode ends. If the location is not tried during the k-th episode, $R_m^i(k) = 0$, otherwise

$$R_m^i(k) = r_{m_f} - j_m(k) *\ c\,, \tag{2}$$

where r_{m_f} is the reward received at the final location of the trajectory, and $j_m(k)$ is the number of locations tried along the trajectory . The constant $j_m(k)\ c$ is the cost associated with visiting $j_m(k)$ occupied locations.

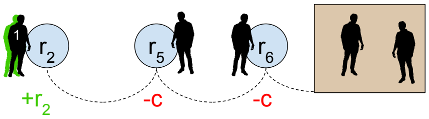
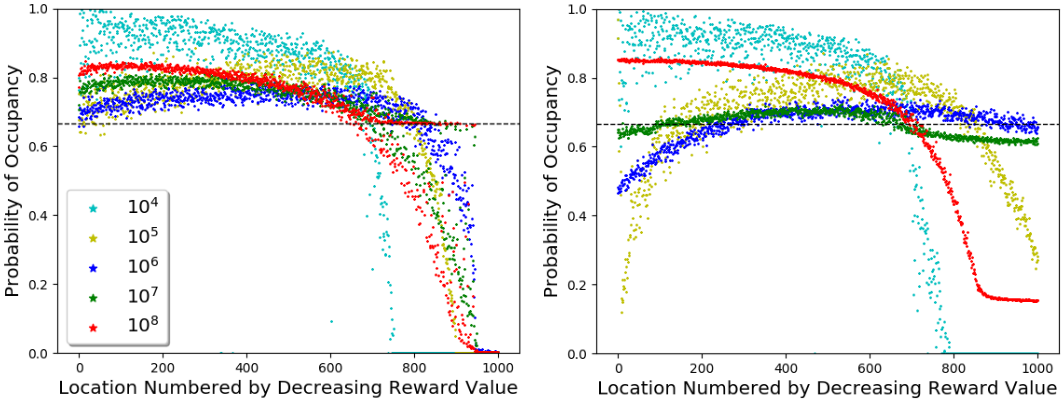


Figure 2: Example of a trajectory performed by agent A_1 during an episode: A_1 leaves home, tries location 6 and 5 both occupied, and finds the available location 2. During this episode, the total extrinsic reward received by A_1 after trying location 6 is $R_6^1 = r_2 - 2\,c$, whereas after trying location 2 it is simply $R_2^1 = r_2$.

The value function $V_m^i(k)$ is an estimate of the average extrinsic reward that agent A_i expects to get by trying location m in the competitive system; this estimate being learnt during the k previous episodes performed by A_i. At the beginning of each new episode, the greedy agent chooses the locations to try based on its current estimate of $V_m^i(k)$. The locations having the highest values in the value function are explored first. The curious agent however, balances the extrinsic exploitation



(a) Greedy Agents

(b) Curious Agents

Figure 3: Evolution of the probability of occupancy Eq. (4) of each location during the learning (a) for a population of greedy agents; (b) for a population of curious agents with parameter $\lambda = 0.5$. Snapshots of the occupancy profile are displayed at successive 10^x learning episodes. Simulations were performed with parameters $N = 1000$, $\eta = 2$, $c = -1$, and a linear distribution of the payoffs between $r_1 = 10$ and $r_{1000} = 1$. Locations are labeled by decreasing reward value, $r_m > r_{m+1}$ for all location m. The dashed line indicates the occupancy probability of locations that are exclusive to a single agent: $p_{out} = \eta/(1 + \eta) \simeq 0.666$ for our simulation with $\eta = 2$.

reward of greedy agents with an intrinsic exploration reward. In this work, exploration is encouraged using a count-based curiosity first proposed in a general form by Kearns & Singh (2002). The value function for the curious agent can then be defined by $W^i_m(k)$:

$$W^i_m(k) = V^i_m(k) + e^{-\lambda N^i_m(k)}\,, \tag{3}$$

where $N^i_m(k)$ is the current number of times agent A_i has visited spot m and λ is a parameter controlling the decay rate of curiosity. At each new episode the curious agent A_i chooses the locations to try based on the value function $W^i_m(k)$. The asymptotically decreasing function $e^{-\lambda N^i_m(k)}$ ensures that the intrinsic reward asymptotically reaches zero for sufficiently large $N^i_m(k)$.

3 Experimental Results

For the system described in Sec. 2.1, Mulatier et al. (2020) found that some agents can establish themselves as property owners of a single resource and thus earn larger payoffs than other agents. They showed that such behavior is allowed theoretically and can emerge naturally during learning. Such systems admit Nash equilibria without property owners, as well as equilibria with coexistence of property owners and other "nomadic" agents (who exploit more than one resource, but earn lower payoffs). Here, we show these equilibria can be discovered by reinforcement learning, and that incentivizing exploration during learning promotes the emergence of agent communities with fewer property owners, and, as a result, with less inequality.

3.1 Comparison of the Observed Learning Dynamics

Figure 3 displays the evolution of the occupancy of each location during learning, with and without curiosity. The occupancy probability of a location m is defined at any given time t during the simulation by:

$$\mathbb{P}[m \text{ is occupied at time } t] = \frac{T_{occ}(t)}{t}\,, \tag{4}$$

where $T_{occ}(t)$ is the total time location m has been occupied since the start of the simulation. The presence of property owners can be identify on the occupancy profile, as locations owned by a single agent have an occupancy probability exactly equal to the probability of that agent to be "out", which is $p_{out} = \eta/(1+\eta)$ (see Sec. 2.1). This probability is highlighted by the dashed horizontal line on both graphs of Fig. 3. For the greedy population (a), we observe the presence of multiple settlers after 10^8 learning episodes in the intermediate range of locations between $m = 700$ and $m = 900$, as indicated by the red dots overlapping with the horizontal dash line in that range. The curious population (b) instead displays no, or very few property owners.

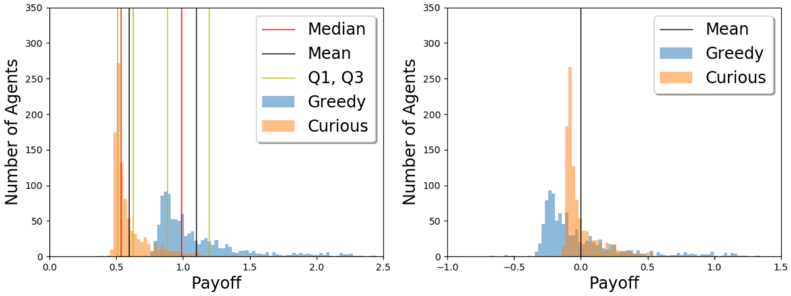


Figure 4: **Left.** Histograms of payoffs across agents after 10^8 iterations for the simulations of Fig. 3. **Right.** Same histograms after subtracting the mean payoff. *Greedy agent reward statistics:* mean (1.099), median (0.986), IQR (0.313), variance (0.105). *Curious agent reward statistics:* mean (0.597), median (0.536), IQR (0.115), variance (0.020).

For both simulations in Fig. 3 the value function was initialised to $V^i_m(0) = r_m$ for all agents, which implies that agents start with an initial knowledge of the system. Thus in the greedy system (a), agents explore first resources with highest payoff and then slowly spread over worse resources, displaying a smooth convergence of the occupancy profile. As a result, only the exploited resources are explored (locations ranging from $m = 950$ to $m = 1000$ are barely ever visited) and this learning dynamics seems to facilitate the emergence of property owners. In contrast, in case (b), curious agents rapidly spread over the entire system, the probability of occupancy of the locations being almost flat at 10^6 episodes. Fig. 3b then shows a sudden move to equilibrium starting around 10^7 episodes, as the curiosity exploration bonus becomes negligible. After exploring the entire system, the whole population suddenly converges towards a stationary solution with almost no property owner. The curious exploration has prevented the emergence of property owners and promoted more equal payoff outcomes for the agents.

3.2 Comparison of the resulting equilibrium strategies

Curiosity decreases the overall payoffs of the agents. Figure 4 compares the distribution of payoffs (average extrinsic rewards) across agents at equilibrium in the curious and greedy system respectively. As an immediate result, we find that greedy agents have a higher mean and median reward than the curious agents. By reducing the number of property owners in the system, curiosity decreases not only the payoff of the property owners, but also of all the agents of the population. This also hints towards commonly seen communities of animals with specialists and generalists (Van Tienderen (1991)), as having specialists increases the rewards of the entire community.

Curiosity increases equality. We also observe in Fig. 4 that curious agents have a lower variance and interquartile range (IQR), indicating greater equality in the system. In the context of an economic system for instance, payoffs can be considered money with which agents may buy goods and services. The real value of money measured against purchasing power for goods and services is determined by the relationship between the amount of money an agent has relative to the other agents. As visualized by the normalized payoffs in Fig. 4, the real value of money for agents with curiosity is higher as well as being more equally distributed.

4 Conclusion

Here we introduced curiosity as a form of intrinsic reward for exploration in a competitive resource allocation problem. We experimentally compared the learning dynamics of this system with another whether agents only receive rewards by exploiting resources. Future work includes exploring more complex curiosity reward functions based on the agent’s model of the environment.

ACKNOWLEDGMENTS

The authors are grateful for support through the Curious Minded Machines project funded by the Honda Research Institute.

REFERENCES

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *CoRR*, abs/1703.01732, 2017. URL `http://arxiv.org/abs/1703.01732`.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

Bernadette Bucher, Anton Arapin, Ramanan Sekar, Feifei Duan, Marc Badger, Kostas Daniilidis, and Oleh Rybkin. Perception-driven curiosity with bayesian surprise. *RSS Workshop on Combining Learning and Reasoning for Human-Level Robot Intelligence*, 2019.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Victor Domansky and Victoria Kreps. Social equilibria for competitive resource allocation models. In *Constructing and Applying Objective Functions*, pp. 408–419. Springer, 2002.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pp. 1109–1117, 2016.

Andrew Jaegle, Vahid Mehrpour, and Nicole Rust. Visual novelty, curiosity, and intrinsic reward in machine learning and the brain. *arXiv preprint arXiv:1901.02478*, 2019.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002. ISSN 08856125. doi: 10.1023/A:1017984413808.

Tetsuya Matsuda, Toshiya Kaihara, and Nobutada Fujii. Resource allocation analysis in perfectly competitive virtual market with demand constraints of consumers. In *Advances in Practical Multi-Agent Systems*, pp. 181–200. Springer, 2010.

Clelia De Mulatier, Cristina Pinneri, and Matteo Marsili. Competing for resources: on the emergence of property rights. *Bulletin of the American Physical Society*, 2020.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pp. 4026–4034, 2016.

Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2721–2730. JMLR. org, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-Supervised Exploration via Disagreement. *International Conference on Machine Learning*, 2019. URL `http://arxiv.org/abs/1906.04161`.

J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990;2013;2010). *IEEE Trans. on Auton. Ment. Dev.*, 2(3):230–247, September 2010. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2056368. URL `https://doi.org/10.1109/TAMD.2010.2056368`.

Jürgen Schmidhuber. Curious model-building control systems. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*, pp. 1458–1463. IEEE, 1991.

Jürgen Schmidhuber. Making the world differentiable: On using self-supervised fully recurrent neural networks for dynamic reinforcement learning and planning in non-stationary environments. Technical report, 1990.

Peter H Van Tienderen. Evolution of generalists and specialists in spatially heterogeneous environments. *Evolution*, 45(6):1317–1331, 1991.