Informationally-Mosaic Reinforcement Learning

Nikunj Gupta[1] , **Tao Li**[1] , **Quanyan Zhu**[1]
[1]New York University
{ng2531, taoli, qz494}@nyu.edu

Abstract

Reinforcement learning (RL) algorithms have recently shown encouraging successes in learning to solve sequential decision-making problems through interactions with the environment. However, to scale them to complex real-world tasks, agents must be able to discover and adapt to the varying information structures in the environment. The issue of learning under unknown, dynamic, and generally amorphous information structures poses a great challenge to current RL studies. To address it, we propose a novel notion, Informationally-Mosaic Reinforcement Learning (IMRL), where the agent relies on a complementary, autonomous module to explore, learn and utilize constructive information from the environment. In particular, the agent's exploration operates in a laissez-faire manner, that is, it voluntarily rewards the autonomous module for discovering helpful information. The proposed framework brings up flexibility with respect to information structures, as well as enhances reinforcement learning efficiency. This paper introduces novel metrics, including *Value of Information* (VoI), quantifying the importance of informational exploration, and *Equilibrium Quotient* (EQ), demonstrating the efficiency and effectiveness of the agent's decision making within IMRL. We present the corresponding numerical evaluation using several procedurally-generated benchmark Minigrid environments.

1 Introduction

Reinforcement Learning (RL) has a vast potential for automating economically relevant tasks such as learning control policies for robots directly from pixel values from cameras in the real world [Levine *et al.*, 2016; Levine *et al.*, 2018], playing video games [Mnih *et al.*, 2015], indoor navigation [Zhu *et al.*, 2017] and even creating agents that can meta-learn ("learn to learn") [Duan *et al.*, 2016; Wang *et al.*, 2016]. However, to scale RL to complex real-world tasks, practical autonomous agents are required to build correct knowledge of the environment using available information flexibly, and utilize it efficiently, which has not been addressed thoroughly.

Informational Flexibility What information agents can observe or acquire at each time instance, which we term the information structure (IS), directly influence the development of agents' decision-making model. Mathematically, IS is defined by a set of random variables that can be observed by agents. Because of its autonomous nature, RL is more challenging than supervised learning and unsupervised learning in terms of information structures. For the later two, the IS, is usually static and prefixed by human operator. For example, the IS for supervised learning is just pairs of data points and associated label. However, in RL applications, agents in general faces unknown, dynamic and amorphous IS. For example, in robot navigation [Zhu *et al.*, 2017], what agents can observe is subject to physical conditions of the environment, which can be dynamic. Besides, in multi-agent systems, message passing among agents can emerge naturally without any preset protocol, creating amorphous IS[Hernandez-Leal *et al.*, 2019]. Therefore, the agents need to dynamically process unstructured information that varies across environments and is generally unknown beforehand. Consequently, the flexibility of learning frameworks with respect to dynamic and amorphous IS is critical to the success of RL applications in reality. However, current studies have primarily focused on RL under prefixed IS, e.g., full/partial state observation [Hernandez-Leal *et al.*, 2019].

Efficiency Moreover, RL is also known to require huge amounts of experience before becoming useful, even when solving relatively small problems, which is a challenge for it to be implemented to solve everyday problems. This bottleneck is primarily due the slow collection and understanding of information. This problem intensifies further when multiple intelligent agents simultaneously learn in the same environment.

In this article, we propose the novel notion of Informationally-Mosaic Reinforcement Learning (IMRL), which explicitly addresses the problem of reinforcement learning under unknown, dynamic, and generally amorphous information structures. IMRL can be applied to various base RL algorithm and information structures (e.g. features, feature groups or raw input). In particular, an intelligent agent that can observe the environment partially, makes use a complementary, autonomous module — the *information explorer* — which explores and learns constructive information from the environment. In IMRL, an agent learns to efficiently arrange together pieces of information (hence the term *mosaic*) both from its own partial observation and informational "messages" from the *information explorer* and chooses sequential actions in the environment. Another fold of mosaic is the agent self is also a small piece, especially in multi-agent systems: it does not require any specific information structure to function, rather it can be implemented under amorphous ones.

We evaluate our algorithm on several procedurally-generated benchmark Minigrid environments for the case with one *information explorer* and one learning agent and show that IMRL achieves better decision-making performances compared to reinforcement learning without explicit information-gathering. We also define two novel metrics - Value of Information and Equilibrium Quotient - to quantify the importance of exploring and gathering information in reinforcement learning and demonstrate the efficiency and effectiveness of the agent's decision making within IMRL.

2 Related Works

The topic of learning information aggregation policies in reinforcement learning has not been much explored, unlike in the supervised learning research [Dulac-Arnold *et al.*, 2013; Mnih *et al.*, 2014]. Many of the proposals in reinforcement learning address this issue by comprehensively providing the learning agents with structured information to speed up their learning. For instance, in action-advising [Griffith *et al.*, 2013; Torrey and Taylor, 2013; Zhan *et al.*, 2016] the learning agent is provided with action suggestions from an expert or a more experienced agent. In Human-focused transfer, an automated agent learns to leverage diverse information transferred from a human and tries to make better use of this costly feedback [Krening *et al.*, 2016; Rosenfeld *et al.*, 2018; Abel *et al.*, 2017]. Learning from Demonstrations is also a well-studied category of methods in which an experienced teacher provides demonstrations to a learning agent [Schaal and others, 1997; Banerjee and Stone, 2007]. Transfer Learning (TL) [Taylor and Stone, 2009] accelerates learning by reusing previous knowledge in Deep RL tasks [Omidshafiei *et al.*, 2019; Devin *et al.*, 2017]. However, all of these works assume availability of prior structured knowledge or an expert which may not be feasible at all times and for all applications. Moreover, as discussed earlier, the information structure in an environment varies dynamically and is generally unstructured. In IMRL, an agent explicitly learns to explore valuable information from the environmental states and goals and is capable of adapting to dynamically changing information structures in an environment.

To address the problem of learning unstructured information from the environment, there is an interesting and growing body of methods which investigates how to best share knowledge by *explicit communication* with another agent (need not be an expert), which can have different sensors and internal representations. This type of transfer is motivated by a simple thought: knowledge which is already available in another agent, need not be relearned from scratch. Recently, there has been a significant revival of emergent communication research using methods from deep reinforcement learning (deep RL) [Hernandez-Leal *et al.*, 2019]. However, the main motivation in these works is to address multi-agent learning and communication is just used as a tool which the agents implicitly use to benefit from. Whereas, in IMRL, we explicitly model informational exploration and enable the agent to learn to value learned knowledge, hence motivating information sharing much more adaptively. Moreover, most works in this field have focused only on defining *what* knowledge to transfer and *how*, whereas, we are interested in also addressing *when* to share information which is a non-trivial task too. Allowing unrestricted communication throughout the learning phase [Gupta *et al.*, 2021] is not feasible for all applications, and heuristics-based limitations are difficult to know prior training and are not generalizable to different settings [Hernandez-Leal *et al.*, 2019].

3 Models

Informationally-Mosaic Reinforcement Learning, different from the typical RL formulation, includes a complementary, autonomous module used by RL agents to explore, learn and utilize constructive information from the environment. Such an autonomous module can be modelled as a separate agent, i.e., *information explorer*, from a multi-agent system viewpoint, and we call this framework as IMMARL (Informationally-Mosaic Multi-Agent Reinforcement Learning). Mathematically, the stochastic game G in IMMARL can now be defined by the following tuple

$$\langle \mathcal{N} \cup \{\mathcal{H}\}, S, \{O^a\}_{a \in N}, (U^a, \mathcal{I})_{a \in N}, (\mathcal{M}, V), T, \{r^a\}_{a \in N} \rangle$$

in which the *information explorer* $\mathcal{H}$ observes the environment's true state $s \in S$, and chooses its actions in the form of a message-value pair (m^a, v^a). IMMARL considers a partially observable setting for the other n agents, identified by $a \in \mathcal{N} \equiv \{1, ..., n\}$. At any given time-step, each agent a, draws observations $o^a \in O^a$ according to the observation function O(s, a) : $\mathcal{S} \times \mathcal{N} \rightarrow O^a$, concatenates it with v^a from $\mathcal{H}$, and chooses sequential actions $i^a \in \mathcal{I} \equiv \{0, 1\}$, to decide whether to inquire for m^a or not. On choosing i^a = 1, the agent observes m^a from $\mathcal{H}$ in addition to o^a. Based on the final observation (o^a with or without m^a), each agent a, chooses sequential actions $u^a \in U^a$, hence forming a joint action $\mathbf{u} \in \mathbf{U} \equiv U^n$ in the environment at each time-step. The joint action $\mathbf{u}$ induces a transition in the environment according to the state transition function T(s'|s, $\mathbf{u}$) : $\mathcal{S} \times \mathbf{U} \times \mathcal{S} \rightarrow$ [0, 1].

The environment returns a reward corresponding to each agent's action u^a in the environment, according to the the reward function r^a(o^a, u^a) : $O^a \times U^a \rightarrow$ R. Now, the *information explorer* must learn to encapsulate available information from state s, into small "messages" m^a, to facilitate the learning of the agent *a* whenever required. The fact that communication enables agents to outperform non-communicating agents in several domains is not very surprising. However, unrestricted sharing of information among agents during the training and execution is not always feasible. The cost of communication in the real world is not negligible, with an increase in the number of agents in the environment it would be best suited for the system to learn to communicate efficiently and in some cases, individual agents may not even have the ability to communicate with other agents. Hence, two main ideas underly IMMARL: 1) sharing learned "useful" knowledge *efficiently* among agents, and 2) learning in a self-adaptive (Laissez-faire) manner. The remainder of this section describes these ideas.

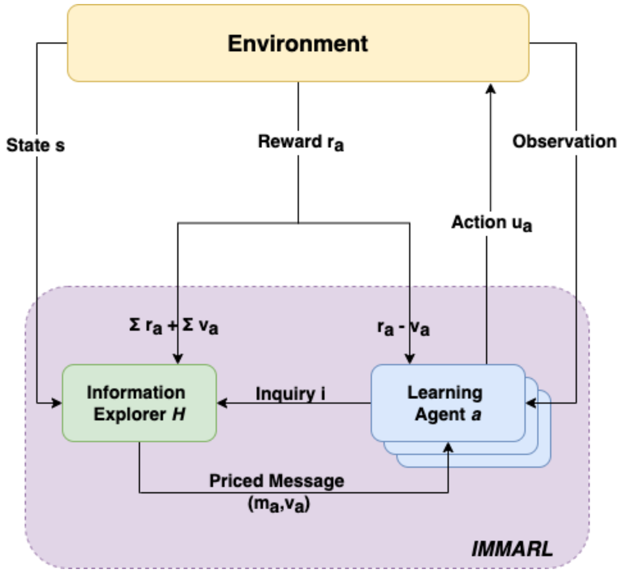


Figure 1: Informationally-Mosaic Reinforcement Learning (modelled as a multi-agent system - IMMARL) to explicitly address reinforcement learning under unknown, dynamic, and amorphous information structures in the environment.

Firstly, to motivate efficient knowledge sharing, the *information explorer* learns a value $v^a \in [0, \infty)$ corresponding to the informational message m^a for each agent $a \in \mathcal{N}$. As mentioned earlier, an agent a must decide whether or not to inquire for a m^a from $\mathcal{H}$ considering this value (or cost) of information v^a along with its partial observation of o^a of the environment. In case the agent decides to inquire, it must incur a cost of v^a as a penalty. Thus, agents $a \in \mathcal{N}$ learn using $\{r^a - v^a\}_{a \in \mathcal{N}}$. To incentivise the *information explorer* to learn to produce better and useful messages for the other agents in the environment it maximizes over not only the sum of their rewards from the environment but also the sum of values it earned from sharing information. Hence, the *information explorer* learns its policy using $\sum_a \{r^a\}_{a \in \mathcal{N}} + \sum_a \{v^a\}_{a \in \mathcal{N}}$. An interesting point to note here that the *information explorer* does not directly interact with the environment to get feedback, instead, it learns to output "valuable" messages by looking at how the other agents performed in the environment using the messages it communicated. This way, *information explorer* does not actually exist in the environment and hence IMMARL is just another perspective to look at IMRL without imposing the inherent challenges of multi-agent systems into the framework.

Secondly, an agent $a \in \mathcal{N}$ is capable of modifying its inquiry policy behavior in order to achieve system objectives in diverse environments. In other words, the agent's exploration operates in a laissez-faire manner, that is, it voluntarily rewards the *information explorer* for discovering helpful information. In most related works, the ***'when-to-communicate'*** aspect of sharing information in a multi-agent system is either ***ignored*** (i.e., free communication is assumed) or limited using ***heuristics*** which do not generalize well. Also, previous works have majorly focused on controlling the communication from the perspective of the information agent, while in IMMARL, the learning agents play a significant role in deciding whether to inquire or not by observing the cost of information along with their partial observation.

4 Numerical Experiments

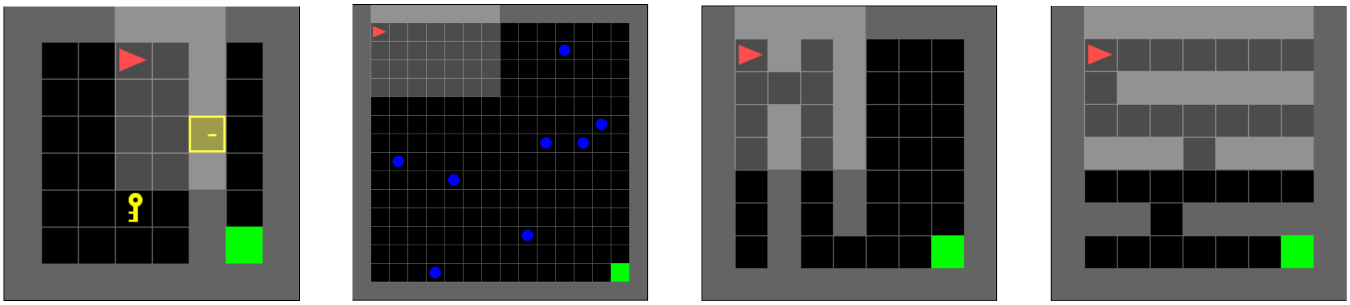
4.1 Environments

MiniGrid environments [Chevalier-Boisvert *et al.*, 2018] are popular, procedurally-generated, and flexible gridworld implementations where the agent can move between adjacent tiles in a rectangular grid and interact with objects, such as keys and doors (Figure 2). Agents can observe the environment partially, rewards are designed to be sparse, and specific actions are needed to visit all states, hence making exploration in these environment quite challenging. With MiniGrid, we can generate several environments that are different in many ways. These test-beds enable evaluation of both the learning abilities and efficiency of IMMARL and its flexibility to deal with diverse tasks.

4.2 Implementation Details

All environments give a 7×7×3 partial observation encoding the contents in front of the agent. The agent cannot observe through walls or closed doors. The position and orientation of the agent are shown by the red pointer, and the grey-highlighted cells comprise the agent's field of view (Figure 2). The action space is discrete: left, right, forward, pick up, drop, toggle (unlocks a door if the agent has the corresponding key and opens/closes a door if unlocked), and done. In all tasks, the extrinsic reward is $r_t = 1 - 0.9(t/T)$ for success and zero for failure. Thus reward is given only for solving the task, making it sparse. The grid is procedurally generated at each episode, and the agent's initial position is random within a fixed area far from the goal (e.g., in DoorKey the agent starts in the area with the key). Below is a list of the tasks used in this article. Everything is as implemented by default in MiniGrid codebase [Chevalier-Boisvert *et al.*, 2018].

- **DoorKey-8x8** (Figure 2a): In this scenario, the agent must pick up the key, toggle the door, and navigate to the green square to receive a reward. This environment is difficult, because of the sparse reward and a requirement of interaction with various objects, to solve using classical RL algorithms (also shown later in Section 4.4).
- **DynamicObstacles-16x16** (Figure 2b): The goal of the agent is to reach the green goal square without colliding with any of the moving obstacles in the room in this environment. Agent is heavily penalized upon colliding with an obstacle and the episode finishes. This environment is useful to test Dynamic Obstacle Avoidance for mobile robots under partial observability. This task is difficult and requires more steps to be completed.



(a) DoorKey (b) DynamicObstacles (c) SimpleCrossingS9N2 (d) SimpleCrossingS9N3

Figure 2: The MiniGrid environments used in this work. The agent has to navigate through a grid and interact with different objects (keys, doors, balls) to fulfil a task. At each episode, the grids are procedurally-generated, changing rooms layout, objects positioning and color.

- **SimpleCrossing** (Figure 2c, 2d): The agent has to reach the green goal square on the other corner of the room which is divided by walls. Each wall runs across the room either horizontally or vertically, and has a single crossing point which can be used.

For the *information explorer* a convolutional neural network was used to encode information into a vector of floating points scaled to [1, 1] representing an informational 'message' (consistent across domains) from the true state it observed and another neural network was used to learn the value of the information it gathered and learned to share with the learning agent. SoftPlus (a smooth approximation to the ReLU function) was used to constrain the output of the *value-of-information* network to always be positive. For the actor and critic networks of the learning agent, 2 fully-connected multi-layer perceptron layers were used to process the input layer and to produce the output from the hidden state. The learning agent used another neural network to output a binary decision for inquiring for a informational message from the *information explorer* at each time-step. To support end-to-end training of IMMARL agent networks, the real-valued output i of the inquiry policy is processed by a discretise/regularise unit (DRU(i)). This unit regularises the output during learning, DRU(i) = Logistic($Gaussian(i, \sigma)$), and binarizes it during execution, DRU(i) = $1\,\{i > 0\}$, where σ is the standard deviation of the noise added to the channel. Proximal policy optimization (PPO) [Schulman *et al.*, 2017] was used to update the decision-making policies for these agents.

4.3 Experimental setups

In this article, we experiment with only one learning agent and one *information explorer*. Nevertheless, it is entirely possible that multiple *information explorer* agents could better assist multiple learning agents. This is left to future works. As discussed earlier, in IMMARL, the *information explorer* $\mathcal{H}$ can observe the entire state of the environment, whereas the learning agent receives a partial observation from the environment. Before, the learning agent takes an action, it has to decide whether or not to inquire for a message from $\mathcal{H}$. We present four setups to study the learning curves (plots of reward collections over time) of an agent with varying strategies of making inquiries (learning-based and heuristic-based) for information from the *information explorer* $\mathcal{H}$.

1. **Standard**: The learning agent has no access to $\mathcal{H}$ throughout the training and execution. This baseline would help us realize the difficulty for an unaided agent learning to perform a given task.
2. **Complete Access**: $\mathcal{H}$ shares information with the learning agent at all times. This acts as another baseline to illustrate how capable and useful $\mathcal{H}$ can be in an environment.
3. **Random**: The learning agent randomly interacts with $\mathcal{H}$. This acts as a heuristic for our experiments to compare against IMMARL. Results and evaluation metrics in the upcoming sections show significant differences in IMMARL's strategy of collecting information compared to a random strategy.
4. **IMMARL**: $\mathcal{H}$ and the learning agent *learn to share* "useful" knowledge *efficiently* using IMRL.

In all the setups except for IMMARL, the learning agent incurs no penalty for receiving messages.

4.4 Results

This section describes the experiments conducted to test *information explorer*'s potential of learning and encapsulating messages, speeding up learning agent's training, and learning an efficient exchange of knowledge in IMRL, on three MiniGrid environments — DoorKey, DynamicObstacles, and SimpleCrossing — whose details are described in the previous section. Learning curves shown in Figure 3 for evaluation are plots of the agent's average reward collected (averaged over 10 episodes) as a function of episodes. The plots show the mean and confidence interval over sixteen random seeds per method, smoothed using a sliding window of 100 episodes for readability.

At first, we conduct experiments to compare the strength of IMRL against RL. For this, we let the learning agent learn to accomplish the tasks alone, by partially observing the environment. Figure 3 shows the corresponding curve (red) for this case in all domains. This experimental setup acts as a baseline to compare with the performance of IMMARL

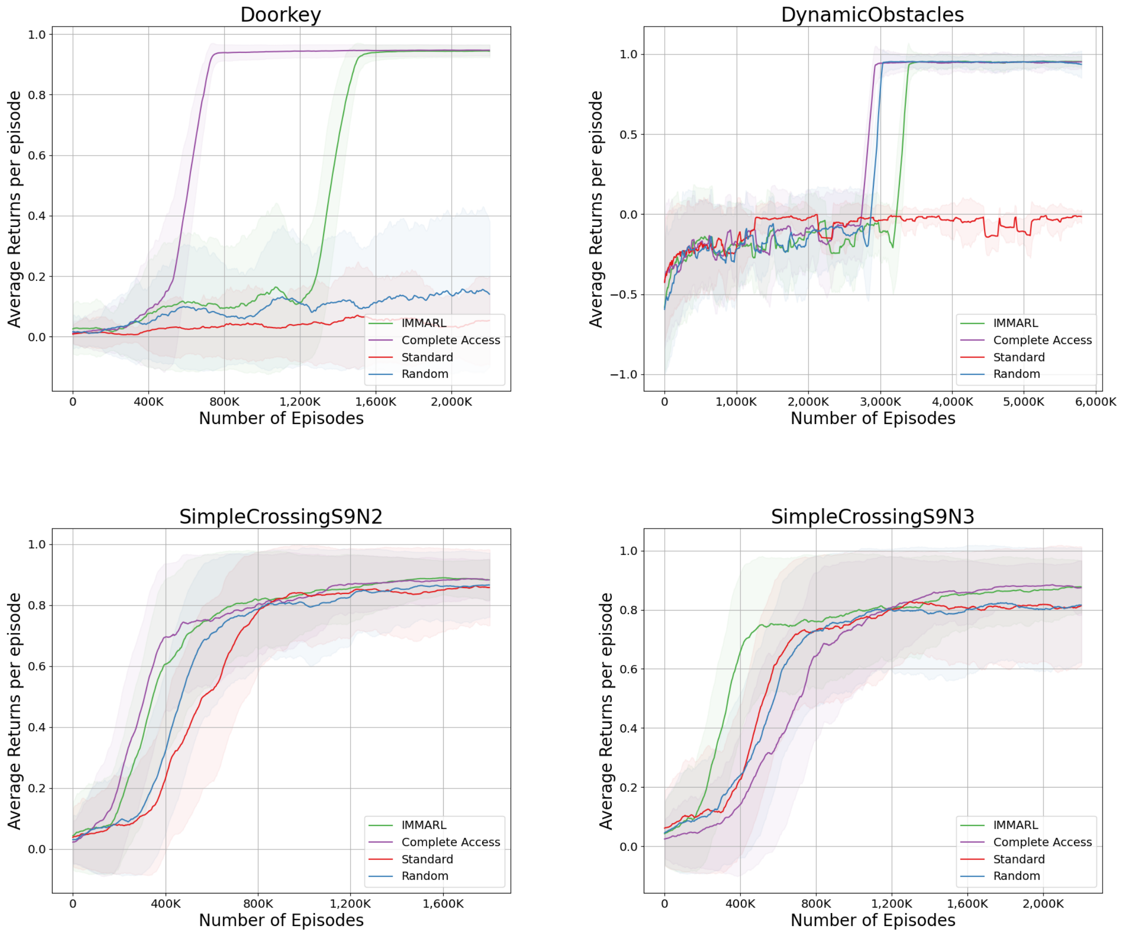


Figure 3: Informationally-mosaic reinforcement learning outperforms regular reinforcement learning and randomly gathering information. IMRL also illustrates similar performances to a learning agent with complete access to true states in all domains, but uses a much more efficient strategy of optimally building *mosaics* of information over time (or episodes). The plots show the mean and confidence interval of the agent's average reward collected (mean over 10 episodes) as a function of episodes, smoothed over 100 episodes for readability.

agents. Next we allow the agent to make unrestricted use of the *information explorer* throughout the learning process. The learning agents performs better than the above baseline, i.e., it is able to collect greater reward values earlier than in other cases, as is shown by corresponding performance curves (purple) in Figure 3. This is not surprising because the learning agent has free access to the true state information of the environment at all times.

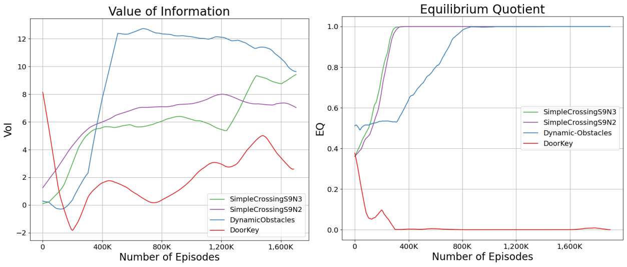
Next, we let the agents learn under the principles of informationally-mosaic reinforcement learning. The learning reward curves for IMMARL agents in all the domains are shown in Figure 3 (green). In all environments, IMMARL agents outperform the independently learning agents. These results suggest that the *information explorer* was able to explore "usable" information structures from the true state of the environment resulting in enhancing reinforcement learning of the agent, thus, reasonably demonstrating that IMRL is superior to standard reinforcement learning.

An interesting point to note in the above results is that, IMMARL agents perform better than the case with non-restricted access to the *information explorer* in the SimpleCrossingS9N3 scenario. We speculate that availability of more knowledge than necessary during the learning made it difficult for the agent to train, whereas IMMARL agents were able to learn an optimal strategy for sharing of only "required" knowledge during training, leading to a better performance in terms of reward collections in the environment. Furthermore, consistent better performance of IMMARL in diverse domains signifies that the notion is indeed flexible and self-adaptive. In other words, the learning agent was able to adapt it inquiry policy to incentivise the *information explorer* to produce more useful or ("valuable") messages and the *information explorer* was able to adapt its value/cost function during the learning in order to promote efficient knowledge sharing in diverse environments without any changes to the framework.

Finally, we enable the agent to randomly inquire for information from the *information explorer* to investigate the fact that IMMARL agents are learning a better, or at least different strategy than a random strategy. Compared to the *Standard* experimental setup, in *Random*, agents performed well in some cases (DynamicObstacles), indifferent in some (SimpleCrossingS9N3, SimpleCrossingS9N3), while very bad in the others (DoorKey). Figure 3 shows corresponding learning curves (blue) in all the domains, illustrating that the random strategy is unpredictable and certainly cannot adapt to varying situations. On the other hand, IMMARL agents learn a significantly different strategy and the learning adjusts to all the diverse test-beds, hence exhibiting adaptability.

In summary, this article's empirical evaluations of IMRL in various MiniGrid domains suggest: 1) IMRL performs better than regular reinforcement learning, randomly gathering information, and in some cases, complete access to true states too, 2) IMRL is flexible and works across domains with different information structures and complexity levels, and 3) IMRL is cost-effective, i.e., does not rely on the *information explorer* all the time (also discussed later in evaluation metrics) and self-adaptive, i.e., adapts the *information explorer*'s value-of-information policy and the learning agent's inquiry policy to adapt to different environmental settings.

4.5 Evaluation Metrics



(a) VoI

(b) EQ

Figure 4: Evaluation Metrics: (a) Value of Information (VoI): to quantify the 'usefulness' of informational exploration through IMRL, (b) Equilibrium Quotient (EQ): to indicate the level of information exchanged among agents, hence demonstrating the efficiency and effectiveness of the agent's decision making within IMRL

We propose the following novel metrics, that illustrate the advantage of informationally-mosaic reinforcement learning.

1. **Value of Information (VoI)**: We define the value of all the information from the *information explorer* with each episode as

$$\nu = log\left(\frac{r}{\sum_t c_t}\right)$$

where r is the reward earned by the learning agent in an episode and the c_t is the cost incurred by the agent for interacting with the *information explorer* at time-step t in an episode. ν indicates how helpful the collection of information sent by the *information explorer* is within each episode.

Now, we can plot ν as a function of time (or episodes). Figure 4a shows that, in all the domains, ν appears to be increasing with time (number of episodes), suggesting that the system is benefiting from the learned information being collected and shared.

2. **Equilibrium Quotient (EQ)**: We define the *equilibrium quotient* for algorithms that make explicit inquiries for additional information before making decisions in a state. This can relate to all RL algorithms where a piece of information could be an action-advice from an expert (human or artificial), communicated 'messages' from other agents in the system (or an expert) or prior knowledge from another source (transfer learning). In IMRL, we consider the number of messages from the *information explorer* as an external source of information for the learning agent. EQ can be defined as

$$\eta = \frac{\text{number of inquiries made per episode}}{\text{horizon of an episode}}$$

We can now plot this efficiency metric as the learning proceeds (Figure 4b). η values indicate the level of information exchanged among agents. η = 1 for *Complete Access* experimental setup and for *Standard*, η = 0. For the Doorkey environment, η decreases over time indicating that enough information has been gathered in the initial stages and there is no further interaction required a certain number of episodes ($\sim$500K in this case). On the other hand, the η plots for the other domains suggest that the agents need not share much knowledge initially and hence save the costs incurred due to such interactions.

5 Conclusion

In this paper we argued how important it is for a framework to be able to discover unknown, unstructured and dynamic information in complex environments, and formulated the task of learning to explore information for solving diverse problem inside the reinforcement learning paradigm in a flexible and adaptable manner. We proposed the novel notion of informationally-mosaic reinforcement learning in which the learning agents bring together multiple pieces of simultaneously learned information for effective and efficient sequential decision-making. We also proposed two novel metrics to formalize the quality of the emergent information (Value of Information) and the efficiency of informational exploration and sharing among agents (or modules) within IMRL (equilibrium quotient) which can be certainly be leveraged by much of the existing related research works.

There are several directions of future work from here. In this article, the *information explorer* was provided with the true state of the environment, which is impractical in various applications. Furthermore, we limited ourselves to only two agents interacting with each other; it would be interesting to extend this to multiple learning agents in the environment. Therefore, fruitful directions for future work would be to develop frameworks where the *information explorer* can also partially observe the environment and one which designs the complicated interactions among multiple learning agents, and possibly multiple *information explorer*s, in the environment.

References

[Abel *et al.*, 2017] David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. Agent-agnostic human-in-the-loop reinforcement learning. *arXiv preprint arXiv:1701.04079*, 2017.

[Banerjee and Stone, 2007] Bikramjit Banerjee and Peter Stone. General game learning using knowledge transfer. In *IJCAI*, pages 672–677, 2007.

[Chevalier-Boisvert *et al.*, 2018] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

[Devin *et al.*, 2017] Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey Levine. Learning modular neural network policies for multi-task and multi-robot transfer. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2169–2176. IEEE, 2017.

[Duan *et al.*, 2016] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl ˆ2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

[Dulac-Arnold *et al.*, 2013] Gabriel Dulac-Arnold, Ludovic Denoyer, Nicolas Thome, Matthieu Cord, and Patrick Gallinari. Sequentially generated instance-dependent image representations for classification. *arXiv preprint arXiv:1312.6594*, 2013.

[Griffith *et al.*, 2013] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. Georgia Institute of Technology, 2013.

[Gupta *et al.*, 2021] Nikunj Gupta, G Srinivasaraghavan, Swarup Kumar Mohalik, and Matthew E Taylor. Hammer: Multi-level coordination of reinforcement learning agents via learned messaging. *arXiv preprint arXiv:2102.00824*, 2021.

[Hernandez-Leal *et al.*, 2019] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

[Krening *et al.*, 2016] Samantha Krening, Brent Harrison, Karen M Feigh, Charles Lee Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 9(1):44–55, 2016.

[Levine *et al.*, 2016] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[Levine *et al.*, 2018] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.

[Mnih *et al.*, 2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

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

[Omidshafiei *et al.*, 2019] Shayegan Omidshafiei, Dong-Ki Kim, Miao Liu, Gerald Tesauro, Matthew Riemer, Christopher Amato, Murray Campbell, and Jonathan P How. Learning to teach in cooperative multiagent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6128–6136, 2019.

[Rosenfeld *et al.*, 2018] Ariel Rosenfeld, Moshe Cohen, Matthew E Taylor, and Sarit Kraus. Leveraging human knowledge in tabular reinforcement learning: A study of human subjects. *The Knowledge Engineering Review*, 33, 2018.

[Schaal and others, 1997] Stefan Schaal et al. Learning from demonstration. *Advances in neural information processing systems*, pages 1040–1046, 1997.

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

[Taylor and Stone, 2009] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

[Torrey and Taylor, 2013] Lisa Torrey and Matthew Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1053–1060, 2013.

[Wang *et al.*, 2016] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[Zhan *et al.*, 2016] Yusen Zhan, Haitham Bou Ammar, et al. Theoretically-grounded policy advice from multiple teachers in reinforcement learning settings with applications to negative transfer. *arXiv preprint arXiv:1604.03986*, 2016.

[Zhu *et al.*, 2017] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE, 2017.