Cost Efficient Exploration for Reinforcement Learning

Pankayaraj Pathmanathan
Department of Computer Science
University of Maryland
College Park
pan@umd.edu

Christabel Acquaye
Department of Computer Science
University of Maryland
College Park
cacquaye@umd.edu

Prannoy Namala
Department of Mechanical Engineering
University of Maryland
College Park
pnamala@umd.edu

Abstract

Exploration exploitation has long been the crux of reinforcement learning problems. As the scale of the problems and the sparsity of the rewards increase, the problem of exploration-exploitation tends to be sample hungry, which increases the complexity of the learning and leads us to find methods that explore thoroughly and efficiently. Recent works using Occupancy Information Ration (OIR) have shown that exploration can be improved by incorporating an entropy maximization of the steady-state probability distribution. In this work, firstly, we extend the work of OIR to perform a validation and comparative study on OIR in different scenarios and algorithms. Furthermore, we propose incorporating an efficiency objective into the OIR objective while maintaining the structured exploration.

1 Introduction

Reinforcement learning (RL), Sutton and Barto is a framework to represent a decision-learning problem in Markov Decision Problem (MDP) environments and learn to optimize the decision-making process to solve those environments. The crux of the optimization relies on finding the delicate balance between exploration and exploitation of the learning agent as it tries to solve the MDP. To elaborate further, exploitation captures the act of following the optimal decision process concerning the knowledge that has been learned so far, and exploration refers to making decisions that are not necessarily optimal for gathering knowledge about the MDP. Finding a balance between these two quantities is essential to ultimately learning an optimal decision process for the MDP. This exploration-exploitation dilemma motivates our current work on striking a balance between these two quantities while minimising the total cost. Recent works of Mnih et al. [2013], Lillicrap et al. [2015], Mnih et al. [2015], Silver et al. [2016] have shown that Deep Reinforcement Learning can solve large and complex MDPs while maintaining this balanced trade-off. However, information theory also presents an interesting approach in ensuring

In this work , we extend upon the work of OIR in two fold.

- Firstly, we extend the OIR work into different RL algorithms, and stress test them on different uncertain environments
- Secondly, we propose a modification to the OIR objective that incorporates the cost of exploration such that while ensuring a diverse exploration, we can also constrain it to be sample efficient.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

2 Related Works

Naive earlier exploration methods were inspired by the bandit theory, where the exploration is induced by a certain type of added noise in the decision process. Here the amount of noise allowed to be added to the decision-making process acted as the exploration vs. exploitation trade-off. However, as the environments become bigger and rewards become sparse, these types of exploration measures increase the sample complexity of the learning algorithms as these methods inherently lack a structured measure of exploration, and in the absence of constant structured feedback from the reward structure, the exploration essentially gets executed by random noise.

One such structured measure used for exploration is the curiosity measure which falls into a wider category of intrinsic motivation based methods. Curiosity has long been used as an intrinsic reward (reward generated internally by the agent) to improve an agent's learning process when the extrinsic reward (the reward given by the environment) is sparse, thus providing a comparatively more guided learning process. These intrinsic rewards can broadly be categorized as those who measure the novelty of a state **?**Lopes et al. [2012] or those who measure the agent's uncertainty in predicting the consequences of its actions Houthooft et al. [2016], Schmidhuber [2010]. The work in Pathak et al. [2017] used self-supervised prediction errors as an effective means to compute the curiosity. In particular, they introduced an inverse and a forward model, which measures the error in predicting its actions and the next states. This error was treated as a curiosity to improve the exploration abilities of RL agents. Variations of the intrinsic motivation incorporated knowledge based intrinsic motivations Aubret et al. [2019], Ladosz et al. [2022], Linke et al. [2020] that tries to maximise the diversity of states and competence based intrinsic motivations which maximise the diversity of skills as measured by the agent Colas et al. [2022].

Another line of work incorporate the information theory based measures in-order to induce a better exploration in the RL agent. Once such work is the work of Haarnoja et al. [2018] which proposes to the maximize the entropy of the policy while optimizing the expected returns which leads to a policy that induces multi modality in the policy while providing a better sample efficiency than the previous works. Rather the works of Hazan et al. [2019], Zhang et al. [2020] proposed to directly estimate and maximize the entropy of the steady sate probabilities itself in an efficent manner. Works of Russo and Roy [2017], Lu et al. [2022] proposed information ratios caprues information gain to a time-horizon in the future in tabular value function settings which required a level of extensions in the parameter space rather than the tabular setting. The work of Suttle et al. [2022] which is related to our line of work proposed an occupancy measure based information ratio that can be extended to the algorithms in parameter space as well. Even though these methods in a structured manner propose the exploration they don't explicitly consider the cost of exploration and assumes the exploration is cost free. But in real world scenarios this is further from the truth as exploration can be much more costlier. Thus we need to not only encure a sufficent exploration but also make a structural setup to make a trade-off such that less number of samples as possible would be used for exploration.

3 Methodology

3.1 Incoporating exploration cost in OIR

As the OIR objective $\rho(\theta) = \frac{J(\theta)}{\kappa + H(d_\theta)}$ is minimized, the agent is induced to explore in a nearly uniform manner as the entropy of the occupancy measure is maximized while minimizing the cost $J(\theta)$. Explicitly this measure does not constrain the agent to consider the cost of exploration; rather, it induces the agent to maximize the state space coverage in as many samples as possible. One potential way of characterizing the exploration cost to OIR is by associating the sample efficiency. To elaborate further, in finite state space MDP with tabular steady-state probability estimation case, as we let the number of samples collected per state go to infinity, we can better estimate the steady state probability and do a better exploration. But that comes at the cost of infinite exploration cost. As we compromise on the accuracy of the steady-state probability estimation, we can improve the exploration cost. Thus if we are to constrain the number of samples collected per state-action pair, we can find a compromise between the exploration and the cost of exploration. For an ergodic time-homogenous Markov chain $s_1, s_2, \ldots$ over a finite state space Ω with stationary distribution d let's define $N^t(\theta, s)$ as at time t the number of times a state s is visited by following a policy π_θ. Let N^t_{tot} be the total number states that have been visted upto time t. Here the steady state probability d_θ can be estimated by $\frac{N^t(\theta,s)}{N^t_{tot}}$.

Let $0 \leq \alpha \leq 1$ be a user defined parameter. Then an alternative objective function can be defied as follows.

$$N(\theta) = \sup_{s \in \Omega} n(s, \theta) \tag{1}$$

where $n(s)$ is the number of times a state is visited by the agent whose policy is defined by parameter θ. Now, we can define an alternative cost-aware ORI objective as follows.

$$\rho(\theta) = \frac{J(\theta) + \alpha.\mathbb{1}_{\sum_s N^t(\theta,s)/N^t_{tot} > N} \cdot \sum_t p}{\kappa + H^t\left(d_\theta\right)} \tag{2}$$

where p refers to a penalty in our case it was fixed negative valued constant. Here α is a predefined parameter that determines a limit on the probability of a certain state being visited over which we need to start implementing the penalty. As the entropy of the steady state probability is forced to be maximized, it would lead the agent to explore all the states. Since there is a penalization that is coming from an overly visited state it would be incentivized against avoiding that state.

3.1.1 Confidence level on the occupancy measure

One issue that we can run into while using the occupancy measure is that the measure can be varying at the beginning of the training thus causing the auxiliary reward we use to be stochastic. This can result in undesired results. In order to alleviate this problem we leverage the Hoeffding inequality in order to formulate confidence bound on the estimation of the occupancy measure. Consider a tabular setting where the occupancy measure is measures as an average $d_\theta = \frac{N(s,theta)}{N_tot}$. Let d^*_θ be the real occupancy measure that is induced by the policy π_θ parameterized by θ. Here $0 \leq d_\theta \leq 1$ since it is a probability.

$$P(|d_\theta - d^*_\theta| > \delta) \leq 2e^{-2n\delta^2} \tag{3}$$

Let's assume that we want d_θ to close as d^*_θ by at least an ϵ with a confidence level of $1 - \alpha$.

$$\begin{aligned} P(|d_\theta - d^*_\theta| > \delta) &\leq \alpha \\ 2e^{-2n\delta^2} &\leq \alpha \\ -2n\delta^2 &\leq \ln \frac{\alpha}{2} \\ n &\geq \frac{1}{2\delta^2} \ln \frac{2}{\alpha} \end{aligned}$$

Thus only when we have collected at least $\frac{1}{2\delta^2} \ln \frac{2}{\alpha}$ number of samples we then employ both the auxiliary rewards in order to avoid the unwanted stochasticity in the auxiliary rewards. Here α, δ are user defined hyperparameters.

3.1.2 Argument for the correlation between the training efficiency and the proposed auxilary reward

In general markovian settings, the convergence of the steady-state probability estimation to the true steady-state probability distribution is characterized by the total variation distance. To be precise for a total variation measure defined between two distributions P, Q

$$D_{\mathrm{TV}}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| \tag{4}$$

We can define a quantity $d_{mix}(t)$ which can in turn be used to define a quantity names mixing time τ_{mix}

$$\begin{aligned} d_{mix}(t) &= \sup_{s \in \Omega} D_{\mathrm{TV}}\left(P^t(s, \cdot), \mu\right) \\ \tau_{mix}(\epsilon) &= \inf t : d_{mix}(t) \leq \epsilon \end{aligned}$$

Essentially the mixing time definition characterizes the minimum number of time steps needed to be taken to get the accuracy of the steady state probability distribution estimation within an ϵ error. Since the definition is defined based on Total Variation distance the convergence won't happen in theory until the least explored state is visited enough. Thus if we are to follow an existing exploration strategy that is partial towards certain states then in-order for us to get a the expected convergence we need to over explore certain states. Thus directly penalizing the states that are over visiting can effectively help us redistribute the exploration budget towards the states that are the least favoured by the exploration mechanism thus effectively helping with a faster convergence.

As per the reward the results didn't show any significant differences. One potential reason we hypothesize for that was the small scale of the current environment.

3.2 OIR Validatation

For OIR validation, we aim to prove empirically that the results obtained by Suttle et al. [2022] are replicable. To achieve the above goal, we perform the following experiment. First, we consider the MiniGrid-Dynamic-Obstacles 16x16 for Dynamic Obstacles model. We use the Actor-Critic method as a baseline and Actor-Critic with OIR cost (referred to as *IDAC*). We expect successful replication of results from the Suttle et al. [2022] paper.

For implementing the `GridWorld`, we will use a python library **MiniGrid** (linked here). For implementing vanilla RL Algorithms, we will use the **StableBaselines3** (linked here), which provides a clean implementation of RL algorithms to expedite the process significantly compared to implementing the algorithms from scratch. The challenge in this part of our project will be implementing the OIR ratio. This can be achieved by implementing a tabular version of OIR by following the instructions in the Suttle et al. [2022] paper.

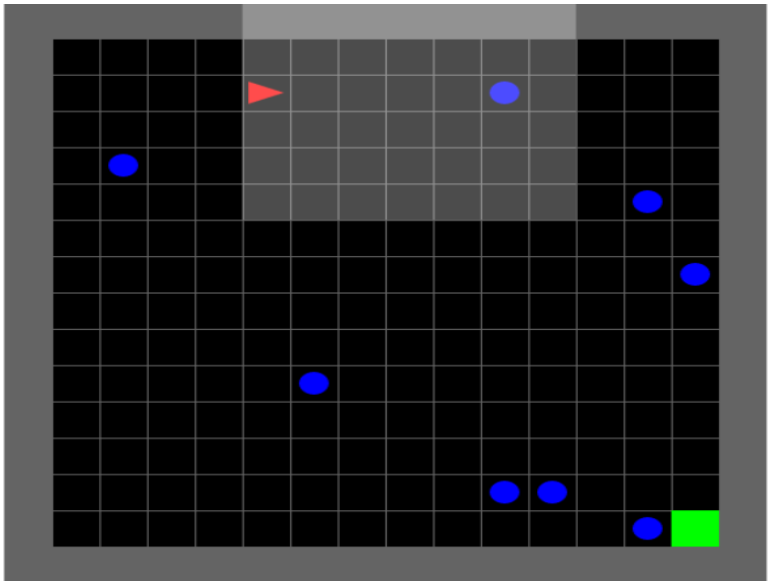

Figure 1: Grid World environment. Here the green block indicate the start state and the yellow block indicate the goal state

3.3 Comparative Study

From the previous part of our study, we attempt to prove that OIR can give us better results when incorporated with Vanilla RL Algorithms in highly uncertain environments. In our project, the aim is to attempt to "stress-test" OIR. This will be done by implementing OIR continuous spaces using density estimation techniques. Testing will be done using `WaterWorld` environment (linked here) as it is significantly more complicated than the `GridWorld` environment.

Similar to the earlier analysis, we will compare vanilla RL algorithms with the OIR versions. Multi-Agent Policy Gradient Algorithms such as APEX DDPG, and so on are used to solve the `WaterWorld`

problem. The challenge here would be integrating density estimation techniques as we are moving from implementing the tabular form of OIR to suit a continuous space problem.

WaterWorld Environment: `WaterWorld Environment` is a multi-agent environment. There are multiple agents whose aim is to gather food particles while avoiding the poison particles. The observation space for each agent contains information about the speed and velocity of the food particles and poison detected by the sensors. The dimensions of the observation space depend on the number of sensors which is a hyper-parameter. The action space has two dimensions and contains the horizontal and vertical thrust. The reward is multi-faceted as it depends on the cooperation between the agents, the capture of food, and the magnitude of thrust applied.

4 Results

4.1 Incorporating exploration cost in OIR

4.1.1 Experiment Setup

Initial experiments were done on a smaller 10x10 grid world environment as shown in Figure 2. Firstly, the original algorithm was evaluated in the Grid Environment without the use of an additional exploration mechanism such as epsilon greedy. Here the assumption from the works of Suttle et al. [2022] about the steady state probability was maintained. That is the steady state probability/ occupancy measure was treated as if it was coming from an oracle. In reality it was measure in a tabular form since the size of the state space was tractable in the problem at hand.

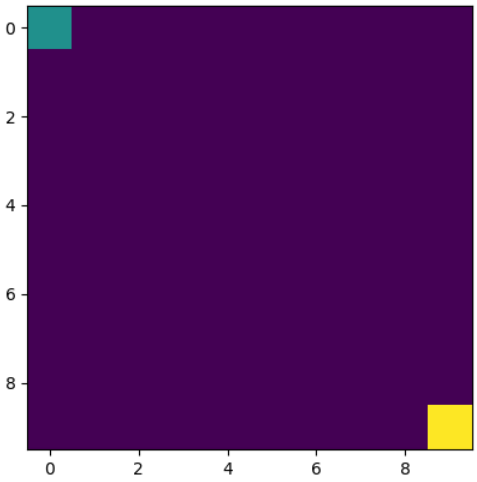

Figure 2: Grid World environment. Here the green block indicate the start state and the yellow block indicate the goal state

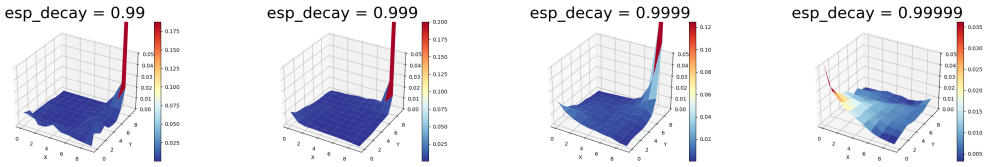

Figure 3: Occupancy measure induced by epsilon greedy algorithm for four epsilon decay values (0.99, 0.999, 0.9999, 0.99999). Here the x, y axis refers to the grid and the z axis refers to the steady state probability that is iduced by the policy and exploration. With further delay the agents does tend to explore more as expected. Also certain recurring states (start , end states) tend to dominate disproportionately more in the occupancy measure.

As we evaluated the ORI algorithm on it's own we found an interesting observation with regards to it's performance. When OIR alone was used an an objective the algorithm did fail to perform much of an exploration (we are yet to conclusively verify this claim via rigorous experimentation). But when paired the least performing epsilon greedy algorithm we where able to get a far better exploration results compared to the epsilong greedy algorithm alone.

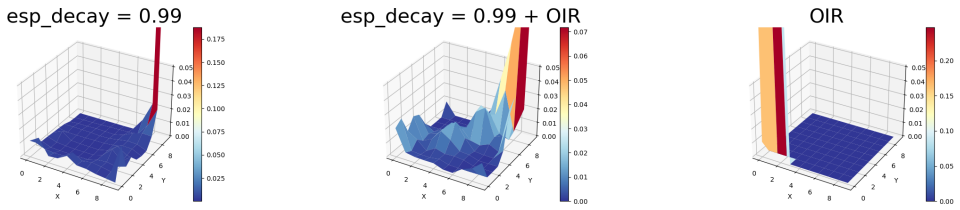

Figure 4: Occupancy measure induced by epsilon greedy algorithm, epsilon greedy + OIR algorithms, OIR algorithm. Here the x, y axis refers to the grid and the z axis refers to the steady state probability that is induced by the policy and exploration. Without some smaller initial exploration algorithm OIR alone tend to fail to generate a favourable exploration. But the when objective was paired with a worse performing epsilon greedy algorithm it did do a better exploration than the epsilon greedy algorithm alone

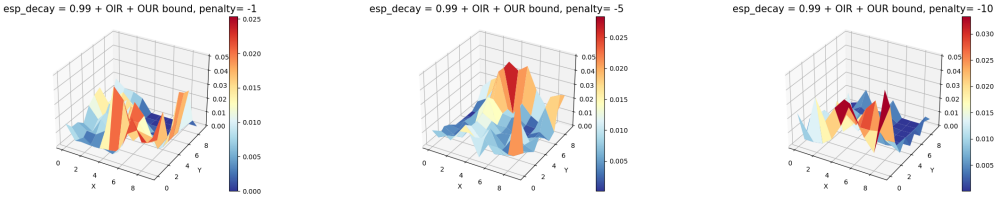

Figure 5: Occupancy measure induced by epsilon greedy + OIR + our penalty across different importance given to the penalty. Here the x, y axis refers to the grid and the z axis refers to the steady state probability that is induced by the policy and exploration. Compared to previous figures the addition of such penalty does enable a better overall exploration. The three figures denote the sensitivity analysis based on different levels of importance given to penalty (-1, -5, -10 respectively being the penalty regret for over visitation of a state).

4.2 OIR Validation and Comparative Study

As discussed in the Methodology section, the validation experiments for OIR was conducted in the `GridWorld` environment. We used the `WaterWorld` environment for the comparative study. We performed experiments to evaluate the baseline performance of RL algorithms without the OIR cost for these environments. The aim is to incorporate the OIR cost in the objective function of the RL algorithms to compare the performance against the baseline. The training plot from the baseline testing of `GridWorld` is in Fig 6. We use Actor-Critic RL Algorithm for `GridWorld`. The training plot from the baseline testing of `WaterWorld` is in Fig 7 for MAPPO and in Fig 8 for MADDPG.

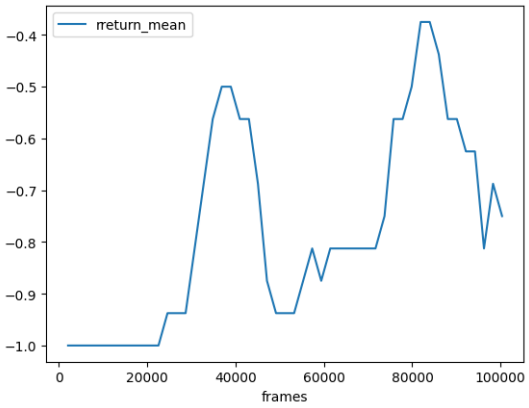

Figure 6: Training Plot for `GridWorld`. Training Conducted for 100,000 frames

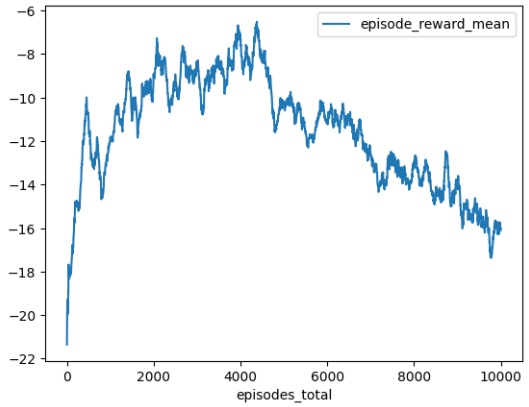

Figure 7: Training Plot for `WaterWorld` with PPO. Training Conducted for 10,000 episodes

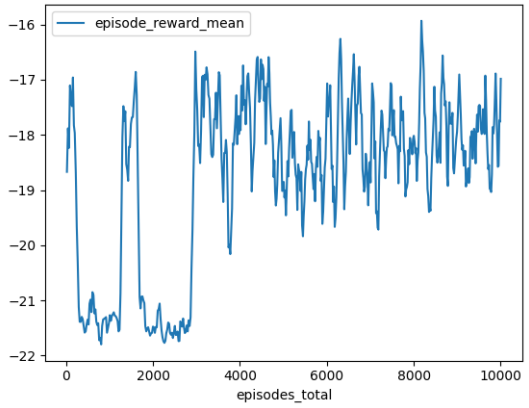

Figure 8: Training Plot for `WaterWorld` with DDPG. Training Conducted for 10,000 episodes

4.2.1 Observation

As we see from Figures 7 and 7, the training plots for MAPPO and MADDPG look vastly different. This difference can be due to the nature of the algorithms. PPO is trained "on-policy" and is a stochastic policy. The training is based on the trajectories generated from the new policy, and the update doesn't depend on the old policy. However, DDPG is trained "off-policy" and is a deterministic

policy. Generally, off-policy RL algorithms are notoriously unstable, contributing to the difference in the training plots for both algorithms.

4.2.2 Challenges

We faced a few challenges in achieving the expected target. The main challenge was implementing the Density Estimator for continuous spaces for the multi-agent setting. As mentioned previously, the paper Suttle et al. [2022] have used an Oracle Density Estimator. We have tried using a learning-based density estimator for a more practical implementation. However, directly implementing such a density estimator for a multi-agent problem was too arduous, and focusing on that didn't give us enough time to focus on the grid world part. Given more time, we would be able to implement these density estimators for the objective function.

5 Future Work

One particular limitation of this work is that the setting and the bounds derived are limited to a tabular setting which may not hold in a setting where the environment can be large and a function approximator is used. As we move away from the tabular setting to a general estimation setting one possible direction worth exploring is the mixing time literature and finding ways to associate the mixing time with the accuracy to find a cost-effective exploration objective. As per the problem of an intractable/continuous state space, one possible solution is exploring the discretization of the state space with some relevancy between the states that fall on a single group. One such relevancy that can be explored is the adjacency measure.

6 Reproducibility Checklist

1. **Code used to eliminate or disprove claims and to conduct and analyze the experiments (2.1)**: `https://github.com/punk95/CMSC742`
2. **Code used to run experiments for PPO and DDPG algorithms in the Waterworld environment (2.2, 2.3)**: `https://github.com/AcquayeChristabel/Cost-Efficient-RL.git`
3. **Hyper Parameters**: In all implementation a network of hidden layer size of [10,10] was used. Seeds : A different randomly generated seed was used for every individual experiment.
4. **Statistics used for the results** : We have used the undiscounted reward as our statistic to evaluate the algorithms along with the occupancy measure to calculate the amount of exploration that was done.

7 Contribution Summary

1. **Section 2.1**: Pankayaraj
2. **Section 2.2, 2.3**: Christabel, Prannoy

References

Arthur Aubret, Laëtitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *ArXiv*, abs/1908.06976, 2019.

Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. 2018. doi: 10.48550/ARXIV.1801.01290. URL `https://arxiv.org/abs/1801.01290`. Publisher: arXiv Version Number: 2.

Elad Hazan, Sham M. Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Curiosity-driven exploration in deep reinforcement learning via bayesian neural networks. *CoRR*, abs/1605.09674, 2016. URL http://arxiv.org/abs/1605.09674.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and HyondongOh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, sep 2022. doi: 10.1016/j.inffus.2022.03.003. URL https://doi.org/10.1016%2Fj.inffus.2022.03.003.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2015. URL https://arxiv.org/abs/1509.02971.

Cam Linke, Nadia M. Ady, Martha White, Thomas Degris, and Adam White. Adapting behaviour via intrinsic reward: A survey and empirical study, 2020.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 206–214, Red Hook, NY, USA, 2012. Curran Associates Inc.

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement learning, bit by bit, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling, 2017.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation(1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368.

David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.

Wesley A. Suttle, Alec Koppel, and Ji Liu. Occupancy information ratio: Infinite-horizon, information-directed, parameterized policy search. *CoRR*, abs/2201.08832, 2022. URL https://arxiv.org/abs/2201.08832.

Richard S Sutton and Andrew G Barto. Reinforcement Learning: An Introduction.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities, 2020.