ExTra: Transfer-guided Exploration*

Extended Abstract

Anirban Santara
Indian Institute of Technology Kharagpur
Kharagpur, WB, India
nrbnsntr@gmail.com

Rishabh Madan
University of Washington
Seattle, Washington
rishabhmadan96@gmail.com

Pabitra Mitra
Indian Institute of Technology Kharagpur
Kharagpur, WB, India
pabitra@cse.iitkgp.ernet.in

Balaraman Ravindran
RBC-DSAI, Indian Institute of Technology Madras
Chennai, TN, India
ravi@cse.iitm.ac.in

ACM Reference Format:
Anirban Santara, Rishabh Madan, Pabitra Mitra, and Balaraman Ravindran. 2020. ExTra: Transfer-guided Exploration. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020*, IFAAMAS, 3 pages.

1 INTRODUCTION

The sample efficiency and convergence time of a Reinforcement Learning (RL) algorithm depend heavily on the exploration method used by the agent. In this work, we formulate an exploration method that uses prior experiences of an agent at similar tasks in other environments for improving the efficiency of exploration in the current task-environment. We show that given an optimal policy in a related task-environment, its bisimulation distance from the current task-environment gives a lower bound on the optimal advantage of state-action pairs in the current task-environment. *Bisimulation*, first introduced for MDP by Givan et al. [4], is a relation that draws equivalence between the states of a Markov Decision Process (MDP) that have the same long-term behavior. It is equivalent to *MDP homomorphism* [5, 6]. Ferns et al. [3] proposed *bisimulation metric* as a quantitative analogue of the bisimulation relation that can be used as a notion of distance between states of an MDP. This was extended by Taylor et al. [10] as the *lax bisimulation metric* in order to measure the distance between state-action pairs of different MDPs. This metric was later used by Castro and Precup [2] for policy transfer between MDPs which motivated this paper.

2 PROPOSED FRAMEWORK

Let us consider two MDPs, $\mathcal{M}_1 = \langle S_1, A_1, P_1, R_1 \rangle$ and $\mathcal{M}_2 = \langle S_2, A_2, P_2, R_2 \rangle$, where S_i, A_i, P_i and R_i respectively denote the state space, action space, transition probability function and reward function of the i^{th} MDP. Let V_i^* and Q_i^* denote the optimum state and state-action value functions and π_i^* denote the optimum policy of the i^{th} MDP. Let $d_\approx : S_1 \times S_2 \times A_2 \rightarrow \mathbb{R}$ and $d'_\approx : S_1 \times S_2 \rightarrow \mathbb{R}$ denote the state-action lax bisimulation and state lax bisimulation metrics respectively [2]. Then the following results hold true. Please refer to the full paper [7] for the proofs.

Algorithm 1 ϵ-greedy Q-learning with Transfer-guided Exploration (ExTra)

```
Compute d_≈(s_1, (s_2, a_2))
step = 0
while step < MAXSTEPS do
    with probability ε
        a_2 ~ π_ExTra(·|s_2, M_1, π_1^*)
    with probability 1 − ε
        a_2 ← arg max_a Q_2(s_2, a)
    reward = take_step(a_2)
    update_Q(Q_2(s_2, a_2), reward)
    step = step + 1
end while
```

Lemma 2.1. $\forall s_1 \in S_1, \forall s_2 \in S_2, \forall a_2 \in A_2, |V_1^*(s_1) - Q_2^*(s_2, a_2)| \leq d_\approx(s_1, (s_2, a_2))$.

Corollary 2.2. $\forall s_1 \in S_1, \forall s_2 \in S_2, |V_1^*(s_1) - V_2^*(s_2)| \leq d_\approx(s_1, (s_2, \pi_2^*(s_2)))$.

This leads us to the following theorem that forms the backbone of our proposed exploration algorithm.

Theorem 2.3. *Given MDPs,* $\mathcal{M}_1 = \langle S_1, A_1, P_1, R_1 \rangle$ *and* $\mathcal{M}_2 = \langle S_2, A_2, P_2, R_2 \rangle$ *and bisimulation metric* $d_\approx : S_1 \times S_2 \times A_2 \rightarrow \mathbb{R}$ *we have* $\forall s_2 \in S_2, \forall a_2 \in A_2$

$$A_2^*(s_2, a_2) \geq -d_\approx(s_{match}, (s_2, a_2)) - \beta(s_2),$$

Where $A_2^*(s_2, a_2)$ *is the optimum advantage function in* $\mathcal{M}_2$, $s_{match} = arg\max_{s_1 \in S_1} V_1^*(s_1) - d'_\approx(s_1, s_2)$ *and* $\beta(s_2) = d_\approx(s_{match}, (s_2, \pi_2^*(s_2)))$.

Thus, given a related MDP $\mathcal{M}_1$, its bisimulation distance from the current MDP $\mathcal{M}_2$ gives a lower bound on the optimal advantage of state-action pairs in the current MDP. We define *bisimulation advantage* of an action $a_2 \in A_2$ in a state $s_2 \in S_2$ as this bound.

Definition 2.4. **Bisimulation Advantage**: Given MDPs, $\mathcal{M}_1 = \langle S_1, A_1, P_1, R_1 \rangle$ and $\mathcal{M}_2 = \langle S_2, A_2, P_2, R_2 \rangle$ and bisimulation metric $d_\approx : S_1 \times S_2 \times A_2 \rightarrow \mathbb{R}$ we define the bisimulation advantage of an action $a_2 \in A_2$ in a state $s_2 \in S_2$ as:

$$A_\approx(s_2, a_2) = -d_\approx(s_{match}, (s_2, a_2)) - \beta(s_2)$$

*Equal contribution by the first two authors. Research by the second author was conducted while the author was a student at IIT Kharagpur.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

In Transfer-guided Exploration (ExTra), the agent samples actions from a maximum entropy distribution [12] over the bisimulation advantages, defined as follows:

$$\pi_{\text{ExTra}}(a_2|s_2, \mathcal{M}_1, \pi_1^*) = \frac{e^{A_{\approx}(s_2,a_2)}}{\sum_{a\in A_2} e^{A_{\approx}(s_2,a)}} \quad (1)$$

Since the optimal policy in the target MDP, π_2^*, is not known during learning, $\beta(s_2)$ can not be known exactly. Since $\beta(s_2)(\geq 0)$ is the same for all actions in a given state s_2, replacing $\beta(s_2)$ with a real positive number preserves the order of probability values assigned to different actions by π_{ExTra}. If the transfer is successful, π_{ExTra} would assign higher probabilities to the optimal actions and thus help the agent arrive at the optimal policy quickly. However, in the event of an unsuccessful transfer, π_{ExTra} may be biased away from the optimal actions. This may cause the agent to remain stuck with the wrong actions for long periods. To help the agent recover from the effect of unsuccessful transfer, we set $\beta = \alpha n$, where $\alpha \in \mathbb{R}^+$ is a tunable hyperparameter and n is the current step number. As n grows, $\pi_{\text{ExTra}}(\cdot|s_2, \mathcal{M}_1, \pi_1^*)$ tends to a uniform distribution over actions, thus annealing the influence of transferred knowledge on exploration with time. This does not hurt the agent's learning in states where the transfer was successful because the agent happens to have explored the optimal actions early on in training in those states. Note that changing β does not affect the rate of exploration; instead, it merely changes the shape of the probability distribution from which the agent samples actions during exploration.

3 EMPIRICAL EVALUATION

In this section, we evaluate the viability of ExTra through empirical analysis of its performance on stochastic grid-world environments. Please refer to the full paper [7] for details of the environments. We use Area under the Mean Average Reward curve (AuC-MAR) as an objective measure of the rate of convergence [11]. We address the following questions:

How does ExTra compare against traditional exploration methods?
We compare ϵ-greedy ExTra (Algorithm 1) with traditional approaches namely ϵ-greedy, MBIE-EB [8], Pursuit [9] and Softmax [9] for navigation in four, six and nine large room environments. The source environment for ExTra has four small rooms. Each large room in the target environments has six small rooms inside it. We observe that the ExTra agent consistently achieves faster convergence in all the three environments. This corroborates our claim that ExTra can achieve faster convergence and hence superior sample efficiency if we have access to the optimal policy in a related task-environment.

How sensitive is ExTra to the choice of source task?
In our *first* study, we consider transfer between tasks that share the same state-action space and reward structure but differ in goal positions. We train source policies for five different goal positions, each in a different room, in the six large room environment and transfer to a sixth goal position. We observe that each of our ExTra agents fetch higher AuC-MAR values than any of the traditional methods, thus demonstrating the efficacy of ExTra. Also, there is a rough trend of the AuC-MAR values decreasing with increasing distance of goal in the source task which demonstrates graceful degradation of performance.
In our *second* study, we compare transfer from source tasks that differ in state space, action space, reward structure, goal structure, goal distribution and transition dynamics. We observe that ExTra is able to leverage knowledge about the transition dynamics of the source MDP even when the reward structures and goal distributions are different. Also, ExTra is more sensitive to the reward structure of the source MDP than transition dynamics or goal distribution and source MDPs with the same action space as the target are more preferable even if the state spaces are different.

Can ExTra enhance the performance of other exploration algorithms that only use local information?
We formulate ϵ-greedy versions of each of our baseline algorithms in the first experiment ($\epsilon = 0.5$), where the agent samples actions from π_{ExTra} with probability $\epsilon = 0.5$ and follows the main algorithm rest of the time. The source environment has four small rooms, and the target has six large rooms. We observe that ExTra achieves gains in performance of traditional exploration algorithms when used in conjunction, which proves its viability as a complementary exploration method for accelerating the rate of convergence of traditional RL algorithms.

How does ExTra compare against Bisimulation Transfer?
We answer this by comparing the rates of convergence of $\epsilon-$greedy Q-learning with ExTra and the bisimulation transfer algorithm of Castro et al. [2] that initializes the Q-matrix with the Q-value of the transferred policy. We choose navigation in four small rooms as the source task. The target tasks are navigation in four large rooms (similar to the source task) and a modified version of the `Taxi-v2` environment of OpenAI Gym [1] (drastically different from the source task). We observe that when the source and target tasks are similar and bisimulation policy transfer is successful, Q-learning initialized with transfer gets an initial jumpstart while ExTra catches up later. But when the source and target tasks are drastically different, it converges slower than ExTra. Since bisimulation distances are larger for dissimilar environments, π_{ExTra} tends to a uniform distribution over target actions. As a result, ExTra falls back to vanilla $\epsilon-$greedy Q-learning with uniform sampling. On the other hand, Q-learning initialized with bisimulation transfer has to first recover from the effect of negative transfer using $\epsilon-$greedy uniform exploration before it can start learning. While bisimulation transfer can be both effective and detrimental depending on how the source and target tasks are related, ExTra does not negatively affect the learning process even when the source and target task-environments are drastically different.

4 CONCLUSION

In this work, we present a novel transfer guided exploration algorithm, ExTra, that achieves faster convergence compared to traditional exploration methods that only use local information, is robust to source task selection with predictable graceful degradation of performance and can compliment traditional exploration methods by improving their rate of convergence. In our future work we plan to extend ExTra to larger state-action spaces and continuous control tasks.

Acknowledgements. Anirban Santara's work was supported by Google India under the Google India PhD Fellowship Award.

REFERENCES

[1] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540* (2016).

[2] Pablo Samuel Castro and Doina Precup. 2010. Using bisimulation for policy transfer in MDPs. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[3] Norm Ferns, Prakash Panangaden, and Doina Precup. 2004. Metrics for finite Markov decision processes. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 162–169.

[4] Robert Givan, Thomas Dean, and Matthew Greig. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147, 1-2 (2003), 163–223.

[5] Balaraman Ravindran. 2003. SMDP homomorphisms: An algebraic approach to abstraction in semi markov decision processes. (2003).

[6] Balaraman Ravindran and Andrew G Barto. 2002. Model minimization in hierarchical reinforcement learning. In *International Symposium on Abstraction, Reformulation, and Approximation*. Springer, 196–211.

[7] Anirban Santara, Rishabh Madan, Balaraman Ravindran, and Pabitra Mitra. 2019. Extra: Transfer-guided exploration. *arXiv preprint arXiv:1906.11785* (2019).

[8] Alexander L Strehl and Michael L Littman. 2005. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 856–863.

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

[10] Jonathan Taylor, Doina Precup, and Prakash Panagaden. 2009. Bounding performance loss in approximate MDP homomorphisms. In *Advances in Neural Information Processing Systems*. 1649–1656.

[11] Matthew E Taylor and Peter Stone. 2009. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10, Jul (2009), 1633–1685.

[12] Joy A Thomas and TM Cover. 1991. Elements of information theory. *John Wiley & Sons, Inc., New York. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, MPH (2009),"Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," Journal of the Royal Society Interface* 6 (1991), 187–202.