

Abstract Demonstrations and Adaptive Exploration for Efficient and Stable Multi-step Sparse Reward Reinforcement Learning

Xintong Yang, Ze Ji
School of Engineering
Cardiff University
Cardiff, UK

yangx66@cardiff.ac.uk, jiz1@cardiff.ac.uk

Jing Wu, Yu-Kun Lai
School of Computer Science and Informatics
Cardiff University
Cardiff, UK

wuj11@cardiff.ac.uk, laiy4@cardiff.ac.uk

Abstract—Although Deep Reinforcement Learning (DRL) has been popular in many disciplines including robotics, state-of-the-art DRL algorithms still struggle to learn long-horizon, multi-step and sparse reward tasks, such as stacking several blocks given only a task-completion reward signal. To improve learning efficiency for such tasks, this paper proposes a DRL exploration technique, termed A^2 , which integrates two components inspired by human experiences: Abstract demonstrations and Adaptive exploration. A^2 starts by decomposing a complex task into subtasks, and then provides the correct orders of subtasks to learn. During training, the agent explores the environment adaptively, acting more deterministically for well-mastered subtasks and more stochastically for ill-learned subtasks. Ablation and comparative experiments are conducted on several grid-world tasks and three robotic manipulation tasks. We demonstrate that A^2 can aid popular DRL algorithms (DQN, DDPG, and SAC) to learn more efficiently and stably in these environments.

I. INTRODUCTION

Deep Reinforcement Learning (DRL) has achieved exciting advances recently in robotics [1]. However, DRL agents still struggle to solve many robotic tasks with long horizon, multiple steps and sparse rewards. For example, in Fig. 1, a robot is asked to push a block into a closed chest, where the chest needs to be opened before the block can be pushed into it, and only a reward is given when the task is done. Fortunately, humans provide many useful insights for mastering such tasks. In this paper, we revisit and integrate two ideas inspired from human experiences to make DRL more efficient and stable in these difficult tasks.

In the real world, humans benefit substantially from decomposing a complex task into a sequence of subtasks. This helps divide the task into a number of steps that are easier and faster to solve [2]–[4]. We can then accomplish the complex task by achieving each step in a specific order, such as following an instruction manual to assemble different parts of a piece of furniture. In addition, skills developed during learning one step can be reused in other steps, such as skills to stack a specific block can be reused to stack others. The example in

Xintong Yang thanks the Chinese Scholarship Council (CSC) for providing the living stipend for his Ph.D. programme (No. 201908440400).



Fig. 1: Visualisation of a pushing task.

Fig. 1 shows that the three steps of the pushing task are easier to achieve individually and they share similar arm motions (reaching and pushing).

Inspired by this, we propose firstly to leverage **abstract demonstrations**, which provide the correct order of steps to a learning agent instead of low-level motions. It has been shown that, given an efficient enough algorithm to learn the subtasks such as hindsight experience replay, abstract demonstrations can accelerate learning for multi-step tasks [5]. The main benefits of abstract demonstrations when compared to demonstrations of motion trajectories [6], [7] are: 1) they do not encode a specific pattern of the behaviours when solving a task; 2) they release robotic operators from the tedious processes of collecting motion trajectory data.

The second idea is inspired from humans’ **adaptive exploration** strategies. For example, children do not try to pick and stack Lego blocks randomly when they have already known which blocks they need and how to stack them together. This relates closely to the trade off between exploration and exploitation, a core issue in DRL [8]. Although exploration methods have been proposed for long-horizon tasks [9], its adaptivity has rarely been studied in the context of multi-step tasks. In multi-step tasks, the learning progress of later steps depends heavily on former steps. For example, stacking the fourth block requires the previous three blocks to have been stacked well. As a result, an agent that explores constantly or decays exploration in a task-agnostic way, which is the default setting in most works [10], will have difficulty approaching later steps. With this in mind, we propose to adjust exploration adaptively based on an agent’s performance on each of the task

steps.

Previously, these ideas have been explored individually (see a detailed review in Section II). In this paper, we highlight the effectiveness of their integration into DRL algorithms for learning multi-step, sparse reward tasks. We conducted a series of ablation studies on three representative off-policy DRL agents, i.e., DQN (deterministic, discrete actions) [11], DDPG (deterministic, continuous actions) [12] and SAC (stochastic, continuous actions) [13], using the Mini-grid [14] and Pybullet Multigoal environments [15].

The results show that **abstract demonstrations** can accelerate learning and improve performance slightly in the grid-world tasks (DQN) and significantly in continuous control tasks (DDPG and SAC). **Adaptive exploration** does not further improve task success rates in general, but it does stabilise learning with reduced performance variances. In short, our contributions include:

- We propose **A²**, a method which integrates Abstract demonstrations and Adaptive exploration, for more efficient and stable multi-step sparse reward DRL.
- We discuss possible implementations of the adaptive exploration component of **A²** for modern DRL algorithms, including discrete and continuous actions, deterministic and stochastic policies.
- We validate the effectiveness of **A²** in various settings: with a deterministic DQN agent (discrete actions, grid-world tasks), a deterministic DDPG agent (continuous actions, robotic tasks) and a stochastic SAC agent (continuous actions, robotic tasks).

Paper organisation: Section II draws connections between **A²** and existing works. Section III briefly reviews important preliminaries for using **A²**. Section IV describes **A²** in details. Section V presents experimental results and discussions. Section VI concludes the paper.

II. RELATED WORK

Long-horizon multi-step tasks have been addressed by task and motion planning (TAMP) [2]. TAMP typically generates high-level task sequences and plans motions for each subtask individually [3], [16]. Although the idea of task decomposition based on human priors has been around for decades [17], researchers have just started recently to study how it can help learning-based methods [18], [19]. Many works in RL focus on learning sub-policies (or options) that are specific to different subtasks or manifest different skills [20], [21]. Our work differs from them in that we decompose a task into subtasks to guide the learning of a control policy (as demonstrations), while they require an agent to learn to decompose a task during learning. In principle, our method can be incorporated with any task decomposition method, be it based on human priors or learning [22].

Learning from demonstrations (LfD) is a practical approach to teach robots behaviours using trajectories collected by another agent, typically a human. These behaviours are normally difficult to program [7]. DRL also benefits from demonstrations in sparse reward and/or long-horizon tasks

(also known as imitation learning) [6]. Different from the mainstream works, which use trajectories at the control command level [9], we use an abstract form of demonstrations, i.e., the correct sequence in which task steps should be performed and learnt. This reduces the human labour spent on collecting long robot motion trajectories. For cases where some subtasks are too difficult to learn, motion trajectories can still be used to further accelerate learning for the subtasks. In order words, abstract demonstrations could be an alternative or a supplement to motion trajectories demonstrations.

Automatic curriculum learning (ACL) has become increasingly active in the field of RL and robotics [23]. Briefly, ACL methods automatically adapt the distribution of training data for a learning agent [23]. Our method can be seen as a contribution to ACL, which 1) prompts the agent to learn from easier subtasks to harder subtasks in the correct sequences (abstract demonstrations) and 2) adjusts the agent’s exploration behaviours based on its performance on each subtask.

III. PRELIMINARIES

Markov Decision Process is a tuple $\langle S, A, R, p, \gamma, \rho_0 \rangle$, where S is the state space, A the action space, $R(s, a)$ the reward function, $p(s'|s, a)$ the system transition model, γ the discount factor and ρ_0 the initial state distribution. A policy $\pi(a|s)$ is a mapping from states to actions. A state-action (Q) value function, $Q^\pi(s_0, a_0)$, is defined as the expected, discounted and accumulated rewards starting from taking an action a at state s and following a policy π thereafter, i.e., $Q^\pi(s_0, a_0) = \mathbb{E}_{a \sim \pi, s \sim p} [\sum_{t=0}^T \gamma^t R(s_t, a_t)]$, where T is the maximal number of task timesteps. The goal of an RL algorithm is to find an optimal policy that maximises the value function [8].

Goal-conditioned Reinforcement Learning: The GRL problem operates on an MDP augmented with a goal space G [24]. Instead of pursuing a single goal (conveyed by a single-objective reward function) as in standard RL, a GRL agent seeks to maximise a universal value function, $Q^\pi(s, g, a)$. Commonly, a goal is defined as some kind of transformation of a state, $g = m(s)$, assuming that, at any state, there is always a goal that is achieved if the agent arrives at that state.

Deep Q Network: DQN is an off-policy, deterministic RL algorithm that uses neural networks to approximate the Q function for discrete action tasks [11]. Commonly, DQN uses the *epsilon-greedy* (EGr) method for exploration. It takes a random action with a probability ϵ and takes an action according to the learnt Q function with a probability $1 - \epsilon$.

Deep Deterministic Policy Gradient: DDPG is an off-policy actor-critic style DRL algorithm developed for continuous control tasks [12]. It approximates a Q-function and a deterministic policy using separate neural networks, alternating the updates of the two with some interval. Originally, Ornstein–Uhlenbeck noises are added to the learnt policy to form exploratory actions [12]. Recently, researchers tend to use a simpler, but efficient exploration strategy, which takes random actions with a probability of ϵ and takes learnt actions with Gaussian noises with a probability of $1 - \epsilon$ [24]. For

convenient reading, we name this strategy *epsilon-Gaussian* (EGa).

Soft Actor Critic: SAC is an off-policy, stochastic DRL algorithm [13]. It updates the policy and value network with entropy regularisation. Typically, the policy network outputs the mean and standard deviation for a Gaussian policy. SAC controls its exploration by adjusting the temperature parameter α either manually or automatically. One can also alter the standard deviation of the policy to directly control exploration in a heuristic manner.

Hindsight Experience Replay: HER is a goal-relabelling technique for GRL, which copies transitions and replaces their desired goals with some other goals obtained by a sampling strategy [24]. It improves learning significantly in goal-conditioned sparse reward tasks. Originally, Andrychowicz et al. proposed in [24] the `final`, `random`, `episode` and `future` strategies to sample goals for replacement, and demonstrated that the `future` strategy is the most efficient. It copies a transition k times, samples k future transitions, and replaces the desired goals for each copied transition with the sampled ones. We adopt the `future` strategy in all our experiments, with $k = 4$.

IV. METHODS

This section illustrates our proposed \mathbf{A}^2 method in details. Overall, it includes two components. First, abstract demonstrations are provided to guide the robot to learn subtasks in the correct order, leading to the completion of a task. Secondly, an adaptive exploration technique adjusts the exploration parameters to achieve faster and more stable learning.

A. Abstract demonstrations

We propose to first decompose a multi-step task into a sequence of subtasks that are easier to accomplish. We observed that, in many tasks (e.g., stacking blocks), although the decomposed steps may have different purposes (e.g., stacking each of the blocks), the underlying behaviours achieving them often share common characteristics (e.g., move-pick-move-place). This inspired us to use goal-conditioned reinforcement learning (GRL) [24] as our main learning framework. The reason is that GRL was proved to have the potential to learn shared representations and knowledge between the steps [18], and Hindsight Experience Replay (HER) enables a GRL agent to learn short-term goals efficiently with sparse reward signals [24].

The motivation of using abstract demonstrations over demonstrations at the control command level (kinematic demonstrations) is twofold. First, abstract demonstrations do not encode a specific pattern of behaviours for solving a subtask. Second, it releases humans from the tedious processes of collecting demonstrations of motion trajectories. However, in theory these two kinds of demonstrations could be used together in cases where the subtasks are difficult to learn without kinematic demonstrations. Given a task decomposition scheme, we then leverage human priors to label the correct sequence to achieve the subtasks. Regarding implementation, an

abstract demonstration is represented by an ordered sequence of indexes, each corresponding to a subtask.

At the beginning of a learning episode, a final goal is sampled and given to the GRL agent. When abstract demonstrations are used, the agent is asked to learn the subtasks following the order given by the demonstration, instead of being directly given the final goal. The agent is given the next subtask if the current goal has been achieved. When given a new goal, the previous transitions within this episode are copied as a new trajectory and their desired goals are replaced with the new goal. New transitions associated with the new goals are then appended to the new trajectory, leaving the previous trajectory unchanged. This is to guarantee that the agent can finally learn to achieve all the goals without demonstrations. The episode ends when the number of maximum timesteps is reached.

We use a hyperparameter, $\eta \in [0, 1]$, to control the proportion of training episodes that are demonstrated. We follow the training procedure used in [24], which is organised in epochs, cycles and episodes. Each epoch has a number of cycles, I , and each cycle has a number of episodes, J . The number of demonstrated episodes in a cycle is computed by $episode_{demo} = \eta \times J$.

B. Adaptive exploration

As mentioned above, the idea of adaptive exploration in multi-step task learning is to reduce unnecessary exploration for well-learnt steps and increase on unfamiliar steps. Coupled with abstract demonstrations, this leads to an adaptive curriculum that guarantees the GRL agent to proceed learning towards the final goal as quickly as possible. Reducing unnecessary exploration also leads to more stable performance. Since deterministic and stochastic agents explore differently, we provide different implementations for them as follows.

Deterministic agents explore the environment using a base behavioural policy, $\pi_b(a|s, g)$, that contains some hyperparameters to control the exploratory behaviours. In this paper, we use the *epsilon-Gaussian* (EGa) strategy as our base behavioural policy for the DDPG agent, and the *epsilon-greedy* (EGr) strategy for the DQN agent.

The EGr strategy takes a random action with a probability ϵ or takes an action that maximises the learnt Q-function. In practice, ϵ will decay to a lower bound during the course of training, commonly using the following equation:

$$\epsilon = \epsilon_{end} + (\epsilon_{start} - \epsilon_{end}) \times e^{-\frac{n}{\beta}}$$

where, ϵ_{start} and ϵ_{end} are the upper and lower bounds, n is the total elapsed environment timesteps and β is the decay coefficient parameter. Intuitively, ϵ decays as the number of elapsed timesteps grows and its speed is controlled by β . To make it adaptive, we simply replace the exponential term by a performance metric. Specifically, we construct an N -dimensional vector to store the value of ϵ for each task step at the m -th epoch, ϵ_m . We then use an N -dimensional vector to record the test success rates for all task steps at the m -th

epoch, denoted as S_m , initialised to 0. We update ϵ_m using the following equations:

$$\epsilon_{m+1} = \epsilon_{end} + (\epsilon_{start} - \epsilon_{end}) \times (1 - S_m). \quad (1)$$

For continuous action cases, the EGa strategy samples a random action uniformly with a probability ϵ or takes an action generated by a learnt policy with noises from a fixed Gaussian distribution. That is,

$$\pi_b(a|s, g) = \begin{cases} a \sim \mathcal{U}(|A|, -|A|), & \delta \leq \epsilon \\ a \sim \mathcal{N}(\pi(a|s, g), \sigma), & \delta > \epsilon \end{cases}$$

where $\delta \sim \mathcal{U}(0, 1)$ and $\pi(a|s, g)$ is a learnt, deterministic, goal-conditioned policy. By varying the hyperparameters, ϵ and σ , based on the performance of each task step at different training epochs, we have an adaptive exploration strategy. Again, we use N -dimensional vectors to represent these two hyperparameters for the N task steps at the m -th epoch, ϵ_m and σ_m . At the beginning of a training process, these two vectors are initialised to their initial values, ϵ_0 and σ_0 ($\epsilon_0 = 0.2$ and $\sigma_0 = 0.05$ in all our experiments). We use another vector to record the test success rates and update ϵ_m and σ_m by:

$$\epsilon_{m+1} = \epsilon_0 \times (1 - S_m), \quad \sigma_{m+1} = \sigma_0 \times (1 - S_m). \quad (2)$$

Stochastic agents explore the environment by sampling from a learnt stochastic policy. As an example, the SAC agent uses a Gaussian policy whose mean and deviation are produced by a neural network. We propose to use the success rate as a scaling factor for the deviation for different task steps. Similar to Eq. 2:

$$\sigma_{m+1} = \sigma_0 \times (1 - S_m), \quad (3)$$

where σ_m is the deviations for the learnt stochastic policy for different task steps, and σ_0 can be pre-defined constants or learnt by a neural network.

Performance metrics. To obtain the test success rates, we perform K testing episodes for each task step after each training epoch and calculate the average success rate for each step. This test is run with abstract demonstrations to reflect the performance of achieving a step starting from the previous one. However, computing Eqs. 1, 2 and 3 with the vanilla success rate would result in bumping changes of the exploration hyperparameters. As the success rate is calculated only from the K testing episodes, it may not reflect the true performance of the current policy, because K is usually not large for the sake of reducing computations. Thus, we use the polyak-average [25] of the test success rate vector, S_m^- to calculate Eqs. 1, 2 and 3 instead of the vanilla success rates. The running average is calculated with a parameter $\tau_S \in [0, 1]$ as follows:

$$S_m^- = (1 - \tau_S) \times S_{m-1}^- + \tau_S \times S_m \quad (4)$$

To demonstrate the proposed A^2 method, we incorporate it with the DQN, DDPG and the SAC agents. For the DQN and DDPG agents, we use the EGr and EGa as the base methods for adaptive exploration and use Eqs. 1 and 2 to update their parameters. For the SAC agent, we use Eq. 3 to scale the policy deviation learnt by a neural network.

V. RESULTS

To investigate the effect of the proposed A^2 method, we conduct a series of simulation experiments using the Mini-Grid [14] and PMG environments [15]. All performances displayed were the success rates of achieving the final goal without demonstrations, averaged over five random seeds. Specifically, we experimented on six multi-step, sparse reward tasks, including GridDoorKey (3 sizes), ChestPush, ChestPickAndPlace and BlockStack (see Fig. 2 for a visualisation). Codes available at <https://github.com/IanYangChina/A-2-paper-code>.

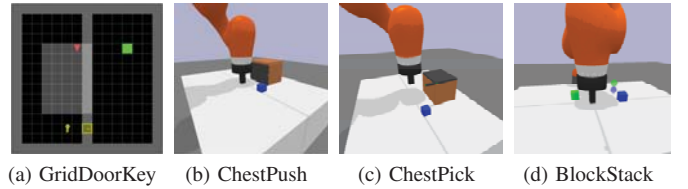


Fig. 2: Tasks. (a) The agent (red) needs to reach the goal (green). (b-c) The robot needs to push or pick-and-drop the block into the chest. (d) The robot needs to stack the blocks.

A. Task and implementation details

GridDoorKey task: The state representation consists of the x and y coordinates of the agent, the key and the door, the heading direction of the agent, and two binary variables indicating whether the agent is carrying the key and whether the door is opened. The goal representation is the x and y coordinates of the target. The reward function gives a value of 0 when the goal is reached and -1 otherwise.

Manipulation tasks: We use the ChestPush and ChestPick tasks with one block and the BlockStack task with two blocks from the Pybullet Multigoal (PMG) environments [15]. The state representation and reward function remain the same as the original paper. The original goal representation only consists of the target coordinates of the blocks in the world frame. We add the gripper location and its finger width to the goal representation for specifying the task steps regarding grasping and placing.

At the beginning of each task, the agent is given a desired goal which, for example, specifies the desired location of a block. If abstract demonstrations are used in an episode, the agent will instead be given a subgoal according to the step at hand. The decomposition schemes for the tasks are as follows:

- Gridworld: three subgoals, including 1) reach the key, 2) reach the door and 3) reach the target location;
- ChestPush: three subgoals, including 1) open the chest, 2) reach the block and 3) push the block into the chest;

- ChestPick: four subgoals, including 1) open the chest, 2) grasp the block, 3) move to the top of the chest and 4) drop the block into the chest;
- BlockStack: four subgoals, including 1) grasp the base block, 2) move the base block to the target location, 3) grasp the second block and 4) stack the second block.

Network architecture: The DQN network has three MLPs of sizes 64, 128 and 64. The actor and critic networks for the DDPG and SAC have three MLPs of sizes 256. All layers use ReLU activation, except for the output. All actors use hyperbolic tangent to activate the final layer and all critics (include DQN) have no activation on the output.

B. Ablations

This subsection examines the effect of different parameter values of the A^2 method, namely, the percentage of demonstrated episodes η and the adaptive exploration update rate τ_S . We perform ablations on the GridDoorKey25x25 tasks with the DQN agent, ChestPush task with the DDPG agent and BlockStack task with the SAC agent.

From Figs. 3a to 3c, we see that adding demonstrations improves convergence speeds and performances in general. Adding demonstrations to 50% or 75% of training episodes has the highest performance gains across all agents and tasks (green and red lines). Interestingly, providing demonstrations to all the training episodes tends to hurt the performances (purple lines). This could be due to a lack of negative learning signals when trained in a fully-demonstrated way, given that the value function normally requires negative examples to distinguish good transitions from bad ones for more accurate value prediction.

The ablations of the adaptive exploration update rate were conducted with a fixed percentage of demonstrated episodes ($\eta = 0.75$). From Figs. 3d to 3f, we see that, varying the value of τ_S tends to have no obvious effects. However, a small value may slow down learning as shown by the gridworld experiment (the orange line in Fig. 3d).

C. General performance

This subsection examines the general improvements gained by the proposed A^2 method. According to the ablation studies (subsection V-B), we add abstract demonstrations into 75% of the training episodes and use 0.3 for the adaptive exploration parameter, τ_S , for improvement evaluation.

As shown by Fig. 4, in all experiments, by adding abstract demonstrations in 75% of the training episodes, the agent learns faster with higher performances. The improvement is more obvious in the robotic tasks. Notice that as the robotic task becomes more difficult (from subfigures 4d to 4f), the gap of success rates becomes larger compared to the vanilla algorithm (blue lines). This demonstrates that abstract demonstrations can provide vast improvement on multi-step tasks, because it significantly eases the agent from the difficulty of exploration in long task horizon with sparse rewards and subtask dependencies.

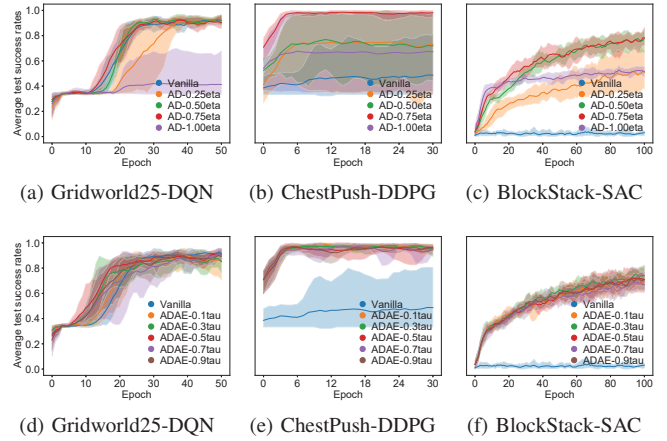


Fig. 3: Test success rates of achieving the final goal with different proportions of demonstrated episodes η (a-c) and adaptive exploration update rate τ_S (d-f). *AD*: abstract demonstrations; *ADAE*: abstract demonstrations and adaptive exploration.

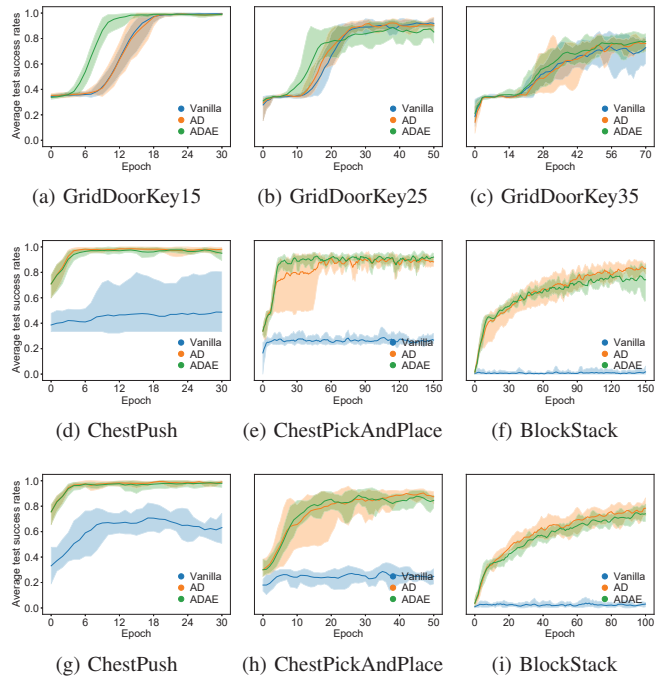


Fig. 4: Test success rates of achieving the final goal in all tasks: DQN on gridworld tasks (a-c), DDPG (d-f) and SAC (g-i) on robotic tasks. *AD*: abstract demonstrations; *ADAE*: abstract demonstrations and adaptive exploration.

Adaptive exploration provides less obvious improvements on top of abstract demonstrations in terms of success rates. However, it clearly stabilises the learning performance, as it shows a smaller variance. This is probably due to its effect on reducing unnecessary exploration such that the agent could act more decisively on well-mastered subtasks.

VI. CONCLUSION

We introduced A^2 – abstract demonstration and adaptive exploration – to aid reinforcement learning algorithms in multi-

step, long-horizon and sparse reward tasks. We showed in section IV that A^2 can be integrated in value-based algorithms with discrete actions (e.g., DQN) and actor-critic algorithms with continuous actions with a deterministic (e.g., DDPG) or stochastic (e.g., SAC) policy. We evaluated A^2 in discrete gridworld and continuous robot manipulation environments. Results in section V showed that abstract demonstrations in general speed up learning with higher success rates, with a more significant gain in continuous robot manipulation tasks, and the adaptive exploration module helps the agent to learn more stably.

A limitation of our method is the requirement of a manually designed task decomposition scheme. This could be addressed by learning to decompose a long-horizon task into subtasks. More sophisticated task assumptions such as image observations may also be considered to evaluate the effects of our method.

REFERENCES

- [1] A. Lazaridis, A. Fachantidis, and I. Vlahavas, “Deep reinforcement learning: A state-of-the-art walkthrough,” *Journal of Artificial Intelligence Research*, vol. 69, pp. 1421–1471, 2020.
- [2] N. T. Dantam, Z. K. Kingston, S. Chaudhuri, and L. E. Kavraki, “Incremental task and motion planning: A constraint-based approach,” in *RSS*, vol. 12, 2016, p. 00052.
- [3] S. J. Levine, “Monitoring the execution of temporal plans for robotic systems,” Ph.D. dissertation, MIT, 2012.
- [4] L. C. Cobo, C. L. Isbell Jr, and A. L. Thomaz, “Automatic task decomposition and state abstraction from demonstration,” Georgia Institute of Technology, 2012.
- [5] X. Yang, Z. Ji, J. Wu, Y.-K. Lai, C. Wei, G. Liu, and R. Setchi, “Hierarchical reinforcement learning with universal policies for multistep robotic manipulation,” *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [6] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, “Imitation learning: A survey of learning methods,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.
- [7] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, “Recent advances in robot learning from demonstration,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 297–330, 2020.
- [8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [9] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, “Overcoming exploration in reinforcement learning with demonstrations,” in *ICRA*, IEEE, 2018, pp. 6292–6299.
- [10] H.-n. Wang, N. Liu, Y.-y. Zhang, D.-w. Feng, F. Huang, D.-s. Li, and Y.-m. Zhang, “Deep reinforcement learning: A survey,” *Frontiers of Information Technology & Electronic Engineering*, pp. 1–19, 2020.
- [11] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [12] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
- [13] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, “Soft actor-critic algorithms and applications,” *arXiv preprint arXiv:1812.05905*, 2018.
- [14] M. Chevalier-Boisvert, L. Willems, and S. Pal, *Minimalistic gridworld environment for openai gym*, <https://github.com/maximecb/gym-minigrid>, 2018.
- [15] X. Yang, Z. Ji, J. Wu, and Y.-K. Lai, “An open-source multi-goal reinforcement learning environment for robotic manipulation with pybullet,” in *Towards Autonomous Robotic Systems*, Springer International Publishing, 2021, pp. 14–24, ISBN: 978-3-030-89177-0.
- [16] M. Fox and D. Long, “Pddl2. 1: An extension to pddl for expressing temporal planning domains,” *Journal of artificial intelligence research*, vol. 20, pp. 61–124, 2003.
- [17] R. E. Fikes and N. J. Nilsson, “Strips: A new approach to the application of theorem proving to problem solving,” *Artificial intelligence*, vol. 2, no. 3-4, pp. 189–208, 1971.
- [18] T. Jurgenson, O. Avner, E. Groshev, and A. Tamar, “Sub-goal trees a framework for goal-based reinforcement learning,” in *ICML*, 2020, pp. 5020–5030.
- [19] J. Andreas, D. Klein, and S. Levine, “Modular multitask reinforcement learning with policy sketches,” in *ICML*, 2017, pp. 166–175.
- [20] R. S. Sutton, D. Precup, and S. Singh, “Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning,” *Artificial intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999.
- [21] S. Krishnan, R. Fox, I. Stoica, and K. Goldberg, “Ddco: Discovery of deep continuous options for robot learning from demonstrations,” in *CoRL*, 2017, pp. 418–437.
- [22] K. Shiarlis, M. Wulfmeier, S. Salter, S. Whiteson, and I. Posner, “TACO: Learning task decomposition via temporal alignment for control,” in *ICML*, ser. PMLR, vol. 80, 2018, pp. 4654–4663.
- [23] R. Portelas, C. Colas, L. Weng, K. Hofmann, and P.-Y. Oudeyer, “Automatic curriculum learning for deep rl: A short survey,” in *IJCAI*, Jul. 2020, pp. 4819–4825.
- [24] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba, “Hindsight experience replay,” in *NIPS*, 2017, pp. 5048–5058.
- [25] D. Scieur and F. Pedregosa, “Universal asymptotic optimality of polyak momentum,” in *ICML*, 2020, pp. 8565–8572.