

# Creativity in Robot Manipulation with Deep Reinforcement Learning

Juan Carlos Vargas,<sup>1</sup> Malhar Bhoite,<sup>1</sup> Amir Barati Farimani<sup>1,2\*</sup>

<sup>1</sup>Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh PA 15213, USA

<sup>2</sup>Machine Learning Department, Carnegie Mellon University, Pittsburgh PA 15213, USA

\*To whom correspondence should be addressed; E-mail: barati@cmu.edu.

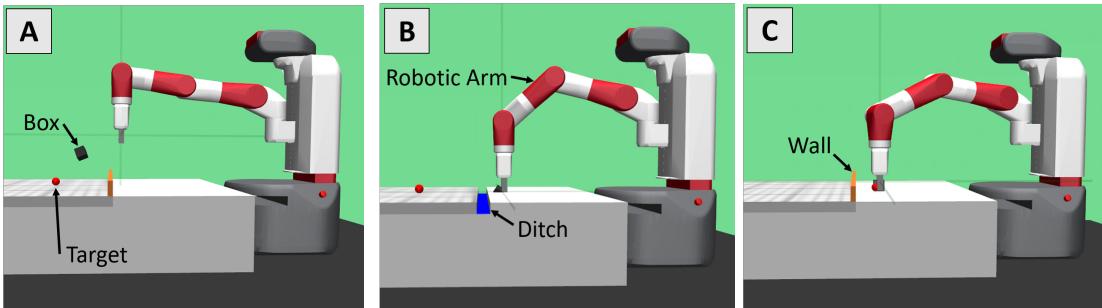
**Deep Reinforcement Learning (DRL) has emerged as a powerful control technique in robotic science. In contrast to control theory, DRL is more robust in the thorough exploration of the environment. This capability of DRL generates more human-like behaviour and intelligence when applied to the robots. To explore this capability, we designed challenging manipulation tasks to observe robots strategy to handle complex scenarios. We observed that robots not only perform tasks successfully, but also transpire a creative and non intuitive solution. We also observed robot's persistence in tasks that are close to success and its striking ability in discerning to continue or give up.**

## Summary

Creative robot manipulation can be achieved with deep reinforcement learning thorough exhaustive exploration of the environment.

# Introduction

With the emergence of novel artificial intelligence (AI) algorithms, and more specifically those in deep reinforcement learning (DRL), there will be a paradigm shift in robot perception and intelligence. In the future, AI enabled robots will be able to learn more quickly, generate complex behavioral strategies, and even achieve human level performance (1–4). For many years, robots have been using control theory successfully to perform tasks. However, control theory has certain limitations. For example, robots depend heavily on human understanding of the task, the robot itself, and the environment. While there are promising results showing adaptability improvements (5), robots still assume perfect knowledge of the system’s description and the environment. On the other hand, DRL techniques, combined with proper training, enable robots to comprehensively explore the environment and learn appropriate solutions (6). The process of exploring the environment is a systematic advantage that we seek to exploit in order to generate creative solutions. To this end, we define creativity as the ability to autonomously find novel solutions to overcome obstacles in order to achieve a given goal.

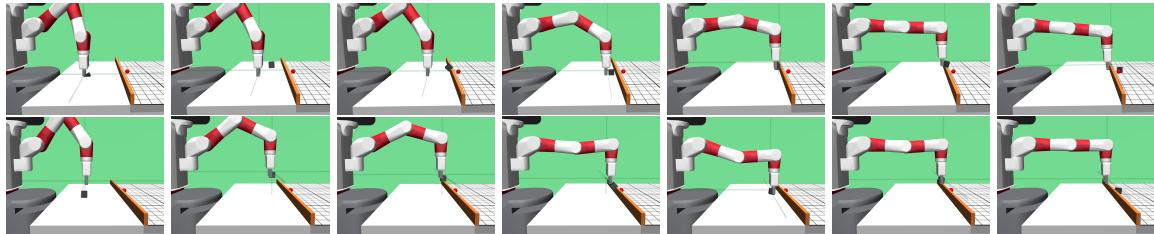


**Figure 1:** Representation of the Wall, Ditch, and Wall-TargetNear environments. On the left the robot can be seen throwing the box to the target location. In the middle, the robot is about to make the box bounce over the ditch to the target by punching. On the right, the robot simply moves the box to the desired location, within the length of its arm. These figures are created using the MuJoCo physics engine (7).

In this paper, we seek to show that a robot can find striking creative solutions during training.

To demonstrate such capabilities, we performed 3 types of experiments using DRL applied to an industrial manipulator in a simulated experiment (Fig. 1). All three experiments require the robot to move a box to a the target location within the same table. We named the experiments Ditch, Wall, and Wall–TargetNear (See Supporting Information (SI)). We have specifically designed these experiments with different spatial constraints and challenges for the robot to test its problem solving strategy. To maximize the challenge, the target location is placed beyond the physical reach of the robot at all times, except for Wall–TargetNear, in which the target location is on an expanded range that includes closer locations. We used negative sparse rewards as an incentive for quick exploration. Through these experiments, we observed that the robot makes surprising decisions and finds creative solutions.

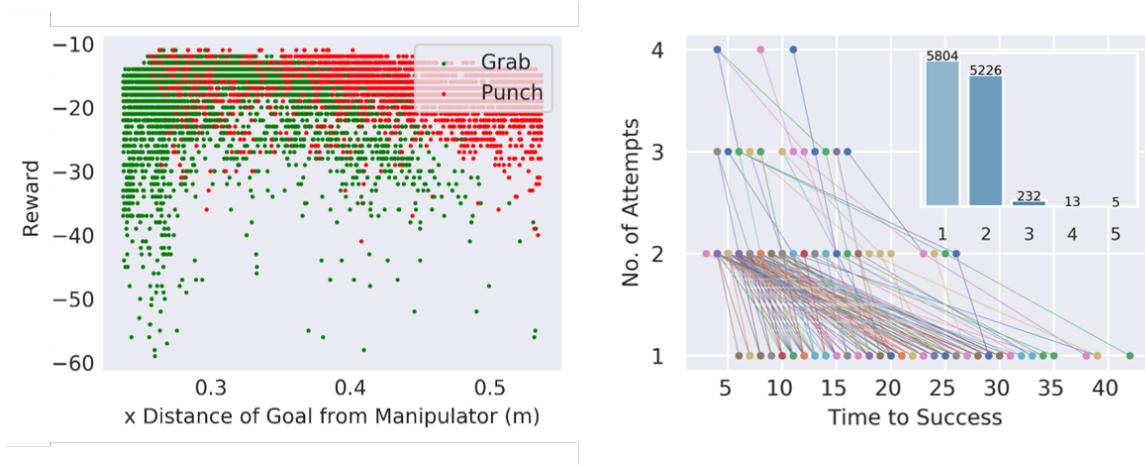
## Results



**Figure 2:** The first row (a) shows a combined action of bouncing the box off the ground and then throwing over the wall. The second row (b) shows two subsequent throws after the first one misses.

We observed that the robot became creative and found a solution that we did not expect. For all three experiments, due to the wall and ditch constraints, the robot needed to devise a solution to placing the box at the target location other than sliding or rolling it. Obviously, sliding or rolling would cause the box to either hit the wall or fall into the ditch. Our intuition was that the robot would learn to throw in order to get the box to the target location. However, the robot discovered that the box material is flexible and utilized that property to come up with a creative solution that is quicker to learn than throwing (throwing requires more dexterity with

the gripper). Surprisingly, the robot learns to punch the flexible box against the table and, upon release, use the stored elastic energy to catapult the box over the constraint and into the target area. (Figs. 1b and 2a and see supporting movies). The intriguing fact is that the physical properties of the box, such as the elasticity, weight, or friction between surfaces; are not part of the observation space. We believe discovering this hidden solution is valuable, creative, and well conceived (see supporting videos). For instance, the robot learned where and how much to press the box to store sufficient elastic energy to land it into the target.

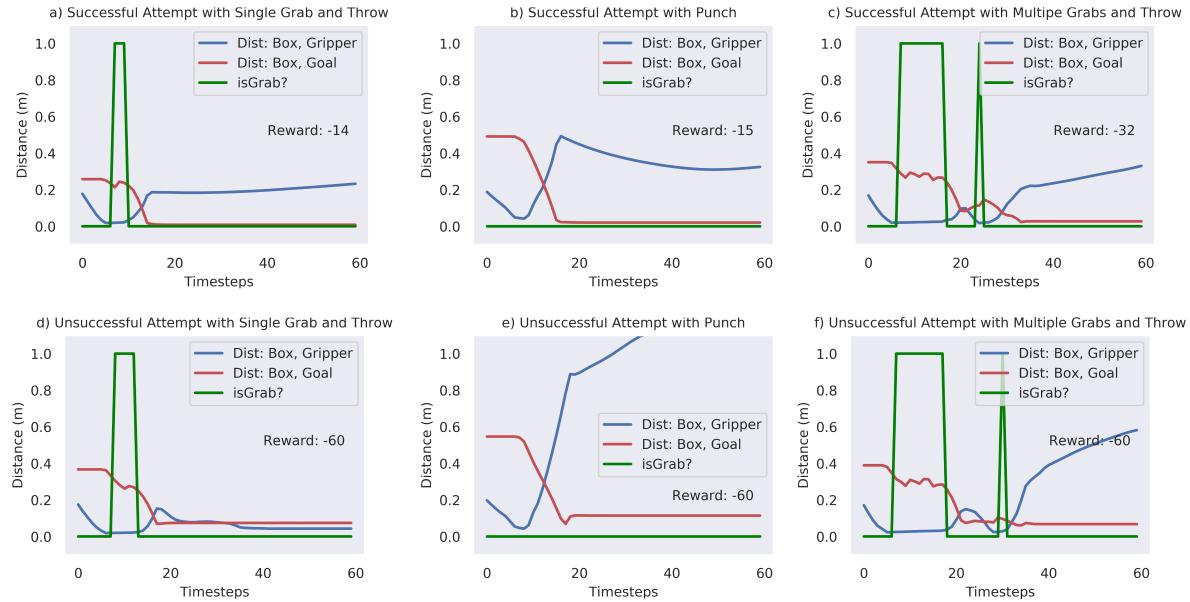


**Figure 3:** (a) Sample policy learned for the Wall environment, filtered for successful episodes, and shown as a function of the distance of target to goal (x-axis) and the rewards obtained (y-axis). It shows the distribution of *punch* vs. *grab* as a function of distance. (b) Diagram of the number of attempts per episode vs. time steps left in an episode (extracted from policy in (a)) shows majority of episodes are successful within the first two attempts, although more attempts are possible if there are sufficient time steps left.

As the robot gains more experience the actions become more complex. For instance, even as the robot already knows how to punch, it learns how to use its gripper to grab and eventually to throw the box. Evidence that it continues to explore the space even as a solution has already been found. Throwing often (but not always) generates more rewards than punching. This is very noticeable in the Wall experiment, in which the constraint is on the way of the trajectory for nearby punches (Fig. 3a). The further away the target is placed, the robot is more likely to

choose to punch. For other regions in which both actions are as likely to succeed, the robot is constantly and persistently looking for whichever action yields the highest reward, and in our case a hard preference based on distance was not clearly defined. Precision is also a factor that improved with training. The robot over time learns to accomplish the tasks using less individual actions (getting closer to a reward of  $-10$ ), as opposed to multiple attempts (closer to a reward of  $-60$ ) (Fig. 3a). Therefore, the amount of reward that is accumulated on each episode is not only a metric of success, but also a metric of how quickly the robot can complete the task. In the event that the first attempt was unsuccessful (Fig. 3b), and the box is still within reach of the gripper, the robot will pursue multiple attempts until the time runs out or it is successful.

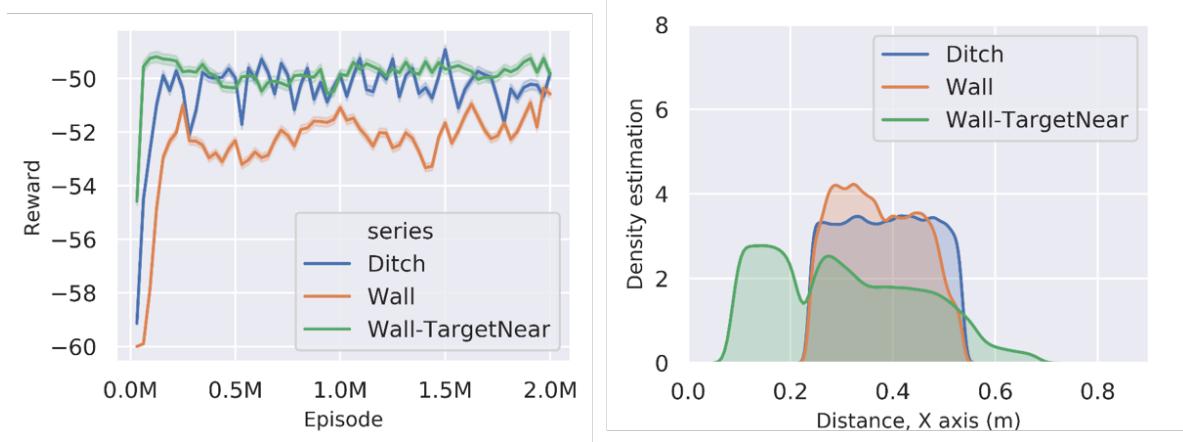
This behavior is a byproduct of using negative sparse rewards.



**Figure 4:** Analysis of single episodes for the Wall experiment when the robot (a) grabs and throws the box right on target, (b) punches the box on target, (c) executes two grabs and throws on target, (d) throws the box and misses, (e) punches the box and misses, and (f) grabs a couple of times before throwing but misses. The tolerance for being on target is  $0.05m$ .

For the robot, a creative solution is not sufficient. Mastery of each one of the actions is paramount to achieve higher rewards. Deep inside each episode we can see that the amount of

time each action takes has to be small in order to accumulate higher rewards. All combined, the reward is unlikely to be ever higher than  $-10$ . The worst reward,  $-60$ , comes from the maximum number of time steps that we allow per episode, and each episode lasts 2.76 seconds. Similarly, for the best case, the gripper takes at least about 6 time steps, or 0.276 seconds, to move from its starting position to the box, leaving an additional 4 or 5 steps, merely a fraction of a second, to throw or punch the box over the constraint (Fig. 4), and have it land (and stay) at the right location. Some high reward ideal scenarios are shown on Fig. 4a and Fig. 4b. Conversely, the scenarios on Fig. 4d and Fig. 4e are likely to be the worst, since a single (hasty) action that lands the box beyond the constraint cannot be corrected (the box sits beyond the reach of the robot). When the action is unlikely to be successful, the best outcome comes from performing multiple successive actions to re-position the box before sending it over the constraint. This improves the chances of success (and collect some reward), even if it is at the expense of time.



**Figure 5:** (a) The average episodic rewards for each one of the three experiments during training (for both successful and unsuccessful episodes). (b) Density estimate plots along the X axis show the effect of constraints in the learning and creative process. Given the uniform target sampling distribution, horizontal densities would be ideally expected. The robot on the Wall-TargetNear experiment has become more successful at clear proximity. The robot on the Wall experiment performed better close to the barrier. The robot on the Ditch experiment performed close to uniform throughout the range.

The robot displayed a similar level of creativity for the Ditch, Wall, and Wall-TargetNear

experiments, as measured by learning to grab and punch the box. Adding environmental constraints to the process of reaching the goal affected only marginally the time needed to learn how to punch or grab the box (Fig. 5a). However when faced with different environmental constraints, the robot modifies its policy to maximize its rewards (Fig. 5b). For instance, for the Wall–TargetNear experiment, the robot was consistently successful at the closer range, but less so at distances farther away beyond the wall. On the other hand, in the Ditch experiment, the robot was almost equally successful across the entire range (25cm to 55cm range along the long side of the table), which we interpret as the target having a uniform difficulty level. In all cases we utilized uniform sampling of the target space.

## Discussion

We have shown that deep reinforcement learning can be used as a technique to explore the environment and construct creative solutions for complex scenarios. We demonstrated that the robot is able to learn and use physical properties of environment, such as the elasticity, friction, and weight of the box for solving difficult manipulation tasks. These properties are not given explicitly as the states, however; robot was able to discover and use those to achieve its goal. Finding the punch strategy to throw the box to the target location is novel, creative and non intuitive to human. This makes us reflect about the importance of allowing robots to learn and be creative with as little guidance as possible. In addition, we observed robot persistence, another human like behaviour, in its confrontation with the failure. In many instances, the robot assess that it can be successful if it tries more. The time to success analysis in this paper shows how robot iteratively gets closer to success and is persistant.

In this study, we are limited by the quality of the simulated environment. For our experiments we also rely on an unchanging test framework in which, for instance, the target area remains the same to enable uniform transition sampling across the observation space and avoid

catastrophic forgetfulness. These issues fall outside of the scope of this set of experiments.

## Materials and Methods

We conduct the experiments on a modified version of OpenAI’s Fetch-Slide environments (8), powered by the MuJoCo physics engine (7). The setup involves a robotic manipulator trying to move a cubic object (box) to a particular location (target) (Fig. 1). The experiments involve adding different constraints to this environment, modifying the goal, and thereon exploring the robot’s behavior. We want to see if the robot is able to adapt to the additional constraints, and whether it is creative enough to learn new policies by comprehensively exploring its action space. For this purpose we created 3 variations of two slightly different environments. For the first environment we created a *ditch* to constrain the robot from rolling or pushing the box to the goal, or target location (the robot should learn to avoid dropping the box into the ditch). (Fig. 1) Second, we created a *wall* to constrain the robot and increase the difficulty of the task (the wall is far enough that it cannot be reached by the robot, forcing it to find a way to pass the box over the wall). The environment with the wall constraint was modified further by creating additional distance variations (see SI).

For the architecture of DRL, we use Deep Deterministic Policy Gradients (DDPG) (9) with Hindsight Experience Replay (HER) (10). We chose DDPG with HER because our experiments require both a continuous action space and a continuous state space. DDPG uses an Actor-Critic network (see SI), in which the actor’s policy is a Deterministic Policy Gradient network and the critic’s policy is a Q-Value network. The state space is a vector of 25 parameters which include the position of the gripper [3 parameters], position of the box [3], rotation of the box [3], velocity of the box [3], rotational velocity of the box [3], relative position between the gripper and the box [3], state of the gripper [2], gripper positional velocity [3], and gripper velocity [2]. The *goal* in the input space refers to the coordinates of the target [3]. The size of the action is

four- comprising the 3D coordinates of the gripper and the gripper state (distance to the center).

The actor network is comprised of 3 hidden layers of fully connected neurons (see SI). The input to the actor network is the combination of the state, the achieved goal, and the desired goal. The output is the action. The critic network is comprised of also 3 hidden layers of fully connected neurons. The input of the critic network is the state, the achieved goal, the desired goal, and the action (the output from the actor network). The output is the Q value evaluated for the current reward. The combination of DDPG and HER enables the robot to learn faster by learning from previous failure. We chose to use *sparse* rewards for several reasons. First, the outcome of the robot’s actions is either right or wrong (within a certain tolerance). Using *dense* rewards could be misguiding the robot into performing incomplete actions to get at least some reward (11). Second, we didn’t want to invest time engineering rewards for specific use cases. By engineering a reward, we would be leading the robot into learning *our* solution to the problem, hindering exploration (10). Instead of leading the robot, by adopting a sparse reward strategy, we encourage an explore-first discover-later strategy for each episode. Lastly, providing negative rewards works as an incentive to achieve the goal as quickly as possible. Each episode lasts 60 time steps, and for each step the robot receives a  $-1$  reward if the goal has not been achieved, or 0 otherwise. As a consequence of this reward structure, the minimum cumulative reward per episode is  $-60$ , and the maximum theoretical cumulative reward is 0.

We train on a single Intel Core i7-8700K CPU at 3.70GHz 12 cores server with a GeForce GTX 1080 Ti/PCIe/SSE2 GPU and 16 Gb of RAM, running Ubuntu 16.04. For all environments, we trained for  $100 \times 10^6$  time steps. For policy analysis, approximately 500,000 test episode samples were collected from each of the experiments. Additionally, 20,000 extra test episode samples were collected with step by step information from the Wall experiment for episodic analysis.

## References

1. V. Mnih, *et al.*, *Nature* **518**, 529 EP (2015).
2. D. Silver, *et al.*, *Science* **362**, 1140 (2018).
3. OpenAI, Openai five (2018).
4. O. Vinyals, *et al.*, Alphastar: Mastering the real-time strategy game starcraft ii (2019).
5. D. Zhang, B. Wei, *Annual Reviews in Control* **43** (2017).
6. J. Kober, J. A. Bagnell, J. Peters, *The International Journal of Robotics Research* **32**, 1238 (2013).
7. E. Todorov, T. Erez, Y. Tassa, *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2012), pp. 5026–5033.
8. M. Plappert, *et al.*, *CoRR* **abs/1802.09464** (2018).
9. T. P. Lillicrap, *et al.*, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio, Y. LeCun, eds. (2016).
10. M. Andrychowicz, *et al.*, *CoRR* **abs/1707.01495** (2017).
11. X. Guo, Deep learning and reward design for reinforcement learning (2017).
12. V. Mnih, *et al.*, *CoRR* **abs/1312.5602** (2013).
13. L.-J. Lin, Reinforcement learning for robots using neural networks. technical report, dtic document (1993).

14. T. Degris, M. White, R. S. Sutton, *arXiv preprint arXiv:1205.4839* (2012).
15. D. Pathak, P. Agrawal, A. A. Efros, T. Darrell, *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* pp. 488–489 (2017).
16. A. Y. Ng, S. J. Russell, *ICML* (2000).
17. B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, *AAAI* (2008).
18. P. Abbeel, A. Y. Ng, *ICML* (2004).
19. Y. LeCun, Y. Bengio, G. Hinton, *nature* **521**, 436 (2015).
20. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
21. R. S. Sutton, A. G. Barto, *Reinforcement learning an introduction. 2nd ed.* (The MIT Press, Cambridge, Mass, 2018), second edn.
22. R. Liu, J. Zou, *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (2018), pp. 478–485.
23. D. Silver, *et al.*, *ICML* (2014).
24. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, *CoRR abs/1707.06347* (2017).
25. Z. Wang, *et al.*, *CoRR abs/1611.01224* (2016).
26. H. Liu, A. Trott, R. Socher, C. Xiong, *CoRR abs/1902.00528* (2019).
27. J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, *arXiv preprint arXiv:1506.02438* (2015).
28. T. Schaul, J. Quan, I. Antonoglou, D. Silver, *arXiv preprint arXiv:1511.05952* (2015).

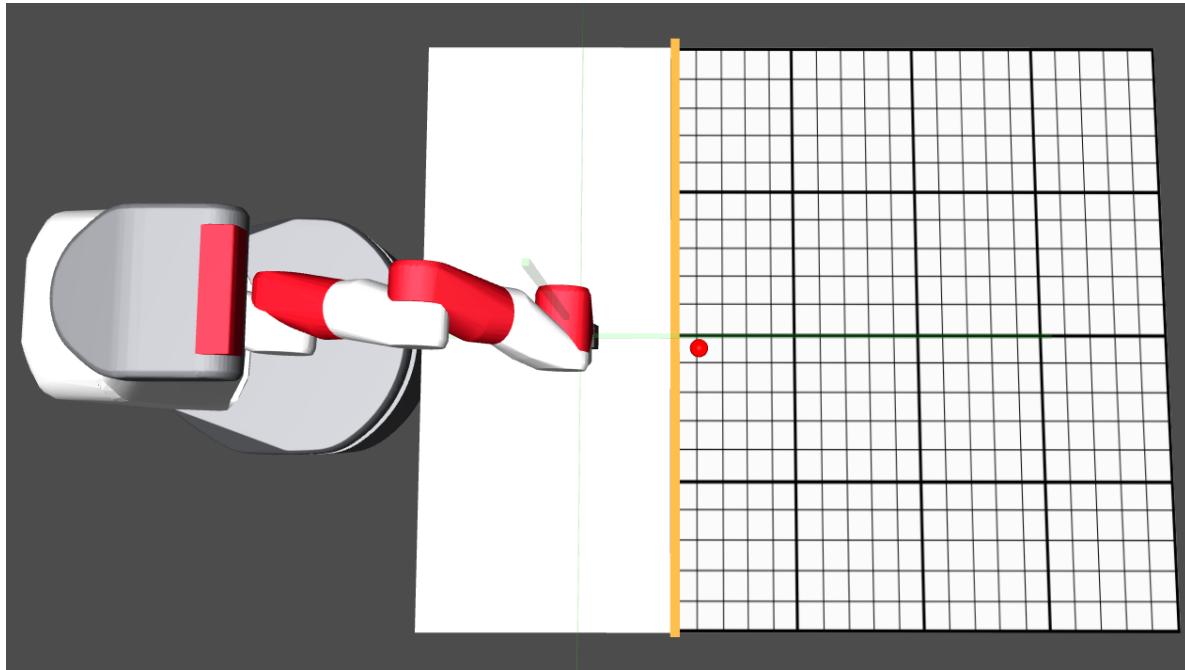
29. P. WawrzyńSki, A. K. Tanwani, *Neural Networks* **41**, 156 (2013).
30. S. Ioffe, C. Szegedy, *arXiv preprint arXiv:1502.03167* (2015).
31. H. V. Hasselt, *Advances in Neural Information Processing Systems* (2010), pp. 2613–2621.
32. P. Wawrzyński, *Neural Networks* **22**, 1484 (2009).
33. A. Gudimella, *et al.*, *CoRR* **abs/1709.06977** (2017).
34. P. Dhariwal, *et al.*, Openai baselines (2017).
35. OpenAI, *et al.*, *CoRR* **abs/1808.00177** (2018).
36. C. Smith, *et al.*, *Robotics and Autonomous systems* **60**, 1340 (2012).
37. S. Ekvall, D. Kragic, *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004* (IEEE, 2004), vol. 4, pp. 3519–3524.

# Supporting Information

This section contains ancillary information beyond what is included in the main paper.

## Environment

The robot sits approximately at the middle of the table on the 'y' direction (Fig. 6) and cannot reach farther than the distance to the wall, shown in yellow (or the ditch, not shown). The range for the Wall and Ditch experiments is contained within the first two gridded regions from the left, and the middle two regions in the vertical direction. The entire target area is approximately a 4 region area. The area for the Wall-TargetNear experiment, on the other hand, is centered at the same point but the coverage has been expanded for about an entire region all around, including the area without a grid.



**Figure 6:** Top view of environment used for Wall and Wall-TargetNear experiments. The initial position of the robot's gripper is the coordinate origin, while the direction left and right is along the 'x' axis, and the direction up and down is along the 'y' axis. Each one of the grid squares has an area of approximately  $2.75 \times 3.25\text{cm}^2$

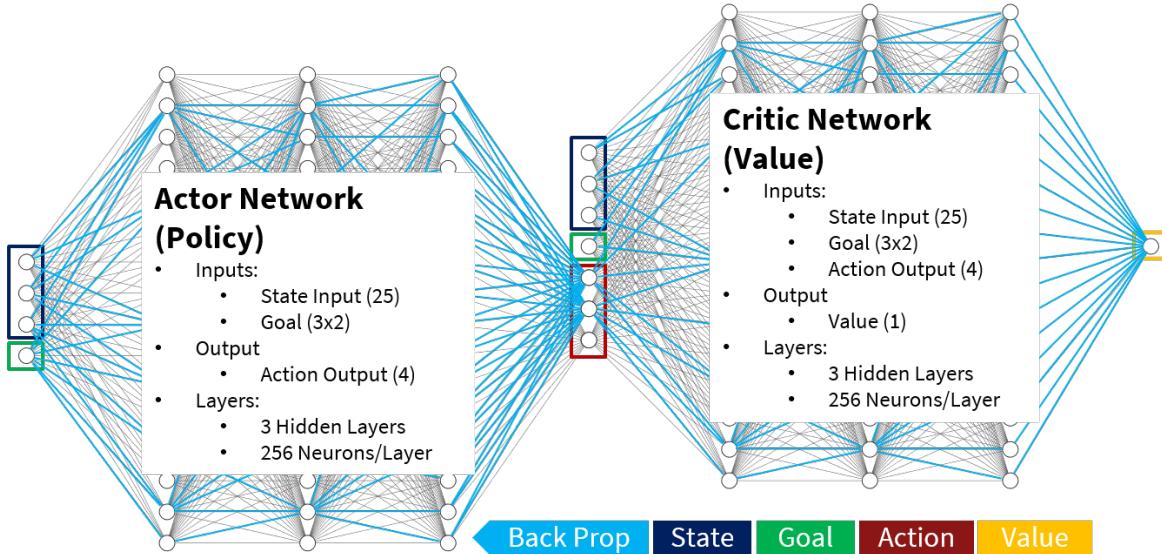
## Configuration

We use a standard DDPG+HER setup. Some of the most important hyperparameters are described below:

- Buffer size: number of transitions  $(s_t, a_t, g_t, r_t, s_{t+1})$  that are stored in the replay buffer.
- Neurons per hidden layer: a definition of how many neurons will be available per hidden layer in the actor-critic network.
- Number of layers: the number of hidden layers in the actor-critic network
- Network class: a definition of a neural network passed down as a class. For our experiments we used a single multi-layer perceptron (mlp) architecture, although other architectures, such as a CNN or LSTM, are also possible.
- Polyak coefficient: for stability, the network may only be allowed to change at a certain rate. A simple way to do this is by making a copy of the main network into a target network and averaging the parameters. The coefficient for Polyak-averaging indicates the weight for such an average.
- Batch size: the batch size from the experience buffers that is used for training
- Gamma: the discount gamma parameter value used for  $Q$  learning updates
- $Q$ -network learning rate: the learning rate for the  $Q$  value update function on the critic network
- $\pi$ -network learning rate: learning rate for the actor network
- Normalization  $\epsilon$ : epsilon used for observation normalization meant to avoid numerical instabilities

- Normalization clipping: magnitude of normalized output is clipped to this value
- Maximum action value: maximum magnitude for the each of the action coordinates
- Action L2 penalty: quadratic penalty on the actions before rescaling by the maximum action value
- Observation clipping: observation clipping before normalization
- Size of rollout batch: refers to the number of parallel rollouts per DDPG agent/thread

We modified both the environment and the agent to record the instances when the robot grabs the box, the moment when it moves the box without grabbing it, and the (x,y) coordinates of the target location for each episode. To perform intra-episode analysis, we also made changes to collect and record separately the state space parameters for each step inside an episode.



**Figure 7:** Diagram of the neural network

Experiment	Description
Ditch	Target goal on the $40 \pm 15\text{cm}$ range from the x axis, beyond the reach of the hand manipulator. The region is separated by a short ditch 2 cm deep and 2.5 cm wide, located at 26.25 cm from the origin.
Wall	Target goal on the $40 \pm 15\text{cm}$ range from the x axis, beyond the reach of the hand manipulator. The region is separated by a short wall 3 cm high and 1 cm wide, located 28 cm from the origin.
TargetNear	Target goal on the $40 \pm 30\text{cm}$ range from the x axis, often but not always within the reach of the hand manipulator. The region is separated by a short wall 3 cm high and 1 cm wide, located 28 cm from the origin.

**Table 1:** Description of the main experiments

## Main Experiments

All experiments use the same *Fetch* robot that was used by OpenAI (8), but the table has been modified to be bigger ( $82.5\text{cm}$  long and  $65.0\text{cm}$  wide) and include either the ditch or the wall constraint. The box has a mass of  $2\text{kg}$ . Table 1 contains additional details about each experiment setup.

## Ancillary Experiments

Experiment	Description
TargetMoving	Same as the Wall experiment, however the target region is moved backwards slowly at a rate of $2.0 \times 10^{-6}\text{cm}$ per step.
TargetExpanding	Same as the Wall experiment, however the target region expands backwards and sideways slowly at a rate of $6.67 \times 10^{-7}\text{cm}$ per step.
RStateSp	Same as the Wall experiment, however the box's orientation and rotational velocity have been removed from the state space.

**Table 2:** Description of the ancillary experiments

Three smaller experiments were created for ancillary purposes (Table 2). The Wall-TargetMoving and Wall-TargetExpanding environments were created to assess the impact of non-uniform distributions for target sampling. Both experiments use the pre-trained policy from the Wall experiment as a baseline. We correctly expected Wall-TargetMoving to perform very poorly as the experiment allows for a time dependence on the sampling, which is one of the more important issues that the experience buffer tries to solve. Moving the target area also breaks the uniform sampling. In this case, new information recently learned came to overwrite the old and performance deteriorates catastrophically. The Wall-TargetExpanding experiment performed better, as the uniform sampling was maintained despite the time dependence of the sampling. The Wall-RStateSp experiment repeated the Wall experiment from the beginning, removing the box’s orientation and rotational velocity from the state space. Originally we anticipated little change, as the gripper cannot really rotate to pick up the box in a different orientation, and the starting orientation of the box is always aligned to the gripper’s. However, the results show that learning did slow down when compared to the regular Wall experiment.