Sim-to-Real transfer of Reinforcement Learning policies in robotics

TA: Gabriele Tiboni (gabriele.tiboni@polito.it)
Link to this file: shorturl.at/a1246

Advanced Machine Learning 2022
Politecnico di Torino

OVERVIEW

The main goal of this project is to get familiar with the reinforcement learning (RL) paradigm and apply it in the context of robot learning. To this end, the student will be introduced to the problem of learning a control policy for a robot in simulation using state-of-the-art reinforcement learning algorithms and methods. In particular, the student will learn about the *sim-to-real transfer* problem in robot learning literature, namely the task of learning policies in simulation through RL that can be directly transferred to real-world hardware, avoiding costly interactions with real setups and speeding up the training time. During this project, the student will be simulating the sim-to-real transfer task in a sim-to-sim scenario, where a discrepancy between source (training) and target (test) domains is manually injected. The student will implement domain randomization of dynamics parameters (e.g. masses, friction coefficients), a popular recent strategy to learn robust policies that transfer well to the target domain. Finally, the student will also attempt to train control policies directly from visual input (i.e. images) through convolutional neural networks, allowing the robot to autonomously interact with the environment from raw images as observations.

GOALS

1. Read the provided material to get familiar with the Reinforcement Learning framework, the sim-to-real transfer challenge and the common techniques to perform an efficient transfer from simulation to reality;
2. Study the code and implement an RL training pipeline via third-party APIs to state-of-the-art reinforcement learning algorithms such as PPO, TRPO, and SAC.
3. Implement Uniform Domain Randomization (UDR) to learn robust policies in the source domain and limit the loss of performance during the sim-to-real transfer;

4. Train vision-based control policies from raw input images

1st STEP) STUDY BACKGROUND & RELATED WORKS

Getting familiar with the reinforcement learning framework in robotics

Before starting you're asked to take some time to familiarize yourself with the framework of Reinforcement Learning, the sim-to-real transfer challenge and SOTA strategies to overcome it. More in detail, read:

- Read sections 1.-1.4, 1.6, 3.-3.8, of [1] to understand the general Reinforcement Learning framework;
- Watch introductory video on Reinforcement learning by DeepMind **video**
- Read **article** on introduction to Reinforcement learning by OpenAI [part 1, part 2, part 3]
- Read sections 1.-1.3, 3.-3.4, of [2]
- Read sections 1., 2., 3. of [3]
- Read debate on the sim-to-real transfer paradigm [4]
- Read [5], [6], **blog post** to understand domain randomization for sim-to-real transfer
- Read this **set of slides** by Josh Tobin and this **article** regarding domain randomization for both vision and dynamics properties

2nd STEP) IMPLEMENTING AND TESTING DIFFERENT ALGORITHMS FOR REINFORCEMENT LEARNING

Defining the baseline/upper bound for the next phase

Train an RL agent on the gym Hopper environment. This environment comes with an easy-to-use python interface which controls the underlying physics engine — MuJoCo — to model the robot.
The hopper is a one-legged robot model whose task is to learn how to jump without falling, while achieving the highest possible horizontal speed.
The student will simulate a transfer scenario on this robot, given two custom variants which have been created ad-hoc: policy training takes place in the *source environment* and the student will transfer and test the policy on the *target environment* — which technically represents the real world. To simulate the reality gap, the source domain Hopper has been generated by shifting the torso mass by 1kg with respect to the target domain.

1. Check out the provided code template and start playing around with the underlying Hopper environment. Get familiar with the *test_random_policy.py* script, the python interface for MuJoCo, the gym documentation, and the hopper environment overall. Finally answer the questions in the table below.

State space	**Action space**	**Mass values**
What is the state space in the Hopper environment? Is it discrete or continuous? The state space is the set of possible states the agent can assume. It is continuous because a state/observation is described by a vector of 11 float elements in the range of [-inf, inf], but they're normalized such that they're included in the range [-1, 1].	What is the action space in the Hopper environment? Is it discrete or continuous? The action space is the set of possible actions the agent can perform. It is continuous because an action is described by a vector of 3 float elements in the range of [-1, 1].	What is the mass value of each link of the Hopper environment, in the source and target variants respectively? Source domain: [torso: 2.53429174 thigh: 3.92699082 leg: 2.71433605 foot: 5.0893801] Target domain: [torso: 3.53429174 thigh: 3.92699082 leg: 2.71433605 foot: 5.0893801]

If you need any help answering the above questions try looking at the Mujoco documentation or the gym documentation.
A few hints:

- *Bodies defined in the environment: env.sim.model.body_names*
- *Mass of all the corresponding bodies: env.sim.model.body_mass*
- *Number of degrees of freedom (DoFs) of the robot: env.sim.model.nv*
- *Number of DoFs for each body: env.sim.model.body_dofnum*
- *Number of actuators: env.sim.model.nu*
- *See other attributes here*

2. Implement a reinforcement learning pipeline to train a control policy for the Hopper environment. To this end, you'll make use of a third-party library to train an agent with state-of-the-art reinforcement learning algorithms such as TRPO, PPO, and SAC. In particular, follow the steps below, and make sure to go through the provided external resources:

a. Create a script using the third-party library stable-baselines3 (sb3) and train the Hopper agent with **one** algorithm of choice between TRPO [8], PPO [9] and SAC [7].
   i. openAI article on TRPO
   ii. openAI article on PPO
   iii. openAI article on SAC
   iv. Explanation video on TRPO and PPO, explanation video on SAC.

   Use the provided template in *train.py* as a starting point. It is okay to look at publicly available code for reference, but it's likely easier and more helpful to study the sb3 documentation and understand how to implement the code by yourself.

3. Train two agents with your algorithm of choice, on the source and target domain Hoppers respectively. Then, test each model and report its average return over 50 test episodes. In particular, report results for the following "training→test" configurations: source→source, source→target (**lower bound**), target→target (**upper bound**).

   Test with different hyperparameters and report the best results found together with the parameters used:

Table 1) RL algorithm	**Hyperparameters**	**Source-Source average return**	**Source-Target average return**	**Target-Target average return**
TRPO	*TS=200K*	1519.37 +/- 11.66	1440.56 +/- 185.18	1465.65 +/- 7.18
TRPO	*TS=300K*	1667.76 +/- 7.55	977.76 +/- 16.55	1545.72 +/- 139.68
TRPO	*TS=500K*	1660.24 +/- 2.87	1002.65 +/- 5.69	**1652.47 +/- 5.33**
TRPO	*TS=200K, LR=3E-4*	1276.75 +/- 153.49	889.60 +/- 165.36	873.66 +/- 92.97

TRPO	*TS=300K, LR=3E-4*	1160.93 +/- 243.49	**1463.48 +/- 38.29**	769.00 +/- 210.86
TRPO	*TS=500K, LR=3E-4*	1626.47 +/- 9.49	1095.82 +/- 105.24	1541.97 +/- 5.37
PPO	*TS=150K*	1421.12 +/- 31.52	844.19 +/- 261.51	1071.64 +/- 26.01
PPO	*TS=175K*	1387.40 +/- 5.54	1316.47 +/- 10.34	715.40 +/- 37.81

The results above will serve as the upper bound and lower bound for the Domain adaptation phase using domain randomization.

3rd STEP) IMPLEMENTING DOMAIN RANDOMIZATION

Implement Uniform Domain Randomization to narrow the gap between source and target domains and make your agent more robust to different environment dynamics.

Implement Uniform Domain Randomization (UDR) on the mass of each link of the Hopper.

Uniform domain randomization (in the proposed setting) refers to manually designing a uniform distribution over the three remaining masses in the source environment (considering that the torso mass is fixed at -1 kg w.r.t. the true one) and doing training with values that vary at each episode (sampled appropriately from the chosen distributions).
The underlying idea is to force the agent to maximize its reward and solve the task for a range of multiple environments at the same time, such that its learned behavior may be robust to slight dynamics variations.
Note that, since the choice of the distribution is a hyperparameter of the method, the student has to manually try different distributions in order to expect good results on the target environment.

1. Train a UDR agent on the source environment with the same RL algorithm previously used. Later test the policy obtained on both the

source and target environments. Finally, report their average reward in the Table below.

Is UDR able to overcome the unmodelled effect (shift of torso mass) and lead to more robust policies?

Suggestions:

- env.sim.model.body_mass[i] controls the mass of the i-th body in the Hopper environment. In particular, the torso mass value is env.sim.model.body_mass[1], the thigh mass value is env.sim.model.body_mass[2], and so on.

- To check out all body names: env.sim.model.body_names

- Remember not to randomize the torso mass!

Table 2) **RL Algorithm**	**Hyperparameters**	**UDR distribution**	**Source-Source average return**	**Source-Target average return**
TRPO	*TS=200K, LR=1E-3*	*Uniform (10% mass variance)* *Range for mass 1:* *[3.534, 4.319)* *Range for mass 2:* *[2.442, 2.985)* *Range for mass 3:* *[4.580, 5.598)*	1283.27 +/- 122.25	719.36 +/- 13.86
TRPO	*TS=200K, LR=1E-3*	*Uniform (20% mass variance)* *Range for mass 1:* *[3.142, 4.712)* *Range for mass 2:*	1486.67 +/- 7.05	1116.52 +/- 20.88

		[2.171, 3.257) *Range for mass 3: [4.072, 6.107)*		
TRPO	*TS=200K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	1472.13 +/- 113.50	906.47 +/- 16.78
TRPO	*TS=200K, LR=1E-3*	*Uniform (40% mass variance)* *Range for mass 1: [2.356, 5.498)* *Range for mass 2: [1.629, 3.800)* *Range for mass 3: [3.054, 7.125)*	1606.69 +/- 72.88	901.86 +/- 20.69
TRPO	*TS=500K, LR=1E-3*	*Uniform (10% mass variance)* *Range for mass 1: [3.534, 4.320)*	1589.22 +/- 2.73	824.54 +/- 6.95

		Range for mass 2: [2.443, 2.986) *Range for mass 3: [4.580, 5.598)*		
TRPO	*TS=500K, LR=1E-3*	*Uniform (20% mass variance)* *Range for mass 1: [3.142, 4.712)* *Range for mass 2: [2.171, 3.257)* *Range for mass 3: [4.072, 6.107)*	1582.34 +/- 1.78	950.52 +/- 12.58
TRPO	*TS=500K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	1683.84 +/- 10.2	**1641.62 +/- 181.69**
TRPO	*TS=500K, LR=1E-3*	*Uniform (40% mass variance)*	1647.88 +/- 2.54	944.78 +/- 11.68

		Range for mass 1: [2.356, 5.498) *Range for mass 2: [1.629, 3.800)* *Range for mass 3: [3.054, 7.125)*		
TRPO	*TS=500K, LR=1E-3*	*Uniform (60% mass variance)* *Range for mass 1: [1.571, 6.283)* *Range for mass 2: [1.086, 4.343)* *Range for mass 3: [2.036, 8.143)*	1571.66 +/- 4.91	877.05 +/- 13.82
TRPO	*TS=1M, LR=1E-3*	*Uniform (10% mass variance)* *Range for mass 1: [3.534, 4.320)* *Range for mass 2: [2.443, 2.986)* *Range for mass 3: [4.580, 5.598)*	1693.16 +/- 9.66	909.72 +/- 16.00

TRPO	*TS=1M, LR=1E-3*	*Uniform (20% mass variance)* *Range for mass 1: [3.142, 4.712)* *Range for mass 2: [2.171, 3.257)* *Range for mass 3: [4.072, 6.107)*	1681.77 +/- 6.90	1193.56 +/- 122.39
TRPO	*TS=1M, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	**1693.54 +/- 3.55**	1259.21 +/- 81.36
TRPO	*TS=1M, LR=1E-3*	*Uniform (60% mass variance)* *Range for mass 1: [1.571, 6.283)* *Range for mass 2:*	1676.72 +/- 32.21	1018.84 +/- 22.32

		[1.086, 4.343) *Range for mass 3: [2.036, 8.143)*		
TRPO	*TS=300K, LR=3E-4*	*Uniform (10% mass variance)* *Range for mass 1:* *[3.534, 4.319) Range for mass 2: [2.442, 2.985) Range for mass 3: [4.580, 5.598)*	1097.42 +/- 4.43	959.15 +/- 3.90
TRPO	*TS=300K, LR=3E-4*	*Uniform (20% mass variance)* *Range for mass 1: [3.142, 4.712)* *Range for mass 2: [2.171, 3.257)* *Range for mass 3: [4.072, 6.107)*	1176.45 +/- 100.99	753.40 +/- 7.63
TRPO	*TS=300K, LR=3E-4*	*Uniform (30% mass variance)* *Range for mass 1:*	971.38 +/- 35.90	1019.81 +/- 62.84

		[2.749, 5.105) *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*		
PPO	*TS=175K, LR=3E-4*	*Uniform (10% mass variance)* *Range for mass 1:* *[3.534, 4.319) Range for mass 2: [2.442, 2.985) Range for mass 3: [4.580, 5.598)*	1130.73 +/- 29.70	1055.48 +/- 20.11
PPO	*TS=175K, LR=3E-4*	*Uniform (20% mass variance)* *Range for mass 1: [3.142, 4.712)* *Range for mass 2: [2.171, 3.257)* *Range for mass 3: [4.072, 6.107)*	1412.45 +/- 229.11	801.16 +/- 151.02
PPO	*TS=175K, LR=3E-4*	*Uniform (30% mass*	1129.56 +/- 51.17	1267.33 +/- 48.50

		variance) *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*		
PPO	*TS=500K, LR=3E-4*	*Uniform (10% mass variance)* *Range for mass 1: [3.534, 4.320)* *Range for mass 2: [2.443, 2.986)* *Range for mass 3: [4.580, 5.598)*	1658.27 +/- 44.44	1058.04 +/- 79.65
PPO	*TS=500K, LR=3E-4*	*Uniform (20% mass variance)* *Range for mass 1: [3.142, 4.712)* *Range for mass 2: [2.171, 3.257)* *Range for mass 3:*	1652.04 +/- 44.26	907.13 +/- 77.20

		[4.072, 6.107)		
PPO	*TS=500K, LR=3E-4*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	1734.12 +/- 7.45	1299.25 +/- 79.71
PPO	*TS=500K, LR=3E-4*	*Uniform (40% mass variance)* *Range for mass 1: [2.356, 5.498)* *Range for mass 2: [1.629, 3.800)* *Range for mass 3: [3.054, 7.125)*	1533.96 +/- 59.65	1041.09 +/- 170.18
PPO	*TS=500K, LR=3E-4*	*Uniform (60% mass variance)* *Range for mass 1: [1.571, 6.283)* *Range for mass 2:*	1573.85 +/- 151.28	1236.25 +/- 204.97

		[1.086, 4.343) *Range for mass 3: [2.036, 8.143)*		
PPO	*TS=1M, LR=3E-4*	*Uniform (10% mass variance)* *Range for mass 1: [3.534, 4.320)* *Range for mass 2: [2.443, 2.986)* *Range for mass 3: [4.580, 5.598)*	1063.45 +/- 35.99	752.20 +/- 19.77

4th STEP) VISION-BASED REINFORCEMENT LEARNING

Implement an RL training pipeline that uses raw images as state observations to the agent. In this final step, you're given more flexibility on how you can implement the pipeline, leaving room for variants and group's own implementations.

1. Implement your own new training script so as to change the observations of the environment to only use raw images (i.e. 2D renders) of the environment rather than low-dimensional configuration vectors, and to use a convolutional neural network of your choice as the underlying policy structure.

   Train an agent on raw images with your RL algorithm of choice and test it on source-source, source-target, target-target configurations.

Suggestions:

- You can retrieve the current Mujoco render with: img_state = env.render(mode="rgb_array", width=224, height=224)

- Note: if you get a "RuntimeError: Failed to initialize OpenGL" error, have a look at this issue

- Stacking renders from previous timesteps often helps to keep information on the velocity of moving objects! See [14] for more details.

Table 3) **RL Algorithm**	**Hyperparameters**	**UDR distribution**	**Source-Source average return**	**Source-Target average return**
TRPO	*TS=10K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	58.78 +/- 4.26	67.27 +/- 2.09
TRPO	*TS=20K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	51.98 +/- 2.03	51.80 +/- 1.00
TRPO	*TS=30K, LR=1E-3*	*Uniform (30% mass variance)*	61.51 +/- 7.88	79.13 +/- 1.51

		Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)		
TRPO	*TS=40K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	53.25 +/- 2.87	48.72 +/- 0.75
TRPO	*TS=50K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	51.62 +/- 1.22	53.08 +/- 1.01
TRPO	*TS=60K, LR=1E-3*	*Uniform (30% mass variance)*	48.90 +/- 2.78	47.20 +/- 0.78

		Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)		
TRPO	*TS=70K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	83.19 +/- 28.56	79.49 +/- 28.35
TRPO	*TS=80K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	115.24 +/- 31.57	99.89 +/- 3.72
TRPO	*TS=125K, LR=1E-3*	*Uniform (30% mass variance)*	71.48 +/- 38.44	122.79 +/- 23.77

		Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)		
TRPO	*TS=200K, LR=1E-3*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	92.62 +/- 1.07	88.05 +/- 1.12
TRPO	*TS=25K, LR=1E-4*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	66.30 +/- 2.09	61.55 +/- 1.25
TRPO	*TS=50K, LR=1E-4*	*Uniform (30% mass variance)*	50.92 +/- 1.01	56.97 +/- 19.68

		Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)		
TRPO	*TS=75K, LR=1E-4*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	50.08 +/- 1.16	54.32 +/- 1.36
TRPO	*TS=100K, LR=1E-4*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	122.81 +/- 1.27	**138.08 +/- 4.50**
TRPO	*TS=125K, LR=1E-4*	*Uniform (30% mass variance)*	118.76 +/- 1.30	121.72 +/- 1.37

		Range for mass 1: [2.749, 5.105) *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*		
TRPO	*TS=150K, LR=1E-4*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	**126.19 +/- 9.20**	127.64 +/- 3.47
TRPO	*TS=25K,* *LR= 1-E2*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	54.87 +/- 1.00	55.52 +/- 13.96
TRPO	*TS=50K,* *LR= 1-E2*	*Uniform (30% mass variance)*	50.97 +/- 0.97	47.32 +/- 26.00

		Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)		
TRPO	*TS=75K,* *LR= 1-E2*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	115.04 +/- 3.27	106.75 +/- 26.00
TRPO	*TS=100K,* *LR= 1-E2*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105) Range for mass 2: [1.900, 3.529) Range for mass 3: [3.563, 6.616)*	49.42 +/- 0.97	111.29 +/- 4.55
TRPO	*TS=125K,* *LR= 1-E2*	*Uniform (30% mass variance)*	99.50 +/- 8.12	85.01 +/- 5.53

		Range for mass 1: [2.749, 5.105) *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*		
TRPO	*TS=150K,* *LR= 1-E2*	*Uniform (30% mass variance)* *Range for mass 1: [2.749, 5.105)* *Range for mass 2: [1.900, 3.529)* *Range for mass 3: [3.563, 6.616)*	46.87 +/- 0.98	47.21 +/- 1.15

2. **(optional)** Variants: at this stage, you may feel free to implement any idea to attempt at further improving the sim-to-real transfer in our simple scenario, for the vision-based RL task.
   For example, variants may try to investigate domain randomization on the appearance of the images to improve the transfer from source to target domain. Other possible directions may include investigating better representation learning techniques (such as transfer learning from pre-trained convolutional neural networks) or domain augmentation.
   Rather than requiring you to obtain actual improvements, this step is for you to go beyond the guidelines and get a feeling of a research-like approach.

AT THE END

- Deliver PyTorch scripts for all the required steps.
- Deliver this file with the tables compiled.
- Write a complete PDF report (with paper-style). The report should contain a brief introduction, a related work section, a methodological section for describing the algorithm that you're going to use, an experimental section with all the results and discussions, and a final brief conclusion. Follow this **link** to open and create the template for the report.

EXAMPLE OF QUESTIONS YOU SHOULD BE ABLE TO ANSWER AT THE END OF THE PROJECT

- What is Reinforcement Learning?
- Why is Reinforcement Learning appealing for the robotics field?
- Why are physics simulators popularly used to learn Reinforcement Learning policies for real robots?
- What is the task given in the Hopper environment?
- What is the reward function in the Hopper environment?
- What are the challenges of the sim-to-real transfer paradigm?
- What is the reality gap?
- What are the popular strategies for performing an efficient transfer?
- What is Domain Randomization?
- What is Uniform Domain Randomization?
- What are the limitations of UDR?
- What is vision-based RL?

REFERENCES

[1] “Reinforcement Learning: An introduction (Second Edition)” by Richard S. Sutton and Andrew G. Barto, PDF

[2] Kober, J., Bagnell, J. A., & Peters, J. (2013). “Reinforcement learning in robotics: A survey”. The International Journal of Robotics Research, PDF

[3] Kormushev, P., Calinon, S., & Caldwell, D. G. (2013). “Reinforcement learning in robotics: Applications and real-world challenges”, PDF
[4] Höfer, S., Bekris, K., Handa, A., Gamboa, J. C., Golemo, F., Mozifian, M., ... & White, M. (2020). “Perspectives on sim2real transfer for robotics: A summary of the R: SS 2020 workshop”, PDF
[5] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World.” arXiv, Mar. 20, 2017. PDF
[6] Peng, X. B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018, May). “Sim-to-real transfer of robotic control with dynamics randomization”, PDF
[7] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.”, PDF
[8] Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015, June). “Trust region policy optimization”, PDF
[9] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). “Proximal policy optimization algorithms”, PDF
[10] Chebotar, Y., Handa, A., Makoviychuk, V., Macklin, M., Issac, J., Ratliff, N., & Fox, D. (2019, May). “Closing the sim-to-real loop: Adapting simulation randomization with real world experience”, PDF
[11] Tiboni, G., Arndt, K., & Kyrki, V. (2022). “DROPO: Sim-to-Real Transfer with Offline Domain Randomization”, PDF
[12] Tsai, Y. Y., Xu, H., Ding, Z., Zhang, C., Johns, E., & Huang, B. (2021). “Droid: Minimizing the reality gap using single-shot human demonstration”, PDF
[13] Muratore, F., Eilers, C., Gienger, M., & Peters, J. (2021). “Data-efficient domain randomization with bayesian optimization”, PDF
[14] V. Mnih et al., “Playing Atari with Deep Reinforcement Learning.” arXiv, Dec. 19, 2013. PDF