A PYBULLET EXPERIMENT CONFIGURATIONS

Table 3 shows the configuration used for each of the experiments wiht the PyBullet environments (Coumans & Bai, 2016). Then, specific details about the environments' feature spaces are shown in Table 4.

Table 3: Configuration used for the PyBullet experiments.

Parameter	**Value**
Discount factor γ	0.99
Maximum episode length	1,000
Algorithm	PPO (Schulman et al., 2017)
Epochs	200
Steps per epoch	4,000
Clip ratio	0.2
Learning rate policy π	0.0003
Learning rate value function v	0.001
Training iterations π	80
Training iterations v	80
λ	0.97
Target KL	0.1
Policy π	MLP, 64×64
Value function v	MLP, 64×64
Activation	Tanh

Table 4: Feature space details for PyBullet environments.

Parameter	**Number of features**
XYZ body position (all)	3
XYZ body velocity (all)	3
Roll (all)	1
Pitch (all)	1
Joint positions Hopper	3
Joint velocities Hopper	3
Contact points Hopper	1
Joint positions Ant	8
Joint velocities Ant	8
Contact points Ant	4
Joint positions Humanoid	17
Joint velocities Humanoid	17
Contact points Humanoid	2

B PADDLE ENVIRONMENT EXPERIMENT CONFIGURATIONS

Table 5 shows the environment configuration used for each of the experiments with the Paddle use case. Then, specific details about the random policy sampling experiments' configuration are shown in Table 6.

C TRAFFIC SIGNAL CONTROL EXPERIMENT CONFIGURATIONS

Table 7 shows the configuration used for each of the experiments with the RESCO benchmark (Ault & Sharon, 2021) for the Traffic Signal Control problem. Then, specific details about the experiments' configuration are shown in Table 8.

Table 5: Configuration used in the Paddle Environment.

Parameter	**Value**
Discount factor γ	0.95
Algorithm	DQN
Batch size	32
Number of episodes	600
$\epsilon_{\max}$	1.0
$\epsilon_{\min}$	0.01
ϵ decay	0.995
Policy	2-layer MLP
Hidden size	16
Δ_x ball	3 (baseline), 9 (hard)
Δ_y ball	3 (baseline), 9 (hard)
Δ_x paddle	20 (baseline), 60 (hard)

Table 6: Paddle experiments configuration.

Parameter	**Value**
Random policies	1,000
Rollouts per policy	150
Right tail filter	Top 5%

Table 7: Configuration used for the Traffic Signal Control environments.

Parameter	**Value**
Discount factor γ	0.99
Algorithm	DQN
Batch size	32
Number of episodes	100
Target network update frequency	500
Policy	2×2 conv. layer + MLP 64 $\times$ 64
Step length	10 seconds
Length yellow signal	3 seconds
Simulation length	1 hour

Table 8: Traffic Signal Control experiment configuration.

Parameter	**Value**
Random policies	500
Rollouts per policy	7
Right tail filter	Top 5%