A TRAINING HYPERPARAMETERS

This supplemental material contains details of the implementation to improve reproducibility of the research work. Table 2 summarize all Cycle-of-Learning hyperparameters used on the LunarLander-Continuous-v2 and Microsoft AirSim experiments.

In terms of hyperparameter tuning, due to sharing similar loss function structure, we started with the same hyperparameters used for a tuned DDPG algorithm for the same task and explored the algorithm performance by slightly varying the hyperparameters in the vicinity of those values. The lambda parameters that weight the combined loss were left at the default value of 1.0 but could be adjusted using standard hyperparameter tuning routines or adjusted based on the expertise of the demonstrator (higher λ_{BC} for high performing demonstrators, correcting for suboptimal demonstrations), type of reward function (higher λ_{Q_1} and λ_A for dense reward schemes), or the level of stochasticity and size of action- and state-spaces (lower λ_{L2Q} and $\lambda_{L2\pi}$ for reduced regularization on smaller and deterministic environments). With respect to how many demonstrations of the task to be used in the CoL, we understand there is no definite method to compute this number as in practice one would use as many as is feasible to collect. By using between 5 and 20 demonstrations we wanted to illustrate that the CoL could leverage even a small number of samples.

Table 2: Cycle-of-Learning hyperparemeters for each environment: (a) LunarLanderContinuous-v2 and (b) Microsoft AirSim.

	Environments	
Hyperparameter	(a)	(b)
λ_{Q_1} factor	1.0	1.0
λ_{BC} factor	1.0	1.0
λ_A factor	1.0	1.0
λ_{L2Q} factor	$1.0e^{-5}$	$1.0e^{-5}$
$\lambda_{L2\pi}$ factor	$1.0e^{-5}$	$1.0e^{-5}$
Batch size	512	512
Actor learning rate	$1.0e^{-3}$	$1.0e^{-3}$
Critic learning rate	$1.0e^{-4}$	$1.0e^{-4}$
Memory size	$5.0e^{5}$	$5.0e^{5}$
Expert trajectories	20	5
Pre-training steps	$2.0e^{4}$	$2.0e^{4}$
Training steps	$5.0e^{6}$	$5.0e^{5}$
Discount factor γ	0.99	0.99
Hidden layers	3	3
Neurons per layer	128	128
Activation function	ELU	ELU