A APPENDIX

Infinite horizon discounted LQR. Given a deterministic discrete-time linear system as in (6) and a quadratic γ-discounted cost function (1), the optimal policy is (see e.g., [4, Chapter 4.3])

$$\pi^\star(x) = K^\star(x - x^\star), \quad K^\star = -\gamma(R_\gamma + \gamma B^\top P B)^{-1} B^\top P A, \qquad (12)$$

where $K^\star$ is the discounted optimal gain and P is the unique positive solution of the discounted Discrete-time Algebraic Riccati Equation (DARE)

$$P = Q_\gamma + \gamma A^\top (P - \gamma P B(R_\gamma + \gamma B^\top P B)^{-1} B^\top P)A. \qquad (13)$$

From [6, Section 3], the value functions for the discounted LQR problem under the optimal controller are

$$\mathbf{J}^\star(x) = x^\top P x, \qquad \mathbf{Q}^\star(x, u) = z^\top H z, \qquad (14)$$

with $z := \mathrm{col}(x, u) \in \mathbb{R}^{n+m}$ and H given by

$$H = \begin{pmatrix} Q_\gamma + \gamma A^\top P A & \gamma A^\top P B \\ \gamma B^\top P A & R_\gamma + \gamma B^\top P B \end{pmatrix}.$$

Table 4: Pendulum swing-up environment parameters

Parameters	**Training**	**Corrupted**
pole mass m	1	1.2
pole length l	1	1
gravity acceleration g	9.81	9.81
episode max length T	200	1000
input bounds	[-2,2]	[-2,2]
noise w	0	$w \sim \mathcal{N}(0, 0.1)$
disturbance d	0	$0.2 \sin(\frac{2\pi}{100} t)$
steady-state threshold t_0	-	500

Table 6: Inverted pendulum swing-up environment parameters

Parameters	**Training**	**Corrupted**
cart mass m_c	10.47	10.47
pole mass m_p	5	6.53
pole length l_p	0.6	0.8
rail bounds l_r	[-1,1]	[-1,1]
episode max length T	1000	1000
input bounds	[-100,100]	[-100,100]
noise w	0	$w \sim \mathcal{N}(0, 0.173)$
disturbance d	0	$20 \sin(\frac{2\pi}{50} t)$
steady-state threshold t_0	-	500

Table 7: Double inverted pendulum swing-up environment parameters

Parameters	**Training**	**Corrupted**
cart mass m_c	10	10
first pole mass m_{p1}	1	1
second pole mass m_{p2}	1	1.2
first pole length l_{p1}	0.6	0.6
second pole length l_{p2}	0.6	0.7
rail bounds l_r	[-2,2]	[-2,2]
episode max length T	1000	2000
input bounds	[-200,200]	[-200,200]
noise w	0	$w \sim \mathcal{N}(0, 0.173)$
disturbance d	0	$20 \sin(\frac{2\pi}{100} t)$
steady-state threshold t_0	-	1500

Table 5: Hyperparameters. lr: learning rate, af: activation function, NN: hidden layer sizes

	PSU			**IPSU**			**DIPSU**		
	lr	**af**	**NN**	**lr**	**af**	**NN**	**lr**	**af**	**NN**
DDPG	0.001	ReLU	400,300	0.001	ReLU	400,300	0.0001	ReLU	400,300
LAS-DDPG	0.001	ReLU	400,300	0.001	ReLU	400,300	0.0001	Tanh	400,300
PPO	0.003	Tanh	64,64	0.00025	Tanh	256,256	0.0001	Tanh	64,64
LAS-PPO	0.003	Tanh	64,64	0.00025	Tanh	256,256	0.0001	Tanh	64,64
TD3	0.001	ReLU	400,300	0.001	ReLU	400,300	0.0006	ReLU	400,300
LAS-TD3	0.001	ReLU	400,300	0.001	ReLU	400,300	0.0002	Tanh	256,256
SAC	0.001	ReLU	256,256	0.001	ReLU	256,256	0.001	ReLU	256,256
LAS-SAC	0.001	ReLU	256,256	0.001	ReLU	256,256	0.0001	Tanh	256,256