Supplementary Material

Tables 1 and 2 describe parameter settings used in the experimentation for SARSA and DQN, respectively.

Table 1: Parameter settings for the tabular expected SARSA algorithm.

PARAMETER	DESCRIPTION	GRID-WORLD	CART-POLE	SUPPLY-CHAIN
	Table initialization	uniform on [0, 0.1]	zeros	uniform on [0, 0.1]
η_t	Learning rate (t episode #)	0.7	$\max\left\{\frac{1}{2}0.99^t, 0.01\right\}$	0.6
T	Max. episode length	200	200	200
μ_0	Prior parameter in (8)	0	0	0
τ_0	Prior parameter in (8)	1	1	1
a_0	Prior parameter in (8)	500	500	500
b_0	Prior parameter in (8)	500	500	500
α_0	Prior parameter for ε	1	10	1000
β_0	Prior parameter for ε	1 + 0.01	10 + 0.01	1000 + 0.01

Table 2: Parameter settings for the deep Q-learning algorithm.

PARAMETER	DESCRIPTION	GRID-WORLD	CART-POLE	SUPPLY-CHAIN
	Network initialization	Glorot uniform	Glorot uniform	Glorot uniform
	Network topology	16-25-25-4	4-12-12-2	102-100-100-100
f	Hidden activation	ReLU	ReLU	ReLU
	Regularization	none	L2(10^{-6})	none
ϕ	State encoding	one-hot	none	one-hot
η_t	Learning rate	0.001	0.0005	0.001
N	Replay buffer size	2000	2000	3000
B	Batch size	24	32	64
	Training epochs per batch	5	3	2
T	Max. episode length	200	200	200
μ_0	Prior parameter in (8)	0	0	0
τ_0	Prior parameter in (8)	1	1	1
a_0	Prior parameter in (8)	500	500	500
b_0	Prior parameter in (8)	500	500	500
α_0	Prior parameter for ε	1	5	25
β_0	Prior parameter for ε	1 + 0.01	5 + 0.01	25 + 0.01