RL applied to the CartPole Environment

Final Project for the Reinforcement Learning Lecture

Arce y de la Borbolla, José
Sälinger, Andreas

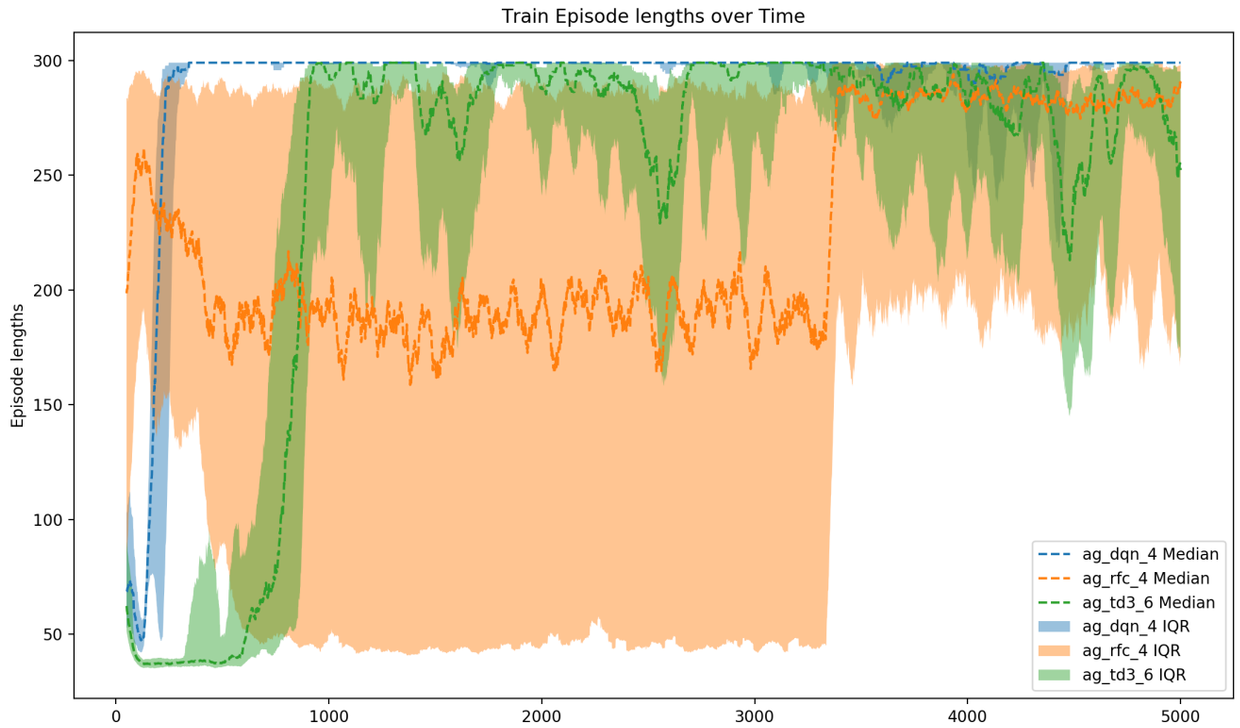
Algorithm

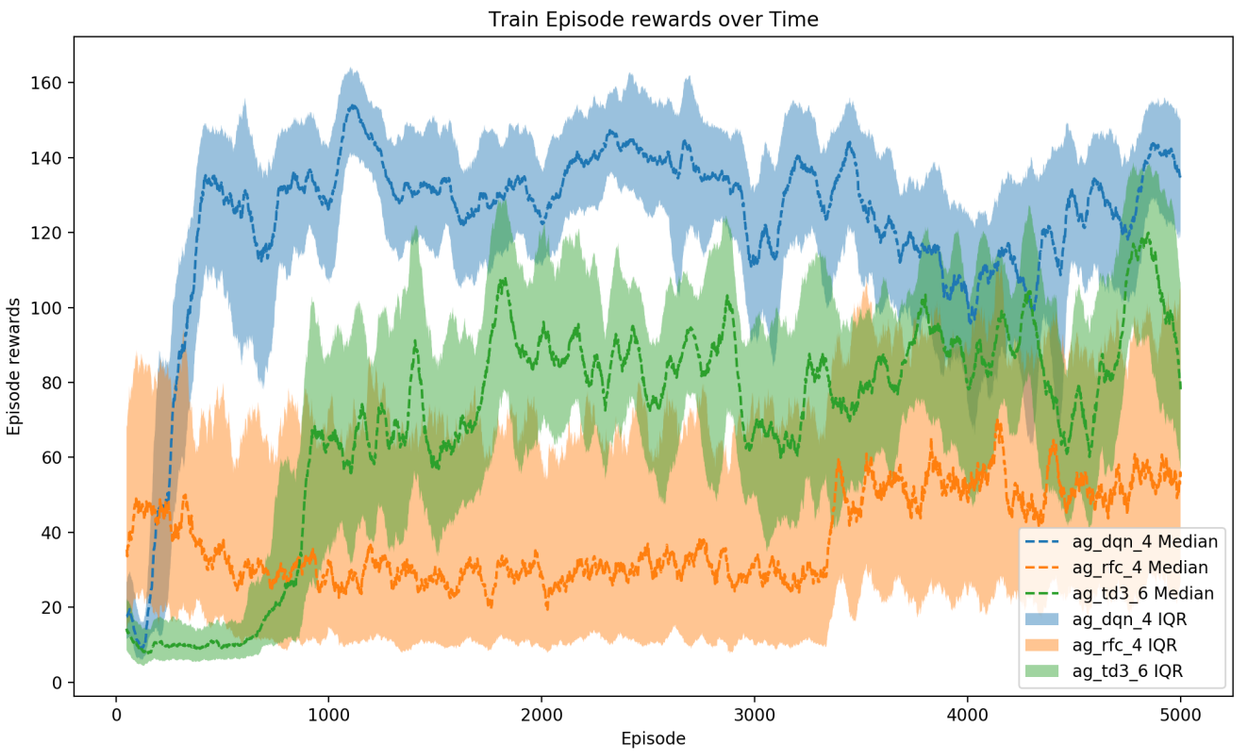
- We tested out DQN, REINFORCE and TD3

- All tests were performed with:
  - 5 runs x 5000 training episodes + 10 test episodes x 300 time steps

- The best results were obtained with Double-Q Deep Q-Network off-policy algorithm with Replay Buffer:
- 2 Discrete Actions {-1, 1}
- 2D Informative Positive reward function (see supplement)
- ±π noise for initial angular position, ± 0.5 noise for initial position, velocity and angular velocity.
- Learning rate 1e-4, discount factor gamma 0.98
- Epsilon-Greedy policy with cosine-annealed exponentially decaying schedule
- Q-Value Function Approximator:
  - 1 FC input Layer + ReLU, 1 FC hidden Layer with 20 units + ReLU, 1 FC output layer

Results





- The best results of each algorithm can be seen in this slide. Plots use a smoothing window of 50.
- A video for the best DQN agent is in `/slides/cartpole.mp4`
- All of the experiments can be seen in the supplement slides.
- All of the trained agents can be tested using `/scripts/test_agent.py`

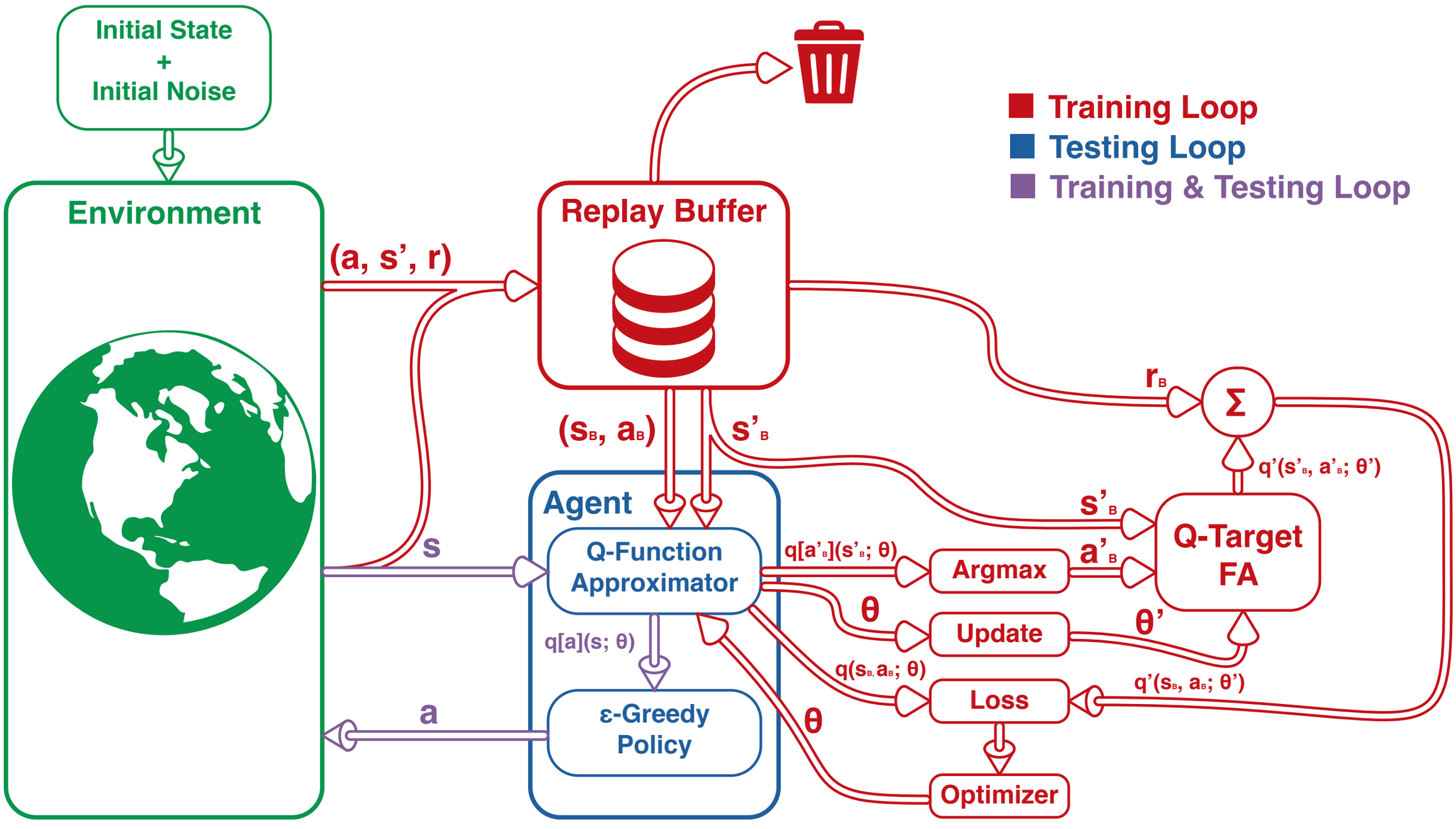
Test Episode Rewards and Lengths

Algorithm	Rewards			Lengths		
	Mean		σ	Mean		σ
dqn_4	135.6167	±	34.7674	300.0000	±	0
rfc_4	12.9702	±	13.5085	62.4800	±	56.3738
td3_6	75.4009	±	41.7202	216.7600	±	88.4225

Supplemental Slides

DQN Algorithm

- Double-Q Deep Q-Network off-policy algorithm with Replay Buffer



First agent

- For each algorithm, we started out with one particular configuration and created new agents by making one change in the design hyper-parameters and recording the results in a table.
- In order to keep it concise, we show the initial configuration here, and the series of changes with their results in the next slide.

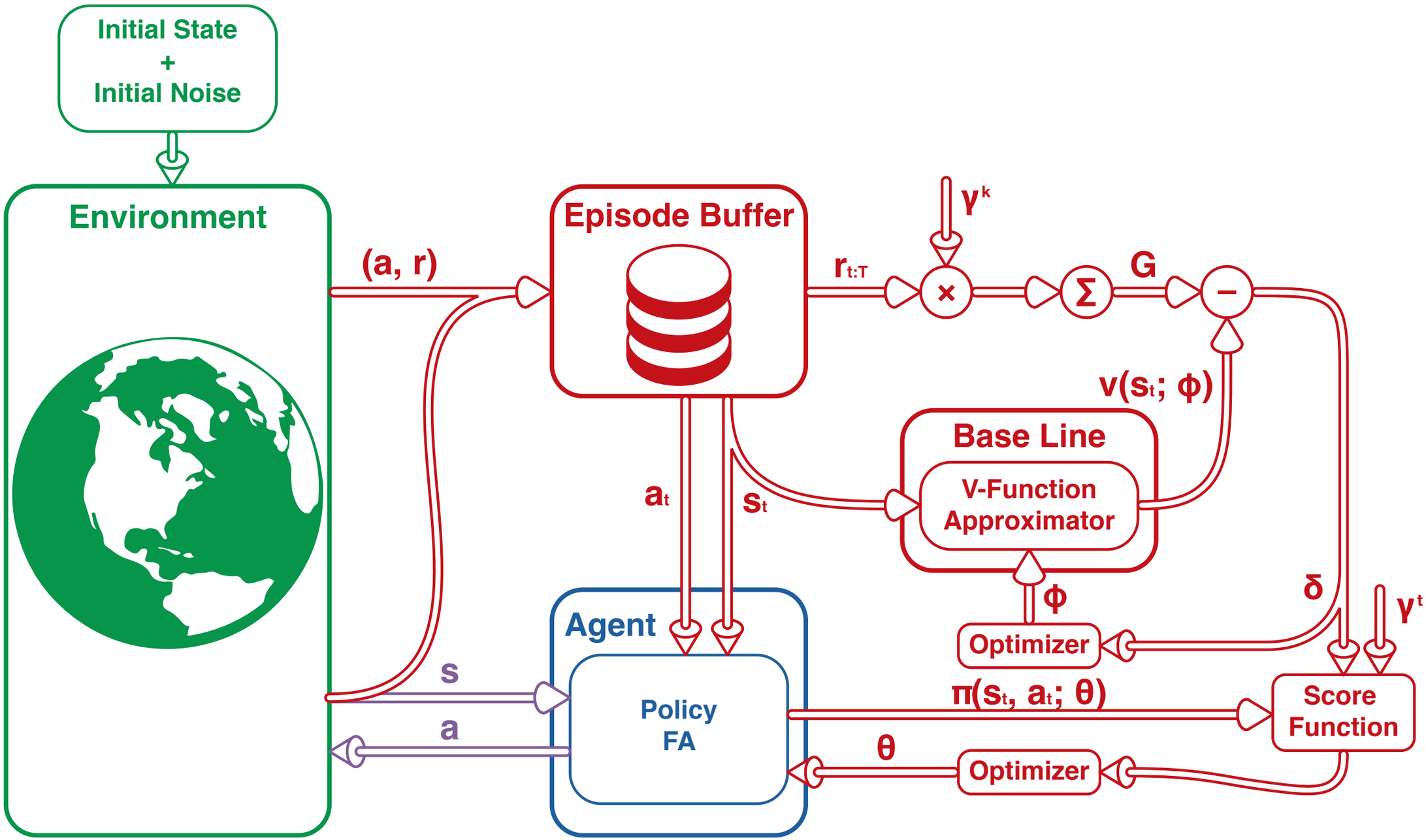
- Double-Q Deep Q-Network off-policy algorithm with Replay Buffer:
- 2 Discrete Actions {-1, 1}
- 2D Informative Positive reward function
- ±π noise for initial angular position, ± 0.5 noise for initial position, velocity and angular velocity.
- Learning rate 1e-4, discount factor gamma 0.99
- Epsilon-Greedy policy with cosine-annealed exponentially decaying schedule
- Q-Value Function Approximator:
  - 1 FC input Layer + ReLU, 1 FC hidden Layer with 20 units + ReLU, 1 FC output layer

DQN Experiments

No	Agent	Parent	Timestamp	Reward Function	Avg Test Score	StdDev	Change
1	agn_dqn_1	-	200207 044139	rf_info2d_pos	129.300	± 35.89	-
2	agn_dqn_2	1	200207 221202	rf_info2d_pos	116.610	± 48.83	t1:1200->2000
3	agn_dqn_3	1	200208 005040	rf_info2d_pos	54.000	± 63.31	gamma:0.99->0.999
4	agn_dqn_4	1	200208 105916	rf_info2d_pos	129.740	± 34.76	gamma:0.99->0.98
5	agn_dqn_5	1	200208 163200	rf_spar_pos	74.700	± 103.04	rf
6	agn_dqn_6	1	200209 000400	None	5.860	± 9.16	rf + lnNoise=0.5
7	agn_dqn_7	6	200209 100600	None	17.400	± 75.47	lr:1e-4->5e-4
8	agn_dqn_8	6	200209 204454	None	-0.800	± 0.4	gamma:0.99->1
9	agn_dqn_9	6	200210 054024	None	21.220	± 55.98	lnNoise=360
10	agn_dqn_10	4	200210 054117	rf_info2d_shp_pos	49.660	± 35.18	rf + lnNoise=360
11	agn_dqn_11	10	200210 114300	rf_info2d_shp_pos	22.110	± 26.07	lnNoise->None
12	agn_dqn_12	1	200210 181100	rf_info2d_pos	7.710	± 9.35	lnNoise->None
13	agn_dqn_13	1	200210 230800	rf_info2d_pos	44.160	± 23.64	rb:True->False
14	agn_dqn_14	1	200211 015000	rf_info2d_pos	100.500	± 55.36	edecay:exp->lin
15	agn_dqn_15	1	200211 193400	rf_info2d_pos	98.190	± 52.07	edecay:no ann
16	agn_dqn_16	4	200212 124100	rf_info2d_pos	110.290	± 38.94	gamma:0.98->0.95

REINFORCE Algorithm

- REINFORCE policy-gradient algorithm with State-Value Baseline



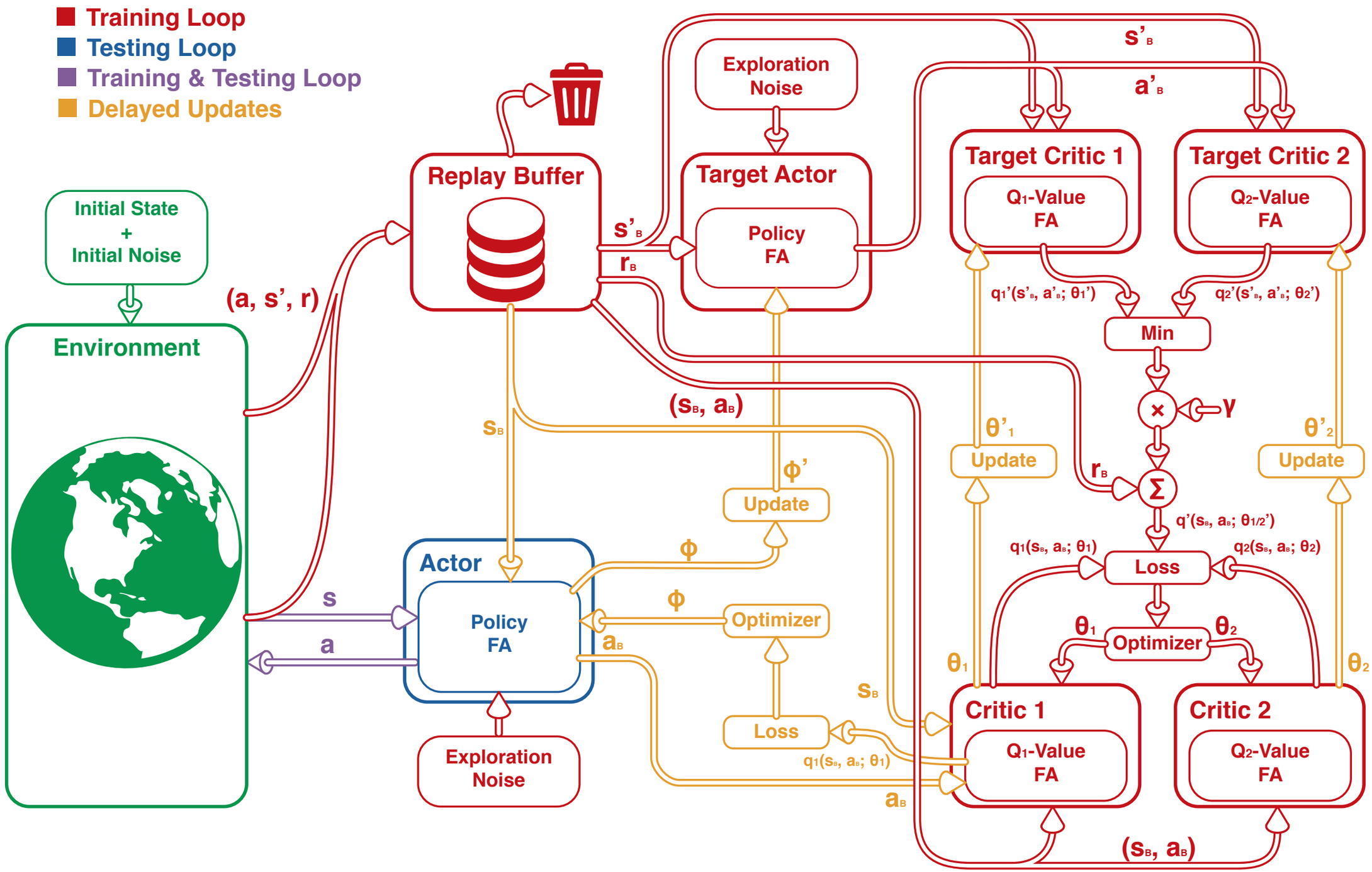
REINFORCE Experiments

- The starting agent had the following configuration
  - Continuous-Action REINFORCE agent with parameterized Beta Policy
  - With State-Value Function as baseline
  - 2D Informative Sharp Positive Reward Function,
  - Initial noise of ± 0.5 for all state variables
  - Learning rate for Pi and Baseline 1e-4, gamma 0.99

No	Agent	Parent	Timestamp	Reward Function	Avg Test Score	StdDev	Change
1	agn_rfc_1	-	200210 120500	rf_info2d_shp_pos	0.001	± 0.0007	-
2	agn_rfc_2	1	200210 180600	rf_info2d_pos	2.730	± 0.1142	rf
3	agn_rfc_3	2	200211 190300	rf_info2d_pos	10.770	± 7.16	code:multi-head
4	agn_rfc_4	3	200212 014200	rf_info2d_pos	12.970	± 13.50	InNoise=360
5	agn_rfc_5	4	200212 121400	rf_info2d_pos	9.540	± 6.49	pi=Gauss,actf=tanh
6	agn_rfc_6	4	200212 00	rf_info2d_pos	9.920	± 6.92	pi=MLP

TD3 Algorithm

- Twin Delayed Deep Deterministic policy-gradient algorithm (TD3)



TD3 Experiments

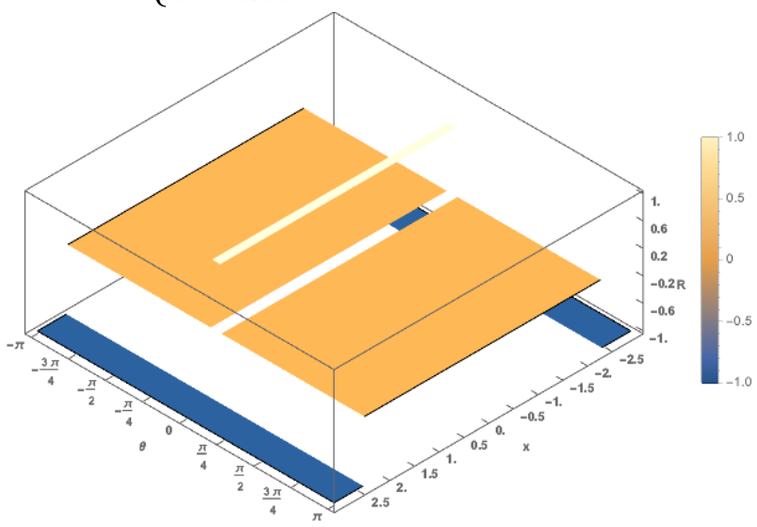
- The starting agent had the following configuration
  - Continuous-Action TD3 agent with MLP actor and critics function approximators.
  - 2D Informative Positive Reward Function,
  - Initial noise of ± 0.5 for all state variables
  - Learning rate of 1e-4, gamma 0.99
  - Soft-update weight tau 0.01, Policy update frequency of 2 steps

No	Agent	Parent	Timestamp	Reward Function	Avg Test Score	StdDev	Change
1	agn_td3_1	-	200210 024752	rf_info2d_pos	20.440	± 25.14	-
2	agn_td3_2	1	200210 114100	rf_info2d_shp_pos	0.002	± 0.0010	rf
3	agn_td3_3	1	200210 180400	rf_info2d_pos	2.710	± 0.13	tau:0.01->0.1
4	agn_td3_4	1	200210 191900	rf_info2d_pos	19.840	± 34.24	actnoise:0.2->0.1
5	agn_td3_5	1	200210 231000	rf_info2d_pos	38.300	± 32.35	lnNoise=360
6	agn_td3_6	5	200211 195800	rf_info2d_pos	75.400	± 41.72	lr:1e-4->5e-5
7	agn_td3_7	6	200211 122200	rf_info2d_pos	51.550	± 57.17	lr:5e-5->3e-5

Used Reward Functions

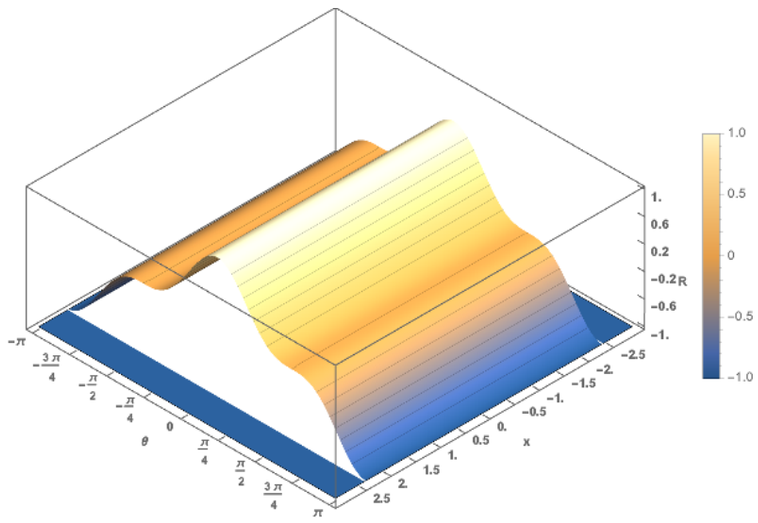
Default
None

$$R(x,\theta) = \begin{cases} -1 & \textbf{if } x < -x_{\text{lim}} \textbf{ or } x > x_{\text{lim}} \\ 1 & \textbf{if } -x_{\text{lim}} \le x \le x_{\text{lim}} \textbf{ and } -0.1 \le \theta \le 0.1 \\ 0 & \textbf{else} \end{cases}$$



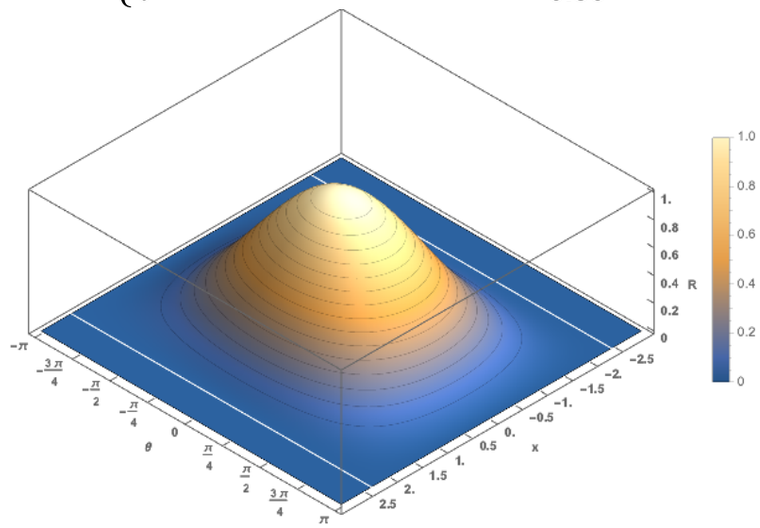
1D Informative
`rf_inf`

$$R(x,\theta) = \begin{cases} \cos^3\theta & \textbf{if } -x_{\text{lim}} \le x \le x_{\text{lim}} \\ -1 & \textbf{else} \end{cases}$$



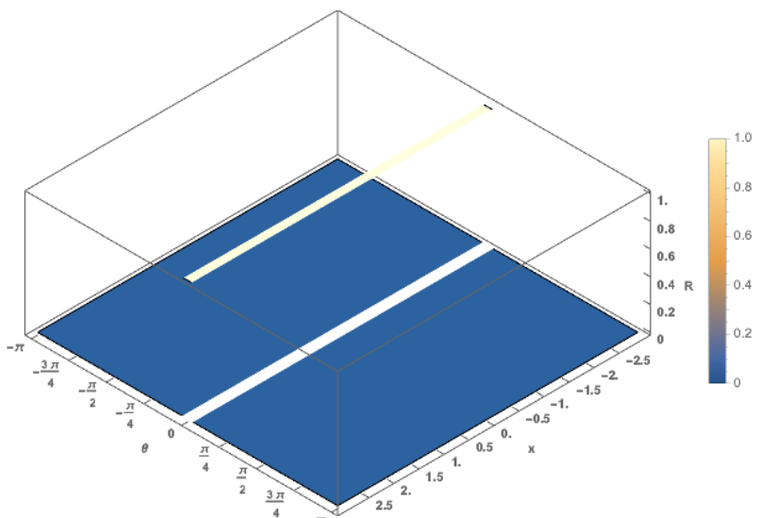
2D Informative Positive
`rf_info2d_pos`

$$R(x,\theta) = \begin{cases} \frac{1}{4}(\cos\theta + 1)\left(\cos\left(\frac{\pi x}{x_{\text{lim}}}\right) + 1\right) & \textbf{if } -x_{\text{lim}} \le x \le x_{\text{lim}} \\ 0 & \textbf{else} \end{cases}$$



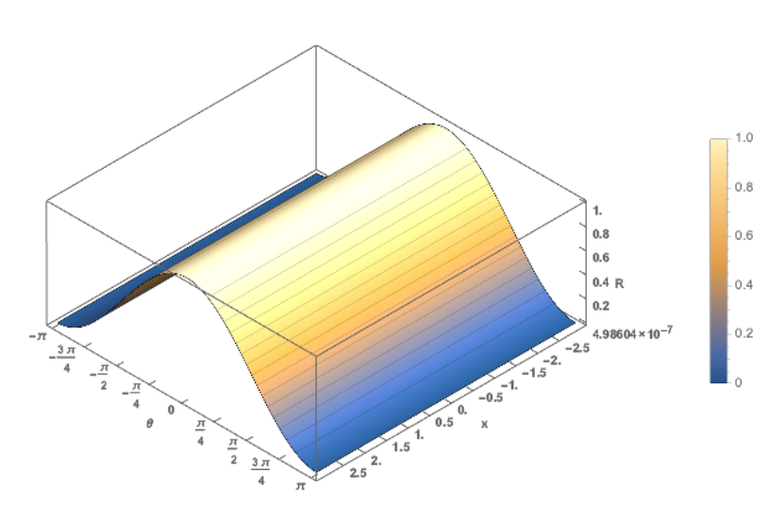
Sparse Positive
`rf_spar_pos`

$$R(x,\theta) = \begin{cases} 1 & \textbf{if } -0.1 \le \theta \le 0.1 \\ 0 & \textbf{else} \end{cases}$$



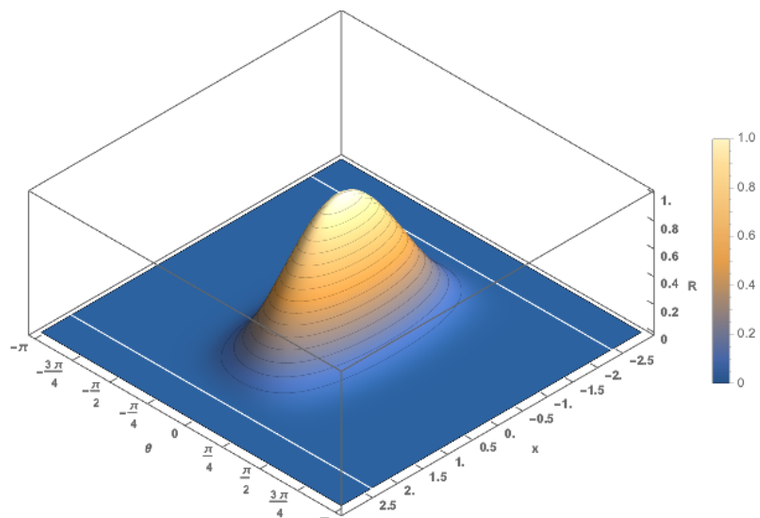
1D Informative Positive
`rf_inf_pos`

$$R(x,\theta) = \frac{1}{2}(\cos\theta + 1)$$



2D Informative Sharp Positive
`rf_info2d_sharp_pos`

$$R(x,\theta) = \begin{cases} \frac{1}{2}\left(\cos^{17}\frac{\theta}{2}\right)\left(\cos\left(\frac{\pi x}{x_{\text{lim}}}\right) + 1\right) & \textbf{if } -x_{\text{lim}} \le x \le x_{\text{lim}} \\ 0 & \textbf{else} \end{cases}$$



Schedules for ε-Greedy Policy