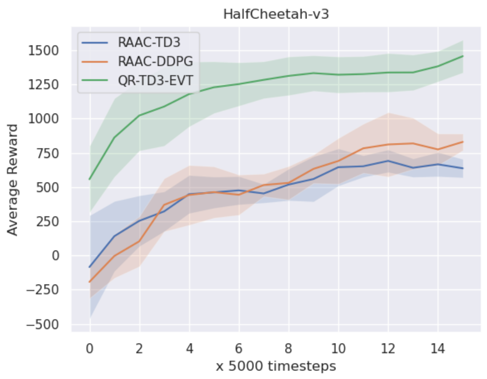
Mujoco Experiments:
Number of training steps: 100,000

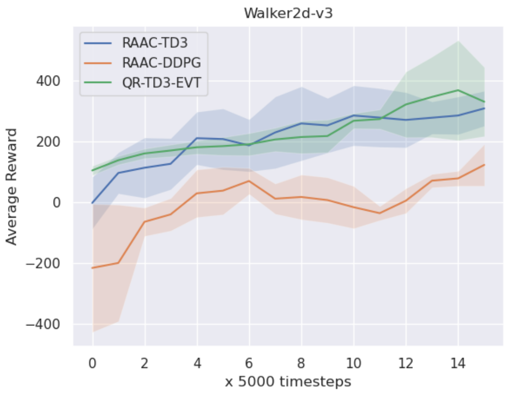
Environment	Algorithm	Mean Overshoot	Percentage Failure	Cumulative Reward
HalfCheetah (penalization rate p=0.05)	DDPG-RAAC	0.32 ± 0.08	8.37 ± 5.77	637.81 ± 319.78
	TD3-RAAC	0.52 ± 0.15	41.3 ± 16.6	836.8 ± 195.85
	EVT-RL	0.04 ± 0.03	**4.43 ± 2.29**	**1498.73 ± 277.4**
Hopper (penalization rate p=0.05)	DDPG-RAAC	0.05 ± 0.03	56.2 ± 8.98	272.39 ± 168.27
	TD3-RAAC	0.01 ± 0.01	6.39 ± 5.56	307.69 ± 48.49
	EVT-RL	0.03 ± 0.04	**1.29 ± 1.82**	**252.21 ± 17.82**
Walker (penalization rate p=0.05)	DDPG-RAAC	0.21 ± 0.05	18.94 ± 11.5	189.53 ± 87.93
	TD3-RAAC	0.21 ± 0.07	14.76 ± 5.8	273.21 ± 77.84
	EVT-RL	0.17 ± 0.24	**1.15 ± 1.63**	**496.35 ± 348.92**

Safety Gym Experiments:
Number of training steps: 70,000

Environment	Algorithm	Episode Length	Percentage Failure	Cumulative Reward
Point-Goal (Many smaller hazard regions) (penalization rate p=0.03)	DDPG-RAAC	210.0 ± 8.64	18.28 ± 4.03	4.2 ± 0.41
	TD3-RAAC	197.67 ± 8.81	12.13 ± 5.94	5.55 ± 0.23
	EVT-RL	229.67 ± 12.81	**0.0 ± 0.0**	**5.71 ± 0.0**
Point-Goal (One large hazard region) (penalization rate p=0.03)	DDPG-RAAC	231.0 ± 19.82	20.2 ± 14.66	5.22 ± 0.41
	TD3-RAAC	216.0 ± 9.42	22.34 ± 9.93	4.87 ± 0.47
	EVT-RL	243.67 ± 6.6	**0.0 ± 0.0**	**5.71 ± 0.0**

Training Curves:


HalfCheetah-v3
RAAC-TD3
RAAC-DDPG
QR-TD3-EVT
Average Reward
x 5000 timesteps


Walker2d-v3
RAAC-TD3
RAAC-DDPG
QR-TD3-EVT
Average Reward
x 5000 timesteps