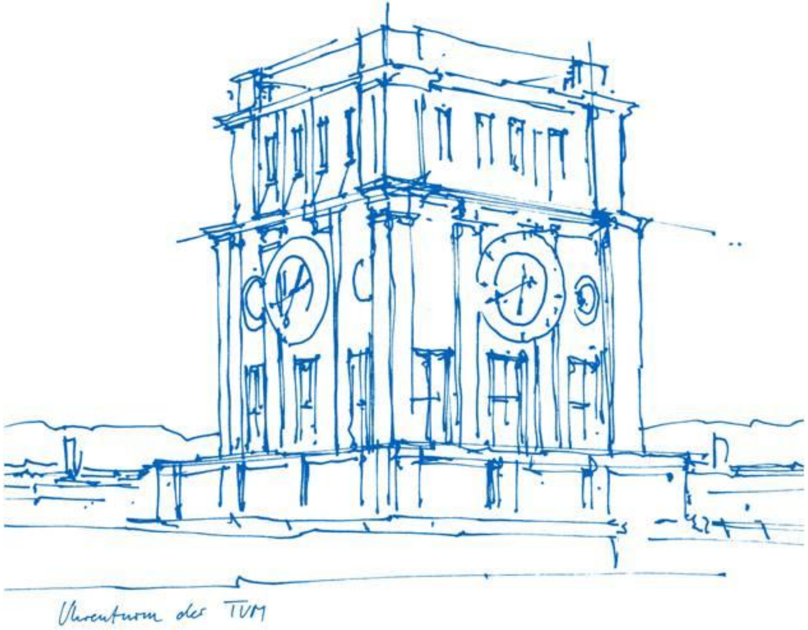
Munchausen RL with Continuous Action Space

Marcel Brucker

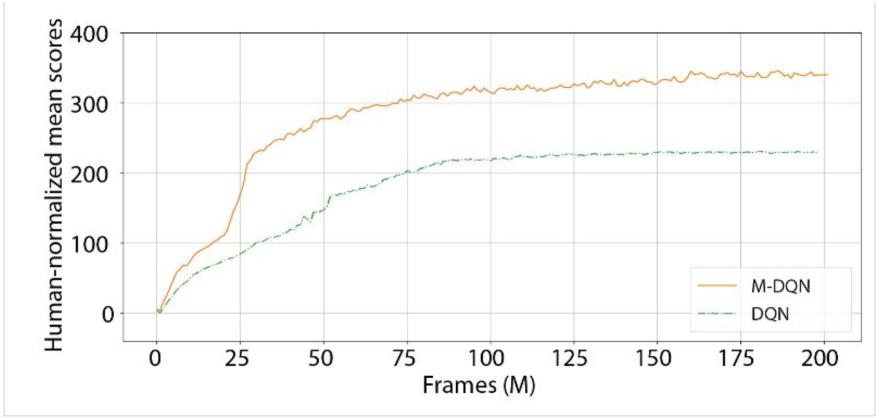
Finn Süberkrüb

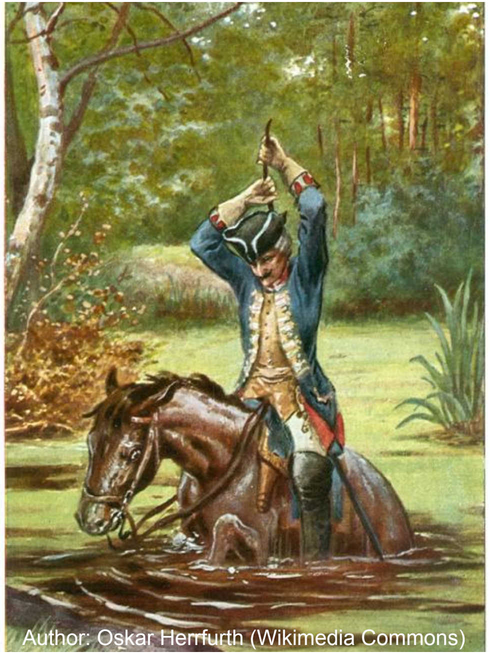
Technische Universität München

München, 22.07.2021



Munchausen RL[1]

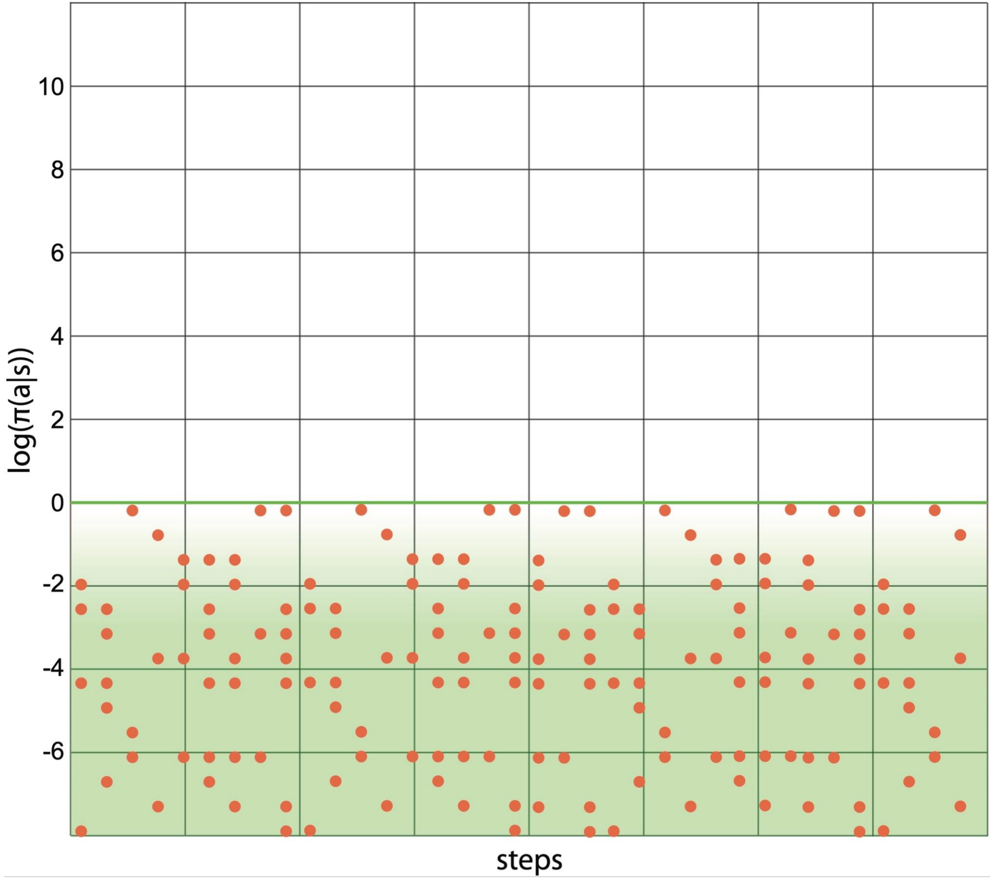




$$Q(s_t, a_t) = r_t + \tau[\alpha \ln \pi_\theta(a_t|s_t)]_{l_0}^{0} + \gamma \mathop{\mathbb{E}}_{s_{t+1} \sim p}[V(s_{t+1})]$$

Discreet actions

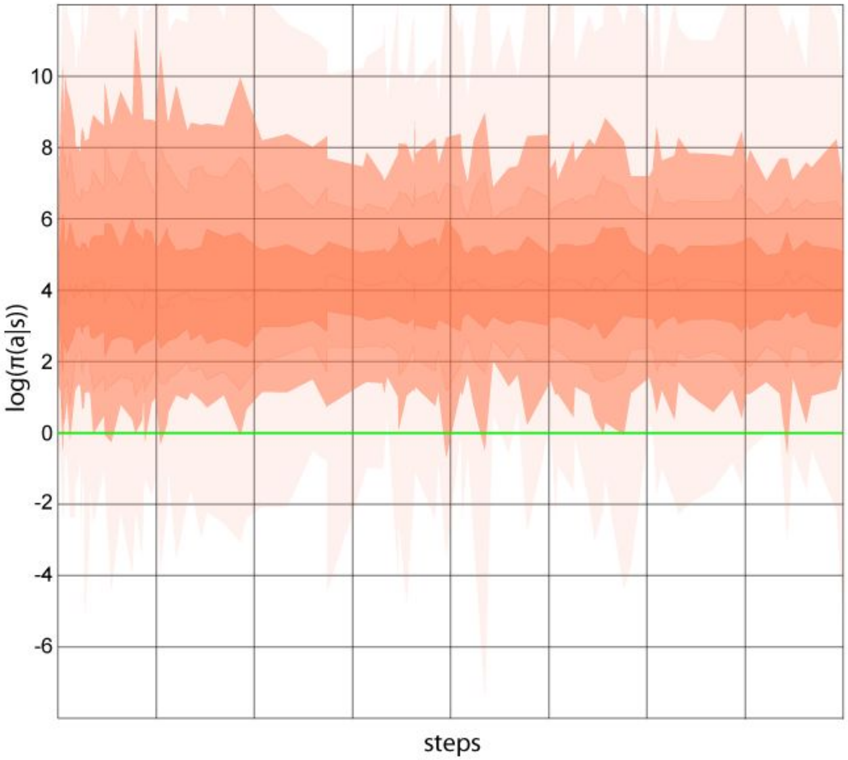
$$+\,\tau[\alpha \ln \pi_\theta(a_t|s_t)]_{l_0}^{0}$$



Continuous actions

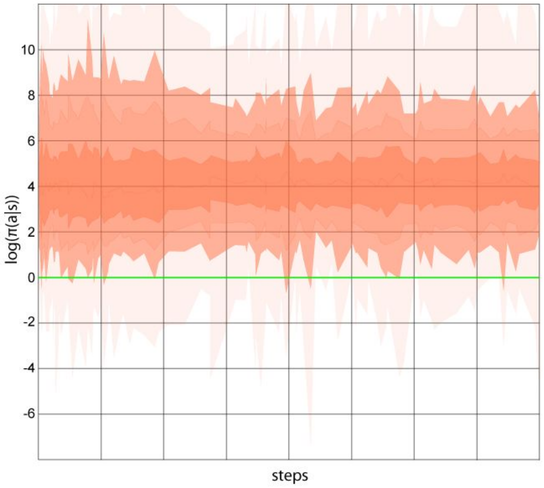
$$+\tau\alpha[\ln \overline{\pi}_\theta(a_t|s_t) + \beta]$$

$$\overline{\pi}_\theta(a|s) = \pi_\theta(a|s) - \mathop{\mathbb{E}}_{\substack{a' \in A \\ s' \in S}} [\pi_\theta(a'|s')]$$



M-SAC

$$\overline{\pi}_\theta(a|s) = \pi_\theta(a|s) - \mathop{\mathbb{E}}_{\substack{a' \in A \\ s' \in S}} [\pi_\theta(a'|s')]$$



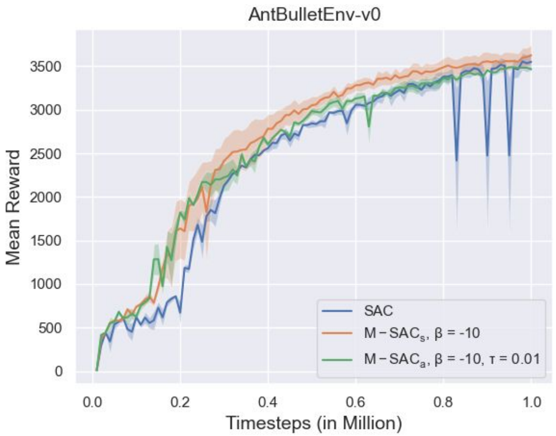
Action based

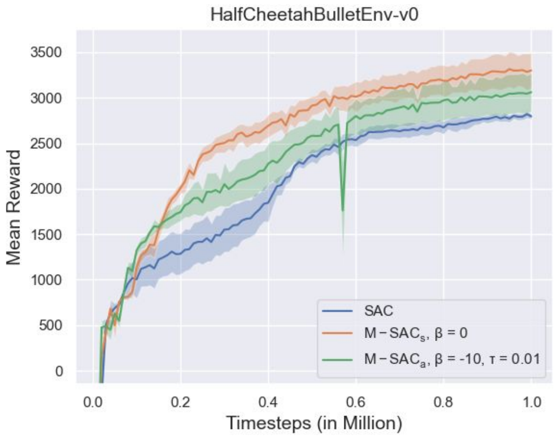
$$Q(s_t, a_t) = r_t + \tau\alpha[\ln \overline{\pi}_\theta(a_t|s_t) + \beta] + \gamma \mathop{\mathbb{E}}_{s_{t+1} \sim p} [V(s_{t+1})]$$

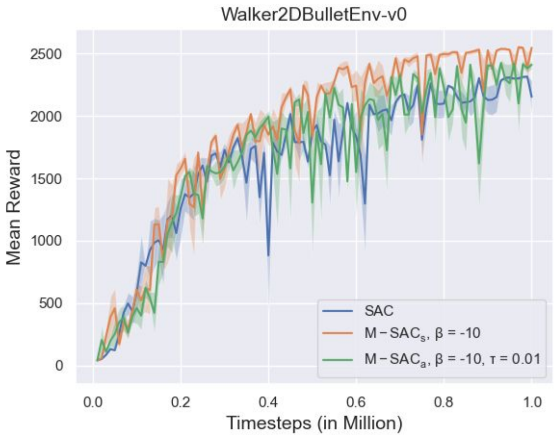
State based

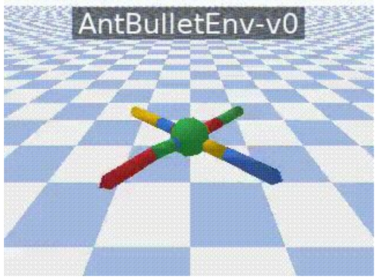
$$Q(s_t, a_t) = r_t + \tau\alpha \mathop{\mathbb{E}}_{\tilde{a} \sim \pi_\theta(\cdot|s)} [\ln \overline{\pi}_\theta(\tilde{a}|s_t) + \beta] + \gamma \mathop{\mathbb{E}}_{s_{t+1} \sim p} [V(s_{t+1})]$$

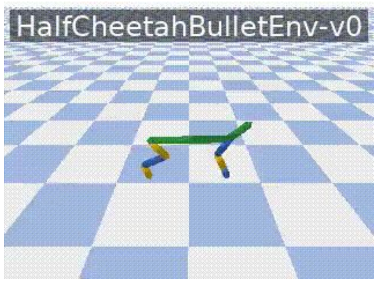
Results

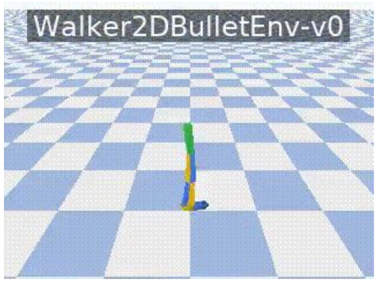

AntBulletEnv-v0
Mean Reward
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Timesteps (in Million)
SAC
M−SAC$_s$, β = -10
M−SAC$_a$, β = -10, τ = 0.01


HalfCheetahBulletEnv-v0
Mean Reward
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Timesteps (in Million)
SAC
M−SAC$_s$, β = 0
M−SAC$_a$, β = -10, τ = 0.01

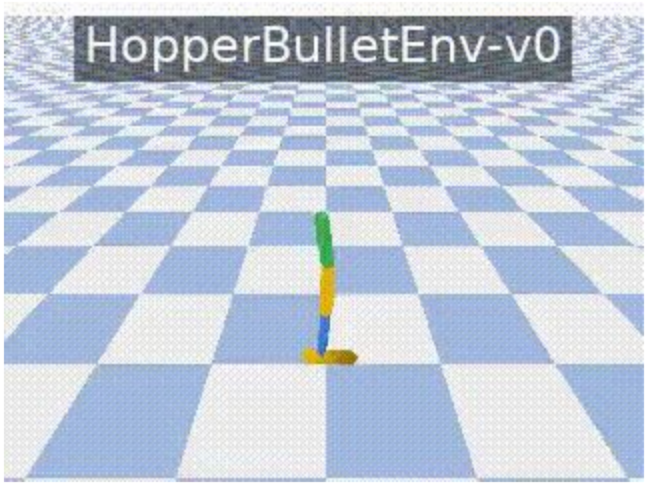

Walker2DBulletEnv-v0
Mean Reward
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Timesteps (in Million)
SAC
M−SAC$_s$, β = -10
M−SAC$_a$, β = -10, τ = 0.01

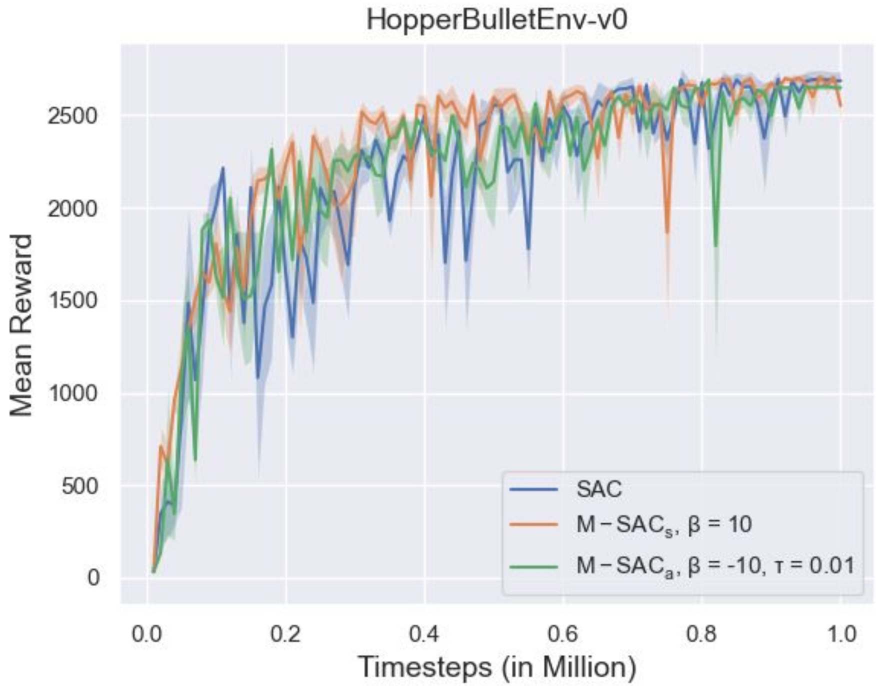

AntBulletEnv-v0


HalfCheetahBulletEnv-v0

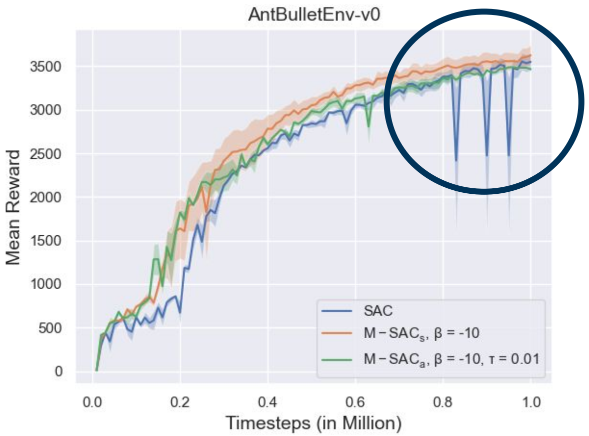

Walker2DBulletEnv-v0

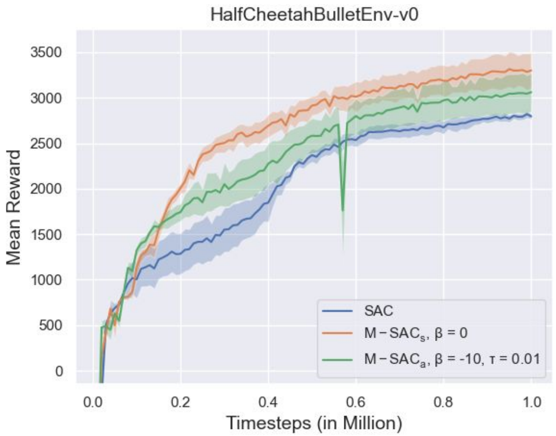
Results

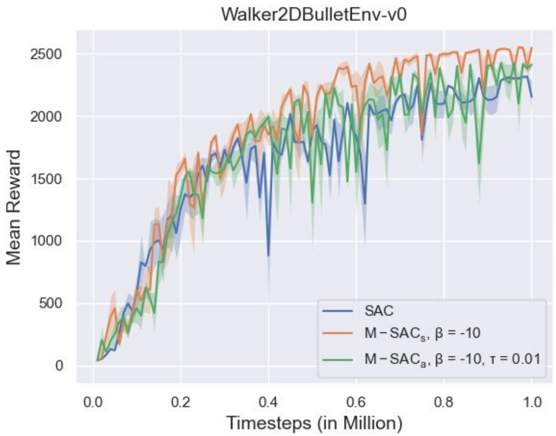

HopperBulletEnv-v0

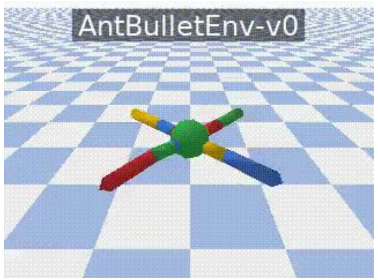

HopperBulletEnv-v0
Mean Reward
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Timesteps (in Million)
SAC
M−SAC$_s$, β = 10
M−SAC$_a$, β = -10, τ = 0.01

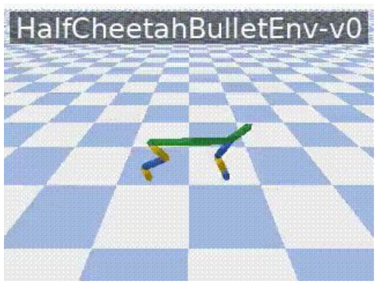
Results

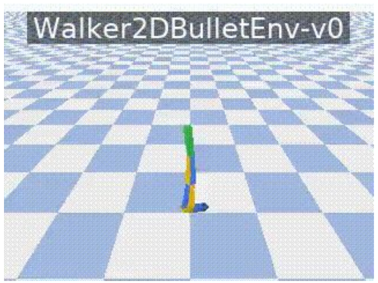

AntBulletEnv-v0
Mean Reward
Timesteps (in Million)
SAC
M−SAC$_s$, β = -10
M−SAC$_a$, β = -10, τ = 0.01


HalfCheetahBulletEnv-v0
Mean Reward
Timesteps (in Million)
SAC
M−SAC$_s$, β = 0
M−SAC$_a$, β = -10, τ = 0.01

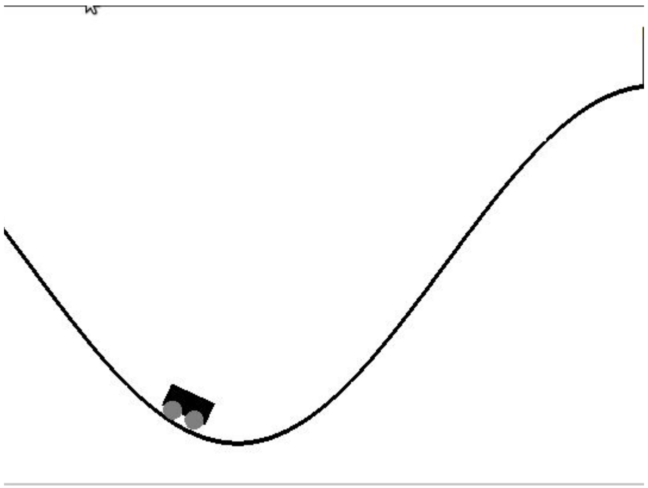

Walker2DBulletEnv-v0
Mean Reward
Timesteps (in Million)
SAC
M−SAC$_s$, β = -10
M−SAC$_a$, β = -10, τ = 0.01

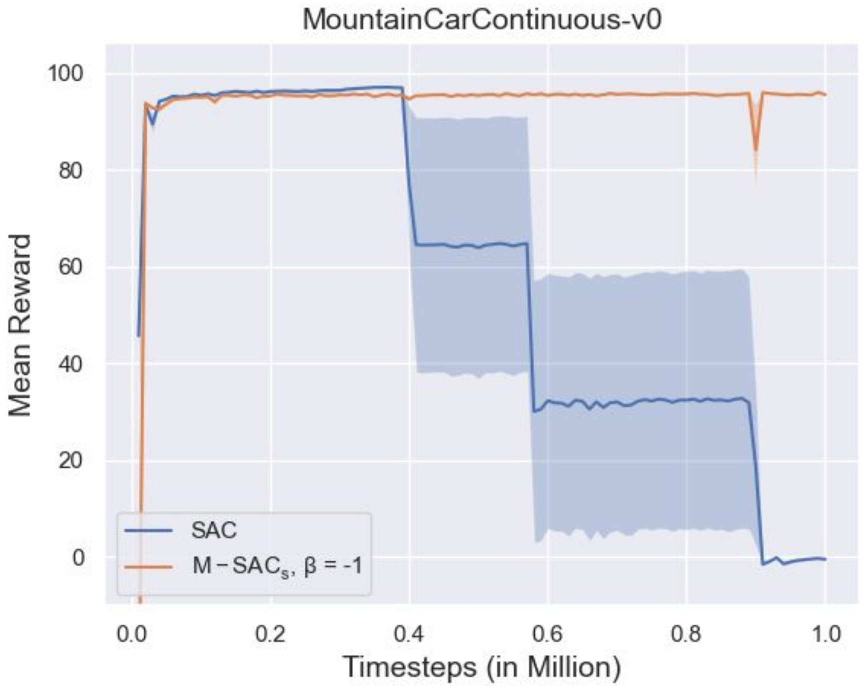

AntBulletEnv-v0


HalfCheetahBulletEnv-v0


Walker2DBulletEnv-v0

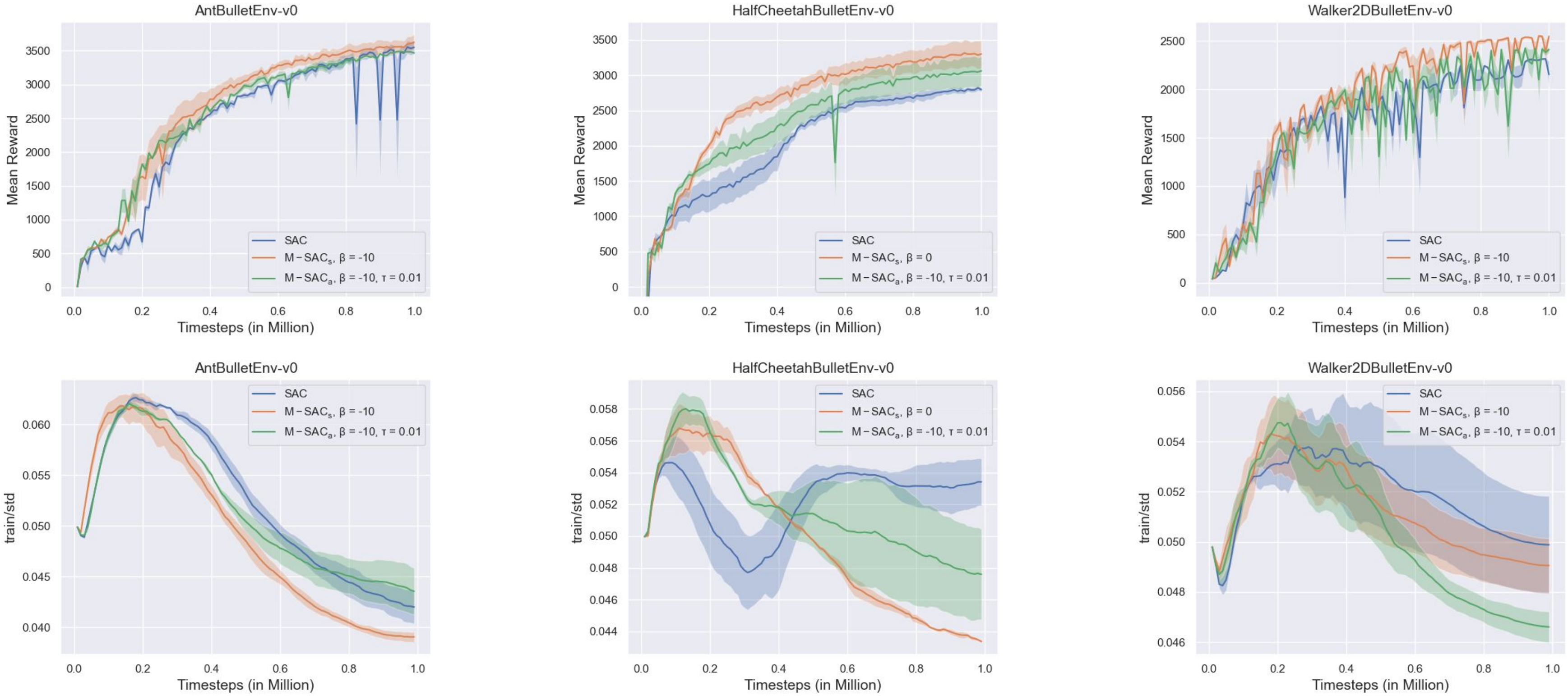
Results




MountainCarContinuous-v0
Mean Reward
Timesteps (in Million)
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
SAC
M − SAC$_s$, β = -1

Std. of Policy

→ Policy tends to become more deterministic if agent performs well



References

[1] Nino Vieillard, Olivier Pietquin and Matthieu Geist (2020). Munchausen Reinforcement Learning. arXiv:2007.14430.

[2] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel and Sergey Levine (2019). Soft Actor-Critic Algorithms and Applications. arXiv:1812.05905.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv:1801.01290.

[4] Make a four-legged creature walk forward as fast as possible. https://gym.openai.com/envs/Ant-v2/.

[5] Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto and Noah Dormann (2019). Stable Baselines3. GitHub repository, https://github.com/DLR-RM/stable-baselines3

[6] Antonin Raffin (2020). RL Baselines3 Zoo. GitHub repository, https://github.com/DLR-RM/rl-baselines3-zoo

[7] Benjamin Ellenberger (2018-2019). PyBullet Gymperium. GitHub repository, https://github.com/benelot/pybullet-gym

[8] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang and Wojciech Zaremba (2016). OpenAI Gym. arXiv:1606.01540.

Results (deterministic)

Environments	SAC	M-SAC$_a$	Improvement	M-SAC$_s$	Improvement
AntPyBulletEnv-v0	**3550 +/- 97**	3466 +/- 10	-2 %	**3625 +/- 116**	+2 %
HalfCheetahPyBulletEnv-v0	2796 +/- 27	**3063 +/- 216**	+10 %	**3301 +/- 188**	+18 %
HopperPyBulletEnv-v0	**2686 +/- 51**	2648 +/- 11	-1 %	2551 +/- 102	-5 %
Walker2DPyBulletEnv-v0	2154 +/- 131	**2411 +/- 32**	+12 %	**2548 +/- 15**	+18 %