

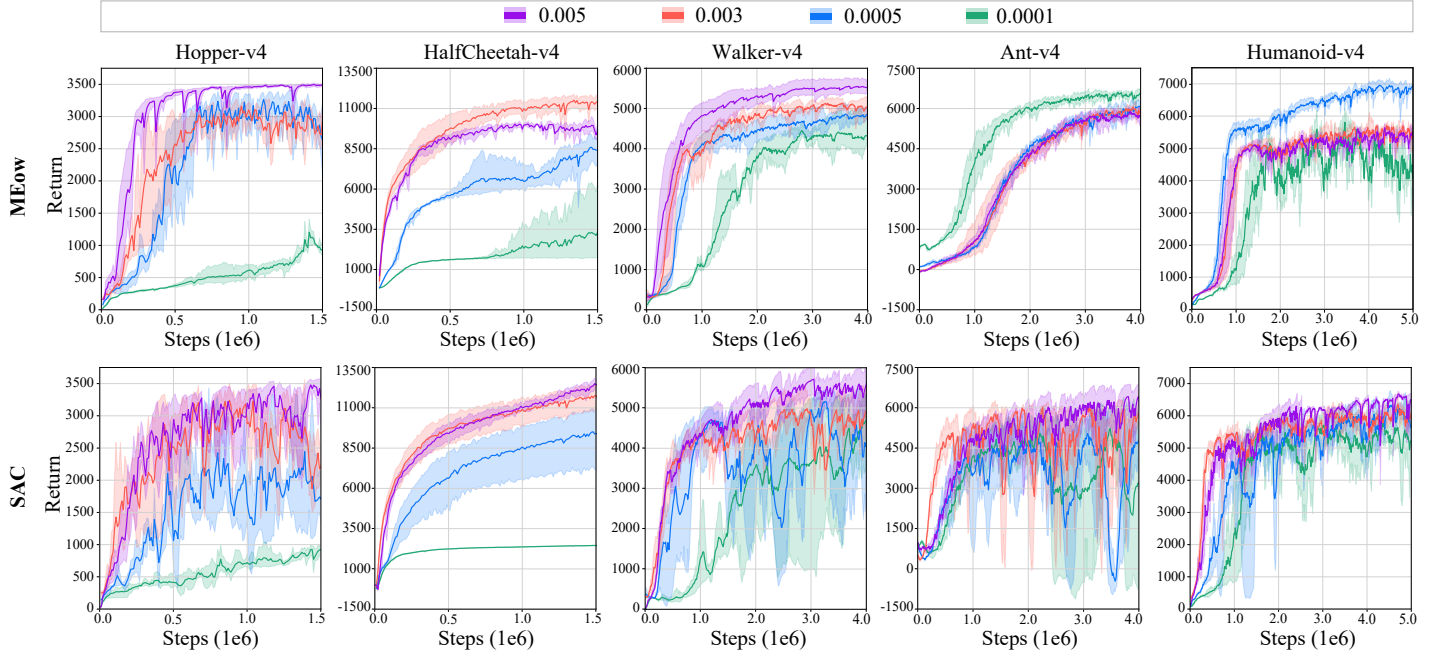


Figure 1. A performance comparison of MEow and SAC trained with different target smoothing values (i.e., $\tau = 0.005, 0.003, 0.0005, 0.0001$) across five MuJoCo environments.

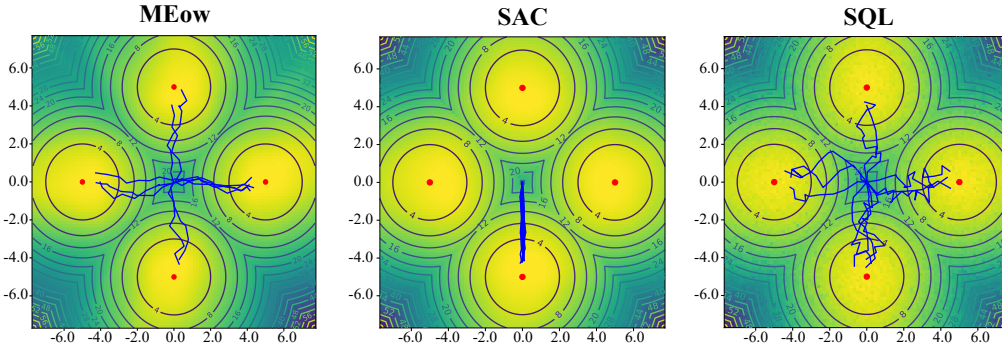


Figure 2. The soft value functions and the trajectories generated using MEow, SAC [1], and SQL [2] on the multi-goal environment.

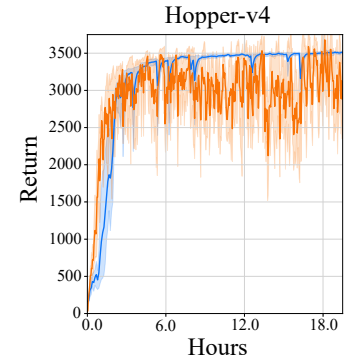


Figure 3. The return versus training time comparison of MEow (blue line) and SAC (orange line).

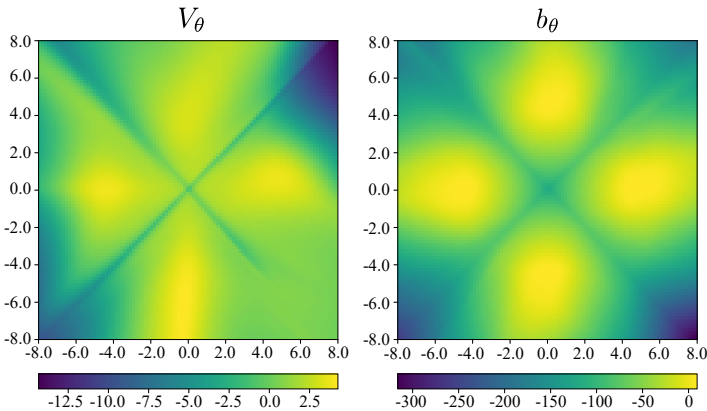


Figure 4. V_θ and b_θ of MEow evaluated on the multi-goal environment.

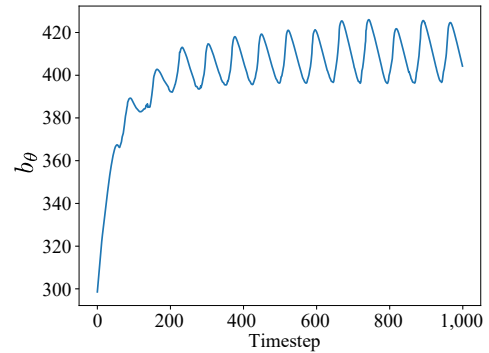


Figure 5. The learnable reward shifting values ($b_\theta(s_t)$) evaluated along a trajectory of states in Hopper-v4.