synthetic - N=10, budget=3, discount=0.9, 1 epochs

average reward

timestep t (50 episodes of length 20

optimal
ucw_value
ucw_qp
ucw_extreme
wiql
random