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

average cumulative reward

timestep $t$ (50 episodes of length 20)

Legend:
- optimal
- ucw_value
- ucw_qp
- ucw_extreme
- wiql
- random