1. Effect of α on Performance

Fig. 1 shows the effect of varying α in the indicator function. We can see if α is too low, performance degrades considerably because UTRAN degenerates to QMIX (the indicator function mistakes all targets for stochastic samples, and thus no target is transformed). In contrast, UTRAN replaces all targets with the best per-agent value if α is too high. Its poor performance in Fig. 1-a and c suggest difficulties in solving the stochastic state-action pairs.

Therefore, α is a task-specific hyperparameter related to the reward and the state scale. Our recommendation is to normalize the state, observation, and reward when we create an environment.

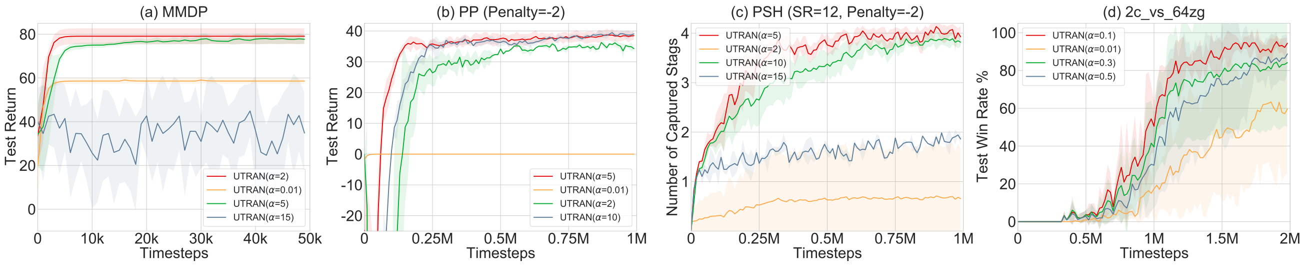

Figure 1. The results of UTRAN with different α on MMDP, PP, PSH, and SMAC.