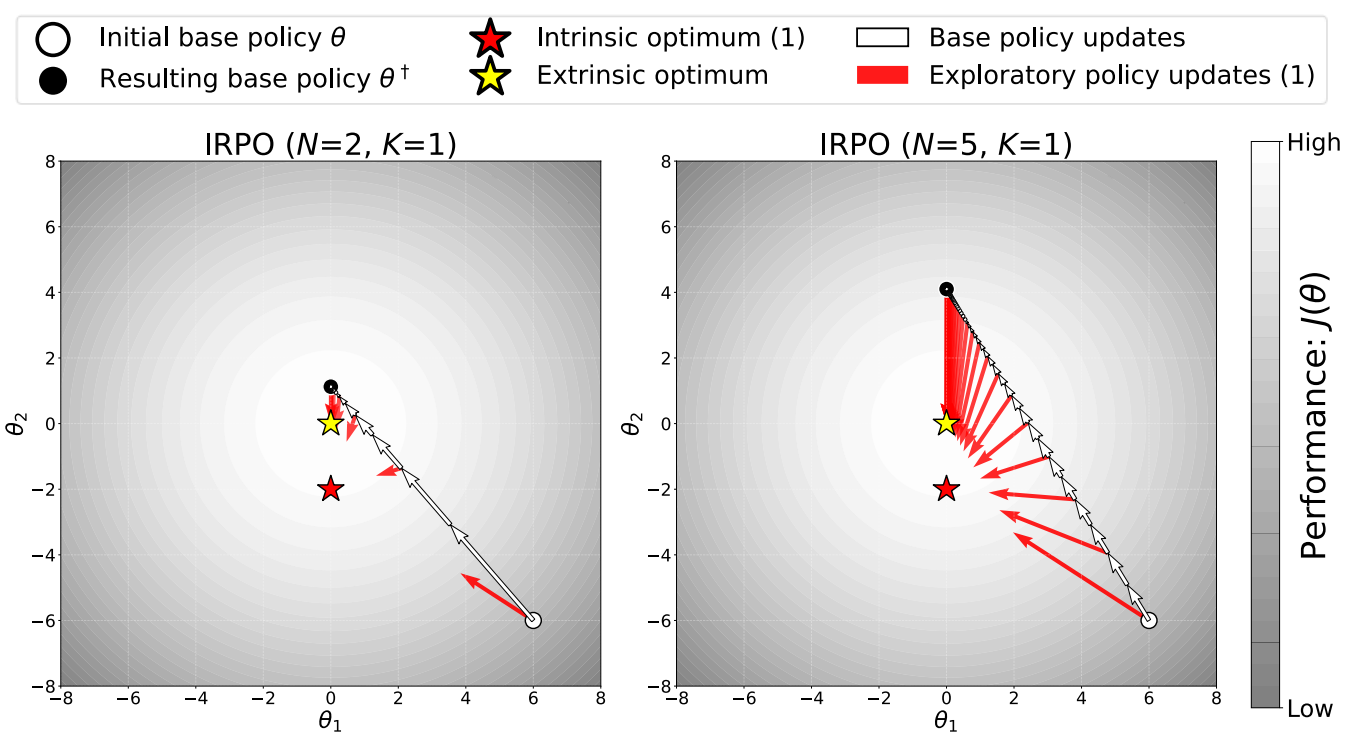
Initial base policy θ
Resulting base policy $\theta^{\dagger}$
Intrinsic optimum (1)
Extrinsic optimum
Base policy updates
Exploratory policy updates (1)
IRPO (N=2, K=1)
IRPO (N=5, K=1)
θ_1
θ_2
Performance: $J(\theta)$
High
Low