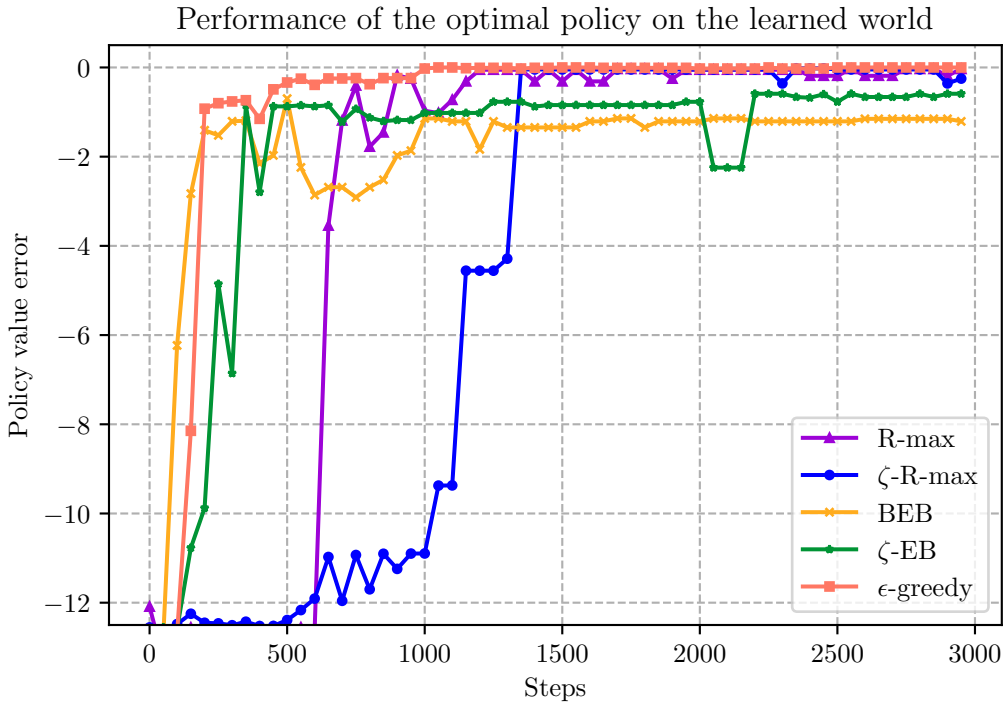
Performance of the optimal policy on the learned world
Policy value error
Steps
0
−2
−4
−6
−8
−10
−12
0
500
1000
1500
2000
2500
3000
R-max
ζ-R-max
BEB
ζ-EB
ϵ-greedy