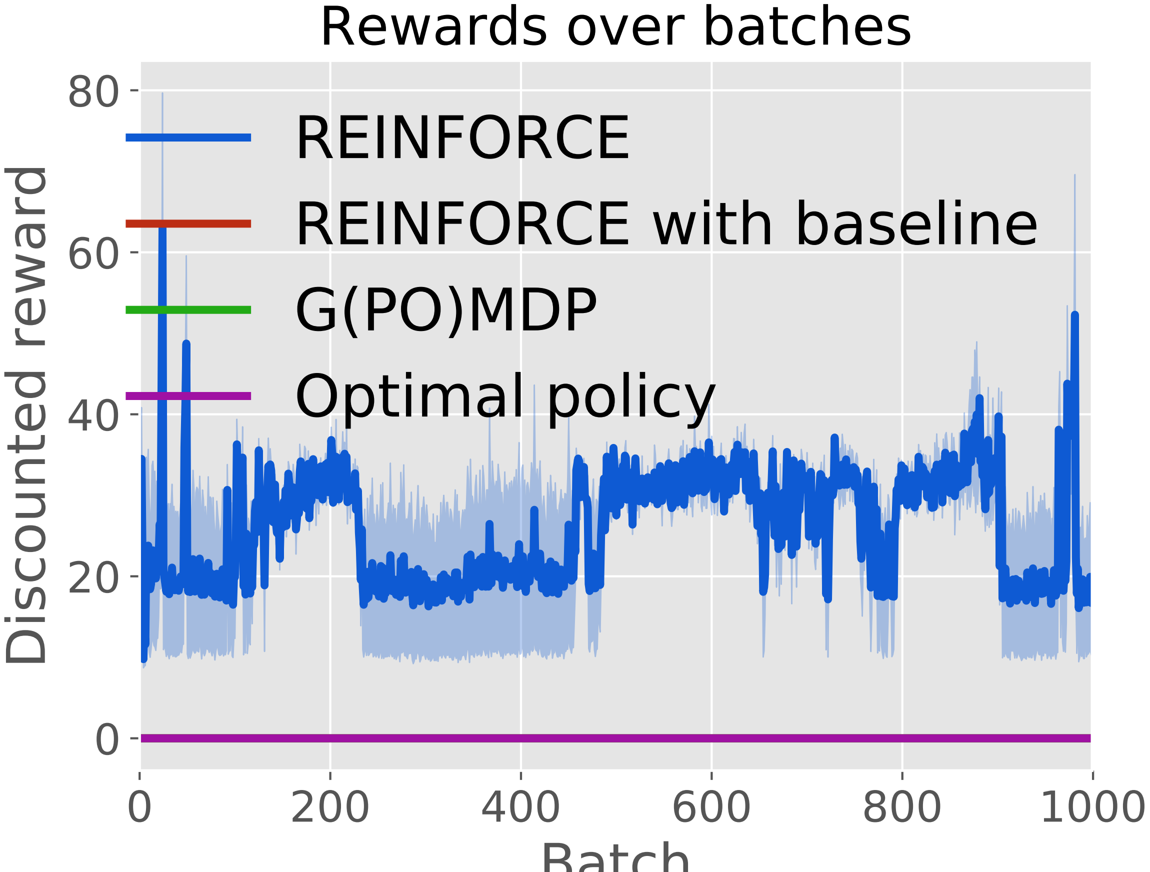
Rewards over batches
REINFORCE
REINFORCE with baseline
G(PO)MDP
Optimal policy
Discounted reward
80
60
40
20
0
0
200
400
600
800
1000
Batch