IDALAB

PARIS
LODRON
UNIVERSITÄT
SALZBURG

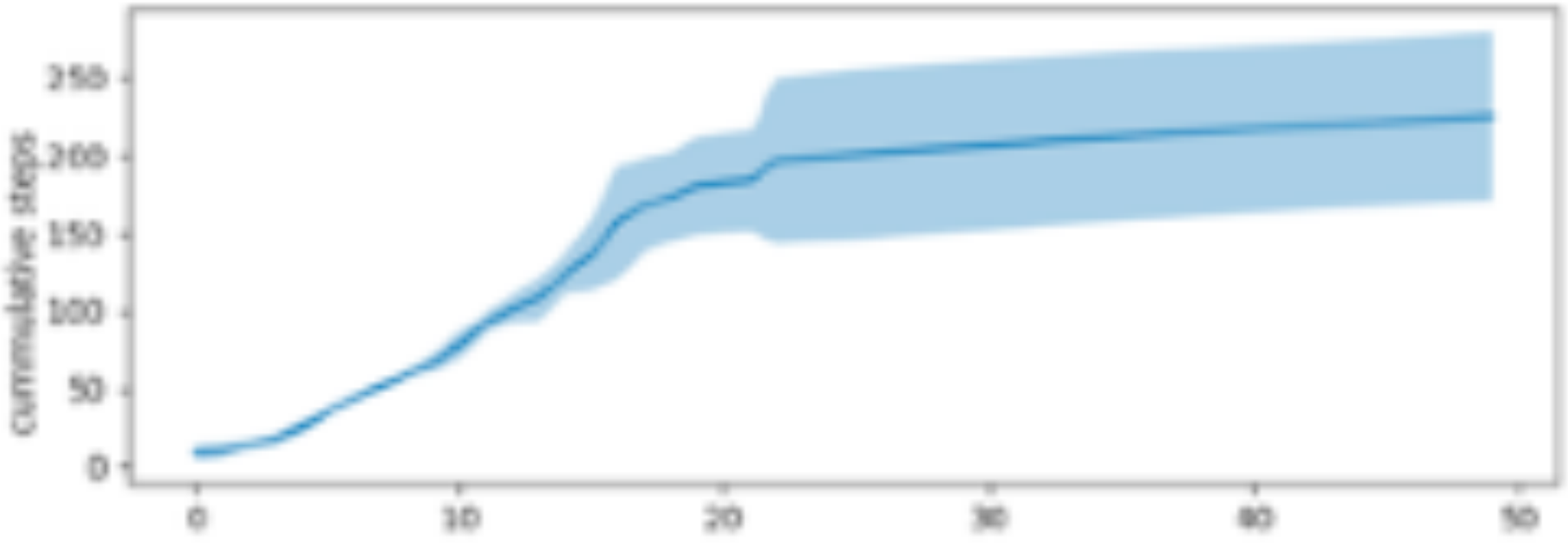
Simon Hirländer

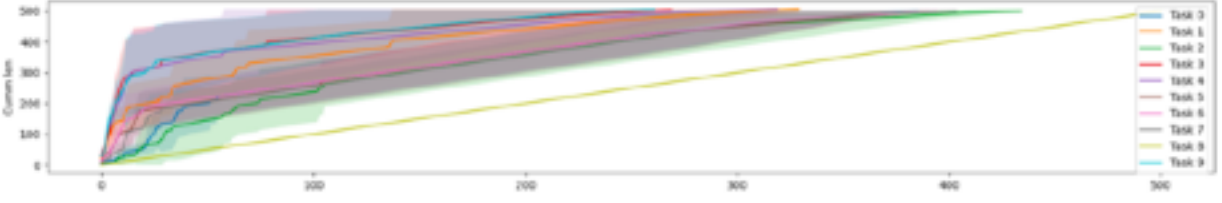
Tutorial RL4AA

Sample efficiency

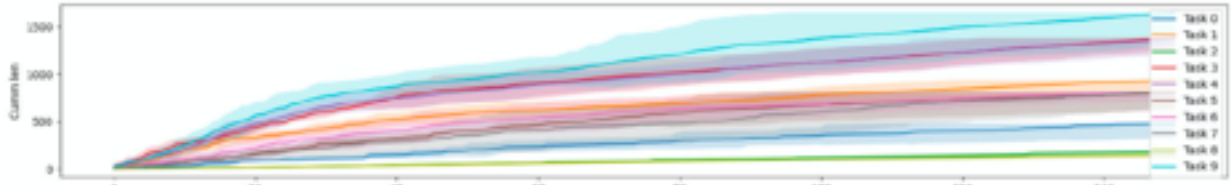
- Derivative free methods: (NES, CMA,..)
- 10 x Online methods (A3C)
- 10 x Policy-gradient methods (TRPO)
- 10 x Replay-Buffer + Value function estimation (Q-Learning, DDPG, TD3, NAF, SAC,…)
- 10 x Model-based RL methods (MPO, Guided Policy Search, Dyna)
- 10 x Model-based shallow methods (no NNs) Few shot GPs...

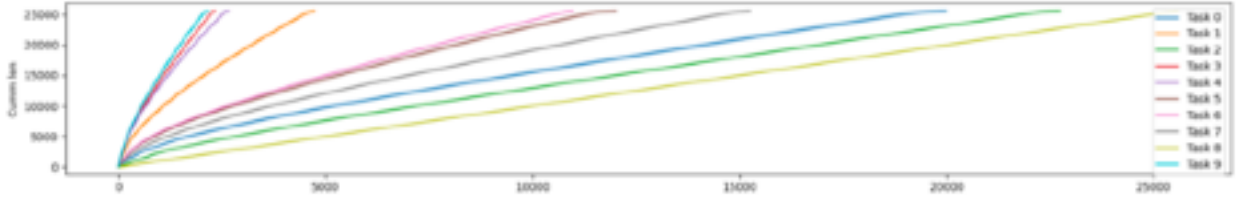
20

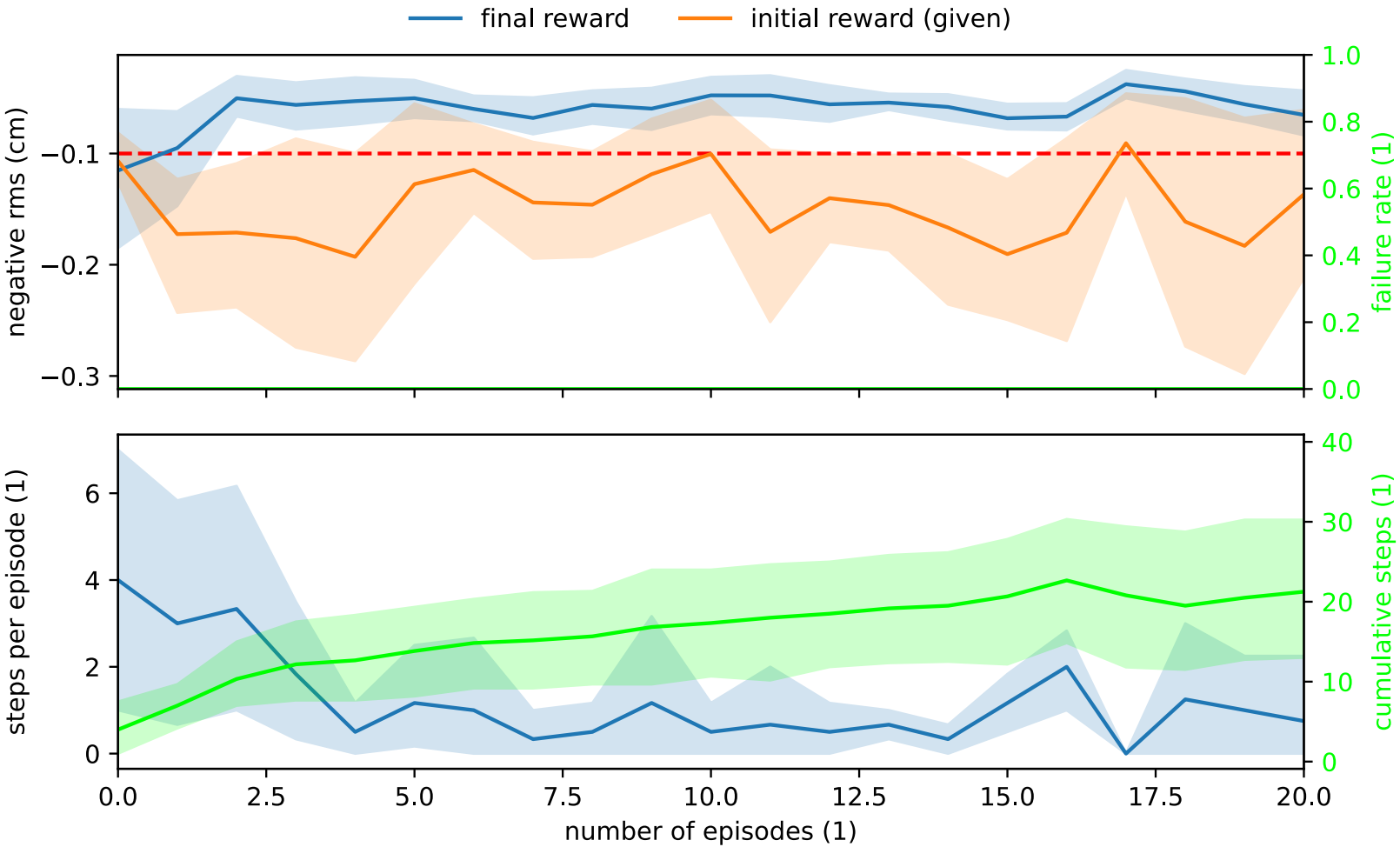

cummulative steps
250
200
150
100
50
0
0
10
20
30
40
50



Cumm km
0
100
200
300
400
500
Task 3
Task 1
Task 2
Task 3
Task 4
Task 5
Task 6
Task 7
Task 8
Task 9


Cumulative
1500
1000
500
0
0
20
40
60
80
100
120
140
Task 0
Task 1
Task 2
Task 3
Task 4
Task 5
Task 6
Task 7
Task 8
Task 9


Cumm loss
25000
20000
15000
10000
5000
0
0
5000
10000
15000
20000
25000
Task 0
Task 1
Task 2
Task 3
Task 4
Task 5
Task 6
Task 7
Task 8
Task 9


final reward
initial reward (given)
negative rms (cm)
−0.1
−0.2
−0.3
failure rate (1)
1.0
0.8
0.6
0.4
0.2
0.0
steps per episode (1)
6
4
2
0
cumulative steps (1)
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
number of episodes (1)

TRPO~50000

MPO~5000

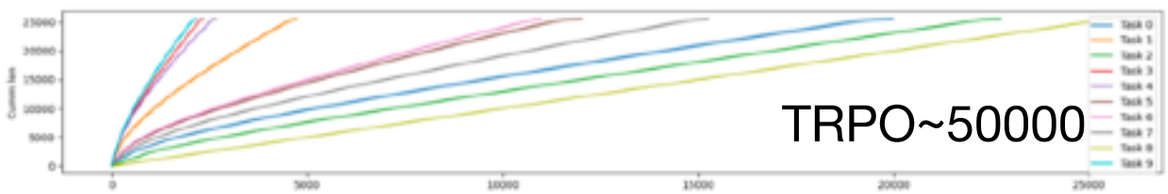
SAC<500

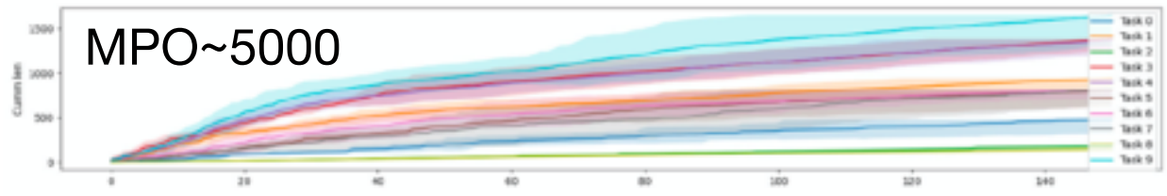
MBPO<100

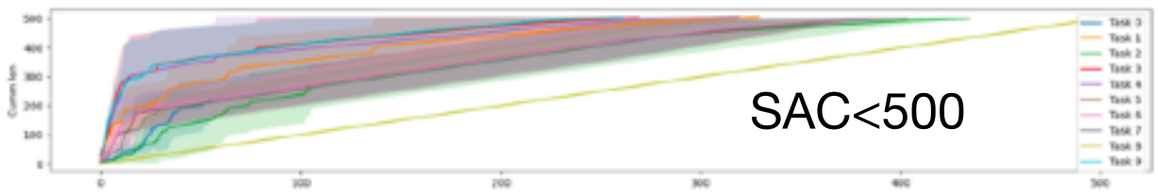
GP-MPC < 50

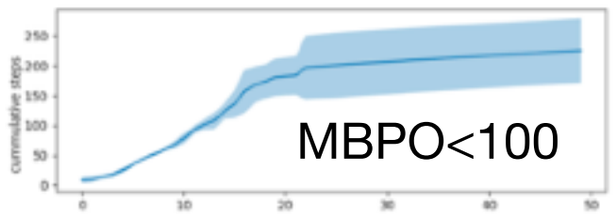
Sample efficiency

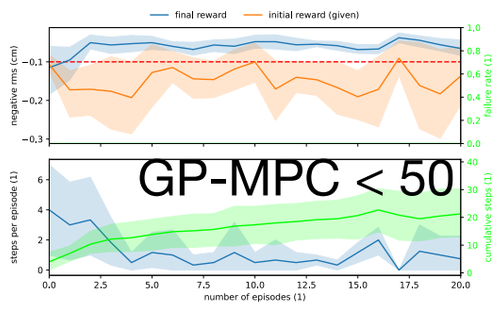
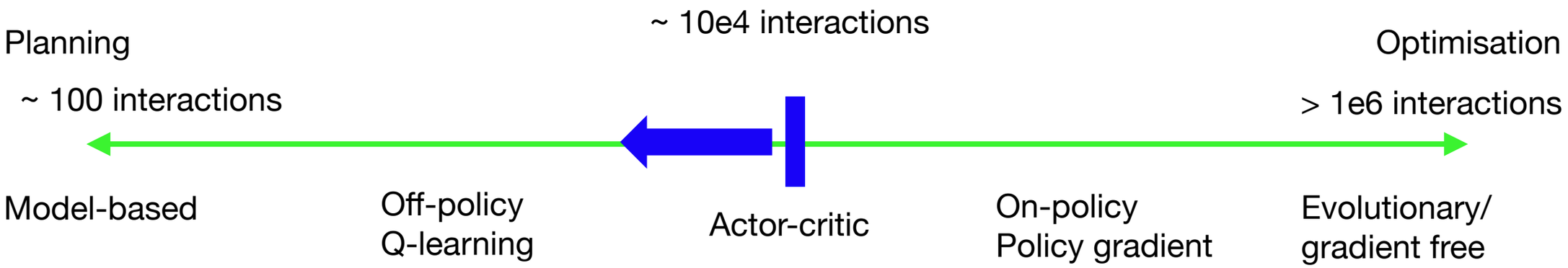
- Derivative free methods: (NES, CMA,..)
- 10 x Online methods (A3C)
- 10 x Policy-gradient methods (TRPO)
- 10 x Replay-Buffer + Value function estimation (Q-Learning, DDPG, TD3, NAF, SAC,…)
- 10 x Model-based RL methods (MPO, Guided Policy Search, Dyna)
- 10 x Model-based shallow methods (no NNs) Few shot GPs...











But sample efficiency is not all