RL benchmark - Manual hyperparameter tuning

1 Introduction and methods

This is a benchmark on the reaching WidowX arm. We vary the hyperparameters and keep the same training environment.

- Algorithms: PPO2
- Environment: widowx-reacher-v5
- 6 joints
- Fixed goal
- Dense reward: -dist**2

2 Results

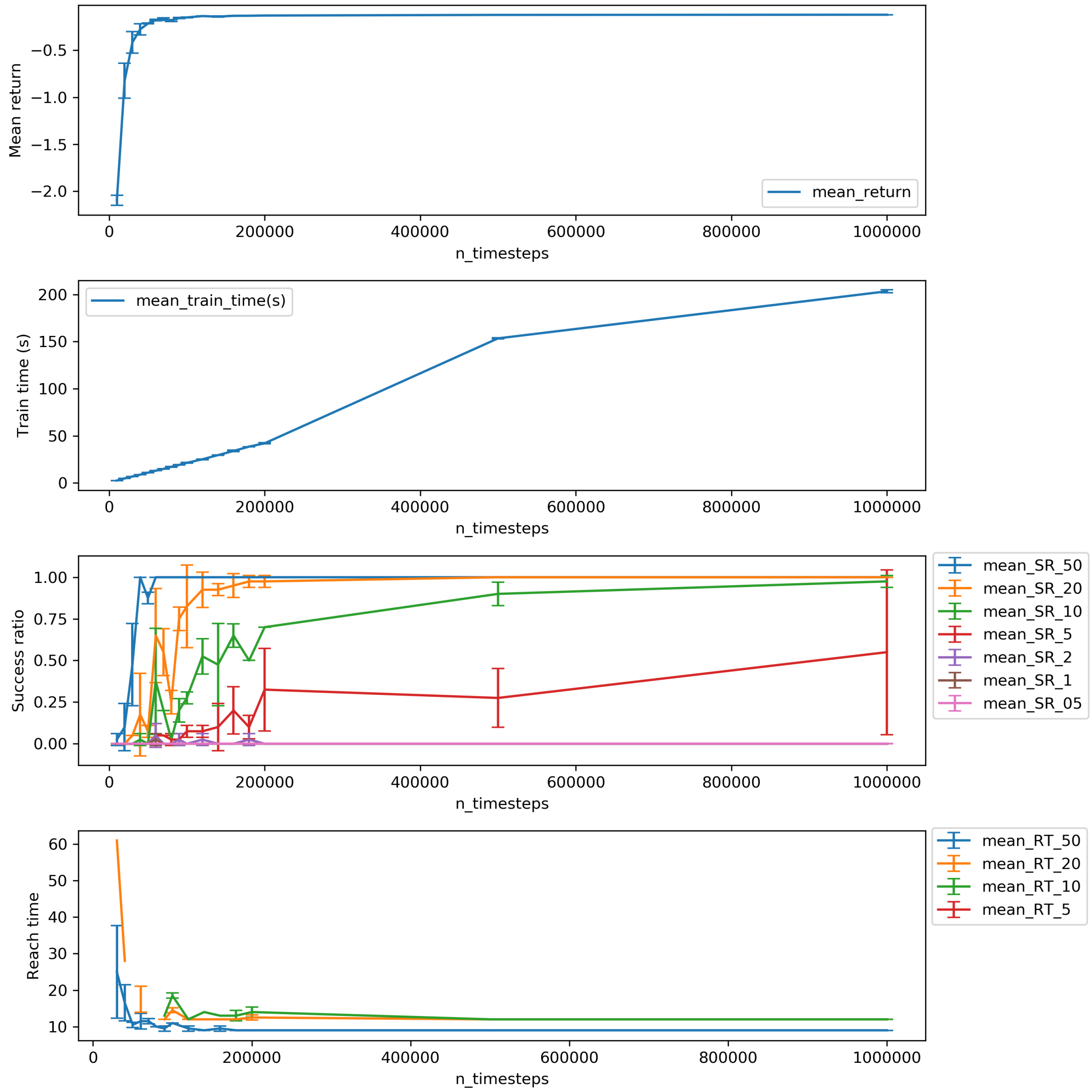

Figure 1: Number of training steps

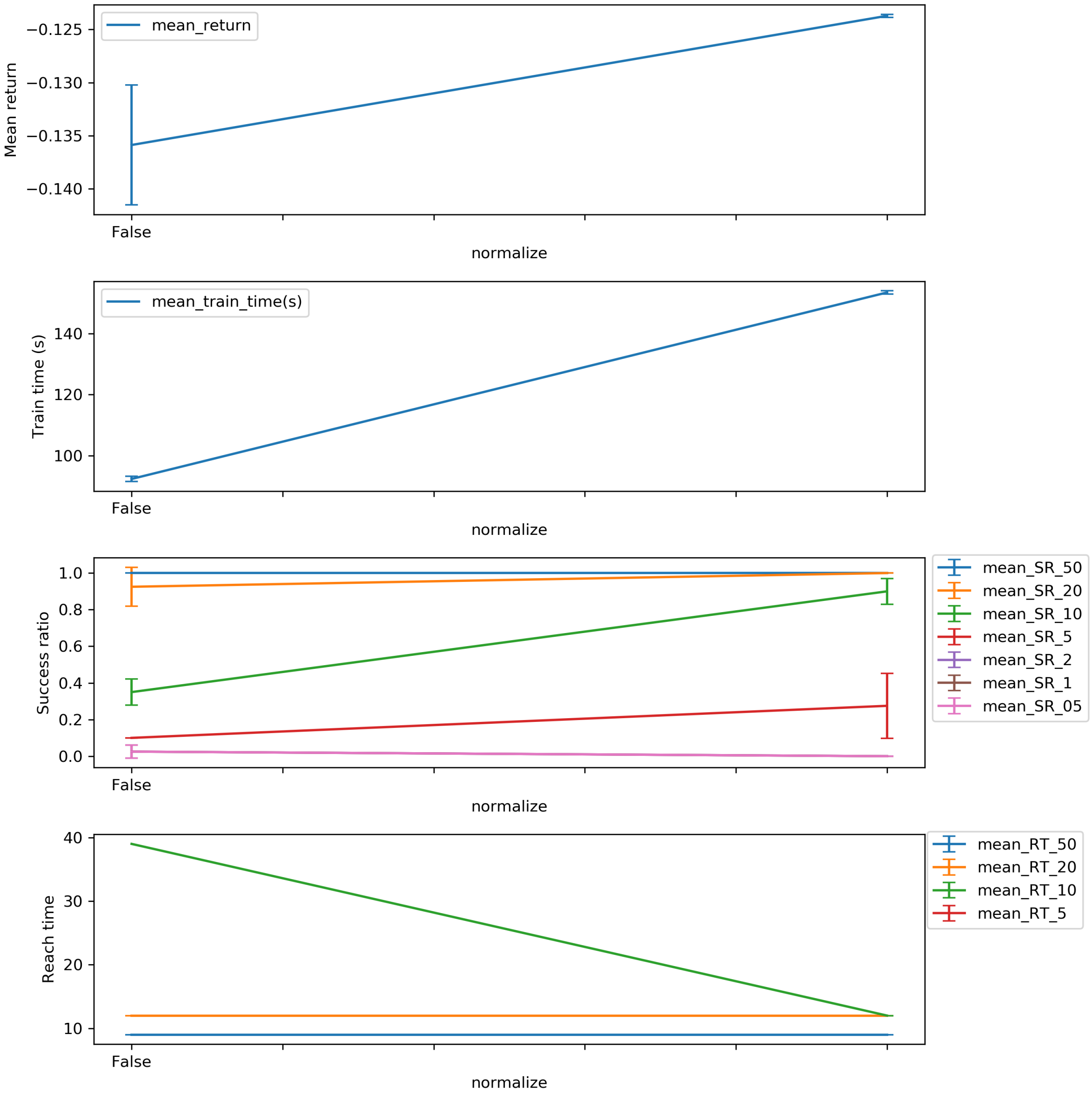

Figure 2: Normalise observation and reward

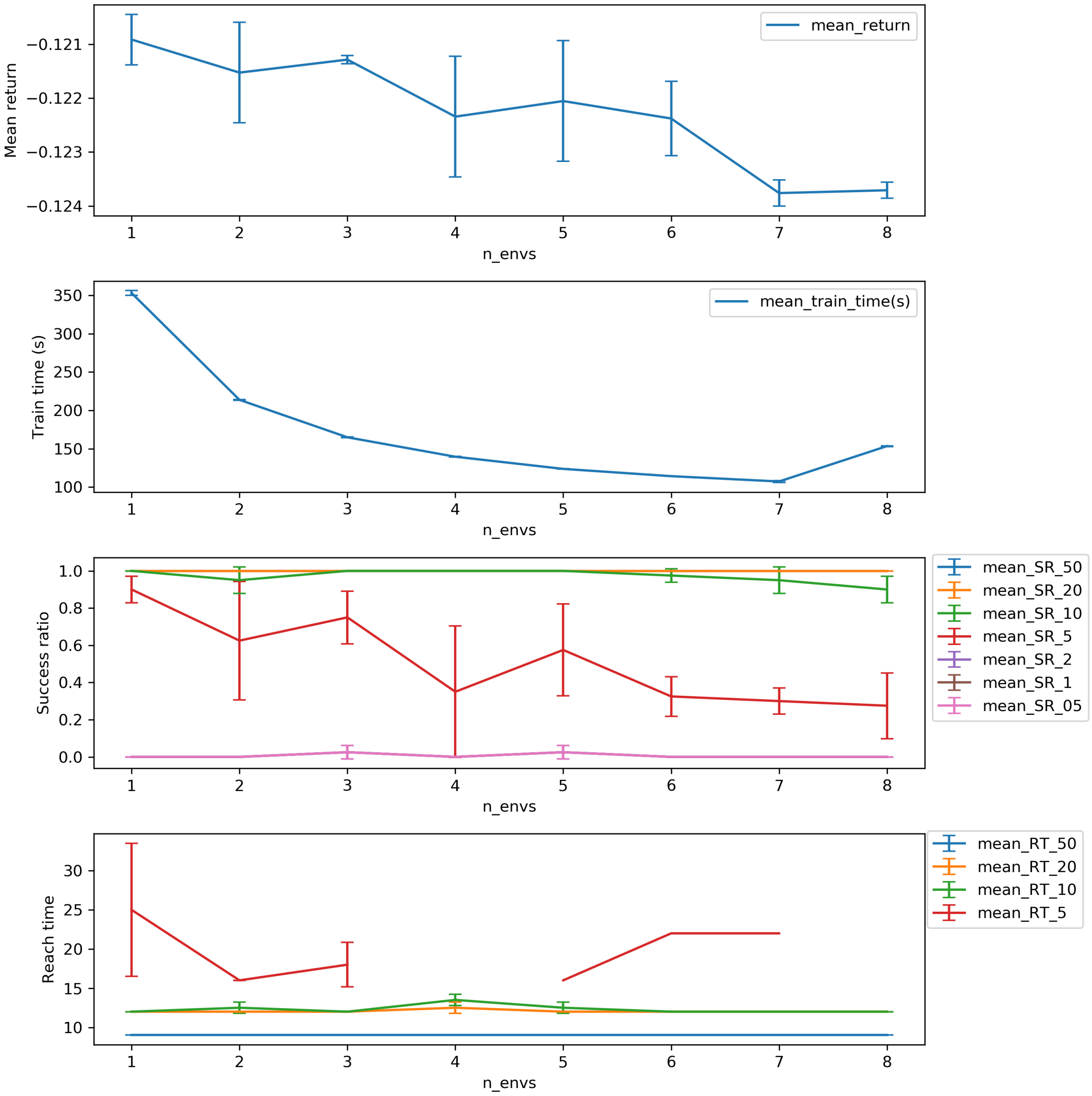

Figure 3: Number of parallel environments

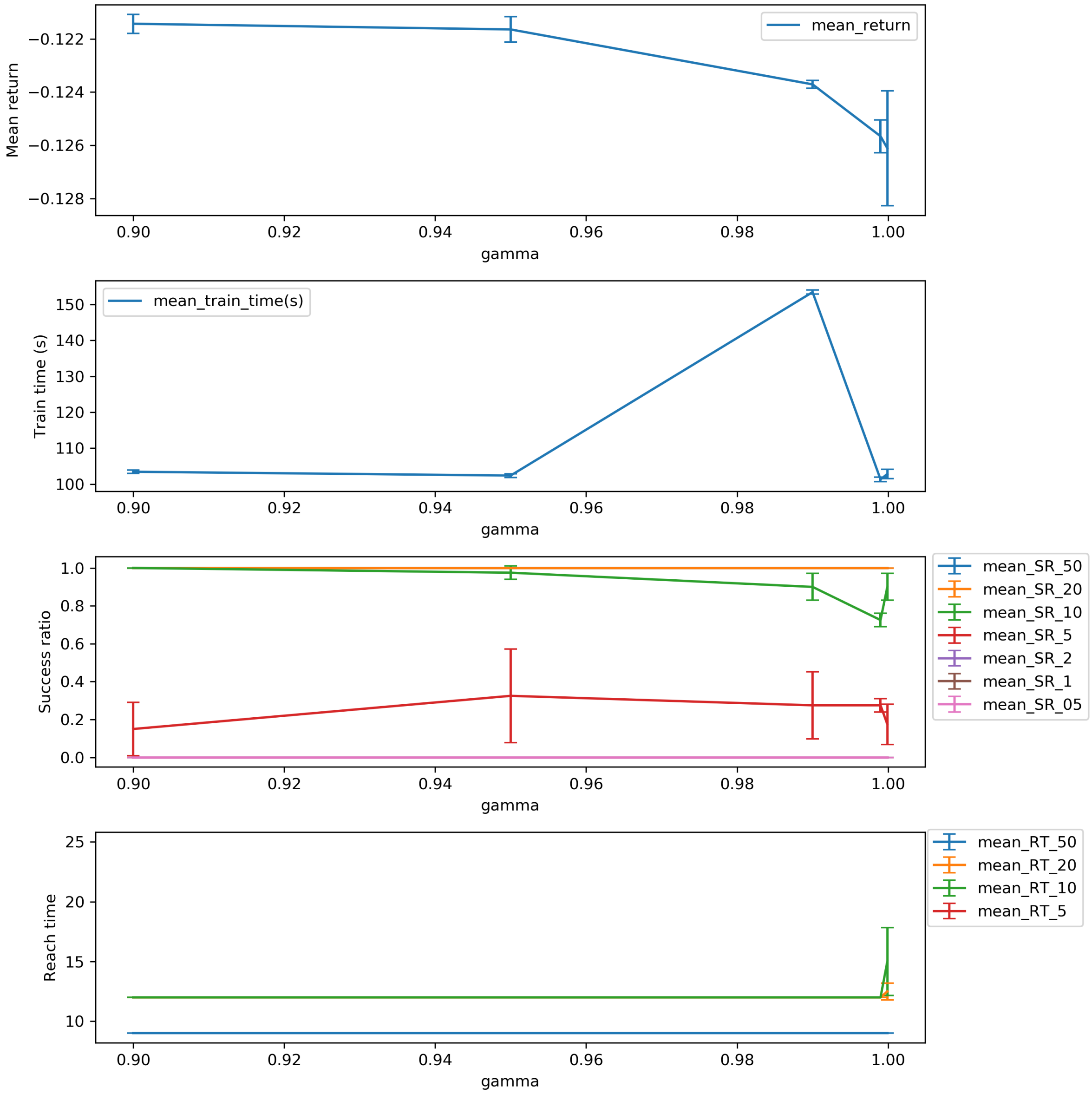

Figure 4: Gamma

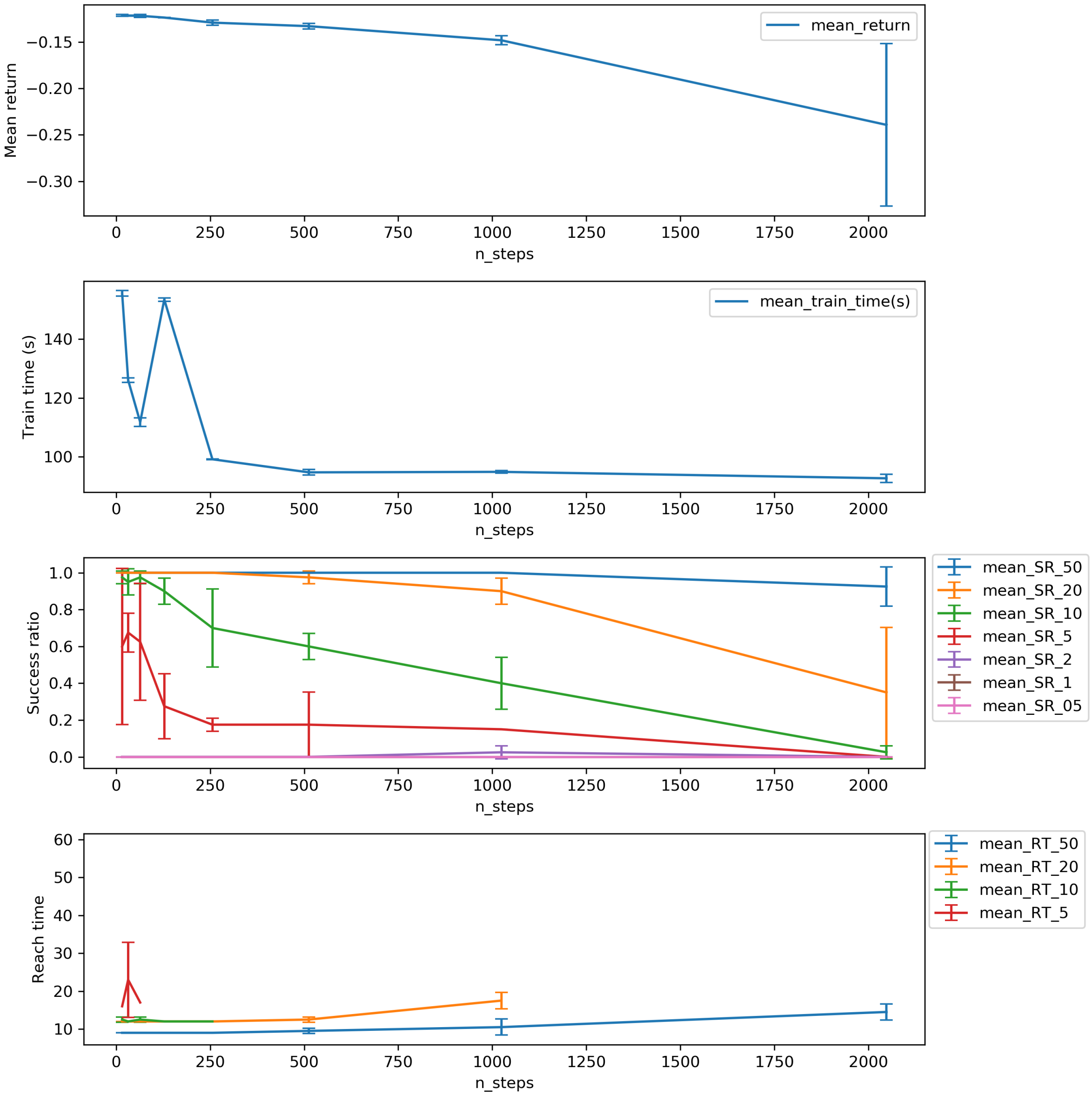

Figure 5: Number of steps to run for each environment per update

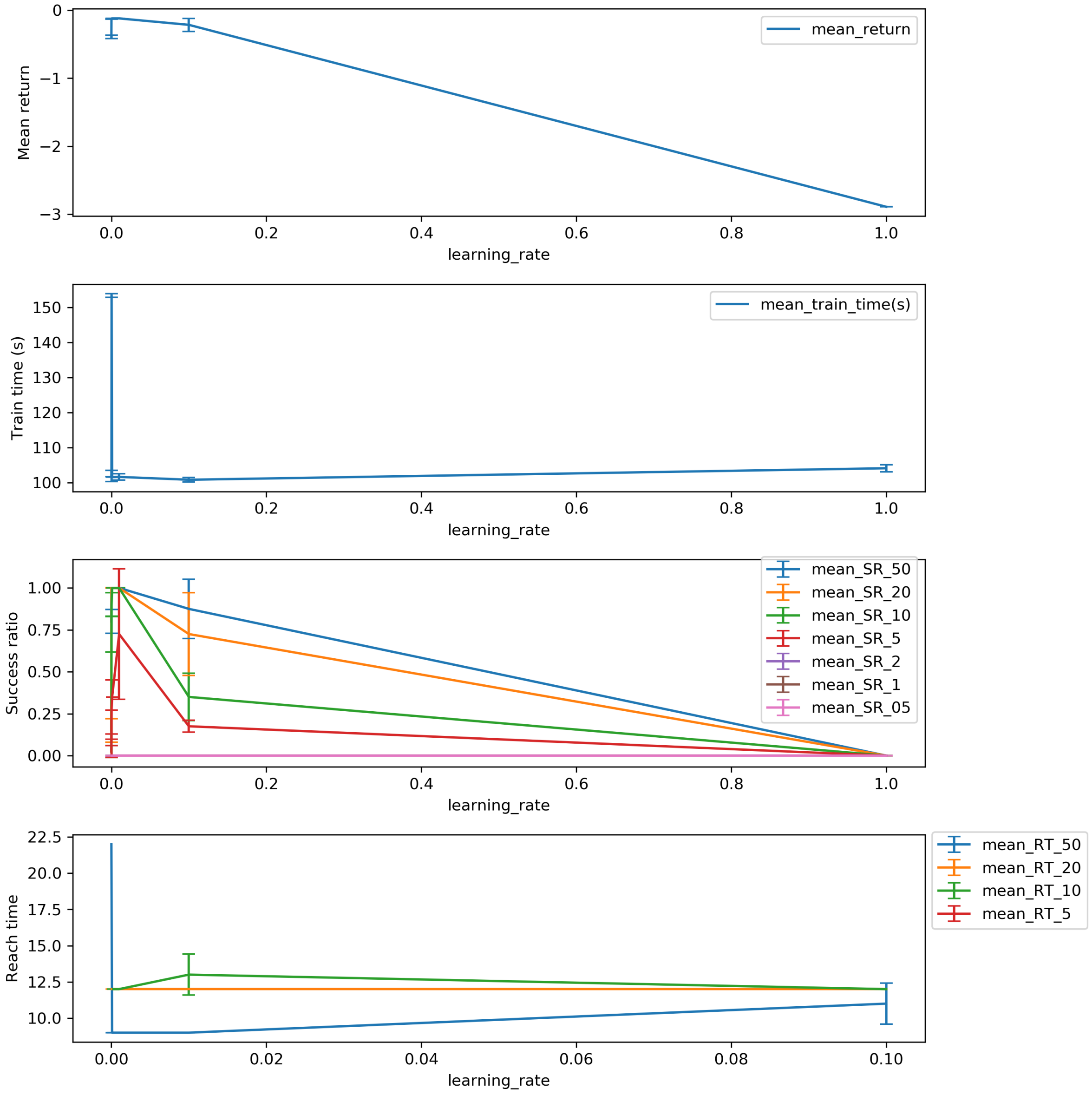

Figure 6: Learning rate

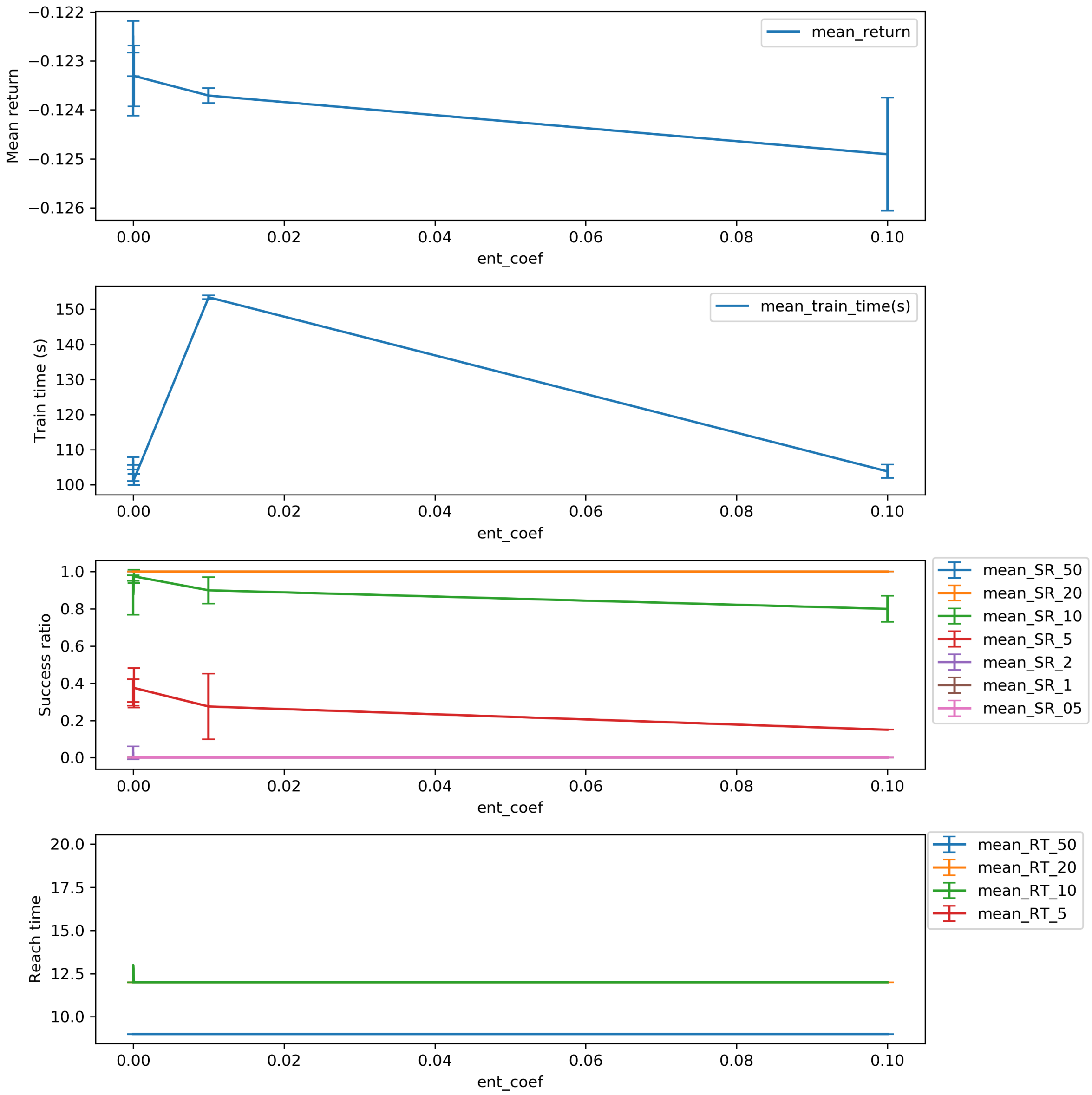

Figure 7: Entropy coefficient for the loss calculation

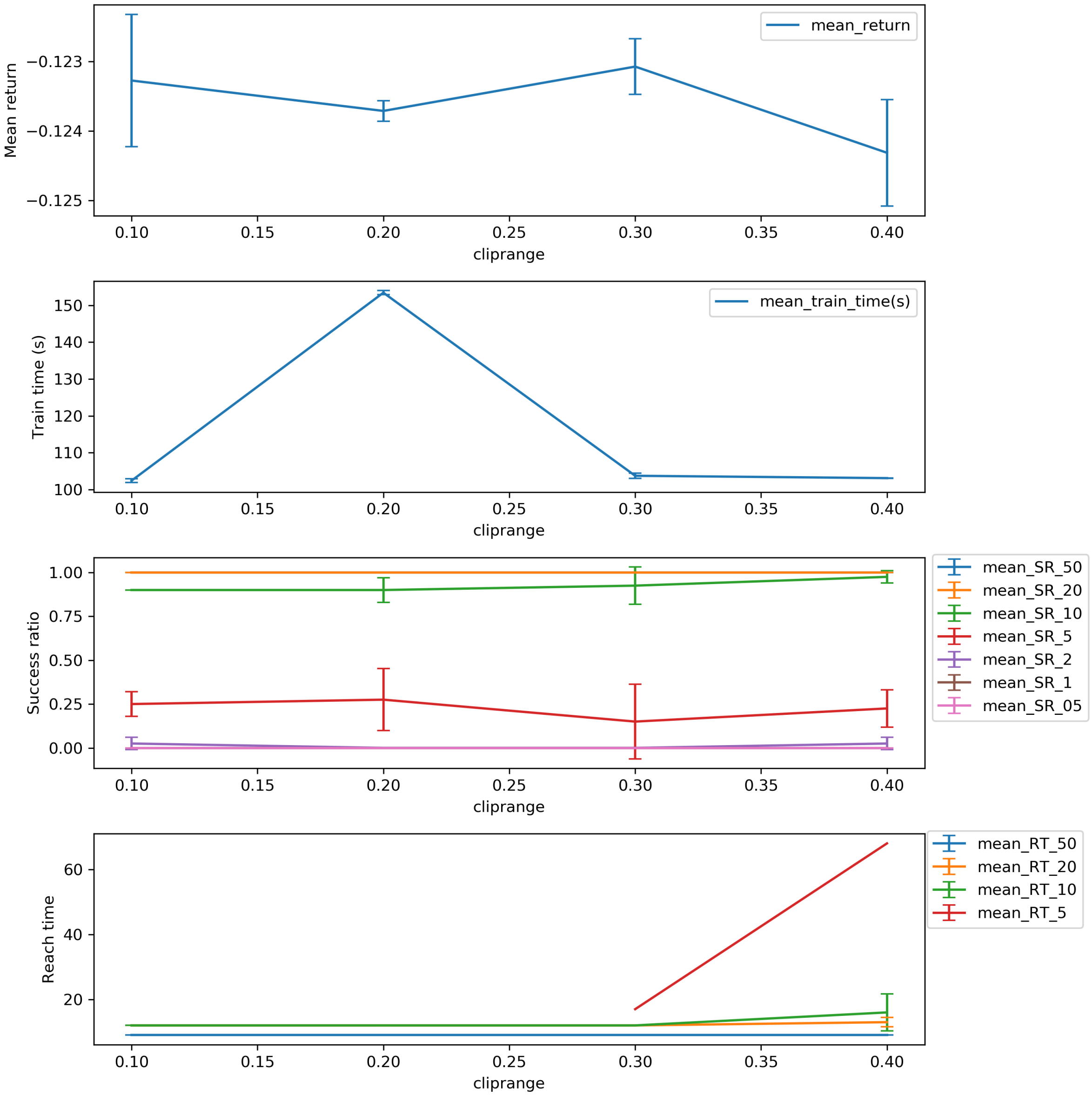

Figure 8: Clipping parameter

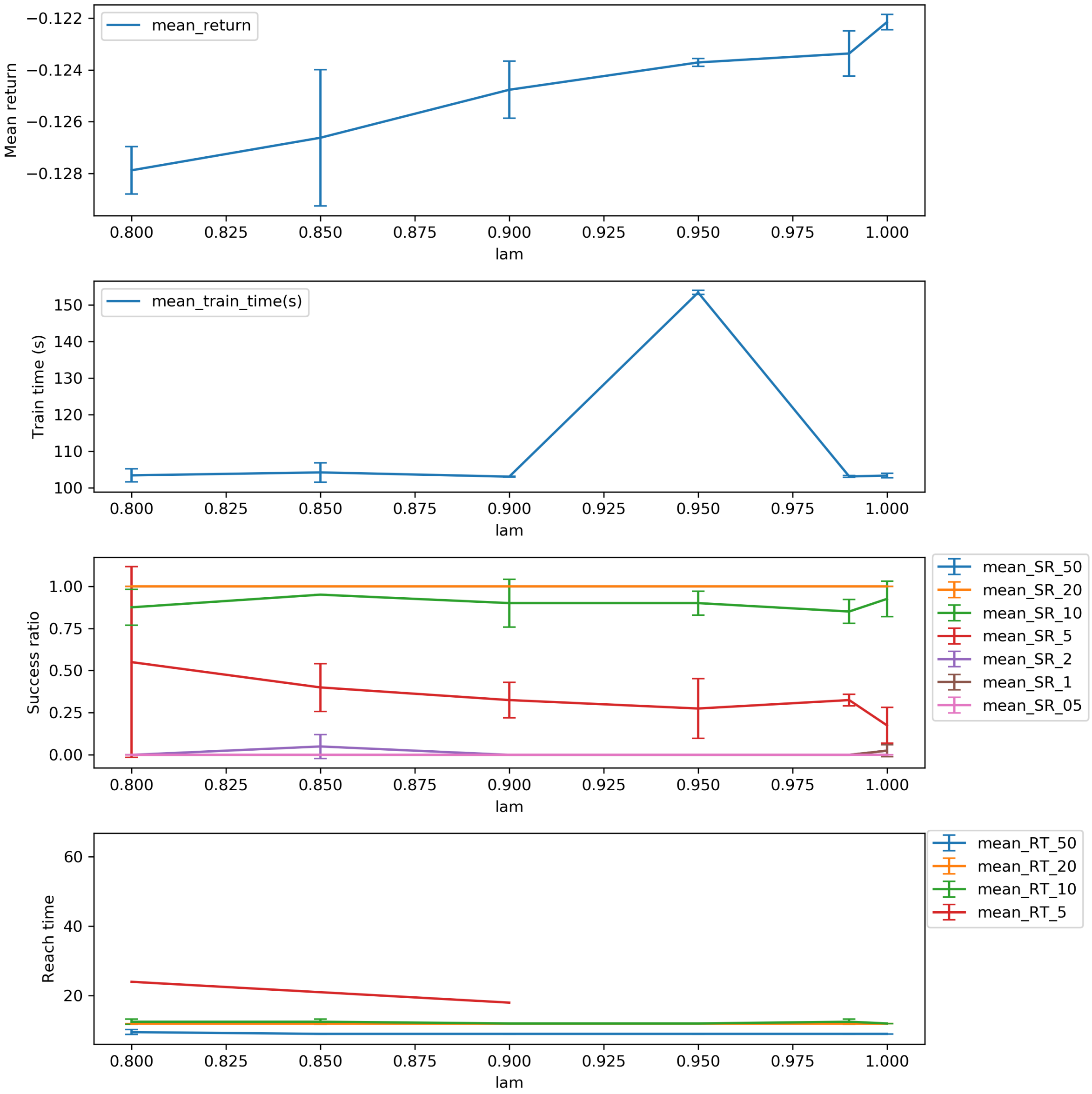

Figure 9: Factor for trade-off of bias vs variance for Generalized Advantage Estimator

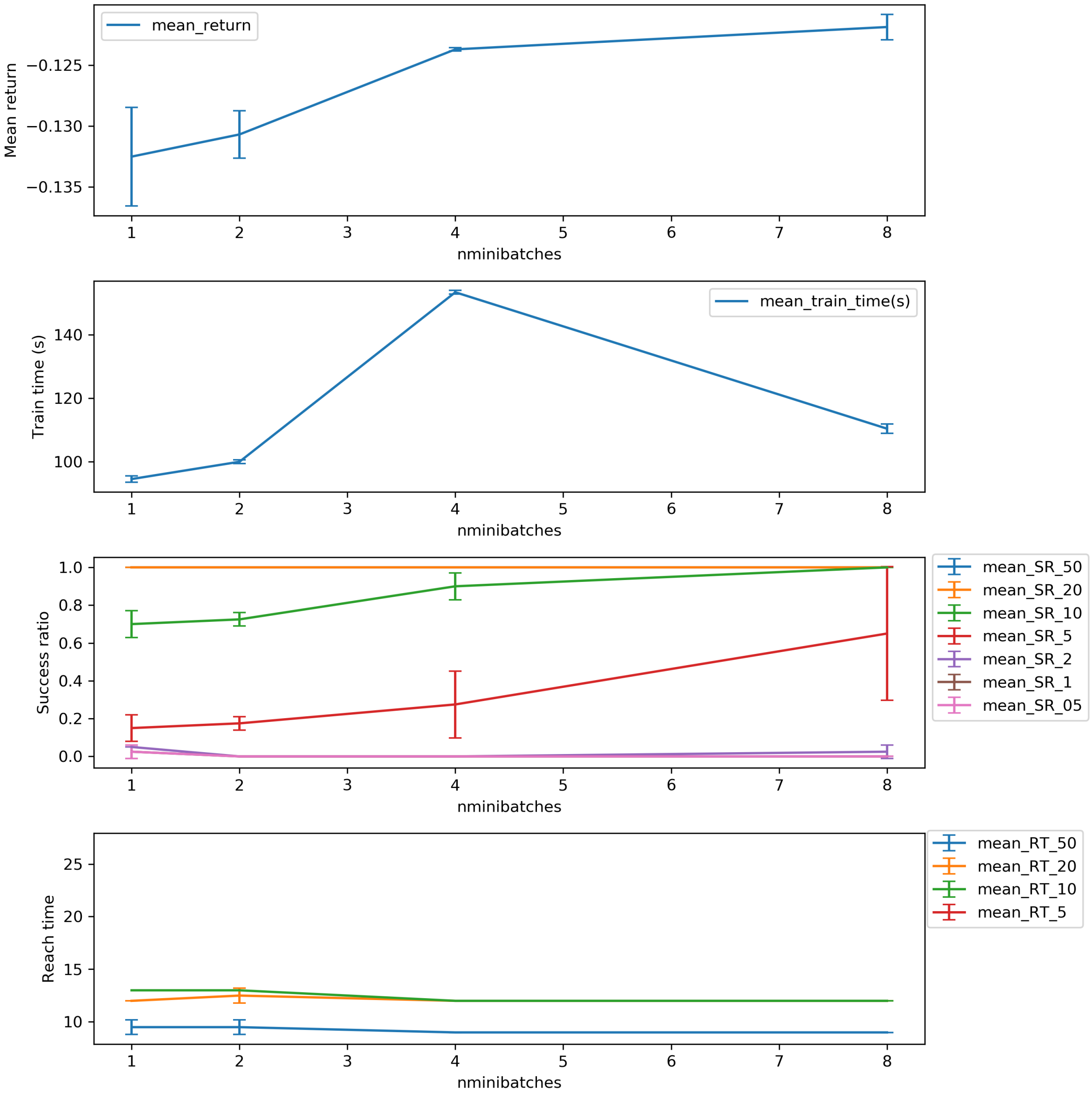

Figure 10: Number of training minibatches per update

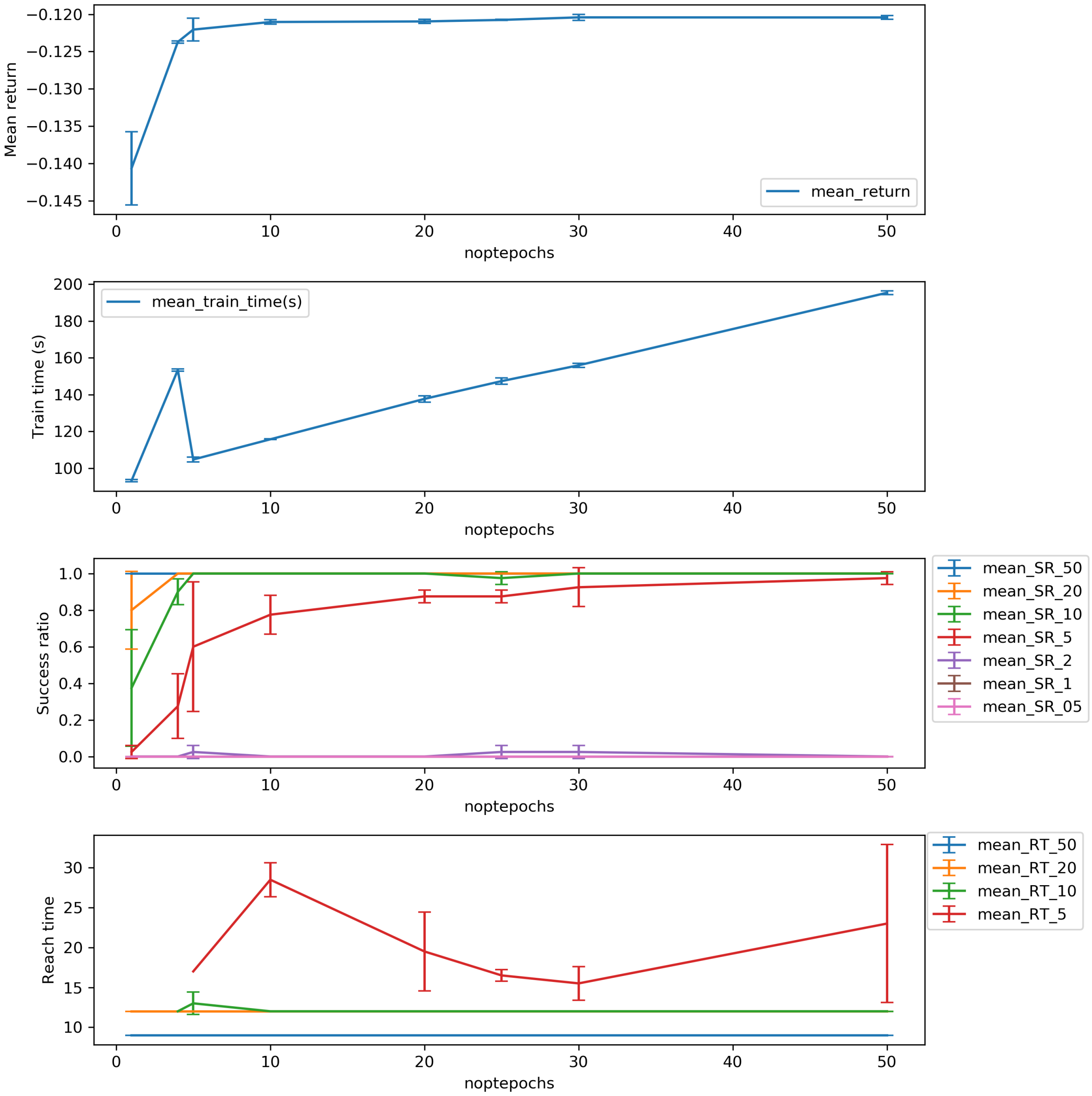

Figure 11: Number of epoch when optimizing the surrogate

3 Findings summary

- 200,000 timesteps are enough for the return to reach a plateau, however 500, 000 timesteps are required to reach the highest success ratio at 5mm. This means that the reward may not describe sufficiently well the objective we want to achieve.
- Best cliprange: 0.2
- Best ent coef: 0.01
- Best gamma: 0.95
- Best lam: 0.95 (note that the best return is not the same as the best success ration @ 5mm)
- Best learning rate: 0.01
- Best nb envs: 1 (but try also 8 since many hyperparams are fitted to this value)
- Best nminibatches: 8

- Best noptepochs: 50
- Best normalize: True
- Best nsteps: 16

These parameters take too long to train. The best trained agent has the following parameters:

- Timesteps: 500, 000
- cliprange: 0.2
- ent coef: 0.01
- gamma: 0.99
- lam: 0.95
- learning rate: 0.00025
- nb envs: 8
- nminibatches: 4
- noptepochs: 50
- normalize: True
- nsteps: 128