

RL benchmark - WidowX

1 Compare RL algorithms

We train the WidowX arm with fixed goal with the different RL algorithm and their default hyperparameters.

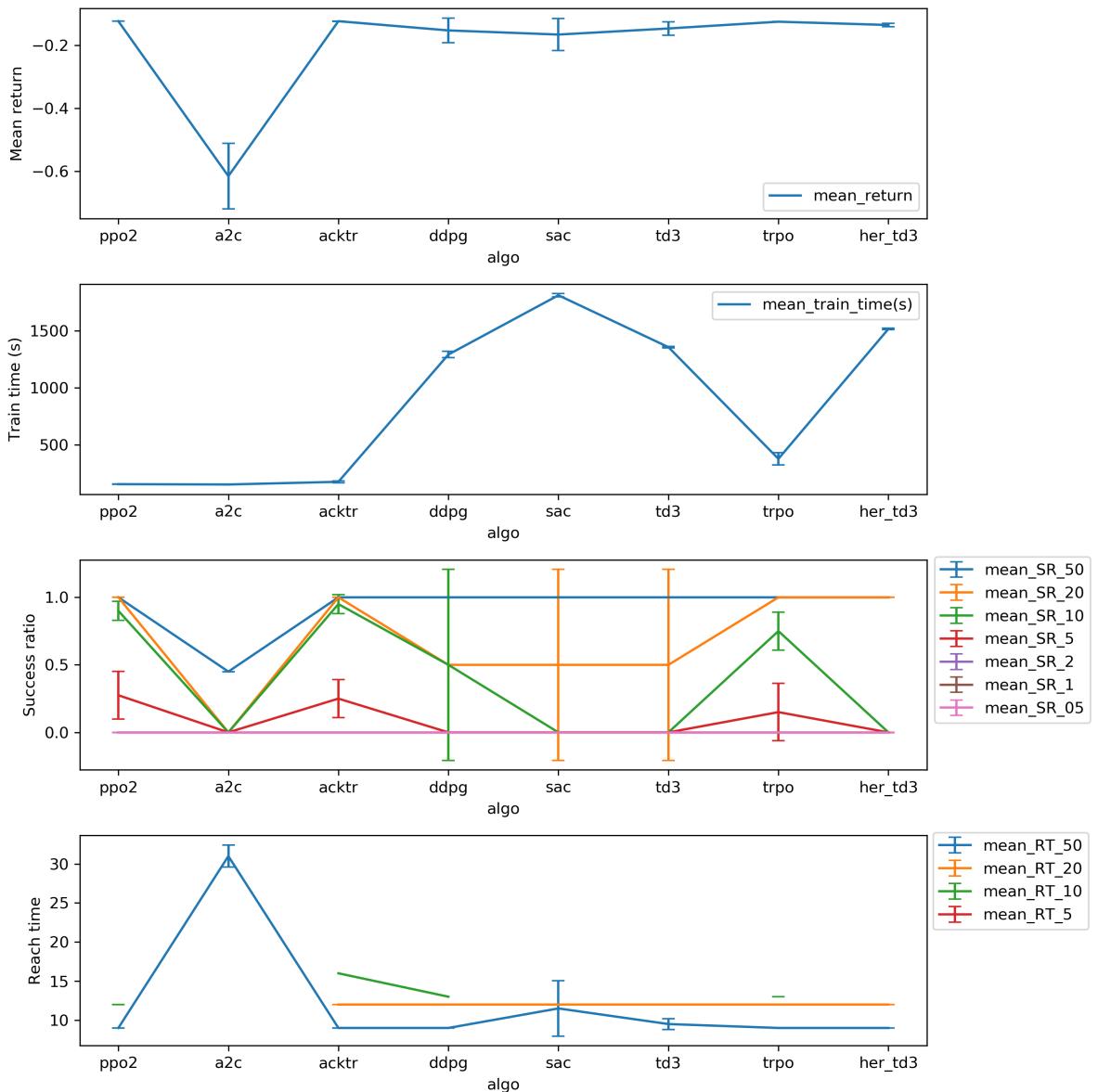


Figure 1: Algorithms applied to widowx-reacher-v5

PPO2 gives the best performance and train time.

2 Manual hyperparameter tuning

We train with PPO2 and change the hyperparameters. The training environment is:

- Environment: widowx-reacher-v5
- 6 joints
- Fixed goal
- Dense reward: $-\text{dist}^{**2}$

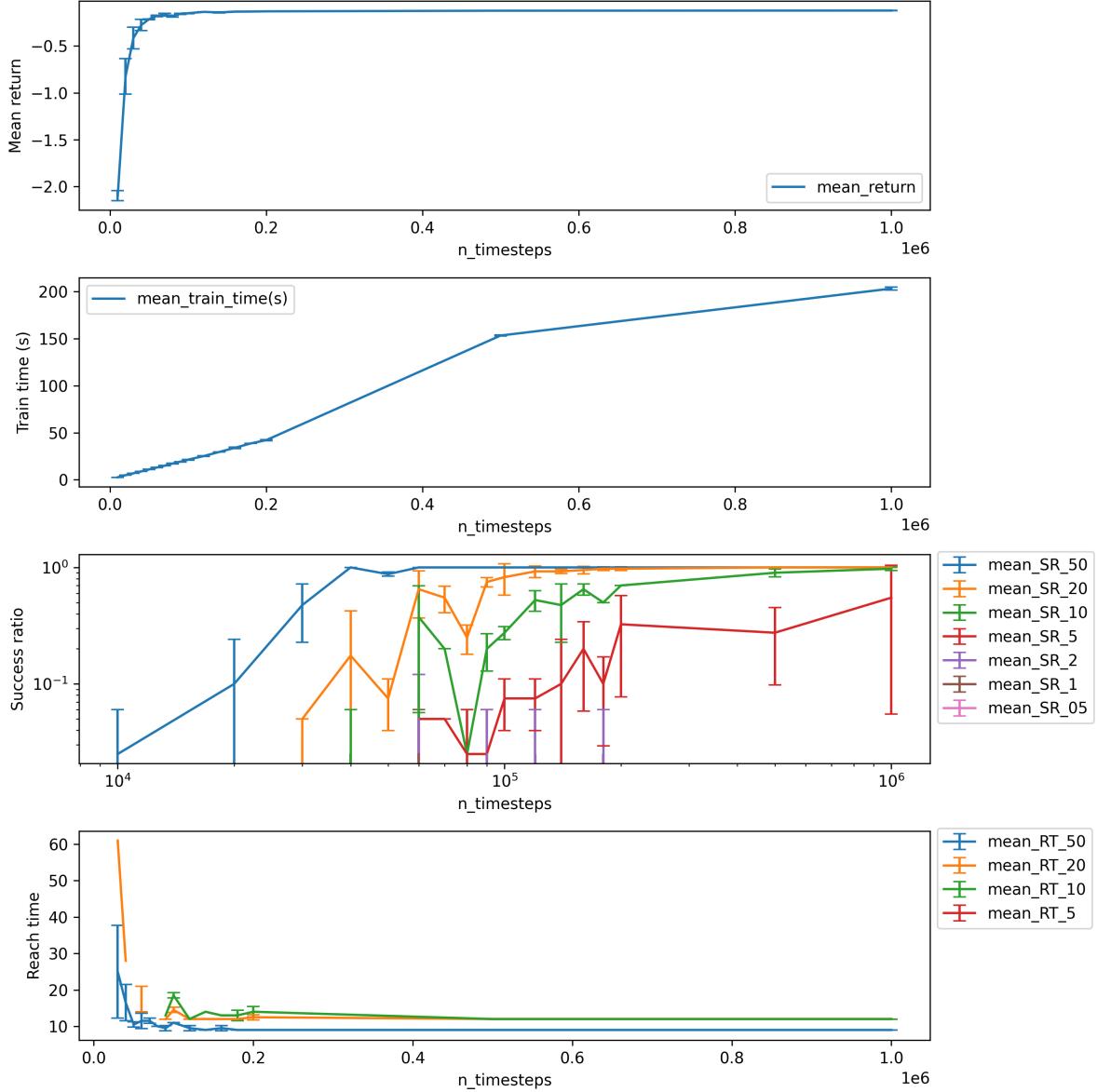


Figure 2: Number of training steps

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.

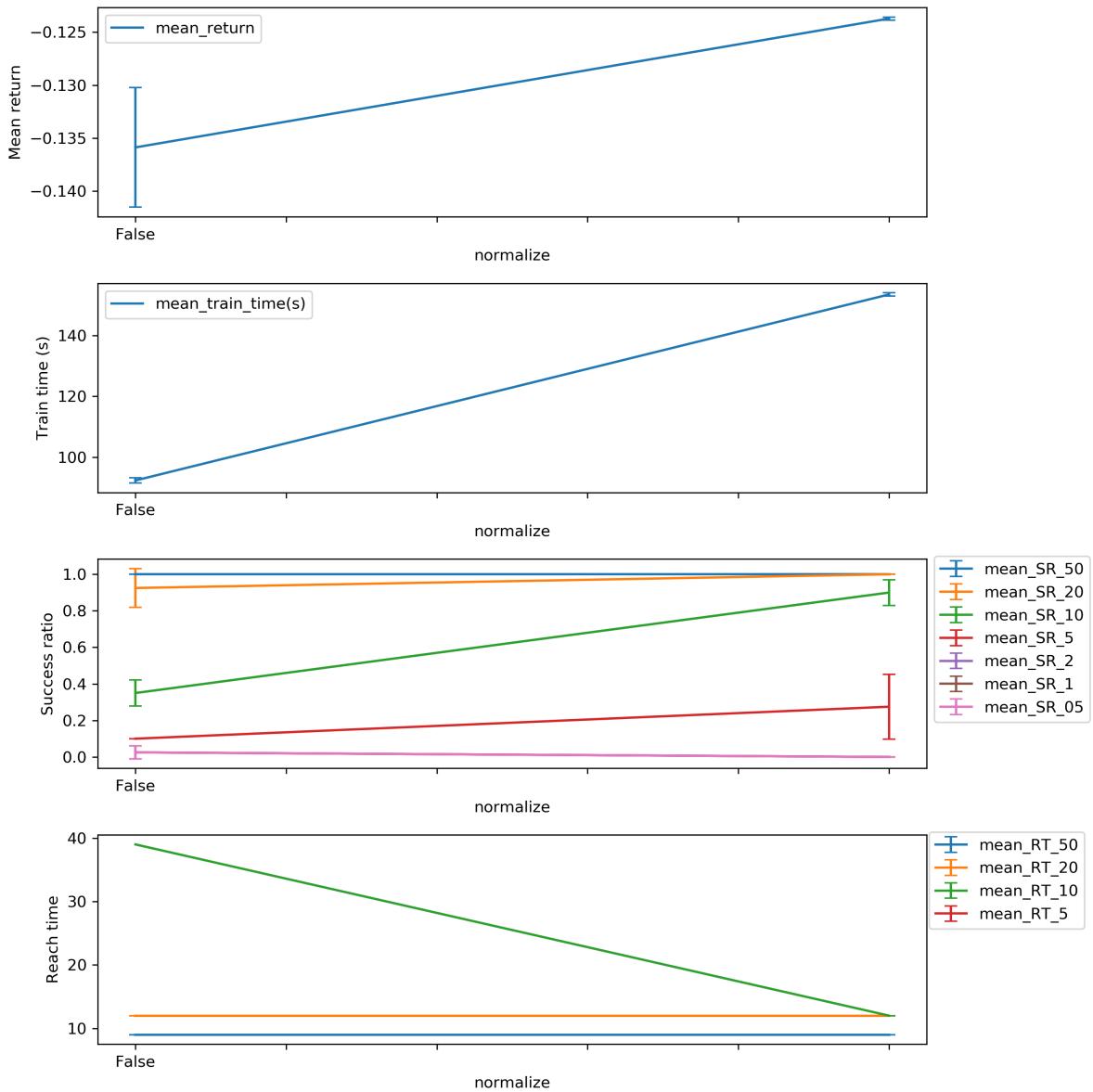


Figure 3: Normalise observation and reward

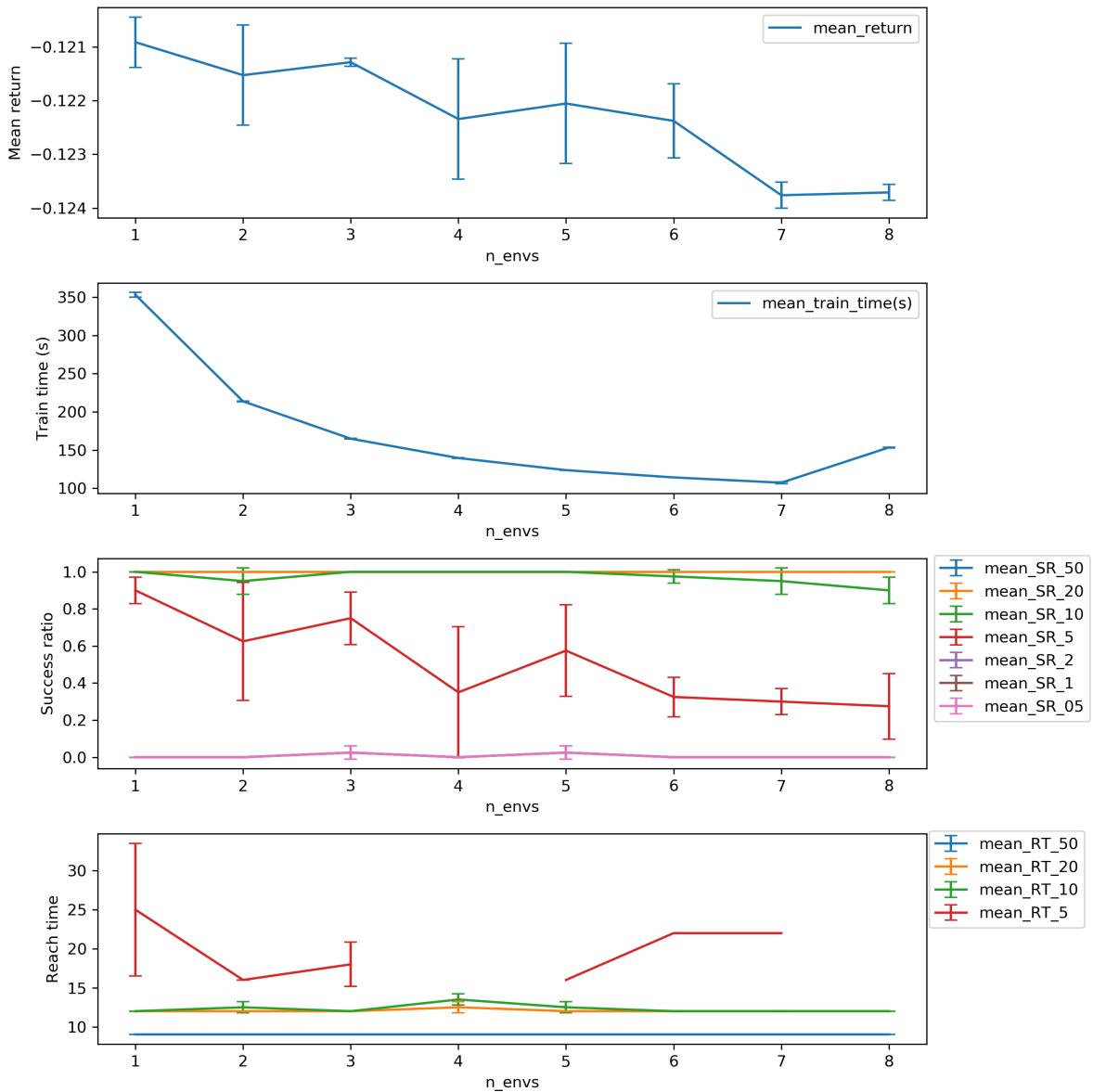


Figure 4: Number of parallel environments

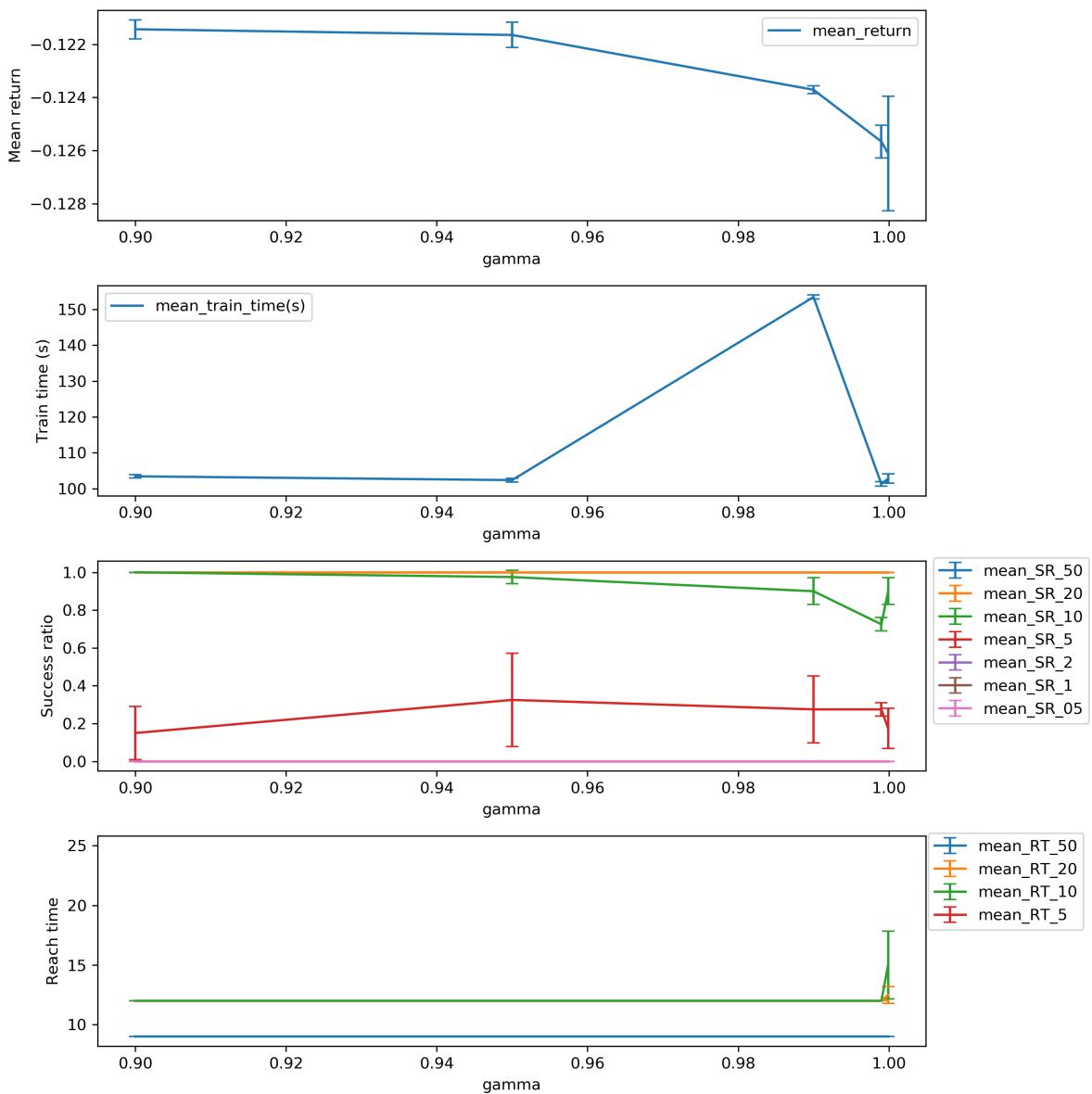


Figure 5: Gamma

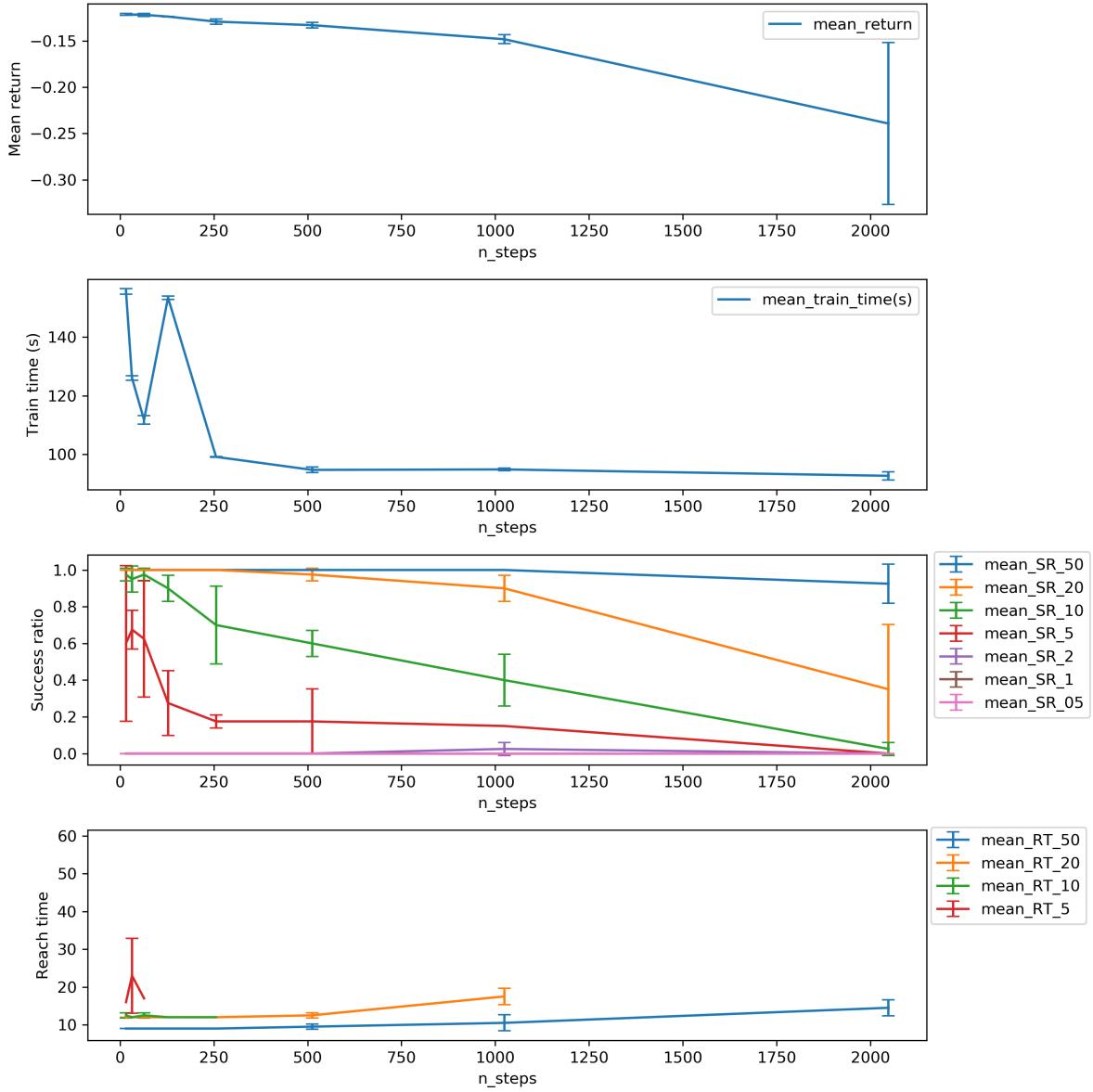


Figure 6: Number of steps to run for each environment per update

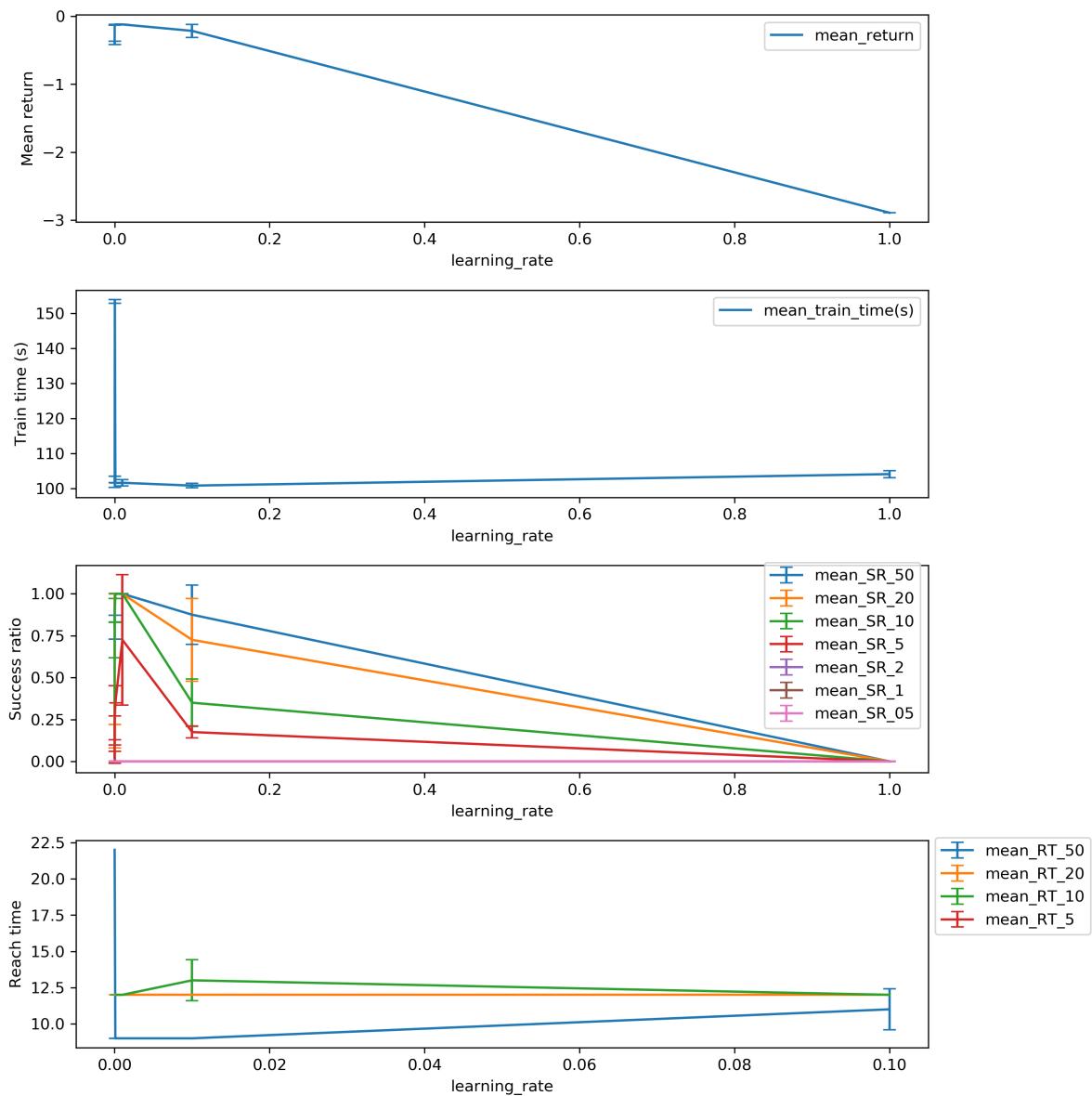


Figure 7: Learning rate

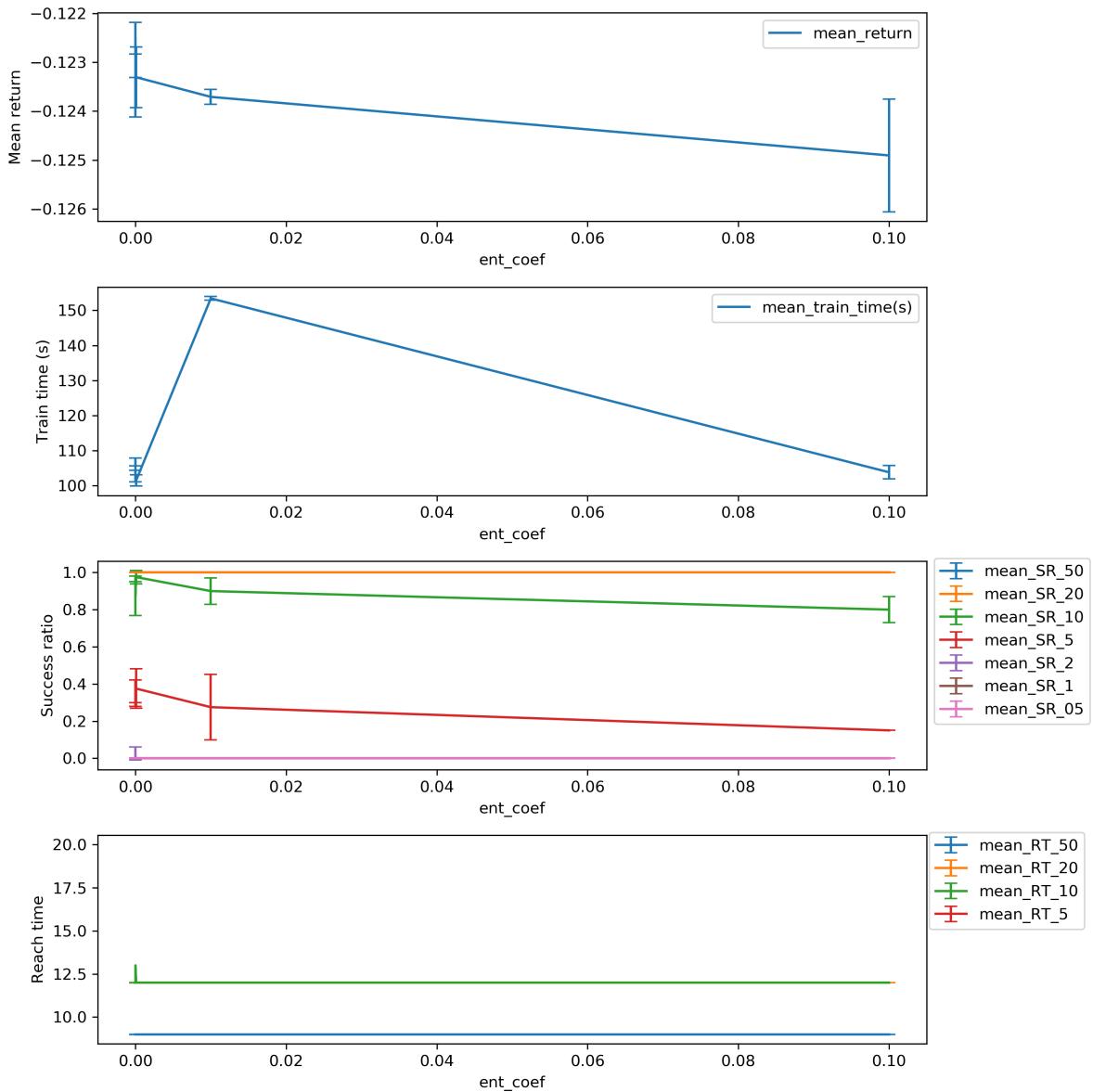


Figure 8: Entropy coefficient for the loss calculation

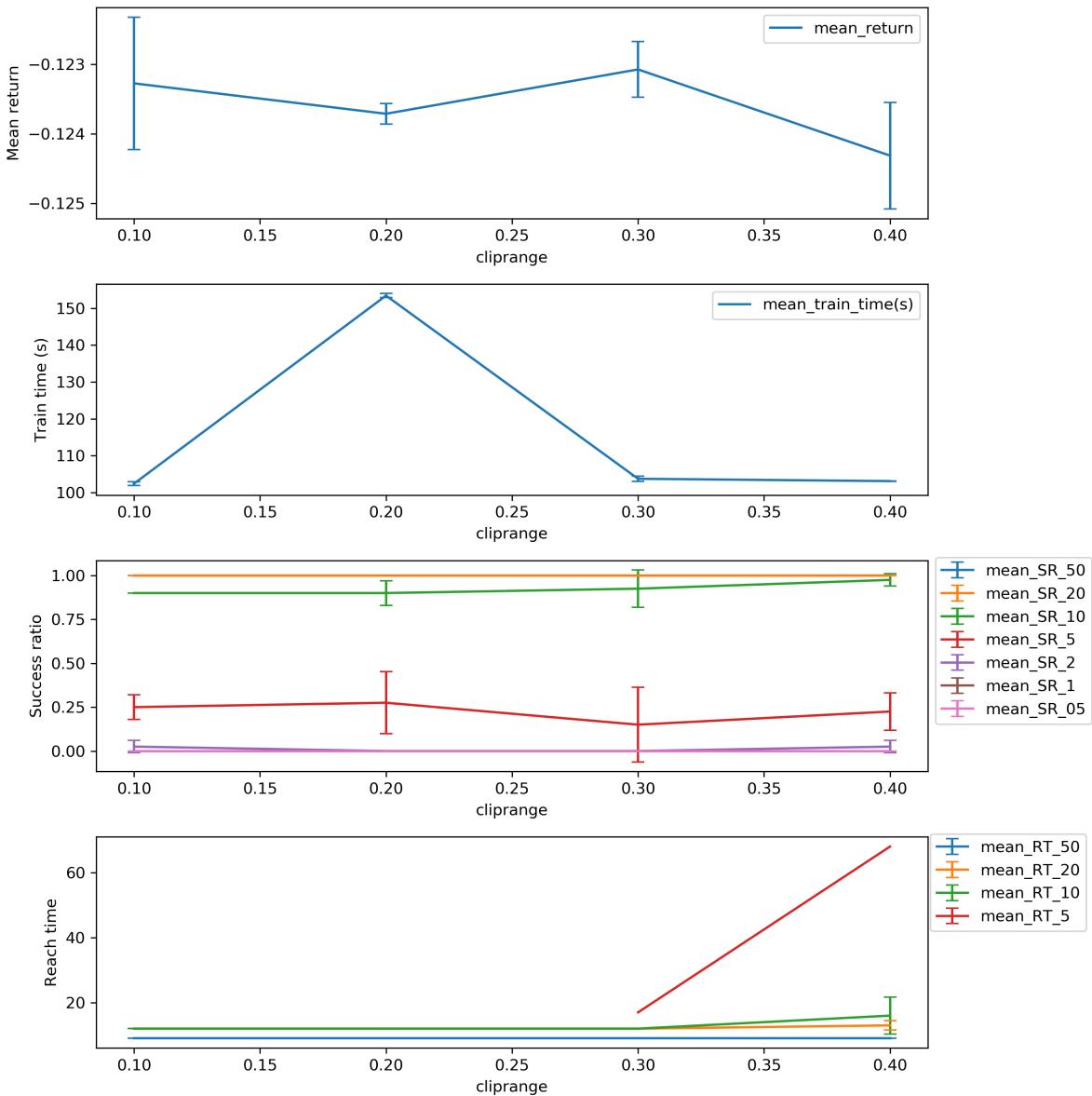


Figure 9: Clipping parameter

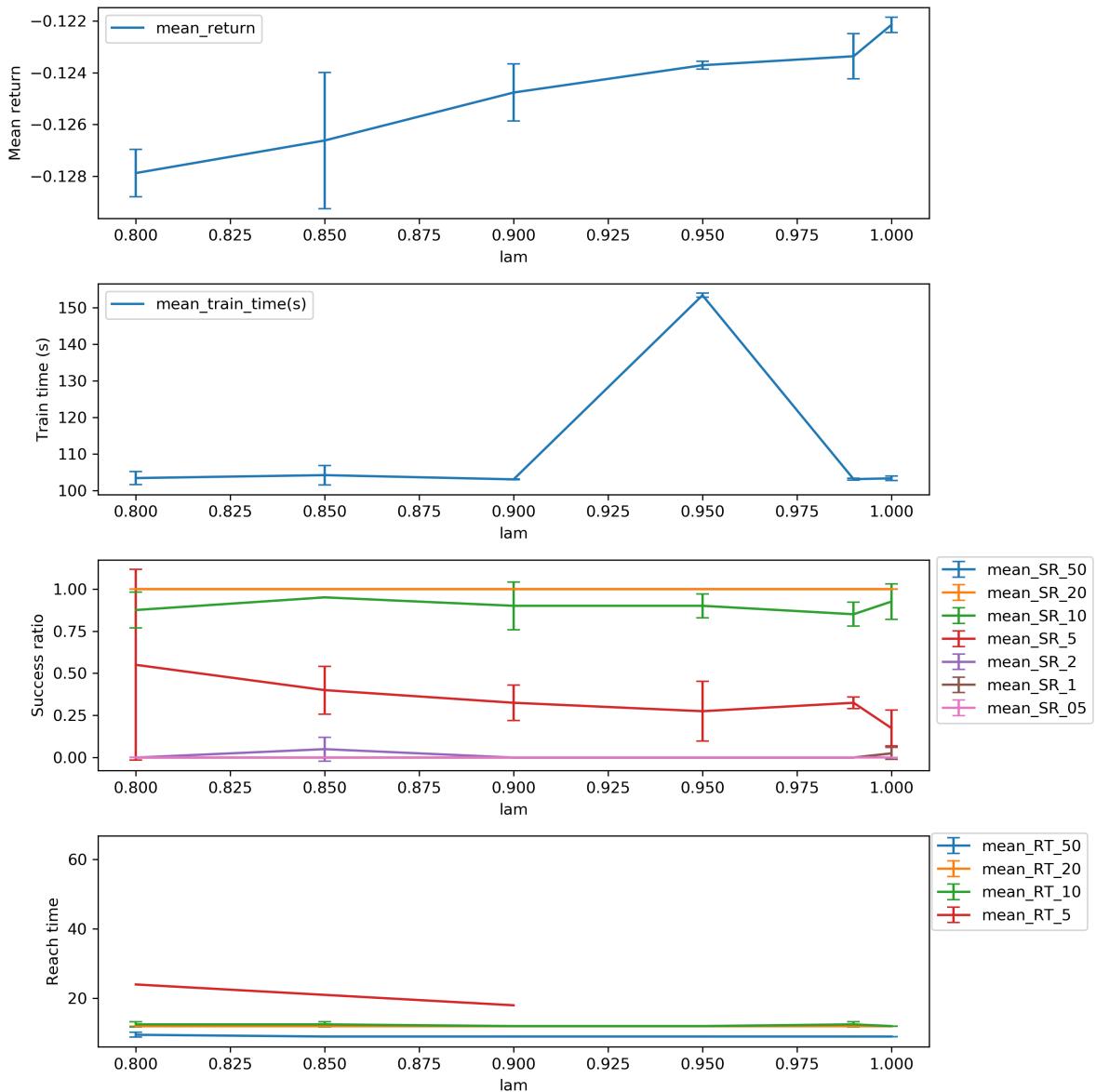


Figure 10: Factor for trade-off of bias vs variance for Generalized Advantage Estimator

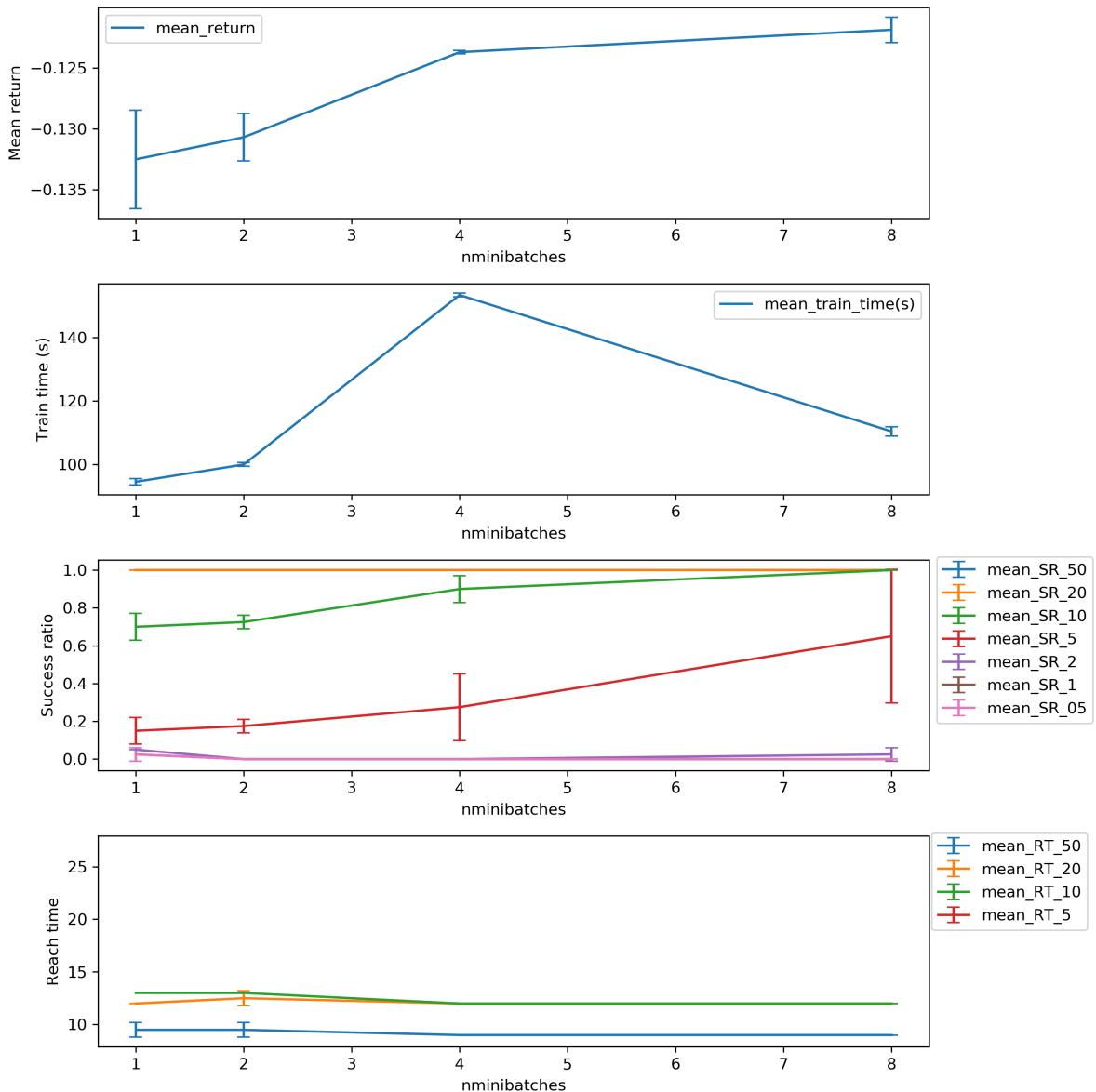


Figure 11: Number of training minibatches per update

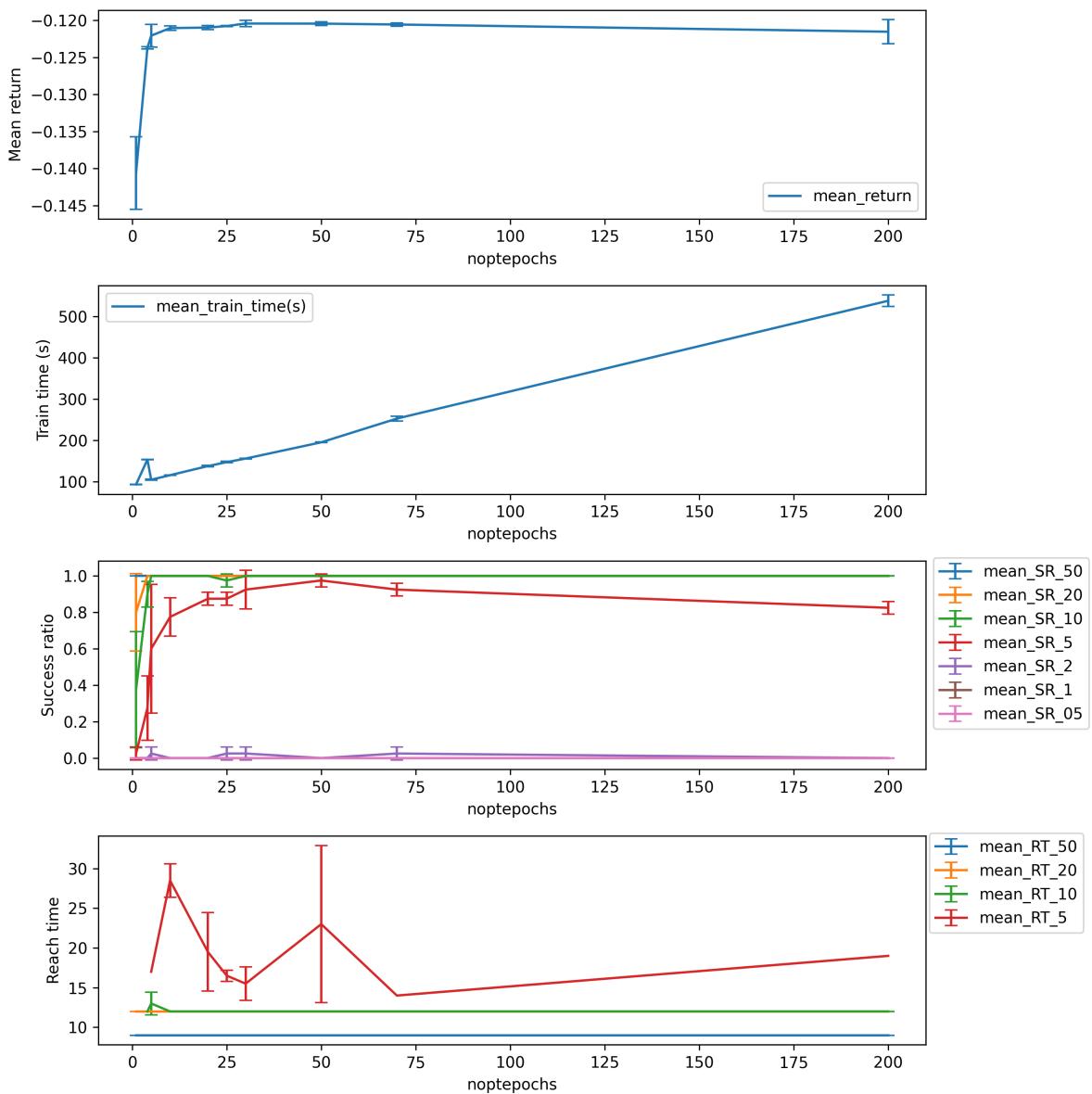


Figure 12: Number of epoch when optimizing the surrogate

3 Environment tuning

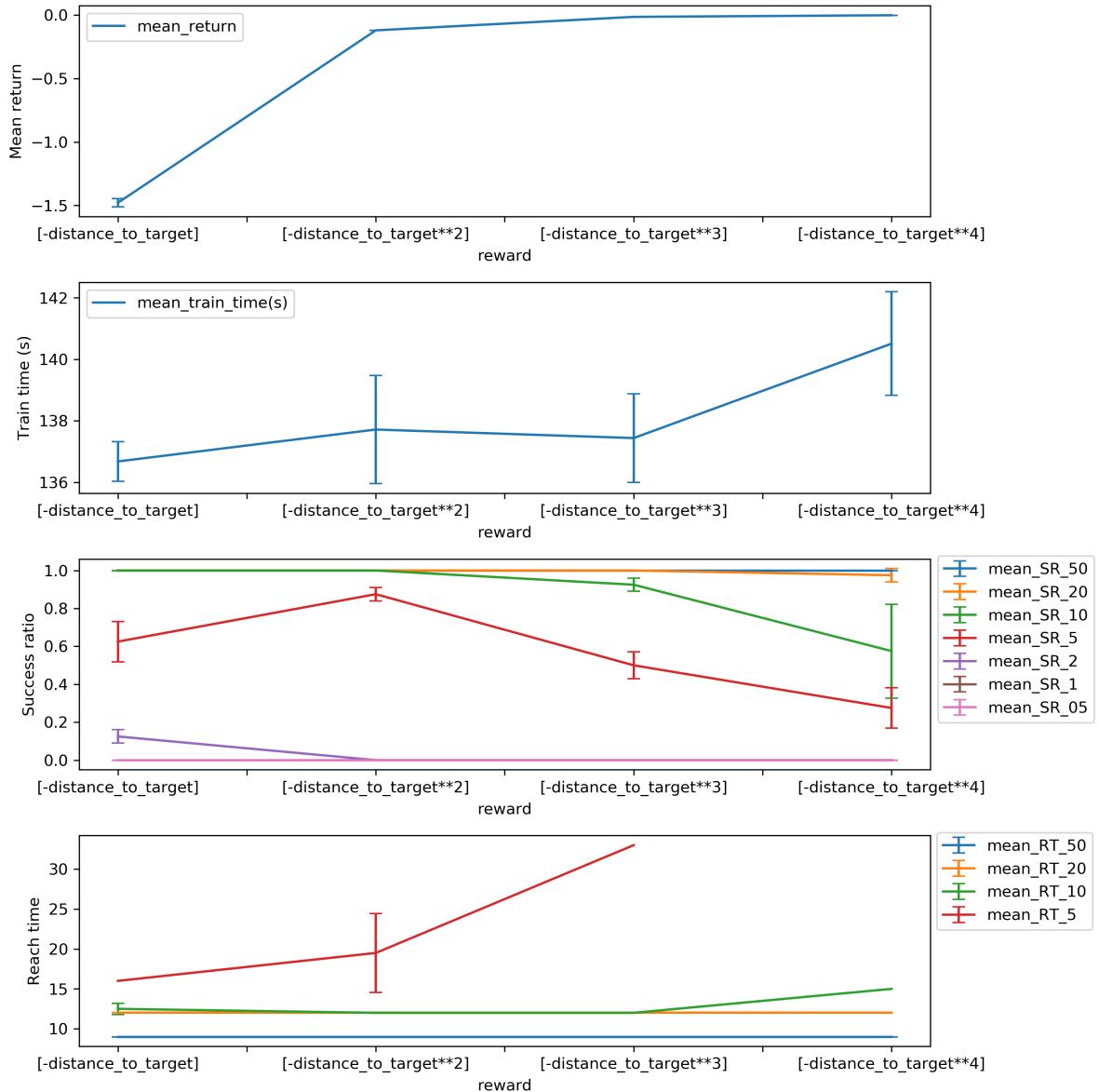


Figure 13: Metrics vs reward shape (distance).

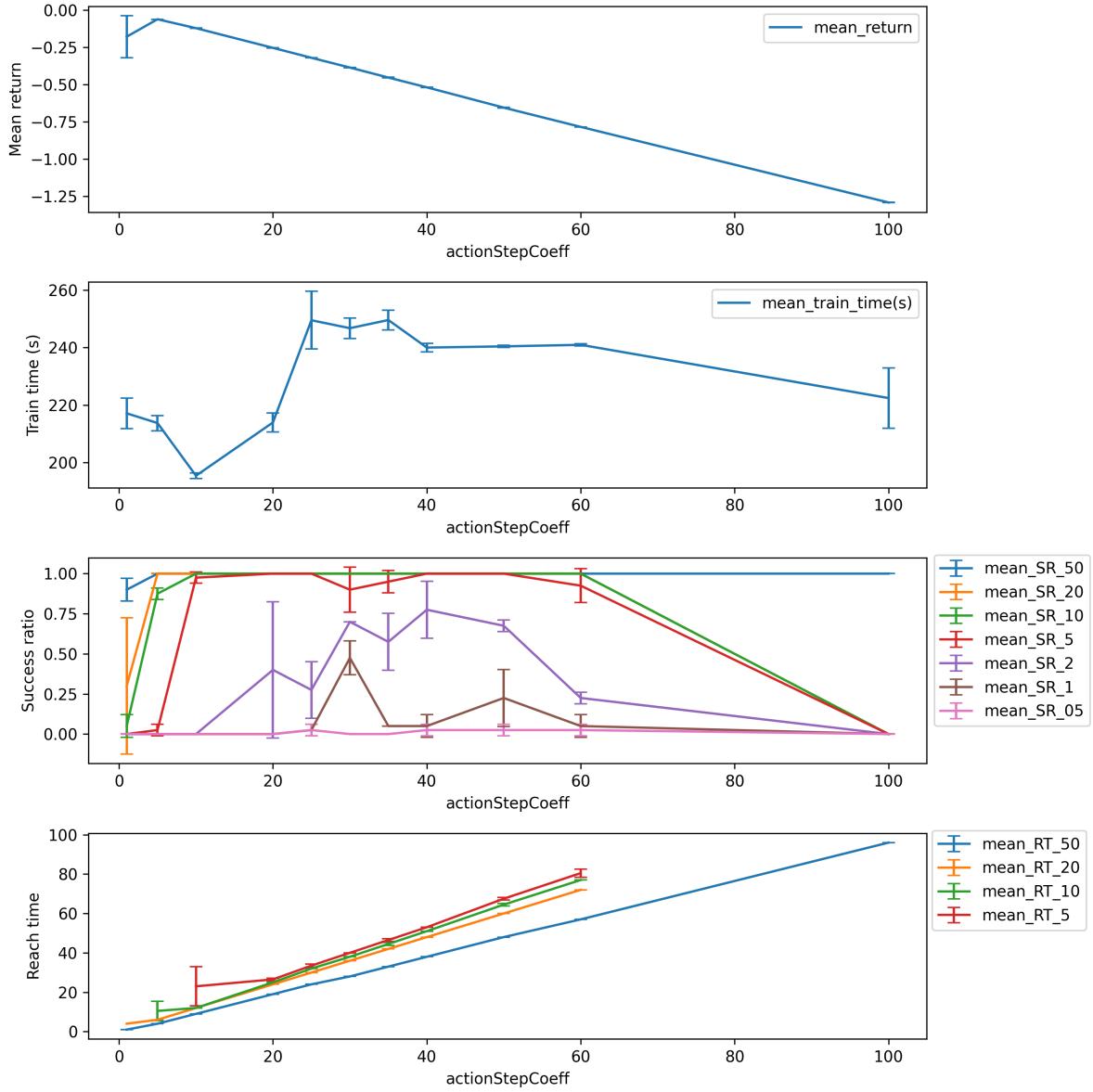


Figure 14: Metrics vs actionStepCoeff (higher = smaller steps).

Best success ratio @1mm achieved: 0.55!

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
- actionStepCoeff: 30

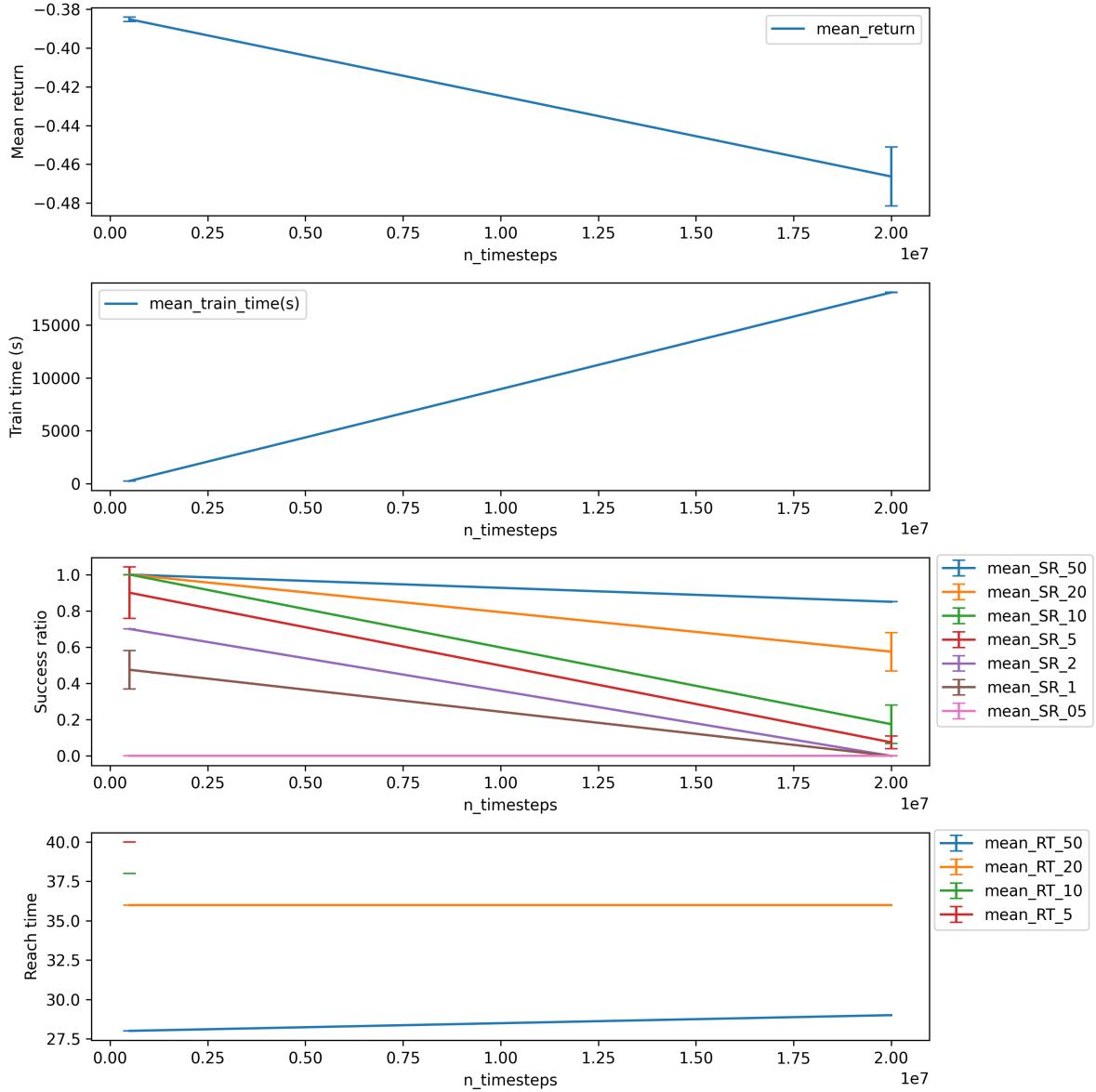


Figure 15: Metrics vs timesteps for env-v22 (actionStepCoeff=30, noptepoch = 50).

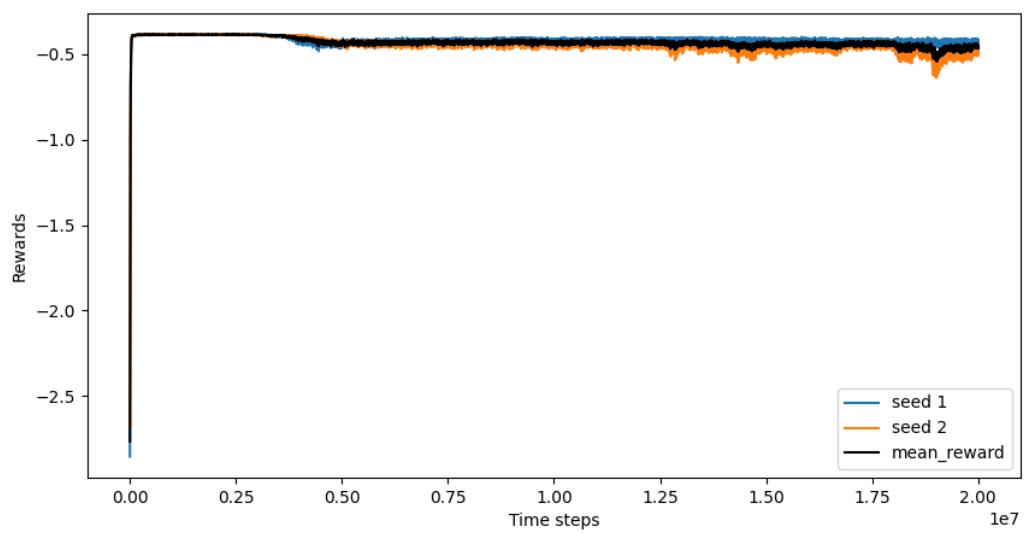


Figure 16: Reward vs timesteps for Env-v22 after 20M timesteps.

The training becomes unstable after 3M timesteps.