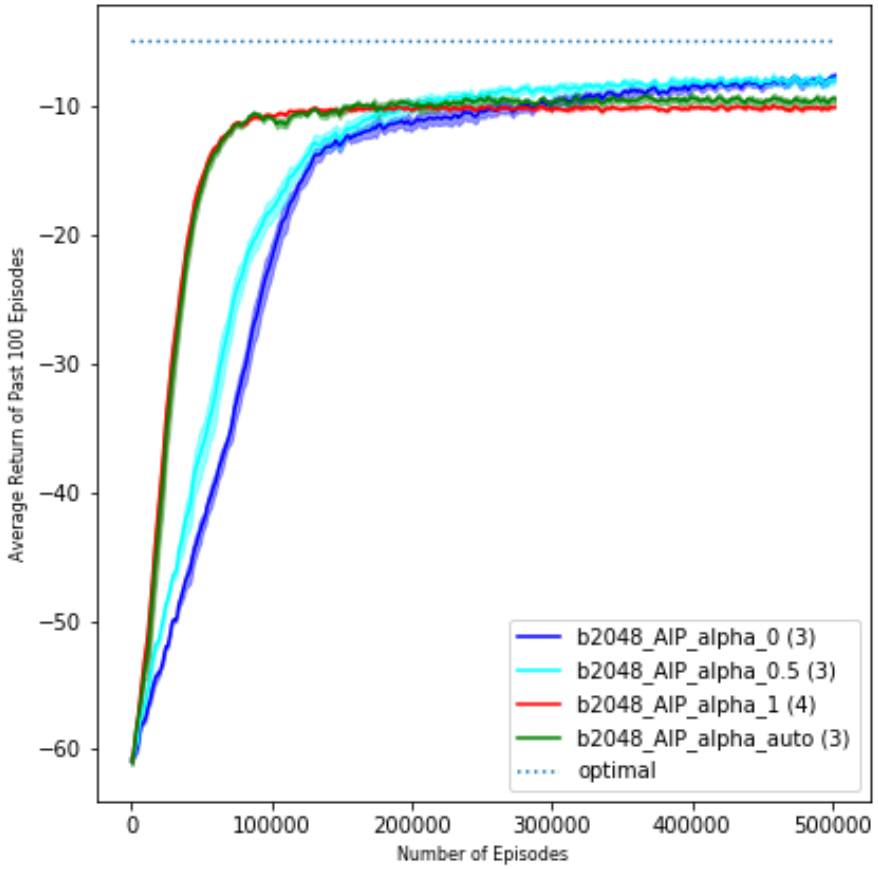
Meeting 2021/08/11

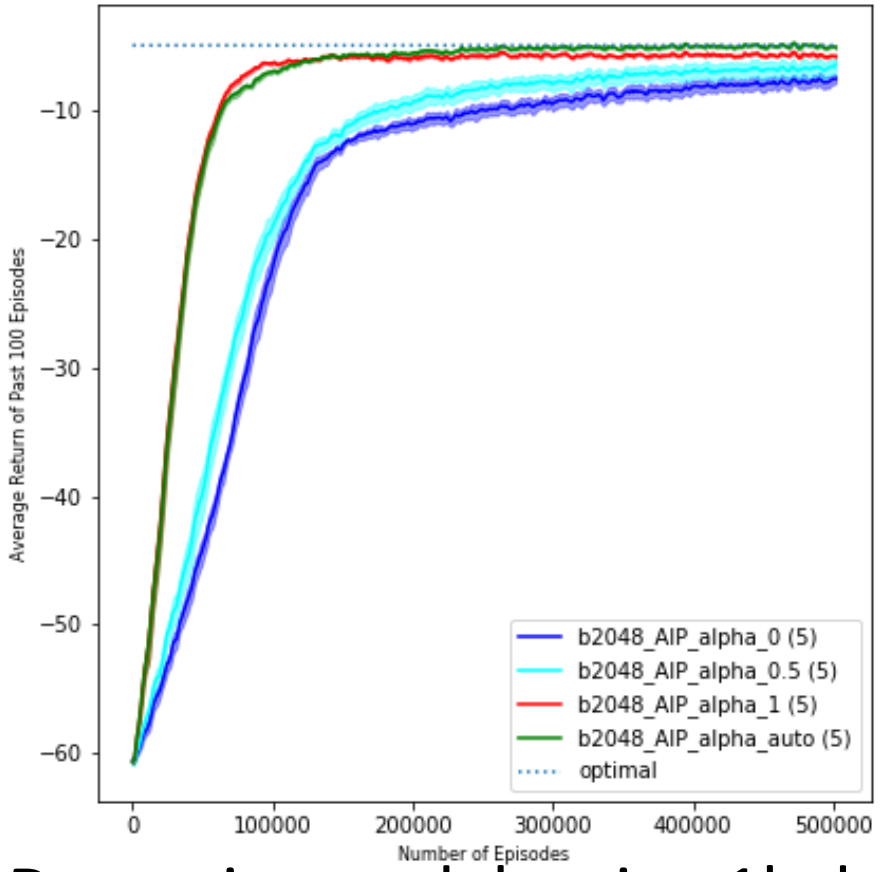
Shuo

Last time

α	0(model-free)	0.5	1(model-based)	automatic
Final Performance(1k)	-7.49	-6.36	-5.74	-5.04
Final Performance(100)	-7.61	-7.81	-9.99	-9.39



Dynamics model: using 100 data



Dynamics model: using 1k data

Problem: AIP performs better than model-based policy not significantly!

Have done so far

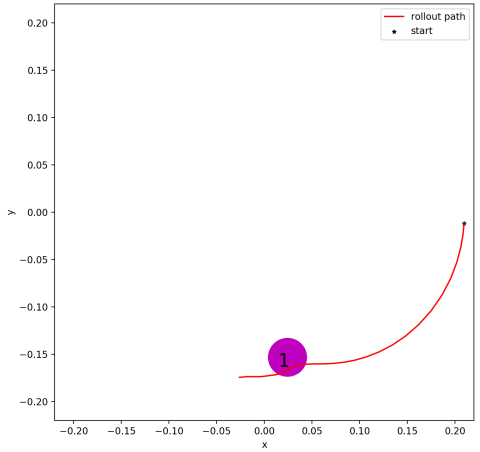
- Implemented the classification version of alpha training/prediction
- Deleted the meaningless exploration of model-based policy
- Investigated comprehensively the AIP performance with 3 different dynamics models and 8 model-based policy for each models

In Conclusion

- Classification generally works much better and more stable than regression
- Stochastic model-based policy has much lower returns than deterministic due to the bigger action value. (Reacher: Reward has an action penalty)
- AIP outperforms model-based policy generally, especially for stochastic model-based policy
- AIP has a much faster convergence generally than model-free policy at the beginning (Sometimes, AIP's final performance outperforms model-free policy)

Experiments

- Dynamics Model 1(DM1): Trained with 1k data without bias
- Dynamics Model 2(DM2): Trained with 100 data without bias
- Dynamics Model 3(DM3): Trained with 50 data with bias



Bias model

- Reference model-based policy: 2 versions; deterministic(det) or stochastic(sto)
- Reference model-based policy: 4 degrees of pre-training using dynamics model

1. Reference policy 1 (RP1): most well-trained, with 1e6 data
2. Reference policy 2 (RP2): with 1.5e5 data
3. Reference policy 3 (RP3): with 1e5 data
4. Reference policy 4 (RP4): most slightly-trained, with 5e4 data

DM1(RF det)

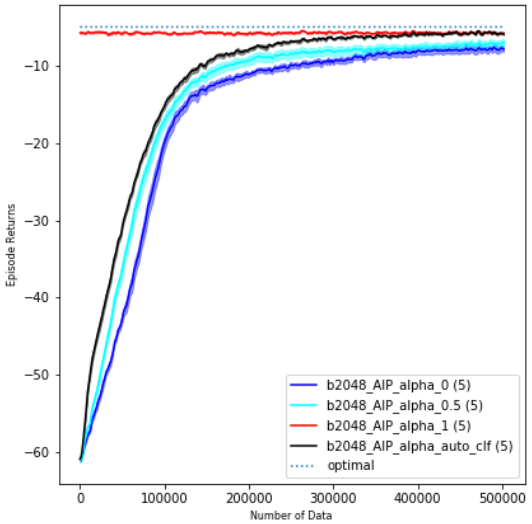


Fig 1.1 RF1

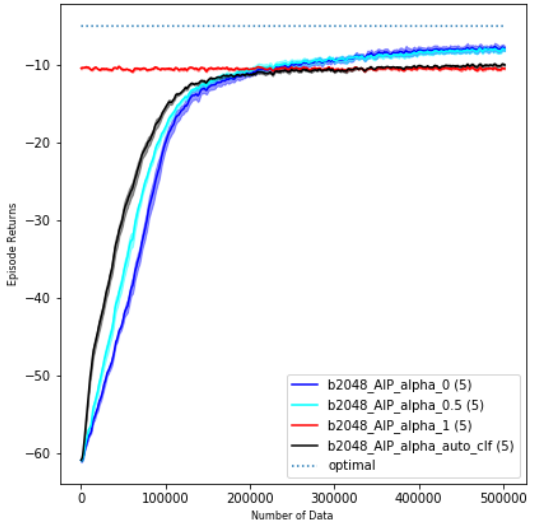


Fig 1.2 RF2

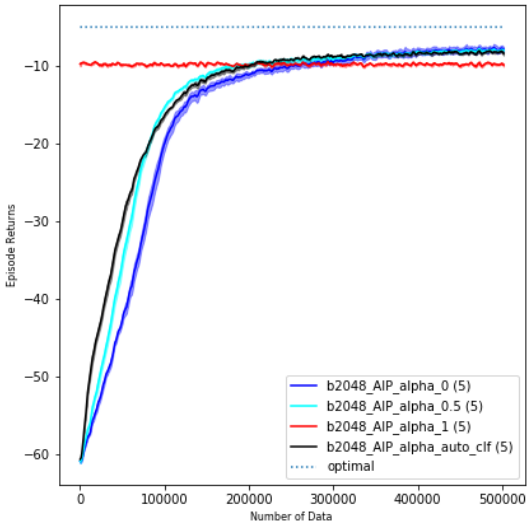


Fig 1.3 RF3

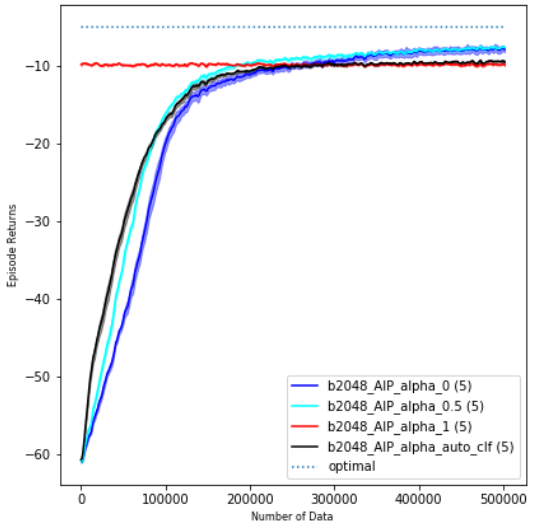


Fig 1.4 RF4

DM1(RF sto)

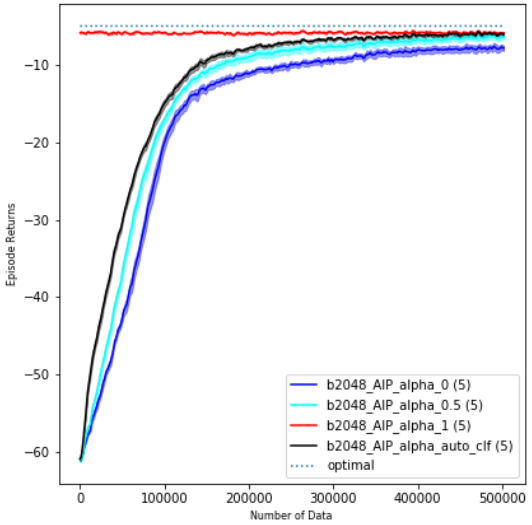


Fig 1.5 RF1

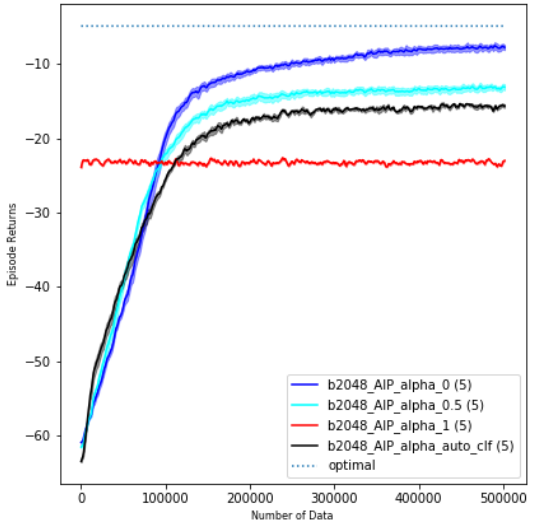


Fig 1.6 RF2

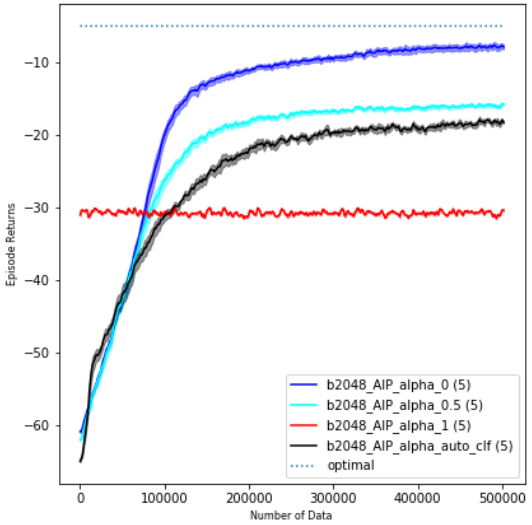


Fig 1.7 RF3

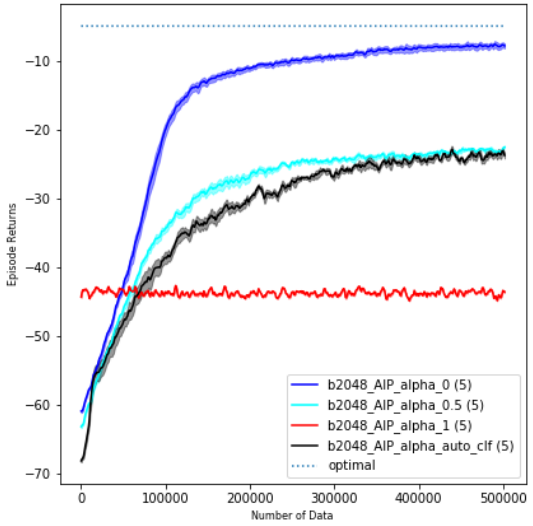


Fig 1.8 RF4

DM2(RF det)

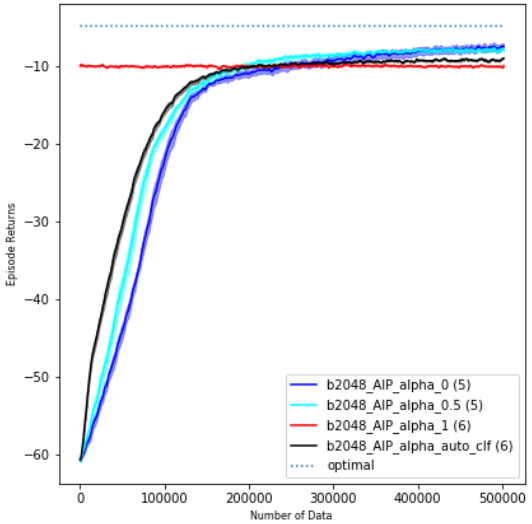


Fig 2.1 RF1

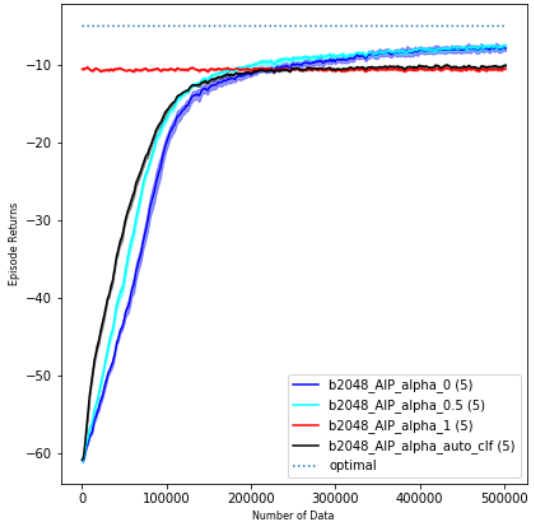


Fig 2.2 RF2

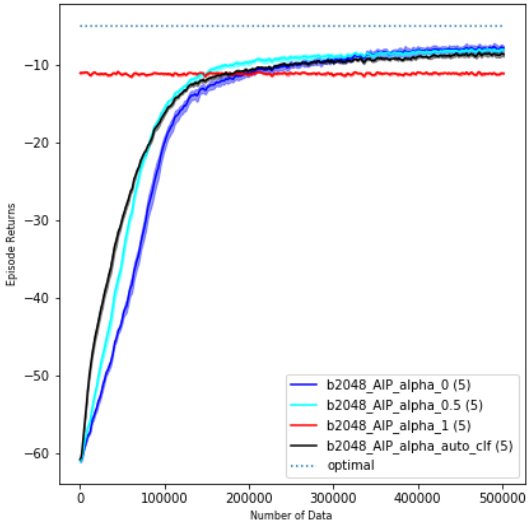


Fig 2.3 RF3

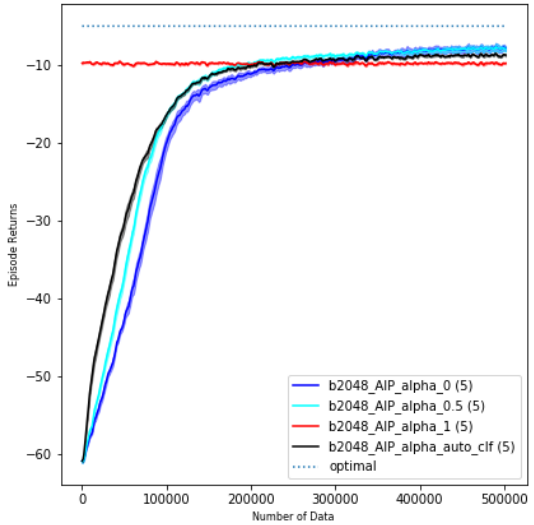


Fig 2.4 RF4

DM2(RF sto)

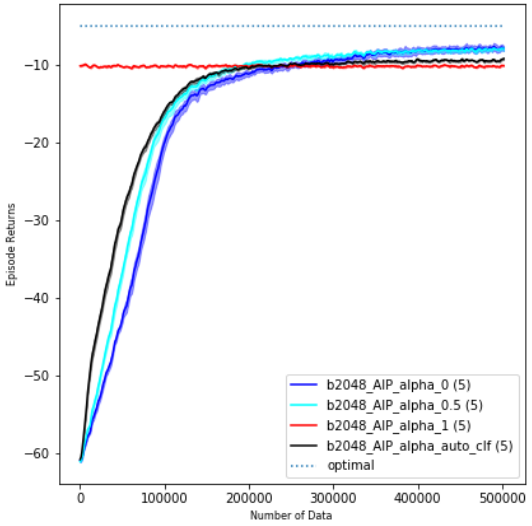


Fig 2.5 RF1

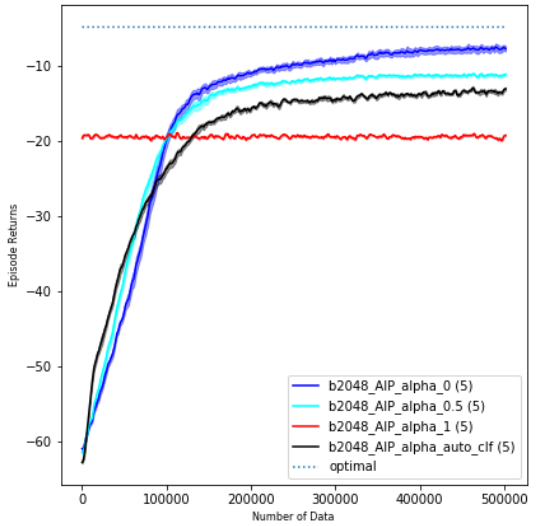


Fig 2.6 RF2

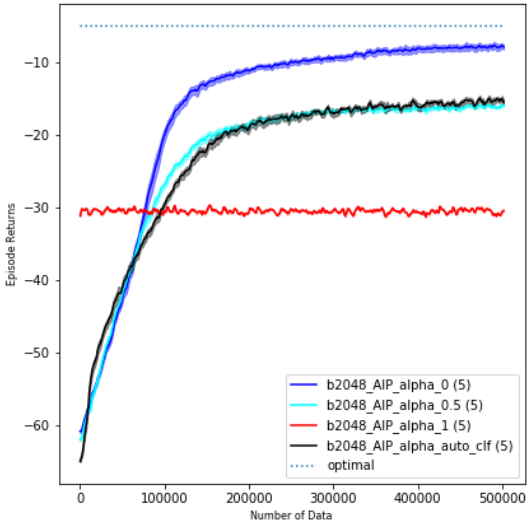


Fig 2.7 RF3

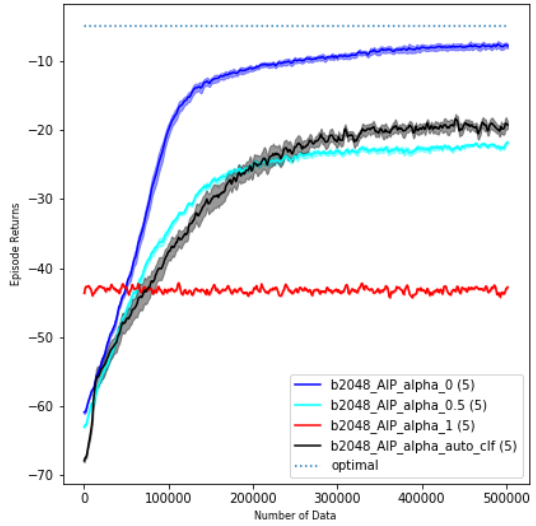


Fig 2.8 RF4

DM3(RF det)

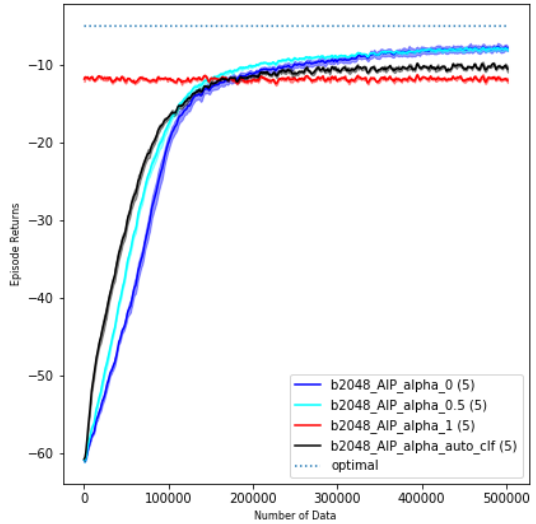


Fig 3.1 RF1

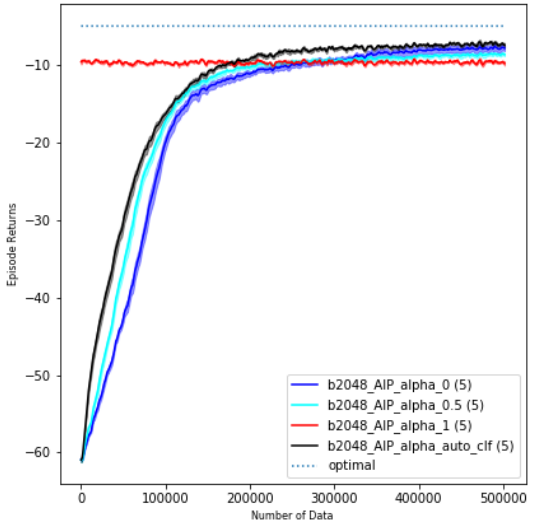


Fig 3.2 RF2

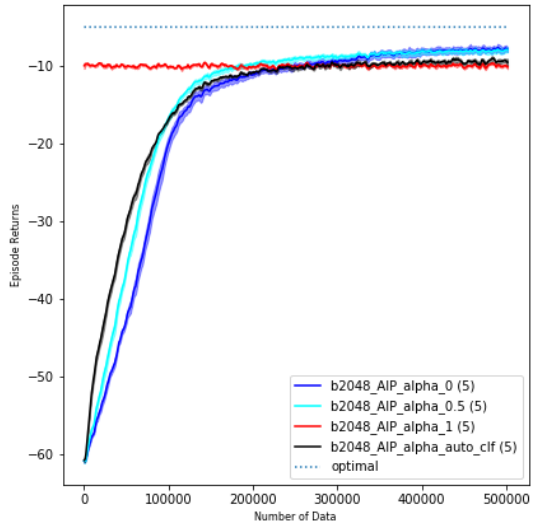


Fig 3.3 RF3

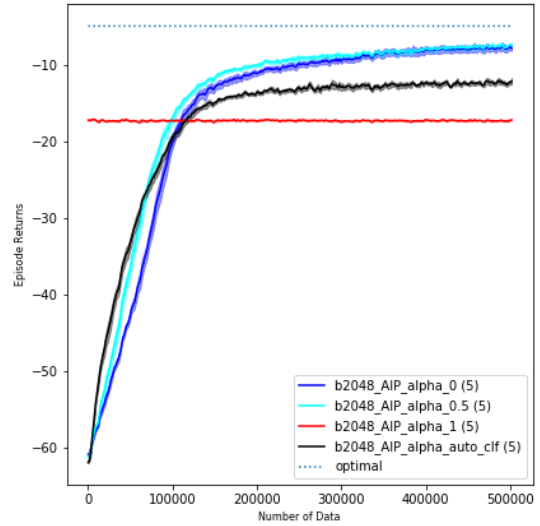


Fig 3.4 RF4

DM3(RF sto)

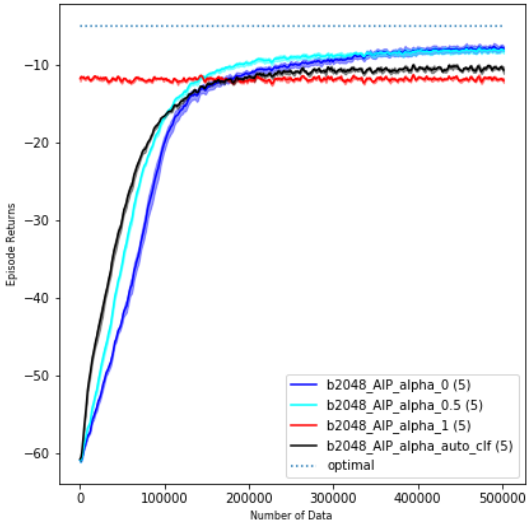


Fig 3.5 RF1

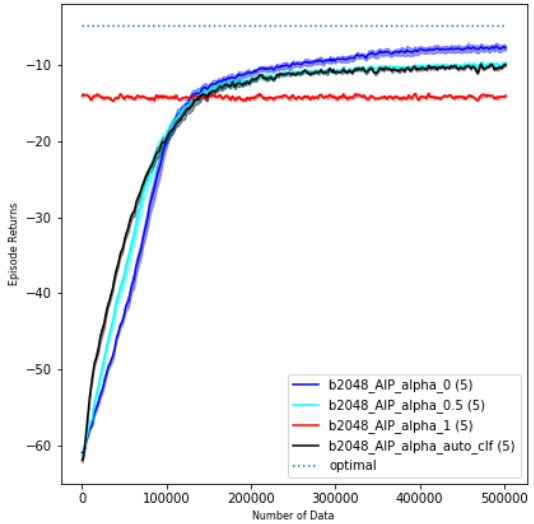


Fig 3.6 RF2

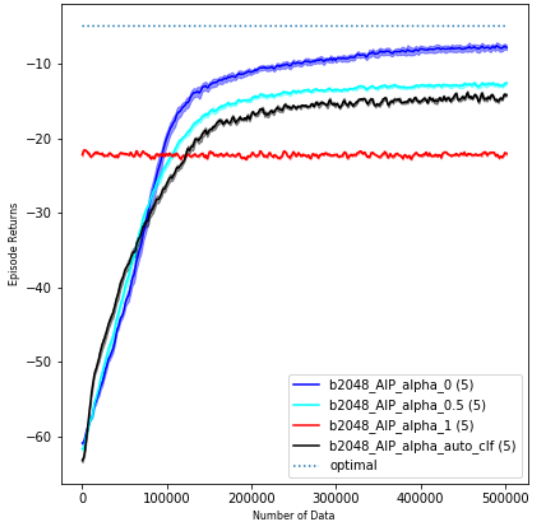


Fig 3.7 RF3

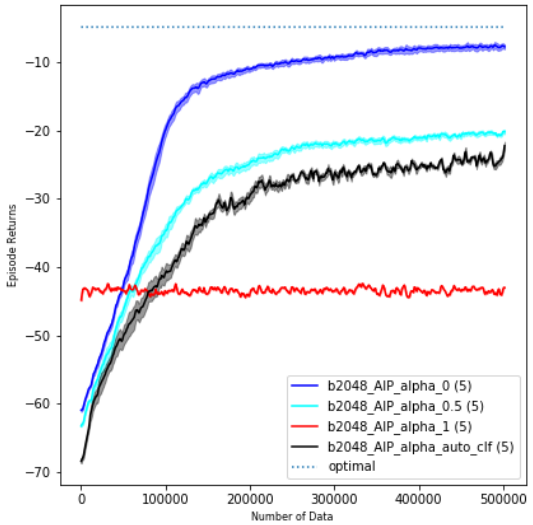
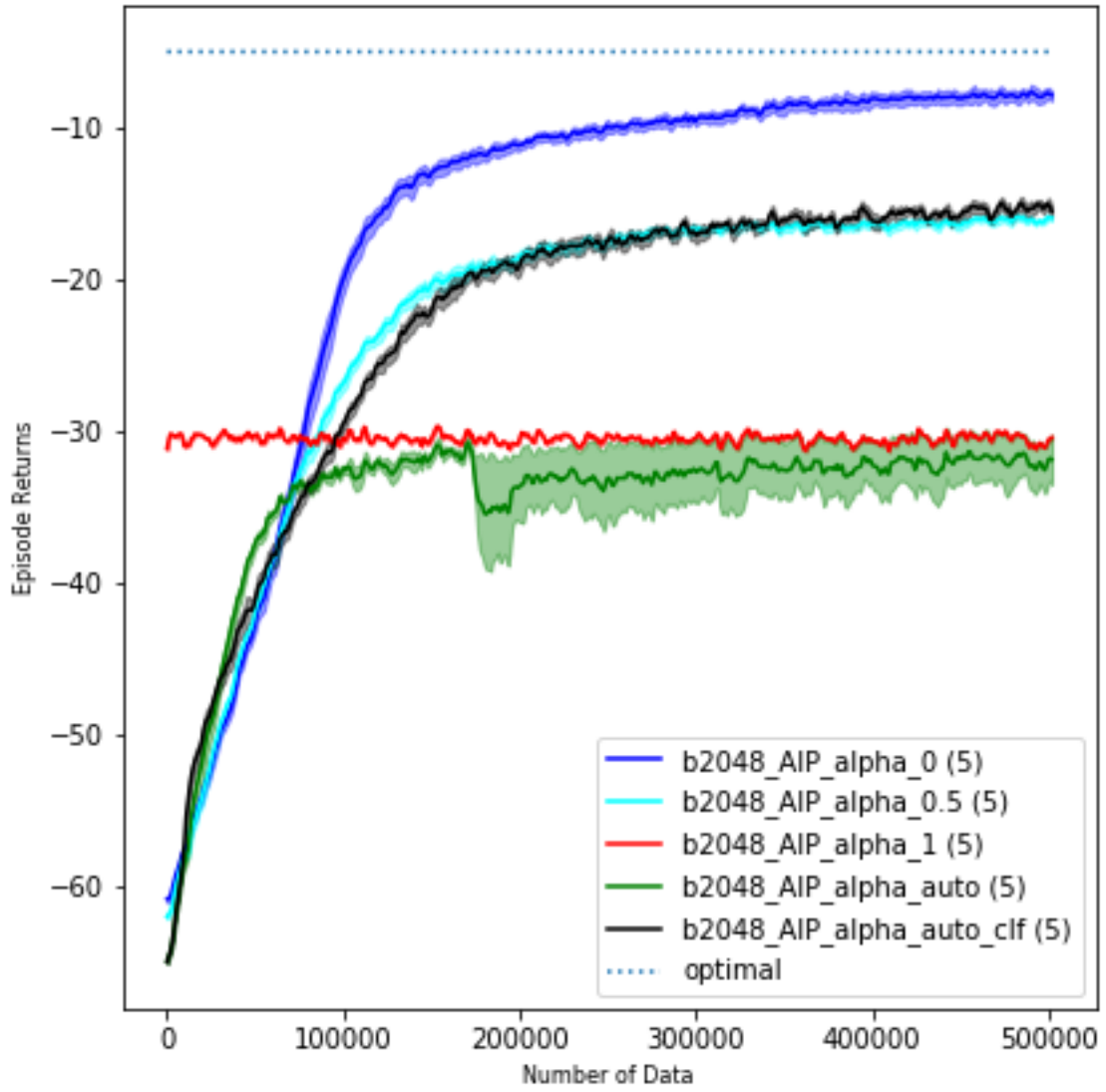
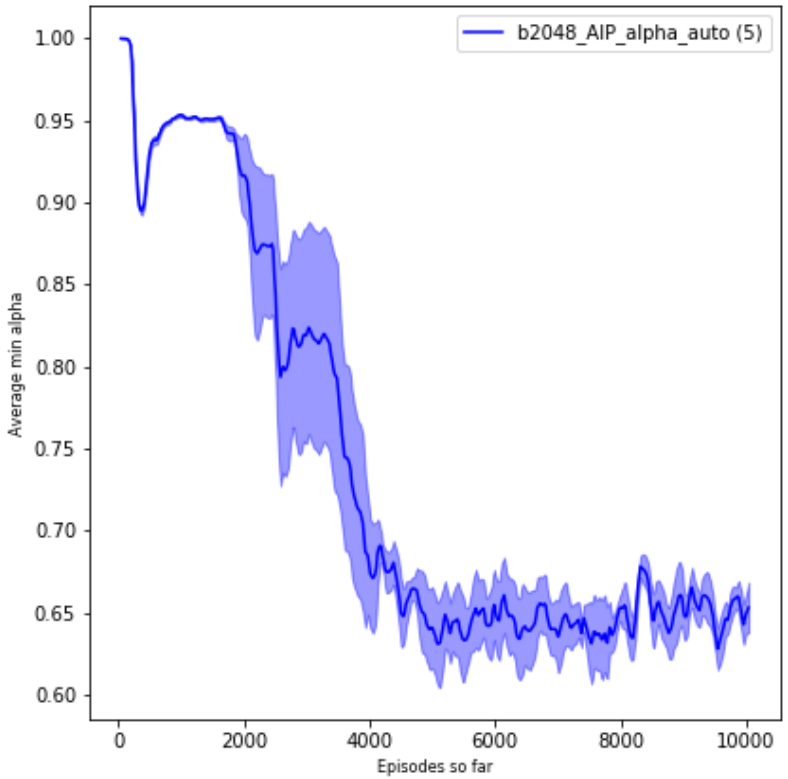


Fig 3.8 RF4

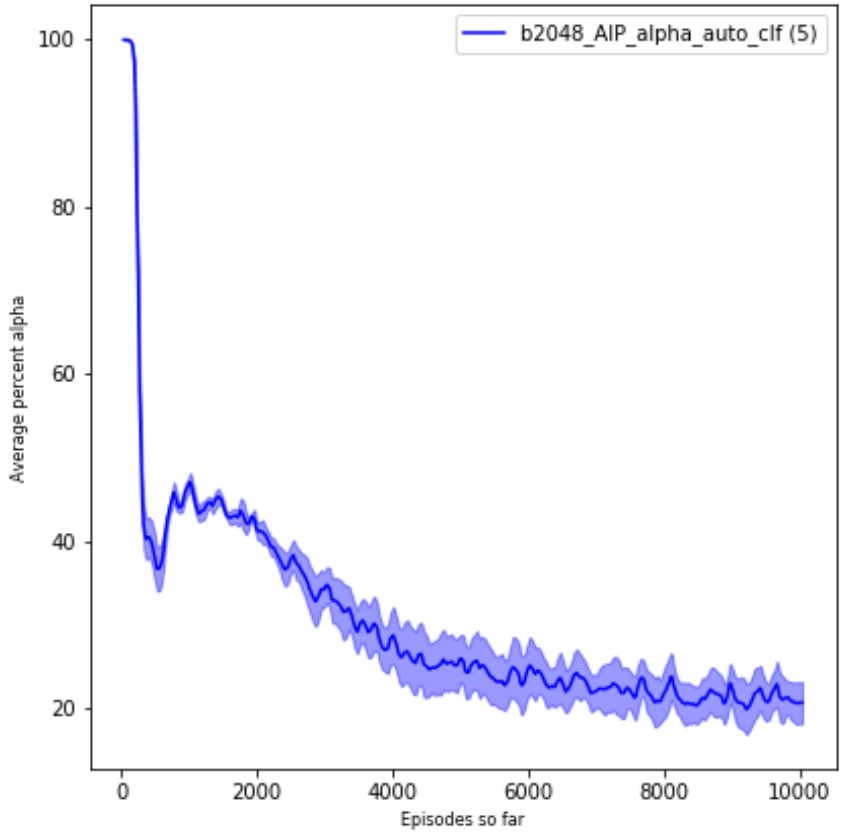
Alpha comparison (DM2 RF3 sto)



Alpha comparison (DM2 RF3 sto)

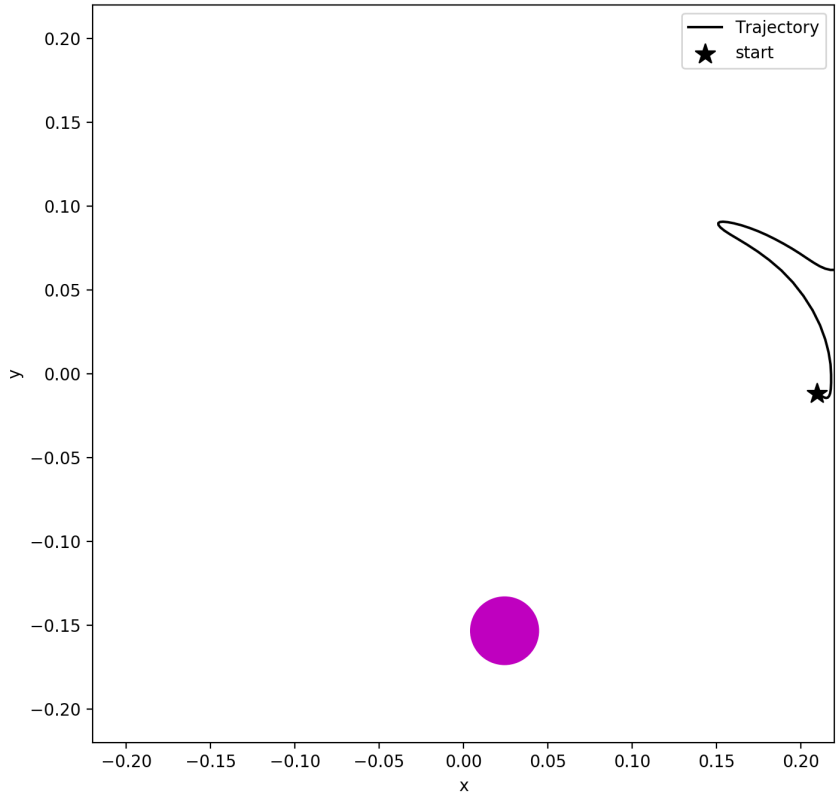


Regression



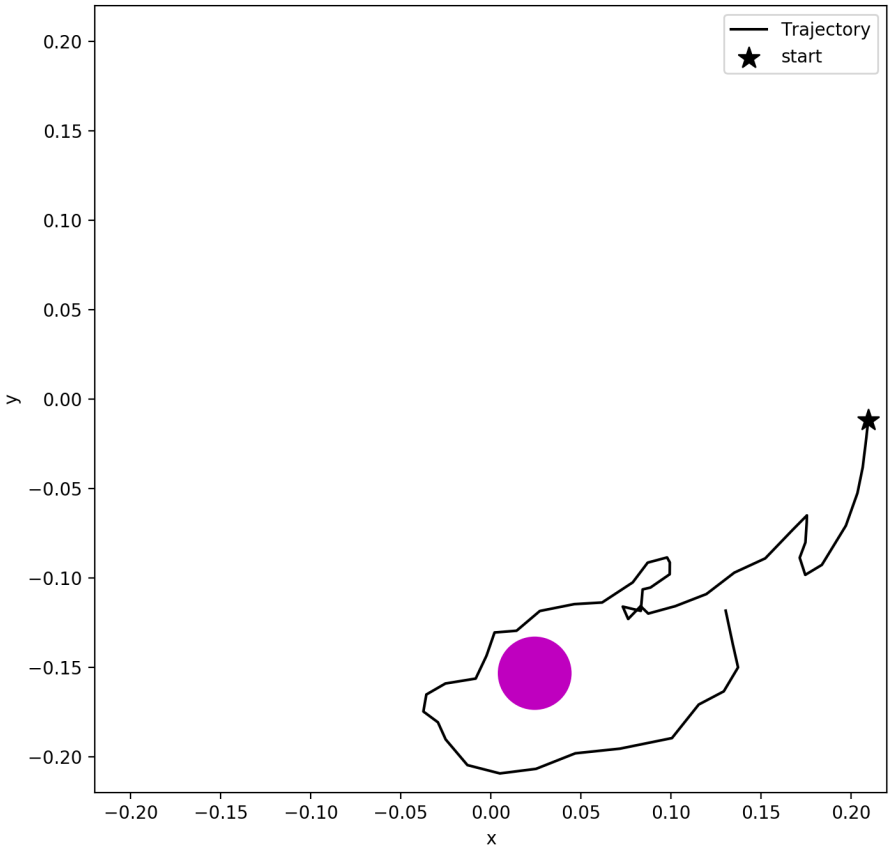
Classification

DM2 (Reference policy rollout in DM2)
Action value matters much in return



Return -11

Reference policy deterministic



Return -60

Reference policy stochastic

Next plan

- Reacher with obstacles?
- Other environments: such as Fetch-Reacher in openai?