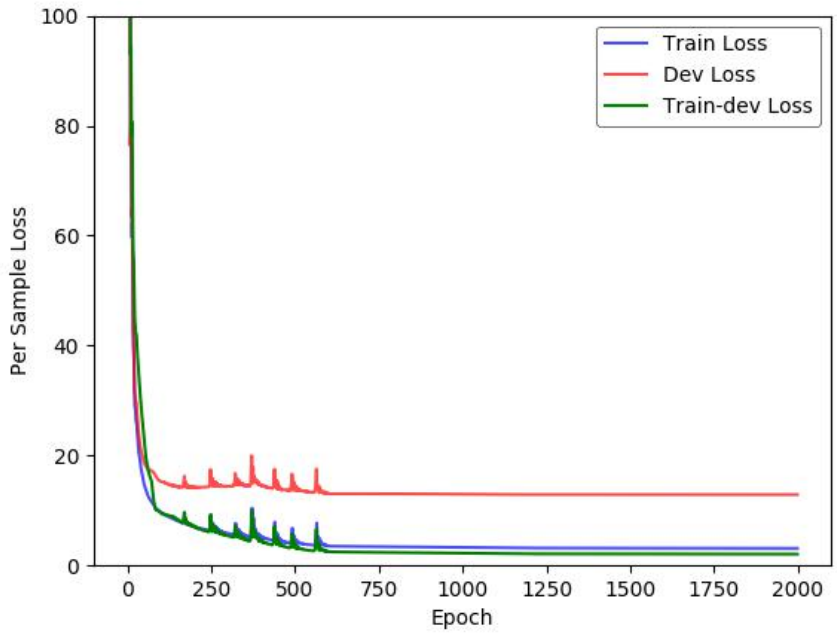
Comparison of Initialized and Non-initialized policy for RL

The training curve in supervised learning:

Y-Axis: per sample loss (Mean Square Error), the average loss of each joint angle, angle value ranges in [-360, 360] deg.
X-axis: the epoch.
Training data: 2000*50*3=3e5 samples; generated from inverse kinematics; initial joint angles of each episode is [0.1,0.1,0.1] (if set [0.0,0.0,0.0], the inverse jacobian will have large values); action is calculated through the difference of joint angle values in target position and initial position divided by N; N is set to be 10 (i.e. 10 steps per episode).
Loss:
(1) Train Loss: average loss of single training epoch;
(2) Dev Loss: loss of test with a separate data of 45000 (300*50*3) samples, not used in training;
(3) Train-dev Loss: average loss of test with random 2% (6e3) of the whole training data.
Network structure:
5 layers of fully connected with 100 nodes each, relu as hidden layer activation, and output layer activation is 360*tanh to match the range of angle value.



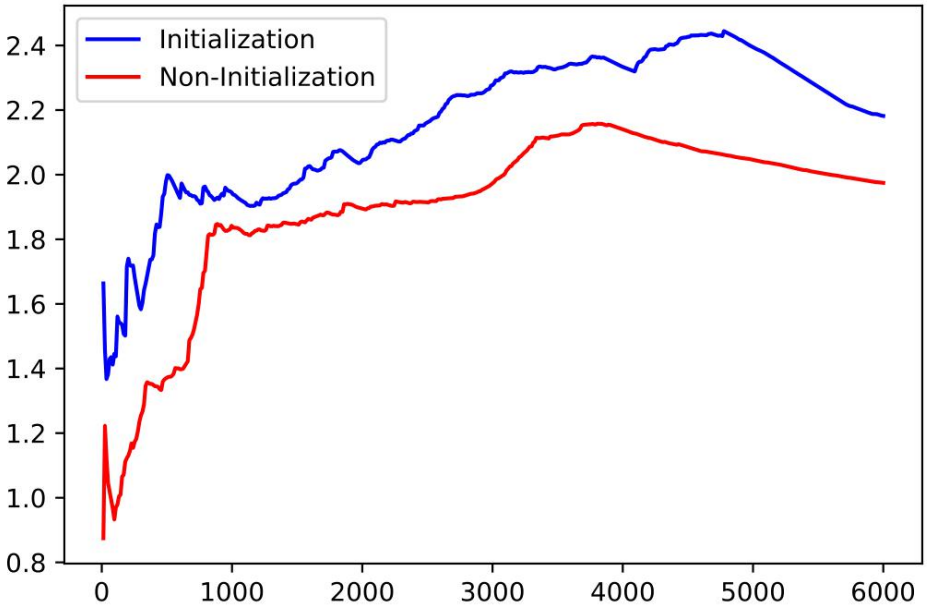
Results: as the steps per episode is set to be 10 in training data, the reacher will almost always reach a position near the target position with around 10 steps; but after that it will not converge at this position, it starts to move around in quite a

large range. However, we could set the steps per episode in RL training process to be 10 as well.

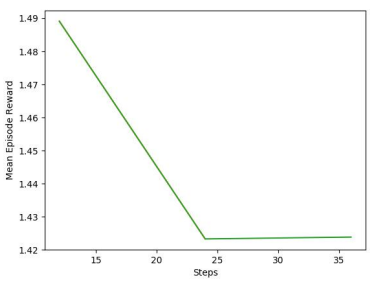
RL training with initialized policy:

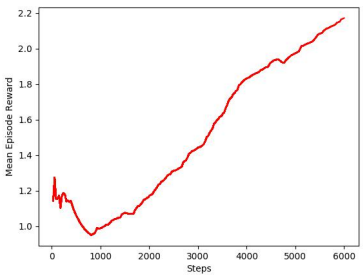
General settings: the actor network has the same structure with the supervised learning network. If training RL with initialized policy, the actor is preloaded with pretrained policy and target-actor is made a copy of the actor. The critic is initialized randomly. The steps per episode is set to be 12, similar as in training data in supervised learning.

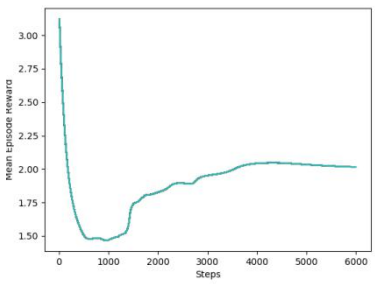
Comparison of initialized and non-initialized RL learning process: (y-axis is reward and x-axis is learning epoch) initialized policy will have larger reward at beginning stage, and always keep larger reward at each epoch.



Here are 3 training processes of initialized RL: although beginning rewards are large with initialized policy, it always falls down at the beginning period after some training or exploration, and then increases to larger rewards.







Some parameters in RL will affect the effects of initialization: exploration noise, normalization of observation, learning rate of actor and critic, number of steps per episode and so on in RL algorithms will all affect the initialization policy to have different influences for RL's training. If we want to show more effects in learning curve caused by initialized policy (to change the initialized policy more smoothly), it needs to set smaller noise (wont affect original action); remove the normalization of observation (this is important, make sure inputs of neural networks are same as in supervised learning); set smaller actor and critic learning rate (not change the actor too much); the number of steps of one episode needs to be same as the steps of epoch in training data of supervised learning for initialization policy.

Comparison of initialized and non-initialized RL learning process with smaller learning rate and smaller exploration noise is shown below: the decrease of reward with initialized policy slows down, but still exists. Smaller exploration noise makes the reward increase slower.