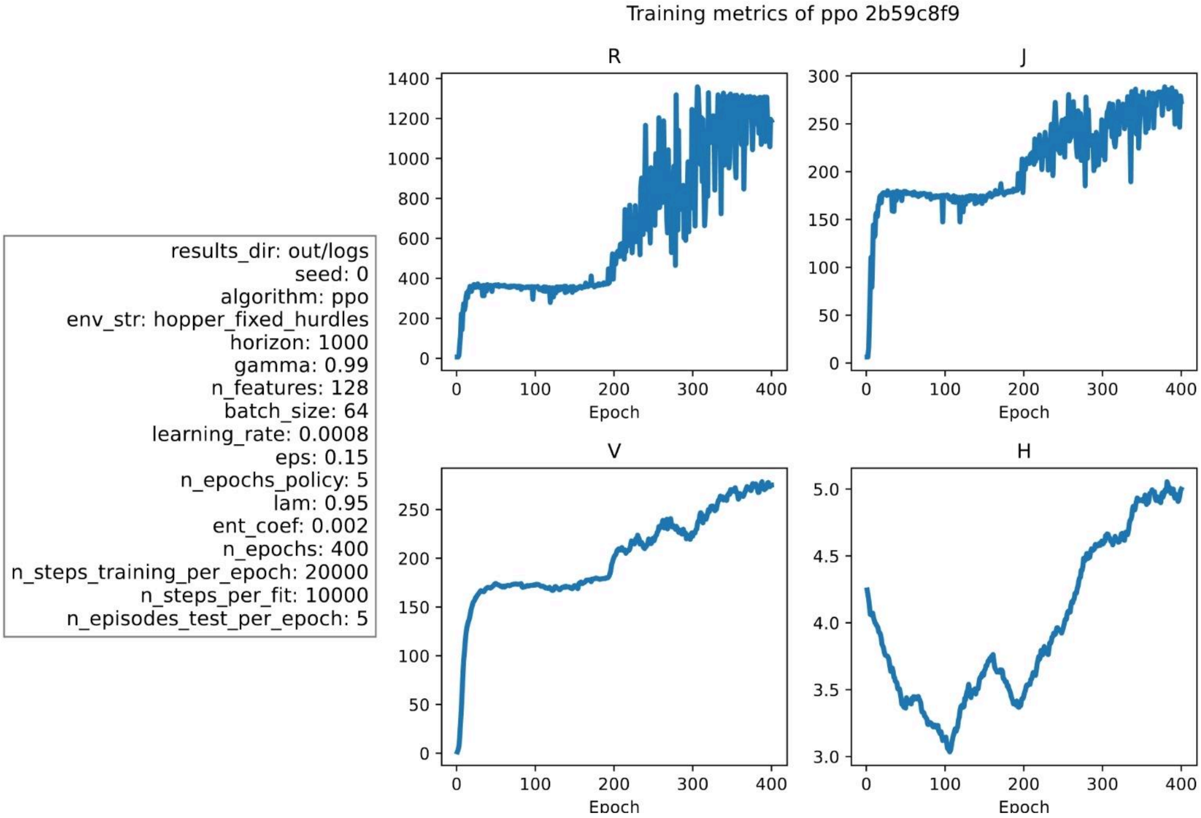
3.1 Code Implementation done

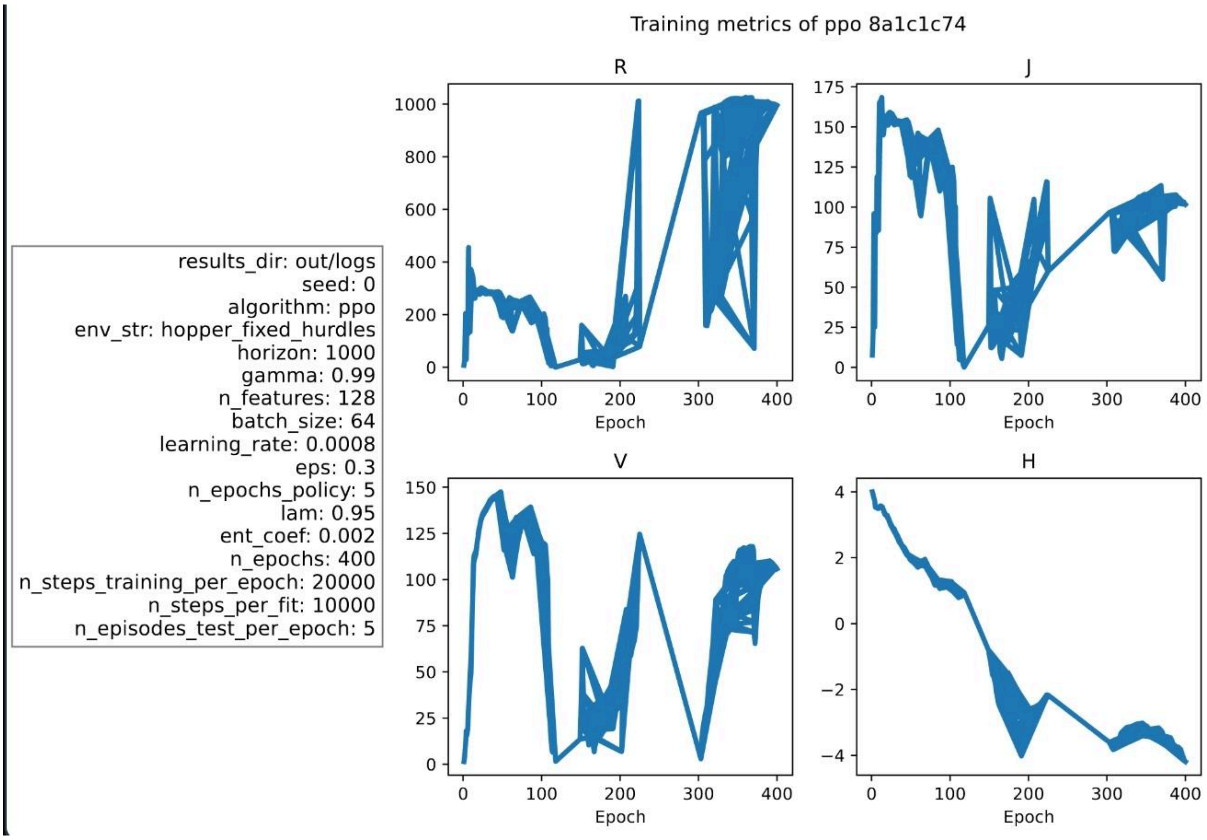
3.2

PPO results:

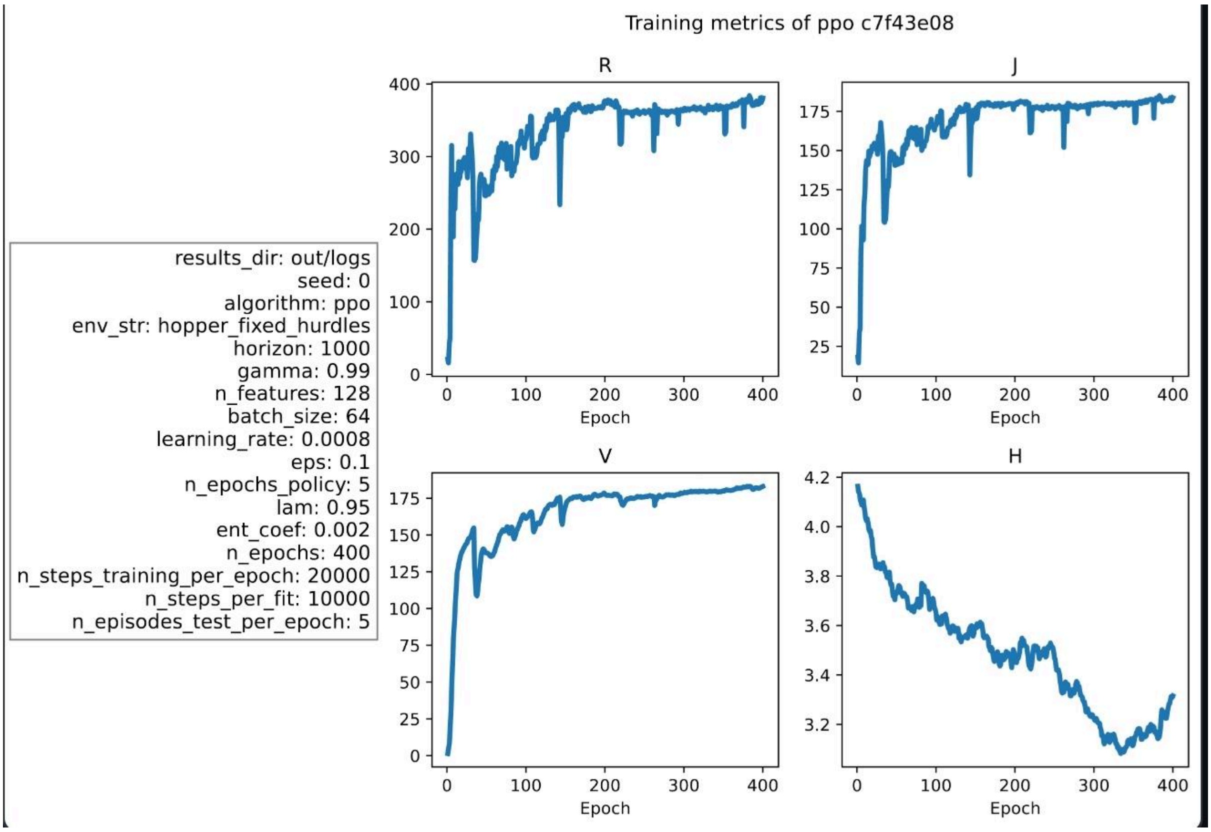
Best agent: eps = 0.15



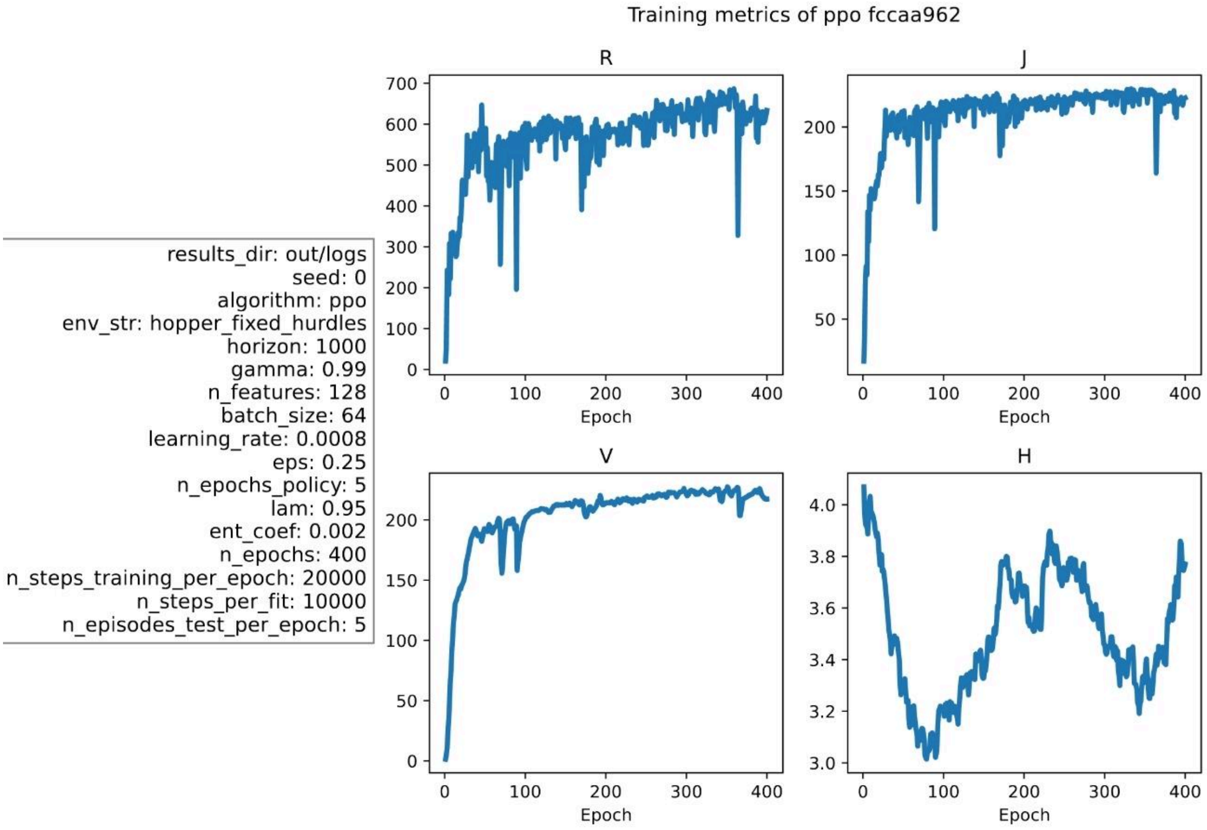
```
results_dir: out/logs
seed: 0
algorithm: ppo
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 64
learning_rate: 0.0008
eps: 0.15
n_epochs_policy: 5
lam: 0.95
ent_coef: 0.002
n_epochs: 400
n_steps_training_per_epoch: 20000
n_steps_per_fit: 10000
n_episodes_test_per_epoch: 5
```



```
results_dir: out/logs
seed: 0
algorithm: ppo
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 64
learning_rate: 0.0008
eps: 0.3
n_epochs_policy: 5
lam: 0.95
ent_coef: 0.002
n_epochs: 400
n_steps_training_per_epoch: 20000
n_steps_per_fit: 10000
n_episodes_test_per_epoch: 5
```

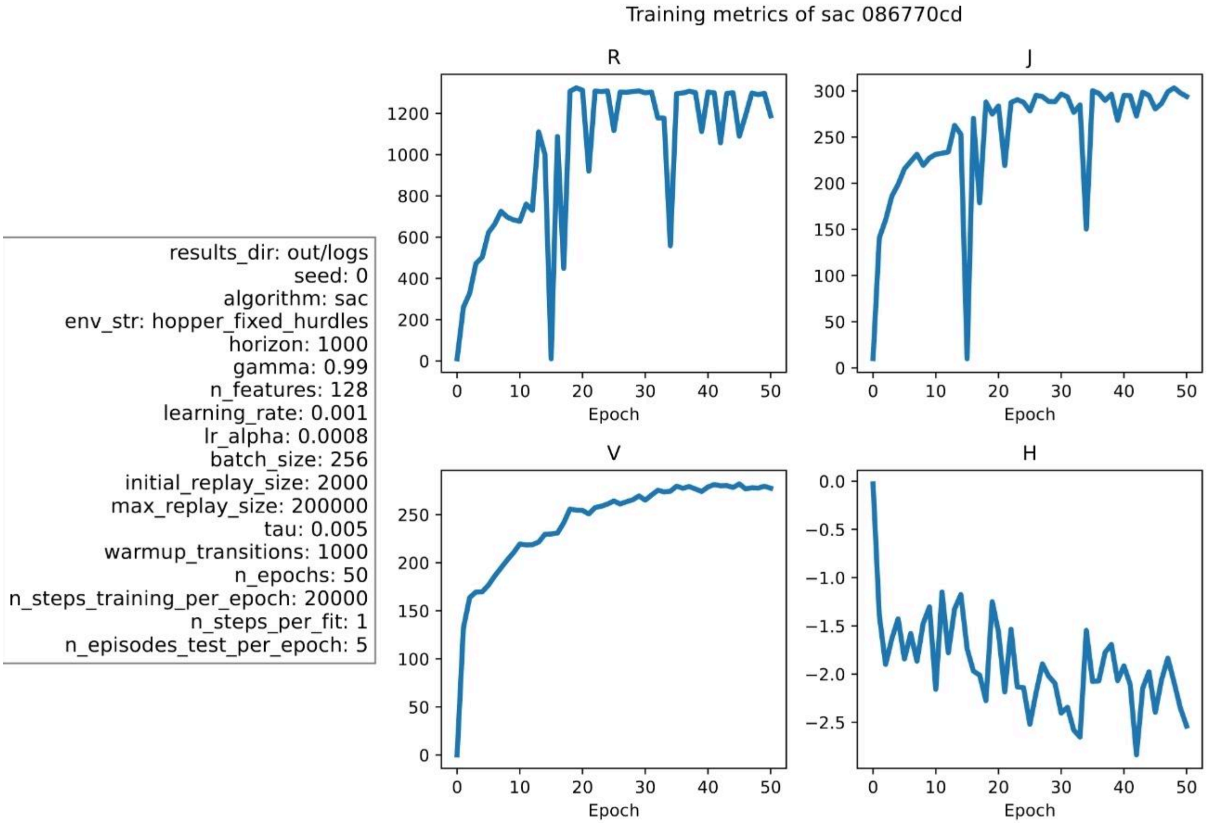


```
results_dir: out/logs
seed: 0
algorithm: ppo
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 64
learning_rate: 0.0008
eps: 0.1
n_epochs_policy: 5
lam: 0.95
ent_coef: 0.002
n_epochs: 400
n_steps_training_per_epoch: 20000
n_steps_per_fit: 10000
n_episodes_test_per_epoch: 5
```

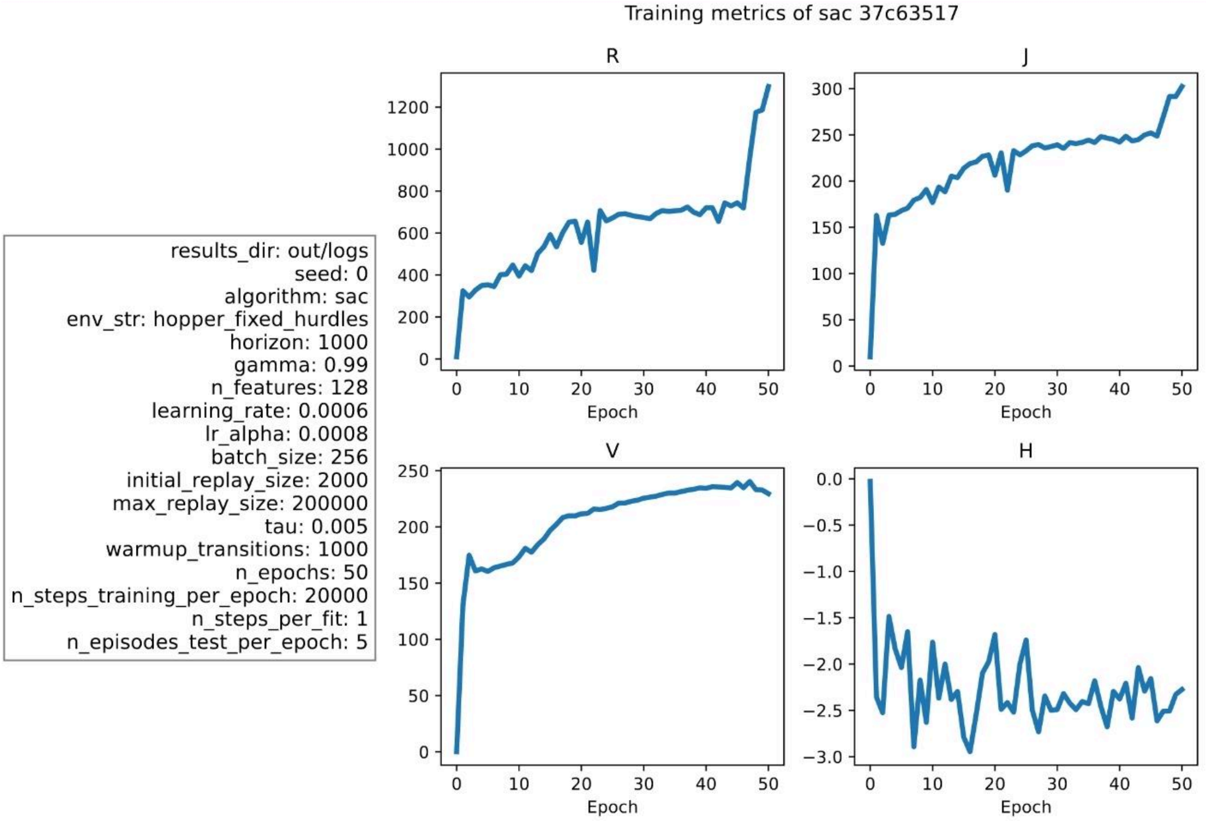


```
results_dir: out/logs
seed: 0
algorithm: ppo
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 64
learning_rate: 0.0008
eps: 0.25
n_epochs_policy: 5
lam: 0.95
ent_coef: 0.002
n_epochs: 400
n_steps_training_per_epoch: 20000
n_steps_per_fit: 10000
n_episodes_test_per_epoch: 5
```

SAC Results:

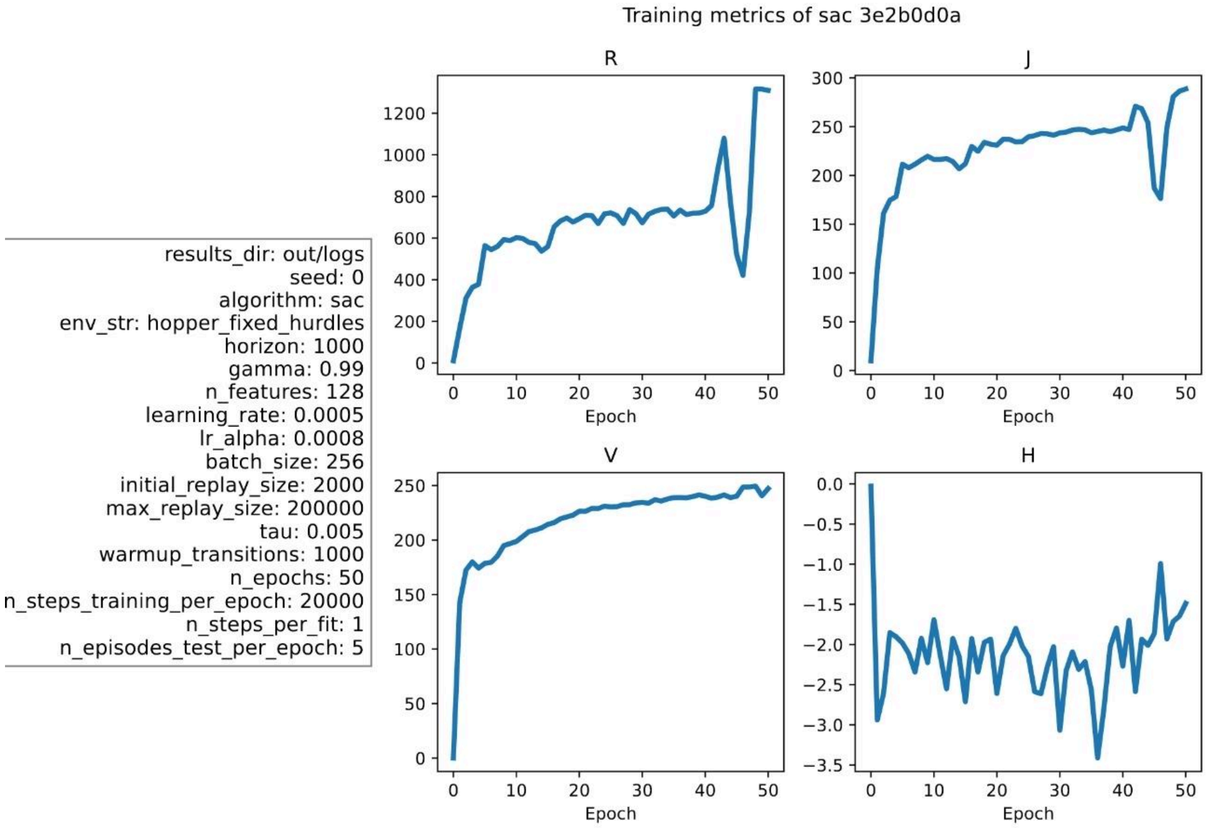


```
results_dir: out/logs
seed: 0
algorithm: sac
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
learning_rate: 0.001
lr_alpha: 0.0008
batch_size: 256
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
warmup_transitions: 1000
n_epochs: 50
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```



Training metrics of sac 37c63517

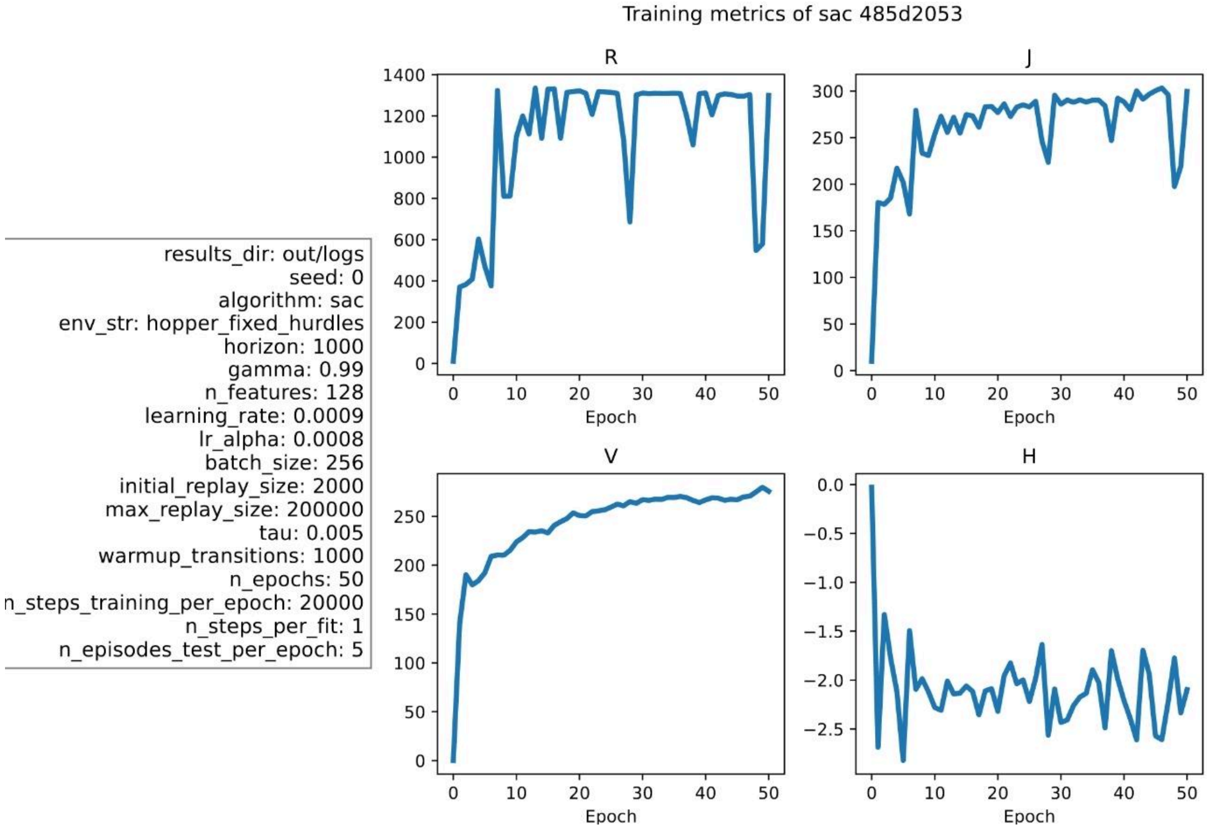
```
results_dir: out/logs
seed: 0
algorithm: sac
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
learning_rate: 0.0006
lr_alpha: 0.0008
batch_size: 256
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
warmup_transitions: 1000
n_epochs: 50
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```



Training metrics of sac 3e2b0d0a

```
results_dir: out/logs
seed: 0
algorithm: sac
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
learning_rate: 0.0005
lr_alpha: 0.0008
batch_size: 256
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
warmup_transitions: 1000
n_epochs: 50
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```

Best agent: learning rate = 0.0009

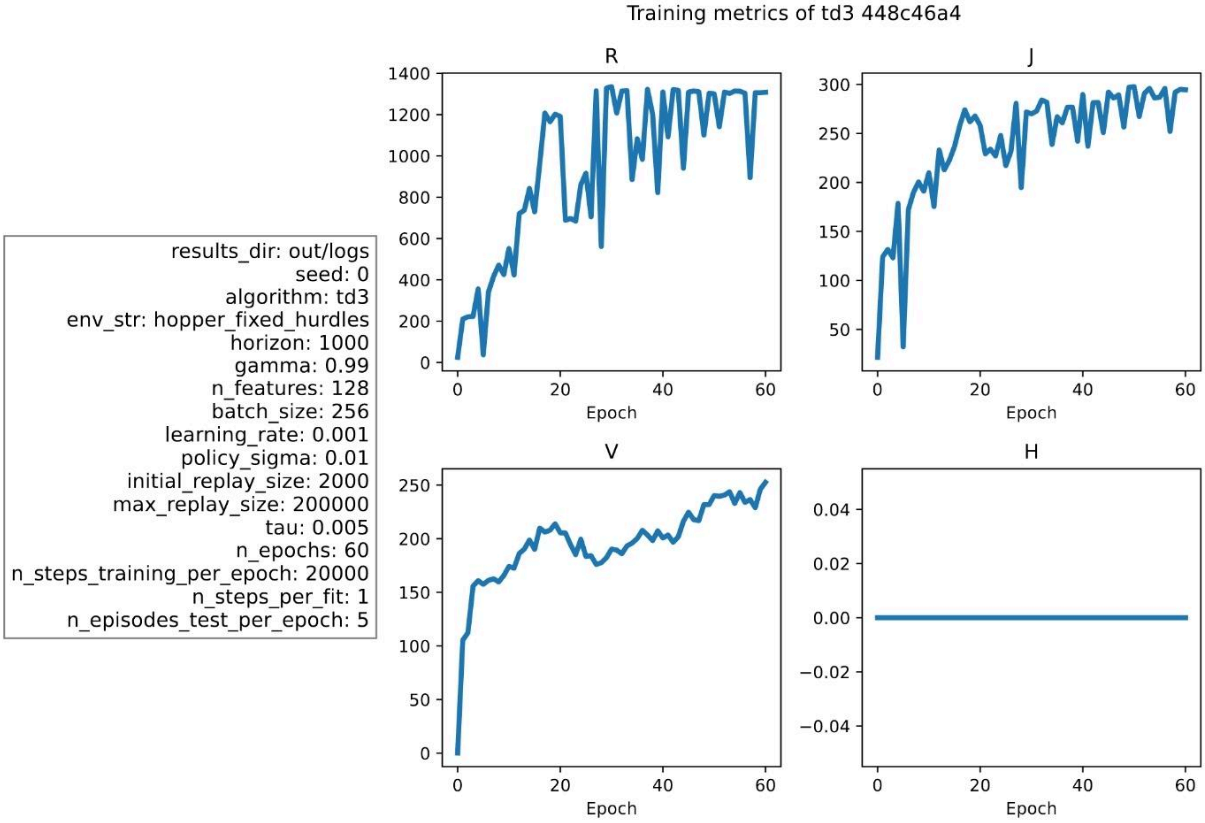


Training metrics of sac 485d2053

```
results_dir: out/logs
seed: 0
algorithm: sac
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
learning_rate: 0.0009
lr_alpha: 0.0008
batch_size: 256
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
warmup_transitions: 1000
n_epochs: 50
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```

TD3 Results:

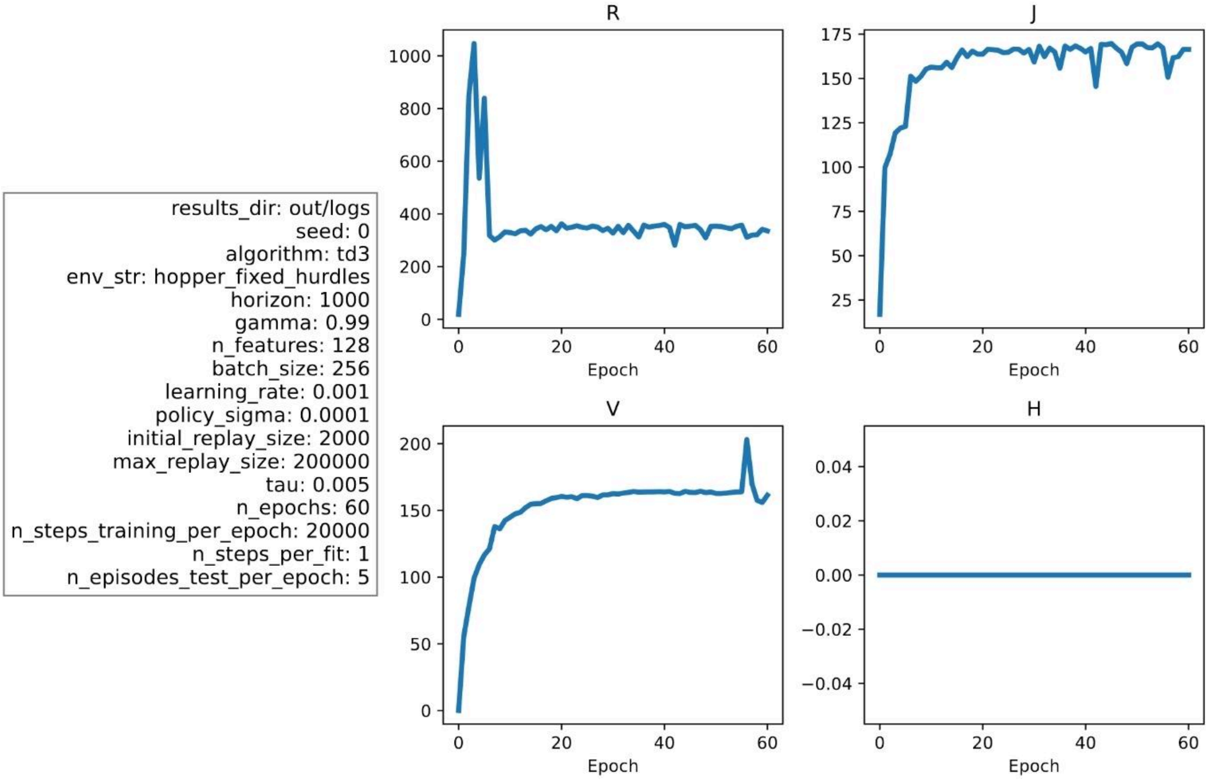
Best agent: policy sigma = 0.01



Training metrics of td3 448c46a4

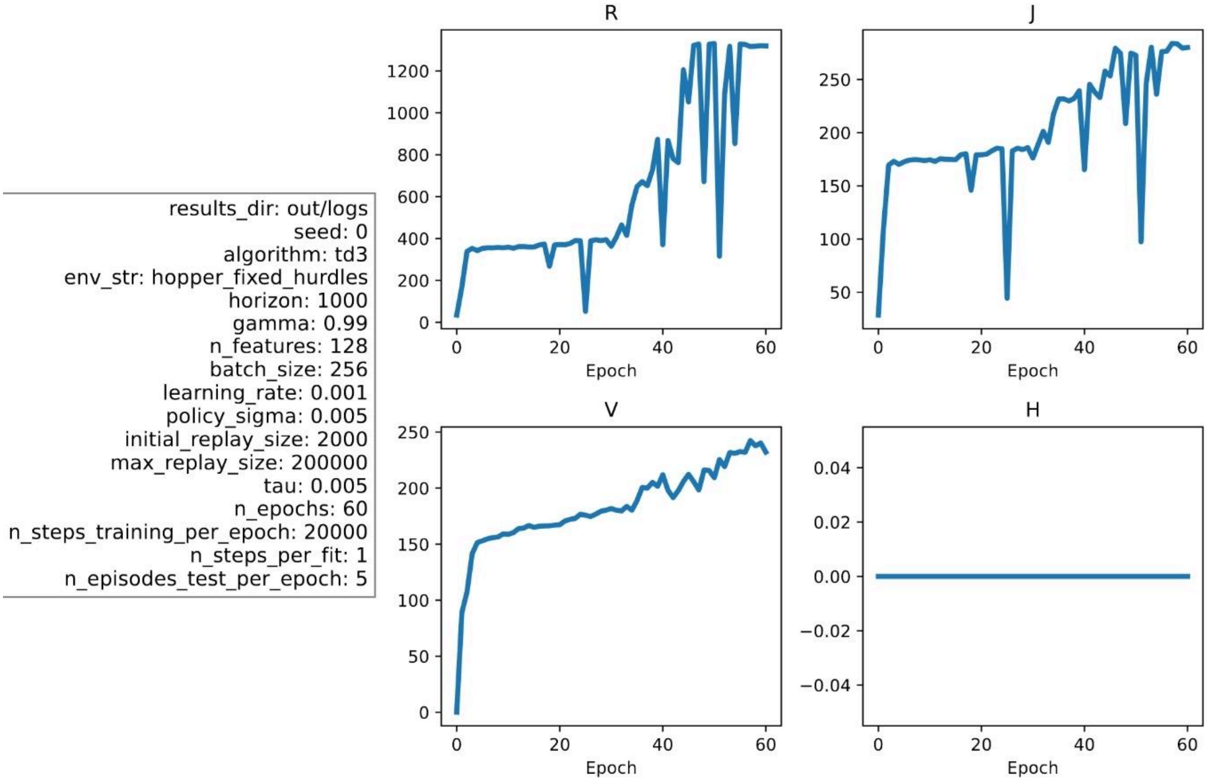
```
results_dir: out/logs
seed: 0
algorithm: td3
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 256
learning_rate: 0.001
policy_sigma: 0.01
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
n_epochs: 60
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```

Training metrics of td3 98d8617d

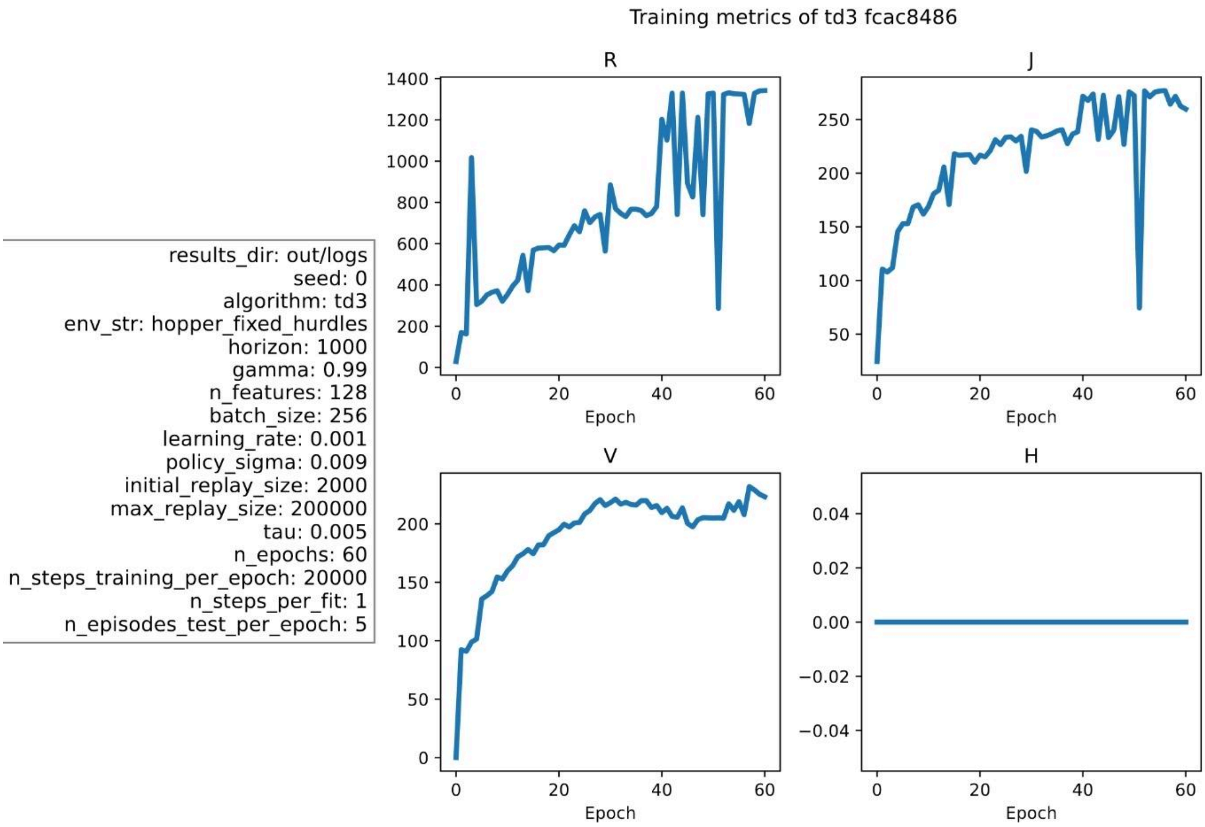


```
results_dir: out/logs
seed: 0
algorithm: td3
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 256
learning_rate: 0.001
policy_sigma: 0.0001
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
n_epochs: 60
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```

Training metrics of td3 f30ccbee



```
results_dir: out/logs
seed: 0
algorithm: td3
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 256
learning_rate: 0.001
policy_sigma: 0.005
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
n_epochs: 60
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```



```
results_dir: out/logs
seed: 0
algorithm: td3
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 256
learning_rate: 0.001
policy_sigma: 0.009
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
n_epochs: 60
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```

3.3

1 a)

action[:, 2] : It is the sequence of torque applied at the foot of the hopper

state[:, 0]: It is the sequence of z-coordinates of the torso (height of hopper).

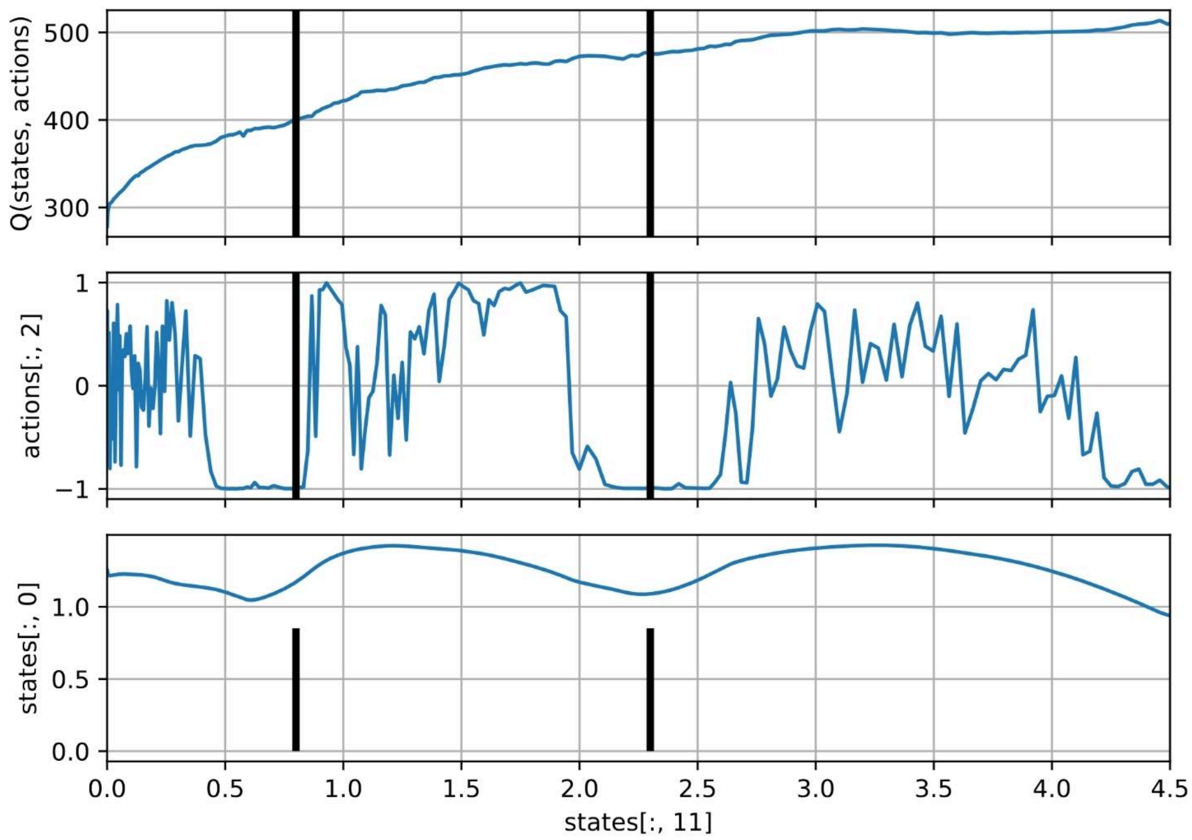
states[:, 11] : It is the sequence of x-coordinates of the torso.

b) Q value increases as the x value increases because of the forward reward.

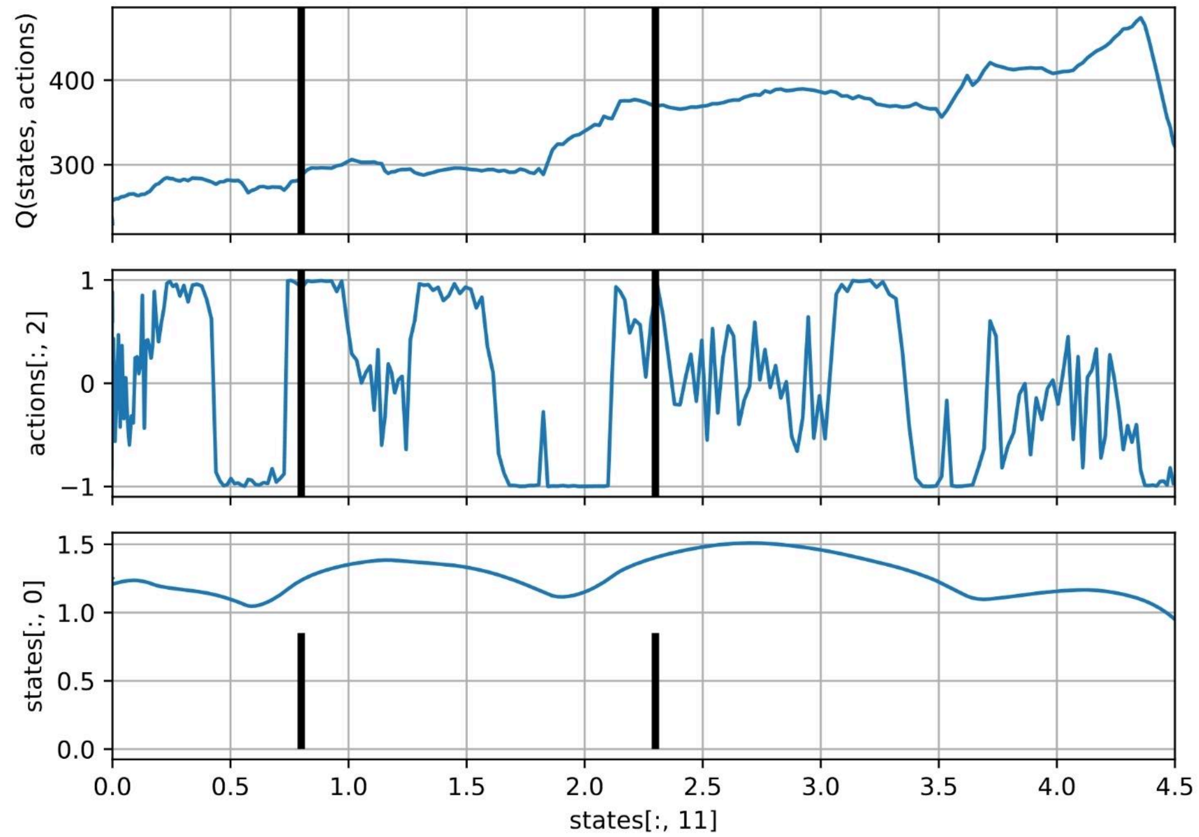
c) action[:, 2] is negative inorder to cross the hurdle.(Maybe creating space in the front part of the foot to go over the hurdle)

1 i)

hash : 086770cd

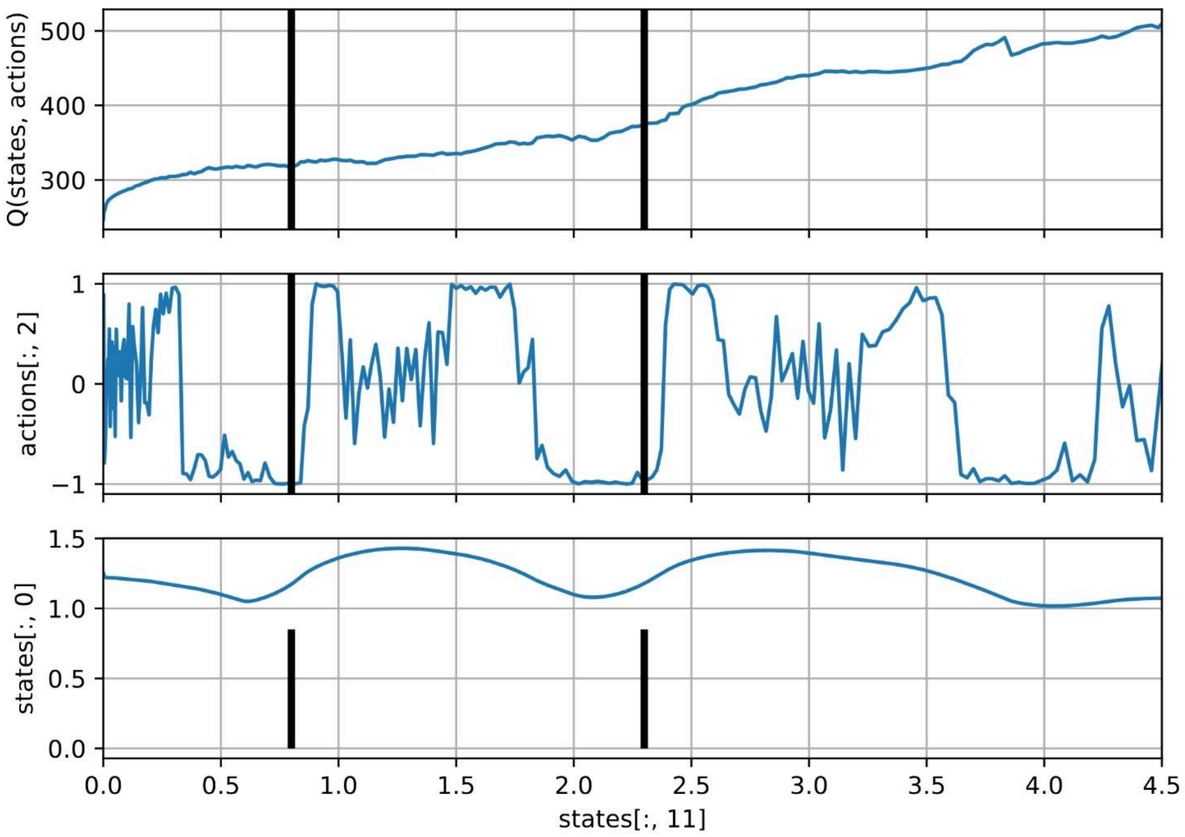


1 ii)
hash: 37c63517



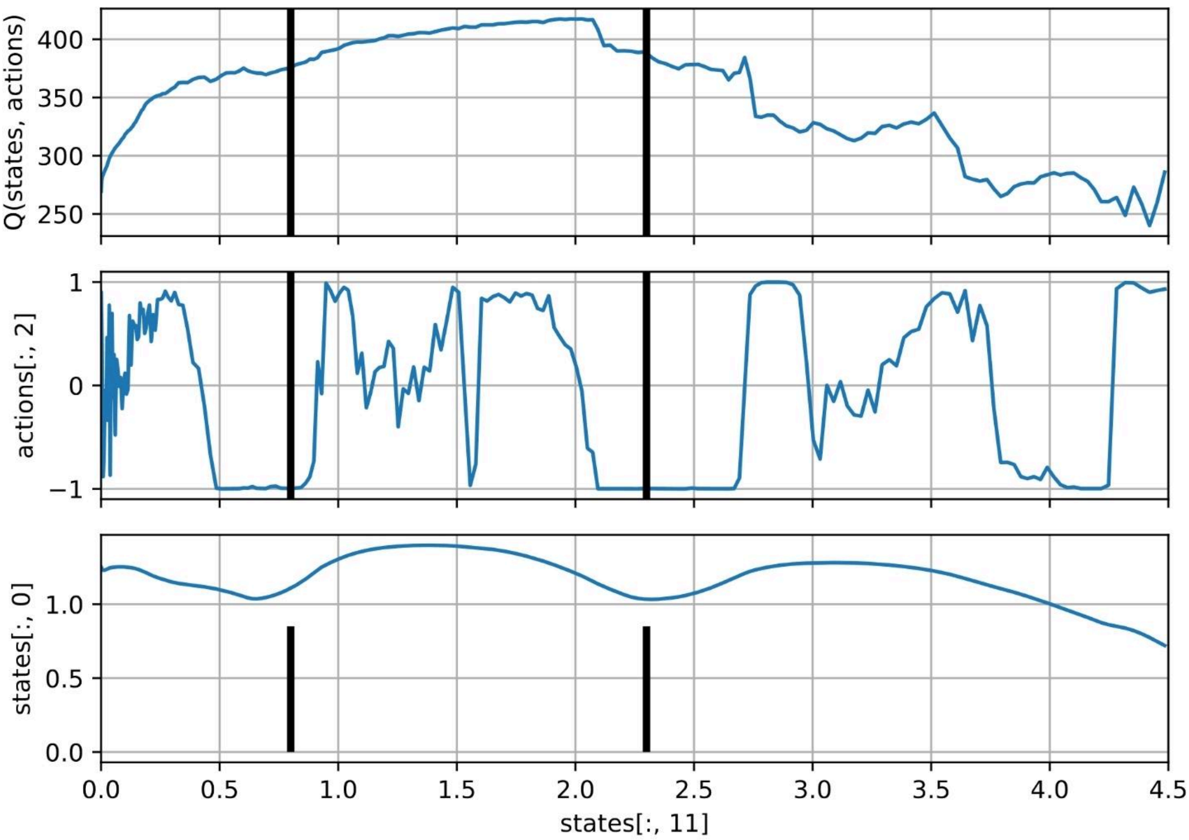
1 iii)

hash: 3e2b0d0a



1 iv)

hash: 485d2053



2

action[1]: Torque applied on the leg rotor

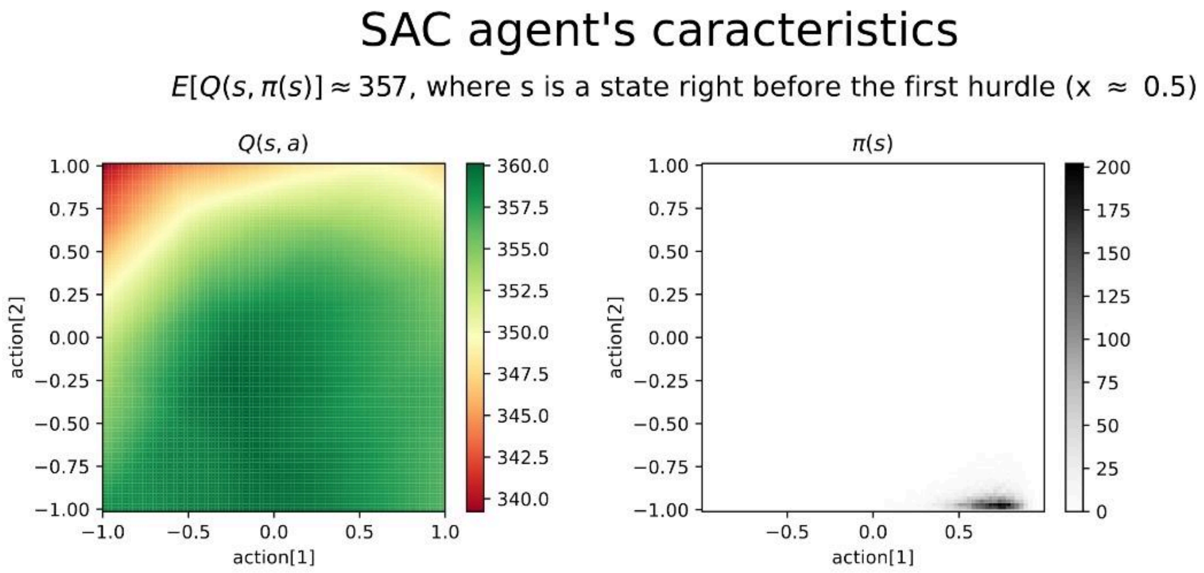
action[2]: Torque applied at the foot of the hopper

a) In all the graphs below, for action[1] +ve and action[2] being -ve provides higher Q value Reasons:
   i) as it helps to go over the hurdle and moving towards right gives higher rewards
   ii) Q-values are lower other actions because it may not be able to go over the hurdle at x ~ 0.5

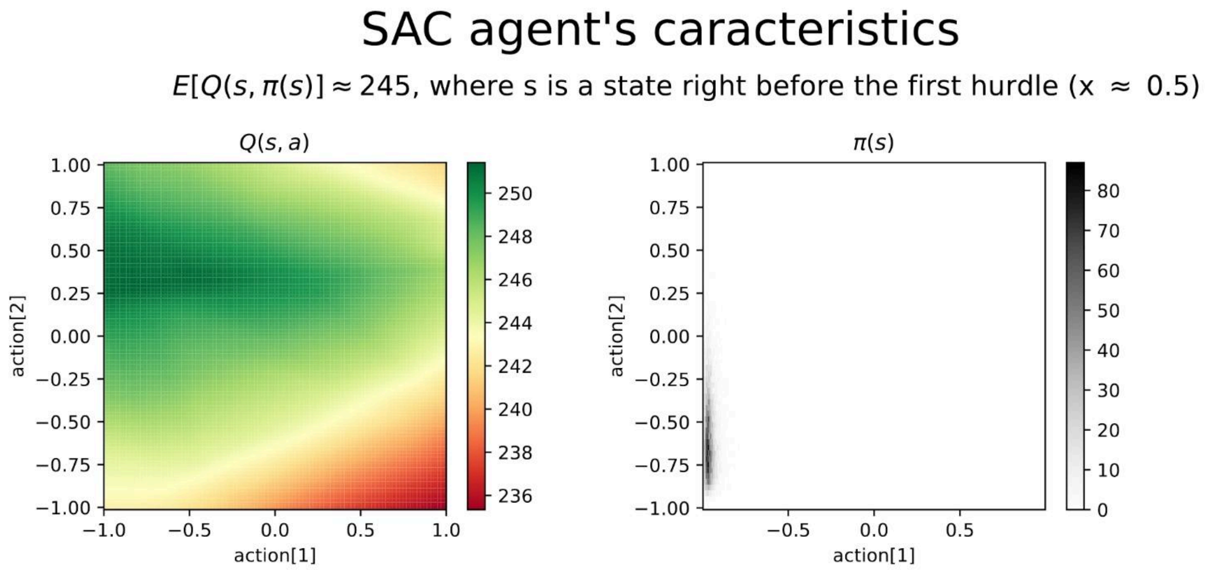
b) Here is the analysis of the all the graphs whether policy is well-trained or not:
   i) hash: 086770cd Yes:
      1) Q value is higher in the region of pi(s).
      2) Could also be seen through E[Q(s, pi(s))]
   ii) hash : 37c63517 No:
      1) Q-value is much lower than other graphs
      2) These actions probably lead to the agent being stuck with this hurdle.
   iii) hash: 3e2b0d0a No, because of similar reasons to ii)
   iv) hash: 485d2053 Yes, because of similar reasons to i). Policy learnt is also very similar.

2. i)

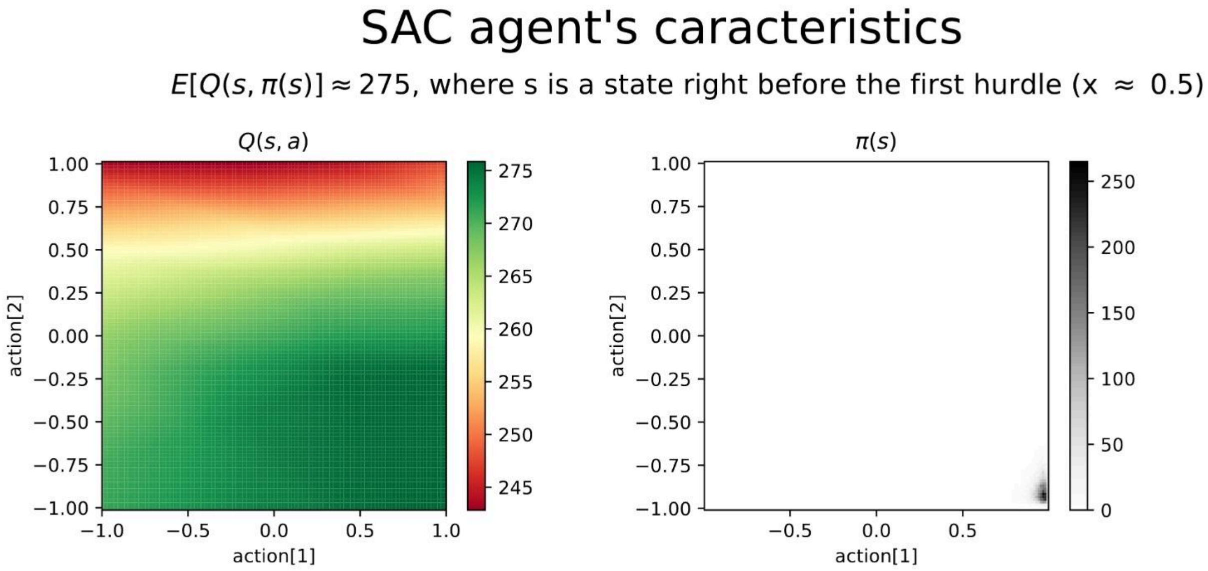
hash: 086770cd



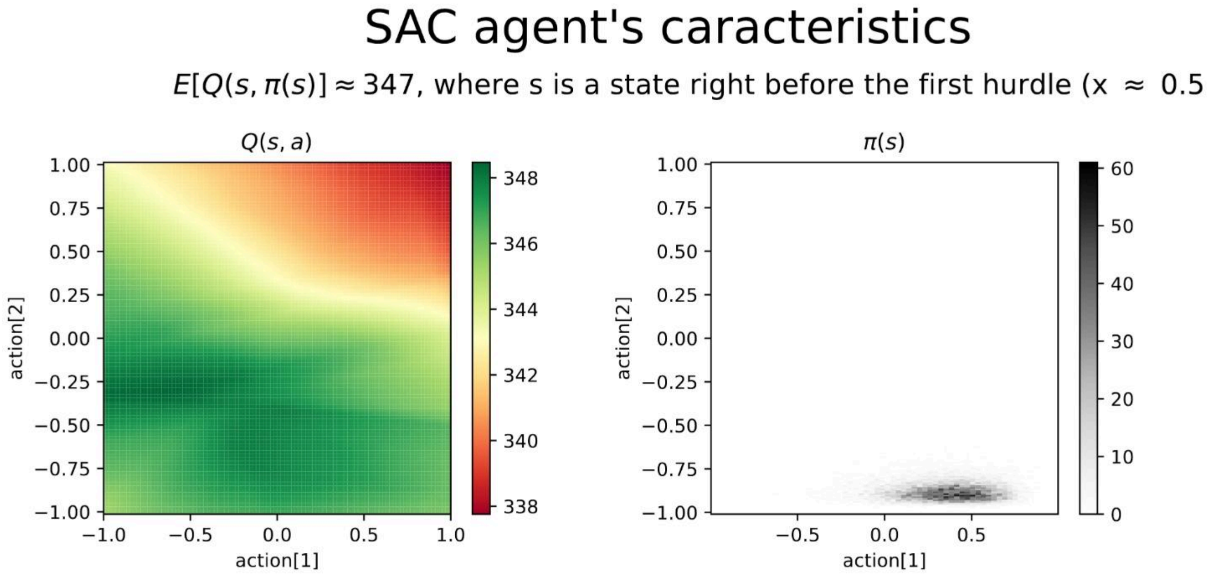
2 ii)
hash : 37c63517



2 iii)
hash: 3e2b0d0a

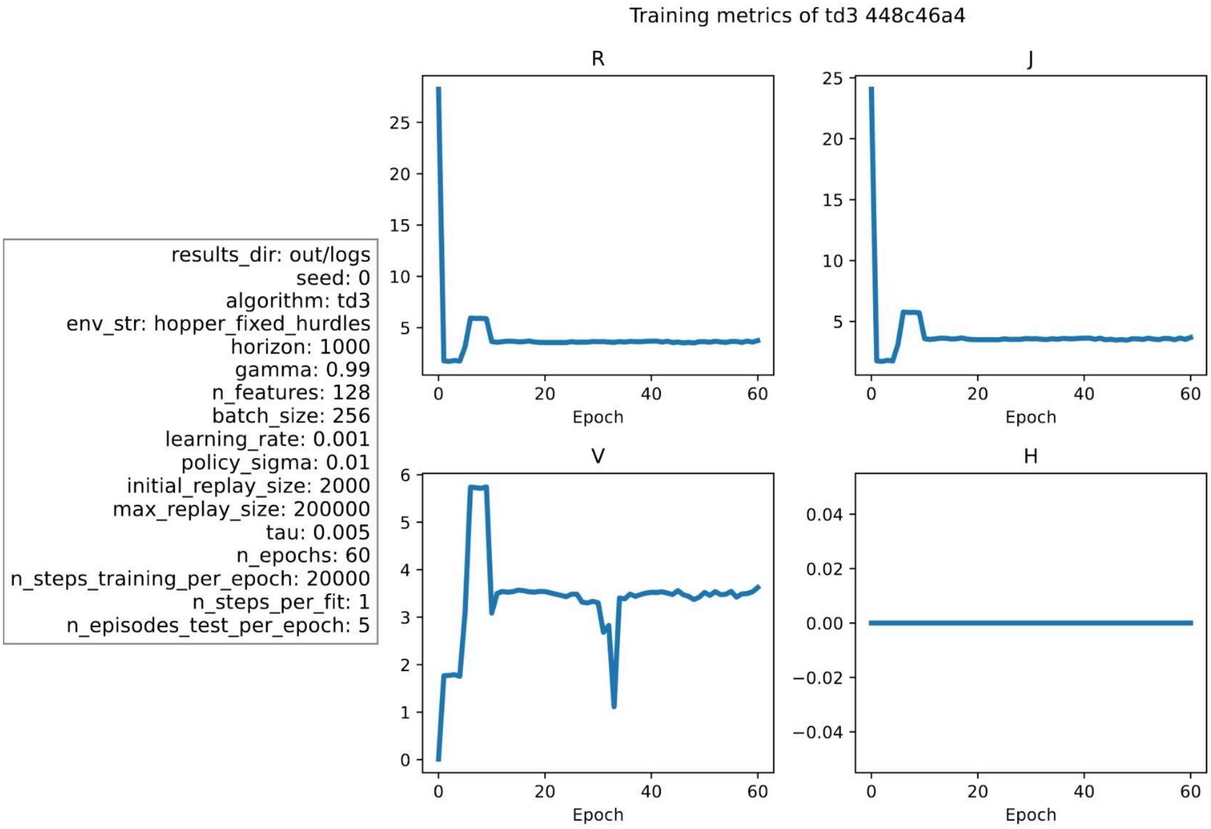


2 iv)
hash: 485d2053



3.4

TD3 without tanh



```
results_dir: out/logs
seed: 0
algorithm: td3
env_str: hopper_fixed_hurdles
horizon: 1000
gamma: 0.99
n_features: 128
batch_size: 256
learning_rate: 0.001
policy_sigma: 0.01
initial_replay_size: 2000
max_replay_size: 200000
tau: 0.005
n_epochs: 60
n_steps_training_per_epoch: 20000
n_steps_per_fit: 1
n_episodes_test_per_epoch: 5
```

Main issue affecting the performance is that action values of the network are not bounded within [-1, 1] and making it difficult for the action network to learn that actions values beyond this range won’t help.