Environment -

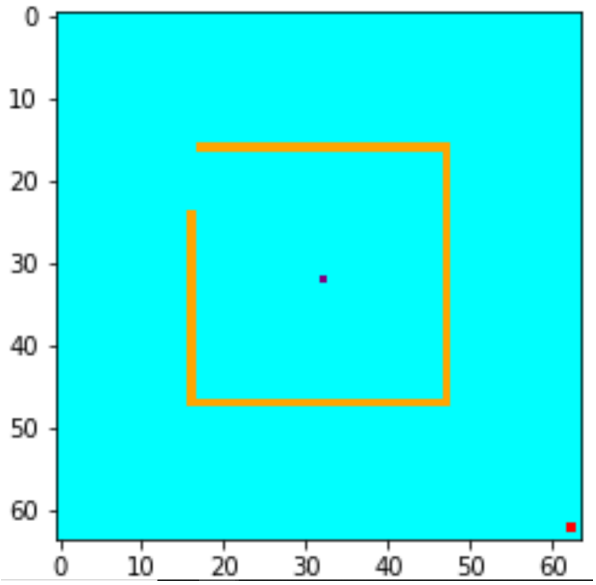
The experiment setup is a gridworld of dimensions 64x64. The environment has a fixed starting point for the agent at location (63,63). There is one terminal state located at the centre at position (32,32). The terminal state is in a mostly enclosed region separated from the rest of the grid through a wall. The figure below shows the environment.

The colors mean the following -

1. Red - agent position (here start position)
2. Purple - Terminal state
3. Orange - Wall
4. Cyan - Navigable grid squares

Rewards -

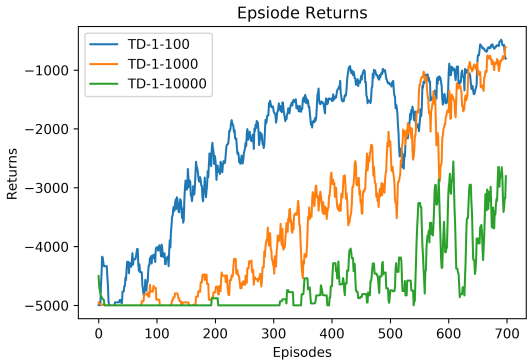
1. For all moves leading to a non-terminal state, there is a reward of -1.
2. For a move into a terminal state, the reward is 0.

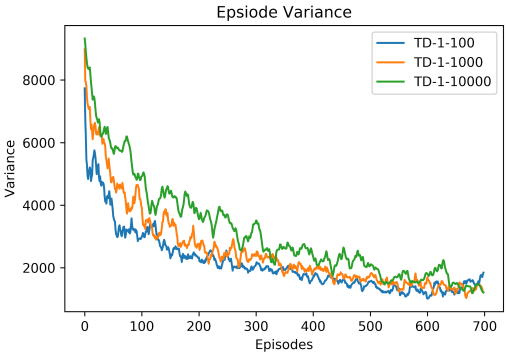


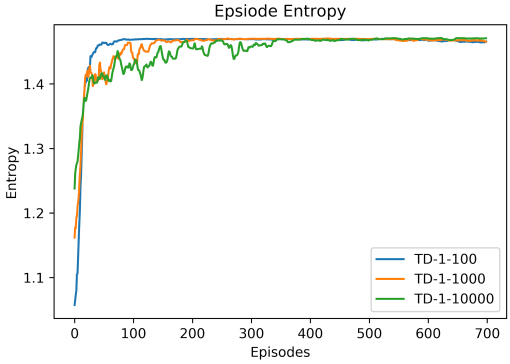
Comparison within uncorrected n-step reward setting -

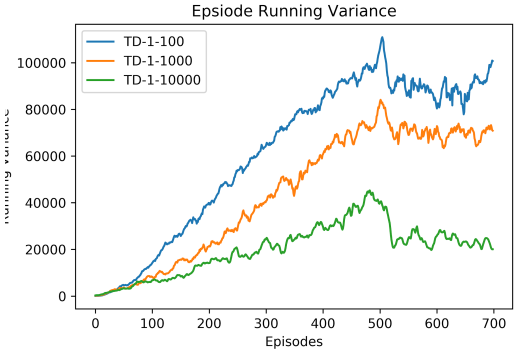
a. Constant n, varying buffer size

1. N = 1

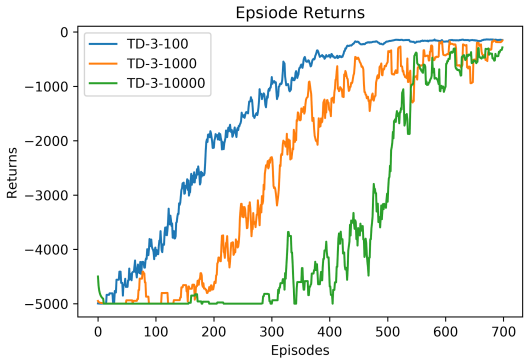


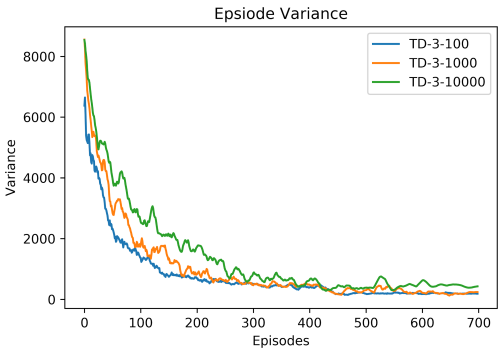


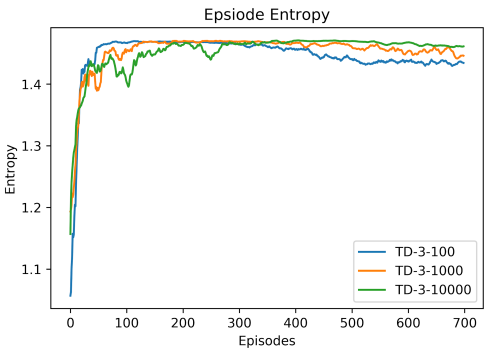


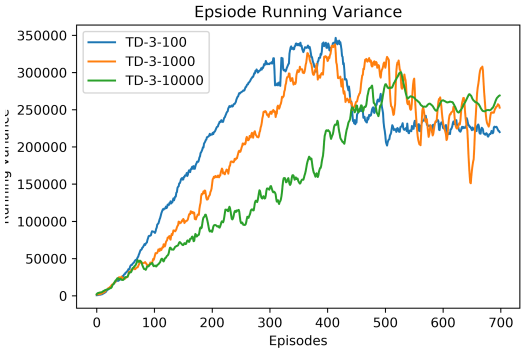


2. N = 3





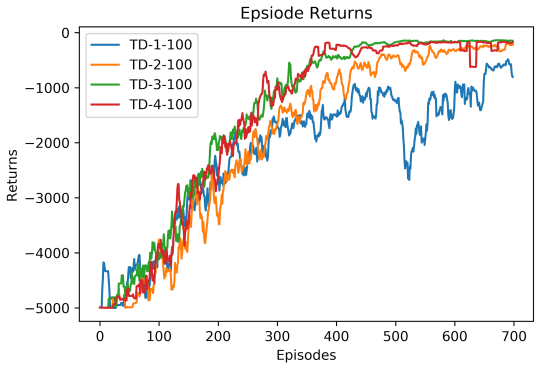


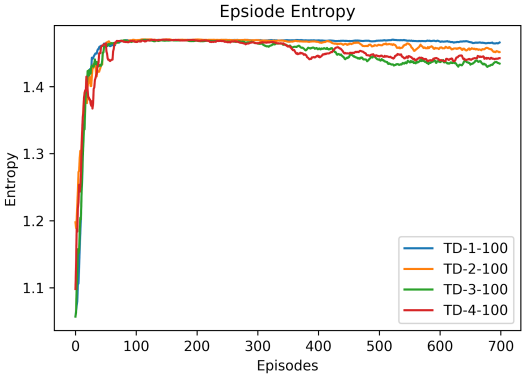


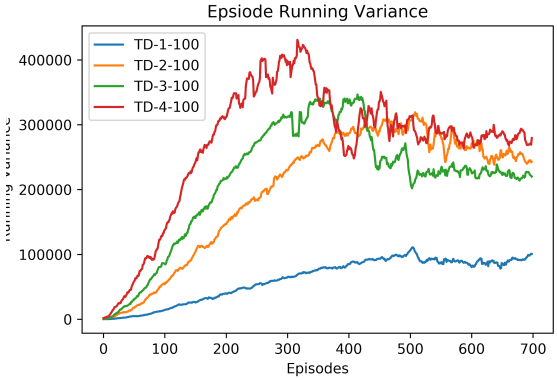
- TRY - REFRESHING
- Larger Replay buffers help mitigate the variance of the target update. This is true irrespective of the length of the n step as the phenomenon is seen even at n=1.
- The running variance decreases as the q values converge to an optimum.
- In the grid world setup all methods guaranteed to converge. However, a larger replay buffer slows this process down. This can viewed from the lens of on policyness as smaller replay buffers guarantee more up to date transitions.
- TODO - Vary the replay capacity: This for less frequent policy updates, the policy should improve for larger replay buffers?

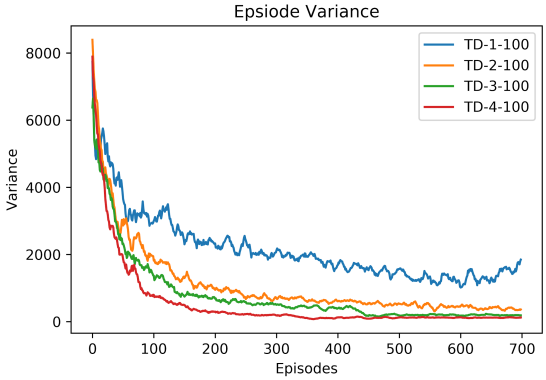
b. Constant Buff size, Varying n

1. Buffer Sz = 100

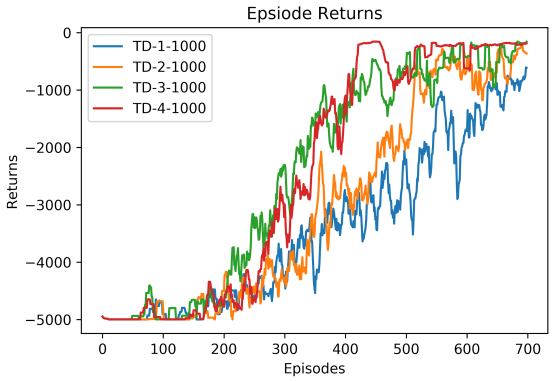


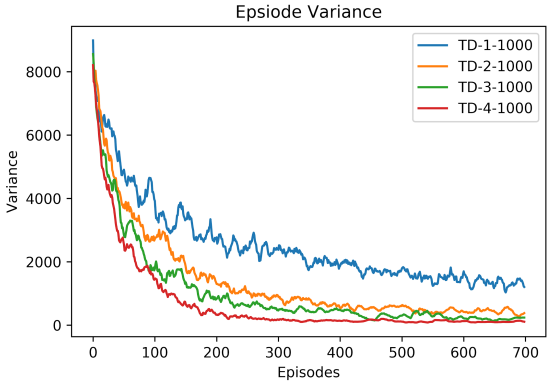


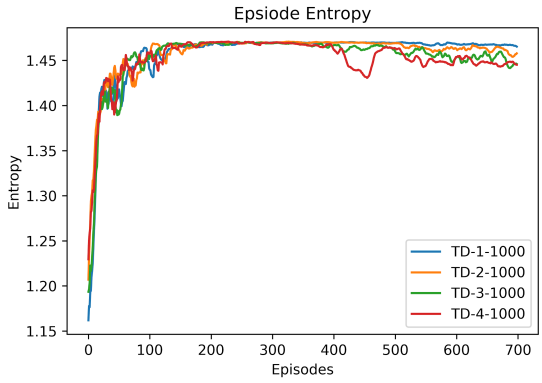


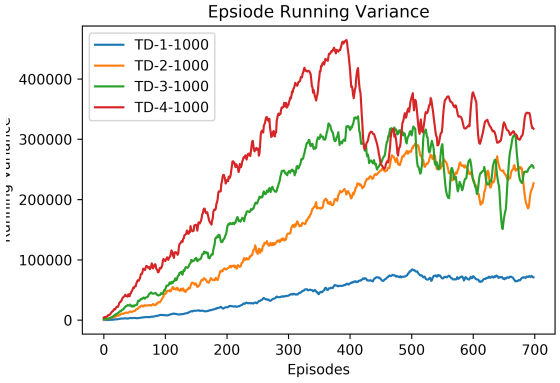


2. Buffer Sz = 1000

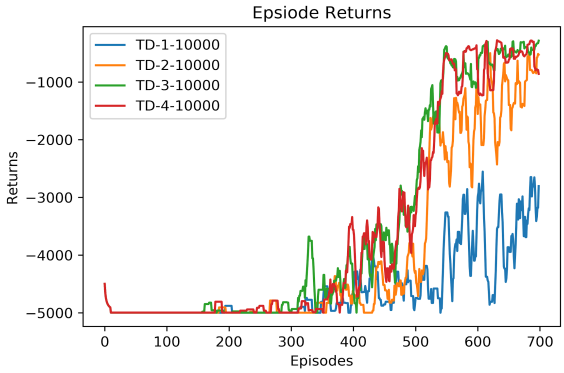


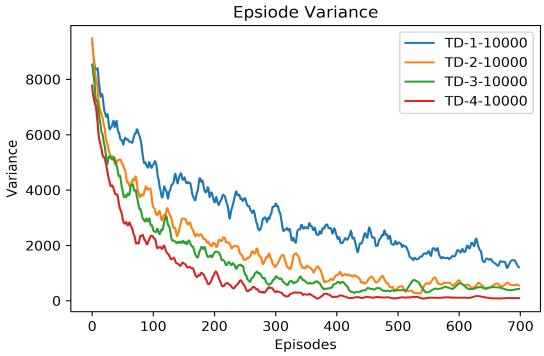


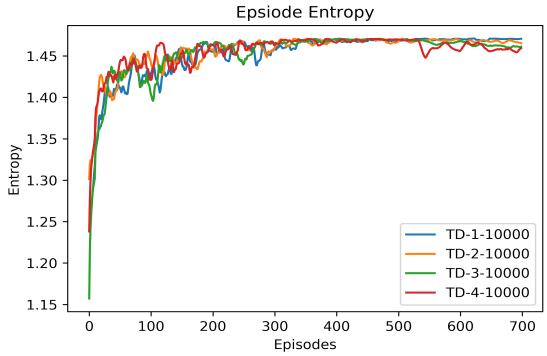


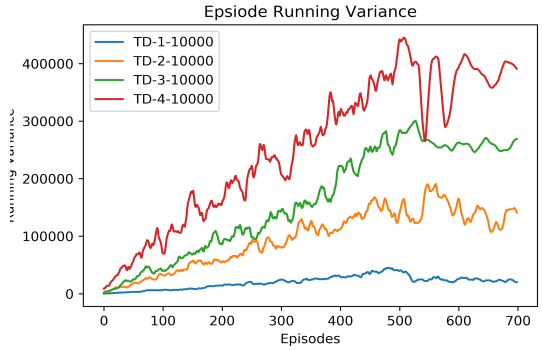


3. Buffer Sz = 10000







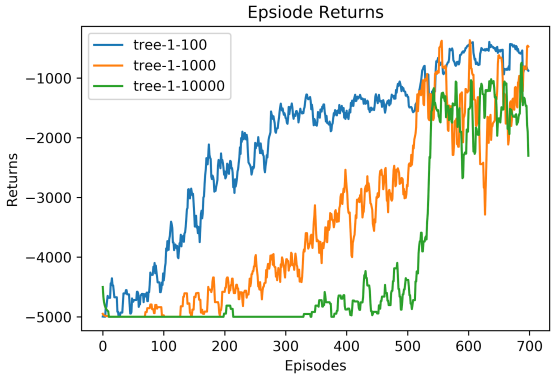


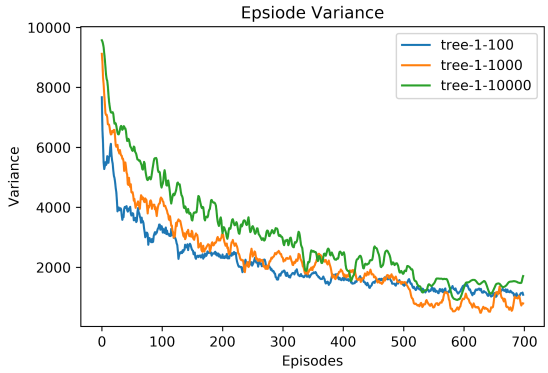
- Larger *n* values lead to faster convergence. This effect diminishes after n=3. This is expected as reward of a terminal state is propagated to more states each time it is encountered until its yield is diminished.
- Larger n values register a sharper drop in variance on approaching an optimal policy / value function.
- While the entropy (in this case of the normalized q value) of state remains constant but registers a slight decrease as the running variance begins to drop.
- QUESTION - In more complex environments, would a flushing heuristic for clearing larger replay buffers demonstrate any visible gains in performance? Could this amount of samples flushed be tied to the running variance graph? Could this be a remedy to catastrophic interference - https://arxiv.org/pdf/2002.12499.pdf

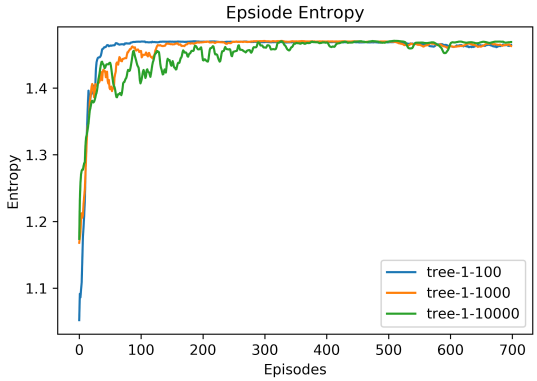
Comparison within n-step-treebackup reward setting -

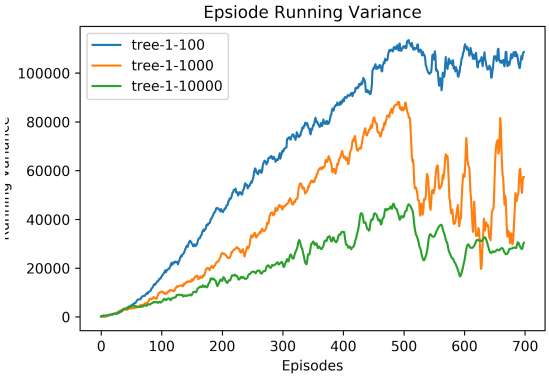
a. Constant n, Varying buffer Size

1. N = 1

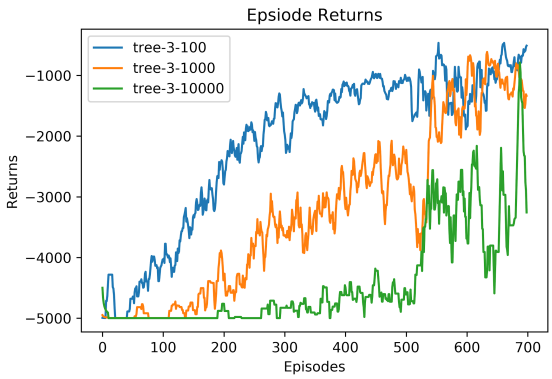


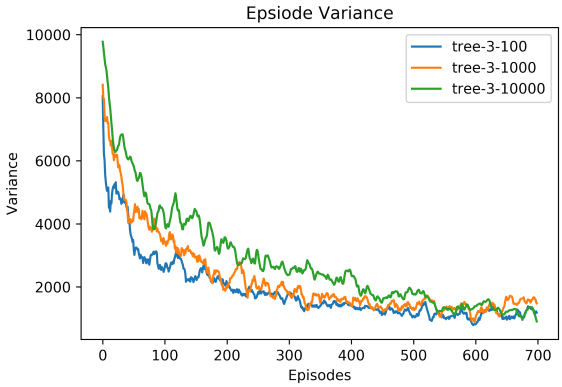


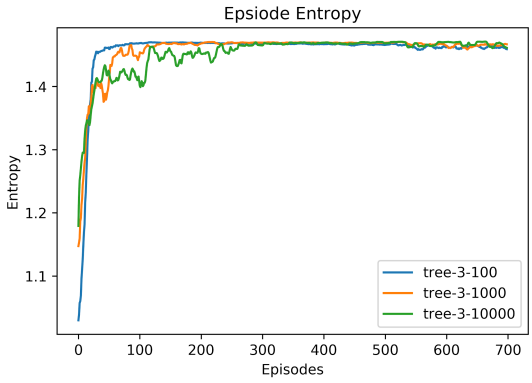


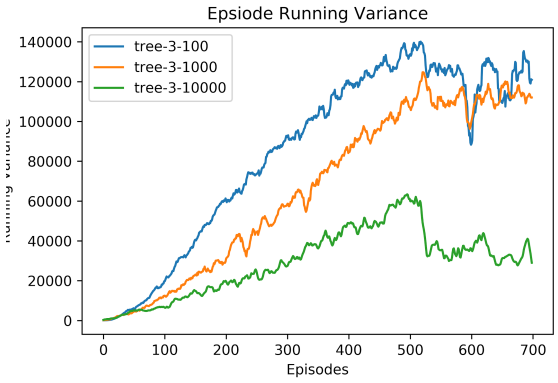


2. N = 3





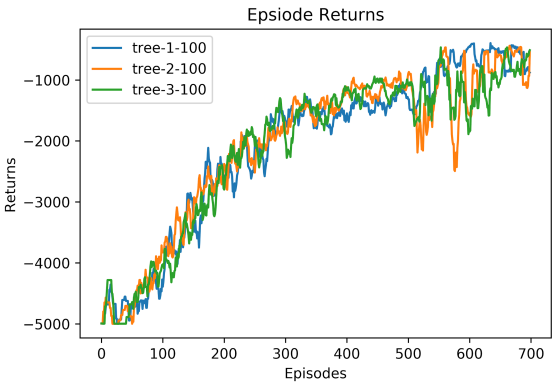


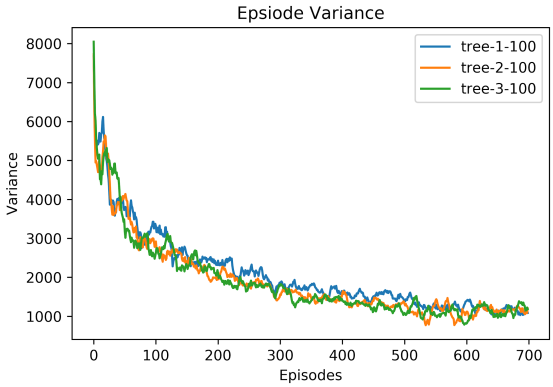


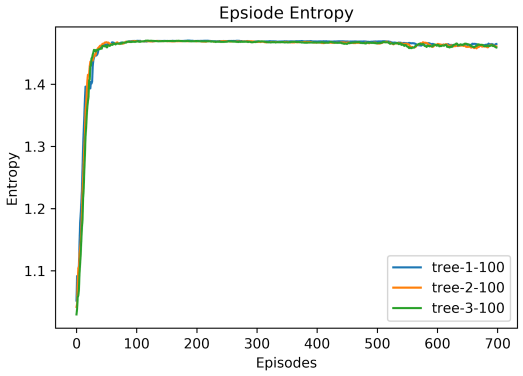
- Tree backup largely mirrors the results of the uncorrected n step, showcasing lower variance with larger buffers albeit at the cost of slower convergence..

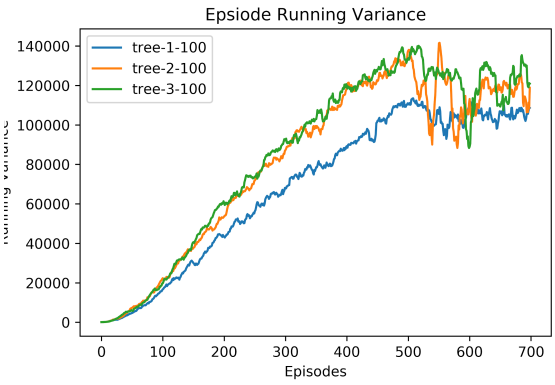
b. Constant Buff size, Varying n

1. Buffer Sz = 100

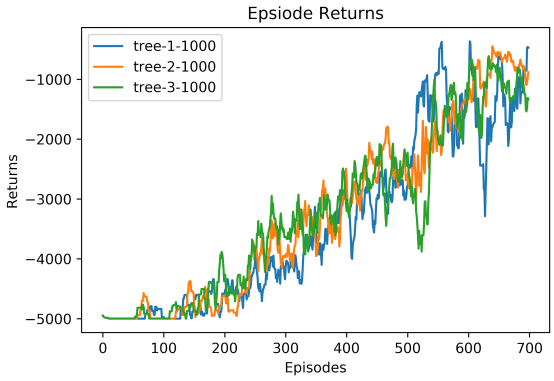


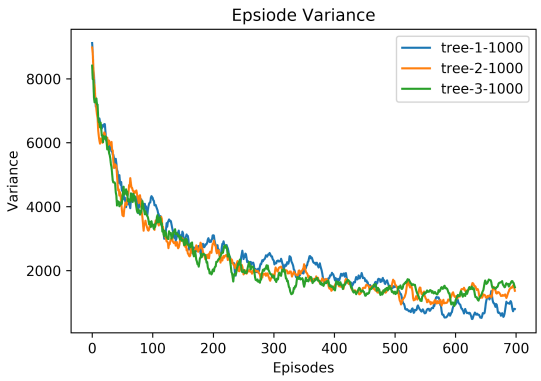


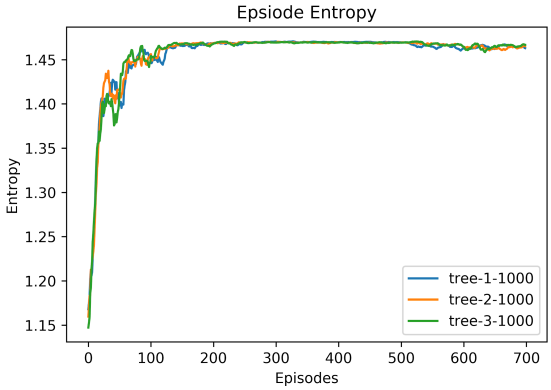


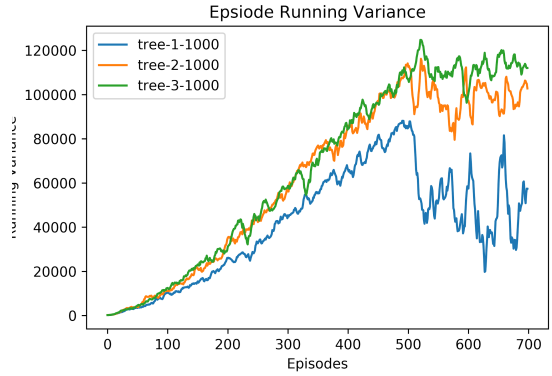


2. Buffer Sz = 1000

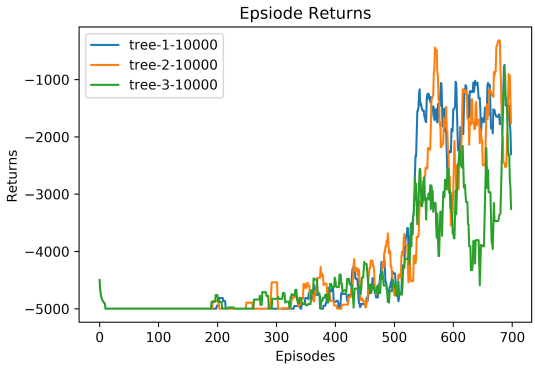


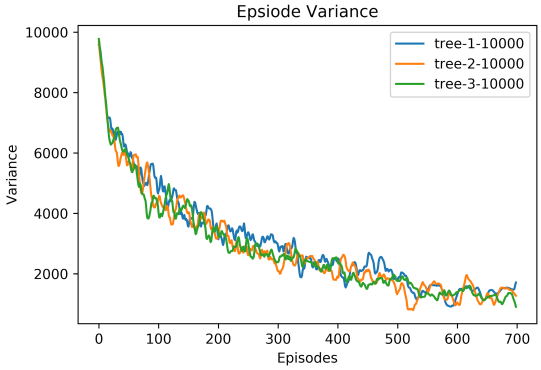


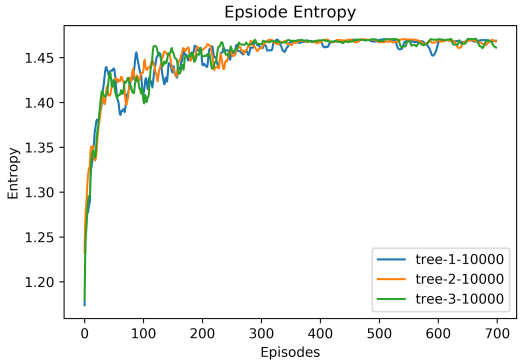


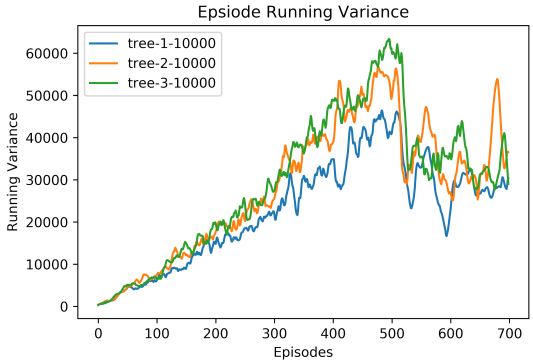


3. Buffer Sz = 10000







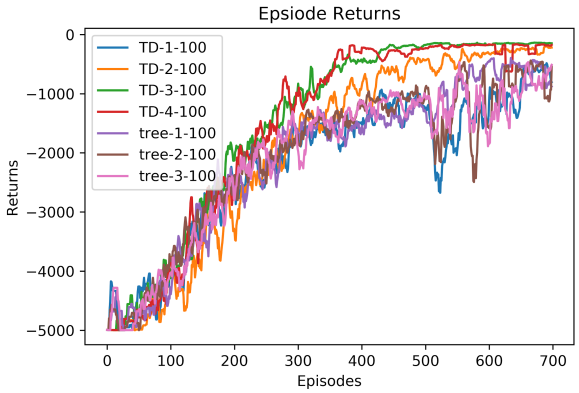


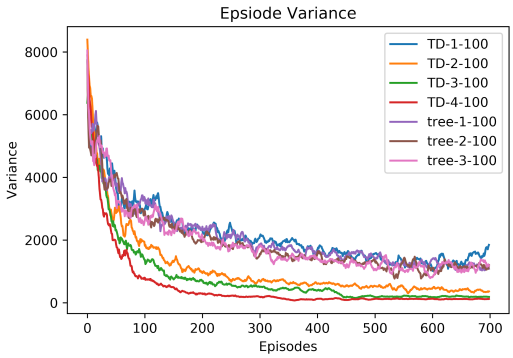
- This section yields interesting results. As hypothesized earlier, the lower variance of the tree backup return as compared to the uncorrected n step return diminishes the gains from a larger replay buffer on moving from a 1 step to a 3 step return.
- While the resulting running variance of a 3 step is higher than a 1 step return the gap is quite small and more importantly, the episode returns are nearly the same.
- TODO - This can now be run on an atari environment to validate the hypothesis further.

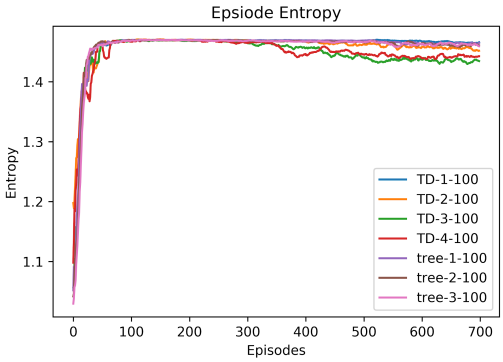
Comparison between uncorrected n-step and treebackup -

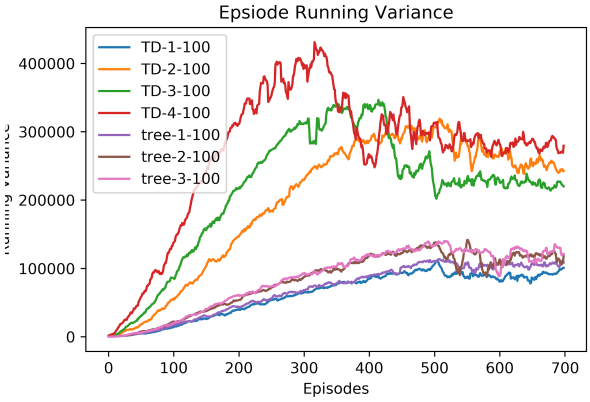
a. Constant Buffer Size

1. Buffer Sz = 100

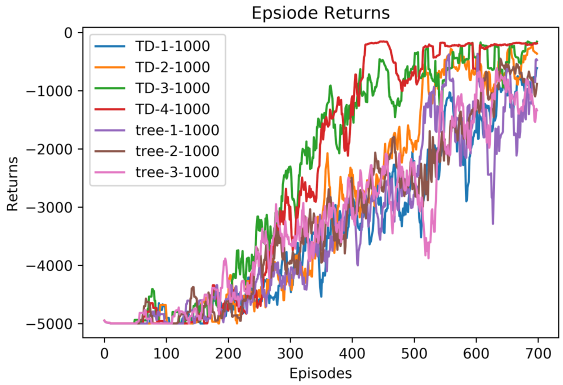


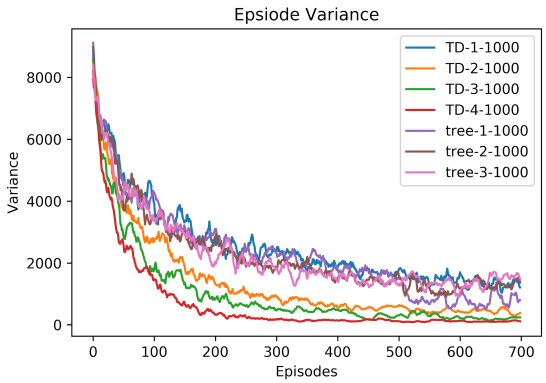


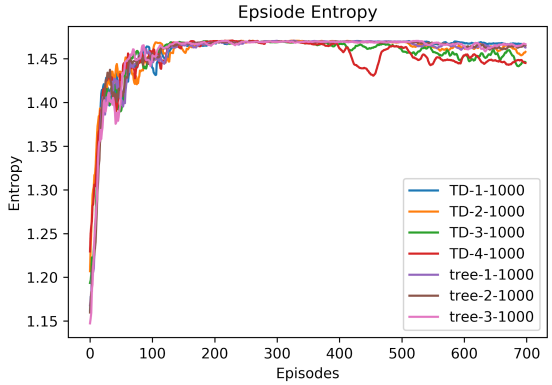


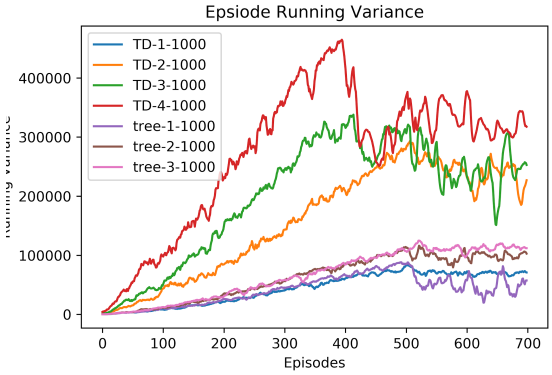


2. Buffer Sz = 1000

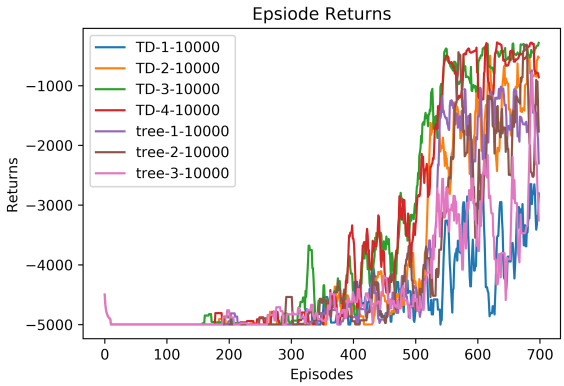


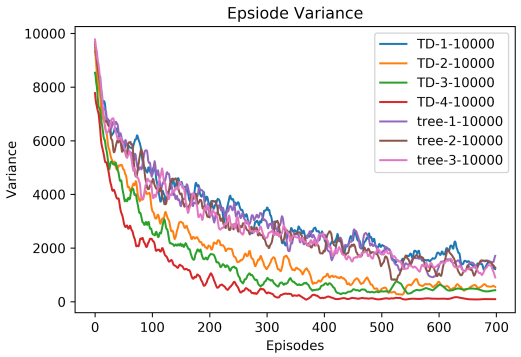


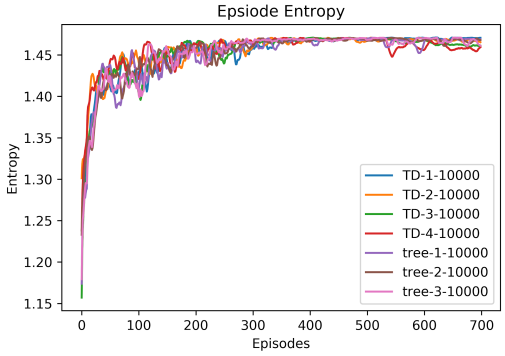


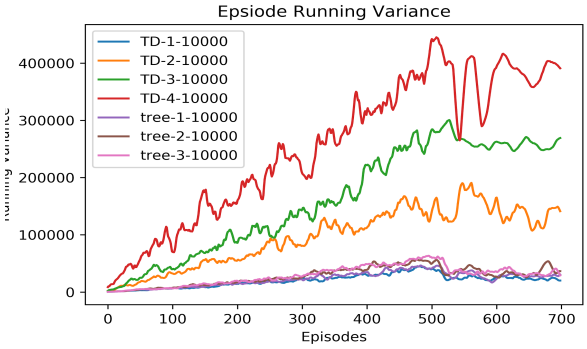


3. Buffer Sz = 10000





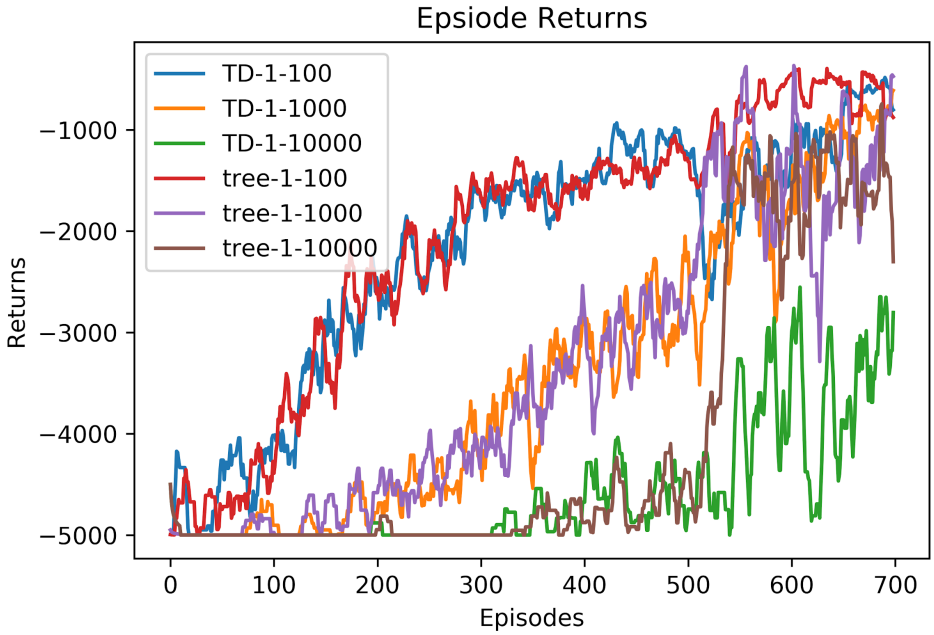


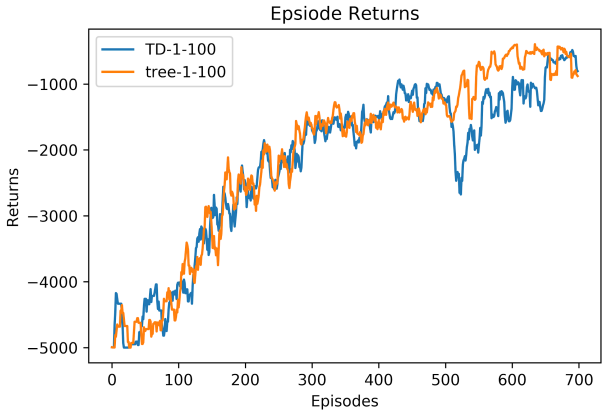


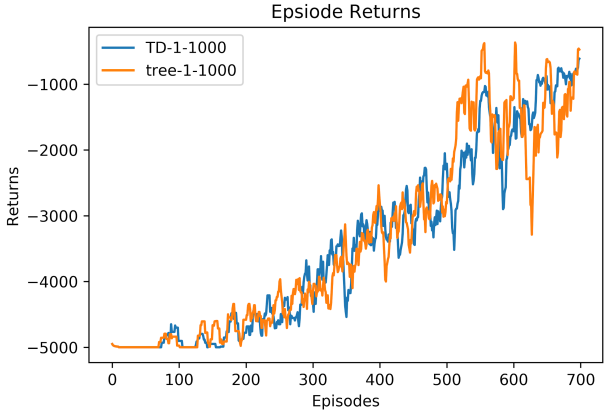
- The initial hypothesis are held - lower variance in tree back vs uncorrected n step.
- This comes at the cost of slower convergence.
- INSIGHT - More variance in the initial stages leads to more exploration. Better not to curb it as it helps in faster convergence.
- TODO - Experiments on larger stochastic environments necessary to determine the maximum achievable rewards across methods. And further reinforce the hypothesis on convergence speed and variance.

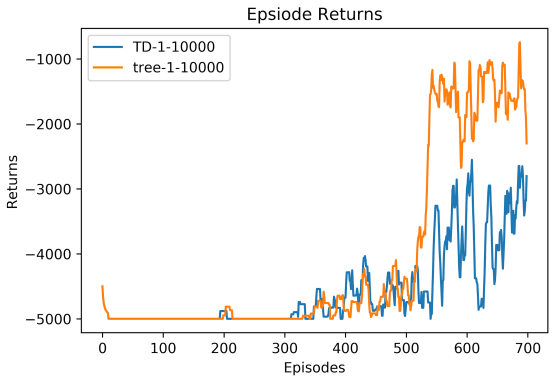
b. Constant n

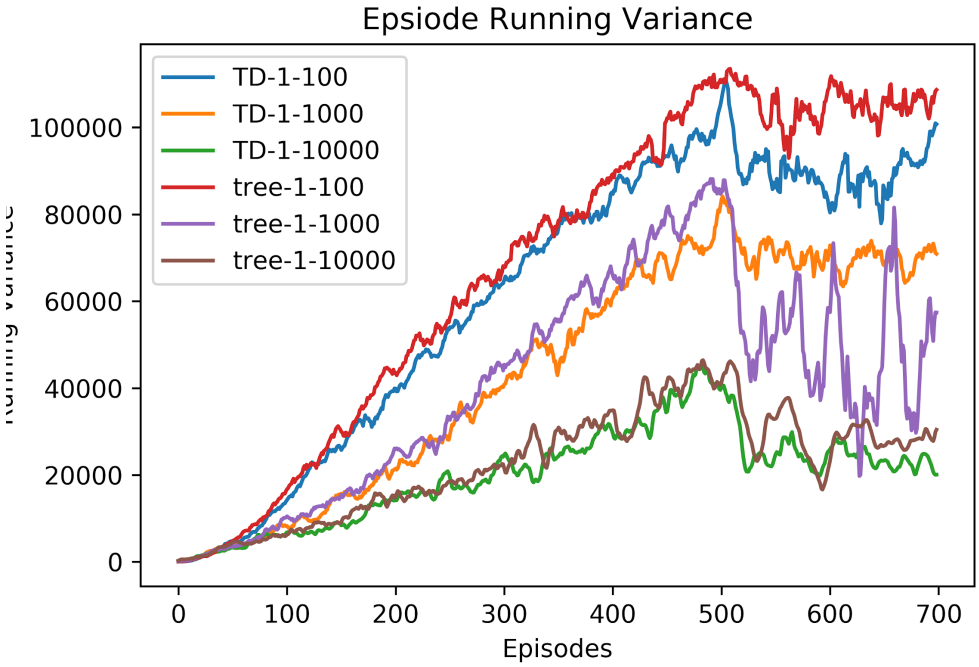
1. N = 1

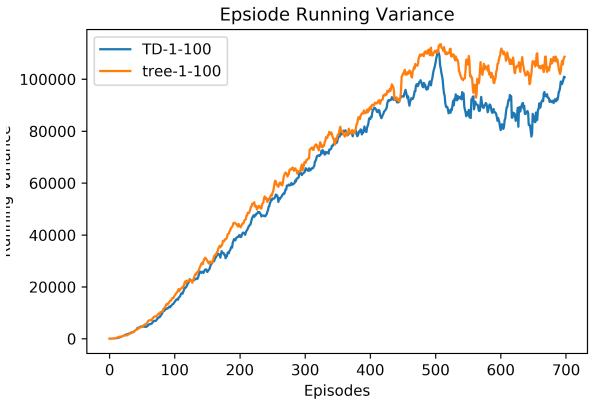

Epsiode Returns
TD-1-100
TD-1-1000
TD-1-10000
tree-1-100
tree-1-1000
tree-1-10000
Returns
−1000
−2000
−3000
−4000
−5000
0
100
200
300
400
500
600
700
Episodes

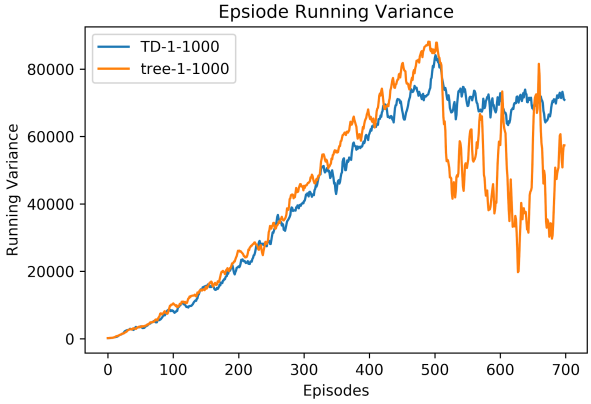

Epsiode Returns
TD-1-100
tree-1-100
Returns
−1000
−2000
−3000
−4000
−5000
0
100
200
300
400
500
600
700
Episodes

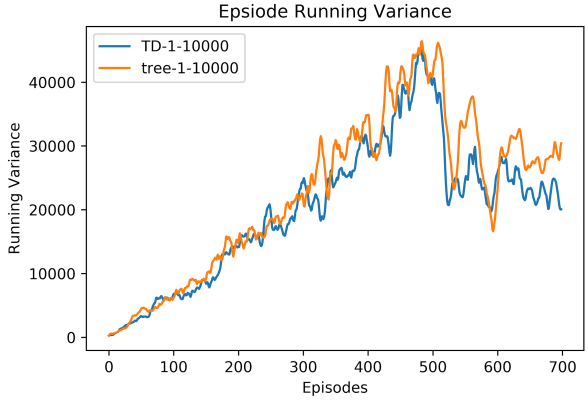

Epsiode Returns
TD-1-1000
tree-1-1000
Returns
−1000
−2000
−3000
−4000
−5000
0
100
200
300
400
500
600
700
Episodes

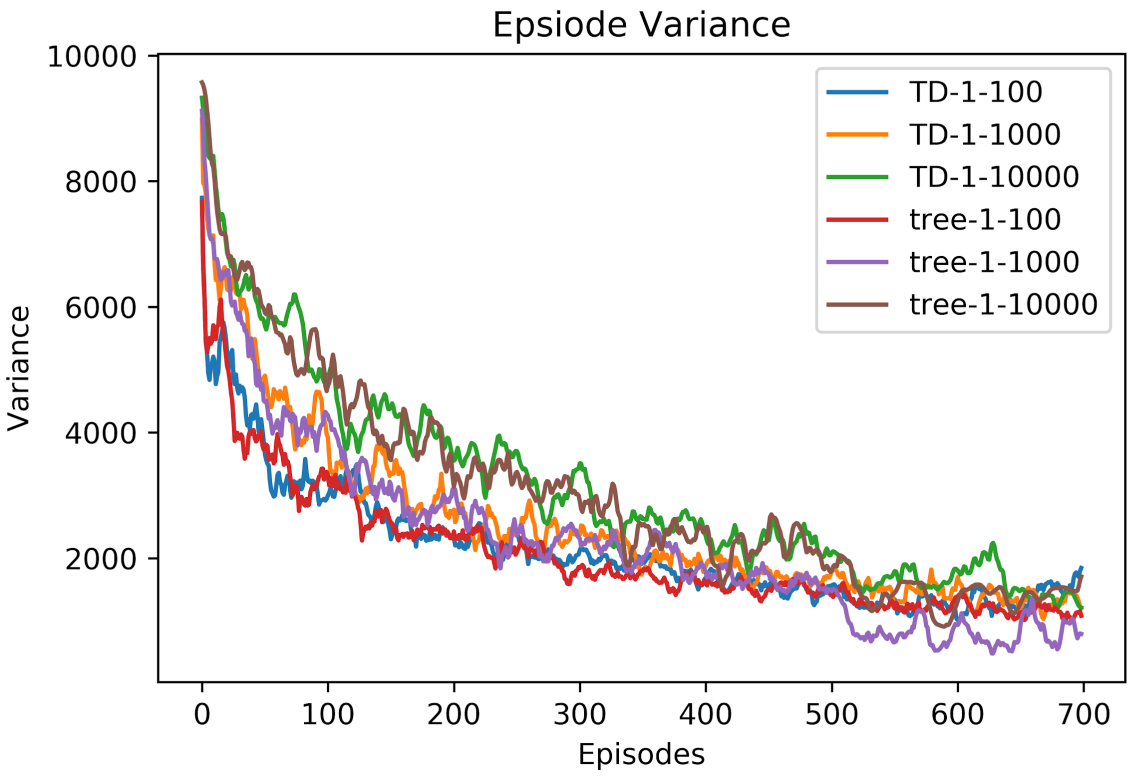

Epsiode Returns
TD-1-10000
tree-1-10000
Returns
−1000
−2000
−3000
−4000
−5000
0
100
200
300
400
500
600
700
Episodes

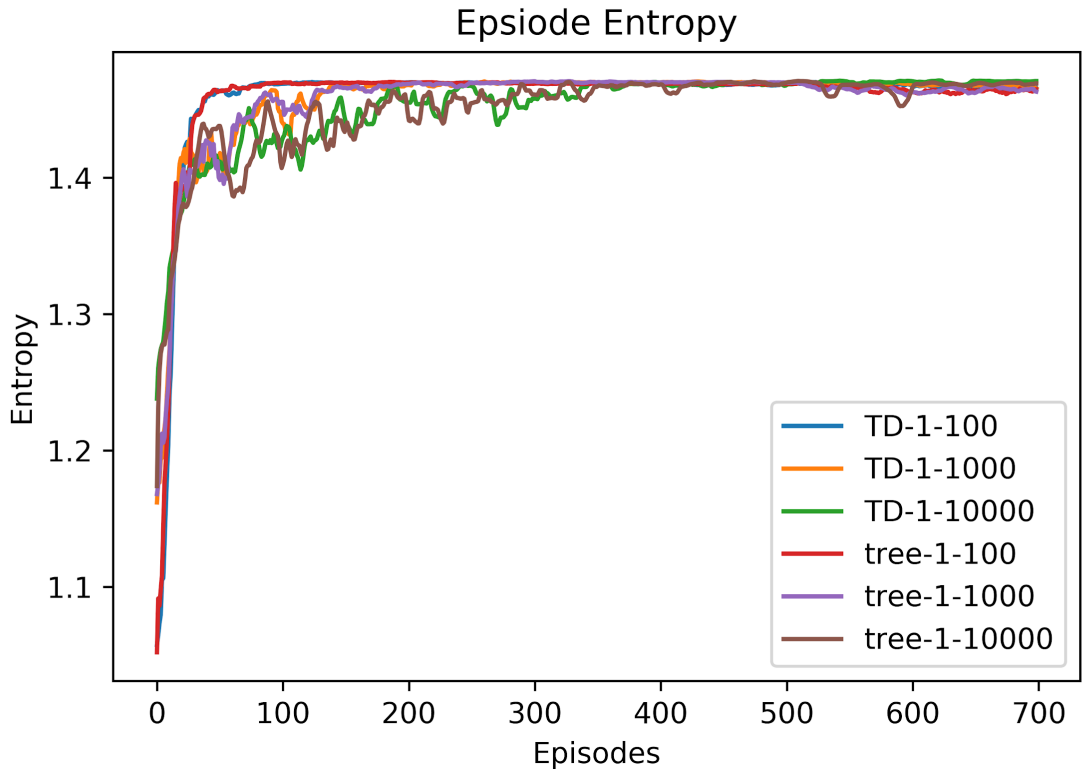

Epsiode Running Variance
TD-1-100
TD-1-1000
TD-1-10000
tree-1-100
tree-1-1000
tree-1-10000
100000
80000
60000
40000
20000
0
0
100
200
300
400
500
600
700
Episodes


Epsiode Running Variance
TD-1-100
tree-1-100
100000
80000
60000
40000
20000
0
0
100
200
300
400
500
600
700
Episodes

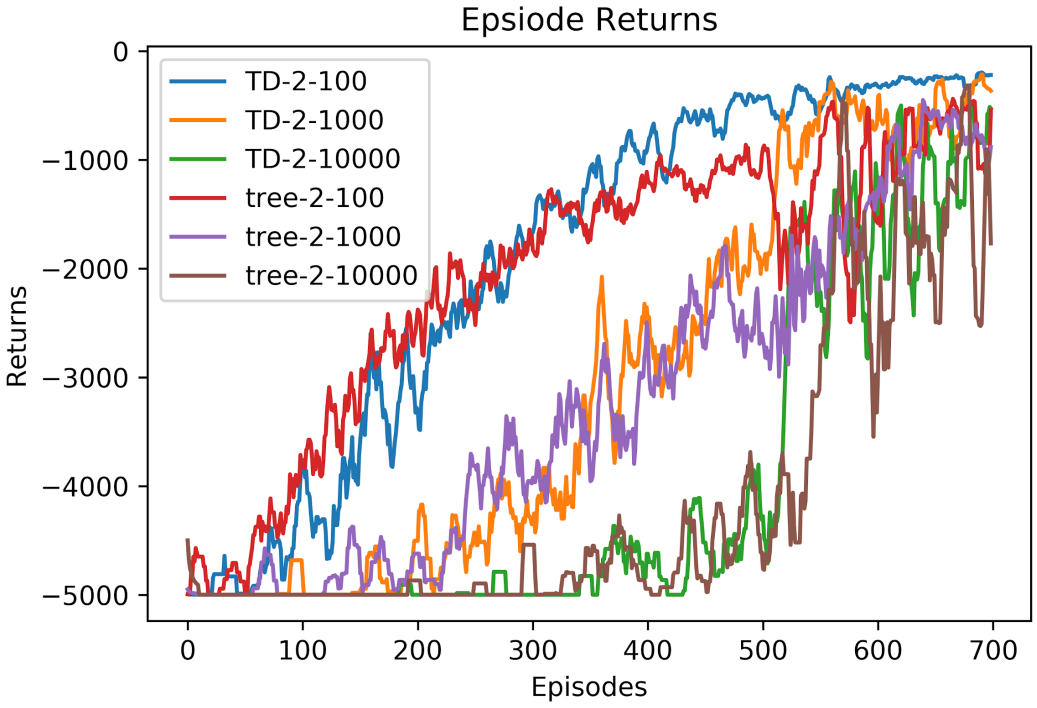

Epsiode Running Variance
TD-1-1000
tree-1-1000
Running Variance
80000
60000
40000
20000
0
0
100
200
300
400
500
600
700
Episodes

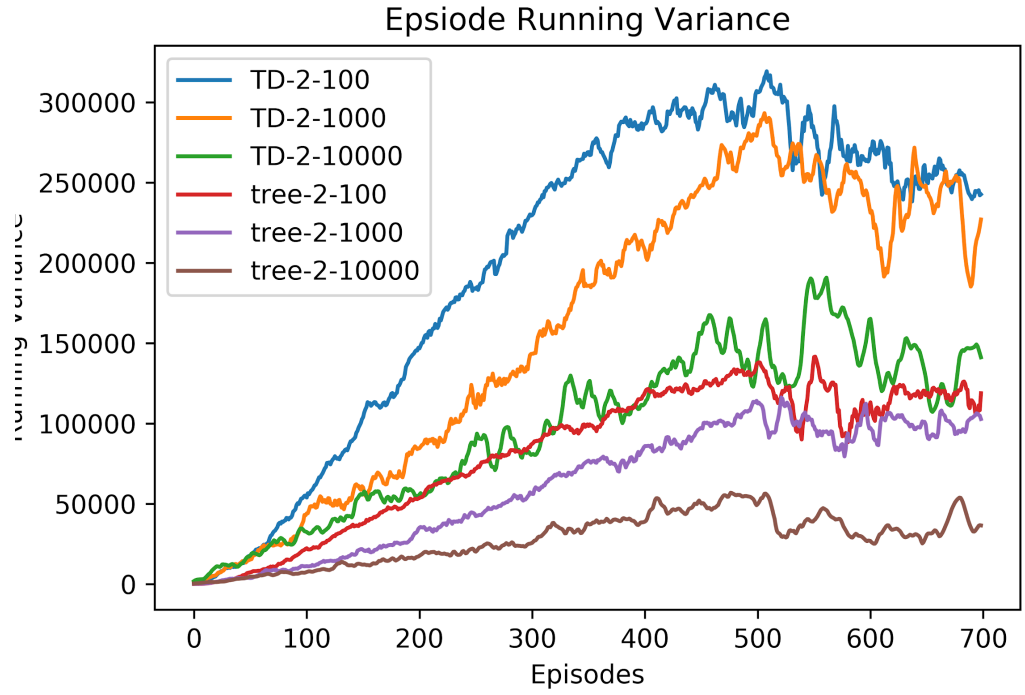

Epsiode Running Variance
TD-1-10000
tree-1-10000
Running Variance
40000
30000
20000
10000
0
0
100
200
300
400
500
600
700
Episodes

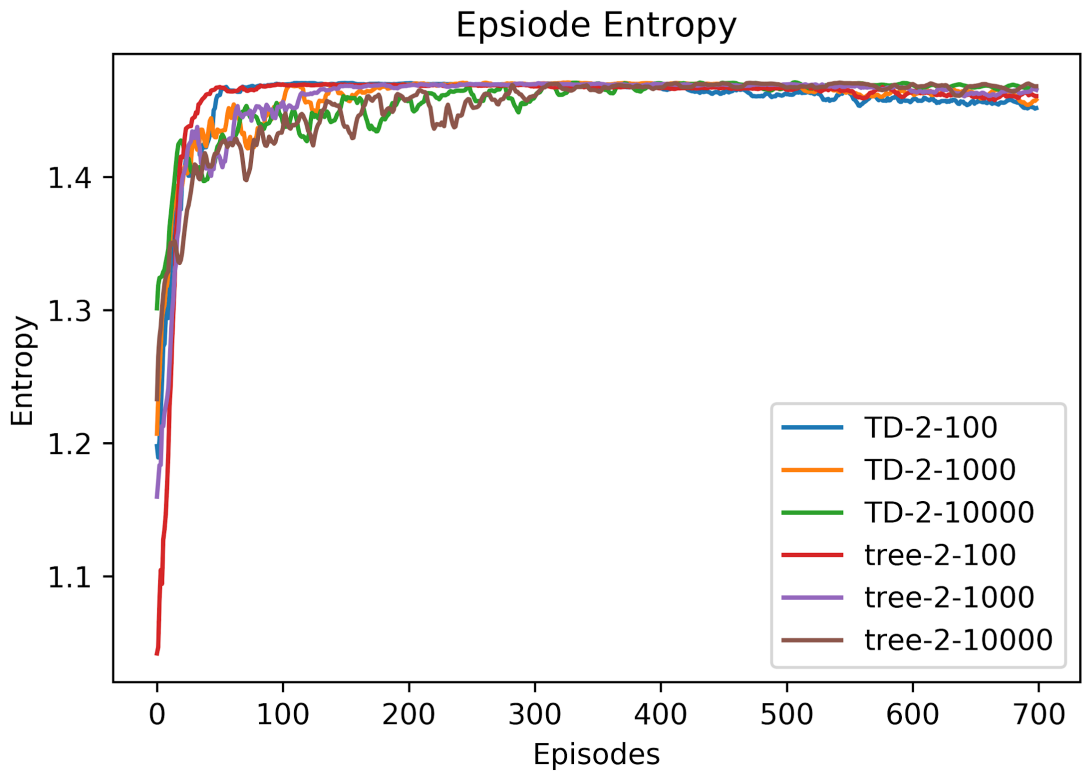

Epsiode Variance
Variance
Episodes
TD-1-100
TD-1-1000
TD-1-10000
tree-1-100
tree-1-1000
tree-1-10000
10000
8000
6000
4000
2000
0
100
200
300
400
500
600
700

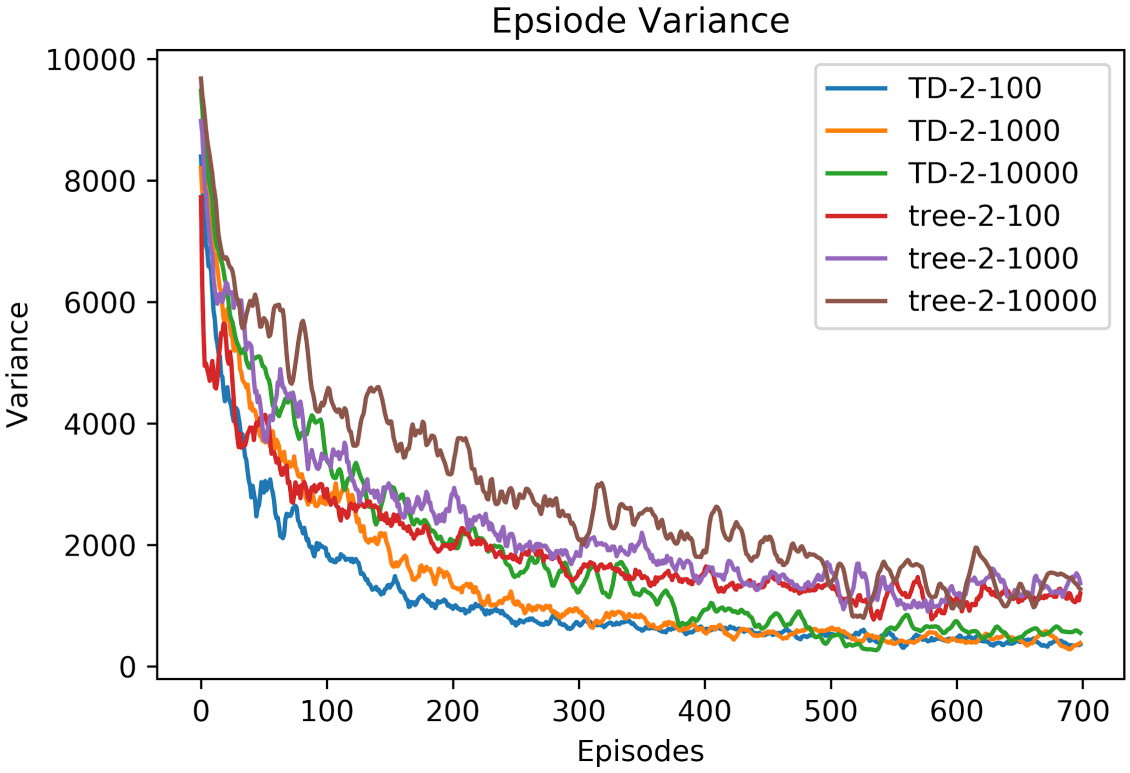

Epsiode Entropy
Entropy
Episodes
TD-1-100
TD-1-1000
TD-1-10000
tree-1-100
tree-1-1000
tree-1-10000
1.4
1.3
1.2
1.1
0
100
200
300
400
500
600
700

2. N = 2

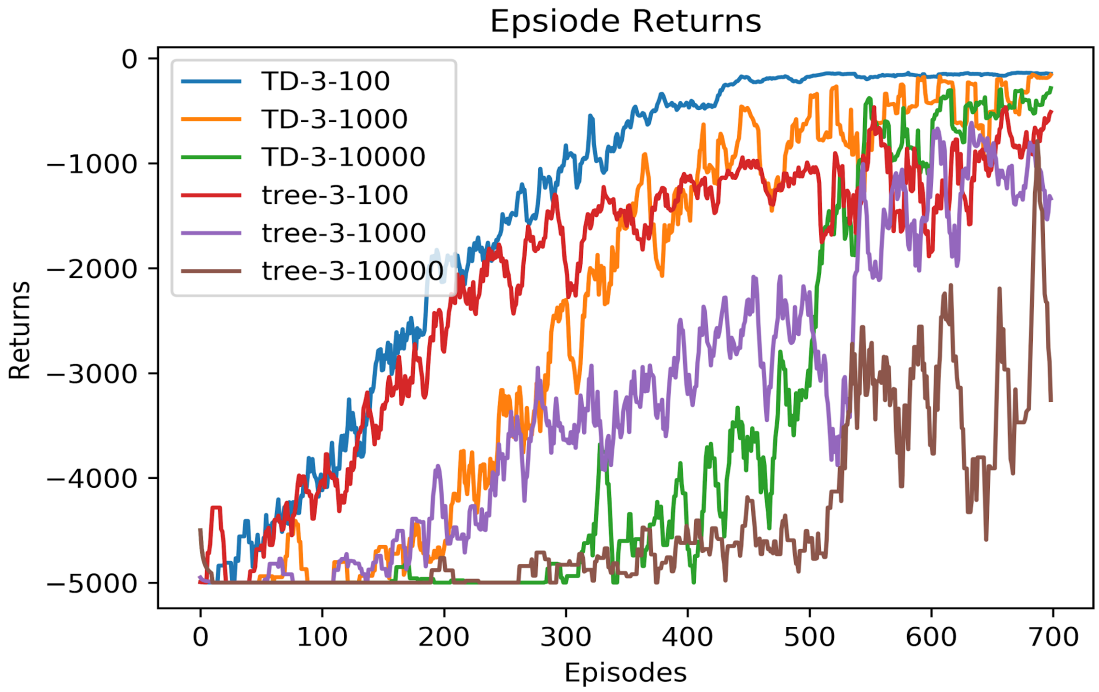

Epsiode Returns
TD-2-100
TD-2-1000
TD-2-10000
tree-2-100
tree-2-1000
tree-2-10000
Returns
0
−1000
−2000
−3000
−4000
−5000
0
100
200
300
400
500
600
700
Episodes

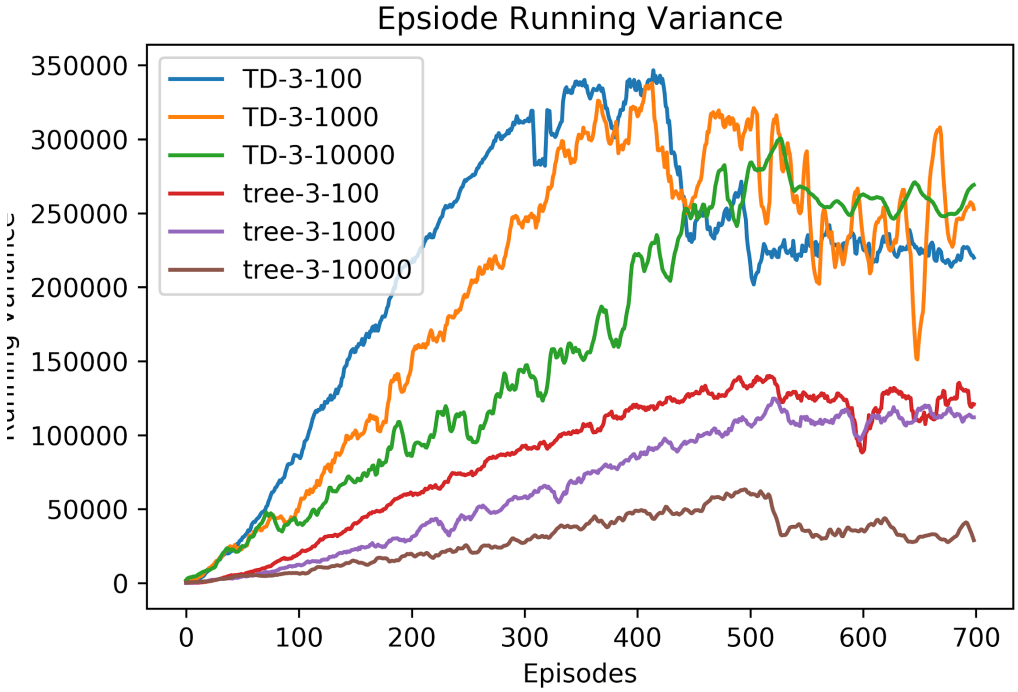

Epsiode Running Variance
TD-2-100
TD-2-1000
TD-2-10000
tree-2-100
tree-2-1000
tree-2-10000
300000
250000
200000
150000
100000
50000
0
0
100
200
300
400
500
600
700
Episodes

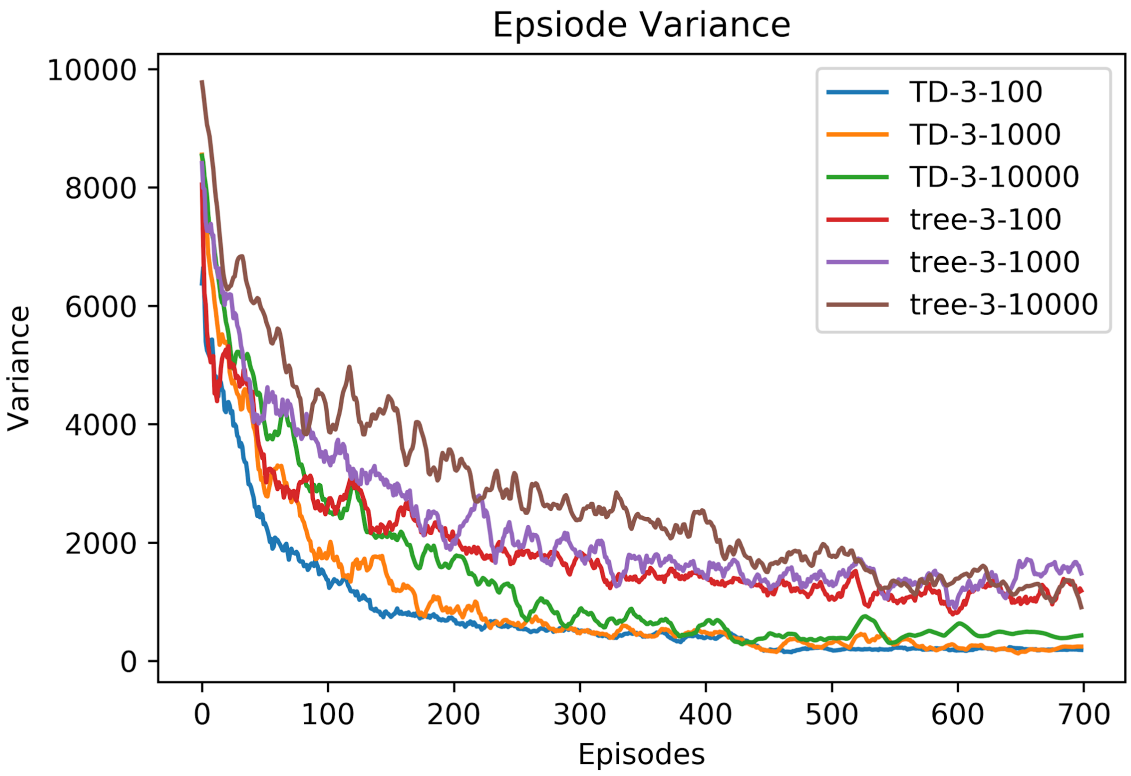

Epsiode Entropy
Entropy
1.4
1.3
1.2
1.1
0
100
200
300
400
500
600
700
Episodes
TD-2-100
TD-2-1000
TD-2-10000
tree-2-100
tree-2-1000
tree-2-10000

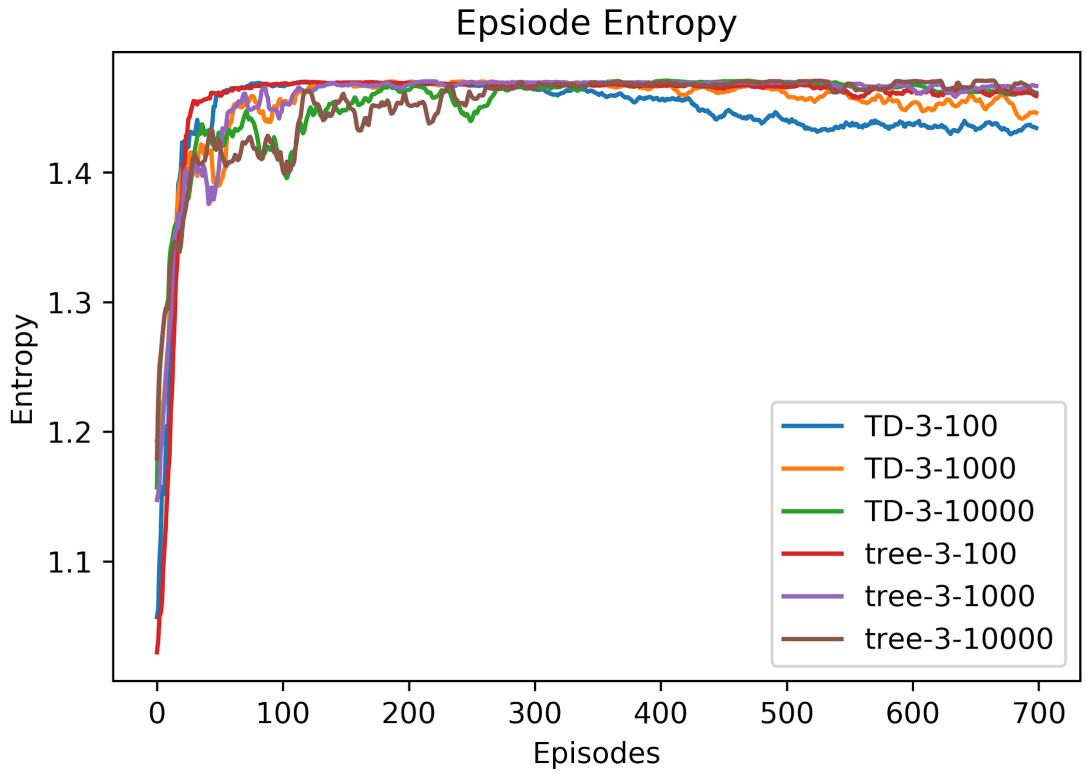

Epsiode Variance
Variance
10000
8000
6000
4000
2000
0
0
100
200
300
400
500
600
700
Episodes
TD-2-100
TD-2-1000
TD-2-10000
tree-2-100
tree-2-1000
tree-2-10000

3. N = 3


Epsiode Returns
TD-3-100
TD-3-1000
TD-3-10000
tree-3-100
tree-3-1000
tree-3-10000
Returns
0
−1000
−2000
−3000
−4000
−5000
0
100
200
300
400
500
600
700
Episodes


Epsiode Running Variance
TD-3-100
TD-3-1000
TD-3-10000
tree-3-100
tree-3-1000
tree-3-10000
350000
300000
250000
200000
150000
100000
50000
0
0
100
200
300
400
500
600
700
Episodes


Epsiode Variance
Variance
Episodes
TD-3-100
TD-3-1000
TD-3-10000
tree-3-100
tree-3-1000
tree-3-10000
0
2000
4000
6000
8000
10000
0
100
200
300
400
500
600
700


Epsiode Entropy
Entropy
Episodes
TD-3-100
TD-3-1000
TD-3-10000
tree-3-100
tree-3-1000
tree-3-10000
1.1
1.2
1.3
1.4
0
100
200
300
400
500
600
700

- Larger n appear to slow down tree backup in comparison to uncorrected n step. The reduced variance appears to play a role in this.

Dissimilarity of transition
Intrinsic motivation