Appendix

Proof of Lemma 1

For ease of exposition in this derivation we shall use the notation $\mathbf{E}_t(\cdot)$ to denote the expectation of a random variable conditioned on the history $\mathcal{F}_t$, rather than the usual $\mathbf{E}(\cdot|\mathcal{F}_t)$.

Recall that if we sample $\hat{\mu}$ from the posterior over the mean rewards $\phi_{\mu|\mathcal{F}_t}$, and $\hat{P}$ from the posterior over the transition probability matrix $\phi_{P|\mathcal{F}_t}$ then the Q-values that are the unique solution to

$$\hat{Q}^\pi_{sa} = \hat{\mu}_{sa} + \gamma \sum_{s'} \pi_{s'a'} \hat{P}_{s'sa} \hat{Q}^\pi_{s'a'}$$

are a sample from the (implicit) posterior over Q-values, conditioned on $\mathcal{F}_t$ (Strens, 2000). Using the definition of the conditional variance

$$\begin{aligned}
\mathbf{var}_t \hat{Q}^\pi_{sa} &= \mathbf{E}_t \left(\hat{Q}^\pi_{sa} - \mathbf{E}_t \hat{Q}^\pi_{sa} \right)^2 \\
&= \mathbf{E}_t \left(\hat{\mu}_{sa} - \mathbf{E}_t \hat{\mu}_{sa} + \gamma \textstyle\sum_{s',a'} \pi_{s'a'} (\hat{P}_{s'sa} \hat{Q}^\pi_{s'a'} - \mathbf{E}_t \hat{P}_{s'sa} \hat{Q}^\pi_{s'a'}) \right)^2 \\
&= \mathbf{E}_t \left(\hat{\mu}_{sa} - \mathbf{E}_t \hat{\mu}_{sa} \right)^2 + \gamma^2 \mathbf{E}_t \left(\textstyle\sum_{s',a'} \pi_{s'a'} (\hat{P}_{s'sa} \hat{Q}^\pi_{s'a'} - \mathbf{E}_t \hat{P}_{s'sa} \hat{Q}^\pi_{s'a'}) \right)^2 \\
&\leq \mathbf{var}_t \hat{\mu}_{sa} + \gamma^2 \textstyle\sum_{s',a'} \pi_{s'a'} \mathbf{E}_t \left(\hat{P}_{s'sa} \hat{Q}^\pi_{s'a'} - \mathbf{E}_t \hat{P}_{s'sa} \hat{Q}^\pi_{s'a'} \right)^2
\end{aligned}$$

where we have used the fact that $\hat{\mu}_{sa}$ is conditionally independent (conditioned on $\mathcal{F}_t$) of $\hat{P}_{s'sa} \hat{Q}^\pi_{s'a'}$ because assumption 1 implies that $\hat{Q}_{s'a'}$ depends only on downstream quantities, and in the last line we used Jensen's inequality. Assumption 1 also implies that that $\hat{P}_{s'sa}$ and $\hat{Q}_{s'a'}$ are conditionally independent, and so we can write the expectation in the last term above as

$$\begin{aligned}
\mathbf{E}_t \left(\hat{P}_{s'sa} \hat{Q}^\pi_{s'a'} - \mathbf{E}_t \hat{P}_{s'sa} \hat{Q}^\pi_{s'a'} \right)^2 &= \mathbf{E}_t \left((\hat{P}_{s'sa} - \mathbf{E}_t \hat{P}_{s'sa}) \hat{Q}^\pi_{s'a'} + (\mathbf{E}_t \hat{P}_{s'sa})(\hat{Q}^\pi_{s'a'} - \mathbf{E}_t \hat{Q}^\pi_{s'a'}) \right)^2 \\
&= \mathbf{E}_t \left((\hat{P}_{s'sa} - \mathbf{E}_t \hat{P}_{s'sa}) \hat{Q}^\pi_{s'a'} \right)^2 + \mathbf{E}_t \left((\mathbf{E}_t \hat{P}_{s'sa})(\hat{Q}^\pi_{s'a'} - \mathbf{E}_t \hat{Q}^\pi_{s'a'}) \right)^2.
\end{aligned}$$

Now using the conditional independence property again and the fact that the Q-values are bounded, as implied by assumption 2, we have

$$\mathop{\mathbf{E}}_t \left((\hat{P}_{s'sa} - \mathop{\mathbf{E}}_t \hat{P}_{s'sa}) \hat{Q}^\pi_{s'a'} \right)^2 = \mathop{\mathbf{E}}_t \left(\hat{P}_{s'sa} - \mathop{\mathbf{E}}_t \hat{P}_{s'sa} \right)^2 \mathop{\mathbf{E}}_t \left(\hat{Q}^\pi_{s'a'} \right)^2 \leq Q^2_{\max} \mathbf{var}_t \hat{P}_{s'sa},$$

and since $P_{s'sa} \in [0, 1]$ we have

$$\mathop{\mathbf{E}}_t \left((\mathop{\mathbf{E}}_t \hat{P}_{s'sa})(\hat{Q}^\pi_{s'a'} - \mathop{\mathbf{E}}_t \hat{Q}^\pi_{s'a'}) \right)^2 = (\mathop{\mathbf{E}}_t \hat{P}_{s'sa})^2 \mathop{\mathbf{E}}_t \left(\hat{Q}^\pi_{s'a'} - \mathop{\mathbf{E}}_t \hat{Q}^\pi_{s'a'} \right)^2 \leq \mathop{\mathbf{E}}_t (\hat{P}_{s'sa}) \mathbf{var}_t \hat{Q}^\pi_{s'a'}.$$

Putting it all together we obtain

$$\mathbf{var}_t \hat{Q}^\pi_{sa} \leq \sigma^2_{sa} + \gamma^2 \sum_{s',a'} \pi_{s',a'} \mathop{\mathbf{E}}_t (\hat{P}_{s'sa}) \mathbf{var}_t \hat{Q}^\pi_{s'a'}$$

where σ^2_{sa} is the *local* uncertainty, and is given by

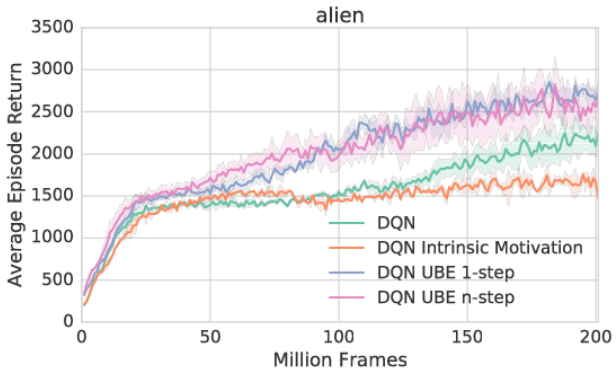
$$\sigma^2_{sa} = \mathbf{var}_t \hat{\mu}_{sa} + \gamma^2 Q^2_{\max} \sum_{s'} \mathbf{var}_t \hat{P}_{s'sa}.$$

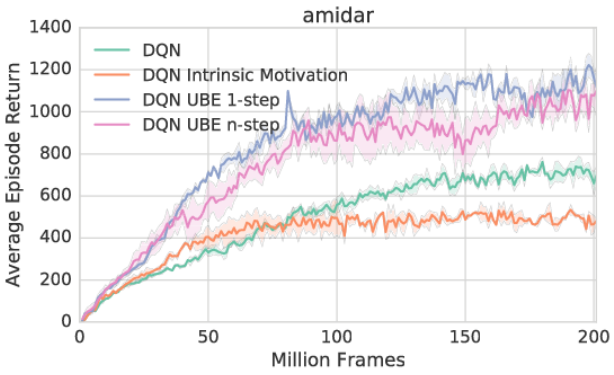
Atari suite scores

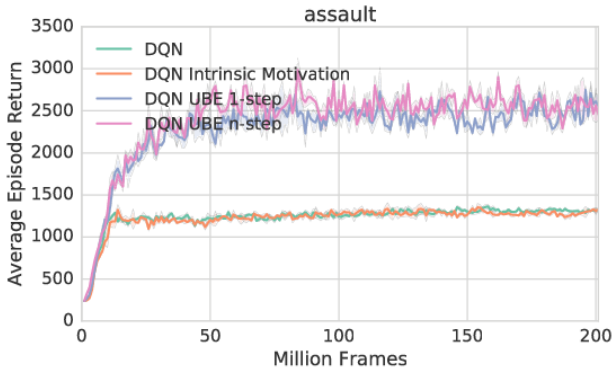
Game	DQN	DQN Intrinsic Motivation	DQN UBE 1-step	DQN UBE n-step
alien	40.96	28.13	**46.90**	43.61
amidar	58.17	41.50	**83.48**	83.14
assault	479.34	647.16	887.18	**1112.28**
asterix	67.26	74.01	82.34	**130.95**
asteroids	1.86	2.25	**3.54**	2.61
atlantis	**28662.58**	14382.27	**28655.71**	6889.77
bank heist	56.43	54.20	97.72	**162.90**
battle zone	70.47	77.80	49.63	**101.68**
beam rider	63.22	58.60	106.79	**149.93**
berzerk	18.29	21.30	44.92	**2284.45**
bowling	13.03	**20.10**	-3.64	14.12
boxing	782.87	784.74	811.68	**816.42**
breakout	1438.45	1377.36	**2045.13**	1474.66
centipede	**34.54**	25.04	22.62	18.94
chopper command	58.06	74.19	69.67	**75.90**
crazy climber	428.59	431.10	496.21	**499.34**
defender	83.94	78.20	94.54	**209.70**
demon attack	338.43	372.29	765.46	**897.89**
double dunk	481.10	575.90	750.00	**1031.25**
enduro	92.01	102.36	9.94	**154.21**
fishing derby	85.03	95.68	97.51	**105.80**
freeway	81.08	121.85	0.02	**126.41**
frostbite	9.86	16.02	11.26	**21.22**
gopher	392.09	558.74	656.70	**867.45**
gravitar	2.38	4.07	2.06	**4.91**
hero	49.27	**63.34**	47.63	34.58
ice hockey	64.17	**64.49**	38.22	58.43
jamesbond	193.96	230.36	198.40	**430.89**
kangaroo	266.39	311.21	202.79	**537.69**
krull	656.32	702.15	**1033.61**	838.45
kung fu master	103.20	112.79	128.30	**153.40**
montezuma revenge	-0.49	4.21	**11.43**	0.80
ms pacman	16.11	16.17	18.42	**19.82**
name this game	114.24	101.23	**129.62**	127.99
phoenix	115.17	157.34	**199.39**	167.51
pitfall	**5.49**	**5.49**	**5.49**	**5.49**
pong	112.04	112.06	**116.32**	**116.37**
private eye	-0.04	**0.44**	-0.44	0.33
qbert	79.41	103.53	124.96	**125.85**
riverraid	43.41	46.33	60.73	**68.24**
road runner	524.43	531.48	722.15	**732.09**
robotank	770.41	779.94	414.29	**803.06**
seaquest	7.93	**10.61**	9.01	9.31
skiing	11.47	15.26	31.55	**54.31**
solaris	3.48	8.23	**18.56**	-6.53
space invaders	87.42	72.95	125.29	**138.65**
star gunner	309.47	398.98	456.86	**547.39**
surround	16.67	19.44	37.42	**60.98**
tennis	145.58	145.58	**227.93**	145.58
time pilot	92.53	76.73	88.36	**121.64**
tutankham	148.29	**191.72**	132.62	138.12
up n down	92.62	95.73	139.35	**142.12**
venture	4.05	**10.77**	-1.24	8.73
video pinball	1230.64	2393.54	**3354.58**	1992.87
wizard of wor	73.26	73.96	**187.48**	118.76
yars revenge	37.93	23.24	**46.21**	44.36
zaxxon	35.24	53.14	**62.20**	56.44

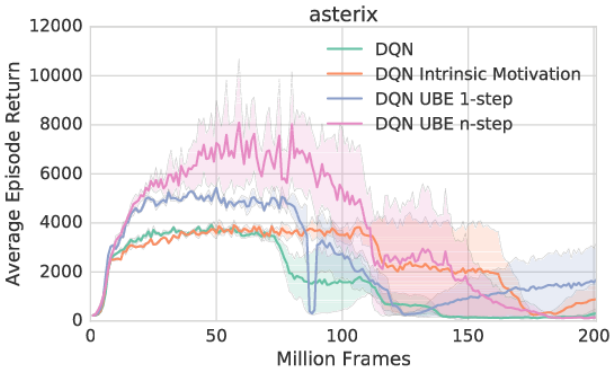
Table 2: Normalized scores for the Atari suite from random starts, as a percentage of human normalized score.

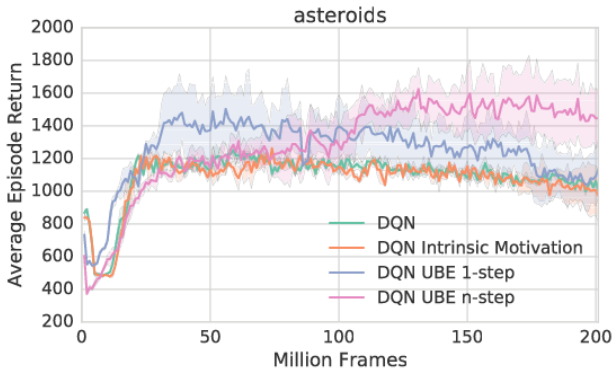
Atari suite learning curves

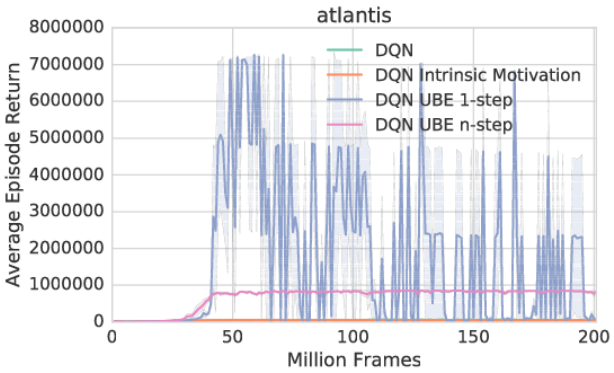


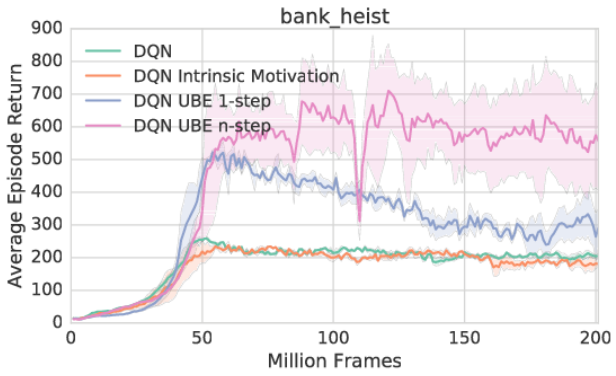


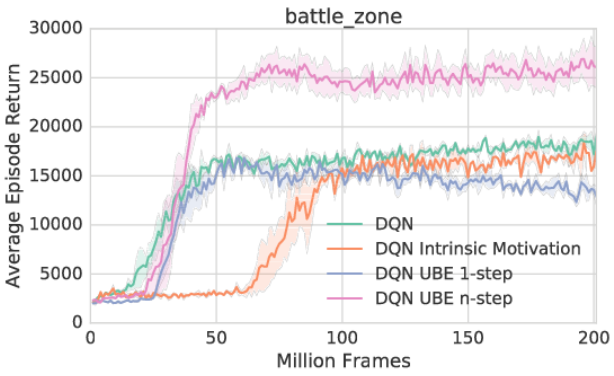


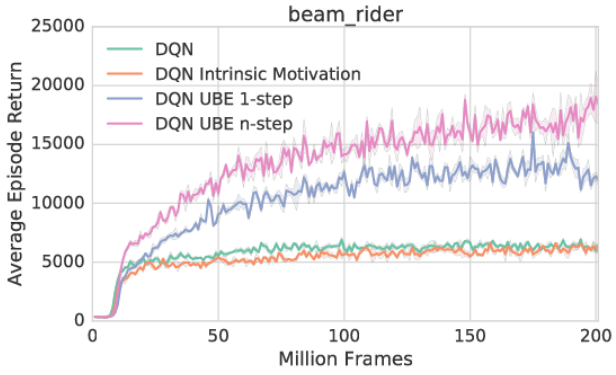


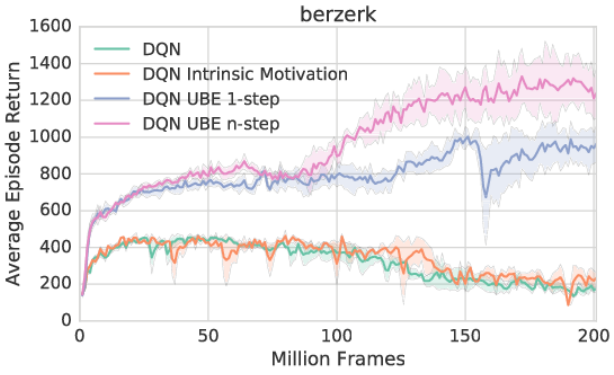


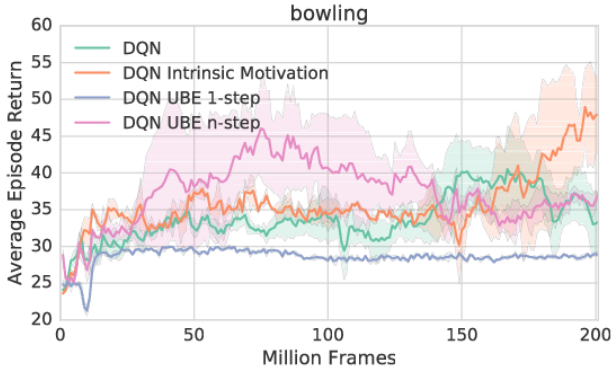


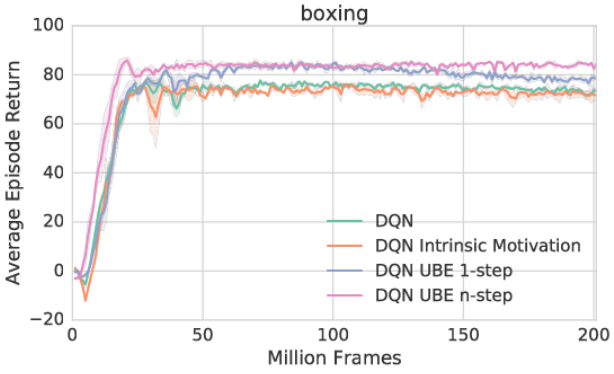

bank_heist
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

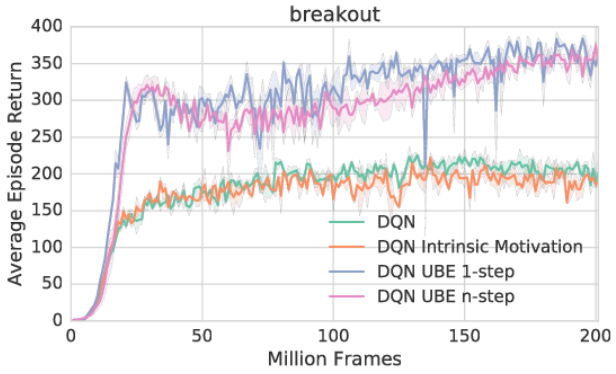

battle_zone
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

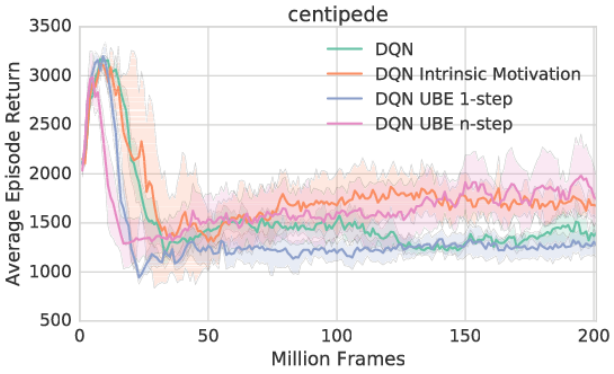

beam_rider
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

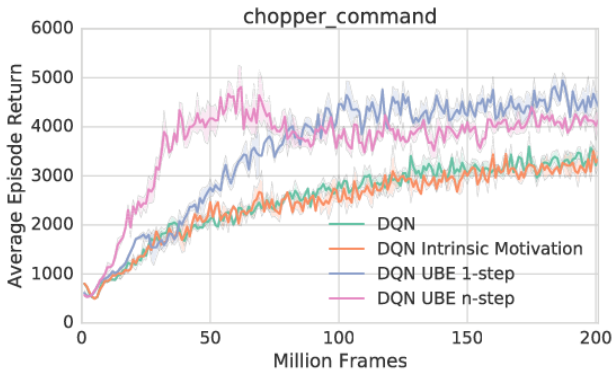

berzerk
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

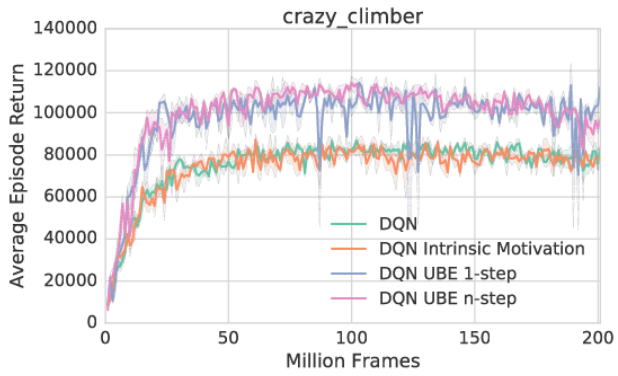

bowling
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

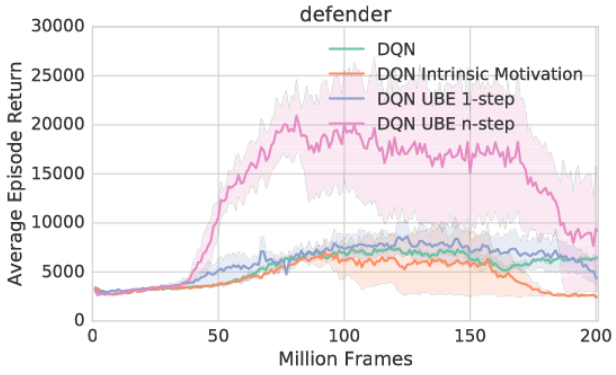

boxing
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

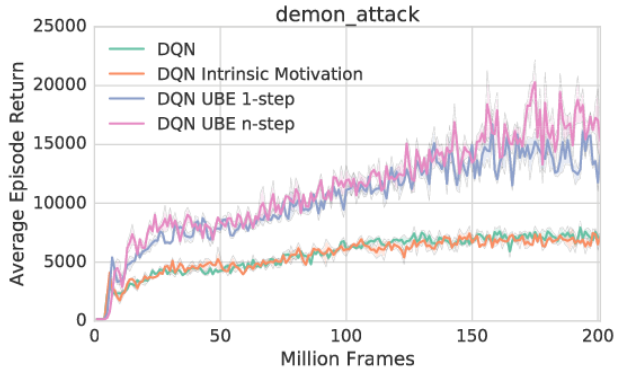

breakout
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

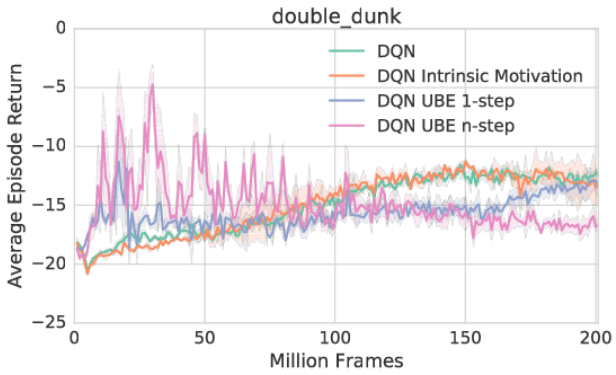

centipede
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

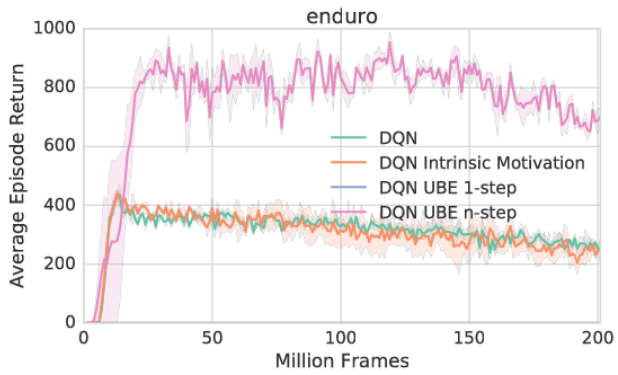

chopper_command
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

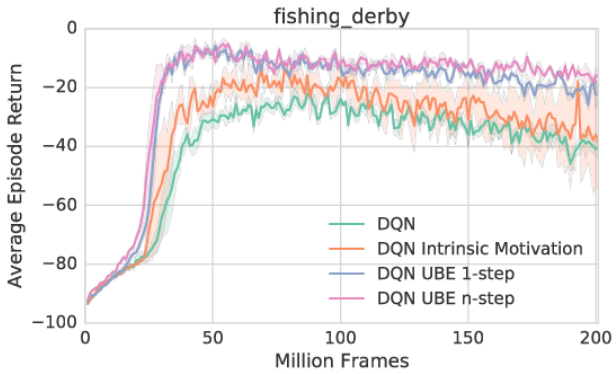

crazy_climber
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

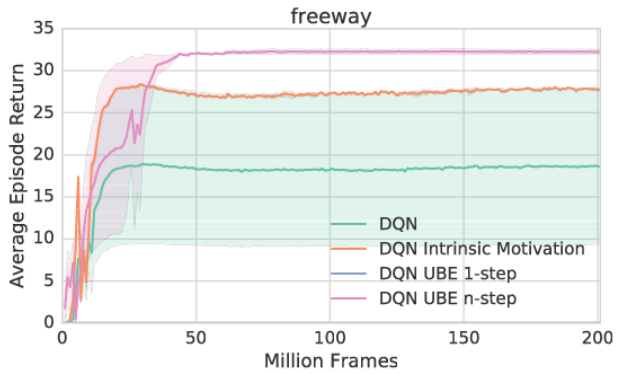

defender
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

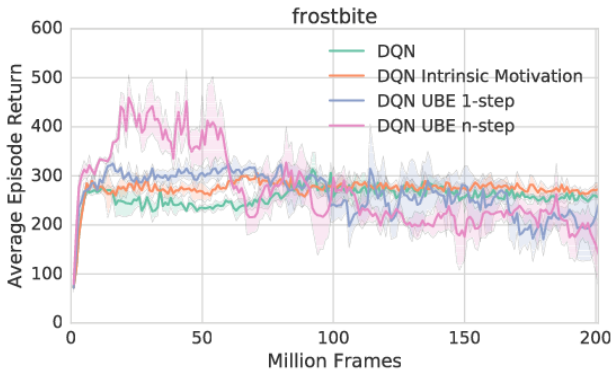

demon_attack
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

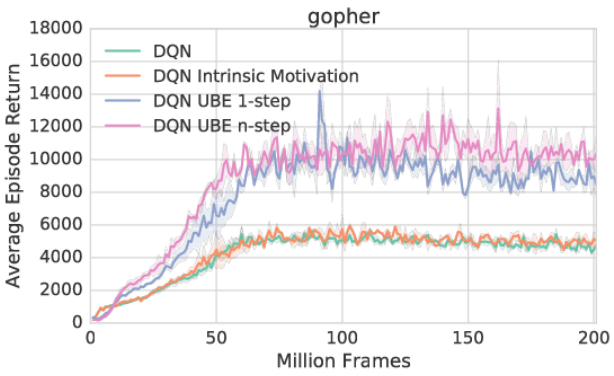

double_dunk
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

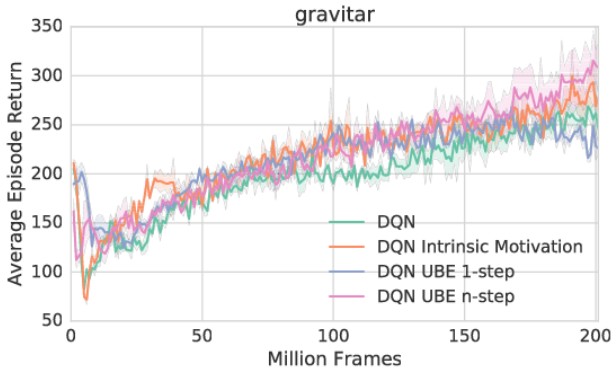

enduro
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

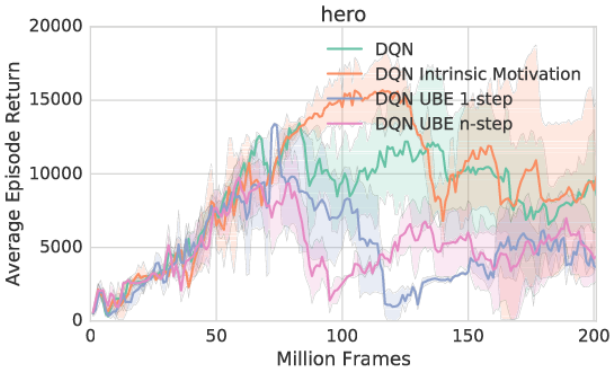

fishing_derby
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

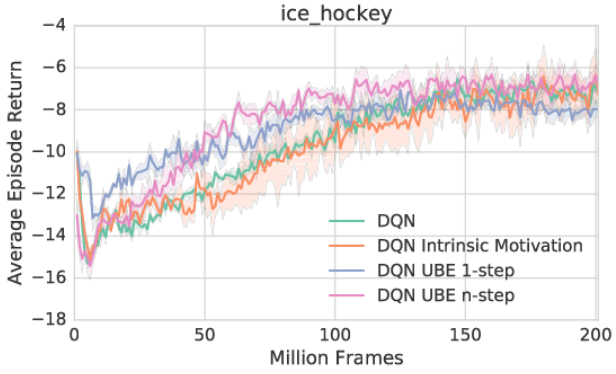

freeway
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

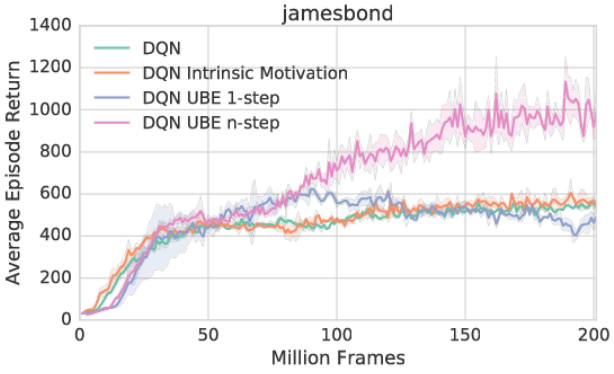

frostbite
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

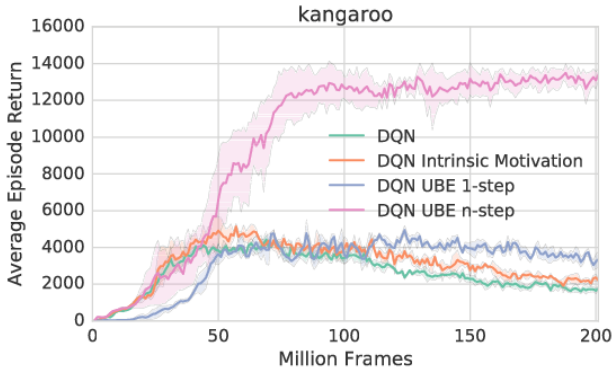

gopher
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

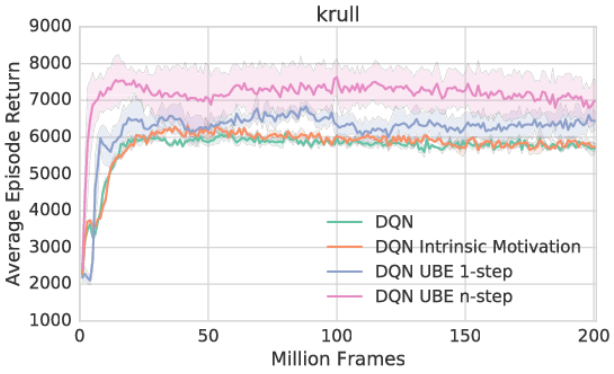

gravitar
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

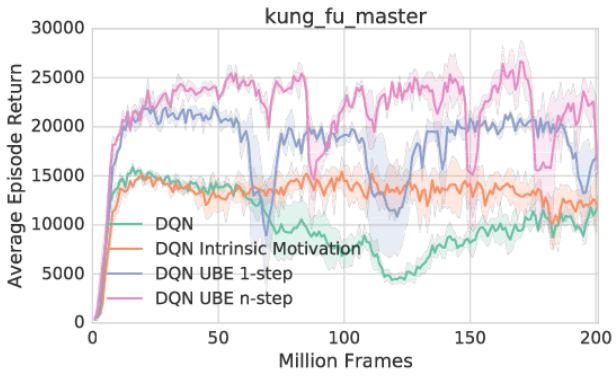

hero
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

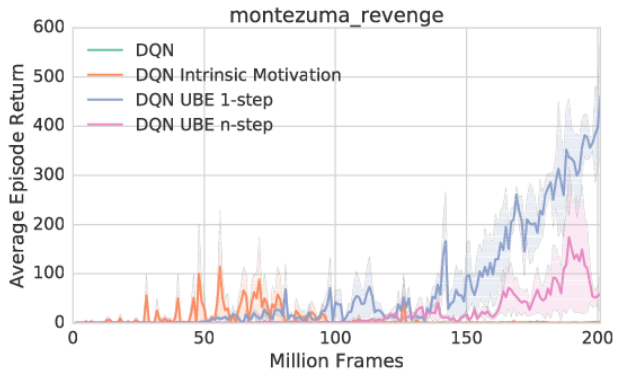

ice_hockey
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

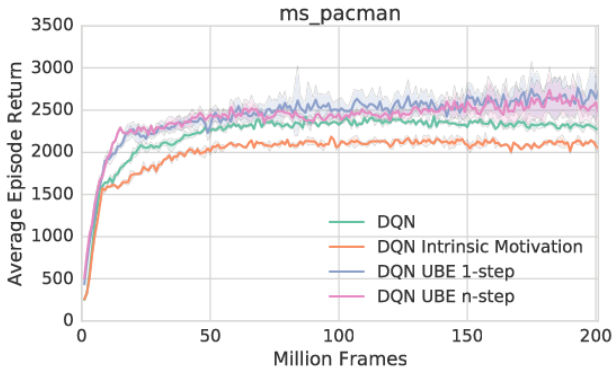

jamesbond
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

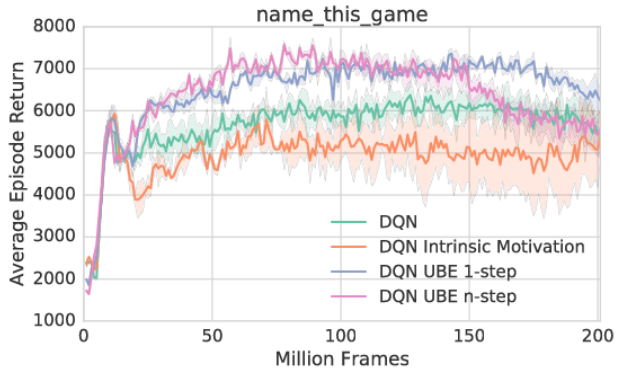

kangaroo
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

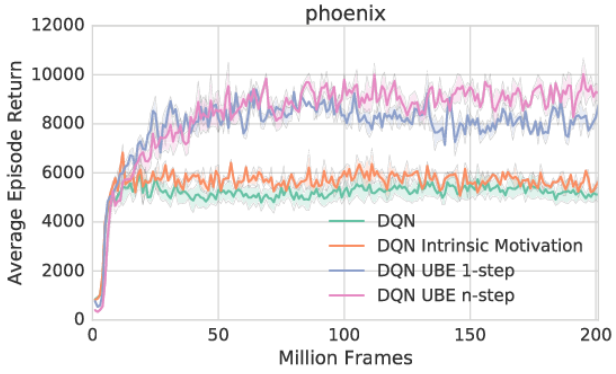

krull
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

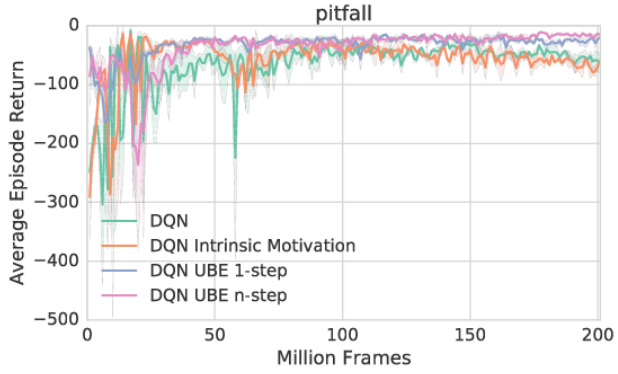

kung_fu_master
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

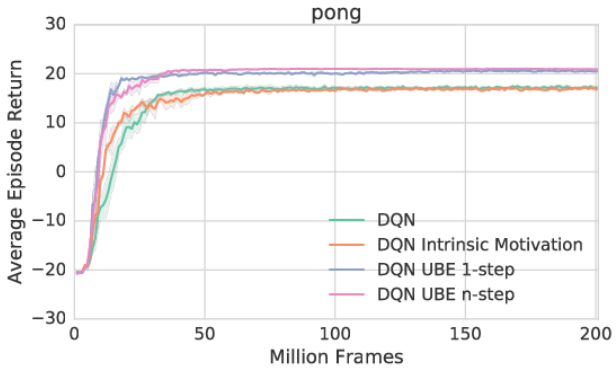

montezuma_revenge
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

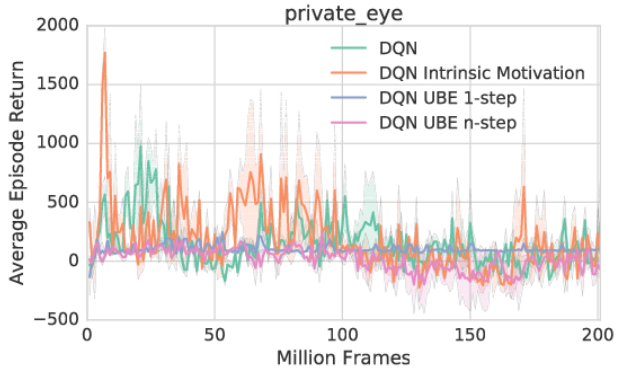

ms_pacman
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

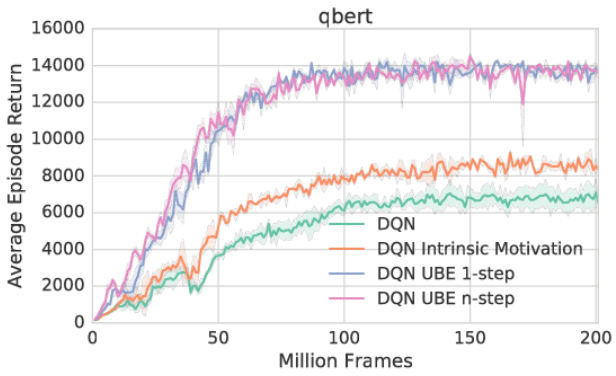

name_this_game
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

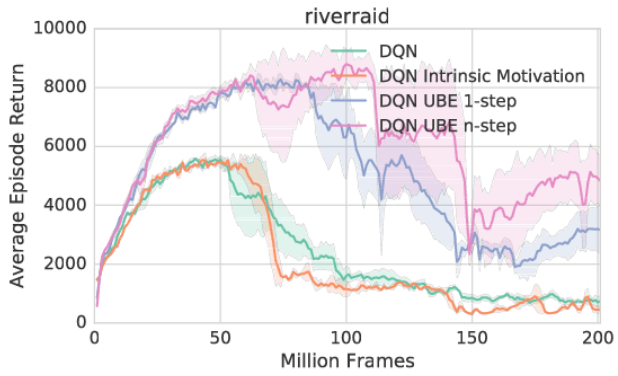

phoenix
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

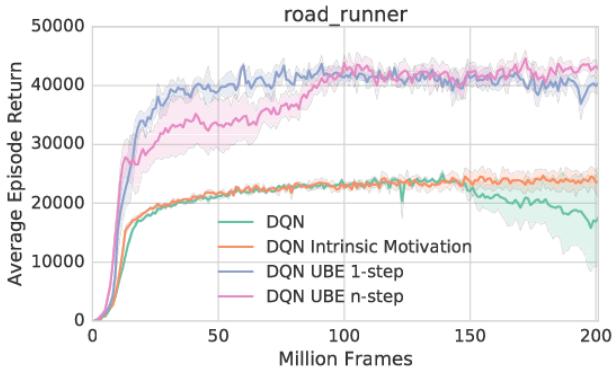

pitfall
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

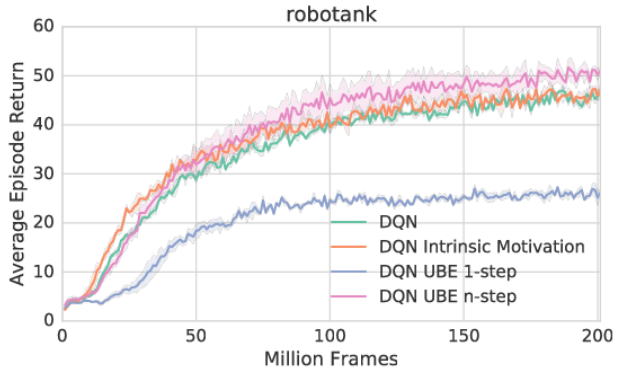

pong
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

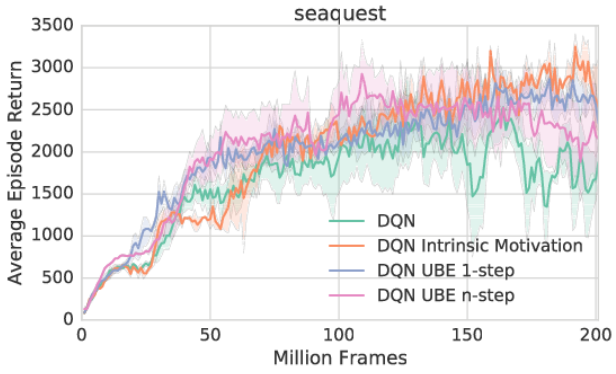

private_eye
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

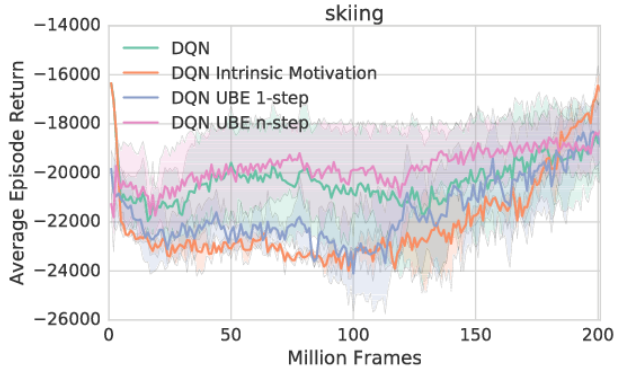

qbert
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

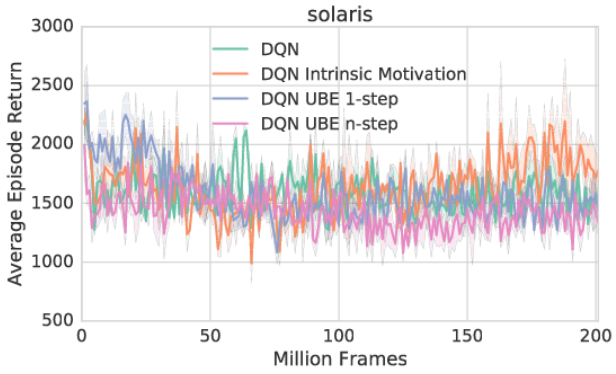

riverraid
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

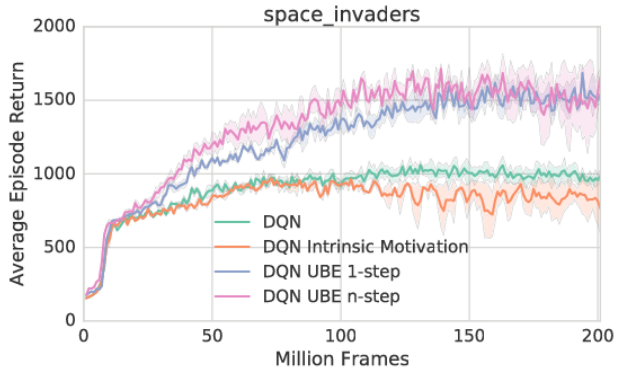

road_runner
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

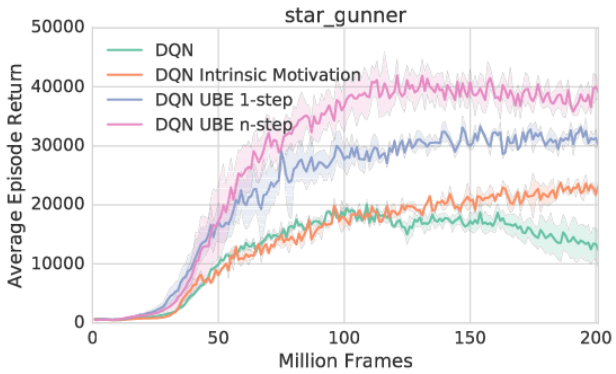

robotank
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

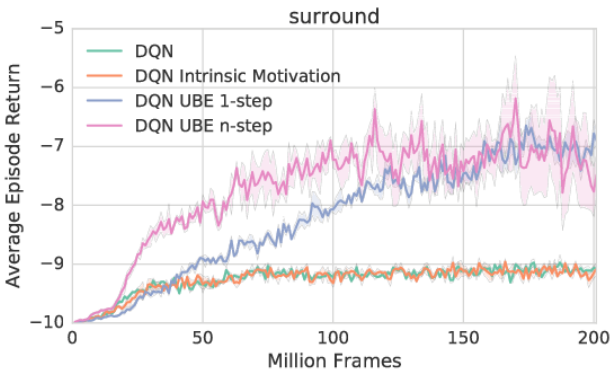

seaquest
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

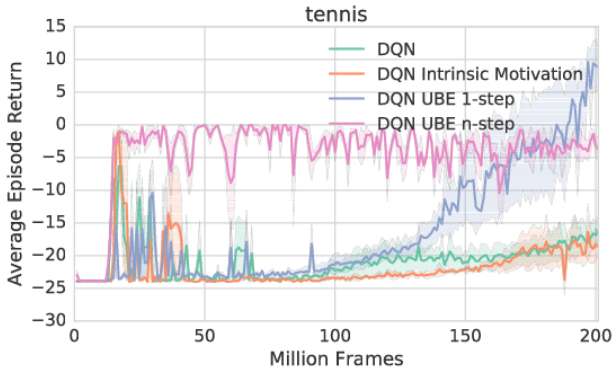

skiing
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

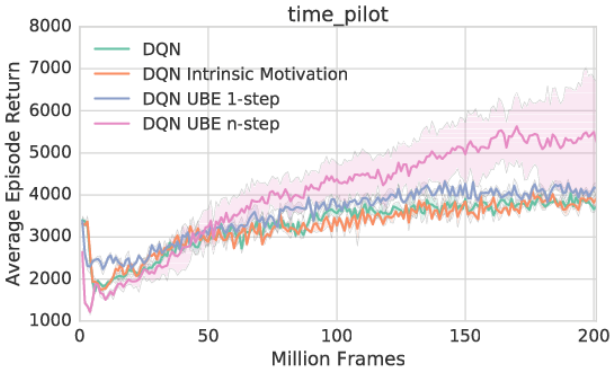

solaris
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

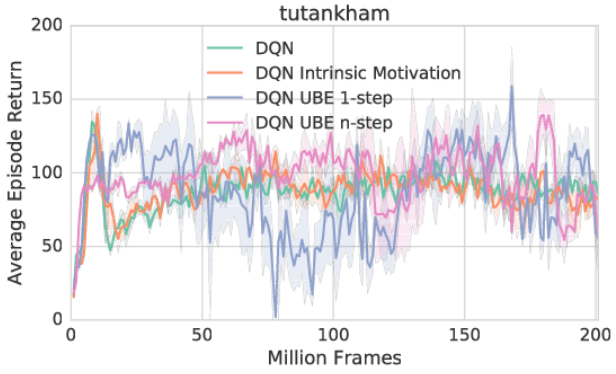

space_invaders
Average Episode Return
Million Frames
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step

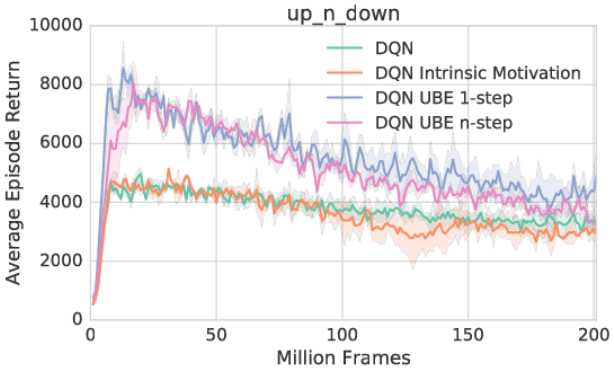

star_gunner
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

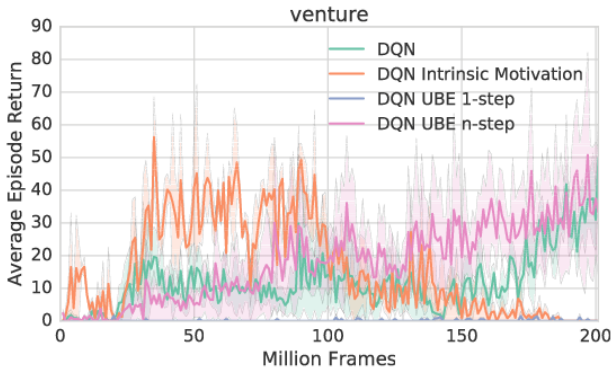

surround
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

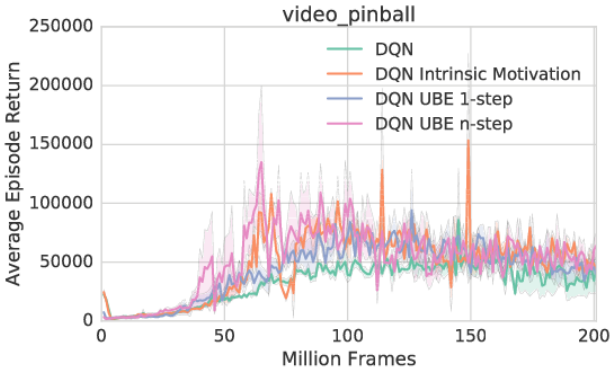

tennis
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

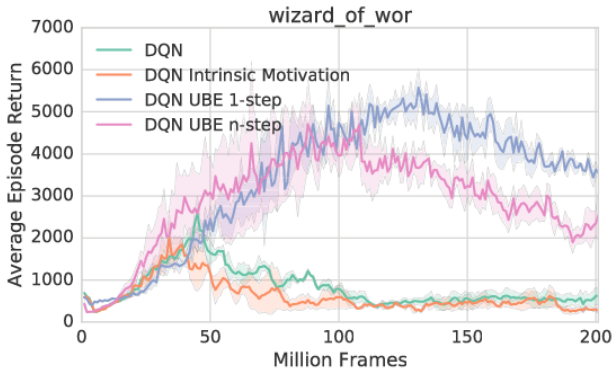

time_pilot
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

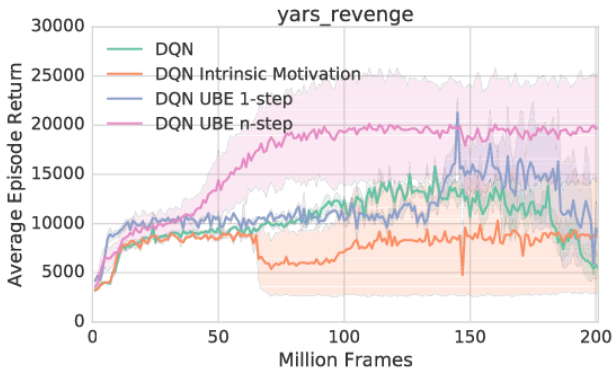

tutankham
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames

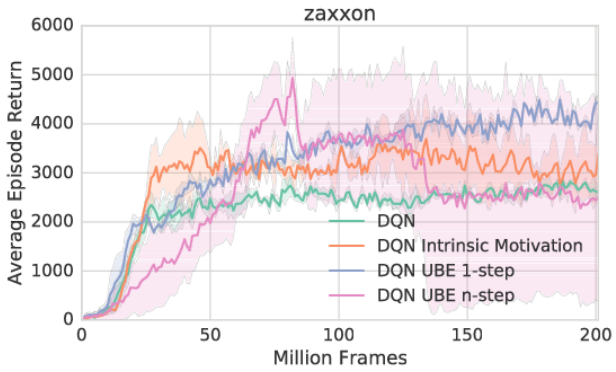

up_n_down
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames


venture
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames


video_pinball
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames


wizard_of_wor
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames
0
1000
2000
3000
4000
5000
6000
7000
0
50
100
150
200


yars_revenge
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames
0
5000
10000
15000
20000
25000
30000
0
50
100
150
200


zaxxon
DQN
DQN Intrinsic Motivation
DQN UBE 1-step
DQN UBE n-step
Average Episode Return
Million Frames
0
1000
2000
3000
4000
5000
6000
0
50
100
150
200