

SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning

Kimin Lee, Michael Laskin, Aravind Srinivas, Pieter Abbeel

UC Berkeley

ICML 2021

Issues on Off-Policy RL

- Current off-policy RL algorithms suffer from several issues:

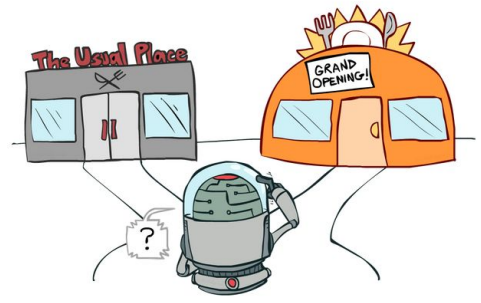
1. Instability in Q-learning

unseen (s,a) → high error

$$Q(s_t, a_t) \leftarrow r_t + \gamma \max_a Q(s_{t+1}, a)$$

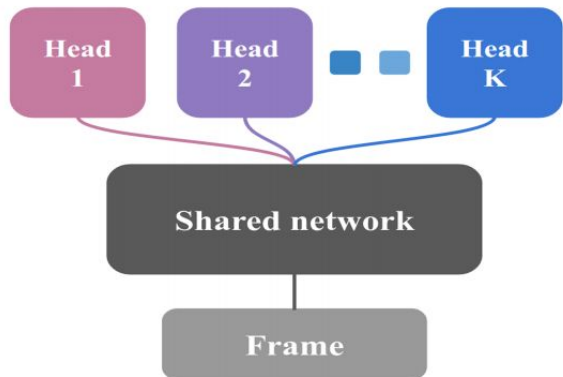


2. Balancing exploration and exploitation



Can Ensemble Improve Off-policy RL?

- Current off-policy RL algorithms suffer from several issues:
- Ensembles have been studied to handle them



Bootstrap DQN
[Osband et al., 2016]



$$a_t = \max_a \{ \underbrace{Q_{\text{mean}}(s_t, a)}_{\text{exploit}} + \lambda \underbrace{Q_{\text{std}}(s_t, a)}_{\text{explore}} \}$$

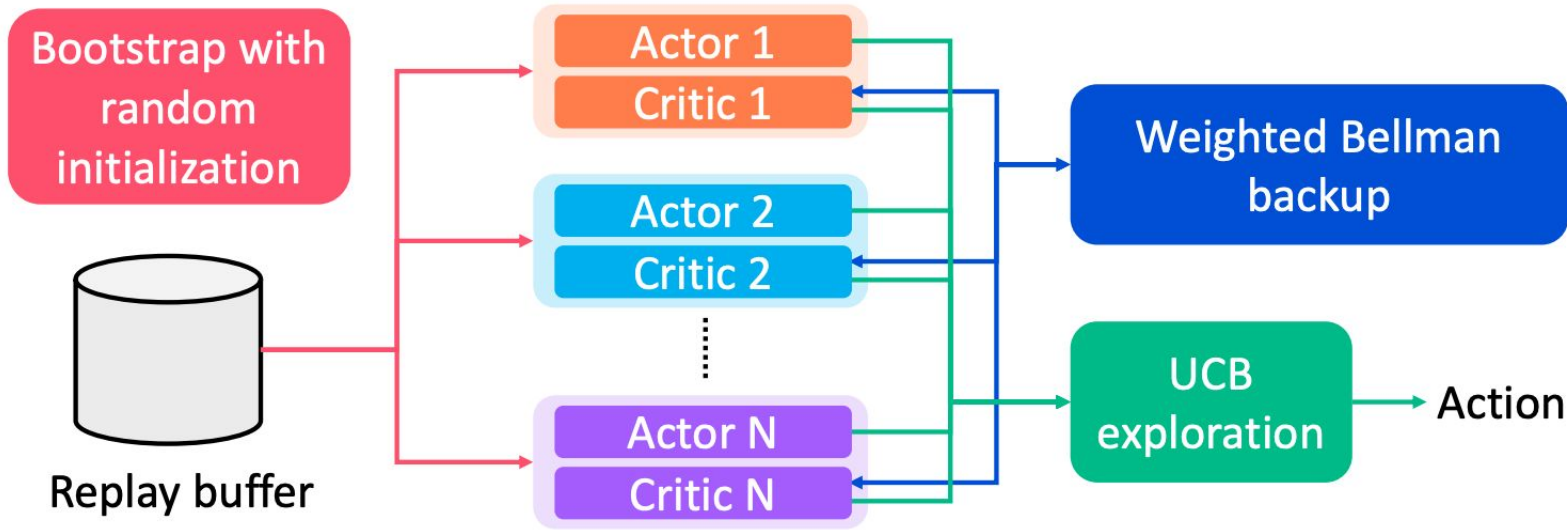
UCB exploration [Chen et al., 2017]

Limitations

1. No method to handle error propagation
2. Most prior work has studied in isolation

SUNRISE

- Weighted Bellman backups based on Q-ensembles
- Unified framework to combine several techniques that use ensembles



SUNRISE: Actor-critic version


Weighted Bellman Backup

- Error propagation issue in Q-learning

$$Q(s_t, a_t) \leftarrow r_t + \gamma \max_a Q(s_{t+1}, a)$$

unseen (s,a) → high error

error propagates




Weighted Bellman Backup

- Error propagation issue in Q-learning


$$Q(s_t, a_t) \leftarrow r_t + \gamma \max_a Q(s_{t+1}, a)$$

unseen (s,a) → high error



error propagates

- Reweighting Bellman backup can handle this issue

$$w(s, a) \left(Q(s, a) - [r(s, a) + \gamma \hat{Q}(s', a')] \right)^2$$



Some confidence score about target value

Weighted Bellman Backup

- Error propagation issue in Q-learning

$$Q(s_t, a_t) \leftarrow r_t + \gamma \max_a Q(s_{t+1}, a)$$

unseen (s,a) → high error



- Reweighting Bellman backup can handle this issue

$$w(s, a) \left(Q(s, a) - [r(s, a) + \gamma \hat{Q}(s', a')] \right)^2$$

Some confidence score about target value

How to quantify the uncertainty on target value?

Weighted Bellman Backup

- Main idea: **uncertainty estimation using ensembles** [Osband et al., 2016, Lakshminarayanan et al., 2017]
- Definition of confidence score

$$w(s, a) = \underbrace{\sigma}_{\text{Sigmoid}} \left(-\bar{Q}_{\text{std}}(s, a) * \underbrace{T}_{\text{Temperature}} \right) + \underbrace{0.5}_{\text{}}$$

- Small variance: weight $\rightarrow 1.0$
- High variance: weight $\rightarrow 0.5$
- Weighted Bellman backup: $w(s, a) \left(Q(s, a) - [r(s, a) + \gamma \hat{Q}(s', a')] \right)^2$

UCB Exploration

- Main idea: utilize uncertainty estimation for exploration
- UCB exploration based on Q-ensemble [Chen et al., 2017]

$$a_t = \max_a \{ \underbrace{Q_{\text{mean}}(s_t, a)}_{\text{exploit}} + \lambda \underbrace{Q_{\text{std}}(s_t, a)}_{\text{explore}} \}$$

- We further extend to continuous action space and apply to more advanced off-policy RL algorithms

Experimental Results

[SUNRISE + Rainbow on Atari]

Game	Human	Random	SimPLe	CURL	DrQ	Rainbow	SUNRISE
Alien	7127.7	227.8	616.9	558.2	761.4	789.0	872.0
Amidar	1719.5	5.8	88.0	142.1	97.3	118.5	122.6
Assault	742.0	222.4	527.2	600.6	489.1	413.0	594.8
Asterix	8503.3	210.0	1128.3	734.5	637.5	533.3	755.0
BankHeist	753.1	14.2	34.2	131.6	196.6	97.7	266.7
BattleZone	37187.5	2360.0	5184.4	14870.0	13520.6	7833.3	15700.0
Boxing	12.1	0.1	9.1	1.2	6.9	0.6	6.7
Breakout	30.5	1.7	16.4	4.9	14.5	2.3	1.8
ChopperCommand	7387.8	811.0	1246.9	1058.5	646.6	590.0	1040.0
CrazyClimber	35829.4	10780.5	62583.6	12146.5	19694.1	25426.7	22230.0
DemonAttack	1971.0	152.1	208.1	817.6	1222.2	688.2	919.8
Freeway	29.6	0.0	20.3	26.7	15.4	28.7	30.2
Frostbite	4334.7	65.2	254.7	1181.3	449.7	1478.3	2026.7
Gopher	2412.5	257.6	771.0	669.3	598.4	348.7	654.7
Hero	30826.4	1027.0	2656.6	6279.3	4001.6	3675.7	8072.5
Jamesbond	302.8	29.0	125.3	471.0	272.3	300.0	390.0
Kangaroo	3035.0	52.0	323.1	872.5	1052.4	1060.0	2000.0
Krull	2665.5	1598.0	4539.9	4229.6	4002.3	2592.1	3087.2
KungFuMaster	22736.3	258.5	17257.2	14307.8	7106.4	8600.0	10306.7
MsPacman	6951.6	307.3	1480.0	1465.5	1065.6	1118.7	1482.3
Pong	14.6	-20.7	12.8	-16.5	-11.4	-19.0	-19.3
PrivateEye	69571.3	24.9	58.3	218.4	49.2	97.8	100.0
Qbert	13455.0	163.9	1288.8	1042.4	1100.9	646.7	1830.8
RoadRunner	7845.0	11.5	5640.6	5661.0	8069.8	9923.3	11913.3
Seaquest	42054.7	68.4	683.3	384.5	321.8	396.0	570.7
UpNDown	11693.2	533.4	3350.3	2955.2	3924.9	3816.0	5074.0

[SUNRISE + SAC on OpenAI Gym]

	Cheetah	Walker
PETS	2288.4 ± 1019.0	282.5 ± 501.6
POPLIN-A	1562.8 ± 1136.7	-105.0 ± 249.8
POPLIN-P	4235.0 ± 1133.0	597.0 ± 478.8
METRPO	2283.7 ± 900.4	-1609.3 ± 657.5
TD3	3015.7 ± 969.8	-516.4 ± 812.2
SAC	4474.4 ± 700.9	299.5 ± 921.9
SUNRISE	4501.8 ± 443.8	1236.5 ± 1123.9

Experimental Results

[SUNRISE + Rainbow on Atari]

Game	Human	Random	SimPLe	CURL	DrQ	Rainbow	SUNRISE
Alien	7127.7	227.8	616.9	558.2	761.4	789.0	872.0
Amidar	1719.5	5.8	88.0	142.1	97.3	118.5	122.6
Assault	742.0	222.4	527.2	600.6	489.1	413.0	594.8
Asterix	8503.3	210.0	1128.3	734.5	637.5	533.3	755.0
BankHeist	753.1	14.2	34.2	131.6	196.6	97.7	266.7
BattleZone	37187.5	2360.0	5184.4	14870.0	13520.6	7833.3	15700.0
Boxing	12.1	0.1	9.1	1.2	6.9	0.6	6.7
Breakout	30.5	1.7	16.4	4.9	14.5	2.3	1.8
ChopperCommand	7387.8	811.0	1246.9	1058.5	646.6	590.0	1040.0
CrazyClimber	35829.4	10780.5	62583.6	12146.5	19694.1	25426.7	22230.0
DemonAttack	1971.0	152.1	208.1	817.6	1222.2	688.2	919.8
Freeway	29.6	0.0	20.3	26.7	15.4	28.7	30.2
Frostbite	4334.7	65.2	254.7	1181.3	449.7	1478.3	2026.7
Gopher	2412.5	257.6	771.0	669.3	598.4	348.7	654.7
Hero	30826.4	1027.0	2656.6	6279.3	4001.6	3675.7	8072.5
Jamesbond	302.8	29.0	125.3	471.0	272.3	300.0	390.0
Kangaroo	3035.0	52.0	323.1	872.5	1052.4	1060.0	2000.0
Krull	2665.5	1598.0	4539.9	4229.6	4002.3	2592.1	3087.2
KungFuMaster	22736.3	258.5	17257.2	14307.8	7106.4	8600.0	10306.7
MsPacman	6951.6	307.3	1480.0	1465.5	1065.6	1118.7	1482.3
Pong	14.6	-20.7	12.8	-16.5	-11.4	-19.0	-19.3
PrivateEye	69571.3	24.9	58.3	218.4	49.2	97.8	100.0
Qbert	13455.0	163.9	1288.8	1042.4	1100.9	646.7	1830.8
RoadRunner	7845.0	11.5	5640.6	5661.0	8069.8	9923.3	11913.3
Seaquest	42054.7	68.4	683.3	384.5	321.8	396.0	570.7
UpNDown	11693.2	533.4	3350.3	2955.2	3924.9	3816.0	5074.0

[SUNRISE + SAC on OpenAI Gym]

	Cheetah	Walker
PETS	2288.4 ± 1019.0	282.5 ± 501.6
POPLIN-A	1562.8 ± 1136.7	-105.0 ± 249.8
POPLIN-P	4235.0 ± 1133.0	597.0 ± 478.8
METRPO	2283.7 ± 900.4	-1609.3 ± 657.5
TD3	3015.7 ± 969.8	-516.4 ± 812.2
SAC	4474.4 ± 700.9	299.5 ± 921.9
SUNRISE	4501.8 ± 443.8	1236.5 ± 1123.9

1. Consistently improve the performance of previous off-policy RL methods

Experimental Results

[SUNRISE + Rainbow on Atari]

Game	Human	Random	SimPLe	CURL	DrQ	Rainbow	SUNRISE
Alien	7127.7	227.8	616.9	558.2	761.4	789.0	872.0
Amidar	1719.5	5.8	88.0	142.1	97.3	118.5	122.6
Assault	742.0	222.4	527.2	600.6	489.1	413.0	594.8
Asterix	8503.3	210.0	1128.3	734.5	637.5	533.3	755.0
BankHeist	753.1	14.2	34.2	131.6	196.6	97.7	266.7
BattleZone	37187.5	2360.0	5184.4	14870.0	13520.6	7833.3	15700.0
Boxing	12.1	0.1	9.1	1.2	6.9	0.6	6.7
Breakout	30.5	1.7	16.4	4.9	14.5	2.3	1.8
ChopperCommand	7387.8	811.0	1246.9	1058.5	646.6	590.0	1040.0
CrazyClimber	35829.4	10780.5	62583.6	12146.5	19694.1	25426.7	22230.0
DemonAttack	1971.0	152.1	208.1	817.6	1222.2	688.2	919.8
Freeway	29.6	0.0	20.3	26.7	15.4	28.7	30.2
Frostbite	4334.7	65.2	254.7	1181.3	449.7	1478.3	2026.7
Gopher	2412.5	257.6	771.0	669.3	598.4	348.7	654.7
Hero	30826.4	1027.0	2656.6	6279.3	4001.6	3675.7	8072.5
Jamesbond	302.8	29.0	125.3	471.0	272.3	300.0	390.0
Kangaroo	3035.0	52.0	323.1	872.5	1052.4	1060.0	2000.0
Krull	2665.5	1598.0	4539.9	4229.6	4002.3	2592.1	3087.2
KungFuMaster	22736.3	258.5	17257.2	14307.8	7106.4	8600.0	10306.7
MsPacman	6951.6	307.3	1480.0	1465.5	1065.6	1118.7	1482.3
Pong	14.6	-20.7	12.8	-16.5	-11.4	-19.0	-19.3
PrivateEye	69571.3	24.9	58.3	218.4	49.2	97.8	100.0
Qbert	13455.0	163.9	1288.8	1042.4	1100.9	646.7	1830.8
RoadRunner	7845.0	11.5	5640.6	5661.0	8069.8	9923.3	11913.3
Seaquest	42054.7	68.4	683.3	384.5	321.8	396.0	570.7
UpNDown	11693.2	533.4	3350.3	2955.2	3924.9	3816.0	5074.0

[SUNRISE + SAC on OpenAI Gym]

	Cheetah	Walker
PETS	2288.4 ± 1019.0	282.5 ± 501.6
POPLIN-A	1562.8 ± 1136.7	-105.0 ± 249.8
POPLIN-P	4235.0 ± 1133.0	597.0 ± 478.8
METRPO	2283.7 ± 900.4	-1609.3 ± 657.5
TD3	3015.7 ± 969.8	-516.4 ± 812.2
SAC	4474.4 ± 700.9	299.5 ± 921.9
SUNRISE	4501.8 ± 443.8	1236.5 ± 1123.9

1. Consistently improve the performance of previous off-policy RL methods
2. Outperform previous SOTA model-based RL methods

Conclusion

- Ensembles can be used to prevent error propagation in Q-update
- Several techniques can be fruitfully integrated
- For more details,
 - Please check out the paper: <https://arxiv.org/abs/2007.04938>
 - Code: <https://github.com/pokaxpoka/sunrise>

**Thank you for your
attention!**