强化学习原理及应用
Reinforcement Learning (RL): Theories & Applications

DCS6289 Spring 2022

Yucong Zhang （张宇聪）

School of Computer Science and Engineering
Sun Yat-Sen University

Lecture 13： Exploration and Exploitation

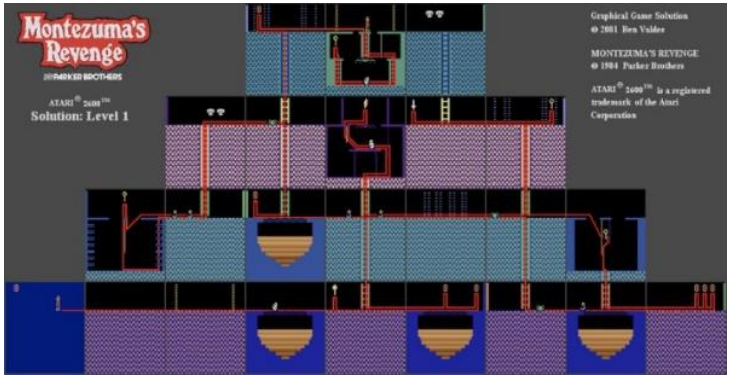
7th June. 2022

Exploration and Exploitation

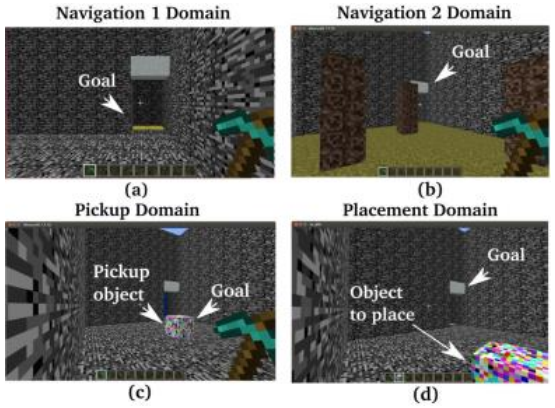
- Introduction of Exploration in RL
- Basic Exploration Technique
- Taxonomy of Exploration
- Exploration in Single-agent RL
- Exploration in Multi-agent RL

Introduction of Exploration in RL

- Challenge
  - Sample-inefficiency
  - Millions of interactions
  - How to efficiently explore the unknown environments and collect informative experiences that could benefit the policy learning most towards optimal ones



- Complex environment
  - Large state-action space
  - Sparse or delayed rewards
  - Noisy
  - Long horizons
  - Non-stationary
  - Local- and global- exploration



Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Introduction of Exploration in RL

- Example： Montezuma’s revenge
  - Getting key = reward
  - Opening door = reward
  - Getting killed by skull = nothing (is it good ? bad?)
  - Finishing the game only weakly correlates with rewarding events
  - We know what to do because we understand what these sprites mean!

Introduction of Exploration in RL

- Two potential definitions of exploration problem
  - How can an agent discover high-reward strategies that require a temporally extended sequence of complex behaviors that, individually, are not rewarding?
  - How can an agent decide whether to attempt new behaviors (to discover ones with higher reward) or continue to do the best thing it knows so far?
- Actually the same problem
  - Exploitation: doing what you know will yield highest reward
  - Exploration: doing things you haven't done before, in the hopes of getting even higher reward
- Exploration and exploitation examples
  - Restaurant selection
    - Exploitation: go to your favorite restaurant
    - Exploration: try to a new restaurant

Basic Exploration Technique

1. Epsilon-greedy
   - With a probability $1-\epsilon$, the agent chooses the action greedily (i.e., exploitation; and a random choice is made otherwise (i.e., exploration)
2. Boltzmann exploration
   - Agent draws actions from Boltzmann distribution over its Q-values: $p(a) = e^{Q(s,a^i)/\tau} \Big/ \sum_i e^{Q(s,a^i)/\tau}$

   Where the temperature parameter τ controls the degree of the selection strategy towards a purely random strategy
3. Upper confidence bounds(UCB)
   - To measure the potential of each action by an upper confidence bound of the reward expectation: $a = \arg\max_a Q(a) + \sqrt{-\ln k / 2N(a)}$
4. Entropy Regularization
   - The policy entropy $H\big(\pi(a|s)\big)$ is added to the objective function as a regularization term, encouraging the policy to take diverse actions
5. Noise Perturbation
   - As to deterministic policies, noise perturbation, $\pi'(s) = \pi(s) + N$, is a natural way to induce exploration
6. Thompson sampling
   - Sample a function f from Posterior(f) each time, and takes actions greedily with respect to such randomly drawn belief f: $\alpha(x) = f(x), s.t. f \sim Posterior(f)$

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Taxonomy of Exploration

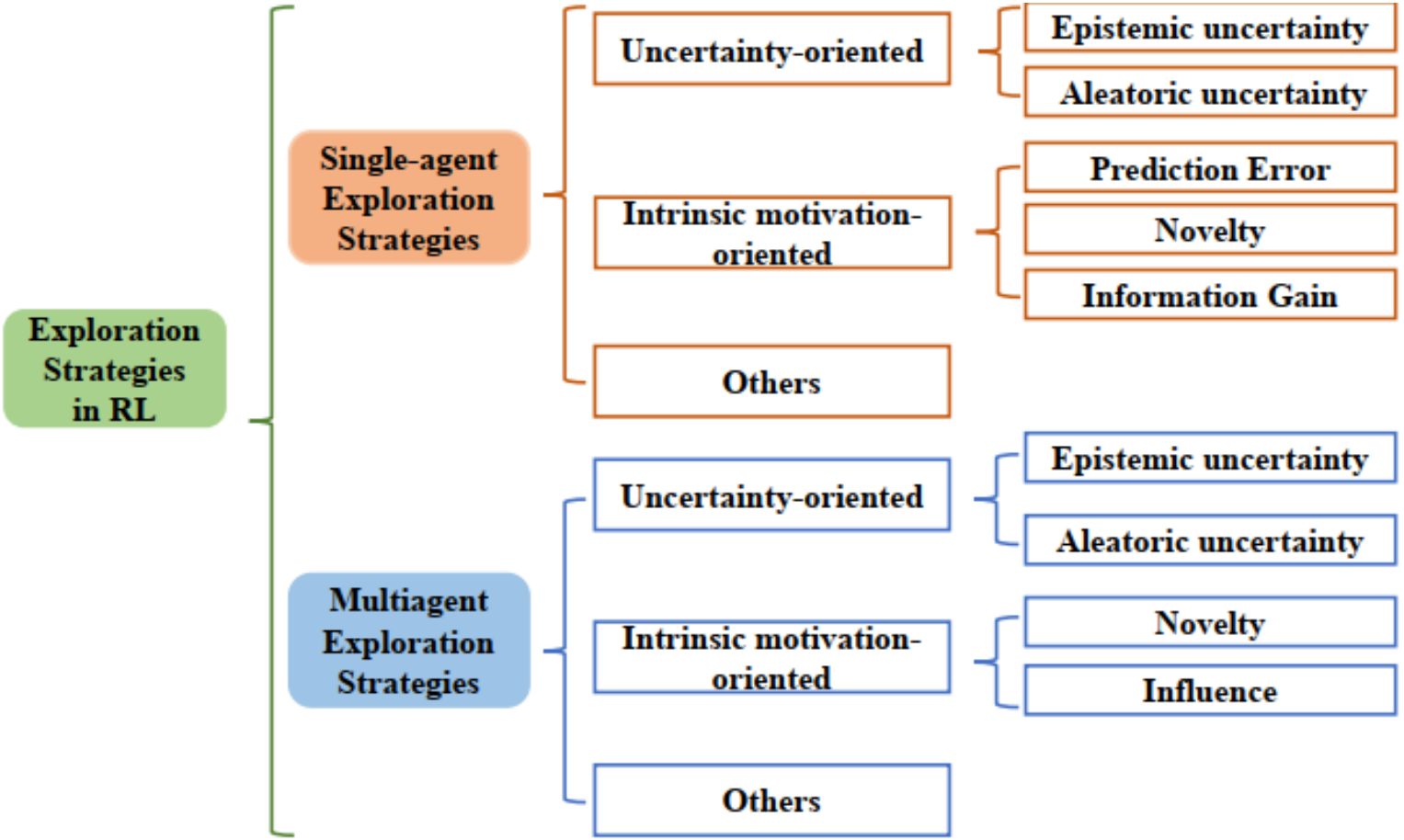
- Uncertain-oriented
  - Optimism in the Face of Uncertainty(OFU)
  - To leverage the quantification of epistemic and aleatoric uncertainty as a general means to measure the sufficiency of learning and the intrinsic stochasticity

- Intrinsic motivation-oriented
  - In developmental psychology, intrinsic motivation is considered as the primary driver in the early stages of human development, e.g., children often employ less goal-oriented exploration but by curiosity to gain the knowledge of the world
  - To make use of various reward-agnostic information as intrinsic motivation to guide exploration

- Other

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Taxonomy of Exploration



Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Uncertain-oriented Exploration

- Epistemic uncertainty (parametric uncertainty)
  - Considered as the errors that arise from insufficient and inaccurate knowledge about environment
  - OFU encourages the agent to visit states and actions with higher epistemic uncertainty to explore the unknown environment
  - How to estimate the epistemic uncertainty: MC dropout, bootstrap sampling, ensemble estimators
  - Common methods: RLSVI, Bayesian DQN, Bootstrap DQN, SUNRISE
- Aleatoric uncertainty (return uncertainty)
  - Represents the intrinsic randomness of environment, and can be captured by the return distribution
  - Two ways:
    - UCB, a directly method for exploring states and actions with high uncertainty is performing optimistic action-selection by choosing the action to maximize the optimistic value function in each time step
    - Thompson Sampling, the action-selection is greedy to a sampled value function from the Q-posterior
  - Common methods: DUVN, IDS

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Uncertain-oriented Exploration

- Parametric Posterior: learned by Bayesian regression in linear MDPS, where the transition and reward functions are assumed to be linear to state-action features

		Method	Large State Space	Continuous Control	Long-horizon	White-noise
Epistemic Uncertainty	Parametric Posterior	RLSVI [73] [74]			high	partially
		Bayesian DQN [75]			high	partially
		Successor Uncertainty [63]	✓		high	partially
		Wasserstein DQN [76]	✓		high	partially
		UBE [67]	✓		high	partially
	Non-parametric Posterior	Bootstrapped DQN [77]	✓		high	partially
		OAC [78]	✓	✓	high	partially
		SUNRISE [79]	✓	✓	high	partially
		OB2I [80]	✓		high	partially
Epistemic & Aleatoric Uncertainty	Non-parametric Posterior	DUVN [81]	✓		partially	partially
		IDS [82]	✓		high	high
		DLTV with QR-DQN [83]	✓			high
		DLTV with NC-QR-DQN [84]	✓			high

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- is originated from humans inherent tendency to be active, to interact with the world in an attempt to have an effect, and to feel a sense of accomplishment

- It is usually accompanied with positive effects (rewards), thus intrinsic motivation-oriented exploration methods often design intrinsic rewards to create the sense of accomplishment for agents

- Three categories:
  - Estimate prediction errors of the environmental dynamics
  - Estimate the state novelty
  - Based on information gain

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

	Method	Continuous Control	Long-horizon	White-noise
Prediction error	Dynamic-AE [105]			
	ICM [70]			high
	Curiosity-Driven [71]			
	AR4E [106]			high
	VDM [107]	✓		high
	EMI [108]	✓		
Novelty	TRPO-AE-hash [109]	✓		partially
	A3C+ [110]	✓		partially
	DQN-PixelCNN [111]			partially
	ϕ-EB [112]	✓		partially
	VAE+ME [113]	✓		
	DQN+SR [114]			high
	DORA [115]			high
	A2C+CoEX [65]	✓		
	RND [64]	✓		partially
	Action balance RND [116]	✓		partially
	Informed exploration [117]	✓	partially	
	EX^2 [118]			high
	SFC [119]	✓	partially	partially
	CB [72]	✓	partially	high
	VSIMR [120]	✓		high
	ECO [121]	✓	partially	high
	SMM [122]	✓		
	DeepCS [123]			
	Novelty Search [124]	✓		high
Information gain	VIME [125]	✓		high
	AKL [126]	✓		high
	Disagreement [127]	✓		high
	MAX [128]	✓		high

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- Prediction Errors
  - Encourage agents to explore states with higher prediction errors.
  - For each state, the intrinsic reward is designed using its prediction error for the next state, which can be measured as the distance between the predicted next state and true one: $R(s_t, s_{t+1}) = dist(\phi(s_{t+1}), \hat{f}(\phi(s_t), a_t))$
  - Common methods：
    - Dynamic Auto-Encoder (DAE): is proposed to compute the distance between the predicted state and the true state in the latent state space which is learned by an auto-encoder
    - Intrinsic Curiosity Module (ICM): by a self-supervised inverse model using states pair (s_t, s_{t+1}) to predict the action a_t done between them
    - AR4E: learns the state transition model via a self-supervised reverse model and contains an action representation module that expands the input low-dimension actions to high dimension representations, and inputs the action representation into the dynamics model together with the current state.
    - EMI: the state and action latent spaces are trained by maximizing the Mutual Information (MI) with the variational divergence lower bound of MI

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- Novelty
  - Motivate agents to approach states they have never visited (a high novelty) by assigning agents intrinsic rewards as bonuses
  - Count based: intrinsic reward is in inverse proportion to the visiting time of states $N(s_t)$: $R(s_t) = \frac{1}{N(s_t)}$. It is hard to apply these methods to very large or continuous state space since an agent is impossible to cover the whole state space
  - Common methods：
    - TRPO-AE-hash: use hash function
    - A3C+, DQN-PixelCNN: rely on density models which compute the pseudo-count, $\rho(s_t)$ is the density model which probability of observing s_t, and $\rho'(s_t)$ is the probability of observing s_t：

$$\hat{N}(s_t) = \frac{\rho(s_t)(1-\rho'(s_t))}{\rho'(s_t)-\rho(s_t)}$$

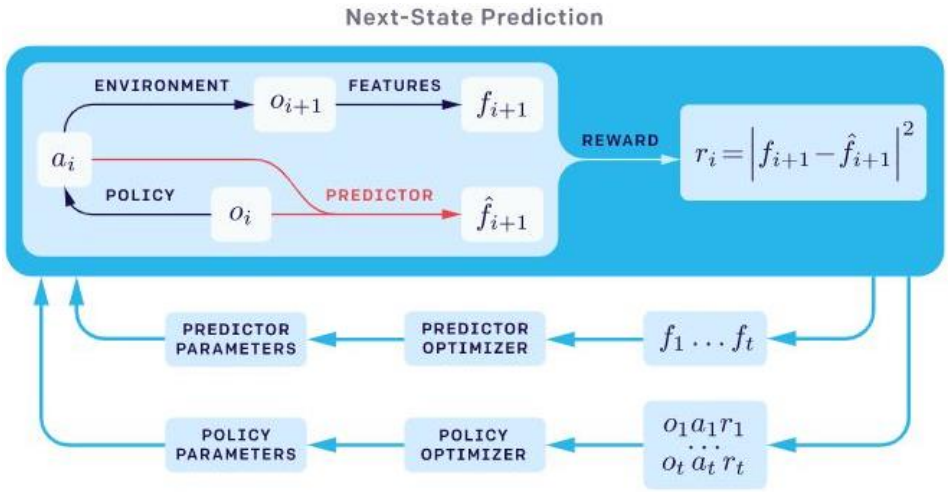
    - RND: by distilling a fixed random network (target network) into another network (predictor network)
    - Never Give Up (NGU): combines both episodic novelty (using k-nearest neighbors) and life-long novelty (RND).

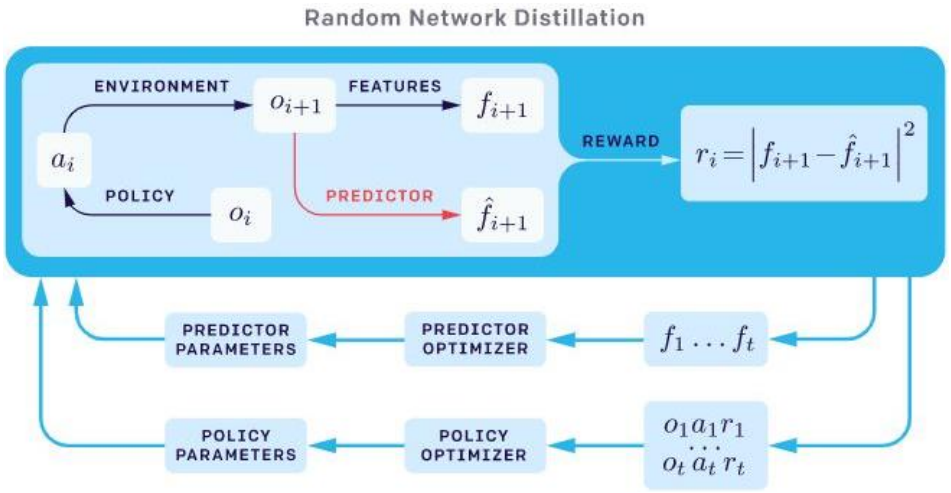
Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- Novelty
  - RND: by distilling a fixed random network (target network) into another network (predictor network)



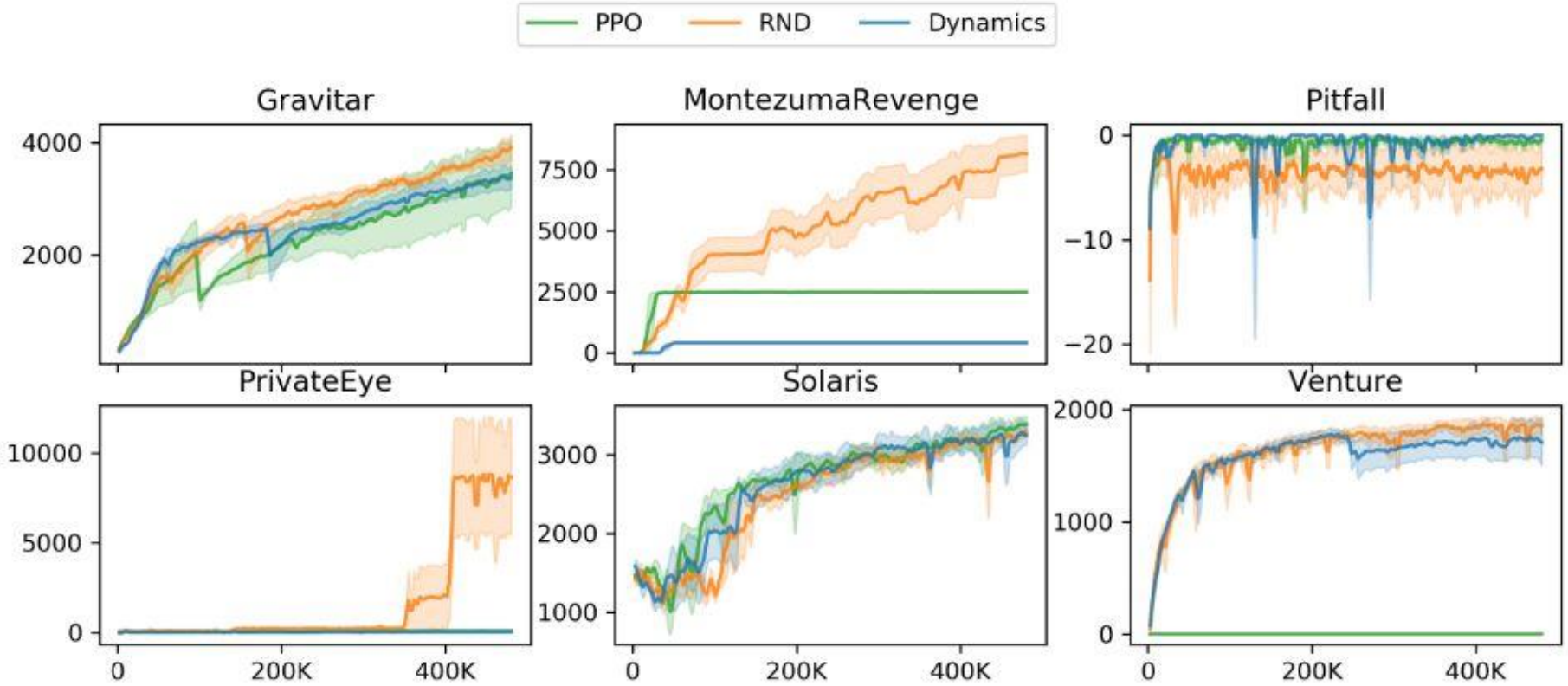


Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- Novelty
  - RND: by distilling a fixed random network (target network) into another network (predictor network)

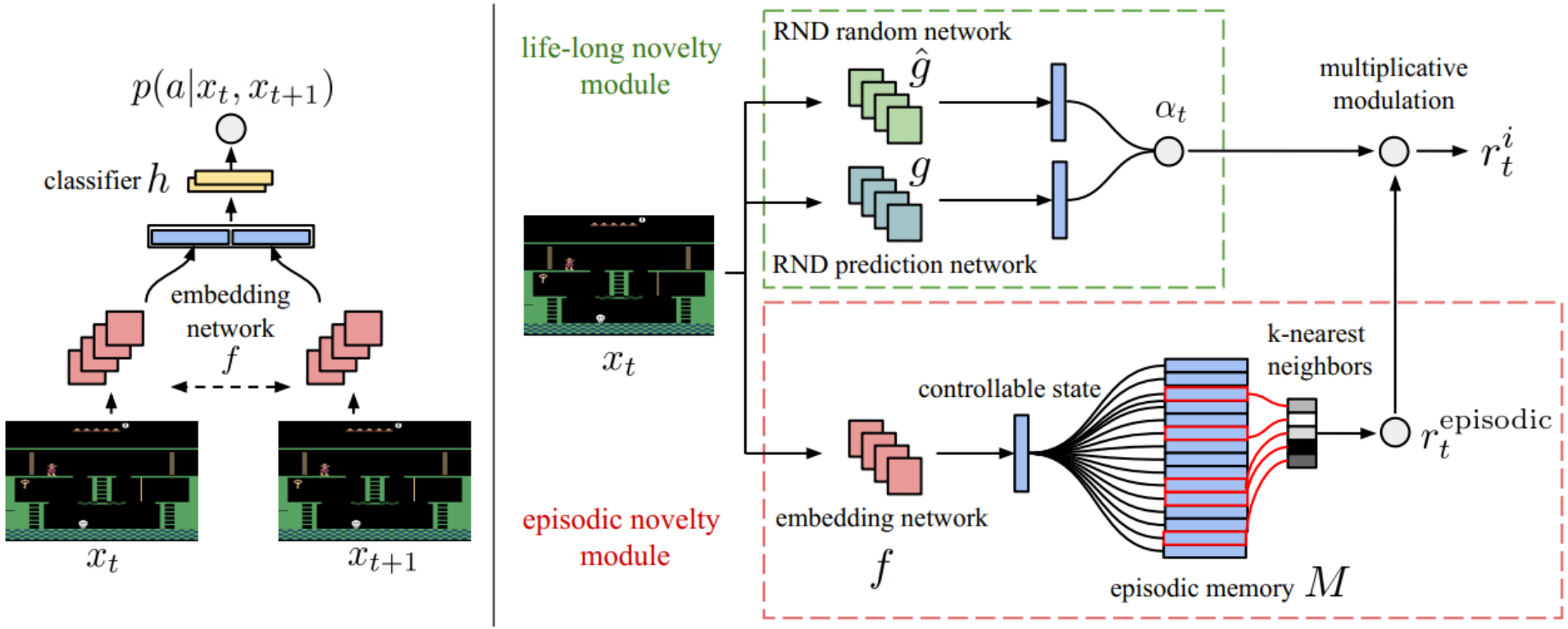


Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- Novelty
  - RND: by distilling a fixed random network (target network) into another network (predictor network)
  - Never Give Up (NGU): combines both episodic novelty (using k-nearest neighbors) and life-long novelty (RND).



Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

Intrinsic Motivation-oriented Exploration

- Information Gain
  - Lead the agents towards unknown areas, as well as to prevent agents paying much attention to stochastic areas
  - Use the information gain as an intrinsic reward
  - $R(s_t, s_{t+k}) = Uncertainty_{t+k}(\theta) - Uncertainty_t(\theta)$, where θ denotes the parameter of a dynamics model, and Uncertainty refers to the model uncertainty
  - Common methods：
    - Variational Information Maximizing Exploration(VIME): encourage agent to take actions that maximize the information gain about its belief of environment dynamics and measures the information gain using variational inference
    - AKL
    - MAX

Other Advanced Methods for Exploration

- Distributed Exploration
- Exploration with Parametric Noise
- Go-Explore

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).

Exploration in single-agent RL

- SUNRISE
  - Challenge
    - Balancing exploration and exploitation.
    - Q-learning often converges to sub-optimal solutions due to error propagation in the target value can lead to an increase in overall error in the Q-function
  - Main idea
    - A simple unified ensemble method
    - Two ingredients:
      - Ensemble-based weighted Bellman back-ups, which re-weight target Q-values based on uncertainty estimates from a Q-ensemble
      - An inference method that selects actions using the highest upper-confidence bounds for efficient exploration. By enforcing the diversity between agents using Bootstrap with random initialization.
    - Is compatible with various off-policy RL algorithms
    - SAC

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in single-agent RL

SUNRISE

- Soft Actor-Critic (SAC): is an off-policy actor-critic method based on the maximum entropy RL framework, which encourages the robustness to noise and exploration by maximizing a weighted objective of the reward and the policy entropy

$$\mathcal{L}_{\text{critic}}^{\text{SAC}}(\theta) = \mathbb{E}_{\tau_t \sim \mathcal{B}}\left[\mathcal{L}_Q\left(\tau_t, \theta\right)\right]$$

$$\mathcal{L}_Q\left(\tau_t, \theta\right) = \left(Q_\theta\left(s_t, a_t\right) - r_t - \gamma \bar{V}\left(s_{t+1}\right)\right)^2$$

$$\text{with } \bar{V}\left(s_t\right) = \mathbb{E}_{a_t \sim \pi_\phi}\left[Q_{\bar{\theta}}\left(s_t, a_t\right) - \alpha \log \pi_\phi\left(a_t \mid s_t\right)\right]$$

$$\mathcal{L}_{\text{actor}}^{\text{SAC}}(\phi) = \mathbb{E}_{s_t \sim \mathcal{B}}\left[\mathcal{L}_\pi\left(s_t, \phi\right)\right]$$

$$\mathcal{L}_\pi\left(s_t, \phi\right) = \mathbb{E}_{a_t \sim \pi_\phi}\left[\alpha \log \pi_\phi\left(a_t \mid s_t\right) - Q_\theta\left(s_t, a_t\right)\right]$$

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in single-agent RL

SUNRISE

- Weighted Bellman backups
  - Consider an ensemble of N SAC agents: $\{Q_{\theta_i}, \pi_{\phi_i}\}_{i=1}^{N}$
  - Error in the target Q-function $Q_{\bar{\theta}}(s_{t+1}, a_{t+1})$ get propagated into the Q-function $Q_{\theta}(s_t, a_t)$ at current state. In other words, errors in the previous Q-function induce the "noise" to the learning "signal" (i.e. true Q-value) of the current Q-function
  - Error propagation can cause inconsistency and unstable convergence

  - For each agent i, considering a weighted Bellman backup:

$$\mathcal{L}_{WQ}(\tau_t, \theta_i) = w(s_{t+1}, a_{t+1})\left(Q_{\theta_i}(s_t, a_t) - r_t - \gamma \bar{V}(s_{t+1})\right)^2$$

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

    where $\bar{Q}_{std}(s,a)$ is the empirical standard deviation of all target Q-function: $\{Q_{\bar{\theta}_i}\}_{i=1}^{N}$

  - The proposed objective $\mathcal{L}_{WQ}$ down-weights the sample transitions with high variance across target Q-functions, resulting in a loss function for the Q-updates that has a better signal-to-noise ratio

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in single-agent RL

SUNRISE

- Bootstrap with random initialization
  - To train the ensemble of agents, using the bootstrap with random initialization, which enforces the diversity between agents through two simple ideas:
    - Initialize the model parameters of all agents with random parameter values for inducing an initial diversity in the models
    - Apply different samples to train each agents
  - For Each SAC agent i in each timestep t, drawing the binary masks $m_{t,i}$ from Bernoulli distribution with parameters $\beta \in (0,1]$, and storing them in the replay buffer, multiplying the bootstrap mask to each objective function, $m_{t,i}\mathcal{L}_\pi$ and $m_{t,i}\mathcal{L}_{WQ}$
- UCB exploration

$$a_t = \max_a \left\{ Q_{\text{mean}}\left(s_t, a\right) + \lambda Q_{\text{std}}\left(s_t, a\right) \right\}$$

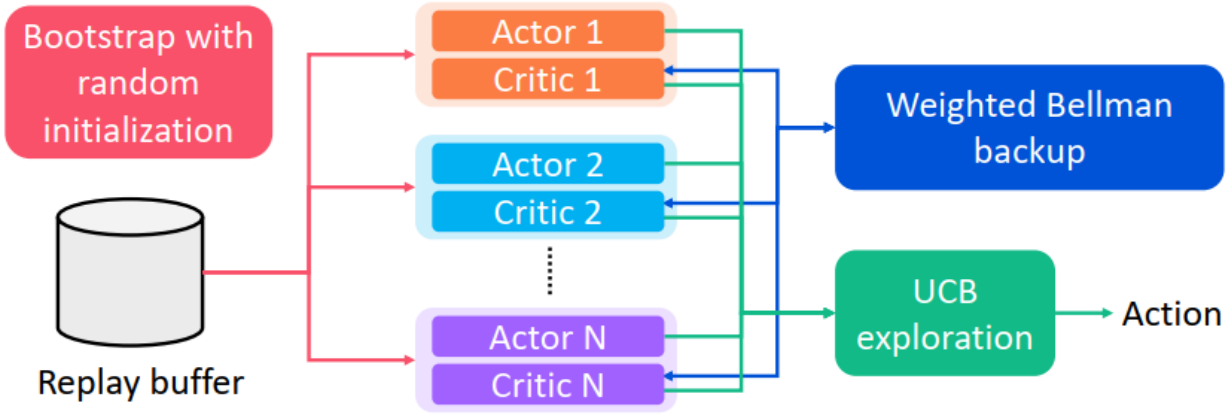
  - Where $\mathrm{Q}_{mean}(s,a)$and $\mathrm{Q}_{std}(s,a)$ are the empirical mean and standard deviation of all Q-function: $\{Q_{\theta_i}\}_{i=1}^{N}$

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

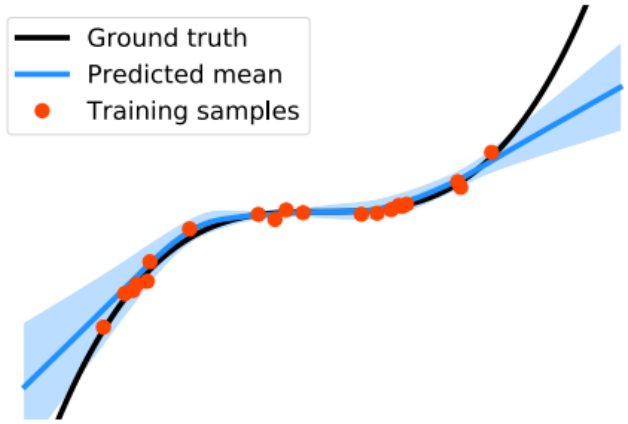
Exploration in single-agent RL

SUNRISE

- Architecture



(a) SUNRISE: actor-critic version



(b) Uncertainty estimates

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in single-agent RL

SUNRISE

- Performance

500K step	PlaNet	Dreamer	SLAC	CURL	DrQ	RAD	SUNRISE
Finger-spin	561 ± 284	796 ± 183	673 ± 92	926 ± 45	938 ± 103	975 ± 16	983 ±1
Cartpole-swing	475 ± 71	762 ± 27	-	845 ± 45	868 ± 10	873 ± 3	876 ± 4
Reacher-easy	210 ± 44	793 ± 164	-	929 ± 44	942 ± 71	916 ± 49	982 ± 3
Cheetah-run	305 ± 131	570 ± 253	640 ± 19	518 ± 28	660 ± 96	624 ± 10	678 ± 46
Walker-walk	351 ± 58	897 ± 49	842 ± 51	902 ± 43	921 ± 45	938 ± 9	953 ± 13
Cup-catch	460 ± 380	879 ± 87	852 ± 71	959 ± 27	963 ± 9	966 ± 9	969 ± 5
100K step							
Finger-spin	136 ± 216	341 ± 70	693 ± 141	767 ± 56	901 ± 104	811 ± 146	905 ± 57
Cartpole-swing	297 ± 39	326 ± 27	-	582 ± 146	759 ± 92	373 ± 90	591 ± 55
Reacher-easy	20 ± 50	314 ± 155	-	538 ± 233	601 ± 213	567 ± 54	722 ± 50
Cheetah-run	138 ± 88	235 ± 137	319 ± 56	299 ± 48	344 ± 67	381 ± 79	413 ± 35
Walker-walk	224 ± 48	277 ± 12	361 ± 73	403 ± 24	612 ± 164	641 ± 89	667 ± 147
Cup-catch	0 ± 0	246 ± 174	512 ± 110	769 ± 43	913± 53	666 ± 181	633 ± 241

Table 2. Performance on DeepMind Control Suite at 100K and 500K environment steps. The results show the mean and standard deviation averaged five runs. For baseline methods, we report the best numbers reported in prior works (Kostrikov et al., 2021).

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in single-agent RL

SUNRISE

- Performance

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

Table 3. Performance on Atari games at 100K interactions. The results show the scores averaged three runs. For baseline methods, we report the best numbers reported in prior works (Kaiser et al., 2020; van Hasselt et al., 2019).

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

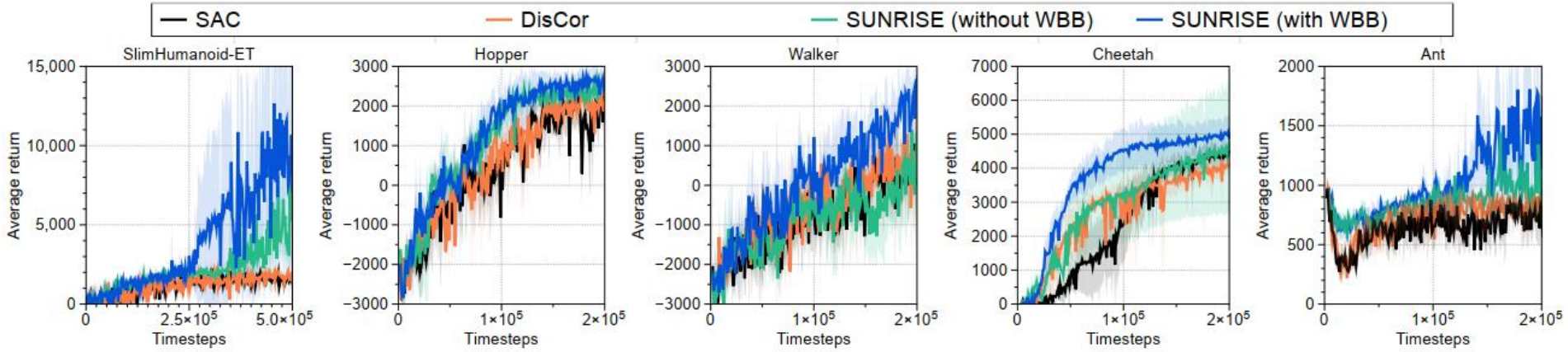
Exploration in single-agent RL

SUNRISE

- Performance

	Cheetah	Walker	Hopper	Ant	SlimHumanoid-ET
PETS	2288.4 ± 1019.0	282.5 ± 501.6	114.9 ± 621.0	1165.5 ± 226.9	2055.1 ± 771.5
POPLIN-A	1562.8 ± 1136.7	-105.0 ± 249.8	202.5 ± 962.5	1148.4 ± 438.3	-
POPLIN-P	4235.0 ± 1133.0	597.0 ± 478.8	2055.2 ± 613.8	2330.1 ± 320.9	-
METRPO	2283.7 ± 900.4	-1609.3 ± 657.5	1272.5 ± 500.9	282.2 ± 18.0	76.1 ± 8.8
TD3	3015.7 ± 969.8	-516.4 ± 812.2	1816.6 ± 994.8	870.1 ± 283.8	1070.0 ± 168.3
SAC	4474.4 ± 700.9	299.5 ± 921.9	1781.3 ± 737.2	979.5 ± 253.2	1371.8 ± 473.4
SUNRISE	4501.8 ± 443.8	1236.5 ± 1123.9	2643.2 ± 472.3	1502.4 ± 483.5	1926.6 ± 375.0

Table 1. Performance on OpenAI Gym at 200K timesteps. The results show the mean and standard deviation averaged over ten runs. For baseline methods, we report the best number in prior works (Wang & Ba, 2020; Wang et al., 2019).

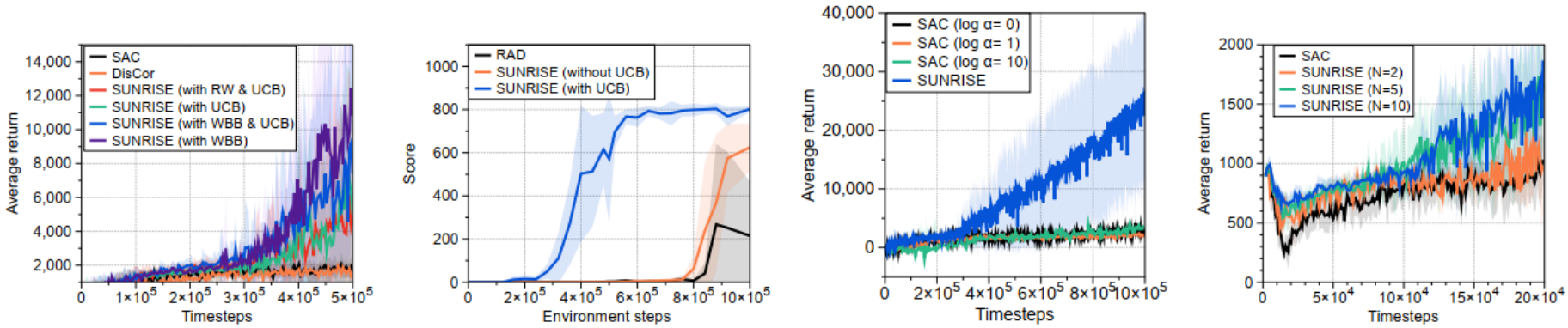


Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in single-agent RL

SUNRISE

- Performance



(a) Large noise (b) Sparse reward (c) Gradient update (d) Ensemble size

Figure 3. (a) Learning curves of SUNRISE with random weight (RW) and the proposed weighted Bellman backups (WBB) on the SlimHumanoid-ET environment with noisy rewards. (b) Effects of UCB exploration on the Cartpole environment with sparse reward. (c) Learning curves of SUNRISE and single agent with h hidden units and five gradient updates per each timestep on the SlimHumanoid-ET environment. (d) Learning curves of SUNRISE with varying values of ensemble size N on the Ant environment.

Lee, Kimin, et al. "Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning." *International Conference on Machine Learning*. PMLR, 2021.

Exploration in Multi-agent RL

- Uncertain-oriented Exploration
  - Epistemic uncertainty
    - Multi-agent safe q-learning
  - Aleatoric uncertainty
    - Distributional value function
- Intrinsic motivated-oriented
  - LIIR
    - Learn individual intrinsic reward to update a proxy critic for each agent
    - Update intrinsic reward network by global critic
  - EMC
    - Two networks: exploration and exploitation
    - Exploration network: make use of experience in replay buffer to update itself network and calculate the intrinsic reward which is measured by distance between exploration network and prediction network
    - Exploitation network: utilize the intrinsic reward and global reward to update the network and interact with environment
  - Other

Yang, Tianpei, et al. "Exploration in deep reinforcement learning: a comprehensive survey." *arXiv preprint arXiv:2109.06668* (2021).