A Non-Monolithic Policy Approach of Offline-to-Online Reinforcement Learning

JaeYoon Kim[1] [0000−0003−0898−5097], Junyu Xuan[1] [0000−0002−8367−6908],
Christy Liang[2] [0000−0001−7179−5208], and Farookh Hussain[1] [0000−0003−1513−8072]

[1] Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia
JaeYoon.Kim@student.uts.edu.au,
{Junyu.Xuan, Farookh.Hussain}@uts.edu.au
[2] Visualisation Institute, University of Technology Sydney, Australia
Jie.Liang@uts.edu.au

A Summary of reference paper [1]

The reference paper [1] is summarized in this research.

A.1 Exploration modes

- Exploit mode (G) : the greedy pursuit of external reward
- Explore mode (X) : uniform random (XU), intrinsic reward based on random network distillation (XI)

A.2 Granularity

Four choices of temporal granularity are considered for exploratory periods.

- Step-level : each step
- Experiment-level : all behaviours produced in explore mode during off-policy training
- Episode-level : each episode
- Intra-episodic : between step-level and episode-level

A.3 Switching mechanisms

Four choices of temporal granularity are considered for exploratory periods. 'Homeostasis' (See A.5) leverages the 'value promise discrepancy', $D_{promise}(t-k,t)$, which represents the difference in the value function over k steps:

$$D_{promise}(t-k,t) := \left|V(s_{t-k}) - \sum_{i=0}^{k-1} \gamma^i R_{t-i} - \gamma^k V(s_t)\right| \quad (1)$$

where $V(s)$ is an agent's value estimate at state s, R is a reward and γ is a discount factor.

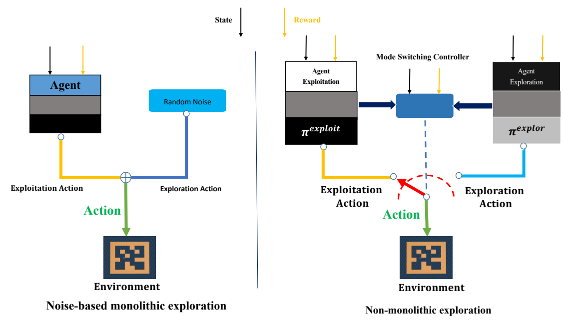


Fig. 1: The diagram illustrates noise-based monolithic exploration (left) and non-monolithic exploration (right). In noise-based monolithic exploration, an agent's action and noise serve roles in exploitation and exploration, respectively. 'Action' represents the final action taken by the agent in the environment, resulting from the addition of the action and noise. In non-monolithic exploration, separate agents for exploitation and exploration operate independently, facilitated by a mode-switching controller. This controller assesses the state of either one agent or both agents, allowing them to pursue their respective objectives.

- Blind switching : not considering any state
- Informed switching : the opposite of blind switching and switching informed by the agent's internal state
- Value promise trigger : triggering according to Eq. 1
- Starting mode : an episode started in explore mode or in exploit mode

A.4 Flexibility to the exploration process

- Bandit adaptation : a meta-controller parameterised by termination probability or target rate
- Homeostasis : the binary switching mechanism based on Algorithm 1

A.5 Homeostasis

The Algorithm 1 of Homeostasis is used for the evaluation implementation of our model. A sequence of scalar signals $x_t \in \mathbb{R}$ is transformed to a sequence of binary switching decisions $y_t \in \{0, 1\}$ for $1 \leq t \leq T$. The binary switching decisions y_t are used for target rate ρ which is the average number of switches approximated to $\frac{1}{T}\Sigma_t y_t \approx \rho \in \{0.1, 0.01, 0.001, 0.0001\}$.

A.6 Variants of intra-episodic exploration

Here is a classification for two-mode intra-episodic exploration for non-monolithic exploration model.

- Explore mode : 1. uniform random (XU). 2. intrinsic reward (XI).

Table 1: The tasks of D4RL environment for experiments.

Task	**D4RL Environment Name**
antmaze-umaze	antmaze-umaze-v2
antmaze-umaze-diverse	antmaze-umaze-diverse-v2
antmaze-medium-play	antmaze-medium-play-v2
antmaze-medium-diverse	antmaze-medium-diverse-v2
antmaze-large-play	antmaze-large-play-v2
antmaze-large-diverse	antmaze-large-diverse-v2
halfcheetah-random	halfcheetah-random-v2
hopper-random	hopper-random-v2
walker-random	walker-random-v2
halfcheetah-medium	halfcheetah-medium-v2
hopper-medium	hopper-medium-v2
walker-medium	walker-medium-v2
halfcheetah-medium-replay	halfcheetah-medium-replay-v2
hopper-medium-replay	hopper-medium-replay-v2
walker-medium-replay	walker-medium-replay-v2

- Explore duration : 1. fixed number of steps (1, 10, 100). 2. adaptively picked by a bandit (represented by *). 3. symmetric switching between entering and exiting explore mode (represented by =).
- Trigger type : 1. blind. 2. informed (based on value promise).
- Exploit duration : 1. blind triggers : a. fixed number of steps (10, 100, 1000, 10000), indirectly defined by a probability of terminating (0.1, 0.01, 0.001, 0.0001) (represented by n*). b. adaptively picked by a bandit over these choices (represented by p*). 2. informed triggers : a. indirectly parameterised by a target rate in (0.1, 0.01, 0.001, 0.0001). b. a bandit over them (represented by p*). 'Informed triggers' (a or b) is transformed into an adaptive switching threshold by homeostasis.
- Starting mode : 1. greedy (G). 2. explore (X).

XU-intra(100,informed,p*,X) is an example instance based on the above classification for the implementation.

Algorithm 1 Homeostasis algorithm taken from [1]

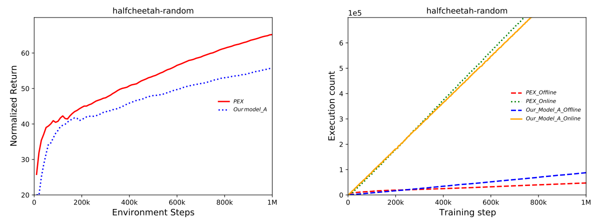
Require:
  Target rate ρ
Initialize:
  $\overline{x} \leftarrow 0$, $\overline{x^2} \leftarrow 1$, , $\overline{x^+} \leftarrow 1$
for $t \in \{1, \dots, T\}$ **do**
  obtain next scalar signal return x_t
  set time-scale $\tau \leftarrow \min(t, \frac{100}{\rho})$
  update moving average $\overline{x} \leftarrow (1 - \frac{1}{\tau})\overline{x} + \frac{1}{\tau} x_t$
  update moving variance $\overline{x^2} \leftarrow (1 - \frac{1}{\tau})\overline{x^2} + \frac{1}{\tau}(x_t - \overline{x})^2$
  standardise and exponentiate $x^+ \leftarrow \exp\left(\frac{x_t - \overline{x}}{\sqrt{\overline{x^2}}}\right)$
  update transformed moving average
    $\overline{x^+} \leftarrow (1 - \frac{1}{\tau})\overline{x^+} + \frac{1}{\tau} x^+$
  sample $y_t \sim$ Bernoulli$\left(\min\left(1, \rho \frac{x^+}{\overline{x^+}}\right)\right)$
end for

Table 2: Hyper-parameters used in the implementation.

Hyper-parameters	**Value**
number of parallel env	1
discount	0.99
replay buffer size	1e6
batch size	256
MLP hidden layer size	[256, 256]
learning rate	3e-4
initial collection steps	5000
target update speed	5e-3
expectile value τ	0.9 (0.7)
inverse temperature α^{-1}	10(3)
number of offline iterations	1M
number of online iterations	1M
number of iteration per rollout step	1
target entropy (SAC)	-*d*

B Tasks and Hyper-Parameters

Task and hyper-parameters are adopted from PEX without modification. The following Table 1 and 2 represents the tasks of D4RL environment and the hyper-parameters for experiments, respectively.



(a) (b)

Fig. 2: Normalized return (left) and execution count (right) of PEX and our model (to stress an arbitrary agent referred to as $Ourmodel\_A$) on 'halfcheetah-random' beyond the execution count of PEX. In (b), the execution counts of the offline policy and online policy of PEX or our model are referred to as $PEX\_Offline$ and PEX_Online or $Our\_Model\_A\_Offline$ and $Our\_Model\_A\_Online$, respectively.

C Ablation study

According to the experiments of our model, the performance drop of our model certainly occurs, as shown in Fig. 2, when the execution count of each policy of our model exceeds that of PEX with the notation used in manuscript as follows.

$$\begin{gathered} OurModel\_Online \leq PEX\_Online \\ and \\ OurModel\_Offline \geq PEX\_Offline \end{gathered} \tag{2}$$

'Ablation study' shows that the execution count of each policy relying on the adjustment of three key parameters—ρ, '$explore\_fixed\_steps$', and '$update\_timestep$'—has an impact on the performance of our model.

References

1. Pislar, M., Szepesvari, D., Ostrovski, G., Borsa, D., Schaul, T.: When should agents explore? arXiv preprint arXiv:2108.11811 (2021)