



Efficient Potential-based Exploration in Reinforcement Learning using Inverse Dynamic Bisimulation Metric

¹State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China



Yiming Wang¹, Ming Yang¹, Renzhi Dong¹, Binbin Sun¹, Furui Liu², Leong Hou U^{1*}

²Zhejiang Lab, Hangzhou, China





Introduction to exploration in RL

• Reward shaping methods

> Rely on human prior knowledge > Introduce human cognitive biases

• Curiosity-driven exploration methods

- > Lack of scalability
- > Rely on count-based episodic term
- Cause policy variance of original MDP

• LIBERTY: expLoration vIa Bisimulation mEtRic-based sTate discrepancY

> Our method (LIBERTY) uses the bisimulation metric to measure state discrepancy and propose a potential function based on the inverse dynamic bisimulation metric, which promotes effective exploration while preserving the optimal policy of the original MDP



Motivation: state discrepancy as exploration bonus



• Using the difference between states $d(s_i, s_j)$ as exploration bonus • Many spikes are related to significant occurrences, e.g., moving forward (1), attacking enemies (4), collecting coins (5) etc. The reward is close to 0 when the agent is stuck (2,3)



Method: bisimulation metric measuring state discrepancy • Bisimulation metric: $d_{\pi}(s_i, s_j) = |r_i^{\pi} - r_j^{\pi}| + \gamma W_1(d_{\pi}) \left(\mathcal{P}^{\pi}(\cdot | s_i), \mathcal{P}^{\pi}(\cdot | s_j) \right)$



Bisimulation Metric: — Euclidean (ℓ_2) norm: \rightarrow Value difference: Contour lines:

X

- denotes value
- ➢ Bisimulation metric identifies the value , starting from initial state s_0





Project the state into 3D latent space and Z axis

differences between states, enabling the agent to reach state s_1 with a higher value compared to s_2

Method: inverse dynamic bisimulation metric

• Meaningless exploration: state difference is caused by background changing without taking actions



• Add inverse dynamic module $(I: S \times S \rightarrow A)$ to avoid meaningless exploration

 $d_{inv}(s_i, s_j) = |r_i^{\pi} - r_j^{\pi}| + \gamma W_2(d_{inv}) \left(\mathcal{P}^{\pi}(\cdot | s_i), \mathcal{P}^{\pi}(\cdot | s_j) \right)$ $+\gamma \|I(\cdot | s_i, s_{i+1}) - I(\cdot | s_j, s_{j+1})\|_1$



Method: inverse dynamic bisimulation metric (cont.)

• Potential function based on d_{inv}

• Potential-based shaping reward function:

$\Phi(s) = d_{inv}(s, s_0)$

 $\mathcal{F}(s_t, a, s_{t+1}) = \gamma d_{in\nu}(s_{t+1}, s_0) - d_{in\nu}(s_t, s_0)$



• Results on MuJoCo continuous control





• Results on Atari games with discrete actions







• Results on the delayed reward setting

Table 1: Quantitative results comparison between LIBERTY and other baseline methods in different environments					
of Mujoco with the delayed reward setting. The best and the runner-up results are (bold) and (underline)					
Delay = 10					
HalfCheetah	Hopper	Walker2d	Ant	Humanoid	Swimmer
1374 ± 368	1258 ± 325	1127 ± 225	-105 ± 43	462 ± 54	27 ± 11
1694 ± 495	1976 ± 458	1405 ± 262	143 ± 17	532 ± 29	32 ± 15
1180 ± 513	989 ± 262	1275 ± 480	-164 ± 35	413 ± 78	24 ± 12
2467 ± 456	1876 ± 431	1651 ± 325	92 ± 31	570 ± 45	65 ± 16
1514 ± 365	2103 ± 129	1997 ± 115	592 ± 67	518 ± 23	43 ± 17
$\textbf{2973} \pm \textbf{437}$	$\textbf{2479} \pm \textbf{315}$	$\textbf{2766} \pm \textbf{487}$	292 ± 68	681 ± 73	73 ± 21
1783 ± 412	1676 ± 275	1732 ± 392	131 ± 22	505 ± 37	46 ± 11
Delay = 40					
HalfCheetah	Hopper	Walker2d	Ant	Humanoid	Swimmer
919 ± 199	857 ± 175	697 ± 172	-213 ± 27	403 ± 34	13 ± 7
1276 ± 387	$\textbf{1683} \pm \textbf{338}$	968 ± 168	71 + 15	483 ± 25	17 ± 11
1028 ± 405	879 ± 155	997 ± 280	-198 ± 27	387 ± 27	11 ± 6
1798 ± 355	1235 ± 269	1025 ± 282	63 ± 18	468 ± 23	32 ± 11
883 ± 275	1382 ± 85	1016 ± 129	105 ± 31	405 ± 15	9 ± 3
$\textbf{2039} \pm \textbf{315}$	$\underline{1612 \pm 215}$	1921 ± 372	142 ± 45	566 ± 35	31 ± 13
1231 ± 253	1213 ± 207	1012 ± 358	58 ± 13	455 ± 27	17 ± 8
	esults comparison layed reward set HalfCheetah 1374 ± 368 1694 ± 495 1180 ± 513 2467 ± 456 1514 ± 365 2973 ± 437 1783 ± 412 HalfCheetah 919 ± 199 1276 ± 387 1028 ± 405 1798 ± 355 883 ± 275 2039 ± 315 1231 ± 253	sults comparison between LIBElayed reward setting. The best andHalfCheetahHopper1374 \pm 3681258 \pm 3251694 \pm 4951976 \pm 4581180 \pm 513989 \pm 2622467 \pm 4561876 \pm 4311514 \pm 3652103 \pm 1292973 \pm 4372479 \pm 3151783 \pm 4121676 \pm 275HalfCheetahHopper919 \pm 199857 \pm 1751276 \pm 3871683 \pm 3381028 \pm 405879 \pm 1551798 \pm 3551235 \pm 269883 \pm 2751382 \pm 852039 \pm 3151612 \pm 2151231 \pm 2531213 \pm 207	sults comparison between LIBERTY and other layed reward setting. The best and the runner-up Delay =HalfCheetahHopperWalker2d 1374 ± 368 1258 ± 325 1127 ± 225 1694 ± 495 1976 ± 458 1405 ± 262 1180 ± 513 989 ± 262 1275 ± 480 2467 ± 456 1876 ± 431 1651 ± 325 1514 ± 365 2103 ± 129 1997 ± 115 2973 ± 437 2479 ± 315 2766 ± 487 1783 ± 412 1676 ± 275 1732 ± 392 Delay =HalfCheetahHopperWalker2d 919 ± 199 857 ± 175 697 ± 172 1276 ± 387 1683 ± 338 968 ± 168 1028 ± 405 879 ± 155 1798 ± 355 1235 ± 269 1025 ± 282 883 ± 275 1382 ± 85 1016 ± 129 2039 ± 315 1612 ± 215 1921 ± 372 1231 ± 253 1213 ± 207 1012 ± 358	sults comparison between LIBERTY and other baseline methodDelay = 10Delay = 10HalfCheetahHopperWalker2dAnt 1374 ± 368 1258 ± 325 1127 ± 225 -105 ± 43 1694 ± 495 1976 ± 458 1405 ± 262 143 ± 17 1180 ± 513 989 ± 262 1275 ± 480 -164 ± 35 2467 ± 456 1876 ± 431 1651 ± 325 92 ± 31 1514 ± 365 2103 ± 129 1997 ± 115 592 ± 67 2973 ± 437 2479 ± 315 2766 ± 487 292 ± 68 1783 ± 412 1676 ± 275 1732 ± 392 131 ± 22 Delay = 40HalfCheetahHopperWalker2dAnt 919 ± 199 857 ± 175 697 ± 172 -213 ± 27 1276 ± 387 1683 ± 338 968 ± 168 $71 + 15$ 1028 ± 405 879 ± 155 997 ± 280 -198 ± 27 $\frac{1798 \pm 355}{1235 \pm 269}$ 1025 ± 282 63 ± 18 883 ± 275 1382 ± 85 1016 ± 129 105 ± 31 2039 ± 315 1612 ± 215 1921 ± 372 142 ± 45 1231 ± 253 1213 ± 207 1012 ± 358 58 ± 13	sults comparison between LIBERTY and other baseline methods in different for a periodDelay = 10HalfCheetahHopperWalker2dAntHumanoid1374 \pm 3681258 \pm 3251127 \pm 225 $-105 \pm$ 43462 \pm 541694 \pm 4951976 \pm 4581405 \pm 262143 \pm 17532 \pm 291180 \pm 513989 \pm 2621275 \pm 480 $-164 \pm$ 35413 \pm 782467 \pm 4561876 \pm 4311651 \pm 32592 \pm 31 $570 \pm$ 451514 \pm 3652103 \pm 1291997 \pm 115 592 \pm 67518 \pm 232973 \pm 4372479 \pm 3152766 \pm 487292 \pm 68681 \pm 731783 \pm 4121676 \pm 2751732 \pm 392131 \pm 22505 \pm 37Delay = 40HalfCheetahHopperWalker2dAntHumanoid919 \pm 199857 \pm 175697 \pm 172 $-213 \pm$ 27403 \pm 341276 \pm 3871683 \pm 338968 \pm 16871 \pm 15483 \pm 251028 \pm 405879 \pm 155997 \pm 280 $-198 \pm$ 27387 \pm 271798 \pm 3551235 \pm 2691025 \pm 28263 \pm 18468 \pm 2383 \pm 2751382 \pm 851016 \pm 129105 \pm 31405 \pm 152039 \pm 3151612 \pm 2151921 \pm 372142 \pm 45566 \pm 351231 \pm 2531213 \pm 2071012 \pm 35858 \pm 13455 \pm 27



• Results on the reward-free setting





Contribution

- need for prior human knowledge
- explore states with higher TD-error
 - > Theorem 1(value difference bound):

with other methods

• We develop a new potential function to ensure policy invariance without the

• Our approach achieves more efficient exploration by encouraging agents to

 $\left|V^{\pi}(s_i) - V^{\pi}(s_j)\right| \leq d_{in\nu}(s_i, s_j)$ > Theorem 2(approximation of optimal value function): $d_{inv}(s,s_0) \approx V^*(s)$

• Our method achieves best performance across various settings in extensive environments, which demonstrate its scalability and superiority compared



Thank you for your attention!



澳門大學

UNIVERSIDADE DE MACAU UNIVERSITY OF MACAU









Code is available