MDPMorph: An MDP-Based Metamorphic Testing Framework for Deep Reinforcement Learning Agents

Anonymous Author(s)

I. Appendix

Based on the assumptions and definitions outlined in Section III-A, we now provide sketch proofs to verify that the proposed MRs satisfy the necessary properties of DRL systems.

MR1.

In the work by Kos et al. [1], the authors conducted extensive experiments involving random perturbations to the state with varying noise levels. They observed that when the noise intensity was less than or equal to 0.02, it had virtually no impact on the agent's performance. Therefore, for a well-trained agent, introducing small-scale random variations in a continuous action space does not significantly alter the action vector—there exists a small threshold τ_1 such that $||\pi(s_0^1) - \pi(s_0^2)|| < \tau_1$. In discrete action spaces, the selected action category remains nearly unchanged, i.e., $\pi(s_0^1) = \pi(s_0^2)$. □

MR2.

Based on Definition 6, it is evident that if an agent has been fully trained and converged to the optimal policy, it should, at each decision step, select the action that maximizes the expected return in the current state, thereby ensuring the maximization of rewards at every step.

When the source test case and the follow-up test case are identical, the agent gradually approaches the optimal long-term behavior as the steps progress. Therefore, the cumulative reward at the i-th step of the agent under the source test case, denoted as $\sum_{k=0}^{i} r_k^1$, accumulates more reward through a series of optimal decisions and is expected to differ from the cumulative reward at the j-th step ($i > j$) under the follow-up test case $\sum_{k=0}^{j} r_k^2$. Considering the noise in the environment, the agent might execute an action that is "close to optimal" but not strictly optimal in certain states. Therefore, we introduce a threshold τ_2 in MR2, denoted as $\sum_{k=0}^{i} r_k^1 - \sum_{k=0}^{j} r_k^2 > \tau_2$. □

MR3.

Based on Assumptions 1, 2, and 3, we use the method of recurrence and telescoping summation to establish the proof.

$$\begin{aligned}\Delta_{s_t} &= ||s_t^1 - s_t^2|| \\ &= ||f_t^1(s_{t-1}^1, a_{t-1}^1) + \eta_t^1 - f_t^2(s_{t-1}^2, a_{t-1}^2) - \eta_{t-1}^2|| \\ &\le ||f_t^1(s_{t-1}^1, a_{t-1}^1) - f_t^2(s_{t-1}^2, a_{t-1}^2)|| + ||\eta_{t-1}^1 - \eta_{t-1}^2|| \\ &\le L_p(||s_{t-1}^1 - s_{t-1}^2|| + ||a_{t-1}^1 - a_{t-1}^2||) + 2\xi \\ &= L_p\Delta_{s_{t-1}} + 2L_pM + 2\xi\end{aligned} \tag{1}$$

where, M denotes the boundary value of the optimal policy set.

Since the follow-up test cases are derived from the original test case through slight modifications—such as perturbation, translation, or scaling—therefore, $s_0^1 - s_0^2 = e_0$. By applying Equation 1, and using the method of constant-coefficient nonhomogeneous linear recurrence, we obtain:

$$\Delta_t \le L_p^t||e_0|| + 2(\xi + L_pM)\sum_{i=0}^{t-1} L_p^i \tag{2}$$

if f is the total number of timestep, then:

$$\begin{aligned}&D(\{s_{f-k}^1, s_{f-k+1}^1 ... s_f^1\}, \{s_{f-k}^2, s_{f-k+1}^2 ... s_f^2\}) \\ &= \Delta_{s_f} + \Delta_{s_{f-1}} + ... + \Delta_{s_{f-k}} \\ &\le ||e_0|| \sum_{i=f-k}^{f} L_p^i + 2(\xi + L_pM)\sum_{j=0}^{k}\sum_{i=0}^{f-k-1+j} L_p^i \\ &= B_f\end{aligned} \tag{3}$$

where,

$$B_f = \begin{cases} ||e_0|| \sum_{i=f-k}^{f} L_p^i + 2(\xi + L_pM)\left(\frac{k+1}{1-L_p} - \frac{L_p^{f-k}(1-L_p^{k+1})}{(1-L_p)^2}\right) & \text{if } L_p \neq 1 \\ ||e_0|| \sum_{i=f-k}^{f} L_p^i + (\xi + L_pM)(2f-k)(k+1) & \text{if } L_p = 1 \end{cases}$$

if $\tau_3 > B_f$, then $D(\{s_{f-k}^1, s_{f-k+1}^1 ... s_f^1\}, \{s_{f-k}^2, s_{f-k+1}^2 ... s_f^2\}) < \tau_3$. □

MR4.

Based on Assumption 1, 3 and Definition 6, we use induction to prove.

The state transition at time t can be represented by a deterministic function f_t along with a noise term, where the noise is bounded above by ξ. That is, for any step i, there exists $||\eta_i^1 - \eta_i^2|| \le 2\xi$. For the state at $t+1$ step, there exists: $s_t^1 = f_t^1(s_{t-1}^1, a_{t-1}^1) + \eta_{t-1}^1$, $s_t^2 = f_t^2(s_{t-1}^2, a_{t-1}^2) + \eta_{t-1}^2$.

Therefore,

$$\begin{aligned}\Delta_{s_t} &= ||s_t^1 - s_t^2|| \\ &= ||f_t^1(s_{t-1}^1, a_{t-1}^1) + \eta_{t-1}^1 - f_t^2(s_{t-1}^2, a_{t-1}^2) - \eta_{t-1}^2|| \\ &\le ||f_t^1(s_{t-1}^1, a_{t-1}^1) - f_t^2(s_{t-1}^2, a_{t-1}^2)|| + ||\eta_{t-1}^1 - \eta_{t-1}^2|| \\ &\le L_p(||s_{t-1}^1 - s_{t-1}^2|| + ||a_{t-1}^1 - a_{t-1}^2||) + 2\xi \\ &= L_p\Delta_{s_{t-1}} + 2L_pM + 2\xi\end{aligned} \tag{4}$$

where, M denotes the upper bound of the maximum norm of action output differences between any two optimal policies at each state.

Through $s_0^1 = s_0^2$, we can derive the following through recursion:

$$\Delta_{s_t} \leq 2(\xi + L_p M) \sum_{i=0}^{t-1} L_p^i \tag{5}$$

If f is the total number of time step, then:

$$\begin{aligned} D(S^1, S^2) &= \Delta_{s_0} + \Delta_{s_1} + ... + \Delta_{s_f} \\ &\leq 2(\xi + L_p M)(f + (f-1)L_p + ... + (L_p)^{f-1}) \\ &= 2(\xi + L_p M) \sum_{i=0}^{f-1} (f-k)(L_p)^i \end{aligned} \tag{6}$$

If $\tau_4 > 2(\xi + L_p M) \sum_{i=0}^{f-1} (f-k)(L_p)^i$, then $D(S^1, S^2) < \tau_4$. □

MR5

Based on Assumption 1 and Definition 6, we use induction to prove.

The reward at time t can be represented by a deterministic function $R(s_t, a_t)$. Therefore, for the reward at step $t+1$, there exists: $r_{t+1}^1 = R(s_t^1, a_t^1)$, $r_{t+1}^2 = R(s_t^2, a_t^2)$.

Therefore, according to Equation 5, we have:

$$\begin{aligned} \Delta_{r_{t+1}} &= ||r_{t+1}^1 - r_{t+1}^2|| \\ &= ||R(s_t^1, a_t^1) - R(s_t^2, a_t^2)|| \\ &\leq L_r(||s_t^1 - s_t^2|| + ||a_t^1 - a_t^2||) \\ &= L_r \Delta_{s_t} + 2L_r M \\ &\leq 2L_r(\xi + L_p M) \sum_{i=0}^{t-1} L_p^i + 2L_r M \end{aligned} \tag{7}$$

where, M denotes the upper bound of the maximum norm of action output differences between any two optimal policies at each state.

If f is the total number of time step, then:

$$\begin{aligned} D(R^1, R^2) &= \Delta_{r_0} + ... + \Delta_{r_f} \\ &\leq 2L_r M + ... + 2L_r(\xi + L_p M) \sum_{i=0}^{f-1} L_p^i + 2L_r M \\ &= 2L_r M + \sum_{t=1}^{f} \left[2L_r(\xi + L_p M) \sum_{i=0}^{t-1} L_p^i + 2L_r M \right] \\ &= 2L_r M(f+1) + 2L_r(\xi + L_p M) \sum_{i=0}^{f-1} (f-i) L_p^i \end{aligned} \tag{8}$$

if $\tau_5 > 2L_r M(f+1) + 2L_r(\xi + L_p M) \sum_{i=0}^{f-1} (f-i) L_p^i$, then $D(R^1, R^2) < \tau_5$ □

REFERENCES

[1] J. Kos and D. Song, "Delving into adversarial attacks on deep policies," *arXiv preprint arXiv:1705.06452*, 2017.