Any-step Dynamics Model Improves Future Predictions for Online and Offline Reinforcement Learning

Haoxin Lin[1,2,3], **Yu-Yan Xu**[3], **Yihao Sun**[1,2], **Zhilong Zhang**[1,2,3], **Yi-Chen Li**[1,2,3], **Chengxing Jia**[1,2,3], **Junyin Ye**[1,2,3], **Jiaji Zhang**[1,2], **Yang Yu**[1,2,3]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

[2]School of Artificial Intelligence, Nanjing University, Nanjing, China

[3]Polixir Technologies, Nanjing, China

Presented by **Haoxin Lin**

Background of MBRL

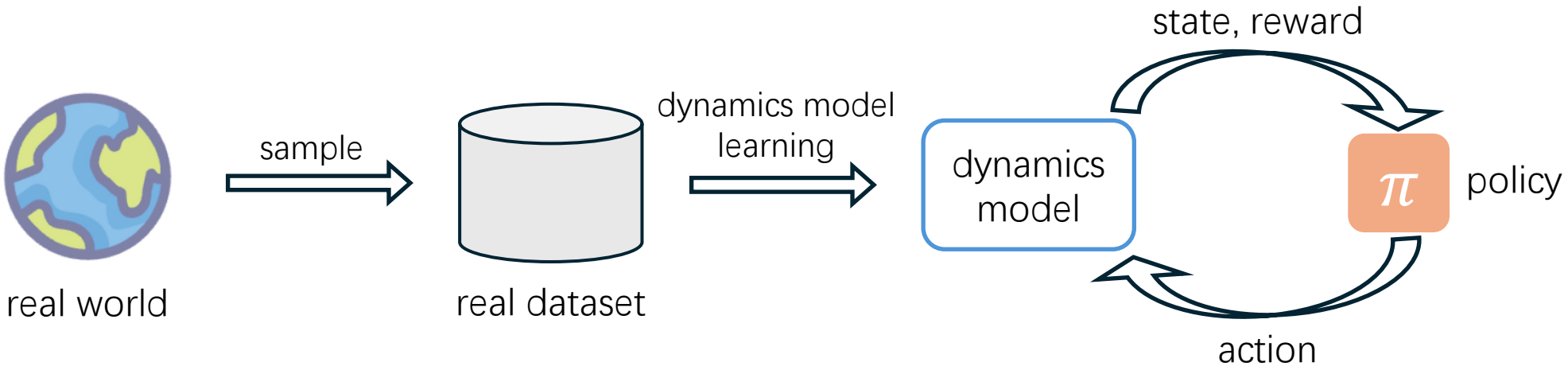


Figure 1. MBRL pipeline

- Model-based Reinforcement Learning (**MBRL**) can reduce the reliance of policy optimization on real-world data
  - learning a dynamics model $T(s', r|s, a)$ from real-world data
  - facilitating policy exploration within the learned dynamics model

Background of MBRL

- Bootstrapping prediction is unavoidable during roll-out in the single-step dynamics model,
  - which attributes the next state to the prediction of the current state at each step

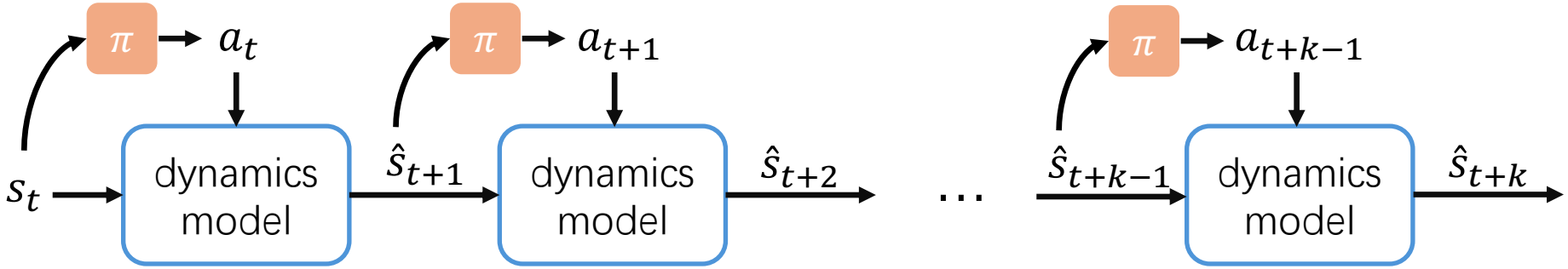
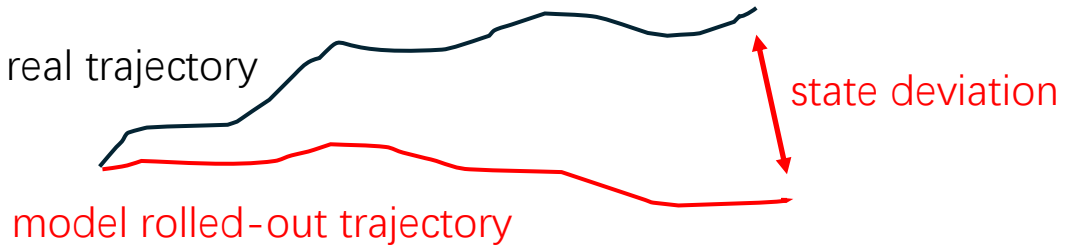


Figure 2. k-step roll-outs in the single-step dynamics model.

- Bootstrapping predictions lead to compounding errors[1] during model roll-out
  - the deviation of generated states increases with the roll-out length



[1] Tian Xu et al. Error bounds of imitating policies and environments for reinforcement learning. IEEE TPAMI, 2021.

Our Method

- We propose the **A**ny-step **D**ynamics **M**odel (ADM) that allows for the use of variable-length plans as inputs for predicting future states without frequent bootstrapping.

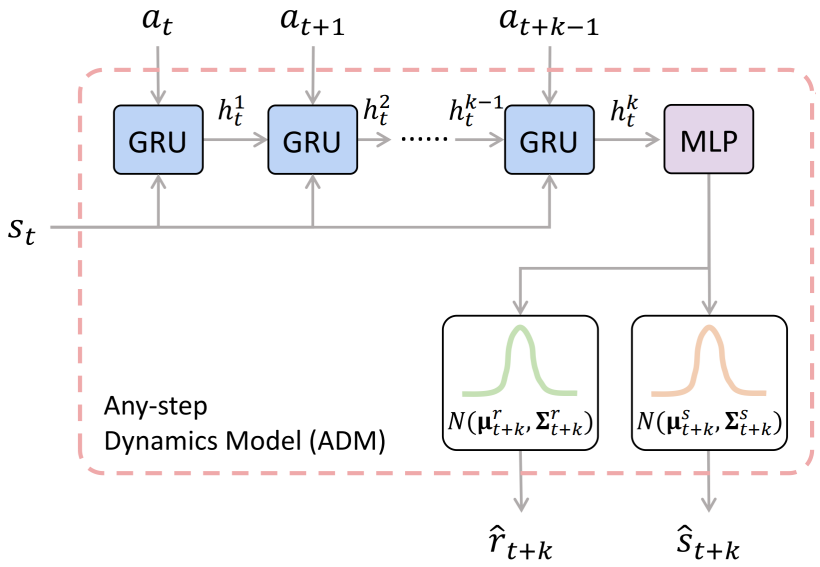


Figure 3. Any-step dynamics model structured using an RNN.

- ADM can predict future states across multiple steps, reducing the number of bootstrapping during the model roll-out.

Our Method

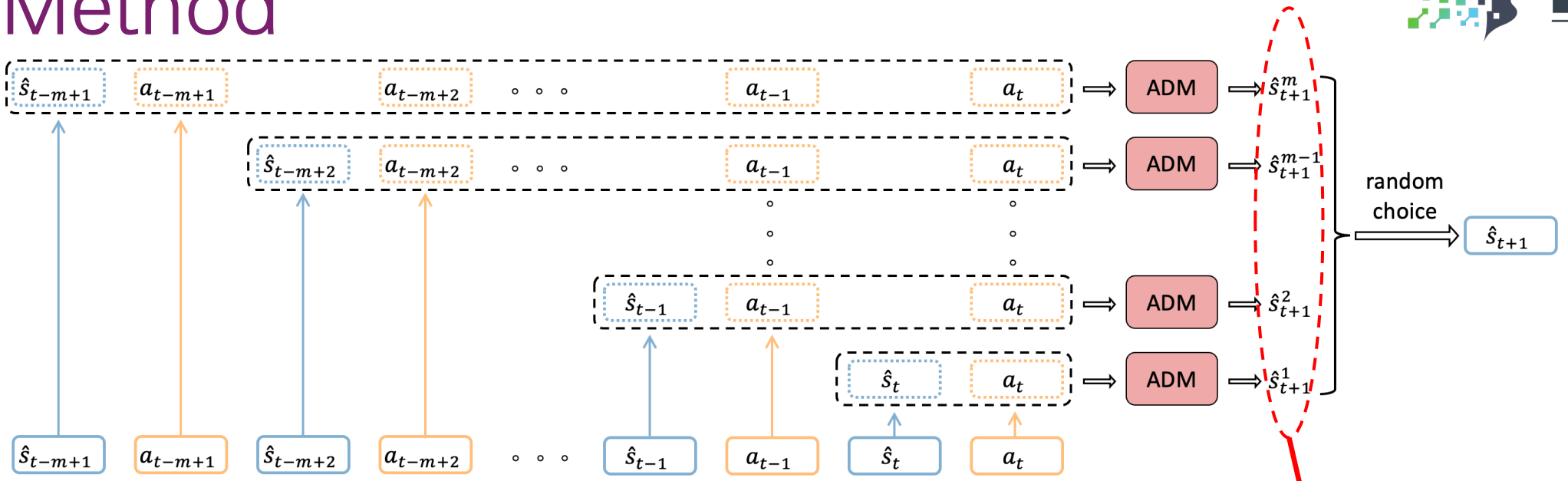


Figure 4. Next-state prediction of ADM.

- ADM naturally possesses the ability to generate diverse predictions.

  While predicting $\hat{s}_{t+1}$, ADM has m choices:

  - using $(\hat{s}_{t-m+1}, a_{t-m+1}, a_{t-m+2}, \ldots, a_t)$ as inputs;
  - using $(\hat{s}_{t-m+2}, a_{t-m+2}, \ldots, a_t)$ as inputs;
  - ...
  - using $(\hat{s}_t, a_t)$ as inputs.

The variance among several prediction choices can be used as an uncertainty quantifier.

Experiments

- Evaluation of Compounding Error

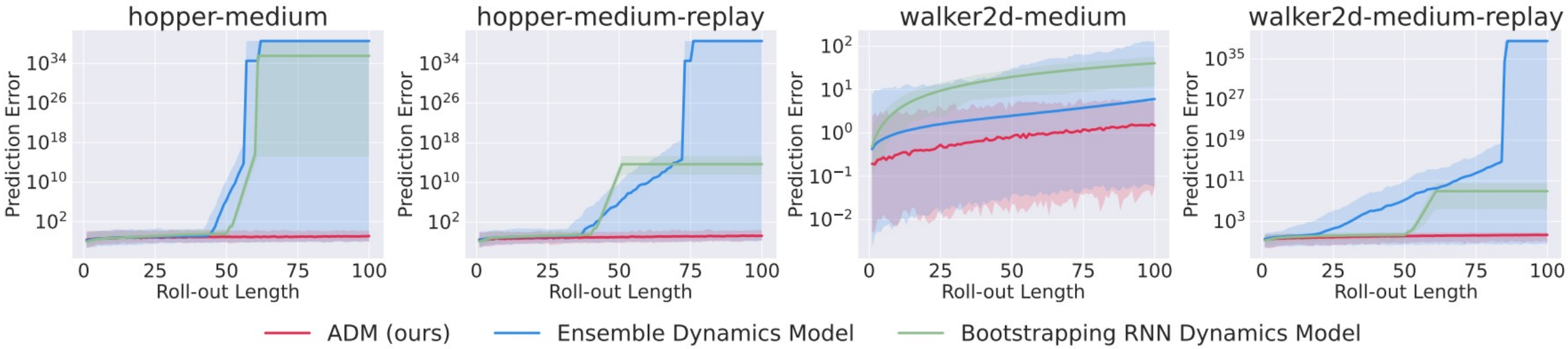


Figure 5. Growth curves of the compounding error (in log scale).

- the growth rate of ADM's compounding error curve is smaller compared to other dynamics models

Experiments

- Offline Performance (policy optimization within a trained ADM)
  - D4RL results

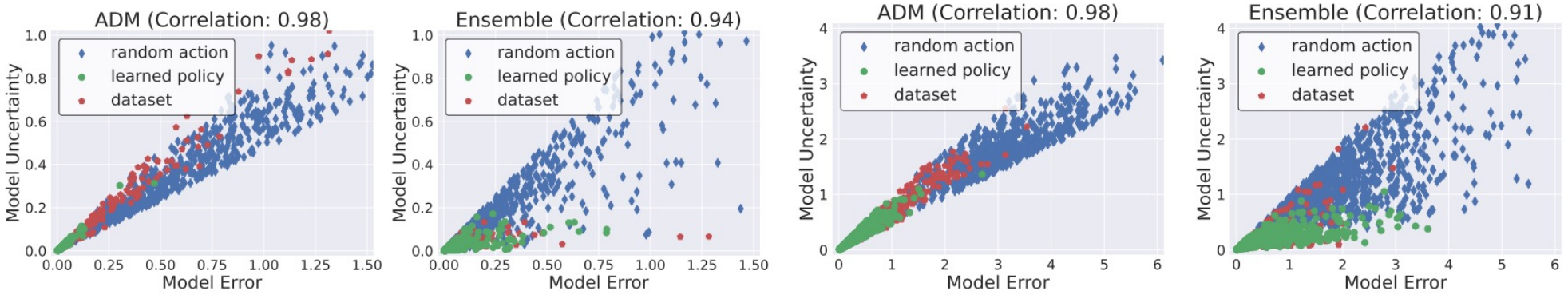
Task Name	BC	CQL	TD3+BC	EDAC	MOPO	COMBO	RAMBO	CBOP	MOBILE	ADMPO-OFF (ours)
hopper-random	3.7	5.3	8.5	25.3	31.7	17.9	25.4	31.4	31.9	**32.7±0.2**
halfcheetah-random	2.2	31.3	11.0	28.4	38.5	38.8	39.5	32.8	39.3	**45.4±2.8**
walker2d-random	1.3	5.4	1.6	16.6	7.4	7.0	0.0	17.8	17.9	**22.2±0.2**
hopper-medium	54.1	61.9	59.3	101.6	62.8	97.2	87.0	102.6	106.6	**107.4±0.6**
halfcheetah-medium	43.2	46.9	48.3	65.9	73.0	54.2	**77.9**	74.3	74.6	72.2±0.6
walker2d-medium	70.9	79.5	83.7	92.5	84.1	81.9	84.9	**95.5**	87.7	93.2±1.1
hopper-medium-replay	16.6	86.3	60.9	101.0	103.5	89.5	99.5	**104.3**	103.9	**104.4±0.4**
halfcheetah-medium-replay	37.6	45.3	44.6	61.3	**72.1**	55.1	68.7	66.4	71.7	67.6±3.4
walker2d-medium-replay	20.3	76.8	81.8	87.1	85.6	56.0	89.2	92.7	89.9	**95.6±2.1**
hopper-medium-expert	53.9	96.9	98.0	110.7	81.6	111.1	88.2	111.6	**112.6**	**112.7±0.3**
halfcheetah-medium-expert	44.0	95.0	90.7	106.3	90.8	90.0	95.4	105.4	**108.2**	103.7±0.2
walker2d-medium-expert	90.1	109.1	110.1	114.7	112.9	103.3	56.7	**117.2**	115.2	114.9±0.3
Average	36.5	61.6	58.2	76.0	70.3	66.8	67.7	79.3	80.0	**81.0**

  - NeoRL results

Task Name	BC	CQL	TD3+BC	EDAC	MOPO	MOBILE	ADMPO-OFF (ours)
neorl-hopper-low	15.1	16.0	15.8	18.3	6.2	17.4	**22.3±0.1**
neorl-halfcheetah-low	29.1	38.2	30.0	31.3	40.1	**54.7**	52.8±1.2
neorl-walker2d-low	28.5	44.7	43.0	40.2	11.6	37.6	**55.9±3.8**
neorl-hopper-medium	51.3	64.5	**70.3**	44.9	1.0	51.1	51.5±5.0
neorl-halfcheetah-medium	49.0	54.6	52.3	54.9	62.3	**77.8**	69.3±1.7
neorl-walker2d-medium	48.7	57.3	58.5	57.6	39.9	62.2	**70.1±2.4**
neorl-hopper-high	43.1	76.6	75.3	52.5	11.5	**87.8**	**87.6±4.9**
neorl-halfcheetah-high	71.3	77.4	75.3	81.4	65.9	83.0	**84.0±0.8**
neorl-walker2d-high	72.6	75.3	69.6	75.5	18.0	74.9	**82.2±1.9**
Average	45.4	56.1	54.5	50.7	28.5	60.7	**64.0**

Experiments

- Uncertainty Quantification
  - Comparison between ADM and ensemble dynamics model



(a) hopper-medium-replay-v2 (b) walker2d-medium-replay-v2

Figure 6. Comparison between ADM and ensemble model in uncertainty quantification.

  - ADM has better consistency between uncertainty and model error.
  - ADM better distinguishes different types of policies.

ICLR

Thank you!