Categorical Decision Mamba : On Tractable Binning in Sequence Model based Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, sequence modeling methods have been applied to solve the problem of off-policy reinforcement learning. One notable example is the work on Decision Mamba, incorporating Mamba block into the Decision-Transformer-type neural network architecture. In this work, we begin our exploration with the latest sequential decision-making model, leveraging its strengths as a foundation for further development. We propose a theoretical measure of alignment on the policy of the agent with the human expert, known as Expected Agent Alignment Error (EA2E). Furthermore, we provide a complete theoretical proof that reducing the Wasserstein-1 distance between distributions of the present model (agent) and the target model (agent) effectively aligns the agent's policy with the potential expert's. Building upon theoretical results, we propose Categorical Decision Mamba (CDMamba), which originates from Decision Mamba (DMamba). The core improvements of CDMamba involve utilizing histograms of categorical distributions as inputs to the Mamba model, minimizing the Wasserstein-1 distance between distributions, which ultimately yields a trained model with aligned policy and superior performance.

1 Introduction

Offline Reinforcement Learning [1, 2, 3, 4] has been a promising approach for training agents that does not necessitate online experience in an environment, which is advantageous when online experience is expensive or when offline experience is abundant. In the past few years, Transformers [5] have shown impressive results across a number of problem domains in Natural Language Processing [6, 7, 8] and Computer Vision [9, 10]. Inspired by these recent successes, formulating offline reinforcement learning as a sequence modeling problem [1, 2, 3, 4, 11] has become a novel idea for solution, where the Transformer model predicts the next element in a sequence of states, actions, rewards, and then tackled the issue with techniques similar to those employed in large language models. The attention mechanism in Transformer does have produced several impressive results [1, 11]; however, investigating alternative mechanisms to further enhance model performance remains an open and intriguing research question [3].

Recently, state space sequence models (SSMs) have gained popularity as efficient and effective building blocks for constructing deep networks, achieving great performance in analyzing continuous long-sequence data [12, 13]. In particular, structured state space sequence models (S4) have been effective in various applications [12, 14]. Mamba [15] enhances S4 [12] by incorporating a selective mechanism, allowing the model to selectively focus on input-dependent, relevant information. This improvement, combined with hardware-aware implementation, enables Mamba to outperform Transformers on dense modalities, such as language and genomics. Leveraging the numerous advantages of the Mamba architecture and significant success achieved in various domains [16, 17, 18, 19], **Decision Mamba** [3] investigate the integration of the Mamba framework as a new architectural

Submitted to Automated Reinforcement Learning Workshop at International Conference on Machine Learning (ICML). Do not distribute.

choice within the Decision Transformer. Empirical study in the work Ota [3] shows that Decision Mamba is competitive to existing DT-type models, suggesting the effectiveness of Mamba framework for RL tasks.

In work [20], the exploration problem is reformulated as a State Marginal Matching (SMM) issue, in which a target state distribution is given, and a policy is learned to make the state marginal distribution match this target distribution. Inspired by such notion, Furuta et al. [2] proposed the Categorical Decision Transformer (CDT) [2], taking histograms of categorical distribution (i.e. discrete approximations of feature distributions; $\mathcal{B}$-dim vector) as the inputs of the transformer. With obtaining the desired information statistics for all trajectories, they are fed to the Categorical Decision Transformer during training/test time. While such approach is effective for reinforcement learning in multiple tasks, it is crucial to consider the limitations of training a black-box model. Specifically, for an agent generated by such large model, we cannot help but question: What about the alignment of the trained agent? Furthermore, does the reduction in Wasserstein-1 distance between distributions of existing model (agent) and target model (agent) effectively aligns the agent's policy with the potential human expert's?

Inspired by the innovative frameworks in sequential decision-making, such as Decision-Mamba [3], we would like to reevaluating the role of SMM method in sequence-based decision-making. Therefore, in this paper, we propose **Categorical Decision Mamba: further taking the histograms of categorical distribution as the input of Mamba, minimizing the distance between distributions and finally get a trained model with excellent and aligned policy.**

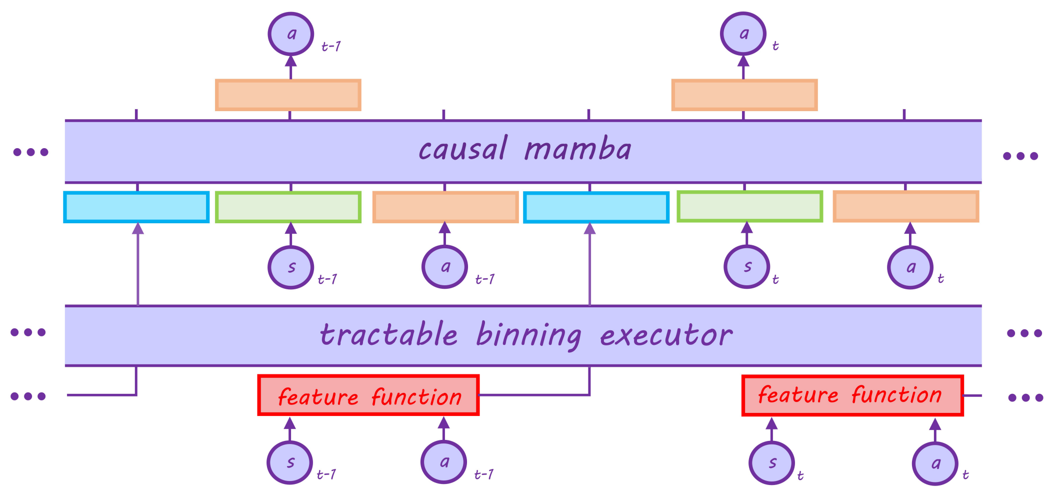

Figure 1: Categorical Decision Mamba (CDMamba) architecture

To summary, the contributions of this work are as follows:

- We propose a theoretical measure of alignment towards the policy of an agent embedded in a large model within the background of state marginal matching.
- We provide a theoretical proof that reducing Wasserstein-1 distance between distributions of existing and target models aligns the agent's policy with the potential human expert's.
- We present an enhanced approach for Decision Mamba (DMamba) termed as **Categorical Decision Mamba (CDMamba).**

2 Methodology

2.1 Expected Agent Alignment Error (EA2E)

Offline reinforcement learning (RL) [1, 2, 3, 4] has been a promising approach for training agents that does not necessitate online experience in an environment, which is advantageous when online experience is expensive or when offline experience is abundant. Notwithstanding, it also gives rise to several other concerns. One notable issue is that the policy learned by an agent may deviate significantly from the underlying human strategy, yet still achieve a higher return, leading to the misconception that a good policy has been learned. To facilitate a comprehensive evaluation of the learned policy, it is essential to define a metric that explains the confidence of the model (agent) in its strategy. Similar to ECE [21], we define the Expected Agent Alignment Error (EA2E).

Definition 2.1. *Defining $U = \mathbb{I}\left[dis\left(\pi_1, \pi_2\right) \leq \delta\right]$, where $\mathbb{I}(\cdot)$ is the the indicator function, U = 1 denotes that the distance between policy π_1 and policy π_2 is smaller than the threshold δ. Hence, for the model (agent), a perfect alignment (full confidence) can be expressed as:*

$$\mathbb{P}\left(\mathrm{U} = 1 | \hat{s} = s\right) = s \qquad \forall s \in [0,1] \tag{1}$$

Definition 2.2. *Expected Agent Alignment Error (EA2E): Define the misalignment of the model (agent) by computing the expectation of alignment error over predicted confidence $\hat{s}$:*

$$\mathrm{EA2E} = \mathbb{E}_{\hat{s}}\left[\left|\mathbb{P}(\mathrm{U} = 1 | \hat{s} = s) - s\right|\right] \tag{2}$$

2.2 Theoretical Analysis

In this section, we theoretically delve deeper into the benefits of mitigating distribution discrepancies, demonstrating the superiority of our approach.

Theorem 1. *Suppose $\pi_{\theta_1}(a|s)$ and $\pi_{\theta_2}(a|s)$ are two policies of the model (agent), $\rho^{\pi_{\theta_1}}(s,a)$ and $\rho^{\pi_{\theta_2}}(s,a)$ are the state-action marginal distributions of two agents (models), then:*

$$\mathrm{EA2E}(\pi_{\theta_1}) - \mathrm{EA2E}(\pi_{\theta_2}) \leq \mathbb{E}\left[4 \cdot \mathrm{TV}(\rho^{\pi_{\theta_1}}(s,a), \rho^{\pi_{\theta_2}}(s,a)\right] \tag{3}$$

, where $\mathrm{TV}(\cdot)$ *is the total variation distance.*

Proposition 1. *Suppose $\pi(a|s)$ is the potential policy of human expert , $\pi_\theta(a|s)$ is the policy of model (agent), $\rho^{\pi}(s,a)$ is the state-action marginal distribution of potential human expert agent and $\rho^{\pi_\theta}(s,a)$ is the state-action marginal distribution of current agent (model), then:*

$$\mathrm{EA2E}(\pi) - \mathrm{EA2E}(\pi_\theta) \leq \mathbb{E}\left[\frac{4}{d_{min}} \cdot \mathrm{W}_1\left(\rho^{\pi}(s,a), \rho^{\pi_\theta}(s,a)\right)\right] \tag{4}$$

, where $\mathrm{W}_1\left(\rho^{\pi}(s,a), \rho^{\pi_\theta}(s,a)\right)$ is the Wasserstein-1 distance between $\rho^{\pi}(s,a)$ and $\rho^{\pi_\theta}(s,a)$; let $\rho^{\pi}(s,a) \in \mu$ and $\rho^{\pi_\theta}(s,a) \in \nu$, setting $\Omega = supp(\mu) \cup supp(\nu)$, $d_{min} = \inf\limits_{\rho^{\pi} \neq \rho^{\pi_\theta} \in \Omega} \|\rho^{\pi} - \rho^{\pi_\theta}\|$.

In Theorem 1, it is shown that the differences of EA2E is bounded by total variation distance of the policies. And in Proposition 1, we can observe that the deviation of EA2E between the potential human expert and the current model (agent) is bounded by Wasserstein-1 distance between the distribution of the expert agent and the current agent (model). Detailed proofs are available in Appendix A and Appendix B.

2.3 Categorical Decision Mamba

Recently, the Mamba framework have been introduced [15], known as an sequence modeling framework that leverages a selective structured state space model to achieve efficient and effective performance. Decision Mamba is a novel approach that replaces traditional self-attention with Mamba block. Empirical verification has shown that such modification can improve the model's capacity of capturing complex dependencies in sequential decision-making tasks, thereby potentially enhancing its decision-making capabilities in diverse and challenging environments. Based on such preliminary, we construct categorical approximations of continuous distributions by leveraging the discretization of feature spaces as a substitute for return-to-go (RTG). The architecture of Categorical Decision Mamba (CDMamba) is shown in Figure 1. From the architecture, we can see that CDMamba takes binnings of distribution (rewards or state dimensions like xyz-velocities), states and actions as input and actions of future as output. And furthermore, we also evaluate the model with Wasserstein-1 distance between categorical distributions of features, in order to demonstrate the effectiveness of state-feature distribution matching.

3 Experiments

We conduct the experiments on the MuJoCo tasks (Halfcheetah, Hopper and Walker2d) from the widely-used D4RL [22] benchmarks. Firstly, we sort all the trajectories by their cumulative rewards. For comparison with CDT [2], we similarly hold out five best trajectories and five 50 percentile trajectories as a test set (10 trajectories in total), and use the rest as a train set. We report the results averaged over 20 rollouts every 3 random seeds in Table 1. We select DT [1], RvS [23], DS4 [24], DC [25] and DMamba [3] as our baselines. The results show that CDMamba is competitive to DMamba and existing DT-type models, suggesting an effectiveness of CDMamba architecture for RL tasks. Furthermore, we evaluate the Wasserstein-1 distance between categorical distributions of features in Table 2. The practical computation of Wasserstein-1 distance is conducted by package in [26]. Distance datas of DT and CDT are from work [8]. The results show that CDMamba matches approximate distribution much better than CDT and DT. For reproducibility, we provide the hyperparameter of experiments in Table 3 (Appendix D). And for intuitive understanding, we provide typical visualizations in Figure 2 (Appendix C).

Table 1: The offline results of CDMamba and baselines in MuJoCo domain. We abbreviate dataset names as follows: 'medium' as 'm', 'medium-replay' as 'm-r', 'medium-expert' as 'm-e'.

Dataset	DT	RvS	DS4	DC	DMamba	CDMamba
halfcheetah-m	42.6	41.6	42.5	43.0	42.8±0.08	43.2±0.06
hopper-m	68.4	60.2	54.2	92.5	83.5±12.5	70.7±1.66
walker2d-m	75.5	71.7	78.0	79.2	78.2±0.6	79.2±0.12
halfcheetah-m-r	37.0	38.0	15.2	41.3	39.6±0.1	39.9±0.05
hopper-m-r	85.6	73.5	49.6	94.2	82.6±4.6	88.8±2.17
walker2d-m-r	71.2	60.6	69.0	76.6	70.9±4.3	78.2±2.88
halfcheetah-m-e	88.8	92.2	92.7	93.0	91.9±0.6	88.4±1.89
hopper-m-e	109.6	101.7	110.8	110.4	111.1±0.3	110.6±0.53
walker2d-m-e	109.3	106.0	105.7	109.6	108.3±0.5	108.6±0.39

Table 2: Quantitative evaluation of reward distribution matching via measuring Wasserstein-1 distance between the rollout and target distributions.

Method	Halfcheetah		Hopper		Walker2d	
	m	m-e	m	m-e	m	m-e
DT	1.039±1.548	0.846±1.134	0.091±0.035	0.159±0.111	0.626±0.495	0.341±0.452
CDT	1.002±1.458	0.838±1.054	0.064±0.017	0.111±0.077	0.114±0.037	0.105±0.030
CDMamba	0.166±0.044	0.469±0.229	0.081±0.025	0.134±0.026	0.215±0.014	0.028±0.019

4 Discussion

In this work, we provide a novel perspective on the agent policy alignment: decreasing Wasserstein-1 distance between distributions of existing and target models aligns the agent's policy with that of potential human experts. Building upon the theoretical foundations, we introduce CDMamba, a novel approach that synergistically integrates the superior performance of Mamba in handling sequential problems with strengths of distributional matching for agent alignment. Hence, this work makes theoretical contributions to the field of agent alignment in reinforcement learning; and further validates the Mamba architecture's robust representational capabilities in sequence model-based reinforcement learning tasks, offering valuable insights for future research in this area.

In future work, we would like to further explore the agent policy alignment in reinforcement learning based on sequence models, foundation models and large language models; and further study practical methods of agent policy alignment. Moreover, we also would like to study additional implementations that better effectively leverage Mamba's or other sequence modeling architectures' advantages in the sequence model based reinforcement learning.

References

[1] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021.

[2] Hiroki Furuta, Yutaka Matsuo, and Shixiang Shane Gu. Generalized decision transformer for offline hindsight information matching, 2022.

[3] Toshihiro Ota. Decision mamba: Reinforcement learning via sequence modeling with selective state spaces, 2024.

[4] Hao Liu and Pieter Abbeel. Emergent agentic transformer from chain of hindsight experience, 2023.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[6] OpenAI et.al. Gpt-4 technical report, 2024.

[7] Hugo Touvron and Louis Martin et.al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[8] Gemini Team et.al. Gemini: A family of highly capable multimodal models, 2024.

[9] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys*, 54(10s):1–41, January 2022. ISSN 1557-7341. doi: 10.1145/3505244.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[11] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems*, 2021.

[12] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022.

[13] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state-space layers, 2021.

[14] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models, 2022.

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

[16] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation, 2024.

[17] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024.

[18] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024.

[19] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation, 2024.

[20] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching, 2020.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[22] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021.

[23] Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. Rvs: What is essential for offline rl via supervised learning?, 2022.

[24] Shmuel Bar-David, Itamar Zimerman, Eliya Nachmani, and Lior Wolf. Decision s4: Efficient sequence-based rl via state spaces layers, 2023.

[25] Jeonghye Kim, Suyoung Lee, Woojun Kim, and Youngchul Sung. Decision convformer: Local filtering in metaformer is sufficient for decision making, 2023.

[26] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. The hidden geometry of particle collisions. *Journal of High Energy Physics*, 2020(7), July 2020. ISSN 1029-8479. doi: 10.1007/jhep07(2020)006.

[27] Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, March 2019. ISSN 2326-831X. doi: 10.1146/annurev-statistics-030718-104938.

A Proof of Theorem 1

Before the prove of the main theorem, let's consider two lemmas.

Lemma 2. *Suppose a, b, c are three vectors, then we have:*

$$\langle |a-b|, 2b\rangle - \langle |a-c|, 2c\rangle \le \langle b+c+|a-b|+|a-c|, |b-c|\rangle$$

Proof. Consider the following inequation:

$$\langle |a-c|, b-c-|b-c|\rangle \le \langle |a-b|, c-b+|c-b|\rangle \tag{5}$$

According to the nature of the absolute value, it is obvious that $b-c-|b-c| \le 0$ and that $c-b+|c-b| \ge 0$, which shows the constancy of (5) is obvious.
Transform (5), we can get:

$$\langle |a-c|, b\rangle + \langle |a-b|, b-|b-c|\rangle \le \langle |a-c|, c+|b-c|\rangle + \langle c, |a-b|\rangle$$

Based on the absolute value inequality, we can get:

$$|a-b|-|b-c| \le |a-c|; |a-c|+|b-c| \ge |a-b|$$

Hence:

$$\begin{aligned}\langle |a-b|-|b-c|, b\rangle + \langle |a-b|, b-|b-c|\rangle &\le \langle |a-c|, b\rangle + \langle |a-b|, b-|b-c|\rangle \\ &\le \langle |a-c|, c+|b-c|\rangle + \langle c, |a-b|\rangle \\ &\le \langle |a-c|, c+|b-c|\rangle + \langle c, |a-c|+|b-c|\rangle\end{aligned}$$

Combine some items of the same kind:

$$\langle |a-b|-|b-c|, b\rangle + \langle |a-b|, b-|b-c|\rangle \le \langle |a-c|+|b-c|, c\rangle + \langle |a-c|, c+|b-c|\rangle$$

Thus,

$$2\langle |a-b|, b\rangle - 2\langle |a-c|, c\rangle \le \langle b, |b-c|\rangle + \langle c, |b-c|\rangle + \langle |a-b|, |b-c|\rangle + \langle |a-c|, |b-c|\rangle$$

Then,

$$\langle |a-b|, 2b\rangle - \langle |a-c|, 2c\rangle \le \langle b+c+|a-b|+|a-c|, |b-c|\rangle$$

, which completes the proof. □

Remark A.1. *Suppose a, b, c are three vectors, then we have:*

$$\langle |a-b|, b\rangle - \langle |a-c|, c\rangle \le \left\langle \frac{b+c+|a-b|+|a-c|}{2}, |b-c|\right\rangle$$

Lemma 3 (Holder's inequation)**.** *Set $p>1, 1/p+1/q=1$, if $a_1, a_2...a_n$ and $b_1, b_2...b_n$ is nonnegative, then we have:*

$$\sum_{i=1}^{n} a_i b_i \le \left(\sum_{i=1} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1} b_i^q\right)^{\frac{1}{q}}$$

Remark A.2. *a and b are two vectors, and each of their terms is nonnegative. Then, we can get:*

$$\langle a, b\rangle \le \langle \|a\|_1, \|b\|_\infty\rangle$$

, where $\|a\|_1$ represents the L_1-norm of vector a and $\|b\|_\infty$ represents the L_∞-norm of vector b.

Proof. Setting p as ∞ and q as 1, then according to:

$$\sum_{i=1}^{n} a_i b_i \le \left(\sum_{i=1} b_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1} a_i^q\right)^{\frac{1}{q}}$$

We can have:

$$\sum_{i=1}^{n} a_i b_i \le \|b\|_\infty \cdot \|a\|_1$$

Thus,

$$\langle a, b\rangle \le \langle \|a\|_1, \|b\|_\infty\rangle$$

, which completes the proof □

Now, let's start considering the proof of **Theorem 1** as follows:

Proof. For any state-action marginal distribution, we set that $\hat{s} = \rho^{\pi_\theta}(s,a)$. We know that $\rho^{\pi_\theta}(s,a) = \rho^{\pi_\theta}(s) \cdot \pi_\theta(a|s)$, and given policy π, the action is sampled with the policy $\pi_\theta(a|s)$. Hence, we can get:

$$\begin{aligned}
\mathrm{EA2E}(\pi_\theta) =& \mathbb{E}_A \left[\mathbb{E}_{\hat{s}|A}\left[|\mathbb{P}(\mathrm{U}=1|\hat{s}=s, A=a) - s|\right]\right] \\
=& \mathbb{E}_A \left[\mathbb{E}_{\rho^{\pi_\theta}(s,a)}\left[|\mathbb{P}(\mathrm{U}=1|\hat{s}=\rho^{\pi_\theta}(s,a), A=a) - \rho^{\pi_\theta}(s,a)|\right]\right] \\
=& \mathbb{E}_A \left[\sum \rho^{\pi_\theta}(s,a) \cdot |\mathbb{P}(\mathrm{U}=1|\hat{s}=\rho^{\pi_\theta}(s,a), A=a) - \rho^{\pi_\theta}(s,a)|\right]
\end{aligned}$$

Setting ρ^π be the conditional distribution $\mathbb{P}(\mathrm{U}=1|\hat{s}=\rho^{\pi_\theta}(s,a), A=a)$. Thus,

$$\mathrm{EA2E}(\pi_\theta) = \mathbb{E}_A \left[\langle |\rho^\pi - \rho^{\pi_\theta}(s,a)|, \rho^{\pi_\theta}(s,a)\rangle\right]$$

,where $|\rho^\pi - \rho^{\pi_\theta}(s,a)|$ and $\rho^{\pi_\theta}(s,a)$ are vectors and $\langle|\rho^\pi - \rho^{\pi_\theta}(s,a)|, \rho^{\pi_\theta}(s,a)\rangle$ represents the inner product of $|\rho^\pi - \rho^{\pi_\theta}(s,a)|$ and $\rho^{\pi_\theta}(s,a)$. Further, let's compare the EA2E of two models (agents) $\theta_1, \theta_2 \in \Theta$:

$$\mathrm{EA2E}(\pi_{\theta_1}) - \mathrm{EA2E}(\pi_{\theta_2}) = \mathbb{E}_A \left[\left\langle\left|\rho^\pi - \rho^{\pi_{\theta_1}}(s,a)\right|, \rho^{\pi_{\theta_1}}(s,a)\right\rangle - \left\langle\left|\rho^\pi - \rho^{\pi_{\theta_2}}(s,a)\right|, \rho^{\pi_{\theta_2}}(s,a)\right\rangle\right]$$

According to the Lemma 2, we can get:

$$\begin{aligned}
&\mathrm{EA2E}(\pi_{\theta_1}) - \mathrm{EA2E}(\pi_{\theta_2}) = \mathbb{E}_A \left[\left\langle\left|\rho^\pi - \rho^{\pi_{\theta_1}}(s,a)\right|, \rho^{\pi_{\theta_1}}(s,a)\right\rangle - \left\langle\left|\rho^\pi - \rho^{\pi_{\theta_2}}(s,a)\right|, \rho^{\pi_{\theta_2}}(s,a)\right\rangle\right] \\
&\leq \mathbb{E}_A \left[\left\langle \frac{\rho^{\pi_{\theta_1}}(s,a) + \rho^{\pi_{\theta_2}}(s,a) + \left|\rho^\pi - \rho^{\pi_{\theta_1}}(s,a)\right| + \left|\rho^\pi - \rho^{\pi_{\theta_2}}(s,a)\right|}{2}, |\rho^{\pi_{\theta_1}}(s,a) - \rho^{\pi_{\theta_2}}(s,a)| \right\rangle\right]
\end{aligned}$$

According to Lemma 3(Holder's inequality), we have that:

$$\begin{aligned}
\mathrm{EA2E}(\pi_{\theta_1}) - \mathrm{EA2E}(\pi_{\theta_2}) \leq& \mathbb{E}_A \left[\left\langle \frac{\rho^{\pi_{\theta_1}} + \rho^{\pi_{\theta_2}} + \left|\rho^\pi - \rho^{\pi_{\theta_1}}\right| + \left|\rho^\pi - \rho^{\pi_{\theta_2}}\right|}{2}, |\rho^{\pi_{\theta_1}} - \rho^{\pi_{\theta_2}}| \right\rangle\right] \\
\leq& \mathbb{E}_A \left[\left\|\rho^{\pi_{\theta_1}} - \rho^{\pi_{\theta_2}}\right\|_1 \cdot \left\| \frac{\rho^{\pi_{\theta_1}} + \rho^{\pi_{\theta_2}} + \left|\rho^\pi - \rho^{\pi_{\theta_1}}\right| + \left|\rho^\pi - \rho^{\pi_{\theta_2}}\right|}{2} \right\|_\infty\right]
\end{aligned}$$

Setting

$$m(\pi_{\theta_1}, \pi_{\theta_2}, \pi) = \left\| \frac{\rho^{\pi_{\theta_1}} + \rho^{\pi_{\theta_2}} + \left|\rho^\pi - \rho^{\pi_{\theta_1}}\right| + \left|\rho^\pi - \rho^{\pi_{\theta_2}}\right|}{2} \right\|_\infty$$

For the sake that each term of the distributions $\rho^{\pi_{\theta_1}}, \rho^{\pi_{\theta_2}}$ and ρ^π are bounded in $[0,1]$, hence it is evident that $m(\pi_{\theta_1}, \pi_{\theta_2}, \pi) \leq 2$. Therefore, we can get:

$$\mathrm{EA2E}(\pi_{\theta_1}) - \mathrm{EA2E}(\pi_{\theta_2}) \leq \mathbb{E}_A \left[2 \cdot \left\|\rho^{\pi_{\theta_1}}(s,a), \rho^{\pi_{\theta_2}}(s,a)\right\|_1\right] = \mathbb{E}\left[4 \cdot \mathrm{TV}\left(\rho^{\pi_{\theta_1}}(s,a), \rho^{\pi_{\theta_2}}(s,a)\right)\right]$$

, which completes the prove.

□

B Proof of Proposition 1

According to Theorem 1, we have that:

$$\mathrm{EA2E}(\pi_{\theta_1}) - \mathrm{EA2E}(\pi_{\theta_2}) \leq \mathbb{E}\left[4 \cdot \mathrm{TV}(\rho^{\pi_{\theta_1}}(s,a), \rho^{\pi_{\theta_2}}(s,a)\right]$$

In work [27], the following conclusion has been given:

Lemma 4. *Setting X, Y are finitely discrete random variables, and they are bounded; $X \in \mu$ and $Y \in \nu$, $\Omega = supp(\mu) \cup supp(\nu)$; $d_{min} = \inf_{x \neq y \in \Omega} \|x - y\|$, then we have :*

$$\mathrm{TV}(X,Y) \leq \frac{1}{d_{min}} \cdot \mathrm{W}_1(X,Y)$$

Hence, let's consider the proof of **Proposition 1**.

Proof. According to Theorem 1, we have:

$$\mathrm{EA2E}(\pi) - \mathrm{EA2E}(\pi_\theta) \leq \mathbb{E}\left[4 \cdot \mathrm{TV}(\rho^\pi(s,a), \rho^{\pi_\theta}(s,a)\right]$$

According to Lemma 4, setting $\rho^\pi(s,a) \in \mu$, $\rho^{\pi_\theta}(s,a) \in \nu$, $\Omega = supp(\mu) \cup supp(\nu)$, $d_{min} = \inf_{\rho^\pi \neq \rho^{\pi_\theta} \in \Omega} \|\rho^\pi - \rho^{\pi_\theta}\|$, we can derive that:

$$\begin{aligned} \mathrm{EA2E}(\pi) - \mathrm{EA2E}(\pi_\theta) &\leq \mathbb{E}\left[4 \cdot \mathrm{TV}(\rho^\pi(s,a), \rho^{\pi_\theta}(s,a)\right] \\ &\leq \mathbb{E}\left[\frac{4}{d_{min}} \cdot \mathrm{W}_1\left(\rho^\pi(s,a), \rho^{\pi_\theta}(s,a)\right)\right] \end{aligned}$$

,which completes the proof. $\square$

C Visualizations

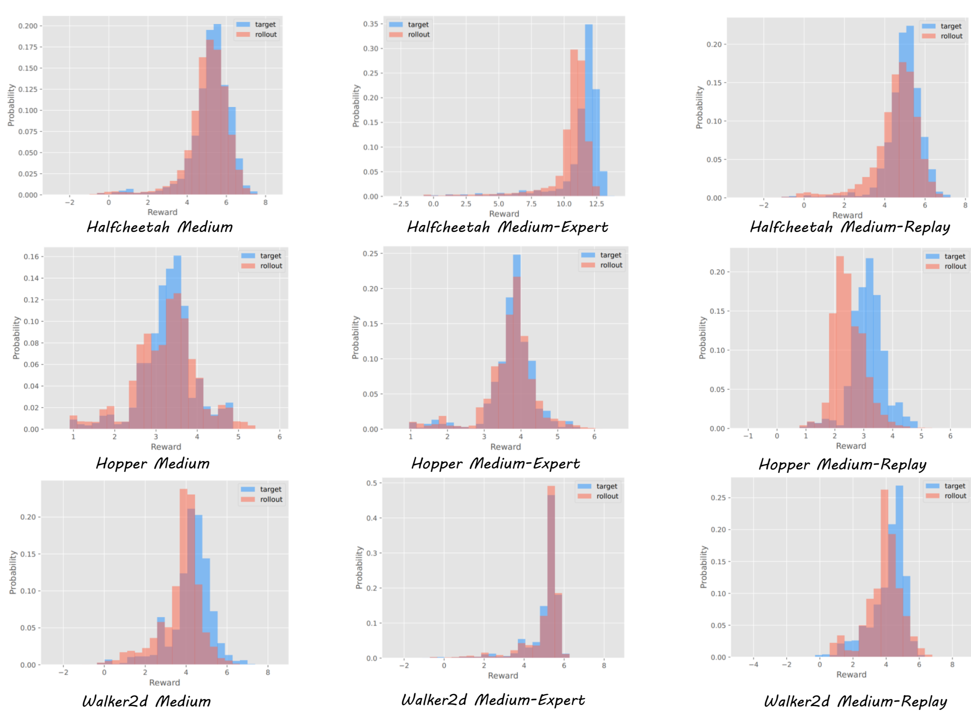

Figure 2: Visualizations of the distributional matching results on offline datasets.

D Experimental Hyperparameters

Table 3: Hyperparameters of CDMamba on the D4RL datasets.

Hyperparameter	Value
Number of Layers	3
Batch Size	64
Context Length K	20
Embedding Dimension	128
Distribution Dimension	30
Number of bins for categorical distribution	31
Learning Rate	1×10^{-4}
Learning Rate Decay	Linear warmup for first 100k training steps
Grad Norm Clip	0.25
Weight Decay	1×10^{-4}