Rebalancing Return Coverage for Conditional Sequence Modeling in Offline Reinforcement Learning

Wensong Bai, Chufan Chen, Yichao Fu, Qihang Xu, Chao Zhang, Hui Qian

Zhejiang University

Background

- **Challenge:** CSM-based offline RL often suffers from distributional shift during high-return inference.
- **Cause:** Imbalanced training data return distribution.
- **Effect:** Models are poorly trained on rare high-return trajectories, leading to suboptimal exploitation at test time.

Contribution

- A theoretical characterization is established, showing that the performance of CSM-based policies is governed jointly by the coverage of expert-level returns and full-spectrum runtime returns in the offline dataset.
- A return-coverage rebalancing mechanism is introduced as a simple plug-in module that can be integrated into existing CSM-based methods to enhance robustness and performance.
- A new algorithm, RVDT, is developed on top of Decision Transformer, combining Q-value guidance with expert-policy KL regularization to more closely align sampled actions with high-return behaviors.

Problem Formulation

- **Setting.** Offline RL operates on a static dataset $\mathcal{D} = \{\tau = (s_t, a_t, r_t)_{t=1}^{H}\}$ collected by a behavior policy π_β, with no further interaction.
- **Objective.** Learn a policy π that maximizes $\mathcal{J}(\pi) = \mathbb{E}_{\tau\sim\pi}\big[g(\tau)\big], g(\tau_t) = \sum_{i=t}^{H} r_i$.
- **CSM paradigm.** Policies are trained via the return-conditioned NLL objective

$$\mathcal{L}(\pi) = -\sum_{\tau\in\mathcal{D}}\sum_{t=1}^{H}\log\pi\big(a_t \,\big|\, s_t,\ g(\tau_t),\ \bar{\tau}_{t-1}^{K}\big),$$

and deployed as $\pi_f(a\,|\,s) = \pi\big(a \,\big|\, s,\ f(s),\ \bar{\tau}^{K}\big)$.

Methodology

Return-rebalanced Decision Transformer (RDT): $\mathcal{L}_{\text{RDT}}(\theta) =$

$$\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\sum_{i=1}^{H} -\log \pi_\theta(a_i|s_i, g(\tau_i), \bar{\tau}_{t-1}^{K})\Big] + \alpha\mathbb{E}_{\tau\sim\mathcal{D}_e}\sum_{i=1}^{H}\text{KL}\big[\pi_\theta(\cdot|s_i, g(\tau_i), \bar{\tau}_{t-1}^{K})\|\pi^e(\cdot|s_i)\big] \tag{1}$$

Proposition (KL regularization as weighted sampling strategy)

Assume the policy π_θ is parameterized by a factorized Gaussian distribution with a fixed standard deviation. Optimizing (1) *is equivalent to optimizing the following weighted NLL loss:*

$$\begin{aligned}
&\arg\min_{\pi_\theta} \mathcal{L}_{RDT}(\theta) = \\
&\arg\min_{\pi_\theta} \mathbb{E}_{\tau\sim\mathcal{D}}\Big[(1+\alpha\cdot\mathbb{I}\left[\tau\in\mathcal{D}_e\right])\cdot\big(\textstyle\sum_{i=1}^{H} -\log \pi_\theta(a_i|s_i, g(\tau_i), \bar{\tau}_{t-1}^{K})\big)\Big].
\end{aligned} \tag{2}$$

Return-rebalanced Value-regularized Decision Transformer (RVDT):

$$\mathcal{L}_{\mathrm{RVDT}}(\theta) =$$
$$\mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{i=1}^{H} - \log \pi_\theta(a_i|s_i, g(\tau_i), \bar{\tau}_{t-1}^{K})\right] - \eta\mathbb{E}_{\tau\sim\mathcal{D}}\mathbb{E}_{s_i\sim\tau, a_i\sim\pi_\theta}[Q^{\pi_\theta}(s_i, a_i)]$$
$$+ \alpha\mathbb{E}_{\tau\sim\mathcal{D}_e}\sum_{i=1}^{H}\mathrm{KL}\left[\pi_\theta(\cdot|s_i, g(\tau_i), \bar{\tau}_{t-1}^{K})\|\pi^e(\cdot|s_i)\right]. \quad (3)$$

Analysis I

- **Notation:**
  - Non-optimal runtime conditioning functions: f
  - Optimal (or near-optimal) conditioning functions: f^*
- **Runtime conditional return function:**
  - f corresponds to arbitrary possible returns that may be encountered during policy execution.
  - Let $\mathcal{G}$ denote the collection of all possible returns collected by $\pi \in \Pi$, then:

$$f : \mathcal{S} \to \mathcal{G}.$$

  - The target conditional function f^* we aim to find is the RTG under the optimal policy π^*:

$$f(s) = \max_{\pi} \mathbb{E}_{\pi}[g(s)].$$

Analysis II

- **Return-coverage definitions:**
  - **Expert-level return-coverage:**

    $$P_{\pi_\beta}(g = f^*(s_1) \mid s_1),$$

    where f^* corresponds to the optimal policy π^* of the underlying MDP.
  - **Full-spectrum return-coverage:**

    $$P_{\pi_\beta}(g = f(s_1) \mid s_1).$$

Analysis

Theorem (Performance gap with respect to return-coverage)

Consider a finite-horizon MDP with horizon H, behavior policy π_β, a runtime conditioning function f, and the optimal conditioning function for π^* is f^*. Assume the following assumptions hold:
(i) ***Return-coverage:*** $P_{\pi_\beta}(g = f(s_1)|s_1) \geq \alpha_f$ *and* $P_{\pi_\beta}(g = f^*(s_1)|s_1) \geq \alpha_f^*$ *for all initial states* s_1*.*
(ii) ***Near determinism:*** $P(r \neq \mathcal{R}(s,a)$ *or* $s' \neq \mathcal{T}(s,a)|s,a) \leq \epsilon$ *at all* (s,a) *for some* $\mathcal{T}$ *and* $\mathcal{R}$*.*
(iii) ***Consistency of f:*** $f(s) = f(s') + r$ *for all* s*.*
Then the following upper bound holds:

$$J(\pi^*) - J(\pi_f^{CSM}) \leq (\frac{1}{\alpha_f^*} + 3)H^2\epsilon + (\frac{1}{\alpha_f} + \frac{1}{\alpha_f^*})H^2 C, \qquad (4)$$

where $C \in (0,1)$ *is a constant.*

Analysis

Theorem (Sample complexity)

To get finite data guarantees, add to the above assumptions:
(i) ***Bounded occupancy mismatch****:* $P_{\pi_f^{CSM}}(s) \leq C_f \cdot P_{\pi_\beta}(s)$ *for all* s*;*
(ii) ***Finite policy class*** Π*;*
(iii) ***Bounded log-likelihood variation****:*
$|\log \pi(a|s,g) - \log \pi(a'|s',g')| \leq c$ *for any* (a,s,g,a',s',g') *and all* $\pi \in \Pi$*;*
(iv) ***Bounded approximation error of*** Π, *i.e.,* $\min_{\pi\in\Pi} L(\pi) \leq \epsilon_{approx.}$ *.*
Define the expected loss as
$L(\hat{\pi}) = \mathbb{E}_{s\sim P_{\pi_\beta}} \mathbb{E}_{g\sim P_{\pi_\beta}(\cdot|s)} \left[\mathrm{KL}\left(P_{\pi_\beta}(\cdot|s,g) \,\|\, \hat{\pi}(\cdot|s,g)\right)\right].$
Then for any estimated CSM policy $\hat{\pi}_f$ *that conditions on* f *at inference time, with probability at least* $1-\delta$*,*

$$J(\pi^*) - J(\hat{\pi}_f) \leq O\left(\left[\frac{C_f}{\alpha_f}\sqrt{c}\left(\frac{\log|\Pi|/\delta}{N}\right)^{1/4} + \frac{C_f}{\alpha_f}\sqrt{\epsilon_{approx}} + \frac{\epsilon + C}{\alpha_f^*} + \frac{C}{\alpha_f}\right] H^2\right). \tag{5}$$

Experiments

Results on D4RL Benchmark:

Gym Tasks	CQL	IQL	BCQ	TD3+BC	MoRel	BC	DD	DT	StAR	GDT	CGDT	QT	**RVDT**
halfcheetah-m-e	91.6	86.7	69.6	90.7	53.3	55.2	90.6	86.8	93.7	93.2	93.6	93.2	**94.4** ± 0.1
hopper-m-e	105.4	91.5	109.1	98.0	108.7	52.5	111.8	107.6	111.1	111.1	107.6	113.0	**113.1** ± 0.5
walker2d-m-e	108.8	109.6	67.3	110.1	95.6	107.5	108.8	108.1	109.0	107.7	109.3	112.0	**112.7** ± 1.6
halfcheetah-m	49.2	47.4	41.5	48.4	42.1	42.6	49.1	42.6	42.9	42.9	43.0	51.0	**51.9** ± 0.3
hopper-m	69.4	66.3	65.1	59.3	95.4	52.9	79.3	67.6	59.5	77.1	96.9	99.6	**100.2** ± 0.1
walker2d-m	83.0	78.3	52.0	83.7	77.8	75.3	82.5	74.0	73.8	76.5	79.1	87.2	**90.2** ± 0.1
halfcheetah-m-r	45.5	44.2	34.8	44.6	40.2	36.6	39.3	36.6	36.8	40.5	40.4	48.8	**53.8** ± 2.0
hopper-m-r	95.0	94.7	31.1	60.9	93.6	18.1	100.0	82.7	29.2	85.3	93.4	102.1	**103.2** ± 1.9
walker2d-m-r	77.2	73.9	13.7	81.8	49.8	32.3	75.0	79.4	39.8	77.5	78.1	97.8	**99.3** ± 0.8
Average	80.6	77.0	53.8	75.3	72.9	52.6	81.8	76.2	66.2	79.1	82.4	89.4	**91.2**

Adroit Tasks	CQL	IQL	BCQ	BEAR	O-RL	BC	DD	D-QL	DT	StAR	GDT	QT	**RVDT**
pen-human	37.5	71.5	66.9	-1.0	90.7	63.9	66.7	72.8	79.5	77.9	92.5	111.9	**127.2** ± 5.5
hammer-human	4.4	1.4	0.9	0.3	0.2	1.2	1.9	0.2	3.7	3.7	5.5	10.4	**24.0** ± 1.5
pen-cloned	39.2	37.3	50.9	26.5	60.0	37.0	42.8	57.3	75.8	33.1	86.2	85.8	**117.8** ± 8.6
hammer-cloned	2.1	2.1	0.4	0.3	2.0	0.6	1.7	3.1	3.0	0.3	8.9	11.8	**21.3** ± 2.7
Average	20.8	28.1	29.8	6.5	38.2	25.7	28.3	33.4	40.5	28.8	48.3	55.0	**72.6**

Kitchen Tasks	CQL	IQL	BCQ	BEAR	O-RL	BC	DD	D-QL	DT	StAR	GDT	QT	**RVDT**
kitchen-Comp.	43.8	62.5	8.1	0.0	2.0	65.0	65.0	84.0	50.8	40.8	43.8	81.7	**84.5** ± 2.3
kitchen-partial	49.8	46.3	18.9	13.1	35.5	33.8	57.0	60.5	57.9	12.3	73.3	72.5	**75.0** ± 2.5
Average	46.8	54.4	13.5	6.6	18.8	49.4	61.0	72.2	54.4	26.6	58.6	77.1	**79.8**

Maze2D Tasks	CQL	IQL	BCQ	BEAR	TD3+BC	BC	Diffuser	DD	DT	GDT	QDT	QT	**RVDT**
maze2d-u	94.7	42.1	49.1	65.7	14.8	88.9	113.9	116.2	31.0	50.4	57.3	99.2	**145.1** ± 3.8
maze2d-m	41.8	34.9	17.1	25.0	62.1	38.3	121.5	122.3	8.2	7.8	13.3	168.8	**183.5** ± 4.5
maze2d-l	49.6	61.7	30.8	81.0	88.6	1.5	123.0	125.9	2.3	0.7	31.0	242.7	**254.3** ± 4.6
Average	62.0	46.2	32.3	57.2	55.2	42.9	119.5	121.5	13.8	19.6	33.9	170.2	**194.3**

AntMaze Tasks	CQL	IQL	BCQ	BEAR	TD3+BC	BC	DD	D-QL	DT	StAR	GDT	QT	**RVDT**
antmaze-u	74.0	87.5	78.9	73.0	78.6	54.6	73.1	93.4	59.2	51.3	76.0	96.0	**98.0** ± 4.0
antmaze-u-d	84.0	62.2	55.0	61.0	71.4	45.6	49.2	66.2	53.0	45.6	69.0	92.0	**98.0** ± 4.0
antmaze-m-d	53.7	70.0	0.0	8.0	3.0	0.0	24.6	**78.6**	0.0	0.0	6.0	24.0	30.0 ± 6.3
antmaze-l-d	14.9	47.5	2.2	0.0	0.0	0.0	7.5	**56.6**	0.0	0.0	0.0	10.0	10.0 ± 0.0
Average	56.6	66.8	34.0	35.5	38.2	25.0	38.6	**73.7**	28.0	24.2	37.8	57	59.0

Experiments

Results on Maze2D Environments:

	Dataset	CQL	DT	QDT	QT	RVDT
Sparse R	maze2d-open-v0	216.7 ± 80.7	196.4 ± 39.6	190.1 ± 37.8	497.9 ± 12.3	**634.6** ± 12.3
	maze2d-umaze-v1	94.7 ± 23.1	31.0 ± 21.3	57.3 ± 8.2	105.4 ± 4.8	**145.1** ± 3.8
	maze2d-medium-v1	41.8 ± 13.6	8.2 ± 4.4	13.3 ± 5.6	172.0 ± 6.2	**183.5** ± 4.5
	maze2d-large-v1	49.6 ± 8.4	2.3 ± 0.9	31.0 ± 19.8	240.1 ± 2.5	**254.3** ± 4.6
Dense R	maze2d-open-v0	307.6 ± 43.5	346.2 ± 14.3	325.7 ± 61.4	608.4 ± 1.9	**663.9** ± 15.9
	maze2d-umaze-v1	72.7 ± 10.1	−6.8 ± 10.9	58.6 ± 3.3	**103.1 ± 7.8**	99.5 ± 4.3
	maze2d-medium-v1	70.9 ± 9.2	31.5 ± 3.7	42.3 ± 7.1	111.9 ± 1.9	**126.9** ± 8.7
	maze2d-large-v1	90.9 ± 19.4	45.3 ± 11.2	62.2 ± 9.9	177.2 ± 7.8	**197.9** ± 2.0

Performance Comparison in Low-data Regimes:

Task (sparse R)	$\mathcal{D}_1$		$\mathcal{D}_2$		$\mathcal{D}_3$		$\mathcal{D}_4$	
	QT	RVDT	QT	RVDT	QT	RVDT	QT	RVDT
maze2d-umaze	100.3	**171.7**	81.8	**101.8**	73.1	**76.9**	61.4	**100.5**
maze2d-medium	137.1	**187.4**	175.2	**190.0**	163.2	**182.0**	98.3	**175.3**
maze2d-large	109.5	**140.4**	81.5	**90.1**	104.4	**131.3**	**100.2**	95.1
Average	115.6	**166.5**	112.8	**127.3**	113.6	**130.1**	86.6	**123.6**

Ablation Results:

Task	DT	DT-Dup	RDT	QT	VDT	RVDT-Dup	RVDT-Determ	**RVDT**
halfcheetah	84.2	90.3	90.5	91.2	89.5	93.4	91.5	**94.9**
hopper	109.5	112.1	111.9	112.3	112.6	112.1	113.6	**113.8**
walker2d	108.2	108.8	109.7	113.2	110.3	110.9	113.1	**118.7**
Average	100.6	103.7	104.0	105.6	104.1	105.5	106.1	**109.1**

Component	DT	DT-Dup	RDT	QT	VDT	RVDT-Dup	RVDT-Determ	**RVDT**
Explicit Rebal.	None	Dup.	KL	None	None	Dup.	KL	KL
Implicit Rebal.	None	None	None	Q-value	Q-value	Q-value	Q-value	Q-value
Policy Type	Determ.	Stoch.	Stoch.	Determ.	Stoch.	Stoch.	Determ.	Stoch.

Thank you!

Questions?

Contact: Wensong Bai | Zhejiang University