Distributional Reinforcement Learning with Regularized Wasserstein Loss

Ke Sun, Yingnan Zhao, Wulong Liu, Bei Jiang, Linglong Kong

University of Alberta
Alberta Machine Intelligence Institute (Amii)

NeurIPS 2024

Ke Sun

Outline

Return: Cumulative Rewards

$$Z^{\pi} = \sum_{t=0}^{\infty} \gamma^{t} R_{t}$$

A Fundamental Problem: Value Function?

- Classical RL learns **value function**, the expectation of returns:

$$
\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}\left[Z^\pi(s,a)\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t) | s_0 = s, a_0 = a\right]
\end{aligned}
$$

A Fundamental Problem: Value Function?

- Classical RL learns **value function**, the expectation of returns:

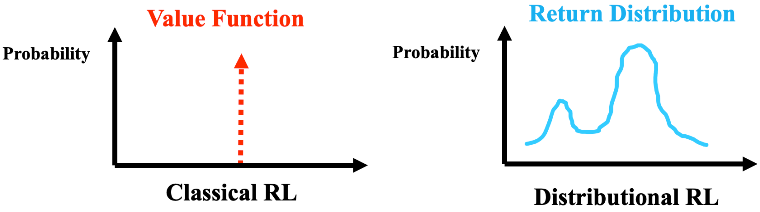
$$Q^\pi(s,a) = \mathbb{E}\left[Z^\pi(s,a)\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s, a_0 = a\right]$$

- Distributional RL learns the whole distribution of returns:

$$\mathcal{D}(Z^\pi(s,a))$$

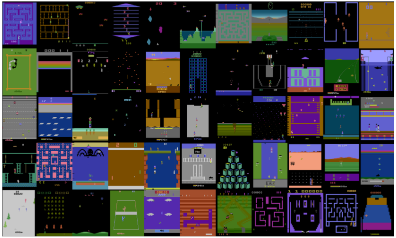
where $\mathcal{D}$ extracts the distribution of a random variable.

Distributional Learning: Beyond Expectation



Performance Improvement of Distributional RL

Classical RL vs Distributional RL



Atari Games

Distributional RL: A Well-Defined RL Area

- **Classical RL:** Classical Bellman operator $\mathcal{T}^{\pi}$ is defined as

$$\mathcal{T}^{\pi} Q(s,a) = \mathbb{E}[R(s,a)] + \gamma \mathbb{E}_{s' \sim p, \pi}\left[Q\left(s', a'\right)\right], \tag{1}$$

where $\mathcal{T}^{\pi}$ is a γ-contractive operator.

Distributional RL: A Well-Defined RL Area

- **Classical RL:** Classical Bellman operator $\mathcal{T}^{\pi}$ is defined as

$$\mathcal{T}^{\pi} Q(s,a) = \mathbb{E}[R(s,a)] + \gamma \mathbb{E}_{s' \sim p, \pi}\left[Q\left(s', a'\right)\right], \quad (1)$$

where $\mathcal{T}^{\pi}$ is a γ-contractive operator.

- **Distributional RL:** Distributional Bellman operator $\mathfrak{T}^{\pi}$ is defined as

$$\mathfrak{T}^{\pi} Z(s,a) \stackrel{D}{:=} R(s,a) + \gamma Z\left(s', a'\right), \quad (2)$$

where $\mathfrak{T}^{\pi}$ is a contractive operator under some proper distribution divergence / statistical distances, e.g., Wasserstein distance.

Distributional RL: A Well-Defined RL Area

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Classical RL:** Classical Bellman operator $\mathcal{T}^{\pi}$ is defined as

$$\mathcal{T}^{\pi} Q(s,a) = \mathbb{E}[R(s,a)] + \gamma \mathbb{E}_{s' \sim p,\pi}\left[Q\left(s',a'\right)\right], \quad (1)$$

  where $\mathcal{T}^{\pi}$ is a γ-contractive operator.

- **Distributional RL:** Distributional Bellman operator $\mathfrak{T}^{\pi}$ is defined as

$$\mathfrak{T}^{\pi} Z(s,a) :\stackrel{D}{=} R(s,a) + \gamma Z\left(s',a'\right), \quad (2)$$

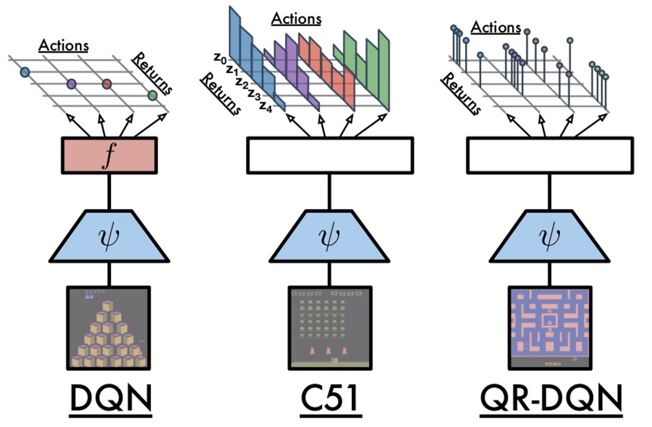
  where $\mathfrak{T}^{\pi}$ is a contractive operator under some proper distribution divergence / statistical distances, e.g., Wasserstein distance.

- **Two key factors** in Distributional RL:
  ① How to parameterize Z^{π}?
  ② How to choose the statistical distance?

Outline

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Existing Algorithms



Two Limitations of Existing Algorithms

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

① Inaccuracy in Capturing Return Distribution Characteristics

- Non-crossing issue of learned quantile curves
- Restricted expressiveness of pre-specified statistics

② Difficulties in Extension to Multi-dimensional Rewards

- Many RL tasks learn a multi-dimensional return distribution
  - multi-source rewards
  - hybrid reward architecture
  - sub-reward architecture
- Difficult to extend existing algorithms to multi-dimensional setting
  - multi-dimensional categorical representation?
  - multi-dimensional quantile regression?

Outline

Our Contribution

UNIVERSITY OF ALBERTA
EDMONTON · ALBERTA · CANADA

- Algorithm. We introduce a new distributional RL algorithm based on Sinkhorn divergence, a regularized Wasserstein loss.
- Theory. We prove the contraction properties of Bellman operators under Sinkhorn divergence, revealing an interpolation relationship between Wasserstein distance and MMD.
- Experiments. We conduct extensive experiments over 55 Atari games, investigating
  - superiority in multi-dimensional reward setting
  - Comprehensive comparison with existing algorithms
  - Sensitivity analysis and computational cost

Outline

Popular Statistical Distances

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Optimal Transport**

$$W_c = \inf_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \int c(x,y)\mathrm{d}\Pi(x,y), \tag{3}$$

where the minimizer Π^* is called the *optimal transport plan* or *optimal coupling*.

- ***p*-Wasserstein Distance**

$$W_p = \left(\inf_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \int \|x-y\|^p \mathrm{d}\Pi(x,y) \right)^{1/p}. \tag{4}$$

- **Maximum Mean Discrepancy (MMD)**

$$\mathrm{MMD}_k^2 = \mathbb{E}\left[k\left(X,X'\right)\right] + \mathbb{E}\left[k\left(Y,Y'\right)\right] - 2\mathbb{E}\left[k(X,Y)\right], \tag{5}$$

where $k(\cdot,\cdot)$ is a continuous kernel and X' (resp. Y') is a random variable independent of X (resp. Y).

Sinkhorn Divergence

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Sinkhorn divergence is an entropic regularized Wasserstein distance. We first define $\mathcal{W}_{c,\varepsilon}(\mu,\nu)$ as

$$\mathcal{W}_{c,\varepsilon}(\mu,\nu) = \min_{\Pi\in\mathbf{\Pi}(\mu,\nu)} \int c(x,y)\mathrm{d}\Pi(x,y) + \varepsilon\mathrm{KL}(\Pi|\mu\otimes\nu), \tag{6}$$

  where the regularization $\mathrm{KL}(\Pi|\mu\otimes\nu) = \int \log\left(\frac{\Pi(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}\right)\mathrm{d}\Pi(x,y)$, is also known as **mutual information**.
- Sinkhorn divergence $\overline{\mathcal{W}}_{c,\varepsilon}$ is defined as

$$\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) = 2\mathcal{W}_{c,\varepsilon}(\mu,\nu) - \mathcal{W}_{c,\varepsilon}(\mu,\mu) - \mathcal{W}_{c,\varepsilon}(\nu,\nu). \tag{7}$$

Benefits and Regularization Effect

① **Addressing Limitation 1:** Efficient approximation of a multi-dimensional Wasserstein distance

② **Addressing Limitation 2:** Leveraging samples, un-restricted statistics, to represent return distributions

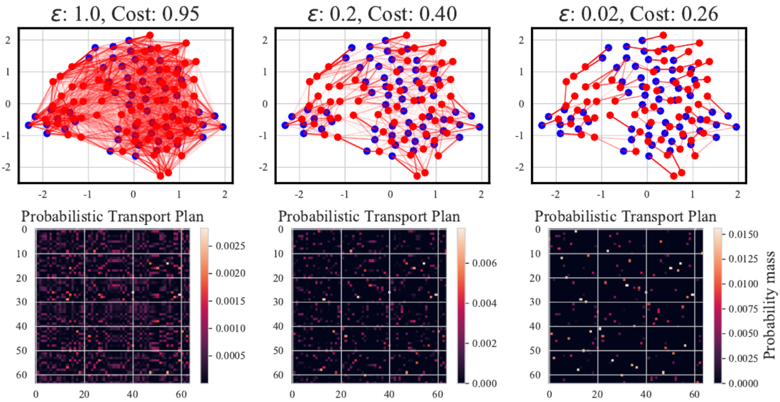
③ **Regularization Effects**

- "Smoother" transport plan
- Maximum entropy principle
- Stable optimization: strongly convexity and smoothness

Smoother Transport Plan

- **Recap.** Regularized Wasserstein distance:

$$\mathcal{W}_{c,\varepsilon}(\mu,\nu) = \min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \int c(x,y)\mathrm{d}\Pi(x,y) + \varepsilon \mathrm{KL}(\Pi | \mu \otimes \nu) \quad (8)$$



Outline

Basic Contraction Properties

The contraction analysis of $\mathfrak{T}^\pi$ depends on two properties of the statistical distance d_p.

Contraction Properties of statistical distance d_p

1. **Scale Sensitive (S)**:

$$d_p(aX, aY) \leq |a|^\tau d_p(X, Y), \tag{9}$$

where $\tau > 0$.

2. **Sum Invariant (I)**:

$$d_p(A + X, A + Y) \leq d_p(X, Y), \tag{10}$$

where the random variable A is independent of X and Y.

Contraction Property of Regularization

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Recap.** Regularized Wasserstein distance:

$$\mathcal{W}_{c,\varepsilon}(\mu,\nu) = \min_{\Pi \in \mathbf{\Pi}(\mu,\nu)} \int c(x,y)\mathrm{d}\Pi(x,y) + \varepsilon \mathrm{KL}(\Pi|\mu \otimes \nu) \tag{11}$$

- Given a joint distribution Π, we define the supremal form of the regularization term:

$$\mathrm{MI}_{\Pi}^{\infty}(\mu,\nu) = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathrm{KL}(\Pi|\mu(s,a)\otimes\nu(s,a)) \tag{12}$$

Proposition 1. Contraction under $\mathrm{MI}_{\Pi}^{\infty}(\mu,\nu)$.

The distributional Bellman operator $\mathfrak{T}^{\pi}$ is non-expansive under $\mathrm{MI}_{\Pi}^{\infty}$ for any non-trivial joint distribution Π.

Contraction Property of $\mathcal{W}_{c,\varepsilon}$

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Two Basic Contraction Properties of $\mathcal{W}_{c,\varepsilon}$

Considering $\mathcal{W}_{c,\varepsilon}$ with the unrectified kernel $k_\alpha := -\|x-y\|^\alpha$ as $-c$ $(\alpha > 0)$ and a scaling factor $a \in (0,1)$, we have:

- **(I)** $\mathcal{W}_{c,\varepsilon}$ is sum-invariant
- **(S)** $\mathcal{W}_{c,\varepsilon}(a\mu, a\nu) \leq \Delta_\varepsilon(a,\alpha)\mathcal{W}_{c,\varepsilon}(\mu,\nu)$, with a scaling constant $\Delta_\varepsilon(a,\alpha) \in (|a|^\alpha, 1)$ for any μ and ν in a finite set of probability measures.

Remark. The scaling factor $\Delta_\varepsilon(a,\alpha)$ has no explicit form, but it is determined by the scale factor a, the order α, the hyperparameter ε, and the set of interested probability distributions.

Contraction Property of $\overline{\mathcal{W}}_{c,\varepsilon}$

- We consider the supremal form of statistical distance.

$$\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}(\mu,\nu) = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \overline{\mathcal{W}}_{c,\varepsilon}(\mu(s,a),\nu(s,a)). \qquad (13)$$

Thm 1. Contraction under $\overline{\mathcal{W}}_{c,\varepsilon}$ and Interpolation Relationship.

Considering $\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu)$ with an unrectified kernel $k_\alpha := -\|x-y\|^\alpha$ as $-c$ ($\alpha > 0$), where $\mu,\nu \in$ the distribution set of $\{Z^\pi(s,a)\}$ for $s\in\mathcal{S}, a\in\mathcal{A}$ in a finite MDP. Then, we have:

① ($\varepsilon \to 0$) $\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) \to 2W_\alpha^\alpha(\mu,\nu)$. When $\varepsilon = 0$, $\mathfrak{T}^\pi$ is γ^α-contractive under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$.

② ($\varepsilon \to +\infty$) $\overline{\mathcal{W}}_{c,\varepsilon}(\mu,\nu) \to \text{MMD}_{k_\alpha}^2(\mu,\nu)$. When $\varepsilon = +\infty$, $\mathfrak{T}^\pi$ is γ^α-contractive under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$.

③ ($\varepsilon \in (0,+\infty)$) $\mathfrak{T}^\pi$ is at least $\overline{\Delta}_\varepsilon(\gamma,\alpha)$-**contractive** under $\overline{\mathcal{W}}_{c,\varepsilon}^{\infty}$, where $\overline{\Delta}_\varepsilon(\gamma,\alpha) \in (\gamma^\alpha, 1)$ is an MDP-dependent constant.

A Brief Summary

- **Interpolation Property.** Sinkhorn divergence interpolates between Wasserstein distance and MMD by varying ϵ.
  $\Rightarrow$ Contraction of $\mathfrak{T}^\pi$ in distributional RL !

A Brief Summary

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Interpolation Property.** Sinkhorn divergence interpolates between Wasserstein distance and MMD by varying ϵ.
  $\Rightarrow$ Contraction of $\mathfrak{T}^\pi$ in distributional RL !
- **Consistency with Existing Contraction Conclusions.**
  - QR-DQN with contraction guarantee under Wasserstein distance
  - MMD-DQN with contraction guarantee under MMD if
    1. Unrectified kernel (energy distance or Cramer distance)
    2. Gaussian kernel: no contraction guarantee...

A Brief Summary

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Algorithm	d_p Distribution Divergence	Representation Z_θ	Convergence Rate of $\mathfrak{T}^\pi$	Sample Complexity of d_p
C51	Cramér distance	Categorical Distribution	$\sqrt{\gamma}$	\
QR-DQN-1	Wasserstein distance	Quantiles	γ	$\mathcal{O}(n^{-\frac{1}{d}})$
MMD-DQN	MMD	Samples	$\gamma^{\alpha/2}$ (k_α)	$\mathcal{O}(n^{-1})$
SinkhornDRL (ours)	Sinkhorn divergence ($c = -k_\alpha$)	Samples	γ $(\varepsilon \to 0)$ $\gamma^{\alpha/2}$ $(\varepsilon \to \infty)$	$\mathcal{O}(n^{\frac{e^{\frac{\kappa}{\varepsilon}}}{\varepsilon^{\lfloor d/2 \rfloor}\sqrt{n}}})$ $(\varepsilon \to 0)$ $\mathcal{O}(n^{-\frac{1}{2}})$ $(\varepsilon \to \infty)$

Table 1: Properties of different distribution divergences in typical distributional RL algorithms. d is the sample dimension and $\kappa = 2\beta d + \|c\|_\infty$, where the cost function c is β-Lipschitz [24]. Sample complexity is improved to $\mathcal{O}(1/n)$ using the kernel herding technique [10] in MMD.

Outline

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Extension to Multi-dimensional Return

- We define a d-dimensional reward function $\mathbf{R} : \mathcal{S} \times \mathcal{A} \to P(\mathbb{R}^d)$.
- We have a d-dimensional return vector $\mathbf{Z}^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t \mathbf{R}(s_t, a_t)$, with $\mathbf{Z}^\pi(s, a) = (Z_1^\pi(s, a), \cdots, Z_d^\pi(s, a))^\top$.
- The joint distributional Bellman operator $\mathfrak{T}_d^\pi$ is defined as

$$\mathfrak{T}_d^\pi \mathbf{Z}(s, a) :\overset{D}{=} \mathbf{R}(s, a) + \gamma \mathbf{Z}\left(s', a'\right)$$

Corollary 1.

For two joint distributions $\mathbf{Z}_1$ and $\mathbf{Z}_2$, $\mathfrak{T}_d^\pi$ is $\overline{\Delta}_\varepsilon(\gamma, \alpha)$-contractive under $\overline{\mathcal{W}}_{c,\varepsilon}^\infty$, i.e.,

$$\overline{\mathcal{W}}_{c,\varepsilon}^\infty(\mathfrak{T}^\pi \mathbf{Z}_1, \mathfrak{T}^\pi \mathbf{Z}_2) \leq \overline{\Delta}_\varepsilon(\gamma, \alpha) \overline{\mathcal{W}}_{c,\varepsilon}^\infty(\mathbf{Z}_1, \mathbf{Z}_2). \tag{14}$$

Outline

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Generic Algorithm Update

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Two key factors in distributional RL:

- Samples to represent the return distribution
- Sinkhorn divergence as the statistical distance

Algorithm 1 Generic Sinkhorn distributional RL Update

Require: Number of generated samples N, the cost function c, hyperparameter ε and the target network Z_{θ^*}.
Input: Sample transition (s, a, r', s')

1: **Policy evaluation**: $a^* \sim \pi(\cdot|s')$ or **Control**: $a^* \leftarrow \arg\max_{a' \in \mathcal{A}} \frac{1}{N} \sum_{i=1}^{N} Z_\theta(s', a')_i$
2: $\mathfrak{T}Z_i \leftarrow r + \gamma Z_{\theta^*}(s', a^*)_i, \forall 1 \leq i \leq N$

Output: $\overline{\mathcal{W}}_{c,\varepsilon}\left(\{Z_\theta(s,a)_i\}_{i=1}^{N}, \{\mathfrak{T}Z_j\}_{j=1}^{N}\right)$

$$\overline{\mathcal{W}}_{c,\varepsilon}(Z_\theta(s,a), \mathfrak{T}^\pi Z_\theta(s,a))$$

Sinkhorn Iteration: Approximation

Sinkhorn Iteration with L steps for approximation

- Differentiable and Efficient, e.g., matrix-vector multiplication
- Approximation guarantee with a linear rate
- Easy to implement: adding extra differential layers in existing network architecture

Algorithm 2 Sinkhorn Iterations to Approximate $\overline{\mathcal{W}}_{c,\varepsilon}\left(\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N\right)$

Input: Two samples sequences $\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N$, number of iterations L and hyperparameter ε.

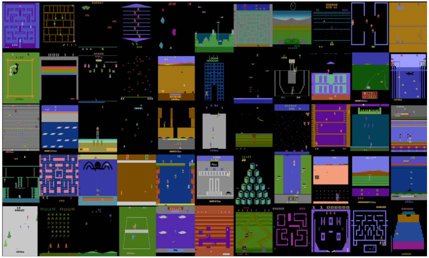
$\hat{c}_{i,j} = c(Z_i, \mathfrak{T}Z_j)$ for $\forall i = 1, ..., N, j = 1, ..., N$
$\mathcal{K}_{i,j} = \exp(-\hat{c}_{i,j}/\varepsilon)$
$b_0 \leftarrow \mathbf{1}_N$
for $l = 1, 2, ..., L$ **do**
    $a_l \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}b_{l-1}}, b_l \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}a_l}$
end for
$\widehat{\overline{\mathcal{W}}}_{c,\varepsilon}\left(\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N\right) = \langle (K \odot \hat{c})b, a\rangle$

Return: $\widehat{\overline{\mathcal{W}}}_{c,\varepsilon}\left(\{Z_i\}_{i=1}^N, \{\mathfrak{T}Z_j\}_{j=1}^N\right)$

Outline

Experiment Setting

- Environments:
  55 Atari Games
- Algorithms:
  - DQN
  - C51
  - QR-DQN
  - MMD-DQN
  - SinkhornDRL (ours)
- The unrectified kernel $k_\alpha := -\|x - y\|^\alpha$ in SinkhornDRL (consistent with Theorem 1)



Atari Games

Comparison with Existing Algorithms

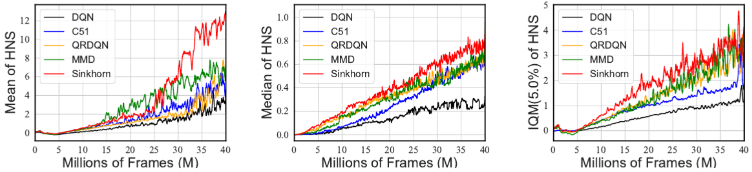
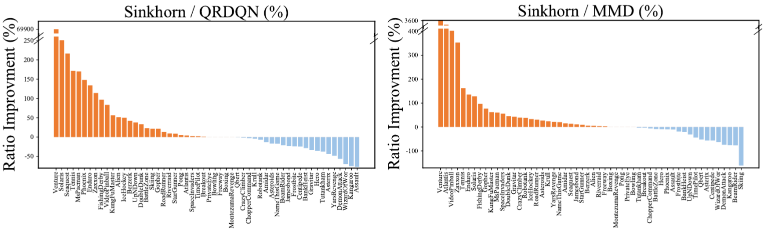


Figure 1: Mean (left), Median (middle), and IQM (5%) (right) of Human-Normalized Scores (HNS) summarized over 55 Atari games. We run 3 seeds for each algorithm.

Evaluation Metric: Human Normalized Score (HNS)

- Mean
- Median
- Interquartile Mean (%)

Ratio Improvement Analysis across All Games



(a) SinkhornDRL vs QR-DQN

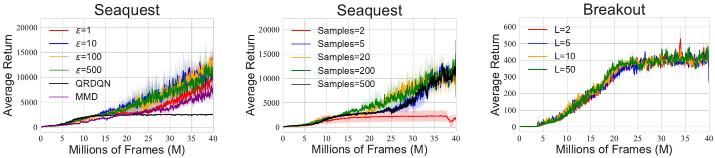
(b) SinkhornDRL vs MMD-DQN

Figure 2: Ratio improvement of return for SinkhornDRL over QR-DQN (left) and MMD-DQN (right) averaged over 3 seeds. The ratio improvement is calculated by (SinkhornDRL - QR-DQN) / QR-DQN in (a) and (SinkhornDRL - MMD-DQN) / MMD-DQN in (b), respectively.

Sensitivity Analysis and Computational Cost

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Sensitivity Analysis**



(a) Hyper-parameter ε (b) Number of Samples (c) Sinkhorn Iterations L

Figure 3: Sensitivity analysis of SinkhornDRL on Breakout and Seaquest in terms of ε, number of samples, and number of iteration L. Learning curves are reported over three seeds.

- **Computational Cost.** SinkhornDRL improves performance over baselines at the cost of slightly increasing computational burden.

Multi-Dimensional Reward Functions

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- **Reward Decomposition.** We decompose the scalar-based rewards to multi-dimensional vectors based on the respective reward structures.
- **Algorithms.**
  1. SinkhornDRL
  2. MMD-DQN
  3. Multi-dimensional Quantile Regression DQN? (not clear)

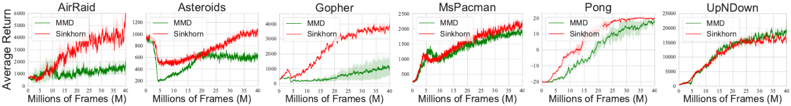


Figure 4: Performance of SinkhornDRL on six Atari games with multi-dimensional reward functions.

Outline

Conclusion: Take-away Messages

1. Sinkhorn divergence can efficiently approximate a multi-dimensional Wasserstein distance by introducing an entropic regularization, interpolation between Wasserstein distance and MMD.
2. Distributional RL under Sinkhorn divergence can also guarantee a contraction with an MDP-dependent contraction factor.
3. Distributional RL with Sinkhorn divergence can
   - Address two major limitations: unrestricted distribution representation and extension to multi-dimensional reward setting
   - Regularization effect: "smoother" transport plan and stable optimization
   - Competitive performance in extensive experiments

Open Problems and Future Work

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

1. The gap exists between theoretical properties of statistical distances and performance in RL environments.
2. It lacks a quantitative criterion to recommend in choosing an RL algorithm, given an environment.
3. Connection and discrepancy between generative models and distributional RL.

Thank You!
Questions?