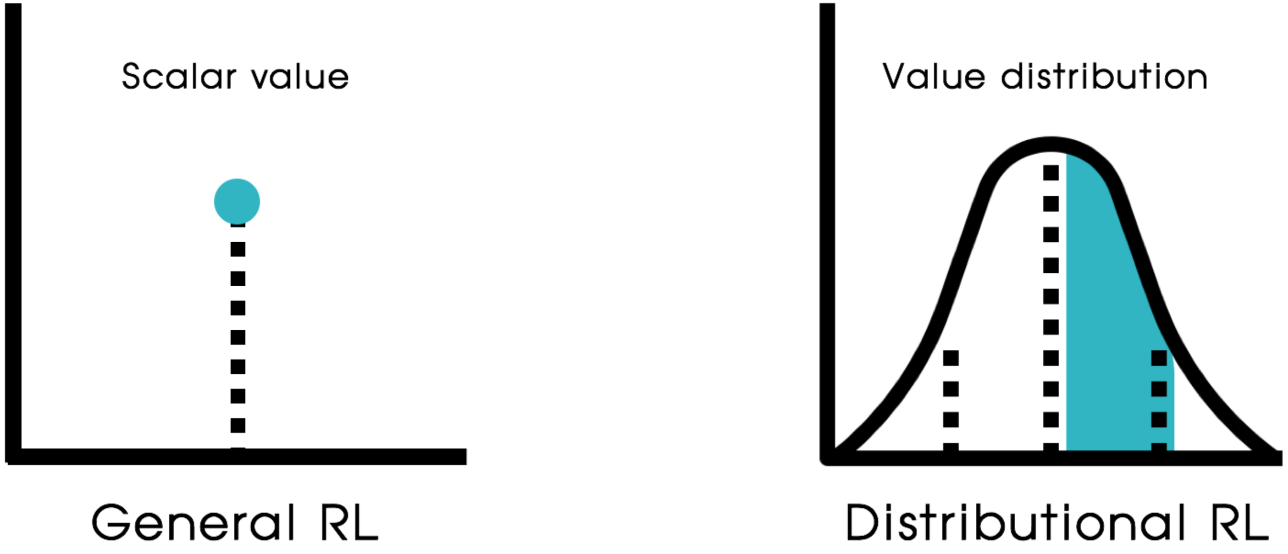
Statistics and Samples in Distributional Reinforcement Learning

Rowland, Dadashi, Kumar, Munos, Bellemare, Dabney

Topic: Distributional RL
Presenter: Isaac Waller

Distributional RL

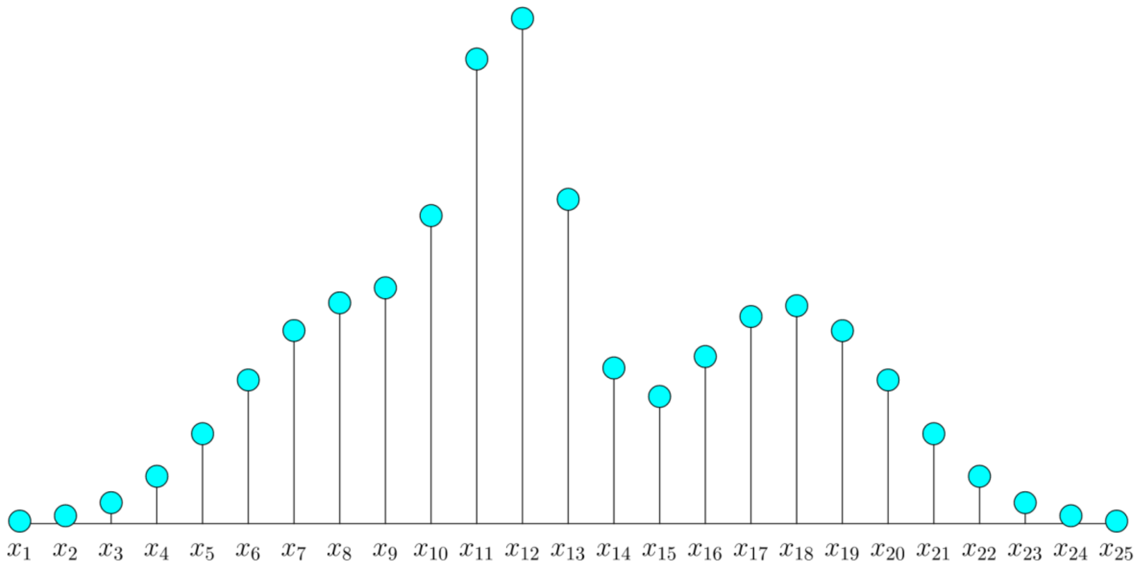


Instead of approximating the return with a value function, learn the distribution of the return $= \eta(x, a)$.

➢ A better model for multi-modal return distributions

Image https://reinforcement-learning-kr.github.io/2018/09/27/Distributional_intro/

Categorical Distributional RL (CDRL)



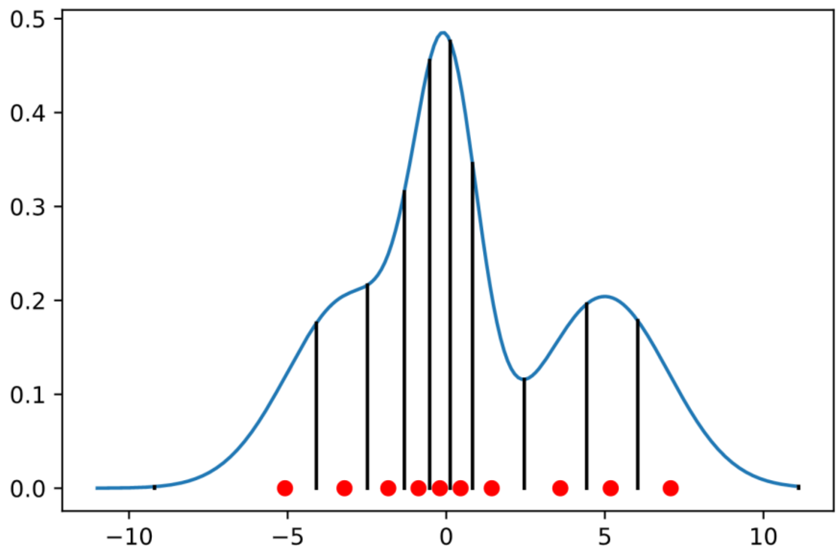
Assumes a categorical form for return distributions $\eta(x, a)$

Fixed set of supports $z_1 \dots z_K$

Learn probability $p_k(x, a)$ for each k

Image https://joshgreaves.com/reinforcement-learning/understanding-rl-the-bellman-equations/

Quantile Distributional RL (QDRL)



Learn K quantiles of the return distributions $\eta(x, a)$

Each learnable parameter z_k has equal probability mass

Image https://joshgreaves.com/reinforcement-learning/understanding-rl-the-bellman-equations/

Motivation

Lack of a **unifying framework** for these distributional RL algorithms

A general approach will

- Assess how well these algorithms model return distributions
- Inform the development of new distributional RL algorithms

Contributions

- Demonstrates that distributional RL algorithms can be decomposed into some statistics and an imputation mechanism
- Shows that CDRL and QDRL inherently cannot learn exactly the true statistics of the return distribution
- Develops a new algorithm – EDRL – which can exactly learn the true *expectiles* of the return distribution
- Empirically demonstrates that EDRL is competitive and sometimes an improvement on past algorithms

Bellman equations

$$Q^\pi(x,a) = \mathbb{E}_\pi[R_0 + \gamma Q^\pi(X_1, A_1) | X_0 = x, A_0 = a]$$

Bellman equation

$$Z^\pi(x,a) \stackrel{D}{=} R_0 + \gamma Z^\pi(X_1, A_1)$$

Distributional Bellman equation?

CDRL and QDRL Bellman updates

$$Z^\pi(x,a) \stackrel{D}{=} R_0 + \gamma Z^\pi(X_1, A_1)$$

CDRL

Update $p_k(x,a)$ to the probability mass for z_k when $Z^\pi(x,a)$ is projected onto only $z_1 \ldots z_k$.

(See Appendix A.2)

QDRL

Update quantiles z_k to the observed quantiles of $Z^\pi(x,a)$.

(See Appendix A.3)

Any algorithm = Statistics + imputation strategies

CDRL

Statistics: $\boldsymbol{s_1 \dots s_K}$
K probability masses of return distribution projected onto supports $z_1 \dots z_k$

Imputation strategy $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi(\hat{s}_{1\dots K}) = \sum^{K} \hat{s}_k \delta_{z_k}}$$

QDRL

Statistics: $\boldsymbol{s_1 \dots s_K}$
K quantiles of return distribution

Imputation strategy $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi(\hat{s}_{1\dots K}) = \frac{1}{K} \sum^{K} \delta_{\hat{s}_k}}$$

Bellman update:

$$\hat{s}_k(x,a) \leftarrow s_k\left((\mathcal{T}^{\pi}\eta)(x,a)\right)$$

Any algorithm = Statistics + imputation strategies

Algorithm 1 Generic DRL update algorithm.

Require: Statistic estimates $\hat{s}_{1:K}(x,a)\ \forall (x,a) \in \mathcal{X} \times \mathcal{A}$ and $k = 1, \dots, K$, imputation strategy Ψ.

Select state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$ to update.

Impute distribution at each possible next state-action pair:

$$\eta(x', a') = \Psi(\hat{s}_{1:K}(x', a')), \quad \forall (x', a') \in \mathcal{X} \times \mathcal{A}.$$

Update statistics at $(x,a) \in \mathcal{X} \times \mathcal{A}$:

$$\hat{s}_k(x,a) \leftarrow s_k\left((\mathcal{T}^\pi \eta)(x,a)\right).$$

Bellman closedness

Bellman closedness: a set of statistics is *Bellman closed* if, for each $(x, a) \in X \times A$, the statistics $s_{1\ldots K}\big(\eta_\pi(x, a)\big)$ can be expressed purely in terms of the random variables R_0 and $s_{1\ldots K}\big(\eta_\pi(X_1, A_1)\big) | X_0 = x, A_0 = a$ and the discount factor γ.

Theorem 4.3: Collections of moments are "effectively" the only finite sets of statistics that are Bellman closed. *Proof in Appendix B.2*

Bellman closedness

The sets of statistics used by CDRL and QDRL are not Bellman closed

Those algorithms are not capable of exactly learning their statistics (* but in practice seem to be effective anyways…)

Does not imply that they are incapable of correctly learning *expected* returns, only distribution

New algorithm: EDRL

Using expectiles

Definition 3.3 (Expectiles). *Given a distribution* $\mu \in \mathscr{P}(\mathbb{R})$ *with finite second moment, and* $\tau \in [0,1]$, *the* τ*-expectile of* μ *is defined to be the minimiser* $q^* \in \mathbb{R}$ *of the expectile regression loss* $\mathrm{ER}(q;\mu,\tau)$, *given by*

$$\mathrm{ER}(q;\mu,\tau) = \mathbb{E}_{Z\sim\mu}\left[[\tau \mathbb{1}_{Z>q} + (1-\tau)\mathbb{1}_{Z\leq q}](Z-q)^2\right].$$

For each $\tau \in [0,1]$, *we denote the* τ*-expectile of* μ *by* $e_\tau(\mu)$.

Can be **exactly** learned using Bellman updates

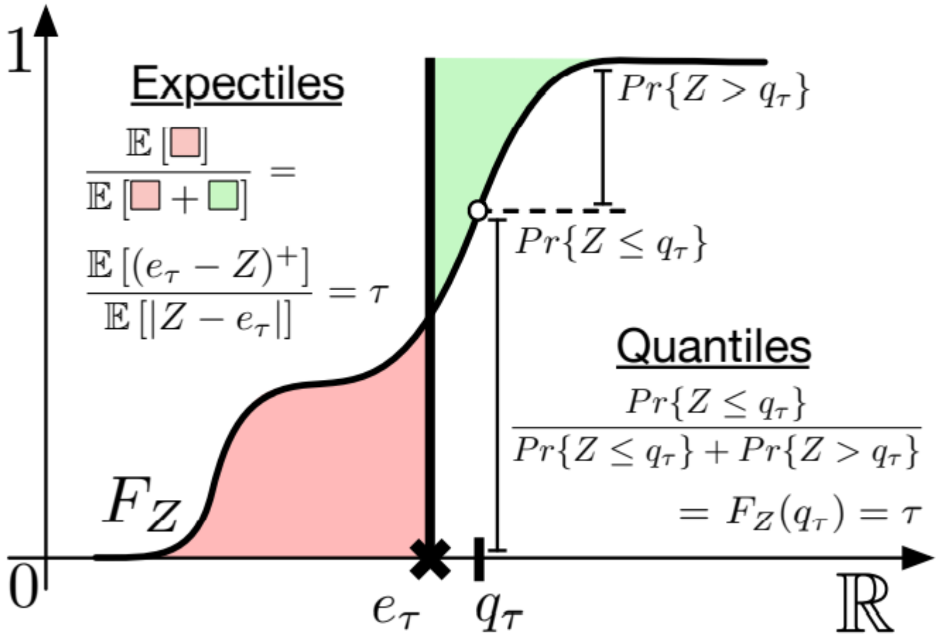


Figure 9. Diagram illustrating the similarities and differences of quantiles and expectiles.

New algorithm: EDRL

Imputation strategy:

Find a distribution satisfying (7)

$$\nabla_q \text{ER}(q; \mu, \tau_i)\big|_{q=\epsilon_i} = 0 \quad \forall i \in [K]\,. \tag{7}$$

Or (equivalently) that minimizes (8)

$$\sum_{i=1}^{K} \left(\nabla_q \text{ER}(q; \mu, \tau_i)\big|_{q=\epsilon_i} \right)^2 . \tag{8}$$

Algorithm 2 Stochastic EDRL update algorithm.

Require: Expectile estimates $\hat{s}_k(x, a)$ for each $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $k = 1, \ldots, K$.

Collect sample (x, a, r, x', a').

Impute distribution $\frac{1}{K} \sum_{k=1}^{K} \delta_{z_k}$ from target expectiles $\hat{s}_{1:K}(x', a')$ by solving (7) or minimising (8).

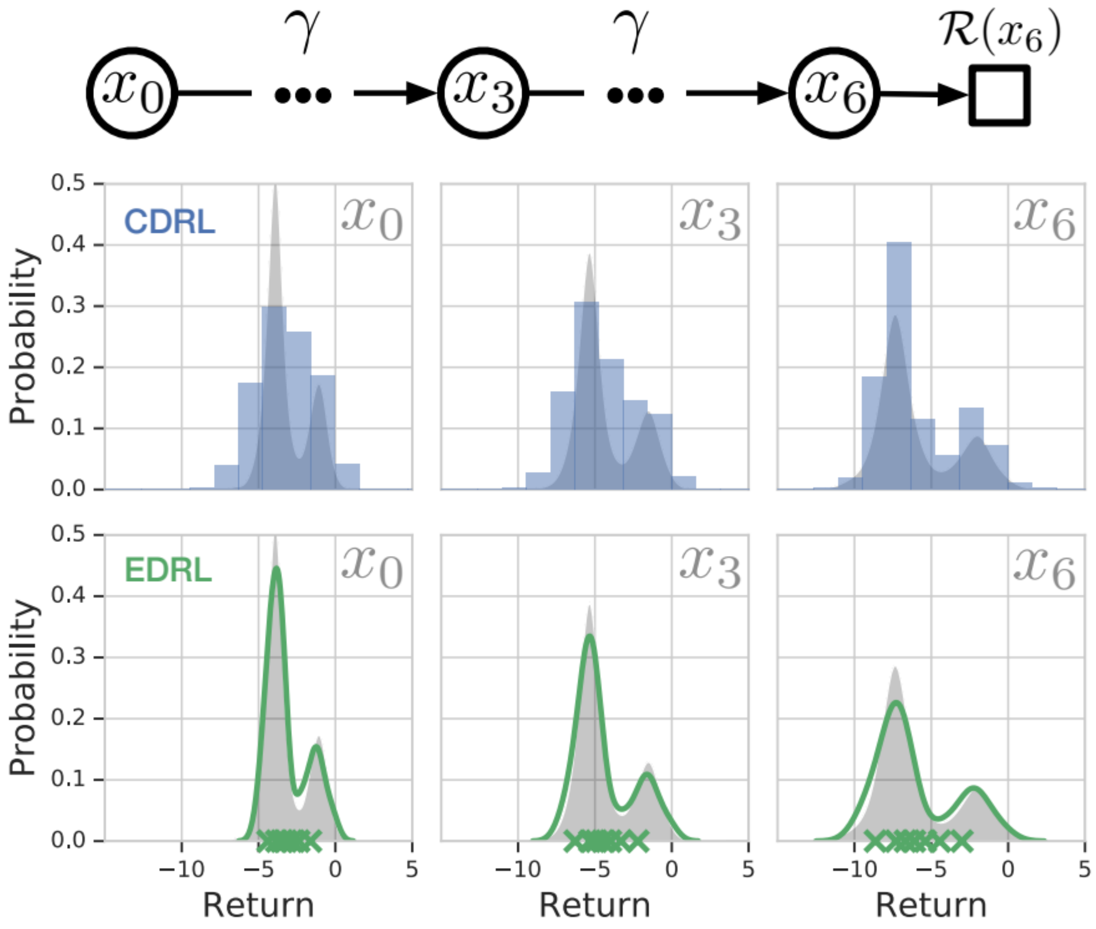
Scale/translate samples $z_i \leftarrow r + \gamma z_i \; \forall i$.

Update estimated expectiles at $(x, a) \in \mathcal{X} \times \mathcal{A}$ by computing the gradients

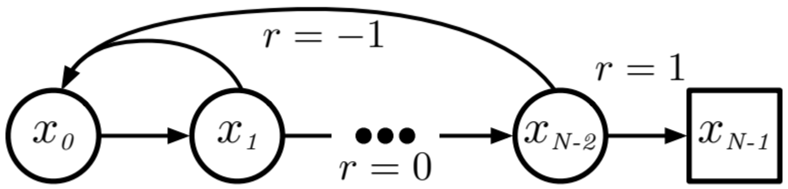
$$\nabla_{\hat{s}_k(x,a)} \sum_{k=1}^{K} \text{ER}(\hat{s}_k(x, a); \tfrac{1}{N} \sum_{n=1}^{N} \delta_{z_n}, \tau_k)$$

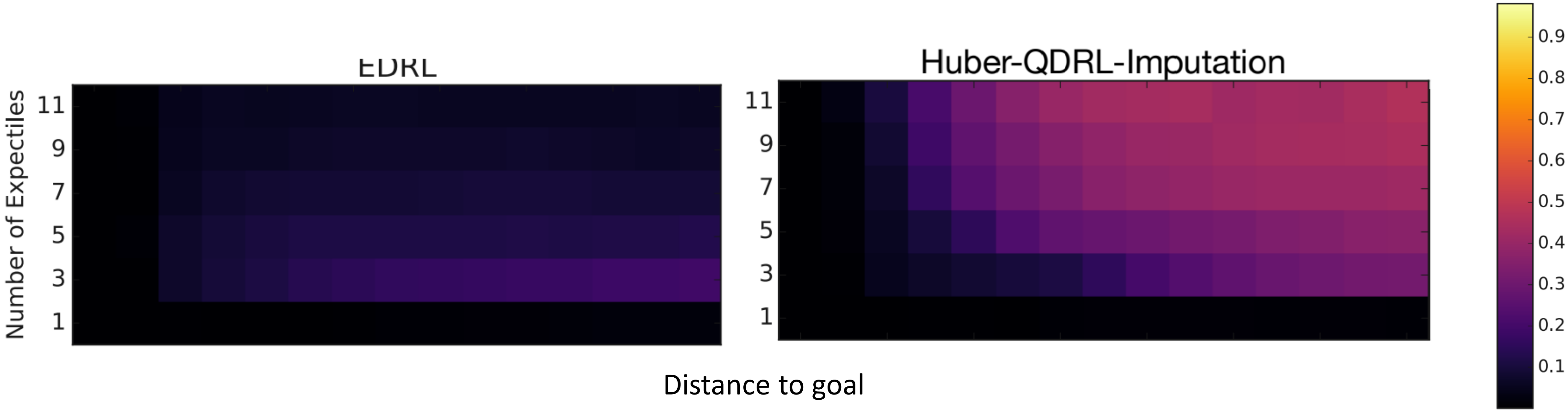
for each $k = 1, \ldots, K$.

Learnt return distributions



Experimental Results

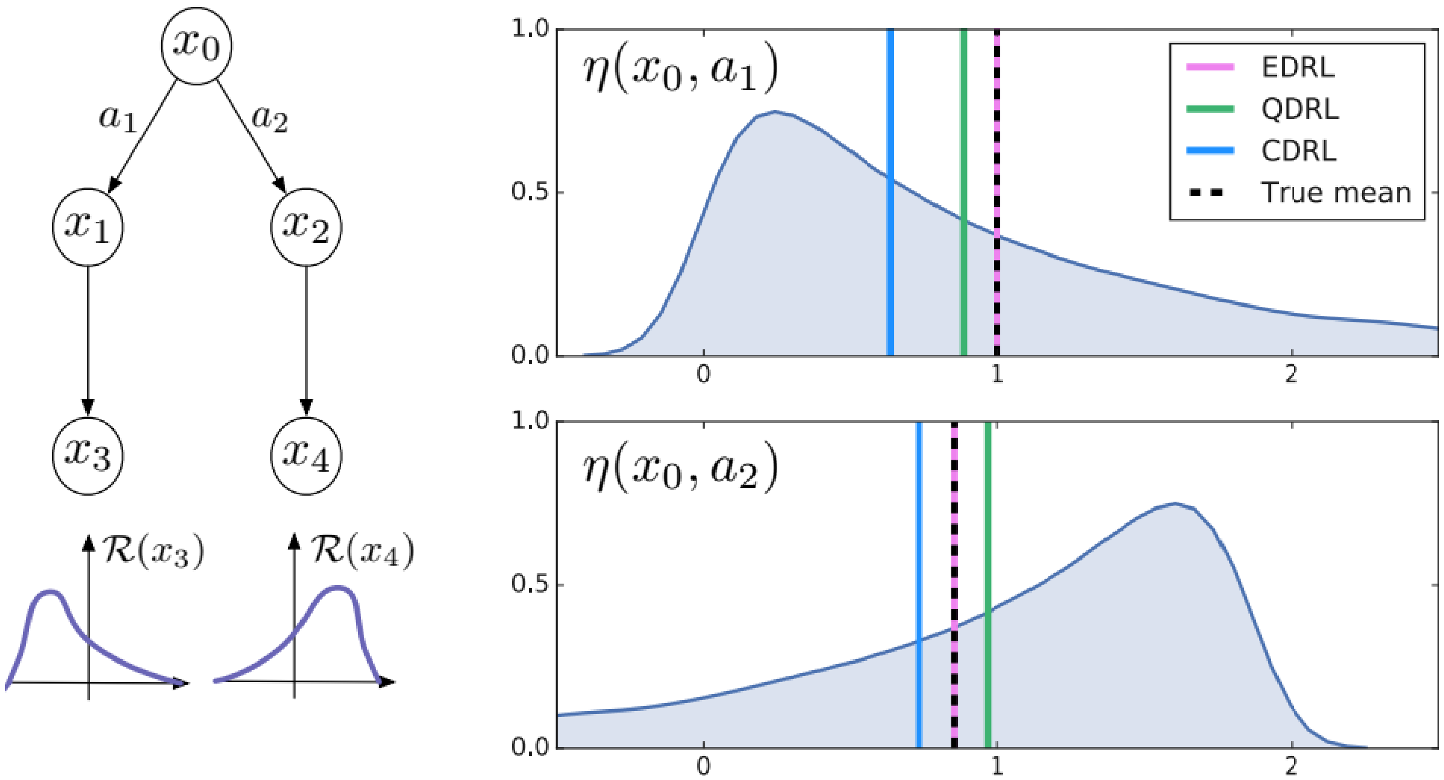




Above: estimation error

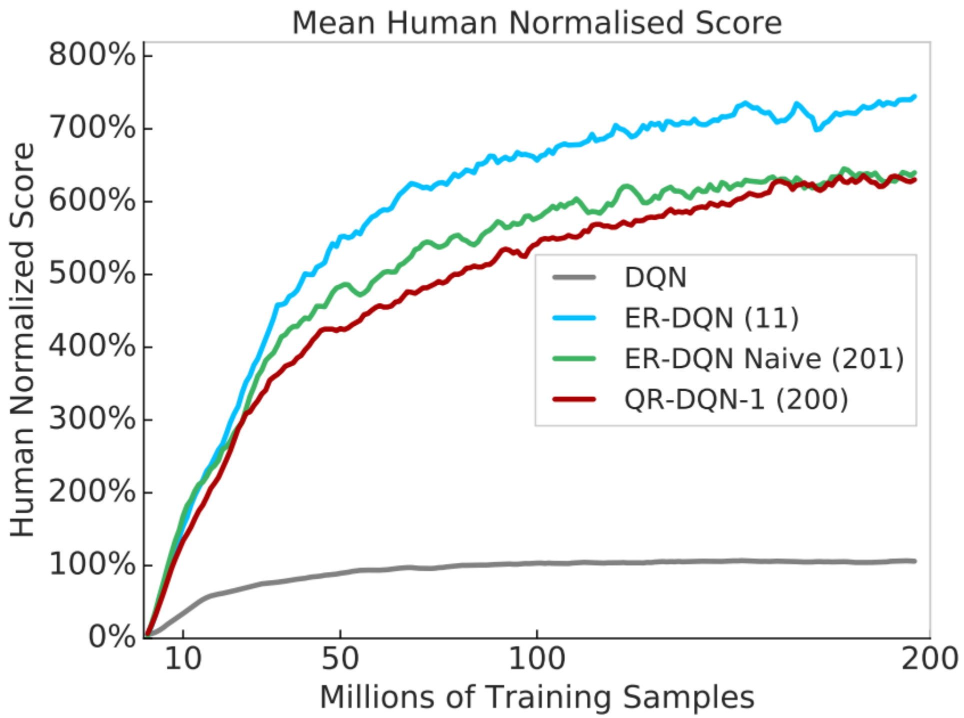
EDRL best approximates statistics

Experimental Results



EDRL does best job at estimating true mean

Experimental Results



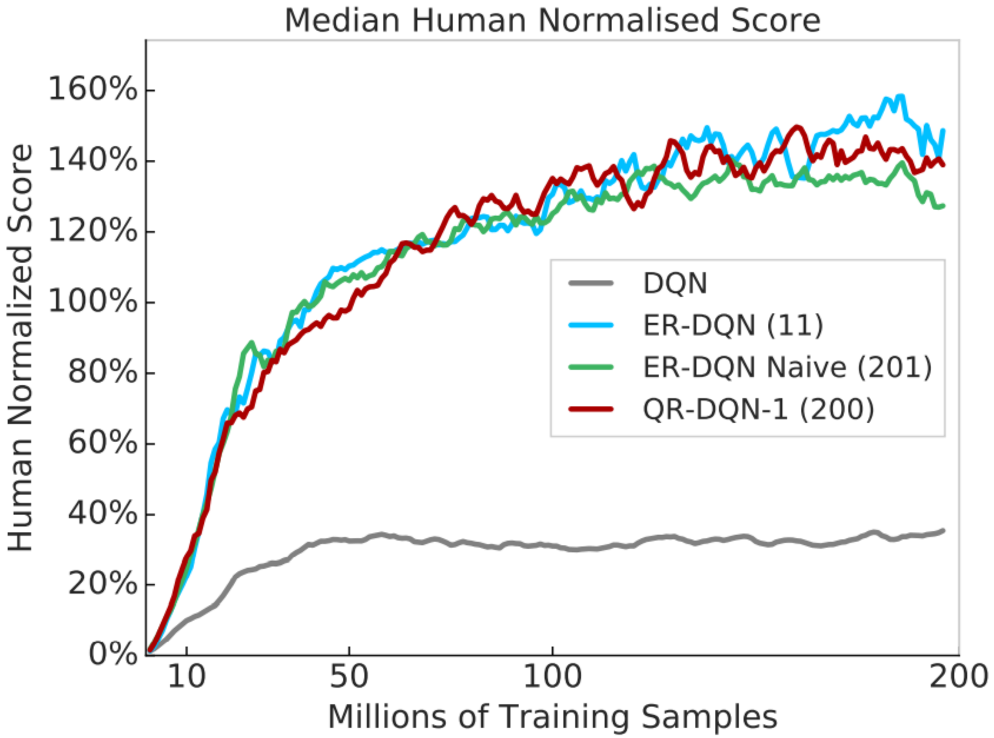


Figure 8. Mean and median human normalised scores across all 57 Atari games. Number of statistics learnt for each algorithm indicated in parentheses.

Discussion of results

- EDRL matches or exceeds performance of the other distributional RL algorithms
- Using imputation strategies grounded in the theoretical framework can improve accuracy of learned statistics

- Conclusion: the theoretical framework is sound and useful. Should be incorporated into future study in distributional RL.

Critique / Limitations / Open Issues

- EDRL does not give enormous improvements in performance over other DRL algorithms and is significantly more complex.
- Is it truly important to learn the exact return distribution? Learning an inexact distribution appears to perform fine with regards to policy performance, which is what matters in the end.
- Or: perhaps test scenarios are not complex enough to allow distributional RL to showcase true power

Contributions (Recap)

- Demonstrates that distributional RL algorithms can be decomposed into some statistics and an imputation mechanism
- Shows that CDRL and QDRL inherently cannot learn exactly the true statistics of the return distribution
- Develops a new algorithm – EDRL – which can exactly learn the true *expectiles* of the return distribution
- Empirically demonstrates that EDRL is competitive and sometimes an improvement on past algorithms

Practice questions

1. Prove the set of statistics learned under QDRL is not Bellman closed. (Hint: prove by counterexample)

2. Give an example of a set of statistics which is Bellman closed that is not expectiles or the mean.