Appendices

A. Distributional Reinforcement Learning Algorithms

For completeness, we give full descriptions of CDRL and QDRL algorithms in this section, complementing the details given in Section 2.2. We also summarise CDRL, QDRL, the exact approach to distributional RL, and our proposed algorithm EDRL, in Figure 9 at the end of this section.

A.1. The Distributional Bellman Operator

In accordance with the distributional Bellman equation (3), the distributional Bellman operator $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is defined by Bellemare et al. (2017) as

$$(\mathcal{T}^\pi\eta)(x,a) = \mathbb{E}_\pi\left[(f_{R_0,\gamma})_\#\eta(X_1,A_1)|X_0{=}x, A_0{=}a\right] ,$$

for all $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$.

A.2. Categorical Distributional Reinforcement Learning

As described in Section 2.2, CDRL algorithms are an approach to distributional RL that restrict approximate distributions to the parametric family of the form $\{\sum_{k=1}^K p_k\delta_{z_k} | \sum_{k=1}^K p_k = 1,\ p_k \geq 0 \forall k\} \subseteq \mathscr{P}(\mathbb{R})$, where $z_1 < \cdots < z_K$ are an evenly spaced, fixed set of supports. For evaluation of a policy $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$, given a collection of approximations $(\eta(x,a)|(x,a) \in \mathcal{X}\times\mathcal{A})$, the approximation at $(x,a) \in \mathcal{X}\times\mathcal{A}$ is updated according to:

$$\eta(x,a) \leftarrow \Pi_\mathcal{C}\mathbb{E}_\pi\left[(f_{R_0,\gamma})_\#\eta(X_1,A_1)|X_0{=}x, A_0{=}a\right] .$$

Here, $\Pi_\mathcal{C} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\{z_1,\ldots,z_K\})$ is a projection operator defined for a single Dirac delta as

$$\Pi_\mathcal{C}(\delta_w) = \begin{cases} \delta_{z_1} & w \leq z_1 \\ \frac{w-z_{k+1}}{z_k-z_{k+1}}\delta_{z_k} + \frac{z_k-w}{z_k-z_{k+1}}\delta_{k+1} & z_k \leq w \leq z_{k+1} \\ \delta_{z_K} & w \geq z_K , \end{cases} \tag{11}$$

and extended affinely and continuously. In the language of operators, the CDRL update may be neatly described as $\eta \leftarrow \Pi_\mathcal{C}\mathcal{T}^\pi\eta$, where we abuse notation by interpreting $\Pi_\mathcal{C}$ as an operator on collections of distributions indexed by state-action pairs, applying the transformation in Expression (11) to each distribution. The supremum-Cramér distance is defined as

$$\bar{\ell}_2(\eta_1,\eta_2) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \ell_2(\eta_1(x,a),\eta_2(x,a)) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left(\int_\mathbb{R} |F_{\eta_1(x,a)}(t) - F_{\eta_2(x,a)}(t)|^2 \mathrm{d}t\right)^{\frac{1}{2}} .$$

for all $\eta_1,\eta_2 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, where for any $\mu \in \mathscr{P}(\mathbb{R})$, F_μ denotes the CDF of μ. The operator $\Pi_\mathcal{C}\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in the supremum-Cramér distance, and so by the contraction mapping theorem, repeated CDRL updates converge to a unique limit point, regardless of the initial approximate distributions. For more details on these results and further background, see Bellemare et al. (2017); Rowland et al. (2018).

Stochastic approximation. The update $\eta \leftarrow \Pi_\mathcal{C}\mathcal{T}^\pi\eta$ is typically not computable in practice, due to unknown/intractable dynamics. An unbiased approximation to $(\mathcal{T}^\pi\eta)(x,a)$ may be obtained by interacting with the environment to obtain a transition (x,a,r,x',a'), and computing the target

$$(f_{r,\gamma})_\#\eta(x',a') .$$

It can be shown (Rowland et al., 2018) that the following is an unbiased estimator for the CDRL update $(\Pi_\mathcal{C}\mathcal{T}^\pi\eta)(x,a)$:

$$\Pi_\mathcal{C}(f_{r,\gamma})_\#\eta(x',a') .$$

Finally, the current estimate $\eta(x,a)$ can be moved towards the stochastic target by following the (semi-)gradient of some loss, in analogy with semi-gradient methods in classical RL. Bellemare et al. (2017) consider the KL loss

$$\mathrm{KL}(\Pi_\mathcal{C}(f_{r,\gamma})_\#\eta(x',a') \,||\, \eta(x,a)) ,$$

and update $\eta(x,a)$ by taking the gradient of the loss through the second argument with respect to the parameters $p_{1:K}(x,a)$. Other losses, such as the Cramér distance, may also be considered (Rowland et al., 2018).

Control. All variants of CDRL for evaluation may be modified to become control algorithms. This is achieved by adjusting the distribution of the action A_1 in the backup in an analogous way to classical RL algorithms. Instead of having $A_1 \sim \pi(\cdot|X_1)$, we instead select A_1 based on the currently estimated expected returns for each of the actions at the state X_1. For Q-learning-style algorithms, the action corresponding to the highest estimated expected return is selected:

$$A_1 = \operatorname*{arg\,max}_{a \in \mathcal{A}} \mathbb{E}_{Z \sim \eta(X_1,a)}[Z] \ .$$

However, other choices are possible, such as SARSA-style ε-greedy action selection.

A.3. Quantile Distributional Reinforcement Learning

As described in Section 2.2, QDRL algorithms are an approach to distributional RL that restrict approximate distributions to the parametric family of the form $\{\frac{1}{K}\sum_{k=1}^{K}\delta_{z_k}|z_{1:K} \in \mathbb{R}^K\} \subseteq \mathscr{P}(\mathbb{R})$. For evaluation of a policy $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$, given a collection of approximations $(\eta(x,a)|(x,a) \in \mathcal{X} \times \mathcal{A})$, the approximation at $(x,a) \in \mathcal{X} \times \mathcal{A}$ is updated according to:

$$\eta(x,a) \leftarrow \Pi_{W_1}\mathbb{E}_\pi\left[(f_{R_0,\gamma})_\# \eta(X_1,A_1)|X_0 = x, A_0 = a\right] \ ,.$$

Here, $\Pi_{W_1} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ is a projection operator defined by

$$\Pi_{\mathcal{C}}(\mu) = \frac{1}{K}\sum_{k=1}^{K}\delta_{F_\mu^{-1}(\tau_k)} \ ,$$

where $\tau_k = \frac{2k-1}{2K}$, and F_μ is the CDF of of μ. As noted in Section 2.2, $F_\mu^{-1}(\tau)$ may also be characterised as the minimiser (over $q \in \mathbb{R}$) of the quantile regression loss $\mathrm{QR}(q;\mu,\tau) = \mathbb{E}_{Z\sim\mu}\left[[\tau \mathbb{1}_{Z>q} + (1-\tau)\mathbb{1}_{Z\leq q}]\,|Z-q|\right]$; this perspective turns out to be crucial in deriving a stochastic approximation version of the algorithm.

Stochastic approximation. As for CDRL, the update $\eta \leftarrow \Pi_{W_1}\mathcal{T}^\pi\eta$ is typically not computable in practice, due to unknown/intractable dynamics. Instead, a stochastic target may be computed by using a transition (x,a,r,x',a'), and updating each atom location $z_k(x,a)$ at the current state-action pair (x,a) by following the gradient of the QR loss:

$$\nabla_q \mathrm{QR}(q;(f_{r,\gamma})_\#\eta(x',a'),\tau_k)\big|_{q=z_k(x,a)} \ .$$

Because the QR loss is affine in its second argument, this yields an unbiased estimator of the true gradient

$$\nabla_q \mathrm{QR}(q;(\mathcal{T}^\pi\eta)(x,a),\tau_k)\big|_{q=z_k(x,a)} \ .$$

Control. The methods for evaluation described above may be modified to yield control methods in exactly the same as described for CDRL in Section A.2.

A.4. Quantiles versus Expectiles

Quantiles of a distribution are given by the inverse of the cumulative distribution function. As such, they fundamentally represent threshold values for the cumulative probabilities. That is, the quantile at τ, q_τ, is greater than or equal to $\tau \times 100\%$ of the outcome values. In contrast, expectiles also take into account the *magnitude* of outcomes; the expectile at τ, e_τ, is such that the expectation of the deviations below e_τ of the random variable Z is equal to $\frac{\tau}{1-\tau}$ of the expectation of the deivations above e_τ. We illustrate these points in Figure 8.

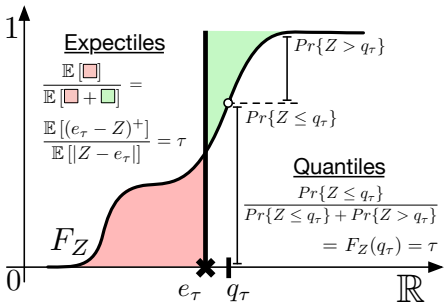


Figure 8. Diagram illustrating the similarities and differences of quantiles and expectiles.

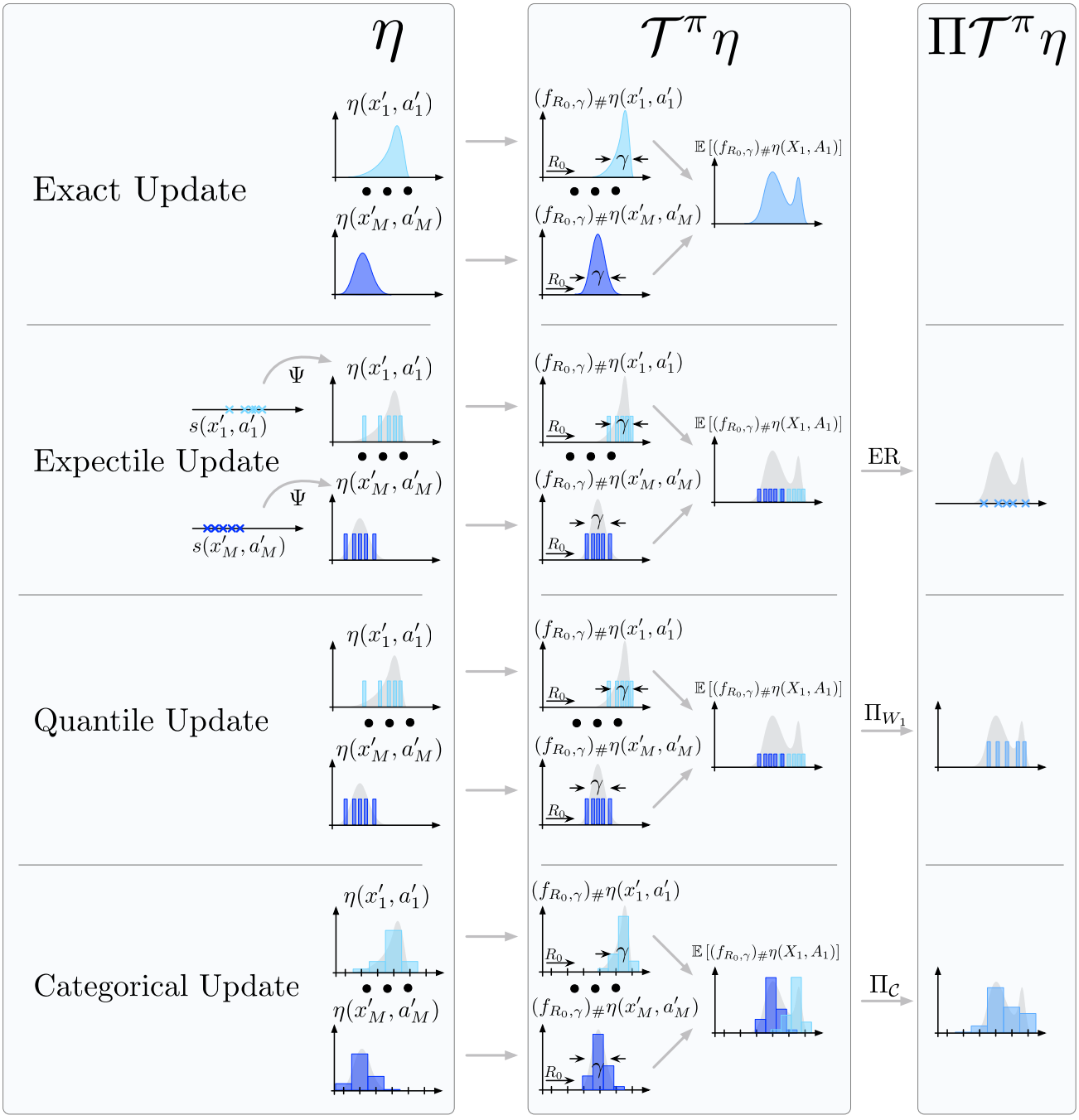


Figure 9. Illustration of distributional RL, with exact updates, expectile updates (EDRL), quantile updates (QDRL), and categorical updates (CDRL).

B. Proofs

B.1. Proofs of Results from Section 3

Lemma 3.2. *CDRL updates, with distributions supported on $z_1 < \ldots < z_K$, can be interpreted as learning the values of the following statistical functionals of return distributions:*

$$s_{z_k,z_{k+1}}(\mu) = \mathbb{E}_{Z\sim\mu}\left[h_{z_k,z_{k+1}}(Z)\right] \text{ for } k=1,\ldots,K-1\,,$$

where for $a < b$, $h_{a,b} : \mathbb{R} \to \mathbb{R}$ is a piecewise linear function defined so that $h_{a,b}(x)$ is equal to 1 for $x \leq a$, equal to 0 for $x \geq b$, and linearly interpolating between $h_{a,b}(a)$ and $h_{a,b}(b)$ for $x \in [a, b]$.

Proof. We first observe that the projection operator Π_C, defined in Section A.2, preserves the values of each of the statistical functionals $s_{z_1,z_2},\ldots,s_{z_{K-1},z_K}$, in the sense that for any distribution μ, we have $s_{z_k,z_{k+1}}(\mu) = s_{z_k,z_{k+1}}(\Pi_C\mu)$ for all $k = 1,\ldots,K$. Secondly, we observe that that the map $\{\sum_{k=1}^K p_k\delta_{z_k} | \sum_{k=1}^K p_k = 1,\ p_k \geq 0 \forall k\} \ni \mu \mapsto (s_{z_1,z_2}(\mu),\ldots,s_{z_{K-1},z_K}(\mu)) \in \mathbb{R}^{K-1}$ is injective; each distribution has a unique vector of statistics. Thus, CDRL can indeed be interpreted as learning precisely the set of statistical functionals $s_{z_1,z_2},\ldots,s_{z_{K-1},z_K}$. □

B.2. Proofs of Results from Section 4.1

Lemma 4.2. *For each $K \in \mathbb{N}$, the set of statistical functionals consisting of the first K moments is Bellman closed.*

Proof. We begin by introducing notation. Let $s_k : \mu \mapsto \mathbb{E}_{Z\sim\mu}\left[Z^k\right]$ be the k^{th} moment functional, for $k = 1,\ldots,K$. We now compute

$$\begin{aligned}
s_k(\eta_\pi(x,a)) &= \mathbb{E}_{Z\sim\eta_\pi(x,a)}\left[Z^k\right] \\
&= \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \int_{\mathbb{R}} \mathcal{R}(\mathrm{d}r|x,a)p(x'|x,a)\pi(a'|x')\mathbb{E}_{Z\sim\eta_\pi(x',a')}\left[(r+\gamma Z)^k\right] \\
&= \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \int_{\mathbb{R}} \mathcal{R}(\mathrm{d}r|x,a)p(x'|x,a)\pi(a'|x') \sum_{m=0}^{k} \binom{k}{m} \gamma^{k-m}\mathbb{E}_{Z\sim\eta_\pi(x',a')}\left[Z^{k-m}\right] r^m \\
&= \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \int_{\mathbb{R}} \mathcal{R}(\mathrm{d}r|x,a)p(x'|x,a)\pi(a'|x') \sum_{m=0}^{k} \binom{k}{m} \gamma^{k-m} s_{k-m}(\eta_\pi(x',a'))r^m \\
&= \mathbb{E}\left[\sum_{m=0}^{k} \binom{k}{m} \gamma^{k-m} s_{k-m}(\eta_\pi(X_1,A_1))R_0^m \middle| X_0 = x, A_0 = a\right].
\end{aligned}$$

Thus, $s_k(\eta_\pi(x,a))$ can be expressed in terms of R_0 and $s_{1:K}(\eta_\pi(X_1,A_1))$, as required. □

Theorem 4.3. *The only finite sets of SFs of the form $s(\mu) = \mathbb{E}_{Z\sim\mu}[h(Z)]$ that are Bellman closed are given by collections of SFs $s_1,\ldots,s_K : \mathscr{P}(\mathbb{R}) \to \mathbb{R}$ with the property that the linear span $\{\sum_{k=0}^K \alpha_k s_k | \alpha_k \in \mathbb{R}\ \forall k\}$ is equal to the linear span of the set of moment functionals $\{\mu \mapsto \mathbb{E}_{Z\sim\mu}\left[Z^l\right] | l = 0,\ldots,L\}$, for some $L \leq K$, where s_0 is the constant functional equal to 1.*

Proof. Suppose $s_1,\ldots,s_K : \mathscr{P}(\mathbb{R}) \to \mathbb{R}$ form a Bellman closed set of statistical functionals of the form $s_k(\mu) = \mathbb{E}_{Z\sim\mu}[h_k(Z)]$ for some measurable $h_k : \mathbb{R} \to \mathbb{R}$, for each $k = 1,\ldots,K$. Now note that for any MDP $(\mathcal{X},\mathcal{A},p,\gamma,\mathcal{R})$, we have the following equation:

$$s_k(\eta_\pi(x,a)) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} \int_{\mathbb{R}} \mathcal{R}(\mathrm{d}r|x,a)p(x'|x,a)\pi(a'|x')s_k((f_{r,\gamma})_\#\eta_\pi(x',a'))\,,$$

for all $(x,a) \in \mathcal{X}\times\mathcal{A}$, and for each $k = 1,\ldots,K$. By assumption of Bellman closedness, the right-hand side of this equation may be written as a function of $\mathcal{R}(x,a)$, γ, and the collection of statistics $(s_{1:K}(\eta_\pi(x',a'))|(x',a') \in \mathcal{X}\times\mathcal{A})$. Since this must hold across all valid sets of return distributions, it must the case that each $s_k((f_{r,\gamma})_\#\eta_\pi(x',a'))$ may be

written as a function of r, γ and $s_{1:K}(\eta_\pi(x', a'))$; we will write $s_k((f_{r,\gamma})_\# \eta_\pi(x', a')) = g(r, \gamma, s_{1:K}(\eta_\pi(x', a')))$ for some g.

We next claim that $g(r, \gamma, s_{1:K}(\eta_\pi(x', a')))$ is affine in $s_{1:K}(\eta_\pi(x', a'))$. To see this, note that both $s_k((f_{r,\gamma})_\# \eta_\pi(x', a'))$ and $s_{1:K}(\eta_\pi(x', a'))$ are affine as functions of the distribution $\eta_\pi(x', a')$, by assumption on the form of the statistical functionals $s_{1:K}$. Therefore $g(r, \gamma, \cdot)$ too is affine on the (convex) codomain of $s_{1:K}$.

Thus, we have

$$\mathbb{E}_{Z\sim\eta_\pi(x',a')}\left[h_k(r+\gamma Z)\right] = a_0(r,\gamma) + \sum_{k'=1}^{K} a_{k'}(r,\gamma)\mathbb{E}_{Z\sim\eta_\pi(x',a')}\left[h_{k'}(Z)\right] , \tag{12}$$

for some functions $a_{0:K} : \mathbb{R} \times [0, 1) \to \mathbb{R}$. By taking $\eta_\pi(x', a')$ to be a Dirac delta at an arbitrary real number, we obtain

$$h_k(r + \gamma x) = a_0(r, \gamma) + \sum_{k'=1}^{K} a_{k'}(r, \gamma)h_{k'}(x) \quad \text{for all } x \in \mathbb{R}\,. \tag{13}$$

In particular, the function $h_k(\gamma x)$ lies in the span of the functions $h_1, \ldots, h_K, \mathbb{1}$, where $\mathbb{1}$ is the constant function at 1. Further, $h_k(r + \gamma x)$ lies in this span for all $r \in \mathbb{R}$, and so the collection of functions $\{x \mapsto h_k(r + \gamma x)|r \in \mathbb{R}\}$ lies in a finite-dimensional subspace of functions. We may now appeal to Theorem 1 of Engert (1970), which states that any finite-dimensional space of functions which is closed under translation is spanned by a set of functions of the form

$$\bigcup_{j=1}^{J} \{x \mapsto x^\ell \exp(\lambda_j x) \mid 0 \le \ell \le L_j\}\,, \tag{14}$$

for some finite subset $\{\lambda_1, \ldots, \lambda_J\}$ of $\mathbb{C}$. From this, we deduce that each function $x \mapsto h_k(x)$ may be expressed as a linear combination of functions of the form appearing in the set in expression (14). Further, enforcing the condition that the linear span must be closed under composition with $f_{r,\gamma}$ with $\gamma \in [0, 1)$ rules out any values of λ_j above which are not zero. Therefore, the linear span of the functions $h_1, \ldots, h_K, \mathbb{1}$ must be equal to the span of some set of monomials $x \mapsto x^\ell$, $0 \le \ell \le L$, for some $L \in \mathbb{N}$, and hence the statement of the theorem follows. □

Lemma 4.4. *The sets of statistical functionals learnt under (i) CDRL, and (ii) QDRL, are not Bellman closed.*

Proof. (i) This follows as a special case of Theorem 4.3, since the statistical functionals learnt by CDRL are expectations, as shown in Lemma 3.2.

(ii) Quantiles cannot be expressed as expectations, and so we cannot appeal to Theorem 4.3. We instead proceed by describing a concrete counterexample to Bellman closedness. Fix a number $K \in \mathbb{N}$ of quantiles. Consider an MDP with a single action, and an initial state x_0 which transitions to one of two terminal states x_1, x_2 with equal probability. Suppose there is no immediate reward at state x_0. We consider two different possibilities for reward distributions at states x_1, x_2, and show that these two possibilities yield the same quantiles for the return distributions at states x_1 and x_2, but different quantiles for the return distribution at state x_0; thus demonstrating that finite sets of quantiles are not Bellman closed.

Firstly, suppose rewards are drawn from $\mathrm{Unif}([0, 1])$ at state x_1 and $\mathrm{Unif}([1/K, 1 + 1/K])$ at x_2, so that the $\frac{2k-1}{2K}$-quantile of the return at states x_1 and x_2 are $\frac{2k-1}{2K}$ and $\frac{2k+1}{2K}$, for each $k = 1, \ldots, K$. Then the return distribution at state x_0 is the mixture $\frac{1}{2}\mathrm{Unif}([0, \gamma]) + \frac{1}{2}\mathrm{Unif}([\gamma/K, \gamma + \gamma/K])$, and hence the $\frac{1}{2K}$-quantile is $\frac{\gamma}{K}$. Now, suppose instead that the reward distribution at state x_1 is $\frac{1}{K}\sum_{k=1}^{K} \delta_{\frac{2k-1}{2K}}$ and the reward distribution at x_2 is $\frac{1}{K}\sum_{k=1}^{K} \delta_{\frac{2k+1}{2K}}$. Then the $\frac{1}{2K}$-quantile of the return distribution at state x_0 is $\frac{3\gamma}{2K}$. □

B.3. Proofs of Results from Section 4.2

In this section, we use operator notation reviewed in Section A. In both proofs, the supremum-Wasserstein distance will be of use, defined as $\overline{W}_1(\mu_1, \mu_2) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\mu_1(x, a), \mu_2(x, a))$ for all $\mu_1, \mu_2 \in \mathscr{P}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. Before proving theorem 4.6, we state and prove an auxiliary lemma.

Lemma B.1. *Let $\Pi_\mathcal{C}$ be the Cramér projection for equally-spaced support points $z_1 < \cdots < z_K$, defined in Appendix Section A.2. (i) $\Pi_\mathcal{C}$ is a non-expansion in W_1. (ii) For any distribution $\mu \in \mathscr{P}(\mathbb{R})$ supported on $[z_1, z_K]$, we have $W_1(\Pi_\mathcal{C}\mu, \mu) \le \frac{z_K - z_1}{2(K-1)}$.*

Proof. In the proof of the first claim, we use the following characeterisation of the Cramér projection (Rowland et al., 2018). For any distribution $\mu \in \mathscr{P}(\mathbb{R})$ with CDF F_μ, the CDF of $\Pi_\mathcal{C}\mu$ is given by $F_{\Pi_\mathcal{C}\mu}(v) = \frac{1}{z_{k+1}-z_k}\int_{z_k}^{z_{k+1}} F_\mu(t)\mathrm{d}t$ for $v \in [z_k, z_{k+1})$, $k = 1, \ldots, K-1$, with $F_{\Pi_\mathcal{C}\mu}$ equal to 0 on (∞, z_1) and equal to 1 on $[z_K, \infty)$.

(i) Let $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R})$. We compute

$$W_1(\mu_1, \mu_2) \geq \sum_{k=1}^{K-1}\int_{z_k}^{z_{k+1}} |F_{\mu_1}(t) - F_{\mu_1}(t)|\mathrm{d}t \geq \sum_{k=1}^{K}(z_{k+1}-z_k)|F_{\Pi_\mathcal{C}\mu_1}(z_k) - F_{\Pi_\mathcal{C}\mu_1}(k)| = W_1(\Pi_\mathcal{C}\mu_1, \Pi_\mathcal{C}\mu_2)\,,$$

as required. The first inequality comes from expressing the Wasserstein distance between two distributions as the L^1 distance between their CDFs, and truncating the corresponding integral at z_1 and z_K. The second inequality follows from Jensen's inequality.

(ii) We first introduce some notation. Let $l, u : [z_1, z_K] \to \{z_1, \ldots, z_K\}$ be functions such that $l(y)$ is the largest element of $\{z_1, \ldots, z_K\}$ which is less than or equal to y, and $u(y)$ is the smallest element of $\{z_1, \ldots, z_K\}$ which is greater than or equal to y, for all $y \in [z_1, z_K]$. A valid coupling between μ and $\Pi_\mathcal{C}$ is then given as follows. Let $Y \sim \mu$, and conditional on Y, let $p \sim \text{Bernoulli}\left(\frac{Y-l(Y)}{u(Y)-l(Y)}\right)$ if $Y \notin \{z_1, \ldots, z_K\}$, and $p = 1$ almost surely conditional on $Y \in \{z_1, \ldots, z_K\}$. Then define $Z = pl(Y) + (1-p)u(Y)$. It is straightforward to check that the marginal distribution of Z is $\Pi_\mathcal{C}\mu$, and we can straightforwardly upper-bound the transport cost associated with this coupling, by observing that for each possible value y of Y, the contribution to the transport cost is 0 if $y \in \{z_1, \ldots, z_K\}$, and $\frac{u(y)-y}{u(y)-l(y)}(y-l(y)) + \frac{y-l(y)}{u(y)-l(y)}(u(y)-y) \leq \frac{u(y)-l(y)}{2} = \frac{z_K - z_1}{2(K-1)}$. Therefore, integrating over the distribution of Y gives a transport cost of at most $\frac{z_K-z_1}{2(K-1)}$, which gives the required bound on the Wasserstein distance. □

Theorem 4.6. *Consider the class $\mathcal{M}$ of MDPs with a fixed discount factor $\gamma \in [0, 1)$, and immediate reward distributions supported on $[-R_{max}, R_{max}]$. The set of SFs and imputation strategy corresponding to CDRL with evenly spaced bin locations at $-R_{max}/(1-\gamma) = z_1 < \cdots < z_K = R_{max}/(1-\gamma)$ is ε-approximately Bellman closed for $\mathcal{M}$, where $\varepsilon = \frac{\gamma}{2(1-\gamma)(K-1)}$.*

Proof. For the CDRL statistical functionals, we have $s_{z_k, z_{k+1}}(\eta_\pi(x,a)) = s_{z_k, z_{k+1}}(\Pi_\mathcal{C}\eta_\pi(x,a))$ for $k = 1, \ldots, K$ and all $(x,a) \in \mathcal{X} \times \mathcal{A}$. Further, since $\Pi_\mathcal{C}\eta_\pi(x,a)$ is supported on $\{z_1, \ldots, z_K\}$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have that $s_{z_k, z_{k+1}}(\Pi_\mathcal{C}\eta_\pi(x,a)) = F^{-1}_{\Pi_\mathcal{C}\eta_\pi(x,a)}(z_k)$. Let $(\eta(x,a)|(x,a) \in \mathcal{X} \times \mathcal{A})$ be the set of approximate distributions learnt by CDRL. As noted in Appendix Section A.2, η is the fixed point of the projected Bellman operator $\Pi_\mathcal{C}\mathcal{T}^\pi$, and η_π is the fixed point of the Bellman operator $\mathcal{T}^\pi$. We now compute:

$$\begin{aligned}
&\frac{1}{K-1}\sum_{k=1}^{K-1}\left|s_{z_k, z_{k+1}}(\eta(x,a)) - s_{z_k, z_{k+1}}(\eta_\pi(x,a))\right| \\
=&\frac{1}{K-1}\sum_{k=1}^{K-1}\left|s_{z_k, z_{k+1}}(\eta(x,a)) - s_{z_k, z_{k+1}}(\Pi_\mathcal{C}\eta_\pi(x,a))\right| \\
=&\frac{1}{K-1}\sum_{k=1}^{K-1}\left|F^{-1}_{\eta(x,a)}(z_k) - F^{-1}_{\Pi_\mathcal{C}\eta_\pi(x,a)}(z_k)\right| \\
=&\frac{1}{2R_{\max}/(1-\gamma)}\frac{2R_{\max}/(1-\gamma)}{K-1}\sum_{k=1}^{K-1}\left|F^{-1}_{\eta(x,a)}(z_k) - F^{-1}_{\Pi_\mathcal{C}\eta_\pi(x,a)}(z_k)\right| \\
=&\frac{1}{2R_{\max}/(1-\gamma)}W_1(\eta(x,a), \Pi_\mathcal{C}\eta_\pi(x,a))\,.
\end{aligned}$$

Further, note that

$$\begin{aligned}\frac{1}{2R_{\max}/(1-\gamma)}W_1(\eta(x,a),\Pi_{\mathcal{C}}\eta_\pi(x,a)) &\overset{(a)}{=}\frac{1}{2R_{\max}/(1-\gamma)}W_1(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta(x,a),\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_\pi(x,a))\\ &\overset{(b)}{\leq}\frac{1}{2R_{\max}/(1-\gamma)}\gamma\overline{W}_1(\eta,\eta_\pi)\\ &\overset{(c)}{\leq}\frac{1}{2R_{\max}/(1-\gamma)}\gamma\frac{1}{1-\gamma}\overline{W}_1(\Pi_{\mathcal{C}}\eta_\pi,\eta_\pi)\\ &\overset{(d)}{\leq}\frac{1}{2R_{\max}/(1-\gamma)}\gamma\frac{1}{1-\gamma}\frac{R_{\max}}{(1-\gamma)(K-1)}\\ &=\frac{\gamma}{2(1-\gamma)(K-1)}\,,\end{aligned}$$

as required. Here, (a) follows since η is the fixed point of $\Pi_{\mathcal{C}}\mathcal{T}^\pi$ and η_π is the fixed point of $\mathcal{T}^\pi$. (b) follows since $\Pi_{\mathcal{C}}$ is a non-expansion in $\overline{W}_1$, by Lemma B.1.(i), and $\mathcal{T}^\pi$ is a γ-contraction in $\overline{W}_1$. (c) follows from the following argument:

$$\begin{aligned}\overline{W}_1(\eta,\eta_\pi) &\leq \overline{W}_1(\eta,\Pi_{\mathcal{C}}\eta_\pi)+\overline{W}_1(\Pi_{\mathcal{C}}\eta_\pi,\eta_\pi)\\ &= \overline{W}_1(\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta,\Pi_{\mathcal{C}}\mathcal{T}^\pi\eta_\pi)+\overline{W}_1(\Pi_{\mathcal{C}}\eta_\pi,\eta_\pi)\\ &\leq \gamma\overline{W}_1(\eta,\eta_\pi)+\overline{W}_1(\Pi_{\mathcal{C}}\eta_\pi,\eta_\pi)\\ \implies \overline{W}_1(\eta,\eta_\pi) &\leq \frac{1}{1-\gamma}\overline{W}_1(\Pi_{\mathcal{C}}\eta_\pi,\eta_\pi)\,.\end{aligned}$$

Finally, (d) follows from Lemma B.1.(ii). □

Before giving a proof of Theorem 4.7, we first state and prove a lemma that will be useful.

Lemma B.2. *Let $\tau_k = \frac{2k-1}{2K}$ for $k = 1,\dots,K$, and consider the corresponding Wasserstein-1 projection operator $\Pi_{W_1}:\mathscr{P}(\mathbb{R})\to\mathscr{P}(\mathbb{R})$, defined by*

$$\Pi_{W_1}(\mu) = \frac{1}{K}\sum_{k=1}^{K}\delta_{F_\mu^{-1}(\tau_k)}\,,$$

for all $\mu\in\mathscr{P}(\mathbb{R})$, where F_μ^{-1} is the inverse c.d.f. of μ. Let $\eta_1,\eta_2\in\mathscr{P}(\mathbb{R})$, such that $\sup(\textit{supp}(\eta_i)) - \inf(\textit{supp}(\eta_i)) \leq I$ for $i=1,2$. Then we have:

$$\begin{aligned}&(i)\ W_1(\Pi_{W_1}\eta_1,\eta_1)\leq\frac{I}{K}\,;\\ &(ii)\ W_1(\Pi_{W_1}\eta_1,\Pi_{W_1}\eta_2)\leq W_1(\eta_1,\eta_2)+\frac{2I}{K}\,.\end{aligned}$$

Proof. We start by proving (i). Let $F_{\eta_1}^{-1}$ be the inverse c.d.f of η_1. We have

$$\begin{aligned}W_1(\mu,\Pi_{W_1}\mu) &= \sum_{i=0}^{K-1}\frac{1}{K}\mathbb{E}_{X\sim\mu}\left[\left|X-F_{\eta_1}^{-1}\left(\frac{2i+1}{2K}\right)\right|\,\middle|\,F_{\eta_1}^{-1}\left(\frac{i}{K}\right)\leq X\leq F_{\eta_1}^{-1}\left(\frac{i+1}{K}\right)\right]\\ &\leq\frac{1}{K}\left(F_{\eta_1}^{-1}(1)-F_{\eta_1}^{-1}(0)\right)\\ &=\frac{I}{K}\end{aligned}$$

We can now prove (ii), using the triangle inequality and (i):

$$\begin{aligned}W_1(\Pi_{W_1}\eta_1,\Pi_{W_1}\eta_2) &\leq W_1(\Pi_{W_1}\eta_1,\eta_1)+W_1(\eta_1,\eta_2)+W_1(\eta_2,\Pi_{W_1}\eta_2)\\ &\leq W_1(\eta_1,\eta_2)+\frac{2I}{K}\,.\end{aligned}$$

□

Theorem 4.7. *Consider the class of MDPs $\mathcal{M}$ with a fixed discount factor $\gamma \in [0, 1)$, and immediate reward distributions supported on $[-R_{max}, R_{max}]$. Then the collection of quantile SFs $s_k(\mu) = F_\mu^{-1}(\frac{2k-1}{2K})$ for $k = 1, \ldots, K$, with the standard QDRL imputation strategy, is ε-approximately Bellman closed for $\mathcal{M}$, where $\varepsilon = \frac{2R_{max}(5-2\gamma)}{(1-\gamma)^2 K}$.*

Proof. Let $(\hat{s}_{1:K}(x, a)|(x, a) \in \mathcal{X} \times \mathcal{A})$ be the collection of statistics learnt under QDRL. We denote by $\eta(x, a)$ the distribution imputed from the statistics $\hat{s}_{1:K}(x, a)$, for each $(x, a) \in \mathcal{X} \times \mathcal{A}$. As noted in Appendix Section A.3, η is the fixed point of the projected Bellman operator $\Pi_{W_1}\mathcal{T}^\pi$, and η_π is the fixed point of $\mathcal{T}^\pi$. We begin by noting that if all immediate reward distributions have support contained within $[-R_{\max}, R_{\max}]$, then the true and learnt reward distributions are supported on $[-R_{\max}/(1-\gamma), R_{\max}/(1-\gamma)]$, and further, so are the distributions $\mathcal{T}^\pi\eta(x, a)$ for each $(x, a) \in \mathcal{X} \times \mathcal{A}$. We thus compute

$$\begin{aligned}
&\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \frac{1}{K}\sum_{k=1}^{K} |s_k(\eta_\pi(x, a)) - \hat{s}_k(x, a)| \\
&= \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\Pi_{W_1}\eta(x, a), \Pi_{W_1}\eta_\pi(x, a)) \\
&\leq \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\eta(x, a), \eta_\pi(x, a)) + \frac{4R_{\max}}{K(1-\gamma)},
\end{aligned}$$

with the inequality following from Lemma B.2(ii). From here, we note that

$$\begin{aligned}
\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\eta(x, a), \eta_\pi(x, a)) &\overset{(a)}{\leq} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} [W_1(\eta(x, a), \Pi_{W_1}\eta_\pi(x, a)) + W_1(\Pi_{W_1}\eta_\pi(x, a), \eta_\pi(x, a))] \\
&\overset{(b)}{\leq} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\eta(x, a), \Pi_{W_1}\eta_\pi(x, a)) + \frac{2R_{\max}}{K(1-\gamma)} \\
&\overset{(c)}{=} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\Pi_{W_1}\mathcal{T}^\pi\eta(x, a), \Pi_{W_1}\mathcal{T}^\pi\eta_\pi(x, a)) + \frac{2R_{\max}}{K(1-\gamma)} \\
&\overset{(d)}{\leq} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\mathcal{T}^\pi\eta(x, a), \mathcal{T}^\pi\eta_\pi(x, a)) + \frac{4R_{\max}}{K(1-\gamma)} + \frac{2R_{\max}}{K(1-\gamma)} \\
&\overset{(e)}{\leq} \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \gamma W_1(\eta(x, a), \eta_\pi(x, a)) + \frac{6R_{\max}}{K(1-\gamma)} \\
\implies \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\eta(x, a), \eta_\pi(x, a)) &\leq \frac{6R_{\max}}{K(1-\gamma)^2}.
\end{aligned}$$

Here, (a) follows from the triangle inequality, (b) follows from Lemma B.2(i). (c) follows since η is the fixed point of $\Pi_{W_1}\mathcal{T}^\pi$ and η_π is the fixed point of $\mathcal{T}^\pi$. (d) follows from Lemma B.2(ii), where we use the fact that the support of the distributions constituting the fixed points of $\Pi_{W_1}\mathcal{T}^\pi$ and $\mathcal{T}^\pi$ necessarily are supported on $[-R_{\max}/(1-\gamma), R_{\max}/(1-\gamma)]$. (e) follows from the γ-contractivity of the Bellman operator $\mathcal{T}^\pi$ with respect to the metric $\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} W_1(\mu_1(x, a), \mu_2(x, a))$, for $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ (Bellemare et al., 2017). Hence, we obtain

$$\begin{aligned}
\sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \frac{1}{K}\sum_{k=1}^{K} |s_k(\eta_\pi(x, a)) - \hat{s}_k(x, a)| &\leq \frac{6R_{\max}}{K(1-\gamma)^2} + \frac{4R_{\max}}{K(1-\gamma)} \\
&= \frac{2R_{\max}(5-2\gamma)}{K(1-\gamma)^2}.
\end{aligned}$$

□

B.4. Proofs of Results from Section 4.3

Lemma 4.8. *(i) Under CDRL updates using support locations $z_1 < \cdots < z_K$, if all approximate reward distributions have support bounded in $[z_1, z_K]$, expected returns are exactly learnt. (ii) Under QDRL updates, expected returns are not exactly learnt.*

Proof. (i) The statistical functionals learnt by CDRL are of the form $s_k(\mu) = \mathbb{E}_{Z\sim\mu}\left[h_{z_k,z_{k+1}}(Z)\right]$, for $k = 1, \ldots, K-1$. We observe that the mean functional $m(\mu) = \mathbb{E}_{Z\sim\mu}[Z]$ is contained in the linear span of $s_{0:K-1}$, where $s_0(\mu) = 1$ for all μ. Indeed,

$$m = R_{\max}s_0 - \left(\frac{R_{\max} - R_{\min}}{K}\right)\sum_{k=1}^{K-1} s_k\,,$$

since

$$x = R_{\max} - \left(\frac{R_{\max} - R_{\min}}{K}\right)\sum_{k=1}^{K-1} h_{z_k,z_{k+1}}(x)$$

for all $x \in [-R_{\min}, R_{\max}]$. Since the singleton set consisting of the mean functional is Bellman closed, it follows that whatever distribution is imputed, the effective update to the mean of the distribution at the current state is the same as updating according to the classical Bellman update for the mean.

(ii) We note that the mean is not encoded by a finite set of quantiles, and hence it is impossible for expected returns to be correctly in general. To make this concrete, fix a number K of quantiles to be learnt, and consider a single state, two action MDP, with reward distribution $\frac{4K-1}{4K}\delta_0 + \frac{1}{4K}\delta_1$ for the first action, and reward distribution $\delta_{1/8K}$ for the second action. Fitting quantiles at $\tau \in \{\frac{2k-1}{2K} | k = 1, \ldots, K\}$ results in all quantiles for the first distribution being equal to 0, and thus the imputed distribution is δ_0, resulting in a imputed mean of 0. By constrast, for the second distribution, all quantiles are fitted at $1/8K$, resulting in an imputed distribution of $\delta_{1/8K}$ and an imputed mean of $1/8K$. Thus, a QDRL control algorithm will act greedily with respect to these imputed means and select the second action, which is sub-optimal as the first action has higher expected reward. □

C. Additional Theoretical Results

In this section, we provide several examples to illustrate the point made in Section 4.2 that in general, it is not possible to simultaneously achieve low approximation error on all statistics in a non-Bellman closed collection.

Lemma C.1. *For a fixed $K \in \mathbb{N}$, let $s_{1:K-1}$ be the statistical functionals corresponding to CDRL (with fixed discount factor $\gamma \in [0, 1)$) with equally spaced support $R_{min} = z_1 < \ldots < z_K = R_{max}$. As earlier in the paper, we denote by $\hat{s}_k(x, a)$ the relevant* learnt *value of the statistical functional concerned. Then we have:*

$$\sup_{\substack{\mathcal{M}\ MDP\\ \pi\ policy}} \sup_{x\in\mathcal{X}} \sup_{k=1,\ldots,K-1} |\hat{s}_k(x, a) - s_k(\eta_\pi(x, a))| \not\to 0$$

as $K \to \infty$.

Proof. We work with a particular family of MDPs with two states x_1, x_2, one action in each state, with x_1 transitioning to x_2 with probability 1, and x_2 terminal. In such MDPs, there is only one policy, which we denote by π; and we drop notational dependence on actions for clarity. No rewards are received at state x_1; we specify the rewards received at state x_2 below. We take a discount factor $\gamma = \frac{2^m}{2^m+1}$ for some $k \in \mathbb{N}$. Fix $L \in \mathbb{N}$, and consider CDRL updates with bin locations at $z_k = \frac{k}{2^L}$ for $k = 0, \ldots, 2^L$. Specifically, consider learning the statistical functional

$$\mathbb{E}_{Z\sim\eta_\pi(x_1)}\left[h_{\frac{1}{2},\frac{1}{2}+\frac{1}{2^L}}(Z)\right].$$

Since there are no rewards received at state x_1, at convergence the estimate of this statistic (which we denote by $\hat{s}(x_1)$) is equal to

$$\mathbb{E}_{Z\sim\hat{\eta}(x_2)}\left[h_{\frac{1}{2},\frac{1}{2}+\frac{1}{2^L}}(\gamma Z)\right] = \mathbb{E}_{Z\sim\hat{\eta}(x_2)}\left[h_{\frac{\gamma-1}{2},\frac{\gamma-1}{2}+\frac{\gamma-1}{2^L}}(Z)\right] = \mathbb{E}_{Z\sim\hat{\eta}(x_2)}\left[h_{\frac{1}{2}+\frac{1}{2^{m+1}},\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L}+\frac{1}{2^{L+m}}}(Z)\right]$$

where $\hat{\eta}(x_2)$ is the approximate return distribution learnt at state x_2. Now, consider two possible reward distributions at state x_2:

$$\rho_A = \delta_{\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{3}{2^{L+1}}}\,,\ \text{and}\ \rho_B = \frac{1}{2}\left(\delta_{\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L}} + \delta_{\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{2}{2^L}}\right).$$

Under these two reward distributions, the fitted distribution $\eta(x_2)$ is the same, namely ρ_B, and thus the estimate $\hat{s}(x_1)$ is the same. Our aim is to show that for these two different reward distributions, the difference of the true values of the statistical functional $\hat{s}(x_1)$ is independent of L, and hence the value of $\hat{s}(x_1)$ cannot converge to the true statistic as $L \to \infty$. To achieve this, and finish the proof, we calculate directly. In the case where the reward distribution at state x_2 is ρ_A, we have (assuming $L > m+1$)

$$s(\eta_\pi(x_1)) = \mathbb{E}_{Z\sim\rho_A}\left[h_{\frac{1}{2}+\frac{1}{2^{m+1}},\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L}+\frac{1}{2^{L+m}}}(Z)\right] = 0\,.$$

In the case where the reward distribution at state x_2 is ρ_B, we have

$$\begin{aligned} s(\eta_\pi(x_1)) &= \mathbb{E}_{Z\sim\rho_B}\left[h_{\frac{1}{2}+\frac{1}{2^{m+1}},\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L}+\frac{1}{2^{L+m}}}(Z)\right] = \\ &= \frac{1}{2}\left(\frac{(\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L})-(\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L}+\frac{1}{2^{L+m}})}{(\frac{1}{2}+\frac{1}{2^{m+1}})-(\frac{1}{2}+\frac{1}{2^{m+1}}+\frac{1}{2^L}+\frac{1}{2^{L+m}})}\right) \\ &= \frac{1}{2}\left(\frac{1}{2^m+1}\right)\,. \end{aligned}$$

□

Lemma C.2. *For a fixed $K \in \mathbb{N}$, let $s_{1:K-1}$ be the statistical functionals corresponding to by QDRL (with fixed discount factor $\gamma \in [0,1)$). As earlier in the paper, we denote by $\hat{s}_k(x,a)$ the relevant* learnt *value of the statistical functional concerned. Then we have:*

$$\sup_{\substack{\mathcal{M}\ MDP\\ \pi\ policy}}\sup_{x\in\mathcal{X}}\sup_{a\in\mathcal{A}}\sup_{k=1,\ldots,K}|\hat{s}_k(x,a) - s_k(\eta_\pi(x,a))| \not\to 0$$

as $K \to \infty$.

Proof. We work with a particular family of MDPs with three states x_0, x_1, x_2, one action in each state, with x_0 transitioning to x_1 with probability $\frac{1}{2}-\varepsilon$ and with x_0 transitioning to x_2 with probability $\frac{1}{2}+\varepsilon$ with $\varepsilon \ll 1$. We take x_1 and x_2 to be terminal, no rewards are received at state x_0; we specify the rewards received at state x_1 and x_2 below. We suppose in the following that K is odd.

The reward distributions at state x_1 and x_2 are given by

$$\rho_1 = \left(\frac{1}{2K}-\varepsilon\right)\delta_0 + \left(\frac{2K-1}{2K}+\varepsilon\right)\delta_1\,,\ \text{and}\ \rho_2 = \left(\frac{1}{2K}-\varepsilon\right)\delta_0 + \left(\frac{2K-1}{2K}+\varepsilon\right)\delta_{-1}\,.$$

Under these reward distributions the fitted return distributions are:

$$\eta(x_1) = \delta_1\,,\ \text{and}\ \eta(x_2) = \delta_{-1}\ .$$

Therefore, we have

$$s_{\frac{K+1}{2}}(\eta_\pi(x_0)) = 0\,,\ \text{and}\ \hat{s}_{\frac{K+1}{2}}(x_0) = -\gamma\,.$$

□

D. ER-DQN Experimental Details

D.1. ER-DQN Architecture

As discussed in Section 5.3, the ER-DQN architecture matches the exact architecture of QR-DQN (Dabney et al., 2017). The Q-network, for a given input x, outputs expectiles $e_{\tau_{1:K}}(x,a)$ for each $a \in \mathcal{A}$. In our experiments with 11 expectiles, we take $\tau_{1:11}$ to be linearly spaced with $\tau_1 = 0.01$, $\tau_{11} = 0.99$. Note that we have $\tau_6 = 0.5$, and thus this expectile is in fact the mean. For the purposes of control, greedy actions at a state $x \in \mathcal{X}$ are thus selected according to $\arg\max_{a\in\mathcal{A}} e_{\tau_6}(x,a)$, rather than averaging over statistics as in QR-DQN. For the imputation strategy, we take the root-finding problem in Expression (6), and use a call to the SciPy `root` routine with default parameters.

D.2. Training Details

We use the Adam optimiser with a learning rate of 0.00005, after testing learning rates 0.00001, 0.00003, 0.00005, 0.00007, and 0.0001 on a subset of 6 Atari games. All other hyperparameters in training correspond to those used in (Dabney et al., 2017). In particular, the target distribution is computed from a target network. Note that each training pass requires a call to the SciPy optimiser to compute the imputed samples, and thus in general will be more computationally expensive than other deep distributional Q-learning-style agents, such as C51 and QR-DQN. However, by parallelising the optimiser calls for a minibatch of transitions across several CPUs, we found that training times when using 11 expectiles to be comparable to training times of QR-DQN.

For ER-DQN Naive, we found that results were slightly improved by using 201 expectiles compared to 11, so include results with this larger number of statistical functionals in the main paper. We take $\tau_{1:201}$ according to the same prescription as for QR-DQN: linearly spaced, with $\tau_1 = 1/(2 \times 201)$ and $\tau_{201} = 1 - 1/(2 \times 201)$.

D.3. Environment Details

We use the Arcade Learning Environment (Bellemare et al., 2013) to train and evaluate ER-DQN on a selection of 57 Atari games. The precise parameter settings of the environment are exactly the same as in the experiments performed on QR-DQN, to allow for direct comparison.

D.4. Detailed Results

In addition to the human normalised mean/median results presented in the main paper, we include training curves for all 4 evaluated agents on all 57 Atari games in Figure 10, and raw maximum scores attained in Table 1.

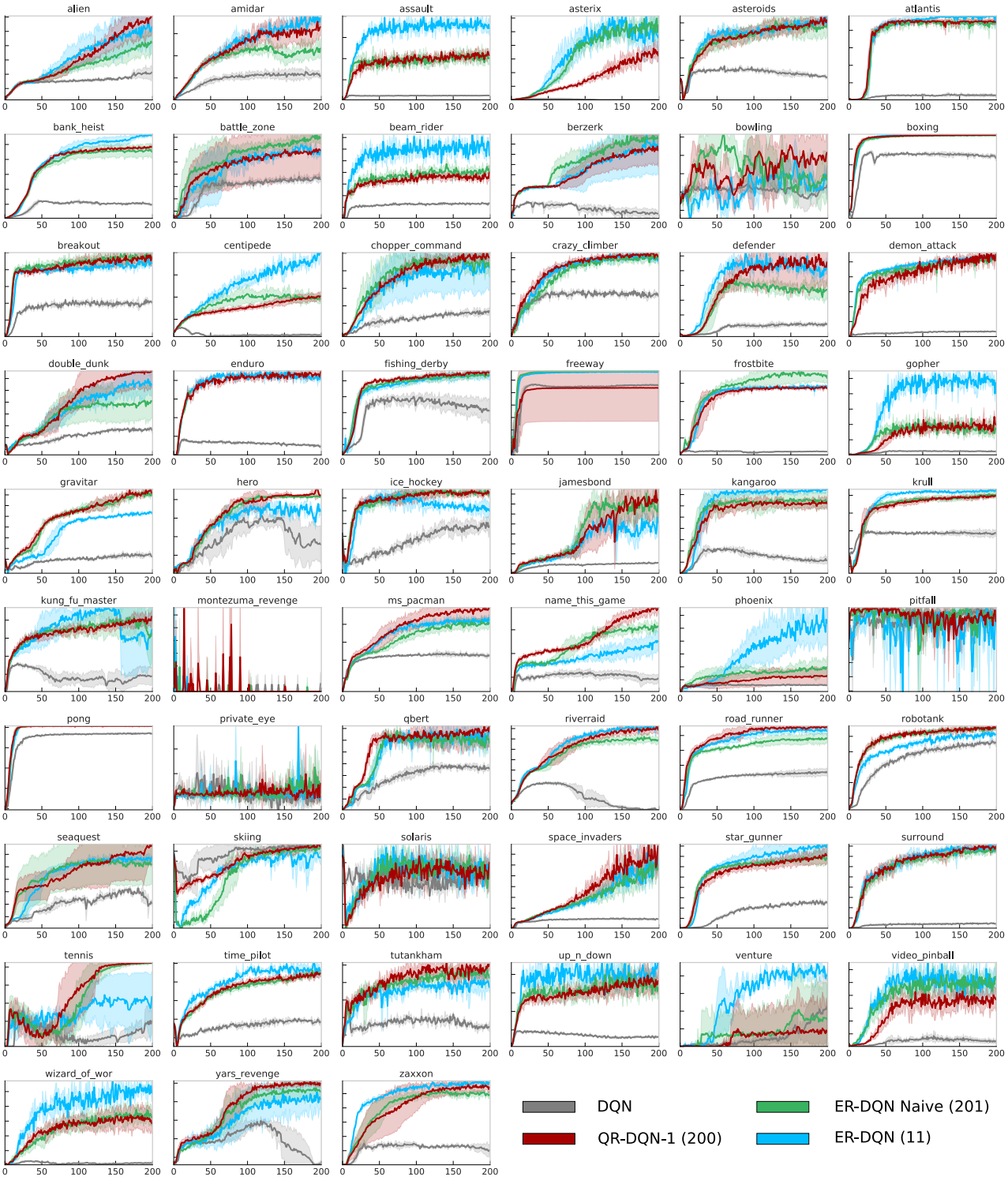


Figure 10. Training curves for DQN, QR-DQN-1, ER-DQN Naive, and ER-DQN on all 57 Atari games.

GAMES	QR-DQN-1	ER-DQN NAIVE	ER-DQN
alien	**7279.5**	5056.2	6212.0
amidar	2235.8	1528.8	**2313.0**
assault	17653.9	19156.2	**25826.8**
asterix	306055.9	366152.1	**434743.6**
asteroids	3484.4	3250.9	**3793.2**
atlantis	947995.0	939050.0	**974408.3**
bank_heist	1185.7	1132.5	**1326.5**
battle_zone	33987.2	**40805.3**	35098.5
beam_rider	25095.7	29542.5	**48230.1**
berzerk	2151.2	2626.6	**2749.8**
bowling	58.0	**63.4**	53.1
boxing	99.5	99.4	**99.9**
breakout	505.2	**538.6**	509.8
centipede	11465.1	12325.3	**22505.9**
chopper_command	**12767.2**	11765.8	11886.1
crazy_climber	159244.2	158369.9	**161040.2**
defender	**41098.7**	32225.2	36473.5
demon_attack	**114530.2**	108496.2	111921.2
double_dunk	**16.5**	4.0	16.3
enduro	2294.1	1923.9	**2339.5**
fishing_derby	**21.6**	18.4	20.2
freeway	27.2	**34.0**	33.9
frostbite	4068.1	**5408.0**	4233.7
gopher	82060.6	86874.1	**115828.3**
gravitar	937.0	**942.8**	680.9
hero	**23799.1**	21916.6	20374.5
ice_hockey	**-1.7**	-1.9	-2.7
jamesbond	5298.5	**5440.4**	4113.6
kangaroo	14827.6	15371.1	**15954.4**
krull	10591.2	10738.0	**11318.5**
kung_fu_master	49695.5	52080.6	**58802.2**
montezuma_revenge	**0.1**	0.0	0.0
ms_pacman	**5860.4**	4856.1	5048.5
name_this_game	**20509.1**	17064.9	13090.9
phoenix	15475.2	25177.3	**91189.4**
pitfall	**0.0**	**0.0**	**0.0**
pong	21.0	**21.0**	21.0
private_eye	**531.3**	388.3	176.3
qbert	**17573.5**	14536.0	17418.4
riverraid	18125.3	15726.4	**18472.2**
road_runner	**67084.8**	57168.0	64577.7
robotank	**58.0**	56.7	54.8
seaquest	16143.3	13501.0	**19401.0**
skiing	-16869.1	-15085.4	**-10528.6**
solaris	2615.3	2483.3	**2810.6**
space_invaders	11873.3	10099.6	**14265.7**
star_gunner	76556.3	75404.8	**88900.3**
surround	8.4	8.2	**8.6**
tennis	**22.8**	22.7	5.8
time_pilot	9902.0	10009.6	**11675.5**
tutankham	**282.8**	256.7	237.9
up_n_down	**44893.6**	35169.7	32083.3
venture	266.5	476.7	**1107.0**
video_pinball	570852.7	603852.1	**727091.1**
wizard_of_wor	21667.1	24397.5	**36049.8**
yars_revenge	**27264.3**	26056.7	24099.4
zaxxon	11707.1	11120.2	**12264.4**

Table 1. Raw max test scores across all 57 Atari games, starting with 30 no-op actions.

E. Additional Material on Tabular Experiments

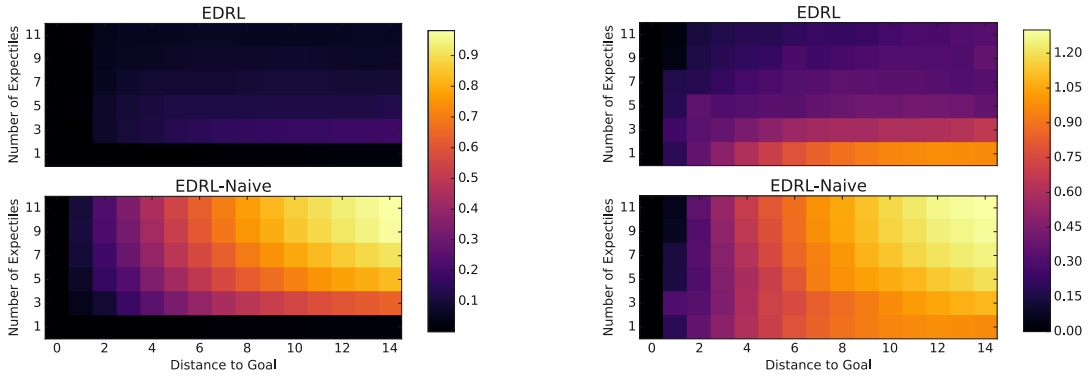
E.1. Full Environment Description

We give a full description of the N-chain environment used in Section 5.1 and illustrated in Figure 3 to investigate the properties of tabular EDRL and QDRL. This environment is a chain of length N with two possible actions at each state: (i) `forward`, which moves the agent right by one step with probability 0.95 and to x_0 with probability 0.05, and `backward`, which moves the agent to x_0 with probability 0.95 and one step to the right with probability 0.05. The reward is -1 when transitioning to the leftmost state, $+1$ when transitioning to the rightmost state, and zero elsewhere. Episodes begin in the leftmost state and terminate when the rightmost state is reached. The discount factor is $\gamma = 0.99$. For an N-Chain with length 15, we compute the return distribution of the optimal policy π^* which selects the `forward` action at each state. This environment formulation induces an increasingly multimodal return distribution under the policy as the distance from the goal state increases. We compute the ground truth start state expectiles from the empirical distribution of 1,000 Monte Carlo rollouts under the policy π^*. We set the learning rate to $\alpha = 0.05$, and perform 30,000 training steps.

E.2. Additional Experimental Results

In Section 5.1, we saw that the expectiles learned by EDRL-Naive on the N-Chain with length 15 collapsed, whereas the expectiles learnt by EDRL were reasonable approximations to the true expectiles of the return distribution. This resulted in lower average expectile estimation error with the latter expectiles, as described in Definition 4.5. In Figure 11, we supplement this by plotting Wasserstein distance between an imputed distribution for the learnt statistics and the true return distribution. This gives an alternate metric which additionally indicates how well the collection of learnt statistics summarises the full return distribution. Under this metric, we observe that increasing the number of expectiles always leads to improved performance under EDRL, whilst for EDRL-Naive, poor Wasserstein reconstruction error is observed for large numbers of expectiles and/or distance from the goal state.

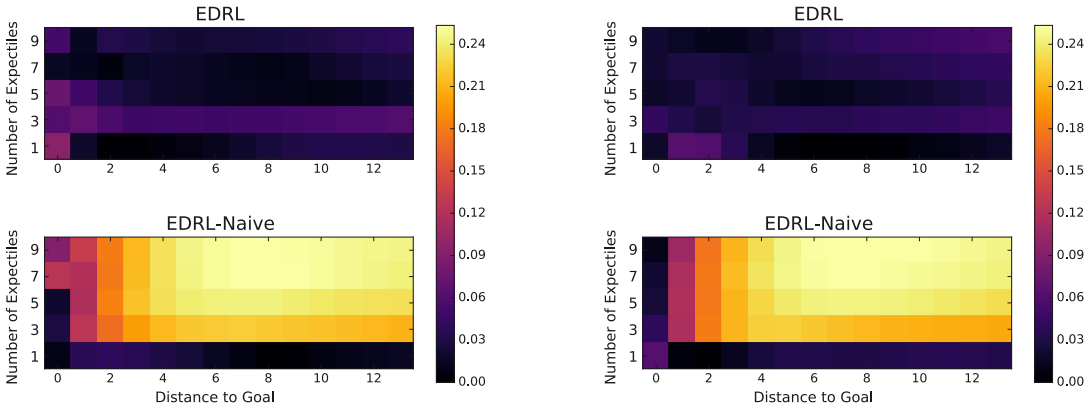
We also include results for N-chain environments with different reward distributions, observing qualitatively similar phenomena as those noted for Figure 11. Specifically, we use two additional variants of the reward distribution at the goal state: uniform and Gaussian. We plot average expectile error in Figure 12, and Wasserstein distance between imputed and true return distributions in Figure 13.



(a) Expectile estimation error

(b) 1-Wasserstein distance on imputed samples

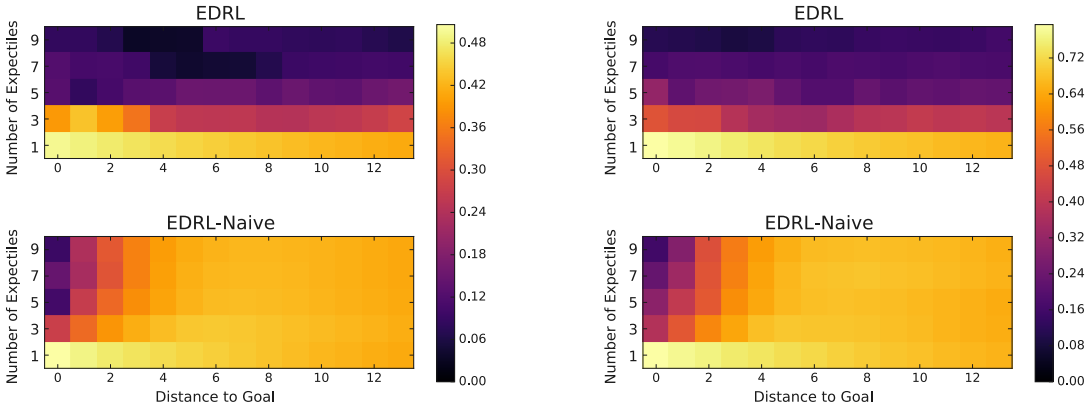
Figure 11. Expectile estimation error and 1-Wasserstein distance between imputed samples and the true return distribution for varying numbers of learned expectiles and different N-Chain lengths.



(a) Uniform (-1, 1) reward distribution.

(b) Standard Gaussian reward distribution.

Figure 12. Expectile estimation error for varying number of expectiles and different chain lengths. Different terminal reward distributions.



(a) Uniform (-1, 1) reward distribution.

(b) Standard Gaussian reward distribution.

Figure 13. 1-Wasserstein distance for varying number of expectiles and different chain lengths. Different terminal reward distributions.