Variance Control for Distributional Reinforcement Learning

Qi Kuang [*1] **Zhoufan Zhu** [*1] **Liwen Zhang** [1] **Fan Zhou** [1]

Abstract

Although distributional reinforcement learning (DRL) has been widely examined in the past few years, very few studies investigate the validity of the obtained Q-function estimator in the distributional setting. To fully understand how the approximation errors of the Q-function affect the whole training process, we do some error analysis and theoretically show how to reduce both the bias and the variance of the error terms. With this new understanding, we construct a new estimator *Quantiled Expansion Mean* (QEM) and introduce a new DRL algorithm (QEMRL) from the statistical perspective. We extensively evaluate our QEMRL algorithm on a variety of Atari and Mujoco benchmark tasks and demonstrate that QEMRL achieves significant improvement over baseline algorithms in terms of sample efficiency and convergence performance.

1. Introduction

Distributional Reinforcement Learning (DRL) algorithms have been shown to achieve state-of-art performance in RL benchmark tasks (Bellemare et al., 2017; Dabney et al., 2018b;a; Yang et al., 2019; Zhou et al., 2020; 2021). The core idea of DRL is to estimate the entire distribution of the future return instead of its expectation value, i.e. the Q-function, which captures the intrinsic uncertainty of the whole process in three folds: (i) the stochasticity of rewards, (ii) the indeterminacy of the policy, and (iii) the inherent randomness of transition dynamics. Existing DRL algorithms parameterize the return distribution in different ways, including categorical return atoms (Bellemare et al., 2017), expectiles (Rowland et al., 2019), particles (Nguyen-Tang et al., 2021), and quantiles (Dabney et al., 2018b;a). Among these works, the quantile-based algorithm is widely used due to its simplicity, efficiency of training, and flexibility in modeling the return distribution.

Although the existing quantile-based algorithms achieve remarkable empirical success, the approximated distribution still requires further understanding and investigation. One aspect is the crossing issue, namely, a violation of the monotonicity of the obtained quantile estimations. Zhou et al. (2020; 2021) solves this issue by enforcing the monotonicity of the estimated quantiles using some well-designed neural networks. However, these methods may suffer from some underestimation or overestimation issues. In other words, the estimated quantiles tend to be higher or lower than their true values. Considering this shortcoming, Luo et al. (2021) applies monotonic rational-quadratic splines to ensure monotonicity, but their algorithm is computationally expensive and hard to implement in large-scale tasks.

Another aspect is regard to the tail behavior of the return distribution. It is widely acknowledged that the precision of tail estimation highly depends on the frequency of tail observations (Koenker, 2005). Due to data sparsity, the quantile estimation is often unstable at the tails. To alleviate this instability, Kuznetsov et al. (2020) proposes to truncate the right tail of the approximated return distribution by discarding some topmost atoms. However, this approach lacks theoretical support and ignores the potentially useful information hidden in the tail.

The crossing issue and tail unrealization illustrate that there is a substantial gap between the quantile estimation and its true value. This finding reduces the reliability of the Q-function estimator obtained by quantile-based algorithms and inspires us to further minimize the difference between the estimated Q-function and its true value. In particular, the error associated with Q-function approximation can be decomposed into three parts:

$$
\begin{aligned}
\Delta &\equiv Q_\theta^\pi(x,a) - Q^\pi(x,a) = \mathbb{E}Z_\theta^\pi(x,a) - \mathbb{E}Z^\pi(x,a) \\
&= \underbrace{\mathbb{E}Z_\theta^\pi(x,a) - \mathbb{E}_{x'\sim\mathcal{D}}[R+\gamma Z_\theta^\pi(x',a')]}_{\text{Target Approximation Error } \mathcal{E}_1} \\
&\quad + \underbrace{\mathbb{E}_{x'\sim\mathcal{D}}[R+\gamma Z_\theta^\pi(x',a')] - \mathbb{E}_{x'\sim P}[R+\gamma Z_\theta^\pi(x',a')]}_{\text{Bellman operator Approximation Error } \mathcal{E}_2} \\
&\quad + \underbrace{\mathbb{E}_{x'\sim P}[R+\gamma Z_\theta^\pi(x',a')] - \mathbb{E}Z^\pi(x,a)}_{\text{Parametrization Induced Error } \mathcal{E}_3}, \qquad (1)
\end{aligned}
$$

[*]Equal contribution [1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China. Correspondence to: Fan Zhou <zhoufan@mail.shufe.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

where $Q^\pi(\cdot)$ is the true Q-function, $Q_\theta^\pi(\cdot)$ is the approximated Q-function, Z^π is the random variable with the true return distribution, Z_θ^π is the random variable with the approximated quantile function parameterized by a set of quantiles θ, $\mathcal{D}$ is the replay buffer, and P is the transition kernel. These errors can be attributed to different kinds of approximations in DRL (Rowland et al., 2018), including (i) parameterization and its associated projection operators, (ii) stochastic approximation of the Bellman operator, and (iii) gradient updates through quantile loss.

We elaborate on the properties of the three error terms in (1). $\mathcal{E}_1$ is derived from the target approximation in quantile loss. $\mathcal{E}_2$ is caused by the stochastic approximation of the Bellman operator. $\mathcal{E}_3$ results from the parametrization of quantiles and the corresponding projection operator. Among the three, $\mathcal{E}_3$ can be theoretically eliminated if the representation size is large enough, whereas $\mathcal{E}_1 + \mathcal{E}_2$ is inevitable in practice due to the batch-based optimization procedure. Therefore, controlling the variance $\mathrm{Var}(\mathcal{E}_1+\mathcal{E}_2)$ can significantly speed up the training convergence (see an illustrating example in Figure 1). Thus, one main target of this work is to reduce the two inevitable errors $\mathcal{E}_1$ and $\mathcal{E}_2$, and subsequently improve the existing DRL algorithms.

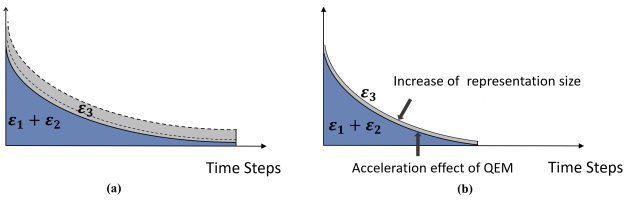


Figure 1. Error decay during training. (a) The parameterization-induced error $\mathcal{E}_3$ (grey areas) remains constant over time with a fixed representation size. The approximation errors $\mathcal{E}_1$ and $\mathcal{E}_2$ (blue areas) decrease slowly with time steps. (b) Increase the size of the representation (i.e., the number of quantiles), $\mathcal{E}_3$ can be theoretically eliminated. By applying the variance reduction technique QEM estimator, $\mathcal{E}_1 + \mathcal{E}_2$ can be quickly decreased, resulting in faster convergence of algorithms.

The contributions of this work are summarized as follows,

- We offer a rigorous investigation on the three error terms $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ in DRL, and find that the approximation errors result from the heteroskedasticity of quantile estimates, especially tail estimates.
- We borrow the idea from the Cornish-Fisher Expansion (Cornish & Fisher, 1938), and propose a statistically robust DRL algorithm, called QEMRL, to reduce the variance of the estimated Q-function.
- We show that QEMRL achieves a higher stability and a faster convergence rate from both theoretical and empirical perspectives.

2. Background

2.1. Reinforcement Learning

Consider a finite Markov Decision Process (MDP) $(\mathcal{X}, \mathcal{A}, P, \gamma, \mathcal{R})$, with a finite set of states $\mathcal{X}$, a finite set of actions $\mathcal{A}$, the transition kernel $P : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$, the discounted factor $\gamma \in [0, 1)$, and the bounded reward function $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathscr{P}([-R_{max}, R_{max}])$. At each timestep, an agent observes state $X_t \in \mathcal{X}$, takes an action $A_t \in \mathcal{A}$, transfers to the next state $X_{t+1} \sim P(\cdot \mid X_t, A_t)$, and receives a reward $R_t \sim \mathcal{R}(X_t, A_t)$. The state-action value function $Q^\pi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$ of a policy $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ is the expected discounted sum of rewards starting from x, taking an action a and following a policy π. $\mathscr{P}(\mathcal{X})$ denotes the set of probability distributions on a space $\mathcal{X}$.

The classic Bellman equation (Bellman, 1966) relates expected return at each state-action pair (x, a) to the expected returns at possible next states by:

$$Q^\pi(x, a) = \mathbb{E}_\pi \left[R_0 + \gamma Q^\pi (X_1, A_1) \mid X_0 = x, A_0 = a\right]. \quad (2)$$

In the learning task, Q-Learning (Watkins, 1989) employs a common way to obtain π^*, which is to find the unique fixed point $Q^* = Q^{\pi^*}$ of the Bellman optimality equation:

$$Q^*(x, a) = \mathbb{E}\left[R_0 + \gamma \max_{a' \in \mathcal{A}} Q^*\left(X_1, a'\right) \mid X_0 = x, A_0 = a\right].$$

2.2. Distributional Reinforcement Learning

Instead of directly estimating the expectation $Q^\pi(x, a)$, DRL focuses on estimating the distribution of the sum of discounted rewards $\eta_\pi(x, a) = \mathcal{D}(\sum_{t=0}^\infty \gamma^t R_t \mid X_0 = x, A_0 = a)$ to sufficiently capture the intrinsic randomness, where $\mathcal{D}$ extract the probability distribution of a random variable. In analogy with Equation (2), η_π satisfies the distributional Bellman equation (Bellemare et al., 2017) as follows,

$$\begin{aligned}\eta_\pi(x, a) &= \left(\mathcal{T}^\pi \eta_\pi\right)(x, a) \\ &= \mathbb{E}_\pi \left[(f_{\gamma,r})_\# \eta_\pi(X_1, A_1) \mid X_0 = x, A_0 = a\right]\end{aligned}$$

where $f_{\gamma,r} : \mathbb{R} \to \mathbb{R}$ is defined by $f_{\gamma,r}(x) = r + \gamma x$, and $(f_{\gamma,r})_\# \eta$ is the pushforward measure of η by $f_{\gamma,r}$. Note that η_π is the fixed point of distributional Bellman operator $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, i.e., $\mathcal{T}^\pi \eta_\pi = \eta_\pi$.

In general, the return distribution supports a wide range of possible returns and its shape can be quite complex. Moreover, the transition dynamics are usually unknown in practice, and thus the full computation of the distributional Bellman operator is usually either impossible or computationally infeasible. In the following subsections, we review two main categories of DRL algorithms relying on parametric approximations and projection operators.

2.2.1. CATEGORICAL DISTRIBUTIONAL RL

Categorical distributional RL (CDRL, Bellemare et al., 2017) represents the return distribution η with a categorical form $\eta(x,a) = \sum_{i=1}^{N} p_i(x,a)\delta_{z_i}$, where δ_z denotes the Dirac distribution at z. $z_1 \leq z_2 \leq \ldots \leq z_N$ are evenly spaced locations, and $\{p_i\}_{i=1}^{N}$ are the corresponding probabilities learned using the Bellman update,

$$\eta(x,a) \leftarrow \left(\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta\right)(x,a),$$

where $\Pi_{\mathcal{C}} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\{z_1, z_2 \ldots z_N\})$ is a *categorical projection* operator which ensures the return distribution supported only on $\{z_1, \ldots, z_N\}$. In practice, CDRL with $N = 51$ has been shown to achieve significant improvement in Atari games.

2.2.2. QUANTILED DISTRIBUTIONAL RL

Quantiled distributional RL (QDRL, Dabney et al., 2018b) represents the return distribution with a mixture of Diracs $\eta(x,a) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\theta_i(x,a)}$, where $\{\theta_i(x,a)\}_{i=1}^{N}$ are learnable parameters. The Bellman operator moves each atom location θ_i towards τ_i-th quantile of the target distribution $\eta'(x,a) := \mathcal{T}^{\pi}\eta(x,a)$, where $\tau_i = \frac{2i-1}{2N}$. The corresponding Bellman update form is:

$$\eta(x,a) \leftarrow \left(\Pi_{\mathcal{W}_1}\mathcal{T}^{\pi}\eta\right)(x,a),$$

where $\Pi_{\mathcal{W}_1} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ is a quantile projection operator defined by $\Pi_{\mathcal{W}_1}\mu = \frac{1}{N}\sum_{i=1}^{N}\delta_{F_{\mu}^{-1}(\tau_i)}$, and F_{μ} is the cumulative distribution function (CDF) of μ. $F_{\eta'}^{-1}(\tau)$ can be characterized as the minimizer of the quantile regression loss, while the atom locations θ can be updated by minimizing the following loss function

$$\mathcal{L}_{QR}(\theta;\eta',\tau) = \mathbb{E}_{Z\sim\eta'}\left(\left[\tau\mathbf{1}_{Z>\theta} + (1-\tau)\mathbf{1}_{Z\leq\theta}\right]|Z-\theta|\right). \tag{3}$$

3. Error Analysis of Distributional RL

As mentioned in Section 1, the parametrization induced error $\mathcal{E}_3$ in Equation (1) comes from quantile representation and its projection operator, which can be eliminated as $N \to \infty$. However, as illustrated in Figure 1, the approximation errors $\mathcal{E}_1$ and $\mathcal{E}_2$ are unavoidable in practice and a high variance $\mathrm{Var}(\mathcal{E}_1 + \mathcal{E}_2)$ may lead to unstable performance of DRL algorithms. Thus, in this section, we further study the three error terms $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$, and show why it is important to control them in practice.

3.1. Parametrization Induced Error

We first examine the convergence of both the expectation and the variance of the distributional Bellman operator $\mathcal{T}^{\pi}$. Then, we take parametric representation and projection operator into consideration.

Proposition 3.1. *Suppose there are two value distributions $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$, and random variables $Z_i^{k+1} \sim \mathcal{T}^{\pi}\nu_i, Z_i^k \sim \nu_i$. Then, we have*

$$\left\|\mathbb{E}Z_1^{k+1} - \mathbb{E}Z_2^{k+1}\right\|_{\infty} \leq \gamma\left\|\mathbb{E}Z_1^{k} - \mathbb{E}Z_2^{k}\right\|_{\infty}, \text{ and}$$
$$\left\|\mathrm{Var}Z_1^{k+1} - \mathrm{Var}Z_2^{k+1}\right\|_{\infty} \leq \gamma^2\left\|\mathrm{Var}Z_1^{k} - \mathrm{Var}Z_2^{k}\right\|_{\infty}.$$

Based on the fact that $\mathcal{T}^{\pi}$ is a γ-contraction in $\bar{d}_p$ metric (Bellemare et al., 2017), where $\bar{d}_p$ is the maximal form of the Wasserstein metric, Proposition 3.1 implies that $\mathcal{T}^{\pi}$ is a contraction for both the expectation and the variance. The two converge exponentially to their true values by iteratively applying the distributional Bellman operator (Sobel, 1982).

However, in practice, employing parametric representation for the return distribution leaves a theory-practice gap, which makes neither the expectation nor the variance converge to the true values. To better understand the bias in the Q-function approximation caused by the parametric representation, we introduce the concept of *mean-preserving* to describe the relationship between the expectations of the original distribution and the projected distribution:

Definition 3.2. (**Mean-preserving**) Let $\Pi_{\mathscr{F}} : \mathscr{P}(\mathbb{R}) \to \mathscr{F}$ be a projection operator that maps the space of probability distributions to the desired representation. Suppose there is a representation $\mathscr{F} \in \mathscr{P}(\mathbb{R})$ and its associated projection operator $\Pi_{\mathscr{F}}$ are mean-preserving if for any distribution $\nu \in \mathscr{F}$, the expectation of $\Pi_{\mathscr{F}}\nu$ is the same as that of ν.

For CDRL, a discussion of the *mean-preserving* property is given by Lyle et al. (2019) and Rowland et al. (2019). It can be shown that for any $\nu \in \mathscr{F}_{\mathcal{C}}$, where $\mathscr{F}_{\mathcal{C}}$ is a N-categorical representation, the projection $\Pi_{\mathcal{C}}$ preserves the distribution's expectation when its support is contained in the interval $[z_1, z_N]$. However, these practitioners usually employ a wide predefined interval for return which makes the projection operator typically overestimate the variance.

For QDRL, $\Pi_{\mathcal{W}_1}$ is not *mean-preserving*. Given any distribution $\nu \in \mathscr{F}_{\mathcal{W}_1}$, where $\mathscr{F}_{\mathcal{W}_1}$ is a N-quantile representation, there is no unique N-quantile distribution $\Pi_{\mathcal{W}_1}\nu$ in most cases, as the projection operator $\Pi_{\mathcal{W}_1}$ is not a non-expansion in 1-Wasserstein distance (See Appendix B for details). This means that the expectation, variance, and higher-order moments are not preserved. To make this concrete, a simple MDP example is used to illustrate the bias in the learned quantile estimates.

In Figure 2 (a), rewards R_1 and R_2 are randomly sampled from $\mathrm{Unif}(0,1)$ and $\mathrm{Unif}(1/N, 1+1/N)$ at states x_1 and x_2 respectively, and no rewards are received at x_0. Clearly, the true return distribution at state x_0 is the mixture $\frac{\gamma}{2}(R_1 + R_2)$, hence the $\frac{1}{2N}$-th quantile is $\frac{\gamma}{N}$. When using the QDRL algorithm with N quantile estimates, the approximated return distribution $\hat{\eta}(x_1,a) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\frac{2i-1}{2N}}$

and $\hat{\eta}(x_2, a) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\frac{2i+1}{2N}}$. In this case, the $\frac{1}{2N}$-th quantile of the approximated return distribution at state x_0 is $\frac{3\gamma}{2N}$, whereas the true value is $\frac{\gamma}{N}$. Moreover, for each $i = 1, \ldots, N$, the $\frac{2i-1}{2N}$-th quantile estimate at state x_0 is not equal to the true value.

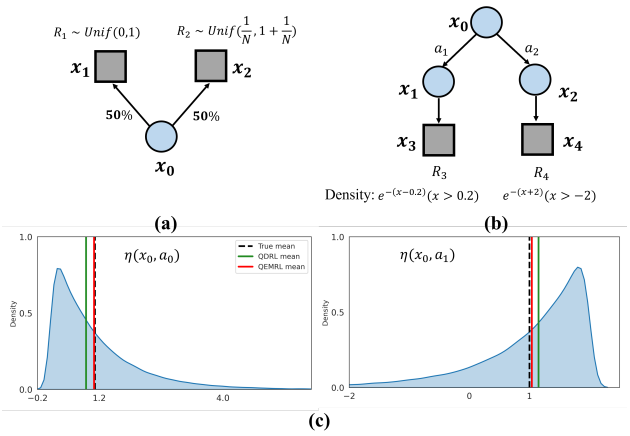


Figure 2. (a) Example MDP, with a single action, equal transition probability, an initial state x_0, and two terminal states x_1, x_2 where rewards are drawn from uniform. (b) 5-state MDP, with two actions at initial state x_0, deterministic transition, and stochastic rewards are exponential at terminal states x_3, x_4. (c) We show the true return distributions $\eta(x_0, a_1)$ and $\eta(x_0, a_2)$, and the expected returns estimated by QDRL and QEMRL.

These biased quantile estimates illustrated in Figure 2 (a) are caused by the use of quantile representation and its projection operator $\Pi_{\mathcal{W}_1}$. This undesirable property in turn affects the QDRL update, as the combined operator $\Pi_{\mathcal{W}_1}\mathcal{T}^\pi$ is in general not a non-expansion in $\bar{d}_p$, for $p \in [1, \infty)$ (Dabney et al., 2018b), which means that the learned quantile estimates may not converge to the true quantiles of the return distribution [1]. The projection operator $\Pi_{\mathcal{W}_1}$ is not *mean-preserving* which inevitably leads to bias in the expectation of return distribution when iteratively applying the projected Bellman operator $\Pi_{\mathcal{W}_1}\mathcal{T}^\pi$ during the training process, resulting in a deviation between the estimate and the true value of the Q-function in the end. We now derive an upper bound to quantify this deviation, i.e. $\mathcal{E}_3$.

Theorem 3.3. ***(Parameterization induced error bound)*** *Let $\Pi_{\mathcal{W}_1}$ be a projection operator onto evenly spaced quantiles τ_i's where each $\tau_i = \frac{2i-1}{2N}$ for $i = 1, \ldots, N$, and $\eta_k \in \mathscr{P}(\mathbb{R})$ be the return distribution of k-th iteration. Let random variables $Z_\theta^k \sim \Pi_{\mathcal{W}_1}\mathcal{T}^\pi\eta_k$ and $Z^k \sim \mathcal{T}^\pi\eta_k$. Assume that the distribution of the immediate reward is supported on $[-R_{max}, R_{max}]$, then we have*

$$\lim_{k\to\infty}\left\|\mathcal{E}_3^k\right\|_\infty = \lim_{k\to\infty}\left\|\mathbb{E}Z_\theta^k - \mathbb{E}Z^k\right\|_\infty \leq \frac{2R_{max}}{N(1-\gamma)},$$

where $\mathcal{E}_3^k$ is parametrization induced error at k-th iteration.

Theorem 3.3 implies that the convergence of expectation with projected Bellman operator $\Pi_{\mathcal{W}_1}\mathcal{T}^\pi$ cannot be guaranteed after quantile representation and its projection operator are applied. Note that the bound will tend to zero with $N \to \infty$, thus it is reasonable to use a relatively large representation size N to reduct $\mathcal{E}_3$ in practice.

3.2. Approximation Error

The other two types of errors $\mathcal{E}_1$ and $\mathcal{E}_2$, which determine the variance of the Q-function estimate, are accumulated during the training process by keeping encountering unseen state-action pairs. The target approximation error $\mathcal{E}_1$ affects action selections, while the Bellman operator approximation error $\mathcal{E}_2$ leads to the accumulated error of the Q-function estimate, which can be amplified by using the temporal difference updates (Sutton, 1988). The accumulated errors of the Q-function estimate with high uncertainty can make some certain states to be incorrectly estimated, leading to suboptimal policies and potentially divergent behaviors.

Using a simple 5-state MDP example, we illustrate how QDRL fails to learn an optimal policy due to a high variance of the approximation error, see Figure 2 (b). In this case, $\eta(x_0, a_1)$ and $\eta(x_0, a_2)$ follow exponential distributions, and the expectations of them are 1.2 and 1, respectively. We consider a tabular setting, which uniquely represents the approximated return distribution at each state-action pair. Figure 2 (c) demonstrates that in policy evaluation, QDRL inaccurately approximates the Q-function, as it underestimates the expectation of $\eta(x_0, a_1)$ and overestimates the other. This is caused by the poor capture of tail events, which results in high uncertainty in the Q-function estimate. Due to the high variance, QDRL fails to learn the optimal policy and chooses a non-optimal action a_2 at the initial state x_0. On the contrary, our proposed algorithm, QEMRL, employs a statistically robust estimator of the Q-function to reduce its variance, relieves the underestimation and overestimation issues, and ultimately allows for more efficient policy learning.

Different from previous QDRL studies that focus on exploiting the distribution information to further improve the model performance, this work highlights the importance of controlling the variance of the approximation error to obtain a more accurate estimate of the Q-function. More discussion about this is given in the following section.

[1] A recent study (Rowland et al., 2023) proves that QDRL update may have multiple fixed points, indicating quantiles may not converge to the truth. Despite this, Proposition 2 (Dabney et al., 2018b) concludes that the projected Bellman operator $\Pi_{\mathcal{W}_1}\mathcal{T}^\pi$ remains a contraction in $\bar{d}_\infty$. This implies that quantile convergence is guaranteed for all $p \in [1, \infty]$.

4. Quantiled Expansion Mean

This section introduces a novel variance reduction technique to estimate the Q-function. In traditional statistics, estimators with lower variance are considered to be more efficient. In RL, variance reduction is also an effective technique for achieving fast convergence in both policy-based and value-based RL algorithms, especially for large-scale tasks (Greensmith et al., 2004; Anschel et al., 2017). Motivated by these findings, we introduce QEM as an estimator that is more robust and has a lower variance than that of QDRL under the heteroskedasticity assumption. Furthermore, we demonstrate the potential benefits of QEM for the distribution approximation in DRL.

4.1. Heteroskedasticity of quantiles

In the context of quantile-based DRL, Q-function is the integral of the quantiles. To approximate this, QDRL employs a simple empirical mean (EM) estimator $\frac{1}{N}\Sigma_i \hat{q}(\tau_i)$, and it is natural to assume that the estimated quantile satisfies

$$\hat{q}(\tau) = q(\tau) + \varepsilon(\tau), \tag{4}$$

where $\varepsilon(\tau)$ is a zero-mean error. In this case, considering the crossing issue and the biased tail estimates, we assume that the variance of $\varepsilon(\tau)$ is non-constant and depends on τ, which is usually called heteroskedasticity in statistics.

For a direct understanding, we conduct a simple simulation using a Chain MDP to illustrate how QDRL can fail to fit the quantile function. As shown in Figure 3(b), QDRL fits well in the peak area but struggles at the bottom and the tail. Moreover, the non-monotonicity of the quantile estimates in the poorly fitted areas is more severe than the others. As the deviations of the quantile estimates from the truths is significantly larger in the low probability region and the tail, we can make the heteroskedasticity assumption in this case. This phenomenon can be explained since samples near the bottom and the tail are less likely to be drawn. In real-world situations, multimodal distributions are commonly encountered and the heteroskedasticity problem may result in imprecise distribution approximations and consequently poor Q-function approximations. In the next part, we will discuss how to enhance the stability of the Q-function estimate.

4.2. Cornish-Fisher Expansion

It is well-known that quantile can be expressed by the Cornish-Fisher Expansion (CFE, Cornish & Fisher, 1938):

$$\begin{aligned} q(\tau) &= \mu + \sigma x'_\tau, \\ x'_\tau &= z_\tau + (z_\tau^2 - 1)\frac{s}{6} + (z_\tau^3 - 3z_\tau)\frac{k}{24} + \cdots, \end{aligned} \tag{5}$$

where z_τ is the τ-th quantile of the standard normal distribution, μ is the mean, σ is the standard deviation, s and k are the skewness and kurtosis of the interested distribution, and the remaining terms in the ellipsis are higher-order moments (See Appendix C for more details). The CFE theoretically determines the distribution with known moments and is widely used in financial studies. Recently, Zhang & Zhu (2023) employ CFE to estimate higher-order moments of financial time series data, which are not directly observable. Our method utilizes a truncated version of CFE framework and employs a linear regression model to construct efficient estimators for distribution moments based on known quantiles. Consequently, we apply this approach within the context of quantile-based DRL.

To be more specific, we plug in the estimate $\hat{q}(\tau)$ of the the τ-th quantile to Equation (5) and expand it by the first order:

$$\hat{q}(\tau) = m_1 + \omega_1(\tau) + \varepsilon(\tau), \tag{6}$$

where m_1 is the mean (say, 1-th moment) of the return distribution, i.e., the Q-function, and $\omega_1(\tau)$ is the remaining term associated with the higher-order ($>$ 1-th) moments. If $\omega_1(\tau)$ is negligible, m_1 can be estimated by averaging the N quantile estimates in QDRL.

When the estimated quantile is expanded to the second order, we particularly have the following representation:

$$\hat{q}(\tau) = m_1 + z_\tau\sqrt{m_2} + \sqrt{m_2}\omega_2(\tau) + \varepsilon(\tau), \tag{7}$$

where $\omega_2(\tau)$ is the remaining term associated with the higher-order ($>$ 2-th) moments. Assume that $\omega_2(\tau)$ is negligible, we can derive a regression model by plugging in the N quantile estimates, such that

$$\begin{pmatrix} \hat{q}(\tau_1) \\ \hat{q}(\tau_2) \\ \vdots \\ \hat{q}(\tau_N) \end{pmatrix} = \begin{pmatrix} 1 & z_{\tau_1} \\ 1 & z_{\tau_2} \\ \vdots & \vdots \\ 1 & z_{\tau_N} \end{pmatrix} \begin{pmatrix} m_1 \\ \sqrt{m_2} \end{pmatrix} + \begin{pmatrix} \varepsilon(\tau_1) \\ \varepsilon(\tau_2) \\ \vdots \\ \varepsilon(\tau_N) \end{pmatrix}. \tag{8}$$

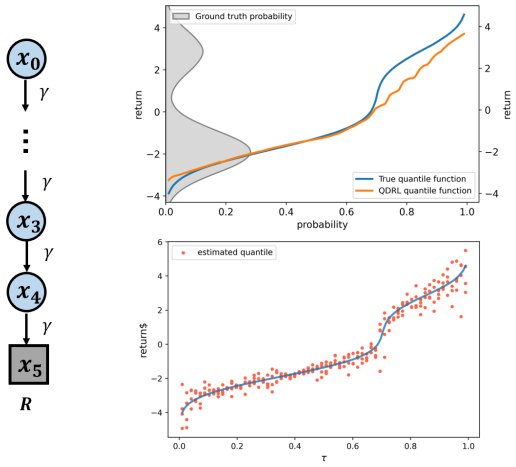
The higher-order expansions can be conducted in the same manner. Note that the remaining term is omitted for constructing a regression model, and a more in-depth analysis of the remaining term is available in Appendix C.2.

For notation simplicity, we rewrite (8) in a matrix form,

$$\hat{\boldsymbol{Q}} = \mathbf{X}_2\boldsymbol{M}_2 + \mathcal{E}, \tag{9}$$

where $\hat{\boldsymbol{Q}} \in \mathbb{R}^N$ is the vector of estimated quantiles, $\mathbf{X}_2 \in \mathbb{R}^{N\times 2}$ and $\boldsymbol{M}_2 \in \mathbb{R}^2$ are the design matrix and the moments respectively, and $\mathcal{E}$ is the vector of error terms.

For this bivariate regression model (9), the traditional ordinary least squares method (OLS) can be used to estimate $\boldsymbol{M}_2 = (m_1, \sqrt{m_2})'$ when the variances of the errors are invariant across different quantile locations, also known as



(a) (b)

Figure 3. (a) Chain MDP, with six states, one action, $\gamma = 0.99$ and gaussian mixture reward distribution at terminal state x_5. (b) True quantile function (top) and QDRL quantile function at state x_0 after 10K steps iterate. Scatter diagram (bottom) of approximated quantile from training process.

the homoscedasticity assumption. The estimator $\hat{m}_1$ is denoted as Quantiled Expansion Mean (QEM) in this work. However, since the homoscedasticity assumption required by OLS is always violated in real cases, we may consider using the weighted ordinary least squares method (WLS) instead. Under the normality assumption, the following results tell that the WLS estimator $\hat{m}_1$ has a lower variance than the direct empirical mean.

Lemma 4.1. *Consider the linear regression model $\hat{\boldsymbol{Q}} = \mathbf{X}_2\boldsymbol{M}_2 + \mathcal{E}$, $\mathcal{E}$ is distributed on $\mathcal{N}(\mathbf{0}, \sigma^2 V)$, where $V = diag(v_1, v_2, \cdots, v_N), v_i \geq 1, i = 1, \cdots, N$, and we set noise variance $\sigma^2 = 1$ without loss of generality. The WLS estimator is*

$$\widehat{\boldsymbol{M}}_2 = (\mathbf{X}_2^\top V^{-1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top V^{-1}\hat{\boldsymbol{Q}}, \tag{10}$$

and the QEM estimator $\hat{m}_1$ is the first component of $\widehat{\boldsymbol{M}}_2$.

Remark: Note that it is impossible to determine the weight matrix V for each state-action pair in practice. Hence, we focus on capturing the relatively high variance in the tail, specifically in the range of $\tau \in (0, 0.1] \cup [0.9, 1)$. To achieve this, we use a constant v_i, which is set to a value greater than 1 in the tail and equal to 1 in the rest. v_i is treated as a hyperparameter to be tuned in practice (See Appendix E).

With Lemma 4.1, the reduction of variance can be guaranteed by the following Proposition 4.2. Throughout the training process, heteroskedasticity is inevitable, and thus the QEM estimator always exhibits a lower variance than the standard EM estimator $\hat{m}_1^* = \frac{1}{N}\sum_{i=1}^N \hat{q}(\tau_i)$.

Proposition 4.2. *Suppose the noise ε_i independently follows $\mathcal{N}(0, v_i)$ where $v_i \geq 1$ for $i = 1, \cdots, N$, then,*

(i) In the homoskedastic case where $v_i = 1$ for $i = 1, \ldots N$, the empirical mean estimator $\hat{m}_1^*$ has a lower variance, $\mathrm{Var}(\hat{m}_1^*) < \mathrm{Var}(\hat{m}_1)$ *;*

(ii) In the heteroskedastic case where v_i's are not eaqul, the QEM estimator $\hat{m}_1$ achieves a lower variance, i.e. $\mathrm{Var}(\hat{m}_1) < \mathrm{Var}(\hat{m}_1^*)$*, if and only if* $\bar{v}^2 - 1 - 1/(\frac{(\sum_i v_i \sum_i v_i z_{\tau_i}^2)}{(\sum_i v_i z_{\tau_i})^2} - 1) > 0$*, where* $\bar{v} = \frac{1}{N}\sum_i v_i$*. This inequality holds when* $z_{\tau_i} = -z_{\tau_{N-i}}$*, which can be guaranteed in QDRL.*

We also try to explore the potential benefits of the variance reduction technique QEM in improving the approximation accuracy. The Q-function estimate with higher variance can lead to noisy policy gradients in policy-based algorithms (Fujimoto et al., 2018) and prevent selection optimal actions in value-based algorithms (Anschel et al., 2017). These issues can slow down the learning process and negatively impact the algorithm performance. By the following theorem, we are able to show that QEM can reduce the variance and thus improve the approximation performance.

Theorem 4.3. *Consider the policy $\hat{\pi}$ that is learned policy, and denote the optimal policy to be π_{opt}, $\alpha = \max_{x'} D_{TV}(\hat{\pi}(\cdot \mid x') \| \pi_{opt}(\cdot \mid x'))$, and $n(x,a) = |\mathcal{D}|$. For all $\delta \in \mathbb{R}$, with probability at least $1-\delta$, for any $\eta(x,a) \in \mathscr{P}(\mathbb{R})$, and all $(x,a) \in \mathcal{D}$,*

$$\left\| F_{\hat{\mathcal{T}}^{\hat{\pi}}\eta(x,a)} - F_{\mathcal{T}^{\pi_{opt}}\eta(x,a)} \right\|_\infty \leq (\alpha+1)\sqrt{\frac{2|\mathcal{X}|}{n(x,a)} \log \frac{4|\mathcal{X}||\mathcal{A}|}{\delta}}.$$

Theorem 4.3 indicates that a lower concentration bound can be obtained with a smaller α value. The decrease in α can be attributed to the benefits of QEM. Specifically, QEM helps to decrease the perturbations on the Q-function and reduce the variance of the policy gradients, which allows for faster convergence of the policy training and a more accurate distribution approximation. To conclude, QEM relieves the error accumulation within the Q-function update, improves the estimation accuracy, reduces the risk of underestimation and overestimation, and thus ultimately enhances the stability of the whole training process.

5. Experimental Results

In this section, we do some empirical studies to demonstrate the advantage of our QEMRL method. First, a simple tabular experiment is conducted to validate some of the theoretical results presented in Sections 3 and 4. Then we apply the proposed QEMRL update strategy in Algorithm 1 to both the DQN-style and SAC-style DRL algorithms, which are evaluated on the Atari and MuJoCo environments. The detailed architectures of these methods and the hyperparameter selections can be found in Appendix D, and the additional experimental results are included in Appendix E.

Algorithm 1 QEMRL update algorithm

Require: Quantile estimates $\hat{q}_i(x, a)$ for each (x, a)
Collect sample (x, a, r, x')
Compute distributional Bellman target
Compute $Q(x', a)$ using Equation (10)
if policy evaluation **then**
  $a^* \sim \pi(\cdot|x')$
else if Q-Learning **then**
  $a^* \leftarrow \arg\max_a Q(x', a)$
end if
Scale samples $\hat{q}_i^*(x', a^*) \leftarrow r + \gamma \hat{q}_i(x', a^*), \forall i.$
Compute quantile loss
Update estimated quantiles $\hat{q}_i(x, a)$ by computing the gradients for each $i = 1, \ldots, N$, $\nabla_{\hat{q}_i(x,a)} \sum_{i=1}^{N} \mathcal{L}_{QR}(\hat{q}_i(x, a); \frac{1}{N} \sum_{j=1}^{N} \delta_{\hat{q}_j^*(x',a^*)}, \tau_i)$.

In this work, we implement QEM using a 4-th order expansion that includes mean, variance, skewness, and kurtosis in this work. The effects of a higher-order expansion on model estimation are discussed in Appendix C.1. Intuitively, including more terms in the expansion improves the estimation accuracy of quantiles, but the overfitting risk and the computational cost are also increased. Hence, there is a trade-off between explainability and learning efficiency. We evaluate different expansion orders using the R^2 statistic, which measures the goodness of model fitting. The simulation results (Figure 9) show that a 4-th order expansion seems to be the optimal choice while a higher-order (> 4-th) expansion does not show a significant increase in R^2.

5.1. A Tabular Example

FrozenLake (Brockman et al., 2016) is a classic benchmark problem for Q-learning control with high stochasticity and sparse rewards, in which an agent controls the movement of a character in an $n \times n$ grid world. As shown in Figure 4 with a FrozenLake-4×4 task, "S" is the starting point, "H" is the hole that terminates the game, "G" is the goal state with a reward of 1. All the blue grids stand for the frozen surface where the agent can slide to adjacent grids based on some underlying unknown probabilities when taking a certain movement direction. The reward received by the agent is always zero unless the goal state is reached.

We first approximate the return distribution under the optimal policy π^*, which can be realized using the value iteration approach. To be specific, we start from the "S" state and perform 1K Monte-Carlo (MC) rollouts. An empirical distribution can be obtained by summarizing all these recording trajectories. With the approximation of the distribution, we can draw a curve of quantile estimates shown in Figure 5. Both QEMRL and QDRL were run for 150K training steps and the ϵ-greedy exploration strategy is applied in the first 1K steps. For both methods, we set the total number of quantiles to be $N = 128$.

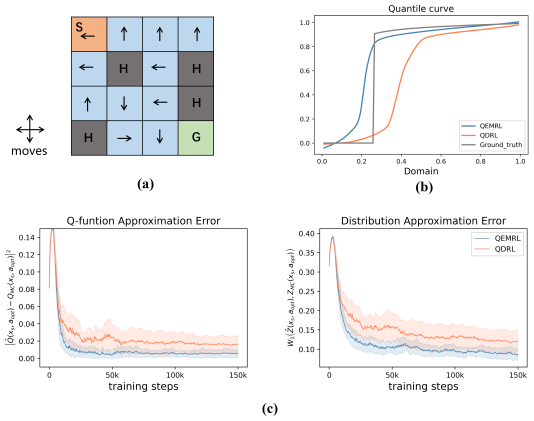


Figure 4. (a) The optimal direction of movement at each grid. (b) Quantile estimates by MC, QDRL, and QEMRL at the start state. (c) Approximation errors of Q-function estimate and distribution approximation error of QEMRL and QDRL (results are averaged over 10 random seeds).

Although both QEMRL and QDRL can eventually find the optimal movement at the start state, their approximations of the return distribution are quite different. Figure 4 (b) visualizes the approximation errors of the Q-function and the distribution for QEMRL and QDRL with respect to the number of training steps. The Q-function estimates of QEMRL converge correctly in average, whereas the estimates of QDRL do not converge exactly to the truth. A similar pattern can also be found when it comes to the distribution approximation error. Besides, the reduction of variance by using QEM can be verified by the fact that the curves of QEMRL are more stable and decline faster. In Figure 4 (c), we show that the distribution at the start state estimated by QEMRL is eventually closer to the ground truth.

5.2. Evaluation on MuJoCo and Atari 2600

We do some experiments using the MuJoCo benchmark to further verify the analysis results in Section 4. Our implementation is based on the Distributional Soft Actor-Critic (DSAC, Ma et al., 2020) algorithm, which is a distributional version of SAC. Figure 5 demonstrate that both DSAC and QEM-DSAC significantly outperform the baseline SAC. Among the two, QEM-DSAC performs better than DSAC and the learning curves are more stable, which demonstrates that QEM-DSAC can achieve a higher sample efficiency.

We also do some comparison between QEM and the baseline method QR-DQN on the Atari 2600 platform. Figure 8 plots the final results of these two algorithms in six Atari games. At the early training stage, QEM-DQN exhibits significant gain in sampling efficiency, resulting in faster convergence and better performance.

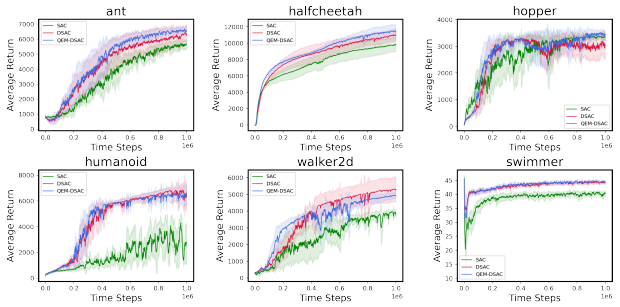


Figure 5. Learning curves of SAC, DSAC, and QEM-DSAC across six MuJoCo games. Each curve is averaged over 5 random seeds and shaded by their confidence intervals.

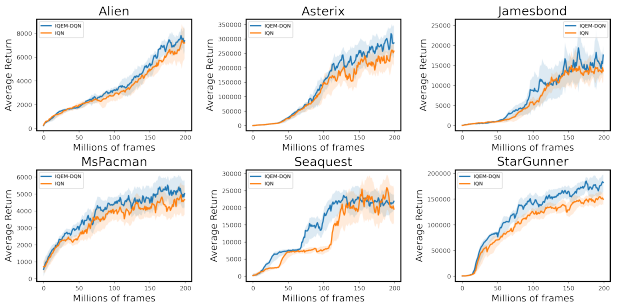


Figure 7. Learning curves of IQN and IQEM-DQN across six Atari games. Each curve is averaged over 3 random seeds and shaded by their confidence intervals.

Extension to IQN. Some great efforts have been made by the community of DRL to more precisely parameterize the entire distribution with a limited number of quantile locations. One notable example is the introduction of Implicit Quantile Networks (IQN, Dabney et al., 2018a), which tries to recover the continuous map of the entire quantile curve by sampling a different set of quantile values from a uniform distribution $\mathrm{Unif}(0, 1)$ each time.

Our method can also be applied to IQN as it uses the EM approach to estimate the Q-function. It is noted that the design matrix $\mathbf{X}$ must be updated after re-sampling all the quantile fractions at each training step. Moreover, one important sufficient condition $z_{\tau_i} = -z_{\tau_{N-i}}$ which ensures the reduction of variance does not hold in the IQN case as τ's are sampled from a uniform distribution. However, according to the simulation results in Table 4, the variance reduction still remains valid in practice. In this case, all the baseline methods are modified to the IQN version. As Figure 6 and Figure 7 demonstrate, QEM can achieve some performance gain in most scenarios and the convergence speeds can be slightly increased.

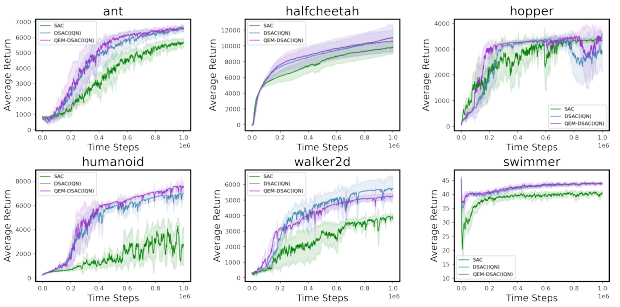


Figure 6. Learning curves of SAC, DSAC (IQN), and QEM-DSAC (IQN) across six MuJoCo games. Each curve is averaged over 5 random seeds and shaded by their confidence intervals.

5.3. Exploration

Since QEM also provides an estimate of the variance, we may consider using it to develop an efficient exploration strategy. In some recent study studies, to more sufficiently utilize the distribution information, Mavrin et al. (2019) proposes a novel exploration strategy, Decaying Left Truncated Variance (DLTV) by using the left truncated variance of the estimated distribution as a bonus term to encourage exploration in unknown states. The optimal action a^* at state x is selected according to $a^* = \arg\max_{a'} \left(Q(x, a') + c_t \sqrt{\sigma_+^2} \right)$, where c_t is a decay factor to suppress the intrinsic uncertainty, and σ_+^2 denotes the estimation of variance. Although DLTV is effective, the validity of the computed truncation lacks a theoretical guarantee. In this work, we follow the idea of DLTV and examine the model performance by using either the variance estimate obtained by QEM or the original DLTV estimation in some hard-explored games. As Figure 8 shows, by using QEM, the exploration efficiency is significantly improved compared to QR-DQN+DLTV since QEM enhances the accuracy of the quantile estimates and thus the accuracy of the distribution variance.

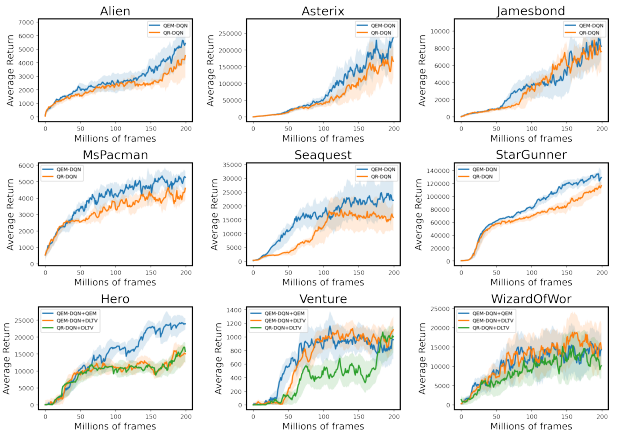


Figure 8. Learning curves (top and middle) of QR-DQN and QEM-DQN across six Atari games. Learning curves (bottom) of QR-DQN and QEM-DQN with exploration across three games.

6. Conclusion and Discussion

In this work, we systematically study the three error terms associated with the Q-function estimate and propose a novel DRL algorithm QEMRL, which can be applied to any quantile-based DRL algorithm regardless of whether the quantile locations are fixed or not. We found that a more robust estimate of the Q-function can improve the distribution approximation and speed up the algorithm convergence. We can also utilize the more precise estimate of the distribution variance to optimize the existing exploration strategy.

Finally, there are some open questions we would like to have further discussions here.

Improving the estimation of weight matrix V. The challenge of estimating the weight matrix V was recognized from the outset of the method proposal since it is unlikely to know the exact value of V in practice. In this work, we treat V as a predefined value that can be tuned, taking into account the computational cost of estimating it across all state-action pairs and time steps. As for future work, we believe a robust and easy-to-implement estimation of weight matrix V is necessary. Given that the variance of quantile estimation errors varies with state-action pairs and algorithm iterations, we consider two approaches for future investigation. The first approach considers a decay value of v_i instead of the constant. It is worth noting that the variance of poorly estimated quantiles tends to decrease gradually as the number of training samples increases, which motivates us to decrease the value of v_i as training epochs increase. The second approach involves assigning different values of v_i to different state-action pairs. Ideas from the exploration field, specifically the count-based method (Ostrovski et al., 2017), can be borrowed to measure the novelty of state-action pairs. Accordingly, for familiar state-action pairs, a smaller value of v_i should be assigned, while unfamiliar pairs should be assigned a larger value of v_i.

Statistical variance reduction. Our variance reduction method is based on a statistical modeling perspective, and the core insight of our method is that performance might be improved through more careful use of the quantiles to construct a Q-function estimator. While alternative ensembling methods can be directly applied to DRL to reduce the uncertainty in Q-function estimator, commonly used in existing works (Osband et al., 2016; Anschel et al., 2017), it undoubtedly increases model complexity. In this work, we transform the Q value estimation into a linear regression problem, where the Q value is the coefficient of the regression model. In this way, we can leverage the weighted least squares (WLS) method to effectively capture the heteroscedasticity of quantiles and obtain a more efficient and robust Q-function estimator.

Acknowledgements

We thank anonymous reviewers for valuable and constructive feedback on an early version of this manuscript. This work is supported by National Social Science Foundation of China (Grant No.22BTJ031) and Postgraduate Innovation Foundation of SUFE. Dr. Fan Zhou's work is supported by National Natural Science Foundation of China (12001356), Shanghai Sailing Program (20YF1412300), "Chenguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission, Open Research Projects of Zhejiang Lab (NO.2022RC0AB06), Shanghai Research Center for Data Science and Decision Technology, Innovative Research Team of Shanghai University of Finance and Economics.

References

Anschel, O., Baram, N., and Shimkin, N. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 176–185. PMLR, 2017.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458. PMLR, 2017.

Bellman, R. Dynamic programming. *Science*, 153(3731): 34–37, 1966.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Cornish, E. A. and Fisher, R. A. Moments and cumulants in the specification of distributions. *Revue de l'Institut international de Statistique*, pp. 307–320, 1938.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 1096–1105, 2018a.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018b.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pp. 1587–1596. PMLR, 2018.

Greensmith, E., Bartlett, P. L., and Baxter, J. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004.

Hsu, D., Kakade, S. M., and Zhang, T. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 2011.

Koenker. *Quantile regression*. Cambridge University Press, 2005.

Kuznetsov, A., Shvechikov, P., Grishin, A., and Vetrov, D. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, pp. 5556–5566. PMLR, 2020.

Luo, Y., Liu, G., Duan, H., Schulte, O., and Poupart, P. Distributional reinforcement learning with monotonic splines. In *International Conference on Learning Representations*, 2021.

Lyle, C., Bellemare, M. G., and Castro, P. S. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4504–4511, 2019.

Ma, X., Xia, L., Zhou, Z., Yang, J., and Zhao, Q. Dsac: distributional soft actor critic for risk-sensitive reinforcement learning. *arXiv preprint arXiv:2004.14547*, 2020.

Mavrin, B., Yao, H., Kong, L., Wu, K., and Yu, Y. Distributional reinforcement learning for efficient exploration. In *International Conference on Machine Learning*, pp. 4424–4434, 2019.

Nguyen-Tang, T., Gupta, S., and Venkatesh, S. Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9144–9152, 2021.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. *Advances in Neural Information Processing Systems*, 29, 2016.

Ostrovski, G., Bellemare, M. G., Oord, A., and Munos, R. Count-based exploration with neural density models. In *International Conference on Machine Learning*, pp. 2721–2730. PMLR, 2017.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., and Teh, Y. W. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M. G., and Dabney, W. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 5528–5536. PMLR, 2019.

Rowland, M., Munos, R., Azar, M. G., Tang, Y., Ostrovski, G., Harutyunyan, A., Tuyls, K., Bellemare, M. G., and Dabney, W. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.

Sobel, M. J. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Watkins, C. J. C. H. Learning from delayed rewards. *PhD thesis*, 1989.

Yang, D., Zhao, L., Lin, Z., Qin, T., Bian, J., and Liu, T.-Y. Fully parameterized quantile function for distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 6193–6202, 2019.

Zhang, N. and Zhu, K. Quantiled conditional variance, skewness, and kurtosis by cornish-fisher expansion. *arXiv preprint arXiv:2302.06799*, 2023.

Zhou, F., Wang, J., and Feng, X. Non-crossing quantile regression for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 15909–15919, 2020.

Zhou, F., Zhu, Z., Kuang, Q., and Zhang, L. Non-decreasing quantile function network with efficient exploration for distributional reinforcement learning. *International Joint Conference on Artificial Intelligence*, pp. 3455–3461, 2021.

A. Projection Operator

A.1. Categorical projection operator

CDRL algorithm uses a categorical projection operator $\Pi_{\mathcal{C}} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\{z_1, \dots, z_N\})$ to restrict approximated distributions to the parametric family of the form $\mathscr{F}_{\mathcal{C}} := \left\{ \sum_{i=1}^N p_i \delta_{z_i} \mid \sum_{i=1}^N p_i = 1, p_i \geq 0 \right\} \subseteq \mathscr{P}(\mathbb{R})$, where $z_1 < \cdots < z_N$ are evenly spaced, fixed supports. The operator $\Pi_{\mathcal{C}}$ is defined for a single Dirac delta as

$$\Pi_{\mathcal{C}}(\delta_w) = \begin{cases} \delta_{z_1} & w \leq z_1 \\ \frac{w - z_{i+1}}{z_i - z_{i+1}} \delta_{z_i} + \frac{z_i - w}{z_i - z_{i+1}} \delta_{i+1} & z_i \leq w \leq z_{i+1} \\ \delta_{z_N} & w \geq z_N. \end{cases}$$

A.2. Quantile projection operator

QDRL algorithm uses a quantile projection operator $\Pi_{\mathcal{W}_1} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ to restrict approximated distributions to the parametric family of the form $\mathscr{F}_{\mathcal{W}_1} := \left\{ \frac{1}{N} \sum_{i=1}^N \delta_{z_i} \mid z_{1:N} \in \mathbb{R}^N \right\} \subseteq \mathscr{P}(\mathbb{R})$. The operator Π_{W_1} is defined as

$$\Pi_{\mathcal{W}_1}(\mu) = \frac{1}{N} \sum_{k=1}^N \delta_{F_\mu^{-1}(\tau_i)},$$

where $\tau_i = \frac{2i-1}{2N}$, and F_μ is the CDF of μ. The midpoint $\frac{2i-1}{2N}$ of the interval $[\frac{i-1}{N}, \frac{i}{N}]$ minimizes the 1-Wasserstein distance $W_1(\mu, \Pi_{W_1}\mu)$ between the distribution, μ, and its projection $\Pi_{W_1}\mu$ (a N-quantile distribution with evenly spaced τ_i), as demonstrated in Lemma 2 (Dabney et al., 2018b).

B. Proofs

In this section, we provide the proofs of the theorems discussed in the main manuscript.

B.1. Proof of Section 3

Proposition B.1. *Suppose there are value distributions $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$, and random variables $Z_i^{k+1} \sim \mathcal{T}^\pi \nu_i, Z_i^k \sim \nu_i$. Then, we have*

$$\left\| \mathbb{E}Z_1^{k+1} - \mathbb{E}Z_2^{k+1} \right\|_\infty \leq \gamma \left\| \mathbb{E}Z_1^k - \mathbb{E}Z_2^k \right\|_\infty, \text{ and}$$
$$\left\| \mathrm{Var}Z_1^{k+1} - \mathrm{Var}Z_2^{k+1} \right\|_\infty \leq \gamma^2 \left\| \mathrm{Var}Z_1^k - \mathrm{Var}Z_2^k \right\|_\infty.$$

Proof. The first statement can be proved using the exchange of $\mathbb{E}\mathcal{T}^\pi = \mathcal{T}^\pi \mathbb{E}$. By independence of R and $P^\pi Z_i$, where P^π is the transition operator, we have

$$Z_i^{k+1}(x,a) \stackrel{D}{:=} R(x,a) + \gamma P^\pi Z_i^k(x,a)$$
$$\mathrm{Var}(Z_i^{k+1}(x,a)) = \mathrm{Var}(R(x,a)) + \gamma^2 \mathrm{Var}\left(P^\pi Z_i^k(x,a)\right).$$

Thus, we have

$$\begin{aligned} &\left\| \mathrm{Var}Z_1^{k+1} - \mathrm{Var}Z_2^{k+1} \right\|_\infty \\ &\quad = \sup_{x,a} \left| \mathrm{Var}Z_1^{k+1}(x,a) - \mathrm{Var}Z_2^{k+1}(x,a) \right| \\ &\quad = \sup_{x,a} \gamma^2 \left| \mathrm{Var}\left(P^\pi Z_1^k(x,a)\right) - \mathrm{Var}\left(P^\pi Z_2^k(x,a)\right) \right| \\ &\quad = \sup_{x,a} \gamma^2 \left| \mathbb{E}\left[\mathrm{Var}\left(Z_1^k(X',A')\right) - \mathrm{Var}\left(Z_2^k(X',A')\right) \right] \right| \\ &\quad \leq \sup_{x',a'} \gamma^2 \left| \mathrm{Var}\left(Z_1^k(x',a')\right) - \mathrm{Var}\left(Z_2^k(x',a')\right) \right| \\ &\quad \leq \gamma^2 \left\| \mathrm{Var}Z_1^k - \mathrm{Var}Z_2^k \right\|_\infty. \end{aligned}$$

□

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

$$\Pi_{W_1}\mu_i = \frac{1}{K}\sum_{k=1}^{K}\delta_{F^{-1}_{\mu_i}(\tau_k)},$$

for all $\mu_i \in \mathscr{P}(\mathbb{R})$, where $F^{-1}_{\mu_i}$ is the inverse CDF of μ_i. Let random variable $X \sim \mu_1$, $X^2 \sim \mu_2$, and $\eta_1, \eta_2 \in \mathscr{P}(\mathbb{R})$. Suppose immediate reward distributions supported on $[-R_{max}, R_{max}]$. Then, we have:
(i) $W_1\left(\Pi_{W_1}\mu_1, \mu_1\right) \leq \frac{2R_{max}}{K(1-\gamma)}$;
(ii) $W_1\left(\Pi_{W_1}\eta_1, \Pi_{W_1}\eta_2\right) \leq W_1\left(\eta_1, \eta_2\right) + \frac{4R_{max}}{K(1-\gamma)}$;
(iii) $W_1\left(\Pi_{W_1}\mu_2, \mu_2\right) \leq \frac{R^2_{max}}{K(1-\gamma)}$.

Proof. For proving (i), let $F^{-1}_{\mu_1}$ be the inverse CDF of μ_1. We have

$$\begin{aligned}
W_1\left(\Pi_{W_1}\mu_1, \mu_1\right) &= \sum_{i=0}^{K-1}\frac{1}{K}\int_{F^{-1}_{\mu_1}(\frac{i}{K})}^{F^{-1}_{\mu_1}(\frac{i+1}{K})} \left|x - F^{-1}_{\mu_1}(\frac{2i+1}{2K})\right| \quad \mu_1(dx) \\
&\leq \frac{1}{K}\left(F^{-1}_{\mu_1}(1) - F^{-1}_{\mu_1}(0)\right) \quad \text{(return distribution } \mu_1 \text{ is bounded on } [-\frac{R_{max}}{1-\gamma}, \frac{R_{max}}{1-\gamma}]) \\
&= \frac{2R_{max}}{K(1-\gamma)}.
\end{aligned}$$

For proving (ii), using the triangle inequality and statement (i):

$$\begin{aligned}
W_1\left(\Pi_{W_1}\eta_1, \Pi_{W_1}\eta_2\right) &\leq W_1\left(\Pi_{W_1}\eta_1, \eta_1\right) + W_1\left(\eta_1, \eta_2\right) + W_1\left(\eta_2, \Pi_{W_1}\eta_2\right) \\
&\leq W_1\left(\eta_1, \eta_2\right) + \frac{4R_{max}}{K(1-\gamma)}.
\end{aligned}$$

(ii) implies the fact that the quantile projection operator Π_{W_1} is not a non-expansion under 1-Wasserstein distance, which is important for the uniqueness of the fixed point and the convergence of the algorithm.

The proof of (iii) is similar to (i), using the fact that the return distribution μ_2 is bounded on $[0, \frac{R^2_{max}}{1-\gamma}]$ to obtain the following inequality:

$$W_1\left(\Pi_{W_1}\mu_2, \mu_2\right) \leq \frac{R^2_{max}}{K(1-\gamma)}.$$

□

Theorem B.3. ***(Parameterization induced error bound)*** *Let $\Pi_{\mathcal{W}_1}$ be a projection operator onto evenly spaced quantiles τ_i's where each $\tau_i = \frac{2i-1}{2N}$ for $i = 1, \ldots, N$, and $\eta_k \in \mathscr{P}(\mathbb{R})$ be the return distribution of k-th iteration. Let random variables $Z^k_\theta \sim \Pi_{\mathcal{W}_1}\mathcal{T}^\pi\eta_k$ and $Z^k \sim \mathcal{T}^\pi\eta_k$. Assume that the distribution of the immediate reward is supported on $[-R_{max}, R_{max}]$, then we have*

$$\lim_{k\to\infty}\left\|\mathcal{E}^k_3\right\|_\infty = \lim_{k\to\infty}\left\|\mathbb{E}Z^k_\theta - \mathbb{E}Z^k\right\|_\infty \leq \frac{2R_{max}}{N(1-\gamma)},$$

where $\mathcal{E}^k_3$ is parametrization induced error at k-th iteration.

Proof. Using the dual representation of the Wasserstein distance (Villani, 2009) and Lemma B.2, $\forall(x, a)$, we have

$$\begin{aligned}
\left|\mathbb{E}Z^k_\theta(x,a) - \mathbb{E}Z^k(x,a)\right| &\leq W_1\left(\Pi_{W_1}\mathcal{T}^\pi\eta_k(x,a), \mathcal{T}^\pi\eta_k(x,a)\right) \\
&\leq \frac{2R_{max}}{N(1-\gamma)}.
\end{aligned}$$

By taking the limitation over (x, a) and iteration k on the left-hand side, we obtain

$$\lim_{k\to\infty} \left\|\mathcal{E}_3^k\right\|_\infty = \lim_{k\to\infty} \left\|\mathbb{E}Z_\theta^k - \mathbb{E}Z^k\right\|_\infty \leq \frac{2R_{max}}{N(1-\gamma)}.$$

In a similar way, the second-order moment can be bounded by,

$$\lim_{k\to\infty} \left\|\mathbb{E}[Z_\theta^k]^2 - \mathbb{E}[Z^k]^2\right\|_\infty \leq \frac{R_{max}^2}{N(1-\gamma)}.$$

It suggests that higher-order moments are not preserved after quantile representation is applied. □

B.2. Proof of Section 4

Lemma B.4. *(expectation by quantiles). Let $Z \sim \nu$ be a random variable with CDF F_ν and quantile function F_ν^{-1}. Then,*

$$\mathbb{E}[Z] = \int_0^1 F_\nu^{-1}(\tau)d\tau.$$

Proof. As any CDF is non-decreasing and right continuous, we have for all $(\tau, z) \in (0, 1) \times \mathbb{R}$:

$$F_\nu^{-1}(\tau) \leq z \Longleftrightarrow \tau \leq F_\nu(z).$$

Then, denoting U by a uniformly distributed random variable over $[0, 1]$,

$$\mathbb{P}(F_\nu^{-1}(U) \leq z) = \mathbb{P}(U \leq F_\nu(z)) = F_\nu(z),$$

which shows that the random variable $F_\nu^{-1}(U)$ has the same distribution as Z. Hence,

$$\mathbb{E}[Z] = \mathbb{E}\left[F_\nu^{-1}(U)\right] = \int_0^1 F_\nu^{-1}(\tau)d\tau$$

□

Lemma B.5. *Consider the linear regression model $\hat{\boldsymbol{Q}} = \mathbf{X}_2\boldsymbol{M}_2 + \mathcal{E}$, $\mathcal{E}$ is distributed on $\mathcal{N}(\mathbf{0}, \sigma^2 V)$, where $V = diag(v_1, v_2, \cdots, v_N), v_i \geq 1, i = 1, \cdots, N$, and we set noise variance $\sigma^2 = 1$ without loss of generality. The WLS estimator is*

$$\widehat{\boldsymbol{M}}_2 = (\mathbf{X}_2^\top V^{-1}\mathbf{X}_2)^{-1}\mathbf{X}_2^\top V^{-1}\hat{\boldsymbol{Q}}, \tag{11}$$

and the distribution of mean estimator takes the form,

$$\hat{m}_1 \sim \mathcal{N}\left(m_1, \frac{1}{\sum_i v_i} + \frac{(\frac{\sum_i v_i z_{\tau_i}}{\sum_i v_i})^2}{\sum_i v_i z_{\tau_i}^2 - \frac{(\sum_i v_i z_{\tau_i})^2}{\sum_i v_i}}\right).$$

When V equals identity matrix I,

$$\hat{m}_1 \sim \mathcal{N}\left(m_1, \frac{1}{N} + \frac{\bar{z}^2}{\sum_i (z_{\tau_i} - \bar{z})^2}\right).$$

Proof. Premultiplying by $V^{-1/2}$, we get the transformed model

$$V^{-1/2}\hat{\boldsymbol{Q}} = V^{-1/2}\mathbf{X}_2\boldsymbol{M}_2 + V^{-1/2}\mathcal{E}.$$

Now, set $\hat{\boldsymbol{Q}}^* = V^{-1/2}\boldsymbol{Q}, X_2^* = V^{-1/2}X_2$, and $\mathcal{E}^* = V^{-1/2}\mathcal{E}$, so that the transformed model can be written as $\hat{\boldsymbol{Q}}^* = \mathbf{X}_2^*\boldsymbol{M}_2 + \mathcal{E}^*$. The transformed model is a Gaussian-Markov model, satisfying OLS assumptions. Thus, the unique OLS solution is $\widehat{\boldsymbol{M}}_2 = \left(X_2^\top V^{-1}X_2\right)^{-1} X_2^\top V^{-1}\hat{\boldsymbol{Q}}$, and $\widehat{\boldsymbol{M}}_2 \sim \mathcal{N}\left(\boldsymbol{M}_2, \sigma^2(X_2^\top V^{-1}X_2)^{-1}\right)$. By computing $(X_2^\top V^{-1}X_2)^{-1}$, we derive $\hat{m}_1 \sim \mathcal{N}\left(m_1, \frac{1}{\sum_i v_i} + \frac{(\frac{\sum_i v_i z_{\tau_i}}{\sum_i v_i})^2}{\sum_i v_i z_{\tau_i}^2 - \frac{(\sum_i v_i z_{\tau_i})^2}{\sum_i v_i}}\right)$. □

Proposition B.6. *Suppose the noise ε_i independently follows $\mathcal{N}(0, v_i)$ where $v_i \geq 1$ for $i = 1, \cdots, N$, then,*

(i) In the homoskedastic case where $v_i = 1$ for $i = 1, \dots N$, the empirical mean estimator $\hat{m}_1^*$ has a lower variance, $\mathrm{Var}(\hat{m}_1^*) < \mathrm{Var}(\hat{m}_1)$ *;*

(ii) In the heteroskedastic case where v_i's are not equal, the QEM estimator $\hat{m}_1$ achieves a lower variance, i.e. $\mathrm{Var}(\hat{m}_1) < \mathrm{Var}(\hat{m}_1^*)$*, if and only if $\bar{v}^2 - 1 - 1/(\frac{(\sum_i v_i \sum_i v_i z_{\tau_i}^2)}{(\sum_i v_i z_{\tau_i})^2} - 1) > 0$, where $\bar{v} = \frac{1}{N}\sum_i v_i$. This inequality holds when $z_{\tau_i} = -z_{\tau_{N-i}}$, which can be guaranteed in QDRL.*

Proof. The proof of (i) comes directly from the comparison of variances, i.e. $\mathrm{Var}(\hat{m}_1) = \frac{1}{N} < \frac{1}{N} + \frac{\bar{z}^2}{\sum_i (z_{\tau_i} - \bar{z})^2} = \mathrm{Var}(\hat{m}_1^*)$. Next, we prove that (ii) holds under a sufficient condition $z_{\tau_i} = -z_{\tau_{N-i}}$. In QDRL, the quantile levels $\tau_i = \frac{2i-1}{2N}$ are equally spaced around 0.5. Under this setup, the condition $z_{\tau_i} = -z_{\tau_{N-i}}$ indeed holds, where z_{τ_i} is the τ_i-th quantile of standard normal distribution. For $N = 2$, we need to validate the inequality $\bar{v}^2 - 1 - 1/(\frac{(\sum_i v_i \sum_i v_i z_{\tau_i}^2)}{(\sum_i v_i z_{\tau_i})^2} - 1) > 0$. This can be transformed into a multivariate extreme value problem. By analyzing the function $f(v_1, v_2) = \frac{(v_1+v_2)^2}{4} - 1 - \frac{1}{\frac{(v_1+v_2)^2}{(v_1-v_2)^2} - 1}$, the infimum of $f(v_1, v_2)$ is 0 when $v_1, v_2 > 1$, and $f(v_1, v_2)$ reaches 0 at the limit $\lim_{(v_1,v_2)\to(1,1)} f(v_1, v_2) = 0$. For $N = 3$, this case is identical to $N = 2$ since $z_{0.5} = 0$. For $N = 4$, $f(v_1, v_2, v_3, v_4) = \frac{(v_1+v_2+v_3+v_4)^2}{N^2} - 1 - \frac{1}{\frac{(v_1+v_2+v_3+v_4)(k^2v_1+v_2+v_3+k^2v_4)}{(kv_1+v_2-v_3-kv_4)^2} - 1}$, and this expression can be factored as, $f(v_1, v_2, v_3, v_4) = \frac{v_1+v_2+v_3+v_4}{N^2 C}\big((v_1 + v_2 + v_3 + v_4)C - N^2(k^2v_1 + v_2 + v_3 + k^2v_4)\big)$, where $C = (k-1)^2 v_1 v_2 + (k+1)^2 v_1 v_3 + 4k^2 v_1 v_4 + 4v_2 v_3 + (k+1)^2 v_2 v_4 + (k+1)^2 v_3 v_4$, and $k = \frac{\Phi^{-1}(7/8)}{\Phi^{-1}(5/8)} > 3$. By comparing the coefficient corresponding to the same terms, we can verify that $f(v_1, v_2, v_3, v_4) > 0$ when $v_i > 1$. Finally, the remaining cases can be proven in the same manner.

□

Theorem B.7. *Consider the policy $\hat{\pi}$ that is learned policy, and denote the optimal policy to be π_{opt}, $\alpha = \max_{x'} D_{TV}(\hat{\pi}(\cdot \mid x') \| \pi_{opt}(\cdot \mid x'))$, and $n(x, a) = |\mathcal{D}|$. For all $\delta \in \mathbb{R}$, with probability at least $1 - \delta$, for any $\eta(x, a) \in \mathscr{P}(\mathbb{R})$, and all $(x, a) \in \mathcal{D}$,*

$$\left\| F_{\hat{\mathcal{T}}^{\hat{\pi}}\eta(x,a)} - F_{\mathcal{T}^{\pi_{opt}}\eta(x,a)} \right\|_\infty \leq (\alpha + 1)\sqrt{\frac{2|\mathcal{X}|}{n(x,a)} \log \frac{4|\mathcal{X}||\mathcal{A}|}{\delta}}.$$

Proof. We give this proof in a tabular MDP. Directly following from the definition of the distributional Bellman operator applied to the CDF, we have that

$$\begin{aligned}
&F_{\hat{\mathcal{T}}^{\hat{\pi}}\eta(x,a)}(u) - F_{\mathcal{T}^{\pi_{opt}}\eta(x,a)}(u) \\
&= \sum_{x',a'} \hat{P}(x' \mid x, a)\hat{\pi}(a' \mid x') F_{\gamma Z(x',a') + \hat{R}(x,a)}(u) - \sum_{x',a'} P(x' \mid x, a)\pi_{opt}(a' \mid x') F_{\gamma Z(x',a') + R(x,a)}(u).
\end{aligned}$$

For notation convenience, we use random variables instead of measures. $\hat{P}$ and $\hat{R}$ are the maximum likelihood estimates of the transition and the reward functions, respectively. Adding and subtracting $\sum_{x',a'} \hat{P}(x' \mid x, a)\pi_{opt}(a' \mid x') F_{\gamma Z(x',a') + R(x,a)}(u)$, then we have

$$\begin{aligned}
&\sum_{x'} \hat{P}(x' \mid x, a) \sum_{a'} \left(\hat{\pi}(a' \mid x') F_{\gamma Z(x',a') + \hat{R}(x,a)}(u) - \pi_{opt}(a' \mid x') F_{\gamma Z(x',a') + R(x,a)}(u) \right) \\
&+ \sum_{x',a'} \left(\hat{P}(x' \mid x, a) - P(x' \mid x, a) \right) \pi_{opt}(a' \mid x') F_{\gamma Z(x',a') + R(x,a)}(u).
\end{aligned}$$

For the first term, note that

$$\begin{aligned}
&\sum_{x'} \hat{P}(x' \mid x,a) \sum_{a'} \left(\hat{\pi}(a' \mid x') F_{\gamma Z(x',a')+\hat{R}(x,a)}(u) - \pi_{opt}(a' \mid x') F_{\gamma Z(x',a')+R(x,a)}(u) \right) \\
&\leq \sum_{x'} \hat{P}(x' \mid x,a) \sum_{a'} |\hat{\pi}(a' \mid x') - \pi_{opt}(a' \mid x')| \cdot \left| F_{\gamma Z(x',a')+\hat{R}(x,a)}(u) - F_{\gamma Z(x',a')+R(x,a)}(u) \right| \\
&= \sum_{x'} \hat{P}(x' \mid x,a) \sum_{a'} |\hat{\pi}(a' \mid x') - \pi_{opt}(a' \mid x')| \cdot \int \left| F_{\hat{R}(x,a)}(r) - F_{R(x,a)}(r) \right| dF_{\gamma Z(x',a')}(u-r) \\
&\leq \sum_{x'} \hat{P}(x' \mid x,a) \sum_{a'} |\hat{\pi}(a' \mid x') - \pi_{opt}(a' \mid x')| \cdot \sup_r \left| F_{\hat{R}(x,a)}(r) - F_{R(x,a)}(r) \right| \int dF_{\gamma Z(x',a')}(u-r) \\
&= 2 \sum_{x'} \hat{P}(x' \mid x,a) D_{TV}(\hat{\pi}(\cdot \mid x') || \pi_{opt}(\cdot \mid x')) \cdot \left\| F_{\hat{R}(x,a)}(\cdot) - F_{R(x,a)}(\cdot) \right\|_\infty \\
&\leq 2\alpha \left\| F_{\hat{R}(x,a)}(\cdot) - F_{R(x,a)}(\cdot) \right\|_\infty .
\end{aligned}$$

The second term can be bounded as follows:

$$\begin{aligned}
&\sum_{x',a'} \left(\hat{P}(x' \mid x,a) - P(x' \mid x,a) \right) \pi_{opt}(a' \mid x') F_{\gamma Z(x',a')+R(x,a)}(u) \\
&\leq \sum_{x'} \left(\hat{P}(x' \mid x,a) - P(x' \mid x,a) \right) \sum_{a'} \pi_{opt}(a' \mid x') \\
&\leq \left\| \hat{P}(\cdot \mid x,a) - P(\cdot \mid x,a) \right\|_1 \cdot \left\| \sum_{a'} \pi_{opt}(a' \mid \cdot) \right\|_\infty \\
&= \left\| \hat{P}(\cdot \mid x,a) - P(\cdot \mid x,a) \right\|_1 .
\end{aligned}$$

Next, we show the two norms can be bounded. By the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, the following inequality holds with probability at least $1 - \delta/2$, for all $(x,a) \in \mathcal{D}$,

$$\left\| F_{\hat{R}(x,a)}(\cdot) - F_{R(x,a)}(\cdot) \right\|_\infty \leq \sqrt{\frac{1}{2n(x,a)} \log \frac{4|\mathcal{X}||\mathcal{A}|}{\delta}}.$$

By Hoeffding's inequality and an l_1 concentration bound for multinomial distribution[2], the following inequality holds with probability at least $1 - \delta/2$,

$$\max_{x,a} \left\| \hat{P}(\cdot \mid x,a) - P(\cdot \mid x,a) \right\|_1 \leq \sqrt{\frac{2|\mathcal{X}|}{n(x,a)} \log \frac{4|\mathcal{X}||\mathcal{A}|}{\delta}}.$$

Consequently, the claim follows from combining the two inequalities. □

C. Cornish-Fisher Expansion

The Cornish-Fisher Expansion (Cornish & Fisher, 1938) is an asymptotic expansion used to approximate the quantiles of a probability distribution based on its cumulants. To be more explicit, let X^* be a non-gaussian variable with mean 0 and variance 1. Then, the Cornish-Fisher Expansion can be represented as a polynomial expansion:

$$F_{X^*}^{-1}(\tau) = \sum_{i=0}^{\infty} a_i (\Phi^{-1}(\tau))^i,$$

[2] see https://nanjiang.cs.illinois.edu/files/cs598/note3.pdf.

where the parameters a_i depend on the cumulants of the X^* and Φ is the standard normal distribution function. To use this expansion in practice, we need to truncate the series. According to Cornish & Fisher (1938), the highest power of i must be odd, and the fourth order ($i = 3$) approximation is commonly used in practice. The parameters for the fourth order expansion are $a_2 = a_0 = \frac{\kappa_3}{6}$, $a_1 = 1 + 5(\frac{\kappa_3}{6})^2 - 3\frac{\kappa_4}{24}$ and $a_3 = \frac{\kappa_4}{24} - 2(\frac{\kappa_3}{6})^2$, where κ_i denotes i-th cumulant. Therefore, the fourth order expansion is

$$F_{X^*}^{-1}(\tau) = -\frac{\kappa_3}{6} + (1 + 5(\frac{\kappa_3}{6})^2 - 3\frac{\kappa_4}{24})\Phi^{-1}(\tau) + \frac{\kappa_3}{6}(\Phi^{-1}(\tau))^2 + (\frac{\kappa_4}{24} - 2(\frac{\kappa_3}{6})^2)(\Phi^{-1}(\tau))^3 + \cdots.$$

Now, simply define the X^* as the normalization of X, $X = \mu + \sigma X^*$, with mean μ and variance σ^2. $F_X^{-1}(\tau)$ can be approximated by

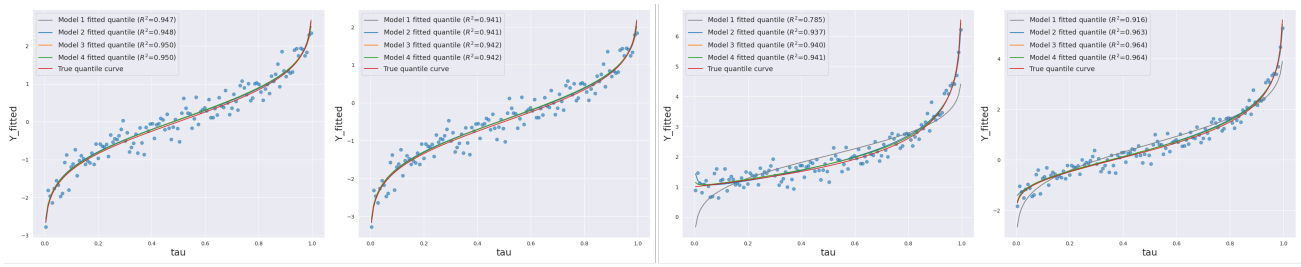
$$F_X^{-1}(\tau) = \mu + \sigma\left(-\frac{\kappa_3}{6\sigma^3} + (1 + 5(\frac{\kappa_3}{6\sigma^3})^2 - 3\frac{\kappa_4}{24\sigma^4})\Phi^{-1}(\tau) + \frac{\kappa_3}{6\sigma^3}(\Phi^{-1}(\tau))^2 + (\frac{\kappa_4}{24\sigma^4} - 2(\frac{\kappa_3}{6\sigma^3})^2)(\Phi^{-1}(\tau))^3 + \cdots\right).$$

Denote skewness $s = \frac{\kappa_3}{\sigma^3}$, kurtosis $k = \frac{\kappa_4}{\sigma^4}$ and normal distribution quantile $z_\tau = \Phi^{-1}(\tau)$. Then, we can rewrite the above equation

$$F_X^{-1}(\tau) = \mu + \sigma\left(z_\tau + (z_\tau^2 - 1)\frac{s}{6} + (z_\tau^3 - 2z_\tau)\frac{k}{24} + (-2z_\tau^3 + 5z_\tau)(\frac{s}{6})^2 + \cdots\right). \tag{12}$$

C.1. Regression model selection

We use the R-Squared (R^2) statistic to determine the number of terms in Equation (12) that should be included in the regression model. R^2, also known as the coefficient of determination, is a statistical measure that shows how well the independent variables explain the variance in the dependent variable. In other words, it is a measure of how well the data fit the regression model.



(a) (b) (c) (d)

Figure 9. Fitted quantile plot. (a) Normal, $\mathcal{N}(0, 1)$. (b) Mixture Gaussian, $0.7\mathcal{N}(-2, 1) + 0.3\mathcal{N}(3, 1)$. (c) Exponential, $Exp(1) = e^{-x}$. (d) Gumbel, $G(0, 1) = e^{-(x+e^{-x})}$.

Consider the linear regression model,

$$\hat{\boldsymbol{Y}} = \mathbf{X}_i\boldsymbol{\beta}_i + \mathcal{E}.$$

The dependent variable $\boldsymbol{Y} = (F_X^{-1}(\tau_1), \ldots, F_X^{-1}(\tau_N))^T$ is composed of the quantiles from distribution of X, and $\mathcal{E}$ is the noise vector sampled from $\mathcal{N}(0, 0.25)$. When the design matrix $\mathbf{X}_1 = (1, \cdots, 1)'$, this regression model reduces to a one-sample problem, and $\boldsymbol{\beta}_1$ can be directly estimated by $\frac{1}{N}\sum_{n=1}^{N} F_X^{-1}(\tau_n)$. We then investigate the following four types of regression models,

Model 1:

$$\mathbf{X}_2 = \begin{pmatrix} 1, & \cdots & ,1 \\ z_{\tau_1}, & \cdots & ,z_{\tau_N} \end{pmatrix}^T, \boldsymbol{\beta}_2 = (\mu, \sigma)^T,$$

Model 2:

$$\mathbf{X}_3 = \begin{pmatrix} 1, & \cdots & ,1 \\ z_{\tau_1}, & \cdots & ,z_{\tau_N} \\ z_{\tau_1}^2 - 1, & \cdots & ,z_{\tau_N}^2 - 1 \end{pmatrix}^T, \boldsymbol{\beta}_3 = \left(\mu, \sigma, \sigma\frac{s}{6}\right)^T,$$

Model 3:

$$\mathbf{X}_4 = \begin{pmatrix} 1, & \cdots & ,1 \\ z_{\tau_1}, & \cdots & ,z_{\tau_N} \\ z_{\tau_1}^2 - 1, & \cdots & ,z_{\tau_N}^2 - 1 \\ z_{\tau_1}^3 - 3z_{\tau_1}, & \cdots & ,z_{\tau_N}^3 - 3z_{\tau_N} \end{pmatrix}^T, \boldsymbol{\beta}_4 = \left(\mu, \sigma, \sigma\frac{s}{6}, \sigma\frac{k}{24}\right)^T,$$

Model 4:

$$\mathbf{X}_5 = \begin{pmatrix} 1, & \cdots & ,1 \\ z_{\tau_1}, & \cdots & ,z_{\tau_N} \\ z_{\tau_1}^2 - 1, & \cdots & ,z_{\tau_N}^2 - 1 \\ z_{\tau_1}^3 - 3z_{\tau_1}, & \cdots & ,z_{\tau_N}^3 - 3z_{\tau_N} \\ -2z_{\tau_1}^3 + 5z_{\tau_1}, & \cdots & ,-2z_{\tau_N}^3 + 5z_{\tau_N} \end{pmatrix}^T, \boldsymbol{\beta}_5 = \left(\mu, \sigma, \sigma\frac{s}{6}, \sigma\frac{k}{24}, \sigma(\frac{s}{6})^2\right)^T.$$

Figure 9 shows that the regression fitted values and corresponding R^2 across several distributions of X. As the number of independent variables increases, more variance in the error can be explained. However, having too many independent variables increases the risk of multicollinearity and overfitting. Based on practical considerations, we choose Model 3 as our regression model due to its satisfactory level of explainability. In the subsequent section, we will give a more in-depth interpretation of this regression model.

C.2. Interpretation of the remaining term $\omega(\tau)$

In this section, we explore the role of the remaining term $\omega(\tau)$ in the context of random design regression. As discussed in Section 4, we present a decomposition of the estimate $\hat{q}(\tau)$ of the τ-th quantile, which includes contributions from the mean, noise error, and misspecified error. Specifically, we expressed the estimate as follows:

$$\hat{q}(\tau) = \mu + \omega_1(\tau) + \varepsilon(\tau).$$

where μ can be estimated using the mean estimator $\frac{1}{N}\sum q(\tau_i)$, which is commonly used in QDRL and IQN settings. However, this simple model fails to capture important information in the $\omega_1(\tau)$. To address this limitation, we employ the Cornish-Fisher Expansion to expand the equation, resulting in the following expression:

$$\begin{aligned} \hat{q}(\tau) =& \mu + z_\tau\sigma + \sigma\omega_2(\tau) + \varepsilon(\tau), \\ \hat{q}(\tau) =& \mu + z_\tau\sigma + (z_\tau^2 - 1)\sigma\frac{s}{6} + \sigma\omega_3(\tau) + \varepsilon(\tau), \\ & \cdots \end{aligned}$$

where μ can be estimated by linear regression estimator given multiple quantile levels $\{\tau_i\}$, which can be sampled from a uniform distribution or predefined to be evenly spaced in $(0, 1)$. In theory, higher-order expansions can capture more misspecified information in $\omega(\tau)$, leading to a more accurate representation of the quantile. However, as discussed before, expansions are typically limited to the fourth order in practice to balance the trade-off between model complexity and estimation accuracy.

To gain a better understanding of the remaining term $\omega(\tau)$ and its impact on the regression estimator, consider the linear model,

$$\hat{q}(\tau) = \mathbf{x}_\tau'\beta + \underbrace{\omega_\tau}_{\text{Misspecified error}} + \underbrace{\varepsilon}_{\text{Noise error}},$$

where τ can be generally considered a uniform, $\mathbf{x}_\tau = (1, z_\tau, z_\tau^2 - 1, ...)' \in \mathbb{R}^d$, and $\beta = (\mu, \sigma, \sigma\frac{s}{6}, ...)' \in \mathbb{R}^d$. In particular, define the random variables,

$$\varepsilon := \hat{q}(\tau) - \mathbb{E}[\hat{q}(\tau) \mid \mathbf{x}_\tau] \quad \text{and} \quad \omega_\tau := \mathbb{E}[\hat{q}(\tau) \mid \mathbf{x}_\tau] - \mathbf{x}_\tau{}'\beta,$$

where ε corresponds to the noise with zero mean, σ^2_{noise} variance and independent across different level of τ, and ω_τ corresponds to the misspecified error of β. Under the following conditions, we can derive a bound for the regression estimator in the misspecified model.

Condition 1 (Subgaussian noise). There exist a finite constant $\sigma_{\text{noise}} \geq 0$ such that for all $\lambda \in \mathbb{R}$, almost surely:

$$\mathbb{E}[\exp(\lambda\varepsilon) \mid \mathbf{x}_\tau] \leq \exp\left(\lambda^2\sigma_{\text{noise}}^2/2\right).$$

Condition 2 (Bounded approximation error). There exist a finite constant $C_{\text{bias}} \geq 0$, almost surely:

$$\left\|\Sigma^{-1/2}\mathbf{x}_\tau\omega_\tau\right\|_2 \leq C_{\text{bias}}\sqrt{d},$$

where $\Sigma = \mathbb{E}[\mathbf{x}_\tau\mathbf{x}_\tau{}']$.

Condition 3 (Subgaussian projections). There exists a finite constant $\rho \geq 1$ such that:

$$\mathbb{E}\left[\exp(\alpha^\top\Sigma^{-1/2}\mathbf{x}_\tau)\right] \leq \exp\left(\rho\cdot\|\alpha\|_2^2/2\right), \quad \forall\alpha \in \mathbb{R}^d.$$

Theorem C.1. *Suppose that Conditions 1, 2, and 3 hold. Then for any $\delta \in (0,1)$ and with probability at least $1-3\delta$, the following holds:*

$$\left\|\hat{\beta}_{\text{ols}} - \beta\right\|_\Sigma^2 \leq \underbrace{K_{\rho,\delta,N}^2\left(\frac{4\mathbb{E}\left\|\Sigma^{-1/2}\mathbf{x}_\tau\omega_\tau\right\|_2^2(1+8\log(1/\delta))}{N} + \frac{3C_{bias}^2\, d\log^2(1/\delta)}{N^2}\right)}_{\textit{Misspecified error contribution}}$$
$$+ \underbrace{K_{\rho,\delta,N}\cdot\frac{\sigma_{noise}^2\cdot(d+2\sqrt{d\log(1/\delta)}+2\log(1/\delta))}{N}}_{\textit{Noise error contribution}},$$

where $K_{\rho,\delta,N}$ is a constant depending on ρ, δ and N.

Proof. The proof of the above theorem can be easily adapted from Theorem 2 in Hsu et al. (2011). □

The first term on the right-hand side represents the error due to model misspecification, which occurs when the true model differs from the assumed model. Intuitively, incorporating more relevant information in $\omega(\tau)$ into explanation variables could decrease the quantity of $\mathbb{E}\left\|\Sigma^{-1/2}\mathbf{x}_\tau\omega_\tau\right\|_2^2$ and C_{bias}. Therefore, the accuracy of the estimator may be potentially improved by reducing the magnitude of the misspecified error. The second term represents the noise error contribution, which is inevitable and can only be controlled by increasing the sample size N.

D. Experimental Details

D.1. Tabular experiment

The parameter settings used for tabular control are presented in Table 1. In the QEMRL case, the weight matrix V is set as shown in the table based on domain knowledge indicating that the distribution has low probability support around its median. The greedy parameter decreases exponentially every 100 steps, and the learning rate decrease in segments every 50K steps.

D.2. Atari experiment

We extend QEMRL to a DQN-like architecture, and we use the same architecture as QR-DQN, which we refer to as QEM-DQN [3]. Our hyperparameter settings (Table 2) are aligned with Dabney et al. (2018b) for a fair comparison. Additionally, we extend QEMRL to the unfixed quantile fraction algorithm IQN, which embeds quantile fraction τ into the quantile value network on the top of QR-DQN. In Atari, it is infeasible to determine the low probability supports for every state-action pair, therefore we only consider the heteroskedasticity that occurs in the tail and treat V as a tuning parameter to select an appropriate value. For exploration experiments, we follow the settings of Mavrin et al. (2019) and set the decay factor $c_t = c\sqrt{\frac{\log t}{t}}$, where $c = 50$.

[3] Code is available at https://github.com/Kuangqi927/QEM

Table 1. The (hyper-)parameters of QEMRL and QDRL used in the tabular control experiment.

Hyperparameter	Value
Learning rate schedule	{0.05,0.025,0.0125}
Discount factor	0.999
Quantile initialization	$\mathrm{Unif}(-0.5, 0.5)$
Number of quantiles	128
Number of training steps	150K
ϵ-greedy schedule	$0.9^{\lfloor t/100 \rfloor}$
Number of MC rollouts	10000
Weight matrix V (QEMRL only)	$diag\{1, 1, \cdots, \underbrace{1.5, \cdots, 1.5}_{\tau \in [0.45, 0.55]}, \cdots, 1, 1\}$

Table 2. The hyperparameters of QEM-DQN and QR-DQN used in the Atari experiments.

Hyperparameter	Value
Learning rate	0.00005
Discount factor	0.99
Optimizer	Adam
Bath size	32
Number of quantiles	200
Number of quantiles (IQN)	32
Weight matrix V (QEM-DQN only)	$diag\{\underbrace{1.5, \cdots, 1.5}_{\tau \in [0.9, 1)}, \cdots, 1, 1, \cdots, \underbrace{1.5, \cdots, 1.5}_{\tau \in (0, 0.1]}\}$

D.3. MuJoCo experiment

We extend QEMRL to a SAC-like architecture, and we use the same architecture of DSAC, named QEM-DSAC. Similarly, we extend QEMRL to an IQN version of DSAC. Hyperparameters and environment-specific parameters are listed in Table 3. In addition, SAC has a variant that introduces a mechanism of fine-tuning α to achieve target entropy adaptively. While this adaptive mechanism performs well, we follow the use of fixed α suggested in the original SAC paper to reduce irrelevant factors.

E. Additional Experimental Results

E.1. Variance reduction for IQN

IQN does not satisfy the sufficient condition $z_{\tau_i} = -z_{\tau_{N-i}}$ since τ is sampled from a uniform distribution, rather than evenly spaced as in QDRL. To examine the impact of this on the inequality $(\frac{\sum_i v_i}{N})^2 - 1 - 1/(\frac{(\sum_i v_i \sum_i v_i z_i^2)}{(\sum_i v_i z_i)^2} - 1) > 0$ in Proposition 4.2, simulation experiments are conducted. We use the function $f(v_1, \cdots, v_N) = (\frac{\sum_i v_i}{N})^2 - 1 - 1/(\frac{(\sum_i v_i \sum_i v_i z_{\tau_i}^2)}{(\sum_i v_i z_{\tau_i})^2} - 1)$ to examine this inequality, where $v_i > 1$ and τ_i are sampled uniformly. In every trial, v_i are randomly sampled from $[1, M]$, repeating the process 100,000 times. The minimum values of $f(v_1, \cdots, v_N)$ are shown in the following Table 4 for varying values of N and M. The results indicate that the minimum of f is always greater than 0, which demonstrates that the inequality holds in practice.

E.2. Weight V tuning experiments

E.3. Additional Atari results

Table 3. The hyperparameters of QEM-DSAC and DSAC used in the MuJoCo experiments.

Hyperparameter	Value
Policy network learning rate	0.0003
Quantile Value network learning rate	0.0003
Discount factor	0.99
Optimization	Adam
Target smoothing	0.005
Batch size	256
Minimum steps before training	10000
Number of quantiles	32
Quantile fraction embedding size (IQN)	64
Weight matrix V (QEM-DSAC only)	$diag\{\underbrace{1.2,\cdots,1.2}_{\tau\in[0.9,1)},\cdots,1,1,\cdots,\underbrace{1.2,\cdots,1.2}_{\tau\in(0,0.1]}\}$

Environment	Temperature Parameter
Ant-v2	0.2
HalfCheetah-v2	0.2
Hopper-v2	0.2
Walker2d-v2	0.2
Swimmer-v2	0.2
Humanoid-v2	0.05

Table 4. Minimum of f.

Minimum of f	M	N
0.614	2	32
4.778	5	32
43.143	20	32
0.932	2	128
7.707	5	128
76.489	20	128
1.082	2	500
9.357	5	500
96.473	20	500

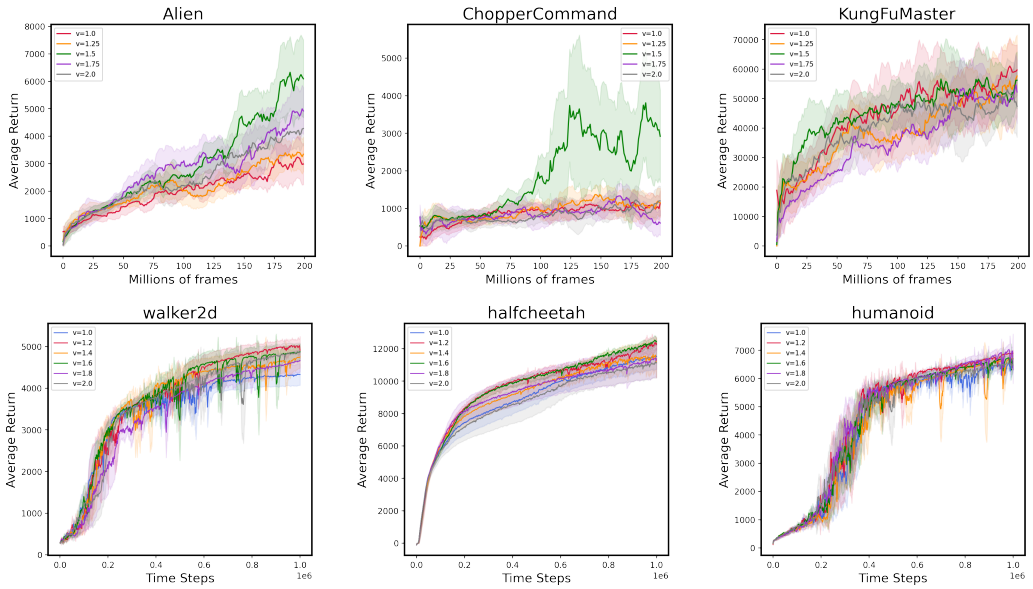


Figure 10. Comparison of different weight v in QEM-DSAC and QEM-DQN experiments

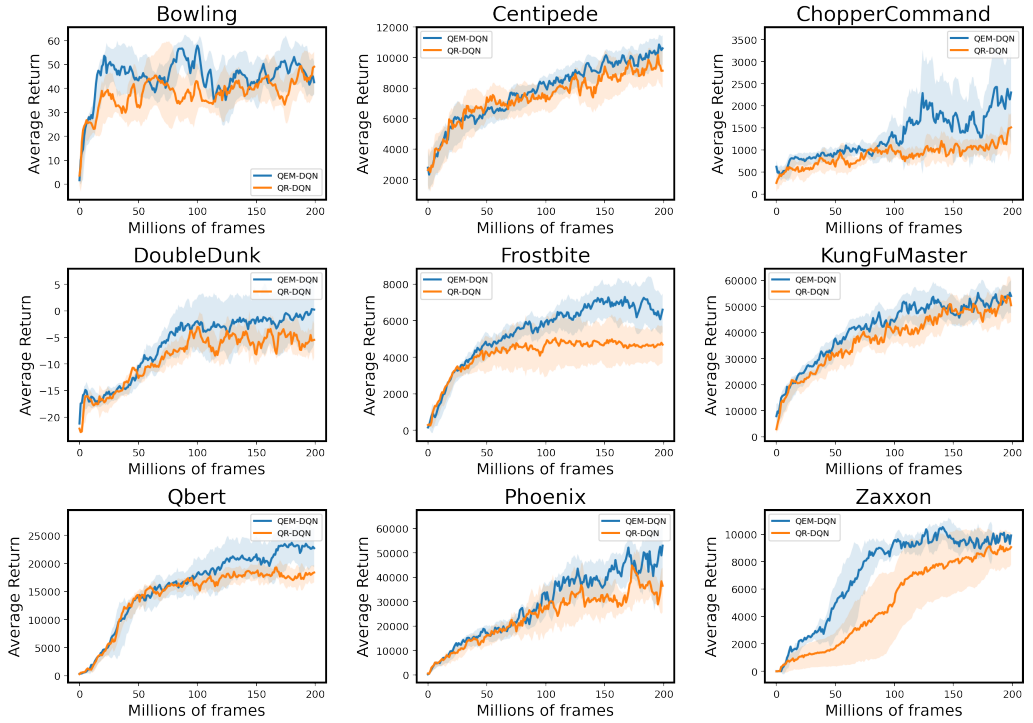


Figure 11. Comparison of QEM-DQN and QR-DQN across 9 Atari games

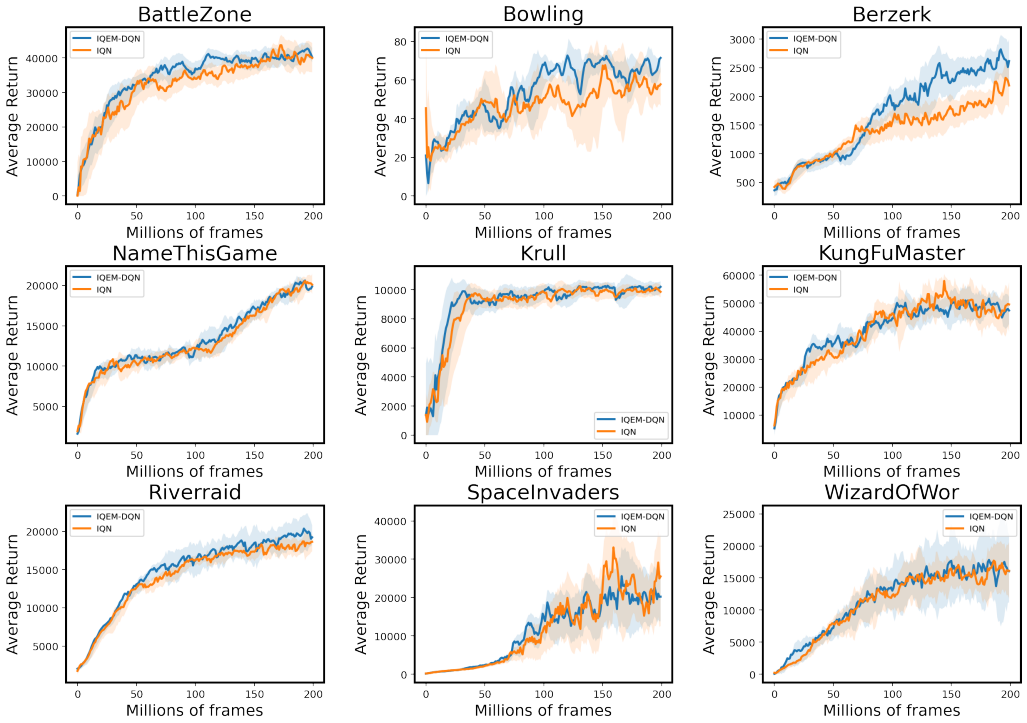


Figure 12. Comparison of IQEM-DQN and IQN across 9 Atari games

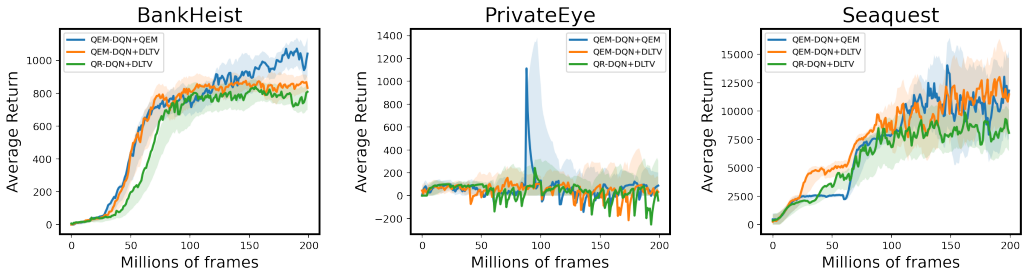


Figure 13. Comparison of QEM and DLTV across 3 hard-explored Atari games