Distributional Off-Policy Evaluation with Deep Quantile Process Regression

Qi Kuang[1]*, Chao Wang[2]*, Yuling Jiao[3]† and Fan Zhou[2]†

[1] School of Statistics and Data Science, and Philosophy and Social Sciences Laboratory of Data Science in Finance and Economics at the Ministry of Education, Jiangxi University of Finance and Economics

[2] School of Statistics and Data Science, Shanghai University of Finance and Economics

[3] School of Artificial Intelligence, and Hubei Key Laboratory of Computational Science, Wuhan University

Abstract

This paper investigates the off-policy evaluation (OPE) problem from a distributional perspective. Rather than focusing solely on the expectation of the total return, as in most existing OPE methods, we aim to estimate the entire return distribution. To this end, we introduce a quantile-based approach for OPE using deep quantile process regression, presenting a novel algorithm called Deep Quantile Process regression-based Off-Policy Evaluation (DQPOPE). We provide new theoretical insights into the deep quantile process regression technique, extending existing approaches that estimate discrete quantiles to estimate a continuous quantile function. A key contribution of our work is the rigorous sample complexity analysis for distributional OPE with deep neural networks, bridging theoretical analysis with practical algorithmic implementations. We show that DQPOPE achieves statistical advantages by estimating the full return distribution using the same sample size required to estimate a single policy value using conventional methods. Empirical studies further show that DQPOPE provides significantly more precise and robust policy value estimates than standard methods, thereby enhancing the practical applicability and effectiveness of distributional reinforcement learning approaches.

Keywords: Distributional off-policy evaluation, Distributional reinforcement learning, Deep quantile process regression, Deep ReLU networks, Sample complexity

*The first two authors contribute equally to this paper

†Joint corresponding authors

1 Introduction

Off-policy evaluation (OPE) is a fundamental problem in reinforcement learning (RL) that seeks to estimate the value of a target policy from data collected under a different behavior policy. Its importance is especially pronounced in applications where online experimentation is costly, risky, or ethically constrained. OPE arises in a broad range of settings, including contextual bandits and more general sequential decision-making problems (Liao et al., 2021). In fields like healthcare, OPE enables the assessment of dynamic treatment policies using historical electronic health records, where real-world experimentation may be impractical or ethically sensitive (Wang et al., 2012; Zhu et al., 2019). In such settings, applying a new treatment policy without offline validation may lead to ethical concerns. These considerations make OPE a particularly vital tool in offline RL.

In recent years, distributional reinforcement learning (DRL) has gained significant traction as an alternative to traditional RL methods. Rather than estimating just the expected value of future returns, DRL models the entire distribution of returns (Bellemare et al., 2017; Dabney et al., 2018b,a), capturing the inherent randomness in dynamic environments. This approach has shown promise, particularly when the mean information is insufficient to represent the full complexity of decision-making. By accounting for the distribution, DRL offers advantages in mean estimation. Rowland et al. (2023) empirically demonstrate that quantile-based distributional RL (QDRL), through quantile averaging, can achieve a lower mean squared error (MSE) in value estimation compared to standard RL approaches.

The potential of DRL methods has been emphasized in various real-world decision-making scenarios (Bodnar et al., 2020; Bellemare et al., 2020). For instance, in neuroscience, recent work (Dabney et al., 2020; Muller et al., 2024; Lowet et al., 2025) demonstrate that in these complex biological environments, the distribution learning perspective is biologically

plausible. These studies reveal that animals may encode not only the mean of return but the entire distribution of possible outcomes, enabling more flexible, nuanced, and context-sensitive decision-making. Similarly, in healthcare, Jin et al. (2023) introduced a Bayesian framework that incorporates distributional methods to optimize sequential combination antiretroviral therapy (cART) for HIV patients, accounting for uncertainties in patient outcomes over time. In the order dispatching systems of ride-sharing platforms, where balancing driver workloads with maximizing customer satisfaction is a core challenge (Qin et al., 2025), distributional methods can effectively capture the distributional characteristics of these objectives, enabling better management of trade-offs and leading to more informed decision-making based on value functions (Zhou et al., 2021).

While DRL has been developed largely for online control and policy optimization, these approaches are not directly applicable to offline settings. Nevertheless, given the remarkable success of distributional reinforcement learning (DRL) in online scenarios, its potential in offline tasks, such as off-policy evaluation (OPE), is promising. In principle, a distributional approach to OPE can provide a richer characterization of policy performance and may also improve the accuracy of policy value estimation.

Additionally, this paper contributes to the theoretical understanding of distributional reinforcement learning. Despite its strong empirical performance, the theoretical foundations underlying the advantages of distributional methods over standard reinforcement learning remain limited. Existing analyses are largely confined either to maximum likelihood estimation (MLE) frameworks (Wu et al., 2023) or to relatively simple tabular cases (Rowland et al., 2024a; Zhang et al., 2025). As a result, there is still limited understanding of how distributional methods, especially when combined with deep neural network approximation, can yield statistical and practical benefits in complex, real-world applica-

tions. To address this gap, we propose **D**eep **Q**uantile **P**rocess regression-based **O**ff-**P**olicy **E**valuation (DQPOPE), a novel approach that applies distributional methods to off-policy evaluation (OPE) using deep quantile process regression. This paper aims to establish a theoretical foundation for distributional OPE, providing insights that are relevant to the broader DRL landscape. Our contributions can be summarized as follows:

- **Introduction of quantile process regression for OPE.** Unlike prior QDRL approaches that focus on estimating discrete quantiles, our method employs the quantile process to model the entire return distribution, thus avoiding the representation error inherent in discrete quantile approximations. Essentially, the introduction of quantile process regression effectively transforms the inherently infinite-dimensional distribution learning task into a finite-dimensional regression task, enhancing both theoretical tractability and practical implementation.

- **Advancing the theory of deep distributional OPE.** We provide a theoretical framework for analyzing distributional OPE through the lens of the quantile process. Particularly, we employ deep neural network (DNN) approximations, which are essential both theoretically and practically due to the inherently nonlinear structure of the distributional Bellman operator (Rowland et al., 2024a).

- **Advantages over value-based OPE methods.** We theoretically demonstrate that DQPOPE can estimate the entire return distribution with the same sample size required by value-based OPE methods for estimating only the distribution mean. Our empirical results further indicate that DQPOPE consistently achieves more accurate mean value estimates compared to value-based OPE methods by better capturing the randomness.

2 Problem Setup and Notations

Notations. For any measurable space $\mathcal{X}$, let $\Delta(\mathcal{X})$ denote the set of all probability measures on $\mathcal{X}$. For a measurable function $f : \mathcal{X} \to \mathbb{R}$ and $\nu \in \Delta(\mathcal{X})$, define $\|f\|_{p,\nu} = \big(\int_{\mathcal{X}} |f(x)|^p \, d\nu(x)\big)^{\frac{1}{p}}$ for $p \geq 1$ if it exists. For any $\nu, \mu \in \Delta(\mathbb{R})$, the p-Wasserstein distance between ν and μ is defined by $\mathcal{W}_p(\nu, \mu) := \big(\int_0^1 |F_\nu^{-1}(t) - F_\mu^{-1}(t)|^p dt\big)^{1/p}$, where F_ν^{-1} and F_μ^{-1} denote the quantile functions of ν and μ, respectively. For a measurable map $\eta : \mathcal{X} \to \Delta(\mathbb{R})$ and a random variable $X \sim \nu \in \Delta(\mathcal{X})$, we use boldface to denote the mixture distribution induced by $\eta(X)$, namely $\boldsymbol{\eta} := \mathbb{E}[\eta(X)]$. For two positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some positive constant $C > 0$ independent of n. For $v = (v_1, ..., v_d)^\top \in \mathbb{R}^d$, define $\max\{v, 0\} = (\max\{v_1, 0\}, ..., \max\{v_d, 0\})^\top$. Let $\mathbb{N}$ denote the natural numbers, $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$, and $\lfloor x \rfloor$ the floor function. The $\mathcal{O}$ notation omits constants and lower-order terms for clarity.

Markov Decision Processes. Consider a Markov Decision Process $(\mathcal{S}, \mathcal{A}, P, \gamma, \mathcal{R})$. $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $\gamma \in (0, 1)$ is some pre-specified discounted factor, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the distribution of reward. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the distribution $\pi(\cdot|s)$ of taking action given $s \in \mathcal{S}$. Starting from an initial state $S_0 \sim \rho \in \Delta(\mathcal{S})$, the trajectory $\{S_t, A_t, R_t\}_{t \geq 0}$ evolves according to $A_t \sim \pi(\cdot|\, S_t)$, $R_t \sim \mathcal{R}(\cdot|S_t, A_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$.

Standard RL estimates the expected return $Q^\pi(s, a) := \mathbb{E}_\pi\big[\sum_{t=0}^\infty \gamma^t R_t \mid S_0 = s, A_0 = a\big]$, where $\mathbb{E}_\pi$ takes expectation over $\{R_t\}_{t \geq 0}$ given $S_0 = s, A_0 = a$ under the policy π. DRL instead studies the law of the discounted return $Z^\pi(s, a) := \sum_{t=0}^\infty \gamma^t R_t \mid S_0 = s, A_0 = a$, regarded as a random variable indexed by (s, a). Denote the collection of maps from $\mathcal{S} \times \mathcal{A}$ to $\Delta(\mathbb{R})$ as $\Delta(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}$. The return distribution $\eta^\pi \in \Delta(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}$ is defined by $\eta^\pi(s, a) = \text{law}(Z^\pi(s, a))$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $\text{law}(\cdot)$ extracts the distribution of the input

random variable. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, let (s,a,R,S') be a random transition according to $R \sim \mathcal{R}(\cdot|s,a), S' \sim P(\cdot|s,a)$. The distributional Bellman operator $\mathcal{T}^\pi : \Delta(\mathbb{R})^{\mathcal{S}\times\mathcal{A}} \to \Delta(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ is the mapping defined by

$$(\mathcal{T}^\pi \eta)(s,a) := \mathbb{E}_\pi\big[(g_{\gamma,R})_\# \eta(S',A') \mid s,a\big], \tag{1}$$

where $g_{\gamma,r} : \mathbb{R} \to \mathbb{R}$ is defined by $g_{\gamma,r}(s) = r + \gamma s$, and $(g_{\gamma,r})_\# \nu$ is the pushforward on distribution ν defined as $(g_{\gamma,r})_\# \nu = \mathrm{law}(g_{\gamma,r}(Z))$, with $Z \sim \nu$. Equivalently, for $Z(s,a) \sim \eta(s,a)$, the $\mathcal{T}^\pi$ can also be defined in terms of random variables (Bellemare et al., 2023)[1]

$$\big(\mathcal{T}^\pi \eta\big)(s,a) := \mathrm{law}\big(R + \gamma Z(S',A') \mid s,a\big). \tag{2}$$

The return distribution η^π is the unique fixed point of $\mathcal{T}^\pi$, namely $\mathcal{T}^\pi \eta^\pi = \eta^\pi$ (Bellemare et al., 2023). Finally, define the discounted occupancy distribution under policy π and initial state distribution ρ by $d^\pi(s,a) := (1-\gamma)\ \sum_{t=0}^\infty \gamma^t\ \mathbb{P}\big[S_t = s, A_t = a \mid \pi, \rho\big]$.

Off-policy evaluation. In standard off-policy evaluation (OPE), the goal is to estimate the value of a target policy π under an initial distribution ρ, defined by $V^\pi := \mathbb{E}_\pi\big[\sum_{t=0}^\infty \gamma^t R_t\big] = \mathbb{E}_{(S,A)\sim\rho\times\pi}\big[Q^\pi(S,A)\big]$, where $(\rho \times \pi)(s,a) := \rho(s)\pi(a|s)$. For any estimate $\widehat{Q}$ of Q^π, its estimation accuracy is typically assessed by the absolute error $|V^\pi - \widehat{V}|$, where $\widehat{V} = \mathbb{E}_{(S,A)\sim\rho\times\pi}\big[\widehat{Q}(S,A)\big]$. Distributional OPE instead targets the full return distribution. The target of interest is the performance of the target policy π, defined by $\boldsymbol{\eta^\pi} := \mathbb{E}_{(S,A)\sim\rho\times\pi}\big[\eta^\pi(S,A)\big]$, which is the mixture distribution obtained by averaging $\eta^\pi(S,A)$ over $(S,A) \sim \rho \times \pi$. For any estimate $\widehat{\eta} \in \Delta(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ of η^π, the estimation accuracy is assessed by a distributional discrepancy, specifically the p-Wasserstein distance $\mathcal{W}_p(\boldsymbol{\eta^\pi}, \widehat{\boldsymbol{\eta}})$, where $\widehat{\boldsymbol{\eta}} = \mathbb{E}_{(S,A)\sim\rho\times\pi}\big[\widehat{\eta}(S,A)\big]$. The offline dataset can be summarized as $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, which is collected under another unknown behavior policy π^b with the data generating pro-

[1] One can refer to Section 4 of this book for more details.

cedure that $a_i \sim \pi^b(\cdot|s_i)$, $r_i \sim \mathcal{R}(\cdot|s_i, a_i)$, and $s_i' \sim P(\cdot|s_i, a_i)$. We assume that (s_i, a_i) are i.i.d. draws from the data distribution $\mu(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\big[S_t = s, A_t = a \mid \pi^b, \rho\big]$.

We aim to analyze the sample complexity of distributional OPE by bounding $\mathcal{W}_p(\boldsymbol{\eta}^\pi, \widehat{\boldsymbol{\eta}})$. We adopt an expected Wasserstein metric equipped with an L_2-norm structure, following recent literature (Abdullah et al., 2019; Wu et al., 2023). Given a distribution ν over (S, A), for any $\eta, \eta' \in \Delta(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$, define

$$\overline{\mathcal{W}}_{p,\nu}(\eta, \eta') := \Big(\mathbb{E}_{(S,A)\sim\nu}\left[\mathcal{W}_p^{2p}(\eta(S,A), \eta'(S,A))\right]\Big)^{\frac{1}{2p}}.$$

In particular, when $\nu = d^\pi$, Wu et al. (2023) shows that for any $\eta, \eta' \in \Delta(\mathbb{R})^{\mathcal{S}\times\mathcal{A}}$ and $p \geq 1$, the distributional Bellman operator is $\gamma^{1-\frac{1}{2p}}$-contractive under the metric $\overline{\mathcal{W}}_{p,d^\pi}(\cdot,\cdot)$,

$$\overline{\mathcal{W}}_{p,d^\pi}\big(\mathcal{T}^\pi\eta, \mathcal{T}^\pi\eta'\big) \;\leq\; \gamma^{1-\frac{1}{2p}}\; \overline{\mathcal{W}}_{p,d^\pi}\big(\eta, \eta'\big).$$

This property is crucial for deriving the sub-optimality decomposition in Lemma 4.6.

3 Deep Quantile Process Regression-based Off-Policy Evaluation (DQPOPE)

In this section, we provide a comprehensive description of the proposed DQPOPE method and demonstrate its advantages from both analytical and practical perspectives.

3.1 Foundations and Implementation of DQPOPE

The goal of distributional OPE is to estimate the target return distribution η^π. Leveraging the contraction property of $\mathcal{T}^\pi$, a natural approach is to consider the iterative scheme $\eta_t = \mathcal{T}^\pi\eta_{t-1}$, initialized from some η_0. This sequence converges to the fixed-point η^π under the metric $\overline{\mathcal{W}}_{p,d^\pi}(\cdot,\cdot)$. However, directly applying the operation $\mathcal{T}^\pi\eta_t$ is infeasible

in implementation, necessitating an approximation of $\mathcal{T}^\pi$ at each step. To this end, we exploit the one-to-one correspondence between a distribution and its quantile function and reformulate the distributional iteration as a sequence of quantile function estimation problems. For any fixed $(s,a) \in \mathcal{S}\times\mathcal{A}$, denote $f(s,a,\cdot) : (0,1) \to \mathbb{R}$ as the quantile function of its corresponding distribution $\eta(s,a)$. The learning task therefore gets transferred into recovering the quantile function $f^* : \mathcal{S} \times \mathcal{A} \times (0,1) \to \mathbb{R}$ of the target return distribution $\eta^\pi \in \Delta^{\mathcal{S}\times\mathcal{A}}$. We proceed to detail how to implement a one-step distributional Bellman update using quantile process regression.

Specifically, starting from the quantile function $\widehat{f}_0 : \mathcal{S} \times \mathcal{A} \times (0,1) \to \mathbb{R}$, the algorithm recursively produces a sequence of quantile functions $\widehat{f}_1, \widehat{f}_2, \cdots, \widehat{f}_T$, where each quantile function $\widehat{f}_t$ is searched from certain function space $\mathcal{F}$. For each $t \in [T]$, the quantile function $\widehat{f}_t$ induces a distribution $\widehat{\eta}_t$, and hence provides the full distributional information at iteration t. Recall that given any (s,a), $R \sim \mathcal{R}(\cdot|s,a)$, $S' \sim P(\cdot|s,a)$, and let $Z_t(s,a) \sim \widehat{\eta}_t(s,a)$. For the transition (s,a,R,S'), by definition of $\mathcal{T}^\pi$ in (2), $\mathcal{T}^\pi\widehat{\eta}_{t-1}$ is given by

$$\left(\mathcal{T}^\pi\widehat{\eta}_{t-1}\right)(s,a) := \mathrm{law}\Big(R + \gamma Z_{t-1}(S',A') \mid s,a\Big).$$

This reformulates the problem into finding the quantile function of the random variable $Y_t = R + \gamma Z_{t-1}(S',A')$ conditional on (s,a). Writing $X = (S,A)$, the conditional quantile function of Y_t given X is the solution to the risk minimization problem $f_t^* = \arg\min_f \mathcal{L}_t(f)$:

$$\mathcal{L}_t(f) = \mathbb{E}_{X,Y_t,\tau}\Big(\rho_\tau(Y_t - f(X,\tau))\Big), \tag{3}$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}_{u\le 0})$ is the check loss, and $\tau \sim \mathrm{Unif}(0,1)$ is independent of (X,Y_t). Here the quantile level τ is treated as a random input to the function f, a formulation referred to as a quantile process (Volgushev et al., 2019). Henceforth, we slightly abuse the notation of τ as both random variable and fixed quantile level when clear from context.

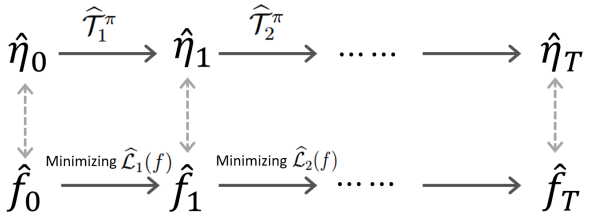

Figure 1: Illustration of the equivalence between distribution iteration in DRL and quantile process training procedure in statistics. Herein, $\widehat{\mathcal{T}}_t^{\pi}$ denotes the transition operator from $\widehat{\eta}_{t-1}$ to $\widehat{\eta}_t$, such that $\widehat{\eta}_t = \widehat{\mathcal{T}}_t^{\pi}\widehat{\eta}_{t-1}$. See Section C of the Supplemental Material for details.

In implementation, the dataset is split into T equal subsets $\{\mathcal{D}_t\}_{t=1}^{T}$, with each of size $n = |\mathcal{D}_t|$, so that $N = nT$. At iteration t, we work with the empirical counterpart of $\mathcal{L}_t(\cdot)$ constructed from the dataset $\mathcal{D}_t$. However, the random variable $Y_t = R + \gamma Z_{t-1}(S', A')$ cannot be directly observed since Z_{t-1} depends on future rewards. To address this issue, we recover $Z_{t-1}(S', A')$ through its quantile function $\widehat{f}_{t-1}(S', A', U)$, where $U \sim \text{Unif}(0, 1)$ is independent of (s, a, R, S'). Specifically, a useful property states that $\widehat{f}_{t-1}(S', A', U)$ has the same distribution as $Z_{t-1}(S', A')$ [2] (see Proposition D.1 of Supplemental Material). For each transition sample (s_i, a_i, r_i, s_i'), sampling $a_i' \sim \pi(\cdot|s_i')$ and $u_i \sim \text{Unif}(0, 1)$, we generate $\widehat{f}_{t-1}(s_i', a_i', u_i)$ by plugging (s_i', a_i', u_i) into $\widehat{f}_{t-1}$. This allows us to generate the exact samples $r_i + \gamma\widehat{f}_{t-1}(s_i', a_i', u_i)$ from Y_t. Consequently, we obtain a collection of i.i.d. samples (x_i, y_i, τ_i), which we use to define the empirical risk:

$$\widehat{\mathcal{L}}_t(f) = \frac{1}{|\mathcal{D}_t|} \sum_{(s_i, a_i, r_i, s_i') \in \mathcal{D}_t} \rho_{\tau_i}\big(y_i - f(x_i, \tau_i)\big), \tag{4}$$

where $x_i = (s_i, a_i)$, $\tau_i \sim \text{Unif}(0, 1)$, and $y_i = r_i + \gamma\widehat{f}_{t-1}(s_i', a_i', u_i)$. The quantile function estimator $\widehat{f}_t$ is then obtained by minimizing the empirical risk over certain function space $\mathcal{F}$, such that $\widehat{f}_t = \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{L}}_t(f)$. In the context of deep quantile regression, the function

[2] This property enables us recover the sample from $Z_t(S, A)$ by sampling $u_i \sim \text{Unif}(0, 1)$ and plugging into $\widehat{f}_t(S, A, u_i)$. Furthermore, selecting a uniform distribution naturally facilitates mean estimation, as $\mathbb{E}[Z(s, a)] = \int_0^1 f(s, a, u)du = \mathbb{E}[f(s, a, U)]$.

Algorithm 1 Deep Quantile Process regression-based OPE (DQPOPE)

1: **Initialize:** DNN class $\mathcal{F}$, $\widehat{f}_0 \in \mathcal{F}$, datasets $\{\mathcal{D}_t\}_{t=1}^T$, target policy π.

2: **for** $t = 1$ to T **do**

3: Collect sample $(s_i, a_i, r_i, s_i') \in \mathcal{D}_t$, and sample quantile level $\tau_i \sim \text{Unif}(0,1)$ for all (s_i, a_i).

4: Generate target sample from $\widehat{\eta}_{t-1}(s', a')$: Sample $u_i \sim \text{Unif}(0,1)$ for each (s_i', a_i') with $a_i' \sim \pi(\cdot|s_i')$, and plug (s_i', a_i', u_i) into $\widehat{f}_{t-1}(s', a', U)$

5: Compute target sample: $y_i \leftarrow r_i + \gamma \widehat{f}_{t-1}(s_i', a_i', u_i)$.

6: Update: $\widehat{f}_t \leftarrow \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}_t|} \sum_{(s_i,a_i,r_i,s_i') \in \mathcal{D}_t} \rho_{\tau_i}\big(y_i - f(s_i, a_i, \tau_i)\big)$.

7: **end for**

8: **Output:** $\widehat{f}_T(s, a, \tau)$ (i.e., $\widehat{\eta}_T(s, a)$).

space $\mathcal{F}$ is typically represented by Deep Neural Networks (DNNs). The iterative nature of this procedure is summarized in Algorithm 1. Figure 1 provides a schematic representation of this dynamic iterative process, demonstrating the equivalence between distributional Bellman updates in DRL and the quantile process regression training procedure.

3.2 Advantages of Quantile Process Regression over Discrete Quantile Estimation

To more clearly illustrate the advantages of introducing quantile process regression, we compare our method with previous QDRL approaches, such as QR-DQN (Dabney et al., 2018b) and related approaches (Rowland et al., 2024a). QDRL methods aim to estimate the conditional τ-th quantile of Y_t given X by minimizing the population risk

$$\mathcal{L}_{t,\tau}(f) = \mathbb{E}_{X,Y_t}\big(\rho_\tau(Y_t - f(X))\big), \tag{5}$$

where we recall that $X = (S, A)$ and the target response is $Y_t = R + \gamma Z_{t-1}(S', A')$. To implement this, prior methods estimate multiple quantiles at a set of fixed levels $\{\tau_i\}_{i=1}^m$ by

minimizing the aggregate loss $\sum_{i=1}^{m} \widehat{\mathcal{L}}_{t,\tau_i}(f)$, where $\widehat{\mathcal{L}}_{t,\tau_i}(\cdot)$ is some empirical approximation to (5). Importantly, in these methods, the quantile levels $\{\tau_i\}_{i=1}^{m}$ are pre-determined and remain fixed throughout the learning process. In contrast, our quantile regression-based method introduces a fundamental innovation by embedding the quantile level τ directly as model input. This design enables the model to learn the continuous representation of the quantile function, producing estimates for any quantile level $\tau \in (0, 1)$.

Addressing pseudo sample issue. In the empirical formulation $\widehat{\mathcal{L}}_{t,\tau_i}(\cdot)$ of (5), a major challenge arises due to the lack of direct access to the true distribution of Z_{t-1}. Since $f(x)$ cannot fully represent a quantile process that reconstructs the original distribution, it is impossible to generate exact samples for the target response Y_t. To address this, previous QDRL methods rely on a "pseudo-sample" construction, where $y_i^p = r_i + \gamma \widehat{f}_{t-1,\tau_i}(s_i', a_i')$. Here, $\widehat{f}_{t-1,\tau_i}$ is the estimated conditional τ_i-th quantile at the previous step $t-1$. These pseudo-samples are then used to approximate Y_t with a mixture of Dirac distributions $\frac{1}{m}\sum_{i=1}^{m} \delta_{y_i^p}$. While this approximation becomes exact as $m \to \infty$, finite m introduces unavoidable representation errors, making it challenging to fully recover the target response Y_t. This limitation is reflected in the empirical performance of QDRL methods, where performance degradation is more pronounced for small m, while increasing m significantly raises computational costs, especially in complex environments. Besides, this discretization issue also complicates the convergence analysis of these methods (Bellemare et al., 2023).

In contrast, our method resolves this issue through quantile process regression, which acts as a "generator." By embedding the quantile level τ as an input and treating it as a continuous random variable, our method directly learns the full quantile function, enabling yields of the exact samples for target response Y_t without requiring additional imputation or discretization [3]. This approach eliminates the need for pseudo-samples and

[3]Previous work use an imputation step, which imposes extra computational burden but fails to generate

enables a seamless representation of the full return distribution, addressing the limitations of discretized quantile methods by prior QDRL methods.

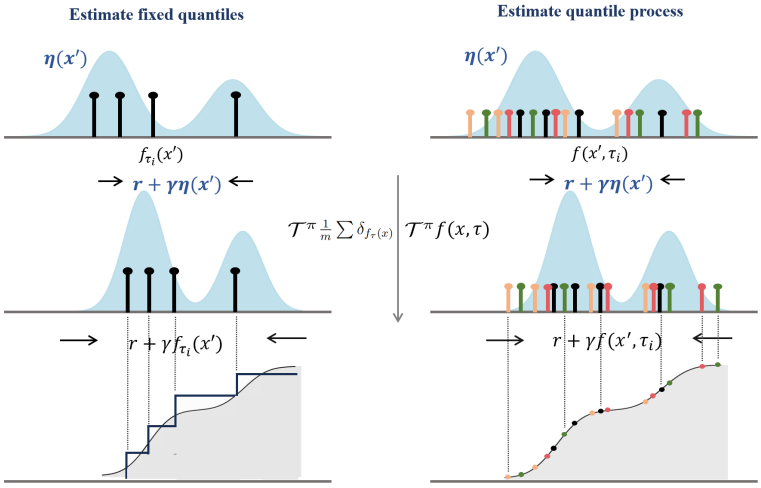

Figure 2: Illustration of estimating quantile process effectively captures the mapping of the distributional Bellman operator. The blue areas symbolize the return distribution, and the markers denote the quantile estimations at certain quantile levels. The bottom line compares the CDF of the true return distribution and the approximated ones

Key advantages of quantile process regression. From a practical perspective, a key advantage of learning a quantile process is transforming an infinite-dimensional distribution-learning problem into a finite-dimensional regression task. By embedding the quantile level as a model input, our method effectively captures the behavior of the distributional Bellman operator. This innovation ensures that the theoretical framework of distributional RL aligns with its practical implementation.

To illustrate this, Figure 2 contrasts traditional QDRL methods with quantile process regression-based approaches. The left panel depicts one-step updates for fixed quantile levels, where the estimates are discrete and susceptible to representation gaps. In contrast, the right panel demonstrates the estimation of quantiles at continuously sampled levels,

exact samples, as discussed in Section A.4 of Supplementary Material.

progressively recovering the entire continuous quantile function over τ during training. This enables quantile process regression-based methods to update the entire distribution seamlessly by operating directly on its quantiles, effectively bridging the gap between theoretical distributional Bellman operators and practical implementation.

4 Theoretical Results

In this section, we present non-asymptotic statistical guarantees for DQPOPE when implemented with DNNs. Building on previous studies in deep nonparametric regression (Schmidt-Hieber, 2020; Farrell et al., 2021), our analysis focuses on neural networks with rectified linear unit (ReLU) activation. While our theoretical analysis is grounded in the ReLU setting, it can be extended to more general DNN architectures.

Definition 4.1 (ReLU network)**.** *The ReLU network $f : \mathbb{R}^{m_0} \to \mathbb{R}^{m_{L+1}}$ is defined by*

$$f(x) = \phi_L(\sigma(\phi_{L-1}(\cdots \sigma(\phi_0(x))))), \tag{6}$$

where $\sigma(x) = \max\{x, 0\}$ is the ReLU activation function, $\phi_\ell(x) = A_\ell x + b_\ell$, and $A_\ell \in \mathbb{R}^{m_l \times m_{\ell-1}}, b_\ell \in \mathbb{R}^{m_\ell}$ are the weight matrix and bias vector in ℓ-th layer, respectively. Particularly, we consider the first layer width $m_0 = d$ and the last $m_{L+1} = 1$, the maximum width $W = \max_{\ell \in [L]} m_\ell$, and the sup-norm of the function $\|f\|_\infty \leq F$. We denote the class of such functions by $\mathcal{F} := \mathcal{F}(W, L)$.

Recent studies (Fan et al., 2020; Nguyen-Tang et al., 2022; Ji et al., 2023) have investigated RL algorithms using ReLU neural networks, where the Bellman target is assumed to belong to the Hölder or Besov space. To capture the smoothness of the Bellman operator $\mathcal{T}^\pi$, we define the Hölder class $\mathcal{G} := \mathcal{G}([0, 1]^d, \beta, H)$ as follows.

Definition 4.2 (Hölder class)**.** *Let $\beta = r + s$, where $s \in \mathbb{N}, r \in (0, 1]$. The class of Hölder smooth functions is defined by*

$$\mathcal{G}([0,1]^d, \beta, H) = \left\{ f : [0,1]^d \to \mathbb{R} : \max_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_1 \leq s} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \max_{\|\boldsymbol{\alpha}\|_1 = s} \sup_{x \neq y} \frac{|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}}(y)|}{\|x - y\|_2^r} \leq H \right\}, \quad (7)$$

where $H > 0$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{N}^d$, and $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \ldots \partial^{\alpha_d}$ is the multi-index notation.

Throughout this section, write $\widetilde{\mu} := \mu \times \text{Unif}(0, 1)$ as the product measure of μ and $\text{Unif}(0, 1)$. The following technical assumptions are introduced to support analysis.

Assumption 4.3 (Coverage)**.** *Given d^π and μ in Section 2, there exists a constant C_μ such that*

$$\sup_{\eta, \eta' \in \Delta(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}} \frac{\overline{\mathcal{W}}_{p, d^\pi}\big(\eta, \mathcal{T}^\pi \eta'\big)}{\overline{\mathcal{W}}_{p, \mu}\big(\eta, \mathcal{T}^\pi \eta'\big)} \leq C_\mu.$$

Assumption 4.4 (Bellman completeness)**.** *We assume that for any $\eta \in \Delta(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}$, if its corresponding quantile function f belongs to $\mathcal{F}$, the quantile function of $\mathcal{T}^\pi \eta$ belongs to $\mathcal{G}$.*

The data coverage and completeness assumptions are standard and widely employed in RL theory literature (Munos and Szepesvári, 2008; Chen and Jiang, 2019). Unlike the classical data coverage, which bounds the distribution ratio $\|\frac{d^\pi}{\mu}\|_\infty := \sup_{s,a} \frac{d^\pi(s,a)}{\mu(s,a)}$ across state-action pairs (s, a), we measure how well Bellman errors transfer between the distributions d^π and μ, offering a tighter measure than $\|d^\pi / \mu\|_\infty$. Even though two distributions d^π and μ are substantially disparate, this discrepancy can still be effectively quantified. The completeness assumption specifies that if quantile function of any $\eta \in \Delta(\mathbb{R})^{\mathcal{S} \times \mathcal{A}}$ belongs to $\mathcal{F}$, then the Bellman operator $\mathcal{T}^\pi$ applied on η results in the quantile function of $\mathcal{T}^\pi \eta$ sitting in $\mathcal{G}$. This assumption is mild and holds for most common smooth dynamics with concrete examples elucidated in Fan et al. (2020). Please refer to Section G of the Supplementary Material for a detailed justification.

Assumption 4.5 (Strong convexity). *There exists a universal constant $c_0 > 0$ such that for any $t \in [T]$ and any function $f \in \mathcal{F}$, we have*

$$\mathcal{L}_t(f) - \mathcal{L}_t(f_t^*) \ \geq \ c_0 \ \|f - f_t^*\|_{2,\widetilde{\mu}}^2.$$

In contrast to the previous two standard assumptions in RL literature, Assumption 4.5 introduces a c_0-strong convexity condition of the population risk of quantile loss. Under a mild Assumption G.1, which requires the density of conditional distribution Y_t given (s, a) near f_t^* to be bounded away from zero, one can ensure that Assumption 4.5 always holds. It is worth mentioning that Assumption 4.5 is crucial for establishing Lemma 4.7. This convexity property is widely used in non-parametric quantile regression literature (Belloni and Chernozhukov, 2011). See Section G of the Supplementary Material for more details.

4.1 Preliminary Results

This part introduces the analytical framework for distributional OPE and presents the preliminary results of DQPOPE. We assume $\sup_{t\geq 0} \ |R_t| \leq R_{max}$, a common assumption in the RL literature for simplifying the analysis. This condition is not essential and can be relaxed to more general assumptions, such as sub-Gaussian tails. We provide a detailed discussion in Section G.1 of the Supplementary Material. Let $C_{F,R} = F + R_{max}$ and define $\widehat{\varepsilon}_{p,t} := \overline{\mathcal{W}}_{p,\mu}(\widehat{\eta}_t, \mathcal{T}^\pi \widehat{\eta}_{t-1})$ for each $t \in [T]$, where $\{\widehat{\eta}_t\}_{t\in[T]}$ are obtained by Algorithm 1.

To the aim of bounding $\mathcal{W}_p(\boldsymbol{\eta}^\pi, \widehat{\boldsymbol{\eta}})$, we start with the following decomposition.

Lemma 4.6 (Sub-optimality decomposition). *Suppose that Assumption 4.3 is satisfied. Then, the sub-optimality of $\widehat{\boldsymbol{\eta}}_T$ satisfies*

$$\mathcal{W}_p(\boldsymbol{\eta}^\pi, \widehat{\boldsymbol{\eta}}_T) \ \leq \ \frac{2C_\mu^{\frac{1}{2p}}}{(1-\gamma)^{\frac{3}{2}}} \ \max_{0<t\leq T} \widehat{\varepsilon}_{p,t} + \frac{\gamma^{\frac{T}{2}}}{(1-\gamma)^{\frac{3}{2}}} \ C_{F,R}$$

Lemma 4.6 is motivated by the error propagation analysis from the RL literature (Antos et al., 2008; Wu et al., 2023). This error propagation illustrates that the total error of distributional OPE can be interpreted as a sum of statistical error and algorithmic error. The statistical error arises from the one-step Bellman error, $\widehat{\varepsilon}_{p,t}$, which can be explicitly associated with the excess risk of the quantile loss by applying Lemma 4.7. The algorithmic error is specific to the iterative nature of the RL setting and has no counterpart in regression settings. It reflects the error that remains after executing a finite number of iterations T. In conclusion, the key component in the analysis hinges on the one-step Bellman error, $\widehat{\varepsilon}_{p,t}$, as it ultimately controls the error propagation.

For $p = 1$, the one-step Bellman error can be bounded by excess risk, stated as follows.

Lemma 4.7. *Suppose Assumption 4.5 is satisfied. Then for each $t \in [T]$, we have*

$$\widehat{\varepsilon}_{1,t} = \overline{\mathcal{W}}_{1,\mu}(\widehat{\eta}_t, \mathcal{T}^\pi \widehat{\eta}_{t-1}) \;\le\; c_0^{-\frac{1}{2}} \left(\mathcal{L}_t(\widehat{f}_t) - \mathcal{L}_t(f_t^*)\right)^{\frac{1}{2}},$$

where $c_0 > 0$ is a constant introduced in Assumption 4.5.

Lemma 4.7 provides a key step in linking the bound for the Wasserstein metric to excess risk, enabling the analysis of distributional OPE in a non-parametric regression manner. Combined with Lemma 4.6, it reduces the problem to bound excess risk for each step.

Theorem 4.8 (Excess risk bound, slow rate)**.** *Suppose Assumption 4.4 is satisfied. With probability at least $1 - 2n^{-1}$, the excess risk satisfies*

$$\mathcal{L}_t(\widehat{f}_t) - \mathcal{L}_t(f_t^*) \;\le\; C\sqrt{\frac{W^2 L^2 \log(W^2 L) \log n}{n}} + \inf_{f \in \mathcal{F}} \left(\mathcal{L}_t(f) - \mathcal{L}_t(f_t^*)\right), \tag{8}$$

where C is a constant independent of W, L, n. Furthermore, for sufficiently large $U, V \in \mathbb{N}^+$, setting width and length to be $W = \mathcal{O}\left((s+1)^2 d^{s+1} U \log U\right)$ and $L = \mathcal{O}\left((s+1)^2 V \log V\right)$, if we choose $UV = \lfloor n^{\frac{d}{4\beta+2d}} \rfloor$, when n is sufficiently large, with probability at least $1 - 2n^{-1}$,

the excess risk has upper bound that

$$\mathcal{L}_t(\widehat{f}_t) - \mathcal{L}_t(f_t^*) \;\leq\; C(s+1)^4 d^{s+(\frac{\beta}{2}\vee 1)} (\log n)^3 n^{-\frac{\beta}{2\beta+d}}, \tag{9}$$

where C is a constant independent of s, β, d, n.

The bound in (8) tracks the usual decomposition into a stochastic term and an approximation term. The first term reflects estimation error in nonparametric quantile process regression-based on ReLU networks, whereas the second term captures the approximation bias induced by restricting the estimator to the class $\mathcal{F}$. Combined with Lemma 4.7, which translates excess risk into one-step Bellman error, this yields a slower rate of order $n^{-1/4}$, in contrast to the $n^{-1/2}$ rate typically obtained in standard OPE (Chen and Jiang, 2019).

The approximation term quantifies the error incurred by approximating the Hölder class $\mathcal{G}$ with the ReLU class $\mathcal{F}$. it is closely related to the *inherent Bellman error* (Munos and Szepesvári, 2008), $\sup_{g\in\mathcal{G}} \inf_{f\in\mathcal{F}} \|g - f\|_{1,\widetilde{\mu}}$, which reflects how well the function class $\mathcal{F}$ is aligned with the Bellman image of the target class. By utilizing approximation theory (Lemma H.3), $\inf_{f\in\mathcal{F}}(\mathcal{L}_t(f) - \mathcal{L}_t(f_t^*)) \leq \inf_{f\in\mathcal{F}} \|f - f_t^*\|_{1,\widetilde{\mu}} \leq C(UV)^{-\frac{2\beta}{d}}$. It is also worth mentioning that the control of approximation error can be further refined with additional smoothness conditions. We will provide a tighter bound in Theorem 4.11.

The stochastic error characterizes the variance in estimating the quantile functions, depending on both the richness of network and sample size. A smaller network reduces variance but increases approximation bias, whereas a larger network improves expressivity at the price of a larger stochastic error. The scaling of W and L in Theorem 4.8 balances these two effects and leads to the convergence rate $n^{-\frac{\beta}{2\beta+d}}$.

Theorem 4.9. *Suppose Assumptions 4.4, and 4.5 are satisfied. For each $t \in [T]$, using the same choice of L and W as in Theorem 4.8, when n is sufficiently large, with probability*

at least $1-2n^{-1}$*, the one-step Bellman error has upper bound that*

$$\overline{\mathcal{W}}_{1,\mu}(\widehat{\eta}_t, \mathcal{T}^\pi \widehat{\eta}_{t-1}) \;\le\; C(s+1)^2 d^{s/2+(\frac{\beta}{4}\vee\frac{1}{2})} (\log n)^{\frac{3}{2}} n^{-\frac{\beta}{4\beta+2d}},$$

where C *is a constant independent of* s, β, d, n*. If Assumption 4.3 further holds and* $T = \mathcal{O}(\xi \log N)$ *for some constant* $\xi > 0$*, when* N *is sufficiently large, with probability at least* $1 - cN^{-1}(\log N)^2$*, the sub-optimality of* $\widehat{\boldsymbol{\eta}}_T$ *has upper bound that*

$$\mathcal{W}_1(\boldsymbol{\eta}^\pi, \widehat{\boldsymbol{\eta}}_T) \;\le\; \frac{C C_\mu^{\frac{1}{2}}}{(1-\gamma)^{\frac{3}{2}}} (\log N)^{\frac{5}{2}} N^{-\frac{\beta}{4\beta+2d}} + \frac{N^{\frac{\xi \log \gamma}{2}}}{(1-\gamma)^{\frac{3}{2}}}\, C_{F,R}, \tag{10}$$

where $0 < \gamma < 1$*, and* c, C *are constants independent of* C_μ, N, γ*.*

Theorem 4.9 follows by combining the excess risk bound with the control of the one-step Bellman error and then propagating the error across T iterations, with the substitution $n = N/T$. The constant C_μ arises from the change-of-measure argument used to propagate the error across iterations, and the multiplier $(1-\gamma)^{-3/2}$ reflects the impact of the metric $\overline{\mathcal{W}}_{p,\mu}(\cdot,\cdot)$ on the performance of the distribution iteration. For clarity, we suppress in C its dependence on the Hölder smoothness parameters s, β and the dimension d.

The non-asymptotic statistical error bound attains $N^{-\frac{\beta}{4\beta+2d}}$, which is slightly slower than $N^{-\frac{\beta}{2\beta+2d}}$ (Nguyen-Tang et al., 2022) with respect to β, and slower than $N^{-\frac{\beta}{2\beta+d}}$ (Fan et al., 2020; Ji et al., 2023) with respect to both β and d. Here, β represents the smooth parameter of the Bellman operator, and d is the input dimension. Consequently, this error bound does not achieve the best possible rate within the OPE setting. The slower rate can be intuitively attributed to the increased complexity of learning a distribution rather than a scalar. Technically, it stems from a coarse error decomposition of excess risk and the reliance on Rademacher complexity to control the stochastic error. This motivates a sharper localized analysis. Furthermore, we extend our results beyond the i.i.d. setting in Section H of the Supplementary Material.

Remark 1. *Note that $N = nT$ is the total sample size across T steps, where the choice of $T = \mathcal{O}(\xi \log N)$ results from data splitting strategy. Since the second term in* (10) *is polynomial dependent on N as $\gamma^{\frac{T}{2}} = \gamma^{\frac{\xi \log N}{2}} = N^{\frac{\xi \log \gamma}{2}}$, we require $\xi \geq \frac{2\beta}{(4\beta+d)\log(1/\gamma)}$ to ensure that $\gamma^{\frac{\xi \log N}{2}} \lesssim N^{-\frac{\beta}{4\beta+2d}}$ in* (10)*. Data splitting removes the correlation between $\widehat{f}_t$ and data that may arise from reusing the same sample batch. Although this simplifies the theoretical analysis, we do not apply data splitting in our experiments. While a more refined analysis could potentially eliminate the need for this technique, we leave this for future work.*

4.2 Fast Rate of Excess Risk Bound

In this subsection, we derive a sharper excess risk bound. Recent studies (Rowland et al., 2024b; Zhang et al., 2025) reveal that estimating the return distribution requires no more samples than estimating the mean value. However, these works come with trade-offs in analysis. In particular, they focus on categorical distributional temporal difference (TD) learning, which approximates the return distribution with a discrete distribution. Their analysis relies on access to a generative model and is confined to the tabular setting, which is impractical for modern RL applications involving large state spaces and limited offline data. This highlights the importance of showing the sample complexity results of distributional OPE that are comparable to those available for standard OPE (Nguyen-Tang et al., 2022; Ji et al., 2023) under more realistic settings. Our goal is to show that the excess risk bound can attain an optimal rate of $N^{-\frac{2\beta}{2\beta+d}}$, consistent with theoretical expectations.

Achieving such a fast rate is substantially more delicate for quantile regression than for squared-loss regression. Recent work on non-parametric quantile regression with ReLU networks, such as Padilla et al. (2022), show that the prediction error of conditional quantile estimator at the quantile level $\tau = 0.5$ can achieve the minimax rate as shown in Schmidt-

Hieber (2020), given Gaussian errors and uniformly distributed covariates in $[0,1]^d$; Shen et al. (2024) establish a minimax lower bound under a strong convexity condition (Assumption 4.5), whereas the corresponding upper bound remains slower than the minimax rate. To derive a faster rate, one may expect that the excess risk of quantile loss could exhibit some nice local quadratic structure, analogous to that of squared loss.

Assumption 4.10 (Local strong convexity and smoothness)**.** *There exist two universal constants $c_0' \geq c_0 > 0$ such that for any $t \in [T]$, and for any $f \in \mathcal{F}$ satisfying $\|f - f_t^*\|_{2,\widetilde{\mu}}^2 \leq b_n$, where $b_n = C_b n^{-\frac{2\beta}{2\beta+d}}$ with C_b being some constant independent of n, we have*

$$c_0\|f - f_t^*\|_{2,\widetilde{\mu}}^2 \;\leq\; \mathcal{L}_t(f) - \mathcal{L}_t(f_t^*) \;\leq\; c_0'\|f - f_t^*\|_{2,\widetilde{\mu}}^2. \tag{11}$$

Assumption 4.10 further imposes a smoothness condition by the RHS inequality, implying that the population risk of quantile loss maintains curvature characteristics similar to the squared loss nearby around the target quantile function f_t^*. For squared loss, it naturally holds with $c_0 = c_0' = 1$. Intuitively, this assumption establishes a pivotal link between quantile loss and squared loss in the neighborhood around the target function f_t^*, allowing a faster convergence rate for excess risk. Technically, a similar assumption is also employed by Farrell et al. (2021), where condition (11) is required for any $f \in \mathcal{F}$. However, our analysis relaxes this condition by requiring (11) only within a neighborhood around the target function f_t^*, where the radius of the neighborhood is controlled by a shrinking sequence b_n. For additional details, refer to Remark E.1 in the Supplementary Material.

Remark 2. *Similar to Assumption 4.5, Assumption 4.10 holds under a mild Assumption G.1 in the Supplementary Material. Specifically, the smoothness condition requires that the density of conditional distribution Y_t given (s,a) near f_t^* is bounded.*

We assume the existence of $f_{t,\mathcal{F}} = \operatorname{argmin}_{f\in\mathcal{F}}\|f - f_t^*\|_{2,\widetilde{\mu}}^2$ for each $t \in [T]$ in the rest of

main text, which is a common requirement in the relevant literature (Farrell et al., 2021). We establish the fast rate of excess risk bound in Theorem 4.11.

Theorem 4.11 (Excess risk bound, fast rate)**.** *Suppose Assumption 4.4 and 4.10 are satisfied. With probability at least* $1-c\exp(-W^2L^2\log(W^2L)\log n)$*, the excess risk satisfies*

$$\mathcal{L}_t(\widehat{f}_t) - \mathcal{L}_t(f_t^*) \;\le\; C\;\frac{W^2L^2\log(W^2L)\log n}{n} + 2\;\mathcal{A}_t, \tag{12}$$

where c, C *are constants independent of* W, L, n*, and* $\mathcal{A}_t := \|f_{t,\mathcal{F}} - f_t^*\|_{2,\widetilde{\mu}}^2$*. Furthermore, using the same choice of* L *and* W *as in Theorem 4.8, when* n *is sufficiently large, with probability at least* $1 - c\exp(-n^{\frac{2d}{2d+4\beta}}\log n)$*, the excess risk has upper bound that*

$$\mathcal{L}_t(\widehat{f}_t) - \mathcal{L}_t(f_t^*) \;\le\; C\;(s+1)^8 d^{2s+(\beta\vee 2)}\;(\log n)^6 n^{-\frac{2\beta}{2\beta+d}},$$

where c, C *are constants independent of* s, β, d, n*.*

In (12), the first term of the RHS is derived to control the stochastic error, and the second term measures the approximation error. For fixed width W and length L, the stochastic error scales as n^{-1}, an improvement over $n^{-1/2}$ in Theorem 4.8. By selecting the appropriate width W and length L, both errors scale as $n^{-\frac{2\beta}{2\beta+d}}$. Theorem 4.11 thus yields a faster rate, which attains the minimax rate $n^{-\frac{2\beta}{2\beta+d}}$ as established in Stone (1982) for the d-dimensional non-parametric regression function with smoothness index β.

Theorem 4.12. *Suppose Assumptions 4.4, and 4.10 are satisfied. For each* $t \in [T]$*, using the same choice of* L *and* W *as in Theorem 4.8, when* n *is sufficiently large, with probability at least* $1 - c\exp(-n^{\frac{2d}{2d+4\beta}}\log n)$*, the one-step Bellman error has upper bound that*

$$\overline{\mathcal{W}}_{1,\mu}(\widehat{\eta}_t, \mathcal{T}^\pi\widehat{\eta}_{t-1}) \;\le\; C\;(s+1)^4 d^{s+(\frac{\beta}{2}\vee 1)}(\log n)^3 n^{-\frac{\beta}{2\beta+d}},$$

where c, C *are constants independent of* s, β, d, n*. If Assumption 4.3 further holds and* $T = \mathcal{O}(\xi\log N)$ *for some constant* $\xi > 0$*, when* N *is sufficiently large, with probability at*

least $1 - c \log N \exp(-(N/\log N)^{\frac{2d}{2d+4\beta}})$*, the sub-optimality of* $\widehat{\boldsymbol{\eta}}_T$ *has an upper bound that*

$$\mathcal{W}_1(\boldsymbol{\eta}^\pi, \widehat{\boldsymbol{\eta}}_T) \;\leq\; \frac{C C_\mu^{\frac{1}{2}}}{(1-\gamma)^{\frac{3}{2}}} \, (\log N)^4 N^{-\frac{\beta}{2\beta+d}} + \frac{N^{\frac{\xi \log \gamma}{2}} \, C_{F,R}}{(1-\gamma)^{\frac{3}{2}}}, \tag{13}$$

where $0 < \gamma < 1$*, and* c, C *are constants independent of* C_μ, N, γ*.*

To ensure that the error rate matches $N^{-\frac{\beta}{2\beta+d}}$, the second term in (13) must be dominated by the first. This requires $N^{\frac{\xi \log \gamma}{2}} \lesssim N^{-\frac{\beta}{2\beta+d}}$, which holds whenever $\xi \geq \frac{2\beta}{(2\beta+d)\log(1/\gamma)}$. The non-asymptotic statistical error bound of sub-optimality attains $N^{-\frac{\beta}{2\beta+d}}$. Relative to existing results for standard OPE, this rate is faster than $N^{-\frac{\beta}{2\beta+2d}}$ (Nguyen-Tang et al., 2022) in its dependence on d, and is comparable to $N^{-\frac{\beta}{2\beta+d}}$ (Ji et al., 2023) in its dependence on both β and d. For a pre-specified error ϵ, the DQPOPE requires a sample complexity of $\mathcal{O}\big((1-\gamma)^{-(3+\frac{3d}{2\beta})} C_\mu^{1+\frac{d}{2\beta}} \epsilon^{-(2+\frac{d}{\beta})}\big)$. Compared to $\mathcal{O}\big((1-\gamma)^{-(4+2\frac{d}{\beta})} C_\mu^{1+\frac{d}{2\beta}} \epsilon^{-(2+\frac{d}{\beta})}\big)$ by Ji et al. (2023), the sample complexity of DQPOPE has the same dependence on the distribution shift constant C_μ and the pre-specified error ϵ, while exhibiting a milder dependence on the horizon due to data splitting. Compared to $\mathcal{O}\big((1-\gamma)^{-(2+2\frac{d}{\beta})} \kappa^{1+\frac{d}{\beta}} \epsilon^{-(2+2\frac{d}{\beta})}\big)$ by Nguyen-Tang et al. (2022) where κ is the bound on distribution ratio $\|d^\pi/\mu\|_\infty$, our result has a weaker dependence on ϵ and κ, as C_μ is often substantially smaller than κ. Our result could achieve comparable sample efficiency as long as $(\epsilon/(1-\gamma)^{\frac{1}{2}})^{\frac{d}{\beta}} \leq 1$. Additionally, our result has a stronger dependence on the horizon when $\beta/d > 1/2$. This difference is attributable to the fact that the distributional Bellman operator contracts at a rate $\gamma^{1-\frac{1}{2p}}$ under $\overline{\mathcal{W}}_{p,d^\pi}(\cdot,\cdot)$ metric, whereas the standard Bellman operator contracts at a rate γ under $\|\cdot\|_\infty$ metric.

Estimating the full return distribution is inherently more challenging than its mean, as demonstrated by the fact that $|\mathbb{E}Z_1 - \mathbb{E}Z_2| \leq \mathcal{W}_p(\nu_1, \nu_2)$ for $p \geq 1$, where $Z_1 \sim \nu_1$ and $Z_2 \sim \nu_2$. Despite increased complexity, DQPOPE learns the entire quantile curve, capturing the full distributional information without sacrificing the convergence rate. Notably, our results are the first to show that distributional OPE with ReLU neural network approximation

achieves sample efficiency comparable to standard OPE. Compared to model-based DRL in the tabular case Zhang et al. (2025); Rowland et al. (2024b), our analysis aligns closely with the most practical model-free QDRL algorithms Dabney et al. (2018b,a).

4.3 Estimating Policy Value through Quantile Process

The preceding analysis characterizes the statistical error of the estimated return distribution. We next show that the learned quantile process also yields a natural estimator of the policy value, which remains a central target in practice for guiding decision-making. Recall in Algorithm 1, the quantile process $\widehat{f}_t(x,\tau)$ is used as a generator by sampling $\tau_i \sim \mathrm{Unif}(0,1)$ and plugging it into $\widehat{f}_t(x,\tau)$ to recover samples from the estimated return distribution $\widehat{\eta}_t(x)$ for a given x. Given quantile levels $\{\tau_k\}_{k=1}^K$, the sample average $\frac{1}{K}\sum_{k=1}^K \widehat{f}_t(x,\tau_k)$ provides a natural estimator of the value function at x.

Proposition 4.13. *For the policy value estimator $\widehat{V}_K = \frac{1}{K}\sum_{k=1}^K \widehat{f}_T(s_{0,k}, a_{0,k}, \tau_k)$, where $s_{0,k} \sim \rho, a_{0,k} \sim \pi(\cdot|s_{0,k})$ and $\tau_k \sim \mathrm{Unif}(0,1)$, under the same assumptions and network size in Theorem 4.12, if $T = \mathcal{O}(\xi \log N)$ with some constant $\xi > 0$ and sufficiently large N, the following bound holds with probability at least $1 - c\log N \exp(-(N/\log N)^{\frac{2d}{2d+4\beta}}) - K^{-1}$,*

$$\left|\widehat{V}_K - V^\pi\right| \;\le\; \frac{C C_\mu^{\frac{1}{2}}}{(1-\gamma)^{\frac{3}{2}}}(\log N)^4 N^{-\frac{\beta}{2\beta+d}} + CF\sqrt{\frac{\log K}{K}} + \frac{N^{\frac{\xi\log\gamma}{2}} C_{F,R}}{(1-\gamma)^{\frac{3}{2}}}, \tag{14}$$

where c, C is a constant independent of C_μ, N, K, γ.

Proposition 4.13 follows from the decomposition $|\widehat{V}_K - V^\pi| \le |\widehat{V}_K - \widehat{V}| + |\widehat{V} - V^\pi|$, where $\widehat{V} = \mathbb{E}_{Z\sim\widehat{\boldsymbol{\eta}}_{\boldsymbol{T}}}[Z]$ is expectation from the output of Algorithm 1 and $V^\pi = \mathbb{E}_{Z\sim\boldsymbol{\eta}^{\boldsymbol{\pi}}}[Z]$ is the target value. The first term on the RHS of (14) represents the finite sample error that can be directly related to the result in Theorem 4.12, and the second term reflects the computational error arising from the empirical average over K draws, which scales as

$\sqrt{1/K}$. Moreover, choosing $K \geq CN^{\frac{2\beta}{2\beta+d}}(\log N)^{-6}$ and $\xi \geq \frac{2\beta}{(2\beta+d)\log(1/\gamma)}$ ensures that the last two terms in (14) are dominated by the first term. Thus, the resulting bound for policy value estimation ultimately attains $N^{-\frac{\beta}{2\beta+d}}$ up to log factors, matching the optimal rate in the standard OPE setting (Fan et al., 2020; Ji et al., 2023).

A key benefit of estimating policy value through a quantile process is the robustness gained from averaging multiple quantiles compared to directly estimating a single distribution mean. In particular, quantile loss is less sensitive to outliers, as it does not disproportionately penalize large deviations, and its gradient scales as $\mathcal{O}(1)$. By contrast, squared loss places disproportionate weight on extreme observations, which can make optimization more sensitive to heavy tails and atypical rewards.

5 Experiments

This section validates the theoretical analysis of DQPOPE across various scenarios. Section 5.1 presents experiments that validate the sample complexity results in Theorem 4.12. Sections 5.2 and 5.3 demonstrate the advantage of DQPOPE in estimating policy values via quantile averaging, leveraging both a simple one-step toy example and a real-world dataset. Throughout these experiments, we compare DQPOPE with value-based OPE [4] implemented using deep ReLU networks (referred to as DOPE). Additional experimental details are provided in Section J of the Supplementary Material.

5.1 Sample Complexity Analysis in CartPole

To provide a clearer understanding of the sample complexity results in Theorem 4.12, we conducted simulation studies in the CartPole environment. The target policy was obtained

[4] Value-based OPE estimates the value function Q^π by minimizing squared loss (see Section I for details).

by training a DQN agent(Mnih et al., 2015) for 10,000 update steps. The target return from the initial state was estimated using 1,000 Monte Carlo (MC) rollouts with discounted cumulative rewards ($\gamma = 0.99$) for each rollout, which is visualized in Figure 3 (b).

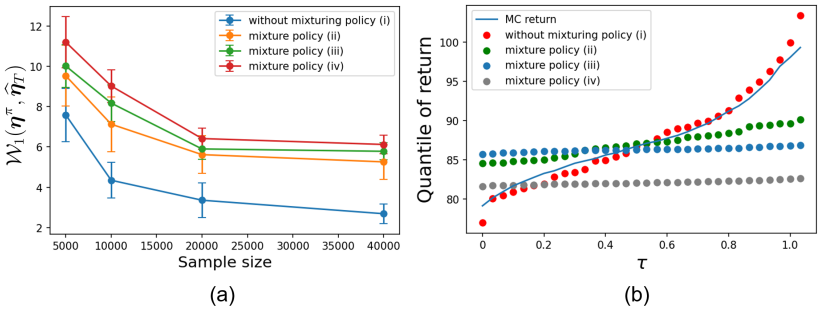

Figure 3: (a) The performance metric versus different sample size. (b) Quantile estimation performance, where the red dots represent the estimated quantiles without mixture policy, other colored dots from the mixture policy (ii)-(iv), and the blue line represents the ground truth quantile function calculated by MC rollouts.

We consider four data-generation schemes: (i) data collected entirely from the target policy; (ii) data collected from a mixture policy with 80% target policy actions and 20% random; (iii) data collected from a mixture policy with 60% target policy actions; (iv) data collected from a mixture policy with 40% target policy actions. We use a three-layer fully connected network with 64 units per layer and ReLU activation. The learning rate was set to 0.0005, the batch size to 64, and the target network was updated every 15 steps.

Figure 3 (a) shows the performance metric $\mathcal{W}_1(\boldsymbol{\eta}^\pi, \widehat{\boldsymbol{\eta}}_T)$ for DQPOPE across varying sample sizes, under the four data generation schemes. In all cases, the error is consistent with a polynomial decay as the sample size increases. As expected, estimation accuracy improves with a larger proportion of data generated from the target policy, reflecting the reduced distribution shift, as quantified by C_μ. Figure 3 (b) highlights the quantile estimation performance under the different distribution shifts. When the data are generated entirely from the target policy, DQPOPE effectively captures the true return distribution.

5.2 Simulation: A One-step Toy Example

In this subsection, we compare DQPOPE and DOPE for policy value estimation within a toy environment (Figure 4(a)). To examine robustness under heavy-tailed rewards, the rewards are generated from Student's t-distributions with different degrees of freedom, thereby controlling the tail heaviness. For DQPOPE, as described in Section 4.3, the policy values are estimated by averaging over K quantile levels. We evaluate the accuracy of both methods using the Mean Squared Error (MSE).

To ensure a fair comparison, both algorithms use the same network architecture, namely a two-layer fully connected neural network with 12 hidden units per layer and ReLU activation. In DQPOPE, the quantile level is concatenated with the state and used as the network input. The training parameters are identical across both methods, with a learning rate of 0.002, a batch size of 32, and 100 update iterations (corresponding to a total sample size of 3200). Numerical results are reported in Table 1.

Table 1: Comparison of MSE ($\times 10^{-3}$) of policy value estimation between DOPE and DQPOPE, based on 100 times of replicates with standard deviations.

	DOPE	DQPOPE			
heavy-tailness		$K = 4$	$K = 8$	$K = 16$	$K = 32$
t(2)	11.3(17.9)	7.21(11.0)	3.53(7.42)	1.94(3.68)	0.93(1.13)
t(4)	6.51(9.94)	4.60(8.97)	3.71(5.66)	2.20(4.64)	0.59(1.07)
t(6)	6.73(8.14)	4.90(5.25)	3.19(4.47)	2.71(3.55)	0.78(1.26)
t(8)	4.37(9.15)	3.98(7.95)	2.91(4.53)	1.01(1.52)	0.82(1.41)
t(10)	8.05(10.4)	4.50(9.52)	3.82(5.14)	1.39(2.25)	0.61(0.95)
$\mathcal{N}(0,1)$	3.75(10.0)	4.73(9.71)	3.87(6.24)	1.41(3.26)	0.62(1.26)

Our findings reveal that DQPOPE consistently achieves lower MSE than DOPE, with

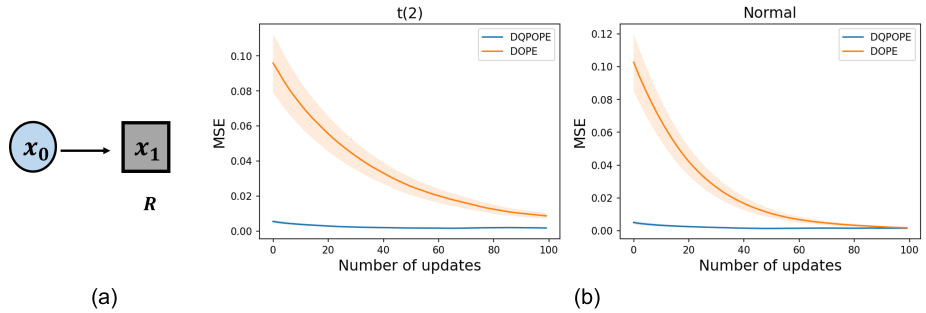

Figure 4: (a) Two state environment with the reward receiving at the terminal state x_1. (b) MSE of the policy value estimation under $t(2)$ and $\mathcal{N}(0,1)$ distribution, where each curve is computed based on 100 times replicates and shaded by their confidence intervals.

the advantage becoming more pronounced as the reward distribution becomes heavier tailed. Moreover, increasing the number of quantiles K enhances the accuracy of policy value estimation, aligning with our theoretical analysis. To further illustrate the robustness of quantile averaging estimators, Figure 4(b) visualizes the MSE against the number of updates under both the $t(2)$ and $\mathcal{N}(0,1)$ distributions. In both cases, DQPOPE exhibits significantly faster and more stable convergence.

5.3 Real Data Analysis: MIMIC-III Dataset

We next examine the performance of DQPOPE on the MIMIC-III v1.4 [5] dataset, which contains critical care records for over 40,000 patients admitted to the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2016). Our analysis focuses on **sepsis**, a high-stakes medical condition that requires sequential treatment decisions under substantial uncertainty. In particular, septic patients often require repeated administration of intravenous fluids and vasopressors to maintain hemodynamic stability. The substantial heterogeneity in patient responses and the complexity of longitudinal clinical trajectories make this setting a natural testbed for distributional OPE.

[5] https://physionet.org/content/mimiciii/1.4/

Recent studies (Raghu et al., 2017; Komorowski et al., 2018) suggest that reinforcement learning can provide useful treatment policies for sepsis management, which further motivates the use of distributional RL methods for quantifying uncertainty in such sequential decision problems. In our analysis, patient trajectories are modeled as Markov decision processes (MDPs). The detailed MDPs specification, together with additional implementation details, is provided in Section I.2 of the Supplementary Material.

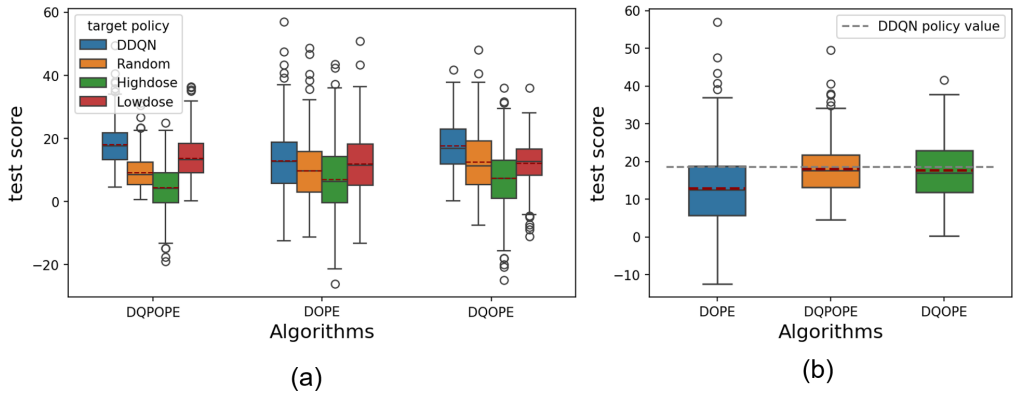

Figure 5: Boxplot of the estimated policy value for different algorithms. In each box, the red dashed line represents the mean, the central solid line indicates the median, and the bottom and top edges of the box correspond to the 25-th and 75-th percentiles, respectively. The bottom and top line outside box correspond to the lower and upper extremes. (a) Comparison of DQPOPE, DOPE and DQOPE across 4 target policies. (b) Comparison of DQPOPE, DOPE and DQOPE under DDQN target policy.

To implement the OPE experiments, we split the dataset into training (75%), validation (5%), and test (15%) sets, ensuring that each subset maintained the same proportion of surviving and non-surviving patients. We consider four target policies: (i) DDQN: A policy trained using Dueling Double Deep Q-Networks (DDQN), which serves as the benchmark policy[6]; (ii) Random: A policy with random dose selection; (iii) High-dose: A policy that

[6]DDQN is a state-of-the-art reinforcement learning (RL) method that has demonstrated significant success in scaling RL to clinical decision-making problems (Lu et al., 2020)

always administers high doses of treatment; (iv) Low-dose: A policy that always administers low doses of treatment.

We compare DQPOPE ($K = 32$), Deep Quantile-based OPE (DQOPE, $K = 32$), and DOPE under these four target policies. Here DQOPE estimates a finite set of quantiles using the QR-DQN approach (Dabney et al., 2018b), as described in Section 3. All methods are trained on the training set, and model selection is performed using the validation loss. Weighted importance sampling (WIS) is then applied on the test set to estimate the policy value for each target policy.

Figure 5 presents the estimated policy values for four target policies over 500 bootstrapped replications with 90% resampling on the test set. In Figure 5(a), the two quantile-based methods, especially DQPOPE, more clearly separate the DDQN policy from the remaining policies, with DDQN consistently receiving the highest test scores. Figure 5(b) further indicates that DQPOPE yields both more accurate and more stable value estimates than DQOPE, emphasizing the advantages of estimating the entire quantile process rather than discrete quantiles.

To further examine the policies induced by the different methods, Figure 6 displays their action-selection distributions. The policy estimated by DQPOPE aligns more closely with the target DDQN policy, with the most frequent treatment selections being (0,0) and (2,1). This pattern is clinically plausible, as many sepsis patients are not critically ill and thus do not require aggressive doses of intravenous fluids or vasopressors. In contrast, the policies estimated by DOPE and DQOPE exhibit larger deviations from the DDQN policy, leading to invalid or suboptimal recommendations.

Overall, these results illustrate the practical advantage of DQPOPE in capturing the uncertainties inherent in complex clinical decision-making scenarios. The results demonstrate

that integrating distributional methods into off-policy evaluation (OPE) can significantly improve the performance of policy value estimation. In particular, these results emphasize the strength of estimating a quantile process, which allows for more robust and accurate policy value estimation by fully capturing the behavior of the distributional Bellman operator and better accounting for the inherent randomness in MDPs.

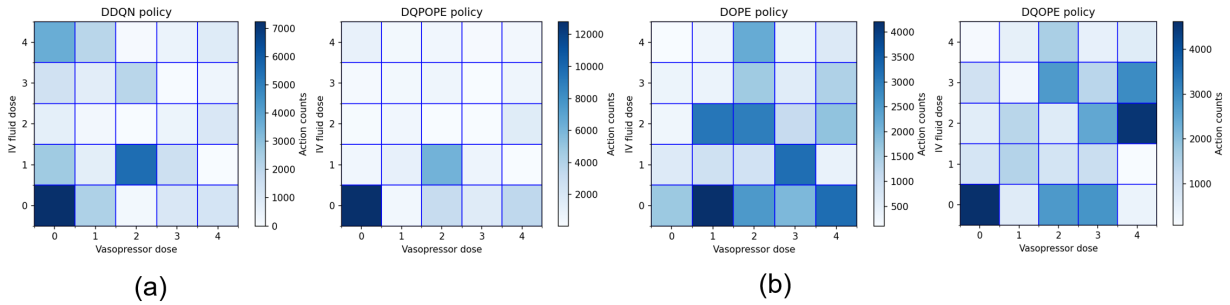

Figure 6: A 2D histogram visualizing the action selection frequency across 4 policies, with each grid representing the count of actions selected over all time steps in the test set. (a) Target policy of DDQN. (b) Estimated policies by DQPOPE, DOPE, and DQOPE.

6 Conclusion

This paper introduces a deep quantile process regression method for OPE, which incorporates the full return distribution rather than focusing solely on the expected return of a target policy. It provides novel insights into both theoretical and practical aspects. Theoretically, we develop a rigorous statistical framework for distributional OPE and establish sample complexity guarantees for DQPOPE with deep neural network approximation. We further show that estimating the full return distribution can be as sample-efficient as estimating only the mean policy value, thereby clarifying the advantage of DQPOPE over standard value-based OPE methods. From a practical perspective, we demonstrate how quantile process regression effectively implements the distributional Bellman update and

how DQPOPE addresses the pseudo-sample issue encountered in existing quantile-based DRL methods. Extensive experiments further demonstrate the empirical advantage of DQPOPE in OPE tasks.

7 Acknowledgements

The work of Qi Kuang is supported by the National Natural Science Foundation of China (Grant 12571286), the Jiangxi Provincial Natural Science Foundation (Grant 20242BAB26002) and the Early-Career Young Scientists and Technologists Project of Jiangxi Province (Grant 20252BEJ730126). The work of Fan Zhou is supported by the Shanghai Research Center for Data Science and Decision Technology, CCF-DiDi GAIA Collaborative Research Funds, and the "Chenguang Program" supported by Shanghai Education Development and Shanghai Municipal Education Commission. The work of Yuling Jiao is supported in part by the National Key Research and Development Program of China (Grant 2024YFA1014202), the National Natural Science Foundation of China (Grants 12371441 and 12526216), and the Fundamental Research Funds for the Central Universities.

References

Abdullah, M. A., Ren, H., Ammar, H. B., Milenkovic, V., Luo, R., Zhang, M., and Wang, J. (2019). Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*.

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129.

Bellemare, M. G., Candido, S., Castro, P. S., Gong, J., Machado, M. C., Moitra, S., Ponda, S. S., and Wang, Z. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588:77 – 82.

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

Bellemare, M. G., Dabney, W., and Rowland, M. (2023). *Distributional Reinforcement Learning*. MIT Press. `http://www.distributional-rl.org`.

Belloni, A. and Chernozhukov, V. (2011). l_1-penalized quantile regression in high dimensional sparse models. *The Annals of Statistics*, pages 82–130.

Bodnar, C., Li, A., Hausman, K., Pastor, P., and Kalakrishnan, M. (2020). Quantile qt-opt for risk-aware vision-based robotic grasping. *In Robotics: Science and Systems*.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675.

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

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

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for dynamics and control*, pages 486–489. PMLR.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Ji, X., Chen, M., Wang, M., and Zhao, T. (2023). Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. *International Conference on Learning Representations*.

Jin, W., Ni, Y., O'halloran, J., Spence, A. B., Rubin, L. H., and Xu, Y. (2023). A bayesian decision framework for optimizing sequential combination antiretroviral therapy in people with hiv. *The Annals of Applied Statistics*, 17(4):3035–3055.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720.

Liao, P., Klasnja, P., and Murphy, S. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391.

Lowet, A. S., Zheng, Q., Meng, M., Matias, S., Drugowitsch, J., and Uchida, N. (2025). An opponent striatal circuit for distributional reinforcement learning. *Nature*, pages 1–10.

Lu, M., Shahn, Z., Sow, D., Doshi-Velez, F., and Li-wei, H. L. (2020). Is deep reinforcement learning ready for practical applications in healthcare? a sensitivity analysis of duel-ddqn for hemodynamic management in sepsis patients. In *AMIA Annual Symposium Proceedings*, volume 2020, page 773. American Medical Informatics Association.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Muller, T. H., Butler, J. L., Veselic, S., Miranda, B., Wallis, J. D., Dayan, P., Behrens, T. E., Kurth-Nelson, Z., and Kennerley, S. W. (2024). Distributional reinforcement learning in prefrontal cortex. *Nature Neuroscience*, 27(3):403–408.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5).

Nguyen-Tang, T., Gupta, S., Venkatesh, S., et al. (2022). On sample complexity of offline reinforcement learning with deep relu networks in besov spaces. *Transactions on Machine Learning Research*.

Padilla, O. H. M., Tansey, W., and Chen, Y. (2022). Quantile regression with relu networks: Estimators and minimax rates. *Journal of Machine Learning Research*, 23(1):11251–11292.

Qin, Z. T., Tang, X., Li, Q., Zhu, H., and Ye, J. (2025). *Reinforcement Learning in the Ridesharing Marketplace*. Springer Nature.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. (2017). Deep reinforcement learning for sepsis treatment. *Workshop on Machine Learning For Health at the conference on Neural Information Processing Systems*.

Rowland, M., Munos, R., Azar, M. G., Tang, Y., Ostrovski, G., Harutyunyan, A., Tuyls, K., Bellemare, M. G., and Dabney, W. (2024a). An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*.

Rowland, M., Tang, Y., Lyle, C., Munos, R., Bellemare, M. G., and Dabney, W. (2023). The statistical benefits of quantile temporal-difference learning for value estimation. *International Conference on Machine Learning*.

Rowland, M., Wenliang, L. K., Munos, R., Lyle, C., Tang, Y., and Dabney, W. (2024b). Near-minimax-optimal distributional reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 4(48):1875—-1897.

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. (2024). Nonparametric estimation of non-crossing quantile regression process with deep requ neural networks. *Journal of Machine Learning Research*, 25(88):1–75.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053.

Volgushev, S., Chao, S.-K., and Cheng, G. (2019). Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634 – 1662.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, 107(498):493–508.

Wu, R., Uehara, M., and Sun, W. (2023). Distributional offline policy evaluation with predictive error guarantees. In *International Conference on Machine Learning*. PMLR.

Zhang, L., Peng, Y., Liang, J., Yang, W., and Zhang, Z. (2025). Estimation and inference in distributional reinforcement learning. *The Annals of Statistics*, 53(5):1987 – 2011.

Zhou, F., Lu, C., Tang, X., Zhang, F., Qin, Z., Ye, J., and Zhu, H. (2021). Multi-objective distributional reinforcement learning for large-scale order dispatching. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1541–1546.

Zhu, W., Zeng, D., and Song, R. (2019). Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, 114(527):1404–1417.