Thank you for all the valuable feedback. One common comment is that some results may not be comparable to the original paper. Since we are not researchers from the industry and do not have strong computing power, we set the number of quantiles to be 100 and evaluate all methods using 40 million training frames for the experiment. We will definitely try to add the **complete comparison (200M, 57 games)** and more evaluation results **(stochastic Atari, different quantile numbers etc)** into the final version of the paper. We are also working on a modified version of the NC architecture which can be extended to **IQN**, and its comparison with the baseline IQN (using Google 'Dopamine') are provided in Figure (a) (last two with exploration). Our response to other specific comments are provided as follows.

Reviewer #1:

Q1. ***Can authors explain why only 49 out of 57 games are used for evaluation?.***

We choose the 49 Atari game (initially proposed by Mnih et al., 2015) since our main contrast DLTV is evaluated using the same environment. But we will definitely include all the 57 games in the final version.

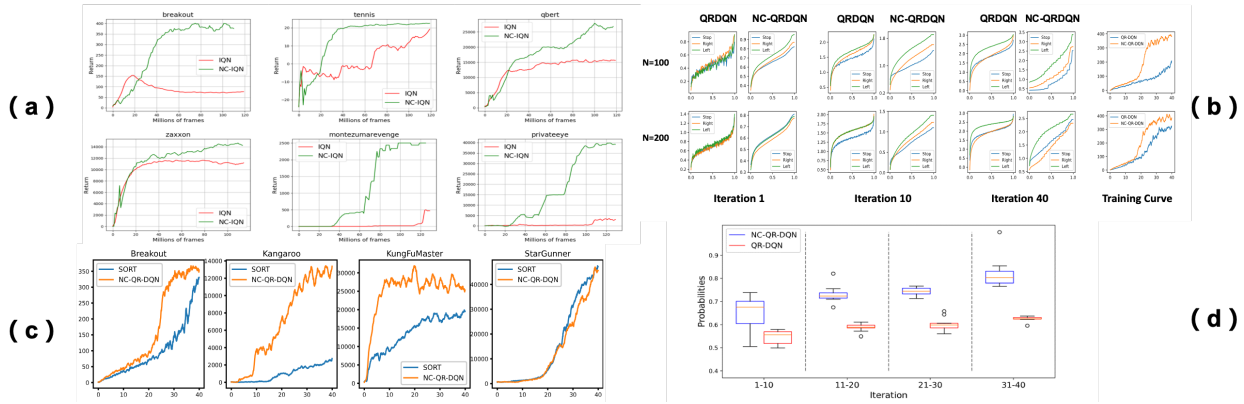
Q2. ***If the authors can commit to providing detailed evaluation results over different number of quantiles.***

We will provide detailed evaluations in the final version of the paper. Figure (b) compares N = 100, 200 on Breakout. The crossing issue is more severe with smaller N at early stage, and the improvement of our approach is more significant.

Q3. ***Hard exploration games: I would love to see complete results on full 57 games.***

Sorry for the misunderstanding. Figure 5(a) in the paper is actually the comparison between our method (with exploration) and DLTV on all the 49 games. The improvement in Motezuma's Revenge, and Private Eye are 6545% (60M) and 25% (40M) respectively. We will improve the presentation of this part in the final version.

Reviewer #2:



Q1. ***It seems much natural to simply sort the atoms before calculating the quantile Huber loss.***

As Figure (c) shows, our approach performs better than 'sort' with a higher optimal return. The key reason is that, by sorting, the same dimension of the network output may be paired to a different quantile location 'τ' each time without employing the non-decreasing constraint, which highly decreases the training stability and efficiency. With the monotonicity restriction, we can make better use of global information in training.

Q2. ***The idea of using cumulated softmax is very similar to the Fully Parameterized Quantile Function (FQF).***

Sorry for ignoring this important reference. We will cite it in the final draft of the paper.

Reviewer #3:

Q1. ***Please expand upon how the non-crossing fix improves model interpretability.***

Sometimes, the upper quantiles instead of the mean (Q-value) are of specific interests when examining some risk-appetite policy. Crossing issue will lead to awkward interpretation due to the abnormal ranking of the quantile points. Also, the non-crossing fix can bring a more precise estimation of distribution variance when doing exploration.

Q2. ***A plot showing how the ranking of Q-function changes on a variety of states would be more convincing.***

We randomly pick 4000 states, and compute the probabilities that QRDQN or NC-QRDQN chooses the same action with the optimal policy within each of the four training period. As the boxplots in Figure (d) shows, our method performs much more stable especially in the early stage with an overall higher consistency with the optimal policy.

Reviewer #4:

Q1. ***"When the sample size goes to infinity, ... " and the statement in line 122 don't seem like clear results.***

When N is fixed and sample size goes to infinity, $\theta_i(s,a)$ converges to the real quantile function F_Z^{-1} at level $\hat{\tau}_i$ as defined in line 105. Thus, the monotonicity of quantile function guarantees the monotonic constraint in equation (9). The projection operator Π_{W_1} is to find a Z_q in non-crossing space $\mathcal{Z}_Q$ to minimize its W_1 distance from Z, which is equivalent to finding a model parameterized by θ_i's under the monotonicity constraint.

Q2. ***Do we require $\mathcal{Z}_Q$ to be a Banach space, to ensure the proper convergence of the operator to the fixed point.***

Yes, we need $\mathcal{Z}_Q$ to be a complete space to ensure the proper convergence.