Perturbed Quantile Regression for Distributional Reinforcement Learning

Taehyun Cho[1], **Sungyeob Han**[1], **Heesoo Lee**[1], **Kyungjae Lee**[2], **Jungwoo Lee**[1]*
[1] Seoul National University, [2] Chung-Ang University
{talium, yubise7en, algol7240, junglee}@snu.ac.kr
{kyungjae.lee}@ai.cau.ac.kr

Abstract

Distributional reinforcement learning aims to learn distribution of return under stochastic environments. Since the learned distribution of return contains rich information about the stochasticity of the environment, previous studies have relied on descriptive statistics, such as standard deviation, for optimism in the face of uncertainty. However, using the uncertainty from an empirical distribution can hinder convergence and performance when exploring with the certain criterion that has an one-sided tendency on risk in these methods. In this paper, we propose a novel distributional reinforcement learning that explores by randomizing risk criterion to reach a risk-neutral optimal policy. First, we provide a perturbed distributional Bellman optimality operator by distorting the risk measure in action selection. Second, we prove the convergence and optimality of the proposed method by using the weaker contraction property. Our theoretical results support that the proposed method does not fall into biased exploration and is guaranteed to converge to an optimal return distribution. Finally, we empirically show that our method outperforms other existing distribution-based algorithms in various environments including 55 Atari games.

1 Introduction

Distributional reinforcement learning (DRL) learns the stochasticity of returns in the reinforcement learning environments and has shown remarkable performance in several benchmark tasks. Its model generates the approximated distribution of returns, where the mean value implies the traditional Q-value [1, 4, 10]. Learning procedure with stochasticity through return distribution is represented by *parametric (epistemic) uncertainty*, which is due to insufficient or inaccurate data, and *intrinsic (aleatoric) uncertainty*, which is inherently possessed randomness in the environment [5, 9]. The learned stochasticity gives rise to the notion of risk-sensitivity, and some distributional reinforcement learning algorithms distort the learned distribution to create a risk-averse or risk-seeking policy.

Another way to employ the uncertainty is to design an efficient exploration method which is essential to find an optimal behavior with a few trials. *Optimism in the face of uncertainty* (OFU) is one of the fundamental exploration principles that employs parametric uncertainty to promote exploring less understood behaviors and to construct confidence set. Most OFU algorithms select an action with the highest upper-confidence bound (UCB) of parametric uncertainty which can be considered as the optimism at the moment [3, 7]. In deep RL, several OFU studies often model the parametric uncertainty explicitly through the Bayesian posterior, which is estimated by using neural networks. However, learning the representation of high-dimensional state-action space and Bellman update simultaneously leads to unstable propagation [32].

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

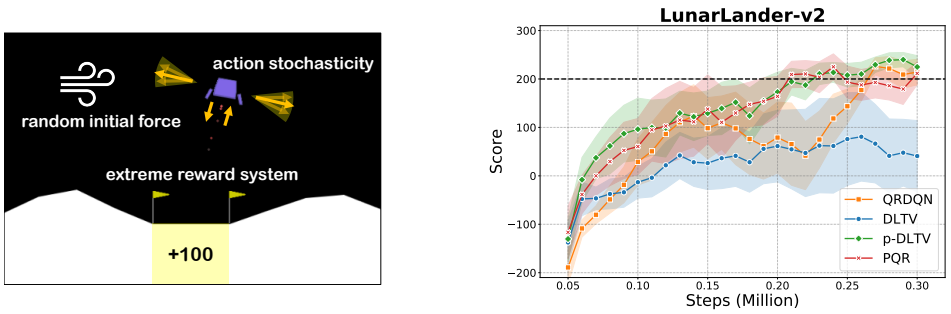


Figure 1: **Environmental stochasticity:** LunarLander-v2 is a simple RL benchmarks, but RL agents suffer from random initial force, action stochasticity by random dispersion, and extreme reward distribution. (**Left**) Three environmental factors cause high intrinsic uncertainty during episode. (**Right**) The proposed perturbation-based algorithms (PQR, p-DLTV) achieve the threshold (score 200) for safe landing.

On the other hand, DRL, which aims to capture the intrinsic uncertainty can provide more statistical information during control such as mode, median, or variance by addressing full characteristics of the return distribution. Despite the richness of risk-sensitive information for return distribution, only a few DRL methods have tried to employ the benefits of distributional perspective for exploration [8, 17, 19, 29, 34] by utilizing the estimated uncertainty from distributional output that is composed of a mixture of intrinsic and parametric uncertainty.

Unfortunately, separating these two types of uncertainty during learning is not a trivial task. Mavrin et al. [17] propose a distribution-based OFU exploration that schedules a decaying bonus rate to suppress the effect of intrinsic uncertainty, which unintentionally induces a risk-seeking policy. Although OFU based approaches try to reduce parametric uncertainty by revisiting the state with high uncertainty, there exists the side effect that the criteria unfortunately force the agent to chase the intrinsic uncertainty (risk) simultaneously due to the indistinguishability of two distinct uncertainties during updates. In Figure 1, DLTV which is based on optimism fails to reach the threshold while its baseline algorithm, QR-DQN, can achieve the goal in an environment with high intrinsic uncertainty. In the entangled case, relying on specific criteria causes a one-sided tendency on risk and makes an agent consistently select certain actions during exploration that degrades performance. We call this phenomenon *'fixedness'* and present a simple, yet effective approach to resolve such an issue.

In this paper, we propose *perturbed quantile regression* (PQR) which perturbs the criterion on uncertainty by randomizing the risk criterion in action selection to avoid a one-sided tendency on risk. We define the distributional perturbation on return distribution to re-evaluate the estimate of return by distorting the learned distribution with perturbation weight. Unlike the typical worst-case approach in risk-sensitive settings or OFU based approaches, we instead randomly sample a risk measure from an ambiguity set, which represents that the risk setting is ambiguous when the characteristics of a given environment are unknown.

In summary, our contributions are as follows.

- A risk-neutral strategy called perturbed quantile regression (PQR) is proposed, which improves over naive risk-seeking strategies.
- A sufficient condition for convergence is provided for the proposed Bellman operator with a weaker contraction property.

2 Backgrounds & Related works

2.1 Distributional RL

We consider a Markov decision process (MDP) which is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability, R is the random variable of rewards in $[-R_{\max}, R_{\max}]$, and $\gamma \in [0, 1)$ is the discount factor. We define a stochastic policy $\pi(\cdot|s)$ which is a conditional distribution over $\mathcal{A}$ given state s. For a fixed policy π, we denote $Z^\pi(s, a)$ as a random variable of return distribution of state-action pair (s, a) following

the policy π. We attain $Z^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t)$, where $S_{t+1} \sim P(\cdot|S_t, A_t),\ A_t \sim \pi(\cdot|S_t)$ and $S_0 = s,\ A_0 = a$. Then, we define an action-value function as $Q^\pi(s,a) = \mathbb{E}[Z^\pi(s,a)]$ in $[-V_{\max}, V_{\max}]$ where $V_{\max} = R_{\max}/(1-\gamma)$. For regularity, we further notice that the space of action-value distributions $\mathcal{Z}$ has the first moment bounded by $V_{\max}$:

$$\mathcal{Z} = \left\{ Z : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathbb{R}) \middle|\ \mathbb{E}[|Z(s,a)|] \leq V_{\max}, \forall (s,a) \right\}.$$

In distributional RL, the return distribution for the fixed π can be computed via dynamic programming with the distributional Bellman operator defined as,

$$\mathcal{T}^\pi Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(S', A'), \quad S' \sim P(\cdot|s,a),\ A' \sim \pi(\cdot|S')$$

where $\overset{D}{=}$ denotes that both random variables share the same probability distribution. We can compute the optimal return distribution by using the distributional Bellman optimality operator defined as,

$$\mathcal{T} Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(S', a^*), \quad S' \sim P(\cdot|s,a),\ a^* = \operatorname*{argmax}_{a'} \mathbb{E}_Z[Z(S', a')].$$

Bellemare et al. [1] have shown that $\mathcal{T}^\pi$ is a contraction in a maximal form of the Wasserstein metric but $\mathcal{T}$ is not a contraction in any metric. Combining with the expectation operator, $\mathbb{E}\mathcal{T}$ is a contraction so that we can guarantee that the expectation of Z converges to the optimal state-action value, while the convergence of a return distribution itself is not guaranteed.

2.2 Exploration on Distributional Reinforcement Learning

To combine with deep RL, a parametric distribution Z_θ is used to learn a return distribution by using $\mathcal{T}$. Dabney et al. [10] have employed a quantile regression to approximate the full distribution by letting $Z_\theta(s,a) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i(s,a)}$ where the parameter θ represents the locations of a mixture of N Dirac delta functions. Each θ_i represents the value where the cumulative probability is $\tau_i = \frac{i}{N}$. By using the quantile representation with the distributional Bellman optimality operator, the problem can be formulated as a minimization problem as,

$$\theta = \arg\min_{\theta'} D\left(Z_{\theta'}(s_t, a_t), \mathcal{T} Z_{\theta^-}(s_t, a_t)\right) = \arg\min_{\theta'} \sum_{i,j=1}^{N} \frac{\rho^\kappa_{\hat{\tau}_i}(r_t + \gamma \theta^-_j(s_{t+1}, a') - \theta'_i(s_t, a_t))}{N}$$

where (s_t, a_t, r_t, s_{t+1}) is a given transition pair, $\hat{\tau}_i = \frac{\tau_{i-1}+\tau_i}{2}$, $a' := \operatorname{argmax}_{a'} \mathbb{E}_Z[Z_\theta(s_{t+1}, a')]$, $\rho^\kappa_{\hat{\tau}_i}(x) := |\hat{\tau}_i - \delta_{\{x<0\}}|\mathcal{L}_\kappa(x)$, and $\mathcal{L}_\kappa(x) := x^2/2$ for $|x| \leq \kappa$ and $\mathcal{L}_\kappa(x) := \kappa(|x| - \frac{1}{2}\kappa)$, otherwise.

Based on the quantile regression, Dabney et al. [10] have proposed a quantile regression deep Q network (QR-DQN) that shows better empirical performance than the categorical approach [1], since the quantile regression does not restrict the bounds for return. As deep RL typically did, QR-DQN adjusts ϵ-greedy schedule, which selects the greedy action with probability $1-\epsilon$ and otherwise selects random available actions uniformly. The majority of QR-DQN variants [9, 30] rely on the same exploration method. However, such approaches do not put aside inferior actions from the selection list and thus suffers from a loss [21]. Hence, selecting a statistically plausible action is crucial for efficient exploration.

In recent studies, Mavrin et al. [17] modifies the criterion of selecting an action for efficient exploration with optimism in the face of uncertainty. Using left truncated variance as a bonus term to estimate optimistic way and decaying ratio c_t to suppress the intrinsic uncertainty, DLTV was proposed as an uncertainty-based exploration in DRL without using ϵ-greedy exploration. At timestep t, the action selection of DLTV can be described as:

$$a^* = \operatorname*{argmax}_{a'} \left(\mathbb{E}_P[Z(s', a')] + c_t \sqrt{\sigma^2_+(s', a')} \right), \quad c_t = c\sqrt{\frac{\log t}{t}},\ \sigma^2_+ = \frac{1}{2N} \sum_{i=\frac{N}{2}}^{N} (\theta_{\frac{N}{2}} - \theta_i)^2,$$

where θ_i's are the values of quantile level τ_i. DLTV shows that a constant schedule degrades the performance significantly compared to a decaying schedule.

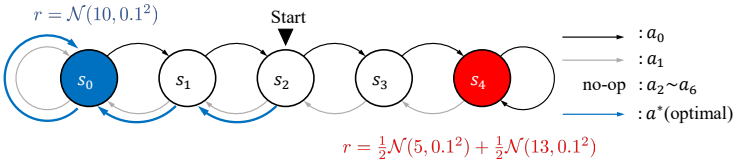


Figure 2: Illustration of the N-Chain environment starting from state s_2. To emphasize the stochasticity, the reward of state s_4 was set as a mixture model composed of two Gaussian distributions. Blue arrows indicate the risk-neutral optimal policy in this MDPs.

2.3 Risk in Distributional RL

Instead of an expected value, risk-sensitive RL tries to maximize a certain risk measure such as Mean-Variance [33], Value-at-Risk (VaR) [6], or Conditional Value-at-Risk (CVaR) [24, 25], which result in different classes of optimal policy. Especially, Dabney et al. [9] interprets risk measures as the expected utility function of the return, i.e., $\mathbb{E}_Z[U(Z(s,a))]$. Under this interpretation, risk-sensitive RL can be formulated as the maximization problem with various types of utility functions. If the utility function U is linear, the policy obtained under such risk measure is called *risk-neutral*. If U is concave or convex, the resulting policy is termed as *risk-averse* or *risk-seeking*, respectively. In general, a *distortion risk measure* is a generalized expression of risk measure generated from the distortion function.

Definition 2.1. Let $h : [0,1] \to [0,1]$ be a **distortion function** such that $h(0) = 0, h(1) = 1$ and non-decreasing. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $Z : \Omega \to \mathbb{R}$, a **distortion risk measure** ρ_h corresponding to a distortion function h is defined by:

$$\rho_h(Z) \coloneqq \mathbb{E}^{h(\mathbb{P})}[Z] = \int_{-\infty}^{\infty} z \frac{\partial}{\partial z}(h \circ F_Z)(z) dz,$$

where F_Z is the cumulative distribution function of Z.

In fact, non-decreasing property of h makes it possible to distort the distribution of Z while satisfying the fundamental property of CDF. Note that the concavity or the convexity of distortion function also implies risk-averse or seeking behavior, respectively. Dhaene et al. [11] showed that any distorted expectation can be expressed as weighted averages of quantiles. In other words, generating a distortion risk measure is equivalent to choosing a reweighting distribution.

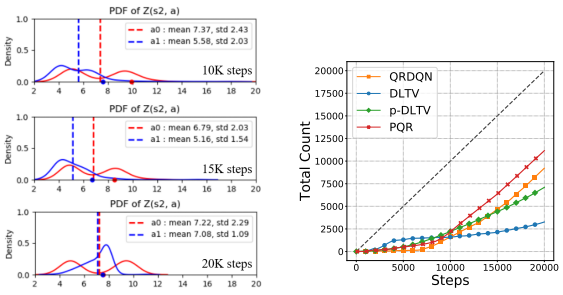
Fortunately, distributional RL has a suitable configuration to apply those uncertainty-based approaches that could naturally expand the class of policies. Chow et al. [5] and Stanko and Macek [28] considered risk-sensitive RL with a CVaR objective, where risk is related to robust decision making. Dabney et al. [9] expanded the class of policies on arbitrary distortion risk measures and investigated the effects of a distinct distortion risk measures by changing the sampling distribution for quantile targets τ. Unlike the usual risk-sensitive RL, DLTV applied the risk measure only on action selection, while it keeps the standard objective to obtain a risk-neutral optimal policy. Our paper will also utilize the risk measure only to select action, and focus on achieving the original risk-neutral purpose.

3 Perturbation in Distributional RL

3.1 Motivation

Distribution-based OFU exploration [15, 18] was proposed to give a bonus for the uncertainty that can be extracted from the distribution. However, we found that keeping optimism on uncertainty tends to select sub-optimal behaviors over a long period. For example, suppose we choose a criterion based on mean-standard deviation with decaying coefficient c_t. Consider two actions a_1, a_2 with mean μ_1, μ_2 and variance σ_1, σ_2 respectively, under the following conditions: $\mu_1 \geq \mu_2, \sigma_1 \leq \sigma_2$, and $\mu_1 + c_t\sigma_1 \leq \mu_2 + c_t\sigma_2$. Then, the agent prefers to select a_2 based on OFU. To change the decision towards the true optimal action a_1, the following steps $\eta = \min\left\{t' > t : c_{t'} \leq \frac{\mu_1 - \mu_2}{\sigma_2 - \sigma_1}\right\} - t$. need to be spent. Hence, if there is a bias in the criterion itself, such fixedness often occurs and degrades the performance as the agent has no experience with the optimal policy during that period.

To empirically demonstrate shortcomings of OFU exploration in distributional RL, we build a representative environment that is easy to interpret intuitively among the cases in which intrinsic



(a) (b)

Figure 3: (a) Empirical return distribution of DLTV during training in N-Chain environment. The dashed lines denote the exact mean, and the dots on the x-axis denote the perturbed mean of each action. No-op actions are not shown for visibility. (b) Total count of performing true optimal action. The oracle (dashed line) is to perform the true optimal action from start to end.

uncertainty exists. We experiment on the stochastic variant of N-Chain environment used in Osband et al. [20] as a toy experiment. A schematic diagram of the N-Chain environment is shown in Figure 2. The reward is only given in the leftmost and rightmost states and the game terminates when one of the reward states is reached. We set the leftmost reward as $\mathcal{N}(10, 0.1^2)$ and the rightmost reward as $\frac{1}{2}\mathcal{N}(5, 0.1^2) + \frac{1}{2}\mathcal{N}(13, 0.1^2)$ which has a lower mean as 9 but higher variance. The agent always starts from the middle state s_2 and should move toward the leftmost state s_0 to achieve the greatest expected return. For each state, the agent can take one of six available actions: left, right, and 4 no-op actions. The optimal policy with respect to mean is to move left twice from the start. Despite the simple configuration, the possibility to obtain a higher reward in the suboptimal state than the optimal state makes the agent difficult which policy is optimal until it experiences enough to detect the characteristics of each distribution. Thus, the goal of our toy experiment is to evaluate how quickly each algorithm could find a risk-neutral optimal policy.

Rather than maintaining optimism with decaying schedule to suppress intrinsic uncertainty, we modify the optimism into a randomized risk criterion and named the perturbed variant as p-DLTV. In short, p-DLTV is a simple modification of DLTV where randomness is given to the coefficient c_t through normal distribution. We compare QR-DQN and DLTV with our randomized variants of two algorithms PQR and p-DLTV to examine the effect of randomized risk criteria. We describe our main algorithm, PQR, in detail in Section 3.4. Pseudocode are given in Appendix C.5.

In Figure 3(a), DLTV fails to estimate the true optimal return distribution of action a_1. Due to the erroneous estimation, the agent takes longer to recognize its error. Hence, the deterministic selection based on a fixed criterion could mislead toward exploitation rather than exploration. This indicates that DLTV may not gather experiences well in stochastic environment. Hence, decaying schedule while maintaining optimism is not sufficient to avoid risk-seeking behavior.

In Figure 3(b), we count the number of timesteps when the optimal policy was actually performed for each algorithm to show the occurrence of fixedness. Since the optimal policy consists of the same index a_1, we plot the total count of performing the optimal action with 10 different seeds. The interval with a slope of 1 implies that the optimal policy was performed every time. From the slope of each line, it is observed that DLTV selects the suboptimal action even if the optimal policy was initially performed. Although the mean return of a_1 (move left) is estimated to be superior, the agent only selects a_2 (move right) during training due to its consistent optimism on uncertainty. Even if DLTV has spent enough number of time steps to choose the true optimal policy, the remaining procedure is already close to greedy selection as it starts from a decreased coefficient. In contrast, p-DLTV alleviates the fixedness early and finds the true optimal policy, which implies that a randomized criterion is a simple but effective on training process.

By applying this approach to risk measure, we propose a novel distributional Bellman operator which converges with a weaker contraction property, and build a practical algorithm called PQR, which produces steeper line by quickly obtaining the optimal policy. Our key idea is to select the statistically plausible action which can be maximal in a certain risk measure. Compared to QR-DQN, PQR improves efficiency by excluding inferior actions implicitly that cannot be maximal in any risk criterion. In addition, Even-Dar et al. [13] theoretically showed *action elimination*, which reduces the size of the action sets to be searched by explicitly eliminating sub-optimal action early, can speedup learning to find an optimal policy. In Section 3.2, we derive the first theoretical sufficient condition

for the convergence of exploration method in DRL, which implies that the proposed operator has the same unique fixed point as the standard distributional Bellman equation.

3.2 Perturbed Distributional Bellman Optimality Operator

To choose statistically plausible actions which may be maximal for certain risk criterion, we will generate a distortion risk measure involved in a pre-defined constraint set called an *ambiguity set*. The ambiguity set, originated from distributionally robust optimization (DRO) literature, is a family of distribution characterized by a certain statistical distance such as *ϕ-divergence* or *Wasserstein distance* [12, 26]. In this paper, we will examine the ambiguity set defined by the discrepancy between distortion risk measure and expectation. We say the sampled reweighting distribution ξ as *(distributional) perturbation* and define it as follows:

Definition 3.1. (Perturbation, Perturbation Gap, and Ambiguity Set) Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $X : \Omega \to \mathbb{R}$ be a random variable and $\Xi = \left\{\xi : \xi(w) \geq 0,\ \int_{w\in\Omega} \xi(w)\mathbb{P}(dw) = 1\right\}$ be a set of probability density functions. For a given constraint set $\mathcal{U} \subset \Xi$, we say $\xi \in \mathcal{U}$ as a **(distributional) perturbation** from $\mathcal{U}$ and denote the $\xi-$weighted expectation of X as follows:

$$\mathbb{E}_\xi[X] \coloneqq \int_{w\in\Omega} X(w)\xi(w)\mathbb{P}(dw),$$

which can be interpreted as the expectation of X under perturbed probability distribution $\xi\mathbb{P}$. We further define $d(X;\xi) = |\mathbb{E}[X] - \mathbb{E}_\xi[X]|$ as **perturbation gap** of X with respect to ξ. Then, for a given constant $\Delta \geq 0$, we define the **ambiguity set** with the bound Δ as

$$\mathcal{U}_\Delta(X) = \Big\{\xi \in \Xi : d(X;\xi) \leq \Delta\Big\}.$$

For brevity, we omit the input w from a random variable unless confusing. Since ξ is a probability density function, $\mathbb{E}_\xi[X]$ is an induced risk measure with respect to a reference measure $\mathbb{P}$. Intuitively, $\xi(w)$ can be viewed as a distortion to generate a different probability measure and allow to vary the risk tendency. The aspect of using distortion risk measures looks similar to IQN [9]. However, instead of changing the sampling distribution of quantile level τ implicitly, we reweight each quantile from the ambiguity set. This allows us to control the maximum allowable distortion with bound Δ, whereas in IQN the risk measure does not change throughout learning. In Section 3.4, we suggest a practical method to construct the ambiguity set.

Now, we characterize *perturbed distributional Bellman optimality operator* (PDBOO) $\mathcal{T}_\xi$ for a fixed perturbation $\xi \in \mathcal{U}_\Delta(Z)$ written as below:

$$\mathcal{T}_\xi Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(S', a^*(\xi)),$$
$$S' \sim P(\cdot|s,a),\ a^*(\xi) = \operatorname*{argmax}_{a'} \mathbb{E}_{\xi,P}[Z(s',a')].$$

Notice that $\xi \equiv 1$ corresponds to a base expectation, i.e., $\mathbb{E}_{\xi,P} = \mathbb{E}_P$, which recovers the standard distributional Bellman optimality operator $\mathcal{T}$. In risk-sensitive DRL or distributionally robust RL, the Bellman optimality equation is reformulated for a pre-defined risk measure [5, 27, 31]. PDBOO has a significant distinction in that it performs dynamic programming that adheres to the risk-neutral optimal policy while randomizing the risk criterion at every step.

If we consider the time-varying bound of ambiguity set, scheduling Δ_t is a key ingredient to determine whether PDBOO will efficiently explore or converge. Intuitively, if an agent continues to sample the distortion risk measure from a fixed ambiguity set with a constant Δ, there is a possibility of selecting sub-optimal actions after sufficient exploration, which may not guarantee eventual convergence.

Based on the quantile model Z_θ, our algorithm can be summarized into two parts. First, we aim to minimize the expected discrepancy between Z_θ and $\mathcal{T}_\xi Z_{\theta^-}$ where ξ is sampled from ambiguity set $\mathcal{U}_\Delta$. To clarify notation, we write $\mathbb{E}_\xi[\cdot]$ as a $\xi-$weighted expectation and $\mathbb{E}_{\xi\sim\mathscr{P}(\mathcal{U}_\Delta)}[\cdot]$ as an expectation with respect to ξ which is sampled from $\mathcal{U}_\Delta$. Then, our goal is to minimize the perturbed distributional Bellman objective with sampling procedure $\mathscr{P}$:

$$\min_{\theta'} \mathbb{E}_{\xi_t\sim\mathscr{P}(\mathcal{U}_{\Delta_t})}[D(Z_{\theta'}(s,a), \mathcal{T}_{\xi_t} Z_{\theta^-}(s,a))] \tag{1}$$

where we use the Huber quantile loss as a discrepancy on $Z_{\theta'}$ and $\mathcal{T}_\xi Z_{\theta^-}$ at timestep t. It is clearly different from DRO which performs the worst-case optimization by using a minimax objective. By using min-expectation instead of min-max operator, we investigate risk-neutral exploration that can avoid overly pessimistic policies. Second, considering a sequence ξ_t which converges uniformly to 1 so that $\mathcal{T}_{\xi_t}$ converges uniformly to standard $\mathcal{T}$, we derive a sufficient condition of Δ_t that the expectation of any composition of the operators $\mathbb{E}\mathcal{T}_{\xi_{n:1}} \coloneqq \mathbb{E}\mathcal{T}_{\xi_n}\mathcal{T}_{\xi_{n-1}}\cdots\mathcal{T}_{\xi_1}$ has the same unique fixed point as the standard.

3.3 Convergence of the perturbed distributional Bellman optimality operator

In this section, we provide the theoretical result of PDBOO about its convergence through $\mathbb{E}[Z^{(n)}]$. We denote the iteration as $Z^{(n+1)} \coloneqq \mathcal{T}_{\xi_{n+1}} Z^{(n)}$, $Z^{(0)} = Z$ for each timestep $n > 0$, and the intersection of ambiguity set as $\bar{\mathcal{U}}_{\Delta_n}(Z^{(n-1)}) \coloneqq \bigcap_{s,a} \mathcal{U}_{\Delta_n}\left(Z^{(n-1)}(s,a)\right)$.

Assumption 3.2. Suppose that the bound Δ_n satisfies the condition, $\sum_{n=1}^{\infty} \Delta_n < \infty$.

Practically, satisfying the above assumption is not strict to characterize the landscape of scheduling. For this update rule, we first state our main theorem below.

Theorem 3.3. *(Weaker Contraction Property) Let ξ_n be sampled from $\bar{\mathcal{U}}_{\Delta_n}(Z^{(n-1)})$ for every iteration. If Assumption 3.2 holds, then the expectation of any composition of operators $\mathbb{E}\mathcal{T}_{\xi_{n:1}}$ converges, i.e., $\mathbb{E}\mathcal{T}_{\xi_{n:1}}[Z] \to \mathbb{E}[Z^*]$. Moreover, the following bound holds,*

$$\sup_{s,a} \left|\mathbb{E}[Z^{(n)}(s,a)] - \mathbb{E}[Z^*(s,a)]\right| \leq \sum_{k=n}^{\infty} \left(2\gamma^{k-1} V_{max} + 2\sum_{i=1}^{k} \gamma^i (\Delta_{k+2-i} + \Delta_{k+1-i})\right).$$

Although the above theorem does not imply γ-contraction property which guarantees an unique fixed point generally, we can show that $\mathbb{E}[Z^*]$ is the unique fixed point for the operator $\mathbb{E}\mathcal{T}_{\xi_{n:1}}$ in the following theorem.

Theorem 3.4. *If Assumption 3.2 holds, $\mathbb{E}[Z^*]$ is the unique fixed point of Bellman optimality equation for any $Z \in \mathcal{Z}$.*

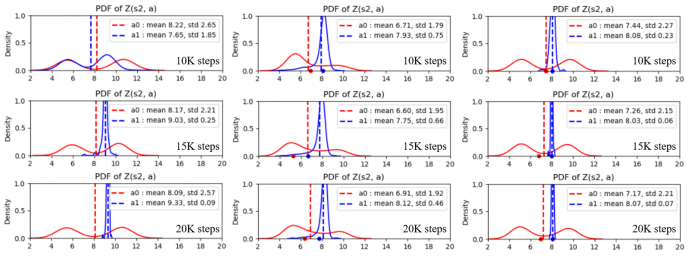
We now know that the converged value $\mathbb{E}[Z^*]$ is the unique solution of the standard Bellman optimality equation. It means that PDBOO, which only has weaker contraction property, can achieve the unique fixed point of standard Bellman operator by Assumption 3.2. Unlike the previous distribution-based or risk-sensitive approaches, PDBOO can be considered as a novel operator which has the compatibility for obtaining a risk-neutral optimal policy by randomizing risk measure during exploration.

3.4 Practical Algorithm with Distributional Perturbation

We propose a perturbed quantile regression (PQR) that is a practical algorithm for distributional reinforcement learning. Our quantile model is updated by minimizing the objective function (1) induced by PDBOO. To compute the target distribution of (1), we propose a sampling method of ξ from ambiguity set $\mathcal{U}_\Delta$. Since we employ a quantile model, sampling a reweight function ξ can be reduced into sampling an N-dimensional weight vector $\boldsymbol{\xi} \coloneqq [\xi_1, \cdots, \xi_N]$ where $\sum_{i=1}^{N} \xi_i = N$ and $\xi_i \geq 0$ for all $i \in \{1, \cdots, N\}$. Based on the QR-DQN setup, note that the condition $\int_{w\in\Omega} \xi(w)\mathbb{P}(dw) = 1$ turns into $\sum_{i=1}^{N} \frac{1}{N}\xi_i = 1$, since the quantile level is set as $\tau_i = \frac{i}{N}$.

A key issue is how to construct an ambiguity set with bound Δ_t and then sample $\boldsymbol{\xi}$. A natural class of distribution for practical use is the *symmetric Dirichlet distribution* with concentration $\boldsymbol{\beta}$, which represents distribution over distributions. (i.e. $\boldsymbol{x} \sim \text{Dir}(\boldsymbol{\beta})$.) If β is small, most of the mass is concentrated on a few elements. Otherwise, all elements are similar to each other and produce evenly distributed weight. By using the Dirichlet distribution, we sample a random vector, $\boldsymbol{x} \sim \text{Dir}(\boldsymbol{\beta})$, and define the reweight distribution as $\boldsymbol{\xi} \coloneqq \mathbf{1}^N + \alpha(N\boldsymbol{x} - \mathbf{1}^N)$. From the construction of $\boldsymbol{\xi}$, we have $1 - \alpha \leq \xi_i \leq 1 + \alpha(N-1)$ for all i and it follows that $|1 - \xi_i| \leq \alpha(N-1)$. By controlling α, we can bound the deviation of ξ_i from 1 and bound the perturbation gap as

$$\begin{aligned}
&\sup_{s,a} |\mathbb{E}[Z(s,a)] - \mathbb{E}_\xi[Z(s,a)]| = \sup_{s,a} \left|\int_{w\in\Omega} Z(w;s,a)(1-\xi(w))\mathbb{P}(dw)\right| \\
&\leq \sup_{w\in\Omega} |1-\xi(w)| \sup_{s,a} \mathbb{E}[|Z(s,a)|] \leq \sup_{w\in\Omega} |1-\xi(w)| V_{\max} \leq \alpha(N-1)V_{\max}.
\end{aligned}$$



(a) QR-DQN (b) p-DLTV (c) PQR(ours)

Ground Truth	$\mu = 8.1$	$\sigma = 0.081$
	$\Delta\mu$ $(\hat{\mu} - \mu)$	$\Delta\sigma$ $(\hat{\sigma} - \sigma)$
QR-DQN	1.23	**0.01**
DLTV	-1.02	1.01
p-DLTV	**0.02**	0.38
PQR(ours)	-0.03	**-0.01**

Figure 4: **(Left)** Empirical return distribution plot in N-Chain environment. Since QR-DQN does not depend on other criterion, the dots are omitted. **(Right)** Mean and standard-deviation difference between each algorithm and ground truth $\mathcal{N}(8.1, 0.081^2)$.

Hence, letting $\alpha \leq \frac{\Delta}{(N-1)V_{\max}}$ is sufficient to obtain $d(Z;\xi) \leq \Delta$ in the quantile setting. We set $\boldsymbol{\beta} = 0.05 \cdot \mathbf{1}^N$ to generate a constructive perturbation ξ_n which gap is close to the bound Δ_n. To satisfy Assumption 3.2, we set $\Delta_t = \Delta_0 t^{-(1+\epsilon)}$ where Δ_0 is a hyperparameter. The detailed procedure is summarized in Algorithm 2 in Appendix C.5.

4 Experimental Results and Details

Our experiments aim to answer the following questions: (1) Does our PQR method successfully escape from the fixedness phenomenon in stochastic environments? (2) Can PQR with perturbed action selection accurately capture a stochastic return distribution? (3) Can a randomized perturbation based exploration serve as a good behavior policy for the full Atari benckmark? We compare our algorithm to various DRL baselines, which has been shown to achieve better performance on stochastic RL environments. This comparison is particularly interesting since the proposed methods outperform the advanced DRL models, such as ϵ-greedy for QRDQN, IQN, and Rainbow by only applying the randomized exploration strategy. The detailed experimental setup and implementation details can be found in Appendix C.

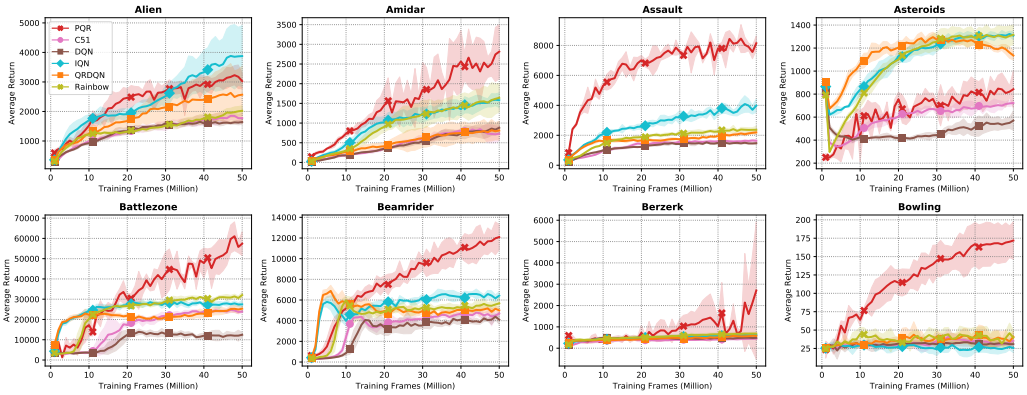


Figure 5: Evaluation curves on 8 Atari games with 3 random seeds for 50 million frames following *sticky actions* protocol [16]. Reference values are from Castro et al. [2].

4.1 Learning on Stochastic Environments with High Intrinsic Uncertainty

N-Chain. As the mean of each return is designed to be similar, it is useful to examine the learning behavior of the empirical return distribution for each algorithm. Figure 4 shows the empirical PDF of return distribution by using Gaussian kernel density estimation. Notably, p-DLTV made a much better estimate than DLTV only by changing from optimism to a randomized scheme. Although the optimal policy was performed, QR-DQN overestimates the optimal Q-value of (s_2, a_1) as $\hat{\mu} = 9.33$, while the ground truth is computed as $\mu = 10\gamma^2 = 8.1$. However, PQR estimates the ground truth much better than other baselines with much closer mean and standard-deviation.

LunarLander-v2. Figure 1 shows that p-DLTV and PQR reach the threshold faster than the other baselines. Surprisingly, p-DLTV with the randomized approach successfully reach the goal with high extreme reward, but DLTV failed to reach the landing pad. The result implies that maintaining optimism on intrinsic uncertainty leads to longer fixedness during training and therefore randomized approach could alleviate fixedness within an affordable time budget.

Atari Games with sticky actions. To avoid the deterministic dynamics of Atari games, [16] proposes injecting stochasticity scheme, called *sticky actions*, by forcing to repeat the previous action with probability $p = 0.25$. Sticky actions protocol prevents agents from relying on memorization and allows to evaluate the robustness during training. In Figure 5, as we expected, PQR shows the steeper learning curves by escaping fixedness earlier without any support of advanced schemes, such as n-step updates for Rainbow. PQR dramatically improves over IQN and Rainbow in ASSAULT, BATTLEZONE, BEAMRIDER, BEZERK and BOWLING at 50 million steps.

4.2 Full Atari Results

Finally, we evaluate the performance on full set of Atari games, each of which contained intrinsic uncertainty in different ways. Even if it is well known that Atari benchmarks without sticky actions do not have 'enough' stochasticity, intrinsic uncertainty is still prevalent in various manners. In Table 1, we evaluated 55 Atari results at 50M frames comparing with published score of QR-DQN [10], IQN [9], and Rainbow [14] via the report of DQN-Zoo benchmark [22] for reliability.

	Mean	Median	> human	> DQN
DQN-zoo(50M)	314%	55%	18	0
QR-DQN-zoo(50M)	559%	118%	29	47
IQN-zoo(50M)	902%	131%	21	50
RAINBOW-zoo(50M)	1160%	154%	37	52
PQR(50M)	1121%	124%	33	53

Table 1: Mean and median of best scores across 55 games on 50M frames, measured as percentages of human baseline. Reference values are from Quan and Ostrovski [22].

While PQR cannot enjoy the environmental stochasticity by the deterministic dynamics compared to sticky action protocol, PQR achieved 562% performance gain in the mean of human-normalized score over QR-DQN, which is comparable results to IQN. From the raw scores of 55 games, PQR wins 39 games against QR-DQN and 34 games against IQN. Note that IQN benefits from the generalized form of distributional outputs which reduce the approximation error from the number of quantiles output. While Rainbow is a combination of several orthogonal improvements such as double q-learning, prioritized replay, dueling networks, and n-step updates, PQR has another orthogonal benefit from **exploration strategies** which are based on the rich information of distributional output and shows the competitive performance with Rainbow.

4.3 Discussion

In section 4.1, a notable consistent result is that just adding randomness to the coefficient c_t on DLTV shows the significant improvement supporting that the randomized risk criterion was superior to OFU in distributional RL. In section 4.2, PQR successfully escapes the fixedness better than ϵ-greedy methods. In most RL environments with intrinsic uncertainty, we observe that OFU and ϵ-greedy have difficulty in making a decision that matches risk-neutral purpose, because two uncertainties are intertwined during learning.

5 Conclusions

In this paper, we proposed a general framework of perturbation in distributional RL which is based on the characteristics of a return distribution. Without resorting to a pre-defined risk criterion, we revealed and resolved fixedness where one-sided tendency on risk can lead to biased action selection under the stochastic environment. To our best knowledge, this paper is the first attempt to integrate risk-sensitivity and exploration by using time-varying Bellman objective with theoretical analysis. In order to validate the effectiveness of PQR, we evaluate on various environments including 55 Atari games with several distributional RL baselines. Without separating the two uncertainties, the results show that perturbing the risk criterion is an effective approach to resolve the issue of fixedness. We believe that PQR can be combined with other distributional RL or risk-sensitive algorithms as a perturbation-based exploration method without sacrificing their original objectives.

6 Acknowledgements

This work is in part supported by National Research Foundation of Korea (NRF, 2021R1A2C2014504), Artificial Intelligence Innovation Hub (2021-0-02068) grant funded by the Ministry of Science and ICT (MSIT), Center for Applied Research in Artificial Intelligence(CARAI, UD190031RD) grant Funded by Defense Acquisition Program Administration(DAPA), Agency for Defense Development(ADD), INMAC Seoul National University, and BK21-plus.

References

[1] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[2] Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G Bellemare. Dopamine: A research framework for deep reinforcement learning. *arXiv preprint arXiv:1812.06110*, 2018.

[3] Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

[4] Yunho Choi, Kyungjae Lee, and Songhwai Oh. Distributional deep reinforcement learning with a mixture of gaussians. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9791–9797. IEEE, 2019.

[5] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *arXiv preprint arXiv:1506.02188*, 2015.

[6] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[7] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor-critic. *arXiv preprint arXiv:1910.12807*, 2019.

[8] William R Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

[9] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[10] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[11] Jan Dhaene, Alexander Kukush, Daniël Linders, and Qihe Tang. Remarks on quantiles and distortion risk measures. *European Actuarial Journal*, 2(2):319–328, 2012.

[12] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[13] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[14] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[15] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[16] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

[17] Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *International conference on machine learning*, pages 4424–4434. PMLR, 2019.

[18] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. The potential of the return distribution for exploration in rl. *arXiv preprint arXiv:1806.04242*, 2018.

[19] Jihwan Oh, Joonkee Kim, and Se-Young Yun. Risk perspective exploration in distributional reinforcement learning. *arXiv preprint arXiv:2206.14170*, 2022.

[20] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29:4026–4034, 2016.

[21] Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

[22] John Quan and Georg Ostrovski. DQN Zoo: Reference implementations of DQN-based agents, 2020. URL `http://github.com/deepmind/dqn_zoo`.

[23] Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Noah Dormann. Stable baselines3. `https://github.com/DLR-RM/stable-baselines3`, 2019.

[24] Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse bayes-adaptive reinforcement learning. *arXiv preprint arXiv:2102.05762*, 2021.

[25] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[26] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[27] Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

[28] Silvestr Stanko and Karel Macek. Risk-averse distributional reinforcement learning: A cvar optimization approach. In *IJCCI*, pages 412–423, 2019.

[29] Yunhao Tang and Shipra Agrawal. Exploration by distributional reinforcement learning. *arXiv preprint arXiv:1805.01907*, 2018.

[30] Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32:6193–6202, 2019.

[31] Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 2020.

[32] Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, and Peng Liu. Exploration in deep reinforcement learning: A comprehensive survey. *arXiv preprint arXiv:2109.06668*, 2021.

[33] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. *arXiv preprint arXiv:2004.10888*, 2020.

[34] Fan Zhou, Zhoufan Zhu, Qi Kuang, and Liwen Zhang. Non-decreasing quantile function network with efficient exploration for distributional reinforcement learning. *arXiv preprint arXiv:2105.06696*, 2021.