A Proof

A.1 Technical Lemma

Before proving our theoretical results, we present two inequalities for supremum to clear the description.

1. $\sup_{x\in X}|f(x)+g(x)| \le \sup_{x\in X}|f(x)| + \sup_{x\in X}|g(x)|$
2. $\left|\sup_{x\in X} f(x) - \sup_{x'\in X} g(x')\right| \le \sup_{x,x'\in X}|f(x)-g(x')|$

Proof of 1. Since $|f(x)+g(x)| \le |f(x)|+|g(x)|$ holds for all $x \in X$,

$$\begin{aligned}\sup_{x\in X}|f(x)+g(x)| &\le \sup_{x\in X}(|f(x)|+|g(x)|)\\ &\le \sup_{x\in X}|f(x)| + \sup_{x\in X}|g(x)|\end{aligned}$$

■

Proof of 2. Since $\left|\|a\|-\|b\|\right| \le \|a-b\|$ for any norm $\|\cdot\|$ and for a large enough M,

$$\begin{aligned}\sup_{x,x'\in X}|f(x)-g(x')| &\ge \sup_{x\in X}|f(x)-g(x)|\\ &= \sup_{x\in X}|(f(x)+M)-(g(x)+M)|\\ &\ge \left|\sup_{x\in X}(f(x)+M) - \sup_{x\in X}(g(x)+M)\right|\\ &= \left|\sup_{x\in X} f(x) - \sup_{x'\in X} g(x')\right|\end{aligned}$$

■

A.2 Proof of Theorem A.3

Theorem A.3. If ξ_t converges to 1 uniformly on Ω, then $\mathbb{E}\mathcal{T}_{\xi_t}$ also converges to $\mathbb{E}\mathcal{T}$ uniformly on $\mathcal{Z}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Proof. Recall that $\mathcal{Z} = \left\{Z : \mathcal{S}\times\mathcal{A} \to \mathscr{P}(\mathbb{R}) \middle|\ \mathbb{E}[|Z(s,a)|] \le V_{\max}, \forall(s,a)\right\}$. Then for any $Z \in \mathcal{Z}$ and $\xi \in \Xi$,

$$\mathbb{E}[|\mathcal{T}_\xi Z|] \le R_{\max} + \gamma\frac{R_{\max}}{1-\gamma} = \frac{R_{\max}}{1-\gamma} = V_{\max}.$$

which implies PDBOO is closed in $\mathcal{Z}$, i.e. $\mathcal{T}_\xi Z \in \mathcal{Z}$ for all $\xi \in \Xi$. Hence, for any sequence ξ_t, $Z^{(n)} = \mathcal{T}_{\xi_{n:1}} Z \in \mathcal{Z}$ for any $n \ge 0$.

Since ξ_t converges to 1 uniformly on Ω, there exists T such that for any $t > T$,

$$\sup_{w\in\Omega}|\xi_t(w)-1| \le \epsilon.$$

For any $Z \in \mathcal{Z}$, $s \in \mathcal{S}, a \in \mathcal{A}$, and $t > T$, by using Hölder's inequality,

$$\begin{aligned}\sup_{Z\in\mathcal{Z}}\sup_{s,a}|\mathbb{E}_{\xi_t}[Z(s,a)] - \mathbb{E}[Z(s,a)]| &= \sup_{Z\in\mathcal{Z}}\sup_{s,a}\left|\int_{w\in\Omega}(1-\xi_t(w))Z(s,a,w)\mathbb{P}(w)dw\right|\\ &\le \sup_{w\in\Omega}|\xi_t(w)-1|\sup_{Z\in\mathcal{Z}}\sup_{s,a}\left|\int_{w\in\Omega}|Z(s,a,w)|\mathbb{P}(w)dw\right|\\ &\le \epsilon V_{\max}\end{aligned}$$

which implies that $\mathbb{E}_{\xi_t}$ converges to $\mathbb{E}$ uniformly on $\mathcal{Z}$ for all s, a.

By using A.1, we can get the desired result.

$$
\begin{aligned}
&\sup_{Z\in\mathcal{Z}}\sup_{s,a}|\mathbb{E}[\mathcal{T}_{\xi_t}Z(s,a)]-\mathbb{E}[\mathcal{T}Z(s,a)]| \\
&\le \sup_{Z\in\mathcal{Z}}\sup_{s,a}|\mathbb{E}[\mathcal{T}_{\xi_t}Z(s,a)]-\mathbb{E}_{\xi_t}[\mathcal{T}_{\xi_t}Z(s,a)]|+\sup_{Z\in\mathcal{Z}}\sup_{s,a}|\mathbb{E}_{\xi_t}[\mathcal{T}_{\xi_t}Z(s,a)]-\mathbb{E}[\mathcal{T}Z(s,a)]| \\
&\le \epsilon V_{\max}+\gamma\sup_{Z\in\mathcal{Z}}\sup_{s,a}\mathbb{E}_{s'}\left[\left|\sup_{a'}\mathbb{E}_{\xi_t}[Z(s',a')]-\sup_{a''}\mathbb{E}[Z(s',a'')]\right|\right] \\
&\le \epsilon V_{\max}+\gamma\sup_{Z\in\mathcal{Z}}\sup_{s',a'}|\mathbb{E}_{\xi_t}[Z(s',a')]-\mathbb{E}[Z(s',a')]| \\
&\le \epsilon V_{\max}+\gamma\epsilon V_{\max} \\
&= (1+\gamma)\epsilon V_{\max}.
\end{aligned}
$$

■

A.3 Proof of Theorem 3.3

Theorem 3.3. 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,

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

Proof. We denote $a_i^*(\xi_n)=\operatorname{argmax}_{a'}\mathbb{E}_{\xi_n}[Z_i^{(n-1)}(s',a')]$ as the greedy action of $Z_i^{(n-1)}$ under perturbation ξ_n. Also, we denote $\sup_{s,a}|\cdot|$ which is the supremum norm over s and a as $\|\cdot\|_{sa}$.

Before we start from the term $\left\|\mathbb{E}[Z^{(k+1)}]-\mathbb{E}[Z^{(k)}]\right\|_{sa}$, for a given (s,a),

$$
\begin{aligned}
&\left|\mathbb{E}[Z^{(k+1)}(s,a)]-\mathbb{E}[Z^{(k)}(s,a)]\right| \\
&\le\gamma\sup_{s'}\left|\mathbb{E}[Z^{(k)}(s',a^*(\xi_{k+1}))]-\mathbb{E}[Z^{(k-1)}(s',a^*(\xi_k))]\right| \\
&\le\gamma\sup_{s'}\Big(\left|\mathbb{E}[Z^{(k)}(s',a^*(\xi_{k+1})]-\max_{a'}\mathbb{E}[Z^{(k)}(s',a')]\right|+\left|\max_{a'}\mathbb{E}[Z^{(k)}(s',a')]-\max_{a'}\mathbb{E}[Z^{(k-1)}(s',a')]\right| \\
&\quad+\left|\max_{a'}\mathbb{E}[Z^{(k-1)}(s',a')]-\mathbb{E}[Z^{(k-1)}(s',a^*(\xi_k))]\right|\Big) \\
&\le\gamma\sup_{s',a'}\left|\mathbb{E}[Z^{(k)}(s',a')]-\mathbb{E}[Z^{(k-1)}(s',a')]\right|+\gamma\sum_{i=k-1}^{k}\sup_{s'}\left|\mathbb{E}[Z^{(i)}(s',a^*(\xi_{i+1}))]-\max_{a'}\mathbb{E}[Z^{(i)}(s',a')]\right| \\
&\le\gamma\left\|\mathbb{E}[Z^{(k)}]-\mathbb{E}[Z^{(k-1)}]\right\|_{sa}+\gamma\sum_{i=k-1}^{k}\Big[\sup_{s'}\Big(\left|\mathbb{E}[Z^{(i)}(s',a^*(\xi_{i+1}))]-\mathbb{E}_{\xi_{i+1}}[Z^{(i)}(s',a^*(\xi_{i+1}))]\right| \\
&\quad+\left|\max_{a'}\mathbb{E}_{\xi_{i+1}}[Z^{(i)}(s',a')]-\max_{a''}\mathbb{E}[Z^{(i)}(s',a''))]\right|\Big)\Big] \\
&\le\gamma\left\|\mathbb{E}[Z^{(k)}]-\mathbb{E}[Z^{(k-1)}]\right\|_{sa}+2\gamma\sum_{i=k-1}^{k}\sup_{s',a'}\left(\left|\mathbb{E}[Z^{(i)}(s',a')]-\mathbb{E}_{\xi_{i+1}}[Z^{(i)}(s',a')]\right|\right) \\
&\le\gamma\left\|\mathbb{E}[Z^{(k)}]-\mathbb{E}[Z^{(k-1)}]\right\|_{sa}+2\gamma\sum_{i=k-1}^{k}\Delta_{i+1}
\end{aligned}
$$

where we use A.1.1 in third and fifth line and A.1.2 in sixth line.

Taking a supremum over s and a, then for all $k > 0$,

$$\begin{aligned}
&\left\|\mathbb{E}[Z^{(k+1)}] - \mathbb{E}[Z^{(k)}]\right\|_{sa} \\
&\leq \gamma \left\|\mathbb{E}[Z^{(k)}] - \mathbb{E}[Z^{(k-1)}]\right\|_{sa} + 2\sum_{i=k-1}^{k} \gamma\Delta_{i+1} \\
&\leq \gamma^2 \left\|\mathbb{E}[Z^{(k-1)}] - \mathbb{E}[Z^{(k-2)}]\right\|_{sa} + 2\sum_{i=k-2}^{k-1} \gamma^2\Delta_{i+1} + 2\sum_{i=k-1}^{k} \gamma\Delta_{i+1} \\
&\quad\vdots \\
&\leq \gamma^k \left\|\mathbb{E}[Z^{(1)}] - \mathbb{E}[Z]\right\|_{sa} + 2\sum_{i=1}^{k} \gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i}) \\
&\leq 2\gamma^k V_{\max} + 2\sum_{i=1}^{k} \gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i})
\end{aligned}$$

Since $\sum_{i=1}^{\infty} \gamma^i = \frac{\gamma}{1-\gamma} < \infty$ and $\sum_{i=1}^{\infty} \Delta_i < \infty$ by assumption, we have

$$\sum_{i=1}^{k} \gamma^i \Delta_{k+1-i} \to 0$$

which is resulted from the convergence of Cauchy product of two sequences $\{\gamma^i\}$ and $\{\Delta_i\}$. Hence, $\{\mathbb{E}[Z^{(k)}]\}$ is a Cauchy sequence and therefore converges for every $Z \in \mathcal{Z}$.

Let $\mathbb{E}[Z^*]$ be the limit point of the sequence $\{\mathbb{E}[Z^{(n)}]\}$. Then,

$$\begin{aligned}
\left\|\mathbb{E}[Z^*] - \mathbb{E}[Z^{(n)}]\right\|_{sa} &= \lim_{l\to\infty} \left\|\mathbb{E}[Z^{(n+l)}] - \mathbb{E}[Z^{(n)}]\right\|_{sa} \\
&\leq \sum_{k=n}^{\infty} \left\|\mathbb{E}[Z^{(k+1)}] - \mathbb{E}[Z^{(k)}]\right\|_{sa} \\
&= \sum_{k=n}^{\infty} \Big(2\gamma^k V_{\max} + 2\sum_{i=1}^{k} \gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i})\Big).
\end{aligned}$$

■

A.4 Proof of Theorem 3.4

Theorem 3.4. If $\{\Delta_n\}$ follows the assumption in Theorem 3.3, then $\mathbb{E}[Z^*]$ is the unique solution of Bellman optimality equation.

Proof. The proof follows by linearity of expectation. Denote the Q-value based operator as $\bar{\mathcal{T}}$. Note that Δ_n converges to 0 with regularity of $\mathcal{Z}$ implies that ξ_n converges to 1 uniformly on Ω. By Theorem A.3, for a given $\epsilon > 0$, there exists a constant $K = \max(K_1, K_2)$ such that for every $k \geq K_1$,

$$\sup_{Z\in\mathcal{Z}} \|\bar{\mathcal{T}}_{\xi_k}\mathbb{E}[Z] - \bar{\mathcal{T}}\mathbb{E}[Z]\|_{sa} \leq \frac{\epsilon}{2}.$$

Since $\bar{\mathcal{T}}$ is continuous, for every $k \geq K_2$,

$$\|\bar{\mathcal{T}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa} \leq \frac{\epsilon}{2}.$$

Thus, it holds that

$$\begin{aligned}\|\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa} &\leq \|\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^{(k)}]\|_{sa} + \|\bar{\mathcal{T}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa} \\ &\leq \sup_{Z\in\mathcal{Z}} \|\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z] - \bar{\mathcal{T}}\mathbb{E}[Z]\|_{sa} + \|\bar{\mathcal{T}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa} \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \\ &= \epsilon.\end{aligned}$$

Therefore, we have

$$\mathbb{E}[Z^*] = \lim_{k\to\infty} \mathbb{E}[Z^{(k)}] = \lim_{k\to\infty} \mathbb{E}[Z^{(k+1)}] = \lim_{k\to\infty} \mathbb{E}[\mathcal{T}_{\xi_{k+1}} Z^{(k)}] = \lim_{k\to\infty} \bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z^{(k)}] = \bar{\mathcal{T}}\mathbb{E}[Z^*]$$

Since the standard Bellman optimality operator has a unique solution, we derived the desired result. ■

B Algorithm Pipeline and Illustrative Example

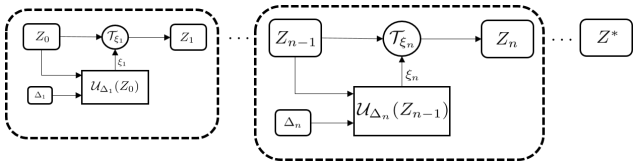

Figure 6: Pipeline of PDBOO.

Figure 6 shows the pipeline of our algorithm. With the schedule of perturbation bound $\{\Delta_n\}$, the ambiguity set $\mathcal{U}_{\Delta_n}(Z_{n-1})$ can be defined by previous Z_{n-1}. For each step, (distributional) perturbation ξ_n is sampled from $\mathcal{U}_{\Delta_n}(Z_{n-1})$ by the symmetric Dirichlet distribution and then PDBOO $\mathcal{T}_{\xi_n}$ can be performed.

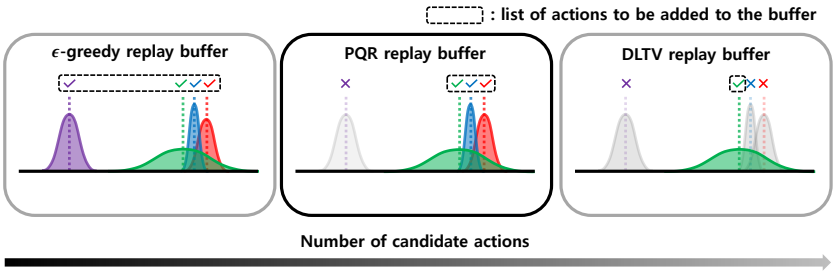

Figure 7: An illustrative example of proposed algorithm (PQR). Each distribution represents the empirical PDF of return. PQR benefits from excluding inferior actions and promoting unbiased selection with regards to high intrinsic uncertainty through randomized risk criterion.

C Implementation details

Except for each own hyperparameter, our algorithms and DLTV shares the same hyperparameter and network architecture with QR-DQN [10] for a fair comparison. Also, we set up p-DLTV by only multiplying a gaussian noise $\mathcal{N}(0, 1)$ to the coefficient of DLTV. We do not combine any additional improvements of Rainbow such as double Q-learning, dueling network, prioritized replay, and n-step update. Experiments on LunarLander-v2 and Atari games were performed with 3 random seeds. The training process is 0-2% slower than QR-DQN due to the sampling ξ and reweighting procedures.

C.1 Hyperparameter Setting

We report the hyperparameters for each environments we used in our experiments.

- Batch size : 32 (Atari games), 64 (N-Chain), 128 (LunarLander-v2)
- Number of quantiles N: 170 (LunarLander-v2), 200 (N-Chain, Atari games)
- Step for $n-$step updates : 1
- Network optimizer : Adam
- β : Grid search[**0.05**, 0.1, 0.5, 1] $\times \mathbf{1}^N$
- κ : 1
- Memory size : 1×10^5 (LunarLander-v2), 1×10^6 (Atari games)
- learning rate : 5×10^{-5} (N-Chain, Atari games), 1.5×10^{-3} (LunarLander-v2)
- Gamma : 0.9 (N-Chain), 0.99 (Atari games), 0.995 (LunarLander-v2)
- Update interval : 1 (N-Chain, LunarLander-v2), 4 (Atari games)
- Target update interval : 1 (LunarLander-v2), 25 (N-Chain), 1×10^4 (Atari games)
- Start steps : 5×10^2 (N-Chain), 1×10^4 (LunarLander-v2), 5×10^4 (Atari games)
- ϵ (train)/(test) : LinearAnnealer($1 \rightarrow 1 \times 10^{-2}$) / 1×10^{-3}
- ϵ decay steps : 2.5×10^3 (N-Chain), 1×10^5 (LunarLander-v2), 2.5×10^5 (Atari games)
- Coefficient c : Grid search[**0.05** (LunarLander-v2), 0.5, 5, **50** (N-Chain, Atari games)]
- Δ_0 : 5×10^2 (N-Chain), 5×10^4 (LunarLander-v2), 1×10^6 (Atari games)

C.2 N-Chain

We used $c = 50$ which was implemented in Mavrin et al. [17]. Although this c may not be optimal in a given environment, noted that its perturb variant, p-DLTV, learns successfully in the same settings. The ground truth of return distribution at state s_0 and s_4 are computed as $\gamma^2\mathcal{N}(10, 0.1^2)$ and $\gamma^2(\frac{1}{2}\mathcal{N}(5, 0.1^2) + \frac{1}{2}\mathcal{N}(13, 0.1^2))$, respectively.

C.3 LunarLander-v2

The hyperparameters of QR-DQN were followed by the settings reported in Raffin et al. [23] for a fair comparison. Our experiments used 2 layers of MLP with 256 hidden units. As a stochastic gradient optimizer, we adopt Adam with a learning rate 1.5×10^{-3} with a linear decaying schedule. For the rest, the best value of $c = 0.05$ was chosen from $[50, 5, 0.5, 0.05]$ where p-DLTV succeeded to learn while DLTV failed in all cases.

C.4 Atari games

For a fair comparison, our hyperparameter setting is aligned with Dabney et al. [10].

C.5 Pseudocode of p-DLTV and PQR

Algorithm 1 Perturbed DLTV (p-DLTV)

Input: transition (s, a, r, s'), discount $\gamma \in [0, 1)$
$Q(s', a') = \frac{1}{N}\sum_j \theta_j(s', a')$
Randomize the coefficient
$c_t \sim c\,\mathcal{N}(0, \frac{\ln t}{t})$
$a^* \leftarrow \mathrm{argmax}_{a'}(Q(s', a') + c_t\sqrt{\sigma_+^2(s', a')})$
$\mathcal{T}\theta_j \leftarrow r + \gamma\theta_j(s', a^*), \quad \forall j$
Output: $\sum_{i=1}^N \mathbb{E}_j[\rho_{\hat{\tau}_i}^{\kappa}(\mathcal{T}\theta_j - \theta_i(s, a))]$

Algorithm 2 Perturbed QR-DQN (PQR)

Input: (s, a, r, s'), $\gamma \in [0, 1)$, timestep $t > 0$, $\epsilon > 0$, concentration $\boldsymbol{\beta}$

Initialize $\Delta_0 > 0$.

$\Delta_t \leftarrow \Delta_0 t^{-(1+\epsilon)}$.

Sample $\xi \sim \bar{\mathcal{U}}_{\Delta_t}(Z^{(t)})$

$\boldsymbol{\xi} \leftarrow \max\left(\mathbf{1}^N + \Delta_t(N\boldsymbol{x} - \mathbf{1}^N), 0\right)$ where $\boldsymbol{x} \sim \text{Dir}(\boldsymbol{\beta})$

$\boldsymbol{\xi} \leftarrow N\boldsymbol{\xi}/\sum \xi_i$

Select greedy action with distorted expectation

$a^* \leftarrow \text{argmax}_{a'} \mathbb{E}_{\xi}[Z(s', a')]$

$\mathcal{T}\theta_j \leftarrow r + \gamma\theta_j(s', a^*), \quad \forall j$

$t \leftarrow t + 1$

Output: $\sum_{i=1}^{N} \mathbb{E}_j[\rho^{\kappa}_{\hat{\tau}_i}(\mathcal{T}\theta_j - \theta_i(s, a))]$

D Further experimental results & Discussion

D.1 N-Chain

Total Count	(8,10)	(7,11)	(6,12)	(5,13)	(4,14)	(3,15)	(2,16)	(1,17)
QR-DQN	12293	11381	11827	12108	10041	11419	9696	11619
DLTV	9997	9172	9646	9251	7941	6964	7896	7257
p-DLTV	14344	14497	13769	**15507**	14469	**14034**	14068	13404
PQR	**14546**	**15018**	**14693**	15142	**15361**	13859	**14602**	**14354**

Table 2: Total counts of performing true optimal action with 4 seeds.

To explore the effect of intrinsic uncertainty, we run multiple experiments with various reward settings for the rightmost state as keeping their mean at 9. As the distance between two Gaussians was increased, the performance of DLTV decrease gradually, while other algorithms show consistent results. The result implies the interference of fixedness is proportional to the magnitude of the intrinsic uncertainty and the randomized criterion is effective in escaping from the issue.

D.2 LunarLander-v2

To verify the effectiveness of the proposed algorithm in the complex environment with **high intrinsic uncertainty**, we conduct the experiment on LunarLander-v2. We have focused on three main factors that increase the intrinsic uncertainty from the structural design of LunarLander environment:

- **Random initial force:** The lander starts at the top center with an random initial force.
- **Action stochasticity:** The noise of engines causes different transitions with same action.
- **Extreme reward system:** If the lander crashes, it receives -100 points. If the lander comes to rest, it receives +100 points.

Therefore, several returns with a fixed policy have a high variance. As previously discussed about the fixedness from N-Chain environment, we can demonstrate that randomized approaches, PQR and p-DLTV, outperform other baselines in LunarLander-v2.

D.3 Atari games

We test our algorithm under 30 no-op settings to align with previous works. We compare our baseline results with results from the DQN Zoo framework [22], which provides the full benchmark results on 55 Atari games at 50M and 200M frames. We report the average of the best scores over 5 seeds for each baseline algorithms up to 50M frames.

However, recent studies tried to follow the setting proposed by Machado et al. [16] for reproducibility, where they recommended using sticky actions. Hence, we provide all human normalized scores

results across 55 Atari games for 50M frames including previous report of Dopamine and DQN Zoo framework to help the follow-up researchers as a reference. We exclude `Defender` and `Surround` which is not reported on Yang et al. [30] because of relialbility issues in the Dopamine framework. In summary,

- DQN Zoo framework corresponds to 30 no-op settings (version **v4**).
- Dopamine framework corresponds to sticky actions protocol (version **v0**).

For the expected concerns about the comparison with DLTV, we address some technical issues to correct misconceptions of their performance. Before we reproduce the empirical results of DLTV, Mavrin et al. [17] did not report each raw scores of Atari games, but only the relative performance with cumulative rewards comparing with QR-DQN. While DLTV was reported to have a cumulative reward 4.8 times greater than QR-DQN, such gain mainly comes from `VENTURE` which is evaluated as 22,700% from their metric (i.e., 463% performance gain solely). From their training curves, however, the approximate raw score of `VENTURE` was 900 which is lower than our score of 993.3. So, the report with cumulative rewards causes a misconception that can be overestimated where the human-normalized score is commonly used for evaluation metrics. Due to the absence of public results, DLTV were inevitably excluded from the comparison with human-normalized score for reliability.

	Mean	**Median**
DQN-dopamine(50M)	401%	51%
DQN-zoo(50M)	314%	55%
QR-DQN-dopamine(50M)	562%	93%
QR-DQN-zoo(50M)	559%	118%
IQN*-dopamine(50M)	940%	124%
IQN-zoo(50M)	902%	131%
RAINBOW-dopamine(50M)	965%	123%
RAINBOW-zoo(50M)	1160%	154%
PQR(50M)	1121%	124%

Table 3: Mean and median of best scores across 55 Atari games on 50M frames, measured as percentages of human baseline [2, 22, 30]. IQN* implemented in Dopamine framework is the improved version of original IQN by applying n-step updates with $n = 3$. **Figure 5** also uses the data from IQN*.

Table 3 provides the mean and median human normalized scores across 55 Atari games. Due to the high computational cost, our algorithm was evaluated on 50M frames to provide results over as many environments as possible. It is observed that PQR shows better performance in terms of both mean and median metrics than QR-DQN. Since our method is based on QR-DQN, we expect that PDBOO can be combined with IQN [9] or techniques in Rainbow [14] as an efficient exploration method, and the performance can be further improved.

GAMES	RANDOM	HUMAN	DQN(50M)	QR-DQN(50M)	IQN(50M)	RAINBOW(50M)	PQR(50M)
Alien	227.8	7127.7	1541.5	1645.7	1769.2	4356.9	**2455.8**
Amidar	5.8	1719.5	324.2	683.4	799.2	2549.2	**938.4**
Assault	222.4	742.0	2387.8	11684.2	15152.4	9737.0	10759.2
Asterix	210.0	8503.3	5249.5	18373.4	32598.2	33378.6	10490.5
Asteroids	719.1	47388.7	1106.3	1503.9	1972.6	1825.4	1662.0
Atlantis	12850.0	29028.1	283392.2	937275.0	865360.0	941740.0	897640.0
BankHeist	14.2	753.1	389.0	1223.9	1266.8	1081.7	1038.8
BattleZone	2360.0	37187.5	19092.4	26325.0	30253.9	35467.1	28470.5
BeamRider	363.9	16926.5	7133.1	12912.0	19251.4	15421.9	10224.9
Berzerk	123.7	2630.4	577.4	826.5	918.9	2061.6	***137873.1**
Bowling	23.1	160.7	34.4	45.4	41.5	54.7	***86.9**
Boxing	0.1	12.1	87.2	99.6	99.2	99.8	97.1
Breakout	1.7	30.5	316.8	426.5	468.0	335.3	380.3
Centipede	2090.9	12017.0	4935.7	7124.0	7008.3	5691.4	***7291.2**
ChopperCommand	811.0	7387.8	974.2	1187.8	1549.0	5525.1	1300.0
CrazyClimber	10780.5	35829.4	96939.0	93499.1	127156.5	160757.7	84390.9
DemonAttack	152.1	1971.0	8325.6	106401.8	110773.1	85776.5	73794.0
DoubleDunk	-18.6	-16.4	-15.7	-10.5	-12.1	-0.3	**-7.5**
Enduro	0.0	860.5	750.6	2105.7	2280.6	2318.3	***2341.2**
FishingDerby	-91.7	-38.7	8.2	25.7	23.4	35.5	**31.7**
Freeway	0.0	29.6	24.4	33.3	33.7	34.0	**34.0**
Frostbite	65.2	4334.7	408.2	3859.2	5650.8	9672.6	4148.2
Gopher	257.6	2412.5	3439.4	6561.9	26768.9	32081.3	***47054.5**
Gravitar	173.0	3351.4	180.9	548.1	470.2	2236.8	**635.8**
Hero	1027.0	30826.4	9948.3	9909.8	12491.1	38017.9	**12579.2**
IceHockey	-11.2	0.9	-11.4	-2.1	-4.2	1.9	**-1.4**
Jamesbond	29.0	302.8	486.4	1163.8	1058.0	14415.5	**2121.8**
Kangaroo	52.0	3035.0	6720.7	14558.2	14256.0	14383.6	***14617.1**
Krull	1598.0	2665.5	7130.5	9612.5	9616.7	8328.5	***9746.1**
KungFuMaster	258.5	22736.3	21330.9	27764.3	39450.1	30506.9	***43258.6**
MontezumaRevenge	0.0	4753.3	0.3	0.0	0.2	80.0	0.0
MsPacman	307.3	6951.6	2362.9	2877.5	2737.4	3703.4	**2928.9**
NameThisGame	2292.3	8049.0	6328.0	11843.3	11582.2	11341.5	10298.2
Phoenix	761.4	7242.6	10153.6	35128.6	29138.9	49138.8	20453.8
Pitfall	-229.4	6463.7	-9.5	0.0	0.0	0.0	**0.0**
Pong	-20.7	14.6	18.7	20.9	20.9	21.0	**21.0**
PrivateEye	24.9	69571.3	266.6	100.0	100.0	160.0	***372.4**
Qbert	163.9	13455.0	5567.9	12808.4	15101.8	24484.9	**15267.4**
Riverraid	1338.5	17118.0	6782.8	9721.9	13555.9	17522.9	11175.3
RoadRunner	11.5	7845.0	29137.5	54276.3	53850.9	52222.6	50854.7
Robotank	2.2	11.9	31.4	54.5	53.8	64.5	**60.3**
Seaquest	68.4	42054.7	2525.8	7608.2	17085.6	3048.9	***19652.5**
Skiing	-17098.1	-4336.9	-13930.8	-14589.7	-19191.1	-15232.3	***-9299.3**
Solaris	1236.3	12326.7	2031.5	1857.3	1301.5	2522.6	***2640.0**
SpaceInvaders	148.0	1668.7	1179.1	1753.2	2906.7	2715.3	1749.4
StarGunner	664.0	10250.0	24532.5	63717.3	78503.4	107177.8	62920.6
Tennis	-23.8	-8.3	-0.9	0.0	0.0	0.0	-1.0
TimePilot	3568.0	5229.2	2091.8	6266.8	6379.1	12082.1	**6506.4**
Tutankham	11.4	167.6	138.7	210.2	204.4	194.3	***231.3**
UpNDown	533.4	11693.2	6724.5	27311.3	35797.6	65174.2	**36008.1**
Venture	0.0	1187.5	53.3	12.5	17.4	1.1	***993.3**
VideoPinball	16256.9	17667.9	140528.4	104405.8	341767.5	465636.5	**465578.3**
WizardOfWor	563.5	4756.5	3459.9	14370.2	10612.1	12056.1	6132.8
YarsRevenge	3092.9	54576.9	16433.7	21641.4	21645.0	67893.3	**27674.4**
Zaxxon	32.5	9173.3	3244.9	9172.1	8205.2	22045.8	**10806.6**

Table 4: Raw scores across all 55 games, starting with 30 no-op actions. We report the best scores for DQN, QR-DQN, IQN and Rainbow on 50M frames, averaged by 5 seeds. Reference values were provided by DQN Zoo framework [22]. **Bold** are wins against DQN, QR-DQN and IQN, and *asterisk are wins over Rainbow.

GAMES	RANDOM	HUMAN	DQN(50M)	QR-DQN(50M)	IQN*(50M)	RAINBOW(50M)	PQR(50M)
Alien	227.8	7127.7	1688.1	2754.2	4016.3	2076.2	**3173.9**
Amidar	5.8	1719.5	888.2	841.6	1642.8	1669.6	***2814.7**
Assault	222.4	742	1615.9	2233.1	4305.6	2535.9	***8456.5**
Asteroids	719.1	47388.7	828.2	1333.4	1336.3	1345.1	904.6
BankHeist	14.2	753.1	720.2	964.1	1082.8	1104.9	501.1
BattleZone	2360.0	37187.5	15110.3	25845.6	29959.7	32862.1	***61494.4**
BeamRider	343.9	16926.5	4771.3	7143.0	7113.7	6331.9	***12217.6**
Berzerk	123.7	2630.4	529.2	603.2	627.3	697.8	***2707.2**
Bowling	23.1	160.7	38.5	55.3	33.6	55.0	***174.1**
Boxing	0.1	12.1	80.0	96.6	97.8	96.3	**96.7**
Breakout	1.7	30.5	113.5	40.7	164.4	69.8	41.7
Centipede	2090.9	12017.0	3403.7	3562.5	3746.1	5087.6	***31079.8**
ChopperCommand	811	7387.8	1615.3	1600.3	6654.1	5982.0	**4480.8**
DemonAttack	152.1	1971.0	4396.7	3182.6	7715.5	6346.4	***19530.2**
DoubleDunk	-18.6	-16.4	-16.7	7.4	20.2	17.4	1.1
Enduro	0	860.5	799.5	2062.5	766.5	2255.6	2341.2
FishingDerby	-91.7	-38.7	12.3	48.4	41.9	37.6	31.4
Freeway	0	29.6	25.8	33.5	33.5	33.2	32.8
Frostbite	65.2	4334.7	760.2	8022.8	7824.9	5697.2	***8401.5**
Gopher	257.6	2412.5	3495.8	3917.1	11192.6	7102.1	***12252.9**
Gravitar	173.0	3351.4	250.7	821.3	1083.5	926.2	703.5
Qbert	163.9	13455.0	8216.2	17228.0	15045.5	17121.4	15806.9
Robotank	2.2	11.9	25.8	53.6	65.9	63.6	48.7
Seaquest	68.4	42054.7	1585.9	4667.9	20081.3	3916.2	**4791.5**
Skiing	-17098.1	-4336.9	-17038.2	-14401.6	-13755.6	-17960.1	***-9021.2**
Solaris	1236.3	12326.7	2029.5	2361.7	2234.5	2922.2	***7145.3**
SpaceInvaders	148.0	1668.7	1361.1	940.2	3115.0	1908.0	**1602.4**
Tennis	-23.8	-9.3	-0.1	19.2	3.5	0.0	*15.4
TimePilot	3568.0	5229.2	3200.9	6622.8	9820.6	9324.4	5304.1
Tutankham	11.4	167.6	138.8	209.9	250.4	252.2	145.4
UpNDown	533.4	11693.2	10405.6	29890.1	44327.6	18790.7	**32155.5**
Venture	0	1187.5	50.8	1099.6	1134.5	1488.9	1000.0
YarsRevenge	3092.9	54576.9	20375.7	66055.9	57960.3	31864.4	***67545.8**
Zaxxon	32.5	9173.3	1928.6	8177.2	12048.6	14117.5	7010.3

Table 5: Raw scores across 34 games. We report the best scores for DQN, QR-DQN, IQN*, and Rainbow on 50M frames, averaged by 5 seeds. Reference values were provided by Dopamine framework [2]. **Bolds** are wins against DQN, QR-DQN, and *asterisk are wins over IQN* and Rainbow. **Note that IQN* and Rainbow implemented in Dopamine framework applied n-step updates with $n = 3$ which improves performance.**