The Nature of Temporal Difference Errors in Multi-step Distributional Reinforcement Learning

Yunhao Tang
DeepMind
robintyh@deepmind.com

Mark Rowland
DeepMind
markrowland@deepmind.com

Rémi Munos
DeepMind
munos@deepmind.com

Bernardo Ávila Pires
DeepMind
bavilapires@deepmind.com

Will Dabney
DeepMind
wdabney@deepmind.com

Marc G. Bellemare
Google Brain
bellemare@google.com

Abstract

We study the multi-step off-policy learning approach to distributional RL. Despite the apparent similarity between value-based RL and distributional RL, our study reveals intriguing and fundamental differences between the two cases in the multi-step setting. We identify a novel notion of path-dependent distributional TD error, which is indispensable for principled multi-step distributional RL. The distinction from the value-based case bears important implications on concepts such as backward-view algorithms. Our work provides the first theoretical guarantees on multi-step off-policy distributional RL algorithms, including results that apply to the small number of existing approaches to multi-step distributional RL. In addition, we derive a novel algorithm, Quantile Regression-Retrace, which leads to a deep RL agent QR-DQN-Retrace that shows empirical improvements over QR-DQN on the Atari-57 benchmark. Collectively, we shed light on how unique challenges in multi-step distributional RL can be addressed both in theory and practice.

1 Introduction

The return $\sum_{t=0}^{\infty} \gamma^t R_t$ is a fundamental concept in reinforcement learning (RL). In general, the return is a random variable, whose distribution captures important information such as the stochasticity in future events. While the classic view of value-based RL typically focuses on the expected return [1–3], learning the full return distribution is of both theoretical and practical importance [4–10].

To design efficient algorithms for learning return distributions, a natural idea is to construct distributional equivalents of existing multi-step off-policy value-based algorithms. In value-based RL, multi-step learning tends to propagate useful information more efficiently and off-policy learning is ubiquitous in modern RL systems. Meanwhile, the return distribution shares inherent commonalities with the expected return, thanks to the close connection between the distributional Bellman equation [4–6, 10] and the celebrated value-based Bellman equation [2]. The Bellman equation is foundational to value-based RL algorithms, including many multi-step off-policy methods [11–14]. Due to the apparent similarity between distributional and value-based Bellman equations, should we expect key value-based concepts and algorithms to seamlessly transfer to distributional learning?

Our study indicates that the answer is no. There are critical differences between distributional and value-based RL, which requires a distinct treatment of multi-step learning. Indeed, thanks to the focus on expected returns, the value-based setup offers many unique conceptual and computational simplifications in algorithmic design. However, we find that such simplifications do not hold for distributional learning. Multi-step distributional RL requires a deeper look at the connections between fundamental concepts such as n-step returns, TD errors and importance weights for off-policy learning. To this end, we make the following conceptual, theoretical and algorithmic contributions:

Preprint.

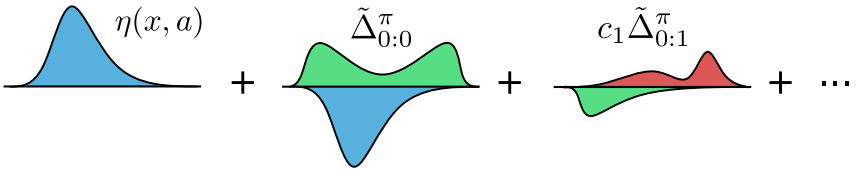


Figure 1: Illustration of a multi-step distributional RL target, constructed as a sum of the initial distribution (left) and weighted distributional TD errors $\widetilde{\Delta}^{\pi}_{0:0}, c_1\widetilde{\Delta}^{\pi}_{0:1}, \ldots$ across multiple time steps (middle and right); see Section 3 for further details and notation. In general, distributional TD errors are signed measures, as reflected by the downwards probability mass; they are also scaled by trace coefficients c_1 to correct for off-policy discrepancies between target and behavior policy.

Distributional TD error. We demonstrate the emergence of a novel notion of path-dependent distributional TD error (Section 4). Intriguingly, as the name suggests, path-dependent distributional TD errors are *path-dependent*, i.e., distributional TD errors at time t depend on the sequence of immediate rewards $(R_s)_{s=0}^{t-1}$. This differs from value-based TD errors, which are path-independent. We will show that the path-dependency property is not an artifact, but rather a fundamental property of distributional learning. We show numerically that naively constructing certain path-independent distributional TD errors does not produce convergent algorithms. The path-dependency property also has conceptual and computational impacts on forward-view estimates and backward-view algorithms.

Theory of multi-step distributional RL. We derive distributional Retrace, a novel and generic multi-step off-policy operator for distributional learning. We prove that distributional Retrace is contractive and has the target return distribution as its fixed point. Distributional Retrace interpolates between the one-step distributional Bellman operator [6] and Monte-Carlo (MC) estimation with importance weighting [15], trading-off the strengths from the two extremes.

Approximate multi-step distributional RL. Finally, we derive Quantile Regression-Retrace, a novel algorithm combining distributional Retrace with quantile representations of distributions [16] (Section 5). One major technical challenge is to define the quantile regression (QR) loss against signed measures, which are unavoidable in sample-based settings. We bypass the issue of ill-defined QR loss and derive unbiased stochastic estimates to the QR loss gradient. This leads up to QR-DQN-Retrace, a deep RL agent with performance improvements over QR-DQN on Atari-57 games.

In Figure 1, we illustrate how the back-up target is computed for multi-step distributional RL. In summary, we take our findings to demonstrate how the set of unique challenges presented by multi-step distributional RL can be addressed both theoretically and empirically. Our study also opens up many exciting research pathways in this domain, paving the way for future investigations.

2 Background

Consider a Markov decision process (MDP) represented as the tuple $(\mathcal{X}, \mathcal{A}, P_R, P, \gamma)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ the action space, $P_R : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathscr{R})$ the reward kernel (with $\mathscr{R}$ a finite set of possible rewards), $P : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$ the transition kernel and $\gamma \in [0, 1)$ the discount factor. In general, we use $\mathscr{P}(A)$ denote a distribution over set A. We assume the reward to take a finite set of values mainly because it is notationally simpler to present results; it is straightforward to extend our results to the general case. Let $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ be a fixed policy. We use $(X_t, A_t, R_t)_{t=0}^{\infty} \sim \pi$ to denote a random trajectory sampled from π, such that $A_t \sim \pi(\cdot|X_t), R_t \sim P_R(\cdot|X_t, A_t), X_{t+1} \sim P(\cdot|X_t, A_t)$. Define $G^{\pi}(x, a) \coloneqq \sum_{t=0}^{\infty} \gamma^t R_t$ as the random return, obtained by following π starting from (x, a). The Q-function $Q^{\pi}(x, a) \coloneqq \mathbb{E}[G^{\pi}(x, a)]$ is defined as the expected return under policy π. For convenience, we also adopt the vector notation $Q \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$. Define the one-step value-based Bellman operator $T^{\pi} : \mathbb{R}^{\mathcal{X}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ such that $T^{\pi}Q(x, a) \coloneqq \mathbb{E}[R_0 + \gamma Q\left(X_1, A_1^{\pi}\right) | X_0 = x, A_0 = a]$ where $Q(X_t, A_t^{\pi}) \coloneqq \sum_a \pi(a|X_t)Q(X_t, a)$. The Q-function Q^{π} satisfies $Q^{\pi} = T^{\pi}Q^{\pi}$ and is also the unique fixed point of T^{π}.

2.1 Distributional reinforcement learning

In general, the return $G^\pi(x,a)$ is a random variable and we define its distribution as $\eta^\pi(x,a) \coloneqq \text{Law}_\pi\left(G^\pi(x,a)\right)$. The return distribution satisfies the distributional Bellman equation [4–6, 17, 10],

$$\eta^\pi(x,a) = \mathbb{E}_\pi\left[\left(\mathsf{b}_{R_0,\gamma}\right)_\# \eta^\pi\left(X_1, A_1^\pi\right) \,\middle|\, X_0 = x, A_0 = a\right], \tag{1}$$

where $\left(\mathsf{b}_{r,\gamma}\right)_\# : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ is the pushforward operation defined through the function $\mathsf{b}_{r,\gamma}(z) = r + \gamma z$ [17]. For convenience, we adopt the notation $\eta^\pi(X_t, A_t^\pi) \coloneqq \sum_a \pi(a|X_t)\eta^\pi(X_t,a)$. Throughout the paper, we focus on the space of distributions with bounded support $\mathscr{P}_\infty(\mathbb{R})$. Let $\eta \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ be any distribution vector, we define the *distributional Bellman operator* $\mathcal{T}^\pi : \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ as follows [17, 10],

$$\mathcal{T}^\pi\eta(x,a) \coloneqq \mathbb{E}\left[\left(\mathsf{b}_{R_0,\gamma}\right)_\#\eta(X_1, A_1^\pi) \mid X_0 = x, A_0 = a\right]. \tag{2}$$

Let η^π be the collection of return distributions under π; the distributional Bellman equation can then be rewritten as $\eta^\pi = \mathcal{T}^\pi\eta^\pi$. The distributional Bellman operator $\mathcal{T}^\pi$ is γ-contractive under the supremum p-Wasserstein distance [16, 10], so that η^π is the unique fixed point of $\mathcal{T}^\pi$. See Appendix B for details of the distance metrics.

2.2 Multi-step off-policy value-based learning

We provide a brief background on the value-based multi-step off-policy setting as a reference for the distributional case discussed below. In off-policy learning, the data is generated under a behavior policy μ, which potentially differs from target policy π. The aim is to evaluate the target Q-function Q^π. As a standard assumption, we require $\text{supp}(\pi(\cdot|x)) \subseteq \text{supp}(\mu(\cdot|x)), \forall x \in \mathcal{X}$. Let $\rho_t \coloneqq \pi(A_t|X_t)/\mu(A_t|X_t)$ be the step-wise importance sampling (IS) ratio at time step t. Step-wise IS ratios are critical in correcting for the off-policy discrepancy between π and μ.

Let $c_t \in [0, \rho_t]$ be a time-dependent trace coefficient. We denote $c_{1:t} = c_1 \cdots c_t$ and define $c_{1:0} = 1$ by convention. Consider a generic form of the return-based off-policy operator $R^{\pi,\mu}$ as in [13],

$$R^{\pi,\mu}Q(x,a) \coloneqq Q(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{\infty} c_{1:t}\gamma^t \underbrace{\left(R_t + \gamma Q\left(X_{t+1}, A_{t+1}^\pi\right) - Q(X_t, A_t)\right)}_{\delta_t^\pi = \text{value-based TD error}}\right], \tag{3}$$

In the above and below, we omit the notation conditioning on $X_0 = x, A_0 = a$ for conciseness. The general form of $R^{\pi,\mu}$ encompasses many important special cases: when on-policy and $c_t = \lambda$, it recovers the Q-function variant of TD(λ) [2, 12]; when $c_t = \lambda\min(\bar{c}, \rho_t)$, it recovers a specific form of Retrace [13]; when $c_t = \rho_t$, it recovers the importance sampling (IS) operator. The back-up target is computed as a mixture over TD errors δ_t^π, each calculated from the one-step transition data. We also define the *discounted TD error* $\widetilde{\delta}_t^\pi = \gamma^t\delta_t^\pi$, which can be interpreted as the difference between n-step returns from two time steps t and $t+1$, as we discuss in Section 4. As we will detail, the property of $\widetilde{\delta}_t^\pi$ marks a significant difference from the distributional RL setting.

By design, $R^{\pi,\mu}$ has Q^π as the unique fixed point. Multi-step updates make use of rewards from multiple time steps, propagating learning signals more efficiently. This is reflected by the fact that $R^{\pi,\mu}$ is β-contractive with $\beta \in [0,\gamma]$ [13] and often contracts to Q^π faster than the one-step Bellman operator T^π. Our goal is to design distributional equivalents of multi-step off-policy operators, which can lead to concrete algorithms with sample-based learning.

3 Multi-step off-policy distributional reinforcement learning

We now present the core theoretical results relating to multi-step distributional operators. In general, the aim is to evaluate the target distribution η^π with access to off-policy data generated under μ.

Below, we use $G_{t':t} = \sum_{s=t'}^{t}\gamma^{s-t'}R_s$ to denote the partial sum of discounted rewards between two time steps $t' \le t$. We define the generic form of multi-step off-policy distributional operator $\mathcal{R}^{\pi,\mu}$ such that for any $\eta \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, its back-up target $\mathcal{R}^{\pi,\mu}\eta(x,a)$ is computed as

$$\eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{\infty} c_{1:t}\cdot\left(\underbrace{\left(\mathsf{b}_{G_{0:t},\gamma^{t+1}}\right)_\#\eta\left(X_{t+1}, A_{t+1}^\pi\right) - \left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_\#\eta(X_t, A_t)}_{\widetilde{\Delta}_{0:t}^\pi = \text{Multi-step Distributional TD error}}\right)\right]. \tag{4}$$

As an effort to simplify the naming, we call $\mathcal{R}^{\pi,\mu}$ the *distributional Retrace* operator. Distributional Retrace only requires $c_t \in [0, \rho_t]$ and represents a large family of distributional operators. Throughout, we will heavily adopt the pushforward notations. This is mainly because instead of directly working with the random variable G^π, we find it much more convenient to express various important multi-step operations with pushfoward notations.

The back-up target $\mathcal{R}^{\pi,\mu}\eta(x,a)$ is written as a weighted sum of the path-dependent distributional TD errors $\widetilde{\Delta}^\pi_{0:t}$, which we extensively discuss in Section 4. Though the form of $\mathcal{R}^{\pi,\mu}$ seems to bear certain similarities to the value-based operator in Equation (3), the critical differences lie in subtle definitions of the distributional TD errors $\widetilde{\Delta}^\pi_{0:t}$ and where to place the traces $c_{1:t}$ for off-policy corrections. We resume to unpack the insights entailed by the design of the operator in Section 4.

Below, we first present theoretical properties of the distributional Retrace operator. We start with a key property which underlies many ensuing theoretical results. Given a fixed n-step reward sequence $r_{0:n-1}$ and a fixed state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$, we call pushfoward distributions of the form $\left(\mathrm{b}_{\sum_{s=0}^{n-1}\gamma^s r_s, \gamma^n}\right)_{\#}\eta(x,a)$ the *n-step target distributions*. Our result shows that the back-up target of Retrace is a convex combination of n-step target distributions with varying values of n.

Lemma 3.1. (**Convex combination**) The Retrace back-up target is a convex combination of n-step target distributions. Formally, there exists an index set $I(x,a)$ such that $\mathcal{R}^{\pi,\mu}\eta(x,a) = \sum_{i\in I(x,a)} w_i \eta_i$ where $w_i \geq 0$, $\sum_{i\in I(x,a)} w_i = 1$ and $(\eta_i)_{i\in I(x,a)}$ are n_i-return target distributions.

Since $\mathcal{R}^{\pi,\mu}\eta \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we can measure the contraction of $\mathcal{R}^{\pi,\mu}$ under probability metrics.

Proposition 3.2. (**Contraction**) $\mathcal{R}^{\pi,\mu}$ is β-contractive under supremum p-Wasserstein distance, where $\beta = \max_{x\in\mathcal{X},a\in\mathcal{A}} \sum_{t=1}^\infty \mathbb{E}_\mu\left[c_1 ... c_{t-1}(1-c_t)\right]\gamma^t \leq \gamma$.

The contraction rate of the distributional Retrace operator is determined by its effective horizon. At one extreme, when $c_t = 0$, the effective horizon is 1 and $\beta = \gamma$, in which case Retrace recovers the one-step operator. At the other extreme, when $c_t = \rho_t$, the effective horizon is infinite which gives $\beta = 0$. This latter case can be understood as correcting for all the off-policy discrepancies with IS, which is very efficient *in expectation* but incurs high variance under sample-based approximations. Proposition 3.2 also implies that the distributional Retrace operator has a unique fixed point.

Proposition 3.3. (**Unique fixed point**) $\mathcal{R}^{\pi,\mu}$ has η^π as the unique fixed point in $\mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$.

The above result suggests that starting with $\eta_0 \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, the recursion $\eta_{k+1} = \mathcal{R}^{\pi,\mu}\eta_k$ produces iterates $(\eta_k)_{k=0}^\infty \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ which converge to η^π in $\overline{W}_p$ at a rate of $\mathcal{O}(\beta^k)$.

4 Understanding multi-step distributional reinforcement learning

Now, we pause and take a closer look at the construction of the distributional Retrace operator. We present a number of insights that distinguish distributional learning from value-based learning.

4.1 Path-dependent TD error

The value-based Retrace back-up target can be written as a mixture of value-based TD errors. To better parse the distributional Retrace operator and draw comparison to the value-based setting, we seek to rewrite the distributional back-up target $\mathcal{R}^{\pi,\mu}\eta(x,a)$ into a weighted sum of some notion of distributional TD errors. To this end, we start with a natural analogy to the value-based TD error.

Definition 4.1. (**Distributional TD error**) Given a transition (X_t, A_t, R_t, X_{t+1}), define the associated distributional TD error as $\Delta^\pi(X_t, A_t, R_t, X_{t+1}) \coloneqq (\mathrm{b}_{R_t,\gamma})_\# \eta\left(X_{t+1}, A^\pi_{t+1}\right) - \eta(X_t, A_t)$.

When the context is clear, we also adopt the concise notation $\Delta^\pi_t = \Delta^\pi(X_t, A_t, R_t, X_{t+1})$. By construction, distributional TD errors are signed measures with zero total mass [10]. The distributional TD error is a natural counterpart to the value-based TD error, because they both stem directly from the corresponding one-step Bellman operators. However, unlike in value-based RL, where TD errors alone suffice to specify the multi-step learning operator (Equation (3)), in distributional RL this is not enough. We introduce the path-dependent distributional TD error, which serves as the building block to distributional Retrace.

Definition 4.2. (**Path-dependent distributional TD error**) Given a trajectory $(X_s, A_s, R_s)_{s=0}^{\infty}$, define the path-dependent distributional TD error at time $t \geq 0$ as follows,

$$\widetilde{\Delta}_{0:t}^{\pi} \coloneqq \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#} \Delta_t^{\pi}. \tag{5}$$

Path-dependent distributional TD errors are defined as a pushforward measures from Δ_t^{π}, where the pushforward operations depend on $G_{0:t-1}$. This equips $\widetilde{\Delta}_{0:t}^{\pi}$ with an intriguing property, *path-dependency*. Concretely, this means that the path-dependent distributional TD error depends on the sequence of rewards $(R_s)_{s=0}^{t-1}$ leading up to step t. With the above definitions, we can finally rewrite the back-up target of distributional Retrace as a weighted sum of path-dependent distributional TD errors $\mathcal{R}^{\pi,\mu}\eta(x,a) = \eta(x,a) + \mathbb{E}_{\mu}[\sum_{t=0}^{\infty} c_{1:t}\widetilde{\Delta}_{0:t}^{\pi}]$. We now illustrate the difference between value-based and distributional TD errors.

Comparison with value-based TD equivalents. The value-based equivalent to the path-dependent distributional TD error is the discounted value-based TD error $\widetilde{\delta}_t^{\pi} = \gamma^t \delta_t^{\pi}$ which we briefly mentioned in Section 2. To see why, note that discounted value-based TD errors allow us to rewrite the value-based Retrace back-up target as $R^{\pi,\mu}Q(x,a) = Q(x,a) + \mathbb{E}_{\mu}[\sum_{t=0}^{\infty} c_{1:t}\widetilde{\delta}_t^{\pi}]$. For direct comparison between the two settings, we rewrite both $\widetilde{\Delta}_{0:t}^{\pi}$ and $\widetilde{\delta}_t^{\pi}$ as the difference between two n-step predictions evaluated at two time steps t and $t+1$,

$$\widetilde{\Delta}_{0:t}^{\pi} = \left(\mathrm{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#} \eta\left(X_{t+1}, A_{t+1}^{\pi}\right) - \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#} \eta(X_t, A_t), \qquad \text{(Distributional)}$$

$$\widetilde{\delta}_t^{\pi} = \left(G_{0:t} + \gamma^{t+1} Q\left(X_{t+1}, A_{t+1}^{\pi}\right)\right) - \left(G_{0:t-1} + \gamma^t Q(X_t, A_t)\right). \qquad \text{(Value-based)}$$

The above rewriting attributes the path-dependency to the fact that the n-step distributional prediction $\left(\mathrm{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#} \eta\left(X_{t+1}, A_{t+1}^{\pi}\right)$ is non-linear in $G_{0:n-1}$. Indeed, in the value-based setting, because $G_{0:t} = G_{0:t-1} + \gamma^t R_t$ the partial sum of rewards $G_{0:t-1}$ cancels out as a common term. This leaves the discounted TD error $\widetilde{\delta}_t^{\pi}$ *path-independent*. In other words, the computation of $\widetilde{\delta}_t^{\pi}$ does not depend on past rewards $(R_s)_{s=0}^{t-1}$. In contrast, in the distributional setting, the pushforward operations are non-linear in the partial sum of rewards $G_{0:t-1}$. As a result, $G_{0:t-1}$ does not cancel out in the definition of $\widetilde{\Delta}_{0:t}^{\pi}$, making the path-dependent TD error $\widetilde{\Delta}_{0:t}^{\pi}$ depend on the past rewards $(R_s)_{s=0}^{t-1}$.

The path-dependent property is not an artifact of the distributional Retrace operator $\mathcal{R}^{\pi,\mu}$; instead, it is an indispensable element for convergent multi-step distributional learning in general. We show this by empirically verifying that multi-step learning operators based on alternative definitions of *path-independent* distributional TD errors are non-convergent even for simple problems.

Numerically non-convergent path-independent operators. Consider the *path-independent* distributional TD error $\overline{\Delta}_t^{\pi} \coloneqq \left(\mathrm{b}_{0,\gamma^t}\right)_{\#} \Delta_t^{\pi}$. We arrived at this definition by dropping the path-dependent term $G_{0:t-1}$ in the pushforward of $\widetilde{\Delta}_{0:t}^{\pi}$. Such a definition seems appealing because when $\eta = \eta^{\pi}$, the error is zero in expectation $\mathbb{E}_{\mu}\left[\overline{\Delta}_t^{\pi} | X_t, A_t\right] = 0$. This implies that we can construct a multi-step operator by a weighted sum of the alternative path-independent TD error $\overline{\mathcal{R}}_n^{\pi,\mu}\eta(x,a) \coloneqq \eta(x,a) + \mathbb{E}_{\mu}\left[\sum_{t=0}^{\infty} c_{1:t}\overline{\Delta}_t^{\pi}\right]$. By construction, $\overline{\mathcal{R}}_n^{\pi,\mu}$ has η^{π} as one fixed point.

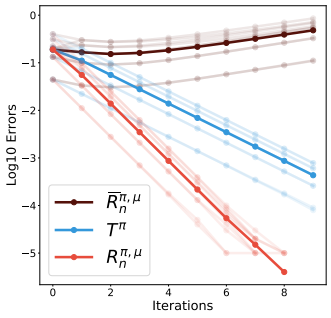


Figure 2: Non-convergent example: comparing $L_p(\mathcal{R}^k\eta_0, \eta^{\pi})$ across iterations. We plot 10 randomly initialized runs. Note $(\overline{\mathcal{R}}^{\pi,\mu})^k\eta_0$ does not converge to η^{π} while others do.

We provide a very simple counterexample on which $\overline{\mathcal{R}}^{\pi,\mu}$ is not contractive: consider an MDP with one state and one action. The state transitions back to itself with a deterministic reward $R_t = 1$. When the discount factor is $\gamma = 0.5$, η^{π} is a Dirac distribution centered at 2. We consider the simple case $c_1 = \rho_1$ and $c_t = 0, \forall t \geq 2$. We use the L_p distance to measure the convergence of the distribution iterates [10]. Figure 2 shows that $(\overline{\mathcal{R}}_n^{\pi,\mu})^k\eta_0$ does not converge to η^{π}, while the one-step Bellman operator $\mathcal{T}^{\pi}$ and distributional Retrace $\mathcal{R}^{\pi,\mu}$ are convergent.

In Appendix C, we discuss yet another alternative to $\widetilde{\Delta}_{0:t}^{\pi}$ designed to be path-independent $\gamma^t\Delta_t^{\pi}$. Though the resulting multi-step operator still has η^{π} as one fixed point, we show numerically that it

is not contractive on the same simple example. These results demonstrate that naively removing the path-dependency might lead to non-convergent multi-step operators.

4.2 Backward-view of distributional multi-step learning

To highlight the difference between distributional and value-based multi-step learning, we discuss the impact that path-dependent distributional TD errors have on the backward-view distributional algorithm. Thus far, distributional back-up targets are expressed in the *forward-view*, i.e., the back-up target at time t is calculated as a function of future transition tuples $(X_s, A_s, R_s)_{s \leq t}$. The forward-view algorithms, unless truncated, wait until the episode finishes to carry out the update, which might be undesirable when the problem is non-episodic or has a very long horizon.

In the *backward-view*, when encountering a distributional TD error Δ_t^π, the algorithm carries out updates for all predictions at time $t' \leq t$ [2]. To this end, the algorithm needs to maintain additional *partial return traces*, i.e., the partial sum of rewards $G_{t':t}$, in order to calculate the path-dependent TD error $\widetilde{\Delta}_t^\pi$. Unlike the value-based state-dependent eligibility traces [2, 18], partial return traces are time-dependent. This implies that in an episode of T steps, value-based backward-view algorithms require memory of size $\min(|\mathcal{X}||\mathcal{A}|, \mathcal{O}(T))$ while the distributional algorithms requires $\mathcal{O}(T)$.

In addition to the added memory complexity, the incremental updates of distributional algorithms are also much more complicated due to the path-dependent TD errors. We remark that the path-independent nature of value-based TD errors greatly simplify the value-based backward-view algorithm. For a more detailed discussion, see Appendix D.

4.3 Importance sampling for multi-step distributional RL

In our initial derivation, we arrived at $\mathcal{R}^{\pi,\mu}$ through the application of importance sampling (IS) in a different way from the value-based setting. We now highlight the subtle differences and caveats.

For a fixed $n \geq 1$, consider the trace coefficient $c_t = \rho_t \mathbb{I}[t < n]$. The back-up target of the resulting Retrace operator reduces to $\mathbb{E}_\mu \left[\rho_{1:n-1} \cdot \left(\mathrm{b}_{G_{0:n-1},\gamma^n}\right)_\# \eta\left(X_n, A_n^\pi\right)\right]$. This can be seen as applying IS to the n-step prediction $\left(\mathrm{b}_{G_{0:n-1},\gamma^n}\right)_\# \eta\left(X_n, A_n^\pi\right)$. As a caveat, note that an appealing alternative approach is to apply IS to $G_{0:n-1}$, producing the estimate $\left(\mathrm{b}_{\rho_{1:n-1}G_{0:n-1},\gamma^n}\right)_\# \eta\left(X_n, A_n^\pi\right)$. This latter estimate does not properly correct for the off-policy discrepancy between π and μ. To see why, note that applying the IS ratio to $G_{0:n-1}$, instead of to the probability of its occurrence, is an artifact of value-based RL because the expected return is linear in $G_{0:t}$ [11]. In general for distributional RL, one should importance weigh the measures instead of sum of rewards.

5 Approximate multi-step distributional reinforcement learning algorithm

We now discuss how the distributional Retrace operator combines with parametric distributions, using the construction of the novel Quantile Regression-Retrace algorithm as a practical example. We focus on the quantile representation because it entails the best empirical performance of large-scale distributional RL [16, 19]. Speficially, we present an application of quantile regression with signed measures, which is interesting in its own right. Below, we start with a brief background on quantile representations [16], followed by details on the proposed algorithm.

Consider parametric distributions of the form: $\frac{1}{m}\sum_{i=1}^m \delta_{z_i}$ for a fixed $m \geq 1$, where $(z_i)_{i=1}^m \in \mathbb{R}$ are a set of parameters indicating the support of the distribution. Let $\mathscr{P}_\mathcal{Q}(\mathbb{R})$ denote the family of distribution $\mathscr{P}_\mathcal{Q}(\mathbb{R}) \coloneqq \{\frac{1}{m}\sum_{i=1}^m \delta_{z_i} | z_i \in \mathbb{R}\}$. We define the projection $\Pi_\mathcal{Q} : \mathscr{P}_\infty(\mathbb{R}) \to \mathscr{P}_\mathcal{Q}(\mathbb{R})$ as $\Pi_\mathcal{Q}\eta = \arg\min_{\nu \in \mathscr{P}_\mathcal{Q}(\mathbb{R})} W_1(\eta, \nu)$, which projects any distribution onto the space of representable distributions in the parametric class under the W_1 distance. With an abuse of notation, we also let $\Pi_\mathcal{Q}$ denote the component-wise projection when applied to vectors. See [16, 10] for more details.

Gradient-based learning via quantile regression. We can use quantile regression [20–22] to calculate the projection $\Pi_\mathcal{Q}\eta$. Let $F_\eta(z), z \in \mathbb{R}$ denote the CDF of a given distribution η. Let F_η^{-1} be the generalized CDF inverse, we define the τ-th quantile as $F_\eta^{-1}(\tau)$ for $\tau \in [0, 1]$. The projection $\Pi_\mathcal{Q}$ is equivalent to computing $z_i = F_\eta^{-1}(\tau_i)$ for $\tau \in (\frac{2i-1}{2m})_{i=1}^m$ [16]. To learn the τ-th quantile for any $\tau \in [0, 1]$, it suffices to solve the quantile regression problem whose optimal solution is $F_\eta^{-1}(\tau)$: $\min_\theta L_\theta^\tau(\eta) \coloneqq \mathbb{E}_{Z\sim\eta}\left[f_\tau(Z - \theta)\right]$ where $f_\tau(u) = u(\tau - \mathbb{I}[u < 0])$. In practice, we carry out the gradient update $\theta \leftarrow \theta - \alpha\nabla_\theta L_\theta^\tau(\eta)$ to find the optimal solution and learn the quantile $\theta \approx F_\eta^{-1}(\tau)$.

5.1 Distributional Retrace with quantile representations

Given an input distribution vector η, we use the distributional Retrace operator to construct the back-up target $\mathcal{R}^{\pi,\mu}\eta$. Then, we use the quantile projection to map the back-up target onto the space of representations $\Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}\eta$. Overall, we are interested in the recursive update: start with any $\eta_0 \in \mathscr{P}_{\mathcal{Q}}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, consider the sequence of distributions generated via $\eta_{k+1} = \Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}\eta_k$. A direct application of Proposition 3.2 allows us to characterize the convergence of the sequence, following the approach of [10].

Theorem 5.1. (**Convergence of quantile distributions**) The projected distributional Retrace operator $\Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}$ is β-contractive under $\overline{W}_\infty$ distance in $\mathscr{P}_{\mathcal{Q}}(\mathbb{R})$. As a result, the above η_k converges to a limiting distribution $\eta^\pi_{\mathcal{R}}$ in $\overline{W}_\infty$, such that $\overline{W}_\infty(\eta_k, \eta^\pi_{\mathcal{R}}) \leq (\beta)^k \overline{W}_\infty(\eta_0, \eta^\pi_{\mathcal{R}})$. Further, the quality of the fixed point is characterized as $\overline{W}_\infty(\eta^\pi_{\mathcal{R}}, \eta^\pi) \leq (1-\beta)^{-1}\overline{W}_\infty(\Pi_{\mathcal{Q}}\eta^\pi, \eta^\pi)$.

Thanks to the faster contraction rate $\beta \leq \gamma$, the advantage of the projected operator $\Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}$ is two-fold: (1) the operator often contracts faster to the limiting distribution $\eta^\pi_{\mathcal{R}}$ than the one-step operator $\mathcal{T}^\pi$ contracts to its own limiting distribution $\eta_{\mathcal{T}^\pi}$ [16]; (2) the limiting distribution $\eta^\pi_{\mathcal{R}}$ also enjoys a better approximation bound to the target distribution. We verify such results in Section 7.

5.2 Quantile Regression-Retrace: distributional Retrace with quantile regression

Below, we use $z_i(x,a)$ to represent the i-th quantile of the distribution at (x,a). Overall, we have a tabular quantile representation $\eta_z(x,a) = \frac{1}{m}\sum_{i=1}^m \delta_{z_i(x,a)}, \forall (x,a) \in \mathcal{X}\times\mathcal{A}$, where we use the notation η_z to stress the distribution's dependency on parameter $z_i(x,a)$. For any given bootstrapping distribution vector $\eta \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, in order to approximate the projected back-up target $\Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}\eta$ with the parameterized quantile distribution η_z, we solve the set of quantile regression problems for all $1 \leq i \leq m, (x,a) \in \mathcal{X}\times\mathcal{A}$,

$$\min_{z_i(x,a)} L^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right), \text{ where } \tau_i = (2i-1)/2m\,.$$

For any fixed (x,a,i), to solve the quantile regression problem, we apply gradient descent on $z_i(x,a)$. In practice, with one sampled trajectory $(X_s, A_s, R_s)_{s=0}^\infty \sim \mu$, the aim is to construct an unbiased stochastic gradient estimate of the QR loss $L^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$. Below, let $\mathrm{b}_t = \mathrm{b}_{G_{0:t-1},\gamma^t}$ for simplicity. We start with a stochastic estimate $\widehat{L}^{\tau_i}_{z_i(x,a)}(\mathcal{R}^{\pi,\mu}\eta(x,a))$ for the QR loss,

$$L^{\tau_i}_{z_i(x,a)}\left(\eta(x,a)\right) + \sum_{t=0}^\infty c_{1:t}\left(L^{\tau_i}_{z_i(x,a)}\left((\mathrm{b}_{t+1})_\#\,\eta\left(X_{t+1}, A^\pi_{t+1}\right)\right) - L^{\tau_i}_{z_i(x,a)}\left((\mathrm{b}_t)_\#\,\eta(X_t,A_t)\right)\right).$$

Since $\widehat{L}^{\tau_i}_{z_i(x,a)}(\mathcal{R}^{\pi,\mu}\eta(x,a))$ is differentiable with $z_i(x,a)$, we use $\nabla_{z_i(x,a)}\widehat{L}^{\tau_i}_{z_i(x,a)}(\mathcal{R}^{\pi,\mu}\eta(x,a))$ as the stochastic gradient estimate. This gradient estimate is unbiased under mild conditions.

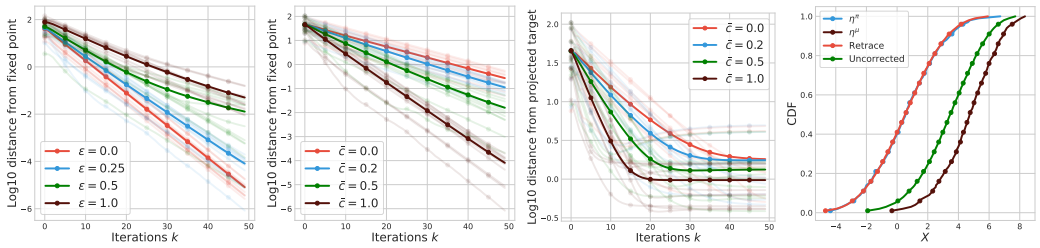
Lemma 5.2. (**Unbiased stochastic QR loss gradient estimate**) Assume that the trajectory terminates within $H < \infty$ steps almost surely, then we have $\mathbb{E}_\mu[\widehat{L}^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)] = L^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$ and $\mathbb{E}_\mu[\nabla_{z_i(x,a)}\widehat{L}^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)] = \nabla_{z_i(x,a)}L^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$.

The above stochastic estimate bypasses the challenge that the QR loss is only defined against distributions, whereas sampled back-up targets $\widehat{R}^{\pi,\mu}\eta(x,a) = \eta(x,a) + \sum_{t=0}^\infty c_{1:t}\widetilde{\Delta}^\pi_{0:t}$ are signed measures in general. In Quantile Regression-Retrace, we use η_z itself as the bootstrapping distribution, such that the algorithm approximates the fixed point iteration $\eta_z \leftarrow \Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}\eta_z$. Concretely, we carry out the following sample-based update

$$z_i(x,a) \leftarrow z_i(x,a) - \alpha\nabla_{z_i(x,a)}\widehat{L}^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta_z(x,a)\right), \text{ for } \forall\ 1 \leq i \leq m, (x,a) \in \mathcal{X}\times\mathcal{A}.$$

5.3 Deep reinforcement learning: QR-DQN-Retrace

We introduce a deep RL implementation of the Quantile Regression-Retrace: QR-DQN-Retrace, where the parametric representation is combined with function approximations [23, 16, 19]. The base agent QR-DQN [23] parameterizes the quantile locations $z_i(x,a;w)$ with the output of a neural network with weights w. Let $\eta(x,a;w) = \frac{1}{m}\sum_{i=1}^m \delta_{z_i(x,a;w)}$ denote the parameterized distribution. QR-DQN-Retrace updates its parameters by stochastic gradient descent on the estimated QR loss, averaged across all m quantile levels $w \leftarrow w - \alpha\frac{1}{m}\sum_{i=1}^m \nabla_w \widehat{L}^{\tau_i}_{z_i(x,a;w)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a;w)\right)$. In practice, the update is further averaged over state-action pairs sampled from a replay buffer.



(a) Off-policyness ε (b) Trace coefficient c_t (c) Fixed point quality (d) Uncorrected

Figure 3: Tabular experiments to illustrate properties of the distributional Retrace operator: we show average results across 10 randomly sampled MDPs. (a) Contraction rate vs. off-policyness; (b) Contraction rate vs. trace coefficient $c_t = \min(\rho_t, \bar{c})$; (c) Fixed point quality vs. trace coefficient c_t; (d) The uncorrected operator introduces bias to the fixed point while Retrace is unbiased.

6 Discussions

Categorical representations. The categorical representation is another commonly used class of parameterized distributions in prior literature [23, 17, 24, 10]. We obtain contractive guarantees for the categorical representation similar to Theorem 5.1. As with QR, this leads both to improved fixed-point approximations and faster convergence. Further, this leads to a deep RL algorithm C51-Retrace. The actor-critic Reactor agent [25] uses C51-Retrace as a critic training algorithm, although without explicit consideration or analysis of the associated distributional operator. See Appendix E for details. We empirically evaluate the stand-alone improvements of C51-Retrace over C51 in Section 7.

Uncorrected methods. The uncorrected methods do not correct for the off-policyness and hence obtain a biased fixed point [26–28]. The Rainbow agent [26] combined n-step uncorrected learning with C51, effectively implementing a distributional operator whose fixed point differs from η^π.

On-policy distributional TD(λ). Nam et al. [29] propose SR(λ), a distributional version of on-policy TD(λ) [30]. In operator form, this can be viewed as a special case of Equation (4) with $\mu = \pi$, $c_t = \lambda$; [29] also introduce a sample-replacement technique for more efficient implementation.

7 Experiments

We carry out a number of experiments to validate the theoretical insights and empirical improvements.

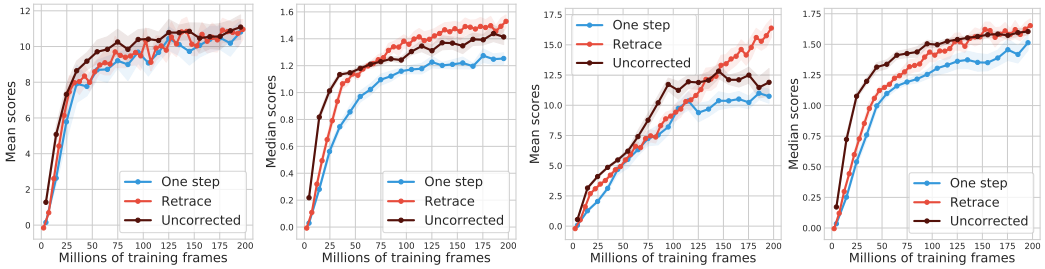
7.1 Illustration of distributional Retrace properties on tabular MDPs

We verify a few important properties of the distributional Retrace operator on a tabular MDP. The results corroborate the theoretical results from previous sections. Throughout, we use quantile representations with $m = 100$ atoms; we obtain similar results for categorical representations. See Appendix F for details on the experiment setup. Let η_0 be the initial distribution, we carry out dynamic programming with $\mathcal{R}^{\pi,\mu}$ and denote $\eta_k = (\mathcal{R}^{\pi,\mu})^k \eta_0$ as the k^{th} distribution iterate.

Impact of off-policyness. We control the level of off-policyness by setting the behavior policy μ to be a uniform policy and the target policy to $\pi = (1-\varepsilon)\mu + \varepsilon\pi_d$ where π_d is a fixed deterministic policy. Moving from $\varepsilon = 0$ to $\varepsilon = 1$, we transition from on-policy to very off-policy. We use $L_p(\eta_k, \eta^\pi_\mathcal{R})$ to measure the contraction rate to the fixed point. Figure 3 shows that as the behavior becomes more off-policy, the contraction slows down, degrading the efficiency of multi-step learning.

Impact of trace coefficient c_t. Throughout, we set $c_t = \min(\rho_t, \bar{c})$ with $\bar{c}$ to control the effective trace length. With a fixed level of off-policyness $\varepsilon = 0.5$, Figure 3(b) shows that increasing $\bar{c}$ speeds up the contraction to the fixed point as predicted by Proposition 3.2.

Quality of fixed point. We next examine how the quality of the fixed point is impacted by $\bar{c}$, by measuring $L_p(\eta_k, \Pi_\mathcal{Q}\eta^\pi)$ as a proxy to $L_p(\eta_k, \eta^\pi)$. As k increases the error flattens, at which point we take the converged value to be $L_p(\eta^\pi_\mathcal{R}, \Pi_\mathcal{Q}\eta^\pi)$ which measures the fixed point quality. Figure 3(c) shows when $\bar{c}$ increases, the fixed point quality improves, in line with the Theorem 5.1. This phenomenon does not arise in *tabular* non-distributional reinforcement learning, although related phenomena do occur when using function approximation techniques.



(a) C51 (b) QR-DQN

Figure 4: Deep RL experiments on Atari-57 games for (a) C51 and (b) QR-DQN. We compare the one-step baseline agent against the multi-step variants (Retrace and uncorrected n-step). For all multi-step variants, we use $n = 3$. For each agent, we calculate the mean and median performance across all games, and we plot the mean $\pm$ standard error across 3 seeds. In almost all settings, multi-step variants provide clear advantage over the one-step baseline algorithm.

Bias of uncorrected methods. Finally, we illustrate a critical difference between Retrace and uncorrected n-step methods [26]: the bias to the fixed point. Figure 3(d) shows that uncorrected n-step arrives at a fixed point in between η^π and η^μ, showing an obvious bias from η^π.

7.2 Deep reinforcement learning

We consider the control setting where the target policy π is the greedy policy with respect to the Q-function induced by the parameterized distribution. Because the training data is sampled from a replay, the behavior policy μ is ε-greedy with respect to Q-functions induced by previous copies of the parameterized distribution. We evaluate the performance of deep RL agents on 57 Atari games [31]. To ensure fair comparison across games, we compute the human normalized scores for each agent, and compare their evaluated mean and median scores across all 57 games during training.

Deep RL agents. The multi-step agents adopt exactly the same hyperparameters as the baseline agents. The only difference is the back-up target. For completeness of results, we show the combination of Retrace with both C51 and QR-DQN. For QR-DQN, we use the Huber loss for quantile regression, which is a thresholded variant of the QR loss [16]. Throughout, we use $c_t = \lambda \min(\rho_t, \bar{c})$ with $\bar{c} = 1$ as in [13]. See Appendix F for details. In practice, sampled trajectories are truncated at length n. We also adapt Retrace to the n-step case, see Appendix A.

Results. Figure 4 compares one-step baseline, Retrace and uncorrected n-step [26]. For C51, both multi-step methods clearly improve the median performance over the one-step baseline. Retrace slightly outperforms uncorrected n-step towards the end of learning. For QR-DQN, all multi-step algorithms achieve clear performance gains. Retrace significantly outperforms the uncorrected n-step with the mean performance, while obtaining similar results on the median performance. Overall, distributional Retrace achieves a clear improvement over the one-step baselines. The uncorrected n-step method typically takes off faster than Retrace but may to slightly worse performance.

Finally, note that in the value-based setting, uncorrected methods are generally more high-performing than Retrace, potentially due to a favorable trade-off between contraction rate and fixed-point bias [32]. Our results add to the benefits of off-policy corrections in the control setting.

8 Conclusion

We have identified a number of fundamental conceptual differences between value-based and distributional RL in multi-step settings. Central to such differences is the novel notion of path-dependent distributional TD error, which naturally arises from the multi-step distributional RL problem. Building on this understanding, we have developed the first principled multi-step off-policy distributional operator Retrace. We have also developed an approximate distributional RL algorithm, Quantile Regression-Retrace, which makes distributional Retrace highly competitive in both tabular and high-dimensional setups. This paper also opens up a several avenues for future research, such as the interaction between multi-step distributional RL and signed measures, and the convergence theory of stochastic approximations for multi-step distributional RL.

References

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[2] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[3] Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2010.

[4] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2010.

[5] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

[6] Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2017.

[7] 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*, 2019.

[8] Cristian Bodnar, Adrian Li, Karol Hausman, Peter Pastor, and Mrinal Kalakrishnan. Quantile QT-OPT for risk-aware vision-based robotic grasping. In *Proceedings of Robotics: Science and Systems*, 2020.

[9] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[10] Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2022. http://www.distributional-rl.org.

[11] Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the International Conference on Machine Learning*, 2001.

[12] Anna Harutyunyan, Marc G. Bellemare, Tom Stepleton, and Rémi Munos. Q(λ) with off-policy corrections. In *Proceedings of the International Conference on Algorithmic Learning Theory*, 2016.

[13] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, 2016.

[14] Ashique Rupam Mahmood, Huizhen Yu, and Richard S Sutton. Multi-step off-policy learning without importance sampling ratios. *arXiv*, 2017.

[15] Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 2021.

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

[17] Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018.

[18] Hado van Hasselt, Sephora Madjiheurem, Matteo Hessel, David Silver, André Barreto, and Diana Borsa. Expected eligibility traces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[19] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2018.

[20] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[21] Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005.

[22] Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of Quantile Regression*. CRC Press, 2017.

[23] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

[24] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2019.

[25] Audrunas Gruslys, Will Dabney, Mohammad Gheshlaghi Azar, Bilal Piot, Marc G. Bellemare, and Rémi Munos. The Reactor: A fast and sample-efficient actor-critic agent for reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2018.

[26] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[27] Steven Kapturowski, Georg Ostrovski, John Quan, Rémi Munos, and Will Dabney. Recurrent experience replay in distributed reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2019.

[28] Tadashi Kozuno, Yunhao Tang, Mark Rowland, Rémi Munos, Steven Kapturowski, Will Dabney, Michal Valko, and David Abel. Revisiting Peng's Q(λ) for modern reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2021.

[29] Daniel W. Nam, Younghoon Kim, and Chan Y. Park. GMAC: A distributional perspective on actor-critic framework. In *Proceedings of the International Conference on Machine Learning*, 2021.

[30] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[31] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[32] Mark Rowland, Will Dabney, and Rémi Munos. Adaptive trade-offs in off-policy learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020.

[33] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[34] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill New York, 1976.

[35] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[36] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[37] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. Jax: composable transformations of Python+NumPy programs. 2018.

[38] Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Claudio Fantacci, Jonathan Godwin, Chris Jones, Tom Hennigan, Matteo Hessel, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Lena Martens, Vladimir Mikulik, Tamara Norman, John Quan, George Papamakarios, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Wojciech Stokowiec, and Fabio Viola. The DeepMind JAX ecosystem. 2010.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[40] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

The Nature of Temporal Difference Errors in Multi-step Distributional Reinforcement Learning: Appendices

A Extension of distributional Retrace to n-step truncated trajectories

The n-step truncated version of distributional Retrace is defined as

$$\mathcal{R}_n^{\pi,\mu}\eta(x,a) = \eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{n} c_{1:t}\widetilde{\Delta}_{0:t}^{\pi}\right],$$

which sums the path-dependent distributional TD errors up to time n. Compared to the original definition of distributional Retrace, this n-step operator is more practical to implement. This operator enjoys all the theoretical properties of the original distributional Retrace, with a slight difference on the contraction rate. Intuitively, the operator bootstraps with at most n steps, which limits the effective horizon of the operator to be $\leq n$. It is straightforward to show that the operator is β_n-contractive under $\overline{W}_p$ with $\beta_n \in (\beta, \gamma]$. As $n \to \infty$, $\beta_n \to \beta$.

B Distance metrics

We provide a brief review on the distance metrics used in this work. We refer readers to [10] for a complete background.

B.1 Wasserstein distance

Let $\eta_1, \eta_2 \in \mathscr{P}_\infty(\mathbb{R})$ be two distribution measures. Let F_η be the CDF of η. The p-Wasserstein distance can be computed as

$$W_p(\eta_1,\eta_2) \coloneqq \left(\int_{[0,1]} |F_{\eta_1}^{-1}(z) - F_{\eta_2}^{-1}(z)|^p dz\right)^{1/p}.$$

Note that the above definition is equivalent to the more traditional definition based on optimal transport; indeed, $F_{\eta_i}^{-1}(z), z \sim \text{Uniform}(0,1), i \in \{1,2\}$ can be understood as the optimal coupling between the two distributions. The above definition is a proper distance metric if $p \geq 1$.

For any distribution vector $\eta_1, \eta_2 \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we can define the supremum p-Wasserstein distance as

$$\overline{W}_p(\eta_1,\eta_2) \coloneqq \max_{x,a} W_p(\eta_1(x,a), \eta_2(x,a)).$$

B.2 L_p distance

Let $\eta_1, \eta_2 \in \mathscr{P}_\infty(\mathbb{R})$ be two distribution measures. Let F_η be the CDF of η. The L_p distance is defined as

$$L_p(\eta_1,\eta_2) \coloneqq \left(\int_{\mathbb{R}} |F_{\eta_1}(z) - F_{\eta_2}(z)|^p \, dz\right)^{1/p}.$$

The above definition is a proper distance metric when $p \geq 1$.

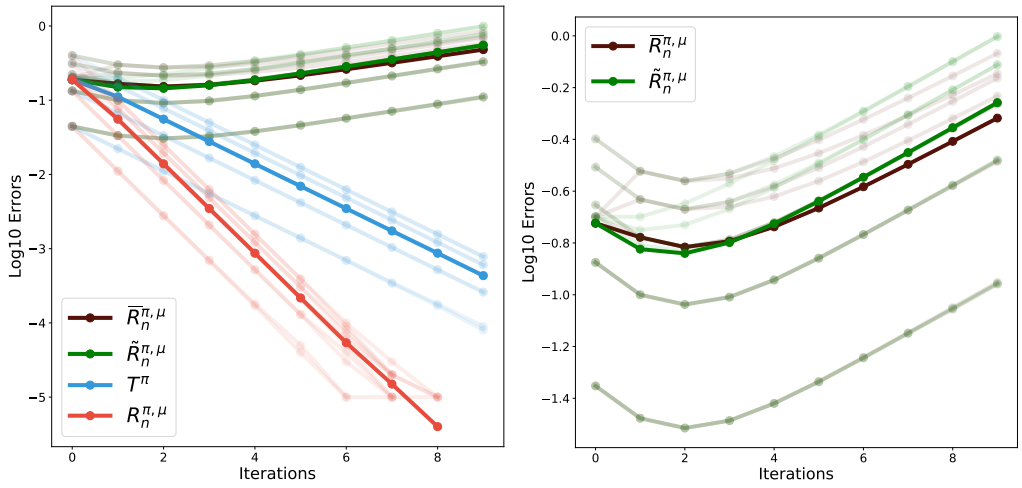
For any distribution vector $\eta_1, \eta_2 \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ or signed measure vector $\eta_1, \eta_2 \in \mathcal{M}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we can define the supremum Cramér-p distance as

$$\overline{L}_p(\eta_1,\eta_2) \coloneqq \max_{x,a} L_p(\eta_1(x,a), \eta_2(x,a)).$$

C Numerically non-convergent behavior of alternative multi-step operators

We consider another alternative definition of path-independent alternative to the path-dependent TD error $\gamma^t\Delta_t^\pi$. The primary motivation for such a path-dependent TD error is that the discounted value-based TD error takes the form $\widetilde{\delta}_t^\pi = \gamma^t\delta_t^\pi$. The resulting multi-step operator is

$$\widetilde{\mathcal{R}}^{\pi,\mu}\eta(x,a) = \eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{\infty} c_{1:t}\gamma^t\Delta_t^\pi\right].$$



(a) Full results for all operators

(b) Comparing two alternatives

Figure 5: Illustration of non-convergent behavior of alternative multi-step operators: for both plots, we show the mean and per-run results across 10 different initial Dirac distributions η_0. (a) the full comparison between all operators. Two alternative operators do not converge while one-step Bellman operator and distributional Retrace both converge; (b) we zoom in on the difference between the two alternative operators.

With the same toy example as in the paper: an one-state one-action MDP with a deterministic reward $R_t = 1$ and discount factor $\gamma = 0.5$. The target distribution η^π is a Dirac distribution centering at 2. Let $\eta_k = (\mathcal{R})^k \eta_0$ be the k-th distribution iterate by applying the operator $\mathcal{R} \in \{\mathcal{R}^{\pi,\mu}, \widetilde{\mathcal{R}}^{\pi,\mu}, \widetilde{\mathcal{R}}^{\pi,\mu}, \mathcal{T}^\pi\}$, we show the L_p distance between the iterates and η^π in Figure 5. It is clear that alternative multi-step operators do not converge to the correct fixed point.

D Backward-view algorithm for multi-step distributional RL

We now describe a backward-view algorithm for multi-step distributional RL with quantile representations. For simplicity, we consider the on-policy case $\pi = \mu$ and $c_t = \lambda$. To implement $\mathcal{R}^{\pi,\mu}$ in the backward-view, at each time step t and a past time step $t' \leq t$, the algorithm needs to maintain two novel traces distinct from the classic eligibility traces [2]: (1) partial return traces $G_{t':t}$, which correspond to the partial sum of rewards between two time steps $t' \leq t$; (2) modified eligibility traces, defined as $e_{t',t} := \lambda^{t-t'}$, which measures the trace decay between two time steps $t' \leq t$. At a new time step $t+1$, the new traces are computed recursively: $G_{t':t+1} = R_{t+1} + \gamma G_{t',t}, e_{t',t+1} = \lambda e_{t',t}$.

We assume the algorithm maintains a table of quantile distributions with m atoms: $\eta(x,a) = \frac{1}{m}\sum_{i=1}^m \delta_{z_i(x,a)}, \forall (x,a) \in \mathcal{X} \times \mathcal{A}$. For any fixed (x,a), define $T_t(x,a) := \{s | X_s = x, A_s = a, 0 \leq s \leq t\}$ be the set of time steps before time t at which (x,a) is visited. Now, upon arriving at X_{t+1}, we observe the TD error Δ_t^π. Recall that $L_\theta^\tau(\eta)$ denote the QR loss of parameter θ at quantile level τ and against the distrbution η. To more conveniently describe the update, we define the QR loss against the path-dependent TD error

$$\left(\mathrm{b}_{G_{s:t-1},\gamma^{t-s}}\right)_\# \widetilde{\Delta}_{0:t}^\pi = \left(\mathrm{b}_{G_{s:t},\gamma^{t+1-s}}\right)_\# \eta(X_{t+1}, A_{t+1}^\pi) - \left(\mathrm{b}_{G_{s:t-1},\gamma^{t-s}}\right)_\# \eta(X_t, A_t)$$

as the difference of the QR losses against the individual distributions,

$$L_\theta^\tau\left(\left(\mathrm{b}_{G_{s:t-1},\gamma^{t-s}}\right)_\# \widetilde{\Delta}_{0:t}^\pi\right) := L_\theta^\tau\left(\left(\mathrm{b}_{G_{s:t},\gamma^{t+1-s}}\right)_\# \eta(X_{t+1}, A_{t+1}^\pi)\right) - L_\theta^\tau\left(\left(\mathrm{b}_{G_{s:t-1},\gamma^{t-s}}\right)_\# \eta(X_t, A_t)\right).$$

Note that the QR loss can be computed using the transition data we have seen so far. We now perform the a gradient update for all entries in the table $(x,a) \in \mathcal{X} \times \mathcal{A}$ and $1 \leq i \leq m$ (in practice, we update entries that correspond to visited state-action pairs):

$$z_i(x,a) \leftarrow z_i(x,a) - \alpha \sum_{s \in T_t(x,a)} e_{s,t} \nabla_{z_i(x,a)} L_\theta^{\tau_i}\left(\left(\mathrm{b}_{G_{s:t-1},\gamma^{t-s}}\right)_\# \widetilde{\Delta}_{0:t}^\pi\right),$$

where $\tau_i = \frac{2i-1}{2m}$. For any fixed (x, a), the above algorithm effectively aggregates updates from time steps $s \in T_t(x, a)$ at which (x, a) is visited.

D.1 Simplifications for value-based RL

We now discuss how the path-independent value-based TD errors greatly simplify the value-based backward-view algorithm. Following the above notations, assume the algorithm maintains a table of Q-function $Q(x, a)$, we can construct incremental backward-view update for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ as follows, by replacing the path-dependent distributional TD error $\widetilde{\Delta}^{\pi}_{0:t}$ by the discounted TD error $\widetilde{\delta}^{\pi}_t$

$$Q(x, a) \leftarrow Q(x, a) - \alpha \sum_{s \in T_t(x,a)} e_{s,t} \widetilde{\delta}^{\pi}_t.$$

Since δ^{π}_t does not depend on the past rewards and is state-action dependent, we can simplify the summation over $s \in T_t(x, a)$ by defining the state-depedent eligibility traces [2] as a replacement to $e_{s,t}$,

$$\widetilde{e}(x, a) \leftarrow \gamma\lambda\widetilde{e}(x, a) + \mathbb{I}[X_t = x, A_t = a].$$

As a result, the above update reduces to

$$Q(x, a) \leftarrow Q(x, a) - \alpha\widetilde{e}(x, a)\delta^{\pi}_t,$$

which recovers the classic backward-view update.

D.2 Non-equivalence of forward-view and backward-view algorithms

In value-based RL, forward-view and backward-view algorithms are equivalent given that the trajectory does not visit the same state twice [2]. However, such an equivalence does not generally hold in distributional RL. Indeed, consider the following counterexample in the case of the quantile representation.

Consider a three-step MDP with deterministic transition $x_1 \to x_2 \to x_3$. There is no action and no reward on the transition. The state x_3 is terminal with a deterministic terminal value $r_3 = 1$. We consider $m = 1$ atom and let the quantile parameters be $\theta_1 = 0$ and $\theta_2 = 1$ at states x_1, x_2 respectively. In this case, the quantile representation learns the median of the target distribution with $\tau = 0.5$.

Now, we consider the update at θ_1 with both forward-view and backward-view implementation of the two-step Bellman operator $\mathcal{T}^{\pi}_2\eta(x) = \mathbb{E}\left[(\mathsf{b}_{0,\gamma^2})_{\#}\eta(X_2, \pi(X_2))|X_0 = x\right]$, which can be obtained from distributional Retrace by setting $c_t = \rho_t$. The target distribution at x_1 is a Dirac distribution centering at γ^2.

Forward-view update. Below, we use δ_x to denote a Dirac distribution at x. In the forward-view, the back-up distribution is

$$\mathbb{E}\left[(\mathsf{b}_{0,\gamma^2})_{\#}\eta(X_2, \pi(X_2))\right] = \delta_{\gamma^2}.$$

The gradient update to θ_1 is thus

$$\theta_1^{(\text{fwd})} = \theta_1 - \alpha\nabla_{\theta_1} L^{0.5}_{\theta_1}\left(\delta_{\gamma^2}\right) = \theta_1 + \alpha\left(0.5 - \mathbb{I}\left[\gamma^2 < \theta_1\right]\right).$$

Backward-view update. To implement the backward-view update, we make clear of the two path-dependent distributional TD errors at two consecutive time steps

$$\widetilde{\Delta}^{\pi}_0 = \delta_\gamma - \delta_0, \;\; \widetilde{\Delta}^{\pi}_1 = (\mathsf{b}_{0,\gamma})_{\#}\left(\delta_{\gamma\theta_2} - \delta_{\theta_1}\right) = \delta_{\gamma^2} - \delta_\gamma$$

The update consists of two steps:

$$\begin{aligned}
\theta_1' &= \theta_1 - \alpha\nabla_{\theta_1} L^{0.5}_{\theta_1}\left(\delta_\gamma\right) = \theta_1 + \alpha\left(0.5 - \mathbb{I}\left[\gamma < \theta_1\right]\right), \\
\theta_1^{(\text{bwd})} &= \theta_1' - \alpha\left(\nabla_{\theta_1'} L^{0.5}_{\theta_1'}\left(\delta^2_\gamma\right) - \nabla_{\theta_1'} L^{0.5}_{\theta_1'}\left(\delta_\gamma\right)\right) \\
&= \theta_1' + \alpha\left(0.5 - \mathbb{I}[\gamma^2 < \theta_1']\right) - \alpha\left(0.5 - \mathbb{I}[\gamma < \theta_1']\right).
\end{aligned}$$

Overall, we have

$$\begin{aligned}
\theta_1^{(\text{bwd})} &= \theta_1 + \alpha\left(0.5 - \mathbb{I}\left[\gamma < \theta_1\right]\right) + \alpha\left(0.5 - \mathbb{I}[\gamma^2 < \theta_1']\right) - \alpha\left(0.5 - \mathbb{I}[\gamma < \theta_1']\right) \\
&= 0.5\alpha - \alpha\mathbb{I}[\gamma^2 < 0.5\alpha] + \mathbb{I}[\gamma < 0.5\alpha].
\end{aligned}$$

Now, let $\alpha \in (2\gamma^2, 2\gamma)$ such that $0.5\alpha \in (\gamma^2, \gamma)$, we have $\theta_1^{\text{bwd}} = 0.5\alpha - \alpha = -0.5\alpha \neq \theta_1^{(\text{fwd})}$.

D.3 Discussion on memory complexity

The return traces $G_{t',t}$ and modified eligibility traces $e_{t',t}$ are time-dependent, which is a direct implication from the fact that distributional TD errors are path-dependent. Indeed, to calculate the distributional TD error $\widetilde{\Delta}^{\pi}_{t':t}$, it is necessary to keep track $G_{t',t}$ in the backward-view algorithm. This differs from the classic eligibility traces, which are state-action-dependent [2, 18]. We remark that the state-action-dependency of eligibility traces result from the fact that value-based TD errors Δ^{π}_{t} are path-independent. The time-dependency greatly influences the memory complexity of the algorithm: when an episode is of length T, value-based backward-view algorithm requires memory of size $\min(|\mathcal{X}||\mathcal{A}|, T)$ to store all eligibility traces. On the other hand, the distributional backward-view algorithm requires $\mathcal{O}(T)$.

E Distributional Retrace with categorical representations

We start by showing that the distributional Retrace operator is β_{L_p}-contractive under the $\overline{L}_p$ distance for $p \geq 1$. As a comparison, the one-step distributional Bellman operator $\mathcal{T}^{\pi}$ is $\gamma^{1/p}$-contractive under $\overline{L}_p$ [17].

Lemma E.1. (**Contraction in** $\overline{L}_p$) $\mathcal{R}^{\pi,\mu}$ is β_{L_p}-contractive under supremum L_p distance for $p \geq 1$, where $\beta_{L_p} \in [0, \gamma]$. Specifically, we have $\beta_{L_p} = \max_{x\in\mathcal{X}, a\in\mathcal{A}} \left(\sum_{t=1}^{\infty} \mathbb{E}_{\mu}\left[c_1 ... c_{t-1}(1-c_t)\right] \gamma^t\right)^{1/p}$.

Proof. The proof is similar to the proof of Proposition 3.2: the result follows by combining the convex combination property of distributional Retrace in Lemma 3.1 with the p-convexity of L_p distance [10]. □

E.1 Categorical representation

In categorical representations [23], we consider parametric distributions of the form for a fixed $m \geq 1$, $\sum_{i=1}^{m} p_i \delta_{z_i}$, where $(z_i)_{i=1}^{m} \in \mathbb{R}$ are a fixed set of atoms and $(p_i)_{i=1}^{m}$ is a categorical distribution such that $\sum_{i=1}^{m} p_i = 1$ and $p_i \geq 0$. Denote the class of such distributions as $\mathscr{P}_{\mathcal{C}}(\mathbb{R}) := \{\sum_{i=1}^{m} p_i \delta_{z_i} | \sum_{i=1}^{m} p_i = 1, p_i \geq 0\}$. For simplicity, we assume that the target return is supported on the set of atoms $[R_{\text{MIN}}/(1-\gamma), R_{\text{MAX}}/(1-\gamma)] \subset [z_1, z_m]$.

We introduce the projection that maps from an initial back-up distribution to the categorical parametric class: $\Pi_{\mathcal{C}} : \mathscr{P}_{\infty}(\mathbb{R}) \to \mathscr{P}_{\mathcal{C}}(\mathbb{R})$ defined as $\Pi_{\mathcal{C}}\eta := \arg\min_{\nu\in\mathscr{P}_{\mathcal{C}}(\mathbb{R})} L_2(\nu, \eta), \forall \nu \in \mathscr{P}_{\infty}(\mathbb{R})$. The projection can be easily calculated as described in [6, 17]. For any distribution vector $\eta \in \mathscr{P}_{\infty}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, define $\Pi_{\mathcal{C}}\eta$ as the component-wise projection. Now, given the composed operator $\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu} : \mathscr{P}_{\infty}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}_{\mathcal{C}}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we characterize the convergence of the seququence $\eta_k = (\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu})^k \eta_0$.

Theorem E.2. (**Convergence of categorical distributions**) The projected distributional Retrace operator $\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}$ is β_{L_2}-contractive under $\overline{L}_2$ distance in $\mathscr{P}_{\mathcal{Q}}(\mathbb{R})$. As a result, the above η_k converges to a limiting distribution $\eta^{\pi}_{\mathcal{R}}$ in $\overline{L}_2$, such that $\overline{L}_2(\eta_k, \eta^{\pi}_{\mathcal{R}}) \leq (\beta_{L_2})^k \overline{L}_2(\eta_0, \eta^{\pi}_{\mathcal{R}})$. Further, the quality of the fixed point is characterized as $\overline{L}_2(\eta^{\pi}_{\mathcal{R}}, \eta^{\pi}) \leq (1-\beta_{L_2})^{-1} \overline{L}_2(\Pi_{\mathcal{C}}\eta^{\pi}, \eta^{\pi})$.

Proof. The above theorem follows from Lemma E.1. Indeed, since $\Pi_{\mathcal{Q}}$ is a non-expansion in supremum Cramér distance $\overline{L}_2$ [17], the composed operator $\Pi_{\mathcal{Q}}\mathcal{R}^{\pi,\mu}$ is β_{L_2}-contractive in $\overline{L}_2$. Following the same argument as the proof of Theorem 5.1, we obtain the remaining desired results. □

The distributional Retrace operator also improves over one-step distributional Bellman operator in two aspects: (1) the bound on the contraction rate $\beta_{L_2} \leq \sqrt{\gamma}$ is smaller, usually leading to faster contraction to the fixed point; (2) the bound on the quality of the fixed point is improved.

E.2 Cross-entropy update and C51-Retrace

Unlike in the quantile projection case, where calculating $\Pi_{\mathcal{Q}}\eta$ requires solving a quantile regression minimization problem, the categorical projection can be calculated in an analytic way [17, 10]. Assume the categorical distribution is parameterized as $\eta_w(x, a) = \sum_{i=1}^{m} p_i(x, a; w)\delta_{z_i}$. After

computing the back-up target distribution $\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)$ for a given distribution vector η, the algorithm carries out a gradient-based incremental update

$$w \leftarrow w - \alpha \nabla_w \mathbb{CE}\left[\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a) | \eta_w(x,a)\right],$$

where $\mathbb{CE}(p|q) \coloneqq -\sum_i p_i \log q_i$ denotes the cross-entropy between distribution p and q. For simplicity, we adopt a short-hand notation $\mathbb{CE}(\eta|\eta_w) = \mathbb{CE}_w(\eta)$. Note also that in practice, η can be a slowly updated copy of η_w [33]. As such, the gradient-based update can be understood as approximating the iteration $\eta_{k+1} = \mathcal{R}^{\pi,\mu}\eta_k$. We propose the following unbiased estimate to the cross-entropy $\widehat{\mathbb{CE}}_w\left[\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)\right]$, calculated as follows

$$\mathbb{CE}_w\left(\eta(x,a)\right) + \sum_{t=0}^{\infty} c_{1:t}\left(\mathbb{CE}_w\left(\left(\mathfrak{b}_{t+1}\right)_{\#}\eta\left(X_{t+1}, A^{\pi}_{t+1}\right)\right) - \mathbb{CE}_w\left(\left(\mathfrak{b}_t\right)_{\#}\eta(X_t, A_t)\right)\right).$$

Lemma E.3. (**Unbiased stochastic estimate for categorical update**) Assume that the trajectory terminates within $H < \infty$ steps almost surely, then we have $\mathbb{E}_\mu\left[\widehat{\mathbb{CE}}_w\left(\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right] = \mathbb{CE}_w\left(\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$. Without loss of generality, assume w is a scalar parameter. If there exists a constant $M > 0$ such that $|\nabla_w \mathbb{CE}_w(\eta)| \le M, \forall \eta \in \mathscr{P}_\infty(\mathbb{R})$, then the gradient estimate is also unbiased $\mathbb{E}_\mu\left[\nabla_w\widehat{\mathbb{CE}}_w\left(\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right] = \nabla_w\mathbb{CE}_w\left(\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$.

Proof. The cross-entropy is defined for any distribution $\mathbb{CE}_w(\eta)$. For any signed measure $\nu = \sum_{i=1}^m w_i\eta_i$ with $\eta_i \in \mathscr{P}_\infty(\mathbb{R})$, we define the generalized cross-entropy as

$$\mathbb{CE}_w(\nu) \coloneqq \sum_{i=1}^m w_i \mathbb{CE}_w(\eta_i),$$

Next, we note the cross-entropy is linear in the input distribution (or signed measure). In particular, for a set of N (potentially infinite) coefficients and distributions (signed measures) (a_i, η_i),

$$\mathbb{CE}_w\left(\sum_{i=1}^N a_i\eta_i\right) \coloneqq \sum_{i=1}^m a_i \mathbb{CE}_w(\eta_i).$$

When a_i denotes a distribution, the above rewrites as $\mathbb{CE}_w\left(\mathbb{E}[\eta_i]\right) = \mathbb{E}[\mathbb{CE}(\eta_i)]$. Finally, combining everything together, we have $\mathbb{E}_\mu\left[\widehat{\mathbb{CE}}_w\left(\Pi_{\mathcal{C}}\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right]$ evaluate to

$$= \mathbb{E}_\mu\left[\mathbb{CE}_w\left(\eta(x,a)\right) + \sum_{t=0}^{\infty} c_{1:t}\left(\mathbb{CE}_w\left(\left(\mathfrak{b}_{t+1}\right)_{\#}\eta\left(X_{t+1}, A^{\pi}_{t+1}\right)\right) - \mathbb{CE}_w\left(\left(\mathfrak{b}_t\right)_{\#}\eta(X_t, A_t)\right)\right)\right]$$

$$=_{(a)} \mathbb{E}_\mu\left[\mathbb{CE}\left(\widehat{\mathcal{R}}^{\pi,\mu}\eta(x,a)\right)\right] =_{(b)} \mathbb{E}_\mu\left[\mathbb{CE}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right].$$

In the above, (a) follows from the definition of the cross-entropy with signed measure $\widehat{\mathcal{R}}^{\pi,\mu}\eta(x,a)$ and (b) follows from the linearity property of cross-entropy.

Next, to show that the gradient estimate is unbiased too, the high level idea is to apply dominated convergence theorem (DCT) to justify the exhchange of gradient and expectation [34]. This is similar to the quantile representation case (see proof for Lemma 5.2). To this end, consider the absolute value of the gradient estimate $\left|\nabla_w\widehat{\mathbb{CE}}_w\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right|$, which serves as an upper bound to the gradient estimate. In order to apply DCT, we need to show the expectation of the absolute gradient is finite. Note we have

$$\mathbb{E}_\mu\left[\left|\nabla_w\widehat{\mathbb{CE}}_w\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right|\right]$$

$$= \mathbb{E}_\mu\left[\left|\nabla_w\mathbb{CE}_w\left(\eta(x,a)\right) + \sum_{t=0}^{H} c_{1:t}\left(\nabla_w\mathbb{CE}_w\left(\left(\mathfrak{b}_{t+1}\right)_{\#}\eta\left(X_{t+1}, A^{\pi}_{t+1}\right)\right) - \nabla_w\mathbb{CE}_w\left(\left(\mathfrak{b}_t\right)_{\#}\eta(X_t, A_t)\right)\right)\right|\right]$$

$$\le_{(a)} \mathbb{E}_\mu\left[\left|\nabla_w\mathbb{CE}_w\left(\eta(x,a)\right)\right| + \sum_{t=0}^{H} c_{1:t}\left|\nabla_w\mathbb{CE}_w\left(\left(\mathfrak{b}_{t+1}\right)_{\#}\eta\left(X_{t+1}, A^{\pi}_{t+1}\right)\right) - \nabla_w\mathbb{CE}_w\left(\left(\mathfrak{b}_t\right)_{\#}\eta(X_t, A_t)\right)\right|\right]$$

$$\le_{(b)} \mathbb{E}_\mu\left[M + \sum_{t=0}^{H}\rho^t \cdot M\right] < \infty,$$

where (a) follows from the application of triangle inequality; (b) follows from the fact that the QR loss gradient against a fixed distribution is bounded $\nabla_w \mathbb{CE}_w(\nu) \in [-M, M], \forall \nu \in \mathscr{P}_\infty(\mathbb{R})$ [16].

Hence, with the application DCT, we can exchange the gradient and expectation operator, which yields $\mathbb{E}_\mu\left[\nabla_w \widehat{\mathbb{CE}}_w^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right] = \nabla_w \mathbb{E}_\mu\left[\widehat{\mathbb{CE}}_w^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right] = \nabla_w \mathbb{CE}_w\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$.

□

We remark that the condition on the bounded gradient $|\nabla_w \mathbb{CE}_w(\eta)| \leq M$ is not restrictive. When η_w is adopts a softmax parameterization and w represents the logits, $M = 1$.

Finally, the deep RL agent C51 parameterizes the categorical distribution $p_i(x, a; w)$ with a neural network w at each state action pair (x, a) [23]. When combined with the above algorithm, this produces C51-Retrace.

F Additional experiment details

In this section, we provide detailed information about experiment setups and additional results. All experiments are carried out in Python, using NumPy for numerical computations [35] and Matplotlib for visualization [36]. All deep RL experiments are carried out with Jax [37], specifically making use of the DeepMind Jax ecosystem [38].

F.1 Tabular

We provide additional details on the tabular RL experiments.

Setup. We consider a tabular MDP with $|\mathcal{X}| = 3$ states and $|\mathcal{A}| = 2$ actions. The reward $r(x, a)$ is deterministic and generated from a standard Gaussian distribution. The transition probability $P(\cdot|x, a)$ is sampled from a Dirichlet distribution with parameter $(\Gamma, \Gamma...\Gamma)$ for $\Gamma = 0.5$. The discount factor is fixed as $\gamma = 0.9$. The MDP has a starting state-action pair (x_0, a_0). The behavior policy μ is a uniform policy. The target policy is generated as follows: we first sample a deterministic policy π_d and then compute $\pi = (1 - \varepsilon)\pi_d + \varepsilon\mu$, with parameter ε to control the level of off-policyness.

Quantile distribution and projection. We use $m = 100$ atoms throughout the experiments. Assuming access to the MDP parameters (e.g., reward and transition probability), we can analytically compute the projection $\Pi_\mathcal{Q}$ using a sorting algorithm. See [16, 10] for details.

Evaluation metrics. Let $\eta_k = (\mathcal{R}^{\pi,\mu})^k \eta_0$ be the k-th iterate. We use a few different metrics in Figure 3. Given any particular distributional Retrace operator $\mathcal{R}^{\pi,\mu}$, there exists a fixed point to the composed operator $\Pi_\mathcal{Q}\mathcal{R}^{\pi,\mu}$. Recall that we denote this distribution as $\eta^\pi_\mathcal{R}$. Fig 3(a)-(b) calculates the iterates' distance from the fixed point, evaluated at (x_0, a_0).

$$L_p\left(\eta_k(x_0, a_0), \eta^\pi_\mathcal{R}(x_0, a_0)\right).$$

Fig 3(c) calculates the distance from the projected target distribution $\Pi_\mathcal{Q}\eta^\pi$. Recall that $\Pi_\mathcal{Q}\eta^\pi$ is in some sense the best possible approximation that the current quantile representation can obtain.

$$L_p\left(\eta_k(x_0, a_0), \Pi_\mathcal{Q}\eta^\pi(x_0, a_0)\right).$$

F.2 Deep reinforcement learning

We provide additional details on the deep RL experiments.

Evaluation metrics. For the i-th of the 57 Atari games, we obtain the performance of the agent G_i at any given point in training. The normalized performance is computed as $Z_i = (G_i - U_i)/(H_i - U_i)$ where H_i is the human performance and U_i is the performance of a random policy. Then the mean/median metric is calculated as the mean or median statistics over $(Z_i)_{i=1}^{57}$.

The super human ratio is computed as the number of games such as $Z_i \geq 1$, i.e., $G_i \geq H_i$ where the agent obtains super human performance on the game. Formally, it is compute as $\frac{1}{57}\sum_{i=1}^{57} \mathbb{I}[Z_i \geq 1]$.

Shared properties of all baseline agents. All baseline agents use the same torso architecture as DQN [33] and differ in the head outputs, which we specify below. All agents an Adam optimizer [39] with a fixed learning rate; the optimization is carried out on mini-batches of size 32 uniformly sampled from the replay buffer. For exploration, the agent acts ε-greedy with respect to induced Q-functions, the details of which we specify below. The exploration policy adopts ε that starts with $\varepsilon_{\text{max}} = 1$ and linearly decays to $\varepsilon_{\text{min}} = 0.01$ over training. At evaluation time, the agent adopts $\varepsilon = 0.001$; the small exploration probability is to prevent the agent from getting stuck.

Details of baseline C51 agent. The agent head outputs a matrix of size $|\mathcal{A}| \times m$, which represents the logits to $(p_i(x, a; \theta))_{i=1}^m$. The support $(z_i)_{i=1}^m$ is generated as a uniform array over $[-V_{\text{MAX}}, V_{\text{MAX}}]$. Though V_{MAX} should in theory be determined by R_{MAX}; in practice, it has been found that setting $V_{\text{MAX}} = R_{\text{MAX}}/(1-\gamma)$ leads to highly sub-optimal performance. This is potentially because usually the random returns are far from the extreme values $R_{\text{MAX}}/(1-\gamma)$, and it is better to set V_{MAX} at a smaller value. Here, we set $V_{\text{MAX}} = 10$ and $m = 51$. For details of other hyperparameters, see [6]. The induced Q-function is computed as $Q_\theta(x, a) = \sum_{i=1}^m p_i(x, a; \theta) z_i$.

Details of baseline QR-DQN agent. The agent head outputs a matrix of size $|\mathcal{A}| \times m$, which represents the quantile locations $(z_i(x, a; \theta))_{i=1}^m$. Here, we set $m = 201$. For details of other hyperparameters, see [16]. The induced Q-function is computed as $Q_\theta(x, a) = \frac{1}{m} \sum_{i=1}^m z_i(x, a; \theta)$.

Details of multi-step agents. Multi-step variants use exactly the same hyperparameters as the one-step baseline agent. The only difference is that the agent uses multi-step back-up targets.

The agent stores partial trajectories $(X_t, A_t, R_t, x_t)_{t=0}^{n-1} \sim \mu$ generated under the behavior policy. Here, the behavior policy μ is the ε-greedy policy with respect to a potentially old Q-function (this is because the data at training time is sampled from the replay); the target policy π is the greedy policy with respect to the current Q-function.

G Proof

To simplify the proof, we assume that the immediate random reward takes a finite number of values. It is straightforward to generalize results to the case where the reward takes an infinite number of values (e.g., the random reward has a continuous distribution).

Assumption G.1. (Reward takes a finite number of values) For all state-action pair (x, a), we assume the random reward $R(x, a)$ takes a finite number of values. Let $\widetilde{R}$ be the finite set of values that the reward $\{R(x, a), (x, a) \in \mathcal{X} \times \mathcal{A}\}$ can take.

For any integer $t \geq 1$, Let $\widetilde{R}^t$ denotes the Cartesian product of t copies of $\widetilde{R}$:

$$\widetilde{R}^t := \underbrace{\widetilde{R} \times \widetilde{R} \times ... \times \widetilde{R}}_{t \text{ copies of } \widetilde{R}}.$$

For any fixed t, we let $r_{0:t-1}$ denote the sequence of realizable rewards from time 0 to time $t-1$. Since $\widetilde{R}$ is a finite set, $\widetilde{R}^t$ is also a finite set.

Lemma 3.1. (**Convex combination**) The Retrace back-up target is a convex combination of n-step target distributions. Formally, there exists an index set $I(x, a)$ such that $\mathcal{R}^{\pi,\mu}\eta(x, a) = \sum_{i \in I(x,a)} w_i \eta_i$ where $w_i \geq 0$, $\sum_{i \in I(x,a)} w_i = 1$ and $(\eta_i)_{i \in I(x,a)}$ are n_i-return target distributions.

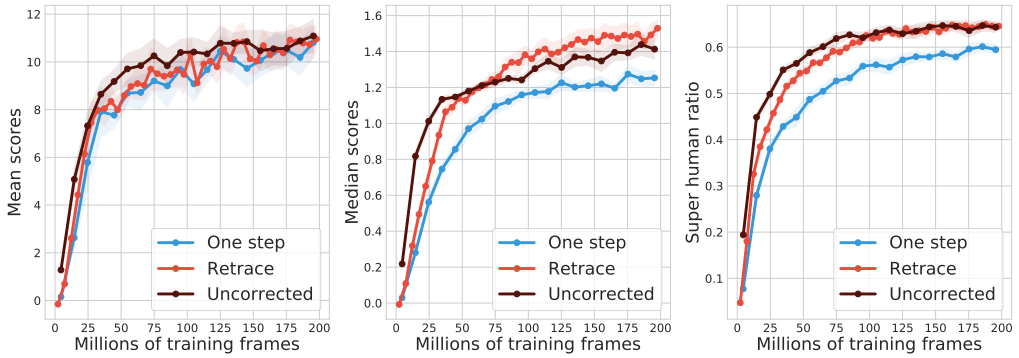
Proof. In general $c_t = c(F_t, A_t)$ where F_t is a filtration of $(X_s, A_s)_{s=0}^t$. To start with, we assume $c_t = c(X_t, A_t)$ to be a Markovian trace coefficient [13]. We start with the simpler case because the proof is greatly simplified with notations and can extend to the general case with some care. We discuss the extension to the general case where $c_t = c(F_t, A_t)$ towards the end of the proof.

For all $t \geq 1$, we define the coefficient

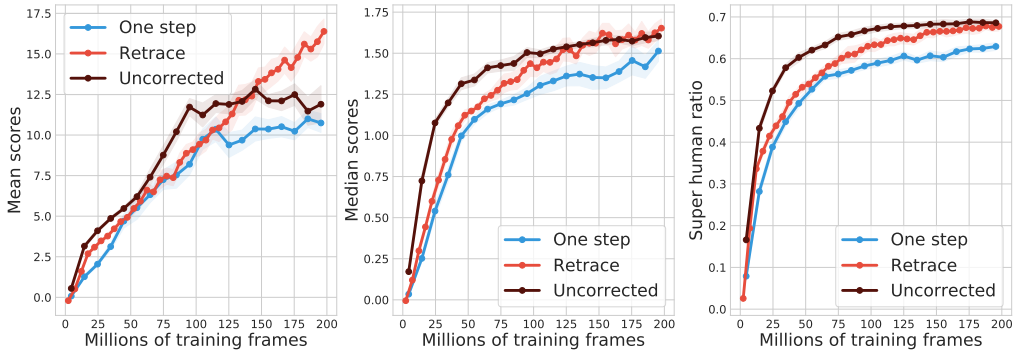
$$w_{y,b,r_{0:t-1}} := \mathbb{E}_\mu \left[c_1 ... c_{t-1} \left(\pi(b|X_t) - c(X_t, b)\mu(b|X_t) \right) \cdot \mathbb{I}[X_t = y] \Pi_{s=0}^{t-1} \mathbb{I}[R_s = r_s] \right].$$

Through careful algebra, we can rewrite the Retrace operator as follows

$$\mathcal{R}^{\pi,\mu}\eta(x, a) = \sum_{t=1}^{\infty} \sum_{y \in \mathcal{X}} \sum_{b \in \mathcal{A}} \sum_{r_{0:t-1} \in \widetilde{R}^t} w_{y,b,r_{0:t-1}} \left(\mathsf{b}_{G_{0:t-1}, \gamma^t} \right)_{\#} \eta(y, b).$$

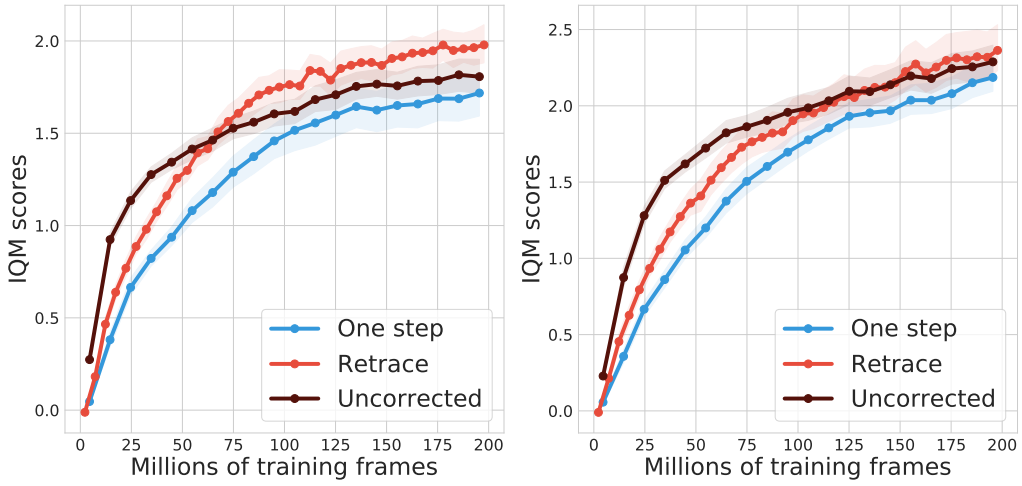


(a) C51



(b) QR-DQN

Figure 6: Deep RL experiments on Atari-57 games for (a) C51 and (b) QR-DQN. We compare the one-step baseline agent against the multi-step variants (Retrace and uncorrected n-step). For all multi-step variants, we use $n = 3$. For each agent, we calculate the mean, median and super human ratio performance across all games, and we plot the mean $\pm$ standard error across 3 seeds. In almost all settings, Multi-step variants provide clear advantage over the one-step baseline algorithm.



(a) C51 (b) QR-DQN

Figure 7: Deep RL experiments on Atari-57 games for (a) C51 and (b) QR-DQN, with the same setup as in Figure 6. Here, we compute the interquartile mean (IQM) with 95% bootstrapped confidence interval [40]. In a nutshell, IQM calculates the mean scores after removing extreme score values, making the performance statistics more robust. Even after excluding extreme scores, Retrace obtains favorable performance compared to the uncorrected and one-step algorithm.

Note that each term of the form $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(y,b)$ corresponds to applying a pushforward operation $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}$ on the distribution $\eta(x,a)$, which means $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(y,b) \in \mathscr{P}_\infty(\mathbb{R})$. Now that we have expressed $\mathcal{R}^{\pi,\mu}\eta(x,a)$ as a linear combination of distributions, we proceed to show that the combination is in fact convex.

Under the assumption $c_t \in [0,\rho_t]$, we have $\pi(b|y) - c(y,b)\mu(b|y) \geq 0$ for all $(y,b) \in \mathcal{X}\times\mathcal{A}$. Therefore, all weights are non-negative. Next, we examine the sum of all coefficients $\sum w_{y,b,r_{0:t-1}} = \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{b\in\mathcal{A}}\sum_{r_{0:t-1}\in\widetilde{R}^t} w_{y,b,r_{0:t-1}}$.

$$\begin{aligned}\sum w_{y,b,r_{0:t-1}} &=_{(a)} \sum_{t=1}^{\infty}\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}}\mathbb{E}_\mu\left[c_1...c_{t-1}\left(\pi(b|X_t) - c(X_t,b)\mu(b|X_t)\right)\cdot\mathbb{I}[X_t=y]\right]\\ &=_{(b)} \sum_{t=1}^{\infty}\mathbb{E}_\mu\left[c_1...c_{t-1}(1-c_t)\right] =_{(c)} 1.\end{aligned}$$

In the above, (a) follows from the fact that $\sum_{r_s\in\widetilde{R}}\mathbb{E}[\mathbb{I}[R_s=r_s]] = 1$; (b) follows from the fact that for all time steps $t \geq 1$, the following is true,

$$\begin{aligned}&\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}}\mathbb{E}_\mu\left[c_1...c_{t-1}\left(\pi(b|X_t) - c(X_t,b)\mu(b|X_t)\right)\cdot\mathbb{I}[X_t=y]\right]\\ &= \sum_{b\in\mathcal{A}}\mathbb{E}_\mu\left[c_1...c_{t-1}\left(\pi(b|X_t) - c(X_t,b)\mu(b|X_t)\right)\right]\\ &= \mathbb{E}_\mu\left[c_1...c_{t-1}\left(1 - \sum_{b\in\mathcal{A}}c(X_t,b)\mu(b|X_t)\right)\right]\\ &= \mathbb{E}_\mu\left[c_1...c_{t-1}(1-c_t)\right].\end{aligned}$$

Finally, (c) is based on the observation that the summation telescopes. Now, by taking the index set to be the set of indices that parameterize $w_{y,b,r_{0:t-1}}$,

$$I(x,a) = \cup_{t=1}^{\infty}\left(y,b,r_{0:t-1}\right)_{y\in\mathcal{X},b\in\mathcal{A},r_{0:t-1}\in\widetilde{R}^t}.$$

We can write $\mathcal{R}^{\pi,\mu}\eta(x,a) = \sum_{i\in I(x,a)} w_i\eta_i$. Note further that for any $i \in I(x,a)$, $\eta_i = (\mathrm{b}_{G_{0:t-1},\gamma^t})_{\#}\eta(y,b)$ is a fixed distribution. The above result suggests that $\mathcal{R}^{\pi,\mu}\eta(x,a)$ is a convex combination of fixed distributions.

Extension to the general case. When $c_t = c(F_t,A_t)$ is filtration dependent, we let $\mathcal{F}_t$ to be the space of the filtration value up to time t. For simplicity with the notation, we assume $\mathcal{F}_t$ contains a finite number of elements, such that below we can adopt the summation notation instead of integral. Define the combination coefficient

$$w_{y,b,f_t,r_{0:t-1}} := \mathbb{E}_\mu\left[c_1...c_{t-1}\left(\pi(b|X_t) - c(F_t,b)\mu(b|X_t)\right)\cdot\mathbb{I}[X_t=y]\Pi_{s=0}^{t-1}\mathbb{I}[R_s=r_s]\right].$$

It is straightforward to verify the following

$$\mathcal{R}^{\pi,\mu}\eta(x,a) = \sum_{t=1}^{\infty}\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}}\sum_{f_t\in\mathcal{F}_t}\sum_{r_{0:t-1}\in\widetilde{R}^t} w_{y,b,f_t,r_{0:t-1}}\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(y,b).$$

In addition, the combination coefficients $w_{y,b,f_t,r_{0:t-1}}$ sum to 1 and are all non-negative. □

Proposition 3.2. (**Contraction**) $\mathcal{R}^{\pi,\mu}$ is β-contractive under supremum p-Wasserstein distance, where $\beta = \max_{x\in\mathcal{X},a\in\mathcal{A}}\sum_{t=1}^{\infty}\mathbb{E}_\mu\left[c_1...c_{t-1}(1-c_t)\right]\gamma^t \leq \gamma$.

Proof. From the proof of Lemma 3.1, we have

$$\mathcal{R}^{\pi,\mu}\eta(x,a) = \sum_{t=1}^{\infty}\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}}\sum_{r_{0:t-1}\in\widetilde{R}^t} w_{y,b,r_{0:t-1}}\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(y,b).$$

Now, we have for any $\eta_1, \eta_2 \in \mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, for any fixed (x, a), we have $W_p\left(\mathcal{R}^{\pi,\mu}\eta_1(x,a), \mathcal{R}^{\pi,\mu}\eta_2(x,a)\right)$ upper bounded as follows

$$
\begin{aligned}
&\leq_{(a)} \sum_{t=1}^{\infty}\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}} w_{y,b,r_{0:t-1}} W_p\left(\left(\mathrm{b}_{\sum_{s=0}^{t-1}\gamma^s r_s,\gamma^t}\right)_{\#}\eta_1(y,b), \left(\mathrm{b}_{\sum_{s=0}^{t-1}\gamma^s r_s,\gamma^t}\right)_{\#}\eta_2(y,b)\right)\\
&\leq_{(b)} \sum_{t=1}^{\infty}\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}} w_{y,b,r_{0:t-1}}\gamma^t W_p\left(\eta_1(y,b), \eta_2(y,b)\right)\\
&\leq_{(c)} \sum_{t=1}^{\infty}\sum_{y\in\mathcal{X}}\sum_{b\in\mathcal{A}} w_{y,b,r_{0:t-1}}\gamma^t \overline{W}_p\left(\eta_1, \eta_2\right)
\end{aligned}
$$

In the above, (a) follows by applying the convexity of the p-Wasserstein distance [10]; (b) follows by the contraction property of the pushforward operation and W_p [10]; (c) follows from the definition of $\overline{W}_p$. By taking the maixmum over (x, a) on both sides of the inequality, we obtain

$$\overline{W}_p(\mathcal{R}^{\pi,\mu}\eta_1, \mathcal{R}^{\pi,\mu}\eta_2) \leq \beta\overline{W}_p(\eta_1, \eta_2).$$

This concludes the proof. □

Lemma G.2. For any fixed (x, a) and scalar $c \in \mathbb{R}$,

$$\left(\mathrm{b}_{c,1}\right)_{\#}\eta^\pi(x,a) = \mathbb{E}_\pi\left[\left(\mathrm{b}_{c+R_0,\gamma}\right)_{\#}\eta^\pi(X_1,A_1) \;\middle|\; X_0 = x, A_0 = a\right]. \tag{6}$$

Proof. Let $B_y := \{x < y | x \in \mathbb{R}\}$ be a subset of $\mathbb{R}$ indexed by $y \in \mathbb{R}$. Since the set of all such sets $\{B_y, y \in \mathbb{R}\}$ is dense in the sigma-field of $\mathbb{R}$ [34], if we can show for two measures η_1, η_2

$$\eta_1\left(B_y\right) = \eta_2\left(B_y\right), \forall y$$

then, $\eta_1(B) = \eta_2(B)$ for all Borel sets in $\mathbb{R}$. Hence, in the following, we seek to show

$$\left(\left(\mathrm{b}_{c,1}\right)_{\#}\eta^\pi(x,a)\right)B_y = \left(\mathbb{E}_\pi\left[\left(\mathrm{b}_{c+R_0,\gamma}\right)_{\#}\eta^\pi(X_1,A_1)\right]\right)B_y, \forall y \in \mathbb{R} \tag{7}$$

Let $F^\pi(y; x, a) := P^\pi(G^\pi(x,a) \leq y) = \eta^\pi(x,a)(B_y), y \in \mathbb{R}$ be the CDF of random variable $G^\pi(x,a)$. The distributional Bellman equation in Equation (1) implies

$$F^\pi(y; x, a) = \mathbb{E}_\pi\left[F^\pi\left(\frac{y - R_0}{\gamma}; X_1, A_1\right)\right], \forall y \in \mathbb{R}.$$

For any constant $c \in \mathbb{R}$, let $y = y' - c$ and plug into the above equality,

$$F^\pi(y' - c; x, a) = \mathbb{E}_\pi\left[F^\pi\left(\frac{y' - c - R_0}{\gamma}; X_1, A_1\right)\right], \forall y' \in \mathbb{R}.$$

Note the LHS is $\left(\left(\mathrm{b}_{c,1}\right)_{\#}\eta^\pi(x,a)\right)(B_y)$ while the RHS is $\left(\mathbb{E}_\pi\left[\left(\mathrm{b}_{c+R_0,\gamma}\right)_{\#}\eta^\pi(X_1,A_1)\right]\right)(B_y)$. This implies that Equation (7) holds and we conclude the proof. □

Proposition 3.3. (**Unique fixed point**) $\mathcal{R}^{\pi,\mu}$ has η^π as the unique fixed point in $\mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$.

Proof. To verify that η^π is a fixed point, it is equivalent to show

$$\mathbb{E}_\mu\left[\sum_{t=0}^{n} c_{1:t}\left(\left(\mathrm{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#}\eta^\pi(X_{t+1}, A^\pi_{t+1}) - \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t, A_t)\right)\right] = \mathbf{0}.$$

Here, the RHS term $\mathbf{0}$ denotes the zero measure, a measure such that for all Borel sets $B \subset \mathbb{R}$, $\mathbf{0}(B) = 0$. We now verify that each of the summation term is a zero measure, i.e.,

$$\mathbb{E}_\mu\left[c_{1:t}\left(\left(\mathrm{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#}\eta^\pi(X_{t+1}, A^\pi_{t+1}) - \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t, A_t)\right)\right] = \mathbf{0}.$$

To see this, we follow the derivation below,

$$
\begin{aligned}
&\mathbb{E}_\mu\left[c_{1:t}\left(\left(\mathsf{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#}\eta^\pi(X_{t+1},A^\pi_{t+1})-\left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t,A_t)\right)\right]\\
&=_{(a)}\mathbb{E}\left[\mathbb{E}\left[c_{1:t}\left[\left(\left(\mathsf{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#}\eta^\pi(X_{t+1},A^\pi_{t+1})\right]-\left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t,A_t)\right)\,\middle|\,(X_s,A_s,R_{s-1})_{s=1}^t\right]\right]\\
&=_{(b)}\mathbb{E}\left[c_{1:t}\mathbb{E}\left[\left(\mathsf{b}_{G_{0:t},\gamma^{t+1}}\right)_{\#}\eta^\pi(X_{t+1},A^\pi_{t+1})\,\middle|\,(X_s,A_s,R_{s-1})_{s=1}^t\right]-c_{1:t}\left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t,A_t)\right]\\
&=_{(c)}\mathbb{E}\left[c_{1:t}\underbrace{\mathbb{E}\left[\left(\mathsf{b}_{G_{0:t-1}+\gamma^tR_t,\gamma^{t+1}}\right)_{\#}\eta^\pi(X_{t+1},A^\pi_{t+1})\,\middle|\,(X_s,A_s,R_{s-1})_{s=1}^t\right]}_{\text{first term}}-c_{1:t}\left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t,A_t)\right].
\end{aligned}
\tag{8}
$$

In the above, in (a) we condition on $(X_s,A_s,R_s)_{s=1}^t$ and the equality follows from the tower property of expectations; in (b), we use the fact that the trace product $c_{1:t}$ and $\left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t,A_t)$ are deterministic function of the conditioning variable $(X_s,A_s,R_s)_{s=1}^t$; in (c), we split the summation $G_{0:t}=G_{0:t-1}+\gamma R_t$. Now we examine the first term in Equation (8), by applying Lemma G.2, we have

$$\text{first term}=\left(\mathsf{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta^\pi(X_t,A_t).$$

This implies Equation (8) evaluates to a zero measure. Hence η^π is a fixed point of the operator $\mathcal{R}^{\pi,\mu}$. Because $\mathcal{R}^{\pi,\mu}$ is also contractive by Proposition 3.2, the fixed point is unique. □

Theorem 5.1. (**Convergence of quantile distributions**) The projected distributional Retrace operator $\Pi_\mathcal{Q}\mathcal{R}^{\pi,\mu}$ is β-contractive under $\overline{W}_\infty$ distance in $\mathscr{P}_\mathcal{Q}(\mathbb{R})$. As a result, the above η_k converges to a limiting distribution $\eta^\pi_\mathcal{R}$ in $\overline{W}_\infty$, such that $\overline{W}_\infty(\eta_k,\eta^\pi_\mathcal{R})\le(\beta)^k\overline{W}_\infty(\eta_0,\eta^\pi_\mathcal{R})$. Further, the quality of the fixed point is characterized as $\overline{W}_\infty(\eta^\pi_\mathcal{R},\eta^\pi)\le(1-\beta)^{-1}\overline{W}_\infty(\Pi_\mathcal{Q}\eta^\pi,\eta^\pi)$.

Proof. The quantile projection $\Pi_\mathcal{Q}$ is a non-expansion under $\overline{W}_\infty$ [16]. Since $\mathcal{R}^{\pi,\mu}$ is β-contractive under $\overline{W}_p$ for all $p\ge1$, the composed operator $\Pi_\mathcal{Q}\mathcal{R}^{\pi,\mu}$ is β-contractive under $\overline{W}_\infty$. Now, because (1) $\Pi_\mathcal{Q}\mathcal{R}^{\pi,\mu}\in\mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$; (2) the space $\Pi_\mathcal{Q}\mathcal{R}^{\pi,\mu}\in\mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is closed [10]; (3) the operator is contractive, the iterate $\eta_k=\left(\Pi_\mathcal{Q}\mathcal{R}^{\pi,\mu}\right)^k\eta_0$ converges to a limiting distribution $\eta^\pi_\mathcal{R}\in\mathscr{P}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. Finally, by Proposition 5.28 in [10], we have $\overline{W}_\infty(\eta^\pi_\mathcal{R},\eta^\pi)\le(1-\beta)^{-1}\overline{W}_\infty(\Pi_\mathcal{Q}\eta^\pi,\eta^\pi)$. □

Lemma 5.2. (**Unbiased stochastic QR loss gradient estimate**) Assume that the trajectory terminates within $H<\infty$ steps almost surely, then we have $\mathbb{E}_\mu[\widehat{L}^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)]=L^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$ and $\mathbb{E}_\mu[\nabla_{z_i(x,a)}\widehat{L}^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)]=\nabla_{z_i(x,a)}L^{\tau_i}_{z_i(x,a)}\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$.

Proof. The QR loss $L^\tau_\theta(\eta)$ is defined for any distribution η and scalar parameter θ. Let $\nu=\sum_{i=1}^m w_i\eta_i$ be the linear combination of distributions $(\eta_i)_{i=1}^m$ where w_is are potentially negative coefficients. In this case, ν is a signed measure. We define the generalized QR loss for ν as the linear combination of QR losses against η_i weighted by w_i,

$$L^\tau_\theta(\nu):=\sum_{i=1}^m w_iL^\tau_\theta(\eta_i).$$

Next, we note that the QR loss is linear in the input distribution (or signed measure). This means given any (potentially infinite) set of N distributions or signed measures ν_i with coefficients a_i,

$$L^\tau_\theta\left(\sum_{i=1}^N a_i\nu_i\right)=\sum_{i=1}^N a_iL^\tau_\theta(\nu_i).$$

When $(a_i)_{i=1}^N$ denotes a distribution, the above is equivalently expressed as an exchange between expectation and the QR loss $L^\tau_\theta(\mathbb{E}[\nu_i])=\mathbb{E}[L^\tau_\theta(\nu_i)]$. For notational convenience, we let $\theta=z_i(x,a)$ and $\tau=\tau_i$. Because the trajectory terminates within H steps almost surely, since $c_{1:t}\le\rho_{1:t}\le\rho^H$

where $\rho \coloneqq \max_{x\in\mathcal{X},\mathcal{A}} \frac{\pi(a|x)}{\mu(a|x)}$, the estimate $\widehat{L}_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$ is finite almost surely. Combining all results from above we obtain the following

$$\begin{aligned}\mathbb{E}_\mu\left[\mathcal{R}^{\pi,\mu}\eta(x,a)\right] &= \mathbb{E}_\mu\left[L_\theta^\tau\left(\eta(x,a)\right) + \sum_{t=0}^{\infty} c_{1:t}\left(L_\theta^\tau\left(\left(\mathrm{b}_{t+1}\right)_{\#}\eta\left(X_{t+1},A_{t+1}^\pi\right)\right) - L_\theta^\tau\left(\left(\mathrm{b}_t\right)_{\#}\eta(X_t,A_t)\right)\right)\right] \\ &=_{(a)} \mathbb{E}_\mu\left[L_\theta^\tau\left(\widehat{\mathcal{R}}^{\pi,\mu}\eta(x,a)\right)\right] =_{(b)} L_\theta^\tau\left(\mathbb{E}_\mu\left[\widehat{\mathcal{R}}^{\pi,\mu}\eta(x,a)\right]\right) = L_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right).\end{aligned}$$

In the above, (a) follows from the definition of the generalized QR loss against signed measure the definition of $\mathcal{R}^{\pi,\mu}\eta(x,a)$; (c) follows from the linearity of the QR loss.

Next, to show that the gradient estimate is unbiased too, the high level idea is to apply dominated convergence theorem (DCT) to justify the exhchange of gradient and expectation [34]. Since the expected QR loss gradient $\nabla_\theta L_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$ exists, we deduce that the estimate $\nabla_\theta\widehat{L}_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$ exists almost surely. Consider the absolute value of the gradient estimate $\left|\nabla_\theta\widehat{L}_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right|$, which serves as an upper bound to the gradient estimate. In order to apply DCT, we need to show the expectation of the absolute gradient is finite. Note we have

$$\begin{aligned}&\mathbb{E}_\mu\left[\left|\nabla_\theta\widehat{L}_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right|\right] \\ &= \mathbb{E}_\mu\left[\left|\nabla_\theta L_\theta^\tau\left(\eta(x,a)\right) + \sum_{t=0}^{H} c_{1:t}\left(\nabla_\theta L_\theta^\tau\left(\left(\mathrm{b}_{t+1}\right)_{\#}\eta\left(X_{t+1},A_{t+1}^\pi\right)\right) - \nabla_\theta L_\theta^\tau\left(\left(\mathrm{b}_t\right)_{\#}\eta(X_t,A_t)\right)\right)\right|\right] \\ &\le_{(a)} \mathbb{E}_\mu\left[\left|\nabla_\theta L_\theta^\tau\left(\eta(x,a)\right)\right| + \sum_{t=0}^{H} c_{1:t}\left|\nabla_\theta L_\theta^\tau\left(\left(\mathrm{b}_{t+1}\right)_{\#}\eta\left(X_{t+1},A_{t+1}^\pi\right)\right) - \nabla_\theta L_\theta^\tau\left(\left(\mathrm{b}_t\right)_{\#}\eta(X_t,A_t)\right)\right|\right] \\ &\le_{(b)} \mathbb{E}_\mu\left[1 + \sum_{t=0}^{H}\rho^t\cdot 2\right] < \infty,\end{aligned}$$

where (a) follows from the application of triangle inequality; (b) follows from the fact that the QR loss gradient against a fixed distribution is bounded $\nabla_\theta L_\theta^\tau\left(\nu\right) \in [-1,1], \forall\nu\in\mathscr{P}_\infty(\mathbb{R})$ [16].

With the application of DCT, we can exchange the gradient and expectation operator, which yields $\mathbb{E}_\mu\left[\nabla_\theta\widehat{L}_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right] = \nabla_\theta\mathbb{E}_\mu\left[\widehat{L}_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)\right] = \nabla_\theta L_\theta^\tau\left(\mathcal{R}^{\pi,\mu}\eta(x,a)\right)$. □