Off-policy Distributional Q(λ): Distributional RL Without Importance Sampling

Yunhao Tang* **Mark Rowland** Google DeepMind **Rémi Munos*** **Bernardo Avila Pires** Google DeepMind **Will Dabney** Google DeepMind

Abstract

We introduce off-policy distributional Q(λ), a new addition to the family of off-policy distributional evaluation algorithms. Off-policy distributional Q(λ) does not apply importance sampling for off-policy learning, which introduces intriguing interactions with signed measures. Such unique properties distributional Q(λ) from other existing alternatives such as distributional Retrace. We characterize the algorithmic properties of distributional Q(λ) and validate theoretical insights with tabular experiments. We show how distributional Q(λ)-C51, a combination of Q(λ) with the C51 agent, exhibits promising results on deep RL benchmarks.

1 INTRODUCTION

Random returns $\sum_{t=0}^{\infty} \gamma^t R_t$ are of fundamental importance to reinforcement learning (RL). While value-based RL focuses on learning the expectation of random returns $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]$ (Sutton and Barto, 1998), distributional RL has demonstrated benefits to approximate the full distribution of the random return (Bellemare et al., 2017a).

There is a continuous spectrum of algorithms for learning return distributions of a target policy. At one end of the spectrum is Monte-Carlo simulation, where one generates the full path of returns $(R_t)_{t=0}^{\infty}$ along a trajectory, building a direct estimate of the random return (Bellemare et al., 2023). At another end of the spectrum lies one-step temporal difference (TD) algorithms, where one approximate the random return by bootstrapping from the next state distribution (Bellemare et al., 2017a). The bias and variance trade-off between these two extreme cases are akin to their counterparts in the case of value-based RL (Sutton, 1988).

By balancing the bias-variance trade-off, the best performing algorithm is usually found by interpolating the two extremes. Combining full Monte-Carlo simulation and one-step distributional TD learning, one can obtain distributional Retrace, a family of multi-step distributional learning algorithms (Tang et al., 2022). For on-policy case where the data generation policy is the same as the target policy, distributional Retrace recovers a distributional equivalent of TD(λ) (Sutton, 1988; Nam et al., 2021). For the off-policy case, distributional Retrace can also approximate the target distribution efficiently, by adjusting for the discrepancy between the data collection policy and target policy using importance sampling. Prior work has established the efficiency of such multi-step distributional RL algorithms in large-scale practices, which have enabled significant improvements sin agent performance (Gruslys et al., 2018; Tang et al., 2022).

Importance sampling is fundamental to off-policy distributional RL, and to off-policy RL in general (Precup et al., 2001; Munos et al., 2016; Espeholt et al., 2018). By applying a careful reweighting with the probability ratios, it allows for learning from an off-policy trajectory *as if* it were generated on-policy. Importance sampling has a number of critical limitations: it often introduces high variance; it is not applicable when the probability of the data collection policy is not available, which is the case for many applications.

In this work, we introduce off-policy distributional Q(λ), a multi-step distributional RL algorithm *without* the need for importance sampling. Off-policy distributional Q(λ) draws inspirations from value-based off-policy Q(λ) (Harutyunyan et al., 2016) and adopts a single trace coefficient $\lambda \in [0, 1]$ to mediate various properties of the algorithm, such as the bias and variance trade-off. Without importance sampling, dis-

*Work done while at Google DeepMind.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

tributional Q(λ) is fundamentally different from distributional Retrace. Below, we highlight a few intriguing and important properties of distributional Q(λ).

Contraction and fixed point. By design, off-policy distributional Q(λ) operator has the target return distribution as a fixed point. The operator is also contractive when the target and the data collection policy is close enough, i.e., when the data is not too off-policy; we provide precise characterizations in Section 3. As a result when these conditions are met, dynamic programming based or sample based distributional Q(λ) will converge to the target distribution. When on-policy, distributional Q(λ) reduces to the distributional equivalent of value-based TD(λ) or Q(λ).

Signed measures and representations. A distinguishing feature of off-policy distributional Q(λ) is that applications of the operator produces signed measures. This contrasts with prior operators such as distributional Retrace (Tang et al., 2022) or one-step Bellman operator (Bellemare et al., 2017a), where the iterates are naturally confined to be distributions. Intriguingly, this also implies that a convergent algorithm would require representing signed measures. Intuitively, this feels like an unnecessary burden since the fixed point itself is just a distribution. We note that this is the result of unique interaction between distributional learning and off-policy learning without importance sampling. See Figure 1 for an illustration of how the signed measure iterates evolve under the operator: it starts as a distribution, evolves into signed measures during intermediary iterations, and finally returns to a distribution.

To derive a practical algorithm based on the operator, we introduce an extension of the categorical distributional RL algorithm (Rowland et al., 2018) to approximations the signed measure iterates produced by distributional Q(λ) (Section 4).

Trust region interpretation and deep RL. The contraction property of off-policy Q(λ) naturally introduces a form of trust region constraint between target and data collection policy, from which we derive a heuristic to adapt the target policy for optimal control. This new implementation, in combination with C51 (Bellemare et al., 2017a), improves over both baseline off-policy Q(λ) and distributional Retrace over Atari-57 benchmarks.

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. We also assume the rewards are bounded $R_t \in [R_{\min}, R_{\max}]$.

Let $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ be a fixed policy and 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. Define the one-step 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 $Q^\pi = T^\pi Q^\pi$ as the unique fixed point.

2.1 Multi-step value-based learning and off-policy Q(λ)

In value-based learning, the fixed point Q^π can be obtained by repeatedly applying the Bellman operator T^π: $Q_{k+1} = T^\pi Q_k$ such that the sequence of iterate Q_k converges to Q^π at a rate of γ^k. To accelerate the convergence, Q(λ) proposes the geometrically weighted average (Sutton and Barto, 1998) across multi-step bootstrapped targets. Define $\delta_t^\pi \coloneqq R_t + \gamma Q\left(X_{t+1}, A_{t+1}^\pi\right) - Q(X_t, A_t)$ as the value-based TD error, the Q(λ) back-up target is

$$T_\lambda^\pi Q(x, a) = Q(x, a) + \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \lambda^t \gamma^t \delta_t^\pi \right].$$

The Q(λ) operator T_λ^π is $\frac{\gamma(1-\lambda)}{1-\lambda\gamma}$-contractive and improves upon the one-step Bellman operator. The Q(λ) operator T_λ^π is on-policy, as the above expectation is taken under the target policy π. In off-policy learning, the data is generated under a behavior policy μ (i.e., the data collection policy), which generally differs from the target policy π. As a standard assumption, we require $\text{supp}(\pi(\cdot|x)) \subseteq \text{supp}(\mu(\cdot|x)), \forall x \in \mathcal{X}$. Off-policy Q($\lambda$) is a straightforward extension of Q(λ) to the off-policy case (Harutyunyan et al., 2016), whose back-up target is now

$$A_\lambda^{\pi,\mu} Q(x, a) = Q(x, a) + \mathbb{E}_\mu \left[\sum_{t=0}^{\infty} \lambda^t \gamma^t \delta_t^\pi \right].$$

By construction, the operator $A_\lambda^{\pi,\mu}$ has Q^π as a fixed point. Unlike importance sampling (IS) based methods such as Retrace (Precup et al., 2001; Munos et al., 2016), off-policy Q(λ) does not apply IS for off-policy corrections. As a result, $A_\lambda^{\pi,\mu}$ is generally only a contraction

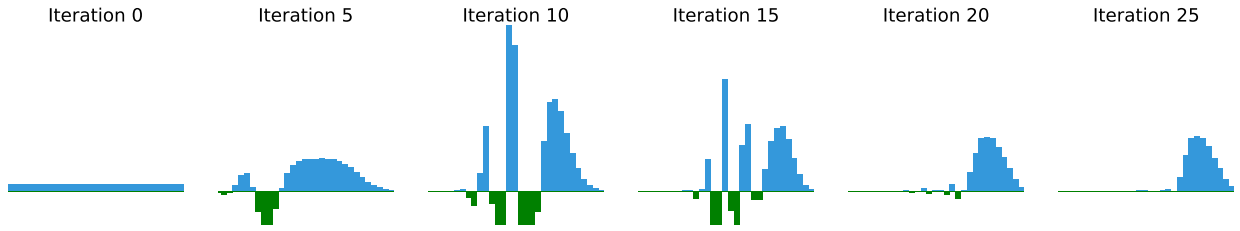

Figure 1: An illustration of the signed measure properties specific to the off-policy Q(λ) operator. The blue and green bars represent the positive and negative probability masses of unit mass signed measures. We visualize the iterate $\eta_{k+1} = \mathcal{A}_\lambda^{\pi,\mu}\eta_k$ for a fixed state-action pair over time on a tabular MDP. The iterate starts as a distribution (an element in $\mathcal{P}(\mathbb{R})$, transitions into a signed measure with unit mass (an element in $\mathcal{M}_1(\mathbb{R})$, and eventually converge to the target return distribution η^π, which is itself a distribution. Any prior distributional RL policy evaluation operators will not exhibit such intriguing behavior, as their iterates are always distributions.

Table 1: A comparison between different distributional operators for policy evaluation with target policy π. The distributional on-policy Q(λ) is an extension of the value-based on-policy Q(λ) operator to the distributional case; its details can be found in Appendix D. All operators preserve the space of probability distribution vector $\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ except off-policy distributional Q(λ). Off-policy distributional Q(λ) is contractive when π and μ are close enough (Lemma 3) and has η^π as the unique fixed point.

Distributional operators	Closed for which space	Contraction rate under $\bar{\ell}_p$	Fixed point
Dist. Bellman operator $\mathcal{T}^\pi$	$\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$	$\gamma^{1/p}$	η^π
Dist. on-policy Q(λ) $\mathcal{T}_\lambda^\pi$	$\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$	$\left(\frac{\gamma(1-\lambda)}{1-\lambda\gamma}\right)^{1/p}$	η^π
Dist. Retrace $\mathcal{R}^{\pi,\mu}$	$\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$	$[0, \gamma^{1/p}]$	η^π
Dist. off-policy Q(λ) $\mathcal{A}_\lambda^{\pi,\mu}$	$\mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$	β_p in Lemma 3	η^π if contractive

under certain conditions on π, μ, and λ (Harutyunyan et al., 2016)

2.2 Distributional reinforcement learning

In general, the return $G^\pi(x,a)$ is a random variable and we define its distribution as $\eta^\pi(x,a) := \text{Law}_\pi\left(G^\pi(x,a)\right)$. The return distribution satisfies the distributional Bellman equation (Morimura et al., 2010a,b; Bellemare et al., 2017a; Rowland et al., 2018; Bellemare et al., 2023),

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

where $\left(\mathrm{b}_{r,\gamma}\right)_\# : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ is the pushforward operation defined through the function $\mathrm{b}_{r,\gamma}(z) = r + \gamma z$ (Rowland et al., 2018). For convenience, we adopt the notation $\eta^\pi(X_t, A_t^\pi) := \sum_a \pi(a|X_t)\eta^\pi(X_t,a)$. Throughout the paper, we focus on the space of distributions with *bounded* support.

Let $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ be any return distribution function, 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 (Rowland et al., 2018; Bellemare et al., 2023),

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

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 $\gamma^{1/p}$-contractive under the ℓ_p distance (Rowland et al., 2018; Bellemare et al., 2023) for any $p \geq 1$, so that η^π is its unique fixed point.

As a remark for technically minded readers, we note that since ℓ_p is initially defined between CDFs of the distributions, it is naturally extended between signed measures too. Meanwhile, it is more challenging to extend Wasserstein distance, another commonly used metric in distributional RL (Bellemare et al., 2023), to signed measures. Hence all the results in this work are stated in terms of the ℓ_p distance.

2.3 Multi-step distributional RL

The multi-step distributional bootstrapping bears qualitative differences from value-based multi-step learning (Gruslys et al., 2018; Tang et al., 2022). In off-policy learning, let $\rho_t := \pi(A_t|X_t)/\mu(A_t|X_t)$ be the step-wise importance sampling (IS) ratio at time step t. 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. Tang et al. (2022) shows that distributional Retrace operator $\mathcal{R}^{\pi,\mu} : \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is

$$\mathcal{R}^{\pi,\mu}\eta(x,a) \coloneqq \eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{\infty} c_{1:t} \cdot \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#} \Delta_t^\pi\right], \tag{3}$$

where $\Delta_t^\pi = \mathcal{T}^\pi\eta(X_t, A_t) - \eta(X_t, A_t)$ is the distributional one-step TD error. Distributional Retrace has η^π as the unique fixed point and in general contracts faster than the one-step distributional Bellman operator $\mathcal{T}^\pi$.

3 OFF-POLICY DISTRIBUTIONAL Q(λ)

We now discuss a number of essential properties of off-policy distributional Q(λ) operator: its origin of derivation, its fixed point and contraction property, and its unique interaction with signed measures.

There are a few equivalent ways to arrive at the operator: to better draw connections to existing multi-step off-policy operators, we start with the mathematical form of the distributional Retrace operator in Eqn (3), and define distributional Q(λ) operator with the trace coefficient $c_t = \lambda \in [0, 1]$:

$$\mathcal{A}_\lambda^{\pi,\mu}\eta(x,a) \coloneqq \eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{\infty} \lambda^t \cdot \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#} \Delta_t^\pi\right]. \tag{4}$$

The off-policy distributional Q(λ) is *not* a special case of distributional Retrace, despite their apparent similarities. Critically, distributional Retrace requires the trace coefficient $c_t \in [0, \rho_t]$ to ensure conservative trace cutting (Munos et al., 2016), whereas setting $c_t = \lambda$ can violate such a condition. A detailed derivation of off-policy distributional Q(λ) can extend from the distributional on-policy Q(λ) operator (Nam et al., 2021), similar to how one arrives at value-based off-policy Q(λ) from on-policy Q(λ) in Section 2. We detail such a derivation in Appendix D.

Before we will elaborate on the connection between distributional Q(λ) with signed measure, a fundamental difference from previous operators.

3.1 Off-policy distributional targets are signed measures

To facilitate the discussion, we introduce the notation $\mathscr{M}_1(\mathbb{R})$ of the space of signed measures with total mass of 1. This particular space of signed measure is a natural generalization of (and a superset to) the space of probability measures by allowing for negative probability mass in certain locations of the distribution, while still requiring a unit total mass. See Bellemare et al. (2023) for a more formal definition of the signed measure space.

While previous distributional operators such as the one-step operator $\mathcal{T}^\pi$ and the Retrace operator $\mathcal{R}^{\pi,\mu}$ map within the space of distributions, this is not the case for the off-policy distributional Q(λ) operator. Concretely, it is possible to find a vector of distribution $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ such that $\mathcal{A}_\lambda^{\pi,\mu}\eta(x,a)$ is not a distribution (but a signed measure in general). In other words, the space of probability distribution is not closed under the operator $\mathcal{A}_\lambda^{\pi,\mu}$. Fortunately, the space of signed measure is closed.

Lemma 1. ***(Closeness of the space of signed measures)*** *Given any $\eta \in \mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we have $\mathcal{A}_\lambda^{\pi,\mu}\eta \in \mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$.*

Figure 1 also illustrate the closeness property of the operator: all iterates are unit mass signed measures.

3.2 Fixed point and contraction property

We now discuss the fixed point and contraction property of the distributional Q(λ) operator. The technical approach is mostly motivated by the value-based case (Harutyunyan et al., 2016). First note that by construction, the off-policy distributional Q(λ) operator has η^π as one fixed point.

Lemma 2. ***(Fixed point)*** *η^π is a fixed point of $\mathcal{A}_\lambda^{\pi,\mu}$.*

Hence, we can evaluate the target return distribution η^π by an algorithm that converges to the fixed point of the operator. A sufficient condition for the approximate dynamic programming algorithms to converge is that the operator be contractive. We consider contraction under the the ℓ_p supremum distance, defined as

$$\bar{\ell}_p(\eta_1, \eta_2) = \max_{x,a} \ell_p\left(\eta_1(x,a), \eta_2(x,a)\right),$$

for any signed measure vectors $\eta_1, \eta_2 \in \mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. The contraction rate critically depends on the distance between π and μ, which we define as $\|\pi - \mu\|_1 = \max_x \sum_a |\pi(a|x) - \mu(a|x)|$.

Lemma 3. ***(Contraction)*** *Let $\epsilon \coloneqq \|\pi - \mu\|_1$, then for any $p \geq 1$ and signed measures $\forall \eta_1, \eta_2 \in \mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$,*

$$\bar{\ell}_p\left(\mathcal{A}_\lambda^{\pi,\mu}\eta_1, \mathcal{A}_\lambda^{\pi,\mu}\eta_2\right) \leq \beta_p \bar{\ell}_p\left(\eta_1, \eta_2\right),$$

where $\beta_p = \gamma^{1/p}\frac{1-\lambda+\lambda\epsilon}{(1-\lambda)^{(p-1)/p}(1-\lambda\gamma)^{1/p}}$ is the contraction rate under the supremum ℓ_p distance.

The contraction rate β_p depends on $\|\pi - \mu\|_1$ which is a measure of off-policyness; and the value of γ and λ. When π, μ are close enough, the off-policy distributional Q(λ) operator is contractive.

Corollary 1. *When* $\|\pi - \mu\|_1 < \frac{1-\gamma}{\lambda\gamma}$*, we have* $\beta_1 < 1$ *and the operator* $\mathcal{A}_\lambda^{\pi,\mu}$ *is contractive under the* L_1 *distance. This also implies that* η^π *is the unique fixed point to* $\mathcal{A}_\lambda^{\pi,\mu}$*.*

A few remarks are in order. We call $\frac{1-\gamma}{\lambda\gamma}$ the contraction radius, the bound on the distance between π and μ to ensure that the operator is contractive, which also coincides with similar quantities in the value-based off-policy Q(λ) (Harutyunyan et al., 2016). Inverting the radius condition, we derive a bound on the trace coefficient λ for the operator to be contractive: $\lambda < \frac{1-\gamma}{\gamma\|\pi-\mu\|_1}$. In other words, when π and μ are far from each other, λ can only take small value close to 0 in order to ensure that the operator is contractive. In such cases, operator cuts traces very quickly, and can only make use of bootstrapped values in the very near future. Meanwhile, in the limiting on-policy case when $\pi = \mu$, the radius condition is always satisfied and any value of $\lambda \in [0, 1]$ makes the operator contractive. The trade-off is that larger value of λ will lead to faster contrction in expectation, but induces higher variance.

The operator $\mathcal{A}_\lambda^{\pi,\mu}$ does not preserve the space of probability distributions, yet it still has η^π as the unique fixed point when π and μ are close enough. Consider initializing a distribution vector $\eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ and generate a sequence of iterate based on $\eta_{k+1} = \mathcal{A}_\lambda^{\pi,\mu}\eta_k$. Lemma 1 suggests that in general, the intermediate iterates $(\eta_k)_{k\geq 1}$ are signed measures. However, as $k \to \infty$, we can expect the signed measure sequence to converge back to a probability distribution η^π (see Figure 1 for the illustration). This example bears important implications on the parametric representations of intermediate iterates in algorithm designs.

Alternative way to construct distributional Q(λ). For interested readers, we also note that there are alternative ways to construct the distributional Q(λ) operator. We provide such a natural alternative in Appendix D, which is closely related to the *path-dependent* distributional TD errors discussed in Tang et al. (2022). Our construction of $\mathcal{A}_\lambda^{\pi,\mu}$ leads to better theoretical properties compared to the alternative.

4 LEARNING WITH CATEGORICAL REPRESENTATION

Return distributions are in general infinite dimensional objects. In practice, it is necessary to approximate the target return distribution with parametric approximations. One commonly used family of parameterization is the categorical representation (Bellemare et al., 2017a, 2023). We provide a brief background below.

In categorical representation, we consider parametric distributions, for a fixed $m \geq 1$, of the form: $\sum_{i=1}^m p_i\delta_{z_i}$, 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 $p_i \geq 0$. For simplicity, we assume z_i to be strictly monotonic $z_i < z_{i+1}$ and the range of atoms covers all possible returns from the MDP $[(1-\gamma)^{-1}R_{\min}, (1-\gamma)^{-1}R_{\max}] \subset [z_1, z_m]$. Let $\mathscr{P}_c(\mathbb{R})$ denote the class of distributions

$$\mathscr{P}_c(\mathbb{R}) \coloneqq \left\{ \sum_{i=1}^m p_i\delta_{z_i} | \sum_{i=1}^m p_i = 1, p_i \geq 0 \right\}.$$

When combining parametric representation with the off-policy distributional Q(λ), it is important to account for the fact that signed measures can arise by applying the operator $\mathcal{A}_\lambda^{\pi,\mu}$. We can extend the categorical representation by dropping the non-negativity constraints on p_i

$$\mathscr{M}_{1,c}(\mathbb{R}) \coloneqq \left\{ \sum_{i=1}^m p_i\delta_{z_i} | \sum_{i=1}^m p_i = 1 \right\}.$$

Given a target signed measure $\eta \in \mathcal{M}_1(\mathbb{R})$, we define the projection $\Pi_c : \mathscr{M}_\infty(\mathbb{R}) \to \mathscr{M}_{1,c}(\mathbb{R})$ onto the space of categorical distributions by minimizing the ℓ_2 distance $\Pi_c\eta \coloneqq \arg\min_{\nu\in\mathscr{M}_{1,c}(\mathbb{R})} \ell_2(\nu, \eta)$. Note a major difference here is that the projection operation also produces a signed measure, which can also be computed in a natural and efficient way (Rowland et al., 2018; Bellemare et al., 2017b). Visually, we can understand the categorical projection $\Pi_c\eta$ as a discretized approximation to signed measure η, see Figure 4 for an illustration. The approximation becomes more accurate as the number of atoms increases. We refer readers to Chapter 9 of Bellemare et al. (2023) for further details.

4.1 Fixed point and contraction property

To implement the off-policy distributional Q(λ) in practice, we represent the distribution iterate as a vector of categorical signed measures $\eta_k \in \mathscr{M}_{1,c}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. After applying the operator $\mathcal{A}_\lambda^{\pi,\mu}\eta_k$, we project the back-up target to the space of categorical signed measures. This yields the following recursion

$$\eta_{k+1} = \Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_k. \tag{5}$$

Understanding the behavior of the above recursion consists in characterizing the composed operator $\Pi_c\mathcal{A}_\lambda^{\pi,\mu}$. We have the following characterization

Lemma 4. ***(Contraction of composed operator)*** *The composed operator* $\Pi_c\mathcal{A}_\lambda^{\pi,\mu}$ *is* β_2*-contractive under the* $\bar{L}_2$ *distance in the space of signed measure vectors, i.e.,* $\forall\eta_1, \eta_2 \in \mathscr{M}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$*,*

$$\bar{L}_2\left(\Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_1, \Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_2\right) \leq \beta_2\bar{L}_2\left(\eta_1, \eta_2\right),$$

where β_2 is defined in Lemma 3. The operator is guaranteed to be contractive when π, μ is within the contraction radius defined below

$$\|\pi - \mu\|_1 < \lambda^{-1}\left(\sqrt{(1-\lambda)(\gamma^{-1}-\lambda)} + \lambda - 1\right)$$

Under the categorical representation, there is an inherent limit on how well one can approximate the true return distribution η^π. The best possible approximation is $\Pi_c\eta^\pi$, the direct projection of the target return to the representation space. The irreducible approximation error is $\bar{L}_2(\eta^\pi, \Pi_c\eta^\pi)$. When using bootstrapping, the approximation error can compound over time. When $\Pi_c\mathcal{A}_\lambda^{\pi,\mu}$ is contractive, the recursion in Eqn (5) converges and let $\eta_{\mathcal{A}}^\pi$ be the fixed point of the composed operator. We can characterize its approximation error to the target return, by extending Rowland et al. (2018, Proposition 3) below.

Lemma 5. ***(Approximation error)*** *When $\beta_2 < 1$, we have*

$$\bar{L}_2(\eta^\pi, \eta_{\mathcal{A}}^\pi) \leq \frac{\bar{L}_2(\eta^\pi, \Pi_c\eta^\pi)}{\sqrt{1-\beta_2^2}}.$$

We see that the parameter β_2 appear both as the contraction rate as well as a factor in the approximation error. An operator with a fast contraction rate will also have smaller approximation error, which is a recurrent observation made in prior work (Bellemare et al., 2023). We will validate this theoretical insight later.

5 DEEP RL IMPLEMENTATION: Q(λ)-C51

We now describe the deep RL implementation of Q(λ)-C51, an extension of the C51 agent (Bellemare et al., 2017a) to the Q(λ) operator. By design, the agent parameterizes the categorical return distribution $(p_i(x,a;\theta))_{i=1}^m$ with neural network parameter θ. Let $\eta_\theta(x,a) = \sum_{i=1}^m p_i(x,a;\theta)\delta_{z_i}$ denote the parameterized categorical distribution. In the original C51 agent, given a target policy π and behavior policy μ, the aim is to minimize the KL-divergence between the parameterized categorical distribution and the one-step back-up target $\mathbb{KL}(\Pi_c\mathcal{T}^\pi\eta_\theta(x,a), \eta_\theta(x,a))$, where $\mathbb{KL}(p,q) \coloneqq \sum_i p_i \log q_i/p_i$. To this end, the algorithm carries out gradient descent on the KL-divergence

$$\theta \leftarrow \theta - \kappa \cdot \nabla_\theta \mathbb{KL}(\Pi_c\mathcal{T}^\pi\eta_{\theta^-}(x,a), \eta_\theta(x,a))$$

with learning rate parameter $\kappa > 0$ and θ^- is the target network parameter, usually computed as a slow moving average of θ (Mnih et al., 2013). To adapt C51 for off-policy distributional Q(λ), we carry out the gradient update, *heuristically* written as

$$\theta \leftarrow \theta - \kappa \cdot \nabla_\theta \mathbb{KL}(\Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a), \eta_\theta(x,a)).$$

Here, *heuristically* refers to the fact that the KL-divergence might not be well defined since $\Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a)$ can be a sigend measure. To make the implementation more concrete, note that we can always write $\mathcal{A}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a)$ as a linear combination of proper distributions generated from the future time step

$$\mathcal{A}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a) = \mathbb{E}_\mu\left[\sum_{t=0}^\infty w_t \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_\# \eta_{\theta^-}(X_t,A_t)\right]$$

where the combination coefficient w_t can be expressed as

$$w_t \coloneqq \mathbb{E}_\mu\left[c_1 \ldots c_{t-1}\left(\pi(b|X_t) - c(X_t,b)\mu(b|X_t)\right)\right].$$

Note that w_t can be negative, whereas for the case of distributional Retrace the coefficient yields the same form but with $w_t \geq 0$; this echos the unique interaction that Q(λ) introduces with signed measures. For more detailed on the derivations of w_t, see Tang et al. (2022) and Appendix C.

Next, we can construct unbiased estimate to the above back-up target by sampling trajectories with the behavior policy $(X_t, A_t, R_t)_{t=0}^\infty \sim \mu$ and calculate the signed measure back-up using the linear combination

$$\widehat{\mathcal{A}}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a) = \sum_{t=0}^\infty \widehat{w}_t \underbrace{\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_\# \eta_{\theta^-}(X_t,A_t)}_{\text{proper distribution}},$$

where the scalar weights can be computed along the sampled trajectory $\widehat{w}_t = c_1 \ldots c_{t-1}\left(\pi(b|X_t) - c(X_t,b)\mu(b|X_t)\right)$. It is straightforward to see that these are unbiased estimates to the coefficients w_t. As a result, $\widehat{\mathcal{A}}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a)$ is an unbiased estimate to the signed measure target $\mathcal{A}_\lambda^{\pi,\mu}\eta_{\theta^-}(x,a)$, and is in general a signed measure though each of the unweighted summand $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_\# \eta_{\theta^-}(X_t,A_t)$ is a proper distribution. Finally, we compute the gradient update $\widehat{g}_\theta$ as a weighted average of gradients through individual well-defined KL divergences against the prediction $\eta_\theta(x,a)$,

$$\sum_{t=0}^\infty \widehat{w}_t \nabla_\theta \mathbb{KL}\left(\eta_\theta(x,a), \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_\# \eta_{\theta^-}(X_t,A_t)\right) \quad (6)$$

See Algorithm 1 for the full algorithmic process.

5.1 Adapting target policy for optimal control

A practical objective is to maximize the agent performance over time, i.e., optimal control. In this case, we let the target policy be the greedy policy with respect to Q_{η_k} where Q_{η_k} is the Q-function induced by η_k. Due

to space limit, we state results for optimal control in Appendix E.

One way to interpret the results on policy evaluation is that we can impose constraints on the target policy π_k such that it stays within the contraction radius. Concretely, we can let the target policy be a mixture between the greedy policy and behavior policy μ with $\alpha \in [0, 1]$.

$$\pi_k = \alpha \mathcal{G}\left(Q_{\eta_k}\right) + (1-\alpha)\mu \tag{7}$$

In practice, when the behavior policy is typically defined through a replay buffer, μ slowly varies over time. In that case, α would be introduced as an extra hyper-parameter. By regularizing the target policy π_k towards the behavior policy, we have made the off-policy evaluation problem more on-policy, which effectively speeds up the contraction rate of the distributional Q(λ) operator. A similar implementation has been considered in Rowland et al. (2020), with a a similar mixing strategy to speed up the contraction rate of the value-based Retrace algorithm.

6 RELATED WORK

Distributional Q(λ) bears close connections to Peng's Q(λ) in the value-based case (Peng and Williams, 1994; Kozuno et al., 2021). Due to space limit, we provide an extended discussion in Appendix B.

Many theoretical results on distributional Q(λ) echo value-based Q(λ) (Harutyunyan et al., 2016), such as the contraction radius between target and behavior policy. A key difference between the value-based and distributional setting is the representation; while value-based Q(λ) requires representing a scalar per state-action pair, distributional Q(λ) requires a parameterized signed measure per state-action pair. This latter property also precludes distributional Q(λ) from being a special case of the distributional Retrace (Tang et al., 2022). We also discuss in Appendix D how the alternative constructs of distributional Q(λ) differ from that of distributional Retrace.

7 EXPERIMENTS

We start with tabular experiments which validate a number of theoretical insights, followed by an assessment of off-policy Q(λ) in large-scale experiments.

7.1 Tabular experiments

We consider a tabular MDP setting with $|\mathcal{X}| = 5$ states and $|\mathcal{A}| = 20$ actions with discount $\gamma = 0.9$. Both the transitions are the reward functions are randomly generated and fixed. The data collection policy μ is uniform for all time. We compare distributional one-step, Retrace and off-policy Q(λ) with categorical representations for $m = 10$, and in the optimal control setting. Throughout, the target policy is the greedy policy induced by the iterate η_k with $\eta_{k+1} = \Pi_c \mathcal{R} \eta_k$ where $\mathcal{R}$ is the distributional operator of interest. For Retrace we sweep over the truncation coefficient $\bar{c} \in \{1, 2, 4\}$ and for Q(λ) over $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Each experiment starts with the same initialization and is repeated 20 times to show standard errors across seeds. In Figure 2, we show the distance $L(\eta_k, \eta^*)$ as a function of iteration k for different algorithmic variants.

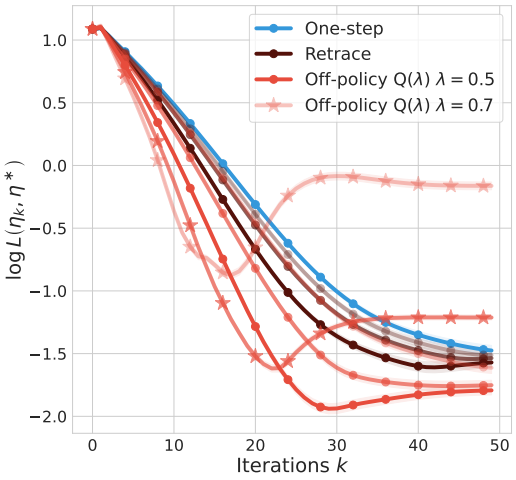

Figure 2: The distance between the algorithmic iterate η_k and return distribution for the optimal policy η^*, as we run control algorithms with distributional one-step, Retrace and off-policy Q(λ). All algorithms use categorical representations and set greedy policy as the target policy. Different curves show an algorithmic variant with a different hyper-parameter setting ($\bar{c}$ for Retrace and λ for Q(λ)). Note that Q(λ) can obtain better performance than Retrace when λ is chosen properly; when λ is too large (≥ 0.7 in this case), the algorithm diverges – despite the initial fast improvement, will not converge to the correct fixed point.

We make a few observations: (1) Both Retrace and off-policy Q(λ) improve over one-step both in terms of asymptotic performance and contraction speed. This corroborates the theoretical insight that the contraction rate and the distance to the final fixed point is related; (2) In this particular case, Q(λ) outperforms Retrace when λ is chosen properly. However, when λ is too large (≥ 0.7 here), the algorithm becomes divergent - despite a fast initial decay in the distance, will converge to a clearly sub-optimal point. Here, a caveat is that since we are testing the case for dynamic programming, we have not considered the variance effect of setting large

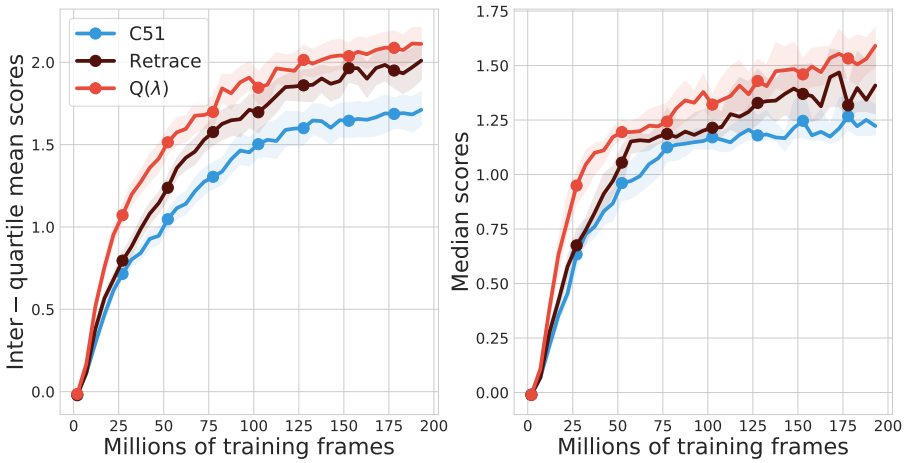

Figure 3: Comparison of C51 (Bellemare et al., 2017a), Retrace-C51 (Tang et al., 2022) and off-policy distributional Q(λ) with target mixing $\alpha = 0.6$ based on Eqn (7) and $\lambda = 0.4$. We show the agents' average performance metrics evaluated throughout training: the inter-quartile mean score (Agarwal et al., 2021), which can be understood as a more robust estimate to the mean score; and the median score, calculated across all 57 games. All scores show the mean and bootstrapped confidence intervals across 5 seeds (Agarwal et al., 2021). Off-policy distributional Q(λ) obtains performance improvements over Retrace-C51 when using target mixing.

values for $\bar{c}$ and λ. Such factors should be accounted for in practice.

Note that with $|\mathcal{A}| = 20$, we have made the algorithm effectively more off-policy. When we decrease $|\mathcal{A}|$, we see that both Retrace and off-policy Q(λ) become better behaved: Q(λ) becomes convergent even with large λ, and Retrace may outperform Q(λ) by benefiting from the full trace with IS corrections. See more results in Appendix F.

7.2 Deep RL experiments

For the deep RL experiments, we use the Atari game suite of 57 games as the testbed and compare distributional RL agents: C51 (Bellemare et al., 2017b), Retrace-C51 (Tang et al., 2022) and Q(λ)-C51. All three algorithms share exactly the same network architecture θ: they parameterize the image inputs via a convnet, followed by MLPs that transform the convnet embeddings into m-logits $l(x, a, i), 1 \leq i \leq m$ per action a. The final prediction is computed as a softmax distribution over the logits $p_i(x, a; \theta) \propto \exp(l(x, a, i))$. The algorithms and only differ in the back-up targets used for updating the prediction $\eta_\theta(x, a)$.

For both distributional Retrace and distributional Q(λ), we calculate the back-up targets with partial trajectories of length $n = 3$ sampled from the replay buffer. This is consistent with practices in prior work (Tang et al., 2022). The C51 baseline can be recovered as a special case with $n = 1$. In general, the acting policy is usually the ϵ-greedy policy with respect to the Q-function induced by the return distribution $Q_{\eta_\theta(x,a)}$. The value of ϵ decays over time, in order to achieve a good balance between exploration and exploitation. The transition tuple (X_t, A_t, R_t) is then put into a replay buffer, which gets sampled when constructing the back-up target. As a result, the effective behavior policy μ is a mixture of ϵ-greedy policy over time. By default, the target policy is greedy with respect to the current Q-function.

Trust region adaptation of target policy. Motivated by the connection between off-policy Q(λ) and trust region updates, we consider a variant of Q(λ) which constructs the target distribution as the mixture between greedy and behavior policy (Eqn (7)). This introduces the mixing coefficient as an extra hyperparameter to the algorithm α, which we find to work best when it is around $0.6 \sim 0.8$. Meanwhile, we find $\alpha = 1$ (i.e., greedy policy) to work generally suboptimally. In Figure 3, we compare such a variant of off-policy distributional Q(λ) against baseline C51 and Retrace-C51. In this case, distributional Q(λ) obtains certain performance improvements over Retrace-C51, which further improves over C51 as shown in (Tang et al., 2022).

8 CONCLUSION

We have proposed off-policy distributional Q(λ), a new addition to the distributional RL arsenal. Without importance sampling distributional Q(λ) has a few intriguing theoretical properties which set it aside from previous approaches. Distributional Q(λ) also enjoys promising empirical performance in various domains.

References

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.

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

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*, 2017b.

Marc G Bellemare, Nicolas Le Roux, Pablo Samuel Castro, and Subhodeep Moitra. Distributional reinforcement learning with linear function approximation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2203–2211. PMLR, 2019.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

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.

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.

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

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.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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.

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*, 2010a.

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*, 2010b.

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.

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.

Jing Peng and Ronald J Williams. Incremental multi-step q-learning. In *Machine Learning Proceedings 1994*, pages 226–232. Elsevier, 1994.

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.

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.

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.

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

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

Yunhao Tang, Remi Munos, Mark Rowland, Bernardo Avila Pires, Will Dabney, and Marc G Bellemare. The nature of temporal difference errors in multi-step distributional reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Mn4IkuWamy.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7 and the conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a foundational reinforcement learning contribution with no direct societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 7 and the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

APPENDICES: Off-policy Q(λ) for distributional reinforcement learning

A Detailed derivations of off-policy distributional Q(λ)

Here, we provide a detailed alternative derivation of off-policy distributional Q(λ) operator. We start with the on-policy n-step distributional operator

$$\mathcal{T}_n^\pi \eta(x,a) = \mathbb{E}_\pi \left[\left(\mathrm{b}_{G_{0:n-1}, \gamma^n} \right)_\# \eta(X_n, A_n^\pi) \right],$$

which reduces to the distributional Bellman operator $\mathcal{T}^\pi$ when $n=1$ (Bellemare et al., 2017a). Note the n-step operator has η^π as the unique fixed point.

The on-policy distributional Q(λ) operator can be constructed as the geometrically weighted mixture of n-step distributional Bellman operator.

$$\begin{aligned}\mathcal{T}_\lambda^\pi \eta(x,a) &\coloneqq (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\mathcal{T}_n^\pi \eta(x,a)\\ &= (1-\lambda)\sum_{n=1}^{\infty}\mathbb{E}_\pi\left[\left(\mathrm{b}_{G_{0:n-1},\gamma^n}\right)_{\#}\eta(X_n,A_n^\pi)\right].\end{aligned}$$

The on-policy Q(λ) operator has η^π as the unique fixed point by design. The on-policy nature of the operator is reflected by the fact that the expectation is taken under target policy π. Now, we rewrite the above operator in the form of distributional TD error,

$$\mathcal{T}_\lambda^\pi \eta(x,a) = \eta(x,a) + \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\lambda^t\cdot\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\Delta_t^\pi\right],$$

where $\Delta_t^\pi = \mathcal{T}^\pi\eta(X_t,A_t) - \eta(X_t,A_t)$ is a signed measure with zero total mass. To derive the off-policy distributional Q(λ) operator, we simply replace the expectation under π by an expectation under behavior policy μ. This yields the operator

$$\mathcal{A}_\lambda^{\pi,\mu}\eta(x,a) = \eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^{\infty}\lambda^t\cdot\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\Delta_t^\pi\right].$$

B Distributional Peng's Q(λ) operator

We provide a more detailed discussion on distributional Peng's Q(λ) operator. Starting from the on-policy n-step distributional operator,

$$\mathcal{T}_n^\pi\eta(x,a) = \mathbb{E}_\pi\left[\left(\mathrm{b}_{G_{0:n-1},\gamma^n}\right)_{\#}\eta(X_n,A_n^\pi)\right],$$

we derive the uncorrected n-step operator, by simply replacing the expectation under π by an expectation under μ,

$$\mathcal{P}_n^{\pi,\mu}\eta(x,a) = \mathbb{E}_\mu\left[\left(\mathrm{b}_{G_{0:n-1},\gamma^n}\right)_{\#}\eta(X_n,A_n^\pi)\right].$$

The uncorrected n-step operator, as its name suggests, does not have η^π as the fixed point in general. This is because the operator does not correct for the off-policyness between π and μ and takes a plain expectation over μ. The distributional Peng's Q(λ) operator, is simply a geometrically weighted mixture of uncorrected n-step operators

$$\begin{aligned}\mathcal{P}_\lambda^{\pi,\mu}\eta(x,a) &\coloneqq (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\mathcal{P}_n^{\pi,\mu}\eta(x,a)\\ &= (1-\lambda)\sum_{n=1}^{\infty}\mathbb{E}_\mu\left[\left(\mathrm{b}_{G_{0:n-1},\gamma^n}\right)_{\#}\eta(X_n,A_n^\pi)\right].\end{aligned}$$

C Proof of theoretical results

Lemma 1. ***(Closeness of the space of signed measures)*** *Given any $\eta\in\mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we have $\mathcal{A}_\lambda^{\pi,\mu}\eta\in\mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$.*

Proof. Our proof follows closely the proof techniques of Lemma 3.1 in Tang et al (Tang et al., 2022). Following their approach, we consider the general notation of trace coefficient c_t which in our case is λ. For all $t\geq 1$, we define the coefficient

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

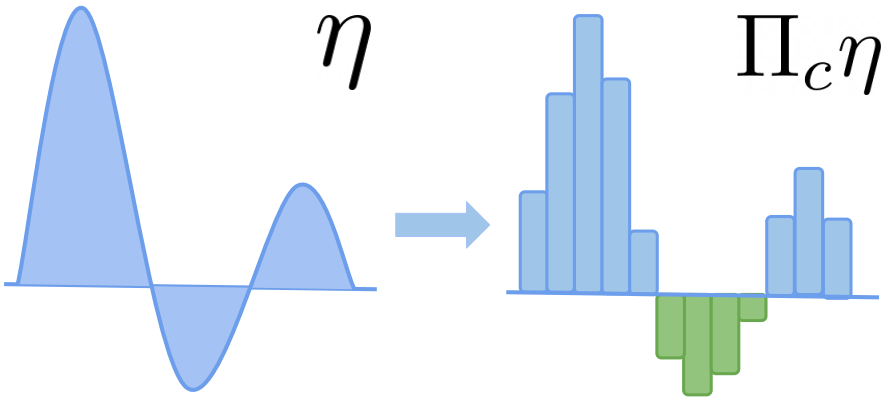

Figure 4: Illustration of categorical projection for the signed measure. On the left, we have a signed measure $\eta \in \mathscr{M}_1(\mathbb{R})$; on the right, we show the categorical projection of the signed measure η onto the space $\mathscr{M}_{1,c}(\mathbb{R})$, with green bars showing the negative mass of the projected measure. The categorical projection is a discretized approximation to the original signed measure, with increasing accuracy as the number of atoms $(z_i)_{i=1}^m$ increases.

Let $\mathscr{R}$ be the set of reward value that random variable R_t can take. Let $\mathcal{R}^t = \mathcal{R} \times \mathcal{R} \times ...\mathcal{R}$ be the Cartesian product of t replicates of $\mathcal{R}$. Through careful algebra, we can rewrite the off-policy Q(λ) operator as follows

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

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 signed measure $\eta(x,a)$, which means $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(y,b) \in \mathscr{M}_\infty(\mathbb{R})$. Now, we examine the sum of all coefficients $\sum w_{y,b,r_{0:t-1}} = \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{b\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} w_{y,b,r_{0:t-1}}$. Tang et al. has showed that for general c_t,

$$\sum w_{y,b,r_{0:t-1}} = 1$$

This implies that $\mathcal{A}_\lambda^{\pi,\mu} \in \mathscr{M}_1(\mathbb{R})$ as it is a linear combination of signed measures with total mass 1. A critical difference here is that since $c_t \notin [0,\rho_t]$ as in the Retrace case, there is no general guarantee that $w_{y,b,r_{0:t-1}} \geq 0$. □

Lemma 6. ***(Fixed point)*** *η^π is a fixed point of $\mathcal{A}_\lambda^{\pi,\mu}$.*

Proof. By construction, $\mathcal{A}_\lambda^{\pi,\mu}$ is a weighted sum of distributional TD error Δ_t^π under policy π. By letting $\eta = \eta^\pi$, we have $\mathbb{E}\left[\Delta_t^\pi \mid X_t, A_t\right] = 0$ (note that here the right hand side is a *zero measure*). This implies $\mathcal{A}_\lambda^{\pi,\mu}\eta^\pi = \eta^\pi$ and verifies η^π as a fixed point of the operator. □

Lemma 3. ***(Contraction)*** *Let $\epsilon := \|\pi - \mu\|_1$, then for any $p \geq 1$ and signed measures $\forall \eta_1, \eta_2 \in \mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$,*

$$\bar{\ell}_p\left(\mathcal{A}_\lambda^{\pi,\mu}\eta_1, \mathcal{A}_\lambda^{\pi,\mu}\eta_2\right) \leq \beta_p \bar{\ell}_p\left(\eta_1,\eta_2\right),$$

where $\beta_p = \gamma^{1/p}\frac{1-\lambda+\lambda\epsilon}{(1-\lambda)^{(p-1)/p}(1-\lambda\gamma)^{1/p}}$ is the contraction rate under the supremum ℓ_p distance.

Proof. From the proof of Lemma 1, we can write

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

For $1 \leq t \leq n$, we have

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

For any $t \geq 1$, we upper bound the absolute value of the weight coefficient $w_{x,a,r_{0:t-1}}$ as follows

$$
\begin{aligned}
&= \left|\lambda^{t-1} \cdot (1-\lambda)\mathbb{E}\left[\pi(a|X_t) \cdot \mathbb{I}[X_t = x]\Pi_{s=0}^{t-1}\mathbb{I}[R_s = r_s]\right] + \lambda^{t-1}\lambda \cdot \mathbb{E}\left[(\pi(a|X_t) - \mu(a|X_t))\,\mathbb{I}[X_t = x]\Pi_{s=0}^{t-1}\mathbb{I}[R_s = r_s]\right]\right| \\
&\leq_{(a)} \lambda^{t-1} \cdot (1-\lambda)\mathbb{E}\left[\pi(a|X_t) \cdot \mathbb{I}[X_t = x]\Pi_{s=0}^{t-1}\mathbb{I}[R_s = r_s]\right] + \lambda^{t-1}\lambda \cdot \mathbb{E}\left[|\pi(a|X_t) - \mu(a|X_t)| \cdot \mathbb{I}[X_t = x]\Pi_{s=0}^{t-1}\mathbb{I}[R_s = r_s]\right] \\
&\leq_{(b)} \lambda^{t-1} \cdot (1-\lambda)\mathbb{E}\left[\pi(a|X_t) \cdot \mathbb{I}[X_t = x]\Pi_{s=0}^{t-1}\mathbb{I}[R_s = r_s]\right] + \lambda^{t-1}\lambda \cdot \mathbb{E}\left[\epsilon \cdot \mathbb{I}[X_t = x]\Pi_{s=0}^{t-1}\mathbb{I}[R_s = r_s]\right] =: |w_{x,a,r_{0:t-1}}|
\end{aligned}
$$

Above, (a) follows from the triangle inequality; (b) follows from the fact that for any random variable Z, $|\mathbb{E}[Z]| \leq \mathbb{E}[|Z|]$; for (b), we also apply the fact that $\|\pi - \mu\|_1 \coloneqq \max_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} |\pi(a|x) - \mu(a|x)| = \epsilon$. Now, we define a signed measure as the negative of the distribution $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(x,a)$

$$\tilde{\eta}_{x,a,r_{0:t-1}} \coloneqq \mathrm{sign}(w_{x,a,r_{0:t-1}}) \cdot \left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_{\#}\eta(x,a),$$

with the signed function $\mathrm{sign}(z) : \mathbb{R} \to \mathbb{R}$. Then we can write

$$\mathcal{A}_\lambda^{\pi,\mu}\eta(x,a) = \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\tilde{\eta}_{x,a,r_{0:t-1}}.$$

Finally, we have

$$
\begin{aligned}
&\ell_p^p\left(\mathcal{A}_\lambda^{\pi,\mu}\eta_1(x,a), \mathcal{A}_\lambda^{\pi,\mu}\eta_2(x,a)\right) \\
&=_{(a)} \ell_p^p\left(\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\tilde{\eta}^{(1)}_{x,a,r_{0:t-1}}, \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\tilde{\eta}^{(2)}_{x,a,r_{0:t-1}}\right) \\
&=_{(b)} \left(\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\right)^{p-1} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\ell_p^p\left(\tilde{\eta}^{(1)}_{x,a,G_{0:t-1}}, \tilde{\eta}^{(2)}_{x,a,G_{0:t-1}}\right) \\
&\leq_{(c)} \left(\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\right)^{p-1} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\ell_p^p\left(\tilde{\eta}_{x,a,r^{(1)}_{0:t-1}}, \tilde{\eta}^{(2)}_{x,a,r_{0:t-1}}\right) \\
&\leq_{(d)} \left(\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\right)^{p-1} \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\gamma^t\bar{\ell}_p^p(\tilde{\eta}_1, \tilde{\eta}_2) \\
&=_{(e)} \left(\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\right)^{p-1} \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\gamma^t\bar{\ell}_p^p(\eta_1, \eta_2)\,.
\end{aligned}
$$

In the above, (a) follows from the definition of $\tilde{\eta}$; (b) follows from the scaling property of the ℓ_p distance; (c) follows from the convex property of the ℓ_p distance, see Bellemare et al. (Bellemare et al., 2023); (d) follows from the definition of the supremum distance $\bar{\ell}_p$; (e) follows from the fact that $\bar{\ell}_p(\eta_1, \eta_2) = \bar{L}(\tilde{\eta}_1, \tilde{\eta}_2)$. Now, we examine the sum over coefficients $|w_{x,a,r_{0:t-1}}|$. For any fixed time step $t \geq 1$,

$$\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}| = \lambda^{t-1}(1-\lambda+\lambda\epsilon).$$

Hence the total sum is

$$\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}| = \frac{1-\lambda+\lambda\epsilon}{1-\lambda}$$

$$\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} |w_{x,a,r_{0:t-1}}|\bar{\ell}_p^p(\eta_1, \eta_2) = \frac{\gamma(1-\lambda+\lambda\epsilon)}{1-\lambda\gamma}\bar{\ell}_p^p(\eta_1, \eta_2).$$

By combing hte above quantities and taking the $1/p$-th root, we obtain the overall result

$$\ell_p\left(\mathcal{A}_\lambda^{\pi,\mu}\eta_1(x,a), \mathcal{A}_\lambda^{\pi,\mu}\eta_2(x,a)\right) \leq \beta_p(x,a)\bar{\ell}_p(\eta_1, \eta_2)$$

with $\beta_p(x,a) = \gamma^{1/p}\frac{1-\lambda+\lambda\epsilon}{(1-\lambda)^{(p-1)/p}(1-\lambda\gamma)^{1/p}}$. We obtain the overall contraction rate by taking $\beta_p = \max_{x,a}\beta_p(x,a)$. □

Corollary 1. *When* $\|\pi - \mu\|_1 < \frac{1-\gamma}{\lambda\gamma}$, *we have* $\beta_1 < 1$ *and the operator* $\mathcal{A}_\lambda^{\pi,\mu}$ *is contractive under the* L_1 *distance. This also implies that* η^π *is the unique fixed point to* $\mathcal{A}_\lambda^{\pi,\mu}$.

Proof. By setting the condition on the contraction rate $\beta_1 < 1$, we obtain $\|\pi - \mu\|_1 < \frac{1-\gamma}{\lambda\gamma}$. Since η^π is a fixed point of $\mathcal{A}_\lambda^{\pi,\mu}$ by Lemma 6, when the operator is contractive $\beta_1 < 1$ under the L_1 distance, we also have η^π as the unique fixed point. □

Lemma 4. ***(Contraction of composed operator)*** *The composed operator* $\Pi_c\mathcal{A}_\lambda^{\pi,\mu}$ *is* β_2*-contractive under the* $\bar{L}_2$ *distance in the space of signed measure vectors, i.e.,* $\forall \eta_1, \eta_2 \in \mathscr{M}_\infty(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$,

$$\bar{L}_2\left(\Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_1, \Pi_c\mathcal{A}_\lambda^{\pi,\mu}\eta_2\right) \le \beta_2\bar{L}_2\left(\eta_1, \eta_2\right),$$

where β_2 *is defined in Lemma 3. The operator is guaranteed to be contractive when* π, μ *is within the contraction radius defined below*

$$\|\pi - \mu\|_1 < \lambda^{-1}\left(\sqrt{(1-\lambda)(\gamma^{-1}-\lambda)} + \lambda - 1\right)$$

Proof. Since $\mathcal{A}_\lambda^{\pi,\mu}$ is β_2-contractive under the $\bar{L}_2$ distance and the categorical projection Π_c is non-expansive under the $\bar{L}_2$ distance (Bellemare et al., 2019), it follows that the composed operator $\Pi_c\mathcal{A}_\lambda^{\pi,\mu}$ is also β_2-contractive. □

Lemma 5. ***(Approximation error)*** *When* $\beta_2 < 1$, *we have*

$$\bar{L}_2\left(\eta^\pi, \eta_{\mathcal{A}}^\pi\right) \le \frac{\bar{L}_2\left(\eta^\pi, \Pi_c\eta^\pi\right)}{\sqrt{1-\beta_2^2}}.$$

Proof. The result follows from an application of Proposition 5.28 in Bellemare et al. (Bellemare et al., 2023) to the off-policy distributional Q(λ) case.

□

D Alternative way to construct distributional Q(λ)

The off-policy distributional Q(λ) depends on the *path dependent* multi-step TD error $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_\# \Delta_t^\pi$. Here, the path-dependency stems from the fact that this TD error depends on the path of reward $R_{0:t}$ (Tang et al., 2022) and is fundamentally different from value-based TD error δ_t^π, which depends on the one-step transition (X_t, A_t, R_t) only.

Nevertheless, we can build an alternative variant of distributional Q(λ) by removing the path-dependency. The derivation might look heuristic: while the transformation $\left(\mathrm{b}_{G_{0:t-1},\gamma^t}\right)_\# \Delta_t^\pi$ shrinks the width of the signed measure Δ_t^π by a factor γ^t, we can instead shrink its height by pulling the factor outside of the pushforward, leading to $\gamma^t(\mathrm{b}_{G_{0:t-1},1})_\#\Delta_t^\pi$. This produces the operator

$$\tilde{\mathcal{A}}_\lambda^{\pi,\mu}\eta(x,a) \coloneqq \eta(x,a) + \mathbb{E}_\mu\left[\sum_{t=0}^\infty \gamma^t\lambda^t \cdot \left(\mathrm{b}_{G_{0:t-1},1}\right)_\# \Delta_t^\pi\right].$$

While the derivation is quite technical, we note that this formulation can be understood as treating the discount factor γ as a termination probability, rather than a scaling factor that defines the cumulative return. This is related to an alternative interpretation of the random return that recovers the same expectation as $\sum_{t=0}^\infty \gamma^t R_t$, see Bellemare et al. (2023) Chapter 2 for more discussions.

By design the operator also has η^π as its fixed point. However, as with $\mathcal{A}_\lambda^{\pi,\mu}$, its contraction property depends on λ.

Lemma 7. ***(Contraction of alternative operator)*** *The alternative operator satisfies the following property: for any* $\eta_1, \eta_2 \in \mathcal{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$,

$$\bar{\ell}_p\left(\tilde{\mathcal{A}}_\lambda^{\pi,\mu}\eta_1, \tilde{\mathcal{A}}_\lambda^{\pi,\mu}\eta_2\right) \le \tilde{\beta}\bar{\ell}_p(\eta_1, \eta_2),$$

with $\tilde{\beta} = \frac{\gamma(1+\lambda)}{1-\gamma\lambda}$. *When* $\lambda < \frac{1-\gamma}{2\gamma}$, *we have* $\tilde{\beta} < 1$ *and the operator is guaranteed to be contractive.*

Proof. The proof idea is similar to Lemma 3, where we seek to write the back-up target as a convex combination of signed measudres. Indeed, for any $\eta \in \mathcal{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$,we can write

$$\tilde{\mathcal{A}}_\lambda^{\pi,\mu}\eta(x,a) = \sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}\in\mathcal{R}^t} \tilde{w}_{x,a,r_{0:t-1}}\tilde{\eta}_{x,a,r_{0:t-1}}.$$

where $\tilde{w}$ can be negative, as before. The individual distribution writes from the pushforward operation that defines the operator

$$\tilde{\eta}_{x,a,r_{0:t-1}} \coloneqq \gamma^t \left(\mathrm{b}_{G_{0:t-1},1}\right)_{\#}\eta(x,a).$$

As a technical note, we see that unlike $\mathcal{A}_\lambda^{\pi,\mu}$, this operator cannot be written in a telescoping form. Hence we have the bound on the coefficient as

$$\sum_{t=1}^{\infty}\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\sum_{r_{0:t-1}} |\tilde{w}_{x,a,r_{0:t-1}}| \le \sum_{t=0}^{\infty}\gamma^t\lambda t\cdot\gamma + \sum_{t=1}^{\infty}\gamma^t\lambda^t = \frac{\gamma(1+\lambda)}{1-\gamma\lambda}.$$

This concludes the proof.

□

We see that the upper bound on the trace coefficient λ is $\frac{1-\gamma}{2\gamma}$ for the operator to be contractive, and strictly worse than the bound for the off-policy distributional Q(λ) operator $\mathcal{A}_\lambda^{\pi,\mu}$ since $\|\pi-\mu\|_1 \le 2$. A direct implication is that $\tilde{\mathcal{A}}_\lambda^{\pi,\mu}$ cannot benefit from multi-step learning even when on-policy, since the bound on λ is constant: for $\gamma = 0.99$, we have $\frac{1-\gamma}{2\gamma} \approx 0.01$, which is almost like one-step learning. The comparison suggests that $\mathcal{A}_\lambda^{\pi,\mu}$ is a much better construct of the multi-step learning operator.

E Optimal control

We can apply distributional Q(λ) for optimal control. We use the notation $Q_{\eta_k} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ to denote the Q-function induced by the signed return distribution η_k. At iteration k, we let the target policy to be the greedy policy with respect to Q_{η_k}, denoted as $\mathcal{G}\left(Q_{\eta_k}\right)$. Consider the recursion

$$\eta_{k+1} = \Pi_c\mathcal{A}^{\mathcal{G}(Q_{\eta_k}),\mu}\eta_k. \tag{8}$$

Under a few regularity conditions, we can guarantee that when λ is small enough, the above recursion converges to a signed return distribution which closely approximates the return distribution $\eta^* \coloneqq \eta^{\pi^*}$ of the optimal policy η^*.

Lemma 8. ***(Optimal control)*** *Assume the MDP has a unique deterministic optimal policy π^*. When $\lambda < \frac{1-\gamma}{2\gamma}$, we have $\eta_k \to \eta_{\mathcal{A}}^* \in \mathscr{M}_1(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ in $\bar{L}_2$ and*

$$\bar{L}_2\left(\eta^*, \eta_{\mathcal{A}}^*\right) \le \frac{\bar{L}_2\left(\eta^*, \Pi_c\eta^*\right)}{\sqrt{1-\gamma^2}}$$

Proof. The proof is a combination of the proof techniques applied in categorical distributional Q-learning (Rowland et al., 2018) and value-based Q(λ) (Harutyunyan et al., 2016).

Let Q_k be the Q-function induced by the iterate η_k, then we can argue that the evolution of Q_k is as if the Q-function iterates are generated under the value-based Q(λ) algorithm. Note that the condition on $\lambda < \frac{1-\gamma}{2\gamma}$ means that $\mathcal{A}_\lambda^{\pi,\mu}$ is contractive for any π,μ. Indeed $\|\pi-\mu\|_1 < 2$ and hence for any π,μ, the corresponding bound on the contraction rate is less than 1. This means the Q-function iterate Q_k will converge to the optimal Q-function Q^* by the unique optimal policy π^*.

Then we follow an identical trace of argument from Rowland et al. (2018) to show the convergence of the signed measure iterate η_k, this includes proving the existence of a limiting signed measure $\eta_{\mathcal{A}}^*$ and its $\bar{L}_2$ distance to the optimal return distribution η^*.

□

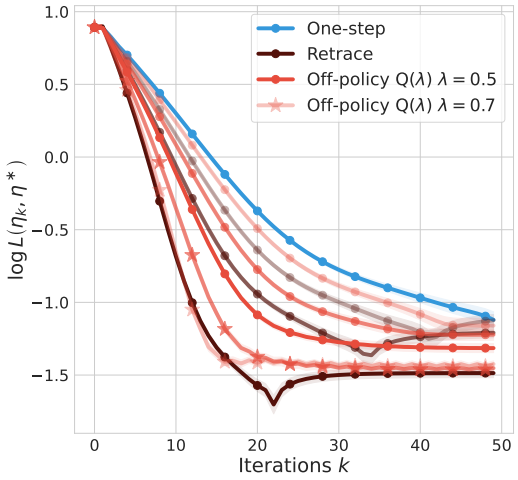

Figure 5: The distance between the algorithmic iterate η_k and return distribution for the optimal policy η^*, as we run control algorithms with distributional one-step, Retrace and off-policy Q(λ). All algorithms use categorical representations and set greedy policy as the target policy. Different curves show an algorithmic variant with a different hyper-parameter setting ($\bar{c}$ for Retrace and λ for Q(λ)). Unlike Figure 2 with $|\mathcal{A}| = 20$, here with $|\mathcal{A}| = 5$ all algorithmic behavior changes slightly. Since the problem effectively becomes less off-policy, Retrace can benefit from the full trace with $\bar{c} = 4$, outperforming Q(λ); meanwhile, Q(λ) becomes more stable across all λ values.

Since the above result is inherited from the policy evaluation case, it puts a fairly conservative restriction on λ. In practice, we find that using a larger value of λ can also lead to stable learning in both tabular and large-scale settings (Section 7).

In Figure 2, we compare the speed of convergence by measuring the Cramer distance $\ell_2\left(\eta^*(x_0, a_0)\right), \eta_k(x_0, a_0))$ at a fixed state action pair (x_0, a_0) throughout iterations and across randomly generated MDPs. When λ is properly chosen (in this case $\lambda = 0.4$), off-policy Q(λ) improves over baselines both in terms of the rate of convergence and the asymptotic accuracy, compatible with observations made in the off-policy evaluation case. However, a caveat is that the improvement of Q(λ) comes at the cost of having to tune trace parameter λ in practice: when λ is too small, the learning is guaranteed to be stable but one does not benefit from multi-step learning ($\lambda = 0$ recovers the one-step algorithm); when λ is too large (e.g., $\lambda \approx 1$), the learning can become unstable as shown in the experiments.

F Experiments

We provide additional experimental details and results.

F.1 Tabular MDP

For the tabular MDP, the transition probability $p(\cdot|x, a)$ is randomly sampled from a Dirichlet distribution with $|\mathcal{X}|$ entries with a rate of 0.1, i.e., Dirichlet $([0.1, 0.1, ...0.1])$, for all (x, a) independently. The reward function $r(x, a)$ is randomly sampled from $\mathcal{N}(0, 1)$ and fixed as a deterministic reward for each (x, a). The discount factor is set as $\gamma = 0.9$. The behavior policy μ is uniform throughout, and target policy is always greedy with respect to the Q-function induced by the current distribution iterate. For all experiments, we generate the iterate as

$$\eta_{k+1} = \Pi_c \mathcal{R} \eta_k$$

where $\mathcal{R}$ is the distributional operator of interest. The initial iterate η_0 assigns uniform weights across all $m = 10$ atoms. We calculate η^* by first finding the optimal policy π^* and then generate Monte-Carlo returns from π^* for an empirical estimate η^*, followed by a projection onto the m atoms.

We have set $|\mathcal{X}| = 5$ throughout but vary $|\mathcal{A} = 10|$ for ablations. For each algorithmic variant, we sweep over certain hyper-parameters, such as $\bar{c} \in \{1, 2, 4\}$ for Retrace and $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for off-policy Q(λ).

Algorithm 1 Q(λ)-C51

Parameterized categorical distribution η_θ with main network parameter θ and target network parameter θ^-
for $k = 1, 2...$ **do**
  Sample trajectory $(X_t, A_t, R_t)_{t=0}^{\infty}$ from the replay.
  Compute gradient estimate $\widehat{g}_\theta$ based on Eqn (6) for the sampled initial state-action pair (X_0, A_0).
  Update parameter $\theta \leftarrow \theta - \kappa \widehat{g}_\theta$.
  Update target parameter $\theta^- \leftarrow (1-\tau)\theta^- + \tau\theta$.
end for
Output final distribution η_θ.

The case with $|\mathcal{A} = 5|$. Figure 5 shows results for when $|\mathcal{A} = 5|$ rather than $|\mathcal{A} = 20|$ in Figure 2. We see that since the number of action gets smaller, the problem has become effectively much less off-policy. As a result, Q(λ) becomes stable for all levels of λ that we sweep. There is also a general trend of improved contraction and fixed point error as λ increases from 0.1 to 0.9. Retrace outperforms Q(λ) when $\bar{c} = 4$, since the algorithm can effectively make use of the full trace for distributional learning by IS.

F.2 Deep RL with Atari

We provide further details on the Atari experiments.

Evaluation. 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. The inter-quartile metric is calculated by dropping out tail samples from across all games and seeds (Agarwal et al., 2021).

Shared settings for all algorithmic variants. All algorithmic variants use the same torso architecture as DQN (Mnih et al., 2015) and differ in the head outputs, which we specify below. All agents an Adam optimizer (Kingma and Ba, 2015) 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 $\epsilon_{\max} = 1$ and linearly decays to $\epsilon_{\min} = 0.01$ over training. At evaluation time, the agent adopts $\epsilon = 0.001$; the small exploration probability is to prevent the agent from getting stuck.

For C51, 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 (Bellemare et al., 2017a). The induced Q-function is computed as $Q_\theta(x, a) = \sum_{i=1}^m p_i(x, a; \theta) z_i$.

Target and behavior policy. Since the behavior policy μ is ϵ-greedy, we have access to the probability distributions $\mu(a)$ for each action a, which gets stored in the replay buffer along with the transition. At learning time, the algorithms sample from the reply buffer and construct back-up targets. The stored probabilities $\mu(a)$ allow for IS techniques applied in distributional Retrace, when combined with a target policy π.

For Retrace and one-step, we set π as the greedy policy with respect to the Q-function induced by the learner return distribution. For Q(λ), with a fixed λ, we find that this works sub-optimally. One potential explanation is that the level of off-policyness changes throughout learning, and so a single λ might not work optimally across the entire learning process. Instead, we borrow inspirations from the trust region literature, and set the target policy as a mixture of the greedy policy and behavior policy as in Eqn (7). This allows for better performance for off-policy Q(λ).