A Differential Perspective on Distributional Reinforcement Learning

Juan Sebastian Rojas[1], **Chi-Guhn Lee**[1]

[1]Department of Mechanical & Industrial Engineering, University of Toronto
juan.rojas@mail.utoronto.ca, cglee@mie.utoronto.ca

Abstract

To date, distributional reinforcement learning (distributional RL) methods have exclusively focused on the *discounted* setting, where an agent aims to optimize a discounted sum of rewards over time. In this work, we extend distributional RL to the *average-reward* setting, where an agent aims to optimize the reward received per time step. In particular, we utilize a quantile-based approach to develop the first set of algorithms that can successfully learn and/or optimize the long-run per-step reward distribution, as well as the differential return distribution of an average-reward MDP. We derive proven-convergent tabular algorithms for both prediction and control, as well as a broader family of algorithms that have appealing scaling properties. Empirically, we find that these algorithms yield competitive and sometimes superior performance when compared to their non-distributional equivalents, while also capturing rich information about the long-run per-step reward and differential return distributions.

1 Introduction

Distributional reinforcement learning (distributional RL) (Bellemare, Dabney, and Rowland 2023) equips decision-making agents with the ability to learn and reason about the probability distribution over a given objective. This approach transcends the traditional RL paradigm of focusing solely on expected values, thereby offering a more insightful and methodical understanding of the variability, uncertainty, and risk associated with a given objective. To date, distributional RL methods have exclusively focused on the *discounted* setting, where an RL agent aims to optimize a potentially-discounted sum of rewards over time (e.g. Bellemare, Dabney, and Munos (2017)).

In this work, we extend distributional RL to the *average-reward* setting, where an RL agent aims to optimize the reward received per time step. This extension offers a timely opportunity to extend the benefits of distributional RL to a promising and growing family of RL methods. In particular, unlike discounted RL methods, average-reward RL methods do not face fundamental challenges in continuing control tasks that require function approximation (Naik et al. 2019). Moreover, average-reward RL methods have been shown to outperform discounted RL methods in some instances (e.g. Adamczyk et al. (2025)). It has even been shown that integrating the average-reward itself into discounted RL methods can increase their performance (e.g. Naik et al. (2024)). Accordingly, extending distributional RL to the average-reward setting allows us to combine the strengths of two increasingly important RL frameworks.

Through this extension, we will see that, from a distributional perspective, we will need to rethink the following foundational questions: *what do we want to learn?* and *how can we learn it?* In this work, we address these questions, and in the process, derive the first distributional framework for the average-reward setting.

2 Related Work

Early works on distributional RL include Sobel (1982), White (1988), and Morimura et al. (2010). Modern distributional RL is typically associated with the distributional Bellman operator, which was introduced in Bellemare, Dabney, and Munos (2017), along with empirical results which showed that such methods could outperform their non-distributional equivalents. Subsequent research has expanded upon this framework by developing methods that can typically be classified as either *categorical* (e.g. Bellemare, Dabney, and Munos (2017)) or *quantile*-based (e.g. Dabney et al. (2018)). While these methods have proven to be effective in the discounted setting, no attention has been given to the average-reward setting. In this work, we address this gap by introducing the first distributional RL algorithms specifically designed for the average-reward setting.

3 Preliminaries

3.1 Average-Reward Reinforcement Learning

A finite average-reward MDP is the tuple $\mathcal{M} \doteq \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, p \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{R} \subset \mathbb{R}$ is a finite set of rewards, and $p : \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S} \to [0, 1]$ is a probabilistic transition function that describes the dynamics of the environment. At each discrete time step, $t = 0, 1, 2, \ldots$, an agent chooses an action, $A_t \in \mathcal{A}$, based on its current state, $S_t \in \mathcal{S}$, and receives a reward, $R_{t+1} \in \mathcal{R}$, while transitioning to a (potentially) new state, S_{t+1}, such that $p(s', r \mid s, a) = \mathbb{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$. In an average-reward MDP, an agent aims to find a policy, $\pi : \mathcal{S} \to \mathcal{A}$, that optimizes the

long-run (or limiting) average-reward, $\bar{r}$, which is defined as follows for a given policy, π:

$$\bar{r}_\pi(s) \doteq \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}[R_t \mid S_0 = s, A_{0:t-1} \sim \pi]. \quad (1)$$

In this work, we limit our discussion to *stationary Markov* policies, which are time-independent policies that satisfy the Markov property.

When working with average-reward MDPs, it is common to simplify Equation (1) into a more workable form by making certain assumptions about the Markov chain induced by following policy π. To this end, a *unichain* assumption is typically used when doing prediction (learning) because it ensures the existence of a unique limiting distribution of states, $\mu_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}(S_t = s \mid A_{0:t-1} \sim \pi)$, that is independent of the initial state, thereby simplifying Equation (1) to the following:

$$\bar{r}_\pi = \sum_{s\in\mathcal{S}} \mu_\pi(s) \sum_{a\in\mathcal{A}} \pi(a \mid s) \sum_{s'\in\mathcal{S}} \sum_{r\in\mathcal{R}} p(s', r \mid s, a) r. \quad (2)$$

Similarly, a *communicating* assumption is typically used when doing control (optimization) because it ensures the existence of a unique optimal average-reward, $\bar{r}*$, that is independent of the initial state.

The *return* of an MDP, G_t, captures how rewards are aggregated over the time horizon. In an average-reward MDP, the return is referred to as the *differential return*, and is defined as follows:

$$G_t \doteq R_{t+1} - \bar{r}_\pi + R_{t+2} - \bar{r}_\pi + R_{t+3} - \bar{r}_\pi + \dots. \quad (3)$$

To solve an average-reward MDP, solution methods such as dynamic programming or RL are typically used in conjunction with the *Bellman equation* (4) for the state-value function, $v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi[R_{t+1} - \bar{r}_\pi + G_{t+1} \mid S_t = s]$, or the *Bellman optimality equation* (5) for the state-action value function, $q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$. Solution methods for average-reward MDPs are typically referred to as *differential* methods because of the reward difference (i.e., $r - \bar{r}_\pi$) operation that occurs in the Bellman equations (4) and (5):

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)[r - \bar{r}_\pi + v_\pi(s')], \quad (4)$$

$$q_\pi(s,a) = \sum_{s',r} p(s', r \mid s, a)[r - \bar{r}_\pi + \max_{a'} q_\pi(s', a')]. \quad (5)$$

3.2 Quantile Regression

Quantile regression (Koenker 2005) refers to the process of estimating a predetermined quantile of a probability distribution from samples. More specifically, for $\tau \in (0,1)$, let $F_w^{-1}(\tau)$ denote the τ^{th} quantile that we are trying to estimate from probability distribution w (where $F_w : \mathbb{R} \to [0,1]$ denotes the CDF of w). Quantile regression maintains an estimate, θ, of this value, and updates the estimate based on samples drawn from w (i.e., $y \sim w$) as follows (Bellemare, Dabney, and Rowland 2023):

$$\theta \leftarrow \theta + \alpha_t(\tau - \mathbb{1}_{\{y<\theta\}}), \quad (6)$$

where α_t denotes the step size for the update. The estimate for θ will continue to adjust until the equilibrium point, θ^*, which corresponds to $F_w^{-1}(\tau)$, is reached (Bellemare, Dabney, and Rowland 2023). In other words, we have that

$$0 = \mathbb{E}[(\tau - \mathbb{1}_{\{y<\theta^*\}})] \quad (7a)$$
$$= \tau - \mathbb{E}[\mathbb{1}_{\{y<\theta^*\}}] \quad (7b)$$
$$= \tau - \mathbb{P}(y < \theta^*) \quad (7c)$$
$$\implies \theta^* = F_w^{-1}(\tau). \quad (7d)$$

3.3 Discounted Distributional RL

Like average-reward MDPs, discounted MDPs have their own return which captures how rewards are aggregated over the time horizon. More specifically, let

$$G_t^\gamma = \sum_{k\geq 0} \gamma^k R_{t+k+1} = R_{t+1} + \gamma G_{t+1}^\gamma \quad (8)$$

denote the *discounted return* from a discounted MDP, where $R_{t+1} \in \mathcal{R}$ denotes the per-step reward and $\gamma \in [0,1]$ denotes the discount factor (Puterman 1994; Sutton and Barto 2018). The aim of *discounted distributional RL* is to learn the probability distribution over discounted returns. More formally, discounted distributional RL aims to learn the discounted return distribution function, Φ_π^γ, such that, with a slight abuse of notation, $\Phi_\pi^\gamma(s)$ denotes the probability distribution over discounted returns when starting from state $s \in \mathcal{S}$ and following policy π, and $\Phi_\pi^\gamma(s,a)$ denotes the probability distribution over discounted returns when starting from state $s \in \mathcal{S}$, taking action $a \in \mathcal{A}$, and following policy π.

Broadly speaking, discounted distributional RL methods can be categorized based on how they approximate (or parameterize) the discounted return distribution function. In this work, we take inspiration from quantile-based methods (Dabney et al. 2018; Rowland et al. 2024), which parameterize the discounted return distribution function as follows:

$$\Phi_\pi^\gamma(s) = \sum_{i=1}^{n} \frac{1}{n} \delta_{\Omega_i^\gamma(s)} \quad \text{or} \quad \Phi_\pi^\gamma(s,a) = \sum_{i=1}^{n} \frac{1}{n} \delta_{\Omega_i^\gamma(s,a)}, \quad (9)$$

where Ω_i^γ denotes the τ_i-quantile of the discounted return distribution, and $\delta_{\Omega_i^\gamma}$ denotes a Dirac at Ω_i^γ.

More formally, the set of τ-quantiles of a probability distribution can be defined as follows:

Definition 3.1 (Rowland et al. (2024))**.** *Let $\mathscr{P}(\mathbb{R})$ denote the set of probability distributions over $\mathbb{R}$. For a probability distribution $w \in \mathscr{P}(\mathbb{R})$ and parameter $\tau \in (0,1)$, the set of τ-quantiles of w is given by the set $\{z \in \mathbb{R} : F_w(z) = \tau\} \cup \inf\{y \in \mathbb{R} : F_w(y) > \tau\}$, where $F_w : \mathbb{R} \to [0,1]$ is the CDF of w, defined by $F_w(t) = \mathbb{P}_{Z\sim w}(Z \leq t)$ for all $t \in \mathbb{R}$.*

Quantile-based discounted distributional RL methods learn the τ-quantiles of the discounted return distribution, $\{\Omega_i^\gamma\}_{i=1}^n$, from samples using quantile regression (see Equation 6) as follows (Dabney et al. 2018):

$$\Omega_{i,t+1}^\gamma = \Omega_{i,t}^\gamma + \alpha_t \frac{1}{n} \sum_{j=1}^{n} \left[\tau_i - \mathbb{1}_{\{\psi_t<0\}}\right], \ \forall i = 1, 2, \dots, n, \quad (10)$$

where $\Omega^{\gamma}_{i,t}$ denotes the estimate of the τ_i-quantile of the discounted return distribution at time t, α_t denotes the step size for the update, and $\{\tau_i \in (0,1)\}_{i=1}^{n}$, such that $\tau_i = \frac{2i-1}{2n}$, $i = 1, 2, \ldots, n$. Here, the choice of ψ_t depends on whether we want to do prediction (learning) or control (optimization). In the case of prediction, $\psi_t = R_{t+1} + \gamma\Omega^{\gamma}_{j,t}(S_{t+1}) - \Omega^{\gamma}_{i,t}(S_t)$. In the case of control (via Q-learning (Watkins and Dayan 1992)), $\psi_t = R_{t+1} + \gamma\Omega^{\gamma}_{j,t}(S_{t+1}, a^*) - \Omega^{\gamma}_{i,t}(S_t, A_t)$, where $a^* \doteq \text{argmax}_{a'} \frac{1}{n}\sum_{k=1}^{n} \Omega^{\gamma}_{k,t}(S_{t+1}, a')$.

Let us take a moment to highlight the definition of the above greedy action, a^*. We are being greedy with respect to the *average* of the τ-quantiles of the discounted return distribution, which, by definition, is equivalent to being greedy with respect to the expected discounted return, and hence, the (discounted) state-action value function. More formally, the *greedy action selection rule* for a distributional RL algorithm can be defined as follows:

Definition 3.2 (Bellemare, Dabney, and Rowland (2023))**.** *Let $\boldsymbol{\pi}_{MS}$ denote the space of stationary Markov policies, and let $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denote a state-action value function. A greedy action selection rule is a mapping, $\mathcal{G} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \boldsymbol{\pi}_{MS}$, with the property that for any $Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $\mathcal{G}(Q)$ is greedy with respect to Q. That is, $\mathcal{G}(Q)(a \mid s) > 0 \implies Q(s,a) = \max_{a'\in\mathcal{A}} Q(s,a')$.*

Importantly, we note that although discounted distributional RL methods make it possible to capture information related to the underlying (discounted) return distribution, much of the theoretical guarantees associated with these methods in the control setting require an action selection rule that satisfies Definition 3.2. That is, they require a greedy action selection rule that is greedy with respect to the expected return, and hence, the state-action value function (Bellemare, Dabney, and Rowland 2023).

Similarly, we note that the convergence of discounted distributional RL approaches, including the above quantile-based approach, is governed by the (discounted) distributional Bellman operator (Bellemare, Dabney, and Munos 2017; Bellemare, Dabney, and Rowland 2023). Importantly, under certain conditions, the distributional Bellman operator is a contraction in the Wasserstein metric, thereby ensuring convergence to a fixed point.

4 Differential Distributional RL

In this section, we derive and present our primary contribution: the first set of distributional RL algorithms specifically designed for the average-reward setting. We call these algorithms, *Differential Distributional RL* algorithms, due to the differential nature of average-reward RL algorithms.

Importantly, this extension of distributional RL into the average-reward setting requires that we move away from using the discounted MDP, which forms the basis of existing distributional RL methods, and, in turn, utilize the average-reward MDP. Consequently, we cannot rely on the (discounted) distributional Bellman operator (Bellemare, Dabney, and Munos 2017; Bellemare, Dabney, and Rowland 2023) when formulating our approach.

In fact, this shift requires that we rethink and answer the following foundational questions: *what do we want to learn?* and *how can we learn it?* In this section, we propose a quantile-based framework that allows us to address these two questions, and in the process, extend distributional RL into the average-reward setting.

4.1 The Differential Distributional Objective

Before we can derive our algorithms, we first need to establish an appropriate distributional objective. In particular, the average-reward formulation suggests *two* distributions that may be of interest. The first corresponds to the probability distribution over differential returns (where the differential return is defined in Equation (3)). The second corresponds to the limiting (or long-run) per-step reward distribution, whose mean yields the regular (non-distributional) average-reward objective (i.e., Equation (1)).

At a first glance, it may seem natural to mirror the approach taken in the discounted setting by pursuing the differential return as the distributional objective. However, such an approach would not be fully aligned with the nature of the average-reward setting. In particular, in the average-reward setting, the differential return serves as a surrogate objective, optimized only to facilitate the optimization of the long-run per-step average-reward (Equation (1)). As such, while it may be feasible to capture information about the differential return distribution, such information would be of little relevance in the average-reward setting, where the long-term, per-step behaviour is what is of interest. Conversely, the limiting per-step reward distribution aligns directly with the average-reward objective (its mean yields the average-reward itself), thereby making it an appealing distributional objective for the average-reward setting.

As such, given the above reasoning, which is formalized as Proposition 4.1 below, the primary focus of this work will be to derive differential algorithms that can learn and/or optimize the *limiting per-step reward distribution*. We will briefly revisit the return distribution from a purely empirical perspective in Section 4.3.

Proposition 4.1. *The limiting per-step reward distribution is the natural distributional objective in the average-reward setting, given that its mean yields the long-run average-reward, which is the primary prediction and control objective of (non-distributional) average-reward RL.*

We are now ready to begin our theoretical treatment of the limiting per-step reward distribution. We begin by formally defining our distributional objective. In particular, for a given policy, π, let

$$\phi_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}(R_t \mid S_0 = s, A_{0:t-1} \sim \pi) \qquad (11)$$

denote the limiting per-step reward distribution induced by following policy π, where $R_t \in \mathcal{R}$ denotes the per-step reward. As is standard practice with the regular average-reward objective, we can simplify Equation (11) by making certain assumptions about the Markov chain induced by following policy π. To this end, we will utilize a *unichain* assumption when doing prediction (learning), because it ensures the existence of a unique limiting distribution of per-step rewards that is independent of the initial state, such

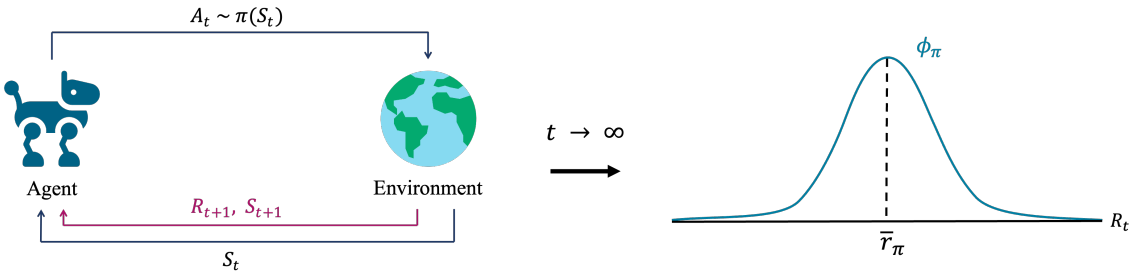

Figure 1: Illustration of the agent-environment interaction in an average-reward MDP. As $t \to \infty$, following policy π yields a limiting per-step reward distribution, ϕ_π, with an average-reward, $\bar{r}_\pi$. Standard average-reward RL methods aim to learn and/or optimize the average-reward, $\bar{r}_\pi$. By contrast, the differential distributional RL methods proposed in this work aim to learn and/or optimize the limiting per-step reward distribution, ϕ_π.

that $\phi_\pi(s) = \phi_\pi$. Similarly, we will utilize a *communicating* assumption when doing control (optimization), because it ensures the existence of a unique optimal distribution of per-step rewards that is independent of the initial state. Figure 1 depicts the agent-environment interaction in an average-reward MDP, where following policy π yields a limiting reward distribution, ϕ_π, whose mean corresponds to the average-reward objective, $\bar{r}_\pi$.

4.2 A Quantile-Based Approach

In this section, we take inspiration from the quantile-based methods utilized for discounted distributional RL to develop a quantile-based approach that can be used to learn and/or optimize the limiting per-step reward distribution of an average-reward MDP. Like the discounted setting (see Section 3.3), our overall goal and strategy is to derive an appropriate set of quantile regression-based updates that approximate our target distribution. More formally, we approximate (or parameterize) the limiting per-step reward distribution as follows:

$$\phi_\pi = \sum_{i=1}^{m} \frac{1}{m} \delta_{\theta_i}, \tag{12}$$

where θ_i denotes the τ_i-quantile of the limiting per-step reward distribution, and δ_{θ_i} denotes a Dirac at θ_i. Here, we adopt the same formal definition for the τ-quantiles of a probability distribution as in the discounted setting (i.e., Definition 3.1).

Now, let us consider the τ-quantiles of the limiting per-step reward distribution, $\{\theta_i\}_{i=1}^m$. As previously mentioned, we wish to approximate these quantiles using quantile regression. To this end, the generic quantile regression framework outlined in Equations (6) and (7) suggest the following set of quantile update rules for the per-step reward quantiles:

$$\theta_{i,t+1} = \theta_{i,t} + \alpha_t \left(\tau_i - \mathbb{1}_{\{R_{t+1} < \theta_{i,t}\}}\right) \ \forall i = 1, 2, \ldots, m, \tag{13}$$

where $\theta_{i,t}$ denotes the estimate of the τ_i-quantile of the limiting per-step reward distribution, α_t denotes the step size for the update, and R_{t+1} denotes the per-step reward.

When comparing this quantile update rule to that of the discounted setting (i.e., Equation (10)), we can see that both rules have a similar structure, with the key difference being that we have replaced ψ_t with $R_{t+1} - \theta_{i,t}$. Most notably, unlike the return-based quantile update rule used in the discounted setting, the per-step reward quantile update rule (13) does not contain the full TD or Q-learning targets and estimates. This is not necessarily an issue, so long as we can find a way to properly incorporate the quantile update rule (13) into a broader algorithm that does contain the full TD or Q-learning targets and estimates.

To this end, we note that, typically, differential RL algorithms will learn and/or optimize the value function and average-reward simultaneously. From an algorithmic perspective, this means that the algorithms start with initial estimates (or guesses) for the value function and average-reward, then update these estimates over time, until they have learned and/or optimized these two objectives. As such, one way through which we can incorporate the quantile updates into a differential algorithm is by incorporating them into the average-reward estimate update. In particular, we can relate the average-reward, $\bar{r}_\pi$, to the τ-quantiles of the limiting per-step reward distribution, $\{\theta_i\}_{i=1}^m$, as follows:

Lemma 4.2. *Given a unichain, communicating, or equivalent assumption, consider the limiting per-step reward distribution parameterized by the τ-quantiles, $\{\theta_i\}_{i=1}^m$, as per Equation (12) for $\{\tau_i \in (0,1)\}_{i=1}^m$, such that $\tau_i = \frac{2i-1}{2m}$, $i = 1, 2, \ldots, m$. The expected value of the τ-quantiles converges to the average-reward of the limiting per-step reward distribution, $\bar{r}_\pi$, as $m \to \infty$.*

Proof. Let $F(r)$ denote the CDF of the limiting per-step reward distribution (which exists and is unique given a unichain, communicating, or equivalent assumption), and let θ_i represent its τ_i-quantile, such that $F(\theta_i) = \tau_i$. The expected value of the τ-quantiles is:

$$\frac{1}{m}\sum_{i=1}^{m} \theta_i = \frac{1}{m}\sum_{i=1}^{m} F^{-1}\left(\frac{2i-1}{2m}\right), \tag{14}$$

where F^{-1} is the inverse CDF.

By the definition of $\tau_i = \frac{2i-1}{2m}$, the τ_i values are evenly spaced over $[0,1]$, creating a uniform partition. Hence, as $m \to \infty$, the summation $\frac{1}{m}\sum_{i=1}^{m} \theta_i$ becomes a Riemann sum for the following integral:

$$\int_0^1 F^{-1}(\tau)\, d\tau. \tag{15}$$

It is a well-known result that for a random variable X with CDF $F(x)$, $E[X] = \int_0^1 F^{-1}(x)dx$. Therefore, as $m \to \infty$,

$$\frac{1}{m}\sum_{i=1}^{m} \theta_i \to \int_0^1 F^{-1}(\tau)\, d\tau = \bar{r}_\pi. \tag{16}$$

Thus, the expected value of the τ-quantiles converges to the average reward, $\bar{r}_\pi$, of the limiting per-step reward distribution as $m \to \infty$. This completes the proof. □

Hence, we now have a way to express the set of quantile updates (13) as an average-reward update:

$$\bar{R}_{t+1} = \frac{1}{m}\sum_{i=1}^{m} \theta_{i,t+1}, \text{ where} \tag{17a}$$

$$\theta_{i,t+1} = \theta_{i,t} + \alpha_{i,t}\left(\tau_i - \mathbb{1}_{\{R_{t+1} < \theta_{i,t}\}}\right),\ \forall i = 1, 2, \ldots, m, \tag{17b}$$

and $\bar{R}_{t+1}$ denotes an estimate of the average-reward, $\bar{r}_\pi$. Importantly, we can show that this quantile-based average-reward update converges to the average-reward of the limiting per-step reward distribution induced by a given policy:

Theorem 4.3. *Given a unichain, communicating, or equivalent assumption, consider the limiting per-step reward distribution induced by following policy π, and parameterized by the τ-quantiles, $\{\theta_i\}_{i=1}^m$, as per Equation* (12) *for $\{\tau_i \in (0,1)\}_{i=1}^m$, such that $\tau_i = \frac{2i-1}{2m},\ i = 1, 2, \ldots, m$. Also consider the quantile-based update rules for the average-reward estimate* (17)*, along with corresponding step sizes that satisfy: $\alpha_t > 0,\ \sup_{t\geq 0}\alpha_t < \infty,\ \sum_{t=0}^{\infty}\alpha_t = \infty$, and $\alpha_t = o(1/log\ t)$. If the quantile estimates, $\{\theta_{i,t}\}_{i=1}^m$, converge a.s. to the τ-quantiles of the limiting per-step reward distribution, then the average-reward estimate, $\bar{R}_t$, converges a.s. to the average-reward of the limiting per-step reward distribution, $\bar{r}_\pi$, as $t \to \infty$.*

Proof. If $\{\theta_{i,t}\}_{i=1}^m \to \{\theta_i\}_{i=1}^m$ as $t \to \infty$, then, by Lemma 4.2, it directly follows that the average of these estimates, which corresponds to $\bar{R}_t$, converges, almost surely, to $\bar{r}_\pi$ as $t \to \infty$. This completes the proof. □

Hence, given the convergence of the τ-quantile estimates, it is intuitive to see that the average-reward estimate will also converge. We provide a full convergence proof for the τ-quantile estimates in Appendix B.

As such, we now have a way to learn the limiting per-step reward distribution induced by a given policy. We can further extend these results into control-based settings by choosing an appropriate action selection rule. As in the discounted setting, we will utilize an action selection rule that is consistent with Definition 3.2. To this end, Algorithm 1 shows the Differential Distributional (or *D2*) Q-learning algorithm, which adopts a greedy action selection rule with respect to the state-action value function. The full set of Differential Distributional RL algorithms is included in Appendix A.

Algorithm 1: Differential Distributional (D2) Q-Learning

```
Obtain initial S
while still time to train do
    A ← action given by π for S
    Take action A, observe R, S'
    for i = 1, 2, ..., m do
        θ_i = θ_i + α_θ (τ_i − 𝟙_{R<θ_i})
    end for
    R̄ = (1/m) Σ_{i=1}^m θ_i
    δ = R − R̄ + max_a Q(S', a) − Q(S, A)
    Q(S, A) = Q(S, A) + αδ
    S = S'
end while
return {θ_i}_{i=1}^m
```

Importantly, with the aforementioned action selection rule, we retain the regular (non-distributional) Q-learning operator for the average-reward setting:

$$TQ(s,a) \doteq \sum_{s',r} p(s', r \mid s, a)(r + \max_{a'} Q(s', a')). \tag{18}$$

As such, we can use existing results to establish that the convergence of the state-action value estimates implies that the average-reward converges to the optimal average-reward:

Theorem 4.4. *Let $\bar{r}_{\pi_t}$ denote the average-reward obtained when following policy π_t, such that π_t is a greedy policy with respect to the state-action value estimates, $Q_t(s,a)$, where $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Also, let q^* denote the solution of the Bellman optimality equation* (5)*. If a communicating or equivalent assumption holds, such that there exists a unique q^* up to an additive constant, then we have that if $Q_t \to q^*$ a.s. as $t \to \infty$, then $\bar{r}_{\pi_t} \to \bar{r}*$ a.s. as $t \to \infty$.*

Proof. We adopt the proof technique used in Wan, Naik, and Sutton (2021) to prove a similar result. In particular, the desired result follows directly from Theorem 8.5.5 of Puterman (1994). More specifically, we have that:

$$\begin{aligned} \min_{s,a}(TQ_t - Q_t) &\leq \bar{r}_{\pi_t} \\ &\leq \bar{r}* \\ &\leq \max_{s,a}(TQ_t - Q_t), \end{aligned} \tag{19}$$

$$\implies |\bar{r}* - \bar{r}_{\pi_t}| \leq sp(TQ_t - Q_t), \tag{20}$$

where $TQ(s,a)$ is defined in Equation (18). Because $Q_t \to q^*$ a.s. as $t \to \infty$, and $sp(TQ_t - Q_t)$ is clearly a continuous function of Q_t, we have, by the continuous mapping theorem, that $sp(TQ_t - Q_t) \to sp(Tq^* - q^*) = 0$ a.s. as $t \to \infty$. Hence, we can conclude that $\bar{r}_{\pi_t} \to \bar{r}*$ a.s. as $t \to \infty$. This completes the proof. □

The above theorem requires that $Q_t \to q^*$ a.s. as $t \to \infty$. In this regard, we provide a full convergence proof for the convergence of Q_t in Appendix B. We also provide a full proof for the convergence of the value function estimates in the prediction setting (i.e., V_t) in Appendix B. Altogether, the theoretical results provided in this section and in Appendix B show the almost sure convergence of the tabular D2 algorithms for both prediction and control.

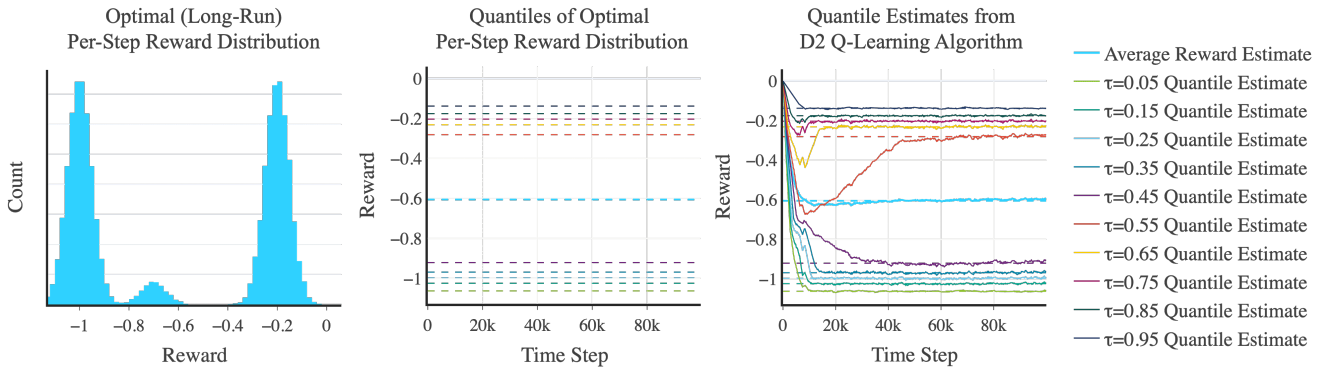

Figure 2: **a)** Histogram showing the empirical (ε-greedy) optimal (long-run) per-step reward distribution in the red-pill blue-pill task. **b)** Quantiles of the optimal per-step reward distribution in the red-pill blue-pill task. **c)** Convergence plot of the per-step reward quantile estimates as learning progresses when using the D2 Q-learning algorithm in the red-pill blue-pill task.

4.3 Learning the Differential Return Distribution

Although we argue in this work that the limiting per-step reward distribution is the natural distributional objective in the average-reward setting, it may still be useful to consider the differential return distribution from an empirical perspective (i.e., it could potentially yield superior empirical performance). As such, in this work we explore such an approach from a purely empirical perspective, and propose a set of algorithms that parameterize the probability distribution over differential returns, as well as the limiting per-step reward distribution. We call the resulting set of algorithms, *Double Differential Distributional*, or *D3*, algorithms because they simultaneously learn and/or optimize both distributions.

To this end, when we apply an analogous framework to the one described in Section 3.3, such that $\psi_t = R_{t+1} - \bar{R}_t + \Omega_{j,t}(S_{t+1}, a^*) - \Omega_{i,t}(S_t, A_t)$, and incorporate the quantile-based average-reward update (17), we arrive at the D3 Q-learning algorithm (2). The full set of Double Differential Distributional RL algorithms is included in Appendix A.

Algorithm 2: D3 Q-Learning

Obtain initial S
while still time to train **do**
  $A \leftarrow$ action given by π for S
  Take action A, observe R, S'
  $\theta_i = \theta_i + \alpha_\theta \left(\tau_i - \mathbb{1}_{\{R<\theta_i\}}\right), \forall i = 1, 2, \ldots, m$
  $\bar{R} = \frac{1}{m}\sum_{i=1}^{m} \theta_i$
  $a^* = \text{argmax}_{a'} \frac{1}{n}\sum_{j=1}^{n} \Omega_j(S', a')$
  for $j = 1, 2, \ldots, n$ **do**
    $\beta = \frac{1}{n}\sum_{k=1}^{n} \left[\tau_j - \mathbb{1}_{\{R-\bar{R}+\Omega_k(S',a^*)-\Omega_j(S,A)<0\}}\right]$
    $\Omega_j(S, A) = \Omega_j(S, A) + \alpha\beta$
  end for
  $S = S'$
end while
return $\{\theta_i\}_{i=1}^m$ and $\{\Omega_j\}_{j=1}^n$

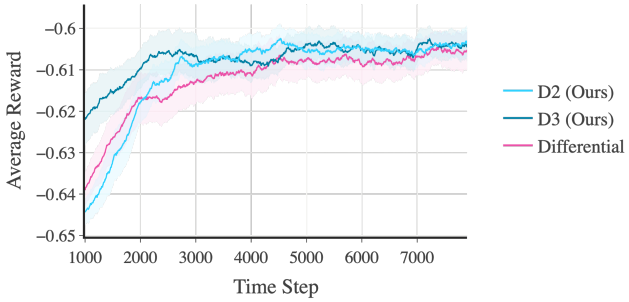

Figure 3: Rolling average-reward when using the D2 and D3 algorithms vs. a non-distributional Differential algorithm in the red-pill blue-pill environment. A solid line denotes the mean average-reward, and the corresponding shaded region denotes a 95% confidence interval over 50 runs.

5 Experimental Results

In this section, we present empirical results obtained when applying our D2 and D3 algorithms on two groups of experiments. In the first group of experiments, we aimed to validate whether the D2 and D3 algorithms could successfully learn the optimal per-step reward distribution by conducting experiments in environments where the optimal per-step reward distribution is known. In the second group of experiments, we aimed to evaluate the empirical performance of the D2 and D3 algorithms in more difficult environments. In both groups, we compared the performance of our algorithms to that of non-distributional Differential algorithms derived from the framework proposed in Wan, Naik, and Sutton (2021). The full set of experimental details and results, including additional experiments performed, is provided in Appendix C. Below, we highlight the key results.

In terms of empirical results from the first group of experiments, Figure 2 shows the agent's (per-step) reward quantile estimates as learning progresses when using the D2

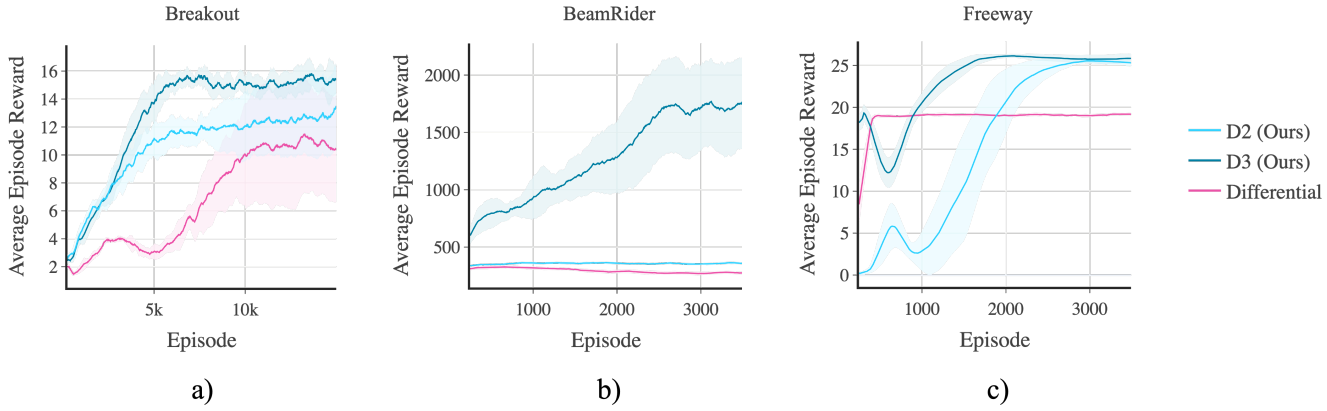

Figure 4: Rolling averages of the total reward per episode when using the D2 and D3 algorithms vs. non-distributional Differential algorithms in the **a)** *Breakout*, **b)** *BeamRider*, and **c)** *Freeway* Atari 2600 environments. A solid line denotes the mean total reward per episode, and the corresponding shaded region denotes a 95% confidence interval over 8 runs.

Q-learning algorithm in the *red-pill blue-pill* environment (Rojas and Lee 2025). As shown in the figure, the agent's quantile estimates converge to the quantiles of the limiting per-step reward distribution induced by the optimal policy, thereby empirically showing that the D2 algorithm not only converges, but converges to the optimal solution.

Subsequently, Figure 3 shows the rolling average-reward as learning progresses when using the D2 and D3 algorithms vs. a non-distributional Differential algorithm in the red-pill blue-pill environment. As shown in the figure, the D2 and D3 algorithms yield competitive performance when compared to their non-distributional counterpart, while also capturing rich information about the long-run reward distribution (i.e., as shown in Figure 2 for the D2 algorithm).

In terms of empirical results from the second group of experiments, Figure 4 shows rolling averages of the total reward per episode as learning progresses when using the D2 and D3 algorithms vs. a non-distributional Differential algorithm in various Atari 2600 games from the *Arcade Learning Environment* (Bellemare et al. 2013). As shown in the figure, the D2 and D3 algorithms consistently outperform the non-distributional baseline, with the D3 algorithm showing better performance than the D2 algorithm.

Our motivation for testing our algorithms in the Arcade Learning Environment (ALE) is twofold. First, ALE is a standard benchmark for evaluating distributional RL algorithms in the discounted setting, thereby making it a potentially-effective tool to use when one aims to measure the relative performance gains of distributional methods over their non-distributional counterparts. Second, there are currently no widely adopted benchmarks for the average-reward setting that match the scale and diversity of ALE. As such, ALE is one of the closest approximations we have for evaluating our algorithms under complex, high-dimensional environments. In this regard, as per Figure 4, we find that our algorithms fare better than the non-distributional baselines in these complex environments that do not strictly satisfy the theoretical assumptions of the average-reward setting.

6 Discussion, Limitations, and Future Work

In this work, we introduced the first distributional RL algorithms specifically designed for the average-reward setting. In particular, we motivated an appropriate distributional objective for the average-reward setting, as well as derived two quantile-based approaches that can learn and/or optimize this objective. We showed, both theoretically and empirically, that these *Differential Distributional* algorithms are able to learn the average-reward-optimal policy, as well as the corresponding (optimal) distributional objective.

In terms of empirical performance, we showed that our algorithms are able to achieve competitive and sometimes superior performance when compared to non-distributional differential algorithms, while also capturing rich information about the long-run reward and return distributions.

Moreover, we note that our choice of distributional objective for the average-reward setting enables the resulting D2 algorithms to be more *scalable* in nature in comparison to distributional RL algorithms in the discounted setting. In particular, the D2 algorithms only require m parameters to parameterize the distributional objective, where m is the number of quantiles. This is in contrast to discounted distributional algorithms, which require $|\mathcal{S}| \times |\mathcal{A}| \times m$ parameters to parameterize the discounted distributional objective. As such, the computational complexity of our D2 algorithms remains constant with respect to the state and action-space sizes, thereby making them more scalable.

In terms of limitations (in the context of the average-reward setting), it remains to be seen how different aspects (beyond the mean) of the chosen distributional objective can be optimized. Similarly, it remains to be seen how a categorical approach could be employed instead of a quantile-based approach. Future work should look to address these limitations, as well as tackle the theoretical aspects of the D3 algorithms. Exploring these directions may yield deeper insights, and in the process, continue unlocking the full potential of distributional approaches in the average-reward setting.

Acknowledgments

We gratefully acknowledge funding from NSERC Discovery Grant # RGPIN-2021-02760. We are also grateful for the computing resources provided by the Digital Research Alliance of Canada. We thank the anonymous AAAI reviewers and area chair for their useful feedback and commentary during the review process.

References

Adamczyk, J.; Makarenko, V.; Tiomkin, S.; and Kulkarni, R. V. 2025. Average-reward soft actor-critic. *Reinforcement Learning Journal*.

Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning*.

Bellemare, M. G.; Dabney, W.; and Rowland, M. 2023. *Distributional Reinforcement Learning*. The MIT Press.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The Arcade Learning Environment: An evaluation platform for general agents. *J. Artif. Intell. Res.*, 253–279.

Dabney, W.; Rowland, M.; Bellemare, M.; and Munos, R. 2018. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*.

Koenker, R. 2005. *Quantile Regression*. Cambridge University Press.

Morimura, T.; Sugiyama, M.; Kashima, H.; Hachiya, H.; and Tanaka, T. 2010. Nonparametric Return Distribution Approximation for Reinforcement Learning. In *Proceedings of the 27th International Conference on Machine Learning*.

Naik, A.; Shariff, R.; Yasui, N.; Yao, H.; and Sutton, R. S. 2019. Discounted reinforcement learning is not an optimization problem. *Optimization Foundations for Reinforcement Learning Workshop at the Conference on Neural Information Processing Systems. Also arXiv:1910.02140*.

Naik, A.; Wan, Y.; Tomar, M.; and Sutton, R. S. 2024. Reward Centering. *Reinforcement Learning Journal*, 4: 1995–2016.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rojas, J. S.; and Lee, C.-G. 2025. Burning RED: Unlocking subtask-driven reinforcement learning and risk-awareness in average-reward Markov decision processes. *Reinforcement Learning Journal*, 6: 431–477.

Rowland, M.; Munos, R.; Azar, M. G.; Tang, Y.; Ostrovski, G.; Harutyunyan, A.; Tuyls, K.; Bellemare, M. G.; and Dabney, W. 2024. An Analysis of Quantile Temporal-Difference Learning. *J. Mach. Learn. Res.*, 25: 1–47.

Sobel, M. J. 1982. The variance of discounted Markov decision processes. *J. Appl. Probab.*, 19(04): 794–802.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction, 2nd edition*. MIT Press.

Wan, Y.; Naik, A.; and Sutton, R. S. 2021. Learning and Planning in Average-Reward Markov Decision Processes. In *Proceedings of the 38th International Conference on Machine Learning*.

Watkins, C. J.; and Dayan, P. 1992. Q-Learning. *Mach. Learn.*, 8: 279–292.

White, D. 1988. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications*, 56(1): 1–29.