Journal of Machine Learning Research 25 (2024) 1-47 Submitted 2/23; Published 5/24

An Analysis of Quantile Temporal-Difference Learning

Mark Rowland MARKROWLAND@GOOGLE.COM
Google DeepMind, London, UK

Rémi Munos MUNOS@GOOGLE.COM
Google DeepMind, Paris, France

Mohammad Gheshlaghi Azar MAZAR@GOOGLE.COM
Google DeepMind, Seattle, USA

Yunhao Tang ROBINTYH@GOOGLE.COM
Google DeepMind, London, UK

Georg Ostrovski OSTROVSKI@GOOGLE.COM
Google DeepMind, London, UK

Anna Harutyunyan HARUTYUNYAN@GOOGLE.COM
Google DeepMind, London, UK

Karl Tuyls KTUYLS@GMAIL.COM
Google DeepMind, Paris, France

Marc G. Bellemare MARC.G.BELLEMARE@GMAIL.COM
Reliant AI & McGill University, Montréal, Canada

Will Dabney WDABNEY@GOOGLE.COM
Google DeepMind, Seattle, USA

Editor: Alexandre Proutiere

Abstract

We analyse quantile temporal-difference learning (QTD), a distributional reinforcement learning algorithm that has proven to be a key component in several successful large-scale applications of reinforcement learning. Despite these empirical successes, a theoretical understanding of QTD has proven elusive until now. Unlike classical TD learning, which can be analysed with standard stochastic approximation tools, QTD updates do not approximate contraction mappings, are highly non-linear, and may have multiple fixed points. The core result of this paper is a proof of convergence to the fixed points of a related family of dynamic programming procedures with probability 1, putting QTD on firm theoretical footing. The proof establishes connections between QTD and non-linear differential inclusions through stochastic approximation theory and non-smooth analysis.

Keywords: Reinforcement learning, temporal-difference learning, distributional reinforcement learning, stochastic approximation, differential inclusion.

1. Introduction

In distributional reinforcement learning, an agent aims to predict the full probability distribution over future returns it will encounter (Morimura et al., 2010b,a; Bellemare et al., 2017, 2023), in contrast to predicting just the mean return, as in classical reinforcement learning (Sutton and Barto, 2018). A widely-used family of algorithms for distributional reinforcement learning is based on the notion of learning *quantiles* of the return distribution, an approach that originated with Dabney et al. (2018b), who introduced the quantile temporal-difference (QTD) learning algorithm. This approach has been particularly successful in combination with deep reinforcement learning, and has been a central component in several recent real-world applications, including sim-to-real stratospheric balloon navigation (Bellemare et al., 2020), robotic manipulation (Bodnar et al., 2020), and algorithm discovery (Fawzi et al., 2022), as well as on benchmark simulated domains such as the Arcade Learning Environment (Bellemare et al., 2013; Machado et al., 2018; Dabney et al., 2018b,a; Yang et al., 2019) and racing simulation (Wurman et al., 2022).

Despite these empirical successes of QTD, little is known about its behaviour from a theoretical viewpoint. In particular, questions regarding the asymptotic behaviour of the algorithm (Do its predictions converge? Under what conditions? What is the qualitative character of the predictions when they do converge?) were left open. A core reason for this is that unlike classical TD, and other distributional reinforcement learning algorithms such as categorical temporal-difference learning (Rowland et al., 2018; Bellemare et al., 2023), the updates of QTD rely on asymmetric L^1 losses. As a result, these updates do not approximate the application of a contraction mapping, are highly non-linear (even in the tabular setting), and also may have multiple fixed points (depending on the exact structure of the reward distributions of the environment), and their analysis requires a distinct set of tools to those typically used to analyse temporal-difference learning algorithms.

In this paper, we prove the convergence of QTD—notably under weaker assumptions than are required in typical proofs of convergence for classical TD learning—establishing it as a sound algorithm with theoretical convergence guarantees, and paving the way for further analysis and investigation. The more general conditions stem from the structure of the QTD updates (namely, their boundedness), and the proof is obtained through the use of stochastic approximation theory with differential inclusions.

We begin by providing background on Markov decision processes, classical TD learning, and quantile regression in Section 2. After motivating the QTD algorithm in Section 3, we describe the related family of quantile dynamic programming (QDP) algorithms, and provide a convergence analysis of these algorithms in Section 4. We then present the main result, a convergence analysis of QTD, in Section 5. The proof relies on the stochastic approximation framework set out by Benaïm et al. (2005), arguing that the QTD algorithm approximates a continuous-time differential inclusion, and then constructing a Lyapunov function to demonstrate that the limiting behaviour of trajectories of the differential inclusion matches that of the QDP algorithms introduced earlier. Finally, in Section 6, we analyse the limit points of QTD, bounding their approximation error to the true return distributions of interest, and investigating the kinds of approximation artefacts that arise empirically.

2. Background

We first introduce background concepts and notation.

2.1 Markov Decision Processes

We consider a Markov decision process specified by finite state and action spaces $\mathcal{X}$ and $\mathcal{A}$, transition kernel $P_{\mathcal{X}} : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X})$, reward distribution function $P_{\mathcal{R}} : \mathcal{X} \times \mathcal{A} \to \mathscr{P}_1(\mathbb{R})$, and discount factor $\gamma \in [0, 1)$. Here, $\mathscr{P}(\mathcal{X})$ is the set of probability distributions over the finite set $\mathcal{X}$, and $\mathscr{P}_1(\mathbb{R})$ is the set of probability distributions over $\mathbb{R}$ (with its usual Borel σ-algebra) with finite mean.

Given a policy $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ and an initial state $x_0 \in \mathcal{X}$, an agent interacting with the environment using the policy π generates a sequence of states, actions and rewards $(X_t, A_t, R_t)_{t=0}^{\infty}$, called a *trajectory*, the joint distribution of which is determined by the transition dynamics and reward distributions of the environment, and the policy of the agent. More precisely, we have

- $X_0 = x_0$, and for each $t \geq 0$:
- $A_t \mid X_{0:t}, A_{0:t-1}, R_{0:t-1} \sim \pi(\cdot|X_t)$;
- $R_t \mid X_{0:t}, A_{0:t}, R_{0:t-1} \sim P_{\mathcal{R}}(\cdot|X_t, A_t)$;
- $X_{t+1} \mid X_{0:t}, A_{0:t}, R_{0:t} \sim P_{\mathcal{X}}(\cdot|X_t, A_t)$.

The distribution of the trajectory is thus parametrised by the initial state x_0, and the policy π. To illustrate this dependency, we use the notation $\mathbb{P}_{x_0}^{\pi}$ and $\mathbb{E}_{x_0}^{\pi}$ to denote the probability distribution and expectation operator corresponding to this distribution, and will write $P^{\pi}(\cdot|x)$ for the joint distribution over a reward–next-state pair when the current state is x.

2.2 Predicting Expected Returns and the Return Distribution

The quality of the agent's performance on the trajectory is quantified by the *discounted return*, or simply the *return*, given by

$$\sum_{t=0}^{\infty} \gamma^t R_t \,. \tag{1}$$

The return is a random variable, whose sources of randomness are the random selections of actions made according to π, the randomness in state transitions, and the randomness in rewards observed. Typically in reinforcement learning, a single scalar summary of performance is given by the expectation of this return over all these sources of randomness. For a given policy, this is summarised across each possible starting state via the *value function* $V^{\pi} : \mathcal{X} \to \mathbb{R}$, defined by

$$V^{\pi}(x) = \mathbb{E}_x^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]. \tag{2}$$

Learning the value function of a policy π from sampled trajectories generated through interaction with the environment is a central problem in reinforcement learning, referred to as the *policy evaluation task*.

Each expected return is a scalar summary of a much more rich, complex object: the probability distributions of the random return in Equation (1) itself. *Distributional reinforcement learning* (Bellemare et al., 2023) is concerned with the problem of learning to predict the *probability distribution* over returns, in contrast to just their expected value. Mathematically, the goal is to learn the return-distribution function $\eta^\pi : \mathcal{X} \to \mathscr{P}(\mathbb{R})$; for each state $x \in \mathcal{X}$, $\eta^\pi(x)$ is the probability distribution of the random return in Expression (1) when the trajectory begins at state x, and the agent acts using policy π. Mathematically, we have

$$\eta^\pi(x) = \mathcal{D}_x^\pi\left(\sum_{t\geq 0}\gamma^t R_t\right),$$

where $\mathcal{D}_x^\pi$ extract the probability distribution of a random variable under $\mathbb{P}_x^\pi$.

There are several distinct motivations for aiming to learn these more complex objects. First, the richness of the distribution provides an abundance of signal for an agent to learn from, in contrast to a single scalar expectation. The strong performance of deep reinforcement learning agents that incorporate distributional predictions is hypothesised to be related to this fact (Dabney et al., 2018b; Barth-Maron et al., 2018; Dabney et al., 2018a; Yang et al., 2019). Second, learning about the full probability distribution of returns makes possible the use of *risk-sensitive* performance criteria; one may be interested in not only the expected return under a policy, but also the variance of the return, or the probability of the return being under a certain threshold.

Unlike the value function V^π, which is an element of $\mathbb{R}^\mathcal{X}$, and can therefore be straightforwardly represented on a computer (up to floating-point precision), the return-distribution function η^π is not representable. Each object $\eta^\pi(x)$ is a probability distribution over the real numbers, and, informally speaking, probability distributions have infinitely many degrees of freedom. Distributional reinforcement learning algorithms therefore typically work with a subset of distributions that *are* amenable to parametrisation on a computer (Bellemare et al., 2023). Common choices of subsets include categorical distributions (Bellemare et al., 2017), exponential families (Morimura et al., 2010b), and mixtures of Gaussian distributions (Barth-Maron et al., 2018). Quantile temporal-difference learning, the core algorithm of study in this paper, aims to learn a particular set of quantiles of the return distribution, as described in Section 3.

2.3 Monte Carlo and Temporal-Difference Learning

To foreshadow our description and motivation of quantile temporal-difference learning, we recall a line of thinking that interprets the classical TD learning update rule as an approximation to Monte Carlo learning; this material is common to many introductory texts on reinforcement learning (Sutton and Barto, 2018), and we present it here to make a direct analogy with QTD. First, we may observe that, under the condition that all reward distributions have finite variance, $V^\pi(x)$ is the unique minimiser of the following loss function

over $u \in \mathbb{R}$, the prediction of mean return at x:

$$\mathcal{L}_x^\pi(u) = \frac{1}{2}\mathbb{E}_x^\pi\Big[\Big(u - \sum_{t=0}^{\infty}\gamma^t R_t\Big)^2\Big].$$

This well-known characterisation of the expectation of a random variable is readily verified by, for example, observing that the loss is convex and differentiable in u, and solving the equation $\partial_u \mathcal{L}_x^\pi(u) = 0$. This motivates an approach to learning $V^\pi(x)$ based on stochastic gradient descent on the loss function $\mathcal{L}_x^\pi$. We maintain an estimate $V \in \mathbb{R}^{\mathcal{X}}$ of the value function, and each time a trajectory $(X_t, A_t, R_t)_{t\geq 0}$ beginning at state x is observed, we can obtain an unbiased estimator of the negative gradient of $\mathcal{L}_x^\pi(V(x))$ as

$$\sum_{t=0}^{\infty}\gamma^t R_t - V(x)\,,$$

and update $V(x)$ by taking a step in the direction of this negative gradient, with some step size α:

$$V(x) \leftarrow V(x) + \alpha\Big(\sum_{t=0}^{\infty}\gamma^t R_t - V(x)\Big). \tag{3}$$

This is a *Monte Carlo* algorithm, so called because it uses Monte Carlo samples of the random return to update the estimate V.

A popular alternative to this Monte Carlo algorithm is temporal-difference learning, which replaces samples from the random return with a *bootstrapped* approximation to the return, obtained from a transition (x, A, R, X') by combining the immediate reward R with the current estimate of the expected return obtained at X', resulting in the return estimate

$$R + \gamma V(X')\,, \tag{4}$$

and the corresponding update rule

$$V(x) \leftarrow V(x) + \alpha(R + \gamma V(X') - V(x))\,. \tag{5}$$

While the mean-return estimator in Expression (4) is generally *biased*, since $V(X')$ is not generally equal to the true expected return $V^\pi(X')$, it is often a lower-variance estimate, since we are replacing the *random* return from X' with an estimate of its expectation (Sutton, 1988; Sutton and Barto, 2018; Kearns and Singh, 2000).

This motivates the TD learning rule in Expression (5) based on the Monte Carlo update rule in Expression (3), with the understanding that this algorithm can be applied more generally, with access only to sampled transitions (rather than full trajectories), and may result in more accurate estimates of the value function, due to lower-variance updates, and the propensity of TD algorithms to "share information" across states. Note however that this does not *prove* anything about the behaviour of temporal-difference learning, and a fully rigorous theory of the asymptotic behaviour emerged several years after TD methods were formally introduced (Sutton, 1984, 1988; Watkins, 1989; Watkins and Dayan, 1992; Dayan, 1992; Dayan and Sejnowski, 1994; Jaakkola et al., 1994; Tsitsiklis, 1994).

3. Quantile Temporal-Difference Learning and Quantile Dynamic Programming

We now present the main algorithms of study in this paper.

3.1 Quantile Regression

To motivate QTD, we begin by considering how we might adapt a Monte Carlo algorithm such as that in Expression (3) to learn about the distribution of returns, rather than just their expected value. We cannot learn the return distribution in its entirety with a finite collection of parameters; the space of return distributions is infinite-dimensional, so we must instead be satisfied with learning an approximation of the return distribution by selecting a *probability distribution representation* (Bellemare et al., 2023, Chapter 5): a subset of probability distributions parametrised by a finite-dimensional set of parameters. The approach of quantile temporal-difference learning is to learn an approximation of the form

$$\eta(x) = \sum_{i=1}^{m} \frac{1}{m} \delta_{\theta(x,i)}\,; \tag{6}$$

an equally-weighted mixture of Dirac deltas, for each state $x \in \mathcal{X}$. The quantile-based approach to distributional reinforcement learning aims to have the particle locations $(\theta(x,i))_{i=1}^{m}$ approximate certain *quantiles* of $\eta^{\pi}(x)$.

Definition 1 *For a probability distribution $\nu \in \mathscr{P}(\mathbb{R})$ and parameter $\tau \in (0,1)$, the set of τ-quantiles of ν is given by the set*

$$\{z \in \mathbb{R} : F_\nu(z) = \tau\} \cup \inf\{y \in \mathbb{R} : F_\nu(y) > \tau\}\,,$$

where $F_\nu : \mathbb{R} \to [0,1]$ is the CDF of ν, defined by $F_\nu(t) = \mathbb{P}_{Z\sim\nu}(Z \leq t)$ for all $t \in \mathbb{R}$.

Expanding on this definition, if the set $\{z \in \mathbb{R} : F_\nu(z) = \tau\}$ is non-empty, then the τ-quantiles are precisely the values z such that $\mathbb{P}_{Z\sim\nu}(Z \leq z) = \tau$. If however this set is empty (which may arise when F_ν has points of discontinuity), then the quantile is the smallest value y such that $\mathbb{P}_{Z\sim\nu}(Z \leq y) > \tau$. Note also that if F_ν is strictly increasing, this guarantees uniqueness of each τ-quantile for $\tau \in (0,1)$; this is often a useful property in the analysis we consider later. These different cases are illustrated in Figure 1. The generalised inverse CDF of ν, $F_\nu^{-1} : (0,1) \to \mathbb{R}$, is defined by

$$F_\nu^{-1}(\tau) = \inf\{y : F_\nu(y) \geq \tau\}\,,$$

and provides a way of uniquely specifying a quantile for each level τ. In cases where there is not a unique τ-quantile (see Figure 1), $F_\nu^{-1}(\tau)$ corresponds to the *left-most* or *least* valid τ-quantile. We also introduce the notation

$$\bar{F}_\nu^{-1}(\tau) = \inf\{y : F_\nu(y) > \tau\}\,,$$

which corresponds to the *right-most* or *greatest* τ-quantile; notice the strict inequality that appears in the definition, in contrast to that of $F_\nu^{-1}(\tau)$. If F_ν^{-1} is continuous at τ, then

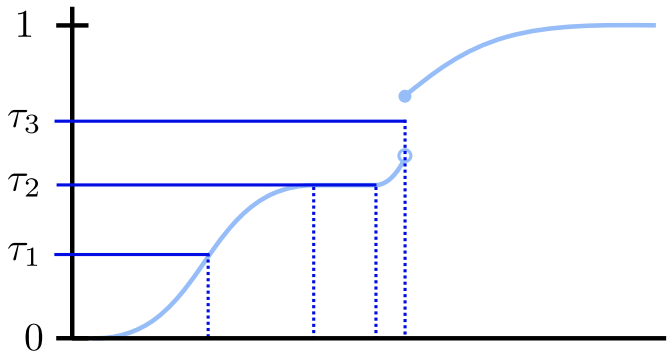

Figure 1: The three distinct scenarios that arise in defining quantiles. Firstly, there is a value z_1 for which $F_\nu(z_1) = \tau_1$ and at which F_ν is strictly increasing. Therefore z_1 is the unique τ_1-quantile of ν. Next, there is an interval $[z_2, z_2']$ on which F_ν equals τ_2, therefore all elements in this interval are τ_2-quantiles of ν. Finally, there is no value z such that $F_\nu(z) = \tau_3$, and the unique τ_3-quantile is therefore defined by the infimum part of the definition.

$F_\nu^{-1}(\tau) = \bar{F}_\nu^{-1}(\tau)$, as is the case for $\tau = \tau_1$ and $\tau = \tau_3$ in Figure 1. However, if F_ν has a flat region for the value τ (as is the case for $\tau = \tau_2$ in Figure 1), then $F_\nu^{-1}(\tau)$ and $\bar{F}_\nu^{-1}(\tau)$ are distinct, and correspond to the boundary points of this flat region.

Algorithmically, we aim for $\theta(x, i)$ to approximate a τ_i-quantile of $\eta^\pi(x)$, where $\tau_i = {}^{2i-1}/_{2m}$. To build learning algorithms that achieve this, we require an incremental algorithm that updates $\theta(x, i)$ in response to samples from the target distribution $\eta^\pi(x)$, which converges to a ${}^{2i-1}/_{2m}$-quantile of $\eta^\pi(x)$.

Such an approach is available by using the quantile regression loss. We define the quantile regression loss associated with distribution $\nu \in \mathscr{P}(\mathbb{R})$ and quantile level $\tau \in (0, 1)$ as a function of v by

$$\mathbb{E}_{Z\sim\nu}[(\tau \mathbb{1}\{Z \geq v\} + (1-\tau)\mathbb{1}\{Z < v\})|Z - v|]\,. \tag{7}$$

This loss is the expectation of an asymmetric absolute value loss, in which positive and negative errors are weighted according to the parameters τ and $1 - \tau$ respectively. Just as the expected squared loss encountered above encodes the mean as its unique minimiser, the quantile regression loss encodes the τ-quantiles of ν as the unique minimisers; see, for example, Koenker (2005) for further background. Thus, applying the quantile regression loss to the problem of estimating τ-quantiles of the return distribution, we arrive at the loss

$$\mathcal{L}_x^{\tau,\pi}(v) = \mathbb{E}_x^\pi\Big[(\tau \mathbb{1}\{\Delta \geq 0\} + (1-\tau)\mathbb{1}\{\Delta < 0\})|\Delta|\Big]\,, \quad \text{where } \Delta = \sum_{t=0}^{\infty} \gamma^t R_t - v\,.$$

Given an observed return $\sum_{t\geq 0}\gamma^t R_t$ from the state x, we therefore have that an unbiased estimator of the negative gradient[1] of this loss is

$$\tau\mathbb{1}\Big\{\sum_{t=0}^{\infty}\gamma^t R_t \geq v\Big\} - (1-\tau)\mathbb{1}\Big\{\sum_{t=0}^{\infty}\gamma^t R_t < v\Big\},$$

which motivates an update rule of the form

$$\theta(x,i) \leftarrow \theta(x,i) + \alpha\Big(\tau_i\mathbb{1}\Big\{\sum_{t=0}^{\infty}\gamma^t R_t \geq \theta(x,i)\Big\} - (1-\tau_i)\mathbb{1}\Big\{\sum_{t=0}^{\infty}\gamma^t R_t < \theta(x,i)\Big\}\Big). \tag{8}$$

This can be rewritten as

$$\theta(x,i) \leftarrow \theta(x,i) + \alpha\Big(\tau_i - \mathbb{1}\Big\{\sum_{t=0}^{\infty}\gamma^t R_t < \theta(x,i)\Big\}\Big). \tag{9}$$

This is essentially the application of the stochastic gradient descent method for quantile regression to learning quantiles of the return distribution.

3.2 Quantile Temporal-Difference Learning

We can motivate and describe the quantile temporal-difference learning algorithm (Dabney et al., 2018b; Bellemare et al., 2023) by modifying the Monte Carlo algorithm in Expression (8) in a similar manner to the modification that led to the TD algorithm in Expression (5). We replace the Monte Carlo return

$$\sum_{t=0}^{\infty}\gamma^t R_t$$

based on a full trajectory, with an approximate sample from the return distribution derived from an observed transition (x, R, X'), and the estimate $\eta(X')$ of the return distribution at state X'. If the return distribution estimate $\eta(X')$ takes the form given in Equation (6), as is the case for the probability distribution representation considered here, then such a sample return is obtained as

$$R + \gamma\theta(X', J),$$

with J sampled uniformly from $\{1,\ldots,m\}$. This yields the update rule

$$\theta(x,i) \leftarrow \theta(x,i) + \alpha\Big(\tau_i - \mathbb{1}\Big\{R + \gamma\theta(X',J) < \theta(x,i)\Big\}\Big).$$

We can consider also a variance-reduced version of this update, in which we average over updates performed under different realisations of J, leading to the update

$$\theta(x,i) \leftarrow \theta(x,i) + \frac{\alpha}{m}\sum_{j=1}^{m}\Big(\tau_i - \mathbb{1}\Big\{R + \gamma\theta(X',j) < \theta(x,i)\Big\}\Big). \tag{10}$$

1. Technically speaking, we are assuming that differentiation and expectation can be interchanged here. Further, under certain circumstances the loss is only *sub*-differentiable. As our principal goal in this section is to provide intuition for QTD, we do not comment further on these technical details here. The convergence results later in the paper deal with these issues carefully.

This is precisely the quantile temporal-difference learning update, presented in Algorithm 1 below, which underlies many recent successful applications of reinforcement learning at scale (Dabney et al., 2018b,a; Yang et al., 2019; Bellemare et al., 2020; Wurman et al., 2022; Fawzi et al., 2022). Similar to other temporal-difference learning algorithms, QTD updates its parameters $((\theta(x,i))_{i=1}^m : x \in \mathcal{X})$ on the basis of sample transitions (x, r, x') generated through interaction with the environment via the policy π, comprising a state, reward, and next state.

Algorithm 1 QTD update

Require: Quantile estimates $\theta \in \mathbb{R}^{\mathcal{X}\times[m]}$,
Observed transition (x, r, x') ,
Learning rate α.

Set $\tau_i = \frac{2i-1}{2m}$ for each $i = 1, \ldots, m$.
for $i = 1, \ldots, m$ **do**
  Set $\theta'(x,i) \leftarrow \theta(x,i) + \alpha\frac{1}{m}\sum_{j=1}^m \left[\tau_i - \mathbb{1}\left\{r + \gamma\theta(x',j) - \theta(x,i) < 0\right\}\right]$
end for
for $i = 1, \ldots, m$ **do**
  Set $\theta(x,i) \leftarrow \theta'(x,i)$
end for
return $((\theta'(x,i))_{i=1}^m : x \in \mathcal{X})$

Whilst the QTD update makes use of temporal-difference errors $r + \gamma\theta(x',j) - \theta(x,i)$, there are two key differences to the use of analogous quantities in classical TD learning. First, the TD errors influence the update only through their sign, not their magnitude. Second, the predictions at each state $(\theta(x,i))_{i=1}^m$ are indexed by i, and each update includes a distinct term τ_i (equal to $2i-1/2m$). The presence of these terms causes the learnt parameters to make distinct predictions, as described in Section 3.1. Practical implementations of QTD use these precise values for τ_i, equally spaced out on $[0, 1]$, as proposed by Dabney et al. (2018b). Much of the analysis in this paper goes through straightforwardly for other values of τ_i, though we will see in Section 6 that this choice is well motivated in that it provides the best bounds on distribution approximation. The tabular QTD algorithm as described in Algorithm 1 uses a factor $O(m)$ times more memory than an analogous classical TD algorithm, owing to the need to store multiple predictions at each state, though the scaling with the size of the state space is the same as for classical TD. For further discussion of the computational complexity of QTD, see Rowland et al. (Appendix A.3; 2023).

The discussion above provides *motivation* for the form of the QTD update given in Algorithm 1, and intuition as to why this algorithm might perform reasonably, and learn a sensible approximation to the return distribution. However, it stops short of providing an explanation of how the algorithm should be expected to behave, or providing any theoretical guarantees as to what the algorithm will in fact converge to. A core goal of the sections that follow is to answer these questions, and put QTD on firm theoretical footing.

3.3 Motivating Examples

Before undertaking an analysis of QTD, we pause to provide several numerical examples of its behaviour in small environments. These examples provide further intuition for the characteristics of the algorithm, illustrate the breadth of qualitative behaviours it can exhibit, and provide motivation for the kinds of theoretical questions we might hope to answer.

Example 2 *Consider the chain MDP illustrated at the top of Figure 2. The random return at each state is a sum of independent Gaussian random variables, and hence the return distribution at each state is Gaussian. The centre plot in Figure 2 illustrates the evolution of $m = 5$ quantile estimates learnt by QTD, using a constant learning rate of 0.01, and updating all states at each update. The estimated quantile values eventually settle after around 6,000 updates, with small oscillations around this point. The bottom of Figure 2 compares the true return distribution at each state (in blue), with the approximation learnt by QTD (in black), and the approximation obtained with the true value of the five quantiles of interest (grey). The behaviour of QTD in this case raises several questions: Can it be shown that QTD is guaranteed to stabilise/converge around a certain point? Can a guarantee be given on the quality of the approximate distributions learnt by QTD?*

Example 3 *For a different perspective on the behaviour of QTD, consider a two-state MDP with transition dynamics as illustrated in the top-left of Figure 3, and discount factor $\gamma = 0.5$. The reward obtained when transitioning from state x_1 is distributed as N(2, 1), and the reward obtained when transitioning from state x_2 is distributed as N(-1, 1); here, we write $N(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2. We consider the case of learning a single quantile (the median) at each of these two states, taking $m = 1$; this allows us to plot the full phase space of the QTD algorithm in a two-dimensional plot.*

The top-right of Figure 3 shows a path taken by QTD under this MDP. In addition, the streamplot illustrates the direction of the expected *update that QTD undertakes at each point in phase space. We empirically observe convergence of the algorithm to a point. Additionally, the expected update direction changes smoothly; the result is a vector field that appears to point towards the point of convergence from all directions.*

The bottom-left of Figure 3 shows a path taken by QTD under a modified version of the MDP, in which the reward distributions N(2, 1) and N(-1, 1) are replaced with δ_2 and δ_{-1}, respectively. We observe that the algorithm still converges to a point, although the vector field of expected update directions is now piecewise constant, with discontinuities along several lines. This behaviour is typical of QTD; the less 'smooth' the reward distributions in the MDP, the more abrupt the changes in behaviour we typically observe with QTD.

Finally, we consider a modified version of the MDP in which all transition probabilities are $1/2$, rewards from state x_1 are always 2, and rewards from state x_2 are always -1. In this case, QTD no longer appears to converge to a point, but instead converges to the set bounded by the four grey lines appearing in the bottom-right of Figure 3, and subsequently performing a random walk over this set. This collection of examples illustrates that QTD can exhibit a fairly wide family of behaviours depending on the characteristics of the environment. In particular, non-uniqueness of quantiles in reward distributions (corresponding to flat regions in reward distribution CDFs) can lead to multiple possible limit points, and

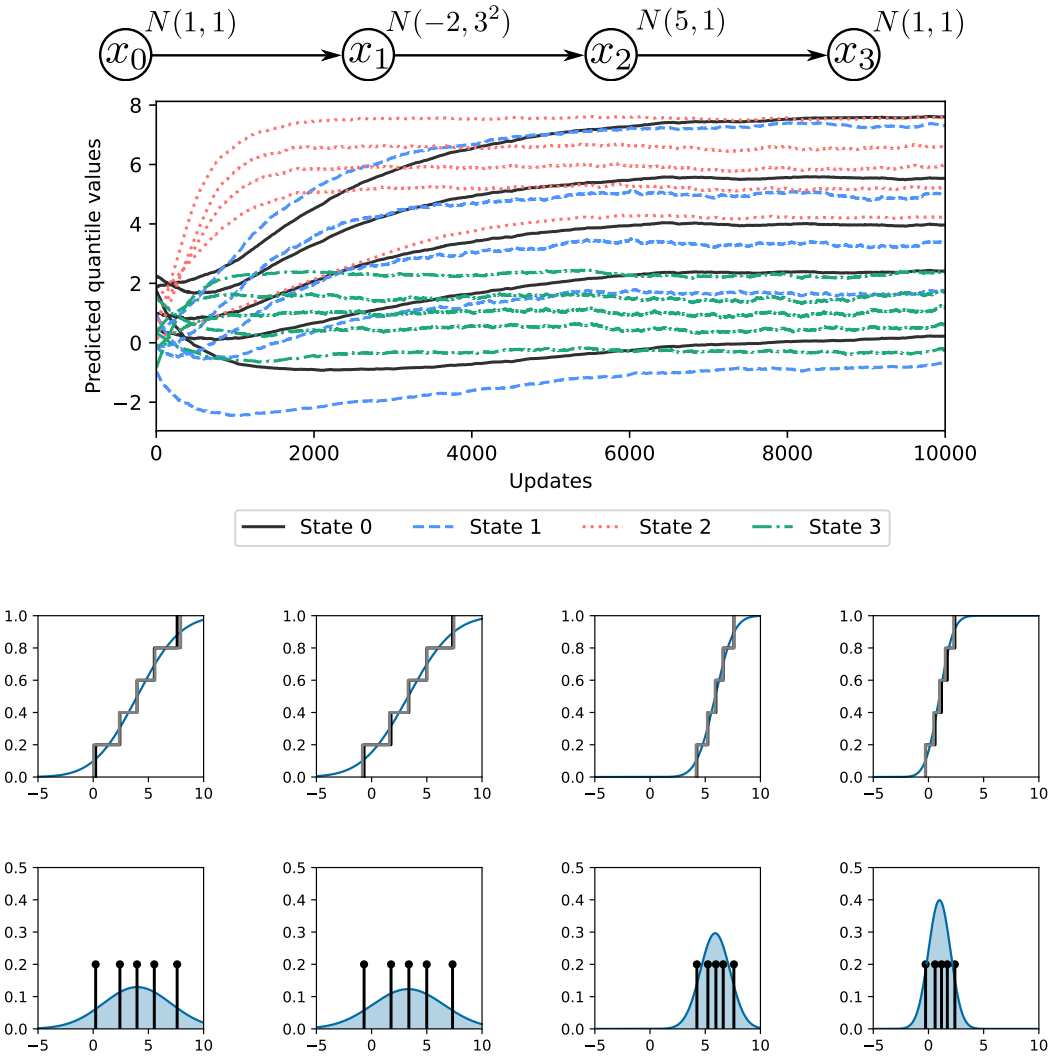

Figure 2: Top: A chain MDP with four states. Each transition yields a normally-distributed reward; from x_3, the episode ends. The discount factor is $\gamma = 0.9$. Centre-top: The progress of QTD, run with $m = 5$ quantiles, over the course of 10,000 updates. The vertical axis corresponds to the predicted quantile values. Centre-bottom: The true CDF of the return distribution (blue) at each state, along with the final estimate produced by QTD (black), and the approximation produced by the quantiles of the return distribution (grey). Bottom: The PDF of the return distribution (blue) at each state, along with the final quantile approximation produced by QTD (black).

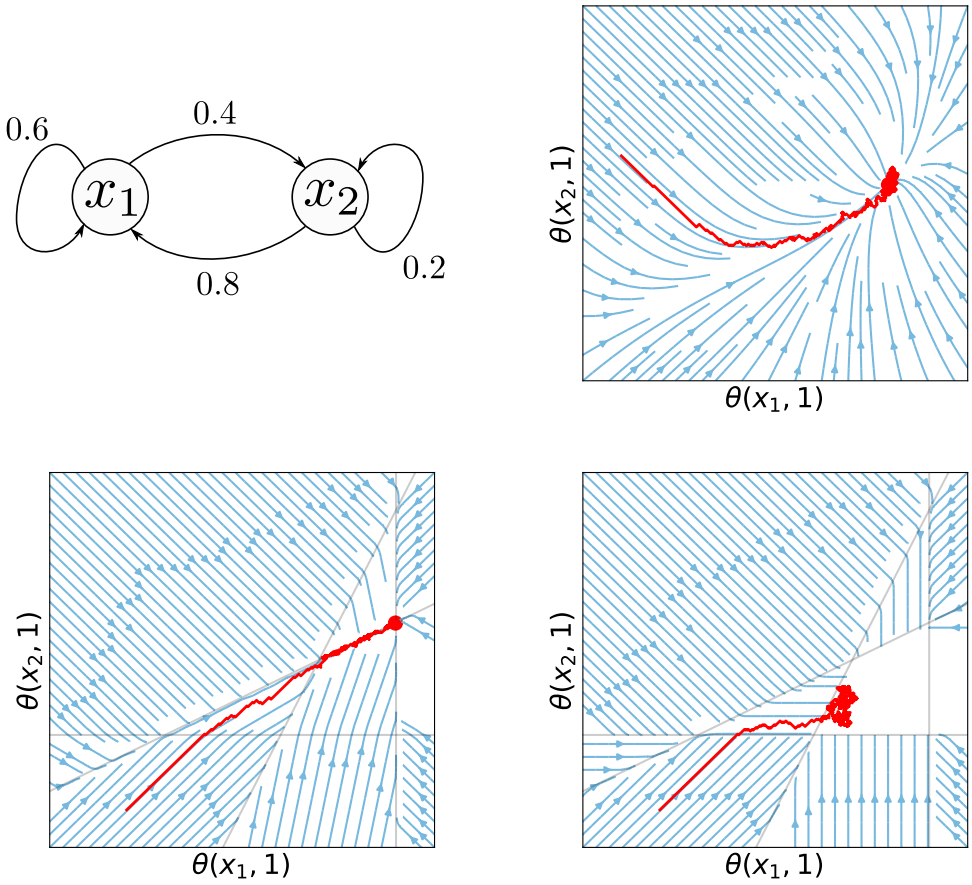

Figure 3: Top left: The example Markov decision process described in Example 3. Top right: Example dynamics of QTD with $m = 1$ in this environment, when reward distributions are Gaussian. Also included are the directions of expected update, in blue. Bottom left: Example dynamics and expected update directions when reward distributions are Dirac deltas. Bottom right: Example dynamics and expected updates with modified environment transition probabilities.

discontinuities in reward distributions can lead to discontinuous changes in expected updates; by contrast, reward distributions that are absolutely continuous lead to smooth changes in expected dynamics.

3.4 Quantile Dynamic Programming

Recall the QTD update given in Equation (10). As described in Section 3.2, this update serves, on average, to move $\theta(x, i)$ in the direction of the τ_i-quantiles of the distribution of the random variable $R + \theta(X', J)$, where (x, R, X') is a random transition generated by interacting with the environment using π, and $J \sim \text{Unif}(\{1, \ldots, m\})$.

Suppose we were able to update $\theta(x, i)$ not just with a single gradient step in this direction, but instead were able to update it to take on exactly this quantile value. This motivates a

dynamic programming alternative to QTD, *quantile dynamic programming* (QDP), which directly calculates these quantiles iteratively, in a similar manner to iterative policy evaluation in classical reinforcement learning (Bertsekas and Tsitsiklis, 1996).

The mathematical structure of such an algorithm is given in Algorithm 2. This stops short of being an implementable algorithm, since we do not describe in what format the transition probabilities and reward distributions are available, which are required to evaluate the inverse CDFs that arise in the algorithm. However, for MDPs in which transition probabilities and reward distributions are available, QDP is an algorithmic framework of interest in its own right, and to this end we provide several concrete implementations in Appendix B.

The QDP template in Algorithm 2 is parametrised by the interpolation parameters $\lambda \in [0,1]^{\mathcal{X}\times[m]}$. These parameters control exactly which quantile is chosen when the desired quantile level τ_i corresponds to a flat region of the CDF for the distribution ν (the second case in Figure 1). QDP was originally presented by Bellemare et al. (2023) in the case $\lambda(x,i) \equiv 0$; the presentation here generalises QDP to a family of algorithms, parametrised by λ.

Our interest in QDP stems from the fact that QTD can be viewed as approximating the behaviour of the QDP algorithms, without requiring access to the transition structure and reward distributions of the environment. In particular, we will show that under appropriate conditions, the asymptotic behaviour of QTD and QDP are equivalent: they both converge to the same limiting points. Figure 4 illustrates the behaviour of the QDP algorithm in the environment described in Example 3; since the reward distributions in this example have strictly increasing CDFs, QDP behaves identically for all choices of interpolation parameters λ. QTD and QDP appear to have the same asymptotic behaviour, converging to the same limiting point. In cases where QTD appears to converge to a set, such as in the bottom-right plot of Figure 3, the relationship is slightly more complicated, and there is a correspondence between the asymptotic behaviour of QTD and the family of dynamic programming algorithms parametrised by λ, as illustrated at the bottom of Figure 4. Thus, to understand the asymptotic behaviour of QTD, we begin by analysing the asymptotic behaviour of QDP.

Algorithm 2 Quantile dynamic programming

Require: Quantile estimates $((\theta(x,i))_{i=1}^m : x \in \mathcal{X})$,
Interpolation parameters $\lambda \in [0,1]^{\mathcal{X}\times[m]}$.

for $x \in \mathcal{X}$ **do**
  Let (x, R, X') be a random transition under π, and $J \sim \text{Unif}(\{1,\dots,m\})$.
  Set ν to be the distribution of $R + \gamma\theta(X', J)$.
  for $i = 1,\dots,m$ **do**
    Set $\theta(x,i) \leftarrow (1-\lambda(x,i))F_\nu^{-1}(\tau_i) + \lambda(x,i)\bar{F}_\nu^{-1}(\tau_i)$.
  end for
end for
return $((\theta(x,i))_{i=1}^m : x \in \mathcal{X})$

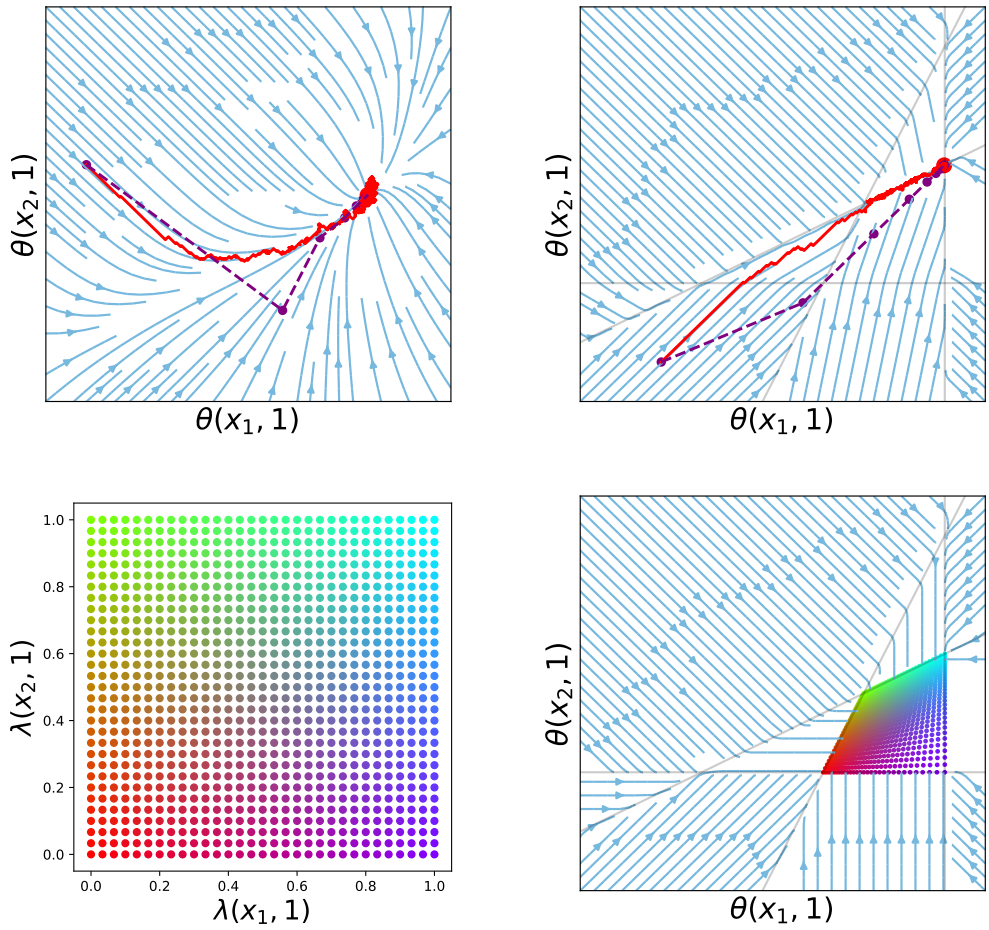

Figure 4: Top left: Illustration of QDP (dashed purple) and QTD (solid red) on the first MDP from Example 3, with Gaussian rewards. Top right: Illustration of QDP and QTD on the second MDP from Example 3, with deterministic rewards. Bottom: Values of λ and corresponding fixed points of QDP in the final MDP from Example 3.

4. Convergence of Quantile Dynamic Programming

We can decompose the update QDP performs into the composition of several operators. Algorithm 2 manipulates tables of the form $((\theta(x,i))_{i=1}^m : x \in \mathcal{X})$. For a given state x, the vector $(\theta(x,i))_{i=1}^m$ represents the estimated ${}^{2i-1}/_{2m}$-quantiles of the return distribution at state x, for $i = 1, \ldots, m$. In mathematically analysing the algorithm, it is useful to be able to refer to the distribution encoded by these quantiles:

$$\frac{1}{m}\sum_{i=1}^{m} \delta_{\theta(x,i)}\,, \tag{11}$$

and reason about the transformations undertaken by Algorithm 2 directly in terms of distributions. To this end, if we write $\eta(x) \in \mathscr{P}(\mathbb{R})$ for the probability distribution associated with the quantile estimates $(\theta(x,i))_{i=1}^m$, we can interpret the transformation performed by Algorithm 2 as comprising two parts, which we now describe in turn.

First, the variable $\eta(x)$ is assigned the distribution of $R + \gamma G(X')$, where R, X' are the random reward and next-state encountered from the initial state x with policy π, and $(G(y) : y \in \mathcal{X})$ is an independent collection of random variables, with each $G(y)$ distributed according to $\eta(y)$.

We write $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{X}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}}$ for this transformation. The function $\mathcal{T}^\pi$ is known as the *distributional Bellman operator* (Bellemare et al., 2017; Rowland et al., 2018; Bellemare et al., 2023). In terms of the above definition via distributions of random variables, $\mathcal{T}^\pi$ can be written

$$(\mathcal{T}^\pi \eta)(x) = \mathcal{D}_\pi(R + \gamma G(X'))\,,$$

where (x, R, X') is a random environment transition beginning at x, independent of $(G(y) : y \in \mathcal{X})$, and $\mathcal{D}_\pi$ extracts the distribution of its argument when (x, R, X') is generated by sampling an action from π. See Bellemare et al. (2023) for further background on the distributional Bellman operator.

In general, $\mathcal{T}^\pi \eta$ may comprise much more complicated distributions than η itself, with many more atoms, or possibly infinite support, if reward distributions are infinitely-supported. Algorithm 2 does not return these full transformed distributions, but rather approximations, or *projections*, of these distributions, obtained by keeping only information about certain quantiles (in the inner for-loop of Algorithm 2); this is the second distribution transformation the algorithm undertakes. Each choice of interpolation parameters λ corresponds to a different projection operator, denoted $\Pi^\lambda : \mathscr{P}(\mathbb{R})^{\mathcal{X}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}}$, and defined by

$$(\Pi^\lambda \eta)(x) = \frac{1}{m}\sum_{i=1}^{m} \delta_{(1-\lambda(x,i))F^{-1}_{\eta(x)}(\tau_i) + \lambda(x,i)\bar{F}^{-1}_{\eta(x)}(\tau_i)}\,. \tag{12}$$

Thus, the composition $\Pi^\lambda \mathcal{T}^\pi$, the projected distributional Bellman operator, is a transformation on the space of return-distribution functions $\mathscr{P}(\mathbb{R})^{\mathcal{X}}$. We will also find it useful to abuse notation slightly and consider $\Pi^\lambda \mathcal{T}^\pi$ as an operator on the space $\mathbb{R}^{\mathcal{X}\times[m]}$ of parameters that QDP and QTD operate over. The understanding is that an input $\theta \in \mathbb{R}^{\mathcal{X}\times[m]}$ is

first re-interpreted as a collection of distributions as in Expression (11), with $\Pi^\lambda \mathcal{T}^\pi$ applied as defined above to this collection of probability distributions, and then finally extracting the support of the output distributions, which take the form

$$\sum_{i=1}^{m} \frac{1}{m} \delta_{z_i} ,$$

to return an element of $\mathbb{R}^{\mathcal{X} \times [m]}$. We will also write $\mathcal{T}^\pi \theta$ for the element of $\mathscr{P}(\mathbb{R})^{\mathcal{X}}$ obtained by applying $\mathcal{T}^\pi$ to the distributions $(\eta(x) : x \in \mathcal{X})$ defined by

$$\eta(x) = \sum_{i=1}^{m} \frac{1}{m} \delta_{\theta(x,i)} .$$

Remark 4 *This convention highlights that there are two complementary views of distributional reinforcement learning algorithms, through finite-dimensional sets of parameters, and through probability distributions. The view in terms of probability distributions is often useful in contraction analysis, and in measuring approximation error, while we will see that the parameter view is key to the stochastic approximation analysis that follows, and is ultimately the way in which these algorithms are implemented.*

With this convention, $\Pi^\lambda \mathcal{T}^\pi \theta$ is precisely the table θ' output by Algorithm 2 on input θ, and so the QDP algorithm is mathematically equivalent to repeated application of the operator $\Pi^\lambda \mathcal{T}^\pi$ to an initial collection of quantile estimates. To understand the long-term behaviour of QDP, we can therefore seek to understand this projected operator $\Pi^\lambda \mathcal{T}^\pi$.

4.1 Convergence Analysis

We will show that $\Pi^\lambda \mathcal{T}^\pi$ is a contraction mapping with respect to an appropriate metric over return-distribution functions. Building on the analysis in the case of $\lambda \equiv 0$ carried out by Dabney et al. (2018b) and Bellemare et al. (2023), we use the Wasserstein-∞ metric $w_\infty : \mathscr{P}(\mathbb{R}) \times \mathscr{P}(\mathbb{R}) \to [0, \infty]$, defined by

$$w_\infty(\nu, \nu') = \sup_{t \in (0,1)} |F_\nu^{-1}(t) - F_{\nu'}^{-1}(t)| ,$$

and its extension to return-distribution functions, $\bar{w}_\infty : \mathscr{P}(\mathbb{R})^{\mathcal{X}} \times \mathscr{P}(\mathbb{R})^{\mathcal{X}} \to [0, \infty]$, given by

$$\bar{w}_\infty(\eta, \eta') = \max_{x \in \mathcal{X}} \sup_{t \in (0,1)} |F_{\eta(x)}^{-1}(t) - F_{\eta'(x)}^{-1}(t)| .$$

Both w_∞ and $\bar{w}_\infty$ fulfil all the requirements of a metric, except that they may assign infinite distances (Villani, 2009; see also Bellemare et al., 2023 for a detailed discussion specifically in the context of reinforcement learning). We must therefore take some care as to when distances are finite. The following is established by Bellemare et al. (2023, Proposition 4.15).

Proposition 5 *The distributional Bellman operator $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{X}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}}$ is a γ-contraction with respect to $\bar{w}_\infty$. That is,*

$$\bar{w}_\infty(\mathcal{T}^\pi \eta, \mathcal{T}^\pi \eta') \leq \gamma \bar{w}_\infty(\eta, \eta') ,$$

for all $\eta, \eta' \in \mathscr{P}(\mathbb{R})^{\mathcal{X}}$.

Next, we show that the projection operator Π^λ cannot expand distances as measured by $\bar{w}_\infty$, generalising the proof given by Bellemare et al. (2023) in the case $\lambda \equiv 0$; the proof is given in Appendix A.1.

Proposition 6 *The projection operator $\Pi^\lambda : \mathscr{P}(\mathbb{R})^{\mathcal{X}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}}$ is a non-expansion with respect to $\bar{w}_\infty$. That is, for any $\eta, \eta' \in \mathscr{P}(\mathbb{R})^{\mathcal{X}}$, we have*

$$\bar{w}_\infty(\Pi^\lambda \eta, \Pi^\lambda \eta') \leq \bar{w}_\infty(\eta, \eta')\,.$$

Finally, we put these two results together to obtain our desired conclusion. In stating this result, it is useful here to introduce the notation

$$\mathscr{F}_{\mathrm{Q},m} = \Big\{ \sum_{i=1}^m \frac{1}{m} \delta_{z_i} : z_i \in \mathbb{R} \text{ for } i = 1, \ldots, m \Big\}\,,$$

for the set of probability distributions representable with m quantile locations.

Proposition 7 *The projected operator $\Pi^\lambda \mathcal{T}^\pi : \mathscr{F}_{Q,m}^{\mathcal{X}} \to \mathscr{F}_{Q,m}^{\mathcal{X}}$ is a γ-contraction with respect to $\bar{w}_\infty$. Hence, $\Pi^\lambda \mathcal{T}^\pi$ has a unique fixed point in $\mathscr{F}_{Q,m}^{\mathcal{X}}$, which we denote $\hat{\eta}_\lambda^\pi$. Further, given any initial $\eta_0 \in \mathscr{F}_{Q,m}^{\mathcal{X}}$, the sequence $(\eta_k)_{k=0}^\infty$ defined iteratively by $\eta_{k+1} = \Pi^\lambda \mathcal{T}^\pi \eta_k$ for $k \geq 0$ satisfies $\bar{w}_\infty(\eta_k, \hat{\eta}_\lambda^\pi) \leq \gamma^k \bar{w}_\infty(\eta_0, \hat{\eta}_\lambda^\pi) \to 0$.*

Proof That $\Pi^\lambda \mathcal{T}^\pi : \mathscr{F}_{\mathrm{Q},m}^{\mathcal{X}} \to \mathscr{F}_{\mathrm{Q},m}^{\mathcal{X}}$ is a γ-contraction with respect to $\bar{w}_\infty$ follows directly from Propositions 5 and 6:

$$\bar{w}_\infty(\Pi^\lambda \mathcal{T}^\pi \eta, \Pi^\lambda \mathcal{T}^\pi \eta') \leq \bar{w}_\infty(\mathcal{T}^\pi \eta, \mathcal{T}^\pi \eta') \leq \gamma \bar{w}_\infty(\eta, \eta')\,.$$

Next, observe that $\bar{w}_\infty$ assigns finite distance to all pairs of return-distribution functions in $\mathscr{F}_{\mathrm{Q},m}^{\mathcal{X}}$, and further, this set is complete with respect to $\bar{w}_\infty$. Hence, we may apply Banach's fixed point theorem to obtain the existence of the unique fixed point $\hat{\eta}_\lambda^\pi$ in $\mathscr{F}_{\mathrm{Q},m}^{\mathcal{X}}$. The final claim follows by induction, and the contraction property established for $\Pi^\lambda \mathcal{T}^\pi$. ■

Note that the fixed point $\hat{\eta}_\lambda^\pi$ depends on λ, and therefore implicitly on m. We also introduce the notation $\hat{\theta}_\lambda^\pi \in \mathbb{R}^{\mathcal{X} \times [m]}$ for the parameters of this collection of distributions, which is what the QDP algorithm really operates over, so that we have

$$\hat{\eta}_\lambda^\pi(x) = \sum_{i=1}^m \frac{1}{m} \delta_{\hat{\theta}_\lambda^\pi(x,i)}\,.$$

Note that the convergence result of Proposition 7 also implies convergence of the estimated quantile locations to $\hat{\theta}_\lambda^\pi$. In Section 6, we will analyse the fixed point $\hat{\eta}_\lambda^\pi$, and understand how closely it approximates the true return-distribution function η^π. For now, having established convergence of QDP through contraction mapping theory, we can return to QTD and demonstrate its own convergence to the same fixed points.

5. Convergence of Quantile Temporal-Difference Learning

We now present the convergence analysis of QTD. We will consider a *synchronous* version of QTD, in which all states are updated using independent transitions at each algorithm step, given by:

$$\theta_{k+1}(x,i) = \theta_k(x,i) + \alpha_k \frac{1}{m} \sum_{j=1}^{m} (\tau_i - \mathbb{1}\{R_k(x) + \gamma\theta_k(X'_k(x), j) < \theta_k(x,i)\}) \,, \qquad (13)$$

where given x and k, we have $(R_k(x), X'_k(x)) \sim P^\pi(\cdot|x)$, independently of the transitions used at all other states/time steps, and $(\alpha_k)_{k=0}^\infty$ is a sequence of step sizes. The assumption of synchronous updates makes the analysis easier to present, and means that our results follow classical approaches to stochastic approximation with differential inclusions (Benaïm et al., 2005). It is also possible to extend the analysis to the asynchronous case, where a single state is updated at each algorithm time step (as would be the case in fully online QTD, or an implementation using a replay buffer); see Section 5.7. We now state the main convergence result of the paper.

Theorem 8 *Consider the sequence $(\theta_k)_{k=0}^\infty$ defined by an initial point $\theta_0 \in \mathbb{R}^{\mathcal{X}\times[m]}$, the iterative update in Equation* (13)*, and non-negative step sizes satisfying the condition*

$$\sum_{t=0}^{\infty} \alpha_k = \infty \,, \quad \alpha_k = o(1/\log k) \,. \qquad (14)$$

Then $(\theta_k)_{k=0}^\infty$ converges almost surely to the set of fixed points of the projected distributional Bellman operators $\{\Pi^\lambda \mathcal{T}^\pi : \lambda \in [0,1]^{\mathcal{X}\times[m]}\}$; that is,

$$\inf_{\lambda \in [0,1]^{\mathcal{X}\times[m]}} \|\theta_k - \hat{\theta}^\pi_\lambda\|_\infty \to 0$$

with probability 1.

Of particular note is the generality of this result. It does not require finite-variance conditions on rewards (as is typically the case with convergence results for classical TD); it holds for any collection of reward distributions with the finite mean property set out at the beginning of the paper. Some intuition as to why this is the case is that the finite-variance conditions typically encountered are to ensure that the updates performed in classical TD learning cannot grow in magnitude too rapidly. Since the updates performed in QTD are bounded, this is not a concern, meaning that the proof does not rely on such conditions. We note also that the step size conditions are weaker than the typical Robbins-Monro conditions used in classical TD analyses (see, for example, Bertsekas and Tsitsiklis, 1996), which enforce square-summability, also to avoid the possibility of divergence due to unbounded noise in the classical TD learning.

The proof is based on the ODE method for stochastic approximation; in particular we use the framework set out by Benaïm (1999) and Benaïm et al. (2005). This involves interpreting the QTD update as a noisy Euler discretisation of a differential equation (or more generally, a differential inclusion). The broad steps are then to argue that the trajectories of the

differential equation/inclusion converge to some set of fixed points in a suitable way (that is, in such a way that is robust to small perturbations), and that the asymptotic behaviour of QTD, forming a noisy Euler discretisation, matches the asymptotic behaviour of the true trajectories. This then allows us to deduce that the QTD iterates converge to the same set of fixed points as the true trajectories. We begin by elucidating the connection to differential equations and differential inclusions.

5.1 The QTD Differential Equation

Taking the expectation over the random variables $R_k(x)$ and $X'_k(x)$ in Equation (13) conditional on the algorithm history up to time k yields an expected increment of

$$\alpha_k \left(\tau_i - \mathbb{P}^{\pi}_x(R + \theta_k(X', J) < \theta_k(x, i))\right) . \tag{15}$$

We now briefly introduce an assumption on the MDP reward structure that simplifies the analysis that follows. This assumption guarantees that the two "difficult" cases of flat and vertical regions of CDFs (see Figure 1) do not arise; note that this assumptions removes the possibility of multiple fixed points or discontinuous expected dynamics, as described in Example 3. We will lift this assumption later.

Assumption 9 *For each state $x \in \mathcal{X}$, the reward distribution at x has a CDF which is strictly increasing, and Lipschitz continuous.*

As described in Section 4, the distribution of $R + \theta_k(X', J)$ given the initial state x is in fact equal to the application of the distributional Bellman operator $\mathcal{T}^{\pi}$ applied to the return-distribution function $\eta_k \in \mathscr{P}(\mathbb{R})^{\mathcal{X}}$ given by

$$\eta_k(x) = \frac{1}{m} \sum_{i=1}^{m} \delta_{\theta_k(x,i)} .$$

Under Assumption 9, and in particular the assumption of continuous reward CDFs, this yields a concise rewriting of the increment as

$$\alpha_k \left(\tau_i - F_{(\mathcal{T}^{\pi}\theta_k)(x)}(\theta_k(x, i))\right) .$$

We may therefore intuitively interpret Equation (13) as a noisy discretisation of the differential equation

$$\partial_t \vartheta_t(x, i) = \tau_i - F_{(\mathcal{T}^{\pi}\vartheta_t)(x)}(\vartheta_t(x, i)) , \tag{16}$$

which we refer to as the QTD differential equation (or QTD ODE). Note also that Assumption 9 guarantees the global existence and uniqueness of solutions to this differential equation, by the Cauchy-Lipschitz theorem.

Remark 10 *Calling back to Figure 3, the trajectories of the QTD ODE are obtained precisely by integrating the vector fields that appear in these plots. In contrast to the ODE that emerges when analysing classical TD learning (both in tabular and linear function approximation settings) (Tsitsiklis and Van Roy, 1997), the right-hand side of Equation* (16) *is non-linear in the parameters ϑ_t, meaning that we are outside the domain of linear stochastic approximation methods.*

5.2 The QTD Differential Inclusion

In lifting Assumption 9, a few complications arise. Firstly, if $F_{(\mathcal{T}^\pi\theta)(x)}$ is not continuous at $\theta(x,i)$, then the right-hand side of the QTD ODE in Equation (16) is modified to

$$\tau_i - \mathbb{P}_{Z\sim(\mathcal{T}^\pi\vartheta_t)(x)}(Z < \vartheta_t(x,i))\,;$$

the difference is the strict inequality. Now the right-hand side of the differential equation itself is not continuous; in general, solutions may not even exist for this differential equation. The situation is illustrated in the bottom-left panel of Figure 3; the lines in this plot illustrate points of discontinuity of the vector field to be integrated, and there are instances where the vector field either side of such a line of discontinuity "pushes" back into the discontinuity. In such cases, the differential equation has no solution in the usual sense. This phenomenon is known as sliding, or sticking, from cases when it arises in the modelling of physical systems with potentially discontinuous forces (such as static friction models in mechanics).

Filippov (1960) proposed a method to deal with such non-existence issues, by relaxing the definition of the dynamics at points of discontinuity. Technically, Filippov's proposal is to allow the derivative to take on any value in the convex hull of possible limiting values as we approach the point of discontinuity. In our case, we consider redefining the dynamics at points of discontinuity as follows:

$$\partial_t\vartheta_t(x,i) \in [\tau_i - F_{(\mathcal{T}^\pi\vartheta_t)(x)}(\vartheta_t(x,i)), \tau_i - F_{(\mathcal{T}^\pi\vartheta_t)(x)}(\vartheta_t(x,i)-)]\,, \tag{17}$$

where $F_{(\mathcal{T}^\pi\vartheta_t)(x)}(\vartheta_t(x,i)-)$ denotes $\lim_{s\uparrow\vartheta_t(x,i)} F_{(\mathcal{T}^\pi\vartheta_t)(x)}(s)$. This refines the dynamics so that for each coordinate (x,i), the derivative may take on either the left or right limit around $\vartheta_t(x,i)$, or any value in between; this is a looser relaxation than Filippov's proposal, and is easier to work with in our analysis.

Equation (17) is a *differential inclusion*, as opposed to a differential equation; the derivative is constrained to a set at each instant, rather than constrained to a single value. We refer to Equation (17) specifically as the QTD differential inclusion (or QTD DI). Note that if $F_{(\mathcal{T}^\pi\theta)(x)}$ is continuous at $\theta(x,i)$, then the right-hand side of Equation (17) reduces to the singleton $\{\tau_i - F_{(\mathcal{T}^\pi\theta)(x)}(\theta(x,i))\}$, and we thus obtain the ODE dynamics considered previously.

5.3 Solutions of Differential Inclusions

We briefly recall some key concepts regarding solutions of differential inclusions; a full review of the theory of differential inclusions is beyond the scope of this article, and we refer the reader to the standard references by Aubin and Cellina (1984), Clarke et al. (1998), and Smirnov (2002).

Definition 11 *Let $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map. The path $(z_t)_{t\geq 0}$ is a solution to the differential inclusion $\partial_t z_t \in H(z_t)$ if there exists an integrable function $g : [0,\infty) \to \mathbb{R}^n$ such that*

$$z_t = \int_0^t g_s \mathrm{d}s \tag{18}$$

for all $t \geq 0$, and $g_t \in H(z_t)$ for almost all $t \geq 0$.

Note that Definition 11 does not require that z_t is *differentiable* with derivative g_t, but only the weaker integration condition in Equation (18). We then have the following existence result (see, for example, Smirnov, 2002 for a proof).

Proposition 12 *Consider a set-valued map $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and suppose that H is a* Marchaud map*: that is,*

- *the set $\{(z, h) : z \in \mathbb{R}^n, h \in H(z)\}$ is closed.*
- *For all $z \in \mathbb{R}^n$, $H(z)$ is non-empty, compact, and convex.*
- *There exists a constant $C > 0$ such that for all $z \in \mathbb{R}^n$,*

$$\max_{h \in H(z)} \|h\| \leq C(1 + \|z\|)\,.$$

Then the differential inclusion $\partial_t z_t \in H(z_t)$ has a global solution, for any initial condition.

It is readily verified that the QTD DI satisfies the requirements of this result, and we are therefore guaranteed global solutions to this differential inclusion, under any initial conditions.

5.4 Asymptotic Behaviour of Differential Inclusion Trajectories

Recall that our goal is to show that the trajectories of the QTD differential inclusion must approach the fixed points of QDP. A key tool in doing so is the notion of a Lyapunov function; the following definition is based on Benaïm et al. (2005).

Definition 13 *Consider a Marchaud map $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and a subset $\Lambda \subseteq \mathbb{R}^n$. A continuous function $L : \mathbb{R}^n \to [0, \infty)$ is said to be a* Lyapunov function *for the differential inclusion $\partial_t z_t \in H(z_t)$ and subset Λ if for any solution $(z_t)_{t \geq 0}$ of the differential inclusion and $0 \leq s < t$, we have $L(z_t) < L(z_s)$ for all $z_s \notin \Lambda$ and $L(z) = 0$ for all $z \in \Lambda$.*

Intuitively, L is a Lyapunov function if it decreases along trajectories of the differential inclusion, and is minimal precisely on Λ. Lyapunov functions are a central tool in dynamical systems for demonstrating convergence, and in the sections that follow, we will consider the QTD differential inclusion, and take Λ to be the set of fixed points of the family of QDP algorithms.

5.5 QTD as a Stochastic Approximation to the QTD Differential Inclusion

We can now give the proof of our core result, Theorem 8. The abstract stochastic approximation result at the heart of the convergence proof of QTD is presented below. It is a special case of the general framework described by Benaïm et al. (2005), the proof of which is given in Appendix A.2.

Theorem 14 *Consider a Marchaud map $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and the corresponding differential inclusion $\partial_t z_t \in H(z_t)$. Suppose there exists a Lyapunov function L for this differential inclusion and a subset $\Lambda \subseteq \mathbb{R}^n$. Suppose also that we have a sequence $(\theta_k)_{k \geq 0}$ satisfying*

$$\theta_{k+1} = \theta_k + \alpha_k(g_k + w_k)\,,$$

where:

- $(\alpha_k)_{k=0}^{\infty}$ *satisfy the conditions* $\sum_{k=0}^{\infty} \alpha_k = \infty$, $\alpha_k = o(1/\log(k))$;
- $g_k \in H(\theta_k)$ *for all* $k \geq 0$;
- $(w_k)_{k=0}^{\infty}$ *is a bounded martingale difference sequence with respect to the natural filtration generated by* $(\theta_k)_{k=0}^{\infty}$; *that is, there is an absolute constant* C *such that* $\|w_k\|_{\infty} < C$ *almost surely, and* $\mathbb{E}[w_k|\theta_0, \ldots, \theta_k] = 0$.

If further $(\theta_k)_{k=0}^{\infty}$ *is bounded almost surely (that is,* $\sup_{k\geq 0} \|\theta_k\|_{\infty} < \infty$ *almost surely), then* $\theta_k \to \Lambda$ *almost surely.*

The intuition behind the conditions of the theorem are as follows. The Marchaud map condition ensures the differential inclusion of interest has global solutions. The existence of the Lyapunov function guarantees that trajectories of the differential inclusion converge in a suitably stable sense to Λ. The step size conditions, martingale difference condition, and boundedness conditions mean that the iterates $(\theta_k)_{k=0}^{\infty}$ will closely track the differential inclusion trajectories, and hence exhibit the same asymptotic behaviour. We can now give the proof of Theorem 8, first requiring the following proposition, which is proven in Appendix A.3.

Proposition 15 *Under the conditions of Theorem 8, the iterates* $(\theta_k)_{k=0}^{\infty}$ *are bounded almost surely.*

Proof (Proof of Theorem 8) We see that for the QTD sequence $(\theta_k)_{k=0}^{\infty}$ and the QTD DI and QDP invariant set $\Lambda = \{\hat{\theta}_{\lambda}^{\pi} : \lambda \in [0,1]^{\mathcal{X}\times[m]}\}$, the conditions of Theorem 14 are satisfied, except perhaps for the boundedness of $(\theta_k)_{k=0}^{\infty}$, and the existence of the Lyapunov function. The fact that the sequence $(\theta_k)_{k=0}^{\infty}$ is bounded almost surely is Proposition 15; its proof is somewhat technical, and given in the appendix. The construction of a valid Lyapunov function is given in Proposition 18 below, which completes the proof. ■

Remark 16 *What makes the relaxation to the differential inclusion work in this analysis? We have already seen that some kind of relaxation of the dynamics is required in order to define a valid continuous-time dynamical system; the original ODE may not have solutions in general. If we relax the dynamics too much (an extreme example would be the differential inclusion* $\vartheta_t(x, i) \in \mathbb{R}$*), what goes wrong? The answer is that there are too many resulting solutions, which do not exhibit the desired asymptotic behaviour. Thus, the differential inclusion in Equation* (17) *is in some sense just the right level of relaxation of the differential equation we started with, since trajectories of the QTD DI are still guaranteed to converge to the QDP fixed points.*

5.6 A Lyapunov Function for the QDP Fixed Points

In this section, we prove the existence of a Lyapunov function required in order to use Theorem 14 to prove Theorem 8. We treat the case when Assumption 9 holds separately as the proof is instructive, and considerably simpler than the general case. Under this assumption, note that all projections Π^{λ} behave identically on the image of $\mathcal{T}^{\pi}$, since all

resulting CDFs are strictly increasing. We therefore introduce the notation Π to refer to any such projection in this case, and the notation $\hat{\theta}^\pi_m$ to refer to the unique fixed point of $\Pi\mathcal{T}^\pi$.

Proposition 17 *Consider the ODE in Equation* (16)*, and suppose Assumption 9 holds. A Lyapunov function for the equilibrium point* $\hat{\theta}^\pi_m$ *is given by*

$$L(\theta) = \max_{x\in\mathcal{X}} \max_{i=1,\dots,m} |\theta(x,i) - \hat{\theta}^\pi_m(x,i)|\,.$$

Proof We immediately observe that L is continuous, non-negative, and takes on the value 0 only at $\hat{\theta}^\pi_m$. To show that $L(\vartheta_t)$ is decreasing, where $(\vartheta_t)_{t\geq 0}$ is an ODE trajectory, suppose (x,i) is a state-index pair attaining the maximum in $L(\vartheta_t)$. It is sufficient to show that $\vartheta_t(x,i)$ is moving towards $\hat{\theta}^\pi_m(x,i)$, or expressed mathematically,

$$\partial_t \vartheta_t(x,i) \overset{\text{S}}{=} \hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i)\,,$$

where we use $a \overset{\text{S}}{=} b$ as shorthand for *equality of signs* $\operatorname{sign}(a) = \operatorname{sign}(b)$, where

$$\operatorname{sign}(z) = \begin{cases} 1 & \text{if } z > 0\,, \\ 0 & \text{if } z = 0\,, \\ -1 & \text{if } z < 0\,. \end{cases}$$

Now note that

$$\begin{aligned}
\partial_t \vartheta_t(x,i) &= \tau_i - F_{(\mathcal{T}^\pi\vartheta_t)(x)}(\vartheta_t(x,i)) \\
&\overset{\text{S}}{=} F^{-1}_{(\mathcal{T}^\pi\vartheta_t)(x)}(\tau_i) - F^{-1}_{(\mathcal{T}^\pi\vartheta_t)(x)}(F_{(\mathcal{T}^\pi\vartheta_t)(x)}(\vartheta_t(x,i))) \\
&= (\Pi\mathcal{T}^\pi\vartheta_t)(x,i) - \vartheta_t(x,i)\,,
\end{aligned}$$

where the sign equality follows from Assumption 9; since all reward CDFs are strictly increasing, so too is $F_{(\mathcal{T}^\pi\vartheta_t)(x)}$, and so $F^{-1}_{(\mathcal{T}^\pi\vartheta_t)(x)}$ is strictly monotonic. Additionally, from the contractivity of $\Pi\mathcal{T}^\pi$ with respect to $\overline{w}_\infty$ (see Proposition 7), we have

$$\begin{aligned}
|(\Pi\mathcal{T}^\pi\vartheta_t)(x,i) - \hat{\theta}^\pi_m(x,i)| &\leq \overline{w}_\infty(\Pi\mathcal{T}^\pi\vartheta_t, \Pi\mathcal{T}^\pi\hat{\theta}^\pi_m) \\
&\leq \gamma \max_{y\in\mathcal{X}} \max_{j=1,\dots,m} |\vartheta_t(y,j) - \hat{\theta}^\pi_m(y,j)| \\
&= \gamma|\vartheta_t(x,i) - \hat{\theta}^\pi_m(x,i)|\,; \qquad (19)
\end{aligned}$$

the equality follows since we selected (x,i) attain the maximum in the definition of $L(\vartheta_t)$. From this, we deduce

$$(\Pi\mathcal{T}^\pi\vartheta_t)(x,i) - \vartheta_t(x,i) \overset{\text{S}}{=} \hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i)\,;$$

which follows by considering the three cases for the sign of $\hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i)$. If the sign equals zero, then since (x,i) was chosen to be maximal in the definition of $L(\vartheta_t)$, we have

$\vartheta_t = \hat{\theta}^\pi_m$, and hence $\Pi\mathcal{T}^\pi\vartheta_t = \hat{\theta}^\pi_m$, and the claim follows; both sides are equal to 0. For the case $\hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i) < 0$, then note we have

$$\begin{aligned}(\Pi\mathcal{T}^\pi\vartheta_t)(x,i) - \vartheta_t(x,i) &\leq \hat{\theta}^\pi_m(x,i) + \gamma(\vartheta_t(x,i) - \hat{\theta}^\pi_m(x,i)) - \vartheta_t(x,i) \\ &= (1-\gamma)(\hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i)) < 0\,,\end{aligned}$$

as required, with the inequality above following from Equation (19). The case $\hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i) > 0$ follow similarly. We therefore have

$$\partial_t\vartheta_t(x,i) \overset{\text{S}}{=} \hat{\theta}^\pi_m(x,i) - \vartheta_t(x,i)\,.$$

We therefore have that $L(\vartheta_t)$ is decreasing at t, strictly so if $\vartheta_t \neq \hat{\theta}^\pi_m$, as required to establish the result. ■

The proof of Proposition 17 also sheds further light on the mechanisms underlying the QTD algorithm. A key step in the argument is to show that for the state-index pairs (x,i) such that $\vartheta_t(x,i)$ is maximally distant from the fixed point $\theta^\pi_m(x,i)$, the expected update under QTD moves this coordinate of the estimate in the same direction as gradient descent on a squared loss from the fixed point. However, the fact that it is only the *sign* of the update that has this property, and not its magnitude, means that the empirical rate of convergence and stability of QTD can be expected to be somewhat different from methods based on an L^2 loss, such as classical TD.

We now state the Lyapunov result in the general case; the proof is somewhat more involved, and is given in Appendix A.4.

Proposition 18 *The function*

$$L(\theta) = \min_{\lambda\in[0,1]^{\mathcal{X}\times[m]}} \max_{(x,i)} |\theta(x,i) - \hat{\theta}^\pi_\lambda(x,i)| \tag{20}$$

is a Lyapunov function for the differential inclusion in Equation (17) *and the set of fixed points* $\{\hat{\theta}^\pi_\lambda : \lambda \in [0,1]^{\mathcal{X}\times[m]}\}$.

5.7 Extension to Asynchronous QTD

Our convergence results have focused on the synchronous case of QTD. However, in practice, it is often of interest to implement *asynchronous* versions of TD algorithms, in which only a single state is updated at a time. More formally, an asynchronous version of QTD computes the sequence $(\theta_k)_{k\geq 0}$ defined by an initial estimate $\theta \in \mathbb{R}^{\mathcal{X}\times[m]}$, a sequence of transitions $(X_k, R_k, X'_k)_{k\geq 0}$, and the update rule

$$\theta_{k+1}(x,i) = \theta_k(x,i) + \beta_{x,k}\frac{1}{m}\sum_{j=1}^{m}(\tau_i - \mathbb{1}\{R_k + \gamma\theta_k(X'_k,j) < \theta_k(x,i)\})$$

for $x = X_k$, and $\theta_{k+1}(x,i) = \theta_k(x,i)$ otherwise. Here, the step size $\beta_{x,k}$ depends on both x and k, and is typically selected so that each state individually makes use of a fixed step size

sequence $(\alpha_k)_{k=0}^{\infty}$, by taking $\beta_{x,k} = \alpha_{\sum_{l=0}^{k} \mathbb{1}\{X_l = x\}}$. This models the online situation where a stream of experience $(X_k, R_k)_{k\geq 0}$ is generated by interacting with the environment using the policy π, and updates are performed setting $X'_k = X_{k+1}$, and also the setting in which the tuples $(X_k, R_k, X'_k)_{k\geq 0}$ are sampled i.i.d. from a replay buffer, among others.

Convergence of QTD in such asynchronous settings can also be proven; Perkins and Leslie (2013) extend the analysis of Benaïm (1999) and Benaïm et al. (2005), incorporating the approach of Borkar (1998), to obtain convergence guarantees for asynchronous stochastic approximation algorithms approximating differential inclusions. In the interest of space, we do not provide the full details of the proof here, but instead sketch the key differences that arise in the analysis in Appendix C.

6. Analysis of the QTD Limit Points

In general, the limiting points $\hat{\eta}^{\pi}_{\lambda}$ for QTD/QDP will not be the same as the true return-distribution function η^{π}. On the one hand, this is clear; each return-distribution function $\hat{\eta}^{\pi}_{\lambda}$ is in the image of the projection Π^{λ}, so each constituent probability distribution must be of the form $\frac{1}{m}\sum_{i=1}^{m} \delta_{z_i}$, whereas the true return distributions need not take on this form. In addition, the magnitude of this approximation error is not immediately clear. Each application of the projection Π^{λ} in the dynamic programming process causes some loss of information, and the quality of the fixed point $\hat{\eta}^{\pi}_{\lambda}$ is affected by the build up of these approximations over time.

Measuring approximation error in $\bar{w}_{\infty}$ typically turns out to be uninformative, as $\bar{w}_{\infty}$ is a particularly strict notion of distance between probability distributions, as discussed in the context of distributional RL by Rowland et al. (2019) and Bellemare et al. (2023). In particular, fixed points $\hat{\eta}^{\pi}_{\lambda}$ that intuitively provide a good approximation to η^{π} may have high $\bar{w}_{\infty}$-distance, and the $\bar{w}_{\infty}$-distance generally does not decrease with m (Bellemare et al., 2023). Instead, we use the Wasserstein-1 metric, and its extension to return-distribution functions, defined by

$$w_1(\nu, \nu') = \int_0^1 |F_{\nu}^{-1}(t) - F_{\nu'}^{-1}(t)| \, \mathrm{d}t \,, \quad \bar{w}_1(\eta, \eta') = \max_{x \in \mathcal{X}} w_1(\eta(x), \eta'(x)) \,,$$

for all $\nu, \nu' \in \mathscr{P}(\mathbb{R})$, and $\eta, \eta' \in \mathscr{P}(\mathbb{R})^{\mathcal{X}}$. The following result improves on the analysis given by Bellemare et al. (2023) for the case of $\lambda \equiv 0$, establishing an upper bound on the $\overline{w}_1$ distance between $\hat{\eta}^{\pi}_{\lambda}$ and η^{π} for any λ, essentially by showing that the errors accumulated in dynamic programming can be made arbitrarily small by increasing m, which controls the richness of the distribution representation.

Proposition 19 *For any $\lambda \in [0,1]^{\mathcal{X}\times[m]}$, if all reward distributions are supported on $[R_{\text{MIN}}, R_{\text{MAX}}]$, then we have*

$$\bar{w}_1(\hat{\eta}^{\pi}_{\lambda}, \eta^{\pi}) \leq \frac{V_{\text{MAX}} - V_{\text{MIN}}}{2m(1-\gamma)} \,,$$

where $V_{\text{MAX}} = R_{\text{MAX}}/(1-\gamma)$, and similarly $V_{\text{MIN}} = R_{\text{MIN}}/(1-\gamma)$.

Remark 20 *This bound also provides motivation for the specific values of $(\tau_i)_{i=1}^m$ that QTD uses. A similar convergence analysis and fixed-point analysis can be straightforwardly carried out for a version of the QTD algorithm with other values for $(\tau_i)_{i=1}^m$; by tracing through the proof of Proposition 19, it can be seen that the bound is proportional to $\max(\tau_1, \max((\tau_{i+1} - \tau_i)/2 : i = 2, \ldots, m-1), 1 - \tau_m)$, which is minimised precisely by the choice of $(\tau_i)_{i=1}^m$ used by QTD.*

6.1 Instance-Dependent Bounds

The result above implicitly assumes the worst-case projection error is incurred at all states with each application of the Bellman operator. In environments where this is not the case, the fixed point can be shown to be of considerably better quality. We describe an example of an instance-dependent quality bound here.

Proposition 21 *Consider an MDP such that for any trajectory, after k time steps all encountered transition distributions and reward distributions are Dirac deltas. If all reward distributions in the MDP are supported on $[R_{\text{MIN}}, R_{\text{MAX}}]$, then for any $\lambda \in [0,1]^{\mathcal{X} \times [m]}$, we have*

$$\bar{w}_1(\hat{\eta}_\lambda^\pi, \eta^\pi) \leq \frac{(V_{\text{MAX}} - V_{\text{MIN}})(1 - \gamma^k)}{2m(1-\gamma)}.$$

Remark 22 *One particular upshot of this bound for practitioners is that for agents in near-deterministic environments using near-deterministic policies, it may be possible to use $m = o((1-\gamma)^{-1})$ quantiles and still obtain accurate approximations to the return-distribution function via QTD and/or QDP. It is interesting to contrast this result for quantile-based distributional reinforcement learning against the case when using categorical distribution representations (Bellemare et al., 2017; Rowland et al., 2018; Bellemare et al., 2023). In this latter case, fixed point error continues to be accumulated even when the environment has solely deterministic transitions and rewards, due to the well-documented phenomenon of the approximate distribution 'spreading its mass out' under the Cramér projection (Bellemare et al., 2017; Rowland et al., 2018; Bellemare et al., 2023). Our observation here leads to immediate practical advice for practitioners (in environments with mostly deterministic transitions, a quantile representation may be preferred to a categorical representation, leading to less approximation error), and raises a general question that warrants further study: how can we use prior knowledge about the structure of the environment to select a good distribution representation?*

We conclude this section by noting that many variants of Proposition 21 are possible; one can for example modify the assumption that rewards are deterministic to an assumption that rewards distributions are supported on a 'small' interval, and still obtain a fixed-point bound that improves over the instance-independent bound of Proposition 19. There are a wide variety of such modifications that could be imagined, and we believe this to be an interesting direction for future research and applications.

6.2 Qualitative Analysis of QDP Fixed Points

The analysis in the previous section establishes quantitative upper bounds on the quality of the fixed point learnt by QTD, and guarantees that with enough atoms an arbitrarily accurate approximation of the return-distribution function (as measured by w_1) can be learnt. We now take a closer look at the way in which approximation errors may manifest in QTD and QDP.

Example 23 *Consider the two-state Markov decision process (with a single action) whose transition probabilities are specified by the left-hand side of Figure 5, such that a deterministic reward of* 2 *is obtained in state* x_1*, and* -1 *in state* x_2*; further, let us take a discount factor* $\gamma = 0.9$*. The centre panel of this figure shows various estimates of the CDF for the return distribution at state* x_1*. The ground truth estimate in black is obtained from Monte Carlo sampling. The CDFs in purple, blue, green, and orange are the points of convergence for QDP with* $m = 2, 5, 10, 100$*, respectively. For* $m = 100$*, a very close fit to the true return distribution is obtained. However, for small* m *in particular, the distribution is heavily skewed to the right. In the case of* $m = 2$*, half of the probability mass is placed on the greatest possible return in this MDP—namely 20—even though with probability 1 the true return is less than this value. What is the cause of this behaviour from QDP? This question is answered by investigating the dynamic programming update itself in more detail.*

In this MDP, the result of the QDP operator applied to the fixed point θ *is to update each particle location with a 'backed-up' particle appearing in the distributions* $\mathcal{T}^\pi\theta$*. When such settings arise, tracking which backed-up particles are allocated to which other particles helps us to understand the behaviour of QDP, and the nature of the approximation incurred. We also gain intuition about the situation, since the QDP operator is behaving like a an affine policy evaluation operator on* $\mathcal{X} \times [m]$ locally *around the fixed point. We can visualise which particles are assigned to one another by a QDP operator application through* local quantile back-up diagrams*; the right-hand side of Figure 5 shows the local quantile back-up diagram for particular MDP. We observe that* $\theta(x_1, 2)$ *backs up from itself, and hence learns a value that corresponds to observing a self-transition at every state, with a reward of 2; under the discount factor of* 0.9*, this corresponds to a return of 20. This is the source of the drastic over-estimation of returns in the approximation obtained with* $m = 2$*, and the fact that all other state-quantile pairs implicitly bootstrap from this estimate leads to the over-estimation leaking out into all quantiles estimated in this case. As* m *increases, we observe from the CDF plot that there is always one particle that learns this maximal return of 20, but that this has less effect on the other quantiles; indeed in the orange curve, we obtain a very good approximation (in* w_1*) to the true return distribution despite this particle with a maximal value of 20 remaining present. We can interpret the increase in* m *as preventing pathological self-loops/small cycles in the quantile backup diagram from "leaking out" and degrading the quality of other quantile estimates; this provides a complementary perspective on the approximation artefacts that occur in QDP/QTD fixed points to the quantitative upper bounds in the previous section.*

We expect the local quantile back-up diagram introduced in Example 23 to be a useful tool for developing intuition, as well as further analysis, of QDP and QTD. As described in the example itself, being able to define the local back-up diagram depends on the structure

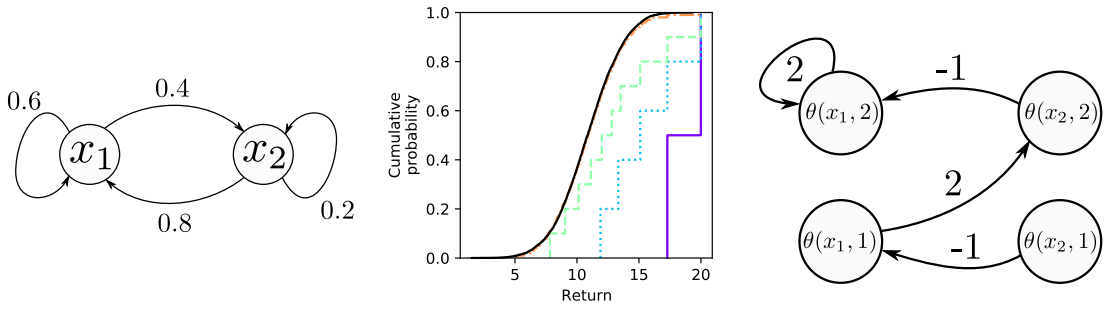

Figure 5: Left: An example MDP. Centre: The fixed point return distribution estimates for state x_1 obtained by QDP for $m = 2, 5, 20, 100$ (solid purple, dotted blue, dashed green, and dash-dotted orange, respectively) compared to ground truth in solid black. Right: The corresponding local quantile backup diagram at the fixed point for $m = 2$, illustrating potential approximation artefacts in QDP fixed points.

of the MDP being such that the QDP operator obtains each new coordinate value from a single backed-up particle location. It is an interesting question as to how the definition of such local back-up diagrams could be generalised to apply in situations where this does not hold, such as with absolutely continuous reward distributions.

7. Related Work

Stochastic approximation theory with differential inclusions. The ODE method was introduced by Ljung (1977) as a means of analysing stochastic approximation algorithms, and was subsequently extended and refined by Kushner and Clark (1978); standard references on the subject include Kushner and Yin (2003); Borkar (2008); Benveniste et al. (2012); see also Meyn (2022) for an overview in the context of reinforcement learning. The framework we follow in this paper is set out by Benaïm (1999), and was extended by Benaïm et al. (2005) to allow for differential inclusions. Perkins and Leslie (2013) later extended this analysis further to allow for asynchronous algorithms, building on the approach introduced by Borkar (1998), and extended, with particular application to reinforcement learning, by Borkar and Meyn (2000).

Differential inclusion theory. Differential inclusions have found application across a wide variety of fields, including control theory (Wazewski, 1961), economics (Aubin, 1991) differential game theory (Krasovskii and Subbotin, 1988), and mechanics (Monteiro Marques, 2013). The approach to modelling differential equations with discontinuous right-hand sides via differential inclusions was introduced by Filippov (1960). Standard references on the theory of differential inclusions include Aubin and Cellina (1984); Clarke et al. (1998); Smirnov (2002); see also Bernardo et al. (2008) on the related field of piecewise-smooth dynamical systems. Joseph and Bhatnagar (2019) also use tools combining stochastic approximation and differential inclusions from Benaïm et al. (2005) to analyse (sub-)gradient descent as a means of estimating quantiles of fixed distributions. Within reinforcement learning and related fields more specifically, differential inclusions have played a key role in

the analysis of game-theoretic algorithms based on fictitious play (Brown, 1951; Robinson, 1951); see Benaïm et al. (2006); Leslie and Collins (2006); Benaïm and Faure (2013) for examples. More recently, Gopalan and Thoppe (2023) used differential inclusions to analyse TD algorithms for control with linear function approximation.

Quantile regression. Quantile regression as a methodology for statistical inference was introduced by Koenker and Bassett (1978). Koenker (2005) and Koenker et al. (2017) provide detailed surveys of the field. Quantile temporal-difference learning may be viewed as fusing quantile regression with the bootstrapping approach (*learning a guess from a guess*, as Sutton and Barto (2018) express it) that is core to much of the reinforcement learning methodology.

Quantiles in reinforcement learning. The approach to distributional reinforcement learning based on quantiles was introduced by Dabney et al. (2018b). A variety of modifications and extensions were then considered in the deep reinforcement learning setting (Dabney et al., 2018a; Yang et al., 2019; Zhou et al., 2020; Luo et al., 2021), as well as further developments on the theoretical side (Lhéritier and Bondoux, 2022). A summary of the approach is presented by Bellemare et al. (2023). Gilbert and Weng (2016) study the problem of optimising quantile criteria in end-state MDPs. Li et al. (2022) consider the risk-sensitive control problem of optimising particular quantiles of the return distribution, and derive a dynamic programming algorithm that maintains a value function over state and the "target quantile-to-go".

8. Conclusion

We have provided the first convergence analysis for QTD, a popular and effective distributional reinforcement learning algorithm. In contrast to the analysis of many classical temporal-difference learning algorithms, this has required the use of tools from the field of differential inclusions and branches of stochastic approximation theory that deal with the associated dynamical systems. Due to the structure of the QTD algorithm, such as its bounded-magnitude updates, these convergence guarantees hold under weaker conditions than are generally used in the analysis of TD algorithms. These results establish the soundness of QTD, representing an important step towards understanding its efficacy and practical successes, and we expect the theoretical tools used here to be useful in further analyses of (distributional) reinforcement learning algorithms.

There are several natural directions for further work building on this analysis. One such direction is to establish finite-sample bounds for the convergence of QTD predictions to the set of QDP fixed points. This is a central theoretical question for developing our understanding of QTD, and may also shed further light on the recently observed empirical phenomenon in which tabular QTD can outperform TD in stochastic environments as a means of value estimation (Rowland et al., 2023). Related to this point, the Lyapunov analysis conducted in this paper provides further intuition for *why* QTD works in general, and we expect this to inform the design of further variants of QTD, for example incorporating multi-step bootstrapping (Watkins, 1989), or Ruppert-Polyak averaging (Ruppert, 1988; Polyak and Juditsky, 1992). Another important direction is to analyse more complex

variants of the QTD algorithm, incorporating more aspects of the large-scale systems in which it has found application. Examples include incorporating function approximation, or control variants of the algorithm based on Q-learning. We believe further research into the theory, practice and applications of QTD, at a variety of scales, are important directions for foundational reinforcement learning research.

Acknowledgments

We thank the anonymous reviewers and action editor for helpful suggestions and feedback. We also thank Tor Lattimore for detailed comments on an earlier draft, and David Abel, Bernardo Avila Pires, Diana Borsa, Yash Chandak, Daniel Guo, Clare Lyle, and Shantanu Thakoor for helpful discussions. Marc G. Bellemare was supported by Canada CIFAR AI Chair funding. The simulations in this paper were generated using the Python 3 language, and made use of the NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), and Matplotlib (Hunter, 2007) libraries.

Appendix A. Proofs

In this section, we provide proofs for results which are not proven in the main text.

A.1 Proof of Proposition 6

Let $\eta, \eta' \in \mathscr{P}(\mathbb{R})$. We have

$$\begin{aligned}\Big|\Big((1-\lambda(x,i))F^{-1}_{\eta(x)}(\tau_i)+\lambda(x,i)\bar{F}^{-1}_{\eta(x)}(\tau_i)\Big)-\Big((1-\lambda(x,i))F^{-1}_{\eta'(x)}(\tau_i)+\lambda(x,i)\bar{F}^{-1}_{\eta'(x)}(\tau_i)\Big)\Big| \\ \leq (1-\lambda(x,i))|F^{-1}_{\eta(x)}(\tau_i)-F^{-1}_{\eta'(x)}(\tau_i)|+\lambda(x,i)|\bar{F}^{-1}_{\eta(x)}(\tau_i)-\bar{F}^{-1}_{\eta'(x)}(\tau_i)|\,.\end{aligned}$$

Clearly

$$|F^{-1}_{\eta(x)}(\tau_i)-F^{-1}_{\eta'(x)}(\tau_i)| \leq \bar{w}_\infty(\eta,\eta')\,.$$

Additionally, we have

$$\begin{aligned}|\bar{F}^{-1}_{\eta(x)}(\tau_i)-\bar{F}^{-1}_{\eta'(x)}(\tau_i)| &= |\lim_{s\downarrow\tau_i} F^{-1}_{\eta(x)}(s)-\lim_{s\downarrow\tau_i} F^{-1}_{\eta'(x)}(s)| \\ &= \lim_{s\downarrow\tau_i}|F^{-1}_{\eta(x)}(s)-F^{-1}_{\eta'(x)}(s)| \\ &\leq \bar{w}_\infty(\eta,\eta')\,.\end{aligned}$$

Putting this together, we obtain

$$\begin{aligned}\Big|\Big((1-\lambda(x,i))F^{-1}_{\eta(x)}(\tau_i)+\lambda(x,i)\bar{F}^{-1}_{\eta(x)}(\tau_i)\Big)-\Big((1-\lambda(x,i))F^{-1}_{\eta'(x)}(\tau_i)+\lambda(x,i)\bar{F}^{-1}_{\eta'(x)}(\tau_i)\Big)\Big| \\ \leq \bar{w}_\infty(\eta,\eta')\,,\end{aligned}$$

as required.

A.2 Proof of Theorem 14

Theorem 14 is essentially a special case of the general results presented in Benaïm et al. (2005), in the form needed for the proof of convergence of QTD. To explain how to obtain Theorem 14 from the results of Benaïm et al. (2005), first, we associate a continuous-time path $(\bar{\theta}(t))_{t\geq 0}$ with the iterates $(\theta_k)_{k=0}^{\infty}$ by linear interpolation, in particular defining $\bar{\theta}(\sum_{k=0}^{s}\alpha_k) = \theta_s$, and linearly interpolating in between. The continuous-time path $(\bar{\theta}(t))_{t\geq 0}$ satisfies the definition of a *perturbed solution* of the Marchaud differential inclusion with probability 1, as defined by Definition II of Benaïm et al. (2005), since: (i) $\bar{\theta}$ is piecewise linear, hence absolutely continuous; (ii) the difference $\|\theta_{k+1} - \theta_k\|_\infty$ is $O(\alpha_k)$, due to the growth condition on H and since $\bar{\theta}$ is bounded by assumption; and (iii) the lim-sup condition holds with probability 1 thanks to the boundedness of the martingale difference sequence $(w_k)_{k=0}^{\infty}$ and Proposition 1.4 of Benaïm et al. (2005); see also Theorem 5.3.3 of Kushner and Yin (2003).

Next, since we assume $\bar{\theta}$ is bounded, Theorem 4.2 of Benaïm et al. (2005) applies so that we deduce that it is an asymptotic pseudotrajectory of the differential inclusion (w.p.1). We then have that $(\bar{\theta}(t))_{t\geq 0}$ is a bounded asymptotic pseudotrajectory (w.p.1), so Theorem 4.3 of Benaïm et al. (2005) applies, and we deduce that the set of limit points of $(\bar{\theta}(t))_{t\geq 0}$ is internally chain transitive (w.p.1). But now by Proposition 3.27 of Benaïm et al. (2005) applied to the Lyapunov function L and the set Λ, all internally chain transitive sets are contained within Λ. Since $(\bar{\theta}(t))_{t\geq 0}$ is bounded, we deduce that it converges to Λ (w.p.1). It therefore follows that the discrete sequence $(\theta_k)_{k=0}^{\infty}$ converges to Λ with probability 1, as required.

A.3 Proof of Proposition 15

Roughly, the intuition of the proof is that the structure of the QTD differential inclusion means that when $\|\theta_k\|_\infty$ is sufficiently large, the coordinates of θ_k furthest from the origin are moved back towards the origin by the differential inclusion. We then argue that the martingale noise cannot cause divergence, which completes the argument.

Differential inclusion update direction. To begin with the analysis of the differential inclusion, fix $\delta > 0$ such that $1 - \delta > \gamma$, and let $M > 0$ be such that for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, we have

$$F_{P_{\mathcal{R}}(x,a)}((1-\delta-\gamma)M) > 1 - 1/(4m)\,, \quad F_{P_{\mathcal{R}}(x,a)}(-(1-\delta-\gamma)M) < 1/(4m)\,.$$

We then introduce the events

$$\begin{aligned} I_k^+(x,i) &= \{\|\theta_k\|_\infty > M\,, \;\; \theta_k(x,i) > (1-\delta)\|\theta_k\|_\infty\}\,, \\ I_k^-(x,i) &= \{\|\theta_k\|_\infty > M\,, \;\; \theta_k(x,i) < -(1-\delta)\|\theta_k\|_\infty\}\,. \end{aligned}$$

which, roughly speaking, hold when θ_k has at least one large coordinate (in absolute value), and $\theta_k(x, i)$ is a positive (respectively, negative) coordinate close to the maximum value.

When $I_k^+(x,i)$ holds, we have

$$\begin{aligned}
&\tau_i - F_{(\mathcal{T}^\pi\theta_k)(x)}(\theta_k(x,i)-) && (21)\\
&= \tau_i - \sum_{x'\in\mathcal{X}} P(x'|x,a)\pi(a|x)\frac{1}{m}\sum_{j=1}^{m}\lim_{s\uparrow 0} F_{\mathcal{R}(x,a)}(\theta_k(x,i)-\gamma\theta_k(x',j)+s)\\
&\overset{(a)}{\leq} \tau_i - \sum_{x'\in\mathcal{X}} P(x'|x,a)\pi(a|x)\frac{1}{m}\sum_{j=1}^{m}(1-1/(4m))\\
&\leq (1-1/(2m)) - (1-(1/(4m))\\
&\leq -1/(4m)\,, && (22)
\end{aligned}$$

and hence the differential inclusion moves $\theta_k(x,i)$ towards the origin. Inequality (a) follows since on $I_k^+(x,i)$, we have

$$\theta_k(x,i) - \gamma\theta_k(x',j) \geq (1-\delta)\|\theta_k\|_\infty - \gamma\|\theta_k\|_\infty \geq (1-\delta-\gamma)M\,.$$

Analogously, we conclude that on $I_k^-(x,i)$, we have

$$\tau_i - F_{(\mathcal{T}^\pi\theta_k)(x)}(\theta_k(x,i)) \geq 1/(4m)\,,$$

and so the differential inclusion moves $\theta_k(x,i)$ towards the origin in this case too.

Chaining updates and reasoning about noise. To describe the relationship between successive iterates in the sequence $(\theta_k)_{k\geq 0}$, we introduce the notation $\theta_{k+1} = \theta_k + \alpha_k g_k + \alpha_k w_k$, where w_j is martingale difference noise, and hence g_j is an expected update direction, from the right-hand side of the QTD differential inclusion. By boundedness of the update noise and the step size assumptions, we have from Proposition 1.4 of Benaïm et al. (2005) (see also Theorem 5.3.3 of Kushner and Yin (2003)) that

$$\limsup_k \left\{ \Big\| \sum_{j=k}^{k+l} \alpha_j w_j \Big\|_\infty : l \geq 0 \text{ and } \sum_{j=k}^{k+l} \alpha_j \leq 8m+1 \right\} = 0\,,$$

almost surely. In particular, letting $\varepsilon \in (0,1)$, there almost-surely exists K (which depends on the realisation of the martingale noise) such that

$$\sup \left\{ \Big\| \sum_{j=k}^{k+l} \alpha_j w_j \Big\|_\infty : l \geq 0 \text{ and } \sum_{j=k}^{k+l} \alpha_j \leq 8m+1 \right\} < \varepsilon$$

for all $k \geq K$, and further such that $\alpha_k < 1$ for all $k \geq K$.

Let us additionally take $\bar{M} \geq M$ such that $\delta\bar{M} \geq 4(8m+1)$. Suppose that for some $k \geq K$, $\|\theta_k\|_\infty \geq \bar{M} + (8m+1)$. Let l be minimal such that $\sum_{j=k}^{k+l}\alpha_j > 8m$. Then we have $\|\theta_{k+j} - \theta_k\|_\infty \leq 8m+1$ for all $0 \leq j \leq l$, and so $\|\theta_{k+j}\|_\infty \geq \bar{M}$ for all $0 \leq j \leq l$. Further, if

$\theta_k(x,i)$ satisfies $\theta_k(x,i) > \|\theta_k\|_\infty(1-\delta) + 2(8m+1)$, then we have

$$\begin{aligned}
&\theta_{k+j}(x,i) \\
&\geq \theta_k(x,i) - (8m+1) \\
&\geq (1-\delta)\|\theta_k\|_\infty + (8m+1) \\
&\geq (1-\delta)(\|\theta_{k+j}\|_\infty - (8m+1)) + (8m+1) \\
&= (1-\delta)\|\theta_{k+j}\|_\infty \,,
\end{aligned}$$

so $I^+_{k+j}(x,i)$ holds for all $0 \leq j \leq l$, and hence

$$|\theta_{k+l+1}(x,i)| \leq \|\theta_k\|_\infty - \sum_{j=k}^{k+l} \alpha_j \times 1/(4m) + \|\sum_{j=k}^{k+l} \alpha_j w_j\|_\infty < \|\theta_k\|_\infty - 2 + \varepsilon < \|\theta_k\| - 1 \,. \tag{23}$$

Similarly, if $\theta_k(x,i) < -\|\theta_k\|_\infty(1-\delta) - 2(8m+1)$, $I^-_{k+j}(x,i)$ holds for all $0 \leq j \leq l$, and we reach the same conclusion as in Equation (23). Finally, if $|\theta_k(x,i)| \leq \|\theta_k\|_\infty(1-\delta) + 2(8m+1)$, then since $\delta\|\theta_k\|_\infty > \delta\bar{M}$, we have $|\theta_k(x,i)| \leq \|\theta_k\|_\infty - 2(8m+1)$, and hence $|\theta_{k+l+1}(x,i)| \leq \|\theta_k\|_\infty - (8m+1)$. Putting these components together, we have

$$\|\theta_{k+l+1}\|_\infty < \|\theta_k\|_\infty + 1 \,, \text{ and } \max_{0\leq j\leq l} \|\theta_{k+j}\|_\infty \leq \|\theta_k\|_\infty + (8m+1) \,,$$

as required to establish boundedness.

A.4 Proof of Proposition 18

We first state and prove a useful lemma that allows us to compare QDP fixed points for different values of λ. Throughout this section, we will adopt the shorthand θ^λ for $\hat{\theta}^\pi_\lambda$.

Lemma 24 *Let $\lambda, \lambda' \in [0,1]^{\mathcal{X}\times[m]}$. Then we have*

$$\|\theta^\lambda - \theta^{\lambda'}\|_\infty \leq C\|\lambda - \lambda'\|_\infty \,,$$

where C is a constant depending only on the reward distributions of the MDP and γ.

Proof By the triangle inequality, we have

$$\begin{aligned}
\|\theta^\lambda - \theta^{\lambda'}\|_\infty &\leq \|\theta^\lambda - \Pi^{\lambda'}\mathcal{T}^\pi\theta^\lambda\|_\infty + \|\Pi^{\lambda'}\mathcal{T}^\pi\theta^\lambda - \theta^{\lambda'}\|_\infty \\
&= \|\Pi^\lambda\mathcal{T}^\pi\theta^\lambda - \Pi^{\lambda'}\mathcal{T}^\pi\theta^\lambda\|_\infty + \|\Pi^{\lambda'}\mathcal{T}^\pi\theta^\lambda - \Pi^{\lambda'}\mathcal{T}^\pi\theta^{\lambda'}\|_\infty \\
&\leq \|(\Pi^\lambda - \Pi^{\lambda'})\mathcal{T}^\pi\theta^\lambda\|_\infty + \gamma\|\theta^\lambda - \theta^{\lambda'}\|_\infty \\
\implies \|\theta^\lambda - \theta^{\lambda'}\|_\infty &\leq \frac{1}{1-\gamma}\|(\Pi^\lambda - \Pi^{\lambda'})\mathcal{T}^\pi\theta^\lambda\|_\infty \,.
\end{aligned}$$

Now we aim to bound $\|\theta^\lambda\|_\infty$, and hence the term on the right-hand side above. Note that in general for a mixture distribution $\nu = \sum_{i=1}^n p_i\nu_i$, we have $F^{-1}_\nu(\tau) \geq \min\{F^{-1}_{\nu_i}(\tau) : i =$

$1, \ldots, n\}$, since

$$\begin{aligned}\mathbb{P}_{Z\sim\nu}(Z \leq \min\{F_{\nu_i}^{-1}(\tau_n) : i = 1, \ldots, n\}) &= \sum_{i=1}^{n} p_i \mathbb{P}_{Z_i\sim\nu_i}(Z_i \leq \min\{F_{\nu_i}^{-1}(\tau) : i = 1, \ldots, n\}) \\ &\leq \sum_{i=1}^{n} p_i \mathbb{P}_{Z_i\sim\nu_i}(Z_i \leq F_{\nu_i}^{-1}(\tau)) \\ &\leq \tau\,.\end{aligned}$$

Thus, it follows that the $1/2m$ quantile of $\mathcal{T}^\pi\theta^\lambda(x)$ is at least as great as

$$\min_x F^{-1}_{\mathcal{R}^\pi(x)}(1/2m) - \gamma\|\theta^\lambda\|_\infty\,.$$

By analogous reasoning, we obtain that $\bar{F}^{-1}_{(\mathcal{T}^\pi\theta^\lambda)(x)}(2m-1/2m)$ is no greater than

$$\max_x \bar{F}^{-1}_{\mathcal{R}^\pi(x)}(2m-1/2m) + \gamma\|\theta^\lambda\|_\infty\,.$$

From these facts, it follows that

$$\|\theta^\lambda\|_\infty \leq \frac{1}{1-\gamma}\max\left(|\min_x F^{-1}_{\mathcal{R}^\pi(x)}(1/2m)|, |\max_x \bar{F}^{-1}_{\mathcal{R}^\pi(x)}(2m-1/2m)|\right)\,,$$

and hence

$$\|(\Pi^\lambda - \Pi^{\lambda'})\mathcal{T}^\pi\theta^\lambda\|_\infty \leq C\|\lambda - \lambda'\|_\infty\,,$$

as required for the statement of the result. ■

We now turn to the proof of Proposition 18. First, we observe that the infimum over λ in Equation (20) is attained, since Lemma 24 establishes that $\lambda \mapsto \theta^\lambda$ is continuous (in fact Lipschitz), and $[0,1]^{\mathcal{X}\times[m]}$ is compact. We therefore have that L is continuous, non-negative, and takes on the value 0 only on the set of fixed points $\{\theta^\lambda : \lambda \in [0,1]^{\mathcal{X}\times[m]}\}$.

For the decreasing property, let $(\vartheta_t)_{t\geq 0}$ be a solution to the differential inclusion in Equation (17), and as in Definition 11, let $g : [0,\infty) \to \mathbb{R}^{\mathcal{X}\times[m]}$ satisfy

$$\vartheta_t = \int_0^t g_s\,\mathrm{d}s\,, \tag{24}$$

with $g_t(x,i) \in H^\pi_{x,i}(\vartheta_t)$ for all (x,i), and for almost all $t \geq 0$, where we have introduced the notation

$$H^\pi_{x,i}(\theta) = [\tau_i - F_{(\mathcal{T}^\pi\theta)(x)}(\theta(x,i)), \tau_i - F_{(\mathcal{T}^\pi\theta)(x)}(\theta(x,i)-)]\,.$$

As in the proof of Proposition 17, we will show that $L(\vartheta_t)$ is locally decreasing outside of the fixed point set, which is enough for the global decreasing property. Further, by continuity of $L(\vartheta_t)$, it is enough to show this property for almost all $t \geq 0$. We will therefore consider a value of $t \geq 0$ at which the above inclusion for g_t holds.

Let $\overline{\lambda}$ attain the minimum in the definition of $L(\vartheta_t)$. Write $\theta^{\overline{\lambda}}$ for the corresponding fixed point for conciseness, and let (x, i) be a *$\overline{\lambda}$-argmax index with respect to ϑ_t*; a state-particle pair achieving the maximum in the definition of the norm $\|\vartheta_t - \theta^{\overline{\lambda}}\|_\infty$. First, we consider the cases where $H^\pi_{x,i}(\vartheta_t)$ is *not* a singleton. Now, if $0 \in H^\pi_{x,i}(\vartheta_t)$, then we have $(\Pi^\lambda \mathcal{T}^\pi \vartheta_t)(x, i) = \vartheta_t(x, i)$, and with the same logic as above, we have $\vartheta_t = \theta^{\overline{\lambda}}$, and hence ϑ_t is in the fixed point set, and $L(\vartheta_t)$ is constant. If $0 \notin H^\pi_{x,i}(\vartheta_t)$, then as in the proof of Proposition 17, it can be shown that any element of $H^\pi_{x,i}(\vartheta_t)$ has the same sign as

$$(\Pi^\lambda \mathcal{T}^\pi \vartheta_t)(x, i) - \vartheta_t(x, i)\,. \tag{25}$$

In the case of Proposition 17, continuity of the derivative then allowed us to deduce that $|\vartheta_t(x, i) - \theta^{\overline{\lambda}}(x, i)|$ is locally decreasing. Here, we require a related concept of continuity for the set-valued map $\theta \mapsto H^\pi_{x,i}(\theta)$, namely that it is upper semicontinuous (see, for example, Smirnov, 2002); for a given $\theta \in \mathbb{R}^{\mathcal{X}\times[m]}$ and any given $\varepsilon > 0$, there exists $\delta > 0$ such that if $\|\theta' - \theta\|_\infty < \delta$, then $H^\pi_{x,i}(\theta') \subseteq \{h + v : h \in H^\pi_{x,i}(\theta)\,, |v| < \varepsilon\}$. From this, it follows that any element of $H^\pi_{x,i}(\vartheta_{t+s})$, for sufficiently small positive s, has the same sign as the expression in Equation (25), and so from Equation (24), we have that $|\vartheta_t(x, i) - \theta^{\overline{\lambda}}(x, i)|$ is locally decreasing, as required.

Now, when $H^\pi_{x,i}(\vartheta_t)$ *is* a singleton, if it is non-zero, then by the same argument as in the proof of Proposition 17, the corresponding element has the same sign as the expression in Equation (25), and so as above, we conclude that $|\vartheta_t(x, i) - \theta^{\overline{\lambda}}(x, i)|$ is locally decreasing.

Finally, the case where there exists an argmax index (x, i) with $H^\pi_{x,i}(\vartheta_t) = \{0\}$ requires more care, and we will need to reason about the effects of perturbing λ to show that the Lyapunov function is decreasing. For some intuition as to what the problem is, if $H^\pi_{x,i}(\vartheta_{t+s}) = \{0\}$ for small positive s, then the coordinate $\vartheta_{t+s}(x, i)$ is static, as it lies on the flat region of the CDF $F_{(\mathcal{T}^\pi \vartheta_{t+s})(x)}$ at level τ_i, and so the distance $|\vartheta_{t+s}(x, i) - \theta^{\overline{\lambda}}(x, i)|$ is not decreasing. We explain how to deal with this case below.

A.4.1 Perturbative Argument

We introduce the notation $J_0 \subseteq \mathcal{X} \times [m]$ for the set of $\overline{\lambda}$-argmax indices with respect to ϑ_t. Assuming that $\|\vartheta_t - \theta^{\overline{\lambda}}\|_\infty$ is not locally decreasing, it must be locally constant (it cannot increase, by the arguments above). Now consider $s > 0$ sufficiently small so that (i) no coordinates not in J_0 can be a $\overline{\lambda}$-argmax index with respect to ϑ_{t+s}, so that J, the set of $\overline{\lambda}$-argmax indices with respect to ϑ_{t+s}, satisfies $J \subseteq J_0$, (ii) all indices $(x, i) \in J$ satisfy $H^\pi_{x,i}(\vartheta_{t+u}) = \{0\}$ for all $u \in [0, 2s]$.

We will now demonstrate the existence of a parameter $\lambda' \in [0, 1]^{\mathcal{X}\times[m]}$ such that $\|\vartheta_{t+s} - \theta^{\lambda'}\|_\infty < \|\vartheta_t - \theta^{\overline{\lambda}}\|_\infty$, which establishes the locally decreasing property of the Lyapnuov function, as required. To do so, we introduce a modification of the fixed point map $\lambda \mapsto \theta^\lambda$. Letting $\mu \in \mathbb{R}^J$, and defining $\overline{\lambda}[\mu] \in \mathbb{R}^{\mathcal{X}\times[m]}$ to be the replacement of the J coordinates of $\overline{\lambda}$ with the corresponding coordinates of μ, we consider the map

$$h_{\overline{\lambda}} : [0, 1]^J \to \mathbb{R}^J\,, \quad h(\mu) = \mathrm{P}_J \theta^{\overline{\lambda}[\mu]}\,,$$

where $\mathrm{P}_J : \mathbb{R}^{\mathcal{X}\times[m]} \to \mathbb{R}^J$ extracts the J coordinates. At an intuitive level, this map allows us to study the effect of perturbing the J coordinates of $\overline{\lambda}$ on the corresponding coordinates of the fixed point.

A.4.2 Case 1: $\overline{\lambda}_J$ is in the Interior of $[0,1]^J$

We now first consider the case where $\overline{\lambda}_J$, the J coordinates of $\overline{\lambda}$, lies in the interior of $[0,1]^J$, that is $(0,1)^J$. By Lemma 24, $h_{\overline{\lambda}}$ is continuous, since it is the composition of the continuous maps $\mu \mapsto \overline{\lambda}[\mu]$, $\lambda \mapsto \theta^\lambda$, and $\theta \mapsto \mathrm{P}_J\theta$. It is also injective in a neighbourhood of $\overline{\lambda}_J$. This can be seen by noting first that the fixed points $\theta^{\overline{\lambda}[\mu]}$ are distinct for distinct values of μ sufficiently close to $\overline{\lambda}_J$; if $\mu \neq \mu'$ are each sufficiently close to $\overline{\lambda}_J$, then we have

$$\Pi^{\overline{\lambda}[\mu']}\mathcal{T}^\pi\theta^{\overline{\lambda}[\mu]} \neq \Pi^{\overline{\lambda}[\mu]}\mathcal{T}^\pi\theta^{\overline{\lambda}[\mu]} = \theta^{\overline{\lambda}[\mu]},$$

where the inequality follows from the fact that since $\theta^{\overline{\lambda}[\mu]}$ is continuous in μ, for μ sufficiently close to J there is a flat region of $F_{(\mathcal{T}^\pi\theta^{\overline{\lambda}[\mu]})(x)}$ at level τ_i, for any $(x,i) \in J$. To complete the injectivity argument, we cannot have $\mathrm{P}_J\theta^{\overline{\lambda}[\mu']} = \mathrm{P}_J\theta^{\overline{\lambda}[\mu]}$ if $\theta^{\overline{\lambda}[\mu']} \neq \theta^{\overline{\lambda}[\mu]}$, as the contraction maps $\Pi^{\overline{\lambda}[\mu]}\mathcal{T}^\pi$ and $\Pi^{\overline{\lambda}[\mu]}\mathcal{T}^\pi$ are equal on coordinates not in J, and these two maps would therefore have the same fixed point, a contradiction.

We may now appeal to the invariance of domain theorem (Brouwer, 1912) to deduce that since $h_{\overline{\lambda}}$ is a continuous injective map between an open subset of $[0,1]^J$ containing $\overline{\lambda}_J$ (here we are using the assumption that $\overline{\lambda}_J$ lies in the interior of $[0,1]^J$) and the Euclidean space $\mathbb{R}^J$ of equal dimension, it is an open map on this domain; that is, it maps open sets to open sets. Hence, we can perturb $\theta^{\overline{\lambda}}$ in the J coordinates in any direction we want by locally modifying the J coordinates of $\overline{\lambda}$. In particular, we can move all J coordinates of $\theta^{\overline{\lambda}}$ closer to those of $(\vartheta_{t+s}(x,i) : (x,i) \in J)$. Let $\lambda' \in (0,1)^{\mathcal{X}\times[m]}$ be such a modification of $\overline{\lambda}$, taken to be close enough to $\overline{\lambda}$ so that all coordinates outside J have sufficiently small perturbations so that they cannot be λ'-argmax indices with respect to ϑ_{t+s}. We then have that $\|\vartheta_{t+s} - \theta^{\lambda'}\|_\infty < \|\vartheta_t - \theta^\lambda\|_\infty$, as required.

A.4.3 Case 2: $\overline{\lambda}_J$ is on the Boundary of $[0,1]^J$

In the more general case when $\overline{\lambda}_J$ may lie on the boundary of $[0,1]^J$, we can apply the same argument to an extension of the function $h_{\overline{\lambda}}$, by increasing its domain from $[0,1]^J$ to an open neighbourhood of this domain in $\mathbb{R}^J$. We define this extension simply by extending the definition of Π^λ in Equation (12) to allow coordinates of λ to lie outside the range $[0,1]$. We lose the non-expansiveness of Π^λ (in L^∞) under this extension, but if $\lambda_{\min}, \lambda_{\max}$ are the minimum and maximum coordinates of λ, respectively, it is easy verified (by modifying the proof of Proposition 6) that Π^λ is $\max(1-\lambda_{\min}, \lambda_{\max})$-Lipschitz, and so if we extend the function to a domain where $\lambda_{\max}, 1-\lambda_{\min} \leq \gamma^{-1/2}$, the composition $\Pi^\lambda\mathcal{T}^\pi$ is a $\gamma^{1/2}$-contraction in L^∞, and hence has a unique fixed point θ^λ.

By the same arguments as above, the extended map $h_{\overline{\lambda}}$ is continuous and injective in a neighbourhood of $\overline{\lambda}_J$ on this extended domain, and hence we may again apply the invariance of domain theorem to obtain that $h_{\overline{\lambda}_{\setminus J}}$ is locally surjective around $\overline{\lambda}_J$. However, since $\overline{\lambda}_J$ lies on the boundary of the original domain, we must additionally check that we can perturb

$\overline{\lambda}_J$ to obtain μ in such a way that we obtain the desired perturbation of $\theta^{\overline{\lambda}}$, without the parameters μ leaving the set $[0,1]^J$. To do this, we first rule out $\overline{\lambda}_J$ lying on certain parts of the boundary.

Lemma 25 *If $(x,i) \in J$ and $\vartheta_{t+s}(x,i) < \theta^{\overline{\lambda}}(x,i)$, then $\overline{\lambda}(x,i) > 0$. Similarly, if $\vartheta_{t+s}(x,i) > \theta^{\overline{\lambda}}(x,i)$, then $\overline{\lambda}(x,i) < 1$.*

Proof We prove the claim when $\vartheta_{t+s}(x,i) < \theta^{\overline{\lambda}}(x,i)$; the other case follows analogously. If $\overline{\lambda}(x,i) = 0$, then since $\vartheta_{t+s}(x,i)$ corresponds to the flat region at level τ_i of the CDF $F_{(\mathcal{T}^\pi \vartheta_{t+s})(x)}$, we must have $(\Pi^{\overline{\lambda}} \mathcal{T}^\pi \vartheta_{t+s})(x,i) \leq \vartheta_{t+s}(x,i)$ since $\overline{\lambda}(x,i) = 0$, and so the chosen quantile at level τ_i by the projection $\Pi^{\overline{\lambda}}$ is the left-most point of this flat region. We therefore have

$$\overline{w}_\infty(\Pi^{\overline{\lambda}} \mathcal{T}^\pi \vartheta_t, \theta^{\overline{\lambda}}) \geq |(\Pi^{\overline{\lambda}} \mathcal{T}^\pi \vartheta_t)(x,i) - \theta^{\overline{\lambda}}(x,i)| \geq |\vartheta_t(x,i) - \theta^{\overline{\lambda}}(x,i)| = \overline{w}_\infty(\vartheta_t, \theta^{\overline{\lambda}}),$$

contradicting contractivity of $\Pi^{\overline{\lambda}} \mathcal{T}^\pi$ around $\theta^{\overline{\lambda}}$. ■

We write $v = \operatorname{sign}((\vartheta_{t+s})_J - \theta^{\overline{\lambda}}_J) \in \mathbb{R}^J$, where the sign mapping is applied elementwise, and introduce the notation $\mathrm{N}(v) = \{\alpha \odot v : \alpha \in \mathbb{R}^n_{>0}\}$ for the (open) orthant containing the vector v. We are therefore seeking a perturbation μ of $\overline{\lambda}_J$ such that $\theta^{\overline{\lambda}[\mu]}_J$ lies in a direction in $\mathrm{N}(v)$ from $\theta^{\overline{\lambda}}_J$, and further such that the perturbation to $\theta^{\overline{\lambda}}$ is sufficiently small that no index that was not an argmax in $\|\vartheta_{t+s} - \theta^{\overline{\lambda}}\|_\infty$ can become one in $\|\vartheta_{t+s} - \theta^{\overline{\lambda}[\mu]}\|_\infty$; under these conditions, we have $\|\vartheta_{t+s} - \theta^{\overline{\lambda}[\mu]}\|_\infty < \|\vartheta_{t+s} - \theta^{\overline{\lambda}}\|_\infty$, as required. Lemma 25 then guarantees that a (sufficiently small) perturbation of $\overline{\lambda}_J$ in any direction in $\mathrm{N}(v)$ remains within $[0,1]^J$, so it is sufficient to show that a perturbation in such a direction achieves the desired perturbation of $\theta^{\overline{\lambda}}$.

Differentiability. Now, if the extended map $\lambda \mapsto \theta^\lambda$ is differentiable at $\overline{\lambda}$, then differentiating through the fixed-point equation $\theta^\lambda = G(\lambda, \theta^\lambda)$ (where we write $G(\lambda, \theta) = \Pi^\lambda \mathcal{T}^\pi \theta$ for conciseness) yields

$$\nabla_\lambda \theta^\lambda = \partial_\lambda G(\lambda, \theta^\lambda) + \partial_\theta G(\lambda, \theta^\lambda) \nabla_\lambda \theta^\lambda ;$$

differentiability of G in θ results from differentiability of the map $\lambda \mapsto G(\lambda, \theta^\lambda)$, and continuous differentiability of G in λ. Since $\theta \mapsto G(\lambda, \theta)$ is contractive in L^∞ with factor $\gamma^{1/2}$ (on the extended domain), and by coordinatewise monotonicity of $\theta \mapsto G(\lambda, \theta)$, it follows that $\partial_\theta G(\lambda, \theta^\lambda)$ is non-negative and strictly substochastic, with row L^1 norms bounded by $\gamma^{1/2}$, the contraction factor for the extended set of contraction mappings. We remark as a point of independent interest that this is a kind of Bellman equation for $\nabla_\lambda \theta^\lambda$, with $\partial_\theta G(\lambda, \theta^\lambda)$ taking the role of the transition matrix, and $\partial_\lambda G(\lambda, \theta^\lambda)$ taking the role of a collection of cumulants; in fact, the structure of $\partial_\theta G(\lambda, \theta^\lambda)$ coincides with the local quantile back-up diagrams described in Example 23. We therefore have

$$\nabla_\lambda \theta^\lambda = (I - \partial_\theta G(\lambda, \theta^\lambda))^{-1} \partial_\lambda G(\lambda, \theta^\lambda) .$$

By extracting the principal submatrix on the J coordinates, we obtain a derivative for $h_{\overline{\lambda}}(\overline{\lambda}_J)$. The following lemma is useful in reasoning about the structure of this principal submatrix.

Lemma 26 *Let $Q_1 \in \mathbb{R}^{n\times n}$ be strictly substochastic, and let $K \subseteq [n]$. Then the principal submatrix on the K coordinates of $(I - Q_1)^{-1}$ can be expressed as $(I - Q_2)^{-1}$, with $Q_2 \in \mathbb{R}^{K\times K}$ strictly substochastic.*

Proof We interpret Q_1 as the transition matrix of a Markov chain $(Z_t)_{t\geq 0}$ that includes a non-zero probability of termination at each state. Each row of the matrix $(I - Q_1)^{-1}$ is then the pre-termination visitation measure associated with a particular initial state in the Markov chain. Now let Q_2 be the strictly substochastic matrix defined by

$$Q_2(z_1, z_2) = \mathbb{P}((Z_t)_{t\geq 0} \text{ does not terminate before returning to } K, \\ \text{first state on return is } z_2 \mid Z_0 = z_1)\,.$$

By construction, the pre-termination visitation distribution $(I - Q_2)^{-1}$ is identical to the principal submatrix of $(I - Q_1)^{-1}$ on the K coordinates, as required. ■

From Lemma 26, we therefore obtain that $\nabla h_{\overline{\lambda}}(\overline{\lambda}_J)$ has the form

$$\nabla h_{\overline{\lambda}}(\overline{\lambda}_J) = (I - Q)^{-1} D\,,$$

with $D \in \mathbb{R}^{J\times J}$ diagonal, with positive elements on the diagonal (from monotonicity of $\lambda \mapsto G(\lambda, \theta)$), with $Q \in \mathbb{R}^{J\times J}$ strictly substochastic. The derivative is therefore invertible, and we obtain the derivative of the inverse of the form

$$\nabla h^{-1}_{\overline{\lambda}}(\theta^\lambda_J) = D^{-1}(I - Q)\,.$$

From strict substochasticity of Q, and since $v \in \{\pm 1\}^J$, it follows that for the desired perturbation direction v, we have

$$\nabla h^{-1}_{\overline{\lambda}}(\theta^\lambda_J) v \overset{\mathrm{S}}{=} v\,,$$

and so $\nabla h^{-1}_{\overline{\lambda}}(\theta^\lambda_J) v \in \mathrm{N}(v)$, where the equality of signs applies elementwise. Therefore, a perturbation of $\overline{\lambda}_J$ in a direction in $\mathrm{N}(v)$ is achieved by a sufficiently small perturbation of $\overline{\lambda}_J$ in a direction in $\mathrm{N}(v)$, as required.

Non-differentiability. If $\lambda \mapsto \theta^\lambda$ is not differentiable at $\overline{\lambda}$, we instead use techniques from non-smooth analysis to complete the argument. First, since $\lambda \mapsto \theta^\lambda$ is Lipschitz (by Lemma 24), it is differentiable almost everywhere by Rademacher's theorem (Rademacher, 1919). By adapting the argument made by Clarke (1976, Lemma 3), by Fubini's theorem, for almost all $\lambda_{\setminus J} \in \mathbb{R}^{\mathcal{X}\times[m]\setminus J}$, the map $\lambda \mapsto \theta^\lambda$ is differentiable at $(\lambda_{\setminus J}, \mu)$ for almost all μ with $(\lambda_{\setminus J}, \mu)$ in the extended domain. The map $(\lambda_{\setminus J}, \mu) \mapsto (\lambda_{\setminus J}, h_{\lambda_{\setminus J}}(\mu))$ is Lipschitz and locally injective around $\overline{\lambda}$, and hence maps sufficiently small open neighbourhoods of $\overline{\lambda}$ to open neighbourhoods of $(\overline{\lambda}_{\setminus J}, h_{\overline{\lambda}_{\setminus J}}(\overline{\lambda}_J))$. Further, since each $h_{\lambda_{\setminus J}}$ is Lipschitz, and so

absolutely continuous, the inverse map $h^{-1}_{\lambda_{\setminus J}}$ is almost-everywhere differentiable within such a neighbourhood. Following the analysis of the differentiable case, we therefore deduce

$$\nabla h^{-1}_{\lambda_{\setminus J}}(\theta)v \overset{\mathrm{S}}{=} v$$

for almost all $\lambda_{\setminus J}$ in a ball B around $\overline{\lambda}_{\setminus J}$, and (for each such $\lambda_{\setminus J}$) for almost all θ in the L^∞ ball B' with centre $h_{\overline{\lambda}_{\setminus J}}(\overline{\lambda}_J)$ and radius ρ, for some radius $\rho > 0$. We further take B and B' to be of small enough radii so that this directional derivative is bounded on this set, so that h^{-1}_{λ} is locally Lipschitz on B' for each $\lambda_{\setminus J} \in B$ (and hence absolutely continuous), and so that for any $\theta \in B'$, we have $\operatorname{sign}(\theta - (\vartheta_{t+s})_J) = \operatorname{sign}(\theta^{\overline{\lambda}}_J - (\vartheta_{t+s})_J)$, and so that for no μ in the preimage of B' under $h_{\lambda_{\setminus J}}$ can have that $\|\theta^{(\lambda_{\setminus J},\mu)} - \vartheta_{t+s}\|_\infty$ has new argmax coordinates outside of J.

Let us consider $\tilde{\lambda} \in B$ at which the almost-everywhere differentiability condition holds. By applying the same argument with Fubini's theorem, for almost all $\bar{\theta}$ in $B(h_{\tilde{\lambda}}(\overline{\lambda}_J), \rho/4)$, the inverse $h^{-1}_{\tilde{\lambda}}$ is differentiable almost everywhere on $\{\bar{\theta} + uv : u \in [0, \rho/2]\}$.

Now, defining $\mu_\tau = h^{-1}_{\tilde{\lambda}}(\bar{\theta} + \tau v)$ for $\tau \in [0, \rho/2]$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mu_\tau = \nabla h^{-1}_{\tilde{\lambda}}(\bar{\theta} + \tau v)v$$

for almost all τ, and by absolute continuity of $h^{-1}_{\tilde{\lambda}}$, it follows that

$$\mu_{\rho/2} = \mu_0 + \int_0^{\rho/2} \frac{\mathrm{d}}{\mathrm{d}\tau}\mu_\tau \,\mathrm{d}\tau\,.$$

Hence, $\mu_\varepsilon - \mu_0 \in \mathrm{N}(v)$, and by construction $h_{\tilde{\lambda}}(\mu_{\rho/2}) = \bar{\theta} + \rho v/2$. By continuity of $\lambda \mapsto \theta^\lambda$ and its inverse, and since $\tilde{\lambda}$ and $\bar{\theta}$ can be chosen above to be arbitrarily close to $\overline{\lambda}_{\setminus J}$ and $h_{\overline{\lambda}_{\setminus J}}(\overline{\lambda}_J)$ respectively, we may consider a sequence of these parameters converging to $\overline{\lambda}_{\setminus J}$ and $h_{\overline{\lambda}_{\setminus J}}(\overline{\lambda}_J)$, such that the values of $\mu_{\rho/2}$ as constructed above also converge (by compactness), and thus conclude the existence of $\bar{\mu}_{\rho/2}$ such that $\bar{\mu}_{\rho/2} - \overline{\lambda}_J \in \overline{\mathrm{N}(v)}$, and $h_{\overline{\lambda}}(\bar{\mu}_{\rho/2}) = h_{\overline{\lambda}}(\overline{\lambda}_J) + \rho v/2$, as required.

A.5 Proof of Proposition 19

We begin with the observation that for any return-distribution function $\eta \in \mathscr{P}([V_{\mathrm{MIN}}, V_{\mathrm{MAX}}])$, for the projection Π^λ onto $\mathscr{F}_{\mathrm{Q},m}$ (for any $\lambda \in [0,1]^{\mathcal{X}\times[m]}$), we have

$$\bar{w}_1(\Pi^\lambda\eta, \eta) \leq \frac{V_{\mathrm{MAX}} - V_{\mathrm{MIN}}}{2m}\,.$$

Using this observation, we have

$$\begin{aligned}
\bar{w}_1(\hat{\eta}^\pi_\lambda, \eta^\pi) &\overset{(a)}{\leq} \bar{w}_1(\hat{\eta}^\pi_\lambda, \mathcal{T}^\pi\hat{\eta}^\pi_\lambda) + \bar{w}_1(\mathcal{T}^\pi\hat{\eta}^\pi_\lambda, \eta^\pi) \\
&\overset{(b)}{=} \bar{w}_1(\Pi^\lambda\mathcal{T}^\pi\hat{\eta}^\pi_\lambda, \mathcal{T}^\pi\hat{\eta}^\pi_\lambda) + \bar{w}_1(\mathcal{T}^\pi\eta^\pi, \mathcal{T}^\pi\eta^\pi) \\
&\overset{(c)}{\leq} \frac{V_{\mathrm{MAX}} - V_{\mathrm{MIN}}}{2m} + \gamma\bar{w}_1(\hat{\eta}^\pi_\lambda, \eta^\pi)\,.
\end{aligned}$$

Here, (a) follows from the triangle inequality, (b) follows as $\hat{\eta}_\lambda^\pi$, η^π are fixed points of $\Pi^\lambda \mathcal{T}^\pi$, $\mathcal{T}^\pi$, respectively, and (c) follows from the application of the inequality at the beginning of the proof and contractivity of $\mathcal{T}^\pi$. Rearranging then gives the desired result.

A.6 Proof of Proposition 21

From the assumptions of the proposition, we have $\hat{\eta}_\lambda^\pi = (\Pi^\lambda \mathcal{T}^\pi)^k \eta^\pi$. Then observe that following the argument for the proof of Proposition 19, we have, for any $l \in \{1, \dots, k\}$,

$$\bar{w}_1((\Pi^\lambda \mathcal{T}^\pi)^l \eta^\pi, \eta^\pi) \leq \gamma \bar{w}_1((\Pi^\lambda \mathcal{T}^\pi)^{l-1} \eta^\pi, \eta^\pi) + \frac{V_{\text{MAX}} - V_{\text{MIN}}}{2m}.$$

Chaining these inequalities yields the required statement.

Appendix B. Implementations of Quantile Dynamic Programming

Here, we describe two concrete implementations of QDP, which may be of independent interest to the reader. Algorithm 3 (Bellemare et al., 2023) describes an implementation when the reward distributions are available as input to the algorithm as a list of outcomes and probabilities.

Algorithm 3 Quantile dynamic programming (finitely-supported rewards)

Require: Quantile estimates $((\theta(x,i)_{i=1}^m : x \in \mathcal{X})$,
Transition and reward probabilities $(P^\pi(x', r \mid x) : x, x' \in \mathcal{X})$,
Interpolation parameters $\lambda \in [0,1]^{\mathcal{X} \times [m]}$.

for $x \in \mathcal{X}$ **do**
  Set `Targets` as empty list {List of outcome/probability pairs}
  for $x' \in \mathcal{X}$ **do**
    for $r \in \mathscr{R}$ **do**
      for $j = 1, \dots, m$ **do**
        Append $(r + \gamma\theta(x', j), P^\pi(x', r|x)/m)$ to `Targets`
      end for
    end for
  end for
  Sort `Targets` ascending according to outcomes.
  for $i = 1, \dots, m$ **do**
    Find minimal outcome q' such that cumulative probability is $\geq {}^{2i-1}/_{2m}$.
    Set $\theta'(x, i) \leftarrow q'$.
  end for
end for
return $((\theta'(x,i)_{i=1}^m : x \in \mathcal{X})$

Algorithm 4 makes use of a root-finding subroutine (such as `scipy.optimize.root_scalar`), and can be used when the CDFs of the reward distributions are available as input, and can be queried at individual points. A common use case for this implementation is the case of Gaussian rewards. Note that the root-finding subroutine is called on a monotonic scalar

function, and therefore strong guarantees can be given on the approximate solution returned when the reward CDFs of the MDP are continuous. Nevertheless, note that Algorithm 4 does not exactly implement the operator $\Pi^{\lambda}\mathcal{T}^{\pi}$ due to this root-finding approximation error. For simplicity, we present the algorithm in the case where the reward and next state in a transition are conditionally independent given the current state, though the algorithm can be straightforwardly extended to the general case, by working with CDFs of reward distributions conditioned on the next state.

Algorithm 4 Quantile dynamic programming (reward CDFs)

Require: Quantile estimates $((\theta(x,i)_{i=1}^{m} : x \in \mathcal{X})$,
Transition probabilities $(P^{\pi}(x' \mid x) : x, x' \in \mathcal{X})$,
Reward CDFs $(F_{\mathcal{R}^{\pi}(x)} : x \in \mathcal{X})$.

1: **for** $x \in \mathcal{X}$ **do**
2: Construct function

$$\phi_x : t \mapsto \sum_{x' \in \mathcal{X}} P^{\pi}(x'|x) \sum_{j=1}^{m} F_{\mathcal{R}^{\pi}(x)}(t - \gamma\theta(x', j))$$

3: **for** $i = 1, \ldots, m$ **do**
4: Use a scalar root-finding subroutine to find $\theta'(x,i)$ approximately satisfying

$$\phi_x(\theta'(x,i)) = \tau_i$$

5: **end for**
6: **end for**
7: **return** $((\theta'(x,i)_{i=1}^{m} : x \in \mathcal{X})$

Appendix C. Convergence of Asynchronous QTD Updates

Here, we describe the key considerations in extending our analysis to a proof of convergence for asynchronous versions of QTD; our discussion follows the approach of Perkins and Leslie (2013).

Step size restrictions. Typically, more restrictive assumptions on step sizes, beyond the Robbins-Monro conditions, are required for asynchronous convergence guarantees. See, for example, Assumption A2 of Perkins and Leslie (2013); note that the typical Robbins-Monro step size schedule of $\alpha_k \propto 1/k^{\rho}$ for $\rho \in (1/2, 1]$ satisfies these requirements.

Conditions on the sequence of states $(X_k)_{k\geq 0}$ to be updated. Additionally, different states are required to be updated "comparably often"; assuming that $(X_k)_{k\geq 0}$ forms an aperiodic irreducible time-homogeneous Markov chain is sufficient, and this conditions holds when either (i) π generates such a Markov chain over the state space of the MDP of interest, or (ii) when the states to be updated are sampled i.i.d. from a fixed distribution supported on the entirety of the state space, amongst other settings. See Assumption A4 of Perkins and Leslie (2013) for further details.

Modified differential inclusion. The QTD differential inclusion in Equation (17) must be broadened to account for the possibility of different states being updated with different frequencies, leading to a differential inclusion of the form

$$\partial_t \vartheta_t(x,i) \in \{\omega h : \omega \in (\delta, 1], h \in H^{\pi}_{x,i}(\vartheta_t)\},$$

where δ represents a minimum relative update frequency for the state x, derived from the conditions on $(X_k)_{k\geq 0}$ described above. Because of the structure of the Lyapunov function for the QTD DI in Equation (20), it is readily verified that this remains a valid Lyapunov function for this broader differential inclusion, for the same invariant set of QDP fixed points.

References

Jean-Pierre Aubin. *Viability theory.* Springer Birkhauser, 1991.

Jean-Pierre Aubin and Arrigo Cellina. *Differential inclusions: Set-valued maps and viability theory.* Springer Science & Business Media, 1984.

Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *Proceedings of the International Conference on Learning Representations*, 2018.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013.

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

Marc G. Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning.* MIT Press, 2023.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de Probabilites XXXIII*, pages 1–68. Springer, 1999.

Michel Benaïm and Mathieu Faure. Consistency of vanishingly smooth fictitious play. *Mathematics of Operations Research*, 38(3):437–450, 2013.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions, Part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*. Springer Science & Business Media, 2012.

Mario Bernardo, Chris Budd, Alan Richard Champneys, and Piotr Kowalczyk. *Piecewise-smooth dynamical systems: Theory and applications*. Springer Science & Business Media, 2008.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

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

Vivek S. Borkar. Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851, 1998.

Vivek S. Borkar. *Stochastic approximation: A dynamical systems viewpoint*. Springer, 2008.

Vivek S. Borkar and Sean P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38 (2):447–469, 2000.

Luitzen E. J. Brouwer. Beweis der Invarianz des n-dimensionalen Gebiets. *Mathematische Annalen*, 71(3):305–313, 1912.

George W. Brown. Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, 13(1):374, 1951.

Francis H. Clarke. On the inverse function theorem. *Pacific Journal of Mathematics*, 64 (1):97–102, 1976.

Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth analysis and control theory*. Springer Science & Business Media, 1998.

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

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

Peter Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8(3-4):341–362, 1992.

Peter Dayan and Terrence J. Sejnowski. TD(λ) converges with probability 1. *Machine Learning*, 14(3):295–301, 1994.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis, and Pushmeet Kohli. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610 (7930):47–53, 2022.

A. F. Filippov. Differential equations with discontinuous right-hand side. *Mat. Sb. (N.S.)*, 51(93):99–128, 1960.

Hugo Gilbert and Paul Weng. Quantile reinforcement learning. In *Proceedings of the Asian Workshop on Reinforcement Leanring*, 2016.

Aditya Gopalan and Gugan Thoppe. Demystifying approximate value-based RL with ϵ-greedy exploration: A differential inclusion analysis. *arXiv*, 2023.

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

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994.

Ajin George Joseph and Shalabh Bhatnagar. An adaptive and incremental approach to quantile estimation. In *IEEE Conference on Decision and Control*, 2019.

Michael J. Kearns and Satinder Singh. Bias-variance error bounds for temporal difference updates. In *Proceedings of the Conference on Learning Theory*, 2000.

Roger Koenker. *Quantile regression*. Econometric Society Monographs. Cambridge University Press, 2005.

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

Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of quantile regression*. CRC press, 2017.

Nikolai Nikolaevich Krasovskii and Andrej I. Subbotin. *Game-theoretical control problems*. Springer, 1988.

Harold Kushner and Dean Clark. *Stochastic approximation methods for constrained and unconstrained systems*. Springer, 1978.

Harold J. Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer Science & Business Media, 2003.

David S. Leslie and Edmund J. Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.

Alix Lhéritier and Nicolas Bondoux. A Cramér distance perspective on quantile regression based distributional reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2022.

Xiaocheng Li, Huaiyang Zhong, and Margaret L Brandeau. Quantile Markov decision processes. *Operations Research*, 70(3):1428–1447, 2022.

Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, 22(4):551–575, 1977.

Yudong Luo, Guiliang Liu, Haonan Duan, Oliver Schulte, and Pascal Poupart. Distributional reinforcement learning with monotonic splines. In *Proceedings of the International Conference on Learning Representations*, 2021.

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

Sean Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.

Manuel D. P. Monteiro Marques. *Differential inclusions in nonsmooth mechanical problems: Shocks and dry friction*. Birkhäuser, 2013.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return density estimation 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.

Steven Perkins and David S. Leslie. Asynchronous stochastic approximation with differential inclusions. *Stochastic Systems*, 2(2):409–446, 2013.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Hans Rademacher. Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variabeln und über die Transformation der Doppelintegrale. *Mathematische Annalen*, 79(4):340–359, 1919.

Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2): 296–301, 1951.

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

Mark Rowland, Yunhao Tang, Clare Lyle, Rémi Munos, Marc G Bellemare, and Will Dabney. The statistical benefits of quantile temporal-difference learning for value estimation. In *Proceedings of the International Conference on Machine Learning*, 2023.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University, 1988.

Georgi V. Smirnov. *Introduction to the theory of differential inclusions*. American Mathematical Society, 2002.

Richard S. Sutton. *Temporal credit assignment in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 1984.

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, 2018.

John N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

Cédric Villani. *Optimal transport: Old and new*. Springer, 2009.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Christopher J. C. H. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge, 1989.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4): 279–292, 1992.

Tadeusz Wazewski. Systemes de commande et equations au contingent. *Bulletin de l'Academie Polonaise des Sciences. Serie des Sciences Mathematiques, Astronomiques et Physiques*, 9:151–155, 1961.

Peter R. Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J. Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, Leilani Gilpin, Piyush Khandelwal, Varun Kompella, HaoChih Lin, Patrick MacAlpine, Declan Oller, Takuma Seno, Craig Sherstan, Michael D. Thomure, Houmehr Aghabozorgi, Leon Barrett, Rory Douglas, Dion Whitehead, Peter Dürr, Peter Stone, Michael Spranger, and Hiroaki Kitano. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, 2019.

Fan Zhou, Jianing Wang, and Xingdong Feng. Non-crossing quantile regression for distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.