An Analysis of Categorical Distributional Reinforcement Learning

Mark Rowland[*1] **Marc G. Bellemare**[†] **Will Dabney**[‡] **Rémi Munos**[‡] **Yee Whye Teh**[‡]
[*]University of Cambridge, [†]Google Brain, [‡]DeepMind

Abstract

Distributional approaches to value-based reinforcement learning model the entire distribution of returns, rather than just their expected values, and have recently been shown to yield state-of-the-art empirical performance. This was demonstrated by the recently proposed `C51` algorithm, based on categorical distributional reinforcement learning (CDRL) [Bellemare et al., 2017a]. However, the theoretical properties of CDRL algorithms are not yet well understood. In this paper, we introduce a framework to analyse CDRL algorithms, establish the importance of the projected distributional Bellman operator in distributional RL, draw fundamental connections between CDRL and the Cramér distance, and give a proof of convergence for sample-based categorical distributional reinforcement learning algorithms.

1 INTRODUCTION

Reinforcement learning (RL) formalises the problems of evaluation and optimisation of an agent's behaviour while interacting with an environment, based upon feedback given through a reward signal [Sutton and Barto, 1998]. A major paradigm for solving these problems is value-based RL, in which the agent predicts the *expected return* – i.e. the expected discounted sum of rewards – in order to guide its behaviour. The moments or distribution of the random return have also been considered in the literature, with a variety of approaches proposing algorithms for estimating more complex distributional information [Morimura et al., 2010b,a, Prashanth and Ghavamzadeh, 2013, Tamar et al., 2016]. Recently, Bellemare et al. [2017a] used the distributional perspective to propose an algorithm, `C51`, which achieved state-of-the-art performance on the Atari 2600 suite of benchmark tasks. `C51` is a deep RL algorithm based on categorical policy evaluation (for evaluation) and categorical Q-learning (for control), also introduced by Bellemare et al. [2017a], and it is these latter two algorithms which are at the centre of our study. We refer to these approaches as categorical distributional reinforcement learning (CDRL).

Given a state x and action a, `C51` approximates the distribution over returns using a uniform grid over a fixed range, i.e. a *categorical* distribution with evenly-spaced outcomes. Analogous to how value-based approaches such as SARSA [Rummery and Niranjan, 1994] learn to predict, `C51` also forms a learning target from sample transitions: reward, next state, and eventually next-state distribution over returns. However, the parallel ends here: because `C51` learns a distribution, it minimises the Kullback-Leibler divergence between its target and its prediction, rather than the usual squared loss. However, the support of the target is in general disjoint from the approximation support; to account for this, Bellemare et al. [2017a] further introduced a projection step normally absent from reinforcement learning algorithms.

As a whole, the particular techniques incorporated in `C51` are not explained by the accompanying theory. While the "mean process" which governs learning within `C51` is described by a contractive distributional Bellman operator, there are not yet any guarantees on the behaviour of sample-based algorithms. To put things in context, such guarantees in case of estimating expected returns require a completely different mathematical formalism [Tsitsiklis, 1994, Jaakkola et al., 1994]. The effect of the discrete approximation and its corresponding projection step also remain to be quantified. In this paper we analyse these issues.

At the centre of our analysis is the Cramér distance between probability distributions. The Cramér distance is of particular interest as it was recently shown to possess many of the same properties as the Wasserstein metric, used to show the contractive nature of

[1]Contributed during an internship at DeepMind.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

the distributional Bellman operator [Bellemare et al., 2017b]. Specifically, using the Cramér distance, we: (i) quantify the approximation error arising from the discrete approximation in CDRL (see Section 4.2); and (ii) develop stochastic approximation results for the sample-based case (see Section 4.3).

One of the main contributions of this paper is to establish a framework for the analysis of CDRL algorithms. This framework reveals a space of possible alternative methods (Sections 3 and 4). We also demonstrate that the fundamental property required for the convergence of distributional RL algorithms is contractivity of a *projected* Bellman operator, in addition to the contractivity of the Bellman operator itself as in non-distributional RL (Proposition 2). This point has parallels with the importance of the (distinct) projection operator in non-tabular RL [Tsitsiklis and Van Roy, 1997].

We begin, in Section 2, with a general introduction to distributional RL, and establish required notation. In Section 3, we give a detailed description of categorical distributional RL, and set it in the context of a new framework in which to view distributional RL algorithms. Finally, in Section 4, we undertake a detailed convergence analysis of CDRL, dealing with the approximations and parametrisations that typically must be introduced into practical algorithms. This culminates in the first proofs of convergence for sample-based CDRL algorithms.

2 BACKGROUND

2.1 Markov decision processes

We consider a Markov decision process (MDP) with a finite state space $\mathcal{X}$, a finite action space $\mathcal{A}$, and a transition kernel $p : \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathbb{R} \times \mathcal{X})$ that defines a joint distribution over immediate reward and next state given a current state-action pair. We will be concerned with stationary policies $\pi : \mathcal{X} \to \mathscr{P}(\mathcal{A})$ that define a probability distribution over the action space given a current state. The full MDP is given by the collection of random variables $(X_t, A_t, R_t)_{t=0}^{\infty}$, where $(X_t)_{t\geq 0}$ is the sequence of states taken by the environment, $(A_t)_{t\geq 0}$ is the sequence of actions taken by the agent, and $(R_t)_{t\geq 0}$ is the sequence of rewards.

2.2 Return distributions

The *return* of a policy π, starting in initial state $x \in \mathcal{X}$ and initially taking action $a \in \mathcal{A}$, is defined as the random variable given by the sum of discounted rewards:

$$\sum_{t=0}^{\infty} \gamma^t R_t \Big| X_0 = x, A_0 = a\,, \tag{1}$$

where $\gamma \in [0, 1)$ is the discount factor. We may implicitly view the distribution of the returns as being parametrised by π [Sutton et al., 1999]. Two common tasks in RL are (i) *evaluation*, in which the expected value of the return is sought for a fixed policy, and (ii) *control*, in which a policy π^* which maximises the expected value of the returns is sought.

In the remainder of this paper, we will write the *distribution* of the return of policy π and initial state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ as

$$\eta_\pi^{(x,a)} = \mathrm{Law}_\pi \left(\sum_{t=0}^{\infty} \gamma^t R_t \,\Big|\, X_0 = x, A_0 = a \right). \tag{2}$$

We write η_π for the collection of distributions $(\eta_\pi^{(x,a)} | (x, a) \in \mathcal{X} \times \mathcal{A})$. We highlight the change in emphasis from discussing random variables, as in (1), to directly referring to probability distributions in their own right. Although Bellemare et al. [2017a] referred to the object η_π as a *value distribution*, here we favour the more technically correct name *return distribution function*, to highlight that η_π is a function mapping state-action pairs to probability distributions over returns. Referring to return distributions in their own right will lead to a clearer statement of the convergence results that appear in Section 4.

2.3 The distributional Bellman operator

It is well known that expected returns satisfy the Bellman equation [Bellman, 1957, Sutton and Barto, 1998]. Bellemare et al. [2017a] showed that the return distribution function η_π satisfies a distributional variant of the Bellman equation. This result was phrased in terms of equality in distribution between random variables. A similar approach was taken by Morimura et al. [2010a], in which cumulative distribution functions were used. To express the Bellman equation in terms of distributions themselves, we will need the notion of pushforward (or image) measures. We first recall the definition of these measures at the level of generality required by the development of our theory; see Billingsley [1986] for further details.

Definition 1. *Given a probability distribution $\nu \in \mathscr{P}(\mathbb{R})$ and a measurable function $f : \mathbb{R} \to \mathbb{R}$, the pushforward measure $f_\# \nu \in \mathscr{P}(\mathbb{R})$ is defined by $f_\# \nu(A) = \nu(f^{-1}(A))$, for all Borel sets $A \subseteq \mathbb{R}$.*

Intuitively, $f_\# \nu$ is obtained from ν by shifting the support of ν according to the map f. Of particular interest in this paper will be pushforward measures obtained via an affine shift map $f_{r,\gamma} : \mathbb{R} \to \mathbb{R}$, defined by

$f_{r,\gamma}(x) = r + \gamma x$. Such transformations also appear, unnamed, in Morimura et al. [2010b].

Using this notation, we can now restate a fundamental result which was shown by Bellemare et al. [2017a] in the language of random variables. The return distribution function η_π associated with a policy π, defined in (2), satisfies the *distributional Bellman equation*:

$$\eta_\pi^{(x,a)} = (\mathcal{T}^\pi \eta_\pi)^{(x,a)} \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A},$$

where $\mathcal{T}^\pi : \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is the *distributional Bellman operator*, defined by:

$$(\mathcal{T}^\pi \eta)^{(x,a)} \tag{3}$$
$$= \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} (f_{r,\gamma})_\# \eta^{(x',a')} \pi(a'|x') p(\mathrm{d}r, x'|x, a),$$

for all $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$. This equation serves as the basis of distributional RL, just as the standard Bellman equation serves as the basis of non-distributional value-based RL. Bellemare et al. [2017a] established a preliminary theoretical result regarding the contractive properties of the operator $\mathcal{T}^\pi$. To further this analysis, we first require a particular notion of distance between collections of probability distributions, introduced in Bellemare et al. [2017a].

Definition 2. *The p-Wasserstein distance d_p, for $p \geq 1$ is defined on $\mathscr{P}_p(\mathbb{R})$, the set of probability distributions with finite p^{th} moments, by:*

$$d_p(\nu_1, \nu_2) = \left(\inf_{\lambda \in \Lambda(\nu_1,\nu_2)} \int_{\mathbb{R}^2} |x - y|^p \lambda(\mathrm{d}x, \mathrm{d}y) \right)^{1/p},$$

for all $\nu_1, \nu_2 \in \mathscr{P}_p(\mathbb{R})$, where $\Lambda(\nu_1, \nu_2)$ is the set of probability distributions on $\mathbb{R}^2$ with marginals ν_1 and ν_2.

The supremum-p-Wasserstein metric $\overline{d}_p$ is defined on $\mathscr{P}_p(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ by

$$\overline{d}_p(\eta, \mu) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} d_p(\eta^{(x,a)}, \mu^{(x,a)}),$$

for all $\eta, \mu \in \mathscr{P}_p(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$.

With these definitions in hand, we may recall the following result.

Lemma 1 (Lemma 3, Bellemare et al. [2017a])**.** *The distributional Bellman operator $\mathcal{T}^\pi$ is a γ-contraction in $\overline{d}_p$, for all $p \geq 1$. Further, we have, for any initial set of distributions $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$:*

$$(\mathcal{T}^\pi)^m \eta \to \eta_\pi \text{ in } \overline{d}_p, \text{ as } m \to \infty.$$

This motivates *distributional* RL algorithms, which attempt to approximately find η_π by taking some initial estimates of the return distributions $\eta_0 = (\eta_0^{(x,a)} | (x,a) \in \mathcal{X} \times \mathcal{A})$, and iteratively computing a sequence of estimates $(\eta_t)_{t\geq 0}$ by approximating the update step

$$\eta_{t+1} \leftarrow \mathcal{T}^\pi \eta_t \quad \text{for } t = 0, 1, \ldots . \tag{4}$$

There is also a control version of these updates, which seeks to find the return distributions associated with an optimal policy π^*, via the following updates

$$\eta_{t+1} \leftarrow \mathcal{T} \eta_t \quad \text{for } t = 0, 1, \ldots , \tag{5}$$

where $\mathcal{T}$ is the control version of the distributional Bellman operator, defined by

$$(\mathcal{T}\eta)^{(x,a)} = \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} (f_{r,\gamma})_\# \eta^{(x',a^*(x'))} p(\mathrm{d}r, x'|x, a),$$
$$\text{where } a^*(x') \in \arg\max_{a'\in\mathcal{A}} \mathbb{E}_{R\sim\eta^{(x',a')}}[R].$$

An ideal policy evaluation algorithm would iteratively compute the exact updates of (4), and inherit the resulting convergence guarantees from Lemma 1. However, full computation of the distributional Bellman operator on a return distribution function is typically either impossible (due to unknown MDP dynamics), or computationally infeasible [Bertsekas and Tsitsiklis, 1996]. In order to take the full updates in (4) or (5) and produce a practical, scalable distributional RL algorithm, several key approximations are required, namely:

(i) *distribution parametrisation*;
(ii) *stochastic approximation* of the Bellman operator;
(iii) *projection* of the Bellman target distribution;
(iv) *gradient updates* via a loss function.

We discuss each of these approximations in Section 3, at the same time describing our two CDRL algorithms, categorical policy evaluation and categorical Q-learning, in detail with this approximation framework in mind.

3 CATEGORICAL POLICY EVALUATION AND CATEGORICAL Q-LEARNING

Our first contribution is to make explicit the various approximations, parametrisations, and assumptions implicit in CDRL algorithms. Categorical poicy evaluation approximates the update scheme (4); it produces an iterative sequence $(\eta_t)_{t\geq 0}$ of approximate return distribution functions, updating the approximations as shown in Algorithm 1. Figure 1 illustrates the salient points of the algorithm, and contrasts them against the full updates of (4). Algorithm 1 also describes categorical Q-learning, which approximates the

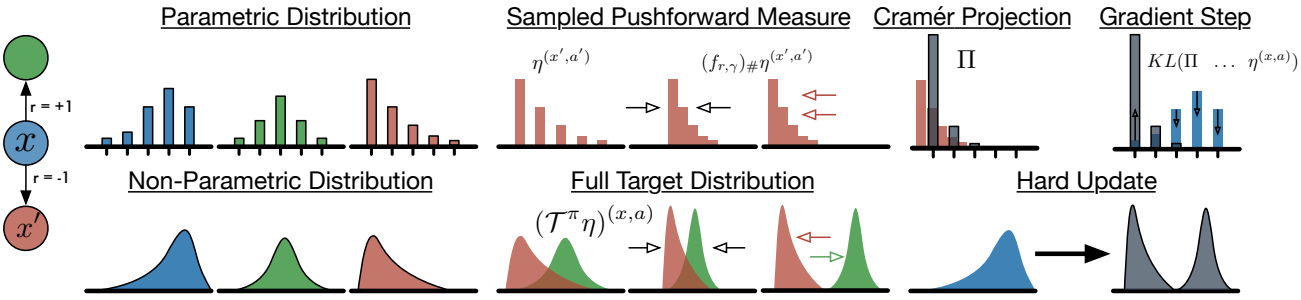

Figure 1: A 3-state MDP with a single action available at each state (shown far left), with full update scheme (4) illustrated on the bottom row, and the categorical policy evaluation update scheme illustrated on the top row. For both update schemes, the current return distribution function is illustrated on the left, the computation of the target distribution in the centre, and finally the update is shown on the right.

full updates in (5). We now discuss the structure of Algorithm 1 in more detail, with reference to the distributional RL framework introduced at the end of Section 2.3.

Algorithm 1 CDRL update [Bellemare et al., 2017a]

Require: $\eta_t^{(x,a)} = \sum_{k=1}^K p_{t,k}^{(x,a)} \delta_{z_k}$ for each (x, a)
1: Sample transition (x_t, a_t, r_t, x_{t+1})
2: # Compute distributional Bellman target
3: **if** Categorical policy evaluation **then**
4: $\quad a^* \sim \pi(\cdot|x_{t+1})$
5: **else if** Categorical Q-learning **then**
6: $\quad a^* \leftarrow \arg\max_a \mathbb{E}_{R \sim \eta_t^{(x_{t+1},a)}}[R]$
7: **end if**
8: $\widehat{\eta}_*^{(x_t,a_t)} \leftarrow (f_{r_t,\gamma})_\# \eta_t^{(x_{t+1},a^*)}$
9: # Project target onto support
10: $\widehat{\eta}_t^{(x_t,a_t)} \leftarrow \Pi_{\mathcal{C}} \widehat{\eta}_*^{(x_t,a_t)}$
11: # Compute KL Loss
12: Find gradient of $\mathrm{KL}(\widehat{\eta}_t^{(x_t,a_t)} || \eta_t^{(x_t,a_t)})$
13: Use gradient to generate new estimate $\eta_{t+1}^{(x_t,a_t)} = \sum_{k=1}^K p_{t+1,k}^{(x_t,a_t)} \delta_{z_k}$
14: **return** $\eta_{t+1}^{(x,a)} = \sum_{k=1}^K p_{t+1,k}^{(x,a)} \delta_{z_k}$ for each (x, a)

3.1 Distribution parametrisation

From an algorithmic perspective, it is impossible to represent the full space of probability distributions $\mathscr{P}(\mathbb{R})$ with a finite collection of parameters. Therefore a first design decision for a general distributional RL algorithm is how probability distributions should be represented in an approximate way. Formally, this requires the selection of a parametric family $\mathcal{P} \subset \mathscr{P}(\mathbb{R})$. CDRL uses the parametric family

$$\mathcal{P} = \left\{ \sum_{i=1}^K p_i \delta_{z_i} \middle| p_1, \ldots, p_K \geq 0, \sum_{k=1}^K p_k = 1 \right\},$$

of *categorical* distributions over some fixed set of equally-spaced supports $z_1 < \cdots < z_K$); see lines 13 and 14 of Algorithm 1. Other parametrisations are of course possible, such as mixtures of Diracs with varying location parameters [Dabney et al., 2018], mixtures of Gaussians, etc.

3.2 Stochastic approximation of Bellman operator

Evaluation of the distributional Bellman operator $\mathcal{T}^\pi$ (see (3)) requires integrating over all possible next state-action-reward combinations. Some approximation is required; a popular way to achieve this in RL is by *sampling* a transition $(x_t, a_t, r_t, x_{t+1}, a^*)$ of the MDP. This is also the approach taken in CDRL, as shown in lines 1-8 of Algorithm 1. Here a^* is selected either by sampling from the policy $\pi(\cdot|x_{t+1})$ in the case of categorical policy evaluation, or as the action with the highest estimated expected returns, in the case of categorical Q-learning. In the context of categorical policy evaluation, this defines a stochastic Bellman operator, given by

$$\begin{aligned} (\widehat{\mathcal{T}}^\pi \eta_t)^{(x_t,a_t)} &= (f_{r_t,\gamma})_\# \eta_t^{(x_{t+1},a^*)}, \\ (\widehat{\mathcal{T}}^\pi \eta_t)^{(x,a)} &= \eta_t^{(x,a)} \qquad \text{if } (x,a) \neq (x_t,a_t), \end{aligned} \tag{6}$$

where the randomness in $\widehat{\mathcal{T}}^\pi$ comes from the randomly sampled transition $(x_t, a_t, r_t, x_{t+1}, a^*)$. Note that this defines a *random measure*, and importantly, this random measure is equal *in expectation* to the true Bellman target $(\mathcal{T}^\pi \eta_t)^{(x_t,a_t)}$.

3.3 Projection of Bellman target distribution

Having computed $(\widehat{\mathcal{T}}^\pi \eta_t)^{(x_t,a_t)}$, this new distribution typically no longer lies in the parametric family $\mathcal{P}$; as shown in (6), the supports of the distributions are transformed by an affine map $f_{r,\gamma}$. We therefore re-

quire a method of mapping the backup distribution function into the parametric family. That is, we require a *projection operator* $\Pi : \mathscr{P}(\mathbb{R}) \to \mathcal{P}$ that may be applied to each real-valued distribution in a return distribution function. CDRL uses the heuristic projection operator $\Pi_{\mathcal{C}}$ (see line 10 of Algorithm 1), which was defined by Bellemare et al. [2017a] as follows for single Dirac measures:

$$\Pi_{\mathcal{C}}(\delta_y) = \begin{cases} \delta_{z_1} & y \leq z_1 \\ \frac{z_{i+1}-y}{z_{i+1}-z_i}\delta_{z_i} + \frac{y-z_i}{z_{i+1}-z_i}\delta_{z_{i+1}} & z_i < y \leq z_{i+1} \\ \delta_{z_K} & y > z_K \end{cases}, \tag{7}$$

and extended affinely to finite mixtures of Dirac measures, so that for a mixture of Diracs $\sum_{i=1}^N p_i \delta_{y_i}$, we have $\Pi_{\mathcal{C}}(\sum_{i=1}^N p_i \delta_{y_i}) = \sum_{i=1}^N p_i \Pi_{\mathcal{C}}(\delta_{y_i})$ - see the right-hand side of Figure 1. In general we will abuse notation, and use $\Pi_{\mathcal{C}}$ to denote the projection operator for individual distributions, and also the operator on return distribution functions $\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathcal{P}^{\mathcal{X}\times\mathcal{A}}$, which applies the former projection to each distribution in the return distribution function.

3.4 Gradient updates

Having computed a stochastic approximation $\widehat{\eta}_t^{(x_t,a_t)} = (\Pi_{\mathcal{C}}\widehat{\mathcal{T}}^{\pi}\eta_t)^{(x_t,a_t)}$ to the full target distribution, the remaining issue is how the next iterate η_{t+1} should be defined. In `C51`, the approach is to perform a single step of gradient descent on the Kullback-Leibler divergence of the prediction $\eta_t^{(x_t,a_t)}$ from the target $\widehat{\eta}_t^{(x_t,a_t)}$:

$$\mathrm{KL}(\widehat{\eta}_t^{(x_t,a_t)} || \eta_t^{(x_t,a_t)}), \tag{8}$$

with respect to the parameters of $\eta_t^{(x_t,a_t)}$ - see line 12 of Algorithm 1. We also consider CDRL algorithms based on a mixture update, described in more detail in Section 4.3. The use of a gradient update, rather than a "hard" update allows for the dissipation of noise introduced in the target by stochastic approximation [Bertsekas and Tsitsiklis, 1996, Kushner and Yin, 2003]. This completes the description of CDRL in the context of the framework introduced at the end of Section 2.3; we now move on to discussing the convergence properties of these algorithms.

4 CONVERGENCE ANALYSIS

The approximations, parametrisations, and heuristics of CDRL discussed in Section 3 yield practical, scalable algorithms for evaluation and control, but the effects of these heuristics on the theoretical guarantees that many non-distributional algorithms enjoy have not yet been addressed. In this section, we set out a variety of theoretical results for CDRL algorithms, and in doing so, emphasise several key ways in which the approximations described in Section 3 must fit together to enjoy good theoretical guarantees.

We begin by drawing a connection between the heuristic projection operator $\Pi_{\mathcal{C}}$ and the Cramér distance in Section 4.1. This connection then paves the way to obtaining the results of Section 4.2, which concern the properties of CDRL policy evaluation algorithms without stochastic approximation and gradient updates, observing only the consequences of the parametrisation and projection steps discussed in Sections 3.1 and 3.3. We then bring these more realistic assumptions into play in Section 4.3, and our analysis culminates in a proof of convergence of categorical policy evaluation and categorical Q-learning in the tabular setting.

4.1 Cramér geometry

We begin by recalling Lemma 1, through which Bellemare et al. [2017a] established that repeated application of the distributional Bellman operator $\mathcal{T}^{\pi}$ to an initial return distribution function guarantees convergence to the true set of return distributions in the supremum-Wasserstein metric. However, once we introduce the parametrisation $\mathcal{P}$ and projection operator $\Pi_{\mathcal{C}}$ of categorica policy evaluation, the operator of concern is now $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$, the composition of the Bellman operator $\mathcal{T}^{\pi}$ with the projection operator $\Pi_{\mathcal{C}}$. Our first result illustrates that the presence of the projection operator is enough to break the contractivity under Wasserstein distances.

Lemma 2. *The operator $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$ is in general not a contraction in $\overline{d}_p$, for $p > 1$.*

Whilst contractivity with respect to $\overline{d}_1$ is in fact maintained, as we shall see there is a much more natural metric, the Cramér distance [Székely, 2002], with which to establish contractivity of the combined operator $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$.

Definition 3. *The Cramér distance ℓ_2 between two distributions $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$, with cumulative distribution functions F_{ν_1}, F_{ν_2} respectively, is defined by:*

$$\ell_2(\nu_1, \nu_2) = \left(\int_{\mathbb{R}} (F_{\nu_1}(x) - F_{\nu_2}(x))^2 \mathrm{d}x\right)^{1/2}.$$

Further, the supremum-Cramér metric $\overline{\ell}_2$ is defined between two distribution functions $\eta, \mu \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ by

$$\overline{\ell}_2(\eta, \mu) = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \ell_2(\eta^{(x,a)}, \mu^{(x,a)}).$$

The Cramér distance was recently studied as an alternative to the Wasserstein distances in the context

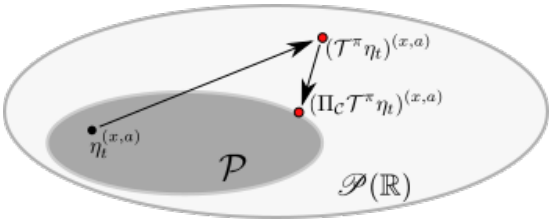

Figure 2: An illustration of the composition of the distributional Bellman operator with the projection $\Pi_\mathcal{C}$, interpreting probability distributions as points in an affine Hilbert space.

of generative modelling [Bellemare et al., 2017b]. The Cramér distance, in fact, induces a useful geometric structure on the space of probability measures. We use this to provide a new interpretation of the heuristic projection $\Pi_\mathcal{C}$ intimately connected with the Cramér distance. The salient points of this connection are stated in Proposition 1, with full mathematical details provided in the corresponding proof in the appendix. We then use this in Section 4.2 to show that $\Pi_\mathcal{C}\mathcal{T}^\pi$ is a contraction in $\bar{\ell}_2$.

Proposition 1. *The Cramér metric ℓ_2 endows a particular subset of $\mathscr{P}(\mathbb{R})$ with a notion of orthogonal projection, and the orthogonal projection onto the subset $\mathcal{P}$ is exactly the heuristic projection $\Pi_\mathcal{C}$. Consequently, $\Pi_\mathcal{C}$ is a non-expansion with respect to ℓ_2.*

A consequence of the result above is the following, which will be useful in later sections.

Lemma 3 (Pythagorean theorem)**.** *Let $\mu \in \mathscr{P}([z_1, z_K])$, and let $\nu \in \mathscr{P}(\{z_1, \ldots, z_K\})$. Then*

$$\ell_2^2(\mu, \nu) = \ell_2^2(\mu, \Pi_\mathcal{C}\mu) + \ell_2^2(\Pi_\mathcal{C}\mu, \nu)\,.$$

A geometric illustration of the action of the composed operator $\Pi_\mathcal{C}\mathcal{T}^\pi$ is given in Figure 2, in light of the interpretation of $\Pi_\mathcal{C}$ as an orthogonal projection.

4.2 Parametrisation and projection

Having established these tools, we can now prove contractivity of the operator $\Pi_\mathcal{C}\mathcal{T}^\pi$, and hence convergence of this variant of distributional RL in the absence of stochastic approximation.

Proposition 2. *The operator $\Pi_\mathcal{C}\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\bar{\ell}_2$. Further, there is a unique distribution function $\eta_\mathcal{C} \in \mathcal{P}^{\mathcal{X}\times\mathcal{A}}$ such that given any initial distribution function $\eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we have*

$$(\Pi_\mathcal{C}\mathcal{T}^\pi)^m \eta_0 \to \eta_\mathcal{C} \text{ in } \bar{\ell}_2 \text{ as } m \to \infty\,.$$

A natural question to ask is how the limiting distribution function $\eta_\mathcal{C}$, established in Proposition 2, differs from the true distribution function η_π. In some sense, this quantifies the "cost" of using the parametrisation $\mathcal{P}$ rather than learning fully non-parametric probability distributions. Reusing the interpretation of $\Pi_\mathcal{C}$ as an orthogonal projection, and using a geometric series argument, we may establish the following result, which echoes existing results for linear function approximation [Tsitsiklis and Van Roy, 1997].

Proposition 3. *Let $\eta_\mathcal{C}$ be the limiting return distribution function of Proposition 2. If $\eta_\pi^{(x,a)}$ is supported on $[z_1, z_K]$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, then we have:*

$$\bar{\ell}_2^2(\eta_\mathcal{C}, \eta_\pi) \leq \frac{1}{1-\gamma} \max_{1\leq i<K} (z_{i+1} - z_i)\,.$$

This establishes that as the fineness of the grid $\{z_1, \ldots, z_K\}$ increases, we gradually recover the true return distribution function. The bound in Proposition 3 relies on a guarantee that the support of the true return distributions lie in the interval $[z_1, z_K]$. Many RL problems come with such a guarantee, but there are also many circumstances where *a priori* knowledge of the scale of rewards is unavailable. It is possible to modify the proof of Proposition 3 to deal with this situation too.

Proposition 4. *Let $\eta_\mathcal{C}$ be the limiting return distribution function of Proposition 2. Suppose $\eta_\pi^{(x,a)}$ is supported on an interval $[z_1 - \delta, z_K + \delta]$ containing $[z_1, z_K]$ for each $(x, a) \in \mathcal{X}\times\mathcal{A}$, and $\eta_\pi^{(x,a)}([z_1-\delta, z_1] \cup [z_K, z_K+\delta]) \leq q$ for some $q \in \mathbb{R}$ and for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ – q bounds the excess mass lying outside the region $[z_1, z_K]$. Then we have*

$$\bar{\ell}_2^2(\eta_\mathcal{C}, \eta_\pi) \leq \frac{1}{1-\gamma} \left(\max_{1\leq i<K} (z_{i+1} - z_i) + 2q^2\delta \right)\,.$$

4.3 Stochastic approximation and gradient updates

In this section, we leverage the theory of stochastic approximation to provide convergence guarantees for sample-based distributional RL algorithms.

We will study a version of categorical policy evaluation that takes a mixture between two distributions, rather than using a KL gradient, as a means of updating the return distribution estimates. The algorithm proceeds by computing the target distribution $\widehat{\eta}_t^{(x_t,a_t)}$ as in Algorithm 1, but then rather than using the gradient of a KL loss, the updated return distribution is produced for some collection of learning rates $(\alpha_t(x,a) | (x,a) \in \mathcal{X}\times\mathcal{A}, t \geq 0)$ according to the following rule:

$$\eta_{t+1}^{(x,a)} \leftarrow (1 - \alpha_t(x,a))\eta_t^{(x,a)} + \alpha_t(x,a)\widehat{\eta}_t^{(x,a)} \;\forall (x,a),$$
$$\text{such that } \alpha_t(x,a) = 0 \text{ if } (x,a) \neq (x_t, a_t)\,. \tag{9}$$

That is, by taking a *mixture* between $\eta_t^{(x_t,a_t)}$ and $\widehat{\eta}_t^{(x_t,a_t)}$. We denote this procedure as Algorithm 2, which for completeness is stated in full in Section 8 of the appendix. The question of whether convergence results hold for the KL update described in Section 3.4 remains open, and is an interesting area for further research.

4.3.1 Convergence of categorical policy evaluation

We first show that, under standard conditions, categorical policy evaluation with the mixture update rule described above is guaranteed to converge to the fixed point of the projected Bellman operator $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$, as described in Proposition 2. We sketch out the main structure of the proof below; the full argument is given in the appendix.

Theorem 1. *In the context of policy evaluation for some policy π, suppose that:*

(i) the stepsizes $(\alpha_t(x,a)|t \geq 0, (x,a) \in \mathcal{X} \times \mathcal{A})$ satisfy the Robbins-Monro conditions:
- $\sum_{t=0}^{\infty} \alpha_t(x,a) = \infty$
- $\sum_{t=0}^{\infty} \alpha_t^2(x,a) < C < \infty$

almost surely, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$;

(ii) we have initial estimates $\eta_0^{(x,a)}$ of the distribution of returns for each state-action pair $(x,a) \in \mathcal{X} \times \mathcal{A}$, each with support contained in $[z_1, z_K]$.

Then, for the updates given by Algorithm 2, in the case of evaluation of the policy π, we have almost sure convergence of η_t to $\eta_{\mathcal{C}}$ in $\overline{\ell}_2$, where $\eta_{\mathcal{C}}$ is the limiting return distribution function of Proposition 2. That is,

$$\overline{\ell}_2(\eta_t, \eta_{\mathcal{C}}) \to 0 \text{ as } t \to \infty \quad \text{almost surely.}$$

The proof follows the approach of Tsitsiklis [1994]; we combine classical stochastic approximation proof techniques with notions of stochastic dominance to prove the almost-sure convergence of the return distribution functions in $\overline{\ell}_2$. Proposition 5 is an interesting result in its own right, as it establishes a formal language to describe the monotonocity of the distributional Bellman operator, which plays an important role in control operators [e.g. Bertsekas, 2012].

We begin by showing that several variants of the Bellman operator are *monotone* with respect to a particular partial ordering over probability distributions known as stochastic dominance [Shaked and Shanthikumar, 1994].

Definition 4. *Given two probability measures $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$, we say that ν_1* stochastically dominates *ν_2, and write $\nu_2 \leq \nu_1$, if there exists a coupling between ν_1 and ν_2 (that is, a probability measure on $\mathbb{R}^2$ with marginals given by ν_1 and ν_2) which is supported on the set $\{(x_1, x_2) \in \mathbb{R}^2 | x_2 \geq x_1\}$. An equivalent characterisation states that $\nu_2 \leq \nu_1$ if for the corresponding CDFs F_{ν_1} and F_{ν_2}, we have*

$$F_{\nu_2}(x) \geq F_{\nu_1}(x) \quad \text{for all } x \in \mathbb{R}\,.$$

Stochastic dominance forms a partial order over the set $\mathscr{P}(\mathbb{R})$. We introduce a related partial order over the space of return distribution functions, $\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, which we refer to as (element-wise) stochastic dominance. Given $\eta, \mu \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$, we say that η stochastically dominates μ element-wise if for each $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\eta^{(x,a)}$ stochastically dominates $\mu^{(x,a)}$.

Proposition 5. *The distributional Bellman operator $\mathcal{T}^{\pi} : \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is a monotone map with respect to the partial ordering on $\mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ given by element-wise stochastic dominance. Further, the Cramér projection $\Pi_{\mathcal{C}} : \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is a monotone map, from which it follows that the Cramér-Bellman operator $\Pi_{\mathcal{C}}\mathcal{T}^{\pi}$ is also monotone.*

The monotonicity of the mappings described in Proposition 5 can then be harnessed to establish a chain of lemmas, given in the appendix, mirroring the chain of reasoning in Tsitsiklis [1994], from which Theorem 1 will follow. In the remainder of this section, we highlight a further important property of the Cramér projection $\Pi_{\mathcal{C}}$ which is crucial in establishing Theorem 1.

We observe from Algorithm 2 that the update rule appearing in Equation (9) can be written

$$\begin{aligned}\eta_{t+1}^{(x,a)} =& \eta_t^{(x,a)} + \alpha_t(x,a)((\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_t)^{(x,a)} - \eta_t^{(x,a)}) \\ &+ \alpha_t(x,a)(\Pi_{\mathcal{C}}(f_{r,\gamma})_{\#}\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_t)^{(x,a)})\,.\end{aligned}$$

for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, for all $t \geq 0$, given that $\alpha_t(x,a) = 0$ if the state $(x,a) \in \mathcal{X} \times \mathcal{A}$ is not selected for update at time t. The second term,

$$\alpha_t(x,a)((\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_t)^{(x,a)} - \eta_t^{(x,a)})\,,$$

may be interpreted as a damped version of the full distributional Bellman update, whilst the third term,

$$\alpha_t(x,a)(\Pi_{\mathcal{C}}(f_{r,\gamma})_{\#}\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_t)^{(x,a)})\,,$$

represents the noise introduced by stochastic approximation. We observe that this noise term is in fact a difference of two probability distributions (one of which is a random measure); thus, this noise term is a particular instance of a random *signed measure*. The Cramér projection leads to an important property of this signed measure, which is crucial in establishing the result of Theorem 1, summarised in Lemma 4.

Lemma 4. *The noise term*

$$\Pi_{\mathcal{C}}(f_{r,\gamma})_{\#}\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_t)^{(x,a)}$$

is a random signed measure with total mass 0 *almost surely, and with the property that when averaged over the next-step reward, state and action tuple* (r, x', a') *it is equal to the zero measure almost surely:*

$$\mathbb{E}_{r,x',a'}\left[(\Pi_{\mathcal{C}}(f_{r,\gamma})_{\#}\eta_t^{(x',a')} - (\Pi_{\mathcal{C}}\mathcal{T}^{\pi}\eta_t)^{(x,a)})\right]((-\infty, y]) = 0\,,$$

for all $y \in \mathbb{R}$.

4.3.2 Convergence of categorical Q-learning

Having established convergence of categorical policy evaluation in Theorem 1, we now leverage this to prove convergence of categorical Q-learning under similar conditions.

Theorem 2. *Suppose that Assumptions (i)–(ii) of Theorem 1 hold, and that all unprojected target distributions* $\widehat{\eta}_*^{(x_t,a_t)}$ *arising in Algorithm 2 are supported within* $[z_1, z_K]$ *almost surely. Assume further that there is a unique optimal policy* π^* *for the MDP. Then, for the updates given in Algorithm 2, in the case of control, we have almost sure convergence of* $(\eta_t^{(x,a)})_{(x,a)\in\mathcal{X}\times\mathcal{A}}$ *in* $\bar{\ell}_2$ *to some limit* $\eta_{\mathcal{C}}^*$*, and furthermore the greedy policy with respect to* $\eta_{\mathcal{C}}^*$ *is the optimal policy* π^*.

Theorem 2 is particularly interesting because it demonstrates that value-based control is not only stable in the distributional case, but also that CDRL *preserves the optimal policy*. This is not a given: for example, if we were to replace $\Pi_{\mathcal{C}}$ with a nearest-neighbour-type projection we could not provide the same guarantee. What makes the CDRL projection step special in this regard is that it preserves the expected value of the unprojected target.

5 DISCUSSION

The `C51` algorithm was empirically successful, but, as we have seen in Lemma 2, is not explained by the initial theoretical results concerning CDRL of Bellemare et al. [2017a]. We have now shown that the projected distributional Bellman operator used in CDRL inherits convergence guarantees from a different metric altogether, the Cramér distance. From Propositions 3 and 4, we see that the limiting approximation error is controlled by the granularity of the parametric distribution and the discount factor γ. Furthermore, we have shown that in the stochastic approximation setting this update converges both for policy evaluation and control.

An important aspect of our analysis is the role of the projection onto the set of parametrised distributions, in distributional RL. Just as existing work has studied the role of the projected Bellman operator in function approximation [Tsitsiklis and Van Roy, 1997], there is a corresponding importance for considering the effects of the projection in distributional RL.

5.1 Function approximation

Our theoretical results in Section 4 treat the problem of tabular distributional RL, with an approximate parametrisation distribution for each state-action pair. Theoretical understanding of function approximation in RL has been the focus of much research, and has significantly improved our understanding of agent behaviour. Although we believe the effects of function approximation on distributional RL are of great theoretical and empirical interest, we leave the function approximation setting as an interesting direction for future work.

5.2 Theoretically grounded algorithms

Turning theoretical results into practical algorithms can often be quite challenging. However, our results do suggest some immediate directions for potential improvements to `C51`. First, the convergence results for stochastic approximation suggest that an improved algorithm could be obtained by either directly minimising the Cramér distance or through a regularised KL minimisation that more closely reflects the mixture updates in Section 4.3. Second, the results of Propositions 3 and 4 indicate that if our support is densely focused around the true range of returns we should expect significantly better performance, due to the effects of the discount factor. Improving this by either prior domain knowledge or adapting the support to reflect the true return range could yield much better empirical performance.

6 CONCLUSION

In this paper we have introduced a framework for distributional RL algorithms, and provided convergence analysis of recently proposed algorithms. We have introduced the notion of the projected distributional Bellman operator and argued for its importance in the theory of distributional RL.

Interesting future directions from an empirical perspective include exploring the space of possible distributional RL algorithms set out in Section 3. From a theoretical perspective, the issue of how function approximation interacts with distributional RL remains an important open question.

Acknowledgements

The authors acknowledge the important contributions of their colleagues at DeepMind. Special thanks to Wojciech Czarnecki, Chris Maddison, Ian Osband and Grzegorz Swirszcz for their early suggestions and discussions. Thanks also to Clare Lyle for useful comments.

References

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

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer Distance as a Solution to Biased Wasserstein Gradients. *arXiv*, 2017b.

R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II: Approximate Dynamic Programming*. Athena Scientific, 2012.

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

P. Billingsley. *Probability and Measure*. John Wiley and Sons, second edition, 1986.

W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional Reinforcement Learning with Quantile Regression. In *Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence*, 2018.

T. Jaakkola, M. I. Jordan, and S. P. Singh. On the Convergence of Stochastic Iterative Dynamic Programming Algorithms. *Neural Computation*, 6(6): 1185–1201, 1994.

H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric Return Distribution Approximation for Reinforcement Learning. In *ICML*, 2010a.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric Return Density Estimation for Reinforcement Learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010b.

L. A. Prashanth and M. Ghavamzadeh. Actor-Critic Algorithms for Risk-Sensitive MDPs. In *NIPS*, pages 252–260, 2013.

G. A. Rummery and M. Niranjan. On-line Q-learning using Connectionist Systems. Technical report, Cambridge University Engineering Department, 1994.

M. Shaked and J. G. Shanthikumar. *Stochastic Orders and their Applications*. Academic Press, 1994.

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

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS)*, 1999.

G. J. Székely. E-statistics: The Energy of Statistical Samples. Technical report, Department of Mathematics and Statistics, Bowling Green State University, 2002.

A. Tamar, D. Di Castro, and S. Mannor. Learning the Variance of the Reward-to-go. *J. Mach. Learn. Res.*, 17(1):361–396, Jan. 2016. ISSN 1532-4435.

J. N. Tsitsiklis. Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning*, 16(3):185–202, Sep 1994.

J. N. Tsitsiklis and B. Van Roy. An Analysis of Temporal-Difference Learning with Function Approximation. Technical report, IEEE Transactions on Automatic Control, 1997.