Distributional reinforcement learning with linear function approximation

Marc G. Bellemare **Nicolas Le Roux** **Pablo Samuel Castro** **Subhodeep Moitra**
Google Brain

Abstract

Despite many algorithmic advances, our theoretical understanding of practical distributional reinforcement learning methods remains limited. One exception is Rowland et al. (2018)'s analysis of the C51 algorithm in terms of the Cramér distance, but their results only apply to the tabular setting and ignore C51's use of a softmax to produce normalized distributions. In this paper we adapt the Cramér distance to deal with arbitrary vectors. From it we derive a new distributional algorithm which is fully Cramér-based and can be combined to linear function approximation, with formal guarantees in the context of policy evaluation. In allowing the model's prediction to be any real vector, we lose the probabilistic interpretation behind the method, but otherwise maintain the appealing properties of distributional approaches. To the best of our knowledge, ours is the first proof of convergence of a distributional algorithm combined with function approximation. Perhaps surprisingly, our results provide evidence that Cramér-based distributional methods may perform worse than directly approximating the value function.

1 Introduction

In reinforcement learning one often seeks to predict the expected sum of discounted rewards, also called return or *value*, of a given state. The distributional perspective on reinforcement learning takes this idea further by suggesting that we should predict the full distribution of this random return, called *value distribution* (Bellemare et al., 2017a). This has produced state-of-the-art performance on a number of deep reinforcement learning benchmarks (e.g. Hessel et al., 2018; Barth-Maron et al., 2018; Dabney et al., 2018a,b).

The original distributional algorithm from this line of work is Bellemare et al.'s C51 algorithm. Core to C51 are 1) the use of a softmax transfer function to represent the value distribution, 2) a heuristic projection step, and finally 3) the minimization of a Kullback-Leibler (KL) loss. Rowland et al. (2018) showed that the heuristic projection minimizes a probability metric called the Cramér distance. However, their work did not explain the role of the KL loss in the algorithm.

The combination of two losses (Cramér and KL) is less than ideal, and makes the learning process technically more challenging to implement than, e.g., the classic Q-Learning algorithm (Watkins, 1989). This combination also makes it difficult to provide theoretical guarantees, both in terms of convergence but also in the quality of the value distribution generated by an approximate learner.

A natural question is whether it is possible to do away with the softmax and KL loss, and derive a "100% Cramér" algorithm, both for simplicity and theoretical understanding. In this paper we seek an algorithm which directly minimizes the Cramér distance between the output of the model, for example a deep network, and a target distribution. As it turns out, we can construct such an algorithm by treating the model outputs as an improper probability distribution, and deriving a variant of the Cramér distance which gracefully handles such distributions.

This new algorithm enables us to derive theoretical guarantees on the behaviour of a distributional algorithm when combined to linear function approximation, in the policy evaluation setting. Although convergence is guaranteed under the usual conditions, our performance bound is worse than that of an algorithm which only approximates the value function. This suggests that predicting the full distribution as an intermediate step in estimating the expected value could hurt

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

performance. As a whole, our results suggest that the good performance of C51 cannot solely be attributed to a better-behaved loss function.

2 Background

We consider an agent acting in an environment described by a finite Markov Decision Process $\langle \mathcal{X}, \mathcal{A}, \Pr, R, \gamma \rangle$ (Puterman, 1994). In this paper we study the policy evaluation setting, in which we assume a fixed policy π mapping states to distributions over actions and consider the resulting state to state transition function $\Pr_\pi$:

$$\Pr_\pi(x' \,|\, x) := \sum_{a \in \mathcal{A}} \pi(a \,|\, x) \Pr(x' \,|\, x, a).$$

We view the reward function R as a collection of random variables describing the bounded, random reward received when an agent exits a state $x \in \mathcal{X}$. The value distribution (Bellemare et al., 2017a) describes the random *return*, or sum of discounted rewards, received when beginning in state x:

$$Z^\pi(x) := \sum_{t=0}^{\infty} \gamma^t R(X_t) \quad X_0 = x, X_{t+1} \sim \Pr_\pi(\cdot \,|\, X_t).$$

The expectation of the value distribution corresponds to the familiar *value function* $V^\pi(x)$ (Sutton & Barto, 1998). Similar to the value function satisfying the Bellman equation, Z^π satisfies the distributional Bellman equation with an equality in distribution:

$$Z^\pi(x) \stackrel{D}{=} R(x) + \gamma Z^\pi(X') \qquad X' \sim \Pr_\pi(\cdot \,|\, x),$$

The distributional Bellman operator $\mathcal{T}^\pi$ over value distributions is defined as

$$\mathcal{T}^\pi Z(x) \stackrel{D}{:=} R(x) + \gamma \Pr_\pi Z(x), \tag{1}$$

where with some abuse of notation we write $\Pr_\pi Z(x) := Z(X'), X' \sim \Pr_\pi(\cdot \,|\, x)$. The operator $\mathcal{T}^\pi$ is a contraction mapping in the following sense: let d be a metric between probability distributions on $\mathbb{R}$, and for two random variables U, V denote by $d(U, V)$ the application of d to their distributions. We define the maximal metric $\bar{d}$ between two value distributions Z_1, Z_2 as

$$\bar{d}(Z_1, Z_2) := \sup_{x \in \mathcal{X}} d(Z_1(x), Z_2(x)).$$

Now, we say that d is 1) *sum invariant* if $d(A + U, A + V) \leq d(U, V)$ for any random variable A independent of U and V, and 2) *scale sensitive* of order β if for all $c \in \mathbb{R}$, $d(cU, cV) \leq c^\beta d(U, V)$ (Bellemare et al., 2017b). For any metric d which satisfies both of these conditions (with $\beta > 0$), then $\mathcal{T}^\pi$ is a contraction mapping with modulus γ^β in the maximal metric $\bar{d}$:

$$\bar{d}(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \leq \gamma^\beta \bar{d}(Z_1, Z_2).$$

Under mild assumptions and as a consequence of Banach's fixed point theorem, the process $Z_{k+1} := \mathcal{T}^\pi Z_k$ converges to Z^π in $\bar{d}$.

2.1 Metrics Over Distributions

Let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions. The Kullback-Leibler (KL) divergence of $\mathbf{q}$ from $\mathbf{p}$ is

$$D_{KL}(\mathbf{p}, \mathbf{q}) = \int_{-\infty}^{\infty} \mathbf{p}(t) \log \frac{\mathbf{p}(t)}{\mathbf{q}(t)} \, dt.$$

Note that the KL divergence is not properly a metric, but does define a loss function. However, the KL divergence is not scale sensitive. Furthermore, it is infinite whenever $\mathbf{p}$ is not absolutely continuous with respect to $\mathbf{q}$, which can be problematic when designing a distributional algorithm with finite support: applying the Bellman operator to a discrete random variable typically changes its support.

The KL divergence is generally used in conjunction with a softmax transfer function which guarantees that $\mathbf{q}$ has unit mass; without this constraint, the minimizer of D_{KL} may not be $\mathbf{q} = \mathbf{p}$. Furthermore, the KL divergence corresponds to the matching loss for the softmax function, guaranteeing that the resulting optimization is convex (with respect to the softmax weights; Auer et al., 1995).

Unlike the KL divergence, the Cramér distance (Székely, 2002) is a proper distance between probability distributions. Given two distributions $\mathbf{p}$ and $\mathbf{q}$ over $\mathbb{R}$ with cumulative distribution functions $F_\mathbf{p}$ and $F_\mathbf{q}$, the Cramér distance is defined as

$$D_C(\mathbf{p}, \mathbf{q}) = \int_{-\infty}^{+\infty} (F_\mathbf{p}(t) - F_\mathbf{q}(t))^2 \, dt. \tag{2}$$

For the purposes of distributional reinforcement learning, the Cramér distance has a number of appealing properties. First, it is both sum invariant and scale sensitive of order $\beta = \frac{1}{2}$. Then, the Cramér distance can be minimized by stochastic gradient methods.

2.2 Approximation in the Distributional Setting

Let us write $\mathbf{P}^\pi(x)$ for the distribution of the random variable $Z^\pi(x)$. There are two common hurdles to learning $\mathbf{P}^\pi(x)$: first, we typically do not have access to a simulator, and must instead rely on sample transitions; second, we cannot in general store the value

distribution exactly, and instead must maintain an approximation. These two issues have been well studied in the expected value setting of reinforcement learning (see, e.g. Bertsekas & Tsitsiklis, 1996; Tsitsiklis & Van Roy, 1997), in particular relating the mean behaviour of sample-based algorithms such as TD (Sutton, 1988) to their operator counterparts, including in the context of linear function approximation. This section provides analogous notation describing sample-based methods for distributional reinforcement learning.

With a tabular representation, where distributions are stored exactly, Rowland et al. (2018) showed the existence of a *mixture update* with step-size α:

$$\mathbf{P}(x) \leftarrow \mathbf{P}(x) + \alpha(f_{r,\gamma}(\mathbf{P}(x')) - \mathbf{P}(x)) \; .$$

In this mixture update, $f_{r,\gamma}(\mathbf{P}(x'))$ is the distribution corresponding to the random variable $r + \gamma Z(x')$, $Z(x') \sim \mathbf{P}(x')$. This update rule converges to $\mathbf{P}^\pi$ under the usual stochastic optimization conditions.

Rowland et al. (2018) also analyzed a mixture update for approximately tabular representations, when $\mathbf{P}$ is constrained to be a distribution over uniformly-spaced atoms (we will describe this parametrization in greater detail in the next section). The modified update incorporates a projection step Π_C which finds the constrained distribution $\mathbf{P}(x)$ closest to $f_{r,\gamma}(\mathbf{P}(x'))$ in Cramér distance:

$$\mathbf{P}(x) \leftarrow \mathbf{P}(x) + \alpha(\Pi_C f_{r,\gamma}(\mathbf{P}(x')) - \mathbf{P}(x)) \; . \qquad (3)$$

This projection step is used in the C51 algorithm, which parametrizes $\mathbf{P}_\Theta$ using a neural network with weights Θ and whose final layer uses a softmax transfer function to generate the vector of probabilities $\mathbf{P}_\Theta(x)$. Ignoring second order optimization terms, the C51 update is

$$\Theta \leftarrow \Theta - \alpha \nabla_\Theta D_{KL}(\Pi_C f_{r,\gamma}(\mathbf{P}_{\tilde{\Theta}}(x')) \,\|\, \mathbf{P}_\Theta(x)), \quad (4)$$

where the use of the KL divergence is justified as the matching loss to the softmax, and $\tilde{\Theta}$ is a "target" copy of Θ (Mnih et al., 2015).

Although the update rule Eq. (4) works well in practice, it is difficult to justify. The KL divergence is not scale sensitive, and it is not clear that its combination with the Cramér projection and the softmax function leads to a convergent algorithm.

To address this issue, here we consider an update rule which directly minimizes the Cramér distance:

$$\Theta \leftarrow \Theta - \alpha \nabla_\Theta D_C(f_{r,\gamma}(\mathbf{P}_{\tilde{\Theta}}(x')), \mathbf{P}_\Theta(x)). \qquad (5)$$

By the matching-loss argument, this suggests doing away with the transfer function and measuring the loss with respect to linear outputs. At first glance this might seem nonsensical, as these may not form a valid probability distribution. Yet, as we will see, the Cramér distance can be extended to deal with arbitrary vectors.

3 Generalizing the Cramér Distance

In this section we generalize the Cramér distance to vectors which do not necessarily describe probability distributions. We then transform this generalized distance to obtain a loss that is suited to the distributional setting. At a high level, our approach is as follows:

1. We rewrite the Cramér distance between distributions with discrete support as a weighted squared distance between vectors;
2. We show that this distance has undesirable properties when generalized beyond the space of probability distributions, and address this by modifying the eigenstructure of the weighting used in defining the distance;
3. We further modify the distance into a loss which regularizes the sum of vectors towards 1. This modification is key in our construction of an algorithm that is theoretically well-behaved when combined with linear function approximation.

We consider the space $\mathcal{D}$ of distributions over returns with finite, common, bounded support $\mathbf{z} = \{z_1, z_2, \ldots, z_k\}$ with $z_i \le z_{i+1}$. In this context, Eq. (2) simplifies to a sum with simple structure:

$$D_C(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k-1} (F_\mathbf{p}(z_i) - F_\mathbf{q}(z_i))^2 (z_{i+1} - z_i)$$

with $\mathbf{p}$, $\mathbf{q}$ in $\mathcal{D}$, where $F_\mathbf{p}$ is the cumulative distribution function of $\mathbf{p}$:

$$F_\mathbf{p}(z_i) = \sum_{j=1}^{i} \mathbf{p}(z_j).$$

We shall also assume that k is odd and $\mathbf{z} = \{\frac{1-k}{2}, \ldots, \frac{k-1}{2}\}$, i.e. $z_i = \frac{2i-1-k}{2}$, $z_{i+1} - z_i = 1$. Without detracting from our results, this simplifies their exposition.

Let $\mathbf{p} := [p_1, p_2, \ldots p_k]$ and $\mathbf{q} := [q_1, q_2, \ldots q_k]$ denote the vectors associated with $z_1, z_2, \ldots, z_k$, and write C for the lower-triangular matrix of 1s:

$$C = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 1 & 0 & \ldots & 0 & 0 & 0 \\ 1 & 1 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots \\ 1 & 1 & 1 & \ldots & 1 & 0 & 0 \\ 1 & 1 & 1 & \ldots & 1 & 1 & 0 \\ 1 & 1 & 1 & \ldots & 1 & 1 & 1 \end{bmatrix} .$$

If $\sum_i p_i = 1, p_i \geq 0$ (resp., $\sum_i q_i = 1, q_i \geq 0$), these can be viewed as the probabilities of a distribution over $\mathbf{z}$. Then, $C\mathbf{p}$ is the cumulative distribution of $\mathbf{p}$, and the Cramér distance between $\mathbf{p}$ and $\mathbf{q}$ becomes

$$l^2_{CC^\top}(\mathbf{p}, \mathbf{q}) := \|C\mathbf{p} - C\mathbf{q}\|^2 = \|\mathbf{p} - \mathbf{q}\|^2_{CC^\top} . \quad (6)$$

One can replace the cumulative distributions with the tail cumulative distributions to get

$$l^2_{C^\top C}(\mathbf{p}, \mathbf{q}) = \left\|C^\top \mathbf{p} - C^\top \mathbf{q}\right\|^2 = \|\mathbf{p} - \mathbf{q}\|^2_{C^\top C} . \quad (7)$$

If $\mathbf{p}$ or $\mathbf{q}$ do not correspond to proper probability distributions, the Cramér distance of Eq. 2 may be infinite, while Eq. 6 and 7 remain finite. This suggests the use of this definition when comparing vector-valued objects that are close to, or attempt to approximate distributions.

However, the two distances can disagree when $\mathbf{p}$ and $\mathbf{q}$ do not correspond to proper probability distributions. Let $\sum_i p_i$ be the "mass" of $\mathbf{p}$, reflecting its relationship to the mass of a probability distribution. If $\mathbf{p}$ and $\mathbf{q}$ have different mass, then $l^2_{CC^\top}(\mathbf{p}, \cdot) \neq l^2_{C^\top C}(\mathbf{p}, \cdot)$. The issue is that Eq. 6 and 7 measure differently the difference in mass.

To resolve this discrepancy, we modify the Cramér distance to deal unambiguously with uneven masses. This leads to a two-part distance: The first is insensitive to differences of total mass while the second only penalizes that difference. Let

$$e = [1/\sqrt{k}, \ldots, 1/\sqrt{k}]^\top \quad \text{and} \quad \Pi_{e^\perp} = I_k - ee^\top ,$$

our distance is

$$l^2_\lambda(\mathbf{p}, \mathbf{q}) := (\mathbf{p} - \mathbf{q})^\top \Pi_{e^\perp} CC^\top \Pi_{e^\perp} (\mathbf{p} - \mathbf{q}) + \lambda \left((\mathbf{p} - \mathbf{q})^\top e\right)^2 . \quad (8)$$

Denoting $C_\lambda = \Pi_{e^\perp} CC^\top \Pi_{e^\perp} + \lambda ee^\top$, we have

$$l^2_\lambda(\mathbf{p}, \mathbf{q}) = (\mathbf{p} - \mathbf{q})^\top C_\lambda (\mathbf{p} - \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|^2_{C_\lambda} .$$

First, one may note that, when $\mathbf{p}$ and $\mathbf{q}$ have the same total mass, we have $l^2_{CC^\top} = l^2_{C^\top C} = l^2_\lambda(\mathbf{p}, \mathbf{q})$ for all values of λ. As such, this new distance clarifies the behaviour for arbitrary vectors while being consistent with the existing Cramér loss for proper distributions. For any given distribution $\mathbf{p}$, the solution to

$$\min_{\mathbf{q} \in \mathbb{R}^k} l^2_\lambda(\mathbf{p}, \mathbf{q})$$

is $\mathbf{p}$. On the other hand, if the minimization is done over a constrained set, λ determines the magnitude of the penalty from the difference in total mass.

As we will later see, the distance l_λ, used as a loss, is not sufficient to guarantee good behaviour with linear function approximation. Instead, we define a related loss but with an explicit normalization penalty:

$$\hat{l}^2_\lambda(\mathbf{p}, \mathbf{q}) = (\mathbf{p} - \mathbf{q})^\top \Pi_{e^\perp} CC^\top \Pi_{e^\perp} (\mathbf{p} - \mathbf{q}) + \lambda \left(\mathbf{q}^\top e - 1\right)^2 . \quad (9)$$

Intuitively, $\hat{l}_\lambda$ recognizes that a distribution-like object should benefit from having unit mass. In the context of the distributional Bellman operator, this is the difference between backing up the mass at successor states versus normalizing the state's distribution to sum to 1. However, $\hat{l}_\lambda$ does not define a distance proper, and our theoretical treatment of it in Section 4.2.2 will require additional care.

4 Analysis

We now explore, through a series of lemmas, properties of the Cramér distance of relevance to distributional reinforcement learning. Two of these results will be related to the minimization of the Cramér loss directly over distributions and two will be related to the use of linear function approximation.

4.1 Optimization properties

We begin by analyzing properties resulting from the minimization of l^2_λ over $\mathbf{q}$, beginning with the approximately tabular setting (Section 2.2).

4.1.1 Impact on optimization speed

A well-known result in convex optimization states that, when minimizing a quadratic function f with positive definite Hessian H using a batch first-order method, e.g., Eq. 5, the convergence to the optimum is linear with a rate of $1 - \frac{1}{\kappa}$ where κ is the condition number of H. Assuming we directly optimize the Cramér loss over $\mathbf{q}$ with such a method, the convergence rate would depend on the condition number of the matrix used, i.e. $CC^\top$ when using the Cramér loss $l^2_{CC^\top}$ or C_λ when using the extended loss l^2_λ.

Lemma 1 (Condition number)**.** *Let $\mathcal{C}$ be the set of symmetric matrices M for which $(\mathbf{p} - \mathbf{q})^\top M(\mathbf{p} - \mathbf{q}) = (\mathbf{p} - \mathbf{q})^\top CC^\top (\mathbf{p} - \mathbf{q})$ for all proper distributions $\mathbf{p}$ and $\mathbf{q}$. Let $\kappa_{\min}(\mathcal{C})$ the lowest condition number attained by matrices M in $\mathcal{C}$. Then all the matrices of the form C_λ with $\lambda \in [\lambda_{k-1}(C_0), \lambda_1(C_0)]$, where $\lambda_{k-1}(C_0)$ and $\lambda_1(C_0)$ are the second smallest and largest eigenvalues of C_0, respectively, have condition number $\kappa_{\min}(\mathcal{C})$.*

The proof of this result and the following may be found in the appendix.

Lemma 1 shows that the optimal convergence rate is obtained for a potentially wide range of values for

λ. As an example, for $k = 51$, the condition number of $CC^\top$ is about 4296 while $\kappa_{\min}(\mathcal{C})$ is around 1053, about 4 times lower, and this is true for λ in the range $[0.250, 263]$.

4.1.2 Preservation of the expectation

Although the prediction $\mathbf{q}$ may not be a distribution, it still makes sense to talk of the dot product between $\mathbf{q}$ and the support $\mathbf{z}$ as its "expectation": indeed, in many cases of interest the optimization procedure does yield valid distributions. In designing a full distributional agent, this generalized notion of expectation is also a natural way to convert $\mathbf{q}$ into a scalar value, e.g. for decision making.

This section discusses potential guarantees on the difference in expected return between $\mathbf{p}$ and $\mathbf{q}$ when $\mathbf{q}$ is the minimizer of the Cramér loss l^2_λ over a restricted set. Typically, we will ask $\mathbf{q}$ to have a specific support but other constraints might include that $\mathbf{q}$ must be normalized or that some values of $\mathbf{q}$ cannot be modified. Specifically, the following lemma studies the impact on the expectation when minimizing the Cramér loss over an affine subset.

Lemma 2 (Expectation preserving)**.** *Let* $\mathbf{p}$ *be an arbitrary distribution over a discrete support. Let* $\Pi_{A,b}(\mathbf{p})$ *the projection of* $\mathbf{p}$ *onto the linear subset* $\mathcal{S}_{A,b} = \{\mathbf{q} | A\mathbf{q} = b\}$*. Then, if the first and the last columns of* A *are equal, i.e.* $A_1 = A_k$*, then* $\mathbf{p}$ *and* $\Pi_{A,b}(\mathbf{p})$ *have the same expectation.*

Lemma 2 covers projections onto specific supports, as used in C51, as well as constraints on the total mass of $\Pi_{A,b}(\mathbf{p})$, for instance that the projection has unit mass. Here, we use $\mathbf{z}$ to denote both the support and the vector containing the elements of that support. More generally, the Cramér projection offers a certain amount of freedom on the constraints that can be enforced while still preserving the expectation. In particular, leaving the two boundaries unconstrained is enough to preserve the expectation.

4.2 Linear function approximation

We next quantify the behaviour of our generalized loss function when combined to linear function approximation. Section 4.2.2 is the main theoretical contribution of this paper: it shows that the combination of a loss based on Equation 9 together with linear approximation produces a stable dynamical system, and quantifies the approximation error that results from it.

4.2.1 Two-step optimization

In categorical distributional RL, the target distribution $\mathbf{p}$ is the product of an application of the distributional Bellman operator and does not usually have the same support as the parametrized output distribution $\mathbf{q}(\theta)$. Recall that $\mathcal{D}$ is the set of distributions with support $\mathbf{z}$. C51 first projects $\mathbf{p}$ onto $\mathcal{D}$, yielding

$$\Pi_{\lambda,\mathcal{D}}(\mathbf{p}) = \arg\min_{\mathbf{u}\in\mathcal{D}} l^2_\lambda(\mathbf{p}, \mathbf{u}),$$

assuming $\mathbf{p}$ is a proper distribution. Then, as a second step in the update process, it minimizes the KL divergence between $\Pi_{\lambda,\mathcal{D}}(\mathbf{p})$ and $\mathbf{q}(\theta)$.

In our experiments we retain the projection onto $\mathbf{z}$ from the C51 algorithm, and subsequently minimize our loss with respect to this projection. Doing so is equivalent to directly minimizing the Cramér loss, even when $\mathbf{p}$ is not a proper distribution. Extending the result from Lemma 3 of Rowland et al. (2018), we note that $\Pi_{\lambda,\mathcal{D}}$ is an orthogonal projection for $\mathbf{q}(\theta) \in \mathcal{D}$:

$$l^2_\lambda(\mathbf{p}, \mathbf{q}(\theta)) = l^2_\lambda(\mathbf{p}, \Pi_{\lambda,\mathcal{D}}(\mathbf{p})) + l^2_\lambda(\Pi_{\lambda,\mathcal{D}}(\mathbf{p}), \mathbf{q}(\theta)).$$

Taking the derivative of the two sides of this equation with respect to θ, the parameters of the model, yields

$$\frac{\partial l^2_\lambda(\mathbf{p}, \mathbf{q}(\theta))}{\partial\theta} = \frac{\partial l^2_\lambda(\Pi_{\lambda,\mathcal{D}}(\mathbf{p}), \mathbf{q}(\theta))}{\partial\theta}$$

and minimizing the distance with the projection of $\mathbf{p}$ onto the support $\mathbf{z}$ of $\mathbf{q}$ leads to the same gradients. With some additional care, the argument extends to the loss with a normalization penalty, $\hat{l}^2_\lambda$.

4.2.2 Convergence to a fixed point

We are now ready to show the convergence of distributional RL in the context of linear function approximation. Recall that a proof of convergence for C51 is hindered by the failure of the KL minimization process to be nonexpansive in the Cramér distance; as we will see, our result critically depends on the loss defined in Equation 9.

We consider a feature matrix $\Phi \in \mathbb{R}^{n\times m}$, with n the number of states and m the number of features, and a weight matrix $\Theta \in \mathbb{R}^{m\times k}$. That is, we consider outputs of the form $\mathbf{Q} = \Phi\Theta \in \mathbb{R}^{n\times k}$, which with some abuse of terminology we call value distributions. As before, we write $\mathbf{Q}(x)$ to denote the k-dimensional output for state $x \in \mathcal{X}$.

We study a stochastic update rule of the form given by Eq. (5), but where D_C is replaced by the loss $\hat{l}^2_\lambda$. When the states to be updated are sampled according to a distribution ξ, the expected behaviour of this update rule corresponds to an operator akin to a projection (Tsitsiklis & Van Roy, 1997). In our setting, the operator minimizes the ξ-weighted Cramér loss derived from $\hat{l}^2_\lambda$, denoted

$$\hat{l}^2_{\xi,\lambda}(\mathbf{P}, \mathbf{Q}) := \sum_{x\in\mathcal{X}} \xi(x)\hat{l}^2_\lambda(\mathbf{P}(x), \mathbf{Q}(x)).$$

We denote this operator by $\hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}$ (the notation is made explicit in the appendix). Given a value distribution $\mathbf{P} \in \mathbb{R}^{n\times k}$, the operator finds the value distribution in the span of Φ which minimizes $\hat{l}^2_{\xi,\lambda}(\mathbf{P}, \cdot)$:

$$\hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathbf{P} = \Phi\Theta^* \quad \text{where} \quad \Theta^* = \arg\min_{\Theta} \hat{l}^2_{\xi,\lambda}(\mathbf{P}, \Phi\Theta).$$

Finally, our analysis is performed with respect the distance l^2_λ, rather than the loss $\hat{l}^2_\lambda$ (which is not a distance). This leads to the ξ-weighted distance

$$l^2_{\xi,\lambda}(\mathbf{P}, \mathbf{Q}) := \sum_{x\in\mathcal{X}} \xi(x) l^2_\lambda(\mathbf{P}(x), \mathbf{Q}(x)),$$

with corresponding projection operator $\mathbf{\Pi}_{\xi,\lambda,\Phi}$.

We now show that the combination of the distributional Bellman operator $\mathcal{T}^\pi$ and the ξ-weighted, projection-like operator describes a convergent algorithm. When $\lambda > 0$, we can further bound the distance of this fixed point to the true value distribution $\mathbf{P}^\pi$ in terms of the best approximation in the class, $\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi$. As is usual, ξ is taken to be the stationary distribution of the Markov chain described by $\Pr_\pi$: $\xi(x') = \sum_{x\in\mathcal{X}} \xi(x) \Pr_\pi(x' \,|\, x)$.

Theorem 1 (Convergence of the projected distributional Bellman process)**.** *Let ξ be the stationary distribution induced by the policy π. The process*

$$\mathbf{P}_0 := \Phi\Theta_0 \quad , \quad \mathbf{P}_{k+1} := \hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{P}_k.$$

converges to a set S such that for any two $\mathbf{P}, \mathbf{P}' \in S$, there is a $\mathcal{X}$-indexed vector of constants α such that

$$\mathbf{P}(x) = \mathbf{P}'(x) + \alpha(x)e.$$

If $\lambda > 0$, S consists of a single point $\tilde{\mathbf{P}}$ which is the fixed point of the process. Furthermore, we can bound the error of this fixed point with respect to the true value distribution $\mathbf{P}^\pi$:

$$l^2_{\xi,\lambda}(\tilde{\mathbf{P}}, \mathbf{P}^\pi) \le \frac{1}{1-\gamma} l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi, \mathbf{P}^\pi) - \frac{\gamma\lambda}{1-\gamma}\left\|\tilde{\mathbf{P}} - \mathbf{P}^\pi\right\|^2_{\xi,ee^\top},$$

where the second term measures the difference in mass between $\tilde{\mathbf{P}}$ and $\mathbf{P}^\pi$.

Theorem 1 is significant for a number of reasons. First, it answers the question left open by Rowland et al. (2018), namely whether a proof of convergence exists for the distributional setting with an approximate representation, and with which representation. Second, it shows that there is a trade-off between the different components of the loss – while our result concerns linear function approximation, it suggests that similar trade-offs must exist within other distributional algorithms.

The parameter λ plays an important role in the theorem, both to guarantee convergence and (indirectly) to determine the approximation error. At a high level, this makes sense: a high value of λ forces the algorithm to output something close to a distribution, at the expense of actual predictions. On the other hand, taking $\lambda = 0$ yields a process which may not converge to a single point. Finally, we note that to guarantee convergence to a unique fixed point, it is not enough to use the loss from Eq. 8: in that case, we can only guarantee convergence to the set S, even for $\lambda > 0$. The following lemma, used to prove Theorem 1, shows why: the distributional Bellman operator $\mathcal{T}$ is only a nonexpansion along the dimension e, which captures the mass of the output vectors.

Lemma 3. *Let ξ be the stationary distribution induced by the policy π. Write $\mathcal{T}^{\pi\prime} := \Pi_{\lambda,\mathcal{D}}\mathcal{T}^\pi$ to mean the distributional Bellman operator followed by a projection onto the support $\mathbf{z} = z_1, \ldots, z_k$. For a matrix $B \in \mathbb{R}^{k\times k}$ and $\Delta \in \mathbb{R}^{n\times k}$, write*

$$\|\Delta\|^2_{\xi,B} = \sum_{x\in\mathcal{X}} \xi(x) \|\Delta(x)\|^2_B .$$

Then for any two value distributions $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times k}$,

$$\left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|^2_{\xi,AA^\top} \le \gamma \|\mathbf{P} - \mathbf{Q}\|^2_{\xi,AA^\top}$$
$$\left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|^2_{\xi,ee^\top} \le \|\mathbf{P} - \mathbf{Q}\|^2_{\xi,ee^\top} .$$

where $A := \Pi_{e^\perp} C$.

When $\mathbf{P}$ and $\mathbf{Q}$ have equal mass, we recover the contraction result by Bellemare et al. (2017b) (albeit in ξ-weighted Cramér distance, rather than maximal Cramér distance) – however, this also shows that our generalization of the distributional Bellman operator deals differently with probability mass itself. This is why Theorem 1 requires the normalization penalty loss $\hat{l}^2_\lambda$, rather than the simpler l^2_λ.

4.2.3 Bound on the approximation error

Our analysis provides us with a partial answer to the question: why and when should distributional reinforcement learning perform better empirically? In the linear approximation case that we study here, one answer is that it might hurt performance, as the following theorem suggests:

Theorem 2 (Error bound for the expected value)**.** *Let $\|\cdot\|_\xi$ be the ξ-weighted norm over value functions. The squared expectation error of the fixed point $\tilde{\mathbf{P}}$ with respect to the true value function V^π is bounded as*

$$\|\mathbb{E}_{\tilde{\mathbf{P}}}\,\mathbf{z} - V^\pi\|^2_\xi \le \|C^{-1/2}_\lambda \mathbf{z}\|^2 \boldsymbol{l}^2_{\xi,\lambda}(\tilde{\mathbf{P}}, \mathbf{P}^\pi).$$

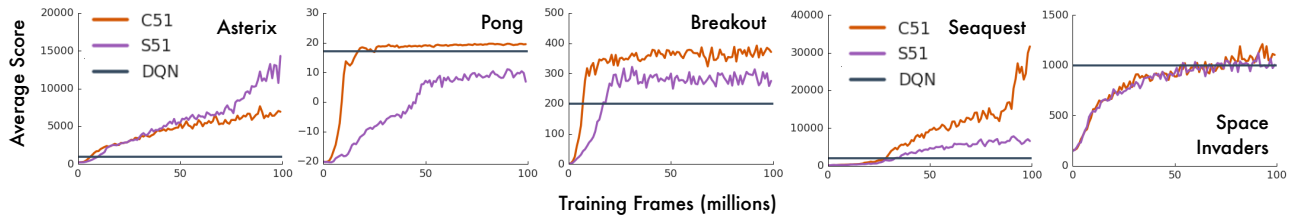

Figure 1: Learning curves (training scores) for C51 and S51 on five Atari 2600 games, and reference score for DQN at 100 million frames as given by Bellemare et al. (2017a).

The proof relies on a Rayleigh quotient argument, and shows that the bound is tight if the error vector $\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)$ is collinear with $C_\lambda^{-1/2}\mathbf{z}$. In particular, if we take λ such that $C_\lambda = CC^T$, then the constant is $\|C_\lambda^{-1/2}\mathbf{z}\|^2 = \|e\sqrt{k}\|^2 = k$. Then, as $\lambda \to 0$, the constant goes to infinity. By contrast, the bound on the approximate value function derived from Tsitsiklis & Van Roy (1997) is better in two respects: first, its equivalent constant is 1. Second, our bound contains an amplification factor $1/\sqrt{1-\gamma}$ from the error term $\boldsymbol{l}^2_{\xi,\lambda}(\tilde{\mathbf{P}}, \mathbf{P}^\pi)$, which in their bound becomes the smaller constant $1/\sqrt{1-\gamma^2}$, because the usual Bellman operator is a γ-contraction in $\|\cdot\|_\xi$, while the Cramér distance is only a $\sqrt{\gamma}$-contraction in the equivalent norm.

However, the bound is slightly misleading. In our analysis we have assumed that the width of the support, i.e. $z_k - z_1$, also grows with k. We can instead normalize the C matrix and the support $\mathbf{z}$ to reflect a fixed width: $C' = C/k$ and $z' = z/k$. In this case, the constant remains but the squared loss may in some cases be k times smaller. Still, it is not unreasonable to expect that, given that the distributional approach models more things, it should be more susceptible to misspecification.

5 Experiments

The Cramér distance enjoys many theoretical properties that the KL divergence used in C51 lacks. To complement our theoretical results in the policy evaluation setting, we now study how our new loss affects the overall performance in the more complex control setting (Sutton & Barto, 1998). Our goals are to demonstrate that we can achieve qualitatively comparable performance to C51 with an algorithm based on this loss, and to study the similarities and differences between the two algorithms.

We compare the original C51 algorithm with our Cramér variant from Eq. 9, dubbed S51, on five games supported by the Arcade Learning Environment (Bellemare et al., 2013), and using the Dopamine framework (Castro et al., 2018). In a nutshell, S51 learns from samples, using the sample-based version of the distributional Bellman operator (Eq. 1), but where the fixed policy is replaced by one which backs up the distribution with maximum expected value (what Rowland et al. (2018) calls "Categorical Q-Learning"). Further experimental details, including on how to transform C51 into S51, are given in Appendix A.

Figure 1 shows that S51 achieves higher scores than DQN, demonstrating that it maintains the empirical benefits of the distributional perspective, and performs as well as C51 in three out of five games. This is especially significant given the relative freedom of the network in outputting arbitrary vectors. Nonetheless, our results suggest that there are benefits to enforcing normalized distributions – possibly in reducing the update variance.

To better understand the qualitative differences between the two algorithms, we studied agents playing through episodes of different games and visualized the predicted distribution for their selected actions (videos available in the supplemental; Figure 3). We find that C51 outputs value distributions which are bell-shaped and may have a separate mode at 0. In contrast, the S51 distributions are much more diverse; we highlight two interesting results:

Double negatives. S51 agents often assign negative mass to negative returns in games where such returns are impossible, such as Pong (Figure 2, left). The total mass in these cases is still close to 1.

Compensation around 0. In Space Invaders (Figure 2, right), the 0 return prediction is bracketed with small negative and positive corrections that cancel each other out. One explanation is that the network compensates for its limited capacity by relying on negative return predictions. This is particularly interesting as this behaviour is not possible under published distributional algorithms.

Noisier predictions (left and right). S51 assigns a small amount of probability to almost all returns. We hypothesize that this effect is visually absent from the C51 histograms because of the squashing effect

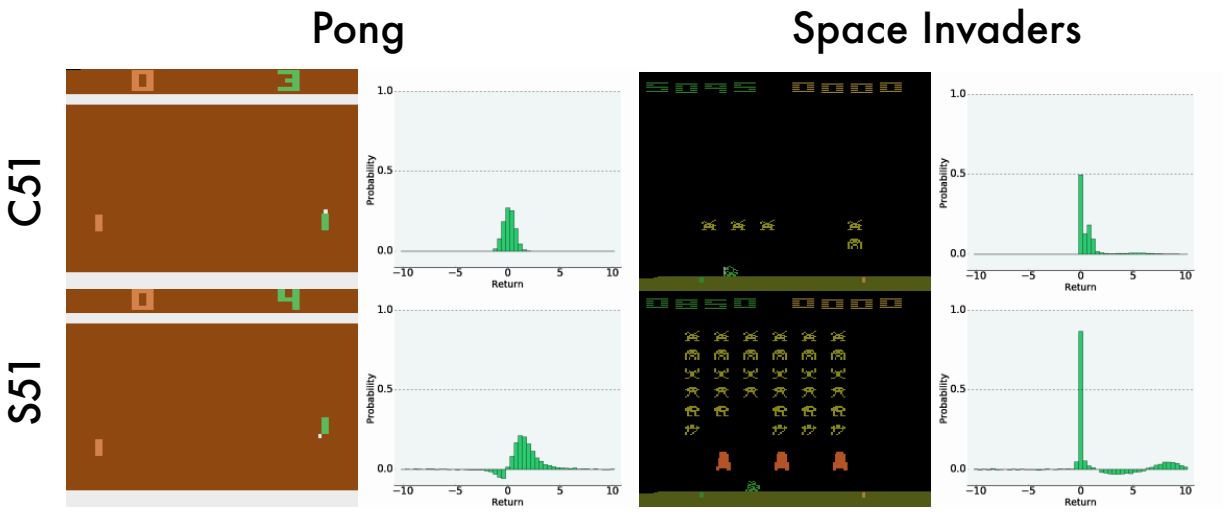

Figure 2: Distributions predicted by both algorithms in similar situations.

of the softmax transfer function, and that this added noise explains some of the difference in performance. In particular, to generate a small probability the C51 network need only output a sufficiently negative logit; by contrast, S51 must output a value which is neither too negative nor too positive (i.e., is actually close to 0).

6 Discussion and Conclusion

While the convergence of the distributional approach with linear approximation may have been predictable, our proof shows that the result is not completely straightforward, and that the normalization penalty plays an important role in convergence. Because the softmax produces bounded outputs, it may still be possible to derive some convergence guarantees for it; however, it seems difficult to bound on its approximation error once we leave the convex regime of the linear outputs/squared loss combination. Another question is whether minimizing the Cramér distance in the context of function approximation for optimal control somehow results in learning dynamics that are more stable than in the expected case, as a wealth of empirical results now suggest.

The Wasserstein distance also plays an important role in distributional reinforcement learning. Dabney et al. (2018b) demonstrated that one can obtain a stable distributional algorithm which minimizes the Wasserstein distance even in the approximate case by performing quantile regression rather than gradient descent on the sample Wasserstein loss. A similar analysis to ours may in fact prove convergence in the approximate setting; we expect that minimizing the Wasserstein metric should also be susceptible to pathological cases yielding a worse approximation of expected values.

Despite our attempts, we could not match the raw performance of C51. While this may only be a matter of hyperparameter tuning, we might have lost other properties when moving away from the KL. One might also wonder if there are other losses even more suited to the problem than our modified Cramér loss. In particular, since the ultimate goal is to preserve the expectation of the target distribution, one could adapt the loss to strengthen the link between loss minimization and expectation preservation.

References

Auer, Peter, Herbster, Mark, and Warmuth, Manfred K. Exponentially many local minima for single neurons. In *Advances in Neural Information Processing*, 1995.

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

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

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

Bellemare, Marc G, Danihelka, Ivo, Dabney, Will, Mohamed, Shakir, Lakshminarayanan, Balaji, Hoyer,

Stephan, and Munos, Rémi. The Cramér distance as a solution to biased Wasserstein gradients. *arXiv*, 2017b.

Bertsekas, Dimitri P. and Tsitsiklis, John N. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Castro, Pablo S., Moitra, Subhodeep, Gelada, Carles, Kumar, Saurabh, and Bellemare, Marc G. Dopamine: A research framework for deep reinforcement learning. *arXiv*, 2018.

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

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

Hessel, Matteo, Modayil, Joseph, van Hasselt, Hado, Schaul, Tom, Ostrovski, Georg, Dabney, Will, Horgan, Dan, Piot, Bilal, Azar, Mohammad, and Silver, David. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015.

Puterman, Martin L. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

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

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

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

Székely, Gabor J. E-statistics: The energy of statistical samples. Technical Report 02-16, Bowling Green State University, Department of Mathematics and Statistics, 2002.

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

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