一种分布视角下的强化学习

Marc G. Bellemare [*] [1] Will Dabney [*] [1] Rémi Munos [1]

摘要

在本文中，我们强调了*value distribution*的基本重要性：强化学习代理接收到的随机回报的分布。这与强化学习中常用的建模这种随机回报的期望值或*value*的方法形成了对比。尽管已经有一系列研究价值分布的文献，但迄今为止，它总是用于特定目的，例如实现风险意识行为。我们从策略评估和控制的理论结果开始，揭示了后者中的显著分布不稳定性。然后，我们从分布的角度设计了一个新算法，该算法将贝尔曼方程应用于近似价值分布的学习。我们使用Arcade Learning Environment中的游戏套件来评估我们的算法。我们获得了最先进的结果，并提供了轶事证据，证明了在近似强化学习中价值分布的重要性。最后，我们结合理论和实证证据，突出了价值分布在近似设置中对学习的影响方式。

1. 引言

强化学习的一个主要原则是，当行为不受其他约束时，智能体应该力求最大化其预期效用 Q，或 *value* (Sutton & Barto, 1998)。贝尔曼方程简明地描述了这一价值，用预期奖励和随机转换的预期结果 $(x, a) \to (X', A')$ 来表示：

$$Q(x,a) = \mathbb{E}\, R(x,a) + \gamma\, \mathbb{E}\, Q(X',A').$$

在本文中，我们旨在超越价值的概念，主张一种分布性的强化视角。机器学习。具体来说，我们研究的主要对象是随机回报 Z，其期望值是价值 Q。这种随机回报还通过一个分布性的递归方程来描述：

$$Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X',A').$$

The *distributional Bellman equation* 表示 Z 的分布由三个随机变量的交互作用来表征：奖励 R、下一个状态-动作 (X', A') 以及其随机回报 $Z(X', A')$。类比于一个熟知的情况，我们将这个量称为 *value distribution*。

尽管分布视角几乎和贝尔曼方程本身一样古老（Jaquette, 1973; Sobel, 1982; White, 1988），在强化学习中，它迄今为止一直被从属于特定目的：用于建模参数不确定性（Dearden等, 1998），设计风险敏感算法（Morimura等, 2010b;a），或者进行理论分析（Azar等, 2012; Lattimore & Hutter, 2012）。相比之下，我们认为价值分布在强化学习中应该扮演核心角色。

收缩策略评估贝尔曼算子。基于Rösler (1992) 的结果，我们证明，在固定策略的情况下，贝尔曼算子对于价值分布是收缩的，这是在Wasserstein（也称为Kantorovich或Mallows）度量的最大形式下。我们特别选择的度量很重要：同样的算子在总变差、Kullback-Leibler散度或Kolmogorov距离下不是收缩的。

控制设置中的不稳定性。我们将展示贝尔曼最优方程在分布版本中的不稳定性，与策略评估情况不同。具体来说，尽管最优性算子在期望值意义上是一个压缩映射（符合通常的最优性结果），但它在任何分布度量下都不是压缩映射。这些结果为学习算法提供了支持，这些算法能够建模非稳态策略的影响。

更好的近似方法。从算法的角度来看，学习一个近似分布而不是其近似期望有许多好处。分布性的贝尔曼算子保留了价值分布的多模态性，我们认为这会导致更稳定的学习。近似整个分布还能减轻从非稳态策略中学习的影响。作为一个整体，

[*] Equal contribution [1] DeepMind, London, UK. Correspondence to: Marc G. Bellemare <bellemare@google.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

we argue that this approach makes approximate reinforcement learning significantly better behaved.

We will illustrate the practical benefits of the distributional perspective in the context of the Arcade Learning Environment (Bellemare et al., 2013). By modelling the value distribution within a DQN agent (Mnih et al., 2015), we obtain considerably increased performance across the gamut of benchmark Atari 2600 games, and in fact achieve state-of-the-art performance on a number of games. Our results echo those of Veness et al. (2015), who obtained extremely fast learning by predicting Monte Carlo returns.

From a supervised learning perspective, learning the full value distribution might seem obvious: why restrict ourselves to the mean? The main distinction, of course, is that in our setting there are no given targets. Instead, we use Bellman's equation to make the learning process tractable; we must, as Sutton & Barto (1998) put it, "learn a guess from a guess". It is our belief that this guesswork ultimately carries more benefits than costs.

2. Setting

We consider an agent interacting with an environment in the standard fashion: at each step, the agent selects an action based on its current state, to which the environment responds with a reward and the next state. We model this interaction as a time-homogeneous Markov Decision Process $(\mathcal{X}, \mathcal{A}, R, P, \gamma)$. As usual, $\mathcal{X}$ and $\mathcal{A}$ are respectively the state and action spaces, P is the transition kernel $P(\cdot \,|\, x, a)$, $\gamma \in [0, 1]$ is the discount factor, and R is the reward function, which in this work we explicitly treat as a random variable. A stationary policy π maps each state $x \in \mathcal{X}$ to a probability distribution over the action space $\mathcal{A}$.

2.1. Bellman's Equations

The *return* Z^π is the sum of discounted rewards along the agent's trajectory of interactions with the environment. The value function Q^π of a policy π describes the expected return from taking action $a \in \mathcal{A}$ from state $x \in \mathcal{X}$, then acting according to π:

$$Q^\pi(x, a) := \mathbb{E}\, Z^\pi(x, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t)\right], \qquad (1)$$

$$x_t \sim P(\cdot \,|\, x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot \,|\, x_t), x_0 = x, a_0 = a.$$

Fundamental to reinforcement learning is the use of Bellman's equation (Bellman, 1957) to describe the value function:

$$Q^\pi(x, a) = \mathbb{E}\, R(x, a) + \gamma \mathop{\mathbb{E}}_{P,\pi} Q^\pi(x', a').$$

In reinforcement learning we are typically interested in acting so as to maximize the return. The most common approach for doing so involves the optimality equation

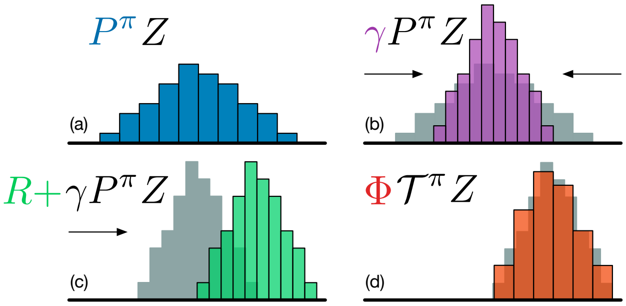


Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy π, (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).

$$Q^*(x, a) = \mathbb{E}\, R(x, a) + \gamma\, \mathbb{E}_P \max_{a' \in \mathcal{A}} Q^*(x', a').$$

This equation has a unique fixed point Q^*, the optimal value function, corresponding to the set of optimal policies Π^* (π^* is optimal if $\mathbb{E}_{a\sim\pi^*} Q^*(x, a) = \max_a Q^*(x, a)$).

We view value functions as vectors in $\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, and the expected reward function as one such vector. In this context, the *Bellman operator* $\mathcal{T}^\pi$ and *optimality operator* $\mathcal{T}$ are

$$\mathcal{T}^\pi Q(x, a) := \mathbb{E}\, R(x, a) + \gamma \mathop{\mathbb{E}}_{P,\pi} Q(x', a') \qquad (2)$$

$$\mathcal{T} Q(x, a) := \mathbb{E}\, R(x, a) + \gamma\, \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \qquad (3)$$

These operators are useful as they describe the expected behaviour of popular learning algorithms such as SARSA and Q-Learning. In particular they are both contraction mappings, and their repeated application to some initial Q_0 converges exponentially to Q^π or Q^*, respectively (Bertsekas & Tsitsiklis, 1996).

3. The Distributional Bellman Operators

In this paper we take away the expectations inside Bellman's equations and consider instead the full distribution of the random variable Z^π. From here on, we will view Z^π as a mapping from state-action pairs to distributions over returns, and call it the *value distribution*.

Our first aim is to gain an understanding of the theoretical behaviour of the distributional analogues of the Bellman operators, in particular in the less well-understood control setting. The reader strictly interested in the algorithmic contribution may choose to skip this section.

3.1. Distributional Equations

It will sometimes be convenient to make use of the probability space $(\Omega, \mathcal{F}, \Pr)$. The reader unfamiliar with measure theory may think of Ω as the space of all possible outcomes of an experiment (Billingsley, 1995). We will write $\|\mathbf{u}\|_p$ to denote the L_p norm of a vector $\mathbf{u} \in \mathbb{R}^{\mathcal{X}}$ for $1 \leq p \leq \infty$; the same applies to vectors in $\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$. The L_p norm of a random vector $U : \Omega \to \mathbb{R}^{\mathcal{X}}$ (or $\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$) is then $\|U\|_p := \left[\mathbb{E}\left[\|U(\omega)\|_p^p\right]\right]^{1/p}$, and for $p = \infty$ we have $\|U\|_\infty = \operatorname{ess\,sup} \|U(\omega)\|_\infty$ (we will omit the dependency on $\omega \in \Omega$ whenever unambiguous). We will denote the c.d.f. of a random variable U by $F_U(y) := \Pr\{U \leq y\}$, and its inverse c.d.f. by $F_U^{-1}(q) := \inf\{y : F_U(y) \geq q\}$.

A distributional equation $U \overset{D}{:=} V$ indicates that the random variable U is distributed according to the same law as V. Without loss of generality, the reader can understand the two sides of a distributional equation as relating the distributions of two independent random variables. Distributional equations have been used in reinforcement learning by Engel et al. (2005); Morimura et al. (2010a) among others, and in operations research by White (1988).

3.2. The Wasserstein Metric

The main tool for our analysis is the Wasserstein metric d_p between cumulative distribution functions (see e.g. Bickel & Freedman, 1981, where it is called the Mallows metric). For F, G two c.d.fs over the reals, it is defined as

$$d_p(F, G) := \inf_{U,V} \|U - V\|_p,$$

where the infimum is taken over all pairs of random variables (U, V) with respective cumulative distributions F and G. The infimum is attained by the inverse c.d.f. transform of a random variable $\mathcal{U}$ uniformly distributed on $[0, 1]$:

$$d_p(F, G) = \|F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U})\|_p.$$

For $p < \infty$ this is more explicitly written as

$$d_p(F, G) = \left(\int_0^1 \left|F^{-1}(u) - G^{-1}(u)\right|^p du\right)^{1/p}.$$

Given two random variables U, V with c.d.fs F_U, F_V, we will write $d_p(U, V) := d_p(F_U, F_V)$. We will find it convenient to conflate the random variables under consideration with their versions under the inf, writing

$$d_p(U, V) = \inf_{U,V} \|U - V\|_p.$$

whenever unambiguous; we believe the greater legibility justifies the technical inaccuracy. Finally, we extend this metric to vectors of random variables, such as value distributions, using the corresponding L_p norm.

Consider a scalar a and a random variable A independent of U, V. The metric d_p has the following properties:

$$d_p(aU, aV) \leq |a| d_p(U, V) \tag{P1}$$
$$d_p(A + U, A + V) \leq d_p(U, V) \tag{P2}$$
$$d_p(AU, AV) \leq \|A\|_p d_p(U, V). \tag{P3}$$

We will need the following additional property, which makes no independence assumptions on its variables. Its proof, and that of later results, is given in the appendix.

Lemma 1 (Partition lemma). *Let $A_1, A_2, \dots$ be a set of random variables describing a partition of Ω, i.e. $A_i(\omega) \in \{0, 1\}$ and for any ω there is exactly one A_i with $A_i(\omega) = 1$. Let U, V be two random variables. Then*

$$d_p(U, V) \leq \sum\nolimits_i d_p(A_i U, A_i V).$$

Let $\mathcal{Z}$ denote the space of value distributions with bounded moments. For two value distributions $Z_1, Z_2 \in \mathcal{Z}$ we will make use of a maximal form of the Wasserstein metric:

$$\bar{d}_p(Z_1, Z_2) := \sup_{x,a} d_p(Z_1(x, a), Z_2(x, a)).$$

We will use $\bar{d}_p$ to establish the convergence of the distributional Bellman operators.

Lemma 2. *$\bar{d}_p$ is a metric over value distributions.*

3.3. Policy Evaluation

In the *policy evaluation* setting (Sutton & Barto, 1998) we are interested in the value function V^π associated with a given policy π. The analogue here is the value distribution Z^π. In this section we characterize Z^π and study the behaviour of the policy evaluation operator $\mathcal{T}^\pi$. We emphasize that Z^π describes the intrinsic randomness of the agent's interactions with its environment, rather than some measure of uncertainty about the environment itself.

We view the reward function as a random vector $R \in \mathcal{Z}$, and define the transition operator $P^\pi : \mathcal{Z} \to \mathcal{Z}$

$$P^\pi Z(x, a) \overset{D}{:=} Z(X', A') \tag{4}$$
$$X' \sim P(\cdot \,|\, x, a),\ A' \sim \pi(\cdot \,|\, X'),$$

where we use capital letters to emphasize the random nature of the next state-action pair (X', A'). We define the distributional Bellman operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ as

$$\mathcal{T}^\pi Z(x, a) \overset{D}{:=} R(x, a) + \gamma P^\pi Z(x, a). \tag{5}$$

While $\mathcal{T}^\pi$ bears a surface resemblance to the usual Bellman operator (2), it is fundamentally different. In particular, three sources of randomness define the compound distribution $\mathcal{T}^\pi Z$:

a) The randomness in the reward R,
b) The randomness in the transition P^π, and
c) The next-state value distribution $Z(X', A')$.

In particular, we make the usual assumption that these three quantities are independent. In this section we will show that (5) is a contraction mapping whose unique fixed point is the random return Z^π.

3.3.1. Contraction in $\bar{d}_p$

Consider the process $Z_{k+1} := \mathcal{T}^\pi Z_k$, starting with some $Z_0 \in \mathcal{Z}$. We may expect the limiting expectation of $\{Z_k\}$ to converge exponentially quickly, as usual, to Q^π. As we now show, the process converges in a stronger sense: $\mathcal{T}^\pi$ is a contraction in $\bar{d}_p$, which implies that all moments also converge exponentially quickly.

Lemma 3. *$\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in $\bar{d}_p$.*

Using Lemma 3, we conclude using Banach's fixed point theorem that $\mathcal{T}^\pi$ has a unique fixed point. By inspection, this fixed point must be Z^π as defined in (1). As we assume all moments are bounded, this is sufficient to conclude that the sequence $\{Z_k\}$ converges to Z^π in $\bar{d}_p$ for $1 \leq p \leq \infty$.

To conclude, we remark that not all distributional metrics are equal; for example, Chung & Sobel (1987) have shown that $\mathcal{T}^\pi$ is not a contraction in total variation distance. Similar results can be derived for the Kullback-Leibler divergence and the Kolmogorov distance.

3.3.2. Contraction in Centered Moments

Observe that $d_2(U, V)$ (and more generally, d_p) relates to a coupling $C(\omega) := U(\omega) - V(\omega)$, in the sense that

$$d_2^2(U, V) \leq \mathbb{E}[(U - V)^2] = \mathbb{V}(C) + \big(\mathbb{E}\, C\big)^2.$$

As a result, we cannot directly use d_2 to bound the variance difference $|\mathbb{V}(\mathcal{T}^\pi Z(x, a)) - \mathbb{V}(Z^\pi(x, a))|$. However, $\mathcal{T}^\pi$ is in fact a contraction in variance (Sobel, 1982, see also appendix). In general, $\mathcal{T}^\pi$ is not a contraction in the p^{th} centered moment, $p > 2$, but the centered moments of the iterates $\{Z_k\}$ still converge exponentially quickly to those of Z^π; the proof extends the result of Rösler (1992).

3.4. Control

Thus far we have considered a fixed policy π, and studied the behaviour of its associated operator $\mathcal{T}^\pi$. We now set out to understand the distributional operators of the *control* setting – where we seek a policy π that maximizes value – and the corresponding notion of an optimal value distribution. As with the optimal value function, this notion is intimately tied to that of an optimal policy. However, while all optimal policies attain the same value Q^*, in our case a difficulty arises: in general there are many optimal value distributions.

In this section we show that the distributional analogue of the Bellman optimality operator converges, in a weak sense, to the set of optimal value distributions. However, this operator is not a contraction in any metric between distributions, and is in general much more temperamental than the policy evaluation operators. We believe the convergence issues we outline here are a symptom of the inherent instability of greedy updates, as highlighted by e.g. Tsitsiklis (2002) and most recently Harutyunyan et al. (2016).

Let Π^* be the set of optimal policies. We begin by characterizing what we mean by an *optimal value distribution*.

Definition 1 (Optimal value distribution)**.** *An optimal value distribution is the v.d. of an optimal policy. The set of optimal value distributions is $\mathcal{Z}^* := \{Z^{\pi^*} : \pi^* \in \Pi^*\}$.*

We emphasize that not all value distributions with expectation Q^* are optimal: they must match the full distribution of the return under some optimal policy.

Definition 2. *A greedy policy π for $Z \in \mathcal{Z}$ maximizes the expectation of Z. The set of greedy policies for Z is*

$$\mathcal{G}_Z := \{\pi : \textstyle\sum_a \pi(a \,|\, x)\, \mathbb{E}\, Z(x, a) = \max_{a' \in \mathcal{A}} \mathbb{E}\, Z(x, a')\}.$$

Recall that the expected Bellman optimality operator $\mathcal{T}$ is

$$\mathcal{T} Q(x, a) = \mathbb{E}\, R(x, a) + \gamma\, \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \quad (6)$$

The maximization at x' corresponds to some greedy policy. Although this policy is implicit in (6), we cannot ignore it in the distributional setting. We will call a *distributional Bellman optimality operator* any operator $\mathcal{T}$ which implements a greedy selection rule, i.e.

$$\mathcal{T} Z = \mathcal{T}^\pi Z \text{ for some } \pi \in \mathcal{G}_Z.$$

As in the policy evaluation setting, we are interested in the behaviour of the iterates $Z_{k+1} := \mathcal{T} Z_k$, $Z_0 \in \mathcal{Z}$. Our first result is to assert that $\mathbb{E}\, Z_k$ behaves as expected.

Lemma 4. *Let $Z_1, Z_2 \in \mathcal{Z}$. Then*

$$\left\| \mathbb{E}\, \mathcal{T} Z_1 - \mathbb{E}\, \mathcal{T} Z_2 \right\|_\infty \leq \gamma \left\| \mathbb{E}\, Z_1 - \mathbb{E}\, Z_2 \right\|_\infty,$$

and in particular $\mathbb{E}\, Z_k \to Q^*$ exponentially quickly.

By inspecting Lemma 4, we might expect that Z_k converges quickly in $\bar{d}_p$ to some fixed point in $\mathcal{Z}^*$. Unfortunately, convergence is neither quick nor assured to reach a fixed point. In fact, the best we can hope for is pointwise convergence, not even to the set $\mathcal{Z}^*$ but to the larger set of *nonstationary optimal value distributions*.

Definition 3. *A nonstationary optimal value distribution Z^{**} is the value distribution corresponding to a sequence of optimal policies. The set of n.o.v.d. is $\mathcal{Z}^{**}$.*

Theorem 1 (Convergence in the control setting)**.** *Let $\mathcal{X}$ be measurable and suppose that $\mathcal{A}$ is finite. Then*

$$\lim_{k\to\infty} \inf_{Z^{**}\in\mathcal{Z}^{**}} d_p(Z_k(x,a), Z^{**}(x,a)) = 0 \quad \forall x, a.$$

If $\mathcal{X}$ is finite, then Z_k converges to $\mathcal{Z}^{**}$ uniformly. Furthermore, if there is a total ordering $\prec$ on Π^*, such that for any $Z^* \in \mathcal{Z}^*$,

$$\mathcal{T}Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*}, \ \pi \prec \pi' \ \ \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\}.$$

Then $\mathcal{T}$ has a unique fixed point $Z^* \in \mathcal{Z}^*$.

Comparing Theorem 1 to Lemma 4 reveals a significant difference between the distributional framework and the usual setting of expected return. While the mean of Z_k converges exponentially quickly to Q^*, its distribution need not be as well-behaved! To emphasize this difference, we now provide a number of negative results concerning $\mathcal{T}$.

Proposition 1. *The operator $\mathcal{T}$ is not a contraction.*

Consider the following example (Figure 2, left). There are two states, x_1 and x_2; a unique transition from x_1 to x_2; from x_2, action a_1 yields no reward, while the optimal action a_2 yields $1 + \epsilon$ or $-1 + \epsilon$ with equal probability. Both actions are terminal. There is a unique optimal policy and therefore a unique fixed point Z^*. Now consider Z as given in Figure 2 (right), and its distance to Z^*:

$$\bar{d}_1(Z, Z^*) = d_1(Z(x_2, a_2), Z^*(x_2, a_2)) = 2\epsilon,$$

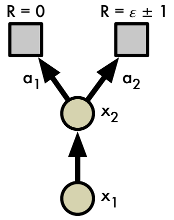
where we made use of the fact that $Z = Z^*$ everywhere except at (x_2, a_2). When we apply $\mathcal{T}$ to Z, however, the greedy action a_1 is selected and $\mathcal{T}Z(x_1) = Z(x_2, a_1)$. But

$$\begin{aligned} d_1(\mathcal{T}Z, \mathcal{T}Z^*) &= d_1(\mathcal{T}Z(x_1), Z^*(x_1)) \\ &= \tfrac{1}{2}|1-\epsilon| + \tfrac{1}{2}|1+\epsilon| > 2\epsilon \end{aligned}$$

for a sufficiently small ϵ. This shows that the undiscounted update is not a nonexpansion: $\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) > \bar{d}_1(Z, Z^*)$. With $\gamma < 1$, the same proof shows it is not a contraction. Using a more technically involved argument, we can extend this result to any metric which separates Z and $\mathcal{T}Z$.

Proposition 2. *Not all optimality operators have a fixed point $Z^* = \mathcal{T}Z^*$.*

To see this, consider the same example, now with $\epsilon = 0$, and a greedy operator $\mathcal{T}$ which breaks ties by picking a_2 if $Z(x_1) = 0$, and a_1 otherwise. Then the sequence $\mathcal{T}Z^*(x_1), (\mathcal{T})^2 Z^*(x_1), \ldots$ alternates between $Z^*(x_2, a_1)$ and $Z^*(x_2, a_2)$.



	x_1	x_2, a_1	x_2, a_2
Z^*	$\epsilon \pm 1$	0	$\epsilon \pm 1$
Z	$\epsilon \pm 1$	0	$-\epsilon \pm 1$
$\mathcal{T}Z$	0	0	$\epsilon \pm 1$

Figure 2. Undiscounted two-state MDP for which the optimality operator $\mathcal{T}$ is not a contraction, with example. The entries that contribute to $\bar{d}_1(Z, Z^*)$ and $\bar{d}_1(\mathcal{T}Z, Z^*)$ are highlighted.

Proposition 3. *That $\mathcal{T}$ has a fixed point $Z^* = \mathcal{T}Z^*$ is insufficient to guarantee the convergence of $\{Z_k\}$ to $\mathcal{Z}^*$.*

Theorem 1 paints a rather bleak picture of the control setting. It remains to be seen whether the dynamical eccentricies highlighted here actually arise in practice. One open question is whether theoretically more stable behaviour can be derived using stochastic policies, for example from conservative policy iteration (Kakade & Langford, 2002).

4. Approximate Distributional Learning

In this section we propose an algorithm based on the distributional Bellman optimality operator. In particular, this will require choosing an approximating distribution. Although the Gaussian case has previously been considered (Morimura et al., 2010a; Tamar et al., 2016), to the best of our knowledge we are the first to use a rich class of parametric distributions.

4.1. Parametric Distribution

We will model the value distribution using a discrete distribution parametrized by $N \in \mathbb{N}$ and $V_{\text{MIN}}, V_{\text{MAX}} \in \mathbb{R}$, and whose support is the set of atoms $\{z_i = V_{\text{MIN}} + i\triangle z : 0 \leq i < N\}$, $\triangle z := \frac{V_{\text{MAX}} - V_{\text{MIN}}}{N-1}$. In a sense, these atoms are the "canonical returns" of our distribution. The atom probabilities are given by a parametric model $\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^N$

$$Z_\theta(x,a) = z_i \quad \text{w.p. } \ p_i(x,a) := \frac{e^{\theta_i(x,a)}}{\sum_j e^{\theta_j(x,a)}}.$$

The discrete distribution has the advantages of being highly expressive and computationally friendly (see e.g. Van den Oord et al., 2016).

4.2. Projected Bellman Update

Using a discrete distribution poses a problem: the Bellman update $\mathcal{T}Z_\theta$ and our parametrization Z_θ almost always have disjoint supports. From the analysis of Section 3 it would seem natural to minimize the Wasserstein metric (viewed as a loss) between $\mathcal{T}Z_\theta$ and Z_θ, which is also conveniently robust to discrepancies in support. However, a second issue prevents this: in practice we are typically restricted to learning from sample transitions, which is not possible under the Wasserstein loss (see Prop. 5 and toy results in the appendix).

Instead, we project the sample Bellman update $\hat{\mathcal{T}}Z_\theta$ onto the support of Z_θ (Figure 1, Algorithm 1), effectively reducing the Bellman update to multiclass classification. Let π be the greedy policy w.r.t. $\mathbb{E}\, Z_\theta$. Given a sample transition (x, a, r, x'), we compute the Bellman update $\hat{\mathcal{T}}z_j := r + \gamma z_j$ for each atom z_j, then distribute its probability $p_j(x', \pi(x'))$ to the immediate neighbours of $\hat{\mathcal{T}}z_j$. The i^{th} component of the projected update $\Phi\hat{\mathcal{T}}Z_\theta(x,a)$ is

$$(\Phi\hat{\mathcal{T}}Z_\theta(x,a))_i = \sum_{j=0}^{N-1}\left[1 - \frac{|[\hat{\mathcal{T}}z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}} - z_i|}{\triangle z}\right]_0^1 p_j(x', \pi(x')), \tag{7}$$

where $[\cdot]_a^b$ bounds its argument in the range $[a, b]$.[1] As is usual, we view the next-state distribution as parametrized by a fixed parameter $\tilde{\theta}$. The sample loss $\mathcal{L}_{x,a}(\theta)$ is the cross-entropy term of the KL divergence

$$D_{\text{KL}}(\Phi\hat{\mathcal{T}}Z_{\tilde{\theta}}(x,a) \,\|\, Z_\theta(x,a)),$$

which is readily minimized e.g. using gradient descent. We call this choice of distribution and loss the *categorical algorithm*. When $N = 2$, a simple one-parameter alternative is $\Phi\hat{\mathcal{T}}Z_\theta(x,a) := [\mathbb{E}[\hat{\mathcal{T}}Z_\theta(x,a)] - V_{\text{MIN}})/\triangle z]_0^1$; we call this the *Bernoulli algorithm*. We note that, while these algorithms appear unrelated to the Wasserstein metric, recent work (Bellemare et al., 2017) hints at a deeper connection.

Algorithm 1 Categorical Algorithm

```
input  A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$
  $Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$
  $a^* \leftarrow \arg\max_a Q(x_{t+1}, a)$
  $m_i = 0, \quad i \in 0, \ldots, N-1$
  for $j \in 0, \ldots, N-1$ do
    # Compute the projection of $\hat{\mathcal{T}}z_j$ onto the support $\{z_i\}$
    $\hat{\mathcal{T}}z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}}$
    $b_j \leftarrow (\hat{\mathcal{T}}z_j - V_{\text{MIN}})/\Delta z$   # $b_j \in [0, N-1]$
    $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$
    # Distribute probability of $\hat{\mathcal{T}}z_j$
    $m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$
    $m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$
  end for
output  $-\sum_i m_i \log p_i(x_t, a_t)$   # Cross-entropy loss
```

5. Evaluation on Atari 2600 Games

To understand the approach in a complex setting, we applied the categorical algorithm to games from the Arcade Learning Environment (ALE; Bellemare et al., 2013). While the ALE is deterministic, stochasticity does occur in a number of guises: 1) from state aliasing, 2) learning from a nonstationary policy, and 3) from approximation errors. We used five training games (Fig 3) and 52 testing games.

For our study, we use the DQN architecture (Mnih et al., 2015), but output the atom probabilities $p_i(x, a)$ instead of action-values, and chose $V_{\text{MAX}} = -V_{\text{MIN}} = 10$ from preliminary experiments over the training games. We call the resulting architecture *Categorical DQN*. We replace the squared loss $(r + \gamma Q(x', \pi(x')) - Q(x, a))^2$ by $\mathcal{L}_{x,a}(\theta)$ and train the network to minimize this loss.[2] As in DQN, we use a simple ϵ-greedy policy over the expected action-values; we leave as future work the many ways in which an agent could select actions on the basis of the full distribution. The rest of our training regime matches Mnih et al.'s, including the use of a target network for $\tilde{\theta}$.

Figure 4 illustrates the typical value distributions we observed in our experiments. In this example, three actions (those including the button press) lead to the agent releasing its laser too early and eventually losing the game. The corresponding distributions reflect this: they assign a significant probability to 0 (the terminal value). The safe actions have similar distributions (LEFT, which tracks the invaders' movement, is slightly favoured). This example helps explain why our approach is so successful: the distributional update keeps separated the low-value, "losing" event from the high-value, "survival" event, rather than average them into one (unrealizable) expectation.[3]

One surprising fact is that the distributions are not concentrated on one or two values, in spite of the ALE's determinism, but are often close to Gaussians. We believe this is due to our discretizing the diffusion process induced by γ.

5.1. Varying the Number of Atoms

We began by studying our algorithm's performance on the training games in relation to the number of atoms (Figure 3). For this experiment, we set $\epsilon = 0.05$. From the data, it is clear that using too few atoms can lead to poor behaviour, and that more always increases performance; this is not immediately obvious as we may have expected to saturate the network capacity. The difference in performance between the 51-atom version and DQN is particularly striking: the latter is outperformed in all five games, and in SEAQUEST we attain state-of-the-art performance. As an additional point of the comparison, the single-parameter Bernoulli algorithm performs better than DQN in 3 games out of 5, and is most notably more robust in ASTERIX.

[1] Algorithm 1 computes this projection in time linear in N.

[2] For $N = 51$, our TensorFlow implementation trains at roughly 75% of DQN's speed.

[3] Video: http://youtu.be/yFBwyPuO2Vg.

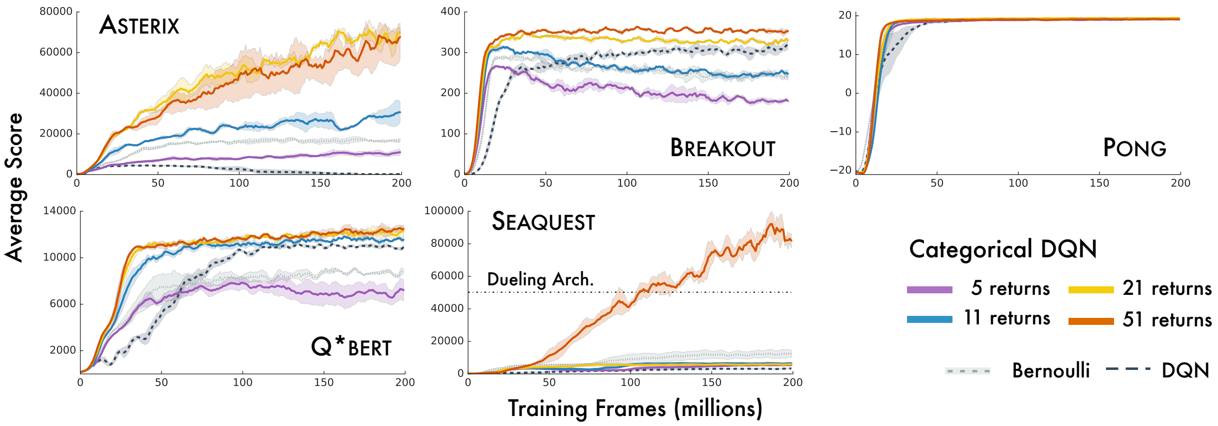


Figure 3. Categorical DQN: Varying number of atoms in the discrete distribution. Scores are moving averages over 5 million frames.

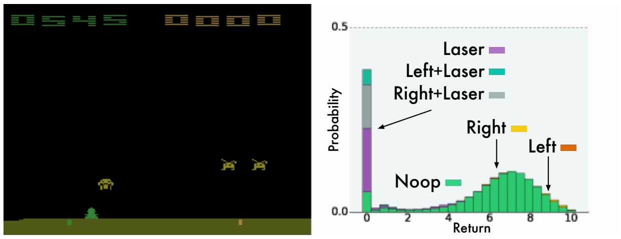


Figure 4. Learned value distribution during an episode of SPACE INVADERS. Different actions are shaded different colours. Returns below 0 (which do not occur in SPACE INVADERS) are not shown here as the agent assigns virtually no probability to them.

One interesting outcome of this experiment was to find out that our method does pick up on stochasticity. PONG exhibits intrinsic randomness: the exact timing of the reward depends on internal registers and is truly unobservable. We see this clearly reflected in the agent's prediction (Figure 5): over five consecutive frames, the value distribution shows two modes indicating the agent's belief that it has yet to receive a reward. Interestingly, since the agent's state does not include past rewards, it cannot even extinguish the prediction after receiving the reward, explaining the relative proportions of the modes.

5.2. State-of-the-Art Results

The performance of the 51-atom agent (from here onwards, C51) on the training games, presented in the last section, is particularly remarkable given that it involved none of the other algorithmic ideas present in state-of-the-art agents. We next asked whether incorporating the most common hyperparameter choice, namely a smaller training ϵ, could lead to even better results. Specifically, we set $\epsilon = 0.01$ (instead of 0.05); furthermore, every 1 million frames, we evaluate our agent's performance with $\epsilon = 0.001$.

We compare our algorithm to DQN ($\epsilon = 0.01$), Double DQN (van Hasselt et al., 2016), the Dueling architecture (Wang et al., 2016), and Prioritized Replay (Schaul et al., 2016), comparing the best evaluation score achieved during training. We see that C51 significantly outperforms these other algorithms (Figures 6 and 7). In fact, C51 surpasses the current state-of-the-art by a large margin in a number of games, most notably SEAQUEST. One particularly striking fact is the algorithm's good performance on sparse reward games, for example VENTURE and PRIVATE EYE. This suggests that value distributions are better able to propagate rarely occurring events. Full results are provided in the appendix.

We also include in the appendix (Figure 12) a comparison, averaged over 3 seeds, showing the number of games in which C51's training performance outperforms fully-trained DQN and human players. These results continue to show dramatic improvements, and are more representative of an agent's average performance. Within 50 million frames, C51 has outperformed a fully trained DQN agent on 45 out of 57 games. This suggests that the full 200 million training frames, and its ensuing computational cost, are unnecessary for evaluating reinforcement learning algorithms within the ALE.

The most recent version of the ALE contains a stochastic execution mechanism designed to ward against trajectory overfitting.Specifically, on each frame the environment rejects the agent's selected action with probability $p = 0.25$. Although DQN is mostly robust to stochastic execution, there are a few games in which its performance is reduced. On a score scale normalized with respect to the random and DQN agents, C51 obtains mean and median score improvements of 126% and 21.5% respectively, confirming the benefits of C51 beyond the deterministic setting.

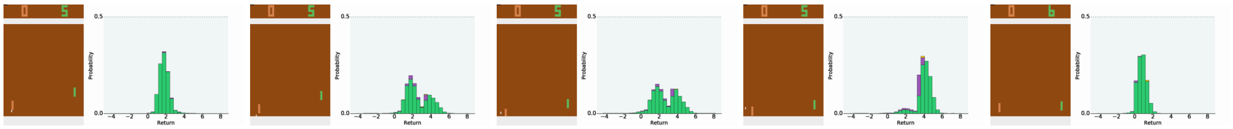


Figure 5. Intrinsic stochasticity in PONG.

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	**701%**	**178%**	**40**	**50**
UNREAL[†]	880%	250%	-	-

Figure 6. Mean and median scores across 57 Atari games, measured as percentages of human baseline (H.B., Nair et al., 2015).

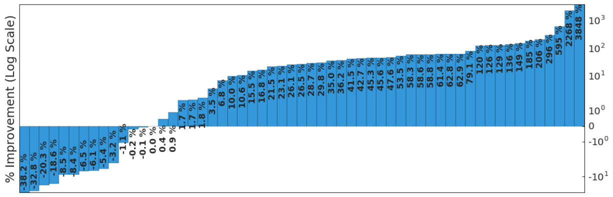


Figure 7. Percentage improvement, per-game, of C51 over Double DQN, computed using van Hasselt et al.'s method.

6. Discussion

In this work we sought a more complete picture of reinforcement learning, one that involves value distributions. We found that learning value distributions is a powerful notion that allows us to surpass most gains previously made on Atari 2600, without further algorithmic adjustments.

6.1. Why does learning a distribution matter?

It is surprising that, when we use a policy which aims to maximize expected return, we should see any difference in performance. The distinction we wish to make is that *learning distributions matters in the presence of approximation*. We now outline some possible reasons.

Reduced chattering. Our results from Section 3.4 highlighted a significant instability in the Bellman optimality operator. When combined with function approximation, this instability may prevent the policy from converging, what Gordon (1995) called *chattering*. We believe the gradient-based categorical algorithm is able to mitigate these effects by effectively averaging the different distributions, similar to conservative policy iteration (Kakade & Langford, 2002). While the chattering persists, it is integrated to the approximate solution.

State aliasing. Even in a deterministic environment, state aliasing may result in effective stochasticity. McCallum (1995), for example, showed the importance of coupling representation learning with policy learning in partially observable domains. We saw an example of state aliasing in PONG, where the agent could not exactly predict the reward timing. Again, by explicitly modelling the resulting distribution we provide a more stable learning target.

A richer set of predictions. A recurring theme in artificial intelligence is the idea of an agent learning from a multitude of predictions (Caruana 1997; Utgoff & Stracuzzi 2002; Sutton et al. 2011; Jaderberg et al. 2017). The distributional approach naturally provides us with a rich set of auxiliary predictions, namely: the probability that the return will take on a particular value. Unlike previously proposed approaches, however, the accuracy of these predictions is tightly coupled with the agent's performance.

Framework for inductive bias. The distributional perspective on reinforcement learning allows a more natural framework within which we can impose assumptions about the domain or the learning problem itself. In this work we used distributions with support bounded in $[V_{\text{MIN}}, V_{\text{MAX}}]$. Treating this support as a hyperparameter allows us to change the optimization problem by treating all extremal returns (e.g. greater than V_{MAX}) as equivalent. Surprisingly, a similar value clipping in DQN significantly degrades performance in most games. To take another example: interpreting the discount factor γ as a proper probability, as some authors have argued, leads to a different algorithm.

Well-behaved optimization. It is well-accepted that the KL divergence between categorical distributions is a reasonably easy loss to minimize. This may explain some of our empirical performance. Yet early experiments with alternative losses, such as KL divergence between continuous densities, were not fruitful, in part because the KL divergence is insensitive to the values of its outcomes. A closer minimization of the Wasserstein metric should yield even better results than what we presented here.

In closing, we believe our results highlight the need to account for distribution in the design, theoretical or otherwise, of algorithms.

[†] The UNREAL results are not altogether comparable, as they were generated in the asynchronous setting with per-game hyperparameter tuning (Jaderberg et al., 2017).

Acknowledgements

The authors acknowledge the important role played by their colleagues at DeepMind throughout the development of this work. Special thanks to Yee Whye Teh, Alex Graves, Joel Veness, Guillaume Desjardins, Tom Schaul, David Silver, Andre Barreto, Max Jaderberg, Mohammad Azar, Georg Ostrovski, Bernardo Avila Pires, Olivier Pietquin, Audrunas Gruslys, Tom Stepleton, Aaron van den Oord; and particularly Chris Maddison for his comprehensive review of an earlier draft. Thanks also to Marek Petrik for pointers to the relevant literature, and Mark Rowland for fine-tuning details in the final version.

Erratum

The camera-ready copy of this paper incorrectly reported a mean score of 1010% for C51. The corrected figure stands at 701%, which remains higher than the other comparable baselines. The median score remains unchanged at 178%.

The error was due to evaluation episodes in one game (Atlantis) lasting over 30 minutes; in comparison, the other results presented here cap episodes at 30 minutes, as is standard. The previously reported score on Atlantis was 3.7 million; our 30-minute score is 841,075, which we believe is close to the achievable maximum in this time frame. Capping at 30 minutes brings our human-normalized score on Atlantis from 22824% to a mere (!) 5199%, unfortunately enough to noticeably affect the mean score, whose sensitivity to outliers is well-documented.

References

Azar, Mohammad Gheshlaghi, Munos, Rémi, and Kappen, Hilbert. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the International Conference on Machine Learning*, 2012.

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*, 47:253–279, 2013.

Bellemare, Marc G., Danihelka, Ivo, Dabney, Will, Mohamed, Shakir, Lakshminarayanan, Balaji, Hoyer, Stephan, and Munos, Rémi. The cramer distance as a solution to biased wasserstein gradients. *arXiv*, 2017.

Bellman, Richard E. *Dynamic programming*. Princeton University Press, Princeton, NJ, 1957.

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

Bickel, Peter J. and Freedman, David A. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pp. 1196–1217, 1981.

Billingsley, Patrick. *Probability and measure*. John Wiley & Sons, 1995.

Caruana, Rich. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.

Chung, Kun-Jen and Sobel, Matthew J. Discounted mdps: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

Dearden, Richard, Friedman, Nir, and Russell, Stuart. Bayesian Q-learning. In *Proceedings of the National Conference on Artificial Intelligence*, 1998.

Engel, Yaakov, Mannor, Shie, and Meir, Ron. Reinforcement learning with gaussian processes. In *Proceedings of the International Conference on Machine Learning*, 2005.

Geist, Matthieu and Pietquin, Olivier. Kalman temporal differences. *Journal of Artificial Intelligence Research*, 39:483–532, 2010.

Gordon, Geoffrey. Stable function approximation in dynamic programming. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

Harutyunyan, Anna, Bellemare, Marc G., Stepleton, Tom, and Munos, Rémi. Q(λ) with off-policy corrections. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2016.

Hoffman, Matthew D., de Freitas, Nando, Doucet, Arnaud, and Peters, Jan. An expectation maximization algorithm for continuous markov decision processes with arbitrary reward. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009.

Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z, Silver, David, and Kavukcuoglu, Koray. Reinforcement learning with unsupervised auxiliary tasks. *Proceedings of the International Conference on Learning Representations*, 2017.

Jaquette, Stratton C. Markov decision processes with a new optimality criterion: Discrete time. *The Annals of Statistics*, 1(3): 496–505, 1973.

Kakade, Sham and Langford, John. Approximately optimal approximate reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2002.

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

Lattimore, Tor and Hutter, Marcus. PAC bounds for discounted MDPs. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2012.

Mannor, Shie and Tsitsiklis, John N. Mean-variance optimization in markov decision processes. 2011.

McCallum, Andrew K. *Reinforcement learning with selective perception and hidden state*. PhD thesis, University of Rochester, 1995.

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.

Morimura, Tetsuro, Hachiya, Hirotaka, Sugiyama, Masashi, Tanaka, Toshiyuki, and Kashima, Hisashi. Parametric return density estimation for reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010a.

Morimura, Tetsuro, Sugiyama, Masashi, Kashima, Hisashi, Hachiya, Hirotaka, and Tanaka, Toshiyuki. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 799–806, 2010b.

Nair, Arun, Srinivasan, Praveen, Blackwell, Sam, Alcicek, Cagdas, Fearon, Rory, De Maria, Alessandro, Panneershelvam, Vedavyas, Suleyman, Mustafa, Beattie, Charles, and Petersen, Stig et al. Massively parallel methods for deep reinforcement learning. In *ICML Workshop on Deep Learning*, 2015.

Prashanth, LA and Ghavamzadeh, Mohammad. Actor-critic algorithms for risk-sensitive mdps. In *Advances in Neural Information Processing Systems*, 2013.

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

Rösler, Uwe. A fixed point theorem for distributions. *Stochastic Processes and their Applications*, 42(2):195–214, 1992.

Schaul, Tom, Quan, John, Antonoglou, Ioannis, and Silver, David. Prioritized experience replay. In *Proceedings of the International Conference on Learning Representations*, 2016.

Sobel, Matthew J. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(04):794–802, 1982.

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

Sutton, R.S., Modayil, J., Delp, M., Degris, T., Pilarski, P.M., White, A., and Precup, D. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the International Conference on Autonomous Agents and Multiagents Systems*, 2011.

Tamar, Aviv, Di Castro, Dotan, and Mannor, Shie. Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36, 2016.

Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2), 2012.

Toussaint, Marc and Storkey, Amos. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the International Conference on Machine Learning*, 2006.

Tsitsiklis, John N. On the convergence of optimistic policy iteration. *Journal of Machine Learning Research*, 3:59–72, 2002.

Utgoff, Paul E. and Stracuzzi, David J. Many-layered learning. *Neural Computation*, 14(10):2497–2529, 2002.

Van den Oord, Aaron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*, 2016.

van Hasselt, Hado, Guez, Arthur, and Silver, David. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Veness, Joel, Bellemare, Marc G., Hutter, Marcus, Chua, Alvin, and Desjardins, Guillaume. Compress and control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Wang, Tao, Lizotte, Daniel, Bowling, Michael, and Schuurmans, Dale. Dual representations for dynamic programming. *Journal of Machine Learning Research*, pp. 1–29, 2008.

Wang, Ziyu, Schaul, Tom, Hessel, Matteo, Hasselt, Hado van, Lanctot, Marc, and de Freitas, Nando. Dueling network architectures for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2016.

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

A. Related Work

To the best of our knowledge, the work closest to ours are two papers (Morimura et al., 2010b;a) studying the distributional Bellman equation from the perspective of its cumulative distribution functions. The authors propose both parametric and nonparametric solutions to learn distributions for risk-sensitive reinforcement learning. They also provide some theoretical analysis for the policy evaluation setting, including a consistency result in the nonparametric case. By contrast, we also analyze the control setting, and emphasize the use of the distributional equations to improve approximate reinforcement learning.

The variance of the return has been extensively studied in the risk-sensitive setting. Of note, Tamar et al. (2016) analyze the use of linear function approximation to learn this variance for policy evaluation, and Prashanth & Ghavamzadeh (2013) estimate the return variance in the design of a risk-sensitive actor-critic algorithm. Mannor & Tsitsiklis (2011) provides negative results regarding the computation of a variance-constrained solution to the optimal control problem.

The distributional formulation also arises when modelling uncertainty. Dearden et al. (1998) considered a Gaussian approximation to the value distribution, and modelled the uncertainty over the parameters of this approximation using a Normal-Gamma prior. Engel et al. (2005) leveraged the distributional Bellman equation to define a Gaussian process over the unknown value function. More recently, Geist & Pietquin (2010) proposed an alternative solution to the same problem based on unscented Kalman filters. We believe much of the analysis we provide here, which deals with the intrinsic randomness of the environment, can also be applied to modelling uncertainty.

Our work here is based on a number of foundational results, in particular concerning alternative optimality criteria. Early on, Jaquette (1973) showed that a *moment optimality* criterion, which imposes a total ordering on distributions, is achievable and defines a stationary optimal policy, echoing the second part of Theorem 1. Sobel (1982) is usually cited as the first reference to Bellman equations for the higher moments (but not the distribution) of the return. Chung & Sobel (1987) provides results concerning the convergence of the distributional Bellman operator in total variation distance. White (1988) studies "nonstandard MDP criteria" from the perspective of optimizing the state-action pair occupancy.

A number of probabilistic frameworks for reinforcement learning have been proposed in recent years. The *planning as inference* approach (Toussaint & Storkey, 2006; Hoffman et al., 2009) embeds the return into a graphical model, and applies probabilistic inference to determine the sequence of actions leading to maximal expected reward. Wang et al. (2008) considered the dual formulation of reinforcement learning, where one optimizes the stationary distribution subject to constraints given by the transition function (Puterman, 1994), in particular its relationship to linear approximation. Related to this dual is the Compress and Control algorithm Veness et al. (2015), which describes a value function by learning a return distribution using density models. One of the aims of this work was to address the question left open by their work of whether one could be design a practical distributional algorithm based on the Bellman equation, rather than Monte Carlo estimation.

B. Proofs

Lemma 1 (Partition lemma). *Let $A_1, A_2, \ldots$ be a set of random variables describing a partition of Ω, i.e. $A_i(\omega) \in \{0, 1\}$ and for any ω there is exactly one A_i with $A_i(\omega) = 1$. Let U, V be two random variables. Then*

$$d_p(U, V) \leq \sum\nolimits_i d_p(A_i U, A_i V).$$

Proof. We will give the proof for $p < \infty$, noting that the same applies to $p = \infty$. Let $Y_i \overset{D}{:=} A_i U$ and $Z_i \overset{D}{:=} A_i V$, respectively. First note that

$$\begin{aligned} d_p^p(A_i U, A_i V) &= \inf_{Y_i, Z_i} \mathbb{E}\left[|Y_i - Z_i|^p\right] \\ &= \inf_{Y_i, Z_i} \mathbb{E}\Big[\mathbb{E}\left[|Y_i - Z_i|^p \,|\, A_i\right]\Big]. \end{aligned}$$

Now, $|A_i U - A_i V|^p = 0$ whenever $A_i = 0$. It follows that we can choose Y_i, Z_i so that also $|Y_i - Z_i|^p = 0$ whenever $A_i = 0$, without increasing the expected norm. Hence

$$d_p^p(A_i U, A_i V) = \inf_{Y_i, Z_i} \Pr\{A_i = 1\} \, \mathbb{E}\left[|Y_i - Z_i|^p \,|\, A_i = 1\right]. \tag{8}$$

Next, we claim that

$$\inf_{U,V} \sum\nolimits_i \Pr\{A_i = 1\} \, \mathbb{E}\Big[\big|A_i U - A_i V\big|^p \,|\, A_i = 1\Big] \tag{9}$$

$$\leq \inf_{\substack{Y_1, Y_2, \ldots \\ Z_1, Z_2, \ldots}} \sum\nolimits_i \Pr\{A_i = 1\} \, \mathbb{E}\Big[\big|Y_i - Z_i\big|^p \,|\, A_i = 1\Big].$$

Specifically, the left-hand side of the equation is an infimum over all r.v.'s whose cumulative distributions are F_U and F_V, respectively, while the right-hand side is an infimum over sequences of r.v's $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$ whose cumulative distributions are $F_{A_i U}, F_{A_i V}$, respectively. To prove this upper bound, consider the c.d.f. of U:

$$\begin{aligned} F_U(y) &= \Pr\{U \leq y\} \\ &= \sum\nolimits_i \Pr\{A_i = 1\} \Pr\{U \leq y \,|\, A_i = 1\} \\ &= \sum\nolimits_i \Pr\{A_i = 1\} \Pr\{A_i U \leq y \,|\, A_i = 1\}. \end{aligned}$$

Hence the distribution F_U is equivalent, in an almost sure sense, to one that first picks an element A_i of the partition, then picks a value for U conditional on the choice A_i. On the other hand, the c.d.f. of $Y_i \overset{D}{=} A_i U$ is

$$\begin{aligned} F_{A_i U}(y) &= \Pr\{A_i = 1\} \Pr\{A_i U \le y \,|\, A_i = 1\} \\ &\quad + \Pr\{A_i = 0\} \Pr\{A_i U \le y \,|\, A_i = 0\} \\ &= \Pr\{A_i = 1\} \Pr\{A_i U \le y \,|\, A_i = 1\} \\ &\quad + \Pr\{A_i = 0\} \mathbb{I}\left[y \ge 0\right]. \end{aligned}$$

Thus the right-hand side infimum in (9) has the additional constraint that it must preserve the conditional c.d.fs, in particular when $y \ge 0$. Put another way, instead of having the freedom to completely reorder the mapping $U : \Omega \to \mathbb{R}$, we can only reorder it within each element of the partition. We now write

$$\begin{aligned} d_p^p(U, V) &= \inf_{U,V} \|U - V\|_p \\ &= \inf_{U,V} \mathbb{E}\left[|U - V|^p\right] \\ &\overset{(a)}{=} \inf_{U,V} \sum\nolimits_i \Pr\{A_i = 1\} \mathbb{E}\left[|U - V|^p \,|\, A_i = 1\right] \\ &= \inf_{U,V} \sum\nolimits_i \Pr\{A_i = 1\} \mathbb{E}\left[|A_i U - A_i V|^p \,|\, A_i = 1\right], \end{aligned}$$

where (a) follows because $A_1, A_2, \dots$ is a partition. Using (9), this implies

$$\begin{aligned} & d_p^p(U, V) \\ &= \inf_{U,V} \sum\nolimits_i \Pr\{A_i = 1\} \mathbb{E}\left[\left|A_i U - A_i V\right|^p \,|\, A_i = 1\right] \\ &\le \inf_{\substack{Y_1, Y_2, \dots \\ Z_1, Z_2, \dots}} \sum\nolimits_i \Pr\{A_i = 1\} \mathbb{E}\left[\left|Y_i - Z_i\right|^p \,|\, A_i = 1\right] \\ &\overset{(b)}{=} \sum\nolimits_i \inf_{Y_i, Z_i} \Pr\{A_i = 1\} \mathbb{E}\left[\left|Y_i - Z_i\right|^p \,|\, A_i = 1\right] \\ &\overset{(c)}{=} \sum\nolimits_i d_p(A_i U, A_i V), \end{aligned}$$

because in (b) the individual components of the sum are independently minimized; and (c) from (8). □

Lemma 2. *$\bar{d}_p$ is a metric over value distributions.*

Proof. The only nontrivial property is the triangle inequality. For any value distribution $Y \in \mathcal{Z}$, write

$$\begin{aligned} & \bar{d}_p(Z_1, Z_2) = \sup_{x,a} d_p(Z_1(x,a), Z_2(x,a)) \\ &\overset{(a)}{\le} \sup_{x,a} \left[d_p(Z_1(x,a), Y(x,a)) + d_p(Y(x,a), Z_2(x,a))\right] \\ &\le \sup_{x,a} d_p(Z_1(x,a), Y(x,a)) + \sup_{x,a} d_p(Y(x,a), Z_2(x,a)) \\ &= \bar{d}_p(Z_1, Y) + \bar{d}_p(Y, Z_2), \end{aligned}$$

where in (a) we used the triangle inequality for d_p. □

Lemma 3. *$\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in $\bar{d}_p$.*

Proof. Consider $Z_1, Z_2 \in \mathcal{Z}$. By definition,

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) = \sup_{x,a} d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)). \tag{10}$$

By the properties of d_p, we have

$$\begin{aligned} & d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)) \\ &\quad = d_p(R(x,a) + \gamma P^\pi Z_1(x,a), R(x,a) + \gamma P^\pi Z_2(x,a)) \\ &\quad \le \gamma d_p(P^\pi Z_1(x,a), P^\pi Z_2(x,a)) \\ &\quad \le \gamma \sup_{x',a'} d_p(Z_1(x',a'), Z_2(x',a')), \end{aligned}$$

where the last line follows from the definition of P^π (see (4)). Combining with (10) we obtain

$$\begin{aligned} \bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \sup_{x,a} d_p(\mathcal{T}^\pi Z_1(x,a), \mathcal{T}^\pi Z_2(x,a)) \\ &\le \gamma \sup_{x',a'} d_p(Z_1(x',a'), Z_2(x',a')) \\ &= \gamma \bar{d}_p(Z_1, Z_2). \end{aligned}$$

□

Proposition 1 (Sobel, 1982)**.** *Consider two value distributions $Z_1, Z_2 \in \mathcal{Z}$, and write $\mathbb{V}(Z_i)$ to be the vector of variances of Z_i. Then*

$$\|\mathbb{E}\, \mathcal{T}^\pi Z_1 - \mathbb{E}\, \mathcal{T}^\pi Z_2\|_\infty \le \gamma \|\mathbb{E}\, Z_1 - \mathbb{E}\, Z_2\|_\infty, \text{ and}$$
$$\|\mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2)\|_\infty \le \gamma^2 \|\mathbb{V} Z_1 - \mathbb{V} Z_2\|_\infty.$$

Proof. The first statement is standard, and its proof follows from $\mathbb{E}\, \mathcal{T}^\pi Z = \mathcal{T}^\pi\, \mathbb{E}\, Z$, where the second $\mathcal{T}^\pi$ denotes the usual operator over value functions. Now, by independence of R and $P^\pi Z_i$:

$$\begin{aligned} \mathbb{V}(\mathcal{T}^\pi Z_i(x,a)) &= \mathbb{V}\Big(R(x,a) + \gamma P^\pi Z_i(x,a)\Big) \\ &= \mathbb{V}(R(x,a)) + \gamma^2 \mathbb{V}(P^\pi Z_i(x,a)). \end{aligned}$$

And now

$$\begin{aligned} & \|\mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2)\|_\infty \\ &\quad = \sup_{x,a} \left|\mathbb{V}(\mathcal{T}^\pi Z_1(x,a)) - \mathbb{V}(\mathcal{T}^\pi Z_2(x,a))\right| \\ &\quad = \sup_{x,a} \gamma^2 \left| \left[\mathbb{V}(P^\pi Z_1(x,a)) - \mathbb{V}(P^\pi Z_2(x,a))\right] \right| \\ &\quad = \sup_{x,a} \gamma^2 \left| \mathbb{E}\left[\mathbb{V}(Z_1(X',A')) - \mathbb{V}(Z_2(X',A'))\right] \right| \\ &\quad \le \sup_{x',a'} \gamma^2 \left|\mathbb{V}(Z_1(x',a')) - \mathbb{V}(Z_2(x',a'))\right| \\ &\quad \le \gamma^2 \|\mathbb{V} Z_1 - \mathbb{V} Z_2\|_\infty. \end{aligned}$$

□

Lemma 4. *Let $Z_1, Z_2 \in \mathcal{Z}$. Then*

$$\|\mathbb{E}\, \mathcal{T} Z_1 - \mathbb{E}\, \mathcal{T} Z_2\|_\infty \le \gamma \|\mathbb{E}\, Z_1 - \mathbb{E}\, Z_2\|_\infty,$$

and in particular $\mathbb{E}\, Z_k \to Q^*$ exponentially quickly.

Proof. The proof follows by linearity of expectation. Write $\mathcal{T}_D$ for the distributional operator and $\mathcal{T}_E$ for the usual operator. Then

$$\begin{aligned}\|\mathbb{E}\,\mathcal{T}_D Z_1 - \mathbb{E}\,\mathcal{T}_D Z_2\|_\infty &= \|\mathcal{T}_E\,\mathbb{E}\,Z_1 - \mathcal{T}_E\,\mathbb{E}\,Z_2\|_\infty \\ &\le \gamma\,\|Z_1 - Z_2\|_\infty . \qquad \square\end{aligned}$$

Theorem 1 (Convergence in the control setting)**.** *Let $Z_k := \mathcal{T} Z_{k-1}$ with $Z_0 \in \mathcal{Z}$. Let $\mathcal{X}$ be measurable and suppose that $\mathcal{A}$ is finite. Then*

$$\lim_{k\to\infty} \inf_{Z^{**}\in\mathcal{Z}^{**}} d_p(Z_k(x,a), Z^{**}(x,a)) = 0 \quad \forall x, a.$$

If $\mathcal{X}$ is finite, then Z_k converges to $\mathcal{Z}^{**}$ uniformly. Furthermore, if there is a total ordering $\prec$ on Π^*, such that for any $Z^* \in \mathcal{Z}^*$,

$$\mathcal{T} Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*},\ \pi \prec \pi' \ \ \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\},$$

then $\mathcal{T}$ has a unique fixed point $Z^* \in \mathcal{Z}^*$.

The gist of the proof of Theorem 1 consists in showing that for every state x, there is a time k after which the greedy policy w.r.t. Q_k is mostly optimal. To clearly expose the steps involved, we will first assume a unique (and therefore deterministic) optimal policy π^*, and later return to the general case; we will denote the optimal action at x by $\pi^*(x)$. For notational convenience, we will write $Q_k := \mathbb{E}\,Z_k$ and $\mathcal{G}_k := \mathcal{G}_{Z_k}$. Let $B := 2\sup_{Z\in\mathcal{Z}} \|Z\|_\infty < \infty$ and let $\epsilon_k := \gamma^k B$. We first define the set of states $\mathcal{X}_k \subseteq \mathcal{X}$ whose values must be sufficiently close to Q^* at time k:

$$\mathcal{X}_k := \Big\{x : Q^*(x, \pi^*(x)) - \max_{a\neq\pi^*(x)} Q^*(x,a) > 2\epsilon_k\Big\}. \tag{11}$$

Indeed, by Lemma 4, we know that after k iterations

$$|Q_k(x,a) - Q^*(x,a)| \le \gamma^k |Q_0(x,a) - Q^*(x,a)| \le \epsilon_k.$$

For $x \in \mathcal{X}$, write $a^* := \pi^*(x)$. For any $a \in \mathcal{A}$, we deduce that

$$Q_k(x,a^*) - Q_k(x,a) \ge Q^*(x,a^*) - Q^*(x,a) - 2\epsilon_k.$$

It follows that if $x \in \mathcal{X}_k$, then also $Q_k(x,a^*) > Q_k(x,a')$ for all $a' \neq \pi^*(x)$: for these states, the greedy policy $\pi_k(x) := \arg\max_a Q_k(x,a)$ corresponds to the optimal policy π^*.

Lemma 5. *For each $x \in \mathcal{X}$ there exists a k such that, for all $k' \ge k$, $x \in \mathcal{X}_{k'}$, and in particular $\arg\max_a Q_k(x,a) = \pi^*(x)$.*

Proof. Because $\mathcal{A}$ is finite, the gap

$$\Delta(x) := Q^*(x, \pi^*(x)) - \max_{a\neq\pi^*(x)} Q^*(x,a)$$

is attained for some strictly positive $\Delta(x) > 0$. By definition, there exists a k such that

$$\epsilon_k = \gamma^k B < \frac{\Delta(x)}{2},$$

and hence every $x \in \mathcal{X}$ must eventually be in $\mathcal{X}_k$. □

This lemma allows us to guarantee the existence of an iteration k after which sufficiently many states are well-behaved, in the sense that the greedy policy at those states chooses the optimal action. We will call these states "solved". We in fact require not only these states to be solved, but also most of their successors, and most of the successors of those, and so on. We formalize this notion as follows: fix some $\delta > 0$, let $\mathcal{X}_{k,0} := \mathcal{X}_k$, and define for $i > 0$ the set

$$\mathcal{X}_{k,i} := \big\{x : x \in \mathcal{X}_k, P(\mathcal{X}_{k-1,i-1} \,|\, x, \pi^*(x)) \ge 1 - \delta\big\},$$

As the following lemma shows, any x is eventually contained in the recursively-defined sets $\mathcal{X}_{k,i}$, for any i.

Lemma 6. *For any $i \in \mathbb{N}$ and any $x \in \mathcal{X}$, there exists a k such that for all $k' \ge k$, $x \in \mathcal{X}_{k',i}$.*

Proof. Fix i and let us suppose that $\mathcal{X}_{k,i} \uparrow \mathcal{X}$. By Lemma 5, this is true for $i = 0$. We infer that for any probability measure P on $\mathcal{X}$, $P(\mathcal{X}_{k,i}) \to P(\mathcal{X}) = 1$. In particular, for a given $x \in \mathcal{X}_k$, this implies that

$$P(\mathcal{X}_{k,i} \,|\, x, \pi^*(x)) \to P(\mathcal{X} \,|\, x, \pi^*(x)) = 1.$$

Therefore, for any x, there exists a time after which it is and remains a member of $\mathcal{X}_{k,i+1}$, the set of states for which $P(\mathcal{X}_{k-1,i} \,|\, x, \pi^*(x)) \ge 1 - \delta$. We conclude that $\mathcal{X}_{k,i+1} \uparrow \mathcal{X}$ also. The statement follows by induction. □

Proof of Theorem 1. The proof is similar to policy iteration-type results, but requires more care in dealing with the metric and the possibly infinite state space. We will write $W_k(x) := Z_k(x, \pi_k(x))$, define W^* similarly and with some overload of notation write $\mathcal{T} W_k(x) := W_{k+1}(x) = \mathcal{T} Z_k(x, \pi_{k+1}(x))$. Finally, let $S_i^k(x) := \mathbb{I}\left[x \in \mathcal{X}_{k,i}\right]$ and $\bar{S}_i^k(x) = 1 - S_i^k(x)$.

Fix $i > 0$ and $x \in \mathcal{X}_{k+1,i+1} \subseteq \mathcal{X}_k$. We begin by using Lemma 1 to separate the transition from x into a solved term and an unsolved term:

$$P^{\pi_k} W_k(x) = S_i^k W_k(X') + \bar{S}_i^k W_k(X'),$$

where X' is the random successor from taking action $\pi_k(x) := \pi^*(x)$, and we write $S_i^k = S_i^k(X'), \bar{S}_i^k = \bar{S}_i^k(X')$ to ease the notation. Similarly,

$$P^{\pi_k} W^*(x) = S_i^k W^*(X') + \bar{S}_i^k W^*(X').$$

Now

$$\begin{aligned} d_p(W_{k+1}(x), W^*(x)) &= d_p(\mathcal{T} W_k(x), \mathcal{T} W^*(x)) \\ &\overset{(a)}{\leq} \gamma d_p(P^{\pi_k} W_k(x), P^{\pi^*} W^*(x)) \\ &\overset{(b)}{\leq} \gamma d_p(S_i^k W_k(X'), S_i^k W^*(X')) \\ &\quad + \gamma d_p(\bar{S}_i^k W_k(X'), \bar{S}_i^k W^*(X')), \end{aligned} \tag{12}$$

where in (a) we used Properties P1 and P2 of the Wasserstein metric, and in (b) we separate states for which $\pi_k = \pi^*$ from the rest using Lemma 1 ($\{S_i^k, \bar{S}_i^k\}$ form a partition of Ω). Let $\delta_i := \Pr\{X' \notin \mathcal{X}_{k,i}\} = \mathbb{E}\{\bar{S}_i^k(X')\} = \|\bar{S}_i^k(X')\|_p$. From property P3 of the Wasserstein metric, we have

$$\begin{aligned} d_p(\bar{S}_i^k W_k(X'), \bar{S}_i^k W^*(X')) & \\ &\leq \sup_{x'} d_p(\bar{S}_i^k(X') W_k(x'), \bar{S}_i^k(X') W^*(x')) \\ &\leq \|\bar{S}_i^k(X')\|_p \sup_{x'} d_p(W_k(x'), W^*(x')) \\ &\leq \delta_i \sup_{x'} d_p(W_k(x'), W^*(x')) \\ &\leq \delta_i B. \end{aligned}$$

Recall that $B < \infty$ is the largest attainable $\|Z\|_\infty$. Since also $\delta_i < \delta$ by our choice of $x \in \mathcal{X}_{k+1,i+1}$, we can upper bound the second term in (12) by $\gamma \delta B$. This yields

$$\begin{aligned} d_p(W_{k+1}(x), W^*(x)) &\leq \\ &\gamma d_p(S_i^k W_k(X'), S_i^k W^*(X')) + \gamma \delta B. \end{aligned}$$

By induction on $i > 0$, we conclude that for $x \in \mathcal{X}_{k+i,i}$ and some random state X'' i steps forward,

$$\begin{aligned} d_p(W_{k+i}(x), W^*(x)) &\leq \\ &\gamma^i d_p(S_0^k W_k(X''), S_0^k W^*(X'')) + \frac{\delta B}{1-\gamma} \\ &\leq \gamma^i B + \frac{\delta B}{1-\gamma}. \end{aligned}$$

Hence for any $x \in \mathcal{X}$, $\epsilon > 0$, we can take δ, i, and finally k large enough to make $d_p(W_k(x), W^*(x)) < \epsilon$. The proof then extends to $Z_k(x, a)$ by considering one additional application of $\mathcal{T}$.

We now consider the more general case where there are multiple optimal policies. We expand the definition of $\mathcal{X}_{k,i}$ as follows:

$$\mathcal{X}_{k,i} := \big\{ x \in \mathcal{X}_k : \forall \pi^* \in \Pi^*, \mathop{\mathbb{E}}_{a^* \sim \pi^*(x)} P(\mathcal{X}_{k-1,i-1} \,|\, x, a^*) \geq 1 - \delta \big\},$$

Because there are finitely many actions, Lemma 6 also holds for this new definition. As before, take $x \in \mathcal{X}_{k,i}$, but now consider the sequence of greedy policies $\pi_k, \pi_{k-1}, \ldots$ selected by successive applications of $\mathcal{T}$, and write

$$\mathcal{T}^{\bar{\pi}_k} := \mathcal{T}^{\pi_k} \mathcal{T}^{\pi_{k-1}} \cdots \mathcal{T}^{\pi_{k-i+1}},$$

such that

$$Z_{k+1} = \mathcal{T}^{\bar{\pi}_k} Z_{k-i+1}.$$

Now denote by $\mathcal{Z}^{**}$ the set of nonstationary optimal policies. If we take any $Z^* \in \mathcal{Z}^*$, we deduce that

$$\inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(\mathcal{T}^{\bar{\pi}_k} Z^*(x, a), Z^{**}(x, a)) \leq \frac{\delta B}{1-\gamma},$$

since Z^* corresponds to some optimal policy π^* and $\bar{\pi}_k$ is optimal along most of the trajectories from (x, a). In effect, $\mathcal{T}^{\bar{\pi}_k} Z^*$ is close to the value distribution of the nonstationary optimal policy $\bar{\pi}_k \pi^*$. Now for this Z^*,

$$\begin{aligned} \inf_{Z^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) & \\ &\leq d_p(Z_k(x, a), \mathcal{T}^{\bar{\pi}_k} Z^*(x, a)) \\ &\quad + \inf_{Z^{**}} d_p(\mathcal{T}^{\bar{\pi}_k} Z^*(x, a), Z^{**}(x, a)) \\ &\leq d_p(\mathcal{T}^{\bar{\pi}_k} Z_{k-i+1}(x, a), \mathcal{T}^{\bar{\pi}_k} Z^*(x, a)) + \frac{\delta B}{1-\gamma} \\ &\leq \gamma^i B + \frac{2\delta B}{1-\gamma}, \end{aligned}$$

using the same argument as before with the newly-defined $\mathcal{X}_{k,i}$. It follows that

$$\inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) \to 0.$$

When $\mathcal{X}$ is finite, there exists a fixed k after which $\mathcal{X}_k = \mathcal{X}$. The uniform convergence result then follows.

To prove the uniqueness of the fixed point Z^* when $\mathcal{T}$ selects its actions according to the ordering $\prec$, we note that for any optimal value distribution Z^*, its set of greedy policies is Π^*. Denote by π^* the policy coming first in the ordering over Π^*. Then $\mathcal{T} = \mathcal{T}^{\pi^*}$, which has a unique fixed point (Section 3.3). □

Proposition 4. *That $\mathcal{T}$ has a fixed point $Z^* = \mathcal{T} Z^*$ is insufficient to guarantee the convergence of $\{Z_k\}$ to $\mathcal{Z}^*$.*

We provide here a sketch of the result. Consider a single state x_1 with two actions, a_1 and a_2 (Figure 8). The first action yields a reward of $1/2$, while the other either yields 0 or 1 with equal probability, and both actions are optimal. Now take $\gamma = 1/2$ and write $R_0, R_1, \ldots$ for the received rewards. Consider a stochastic policy that takes action a_2 with probability p. For $p = 0$, the return is

$$Z_{p=0} = \frac{1}{1-\gamma} \frac{1}{2} = 1.$$

For $p = 1$, on the other hand, the return is random and is given by the following fractional number (in binary):

$$Z_{p=1} = R_0.R_1 R_2 R_3 \cdots.$$

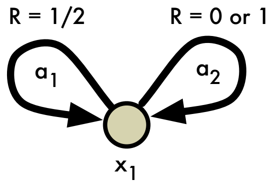


Figure 8. A simple example illustrating the effect of a nonstationary policy on the value distribution.

As a result, $Z_{p=1}$ is uniformly distributed between 0 and 2! In fact, note that

$$Z_{p=0} = 0.11111\cdots = 1.$$

For some intermediary value of p, we obtain a different probability of the different digits, but always putting some probability mass on all returns in $[0, 2]$.

Now suppose we follow the nonstationary policy that takes a_1 on the first step, then a_2 from there on. By inspection, the return will be uniformly distributed on the interval $[1/2, 3/2]$, which does not correspond to the return under any value of p. But now we may imagine an operator $\mathcal{T}$ which alternates between a_1 and a_2 depending on the exact value distribution it is applied to, which would in turn converge to a nonstationary optimal value distribution.

Lemma 7 (Sample Wasserstein distance)**.** *Let $\{P_i\}$ be a collection of random variables, $I \in \mathbb{N}$ a random index independent from $\{P_i\}$, and consider the mixture random variable $P = P_I$. For any random variable Q independent of I,*

$$d_p(P, Q) \leq \mathop{\mathbb{E}}_{i\sim I} d_p(P_i, Q),$$

and in general the inequality is strict and

$$\nabla_Q d_p(P_I, Q) \neq \mathop{\mathbb{E}}_{i\sim I} \nabla_Q d_p(P_i, Q).$$

Proof. We prove this using Lemma 1. Let $A_i := \mathbb{I}\left[I = i\right]$. We write

$$\begin{aligned} d_p(P, Q) &= d_p(P_I, Q) \\ &= d_p\Big(\sum\nolimits_i A_i P_i, \sum\nolimits_i A_i Q\Big) \\ &\leq \sum\nolimits_i d_p(A_i P_i, A_i Q) \\ &\leq \sum\nolimits_i \Pr\{I = i\} d_p(P_i, Q) \\ &= \mathbb{E}_I\, d_P(P_i, Q). \end{aligned}$$

where in the penultimate line we used the independence of I from P_i and Q to appeal to property P3 of the Wasserstein metric.

To show that the bound is in general strict, consider the mixture distribution depicted in Figure 9. We will simply consider the d_1 metric between this distribution P and another distribution Q. The first distribution is

$$P = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2. \end{cases}$$

In this example, $i \in \{1, 2\}$, $P_1 = 0$, and $P_2 = 1$. Now consider the distribution with the same support but that puts probability p on 0:

$$Q = \begin{cases} 0 & \text{w.p. } p \\ 1 & \text{w.p. } 1 - p. \end{cases}$$

The distance between P and Q is

$$d_1(P, Q) = |p - \tfrac{1}{2}|.$$

This is $d_1(P, Q) = \frac{1}{2}$ for $p \in \{0, 1\}$, and strictly less than $\frac{1}{2}$ for any other values of p. On the other hand, the corresponding expected distance (after sampling an outcome x_1 or x_2 with equal probability) is

$$\mathbb{E}_I\, d_1(P_i, Q) = \tfrac{1}{2}p + \tfrac{1}{2}(1 - p) = \tfrac{1}{2}.$$

Hence $d_1(P, Q) < \mathbb{E}_I\, d_1(P_i, Q)$ for $p \in (0, 1)$. This shows that the bound is in general strict. By inspection, it is clear that the two gradients are different. □

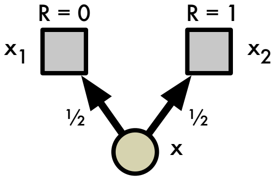


Figure 9. Example MDP in which the expected sample Wasserstein distance is greater than the Wasserstein distance.

Proposition 5. *Fix some next-state distribution Z and policy π. Consider a parametric value distribution Z_θ, and and define the Wasserstein loss*

$$\mathcal{L}_W(\theta) := d_p(Z_\theta(x, a), R(x, a) + \gamma Z(X', \pi(X'))).$$

Let $r \sim R(x, a)$ and $x' \sim P(\cdot\,|\,x, a)$ and consider the sample loss

$$L_W(\theta, r, x') := d_p(Z_\theta(x, a), r + \gamma Z(x', \pi(x'))).$$

Its expectation is an upper bound on the loss $\mathcal{L}_W$:

$$\mathcal{L}_W(\theta) \leq \mathop{\mathbb{E}}_{R,P} L_W(\theta, r, x'),$$

in general with strict inequality.

The result follows directly from the previous lemma.

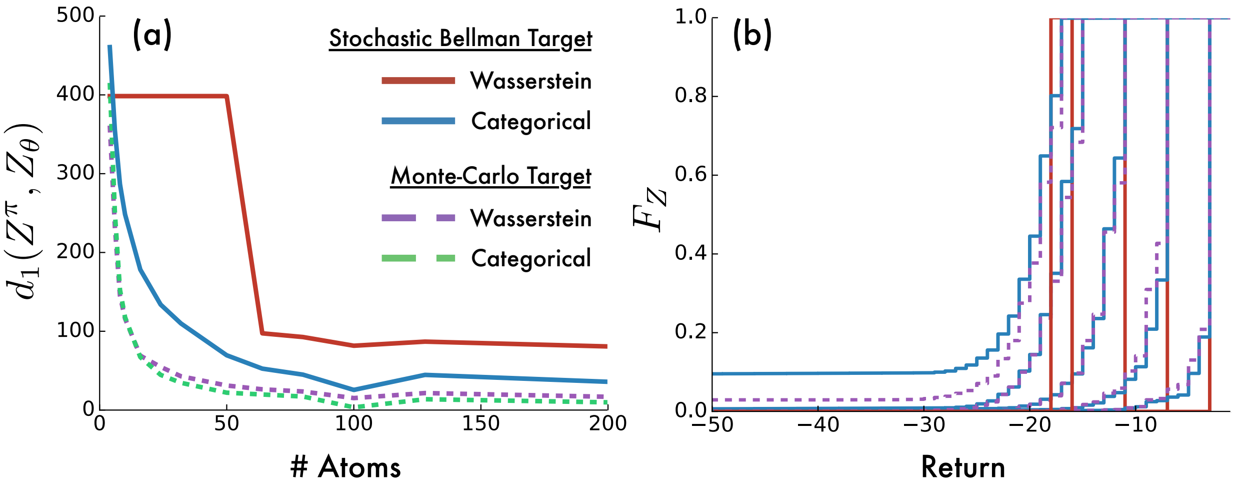


Figure 10. (a) Wasserstein distance between ground truth distribution Z^π and approximating distributions Z_θ. Varying number of atoms in approximation, training target, and loss function. (b) Approximate cumulative distributions for five representative states in CliffWalk.

C. Algorithmic Details

While our training regime closely follows that of DQN (Mnih et al., 2015), we use Adam (Kingma & Ba, 2015) instead of RMSProp (Tieleman & Hinton, 2012) for gradient rescaling. We also performed some hyperparameter tuning for our final results. Specifically, we evaluated two hyperparameters over our five training games and choose the values that performed best. The hyperparameter values we considered were $V_{\text{MAX}} \in \{3, 10, 100\}$ and $\epsilon_{adam} \in \{1/L, 0.1/L, 0.01/L, 0.001/L, 0.0001/L\}$, where $L = 32$ is the minibatch size. We found $V_{\text{MAX}} = 10$ and $\epsilon_{adam} = 0.01/L$ performed best. We used the same step-size value as DQN ($\alpha = 0.00025$).

Pseudo-code for the categorical algorithm is given in Algorithm 1. We apply the Bellman update to each atom separately, and then project it into the two nearest atoms in the original support. Transitions to a terminal state are handled with $\gamma_t = 0$.

D. Comparison of Sampled Wasserstein Loss and Categorical Projection

Lemma 3 proves that for a fixed policy π the distributional Bellman operator is a γ-contraction in $\bar{d}_p$, and therefore that $\mathcal{T}^\pi$ will converge in distribution to the true distribution of returns Z^π. In this section, we empirically validate these results on the CliffWalk domain shown in Figure 11. The dynamics of the problem match those given by Sutton & Barto (1998). We also study the convergence of the distributional Bellman operator under the sampled Wasserstein loss and the categorical projection (Equation 7) while following a policy that tries to take the safe path but has a 10% chance of taking another action uniformly at random.

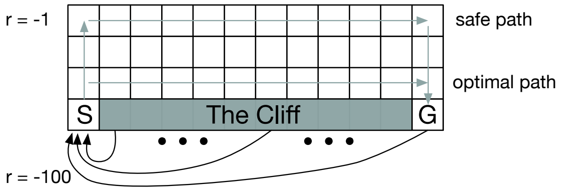


Figure 11. CliffWalk Environment (Sutton & Barto, 1998).

We compute a ground-truth distribution of returns Z^π using 10000 Monte-Carlo (MC) rollouts from each state. We then perform two experiments, approximating the value distribution at each state with our discrete distributions.

In the first experiment, we perform supervised learning using either the Wasserstein loss or categorical projection (Equation 7) with cross-entropy loss. We use Z^π as the supervised target and perform 5000 sweeps over all states to ensure both approaches have converged. In the second experiment, we use the same loss functions, but the training target comes from the one-step distributional Bellman operator with sampled transitions. We use $V_{\text{MIN}} = -100$ and $V_{\text{MAX}} = -1$.[4] For the sample updates we perform 10 times as many sweeps over the state space. Fundamentally, these experiments investigate how well the two training regimes

[4] Because there is a small probability of larger negative returns, some approximation error is unavoidable. However, this effect is relatively negligible in our experiments.

(minimizing the Wasserstein or categorical loss) minimize the Wasserstein metric under both ideal (supervised target) and practical (sampled one-step Bellman target) conditions.

In Figure 10a we show the final Wasserstein distance $d_1(Z^\pi, Z_\theta)$ between the learned distributions and the ground-truth distribution as we vary the number of atoms. The graph shows that the categorical algorithm does indeed minimize the Wasserstein metric in both the supervised and sample Bellman setting. It also highlights that minimizing the Wasserstein loss with stochastic gradient descent is in general flawed, confirming the intuition given by Proposition 5. In repeat experiments the process converged to different values of $d_1(Z^\pi, Z_\theta)$, suggesting the presence of local minima (more prevalent with fewer atoms).

Figure 10 provides additional insight into why the sampled Wasserstein distance may perform poorly. Here, we see the cumulative densities for the approximations learned under these two losses for five different states along the safe path in CliffWalk. The Wasserstein has converged to a fixed-point distribution, but not one that captures the true (Monte Carlo) distribution very well. By comparison, the categorical algorithm captures the variance of the true distribution much more accurately.

E. Supplemental Videos and Results

In Figure 13 we provide links to supplemental videos showing the C51 agent during training on various Atari 2600 games. Figure 12 shows the relative performance of C51 over the course of training. Figure 14 provides a table of evaluation results, comparing C51 to other state-of-the-art agents. Figures 15–18 depict particularly interesting frames.

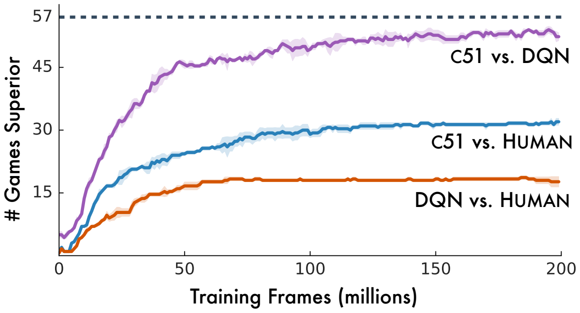


Figure 12. Number of Atari games where an agent's training performance is greater than a baseline (fully trained DQN & human). Error bands give standard deviations, and averages are over number of games.

GAMES	VIDEO URL
Freeway	http://youtu.be/97578n9kFIk
Pong	http://youtu.be/vIz5P6s80qA
Q*Bert	http://youtu.be/v-RbNX4uETw
Seaquest	http://youtu.be/d1yz4PNFUjI
Space Invaders	http://youtu.be/yFBwyPuO2Vg

Figure 13. Supplemental videos of C51 during training.

GAMES	RANDOM	HUMAN	DQN	DDQN	DUEL	PRIOR. DUEL.	C51
Alien	227.8	**7,127.7**	1,620.0	3,747.7	4,461.4	3,941.0	3,166
Amidar	5.8	1,719.5	978.0	1,793.3	**2,354.5**	2,296.8	1,735
Assault	222.4	742.0	4,280.4	5,393.2	4,621.0	**11,477.0**	7,203
Asterix	210.0	8,503.3	4,359.0	17,356.5	28,188.0	375,080.0	**406,211**
Asteroids	719.1	**47,388.7**	1,364.5	734.7	2,837.7	1,192.7	1,516
Atlantis	12,850.0	29,028.1	279,987.0	106,056.0	382,572.0	395,762.0	**841,075**
Bank Heist	14.2	753.1	455.0	1,030.6	**1,611.9**	1,503.1	976
Battle Zone	2,360.0	**37,187.5**	29,900.0	31,700.0	37,150.0	35,520.0	28,742
Beam Rider	363.9	16,926.5	8,627.5	13,772.8	12,164.0	**30,276.5**	14,074
Berzerk	123.7	2,630.4	585.6	1,225.4	1,472.6	**3,409.0**	1,645
Bowling	23.1	**160.7**	50.4	68.1	65.5	46.7	81.8
Boxing	0.1	12.1	88.0	91.6	**99.4**	98.9	97.8
Breakout	1.7	30.5	385.5	418.5	345.3	366.0	**748**
Centipede	2,090.9	**12,017.0**	4,657.7	5,409.4	7,561.4	7,687.5	9,646
Chopper Command	811.0	7,387.8	6,126.0	5,809.0	11,215.0	13,185.0	**15,600**
Crazy Climber	10,780.5	35,829.4	110,763.0	117,282.0	143,570.0	162,224.0	**179,877**
Defender	2,874.5	18,688.9	23,633.0	35,338.5	42,214.0	41,324.5	**47,092**
Demon Attack	152.1	1,971.0	12,149.4	58,044.2	60,813.3	72,878.6	**130,955**
Double Dunk	-18.6	-16.4	-6.6	-5.5	0.1	-12.5	**2.5**
Enduro	0.0	860.5	729.0	1,211.8	2,258.2	2,306.4	**3,454**
Fishing Derby	-91.7	-38.7	-4.9	15.5	**46.4**	41.3	8.9
Freeway	0.0	29.6	30.8	33.3	0.0	33.0	**33.9**
Frostbite	65.2	4,334.7	797.4	1,683.3	4,672.8	**7,413.0**	3,965
Gopher	257.6	2,412.5	8,777.4	14,840.8	15,718.4	**104,368.2**	33,641
Gravitar	173.0	**3,351.4**	473.0	412.0	588.0	238.0	440
H.E.R.O.	1,027.0	30,826.4	20,437.8	20,130.2	20,818.2	21,036.5	**38,874**
Ice Hockey	-11.2	**0.9**	-1.9	-2.7	0.5	-0.4	-3.5
James Bond	29.0	302.8	768.5	1,358.0	1,312.5	812.0	**1,909**
Kangaroo	52.0	3,035.0	7,259.0	12,992.0	**14,854.0**	1,792.0	12,853
Krull	1,598.0	2,665.5	8,422.3	7,920.5	**11,451.9**	10,374.4	9,735
Kung-Fu Master	258.5	22,736.3	26,059.0	29,710.0	34,294.0	**48,375.0**	48,192
Montezuma's Revenge	0.0	**4,753.3**	0.0	0.0	0.0	0.0	0.0
Ms. Pac-Man	307.3	**6,951.6**	3,085.6	2,711.4	6,283.5	3,327.3	3,415
Name This Game	2,292.3	8,049.0	8,207.8	10,616.0	11,971.1	**15,572.5**	12,542
Phoenix	761.4	7,242.6	8,485.2	12,252.5	23,092.2	**70,324.3**	17,490
Pitfall!	-229.4	**6,463.7**	-286.1	-29.9	0.0	0.0	0.0
Pong	-20.7	14.6	19.5	20.9	**21.0**	20.9	20.9
Private Eye	24.9	**69,571.3**	146.7	129.7	103.0	206.0	15,095
Q*Bert	163.9	13,455.0	13,117.3	15,088.5	19,220.3	18,760.3	**23,784**
River Raid	1,338.5	17,118.0	7,377.6	14,884.5	**21,162.6**	20,607.6	17,322
Road Runner	11.5	7,845.0	39,544.0	44,127.0	**69,524.0**	62,151.0	55,839
Robotank	2.2	11.9	63.9	65.1	**65.3**	27.5	52.3
Seaquest	68.4	42,054.7	5,860.6	16,452.7	50,254.2	931.6	**266,434**
Skiing	-17,098.1	**-4,336.9**	-13,062.3	-9,021.8	-8,857.4	-19,949.9	-13,901
Solaris	1,236.3	**12,326.7**	3,482.8	3,067.8	2,250.8	133.4	8,342
Space Invaders	148.0	1,668.7	1,692.3	2,525.5	6,427.3	**15,311.5**	5,747
Star Gunner	664.0	10,250.0	54,282.0	60,142.0	89,238.0	**125,117.0**	49,095
Surround	-10.0	6.5	-5.6	-2.9	4.4	1.2	**6.8**
Tennis	-23.8	-8.3	12.2	-22.8	5.1	0.0	**23.1**
Time Pilot	3,568.0	5,229.2	4,870.0	8,339.0	**11,666.0**	7,553.0	8,329
Tutankham	11.4	167.6	68.1	218.4	211.4	245.9	**280**
Up and Down	533.4	11,693.2	9,989.9	22,972.2	**44,939.6**	33,879.1	15,612
Venture	0.0	1,187.5	163.0	98.0	497.0	48.0	**1,520**
Video Pinball	16,256.9	17,667.9	196,760.4	309,941.9	98,209.5	479,197.0	**949,604**
Wizard Of Wor	563.5	4,756.5	2,704.0	7,492.0	7,855.0	**12,352.0**	9,300
Yars' Revenge	3,092.9	54,576.9	18,098.9	11,712.6	49,622.1	**69,618.1**	35,050
Zaxxon	32.5	9,173.3	5,363.0	10,163.0	12,944.0	**13,886.0**	10,513

Figure 14. Raw scores across all games, starting with 30 no-op actions. Reference values from Wang et al. (2016).

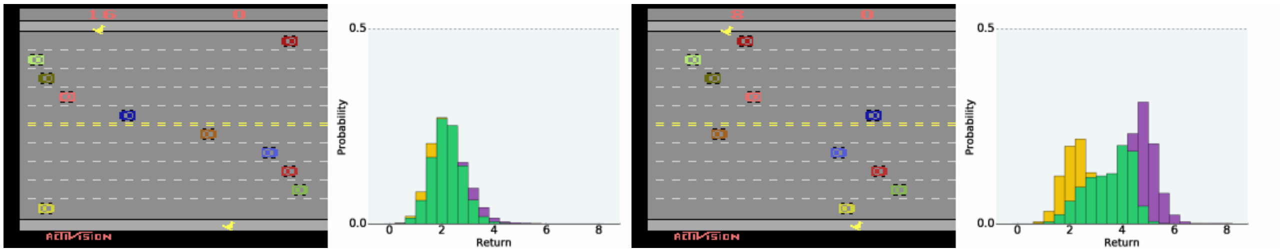


Figure 15. FREEWAY: Agent differentiates action-value distributions under pressure.

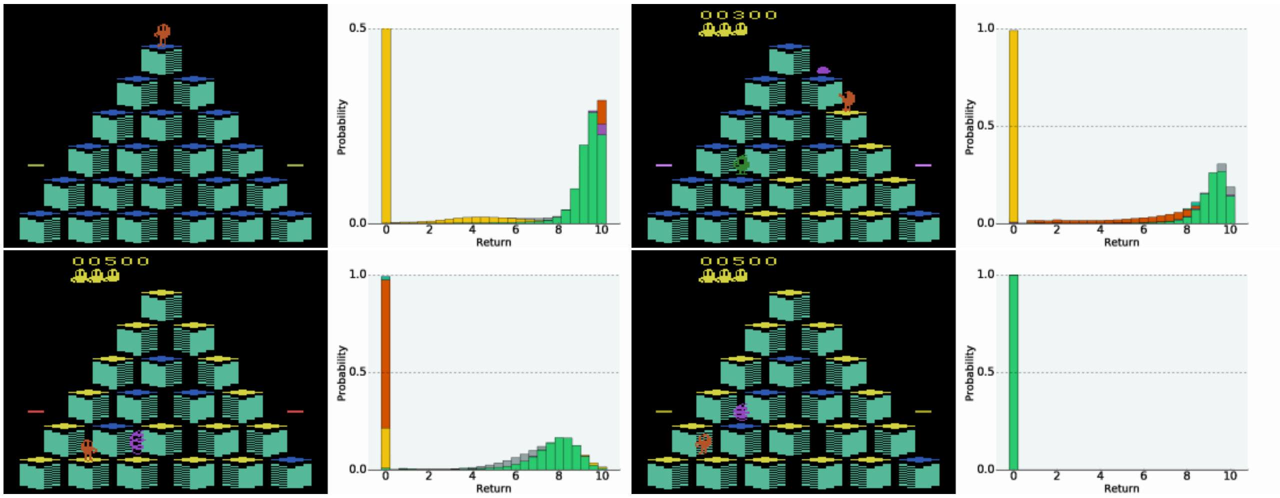


Figure 16. Q*BERT: Top, left and right: Predicting which actions are unrecoverably fatal. Bottom-Left: Value distribution shows steep consequences for wrong actions. Bottom-Right: The agent has made a huge mistake.

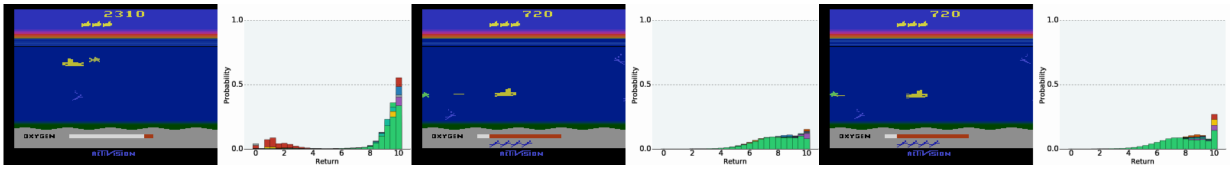


Figure 17. SEAQUEST: Left: Bimodal distribution. Middle: Might hit the fish. Right: Definitely going to hit the fish.

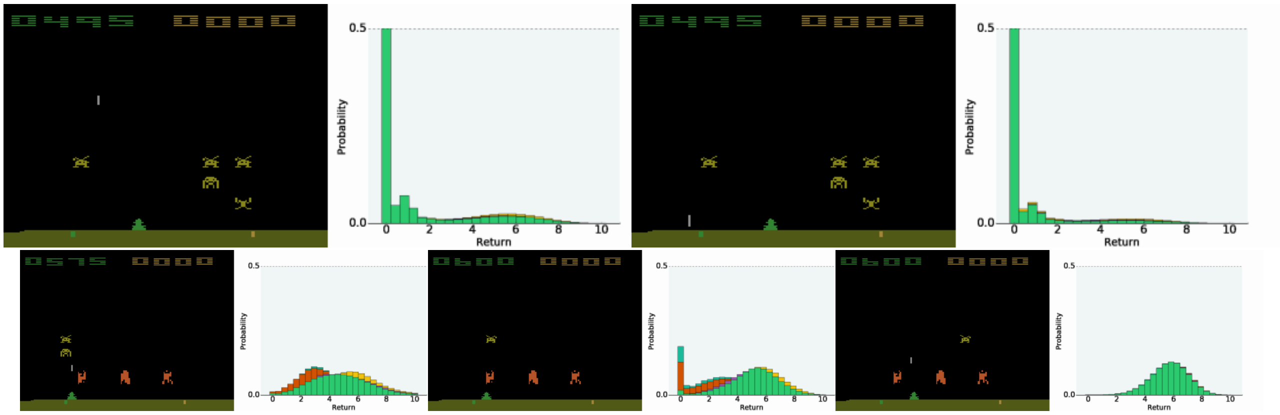


Figure 18. SPACE INVADERS: Top-Left: Multi-modal distribution with high uncertainty. Top-Right: Subsequent frame, a more certain demise. Bottom-Left: Clear difference between actions. Bottom-Middle: Uncertain survival. Bottom-Right: Certain success.