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 $A_1, A_2, \dots$ 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*

$$d_p(U,V) \le \sum\nolimits_i d_p(A_iU, A_iV).$$

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

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

Now, $|A_iU - A_iV|^p = 0$ 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

$$\begin{aligned} &d_p^p(A_iU, A_iV) = \\ &\qquad \inf_{Y_i, Z_i} \Pr\{A_i = 1\}\, \mathbb{E}\left[|Y_i - Z_i|^p \mid A_i = 1\right]. \end{aligned} \tag{8}$$

Next, we claim that

$$\begin{aligned} &\inf_{U,V} \sum\nolimits_i \Pr\{A_i = 1\}\, \mathbb{E}\Big[\big|A_iU - A_iV\big|^p \mid A_i = 1\Big] \qquad (9) \\ &\quad \le \inf_{\substack{Y_1, Y_2, \dots \\ Z_1, Z_2, \dots}} \sum\nolimits_i \Pr\{A_i = 1\}\, \mathbb{E}\Big[\big|Y_i - Z_i\big|^p \mid A_i = 1\Big]. \end{aligned}$$

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 $Y_1, Y_2, \dots$ and $Z_1, Z_2, \dots$ whose cumulative distributions are F_{A_iU}, F_{A_iV}, respectively. To prove this upper bound, consider the c.d.f. of U:

$$\begin{aligned} F_U(y) &= \Pr\{U \le y\} \\ &= \sum\nolimits_i \Pr\{A_i = 1\} \Pr\{U \le y \mid A_i = 1\} \\ &= \sum\nolimits_i \Pr\{A_i = 1\} \Pr\{A_iU \le y \mid A_i = 1\}. \end{aligned}$$

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*

$$\begin{aligned} \left\| \mathbb{E}\,\mathcal{T}^\pi Z_1 - \mathbb{E}\,\mathcal{T}^\pi Z_2 \right\|_\infty &\le \gamma \left\| \mathbb{E}\, Z_1 - \mathbb{E}\, Z_2 \right\|_\infty, \textit{ and} \\ \left\| \mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2) \right\|_\infty &\le \gamma^2 \left\| \mathbb{V} Z_1 - \mathbb{V} Z_2 \right\|_\infty. \end{aligned}$$

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

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

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

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

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

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

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 $\mathcal{X}_k$. $\square$

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. $\square$

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 $S_i^k(x) := \mathbb{I}\,[x \in \mathcal{X}_{k,i}]$ 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):

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

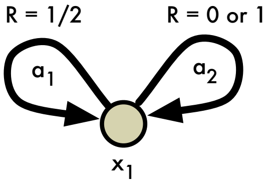

Figure 8. A simple example illustrating the effect of a nonstationary policy on the value distribution.

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 $A_i := \mathbb{I}[I = i]$. 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. □

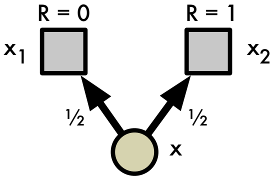

Figure 9. Example MDP in which the expected sample Wasserstein distance is greater than the Wasserstein distance.

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.

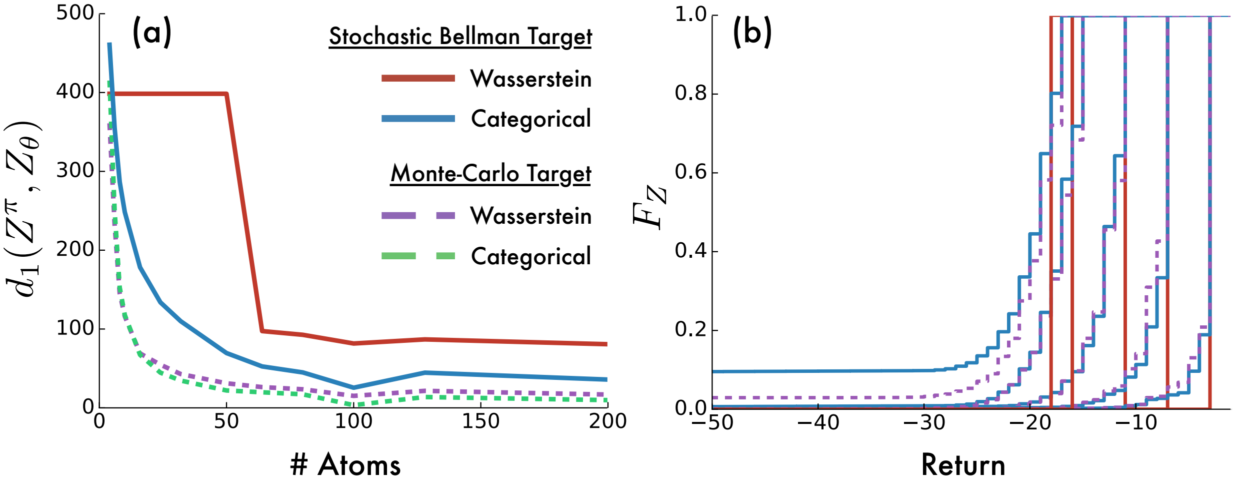

Figure 10. (a) Wasserstein distance between ground truth distribution Z^π and approximating distributions Z_θ. Varying number of atoms in approximation, training target, and loss function. (b) Approximate cumulative distributions for five representative states in CliffWalk.

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.

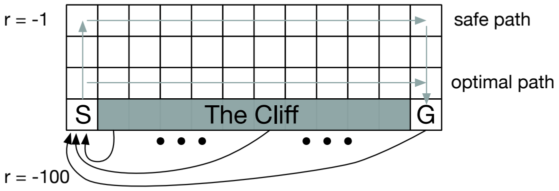

Figure 11. CliffWalk Environment (Sutton & Barto, 1998).

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.

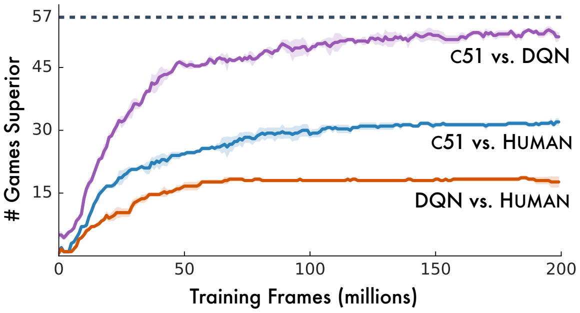

Figure 12. Number of Atari games where an agent's training performance is greater than a baseline (fully trained DQN & human). Error bands give standard deviations, and averages are over number of games.

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	**3,692,500**
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).

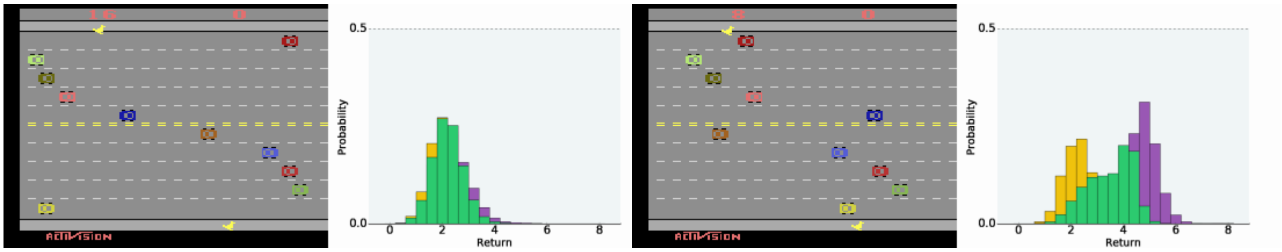

Figure 15. FREEWAY: Agent differentiates action-value distributions under pressure.

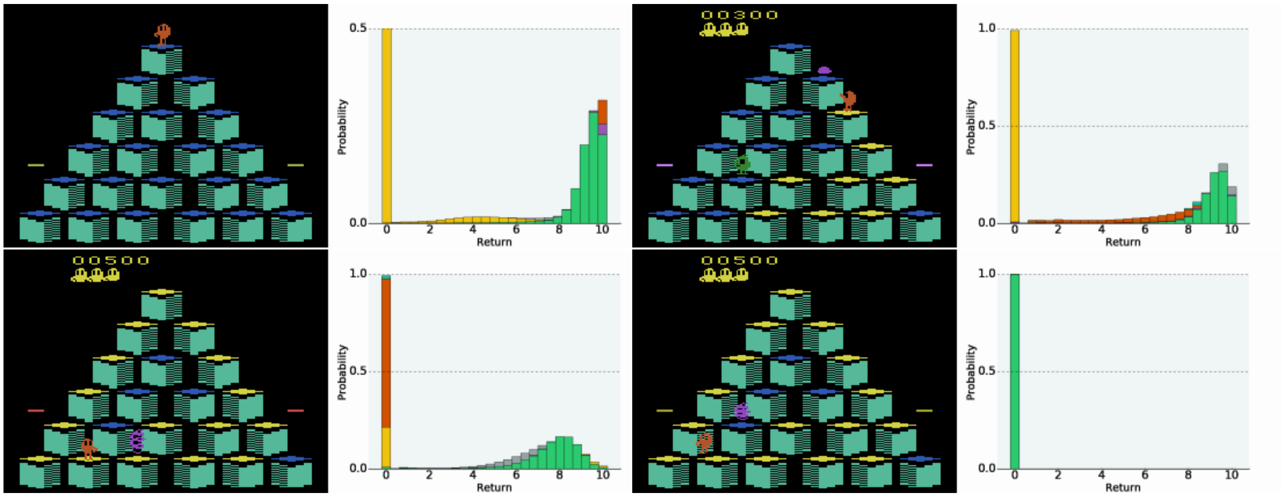

Figure 16. Q*BERT: Top, left and right: Predicting which actions are unrecoverably fatal. Bottom-Left: Value distribution shows steep consequences for wrong actions. Bottom-Right: The agent has made a huge mistake.

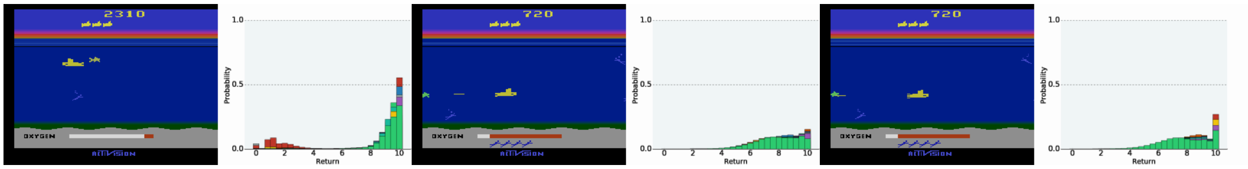

Figure 17. SEAQUEST: Left: Bimodal distribution. Middle: Might hit the fish. Right: Definitely going to hit the fish.

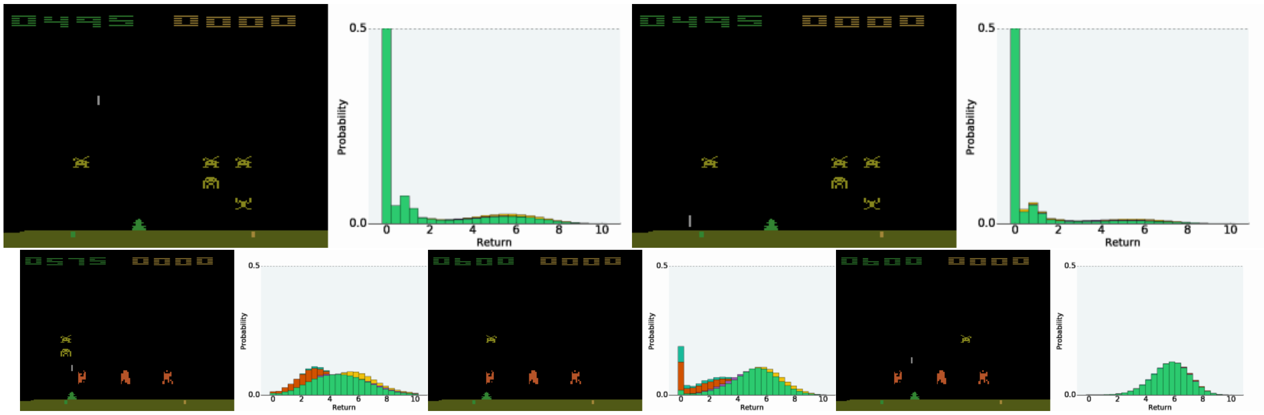

Figure 18. SPACE INVADERS: Top-Left: Multi-modal distribution with high uncertainty. Top-Right: Subsequent frame, a more certain demise. Bottom-Left: Clear difference between actions. Bottom-Middle: Uncertain survival. Bottom-Right: Certain success.