A Distributional Perspective on Reinforcement Learning

Marc G. Bellemare, Will Dabney, Remi Munos (DeepMind)

ICML 2017

Presented by Sahana Ramnath, EE15B109

October 11, 2018

IIT Madras

Table of contents

Introduction

The main idea

The common approach to reinforcement learning is to maximize the **expected** value of the random return received by the agent : the value $Q(s, a)$

$$\underline{\text{Bellman Equation}} : \; Q(s, a) = \mathbb{E}[R(s, a)] + \gamma \, \mathbb{E}[Q(s', a')]$$

The main idea

The common approach to reinforcement learning is to maximize the **expected** value of the random return received by the agent : the value $Q(s, a)$

$$\underline{\text{Bellman Equation}} : \; Q(s, a) = \mathbb{E}[R(s, a)] + \gamma \, \mathbb{E}[Q(s', a')]$$

This paper argues for the fundamental importance of the value **distribution**.

$$\underline{\text{Distributional Bellman Equation}} : \; Z(s, a) \overset{D}{=} R(s, a) + \gamma Z(s', a')$$

where Z is the random return whose expectation is Q.

The main idea

The common approach to reinforcement learning is to maximize the **expected** value of the random return received by the agent : the value $Q(s, a)$

$$\underline{\text{Bellman Equation}} : \; Q(s,a) = \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}[Q(s',a')]$$

This paper argues for the fundamental importance of the value **distribution**.

$$\underline{\text{Distributional Bellman Equation}} : \; Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(s',a')$$

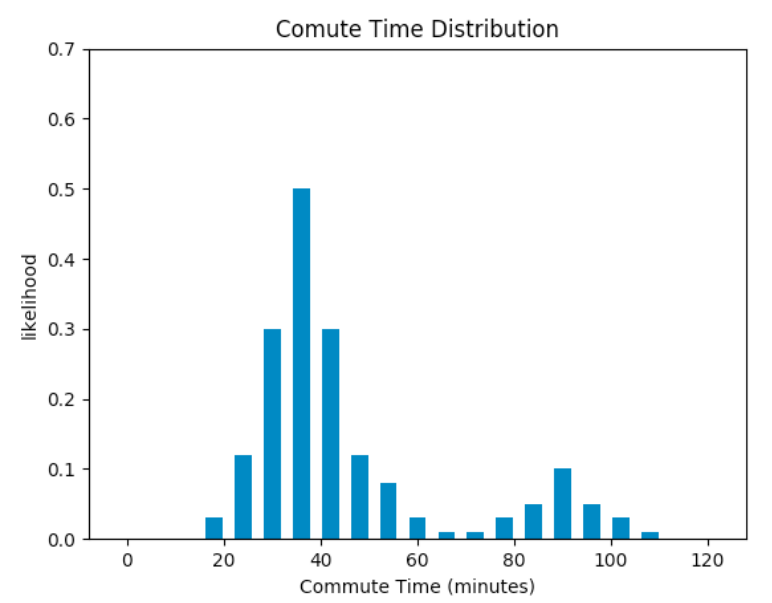
where Z is the random return whose expectation is Q.

The paper claims that modelling the entire value distribution makes the learning much better behaved.

Benefits

PRESERVING THE **MULTIMODALITY** OF THE VALUE DISTRIBUTION

For example, consider a person who takes the train to work, every morning.We want to minimize the total commute time.
Under normal conditions, the train runs on time and the commute time is say, 30 minutes on average.
But once in a while, there are unexpected problems, like mechanical troubles or rain which lead to a much higher commute time.



Given the full knowledge of the distribution, the person can now take more informed decisions.

Benefits

HELPS TO MODEL **RISK AVERSENESS** OF THE AGENT

Say 2 actions have the same expected reward, but widely varying standard deviations.

Or say one action gives a lower expected reward but with a very low variance, and another action gives a higher expected reward but with a high variance.

EASIER TO MODEL **NON-STATIONARY** POLICIES

Has been proven that though the distributional Bellman optimality operator is a contraction in the expected value, it is not a contraction in any metric over distributions.

Short Note on Related Work

Uses so far and now

The distributional perspective of RL is pretty old, but so far, it has been used for specific purposes :

- To model parametric uncertainty
- To design risk sensitive algorithms
- General theoritic analysis

This paper on the other hand, uses the value distribution directly to represent the policy and make decisions in the environment.

Uses so far and now

MODELLING UNCERTAINTY

- Bayesian Q-Learning(Dearden et al.) considers a prior over the distribution of the return $R_{s,a}$ and then updates these priors based on agent's experience.

RISK SENSITIVE ALGORITHMS

- Learning the Variance of the Reward to-go(Tamar et al.) uses linear function approximation to learn the variance of the return for policy evaluation.

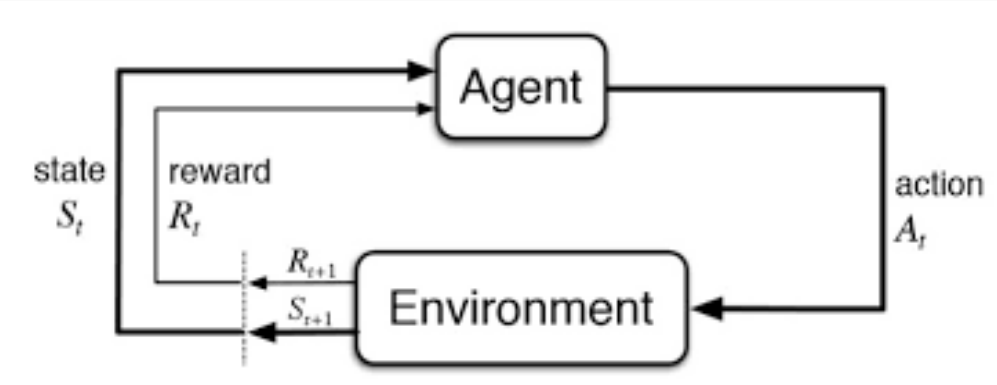
GENERAL THEORITIC ANALYSIS

- MDPs with a new optimality criterion: Discrete time(Jacquet et al.) Showed that a moment optimality criterion which imposes an ordering on distributions is achievable.
- Discounted MDPs: Distribution functions and exponential utility maximization(Chung Sobel) showed that the operator is not a contraction in total variation distance.

The Bellman and Distributional Bellman Operators

Setting and Notation

Agent interacts with the environment in the standard fashion



This is modelled as an MDP (S, A, R, P, γ) where R is now explicitly treated as a random variable.

Bellman and Distributional Bellman

The return Z^π is the sum of discounted rewards along the agent's trajectory in the environment. The value function Q^π is the expected return along the trajectory($\mathbb{E}[Z^\pi(s,a)]$).

Bellman Operator

$$\mathcal{T}^B Q^\pi(s,a) := \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}_{P,\pi}[Q^\pi(s',a')]$$

Distributional Bellman Operator

$$\mathcal{T}^\pi Z(s,a) :\overset{D}{=} R(s,a) + \gamma P^\pi Z(s,a)$$
$$\mathcal{T}^\pi Z(s,a) :\overset{D}{=} R(s,a) + \gamma Z(s',a')$$

While $\mathcal{T}$ is simililar to $\mathcal{T}^B$ on the surface, it is fundamentally different. It is defined by three sources of randomness.

- The randomness of R
- The randomness in the transition from (s,a) to (s',a')
- The next-state value distribution $Z(s',a')$

Contraction of the Operator

Contraction of the Operator

The Bellman operator has been proven to be a contraction in the Banach space.

In this paper, they have proved that, for a fixed policy in the environment, the distributional Bellman operator is a contraction over the maximal form of the **Wasserstein Metric**(explained in the next slide).

The choice of metric matters, since this operator is not a contraction in other distribution metrics such as KL divergence or Kolmogorov distance or total variation distance.

The Wasserstein Metric

The Wasserstein or Mallows metric d_p is between cumulative distribution functions.

For F, G, two c.d.fs over reals,

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

where infimum is taken over all pairs of random variables (U, V) with c.d.fs as F and G.

For $p < \infty$, this is calculated as

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

The Wasserstein Metric

The metric d_p has the following properties

$$d_p(aU, aV) \leq |a| d_p(U, V)$$
$$d_p(A + U, A + V) \leq d_p(U, V)$$
$$d_p(AU, AV) \leq ||A||_p d_p(U, V)$$

Convergence in Policy Evaluation

Let Z denote the space of value distributions with bounded moments.

For $Z_1, Z_2 \in Z$, the maximal form of the Wasserstein metric is :

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

$\bar{d}_p$ is used to establish the convergence of the distributional Bellman operator.

$\mathcal{T}^\pi : Z \to Z$ has been proven to be a γ-contraction in $\bar{d}_p$ with the unique fixed point Z^π.

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

Convergence in Policy Evaluation

$$
\begin{aligned}
& d_p(\mathcal{T}^\pi Z_1(s,a), \mathcal{T}^\pi Z_2(s,a)) \\
& \quad = \mathrm{d}_p(R(s,a) + \gamma P^\pi Z_1(s,a), R(s,a) + \gamma P^\pi Z_2(s,a)) \\
& \quad \leq d_p(\gamma P^\pi Z_1(s,a), \gamma P^\pi Z_2(s,a)) \\
& \quad \leq \gamma d_p(P^\pi Z_1(s,a), P^\pi Z_2(s,a)) \\
& \quad \leq \gamma sup_{s',a'} d_p(Z_1(s',a'), Z_2(s',a'))(1)
\end{aligned}
$$

Hence,

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

(2)

Convergence in Control

While all optimal policies attain the same Q^*, in general, there are many optimal value distributions.

The paper proves that :

- The distributional Bellman optimality operator converges in a weak sense, to the set of optimal value distributions.
- This operator however, is **not** a contraction in any metric for distributions.

Definitions

Optimal Value Distribution

Let Π^* be the set of optimal policies. An optimal value distribution is the value distribution of an optimal policy.

The set of optimal value distributions is $Z^* := \{Z^{\pi^*} : \pi^* \in \Pi^*\}$

Not all value distributions with expectation Q^* are optimal. They must match the full distribution of the return under an optimal policy

Set of Greedy Policies

A greedy policy π for Z maximises the expectation of Z. The set of greedy policies for Z is

$$G_Z := \{\pi : \sum_a \pi(a|s)\, \mathbb{E}[Z(s,a)] = max_{a' \in A}\, \mathbb{E}[Z(s,a')]\}$$

Convergence in Control

Distributional Bellman Optimality Operator

$$\mathcal{T}Q(s,a) := \mathbb{E}[R(s,a)] + \gamma\, \mathbb{E}_P[max_{a' \in A} Q^*(s',a')]$$

The maximization at s' corresponds to some greedy policy.

Any operator which implements a greedy selection rule is a distributional Bellman optimality operator.

$$\mathcal{T}Z = \mathcal{T}^\pi Z \text{ for some } \pi \in G_Z$$

Convergence of $\mathbb{E}[Z]$

Consider the iterates $Z_{k+1} := \mathcal{T}Z_k, Z_0 \in Z$

Let $Z_1, Z_2 \in Z$. Then, $||\,\mathbb{E}[\mathcal{T}Z_1] - \mathbb{E}[\mathcal{T}Z_2]||_\infty \leq \gamma ||\,\mathbb{E}[Z_1] - \mathbb{E}[Z_2]||_\infty$,

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

Convergence in Control

Nonstationary Optimal Value Distributions

A nonstationary optimal value distribution Z^{**} is the value distribution corresponding to a sequence of optimal policies. The set of nonstationary optimal value distributions is Z^{**}. If the state space is finite, then Z_k converges to Z^{**} uniformly.

Also, if there is a total ordering $\prec$ on Π^* such that for any Z^*,

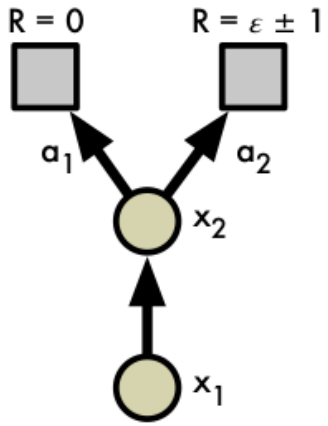
$$\mathcal{T}Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in G_{Z^*}, \pi \prec \pi', \forall \pi' \in G_{Z^*} / \{\pi\}$$

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

While the mean of Z_k converges exponentially quickly to Q^*, its distribution need not be well behaved.

- The operator $\mathcal{T}$ is not a contraction.
- Not all optimality operators have a fixed point.
- The presence of a fixed point is insufficient to guarantee convergence.

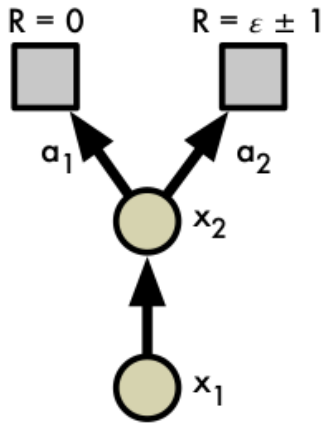
Example



	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	$\epsilon \pm 1$

There is a unique optimal policy and so, a unique fixed point Z^*.

Example



	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	$\epsilon \pm 1$

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

But, when we apply $\mathcal{T}$ to Z, the greedy action a_1 is selected and $\mathcal{T}Z(x_1) = Z(x_2, a_1)$.

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

$$> 2\epsilon$$

Approximate Distributional Learning

Parametric Distribution

In their experiments ,the value distribution is modelled using a discrete distribution parametrized by $N \in \mathbb{N}$ and $V_{MIN}, V_{MAX} \in \mathbb{R}$ and whose support is the set of atoms $\{z_i = V_{MIN} + i\Delta z : 0 \leq i < N\}$, $\Delta z := \frac{V_{MAX} - V_{MIN}}{N-1}$.

The atom probabilities are given by a parametric model $\theta : SxA \rightarrow \mathbb{R}^N$

$$Z_\theta(s,a) = z_i \text{ with probability } p_i(s,a) := \frac{e^{\theta_i(s,a)}}{\sum_j e^{\theta_j(s,a)}}$$

Projected Bellman Updates

- The Bellman update $\mathcal{T}Z_\theta$ and the original parametrization Z_θ almost always have disjoint supports.
- This can be still be used to minimize the Wasserstein metric(viewed as a loss between $\mathcal{T}Z_\theta$ and Z_θ)
- Since learning is restricted to sample transitions, the sample Bellman update $\hat{\mathcal{T}}Z_\theta$ is projected onto the support of Z_θ(similar to multiclass classification).

The Categorical C51 Algorithm

This is the main algorithm proposed by the paper to perform **iterative approximation** of the value distribution Z using the distributional Bellman equation.

The number **51** represents the number of discrete values used to parametrize the value distribution (found using trial).

Basically, at each time step, a transition is sampled from the environment is used to compute the target distribution $R + \gamma Z'$. This is used to update the current distribution Z by minimizing the **cross-entropy** loss between the two.

C51 Pseudocode

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

Experimental Results

Experiments

Key points of the experiments conducted on this model :

- This algorithm was trained and tested on Atari 2600 games.
- A DQN was used to compute the $p_i(s, a)$ atom probabilities.
- Called a **Categorical DQN**
- The squared error loss is replaced by the cross-entropy loss.
- $\epsilon-$greedy policy was used.

Results

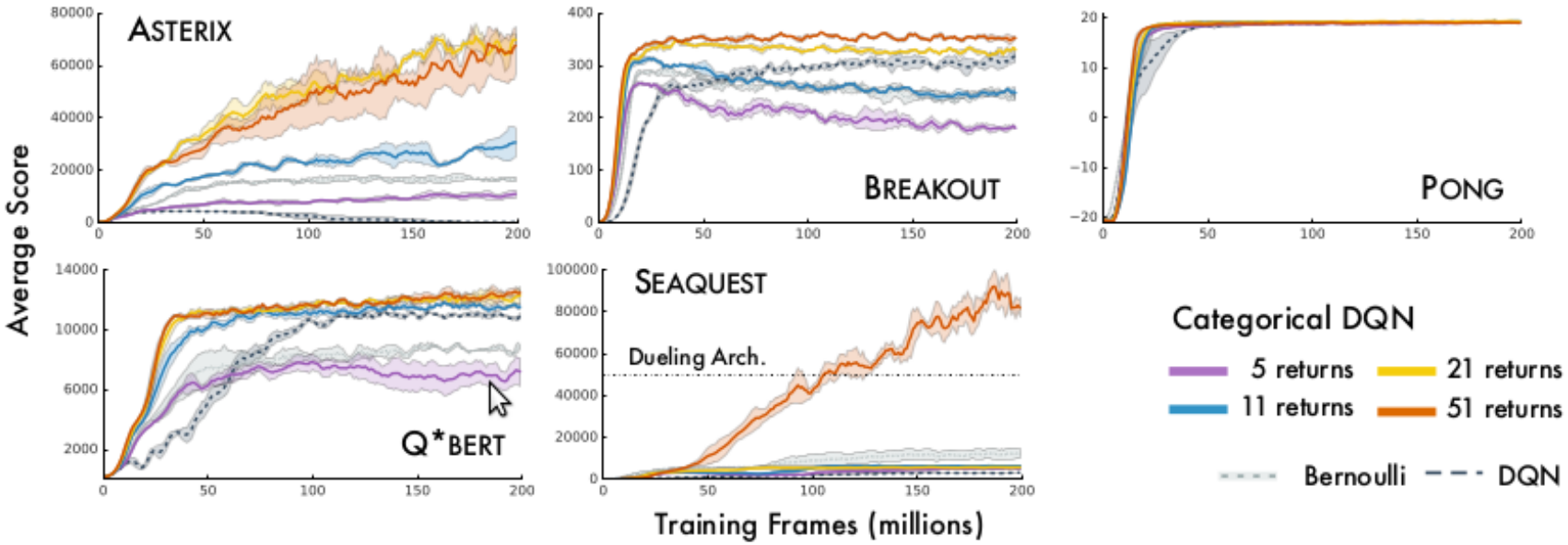
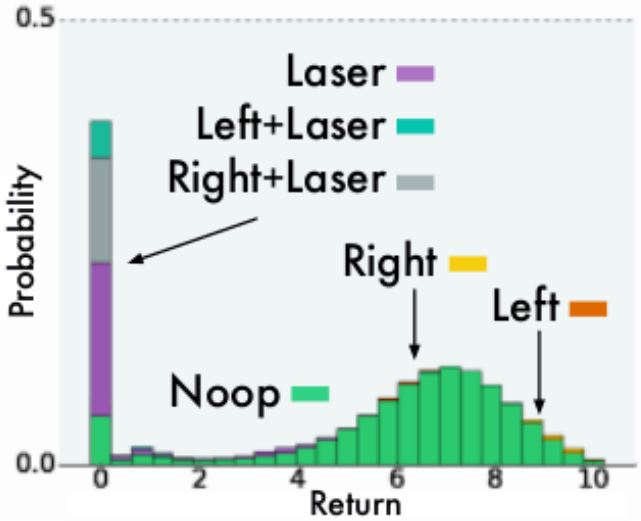


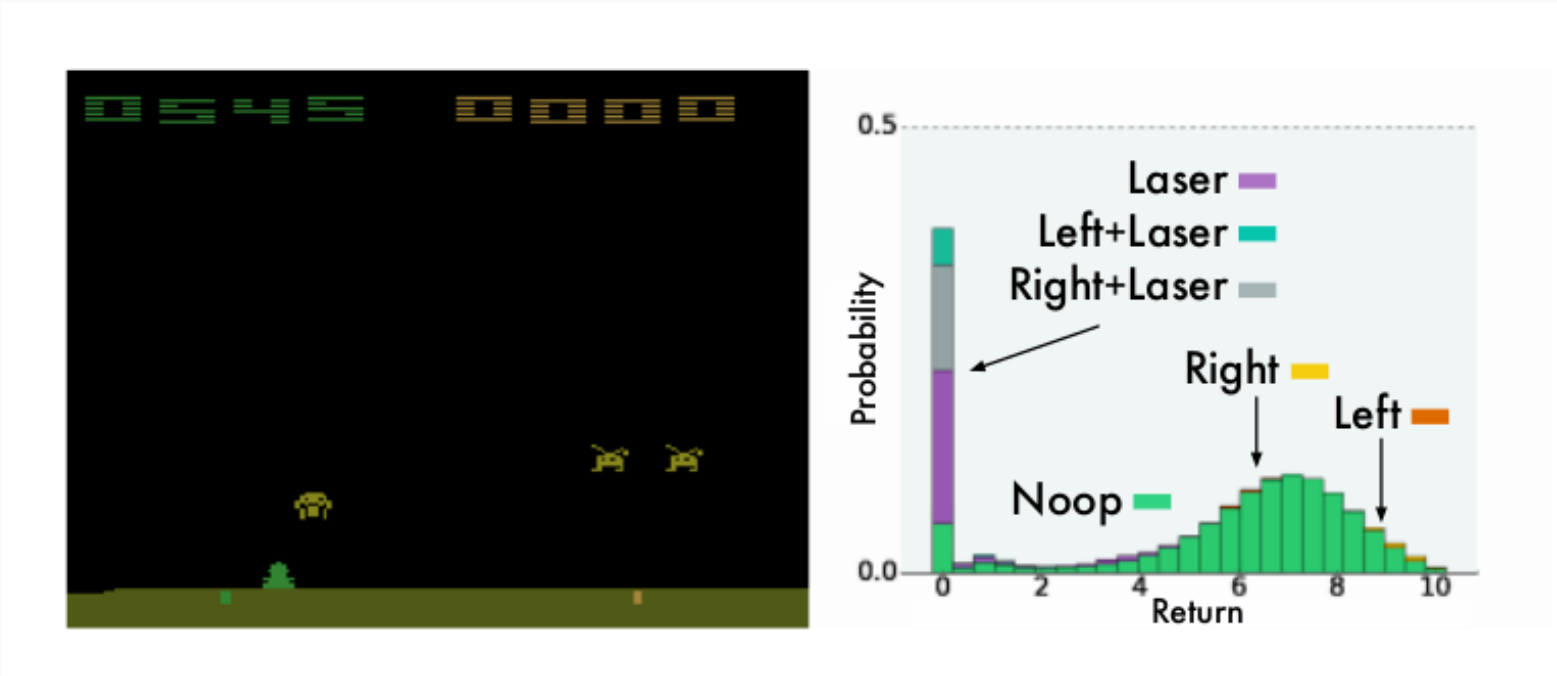
Figure 3. Categorical DQN: Varying number of atoms in the discrete distribution. Scores are moving averages over 5 million frames.

Results : Space Invaders

Learned value distribution over an episode of Space Invaders.

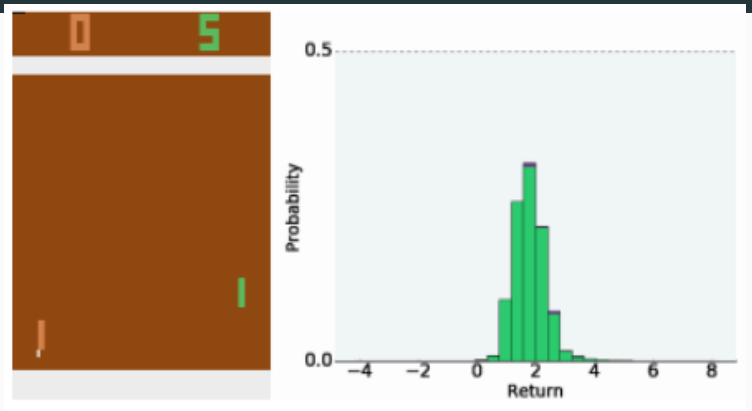


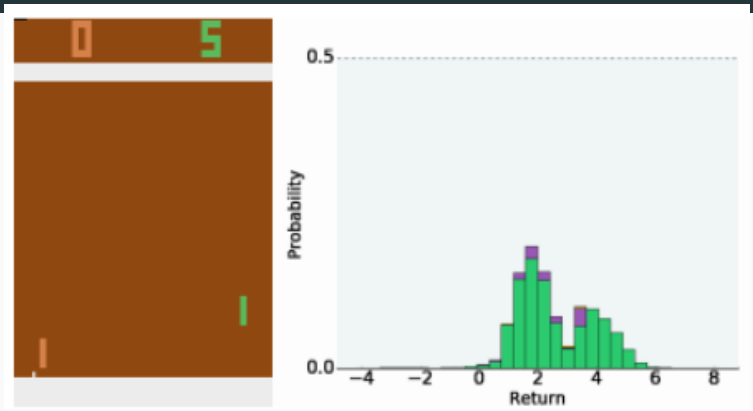
Results : Space Invaders

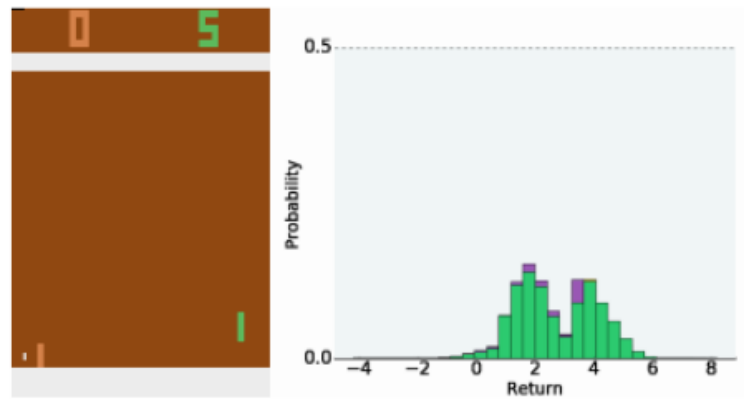


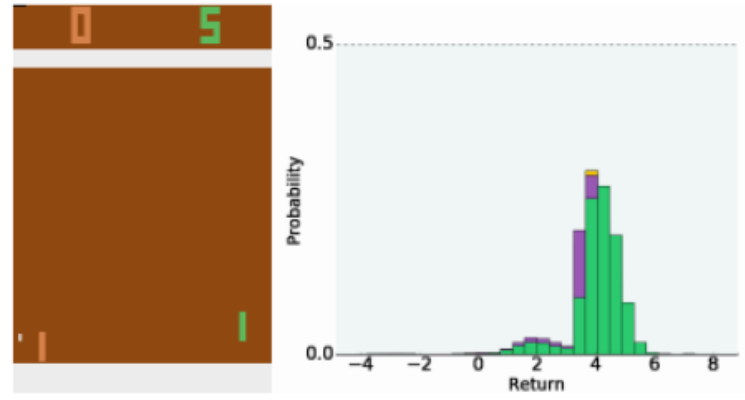
The distribution keeps the low-value 'losing' event separated from the high-value 'survival' event, rather than averaging them into one unrealizable expectation.

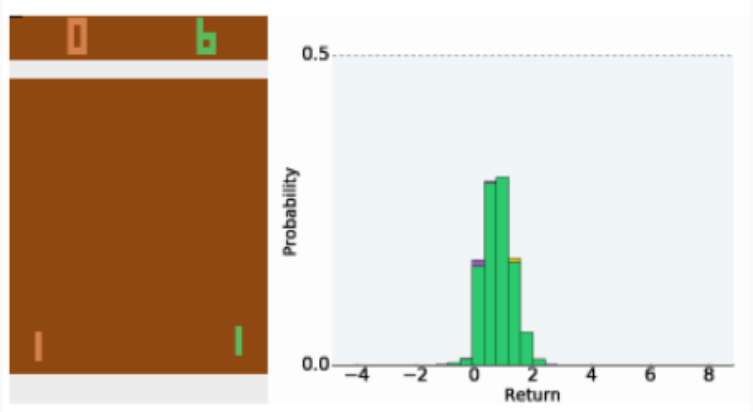
Results : Pong



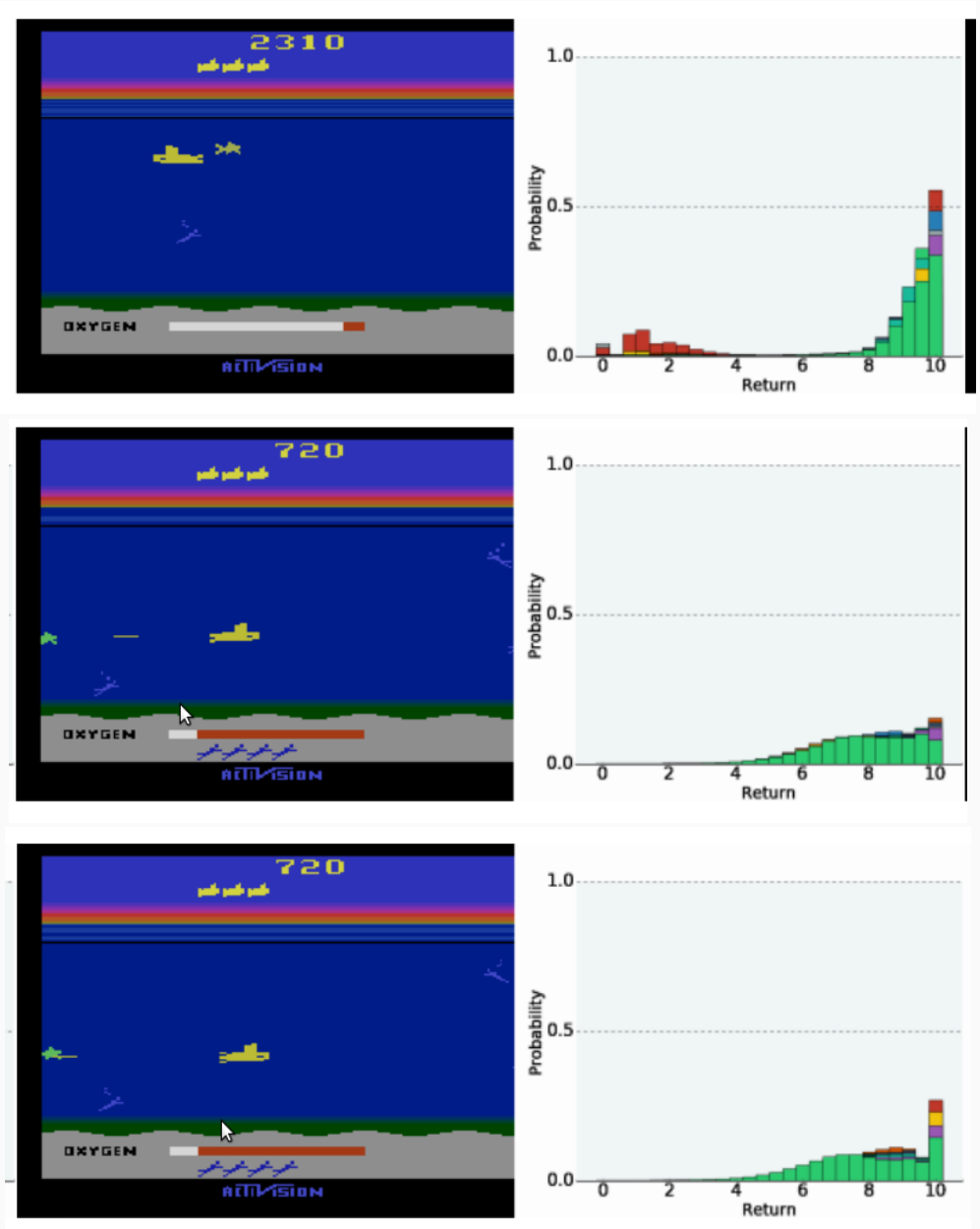








Results : Seaquest



Intuition to learn value distributions

The policy being used is still trying to maximize expected reward. But the difference here is the learning of distributions to model uncertainty and approximation.

- **Reduced chattering :** By averaging the different distributions similar to conservative policy iteration, the oscillations/instability of convergence of the Bellman optimality operator is prevented.
- **State Aliasing :** By explicitly modelling the resulting distribution, a more stable learning target is provided.
- **A richer set of predictions :** Rich set of auxillary predictions, so better learning.
- **Framework for inductive bias :** Can treat all returns greater than V_{MAX} as equivalent.
- **Well-behaved optimization :** Closer minimization of the Wasserstein metric would yield even better results.

THANK YOU