A Distributional Perspective on Reinforcement Learning

- By Marc G.Bellemare, Will Dabney & Remi Munos (2017)

Distributional reinforcement learning

- A distributional perspective on reinforcement learning (2017)
- Distributional reinforcement learning with quantile regression (2017)
- An analysis of categorical distributional reinforcement learning (2018)
- Implicit quantile networks for distributional reinforcement learning (2018)

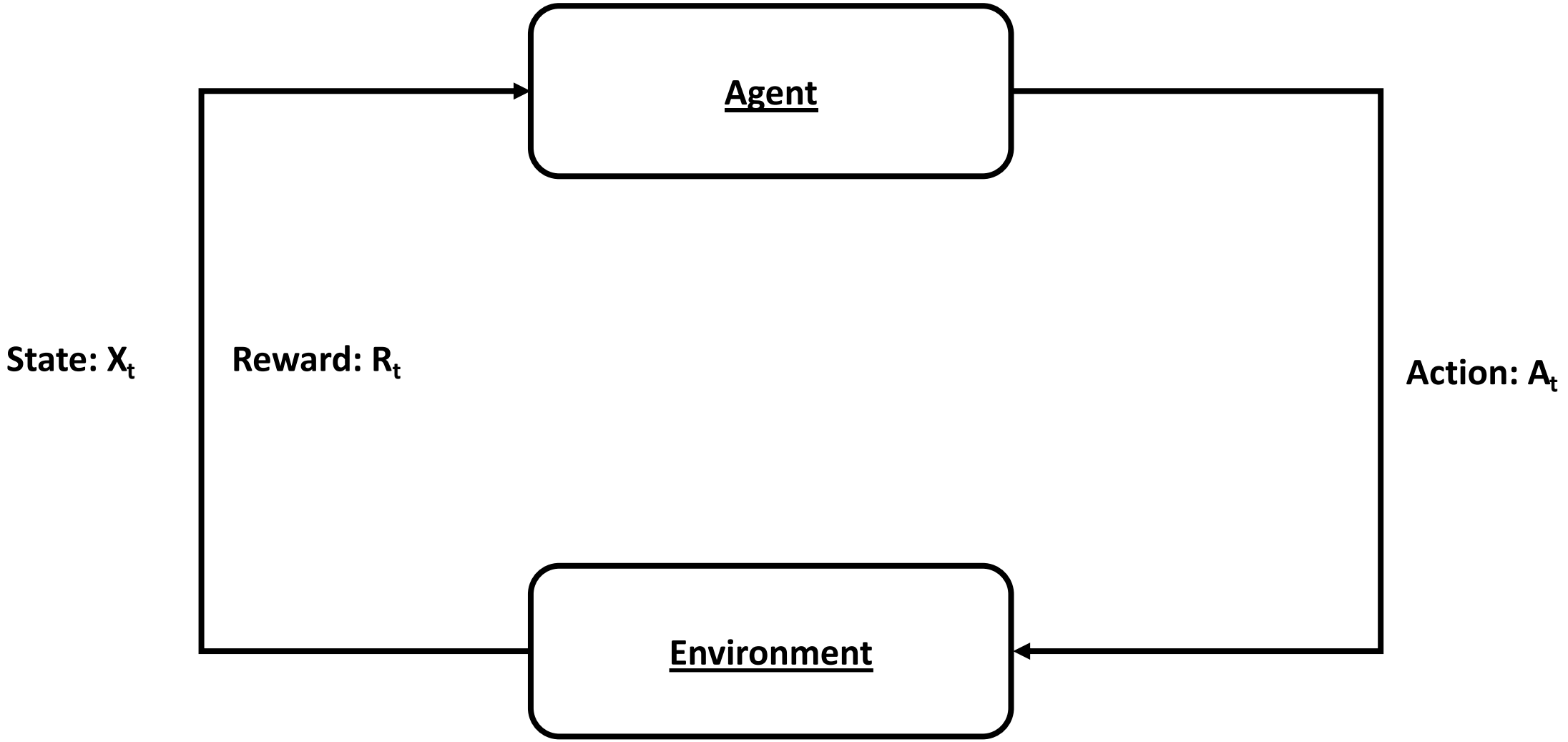
One sentence summary

- Replace q-value with a distribution

Content

- Reinforcement Learning: A very brief review
- Motivation on distributional approach
- Convergence proof
- Implementation: Algorithm – C51
- Examples

RL: Brief Review



RL: Brief Review

Settings Parameters	
X	$state\ space$
A	$action\ space$
R	$reward\ function$
P	$state\ transition$
γ	$discount\ rate$
π	$policy$
$Q(x,a)$	$state-action\ value$
$Z^{\pi}(x)$	Returns

$$Bellman\ Equation:$$
$$Q^{\pi}(x,a) = E\ R(x,a) + \gamma E_{P,\pi} Q^{\pi}(x',a')$$
$$V^{\pi}(x) = E[Z^{\pi}(x)] = ER(x) + \gamma E[Z^{\pi}(X')]$$
$$Optimality\ Equation:$$
$$Q^{*}(x,a) = E\ R(x,a) + \gamma E_{P} \max_{\{a' \in A\}} Q(x',a')$$

Two popular algorithms: SARSA and Q-learning

$$Bellman\ Operator:$$
$$\tau^{\pi} Q(x,a) \coloneqq E\ R(x,a) + \gamma E_{P,\pi} Q^{\pi}(x',a')$$
$$Optimality\ Operator:$$
$$\tau Q(x,a) \coloneqq E\ R(x,a) + \gamma E_{P} \max_{\{a' \in A\}} Q(x',a')$$

They are both contraction mapping =>
$$Q_0\ converges\ to\ Q^{\pi}\ and\ Q^{*}, respectively$$

Why Distributional RL?

- The puzzle when modelling using expectation approach
  - An example: $\mathrm{R(x)} = \begin{cases} +1\ with\ 60\% \\ -1\ with\ 40\% \end{cases} => E(R) = +0.2$
  - The problem is… you never actually get the expectation value (+0.2)
  - Modelling the expected return hides this intrinsic randomness
  - The reality is not actually considered in the model
- Random returns are often complex, multimodal (multiple hypes)
- Why mean? Why not the entire distribution
  - Bayesian Regression framework: Why MAP? Why not the full posterior?

High level understanding

$$Bellman\ Equation:$$
$$Q^{\pi}(x,a) = E\ R(x,a) + \gamma E_{P,\pi} Q^{\pi}(x',a')$$
$$V^{\pi}(x) = E[Z^{\pi}(x)] = ER(x) + \gamma E[Z^{\pi}(X')]$$

Let's remove the expectation from the Bellman Equation

$$Distributional\ Bellman\ Equation:$$
$$Q^{\pi}(x,a) = R(x,a) + \gamma Q^{\pi}(x',a')$$
$$V^{\pi}(x) = Z^{\pi}(x) = R(x) + \gamma [Z^{\pi}(X')]$$

High level understanding

- We have:
  - $V^{\pi}(x) = Z^{\pi}(x) = R(x) + \gamma[Z^{\pi}(X')]$
  - R.V. = R.V. + R.V
  - So what does it mean?
  - Recursive Distributional Equations (Rosler, 1992)
    - Tells us how can we describe the relationship between R.V. on the LHS and R.V. on the RHS of the equation.
    - $X = \frac{1}{\sqrt{2}} X_1 + \frac{1}{\sqrt{2}} X_2$
    - We want: $X_1 \sim X_2 \sim X$, R.V. with same distribution to satisfy the equation
    - $X \sim N(0, 1)$ is the solution
    - Similar idea: Conjugate prior

Fixed point of the distributional Bellman Eq.

- $Z^{\pi}(x) = R(x) + \gamma Z^{\pi}(x') \quad X' \sim P^{\pi}(.|X)$
- Z^{π} is distributed like a reward distribution plus a discounted return distribution in the next state
- Question: is there a fixed point? (Convergence proof)

Fixed point of the distributional Bellman Eq.

- We know that:
  - Bellman (1957): Bellman equation to describe the relationship between one mean and the other mean.
  - Sobel (1982): Bellman equation for variance
  - Engel (2003): Bellman equation for Bayesian Uncertainty
  - Azar et al. (2011), Lattiemore & Hutter (2012):
    - Bellman equation for higher moments

- The difference in this paper is that the authors intend to quantify the distribution all at once
  - NOT individual moments.

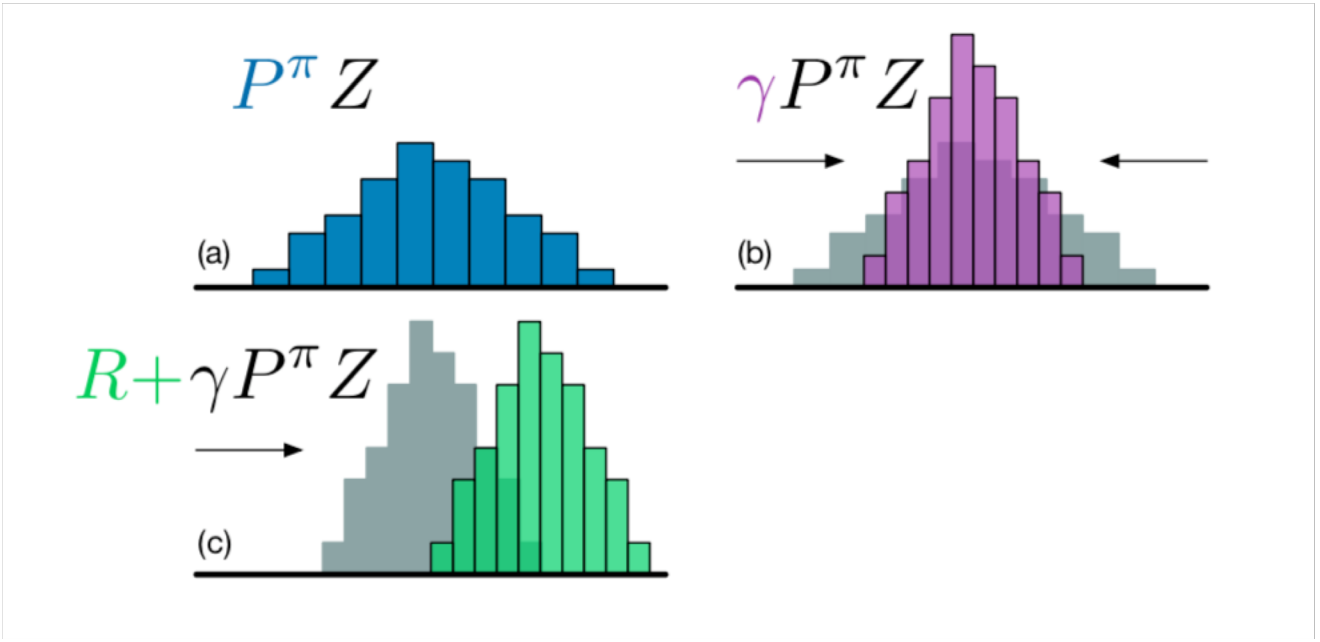
Bellman Operator

- Operator: a way of describing expected behaviour of the learning algorithm.
  - In RL, agents are going to sample the states of the environment, they are going to wonder around and try to learn about this world
  - One way to describe average behaviour is by looking at the operator
  - Explains what is going to happen to their current prediction if the agents do a one-step update.

Bellman Operator

- Bellman Operator
  - $\tau^{\pi} Z(x,a) \coloneqq R(x,a) + \gamma P^{\pi} Z(x,a)$
  - We start with some prediction about the world Z, which is a distribution of the return of every state and action.
  - $P^{\pi} Z(x,a)$
    - Next state action pair return distribution
  - Apply discount factor
  - Add the reward → shift the distribution (Convolutions)

Bellman Operator



Policy Evaluation Setting

- We fix the policy, and we want to look at what happens as the learning process takes place
- Effectively we want to prove the convergence given a policy π

Convergence of a RL algorithm

- Conventional RL convergence relies on contraction
  - $\left|\left|\tau^{\pi}Q_1 - \tau^{\pi}Q_2\right|\right|_{\infty} \leq \gamma \left|\left|Q_1 - Q_2\right|\right|_{\infty}$
  - γ :[0,1) Lipschitz constant
  - The max difference between two value functions Q_1 and $Q_2$is smaller after we apply the Bellman Operator
- $\lim_{k\to\infty} (\tau^{\pi})^{k} Q(x,a) = Q^{\pi}(x,a)$
- One will find the Q value of a particular policy π

Convergence of a RL algorithm

- How about distributional RL?
  - Not so clear, because now we have the entire distribution... Not only an expectation
- However, the authors employ the similar idea
  - We want to start to measure the difference between two distribution
    - Defined as $d(p, q)$
  - Possible solutions
    - KL divergence
    - Total variations
    - But there are problem with those two approaches when two distributions have no joint part

The Wasserstein Metric

Definition: The Wasserstein Metric

For any random variable U, V, and corresponding distributions F, G, the Wasserstein Distance 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 corresponding CDF F and G.

Wasserstein is actually a metric to calculate the distance between two distributions F and G, just like KL divergence [a]. However, it is better than KL divergence when two distributions have no overlaps.

[a] relative entropy, information divergence, information gain

Maximal Wasserstein Metric

Define Maximal Wasserstein Metric as the largest Wasserstein distance between any two value distributions at any state-action

$$\widehat{w}_p(Z_1, Z_2) \coloneqq \sup_{x,a} w_p(Z_1(x,a), Z_2(x,a))$$

We can prove that

$$\begin{aligned}\widehat{w}_p(\tau^\pi Z_1, \tau^\pi Z_2) &= \widehat{w_p}(R + \gamma P^\pi Z_1, R + \gamma P^\pi Z_2) \\ &\leq \widehat{w}_p(\gamma P^\pi Z_1, \gamma P^\pi Z_2) \\ &\leq \gamma\, \widehat{w}_p\,(P^\pi Z_1, P^\pi Z_2) \\ &\leq \gamma \widehat{w}_p(Z_1, Z_2)\end{aligned}$$

The conclusion here is that we can prove that after applying the Bellman Operator, the Wasserstein distance of two value distributions is at least γ smaller.

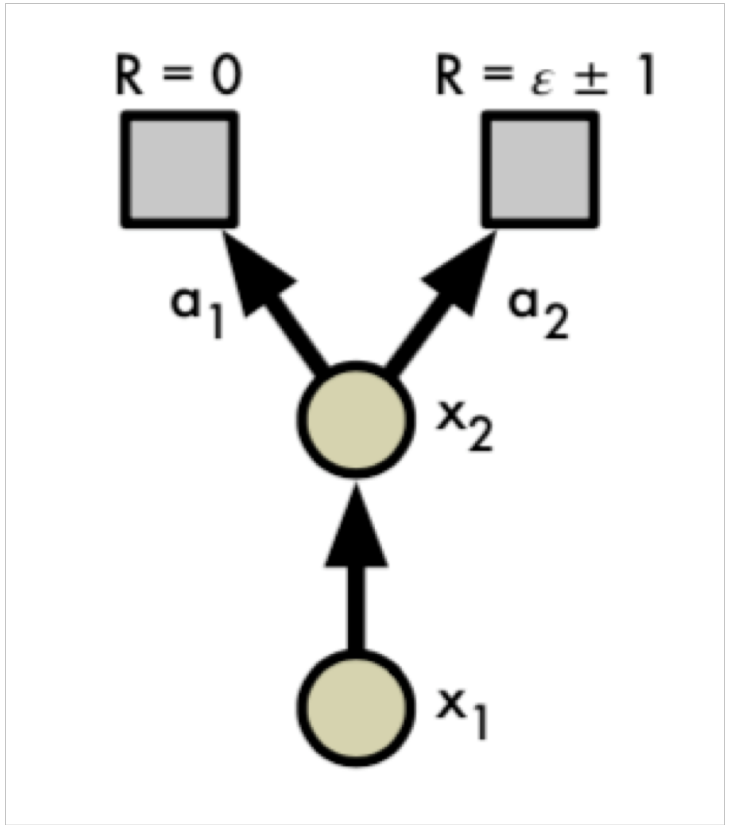
Control setting

- We know that given a policy, a distributional RL converges to a distribution for each of the state-action pairs.
- What the about the control case?
  - We want to learn a bit about the optimal policy
- Again we first look at the conventional RL

Optimality Operator

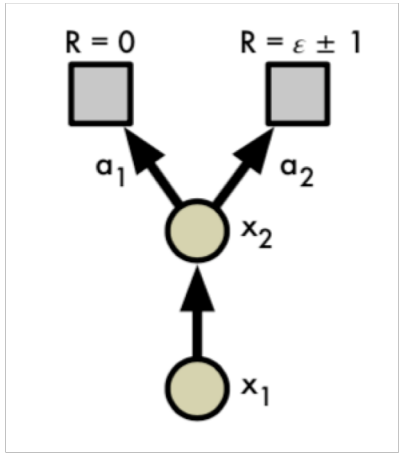
- $\tau^{\pi} Q(x,a) \coloneqq E\ R(x,a) + \gamma E_P \max_{a'} Q^{\pi}(x',a')$
- $\tau Z(x,a) \coloneqq \tau^{\pi} Z(x,a)$ for some π greedy w.r.t $E[Z(x,a)]$
- Some operator corresponds to greedy choice
- We hope that ultimately we can choose a greedy policy so that we can maximize the expectation return: Z(x, a)
- In conventional RL, yes, we will find the optimal value function.
- In distributional RL, no….
- i.e. there is no guarantee that by using the greedy principle we can get the optimal policy.

An Example to illustrate Non-contraction



- From an expectation point of view:
- If $\varepsilon = 0$, both action is equally good
- If $\varepsilon > 0$, action 2 is better
- If $\varepsilon < 0$, action 1 is better

An Example to illustrate Non-contraction



	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	$\epsilon \pm 1$

- We can show non-contraction: Z is further away from Z^*
- Z is my starting guess, let's say in state 1, Z is the same as the optimal policy Z^* except at (x_2, a_2), and I'm underestimating the value that I can get at (x_2, a_2), applying the greedy algorithm, I have to choose a_2 at x_2
- Apply the bellman update, value at x_1 is 0, and value at (x_2, a_2) is now correct.
- Wasserstein Metric captures all the moments of the distribution, even though you are getting closer to the expectation, you are going further away in terms of the variance

Distributional RL Algorithm

- Choice of value distribution
- Projected Bellman Update

Choice of value distribution

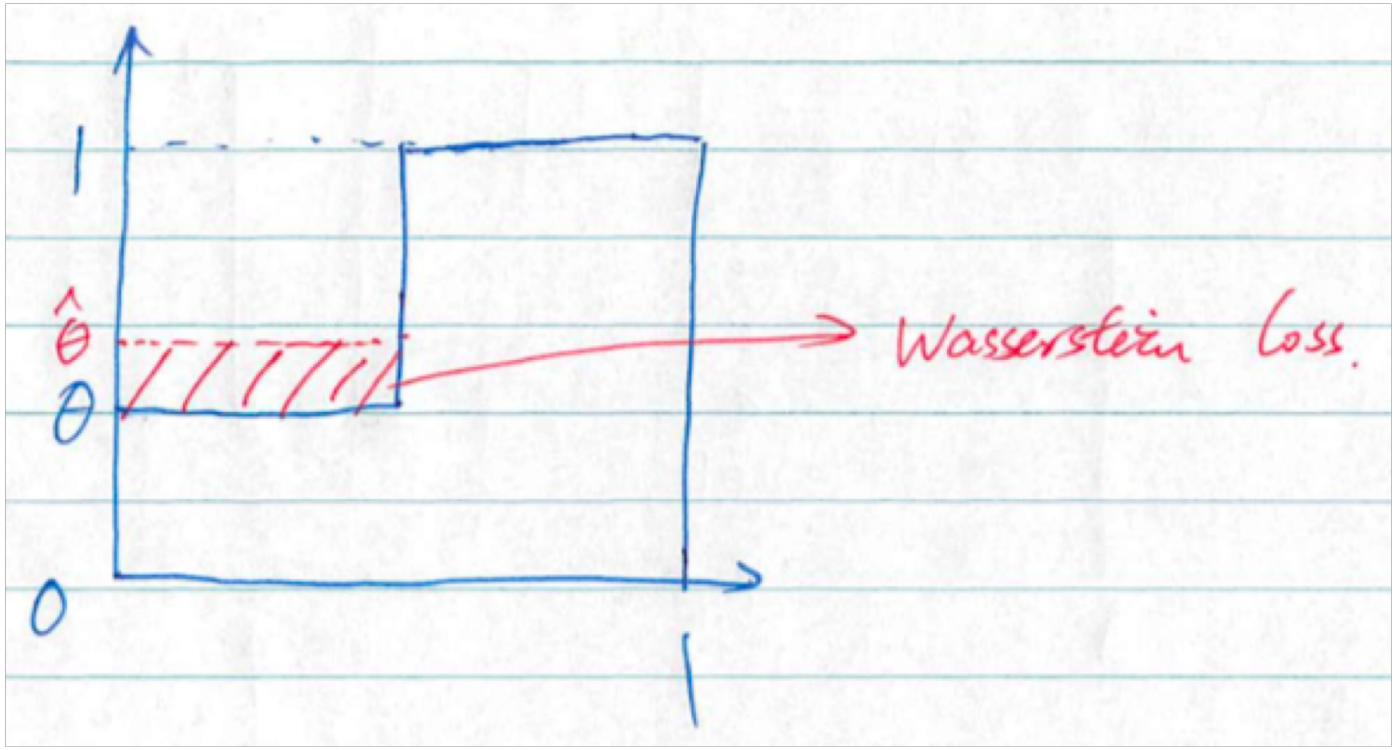
- We have to choose some approximation value distribution
- Discrete distribution (set):
  - It's support is set of atoms: $\{z_i = V_{MIN} + i\,\Delta z : 0 \leq i < N\}$, $\Delta z = \frac{V_{MAX} - V_{MIN}}{N-1}$
  - Referred the atoms as "canonical returns" of a value distribution.
  - Probability of each atom to be chosen:
    - $Z_\theta(x, a) = z_i \;\; w.p. \;\; p_i(x, a) = \frac{e^{\theta_i(x,a)}}{\sum_j e^{\theta_j(x,a)}}$ where $\theta_i(x, a)$ is some parametric model
- The discrete distribution has the advantages of being highly expressive and computationally friendly.

Projected Bellman Update

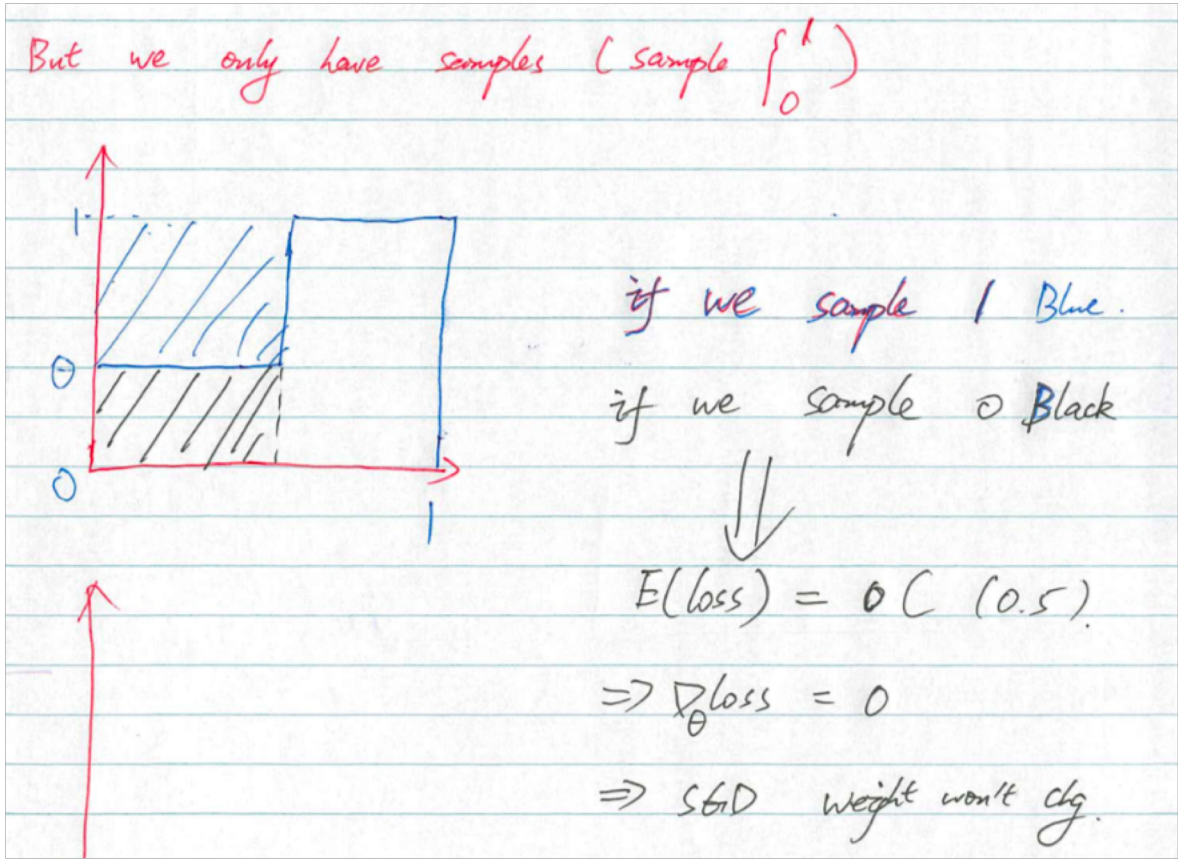
- It is natural to think of an algorithm
- Step 1: From x, a, sample a transition
  - $r, X', A' \sim R(x,a), P(.|x,a), \pi(.|X')$
- Step 2: Compute sample backup
  - $\tau^{\pi} Z(x,a) \coloneqq R(x,a) + \gamma Z(X', A')$
- Step 3: Minimize Wasserstein Loss
  - $l_{x,a}(\theta) \coloneqq w_1\left(\tau^{\pi}(x,a), Z_{\theta}(x,a)\right)$
- Step 4: SGD etc. on $l_{x,a}(\theta)$ w.r.t θ

Projected Bellman Update

- Turns out you can't do step 4 with Wasserstein Loss: Biased Wasserstein Gradient



Projected Bellman Update



Projected Bellman Update

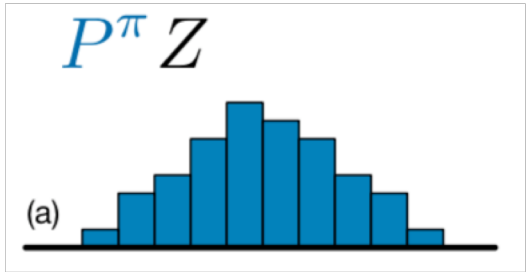
- Let's start again
- Step 1: From x, a, sample a transition
  - $r, X', A' \sim R(x,a), P(.|x,a), \pi(.|X')$
- Step 2: Compute sample backup
  - $\tau^{\pi} Z(x,a) \coloneqq R(x,a) + \gamma Z(X', A')$
- Step 3: Minimize KL divergence
  - But we just mentioned that KL divergence has problem with infinity values when we shift the distribution.

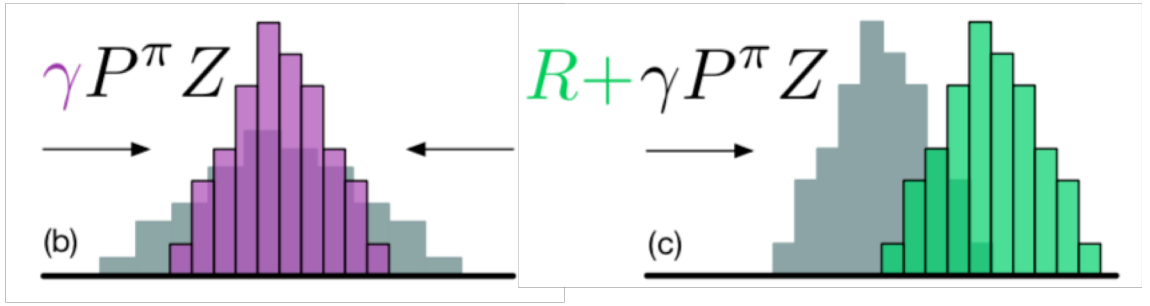
Projected Bellman Update

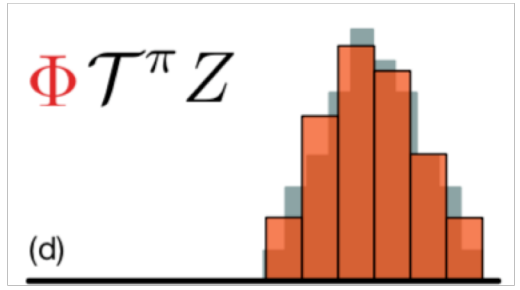
- Let's start again
- Step 1: From x, a, sample a transition
  - $r, X', A' \sim R(x,a), P(.|x,a), \pi(.|X')$
- Step 2: Compute sample backup
  - $\tau^{\pi} Z(x,a) \coloneqq R(x,a) + \gamma Z(X', A')$
- Step 3: Projection step
  - Output the sample backup and bring it back to the sample distribution
  - $\Phi \tau^{\pi} Z(x,a)$
- Step 4: Minimize the KL divergence

Projected Bellman Update

- Let's start again
- Step 1: From x, a, sample a transition
  - $r, X', A' \sim R(x,a), P(.|x,a), \pi(.|X')$
- Step 2: Compute sample backup
  - $\tau^{\pi} Z(x,a) \coloneqq R(x,a) + \gamma Z(X', A')$
- Step 3: Projection step
  - $\Phi \tau^{\pi} Z(x,a)$
- Step 4: Minimize the KL divergence

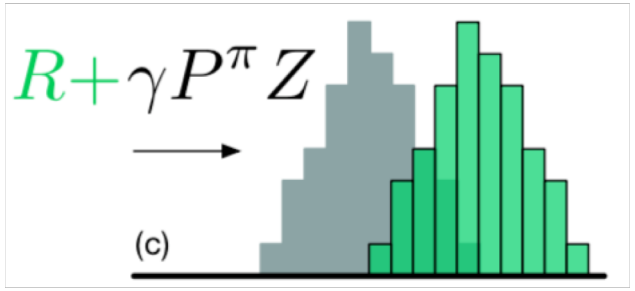


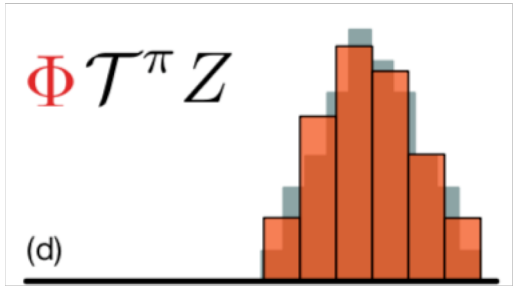




Projected Bellman Update

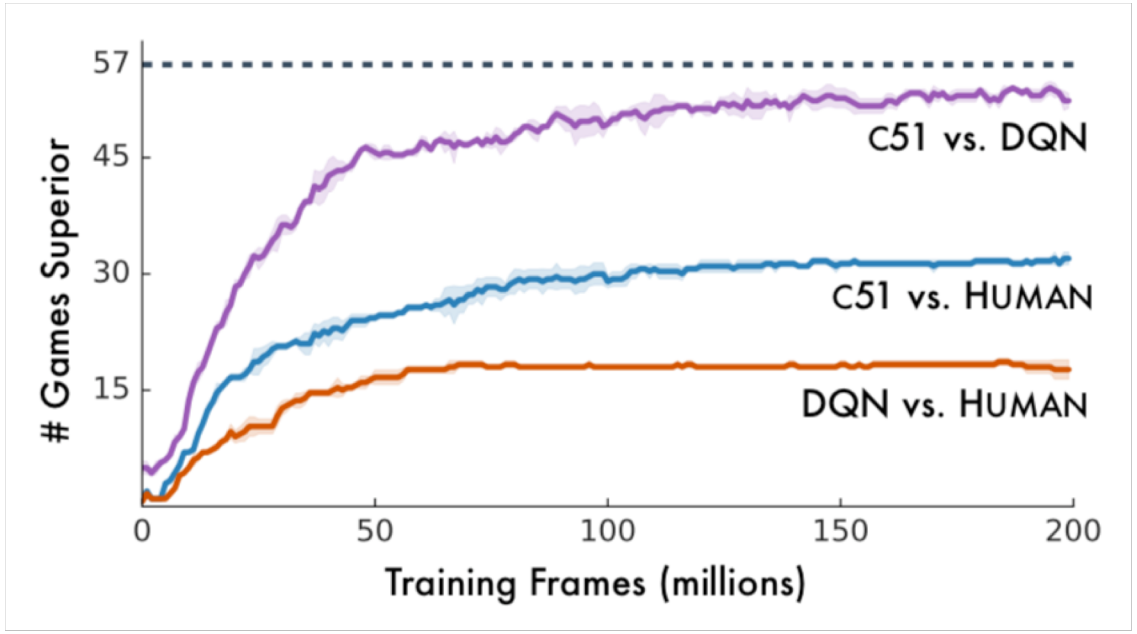
- Step 3: Projection step
  - Output the sample backup and bring it back to the sample distribution
  - $\Phi\tau^{\pi}Z(x,a)$
- Linear interpolation: map the target atoms to nearest neighbours





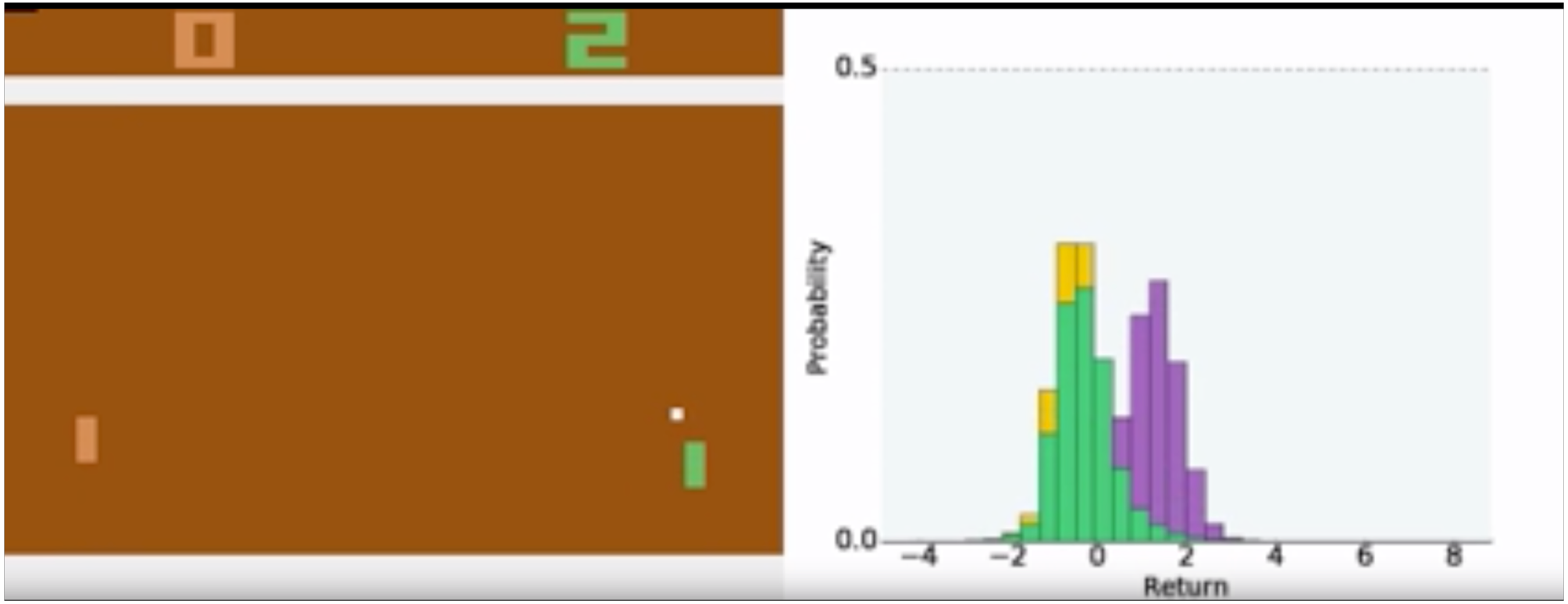
Scratch of the results

- Better performance; speed up learning process

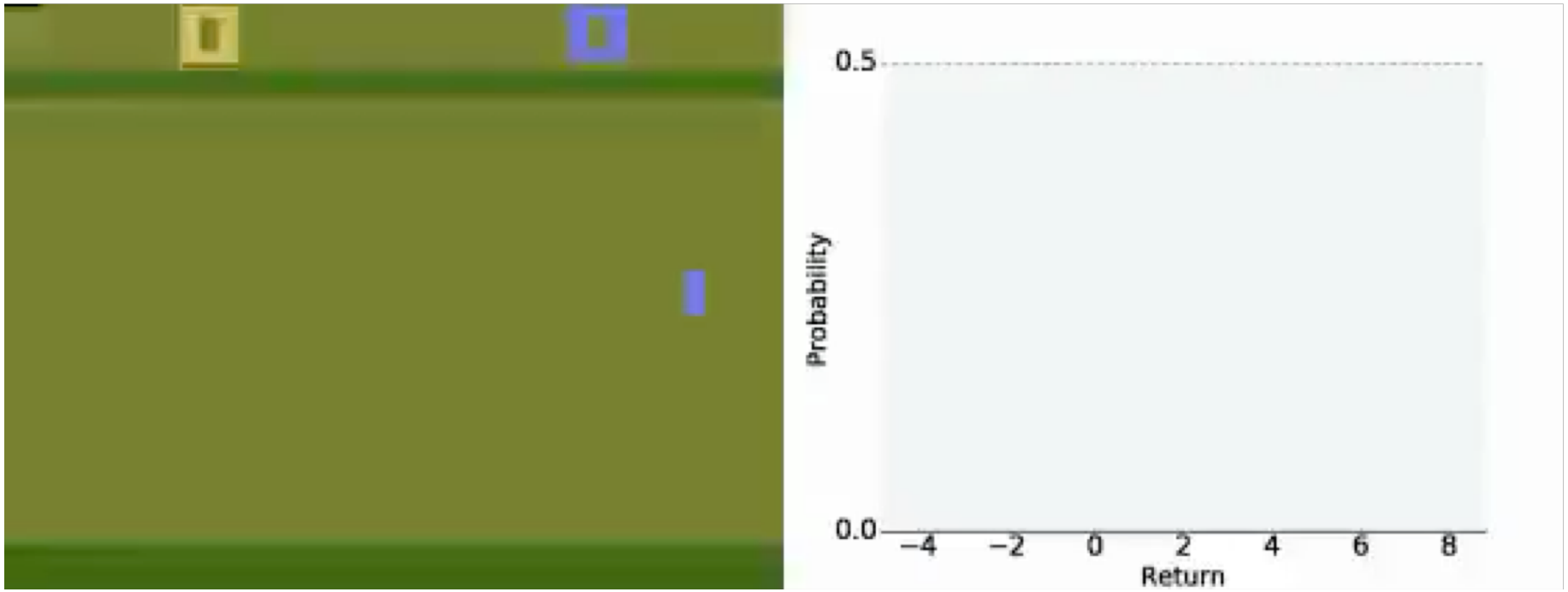


An example

- Atari Pong game: 3 actions: Up, Down, No-op

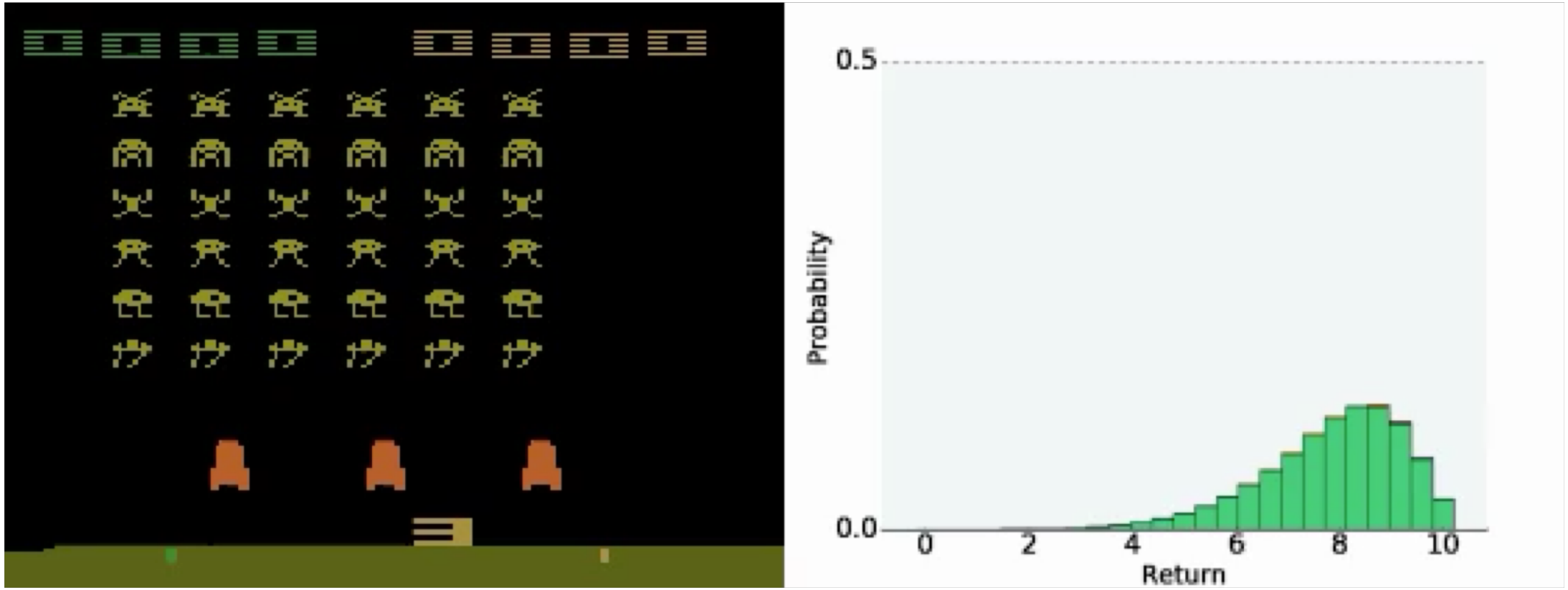


An example



https://www.youtube.com/watch?v=vIz5P6s80qA&t=17s

Another interesting example



https://www.youtube.com/watch?v=yFBwyPuO2Vg&t=2s

References

- Marc Bellemare's talk: https://www.youtube.com/watch?v=ba_l8IKoMvU