Distributional RL

10-403 Recitation 7

Conor Igoe — 04/23/2021

Distributional RL

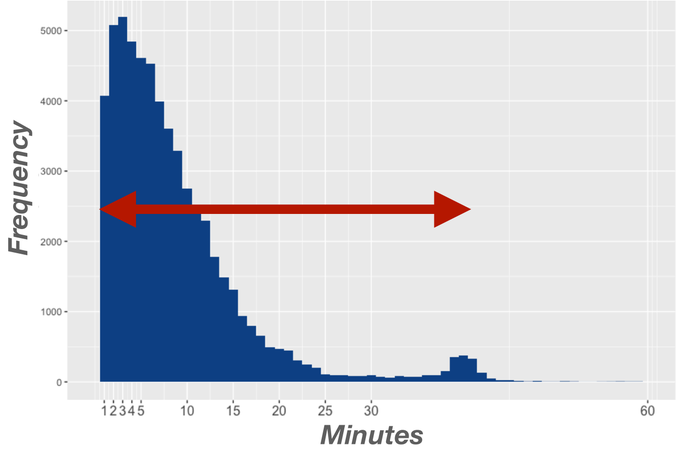
Overview

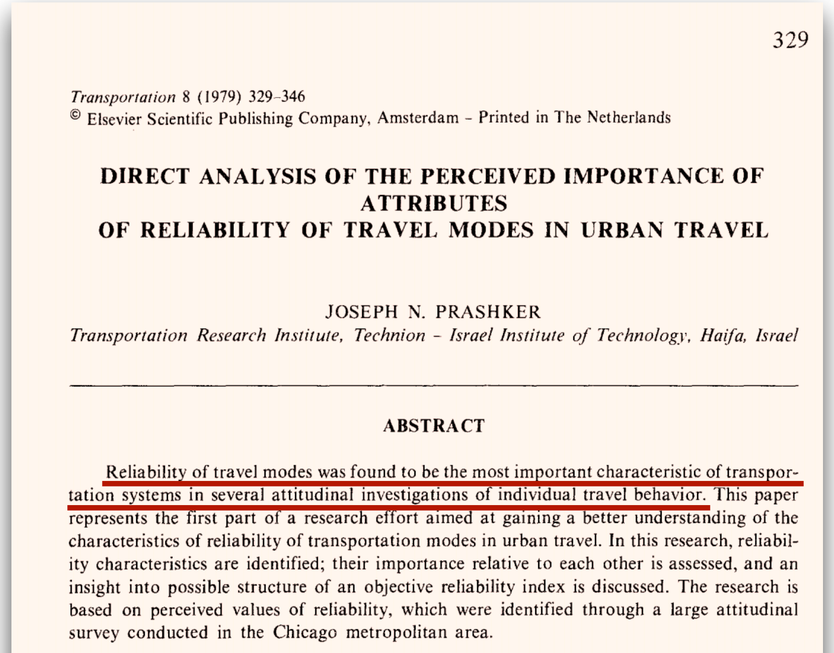
- Motivation
- Preliminaries
- Setting
- Distributional Bellman Equations & Operators
- C51 Algorithm
- Current Research Topics

Distributional RL

Motivation

How long do you have to wait for the bus?





329

Transportation 8 (1979) 329–346
© Elsevier Scientific Publishing Company, Amsterdam – Printed in The Netherlands

DIRECT ANALYSIS OF THE PERCEIVED IMPORTANCE OF ATTRIBUTES OF RELIABILITY OF TRAVEL MODES IN URBAN TRAVEL

JOSEPH N. PRASHKER
Transportation Research Institute, Technion – Israel Institute of Technology, Haifa, Israel

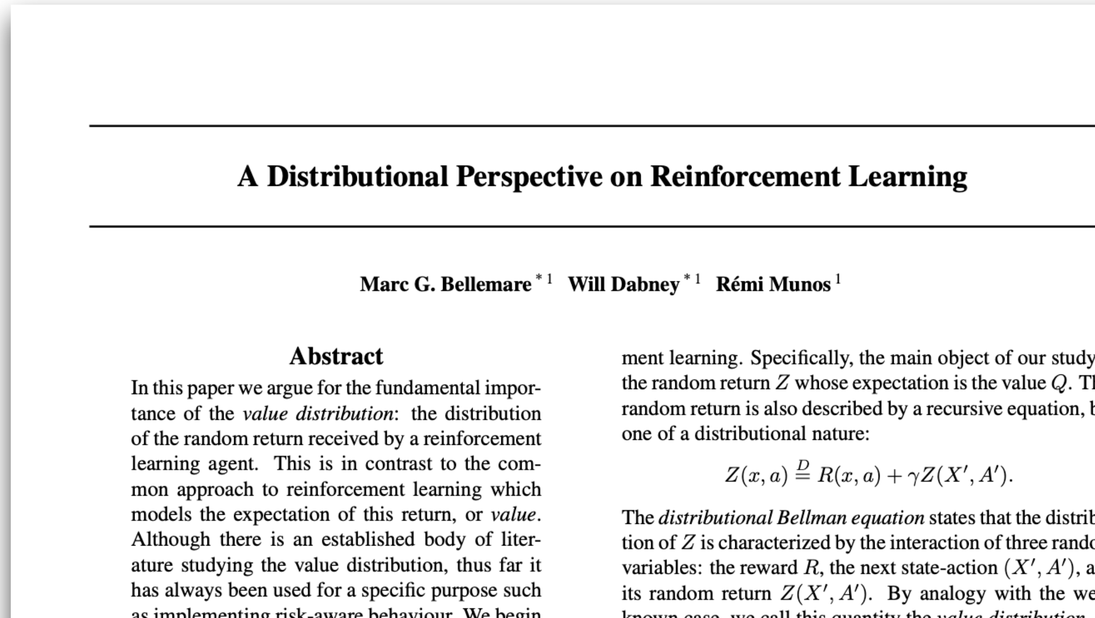
ABSTRACT

Reliability of travel modes was found to be the most important characteristic of transportation systems in several attitudinal investigations of individual travel behavior. This paper represents the first part of a research effort aimed at gaining a better understanding of the characteristics of reliability of transportation modes in urban travel. In this research, reliability characteristics are identified; their importance relative to each other is assessed, and an insight into possible structure of an objective reliability index is discussed. The research is based on perceived values of reliability, which were identified through a large attitudinal survey conducted in the Chicago metropolitan area.

Distributional RL

Motivation

- So far we have been focusing on maximising Expected returns
- Distributional versions of Bellman equations have been studied for almost as long as the usual expected versions, originally motivated by risk-aware decision making settings
- Recently (2017), researchers have found that taking a distributional approach to DRL has benefits even if all we care about is expected return
- Still being studied to gain understanding as to why distributional approach helps the expected setting



A Distributional Perspective on Reinforcement Learning

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

Abstract

In this paper we argue for the fundamental importance of the *value distribution*: the distribution of the random return received by a reinforcement learning agent. This is in contrast to the common approach to reinforcement learning which models the expectation of this return, or *value*. Although there is an established body of literature studying the value distribution, thus far it has always been used for a specific purpose such as implementing risk-aware behaviour. We begin

ment learning. Specifically, the main object of our study the random return Z whose expectation is the value Q. T random return is also described by a recursive equation, b one of a distributional nature:

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

The *distributional Bellman equation* states that the distrib tion of Z is characterized by the interaction of three rando variables: the reward R, the next state-action (X',A'), a its random return $Z(X',A')$. By analogy with the we

Distributional RL

Preliminaries

- *What is an operator?*
- *What does is mean for two discrete random variables to be equivalent?*
- *What is the difference between* $\mathscr{F}$ *and* $\mathscr{Z}$ *?*
- *What is the difference between* $z(s, a)$ *and* $z(S, A)$ *?*

Distributional RL

Preliminaries

- *What is an operator?*
  - Suppose we have a space of functions $\mathscr{F}$, where each $f \in \mathscr{F}$ is a function of the form $f : \mathscr{X} \to \mathscr{Y}$
  - We call T an operator if it is a mapping of the form $T : \mathscr{F} \to \mathscr{F}$

Distributional RL

Preliminaries

- *What is an operator?*
  - We write:
    - $f \in \mathscr{F}$ to denote the function
    - $f(x)$ to denote the function evaluated at point $x \in \mathscr{X}$
    - Tf to denote the function returned by the operator T
    - $(Tf)(x)$ to denote the function returned by the operator T evaluated at point $x \in \mathscr{X}$

Distributional RL

Preliminaries

- *What is an operator?*
  - **Example:** if $\mathscr{F}$ is the space of functions of the form $\mathscr{S} \times \mathscr{A} \to \mathbb{R}$, then the Bellman operator $T^\pi : \mathscr{F} \to \mathscr{F}$ is an operator on this space of functions, and for any $f \in \mathscr{F}$, we know that $(T^\pi f)(s, a)$ will be closer to $q^\pi(s, a)$ than $f(s, a)$ for any $(s, a) \in \mathscr{S} \times \mathscr{A}$ (because T^π is a contraction)
  - In the rest of these slides, $\mathscr{F}$ will denote the space functions of the form $\mathscr{S} \times \mathscr{A} \to \mathbb{R}$

Distributional RL

Preliminaries

- *What does is mean for two discrete random variables to be equivalent?*
  - Suppose $X \sim \mathbb{P}_X$ is a discrete RV with pmf $p_X(x)$
  - Suppose $Y \sim \mathbb{P}_Y$ is a discrete RV with pmf $p_Y(y)$
  - We say that X is *equivalent* to Y, and write $X \overset{D}{=} Y$ if $p_X(x) = p_Y(y)$ whenever $x = y$
  - Note that this says nothing about the dependency between X and Y
  - (We also write $\mathbb{P}_X \overset{D}{=} \mathbb{P}_Y$, which can be interpreted as implying that $X \overset{D}{=} Y$)

Distributional RL

Preliminaries

- *What is the difference between $\mathscr{F}$ and $\mathscr{Z}$?*
  - $\mathscr{F}$ is space of functions where each function $f \in \mathscr{F}$ is a mapping of the form $f : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$
  - $\mathscr{Z}$ is a special space of functions used in Distributional RL
  - Each function $z \in \mathscr{Z}$ is a mapping of the form $z : \mathscr{S} \times \mathscr{A} \to \Delta(\mathbb{R})$
    - Here I am using $\Delta(\mathbb{R})$ to denote the set of discrete probability distributions over the real numbers

Distributional RL

Preliminaries

- *What is the difference between $\mathscr{F}$ and $\mathscr{Z}$?*
  - **Example:** $z_1 \in \mathscr{Z}$ could be defined as
    - $z_1(s, a) =$ Uniform({1,2,...,10}) for all $(s, a) \in \mathscr{S} \times \mathscr{A}$
  - **Example:** $z_2 \in \mathscr{Z}$ could be defined as
    - $z_2(s, a) = \begin{cases} \text{Uniform}(\{1,2,...,10\}) & \text{if } s > 3 \\ \text{Binomial}(10,0.5) & \text{otherwise} \end{cases}$ for all $(s, a) \in \mathscr{S} \times \mathscr{A}$

Distributional RL

Preliminaries

- *What is the difference between $\mathscr{F}$ and $\mathscr{Z}$?*
  - **Example:** $z_R \in \mathscr{Z}$ could be defined as
    - $z_R(s,a) = \mathbb{P}\left(R \in \cdot \,\middle|\, S = s, A = a\right)$ for all $(s,a) \in \mathscr{S} \times \mathscr{A}$

Distributional RL

Preliminaries

- *What is the difference between $z(s, a)$ and $z(S, A)$?*
  - $z \in \mathscr{Z}$ is a deterministic function of the form $z : \mathscr{S} \times \mathscr{A} \to \Delta(\mathbb{R})$
  - $(s, a) \in \mathscr{S} \times \mathscr{A}$ is an arbitrary (i.e. *not* random) point in state-action space
  - $z(s, a)$ lies in $\Delta(\mathbb{R})$, so it is a probability distribution, but it is deterministic in the sense that it is a specific probability distribution in the set $\Delta(\mathbb{R})$.

Distributional RL

Preliminaries

- *What is the difference between $z(s, a)$ and $z(S, A)$?*
  - S, A are two random variables
  - Ordinarily, $z(S, A)$ would be random, and would therefore have a probability distribution defined over $\Delta(\mathbb{R})$ (i.e. a probability distribution *over probability distributions*). However, it is common in Distributional RL to use $z(S, A)$ to denote the *mixture* distribution suggested by the RVs S, A, which is *not* random.

Distributional RL

Preliminaries

- *What is the difference between $z(s, a)$ and $z(S, A)$?*
  - **Example:**
    - Suppose $z_R(s, a) = \mathbb{P}\left(R \in \cdot \,\middle|\, S = s, A = a\right)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$
    - Suppose $S, A = \begin{cases} s_1, a_1 & \text{with probability 0.5} \\ s_2, a_2 & \text{with probability 0.5} \end{cases}$
    - Then $z_R(S, A)$ denotes the mixture distribution combining $z_R(s_1, a_1)$ and $z_R(s_2, a_2)$ with equal weights (so we have $z_R(S, A) \in \Delta(\mathbb{R})$).

Questions?

- *What is an operator?*
- *What does is mean for two discrete random variables to be equivalent?*
- *What is the difference between* $\mathscr{F}$ *and* $\mathscr{Z}$ *?*
- *What is the difference between* $z(s, a)$ *and* $z(S, A)$ *?*

Distributional RL

Preliminaries

- Quick note:
  - There is one point where we revert to the usual stochastic interpretation for $z(S, A)$ instead of its mixture distribution interpretation. I will point this out when it happens.

Distributional RL

Setting

- Finite, episodic MDP $\mathscr{M} = (\mathscr{S}, \mathscr{A}, \mathscr{R}, p, \rho_0, \gamma)$
  - $\implies$ finitely many realisable discounted returns
- Assume that $(R \perp\!\!\!\perp S' \mid S, A)$ i.e. $p(r, s' \mid s, a) = p_R(r \mid s, a) p_{S'}(s' \mid s, a)$
- Stochastic policy $\pi : \mathscr{S} \times \mathscr{A} \to [0,1]$
- q^π is the true state-action value function for policy π
- q is an arbitrary function of the form $q : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ (so we have $q \in \mathscr{F}$)

Distributional RL

Setting

- Bellman Equation:
  - $q^\pi(s, a) = \mathbb{E}\left[R + \gamma q^\pi(S', A') \,\middle|\, S = s, A = a\right]$
- Bellman Operator T^π applied to q and evaluated at (s, a):
  - $(T^\pi q)(s, a) = \mathbb{E}\left[R + \gamma q(S', A') \,\middle|\, S = s, A = a\right]$
  - Note that $T^\pi : \mathscr{F} \to \mathscr{F}$

Distributional RL

Distributional Bellman Equations & Operators

- In Distributional RL, we have something similar, but we need to be careful with our interpretation
- $z^\pi \in \mathscr{Z}$ is the true *distributional* state-action value function for policy π
  - Remember that $z^\pi : \mathscr{S} \times \mathscr{A} \to \Delta(\mathbb{R})$
  - $z^\pi(s, a)$ is the discrete probability distribution

    $\mathbb{P}\left(\sum_{t=0}^{\infty} \gamma^t R_{t+1} \in \cdot \;\middle|\; S_0 = s, A_0 = a \right)$ assuming policy π in MDP $\mathscr{M}$
  - Note that (with abuse of notation) $\mathbb{E}\left[z^\pi(s, a)\right] = q^\pi(s, a)$ and $\mathbb{E}\left[z^\pi\right] = q^\pi$

Distributional RL

Distributional Bellman Equations & Operators

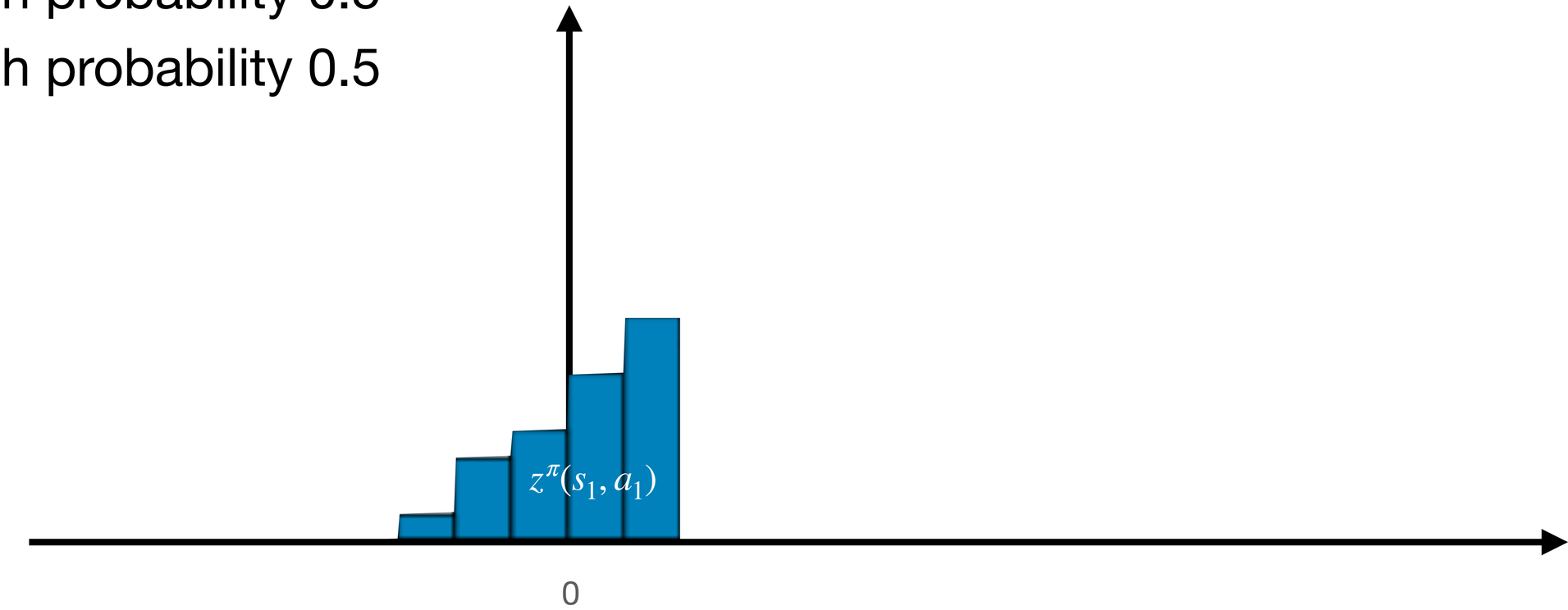
- Distributional Bellman Equation:
  - $z^\pi(s,a) \overset{D}{=} z_R(s,a) + \gamma z^\pi(S',A')$
  - $z_R(s,a)$ is the discrete probability distribution $\mathbb{P}\left(R \in \cdot \,\middle|\, S = s, A = a\right)$
  - $z_R(s,a) + \gamma z^\pi(S',A')$ is abusive notation for the discrete probability distribution of the RV $X + \gamma Y$ where $X \sim z_R(s,a)$ and $Y \sim z^\pi(S',A')$ with S', A' distributed according to MDP $\mathscr{M}$ under policy π from state-action pair (s,a)
    - Remember that $z^\pi(S',A')$ denotes the mixture distribution!

Distributional RL

Distributional Bellman Equations & Operators

- $z^{\pi}(s,a) \overset{D}{=} z_R(s,a) + \gamma z^{\pi}(S',A')$

$$S',A' = \begin{cases} s_1, a_1 & \text{with probability 0.5} \\ s_2, a_2 & \text{with probability 0.5} \end{cases}$$



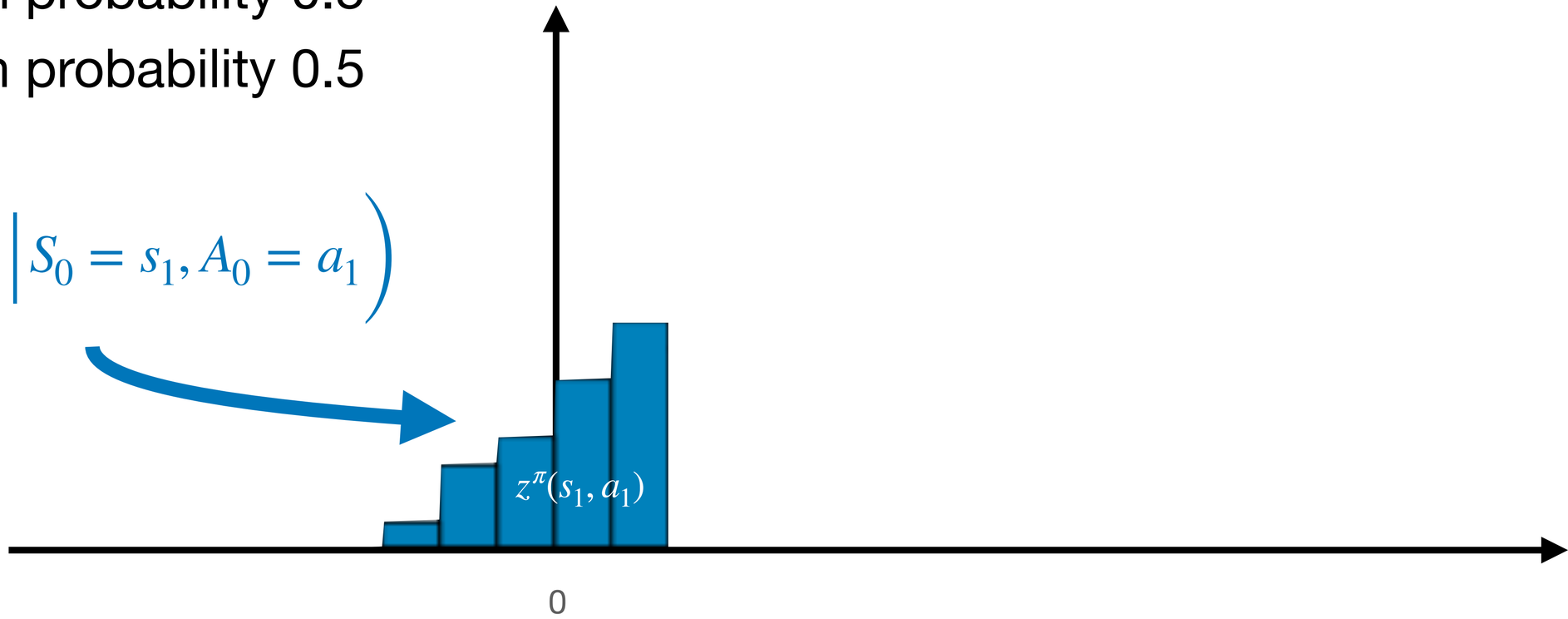
Distributional RL

Distributional Bellman Equations & Operators

- $z^\pi(s,a) \overset{D}{=} z_R(s,a) + \gamma z^\pi(S', A')$

$$S', A' = \begin{cases} s_1, a_1 & \text{with probability 0.5} \\ s_2, a_2 & \text{with probability 0.5} \end{cases}$$

$$\mathbb{P}\left(\sum_{t=0}^{\infty} \gamma^t R_{t+1} \in \cdot \,\middle|\, S_0 = s_1, A_0 = a_1 \right)$$

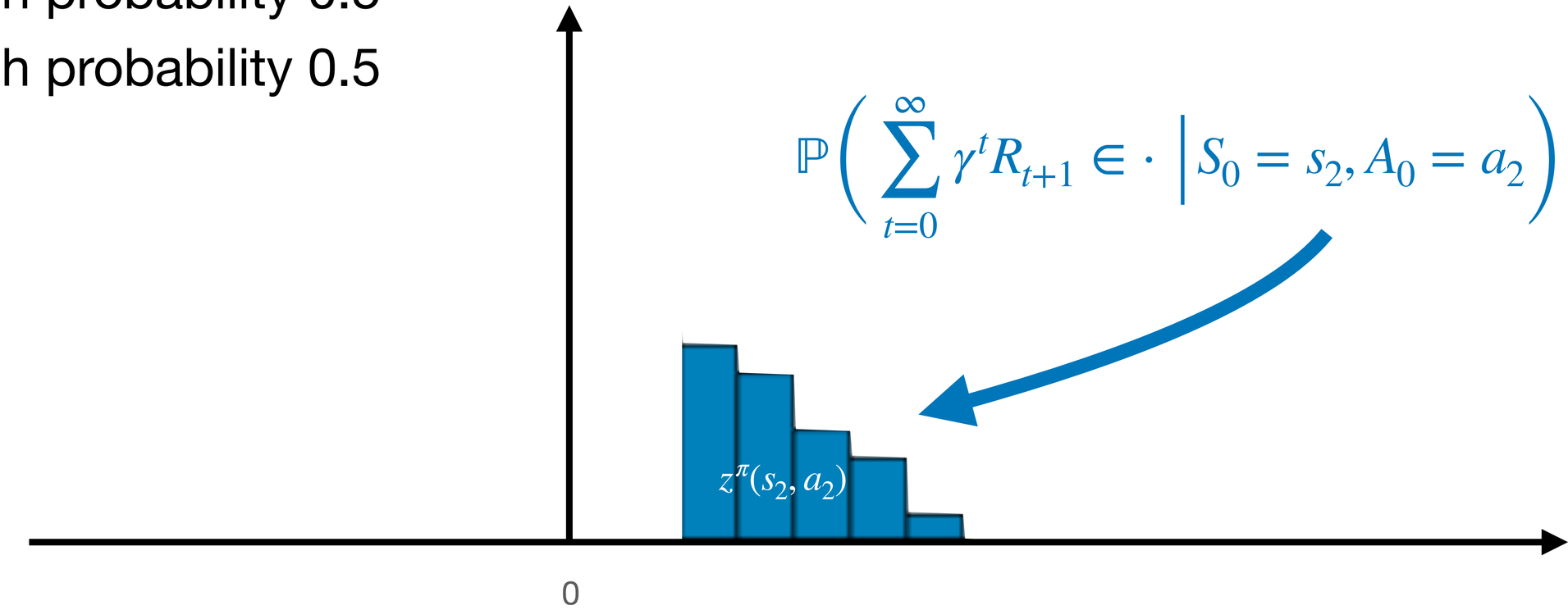


Distributional RL

Distributional Bellman Equations & Operators

- $z^\pi(s,a) \overset{D}{=} z_R(s,a) + \gamma z^\pi(S', A')$

$$S', A' = \begin{cases} s_1, a_1 & \text{with probability 0.5} \\ s_2, a_2 & \text{with probability 0.5} \end{cases}$$



Distributional RL

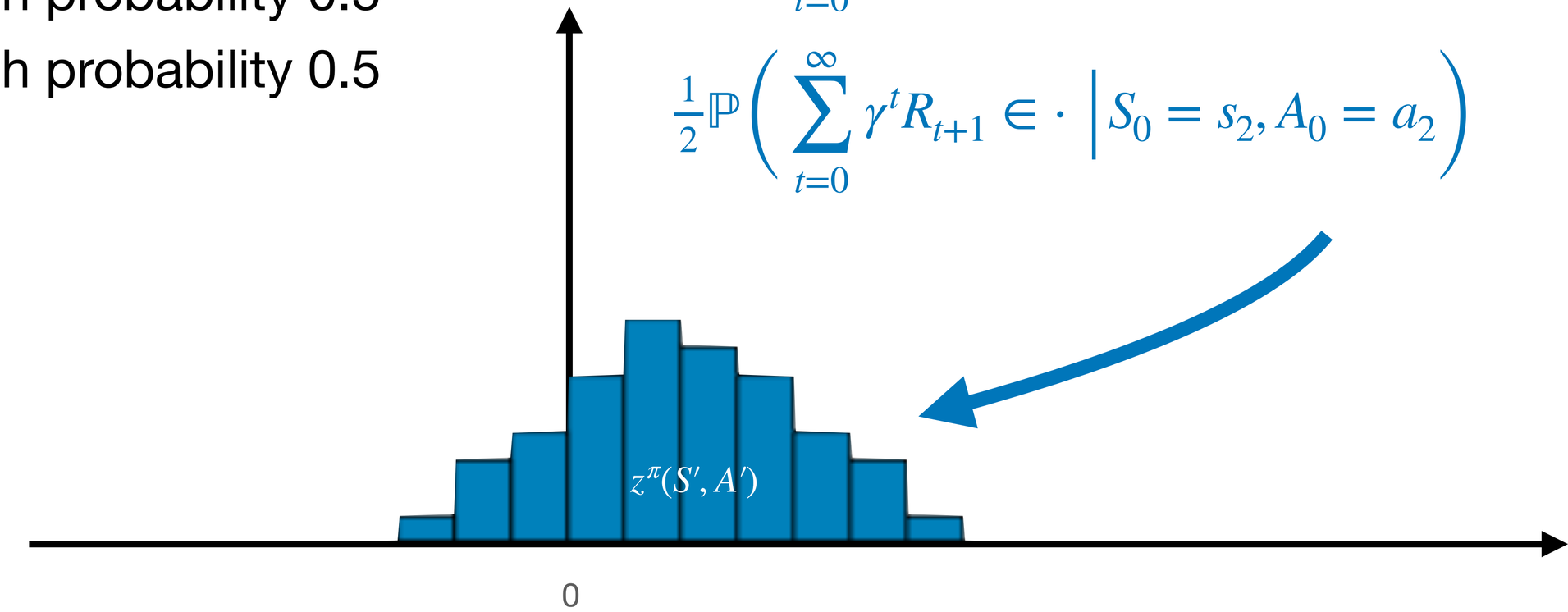
Distributional Bellman Equations & Operators

- $z^\pi(s,a) \overset{D}{=} z_R(s,a) + \gamma z^\pi(S', A')$

$$S', A' = \begin{cases} s_1, a_1 & \text{with probability 0.5} \\ s_2, a_2 & \text{with probability 0.5} \end{cases}$$

$$\frac{1}{2}\mathbb{P}\left(\sum_{t=0}^{\infty} \gamma^t R_{t+1} \in \cdot \,\middle|\, S_0 = s_1, A_0 = a_1\right) +$$

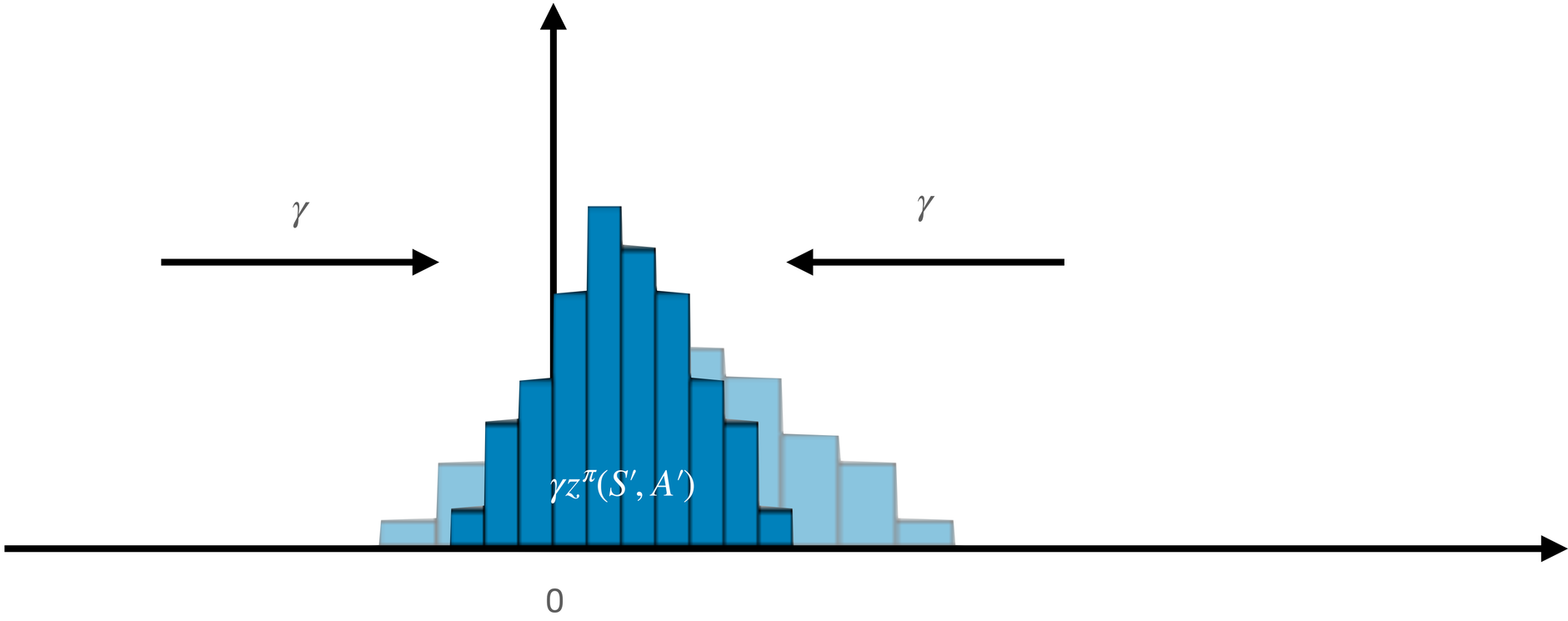
$$\frac{1}{2}\mathbb{P}\left(\sum_{t=0}^{\infty} \gamma^t R_{t+1} \in \cdot \,\middle|\, S_0 = s_2, A_0 = a_2\right)$$



Distributional RL

Distributional Bellman Equations & Operators

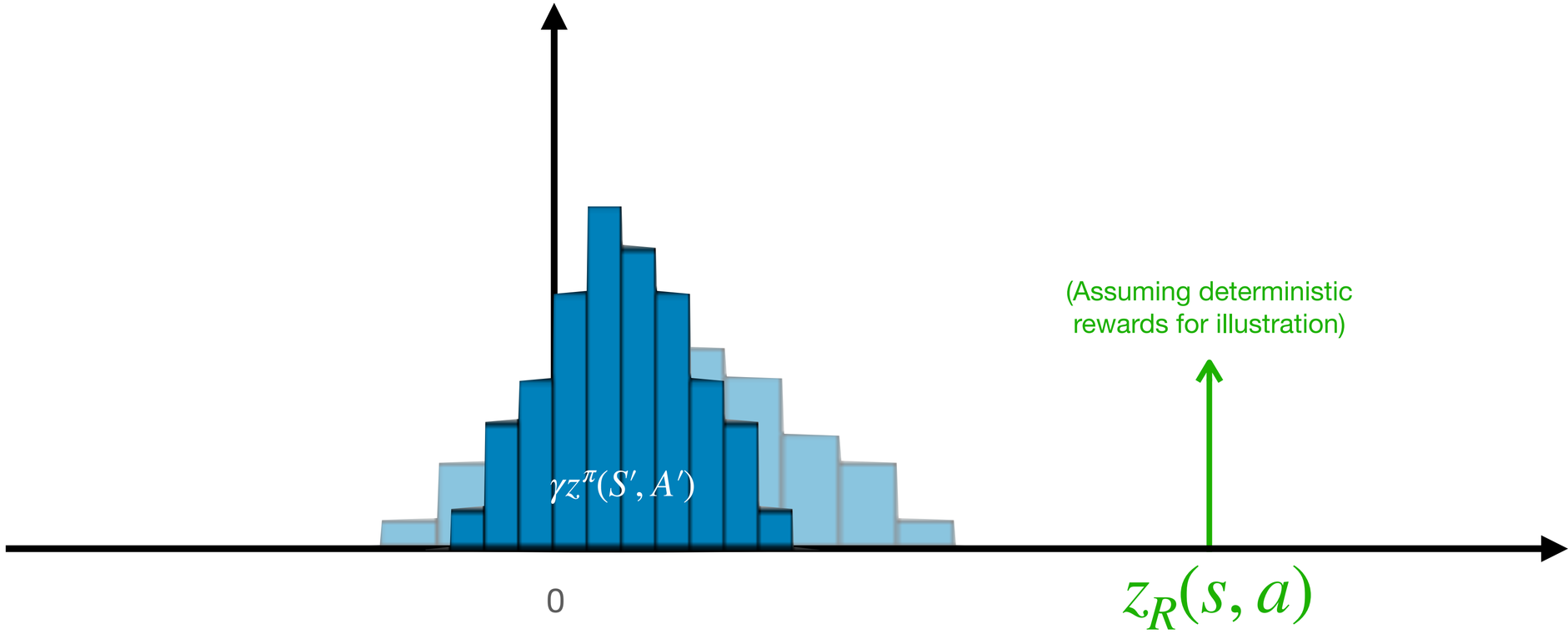
- $z^{\pi}(s,a) \overset{D}{=} z_R(s,a) + \gamma z^{\pi}(S',A')$



Distributional RL

Distributional Bellman Equations & Operators

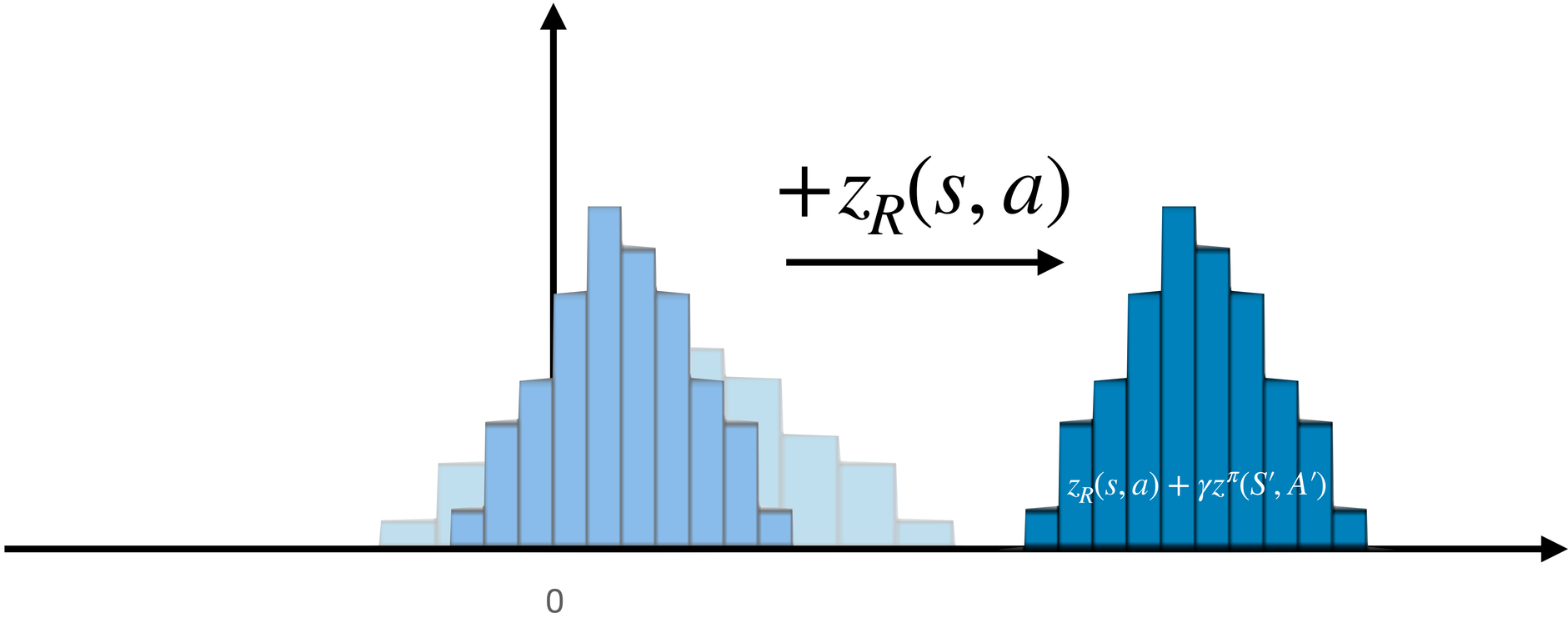
- $z^\pi(s,a) \overset{D}{=} z_R(s,a) + \gamma z^\pi(S',A')$



Distributional RL

Distributional Bellman Equations & Operators

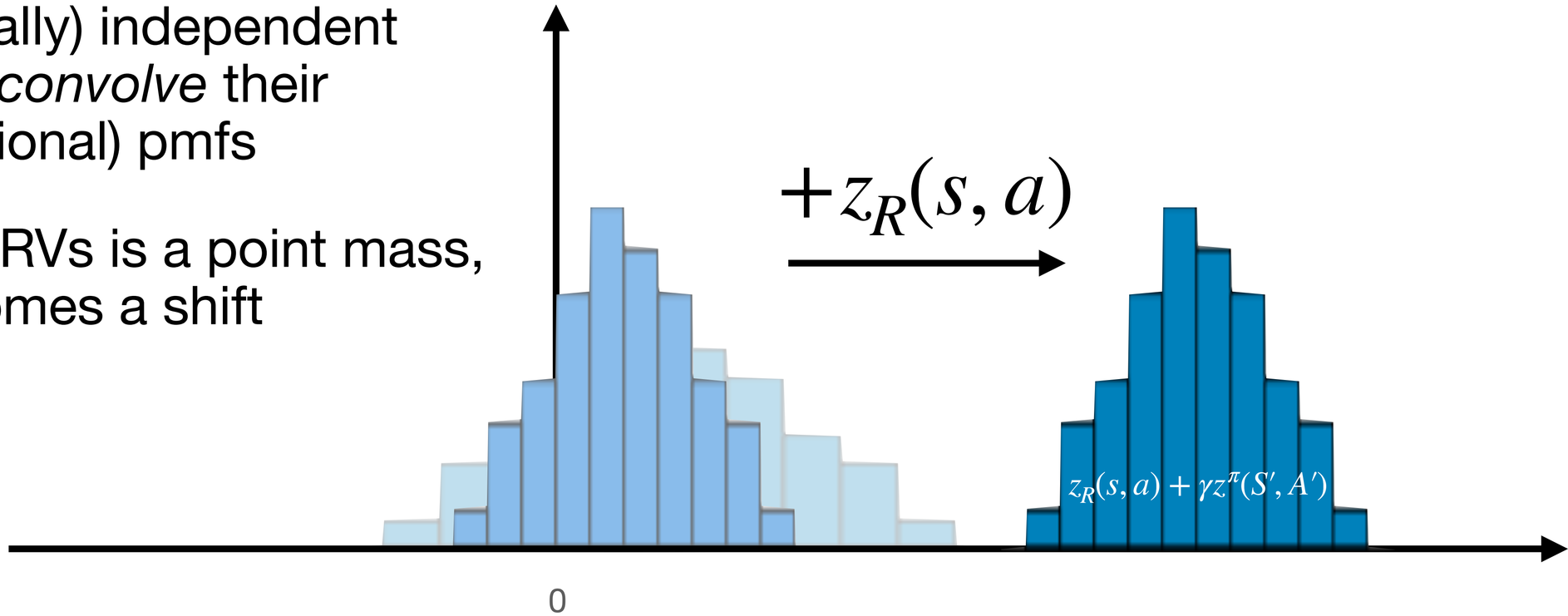
- $z^{\pi}(s,a) \overset{D}{=} z_R(s,a) + \gamma z^{\pi}(S',A')$



Distributional RL

Distributional Bellman Equations & Operators

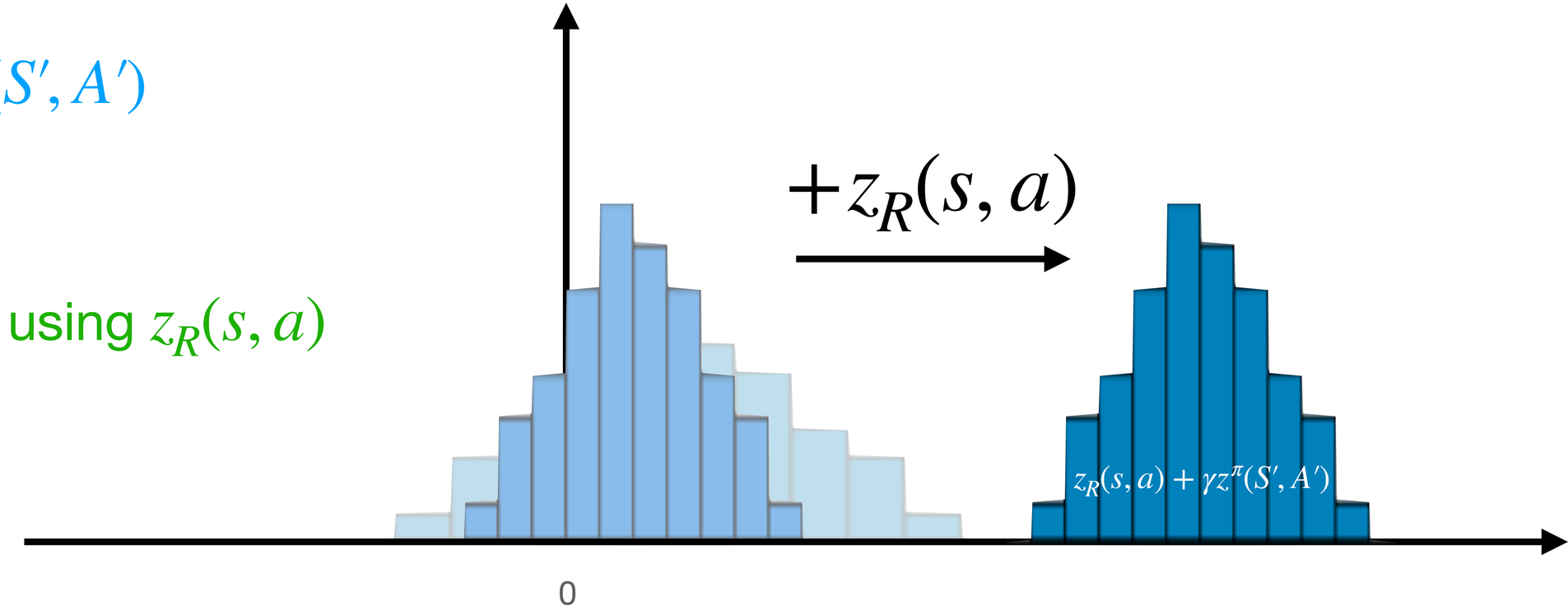
- $z^\pi(s,a) \overset{D}{=} z_R(s,a) + \gamma z^\pi(S',A')$
- In general, to find the pmf of the sum of two (conditionally) independent RVs, we need to *convolve* their individual (conditional) pmfs
- When one of the RVs is a point mass, convolution becomes a shift



Distributional RL

Distributional Bellman Equations & Operators

- $z^\pi(s,a) \stackrel{D}{=} z_R(s,a) + \gamma z^\pi(S',A')$
- RHS has 3 steps:
  - find mixture $z(S',A')$
  - scale by γ
  - convolve/shift using $z_R(s,a)$



Distributional RL

Distributional Bellman Equations & Operators

- Just like in classical Dynamic Programming, we can define a Distributional Bellman *Operator* U^π. Specifically, the Distribuitonal Bellman Operator applied to z and evaluated at (s, a) is given by:
  - $(U^\pi z)(s, a) \overset{D}{=} z_R(s, a) + \gamma z(S', A')$
  - That is, we start with some arbitrary $z \in \mathscr{Z}$, and we update the *distribution* at each (s, a) to be the distribution of the RV $X + \gamma Y$, where $X \sim z_R(s, a)$, and $Y \sim z(S', A')$
  - Note that $U^\pi : \mathscr{Z} \to \mathscr{Z}$, whereas $T^\pi : \mathscr{F} \to \mathscr{F}$

Distributional RL

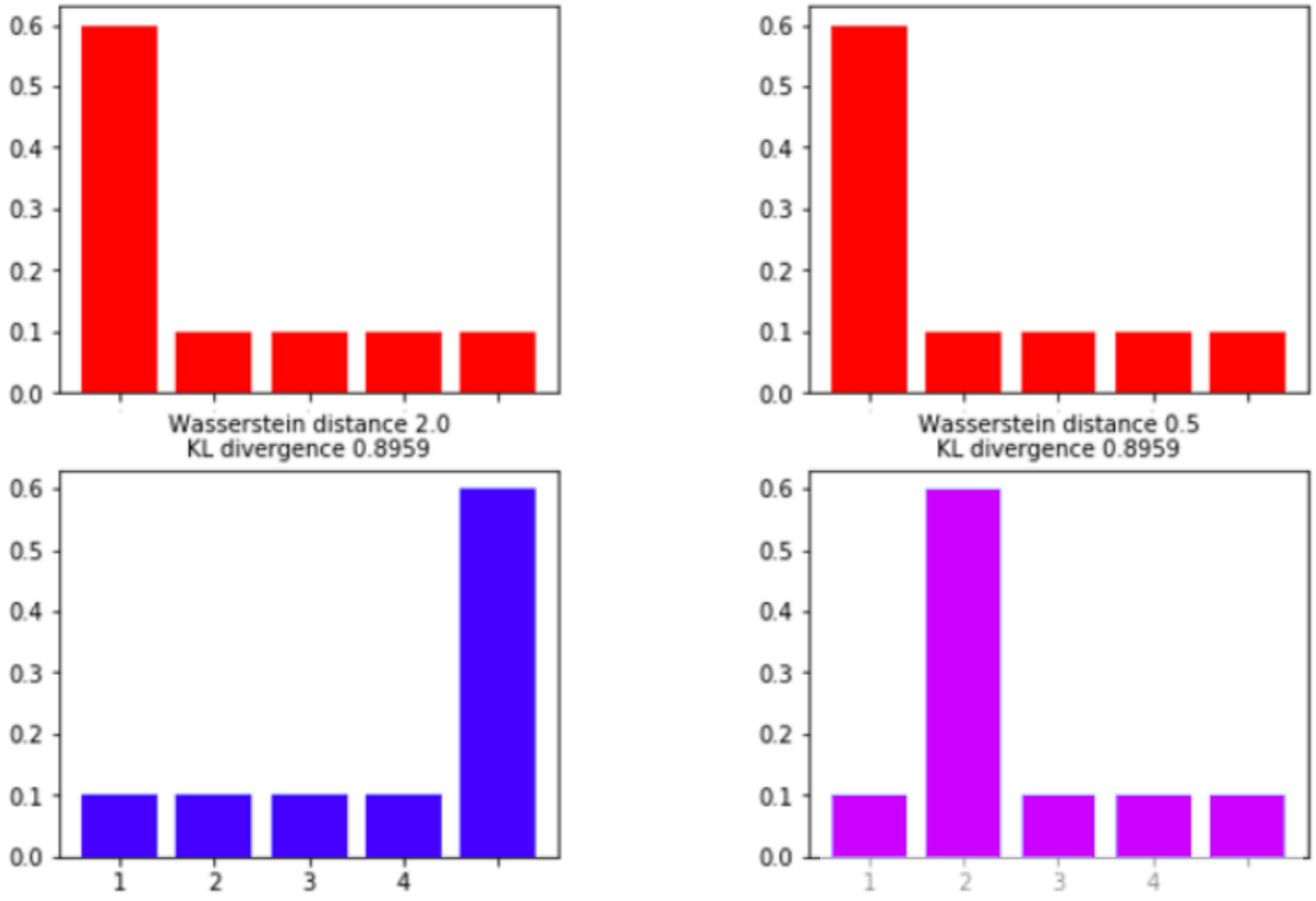
Distributional Bellman Equations & Operators

- In [Bellemare et al. 2017], they showed that U^π is a contraction, similar to how T^π is a contraction
  - For any two $q, q' \in \mathscr{F}$, we needed to find the biggest difference across $\mathscr{S} \times \mathscr{A}$, and "difference" was trivial to calculate: just $|q(s,a) - q'(s,a)|$
  - For any two $z, z' \in \mathscr{Z}$, now it is less straightforward to measure "difference" between $z(s,a)$ and $z'(s,a)$
  - Turns out Wasserstein metric is a convenient way to measure distance for analysing U^π

Distributional RL

Distributional Bellman Equations & Operators

- Intuitively:
  - **KL divergence** only penalises when two distributions disagree on the probability of an event, but there is no notion of “similarity” of events
  - **Wasserstein** distances additionally measure similarity of events and penalise disagreements accordingly



Questions?

- $z^{\pi}(s, a) \stackrel{D}{=} z_R(s, a) + \gamma z^{\pi}(S', A')$
  - find mixture $z(S', A')$
  - scale by γ
  - convolve/shift using $z_R(s, a)$

Distributional RL

Distributional Bellman Equations & Operators

- Unfortunately, things are not so simple for the control case
- Main issue is that it's not clear what is the best way to generalise the maximisation step in the Bellman Optimality Equations: how should we maximise over distributions?
  - $(U^\star z)(s,a) \stackrel{D}{=} z_R(s,a) + \gamma \max_{a'} z(S',a')$ not defined

Distributional RL

Distributional Bellman Equations & Operators

- In [Bellemare et al. '17], they propose the definition $U^\star z := U^{\pi^\star} z$ where $\pi^\star$is any greedy policy w.r.t. $\mathbb{E}[z]$
- Unfortunately, $U^\star$ is not a contraction in any metric
- However, the authors show that $\lim_{k \to \infty} \mathbb{E}[(U^\star)^k z] \to q^\star$ which motivates their proposed algorithm

Distributional RL

C51 Algorithm

- For a tractable algorithm, we need a differentiable space of functions $\mathcal{Z}_\theta$ that interacts well with the Distributional Bellman (Optimality) Operators
- In [Bellemare et al. '17], authors propose using a categorical distribution parameterised by θ and then modify DQN to train approximate distributional value functions
- They have a minimum and maximum return, and then they discretise the real number line into N uniform intervals (they found $N = 51$ to be effective for playing Atari, hence the name)

Distributional RL

C51 Algorithm

- C51 = DQN, but z_θ instead of q_θ; train z_θ by taking stochastic gradient descent steps on loss:
  - $KL\left(\delta(R) + \gamma z_{\bar{\theta}}(S', A^\star) \,\middle\|\, z_\theta(S, A)\right)$
  - with (S, A, R, S') from replay buffer and $A^\star = \arg\max_{a'} \mathbb{E}\left[z_{\bar{\theta}}(S', a')\right]$

Distributional RL

C51 Algorithm

- C51 = DQN, but z_θ instead of q_θ; train z_θ by taking stochastic gradient descent steps on loss:
  - $KL\left(\delta(R) + \gamma z_{\bar{\theta}}(S', A^\star) \,\middle\|\, z_\theta(S, A)\right)$
  - with (S, A, R, S') from replay buffer and $A^\star = \arg\max_{a'} \mathbb{E}\left[z_{\bar{\theta}}(S', a')\right]$
    - Note: we are returning to the more standard *stochastic* interpretation of $z(S, A)$ here instead of the mixture interpretation during operator analysis
      - i.e. $z_\theta(S, A)$, $z_{\bar{\theta}}(S', A^\star)$ are random *not* mixture distributions

Distributional RL

C51 Algorithm

- C51 = DQN, but z_θ instead of q_θ; train z_θ by taking stochastic gradient descent steps on loss:
  - $KL\left(\delta(R) + \gamma z_{\bar{\theta}}(S', A^\star) \,\middle\|\, z_\theta(S, A)\right)$
  - with (S, A, R, S') from replay buffer and $A^\star = \arg\max_{a'} \mathbb{E}\left[z_{\bar{\theta}}(S', a')\right]$
    - Note: $\mathbb{E}$ is *not* taken over randomness in S' but is just w.r.t. the distribution returned by $z_{\bar{\theta}}$ (so $A^\star$ is a RV)

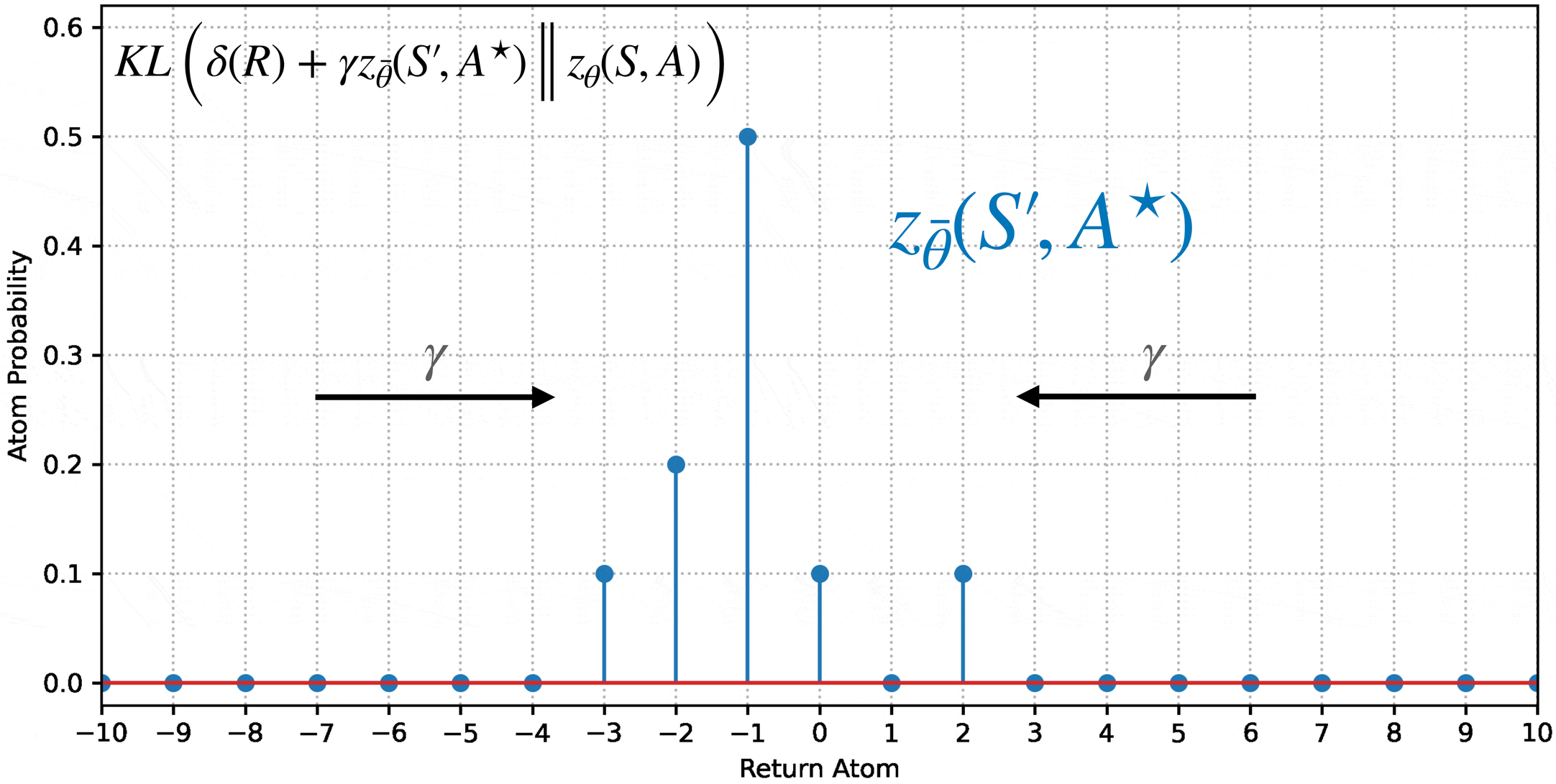
Distributional RL

C51 Algorithm

- C51 = DQN, but z_θ instead of q_θ; train z_θ by taking stochastic gradient descent steps on loss:
  - $KL\left(\delta(R) + \gamma z_{\bar{\theta}}(S', A^\star) \,\middle\|\, z_\theta(S, A)\right)$
  - with (S, A, R, S') from replay buffer and $A^\star = \arg\max_{a'} \mathbb{E}\left[z_{\bar{\theta}}(S', a')\right]$
    - Note: we have replaced $z_R(S, A)$ with point mass $\delta(R)$

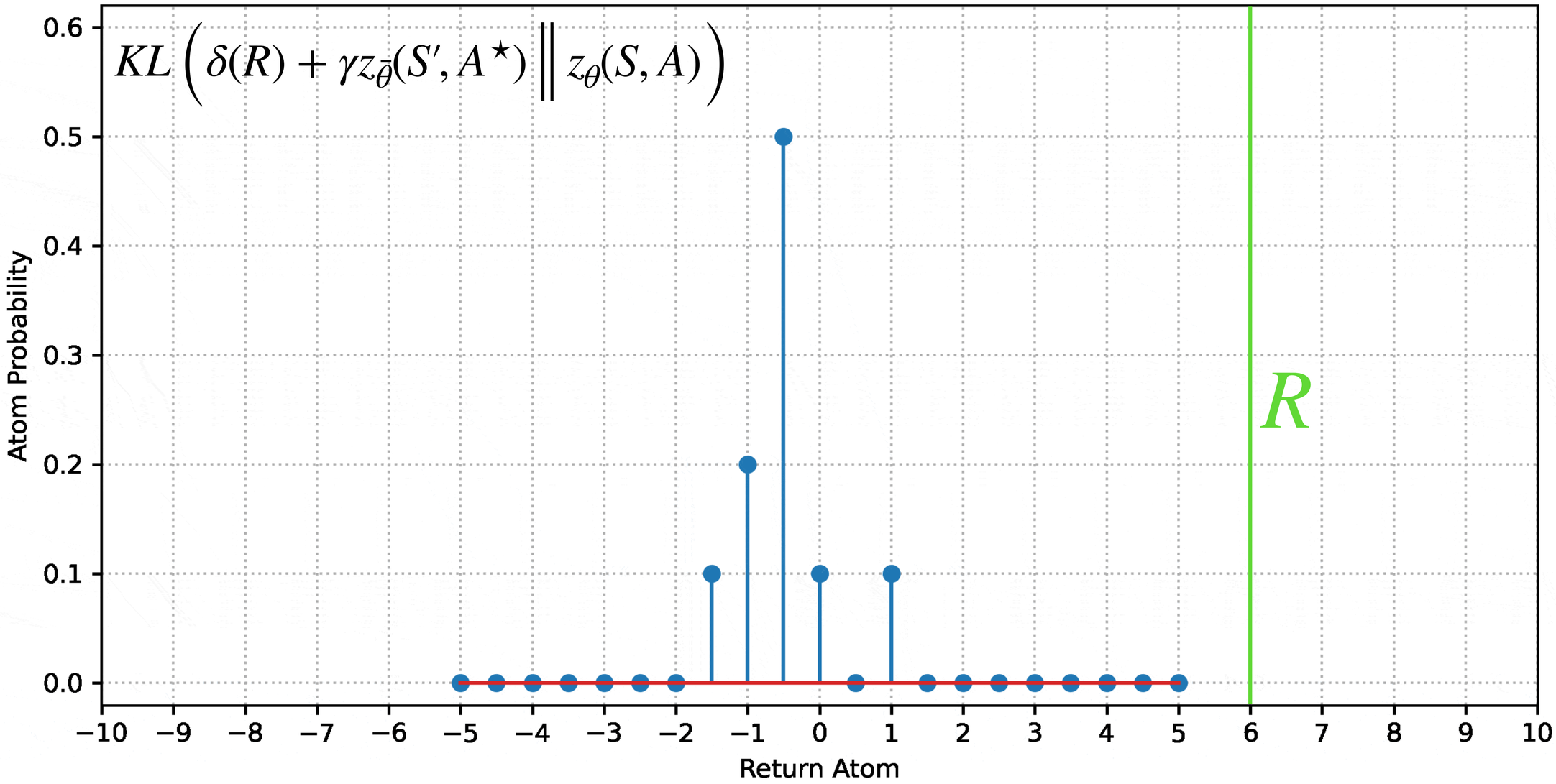
Distributional RL

C51 Algorithm



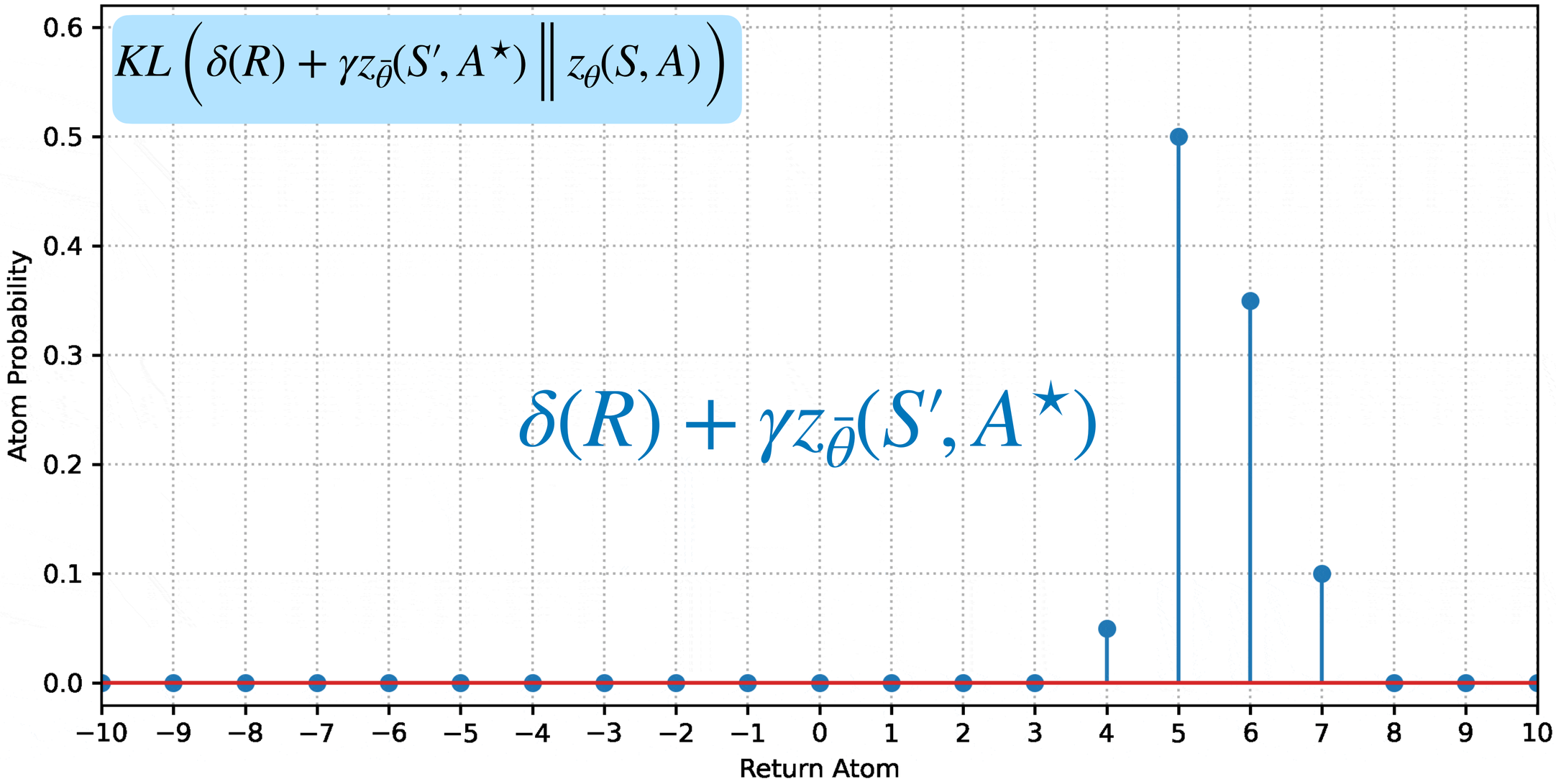
Distributional RL

C51 Algorithm



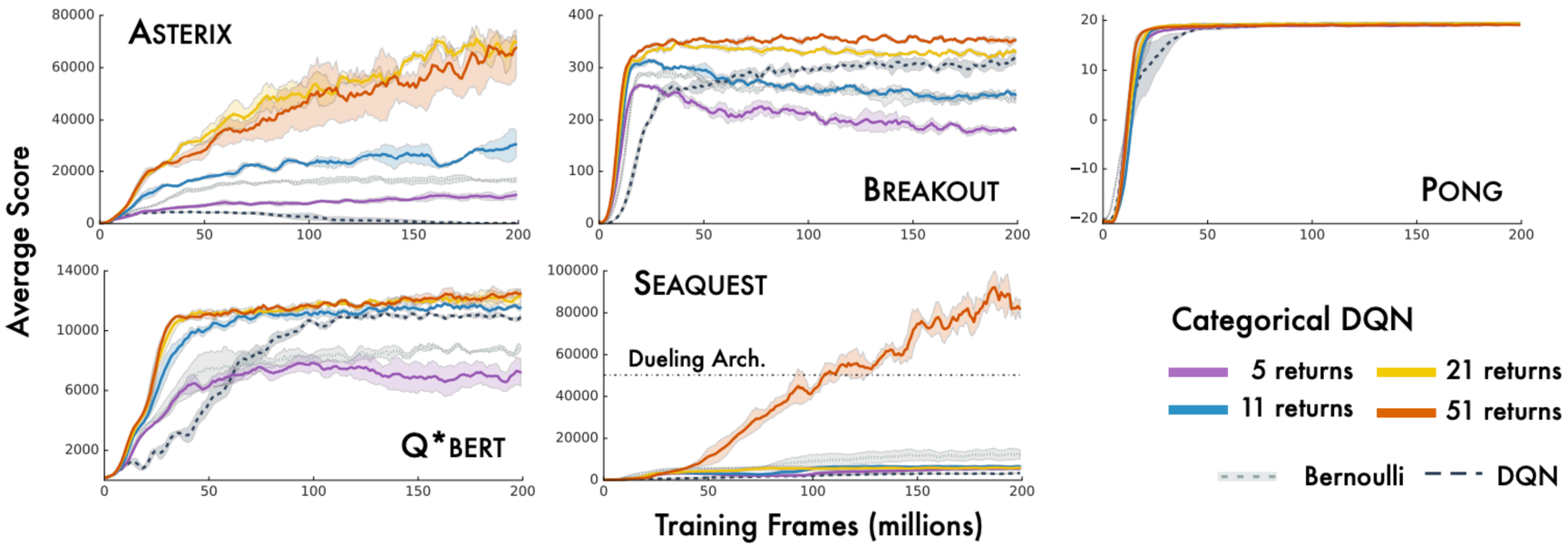
Distributional RL

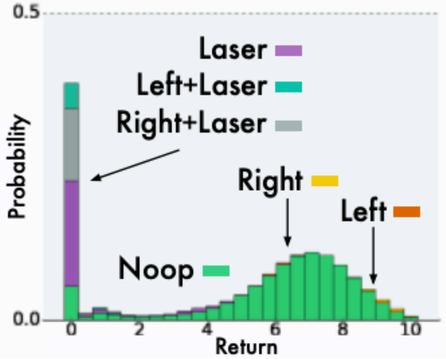
C51 Algorithm

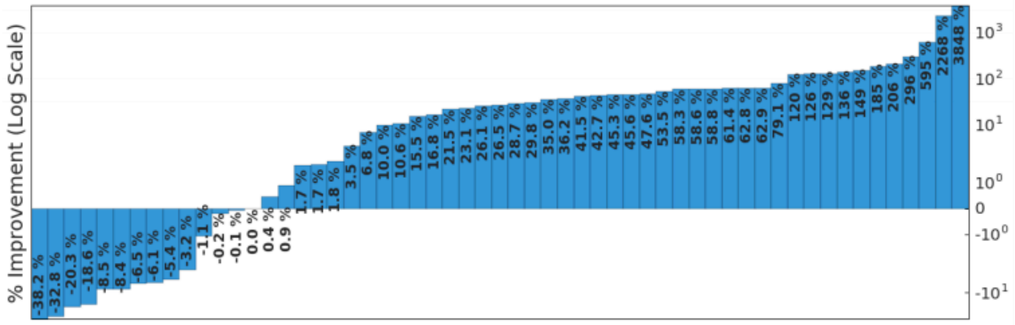


Distributional RL

C51 Algorithm







Distributional RL

Current Research Topics

- Still an active area of research to understand why learning distributions of values helps when all we care about is the expectation
  - Framework for inductive bias?
  - Better-behaved optimisation?
- How can we efficiently optimise the Wasserstein loss instead of the KL divergence?
  - Theory motivates the Wasserstein metric, but empirically, stochastic approximations of Wasserstein loss perform very poorly

Questions?

$$KL\left(\delta(R) + \gamma z_{\bar{\theta}}(S', A^\star) \,\middle\|\, z_\theta(S, A)\right)$$