

A Comparative Analysis of Expected and Distributional Reinforcement Learning

Clare Lyle, Pablo Samuel Castro, Marc G. Bellemare

Presented by,
Jerrod Parker and Shakti Kumar

Outline:

1. Motivation
2. Background
3. Proof Sequence
4. Experiments
5. Limitations

Outline:

1. Motivation
2. Background
3. Proof Sequence
4. Experiments
5. Limitations

Why Distributional RL?

1. Why restrict ourselves to the mean of value distributions?

i.e. Approximate Expectation v/s Approximate Distribution

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

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

Why Distributional RL?

1. Why restrict ourselves to the mean of value distributions?

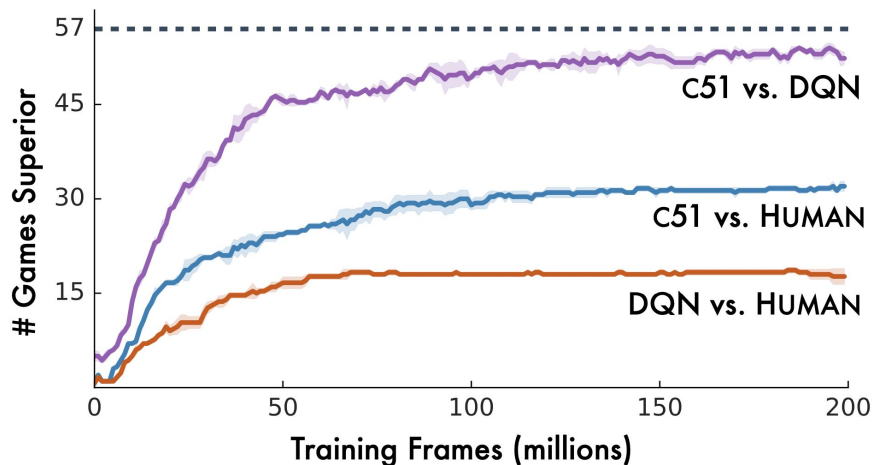
i.e. Approximate Expectation v/s Approximate Distribution

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

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

2. Approximation of multimodal returns?

Why Distributional RL?



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

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

Motivation

- Poor theoretical understanding of distributional RL framework
- Benefits have only been seen in Deep RL architectures and it is not known if simpler architectures have any advantage at all

Contributions

- Distributional RL different than Expected RL?

Contributions

- Distributional RL different than Expected RL?
 - Tabular setting 





Contributions

- Distributional RL different than Expected RL?
 - Tabular setting 
 - Tabular setting with categorical distribution approximator 





Contributions

- Distributional RL different than Expected RL?
 - Tabular setting ✖
 - Tabular setting with categorical distribution approximator ✖
 - Linear function approximation ✖

Contributions

- Distributional RL different than Expected RL?
 - Tabular setting 
 - Tabular setting with categorical distribution approximator 
 - Linear function approximation 
 - Nonlinear function approximation 

Contributions

- Distributional RL different than Expected RL?
 - Tabular setting 
 - Tabular setting with categorical distribution approximator 
 - Linear function approximation 
 - Nonlinear function approximation 
- Insights into nonlinear function approximators' interaction with distributional RL

Outline:

1. Motivation
2. Background
3. Proof Sequence
4. Experiments
5. Limitations

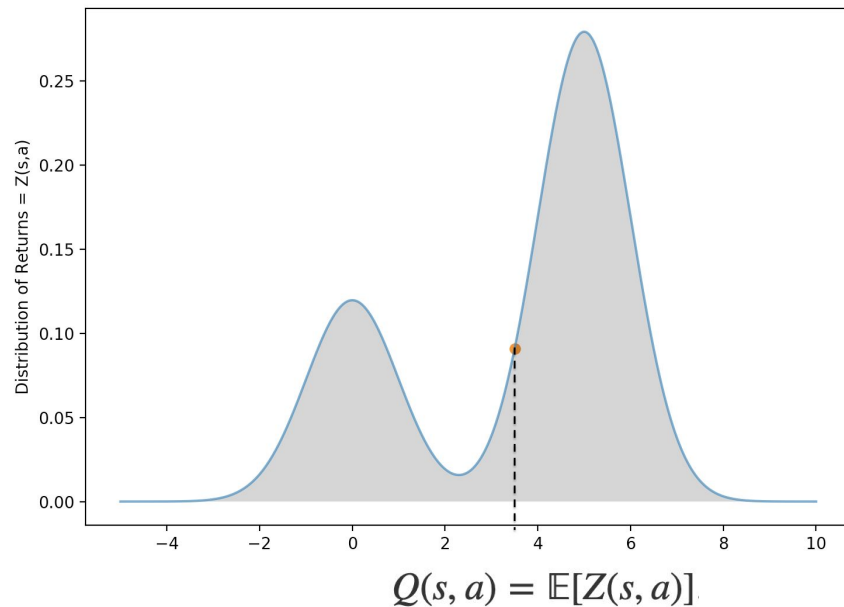
General Background– Formulation

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



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

X', A' are the random variables



General Background– Formulation

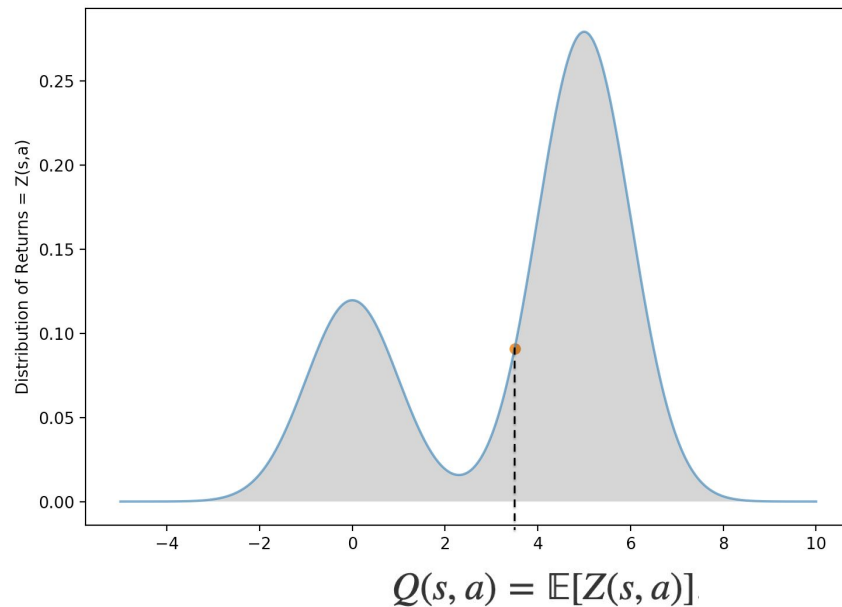
$$Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E} Q(X', A').$$



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

X', A' are the random variables

Sources of randomness in $Z(X', A')$?



General Background– Formulation

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

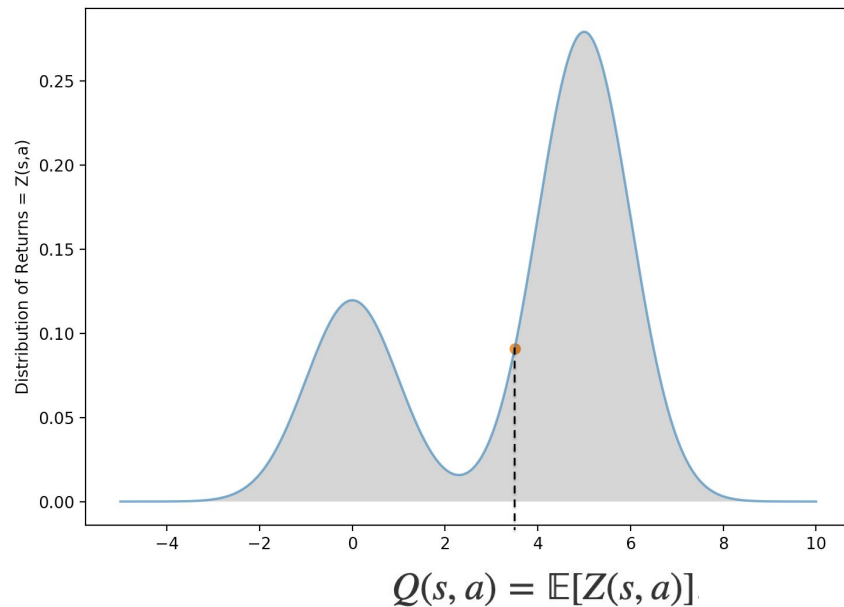


$$Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(X', A')$$

X', A' are the random variables

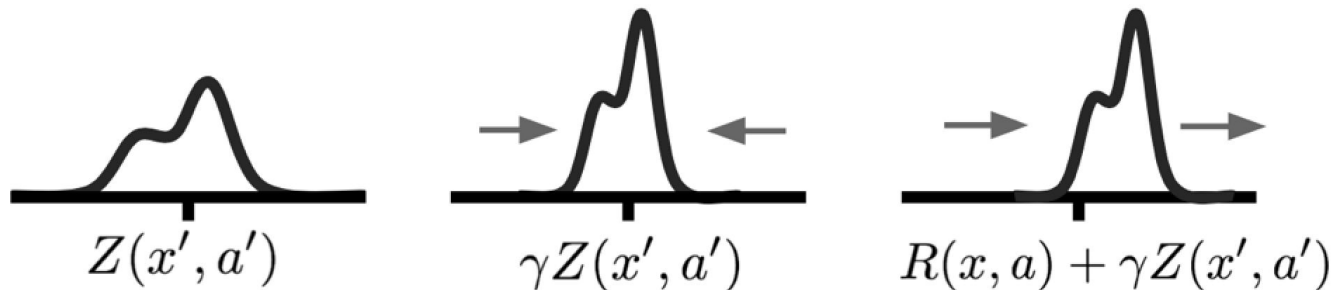
Sources of randomness in $Z(X', A')$?

1. Immediate rewards
2. Dynamics
3. Possibly stochastic policy



General Background– Visualization

$$Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E} Q(X', A') \longrightarrow Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(X', A')$$

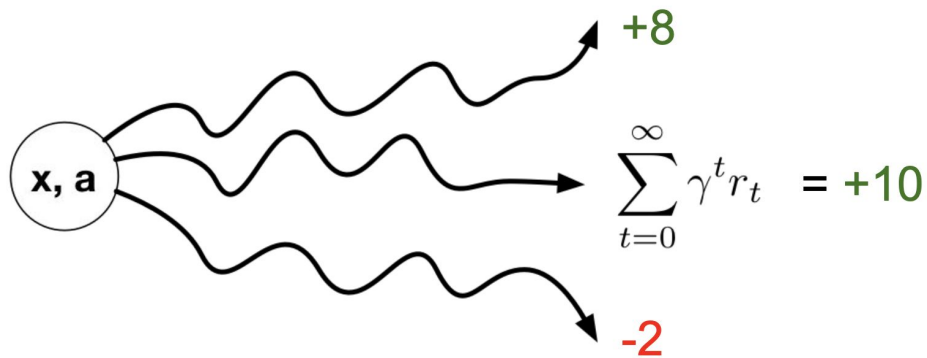


$x' \sim p(\cdot|x, a), a' \sim \pi(\cdot|x')$; $R(x, a)$ denotes the scalar reward obtained for $x \xrightarrow{a'} x'$ transition

General Background: Randomness

Source of randomness $Z^\pi(x, a)$

- Immediate rewards
- Stochastic dynamics
- Possibly stochastic policy



General Background– Contractions?

1. Is the policy evaluation step a contraction operation?
Can I believe that during policy evaluation my distribution is converging to the true return distribution?
2. Is contraction guaranteed in the control case, when I want to improve the current policy?
Can I believe that the Bellman optimality operator will lead me to the optimal policy?

Policy Evaluation Contracts?

Is the policy evaluation step a contraction operation?

Can I believe that during policy evaluation my distribution is converging to the true return distribution?

Formally— given a policy π do iterations $Z \leftarrow T^\pi Z$ converge to Z^π ?

Contraction in Policy Evaluation?

Given a policy π do iterations $Z \leftarrow T^\pi Z$ converge to Z^π ?

$$\begin{aligned} & d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)) \\ &= d_p(R(x, a) + \gamma P^\pi Z_1(x, a), R(x, a) + \gamma P^\pi Z_2(x, a)) \end{aligned}$$

Detour– Wasserstein Metric

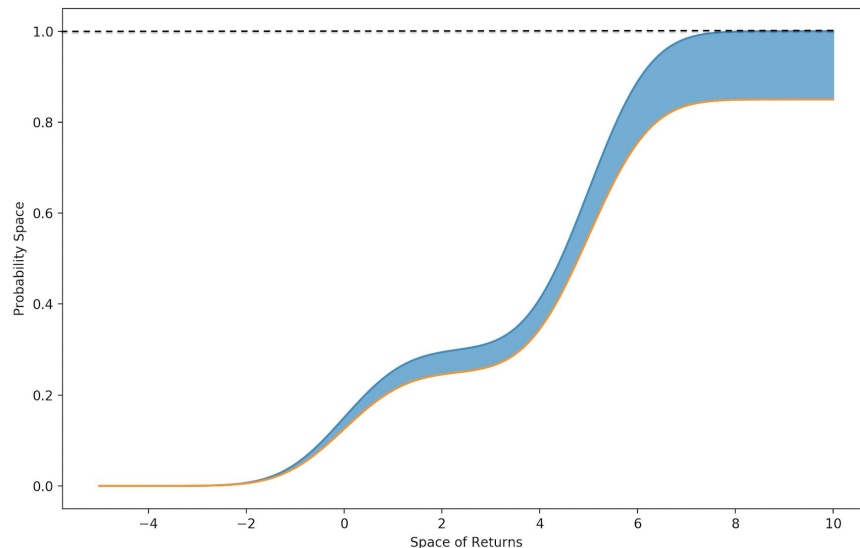
Defined as, $d_p(F, G) = \left(\int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p}$

where F^{-1} and G^{-1} are inverse CDF of F and G respectively

Maximal form of the Wasserstein,

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

Where $Z_1, Z_2 \in \mathcal{Z}$ and \mathcal{Z} denotes the space of value distributions with bounded moments



Contraction in Policy Evaluation?

Given a policy π do iterations $Z \leftarrow T^\pi Z$ converge to Z^π ?

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

$$[d_p(A + U, A + V) \leq d_p(U, V)]$$

$$[d_p(\gamma U, \gamma V) = \gamma d_p(U, V)]$$

Contraction in Policy Evaluation?

Given a policy π do iterations $Z \leftarrow T^\pi Z$ converge to Z^π ?

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

$$[d_p(A + U, A + V) \leq d_p(U, V)]$$

$$[d_p(\gamma U, \gamma V) = \gamma d_p(U, V)]$$

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

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

Contraction in Control/Improvement ?

First give a small background using definitions 1 and 2 from DPRL

Write the equation in the policy iteration of the attached image.

<give equations>

Unfortunately this cannot be guaranteed...

Distributional Bellman optimality operator

$$TZ(x, a) \stackrel{D}{=} r(x, a) + \gamma Z(x', \pi_Z(x'))$$

where $x' \sim p(\cdot|x, a)$ and $\pi_Z(x') = \arg \max_{a'} \mathbb{E}[Z(x', a')]$

Is this operator a contraction mapping?

No!

It's not even continuous



Give a similar equation for the policy evaluation also

Thus T^π has a unique fixed point, and it is Z^π

Policy evaluation:

For a given policy π , iterate $Z \leftarrow T^\pi Z$ converges to Z^π

Policy iteration:

- For current policy π_k , compute Z^{π_k}
- Improve policy

$$\pi_{k+1}(x) = \arg \max_a \mathbb{E}[Z^{\pi_k}(x, a)]$$

Does Z^{π_k} converge to the return distribution for the optimal policy?

General Background– Contractions?

1. Is the policy evaluation step a contraction operation?
Can I believe that during policy evaluation my distribution is converging to the true return distribution? ✓
2. Is contraction guaranteed in the control case, when I want to improve the current policy?
Can I believe that the Bellman optimality operator will lead me to the optimal policy?

Contraction in Policy Improvement?

Starting from a random policy π , do iterations $TZ(x, a) \leftarrow R(x, a) + \gamma Z(x', a')$
converge to Z^* such that $Z^* \leftarrow TZ^*$? $x' \sim p(\cdot|x, a), a' = \arg \max_a \pi(a|x')$
 π^* will be defined by Z^*

Contraction in Policy Improvement?

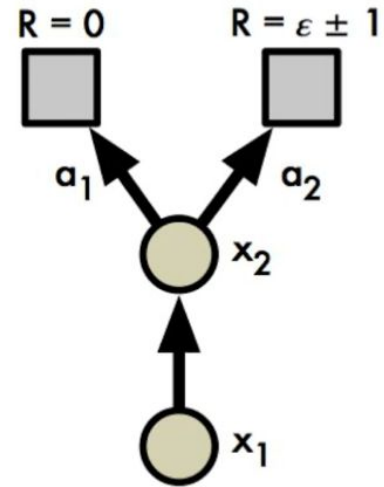
$x_1 \longrightarrow x_2$ transition

At x_2 two actions are possible

$r(a_1)=0$, $r(a_2) = \epsilon+1$ or $\epsilon-1$ with 0.5 probability

Assume a_1 , a_2 are terminal actions and the environment is undiscounted

What is the bellman update $TZ(x_2, a_2)$?



Contraction in Policy Improvement?

$x_1 \longrightarrow x_2$ transition

At x_2 two actions are possible

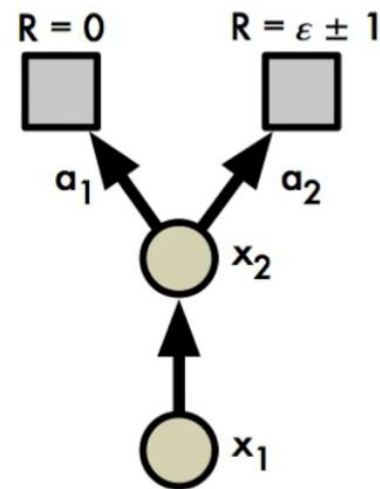
$r(a_1)=0$, $r(a_2) = \epsilon+1$ or $\epsilon-1$ with 0.5 probability

Assume a_1 , a_2 are terminal actions and the environment is undiscounted

What is the bellman update $TZ(x_2, a_2)$?

Since the actions are terminal, the backed up distribution should equal the rewards

Thus $TZ(x_2, a_2) = \epsilon \pm 1$ (or 2 diracs at $\epsilon+1$ and $\epsilon-1$)



Contraction in Policy Improvement?

Recall that if rewards are scalar, then bellman updates are older distributions Z just scaled and translated

Thus the original distribution $Z(x_2, a_2)$ can be considered as a translated version of $\mathcal{T}Z(x_2, a_2)$

Let $Z(x_2, a_2)$ be $-\epsilon \pm 1$

	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	$\epsilon \pm 1$

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

Contraction in Policy Improvement?

When we apply T to Z , then greedy action a_1 is selected,

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

	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	$\epsilon \pm 1$

$$\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) > \bar{d}_1(Z, Z^*)$$

This shows that the undiscounted update is not a contraction.

Thus a contraction cannot be guaranteed in the control case.

Contraction in Policy Improvement?

When we apply T to Z, then greedy action a_1 is selected, thus $\mathcal{T}Z(x_1) = Z(x_2, a_1)$

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

So is distributional RL a dead end?

This shows that the undiscounted update is not a contraction.

$$\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) > \bar{d}_1(Z, Z^*)$$

Thus a contraction cannot be guaranteed in the control case.

	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	$\epsilon \pm 1$

Contraction in Policy Improvement?

When we apply T to Z, then greedy action a_1 is selected, thus $\mathcal{T}Z(x_1) = Z(x_2, a_1)$

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

So is distributional RL a dead end?

	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	$\epsilon \pm 1$

This shows that the undiscounted update is not a contraction.

Bellemare showed that if there is a total ordering on the set of optimal policies, and the state space is finite, then there exists an optimal distribution which is the fixed point of the bellman update in the control case.

And the policy improvement converges to this fixed point [4]

Contraction in Policy Improvement?

So is distributional RL a dead end?

Bellemare showed that if there is a total ordering on the set of optimal policies, and the state space is finite, then there exists an optimal distribution which is the fixed point of the bellman update in the control case

Here Z^{**} is the set of value distributions corresponding to the set of optimal policies. This is a set of non stationary optimal value distributions

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

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

*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 $Z^* \in \mathcal{Z}^*$,*

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

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

The C51 Algorithm

Could have minimized Wasserstein metric between TZ and Z and hence learn an algorithm.

But learning cannot be done with samples in this case.

The expected sample Wasserstein distance between 2 distributions is always greater than the true Wasserstein distance between the 2 distributions.

So how do you develop an algorithm?

Instead project it on some finite supports, (which implicitly minimizes the Cramer distance between the original distribution thus still approximating the original distribution while keeping the expectation the same.)

Project what? Project the updates TZ.

So now we can see the entire algorithm!

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 ,*

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

and in general the inequality is strict and

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

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

$$\begin{aligned} d_p(P, Q) &= d_p(P_I, Q) \\ &= d_p\left(\sum_i A_i P_i, \sum_i A_i Q\right) \\ &\leq \sum_i d_p(A_i P_i, A_i Q) \\ &\leq \sum_i \Pr\{I = i\} d_p(P_i, Q) \\ &= \mathbb{E}_I d_p(P_i, Q). \end{aligned}$$

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.

The C51 Algorithm

Algorithm 1 Categorical Algorithm

input A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$

Compute the Q assuming you have diracs at each z_i

$$Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$$

This is same as a Cramer Projection which we'll see in the next slide

$$a^* \leftarrow \arg \max_a Q(x_{t+1}, a)$$

Take a greedy action

$$m_i = 0, \quad i \in 0, \dots, N-1$$

for $j \in 0, \dots, N-1$ **do**

Compute the projection of $\hat{T} z_j$ onto the support $\{z_i\}$

update the distribution (scale with γ and then translate with z_j)

$$\hat{T} z_j \leftarrow [r_t + \gamma z_j]_{V_{\min}}^{V_{\max}}$$

$$b_j \leftarrow (\hat{T} z_j - V_{\min}) / \Delta z \quad \# b_j \in [0, N-1]$$

find out the neighbour of Tz_j

$$l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$$

Distribute probability of $\hat{T} z_j$ on the found neighbours

$$m_l \leftarrow m_l + p_j(x_{t+1}, a^*) (u - b_j)$$

$$m_u \leftarrow m_u + p_j(x_{t+1}, a^*) (b_j - l)$$

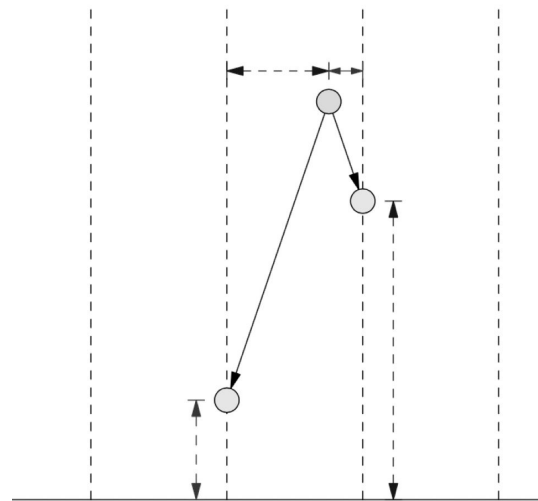
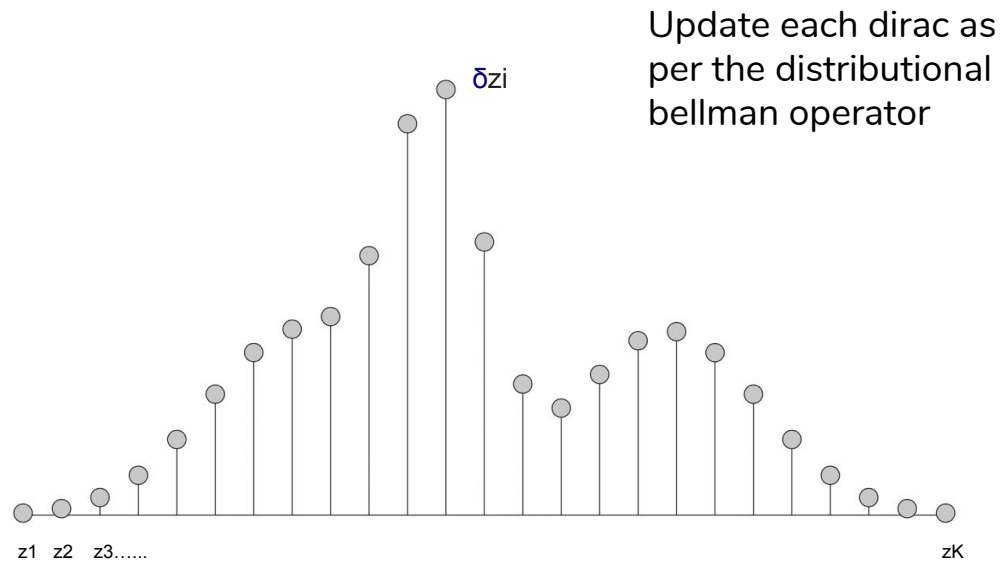
end for

Get the CE Loss

output $-\sum_i m_i \log p_i(x_t, a_t)$ # Cross-entropy loss

Backpropagate the loss in a DQN.
Repeat

C51 Visually



The distribute the mass of misaligned diracs on the supports

Cramèr Distance

- Gradient for the sample Wasserstein distance is biased

$$\left| \mathbb{E}_{\mathbf{x}_m \sim P} [\nabla_{\theta} w_p^p(\hat{P}_m, Q_{\theta})] - \nabla_{\theta} w_p^p(P, Q_{\theta}) \right| \geq 2e^{-2};$$

- For 2 given probability distributions with CDFs, F_P and F_Q , the cramer metric is defined as

$$\ell_2(P, Q) = \sqrt{\int_{\mathbb{R}} (F_P(x) - F_Q(x))^2 dx}$$

Cramèr Distance

- Attractive metric for distributional manipulations
 1. The policy evaluation bellman operator is a contraction in Cramer distance as well as shown by Rowland et. al. 2018
 2. A Cramer projection produces a distribution supported on z which minimizes the Cramer distance to the original distribution

If the support is contained in the interval $[z_1, z_k]$ then it's trivial to show that Cramer projection preserves the distribution expected value

Cramèr Distance

Now as we saw earlier, in distributional RL we need to approximate distributions

One way to do this is to formulate them as a categorical distribution like C51 did

Given some fixed set $\mathbf{z} = z_1, \dots, z_K \in \mathbb{R}^K$ with $z_1 \leq z_2 \leq \dots \leq z_K$, a categorical distribution P with support \mathbf{z} is a mixture of Dirac measures on each of the z_i 's, having the form

$$P \in \mathcal{Z}_{\mathbf{z}} := \left\{ \sum_{i=1}^K \alpha_i \delta_{z_i} : \alpha_i \geq 0, \sum_{i=1}^K \alpha_i = 1 \right\}.$$

Then the cramer distance is given as, $\ell_2(F_P, F_Q) = \sqrt{\sum_{i=1}^{K-1} (z_{i+1} - z_i)(F_P(z_i) - F_Q(z_i))^2}$

This is same as a weighted Euclidean norm between the CDFs of the 2 distributions.

When the atoms of the support are equally spaced apart, we get a scalar multiple of the Euclidean distance between the vectors of the CDFs

Outline:

1. Motivation
2. Background
3. **Proof Sequence**
4. Experiments
5. Limitations

Methods

- Compare policy evaluation in expected RL vs dist RL in several settings (ie tabular, linear approx, non linear approx)
- For each setting, the goal is to show expectation equivalence of expected version vs an analogous distributional version. Expectation equivalence:

$$Z \stackrel{\mathbb{E}}{=} Q \iff \mathbb{E}[Z(x, a)] = Q(x, a) \quad \forall (x, a) \in \mathcal{X} \times \mathcal{A}.$$

- Want to show: $Z_0 \stackrel{\mathbb{E}}{=} Q_0 \iff Z_t \stackrel{\mathbb{E}}{=} Q_t \quad \forall t \in \mathbb{N}.$
- Use same experience in both

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Constrain return distributions to being categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Constrain return distributions to being categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

Proposition 1: Cramer Projection

- If we have a categorical distribution which has support lying between $[z_1, z_k]$ where $z_1 < z_2 < \dots < z_k$, then Cramer project it onto the support z , then the expectation will remain.

Proposition 2: Tabular, Model-Based

$Z(s,a)$ and $Q(s,a)$ defined separately for each (s,a)

Expected bellman operator $T^\pi Q(x, a) := \mathbb{E}(R(x, a)) + \gamma \sum_{x', a'} P(x'|x, a)\pi(a'|x')Q(x', a')$

Distributional bellman $T_D^\pi Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(X', A')$ where $x' \sim p(\cdot|x, a)$, $a' \sim \pi(\cdot|x')$

Proposition 2:

Suppose $Z_0 \stackrel{\mathbb{E}}{=} Q_0$ then if $Z_{t+1} := T_D^\pi Z_t$, and $Q_{t+1} := T^\pi Q_t$, then $Z_t \stackrel{\mathbb{E}}{=} Q_t, \forall t$

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

Proof Proposition 2

Proof. By induction. By construction this is the case for Z_0, Q_0 . Suppose it holds for timestep t . Then for timestep $t + 1$, we have:

$$\begin{aligned}\mathbb{E}[Z_{t+1}(x, a)] &= \mathbb{E}[R(x, a) + Z_t(\mathbf{X}', A')] \\ &= \mathbb{E}[R(x, a)] + \gamma \sum_{x', a'} P(x'|x, a)\pi(a'|x')\mathbb{E}[Z_t(x', a')] \\ &= \mathbb{E}[R(x, a)] + \gamma \sum_{x', a'} P(x'|x, a)\pi(a'|x')Q_t(x', a') \\ &= Q_{t+1}(x, a)\end{aligned}$$

□

Tabular, Contains Model, Categorical Distributions

Suppose Z has finite support $z_1 < z_2 < \dots < z_k$ then applying:

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

can cause the resulting distribution to require a projection back to the support.

Proposition 3:

Suppose $Z_0 \stackrel{\mathbb{E}}{=} Q_0$, for $Z_0 \in \mathcal{Z}_z, Q_0 \in \mathcal{Q}$. If

$$Z_{t+1} := T_C^\pi Z_t \quad Q_{t+1} := T^\pi Q_t,$$

then also $Z_t \stackrel{\mathbb{E}}{=} Q_t \forall t \in \mathbb{N}$.

Proof follows from: $E[T_C^\pi Z_t] = E[T^\pi Z_t] = T^\pi Q_t$

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

SARSA vs Distributional SARSA (Arbitrary Distribution)

Given transition:

$$(x_t, a_t, r_t, x_{t+1}, a_{t+1}) \quad Z'_t(x_t, a_t) \stackrel{D}{=} r_t + \gamma Z_t(x_{t+1}, a_{t+1})$$

$$P_{Z_{t+1}}(x, a) := \begin{cases} (1 - \alpha_t)P_{Z_t}(x, a) + \alpha_t P_{Z'_t}(x_t, a_t) \\ P_{Z_t}(x, a) & \text{if } x, a \neq x_t, a_t \end{cases}$$

and the SARSA update

$$Q_{t+1}(x_t, a_t) := \begin{cases} Q_t(x_t, a_t) + \alpha_t \delta_t \\ Q_t(x, a) & \text{if } x, a \neq x_t, a_t \end{cases}$$

where $\delta_t := (r_t + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(x_t, a_t))$,

Proposition 4: These two policy evaluation methods have expectation equivalence.

Proof: SARSA vs Distributional SARSA

$$\mathbb{E}(Z_{t+1}(x_t, a_t)) = \sum_{i=1}^{k_t} P_{Z_{t+1}}(z_i) z_i$$

Expand $P_{Z_{t+1}}$

$$= \sum_{i=1}^{k_t} (1 - \alpha_t) P_{Z_t}(z_i) z_i + \alpha_t P_{Z'_t}(z_i) z_i$$

Notice similarities between exp SARSA and dist SARSA

$$= (1 - \alpha_t) \sum_{i=1}^{k_t} P_{Z_t}(z_i) z_i + \alpha_t \sum_{i=1}^{k_t} P_{Z'_t}(z_i) z_i$$

$$= (1 - \alpha_t) \mathbb{E}[Z_t(x_t, a_t)] + \alpha_t \mathbb{E}[r_t + \gamma Z_t(x_{t+1}, a_{t+1})]$$

$$= (1 - \alpha_t) Q_t(x_t, a_t) + \alpha_t [r_t + \gamma Q_t(x_{t+1}, a_{t+1})]$$

$$= Q_{t+1}(x_t, a_t)$$

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

SARSA vs Distributional SARSA (with Categorical Dist)

Recall: $Z'_t(x_t, a_t) \stackrel{D}{=} r_t + \gamma Z_t(x_{t+1}, a_{t+1})$

$$P_{Z_{t+1}}(x, a) := \begin{cases} (1 - \alpha_t)P_{Z_t}(x, a) + \alpha_t P_{Z'_t}(x_t, a_t) \\ P_{Z_t}(x, a) \end{cases} \quad \text{if } x, a \neq x_t, a_t$$

Difference:
Project onto support

and the SARSA update

$$Q_{t+1}(x_t, a_t) := \begin{cases} Q_t(x_t, a_t) + \alpha_t \delta_t \\ Q_t(x, a) \end{cases} \quad \text{if } x, a \neq x_t, a_t$$

where $\delta_t := (r_t + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(x_t, a_t))$, then also

Proof: SARSA vs Distributional SARSA (Categorical)

$$\begin{aligned}\mathbb{E}(Z_{t+1}(x_t, a_t)) &= \sum_{i=1}^{k_t} P_{Z_{t+1}}(z_i) z_i \\ &= \sum_{i=1}^{k_t} (1 - \alpha_t) P_{Z_t}(z_i) z_i + \alpha_t P_{Z'_t}(z_i) z_i \\ &= (1 - \alpha_t) \sum_{i=1}^{k_t} P_{Z_t}(z_i) z_i + \alpha_t \sum_{i=1}^{k_t} P_{Z'_t}(z_i) z_i \\ &= (1 - \alpha_t) \mathbb{E}[Z_t(x_t, a_t)] + \alpha_t \mathbb{E}[r_t + \gamma Z_t(x_{t+1}, a_{t+1})] \\ &= (1 - \alpha_t) Q_t(x_t, a_t) + \alpha_t [r_t + \gamma Q_t(x_{t+1}, a_{t+1})] \\ &= Q_{t+1}(x_t, a_t)\end{aligned}$$

Need to Cramer project this variable

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA
(doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

SARSA vs Semi-gradient of Cramer Distance:

Assume approximating distribution with categorical (c-spaced support). Gradient of squared Cramer w.r.t CDF:

$$\nabla_F \ell_2^2(Z, Z')[i] := \frac{\partial}{\partial F(z_i)} c \sum_{i=1}^k (F_{Z(x_t, a_t)}[i] - F_{\Pi_C Z(x_{t+1}, a_{t+1}) + r_t}[i])^2$$

Goal Proposition 6: Showing there is a semi gradient update which maintains expectation equivalence to SARSA (with a slight change in step size).

Results: Semi-gradient w.r.t CDF => Expectation Equivalence

Semi-gradient w.r.t PDF $\not\Rightarrow$ Expectation Equivalence

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

Linear Function Approximation

$$Q^\pi(x, a) \approx \theta^T \phi_{x,a} \quad F_{Z(x,a)}(z_i) \approx W \phi_{x,a}[i]$$

Loss Functions

$$\nabla_{\theta} (Q_t(x_t, a_t) - r(x_t, a_t) - Q_t(x_{t+1}, a_{t+1}))^2$$
$$\nabla_W \text{Cramer}(Z_t(x_t, a_t), r(x_t, a_t) + Z_t(x_{t+1}, a_{t+1}))^2$$

Update Rule

$$W_{t+1} := W_t - \alpha_t (W_t \phi_{x_t, a_t} - F_{Z'_t}(x_t, a_t)) \phi_{x_t, a_t}^T$$
$$\theta_{t+1} := \theta_t - \alpha_t (\theta_t^T \phi_{x_t, a_t} - r_t - \gamma \theta_t^T \phi_{x_{t+1}, a_{t+1}}) \phi_{x_t, a_t}^T$$

Proposition 8. Let $Z_0 \in \mathcal{Z}_\phi$, $Q_0 \in \mathcal{Q}_\phi$, and $\mathbf{z} \in \mathbb{R}^K$ such that \mathbf{z} is 1-spaced. Suppose that $Z_0 \stackrel{\mathbb{E}}{=} Q_0$, and that $Z_0(x, a)[z_K] = 1 \forall x, a$. Let W_t, θ_t respectively denote the weights corresponding to Z_t and Q_t . If Z_{t+1} is computed from the semi-gradient update rule

$$W_{t+1} := W_t + \alpha_t (F_{Z'_t}(x_t, a_t) - W_t \phi_{x_t, a_t}) \phi_{x_t, a_t}^T$$

and Q_{t+1} is computed according to Equation 2 with the same step-size α_t , then also $Z_t \stackrel{\mathbb{E}}{=} Q_t \forall t \in \mathbb{N}$.

Takeaway: If 1. Distributions add to 1
 2. Distance between bins in distribution is 1
 Then:
 Expectation equivalence holds

Theta update from last slide

Methods: Sequence of Proofs

1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Contains Model: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - b. Sample Based: (SARSA based updates, i.e. only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Approximate return distributions as categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

Nonlinear Function Approximation

Created example to show expectation equivalence doesn't always hold:

Let: $F_Z(x_t, a_t) = \sigma(W\phi(x_t, a_t))$

1. Start with expectation equivalence
2. Fix a transition such that the target and prediction have the same expectation but different distributions.

Nonlinear Function Approximation

Created example to show expectation equivalence doesn't always hold:

Let: $F_{Z(x_t, a_t)} = \sigma(W \phi(x_t, a_t))$

1. Start with expectation equivalence
2. Fix a transition such that the target and prediction have the same expectation but different distributions.

Recall Losses Used

$$\nabla_{\theta} (Q_t(x_t, a_t) - r(x_t, a_t) - Q_t(x_{t+1}, a_{t+1}))^2$$
$$\nabla_W \text{Cramer}(Z_t(x_t, a_t), r(x_t, a_t) + Z_t(x_{t+1}, a_{t+1}))^2$$

Nonlinear Function Approximation

Created example to show expectation equivalence doesn't always hold:

Let: $F_Z(x_t, a_t) = \sigma(W \phi(x_t, a_t))$

1. Start with expectation equivalence
2. Fix a transition such that the target and prediction have the same expectation but different distributions.
3. Now show that when we take a gradient step (using gradient of Cramer), the expectation of the predicted distribution changes but the Q-value didn't change expectation equivalence is broken.

Nonlinear Function Approximation

Takeaways:

- This doesn't prove that for all nonlinear functions that this happens.
- Gradient is taken w.r.t Cramer distance which isn't the case in many successful algorithms (Quantile Distributional RL for ex. minimizes Wasserstein).
- Expectation equivalence never breaks in the linear case which might mean that the benefits of distributional RL seen in practice could have to do with it's interplay with nonlinear function approximation.

Recap: Sequence of Proofs

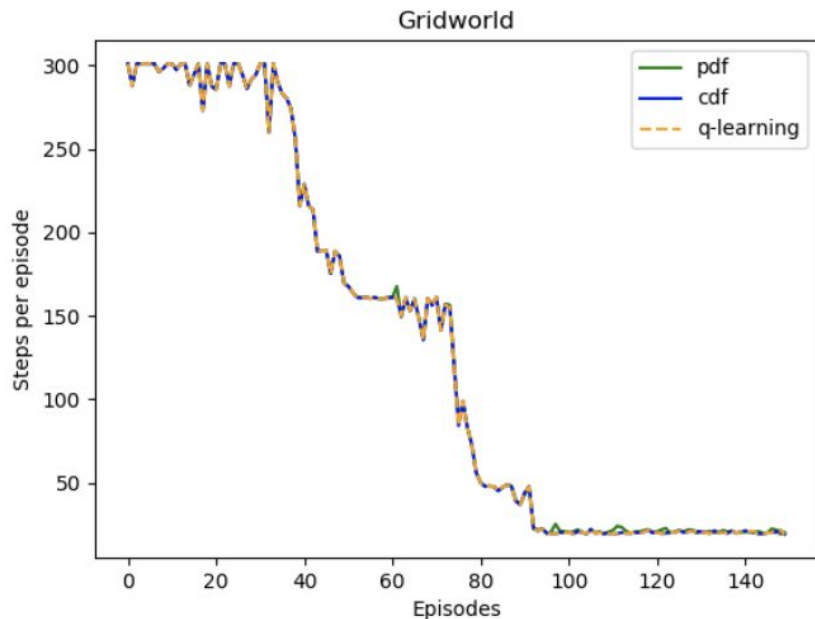
1. Tabular Models: Represent distribution over returns at each (s,a) separately
 - a. Model Based: (Have full knowledge of the transition model and policy)
 - i. No constraint on type of distribution to model returns
 - ii. Constrain return distributions to being categorical on fixed support
 - b. Sample Based: (SARSA based updates, ie only using samples)
 - i. No constraint on type of distribution to model returns
 - ii. Constrain return distributions to being categorical on fixed support
 - iii. Semi gradient w.r.t CDF update for distributional compared to SARSA
 - iv. Semi gradient w.r.t PDF update for distributional compared to SARSA (doesn't hold)
2. Linear Approximations:
 - a. Semi gradient of Cramer distance w.r.t CDF
3. Non linear Approximation:
 - a. There exists a non linear representation of the CDF such that initially we have equivalence but lose it after the first weight update.

Takeaways

1. In cases where they proved expectation equivalence, there isn't anything to gain from dist RL in terms of expected return. For ex:
 - a. Variance of our expected return is same in expected RL and distributional RL since $\text{Var}[E(Z)] = \text{Var}[Q]$
 - b. If using greedy methods, then policy improvement steps will be equivalent since expected value is same for each action.
2. Distributional RL and expected RL are usually expectation-equivalent for tabular representations and linear function approximation.
3. Expectation equivalence doesn't always hold when using non linear function approximation.

Experimental Results: Tabular Case (12x12 Grid)

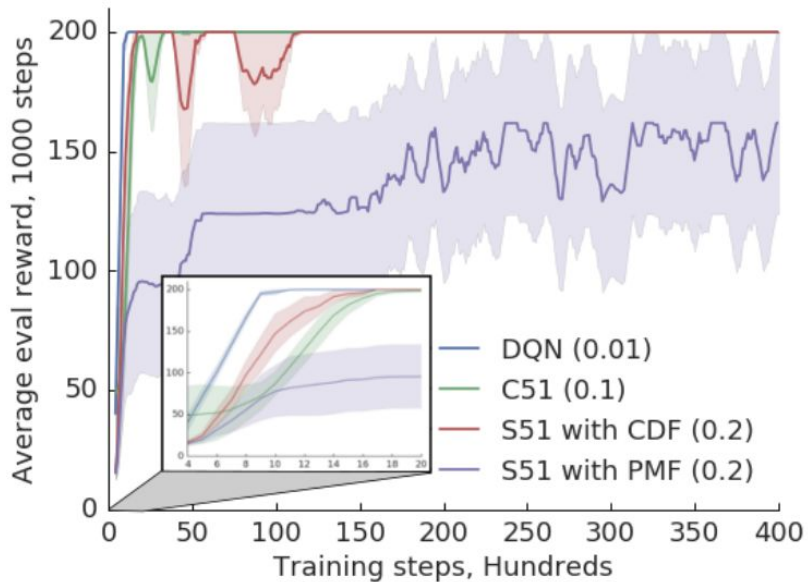
Compare: Q-learning, dist with CDF updates, dist with PDF updates. Using same random seed, eps-greedy actions: (so end with same results if expectation equiv)



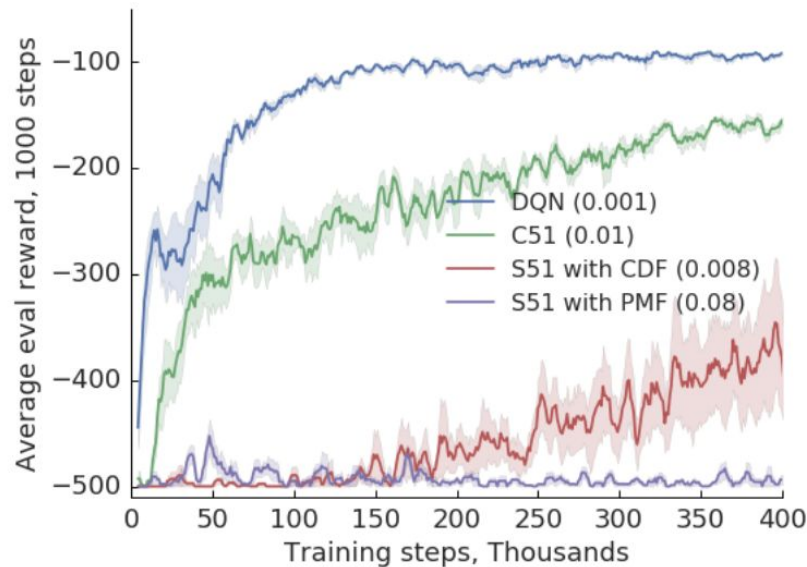
Outline:

1. Motivation
2. Background
3. Proof Sequence
- 4. Experiments**
5. Limitations

Experimental Results: Linear Approximation

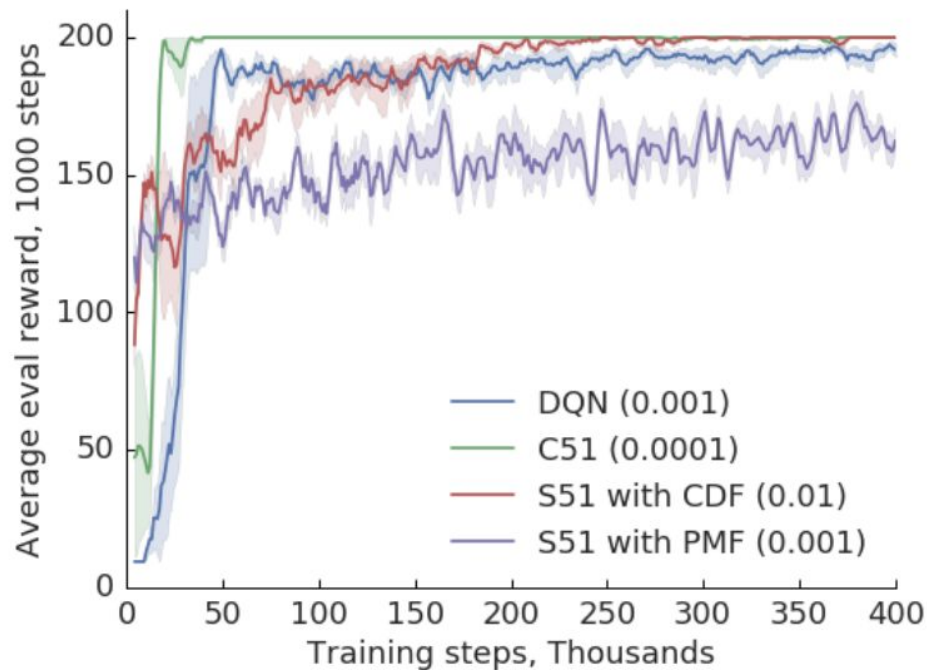


Cart Pole

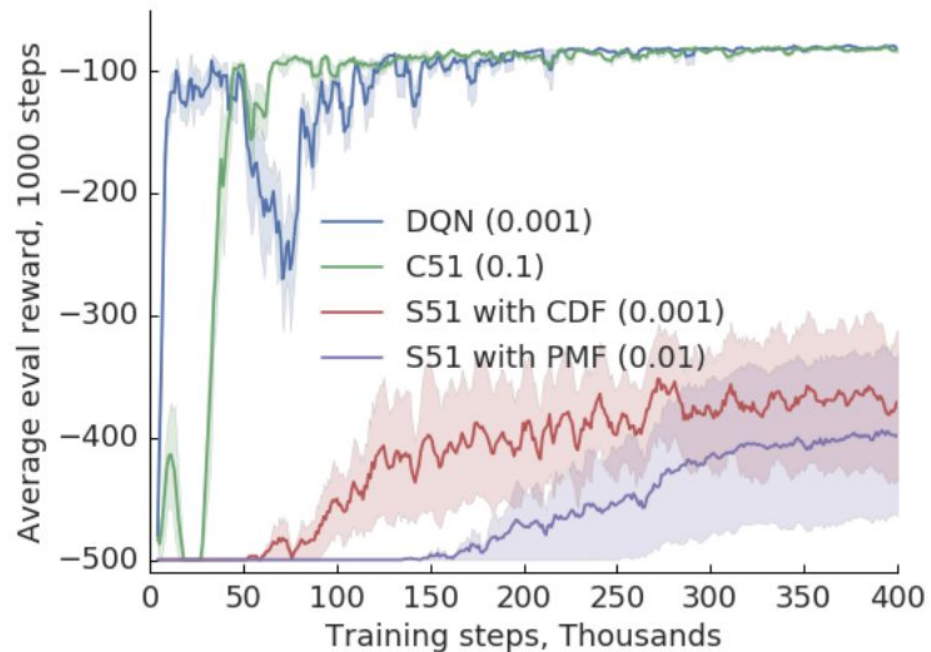


Acrobat

Experimental Results: Nonlinear Approximation



Cart Pole



Acrobat

Outline:

1. Motivation
2. Background
3. Proof Sequence
4. Experiments
5. Limitations

Limitations

- Their results all hold for minimizing Cramer distance but possibly not other metrics that are used in some successful distributional RL algorithms (Wasserstein, cross-entropy)
- The algorithm they use through their proofs doesn't seem to lead to quality results in practice
- Even though the authors prove that Cramer improves on Wasserstein limitations distributional RL [1], the empirical results don't convey this

Open Questions

1. What happens in deep neural networks that benefits most from the distributional perspective?
2. Is there a regularizing effect of modeling a distribution instead of expected value?

Questions

1. Derive: **Proposition 2.** *Let $Z_0 \in \mathcal{Z}$ and $Q_0 \in \mathcal{Q}$, and suppose that $Z_0 \stackrel{\mathbb{E}}{=} Q_0$. If*

$$Z_{t+1} := T_{\mathcal{D}}^{\pi} Z_t \quad Q_{t+1} := T^{\pi} Q_t,$$

then also $Z_t \stackrel{\mathbb{E}}{=} Q_t \forall t \in \mathbb{N}$.

2. What is one of the major benefits of the Cramer projection?
3. What are some possible reasons for the performance improvement of distributional RL over expected value RL when using non linear function approximation?

References

- [1] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer distance as a solution to biased Wasserstein gradients. In *arXiv preprint arXiv:1705.10743*, 2017.
- [2] Rowland, M.; Bellemare, M.; Dabney, W.; Munos, R.; and Teh, Y. W. 2018. An analysis of categorical distributional reinforcement learning. In Storkey, A., and Perez-Cruz, F., eds., Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, 29–37. Playa Blanca, Lanzarote, Canary Islands: PMLR.
- [3] Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A distributional perspective on reinforcement learning. In ICML.
- [4] Will Dabney, Mark Rowland, Marc G. Bellemare, and Remi Munos. Distributional Reinforcement Learning with Quantile Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018b.