1 Distributional RL

Typical Objective fn: Maximize expected discounted return.

In Value-function based,

Bellman Expectat^n Eq^n $\quad Q(s,a) = E[\underbrace{R(s,a) + \gamma \sum_{s'} Q(s',a')}_{\text{return}}]$

Bellman Optimality Eq^n $\quad Q(s,a) = E[R(s,a) + \gamma \max_{a'} Q(s',a')]$

$$= E[R(s,a)] + \gamma E[\max_{a'} Q(s',a')]$$

In Distributed RL, $\quad Q(s,a) = E[Z(s,a)]$

↳ return, random variable from some dist.

$$= E[R(s,a) + \gamma Z(s',a')]$$

We want to learn this!

2 Algorithm (C51)

a Objective functions for learning

In Value based, tabular: TD error

In Approx setting: MSE

In value-dist, propose W. dist. BUT

In practice, KL divergence used instead of W distance.

b Defining $Z(s,a)$ (N=51 in C51)

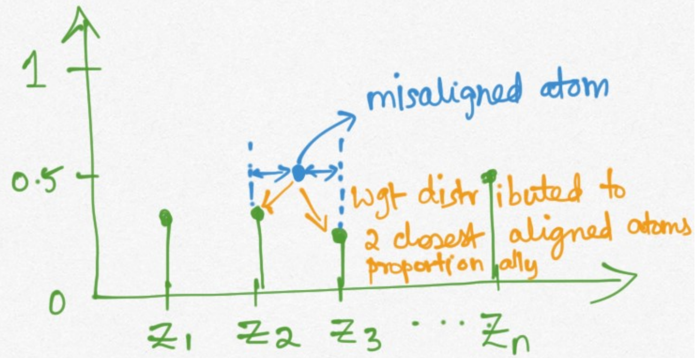
Z defined as discrete dist of N fixed atoms (support)

$$Z_\theta(s,a) = z_i \quad \text{w.p.} \quad p_i = \frac{e^{\theta_i(s,a)}}{\sum_j e^{\theta_j(s,a)}}$$

(softmax)

z_i's bounded in $[V_{min}, V_{max}]$

$$z_i = V_{min} + i\Delta z \qquad \Delta z = \frac{V_{max} - V_{min}}{N-1}$$



c Projection step

Estimated Z: $z_i(s,a)$ $\quad$ Target: $R_i + \gamma z_i(s',a')$

Supports' disjoint, cannot compute KL loss. (goes to ∞)

⇒ Projection: Transfer prob. mass from misaligned target atoms to closest neighboring estimate 'aligned' atoms.

Then compute KL loss. $KL(p \| q)$ b/w target p and estimate q.

Minimize loss using gradient descent.

③ Wasserstein Distance

Dist b/w 2 prob distributions.

$$W(Z_1, Z_2) \equiv \inf_{J \in \Pi(Z_1, Z_2)} \underset{(x,y)\sim J}{E} \|x-y\|$$

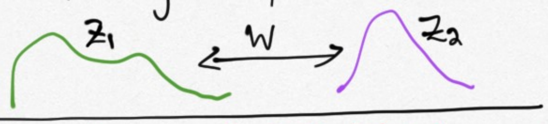
min

Π - set of transport plans to move "earth" from one dist to other

J - transport plan w/ min. cost/dist

(x,y) sampled from J

$\|x-y\|$ - dist b/w x & y. Can also be L_p norm to give p-Wasserstein dist



Primal is intractable, joint dist. We can use dual. Dual also has convergence guarantees in policy evaluation case.

$$W(Z_1, Z_2) = \sup_{f \in \mathcal{F}_{Lip}} \underset{x\sim Z_1}{E}[f(x)] - \underset{y\sim Z_2}{E}[f(y)]$$

Or written as: (HOW?)

$$\bar{W}(Z_1, Z_2) = \sup_{s,a} W(Z_1(s,a), Z_2(s,a))$$

This dual form can be proved to be γ-contraction in W in policy eval case. In control, cannot guarantee.

⑤ Why should this work?

In value function based,

a) Bellman operator T is γ-contraction mapping

$$\|TF - TG\|_\infty \leq \gamma \|F - G\|_\infty$$

(has convergence properties $F^* = TF^*$, $F_t = TF_{t-1}$) in tabular setting. Converges to fixed point.

In value distribut^n, we can use Wasserstein dist; dual form is γ-contraction mapping (PROOF). Only in policy evaluation case.

Banach's fixed pt theorem: $d(T(F), T(G)) \leq \gamma d(F, G)$

(T is contraction mapping in d) d - some distance measure

So, also convergence to fixed point.

[But authors claim W dist cannot be used w/ samples. Hence KL divergence used instead.

But, dual of W dist is in terms of Expectat^n, which means we can use samples to estimate, & same measure is used in WGANs.]