① Distributional RL

Typical objective fn: Maximize expected discounted return.

In Value-function based,

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

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

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

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

↳ return, random variable from some dist.

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

We want to learn this!

② Algorithm (C51)

ⓐ Objective functions for learning

In Value based, tabular: TD error

In Approx setting: MSE

In value-dist, propose W. distance BUT

In practice, KL divergence used instead of W distance.

ⓑ Defining $Z(s,a)$ (N=51 in C51)

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

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

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

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

ⓒ Projection step

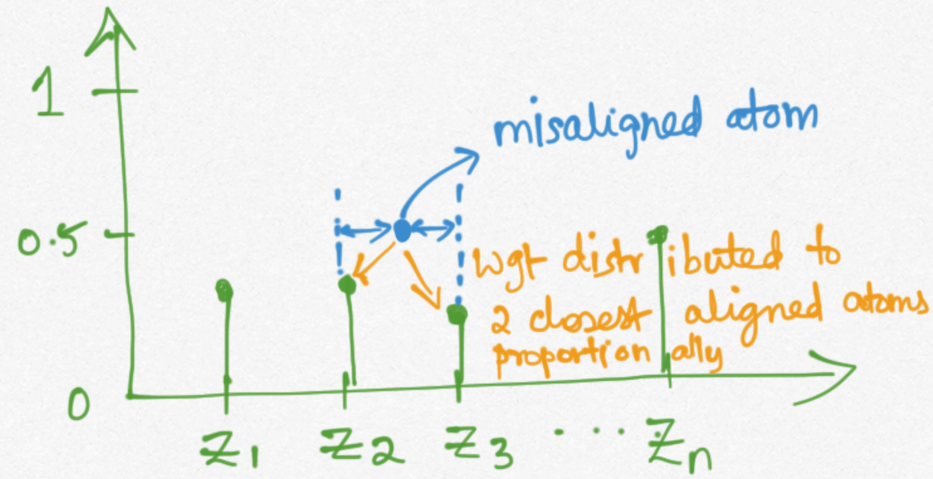
Estimated Z: $Z_i(s,a)$ Target: $R_i + \gamma Z_i(s',a')$

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.

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

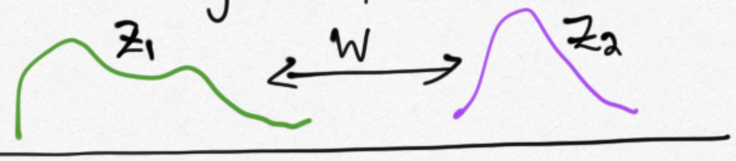
(inf: min)

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

J - transport plan w/ min. cost/distance

(x, y) sampled from J

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



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

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

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", we can use Wasserstein distance, 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 athors claim W distance cannot be used w/ samples. Hence KL divergence used instead.

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