Implicit Quantile Networks (IQN)

- Prev work (QR-DQN) learnt a distribut^n over returns but still used mean returns for policy.
- IQN learns a risk sensitive policy.

Risk-Sensitive RL - I

- Instead of max E[Returns], max E[Utility].
- E.g.s of Utility : Exp/Log (Returns), Measure that includes variance etc.
- In general, utility f^n satisfies 4 axioms
  - ① Monotonicity : lower cost ⇒ lower risk
  - ② Sub-additivity : Diversification ⇒ lower risk
  - ...
- Can be concave (risk averse) / convex (seeking) or both.

Risk-Sensitive RL - II

- One more axiom of utility f^n is of independence:
  If $X \succ$ (preferred over) Y, then $pX + (1-p)Z > pY + (1-p)Z$
  (Any mixture of X & Z preferred over same mixture over Y & Z).

We can think of utility as mixture of distributions, which we want to learn, to learn a risk-sensitive policy.

(As before, we want to use quantiles to learn distribut^n)

Learning mixture of distributions (MoD)

- MoD can be represented in terms of cdf.
- cdf of MoD is convex combination of its individual distribut^{ns}.

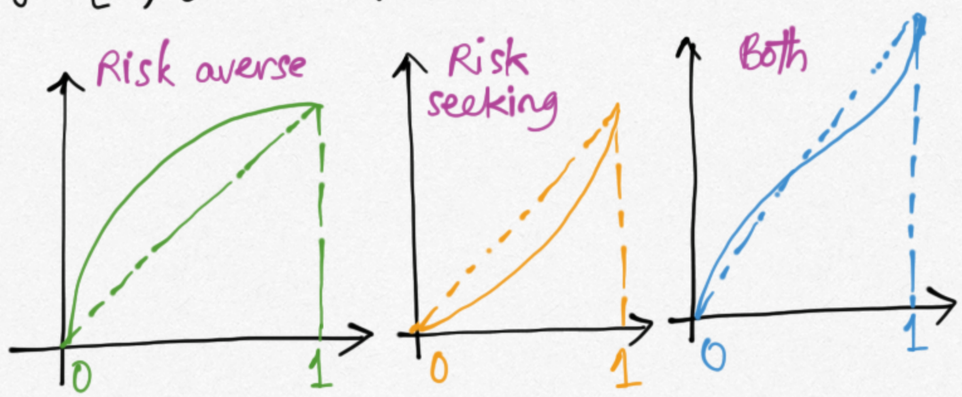
$$F(x) = \sum_i p_i F_i(x)$$

- MoD is also represented as convex comb. of quantiles (in insurance/reinsurance ...) where weighting is some risk f^n. (e.g. VaR)

Distortion Risk Measures (DRM)

- VaR, Conditional VaR are DRMs.
- In General, DRM is a continuous non-decreasing f^n.

$g : [0,1] \to [0,1]$ s.t $g(0)=0$, $g(1)=1$.



Distortion f's.

Dual theory of Choice under Risk

- An alternate objective for risk sensitive policy:
  Instead of maximizing expected utility,
  maximize expectation of distortion risk measure
  $\Rightarrow$ i.e. Distorted Expectation.

Learning a Distorted Expectation

Say $Z(s,a)$ is our usual returns distribut^n.
For some quantile τ, $E[Z_\tau(s,a)]$ is expected return for that quantile.
$\tau \in U(0,1)$

We can reweight this expectat^n w/ β (DRM),

$$Q_\beta(s,a) := E[Z_{\beta(\tau)}(s,a)]$$

[Can think of this as Inverse Transform Theorem: To generate samples from some distribut^n, sample from Unif(0,1), & transform them through inverse CDF (quantile) of the prob. distribut^n.]

$Q_\beta(s,a)$ can be used to follow policy (action-select^n)

$$\pi_\beta(s) = \underset{a \in A}{\text{argmax}}\ Q_\beta(s,a)$$

IQN Loss function

- Similar to QR-DQN we want to minimize loss b/w pairs of quantiles. But we don't have a fixed set of quantiles as in QR-DQN
- Instead, parametric f^n $\phi(\tau)$ used to represent quantiles (embedding layer in NN).
- So we can generate samples, N for τ, N' for τ' & calc. loss.

$$L = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho^k_{\tau_i}(\delta^{\tau_i, \tau_j'})$$

$\delta^{\tau,\tau'}$ - TD error

- Similarly to select actions, use K samples of $\tilde{\tau}$ & take argmax of mean $Q_{\beta(\tilde{\tau})}(s,a)$
- How to generate these samples? Sample from $U(0,1)$ & reweight using the parametric f^n.