CS7011 : Topics in Reinforcement Learning
Summary of Implicit Quantile Networks : Dabney, Ostrovski, Silver, Munos
By Sahana Ramnath : EE15B109

1. Abstract

(Dabney et al., 2018) proposes a novel distributional variant of the DQN : *Implicit Quantile Networks*(IQN) which uses quantile regression to approximate the full quantile function/c.d.f for the state-action return distribution. Experiments have been conducted on Atari games. By reparametrizing a distribution over the sample space, this yields an implicitly defined return distribution that gives rise to a large class of risk-sensitive policies, the effects of which were studied in Atari games.

2. Distributional RL : Recent Advances

In distributional RL, the distribution over returns Z^π is modelled rather than its expected value function Q^π. While the distributional Bellman operator for policy evaluation is a contraction in p-Wasserstein distance, it is not a contraction in any metric for control.

Any distributional RL algorithm is characterized by the parametrization of the return distribution and the distance metric or loss function being minimized. C51/Categorical DQN (Bellemare et al., 2017) combines a categorical distribution and the cross-entropy loss with the Cramer-minimizing projection;this assumes returns are bounded in a known range and trades off mean-preservation at the cost of overestimating variance. C51 and subsequent papers are restricted to assigning probabilities to a fixed, discrete set of possible returns. QR-DQN (Dabney et al., 2017) proposes parametrizing the distribution by a uniform mixture of Diracs whose locations are adjusted using quantile regression; though it is restricted to a discrete set of quantiles, it automatically adapts return quantiles to minimize the Wasserstein distance between the Bellman updated and current return distributions. This flexibility allows the QR-DQN to significantly improve on C51's performance.

2.1. Extending to IQN

This paper extends the approach of QR-DQN from learning a discrete set of quantiles to learning the full quantile function, a continuous map from probabilities to returns. When combined with a base distribution, such as $U([0, 1])$, this forms an implicit distribution capable of approximating any distribution over returns given sufficient network capacity. Advantages :

- The representative power for the distribution is now controlled by the size of the network and the amount of training and not by the number of quantiles fixed apriori. In this way, IQNs are a type of UVFA.
- IQN has improved data efficiency.
- By taking the base distribution to be non-uniform, the return distribution can be expanded to ϵ-greedy policies on arbitrary distortion risk measures.

(Bellemare et al., 2017) showed that the distributional Bellman operator is a contraction in the p-Wasserstein metric, but as the proposed algorithm C51 did not itself minimize this metric, this left a theory-practice gap in distributional RL. By estimating the quantile function at precisely chosen points, QR-DQN minimizes the Wasserstein distance to the distributional Bellman target. IQN on the other hand uses quantile regression, which has been shown to converge to the true quantile function value when minimized using stochastic approximation.

2.2. Risk Sensitivity in RL

All previous work in distributional RL are based on maximizing the mean of the estimated distribution just like in standard RL; moving away from this leads to *risk-sensitivity* in policies. In general, there are two notions of risk-sensitivity. One is by utility theory where the convexity of the utility functions matters(linear is risk-neutral, concave/convex is risk averse/seeking respectively). Alternately, we can compute a reweighting of the distribution under the distortion risk measure as it distorts the cumulative probabilities of the random variable.

3. The Algorithm

IQN is a deterministic parametric function trained to reparameterize samples from a base distribution, e.g. $\tau \sim U([0, 1])$, to the respective quantile values of a target distribution. $Z_\tau = F_Z^{-1}(\tau)$ represents the samples from the implicitly defined return quantile function at any τ. Let $\beta : [0, 1] \rightarrow [0, 1]$ be the risk distortion measure with identity corresponding to risk neutrality. The distorted expectation of $Z(x, a)$ under β is given by $\mathrm{E}_{\tau \sim U([0,1])}[Z_{\beta(\tau)}(x, a)]$. $\pi_\beta(x)$ represents the risk-sensitive greedy policy $argmax_{a \in \mathcal{A}} Q_\beta(x, a)$. For 2 samples $\tau, \tau' \sim U([0, 1])$ and policy π_β the sampled TD error at time t : $\delta_t^{\tau,\tau'} = r_t + \gamma Z_{\tau'}(x_{t+1}, \pi_\beta(x_{t+1})) - Z_\tau(x_t, a_t)$ and the IQN loss is $\mathcal{L}(x_t, a_t, r_t, x_{t+1}) = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa}(\delta_t^{\tau_i,\tau_j'})$ where N, N' denote the number of iid samples τ_i, τ_j' used to estimate loss. A corresponding sample-based risk-sensitive policy is obtained by approximating Q_β above with K samples of $\tilde{\tau} \sim U([0, 1])$: $\tilde{\pi_\beta}(x) = argmax_{a \in \mathcal{A}} \frac{1}{K} \sum_{k=1}^{K} Z_{\beta(\tilde{\tau_k})}(x, a)$

4. Experiments and Results

The basic architecture of DQN(f : convolutional and ψ : fully-connected layers) is used with an additional function $\phi_j(\tau) := ReLU(\sum_{i=0}^{n-1} cos(\pi i \tau) w_{ij} + b_j)$ computing an embedding from the sample point τ. IQN uses $\psi \odot \phi$ to force interaction between the convolutional features and the sample embedding. They hypothesized that N, the number of samples τ would affect the sample complexity, with $N = 1$ being a DQN and higher N being better; this was confirmed in experiments. They also hypothesized N', the number of samples τ' would affect the variance similar to the minibatch size; however it behaved differently; it had a strong effect on early performance, but minimal impact on long-term performance past $N' = 8$.

They also studied the effect of varying β away from identity to evaluate different risk-sensitive policies. They tried cumulative probability weighting(CPW) parameterization from cumulative utility theory, Norm distortion risk(Wang, 2000) and conditional value-at-risk (CVaR). Norm(3) and CPW(.71) reduce the impact of the tails of the distribution, while Wang and CVaR heavily shift the distribution mass towards the tails, creating a risk-averse or risk-seeking preference. They found that risk-averse policies improve performance over the IQN(maybe because risk-aversion en-

codes a heuristic to stay alive longer which in many games means increased rewards). CPW performs almost identically to the standard risk-neutral policy, and the risk-seeking Wang(1.5) performs as well or worse than risk-neutral.
IQN outperforms C51 and QR-DQN which themselves outperform all previous results. However there it still underperforms Rainbow-DQN though it halved the distance between QR-DQN and Rainbow; this is because Rainbow combines six strong variants of the DQN. A corresponding Rainbow-IQN is left as future work.

References

Bellemare, Marc G, Dabney, Will, and Munos, Rémi. A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*, 2017.

Dabney, Will, Rowland, Mark, Bellemare, Marc G, and Munos, Rémi. Distributional reinforcement learning with quantile regression. *arXiv preprint arXiv:1710.10044*, 2017.

Dabney, Will, Ostrovski, Georg, Silver, David, and Munos, Rémi. Implicit quantile networks for distributional reinforcement learning. *arXiv preprint arXiv:1806.06923*, 2018.

Wang, Shaun S. A class of distortion operators for pricing financial and insurance risks. *Journal of risk and insurance*, pp. 15–36, 2000.