Human expert   On-policy RL agent   Database

Data collection

Training

Learn action-value distribution

$Z(x,a; \tau)$

Reward

Gradient

CRITIC

τ

$x_B \ a_B \ x'_B \ r_B$

TRAIN VIA TD-LEARNING:
$$L_{QR}\left[Z(x'_B, a'\,; \tau'), Z(x_B, a_B\,; \tau)\right]$$
$$\tau, \tau' \sim U([0,1])$$

Learn CVaR maximization policy

$Z(x,a; \tau)$

a

Gradient

Reward

$\text{CVaR}_\alpha \propto Z(x, a, \hat{\tau}) \quad \hat{\tau} \sim U([0,\alpha])$

TRAIN VIA DETERMINISTIC POLICY GRADIENT:
$$\nabla_{\text{ACTOR}}\text{CVaR}_\alpha$$

ACTOR

a_{VAE}

x_B

Learn action generative model

$a_{VAE} \leftarrow$ VAE

$x_B \ a_B$

Gradient