Human expert On-policy RL agent

Database

Data collection

Training

Z(x,a)

Return

CRITIC

TRAIN VIA TD-LEARNING:

$$L_{QR}\big[Z(x'_B, a'\,;\tau'), Z(x_B, a_B\,;\tau)\big]$$
$$\tau, \tau' \sim U([0,1])$$

Gradient

τ

$x_B\ a_B\ x'_B\ r_B$

Learn action-value distribution

a

Z(x,a)

Return

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

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

Gradient

ACTOR

a_{VAE}

x_B

a_{VAE} VAE $x_B\ a_B$

Gradient

Learn action generative model

Learn CVaR maximization policy