Human expert — On-policy RL agent — Database

Data collection

Training

Z(x,a; τ)

Reward

CRITIC

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

τ

$x_B\ a_B\ x'_B\ r_B$

Learn action-value distribution

a

Z(x,a; τ)

Reward

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

TRAIN VIA DETERMINISTIC POLICY GRADIENT:
$$\nabla_{ACTOR} CVaR_\alpha$$

ACTOR

a_{VAE}

x_B

Gradient

VAE

a_{VAE} $x_B\ a_B$

Learn action generative model

Learn CVaR maximization policy