PD-DDPGfD Agent

Batch $(s, a, c_1, c_2, c_3, c_4, \acute{s})$

Batch $(s, \acute{s})$

Batch $(s, a, r, \acute{s})$

EH Environment

Dispatch action $a = \mu(s | \theta^\mu)$

Cost Network(4)

...

Cost Network(1)

Loss function $L(\psi_1)$

Update C_{ψ_1}

Online Cost Network 1 $C_{\psi_1}(s, a | \psi_1^C)$

$a = \mu_\theta(s)$

$C_{\psi_1}(s, a)$

Soft update C_{ψ_1}

Target Cost Network 1 $\acute{C}_{\psi_1}(s, a | \psi_1^C)$

$\acute{C}_{\psi_1}(\acute{s}, \acute{a})$

$\begin{matrix} \acute{C}_{\psi_1}(s, a = \mu\theta(s)) \\ (\ldots) \\ \acute{C}_{\psi_4}(s, a = \mu\theta(s)) \end{matrix}$

Actor Network

Policy gradient $\nabla_\theta J(\mu)$

Update θ^μ

$-\sum_{i=1}^{n=4} \lambda_i C_{\psi_i}(s, a = \mu\theta(s))$

Online actor network $\mu_\theta(s | \theta^\mu)$

Soft update θ^μ

$\acute{a} = \acute{\mu}(\acute{s}|\theta^{\acute{\mu}})$

Target actor network $\acute{\mu}_\theta(s | \theta^{\acute{\mu}})$

$\acute{a} = \acute{\mu}(\acute{s}|\theta^{\acute{\mu}})$

$a = \mu_\theta(s)$

$-(a - \mu_\theta(s))^2$

$+$

Critic Network

Loss function $L(\phi)$

Update Q_ϕ

Online Critic Network $Q_\phi(s, a | \phi^Q)$

Soft update Q_ϕ

Target Critic Network $\acute{Q}_\phi(s, a | \phi^Q)$

$Q_\phi(s, a)$

$\acute{Q}_\phi(\acute{s}, \acute{a})$

Update Dual Variables

$\lambda_i^k \longleftarrow \max(0, \lambda_i^t + \beta_k(C_{\psi_i}(s, a) - d_i))$

Store agent's experience $(s, a, r, c_1, c_2, c_3, c_4, \acute{s})$

Experience Replay Memory

Training Days Data

MILP Solver

GUROBI

- **Storing Expert Demonstrations**
- **Pre-training actor, critic, and cost networks**