

Agent

$\mathbf{b}_t = \text{Posterior}(\mathbb{S})$ at time step t , $b_t(E_t) = P(E_t|O_t, a_{t-1}, \mathbf{b}_{t-1})$

Belief Update Rule, this is theory, in practice we may use the Sequential Monte Carlo Simulation to get posterior like the MCMC or Importance Sampling

$$b_{t+\Delta t}(E_{t+\Delta t}) = P(E_{t+\Delta t}|O_{t+\Delta t}, a_t, \mathbf{b}_t) = \frac{P(O_{t+\Delta t}|E_{t+\Delta t}, a_t)P(E_{t+\Delta t}|a_t, \mathbf{b}_t)}{P(O_{t+\Delta t}|a_t, \mathbf{b}_t)}$$

Transition Model $\mathbf{T}_{belief} = P(\mathbf{b}_t|\mathbf{b}_{t-1}, a_{t-1})$

$$P(\mathbf{b}_{t+\Delta t}|\mathbf{b}_t, a_t) = \int_{O_{t+\Delta t} \in \mathbb{O}} P(\mathbf{b}_{t+\Delta t}|\mathbf{b}_t, a_t, O_{t+\Delta t}) dO_{t+\Delta t}$$

Observation Model

$$P(O_{t+\Delta t}|\mathbf{b}_t, a_t) = \int_{E_{t+\Delta t} \in \mathbb{S}} P(O_{t+\Delta t}|E_{t+\Delta t}, a_t) \int_{E_t \in \mathbb{S}} P(E_{t+\Delta t}|E_t, a_t) b_t(E_t) dE_t dE_{t+\Delta t}$$

Reward Model at each time step

$$R_{t+\Delta t} = r_{belief}(a_{t+\Delta t}, \mathbf{b}_{0:t+\Delta t}) = \int_{E_t \in \mathbb{S}} r(a_t, E_t) b_t(E_t) dE_t$$

The accumulated weighted total reward

$$U_t^\pi = \sum_{i=t}^T \gamma^{i-t} r_{belief}(a_i, \mathbf{b}_i)$$

The State-Action Value Function

$$Q^\pi(\mathbf{b}_t, a_t) = \mathbb{E}_{\mathbf{B}_{i>t} \sim \mathbf{T}_{belief}, A_{i>t} \sim \pi} [U^\pi(\mathbf{b}_t, a_t, \mathbf{B}_{t+1}, A_{t+1}, \dots, \mathbf{B}_T, A_T) | \mathbf{b}_t, a_t]$$

The State Value Function

$$V^\pi(\mathbf{b}_t) = \mathbb{E}_{A_t \sim \mu} [Q^\pi(\mathbf{b}_t, A_t)] \approx V^\pi(\mathbf{b}, \theta)$$

The optimal State-Action Value Function

$$Q^*(\mathbf{b}_t, a_t) = \max_\pi Q^\pi(\mathbf{b}_t, a_t)$$

The optimal State Value Function

$$V^*(\mathbf{b}_t) = \max_\pi V^\pi(\mathbf{b}_t)$$

The value of policy

$$J^\pi = \mathbb{E}_{\mathbf{B} \sim \mathbb{B}} [V^\pi(\mathbf{B})] \approx \mathbb{E}_{\mathbf{B} \sim \mathbb{B}} [V^\pi(\mathbf{B}, \theta)] = J^\pi(\theta)$$

Action is selected according to a policy $a_{t+\Delta t} \sim \pi(a|\mathbf{b}_{t+\Delta t})$

The policy is supposed to optimize the expected accumulated weighted total reward

$$\pi = \text{argmax}_\pi \mathbb{E}_\pi [U_t^\pi]$$

Proximal Policy Optimization (PPO)
Soft Actor Critic (SAC)

Policy-based Method: Approximate the policy function $\pi(a_t|\mathbf{b}_t)$ by policy network $\pi(a_t|\mathbf{b}_t, \theta^\pi)$

Then the state Value function could be approximated as $V^\pi(\mathbf{b}_t)$

$$= \mathbb{E}_{A_t \sim \mu} [Q^\pi(\mathbf{b}_t, A_t)] \approx \sum_a \pi(a|\mathbf{b}_t, \theta^\pi) Q^\pi(\mathbf{b}_t, a) = V^\pi(\mathbf{b}_t, \theta^\pi)$$

Then the value of the policy could be approximated as J^π

$$= \mathbb{E}_{\mathbf{B} \sim \mathbb{B}} [V^\pi(\mathbf{B})] \approx \mathbb{E}_{\mathbf{B} \sim \mathbb{B}} [V^\pi(\mathbf{B}, \theta^\pi)] = J^\pi(\theta^\pi)$$

$$\theta^\pi = \text{argmax}_{\theta^\pi} J^\pi(\theta^\pi)$$

Value-based Method:

Approximate the optimal state value function $V^*(\mathbf{b}_t)$ by value network $v(\mathbf{b}_t, \theta^{V^*})$

or the optimal state-action value function $Q^*(\mathbf{b}_t, a_t)$ by value network $q(\mathbf{b}_t, \theta^{Q^*})$

Actor-Critic Method:

Approximate the policy function $\pi(a_t|\mathbf{b}_t)$ by policy network

$$\pi(a_t|\mathbf{b}_t, \theta^\pi)$$

and the corresponding state value function $V^\pi(\mathbf{b}_t)$ by value network $v(\mathbf{b}_t, \theta^{V^\pi})$

World

$$\Delta E_{natural} = \text{natural_deterioration_step}(t, \Delta t, \mu_A, \sigma_A, \mu_B, \sigma_B, \mu_w, \sigma_w, \lambda, \alpha, \beta)$$

$$E_{t+\Delta t} = E_t + \Delta E_{natural}, \text{ if } a_t == 0$$

$$E_{t+\Delta t} = E_t * 1.2, \text{ if } a_t == 1$$

$$E_{t+\Delta t} = E_0 \text{ if } a_t == 2$$

$$K_{t+\Delta t} = \text{assemble_K}(E_{t+\Delta t}, A, L_e, N)$$

$$M = \text{assemble_M}(\rho, A, L_e, N)$$

Static Analysis:

$$U_{t+\Delta t} = \text{Static_solve}(K_{t+\Delta t}, F_{t+\Delta t})$$

$$O_{t+\Delta t} = (U_{t+\Delta t} + \text{noise}) [\text{mid_point}]$$

Dynamic Analysis:

$$\text{acc}_{t+\Delta t} = \text{Dynamic_solve}(K_{t+\Delta t}, M, F_{t+\Delta t})$$

$$\text{DSF}_{t+\Delta t} = \text{DamageSensitiveFeature}(\text{acc}_{0:t+\Delta t})$$

All kinds of observations

Reward Model

$$R_t \sim P(R_t|E_t, a_t, E_{t+\Delta t})$$

For instant reward $R_t \sim P(R_t|E_t, a_t) = r(E_t, a_t)$, e.g. $R_t = R_{repair} + R_{inspect} + R_{\mathbb{E}[P_{Failure}]}$