Ground Truth Comparison Metrics

1 Setup

Given R Monte Carlo runs over T steps, let $\mathcal{A}$ denote the action set. For each run r at step t:

- $s_t^{(r)}$ is the true state.
- $\mathcal{P}_t^{(r)}$ is the LFP belief polytope after propagation.
- $S_{\text{safe}}(a)$ is the set of state indices where action a is safe (from the inverted shield).

Define the **predicted minimum safety probability** for action a in run r at step t:

$$\hat{p}_{t,a}^{(r)} \;=\; \min_{b \in \mathcal{P}_t^{(r)}} \sum_{i \in S_{\text{safe}}(a)} b_i$$

Define the **empirical safety indicator** for action a in run r at step t:

$$\mathbb{1}_{t,a}^{(r)} \;=\; \begin{cases} 1 & \text{if } s_t^{(r)} \in S_{\text{safe}}(a) \\ 0 & \text{otherwise} \end{cases}$$

Let $R_t \le R$ be the number of runs that have not terminated by step t. The step-averaged quantities are:

$$\bar{p}_{t,a} \;=\; \frac{1}{R_t} \sum_{r=1}^{R_t} \hat{p}_{t,a}^{(r)}, \qquad \bar{e}_{t,a} \;=\; \frac{1}{R_t} \sum_{r=1}^{R_t} \mathbb{1}_{t,a}^{(r)}$$

2 Action Coverage Rate

The action coverage rate measures the fraction of (step, action) pairs where the empirical safety frequency meets or exceeds the predicted minimum:

$$\texttt{action\_coverage\_rate} \;=\; \frac{\left|\{(t,a) : \bar{e}_{t,a} \ge \bar{p}_{t,a}\}\right|}{T \cdot |\mathcal{A}|}$$

A value close to 1 indicates that the LFP lower bounds are sound in aggregate: the predicted minimum rarely overestimates the true safety frequency.

3 Mean Conservatism Gap

The mean conservatism gap quantifies how conservative the predicted lower bounds are on average:

$$\texttt{mean\_conservatism\_gap} \;=\; \frac{1}{T \cdot |\mathcal{A}|} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \left(\bar{e}_{t,a} - \bar{p}_{t,a}\right)$$

A positive value indicates the predictions are conservative (the true frequency exceeds the predicted minimum). A value near zero indicates tight bounds.