Department of Electrical and Computer Engineering National University of Singapore Dr Fan Shi, Dr Armin Lederer	**Robotics and Embodied Artificial Intelligence** **CEG5306** Assignment 4: safety and partial observability in RL	Autumn 2025

1. **Safety using Lagrangian Relexations (~ 5 pts)**
   Consider the scenario that the primal-dual method is used to find a safe policy for a CMDP. After running a few iterations of the algorithm with constraint threshold $c = 1.0$, you obtain value functions for the task and safety constraints under the current policy $\boldsymbol{\pi}$ satisfying

$$\mathbb{E}[V^{\pi}_{\text{task}}(\boldsymbol{s})] = -0.9 \tag{1}$$
$$\mathbb{E}[V^{\pi}_{\text{safety}}(\boldsymbol{s})] = 0.5. \tag{2}$$

   (a) In which range is the value of the Lagrange multiplier λ?

   (b) Assume $\lambda = 5$ holds at the considered iteration. Compute a gradient update step of λ with step size 0.1 for the primal-dual algorithm described in the lecture.

   (c) How can you determine that a safe policy has been found?

2. **Safety using Control Barrier Functions (~ 5 pts)**
   Consider a MDP with Gaussian transition probabilities $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{s}, \boldsymbol{a}), 0.1\boldsymbol{I})$ with mean

$$\boldsymbol{\mu}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a})}[\boldsymbol{s}'] = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{B}\boldsymbol{a} \tag{3}$$

   and covariance matrix $0.1\boldsymbol{I}$ for matrix $\boldsymbol{A} = \begin{bmatrix} 0.0 & 0.5 \\ 0.75 & 1.0 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 0.0 \\ 1.0 \end{bmatrix}$. Show that the control barrier function $h(\boldsymbol{s}) = 1 - \|\boldsymbol{s}\|^2$ ensures feasibility of the safety filter

$$\boldsymbol{a}_t = \arg\min_{\boldsymbol{a}} \|\boldsymbol{a} - \boldsymbol{\pi}(\boldsymbol{s})\| \tag{4a}$$
$$\text{s.t. } \mathbb{E}_{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a})}[h(\boldsymbol{s}')] \geq \alpha h(\boldsymbol{s}) \tag{4b}$$

   for arbitrary nominal policies $\boldsymbol{\pi}(\boldsymbol{s})$, constant $\alpha = 0.25$, and all states $\boldsymbol{s} : \|\boldsymbol{s}\| \leq 1$. Be mathematically accurate! Hint: Pick $\boldsymbol{a} = -\begin{bmatrix} 0.75 & 1.0 \end{bmatrix} \boldsymbol{s}$

3. **Partial Observability in RL (~ 4 pts)**
   You are tasked with designing a reinforcement learning algorithm for a partially observable version of the MuJoCo Ant environment. You have decided for an approach relying on a recurrent neural network as an encoder for the policy and value function. Since you are unsure about certain design choices, you have tested different variants of your algorithm resulting in the plots shown in Fig. 1. Which conclusions can you draw in terms of best algorithm (sac, td3), encoder type (gru, lstm), sequence length (5, 64) and inputs (o(bservation),(a)ction, (r)eward)?

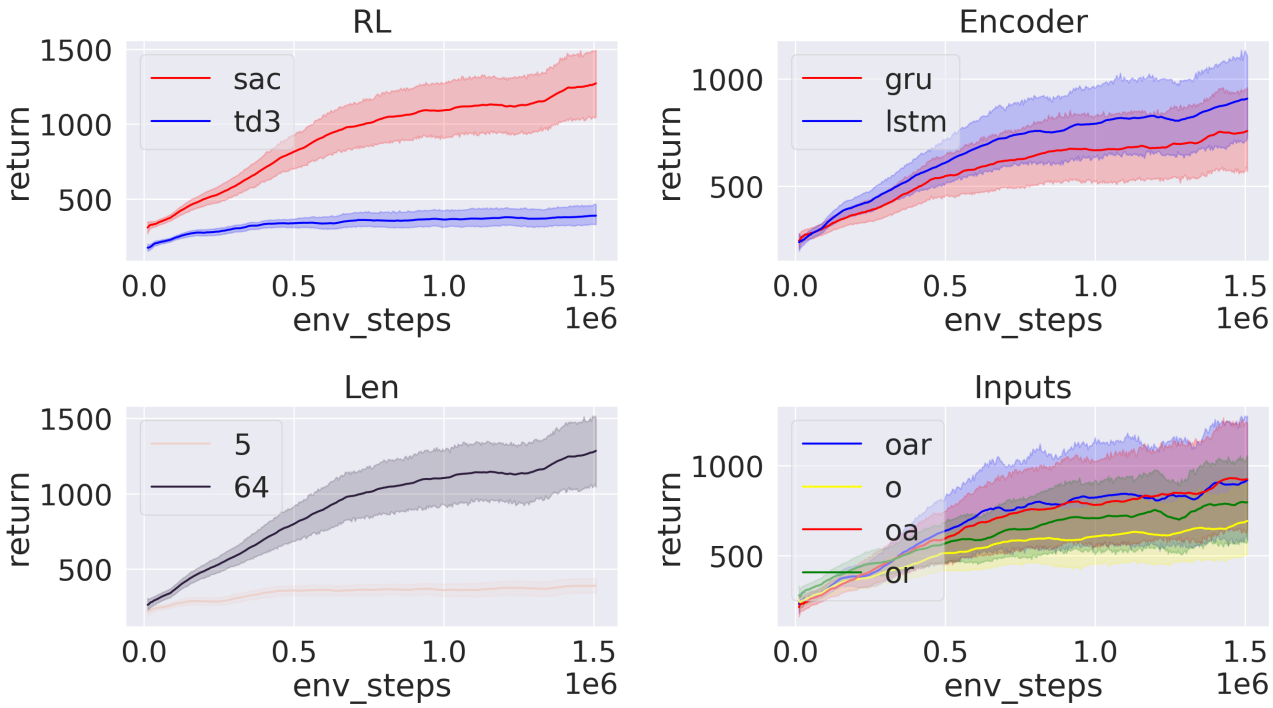

Figure 1: Ablation study on the Ant-P environment. Taken from `https://arxiv.org/abs/2110.05038`