Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 23

Solution By: Arshia Gharooni

Q1

Guided Cost Learning alternates (i) reward updates that maximise a logistic-regression objective and (ii) TRPO steps that minimise expected cost. State the necessary condition on the Jensen–Shannon divergence between learner and expert trajectory distributions at convergence, and argue informally why that implies the learned cost can no longer improve.

Let $p_E(\tau)$ be the expert trajectory distribution and $p_\pi(\tau)$ the learner's. At convergence it is necessary that

$$D_{\mathrm{JS}}(p_\pi \,\|\, p_E) = 0 \quad \Longleftrightarrow \quad p_\pi(\tau) = p_E(\tau) \text{ for all } \tau,$$

so the optimal discriminator satisfies $D^*(\tau) = \frac{1}{2}$ everywhere.
The logistic-regression objective used for the reward update is

$$\max_\theta \; \mathbb{E}_{\tau\sim p_E}[\log D_\theta(\tau)] + \mathbb{E}_{\tau\sim p_\pi}[\log(1 - D_\theta(\tau))]\,,$$

whose optimum equals $-\log 4$ and has *zero gradient* w.r.t. the cost/reward parameters when $p_\pi = p_E$ (since $D^*_\theta \equiv \frac{1}{2}$). Hence the reward update provides no further learning signal.
With the cost fixed, a TRPO step can only reduce expected cost by changing π, but any such change would move p_π away from p_E, making $D_{\mathrm{JS}}(p_\pi \| p_E) > 0$ and reintroducing a discriminator signal—contradicting convergence. Therefore, at $D_{\mathrm{JS}} = 0$ the learned cost cannot improve further (the system is at a stationary point for both the discriminator/reward and the policy).

Q2

Suppose the convex hull of learner feature expectations has already intersected the open ε-ball centered at the expert expectation. Give a geometric argument (no algebra) showing why every subsequent quadratic-program step of the feature-matching algorithm can only return a margin $t \leq \varepsilon$.

Let C be the convex hull of the learner feature expectations and let μ_E be the expert expectation. The QP step seeks the maximum-margin separating hyperplane between the point μ_E and the convex set C. Geometrically, the optimal margin equals the shortest Euclidean distance from μ_E to C:

$$t^{\star} = \operatorname{dist}(\mu_E, C) = \inf_{x \in C} \|\mu_E - x\|.$$

Since C already intersects the open ball $B(\mu_E, \varepsilon)$, there exists $y \in C$ with $\|\mu_E - y\| < \varepsilon$. Hence

$$t^{\star} = \operatorname{dist}(\mu_E, C) \leq \|\mu_E - y\| < \varepsilon,$$

so any subsequent QP step can return only a margin $t \leq \varepsilon$ (indeed, $t < \varepsilon$).

Q3

The feature-matching Apprenticeship-Learning algorithm stops when

$$\left\|\hat{\mu}_E - \mu(\pi)\right\|_2 \leq \varepsilon.$$

Assume the following:

- **Feature dimension** $k = 15$**,**
- **The Euclidean diameter of the feature expectation space is** $D = 4$**,**
- **Desired tolerance** $\varepsilon = 0.05$**.**

Using the tightest bound derived in the lecture, the maximum number of calls to the forward RL solver that the algorithm is *guaranteed* to make is closest to:

1. 3000
2. 24000
3. 96000
4. 8000
5. 4800

Correct Answers: C

The tight bound on the number of iterations (and hence forward RL calls) is

$$N \;\leq\; k\left(\frac{D}{\varepsilon}\right)^2.$$

Plug in $k = 15$, $D = 4$, and $\varepsilon = 0.05$:

$$\left(\frac{4}{0.05}\right)^2 = 80^2 = 6400, \qquad N \leq 15 \times 6400 = 96{,}000.$$

Therefore, the closest option is **96,000**.