Quiz Solutions - Lecture 17

Solution By: Arshia Gharooni

Q1

Give a concise argument showing that, for sufficiently small step-size, a vanilla policy-gradient update is guaranteed to improve performance when the advantage function is uniformly bounded.

Since $A_{\pi_{\theta_k}}$ is bounded by some constant C, the second-order term in the performance-difference expansion can be bounded by a quantity proportional to α_k^2 times C. Choosing a step-size such that $D_{\mathrm{KL}}(\pi_{\theta_k} \| \pi_{\theta_{k+1}}) \leq 2(1-\gamma)\alpha_k^2/C$ ensures that the negative second-order penalty does not outweigh the positive first-order gain, resulting in $J(\theta_{k+1}) \geq J(\theta_k)$.

Q2

Explain why the soft Bellman operator $\mathcal{T}^{\pi}_{\text{soft}}$ is still a γ-contraction in the supremum norm, even though it contains an additional entropy term.

The operator differs from the standard Bellman operator only by an additive term $-\alpha \log \pi(a|s)$, which depends on (s, a) but not on the next-state value estimate. Contraction properties hinge on the γ factor multiplying the future value; since this factor remains unchanged, the operator remains a γ-contraction and enjoys a unique fixed point.

Q3

The actor update in SAC can be viewed as minimising a Kullback–Leibler divergence between the current policy and a Boltzmann distribution derived from Q_ψ. Which KL direction is minimised?

A. $D_{\mathrm{KL}}\big(\exp(Q/\alpha) \,\|\, \pi_\theta\big)$
B. $D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \exp(Q/\alpha)\big)$
C. The symmetric Jensen–Shannon divergence between the two distributions
D. Neither; SAC avoids KL divergence entirely

Correct Answers: B

The policy is updated by information projection onto the Boltzmann target, solving

$$\arg\min_{\pi} \mathbb{E}_s \left[D_{\mathrm{KL}} \left(\pi(\cdot|s) \,\|\, \exp(Q_\psi/\alpha) \right) \right],$$

thus minimising the forward KL with the policy π_θ appearing in the first argument.

Q4

Which statement best describes the role of the temperature parameter α in Soft-Actor-Critic?

A. It rescales the discount factor to balance bias and variance.
B. It controls the trade-off between the expected return and the policy's entropy.
C. It stabilises the target network by Polyak averaging.
D. It enforces a hard trust region on the policy update.

Correct Answers: B

The temperature α weights the entropy bonus $\mathcal{H}(\pi(\cdot|s))$ in the objective J_{soft}, tuning exploration versus exploitation.