Lecture 25, Mar 12, 2026

Q-Learning – Policy Improvement

- We have now solved policy evaluation, but we still need to figure out how to do policy improvement in a tractable way
- Because the Q-function is known, we can now perform the argmin by differentiating the Q-function with respect to u and setting it to zero; i.e. we solve for $\frac{\partial}{\partial u} Q^{\mu^j}(x,u) = 0$ offline
- Example: policy improvement for LQR, where $\mu(x) = -Kx$
  - $$\begin{aligned} Q^\mu(x,u) &= r(x,u) + \gamma V^\mu(f(x,u)) \\ &= x^T Q x + u^T R u + \gamma (Ax+Bu)^T P^\mu (Ax+Bu) \\ &= \begin{bmatrix} x^T & u^T \end{bmatrix} \begin{bmatrix} S_{xx} & S_{xu} \\ S_{ux} & S_{uu} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \\ &= x^T S_{xx} x + x^T S_{ux}^T u + x^T S_{xu} u + u^T S_{uu} u \end{aligned}$$
  - $\frac{\partial Q^\mu}{\partial u} = x^T S_{ux}^T + x^T S_{xu} + 2u^T S_{uu}$
  - Setting to zero, we get $\mu(x) = -\frac{1}{2} S_{uu}^{-1} (S_{ux} + S_{xu}^T) x$
  - Notice that $-\frac{1}{2} S_{uu}^{-1} (S_{ux} + S_{xu}^T)$ is the new gain K^{j+1}
- Overall our Q-learning algorithm is:
  1. Initialization: select any admissible $\mu^0 \in \mathcal{M}$
  2. Policy evaluation:
     - Prediction error: $e_1(k) = (\hat{\psi}^{j+1})^T(k) v(k) - r(x(k), u(k))$
     - Gradient law: $\hat{\psi}^{j+1}(k+1) = \hat{\psi}^{j+1}(k) - \gamma_1(k) e_1(k) v(k)$ where $\gamma_1(k) = \frac{\bar{\gamma}}{1 + \|v(k)\|^2}, \bar{\gamma} \in (0,2)$
     - New regressor: $v(k) = w(x(k), u(k)) - \gamma w(x(k+1), \mu^j(x(k+1)))$
       * $w(x,u)$ is our parametrization for Q-function approximation, so $Q^\mu(x,u) = \psi^T w(x,u)$
       * Importantly here we don't use $u(k+1)$!
     - Note that the plant updates as $x(k+1) = f(x(k), u(k))$
  3. Policy improvement: Solve $\frac{\partial Q^\mu(x,u)}{\partial u} = \frac{\partial}{\partial u} \left(\hat{\psi}^{j+1} w(x,u) \right) = 0$ (for u as a function of x) to get μ^{j+1}
- Note that policy evaluation only converges if we have a persistently exciting $v(k)$, which we do by changing $u(k)$
  - We do this by introducing a probing noise to $u(k)$, instead of just doing $u(k) = \mu(x(k))$ (this is the "exploration")