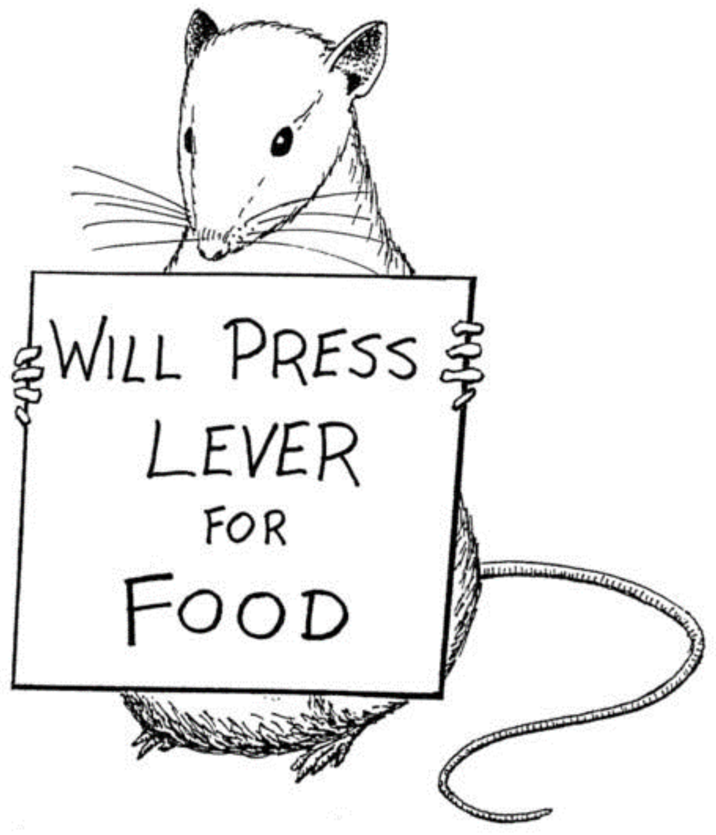
CS 277: Control and Reinforcement Learning

Winter 2021

Lecture 7: Optimal Control

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine



Logistics

assignments

- Assignment 2 due this Friday

Today's lecture

Off-policy evaluation, TRPO

Stability, reachability, stabilizability

Linear Quadratic Regulator

Hamiltonian

n-step DQN

- In DQN, the target y is the 1-step rolled out value predictor:

$$y^1(r_t, s_{t+1}) = r_t + \gamma \max_{a_{t+1}} Q_{\bar\theta}(s_{t+1}, a_{t+1})$$

- Can we instead use the n-step rolled out predictor?

$$y^n(r_t, \ldots, r_{t+n-1}, s_{t+n}) = r_t + \cdots + \gamma^{n-1} r_{t+n-1} + \gamma^n \max_{a_{t+n}} Q_{\bar\theta}(s_{t+n}, a_{t+n})$$

  - The bootstrapped value predictor $\max Q_{\bar\theta}$ is more discounted $\implies$ less bias
  - But we use just one sample of the n-step trajectory, not averaged $\implies$ more variance
  - May be a better tradeoff, allows TD(λ) methods that combine different n

Can we use n-step DQN off-policy?

- n-step DQN target:

$$y^n(r_t, \ldots, r_{t+n-1}, s_{t+n}) = r_t + \cdots + \gamma^{n-1} r_{t+n-1} + \gamma^n \max_{a_{t+n}} Q_{\bar{\theta}}(s_{t+n}, a_{t+n})$$

- Problem: $a_{t+1}, \ldots, a_{t+n-1}$ must all be on-policy for the target to be unbiased
- Solutions:
  - Ignore the problem: just use it off-policy anyway
  - Use Importance Sampling: use data $x \sim p_2$ to estimate $\mathbb{E}_{x\sim p_1}[f(x)]$

$$\mathbb{E}_{x\sim p_1}[f(x)] = \mathbb{E}_{x\sim p_2}\left[\frac{p_1(x)}{p_2(x)} f(x)\right]$$

Off-policy policy evaluation

- How to get an unbiased estimator of $\mathcal{J}_\theta = \mathbb{E}_{\xi \sim p_\theta}[R(\xi)]$

  from data sampled from a different distribution $\xi_1, \ldots, \xi_N \sim p_{\theta'}$?

importance weight

$$\mathcal{J}_\theta = \mathbb{E}_{\xi \sim p_{\theta'}}\left[\frac{p_\theta(\xi)}{p_{\theta'}(\xi)} R(\xi)\right]$$

$$\frac{p_\theta(\xi)}{p_{\theta'}(\xi)} = \frac{p(s_0)}{p(s_0)} \prod_t \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta'}(a_t \mid s_t)} \frac{p(s_{t+1} \mid s_t, a_t)}{p(s_{t+1} \mid s_t, a_t)} = \prod_t \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta'}(a_t \mid s_t)} = w$$

- A reward r_t is not affected by future probability divergence

$$\mathcal{J}_\theta = \sum_t \mathbb{E}_{s_t, a_t \sim p_{\theta'}}[\gamma^t w r_t] = \sum_t \mathbb{E}_{s_t, a_t \sim p_{\theta'}}\left[\gamma^t w_t r_t\right] \qquad w_t = \prod_{t' \leq t} \frac{\pi_\theta(a_{t'} \mid s_{t'})}{\pi_{\theta'}(a_{t'} \mid s_{t'})}$$

Off-policy n-step DQN

- Importance weighted n-step DQN target:

$$y^n(r_t, \dots, r_{t+n-1}, s_{t+n}) = \tilde{r}_t + \cdots + \gamma^{n-1}\tilde{r}_{t+n-1} + \gamma^n \max_{a_{t+n}} Q_{\bar{\theta}}(s_{t+n}, a_{t+n})$$

  - With importance-weighted rewards: $\tilde{r}_{t+k} = \prod_{t'=t+1}^{t+k} \frac{\pi_\theta(a_{t'} \mid s_{t'})}{\pi_{\theta'}(a_{t'} \mid s_{t'})} r_{t+k}$

  - Replay buffer must contain consecutive $\geq n$-step trajectories

    - and the probability $\pi_{\theta'}(a \mid s)$ of taking each action using the policy at that time

Off-policy MC advantage estimation

- MC advantage estimation: $\sum_t \gamma^t r(s_t, a_t) - V_{\pi_\theta}(s)$

$$= \sum_t \gamma^t (r(s_t, a_t) + \gamma V_{\pi_\theta}(s_{t+1}) - V_{\pi_\theta}(s_t)) = \sum_t \gamma^t \hat{A}^1_{\pi_\theta}(s_t, a_t)$$

- What if estimate the advantage of our current π_θ under a proposed new policy $\pi_{\theta'}$?

we want to maximize this!

$$\mathbb{E}_{\xi \sim p_{\theta'}}\left[\sum_t \gamma^t \hat{A}^1_{\pi_\theta}(s_t, a_t)\right] = \mathbb{E}_{\xi \sim p_{\theta'}}\left[\sum_t \gamma^t r(s_t, a_t) - V_{\pi_\theta}(s_0)\right] = \mathcal{J}_{\theta'} - \mathcal{J}_\theta$$

$$= \sum_t \gamma^t \mathbb{E}_{s_t, a_t \sim p_{\theta'}}[\hat{A}^1_{\pi_\theta}(s_t, a_t)]$$

$$= \sum_t \gamma^t \mathbb{E}_{s_t \sim p_{\theta'}}\left[\mathbb{E}_{a_t | s_t \sim \pi_\theta}\left[\frac{\pi_{\theta'}(a_t | s_t)}{\pi_\theta(a_t | s_t)} \hat{A}^1_{\pi_\theta}(s_t, a_t)\right]\right]$$

  - We can't estimate this empirically, because of $s_t \sim p_{\theta'}$; what's the difference if we just use $s_t \sim p_\theta$?

Change of measure

- Intuition: switching from $s_t \sim p_{\theta'}$ to $s_t \sim p_\theta$ isn't too bad if they are similar
  - In Kullback-Leibler (KL) divergence: $\mathbb{D}[p_{\theta'} \| p_\theta]$
  - Or in Total Variation (TV) distance: $\delta(p_{\theta'}, p_\theta) = \frac{1}{2}|p_{\theta'} - p_\theta|_1 \leq \sqrt{\frac{1}{2}\mathbb{D}[p_{\theta'} \| p_\theta]}$
- Suppose we $\pi_{\theta'}$, π_θ are similar: $|\pi_{\theta'}(\cdot|s) - \pi_\theta(\cdot|s)|_1 \leq 2\sqrt{\frac{\epsilon}{2}} = \epsilon'$ for all s
  - Then they induce similar marginal distributions at time t

$$|\mathbb{E}_{s_t \sim p_{\theta'}}[f(s)] - \mathbb{E}_{s_t \sim p_\theta}[f(s)]| \leq |p_{\theta'} - p_\theta|_1 \max_{s_t} f(s_t) \leq t\epsilon' \max_{s_t} f(s_t)$$

Trust-Region Policy Optimization (TRPO)

$$\max_{\theta'} \sum_t \gamma^t \mathbb{E}_{s_t \sim p_\theta} \left[\mathbb{E}_{a_t|s_t \sim \pi_\theta} \left[\frac{\pi_{\theta'}(a_t \mid s_t)}{\pi_\theta(a_t \mid s_t)} \hat{A}^1_{\pi_\theta}(s_t, a_t) \right] \right]$$

s.t. $\mathbb{D}[\pi_{\theta'} \| \pi_\theta] \leq \epsilon$

- For small enough ϵ, the objective is close to $\mathcal{J}_{\theta'} - \mathcal{J}_\theta$
  - Guarantees improvement of our objective; in practice, good ϵ is too large
- TRPO loss:

$$\mathcal{L}_\theta(s, a, r, s') = -\underbrace{\frac{\pi_\theta(a \mid s)}{\pi_{\bar{\theta}}(a \mid s)}}_{\text{importance weight}} \underbrace{(r + \gamma V_\phi(s') - V_\phi(s))}_{\text{advantage estimate}} + \underbrace{\lambda}_{\text{Lagrange multiplier for KL constraint}} \mathbb{D}[\pi_\theta(\cdot \mid s) \| \pi_{\bar{\theta}}(\cdot \mid s)]$$

- The actual algorithm is more complicated; simpler variant: PPO

Today's lecture

Off-policy evaluation, TRPO

Stability, reachability, stabilizability

Linear Quadratic Regulator

Hamiltonian

Linear Time-Invariant (LTI) system

$$\cdots \rightarrow x_{t-1} \xrightarrow{A} x_t \xrightarrow{A} x_{t+1} \rightarrow \cdots$$

- Continuous state space: $x_t \in \mathbb{R}^n$
  - Distributions and values may be hard to represent
- Simplest system — linear: $x_{t+1} = Ax_t$ $\qquad A \in \mathbb{R}^{n \times n}$
  - Linear Time-Invariant (LTI): A does not depend on t
- How does the system evolve over time?

$$x_t = A^t x_0$$

Stability

- To analyze: use eigenvectors $\lambda e = Ae$

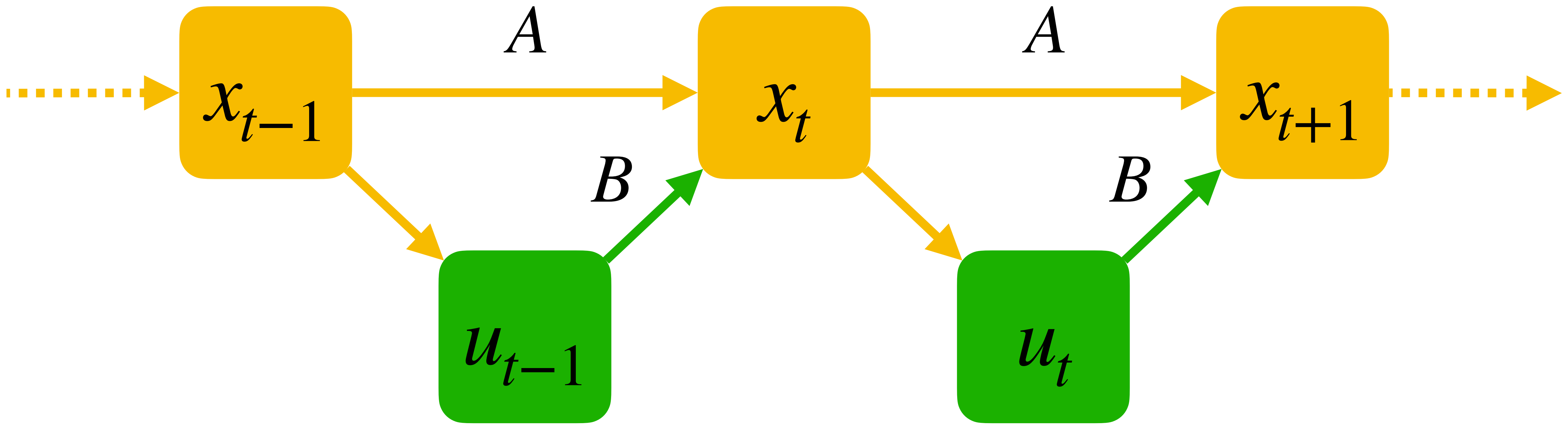
- Consider a basis of eigenvectors $e_1, \ldots, e_n \in \mathbb{C}^n$

$$x_0 = \sum_i \alpha_i e_i \implies x_1 = Ax_0 = \sum_i \alpha_i \lambda_i e_i \implies x_t = \sum_i \alpha_i \lambda_i^t e_i$$

- Stability: all $\|\lambda_i\| < 1$, so that $\lim_{t\to\infty} x_t = 0$

- Instability: some $\|\lambda_i\| > 1$, so that $\lim_{t\to\infty} \|x_t\| \to \infty$
  - When some $\|\lambda_i\| = 1$, that component never vanishes or explodes

Linear control systems



- Continuous action (control) space: $u_t \in \mathbb{R}^m$
- Controlled LTI system: $x_{t+1} = Ax_t + Bu_t \qquad B \in \mathbb{R}^{n \times m}$

$$x_t = A^t x_0 + A^{t-1}Bu_0 + \cdots + ABu_{t-2} + Bu_{t-1}$$

$$x_t - A^t x_0 = \begin{bmatrix} B & AB & \cdots & A^{t-1}B \end{bmatrix} \begin{bmatrix} u_{t-1} \\ u_{t-2} \\ \vdots \\ u_0 \end{bmatrix}$$

Reachability

$$x_t - A^t x_0 = \begin{bmatrix} B & AB & \cdots & A^{t-1}B \end{bmatrix} \begin{bmatrix} u_{t-1} \\ u_{t-2} \\ \vdots \\ u_0 \end{bmatrix}$$

- Can we reach a given state x_t at time t?
  - If and only if $x_t - A^t x_0 \in \operatorname{span} \begin{bmatrix} B & AB & \cdots & A^{t-1}B \end{bmatrix}$
- Can we reach all states eventually?
  - Cayley-Hamilton: A satisfies its characteristic polynomial $\implies A^n \in \operatorname{span}\{I, A, \ldots, A^{n-1}\}$
    - Sufficient to consider controllability matrix: $\mathscr{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$
  - Reachability = $\operatorname{span} \mathscr{C} = \mathbb{R}^n \iff \operatorname{rank} \mathscr{C} = n$

Stabilizability

- Controllability matrix: $\mathscr{C} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$
- Can we eventually reach $x_t = 0$?
    - Some components may be uncontrollable but stable — they reach 0 on their own
    - Stabilizability: $e_i \in \operatorname{span}\mathscr{C}$ for all e_i with $\lambda_i \geq 1$

Today's lecture

Off-policy evaluation, TRPO

Stability, reachability, stabilizability

Linear Quadratic Regulator

Hamiltonian

Quadratic costs

- Simplest rewards: concave quadratic (linear has no maximum)
  - Consider costs: $c(x_t, u_t) = \frac{1}{2}x_t^\intercal Q x_t + \frac{1}{2}u_t^\intercal R u_t$
  - $Q \in \mathbb{R}^{n \times n}$ is positive semidefinite $Q \succeq 0$: $\frac{1}{2}x^\intercal Q x \geq 0$ for all x
    - No incentive to go to infinity in any direction
  - $R \in \mathbb{R}^{m \times m}$ is positive definite $R \succ 0$: $\frac{1}{2}u^\intercal R u > 0$ for all u
    - Incentive against infinite control in any direction
  - Usually no discounting — finite or infinite undiscounted horizon

Linear Quadratic Regulator (LQR)

- Given LTI dynamics + quadratic cost (A, B, Q, R)
- Find the control function $u_t(x_t)$
- That minimizes $\mathcal{J}(x_0, u) = \sum_{t=0}^{T-1} c(x_t, u_t) = \frac{1}{2} \sum_{t=0}^{T-1} x_t^\intercal Q x_t + u_t^\intercal R u_t$
- Such that $x_{t+1} = f(x_t, u_t) = Ax_t + Bu_t$ for all t

Bellman recursion

- $\mathcal{J}_t^*(x_t) = \min_{u_t} c(x_t, u_t) + \mathcal{J}_{t+1}^*(x_{t+1})$
- Let's solve while we prove by induction that $\mathcal{J}_t^*$ is quadratic
  - Base case: $\mathcal{J}_T^* = 0$
  - Assume: $\mathcal{J}_{t+1}^*(x_{t+1}) = \frac{1}{2}x_{t+1}^\intercal S_{t+1} x_{t+1} \qquad S_{t+1} \succeq 0$
  - Solve: $\nabla_{u_t}(c(x_t, u_t) + \mathcal{J}_{t+1}^*(x_{t+1})) = 0$

Bellman optimality

$$
\begin{aligned}
0 &= \nabla_{u_t}(c(x_t, u_t) + \mathcal{J}^*_{t+1}(x_{t+1})) \\
&= \frac{1}{2}\nabla_{u_t}(x_t^\intercal Q x_t + u_t^\intercal R u_t + (Ax_t + Bu_t)^\intercal S_{t+1}(Ax_t + Bu_t)) \\
&= Ru_t + B^\intercal S_{t+1}(Ax_t + Bu_t)
\end{aligned}
$$

$$u_t^* = -(R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A x_t$$

- Plugging u_t^* into the Bellman recursion and rearranging terms:

$$\mathcal{J}_t^*(x_t) = \frac{1}{2} x_t^\intercal (Q + A^\intercal (S_{t+1} - S_{t+1} B (R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1}) A) x_t$$

- Ricatti equation: $S_t = Q + A^\intercal (S_{t+1} - S_{t+1} B (R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1}) A$

Optimal control: properties

- Linear control policy: $u_t = L_t x_t$
  - Feedback gain: $L_t = -(R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A$
- Quadratic value (cost-to-go) function $\mathcal{J}_t(x_t)^* = \frac{1}{2} x_t^\intercal S_t x_t$
  - Cost Hessian $S_t = \nabla^2_{x_t} \mathcal{J}_t^*$ is the same for all x_t
- Ricatti equation for S_t can be solved recursively backward

$$S_t = Q + A^\intercal (S_{t+1} - S_{t+1} B (R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1}) A$$

  - Without knowing any actual states or controls (!) = at system design time
- Woodbury matrix identity shows $S_t = Q + A^\intercal (S_{t+1}^\dagger + B R^{-1} B^\intercal)^\dagger A \geq 0$

Infinite horizon

- Average cost: $\mathcal{J} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x_t, u_t)$
- For each finite T we solve with Bellman recursion, affected by end $\mathcal{J}_T = 0$
- In the limit, the solution must converge to time-independent
  - Discrete-time algebraic Ricatti equation (DARE):

$$S = Q + A^\intercal (S - SB(R + B^\intercal SB)^{-1} B^\intercal S)A$$

Non-homogeneous case

- More generally, LQR can have lower-order terms
  - $x_{t+1} = f_t(x_t, u_t) = A_t x_t + B_t u_t + c_t$
  - $c_t(x_t, u_t) = \frac{1}{2} x_t^\intercal Q_t x_t + \frac{1}{2} u_t^\intercal R_t u_t + u_t^\intercal N_t x_t + q_t^\intercal x_t + r_t^\intercal u_t + s_t$
- More flexible modeling, e.g. tracking a target trajectory $\frac{1}{2}(x_t - \tilde{x}_t)^\intercal Q(x_t - \tilde{x}_t)$
- Solved essentially the same way
  - Cost-to-go $\mathcal{J}$ will also have lower-order terms

Today's lecture

Off-policy evaluation, TRPO

Stability, reachability, stabilizability

Linear Quadratic Regulator

Hamiltonian

Co-state

- Consider the cost-to-go $\mathcal{J}_t(x_t, u) = c(x_t, u_t) + \mathcal{J}_{t+1}(f(x_t, u_t), u)$

- To study its landscape over state space, consider its gradient

$$\nu_t = \nabla_{x_t} \mathcal{J}_t = \nabla_{x_t} c_t + \nabla_{x_{t+1}} \mathcal{J}_{t+1} \nabla_{x_t} f_t = \nabla_{x_t} c_t + \nu_{t+1} \nabla_{x_t} f_t$$

  - Co-state $\nu_t \in \mathbb{R}^n$ = direction of steepest increase in cost-to-go

  - Linear backward recursion, initialization: $\nu_T = 0$

  - $\nabla_{x_t} f_t$ = Jacobian of the dynamics

Lagrangian

- Constrained optimization: $\max_u g(u)$ s.t. $h(u) = 0$
  - Equivalent to Lagrangian (with Lagrange multiplier λ): $\max_u \min_\lambda g(u) + \lambda h(u)$
- Our optimization problem: $\min_u \mathcal{J}$ s.t. $x_{t+1} = f(x_t, u_t)$
  - Lagrangian: $\mathcal{L} = \sum_{t=0}^{T-1} c(x_t, u_t) + \nu_{t+1}(f(x_t, u_t) - x_{t+1})$
  - At the optimum: $\nabla_{x_t} \mathcal{L} = 0 \implies \nu_t = \nabla_{x_t} c_t + \nu_{t+1} \nabla_{x_t} f_t$ = the co-state
- Lagrange multipliers often have their own meaning

Hamiltonian

- Hamiltonian = first-order approximation of cost-to-go

$$\mathscr{H}_t = c(x_t, u_t) + \nu_{t+1} f(x_t, u_t)$$

- At the optimum, defines all x, u, and ν in one equation:

$$\nabla_{x_t} \mathscr{H}_t = \nabla_{x_t} c_t + \nu_{t+1} \nabla_{x_t} f_t = \nu_t$$
$$\nabla_{\nu_{t+1}} \mathscr{H}_t = f(x_t, u_t) = x_{t+1}$$
$$\nabla_{u_t} \mathscr{H}_t = \nabla_{u_t} (\mathscr{L} + \nu_{t+1} x_{t+1}) = 0$$

- Can solve these $(2n + m)T$ equations in $(2n + m)T$ variables
  - Generally, nonlinear with many local optima

Hamiltonian in LQR

- In LQR, the Hamiltonian is quadratic

$$\mathscr{H}_t = \frac{1}{2}x_t^\intercal Q x_t + \frac{1}{2}u_t^\intercal R u_t + \nu_{t+1}(Ax_t + Bu_t)$$

- This suggest forward–backward recursions for x, u, and ν:

$$x_{t+1} = \nabla_{\nu_{t+1}} \mathscr{H}_t = Ax_t + Bu_t$$

$$\nu_t = \nabla_{x_t} \mathscr{H}_t = \nu_{t+1}A + x_t^\intercal Q$$

$$\nabla_{u_t} \mathscr{H}_t = Ru_t + B^\intercal \nu_{t+1}^\intercal = 0$$

- These correspond to the previous approach with $\nu_t^\intercal = S_t x_t \quad u_t = L_t x_t$

Recap

- LQR = simplest dynamics: linear; simplest cost: quadratic
- Can characterize stability, reachability, stabilizability in terms of (A, B)
- Can use Ricatti equation to find cost-to-go Hessian
- Alternatively: Hamiltonian gives state forward / co-state backward recursions