Optimal Control Theory and the Linear Quadratic Regulator

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2022

Today

- Feedback from last lecture
- Recap
- Finite-horizon discrete MDPs
- General optimal control problem
- The linear quadratic regulator (LQR) problem

Feedback from feedback forms

1. Thank you to everyone who filled out the forms!
2.

Today

- Feedback from last lecture
- Recap
- Finite-horizon discrete MDPs
- General optimal control problem
- The linear quadratic regulator (LQR) problem

Recap

- For discrete MDPs, we covered some great algorithms for computing the optimal policy (reminder, we haven't done any *learning* in MDPs yet)
- But all algorithms polynomially in the size of the state and action spaces… what if one or both are infinite?
- In this unit (next 3 lectures), we will discuss computation of good/optimal policies in continuous state and action spaces (still no learning yet!)

Today

- Feedback from last lecture
- Recap
- Finite-horizon discrete MDPs
- General optimal control problem
- The linear quadratic regulator (LQR) problem

Recap
+
Finite Horizon MDPs

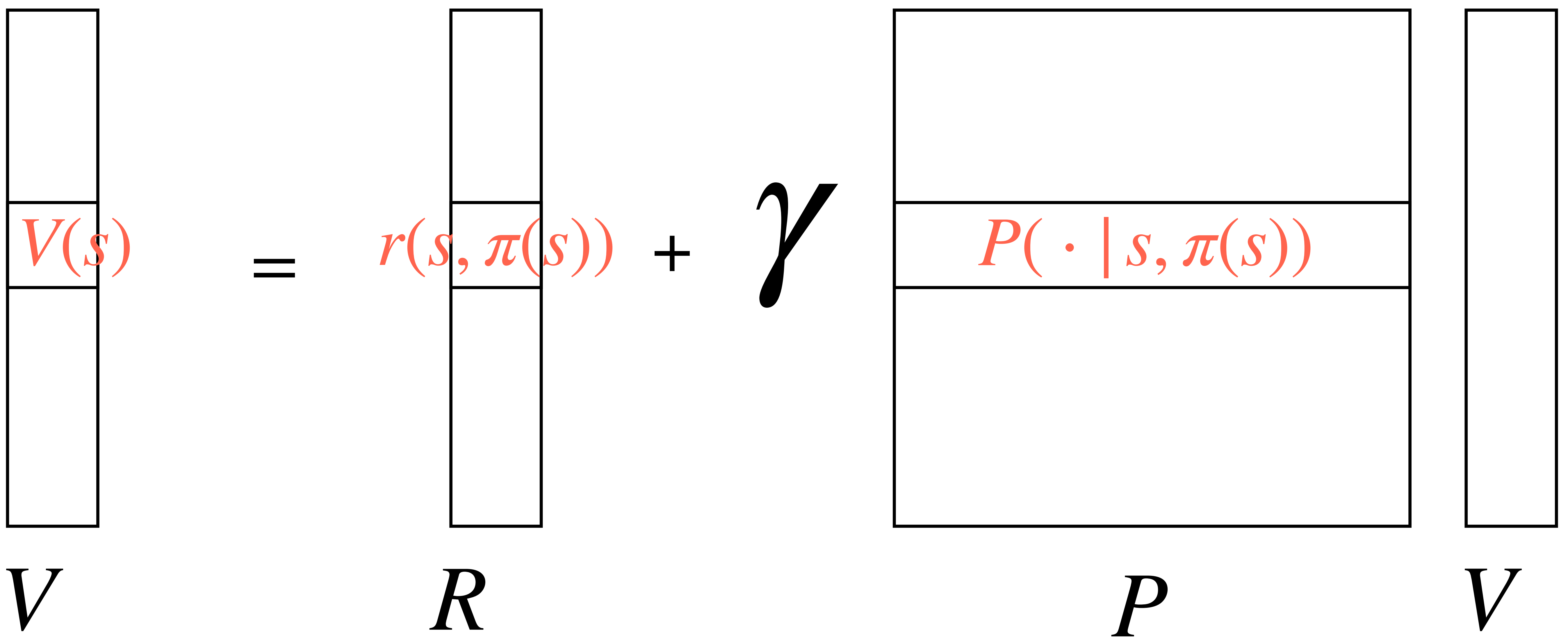
Value Iteration Algorithm:

1. Initialization: $V^0 : \|V^0\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]$

2. Iterate until convergence: $V^{t+1} \leftarrow \mathcal{T} V^t$

Exact Policy Evaluation: Matrix Version

- Define: $R \in \mathbb{R}^{|S|}$, where $R^\pi_s = r(s, \pi(s))$, and $P^\pi \in \mathbb{R}^{|S| \times |S|}$, where $P^\pi_{s,s'} = P(s' | s, \pi(s))$
- So we want to find $V \in \mathbb{R}^{|S|}$, s.t. $V = R^\pi + \gamma P^\pi V$



- **Algo:** compute $V = (I - \gamma P^\pi)^{-1} R^\pi$
  One can show that $I - \gamma P^\pi$ is full rank (thus invertible).
- **Runtime:** This approach runs in time $O(|S|^3)$.

An Iterative Version for Policy Eval

Algorithm (Iterative PE):

1. Initialization: $V^0 : \|V^0\|_\infty \in \left[0, \frac{1}{1-\gamma}\right]$
2. Iterate until convergence: $V^{t+1} \leftarrow R + \gamma P V^t$

Policy Iteration (PI)

- Initialization: choose a policy $\pi^0 : S \mapsto A$
- For $t = 0,1,\ldots$
  1. **Policy Evaluation**: compute $V^{\pi^t}(s)$ and $Q^{\pi^t}(s,a)$, where
  $$Q^{\pi^t}(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^{\pi^t}(s')$$
  2. **Policy Improvement**: set
  $$\pi^{t+1}(s) := \arg\max_a Q^{\pi^t}(s,a)$$

What's computation complexity per iteration?

$O(|S|^3 + |S|^2|A|)$

Recap

+

Finite Horizon MDPs

Finite horizon Markov Decision Process

$$\mathscr{M} = \{S, A, r, P, H\},$$
$$r : S \times A \mapsto [0,1], H \in \mathbb{N}, P : S \times A \mapsto \Delta(S)$$

Note that in finite horizon setting, we will consider time-dependent policies, i.e.,

$$\pi := \left\{\pi_0, \pi_1, \ldots, \pi_{H-1}\right\}, \pi_h : S \mapsto A, \forall h$$

Policy interacts with the MDP as follows to sample a trajectory $\tau = \left\{s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}\right\}$ as follows:

$$a_0 = \pi_0(s_0), s_1 \sim P(\cdot \mid s_0, a_0), a_1 = \pi_1(s_1), \ldots \quad s_h \sim P(\cdot \mid s_{h-1}, a_{h-1}), a_h = \pi_h(s_h) \ldots$$

V/Q functions in Finite horizon MDP

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| s_h = s\right]$$

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{\tau=h}^{H-1} r(s_\tau, a_\tau) \middle| (s_h, a_h) = (s, a)\right]$$

Bellman Consistency Equation:

$$Q_h^\pi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(s,a)}\left[V_{h+1}^\pi(s')\right]$$

Compute Optimal Policy via Dynamic Programming

DP is a backwards in time approach for computing the optimal policy:

$$\pi^\star = \{\pi_0^\star, \pi_1^\star, \ldots, \pi_{H-1}^\star\}$$

1. Start at $H - 1$,

$$Q_{H-1}^\star(s,a) = r(s,a) \qquad \pi_{H-1}^\star(s) = \arg\max_a Q_{H-1}^\star(s,a)$$

$$V_{H-1}^\star = \max_a Q_{H-1}^\star(s,a) = Q_{H-1}^\star(s, \pi_{H-1}^\star(s))$$

2. Assuming we have computed $V_{h+1}^\star$, $h \leq H - 2$, then:
(i.e. assuming we know how to perform optimally starting at $h + 1$)

$$Q_h^\star(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(s,a)} V_{h+1}^\star(s')$$

$$\pi_h^\star(s) = \arg\max_a Q_h^\star(s,a), \qquad V_h^\star = \max_a Q_h^\star(s,a)$$

Summary on Finite horizon MDP

$$\mathcal{M} = \{S, A, r, P, H\},$$
$$r : S \times A \mapsto [0,1], H \in \mathbb{N}, P : S \times A \mapsto \Delta(S)$$

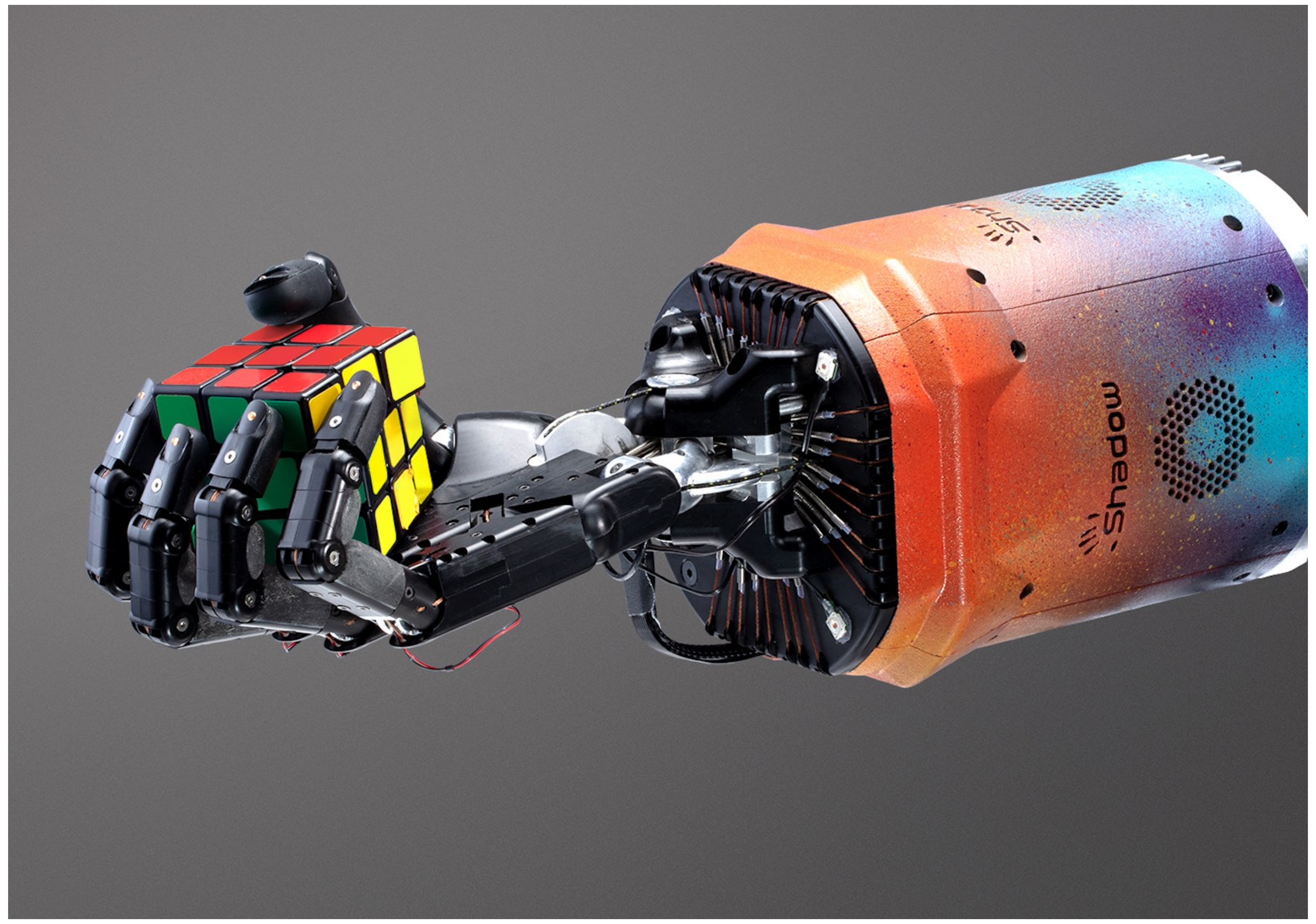
Comparing to the infinite horizon, discounted MDP:

1. Policy will be time dependent
2. DP takes H steps to compute $\pi^\star$
   - total computation time is $O(H|S|^2|A|)$
   - no need to use contraction argument and no discount factor
3. Extension to non-stationary setting works immediately:
   (i.e. with a non-stationary transition model: $P_0(s'|s,a), P_1(s'|s,a), \ldots P_{H-1}(s'|s,a)$)

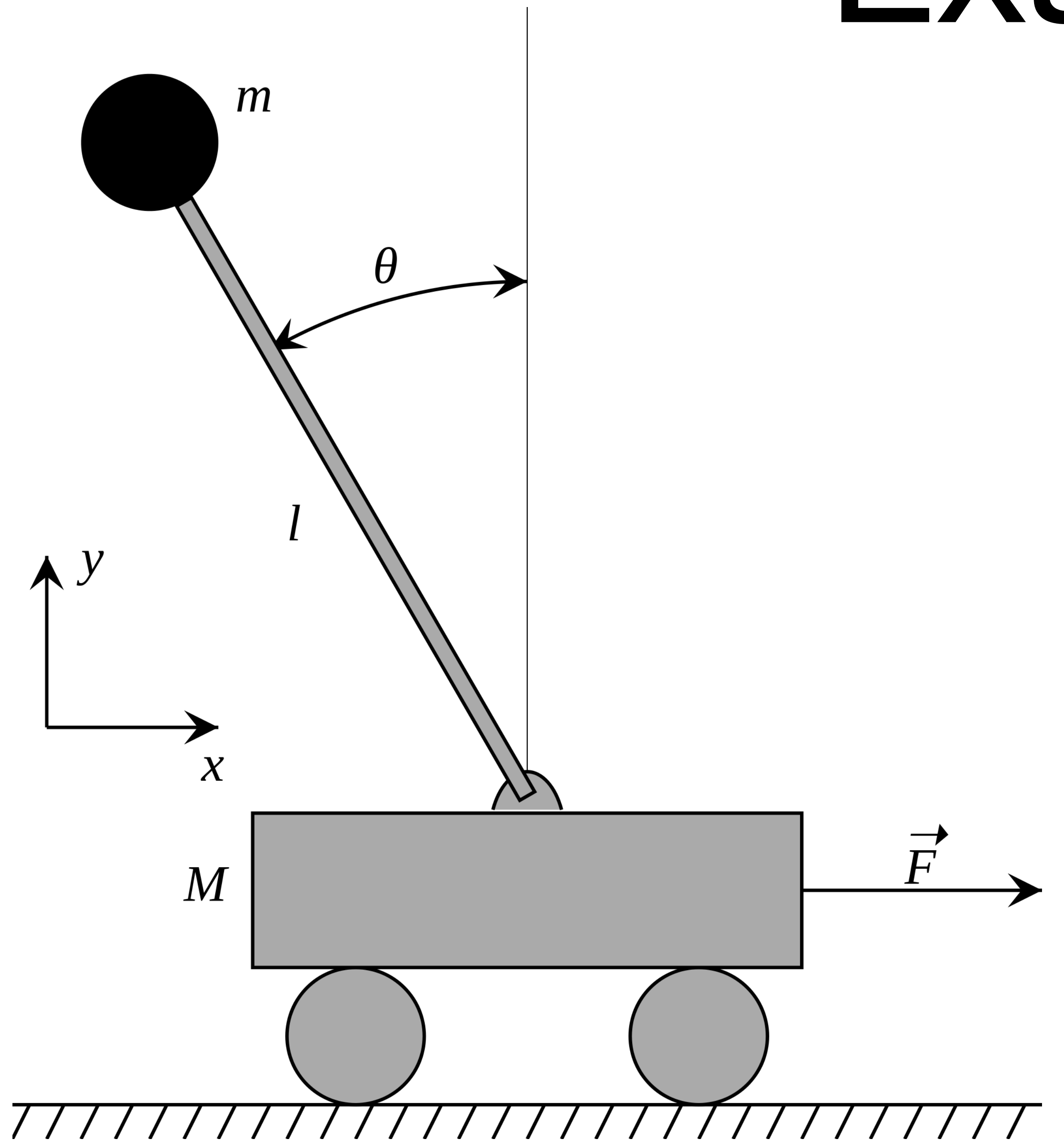
Today

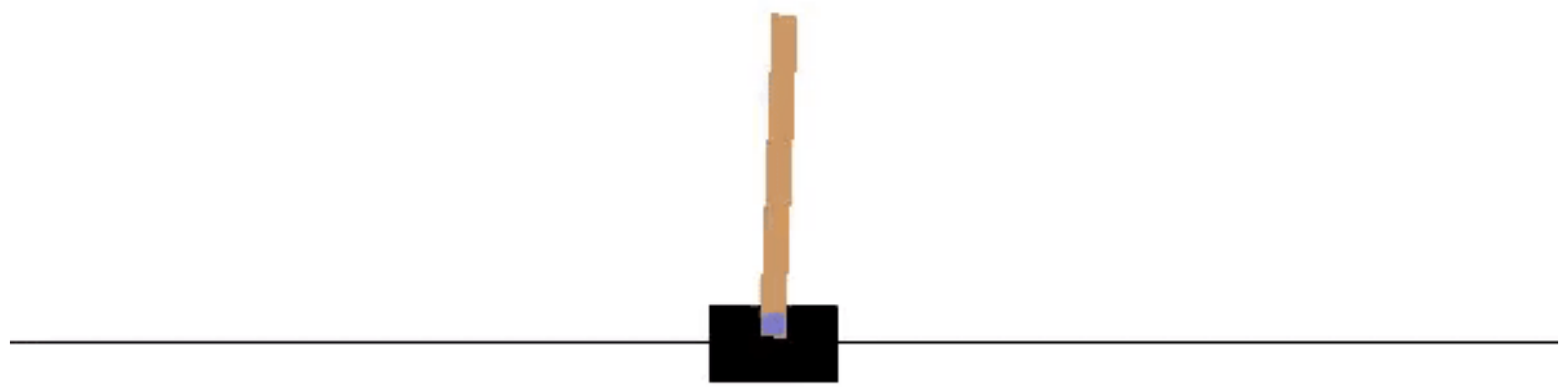
- Feedback from last lecture
- Recap
- Finite-horizon discrete MDPs
- General optimal control problem
- The linear quadratic regulator (LQR) problem

Robotics and Controls



Example: CartPole





State: position and velocity of the cart, angle and angular velocity of the pole

Control: force on the cart

Goal: stabilizing around the point $(s = s^\star, a = 0)$

$$c(s_t, a_t) = a_t^\top R a_t + (s_t - s^\star)^\top Q (s_t - s^\star)$$

Optimal control:

$$\min_{\pi_0, \ldots, \pi_{T-1}: S \to A} \mathbb{E}\left[\sum_{t=0}^{T-1} c(s_t, a_t)\right] \quad \text{s.t. } \; s_{t+1} = f(s_t, a_t),\ s_0 \sim \mu_0$$

More Generally: Optimal Control

General dynamical system is described as $s_{t+1} = f_t(s_t, a_t, w_t)$, where

- $s_t \in \mathbb{R}^d$ is the state which starts at initial value $s_0 \sim \mu_0$,
- $a_t \in \mathbb{R}^k$ is the control (action),
- w_t is the noise/disturbance,
- f_t is a function (the dynamics) that determines the next state $s_{t+1} \in \mathbb{R}^d$

Objective is to find control policy π_t which minimizes the total cost (finite horizon T),

$$\text{minimize } \mathbb{E}\left[c_T(s_T) + \sum_{t=0}^{T-1} c_t(s_t, a_t)\right]$$

$$\text{s.t. } s_{t+1} = f_t(s_t, a_t, w_t),\ a_t = \pi_t(s_t),\ s_0 \sim \mu_0$$

- Randomness (in the expectation) generally enters via w_t, e.g., $w_t \sim \mathcal{N}(0,\Sigma)$
- Note c_T separated out because by convention there is no a_T

Discretize to finite state/action spaces?

$$s \in \mathbb{R}^d, a \in \mathbb{R}^k$$

Idea: Round states and actions onto an ϵ-grid of their spaces; then use tools from finite MDPs

E.g., if $\epsilon = 0.01$, round s and a to 2 decimal places

Assuming state/control spaces are bounded, this makes both finite

Recall: VI/PI computation times scaled polynomially in $|S|$ and $|A|$

But curse of dimensionality means $|S|$ and $|A|$ will scale like $(1/\epsilon)^d$

E.g., $\epsilon = 0.01, d = k = 10$ gives $|S|^2|A|$ on the order of 10^{60}…

Even the idea of discretizing relies on continuity (i.e., rounding nearby values to the same grid point only works if system treats them nearly the same),

So why not rely on this more formally by assuming smoothness/structure on the dynamics f and cost c?

Today

- ✓ Feedback from last lecture
- ✓ Recap
- ✓ Finite-horizon discrete MDPs
- ✓ General optimal control problem
- The linear quadratic regulator (LQR) problem

The Linear Quadratic Regulator (LQR)

Linear dynamics: $s_{t+1} = f(s_t, a_t, w_t) = As_t + Ba_t + w_t$

Quadratic cost function: $c(s_t, a_t) = s_t^\top Q s_t + a_t^\top R a_t, \quad c_T(s_T) = s_T^\top Q s_T$

Gaussian noise: $w_t \sim \mathcal{N}(0, \Sigma)$

- Why not linear for c? Want it bounded below so we can minimize it
- $Q \in \mathbb{R}^{d \times d}$ and $R \in \mathbb{R}^{k \times k}$ are positive definite matrices
- $A \in \mathbb{R}^{d \times d}$, $B \in \mathbb{R}^{d \times k}$, $\Sigma \in \mathbb{R}^{d \times d}$ determine the dynamics
- Note lack of subscripts on c (except at T) and f: time-homogeneous

Is LQR useful?

Surprisingly yes, despite its simplicity!

Any smooth dynamics function is locally approximately linear, and any smooth function with a minimum is locally approximately quadratic near its minimum

E.g., think of heating/cooling a room: if done right, temperature should rarely deviate much from a fixed value, and shouldn't have to do too much heating or cooling, i.e., states and actions stay local to some fixed points!

In fact, because the LQR model is so well-studied in control theory, many human-engineered systems are designed to be approximately linear where possible

That said, it is indeed far too simple for many more complex (nonlinear) systems, yet in a couple lectures we will see how to extend its ideas to such systems to get surprisingly good solutions to apparently intractable nonlinear control problems

Example: 1-d Vehicle

Robot moving in 1-d by choosing to apply force a_t left (negative) or right (positive)

Newton: Force = mass x acceleration, so if vehicle mass = m, acceleration = $\frac{a_t}{m}$

If time steps are separated by δ (small), then we can approximate acceleration (derivative of velocity) by finite difference of velocities v_t:

$$\text{acceleration}_t = \frac{v_t - v_{t-1}}{\delta} = \frac{a_t}{m}$$

Same trick to approximate velocity (derivative of position) via positions p_t:

$$v_t = \frac{p_t - p_{t-1}}{\delta}$$

Preliminaries about LQR

Iterating the dynamics $s_t = As_{t-1} + Ba_{t-1} + w_{t-1}$ all the way back to time 0 gives:

$$s_t = A^t s_0 + \sum_{i=0}^{t-1} A^i (Ba_{t-i-1} + w_{t-i-1})$$

So assuming $\mathbb{E}[w_t] = 0$, we get $\mathbb{E}[s_t \mid s_0, a_0, \ldots, a_{t-1}] = A^t s_0 + \sum_{i=0}^{t-1} A^i Ba_{t-i-1}$

Looking ahead: we will show next lecture that the optimal control/policy is linear:

$$\pi_t^\star(s_t) = -K_t s_t$$

Plugging in gives a relatively simple form for the expected state:

$$\mathbb{E}[s_t \mid s_0, a_t = -K_t s_t] = (\Pi_{i=0}^{t-1}(A - BK_i))\, s_0$$

LQR Value and Q functions

Given a policy π (abbreviating a sequence of policies $\pi_0, \ldots, \pi_{t-1}$, one for each t),

Define the value function $V_t^\pi : \mathbb{R}^d \to \mathbb{R}$ as:

$$V_t^\pi(s) = \mathbb{E}\left[s_T^\top Q s_T + \sum_{i=t}^{T-1} (s_i^\top Q s_i + a_i^\top R a_i) \,\middle|\, a_i = \pi_i(s_i)\ \forall i \geq t,\ s_t = s\right]$$

and the Q function $Q_t^\pi : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ as:

$$Q_t^\pi(s, a) = \mathbb{E}\left[s_T^\top Q s_T + \sum_{i=t}^{T-1} (s_i^\top Q s_i + a_i^\top R a_i) \,\middle|\, a_t = a,\ a_i = \pi_i(s_i)\ \forall i > t,\ s_t = s\right]$$

Next time: we will solve for the optimal policy π via dynamic programming on these

Today

- Feedback from last lecture
- Recap
- Finite-horizon discrete MDPs
- General optimal control problem
- The linear quadratic regulator (LQR) problem

Today’s summary:

Finite-horizon discrete MDPs

- Solvable by dynamic programming

Optimal control problem

- Find optimal policy in MDP with infinite/continuous state and action spaces
- Requires some sort of structure

Linear quadratic regulator (LQR) problem

- Canonical problem in optimal control
- Linear dynamics, Gaussian errors, quadratic costs

Next time:

- Deriving the LQR optimal value and policy

1-minute feedback form: https://bit.ly/3RHtlxy