Optimal Control Theory and Linear Quadratic Regulators
Recap:
Finite horizon Markov Decision Process

\[ \mathcal{M} = \{S, A, r, P(H, \mu_0)\}, \]
\[ r : S \times A \mapsto [0,1], H \in \mathbb{N}^+, P : S \times A \mapsto \Delta(S), s_0 \sim \mu_0 \]

We need to consider time-dependent policies, i.e.,
\[ \pi := \{\pi_0, \pi_1, \ldots, \pi_{H-1}\}, \pi_h : S \mapsto A, \forall h \]

Policy interacts with the MDP as follows:
\[ \tau = \{s_0, a_0, s_1, a_1, \ldots, (H, a_H)\}, s_0 \sim \mu_0, a_0 = \pi_0(s_0), s_1 \sim P(\cdot | s_0, a_0), a_1 = \pi_1(s_1), \ldots \]
Recap: V/Q functions in Finite horizon MDP

\[ V^\pi_h(s) = \mathbb{E} \left[ \sum_{\tau=0}^{H-1} r(s_\tau, a_\tau) \mid s_h = s, a_\tau = \pi_\tau(s_\tau), s_{\tau+1} \sim P(\cdot \mid s_\tau, a_\tau) \right] \]

\[ Q^\pi_h(s, a) = \mathbb{E} \left[ \sum_{\tau=0}^{H-1} r(s_\tau, a_\tau) \mid (s_h, a_h) = (s, a), a_\tau = \pi_\tau(s_\tau), P \right] \]

Bellman Equation:

\[ Q^\pi_h(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(\cdot \mid s, a)} \left[ V^\pi_{h+1}(s') \right] \]
Example: Compute Optimal Policy

5 Acts: \( \uparrow \), \( \downarrow \), \( \rightarrow \), \( \leftarrow \), \(-

\( r((z, o), a) = 1, \forall a \)

\( r(s, a) = 0, \text{ otherwise} \)

\( r(s) = r(s, a), \forall a, \)

\( H = 1: \)

\( Q_{H-1}(s, a) = r(s, a) \)

\( V_{H-1}(s) = \max_a r(s, a) = r(s) \)

\( V_{H-1}(s) \)
Example (Continue)

$H-2$: focus on cell $\{3,4\}$

$Q^+_{H-2}(\{3,4\}, +) = 0 + V^+_{H-1}(\{2,4\})$

$Q^-_{H-2}(\{3,4\}, -) = 0 + V^-_{H-1}(\{3,4\})$

$V^+_{H-2}(\{3,4\}) = \max \left\{ Q^+_{H-2}(\{3,4\}, +) \right\}$
Recap: Formalizing the process

\[ \pi^* = \{\pi_0^*, \pi_1^*, \ldots, \pi_{H-1}^*\} \]

We use Dynamic Programming, and do DP backward in time; start at \( H - 1 \)

\[ Q_{H-1}^*(s, a) = r(s, a) \quad \pi_{H-1}^*(s) = \arg\max_a Q_{H-1}^*(s, a) \]

\[ V_{H-1}^* = \max_a Q_{H-1}^*(s, a) = Q_{H-1}^*(s, \pi_{H-1}^*(s)) \]
Recap: Formalizing the process

\[ \pi^* = \{\pi_0^*, \pi_1^*, \ldots, \pi_{H-1}^*\} \]

We use Dynamic Programming, and do DP backward in time; start at \( H - 1 \)

\[ Q_{H-1}^*(s, a) = r(s, a) \quad \pi_{H-1}^* = \arg \max_a Q_{H-1}^*(s, a) \]

\[ V_{H-1}^* = \max_a Q_{H-1}^*(s, a) = Q_{H-1}^*(s, \pi_{H-1}^*(s)) \]

Now assume that we have already computed \( V_{h+1}^*, h \leq H - 2 \) (i.e., we know how to perform optimally at \( h + 1 \))

\[ Q_h^*(s, a) = r(s, a) + \mathbb{E}_{s' \sim P(\cdot|s, a)} V_{h+1}^*(s') \]

\[ \pi_h^*(s) = \arg \max_a Q_h^*(s, a) \]
State-action Distributions

Given $\pi := \{ \pi_0, \ldots, \pi_{H-1} \}$
Define $\mathbb{P}_h^\pi(s, a; \mu_0)$: the probability of reaching $(s, a)$ at time step $h$ following $\pi$ from $\mu_0$
State-action Distributions

Given \( \pi := \{\pi_0, \ldots, \pi_{H-1}\} \)

Define \( \mathbb{P}_h^\pi(s, a; \mu_0) \): the probability of reaching \((s, a)\) at time step \(h\) following \(\pi\) from \(\mu_0\)

We define average state-action distribution as:

\[
d^\pi(s, a) = \frac{1}{H} \sum_{h=0}^{H-1} \mathbb{P}_h^\pi(s, a; \mu_0)
\]
State-action Distributions

Given $\pi := \{\pi_0, \ldots, \pi_{H-1}\}$
Define $P^\pi_h(s, a; \mu_0)$: the probability of reaching $(s, a)$ at time step $h$ following $\pi$ from $\mu_0$

We define average state-action distribution as:

$$d^\pi(s, a) := \frac{1}{H} \sum_{h=0}^{H-1} P^\pi_h(s, a; \mu_0)$$

Writing Expected total reward using $d^\pi(s, a)$:

$$\mathbb{E}_{s_0 \sim \mu_0} \left[ V^\pi_h(s_0) \right] = H \mathbb{E}_{s, a \sim d^\pi} \left[ r(s, a) \right]$$
Robotics and Controls

Dexterous Robotic Hand Manipulation
OpenAI, 2019
Example: CartPole

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

**Control**: force on the cart
Example: CartPole

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

Control: force on the cart
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 \((x = x^*, u = 0)\)

\[
c(x_h, u_h) = u_h^T R u_h + (x_h - x^*)^T Q (x_h - x^*)
\]

\(\Delta\) a positive definite
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 \((x = x^*, u = 0)\)

\[
c(x_h, u_h) = u_h^\top Ru_h + (x_h - x^*)^\top Q(x_h - x^*)
\]

Optimal control:

\[
\min_{\pi: X \rightarrow U} \mathbb{E} \left[ \sum_{h=0}^{H-1} c(x_h, u_h) \right], \text{s.t., } x_{h+1} = f(x_h, u_h), x_0 \sim \mu_0
\]
More Generally: Optimal Control

- A dynamical system is described as
  \[ x_{h+1} = f_h(x_h, u_h, w_h) \]
  where \( f_h \) maps a state \( x_h \in \mathbb{R}^d \), a control (the action) \( u_h \in \mathbb{R}^k \), and a disturbance \( w_h \), to the next state \( x_{h+1} \in \mathbb{R}^d \), starting from an initial state \( x_0 \sim \mu_0 \).
More Generally: Optimal Control

- a dynamical system is described as
  \[ x_{h+1} = f_h(x_h, u_h, w_h) \]
  where \( f_h \) maps a state \( x_h \in R^d \), a control (the action) \( u_h \in R^k \), and a disturbance \( w_h \), to the next state \( x_{h+1} \in R^d \), starting from an initial state \( x_0 \sim \mu_0 \).

- The objective is to find the control policy \( \pi \) which minimizes the long term cost,
  \[
  \min_{\pi} \mathbb{E}_\pi \left[ c_H(x_H) + \sum_{h=0}^{H-1} c_h(x_h, u_h) \right]
  \]
  such that \( x_{h+1} = f_h(x_h, u_h, w_h), u_h = \pi(x_h), x_0 \sim \mu_0 \)

where \( H \) is the time horizon and where \( w_h \) is either statistical (e.g., Gaussian noise) or deterministic (e.g., constant deviation).
Reduce Continuous control to Discrete MDP?

\[ x \in \mathbb{R}^d, u \in \mathbb{R}^k \]
Reduce Continuous control to Discrete MDP?

\[ x \in \mathbb{R}^d, u \in \mathbb{R}^k \]

Curse of dimensionality:
the number of discretized points are approximately \((1/\epsilon)^d + (1/\epsilon)^k\)
Today:
The LQR Model
The Linear Quadratic Regulator (LQR)

\[
\min_{\pi_0, \ldots, \pi_{H-1}} E \left[ x_H^T Q x_H + \sum_{h=0}^{H-1} (x_h^T Q x_h + u_h^T R u_h) \right]
\]

such that

\[
x_{h+1} = A x_h + B u_h + w_h, \quad u_h = \pi_h(x_h), \quad x_0 \sim \mu_0, \quad w_h \sim N(0, \sigma^2 I),
\]

Here, \( x_h \in \mathbb{R}^d, u_h \in \mathbb{R}^k, \)
The Linear Quadratic Regulator (LQR)

\[
\min_{\pi_0, \ldots, \pi_{H-1}} E \left[ x_H^T Q x_H + \sum_{h=0}^{H-1} (x_h^T Q x_h + u_h^T R u_h) \right]
\]

such that \( x_{h+1} = A x_h + B u_h + w_h, \quad u_h = \pi_h(x_h) \quad x_0 \sim \mu_0, \quad w_h \sim N(0, \sigma^2 I), \)

Here, \( x_h \in \mathbb{R}^d, \ u_h \in \mathbb{R}^k, \)

Studied often in theory, but less relevant in practice (\(?\))
(largely due to that time homogenous, globally linear models are rarely good approximations)
Example: 1-d Vehicle

\[ \dot{x} = \begin{bmatrix} p \\ v \end{bmatrix} \]

State \( x = (p, v) \), i.e., 1-d position and its velocity

[Example credit: Stanford EE 103]
Example: 1-d Vehicle

State $x = (p, v)$, i.e., 1-d position and its velocity

Control $u$, 1-d force, Friction force $-\eta v$, Vehicle mass $m$,
Example: 1-d Vehicle

State \( x = (p, v) \), i.e., 1-d position and its velocity

Control \( u \), 1-d force, Friction force \(-\eta v\), Vehicle mass \( m\),

Consider discrete time \( t = 0, 2\delta, 3\delta, \ldots, \) for small \( \delta \), we have:

[Example credit: Stanford EE 103]
Example: 1-d Vehicle

State $x = (p, v)$, i.e., 1-d position and its velocity

Control $u$, 1-d force, Friction force $-\eta v$, Vehicle mass $m$,

Consider discrete time $t = 0, 2\delta, 3\delta, \ldots$, for small $\delta$, we have:

$$m \frac{v_{h+1} - v_h}{\delta} \approx u - \eta v_h, \quad \frac{p_{h+1} - p_h}{\delta} \approx v_h$$

$$v_{h+1} = \frac{8}{m} \left( \frac{u - \eta v_h}{\Delta} \right) + v_h$$

$$p_{h+1} = s v_h + p_h$$

[Example credit: Stanford EE 103]
**V/Q functions:**

- Define the value function $V^\pi_h : \mathbb{R}^d \to \mathbb{R}$ as
  $$V^\pi_h(x) = \mathbb{E}\left[ x_H^T Q x_H + \sum_{t=h}^{H-1} (x_t^T Q x_t + u_t^T R u_t) \right | \pi, x_h = x],$$

- and the state-action value $Q^\pi_h : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ as:
  $$Q^\pi_h(x, u) = \mathbb{E}\left[ x_H^T Q x_H + \sum_{t=h}^{H-1} (x_t^T Q x_t + u_t^T R u_t) \right | \pi, x_h = x, u_h = u].$$
Optimal Value functions:

\[
V_h^*(x) = \max_{\pi_h, \pi_{h+1}, \ldots, \pi_{H-1}} \min_{h} \mathbb{E} \left[ x_H^T Q x_H + \sum_{i=h}^{H-1} x_i^T Q x_i + u_i^T R u_i \bigg| u_i = \pi_i(x_i), x_h = x \right]
\]
Optimal Value functions:

\[ V_h^*(x) = \max_{\pi_h, \pi_{h+1}, \ldots, \pi_{H-1}} \mathbb{E} \left[ x_H^T Q x_H + \sum_{i=h}^{H-1} x_i^T Q x_i + u_i^T R u_i \mid u_t = \pi_t(x_t), x_h = x \right] \]

**Theorem:**

\( V_h^* \) is a quadratic function, i.e., \( V_h^*(x) = x^T P_h x + p_h \), and optimal policy is linear:

\[ \pi_h^*(x) = -K_h^* x, \]

and \( P_h \) & \( K_h^* \) can be computed exactly.
Optimal Value functions:

$$V_h^*(x) = \max_{\pi_h, \pi_{h+1}, \ldots, \pi_{H-1}} \mathbb{E} \left[ x_H^T Q x_H + \sum_{i=h}^{H-1} x_i^T Q x_i + u_i^T R u_i \middle| u_t = \pi_t(x_t), x_h = x \right]$$

**Theorem:**

$V_h^*$ is a quadratic function, i.e., $V_h^*(x) = x^T P_h x + p_h$, and optimal policy is linear:

$$\pi_h^*(x) = -K_h^* x,$$

and $P_h$ & $K_h^*$ can be computed exactly

(Derivation? We will do it together next Tuesday)
Next Lecture:

How to compute the optimal policy in closed-form solution