RL 2022 - lec 1

Introduction, context

RL $\subset$ Optimal control

Agent $\longleftrightarrow$ Environment

Agent = Controller (decision maker)

Environment = System (dynamical system)

Terminology & notation:

x – state

u – action

y – observation

Remark:

s – state

a – action

o – observation

Environment evolution

$$x_{next} = f(x, u)$$

f – state transition law
(state dynamics function)

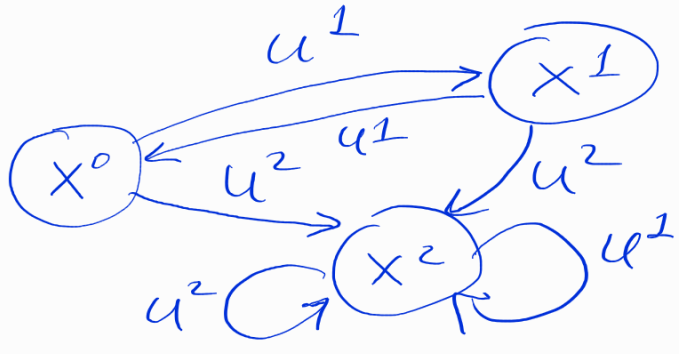
(Discrete-time, deterministic env.)

Now, stochastic case:

$$X_{next} \sim f(x_{next} \mid x, u)$$

like a probability of the event $X_{next} = x_{next}$ given that the env. was in the state x and action u was taken

Example: there is 3 states x^0, x^1, x^2 and two actions u^1, u^2



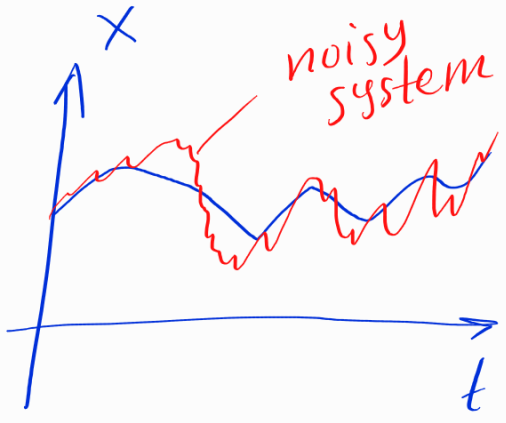
$$P[X = x^2 / X = x^1, U = u^2] = 50\%$$

$$x_{next} = f(x, u) + \eta$$

↑ noise term

$$\eta \sim N(0, \sigma_\eta^2)$$

↑ zero mean



When there is no designated time steps, time is treated continuous.

$$x_{next} = f(x, u)$$

$$\downarrow$$

diff-n w.r.t. time t $\rightarrow$ $\dot{x} = \frac{dx}{dt} = f(x, u)$

Case 1

$$x_{next} = f(x, u)$$

Start at x_0, x_5 - ?

$$x_1 = f(x_0, u_0)$$

$$x_2 = f(x_1, u_2) = f(f(x_0, u_0), \dots)$$

Case 2

$$\dot{x} = f(x, u)$$

What is the state after T seconds starting at x_0?

The answer:

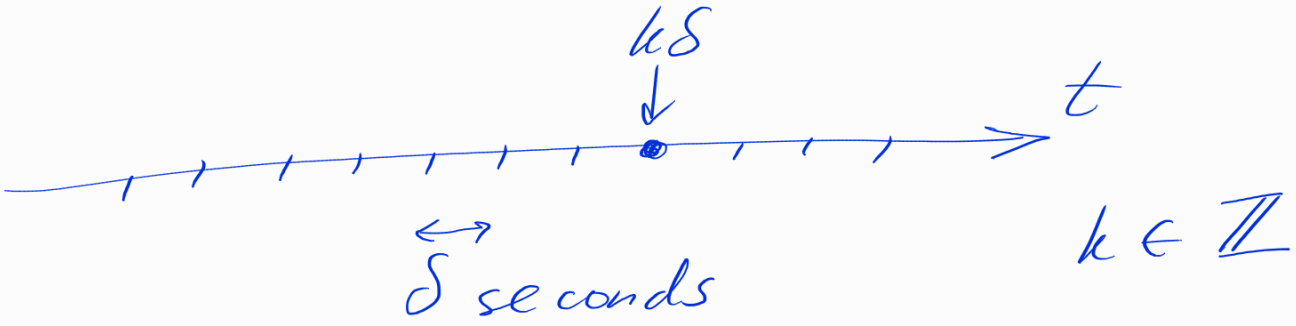
$$x(T) = x_0 + \int_0^T f(x(t), u(t))\, dt$$

scipy.integrate

(ode_solve)

How to get from a cont. time to a discrete time?

$$\dot{x} = f(x, u)$$



Euler discretization

$$x_{k+1} = \underbrace{x_k + \delta f(x_k, u_k)}$$

$$=: \hat{f}(x_k, u_k)$$

$$x_{k+1} = \hat{f}(x_k, u_k) + \eta_k$$

Why RL?

Optimal control (recall)

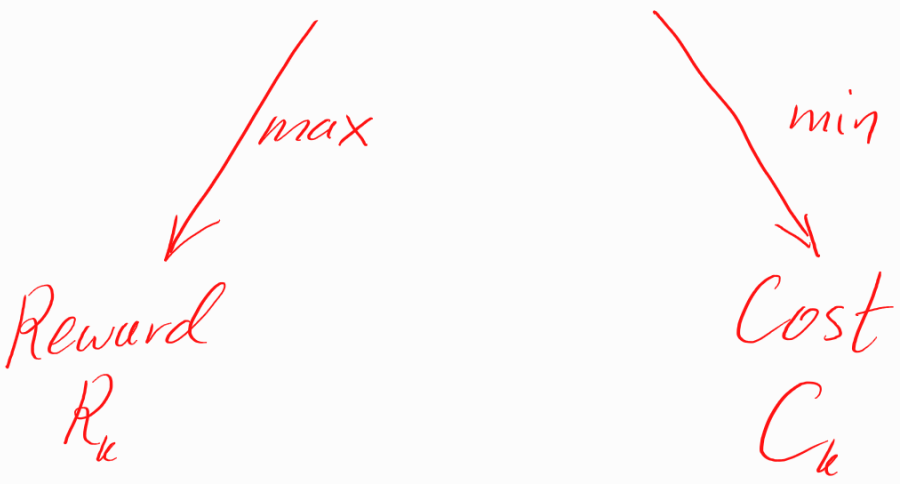
That is to say optimize

an objective

System: $X_{k+1} \sim f(x_{k+1} | x_k, u_k)$

In a moment in time (say, at some step k), there is a running objective

R_k



R is an instantaneous performance mark

(it answers the question: how good am I right now?)

Optimal control is usually concerned with optimization not just at a moment in

time, but in all moments in time.

This can be formulated as:

sequence of actions $\{u\}_0^T$

$$\max_{\rho} \sum_{k=0}^{T} R_k$$

where T may be ∞.

T ↓ horizon

ρ – policy! (control law, decision law, action plan etc.)

Deterministic	Stochastic
$\rho(x)$	$\rho(u/x)$

Let's recap what we have so far:

discount factor

$$\max_{\rho} \sum_{k=0}^{\infty} \mathbb{E}\left[\gamma^k R_k\right], \quad 0 < \gamma \leq 1$$

$$\text{s.t.} \quad X_{k+1} \sim f(x_{k+2} \mid x_k, u_k)$$

$$R_k \sim r(r_k \mid x_k, u_k)$$

$$U_k \sim \rho(u_k \mid x_k)$$

r – reward law

But, remember, the agent in general only sees an observation.

$$\max_{\rho} \sum_{k=0}^{\infty} \mathbb{E}\left[\gamma^k R_k\right]$$

$$\text{s.t.} \quad X_{k+1} \sim f(x_{k+2} \mid x_k, u_k)$$

$$R_k \sim r(r_k \mid x_k, u_k)$$

$$U_k \sim \rho(u_k \mid y_k)$$

$$Y_k \sim h(y_k \mid x_k)$$

For simplicity, assume $\gamma = 1$.

$$G_{k:k+N} = \sum_{i=k}^{k+N-1} R_i \quad \text{or}$$

$$G_k = \sum_{i=k}^{\infty} R_i \quad \text{is called an}$$

outcome

max → Return

min → Loss

Objective $\xrightarrow{\max}$ Value

Objective $\xrightarrow{\min}$ Cost-to-go

Value, Cost-to-go $\rightarrow \mathbb{E}[G_k]$

$$V^{\rho} = \mathbb{E}_{a_k, S_k \sim \rho}[G]$$

action sequence was distributed by ρ

$$V^* = \max_{\rho} V^{\rho}$$

optimal value

In max problems: J^{ρ}, J^*

Optimal policy: ρ^*

$$\rho^* = \arg\max_{\rho} V^{\rho}$$