

# Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 9, 2021

# Overview

- ▶ Vector norms
  - ▶ Induced matrix norms
- ▶ Spectral radius
  - ▶ Invertibility
- ▶ Neumann Lemma
- ▶ Markov Decision Processes
  - ▶ Optimality criteria
  - ▶ Vector notation

## Vector norms: $l_p$ -norms

An important class of vector norms on  $\mathbb{R}^n$  (or  $\mathbb{C}^n$ ) is the class of  $l_p$ -norms:

$$\|x\|_p \triangleq \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad 1 \leq p \leq \infty, \quad x \in \mathbb{R}^n.$$

# Norms of linear mappings

(This definition hinges on that of vector norms.)

**Definition (Operator norm).** Let  $\|\cdot\|$  be a norm on  $\mathbb{R}^n$  and  $\|\cdot\|'$  on  $\mathbb{R}^m$ . The operator norm of  $T \in L(\mathbb{R}^n, \mathbb{R}^m)$  with respect to  $\|\cdot\|$  and  $\|\cdot\|'$  is defined as:

$$\|T\| \triangleq \sup_{\|x\|=1} \|Tx\|' .$$

# Properties of matrix norms

- ▶  $\|A\| \geq 0$ ,  $\forall A \in L(\mathbb{R}^n, \mathbb{R}^m)$ ,  $\|A\| = 0$  only for  $A = 0$ , (positive, definite)
- ▶  $\|\alpha A\| = |\alpha| \|A\|$ ,  $\forall A \in L(\mathbb{R}^n, \mathbb{R}^m)$ , (homogeneous)
- ▶  $\|A + B\| \leq \|A\| + \|B\|$ ,  $\forall A, B \in L(\mathbb{R}^n, \mathbb{R}^m)$ , (sub-additive/triangle inequality)



If  $A, B \in L(\mathbb{R}^n, \mathbb{R}^n)$  (square matrices), we also have:

- ▶  $\|AB\| \leq \|A\| \|B\|$ , (sub-multiplicative)

## Matrix norms induced by $l_p$ vector norms

Let  $A \in L(\mathbb{R}^n, \mathbb{R}^m)$  with  $\mathbb{R}^n$  and  $\mathbb{R}^m$  normed with  $l_p$  norms for  $p = 1, 2, \infty$ .

- ▶  $\|A\|_1 \triangleq \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$ , (column-wise max over the column sums)
- ▶  $\|A\|_\infty \triangleq \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$ , (row-wise max over the row sums)
- ▶  $\|A\|_2 \triangleq \sqrt{\lambda}$ , where  $\lambda$  is the maximum eigenvalue of  $A^\top A$



If  $A \in L(\mathbb{R}^n, \mathbb{R}^n)$  is symmetric with eigenvalues  $\lambda_1, \dots, \lambda_n$ :

- ▶  $\|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i|$  .

## Matrix norms and spectral properties

If  $A \in L(\mathbb{C}^n, \mathbb{C}^n)$  with eigenpair  $u \in \mathbb{C}^n$ ,  $\lambda \in \mathbb{C}$ , then by sub-multiplicativity:

$$\|Au\| = \|\lambda u\| = |\lambda| \|u\| \leq \|A\| \|u\| .$$

Therefore,  $|\lambda| \leq \|A\|$  for any matrix norm and any eigenvalue  $\lambda$  of  $A$ .



This is true for any **consistent** norm. Any induced norm is a consistent norm.

# Spectral radius

**Definition (Spectral radius)** Let  $A \in L(\mathbb{C}^n, \mathbb{C}^n)$ , the **spectral radius** of  $A$   $\sigma(A)$  is the maximum of  $|\lambda_1|, \dots, |\lambda_n|$ , where  $\lambda_1, \dots, \lambda_n$  are eigenvalues of  $A$ .

**Lemma (Spectral radius and matrix norms)**  $\sigma(A) \leq \|A\|$



The spectral radius is not a norm. Not to confuse with the **spectral norm** which is used to refer to  $\|A\|_2$  (maximum eigenvalue of  $A^T A$ )



## Spectral radius

Example of a non-zero matrix whose spectral radius is zero:

$$A = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}, \quad \|A\|_1 = |\alpha| \quad \text{but} \quad \sigma(A) = 0.$$

**Lemma (Invertibility)** Let  $A \in L(\mathbb{C}^n)$ , if  $\sigma(A) < 1$ , then  $I - A$  is invertible.

**Proof** Let  $(x, \lambda)$  be an eigenpair of  $I - A$ :

$$(I - A)x = \lambda x \quad \Leftrightarrow \quad Ax = \underbrace{(1 - \lambda)}_{< 1 \text{ if } \sigma(A) < 1} x$$

If  $\sigma(A) < 1$ , then  $1 - \lambda < 1$  and  $\lambda > 0$  for any eigenvector. Therefore  $I - A$  has no zero eigenvalues and must be invertible.

# Invertibility of linear maps

**Theorem (Neumann Lemma)** Let  $A \in L(\mathbb{R}^n)$ , if  $\sigma(A) < 1$ , then

$$(I - A)^{-1} \text{ exists and } (I - A)^{-1} = \lim_{k \rightarrow \infty} \sum_{i=0}^k A^i.$$

- Proof**
1.  $I - A$  is invertible by above lemma.
  2. The inverse coincides with the series expansion.

$$\text{Note that } (I - A) \sum_{i=0}^{k-1} A^i = I - A^k$$

Therefore:

$$\sum_{i=0}^{k-1} A^i = (I - A)^{-1} - \underbrace{(I - A)^{-1} A^k}_{\text{vanishes}}$$

If  $\sigma(A) < 1$ , then  $\lim_{k \rightarrow \infty} A^k = 0$  (see 2.2.9 in O&R). So as  $k \rightarrow \infty$  the series converges to  $(I - A)^{-1}$ .

## Application to Markov Decision Processes

# Markov Decision Process

A framework for sequential decision making in discrete-time.

The main components of an MDP are:

- ▶ Set of states  $\mathcal{S}$
- ▶ Set of actions  $\mathcal{A}$  (can be state-dependent  $\mathcal{A}(s)$ )
- ▶ Transition probability function  $p(j|i, a), i, j \in \mathcal{S}, a \in \mathcal{A}$  such that  $\sum_{j \in \mathcal{S}} p(j|i, a) = 1$  (for discrete states)
- ▶ A reward function:  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ .

# Policy

**Definition** A **decision rule** is a prescription for the action choice in each state. They can be **deterministic** or **randomized**.

- ▶ If deterministic:  $d_t : \mathcal{S} \rightarrow \mathcal{A}(s)$ .
- ▶ If randomized:  $d_t : \mathcal{S} \rightarrow \text{Dist}(\mathcal{A}(\mathcal{S}))$



A decision rule can also be history-dependent or Markov. For now, we assume Markovian policies, but we'll have to motivate later when this choice is appropriate.

**Definition** A policy is a sequence of decision rules  $\pi = (d_1, \dots, d_T)$  (for a  $T$ -stages finite-horizon MDP). A policy can be **stationary** in which case  $\pi = (d, d, \dots)$  (also denoted by  $d^\infty$ )

# Taxonomy

We denote the possible set of policies by either:

- ▶ History-dependent Randomized (HR)
- ▶ History-dependent Deterministic (HD)
- ▶ Markovian Randomized (MR)
- ▶ Markovian Deterministic (MD)

# Criteria

Goal: find a **policy**  $\pi$  prescribing how to take actions across the state space so as to maximize some given criterion.

In order to talk about maximization, we also need to be able to **order** policies according to their performance. To do this we need to define the notion of **value of a policy**.

The **expected total reward** of a policy  $\pi \in \Pi^{\text{HR}}$  is:

$$v_{\pi}(s) \triangleq \lim_{T \rightarrow \infty} \mathbb{E}_{\pi} \left[ \sum_{t=1}^T r(S_t, A_t) \mid S_1 = s \right] .$$

## Expected total discounted reward

If we assume that  $\sup_{s \in \mathcal{S}} \sup_{a \in \mathcal{A}(s)} |r(s, a)| = M < \infty$ , and by introducing a **discount factor**  $\gamma \in [0, 1)$  then the limit exists and we can also interchange the limit and expectation and get:

$$v_{\gamma, \pi}(s) \triangleq \lim_{T \rightarrow \infty} \mathbb{E} \left[ \sum_{t=1}^T \gamma^{t-1} r(S_t, A_t) \right] = \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t) \mid S_1 = s \right]$$



## Optimal policies (expected total reward)

A policy  $\pi^* \in \Pi^{\text{HR}}$  is **total reward optimal** if:

$$v_{\pi^*}(s) \geq v_{\pi}(s), \quad \forall s \in \mathcal{S}, \pi \in \Pi^{\text{HR}} .$$

The **value of an MDP** (or the **optimal value function**) is defined as:

$$v^*(s) \triangleq \sup_{\pi \in \Pi^{\text{HR}}} v_{\pi}(s) .$$

## Optimal policies (expected total discounted reward)

A policy  $\pi^* \in \Pi^{\text{HR}}$  is **discount optimal** if:

$$v_{\pi^*, \gamma}(s) \geq v_{\pi, \gamma}(s), \quad \forall s \in \mathcal{S}, \pi \in \Pi^{\text{HR}} .$$

The **value of an MDP** (or the **optimal value function**) is defined as:

$$v_{\gamma}^*(s) \triangleq \sup_{\pi \in \Pi^{\text{HR}}} v_{\pi, \gamma}(s), \quad \forall s \in \mathcal{S}, \pi \in \Pi^{\text{HR}} .$$

## Vector notation

Let  $d \in \mathcal{D}^{\text{MD}}$  be a decision rule:

$$[r_d]_i \triangleq r(i, d(i)) \text{ and } [P_d]_{ij} = p(j|i, d(i))$$

If  $d \in \mathcal{D}^{\text{MR}}$  be a decision rule:

$$[r_d]_i \triangleq \sum_{a \in \mathcal{A}(i)} r(i, a) d(a|i) \text{ and } [P_d]_{ij} = \sum_{a \in \mathcal{A}(i)} p(j|i, a) d(a|i)$$