THE UNIVERSITY OF CHICAGO

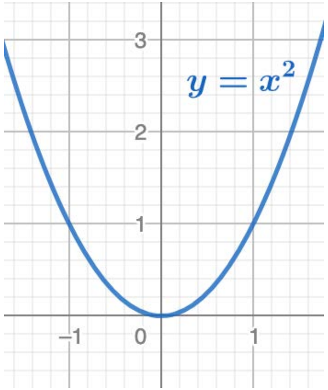
MSCA 37016
Advanced Linear Algebra for Machine Learning

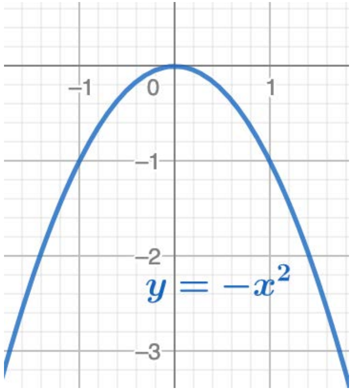
Lecture 5

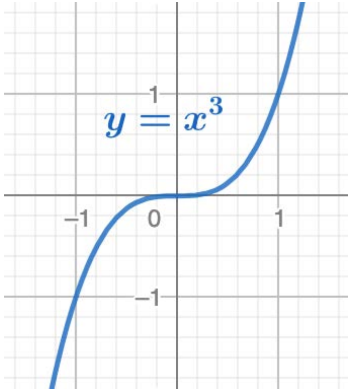
Danny Ng

Local Extremum of 1-Dimensional Function

- **1st Order Condition for Slope:** If a differentiable $f: \mathbb{R} \mapsto \mathbb{R}$ has a **local extremum** / optimum at an interior x^*, then $f'(x^*) = 0$ i.e. x^* is a **stationary / critical point**.



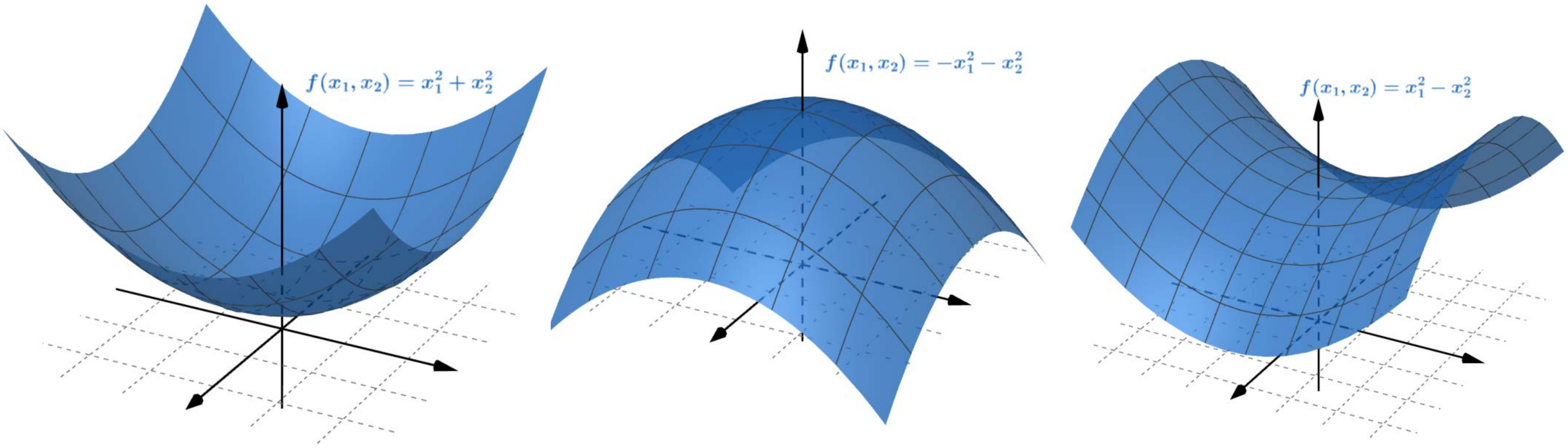




- **2nd Order Condition for Curvature:** Suppose $f'(x^*) = 0$.
  - If $f''(x^*) > 0$, then x^* is a local minimum i.e. **concave up**.
  - If $f''(x^*) < 0$, then x^* is a local maximum i.e. **concave down**.
  - If $f''(x^*) = 0$, then it is inconclusive since x^* can also be an **inflection / saddle point** i.e. curvature changes sign. (Need higher-order derivative test.)

Local Extremum of d-Dimensional Function

- Let $f: \mathbb{R}^d \mapsto \mathbb{R}$ i.e. $y = f(x_1, \dots, x_d)$ with input $\vec{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ and output $y \in \mathbb{R}$



- **1st Order Condition:** If an interior $\vec{x}^*$ is a local extremum, then all 1st order **partial derivatives**

$$\frac{\partial}{\partial x_1} f(\vec{x}^*) = 0 , \dots , \frac{\partial}{\partial x_d} f(\vec{x}^*) = 0$$

- Zero slope along each x_i-coordinate so f is overall "flat" at $\vec{x}^*$

2^{nd} Order Partial Derivative

- **Question:** What about 2^{nd} order condition to distinguish among concave up/down, saddle point, etc?

- **Answer:** Use 2^{nd} order partial derivatives

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \quad \text{for } i,j = 1, \dots, d$$

- **Fact:** If f is smooth enough (i.e. all 2^{nd} order partial derivatives are continuous), then

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

e.g. $f(x_1, x_2) = 3x_1^2 + 2x_1 x_2 + x_2^2 - 4x_1 + 2$

$$\frac{\partial f}{\partial x_1} = 6x_1 + 2x_2 - 4 \qquad \frac{\partial f}{\partial x_2} = 2x_1 + 2x_2$$

$$\frac{\partial^2 f}{\partial x_1^2} = \frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_1} = 6 \qquad \frac{\partial^2 f}{\partial x_2^2} = \frac{\partial}{\partial x_2}\frac{\partial f}{\partial x_2} = 2$$

$$\left.\begin{aligned} \frac{\partial^2 f}{\partial x_1 \partial x_2} &= \frac{\partial}{\partial x_1}\frac{\partial f}{\partial x_2} = 2 \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} &= \frac{\partial}{\partial x_2}\frac{\partial f}{\partial x_1} = 2 \end{aligned}\right\} \text{same}$$

Hessian Matrix

- **Hessian matrix** of a function $f: \mathbb{R}^d \mapsto \mathbb{R}$ is

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{pmatrix}$$

- $H \in \mathbb{R}^{d \times d}$ is usually a symmetric matrix!
- **Question:** How to define "H is positive or negative" when it is not a single number?

Matrix Definiteness

- Symmetric matrix $A \in \mathbb{R}^{d \times d}$ is **positive semi-definite** if

$$\overset{d}{{}^{1}\boxed{\vec{v}^T}} \times \overset{d}{{}^{d}\boxed{A}} \times \overset{1}{{}^{d}\boxed{\vec{v}}} \geq 0 \text{ for all } \vec{v} \in \mathbb{R}^d$$

- Furthermore, A is **positive definite** if

$$\vec{v}^T A \vec{v} > 0 \quad \text{whenever} \quad \vec{v} \neq \vec{0}$$

- Similarly for **negative semi-definite** and **negative definite**
- If A is none of the above, then it is **indefinite**

Matrix Definiteness vs Eigenvalue Spectrum

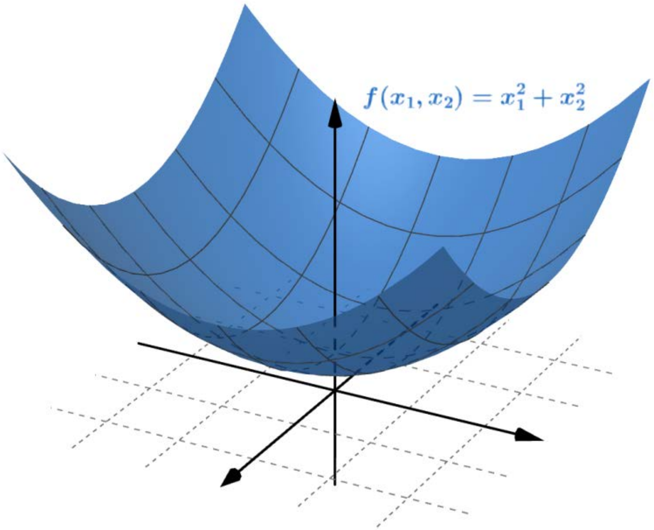
- Let $A = QDQ^T$ be its spectral decomposition

$$\hat{v}_i^T A \hat{v}_i = \underbrace{\begin{pmatrix} - & \hat{v}_i & - \end{pmatrix} \overbrace{\begin{pmatrix} | & & | \\ \hat{v}_1 & \dots & \hat{v}_d \\ | & & | \end{pmatrix}}^{Q}}_{\hat{e}_i^T = (0, \dots, 1, \dots, 0)} \overbrace{\begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}}^{D} \underbrace{\overbrace{\begin{pmatrix} - & \hat{v}_1 & - \\ & \vdots & \\ - & \hat{v}_d & - \end{pmatrix}}^{Q^T} \begin{pmatrix} | \\ \hat{v}_i \\ | \end{pmatrix}}_{\hat{e}_i} = \lambda_i$$

- If A is positive definite, then $\lambda_i > 0$ for all $i = 1, \dots, d$ (Converse is also true!)
- If A is positive semi-definite, then $\lambda_i \geq 0$ for all $i = 1, \dots, d$
- Similarly for negative definite and negative semi-definite
- A is indefinite iff its eigenvalues have mixed signs i.e. some $\lambda_i < 0$ and $\lambda_j > 0$

Spectrum of a symmetric matrix tells us about its definiteness

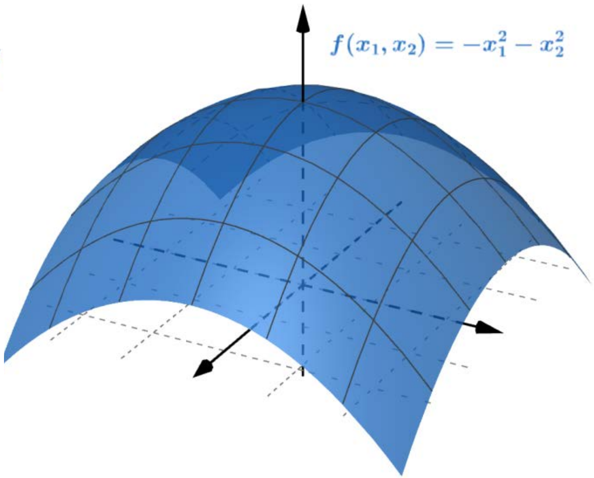
2nd Order Condition for d-Dimensional Function



Positive definite Hessian

$$H = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

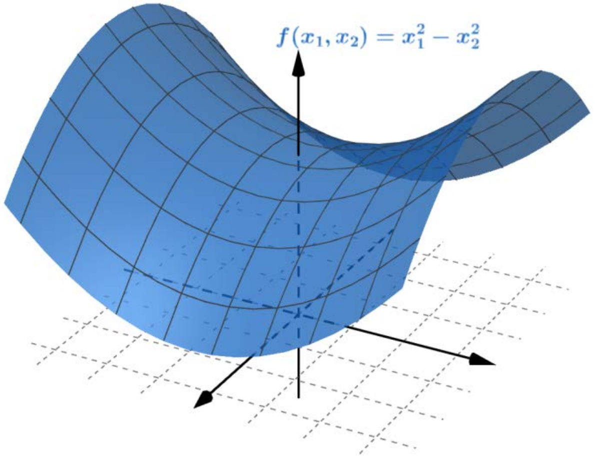
x^* is a local minimum i.e. **concave up**



Negative definite Hessian

$$H = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}$$

x^* is a local maximum i.e. **concave down**



Indefinite Hessian

$$H = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

x^* is a **saddle point**

Quadratic Form

- **Quadratic form** of a symmetric matrix $A \in \mathbb{R}^{d \times d}$ is a function $f: \mathbb{R}^d \mapsto \mathbb{R}$

2nd degree term

$$f(\vec{x}) = \vec{x}^T A \vec{x} = \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij} x_i x_j$$

- It is the high-dimensional analog of the parabola function $f(x) = ax^2$

e.g.

$$A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}$$

$$\begin{aligned} f(x_1, x_2) &= 4x_1x_1 + 1x_1x_2 + 1x_2x_1 + 3x_2x_2 \\ &= 4x_1^2 + 2x_1x_2 + 3x_2^2 \end{aligned}$$

Hessian of Quadratic Form

- The 2nd order partial derivatives are

$$\frac{\partial^2 f}{\partial x_i^2} = 2A_{ii}$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = A_{ij} + A_{ji} = 2A_{ij}$$

- So the Hessian matrix of f is exactly $2A$
- Therefore, we can understand more about definiteness of A by visualizing a "pure" quadratic function f closely related to A

Geometry of Quadratic Form

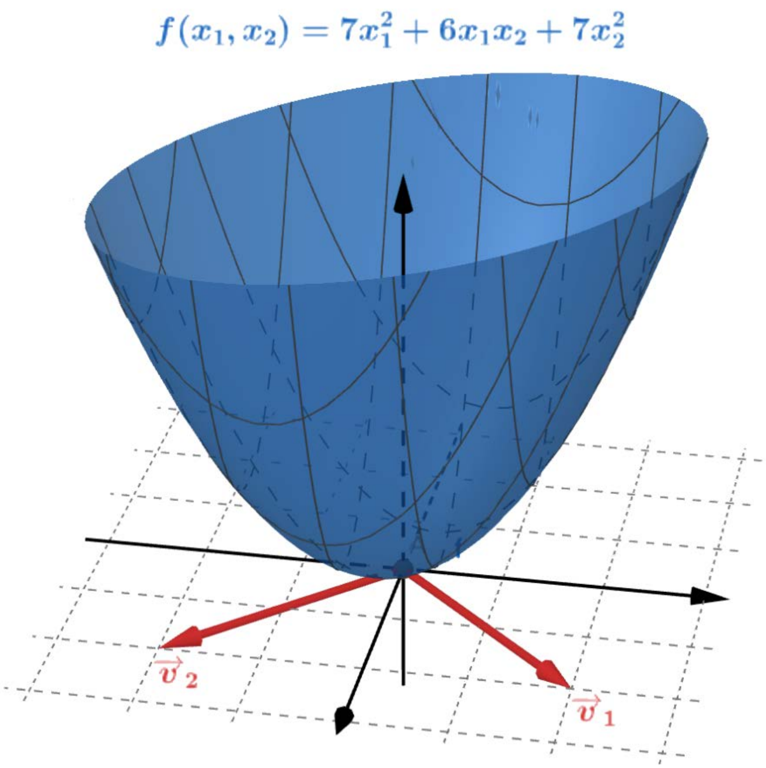
e.g. Positive definite matrix

$$A = \begin{pmatrix} 7 & 3 \\ 3 & 7 \end{pmatrix}$$

$$f(x_1, x_2) = \vec{x}^T A \vec{x} = 7x_1^2 + 6x_1x_2 + 7x_2^2$$

$$\hat{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad \text{with} \quad \lambda_1 = 10$$

$$\hat{v}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \quad \text{with} \quad \lambda_2 = 4$$



Geometry of Quadratic Form

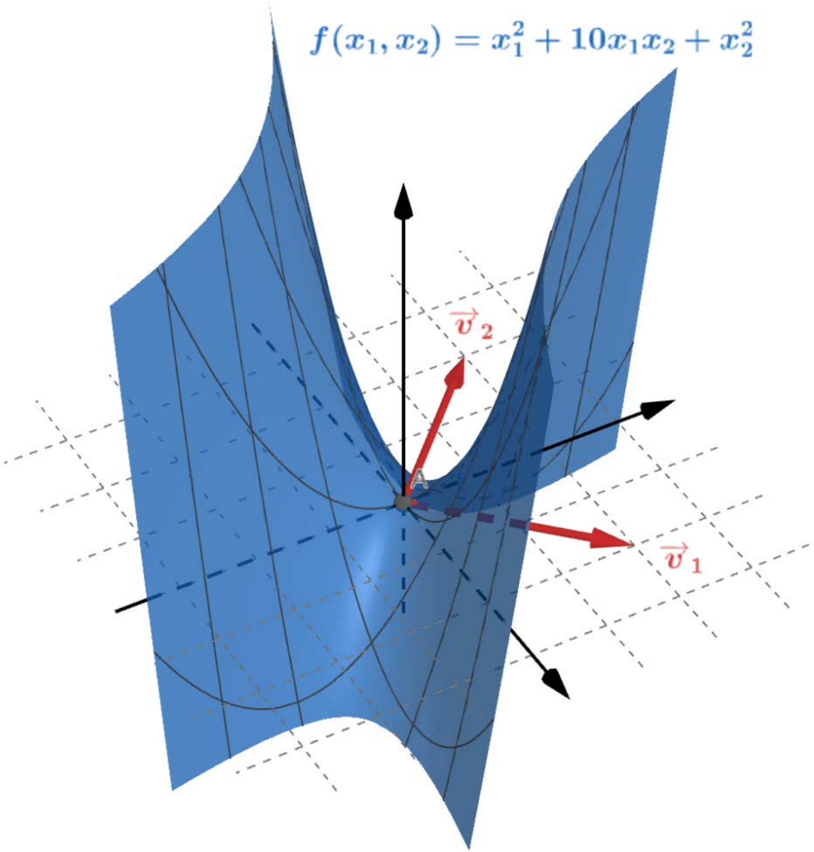
e.g. Indefinite matrix

$$A = \begin{pmatrix} 1 & 5 \\ 5 & 1 \end{pmatrix}$$

$$f(x_1, x_2) = \vec{x}^T A \vec{x} = x_1^2 + 10x_1x_2 + x_2^2$$

$$\hat{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad \text{with} \quad \lambda_1 = 6$$

$$\hat{v}_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \quad \text{with} \quad \lambda_2 = -4$$



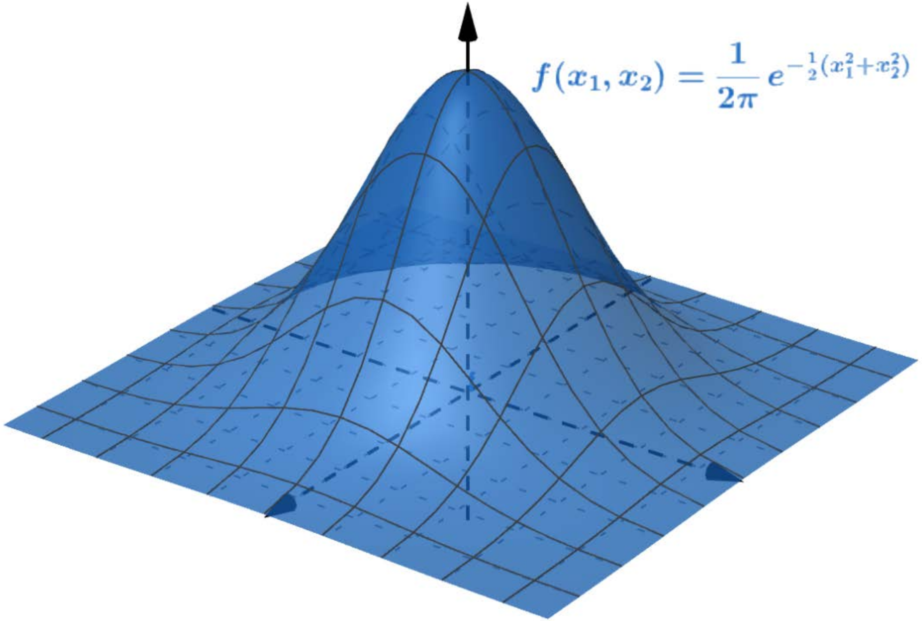
Normal Distribution Density

- Univariate normal $\mathcal{N}(\mu, \sigma^2)$ distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- Multivariate normal $\mathcal{N}_d(\vec{\mu}, C)$ distribution

$$f(\vec{x}) = \underbrace{\frac{1}{\sqrt{(2\pi)^d \det(C)}}}_{\text{density normalization constant}} e^{-\frac{1}{2}\underbrace{(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})}_{\text{quadratic form}}}$$



Gradient Vector

- The **gradient** / Jacobian / "derivative" of $f: \mathbb{R}^d \mapsto \mathbb{R}$ at point $\vec{x} \in \mathbb{R}^d$ is

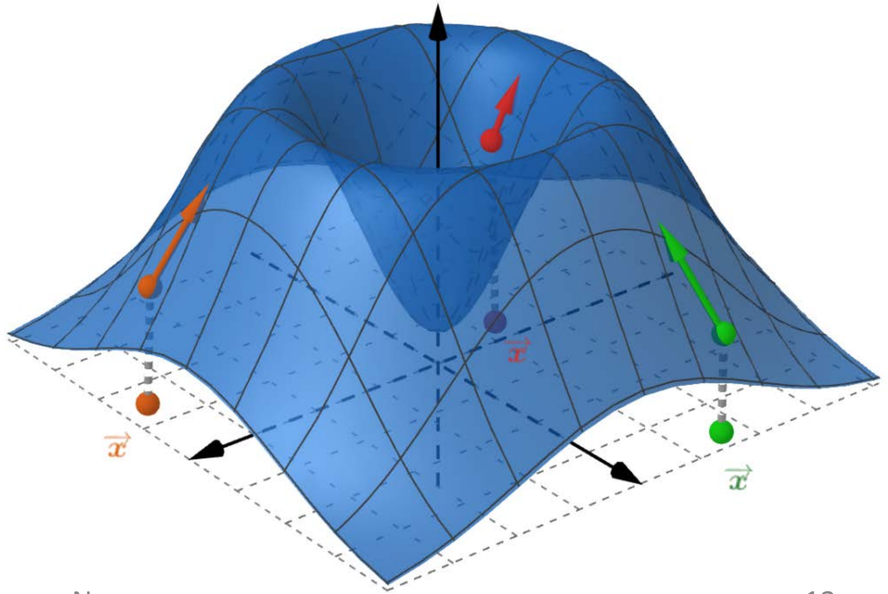
$$\nabla f(\vec{x}) = Df(\vec{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\vec{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\vec{x}) \end{pmatrix} \in \mathbb{R}^d$$

  i.e. vector of all 1st order partial derivatives

- Previously, $\nabla f(\vec{x}^*) = \vec{0}$ when $\vec{x}^*$ is a critical point

Geometry of Gradient Vector

- Gradient vector $\nabla f(\vec{x})$ represents:
  1. Direction of **steepest ascent** along the surface of f at point $\vec{x}$
  2. Length is steepness



Gradient Descent Algorithm

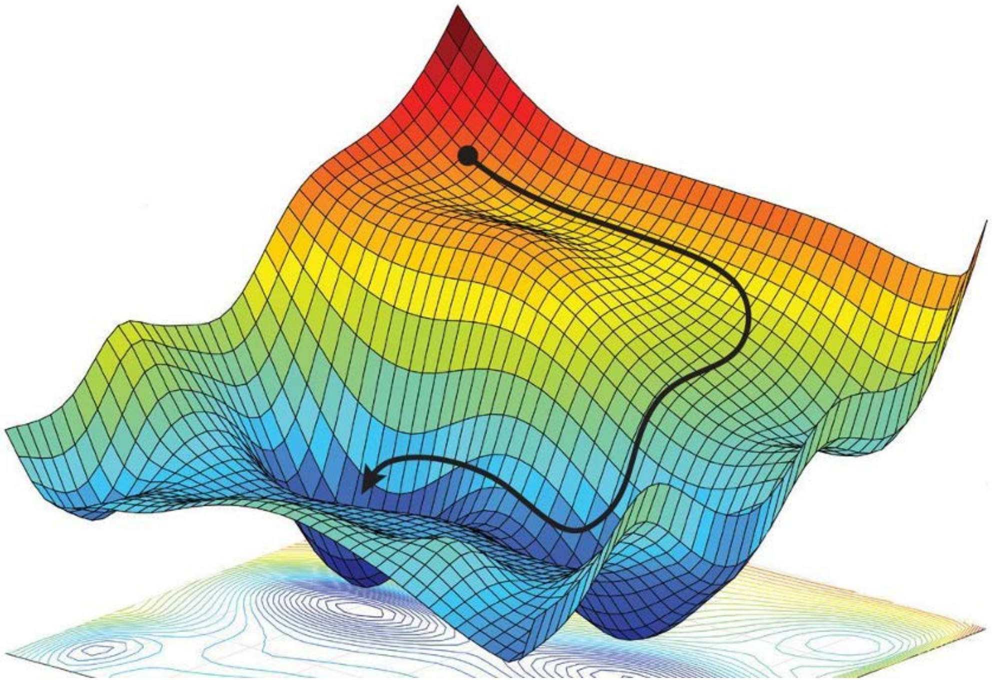
- Numerical minimization problem

$$\min_{\vec{x}} f(\vec{x})$$

- Iterative **gradient descent** algorithm
  1. Initialize $\vec{x}_0$ somehow
  2. Compute

$$\vec{x}_{t+1} = \vec{x}_t \underbrace{- \nabla f(\vec{x}_t)}_{\text{direction of steepest descent}}$$

- Application: Learning the weight parameters $\vec{w}$ of deep neural network models



Source: Alexander Amini, Daniela Rus. Massachusetts Institute of Technology, Adapted by M. Atarod / Science

Markov Chain

- Everyday, I decide among following 4 lunch options:
    1. McDonald's
    2. Pizza Hut
    3. Starbucks
    4. Leftover

- Let random variables $X_1, X_2, \ldots$ be the sequence of daily lunch choices

$$\underset{\substack{\uparrow \\ \textbf{discrete time} \\ t\,=\,1,2,\ldots}}{X_t} = \begin{pmatrix} \text{lunch choice} \\ \text{on } t^{th} \text{ day} \end{pmatrix} \in \underbrace{\{1, 2, 3, 4\}}_{\textbf{finite state space}}$$

Markov Property

- Assume **Markov property** i.e. tomorrow's choice depends only on today but not previous days

$$\mathbb{P}(X_{t+1} = j | \underbrace{X_t, X_{t-1}, \ldots, X_1}_{\text{all history}}) = \mathbb{P}(X_{t+1} = j | X_t\ )$$

e.g. If I choose to "not repeat what I just ate in the last 2 days", then it is not Markovian

Transition Probability Matrix

- Assume **time-homogenous** chain i.e. transition probability does not depend on time index t

$$\mathbb{P}(X_{t+1} = j | X_t = i) = p_{ij}$$

- Let **1-step transition probability matrix** be

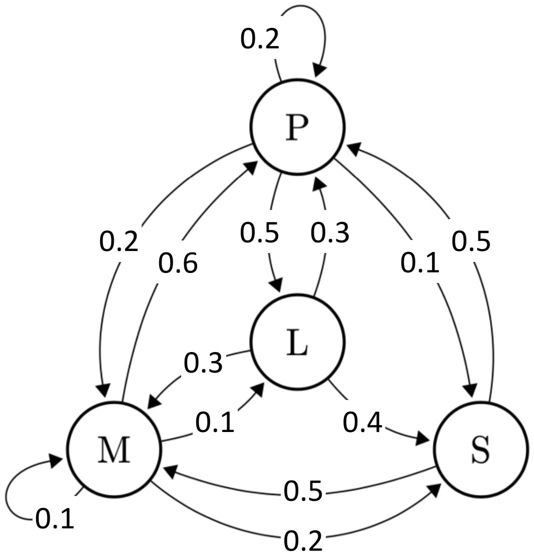
$$P = \begin{pmatrix} 0.1 & 0.6 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.3 & 0.4 & 0 \end{pmatrix} \in [0,1]^{d \times d}$$

row stochastic matrix
i.e. each row sum to 1

number of states

Lunch Example

e.g. McDonald's, Pizza Hut, Starbucks, leftover



Question: In the long run, how often do I visit each lunch option?

Chapman-Kolmogorov Equation

- Let 1-step transition probability matrix

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1d} \\ \vdots & \ddots & \vdots \\ p_{d1} & \cdots & p_{dd} \end{pmatrix} \in [0,1]^{d\times d}$$

- The 2-step transition probability matrix is $P^{(2)} = P \times P = P^2$. Similarly, the **$n$-step transition probability matrix** is

$$P^{(n)} = P \times \cdots \times P = P^n$$

- The 2-step transition probability is

$$P_{ij}^{(2)} = \mathbb{P}(X_{t+2} = j | X_t = i)$$

$$= \sum_{k=1}^{d} \mathbb{P}(X_{t+1} = k | X_t = i)\mathbb{P}(X_{t+2} = j | X_{t+1} = k)$$

$$= \underbrace{(p_{i1} \quad \cdots \quad p_{id})}_{i^{\text{th}} \text{ row of } P} \times \underbrace{\begin{pmatrix} p_{j1} \\ \vdots \\ p_{jd} \end{pmatrix}}_{j^{\text{th}} \text{ column of } P}$$

e.g.

$$P^{(4)} = \begin{pmatrix} 0.1 & 0.6 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.3 & 0.4 & 0 \end{pmatrix}^4$$

$$= \begin{pmatrix} 0.26 & 0.38 & 0.16 & 0.21 \\ 0.24 & 0.37 & 0.17 & 0.22 \\ 0.25 & 0.38 & 0.17 & 0.19 \\ 0.26 & 0.36 & 0.18 & 0.20 \end{pmatrix}$$

Initial Distribution at $t = 1$

- At initial time $t = 1$, let us impose

$$\mathbb{P}(X_1 = i) = p_i^{(1)}$$

$$\vec{p}^{(1)} = \underbrace{(p_1^{(1)}, \dots, p_d^{(1)})}_{\text{sum up to 1}} \in [0,1]^d$$

e.g. McDonald's on 1st day

$$\vec{p}^{(1)} = (1, 0, 0, 0)$$

e.g. Pick one at random uniformly

$$\vec{p}^{(1)} = (1/4, 1/4, 1/4, 1/4)$$

Distribution of X_2

- Marginalize the joint distribution

$$\mathbb{P}(X_2 = j) = \sum_{i=1}^{d} \mathbb{P}(X_1 = i)\mathbb{P}(X_2 = j | X_1 = i)$$

$$= \underbrace{\begin{pmatrix} p_1^{(1)} & \dots & p_d^{(1)} \end{pmatrix}}_{\text{distribution of } X_1} \times \underbrace{\begin{pmatrix} p_{j1} \\ \vdots \\ p_{jd} \end{pmatrix}}_{j^{\text{th}} \text{ column of } P} = p_j^{(2)}$$

- Distribution of X_2 is

$${}_{1}\overset{d}{\boxed{\vec{p}^{(1)}}} \times {}_{d}\overset{d}{\boxed{\;P\;}} = \underbrace{{}_{1}\overset{d}{\boxed{\vec{p}^{(2)}}}}_{\text{also sum up to 1}} \in [0,1]^{1\times d}$$

Distribution of X_t

- By induction or n-step transition probability matrix, X_t has distribution

$$\mathbb{P}(X_t = i) = p_i^{(t)}$$

$$\vec{p}^{(t)} = (p_1^{(t)}, \dots, p_d^{(t)}) \in [0,1]^d$$

$$1\ \overset{d}{\boxed{\vec{p}^{(1)}}} \times \overset{d}{\boxed{{}^{d}\ P^{t-1}}} = \overset{d}{\boxed{1\ \vec{p}^{(t)}}} \in [0,1]^{1\times d}$$

- We denote $X_t \sim \vec{p}^{(t)}$

Day 2 Lunch Example

e.g. McDonald's on 1st day

$$\vec{p}^{(1)} = (1, 0, 0, 0)$$

$$P = \begin{pmatrix} 0.1 & 0.6 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.3 & 0.4 & 0 \end{pmatrix}$$

$$\vec{p}^{(2)} = \vec{p}^{(1)} \times P = (0.1, 0.6, 0.2, 0.1)$$

e.g. Pick one at random uniformly

$$\vec{p}^{(1)} = (1/4, 1/4, 1/4, 1/4)$$

$$\vec{p}^{(2)} = \vec{p}^{(1)} \times P = (0.275, 0.4, 0.175, 0.15)$$

Stationary Distribution

- A **stationary** / **steady-state** / equilibrium distribution is an initial distribution

$$X_1 \sim \vec{\pi}$$

$$\vec{\pi} = (\pi_1, \dots, \pi_d) \in [0,1]^{1\times d}$$

that satisfies

$$\vec{\pi} \times P = \vec{\pi}$$

- In this case, $X_2 \sim \vec{\pi}$ and hence

$$X_t \sim \vec{\pi} \quad \text{for all } t = 1, 2, \dots$$

i.e. all have same distribution

Left Eigenvector of Eigenvalue 1

- Stationary condition says:
  1. $\vec{\pi}$ is a **left eigenvector** of P with eigenvalue 1
  2. We also need $\pi_1 + \dots + \pi_d = 1$

- In other words, column vector $\vec{\pi}^T \in [0,1]^{d\times 1}$ is a (right) eigenvector of P^T

$$P^T \times \vec{\pi}^T = \vec{\pi}^T$$

- **Fact**: If transition probability matrix P has any other (real) eigenvalue λ, then $|\lambda| < 1$

Stationary Lunch Example

e.g.

$$P = \begin{pmatrix} 0.1 & 0.6 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.3 & 0.4 & 0 \end{pmatrix}$$

- Using R to compute

```
v <- eigen(t(P))$vectors[,1]
v / sum(v)
```

- Only one stationary distribution

$$\vec{\pi} = (0.248, 0.371, 0.171, 0.210)$$

Interpretation of Stationary

- If I pick 1st day lunch randomly based on $X_1 \sim \vec{\pi}$ and follow transition probability matrix P, then each lunch X_t has same **marginal distribution**

$$X_t \sim \vec{\pi}$$

- But $X_1, X_2, \ldots$ are **not independent** e.g. if I eat leftover today, then no leftover for tomorrow

- Stationary is not same as choosing everyday lunch X_t from $\vec{\pi}$ separately

$$X_t \sim \vec{\pi} \quad \text{but not i.i.d.}$$

Countable State Space

- **Main questions:**
  1. Does stationary $\vec{\pi}$ always exist?
  2. In this case, is it unique?
  3. Does $\lim_{n\to\infty} P^n = P^\infty$ exist?
  4. Is this case, is $\vec{\pi}$ related to rows of P^∞?

- Finite state space is too simple to see general big picture

- We will also illustrate some **countably infinite state space** examples

Random Walk on $\mathbb{Z}$

e.g. Random walk on integers

$$\mathbb{Z} = \underbrace{\{0, \pm 1, \pm 2, \dots\}}_{\text{countable but infinite state space}}$$

- Let transition probability be

$$\left.\begin{aligned} \mathbb{P}(X_{t+1} = i+1 | X_t = i) &= p \\ \mathbb{P}(X_{t+1} = i-1 | X_t = i) &= 1-p \end{aligned}\right\} \text{sum to 1}$$

- So the chain can only move ± 1 unit to left or right each time

- Application: Brownian motion, stochastic calculus, financial derivative pricing

Accessible and Irreducible

- If we start at state i and there is a chance to visit state j eventually

$$(P^n)_{ij} > 0 \quad \text{for some step } n$$

then j is **accessible** from i.

$$i \to j$$

- If two states i and j are accessible from each other i.e. both $i \to j$ and $j \to i$, then they **communicate**.

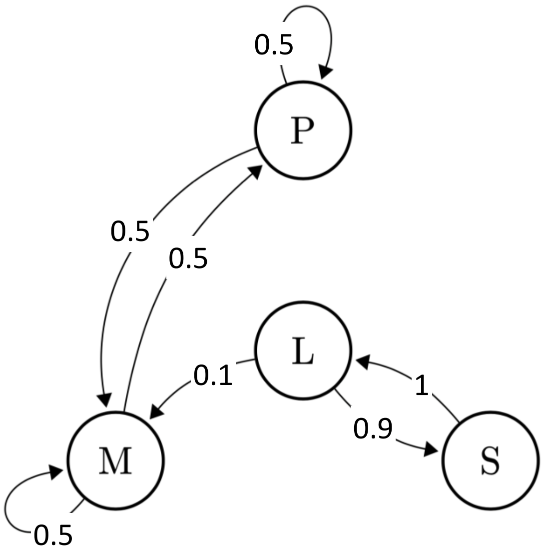
$$i \leftrightarrow j$$

- A Markov chain is **irreducible** if all states communicate with each other

Fast Food Junkie Example

e.g.

$$P_{\text{fast food}} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0.1 & 0 & 0.9 & 0 \end{pmatrix}$$



Absorbing State

- An **absorbing state** i has

$$\mathbb{P}(X_{t+1} = i | X_t = i) = P_{ii} = 1$$

For all other $j \neq i$

$$\mathbb{P}(X_{t+1} = j | X_t = i) = P_{ij} = 0$$

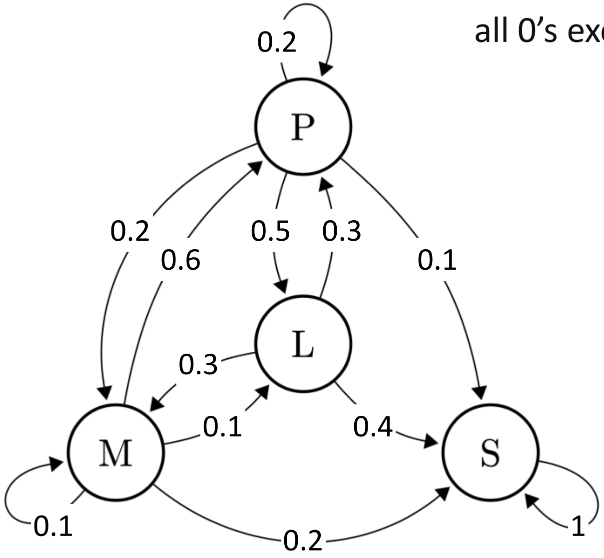
- **Facts:**
  1. Absorbing state does not communicate with other states
  2. Irreducible Markov chain cannot have an absorbing state

Latte-Holic Example

e.g. Once visit Starbucks, then latte everyday

$$P_{\text{latte}} = \begin{pmatrix} 0.1 & 0.6 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.5 \\ 0 & 0 & 1 & 0 \\ 0.3 & 0.3 & 0.4 & 0 \end{pmatrix}$$

all 0's except 1



Transient and Recurrent

- A state is **transient** if

$$\mathbb{P}(\text{never return}|X_1 = i) > 0$$

Otherwise, it is **recurrent** since

$$\mathbb{P}(\text{will return}|X_1 = i) = 1$$

- (For recurrent state) Let T_i be the time it takes to return. State i is **positive recurrent** if

$$\mathbb{E}(T_i) < \infty$$

Otherwise, $\mathbb{E}(T_i) = \infty$ and it is **null recurrent**

Examples

e.g. Fast food junkie

- Positive recurrent: McDonald's, Pizza Hut
- Transient: Starbucks, leftover

e.g. Latte-holic

- Positive recurrent: Starbucks
- Transient: (all other states)

e.g. Asymmetric random walk on $\mathbb{Z}$ i.e. $p \neq 1/2$

- All integers are transient

e.g. Symmetric random walk on $\mathbb{Z}$ i.e. $p = 1/2$

- All integers are actually null recurrent

Period and Aperiodic

- Recall the n-step return probability is

$$(P^n)_{ii} = \mathbb{P}(X_{t+n} = i | X_t = i)$$

- The **period** of state i is

GCD of $\{n \in \mathbb{N}: (P^n)_{ii} > 0\}$

greatest common denominator

- A state is **aperiodic** if it has period 1

e.g. McDonald's has self-loop so is period 1

$$\{n \in \mathbb{N}: (P^n)_{ii} > 0\} = \{1, \ldots\}$$

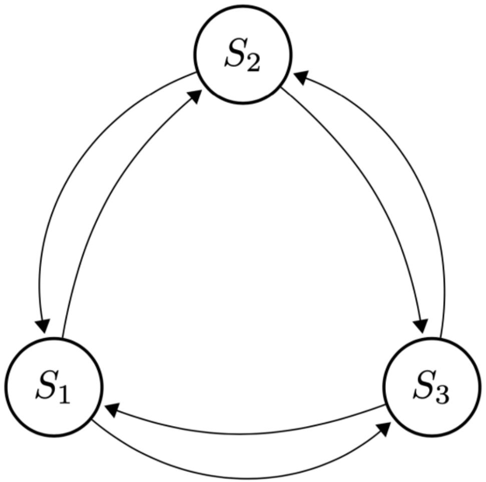
Examples

e.g. Random walk on $\mathbb{Z}$ is period 2 (all states)

$$\{n \in \mathbb{N}: (P^n)_{ii} > 0\} = \{2, 4, 6, \ldots\}$$

e.g. Random walk on 3-cycle is period 1

$$\{n \in \mathbb{N}: (P^n)_{ii} > 0\} = \{2, 3, \ldots\}$$



Shared Characteristic

- **Fact:** If two states communicate, then they are either:
  1. Both transient
  2. Both positive recurrent
  3. Both null recurrent

  Moreover, they have the same period.

e.g. In lunch example, Markov chain is irreducible. Since McDonald's has period 1, all states are aperiodic.

3 Types of Markov Chains

- **Thm 1:** For an irreducible Markov chain on countable state space, the states are either:
  1. All transient and no stationary $\vec{\pi}$
  2. All null recurrent and no stationary $\vec{\pi}$
  3. All positive recurrent and has unique stationary $\vec{\pi}$. In this case,

$$\pi_i = \underbrace{\lim_{t\to\infty} \frac{N_i^{(t)}}{t}}_{\text{long run frequency of visits to state } i} = \frac{1}{\mathbb{E}(T_i)} > 0$$

expected return time

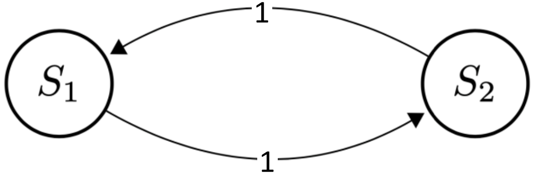
must be positive

- **Thm 2:** For finite state space, must be case #3

Oscillation and Ergordic

e.g.

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{period 2}$$



$$P^{2n} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad P^{2n+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

So P^n does not converge as $n \to \infty$

- A Markov chain is **ergodic** if it is positive recurrent and aperiodic

Convergence of P^n to $\vec{\pi}$

- **Thm 3:** For an irreducible Markov chain on countable state space, if it is ergordic (so $\vec{\pi}$ exists):

    1. n-step transition matrix must converge

    $$\lim_{n\to\infty} P^n = P^\infty$$

    2. All rows of matrix limit are the same

    $$\vec{\pi} = \text{any row of } P^\infty$$

    3. For any initial distribution $X_1 \sim \vec{p}^{(1)}$

    $$X_t \sim \vec{p}^{(1)} P^{t-1} \xrightarrow{\text{as } t \to \infty} \underbrace{\vec{p}^{(1)} P^\infty}_{\text{linear combination of all rows in } P^\infty} = \vec{\pi}$$

Detailed Balance Condition

- Recall 1-step transition probability

$$\mathbb{P}(X_{t+1} = j | X_t = i) = p_{ij}$$

- **Fact:** If $\vec{v} = (v_1, \dots, v_d) \in [0,1]^d$ with $v_1 + \cdots + v_d = 1$ satisfies the **detailed balance equations**

$$v_i \times p_{ij} = v_j \times p_{ji} \quad \text{for all } \; i, j$$

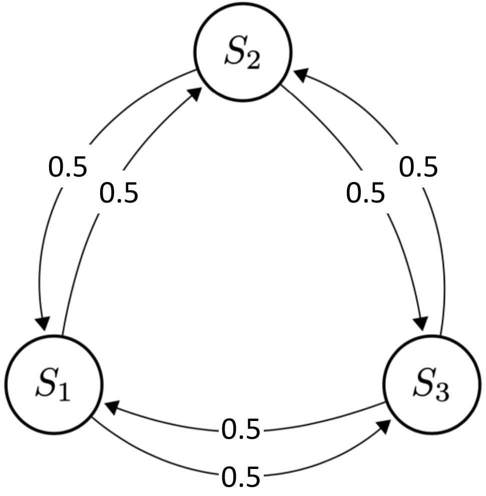
then $\vec{v}$ must be a stationary distribution

Proof?

$$v_i = v_i \times \sum_{j=1}^{d} p_{ij} = \underbrace{\sum_{j=1}^{d} v_j \times p_{ji}}_{\text{dot product of } \vec{v} \text{ with } i^{\text{th}} \text{ column of } P}$$

- In this case, the Markov chain is called **reversible**

e.g. Symmetric random walk on 3-cycle



So $v_i = 1/3$ for $i = 1, \dots, 3$

Markov Chain Monte Carlo

- In Bayesian statistics, we are often given some probability distribution (say $\vec{\pi}$) and need to generate a sample

$$X_1, X_2, \ldots \sim \vec{\pi} \quad \text{approx i.i.d.}$$

- **Metropolis-Hastings / Gibbs sampling** algorithms:

1. Construct transition probability P to satisfy detailed balance condition for $\vec{\pi}$
2. Use P as Markov chain to generate a sample $X_1, X_2, \ldots$
3. We know $X_t \approx \vec{\pi}$ for all large t

Markov Chain Applications

e.g. Sequence modeling

- Hidden Markov model HMM (1989) for speech recognition
- Text prediction / generation e.g. Mark V. Shaney

e.g. Markov decision process

- Markov chain + action + reward
- Bellman equation, optimal policy
- Reinforcement learning

e.g. Google's PageRank (1996)

- Graph of connected web pages with popularity

Low-Rank Matrix Approx.

- Let $A \in \mathbb{R}^{n \times d}$. We wish to find some matrix $B \in \mathbb{R}^{n \times d}$ such that:

  1. Low rank
  $$\operatorname{rank}(B) = k \ll \min(n, d)$$
  2. Approximation
  $$A \approx B$$

- The **Frobenius norm** of a matrix is

$$\|A\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{d} a_{ij}^2\right)^{1/2}$$

Frobenius Norm of Matrix

- Find the "best" rank-k matrix approximation

$$\min_{\operatorname{rank}(B)=k} \|A - B\|_F$$

- **Fact:** Frobenius norm is related to singular values

$$\|A\|_F^2 = \operatorname{tr}(A^T A) = \sum_{i=1}^{\overset{\operatorname{rank}(A)}{r}} \sigma_i^2$$

Proof? Let $A = UDV^T$ be its SVD. Then

$$\begin{aligned} \operatorname{tr}(A^T A) &= \operatorname{tr}(VDU^T UDV^T) \\ &= \operatorname{tr}(V^T V D^2) = \operatorname{tr}(D^2) \end{aligned}$$

Truncation of Singular Value Decomposition

- Let $A = UDV^T$ be its SVD with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$

$$A = \underbrace{\begin{pmatrix} | & & | \\ \hat{u}_1 & \dots & \hat{u}_r \\ | & & | \end{pmatrix}}_{U \in \mathbb{R}^{n\times r}} \underbrace{\begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r \end{pmatrix}}_{D \in \mathbb{R}^{r\times r}} \underbrace{\begin{pmatrix} - & \hat{v}_1 & - \\ & \vdots & \\ - & \hat{v}_r & - \end{pmatrix}}_{V^T \in \mathbb{R}^{r\times d}} = \sum_{i=1}^{r} \sigma_i \underbrace{\begin{pmatrix} | \\ \hat{u}_i \\ | \end{pmatrix} \begin{pmatrix} - & \hat{v}_i & - \end{pmatrix}}_{\text{rank-1 matrix}}$$

$$A \approx \underbrace{\sum_{i=1}^{k} \sigma_i \begin{pmatrix} | \\ \hat{u}_i \\ | \end{pmatrix} \begin{pmatrix} - & \hat{v}_i & - \end{pmatrix}}_{\text{rank-}k\text{ matrix}} = \underbrace{\begin{pmatrix} | & & | \\ \hat{u}_1 & \dots & \hat{u}_k \\ | & & | \end{pmatrix}}_{\in \mathbb{R}^{n\times k}} \underbrace{\begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_k \end{pmatrix}}_{\in \mathbb{R}^{k\times k}} \underbrace{\begin{pmatrix} - & \hat{v}_1 & - \\ & \vdots & \\ - & \hat{v}_k & - \end{pmatrix}}_{\in \mathbb{R}^{k\times d}} = B$$

Latent Factor Analysis

- Let $X \in \mathbb{R}^{n \times d}$ be the movie review ratings by n people on d movies

- **Latent factor model** assumption:
  1. Each person's representation
     $$\vec{p}_i \in \mathbb{R}^k$$
  2. Each movie's representation
     $$\vec{m}_j \in \mathbb{R}^k$$
  3. Movie rating is
     $$x_{ij} = \underbrace{\vec{p}_i \cdot \vec{m}_j}_{\text{combining factors}} + \underbrace{e_{ij}}_{\text{unexplained / noise}}$$

- Wish to learn low-dimensional hidden factors $\vec{p}_i$'s and $\vec{m}_j$'s from data X

- Data matrix X can be decomposed into

$$\underset{n \times d}{X} = \underbrace{\underset{n \times k}{P} \times \underset{k \times d}{M}}_{\text{rank-}k\text{ approximation of }X} + \underbrace{\underset{n \times d}{E}}_{\text{noise}}$$

- Use cases:
  1. Movie similarity $\left\| \vec{m}_{j_1} - \vec{m}_{j_2} \right\|$
  2. Predict rating for unwatched movie $\vec{m}$

- Applications: Collaborative filtering, latent semantic analysis LSA on document-word matrix