Review of Mathematical Foundation – Part 1

Yan Liu
Thomas Lord Department of Computer Science
University of Southern California

Credits to Joseph Chuang-Chieh Lin

Logistics

- Course project and TA preference: https://forms.gle/49emCRwYXdzXeRja9

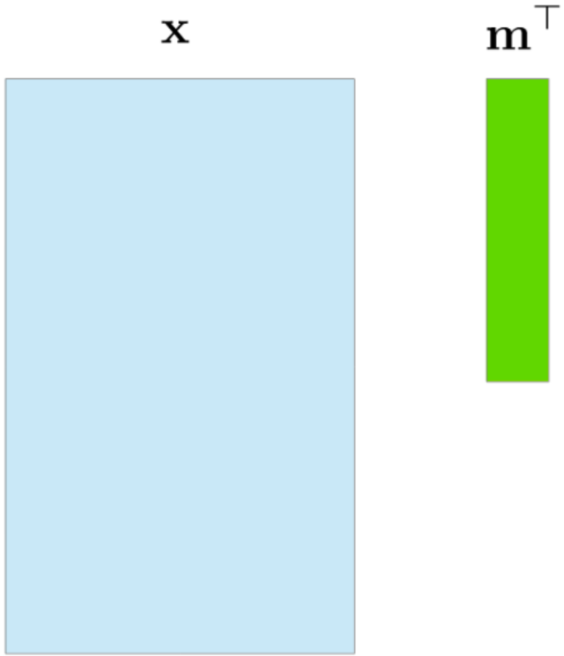
- Piazza is mostly set-up. Questions will be answered shortly

Linear Algebra

Vectorization Example (1/3)

$$
\begin{aligned}
y_i &= \langle \mathbf{m}, \mathbf{x}_i \rangle \\
&= m_1 x_{i,1} + m_2 x_{i,2} + \ldots + m_k x_{i,k}.
\end{aligned}
$$

```
m = np.random.rand(1,5)
x = np.random.rand(5000000,5)
#assume k=5
```



Vector Space

Vector Space

A real-valued vector space $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations:

$$+ : \mathcal{V} \times \mathcal{V} \mapsto \mathcal{V}$$

$$\cdot : \mathbb{R} \times \mathcal{V} \mapsto \mathcal{V}$$

where

- $(\mathcal{V}, +)$ is an Abelian group.
- Distributivity holds:
  - $\forall \lambda \in \mathbb{R}$, $\mathbf{x}, \mathbf{y} \in \mathcal{V}$: $\lambda \cdot (\mathbf{x} + \mathbf{y}) = \lambda \cdot \mathbf{x} + \lambda \cdot \mathbf{y}$.
  - $\forall \lambda, \psi \in \mathbb{R}$, $\mathbf{x} \in \mathcal{V}$: $(\lambda + \psi) \cdot \mathbf{x} = \lambda \cdot \mathbf{x} + \psi \cdot \mathbf{x}$.
- $\forall \lambda, \psi \in \mathbb{R}$, $\mathbf{x} \in \mathcal{V}$: $\lambda \cdot (\psi \cdot \mathbf{x}) = (\lambda\psi) \cdot \mathbf{x}$.
- $\forall \mathbf{x} \in \mathcal{V}$: $1 \cdot \mathbf{x} = \mathbf{x}$.

$\star$ Note: A vector multiplication is not defined.

Linear Combination

Linear Combination

Consider a vector space V and a finite number of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in V$. Then, every $\mathbf{v} \in V$ of the form

$$\mathbf{v} = \lambda_1 \mathbf{x}_1 + \cdots \lambda_k \mathbf{x}_k = \sum_{i=1}^{k} \lambda_i x_i \in V$$

with $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ is a linear combination of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

- **Question**: How to represent $\mathbf{0}$ as a linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$?

Linearly Independent

Linear (In)dependence

Consider a vector space V with $k > 0$ vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in V$.

- If there is a nontrivial linear combination such that $\mathbf{0} = \sum_{i=1}^{k} \lambda_i x_i$ with at least one $\lambda_i \neq 0$, then we say $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly dependent.
- If only the trivial solution exists (i.e., $\lambda_1 = \lambda_2 = \cdots = \lambda_k = 0$), then we say $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are linearly independent.

Remark (1/2)

Consider a vector space V with k linear independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_k$ and m linear combinations

$$\begin{aligned}
\mathbf{x}_1 &= \sum_{i=1}^{k} \lambda_{i,1} \mathbf{b}_i \\
&\vdots \\
\mathbf{x}_m &= \sum_{i=1}^{k} \lambda_{i,m} \mathbf{b}_i
\end{aligned}$$

- Define $\boldsymbol{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_k]$ (i.e., a matrix), then

$$\mathbf{x}_j = \boldsymbol{B}\boldsymbol{\lambda}_j, \text{ for } \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{1j} \\ \vdots \\ \lambda_{kj} \end{bmatrix}, \; j = 1, \ldots, m.$$

Remark (2/2)

We want to test whether $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are linearly independent.

- $\sum_{j=1}^{m} \psi_j \mathbf{x}_j = \mathbf{0}$.
- So,

$$\sum_{j=1}^{m} \psi_j \mathbf{x}_j = \sum_{j=1}^{m} \psi_j \boldsymbol{B} \boldsymbol{\lambda}_j = \boldsymbol{B} \sum_{j=1}^{m} \psi_j \boldsymbol{\lambda}_j.$$

  - Why does the last equality hold?

- $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ are linearly independent iff $\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m\}$ are linearly independent.

Basis

Spanning/Generating

Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and a set $\mathcal{A} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \subseteq \mathcal{V}$.

If every vector $\mathbf{v} \in \mathcal{V}$ can be expressed as a linear combination of vectors in $\mathcal{A}$, then $\mathcal{A}$ is called a spanning set (or generating set) of V.

- $\mathcal{A}$ spans V; $\text{span}(\mathcal{A}) = V$.

Basis

Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and a set $\mathcal{A} \subseteq \mathcal{V}$. Then if one of the following condition holds, we say that $\mathcal{A}$ is a basis of V.

- $\mathcal{A}$ is a minimal generating set of V.

  No smaller set $\mathcal{A}' \subsetneq \mathcal{A} \subseteq \mathcal{V}$ that spans V.
- $\mathcal{A}$ spans V and is also linearly independent.

Dimension

Dimension

The number of basis vectors of a vector space V is the *dimension* of V and denoted by $\dim(V)$.

- For $U \subset V$ a subspace of V, $\dim(U) \leq \dim(V)$

Rank

Rank

Rank: the number of linearly independent columns of a matrix $\boldsymbol{A} = \mathbb{R}^{m \times n}$.
This equals the number of linearly independent rows of $\boldsymbol{A}$.

Denote by rank($\boldsymbol{A}$) the rank of $\boldsymbol{A}$.

Linear Mappings/Linear Transformation

A mapping $\Phi : V \mapsto W$ preserves the structure of the vector space if

- $\Phi(\mathbf{x} + \mathbf{y}) = \Phi(\mathbf{x}) + \Phi(\mathbf{y})$
- $\Phi(\lambda \mathbf{x}) = \lambda \Phi(\mathbf{x})$

for all $\mathbf{x}, \mathbf{y} \in V$ and $\lambda \in \mathbb{R}$.

Linear Mapping

For two vector spaces V, W, a mapping $\Phi : V \mapsto W$ is a linear mapping if

$$\forall \mathbf{x}, \mathbf{y} \in V, \forall \lambda, \psi \in \mathbb{R} : \; \Phi(\lambda \mathbf{x} + \psi \mathbf{y}) = \lambda \Phi(\mathbf{x}) + \psi \Phi(\mathbf{y}).$$

Transformation Matrix

Transformation Matrix

Given vector spaces V, W with corresponding bases $B = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ and $C = (\mathbf{c}_1, \dots, \mathbf{c}_m)$. Consider a linear mapping $\Phi : V \mapsto W$. For $1 \leq j \leq n$,

$$\Phi(\mathbf{b}_j) = \alpha_{1,j}\mathbf{c}_1 + \cdots \alpha_{m,j}\mathbf{c}_m = \sum_{i=1}^{m} \alpha_{ij}\mathbf{c}_i$$

is the unique representation of $\Phi(\mathbf{b}_j)$ w.r.t. C (i.e., coordinate). Then, we call the $m \times n$ matrix $\boldsymbol{A}_\Phi$, whose elements are $A_\Phi(i,j) = \alpha_{ij}$, the transformation matrix of Φ.

- If $\hat{\mathbf{x}}$ is the coordinate of $\mathbf{x} \in V$ w.r.t. B and $\hat{\mathbf{y}} = \Phi(\mathbf{x}) \in W$ w.r.t. C, then
$$\hat{\mathbf{y}} = \boldsymbol{A}_\Phi(\hat{\mathbf{x}}).$$

Transformation Matrix

Transformation Matrix

Given vector spaces V, W with corresponding bases $B = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ and $C = (\mathbf{c}_1, \ldots, \mathbf{c}_m)$. Consider a linear mapping $\Phi : V \mapsto W$. For $1 \leq j \leq n$,

$$\Phi(\mathbf{b}_j) = \alpha_{1,j}\mathbf{c}_1 + \cdots \alpha_{m,j}\mathbf{c}_m = \sum_{i=1}^{m} \alpha_{ij}\mathbf{c}_i$$

is the unique representation of $\Phi(\mathbf{b}_j)$ w.r.t. C (i.e., coordinate). Then, we call the $m \times n$ matrix $\boldsymbol{A}_\Phi$, whose elements are $A_\Phi(i,j) = \alpha_{ij}$, the transformation matrix of Φ.

- If $\hat{\mathbf{x}}$ is the coordinate of $\mathbf{x} \in V$ w.r.t. B and $\hat{\mathbf{y}} = \Phi(\mathbf{x}) \in W$ w.r.t. C, then

$$\hat{\mathbf{y}} = \boldsymbol{A}_\Phi(\hat{\mathbf{x}}).$$

Example

Consider a linear mapping $\Phi : V \mapsto W$ and ordered bases $B = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ of V and $C = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4)$ of W. Assume that

$$
\begin{aligned}
\Phi(\mathbf{b}_1) &= \mathbf{c}_1 - \mathbf{c}_2 + 3\mathbf{c}_3 - \mathbf{c}_4 \\
\Phi(\mathbf{b}_2) &= 2\mathbf{c}_1 + \mathbf{c}_2 + 7\mathbf{c}_3 + 2\mathbf{c}_4 \\
\Phi(\mathbf{b}_3) &= 3\mathbf{c}_2 + \mathbf{c}_3 + 4\mathbf{c}_4.
\end{aligned}
$$

The transformation matrix $\boldsymbol{A}_\Phi$ w.r.t. B and C satisfying $\Phi(\mathbf{b}_k) = \sum_{i=1}^{4} \alpha_{ik}\mathbf{c}_i$ for $k = 1, 2, 3$ is

$$
\boldsymbol{A}_\Phi = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3] = \begin{bmatrix} 1 & 2 & 0 \\ -1 & 1 & 3 \\ 3 & 7 & 1 \\ -1 & 2 & 4 \end{bmatrix}.
$$

Norm

Norm

A norm on a vector space V is a function

$$\begin{aligned} \|\cdot\| : V &\mapsto \mathbb{R} \\ \mathbf{x} &\mapsto \|\mathbf{x}\| \end{aligned}$$

such that for $\lambda \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in V$ the following hold:

- $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$.
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.
- $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$.

ℓ_1 norm & ℓ_2 norm

ℓ_1 norm (Manhattan Norm)

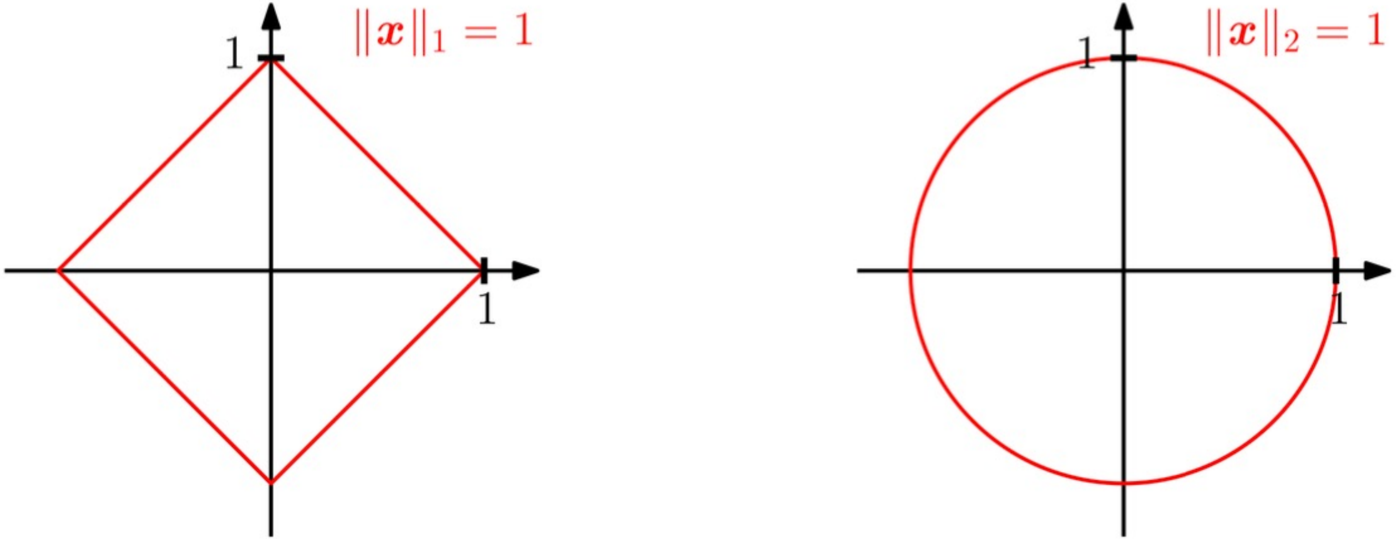
For $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i|.$$

ℓ_2 norm

For $\mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\mathbf{x}^\top \mathbf{x}}.$$


$\|x\|_1 = 1$
1
1
$\|x\|_2 = 1$
1
1

Dot Product

Dot Product

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i .$$

General Inner Products

Bilinear Mapping f

Given a vector space V. For all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, $\lambda, \psi \in \mathbb{R}$, such that

$$f(\lambda\mathbf{x} + \psi\mathbf{y}, \mathbf{z}) = \lambda f(\mathbf{x}, \mathbf{z}) + \psi f(\mathbf{y}, \mathbf{z})$$
$$f(\mathbf{x}, \lambda\mathbf{y} + \psi\mathbf{z}) = \lambda f(\mathbf{x}, \mathbf{y}) + \psi f(\mathbf{x}, \mathbf{z})$$

Symmetric & Positive Definite (1/6)

Symmetric

Let V be a vector space and $f : V \times V \mapsto \mathbb{R}$ be a bilinear mapping. Then f is symmetric if $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$.

Positive Definite

Let V be a vector space and $f : V \times V \mapsto \mathbb{R}$ be a bilinear mapping. Then f is positive definite if $\forall \mathbf{x} \in V \setminus \{\mathbf{0}\}$, we have

$$f(\mathbf{x}, \mathbf{x}) > 0 \quad \text{and} \quad f(\mathbf{0}, \mathbf{0}) = 0.$$

Inner Product

A positive definite & symmetric bilinear mapping $f : V \times V \mapsto \mathbb{R}$ is called an inner product on V and we write $f(\mathbf{x}, \mathbf{y})$ as $\langle \mathbf{x}, \mathbf{y} \rangle$.

Orthogonal Matrix

Orthogonal Matrix

A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is an orthogonal matrix iff its columns are orthonormal so that

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I} = \boldsymbol{A}^\top\boldsymbol{A},$$

which implies

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^\top.$$

Remark

Transformations by orthogonal matrices do NOT change the length of a vector.

$$\|\boldsymbol{A}\mathbf{x}\|^2 = (\boldsymbol{A}\mathbf{x})^\top(\boldsymbol{A}\mathbf{x}) = \mathbf{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\mathbf{x} = \mathbf{x}^\top\boldsymbol{I}\mathbf{x} = \mathbf{x}^\top\mathbf{x} = \|\mathbf{x}\|^2.$$

Let θ be the angle between $\boldsymbol{A}\mathbf{x}$ and $\boldsymbol{A}\mathbf{y}$, what is $\cos\theta$?

Orthogonality

Orthogonality

- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if and only if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.
  - We write $\mathbf{x} \perp \mathbf{y}$.
- If $\mathbf{x}$ and $\mathbf{y}$ are orthogonal and $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, then $\mathbf{x}$ and $\mathbf{y}$ are both orthonormal.

Matrix Decomposition

Eigenvalue Equation

Eigenvalues & Eigenvectors

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a square matrix. Then

- $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$ and
- $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ is the corresponding eigenvector of A

if $\boldsymbol{A}\mathbf{x} = \lambda \mathbf{x}$.

$$\boldsymbol{A}\mathbf{x} = \lambda \mathbf{x} \iff (\boldsymbol{A} - \lambda \boldsymbol{I})\mathbf{x} = \mathbf{0}$$

Equivalent statements:

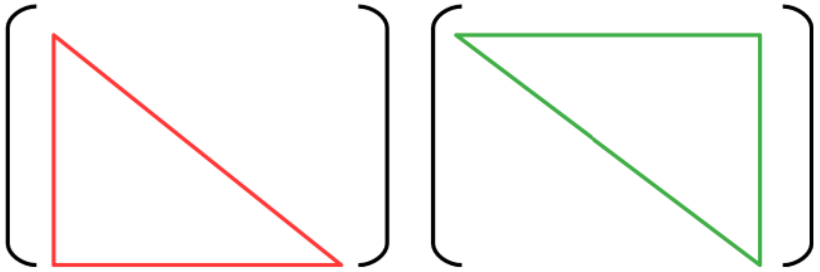
- λ is an eigenvalue of $\boldsymbol{A} \in \mathbb{R}^{n \times n}$.
- There exists an $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ with $\boldsymbol{A}\mathbf{x} = \lambda\mathbf{x}$ (i.e., $(\boldsymbol{A} - \lambda\boldsymbol{I}_n)\mathbf{x} = \mathbf{0}$) that can be solved non-trivially (i.e., $\mathbf{x} \neq \mathbf{0}$).
- $\text{rank}(\boldsymbol{A} - \lambda\boldsymbol{I}_n) < n$.
- $\det(\boldsymbol{A} - \lambda\boldsymbol{I}_n) = 0$.

Cholesky Decomposition

Cholesky Decomposition

A symmetric, positive definite matrix $\boldsymbol{A}$ can be factorized into a product $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower-triangular matrix with positive diagonal elements.

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} =$$



Example of Cholesky Factorization

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \boldsymbol{L}\boldsymbol{L}^\top = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix} \begin{bmatrix} \ell_{11} & \ell_{21} & \ell_{31} \\ 0 & \ell_{22} & \ell_{32} \\ 0 & 0 & \ell_{33} \end{bmatrix}.$$

We have

$$\boldsymbol{A} = \begin{bmatrix} \ell_{11}^2 & \ell_{21}\ell_{11} & \ell_{31}\ell_{11} \\ \ell_{21}\ell_{11} & \ell_{21}^2 + \ell_{22}^2 & \ell_{31}\ell_{21} + \ell_{32}\ell_{22} \\ \ell_{31}\ell_{11} & \ell_{31}\ell_{21} + \ell_{32}\ell_{22} & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \end{bmatrix}$$

Finally, solve $\ell_{11}, \ldots, \ell_{33}$.

Example Steps for Cholesky Factorization

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} \ell_{11}^2 & \ell_{21}\ell_{11} & \ell_{31}\ell_{11} \\ \ell_{21}\ell_{11} & \ell_{21}^2 + \ell_{22}^2 & \ell_{31}\ell_{21} + \ell_{32}\ell_{22} \\ \ell_{31}\ell_{11} & \ell_{31}\ell_{21} + \ell_{32}\ell_{22} & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \end{bmatrix}$$

- $\ell_{11} = \sqrt{a_{11}}, \quad \ell_{21} = \dfrac{a_{21}}{\ell_{11}}, \quad \ell_{22} = \sqrt{a_{22} - \ell_{21}^2}, \quad \ell_{31} = \dfrac{a_{31}}{\ell_{11}},$

  $\ell_{32} = \dfrac{a_{32} - \ell_{31}\ell_{21}}{\ell_{22}}, \quad \ell_{33} = \sqrt{a_{33} - \ell_{31}^2 - \ell_{32}^2}.$

Motivations of Using Cholesky Decomposition

- Symmetric positive definite matrices require frequent manipulation.
  - E.g., Covariance matrix of a multivariate Gaussian variable.
  - The Cholesky factorization of the covariance matrix allows us to generate samples from a Gaussian distribution.
- Computing gradients in deep stochastic models such as variational auto-encoder (VAE).
- Compute determinants efficiently.
  - $\det(\boldsymbol{A}) = \det(\boldsymbol{L})\det(\boldsymbol{L}^\top) = \det(\boldsymbol{L})^2$.
  - Note: $\det(\boldsymbol{L})$ can be computed efficiently ($\because$ triangular).

Eigendecomposition (Diagonalization)

Theorem [Eigendecomposition (Diagonalization)]

A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ can be factored into

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1},$$

where $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{D}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\boldsymbol{A}$

if and only if

the eigenvectors of $\boldsymbol{A}$ form a basis of $\mathbb{R}^n$.

Remark

The spectral theorem tells us that:

We can find an orthonormal basis of the corresponding vector space consisting of eigenvectors of of a symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$.

Theorem

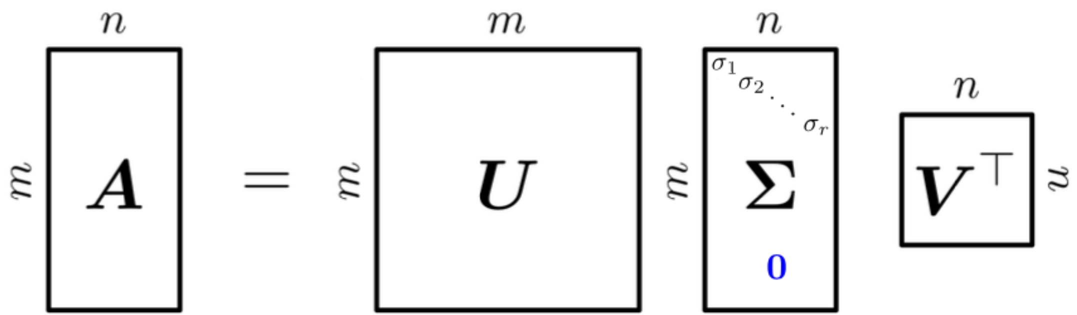
A symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ can be always diagonalized.

Why Singular Value Decomposition?

- It can be applied to all matrices (not only to square matrices).
- It always exists.

Illustration

$\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\text{rank}(\boldsymbol{A}) = r \le \min(m, n)$:



- $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ with orthogonal columns vectors $\boldsymbol{u}_i$, $i = 1, \ldots, m$.
- $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ with orthogonal columns vectors $\boldsymbol{v}_j$, $j = 1, \ldots, n$.
- $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$ with $\Sigma_{ii} = \sigma_i \ge 0$ and $\Sigma_{ij} = 0$ for $i \ne j$.
  - σ_i: singular values; $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_r \ge 0$.
  - $\boldsymbol{u}_i$: left-singular vectors;
    $\boldsymbol{v}_j$: right-singular vectors;

SVD & Eigendecomposition

- Recall the eigendecomposition of a symmetric positive definite matrix

$$\boldsymbol{S} = \boldsymbol{S}^\top = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top.$$

with the corresponding SVD

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

so $\boldsymbol{U} = \boldsymbol{P} = \boldsymbol{V}$, $\boldsymbol{D} = \boldsymbol{\Sigma}$.

The first step: Constructing the right-singular vectors

- **Recall:** Eigenvectors of a *symmetric* matrix form an orthonormal basis (The Spectral theorem).
- Also, we can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{n \times n}$ from any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.
- Thus,

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^\top = \boldsymbol{P} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \boldsymbol{P}^\top,$$

where $\boldsymbol{P}$ is orthogonal and composed of orthonormal eigenbasis.

⋆ $\lambda_i \geq 0$ are the eigenvalues of $\boldsymbol{A}^\top \boldsymbol{A}$.

The first step (2/2)

- Assume the SVD of $\boldsymbol{A}$ exists.

$$\boldsymbol{A}^\top \boldsymbol{A} = (\boldsymbol{U}\Sigma\boldsymbol{V}^\top)^\top(\boldsymbol{U}\Sigma\boldsymbol{V}^\top) = \boldsymbol{V}\Sigma^\top\boldsymbol{U}^\top\boldsymbol{U}\Sigma\boldsymbol{V}^\top$$

where $\boldsymbol{U}$, $\boldsymbol{V}$ are orthonormal matrices $(\because \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I})$. So,

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{V}\Sigma^\top\Sigma\boldsymbol{V}^\top = \boldsymbol{V}\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n^2 \end{bmatrix}\boldsymbol{V}^\top$$

- Hence, we identify $\boldsymbol{V}^\top = \boldsymbol{P}^\top$ (right-singular vectors) and $\sigma_i^2 = \lambda_i$.

The second step: Constructing the left-singular vectors

- Similarly, we can always construct a symmetric, positive semidefinite matrix $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{m\times m}$ from any matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.
- Thus, by assuming the SVD of $\boldsymbol{A}$ exists, we have

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^\top &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top \\
&= \boldsymbol{U}\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_m^2 \end{bmatrix}\boldsymbol{U}^\top
\end{aligned}$$

Note: $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$ have the same eigenvalues.

Vector Calculus

Motivations

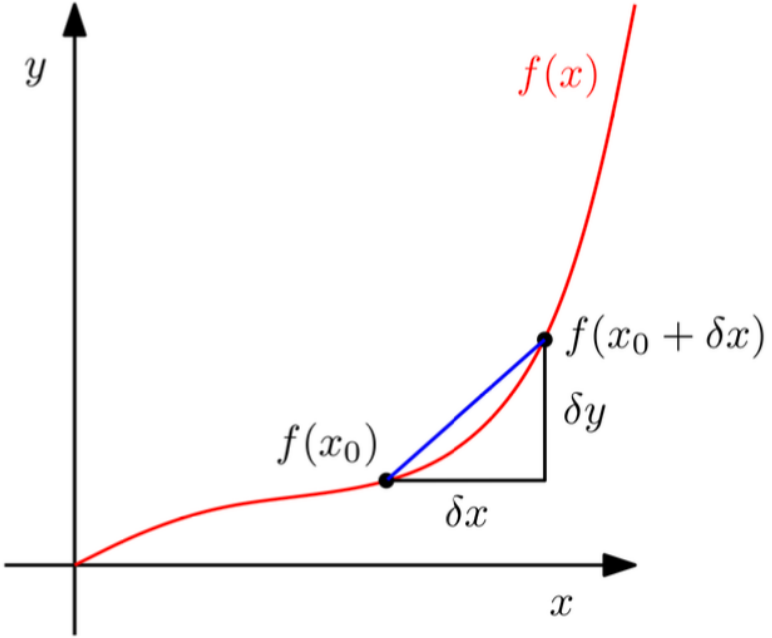
- Machine learning algorithms that optimize an objective function w.r.t. a set of model parameters.
- Examples:
  - Curve-fitting.
  - Neural networks (parameters as weights & biases of layers, repeatedly application of chain rule, etc.)
  - Gaussian mixture models (maximizing the likelihood of the model).
- We focus on functions.
  - $f : \mathbb{R}^D \mapsto \mathbb{R}$ (i.e., $\mathbf{x} \mapsto f(\mathbf{x})$).

Derivative

Consider a univariate function $y = f(x)$, $x, y \in \mathbb{R}$.

Difference Quotient

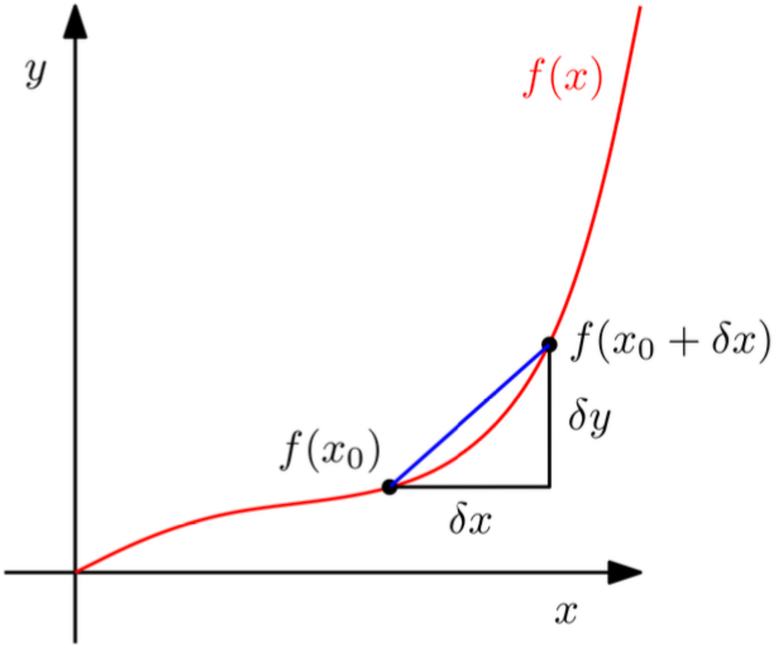
$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x}.$$



Derivative

For $h > 0$, the derivative of f at x:

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$



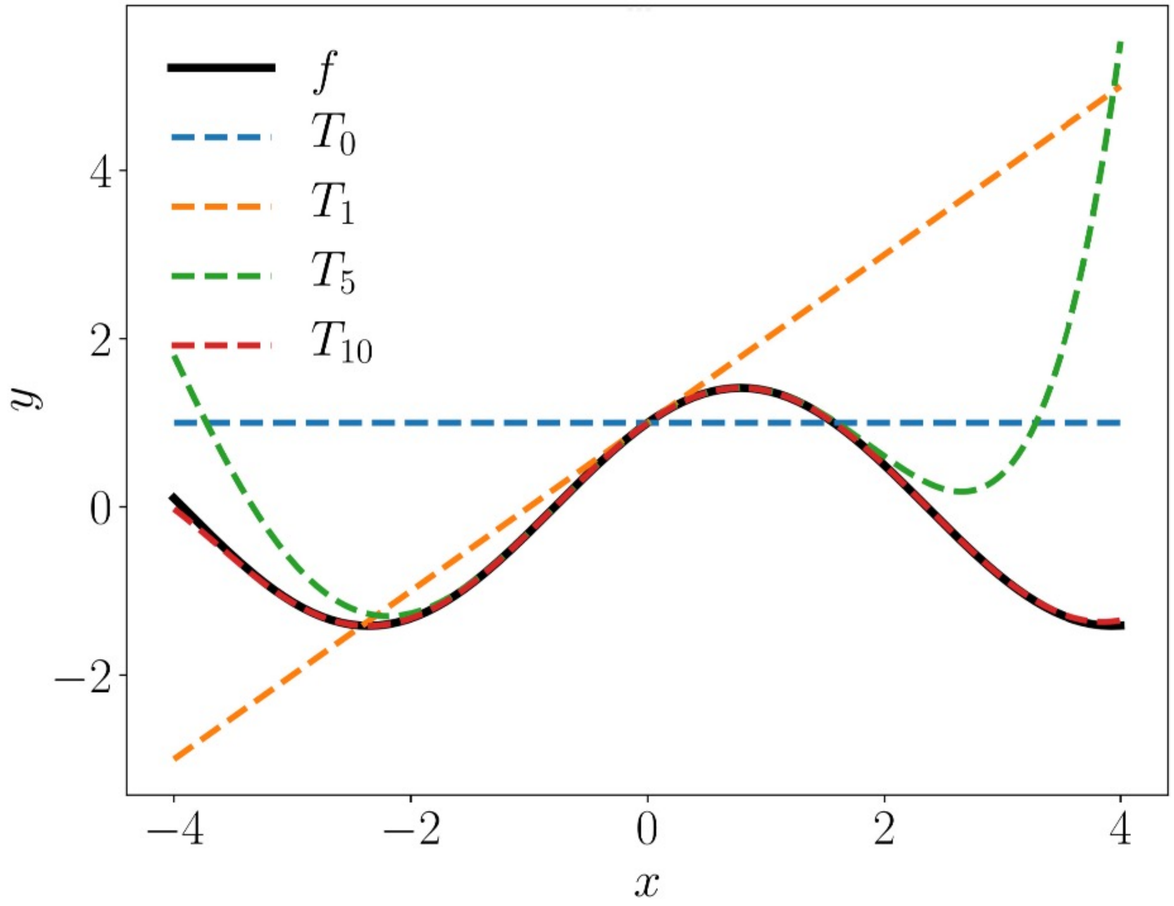
Taylor Series

For a function $f : \mathbb{R} \mapsto \mathbb{R}, f \in \mathcal{C}^\infty$, the Taylor series f at x_0 is:

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k$$

For $x_0 = 0$, it is the *Maclaurin series*.

f is analytic: $f(x) = T_\infty(x)$.



f
T_0
T_1
T_5
T_{10}
y
x
4
2
0
−2
−4
−2
0
2
4

Differentiation Rules

- $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$.
- $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$.
- $(f(x) + g(x))' = f'(x) + g'(x)$.
- $(g(f(x)))' = (g \circ f)'(x) = g'(f(x))f'(x)$.
  - Chain rule.

- **Example:** Compute $h'(x)$ where $h(x) = (2x+1)^4$.

Partial Derivative

Partial Derivative

For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ and $x \in \mathbb{R}^n$ of n variables $x_1, \ldots, x_n$, the partial derivatives are:

$$
\begin{aligned}
\frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \ldots, x_n) - f(\mathbf{x})}{h} \\
&\vdots \\
\frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \ldots, x_{n-1}, x_n + h) - f(\mathbf{x})}{h}
\end{aligned}
$$

We collect them in the row vector:

$$
\nabla_{\mathbf{x}} f = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}
$$

Examples

Example

Given $f(x, y) = (x + 2y^3)^2$, compute $\frac{\partial f(x,y)}{\partial x}$ and $\frac{\partial f(x,y)}{\partial y}$.

Example

Given $f(x, y) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$, compute $\frac{\partial f(x,y)}{\partial x}$, $\frac{\partial f(x,y)}{\partial y}$ and $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$.

Basic Partial Differentiation Rules

- $\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x})g(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} g(x) + f(x)\frac{\partial g}{\partial \mathbf{x}}$.
- $\frac{\partial}{\partial \mathbf{x}}(f(x) + g(x)) = \frac{\partial f}{\partial \mathbf{x}} + \frac{\partial g}{\partial \mathbf{x}}$.
- $\frac{\partial}{\partial \mathbf{x}}(g \circ f)(\mathbf{x}) = \frac{\partial g}{\partial \mathbf{x}}(g(f(\mathbf{x}))) = \frac{\partial g}{\partial f}\frac{\partial f}{\partial \mathbf{x}}$.
    - Chain rule.

Chain Rule (Partial Differentiation)

- Consider a function $f : \mathbb{R}^2 \mapsto \mathbb{R}$ of two variables x_1, x_2.
    - $x_1(t), x_2(t) : \mathbb{R} \mapsto \mathbb{R}$.

Then,

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}.$$

Here 'd' denotes the gradient and '∂' denotes partial derivatives.

- **Note:** Here the 't' in $\mathrm{d}t$ is in $\mathbb{R}^1$.
- Trick: View $[x_1, x_2]^\top$ as $\mathbf{x} \in \mathbb{R}^2$.

  $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}}$: $\mathbb{R}$ w.r.t. $\mathbb{R}^2$.

  $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}$: $\mathbb{R}^2$ w.r.t. $\mathbb{R}$.

Example

Example

Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$. Calculate

$$\frac{\mathrm{d}f}{\mathrm{d}t} = ?$$

Heads up

We will see that

- $f : \mathbb{R}^D \mapsto \mathbb{R}$: the gradient is a $1 \times D$ row vector.
- $\mathbf{f} : \mathbb{R} \mapsto \mathbb{R}^E$: the gradient is a $E \times 1$ column vector.
- $\mathbf{f} : \mathbb{R}^D \mapsto \mathbb{R}^E$: the gradient is a $E \times D$ matrix.

Thank you !