Tutorial 3 : Algebra Review

1 Theory

Vector Space

A vector space over $\mathbb{R}$ is a set $\mathbb{V}$ of vectors together with two operations, **vector addition** and **scalar multiplication**, that satisfy the following for all $u, v, w \in V$:

1. Associativity - $u + (v + w) = (u + v) + w$.
2. Commutativity - $u + v = v + u$.
3. Zero vector - $\exists 0 \in \mathbb{V}$ such that $v + 0 = v \;\; \forall v \in \mathbb{V}$.
4. Inverse - $\forall v \in \mathbb{V} \;\; \exists(-v) \in \mathbb{V}$ such that $v + (-v) = 0$.
5. Compatibility - $a(bv) = (ab)v, \;\; a, b, \in \mathbb{R}$.
6. Identity - $1v = v$.
7. Distributivity - $a(u + v) = au + av$ and $(a + b)v = av + bv \;\; a, b \in \mathbb{R}$.

A set $\{v_1, v_2, ..., v_n\} \subseteq \mathbb{V}$ is **linearly independent** if

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \;\; \Rightarrow \;\; \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

$\{v_1, v_2, ..., v_n\}$ is said to **span** $\mathbb{V}$ if for any $v \in \mathbb{V}$, there exists $\beta_1, \beta_2, ..., \beta_n \in \mathbb{R}$ such that

$$x = \sum_{i=1}^{n} \beta_i v_i.$$

A **basis** of $\mathbb{V}$ is an independent set of vectors that spans $\mathbb{V}$. The number of vectors in all the bases of a vector space $\mathbb{V}$ is the same and called the **dimension** of $\mathbb{V}$ - $dim(\mathbb{V})$.

Inner Products

An inner product of a pair $x, y, \in \mathbb{V}$ is a function denoted by $\langle x, y \rangle$ which satisfies the following properties:

1. Commutativity - $\langle x, y \rangle = \langle y, x \rangle$ for any $x, y, \in \mathbb{V}$.
2. Linearity - $\langle \alpha x_1 + \beta x_2, y \rangle = \alpha \langle x_1, y \rangle + \beta \langle x_2, y \rangle$ for any $\alpha, \beta \in \mathbb{R}$ and $x_1, x_2, y \in \mathbb{V}$.
3. Positive definiteness $\langle x, x \rangle \geq 0$ for any $x \in \mathbb{V}$ and $\langle x, x \rangle = 0$ if and only if (iff) $x = 0$

Examples

- $x, y \in \mathbb{R}^n$ - $\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$.
- $A, B \in \mathbb{R}^{m \times n}$ - $\langle A, B \rangle = Tr(A^T B) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$.
- $x, y \in \mathbb{R}^n, Q \succeq 0$ - $\langle x, y \rangle_Q = x^T Q y$.

Adjoint Transformation

Given a linear transformation $\mathcal{A} : \mathbb{V} \to \mathbb{U}$, the adjoint transformation denoted by $\mathcal{A}^* : \mathbb{U} \to \mathbb{V}$ is a transformation that is defined by the relation

$$\langle \mathcal{A}(x), y \rangle = \langle x, \mathcal{A}^*(y) \rangle$$

for any $x \in \mathbb{V}$ and $y \in \mathbb{U}$. As an example for $x \in \mathbb{R}^m, \; y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$ we have

$$\langle Ax, y \rangle = (Ax)^T y = x^T A^T y = x^T (A^T y) = \langle x, A^T y \rangle \;\; \to \;\; A^* = A^T.$$

Norm

A norm on a vector space $\mathbb{V}$ is a function $||\cdot|| : \mathbb{V} \to \mathbb{R}$ satisfying

- Nonnegativity - $\forall x \in \mathbb{V}$ $||x|| \geq 0$ and $||x|| = 0$ iff $x = 0$.
- Homogeneity - $||\lambda x|| = |\lambda| \cdot ||x||$ $\forall x \in \mathbb{V}$ and $\forall \lambda \in \mathbb{R}$.
- Triangle inequality - $||x + y|| \leq ||x|| + ||y||$ $\forall x, y \in \mathbb{V}$.

Examples

- l_p norm ($p \geq 1$) - $||x||_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
- l_1 norm - $||x||_1 = \sum_{i=1}^{n} |x_i|$.
- l_2 norm - $||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.
- l_∞ norm - $||x||_\infty = \max\limits_{i=1,2,\ldots,n} |x_i| = \lim\limits_{p\to\infty} ||x||_p$.
- Induced norm - $||x|| \equiv \sqrt{\langle x, x\rangle}$.
- Induced matrix norm - $||A||_{a,b} = \max\limits_{x:||x||_a \leq 1} ||Ax||_b$.
- Spectral norm - $||A||_2 = ||A||_{2,2} = \sigma_{\max}(A)$.
- Frobenius - $||A||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2} = \sqrt{Tr(A^T A)}$.

Cauchy-Schwartz Inequality

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

$$|\langle x, y\rangle| \leq ||x|| \cdot ||y||.$$

Matrices

Eigenvalues and Eigenvectors

Let $A \in \mathbb{R}^{n\times n}$. Then a nonzero vector $v \in \mathbb{R}^n$ is called an **eigenvector** of A if there exists a $\lambda \in \mathbb{R}$ for which

$$Av = \lambda v.$$

The scalar λ is the **eigenvalue** corresponding to the eigenvector v.

Positive Definiteness

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. A is said to be **positive semi-definite (PSD)** if it holds that

$$v^T Av \geq 0, \ \forall v \in \mathbb{R}^n.$$

The matrix A is said to be **positive definite (PD)** if $v^T Av > 0$ for every non-zero $v \in \mathbb{R}^n$.

Lemma: The matrix A is PSD/PD if all its eigenvalues are non-negative/positive.

Spectral Decomposition

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then there exists a unitary matrix $U \in \mathbb{R}^{n\times n}$ ($U^T U = UU^T = I$) and a diagonal matrix Λ with $\Lambda_{ii} = \lambda_i$ for which

$$A = U\Lambda U^T.$$

Notice that for an integer $k \geq 0$ it holds that $A^k = U\Lambda^k U^T$.

Matrix Trace

Let $A \in \mathbb{R}^{n \times n}$. Then the trace of A is defined as

$$Tr(A) \triangleq \sum_{i=1}^{n} A_{ii}.$$

Properties:

- Trace is a linear mapping.
- Trace is invariant under cyclic permutations - $Tr(ABC) = Tr(CAB) = Tr(BCA)$.
- $Tr(A) = \sum_{i=1}^{n} \lambda_i$.

Matrix Function

Let $A \in \mathbb{R}^{n \times n}$ and $f(x)$ be a scalar function where it Taylor series is

$$f(x) = \sum_{k=0}^{\infty} c_k x^k.$$

We define the matrix function $f(A)$ as follows

$$f(A) \triangleq \sum_{k=0}^{\infty} c_k A^k = \sum_{k=0}^{\infty} c_k U \Lambda^k U^T = U(\sum_{k=0}^{\infty} c_k \Lambda^k) U^T = U f(\Lambda) U^T$$

where

$$f(\Lambda) = \begin{bmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_2) \end{bmatrix}.$$

Notice that $Tr\big(f(A)\big) = \sum_{i=1}^{n} f(\lambda_i)$.

External Definition of Gradient

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function for which

$$df = \langle g(x), dx \rangle.$$

Then $g(x)$ is the gradient of $f(x)$.

2 Practice

Question 1

Compute the gradient of the following functions:

(a) $f(x) = \sum_{i=1}^{n} f_i(x_i)$ where $f_i(x)$ is a differentiable function.

(b) $f(X) = \frac{1}{2}||Y - AX||_F$.

(c) $f(X) = \sum_{i=1}^{n} \lambda_i$.

(d) $f(X) = ||X^{\frac{k}{2}}||_F^2$, X is symmetric, $k \geq 0$.

(e) $f(X) = Tr\big(h(X)\big)$ where $h(x)$ is a scalar differentiable function and X is symmetric.

(f) $f(X) = \log\det X$, X is PSD.

(g) We define the following transformation $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$:

$$\mathcal{A}(x) \triangleq \sum_{i=1}^{n} x_i V_i,$$

where $V_i \in \mathbb{R}^{n\times n}$ are known symmetric matrices.
For a given matrix $Y \in \mathbb{R}^{n\times n}$ we define $y = \mathcal{A}^*(Y)$. Then $f(Y) = \sum_{i=1}^{n} y_i$.

Solution

(a) $f(x) = \sum_{i=1}^{n} f_i(x_i)$:

$$df = d\left(\sum_{i=1}^{n} f_i(x_i)\right) = \sum_{i=1}^{n} df_i(x_i) = \sum_{i=1}^{n} f_i^{'}(x_i)dx_i = \langle g(x), dx\rangle$$

where

$$g(x) = \begin{bmatrix} f_1^{'}(x_1) \\ f_2^{'}(x_2) \\ \vdots \\ f_n^{'}(x_n) \end{bmatrix}.$$

(b) $f(X) = \frac{1}{2}||Y - AX||_F^2$:

$$\begin{aligned}
df &= d\left(\frac{1}{2}||Y - AX||_F^2\right) = d\left(\frac{1}{2}Tr\Big((Y-AX)^T(Y-AX)\Big)\right) \\
&= \frac{1}{2}Tr\Big(d(Y-AX)^T(Y-AX)\Big) = \frac{1}{2}Tr\Big((-AdX)^T(Y-AX) + (Y-AX)^T(-AdX)\Big) \\
&= \frac{1}{2}Tr\Big((-AdX)^T(Y-AX)\Big) + Tr\Big((Y-AX)^T(-AdX)\Big) \\
&= \frac{1}{2}Tr\Big((Y-AX)^T(-AdX)\Big) + Tr\Big((Y-AX)^T(-AdX)\Big) \\
&= Tr\Big((Y-AX)^T(-AdX)\Big) = Tr\Big(-(Y-AX)^T AdX\Big) \\
&= Tr\left(\Big(-A^T(Y-AX)\Big)^T dX\right) \to g(x) = -A^T(Y-AX).
\end{aligned}$$

(c) $f(X) = \sum_i \lambda_i(X)$:

$$df = d\Big(\sum_i \lambda_i(X)\Big) = d\Big(Tr(X)\Big) = Tr(dX) = Tr(I^T dX) \to g(x) = I.$$

(d) $f(X) = ||X^{\frac{k}{2}}||_F^2$ - Notice that $||X^{\frac{k}{2}}||_F^2 = Tr(X^{\frac{k}{2}}X^{\frac{k}{2}}) = Tr(X^k)$, hence

$$\begin{aligned}
df &= d\Big(Tr(X^k)\Big) = Tr(d(\underbrace{X\cdots X}_{\text{k times}})) = Tr\Big((dX\cdots X) + \cdots(X\cdots dX\cdots X) + \cdots(X\cdots dX)\Big) \\
&= Tr(dX\cdots X) + \cdots Tr(X\cdots dX\cdots X) + \cdots Tr(X\cdots dX) \\
&= Tr(X^{k-1}dX) + \cdots Tr(X^{k-1}dX) + \cdots Tr(X^{k-1}dX) \\
&= Tr(kX^{k-1}dX) = Tr\Big((kX^{k-1})^T dX\Big) \to g(X) = kX^{k-1}.
\end{aligned}$$

(e) $f(X) = Tr\big(h(X)\big)$ - Consider $h(X) = \sum_{k=0}^{\infty} c_k X^k$. Then

$$h'(X) = \sum_{k=1}^{\infty} c_k k X^{k-1} \equiv \sum_{k=0}^{\infty} \tilde{c}_k X^k$$

$$\begin{aligned}
df &= dTr\big(h(X)\big) \\
&= dTr\left(\sum_{k=0}^{\infty} c_k X^k\right) \\
&= Tr\left(d\left(\sum_{k=0}^{\infty} c_k X^k\right)\right) \\
&= Tr\left(\sum_{k=1}^{\infty} c_k k X^{k-1} dX\right) \\
&= Tr\left(\sum_{k=1}^{\infty} \tilde{c}_k X^k dX\right) \\
&= Tr\left(h^{'}(X) dX\right) = \langle h^{'}(X)^T, dX\rangle \to g(X) = h^{'}(X)^T.
\end{aligned}$$

(f) $f(X) = \log\det X$ - Notice that

$$f(X) = \log\prod_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \log\lambda_i = Tr(\log(X)).$$

Since $\log^{'}(x) = x^{-1}$ we get that $g(x) = X^{-1}$.

(g) First, we find an expression for $\mathcal{A}^*(Y)$:

$$\begin{aligned}
\langle \mathcal{A}(x), Y\rangle &= Tr\left(\mathcal{A}(x)^T Y\right) \\
&= Tr\left(\sum_{i=1}^{n} x_i V_i^T Y\right) \\
&= \sum_{i=1}^{n} x_i Tr(V_i^T Y) \equiv \langle x, \mathcal{A}^*(Y)\rangle.
\end{aligned}$$

Hence,

$$\mathcal{A}^*(Y) = \begin{bmatrix} Tr(V_1^T Y) \\ Tr(V_2^T Y) \\ \vdots \\ Tr(V_n^T Y) \end{bmatrix} \Rightarrow f(Y) = \sum_{i=1}^{n} Tr(V_i^T Y) = Tr\left(\left(\sum_{i=1}^{n} V_i\right)^T Y\right).$$

Define $V \triangleq \sum_{i=1}^{n} V_i$. Then, $f(Y) = Tr(V^T Y)$ and the gradient is $g(Y) = V = \sum_{i=1}^{n} V_i$.