Mathematical background review

1) <u>Notation:</u>

$x \in \mathbb{R}^n$: x is a column vector of n-dimension

x^T : transpose of x is a row vector.

$x = (x_1\, x_2 \dots x_n)$
$x^T = (x_1\, x_2 \dots x_n)^T$

$X \in \mathbb{R}^{m \times n}$: X is a matrix of m rows and n columns.

X_{ij} : X_{ij} element at row i column j.

dom f : domain of function f, which is the set where the variable of f comes from

range f : range of function f, the set of values of the function itself.

2) Norms

a) Inner product

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i$$

Euclidean norm or l_2 norm:

$$\|x\|_2 = (x^T x)^{1/2} = (x_1^2 + \ldots + x_n^2)^{1/2}$$

On matrices:

$$\langle X, Y \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij} = \mathrm{tr}(X^T Y)$$

Frobenius norm:

$$\|X\|_F = [\mathrm{tr}(X^T X)]^{1/2} = \left(\sum_{i}^{m} \sum_{j}^{n} X_{ij}^2\right)^{1/2}$$

b) Norm definition:

A function $f: \mathbb{R}^n \to \mathbb{R}$ with $\mathrm{dom} f = \mathbb{R}^n$ is called a norm if

- $f(x) \geq 0 \quad \forall x \in \mathbb{R}^n$
- $f(x) = 0$ only if $x = 0$
- $f(tx) = |t| f(x) \quad \forall x \in \mathbb{R}^n, t \in \mathbb{R}$
- $f(x + y) \leq f(x) + f(y)$

Norm is like a length measure. Usually write $f(x) = \|x\|$.

Example:

+) l_1-norm:

$$\|x\|_1 = |x_1| + \dots + |x_n|$$

+) l_2-norm: Euclidean norm

$$\|x\|_2 = (|x_1|^2 + \dots + |x_n|^2)^{1/2}$$

+) l_p-norm:

$$\|x\|_p = (|x_1|^p + \dots + |x_n|^p)^{1/p}, \quad p \geq 1$$

+) l_∞-norm:

$$\|x\|_\infty = \max\{|x_1|, \dots, |x_n|\}.$$

+) l_2 norm of a matrix:

$$\|X\|_2 = \sigma_{max}(X) = \lambda_{max}^{1/2}(X^T X)$$

also called the spectral norm.

+) distance:

$$\operatorname{dist}(x, y) = \|x - y\|$$

+) Unit ball:

$$B = \{x \in \mathbb{R} \mid \|x\| \leq 1\}$$

3. Analysis:

+) Open and closed sets:

An element $x \in S \subseteq \mathbb{R}^n$ is an interior point of S if $\exists\ \varepsilon > 0$ s.t.

$$\{ y \mid \|y - x\|_2 \leq \varepsilon \} \subseteq S$$

$\rightarrow$ the ball centered at x with radius ε lies entirely in set S.

The closure of set S:

$$\mathrm{cl}\, S = \left(\mathrm{int}(S^c)\right)^c$$

Boundary of set S:

$$\mathrm{bd}\, S = \mathrm{cl}\, S \setminus \mathrm{int}\, S$$

A set is open if : $S = \mathrm{int}\, S$

A set is closed if it contains its boundary.

+) Supremum and infimum:

For a set $S \subseteq \mathbb{R}$, a number a is an upperbound on S if

$$x \leq a \quad \forall\ x \in S.$$

The smallest in the set of upperbound on S is called the supremum of $\sup S$.

$\max S = \sup S$ when the supremum is attainable
usually when S is finite.

4. Linear Algebra:

1) Range and nullspace:

- $A \in \mathbb{R}^{m \times n}$: real matrix with m rows, n columns
- $R(A)$: range of A

$$R(A) = \{ Ax \mid x \in \mathbb{R}^n \}$$

$$R(A) \subseteq \mathbb{R}^m$$ - set of vectors in $\mathbb{R}^m$. is a subspace in $\mathbb{R}^m$.

- $N(A)$: null space of A

$$N(A) = \{ x \mid Ax = 0 \}$$

$$N(A) \subseteq \mathbb{R}^n$$ - a subspace in $\mathbb{R}^n$.

- Orthogonality:

If V is a subspace of $\mathbb{R}^n$, its orthogonal complement is defined as

$$V^{\perp} = \{ x \mid z^T x = 0 \ \forall z \in V \}.$$

A basic result:

$$N(A) = R(A^T)^{\perp} \longleftrightarrow N(A) \overset{\perp}{\oplus} R(A^T) = \mathbb{R}^n$$

orthogonal direct sum.

- Rank: dimension of $R(A)$ is the rank of A.

If A is full rank $\rightarrow$ $\text{rank}(A) = \min(m, n)$.

+) Eigenvalue decomposition. (EVD) also called spectral decomposition

- $A \in S^n$: a real symmetric matrix size $n \times n$

$$A = A^T, \quad A \in \mathbb{R}^{n \times n}$$

Then A can be factored into its EVD as

$$A = Q \Lambda Q^T$$

where $Q \in \mathbb{R}^{n \times n}$ is orthogonal, $Q^T Q = I$

columns (and rows) of Q form a orthonormal basis.

$$Q^T Q = Q Q^T = I.$$

$\Lambda \in \mathbb{R}^{n \times n}$, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$

λ_i are the eigenvalues of A, λ_i real

Notation: assume $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.

$$\lambda_1 = \lambda_{max}(A)$$
$$\lambda_n = \lambda_{min}(A)$$

- Some identities

$$\det A = \prod_{i=1}^{n} \lambda_i$$

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i$$

$$\|A\|_2 = \max_i |\lambda_i| \quad \text{spectral norm}$$

$$\|A\|_F = \left(\sum_{i=1}^{n} \lambda_i^2\right)^{1/2} \quad \text{Frobenius norm.}$$

+) Definiteness :

- $A \in S^n$ is positive definite if
$$x^T A x > 0 \quad \forall x \in \mathbb{R}^n, x \neq 0.$$
Denote as $A \succ 0$, $A \in S^n_{++}$.

- Positive semidefinite if
$$x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n$$
Written $A \succeq 0$, $A \in S^n_+$

- Eigenvalues inequalities :
$$\lambda_{max}(A) = \sup_{x \neq 0} \frac{x^T A x}{x^T x}$$

$$\lambda_{min}(A) = \inf_{x \neq 0} \frac{x^T A x}{x^T x}$$

thus $\forall x \in \mathbb{R}^n$
$$\lambda_{min}(A)\, x^T x \leq x^T A x \leq \lambda_{max}(A)\, x^T x$$

- Matrix inequalities : (associated with the positive semi-definite cone)

$A, B \in S^n$, we say $A \prec B$ if $B - A \succ 0$.

Note : not all matrices can be ordered.

+) Symmetric square root : $A \in S^n_+$.
$$A = Q \Lambda Q^T$$
$$\rightarrow A^{1/2} = Q \Lambda^{1/2} Q^T, \quad \Lambda^{1/2} = \mathrm{diag}(\lambda_1^{1/2}, \dots, \lambda_n^{1/2})$$

+) Singular value decomposition: (SVD)

For a generic matrix $A \in \mathbb{R}^{m \times n}$, rank $A = r$, ($r \le m, n$)
its SVD is

$$A = U \Sigma V^T$$

(columns orthogonal)

where $U \in \mathbb{R}^{m \times r}$, $U^T U = I_r$ → left singular vectors

$V \in \mathbb{R}^{n \times r}$, $V^T V = I_r$ → right singular vectors

$\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ → singular values

Can also write:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \qquad v_i \in \mathbb{R}^n, \; u_i \in \mathbb{R}^m$$

∘ Maximum and minimum singular values:

$$\sigma_{max}(A) = \sup_{x, y \neq 0} \frac{x^T A y}{x^T y}$$

Minimum singular value

$$\sigma_{min}(A) = \begin{cases} \sigma_r(A) & \text{if } r = \min(m,n) \\ 0 & r < \min(m,n) \end{cases}$$

∘ Condition number: for a non-singular $A \in \mathbb{R}^{n \times n}$

$$\mathrm{cond}(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \qquad (= K(A))$$

+) Pseudo-inverse:

Let $A = U\Sigma V^T$, $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank} A = r$.

Define the pseudo inverse or Moore-Penrose Inverse of A as

$$A^{\dagger} = V\Sigma^{-1}U^T, \quad A^{\dagger} \in \mathbb{R}^{n\times m}$$

If $\operatorname{rank} A = n \ (< m)$: $\quad A^{\dagger} = (A^TA)^{-1}A^T$

$\operatorname{rank} A = m \ (< n)$: $\quad A^{\dagger} = A^T(AA^T)^{-1}$

A square, non-singular: $\operatorname{rank} A = n = m$ $\quad A^{\dagger} = A^{-1}$

+) SVD and EVD:

$A \in \mathbb{R}^{m\times n}$, let $B = A^TA$, $B \in S^n$.

If $A = U\Sigma V^T$ as its SVD

then
$$B = A^TA = V\Sigma^2V^T, \quad V \in \mathbb{R}^{n\times r}$$
$$= [V \ \tilde{V}]\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} V^T \\ \tilde{V}^T \end{bmatrix}.$$

where $\tilde{V}$ is any matrix s.t. $[V \ \tilde{V}] \in \mathbb{R}^{n\times n}$ is orthogonal.

Similarly
$$AA^T = U\Sigma^2U^T$$

$\rightarrow \lambda_i(AA^T) = \lambda_i(A^TA) = \sigma_i^2$.

+) Eigenvalues, similarity, EVD and generalized eigenvalue decomposition.

- Eigenvalue definition:

$A \in \mathbb{R}^{n \times n}$ (also applies to complex valued matrices)

If $Ax = \lambda x$, $x \in \mathbb{R}^n$, $x \neq 0$

$\lambda \in \mathbb{R}$

then λ is called an eigenvalue

x " eigenvector.

- Eigenvalues are roots of the following equation:

$(\lambda I - A)x = 0$, $x \neq 0$

$\rightarrow \det(\lambda I - A) = 0$ characteristic polynomial

The set of all eigenvalues of A is called the spectrum of A, denoted as $\sigma(A)$.

Spectral radius: $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$.

- Similarity: Let $M_n \equiv \mathbb{C}^{n \times n}$, $A \in M_n$ complex square matrix

A matrix $B \in M_n$ is said to be "similar" to A if there exists a non-singular $S \in M_n$ such that

$$B = S^{-1} A S$$

Write $B \sim A$

Transformation $A \rightarrow S^{-1} A S$ is called a similarity transform.

- If A and B are similar then they have the same characteristic function and, consequently, the same eigenvalues.

Note: The reverse is not true.

- Diagonalizable:

If $A \in M_n$ is similar to a diagonal matrix then A is said to be diagonalizable.

- $A \in M_n$ is diagonalizable if and only if there is a set of n linearly independent vectors, each of which is an eigenvector of A.

- If A has n distinct eigenvalues then A is diagonalizable. The reverse is not true.

- If A is diagonalizable then $\exists$ nonsingular $S \in M_n$ and a diagonal matrix Λ s.t.

$$A = S^{-1} \Lambda S$$

- Hermitian matrices:

Let $A \in M_n$. Then A is Hermitian iff $\exists$ a unitary matrix $U \in M_n$ and a real diagonal matrix $\Lambda \in M_n$ s.t.

$$A = U \Lambda U^* \qquad (UU^* = U^*U = I)$$

$A \in S^n$ (real & symmetric) iff $\exists$ Q real orthogonal and Λ real diagonal s.t.

$$A = Q \Lambda Q^T \qquad (QQ^T = Q^TQ = I)$$

+) Generalized EVD:

◦ Given a pair of symmetric matrices $(A, B) \in S^n$, their generalized eigenvalues are the roots of the polynomial equation:

$$\det(sB - A) = 0.$$

◦ If B is non singular and positive definite ($B \in S^n_{++}$) then the generalized eigenvalues are the eigenvalues of $B^{-1/2} A B^{-1/2}$.

When $B \in S^n_{++}$, the pair (A, B) can be factored as

$$A = V \Lambda V^T \quad , \quad B = V V^T$$: generalized eig. decomp.

where $V \in \mathbb{R}^{n \times n}$, V non singular

Λ : diagonal matrix of generalized eigenvalues

So if the eigenvalue decomposition of $B^{-1/2} A B^{-1/2}$ is

$$B^{-1/2} A B^{-1/2} = Q \Lambda Q$$

then $V = B^{+1/2} Q$.

5. Derivatives

+) Derivative and gradient:

- Consider $f: \mathbb{R}^n \to \mathbb{R}^m$, $x \in \text{int dom} f$.

  f: function that takes input as a vector in $\mathbb{R}^n$ and produces a vector value in $\mathbb{R}^m$.

- The derivative (or Jacobian) of f at x is the matrix $Df(x) \in \mathbb{R}^{m\times n}$ given by

$$Df(x)_{ij} = \frac{\partial f_i(x)}{\partial x_j} \qquad \begin{matrix} i = 1 \ldots m \\ j = 1 \ldots n, \end{matrix}$$

provided the partial derivatives exist.

- First-order approximation: The affine function of z given by

$$g(z) = f(x) + Df(x)(z - x)$$

is called the first-order approximation of f at x or near x.

$$\lim_{\substack{z \neq x,\, z \to x \\ z \in \text{dom} f}} \frac{\| f(z) - f(x) - Df(x)(z-x) \|_2}{\| x - z \|_2} = 0$$

dom f: the set of all input values for which the function f is defined.

+) Gradient: When f is a scalar real-valued $f: \mathbb{R}^n \to \mathbb{R}$, the derivative $Df(x)$ is a $\underset{\text{row}}{\hat{}}$ vector. Its transpose is called the gradient of f:

$$\nabla f(x) = Df(x)^T.$$

$$\nabla f(x) \in \mathbb{R}^n$$

$$\nabla f(x)_i = \frac{\partial f(x)}{\partial x_i}, \quad i = 1 \ldots n.$$

The first order approximation of f at point $x \in \operatorname{dom} f$ can be written as

$$g(z) = f(x) + \nabla f(x)^T (z - x).$$

Example:

i) $f: \mathbb{R}^n \to \mathbb{R}$ as a quadratic function

$$f(x) = \frac{1}{2} x^T P x + q^T x + r$$

$$P \in S^n, \; q \in \mathbb{R}^n, \; r \in \mathbb{R}.$$

Its derivative is:

$$Df(x) = x^T P + q^T.$$

Its gradient is

$$\nabla f(x) = Px + q.$$

ii) $f: S^n \to \mathbb{R}$ as

$$f(X) = \log\det(X) \quad , \quad \text{dom} f = S^n_{++}$$

To find the gradient of $f(X)$, we will use the first order approximation.

Let $Z = X + \Delta X$, where ΔX is small

$$\begin{aligned}
\log\det Z &= \log\det(X + \Delta X) \\
&= \log\det\left(X^{1/2}[I + X^{-1/2}\Delta X X^{-1/2}]X^{1/2}\right) \\
&= \log\det X + \log\det(I + X^{-1/2}\Delta X X^{-1/2}) \\
&= \log\det X + \sum_{i=1}^{n} \log(1 + \lambda_i)
\end{aligned}$$

where λ_i is the i^{th} eigenvalues of $X^{-1/2}\Delta X X^{-1/2}$.

Since ΔX is small, λ_i are small $\to$ can use the approximation

$$\log(1 + \lambda_i) \approx \lambda_i$$

then

$$\begin{aligned}
\log\det Z &\approx \log\det X + \sum_{i=1}^{n} \lambda_i \\
&= \log\det X + \text{tr}(X^{-1/2}\Delta X X^{-1/2}) \\
&= \log\det X + \text{tr}(X^{-1}\Delta X) \\
&= \log\det X + \text{tr}(X^{-1}(Z - X))
\end{aligned}$$

Thus the first-order approximation of $f(X)$ is

$$f(Z) \approx f(X) + \text{tr}(X^{-1}(Z - X))$$

$\to$ the gradient of $f(X)$ is $\nabla f(X) = X^{-1}$.

+) Chain rule:

Suppose $f: \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $x \in \text{int dom} f$

and $g: \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(x) \in \text{int dom}\, g$

- Define the composition $h: \mathbb{R}^n \to \mathbb{R}^p$

$$h = g(f(x)).$$

then h is differentiable at x with derivatives

$$Dh(x) = Dg(f(x)) Df(x)$$

- If $f: \mathbb{R}^n \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$ then

$$h = g(f(x)) : \mathbb{R}^n \to \mathbb{R}$$

$$\nabla h(x) = g'(f(x)) . \nabla f(x) .$$

$$Dh(x) = Dg\, Df = g'(f(x))\, D(f(x))$$

- Composition with affine function:

$f: \mathbb{R}^n \to \mathbb{R}^m$

$h: \mathbb{R}^p \to \mathbb{R}^m$ where $h(x) = f(Ax + b)$

$A \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^n$

$\text{dom}\, g = \{x \in \mathbb{R}^p \mid Ax + b \in \text{dom} f\}$

By the chain rule:

$$Dh(x) = Df(Ax+b)\, A .$$

$$\to \nabla h(x) = A^T \nabla f(Ax+b) .$$

Example: Consider $f: \mathbb{R}^n \to \mathbb{R}$, $\mathrm{dom} f = \mathbb{R}^n$ as

$$f(x) = \log\left(\sum_{i=1}^{m} \exp(a_i^T x + b_i)\right)$$

where $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}^n$, $i = 1 \ldots m$.

Think of $f(x)$ as a composition of two function:

$$g(y) = \log\left(\sum_{i=1}^{n} e^{y_i}\right) \qquad A = \begin{bmatrix} — a_1^T — \\ — a_2^T — \\ \vdots \end{bmatrix}$$

and $y = Ax + b$, $A \in \mathbb{R}^{m \times n}$ with rows a_i^T.

The gradient of $g(y)$ is:

$$\nabla g(y) = \frac{1}{(\sum e^{y_i})} \begin{bmatrix} e^{y_1} \\ e^{y_2} \\ \vdots \\ e^{y_m} \end{bmatrix}$$

Then

$$f(x) = g(Ax + b)$$

$$\nabla f(x) = A^T \nabla g(Ax + b)$$

$$= \frac{1}{1^T z} A^T z \quad , z_i = e^{a_i^T x + b_i}.$$

+) Second derivatives

Consider a scalar, real-valued function $f : \mathbb{R}^n \to \mathbb{R}$

- The second derivative is the Hessian matrix $\nabla^2 f(x)$:

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad \begin{array}{l} i = 1 \dots n \\ j = 1 \dots n \end{array}$$

provided that f is twice differentiable at x.

- The second-order approximation of f at or near x is a quadratic function defined as:

$$g_2(z) = f(x) + \nabla f(x)^T (z - x) + \frac{1}{2} (z - x)^T \nabla^2 f(x) (z - x)$$

This second order approximation satisfies

$$\lim_{\substack{z \to x, z \neq x \\ z \in \operatorname{dom} f}} \frac{| g_2(z) - f(x) |}{\| z - x \|_2^2} = 0$$

- The second derivative can be interpreted as the derivative of the first derivative:

$$D \nabla f(x) = \nabla^2 f(x).$$

Apply first-order approximation to $\nabla f(x)$ to get

$$g_1(z) = \nabla f(x) + \nabla^2 f(x) (z - x)$$

then the second order approximation becomes

$$g(z) = f(x) + \frac{1}{2} g_1(z)^T (z - x) + \frac{1}{2} \nabla f(x)^T (z - x)$$

Example:

i) Quadratic function:

$$f(x) = \frac{1}{2} x^T P x + q^T x + r \qquad \begin{cases} P \in S^n \\ q \in \mathbb{R}^n \\ r \in \mathbb{R} \end{cases}$$

$$\nabla f(x) = Px + q$$

$$\nabla^2 f(x) = P$$

The second-order approximation for this function is itself.

ii) $$f(X) = \log\det(X), \qquad X \in S^n_{++}$$

From before:

$$\nabla f(X) = X^{-1}$$

To find $\nabla^2 f(X)$, start with the first-order approximation of $\nabla f(X)$ at $Z = X + \Delta X$.

$$\begin{aligned} Z^{-1} &= (X + \Delta X)^{-1} \\ &= \left(X^{+1/2}(I + X^{-1/2}\Delta X X^{-1/2})X^{1/2}\right)^{-1} \\ &= X^{-1/2}(I + X^{-1/2}\Delta X\, X^{-1/2})^{-1} X^{-1/2} \\ &\cong X^{-1/2}(I - X^{-1/2}\Delta X\, X^{-1/2})\, X^{-1/2} \\ &= X^{-1} - X^{-1}\Delta X X^{-1} \end{aligned}$$

where we used $(I+A)^{-1} \approx I - A$ for small A.

Thus the first order approximation of $\nabla f(X)$ is

$$g_1(Z) = X^{-1} - X^{-1}(Z - X)X^{-1}$$

Based on this expression, the second order approximation of $f(X)$ at Z near X is

$$g_2(Z) = f(X) + \mathrm{tr}(X^{-1}(Z-X)) + \mathrm{tr}(X^{-1}(Z-X)X^{-1}(Z-X))$$

+) Chain rule for second derivatives:
The general chain rule here is cumbersome, we will only consider special cases:

- Composition with scalar function

$$f: \mathbb{R}^n \to \mathbb{R}, \quad g: \mathbb{R} \to \mathbb{R}$$
$$h = g(f(x))$$

$$\nabla h(x) = g'(f(x)) \, \nabla f(x)$$

$$\nabla^2 h(x) = g''(f(x)) \, \nabla f(x) \nabla f(x)^T + g'(f(x)) \nabla^2 f(x)$$

- Composition with affine function:

$$f: \mathbb{R}^n \to \mathbb{R}$$
$$g: \mathbb{R}^m \to \mathbb{R} \quad : \quad g(x) = f(Ax+b) \quad \begin{cases} A \in \mathbb{R}^{n\times m} \\ b \in \mathbb{R}^n \end{cases}$$

Then
$$\nabla g(x) = A^T \nabla f(Ax+b)$$
$$\nabla^2 g(x) = A^T \nabla^2 f(Ax+b) A.$$

$\underline{\text{Example:}}$ $f(x) = \log\left(\sum_{i=1}^{m} e^{a_i^T x + b_i}\right)$, $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}^n$

as in the previous example in first derivative chainrule

$$f(x) = g(Ax + b).$$

Recall

$$Dg(y) = \frac{1}{(\sum e^{y_i})} \begin{bmatrix} e^{y_1} \\ e^{y_2} \\ \vdots \\ e^{y_m} \end{bmatrix}$$

Thus

$$\nabla^2 g(y)_{ii} = \frac{e^{y_i}}{(\sum e^{y_k})} - \frac{e^{2y_i}}{(\sum e^{y_k})^2}$$

$$\nabla^2 g(y)_{ij} = \frac{-e^{y_i} e^{y_j}}{(\sum e^{y_k})^2}, \quad i \neq j$$

$$\rightarrow \nabla^2 g(y) = \operatorname{diag}(\nabla g(y)) - \nabla g(y) \nabla g(y)^T.$$

By composition:

$$\nabla^2 f(x) = A^T \left(\frac{1}{1^T z} \operatorname{diag}(z) - \frac{1}{1^T z} z z^T \right) A,$$

where $z_i = e^{a_i^T x + b_i}$,

$$f(x) = \log \sum_{i=1}^{m} z_i.$$