Note A^TA always come out as square and symmetry matrix even if $A \in \mathbb{R}^{m \times n}$

Matrix

$A \in \mathbb{R}^{m \times n}$ where m = row n = column

A^k = k^{th} row on $A \rightarrow$ (row vector) $\begin{bmatrix} a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix}$

A_k = k^{th} column on $A \rightarrow$ (column vector) $\begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{bmatrix}$ where $k \in m, n$

Transpose

$A_{ij} = A_{ji}^T$ and $A^H = \bar{A}^T$

Matrix Multiplication

Vector - Matrix Multiplication

$$(Av)_i = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \sum_{j=1}^{n} a_{ij} v_j = \begin{bmatrix} a_1 v_1 + a_2 v_2 \\ a_3 v_1 + a_4 v_2 \end{bmatrix}$$

$A \in \mathbb{R}^{m \times n}$

$v \in \mathbb{R}^n$

$Av \in \mathbb{R}^{m \times 1}$

Matrix - matrix multiplication

$$(AB)_{ik} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} b_1 & b_2 & b_5 \\ b_3 & b_4 & b_6 \end{bmatrix} = \sum_{j=1}^{n} a_{ij} b_{jk} = \begin{bmatrix} a_1 b_1 + a_2 b_3 & a_1 b_2 + a_2 b_4 & \cdots \\ a_3 b_1 + a_4 b_3 & \cdots & \cdots \end{bmatrix}$$

$A \in \mathbb{R}^{m \times n}$

$B \in \mathbb{R}^{n \times k}$

$AB \in \mathbb{R}^{m \times k}$

Dot product (between 2 vector)

$$\langle x, y \rangle = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3 = \sum_{i=j} x_i \bar{y}_j$$

(can be ignore if it's in real space)

$x \in \mathbb{R}^m$

$y \in \mathbb{R}^m$

$\langle x, y \rangle = y^H x \in \mathbb{R}^{1 \times 1}$

$$\langle x, y \rangle = y^H x = x y^H \text{ (matrix multiplication)} = \sum_{i \in} x_i \bar{y}_i$$

Dot product property

1. $\langle v, v \rangle = 0$ iff $v = 0$
2. $\langle v, w \rangle = \overline{\langle w, v \rangle}$
3. $\langle u + \lambda v, w \rangle = \langle u, w \rangle + \lambda \langle v, w \rangle$
4. $\langle w, u + \lambda v \rangle = \langle w, u \rangle + \bar{\lambda} \langle w, v \rangle$

Kronecker symbol (δ) $= \begin{cases} 1, & \text{if } i=j=1 \\ 0, & \text{if } i \neq 0 \end{cases}$ $\begin{bmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$

standard unit vector $(e^{(i)}) = (\delta_{ij})_{j \in I}$ if $I=3$ and $i=2$

$$\begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} \\ \delta_{31} & \delta_{32} & \delta_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \rightarrow e^{(2)} = [0\ 1\ 0]$$

Range of matrix $A \in \mathbb{K}^{I \times J}$ is the vector space spanned by its column

↳ $C(A)$ $\quad \text{range}(A) = \text{span}(A_{(j)}, j \in J)$ $\begin{bmatrix} a_1 \\ \vdots \\ a_j \end{bmatrix}$ span of

Orthogonal

Two vectors x, y are orthogonal ($\perp$) if $\langle x, y \rangle = 0$

Two set X, Y are mutually orthogonal if $\langle x, y \rangle = 0 \ \forall x \in X$ and $\forall y \in Y$

↳ $\begin{bmatrix} | & | & | \\ x_1 & x_2 & x_3 \cdots \\ | & | & | \end{bmatrix} \begin{bmatrix} | & | & | \\ y_1 & y_2 & y_3 \cdots \\ | & | & | \end{bmatrix}$

Set of vector X are consider Orthonormal if

1. all vector in X is orthogonal to each other
   i.e. $\langle x_1, x_2 \rangle = 0$
2. Length of each vector is 1 (normalise)
   $\| x_i \| = 1$

X is consider Orthonormal basis if

1. has orthonormal property
2. linearly independent

to be L.I
$a_1 x_1 + \dots a_n x_n = 0$
if and only if $a_1 \dots a_n = 0$

to be linear dependent
$x_i = c x_j$
vector x_i can be represent as scaler multiple of another

All Rotation matrix are orthogonal matrix

It's consider Orthogonal matrix if matrix $A \in \mathbb{R}^{I \times J}$

1. the transpose of matrix A = the inverse of matrix A

if A is rectangle → $A^H A = I$, Rectangular matrices are not invertible

if A is square → "Unitary" → $A^H A = AA^H = I$ and $A^{-1} = A^H$, A is invertible

2. $\det[A] = \pm 1$

Proof

since $AA^H = I \rightarrow \det[AA^H] = \det[I] = 1$

$\det[AA^H] = \det[A] \cdot \det[A^H] = 1$

since $\det[A^H] = \det[A] \rightarrow \det[A] \cdot \det[A] = 1$

$\det[A] = \pm 1$

3. In the orthogonal matrix, the dot product of every 2 rows and 2 columns = 0

$$A = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$\langle Col_1(A), Col_2(A) \rangle = \begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \begin{bmatrix} \frac{2}{3} \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix} = \frac{2}{9} - \frac{4}{9} + \frac{2}{9} = 0$$

$$= \langle Row_1(A), Row_2(A) \rangle = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ -\frac{2}{3} \end{bmatrix} \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix} = -\frac{2}{9} + \frac{4}{9} - \frac{2}{9} = 0$$

4. Length of all rows and columns = 1

$$\| Col_1(A) \| = \sqrt{\left(\frac{1}{3}\right)^2 + \left(-\frac{2}{3}\right)^2 + \left(\frac{2}{3}\right)^2} = \sqrt{1} = 1$$

$$\| Row_1(A) \|^2 \quad \sqrt{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 + \left(-\frac{2}{3}\right)^2} = \sqrt{1} = 1$$

Matrix Rank

No. of dimension in the column space → # of L.I columns

For $A \in \mathbb{R}^{m \times n}$

Full column rank = $r = n$
- ↳ $N(A) = \{0\}$, 0 or 1 solution exist

Full Rank = $r = n = m$

Full Row rank = $r = m$
- ↳ there exist a solution x for every b for $Ax = b$
- ↳ $N(A^T) = \{0\}$

Vector Norm ($\|v\|_1, \|v\|_2$)

→ has polynomial of degree at most 2 → magnitude (length) of the vector

A norm on V has following properties

1. $\|v\| = 0$ if and only if $\vec{v} = 0$
2. $\|cv\| = |c| \|v\| \quad \forall v \in V$ and $\forall c \in \mathbb{K}$
3. $\|v + w\| \leq \|v\| + \|w\| \quad \forall v, w \in V$ (triangle inequality)

$$\|x\|_2 = \sqrt{\sum_{i \in I} |x_i|^2} = \sqrt{\langle x, x \rangle}$$

Triangle Inequality $\|x+y\| \leq \|x\| + \|y\|$

Cauchy - Schwarz inequality $|\langle x, y \rangle| \leq \|x\| \|y\|$

Pre-Hilbert Space define by

1. $\|v\| = \sqrt{\langle v, v \rangle}$
2. inner product space (dot product)
   - $\langle v, v \rangle > 0$ for $v \neq 0$
   - $\langle v, w \rangle = \overline{\langle w, v \rangle}$ for $v, w \in V$
   - $\langle u + cv, w \rangle = \langle u, w \rangle + c\langle v, w \rangle$ for $v, u, w \in V$ and $c \in \mathbb{K}$
   - $\langle w, u + cv \rangle = \langle w, u \rangle + \bar{c}\langle w, v \rangle$ for $v, u, w \in V$ and $c \in \mathbb{K}$
3. follow Cauchy-Schwarz inequality

   $|\langle v, w \rangle| \leq \|v\| \|w\|$, $v, w \in V$

If Pre-Hilbert space is complete, then it's called Hilbert Space

What is complete? → if Cauchy Sequence converge to an element in the space and also continuous

$\forall \varepsilon > 0, \exists N \in \mathbb{N}$ where $\forall n, m > N : |a_m - a_n| < \varepsilon$

Convexity

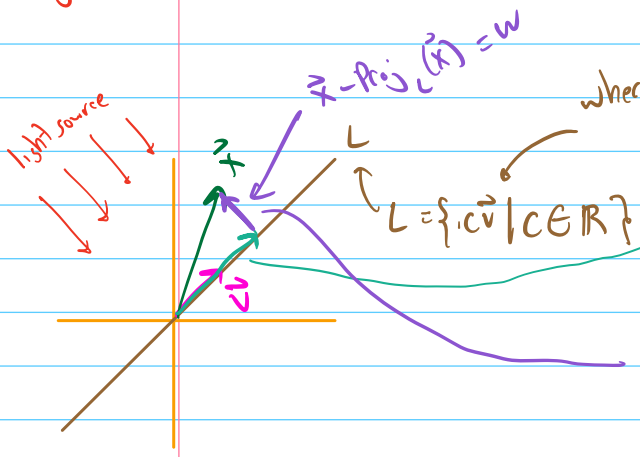
A set C is convex if $\forall v, w \in C$

$\lambda v + (1-\lambda) w \in C$ where $\lambda = [0, 1]$

Projection Explanation

"projection of v onto closed convex set C in Hilbert space"

subspace



where $\vec{v}$ is a spanning vector

$L=\{c\vec{v} \mid c\in\mathbb{R}\}$

$\text{Proj}_L(\vec{x})$ "shadow of $\vec{x}$ on L"

(how much of $\vec{x}$ goes into L direction)

this line will always be $\perp$ to the line L

also mean it's $\perp$ to every vector on the line L

$\text{Proj}_L(\vec{x})$ = some vector in L where $x-\text{Proj}_L(\vec{x})$ is $\perp$ to L

b/c $\text{Proj}_L(\vec{x})$ is some vector in L

$$\text{Proj}_L(\vec{x}) = cv$$

b/c $x-\text{Proj}_L(\vec{x}) \perp L \rightarrow L$ is define by v

$$\langle x-cv, v\rangle = 0$$

$$\langle x,v\rangle - c\langle v,v\rangle = 0$$

$$\langle x,v\rangle = c\langle v,v\rangle$$

$$c = \frac{\langle x,v\rangle}{\langle v,v\rangle}$$

$$\therefore \text{Proj}_L(\vec{x}) = cv = \frac{\langle x,v\rangle}{\langle v,v\rangle} v$$

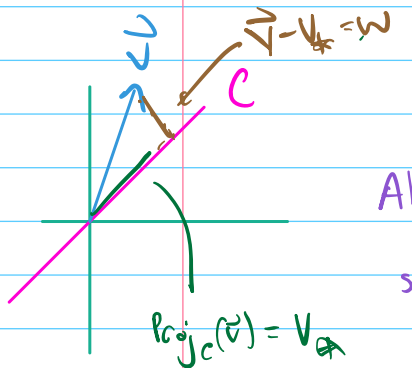
$$= \frac{\langle \vec{x},\vec{v}\rangle}{\|\vec{v}\|^2}\vec{v}$$

So the $\text{Proj}_W(v) = \frac{\langle v,w\rangle}{\|w\|^2} w$ and when $\vec{w}$ is an unit vector then $\|w\|=1$ (orthogonal basis)

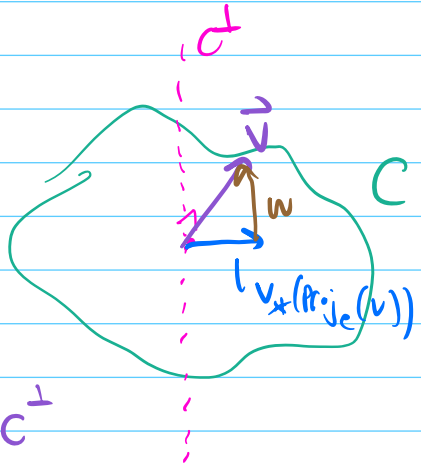
$\therefore \text{Proj}_W(v) = \langle v,w\rangle w$ for $\forall v \in V$

Explaing $P_C(v)$ is define by argmin $(\|v-w\|)$

$P_C(v)$ = is unique vector $\vec{V}_*$ in C such that $\vec{v}-\vec{V}_* = w$ where $w \perp$ every vector in C where w is a unique member of $C^\perp$



$$\vec{v} = \vec{V}_* + \vec{w}$$



Also $P_C(v)$ = some unique vector in C s.t. $\vec{v}-\vec{V}_*(w)$ is orthagonal to every member of C

$$\therefore \vec{v}-\vec{V}_* \in C^\perp \rightarrow w \in C^\perp$$

and $P_C(v)$ is define when $\|v-w\|$ is the shortest (shorter than any other w in C) which turn out to be when $w \perp C$.

(so $P_C v$ is in the set w as well $P_C v = w$)

Projection of a vector in subspace $C \in C$ ← the w that will result in the least length between $\vec{v}$ and $\vec{w}$

$$P_C(\vec{v}) = \arg\min_{w \in C} \|v-w\|$$

where $\vec{w} \in C$ as well, Any other point in $\vec{w}$ that doesn't create least length is not a projection

Writing $Proj_W(v)$ into a Matrix $Proj_W(v) = A\vec{v}$

Changing $P_W(v) = \vec{v}-\vec{w}$ into matrix formula $P_W(v) = A(A^TA)^{-1}A^T\vec{v}$

W is a subspace $\mathbb{R}^n$ and $b = \{b_1, b_2 \ldots b_k\}$

we know that $P_W(v) \in W$ where A = matrix with basis b as a column

so $P_W(v) = A\vec{y}$ for some vector $\vec{y}$ in $\mathbb{R}^k$

we also know that w is $\perp$ to all member of W, and therefore $w \in W^\perp$

matrix A is create from basis vector of W

so $C(A) = W$ and $W^\perp = C(A)^\perp = N(A^T)$ so $w \in N(A^T)$

and $\vec{w} = \vec{v} - P_W(v) \in N(A^T)$

by def. of nullspace $N(A^T) \rightarrow A^T w = 0$

$$A^T(\vec{v} - P_w(v)) = 0$$

$$A^T(\vec{v} - A\vec{y}) = 0$$

$$A^T\vec{v} - A^TA\vec{y} = 0$$

$$A^T\vec{v} = A^TA\vec{y}$$

$$(A^TA)^{-1}A^T\vec{v} = \vec{y}$$

we know A^TA is invertible cuz matrix A are make from basis vector and by def. of basis vector, they are L.I. (matrix A whose column are L.I then A^TA is always invertible)

$\therefore P_w(v) = A\vec{y}$

$$P_w(v) = A(A^TA)^{-1}A^T\vec{v}$$

same matrix that transform $\vec{v}$ into subspace w

$P_w(\vec{v}) = AA^T\vec{v}$ if A has orthonormal basis

If Subspace V has $\{\vec{v}_1, \vec{v}_2 \dots \vec{v}_k\}$ as an orthonormal basis

then $A = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_k \\ | & | & & | \end{bmatrix}$

Where A is the matrix where each of the column vectors are the basis vector for the subspace V

$$A^TA = \begin{bmatrix} -v_1- \\ -v_2- \\ \vdots \\ -v_k- \end{bmatrix}\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_k$$

cuz $v_1 \cdot v_2 = 0$ cuz of orthogonal

so $A(A^TA)^{-1}A^T\vec{x} = A(I_k)^{-1}A^T\vec{x} = AI_kA^T\vec{x}$

$$\therefore P_v(\vec{x}) = AA^T\vec{x}$$

↳ might not equal to I cus it's a dot product in row vector and row vector might not be orthogonal.

$$\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{bmatrix}\begin{bmatrix} -v_1- \\ -v_2- \\ \vdots \\ -v_k- \end{bmatrix}$$

Eigenvector and Eigenvalue

Eigenvector → the vector in transformation that only scale up or down (shrink or stretch) but not change direction

Eigenvalue → Scaling factors of eigenvector

Transformation of $\vec{v}$ = $A\vec{v} = \lambda\vec{v}$ (λ = eigenvalue, $\vec{v}$ = eigenvector)

↳ $0 = \lambda\vec{v} - A\vec{v} \rightarrow v = I_n v$

$\lambda I_n v - A\vec{v} = 0$

this same matrix as a result of I_n

$(\lambda I_n - A)\vec{v} = 0$ ← some matrix

this mean $\vec{v}$ is a member of nullspace ($A\vec{x} = 0$)

$\vec{v} \in N(\lambda I_n - A)$

we know that if $N(\lambda I - A)$ doesn't contain only $v = \{0\}$, then $\lambda I - A$ is not L.I

If $\lambda I - A$ is not L.I, then it has L.D columns → $\lambda I - A$ is not invertible → $\det = 0$

$A\vec{v} = \lambda\vec{v}$ for nonzero $\vec{v}$'s if and only if $\det(\lambda I - A) = 0$ or $\det(A - \lambda I) = 0$

if $A = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}$ then $\det\begin{bmatrix} 1-\lambda & 2 \\ 4 & 3-\lambda \end{bmatrix}$ or $\det\begin{bmatrix} \lambda-1 & -2 \\ 4 & 3-\lambda \end{bmatrix}$

$\Delta = (1-\lambda)(3-\lambda) - (2)(4) = (3 - 4\lambda + \lambda^2) - 8 = \lambda^2 - 4\lambda - 5$

$\therefore (\lambda - 5)(\lambda + 1) = 0 \quad \lambda = 5, -1$

Eigenspace - all of the eigenvectors that correspond to some eigenvalue

↳ $E_\lambda = N(A - \lambda I)$

for $\lambda = 5 \rightarrow E_5 = N\left(\begin{bmatrix} 1-5 & 2 \\ 4 & 3-5 \end{bmatrix}\right) = N\left(\begin{bmatrix} -4 & 2 \\ 4 & -2 \end{bmatrix}\right) \rightarrow N(A) = A\vec{v} = \vec{0}$

$A\vec{v} = \vec{0} = \begin{bmatrix} -4 & 2 \\ 4 & -2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$-4v_1 + 2v_2 = 0$ $\quad$ $4v_1 - 2v_2 = 0$

$4v_1 = 2v_2$ $\quad$ $4v_1 = 2v_2$

$v_1 = \frac{1}{2}v_2$ $\quad$ $v_1 = \frac{1}{2}v_2$

let $v_2 = t$

$E_5 = \left\{\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = t\begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}, t \in \mathbb{R}\right\}$

$E_5 = \text{span}\left(\begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}\right)$ → do this for $\lambda = -1$ as well

↳ this mean that if there's vector that sit along this line, the transformation A will stretch the vector by 4

Linear Operator → Matrix that transform vector from one vector space to another (Linear transformation)

Operator Norm → a measure of the largest value (change) by which the matrix A stretch the vector z

$$\|A\| = \|A\|_{X \to Y} = \sup_{z \neq 0} \frac{\|Az\|_Y}{\|z\|_X}, \quad A \in \mathbb{K}^{I \times J}$$

↑ change space from $X \to Y$

Spectral Norm → if the operator norm is done by the linear operate that transform vector from space X to Y and X and Y is coincide with Euclidean Norm → then it's called Spectral Norm

$\|A\|$ → largest singular value (or λ if A is symmetric)

$$\|A\| = \max_{\|v\|_2=1} \frac{\|Av\|_2}{\|v\|_2} = \|Av_1\|_2 = \left\|\frac{\sigma_1 u_1}{\|u_1\|_2}\right\|_2 = |\sigma_1| \|u_1\|_2$$

v_1 is the largest blow up vector (v_1 = leading singular vector)

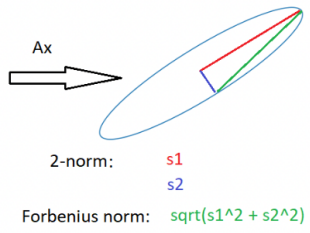
$\|u_1\|_2 = 1$

$\therefore \|A\| = \max \|Av\|_2 = \sigma_1 = \|Av_1\|_2$

v_1 = best fit line (leading principle component) with corresponding maximal value σ_1

Frobenius Norm (Norm over a matrix) → "diagonal" between all the singular value

$$\|A\|_F = \sqrt{\sum_{\substack{i \in I \\ j \in J}} |A_{ij}|^2} = \sqrt{\langle A, A \rangle_F} \quad A \in \mathbb{K}^{I \times J}$$



$$\langle A, B \rangle_F = \sum_{i \in I} \sum_{j \in J} A_{ij} B_{ij} = \mathrm{tr}(AB^H) = \mathrm{tr}(B^H A)$$

$$\|A\|_F^2 = \langle A, A \rangle_F = \mathrm{tr}(A^H A) = \mathrm{tr}(AA^H)$$

$$\|A\|_F = \sqrt{\mathrm{tr}(AA^H)} = \sqrt{\mathrm{tr}(A)^2} = \sqrt{\lambda_1^2 + \lambda_2^2 + \dots} \text{ or } \sqrt{\sum_{k=1}^{r} \sigma_k(A)^2}$$

prove

$$\sum_{i \in I} \|A^{(i)}\|_2^2 = \sum_{i \in I} \sum_{k=1}^{r} |\langle A^i, v_k \rangle|^2 = \sum_{k=1}^{r} \sum_{i \in I} |\langle A^i, v_k \rangle|^2 = \sum_{k=1}^{r} \|Av_k\|_2^2 = \sum_{k=1}^{r} \sigma_k^2 = \|A\|_F^2$$

sum of Projection onto each component of basis vector

Note if $v_1 \dots v_r$ is all row of A

Identifying largest principle component

- finding the line that best fit the data
  - Best line v_1 → Max variance between data

v_1 = best fit line = leading singular vector with σ_1

$$v_k = \max_{\substack{\|v\|_2=1 \\ v \perp v_2 \dots}} \|Av\|_2 \quad k>1$$

v_k is singular vectors and a maximizers

$$\sigma_k = \|Av_k\|_2 \quad k=1 \dots n$$

$Av_i = \sigma_i u_i$

Prove for general case v_k is the best fit subspace with $\|Av_k\| \geq \|Aw_k\|$

Assume: $V_{k-1} = \text{span}(v_1 \dots v_{k-1})$ is the best fit subspace
Prove: $V_k = \text{span}(v_1 \dots v_k)$ is the best fit subspace

let $W_k = \text{span}(w_1 \dots w_k)$ $W \neq V_k$ $w_1 \dots w_k$ is orthonormal basis and $w_k \perp v_{k-1}$

since $w_k \perp v_{k-1}$ and we know that v_k is the maximum $\|Av\|_2^2$ for any $v \perp v_{k-1}$

$\therefore \|Av_k\| \geq \|Aw_k\|$

$$\therefore \underbrace{\|Aw_1\|_2^2 + \dots + \|Aw_{k-1}\|_2^2}_{\text{from assumption}} + \|Aw_k\|_2^2 \leq \|Av_1\|_2^2 + \dots + \|Av_{k-1}\|_2^2 + \|Av_k\|_2^2$$

Prove SVD

for $v_1 \dots v_r$ singular vector let some vector $\vec{v} \in \mathbb{R}^n$

A is make up of $v_1 \dots v_r$

so $\vec{v}$ = liner comb of $v_1 \dots v_r = \sum_{i=1}^{r} \langle v, v_i \rangle v_i + \sum_{i=r+1}^{n} \langle v, v_i^{\perp} \rangle v_i^{\perp}$

borrow from nullspace to make dimension of $v = n$

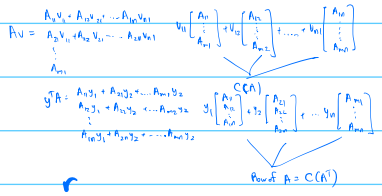
then $Av = A\left(\sum_{i=1}^{r} \langle v, v_i \rangle v_i + \sum_{i=r+1}^{n} \langle v, v_i^{\perp} \rangle v_i^{\perp}\right)$

$= \sum_{i=1}^{r} v_i^H v (Av_i) + \sum_{i=r+1}^{n} \langle v, v_i \rangle^{\perp} (Av_i^{\perp})$ → b/c $v_i^{\perp}$ is from $N(A) \rightarrow Av_i^{\perp} = 0$

$= \sum_{i=1}^{r} v_i^H v (Av_i)$ $\Rightarrow Av_i = \sigma_i u_i$

$= \sum_{i=1}^{r} v_i^H v (\sigma_i u_i) = \sum_{i=1}^{r} (\sigma_i u_i v_i^H)\, v$

$\therefore A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$



SVD

$A_{m \times n} = U_{m \times r}\, \Sigma_{r \times r}\, V^H_{r \times n}$

$A = U\Sigma V^H = \sum_{i=1}^{r} \sigma_i u_i v_i^H$

$Av_i = \sigma_i u_i$ Proof

$AV = U\Sigma$

$Ax = AV = U\Sigma = \text{span}(u_1 \dots u_r) = C(U) = C(A)$

$A^T y = A^T U = \Sigma V = \text{span}(v_1 \dots v_r) = C(V) = C(A^T)$

$C(A) = \text{span}(u_1 \dots u_r)$

$C(A^T) = \text{span}(v_1 \dots v_r)$

$N(A) = \text{span}(v_{r+1} \dots v_n)$

$N(A^T) = \text{span}(u_{r+1} \dots u_m)$

Best k-approximation

If V_k is span by $v_1 \dots v_k$ of A

then V_k is the best fit k-Dimensional subspace of A

$$V_k = \arg\min_{\dim V \le k} \sum_{i \in I} \|A^i - P_V(A^i)\|_2^2 = \arg\max_{\dim V \le k} \sum_{i \in I} \|P_V A^i\|_2^2 = \sum_{l=1}^{k} \sigma_l u_l v_l^H = A_k$$

$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^H$ $\quad k \le r$ $\quad A_k$ is the best approx. of A of any matrix of rank k

A_k are orthogonal projection of rows of A (A^i) onto subspace

$$A_k = \sum_{i \in I} P_{V_{1 \dots k}}(A^i)$$

Prove

$\langle A^i, v_l \rangle = (Av_l)_i$

$$\sum_{i \in I} \sum_{l=1}^{k} P_{v_l}(A^i) = \sum_{i \in I} \sum_{l=1}^{k} \langle A^i, v_l \rangle v_l$$

$$= \sum_{l=1}^{k} \sum_{i \in I} (A^i v_l) v_l^H = \sum_{l=1}^{k} A v_l v_l^H = \sum_{l=1}^{k} \sigma_l u_l v_l^H = A_k$$

Prove

$\|A-A_k\|_F \le \|A-B\|_F$ also $\|A-A_k\| \le \|A-B\|$

let B is of rank at most k and row of $B(B')$ span subspace V, $\dim(V) \le k$

Assume B minimize $\|A-B\|_F$

In order to minimize $B^i = P_V(A^i)$

so $\|A-B\|_F^2 = \sum_{i \in I} \|A^i - P_V(A^i)\|_2^2$

and Since $\dim(V) = k$, V_k is the best-fit subspace when $V_k = \min \sum_{i \in I} \|A^{(i)} - P_V(A^i)\|_2^2$ in k-dimension

so in order to get best fit subspace at k-dimension V_k, $B = P_V(A^i)$ which mean $B = A_k$ as well

$\|A-A_k\|^2 = \sigma_{k+1}^2$

$A - A_k = \sum_{i=k+1}^{r} \sigma_i u_i v_i^H$

$A - A_k = \sum_{j=k+1}^{r} \sigma_j u_j v_j^H$

$\|(A-A_k)v\|_2 = \left\| \sum_{j=k+1}^{r} \sigma_j u_j v_j^H (\alpha_j v_j) \right\|_2 = \left\| \sum_{j=k+1}^{r} \alpha_j \sigma_j u_j v_j^H v_j \right\|_2$

let v = top singular vector of $A - A_k$ and $v = \sum_{j=k+1}^{r} \alpha_j v_j$, $\|v\|_2 = 1$

$= \left\| \sum_{j=k+1}^{r} \alpha_j \sigma_j u_j \right\|_2 = \|u_j\|_2 \left\| \sum_{j=k+1}^{r} \alpha_j \sigma_j \right\|_2$ ($\|u_j\|_2 = 1$)

$= \sqrt{\left(\sum_{j=k+1}^{r} \alpha_j \sigma_j\right)^2}$ $\quad \|v\|_2 = 1 = \sqrt{\sum_{i=1} \alpha_j^2}$ → we can by $\alpha_j = 1$ at $k+1$ and 0 everywhere else (minimize $\|(A-A_k)v\|_2 = \|A-A_k\|$)

$= \sqrt{(1)_{\text{at } k+1} \sum_{j=k+1}^{r} \sigma_j^2} = \sqrt{\sigma_{k+1}^2} = \sigma_{k+1}$ (0 everywhere else)

b/c $k+1$ is where σ is the biggest for $A - A_k$

$\|A-A_k\|_F^2 = \sqrt{\sum_{K=1}^{r} \sigma_K (A-A_k)^2}$

Power of A

if A is square and symmetric → has same left and right singular vector

$A^2 = A^H A = AA = (V \Sigma V^H)(V \Sigma V^H) = V \Sigma^2 V^H = \sum \sigma^2 v v^H$

$\therefore A^K = V \Sigma^K V^H = \sum_{i=1}^{r} \sigma^K v_i v_i^H$

if A is not square and symmetry

let $B = AA^H$ $\quad B = (U \Sigma V^H)(V \Sigma U^H) = U \Sigma^2 U^H = \sum_{i=1}^{r} \sigma^2 u_i u_i^H$ $\quad \therefore B^m = AA^H = \sum \sigma^{2m} u_i u_i^H$

or $B = A^H A$ $\quad B = (V \Sigma U^H)(U \Sigma V^H) = V \Sigma^2 V^H = \sum \sigma^2 v_i v_i^H$ $\quad \therefore B^m = A^H A = \sum \sigma^{2m} v_i v_i^H$

$\lim_{k\to\infty} A^k = \lim_{k\to\infty} \sigma^k V_1 V_1^H$ note that $\sigma_1 \gg \sigma_2$ so if $\sigma_1 = 1$ then all $\sigma < 1$ for sure

if $\sigma > 1$ $k\to\infty$ $\sigma^k \to \infty$

if $\sigma = 1$ $k\to\infty$ $\sigma^k = \sigma_1$ constant (stable state) → most desired

if $\sigma \neq 0$ but $\sigma < 1$ $k\to\infty$ $\sigma^k \to 0$ (stability)

$\|A^m\|_F = \sqrt{\sigma_1^{2m} + \sigma_2^{2m} \dots} \approx \sigma_1^{2m}$ and $\lim_{m\to\infty} \frac{A^m}{\|A\|_F} = V_1 V_1^H$

if $\sigma_1 \gg \sigma_2 > \dots > \sigma_r$

$B^m = \sum \sigma_i^{2m} V_i V_i^H$ Normalize B^m by $\frac{B^m}{\sigma_1^{2m}} = \sum \left(\frac{\sigma_i}{\sigma_1}\right)^{2m} V_i V_i^H$ and $\frac{\sigma_1}{\sigma_1} = 1$ $\frac{\sigma_2}{\sigma_1} < 1$ $\dots$ $\frac{\sigma_r}{\sigma_1} < 1$

so when $m\to\infty$ $\frac{\sigma_2}{\sigma_1} \dots \frac{\sigma_r}{\sigma_1} \to 0$

which mean that $\frac{B^m}{\sigma_1^{2m}} \to V_1 V_1^H$ or $B^m \to \sigma_1^{2m} V_1 V_1^H$ when $m\to\infty$

let x be unit vector with significant component on u_1 direction

$B^m x = \sum \sigma_i^{2m} V_i V_i^H x$

$\frac{B^m x}{\sigma_1^{2m}} = \sum \left(\frac{\sigma_i}{\sigma_1}\right)^{2m} \underbrace{\langle V_i, x\rangle}_{\text{Projection of } x \text{ onto } V_i} V_i$ when $m\to\infty$ $\frac{B^m x}{\sigma_1^{2m}} \to \langle V_1, x\rangle V_1$ $\therefore B^m x \to \sigma_1^{2m} \langle u_1, x\rangle V_1$ when $m\to\infty$

$\therefore B^m x \approx \sigma_1^{2m} \underbrace{\langle x, V_1\rangle}_{\text{projection onto } V_1 \text{ mostly contribute to } B^m x} V_1$ as $m\to\infty$

From this we can find v_1 by normalizing $v_1 = \frac{\sigma_1^{2m}\langle x, v_1\rangle v_1}{\|\sigma_1^{2m}\langle x, v_1\rangle v_1\|}$

$A^T A v_1 = \sigma_1^2 V_1$

σ^2 is the magnitude of v_1 which is the same as the norm of $A^T A v_1$

$\sigma^2 = \|A^T A v_1\|$

$\sigma = \|\sqrt{A^T A v_1}\|$

then we can find σ_1 from $\sigma_1 = \|A v_1\|$ or $\sigma_1 = \|\sqrt{A^T A v_1}\|$

then we can find u_1 from $u_1 = \frac{A v_1}{\sigma_1}$

Randomizing $\vec{x}$ in Power Method

$\vec{x} = |\langle x, u_1 \rangle| \geq \alpha > 0$ with probability at least $1 - c\alpha\sqrt{d}$
$(P(|x| \geq \alpha))$

let V = space spanned by left singular vector for singular value $\sigma_k > (1-\varepsilon)\sigma_1$

$$w^* = \frac{(AA^H)^m x}{\|(AA^H)^m x\|_2} \quad \text{where} \quad m = \Omega\left(\frac{\ln(\frac{d}{\varepsilon})}{\varepsilon}\right)$$

$$w^* = P_V(w^*) + P_{V^\perp}(w^*)$$

$P_{V^\perp}(w^*)$ has $O\left(\frac{\varepsilon}{\alpha\sqrt{d}}\right)$ with probability at least $1 - c\alpha\sqrt{d}$

Prove that $P_{V^\perp}(w^*) \leq O\left(\frac{\varepsilon}{\alpha\sqrt{d}}\right)$

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^H \qquad \text{let } x = \sum_{k=1}^{n} \langle x, u_k \rangle u_k$$

Find Lower bound of $\|(AA^H)^m x\|_2$

1) let $\sigma_k = 0$ for $k > r$ so $A = \sum_{k=1}^{n} \sigma_k u_k v_k^H$

2) $(AA^H)^m x = \sum_{k=1}^{n} (\sigma_k)^{2m} u_k u_k^H x = \sum_{k=1}^{n} (\sigma_k)^{2m} \langle x, u_k \rangle u_k$

3) $\|(AA^H)^m x\|_2^2 = \sum_{k=1}^{n} (\sigma_k)^{4m} |\langle x, u_k \rangle|^2 \|u_k\|_2^2 = \sum_{k=1}^{n} (\sigma_k)^{4m} |\langle x, u_k \rangle|^2$

4) We know that $\sum_{k=1}^{n} \sigma_k > \sigma_1$ so

$$\|(AA^H)^m x\|_2^2 = \sum_{k=1}^{n} (\sigma_k)^{4m} |\langle x, u_k \rangle|^2 \geq \sigma_1^{4m} |\langle x, u_1 \rangle|^2$$

5) We know that $|\langle x, u_1 \rangle| \geq \alpha$

$$\|(AA^H)^m x\|_2^2 \geq \sigma_1^{4m} |\langle x, u_1 \rangle|^2 \geq \sigma_1^{4m} \alpha^2$$

this is the lower bound, it need to be at least this

Find upper bound of $(AA^H)^m x$ (or $P_{v^\perp} w^*$ in my case)

6. $\|(AA^H)^m x\|_2^2 = \sum_{k=1}^{n} (\sigma_k)^{4m} |\langle x, u_k \rangle|^2 \|u_k\|_2^2 = \sum_{k=1}^{n} (\sigma_k)^{4m} |\langle x, u_k \rangle|^2$

6.5 Choose r_2 such that $\sigma_1, \dots \sigma_{r_2}$ are the singular value of A that is greater than $(1-\varepsilon)\sigma_1$ which mean that $\sigma_{r_2+1} \dots \sigma_n < (1-\varepsilon)\sigma_1$

$P_{v^\perp} w^*$ = component of $(AA^H)^m x$ orthogonal to the space $V = \text{span}(u_1, \dots u_{r_2})$
which is the subspace $V_\perp = \text{span}(u_{r_2}, \dots u_n)$

$$\therefore \|P_{v^\perp} w^*\|_2^2 = \sum_{k=r_2+1}^{n} (\sigma_k)^{4m} |\langle x, u_k \rangle|^2$$

7. We know that $\sigma_k < (1-\varepsilon)\sigma_1$ for $k \geq r_2+1$

so $\|P_{v^\perp} w^*\|_2^2 = \sum_{k=r_2+1}^{n} \sigma_k^{4m} |\langle x, u_k \rangle|^2 \leq (1-\varepsilon)^{4m} \sigma_1^{4m} \sum_{k=r_2+1}^{n} |\langle x, u_k \rangle|^2$

8. We know that $\|x\|_2^2 = 1$ and $\|x\|_2^2 = \sum_{k=1}^{n} |\langle x, u_k \rangle|^2$, so $1 = \sum_{k=1}^{n} |\langle x, u_k \rangle|^2 \geq \sum_{k=r_2+1}^{n} |\langle x, u_k \rangle|^2$

so $\|P_{v^\perp} w^*\|_2^2 \leq (1-\varepsilon)^{4m} \sigma_1^{4m} \sum_{k=r_2+1}^{n} |\langle x, u_k \rangle|^2 \leq (1-\varepsilon)^{4m} \sigma_1^{4m}$ — This is the upperbound, it cannot be more than this

9. Now we can normalize $\|P_{v^\perp} w^*\|_2^2$

we know that $w^* = \dfrac{(AA^H)^m x}{\|(AA^H)^m x\|_2}$

so unit vector of $\|P_{v^\perp} w^*\|_2^2 = \dfrac{\|P_{v^\perp} w^*\|_2^2}{\|(AA^H)^m x\|_2^2}$ where $\|P_{v^\perp} w^*\|_2^2 \leq (1-\varepsilon)^{4m} \sigma_1^{4m}$
and $\|(AA^H)^m x\|_2^2 \geq \sigma_1^{4m} \alpha^2$

$$\|P_{v^\perp} w^*\|_2^2 \leq \frac{(1-\varepsilon)^{4m} \sigma_1^{4m}}{\sigma_1^{4m} \alpha^2} = \frac{(1-\varepsilon)^{4m}}{\alpha^2} \quad \text{and} \quad \|P_{v^\perp} w^*\| = \frac{(1-\varepsilon)^{2m}}{\alpha}$$

10. Now we know that the linear approximation of $(1-\varepsilon)$ is $e^{-\varepsilon}$ ✳

so $(1-\varepsilon)^{2m} = e^{-2m\varepsilon}$ so $\|P_{v^\perp} w^*\| = O\left(\alpha^{-1} e^{-2m\varepsilon}\right)$

11. We know that $m = \Omega\left(\frac{\ln(\frac{d}{\varepsilon})}{\varepsilon}\right)$ which mean $m \geq c\frac{\ln(\frac{d}{\varepsilon})}{\varepsilon}$ where c is some constant

we let $c = \frac{1}{2}$ to get rid of 2 in $\alpha^{-1}e^{-2m\varepsilon}$

$$\therefore \|P_W w^*\| = O\left(\frac{e^{-2\varepsilon\left(\frac{\ln(\frac{d}{\varepsilon})}{2\varepsilon}\right)}}{\alpha}\right) = O\left(\frac{\varepsilon}{\alpha d}\right) //$$

Pseudo inverse matrices

if $A \in \mathbb{R}^{m\times n}$, A does not have a real inverse

$$A^{\dagger} = V\Sigma^{\dagger}U^{H}$$

if $m > n$ (tall matrix) $\begin{bmatrix} = \\ = \end{bmatrix}$ and $n = r$ (n independent col)

$A^{\dagger}$ is left inverse $A^{\dagger} = (A^{H}A)^{-1}A^{H}$

$A^{\dagger}A = I$ $AA^{\dagger}$ is the projection onto A

$N(A) = \{0\}$, have one or zero solution to $Ax = y$ ——— $Ax = y$ $x = A^{\dagger}y$

Least Square of $A = \min \|Ax - y\|_2$ where there's only 1 unique solution $x = A^{\dagger}y$

---

if $n > m$ (wide matrix $[\ =\ ]$) and $r = m$ (m independent row)

$A^{\dagger}$ is a right inverse

$A^{\dagger} = A^{H}(AA^{H})^{-1}$ $AA^{\dagger} = AA^{H}(AA^{H})^{-1} = I$

$A^{\dagger}A$ = projection onto A

$N(A^{T}) = \{0\}$ but $N(A) \neq \{0\}$ so there's many solution to $Ax = y$

Least Square of $A = \min \|Ax - y\|_2$ where there can be many solutions $x = A^{\dagger}y$ to $\min \|Ax - y\|_2$

but we pick the solution with $\min \|x\|_2$

Stability of SVD (What's the error from approximation)

Weyl's bound

A perturbed ($\tilde{A}$) $= A + E$
(approx.)

$$\lambda_1(\tilde{A}) = \max v^H(A+E)v \le \max(v^HAv) + \max(v^HEv)$$
$$\therefore \lambda_1(\tilde{A}) \le \lambda_1(A) + \lambda_1(E)$$

spectral decomposition (eigenvalue decomp.)

if A is square and symmetric

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

where S = eigenvector matrix
Λ = eigenvalue diagonal matrix

$$\lambda_k(A) + \lambda_n(E) \le \lambda_k(A+E) \le \lambda_k(A) + \lambda_k(E) \le \lambda_k(A) + \lambda_1(E)$$
$$\lambda_n(E) \le \lambda_k(A+E) - \lambda_k(A) \le \lambda_1(E)$$

we know that $-\sigma_1 \le \lambda_n \le \lambda_1 \le \sigma_1$

$$\text{so } -\sigma_1(E) \le \lambda_n(E) \le \sigma_k(A+E) - \sigma_k(A) \le \lambda_1(E) \le \sigma_1(E)$$
$$-\sigma_1(E) \le \sigma_k(A+E) - \sigma_k(A) \le \sigma_1(E)$$

$$\therefore |\sigma_k(A+E) - \sigma_k(A)| \le \|E\| = |\sigma_1(E)|$$

Weyl's bound

$$\|E\| \ge |\sigma_k(A+E) - \sigma_k(A)|$$

Mirsky's bound

$$\|E\|_F \ge \sqrt{\sum_{k=1}^{n} |\sigma_k(A+E) - \sigma_k(A)|^2}$$

$$\|P_v - P_w\|_F^2 = \|WW^H - ww^H\|_F^2 = tr\langle (VV^H - ww^H)(VV^H - ww^H)\rangle_F$$
$$= tr(V^2)_F - 2tr(VV^HWW^H)_F + tr(W^2)_F$$
$$= \|V\|_F^2 - 2tr(V^HWW^HV) + \|W\|_F^2$$
$$= 2n - 2\|W^HV\|_F^2 = 2n - 2\langle W^HV, W^HV\rangle_F$$
$$= 2n - 2\cos^2\theta(V,W)$$
$$= 2\sum_{k=1}^{n}(1 - \cos^2\theta(v,w)_k) = 2\sum_{k=1}^{n}\sin^2\theta(v,w)_k$$
$$= 2\|\sin\theta(v,w)\|_2^2$$
$$\|P_v - P_w\|_F^2 = 2\|\sin\theta(v,w)\|_2^2$$
$$\|P_v - P_w\|_F = \sqrt{2}\|\sin\theta(v,w)\|_2$$

$\cos\theta(v,w) = w^Tv$

$$\langle v, w\rangle = w^Tv = \|w\|_2\|v\|_2\cos\theta(w,v)$$

since $\|w\|_2 = \|v\|_2 = 1$

$$w^Tv = \cos\theta(w,v) \rightarrow \cos\theta(w,v) = \sum_{i=1}^{n} w_i^Tv_i = \text{sum of projection of all dimension}$$

Stability of Singular Space (Wedin's bound)

Non stable for SVD mean :

→ the matrices that are close in norm can have completely difference singular vector (Hence, the stability of singular space)

Stable SVD

= matrices that are close in norm are close in singular vector (singular space) as well

From Exercise 7 we show that

If A has a non stable SVD

$\tilde{A} = A + E$ adding perturbation E onto A will

1. norm of $\tilde{A}$ is changed based on how big ε in E is. → dependent of ε
   → we can lower the perturbation ε to get norm of $\tilde{A}$ as close as possible to norm of A

2. Singular Vector of A is transform independent of ε in perturbation
   → so Singular Vector of A is forever change and we cannot reduce the effect of change
   ↑ even with the smallest ε

If A has stable SVD

both norm and singular vector is dependent on ε

$\tilde{A} = A + E \qquad E = \tilde{A} - A$

$\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^H \rightarrow \tilde{A}\tilde{V}_1 = \tilde{A}\tilde{V}_1 = \tilde{U}_1\tilde{\Sigma}_1$

$R_{11} = A\tilde{V}_1 - \tilde{A}\tilde{V}_1 = A\tilde{V}_1 - \tilde{\Sigma}_1\tilde{U}_1$

$= \tilde{V}_1(A - \tilde{A})$

$= \tilde{V}_1(A - (A+E))$

$R_{11} = -E\tilde{V}_1$

$R_{12} = A^H\tilde{U}_1 - \tilde{A}^H\tilde{U}_1 = A^H\tilde{U}_1 - \tilde{\Sigma}_1^H\tilde{V}_1$

$= \tilde{U}_1(A^H - \tilde{A}^H)$

$= \tilde{U}_1(A^H - (A^H + E^H))$

$R_{12} = -E^H\tilde{U}_1$

$\|R_{11}\| = \|-E\tilde{V}_1\| = \|E\| = \sigma_1(E)$

Normal distribution

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

$P(|x| \geq t) = P(x \geq t) + P(x \leq -t)$

$P(|x| < t) = P(-t < x < t)$

Basic Probability

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot \phi(t)\, dt \qquad P(a<x<b) = \int_a^b \phi(x)\, dx$$

$P(w \in B) = \frac{\text{Vol}(B)}{\text{Vol}(\Omega)}$ the prob. that x hit a point $w \in B$ is the ratio of area of B to entire area (Ω)
(a whole)

Union Bound = $P(A \cup B) \leq P(A) + P(B)$

$$\phi(X+Y)(t) = \int_{-\infty}^{\infty} \phi(X(u))\, \phi(Y(t-u))\, du \rightarrow \text{if } X, Y \sim N(0,1) \quad X+Y \sim N(0,2) \text{ sd} = \sqrt{2}$$

Gaussian vector - random vector $X = (X_1, \ldots X_n)$ with all its entries x_i $\quad X_i \sim N(0,1)$

Gaussian distribution → rotation invariant → if $X \sim N(0, I)$ and u is random variable
↳ not care about direction, $Xu \sim N(0, \|u\|_2^2)$ if u is fixed or is orthogonal matrix
then $Xu \sim N(0, I)$

Sum of Gaussian variable

if $X = (X_1, \ldots X_n)$, x_i is independent and $x_i \sim N(\mu_i, \sigma_i^2)$

then $Z = \sum_{i=1}^{n} X_i$, $Z \sim N(\sum \mu_i, \sum \sigma_i^2)$

Variance = $\text{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

Moments of $X = \mathbb{E}(X^p)$

Cavalieri's formula

Absolute Moment of $X = \mathbb{E}(|X|^p) = p \int_0^{\infty} P(|x| \geq t)\, t^{p-1}\, dt$

$(\mathbb{E}|X+Y|^p)^{\frac{1}{p}} \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} + (\mathbb{E}|Y|^p)^{\frac{1}{p}}$

$\mathbb{E}(|X+Y|^p)^{\frac{1}{p}} \leq \mathbb{E}(|X|^p)^{\frac{1}{p}} + \mathbb{E}(|Y|^p)^{\frac{1}{p}}$

$\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|Y|^q)^{\frac{1}{q}}$ if $p, q \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$

$\mathbb{E}(|XY|) \leq \mathbb{E}(|X|^p)^{\frac{1}{p}}\, \mathbb{E}(|Y|^q)^{\frac{1}{q}}$ if $\frac{1}{p} + \frac{1}{q} = 1$

Markov Inequality (estimate tail w/ upperbound)

$$P(|X| \geq t) \leq \frac{\mathbb{E}(X)}{t}$$

$$P(|X| \geq t) = P(|X|^p \geq t^p) \leq \frac{\mathbb{E}(|X|^p)}{t^p}$$

$$P(|X| \geq t) = P(\exp(\theta X) \geq \exp(\theta t)) \leq \frac{\mathbb{E}(\exp(\theta X))}{\exp(\theta t)}$$

Moment generating function of X

Cumulative generating function → $C_x(\theta) = \ln \mathbb{E}(\exp(\theta x))$

Cramer's Theorem (estimate tail w/ upper bound) → Proof

if $X = (X_1, \dots X_N)$ and X is independence

$$P\left(\sum_{i=1}^{N} X_i \geq t\right) \leq \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{i=1}^{N} C_{X_i}(\theta)\right\}\right)$$

Hoeffding's inequality (estimate tail if X_i is bounded)

$|X_i| \leq B_i \quad i \in [N]$

$$P\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{N} B_i^2}\right)$$

Taylor Series of exponential

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Bernstein's inequality

Inequality for 1. sum of bound or not bound independent random variable ($|X_\ell| < K$)

2. Subexponential random variable

It's Bernstein

Lif $X_1, \dots X_m$ be independent mean zero random variable

with → $\mathbb{E}|X_\ell|^n \leq \frac{n!}{2} R^{n-2} \sigma_\ell^2$

not

1.1 Sum of bound indep. random variable

then $$P\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + Rt)}\right)$$

where $\sigma = \sum_{\ell=1}^{M} \sigma_\ell^2$

1.2 if $|X_\ell| < K$ X_ℓ is bounded, then

$$P\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + \frac{1}{3}Kt)}\right) \quad (R = \tfrac{1}{3}K)$$

Proof

if $\mathbb{E}|X_\ell|^n \leq \frac{n!}{2} R^{n-2} \sigma_\ell^2 \quad |X_\ell| < K$

for $n=2$, $\mathbb{E}|X_\ell|^2 \leq \frac{2!}{2} R^0 \sigma_\ell^2 = \sigma_\ell^2$

for $n \geq 3$, $\mathbb{E}|X_\ell|^n = \mathbb{E}\left(|X_\ell|^{n-2}|X_\ell|^2\right) < \mathbb{E}\left(K^{n-2}|X_\ell|^2\right) = K^{n-2}\mathbb{E}|X_\ell|^2 \leq K^{n-2}\sigma_\ell^2$

so $\mathbb{E}|X_\ell|^n \leq K^{n-2}\sigma_\ell^2 = \frac{n!}{n!} K^{n-2}\sigma_\ell^2 \leq \frac{n!}{2\cdot 3^{n-2}} \cdot K^{n-2}\sigma_\ell^2 = \frac{n!}{2}\left(\frac{K}{3}\right)^{n-2}\sigma_\ell^2$ $\therefore R = \frac{K}{3}$ and $\sigma_\ell = \sigma_\ell$ □

($n! \geq 2 \cdot 3^{n-2}$)

Proof

$$P\left(\sum_{i=1}^{N} X_i \geq t\right) = P\left(\exp\left(\theta \sum_{i=1}^{N} X_i\right) \geq \exp(\theta t)\right) \leq \frac{\mathbb{E}\left(\exp\left(\sum \theta X_i\right)\right)}{\exp(\theta t)}$$ (Markov)

$$= e^{-\theta t}\,\mathbb{E}\left(\prod_{i=1}^{N} \exp(\theta X_i)\right) \overset{\text{independence}}{=} e^{-\theta t}\prod_{i=1}^{N}\mathbb{E}(\exp(\theta X_i))$$

$$= e^{-\theta t}\prod_{i=1}^{N}\exp\left(\ln\left(\mathbb{E}(\exp(\theta X_i))\right)\right) = e^{-\theta t}\prod_{i=1}^{N}\exp(C_{X_i}\theta) = e^{-\theta t}e^{\sum C_{X_i}\theta}$$

$$= \exp\left(-\theta t + \sum_{i=1}^{N} C_{X_i}\theta\right) = \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{i=1}^{N} C_{X_i}\theta\right\}\right)$$

Proof

$X_i \in [-B_i, B_i]$ so $X_i = t(-B_i) + (1-t)B_i$

Since $f(x) = \exp(\theta x)$ $\quad t = \frac{B_i - X_i}{2B_i}$

$$f(X_i) = f\left(t(-B_i) + (1-t)(B_i)\right) \leq tf(-B_i) + (1-t)f(B_i)$$

$$\exp(\theta X_i) \leq \frac{B_i - X_i}{2B_i}e^{-\theta B_i} + \frac{B_i - X_i}{2B_i}e^{\theta B_i}$$

$$\mathbb{E}(\exp(\theta X_i)) \leq \mathbb{E}\left(\frac{B_i - X_i}{2B_i}e^{-\theta B_i} + \frac{B_i - X_i}{2B_i}e^{\theta B_i}\right)$$

$$= e^{-\theta B_i}\frac{B_i}{2B_i} - \frac{\mathbb{E}(X_i)}{2B_i} + e^{\theta B_i}\frac{B_i}{2B_i} - \frac{\mathbb{E}(X_i)}{2B_i}$$ ($\mathbb{E}(X_i) = 0$)

$$= \frac{1}{2}e^{-\theta B_i} + \frac{1}{2}e^{\theta B_i} \qquad e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$

$$= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\theta B_i)^k}{k!} + \sum_{k=0}^{\infty}\frac{(\theta B_i)^k}{k!}\right)$$

$$= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{2(\theta B_i)^{2k}}{(2k)!}\right) = \sum_{k=0}^{\infty}\frac{(\theta B_i)^{2k}}{(2k)!} \qquad ((2k)! \geq 2^k k!)$$

$$\mathbb{E}(\exp(\theta X_i)) \leq \frac{(\theta B_i)^{2k}}{2^k k!} = \sum_{k=0}^{\infty}\frac{\left(\frac{(\theta B_i)^2}{2}\right)^k}{k!} = e^{\frac{\theta^2 B_i^2}{2}}$$

$$C_{X_i}(\theta) = \ln(\mathbb{E}(e^{\theta X_i})) = \ln\left(e^{\frac{\theta^2 B_i^2}{2}}\right)$$

From Cramer's Rule

$$P\left(\sum_{i=1}^{N} X_i \geq t\right) \leq \exp\left(\inf_{\theta>0}\left(-\theta t + \sum_{i=1}^{N} C_{X_i}\theta\right)\right) \leq \exp\left(\inf_{\theta>0}\left(-\theta t + \sum_{i=1}^{N}\frac{\theta^2 B_i^2}{2}\right)\right)$$

Optimal $\theta = \frac{t}{\sum_{i=1}^{N} B_i^2}$

$$P\left(\sum_{i=1}^{N} X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{N} B_i^2}\right)$$

$P\left(\sum_{i=1}^{N} X_i \leq -t\right)$ also give the same bound

$$\therefore P\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) = P\left(\sum_{i=1}^{N} X_i \geq t\right) + P\left(\sum_{i=1}^{N} X_i \leq -t\right) = 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{N} B_i^2}\right)$$

Note

$$\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,dt = \sqrt{2\pi} \qquad \int_{0}^{\infty} e^{-\frac{t^2}{2}}\,dt = \frac{\sqrt{2\pi}}{2}$$

$$\int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}$$

2) Subexponential random variable

random variable X is subexponential if

$$\mathbb{P}(|X| > t) \le \beta e^{-\kappa t} \qquad \beta, \kappa > 0$$

then $$\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \ge t\right) \le 2\exp\left(-\frac{(\kappa t)^2}{2(2\beta M + \kappa t)}\right)$$

Proof $\mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2(\sigma^2 + Rt)}\right)$

$$\mathbb{E}(\exp(\theta X_\ell)) = 1 + \theta\,\mathbb{E}(X) + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}(X^n)}{n!} = 1 + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}(X^n)}{n!} = 1 + \frac{\theta^2\sigma_\ell^2}{2} \sum_{n=2}^{\infty} \frac{2\theta^{n-2}\mathbb{E}(X^n)}{n!\,\sigma_\ell^2}$$

(Taylor series of exponential; $\mathbb{E}(X) = 0$)

→ let $F_\ell(\theta) = \sum_{n=2}^{\infty} \frac{2\theta^{n-2}\mathbb{E}(X^n)}{n!\,\sigma_\ell^2}$, and $F(\theta) = \max_{\ell \in [M]} F_\ell(\theta)$ so $F_\ell(\theta) \le F(\theta)$

$$\mathbb{E}(\exp(\theta X_\ell)) = 1 + \frac{\theta^2\sigma_\ell^2}{2} F_\ell(\theta) \le 1 + \frac{\theta^2\sigma_\ell^2}{2} F(\theta) \le \exp\left(\frac{\theta^2\sigma_\ell^2 F(\theta)}{2}\right)$$

(Taylor series of exponential)

→ $$C_{X_\ell}\theta = \ln \mathbb{E}(\exp(\theta X_\ell)) \le \ln \exp\left(\frac{\theta^2\sigma_\ell^2 F(\theta)}{2}\right) = \frac{\theta^2\sigma_\ell^2 F(\theta)}{2}$$

So, using Cramer's theorem → $\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \ge t\right) \le \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{\ell=1}^{M} \frac{\theta^2\sigma_\ell^2 F(\theta)}{2}\right\}\right) = \exp\left(\inf_{\theta>0}\left\{-\theta t + \frac{\theta^2\sigma^2 F(\theta)}{2}\right\}\right)$

($\sum_{\ell=1}^{M} \sigma_\ell^2 = \sigma^2$)

We know that $\mathbb{E}(X_\ell^n) \le \mathbb{E}(|X_\ell|^n)$

so $$F_\ell(\theta) \le F(\theta) \le \sum_{n=2}^{\infty} \frac{2\theta^{n-2}\mathbb{E}(|X_\ell|^n)}{n!\,\sigma_\ell^2}$$ and from (def of) Bernstein's inequality $\mathbb{E}(|X_\ell|^n) \le \frac{n!}{2} R^{n-2}\sigma_\ell^2$

$$F(\theta) \le \sum_{n=2}^{\infty} \frac{2\theta^{n-2}}{n!\,\sigma_\ell^2}\left(\frac{n!}{2} R^{n-2}\sigma_\ell^2\right) = \sum_{n=2}^{\infty} (\theta R)^{n-2} = \frac{1}{1-\theta R} \qquad \left(\text{Taylor's Series } x^n = \sum_{n=1}^{\infty} \frac{1}{1-x}\right)$$

So $$\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \ge t\right) \le \exp\left(\inf_{\theta>0}\left\{-\theta t + \frac{\theta^2\sigma^2 F(\theta)}{2}\right\}\right) \le \exp\left(\inf_{0<\theta R<1}\left\{-\theta t + \frac{\theta^2\sigma^2}{2(1-\theta R)}\right\}\right)$$

letting $\theta = \frac{t}{\sigma^2 + Rt}$ will result in $\mathbb{P}\left(\sum_{\ell=1}^{M} X_\ell \ge t\right) \le \exp\left(\frac{t^2}{2(\sigma^2 + Rt)}\right)$

changing X_ℓ to $-X_\ell$ will yield the same result, so applying union bound complete the proof

$$\therefore \mathbb{P}\left(\left|\sum_{\ell=1}^{M} X_\ell\right| \ge t\right) \le 2\exp\left(\frac{t^2}{2(\sigma^2 + Rt)}\right)$$

Dimensionality Reduction : The Johnson - Lindenstrauss Lemma

1. Johnson - Lindenstrauss Lemma (JL) by random Projection

For any $0 < \varepsilon < 1$, any integer n, let k be positive integer such that

$$k \geq 2\beta \left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right)^{-1} \ln n$$

$\hookrightarrow 2[\varepsilon^2 - \varepsilon^3]^{-1}$ for gaussian (to prove easier ??)

then for set P of n points in $\mathbb{R}^d$, there's a map $f : \mathbb{R}^d \to \mathbb{R}^k$ s.t for all point $v, w \in P$

$$(1-\varepsilon)\|v-w\|_2^2 \leq \|f(v)-f(w)\|_2^2 \leq (1+\varepsilon)\|v-w\|_2^2$$

The map f, can be generate at random with prob $= 1 - (n^{2-\beta} - n^{1-\beta})$ (choosing β large will have higher prob)

Considering unit vector $z = \frac{v-w}{\|v-w\|_2}$, A = orthogonal projection matrix onto dim k in $\mathbb{R}^d$

$$(1-\varepsilon) \leq \|Az\|_2^2 \leq (1+\varepsilon)$$
$$\left| \|Az\|_2^2 - 1 \right| \leq \varepsilon$$

2. Johnson - Lindenstrauss for Gaussian random matrices

let matrix A have i.i.d. entries and $N \sim (0,1)$ $A \in \mathbb{R}^{k \times d}$ and $x \in \mathbb{R}^d$

then $$\mathbb{P}\left(\left| \left\| \frac{1}{\sqrt{k}} Ax \right\|_2^2 - \|x\|_2^2 \right| > \varepsilon \|x\|_2^2 \right) \leq 2e^{-(\varepsilon^2 - \varepsilon^3)\frac{k}{4}}$$

or

$$\mathbb{P}\left((1-\varepsilon)\|x\|_2^2 \leq \left\| \frac{1}{\sqrt{k}} Ax \right\|_2^2 \leq (1+\varepsilon)\|x\|_2^2 \right) \geq 1 - 2e^{-(\varepsilon^2 - \varepsilon^3)\frac{k}{4}}$$

Convex

Convex Set $K \subset \mathbb{R}^N$, K is a convex set if for all $x, y \in K$

$tx + (1-t)y \in K$ for all $t \in [0,1]$

Convex Combination $x = t_1x_1 + t_2x_2 + \dots + t_kx_k$ with $t_1 + \dots + t_k = 1$, $\dots \geq 0$

Convex Hull - boundary of the sets of point

- Convex Hull $\text{conv}(T)$ - set of all convex combinations of point in T

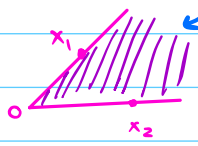
(Smallest convex set containing T)

$$\text{conv}(T) = \left\{ \sum_j t_j x_j : t \geq 0, \ \sum_j t_j = 1, \ x_j \in T \right\}$$

Convex Cone

Conic combination - any point of the form

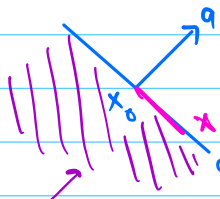
$x = t_1x_1 + t_2x_2$ $\quad t_1, t_2 \geq 0$



Convex Cone - set that contain all conic combination of points in the set

Hyperplane $\{x \mid a^Tx = b\}$ $\quad a \neq 0$

↳ convex



$\langle a, x_0 - x \rangle = 0$ create a hyperplane

$a_1x_1 + a_2x_2 + \dots = 0$

$a^Tx = b$

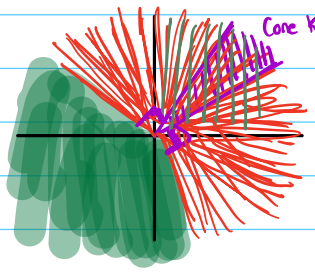
Halfspace $\{x \mid a^Tx \leq b\}$

↳ convex

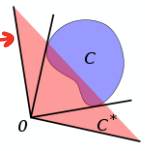
or $a^Tx \geq b$ (depend)

Dual Cones

For cone K, dual cone $K^* = \{z \in \mathbb{R}^N : \langle x, z \rangle \geq 0 \text{ for all } x \in K\}$

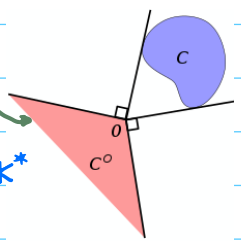


Dual cone K^* ($< 90°$ of cone)



Polar Cone

$K^\circ = \{z \in \mathbb{R}^N : \langle x, z \rangle \leq 0 \text{ for all } x \in K\} = -K^*$



Conic Hull

cone(T) of set $T \subset \mathbb{R}^N$ is the smallest convex cone containing T

$$\text{cone}(T) = \left\{ \sum_j t_j x_j \geq 0,\ x_j \in T \right\}$$

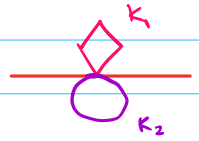
Geometric Hahn-Banach Theorem

non-intersecting convex sets can be seperate by hyperplanes.

Let $K_1, K_2 \subset \mathbb{R}^N$ be convex set, $w \in \mathbb{R}^N$

$$K_1 \subset \{ x \in \mathbb{R}^N : \langle x, w \rangle \leq \lambda \}$$

$$K_2 \subset \{ x \in \mathbb{R}^N : \langle x, w \rangle \geq \lambda \}$$

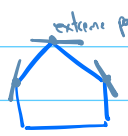


Hyperplane seperate K_1 and K_2 w/ distance λ

Extreme Points

Point on convex that you cannot build a convex combination (convex line is not inside the graph)

(e.g. vertex or corner point)



if K does not lie between 2 distinct points of K

$x = tw + (1-t)z$, $w, z \in K$, $t \in (0,1)$ and $x = w = z$

Proper Functions

Domain of function — set of all possible inputs for the function

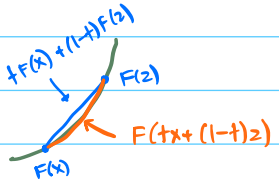
$$\text{dom}(F) = \{ x \in \mathbb{R}^N : F(x) \neq \infty \}$$

$\text{dom}(F) \neq \emptyset$ is called proper.

(input is not empty)

Convex Function $t \in [0,1]$

$$F(tx + (1-t)z) \leq tF(x) + (1-t)F(z)$$



Strictly convex $x \neq z$ and $t \in (0,1)$

$$F(tx + (1-t)z) < tF(x) + (1-t)F(z)$$

is strictly convex

is not

$F(tx + (1-t)z) = tF(x) + (1-t)F(z)$

Strongly convex $\gamma > 0$ $t \in [0,1]$

$$F(tx + (1-t)z) \leq tF(x) + (1-t)F(z) - \frac{\gamma}{2} t(1-t) \|x - z\|_2^2$$

Strongly convex are always strictly convex

Strictly convex are always convex

Note

Tangent line formula

$y - y_1 = m(x - x_1)$ m = slope at tangent point

or

$y = y_1 + \nabla f(x_1)(x - x_1)$ x_1, y_1 = tangent point

Domain of convex function is convex

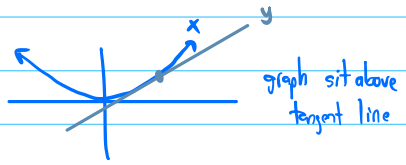
Function F is convex if and only if its epigraph is a convex set

Epigraph - Area (all point) above the graph

$$epi(F) = \{(x,r) : r \geq F(x)\} \subset \mathbb{R}^N \times \mathbb{R}$$

Smooth Convex Function

if F is differentiable



1. F is convex iff $F(x) \geq F(y) + \langle \nabla F(y), x - y \rangle$ (Tangent line)

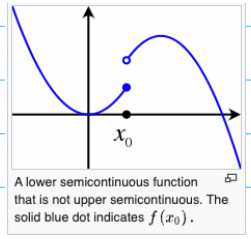
2. F is strongly convex w/ parameter $\gamma > 0$ iff

$$F(x) \geq F(y) + \langle \nabla F(y), x - y \rangle + \frac{\gamma}{2} t(1-t) \| x - z \|_2^2$$

3. If F is twice differentiable then F is convex iff

$$\nabla^2 F(x) \geq 0$$

Lower Semicontinuous



A lower semicontinuous function that is not upper semicontinuous. The solid blue dot indicates $f(x_0)$.

f is LSC iff epi(f) is closed set

$$\lim_{x \to x_0} F(x) \geq F(x_0)$$

sequence of $\{x\}_{x \to x_0}$ converge to $F(x_0)$

Every Norm is convex

ℓ_p-norms are strictly convex (except ℓ_1 and ℓ_∞ — only convex)

$A \in \mathbb{R}^{N \times N}$ $F(x) = x^T A x$ is positive semidefinite = convex

if positive definite = strictly convex

Jointly convex function

$$f(tx_1 + (1-t)x_2, ty_1 + (1-t)y_2) \leq tf(x_1,y_1) + (1-t) f(x_2,y_2)$$

If $g(x) = \inf_{y \in \mathbb{R}^m} f(x,y)$ and $f(x,y)$ is a jointly convex function,

then $g(x)$ is a convex function

Maxima of convex function

F attain its maxima at an extreme point of K

(K is a convex hull (compact convex set))

↳ closed and bounded

closed = contain all boundary point in the set

bounded = no points going to infinity

Convex Conjugate (F^*)

$$F^*(y) := \sup_{x \in \mathbb{R}^N} \{\langle x, y\rangle - F(x)\} \rightarrow \langle x, z\rangle - F(z) \leq \langle x, y\rangle - F(x) \quad \forall z \in \mathbb{R}^n$$

only equal when $z = x$

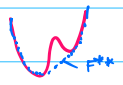
Fenchel's inequality $\Rightarrow \langle x, y\rangle \leq F(x) + F^*(y)$ for all $x, y \in \mathbb{R}^N$

F^* is always a convex function

Property of Convex Conjugate

1. F^* is lower semicontinuous $\rightarrow \lim_{x \to x_0} F(x) \geq F(x_0)$ (the actual point is lower than point that x converge to)
2. F^{**} is the largest lower semicontinuous convex function satisfying $F^{**}(x) \leq F(x)$ for all $x \in \mathbb{R}$
   - fit under the graph
   - convex hull of epigraph of F

   if F is convex and lower semicontinuous, $F = F^{**}$



$\text{epi}(F^{**}) = \text{conv}(\text{epi } F)$

3. For $\tau \neq 0$ let $F_\tau(x) = F(\tau x)$ then $(F_\tau)^*(y) = F^*(\frac{y}{\tau})$
4. For $\tau > 0$, $(\tau F)^*(y) = \tau F^*(\frac{y}{\tau})$
5. For $z \in \mathbb{R}^N$ let $F^{(z)} = F(x - z)$. then $(F^{(z)})^*(y) = \langle z, y\rangle + F^*(y)$

Subdifferential

- There are multiple derivatives at a non-smooth point

  e.g. $F(x) = |x|$ can have slope at any value from -1 to 1 at $x = 0$

- So the generalise gradient

$$\partial F(x) = \{v \in \mathbb{R}^N : F(z) - F(x) \geq \langle v, z - x\rangle \text{ for all } z \in \mathbb{R}^N\}$$

$\partial F(x)$: subgradient of F at x; v: a slope; $v \in \partial F(x)$

$\partial F(x)$ of convex function is always non empty

if F is differentiable (smooth point) in x, then

$\partial F(x) = F(z) - F(x) = \langle v, z - x\rangle$

and $\partial F(x) = \nabla F(x)$

Subdifferentials and Conjugation

F is convex function

All this conditions are equivalent

1. $y \in \partial F(x)$
2. $F(x) + F^*(y) = \langle x, y\rangle$

if F is lower semicontinuous, then also equivalent to

3. $x \in \partial F^*(y)$

Proximal mapping

function $x \longmapsto F(x) + \frac{1}{2}\|x-z\|_2^2$ is strictly convex

because $x \longmapsto \|x\|_2^2$ is strictly convex

B/c it's strictly convex, minimizer is unique

$$P_F(z) := \arg\min \left\{ F(x) + \frac{1}{2}\|x-z\|_2^2 : x \in \mathbb{R}^n \right\}$$

↑ Proximal mapping associate with F

and $x = P_F(z)$ iff $z \in x + \partial F(x)$

Moreau's identity

$$P_F = (I + \partial F)^{-1}$$

if F is a lower semicontinuous convex function

$$P_F(z) + P_{F^*}(z) = z$$

if $P_F(z)$ is easy to compute, then $P_{F^*}(z) = z - P_F(z)$

$$P_{\tau F}(z) + \tau P_{\tau^{-1}F}\left(\frac{z}{\tau}\right) = z$$

Lagrange function (Try to solve minimization problem w/o constrain)

Normal Convex problems has lot of constrain, so we form Lagrange function with the convex problem + constrain instead to have problem with no constrain

if Convex problem (primal problem) $= \min_{x \in R} F_0(x)$ subject to $Ax = b$ → $Ax - b$

and $F_i(x) \le b_i$ → $F_i(x) - b_i \le 0$, $i \in [M]$

The Lagrange function $L(x, \xi, \nu) = F_0(x) + \xi^*(Ax-b) + \sum_{i=1}^{M} \nu_i (F_i(x) - b_i)$ without constrain

(ξ: Lagrange multiplier)

if there's no inequality constrain

then $L(x, \xi) = F_0(x) + \xi^*(Ax-b)$

Lagrange Dual (Provide a lower bound on the optimal value of $F_0(x^\#)$ of the minimization problem)

($x^\#$: minimizer)

Lagrange Dual Function (H) (Find lower bound of minimization using Lagrange)

$$H(\xi, \nu) = \inf_{x \in \mathbb{R}^n} (L(x, \xi, \nu))$$

Lagrange Dual Function is affine function

Infimum of family of affine function is always concave.

$H(\xi, \nu) \le F(x^\#)$ for all $\xi \in \mathbb{R}^m$, $\nu \ge 0$

if primal problem = $\min_{x \in \mathbb{R}^n} F_0(x)$ subject to $\underbrace{Ax = b}_{Ax-b=0}$ and $\underbrace{F_\ell(x) \le b_j}_{F_\ell(x)-b_j \le 0}$ $\ell \in [M]$

then

$$L(x, \xi, \nu) = F_0(x) + \xi^*(Ax-b) + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(x) - b_j)$$

for all $\xi \in \mathbb{R}^m$ and $\nu \ge 0$ and x is feasible

$$\xi^*(Ax-b) = 0 \quad \text{and} \quad \nu_\ell \underbrace{(F_\ell(x)-b_\ell)}_{\le 0} \le 0$$

So $L(x, \xi, \nu) = F_0(x) + \xi^*(Ax-b) + \sum_{\ell=1}^{M} \nu_\ell (F_\ell(x)-b_j) \le F_0(x)$

$$L(x, \xi, \nu) \le F_0(x)$$

$$\inf_{x \in \mathbb{R}^n} L(x, \xi, \nu) \le \inf_{x \in \mathbb{R}^n} F_0(x)$$

$$\underbrace{H(\xi, \nu) \le F(x^\#)}_{\text{supremum}}$$

Lagrange Dual Problem (finding the best lower bound for the optimal value)

$$\max H(\xi, \nu) \quad \text{subject to } \nu \ge 0$$

A feasible maximizer $(\xi^\#, \nu^\#)$ is dual optimal

If $x^\#$ is optimal for primal problem, $(x^\#, \xi^\#, \nu^\#)$ is primal dual optimal

Weak Duality

$H(\xi^\#, \nu^\#) \le F(x^\#)$ best lower bound $\le$ optimal value

Strong Duality

$H(\xi^\#, \nu^\#) = F(x^\#)$

Slater's constraint qualification

if $\exists x$ that is strictly feasible $(F_\ell(x) < b_j)$ then strong duality hold for optimization problem $H(\xi^\#, \nu^\#) = F(\nu^\#)$

Saddle point

$$\sup_{\xi} L(x, \xi) = \sup_{\xi} \left(F_0(x) + \xi^*(Ax-y)\right) = \begin{cases} F_0(x) & \text{if } Ax = y \\ \infty & \text{ow.} \end{cases}$$

in other words, supremum of Lagrange function is ∞ if x is not feasible

get the lowest value of feasible x

minimize value of F

get value of feasible x (not feasible x → value = ∞)

$$F_0(x^\#) = \inf_{x \in \mathbb{R}^n} \sup_{\xi \in \mathbb{R}^m} L(x, \xi)$$

maximize the Lagrange dual function to have the best lower bound to optimal value $F_0(x^\#)$

$$H(\xi^\#) = \sup_{\xi \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^n} L(x, \xi)$$

Get the lower bound for $F_0(x^\#)$ ($H(\xi)$)

Get the best lower bound = $H(\xi^\#)$

Weak duality → $H(\xi^\#) \leq F_0(x^\#)$

Strong duality → $H(\xi^\#) = F_0(x^\#)$

$\inf_{x \in \mathbb{R}^n} \sup_{\xi \in \mathbb{R}^m} L(x, \xi) = \sup_{\xi \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^n} L(x, \xi)$ → this mean, in strong duality, min-max is interchangable. → Saddle point property!

if a primal dual optimal $(x^\#, \xi^\#)$ is a saddle point of Lagrange function

$$L(x^\#, \xi) \leq L(x^\#, \xi^\#) \leq L(x, \xi^\#)$$