08/25/22 Thursday.

Least-Squares.

$\vec{x}^* = \underset{\vec{x}}{\operatorname{argmin}} \|A\vec{x}-\vec{b}\|^2 = (A^TA)^{-1}A^T\vec{b}$.

$A^T(A\vec{x}-\vec{b})=0$, A^TA invertible

08/30/22 Tuesday.

Vector. $\vec{x}\in\mathbb{R}^n$.

2-norm / Euclidean Norm $\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$. distance to the origin.

Vector Norm.

Vector Space X. $X\to\mathbb{R}$ is a norm if $\cdot\ \|\vec{x}\|\geq 0\ \forall \vec{x}\in X$, and $\|\vec{x}\|=0$ iff $\vec{x}=0$.

$\cdot$ triangular inequality. $\|\vec{x}+\vec{y}\|\leq\|\vec{x}\|+\|\vec{y}\|\ \forall \vec{x},\vec{y}\in X$.

$\cdot\ \|\lambda\vec{x}\|=|\lambda|\|\vec{x}\|\ \forall\lambda\in\mathbb{R}, \vec{x}\in X$.

Example: ℓ_p-norm. $\|\vec{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$, $1\leq p<\infty$.

$p=1$, $\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$

$p=\infty$, $\|\vec{x}\|_\infty = \max_{i=1,\dots,n}|x_i|$.

Cauchy-Schwartz Inequality: $\langle\vec{x},\vec{y}\rangle = \vec{x}^T\vec{y} = \|\vec{x}\|_2\|\vec{y}\|_2\cos\theta$.

$\langle\vec{x},\vec{y}\rangle \leq \|\vec{x}\|_2\|\vec{y}\|_2$. $|\vec{x}^T\vec{y}|\leq\|\vec{x}\|_2\|\vec{y}\|_2$.

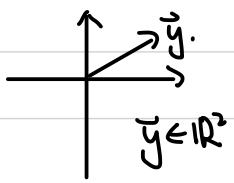
Hölder's Inequality: $p,q\geq 1$ s.t. $\frac{1}{p}+\frac{1}{q}=1$.

$|\vec{x}^T\vec{y}| \leq \sum_{i=1}^n |x_iy_i| \leq \|\vec{x}\|_p\|\vec{y}\|_q$.

↪ $\|\vec{x}\|_2\|\vec{y}\|_2\cos\theta$.

optimization: $\max \vec{x}^T\vec{y}$ $\|\vec{x}\|_p\leq 1$ given $\vec{y}\in\mathbb{R}^n$ fixed.

$p=2$.

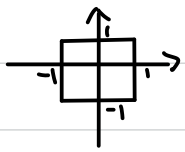


$\|\vec{x}\|_2\leq 1$. $\Rightarrow$ unit circle.

unit vector in the direction of $\vec{y}$.

$\vec{x} = \vec{y}/\|\vec{y}\|_2$.

$p=\infty$.



$\|\vec{x}\|_\infty = \max_{i=1,\cdots,n} |x_i|$

$\vec{x}^T\vec{y} = \vec{x}_1\vec{y}_1 + \vec{x}_2\vec{y}_2 + \cdots + \vec{x}_n\vec{y}_n$. independent: max each term. $\Rightarrow$ max the sum.

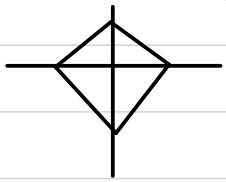
$x_i = \operatorname{sgn}(y_i)$ $\vec{x} = \operatorname{sgn}(\vec{y})$.

$\operatorname{sgn}(x) = \begin{cases} 1 & x>0 \\ 0 & x=0 \\ -1 & x<0 \end{cases}$.

$x_i y = \operatorname{sgn}(y_i)\cdot y_i = |y_i|$

so $\max_{\|\vec{x}\|_\infty=1} \vec{x}^T\vec{y} = \sum_{i=1}^n |y_i| = \|\vec{y}\|_1$.

$p=1$.



$|x_1|+|x_2| \le 1$. $\sum_{i=1}^n |x_i| \le 1$.

triangular inequality.

$\vec{x}^T\vec{y} \le |\vec{x}^T\vec{y}| = |\sum_{i=1}^n x_i y_i| \le \sum_{i=1}^n |x_i y_i| = \sum_{i=1}^n |x_i||y_i| \le \sum_{i=1}^n |x_i||y_{max}| \le |y_{max}| = \|\vec{y}\|_\infty$

$y_1\ y_2 \cdots y_{max} \cdots y_n$.

$\vec{x} = 0 \cdots 0 \cdots \operatorname{sgn}(y_{max}) \cdots 0$.

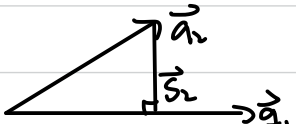
Gram-Schmidt | QR-decomposition.

$X = \{\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_n\}$. Basis.

find orthonormal basis for the vector space.

$\vec{q}_1 = \vec{a}_1 / \|\vec{a}_1\|_2$. $\|\vec{q}_1\|_2 = 1$.

$\vec{q}_2 = \vec{s}_2 / \|\vec{s}_2\|_2$.



must be normalized.

project $\vec{a}_2$ along $\vec{q}_1 = \vec{q}_1 \langle \vec{a}_2, \vec{q}_1 \rangle$.

$\vec{s}_2 = \vec{a}_2 - \vec{q}_1 \langle \vec{a}_2, \vec{q}_1 \rangle$.

$\vec{q}_3 = \vec{s}_3 / \|\vec{s}_3\|_2$. $\vec{s}_3 = \vec{a}_3 - \langle \vec{a}_3, \vec{q}_1 \rangle \vec{q}_1 - \langle \vec{a}_3, \vec{q}_2 \rangle \vec{q}_2$.

$A = QR$. Q: orthonormal. R: upper triangular.

$[\vec{a}_1 \cdots \vec{a}_n] = [\vec{q}_1 \cdots \vec{q}_n] R$.

09/01/22 Thursday.

Fundamental Theorem of Linear Algebra.

$A \in \mathbb{R}^{m \times n}$ $N(A) \oplus R(A^T) = \mathbb{R}^n$.

$R(A) \oplus N(A^T) = \mathbb{R}^m$.

orthogonal decomposition thm.

let X be a vector space, S be a subspace

$\forall \vec{x} \in X \quad \vec{x} = \vec{s} + \vec{r}. \quad \vec{s} \in S. \quad \vec{r} \in S^\perp \quad X = S \oplus S^\perp \quad S^\perp = \{\vec{r} \mid \langle \vec{r}, \vec{s} \rangle = 0, \forall \vec{s} \in S\}$.

prove $N(A) \oplus R(A^T) = \mathbb{R}^n$:

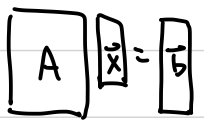
$N(A) \subseteq R(A^T)^\perp$: let $\vec{x} \in N(A)$, $A\vec{x} = 0$. want to show $\vec{x} \in R(A^T)^\perp \iff \forall \vec{w}$ s.t. $\vec{w} = A^T\vec{y}$. $\langle \vec{x}, \vec{w} \rangle = 0$.

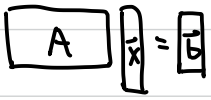
$\langle \vec{x}, \vec{w} \rangle = \langle \vec{x}, A^T\vec{y} \rangle = \vec{x}^T A^T \vec{y} = \vec{y}^T A\vec{x} = 0$. ($\vec{x}^T A^T = 0$)

$R(A^T)^\perp \subseteq N(A)$: let $\vec{x} \in R(A^T)^\perp$ $\forall \vec{w}$ s.t. $\vec{w} = A^T\vec{y}$. $\langle \vec{x}, \vec{w} \rangle = 0$. want to show $A\vec{x} = 0$

$\langle \vec{x}, \vec{w} \rangle = \langle \vec{x}, A^T\vec{y} \rangle = \vec{x}^T A^T \vec{y} = \vec{y}^T A\vec{x} = 0$. since it holds for all $\vec{y}$, $A\vec{x} = 0$

Minimum Norm Problem.

$A\vec{x} = \vec{b}$.  overdetermined. # solutions = 0.

 underdetermined. # solutions = ∞.

$\min \|\vec{x}\|_2^2$ s.t. $A\vec{x} = \vec{b}$

by FTLA, $\vec{x} = \vec{y} + \vec{z}$. s.t. $\vec{y} \in N(A)$. $\vec{z} \in R(A^T)$.

$A(\vec{y} + \vec{z}) = 0 + A\vec{z} = \vec{b}$.

$\|\vec{x}\|_2^2 = \|\vec{y}\|_2^2 + \|\vec{z}\|_2^2$. since $\vec{y} \perp \vec{z}$.

$\vec{z} = A^T\vec{w}$. $A\vec{z} = \vec{b} \Rightarrow AA^T\vec{w} = \vec{b} \Rightarrow \vec{w} = (AA^T)^{-1}\vec{b}$. Hence $\vec{z} = A^T(AA^T)^{-1}\vec{b}$.

Symmetric Matrices.

$A \in S^n$ $A = A^T$. $A_{ij} = A_{ji}$.

e.g. Covariance matrices. $A = B \cdot B^T$. $A^T = (B \cdot B^T)^T = B \cdot B^T$.

Always diagonalizable. $A = U \Lambda U^{-1}$ (Λ: diagonal; U: orthogonal) algebraic multiplicity = geometric multiplicity.

$\det(\lambda I - A)$ / # time λ is a root. dimension of $N(\lambda I - A)$.

e.g. $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \Rightarrow \begin{vmatrix} \lambda - 1 & -1 \\ 0 & \lambda - 1 \end{vmatrix} = (\lambda - 1)^2 = 0 \Rightarrow \lambda = 1$. w/ multiplicity = 2.

$\text{Null}(\lambda I - A) = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ dimension = 1. Hence Not diagonalizable.

Spectral Thm

$A \in S^n$. $\lambda_i \in \mathbb{R}$. with algebraic multiplicity μ_i. $E_i = N(\lambda_i I - A)$.

① $\lambda_i \neq \lambda_j$. $E_i \perp E_j$.

② $\dim(E_i) = \mu_i$.

proof: lemma: (λ, u) be an eigenpair for A. $\exists$ orthonormal U s.t. $U^TAU = \begin{bmatrix} \lambda & 0 \cdots 0 \\ \vdots & B \end{bmatrix}$ $B \in S^{n-1}$.

by construction, $U = [\vec{u} \mid U_1]$ $U^TAU = \begin{bmatrix} \vec{u}^T \\ U_1^T \end{bmatrix} A [\vec{u} \mid U_1] = \begin{bmatrix} \vec{u}^T \\ U_1^T \end{bmatrix} [A\vec{u} \mid AU_1] = \begin{bmatrix} \lambda & 0 \cdots 0 \\ 0 & \\ \vdots & B \\ 0 & \end{bmatrix}$ ($A\vec{u} = \lambda\vec{u}$.) $B = U_1^TAU_1$

let $B = V^T \Lambda V \in \mathbb{R}^{(n-1)\times(n-1)}$ (V: orthonormal. Λ: diagonal) $VU^TAUV^T = \begin{bmatrix} \lambda & \\ & VBV^T \end{bmatrix} = \begin{bmatrix} \lambda & 0 \\ 0 & \Lambda \end{bmatrix}$

For matrix, orthogonal = orthonormal. (→ orthonormal vectors.)

09/06/2022 Tuesday. ① eigenvalue decomposition = diagonalization ② SVD.

Spectral theorem: $A \in S^n$ $A = U\Lambda U^T$. (Λ: diagonal, U: orthonormal (unitary))

$A = [u_1 u_2 \cdots u_r \cdots u_n] \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_r \\ & & 0 \cdots 0 & \end{bmatrix} U^T$. ($u_1 \cdots u_r$: range, $u_r \cdots u_n$: null space)

Variational Characteristics of eigenvalue for symmetric matrix.

$R = \frac{\vec{x}^TA\vec{x}}{\vec{x}^T\vec{x}}$. ↓ Rayleigh coeff. $A \in S$ $\lambda_{min}(A) \leq \frac{\vec{x}^TA\vec{x}}{\|\vec{x}\|_2^2} \leq \lambda_{max}(A)$.

$\lambda_{max}(A) = \max_{\|\vec{x}\|_2=1} \vec{x}^TA\vec{x}$. $= \vec{x}$ be the eigenvector corr. to λ_{max}.

$\lambda_{min}(A) = \min_{\|\vec{x}\|_2=1} \vec{x}^TA\vec{x}$. $= \vec{x}$ be the eigenvector corr. to λ_{min}.

$A = U^T\Lambda U$. $U\vec{x} = \vec{y}$. $\vec{x}^TA\vec{x} = \vec{x}^TU^T\Lambda U\vec{x}$. $= \vec{y}^T\Lambda\vec{y}$. $= \sum_{i=1}^n \lambda_i y_i^2 \leq \lambda_{max} \sum_{i=1}^n y_i^2$. $= \lambda_{max}\|\vec{y}\|_2^2 = \lambda_{max}\|\vec{x}\|_2^2$.

"=" when $\vec{x}$ corr. to λ_{max}. $\|\vec{x}\|_2^2 = 1 \Rightarrow \|\vec{y}\|_2^2 = 1$

Principal Component Analysis.

$\vec{x}_1, \cdots, \vec{x}_n$ $\vec{x}_i \in \mathbb{R}^p$ Data Matrix $X = \begin{bmatrix} -\vec{x}_1^T- \\ -\vec{x}_2^T- \\ \vdots \\ -\vec{x}_n^T- \end{bmatrix}$ $n \times p$.

$C = \frac{1}{n}XX^T = \frac{1}{n}\begin{bmatrix} \|\vec{x}_1\|^2 & \langle \vec{x}_1, \vec{x}_2 \rangle & \cdots \\ & \|\vec{x}_2\|^2 & \\ & & \ddots \\ & & \|\vec{x}_n\|^2 \end{bmatrix}$ Symmetric

$D = \frac{1}{n}X^TX$. Symmetric.

want to find direction of PC1 $\vec{w}$. $\|\vec{w}\|_2 = 1$.

projection onto $\vec{w}$: $\langle \vec{x}_i, \vec{w} \rangle \vec{w}$. Error: $\|\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}\|^2 = e_i^2$. Avg Error: $\frac{1}{n}\sum_{i=1}^n e_i^2$.

$\|\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w}\|^2 = (\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w})^T (\vec{x}_i - \langle \vec{w}, \vec{x}_i \rangle \vec{w})$.

$= \|\vec{x}_i\|_2^2 + \langle \vec{w}, \vec{x}_i \rangle^2 \underbrace{\|\vec{w}\|_2^2}_{=1} - 2\langle \vec{w}, \vec{x}_i \rangle^2$.

$= \|\vec{x}_i\|_2^2 - \langle \vec{w}, \vec{x}_i \rangle^2$.

$\min_{\vec{w}} \frac{1}{n}\sum_{i=1}^n \|\vec{x}_i\|_2^2 - \frac{1}{n}\sum_{i=1}^n \langle \vec{w}, \vec{x}_i \rangle^2 \Rightarrow \max_{\vec{w}} \frac{1}{n}\sum_{i=1}^n \langle \vec{w}, \vec{x}_i \rangle^2 = \frac{1}{n}\|X\cdot\vec{w}\|^2 = \frac{1}{n}(X\vec{w})^T(X\vec{w}) = \frac{1}{n}\vec{w}^T X^T X \vec{w}$.

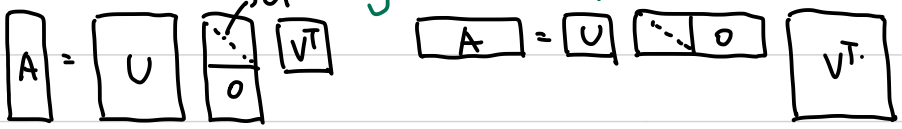
symmetric.

so $\vec{w}$ is the eigenvector corresponding to λ_{max} of X^TX.

Singular Value Decomposition.

U - eigenvectors of AA^T.
V - eigenvectors of A^TA.

$A \in \mathbb{R}^{m\times n}$. $A = U\Sigma V^T$. ($m\times n = m\times m \; m\times n \; n\times n$)



$A = \sigma_1\vec{u}_1\vec{v}_1^T + \sigma_2\vec{u}_2\vec{v}_2^T + \cdots + \sigma_r\vec{u}_r\vec{v}_r^T$ Rank $A = r$.

$A \in \mathbb{R}^{m\times n}$. consider A^TA $\lambda_1 \geq \lambda_2 \cdots \lambda_r \geq 0 \cdot \lambda_{r+1} = \cdots = \lambda_n = 0$.

$\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_n$ eigenvectors.

$A^TA\vec{v}_i = \lambda_i \cdot \vec{v}_i$. $V = \begin{bmatrix} | & | & & | \\ \vec{v}_1 & \vec{v}_2 & \cdots & \vec{v}_n \\ | & | & & | \end{bmatrix}$ define $\sigma_i = \sqrt{\lambda_i}$, $\vec{u}_i$ s.t. $A\vec{v}_i = \sigma_i \cdot \vec{u}_i$. $i \leq r$.

want to prove $\vec{u}_i$ and $\vec{u}_j$ are orthonormal.

$\vec{u}_i^T\vec{u}_j = \frac{(A\vec{v}_i)^T}{\sigma_i}\frac{(A\vec{v}_j)}{\sigma_j} = \frac{1}{\sigma_i\sigma_j}\vec{v}_i^T A^T A\vec{v}_j = \frac{1}{\sigma_i\sigma_j}\vec{v}_i^T\lambda_j\vec{v}_j = \frac{\lambda_j}{\sigma_i\sigma_j}\vec{v}_i^T\vec{v}_j = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$

$\|\vec{u}_i\|_2^2 = 1$.

$A^TA \in S^n$. orthogonal

Gram-Schmidt to find remaining $\vec{u}$s. $U: m\times m$.

$A \cdot V_r = \begin{bmatrix} \vec{u}_1 \cdots \vec{u}_r \end{bmatrix}\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$ $A\vec{v}_i = \sigma_i\vec{u}_i$.

$AV = U\Sigma$

SVD is not unique e.g. GS.

(not unique even if order the λ: repeated $\lambda \Rightarrow$ multiple bases).

09/08/22 Thursday.

Geometry of SVD

$A = U\Sigma V^T$ $A\vec{x} = U\Sigma V^T\vec{x}$.

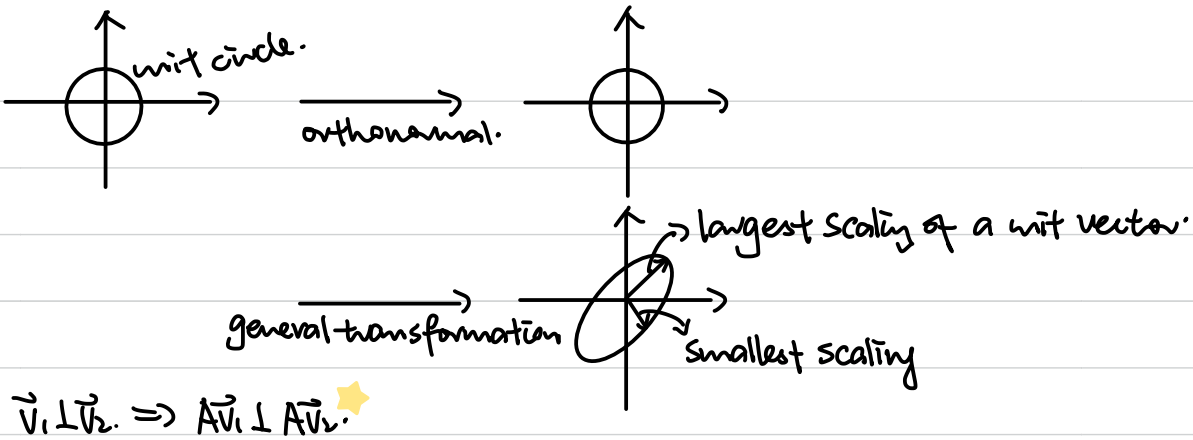
$V^T\vec{x}$: rotation / reflection.

Σ: scaling.

U: rotation / reflection.

If $\lambda_i = 0$ / Not full rank $\Rightarrow$ collapse dimension.

e.g. line.



$\vec{v}_1 \perp \vec{v}_2 \Rightarrow A\vec{v}_1 \perp A\vec{v}_2$.

orthogonal after transformation $\Rightarrow$ rotate / reflect / scale vectors.

$V^T\vec{x}$: change $\vec{x}$ into a basis s.t. A is just scaling.

$\Sigma V^T\vec{x}$: do the scaling.

$U\Sigma V^T\vec{x}$: change back to the original basis.

for arbitrary $\vec{x}$, think of $\vec{x} = a_1\vec{v}_1 + a_2\vec{v}_2$.

Low Rank Approximation. (of matrix.)

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

↳ block of data.

↳ operator.

Matrix-Norm.

- Frobenius Norm. $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} = \sqrt{\text{trace}(A^TA)}$.
  - ↳ invariant to orthonormal trans. $\|UA\|_F = \|AU\|_F = \|A\|_F$.
- Spectral Norm / ℓ_2-norm: $\|A\|_2 = \max_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2$. (max scaling of unit vector.)
  $= \max_{\|\vec{x}\|_2=1} \sqrt{\vec{x}^TA^TA\vec{x}} = \sqrt{\lambda_{max}(A^TA)} = \sigma_{max}(A)$.

$\|A^{-1}\|_2 = \frac{1}{\sigma_{min}(A)}$

Eckart-Young-Mirsky Thm

$A \in \mathbb{R}^{m\times n}$. $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \vec{u}_i \vec{v}_i^T$. $A_k = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T$.

① $\underset{B\in\mathbb{R}^{m\times n}}{\text{argmin}}\ \text{Rank}(B)=k.\ \|A-B\|_F = A_k$.

② $\underset{B\in\mathbb{R}^{m\times n}}{\text{argmin}}\ \text{Rank}(B)=k.\ \|A-B\|_2 = A_k$.

proof: $\|A-A_k\|_2 = \|\sum_{i=k+1}^{n} \sigma_i \vec{u}_i \vec{v}_i^T\|_2 = \sigma_{k+1}$

$\|A-B\|_2 \geq \|(A-B)\vec{w}\|_2 \quad \forall \|\vec{w}\|_2 = 1$. Let $\vec{w} \in \text{Null}(B)$. $\|(A-B)\vec{w}\|_2 = \|A\vec{w}\|_2$.

Consider $V_{k+1} = [\vec{v}_1\ \vec{v}_2\ \cdots\ \vec{v}_{k+1}]$ $\text{Rank}(V_{k+1}) = k+1$. $\dim(\text{Null}(B)) = n-k$.

$(n-k) + (k+1) = n+1 > n$. $\Rightarrow$ ∃ dimension overlap in $\text{Range}(V_{k+1})$ and $\text{Null}(B)$.

Suppose $\vec{w} \in \text{Range}(V_{k+1})$. $\vec{w} = V\cdot\vec{z} = [V_{k+1}\ V_{rest}] \begin{bmatrix} z_1 \\ \vdots \\ z_{k+1} \\ 0 \end{bmatrix} = z_1\vec{v}_1 + z_2\vec{v}_2 + \cdots + z_{k+1}\vec{v}_{k+1}$.

$\|\vec{w}\|^2 = 1 \Rightarrow \sum_{i=1}^{k+1} z_i^2 = 1$. *

$\|A\vec{w}\|_2^2 = \|U\Sigma V^TV\vec{z}\|_2^2 = \|U\Sigma\vec{z}\|_2^2 = \|\Sigma\vec{z}\|_2^2 = z_1^2\sigma_1^2 + \cdots + z_{k+1}^2\sigma_{k+1}^2 \geq \sigma_{k+1}^2 \left(\sum_{i=1}^{k+1} z_i^2\right) = \sigma_{k+1}^2$.

09/13/22 Tuesday.

$\|AU\|_F = \|UA\|_F = \|A\|_F$. $\|A\|_F = \sqrt{\text{trace}(A^TA)}$

Proof: $\|AU\|_F = \sqrt{\text{trace}((AU)^TAU)} = \sqrt{\text{trace}(U^TA^TAU)}$ ($\text{trace}(AB) = \text{trace}(BA)$) $= \sqrt{\text{trace}(UU^TA^TA)} = \sqrt{\text{trace}(A^TA)} = \|A\|_F$.

$\|A\|_F = \|U\Sigma V^T\|_F = \|\Sigma\|_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$

Proof: want $\|A-B\|_F \geq \|A-A_k\|_F$.

$\|A-A_k\|_F = \|\sum_{i=k+1}^{n} \sigma_i \vec{u}_i \vec{v}_i^T\|_F = \sqrt{\sum_{i=k+1}^{n} \sigma_i^2}$

Want $\sum_{i=1}^{n} \sigma_i^2(A-B) \geq \sum_{i=k+1}^{n} \sigma_i^2(A)$

Enough to prove: $\sigma_i(A-B) \geq \sigma_{k+i}(A)$. $\sum_{i=1}^{n-k} \sigma_{k+i}(A)$. $\Rightarrow \sum_{i=1}^{n-k} \sigma_i^2(A-B) \geq \sum_{i=1}^{n-k} \sigma_{k+i}^2(A)$.

$\sigma_{k+i}(A)$ = $k+i$th largest SV of A = $\|A - A_{k+i-1}\|_2$ (top) ← where $A_j = \sum_{i=1}^{j} \sigma_i \vec{u}_i \vec{v}_i^T$

Let $A-B=C$. $\sigma_i(A-B) = \sigma_i(C) = \|C - C_{i-1}\|_2$.

$\text{rank}(B) = k$. $\sigma_{k+1}(B) = 0$. $\|B - B_k\|_2 = 0$.

$\sigma_i(A-B) = \|C - C_{i-1}\|_2 + \|B - B_k\|_2 \geq \|C+B - C_{i-1} - B_k\|_2 = \|A - C_{i-1} - B_k\|_2$.

$\text{rank}(A+B) \leq \text{rank}(A) + \text{rank}(B)$.

$\text{rank}(B_k) = k$. $\text{rank}(C_{i-1}) \leq i-1$. Let $D = C_{i-1} + B_k$. $\text{rank}(D) \leq k+i-1$.

$\sigma_i(A-B) = \|A-D\|_2$ $\sigma_{k+i}(A) = \|A - A_{k+i-1}\|_2$. $\Rightarrow \sigma_i(A-B) \geq \sigma_{k+i}(A)$

Ingredients: $\text{rank}(B) = k$.

Frob. Norm $\Leftrightarrow$ Singular values.

Spetral norm = max singular value.

False Proof: $\min_{\text{rk}(B)=k} \|A-B\|_F = \min \|U\Sigma V^T - B\|_F = \min \|\Sigma - U^TBV\|_F$ ✗ $\min_{\substack{\text{rk}(Z)=k \\ Z=\text{diagonal}}} \|\Sigma - Z\|_F$.

$\text{rk}=1$ $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ $\text{rk}=2$ $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

09/15/2022 Thursday.

Vector Calculus.

Taylor's Theorem: $f(x): \mathbb{R} \to \mathbb{R}$. $x_0 \in \mathbb{R}$. fixed point.

$$f(x_0+\Delta x) = f(x_0) + \frac{df}{dx}\Big|_{x=x_0}(\Delta x) + \frac{1}{2!}\frac{df}{dx^2}\Big|_{x=x_0}(\Delta x)^2 + \cdots$$

Taylor's Thm for Vectors: $f(\vec{x}): \mathbb{R}^n \to \mathbb{R}$

$$f(\vec{x}_0+\Delta\vec{x}) = f(\vec{x}_0) + \underbrace{\nabla f\Big|_{\vec{x}=\vec{x}_0}^T \Delta\vec{x}}_{\text{first-order approximation}} + \frac{1}{2!}(\Delta\vec{x})^T \underbrace{\nabla^2 f\Big|_{\vec{x}=\vec{x}_0}}_{\text{Hessian}} \Delta\vec{x} + \cdots$$

Gradient: always a column vector. same dimension as $\vec{x}$.

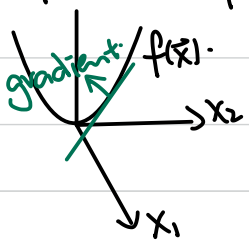
Hessian: often symmetric. $\nabla^2 f(\vec{x})_{ij} = \frac{\partial f}{\partial x_i \partial x_j}$

gradient → $\vec{p}^T \cdot \vec{x} = b$. e.g. $b=0$. → hyperplane.

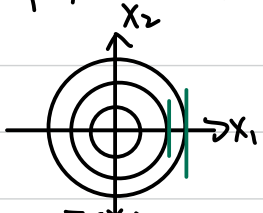
$\nabla f(\vec{x})$ captures the change cov. to all components of $\vec{x}$.

$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}^T$

Example: $f(\vec{x}) = \|\vec{x}\|_2^2$. $f: \mathbb{R}^2 \to \mathbb{R}$.



level sets: $f(x) = \text{constant}$.



$Df(\vec{x}) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 2\vec{x}$. $\to \frac{\partial f(x_1)}{\partial (x_1)} = 0$.

$D^2 f(\vec{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

$f(\vec{x}+\Delta\vec{x}) = (x_1^2 + x_2^2) + \begin{bmatrix} 2x_1 & 2x_2 \end{bmatrix}\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} \Delta x_1 & \Delta x_2 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix}$

$= x_1^2 + x_2^2 + 2x_1\Delta x_1 + 2x_2\Delta x_2 + \Delta x_1^2 + \Delta x_2^2$

$= (x_1+\Delta x_1)^2 + (x_2+\Delta x_2)^2$ Error $= 0$. quadratic approximation to quadratic function.

Example: $f(\vec{x}) = \vec{x}^T\vec{a} = \sum_{i=1}^n x_i a_i$. $Df(\vec{x}) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \vec{a}$.

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

Example: $f(\vec{x}) = \vec{x}^T A\vec{x} = \sum_i \sum_j x_i a_{ij} x_j$ terms including x_i: $\sum_{j\neq i} x_i a_{ij} x_j + \sum_{j\neq i} x_j a_{ji} x_i + x_i^2 a_{ii}$.

$\mathbb{R}^{n\times n} \neq S^n$. $\frac{\partial f}{\partial x_i} = \sum_{j\neq i} a_{ij}x_j + \sum_{j\neq i} x_j a_{ji} + 2x_i a_{ii} = \sum_j (a_{ij}+a_{ji})x_j$ $Df(\vec{x}) = (A+A^T)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$.

$D^2 f = (A+A^T)$

The Main Thm. optimal is obtained at boundary OR derivative $= 0$.

$f: \mathbb{R}^n \to \mathbb{R}$. differentiable everywhere. Ω: open set $\subseteq \mathbb{R}^2$. (↑ no.) minimize $_{\vec{x}\in\Omega} f(\vec{x})$. $\frac{df}{dx}(\vec{x}^*) = 0$.

Proof: $f(x^* + \Delta x) = f(x^*) + \frac{df}{dx}(x^*)\Delta x + \cdots$

$f(x^*)$ is the minimal $\Rightarrow f(x^*) \leq f(x^*+\Delta x)$.

open set $\Rightarrow \exists \varepsilon > 0$ s.t. $\forall x \in \Omega$, $|x^* - x| < \varepsilon$.

Let $0 < \Delta x < \varepsilon$. $f(x^*) \leq f(x^*) + \frac{df}{dx}(x^*)\Delta x + \cdots \Rightarrow 0 \leq \frac{df}{dx}(x^*)\Delta x + \cdots$

$0 \leq \frac{df}{dx}(x^*) + \text{Const.}\frac{\Delta x^2}{\Delta x} + \text{const.}\frac{\Delta x^3}{\Delta x}$ $\lim \Delta x \to 0$, $0 \leq \frac{df}{dx}(x^*)$.

WLOG wl $x^* - \Delta x$, $\frac{df}{dx}(x^*) \leq 0$.

09/20/2022 Tuesday.

Sensitivity / Perturbation Analysis.

data ↑ unknown → measurement.

$A\vec{x} = \vec{y}$. $A \in \mathbb{R}^{n\times n}$, invertible. $\vec{y} \to \vec{y} + \delta\vec{y} \Rightarrow \vec{x} \to \vec{x} + \delta\vec{x}$. How big is $\delta\vec{x}$?

$\frac{\|\delta x\|_2}{\|x\|_2}$ $A(\vec{x}+\delta\vec{x}) = \vec{y} + \delta\vec{y} \Rightarrow A\delta\vec{x} = \delta\vec{y}$. $\Rightarrow \|\delta\vec{x}\|_2 = \|A^{-1}\delta\vec{y}\|_2 \leq \|A^{-1}\|_2 \|\delta\vec{y}\|_2$. by def of spectral Norm.

$\|\vec{y}\|_2 = \|A\vec{x}\|_2 \leq \|A\|_2\|\vec{x}\|_2 \Rightarrow \|\vec{x}\|_2 \geq \|\vec{y}\|_2 / \|A\|_2$.

A is invertible $\Rightarrow A \neq 0$.

$$\frac{\|\delta x\|_2}{\|\vec{x}\|_2} \le \|A\|_2\|A^{-1}\|_2 \left[\frac{\|\delta y\|_2}{\|\vec{y}\|_2}\right] = \frac{\sigma_{max}}{\sigma_{min}} \frac{\|\delta y\|_2}{\|\vec{y}\|_2}$$

A is invertible $\Rightarrow \forall i, \sigma_i \neq 0$.

"Condition Number"

Least-Squares: $\vec{x} = (A^TA)^{-1}A^T\vec{b}$. $\underbrace{(A^TA)}_{\in \mathbb{R}^{n\times n}}\vec{x} = A^T\vec{b}$.

(A doesn't need to be square, A^TA should be invertible).

Condition Number ↑

if Add a constant (e.g. 200) to the dataset (A), the error will explode.

$(A^TA + \lambda I)\vec{x} = A^T\vec{b}$ works. λ-hyperparameter.

eigenvalue of $A+\lambda I$.

Small property of eigenvalues: $(A+\lambda I)\vec{v}_i = A\vec{v}_i + \lambda\vec{v}_i = \lambda_i\vec{v}_i + \lambda\vec{v}_i = (\lambda_i + \lambda)\vec{v}_i$.

$\Rightarrow$ increase λ_{min} and λ_{max} of A $\Rightarrow$ decrease condition number.

Ridge Regression.

$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2 + \lambda^2\|\vec{x}\|_2^2 \quad \lambda > 0$.

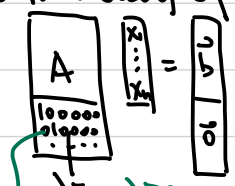
$f(\vec{x}) = (A\vec{x}-\vec{b})^T(A\vec{x}-\vec{b}) + \lambda^2\vec{x}^T\vec{x} = \vec{x}^TA^TA\vec{x} - \vec{x}^TA^T\vec{b} - \vec{b}^TA\vec{x} + \lambda^2\vec{x}^T\vec{x} + \vec{b}^T\vec{b}$.

PD $\Leftrightarrow$ invertible.

invertible since eigenvalues $\ge \lambda^2$.

$\nabla f(\vec{x}) = 2A^TA\vec{x} - 2(\vec{b}^TA)^T + 2\lambda^2\vec{x} = 0 \Rightarrow (A^TA + \lambda^2 I)\vec{x}^* = A^T\vec{b} \Rightarrow \vec{x}^* = (A^TA + \lambda^2 I)^{-1}A^T\vec{b}$.

Another way of interpretation:



$$\vec{x}^* = \left(\begin{bmatrix} A^T & \lambda I \end{bmatrix}\begin{bmatrix} A \\ \lambda I \end{bmatrix}\right)^{-1}\begin{bmatrix} A^T & \lambda I \end{bmatrix}\begin{bmatrix} \vec{b} \\ 0 \end{bmatrix} = (A^TA + \lambda^2 I)^{-1}(A^T\vec{b} + 0)$$

we are confident by λ degree that x_2 is (approx.) 0.

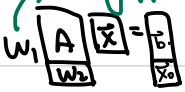
greater λ, greater confidence.

01/22/2022 Thursday.

Tikhonov Regularization.

$\min_{\vec{x}} \|W_1(A\vec{x}-\vec{b})\|_2^2 + \|W_2(\vec{x}-\vec{x}_0)\|_2^2$.

W_1 → weight matrix → importance of each row.



incoporate side info.

$y_i = g(\vec{x}_i) + z_i$ $(\vec{x}_i, y_i)$ data points $z_i \sim N(0, \sigma_i^2)$ $f(z_i) = \frac{e^{-z_i^2/2\sigma_i^2}}{\sqrt{2\pi}\sigma_i}$ i.i.d.

↑ X

$g(\vec{x}_i) = \vec{x}_i^T \vec{w}$. ↳ linear model. Not LS since we know $z_i \sim N(0, \sigma_i^2)$.

Maximum Likelihood Estimation

$\underset{\vec{w}_0}{\operatorname{argmax}}\, f(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n | \vec{w} = \vec{w}_0)$ $\vec{w}$ that makes observed data most likely.

↑ density (integrate to get probability).

$= \underset{\vec{w}_0}{\operatorname{argmax}} \prod_{i=1}^n f(Y_i = y_i | \vec{w} = \vec{w}_0)$ since independent.

$= f(\vec{x}_i^T \vec{w}_0 + z_i = y_i | \vec{w} = \vec{w}_0) = f(z_i = y_i - \vec{x}_i^T \vec{w}_0 | \vec{w} = \vec{w}_0) = \frac{e^{-\frac{(y_i - \vec{x}_i^T \vec{w}_0)^2}{2\sigma_i^2}}}{\sqrt{2\pi}\,\sigma_i}$

$= \underset{\vec{w}_0}{\operatorname{argmax}} \frac{1}{(\sqrt{2\pi})^n \prod_{i=1}^n \sigma_i} \prod_{i=1}^n e^{-\frac{(y_i - \vec{x}_i^T \vec{w}_0)^2}{2\sigma_i^2}}$

$= \underset{\vec{w}_0}{\operatorname{argmin}} \sum_{i=1}^n \frac{(y_i - \vec{x}_i^T \vec{w}_0)^2}{2\sigma_i^2}$ log is strictly increasing.

$= \underset{\vec{w}_0}{\operatorname{argmin}} \|S(X\vec{w}_0 - \vec{y})\|_2^2$. $S = \begin{bmatrix} \frac{1}{\sqrt{2\sigma_1^2}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{2\sigma_n^2}} \end{bmatrix}$ weighted LS.

Maximum a-posteriori (MAP) w/ prior on $\vec{w}$

$y_i = \vec{x}_i^T \vec{w} + z_i$ $\quad z_i \sim N(0, \sigma_i^2)$

$w_i \sim N(\mu_i, s_i^2)$ $\quad \vec{w} \sim N(\vec{\mu}, \Sigma_{\vec{w}})$ $\quad \Sigma_{\vec{w}} = \begin{bmatrix} s_1^2 & & 0 \\ & s_2^2 & \\ 0 & & \ddots\ s_n^2 \end{bmatrix}$. ↳ covariance = 0.

what is most likely $\vec{w}$, given data $y_1, y_2, \cdots, y_n$.

$\underset{\vec{w}}{\operatorname{argmax}}\, f(\vec{w} | \vec{Y} = \vec{y})$. $f(\vec{w} | Y = \vec{y}) = \frac{f(Y = \vec{y} | \vec{w}) f(\vec{w})}{f(\vec{y})}$ constant.

$= \underset{\vec{w}}{\operatorname{argmax}}\, f(\vec{Y} = \vec{y} | \vec{w}) \cdot f(\vec{w}) = \underset{\vec{w}}{\operatorname{argmax}} \left(\prod_{i=1}^n f(Y = y_i | \vec{w}) \right) \cdot f(\vec{w}) = \underset{\vec{w}}{\operatorname{argmax}} \prod_{i=1}^n \frac{\exp\{-\frac{(y_i - \vec{x}_i^T \vec{w})^2}{2\sigma_i^2}\}}{\sqrt{2\pi}\cdot\sigma_i} \cdot \frac{\exp\{-\frac{1}{2}(\vec{w} - \vec{\mu})^T \Sigma_w^{-1} (\vec{w} - \vec{\mu})\}}{(\sqrt{2\pi})^n (\prod s_i)}$

$= \underset{\vec{w}}{\operatorname{argmax}} \exp\{\sum_{i=1}^n \frac{-(y_i - \vec{x}_i^T \vec{w})^2}{2\sigma_i^2} + -(\vec{w} - \vec{\mu})^T \Sigma_w^{-1} (\vec{w} - \vec{\mu})\}$. multivariate Gaussian Distribution. $w \sim \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$.

$= \underset{\vec{w}}{\operatorname{argmin}} \sum_{i=1}^n \frac{(y_i - \vec{x}_i^T \vec{w})^2}{2\sigma_i^2} + (\vec{w} - \vec{\mu})^T \Sigma_w^{-1} (\vec{w} - \vec{\mu})$

$= \underset{\vec{w}}{\operatorname{argmin}} \|S(X\vec{w} - \vec{y})\|_2^2 + \|\sqrt{\Sigma_w^{-1}}(\vec{w} - \vec{\mu})\|_2^2$. $\Sigma_w^{-1} \in S \Rightarrow \Sigma_w^{-1} = \sqrt{\Sigma_w^{-1}} \cdot \sqrt{\Sigma_w^{-1}}$

$\vec{\mu} = 0 \Rightarrow$ Ridge. $\vec{\mu} \neq 0 \Rightarrow$ Tikhonov.

09/27/2022 Tuesday.

Convex Combinations: $\sum_{i=1}^n \lambda_i \vec{x}_i$ is a convex combination of $\vec{x}_i$ if $\lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$.

Convex Set: A set C is convex if the line segment joining any two points in the set is contained in the set.

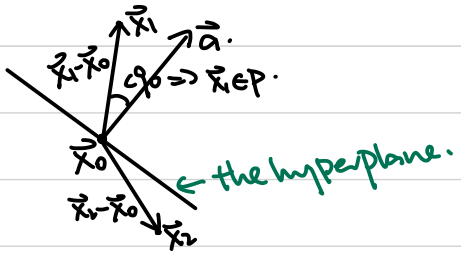
$\theta \vec{x}_1 + (1-\theta)\vec{x}_2 \in C$ $\forall \theta$ if $\vec{x}_1, \vec{x}_2 \in C$ and $\theta \in [0,1]$. traverse along the line segment

Example: $C = \{\vec{x} | \vec{a}^T \vec{x} = b\}$. A.K.A. $\vec{a}^T(\vec{x} - \vec{x}_0) = 0$. A.K.A. hyperplane.

Consider $\vec{x}_1, \vec{x}_2 \in C$, $\vec{x}_3 = \theta\vec{x}_1 + (1-\theta)\vec{x}_2$

$\vec{a}^T\vec{x}_3 = \theta\vec{a}^T\vec{x}_1 + (1-\theta)\vec{a}^T\vec{x}_2 = b \Rightarrow \vec{x}_3 \in C$. $\Rightarrow C$ is convex

$P: \{\vec{x} \mid \vec{a}^T(\vec{x}-\vec{x}_0) \geq 0\}$ $N: \{\vec{x} \mid \vec{a}^T(\vec{x}-\vec{x}_0) \leq 0\}$.



Example: $P = \{A \mid A \in S^n, A \text{ is PSD}\}$.

Consider $A_1, A_2 \in P$. $A_3 = \theta A_1 + (1-\theta)A_2$.

$\vec{x}^T A_3 \vec{x} = \theta(\vec{x}^T A_1 \vec{x}) + (1-\theta)(\vec{x}^T A_2 \vec{x}) \geq 0 \Rightarrow A_3 \in P \Rightarrow P$ is convex.

Separating Hyperplane Theorem: let C, D be convex sets, $C \cap D = \phi$, then $\exists$ hyperplane $\vec{a}^T\vec{x} = b$ s.t.

$\forall \vec{x} \in C, \vec{a}^T\vec{x} \geq b$. $\forall \vec{x} \in D, \vec{a}^T\vec{x} \leq b$. $\Rightarrow$ only.

Proof: $dist(C,D) = \inf\{\|\vec{c}-\vec{d}\|_2 \mid \vec{c} \in C, \vec{d} \in D\}$. (→ largest lower bound. → in case C, D are open sets.) C, D are closed, bounded sets. Assume $\vec{c}, \vec{d}$ exist, are distinct.

Consider $\vec{d}-\vec{c}$ as the normal vector passes through $\frac{\vec{d}+\vec{c}}{2}$

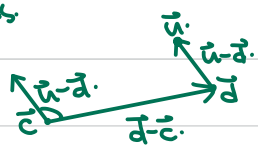
$f(\vec{x}) = (\vec{d}-\vec{c})^T(\vec{x} - \frac{\vec{d}+\vec{c}}{2})$ $f(\vec{x}) = 0$ is the hyperplane.

$f(\vec{c}) = -\frac{1}{2}\|\vec{d}-\vec{c}\|^2$ $f(\vec{x}) \leq 0 \ \forall \vec{x} \in C$. $f(\vec{d}) = \frac{1}{2}\|\vec{d}-\vec{c}\|^2$. $f(\vec{x}) \geq 0 \ \forall \vec{x} \in D$.

If possible, let $\vec{u} \in D$ s.t. $f(\vec{u}) < 0$ (prove by contradiction).

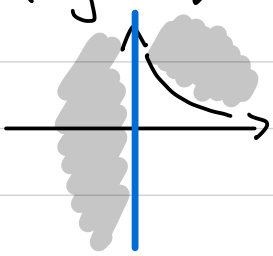
$f(\vec{u}) = (\vec{d}-\vec{c})^T(\vec{u} - \frac{1}{2}(\vec{d}+\vec{c}) + \vec{d} - \vec{d}) = (\vec{d}-\vec{c})^T((\vec{u}-\vec{d}) + \frac{1}{2}(\vec{d}-\vec{c})) = \langle \vec{d}-\vec{c}, \vec{u}-\vec{d} \rangle + \frac{1}{2}\|\vec{d}-\vec{c}\|_2^2 < 0$

So $\langle \vec{d}-\vec{c}, \vec{u}-\vec{d} \rangle < 0$ (directions)

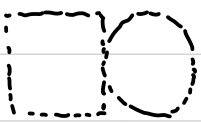


09/29/2022 Thursday.

tricky case: ① $C = \{(x,y) \mid x \geq 0\}$. $D = \{(x,y) \mid xy \geq 1\}$.



* unbounded sets.

②  * open sets.

Consider $\vec{p}=\vec{a}+t(\vec{u}-\vec{a})=t\vec{u}+(1-t)\vec{a}$.

D is a convex set $\vec{u}\in D$, $\vec{a}\in D$, if $t\in[0,1]$, $\vec{p}\in D$.

$\|\vec{c}-\vec{p}\|_2^2=\|\vec{c}-\vec{a}-t(\vec{u}-\vec{a})\|^2=((\vec{c}-\vec{a})-t(\vec{u}-\vec{a}))^T((\vec{c}-\vec{a})-t(\vec{u}-\vec{a}))=\|\vec{c}-\vec{a}\|^2+t^2\|\vec{u}-\vec{a}\|^2-2t\langle\vec{c}-\vec{a},\vec{u}-\vec{a}\rangle$

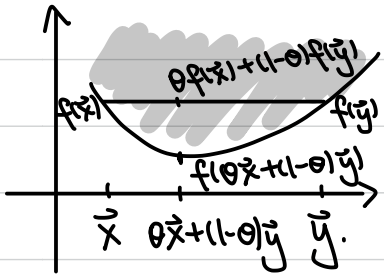
Want $\|\vec{c}-\vec{p}\|_2^2<\|\vec{c}-\vec{a}\|_2^2$. $\Rightarrow t^2\|\vec{u}-\vec{a}\|^2+2t\langle\vec{a}-\vec{c},\vec{u}-\vec{a}\rangle<0$

$\Rightarrow t<\frac{-2\langle\vec{a}-\vec{c},\vec{u}-\vec{a}\rangle}{\|\vec{u}-\vec{a}\|^2}$ t can be arbitrary small.

Convex function.

Def: A function $f:\mathbb{R}^n\to\mathbb{R}$ is convex (upwards) if dom(f) is a convex set and $\forall\vec{x},\vec{y}\in$dom(f) and $\theta\in[0,1]$, we have $f(\theta\vec{x}+(1-\theta)\vec{y})\le\theta f(\vec{x})+(1-\theta)f(\vec{y})$. [Strictly convex].

Def: A function $f:\mathbb{R}^n\to\mathbb{R}$ is concave (downwards) if dom(f) is a convex set and $\forall\vec{x},\vec{y}\in$dom(f) and $\theta\in[0,1]$, we have $f(\theta\vec{x}+(1-\theta)\vec{y})\ge\theta f(\vec{x})+(1-\theta)f(\vec{y})$. [Strictly concave].



Epigraph: $\text{Epi}f=\{(x,t)\mid x\in\text{dom}(f),f(x)\le t\}$.

Prop: f is a convex function $\Leftrightarrow$ Epi f is a convex set.

Jensen's inequality: if f is convex, $\vec{x}_1,\vec{x}_2,\cdots,\vec{x}_k\in$dom(f), and $\theta_1,\theta_2,\cdots,\theta_k\ge0$ with $\sum_{i=1}^k\theta_i=1$,

then $f(\theta_1\vec{x}_1+\theta_2\vec{x}_2+\cdots+\theta_k\vec{x}_k)\le\theta_1f(\vec{x}_1)+\theta_2f(\vec{x}_2)+\cdots+\theta_kf(\vec{x}_k)$.

first order condition (FOC): Suppose f is differentiable. f is convex iff dom(f) is convex and $\forall\vec{x},\vec{y}\in$dom(f), $f(\vec{y})\ge f(\vec{x})+\nabla f(\vec{x})^T(\vec{y}-\vec{x})$ tangent line below the fn.

Second order condition (SOC): Suppose f is twice differentiable. f is convex iff dom(f) is convex and $\nabla^2f(\vec{x})$ (Hessian) is PSD for all $\vec{x}\in$dom(f).

example of convex fn that is not differentiable. $f(\vec{x})=|\vec{x}|$.

implication: if $\nabla f(\vec{x}^*)=0$ and if f is convex, $f(\vec{y})\ge f(\vec{x}^*)+0(\vec{y}-\vec{x}^*)\Rightarrow\vec{x}^*$ is the global minimum.

proof of first order condition:

$\Rightarrow$: $f((1-t)\vec{x}+t\vec{y})\le(1-t)f(\vec{x})+tf(\vec{y})\Rightarrow tf(y)\ge tf(x)-f(x)+f((1-t)\vec{x}+t\vec{y})$.

$\Rightarrow f(y)\ge f(x)+\frac{1}{t}[f(x+t(\vec{y}-\vec{x}))-f(x)]$.

$f'(x)=\lim_{\delta x\to0}\frac{f(x+\delta x)-f(x)}{\delta x}\overset{\delta x=t(y-x)}{=}\lim_{t\to0}\frac{f(x+t(y-x))-f(x)}{t(y-x)}$

take $\lim t\to0$, $f(y)\ge f(x)+f'(x)(y-x)$

$\Leftarrow$: let $z=\theta x+(1-\theta)y$.

$f(x)\geq f(z)+f'(z)(x-z)$. $\quad f(y)\geq f(z)+f'(z)(y-z)$

$\theta f(x)+(1-\theta)f(y)\geq \theta f(z)+\theta f'(z)(x-z)+(1-\theta)f(z)+(1-\theta)f'(z)(y-z)$.

$= f(z)+f'(z)(z\theta-x\theta+y-z-y\theta+z\theta) = f(z)+f'(z)[\theta x+(1-\theta)y-z] = f(z)$.

10/04/2022 Tuesday. Strong convexity => strict convexity => convexity.

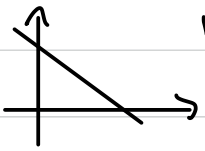
Strict Convexity

domain(f) is convex, $\theta\in(0,1)$ $\forall x,y\in$ domain(f).

Zeroth order condition: $f(\theta\vec{x}+(1-\theta)\vec{y})<\theta f(\vec{x})+(1-\theta)f(\vec{y})$.

First order condition: $f(\vec{y})>f(\vec{x})+\nabla f(\vec{x})^T(y-x)$ Taylor first order approximation.

Second order condition: $\nabla^2 f(\vec{x})\succ 0$ PD. $\quad A\succeq B\Rightarrow A-B\succeq 0$.

Example:  both concave and convex. Not strictly.

Strong Convexity.

f is differentiable, domain f is convex. $\mu>0$

f is μ-strongly convex if $\forall \vec{x},\vec{y}\in$ domain(f), $\vec{x}\neq\vec{y}$ $f(\vec{y})\geq f(\vec{x})+\nabla f(\vec{x})^T(\vec{y}-\vec{x})+\frac{\mu}{2}\|\vec{y}-\vec{x}\|^2$.

quadratic fn under $=\frac{1}{2}(\vec{y}-\vec{x})^T\begin{bmatrix}\mu & & \\ & \ddots & \\ & & \mu\end{bmatrix}(\vec{y}-\vec{x})$

Taylor second order approximation. $f(\vec{y})\approx f(\vec{x})+\nabla f(\vec{x})^T(\vec{y}-\vec{x})+\frac{1}{2}(\vec{y}-\vec{x})^T\nabla^2 f(\vec{x})(\vec{y}-\vec{x})$.

$\mu I\leq \nabla^2 f(\vec{x})$.

the gradient is not changing too slow.

unconstrained optimization problem

Gradient Descent.

$f(\vec{x}+\Delta\vec{x})=f(\vec{x})+\nabla f(\vec{x})^T\Delta\vec{x}$.

stepsize, direction.

$s>0$ $f(\vec{x}+s\vec{v})=f(\vec{x})+s\langle\nabla f(\vec{x}),\vec{v}\rangle$. maximize magnitude => $\vec{v}$ aligned to gradient.

minimize f => $\vec{v}=-\nabla f(\vec{x})$.

* if close to $\vec{x}^*$, $\nabla f(\vec{x})$ is small, step $\langle\nabla f(\vec{x}),\vec{v}\rangle$ is small.

$\vec{x}_{k+1}=\vec{x}-\eta\nabla f(\vec{x}_k)$. $\vec{x}_0$: initial point. η: stepsize. if $\vec{x}_k=\vec{x}^*$, $\vec{x}_{k+1}=\vec{x}^*-0$.

Least squares: $f(x)=\|A\vec{x}-\vec{b}\|_2^2$. $\nabla f(\vec{x})=2A^T(A\vec{x}-\vec{b})$.

$\vec{x}^*=(A^TA)^{-1}A^T\vec{b}$. $\quad \vec{x}_{k+1}=[I-2\eta A^TA]\vec{x}_k+2\eta A^T\vec{b}$.

eigenvalues.

Why GD? - faster.

$$\vec{x}_{k+1}-\vec{x}^* = \vec{x}_{k+1}-(A^TA)^{-1}A^T\vec{b} = (I-2\eta A^TA)\vec{x}_k + 2\eta(A^TA)(A^TA)^{-1}A^T\vec{b} - (A^TA)^{-1}A^T\vec{b}$$

$$= \underbrace{(I-2\eta A^TA)}_{\in S^n}(\vec{x}_k-(A^TA)^{-1}A^T\vec{b}).$$

eigenvalues $|<1| \Rightarrow$ converge

$$= (I-2\eta A^TA)^k(\vec{x}_0-(A^TA)^{-1}A^T\vec{b}).$$

10/06/2022 Thursday.

L-smooth: $\forall \vec{x},\vec{y}\in$ domain, $f(\vec{y}) \le f(\vec{x}) + \nabla f(\vec{x})^T(\vec{y}-\vec{x}) + \frac{L}{2}\|\vec{y}-\vec{x}\|_2^2$.

↑ the gradient is not changing too fast.

Thm: $f:\mathbb{R}^n \to \mathbb{R}$ μ-strong convexity AND L-smooth. $\min_{\vec{x}\in\mathbb{R}^n} f(\vec{x})$ $\vec{x}_{t+1} = \vec{x}_t - \eta\nabla f(\vec{x}_t)$.

then $\|\vec{x}_{t+1}-\vec{x}^*\|_2^2 \le C^{t+1}\|\vec{x}_0-\vec{x}^*\|_2^2$ (choose η s.t. $|C|<1$) $\Rightarrow \lim_{t\to\infty}\vec{x}_t = \vec{x}^*$

Lemma: if f is L-smooth, $\|\nabla f(\vec{x})\|_2^2 \le 2L(f(\vec{x})-f(\vec{x}^*))$ gradient close to optimum is small.

Proof: $f(\vec{x}^*) \le f(\vec{x})$ $f(\vec{x}^*) \le f(\vec{x}-\frac{\nabla f(\vec{x})}{L})$

$$f(\vec{x}-\tfrac{\nabla f(\vec{x})}{L}) \le f(\vec{x}) + \nabla f(\vec{x})^T(-\tfrac{\nabla f(\vec{x})}{L}) + \tfrac{L}{2}\|-\tfrac{\nabla f(\vec{x})}{L}\|_2^2 = f(\vec{x}) - \tfrac{1}{2L}\|\nabla f(\vec{x})\|_2^2$$

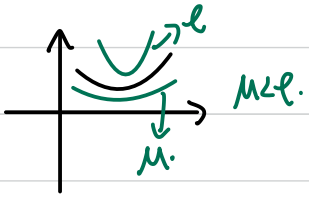
μ-strong convexity: let $\vec{y}=\vec{x}^*$ $\nabla f(\vec{x})^T(\vec{x}-\vec{x}^*) \ge f(\vec{x}) - f(\vec{x}^*) + \frac{\mu}{2}\|\vec{x}^*-\vec{x}\|_2^2$.

Proof: $\|\vec{x}_{t+1}-\vec{x}^*\|_2^2 = \|(\vec{x}_t-\vec{x}^*)-\eta\nabla f(\vec{x}_t)\|_2^2 = \|\vec{x}_t-\vec{x}^*\|_2^2 + \eta^2\|\nabla f(\vec{x}_t)\|_2^2 - 2\eta\nabla f(\vec{x}_t)^T(\vec{x}_t-\vec{x}^*)$.

$$\le \|\vec{x}_t-\vec{x}^*\|_2^2 + \eta^2\cdot 2L\cdot(f(\vec{x}_t)-f(\vec{x}^*)) - 2\eta\left(f(\vec{x}_t)-f(\vec{x}^*)+\frac{\mu}{2}\|\vec{x}^*-\vec{x}_t\|_2^2\right).$$

$$= (1-\eta\cdot\mu)\|\vec{x}_t-\vec{x}^*\|_2^2 + (2\eta^2L-2\eta)[f(\vec{x}_t)-f(\vec{x}^*)].$$

choose $\eta=\frac{1}{L} \Rightarrow 2\eta^2L-2\eta=0$, then $\|\vec{x}_{t+1}-\vec{x}^*\|_2^2 \le \underbrace{(1-\frac{\mu}{L})}_{<1}^{t+1}\|\vec{x}_0-\vec{x}^*\|_2^2$.



Stochastic Gradient Descent.

$f(\vec{x}) = \sum_{i=1}^m \frac{1}{m} f_i(\vec{x})$

$x_{k+1} = x_k - \eta_k \nabla f_i(x_k)$ $E[\nabla f_i(x_k)] = \nabla f(x_k)$.

pointwise maximize convex fns → convex. concave → convex|concave.

pointwise minimize convex fns → convex|concave. concave → concave.

10/13/2022 Thursday.

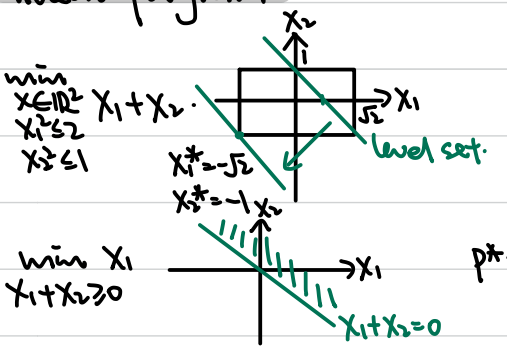
Active Constraint. optimum on the constraint / boundary.

Inactive Constraint.

Convex optimization problem: minimize convex / maximize concave.

infeasible optimization problem → feasible set = ϕ. $p^* = \infty$ | $-\infty$

linear Program.



↳ c^T (orthogonal to level set).

$\begin{bmatrix} 1 \\ 1 \end{bmatrix} x$ minimize ⇒ move in negative $\nabla f(x) = c$ direction.

$\min x_1$
$x_1 + x_2 \ge 0$

$p^* = -\infty$.

$x_1 + x_2 = 0$

Thm: $\min_{\vec{x} \in X} \vec{c}^T \vec{x}$ X: convex set, closed. If $\vec{x}^*$ is an optimum, then $\vec{x}^* \in$ Boundary(X).

Proof: Assume $\vec{x}^*$ is in the interior of the set X.

↳ in direction of $\vec{c}$

$\forall \vec{z}$ $\|\vec{x} - \vec{z}\|_2 \le r \Rightarrow \vec{z} \in X$. (let $\vec{z}_0 = -\lambda \cdot \vec{c}$ $\lambda^2 \frac{r}{\|\vec{c}\|_2} > 0$

$f_0(\vec{x}^* + \vec{z}) = c^T(\vec{x}^* + \vec{z}) = c^T \vec{x}^* - \lambda \vec{c}^T \vec{c} < f_0(\vec{x}^*)$

Problem Transformation.

(maximize concave fn) = −(minimize convex fn). (= −(the concave fn.))

① Addition of "slack" variable. $\min_{\vec{x} \in X} f_0(\vec{x}) \iff \min_{\substack{f(\vec{x}) \le t \\ \vec{x} \in X}} t$

e.g. $\min_x \|Ax - b\|_2^2 + \|x\|_1$ LASSO $\iff \min_{|x_i| \le t_i} \|A\vec{x} - \vec{b}\|_2^2 + \sum_{i=1}^n t_i$.

Not differentiable everywhere.

② $\phi(x)$ conti. + strictly increasing.

$\min_{f_i(\vec{x}) = 0} f_0(\vec{x}) \iff \min_{f_i(\vec{x}) \le 0} \phi(f_0(\vec{x}))$ should be the same.

Not convex.

e.g. $\min_{x_i^{\beta_1} x_2^{\beta_2} \le b, x_i \ge 0} \lambda x_1^{\alpha_1} x_2^{\alpha_2} \iff \min_{\log(x_i) \ge 0} \beta_1 \log x_1 + \beta_2 \log x_2 \le \log b$. Convex. $\log\lambda + \lambda_1 \log x_1 + \lambda_2 \log x_2$

obj. strictly decreasing over domain

f_i nondecreasing over domain

⇓

equality constraint ⇒ inequality ($\le$)

10/18/2022 Tuesday.

Duality.

$p^* = \min_{\text{over } \vec{x}} f_0(\vec{x})$ ↳ inequality constraint.

$f_i(\vec{x}) \le 0 \ \forall i \in [1, m]$

$h_i(\vec{x}) = 0 \ \forall i \in [1, p]$ → equality constraint.

PRIMAL

Lagrangian: $L(\vec{x}, \vec{\lambda}, \vec{\nu}) = f_0(\vec{x}) + \sum_{i=1}^m \lambda_i f_i(\vec{x}) + \sum_{i=1}^p \nu_i h_i(\vec{x})$ $\lambda_i \ge 0$

↓ min

$\min_{\vec{x}} L(\vec{x}, \vec{\lambda}, \vec{\nu})$ s.t. $\vec{\lambda} \ge 0$; $g(\vec{\lambda}, \vec{\nu})$

L is affine fn. → both convex and concave. ✓

g is pointwise min of concave fn ⇒ concave.

prop: $g(\vec{\lambda},\vec{\mu})$ is the lower bound of p^*. $\forall \vec{\lambda}\geq 0, \vec{\mu}$, $g(\vec{\lambda},\vec{\mu}) \leq p^*$.

proof: consider $\vec{x}$ some feasible set for the primal. $f_i(\vec{x}) \leq 0$ $g_i(\vec{x}) = 0$.

$$L(\vec{x},\vec{\lambda},\vec{\mu}) = f_0(\vec{x}) + \underbrace{\sum \lambda_i f_i(\vec{x})}_{\leq 0} + \underbrace{\sum \mu_i h_i(\vec{x})}_{=0} \leq f_0(\vec{x})$$

$$g(\vec{\lambda},\vec{\mu}) = \min_{\vec{x}} L(\vec{x},\vec{\lambda},\vec{\mu}) \Rightarrow g(\vec{\lambda},\vec{\mu}) \leq p^*.$$

$M(\vec{x}) = f_0(\vec{x}) + \sum \mathbb{1}_{\leq 0}\{f_i(\vec{x})\} + \sum \mathbb{1}_0\{h_i(\vec{x})\}$. $\mathbb{1}_{\leq 0}\{f_i(\vec{x})\} = \begin{cases} 0 & \text{if } f_i(\vec{x}) \leq 0 \\ \infty & \text{if } f_i(\vec{x}) > 0 \end{cases}$ $\mathbb{1}_0\{h_i(\vec{x})\} = \begin{cases} 0 & \text{if } h_i(\vec{x}) = 0 \\ \infty & \text{if } h_i(\vec{x}) \neq 0 \end{cases}$

"Hard threshold" $\sum \mathbb{1}_{\leq 0}\{f_i(\vec{x})\}$. $\longleftrightarrow$ $\sum \lambda_i f_i(\vec{x})$. encourage $f_i(\vec{x}) \leq 0$.

$f_i(\vec{x}) \leq 0$	0	$\geq$	≤ 0
$f_i(\vec{x}) > 0$	∞	$\geq$	> 0

$\max_{\vec{\lambda}\geq 0, \vec{\nu}} g(\vec{\lambda},\vec{\mu}) = d^*$. (largest lower bound.) DUAL problem. to primal.

* Maximization of a concave fn, linear constraint. $\Rightarrow$ Convex Program.
* #variable = #constraints of primal.
* Always Convex program regardless of primal ($f_0(\vec{x})$).

WEAK DUALITY: $d^* \leq p^*$. Duality Gap $p^* - d^*$ STRONG DUALITY. $d^* = p^*$
* Always hold regardless of convexity of $f_0(\vec{x}), f_i(\vec{x})$.

Example: $\min_{A\vec{x}=\vec{b}} \vec{x}^T\vec{x}$. $A \in \mathbb{R}^{m\times n}$ $m \geq n$. ($A\vec{x}-\vec{b}=0$)

$L(\vec{x},\vec{\mu}) = \vec{x}^T\vec{x} + \vec{\mu}^T(A\vec{x}-\vec{b})$ $\nabla_{\vec{x}} L(\vec{x},\vec{\mu}) = 2\vec{x} + A^T\vec{\mu}$. $\Rightarrow \vec{x}^* = -\frac{1}{2}A^T\vec{\mu}$.

$g(\vec{\mu}) = \min_{\vec{x}} L(\vec{x},\vec{\mu}) = L(\vec{x}^*,\vec{\mu}) = \frac{1}{4}\vec{\mu}^T AA^T\vec{\mu} + \vec{\mu}^T(-\frac{1}{2}AA^T\vec{\mu} - \vec{b}) = -\frac{1}{4}\vec{\mu}^T AA^T\vec{\mu} - \vec{\mu}^T\vec{b}$.

$\nabla_{\vec{\mu}} g(\vec{\mu}) = -\frac{1}{2}AA^T\vec{\mu} - \vec{b} = 0 \Rightarrow \vec{\mu}^* = -2(AA^T)^{-1}\vec{b} \Rightarrow \vec{x}^* = -\frac{1}{2}A^T(-2)(AA^T)^{-1}\vec{b} = A^T(AA^T)^{-1}\vec{b}$. Strong duality. by slater since no f_i.

Partitioning Problem $\min_{x_i^2=1, i\in\{1,\dots,n\}} \vec{x}^T W\vec{x}$ $W \in S^n$ Not Convex. ← discrete domains.

$\vec{x}^T W\vec{x} = \sum_i\sum_j x_i W_{ij} x_j$. W_{ij} → unhappiness of two people in the same group.

$L(\vec{x},\vec{\mu}) = \vec{x}^T W\vec{x} + \sum \mu_i(x_i^2 - 1) = \vec{x}^T W\vec{x} + \vec{x}^T \mathrm{diag}(\vec{\mu})\vec{x} - \sum \mu_i = \vec{x}^T(W + \mathrm{diag}(\vec{\mu}))\vec{x} - \sum \mu_i$

$g(\vec{\mu}) = \min_{\vec{x}} L(\vec{x},\vec{\mu}) = \begin{cases} -\sum \mu_i & \text{if } W + \mathrm{diag}(\vec{\mu}) \text{ PSD} \\ -\infty & \text{otherwise.} \end{cases}$ have negative eigenvalues.

DUAL: $\max_{W+\mathrm{diag}(\vec{\mu}) \text{ is PSD}} -\sum \mu_i$ $\vec{\mu}^* = \lambda_{\min}(W)$ $p^* \geq n \cdot \lambda_{\min}(W)$.

10/20/2022 Thursday.

MiniMax Inequality: Sets X,Y F is any function. $\min_{x\in X}\max_{y\in Y} F(x,y) \geq \max_{y\in Y}\min_{x\in X} F(x,y)$ (first.)

Proof: Fix $x_0\in X, y_0\in Y$. Define $h(y_0) := \min_{x\in X} F(x,y_0)$ $g(x_0) := \max_{y\in Y} F(x_0,y)$.

$\forall x_0,y_0\ h(y_0)\leq F(x_0,y_0)\leq g(x_0) \Rightarrow \max_{y_0\in Y} h(y_0) \leq \min_{x_0\in X} g(x_0)$

$d^* = \max_{\vec{\lambda}\geq 0,\vec{\mu}} \min_{\vec{x}} L(\vec{x},\vec{\lambda},\vec{\mu})$

$\max_{\vec{\lambda}\geq 0,\vec{\mu}} (f_0(\vec{x}) + \sum_{i=1}^{m}\lambda_i f_i(\vec{x}) + \sum_{i=1}^{r}\nu_i h_i(\vec{x})) = \begin{cases} f_0(\vec{x}) & \text{if } x \text{ is feasible.} \\ \infty & \text{otherwise.} \end{cases} \Rightarrow p^* = \min_{\vec{x}}\max_{\vec{\lambda}\geq 0,\vec{\mu}} L(\vec{x},\vec{\lambda},\vec{\mu})$

Hence $p^*\geq d^*$

weak duality holds with Non-Convex dimension thing.

Slater's Condition Strong duality holds if $\exists \vec{x}_0\in$ relative interior(D) s.t. $\forall i\ f_i(\vec{x}_0)<0$

CONVEX (constraints, obj fns, domains).

refined slater. CONVEX. $f_1,f_2,\cdots,f_k$ are affine.

$\exists \vec{x}_0\in$ relative interior(D) s.t. $f_i(\vec{x}_0)\leq 0$ for $i\in[1,k]$ AND $f_i(\vec{x}_0)<0$ for $i\in[k+1,n]$

↳ Strong duality always hold if feasible.

linear Program & duality.

$\min_{A\vec{x}\leq\vec{b}} \vec{c}^T\vec{x}$ $L(\vec{x},\vec{\lambda}) = \vec{c}^T\vec{x} + \vec{\lambda}^T(A\vec{x}-\vec{b}) = (A^T\vec{\lambda}+\vec{c})^T\vec{x} - \vec{b}^T\vec{\lambda}$.

$g(\vec{\lambda}) = \min_{\vec{x}} L(\vec{x},\vec{\lambda}) = \begin{cases} -\infty & \text{if } A^T\vec{\lambda}+\vec{c}\neq 0 \\ -\vec{b}^T\vec{\lambda} & \text{otherwise} \end{cases} \Rightarrow d^* = \max_{\substack{\vec{\lambda}\geq 0 \\ A^T\vec{\lambda}+\vec{c}=0}} -\vec{b}^T\vec{\lambda}$. Also Linear program.

d^{**} = primal.

200kg merlot. 300kg shiraz. Blend 1: 4kgM + 1kgS \$20. Blend 2: 2kgM + 3kgS \$15.

$\max_{q_1,q_2} 20q_1+15q_2$ s.t. $4q_1+2q_1\leq 200,\ q_1+3q_2\leq 300,\ q_1,q_2\geq 0$

+sell grapes directly: $\max_{q_1,q_2\geq 0} 20q_1+15q_2+\lambda_1(200-4q_1-2q_2)+\lambda_2(300-q_1-3q_2)$ Lagrangian

$= \max_{q_1,q_2\geq 0} \underbrace{(20-4\lambda_1-\lambda_2)}_{<0\Rightarrow\text{No Blend 1}}q_1 + \underbrace{(15-2\lambda_1-3\lambda_2)}_{<0\Rightarrow\text{No Blend 2.}}q_2 + 200\lambda_1+300\lambda_2$

At indifference point: λ_1,λ_2 → shadow prices.

$\min_{\lambda_1,\lambda_2\geq 0} 200\lambda_1+300\lambda_2$ s.t. $20-4\lambda_1-\lambda_2=0$ AND $15-2\lambda_1-3\lambda_2=0$ DUAL.

10/25/2022 Tuesday.

Certificate Property.

(λ_1,μ_1) – feasible dual point x_1 – feasible primal

$p^*\geq g(\lambda_1,\mu_1) \Rightarrow \underbrace{f_0(x_1)-p^*}_{\leq\varepsilon} \leq \underbrace{f_0(x_1)-g(\lambda_1,\mu_1)}_{\leq\varepsilon}$ $\Longleftarrow$

$p^*\in[g(\lambda_1,\mu_1), f_0(x_1)]$ w/ strong duality. $d^*\in[g(\lambda_1,\mu_1), f_0(x_1)]$.

Complementary Slackness. * regardless of the convexity.

primal optimal $(\vec{x}^*)$ Dual Optimal $(\vec{\lambda}^*, \vec{\mu}^*)$

$g(\vec{\lambda}^*, \vec{\mu}^*) = \min_{\vec{x}} \left(f_0(\vec{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\vec{x}) + \sum_{i=1}^{p} \mu_i^* h_i(\vec{x}) \right) \overset{①}{\leq} f_0(\vec{x}^*) + \underbrace{\sum_{i=1}^{m} \lambda_i^* f_i(\vec{x}^*)}_{\leq 0} + \underbrace{\sum_{i=1}^{p} \mu_i^* h_i(\vec{x}^*)}_{=0} \overset{②}{\leq} f_0(\vec{x}^*)$.

Assume strong duality holds: $p^* = d^*$ $p^* = f_0(\vec{x}^*)$ $d^* = g(\vec{\lambda}^*, \vec{\mu}^*)$

① $\Rightarrow \min_{\vec{x}} \left(f_0(\vec{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\vec{x}) + \sum_{i=1}^{p} \mu_i^* h_i(\vec{x}) \right) = f_0(\vec{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\vec{x}^*) + \sum_{i=1}^{p} \mu_i^* h_i(\vec{x}^*) \Rightarrow x^*$ also minimize L.

② $\Rightarrow \sum_{i=1}^{m} \lambda_i^* f_i(\vec{x}^*) \Rightarrow \forall i \ \lambda_i^* f_i(\vec{x}^*) = 0 \Rightarrow$ $\because \lambda_i^* > 0, f_i(\vec{x}^*) = 0$ / $\because f_i(\vec{x}^*) > 0, \lambda_i^* = 0$ } Complementary slackness.

(not necessarily the only minimizer.)

KKT Condition.

Strong duality holds; regardless of convexity; differentiable f_i's, h_i's.

Necessary conditions of $\vec{x}^*, \vec{\lambda}^*, \vec{\mu}^*$ optimality: (if A→B) Contrapositive: if Not B, Not A.

then:

① $\forall i \in [1,m], f_i(\vec{x}^*) \leq 0$.

② $\forall i \in [1,p], h_i(\vec{x}^*) = 0$.

③ $\forall i \in [1,m], \lambda_i^* \geq 0$.

④ $\forall i \in [1,m], \lambda_i^* f_i(\vec{x}^*) = 0$.

⑤ $\nabla f_0(x^*) + \sum \lambda_i^* \nabla f_i(\vec{x}^*) + \sum \mu_i^* \nabla h_i(\vec{x}^*) = 0$ by Main Thm. * respect to x_i for $i \in [1,n]$.

Convex problems ($\forall i \in [1,m]$, f_i convex; h_i affine.); differentiable f_i's, h_i's. (f_0 convex; domain convex.)

Sufficient conditions

$\vec{x}, \vec{\lambda}, \vec{\mu}$ satisfying KKT; convex; h_i affine; slater's condition. $\Leftrightarrow \vec{x}, \vec{\lambda}, \vec{\mu}$ are optimal.

10/27/2022 Thursday.

proof sufficient condition:

consider $L(x, \vec{\lambda}, \vec{\mu}) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x)$ → convex fn in x. by ⑤, $\vec{x}$ is a minimizer

$g(\vec{\lambda}, \vec{\mu}) = \min_{x} L(x, \vec{\lambda}, \vec{\mu}) = f_0(\vec{x}) + \sum \lambda_i f_i(x) + \sum \mu_i h_i(x) \overset{④②}{=} f_0(\vec{x})$

Since $p^* \geq g(\vec{\lambda}, \vec{\mu})$, $p^* = f_0(\vec{x})$ (p^* could $> f_0(\vec{x})$) $\Rightarrow d^* = g(\vec{\lambda}, \vec{\mu}) = f_0(\vec{x})$ Certificate Property.

Linear Program. → Convex. ← Strong duality holds as long as feasible.

$\min_{\vec{x}} \vec{c}^T\vec{x}$ s.t. $A\vec{x}\le\vec{b}$.

- $\vec{a_i}^T\vec{x}=b \Rightarrow \vec{a_i}^T\vec{x}\ge b$ AND $\vec{a_i}^T\vec{x}\le b$
- $\vec{a_i}^T\vec{x}\ge b \Rightarrow -\vec{a_i}^T\vec{x}\le -b$.

Standard format: $\min_{\vec{x}} \vec{c}^T\vec{x}$. s.t. $A\vec{x}=\vec{b}$, $\vec{x}\ge 0$

- $\sum_{j=1}^{n} a_{ij}x_j \le b_i \Rightarrow \sum_{j=1}^{n} a_{ij}x_j + s_i = b_i$, $s_i \ge 0$.
- $x_i = x_i^+ - x_i^-$ where $x_i^+\ge 0, x_i^-\ge 0$.

e.g. $\min_{x_1,x_2} 2x_1+4x_2$ s.t. $3x_1+2x_2=14$, $x_1+x_2\ge 3$, $x_1\ge 0$ $\Rightarrow \min_{x_1,x_2^+,x_2^-,x_3} 2x_1+4x_2^+-4x_2^-$ s.t. $3x_1+2x_2^+-2x_2^-=14$, $x_1+x_2^+-x_2^--x_3=3$, $x_1,x_2^+,x_2^-,x_3\ge 0$.
(↳ slack variable.)

Def: Polyhedron $\{\vec{x}\in\mathbb{R}^n \mid A\vec{x}\ge\vec{b}\}$ $A\in\mathbb{R}^{m\times n}$ $\vec{b}\in\mathbb{R}^m$.

in "standard form" $\{\vec{x}\in\mathbb{R}^n \mid C\vec{x}=\vec{d}; \vec{x}\ge 0\}$.

Def: Extreme Points polyhedron $=P$ $x\in P$ is an extreme point (vertex of P) if we cannot find two vectors $\vec{y},\vec{z}\neq\vec{x}, \vec{y},\vec{z}\in P$ and $\lambda\in[0,1]$ s.t. $\vec{x}=\lambda\vec{y}+(1-\lambda)\vec{z}$.

Thm: P has an extreme point iff P does not contain a line.

11/01/2022 Tuesday.

Thm: Consider a LP $\min_{\vec{x}} \vec{c}^T\vec{x}$ s.t. $A\vec{x}\le\vec{b}$, if P has an extreme point and opt. solution exists and is finite, there exists an optimal solution that is an extreme point of P.

Proof: $P=\{\vec{x}\mid A\vec{x}\le\vec{b}\}$. $p^*=\min_{\vec{x}\in P}\vec{c}^T\vec{x}$. $Q=\{\vec{x}\mid A\vec{x}\le\vec{b}, \vec{c}^T\vec{x}=p^*\}$. Set of all opt. solutions.

Q is also a polyhedron.

P doesn't contain a line, $Q\subset P \Rightarrow Q$ doesn't contain a line $\Rightarrow Q$ has an extreme point.

Suppose $\vec{v}$ is a vertex of Q, WTS $\vec{v}$ is a vertex of P.

Prove by contradiction. $\exists \vec{y},\vec{z}\in P$, $\vec{y},\vec{z}\neq\vec{v}$ s.t. $\lambda\vec{y}+(1-\lambda)\vec{z}=\vec{v}$. $\lambda\in(0,1)$.

$\vec{v}\in Q \Rightarrow \vec{c}^T\vec{v}=p^*$ $\lambda\vec{c}^T\vec{y}+(1-\lambda)\vec{c}^T\vec{z}=\vec{c}^T\vec{v}=p^*$ since $\vec{c}^T\vec{y}\ge p^*, \vec{c}^T\vec{z}\ge p^* \Rightarrow \vec{c}^T\vec{y}=\vec{c}^T\vec{z}=p^*$.

$\exists \vec{y},\vec{z}\in Q, \vec{y},\vec{z}\neq\vec{v}$ s.t. $\lambda\vec{y}+(1-\lambda)\vec{z}=\vec{v}$ contradicts.

Quadratic Programs

$p^*=\min_{\vec{x}} \frac{1}{2}\vec{x}^TH\vec{x}+\vec{c}^T\vec{x}$ s.t. $A\vec{x}\le\vec{b}, C\vec{x}=\vec{d}$.

convex iff H is PSD.

$p^*=\min_{\vec{x}} \frac{1}{2}\vec{x}^TH\vec{x}+\vec{c}^T\vec{x}+d$. s.t. $A_{eq}\vec{x}=\vec{b}_{eq}$. $A\vec{x}\le\vec{b}$.

If $H=H^T$ symmetric, H is PSD $\Rightarrow Q_P$ is convex.

if H is symmetric, at least one negative eigenvalue.

$p^* = -\infty$, choose $\vec{x}$ be the eigenvector corr. to negative eigenvalue.

if H is symmetric, H is PSD, $\vec{c} \in R(H)$

$f(\vec{x}) = \frac{1}{2}\vec{x}^T H\vec{x} + \vec{c}^T\vec{x} = \frac{1}{2}(\vec{x}-\vec{x}_0)^T H(\vec{x}-\vec{x}_0) + d = \frac{1}{2}\vec{x}^T H\vec{x} + \frac{1}{2}\vec{x}_0^T H\vec{x}_0 - \vec{x}_0^T H\vec{x} + d$

$\vec{c} = -H^T\vec{x}_0 = -H\vec{x}_0$ $d = -\frac{1}{2}\vec{x}_0^T H\vec{x}_0$ want to choose $\vec{x} = \vec{x}_0$

if H is invertible, $\vec{x}^* = -H^{-1}\vec{c}$.

if H has null space, $\vec{x}^* =$ any $\vec{x}_0$ s.t. $-H\vec{x}_0 = \vec{c}$.

$H = U\Sigma V^T$ H has rank r. $H = U\begin{bmatrix}\Sigma_{r\times r} & 0\\ 0 & 0\end{bmatrix}V^T$.

Moore-Penrose pseudoinverse $H^\dagger = V\begin{bmatrix}\Sigma_{r\times r}^{-1} & 0\\ 0 & 0\end{bmatrix}U^T$ $HH^\dagger = U_r U_r^T$ $H^\dagger H = V_r V_r^T$ $HH^\dagger H = H$.

$\vec{x}^* = -H^\dagger\vec{c} + \vec{e}_g$ $\vec{e}_g \in N(H)$

if H is symmetric, H is PSD, $\vec{c} \notin R(H)$

$\vec{c} = -H\vec{x}_0 + \vec{r}$ $\vec{r} \in N(H^T) = N(H)$

$\vec{r} \notin R(H)$

$f(\vec{r}) = \frac{1}{2}\vec{r}^T H\vec{r} + \vec{c}^T\vec{r} = 0 + (-H\vec{x}_0 + \vec{r})^T\vec{r} = -\vec{x}_0^T H^T\vec{r} + \vec{r}^T\vec{r} = \|\vec{r}\|_2^2$

$f(k\vec{r}) = k\|\vec{r}\|_2^2$ $k \to -\infty$, $f(k\vec{r}) \to -\infty \Rightarrow p^* = -\infty$.

Equality constrained QP $\Rightarrow$ Unconstrained QP.

Application 1: Linear Control LQR.

$\vec{x}(t+1) = A\vec{x}(t) + Bu(t)$. $\vec{x}(t) = A^t\vec{x}(0) + \sum_{i=0}^{t-1} A^{t-i-1}B\cdot u(i)$

goal: reach $\vec{g}$ by time T. $\Rightarrow$ $\min_{\substack{x(t)\ t=0,\cdots,T\\ u(t)\ t=0,\cdots,T-1}} \underbrace{\|\vec{x}(T)-\vec{g}\|_2^2 + \sum_{t=0}^{T}\|u(t)\|_2^2}_{QP}$ s.t. $\underbrace{\vec{x}(t) = A^t\vec{x}(0) + \sum_{i=0}^{t-1} A^{t-i-1}B\cdot u(i)}_{LP}$.

Application 2: $\vec{y} = \vec{x} + \text{noise}$. find $\vec{x}$ s.t. it does not change much. consecutively.

$D = \begin{bmatrix}1 & -1 & 0 & \cdots & 0\\ 0 & 1 & -1 & \cdots & 0\\ \vdots & & \ddots & & \vdots\\ 0 & \cdots & 0 & 1 & -1\end{bmatrix}$ $D\vec{x} = \begin{bmatrix}x_2-x_1\\ x_3-x_2\\ \vdots\\ x_n-x_{n-1}\end{bmatrix}$

$p^* = \min_{\vec{x}} \|\vec{y}-\vec{x}\|_2^2$ s.t. $\text{card}(D\vec{x}) \le k$. card $\to$ #nonzero entries.

relax to $p^* = \min_{\vec{x}} \|\vec{y}-\vec{x}\|_2^2$ s.t. $\|D\vec{x}\|_1 \le 2$. relax to QP. encourage sparsity.

LP $\subset$ QP $\subset$ convex QP $\subset$ QCQP $\subset$ convex QCQP $\subset$ SOCP $\subset$ {all convex programs}.

Quadratically-constrained quadratic programming. QCQP.

$p^* = \min_{\vec{x}} \vec{x}^T H_0\vec{x} + 2\vec{c}_0^T\vec{x} + d_0$ s.t. $\vec{x}^T H_i\vec{x} + 2\vec{c}_i^T\vec{x} + d_i \le 0$ $i = 1,\cdots,m$

$\vec{x}^T H_j\vec{x} + 2\vec{c}_j^T\vec{x} + d_j = 0$ $j = 1,\cdots,q$.

convex if H_0, H_i PSD, $H_j = 0$.

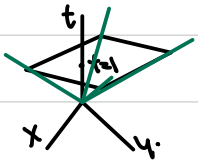
11/03/2022 Thursday.

SOCP

Def: A set of points $C \subseteq \mathbb{R}^n$ is a cone if $\lambda\vec{x} \in C$ if $\vec{x} \in C$ $\forall \lambda \ge 0$.

Def: convex cone if $\lambda\vec{x} \in C$ if $\vec{x} \in C$ $\forall \lambda \ge 0$ AND $\theta_1, \theta_2 \ge 0$ $\theta_1\vec{x}_1 + \theta_2\vec{x}_2 \in C$.

e.g. $C=\{(x,y)\mid |x|\le y\}$. $C=\{(x,y)\mid y\ge 0\}$.



Def. Polyhedron Cone $\{(\vec{x},t)\mid A\vec{x}\le \vec{b}t,\ t\in\mathbb{R},t\ge 0\}$.

Def: Ellipsoidal Cone. $\{(\vec{x},t)\mid \|A\vec{x}+\vec{b}t\|_2\le ct\}$

ellipse $\vec{x}^TP\vec{x}+q^T\vec{x}+r\le 0$ P is PSD.

$\|A\vec{x}+\vec{b}\|_2^2\le c^2 \Rightarrow \vec{x}^T\underbrace{A^TA}_{\text{PSD}}\vec{x}+2\vec{b}^TA\vec{x}+\vec{b}^T\vec{b}-c^2\le 0$

Def: Second-order Cone Special case of Ellipsoidal Cone.

in $\mathbb{R}^3$ $\{(x_1,x_2,t)\mid \sqrt{x_1^2+x_2^2}\le t\}$.

"ice cream cone"

Define: Second Order Cone Program Always Convex.

$\min_{\vec{x}} \vec{c}^T\vec{x}$ s.t. $\|A_i\vec{x}+\vec{b}_i\|_2\le \vec{c}_i^T\vec{x}+d_i$ $i=1,2,\cdots,m$ $\Rightarrow (A_i\vec{x}+\vec{b}_i,\ \vec{c}_i^T\vec{x}+d_i)$ must belong to soc.

Example: $\min_{\vec{x}}\sum_{i=1}^m\|A_i\vec{x}-\vec{b}_i\|_2 \iff \min_{\vec{x},y_i}\sum_{i=1}^m y_i$ s.t. $\|A_i\vec{x}-\vec{b}_i\|_2=y_i \xrightarrow[\text{same } p^*]{\text{relax}} \min_{\vec{x},y_i}\sum_{i=1}^m y_i$ s.t. $\|A_i\vec{x}-\vec{b}_i\|_2\le y_i$

if $A_k\vec{x}-\vec{b}_k=y_k-s$ $s\ge 0$, decrease y_k.

Example: $\min_{\vec{x}}\max_{i=1,\cdots,m}\|A_i\vec{x}-\vec{b}_i\|_2 \iff \min_{\vec{x},y} y$ s.t. $\|A_i\vec{x}-\vec{b}_i\|_2\le y$ $\forall i=1,\cdots,m$.

Facility Location Problem.

Example: Trilateration / GPS.

Packet transmission t_i^T Packet receive t_i^R $\Delta_i=t_i^R-t_i^T$

Offset δ $t_i^{true}=t_i^R+\delta$. unknown.

time of flight $f_i=t_i^{true}-t_i^T=t_i^R+\delta-t_i^T=\Delta_i+\delta$.

Distance $c\cdot f_i=c\Delta_i+c\delta$. c=speed of light.

$\|\vec{x}-\vec{a}_i\|_2=c\Delta_i+c\delta$ $\vec{x}$: my position $\vec{a}_i$: position of satellite i.

w/ 4 satellites $\|\vec{x}-\vec{a}_4\|_2^2-\|\vec{x}-\vec{a}_1\|_2^2$, $\|\vec{x}-\vec{a}_4\|_2^2-\|\vec{x}-\vec{a}_2\|_2^2$, $\|\vec{x}-\vec{a}_4\|_2^2-\|\vec{x}-\vec{a}_3\|_2^2 \Rightarrow$ 3 linear equations.

in 2D, 3 unknowns $\Rightarrow$ Solvable.

w/ 3 satellites $\min_{\vec{x}}\delta$ s.t. SOCP

① $2(a_3-a_1)^T\vec{x}+2c^2(\Delta_3-\Delta_1)\delta=c^2(\Delta_1^2-\Delta_3^2)+\|a_3\|_2^2-\|a_1\|_2^2$

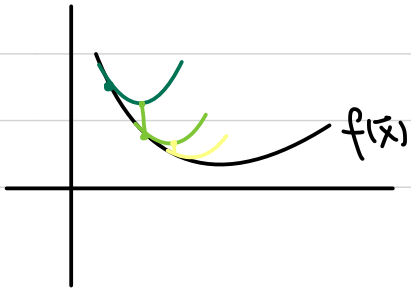
② $2(a_3-a_2)^T\vec{x}+2c^2(\Delta_3-\Delta_2)\delta=c^2(\Delta_2^2-\Delta_3^2)+\|a_3\|_2^2-\|a_2\|_2^2$

③ $\|x-a_3\|_2\le c\Delta_3+c\delta$ relaxed. linear obj. $\Rightarrow \vec{x}^*$ at boundary.

Newtons Method.

$\min_{\vec{x}} f(x)$ want $\vec{x}_0,\vec{x}_1,\cdots$ converging to $\vec{x}^*$.

$f(\vec{x}+\vec{v})=f(\vec{x})+\nabla f(\vec{x})^T\vec{v}+\frac{1}{2}\vec{v}^T\nabla^2 f(\vec{x})\vec{v}+\cdots$ quadratic approximation.



$\vec{x}_0$ initial point. Hessian is PD (invertible) $\Rightarrow \vec{v} = -(\nabla^2 f(\vec{x}))^{-1} \nabla f(\vec{x})$.

Newton step: $\vec{x}_{k+1} = \vec{x}_k - (\nabla^2 f(\vec{x}_k))^{-1} \nabla f(\vec{x}_k)$. No η. go to the actual minimum point instead of that direction

if Hessian is PSD $\Rightarrow$ Quasi-Newton Method.

Newtons Method vs Gradient Descent.

pros: converge faster. if $f(x)$ is quadratic, one step vs. forever.

cons: need to compute ∇^2 vs ∇.

11/08/2022 Tuesday.

The ℓ_1 Norm. ℓ_0 Norm (# nonzero elements) $\Rightarrow$ Not convex $\Rightarrow$ Relax to ℓ_1 Norm.

① min ℓ_1 norm.

$\min_{\vec{x}} \|\vec{x}\|_1$ s.t. $A\vec{x} = \vec{b}$. (full row rank.) Convex objective, linear Constraint. Not differentiable everywhere.

$\|\vec{x}\|_1 = \sum_{i=1}^n |x_i|$ $x_i = x_i^+ - x_i^-$ $|x_i| = x_i^+ + x_i^-$ $x_i^+ = \begin{cases} x_i & \text{if } x_i > 0 \\ 0 & \text{otherwise} \end{cases}$ $x_i^- = \begin{cases} -x_i & \text{if } x_i < 0 \\ 0 & \text{otherwise} \end{cases}$ $x_i^+, x_i^- \geq 0$.

$\min_{x_i^+, x_i^-} \sum_{i=1}^n x_i^+ + \sum_{i=1}^n x_i^-$ s.t. $A[\vec{x}^+ - \vec{x}^-] = \vec{b}$ $x_i^+ \geq 0, x_i^- \geq 0$.

claim: This new program will always choose only one of x_i^+, x_i^- to be nonzero.

proof: Suppose $x_i^+ > 0$ AND $x_i^- > 0$. WLOG, $x_i^+ > x_i^-$

Consider $x_i^+(\text{new}) = x_i^+ - x_i^-$ and $x_i^-(\text{new}) = 0$.

Still under the constraints But strictly decrease obj. $\Rightarrow$ contradicts x_i^+, x_i^- couldn't be the optimal.

② least squares.

$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_1 \iff \min_{\vec{e}} \|\vec{e}\|_1$ s.t. $A\vec{x} - \vec{e} = \vec{b}$ → ① → LP.

$\min_{\vec{x}} \sum_{i=1}^m (\vec{x} - \vec{b}_i)^2$ $\vec{x}^* = \frac{1}{m}\sum_{i=1}^m \vec{b}_i$

ℓ_1 Norm Parallel. $\min_{\vec{x}} \sum_{i=1}^m |\vec{x} - \vec{b}_i|$ = Median (robust than mean) Scalar only. (undefined if m is even.)

Scalar case: $x \in \mathbb{R}$, $b_i \in \mathbb{R}$ $1 \leq i \leq m$. $\min_x \sum_{i=1}^m |x - b_i|$ (outlier.)

$|x - b_i| = \begin{cases} x - b_i & \text{if } x > b_i \\ b_i - x & \text{if } x \leq b_i \end{cases}$ $\frac{d}{dx}|x - b_i| = \begin{cases} 1 & \text{if } x > b_i \\ -1 & \text{if } x < b_i \\ \text{DNE} & \text{if } x = b_i \end{cases}$ doesn't exist.

if Not differentiable everywhere. think of critical points ($\frac{d}{dx} = 0$, $\frac{d}{dx}$ DNE, or boundary).

Consider $\frac{d}{dx}\left(\sum_{i=1}^m |x - b_i|\right) = \frac{-7 \quad -5 \quad -3 \quad -1 \quad 1 \quad 3 \quad 5 \quad 7}{b_1 \quad b_2 \quad b_3 \quad b_4 \quad b_5 \quad b_6 \quad b_7}$ if $b_4 < x < b_5$ DNE otherwise.

$\sum_{i=1}^7 |x - b_i| = |x - b_4| + \sum_{i \neq 4} |x - b_i|$. $\forall x \in (b_3, b_5)$, $|x - b_3| + |x - b_5| = \text{constant}$.

③ LASSO

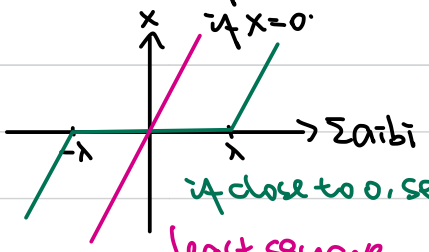
$\min_{\vec{x}} \frac{1}{2}\|A\vec{x}-\vec{b}\|_2^2 + \lambda\|\vec{x}\|_1$

Simple scalar case: $\min_x \frac{1}{2}\sum_{i=1}^{m}(a_i x - b_i)^2 + \lambda|x|$. $\frac{d}{dx}f(x) = \begin{cases} \sum(a_i x - b_i)a_i + \lambda & \text{if } x>0 \\ \sum(a_i x - b_i)a_i - \lambda & \text{if } x<0 \\ 0 & \text{if } x=0. \end{cases}$

if $x>0$, $x_i = \frac{\sum a_i b_i - \lambda}{\sum a_i^2}$ if $\sum a_i b_i > \lambda$.

if $x<0$, $x_i = \frac{\sum a_i b_i + \lambda}{\sum a_i^2}$ if $\sum a_i b_i < -\lambda$.

Soft thresholding. / sparsity.



if close to 0, set to 0.

least square

$\min_x \frac{1}{2}\sum_{i=1}^{n}(a_i x - b_i)^2$ $\vec{x}^* = \frac{\sum a_i b_i}{\sum a_i^2}$

11/10/2022 Thursday.

Coordinate Descent.

work well for LASSO.

(sequential version). $f(\vec{x}) = g(\vec{x}) + \sum_{i=1}^{n} h_i(x_i)$ g: convex, differentiable. h_i: convex.

initial guess: $\vec{x}^{(0)} = [x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)}]$

$x_1^{(1)} = \underset{x_1}{\operatorname{argmin}} f(x_1, x_2^{(0)}, x_3^{(0)}, \cdots, x_n^{(0)})$.

$x_1^{(k)} = \underset{y}{\operatorname{argmin}} f(y, x_2^{(k-1)}, \cdots, x_n^{(k-1)})$ fixed.

$x_2^{(k)} = \underset{y}{\operatorname{argmin}} f(x_1^{(k)}, y, x_3^{(k-1)}, \cdots, x_n^{(k-1)})$.

$\vdots$

$x_n^{(k)} = \underset{y}{\operatorname{argmin}} f(x_1^{(k)}, x_2^{(k)}, \cdots, y)$

Equality Constrained Newton.

equality constrained QP → unconstrained QP.

$\min_{\vec{x}} \frac{1}{2}\vec{x}^T H\vec{x} + \vec{c}^T\vec{x} + d$ s.t. $A\vec{x} = \vec{b}$. H is PD, strong duality.

$L(\vec{x}, \vec{v}) = \frac{1}{2}\vec{x}^T H\vec{x} + \vec{c}^T\vec{x} + d + \vec{v}^T(A\vec{x} - \vec{b})$.

KKT: $A\vec{x}^* = \vec{b}$. $H\vec{x}^* + \vec{c} + A^T\vec{v}^* = 0$. $\Rightarrow \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} \vec{x}^* \\ \vec{v}^* \end{bmatrix} = \begin{bmatrix} -\vec{c} \\ \vec{b} \end{bmatrix}$ invertible / non-singular ⇒ unique sol^n.

$\min_{\vec{x}} f(\vec{x}_0 + \vec{v}) = f(\vec{x}_0) + Df(\vec{x}_0)^T\vec{v} + \frac{1}{2}\vec{v}^T D^2 f(\vec{x}_0)\vec{v} + \cdots$ s.t. $A(\vec{x}_0 + \vec{v}) = \vec{b}$. $A\vec{v} = 0$

Sol^n: $\begin{bmatrix} D^2 f(\vec{x}) & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} \vec{v} \\ \vec{v} \end{bmatrix} = \begin{bmatrix} -Df(\vec{x}_0)^T \\ 0 \end{bmatrix}$

11/15/2022 Tuesday. LQG (linear quadratic Gaussian).

Application: Control (linear quadratic Regulator).

$x(t+1) = f(x(t), u(t))$.

Vertical Rocket. Goal: Maximize height by time T.

$x_1(t)$ = height. $x_2(t)$ = Vertical speed. $x_3(t)$ = weight of the rocket. $(x_1(0), x_2(0), x_3(0)) = (0, 0, M)$.

Force: ↑thrust: $G_T \cdot \dot{x}_3$ ↓inertia: $x_3 \cdot \ddot{x}_1 = x_3 \cdot \dot{x}_2$ ↓drag: $C_D \cdot \rho(x_1) \cdot x_2^2$ ↓gravity: $g x_3$

$\dot{x}_1(t) = x_2(t)$ $x_3(t)\dot{x}_2(t) = -C_D \rho(x_1) x_2^2 - g x_3 + G_T u(t)$ $\dot{x}_2(t) = -\frac{C_D}{x_3(t)} \rho(x_1) x_2^2 - g + \frac{G_T u(t)}{x_3(t)}$ $\dot{x}_3(t) = -u(t)$.

$\max x_1(T)$ s.t. $\dot{x}(t) = f(\vec{x}(t), \vec{u}(t))$ $\forall 0 \le t \le T$. linearlize / approximate to $\vec{x}(t+1) = A\vec{x}(t) + B\vec{u}(t)$.

Cost: $\sum_{t=0}^{N-1} \frac{1}{2}(\vec{x}_t^T Q \vec{x}_t + \vec{u}_t^T R \vec{u}_t) + \frac{1}{2}\vec{x}_N^T Q_f \vec{x}_N$ Q, R: PD matrices. (time invariant).

$\underset{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N, \vec{u}_1, \vec{u}_2, \cdots, \vec{u}_N}{\text{minimize}} \sum_{t=0}^{N-1} \frac{1}{2}(\vec{x}_t^T Q \vec{x}_t + \vec{u}_t^T R \vec{u}_t) + \frac{1}{2}\vec{x}_N^T Q_f \vec{x}_N$ s.t. $\vec{x}_{t+1} = A\vec{x}_t + B\vec{u}_t$ $t = 0, \cdots, N-1$.

Solve by Ricatti Equation

$L(\vec{x}_0, \cdots, \vec{x}_N, \vec{u}_0, \cdots, \vec{u}_N, \vec{\lambda}_1, \cdots, \vec{\lambda}_N) = \sum_{t=0}^{N-1} \frac{1}{2}(\vec{x}_t^T Q \vec{x}_t + \vec{u}_t^T R \vec{u}_t) + \frac{1}{2}\vec{x}_N^T Q_f \vec{x}_N + \sum_{t=0}^{N-1} \vec{\lambda}_{t+1}^T (A\vec{x}_t + B\vec{u}_t - \vec{x}_{t+1})$.

KKT condition: ① $\nabla_{\vec{u}_t} L = R\vec{u}_t + B^T \vec{\lambda}_{t+1} = 0$ $t = 0, 1, \cdots, N-1$. $\Rightarrow$ ⑥ $\vec{u}_t = -R^{-1}B^T\vec{\lambda}_{t+1}$.

② $\nabla_{\vec{x}_t} L = Q\vec{x}_t + A^T\vec{\lambda}_{t+1} - \vec{\lambda}_t = 0$ $t = 1, \cdots, N-1$. $\Rightarrow$ ⑦ $\vec{\lambda}_t = Q\vec{x}_t + A^T\vec{\lambda}_{t+1}$

③ $\nabla_{\vec{x}_N} L = Q_f\vec{x}_N - \vec{\lambda}_N = 0$. $\Rightarrow$ ⑤ $\vec{\lambda}_N = Q_f\vec{x}_N$. ——— Dynamics of the adjoint sys. $\vec{\lambda}$ = co-state.

Approach: Backward Induction. Goal: Find optimal $\vec{u}_t$.

Induction Hypothesis: $\vec{\lambda}_{t+1} = P_{t+1} \cdot \vec{x}_{t+1}$

Base Case: $t = N$. $\vec{\lambda}_N = Q_f \cdot \vec{x}_N \Rightarrow P_N = Q_f$.

Inductive Steps: $\vec{\lambda}_{t+1} = P_{t+1} \cdot \vec{x}_{t+1} = P_{t+1} \cdot (A\vec{x}_t + B(-R^{-1}B^T\vec{\lambda}_{t+1})) = (I + P_{t+1}BR^{-1}B^T)^{-1} \cdot P_{t+1}A\vec{x}_t$.

$\vec{\lambda}_t = A^T\vec{\lambda}_{t+1} + Q\vec{x}_t = \underbrace{[A^T(I + P_{t+1}BR^{-1}B^T)^{-1} \cdot P_{t+1} \cdot A + Q]}_{P_t}\vec{x}_t$

11/17/2022 Thursday. do not depend on every single datapoint, only ones at the boundary/margin ⇒ support vectors.

Support Vector Machine. Classification.

$\max_{\vec{w}, b, m} m$ s.t. $y_i(\vec{w}^T \cdot \vec{x}_i + b) > 0$ $\forall i$. $\frac{|\vec{w}^T\vec{x}_i + b|}{\|\vec{w}\|_2} \ge m$ $\forall i \Leftrightarrow \frac{y_i(\vec{w}^T \cdot \vec{x}_i + b)}{\|\vec{w}\|_2} \ge m$. ($y_i \in \pm 1$)

the separating hyperplane w/ greatest margin.

distance of the hyperplane to the closet datapoint.

$m = \frac{1}{\|\vec{w}\|_2}$

$\Leftrightarrow \max_{\vec{w},b} (\|\vec{w}\|_2)^{-1}$ s.t. $y_i(\vec{w}^T\vec{x}_i + b) \geq 1$. $\Leftrightarrow \min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|_2^2$ s.t. $y_i(\vec{w}^T\vec{x}_i + b) \geq 1$.

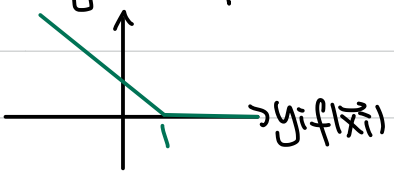
Hard-Margin SVM.

Soft-Margin SVM.

$\min_{\vec{w},b,\xi_i} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i$ s.t. $y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \xi_i$ $\xi_i \geq 0$.

↓ larger C, harder SVM.

Hinge-loss perspective of Soft-Margin SVM.



$L_{hinge}(y_i, \vec{w}^T\vec{x}_i + b) = \max(1 - y_i(\vec{w}^T\vec{x}_i + b), 0)$.

Hinge-loss perspective: $\min_{\vec{w},b} \frac{1}{n}\sum L_{hinge}(y_i, \vec{w}^T\vec{x}_i + b) + \lambda\|\vec{w}\|_2^2$

Soft-Margin SVM: $\min_{\vec{w},b,\xi_i} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n \xi_i$ s.t. $\xi_i \geq \max(1 - y_i(\vec{w}^T\vec{x}_i + b), 0)$

$\Leftrightarrow \min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_{i=1}^n L_{hinge}(y_i, \vec{w}^T\vec{x}_i + b)$ $C = n\lambda$.

11/22/2022 Tuesday.

Dual Perspective of Soft-Margin SVM.

$L(\vec{w}, b, \vec{\xi}, \vec{\lambda}, \vec{\beta}) = \frac{1}{2}\|\vec{w}\|_2^2 + C\cdot\sum_{i=1}^n \xi_i + \sum_{i=1}^n \lambda_i((1 - \xi_i) - y_i(\vec{w}^T\vec{x}_i + b)) + \sum_{i=1}^k \beta_i(-\xi_i)$

$= \frac{1}{2}\|\vec{w}\|_2^2 - \sum_{i=1}^n \lambda_i y_i(\vec{w}^T\vec{x}_i + b) + \sum_{i=1}^n \lambda_i + \sum_{i=1}^n (C - \lambda_i - \beta_i)\cdot\xi_i$

First-order KKT: $\nabla_{\vec{w}} L = \vec{w} - \sum \lambda_i y_i \vec{x}_i = 0 \Rightarrow \vec{w} = \sum \lambda_i y_i \vec{x}_i$

$\frac{\partial L}{\partial b} = -\sum \lambda_i y_i = 0$

$\frac{\partial L}{\partial \xi_i} = C - \lambda_i - \beta_i = 0$.

Complementary Slackness: $\lambda_i((1 - \xi_i) - y_i(\vec{w}^T\vec{x}_i + b)) = 0$ $\beta_i\xi_i = 0$ $\forall i$.

Case ①: $\lambda_i \neq 0$, $\lambda_i \neq C$, $0 < \lambda_i < C$

$(1 - \xi_i) - y_i(\vec{w}^T\vec{x}_i + b) = 0$. $\lambda_i \neq C \Rightarrow \beta_i \neq 0 \Rightarrow \xi_i = 0$.

$y_i(\vec{w}^T\vec{x}_i + b) = 1 \Rightarrow$ All points are exactly on the margin.

Case ②: $\lambda_i \neq 0$, $\lambda_i = C$

$\beta_i = 0$, ξ_i need not be zero. $\Rightarrow y_i(\vec{w}^T\vec{x}_i + b) = 1 - \xi_i \leq 1$

Case ③: $\lambda_i = 0$

$\beta_i = C \neq 0 \Rightarrow \xi_i = 0 \Rightarrow$ No margin violation.

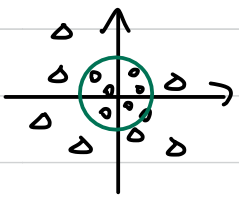
Dual: $L = -\frac{1}{2}\|\vec{w}\|_2^2 + \sum\lambda_i = -\frac{1}{2}\left(\sum_{i=1}^n \lambda_i y_i \vec{x}_i^T\right)\left(\sum_{i=1}^n \lambda_i y_i \vec{x}_i\right) + \sum_{i=1}^n \lambda_i = -\frac{1}{2}\vec{\lambda}^T \underbrace{\text{diag}(\vec{y})X X^T \text{diag}(\vec{y})}_{Q}\vec{\lambda} + \sum\lambda_i$.

↑ by first-order KKT

$d^* = \max_{\lambda,\beta} -\frac{1}{2}\vec{\lambda}^T Q \vec{\lambda} + \sum \lambda_i$ s.t. $\sum \lambda_i y_i = 0$, $\underline{C - \lambda_i - \beta_i = 0, \lambda_i \geq 0, \beta_i \geq 0}$ $\forall i$

$0 \leq \lambda_i \leq C$.

kernel SVM.



$f(\vec{x}) = \vec{w}\phi(\vec{x}) + b$. $\phi: (x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2)$.

e.g. poly, RBF (radial basis function).

10/11/2022 Tuesday.

Stochastic Gradient Descent.

$\vec{x}_{k+1} = \vec{x}_k - \eta \nabla f_i(\vec{x}_k)$. $\nabla f(\vec{x}) = \frac{1}{m}\sum_{i=1}^{n} \nabla f_i(\vec{x})$

Example. $f(\vec{x}) = \frac{1}{m}\sum_{i=1}^{n} \frac{1}{2}\|\vec{x} - \vec{p}_i\|_2^2$

SGD. stepsize: $\eta_t = \frac{1}{t}$, $\vec{x}_0 = 0$ $\nabla f_i(\vec{x}) = \frac{1}{2} \cdot 2(\vec{x} - \vec{p}_i) = \vec{x} - \vec{p}_i$.

$\vec{x}_1 = \vec{x}_0 - (\vec{x}_0 - \vec{p}_1) = \vec{p}_1$ $\vec{x}_2 = \frac{\vec{p}_1 + \vec{p}_2}{2}$

Projected Gradient Descent.

$\vec{x}_{k+1} = \Pi_X(\vec{x}_k - \eta_k \nabla f(\vec{x}_k))$ $\Pi_X(\vec{y}) = \operatorname{argmin}_{\vec{z} \in X} \|\vec{y} - \vec{z}\|_2^2$ Project back into the set X.

Conditional Gradient Descent.

$\vec{y}_k = \operatorname{argmin}_{\vec{y} \in X} \nabla f(x_k)^T \vec{y}$ $\vec{x}_{k+1} = (1 - \gamma_k)\vec{x}_k + \gamma_k \vec{y}_k$

↳ predetermined sequence.