Basic Linear Algebra

Least Squares
minimize residual $\vec{r}$: $\vec{r} = A\vec{x} - \vec{y}$

$\min_x \|Ax-y\|_2 = \min_x \|Ax-y\|_2^2$ (monotonically increasing)

$\|A\vec{x}-\vec{y}\|_2^2 = \sum_{i=1}^{m} (a_i^T x - y_i)^2$

solution (best $\vec{x}$): $\vec{x}^* = (A^TA)^{-1}A^T\vec{b}$

Orthogonality
angle btwn 2 vectors given by:

$\cos\theta = \frac{x^Ty}{\|x\|_2\|y\|_2} = \frac{\langle\vec{x},\vec{y}\rangle}{\|x\|_2\|y\|_2}$

orthogonal iff inner product $=0$, eg that: $\langle\vec{x},\vec{y}\rangle=0$

bc $\cos 90 = 0$

orthonormal vectors have unit norm & are orthogonal to each other

$\|x_i\| = 1$, $\langle x,y\rangle = 0$

$\perp$ is orthogonal complement of S

Parallelogram Law
$\|x+z\|^2 + \|x-z\|^2 = 2\|x\|^2 + 2\|z\|^2$

Norms
properties:

- $\|x\| \geq 0 \ \forall x \in X$
- $\|x\| = 0$ iff $x=0$
- $\|x+y\| \leq \|x\| + \|y\|$ (triangle ineq.)
- $\|\alpha x\| = |\alpha|\|x\|$

p-norms:

$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$, $1 \leq p \leq \infty$

p=2, p=1, p=∞

$p=1$: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ (Manhattan distance)

$p=2$: $\|x\|_2 = (\sum |x_i|^2)^{\frac{1}{2}}$ euclidean norm

$p=\infty$: $\|x\|_\infty = \max |x_i|$ measure of a vector by its largest element

ℓ_0 (cardinality) $\|x\|_0 = \sum_{k=1}^{n} \mathbb{I}(x_k \neq 0)$

$\mathbb{I}(x_k \neq 0) = \begin{cases} 1 & \text{if } x_k \neq 0 \\ 0 & \text{o/w} \end{cases}$

norm balls: $B_p = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$

Cauchy-Schwartz: $\langle x,y\rangle = x^Ty = \|x\|_2\|y\|_2\cos\theta$

$|x^Ty| \leq \|x\|_2\|y\|_2$

Hölders inequality: $|x^Ty| \leq \sum_{i=1}^{n}|x_iy_i| \leq \|x\|_p\|y\|_q$

$\frac{1}{p}+\frac{1}{q} \leq 1$

some properties we proved in homework:

$\|x\|_2 \leq \|x\|_1$, $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$

$\leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty$

$\|x\|_0 \geq \frac{\|x\|_1}{\|x\|_2}$, $x \neq \vec{0}$

Gram-Schmidt
$q_1 = a_1/\|a_1\|$

$s_2 = a_2 - \langle q_1, a_2\rangle q_1$

$q_2 = \frac{s_2}{\|s_2\|}$

$s_3 = a_3 - \langle a_3, q_1\rangle q_1 - \langle a_3, q_2\rangle q_2$

$q_3 = s_3/\|s_3\|$

Projections
$\Pi_S(x) = \arg\min_{y\in S} \|y-x\|$

↳ projection of x onto S

- onto 1-dim subspace: $\Pi_v(x) = \frac{\langle v,x\rangle v}{\|v\|^2}$
- consider $\vec{a},\vec{b} \in \mathbb{R}^n$, $\vec{b}^T\vec{b} = 1$
- vector projection $\vec{a}$ onto $\vec{b}$: $\text{proj}_b(\vec{a}) = (\vec{a}^T\vec{b})\vec{b}$
- scalar proj. $\vec{a}$ onto $\vec{b}$: $\vec{a}^T\vec{b}$
- projection length: $|\vec{a}^T\vec{b}|$

Matrices
$AB = [A_1\ A_2]\begin{bmatrix}B_1\\B_2\end{bmatrix} = A_1B_1 + A_2B_2$

$\begin{bmatrix}x_1^T\\ \vdots \\ x_n^T\end{bmatrix}A = \begin{bmatrix}x_1^TA\\ \vdots \\ x_n^TA\end{bmatrix}$

$CD = \begin{bmatrix}C_1\\C_2\end{bmatrix}[D_1\ D_2] = \begin{bmatrix}C_1D_1 & C_1D_2\\C_2D_1 & C_2D_2\end{bmatrix}$

$\vec{a} = [a_1\ a_2 \cdots a_m]^T$

$\vec{b} = [b_1\ b_2 \cdots b_p]$

$\begin{bmatrix}A & B\end{bmatrix}^T = \begin{bmatrix}A^T & C^T\\ B^T & D^T\end{bmatrix}$

$\begin{bmatrix}A\\B\end{bmatrix}^T = [A^T\ B^T]$

$\vec{a}\vec{b} = \begin{bmatrix}a_1b_1 & \cdots & a_1b_p\\ \vdots & & \vdots\\ a_mb_1 & \cdots & a_mb_p\end{bmatrix}$

$A[B\ C] = [AB\ AC]$

$\begin{bmatrix}A\\B\end{bmatrix}C = \begin{bmatrix}AC\\BC\end{bmatrix}$

Fundamental Thm of Lin Alg
$A \in \mathbb{R}^{m,n}$

- $N(A) \perp R(A^T)$; $R(A) \perp N(A^T)$

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

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

$\Rightarrow \dim(N(A)) + rk(A) = n$

$\dim(N(A^T)) + rk(A) = m$

$\Rightarrow R(A)^\perp = N(A)^T$

$N(A)^\perp = R(A)^T$

$N(A^TA) = N(A)$

$R(A^TA) = R(A^T)$

Determinants
- area (parallelogram) $= |\det(A)|$
- $\det(A) = 0 \iff A \in \mathbb{R}^{n,n}$ singular $\iff N\{A\} \neq \{0\}$
- $\det(A) = \det(A^T)$
- $\det(AB) = \det(BA) = \det(A)\det(B)$
- $\det(\alpha A) = \alpha^n \det(A)$ $B,A \in \mathbb{R}^{n,n}$

Inverses
- $AA^{-1} = A^{-1}A = I_n$
- $\det A = \det A^T = \frac{1}{\det A^{-1}}$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^T)^{-1} = (A^{-1})^T$

Quadratic Forms
$x^TAy = \sum_i\sum_j A_{ij}x_iy_j$

$\begin{bmatrix}x\\y\end{bmatrix}^T\begin{bmatrix}A & B\\C & D\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = x^TAx + x^TBy + y^TCx + y^TDy$

Special Matrices
- symmetric: square, $a_{ij} = a_{ji}$, $A = A^T$
- diagonal: square, $a_{ij} = 0$ if $i \neq j$
  ↳ eigenvalues are elements on diag.

  $A^{-1} = \begin{bmatrix}\frac{1}{a_1} & & \\ & \ddots & \\ & & \frac{1}{a_n}\end{bmatrix}$
- triangular matrices
  ↳ eigenvalues are elements on diagonal
- dyad: $A = \vec{u}\vec{v}^T$, $rk(A) = 1$, $\lambda = \vec{v}^T\vec{u}$, e-vector $= \vec{u}$

Quadratic Functions
- $q: \mathbb{R}^n \to \mathbb{R}$

$q(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j + \sum_{i=1}^{n} c_ix_i + d$

$q(x) = \vec{x}^TA\vec{x} + \vec{c}^T\vec{x} + d$

- quadratic form of diag(a):

$q(x) = x\,\text{diag}(a)\,x = \sum_{i=1}^{n} a_ix_i^2$

Norms (Matrix)
Properties
- $f(A) \geq 0$ $\forall A, B \in \mathbb{R}^{m,n}$
- $f(A) = 0$ iff $A = 0$
- $f(\alpha A) = |\alpha| f(A)$
- $f(A+B) \leq f(A) + f(B)$
- $f(AB) \leq f(A) f(B)$

→ Frobenius Norm

$\|A\|_F = \sqrt{\text{trace}(AA^T)}$

$= \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$

$= \sqrt{\sum_{i=1}^{m} \lambda_i(AA^T)}$

$= \sum_{i=1}^{m} \|\vec{a}_i^T\|_2^2$

→ lp-norms

$\|A\|_p \doteq \max_{u \neq 0} \frac{\|A\vec{u}\|_p}{\|\vec{u}\|_p}$

$= \max_{\|\vec{u}\|_p = 1} \|A\vec{u}\|_p$

→ ℓ_2-norm/spectral

$\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}$

→ ℓ_1-norm: largest absolute column sum

$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$

→ ℓ_∞-norm: largest absolute row sum

$\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$

Ellipsoid
$\mathcal{E} = \{x \in \mathbb{R}^n : \vec{x}^TP^{-1}x \leq 1\}$

Symmetric Matrices
Rayleigh quotient
$A \in \mathbb{S}^n$ (symmetric) $\forall x \neq 0$

$\lambda_{min}(A) \leq \frac{x^TAx}{x^Tx} \leq \lambda_{max}(A)$

Positive Semidefiniteness
$A \in \mathbb{S}^n$, all eigenvalues non-neg $\Rightarrow A \succeq 0$ (A is PSD)

$x^TAx \geq 0$ $\forall x \in \mathbb{R}^n$ $\Rightarrow A \succeq 0$

- $A \succeq 0 \Rightarrow A_{ii} \geq 0$
- if we have that $x^TAx > 0$ $\forall 0 \neq x \in \mathbb{R}^n$ then $A \succ 0$ (A is positive definite)
- $A \succeq 0$ is actually $A \succ 0$ iff A is invertible
- $A \succeq 0 \iff$ we can decompose $A = P^TP$ ($P \succeq 0$)
  ↳ $A^{\frac{1}{2}} = U\Sigma^{\frac{1}{2}}V^T = P$
- $\vec{x}^TA\vec{x} \geq 0 \Rightarrow$ region ~~bounded~~ where $\|\vec{x}\| \leq 1$ is bounded
- $\vec{x}^TA\vec{x} \geq 0 \Rightarrow$ region unbounded (non trivial nullspace)

Eigenvalues of Symmetric Matrices
$A \in \mathbb{S}^n$ → e-vecs/vals $= \{\vec{v}_1,\dots,\vec{v}_n\}$, $\{\lambda_1,\dots,\lambda_n\}$

eigenspaces corr. to diff eigenvals are mutually orthogonal

- $B \doteq A + cI_n$

$\lambda_i(B) = \lambda_i(A) + c$

$\vec{v}_i(B) = \vec{v}_i(B)$

shift property of eigenvalues

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

$\lambda_i(A^TA) = \lambda_i(AA^T)$ $\forall \lambda_i \neq 0$

- $A \doteq \vec{p}\vec{q}^T + \vec{q}\vec{p}^T$ $\vec{p},\vec{q} \in \mathbb{R}^n$ $\|p\|_2 = \|q\|_2 = 1$

$\vec{v}_{1,2}(A) = \vec{p} \pm \vec{q}$ $\lambda_{1,2} = c \pm 1$

Singular Value Decomposition
$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T = U_r \Sigma V_r^T$

$\sigma_1 \geq \sigma_2 \geq \dots \sigma_r \geq 0$

↳ compact form ($r = rk(A)$)

$\underset{n\times m}{A} = \underset{n\times n}{U}\,\underset{n\times m}{\Sigma}\,\underset{m\times m}{V^T}$

orthogonal matrices, orthogonal transformations (rotation/reflection)

scaling

↳ full SVD

U: e-$\vec{v}$ of AA^T

V: e-$\vec{v}$ of A^TA

- $\dim(N(A)) = n - r$
- $N(A) = R(V_{nr})$ $V_{nr} \doteq [V_{r+1} \cdots V_n]$
- $R(A) = R(U_r)$ $U_r = [u_1 \cdots u_r]$
- $\|A\|_F^2 = \text{trace}(A^TA)$

$= \sum_{i=1}^{n} \lambda_i(A^TA) = \sum_{i=1}^{r} \sigma_i^2$

- $\|A\|_2^2 = ~~\max~~ \lambda_{max}(A^TA) = \sigma_1^2$
- condition number $A \in \mathbb{R}^{n,n}$, invertible

$\kappa(A) = \frac{\sigma_1}{\sigma_n} = \|A\|_2 \cdot \|A^{-1}\|_2$

- $R(A^T) = R(V_r)$
- $N(A^T) = R(U_{nr})$

Cross product
$\vec{a} \times \vec{b} = \det\begin{bmatrix}\hat{i} & \hat{j} & \hat{k}\\ a_1 & a_2 & a_3\\ b_1 & b_2 & b_3\end{bmatrix}$

Low-Rank Approximation
$\min_{A_k \in \mathbb{R}^{m,n}} \|A - A_k\|_F^2 = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

s.t. $rk(A_k) = k$, $A = U\Sigma V^T$

↳ same result for spectral norm ($\|A - A_k\|_2^2$)

The Matrix Cookbook
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^T)^{-1} = (A^{-1})^T$
- $\text{Tr}(A+B) = \text{Tr}(A) + \text{Tr}(B)$
- $(A+B)^T = A^T + B^T$
- $\text{Tr}(A) = \sum A_{ii} = \sum \lambda_i = \text{Tr}(A^T)$
- $\nabla\|Ax - b\|_2^2 = 2A^TA x - A^Tb$
- $\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB)$
- $\partial(XY) = (\partial X)Y + X(\partial Y)$
- $\nabla_x a^Txb = ab^T$
- $\partial(X \circ Y) = (\partial X) \circ Y + X \circ (\partial Y)$
- $\nabla_x a^Tx^Tb = ba^T$
- $\nabla^2\|Ax - b\|_2^2 = 2A^TA$
- $\nabla_x x^Ta = \nabla_x a^Tx = a$
- $\nabla_x x^TAx = (A + A^T)x$
- $\nabla^2 x^TAx = A + A^T$

QR decomposition (n×n)
$\underset{n\times n}{A} = QR$ → invertible, upper triangular

↳ orthonormal (G-S) ↳ og matrix

R:

$\begin{bmatrix}\|s_1\| & q_1^Tv_2 & q_1^Tv_3 & \cdots & q_1^Tv_n\\ 0 & \|s_2\| & q_2^Tv_3 & \cdots & q_2^Tv_n\\ \vdots & 0 & \|s_3\| & & \vdots\\ 0 & 0 & 0 & & \|s_n\|\end{bmatrix}$

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

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

$A = QR$

$= Q\begin{bmatrix}R_1\\ 0_{m-n,n}\end{bmatrix}$

Least squares: G-S orthonormalization
$A = QR = [Q_1\ Q_2]\begin{bmatrix}R_1\\0\end{bmatrix}$ $\min_x \|A\vec{x} - \vec{b}\|_2 = x^*$

- squared norm of residual given by:

$\|\vec{r}\|_2^2 = \|\vec{b} - A\vec{x}\|_2^2 = \|Q_1^T\vec{b} - R_1\vec{x}\|_2^2 + \|Q_2^T\vec{b}\|$

- $\vec{x}^* = \arg\min_x \|Q_1^T\vec{b} - R_1\vec{x}\|_2^2 = R_1^{-1}Q_1^T\vec{b}$

PCA
1st PC [redacted] is e-vector corr. to λ_{max} of X^TX

1st PC $\vec{u}_1$ is the eigenvector corresponding to the largest value of C (the P_1) covariance matrix

$X = \begin{bmatrix}-x_1^T-\\ \vdots \\ -x_n^T-\end{bmatrix}$ data matrix ~~space~~ $x_i \in \mathbb{R}^m$ centered (0-mean)

$P_1)$ $\arg\max_{\substack{u \in \mathbb{R}^m\\ \|u\| = 1}} \vec{u}^T\left(\frac{1}{n}\sum_{i=1}^{n} \vec{x}_i\vec{x}_i^T\right)\vec{u}$

↳ finds a line going through the origin that maximizes scalar projections of points onto the line

$P_2)$ $\arg\min_{\substack{u \in \mathbb{R}^m\\ \|u\| = 1}} \sum_{i=1}^{n} \min \|\vec{x}_i - v_i\vec{u}\|_2^2$

↳ finds a line that minimizes sum of squared of distances from their points to their vector projections

↳ vector projection of $\vec{x}$ onto $\vec{u}$ is given by $v^*\vec{u}$ where

$v^* = \arg\min_{v \in \mathbb{R}} \|\vec{x} - v\vec{u}\|_2^2$

$= \langle\vec{x},\vec{u}\rangle$

$P_3)$ $\arg\min_{Y: rk(Y) \leq 1} \|X - Y\|_F$

↳ finds ~~the~~ rk 1 approximation to data matrix X

- P_1, P_2, P_3 are equivalent and find the 1st PC of X

Systems of Linear Equations
$A \in \mathbb{R}^{m,n}$

- $rk(A) = n$ (ie full column rank) iff A^TA is invertible
- $rk(A) = m$ (full row rank) iff AA^T is invertible

min-norm solution
↳ A has more cols than rows & is full row rank, then

$\min_{\vec{x}: Ax = y} \|\vec{x}\|_2$

$\vec{x}^* = A^T(AA^T)^{-1}\vec{y}$

least-squares
↳ interpretations:

- approx. soln of linear eqns
- projection onto R(A)
- linear regression
- minimal perturbation to feasibility $Ax = y + \delta y$
- best linear unbiased estimator

↳ solving via SVD

- $A = U\Sigma V^T$

$Ax = y$

- define: $Uw = y$, $\Sigma z = w$, $V^Tx = z$ ⇒ solve $w = U^Ty$, $z = \Sigma^{-1}w$, $x = Vz$

↳ solving L-S using QR decomp:

- $A = QR$

$Ax = y$

- $Q^TAx = Rx = \tilde{y}$

$\tilde{y} = Q^Ty$

→ solve $Rx = \tilde{y}$ by back-substitution

graphically:

$x \to \boxed{R} \xrightarrow{Rx = \tilde{y}} \boxed{Q}$, $y = Q\tilde{y}$

SVD
- $A \in \mathbb{R}^{m,n}$, non-zero

$\underset{m\times n}{A} = \underset{m\times m}{U}\,\underset{m\times n}{\Sigma}\,\underset{n\times n}{V^T}$

- $\Sigma = \text{diag}(\sigma_1, \dots \sigma_n)$
- $\sigma_i^2 = \lambda_i(AA^T)$

$= \lambda_i(A^TA)$

$\sigma_i = \sqrt{\lambda_i(A^TA)}$

$= \sqrt{\lambda_i(AA^T)}$

Spectral Decomp
- A symmetric

$\sum_{i=1}^{n} \lambda_i\vec{u}_i\vec{u}_i^T = A = U\Lambda U^T$ → orthogonal

↳ $U = \left[\frac{\vec{v}_1}{\|\vec{v}_1\|} \cdots \frac{\vec{v}_n}{\|\vec{v}_n\|}\right]$

↳ normalized e-vecs of A

Eigenvalue Decomposition
- A square, diagonalizable

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

↳ $U = [\vec{v}_1 \cdots \vec{v}_n]$

↳ eigenvecs of A

$\Lambda = \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}$

- $AU = \Lambda U$

Vector Calculus

- gradient – column vector: $\nabla f(\vec{x})_i = \frac{\partial f(\vec{x})}{\partial x_i}$, $j, i = 1, \ldots, n$
- Hessian – $n \times n$ matrix: $\nabla^2 f(\vec{x})_{ij} = \frac{\partial^2 f(\vec{x})}{\partial x_i x_j}$
- Jacobian – $m \times n$ matrix: $(Df)_{ij} = \frac{\partial f_i}{\partial x_j}$
- if f is a scalar-valued fcn, the Jacobian of ∇f is $(\nabla^2 f)^T = D_g(x)$ (so it's the transpose of the Hessian)

$\|A\|_F^2 = tr(A^TA)$

$\nabla_Q tr(Q^TQB)$, $f: \mathbb{R}^n \to \mathbb{R}$: $= QB^T + QB$, B is square
- $\nabla_A tr(AQ) = A^T$
- $\langle A, B \rangle = tr(A^TB)$
- $tr(A) = \sum_i a_{ii}$

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

$f: \mathbb{R}^n \to \mathbb{R}^m$ ↳ vector-valued fcn

$\nabla \vec{x}^TA\vec{x} = \nabla \vec{x}^T[\vec{a}_1 \cdots \vec{a}_n]\vec{x}$
$= \nabla \vec{x}^T[\vec{a}_1x_1 + \vec{a}_2x_2 + \cdots + \vec{a}_nx_n]$
$= \nabla \sum_{i=1}(\vec{x}^T\vec{a}_i)x_i$
$= \frac{\partial}{\partial x_j}((\vec{x}^T\vec{a}_j)x_j + \sum_{i \neq j}(\vec{x}^T\vec{a}_i)x_i)$
$= \vec{x}^T\vec{a}_j + a_{jj}x_j + \sum_{i \neq j} a_{ji}x_i$
$= a_j^Tx + a_j^Tx$
$= (A + A^T)x$

- $\nabla^2(\vec{x}^TA\vec{x}) = A + A^T$
- $\nabla \|A\vec{x} - \vec{b}\|_2^2 = 2(A^TA\vec{x} - A^T\vec{b})$
- $\nabla^2 \|A\vec{x} - \vec{b}\|_2^2 = 2A^TA$
- $\nabla \vec{x}^T\vec{a} = \vec{a}$
- $\nabla_{\vec{y}} \vec{y}^TA\vec{x} = A\vec{x}$ $\quad \nabla^2_{\vec{y}} \vec{y}^TA\vec{x} = 0$
- $\nabla \sum_i (x_i \log(x_i) - x_i) = [\log(x_1) \cdots \log(x_n)]^T$
- $\nabla^2 \sum_i (x_i \log(x_i) - x_i) = diag(\frac{1}{x_i})$
- $\nabla e^{\vec{a}^T\vec{x}+b} = \vec{a}e^{\vec{a}^T\vec{x}+b}$
- $\nabla^2 e^{\vec{a}^T\vec{x}+b} = \vec{a}\vec{a}^Te^{\vec{a}^T\vec{x}+b}$
- $(D_g)(\vec{g}(\vec{x})) = (D_g)(\frac{1}{2}\begin{bmatrix} x_1^2 \\ \vdots \\ x_n^2 \end{bmatrix})$ $= diag(x_i)$

ex of conv. sets
- $C = \{X \in S^n \mid \lambda_{min}(X) \geq 2\}$

ex of conv. fcns
- $f: \mathbb{R}^{m \times n} \to \mathbb{R}, f(X) = \sigma_{max}(X)$

Gradient Descent

$\vec{x}_{k+1} = \vec{x}_k - \eta \nabla f(\vec{x}_k)$

- to show convergence: $\vec{x}_{k+1} - \vec{x}_* = M(\vec{x}_k - \vec{x}_*) = M^k(\vec{x}_0 - \vec{x}_*)$
  ↳ need to show $\lim_{k \to \infty} M^k(\vec{x}_0 - \vec{x}_*) = \vec{0}$
  or $-1 < \lambda(M) < 1$ to show convergence
- μ-strongly convex: shows that gradients are changing fast enough
  $f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T(\vec{y} - \vec{x}) + \frac{\mu}{2}\|\vec{y} - \vec{x}\|_2^2$
- L-smoothly convex: gradients not changing too fast:
  $f(\vec{y}) \leq f(\vec{x}) + \nabla f(\vec{x})^T(\vec{y} - \vec{x}) + \frac{L}{2}\|\vec{y} - \vec{x}\|_2^2$
- if μ-strongly convex & L-smooth:
  $\|\vec{x}_{t+1} - \vec{x}_*\|_2 \leq (1 - \frac{\mu}{L})^{t+1}\|\vec{x}_0 - \vec{x}_*\|$

$\min_{x, y \in B} x^TAy = -\max_{x, y \in B} x^TA\ldots$
$= -\max_{x, y \in B} x^TU\Sigma V^Ty = \max \ldots$
$= -\sigma_{max}$

PCA

C: covariance matrix
$\sigma^2(\vec{u}) = \vec{u}^TC\vec{u}$ (var along any unit vector)
1st PC vector is eigenvector corr to largest eigenvalue of C

Inner prod of Symmetric (B) & skew-symmetric (A) matrices:
$\langle A, B \rangle = tr(A^TB) =$
$\sum_i \sum_j A_{ij}B_{ij} = \sum_i \sum_{j>i} A_{ij}B_{ij} + \sum_{i}\sum_{j<i} A_{ij}B_{ij} = \sum\sum A_{ij}(B_{ij} - B_{ij}) = 0$

Alternatively: $B = UU^T$
$\langle A, B \rangle = \langle A, UU^T \rangle = \langle AU, U \rangle = \langle -A^TU, U \rangle = -\langle U, AU \rangle$ ⇒ only true if $\langle AU, U \rangle = 0 \Rightarrow \langle A, B \rangle = 0$

$A \in \mathbb{R}^{3 \times 3}$, $A^TA = \Sigma$ $\quad S = \{Au : u \in \mathbb{R}^3, \|u\|_2 \ldots$
$H(\vec{\omega}) = \{\vec{z} \in \mathbb{R}^3 : \vec{\omega}^T\vec{z} = 0\}$ $\quad \vec{\omega} \in \mathbb{R}^3$. Find vec $\vec{\omega}$ s.t. proj. of S onto $H(\vec{\omega})$ is a circular disc.
↳ normal vector of hyperplane corr to (a) column of U.

$H(\vec{\omega}) = \{\vec{x} \in \mathbb{R}^n \mid \vec{x}^T\vec{\omega} = 0\}$, $B = \{\vec{x} \in \mathbb{R}^n \mid \|\vec{x}\|_2 \leq 1\}$, $P: \mathbb{R}^n \ldots$
$P(\vec{x}) = \text{argmin}_{\vec{y} \in H(\vec{\omega})} \|\vec{y} - \ldots$
$C = \{P(\vec{x}) \mid \vec{x} \in B\}$. Is C convex?
- say $Q \in \mathbb{R}^{n \times n-1}$ forms basis for $H(\vec{\omega})$
  Then $P(\vec{x}) = Q\,\text{argmin}_{\vec{\omega} \in \mathbb{R}^{n-1}} \|Q\vec{\omega} - \vec{x}\|_2^2$. Has solution $P(\vec{x}) = Q(Q^TQ)^{-1}Q^T\vec{x}$.
  set, P is affine operator, P is linear in x. B is convex sets are convex ⇒ C is convex. & affine transformations of

Taylor's Theorem

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

↳ fixed; variable; first order approximation

$f(\vec{x}_0 + \Delta\vec{x}) = f(\vec{x}_0) + \nabla f\Big|_{\vec{x} = \vec{x}_0}^T \Delta\vec{x} + \frac{1}{2!}(\Delta\vec{x})^T \nabla^2 f\Big|_{\vec{x} = \vec{x}_0} \Delta\vec{x}$

row vector; scalar; matrix (Hessian)

The Main Theorem

$f: \mathbb{R}^n \to \mathbb{R}$, diff'able everywhere
Ω: open set $\subseteq \mathbb{R}^n$

Consider the optimization problem: $\min_{\vec{x} \in \Omega} f(\vec{x})$

Then if x^* is an optimal solution $\frac{\partial f(x^*)}{\partial x} = 0$ (the optimum value is found when the derivative is 0)

Sensitivity of Solutions/Perturbations

$A\vec{x} = \vec{y}$, $A(\vec{x} + \delta\vec{x}) = \vec{y} + \delta\vec{y}$, then

$\frac{\|\delta x\|_2}{\|x\|_2} \leq \chi(A)\frac{\|\delta y\|_2}{\|y\|_2}$ $\quad \chi(A)$: condition number $\quad \chi(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$

↳ sensitivity to input perturbations $\quad \chi(A) = \|A^{-1}\|_2\|A\|_2$

$A\vec{x} = \vec{y}$, $(A + \delta A)(\vec{x} + \delta\vec{x}) = \vec{y}$ for some $\delta\vec{x}$, then

$\frac{\|\delta x\|_2}{\|x + \delta x\|_2} \leq \|A^{-1}\|_2\|A\|_2 \frac{\|\delta A\|_2}{\|A\|_2}$
$= \chi(A) = \frac{\sigma_{max}}{\sigma_{min}}$

↳ sensitivity to perturbations in coefficients matrix

$A\vec{x} = \vec{y}$, $(A + \delta A)(\vec{x} + \delta\vec{x}) = \vec{y} + \delta\vec{y}$, then

$\chi(A) \leq \frac{\gamma}{1 + \gamma\frac{\|\delta\vec{y}\|_2}{\|\vec{y}\|_2}}$ ($\gamma > 1$, given)

$\frac{\|\delta x\|_2}{\|x + \delta x\|_2} \leq \gamma\left(\frac{\|\delta y\|_2}{\|y\|_2} + \frac{\|\delta A\|}{\|A\|_2}\right)$ sensitivity to joint perturbation A, y

Probability

- empirical mean of $\vec{y} \in \mathbb{R}^K$: $\mu(\vec{y}) = \frac{1}{K}\vec{1}^T\vec{y} = \frac{1}{K}\sum_{i=1}^K y_i$
- empirical variance $\vec{y} \in \mathbb{R}^K$: $\sigma^2(\vec{y}) = \frac{1}{K}\|\vec{y} - \mu(\vec{y})\vec{1}\|_2^2$
  $= \frac{1}{K}\sum_i (y_i - \mu(\vec{y}))^2$
- empirical standard deviation $\sigma(\vec{y}) = \sqrt{\sigma^2(\vec{y})}$
- empirical covariance $\sigma(\vec{\omega}, \vec{y})$
  $= \frac{1}{K}(\vec{\omega} - \mu(\vec{\omega})\vec{1})^T(\vec{y} - \mu(\vec{y})\vec{1})$
  $= \frac{1}{K}\sum_{i=1}^K (\omega_i - \mu(\vec{\omega}))(y_i - \mu(\vec{y}))$
  $\sigma(\vec{y}, \vec{y}) = \sigma^2(\vec{y})$
- Bayes rule: $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A})} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Total probability rule: $\mathbb{P}(B) = \sum_{i=1}^n \mathbb{P}(B \cap A_i) = \sum_{i=1}^n \mathbb{P}(B|A_i)\mathbb{P}(A_i)$
- independence: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, $\mathbb{P}(A|B) = \mathbb{P}(A)$
- expectation $\mathbb{E}[X] = \sum_{x \in X} x\mathbb{P}[X = x]$, $\mathbb{E}[g(x)] = \sum_x g(x)\mathbb{P}[X = x]$
  ↳ $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- variance $Var(X) = \mathbb{E}[(X - \mu)^2]$
  $= \mathbb{E}[X^2] - \mathbb{E}[X]^2$
  $Var(cX) = c^2Var(X)$
  $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- covariance $cov(X, Y) = \mathbb{E}[X]\mathbb{E}[Y]$
- continuous distribution
  → Probability Density Fcn
  - $f_X(x) \geq 0 \quad x \in \mathbb{R}$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  → Cumulative Dens. Fcn
  - $F_X(x) = \mathbb{P}(X \leq x)$
    $= \int_{-\infty}^x f_X(t)dt$
  - $f_X(x) = \frac{d}{dx}F_X(x)$
  → Expectation $\mathbb{E}[X] = \int x f \ldots$
- Distributions
  - Binomial ($X \sim Bin(n, p)$)
    $\mathbb{P}[X = i] = \binom{n}{i}p^i(1-p)^{n-i}$
    $\mathbb{E}[X] = np \quad Var(X) = np(1-p)$
  - Poisson ($X \sim Poisson(\lambda)$)
    $\mathbb{P}[X = i] = \frac{\lambda^i}{i!}e^{-\lambda} \quad \mathbb{E}[X] = \lambda \quad Var(X) = \lambda$
  - geometric ($X \sim geometric(p)$)
    $\mathbb{P}[X = i] = (1-p)^{i-1}p$
    $\mathbb{E}[X] = \frac{1}{p} \quad Var(X) = \frac{1-p}{p^2}$
  - exponential ($X \sim Exp(\lambda)$)
    PDF $= \lambda e^{-\lambda x} \quad$ CDF $= 1 - e^{-\lambda x}$
    $\mathbb{E}[X] = 1/\lambda \quad Var(X) = 1/\lambda^2$
  - Uniform (continuous) ($X \sim uniform(a, b)$)
    PDF $= \frac{1}{b-a}$ $\quad$ CDF $= \frac{x-a}{b-a}$
    $\mathbb{E}[X] = (a+b)/2 \quad Var(X) = (b\ldots$
  - Normal/Gaussian ($X \sim N(\mu, \sigma^2)$)
    PDF $= \frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{(x-\ldots}{2\ldots})$
    $\mathbb{E}[X] = \mu \quad Var(x) = \ldots$

MLE

$\text{argmax}\, f(Y_1 = y_1, \ldots, Y_n = y_n \mid \omega = \omega_0)$
$= \text{argmax} \prod f(z_i = y_i - x_i^T\omega \mid \omega = \omega_0)$
$= \text{argmin} \|S(X\omega_0 - y)\|_2^2$
where $S = \begin{bmatrix} \sqrt{\frac{1}{2\sigma_1^2}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\frac{1}{2\sigma_n^2}} \end{bmatrix}$

↳ if iid all σ_i^2 are the same so S is a weight matrix
⇒ MLE becomes least squares

MAP

$Y_i = x_i^T\omega + z_i \quad z_i \sim N(0, \sigma_i^2) \quad \omega_i \sim N(\mu_i, S_i^2)$
$\text{argmax}_{\vec{\omega}}\, f(\omega \mid Y = y)$ $\quad \vec{\omega} \sim N(\vec{\mu}, \Sigma_\omega)$
$= \text{argmax}_{\vec{\omega}}\, f(Y = y \mid \omega)f(\omega)$
$= \text{argmin}_{\vec{\omega}} \|S(X\omega - y)\|_2^2 + \|\Sigma^{-1}(\vec{\omega} - \vec{\mu})\|_2^2$ $\quad \Sigma_\omega = \begin{bmatrix} S_1^2 & 0 \\ 0 & \ddots\, S_n^2 \end{bmatrix}$

Convergence of Gradient Descent for Ridge Regression

$f(x) = \frac{1}{2}(\|A\vec{x} - \vec{y}\|_2^2 + \lambda\|\vec{x}\|_2^2)$, $\vec{x}^* = (A^TA + \lambda I)^{-1}A\ldots$
$x_{t+1} = \vec{x}_t - \eta\nabla f(x) = \vec{x}_t - \eta(A^T(A\vec{x}_t - \vec{y}) + \lambda\vec{x}_t)$
$= (I - \eta(A^TA + \lambda I))\vec{x}_t + \eta A^T\vec{y}$
$\vec{x}_{t+1} - \vec{x}_* = (I - \eta(A^TA + \lambda I))\vec{x}_t + \eta A^T\vec{y} - (A^TA + \lambda I)^{-1}A^T\vec{y}$
$= (I - \eta(A^TA + \lambda I))\vec{x}_t + \eta(A^TA + \lambda I)\vec{x}_* - \vec{x}_*$
$= (I - \eta(A^TA + \lambda I))\vec{x}_t - (I - \eta(A^TA + \lambda I))\vec{x}_*$
$= (I - \eta(A^TA + \lambda I))(\vec{x}_t - \vec{x}_*)$
$= (I - \eta(A^TA + \lambda I))^{t+1}(\vec{x}_0 - \vec{x}_*)$

wts: $\lim_{t \to \infty} \vec{x}_t = \vec{x}^*$ $\quad \forall \vec{x}_0$ iff $\lambda_i(I - \eta(A^\ldots$ are contained in $(-1, 1)$:
- diagonalize $(I - \eta(A^TA) + \lambda I) = V\Gamma V^T$
- iff $\Gamma^t \to 0$ as $t \to \infty$ ⇒ $\vec{x}_t - \vec{x}_* = (V\Gamma V^T)^t(\vec{x}_0 - \vec{x}_*) = V\Gamma^tV^T(\vec{x}_0 - \vec{x}_*)$ → goes t… $\vec{x}_t \to \vec{x}_*$ $\forall \vec{x}_0$ iff $\lambda_i(I - \eta(A^TA + \lambda I))$ are in $(-1, 1)$
- to get quantitative estimate on rate of convergence:
  $\|\vec{x}_t - \vec{x}_*\| = \|V\Gamma^tV^T(\vec{x}_0 - \vec{x}_*)\| = \|\Gamma^t\|\|V^T(\vec{x}_0 - \vec{x}_*)\| = \|\Gamma^t\|\|\vec{x}_0 - \vec{x}_*\|_2 = (\max \ldots$
  $= (\max_{i=1} |\lambda_i(I - \eta(A^TA + \lambda I))|)^t\|\vec{x}_0 - \vec{x}_*\|_2 = (\max_{i=1} |1 - \eta\lambda - \eta\sigma_i(A)^2|)^t\|\vec{x}_0 - \vec{x}_*\|_2$
  $= (\max_{i=1} |1 - \eta\lambda - \eta\lambda_i(A^TA)|)^t\|\vec{x}_0 - \vec{x}_*\|_2 = (\max |1 - \eta\lambda_i(A^TA + \lambda I)|)^t\|\vec{x}_0 - \vec{x}_*\ldots$
- For $\lambda > 0$, values for $\eta > 0$ that allow gradient to converge:
  $1 > 1 - \eta\lambda - \eta\sigma_{min}(A)^2$
  $\vdots$
  $-1 < 1 - \eta\lambda - \eta\sigma_{max}(A)$ ⇒ $0 < \eta < \frac{2}{\sigma_{max}(A)^2 + \lambda}$
  $\max_{i=1} |1 - \eta\lambda - \eta\sigma_i(A)^2| < 1$ which mean…

Ridge Regression

$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|_2^2 + \underbrace{\lambda^2\|\vec{x}\|_2^2}_{\text{regularizer}}$ $\quad \lambda > 0$ $\quad \nabla f = 2A^TAx - 2A^Tb + 2\lambda^2x$
$\vec{x}^* = (A^TA + \lambda^2I)^{-1}A^T\vec{b}$

Tikhanov Regularization

$\min_{\vec{x}} \|W^{\frac{1}{2}}(A\vec{x} - \vec{y})\|_2^2 + \lambda\|S^{1/2}(\vec{x} - \vec{z})\|_2^2$
$\vec{x}^* = (A^TWA + \lambda S)^{-1}(A^TW\vec{y} + \lambda S\vec{z})$

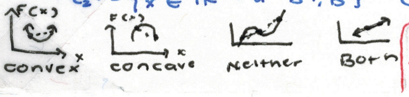
$\frac{d}{dx}\log_a f(x) = \frac{f'(x)}{\ln(a) \cdot f(x)}$ $\quad \frac{d}{dx}\frac{f(x)}{g(x)} = \frac{g(x)f'(x) - f(x)g'(x)}{g(x)^2}$

chain rule: $(f(g(x)))' = f'(g(x)) \cdot g'(x)$

product rule: $\frac{d}{dx}(f(x)g(x)) = f(x)g'(x) + f'(x)\ldots$

Convexity (Sets)

- C convex ⇔ $\forall \vec{x}_1, \vec{x}_2 \in C$ $\forall \theta \in [0, 1]$
  $\theta\vec{x}_1 + (1-\theta)\vec{x}_2 \in C$
- operations preserving convexity:
  - intersection ($C = \bigcap_{i=1,\ldots,m} C_i$)
  - affine transformation: map $f: \mathbb{R}^n \to \mathbb{R}^m$ is affine, $C \subset \mathbb{R}^n$ convex, then $f(C) = \{f(x) : x \in C\}$ is convex
- seperating hyperplane thm: C_1, C_2 convex subsets of $\mathbb{R}^n$ having empty intersection ($C_1 \cap C_2 = \emptyset$), then $\exists\mathcal{H}$, a seperating hyperplane for C_1, C_2
  $\mathcal{H}_{C_1} := \{x \in \mathbb{R}^n : \vec{a}^T\vec{x} \leq b\}$ $\quad C_1 \subseteq \mathcal{H}_{C_1}$
  $\mathcal{H}_{C_2} := \{x \in \mathbb{R}^n : \vec{a}^T\vec{b} \geq b\}$ $\quad C_2 \subseteq \mathcal{H}_{C_2}$



Convex Optimization Problems

$\min_{\vec{x} \in S} f_0(x)$: $f_i(x) \leq 0 \quad i = 1, \ldots, m$
$h_i(x) = 0 \quad i = 1, \ldots, p$

is convex if the objective fcn f_0 is convex, fcns defining inequality constraints (f_i) are convex fcns defining equality constraints (h_i) are affine & the sets it's defined over is convex.

- all local optima are global
- optimal set is convex

Convexity (Functions)

- $f: \mathbb{R}^n \to \mathbb{R}$ convex if $dom(f)$ is a convex set & $\forall \vec{x}, \vec{y} \in dom(f)$, $\forall \theta \in [0, 1]$
  $f(\theta\vec{x} + (1-\theta)\vec{y}) \leq \theta f(\vec{x}) + (1-\theta)f(\vec{y})$
  ↳ strictly convex if inequality is strict
  ↳ concave if inequality is flipped
- any norm is a convex fcn
- Jensen's inequality:
  f convex, $\sum\theta_i = 1$, $\vec{x}_1, \ldots, \vec{x}_k \in dom(f)$
  $f(\theta_1\vec{x}_1 + \theta_2\vec{x}_2 + \cdots + \theta_k\vec{x}_k) \leq \theta_1 f(\vec{x}_1) + \cdots + \theta_k f(\vec{x}_k)$
- pointwise max of affine fcns is convex
- First-order condition:
  f diff'able, convex, $dom(f)$ convex
  $f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T(\vec{y} - \vec{x})$
- Second-order condition:
  f 2x diff'able, convex, $\forall \vec{x} \in dom(f)$
  ↳ strictly convex if Hessian is PD
  $\nabla^2 f(\vec{x}) \succeq 0$
  ↳ Hessian is PSD
- epigraph: f convex ⇔ $epi(f)$ conv. set
  $epi\, f = \{(x, t), x \in dom(f), t \in \mathbb{R} : f(x) \leq t\}$

f_1, f_2 convex ⇒ $f(x) = \max(f_1(x), f\ldots$ convex
$dom(f) = dom(f_1) \cap dom(f_2)$
→ $epi(f) = \{(\vec{x}, t) \mid \vec{x} \in dom(f), f(\vec{x}) \leq t\}$
$= \{(\vec{x}, t) \mid \vec{x} \in dom(f), \max(f_1(\vec{x}), f_2(\vec{x})) \leq t\}$
$= \{(\vec{x}, t) \mid \vec{x} \in dom(f_1) \cap dom(f_2), f_1(\vec{x}) \leq t, f_2\ldots$
$= \{(\vec{x}, t) \mid \vec{x} \in dom(f_1), f_1(\vec{x}) \leq t\} \cap \{(\vec{x}, t) \mid \vec{x} \in dom(f_2), f_2\ldots$
$= epi(f_1) \cap epi(f_2)$
⇒ f convex

f convex, nondecreasing g convex ⇒ $h = f \circ g = f(g(\vec{x}))$ convex
$h(\lambda\vec{x} + (1-\lambda)\vec{y}) = f(g(\lambda\vec{x} + (1-\lambda)\vec{y}))$
$\leq f(\lambda g(\vec{x}) + (1-\lambda)g(\vec{y}))$
$\leq \lambda f(g(\vec{x})) + (1-\lambda)f(g(\vec{y}))$
$= \lambda h(\vec{x}) + (1-\lambda)h(\vec{y})$ ⇒ h convex

Hyperplane $H = \{\vec{x} \mid \vec{a}^T\vec{x} = b\}$ is conv…
$\forall \vec{x}_1, \vec{x}_2 \in H$ $\forall \theta \in [0, 1]$
$\vec{a}^T(\theta\vec{x}_1 + (1-\theta)\vec{x}_2) = \theta(\vec{a}^T\vec{x}_1) + (1-\theta)\ldots$
$= \theta b - (1-\theta)b = b$

Standard Form/Primal Prob

$$p^* = \text{minimize } f_0(x) \quad \text{s.t. } f_i(x) \leq 0 \quad i=1,\dots,m; \quad h_i(x) = 0 \quad i = 1,\dots,p$$

$\rightarrow p^*$: optimal value, x^*: opt. point

Lagrangian

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

· affine in λ, ν

Dual Function

$$g(\lambda,\nu) = \inf_{x \in X} \mathcal{L}(\vec{x},\vec{\lambda},\vec{\nu})$$

· concave in λ, ν

↳ $g(\lambda,\nu) \leq p^* \quad \forall \lambda \geq 0, \nu$

- dual is a lower bound

Dual problem

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

· best lower bound for the primal, is a convex opt. prob. bc objective (to be maximized) is concave & constraints are convex.

Weak Duality

$d^* \leq p^*$

· holds even when primal's non-convex

· $p^* - d^*$ ← duality gap

Strong Duality ($d^* = p^*$)

s.t. $f_i(x) \leq 0 \quad i=1,\dots,m$

$Ax = b$

· $f_0, \dots, f_m$ all convex

Slater's Condition

$\exists x \in \text{relint}(D)$ s.t.

$f_i(x) < 0 \quad i=1,\dots,m$

$Ax = b$

↳ this x is "strictly feasible"

↳ here, strong duality holds

Refined Slater

· if some constraints are affine (ie, $f_1,\dots,f_k$) then strong duality holds if this (weaker) condition holds: $\exists x \in \text{relint}(D)$ s.t.

$f_i(x) \leq 0 \quad i=1,\dots,k$

$f_i(x) < 0 \quad i=k+1,\dots,m$

$Ax = b$

KKT Conditions (Nonconvex problems)

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) = 0$$

↳ where x^* is primal optimal & (λ^*, ν^*) are dual optimal w/ 0 duality gap, and where $\forall f_i, h_i$ are diff'able

Then we have the following KKT conditions:

1. Primal feasibility:
   $f_i(x^*) \leq 0 \quad i=1,\dots,m$
   $h_i(x^*) = 0 \quad i=1,\dots,p$
2. Dual feasibility:
   $\lambda_i^* \geq 0 \quad i=1,\dots,m$
3. Complementary slackness:
   $\lambda_i^* f_i(x^*) = 0 \quad i=1,\dots,m$
4. Stationarity:
   $\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) = 0$

· essentially: for any optimization problem w/ diff'able obj. & constr. fcns, for which strong duality holds, any pair of primal & dual optimal points must sastisfy the KKT conds.

"necessary": strong duality & diff'ability ⇒ KKT

KKT (Convex problems)

· when the primal is convex (ie, f_i, h_i affine) and $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ are points that satisfy the KKT conditions, then $\tilde{x}$, $(\tilde{\lambda}, \tilde{\nu})$ are primal & dual optimal & strong duality holds

↳ "sufficient": convex primal, KKT satisfied ⇒ strong duality

Linear Programs

· optimization problem where objective fcn & constraints are all linear

Standard LP Form:

$$\min c^T x \quad \text{s.t. } Ax = b, \; x \geq 0$$

For $\forall$ LPs, so long as the prob. is feasible, strong duality holds

Inequality Form LP

$$\min c^T x \quad \text{s.t. } Ax \leq b$$

LPs & Duality

Consider $\min c^T x$ s.t. $Ax - b \leq 0$

$\rightarrow L(x,\lambda) = c^T x + \lambda^T(Ax - b) = (A^T\lambda + c^T)x - b^T\lambda$

$\rightarrow g(\lambda) = \min_x L(x,\lambda) = \begin{cases} -\infty & \text{if } A^T\lambda + c \neq 0 \\ -b^T y & \text{o/w} \end{cases}$

$\rightarrow d^* = \max_{\lambda \geq 0} g(\lambda) = \max_{\lambda \geq 0,\ A^T\lambda + c = 0} (-b^T\lambda)$