Algorithm

Good: Finiteness, Definiteness, Input/Output, Effective.

Efficiency → computational complexity → O(·) notation

Fixed point arithmetic

1. Signed integer used. $M=8$, $+21 = 00010101$
2. One's complement: $M=8$, $+21 : 00010101$, $-21 : 11101010$
3. Two's complement: $M=8$, $+21 : 00010101$, $-21 : 11101011$

Positional number system

$\forall z > 0$ (can be -0), same base B.

$z = a_k B^k + \dots + a_1 B + a_0 + a_{-1} B^{-1} + \dots$

coefficients $a_j \in \{0, 1, \dots, B-1\}$, radix point

$z = (a_k \cdots a_1 a_0 . a_{-1} \cdots)_B$

$\binom{n}{k} = n!/k!(n-k)!$ ← overflow/underflow

$= \frac{k+1}{1} \times \frac{k+2}{2} \cdots \times \frac{n}{n-k}$

Trade-off between statistical cost (bias) and computational cost (time/resource)

O(·) notation

Let $f, g : \{1,2,\dots\} \to \mathbb{R}$. $f = O(g)$ if $\exists M, n_0$ s.t. $|f(n)| \le M|g(n)|\ \forall n > n_0$

flops (floating point operations)

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

$x + y = O(n) = n$ flops

$x^T y = O(n) = 2n-1$ flops

$\hat\beta = \arg\min_\beta \|y - X\beta\|^2$

$= (X'X)^{-1}X'y$, $X'X = O(np^2)$, $(X'X)^{-1} = O(p^3)$, $X'y = O(np)$

$= O(np^2)$, $(X'X)^{-1}X'y = O(p^2)$

sample size $n \ge 2$ [illegible]

Recurrence

Pascal's triangle: $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$ with conditions $\binom{-1}{n} = \binom{n}{0} = 1$

Binomial PMF

$P(B=k+1) = \frac{P(B=k)}{\binom{n}{k}}\binom{n}{k+1}\frac{p}{1-p}$

$= P(B=k)\frac{n-k}{k+1}\frac{p}{1-p}$

Partitions

Set of n elements, B_n the number of partitions

$B_{n+1} = \sum_{k=0}^{n}\binom{n}{k} B_{n-k}$

Sample [illegible]

$\bar x_n = \frac{1}{n}\sum x_i$, $\bar x_{n+1} = \frac{1}{n+1}(n\bar x_n + x_{n+1})$

$S_n^2 = \frac{1}{n-1}\sum (x_i - \bar x_n)^2 \Rightarrow S_{n+1}^2 = \frac{n-1}{n}S_n^2 + \frac{1}{n+1}(x_{n+1} - \bar x_n)^2$

$nS_{n+1}^2 = \sum^{n+1}(x_i - \bar x_{n+1})^2$

$= \sum (x_i - \bar x_n)^2 + (x_{n+1} - \bar x_{n+1})^2$

$= \sum (x_i - \bar x_n)^2 + n(\bar x_n - \bar x_{n+1})^2 + (x_{n+1} - \bar x_{n+1})^2$

$= (n-1)S_n^2 + \frac{n}{n+1}(x_{n+1} - \bar x_n)^2$

Polynomials: $p(x) = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n$

$= O(n)$: $3n-1$ flops

Horner's method $= x(a_0x^{n-1} + \dots + a_{n-1}) + a_n$

$= x b_{n-1}(x) + a_n$

$b_0(x) = a_0$, $b_k(x) = x b_{k-1}(x) + a_k$, $= 2n$ flops

$\Rightarrow p'(x) \Rightarrow b_1'(x) = b_0(x)$

$b_k'(x) = x b_{k-1}'(x) + b_{k-1}(x)$, $k = 2, \dots, n$

$\Rightarrow n-1$ multiplications; $n-1$ additions.

Ex: Infinite urn (red/blue balls)

red: p, blue: $q = 1-p$

$N_{rb} = n$ s.t. $R \ge r$, $B \ge b$.

$E(N_{0b}) = \frac{b}{q}$, $E(N_{0b}^2) = \frac{b(1-q)}{q^2} + \frac{b^2}{q^2}$

$E(N_{r0}) = \frac{r}{p}$; $E(N_{r0}^2) = \frac{r(1-p)}{p^2} + \frac{r^2}{p^2}$

$E(N_{rb}) = E(E(N_{rb} \mid \text{last}))$

$= E(N_{rb} \mid r)P(r) + E(N_{rb} \mid b)P(b)$

$= (E(N_{r-1,b}) + 1)p + (E(N_{r,b-1}) + 1)q$

$= E(N_{r-1,b})p + E(N_{r,b-1})q + 1$

$E(N_{rb}^2) = E(E(N_{rb}^2 \mid \text{last}))$

$= E(N_{rb}^2 \mid r)P(r) + E(N_{rb}^2 \mid b)P(b)$

$= E((N_{r-1,b}+1)^2)p + E((N_{r,b-1}+1)^2)q$

$= E(N_{r-1,b}^2)p + 2pE(N_{r-1,b})$

$+ E(N_{r,b-1}^2)q + 2qE(N_{r,b-1}) + 1$

Integration Recursion

Ex: $y_n(a) = \int_0^1 \frac{x^n}{x+a}dx$ satisfies:

$y_0(a) = \int_0^1 \frac{1}{x+a}dx = \log\frac{1+a}{a}$

$y_n(a) = \int_0^1 \frac{x^n}{x+a}dx$

$= \int_0^1 \frac{x^{n-1}(x+a-a)}{x+a}dx$

$= \int_0^1 x^{n-1}dx - a\int_0^1 \frac{x^{n-1}}{x+a}dx$

$= \frac{1}{n} - a y_{n-1}(a)$

With $n \uparrow$, $y_n \approx \frac{1}{n} - \frac{a}{(1+a)n}$

$= \frac{1}{n}\left(\frac{1}{1+a}\right)$

weird, always ⊕ function, may have ⊖ values

Ex: [illegible], Bernoulli(p_i)

W.T.F. $S_n(k) = P(S_n = \sum X_i = k)$

if $p_i = p_j$ ⇒ binomial(p)

if $p_i \ne$, $\sum p_i$ moderate ⇒ Poisson($\sum p_i$)

In general: $q_1(0) = 1 - p_1$, $q_1(1) = p_1$

$q_j(0) = (1-p_j)q_{j-1}(0)$

$q_j(i) = p_j q_{j-1}(i-1) + (1-p_j)q_{j-1}(i)$, $1 \le i \le j-1$

$q_j(j) = p_j q_{j-1}(j-1)$

Power Series Expansions

mgf: $M_X(s) = E(e^{sX})$

$= \sum_{k=0}^{\infty}\frac{m_k}{k!}s^k$

cgf: $K_X(s) = \log(M_X(s))$

$= \sum_{k=0}^{\infty}\frac{c_k}{k!}s^k$

Ex: Standard Normal Distribution

$F(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^x e^{-y^2/2}dy$

$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^x \sum_{n=0}^{\infty}\frac{(-1)^n y^{2n}}{2^n n!}dy$

$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{(2n+1)2^n n!}$

Alternating sequence → roundoff error

Let $g(x) = e^{x^2/2}\int_0^x e^{-y^2/2}dy = \sum_{n=0}^{\infty} c_n x^{2n+1}$

$g(x)$ satisfies $g'(x) = xg(x) + 1$

Thus $c_n = \frac{1}{2n+1}c_{n-1}$ $\forall c_i > 0$, $g(x)$ converges

Since $g(0) = -g(0)$ odd func., only $x > 0$ considered

$F(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}e^{-x^2/2}g(x) = \sum_{n=0}^{\infty} a_n \frac{x}{2}$

with $a_0 = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $a_n = a_{n-1}\frac{x^2}{2n+1}$

Moments of iid sum

X_i iid with kth moment $X \sim F$

Let $S_n = \sum X_i$

moments $b_{nk} = \frac{1}{n}\sum_{j=0}^{k-1}\binom{k}{j}[n(k-j) - j]\,a_{k-j}\,b_{n,j}$

Ex: $X \sim U(0,1)$, $m_k = \frac{1}{k+1}$

Ex: Quick sort

pivot $\to x_1$

M_i: Candidate position after exami[ning]

$M_1 = 1$: for $i = 2 \sim n$,

if $x_i < x_1$, swap x_{M_i+1} and x_i, set $M_{i+1} \to M_i + 1$

Insert pivot at M_n.

Let e_n = # operations

$= (n-1) + \frac{2}{n}\sum_{i=1}^{n} e_{i-1}$

$\lim_{n\to\infty} e_n / 2n\log(n) = 1$

$E(N_n \mid M_n) = (n-1) + E(N_{M_n-1} \mid M_n)$

$+ E(N_{n-M_n} \mid M_n)$

Assume $P(M_n = i) = \frac{1}{n}$ $\forall i = 1 \sim n$

$e_n = E(N_n) = \sum E(N_n \mid M_n = i)P(M_n = i)$

$= \frac{1}{n}\sum E(N_n \mid M_n = i)$

$= \frac{1}{n}\sum[(n-1) + e_{i-1} + e_{n-i}]$

$\frac{e_n}{n+1} = \frac{2(n-1)}{n(n+1)} + \frac{e_{n-1}}{n} = 2\sum\frac{k-1}{k(k+1)}$

$= 2\sum\frac{k+1-2}{k(k+1)} = 2\sum\frac{1}{k} - 4\sum\frac{1}{k(k+1)}$

$\approx 2\log(n) - 4 \approx 2\log n$

Network: random matrix A

Let $A(i,j) \sim$ Bernoulli($P(ij)$) indep

$A_{ii} = 0$, $A_{ji} = A_{ij}$, $= A \sim F$

Degree of node: # node connects

$d_i = \sum_{j=1}^{n} A(i,j) = \sum_j A(j,i)$ for $i = 1 \sim n$

Poisson's Binomial distribution

Exponentials

Let $P(s)$ power series, $Q(s) = e^{P(s)}$

$Q'(s) = P'(s)Q(s)$. Equating the coefficients of s^{k-1}, get:

$q_k = \frac{1}{k}\sum_{j=0}^{k-1}(k-j)p_{k-j}q_j$

With $q_0 = e^{p_0}$. If $\frac{p_k}{k!} \ldots$ $\frac{P_k^s}{k!} = p_k$

$\Rightarrow q_k^s = \sum_{j=0}^{k-1}\binom{k-1}{j}p_{k-j}^s q_j^s$

If $Q(s) = \sum_{k=0}^{\infty}\frac{m_k}{k!}s^k = e^{P(s)}$ & MGF then $P(s)$ corresponding cumulant GF

⇒ recurrence: $m_0 = 1$, $c_0 = 0$

$m_k = \sum_{j=0}^{k-1}\binom{k-1}{j}c_{k-j}m_j$

$c_k = m_k - \sum_{j=1}^{k-1}\binom{k-1}{j}c_{k-j}m_j$

Random Sum of iid var

$S_N = X_1 + \dots + X_N$ compounded poisson

$N \sim$ Poisson(λ), N random

MGF $X_i = R(s)$, MGF $S_N = e^{-\lambda + \lambda R(s)}$

Probgf $X_i = R(s)$, Pgf $S_N = e^{-\lambda + \lambda R(s)}$

Ex: Incomplete Gamma Function

$P(a, bx) = \frac{1}{\Gamma(a)}\int_0^x b^a y^{a-1}e^{-by}dy$

$= \frac{1}{\Gamma(a)}\int_0^{bx} z^{a-1}e^{-z}dz$

We can expand $P(a,x)$ in a power series by repeated integration by parts.

$P(a,x) = \frac{1}{\Gamma(a)}\int_0^x y^{a-1}e^{-y}dy$

$= \frac{1}{\Gamma(a)}(-x^{a-1}e^{-x} + (a-1))\int_0^x y^{a-2}e^{-y}dy$

$= \frac{-e^{-x}x^{a-1}}{\Gamma(a)} + \frac{(a-1)\Gamma(a-1)}{\Gamma(a)}P(a-1, x)$

$= \frac{-e^{-x}x^{a-1}}{\Gamma(a)} + P(a-1, x)$

$= \sum_{n=1} \frac{-e^{-x}x^{a-n}}{\Gamma(a-n-1)}$

W.T.F. X_i then $\hat\beta$

Incomplete Beta Function

$I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x y^{a-1}(1-y)^{b-1}dy$

$= x^a(1-x)^b\sum_{n=0}^{\infty} c_n x^n$

$= \sum_{n=0}^{\infty} c_n x^{n+a}(1-x)^b$

$\frac{d}{dx}I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$

$= \sum_{n=0}^{\infty} c_n[(n+a)(1-x) - bx]$

$\cdot x^{n+a-1}(1-x)^{b-1}$

$\Rightarrow \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \sum_{n=0}^{\infty} c_n[(n+a)(1-x) - bx]x^n$

$\Rightarrow c_0 = \frac{\Gamma(a+b)}{a\Gamma(a)\Gamma(b)} = \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)}$

$\Rightarrow c_n = \frac{n-1+a+b}{n+a}c_{n-1}$ for $n > 0$

⇒ thus $c_n > 0$ $\forall c_n$

Distribution Connections

$X_n^2 = $ Gamma$(\frac{n}{2}, \frac{1}{2})$ / $Pr(X_n^2 \le x) = P(\frac{n}{2}, \frac{x}{2})$

$Poi(\lambda) = $ inc-Gamma / $Pr(X \le k-1) = 1 - P(k, \lambda)$

$Bin(^n_p) = $ inc-beta / $Pr(X \le k-1) = 1 - I_p(k, n-k+1)$

$F_{mn} = \frac{nX_m^2}{mX_n^2}$ / $Pr(F \le x) = Pr(\frac{X_m^2}{X_m^2 + X_n^2} \le \frac{mx}{mx+n})$

• if $P = n/2 \in \mathbb{Z}$, $X_n^2/2 = W_p \sim$ Gamma

• if $q = m/2 \in \mathbb{Z}$, $X_m^2/2 = W_q \sim$ Gamma

$\Rightarrow \frac{X_n^2}{X_m^2 + X_n^2} = \frac{W_p}{W_p + W_q} \ge u$

$\Rightarrow Pr(\frac{W_p}{W_q + W_p} \ge u) = \sum_{j=0}^{q-1}\binom{p+q-1}{j}u^j(1-u)^{p+q-1-j}$

$= 1 - I_u(p, p+q-1-p+1)$

$= I_{1-u}(q, p)$

$Pr(F_{mn} \le x) = I_{\frac{mx}{mx+n}}(\frac{m}{2}, \frac{n}{2})$

$Pr(t_n \le x) = \frac{1}{2} + \frac{1}{2}Pr(t_n^2 \le x^2)$

$= \frac{1}{2} + \frac{1}{2}Pr(F_{1,n} \le x^2)$

for $x \ge 0$ $= \frac{1}{2} + \frac{1}{2}I_{\frac{x^2}{n+x^2}}(\frac{1}{2}, \frac{n}{2})$

Monotone Transformations

$X \sim F(x)$, $h(x)$ strictly ↑, continuous.

$Pr(h(X) \le x) = F(h^{-1}(x))$

$h(x)$ strictly ↓, continuous

$Pr(h(X) < x) = Pr(X > h^{-1}(x)) = 1 - F(h^{-1}(x))$

• X normal, e^X log-normal

• $X \sim X^2$, $1/X$ inverse X^2, $1/\sqrt{X}$ inverse X

$\ln X$ log X^2

• $X \sim F_{mn}$, $\frac{1}{2}\ln X \sim$ Fisher Z.

Asymptotic Expansions

$f(x) = \frac{x^2+1}{x+1}$ ← resembles x as $x \uparrow \infty$

$= \frac{x^2+1}{x(1+1/x)} = (x + \frac{1}{x})\sum_{k=0}^{\infty}(\frac{-1}{x})^k$ resembles $x - 1$ as $x \uparrow \infty$

$= x - 1 - 2\sum_{k=0}^{\infty}(\frac{-1}{x})^k$

Finite Taylor Expansion

$f(x) = \sum_{k=0}^{n}\frac{f^{(k)}}{k!}(0)x^k + O(x^{n+1})$

$= \sum_{k=0}^{n}\frac{1}{k!}f^{(k)}(0)x^k + \frac{x^{n+1}}{n!}\int_0^1 \frac{f^{(n+1)}(tx)}{}(1-t)^n dt$

$R_n(x) = -\frac{x^n}{n!}f^{(n)}(0) + R_{n-1}(x)$

$R_0(x) = f(x) - f(0)$

Property of vector norm:

$\exists k_U, k_L$ s.t.

$k_L\|x\|_1 \le \|x\| \le k_U\|x\|_2$ $\forall x \in \mathbb{R}^n$

⇒ $\forall$ norms, $\sup_{x \ne 0}\frac{\|x\|}{\|x\|_*} < \infty$

⇒ if $\|x_n - x\| \to 0$ then $\|x_n - x\|_* \to 0$ } rate can be different

Matrix norms

5. $\|AXB\| \le \|A\| \times \|B\|$ (induced norm)

Ex: $\|A\| = \sup_{X \ne 0}\frac{\|AX\|}{\|X\|} = \sup_{X: \|X\|=1}\|AX\|$

Frobenius Norm

$\|A\|_F = \sqrt{\sum_i\sum_j A^2(i,j)} = \sqrt{\mathrm{trace}(AA^T)}$

Not an induced norm, since $\|I\|_F \ne 1$

Cauchy-Schwarz:

$|\langle u, v\rangle|^2 \le \|u\|^2\|v\|^2$

$= \langle u, u\rangle\langle v, v\rangle$

$(\sum u_i v_i)^2 \le (\sum u_i^2)(\sum v_i^2)$

Induced matrix norm

1. $\|I\| = 1$ $\forall \|\cdot\|$ induced
2. $\|A\|_\infty = \max_i \sum_{j=1}^{m}|A(i,j)|$ max absolute row sum
3. $\|A\|_1 = \max_j \sum_{i=1}^{m}|A(i,j)|$ max absolute column sum
4. $\|A\|_2 = \sqrt{P(A^TA)}$ with $P(M)$ "spectral radius" largest (absolute) eigenvalue $|\lambda_m|$
5. $\forall \|\cdot\|$ induced, $P(A) \le \|A\|$

Matrix Norm

$A_{m \times n}$, $B_{n \times p}$, $X \in \mathbb{R}^n$

$Ax = O(mn)$, $AB = O(mnp)$

$A_{n \times n} = I_{n \times n} + u(v^T)$, $u, v, x \in \mathbb{R}^n$

$Ax = Ix + u(v'x)$

Linear Regression

$\hat\beta = \arg\min_b (Y - Xb)'(Y - Xb)$

normal equation: $X'X\hat\beta = X'Y$

$\hat\beta = (X'X)^{-1}X'Y$

Vector norm

1. $\|X\| \ge 0$ $\forall X \in \mathbb{R}^m$
2. $\|X\| = 0$ iff $X = 0$
3. $\|cX\| = |c|\|X\|$ $\forall X \in \mathbb{R}^m$, $c \in \mathbb{R}$
4. $\|X + Y\| \le \|X\| + \|Y\|$

Ex: $\|X\|_1 = \sum_{i=1}^{m}|X_i|$ $\|X\|_2 = \sqrt{\sum_1^m X_i^2}$

Iterative Solution of linear equations

$(I - B)X = W \Rightarrow X_{n+1} = BX_n + W$ converges if $\|B\| < 1$

The power method:

Markov Chain $(X_0, X_1, \dots)$

With transition matrix P:

$Prob(X_{t+1} = S_j \mid X_t = S_i) = P(i,j)$

W.T.F. equilibrium dist f: $P^Tf = f$

To solve: start X_0: $X_{n+1} = P'X_n$

Condition Number

If $Ax = b$ & $A(x + \Delta x) = b + \Delta b$ ⇒

$\|b\| = \|Ax\| = \frac{\|Ax\|}{\|x\|}\cdot\|x\| \le \|A\|\cdot\|x\|$ &

$A\Delta x = \Delta b \Rightarrow \Delta x = A^{-1}\Delta b$

$\Rightarrow \|\Delta x\| = \|A^{-1}\Delta b\| \le \|A^{-1}\|\cdot\|\Delta b\|$

$\|b\|\cdot\|\Delta x\| \le \|A\|\cdot\|x\|\cdot\|A^{-1}\|\cdot\|\Delta b\|$

$\|\Delta x\|/\|x\| \le (\|A\|\cdot\|A^{-1}\|)\|\Delta b\|/\|b\|$

$= \mathrm{cond}(A)\|\Delta b\|/\|b\|$

governs the stability of the soln.

Condition Number: Symmetric

$\mathrm{cond}_2(A) = \|A\|_2\cdot\|A^{-1}\|_2$

recall $\|A\|_2 = \sqrt{P(A^TA)} = P(A)$ if A symmetric

$\Rightarrow \mathrm{cond}_2(A) = \lambda_1/\lambda_n$ with $\lambda_1 \ge \dots \ge \lambda_n$

A non-singular

Condition Number: linear Regression

$Y = X\beta + \varepsilon \Rightarrow (X'X)\hat\beta = X'y \Rightarrow AX = b$

$\mathrm{cond}_2(A) = \lambda_1/\lambda_n$ singular if $X'X < 2$

Orthogonal Transformations

O orthogonal, $O'O = OO' = I$

$(Ou)'Ov = u'O'Ov = u'v$

① $\|OA\|_2 = \|AO\|_2 = \|A\|_2$

② $\|OA\|_F = \|AO\|_F = \|A\|_F$

Jacobi's Method

Solve $Ax = b$

$A = (A_{ij})$ satisfies $|A_{ii}| > \sum_{j \ne i}|A_{ij}|$ $\forall i$

$D = \mathrm{diag}(A(i))$, $B = I - D^{-1}A$

converges if $\|B\|_\infty = \max_i \sum_{j \ne i}\frac{|A_{ij}|}{|A_{ii}|} < 1$

Landweber's iteration scheme $\forall \varepsilon > 0$

$B = I - \varepsilon A'A$ → $A'A$ P.D.

$\Rightarrow \|I - \varepsilon A'A\|_2 < 1$ eigenvalues: $0 < \varepsilon$

Matrix Inversion

The sweep operator: on A symmetric

Sweeping on k^{th} diagonal entry:

$\hat A = (\hat a_{ij})$ symmetric:

$\hat a_{kk} = -\frac{1}{a_{kk}}$, $\hat a_{ik} = \frac{a_{ik}}{a_{kk}}$, $\hat a_{kj} = \frac{a_{kj}}{a_{kk}}$

$\hat a_{ij} = a_{ij} - a_{ik}a_{kj}/a_{kk}$ $\forall i \ne k, j \ne k$

The Reverse Sweep:

$\hat a_{kk} = -\frac{1}{a_{kk}}$, $\hat a_{ik} = -\frac{a_{ik}}{a_{kk}}$, $\hat a_{kj} = -\frac{a_{kj}}{a_{kk}}$

$\hat a_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}$

Property 1: $A \in \mathbb{R}^{m \times m}$, $U, V \in \mathbb{R}^{m \times n}$

$V = UA \Rightarrow \hat V = \hat U\hat A$ after sweep a_{kk}

• $\hat U = U$ except U_k, V_k swap

• $\hat V = V$ except U_k, V_k swap

⇒ 3 sweep on $k = 1$, inverse sweep on k

⇒ inverse sweep undo a sweep

Property 2: $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{pmatrix}$ symmetric, sweep on diag of A_{11}

$\hat A = \begin{pmatrix} -A_{11}^{-1} & A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12}\end{pmatrix}$

Property 3: A is P.D iff each diagonal entry can be swept in succession and is positive until it is swept.

Sweep in linear reg: $\hat A = O(m^3)$

$A = \begin{pmatrix} X'X & X'y \\ y'X & y'y\end{pmatrix}$ ⇒ sweep on diagonal of $X'X$

$\begin{pmatrix} -(X'X)^{-1} & (X'X)^{-1}X'y \\ y'X(X'X)^{-1} & y'y - y'X(X'X)^{-1}X'y\end{pmatrix} = \begin{pmatrix} -\mathrm{Var}(\hat\beta)/\sigma^2 & \hat\beta \\ \hat\beta' & \|y - \hat y\|_2^2\end{pmatrix}$

Gaussian elimination:

Solve $AX = b$, W.T.F P s.t. $PA = LU/LU$

Find $M_1M_1A = U$, let $P = M_1M_1$

Solve for $Ux = Pb$ by backward elimination

Cholesky Decomposition

$A_{m \times m}$ is P.D. ⇒ unique L s.t. $A = LL'$ (L lower triangular)

$(X,Y)'(X,Y) = \begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y\end{pmatrix} = \begin{pmatrix} L & 0 \\ l' & d\end{pmatrix}\begin{pmatrix} L' & l \\ 0 & d\end{pmatrix}$

$\Rightarrow Ll = X'y$, $L'\beta = l \Rightarrow (X'X)\beta = X'y$

$\Rightarrow d^2 = Y'Y - l'l$

$= Y'Y - Y'X(X'X)^{-1}X'Y = \|y - \hat y\|_2^2$

Gram-Schmidt Orthogonalization

By Q.R decomposition: $X_{n \times p} = Q_{n \times p}R_{p \times p}$

Q has orthonormal cols, $Q'Q = I$

R $p \times p$ invertible, upper-triangular

Define $U_1 = X_1/\|X_1\|_2$, inductively

$V_k = X_k - \sum_{j=1}^{k-1}(U_j'X_k)U_j$; $U_k = V_k/\|V_k\|_2$

Then $(U_1 \dots U_p)$ forms Q.

$r_{jk} = U_j'X_k$ for $i \le j \le k$, $r_{kk} = \|V_k\|_2$

$\Rightarrow X = QR \Rightarrow X'X = R'Q'QR = R'R$

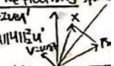
$\& X'y = R'Q'y$, solve by backward

$\Rightarrow R\hat\beta = R'Q'y \Leftrightarrow R\hat\beta = Q'y$

Householder Reflections

Householder: $H = I - 2uu'$

$H'H = I - 4uu' + 4u\|u\|^2u' = I$



H orthogonal, symmetric

$Hx = x$ if $x \in$ plane ⊥ u

$\Rightarrow H_{n-1}\cdots H_1X = OX = \begin{pmatrix} R \\ O\end{pmatrix}$

$\Rightarrow O_{n \times n}$ orthogonal, $R_{p \times p}$ upper tri

⇒ Construct QR decomp of X

Sherman-Morrison Formula

$A_{m \times m}$ invertible, u, v vectors

$A + uv'$ invertible, iff $1 + v'A^{-1}u \ne 0$

$\Rightarrow (A + uv')^{-1} = A^{-1} - \frac{1}{1 + v'A^{-1}u}A^{-1}uv'A^{-1}$

Woodbury's Formula

$A_{m \times m}$ invertible, $U, V_{m \times d}$, rank of $\ldots$

$\Rightarrow (A + UV')^{-1} = A^{-1} - A^{-1}U(I_p + V'A^{-1}U)^{-1}V'A^{-1}$

To solve the $(X'X)\hat\beta = X'y$

Sweep:	$O(np^2 + p^3)$
Gaussian Elimination:	$O(np^2 + p^3)$
Cholesky Decomposition:	$O(np^2 + p^3)$
Householder Orthogonal:	$O(np^2 - p^3)$
Modified Gram-Schmidt:	$O(np^2)$

Eigen computations (SVD/PCA)

Spectral clustering:
Input: Adjacency A, # clusters k
Output: cluster assignment $\{C\}$
1. Compute D with node degrees from A: $D_{ii}=\sum_j A_{ij}$
2. Compute Laplacian $L=D^{-1/2}AD^{-1/2}$
3. Compute λ for L, construct X $n\times k$ eigenvectors w/ top eigenvalues
4. Perform K-means over rows of X

Epidemics on networks:
pathogen spreads: epidemic threshold
β (infection rate), μ (recovery rate)
Let A be adjacency, $\lambda(A)$ spectral radius
Disease dies: $\beta/\mu < 1/\lambda(A)$
Pandemic: $\beta/\mu > 1/\lambda(A)$

Jacobi's method
W.T.F Λ for symmetric Ω
- Gradually transform Ω to diagonal
- Each transform contains rotation matrix $R=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}$, reduces off-diagonal gradually
- A,B transformed diagonal, then $B=U'AU$ with $\|B\|_F^2=\|A\|_F^2$, $tr(A)=tr(B)$
- Rotation effect row k, col l, e.g. 1, 2: $U=\begin{pmatrix}R & 0\\ 0' & I_{n-2}\end{pmatrix}$, $B_{ii}=(U'AU)_{ii}=a_{ii}$ if $i>2$
- $b_{11}=a_{11}\cos^2\theta-2a_{12}\cos\theta\sin\theta+a_{22}\sin^2\theta$
- $b_{12}=(a_{11}-a_{22})\cos\theta\sin\theta+a_{12}(\cos^2\theta-\sin^2\theta)$
- $b_{22}=a_{11}\sin^2\theta+2a_{12}\cos\theta\sin\theta+a_{22}\cos^2\theta$
- let θ s.t. $\tan(2\theta)=\frac{2a_{12}}{a_{22}-a_{11}}$ if $a_{22}-a_{11}\neq 0$
- let $\theta=\pi/4$ otherwise
$\Rightarrow$ ensures $b_{12}=0$, $S(B)>S(A)$
$S(\cdot)$ sum of squares of the diagonal entries

Rayleigh Quotient $R(x)=\frac{x'Ax}{x'x}$
Suppose A has $\lambda_1<\dots<\lambda_m$
$R(x)\le\lambda_m \Rightarrow R(V_m)=\lambda_m$

Courant–Fischer Theorem
$\lambda_k=\min_{V_k}\max_{x\in V_k, x\neq 0}R(x)$ if V_k subspace $\mathbb{R}^m$
$=\max_{V_{m-k+1}}\min_{x\in V_{m-k+1}}R(x)$
min by $V_k=\mathrm{span}\{u_1\cdots u_k\}$
max by $V\ldots=\mathrm{span}\{u_k\cdots u_m\}$

Eigenvalue Perturbation
Let $A,B=A+\Delta A$ symmetric $m\times m$
$\lambda_1\le\dots\le\lambda_m$, $\mu_1\le\dots\le\mu_m$
then $|\lambda_k-\mu_k|\le\|\Delta A\|_2\ \forall k\in\{1,\dots,m\}$

Power Iteration Method
Goal: Find λ_m, V_m s.t. $|\lambda_1|\le|\lambda_2|<\dots<|\lambda_m|$, diagonalizable $A_{m\times m}$
1. Start with initial $u^{(0)}$
2. Recurrence/iteration: $u^{(n)}=\frac{Au^{(n-1)}}{\|Au^{(n-1)}\|}$ for $n=0,1,\dots\to V_m$
3. $\lambda_m=R_A(V_m)$

Observe:
1. if $|\lambda_{m-1}|\approx|\lambda_m|$ slow, rate depends $\frac{|\lambda_{m-1}|}{|\lambda_m|}$
2. $\lambda_m\approx 0$, slow or D.N. converge
3. $\lambda<0$, $u^{(2n)}$ converges

Question:
1. λ_1? Power method with A^{-1}
2. λ_k? Use $(A-\mu I)^{-1}$ with $\mu\approx\lambda_k$
$\Rightarrow(A-\mu I)^{-1}V_i=\frac{1}{\lambda_i-\mu}V_i$

Gershgorin's circle theorem
Every λ satisfies at least one $|\lambda-a_{ii}|\le\sum_{j\neq i}|a_{ij}|$
If A symmetric, w.T.F λ_m, $\lambda_m\approx M=\max_{1\le i\le m}\{a_{ii}+\sum_{j\neq i}|a_{ij}|\}$

Add-on Jacobi's result:
- Similarity transformation: if λ, x pairs for A then λ, Bx pairs for BAB^{-1}
- Gershgorin circle theorem: λ_i for A lies in the union of circles $\bigcup_{i=1}^n D_i$, where: $D_i=\{s:|s-A_{ii}|\le\sum_{j\neq i}|A_{ij}|\}$

Disadvantages of classical approach:
need search through all off-diagonal entries of current matrix
threshold method: only rotate when a_{ij} is sufficiently large in absolute value

Singular Value Decomposition (SVD)
$\forall A\in\mathbb{R}^{n\times m}$ with rank k, $A=\sum_{i=1}^k\sigma_i u_i v_i'=U\Sigma V'$
$u_1\cdots u_k$, $v_1\cdots v_k$ orthonormal, $\sigma_i>0$
$\Rightarrow\|A\|_2=\|\Sigma\|_2=\max\sigma_i$
$\|A\|_F=\|\Sigma\|_F=\sqrt{\sum_i\sigma_i^2}$

low-rank approximation
A has full $U\Sigma V'$ with σ_{ii} descending order. Best rank-k approx of A in F norm is:
$B=\sum_{j=1}^k\sigma_j u_j v_j'$
$\Rightarrow\|A-B\|_F=\sqrt{\sum_{i>k}\sigma_i^2}$; $\|A-B\|_2=\sigma_{k+1}$

SVD through Jacobi:
- $A'A\to$ symbolize w/ orthogonal columns, normalize non-trivial columns
- $AV=U\Sigma\Rightarrow A=U\Sigma V'$ by Jacobi

Ridge Regression:
- $f_\lambda(\beta)=\|y-X\beta\|_2^2+\lambda\|\beta\|_2^2$
$=(y-X\beta)'(y-X\beta)+\lambda\beta'\beta$
- $\nabla f_\lambda(\beta)=-2X'(y-X\beta)+2\lambda\beta$
- $\nabla f_\lambda(\beta)=0\Rightarrow\hat\beta=(X'X+\lambda I)^{-1}X'y$
if $X=\sum_j\sigma_j u_j v_j'$ SVD, then
$X'y=\sum_j\sigma_j v_j u_j' y$, $X'X+\lambda I=\sum_j(\sigma_j^2+\lambda)v_j v_j'$
$\Rightarrow\hat\beta=\sum_j\frac{\sigma_j}{\sigma_j^2+\lambda}u_j'y\,v_j$
$\hat y=\sum_j\frac{\sigma_j^2}{\sigma_j^2+\lambda}u_j'y\,u_j$

HW Problems

HW1 Prob2: Prove $\binom{n}{k}$ integer
Base: $\binom{0}{0}=1$, $\binom{n}{0}=1$
Induction: Suppose $\binom{n-1}{k}\ \forall 0\le k\le n-1$
then $\binom{n}{k}=\binom{n-1}{k}+\binom{n-1}{k-1}$
$=$ integer

Prob3: Find $C(x)=\left|\frac{x f'(x)}{f(x)}\right|$
2). $f(x)=e^x$, $C(x)=|x|$
3). $f(x)=\log x$, $C(x)=|1/\log x|$
... $f(x)=\log(1+x)$, $C(x)=\left|\frac{x}{(1+x)\log(1+x)}\right|$
... $f(x)=\sqrt{1+x}$, $C(x)=\left|\frac{x}{2(1+x)}\right|$

Prob4: V_i, $|x_i-\bar x|\le s\sqrt{n-1}$
$s^2(n-1)=\frac{n-1}{n}\sum(x_i-\bar x)^2$
$=\frac{n-1}{n}(\sum_{j\neq i}(x_j-\bar x)^2+(x_i-\bar x)^2)$
$=\frac{1}{n}\sum_{j\neq i}(x_j-\bar x)^2\sum 1^2+\frac{n-1}{n}(x_i-\bar x)^2$
Cauchy $\ge\frac{1}{n}(\sum_{j\neq i}x_j-\bar x)^2+\frac{n-1}{n}(x_i-\bar x)^2$
$=\frac{1}{n}(-(x_i-\bar x))^2+\frac{n-1}{n}(x_i-\bar x)^2$
$=(x_i-\bar x)^2$
$\Rightarrow s\sqrt{n-1}\ge|x_i-\bar x|$

HW2 Prob5: Tri-diag. alg. with $O(n)$
$E(N_{add})=\min(m,b/q)$
$E(N_{com})=\min(m,r/p)$
$E(N_{stat})=E(E(N_{stat}|\text{last}))$
$=E(N_{stat(last)})+E(N_{stat})$
$+1$

Prob6: Quicksort, randomized
Suppose pivot the smallest.
Step 1: $n-1$ comparison
Step 2: $n-2$ comparison
Total: $\sum_{i=1}^{n-1}i=\frac{(n-1)n}{2}=O(n^2)$

Prob7: Power Series
Gia pmf: $P_n(\theta)=C_n\theta^n/g(\theta)$
with $g(\theta)=\sum_{k=0}^\infty C_k\theta^k$
$M(\theta)=\sum nC_n\theta^n/g(\theta)$
$=\theta/g(\theta)\cdot\sum nC_n\theta^{n-1}$
$=\theta g'(\theta)/g(\theta)$
$E(n^2)=\sum n^2C_n\theta^n/g(\theta)$
$=\sum(n-1)nC_n\theta^n/g(\theta)$
$+nC_n\theta^n/g(\theta)$
$=\frac{\theta}{g(\theta)}(\theta g''(\theta)+g'(\theta))$
$\sigma^2(\theta)=E(n^2)-M(\theta)^2$
$=\frac{\theta}{g(\theta)}(\theta g''(\theta)+g'(\theta))-\frac{\theta^2g'(\theta)^2}{g(\theta)^2}$
$=\theta M'(\theta)$
Let $X_i\sim P_n(\theta)$, $S_m=X_1+\dots+X_m$
$P(X_i=n_i)=C_{n_i}\theta^{n_i}/g(\theta)$
$P(S_m=n)=P(X_1+\dots+X_m=n)$
$=\sum\prod_{i=1}^m C_{n_i}\theta^{n_i}/g(\theta)$
$=\sum(\prod C_{n_i}\theta^{\sum n_i}/g(\theta)^m)$
$=\sum\prod C_{n_i}\cdot\theta^n/g(\theta)^m$
$=a_{mn}\cdot\theta^n/g(\theta)^m$
$E(a_{m,S_m}/a_{m-1,S_m})$
$\sum\frac{a_{m,n}}{a_{m-1,n}}\cdot\frac{a_{m-1,n}\theta^n}{g(\theta)^m}=\theta\sum\frac{a_{m,n}\theta^{n-1}}{g(\theta)^m}$

HW3 Prob1: complexity
$A_{n\times m}$, $B_{m\times m}$, $X\in\mathbb{R}^n$
$(AB)x=O(n^3)+O(n^2)$
$A(Bx)=O(n^2)+O(n^2)$

Prob2: $\|A\|^{-1}\le\|A^{-1}\|\ \forall\|\cdot\|$
$\|I\|=\|A^{-1}A\|\le\|A^{-1}\|\|A\|$
$\|A\|^{-1}\|I\|\le\|A^{-1}\|\|A\|\|A\|^{-1}$
$\|A\|^{-1}\|I\|\le\|A^{-1}\|$
See $\|I\|\le\|I\|\|I\|=\|I\|^2$, $\|I\|\ge 1$

Prob3: Check $\|A\|=\max_{i,j}|A_{ij}|$
(1) $\|A\|\ge 0\ \forall A$ ✓
(2) $\|A\|=0$ iff $A=0$ ✓
(3) $\|cA\|=|c|\|A\|$ ✓
(4) $\|A+B\|\le\|A\|+\|B\|$ ✓
(5) $\|AB\|\le\|A\|\|B\|$ ✗
$A=\begin{pmatrix}1&1\\1&1\end{pmatrix}$ $B=\begin{pmatrix}1&1\\1&1\end{pmatrix}$ $AB=\begin{pmatrix}2&2\\2&2\end{pmatrix}$

Prob4: induced norm inequality
$\frac{1}{\sqrt m}\|A\|_1\le\|A\|_2\le\sqrt m\|A\|_1$
$\|A\|_2=\sup_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}$
$=\sup_{\|x\|_2=1}(\sum_{i=1}^m|(Ax)_i|^2)^{1/2}$
Cauchy Schwarz $\ge\sup_{\|x\|_2=1}\frac{1}{\sqrt m}(\sum_{i=1}^m|Ax|_i)$
$=\frac{1}{\sqrt m}\sup\sum_{i=1}^m|Ax|_i$
$=\frac{1}{\sqrt m}\|A\|_1$
$\|A\|_2=\sup\|Ax\|_2$
$\le\sup\|Ax\|_1$
$\le\sup\|A\|\|x\|_1$
$\le\sup\|A\|_1\sqrt{(\sum|x_i|)^2}$
$\le\sup\|A\|_1\|x\|_2\sqrt m$
$=\|A\|_1\sqrt m$
Since $\forall x\in\mathbb{R}^n$, $\sum x_i^2\le(\sum x_i)^2\le(\sum|x_i|)$ from ①

Exam 1 Practice

Prob 1: Recall Quicksort, let P be the prob that a randomly chosen permutation of n distinct numbers contains at least one pre-existing splitter. Proof use inclusion-exclusion argument.

Ans 1:
Let $A_m=\{$ordering: m-th entry correct$\}$
$W=|\bigcup_{m=1}^n A_m|=$ # of permutations s.t. at least one entry correct.
Apply inclusion-exclusion:
$W=\sum_{m=1}^n|A_m|-\sum_{1\le m_1<m_2\le n}|A_{m_1}\cap A_{m_2}|+\dots+(-1)^{p-1}\sum_{1\le m_1<\dots<m_p\le n}|A_{m_1}\cap\dots\cap A_{m_p}|+\dots$
Each $A_{m_1}\cap\dots\cap A_{m_p}$ represents the set of permutations that at least p values: $m_1,\dots,m_p$ are in the correct position. Since only # matters, exact position doesn't:
$W=\binom{n}{1}|A_1|-\binom{n}{2}|A_1\cap A_2|+\dots+(-1)^{p-1}\binom{n}{p}|A_1\cap\dots\cap A_p|+\dots$
permutations have p correct = # of ordering the rest $(n-p)$
$W=\binom{n}{1}(n-1)!-\binom{n}{2}(n-2)!+\dots+(-1)^{p-1}\binom{n}{p}(n-p)!+\dots$
$=\sum_{i=1}^n\binom{n}{i}(n-i)!(-1)^{i-1}$ $\quad P=\frac{\#\text{ at least one}}{\#\text{ all permutation}}$
$=n!\sum_{i=1}^n\frac{(-1)^{i-1}}{i!}$ $\quad =W/n!$
$=1-e^{-1}$ □

Prob 2: Let $X\in\mathbb{R}^m$ s.t. $\|x\|_2=1$, and $Q=I-2XX'$. Show that $Q=Q'=Q^{-1}$. What are the eigenvalues of Q?
Ans 2:
$Q'=I'-2XX'=Q$ symmetric
$QQ'=II'-I'2XX'-I2XX'+4XXXX'$
$=I-4\|X\|_2^2+4\|X\|_2^4$
$=I$
$QQ'=Q'Q=I\Rightarrow Q'=Q^{-1}$ orthogonal
Given $Q=Q^{-1}$:
$QQV=\lambda QV$, $\lambda V=QV$
$V=\lambda QV$ | thus $\frac{1}{\lambda}=\lambda$, $\lambda=\pm1$
$\frac{1}{\lambda}V=QV$
Q is reflection matrix: $Q'=Q^{-1}$; $\lambda=\pm1$
x is an eigenvector of Q
$\lambda x=Qx\Rightarrow\lambda=-1$
$V\perp x$ is an eigenvector of $Q\Rightarrow\lambda=1$
λ: -1 (multiplicity 1); 1 (multiplicity $m-1$)

Prob 3: Let A & $W\in\mathbb{R}^{m\times m}$, W orthogonal. Prove A & WA have same $\mathrm{cond}_2(\cdot)$.
Ans 3: Recall $\mathrm{cond}_2(A)=\sqrt{\rho(A'A)}=\mathrm{Sh}(A)$
$\mathrm{cond}_2(WA)=\sqrt{\rho(A'W'WA)}=\sqrt{\rho(A'A)}$

Prob 4: Prove A & WA same F
Ans 4: $\|A\|_F=\sqrt{tr(AA')}$
$\|WA\|_F=\sqrt{tr(WAAW')}$
$=\sqrt{tr(AAWW')}$
$=\sqrt{tr(AA')}$
$A=U\Sigma V'$
$WA=(WU)\Sigma V'$
$\sigma(WA)=\sigma(A)$
$\mathrm{cond}(WA)=\frac{\sigma_{max}(WA)}{\sigma_{min}(WA)}$
$=\frac{\sigma_{max}(A)}{\sigma_{min}(A)}=\mathrm{cond}(A)$

Prob 5: If A p.d. nonsingular, proof A^{-1} p.d.
Ans 5: A.P.D. also symmetric,
$A=Q\Lambda Q'\Rightarrow x'Q\Lambda Q'x\ge 0\ \forall x$
$\Rightarrow x'Q\Lambda^{-1}Q'x>0\ \forall x$
$\Rightarrow(Q'x)'\Lambda^{-1}(Q'x)>0$
Since $\Lambda^{-1}>0$, quadratic form

Prob 6: Prove Gerschgorin's Circle theorem
Ans 6: $Av=\lambda v\Rightarrow\sum_j A_{ij}V_j-\lambda V_i=0$
Find some i s.t. largest $|x_i|$ — $|x_i|\ge|x_j|\ \forall j$
$A_{i1}V_1+A_{i2}V_2+\dots+A_{in}V_n-\lambda V_i=0$
$A_{i1}V_1+\dots+(A_{ii}-\lambda)V_i+\dots+A_{in}V_n=0$
$\sum_{j\neq i}A_{ij}V_j=(\lambda-a_{ii})V_i$
$|\sum_{j\neq i}A_{ij}V_j|=|(\lambda-a_{ii})V_i|$
$\sum_{j\neq i}|A_{ij}||V_j|\ge|(\lambda-a_{ii})V_i|=|\lambda-a_{ii}||V_i|$
by triangular inequality
$\sum_{j\neq i}|A_{ij}||V_j|/|V_i|\ge|\lambda-a_{ii}|$
$\sum_{j\neq i}|A_{ij}|\ge|\lambda-a_{ii}|$

Prob 7: Let $u\in\mathbb{R}^n$ and consider the Householder transformation defined as $H(u)=I-uu'/T$ with $T=\|u\|_2$. Prove that $H(u)z=z\ \forall z\in\mathbb{R}^n$ that $z\perp u$.
Let $x=z+u'xu$, where $z\perp u$ and $u'x$ is component of x in direction of u. Then $H(u)x=z-u'xu$.
Ans 7: $H(u)z=(I-uu'/T)z$
$=z-uu'z/T$
$=z$
$H(u)x=(I-uu'/T)x$
$=z+u'xu-uu'u'xu/T$
$=z+u'xu-\frac{2uu'u'xu}{u'u}$
constant $=z+u'xu-(u'x)\frac{2uu'u}{u'u}$
$=z+u'xu-2u'xu$
$=z-u'xu$

(2) $\frac{1}{\sqrt m}\|A\|_\infty\le\|A\|_2\le\sqrt m\|A\|_\infty$
Recall $\|A\|_\infty=\|A'\|_1$, $\|A\|_2=\|A'\|_2$
Thus from (1), proved.
$\|A\|_2=\sqrt{\rho(A'A)}=\sqrt{\rho(AA')}=\|A'\|_2$
$\frac{1}{\sqrt m}\|A\|_F\le\|A\|_2\le\|A\|_F$
$\|A\|_F=\sqrt{trace(AA')}$
$=\sqrt{tr(A'A)}$
$=\sqrt{\sum_{i=1}^m\lambda_i}$ since $A'A$ symmetric diagonalizable
$\le\sqrt{m\cdot\lambda_{max}}$
$=\sqrt m\|A\|_2$
$\le\sqrt m\sqrt{\sum\lambda_i}$
$=\sqrt m\|A\|_F$

Prob5: Condition number
$\|(A+\Delta A)^{-1}-A^{-1}\|=\|A^{-1}(I+\Delta A A^{-1})^{-1}\Delta A A^{-1}\|$
Woodbury $=\|A^{-1}(I+\Delta A A^{-1})^{-1}\Delta A A^{-1}\|$
$=\|(A+\Delta A)^{-1}\Delta A A^{-1}\|$
$\frac{\|(A+\Delta A)^{-1}-A^{-1}\|}{\|(A+\Delta A)^{-1}\|}\le\|(A+\Delta A)^{-1}\|\|\Delta A\|\|A^{-1}\|$
$\le\|\Delta A\|\|A^{-1}\|$
$=\mathrm{cond}_2(A)\|\Delta A\|/\|A\|$

Prob6: Prove $\rho(A)=\lim\|A^n\|^{1/n}\ \forall\|\cdot\|$
$\forall\varepsilon>0$, $A^*=\frac{A}{\rho(A)+\varepsilon}$, $A^{*n}\to 0$
Then $\rho(A^*)=\rho(A)/(\rho(A)+\varepsilon)<1$
Thus $\exists N$ s.t. $\forall n>N$, $\|A^{*n}\|<1$
Symmetric, $A^*=A/\rho$, $\|A^n\|\le\rho(A)+\varepsilon$
Thus, $\|A^n\|^{1/n}\ge\dots$

HW4 Prob1: P.D. matrix
Assume P.D. $B=A^{-1}$
By Schur complement partition
$b_{ii}=(A^{-1})_{ii}$
$=(a_{ii}-A_{i,-i}A_{-i,-i}^{-1}A_{-i,i})^{-1}$
- $A_{i,-i}$ row vector of off-diagonal elements in the i-th row
- A_{-i} the principal submatrix of A obtained by remove i row
- $A_{i,-i}A_{-i,-i}^{-1}A_{-i,i}\ge 0$ since A_{-i} P.D. quadratic
- Thus $b_{ii}=a_{ii}^{-1}$ iff $A_{i,-i}=0$
$\Rightarrow A_{ij}=A_{ji}=0\ \forall j\neq i$

HW5 Prob2
W.T.S. $\sum_j|A_{ij}|<1\Rightarrow|\lambda|<1$
By Gershgorin's Circle,
$|\lambda-a_{ii}|\le\sum_{j\neq i}|a_{ij}|$
$|\lambda|\le|\lambda-a_{ii}|+|a_{ii}|$
$\le\sum_{j\neq i}|a_{ij}|+|a_{ii}|$
<1
W.T.S. $\sum_j|A_{ij}|<|A_{ii}|\Rightarrow A$ singular
A non-singular iff $\det(A)\neq 0$
Suppose $\lambda=0$,
$Au=0u$, for i s.t. $|u_i|\ge|u_j|\ \forall j$
$\sum_j A_{ij}u_j=0$, $|u_j|/|u_i|\le 1$
$A_{ii}u_i=-\sum_{j\neq i}A_{ij}u_j$
$A_{ii}=-\sum A_{ij}u_j/u_i$
$|A_{ii}|\le\sum|A_{ij}||u_j|/|u_i|$
$|A_{ii}|\le\sum_{j\neq i}|A_{ij}|$ Contradict

Prob3: Orthogonal
O has real entries only if $\lambda=\pm1$
$Ov=\lambda v$
$v'O'Ov=v'\lambda'\lambda v$
$v'v=\lambda'\lambda v'v$
Thus $\lambda'\lambda=1$, $|\lambda|^2=1$
If v has real entries, Ov has real entries, thus λ need to be real.
Thus $\lambda^2=1$, $\lambda=\pm1$.

Prob3: matrix sweep
$\begin{pmatrix}x'x & x' & x'y\\ x & I_p & y\\ y'x & y' & y'y\end{pmatrix}$ sweep on $x'x$
$(x'x)^{-1}$ $(x'x)^{-1}x'$ $(x'x)^{-1}x'y$
$x(x'x)^{-1}$ $I_p-x(x'x)^{-1}x'$
$y'x(x'x)^{-1}$ $y'-y'x(x'x)^{-1}x'$
$y-x(x'x)^{-1}x'y$
$y'y-y'x(x'x)^{-1}x'y$

Prob4: $\|A\|_F^2\le\mathrm{rank}(A)\|A\|_2^2$
$\|A\|_F^2=trace(AA')$
$=\sum\lambda_i$
$\le\sum|\lambda_i|$
$\le m\cdot|\lambda_{max}|$
$=\mathrm{rank}(A)\|A\|_2^2$
$\le\mathrm{rank}(A)^2\|A\|_2^2$

Prob5: A symmetric, U orthogonal
$\|U'AU\|_F^2=trace((U'AU)(U'AU)')$
$=trace(UU'AA')$
$=trace(AA')$
$=\|A\|_F^2$
$trace(U(AU))=trace(AUU')$
$=trace(A)$

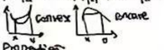
Prob6: $\|A\|_*=\max_{R,S}tr(R'AS)$
$=\max_{R,S}tr(R'AS)$
$f(A)=\sum_{i=1}^{\min(m,n)}a_{ii}$
(a) $\|A\|_*=\|U\Sigma V'\|_*$
$=\max_{R,S}tr(R'U\Sigma V'S)$
$\ge tr(U'U\Sigma V'V)$
$=tr(\Sigma)=\sum\sigma_i$
(b) $\|A\|_*=0$ iff $A=0$
(c) $\|cA\|_*=|c|\|A\|_*$
(d) $\|A+B\|_*\le\|A\|_*+\|B\|_*$
(e) $\|UAV\|_*=\|A\|_*$
(f) $\|A\|_*=\|\Sigma\|_*$
$=\max_{R,S}\sum_{i=1}^{\min(m,n)}(R'\Sigma S)_{ii}$
$=\max_{R,S}\sum_i\sum_j\sigma_j R_{ji}S_{ji}$
$\le\sum_j\sigma_j\max_{R,S}\sum_i S_{ji}R_{ji}$
$\le\sum\sigma_j$
(g) Form (f) & (e), $\|A\|_*=\sum\sigma_i$
$\|A\|_*=tr(\sqrt{A'A})$
$=tr(\sqrt{AA'})$
$=\|A'\|_*$
(h) $\|AB\|_*=\max_{R,S}f(RABS)$
$=\max_{R,S}f(RUDV'EZS)$
$=\max_{R,S}f(RUD\tilde U'\tilde E\tilde S)$
$=\max_{R,S}f(TDOEX)$
$=\max_{R,S}\sum\sum\sum t_{ij}d_j o_{jk}e_k x_{ki}$
Since $|O_{jk}|\le 1$: $\le\max_{R,S}\sum\sum\sum t_{ij}d_j e_k x_{ki}$
$\le\sum_j d_j\sum_k e_k\sum t_{ij}x_{ki}$
$\le\sum_j d_j\sum_k e_k|\sum t_{ij}x_{ki}|$
$\|A\|_*\|B\|_*\Rightarrow\le\sum_j d_j\sum e_k$

Univariate Convex Optimization

Unconstrained Minimization:
Solves $\min_{x \in \mathbb{R}^n} f(x) = f(x^*) = p^*$
With $f: \mathbb{R}^n \to \mathbb{R}$ convex and twice differentiable.

Convex f: $\forall x, y,\ \theta \in (0,1)$
$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y)$
Strictly convex if: $<$

Jensen's inequality:
for $\sum_i \theta_i = 1$, $f(\sum_i \theta_i x_i) \le \sum_i \theta_i f(x_i)$



Properties:
1) convex: local min $\to$ global min
Strict convex: unique global min
2) Convex + differentiable:
$f(y) \ge f(x) + \nabla f(x)'(y-x)$
strictly convex: $>$

Remark: For a convex function, the first-order Taylor approximate is a global underestimator of the function.
Optimal: x^* optimal iff $\nabla f(x^*) = 0$

Proof: Suppose x^* local minimum
- Then $f(x^*) \le f(x)\ \forall x$ s.t. $\|x - x^*\| < \varepsilon$
- $\forall x^{00} \in \mathbb{R}$,
- Find θ s.t. $x = x^* + \theta(x^{00} - x^*)$
- Then $f(x^*) \le f(x) = f((1-\theta)x^* + \theta x^{00})$
since linear $\Rightarrow$ $= (1-\theta) f(x^*) + \theta f(x^{00})$

Proof: $\forall t \in (0,1)$,
$f(x + t(y-x)) \le (1-t) f(x) + t f(y)$
$\Rightarrow f(y) \ge f(x) + \frac{f(x+t(y-x)) - f(x)}{t}$
$\Rightarrow$ As $t \to 0$, $f(y) \ge f(x) + \nabla f(x)'(y-x)$

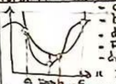
Ex: Show $g(t) = f(ty + (1-t)x)$ convex
Fix $t_1, t_2 \in (0,1)$, $\theta \in (0,1)$
$z_1 = t_1 y + (1-t_1)x$, then $g(t_1) = f(z_1)$
$g(\theta t_1 + (1-\theta)t_2) = f(\Delta y + (1-\Delta)x)$
$= f(\theta z_1 + (1-\theta) z_2)$
$\le \theta f(z_1) + (1-\theta) f(z_2)$
$= \theta g(t_1) + (1-\theta) g(t_2)$
Then, we have $g(t_0) \ge g(t_1) + g'(t_1)(t_0 - t_1)$
Now, put $t_0 = 0$, $t_1 = 1$, $g(1) \ge g(0) + g'(0)$
$\Rightarrow f(y) \ge f(x) + g'(0)$
$= f(x) + \nabla f(x)'(y-x)$

Section Search: Ex $f(x) = \sum_{i=1}^{n} |x - i|$
Known f not differentiable
f unimodal
$x^* \in (x_{min}, x_{max})$

Algorithm: Univariate Section Search
Result: $x^* = \frac{1}{2}(x_L + x_U)$
Initialization: x_L, x_U known, $x^0 \leftarrow \mathrm{Unif}(x_L, x_U)$
while not converged, do:
  $x_{new} \leftarrow \mathrm{Unif}(x_L, x_U)$;
  if $f(x_{new}) > f(x^0)$: $x^0 \leftarrow x_{new}$, end
  if $x_{new} < x^0$ then $x_L \leftarrow x_{new}$, end
  else $x_U \leftarrow x_{new}$, end.
end

Algorithm: Golden Section Search
Result: same
Initialization: x_L, x_U, $x^0 \leftarrow x_U - \phi(x_U - x_L)$, $\phi = \frac{\sqrt{5}-1}{2}$
while not converged, do:
  $x_{new} \leftarrow x_L + \phi(x_U - x_L)$
  ...
end
efficiency ↑
time ↓

Parabolic Interpolation



- $a = x_L$; $c = x_U$
- b: $f(b) < f(a), f(b) < f(c)$
- draw parabola thru a, b, c
- Find x_{min}
- discard one of a, c, keep b
$x_{min} = x_{min}(f, a, b, c)$
$x_{min} = b + \frac{1}{2} \frac{(f(a)-f(b))(c-b)^2 - (f(c)-f(b))(b-a)^2}{(f(a)-f(b))(c-b) + (f(c)-f(b))(b-a)}$

Adaptive Parabolic Interpolation
Result: $x^* = x_{min}$
Initialization: x_L, x_U, $x^0 \leftarrow \frac{1}{2}(x_L + x_U)$
while not converged, do:
  $x_{new} \leftarrow x_{min}(f, x_L, x_U, x^0)$
  ...
end
time ↓
iteration ↓

Brent's method: (no derivative)
when PL stalls, switch to Golden Sec. Search

Newton's Method:
f twice differentiable,
f can be approximated by the minimum of its quadratic Taylor series approx.
For $x \approx x_0$,
$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \equiv g(x)$

Algorithm: Newton's method
Result: x^*
Initialization: $x^{(0)}$
while not converged do
  $x^{(k+1)} \leftarrow x^{(k)} - f'(x^{(k)}) / f''(x^{(k)})$;
  $x^0 = x^{(k+1)}$
end
fastest
Closely near f

Newton's method solves $f'(x^*) = 0$
Newton root-find solves $f(x^*) = 0$
$\Rightarrow x^{(k+1)} \leftarrow x^{(k)} - f(x^{(k)}) / f'(x^{(k)})$

Descent Methods

Algorithm: General Descent
- choose initial $x^{(0)} \in \mathbb{R}^n$
- iteratively update:
  1. descent direction Δx
  2. line search: step size $t > 0$
  3. Update: $x \leftarrow x + t\Delta x$
- Stop when criterion satisfied
① How to choose descent direction
② How to choose step size.

Line Search
Exact line search: let $t = \arg\min_{s \ge 0} f(x + s\Delta x)$
Since $f(x)$ convex, $\phi(s) = f(x + s\Delta x)$ is also convex
- Use last class tool to find t.

Backtracking line search:
- Given descent direction Δx, $\alpha \in (0, 0.5)$, $\beta \in (0,1)$, set: $t = 1$
- while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)' \Delta x$
  $t = \beta t$
- Stopping criterion:
$f(x + t\Delta x) \le f(x) + \alpha t \nabla f(x)' \Delta x$ holds
- Δx is descent direction, $\nabla f(x)'\Delta x < 0$, so small t: $f(x + t\Delta x) \approx f(x) + t\nabla f(x)'\Delta x < f(x) + \alpha t \nabla f(x)'\Delta x$
- α = fraction of the decrease in f predicted by linear extrapolation

Gradient Descent
- Since convex, $\nabla f(x^{(k)})'(y - x^{(k)}) \ge 0$ implies $f(y) \ge f(x^{(k)})$. So $\Delta x^{(k)}$ must satisfy $\nabla f(x^{(k)})' \Delta x^{(k)} < 0$
- When $\Delta x = -\nabla f(x)$, gradient descent.

Algorithm:
- Choose initial $x^{(0)} \in \mathbb{R}^n$
- Iteratively update:
  - $\Delta x := -\nabla f(x)$
  - line search: step size t
  - update: $x \leftarrow x + t\Delta x$
- Stop when stopping criterion satisfied.

Theorem: $x^{(k+1)} \leftarrow x^{(k)} - t^{(k)} \nabla f(x^{(k)})$, $\Delta x = -\nabla f(x)$
Define $\tilde{f}: \mathbb{R} \to \mathbb{R}$, $\tilde{f}(t) = f(x - t\nabla f(x))$
Assume strong convexity, $mI \preceq \nabla^2 f(x) \preceq MI$,
$f(y) \le f(x) + \nabla f(x)'(y - x) + \frac{M}{2}\|y - x\|_2^2$
$\Rightarrow \tilde{f}(t) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$
Exact line search:
$g'(t) = -\|\nabla f(x)\|_2^2 + 2Mt\|\nabla f(x)\|_2^2 \equiv g(t)$
$g'(t) = 0 \Rightarrow t = 1/M$
$\tilde{f}(1/M) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$
$\tilde{f}(t_{exact}) \le \tilde{f}(1/M) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$
$f(x^+) - p^* \le (f(x) - p^*) - \frac{1}{2M}\|\nabla f(x)\|_2^2$
$f(x^+) - p^* \le (1 - m/M)(f(x) - p^*)$
$\Rightarrow f(x^{(k)}) - p^* \le (1 - m/M)^k (f(x^{(0)}) - p^*)$
To achieve ε-approx, $f(x^{(k)}) - p^* \le \varepsilon$
Hence $k = \frac{\log((f(x^{(0)}) - p^*)/\varepsilon)}{\log(1/(1 - m/M))}$
No guarantee to reach global minimum.
But backtrack line search does.

Strong Convexity:
Mathematically: if $\forall x, y$, f is strong convex with $m > 0$
$(\nabla f(x) - \nabla f(y))'(x - y) \ge m\|x - y\|_2^2$
$\equiv \langle \nabla f(x) - \nabla f(y), x - y \rangle \ge m\|x - y\|_2^2$
$\equiv f(y) \ge f(x) + \nabla f(x)'(y - x) + \frac{m}{2}\|y - x\|_2^2$
If differentiable, for $\theta \in (0,1)$
$\equiv f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y) - \frac{1}{2} m\theta(1-\theta)\|x - y\|_2^2$
Strong convexity $\to$ strictly convexity
Strict convexity ↛ strong convexity

For twice continuous differentiable f:
- $f: \mathbb{R} \to \mathbb{R}$:
  - convex iff $f''(x) \ge 0\ \forall x$
  - strictly convex if $f''(x) > 0\ \forall x$
  - strongly convex iff $f''(x) \ge m > 0\ \forall x$
- $f: \mathbb{R}^n \to \mathbb{R}$, twice differentiable
  - strongly convex iff $\nabla^2 f(x) \succeq mI$, $\nabla^2 f(x) \preceq MI$
  - eigenvalue of $\nabla^2 f(x) \ge m\ \forall x$
  - $f(y) = f(x) + \nabla f(x)'(y - x) + \frac{1}{2}(y - x)'\nabla^2 f(z)(y - x)$
  $\ge f(x) + \nabla f(x)'(y - x) + \frac{m}{2}\|y - x\|_2^2$

Convergence of Gradient Descent

Steps:
$f(y) \ge f(x) + \nabla f(x)'(y - x) + \frac{m}{2}\|y - x\|_2^2$ ①
$p^* \ge f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$ ②
$\|x - x^*\|_2 \le \frac{2}{m}\|\nabla f(x)\|_2$ ③
$f(y) \le f(x) + \nabla f(x)'(y - x) + \frac{M}{2}\|y - x\|_2^2$ ④
$f(y) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$

Steepest Descent

Recall $f(x+v) \approx f(x) + \nabla f(x)'v$
$\nabla f(x)'v$ = directional derivative of f at x in the direction v.
- gives approx change in f for a small step v
- v is a descent direction if $\nabla f(x)'v < 0$

Q: How to find v for smallest $\nabla f(x)'v$?
$\Delta x_{nsd} = \arg\min_{\|v\| \le 1} \{\nabla f(x)'v\}$ normalized
$= \arg\min_{\|v\| = 1} \{\nabla f(x)'v\}$

Recall dual norm: $\|x\|_* = \sup_{\|v\| \le 1} x'v$
$\nabla f(x)'\Delta x_{nsd} = \min_{\|v\| \le 1} \{\nabla f(x)'v\}$
$= -\sup_{\|v\| \le 1} x'v$
$= -\|\nabla f(x)\|_*$
$\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$
$\nabla f(x)'\Delta x_{sd} = \|\nabla f(x)\|_* \nabla f(x)'\Delta x_{nsd}$
$= -\|\nabla f(x)\|_*^2$

Algorithm: Steepest Descent
Initialization: $x^{(0)} \in \mathbb{R}^n$
while not converged do:
  compute $\nabla f(x^{(k)})$
  compute $\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$
  $g^{(k)}(t) \leftarrow f(x^{(k)} + t\Delta x_{sd}^{(k)})$
  line search: $t^* = \arg\min_{t \ge 0} g^{(k)}(t)$
  $x^{(k+1)} = x^{(k)} + t^* \Delta x_{sd}^{(k)}$
end

Example: $\Delta x_{nsd} = \arg\min_{\|v\|_2 = 1} \nabla f(x)'v$
$\Rightarrow \nabla f(x)'v \ge -\|\nabla f(x)\|_2 \|v\|_2$ Cauchy-Schwarz
By setting, $v = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \Delta x_{nsd}$, $\|\nabla f(x)\|_2$
For l_2-norm, steepest descent results equals gradient descent result.

Coordinate Descent
In each iteration, we update $x^{(k)}$ to $x^{(k+1)}$ by changing a single coordinate.

Algorithm: Coordinate Descent
Initialization: $x^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$,
while not converged do
  Fix i;
  $g_i^{(k)}(y) = f(x_1^{(k)}, \dots, x_n^{(k)})$
  $y^* = \arg\min_{y \in \mathbb{R}} g_i^{(k)}(y)$ univariate optim
  $x^{(k+1)} = (x_1^{(k)}, \dots, y^*, \dots, x_n^{(k)})$
end
Cyclical coordinate descent: cycle through coordinates following permutation of $\{1, 2, \dots, n\}$
- works well when multivariate fun smooth
- works bad when multivariate fun non-smooth

For l_1-norm, steepest descent result equals coordinate descent for the coordinate on which $\nabla f(x^{(k)})$ has its largest entry

Theorem: let $f: \mathbb{R}^n \to \mathbb{R}$ be differentiable. The steepest direction v at a point $x \in \mathbb{R}^n$ under the l_1 norm, is a vector satisfies
$v_i = -\mathrm{sgn}(\nabla f(x)_i)$
for some index i where $|\nabla f(x)_i|$ is max and $v_j = 0$ for all $j \ne i$.
Proof:
$\Delta x_{nsd} = \arg\min_{\|v\|_1 = 1} \nabla f(x)'v$
$\nabla f(x)'v = \sum_{i=1}^n \nabla f(x)_i v_i \ge -\sum_{i=1}^n |\nabla f(x)_i||v_i|$
$\ge -\max_{1 \le i \le n} |\nabla f(x)_i|$
Since $|v_i| \le 1\ \forall i$ ($\|v\|_1 = \sum_i |v_i| = 1$)
$v_i = \begin{cases} -\mathrm{sgn}(\nabla f(x)_i) & \text{for } i = \arg\max_{1 \le i \le n} |\nabla f(x)_i| \\ 0 & \text{otherwise} \end{cases}$

Newton's Method

The descent direction (Newton Step) is given by $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$
Positive definiteness of $\nabla^2 f(x)$ implies that
$\nabla f(x)'\Delta x_{nt} = -\nabla f(x)'\nabla^2 f(x)^{-1}\nabla f(x) < 0$
unless $\nabla f(x) = 0$, so Newton step is a descent direction (unless $x = x^*$ is optimal)

Algorithm: Newton's Method
Start point $x = x^{(0)}$, tolerance $\varepsilon > 0$
repeat:
1. Compute step + decrement
$\Delta x_{nt} := -\nabla^2 f(x)^{-1}\nabla f(x)$
$\lambda^2 := \nabla f(x)'\nabla^2 f(x)^{-1}\nabla f(x)$
2. Stopping criterion, quit if $\frac{\lambda^2}{2} \le \varepsilon$
3. line search: choose step size t by backtracking line search.
4. Update: $x = x + t\Delta x_{nt}$

① Interpretation
- The 2nd Taylor of f at x is
$\hat{f}(x+v) = f(x) + \nabla f(x)'v + \frac{1}{2}v'\nabla^2 f(x)v$
(convex quadratic of v) is minimized when $v = \Delta x_{nt}$. Newton step minimizes the 2nd order approx of f at x.
- If f is nearly quadratic, then $x + \Delta x_{nt}$ is the exact (approx) minimizer of f.
- As f twice differentiable, $\hat{f}$ is very accurate when x is near x^*. When x is near x^*, the point $x + \Delta x_{nt}$ is a good estimate of x^*

② Interpretation
= If linearize the optimal condition $\nabla f(x^*) = 0$ near x, we obtain
$\nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x)v = 0$
which solved by $v = \Delta x_{nt}$.
= Newton step finds the zero-crossing of a first-order Taylor approx. of f'

③ Interpretation
- Recall Newton is a special case of steepest descent with quadratic norm. $\|u\|_{\nabla^2 f(x)} = (u'\nabla^2 f(x)u)^{\frac{1}{2}}$
- $\|\cdot\|_P$ converges rapidly when the Hessian has a small cond number.
$(\mathrm{cond}(M) = \lambda_1(M)/\lambda_n(M))$
Near x^*, $\nabla^2 f(x) \approx \nabla^2 f(x^*) = P$,
thus Newton step a good choice.

Affine Invariance of Newton Step
- Suppose $T \in \mathbb{R}^{n \times n}$ non-singular, define $\bar{f}(y) = f(Ty)$
$\nabla \bar{f}(y) = T'\nabla f(x)$, $\nabla^2 \bar{f}(y) = T'\nabla^2 f(x)T$, with $x = Ty$
- $\Delta y_{nt} = -(T'\nabla^2 f(x)T)^{-1}(T'\nabla f(x))$
$= -T^{-1}\nabla^2 f(x)^{-1}\nabla f(x)$
$= T^{-1}\Delta x_{nt}$
The Newton steps of f and $\bar{f}$ are related by the same linear transformation, and
$x + \Delta x_{nt} = T(y + \Delta y_{nt})$

Convergence analysis
- ∃ numbers η and γ with $0 < \eta \le m^2/L$, & $\gamma > 0$
Following holds:
- If $\|\nabla f(x^{(k)})\|_2 \ge \eta$, then
$f(x^{(k+1)}) - f(x^{(k)}) \le -\gamma$ { damped Newton phase / linear convergence phase }
- If $\|\nabla f(x^{(k)})\|_2 < \eta$, then backtracking $t^{(k)} = 1$
$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \le \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2$
quadratic convergence phase

Newton Advantages
1. Convergence is rapid in general for Newton Method, and quadratic near x^*. Especially enter the quadratic converge phase.
2. Newton's method is affine invariant.
3. Newton method scales well with problem size.
4. Good performance of Newton doesn't depend on algorithm param choice. But choice of norm for steepest descent matters a lot

Newton Disadvantages
1. Storing Hessian and compute linear equations costs time

Quasi Newton Idea
- Use $f(x^{(k)} + \delta) - f(x^{(k)}) \approx \nabla f(x^{(k)})'\delta + \frac{1}{2}\delta' Q^{(k)}\delta$
where $\delta \in \mathbb{R}^n$ and $\|\delta\|_2$ small, and $Q^{(k)}$ some p.d. matrix
- Choose descent direction as
$\delta = -(Q^{(k)})^{-1}\nabla f(x^{(k)})$
where $Q^{(k)}$ some p.d. matrix
- $Q^{(k)} = I$ gives Gradient Descent
- $Q^{(k)} = \nabla^2 f(x^{(k)})$ gives Newton Method

Quadratic Norms
- Consider quadratic norms, P is p.d.
$\|z\|_P = \sqrt{z'Pz} = \|P^{\frac{1}{2}}z\|_2$
- Dual norm: $\|z\|_* = \|P^{-\frac{1}{2}}z\|_2$
- Steepest direction:
$\Delta x_{sd} = -P^{-1}\nabla f(x)$
- Ex: When $P = \nabla^2 f(x)$, we have:
$\Delta x_{sd} = -\nabla^2 f(x)^{-1}\nabla f(x)$
- Newton's method is the steepest descent method for the quadratic Hessian norm.

Summary of Steepest descent:

norm	Descent Method
l_2	gradient descent
l_1	coordinate (max gradient)
Hessian $\|z\|_P = \sqrt{z'\nabla^2 f(x)z}$	Newton's Method

Quasi Newton Method
- Let $P^{(k)} = (Q^{(k)})^{-1}$ thus descent direction is
$\delta = -P^{(k)}\nabla f(x^{(k)})$
- Start with $P = I$, let $P^{(k+1)} = P^{(k)} + U^{(k)}$
Q: How to update $P^{(k+1)}$?
- SR1: symmetric rank 1 approximation
- DFP: a rank 2 approximation
- BFGS

Algorithm: BFGS
Initial guess x^0, approx $B_0 = \hat{\nabla}^2 f(x_0)$
1. obtain direction $P_k = -B_k \nabla f(x_k)$
2. line search: stepsize α_k in P_k direction
3. Set $S_k = \alpha_k P_k$, update $x_{k+1} = x_k + S_k$
4. $Y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$
5. $B_{k+1} = B_k + \frac{(S_k'Y_k + Y_k'B_kY_k)(S_kS_k')}{(S_k'Y_k)^2} - \frac{B_kY_kS_k' + S_kY_k'B_k}{S_k'Y_k}$

Backtracking Line Search:
$f(t) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$
BLS exit condition:
- General:
$f(x + t\Delta x) \le f(x) + \alpha t \nabla f(x)'\Delta x$
- GD version:
$f(t) \le f(x) + \alpha t\|\nabla f(x)\|_2^2$
$0 \le t \le \frac{1}{M} \Rightarrow -t + \frac{Mt^2}{2} \le -\frac{t}{2}$
$\Rightarrow \tilde{f}(t) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$
$\le f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2$
$\le f(x) - \alpha t\|\nabla f(x)\|_2^2$ when $\alpha < \frac{1}{2}$
Case 1:
$t = 1$, $f(x^+) \le f(x) - \alpha\|\nabla f(x)\|_2^2$
Case 2:
$t \ge \beta/M$, $f(x^+) \le f(x) - \frac{\beta\alpha}{M}\|\nabla f(x)\|_2^2$
$\Rightarrow f(x^+) - p^* \le f(x) - p^* - \min\{\alpha, \frac{\beta\alpha}{M}\}\|\nabla f(x)\|_2^2$
$\Rightarrow f(x^{(k)}) - p^* \le c^k (f(x)^{(0)} - p^*)$ where $c = 1 - \min\{2m\alpha, \frac{2\beta\alpha m}{M}\} < 1$

Convergence Theory
Exact Line Search:
$f(x^+) = \tilde{f}(t_{exact}) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$
$\Rightarrow (f(x^+) - p^*) = (f(x) - p^*) - \frac{1}{2M}\|\nabla f(x)\|_2^2$
$\Rightarrow f(x^+) - p^* \le (1 - \frac{m}{M})(f(x) - p^*)$
From ②, $\|\nabla f(x)\|_2^2 \ge 2m(f(x) - p^*)$
$\Rightarrow f(x^{(k)}) - p^* \le (1 - \frac{m}{M})^k (f(x)^{(0)} - p^*)$
$1 - \frac{m}{M} \to 0$ at rate of $\log(\frac{1}{1 - m/M})$
If $m \ll M$, k needs to be large.

Logistic Regression
$P(Y_i = 1 | X_{i1}, \dots, X_{id}) = P_i$
where $\log(\frac{P_i}{1 - P_i}) = \beta_1 X_{i1} + \dots + \beta_d X_{id}$
$L(\beta) = \prod_{i=1}^n P_i^{Y_i}(1 - P_i)^{1 - Y_i}$
$= \prod_{i=1}^n \left(\frac{\exp x_i'\beta}{1 + \exp x_i'\beta}\right)^{Y_i}\left(\frac{1}{1 + \exp x_i'\beta}\right)^{1 - Y_i}$
$f(\beta) = -\log L(\beta)$ objective function
$= -\sum_{i=1}^n (Y_i x_i'\beta - \log(1 + \exp(x_i'\beta)))$
W.T.F. $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} f(\beta)$

Stochastic Gradient Descent
- Batch training: find $\arg\min_x \frac{1}{n}\sum_{i=1}^n f_i(x)$
Full GD: $x^{(k)} = x^{(k-1)} - t_k \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^{(k-1)})$
- SGD estimates $\frac{1}{n}\sum \nabla f_i(x)$
e.g.: $x^{(k)} = x^{(k-1)} - t_k \nabla f_{i_k}(x^{(k-1)})$
$E(\nabla f_{i_k}(x^{(k-1)})) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^{(k-1)})$
Unbiased, high variance (single point)
- Mini-batch: choose $I_k \subset \{1, \dots, n\}$
$x^{(k)} = x^{(k-1)} - t_k \frac{1}{b}\sum_{i \in I_k} \nabla f_i(x^{(k-1)})$
$|I_k| = b \ll n$ lower variance

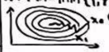
Multivariate Optimization

Choose descent direction:
1. Steepest descent
  a). Coordinate descent: optimize along the coordinate with the largest gradient
  b). Gradient descent: optimize along gradient
  c). Newton's method: Use both gradient, Hessian

Choose Step size:
1. exact line search
2. backtracking line search

Stopping criterion:
1. $\|x^{(k+1)} - x^{(k)}\|_2 < \delta$ for some $\delta > 0$
2. $f(x^{(k+1)}) - f(x^{(k)}) > -\delta$ for some $\delta > 0$
3. $\|\nabla f(x^{(k)})\|_2 < \delta$ for some $\delta > 0$
4. A prefixed # of iterations
5. A prefixed computation time

Example: quadratic problem in $\mathbb{R}^2$
$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2)$ satisfies strong convexity with $m = \min\{1, \gamma\}$; $M = \max\{1, \gamma\}$.



Example: nonquadratic problem
$f(x) = e^{x_1 + 3x_2 - 0.1} + e^{x_1 - 3x_2 - 0.1} + e^{-x_1 - 0.1}$
iter ↓
time ↑

- initial value defines the moving direction
- derivative + Hessian defines the moving direction

HW Problems

HW8 Prob 1:

Define g subgradient of f at x if $\forall y$

$f(y) \geq f(x) + g(x)'(y-x)$

Suppose g is a subgradient of $f(x) = \sum_{i=1}^{r} |x_i|$

Let $g_i(x)$ be the i-th coordinate of $g(x)$.

Prove $g_i(0) \in [-1, 1]$ and $g_i(x) = \text{sign}(x_i)$ when $x \neq 0$ for $i = 1, \cdots, p$.

$g_i(x) = \nabla_{x_i} f(x)$

$= \nabla_{x_i} \sum_{i=1}^{r} |x_i|$

$= \begin{cases} \frac{\partial}{\partial x_i} -x_i & \text{if } x_i < 0 \\ \frac{\partial}{\partial x_i} x_i & \text{if } x_i > 0 \end{cases} \Rightarrow \begin{cases} -1 \\ 1 \end{cases}$

Thus $g_i(x) = \text{Sgn}(x_i)$ when $x \neq 0$

When $x = 0$: [illegible]

$f(y) \geq f(x) + g(x)'(y-x)$

$\sum_{i=1}^{p} |y_i| - \sum_{i=1}^{p} |x_i| \geq (g_1(x) \cdots g_p(x)) \begin{pmatrix} y_1 - x_1 \\ \vdots \\ y_p - x_p \end{pmatrix}$

$\sum_{i=1}^{p} |y_i| \geq \sum_{i=1}^{p} g_i(0) y_i$

$|y_i| \geq g_i(0) y_i$

$g_i(0) \in [-1, 1]$

c). Recall $f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$

$\lim_{x \to a^+} \frac{f(x) - f(a)}{x - a} \leq \lim_{x \to a^+} \frac{f(b) - f(a)}{b - a}$

$\Rightarrow f'(a) \leq \frac{f(b) - f(a)}{b - a}$

Similarly $x \to b^-$

$f'(b) \geq \frac{f(b) - f(a)}{b - a}$

d). Result from c implies f non-decreasing

$f''(x_0) = \lim_{h \to 0} \frac{f'(x_0 + h) - f'(x_0)}{h}$

At $x_0 = a$:

$f'(a+h) - f'(a) \geq 0, \ h > 0$

$\Rightarrow \frac{f'(a+h) - f'(a)}{h} \geq 0$

$\Rightarrow \lim_{h \to 0^+} \frac{f'(a+h) - f'(a)}{h} \geq 0$

$\Rightarrow f''(a) \geq 0$

Similarly at $x_0 = b$

$\lim_{h \to 0^-} \frac{f'(b+h) - f'(b)}{h} \geq 0, \ h < 0$

$\Rightarrow f''(b) \geq 0$

Prob 5:

$x^* = \text{argmin} f(x), \ f(x) = \sum_{i=1}^{m} |x_i - x|$

a). Prove $x^* = \text{median}\{x_1, \cdots, x_m\}$

b). Prove convexity, f continuous but not differentiable.

a). Subgradient of $f(x)$

$f'(x) = \sum_{i=1}^{m} \text{sgn}(x - x_i)$

b). $x_1, x_2 \in \mathbb{R}, \ \theta \in (0, 1)$

$f(\theta x_1 + (1-\theta) x_2) = \sum |x_i - (\theta x_1 + (1-\theta) x_2)|$

$\leq \sum |\theta x_i - \theta x_1| +$

$|(1-\theta) x_i - (1-\theta) x_2|$

$= \theta \sum |x_i - x_1| +$

$(1-\theta) \sum |x_i - x_2|$

$= \theta f(x_1) + (1-\theta) f(x_2)$

- $f(x)$ continuous as it's sum of continuous function
- $f(x)$ NOT differentiable at $x = x_i$ "kink" at 0.

Prob 2:

Suppose f is strongly convex with $mI \preceq \nabla^2 f(x) \preceq MI$

Let Δx be a descent direction at x. Show the backtracking stopping condition holds for $0 < t \leq -\frac{\nabla f(x)' \Delta x}{M \|\Delta x\|_2^2}$

Use this to give an upper bound on t.

Recall strongly convexity gives:

$f(y) \leq f(x) + \nabla f(x)'(y-x) + \frac{M}{2} \|y - x\|_2^2$

$f(y) - f(x) \leq \nabla f(x)'(y-x) + \frac{M}{2} \|y - x\|_2^2$

$f(x + t\Delta x) - f(x) \leq \nabla f(x)' t \Delta x + \frac{M}{2} \|t \Delta x\|_2^2$

BLS exists when

$\nabla f(x)' t \Delta x + \frac{M}{2} \|t \Delta x\|_2^2 \leq \alpha t \nabla f(x)' \Delta x$

$\Rightarrow 0 < t \leq 2(\alpha - 1) \frac{\nabla f(x)' \Delta x}{M \|\Delta x\|_2^2}$

$0 < t \leq -\frac{\nabla f(x)' \Delta x}{M \|\Delta x\|_2^2} \quad \alpha \in (0, 0.5]$

From lecture,

$f(x + t \Delta x) \leq f(x) - \frac{[illegible]}{[illegible]} \nabla f(x)' \Delta x$

$\Rightarrow f(x^+) - p^* \leq f(x) - p^* - (\cdot)$

$\Rightarrow f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*)$

$c = 1 - \frac{2m \nabla f(x)' \Delta x}{M \|\Delta x\|_2^2}$

To achieve ε-approx, $f(x^{(k)}) - p^* \leq \varepsilon$

Need $k = \frac{\log((f(x^{(0)}) - p^*)/\varepsilon)}{\log(1/c)}$

HW6 Prob 1:

Let $f: \mathbb{R}^n \to \mathbb{R}$ be a (vector) norm. Prove f convex.

$\forall x, y \in \mathbb{R}^n, \ \theta \in (0, 1)$

$\|\theta x + (1-\theta) y\| \leq \|\theta x\| + \|(1-\theta) y\|$

$= |\theta| \|x\| + |(1-\theta)| \|y\|$

$= \theta \|x\| + (1-\theta) \|y\|$

Prob 2:

Let $f: \mathbb{R}^n \to \mathbb{R}$ be $f(x_1 \cdots x_n) = \max(x_1 \cdots x_n)$

Prove f is convex.

$\forall x, y \in \mathbb{R}^n, \ \theta \in (0, 1)$

$f(\theta x + (1-\theta) y) = \max(\theta x_1 + (1-\theta) y_1, \cdots \theta x_n + (1-\theta) y_n)$

$\leq \theta x_{max} + (1-\theta) y_{max}$

$= \theta f(x) + (1-\theta) f(y)$

Prob 3:

Prove that the Kullback-Leibler cross-entropy distance $f(x) = x_1 \ln\left(\frac{x_1}{x_2}\right) + x_2 - x_1$ is convex on $x_1 > 0, x_2 > 0$

$Hf(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{x_1} & -\frac{1}{x_2} \\ -\frac{1}{x_2} & \frac{x_1}{x_2^2} \end{pmatrix}$

$\det(Hf(x)) = \frac{1}{x_1} \frac{x_1}{x_2^2} - \left(\frac{1}{x_2}\right)^2$

$= 0$

With first leading principle minor $\frac{1}{x_1} > 0$, $Hf(x)$ is P.S.D, thus $f(x)$ is convex.

Prob 4:

Suppose $f: \mathbb{R} \to \mathbb{R}$ convex, $a, b \in \mathbb{R}$ with $a < b$

a) Show $\forall x \in (a, b)$

$f(x) \leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$

b) Show $\forall x \in (a, b)$

$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}$

c). Suppose f differentiable, show

$f'(a) \leq \frac{f(b) - f(a)}{b - a} \leq f'(b)$

d). Suppose f twice differentiable, show

$f''(a) \geq 0 \quad f''(b) \geq 0$

a). f is convex, $a < b$, $a \leq x \leq b$

$f(x) = f\left(\frac{b-a}{b-a} x\right)$

$= f\left(\frac{bx - ax}{b-a}\right)$

$= f\left(\frac{bx - ax + ab - ab}{b-a}\right)$

$= f\left(\frac{ab - ax}{b-a} + \frac{bx - ab}{b-a}\right)$

$= f\left(\frac{b-x}{b-a} a + \frac{x-a}{b-a} b\right)$

$\leq \frac{b-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$

b). From (a), $f(x) \leq (\cdot)$

$f(x) - f(a) \leq \frac{a-x}{b-a} f(a) + \frac{x-a}{b-a} f(b)$

$\frac{f(x) - f(a)}{x - a} \leq \frac{f(b) - f(a)}{b - a}$ ①

$-\frac{b-x}{b-a} f(a) \leq \frac{x-a}{b-a} f(b) - f(x)$

$\frac{b-x}{b-a} (f(b) - f(a)) \leq f(b) - f(x)$

$\frac{f(b) - f(a)}{b - a} \leq \frac{f(b) - f(x)}{b - x}$ ②

Prob 4:

Show $f(\beta) = -\log L(\beta)$ convex where $L(\beta)$ is the likelihood of a logistic regression model.

$f(\beta) = -\sum_{i=1}^{n} (y_i x_i' \beta - \log(1 + \exp(x_i' \beta)))$

$\frac{\partial f(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} (\cdot)$

$= -\sum_{i=1}^{n} y_i x_i - \frac{x_i \exp(x_i' \beta)}{1 + \exp(x_i' \beta)}$

$\frac{\partial^2 f(\beta)}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \frac{x_i \exp(x_i' \beta)(\cdot) - x_i \exp(x_i' \beta) [illegible]}{(1 + (\cdot))^2}$

$= \sum_{i=1}^{n} \frac{x_i^2 \exp(x_i' \beta)}{(1 + (\cdot))^2}$

≥ 0

Thus convex.

Prob 5:

$\|\cdot\|$ is a norm on $\mathbb{R}^n$. $\|z\|_* = \sup\{z'x \mid \|x\| \leq 1\}$

a). Prove dual norm is a valid norm

b). Find l_2 norm $\|x\|_2 = \sqrt{\sum x_i^2} = [illegible]$

c). Dual of l_1 is l_∞, dual of l_∞ is l_1

$\|x\|_1 = \sum |x_i|$; $\|x\|_\infty = \max |x_i|$

d). Dual of l_p is l_q, $\frac{1}{p} + \frac{1}{q} = 1$

$q = p/(p-1)$

a). $\|z\|_* = \sup\{z'x \mid \|x\| \leq 1\}$

$\geq z'0$

$= 0$

$\|z\|_* = 0$

$\Rightarrow z'x = 0 \quad \forall x = \|x\| \leq 1$

$\Rightarrow x'z = 0$

$\Rightarrow z = 0$ let $x = e_i$

$\|cz\|_* = \sup\{cz'x \mid \|x\| \leq 1\}$

$= |c| \sup\{z'x \mid \|x\| \leq 1\}$

$= |c| \|z\|_*$

$\|z + y\|_* = \sup\{z'x + y'x \mid \|x\| \leq 1\}$

$\leq \sup\{z'x \mid \cdot\} + \sup\{y'x \mid \cdot\}$

$= \|z\|_* + \|y\|_*$

b). $\|z\|_* = \sup\{z'x \mid \|x\| \leq 1\}$

$= \sup\{z'x \mid x'x \leq 1\}$

think $z'x = |\langle z, x \rangle|$

$\leq \|z\|_2 \|x\|_2$ by Cauchy Schwarz

$= \|z\|_2$

Equality achieved when $x = \frac{z}{\|z\|_2}$

Thus, $|\langle z, x \rangle|$

$= |\langle z, \frac{z}{\|z\|_2} \rangle|$

$= \|z\|_2$

c). $\|z\|_* = \sup\{z'x \mid \|x\|_\infty \leq 1\}$

$= \sup\{z'x \mid \max|x_i| \leq 1\}$

$= \sup_{\max|x_i| \leq 1} \left|\sum_{i=1}^{n} z_i x_i\right|$

Since $|\sum z_i x_i| \leq \sum |z_i x_i|$ by triangle

$= \sum |z_i| |x_i|$

$= \sup_{\max|x_i| \leq 1} \sum |z_i| |x_i|$ when $x_i = \text{sgn}(z_i)$

$= \sum |z_i|$

$= \|z\|_1$

$\|z\|_* = \sup\{z'x \mid \|x\|_1 \leq 1\}$

$= \sup\{z'x \mid \sum |x_i| \leq 1\}$

$= \sup_{\sum |x_i| \leq 1} \left|\sum_{i=1}^{n} z_i x_i\right|$

$= \max |z_i|$ with $x_i = e_i$

$= \|z\|_\infty$

d). $\|z\|_* = \sup\{z'x \mid \|x\|_p \leq 1\}$

$= \sup_{\|x\|_p \leq 1} |\sum z_i x_i|$

Where $|\sum z_i x_i| \leq \|z\|_q \|x\|_p$ Holder

$\leq \|z\|_q$

When $x_i = \frac{\text{sgn}(z_i) |z_i|^{q/p}}{\|z\|_q^{q/p}}$

$= \frac{z_i \, \text{sgn}(z_i) |z_i|^{q/p}}{\|z\|_q^{q/p}}$

$= \|z\|_q^{-q/p} \sum_{i=1}^{n} |z_i|^{q/p + 1}$

$= \|z\|_q^{-q/p} \sum_{i=1}^{n} |z_i|^{\frac{p}{p-1}}$

$= \|z\|_q^{-\frac{1}{p-1}} \left(\|z\|_{\frac{p}{p-1}}\right)^{\frac{p}{p-1}}$

$= \|z\|_q$

HW7 Prob 1:

Let Δx_{nsd} and Δx_{sd} be the normalized and unnormalized steepest descent direction at x, for the norm $\|\cdot\|$. Prove

a). $\nabla f(x)' \Delta x_{nsd} = -\|\nabla f(x)\|_*$

b). $\nabla f(x)' \Delta x_{sd} = -\|\nabla f(x)\|_*^2$

c). $\Delta x_{sd} = \text{argmin}_v (\nabla f(x)' v + \frac{1}{2} \|v\|^2)$

a). Recall $\Delta x_{nsd} = \text{argmin}_{\|v\| \leq 1} \{\nabla f(x)' v\}$

$\nabla f(x)' \Delta x_{nsd} = \min_{\|v\| \leq 1} \{\nabla f(x)' v\}$

$= -\sup_{\|v\| \leq 1} \{\nabla f(x)' v\}$

$= -\|\nabla f(x)\|_*$ dual norm

b). $\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$

c). Consider $v = tw$, $t \geq 0$ magnitude, $\|w\| = 1$ unit vector.

$\nabla f(x)' v + \frac{1}{2} \|v\|^2 \Rightarrow t \nabla f(x)' w + \frac{t^2}{2}$

Fix w, $\frac{\partial}{\partial t} f(t) = \nabla f(x)' w + t = 0$

$\Rightarrow t = -\nabla f(x)' w \Rightarrow t = \max(0, -\nabla f(x)' w)$

Find, $w = -\text{argmin}_{\|w\| = 1} \nabla f(x)' w$

$v = tw = \max(0, \nabla f(x)' \, \text{argmax}_{\|w\| = 1} \nabla f(x)' w) \, \text{arg} \ldots$

$= -\|\nabla f(x)\|_* (-\Delta x_{nsd})$

$= \Delta x_{sd}$

Prob 2:

Explain how to find a steepest descent direction in the l_∞ norm.

$\Delta x_{nsd} = \text{argmin}_{v: \|v\|_\infty \leq 1} \{\nabla f(x)' v\}$

$= \text{argmin}_{v: |v_i| \leq 1} \{\nabla f(x)' v\}$

$= \text{argmin}_{v: v_i \in [-1, 1]} \{\nabla f(x)' v\}$

Thus, select $v_i = -1$ if $(\nabla f(x))_i > 0$

$v_i = 1$ if $(\nabla f(x))_i < 0$

$v_i = -\text{sgn}(\nabla f(x)_i)$

The steepest descent direction in the l_∞ norm focus on reducing $f(x)$ by moving direction along the axis of the largest magnitude component of the gradient.

e). Consider the matrix spectral norm on $\mathbb{R}^{m \times n}$. $\|Z\|_{2*} = \sup\{tr(Z'X) \mid \|X\|_2 \leq 1\}$

Prove $\|Z\|_{2*} = \sigma_1(Z) + \cdots + \sigma_r(Z)$

$= tr(Z'Z)^{\frac{1}{2}} \quad \text{rank}(Z) = r$

$\|Z\|_{2*} = \sup\{tr(Z'X) \mid \max \sigma(X) \leq 1\}$

$= \sup\{tr((U \Sigma V')' X) \mid \max \sigma_X \leq 1\}$

$= \sup\{tr(V \Sigma U' X) \mid \max \sigma_X \leq 1\}$

$= \sup\{tr(\Sigma (U' X V)) \mid \max \sigma_X \leq 1\}$

$= \sup\{tr(\Sigma Y) \mid \max \sigma_X \leq 1\}$

$= \sup\{\sum_{i=1}^{r} \sigma_i(Z) Y_{ii} \mid \sigma_X \leq 1\}$

$= \sum_{i=1}^{r} \sigma_i(Z)$

$= tr(Z'Z)^{\frac{1}{2}}$