CHAPTER 1

inner product: an inner product on $\mathbb{R}^n$ is a map $\langle\cdot,\cdot\rangle: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:
1. (symmetry): $\langle x,y\rangle = \langle y,x\rangle$ for any $x,y \in \mathbb{R}^n$
2. (additivity): $\langle x, y+z\rangle = \langle x,y\rangle + \langle x,z\rangle$ for any $x,y,z \in \mathbb{R}^n$
3. (homogeneity): $\langle \lambda x, y\rangle = \lambda\langle x,y\rangle$ for any $\lambda \in \mathbb{R}$ and $x,y \in \mathbb{R}^n$
4. (positive definiteness): $\langle x,x\rangle \geq 0$ for any $x \in \mathbb{R}^n$ and $\langle x,x\rangle = 0$ iff $x = 0$

vector norms: A norm $\|\cdot\|$ on $\mathbb{R}^n$ is a function $\|\cdot\|: \mathbb{R}^n \to \mathbb{R}$ satisfying the following:
1. (nonnegativity): $\|x\| \geq 0$ for any $x \in \mathbb{R}^n$ and $\|x\| = 0$ iff $x = 0$
2. (positive homogeneity): $\|\lambda x\| = |\lambda|\|x\|$ for any $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$
3. (triangle inequality): $\|x+y\| \leq \|x\| + \|y\|$ for any $x,y \in \mathbb{R}^n$

Euclidean norm: $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ for all $x \in \mathbb{R}^n$

p-norm: $\|x\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$

infinity norm: $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$

Cauchy-Schwarz inequality: For any $x,y \in \mathbb{R}^n$, $|x^T y| \leq \|x\|_2 \|y\|_2$ (equality iff x,y linearly dependent)

matrix norms: A norm on $\mathbb{R}^{m\times n}$ is a function $\|\cdot\|: \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfying
1. (nonnegativity): $\|A\| \geq 0$ for any $A \in \mathbb{R}^{m\times n}$ and $\|A\| = 0$ iff $A = 0$
2. (positive homogeneity): $\|\lambda A\| = |\lambda|\|A\|$ for any $A \in \mathbb{R}^{m\times n}$ and $\lambda \in \mathbb{R}$
3. (triangle inequality): $\|A+B\| \leq \|A\| + \|B\|$ for any $A,B \in \mathbb{R}^{m\times n}$

induced matrix norm: $\|A\|_{a,b} = \max_x \{\|Ax\|_b : \|x\|_a \leq 1\}$
→ implies: $\|Ax\|_b \leq \|A\|_{a,b}\|x\|_a$

spectral norm: $\|A\|_2 = \|A\|_{2,2} = \sqrt{\lambda_{max}(A^T A)} = \sigma_{max}(A)$

1-norm: $\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^m |A_{ij}|$ (max abs col sum)

∞ norm: $\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^n |A_{ij}|$ (max abs row sum)

Frobenius norm: $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$, $A \in \mathbb{R}^{m\times n}$

eigenvector: Given $A \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n$ is an eigenvector of A if $\exists \lambda \in \mathbb{C}$ for which $Av = \lambda v$, and λ is the corresponding eigenvalue.
- EIGENVALS OF SYMMETRIC MATRICES ARE REAL
- we denote the eigenvals of a symmetric $n\times n$ matrix $\lambda_1(A) \geq \lambda_2(A) \geq \dots \geq \lambda_n(A)$

Spectral Decomposition Theorem: Let $A \in \mathbb{R}^{n\times n}$ be an $n\times n$ symmetric matrix. $\exists$ an orthogonal matrix $U \in \mathbb{R}^{n\times n}$ ($U^T U = UU^T = I$) and a diagonal matrix D for which $U^T A U = D$.
- columns of U are comprised of eigenvectors of A and D's elements are the corresponding eigenvals
- $Tr(A) = \sum_{i=1}^n \lambda_i(A)$ $\quad \det(A) = \prod_{i=1}^n \lambda_i(A)$

Rayleigh Quotient: $R_A(x) = \frac{x^T A x}{\|x\|^2}$ for any $x \neq 0$

Lemma 1.11: Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then $\lambda_{min}(A) \leq R_A(x) \leq \lambda_{max}(A)$ $\forall x \neq 0$
x = corresponding eigenvecs → $\min_{x\neq 0} R_A(x) = \lambda_{min}(A)$, $\max_{x\neq 0} R_A(x) = \lambda_{max}(A)$

open ball: the open ball with center $c \in \mathbb{R}^n$ and radius r is denoted $B(c,r)$ and defined $B(c,r) = \{x \in \mathbb{R}^n : \|x-c\| < r\}$

closed ball: $B[c,r] = \{x \in \mathbb{R}^n : \|x - c\| \leq r\}$

interior points: Given a set $U \subseteq \mathbb{R}^n$, a point $c \in U$ is an interior point if $\exists r > 0$ for which $B(c,r) \subseteq U$.

interior of a set: $int(U) = \{x \in U : B(x,r) \subseteq U\}$

open set: a set that contains only interior points: for every $x \in U$ $\exists r > 0$ s.t. $B(x,r) \subseteq U$.

closed set: a set $U \subseteq \mathbb{R}^n$ that contains all limits of convergent sequences of points in U, i.e. $\forall \{x_i\}_{i\geq 0} \subseteq U$ satisfying $x_i \to x^*$ as $i \to \infty$, $x^* \in U$.
- A set is closed iff its complement is open.
- Closedness of level & contour sets of cts functions: let f be a cts function defined over the closed set $S \subseteq \mathbb{R}^n$. Then for any $\alpha \in \mathbb{R}$ the sets $Lev(f,\alpha) = \{x \in S : f(x) \leq \alpha\}$ and $Con(f,\alpha) = \{x \in S : f(x) = \alpha\}$ are closed.

boundary points: Given $U \subseteq \mathbb{R}^n$, a boundary point is a point $x \in \mathbb{R}^n$ s.t. any neighbourhood of x contains at least one point in U and one point in U^c.
- **Boundary of U**, $bd(U)$: set of all boundary points of U.
- **Closure of U**, $cl(U)$: smallest closed set containing U: $\bigcap\{T : U \subseteq T, T \text{ closed}\}$
$cl(U) = U \cup bd(U)$

bounded set: set U s.t. $\exists M > 0$ for which $U \subseteq B(0,M)$

compact set: a closed and bounded set

directional derivative: let f be a function defined on $S \subseteq \mathbb{R}^n$. Let $x \in int(S)$, $0 \neq d \in \mathbb{R}^n$. The DD (if exists) is: $f'(x;d) = \lim_{t\to 0^+} \frac{f(x+td) - f(x)}{t} = \nabla f(x)^T d$

ith partial derivative: $\frac{\partial f}{\partial x_i}(x) = \lim_{t\to 0^+} \frac{f(x+te_i) - f(x)}{t}$

gradient of f: $\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x)\end{pmatrix}$

ctsly diffable: a function f defined on open set U for which all partial derivatives exist and are cts on U.

proposition 1.23: Let $f: U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose f is ctsly diffable over U. Then
$\lim_{d\to 0} \frac{f(x+d) - f(x) - \nabla f(x)^T d}{\|d\|} = 0$ for all $x \in U$
equiv: $f(y) = f(x) + \nabla f(x)^T(y-x) + o(\|y-x\|)$ $\quad$ [$o(\cdot): \mathbb{R}_+ \to \mathbb{R}$, $\frac{o(t)}{t} \to 0$ as $t \to 0^+$]

twice ctsly diffable: f defined on open set U s.t. all 2nd order partial derivatives exist and are cts over U.

Hessian: of f at point $x \in U$ is the $n\times n$ matrix
$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x)\end{bmatrix}$

Linear approximation theorem: Let $f: U \to \mathbb{R}$ be a twice ctsly diffable function over an open set $U \subseteq \mathbb{R}^n$, $x \in U$, $r > 0$ satisfy $B(x,r) \subseteq U$. Then for any $y \in B(x,r)$ there exists $\xi \in [x,y]$ s.t.
$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(\xi)(y-x)$

Quadratic approximation theorem: Let $f: U \to \mathbb{R}$ be a twice ctsly diffable function over an open set $U \subseteq \mathbb{R}^n$, $x \in U$, $r > 0$ satisfy $B(x,r) \subseteq U$. Then for any $y \in B(x,r)$
$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x) + o(\|y-x\|^2)$

Global min and max: Let f be defined on $S \subseteq \mathbb{R}^n$.
1. $x^* \in S$ is a global min point of f over S if $f(x) \geq f(x^*)$ for any $x \in S$.
2. $x^* \in S$ is a strict global min point of f over S if $f(x) > f(x^*)$ for any $x^* \neq x \in S$.
3. $x^* \in S$ is a global max point of f over S if $f(x) \leq f(x^*)$ for any $x \in S$.
4. $x^* \in S$ is a strict global max point of f over S if $f(x) < f(x^*)$ for any $x^* \neq x \in S$.

global optimum: $x^* \in S$ is either a global min or global max.

local min and max: Let $f: S \to \mathbb{R}$ be defined on $S \subseteq \mathbb{R}^n$.
1. $x^* \in S$ is a (strict) local min point of f over S if $\exists r > 0$ for which $f(x^*) \leq f(x)$ $(<)$ $\forall x \in S \cap B(x^*,r)$ $(x \neq x^*)$
2. $x^* \in S$ is a (strict) local max point of f over S if $\exists r > 0$ for which $f(x^*) \geq f(x)$ $(>)$ $\forall x \in S \cap B(x^*,r)$ $(x \neq x^*)$

Theorem: first order optimality condition for local optima: Let $f: U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose $x^* \in int(U)$ is a local optimum point and all partial derivatives of f exist at x^*. Then $\nabla f(x^*) = 0$.

stationary point: Let $f: U \to \mathbb{R}$ be defined on a set $U \subseteq \mathbb{R}^n$. Suppose $x^* \in int(U)$ and f is diffable over some neighbourhood of x^*. Then x^* is called a stationary point of f if $\nabla f(x^*) = 0$.

Positive definiteness:
1. positive definite ($A \succ 0$): $x^T A x > 0$ for every $0 \neq x \in \mathbb{R}^n$
2. positive semidefinite ($A \succeq 0$): $x^T A x \geq 0$ for every $x \in \mathbb{R}^n$

Lemma 2.13: Let A be a positive semidefinite matrix. Then its diagonal elements are nonnegative.

negative semidefiniteness & indefiniteness: A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is
1. negative semidefinite: $A \preceq 0$ if $x^T A x \leq 0$ $\forall x \in \mathbb{R}^n$
2. negative definite: $A \prec 0$ if $x^T A x < 0$ $\forall 0 \neq x \in \mathbb{R}^n$
3. indefinite: $\exists x,y \in \mathbb{R}^n$ s.t. $x^T A x > 0$, $y^T A y < 0$

* A is negative (semi)definite iff $-A$ is positive (semi)definite

Lemma 2.15
- Diagonal elements of a negative definite matrix are negative
- If a matrix has positive and negative diagonal elements it is indefinite

Eigenvalue characterization thm: Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then
a) A is pd iff all eigenvals are pos
b) A is psd iff all eigenvals are nonneg
c) A is nd iff all eigenvals are neg
d) A is nsd iff all eigenvals are nonpos
e) A is indef iff it has at least one pos and one neg eigenval

Corollary 2.17: Let A be a positive semidef (def) matrix. Then $Tr(A)$ and $\det(A)$ are nonneg (pos).
(eigenvals of a diag matrix are its diag elements)

Proposition 2.18: Let A be a symmetric 2×2 matrix. Then A is positive semidef (def) iff $Tr(A), \det(A) \geq 0$ (>0)

Principal minors criterion: Let A be an $n\times n$ symmetric matrix. Then A is pos def iff $D_1(A) > 0, D_2(A) > 0, \dots, D_n(A) > 0$ ← determinants of upper submatrices

Diagonally dominant matrices: Let A be a symmetric $n\times n$ matrix.
1. A is diagonally dominant if $|A_{ii}| \geq \sum_{j\neq i}|A_{ij}|$
2. A is strictly diagonally dominant if $|A_{ii}| > \sum_{j\neq i}|A_{ij}|$

positive semidefiniteness of diagonally dominant matrices:
a) Let A be a symmetric $n\times n$ diagonally dominant matrix whose diagonal elements are nonnegative. Then A is psd.
b) Let A be a symmetric $n\times n$ strictly diagonally dominant matrix whose diagonal elements are positive. Then A is pd.

Necessary second order optimality condition: Let $f: U \to \mathbb{R}$ be a function defined over $U \subseteq \mathbb{R}^n$. Suppose f is twice ctsly diffable over U and x^* is a stationary point. Then
1. If x^* is a local min point of f over U, then $\nabla^2 f(x^*) \succeq 0$
2. If x^* is a local max point of f over U, then $\nabla^2 f(x^*) \preceq 0$

Sufficient second order optimality condition: Let $f: U \to \mathbb{R}$ be a function defined over $U \subseteq \mathbb{R}^n$. Suppose f is twice ctsly diffable over U and $x^* \in int(U)$ a stationary point. Then
1. If $\nabla^2 f(x^*) \succ 0$, then x^* is a strict local min point of f over U
2. If $\nabla^2 f(x^*) \prec 0$, then x^* is a strict local max point of f over U

saddle point: Let $f: U \to \mathbb{R}$ be defined on $U \subseteq \mathbb{R}^n$. A stationary point $x^* \in int(U)$ is a saddle point of f over U if it is neither a local min nor a local max point of f over U.

Sufficient condition for a saddle point: Suppose f is twice ctsly diffable and x^* is a stationary point. If $\nabla^2 f(x^*)$ is indefinite, then x^* is a saddle point of f over U.

Weierstrass Theorem: Let f be a cts function defined over a nonempty and compact set $C \subseteq \mathbb{R}^n$. Then $\exists$ a global min point of f over C and a global max point of f over C.

coercive: Let $f: \mathbb{R}^n \to \mathbb{R}$ be a cts function defined over $\mathbb{R}^n$. f is coercive if $\lim_{\|x\|\to\infty} f(x) = \infty$.

Attainment under coerciveness: If $f: \mathbb{R}^n \to \mathbb{R}$, a cts function, is coercive and $S \subseteq \mathbb{R}^n$ a nonempty closed set, then f has a global min point on S.

Theorem 2.38 Global Optimality: Let f be a twice ctsly diffable function over $\mathbb{R}^n$. Let $x^* \in \mathbb{R}^n$ be a stationary point of f. Then if $\nabla^2 f(x) \succeq 0$ for all x, x^* is a global min point of f over $\mathbb{R}^n$.

Quadratic function: function over $\mathbb{R}^n$ of the form $f(x) = x^T A x + 2b^T x + c$, where $A \in \mathbb{R}^{n\times n}$ symmetric, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$
- $\nabla f(x) = 2Ax + 2b$, $\nabla^2 f(x) = 2A$

Lemma 2.41: Let $f(x) = x^T A x + 2b^T x + c$, where $A \in \mathbb{R}^{n\times n}$ is symmetric, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then
a) x is a stationary point of f iff $Ax = -b$
b) If $A \succeq 0$, then x is a global min pt of f iff $Ax = -b$
c) If $A \succ 0$, then $x = -A^{-1}b$ is a strict global min of f.

Lemma 2.42 (coerciveness of quadratic functions): Let f be a quadratic function. Then f is coercive iff $A \succ 0$.

Theorem 2.43 (characterization of nonnegativity of quadratic functions): Let $f(x) = x^T A x + 2b^T x + c$. Then the following two claims are equivalent: a) $f(x) \geq 0$ $\forall x \in \mathbb{R}^n$ b) $\begin{pmatrix} A & b \\ b^T & c\end{pmatrix} \succeq 0$.

CHAPTER 3

Solution of overdetermined systems: Consider $Ax \approx b$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Solution exists if $m > n$ (also assume $rank(A) = n$). Since this system is usually inconsistent (no soln), we can find the best approximate solution.
(LS) $\min \|Ax - b\|^2 = \min\ x^T A^T A x - 2b^T A x + \|b\|^2$
$\nabla f(x) = 2A^T A x - 2A^T b = 0$
→ $A^T A x = A^T b$, $x_{LS} = (A^T A)^{-1} A^T b$ (Lemma 2.41)

regularized Least Squares: helpful if A underdetermined
(RLS) $\min \|Ax - b\|^2 + \lambda R(x)$
Often $R(x) = \|Dx\|^2$
$\min f_{RLS}(x) = x^T(A^T A + \lambda D^T D)x - 2b^T A x + \|b\|^2$
Hessian of objective function: $\nabla^2 f_{RLS}(x) = 2(A^T A + \lambda D^T D) \succeq 0$
$x_{RLS} = (A^T A + \lambda D^T D)^{-1} A^T b$

CHAPTER 4

descent direction: Let $f: \mathbb{R}^n \to \mathbb{R}$ be a ctsly diffable function. $0 \neq d \in \mathbb{R}^n$ is a descent direction of f at x if the directional derivative $f'(x;d)$ is negative, i.e. $f'(x;d) = \nabla f(x)^T d < 0$.

Lemma 4.2 (descent property): Given $f: \mathbb{R}^n \to \mathbb{R}$ ctsly diffable, $x \in \mathbb{R}^n$, and d a descent direction of f at x. Then $\exists \varepsilon > 0$ s.t. $f(x + td) < f(x)$ for any $t \in (0, \varepsilon]$.

Lemma 4.3 (validity of sufficient decrease condition): Let f be a ctsly diffable function over $\mathbb{R}^n$, $x \in \mathbb{R}^n$, and d be a descent direction of f at x. Let $\alpha \in (0,1)$. Then $\exists \varepsilon > 0$ s.t. $f(x) - f(x + td) \geq -\alpha t \nabla f(x)^T d$ holds $\forall t \in [0, \varepsilon]$.

Lemma 4.5: Let f be a ctsly diffable function and $x \in \mathbb{R}^n$ a non-stationary point. Then an optimal soln of $\min\{f'(x;d) : \|d\| = 1\}$ is $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$.

THE GRADIENT METHOD:
- Input: $\varepsilon > 0$ tolerance parameter
- Initialization: Pick arbitrary $x_0 \in \mathbb{R}^n$
- General Step: For any $k = 0,1,\dots$ execute the following steps:
  a) Pick a stepsize t_k by a line search procedure on the function $g(t) = f(x_k - t\nabla f(x_k))$ (can add scaling matrix D_k here)
  b) Set $x_{k+1} = x_k - t_k\nabla f(x_k)$
  c) If $\|\nabla f(x_{k+1})\| \leq \varepsilon$, STOP. Output x_{k+1}

Lemma 4.6: Let $\{x_k\}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a ctsly diffable function f. Then for any $k = 0,1,2,\dots$ $(x_{k+2} - x_{k+1})^T(x_{k+1} - x_k) = 0$.

Kantorovich Inequality: Let $A \in \mathbb{R}^{n\times n}$ be pos def. Then $\forall x \neq 0$
$\frac{(x^T x)^2}{(x^T A x)(x^T A^{-1} x)} \geq \frac{4\lambda_{max}(A)\lambda_{min}(A)}{(\lambda_{max}(A) + \lambda_{min}(A))^2}$

Lemma 4.11: Let $\{x_k\}$ be the sequence generated by the gradient descent method with exact line search for solving $\min x^T A x$. Then for any $k = 0,1,\dots$
$f(x_{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(x_k)$, $M = \lambda_{max}(A)$, $m = \lambda_{min}(A)$

Df (L-Lipschitz): Given a ctsly diffable objective function f, its gradient is Lipschitz cts over $\mathbb{R}^n$ if $\exists L \geq 0$ for which $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ $\forall x,y \in \mathbb{R}^n$.

$C_L^{1,1}$: The class of functions with Lipschitz gradient with constant L.
- Linear functions: $f(x) = a^T x \in C_0^{1,1}$
- Quadratic functions: $f(x) = x^T A x + 2b^T x + c \in C_{2\|A\|}^{1,1}$

Theorem 4.20: Let f be a twice ctsly diffable function over $\mathbb{R}^n$. Then the following claims are equivalent: a) $f \in C_L^{1,1}(\mathbb{R}^n)$ b) $\|\nabla^2 f(x)\| \leq L$ $\forall x \in \mathbb{R}^n$

Descent Lemma: Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any $x,y \in \mathbb{R}^n$
$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|x-y\|^2$

Sufficient decrease lemma: Suppose $f \in C_L^{1,1}(\mathbb{R}^n)$. Then for any $x \in \mathbb{R}^n$, $t > 0$
$f(x) - f(x - t\nabla f(x)) \geq t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x)\|^2$

Lemma 4.23 SUFFICIENT DECREASE OF THE GRADIENT METHOD: Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Let $\{x_k\}$ be the sequence generated by the gradient method for solving $\min f(x)$ with one of the following stepsize strategies:
- constant stepsize $\bar{t} \in (0, \frac{2}{L})$
- exact line search
- backtracking procedure with $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, $\beta \in (0,1)$

Then $f(x_k) - f(x_{k+1}) \geq M\|\nabla f(x_k)\|^2$
where $M = \begin{cases} \bar{t}\left(1 - \frac{\bar{t}L}{2}\right) & \text{constant stepsize} \\ \frac{1}{2L} & \text{exact line search} \\ \alpha\min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\} & \text{backtracking}\end{cases}$

Theorem 4.25 (convergence of gradient method): Let $f \in C_L^{1,1}$ and let $\{x_k\}_{k\geq 0}$ be the sequence generated by the gradient method for solving $\min f(x)$ with one of the following stepsize strategies:
- constant stepsize $\bar{t} \in (0, \frac{2}{L})$
- exact line search
- backtracking procedure

Assume f is bounded below over $\mathbb{R}^n$, i.e. $\exists m \in \mathbb{R}$ s.t. $f(x) > m$ $\forall x \in \mathbb{R}^n$. Then
a) The sequence $\{f(x_k)\}$ is nonincreasing and for any $k \geq 0$ $f(x_{k+1}) < f(x_k)$ unless $\nabla f(x_k) = 0$
b) $\nabla f(x_k) \to 0$ as $k \to \infty$

Theorem 4.26 (rate of convergence of gradient norm): under the setting of Theorem 4.25, let f^* be the limit of the convergent sequence $\{f(x_k)\}$. Then for any $n = 0,1,\dots$
$\min_{k=0,\dots,n}\|\nabla f(x_k)\| \leq \sqrt{\frac{f(x_0) - f^*}{M(n+1)}}$
where M is as in Theorem 4.23.

Chapter 5: Newton's Method

main idea: given iterate x_k, choose x_{k+1} to minimize the quadratic approximation of the function around x_k
$x_{k+1} = \arg\min\{f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2}(x - x_k)^T\nabla^2 f(x_k)(x - x_k)\}$
* Only well defined if $\nabla^2 f(x_k) \succ 0$ *

unique minimizer: $\nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k) = 0$
$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1}\nabla f(x_k)$ (*)

Newton direction: $-(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$

Pure Newton's method: algorithm indicated by update formula (*)

Theorem 5.2: Quadratic local convergence of Newton's method: Let f be a twice ctsly diffable function over $\mathbb{R}^n$. Assume
- $\exists m > 0$ s.t. $\nabla^2 f(x) \succeq mI$ $\forall x \in \mathbb{R}^n$
- $\exists L > 0$ s.t. $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ $\forall x,y \in \mathbb{R}^n$

Let $\{x_k\}_{k\geq 0}$ be the sequence generated by Newton's method and let x^* be the unique minimizer of f over $\mathbb{R}^n$. Then for $k = 0,1,\dots$
$\|x_{k+1} - x^*\| \leq \frac{L}{2m}\|x_k - x^*\|^2$
Also, if $\|x_0 - x^*\| \leq \frac{m}{L}$, then
$\|x_k - x^*\| \leq \frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}$, $k = 0,1,\dots$

Damped Newton's method:
- Input: $\alpha, \beta \in (0,1)$, $\varepsilon > 0$ tolerance param
- Init: Pick $x_0 \in \mathbb{R}^n$ arbitrarily
- General Step: For $k = 0,1,2,\dots$
  a) Compute Newton direction $d_k = (\nabla^2 f(x_k))^{-1}\nabla f(x_k)$
  b) Set $t_k = 1$. While $f(x_k) - f(x_k + t_k d_k) < -\alpha t_k\nabla f(x_k)^T d_k$ set $t_k := \beta t_k$
  c) $x_{k+1} = x_k + t_k d_k$ (can also use constant stepsize here)
  d) If $\|\nabla f(x_{k+1})\| \leq \varepsilon$, STOP, output x_{k+1}.

Hybrid Gradient-Newton: Newton step if Hessian is pd, gradient step otherwise

Level set: $f: S \to \mathbb{R}$ a function on $S \subseteq \mathbb{R}^n$. The level set with level α is
$Lev(f,\alpha) = \{x \in S : f(x) \leq \alpha\}$

CHAPTER 6

convex set: A set $C \subseteq \mathbb{R}^n$ s.t. $\forall x,y \in C$, $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in C$.

Some sets that are convex:
- lines $x + td$, $t \in \mathbb{R}$
- halfspaces: $\{x \in \mathbb{R}^n, a^T x \leq b\}$
- hyperplanes $\{x \in \mathbb{R}^n, a^T x = b\}$
- open halfspaces: $\{x \in \mathbb{R}^n, a^T x < b\}$
- open & closed balls: $\{x \in \mathbb{R}^n : \|x - c\| < r\}$ (for any norm)
- ellipsoids: $\{x \in \mathbb{R}^n : x^T Q x + 2b^T x + c \leq 0\}$, $Q \succeq 0$

Lemma 6.6: Let $C_i \subseteq \mathbb{R}^n$ be a convex set for any $i \in I$, where I is a (possibly infinite) index set. Then $\bigcap_{i\in I} C_i$ is convex.

- Convex polytopes, sets defined by linear inequalities, are convex! (intersection of half spaces)
$P = \{x \in \mathbb{R}^n : Ax \leq b\}$

Theorem 6.8: preservation of convexity
a) Let $C_1, C_2, \dots, C_k \subseteq \mathbb{R}^n$ be convex sets, $\mu_1, \dots, \mu_k \in \mathbb{R}$. Then $\mu_1 C_1 + \dots + \mu_k C_k = \{\sum_{i=1}^k \mu_i x_i : x_i \in C_i, i = 1,\dots,k\}$ is convex.
b) Let $C_i \subseteq \mathbb{R}^{k_i}$ be a convex set for any $i = 1,\dots,m$. Then the Cartesian product $C_1 \times C_2 \times \dots \times C_m = \{(x_1, x_2, \dots, x_m) : x_i \in C_i, i = 1,\dots,m\}$ is convex.
c) Let $M \subseteq \mathbb{R}^n$ be a convex set, $A \in \mathbb{R}^{m\times n}$. Then $A(M) = \{Ax : x \in M\}$ is convex.
d) Let $D \subseteq \mathbb{R}^m$ be a convex set. Let $A \in \mathbb{R}^{m\times n}$. Then the set $A^{-1}(D) = \{x \in \mathbb{R}^n : Ax \in D\}$ is convex.

convex combinations: given k vectors $x_1, x_2, \dots, x_k \in \mathbb{R}^n$, a convex combo of these vectors is a vector of the form $\lambda_1 x_1 + \dots + \lambda_k x_k$ where $\lambda_1, \dots, \lambda_k \geq 0$ and $\sum \lambda_i = 1$.

Theorem 6.10: Let $C \subseteq \mathbb{R}^n$ be a convex set and let $x_1, \dots, x_m \in C$. Then for any $\lambda \in \Delta_m$, $\sum_{i=1}^m \lambda_i x_i \in C$

convex hulls: Let $S \subseteq \mathbb{R}^n$. The convex hull of S, $conv(S)$, is the set comprising all the convex combinations of vectors from S.
$conv(S) = \{\sum_{i=1}^k \lambda_i x_i : x_1, \dots, x_k \in S, \lambda \in \Delta_k, k \in \mathbb{N}\}$

Lemma 6.12: Let $S \subseteq \mathbb{R}^n$. If $S \subseteq T$, T convex, then $conv(S) \subseteq T$.

Caratheodory Theorem: Let $S \subseteq \mathbb{R}^n$, $x \in conv(S)$. Then $\exists$ $x_1, \dots, x_{n+1} \in S$ s.t. $x \in conv(\{x_1, \dots, x_{n+1}\})$, i.e. $\exists \lambda \in \Delta_{n+1}$ s.t. $x = \sum_{i=1}^{n+1}\lambda_i x_i$.

cone: a set $S \subseteq \mathbb{R}^n$ s.t. $\forall x \in S$, $\lambda \geq 0$, $\lambda x \in S$.

convex cone: a set S is a convex cone iff
a) $x, y \in S \Rightarrow x + y \in S$ b) $x \in S, \lambda \geq 0 \Rightarrow \lambda x \in S$

conic combination: given $x_1, \dots, x_k \in \mathbb{R}^n$, their conic combo is any vector of the form $\lambda_1 x_1 + \dots + \lambda_k x_k$, where $\lambda \in \mathbb{R}_+^k$

Lemma 6.20: Let C be a convex cone, $x_1, \dots, x_k \in C$, and $\lambda_1, \dots, \lambda_k \geq 0$. Then $\sum_{i=1}^k \lambda_i x_i \in C$.

conic hull: Let $S \subseteq \mathbb{R}^n$. The conic hull of S, $cone(S)$, is the set comprising all the conic combos of vectors of S.
$cone(S) = \{\sum_{i=1}^k \lambda_i x_i : x_1, \dots, x_k \in S, \lambda \in \mathbb{R}_+^k, k \in \mathbb{N}\}$

Lemma 6.22: Let $S \subseteq \mathbb{R}^n$. If $S \subseteq T$, T convex cone, then $cone(S) \subseteq T$.

conic representation theorem: Let $S \subseteq \mathbb{R}^n$, $x \in cone(S)$. Then $\exists$ k linearly independent vectors $x_1, \dots, x_k \in S$ s.t. $x \in cone(\{x_1, \dots, x_k\})$, i.e. $\exists \lambda \in \mathbb{R}_+^k$ s.t. $x = \sum_{i=1}^k \lambda_i x_i$, where $k \leq n$.

Basic feasible solution: Let $P = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$, where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Suppose rows of A are linearly independent. Then $\bar{x}$ is a BFS of P if the columns of A corresponding to the indices of the positive values of $\bar{x}$ are linearly independent. (Has at most m nonzero elements)

Theorem 6.26: Let $P = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then if $P \neq \emptyset$ it contains at least one BFS.

Theorem 6.27 (convexity under closure): $cl(C)$ of a convex set is convex.

Lemma 6.28 (line segment principle): Let C be a convex set, $int(C) \neq \emptyset$. Suppose $x \in int(C)$, $y \in cl(C)$. Then $(1-\lambda)x + \lambda y \in int(C)$ $\forall \lambda \in (0,1)$

Theorem 6.29: $int(C)$ of convex C is convex

Lemma 6.30: Let C be a convex set with nonempty interior. Then a) $cl(int(C)) = cl(C)$ b) $int(cl(C)) = int(C)$

Prop 6.31: Let $S \subseteq \mathbb{R}^n$ be a compact set. Then $conv(S)$ is compact

Lemma 6.32: Let $a_1, \dots, a_k \in \mathbb{R}^n$. Then $cone(\{a_1, \dots, a_k\})$ is closed

extreme points: Let $S \subseteq \mathbb{R}^n$. $x \in S$ is an extreme point if there do not exist $x_1, x_2 \in S$ $(x_1 \neq x_2)$ and $\lambda \in (0,1)$ s.t. $x = \lambda x_1 + (1-\lambda)x_2$

Theorem 6.34: Let $P = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$, where $A \in \mathbb{R}^{m\times n}$ has linearly independent rows and $b \in \mathbb{R}^m$. Then $\bar{x}$ is a BFS of P iff it's an extreme point of P.

Krein-Milman Theorem: Let $S \subseteq \mathbb{R}^n$ be a compact convex set. Then $S = conv(ext(S))$.

CHAPTER 7

convex function: A function $f: C \to \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^n$ is convex if
$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for any $x,y \in C$, $\lambda \in [0,1]$
(a function is strictly convex if the inequality holds strictly for $\lambda \in (0,1)$)

Jensen's Inequality: Let $f: C \to \mathbb{R}$ be a convex function where $C \subseteq \mathbb{R}^n$ is a convex set. Then for any $x_1, \dots, x_k \in C$ and $\lambda \in \Delta_k$, the following inequality holds:
$f\left(\sum_{i=1}^k \lambda_i x_i\right) \leq \sum_{i=1}^k \lambda_i f(x_i)$

Theorem: Gradient Inequality: Let $f: C \to \mathbb{R}$ be a ctsly diffable function defined on a convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C iff
$f(x) + \nabla f(x)^T(y - x) \leq f(y)$ for any $x,y \in C$
(for strictly convex, this inequality becomes strict)

Proposition 7.8: Let f be a ctsly diffable function that is convex over convex set $C \subseteq \mathbb{R}^n$. Suppose $\nabla f(x^*) = 0$ for some $x^* \in C$. Then x^* is a global minimizer of f over C.

Theorem 7.9: Let $f: \mathbb{R}^n \to \mathbb{R}$ be a ctsly diffable convex function. Then $\nabla f(x^*) = 0$ iff x^* is a global minimum point of f over $\mathbb{R}^n$.

Theorem 7.10: Let $f: \mathbb{R}^n \to \mathbb{R}$ be the quadratic function $f(x) = x^T A x + 2b^T x + c$, where $A \in \mathbb{R}^{n\times n}$ is symmetric, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Then f is (strictly) convex iff $A \succeq 0$ $(A \succ 0)$

Theorem 7.11: Suppose f is ctsly diffable over convex $C \subseteq \mathbb{R}^n$. Then f is convex over C iff
$(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0$ $\forall x,y \in C$.

Theorem 7.12: Let f be a twice ctsly diffable function over an open convex set $C \subseteq \mathbb{R}^n$. Then f is convex iff $\nabla^2 f(x) \succeq 0$ for any $x \in C$.

Theorem 7.13: If $\nabla^2 f(x) \succ 0$ $\forall x \in C$, C convex, f twice ctsly diffable, then f is strictly convex.

Theorem 7.16
a) If f convex over $C \subseteq \mathbb{R}^n$ convex, $\alpha \geq 0$, then αf convex
b) The sum of convex functions is convex.

Theorem 7.17: Let $f: C \to \mathbb{R}$ be a convex function defined on a convex set $C \subseteq \mathbb{R}^n$. Let $A \in \mathbb{R}^{n\times m}$, $b \in \mathbb{R}^n$. Then $g(y) = f(Ay + b)$ is convex over $D = \{y \in \mathbb{R}^m : Ay + b \in C\}$

Theorem 7.22: Let $f: C \to \mathbb{R}$ be a convex function over the convex set $C \subseteq \mathbb{R}^n$. Let $g: I \to \mathbb{R}$ be a 1-dimensional nondecreasing convex function over the interval $I \subseteq \mathbb{R}$. Assume the image of C under f is contained in I: $f(C) \subseteq I$. Then
$h(x) = g(f(x))$, $x \in C$ is convex over C.

Theorem 7.25: Let $f_1, \dots, f_p: C \to \mathbb{R}$ be p convex functions over the convex set $C \subseteq \mathbb{R}^n$. Then the maximum function
$f(x) = \max_{i=1,\dots,p} f_i(x)$ is convex over C.

Theorem 7.27: Let $f: C \times D \to \mathbb{R}$ be a convex function defined over the set $C \times D$ where $C \subseteq \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$ are convex.
$g(x) = \min_{y\in D} f(x,y)$, $x \in C$
where we assume the minimum is finite, g is convex!

Theorem 7.31: Let $f: C \to \mathbb{R}$ be a convex function over $C \subseteq \mathbb{R}^n$. Then for any $\alpha \in \mathbb{R}$ the level set $Lev(f,\alpha)$ is convex.

Theorem 7.36: Let $f: C \to \mathbb{R}$ be a convex function defined over convex $C \subseteq \mathbb{R}^n$. Let $x_0 \in \text{int}(C)$. Then $\exists\, \varepsilon > 0$ and $L > 0$ s.t. $B[x_0, \varepsilon] \subseteq C$ and

$$|f(x) - f(x_0)| \le L\|x - x_0\| \quad \forall x \in B[x_0, \varepsilon]$$

Theorem 7.37: Let $f: C \to \mathbb{R}$ be convex over convex $C \subseteq \mathbb{R}^n$, $x \in \text{int}(C)$. Then for any $d \neq 0$, $f'(x; d) = \nabla f(x)^T d$ exists.

Theorem 7.38: Let $f_i: \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be an extended real-valued convex function for any $i \in I$. Then $\max_{i \in I} f_i(x)$ is an extended real-valued convex function.

Theorem 4.41: Let $f: C \to \mathbb{R}$ be a convex function which is not constant over the convex set C. Then f does not attain a max at a point in $\text{int}(C)$.

Theorem 4.42: Let $f: C \to \mathbb{R}$ be convex & cts over convex & compact $C \subseteq \mathbb{R}^n$. Then $\exists$ at least one maximizer of f over C that is an extreme point of C.

AGM Inequality: For any $x_1, x_2, \dots, x_n \ge 0$,

$$\frac{1}{n}\sum_{i=1}^n x_i \ge \sqrt[n]{\prod_{i=1}^n x_i} \Rightarrow \text{for any } \lambda \in \Delta_n,\ \sum_{i=1}^n \lambda_i x_i \ge \prod_{i=1}^n x_i^{\lambda_i}$$

Young's Inequality: For any $s, t \ge 0$, $p, q > 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$,

$$st \le \frac{s^p}{p} + \frac{t^q}{q}$$

Hölder's Inequality: For any $x, y \in \mathbb{R}^n$ and $p, q \ge 1$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$,

$$|x^T y| \le \|x\|_p \|y\|_q$$

Minkowski's Inequality: Let $p \ge 1$. Then for any $x, y \in \mathbb{R}^n$

$$\|x + y\|_p \le \|x\|_p + \|y\|_p$$

PRACTICE MIDTERM SOLUTIONS

3) Suppose that $\|x_{t+1} - x^*\|^2 \le \|x_t - x^*\|^2 - g_t^2$ for some sequence g_t. Prove $\min_t |g_t|$ approaches 0 at a rate of $1/\sqrt{t}$.

$g_t^2 \le \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2$

$\sum_{t=1}^T g_t^2 \le \|x_0 - x^*\| - \|x_{T+1} - x^*\| \le \|x_0 - x^*\|$

$T \min g_t^2 \le \|x_0 - x^*\|$

$\therefore \min |g_t| \le \frac{\sqrt{\|x_0 - x^*\|}}{\sqrt{T}}$

5) Let $f(x) = \|Ax - b\|^2$. Find the smallest L s.t. $\nabla f(x)$ is L-Lipschitz

$\|Ax - b\|^2$ is a quadratic $\Rightarrow$ its Lipschitz constant is

$2\lambda_{max}(A^T A)$ [crossed out] $= 2 \cdot 3^2 = 18$

$= 2(\lambda_{max}(A))^2$

For symmetric A,

$\|A\| = \lambda_{max}(A)$

PRACTICE SOLUTIONS

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

$A \succeq B \Rightarrow A - B$ is psd

$\nabla^2 f(x) \succeq mI \Rightarrow$ every eigenvalue of $\nabla^2 f(x) \ge m$.

$\Rightarrow$ largest eigenval of $(\nabla^2 f(x))^{-1}$ is $\frac{1}{m}$ at most

$\nabla^2 f(x) \preceq LI \Rightarrow$ every eigenval of $\nabla^2 f(x) \le L$

$\Rightarrow$ smallest eigenval of $(\nabla^2 f(x))^{-1}$ is $\ge \frac{1}{L}$

EIGENVALS OF A 2x2: $|A - \lambda I| = 0$

Fundamental Thm of Calc:

$$g(t) = g(0) + \int_0^t g'(z)\,dz$$

$$\nabla f(x + td) - \nabla f(x) = \int_0^t \nabla^2 f(x + zd)\,d\,dz$$

TRANSPOSE PROPERTIES

$(A^T)^T = A$ $\qquad (A^{-1})^T = (A^T)^{-1}$

$(A + B)^T = A^T + B^T$ $\qquad (AB)^{-1} = B^{-1}A^{-1}$

$(A - B)^T = A^T - B^T$

$(kA)^T = kA^T$

$(AB)^T = B^T A^T$

$A_i = (-1/i, 1/i)$

OPEN/CLOSED SETS:

- intersection of open sets is open if # of sets is finite
- union of open sets is open
- intersection of closed sets is closed
- union of finite # of closed sets is closed

$cl(A \cap B) \subseteq cl(A) \cap cl(B)$

$\text{int}(A \cap B) = \text{int}(A) \cap \text{int}(B)$

$\text{int}(A) \cup \text{int}(B) \subseteq \text{int}(A \cup B)$

Unit balls!

Linear dependence:

$c_1 x_1 + \dots + c_n x_n = 0$ for nonzero c_i's

rank(A) = max # lin independent columns
= lin independent rows

$Ax = b$, $A \in \mathbb{R}^{m \times n}$

overdetermined $m > n$, no solutions, find x to min $\|Ax - b\|$

underdetermined $m < n$, infinitely many solutions,
Find solution $Ax = b$ s.t. $\|x\|_2$ small as possible

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

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

Nullspace $N(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$, $N(A) \subseteq \mathbb{R}^n$

If $Ax = b$ solutions unique, $N(A) = \{0\}$
rank$(A) = n$

HOMEWORK 1

1.6 Show the general norm function $f(x) = \|x\|$ is cts over $\mathbb{R}^n$.

Must show that given $\{x_k\} \in \mathbb{R}^n$, $x_k \to x^*$, $\|x_k\| \to \|x^*\|$.

Let $[x_k]_i$ denote the ith element of x_k.

clearly $[x_k]_i - x_i^* \to 0 \ \forall i$

So, $|[x_k]_i - x_i^*| \to 0 \ \forall i$. Adding over i gives $\|x_k - x^*\|_1 \to 0$

For arbitrary norm

$\|x_k - x^*\| = \|\sum_i ([x_k]_i - x_i^*) e_i\| \le \sum |[x_k]_i - x_i^*| \cdot \|e_i\|$

$\le \sum_i |[x_k]_i - x_i^*| \cdot (\max_j \|e_j\|) \le \|x_k - x^*\|_1 \max_j \|e_j\|$

We showed $\|x_k - x^*\|_1 \to 0$, and $\max_j \|e_j\|$ is just the max of n #s

so $\|x_k - x^*\| \to 0$.

Observe $a = a - b + b$

$\|a\| \le \|a - b\| + \|b\|$ by triangle inequality

$\|a\| - \|b\| \le \|a - b\|$

$\|b\| - \|a\| \le \|a - b\|$

$|\|a\| - \|b\|| \le \|a - b\|$

So $|\|x_k\| - \|x^*\|| \le \|x_k - x^*\| \to 0$.

1.12 Let $A \in \mathbb{R}^{m \times n}$. Prove:

i) $\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\|A\|_\infty$

For any vector $x \in \mathbb{R}^m$

$\|x\|_2^2 = \sum_{i=1}^m x_i^2 \le m \max_{i=1,\dots,m} x_i^2 = m\|x\|_\infty^2$

So $\|x\|_2 \le \sqrt{m}\|x\|_\infty$

And $\|A\|_2 = \max_{\|x\|_2 \le 1} \|Ax\|_2$

$\le \sqrt{m} \max_{\|x\|_2 \le 1} \|Ax\|_2 \le \sqrt{m} \max_{\|x\|_2 \le 1} \|Ax\|_\infty \le \sqrt{m} \max_{\|x\|_\infty \le 1} \|Ax\|_\infty$

$\le \|A\|_2 \le \sqrt{m}\|A\|_\infty$

Also, $\|x\|_2^2 \ge \|x\|_\infty^2$

$\|A\|_\infty = \max_{\|x\|_\infty \le 1} \|Ax\|_\infty \le \max_{\|x\|_\infty \le 1} \|Ax\|_2 \le \max_{\|x\|_2/\sqrt{n} \le 1} \|Ax\|_2$

So $\|A\|_\infty \le \max_{\|x/\sqrt{n}\|_2 \le 1} \sqrt{n}\|Ax/\sqrt{n}\|_2$

$= \sqrt{n}\|A\|_2$

$\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\|A\|_\infty$

ii) $\frac{1}{\sqrt{m}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1$

$\|A\|_\infty = \|A^T\|_1$, $\|A\|_2 = \|A^T\|_2$

Obtain required inequality by taking i) replacing ∞ norm with 1 norm, and transposing n & m.

HOMEWORK 2

2.9 Let A be an $n \times n$ PSD matrix

i) Show for any $i \neq j$, $A_{ii}A_{jj} \ge A_{ij}^2$

Given any matrix consider the quantities $x^T A x$ where x only has nonzero entries at the ith and jth coordinates.

Then $x^T A x = [x]_{ij}^T [A]_{ij} [x]_{ij}$.

So, if A is psd, clearly A_{ij} must also be psd.

This means $\det([A]_{ij} \ge 0) \Rightarrow A_{ii}A_{jj} - A_{ij}A_{ji} \ge 0$

By symmetry of A, $A_{ij} = A_{ji}$, so $A_{ii}A_{jj} \ge A_{ij}A_{ji}$.

ii) Show if for some $i \in \{1, 2, \dots, n\}$, $A_{ii} = 0$, then the ith row of A consists of 0s.

If $A_{ii} = 0$, and $A_{ik} \neq 0$ for some k, we have

$A_{ii}A_{kk} - A_{ik}^2 < 0$, contradicting i)

2.10 Let A^α be the $n \times n$ matrix defined by

$A_{ij}^\alpha = \begin{cases} \alpha, & i = j \\ 1, & i \neq j \end{cases}$

Show it is psd iff $\alpha \ge 1$

$A = 11^T + (\alpha - 1)I$

when $\alpha \ge 1$, both are psd $\Rightarrow A$ psd

If $\alpha < 1$, let x be a nonzero vector whose entries add up to 1. Then

$x^T A x = x^T 11^T x + (\alpha - 1)x^T x = \|x\|^2 + (\alpha - 1)\|x\|^2$

[crossed out], so A not PSD

$= \|x\|^2(1 + (\alpha - 1)) < 0$

Homework 3

4.1 Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and let $\{x_k\}_{k \ge 0}$ be the sequence generated by the gradient method with a constant stepsize $t_k = 1/L$. Assume $x_k \to x^*$. Show if $\nabla f(x_k) \neq 0 \ \forall k \ge 0$ then x^* is not a local max point.

We know $x_k \to x^*$ and $f(x_k)$ is strictly decreasing since $f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$

So, $f(x^*) < f(x_k) \ \forall k$, otherwise if

$f(x^*) \ge f(x_l)$ for some l, it would mean $f(x^*) > f(x_{l+1}) > f(x_{l+2}) > \dots$

and $f(x_k)$ does not converge to $f(x^*)$, contradicting continuity of f.

$f(x^*) < f(x_k) \ \forall k$ and $x_k \to x^*$ implies for any $\varepsilon > 0$, $B(x^*, \varepsilon)$ contains some x_k s.t. $f(x_k) > f(x^*)$, so x^* cannot be a local max.

4.10 Give an example of $f \in C_L^{1,1}(\mathbb{R})$, starting point x_0 s.t. $\min f(x)$ has an optimal solution and the gradient method w/ constant stepsize $t = 2/L$ diverges

Let $f(x) = x^2$. Then $L = 2$, $2/L = 1$. Gradient descent update: $x_{k+1} = x_k - 1 \cdot 2x_k = -x_k$, which diverges!

5.3 Let f be a twice cts differentiable function satisfying $LI \succeq \nabla^2 f(x) \succeq mI$ for some $L > m > 0$. Let x^* be the unique minimizer of f over $\mathbb{R}^n$.

i) Show that $f(x) - f(x^*) \ge \frac{m}{2}\|x - x^*\|^2$ for any $x \in \mathbb{R}^n$

→ Using Taylor's theorem,

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

Since $\nabla f(x^*) = 0$, $\nabla^2 f(x) \succeq mI$,

$f(x) \ge f(x^*) + \frac{m}{2}\|x - x^*\|^2$

(ii) Let $\{x_k\}_{k \ge 0}$ be the sequence generated by the damped Newton's method w/ constant stepsize $t_k = m/L$. Show

$f(x_k) - f(x_{k+1}) \ge \frac{m}{2L}\nabla f(x_k)^T(\nabla^2 f(x_k))^{-1}\nabla f(x_k)$

Lemma: for PSD matrix A, $x^T A^2 x \le \lambda_{max}(A)x^T A x$

If A was diagonal, this is clear b/c

$x^T A x = \sum_i d_{ii} x_i^2$, $x^T A^2 x = \sum_i d_{ii}^2 x_i^2$

Since $d_{11}, \dots, d_{nn}$ are the eigenvalues of A the desired statement holds

$x^T A^2 x = \sum_i d_{ii}^2 x_i^2 \le \max_i d_{ii} \sum_i d_{ii} x_i^2 = \lambda_{max}(A)x^T A x$

If A not diagonal, we can diagonalize it as

$A = U^T D U$, where D is the diagonal matrix of eigenvals of A and U orthogonal. Then

$x^T A x = x^T U^T D U x = y^T D y$, where $y = Ux$. Thus $x^T A x = \sum_i d_{ii} y_i^2$

Also, $A^2 = (U^T D U)(U^T D U) = U^T D^2 U$, so $x^T A^2 x = \sum_i d_{ii}^2 y_i^2$ and the proof proceeds as in the diagonal case.

Returning to the problem, we have $x_{k+1} = x_k - \frac{m}{L}[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$

Using descent lemma:

$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$

$= f(x_k) - \frac{m}{L}\nabla f(x_k)[\nabla^2 f(x_k)]^{-1}\nabla f(x_k) + \frac{m^2}{2L}\nabla f(x_k)^T[\nabla^2 f(x_k)]^{-2}\nabla f(x_k)$

From $\nabla^2 f(x) \succeq mI$, $\nabla^2 f(x)^{-1} \preceq \frac{1}{m}I$, so largest eigenvalue of $[\nabla^2 f(x_k)]^{-1}$ is at most $1/m$, we have from the lemma

$z^T[\nabla^2 f(x_k)]^{-2}z \le z^T[\nabla^2 f(x_k)]^{-1}z\left(\frac{1}{m}\right) \ \forall z$

$m z^T[\nabla^2 f(x_k)]^{-2}z \le z^T[\nabla^2 f(x_k)]^{-1}z$

So $f(x_{k+1}) \le f(x_k) - \frac{m}{L}\nabla f(x_k)[\nabla^2 f(x_k)]^{-1}\nabla f(x_k) + \frac{m}{2L}\nabla f(x_k)^T[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$

$= f(x_k) - \frac{m}{2L}\nabla f(x_k)[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$

So $f(x_k) - f(x_{k+1}) \ge \frac{m}{2L}f(x_k)[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$

iii) ~~Show~~ Show $x_k \to x^*$ as $k \to \infty$

Since $\lambda_{min}(\nabla^2 f(x)^{-1}) \ge 1/L$, we have $f(x_k) - f(x_{k+1}) \ge \frac{m}{2L^2}\|\nabla f(x_k)\|^2$

So $\nabla f(x_k) \to 0$ (else $f(x_k) \to -\infty$).

x_k lies within a bounded region since $f(x_k)$ is nonincreasing, and part i) tells us $\{x \mid f(x) \le f(x_0)\}$ is bounded.

So $x_k \to x^*$. If not, $\exists\, \varepsilon > 0$ s.t. infinitely many x_k are outside $B(x^*, \varepsilon)$. Since the x_k chosen are within a bounded region, $\exists$ a convergent subsequence of them. It must converge to a $y \neq x^*$. $\Rightarrow \nabla f(y) = 0$ by continuity