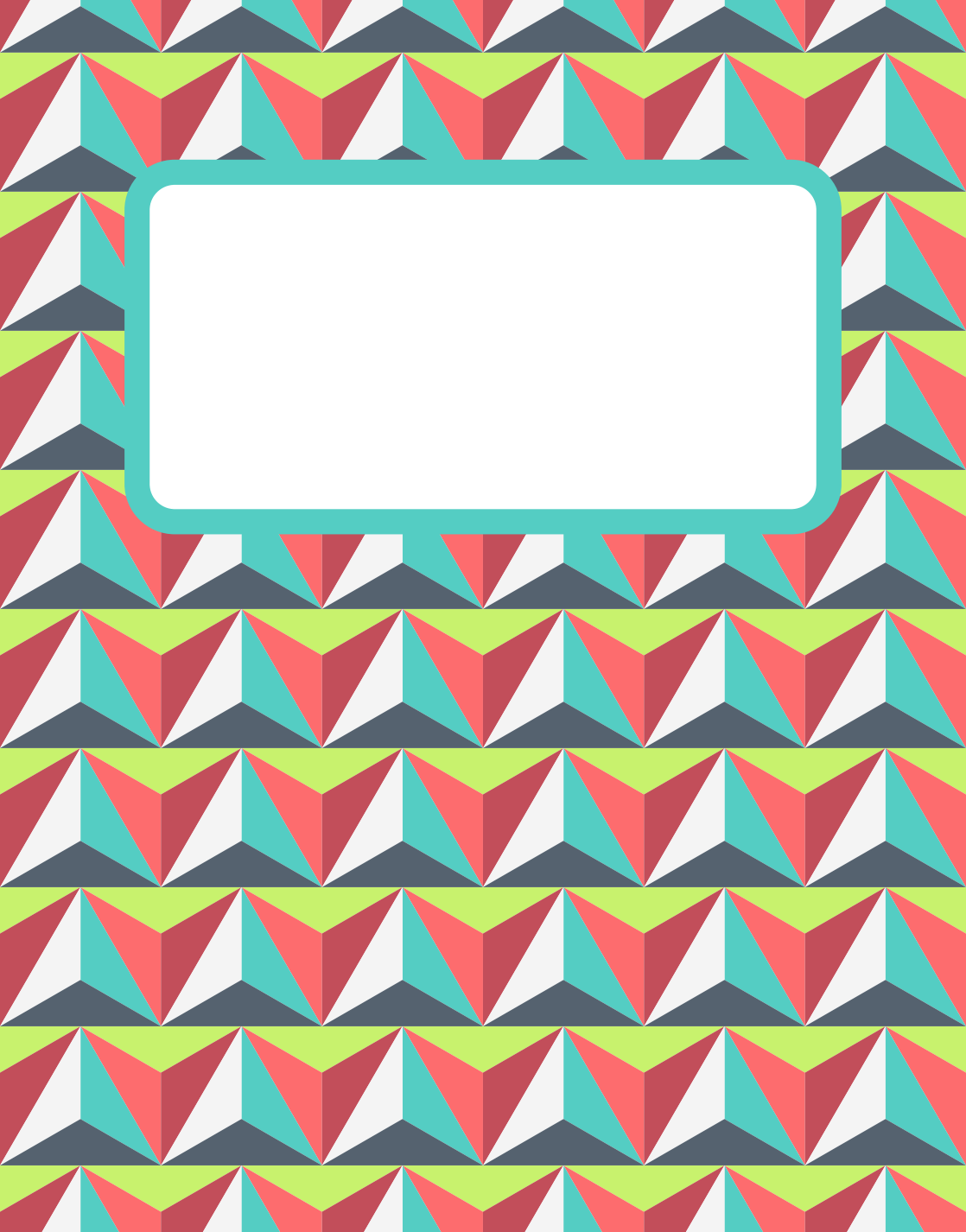






Math 408

Course: Nonlinear Optimization

Webpage: [sites.math.washington.edu/
~dprusv/crs/Math_408_2019/
homepage.html](https://sites.math.washington.edu/~dprusv/crs/Math_408_2019/homepage.html)

(easier: link from my webpage)

Instructor: Dmitry Prusvyatsky (PDL C-434)

Office Hours: TBA

TA: Andrew Pryhuber (PDL C-8M)

Office Hours: TBA

Grading:

- ① HW weekly assignments 35%
- ② Midterm (May 3rd) 30%
- ③ Final Exam (8:30-10:20am, June 12, ECE 105)
35%

Linear Algebra Review:

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

rows: $A_{1\cdot}, A_{2\cdot}, \dots, A_{m\cdot}$

columns: $A_{\cdot 1}, A_{\cdot 2}, \dots, A_{\cdot n}$

Ex: $A = \begin{bmatrix} 3 & 2 \\ -2 & 1 \\ 4 & 3 \end{bmatrix}$

$$A_{3\cdot} = [4 \ 3], \quad A_{\cdot 2} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

Matrix-vector Multiplication

Let $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$

Row space view:

Then $x^T y \triangleq x_1 y_1 + x_2 y_2 + \dots + x_n y_n$

and

$$Ax = \begin{bmatrix} A_{1\cdot} x \\ A_{2\cdot} x \\ \vdots \\ A_{m\cdot} x \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n A_{1j} x_j \\ \vdots \\ \sum_{j=1}^n A_{mj} x_j \end{bmatrix}$$

Column Space View:

$$Ax = x_1 A_{\cdot 1} + x_2 A_{\cdot 2} + \dots + x_n A_{\cdot n}$$

linear combination of cols.

Equivalent Definitions!

Matrix Matrix Multiplication

Fix $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times k}$

$$AB \triangleq [AB_{\cdot 1} \ AB_{\cdot 2} \ \dots \ AB_{\cdot k}]$$

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

Equivalently

$$[AB]_{ij} = A_{i \cdot} B_{\cdot j}$$

Four Fundamental Subspaces

Defn: $W \subseteq \mathbb{R}^n$ is called a subspace if
 $ax + by \in W$ for all $x, y \in W$
 $a, b \in \mathbb{R}$.

[Ex: Fix $v_1, \dots, v_k \in \mathbb{R}^n$. Then

① $\text{span}(v_1, \dots, v_k) \triangleq$
 $\{a_1 v_1 + \dots + a_k v_k : a_1, \dots, a_k \in \mathbb{R}\}$
is a subspace.

② For any set $S \subseteq \mathbb{R}^n$, the
orthogonal complement
 $S^\perp \triangleq \{v \in \mathbb{R}^n : v^T x = 0 \ \forall x \in S\}$
is a subspace.

Thm: If $S \subseteq \mathbb{R}^n$ is a subspace,
Then
 $\mathbb{R}^n = S + S^\perp$ and $(S^\perp)^\perp = S$

Let $A \in \mathbb{R}^{m \times n}$. Define the
four fundamental subspaces.

$$\text{range}(A) \triangleq \{Ax : x \in \mathbb{R}^n\}$$
$$= \text{span}\{A_{\cdot 1}, \dots, A_{\cdot n}\}$$

$$\text{Null}(A) \triangleq \{x \in \mathbb{R}^n : Ax = 0\}$$
$$= \{A_{\cdot 1}, A_{\cdot 2}, \dots, A_{\cdot m}\}^\perp$$

$$\text{range}(A^T) = \dots$$

$$\text{Null}(A^T) = \dots$$

Observation:

$$\begin{aligned}\text{Null}(A) &= \{A_1, A_2, \dots, A_m\}^\perp \\ &= \text{span}\{A_1, A_2, \dots, A_m\}^\perp \\ &= \text{Range}(A^T)^\perp\end{aligned}$$

Thm: $\text{Null}(A) = \text{Range}(A^T)^\perp$
 $\text{Null}(A^T) = \text{Range}(A)^\perp$

Gaussian Elimination:

Review it if you don't remember!

Eigenvalue Decomposition

Fix a matrix $A \in \mathbb{R}^{n \times n}$.

Def: We say λ is an eigenvalue of A if $\exists 0 \neq x$ with

$$Ax = \lambda x$$

$$(\Leftrightarrow (A - \lambda I)x = 0$$

We call x an eigenvector corresponding to λ .

Thus $\{\text{eigenvalues of } A\}$

$\parallel$

$$\{\lambda: \det(A - \lambda I) = 0\}$$

polynomial in λ .

There are n eigenvalues $\lambda \in \mathbb{C}$ counting multiplicity.

Defn: Symmetric Matrices

$$\mathcal{S}^n = \{A \in \mathbb{R}^{n \times n} : A^T = A\}$$

Orthogonal Matrices

$$\mathcal{O}^n = \{U \in \mathbb{R}^{n \times n} : U^T U = I\}$$

Thm: A matrix $A \in \mathcal{S}^n$ has only real eigenvalues and for each real eigenvalue there is a real eigenvector. Moreover if (λ_1, x_1) and (λ_2, x_2) are eigenpairs with $\lambda_1 \neq \lambda_2$, then $x_1^T x_2 = 0$.

Thm: [Eigenvalue Decomposition] Let $A \in \mathcal{S}^n$. Then there exists $U \in \mathcal{O}^n$ st.

$$A = U \Lambda U^T \text{ where } \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}.$$

Moreover, in any such decomposition $(\lambda_i, u_{:,i})$ are eigenpairs.

Remark: $A = U \Lambda U^T = \sum_{i=1}^n \lambda_i u_i u_i^T$
 rank 1 decomposition.

Quadratic forms:

Defn: A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ of the form $f(x) = \sum_{i,j=1}^n a_{ij} x_i x_j$ for some $a_{ij} \in \mathbb{R}$ is called a quadratic form.

Rem: Define the matrix x
 $A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \in \mathbb{R}^{n \times n}$

Then

$$f(x) = \sum_{i,j=1}^n a_{ij} x_i x_j = x^T A x$$

We can always assume $A \in \mathcal{S}^n$
 since

$$\begin{aligned} x^T A x &= \frac{1}{2} (x^T A x) + \frac{1}{2} (x^T A x)^T \\ &= \frac{1}{2} x^T A x + \frac{1}{2} x^T A^T x = x^T \left(\frac{A + A^T}{2} \right) x \end{aligned}$$

So $A \in \mathbb{R}^{n \times n}$ and $\frac{A + A^T}{2} \in \mathcal{S}^n$ define the same quadratic form.

For any $x \in \mathbb{R}^n$, define $\|x\| = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}$

Corollary: [Rayleigh-Ritz] Let $A \in \mathbb{S}^n$.

Then $\lambda_{\min}(A) \leq \frac{x^T A x}{\|x\|^2} \leq \lambda_{\max}(A) \quad \forall x$

Equality holds on the left, iff x is an eigenvector corresponding to $\lambda_{\min}(A)$. Similarly on the right

pf: Fix an eigenvalue decomposition

$$A = U \Lambda U^T$$

Fix $x \in \mathbb{R}^n$. Then $\exists z \in \mathbb{R}^n$ s.t. $x = Uz$

Observe $\|x\|^2 = \|Uz\|^2 = (Uz)^T Uz = z^T U^T U z = \|z\|^2$

and therefore

$$\sum_{i=1}^n \lambda_{\max}(A) z_i^2 = \lambda_{\max}(A) \|z\|^2$$

(*)

$$x^T A x = z^T U^T (U \Lambda U^T) Uz = \sum_{i=1}^n \lambda_i z_i^2$$

$$\sum_{i=1}^n \lambda_{\min}(A) z_i^2 = \lambda_{\min}(A) \|z\|^2$$

If x is an eigenvector corresponding to $\lambda_{\min}(A)$, then

$$x^T(Ax) = x^T(\lambda_{\min}(A)x) = \lambda_{\min}(A) \|x\|^2$$

Conversely, suppose

$$x^T Ax = \lambda_{\min}(A) \|x\|^2$$

Then from $\star$ we learn

$$\lambda_{\min}(A) z_i^2 = \lambda_i(A) z_i^2 \quad \forall i$$

Suppose $\lambda_{\min}(A) \neq \lambda_i(A)$ for some i

Then $z_i = 0$

$$(U^{\parallel T} x)_i = U_{\bullet i} x$$

$$\Rightarrow x \in \left\{ U_{\bullet i} : \lambda_i \neq \lambda_{\min}(A) \right\}^\perp$$

$$\text{span} \left\{ U_{\bullet i}^{\parallel} : \lambda_i = \lambda_{\min}(A) \right\}$$

{eigenvectors corresponding to $\lambda_{\min}(A)$ } $\square$

Defn: Let $A \in \mathbb{S}^n$

① A is positive semidefinite, denoted $A \succeq 0$, if $x^T A x \geq 0 \quad \forall x$.

② A is positive definite, denoted $A \succ 0$, if $x^T A x > 0 \quad \forall \neq x$

③ " " negative semidefinite " $A \preceq 0$, if $x^T A x \leq 0 \quad \forall x$

④ " " negative definite " $A \prec 0$, if $x^T A x < 0 \quad \forall \neq x$

⑤ A is indefinite if it is neither positive nor negative semidefinite.

Thm. $A \succeq 0$ iff $\lambda_{\min}(A) \geq 0$

$A \preceq 0$ iff $\lambda_{\max}(A) \leq 0$

pt: A indefinite, iff $\lambda_{\min}(A) < 0 < \lambda_{\max}(A)$
pt: Apply Rayleigh-Ritz $\square$

Thm.

$A \succeq 0$ iff $\exists L$ s.t. $A = LL^T$

pt: If $A = LL^T$, then

$$x^T A x = x^T L L^T x = (L^T x)^T L^T x = \|L^T x\|^2 \geq 0$$

$\Rightarrow A \succeq 0$

Conversely suppose $A \succeq 0$, and fix an eigendecomposition $A = U \Lambda U^T$

Set $\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \sqrt{\lambda_2} & \\ & & \ddots \\ & & & \sqrt{\lambda_n} \end{bmatrix}$ and

$$L = U \Lambda^{1/2} U^T \quad \square$$

Unconstrained Quadratic Optimization

We will be studying problems of the form:

$$\textcircled{*} \min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} x^T A x + g^T x$$

where $A \in \mathbb{R}^{n \times n}$ and $g \in \mathbb{R}^n$

Recall we can always replace $A \in \mathbb{R}^{n \times n}$ with $\frac{A+A^T}{2} \in \mathcal{S}^n$

Defn: $\bar{x}$ is a local soln of f if

$\exists \epsilon > 0$ s.t. $f(x) \geq f(\bar{x})$ for

all x satisfying $\|x - \bar{x}\| \leq \epsilon$.

(Recall $\|y\| = \sqrt{y^T y} = \sqrt{\sum_i y_i^2}$)

Thm: (Existence and Uniqueness)

Let $A \in \mathbb{S}^n$ and $g \in \mathbb{R}^n$. Then the following are true

- 1) A local soln to $(*)$ exists, iff $A \succ 0$ and $\exists \bar{x}$ s.t. $A\bar{x} + g = 0$, in which case $\bar{x}$ is a global soln of $(*)$.
- 2) If $\bar{x}$ is a local soln of $(*)$, then it is a global soln of $(*)$.
- 3) The problem has a unique global soln, iff $A \succ 0$, in which case this soln is $\bar{x} = -A^{-1}g$.
- 4) If either A is not positive semidefinite or there is no soln to $Ax + g = 0$, then
$$-\infty = \inf_x \frac{1}{2} x^T A x + g^T x.$$

proof: Suppose $\bar{x}$ is a local soln. Fix $v \in \mathbb{R}^n$
 $t \in \mathbb{R}$.

Then

$$\begin{aligned} f(\bar{x} + tv) &= \frac{1}{2} (\bar{x} + tv)^T A (\bar{x} + tv) + g^T (\bar{x} + tv) \\ &= \frac{1}{2} \bar{x}^T A \bar{x} + t v^T A \bar{x} + \frac{1}{2} t^2 v^T A v + g^T \bar{x} + t g^T v \\ &= \underbrace{f(\bar{x})}_{\text{red}} + t [v^T A \bar{x} + g^T v] + \underbrace{\frac{1}{2} t^2 v^T A v}_{\text{red}} \\ &= f(\bar{x}) + t v^T [A \bar{x} + g] + \frac{1}{2} t^2 v^T A v \end{aligned}$$

For all small $t > 0$, then

~~0~~ $0 \leq f(\bar{x} + tv) - f(\bar{x}) = t v^T [A \bar{x} + g] + \frac{1}{2} t^2 v^T A v$
Divide by t and let $t \searrow 0$.

$$\Rightarrow 0 \leq v^T [A \bar{x} + g] \quad \forall v$$

Set $v = -(A \bar{x} + g) \Rightarrow A \bar{x} + g = 0$

Back to ~~0~~:

$$0 \leq \frac{1}{2} t^2 v^T A v \quad \forall t \in \mathbb{R}, v \in \mathbb{R}^n$$

$$\Rightarrow v^T A v \geq 0 \quad \forall v \Rightarrow A \succeq 0.$$

Conversely suppose $A \succeq 0$ and $A\bar{x} + g = 0$ for some $\bar{x}$. From previous computation,

$$f(\bar{x} + tv) = f(\bar{x}) + tv^T[A\bar{x} + g] + \frac{1}{2}t^2 v^T A v$$

$$\stackrel{t \in \mathbb{R}, v \in \mathbb{R}^n}{=} f(\bar{x}) + \frac{1}{2}t^2 v^T A v \geq f(\bar{x})$$

So $\bar{x}$ is a global minimizer of f .

② Clear from ①.

③ Suppose $A \succ 0$, then by ①

$\bar{x} = -A^{-1}g$ is a global soln.

Conversely, suppose there is a unique soln. Then ① implies

$A \succeq 0$ and all solns are characterized by $Ax + g = 0 \iff$ uniqueness A invertible.

④ Suppose A is not positive semidefinite.

Let v be an e-vector corresponding to e-val $\lambda < 0$. Then

$$f(tv) = \frac{1}{2}t^2 v^T A v + tg^T v = \underbrace{\frac{\lambda}{2} \|v\|^2}_{< 0} t^2 + t(g^T v)$$

Let $t \rightarrow \infty$. Then $f(tv) \rightarrow -\infty$. $\square$

Suppose there is no soln to the system

$$Ax + g = 0$$

$$\Rightarrow g \notin \text{Range } A.$$

$$\Rightarrow \text{We can write } g = \underset{\substack{\uparrow \\ \text{Range}(A)}}}{r} + \underset{\substack{\uparrow \\ \text{Range}(A)^\perp}}}{w} \quad \text{#}^0$$

Null A.

Hence

$$f(tw) = \frac{1}{2} (tw)^T A (tw) + t g^T w = t g^T w$$

Notice $g^T w = \cancel{v^T w} + \|w\|^2 > 0.$

$$\text{Let } t \rightarrow -\infty$$

$$\Rightarrow f(tw) \rightarrow -\infty$$

□

Minimization of a quadratic over affine set

Consider

$$\begin{array}{ll} \min & \frac{1}{2} x^T A x + g^T x \\ \text{s.t.} & Bx = c \end{array}$$

Reduction:

$$\text{Find } x_p \text{ s.t. } Bx_p = c.$$

$$\text{Then } \{x: Bx = c\} = x_p + \text{Null}(B)$$

$$\text{Let } H = [H_1 \ H_2 \ \dots \ H_k]$$

↑
basis for $\text{Null}(B)$.

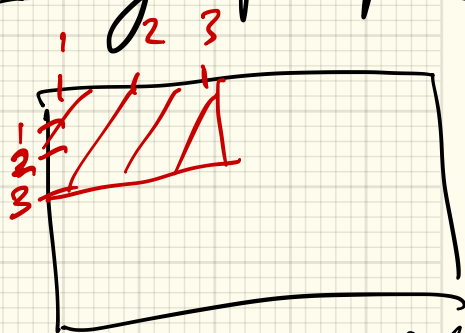
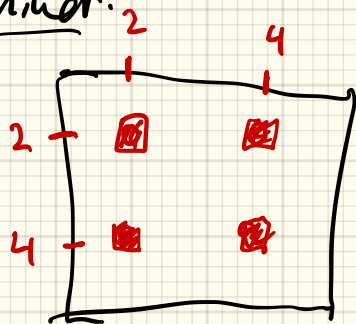
Then the problem is equivalent to

$$\min_y \frac{1}{2} (x_p + Hy)^T A (x_p + Hy) + g^T (x_p + Hy)$$

Unconstrained quadratic optimization problem!

Principal Minor Test

Def: Let $A \in \mathbb{R}^{n \times n}$. For any set $I \subseteq \{1, \dots, n\}$, the determinant of the $I \times I$ submatrix is called a principal minor.
If $I = \{1, \dots, k\}$ for some k , then it is called a leading principal minor.



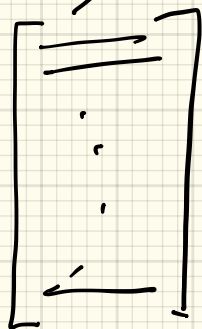
Remark: There are $2^n - 1$ principal minors and n leading principal minors.

Thm: Let $A \in \mathbb{S}^n$. Then the following are true

- ① $A \succeq 0 \Leftrightarrow$ all principal minors ≥ 0
- ② $A \succ 0 \Leftrightarrow$ all leading principal minors > 0 .

Linear Least Squares:

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



tall and skinny

Suppose we wish to solve

$$Ax = b$$

Usually impossible if errors in b

Linear Least squares Problem

$$\begin{aligned} \min_x \frac{1}{2} \|Ax - b\|^2 &= (Ax - b)^T (Ax - b) \\ &= \frac{1}{2} [x^T A^T A x - x^T A^T b - b^T A x + \|b\|^2] \\ &= \frac{1}{2} x^T A^T A x - (A^T b)^T x + \frac{1}{2} \|b\|^2 \end{aligned}$$

So this is a quadratic optimization problem.

Thm: The problem $\min_x \frac{1}{2} \|Ax - b\|^2$ always has a global minimizer and every local minimizer is global. The minimizers of $\frac{1}{2} \|Ax - b\|^2$ are precisely the solutions of the normal equation

$$A^T A x = A^T b$$

pf: Since this is a quadratic optimization problem, every local minimizer is global, and is characterized by $A^T A x - A^T b = 0$ and $A^T A \geq 0$. Notice from previous theorems $A^T A$ is always positive semidefinite while $A^T A x = A^T b$ has a solution, since otherwise $-\infty = \text{opt}(\frac{1}{2} \|Ax - b\|^2) \geq 0$, a contradiction. $\square$

A more direct argument for existence

Lemma: $\text{Null}(A^T A) = \text{Null}(A)$

$$\text{Range}(A^T A) = \text{Range}(A^T)$$

pf: Suppose $x \in \text{Null}(A)$. Then

$$A^T A x = A^T (A x) = A^T \cdot 0 = 0$$

$$\Rightarrow x \in \text{Null}(A^T A)$$

Conversely suppose $A^T A x = 0$

$$\Rightarrow 0 = x^T A^T A x = \|A x\|^2 \Rightarrow A x = 0$$

Thus $\text{Null}(A^T A) = \text{Null}(A) \Rightarrow x \in \text{Null}(A)$

and

$$\text{Range}(A^T A) = \text{Range}(A^T A^T A)$$

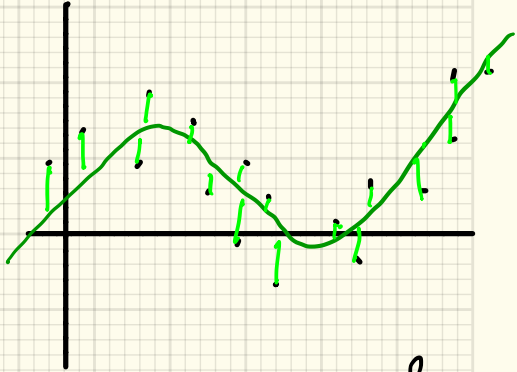
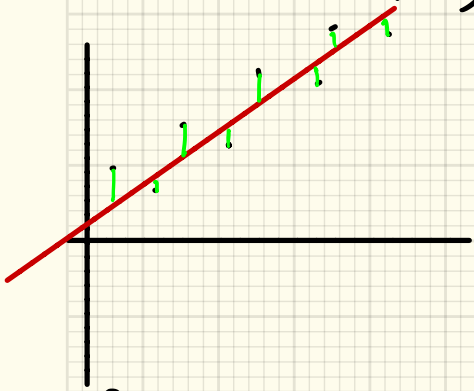
$$= (\text{Null}(A^T A))^{\perp} = (\text{Null}(A))^{\perp}$$

$$= \text{Range } A^T$$

Q

Example: (Data Fitting)

Suppose we have collected data (t_i, r_i) for $i=1, \dots, m$



Suppose $r_i \approx p(t_i) = x_0 + x_1 t_i + x_2 t_i^2 + \dots + x_d t_i^d$

Possible formulations:

Least Angle Regression

$$\min_{x \in \mathbb{R}^{d+1}} \sum_{i=1}^m |p(t_i) - r_i|$$

Useful when there are gross corruptions in data, but not smooth.

Least Squares Regression

$$\min_{x \in \mathbb{R}^{d+1}} \sum_{i=1}^m (p(t_i) - r_i)^2 = \sum_{i=1}^m (x_0 + x_1 t_i + x_2 t_i^2 + \dots + x_d t_i^d - r_i)^2$$

$$= \left\| \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^d \\ 1 & t_2 & t_2^2 & \dots & t_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \dots & t_m^d \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} - \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} \right\|_2^2$$

When d is large we can overfit the data



Common technique: penalize the function if the coefficients are large

$$\min_x \left\| \begin{bmatrix} 1 & t_1 & t_1^2 & \dots & t_1^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \dots & t_m^d \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} - \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} \right\|^2 + \lambda \cdot \left\| \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \right\|^2$$

positive regularization parameter.

In general, quadratically regularized least squares problems take the form
min_x $f(x) := \|Ax - b\|^2 + \lambda \|Dx\|^2$
where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $D \in \mathbb{R}^{n \times k}$,
 $\lambda > 0$.

Observe

$$\begin{aligned} f(x) &= (Ax - b)^T (Ax - b) + \lambda (Dx)^T (Dx) \\ &= x^T (A^T A + \lambda D^T D) x - 2(A^T b)^T x + \|b\|^2 \end{aligned}$$

Clearly $A^T A + D^T D \succeq 0$.

Therefore minimizers x
are characterized by

$$(A^T A + D^T D)x = A^T b$$

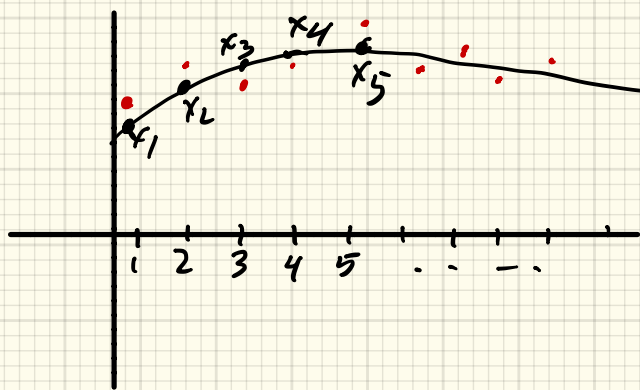
Example: (Denoising)

Suppose b is a good estimate of x , that is

$$b = \underset{\substack{\uparrow \\ \text{signal}}}{x} + \underset{\substack{\uparrow \\ \text{noise}}}{w}$$

observation

Goal: Find a good estimate of x , with the prior information that x is a discretization of a smooth function



- value of b_i
- value of x_i

Common formulation:

$$\min_{x \in \mathbb{R}^n} \|x - b\|^2 + \lambda \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2$$

forces x_i to vary smoothly across indices.

Notice

$$\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 = \|D \cdot x\|^2 \quad \text{where}$$

$$D = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

So solution of the problem is

$$\bar{x} = (I + \lambda D^T D)^{-1} b$$

See the numerics in Section 3.4 of Beck.

Orthogonal Projection as Linear Least Squares.

Defn: Consider a subspace S .

Then the solution of the problem with $\|x - b\|^2$, denoted $P_S(b)^*$ is called the orthogonal projection onto S .

Let $A = [A_1 \ A_2 \ \dots \ A_m]$ be such that $\text{Range}(A) = \text{span}(A_1, A_2, \dots, A_m) = S$.

Then to find $P_S(b)$, we must solve

$$\min_{x \in S} \|x - b\|^2 = \min_{y \in \mathbb{R}^m} \|Ay - b\|^2$$

$$\Rightarrow \bar{y} = (A^T A)^{-1} A^T b \quad [\text{why } (A^T A)^{-1} \text{ exists?}]$$

$$\Rightarrow P_S(b) = A \bar{y} = A(A^T A)^{-1} A^T b.$$

Thus

$$P_S(b) = A(A^T A)^{-1} A^T b$$

Rem: In particular P_S is linear and represented as the matrix

$$A(A^T A)^{-1} A^T$$

From linear Algebra recall $b = P_S(b) + P_{S^\perp}(b)$

Therefore

$$P_{S^\perp}(b) = (I - A(A^T A)^{-1} A^T) b.$$

In particular, if $\{A_1, \dots, A_m\}$ are orthonormal then $A^T A = I$ and $P_S(b) = A A^T b$ and $P_{S^\perp}(b) = (I - A A^T) b$.

Analysis Review

Notation:

Open ball $B_r(x) = \{y \in \mathbb{R}^n : \|x-y\| < r\}$

Defn: Let $Q \subseteq \mathbb{R}^n$.

- Q is open if for every $x \in Q$, there is $r > 0$, such that $B_r(x) \subseteq Q$.
- Q is closed if for any $x_i \in Q$ converging to $\bar{x} \in \mathbb{R}^n$, the point $\bar{x}$ lies in Q .
- Q is bounded if $\exists R > 0$ s.t. $Q \subseteq B_R(0)$.
- Q is compact if Q is closed and bounded.
- $\bar{x}$ is a limit point of a sequence x_i if there exists subsequence $x_{i_j} \rightarrow \bar{x}$.

Thm: Q is compact, iff every infinite sequence in Q has a limit point in Q .

Defn: A function $f: Q \rightarrow \mathbb{R}$ is continuous if for every $\bar{x} \in Q$ we have $\lim_{x \rightarrow \bar{x}} f(x) = f(\bar{x})$.

Thm: Every continuous function on a compact set Q attains its infimum and supremum over Q .

Differentiation

We will use the notation $F(h) = o(\|h\|) \Leftrightarrow \lim_{h \rightarrow 0} \frac{\|F(h)\|}{\|h\|} = 0$

Def: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $\bar{x} \in \mathbb{R}^n$. We say f is differentiable at $\bar{x}$ if $\exists v$ s.t.
 $f(\bar{x}+h) = f(\bar{x}) + v^T h + o(\|h\|)$

Then we call v the gradient of f at $\bar{x}$, and we denote it by $\nabla f(\bar{x})$

Thm: If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable at $\bar{x}$, then
 $\nabla f(\bar{x}) = \left(\frac{\partial f}{\partial x_1}(\bar{x}), \dots, \frac{\partial f}{\partial x_n}(\bar{x}) \right)^T$

We say a map $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is
differentiable at $\bar{x}$ if $\exists A \in \mathbb{R}^{m \times n}$ s.t.

$$F(\bar{x}+h) = F(\bar{x}) + Ah + o(\|h\|)$$

Then we use the notation $A = \nabla F(\bar{x})$.

Thm: If F is differentiable at $\bar{x}$, then

$$\nabla F(\bar{x}) = \left[\frac{\partial F_i}{\partial x_j}(\bar{x}) \right]_{\substack{i=1, \dots, m \\ j=1, \dots, n}}$$

Jacobian

The second derivative of $f: \mathbb{R}^n \rightarrow \mathbb{R}$
is $\nabla^2 f(x) := \nabla(\nabla f)(x)$, that is

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right]_{i,j=1, \dots, n}$$

Hessian matrix

Thm: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be such that all
partials $\frac{\partial^2 f}{\partial x_i \partial x_j}$ exist and are continuous.
Then $\nabla^2 f(x)$ is a real symmetric matrix.

Ex: Let's compute ∇f , $\nabla^2 f$ of the function $f(x) = \frac{1}{2} x^T A x + g^T x$ where $A \in \mathbb{S}^n$, $g \in \mathbb{R}^n$.

$$\begin{aligned} f(x+h) &= \frac{1}{2} (x+h)^T A (x+h) + g^T (x+h) \\ &= \frac{1}{2} x^T A x + \frac{1}{2} x^T A h + \frac{1}{2} h^T A x + \frac{1}{2} h^T A h + g^T x + g^T h \\ &= f(x) + (Ax + g)^T h + \frac{1}{2} h^T A h \\ &= f(x) + (Ax + g)^T h + o(\|h\|) \end{aligned}$$

$$\Rightarrow \nabla f(x) = Ax + g$$

$$\begin{aligned} \nabla f(x+h) &= A(x+h) + g = Ax + g + Ah + o(\|h\|) \\ &= \nabla f(x) + Ah + o(\|h\|) \end{aligned}$$

$$\Rightarrow \nabla^2 f(x) = A$$

Thm (1D-mean value thm)

Let $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ be $k+1$ times differentiable on (a, b) . Then for every $x, y \in (a, b)$, there exists $\xi \in (x, y)$ such that

$$\varphi(y) = \varphi(x) + \varphi'(\xi)(y-x) + \frac{1}{2!} \varphi''(\xi)(y-x)^2 + \frac{1}{(k+1)!} \varphi^{(k+1)}(\xi)(y-x)^{k+1}$$

Thm Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable on an open set containing a closed interval

$$[x, y] := \{(1-\lambda)x + \lambda y : \lambda \in [0, 1]\}$$

Then there exists

$$z, w \in (x, y) := \{(1-\lambda)x + \lambda y : \lambda \in (0, 1)\}$$

satisfying

$$f(y) = f(x) + \nabla f(z)^T (y-x)$$

and

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

pf: Define $\varphi(t) = f(x + t(y-x))$

$$\text{H.W: } \varphi'(t) = \nabla f(x + t(y-x))^T (y-x)$$

$$\varphi''(t) = (y-x)^T \nabla^2 f(x + t(y-x)) (y-x)$$

$$\Rightarrow \exists t_1, t_2 \in (0, 1) \text{ s.t.}$$

$$\varphi(1) = \varphi(0) + \varphi'(t_1)(1-0)$$

$$f(y) = f(x) + \nabla f(x + t_1(y-x))^T (y-x)$$

$$\text{and } \varphi(1) = \varphi(0) + \varphi'(0)(1-0) + \frac{1}{2} \varphi''(t_2)(1-0)^2$$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} (y-x)^T \nabla^2 f(x + t_2(y-x)) (y-x)$$

W R

Thm: (Quadratic Approximation)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and suppose $\nabla^2 f(x)$ exists and varies continuously.

Then

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

pf: MVT shows $\forall y \in B_\varepsilon(x) \exists w \in (x, y)$ s.t.

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} (y-x)^T \nabla^2 f(w) (y-x) \\ &= f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} (y-x)^T \nabla^2 f(x) (y-x) \\ &\quad + \frac{1}{2} (y-x)^T [\nabla^2 f(w) - \nabla^2 f(x)] (y-x) \\ &\quad o(\|y-x\|^2) \quad \square \end{aligned}$$

Optimality Conditions

Consider the problem

$$(P) \min_{x \in \mathbb{R}^n} f(x)$$

Defn: $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is coercive if for every sequence x_i for which $\|x_i\| \rightarrow +\infty$, it must be the case that $f(x_i) \rightarrow +\infty$.

Ex: $f(x) = x^2$, e^{x^2} are coercive

$f(x) = e^x$, x , x^3 are not coercive

Lemma: A continuous function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is coercive iff for every $\alpha \in \mathbb{R}$, the set $\{x: f(x) \leq \alpha\}$ is compact.

pf: continuity $\Rightarrow \{x: f(x) \leq \alpha\}$ is closed. Suppose f coercive but $L = \{x: f(x) \leq \alpha\}$ is not bounded $\Rightarrow \exists x_i \in L$ with $\|x_i\| \rightarrow +\infty$. $\Rightarrow f(x_i) \rightarrow +\infty$ $\nless \alpha$ contradiction.

For converse, see book. $\square$

Thm: If $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuous and coercive, then f has at least one global minimizer.

pf: Let $\alpha \in \mathbb{R}$ be such that $L := \{x: f(x) \leq \alpha\}$ is nonempty. Then $\arg \min_x f(x) = \arg \min_{x \in L} f(x)$ is nonempty. $\square$

Thm: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable and let $\bar{x}$ be a local minimizer of f .

Then $\nabla f(\bar{x}) = 0$

~~pf:~~ $f(\bar{x}+h) = f(\bar{x}) + \nabla f(\bar{x})^T h + o(\|h\|)$

$$\exists \epsilon > 0 \text{ s.t. } \forall h \in B_\epsilon(0), f(\bar{x}+h) \geq f(\bar{x})$$

$$\Rightarrow \nabla f(\bar{x})^T h \geq o(\|h\|)$$

$$\text{Let } h = -t \nabla f(\bar{x})$$

$$\Rightarrow -t \|\nabla f(\bar{x})\| \geq o(t)$$

Divide by t and let $t \rightarrow 0$

$$\Rightarrow \nabla f(\bar{x}) = 0 \quad \square$$

Thm: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable.

1) (Necessity) If $\bar{x}$ is a local minimizer then $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \succeq 0$.

2) (Sufficiency) If $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \succeq 0$, then $\exists \epsilon > 0$ s.t. $f(x) \geq f(\bar{x}) + \alpha \|x - \bar{x}\|^2, \forall x \in B_\epsilon(\bar{x})$

pl: ① We already proved $\nabla f(\bar{x}) = 0$.

So for small h

$$f(\bar{x}) \leq f(\bar{x}+h) = f(\bar{x}) + \cancel{\nabla f(\bar{x})^T h} + \frac{1}{2} h^T \nabla^2 f(\bar{x}) h + o(\|h\|^2)$$

$$\Rightarrow h^T \nabla^2 f(\bar{x}) h \geq o(\|h\|^2)$$

Then v be e-vector corresponding to $\lambda_{\min}$
with $\|v\|=1$

$$\Rightarrow h = tv$$
$$\Rightarrow t^2 \lambda_{\min}(\nabla^2 f(\bar{x})) \geq o(t^2)$$

Divide by t and let $t \rightarrow 0 \Rightarrow \nabla^2 f(\bar{x}) \neq 0$

② Suppose $\nabla f(\bar{x}) = 0$ and $\lambda_{\min}(\nabla^2 f(\bar{x})) \geq \mu > 0$

$$\Rightarrow f(\bar{x}+h) = f(\bar{x}) + \cancel{\nabla f(\bar{x})^T h} + \frac{1}{2} h^T \nabla^2 f(\bar{x}) h + o(\|h\|^2)$$

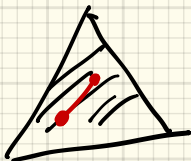
$$\geq f(\bar{x}) + \frac{\mu}{2} \|h\|^2 + o(\|h\|^2)$$

$$\exists \epsilon \text{ s.t. } |o(\|h\|^2)| \leq \frac{\mu}{4} \|h\|^2 \quad \forall h \in B_\epsilon(0)$$

$$\Rightarrow f(\bar{x}+h) \geq f(\bar{x}) + \frac{\mu}{4} \|h\|^2 \quad \square$$

Convexity

Defn: A set $Q \subseteq \mathbb{R}^n$ is convex if
 $\lambda x + (1-\lambda)y \in Q \quad \forall x, y \in Q$
 $\lambda \in [0, 1]$



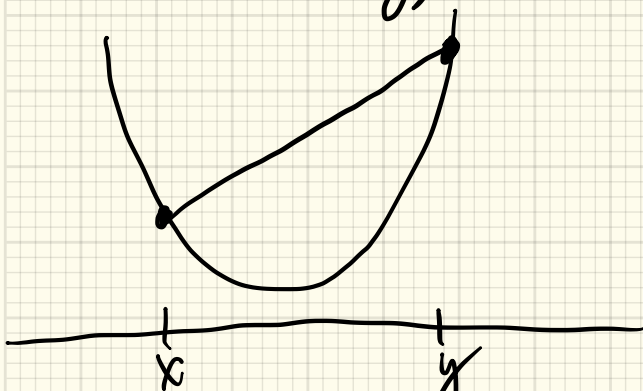
convex



not

convex.

• A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if
 $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad \forall x, y$
 $\lambda \in [0, 1]$



Ex: $c^T x$, $\|x\|$, e^x , x^2

Recall $f'(x, d) := \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t}$

Thm: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be convex.

① Then $t \rightarrow \frac{f(x+td) - f(x)}{t}$
is nondecreasing on $(0, +\infty)$



② The directional derivative $f'(x, d)$
exists and equals
 $f'(x, d) = \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t}.$

pf: ① is trivial from ①.

Now we prove ②: Let $x, d \in \mathbb{R}^n$,
 $0 < t_1 < t_2$.

$$\begin{aligned}
 \text{Then } f(x+t, d) &= f\left(x + \left(\frac{t_1}{t_2}\right) t_2 d\right) \\
 &= f\left(\left(1 - \left(\frac{t_1}{t_2}\right)\right)x + \frac{t_1}{t_2} (x+t_2 d)\right) \\
 &\leq \left(1 - \frac{t_1}{t_2}\right) f(x) + \frac{t_1}{t_2} f(x+t_2 d) \\
 \Rightarrow \underbrace{f(x+t, d) - f(x)}_{t} &\leq \underbrace{f(x+t_2 d) - f(x)}_{t_2} \quad \square
 \end{aligned}$$

Thm: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and suppose f is differentiable at $\bar{x}$. Then $\bar{x}$ is a global minimizer iff $\nabla f(\bar{x}) = 0$

pf: $\frac{f(\bar{x}+td) - f(\bar{x})}{t} \geq f'(\bar{x}, d) = \nabla f(\bar{x})^T d \geq 0 \quad \forall t, d \in \mathbb{R}$

Thm: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$

① If f is differentiable, then the following are equivalent:

② f is convex

③ $f(y) \geq f(x) + \nabla f(x)^T (y-x) \quad \forall x, y$

④ $(\nabla f(x) - \nabla f(y))^T (x-y) \geq 0 \quad \forall x, y$

② If f is twice differentiable, then f is convex, if $\nabla^2 f(x) \succeq 0 \quad \forall x$.

$$a \Rightarrow b \quad \frac{f(y) - f(x)}{1} \geq \nabla f(x)^T (y - x)$$

set $t=1, d=y-x$

b $\Rightarrow$ c From b:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

$$f(x) \geq f(y) + \nabla f(y)^T (x - y)$$

Adding $\Rightarrow$ monotonicity

c $\Rightarrow$ b Let $x, y \in \mathbb{R}^n$. By MVT, $\exists z \in (x, y)$
s.t. $\lambda \in [0, 1]$ s.t.

$$z = \lambda y + (1 - \lambda)x$$

and

$$f(y) - f(x) = \nabla f(z)^T (y - x)$$

$$\begin{aligned} \Rightarrow 0 &\leq (\nabla f(z) - \nabla f(x))^T (z - x) = \lambda (\nabla f(z) - \nabla f(x))^T (y - x) \\ &= \lambda [f(y) - f(x) - \nabla f(x)^T (y - x)] \quad \square \end{aligned}$$

b $\Rightarrow$ a, 3: see book.

Like Search Methods:

Def: Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable at $\bar{x}$. Then $d \neq 0$ is a descent direction of f at $\bar{x}$ if $\nabla f(\bar{x})^T d < 0$.

Thm: If d is a descent direction, then $f(\bar{x} + td) < f(\bar{x}) \quad \forall \text{ small } t$.

pf: $\lim_{t \rightarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} = \nabla f(\bar{x})^T d < 0$.

Ex: Steepest Descent: $d = -\nabla f(\bar{x})$

$$\Rightarrow \nabla f(\bar{x})^T d = -\|\nabla f(\bar{x})\|^2$$

Newton: $d = -\nabla^2 f(\bar{x})^{-1} \nabla f(\bar{x})$ if $\nabla^2 f(\bar{x}) \neq 0$

Comes from $f(x+h) \approx f(x) + \nabla f(x)^T h + \frac{1}{2} h^T \nabla^2 f(x) h$

Newton-like: $d = -\overset{\text{minimal}}{H} \nabla f(x)$ where $H \neq 0$.

back-track (x, d, γ, c) :

Choose $\gamma \in (0, 1)$, $c \in (0, 1)$

while $\{f(x+td) > f(x) + c t \nabla f(x)^T d\}$

$t \leftarrow \gamma t$.

return t

Line search Algorithm:

For $k \geq 0$:

Choose $t = \text{back-track}(x_k, d_k, \gamma, c)$

Set $x_{k+1} = x_k - t d_k$.

Thm. (★): Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $x_0 \in \mathbb{R}^n$ be

such that f is differentiable with

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad \forall x, y.$$

Let x_k be a sequence satisfying

$$x_{k+1} = x_k - t_k d_k$$

where d_k satisfies $\nabla f(x_k)^T d_k < 0$

and t_k is chosen by backtracking.

Then one of the following holds:

- (1) $\exists k_0$ s.t. $\nabla f(x_{k_0}) = 0$
- (2) $\{f(x_k)\}$ diverges
- (3) $\{f(x_k)\}$ is bounded $\nabla f(x_k)^T d_k \rightarrow -\infty$
- (4) For every bounded $\{c_k\}$ we have $\sum_{k \in \mathbb{N}} \nabla f(x_k)^T d_k = 0$

Examples:

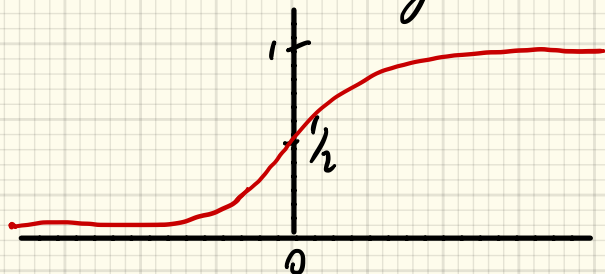
1) Logistic Regression:

Let y be a binary response variable with

$$p[y=1/x] = p(x, \theta)$$

$$p[y=0/x] = 1 - p(x, \theta)$$

Here $x \in \mathbb{R}^n$ are the predictor variables and θ tells us how to combine x together.



Assumption: $\log \frac{p}{1-p} = \theta_0 + \theta^T x$ where, $\theta \in \mathbb{R}^{n+1}$
Solving for p , yields

$$\Rightarrow p(x, \theta) = \frac{e^{-(\theta_0 + \theta^T x)}}{1 + e^{-(\theta_0 + \theta^T x)}}$$

So suppose now you observe $\{(x^i, y^i)\}_{i=1}^m$. How to determine θ ?

$$p[y = y^i | x^i] = p(x^i, \theta)^{y^i} (1 - p(x^i, \theta))^{1-y^i}$$

Define the likelihood function

$$L(\theta) = \prod_{i=1}^m p(x^i, \theta)^{y^i} (1 - p(x^i, \theta))^{1-y^i}$$

Let's maximize it in θ .

Take

$$\begin{aligned} \max_{\theta} \log(L(\theta)) &= \sum [y_i \log(p(x^i, \theta)) + (1-y_i) \log(1-p(x^i, \theta))] \\ &= \sum \left[y_i \left(\log\left(\frac{p(x^i, \theta)}{1-p(x^i, \theta)}\right) + \log(1-p(x^i, \theta)) \right) \right] \\ &= \sum \left[y_i (\theta_0 + \theta^T x^i) + \log\left(\frac{1}{1 + e^{\theta_0 + \theta^T x^i}}\right) \right] \\ &= \sum \left[y_i (\theta_0 + \theta^T x^i) - \log(1 + e^{\theta_0 + \theta^T x^i}) \right] \end{aligned}$$

↗ concave function in θ !

Again to prevent overfitting
 one may instead solve

$$\min_{\theta} \sum_{i=1}^m [\log(1 + e^{\theta_0 + \theta^T x^i}) - y_i(\theta_0 + \theta^T x^i)] + \frac{\lambda}{2} (\theta_0^2 + \|\theta\|^2)$$

↗
 This is called Logistic Regression

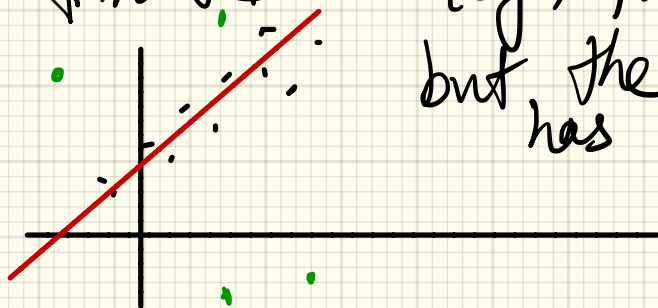
2) Huber Regression:

Suppose you want to "fit"

from data $y = a^T x$

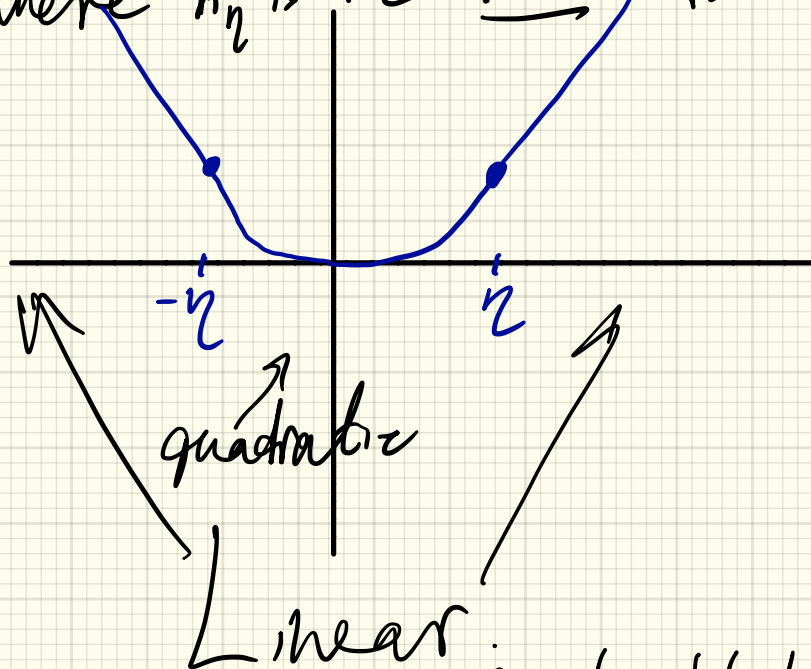
$\{(y_i, a_i)\}_{i=1}^m$

but the data has outliers.



One idea is to solve
$$\min_x \sum_{i=1}^m h_2(a_i^T x - y_i)$$

where h_2 is the Huber function



This is called Huber
regression. Again this
problem is convex.

Back to algorithms
proof of Thm 4:

Assume none of (i), (ii), (iii) and (iv)
hdd. Thus $\nabla f(x_k) \neq 0$ and
 $\{f(x_k)\}$ is a bounded decreasing
sequence. $\Rightarrow \exists l$ s.t. $\lim_{k \rightarrow \infty} f(x_k) = l$.

Now $\exists$ sequence $J \subset \mathbb{N}$ and
a vector d s.t. $d_k \xrightarrow{J} d$ and
$$\sup \nabla f(x_k)^T d_k =: \beta < 0$$

Line search guarantees $t_k \nabla f(x_k)^T d_k \rightarrow 0$

Thus $t_k \xrightarrow{J} 0$. Observe

$$f(x_k + t_k \delta^{-1} d_k) > f(x_k) + c \delta^{-1} t_k \nabla f(x_k)^T d_k$$

By MVT, $\exists \theta \in (0, 1)$ s.t.

$$f(x_k + t_k \delta^{-1} d_k) - f(x_k) = t_k \delta^{-1} \nabla f(\hat{x}_k)^T d_k$$

where

$$\hat{x}_k = (1-\theta)x_k + \theta(x_k + t_k \delta^{-1} d_k)$$
$$= x_k + \theta t_k \delta^{-1} d_k$$

$$\begin{aligned}
 \Rightarrow f(x_k + t_k \gamma^{-1} d_k) - f(x_k) &= \\
 &= t_k \gamma^{-1} \nabla f(\tilde{x}_k)^T d_k \\
 &= t_k \gamma^{-1} \nabla f(x_k)^T d_k + t_k \gamma^{-1} [\nabla f(x_k)^T - \nabla f(\tilde{x}_k)^T] d_k \\
 &\leq t_k \gamma^{-1} \nabla f(x_k)^T d_k + t_k \gamma^{-1} L \|x_k - \hat{x}_k\| \|d_k\| \\
 &= t_k \gamma^{-1} \nabla f(x_k)^T d_k + (t_k \gamma^{-1})^2 L \theta \|d_k\|^2
 \end{aligned}$$

$$\text{So } c \gamma^{-1} t_k \nabla f(x_k)^T d_k < t_k \gamma^{-1} \nabla f(x_k)^T d_k + (t_k \gamma^{-1})^2 L \theta \|d_k\|^2$$

So

$$\begin{aligned}
 0 &\leq (1-c) \nabla f(x_k)^T d_k + t_k \gamma^{-1} L \|d_k\|^2 \\
 &\leq (1-c) \beta + t_k \gamma^{-1} L \|d_k\|^2
 \end{aligned}$$

Let $k \rightarrow \infty$

$$\Rightarrow 0 \leq (1-c) \beta < 0$$

a contradiction.

□

Cor: If $\{d_k\}$ and $\{f(x_k)\}$ are bounded, then $\lim_{k \rightarrow \infty} \nabla f(x_k)^T d_k = 0$.

Cor: If $d_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$, then every limit point $\bar{x}$ of $\{x_k\}$ satisfies $\nabla f(\bar{x}) = 0$.

Cor: Suppose f is twice differentiable and there is $\beta > 0$ s.t. $\lambda_{\min}(\nabla^2 f(x)) \geq \beta$ $\forall x$. Suppose we set $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$. Then every limit point $\bar{x}$ satisfies $\nabla f(\bar{x}) = 0$.

sketch:

$$\begin{aligned} \underbrace{\|\nabla f(x_k)^T d_k\|}_{\downarrow 0} &= \left| -\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) \right| \\ &\geq \frac{1}{\beta} \|\nabla f(x_k)\|^2 \quad \square \end{aligned}$$

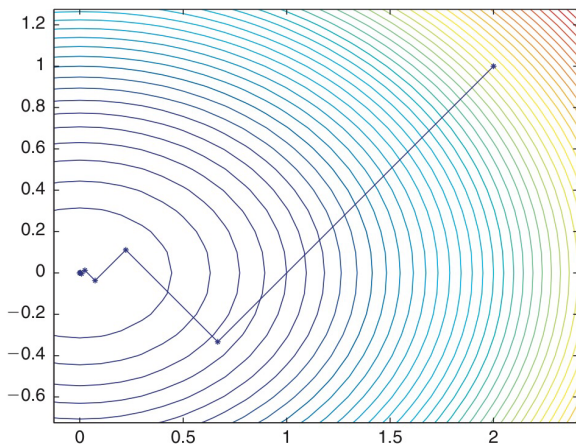
The Gradient Method:

Gradient method on

$$\min_{x,y} x^2 + 2y^2$$

with an "exact" line search

$$t_{k+1} = \arg \min_{t \geq 0} f(x_k + t d_k)$$



Lemma: Let $\{x_k\}$ be generated by gradient descent with exact line search on a differentiable $f: \mathbb{R}^n \rightarrow \mathbb{R}$.

Then

$$(x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0$$

$$\text{pf: } x_{k+2} - x_{k+1} = -t_{k+1} \nabla f(x_{k+1})$$

$$t_k = \arg \min_{t \geq 0} f(x_k - t \nabla f(x_k))$$

$$\Rightarrow \nabla f(x_k)^T \nabla f(x_{k+1}) = \frac{(x_{k+1} - x_k)^T (x_{k+2} - x_{k+1})}{-t_k - t_{k+1}} \quad \square$$

Thm: Suppose $f: \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies
 $0 < \alpha \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}(\nabla^2 f(x)) \leq \beta \quad \forall x$

Then gradient descent finds
 a point x with $\|x - x^*\|^2 \leq \epsilon$
 after $O\left(\frac{\beta}{\alpha} \ln\left(\frac{\|x_0 - x^*\|^2}{\epsilon}\right)\right)$
 iterations.

condition number

Ex: (Rosenbrock)

$$f(x, y) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

$(x_1, x_2) = (1, 1)$ is the unique stationary
 point and $\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$.

$$\Rightarrow \text{condition \#} \geq 2.5 \times 10^3$$

```
>> f=@(x) 100*(x(2)-x(1)^2)^2+(1-x(1))^2;
>> g=@(x) [-400*(x(2)-x(1)^2)*x(1)-2*(1-x(1)); 200*(x(2)-x(1)^2)];
>> [x,fun_val]=gradient_method_backtracking(f,g,[2;5],2,0.25,0.5,1e-5);
iter_number = 1 norm_grad = 118.254478 fun_val = 3.221022
iter_number = 2 norm_grad = 0.723051 fun_val = 1.496586
:
:
iter_number = 6889 norm_grad = 0.000019 fun_val = 0.000000
iter_number = 6890 norm_grad = 0.000009 fun_val = 0.000000
```

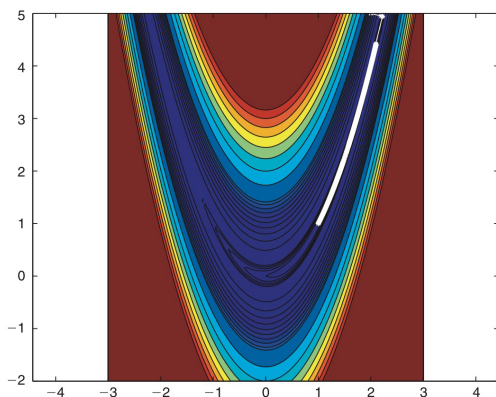


Figure 4.2. Contour lines of the Rosenbrock function along with thousands of iterations of the gradient method.

Can the convergence rate of grad method be improved?

Let's look at

$$\min_x f(x) = \frac{1}{2} x^T A x - b^T x$$

with $A \succ 0$.

Defn: Vectors $v, w \in \mathbb{R}^n$ are A-conjugate if $v^T A w = 0$.

Lem: If $d_0, d_1, \dots, d_k \neq 0$ are A-conjugate, then they are linearly independent.

pt: Suppose $0 = \sum_{i=0}^k \mu_i d_i$. Then

$$0 = \sum_{i=0}^k \mu_i d_i^T A d_i = \sum_{i=0}^k \mu_i \underbrace{d_i^T A d_i}_{\neq 0}$$

So $\mu_i = 0 \forall i$ $\square$

$\neq 0$ because $A \succ 0$.

Suppose we have available
 A -conjugate vectors $d_1, d_2, \dots, d_n$.
 and consider the sequence

$$\left\{ \begin{array}{l} t_k = \arg \min f(x_k + t d_k) \\ x_{k+1} = x_k + t_k d_k \end{array} \right\}$$

Set $r_k = b - Ax_k$

Note $r_{k+1} = r_k - t_k A d_k$

Lemma: $t_k = \frac{r_k^T d_k}{d_k^T A d_k}$

Thm: We have $r_{k+1}^T d_i = 0 \quad \forall i=1, \dots, k$
 and x_{k+1} is a minimize of f on
 $x_1 + \text{span}\{d_1, d_2, \dots, d_k\}$

pf: Induction: Assume $r_k^T d_i = 0 \quad \forall i=1, \dots, k-1$

Then $r_{k+1}^T d_k = r_k^T d_k - t_k d_k^T A d_k = 0$

For $i=1, \dots, k-1$,
 $r_{k+1}^T d_i = \underbrace{r_k^T d_i}_0 - \underbrace{d_k^T A d_i}_0 = 0$

Thus $r_{k+1} = \nabla f(x_{k+1}) \perp \{d_1, d_2, \dots, d_k\} \quad \square$

How do we generate conjugate directions on the fly?

Conjugate Gradient Method

Given x_0 ;

Set $r_0 = b - Ax_0$, $d_0 \leftarrow r_0$, $k \leftarrow 0$

While $r_k \neq 0$ do

$$t_k = \frac{r_k^T d_k}{d_k^T A d_k}$$

$$x_{k+1} = x_k + t_k d_k$$

$$r_{k+1} = b - Ax_{k+1}$$

$$\beta_{k+1} = -r_{k+1}^T A d_k / d_k^T A d_k$$

$$d_{k+1} = r_{k+1} + \beta_{k+1} d_k$$

$k \leftarrow k+1$

let's check that d_i are A -conj.

$$d_{k+1}^T A d_k = r_{k+1}^T A d_k + \beta_{k+1} d_k^T A d_k = 0$$

For $i = 1, \dots, k-1$

$$d_{k+1}^T A d_i = r_{k+1}^T A d_i + \beta_{k+1} \underbrace{d_k^T A d_i}_0$$

$$= r_{k+1}^T A d_i = r_{k+1}^T (r_i - r_{i+1}) t_i^{-1}$$

Notice $r_i \in \text{span}\{d_i, d_{i-1}\}$

$$r_{i+1} \in \text{span}\{d_i, d_{i+1}\}$$

So $r_{k+1} \perp (r_i, r_{i+1})$.

$$\Rightarrow d_{k+1}^T A d_i = 0$$

Thm: If A has m distinct eigenvalues, then CG terminates after at most m iterations.

Thm: CG will find a point x satisfying $\|x_k - x^*\| \leq \epsilon$ after at most

$$O\left(\sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}} \ln\left(\frac{\|x_0 - x^*\|^2}{\epsilon}\right)\right)$$

iterations.

Newton's Method:

Consider solving

where $G: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is differentiable

non example: $G = \nabla f$ for a twice differentiable function.

Thm: Suppose $\|DG(x) - DG(y)\| \leq L\|x - y\| \forall x, y$

Suppose at some point x , $DG(x)$ is invertible. Then

$x_N := x - [DG(x)]^{-1}G(x)$ satisfies

$$\|x_N - x^*\| \leq \frac{L}{2} \|DG(x)^{-1}\| \|x - x^*\|^2$$

where x^* is any point satisfying $G(x^*) = 0$.

pf: $x_N - x^* = x - x^* - DG(x)^{-1}G(x)$

$$= DG(x)^{-1} [DG(x)(x - x^*) + G(x^*) - G(x)]$$
$$= DG(x)^{-1} \left[\int_0^1 (DG(x) - DG(x + t(x - x^*))) (x - x^*) dt \right]$$

$$\Rightarrow \|x_N - x^*\| \leq \|DG(x)^{-1}\| \cdot \|x - x^*\| \cdot L \|x - x^*\| \cdot \int_0^1 t dt$$
$$= \frac{L}{2} \|DG(x)^{-1}\| \|x - x^*\|^2 \quad \square$$

Constrained Optimization

$$(P) \quad \min_{\text{s.t.}} \quad f(x) \\ x \in \Omega.$$

where Ω is a closed set.

Defn: Tangent Cone

$$T_{\Omega}(x) = \left\{ d : \exists \tau_i \rightarrow 0, \{x_i\} \subset \Omega, \right. \\ \left. \text{with } x_i \rightarrow x \text{ and } \tau_i^{-1}(x_i - x) \rightarrow d \right\}$$

Ex: If $\Omega = \{x : Ax = b\}$, then

$$T_{\Omega}(x) = \text{Null}(A) \quad \forall x \in \Omega.$$

• If $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$

$$\text{then } T_{\Omega}(x, y) = \begin{bmatrix} x \\ y \end{bmatrix}^{\perp}.$$

Thm: If Ω is convex, then

$$T_{\Omega}(x) = \text{cl} \left\{ \lambda(y - x) : \lambda \geq 0, y \in \Omega \right\}$$

Thm: Suppose $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable.

If $\bar{x}$ is a local soln to (P), then $f'(\bar{x}, d) \geq 0 \quad \forall d \in T_{\Omega}(\bar{x})$

pf: By MVT,

$$f'(\bar{x}, d) = \lim_{t \downarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

$$\text{MVT} \rightarrow = \lim_{\substack{v \rightarrow d \\ t \downarrow 0}} \frac{f(\bar{x} + tv) - f(\bar{x})}{t}$$

$$\text{Let } d \in T_{\Omega}(\bar{x}) \Rightarrow \exists t_i \downarrow 0, x_i \xrightarrow{d} \bar{x}$$

$$\text{s.t. } t_i^{-1}(x_i - \bar{x}) \rightarrow d.$$

$$\begin{aligned} \text{So } f'(\bar{x}, d) &= \lim_{i \rightarrow \infty} \frac{f(\bar{x} + t_i^{-1}(x_i - \bar{x})) - f(\bar{x})}{t_i^{-1}(x_i - \bar{x})} \\ &= \lim_{i \rightarrow \infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}} \geq 0. \quad \square \end{aligned}$$

Suppose

$$\Omega = \left\{ x: \begin{array}{l} f_i(x) \leq 0 \quad \forall i=1, \dots, s \\ f_i(x) = 0 \quad \forall i=s+1, \dots, m \end{array} \right\}$$

Natural candidate for tangent cone

$$\hat{T}_\Omega(\bar{x}) = \left\{ d: \begin{array}{l} \nabla f_i(\bar{x})^T d \leq 0 \quad \forall i \in I(\bar{x}) \\ \nabla f_i(\bar{x})^T d = 0 \quad \forall i=s+1, \dots, m \end{array} \right\}$$

where $I(\bar{x})$ are the active indices

$$I(\bar{x}) := \{ i \in \{1, \dots, s\} : f_i(\bar{x}) = 0 \}$$

Thm: $T_\Omega(\bar{x}) \subset \hat{T}_\Omega(\bar{x})$.

pf: Let $d \in T_\Omega(\bar{x}) \Rightarrow \exists t_i \downarrow 0, x_i \xrightarrow{e} \bar{x}$, with $t_i'(x_i - \bar{x}) \rightarrow d$.

$$\Rightarrow \nabla f_i(\bar{x})^T d = \lim_{i \rightarrow \infty} \frac{f(x_i) - f(\bar{x})}{t_i}$$

If $i \in I(\bar{x})$, then $f(x_i) - f(\bar{x}) = f(x_i) \leq 0$

If $i = s+1, \dots, m$, then $f(x_i) - f(\bar{x}) = 0$

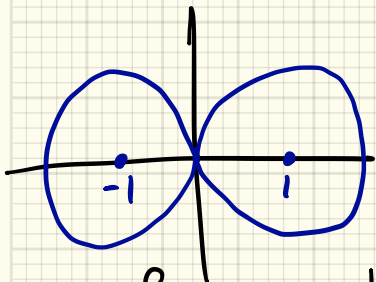


$$\text{So } T_{\Omega}(\bar{x}) \subset \left\{ d: \nabla f_i(\bar{x})^T d \leq 0 \quad \forall i \in I(\bar{x}) \right. \\ \left. \nabla f_i(\bar{x})^T d = 0 \quad \forall i = s+1, \dots, m \right\}$$

$\parallel$
 $\hat{T}(\bar{x})$

Bad Example:

$$\Omega = \{(x, y): (x-1)^2 + y^2 = 1, (x+1)^2 + y^2 = 1\} = \{(0, 0)\}$$



$$T_{\Omega}(\bar{x}) = \{(0, 0)\}$$

$$T_{\hat{\Omega}}(\bar{x}) = \{0\} \times \mathbb{R}.$$

Thm: Suppose Mangasarian-Fromovitz
constraint qualification holds (MFCQ):

$$\left\{ \begin{array}{l} \sum_{i=1}^m \lambda_i \nabla f_i(\bar{x}) = 0, \quad \lambda_i f_i(\bar{x}) = 0 \quad \forall i = 1, \dots, s \\ \lambda_1, \dots, \lambda_s \geq 0 \end{array} \right\}$$

$$\lambda \perp \hat{T}(\bar{x})$$

Then $T_{\Omega}(\bar{x}) = \hat{T}_{\Omega}(\bar{x})$

Thm: Suppose MFCQ holds at $\bar{x}$ and $\bar{x}$ is a local minimizer.

Then $\exists \lambda \in \mathbb{R}^m$ such that

$$0 = \nabla f_0(\bar{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\bar{x})$$

$$0 = \lambda_i f_i(\bar{x}) \quad \forall i=1, \dots, s$$

$$0 \leq \lambda_1, \dots, \lambda_s$$

We call λ the Lagrange multiplier vector.

pt: Consider $\max_{s.t.} -\nabla f_0(\bar{x})^T d$
 $\nabla f_i(\bar{x})^T d \leq 0 \quad \forall i \in I(\bar{x})$
 $\nabla f_i(\bar{x})^T d = 0 \quad \forall i = s+1, \dots, m$

opt val $\nearrow = 0$

Dual problem

max
s.t.

$$\sum_{i \in I(\bar{x})} \lambda_i \nabla f_i(\bar{x}) + \sum_{i=s+1}^m \lambda_i \nabla f_i(\bar{x}) = -\nabla f_0(\bar{x})$$

$$0 \leq \lambda_i \quad \forall i \in I(\bar{x}).$$

By Linear Programming duality,
the dual is feasible $\square$

Consider the problem

$$(CP) \min_{x \in \Omega} f(x)$$

where f is convex
and Ω is a convex set.

Thm: The following are equivalent.

- (1) $\bar{x}$ is a local minimizer of (CP)
- (2) $\bar{x}$ is a global minimizer of (CP)
- (3) $f'(\bar{x}, v) \geq 0 \quad \forall v \in T_{\Omega}(\bar{x})$.

Fix a set

$$\Omega = \left\{ x : \begin{array}{l} f_i(x) \leq 0 \quad \forall i=1, \dots, s \\ f_i(x) = 0 \quad \forall i=s+1, \dots, m \end{array} \right\}$$

Then Ω is convex provided f_i ($i=1, \dots, s$) are convex and f_i ($i=s+1, \dots, m$) are affine meaning $f_i(x) = a_i^T x + b_i$ for some $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$

Thm: Suppose f_i ($i=1, \dots, s$) are convex and f_i ($i=s+1, \dots, m$) are affine.

Suppose $\exists \hat{x} \in \Omega$ with

$$f_i(\hat{x}) < 0 \quad \forall i=1, \dots, s$$

(This is called Slater condition)
Then MFCQ holds at any $x \in \Omega$.

Optimization over Simple Constraints

Consider

$$\begin{array}{ll} \min_x & f(x) \\ \text{s.t.} & x \in \Omega \end{array}$$

where Ω is a closed convex set.
and f is differentiable

Defn: Orthogonal projection

$$\text{proj}_{\Omega}(x) = \arg \min_{z \in \Omega} \|z - x\|$$

Often $\text{proj}_{\Omega}(\cdot)$ is easy to compute

Examples: $\Omega = [0, 1]^n$, $\Omega = \{x \geq 0 : \sum_{i=1}^n x_i = 1\}$

$$\Omega = \{x : \|x\| \leq 1\}, \Omega = \{x : Ax = b\}$$

$$\Omega = \{x : \sum_{i=1}^n |x_i| \leq 1\}$$

Thm: $z = P_{\Omega}(x)$ if $z \in \Omega$ and

$$(x - z)^T (y - z) \leq 0 \quad \forall y \in \Omega$$

pf: $z = P_{\Omega}(x) \Leftrightarrow z$ solves $\min_{y \in \Omega} \|y - x\|_2^2$

$$\Leftrightarrow (z - x)^T w \geq 0 \quad \forall w \in T_{\Omega}(x)$$

$$\Leftrightarrow (z - x)^T (y - z) \geq 0 \quad \forall y \in \Omega.$$

□

Thm: $\bar{x} \in \Omega$ satisfies

$$f'(\bar{x}, u) \geq 0 \quad \forall u \in T_{\Omega}(\bar{x})$$

iff

$$\bar{x} = \text{Proj}_{\Omega}(\bar{x} - t \nabla f(\bar{x}))$$

where $t > 0$ is arbitrary

pf: By previous thm,

$$\bar{x} = \text{proj}_{\Omega}(\bar{x} - t \nabla f(\bar{x}))$$

$$\Leftrightarrow (\bar{x} - t \nabla f(\bar{x}) - \bar{x})^T (y - \bar{x}) \leq 0 \quad \forall y \in \Omega$$

$$\Leftrightarrow \nabla f(\bar{x})^T (y - \bar{x}) \geq 0 \quad \forall y \in \Omega.$$

Gradient projection method 

Input: $\epsilon > 0, x_0 \in \Omega$

While $\|x_k - x_{k-1}\| > \epsilon$

• pick a step-size t_k by a

• like-search
set $x_{k+1} = P_{\Omega}(x_k - t_k \nabla f(x_k))$

end

Define

$$G_m(x) = M\left[x - P_\Omega\left(x - \frac{1}{m} \nabla f(x)\right)\right]$$

Back-tracking (s, α, β)

Require: $s > 0$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$

Set $t_k = s$.

While $f(x_k) - f(P_\Omega(x_k - t_k \nabla f(x_k))) < \alpha t_k \|G_m(x_k)\|_{t_k}^2$

$$t_{k+1} = \beta t_k$$

Thm: Consider the problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in \Omega \end{array}$$

where Ω is a closed convex set and f is differentiable with

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y$$

and $\inf_x f > -\infty$. Let $\{x_k\}$ be chosen

by projected gradient method with backtracking. Then every limit point of $\{x_k\}$ satisfies first-order necessary conditions for optimality.

Convex Duality:

$$(P) \quad \min_{s.t.} \quad f_0(x) \\ f_i(x) \leq 0 \quad \text{for } i=1, \dots, s \\ f_i(x) = 0 \quad \text{for } i=s+1, \dots, m$$

where $f_0, f_1, \dots, f_s(x)$ are convex
and $f_i (i=s+1, \dots, m)$ are affine.
Let x^* be feasible of P .

Defn: Lagrangian function

$$L(x, y) = f_0(x) + y_1 f_1(x) + \dots + y_m f_m(x)$$

defined on $\mathbb{R}^n \times K$ where

$K := \mathbb{R}_+^s \times \mathbb{R}^{m-s}$. A pair $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times K$
is a saddle-point of L if

$$L(\bar{x}, y) \leq L(\bar{x}, \bar{y}) \leq L(x, \bar{y}) \quad \forall (x, y) \in \mathbb{R}^n \times K$$

Thm: If $(\bar{x}, \bar{y})$ is a saddle point of L , then $\bar{x}$ solves P . Conversely, if $\bar{x}$ solves P and Slater holds, then $\exists \bar{y}$ s.t. $(\bar{x}, \bar{y})$ is a saddle point of L .

pt: Suppose $(\bar{x}, \bar{y})$ is a saddlepoint,
then

$$\begin{aligned} \sup_{y \in K} L(\bar{x}, y) &\leq L(\bar{x}, \bar{y}) = f_0(\bar{x}) + \sum_{i=1}^m \bar{y}_i f_i(\bar{x}) \\ &= \sup_{y \in K} \{f_0(\bar{x}) + y_1 f_1(\bar{x}) + \dots + y_m f_m(\bar{x})\} \\ &= \begin{cases} f_0(\bar{x}), & \text{if } \bar{x} \in \Omega \\ +\infty, & \text{o.w.} \end{cases} \end{aligned}$$

$$\Rightarrow \bar{x} \in \Omega \quad \text{and} \quad \sum_{i=1}^m \bar{y}_i f_i(\bar{x}) = 0.$$

$$\begin{aligned} \hookrightarrow L(\bar{x}, \bar{y}) &\leq \inf_{\bar{x}} L(\bar{x}, \bar{y}) \leq \inf_{x \in \Omega} L(x, \bar{y}) \\ &\stackrel{!}{=} f_0(\bar{x}) = \inf_{x \in \Omega} \left\{ f_0(x) + \underbrace{\sum_{i=1}^m \bar{y}_i f_i(x)}_{\leq 0} \right\} \end{aligned}$$

Thus $\bar{x}$ solves f_0 .

Conversely suppose Slater holds and $\bar{x}$ solves P . Then by KKT $\exists y$ s.t.

$$0 = \nabla f_0(\bar{x}) + \sum_{i=1}^m \bar{y}_i \nabla f_i(\bar{x}) \quad (= \nabla_x L(\bar{x}, \bar{y}))$$

$$0 = \bar{y}_i f_i(\bar{x}) \quad \forall i$$

$$\bar{y}_i \geq 0 \quad \forall i=1, \dots, s$$

Thus $\forall x \in \mathbb{R}^n$, we have

$$L(\bar{x}, \bar{y}) = f_0(\bar{x}) + \sum \bar{y}_i f_i(\bar{x})$$

$$\leq L(x, \bar{y})$$

$L(\cdot, \bar{y})$ is convex with $\bar{x}$ its minimizer

Moreover $\forall y \in K$

$$L(\bar{x}, y) = f_0(\bar{x}) + \sum_{i=1}^m y_i f_i(\bar{x}) \leq f_0(\bar{x}) = L(\bar{x}, \bar{y})$$

□

Observe

$$\sup_{y \in K} L(x, y) = \begin{cases} f_0(x), & \text{if } x \in \mathcal{R} \\ +\infty, & \text{o.w} \end{cases}$$

Hence

$$\inf_x \sup_{y \in K} L(x, y) = \inf_{x \in \mathcal{R}} f_0(x)$$

Primal Problem.

Define the dual problem

$$(D) \sup_{y \in K} \inf_x L(x, y)$$

Thm: If L has a saddle-point $(\bar{x}, \bar{y})$
then $\bar{x}$ solves the primal,
 $\bar{y}$ solves the dual and
 $\text{val}(P) = \text{val}(D)$.

pt: Observe For any $\hat{y} \in K$,

$$L(x, \hat{y}) \leq \sup_{y \in K} L(x, y)$$

$$\Rightarrow \inf_x L(x, \hat{y}) \leq \inf_x \sup_{y \in K} L(x, y)$$

$$\Rightarrow \sup_{y \in K} \inf_x L(x, y) \leq \inf_x \sup_{y \in K} L(x, y)$$

weak duality inequality, requires no assumptions. Suppose $(\bar{x}, \bar{y})$ is a saddle point of L . Then

$$\inf_x \sup_{y \in K} L(x, y) \leq \sup_{y \in K} L(\bar{x}, y)$$

$$\leq L(\bar{x}, \bar{y}) \leq \inf_x L(x, \bar{y})$$

$$\leq \sup_{y \in K} \inf_x L(x, y)$$

Equality must hold throughout □

Ex: (Linear Programming)

$$(P) \quad \min_x b^T x$$

s.t. $A^T x \geq c$
 $x \geq 0$

$$L(x, y, v) = b^T x + y^T (c - A^T x) - v^T x$$

where $y \geq 0, v \geq 0$.

$$(D) \quad \max_{y \geq 0, v \geq 0} \min_x L(x, y) =$$

$$= \max_{y \geq 0, v \geq 0} \min_x \left\{ (b - Ay - v)^T x + y^T c \right\}$$

$$= \begin{cases} y^T c, & \text{if } b - Ay - v = 0 \\ -\infty, & \text{o.w.} \end{cases}$$

$$= \max_{y, v \text{ s.t.}} y^T c$$
$$A^T y + v = b$$
$$y \geq 0, v \geq 0$$

dual LP

↖ stock variable

Ex: (Convex Quadratic Primal)

$$(P) \min_x \frac{1}{2} x^T Q x + c^T x \quad \text{where } Q \succ 0$$

$$\text{s.t. } Ax \leq b, \quad 0 \leq x$$

$$L(x, y, v) = \frac{1}{2} x^T Q x + c^T x + y^T (Ax - b) - v^T x$$

Dual

$$\max_{\substack{y \geq 0 \\ v \geq 0}} \min_x \frac{1}{2} x^T Q x + c^T x + y^T (Ax - b) - v^T x$$

Solve for x

$$Qx + c + A^T y - v = 0$$

$$\Rightarrow x = Q^{-1} (v - c - A^T y)$$

$\Rightarrow$ Plug back in to get (after cancellation)

$$\max_{\substack{y \geq 0 \\ v \geq 0}} -\frac{1}{2} (v - A^T y - c)^T Q^{-1} (v - A^T y - c) - b^T y$$

$\uparrow$ Dual QP.

Penalty Methods

Penalty Algorithms

$$(*) \min_x f_0(x) \quad \text{s.t.} \quad x \in \Omega = \{x: \begin{matrix} f_i(x) \leq 0 & \forall i=1, \dots, s \\ f_j(x) = 0 & \forall j=s+1, \dots, m \end{matrix}\}$$

Idea: We will replace $(*)$ by a sequence of unconstrained problems

$$\min_{x \in \mathbb{R}^n} P_d(x) = f_0(x) + d \Theta(x)$$

where $d > 0$ and $\Theta: \mathbb{R}^n \rightarrow \mathbb{R}$ satisfies

1) Θ is continuous

2) $\Theta(x) \geq 0 \quad \forall x \in \mathbb{R}^n$

3) $\Theta(x) = 0$ iff $x \in \Omega$.

We will use notation $r^+ = \max\{0, r\}$

Examples:

$$1) \Theta(x) = \sum_{i=1}^s [f_i(x)]^+ + \sum_{j=s+1}^m |f_j(x)|$$

$$2) \Theta(x) = \frac{1}{2} \sum_{i=1}^s [f_i(x)]^+ + \sum_{j=s+1}^m [f_j(x)]^2$$

$$3) \Theta(x) = \max \left\{ 0, f_i(x), |f_j(x)|, i=1, \dots, s, j=s+1, \dots, m \right\}$$

Algorithm:

Let $\alpha_i > 0$ be such that $\alpha_i \rightarrow +\infty$
with $\alpha_i \leq \alpha_{i+1}$. Let x_n solve
 $\min_{x \in \mathbb{R}^n} p_{\alpha_n}(x)$.

Ex: $\min_x$
s.t. $x=0$

Consider $p_\alpha(x) = x + \frac{\alpha}{2} x^2$
minimizing get

$$0 = 1 + \alpha \hat{x} \Rightarrow \boxed{\hat{x} = -\frac{1}{\alpha}}$$

So need $\alpha \sim \frac{1}{\epsilon}$ to guarantee $\|\hat{x}\| \leq \epsilon$.

Consider instead

$$p_\alpha(x) = x + \alpha |x| = \begin{cases} (1+\alpha)x & , \text{ if } x \geq 0 \\ (1-\alpha)x & , \text{ if } x < 0 \end{cases}$$

As soon as $\alpha > 1$, $p_\alpha \geq 0$

So minimizer is $\boxed{x=0}$.

Lemma: If $\{x_k\}$ is generated by the penalty method, then:

$$1) \rho_{\alpha_k}(x_k) \leq \rho_{\alpha_{k+1}}(x_{k+1})$$

$$2) \theta(x_k) \geq \theta(x_{k+1})$$

$$3) f_0(x_k) \leq f_0(x_{k+1})$$

Corollary: Every cluster point $\bar{x}$ of $\{x_k\}$ is a solution to $(\star)$.

pf of Cor: Let $x \in \Omega$. Then

$$f_0(x) = \rho_{\alpha_k}(x) \geq \rho_{\alpha_k}(x_k) \geq f_0(x_k)$$

$$\Rightarrow f_0(x) \geq f_0(\bar{x}) \quad \forall x \in \Omega.$$

$\rho_{\alpha_k}(x_k)$ is increasing and bounded from above so has limit p^* .

$$\text{Then } f_0(x) \geq \lim_{k \rightarrow \infty} \rho_{\alpha_k}(x_k) =$$

$$= \lim_{k \rightarrow \infty} f_0(x_k) + \alpha_k \theta(x_k)$$

$$\Rightarrow \theta(x_k) \rightarrow 0. \quad \square$$

pf of Lemma:

$$P_{\alpha_k}(x_k) \leq P_{\alpha_k}(x_{k+1}) \leq P_{\alpha_{k+1}}(x_{k+1})$$

So ① holds. Next

$$f_0(x_k) + \alpha_k \Theta(x_k) \leq f_0(x_{k+1}) + \alpha_k \Theta(x_{k+1})$$

and

$$f_0(x_{k+1}) + \alpha_{k+1} \Theta(x_{k+1}) \leq f_0(x_k) + \alpha_{k+1} \Theta(x_k)$$

Adding

$$\Rightarrow (\alpha_{k+1} - \alpha_k) \Theta(x_{k+1}) \leq (\alpha_{k+1} - \alpha_k) \Theta(x_k)$$

$$\Rightarrow \Theta(x_{k+1}) \leq \Theta(x_k).$$

Finally

$$\begin{aligned} f_0(x_k) + \alpha_k \Theta(x_k) &\leq f_0(x_{k+1}) + \alpha_k \Theta(x_{k+1}) \\ &\leq f_0(x_{k+1}) + \alpha_{k+1} \Theta(x_k) \end{aligned}$$

□