日期:

- $\star$ ℓ_1 norm: $\|x\|_1 := \sum_{i=1}^n |x_i|$.
- $\star$ ℓ_2 norm: $\|x\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2}$.
- $\star$ ℓ_∞ norm: $\|x\|_\infty := \max_{1\le i\le n} |x_i|$.

Theorem 1.1: Let $\|\cdot\|$ be a norm in $\mathbb{R}^n$. Then there exist positive numbers C_1 and C_2 so that for all $x \in \mathbb{R}^n$,

$$C_1 \sum_{i=1}^n |x_i| \le \|x\| \le C_2 \sum_{i=1}^n |x_i|$$

Theorem 1.2: Let $\|\cdot\|$ be a norm. Then the following function defines a matrix norm

$$|||A||| := \max_{\|x\|=1} \|Ax\|.$$

- $\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$ (maximum of the ℓ_1 norms of columns). Moreover, this is an induced matrix norm:

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1.$$

- $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$. Moreover, this is an induced matrix norm:

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

- $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$ (maximum of the ℓ_1 norms of rows). Moreover, this is an induced matrix norm:

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty.$$

- $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |a_{ij}|^2}$. This is known as the Frobenius norm.

Theorem 1.9. (Taylor's theorem in $\mathbb{R}^n$ with remainder term)

- Let $f \in C^1(\mathbb{R}^n)$, x and $y \in \mathbb{R}^n$. Then there exists $\xi \in \{(1-s)x + sy : \ s \in (0,1)\}$ such that
$$f(y) = f(x) + [\nabla f(\xi)]^T (y - x).$$
- Let $f \in C^2(\mathbb{R}^n)$, x and $y \in \mathbb{R}^n$. Then there exists $\xi \in \{(1-s)x + sy : \ s \in (0,1)\}$ such that
$$f(y) = f(x) + [\nabla f(x)]^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(\xi)(y - x).$$

Definition: We say that x^* is a stationary point of f if $\nabla f(x^*) = 0$.

Theorem 2.2. Let $f \in C^2(\mathbb{R}^n)$.

1. If x^* is a local minimizer of f, then $\nabla^2 f(x^*) \succeq 0$.
2. If x^* is a stationary point of f and $\nabla^2 f(x^*) \succ 0$, then x^* is a local minimizer.

日期

Newton's method

Let $x^0 \in \mathbb{R}^n$. For $k = 0, 1, 2, \ldots$, update

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k).$$

Solve for d

$\nabla^2 f(x^k) d = \nabla f(x^k)$

Newton direction

Solve

$f(x) = 0 \iff g(x) = f'(x) = 0$

$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$

Theorem 2.3. (Quadratic convergence of Newton's method)
Let $g \in C^2(\mathbb{R})$ and x_* satisfies $g(x_*) = 0$ and $g'(x_*) \neq 0$. Then there exists $\epsilon > 0$ so that if $|x_0 - x_*| < \epsilon$ then the Newton's iterate $x_{k+1} = x_k - g(x_k)/g'(x_k)$ is well defined and there exists $M > 0$ so that

$$|x_{k+1} - x_*| \leq M |x_k - x_*|^2.$$

0.1 0.2

so $\{x_k\} \to x_*$ convergent

Note:

Steepest descent cont.

- $-\nabla f(x)$ is called the steepest descent direction.
- A natural greedy algorithm is

Steepest descent with exact line search
Start at $x^0 \in \mathbb{R}^n$. For each $k = 0, 1, 2, \ldots$,

- ⋆ Set $d^k = -\nabla f(x^k)$.
- ⋆ Pick α_k so that

the set of minimizers

$$\alpha_k \in \operatorname{Arg\,min}\{f(x^k + \alpha d^k) : \alpha \geq 0\}. \quad (1)$$

- ⋆ Set $x^{k+1} = x^k + \alpha_k d^k$.

Note: The update $x^{k+1} = x^k + \alpha_k d^k$ is prototypical in optimization.

- d^k is called the search direction. In the above algorithm, $d^k = -\nabla f(x^k)$.
- α_k is called the step size. In the above algorithm, it is chosen according to the exact line search criterion (1).

learning rate

$f(x+d) \approx f(x) + (\nabla f(x))^T d$

$f(x - 2 \cdot \nabla f(x)) \approx f(x) - 2 \cdot \nabla f(x)^T \cdot \nabla f(x) < f(x).$

$d^k = -\nabla f(x^k).$

日期:

Armijo rule

In contrast to exact line search, usually inexact line search strategy is performed. One commonly used rule is:

Armijo rule:
Let $\sigma \in (0,1)$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x+\alpha d) \le f(x) + \alpha\sigma[\nabla f(x)]^T d.$$

Definition. Let $f \in C^1(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. A $d \in \mathbb{R}^n$ is said to be a descent direction of f at x if

$$[\nabla f(x)]^T d < 0.$$

Examples: At an x that is not stationary,

- $d = -\nabla f(x)$ is a descent direction;
- More generally, if $D \succ 0$, then $d = -D\nabla f(x)$ is a descent direction.

Is the Newton direction $-[\nabla^2 f(x)]^{-1}\nabla f(x)$ a descent direction?

Newton direction is in general not a descent direction.

Armijo rule cont.

Trial list $\{\bar\alpha, \bar\alpha\beta, \bar\alpha\beta^2, \dots\}$

How to execute Armijo rule in practice?

$\alpha^k = \bar\alpha \cdot \beta^{j_k}$, $\beta < 1$, $j_k = 0, 1, 2 \dots$

10^{-4} $\quad \frac{1}{2}$

Armijo line search by backtracking:
Fix $\sigma \in (0,1)$ and $\beta \in (0,1)$. Given $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ and $\bar\alpha > 0$.
Find the smallest nonnegative integer $j = j_0$ so that

$$f(x+\bar\alpha\beta^j d) \le f(x) + \bar\alpha\beta^j\sigma[\nabla f(x)]^T d. \quad (2)$$

The stepsize generated is then $\bar\alpha\beta^{j_0}$.

β to the power j

$f(x+\bar\alpha\cdot\beta^j d) \le f(x) + \bar\alpha\cdot\beta^j\cdot\sigma(\nabla f(x))^T d$, find j satisfies it.

Convergence under Armijo rule

Theorem 2.5:
Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$. Let $\{\bar\alpha_k\} \subset \mathbb{R}$ satisfy $0 < \inf_k \bar\alpha_k \le \sup_k \bar\alpha_k < \infty$, and fix $\sigma \in (0,1)$ and $\beta \in (0,1)$. Suppose $\{x^k\}$ is generated as

① $\sigma = 10^{-4}$, $\beta = \frac{1}{2}$

② $x^{k+1} = x^k + \alpha_k d^k$,

where

- ② $d^k := -D_k\nabla f(x^k)$; here $\{D_k\}$ is a bounded sequence of positive definite matrices with $D_k - \delta I \succeq 0$ for some $\delta > 0$; (③ indep of k)
- α_k is generated via the Armijo line search by backtracking with $x = x^k$, $d = d^k$ and $\bar\alpha = \bar\alpha_k$, and σ and β defined above. (④)

Then any accumulation point of $\{x^k\}$ is a stationary point of f.

descent direction

⑤

③ $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ is bounded.

$\|A\|_2 = \sqrt{\lambda_{\max}|A|}$.

Any convergent subsequence (if exists) will converge to a stationary point of f.

$\lambda_{\min}(D_k) \ge \delta > 0$

日期:

Special case

Corollary 2.1: (Steepest descent with constant stepsize)
Let $f \in C^2(\mathbb{R}^n)$ with $\inf f > -\infty$. Suppose that there exists $L > 0$ so that

$$L \geq \|\nabla^2 f(x)\|_2 \quad \text{for all } x.$$

Fix any $\gamma \in (0, 2)$ and consider the sequence generated as

$$x^{k+1} = x^k - \frac{\gamma}{L}\nabla f(x^k).$$

Then any accumulation point of $\{x^k\}$ is a stationary point of f.

A chain rule

Let $h \in C^2(\mathbb{R}^m)$ and let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$.

Define $f(x) := h(Ax - b)$. Then $f \in C^2(\mathbb{R}^n)$ and

$$\nabla f(x) = A^T \nabla h(Ax - b) \quad \text{and} \quad \nabla^2 f(x) = A^T \nabla^2 h(Ax - b) A.$$

Quasi-Newton method: Basic version

Given $f \in C^1(\mathbb{R}^n)$.

Quasi-Newton based on B_k:
Initialize at $x^0 \in \mathbb{R}^n$ and $B_0 \succ 0$.
For $k = 0, 1, 2, \ldots$

1. Find d^k via $B_k d^k = -\nabla f(x^k)$.
2. Update $x^{k+1} = x^k + d^k$. Or, more generally, $x^{k+1} = x^k + \alpha_k d^k$ for some $\alpha_k > 0$.
3. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ and $s^k = x^{k+1} - x^k$. Compute B_{k+1}.

Quasi-Newton based on H_k:
Initialize at $x^0 \in \mathbb{R}^n$ and $H_0 \succ 0$.
For $k = 0, 1, 2, \ldots$

1. Find d^k via $d^k = -H_k \nabla f(x^k)$.
2. Update $x^{k+1} = x^k + d^k$. Or, more generally, $x^{k+1} = x^k + \alpha_k d^k$ for some $\alpha_k > 0$.
3. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ and $s^k = x^{k+1} - x^k$. Compute H_{k+1}.

$H_k \succ 0$?

Proposition 3.1
Let $H_k \succ 0$ and $y^{k^T} s^k > 0$. Let H_{k+1} be given by BFGS update. Then $H_{k+1} \succ 0$. The same conclusion holds if H_k and H_{k+1} are replaced by B_k and B_{k+1}, respectively.

Popular update formulae

Initialize B_0 (or H_0) at a positive definite matrix.

Method	$B_{k+1} =$	$H_{k+1} =$
DFP	$\left(I - \frac{y^k s^{k\,T}}{y^{k\,T} s^k}\right) B_k \left(I - \frac{s^k y^{k\,T}}{y^{k\,T} s^k}\right) + \frac{y^k y^{k\,T}}{y^{k\,T} s^k}$	$H_k + \frac{s^k s^{k\,T}}{y^{k\,T} s^k} - \frac{H_k y^k y^{k\,T} H_k}{y^{k\,T} H_k y^k}$
BFGS	$B_k + \frac{y^k y^{k\,T}}{y^{k\,T} s^k} - \frac{B_k s^k s^{k\,T} B_k}{s^{k\,T} B_k s^k}$	$\left(I - \frac{s^k y^{k\,T}}{y^{k\,T} s^k}\right) H_k \left(I - \frac{y^k s^{k\,T}}{y^{k\,T} s^k}\right) + \frac{s^k s^{k\,T}}{y^{k\,T} s^k}$
SR1	$B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$	$H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^T}{(s^k - H_k y^k)^T y^k}$

Remark: symmetric rank 1.

Wolfe conditions

In view of Proposition 3.1, it suffices to guarantee that $H_0 \succ 0$ and make sure that $y^{k\,T} s^k > 0$ for each $k \geq 0$.

The latter can be guaranteed if line search is performed to guarantee the Wolfe conditions.

Wolfe conditions:
Let $0 < c_1 < c_2 < 1$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x + \alpha d) \leq f(x) + \alpha c_1 [\nabla f(x)]^T d,$$
$$-[\nabla f(x + \alpha d)]^T d \leq -c_2 [\nabla f(x)]^T d.$$

Armijo rule

Remark:

- The first inequality in Wolfe conditions is the Armijo rule.
- The second relation is called curvature condition.

Wolfe conditions cont.

Theorem 3.3 (Wolfe conditions are not void)
Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$, $x \in \mathbb{R}^n$, and $d \in \mathbb{R}^n$ be a descent direction at x. Let $0 < c_1 < c_2 < 1$. Then there exists $\alpha > 0$ with

$$f(x + \alpha d) \leq f(x) + \alpha c_1 [\nabla f(x)]^T d,$$
$$-[\nabla f(x + \alpha d)]^T d \leq -c_2 [\nabla f(x)]^T d.$$

日期:

Quasi-Newton method: Wolfe line search

Quasi-Newton using H_k in BFGS for $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$:
Pick $0 < c_1 < c_2 < 1$, $x^0 \in \mathbb{R}^n$, $H_0 = \eta I$ for some $\eta > 0$.
For $k = 0, 1, 2, \ldots$

1. Find d^k via $d^k = -H_k \nabla f(x^k)$.
2. Compute α_k that satisfies the Wolfe conditions. ← only difference
3. Update $x^{k+1} = x^k + \alpha_k d^k$.
4. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$, $s^k = x^{k+1} - x^k$ and compute H_{k+1} as in BFGS.

Convergence under Wolfe conditions

Theorem 3.4: (Zoutendijk's theorem)
Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$, $x^0 \in \mathbb{R}^n$ and $\exists \ell > 0$ so that

conditions: ① ② $\|\nabla f(x) - \nabla f(y)\|_2 \le \ell \|x - y\|_2$ Lipchitz's condition

whenever $\max\{f(x), f(y)\} \le f(x^0)$. Let $\{x^k\}$ be a sequence of non-stationary points generated as

$$x^{k+1} = x^k + \alpha_k d^k,$$

with d^k being a descent direction and α_k satisfying the Wolfe conditions. Then it holds that

conclusion: $\cos\theta_k$ is bounded $\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^k)\|_2^2 < \infty,$

where $\cos\theta_k = \frac{-[\nabla f(x^k)]^T d^k}{\|\nabla f(x^k)\|_2 \|d^k\|_2}$. (ZM)

Then we proove sequence $\{x^k\}$, $\|\nabla f(x^k)\|_2 \to 0$ $(k \to \infty)$.

$\|x^k\|$ bounded $\Rightarrow$ $x^{k_j} \to x^*$ $(j \to +\infty)$

Thm: $f \in C^2(\mathbb{R}^n)$, $\|\nabla f(x) - \nabla f(y)\|_2 \le \max \|\nabla^2 f(\xi)\|_2 \cdot \|x - y\|_2$

A closedness result

Theorem 4.4:
Let $A \in \mathbb{R}^{m \times n}$. Then the set

$$S := \{Ay : y \ge 0\}$$

is closed and convex, where $y \ge 0$ means $y_i \ge 0$ for each i.

$A(\lambda y_1 + (1-\lambda) y_2) \in S$

Def: $f: \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$

① $\operatorname{dom} f := \{x \mid f(x) < \infty\}$

② $\operatorname{epi} f := \{(x, r) \in \mathbb{R}^{n+1} \mid r \ge f(x)\}$

epi f convex $\Leftrightarrow$ f convex

日期:

Theorem 4.7:
Suppose that $f \in C^2(\mathbb{R}^n)$. Then f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$.

Thm: f convex $\Leftrightarrow$ $\nabla^2 f(x) \succeq 0$

x^* is local minimizer $\Rightarrow$ $\nabla^2 f(x^*) \succeq 0$

$\nabla^2 f(x^*) \succ 0$, $\nabla f(x^*) = 0$ $\Rightarrow$ x^* is local minimizer.

Calculus of convex functions

Proposition 4.1:
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ both be convex, $A \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^n$, $H(x) := Ax - b$, and $\alpha > 0$. Then the following functions are convex:

- $f + g$;
- αf; $\alpha > 0$
- $f \circ H$; (1)
- $\max\{f, g\}$. (2)

f, $\Rightarrow f(H)$

f, g convex

Example: proove $\sum_{i=1}^{n-1} |x_{i+1} - x_i|$ convex,

$\sum_{i=1}^{n-1} |x_{i+1} - x_i| = \|Dx\|_1 = \|\cdot\|_1 \circ D\,(x)$

$\|\cdot\|_1$ is convex, D linear $\Rightarrow$ $\|Dx\|_1$ convex

---

Thm: if h is convex

g is convex and non-decreasing

$\Rightarrow$ $f = g \circ h$ is convex.

日期:

Strong duality I (necessary)

Theorem 5.1 (Strong duality for LP, version I)
Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$. Consider

$$v_p := \sup_x\{c^T x : \ Ax = b, x \geq 0\}, \quad v_d := \inf_y\{b^T y : c \leq A^T y\}.$$

Suppose that there exists $\hat{x} \geq 0$ with $A\hat{x} = b$. Then $v_p = v_d$.

implies ⇕ 0

primal: $\max\ c^T x$ s.t. $Ax = b$, $x \geq 0$

dual: $\min\ b^T y$ s.t. $c \leq A^T y$

$\exists\ \hat{x} \geq 0$, and $A\hat{x} = b \Rightarrow v_p = v_d$.
(constraints)

Strong duality II

Theorem 5.2 (Strong duality theorem for LP)
Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$. Consider

$$v_p := \sup_x\{c^T x : \ Ax = b, x \geq 0\}, \quad v_d := \inf_y\{b^T y : c \leq A^T y\}.$$

Suppose that either

- there exists $\hat{x} \geq 0$ with $A\hat{x} = b$; or
- there exists $\hat{y}$ with $c \leq A^T \hat{y}$.

Then $v_p = v_d$ and both optimal values are attained when finite.

$f(\hat{x}) = v_p$, since $\hat{x}$ is unique feasible element.

if $\exists\ \hat{x}$ or $\hat{y}$ obey constraints $\Rightarrow v_p = v_d$ and finite.

if $\nexists\ \hat{x}$ obey constraints, and $\exists\ \hat{y}$ obey constraints

$\Rightarrow v_p = \max c^T x = -\infty$

and $v_d = \min b^T y = -\infty$

if $\nexists\ \hat{x}, \hat{y} \Rightarrow v_p = v_d = \infty$.

	Maximize $c^T x$ subject to $Ax \clubsuit b$ $x \diamondsuit$	Minimize $b^T y$ subject to $A^T y \diamondsuit c$ $y \clubsuit$
$\clubsuit$	ith constraint $\leq$ ith constraint $\geq$ ith constraint $=$	ith variable ≥ 0 ith variable ≤ 0 ith variable unrestricted
$\diamondsuit$	jth variable ≥ 0 jth variable ≤ 0 jth variable unrestricted	jth constraint $\geq$ jth constraint $\leq$ jth constraint $=$

日期: /

Semi definite Programming

SDP

primal: $\min_{X \in S^n} \operatorname{tr}(CX)$

s.t. $\operatorname{tr}(A_i X) = b_i, \quad i = 1, \cdots, m$

$X \succcurlyeq 0.$

dual: $\max_{y \in \mathbb{R}^m} b^T y$

s.t. $C - \sum_i^m y_i \cdot A_i \succcurlyeq 0$

Thm: if $\exists\, \bar{y}$, s.t. $C - \sum \bar{y}_i \bar{A}_i \succ 0$

$\{(\operatorname{tr}(CX), \operatorname{tr}(A_1X), \cdots \operatorname{tr}(A_mX))^T \mid X \succcurlyeq 0\}$ is closed, $\Rightarrow$ primal optimal value is attainable.

Strong duality

Theorem 6.1 (Strong duality for SDPs)

Consider the following primal-dual SDP pairs:

$$\text{Primal}: \begin{cases} \underset{X \in S^n}{\text{Minimize}} & \operatorname{tr}(CX) \\ \text{subject to} & \operatorname{tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \\ & X \succeq 0, \end{cases}$$

$$\text{Dual}: \begin{cases} \underset{y \in \mathbb{R}^m}{\text{Maximize}} & b^T y \\ \text{subject to} & C - \sum_{i=1}^m y_i A_i \succeq 0, \end{cases}$$

where $C \in S^n$ and $A_i \in S^n$ for all i. Let v_p and v_d denote their optimal values respectively. Then the following statements hold.

1. If there exists $\bar{X} \succ 0$ such that $\operatorname{tr}(A_i \bar{X}) = b_i$ for all i, then $v_p = v_d$ and v_d is attained when finite.
2. If there exists $\bar{y} \in \mathbb{R}^m$ such that $C - \sum_{i=1}^m \bar{y}_i A_i \succ 0$, then $v_p = v_d$ and v_p is attained when finite.

if $\bar{X}$ is unique element in feasible region

Def: (slater point): $\bar{X}$ satisfies $\succ 0$

Thm: ① if $\exists\, \bar{x}$ obey primal constraints and $(X \succ 0)$

$\Rightarrow v_p = v_d$, and v_d is attainable when finite

② if $\exists\, \bar{y}$ obey $C - \sum \bar{y}_i A_i \succ 0$ (not $\succcurlyeq 0$)

$\Rightarrow v_p = v_d$, and v_p is attained when finite

日期

To understand Theorem 6.1, we need the following.

Theorem 6.2
Let $A \in \mathcal{S}_+^n$ and $C \in \mathcal{S}_+^n$. Then $\operatorname{tr}(AC) \geq 0$.

$\succeq 0$

S_+^n : symmetric , positive semi-definite

Def : principal matrices.
(ordered)

Schur complement

The following result is crucial in reformulating problems into SDPs.

Theorem 7.1
Let $A \in \mathcal{S}^m$, $C \in \mathcal{S}^n$, $B \in \mathbb{R}^{m \times n}$, and $A \succ 0$. Then

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0 \iff C - B^T A^{-1} B \succeq 0.$$

Note: We call $C - B^T A^{-1} B$ the Schur complement of A in $\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$.

Thm: $C - B^T A^{-1} B \succeq 0$, $A \succ 0$, $\iff A \succeq 0$, $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succeq 0$.

Use: primal : $\min_x \; x^T Q x$

$\iff \min \; \operatorname{tr}(x^T Q x) = \operatorname{tr}(Q x \cdot x^T)$

$\iff \min \; \operatorname{tr}(QY)$ (Thm: ——)
s.t. $Y \succeq x \cdot x^T$

$\iff \min \; \operatorname{tr}(QY)$
s.t. $\begin{pmatrix} Y & x \\ x^T & 1 \end{pmatrix} \succeq 0$.

日期: /

Thm: ① $t \geq \|x\|_2$, $(t \geq 0)$

$\Leftrightarrow \begin{pmatrix} t \cdot I & x \\ x^T & t \end{pmatrix} \succeq 0$

② $ab \geq \|x\|_2$, $a \geq 0$, $b \geq 0$

$\Leftrightarrow \begin{pmatrix} aI & x \\ x^T & b \end{pmatrix} \succeq 0.$

③ $y \geq x^6 \Leftrightarrow y \geq t^2 \geq x^6$, $t \geq 0$

$\Leftrightarrow y \geq t^2$, $t \geq |x|^3 \Leftrightarrow y \geq t^2$, $t \geq z^3 \geq |x|^3$, $z \geq 0$

$\Leftrightarrow y \geq t^2$, $z \geq |x|$, $z \geq 0$, $t \cdot z \geq s^2 \geq z^4$, $s \geq 0$

Thm: convex QP are SDP

Pf: QP: Min $x^T A x$

$\Leftrightarrow$ Min $tr(x^T A x) = tr(A x \cdot x^T)$

since $A \succeq 0$, $\Leftrightarrow$ Min $tr(AY)$ s.t. $Y \succeq x \cdot x^T$

$\Leftrightarrow$ Min $tr(AY)$

s.t. $\begin{pmatrix} Y & x \\ x^T & 1 \end{pmatrix} \succeq 0$

it's SDP.

Def: $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$ (F norm)

日期：

KKT conditions for LP

Theorem 8.1: (Karush-Kuhn-Tucker conditions for the LP)
Consider the linear program

primal

$$\begin{array}{lll} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & c^T x & \\ \text{subject to} & Bx = d, & (2) \\ & Ax \le b. & \end{array}$$

dual : $(\mu, \lambda) = y$

where $c \in \mathbb{R}^n$, $B \in \mathbb{R}^{p\times n}$ and $A \in \mathbb{R}^{q\times n}$. Then $x^* \in \mathbb{R}^n$ is an optimal solution of (2) if and only if there exist $\lambda^* \in \mathbb{R}^q$ and $\mu^* \in \mathbb{R}^p$ so that the following conditions hold:

- (Primal feasibility) $Bx^* = d$ and $Ax^* \le b$; and
- (Dual feasibility) $B^T\mu^* + A^T\lambda^* + c = 0$ and $\lambda^* \ge 0$; and
- (Complementary slackness) $\lambda^{*T}(Ax^* - b) = 0$. ← LP : means

Remark: Since $\lambda^* \ge 0$ and $Ax^* \le b$, we note that $\lambda^{*T}(Ax^* - b) = 0$ means $\lambda_i^*(Ax^* - b)_i = 0$ for each $i = 1, \dots, q$.

$V_d = V_p$

$$\begin{array}{lll} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & f(x) & \\ \text{subject to} & h_j(x) = 0, & j \in J, \\ & g_i(x) \le 0, & i \in I. \end{array}$$

I = all i

$I(x) = \{i \mid g_i(x) = 0\}$

MFCQ is for explanation to complementary slackness

When is the approximation good?

Definition: (Mangasarian-Fromovitz constraint qualification, MFCQ)
Consider the feasible set of (1) and let x^* be feasible. We say that the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at x^* if the following conditions hold:

- **if**
$$\sum_{i\in J} \mu_j \nabla h_j(x^*) + \sum_{i\in I(x^*)} \lambda_i \nabla g_i(x^*) = 0 \text{ and } \lambda_i \ge 0 \ \forall i \in I(x^*),$$
then $\lambda_i = 0$ for all $i \in I(x^*)$ and $\mu_j = 0$ for all $j \in J$.

$g_i(x^*) \cdot \lambda_i = 0$

$I(x^*) = \{i \mid g_i(x^*) = 0\}$

Remark:

- If $g_i(x^*) < 0$ for all $i \in I$ so that $I(x^*) = \emptyset$, then MFCQ means $\{\nabla h_j(x^*) : j \in J\}$ is linearly independent.
- If $J = \emptyset$, then MFCQ means $\{\nabla g_i(x^*) : i \in I(x^*)\}$ is positively independent.

since $\nabla g_i(x^*) \le 0$ or ≥ 0

$\nabla h_j(x^*)$ means tangent

: Need to consider if $g_i(x^*) = 0$, $g_i(x^*) \ne 0$.

KKT conditions for NLP

Theorem 8.2: (KKT conditions for NLP)
Consider (1) and let x^* be a local minimizer. Suppose that MFCQ holds at x^*. Then there exist $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that

- $\nabla f(x^*) + \sum_{i \in I} \lambda_i^* \nabla g_i(x^*) + \sum_{j \in J} \mu_j^* \nabla h_j(x^*) = 0$; and
- $\lambda_i^* \geq 0$ and $\lambda_i^* g_i(x^*) = 0$ for all $i \in I$.

Remarks:

- Under MFCQ, the approximating LP is "good" around x^*. We hence look at the KKT of this LP.
- The first bullet point follows from the dual feasibility condition of the approximating LP, and by defining $\lambda_i^* = 0$ for $i \notin I(x^*)$.
- The second bullet point follows from the definition of $I(x^*)$ and by defining $\lambda_i^* = 0$ for $i \notin I(x^*)$.

Remarks cont.:

- The λ^* and μ^* are called Lagrange multipliers at x^*.
- We consider the so-called stationary points:

Definition: (Stationary points)
Consider (1). An $\bar{x}$ is called a stationary point of (1) if it is feasible and there exist $\bar{\lambda} \in \mathbb{R}^m$ and $\bar{\mu} \in \mathbb{R}^p$ so that

- $\nabla f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j \in J} \bar{\mu}_j \nabla h_j(\bar{x}) = 0$; and
- $\bar{\lambda} \geq 0$ and $\bar{\lambda}_i g_i(\bar{x}) = 0$ for all $i \in I$.

- According to Theorem 8.2, if x^* is a local minimizer and if the MFCQ holds at x^*, then x^* is a stationary point.

日期: /

Min $f(x)$

s.t. $h_j(x) = 0$, $j \in J$

$g_i(x) \le 0$, $i \in I$

KKT conditions for (1):

$$\begin{cases} \star\ g_i(x) \le 0 \text{ for all } i \in I \text{ and } h_j(x) = 0 \text{ for all } j \in J; \text{ and} \\ \star\ \nabla f(x) + \sum_{i \in I} \lambda_i \nabla g_i(x) + \sum_{j \in J} \mu_j \nabla h_j(x) = 0; \text{ and} \\ \star\ \lambda \ge 0 \text{ and } \lambda_i g_i(x) = 0 \text{ for all } i \in I. \end{cases}$$

x satisfy KKT $\Leftrightarrow$ x is stationary point

Thm: x^* is local minimizer, MFCQ $\Rightarrow$ stationary point

18/32

① check MFCQ holds for $\forall x$, $\Rightarrow$ x^* is stationary. ($\Leftrightarrow$ KKT satisfies)

② f can be not convex

Slater's condition

Theorem 8.3: (MFCQ from Slater)
Consider the set defined by

$I = (\text{all } i)$, $I(x^*) \neq I$

$$S := \{x \in \mathbb{R}^n : g_i(x) \le 0 \ \ \forall i \in I\},$$

where g_i are convex C^1. Suppose that there exists $\bar{x}$ satisfying

$$g_i(\bar{x}) < 0 \ \ \forall i \in I.$$

Then MFCQ holds at every point in S.

① convex g

② slater point $\bar{x}$.

↑ Show in exam

Remark: $\lambda_i \cdot \nabla g_i(x^*) = 0$, $\lambda_i \ge 0 \Rightarrow \lambda_i = 0$ (MFCQ)

- The set S in the above theorem is closed and convex.
- The condition that "there exists $\bar{x}$ satisfying $g_i(\bar{x}) < 0$ for all $i \in I$" is called the Slater's condition. The $\bar{x}$ is called a Slater point.
- One can indeed show that for the above S, the MFCQ holds at every point in S if and only if Slater's condition holds.

日期: /

Generalized Slater's condition

Theorem 8.4: (MFCQ from generalized Slater)
Consider the set defined by

$$S := \{x \in \mathbb{R}^n : \ g_i(x) \le 0 \ \ \forall i \in I, \ Ax = b\},$$

where g_i are convex C^1 and $A \in \mathbb{R}^{p\times n}$. Suppose that there exists $\bar{x}$ satisfying

$$A\bar{x} = b, \ \ g_i(\bar{x}) < 0 \ \ \forall i \in I,$$

and A has full row rank. Then MFCQ holds at every point in S.

$\bar{x}$ is called slater point

Remark: $\sum \mu_j \nabla h_j(x^*) = \sum \mu \cdot A = 0$, A full row rank $\Rightarrow \mu = 0$.

- The set S in the above theorem is closed and convex.

Role of convexity

Consider the following special instance of (1)

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{Minimize}} \quad & f(x) \\ \text{subject to} \quad & Ax = b, \\ & g_i(x) \le 0, \ \ i \in I, \end{aligned} \tag{3}$$

here f and g_i are all convex C^1 functions, $A \in \mathbb{R}^{p\times n}$.

Theorem 8.5: (Sufficiency under convexity)
Consider (3). Suppose that there exist $x^* \in \mathbb{R}^n$, $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ so that

- $Ax^* = b$ and $g_i(x^*) \le 0$ for all $i \in I$; and
- $\nabla f(x^*) + \sum_{i\in I} \lambda_i^* \nabla g_i(x^*) + A^T\mu^* = 0$; and
- $\lambda_i^* \ge 0$ and $\lambda_i^* g_i(x^*) = 0$ for all $i \in I$.

Then x^* is a global minimizer of (3).

KKT

Thm: f, g convex, stationary point (KKT in I) $\Rightarrow$ global minimizer

Only one minimizer

日期: /

△ : Thm: local minimizer + MFCQ

⇒ stationary point ?

Thm: KKT ⇔ stationary point

★ Penalty method (exterior type);
★ Barrier method (interior type).

Penalty method (exterior type)

- Initial idea: Define

$$P(x) := \begin{cases} 0 & \text{if } g_i(x) \le 0 \ \ \forall i \in I, \\ \infty & \text{otherwise.} \end{cases}$$

Then we can consider the unconstrained problem:

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \ f(x) + P(x).$$

Examples:

- $P(x) = \sum_{i=1}^{m} \max\{g_i(x), 0\}$.
- $P(x) = \frac{1}{2} \sum_{i=1}^{m} \max\{g_i(x), 0\}^2$. Courant-Beltrami penalty function: C^1 function

numerator
denominator

Penalty method: basic version

Penalty method for (1): basic version

Let $x^0 \in \mathbb{R}^n$, $c > 0$ and $\eta > 1$. Set $c_1 = c$. For $k = 1, \ldots,$

- Find a minimizer x^k of

$$q_{c_k}(x) := f(x) + \frac{c_k}{2} \sum_{i=1}^{m} (\max\{g_i(x), 0\})^2,$$

using x^{k-1} as the initial point for the iterative method.

- Update $c_{k+1} = \eta c_k$.

$\{c_k\}$

Remark:

- As c increases, q_c becomes more ill-conditioned. The choice of x^{k-1} as a starting point for the iterative method helps alleviate the ill-conditioning.
- The above algorithm is only conceptual because finding global minimizers can be challenging if q_{c_k} is not convex. Global minimizers also may

if $c_k \to +\infty$, — compute x^*

Penalty method: basic version cont.

Theorem 9.1: (Convergence of penalty method: basic version)
Consider (1) and suppose that $\inf f > -\infty$. Let $\{x^k\}$ be generated by the penalty method on Slide 7. Then any accumulation point x^* of $\{x^k\}$ is a globally optimal solution of (1).

日期:

Barrier method

(interior)

Recall that

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{Minimize}} & f(x) \\ \text{subject to} & g_i(x) \le 0, \quad i \in I. \end{array} \tag{2}$$

One standard way is to make use of the log-barrier function:

$$B(x) := -\sum_{i=1}^{m} \ln[-g_i(x)]. \qquad -g_i(x) > 0$$

Then one minimizes $f_\mu(x) := f(x) + \mu B(x)$ for some $\mu > 0$.

x_0

$S^0 \triangleq \{x \mid g_i(x) < 0\}$

slater point $g_i(x^0) < 0$

Barrier method: basic version

Barrier method for (2): basic version

Let $x^0 \in S^0$, $\mu > 0$ and $\eta > 1$. Set $\mu_1 = \mu$. For $k = 1, \ldots,$

- Find a minimizer x^k of
$$f_{\mu_k}(x) := f(x) - \mu_k \sum_{i=1}^{m} \ln[-g_i(x)],$$
using x^{k-1} as the initial point for the iterative method.
- Update $\mu_{k+1} = \mu_k/\eta$.

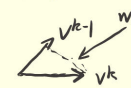
$\{\mu_k\} \searrow$ to 0

日期: /

Gram-Schmidt process

Theorem 10.1: (Gram-Schmidt process)
Given a set of linearly independent vectors $\{v^0, \dots, v^k\} \subset \mathbb{R}^n$.
Set $w^0 = v^0$ and for each $j = 1, \dots, k$

$$w^k = v^k - \sum_{j=0}^{k-1} \frac{v^{k\,T} w^j}{\|w^j\|_2^2} w^j$$



Then $w^i \neq 0$ for all i. Moreover, ${w^i}^T w^j = 0$ whenever $i \neq j$, and for each $i = 0, 1, \dots, k$, it holds that

$$\text{Span}\{v^0, \dots, v^i\} = \text{Span}\{w^0, \dots, w^i\}.$$

seem as inner product $\langle v, w \rangle = v^T \cdot A \cdot w$

(Generalized) Gram-Schmidt process

Theorem 10.2: ((Generalized) Gram-Schmidt process)
Given $A \in \mathbb{R}^{n \times n}$ with $A \succ 0$ and a set of linearly independent vectors $\{v^0, \dots, v^k\} \subset \mathbb{R}^n$. Set $w^0 = v^0$ and for each $j = 1, \dots, k$

$$w^k = v^k - \sum_{j=0}^{k-1} \frac{v^{k\,T} A w^j}{{w^j}^T A w^j} w^j \quad \Rightarrow$$

that means, $w^i \perp w^j$ in $\langle \cdot, \cdot \rangle$ inner product def

Then $w^i \neq 0$ for all i. Moreover, ${w^i}^T A w^j = 0$ whenever $i \neq j$, and for each $i = 0, 1, \dots, k$, it holds that

$$\text{Span}\{v^0, \dots, v^i\} = \text{Span}\{w^0, \dots, w^i\}.$$

Conjugate gradient method: Conceptual version

Conjugate gradient method: Conceptual version
Start at $x^0 \in \mathbb{R}^n$ and $d^0 = -\nabla f(x^0) = b - Ax^0$. (see note)
For each $k = 0, 1, 2, \dots$,

- If $d^k = 0$, terminate.
- Pick α_k so that

$$\alpha_k \in \text{Arg min}\{f(x^k + \alpha d^k) : \alpha \geq 0\}.$$

- Set $x^{k+1} = x^k + \alpha_k d^k$ and

$$d^{k+1} = -\nabla f(x^{k+1}) - \sum_{j=0}^{k} \frac{[-\nabla f(x^{k+1})]^T A d^j}{{d^j}^T A d^j} d^j$$

To do: upper to n orthogonal $\{d^k\}_{k=1}^n$.

- Prove the correctness: i.e. when $d^k = 0$, what happens?

Thm: $\nabla f(x^i)^T \nabla f(x^j) = 0$, ${d^i}^T \cdot \nabla f(x^j) = 0$, $\forall\, i < j$

$\Rightarrow$ span $\{\nabla f(x^1), \dots, \nabla f(x^n)\}$ = span $\{d^1, \dots, d^n\}$

$\Rightarrow$ at most, n iteration can solve the problem.

日期:

Theorem 10.3:
Let $A \succ 0$ and $x^0 \in \mathbb{R}^n$. Set $d^0 = -\nabla f(x^0)$. For $k = 0, 1, \ldots,$ suppose that $d^0, \ldots, d^k \neq 0$, where for each $i = 0, \ldots, k-1$,

$$d^{i+1} = -\nabla f(x^{i+1}) - \sum_{j=0}^{i} \frac{[-\nabla f(x^{i+1})]^T A d^j}{d^{j\,T} A d^j} d^j,$$

with $x^{i+1} = x^i + \alpha_i d^i$ and α_i coming from exact line search. Then $[\nabla f(x^i)]^T \nabla f(x^{k+1}) = 0$ and $d^{i\,T} \nabla f(x^{k+1}) = 0$ for $j < k+1$.

Proof: The proof is by induction. Let $k = 0$ and $d^0 \neq 0$. Then $d^0 = -\nabla f(x^0)$, and $d^{0\,T} \nabla f(x^1) = 0$ holds because of exact line search.

$\nabla f(x^0)^T \cdot \nabla f(x^1) = 0$

Conjugate gradient method: Formal version
Start at $x^0 \in \mathbb{R}^n$ and $d^0 = -\nabla f(x^0) = b - Ax^0$.
For each $k = 0, 1, 2, \ldots,$

- If $d^k = 0$, terminate.
- Pick α_k so that

$$\alpha_k \in \operatorname{Arg\,min}\{f(x^k + \alpha d^k) : \alpha \geq 0\}.$$

- Set $x^{k+1} = x^k + \alpha_k d^k$ and

$$d^{k+1} = -\nabla f(x^{k+1}) + \frac{\|\nabla f(x^{k+1})\|_2^2}{\|\nabla f(x^k)\|_2^2} d^k$$

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

Conjugate gradient method: Actual version

Conjugate gradient method: Actual version
Start at $x^0 \in \mathbb{R}^n$ and $r^0 = d^0 = b - Ax^0$.
For each $k = 0, 1, 2, \ldots,$

- If $\|r^k\|$ (or, less commonly, $\|d^k\|$) is below a tolerance, terminate.
- (Exact line search) Compute

$$\alpha_k = \frac{r^{k\,T} r^k}{d^{k\,T} A d^k}, \quad x^{k+1} = x^k + \alpha_k d^k, \quad r^{k+1} = r^k - \alpha_k A d^k.$$

- (Update d^{k+1}) Compute

$$\beta_k = \frac{r^{k+1\,T} r^{k+1}}{r^{k\,T} r^k}, \quad d^{k+1} = r^{k+1} + \beta_k d^k.$$

Remark: β_k not saved

- One matrix-vector multiplication per iteration if Ad^k is saved.
- Keeping track of four vectors, x^k, r^k, d^k and the Ad^k saved.

日期:

Convergence rate

Conjugate gradient method must terminate in at most n iterations: because $\{d^0, \ldots, d^{n-1}\}$ must be a basis of $\mathbb{R}^n$, if $d^k \neq 0$ for each k. Then necessarily, $d^n = 0$.

In practice, conjugate gradient method can converge much more quickly. We state the following result without proof. See Theorem 5.5 of Ref 2.

Theorem 10.4: (Luenberger)
Consider the conjugate gradient method for minimizing $f(x) = \frac{1}{2}x^T Ax - b^T x$ for some $b \in \mathbb{R}^n$ and $A \succ 0$. Let $\{x^k\}$ be the sequence generated and let x^* be the minimizer of f. If A has eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, then

$$(x^{k+1} - x^*)^T A(x^{k+1} - x^*) \leq \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 (x^0 - x^*)^T A(x^0 - x^*)$$

↳ $\geq \lambda_1 \cdot \|x^{k+1} - x^*\|_2^2$

so if $\lambda_{n-k} - \lambda_1 = 0$, we have $x^{k+1} = x^*$.

Thm: ↗ that means $\{x^k\} = x^*$ from x^{k+1}.