— Lecture Notes

L- Lipschitz Continuity $\|f(x)-f(y)\| \le L\|x-y\|$

L- smothness $\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|$ iff $-LI \le \nabla^2 f(x) \le LI$ iff $|\lambda| \le L$

proof : $\|\nabla f(x)-\nabla f(y)\| = \|\nabla^2 f(z)(x-y)\| \le L\|x-y\|$

Thm. f is L- smooth $\longrightarrow$ $f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$ (Taylor Expanation)

Thm. GD : $f(x_{k+1}) - f(x_k) \le -t\left(1-\frac{Lt}{2}\right)\|\nabla f(x_k)\|^2 \overset{t\in(0,\frac{1}{L})}{\le} -\frac{t}{2}\|\nabla f(x_k)\|^2$ Significance : $\|\nabla f(x_k)\|^2 \le \frac{2}{t}[f(x_k)-f(x_{k+1})]$

$f(x_k) - f(x^*) \le \frac{\|x_0-x^*\|^2}{2tk}$ Using $\|x_{k+1}-x^*\|^2$

$x = x^*$ is useful

m - strongly convex $f(x) - \frac{m}{2}x^2$ is convex iff $f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2$ iff $\nabla^2 f(x) \succeq mI$ iff $\lambda \ge m$

Thm. $\|x_{k+1}-x^*\|^2 \le (1-mt)\|x_k-x^*\|^2 \le (1-mt)^{k+1}\|x_0-x^*\|^2$

Some Tuning

1、$t_k = \operatorname{argmin}_t \{f[x_k - t\nabla f(x_k)]\}$ $f(x_k)-f(x^*) \le (1-\frac{m}{L})^k[f(x_0)-f(x^*)]$ (2 Inequalities + Minimization)

2、Armijo's rule : backtracking line search $f(x_k) - f(x_k+td_k) \ge \alpha t\|\nabla f(x_k)^T d_k\|$ (usually using $t \longleftarrow \beta t$)

3、AGD $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$, $y_{k+1} = x_{k+1} + \beta_k(x_{k+1}-x_k)$

4、Newton $x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$

Thm. $\|x_k - x^*\| \le \frac{2m}{M}\left(\frac{2M}{m}\|x_0-x^*\|\right)^{2k}$ $\|A\|_{a,b} = \max_x \frac{\|Ax\|_b}{\|x\|_a}$

Damped Newton's method : $f(x+td) > f(x) + \alpha t\,\nabla f(x)^T d$

Proximal GD

$x_{k+1} = \operatorname{argmin}_x \frac{1}{2t_k}\|x-(x_k-\nabla F(x_k))\|_2^2 \longrightarrow x_{k+1} = \operatorname{argmin}_x\{\frac{1}{2}\|x-(x_k-t_k f(x_k))\|_2^2 + t_k h(x)\}$

$\operatorname{prox}_h(x) \triangleq \operatorname{argmin}_z\{\frac{1}{2}\|z-x\|_2^2 + h(z)\}$, then $x_{k+1} = \operatorname{prox}_{t_k h}(x_k - t_k\nabla f(x_k))$

Convergence Analysis $\|x^*-x_k\|_2^2 \le (1-\frac{m}{L})^k\|x^*-x_0\|_2^2$

Lagrange ($\min f(x)$, $Ax=b$)

Thm. x^* is optimal iff $\nabla f(x^*) \perp \mathrm{Null}(A)$

$\mathrm{Null}(A) = \{x | Ax = 0\}$, $\mathrm{Range}(A) = \{Av : v\in\mathbb{R}^k\}$, $\mathrm{Null}(A)^\perp = \mathrm{Range}(A^T)$

$\mathcal{L}(x,\lambda) = f(x) + \lambda^T(Ax-b)$, Lagrange Equation $\nabla\mathcal{L} = 0$

A Regular point x of a function h if $\nabla h(x) \ne 0$; $\vec{h}$ if $\nabla h_i(x)$ are linearly independent.

ICP : $g_j(x) \le 0$

Karush-Kuhn-Tucker conditions

If x^* is a local minimum of ICP and also a regular point , then :

1、$\nabla f(x^*) + \sum_{i=1}^{k}\lambda_i\nabla h_i(x^*) + \sum_{j=1}^{m}\mu_j\nabla g_j(x^*) = 0$

2、$\mu_j \ge 0$

3、$\mu_j g_j(x^*) = 0$ —— complementary slackness condition

Sufficiency of KKT conditions for convex problem

Projected GD $x_{k+1} = P_X(x_k - t_k\nabla f(x_k)) \triangleq x_k - tg(x_k)$

Connection to proximal GD $I_X \triangleq \begin{cases} 0 & ,\text{if } x\in X \\ +\infty & ,\text{if } x\notin X \end{cases}$, $\operatorname{prox}_{I_X}(y) = P_X(y)$

Newton's method for ECP : $\min_x f(x) = \frac{1}{2}x^TQx + g^Tx + c$ s.t. $Ax=b$ ($Q \succeq 0$, A has full column rank)

KKT system $\underbrace{\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix}}_{\text{KKT matrix}}\begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -g \\ b \end{bmatrix}$

KKT matrix $\triangleq K$ Lemma : $\mathrm{Null}(K) = \{\begin{bmatrix} x \\ 0 \end{bmatrix} : x\in \mathrm{Null}(Q)\cap\mathrm{Null}(A)\}$

Thm. if KKT system has no solution , then QP is either infeasible or unbounded below

Thm. if K is nonsingular , then Unique solution .

Following conditions are equivalent :

1、K is nonsingular

2、$\mathrm{Null}(Q)\cap\mathrm{Null}(A) = \{0\}$

3、$Ax=0 \wedge x\ne 0 \Rightarrow x^TQx > 0$

4、$F^TQF \succ 0$ for any $F\in\mathbb{R}^{n\times(n-k)}$ s.t. $\mathrm{Range}(F) = \mathrm{Null}(A)$

Particularly , if $Q \succ 0$, then K is nonsingular

$\min_d \; h(d) \triangleq \hat{f}(x_k+d) + \nabla f(x_k)^T d + \frac{1}{2}d^T\nabla^2 f(x_k)d$

s.t. $A(x_k+d) = b \longleftrightarrow Ad = 0$

Now we have $\begin{bmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x_k) \\ 0 \end{bmatrix}$

Elimination $x = \hat{x} + Fz$

Log barrier function $B(x) = -\sum_{i=1}^{m}\log(-g_i(x))$

$\Longrightarrow f(x) + \frac{1}{t}B(x)$ central point and central path $x^*(t)$, $t \to +\infty$

Seems $\frac{m}{t}$ - suboptimal

solution : an increasing sequence $\{t_k\}$ with $x^*(t_{k-1})$ as the initial point

Dual LP

$$\begin{aligned} &\min_x \; f(x) = c^Tx & & \max_{\lambda,\mu} \; \psi(\lambda,\mu) = \lambda^Tb + \mu^Th \\ &\text{s.t.} \quad Ax = b & \xleftrightarrow{\text{Dual}} \quad & \text{s.t.} \quad A^T\lambda + G^T\mu = 0 \\ &\qquad\quad Gx \ge h & & \qquad\quad \mu \ge 0 \end{aligned}$$

weak duality $f(x) \ge \psi(\lambda,\mu)$ Strong duality $f(x^*) = \psi(\lambda^*,\mu^*)$

$W_1(a,b) = \max_{h \text{ is 1-Lipschitz}} E_{Y\sim b}h(Y) - E_{X\sim a}h(X)$ wasserstein distance

Lagrange dual function

$\phi(\vec{\lambda},\vec{\mu}) = \inf_{x\in D}\left(f(x) + \sum\lambda h(x) + \sum\mu g(x)\right)$ downplay the role of D

Always concave . (λ,μ) is feasible if $\mu \ge 0$ and $\phi(\lambda,\mu) > -\infty$

Slater's condition

convex problem :

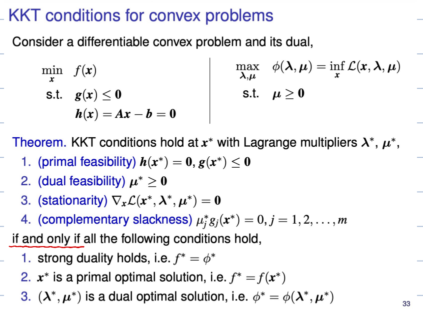
$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_j(x) \le 0 , \; j = 1.2.\dots.m \\ & Ax = b \end{aligned}$$

The above problem is strictly feasible , i.e. $\exists x\in \operatorname{int} D$, s.t. $g_j(x) < 0 \wedge Ax = b$.

Refined : g_i is affine $\longrightarrow g_i(x) \le 0$

Thm. Strong duality holds for (CP) under Slater's condition.

KKT conditions



KKT conditions for convex problems

Consider a differentiable convex problem and its dual,

$$\begin{aligned} \min_x \quad & f(x) \\ \text{s.t.} \quad & g(x) \le 0 \\ & h(x) = Ax - b = 0 \end{aligned} \qquad \begin{aligned} \max_{\lambda,\mu} \quad & \phi(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) \\ \text{s.t.} \quad & \mu \ge 0 \end{aligned}$$

Theorem. KKT conditions hold at x^* with Lagrange multipliers λ^*, μ^*,

1. (primal feasibility) $h(x^*) = 0, g(x^*) \le 0$
2. (dual feasibility) $\mu^* \ge 0$
3. (stationarity) $\nabla_x\mathcal{L}(x^*,\lambda^*,\mu^*) = 0$
4. (complementary slackness) $\mu_j^* g_j(x^*) = 0, j = 1, 2, \dots, m$

if and only if all the following conditions hold,

1. strong duality holds, i.e. $f^* = \phi^*$
2. x^* is a primal optimal solution, i.e. $f^* = f(x^*)$
3. (λ^*,μ^*) is a dual optimal solution, i.e. $\phi^* = \phi(\lambda^*,\mu^*)$

33

— For Review

Positive definite $A \succ 0$ iff $v^TAv > 0$ iff $\lambda_i > 0$ iff $|$顺序主子式$| > 0$

[Convexity] $\theta x + (1-\theta)y$

Affine set $\{x \mid Ax = b\}$ Hyperplane $a^Tx = b$ Halfspace $a^Tx \leq b$

Ellipsoids $\{x \mid (x-x_c)P^{-1}(x-x_c) \leq 1\} = \{x_c + Au \mid \|u\|_2 \leq 1\}$ Polyhedron $\{x \mid Ax = b \wedge A'x \leq b'\}$

Convex combination $\sum \theta_i x_i$, where $\theta_i \geq 0$, $\sum \theta_i = 1$

Convex Hull $\text{conv}\, S = \{\sum_{i=1}^{m} \theta_i x_i : \exists m \in \mathbb{N}, x_{1 \cdots m} \in S, \sum_{i=1}^{m} \theta_i = 1\}$

Convexity-preserving operations : Intersections , affine transformation

Def. $\text{int}\, C$. $\overline{C}$ ($\text{cl}\, C$) . ∂C (boundary) , $P_C(x) = \text{argmin}_{x \in C} \|x - z\|_2$

Thm. Supporting / Separating hyperplane (BW Thm. , $C_0 = C_1 - C_2$)

[Convex functions] $f(\theta x + \bar{\theta} y) \leq \theta f(x) + \bar{\theta} f(y)$

Def $\text{epi}(f) = \{(x,t) \mid t \geq f(x)\}$; f is m-strongly convex if $f - \frac{m}{2}\|x\|_2^2$ is convex

Criteria : Zero-th order condition $\forall x, v$ $g(t) = f(x + tv)$ is convex . $D = \{t \mid x + tv \in D_f\}$

First order condition $\forall x, x_0$ $f(x) \geq f(x_0) + [\nabla f(x_0)]^T (x - x_0)$

Second order condition $\forall x$ $\nabla^2 f(x) \succeq 0$ ($\nabla^2 f(x) \succ 0$ 不必要 $f(x) = x^4$)

Convexity preserving operations

$\sum c_i f_i$ | $f(Ax+b)$ | $f(g_1(x), \cdots, g_m(x))$, f is increasing , g_i is convex | $\max_i f_i$ or $\sup_i f_i(x)$ | $\inf_{y \in C} g(x,y)$, convex g , convex C

[Convex optimization] $\min f(x)$

s.t. $g(x) \leq 0$, $Ax = b$ ⟶ Generating feasible set

Methods of Transformations : Eliminating | Introducing slack variables |

Optimality conditions $\nabla f(x^*) = 0$, $\nabla f(x^*)(x - x^*) \geq 0$, $\begin{cases} \nabla f(x^*) + A^T \lambda^* = 0 \\ Ax^* = b \end{cases}$

LP

	Standard	Inequality
$\min_x \; c^Tx$	$\min_x \; c^Tx$	$\min_x \; c^Tx$
s.t. $Bx \leq d$ $Ax = b$	s.t. $Ax = b$ $x \geq 0$	s.t. $Ax \leq b$

$Ax \leq b \longrightarrow Ax + c = b$, $c \geq 0$

$x \longrightarrow x^+ - x^-$, $x^+ \geq 0$, $x^- \geq 0$

$Ax = b \longrightarrow b \leq Ax \leq b$

regular point x : $\nabla g_1(x), \nabla g_2(x), \cdots, \nabla g_m(x)$ linear independent

[Lagrange] $\mathcal{L}(x, \lambda, u) = f(x) + \sum \lambda_i h_i(x) + \sum u_i g_i(x) = f(x) + \lambda^T h(x) + u^T g(x)$

$\mathcal{L}(x, \lambda, u) \leq f(x)$ for $u \geq 0$

KKT conditions :

1、 primal feasibility $h(x^*) = 0$, $g(x^*) \leq 0$

2、 dual feasibility $u^* \geq 0$

3、 stationarity $\nabla_x \mathcal{L}(x^*, \lambda^*, u^*) = 0$

4、 complementary slackness $u_j^* g_j(x^*) = 0$

$\min_x f(x)$ s.t. $g_j(x) \overset{\text{significant}}{\leq} 0$, $h_i(x) = 0$

Duality : $\phi(\lambda, u) = \inf_{x \in D} \mathcal{L}(x, \lambda, u)$ Domain : $\{(\lambda, u) \mid \phi(\lambda, u) > -\infty\}$

Dual problem $\max_{\lambda, u} \phi(\lambda, u)$ s.t. $u \geq 0$

Strong daulity $\phi^* = f^*$ (holds for LP)

Slater's condition ($\exists x_0 \in \text{int}\, D$ s.t. $g_j(x_0) < 0$ for all nonlinear inequality constraints g_i) $\Longrightarrow$ strong duality

Thm. KKT $\longleftrightarrow$ Strong duality + primal optimality + dual optimality

[Algorithms]

— smooth f

descent method $x_{k+1} = x_k + t d_k$

descent direction $\begin{cases} d_k = -\nabla f(x_k) \\ d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k) \end{cases}$

stepsize t_k $\begin{cases} \text{constant } (\frac{1}{L} \text{ for } L\text{-smooth}) \\ \text{exact line search} \text{ —— } t_k = \text{argmin}_t f(x_k + t d_k) \\ \text{backtracing line search} \text{ —— } t \longleftarrow \beta t \text{ until } f(x_k + t d_k) > f(x_k) + \alpha t \nabla f(x_k)^T d_k \end{cases}$

— constrained problem

elimination

Newton's method $\longrightarrow$ $\overset{\text{KKT matri}}{\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix}} \begin{bmatrix} d \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}$ KKT system