Lecture 18. KKT system and Newton's method.

Optimization problem: $\min f(x)$ subject to $g_1(x) = \cdots = g_m(x) = 0$

Lagrangian function $L(x, \lambda) = f(x) + \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x)$.

Lagrange condition: for ECP. x^* is optimal iff $\exists \lambda^*$. $\nabla L(x^*, \lambda^*) = 0$

in general. if x^* is regular. optimal only if $\exists \lambda^*$. $\nabla L(x^*, \lambda^*) = 0$.

Remark. For ECP, any feasible solution is regular.

Question: Does $\nabla L(x^*, \lambda^*) = 0$ mean $(x^*, \lambda^*) = \operatorname{argmin} L(x, \lambda)$?

The answer is no. since L is not convex. Although for any fixed

λ, $L(x, \lambda)$ is a convex function of x.

In fact, L is unbounded below. consider $\begin{array}{ll} \min & (x_1 - x_2)^2 \\ \text{s.t.} & x_1 - x_2 - 5 = 0 \end{array}$.

If $\nabla L(x^*, \lambda^*) = 0$. (x^*, λ^*) is a saddle point.

Proposition. If $\nabla L(x^*, \lambda^*) = 0$. $f(x^*) = L(x^*, \lambda^*) = \max\limits_{\lambda} \min\limits_{x} L(x, \lambda)$.

Fix λ^* (e.g. $\lambda^* = -10$). $\hat{L}(x_1, x_2) = L(x_1, x_2, \lambda^*)$ is convex.

$\nabla \hat{L}(x_1, x_2) = \nabla_x L(x_1, x_2, \lambda^*)$. So $\nabla \hat{L}(x^*) = 0$. $x^* = \operatorname{argmin} L(x, \lambda^*)$

Let Ω be the feasible set $\{x : g_1(x) = g_2(x) = \cdots = g_m(x) = 0\}$.

$\forall \lambda \in \mathbb{R}^m$. $\min\limits_{x} L(x, \lambda) \le \min\limits_{x \in \Omega} L(x, \lambda) = \min\limits_{x \in \Omega} f(x) = f(x^*)$.

$\exists \lambda = \lambda^*$. $\min\limits_x L(x,\lambda) = f(x^*)$ So $f(x^*) = \max\limits_\lambda \min\limits_x L(x,\lambda)$.

Solving equality constrained optimization.

Warmup: quadratic problem $\begin{cases} \min & \frac{1}{2}x^TQx + w^Tx. \quad Q \geq 0 \\ \text{subject to} & Ax - b = 0 \end{cases}$

Lagrangian. $L(x,\lambda) = \frac{1}{2}x^TQx + w^Tx + \lambda^T(Ax-b)$.

Lagrange condition: $\begin{cases} \nabla_x L(x,\lambda) = Qx + w - A^T\lambda = 0 \\ \nabla_\lambda L(x,\lambda) = Ax - b = 0 \end{cases}$ $\longrightarrow$ $\begin{cases} \text{Karush} \\ \text{Kuhn} \\ \text{Tucker.} \end{cases}$

$\Rightarrow \begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} -w \\ b \end{pmatrix}$ called KKT system

$\nwarrow$ KKT matrix. solving KKT system: Gaussian elimination

row 2 $\leftarrow AQ^{-1}$ row 1 $-$ row 2 : $(0, AQ^{-1}A^T)\begin{pmatrix} x \\ \lambda \end{pmatrix} = b + AQ^{-1}w$

$\Rightarrow \lambda = (AQ^{-1}A^T)^{-1}(b + AQ^{-1}w)$. provided $AQ^{-1}A^T$ invertible

$Qx + A^T\lambda = -w \Rightarrow x = -Q^{-1}w - Q^{-1}A^T(AQ^{-1}A^T)^{-1}(b + AQ^{-1}w)$.

We hope $(AQ^{-1}A^T)^{-1}$ exists. It is always true if $Q > 0$, $\text{rank}(A) = m$.

(since $Q^{-1} > 0$, $AQ^{-1}A^Tv = 0 \Rightarrow v^TAQ^{-1}A^Tv = 0 \Rightarrow A^Tv = 0 \Rightarrow v = 0$)

Theorem. The KKT matrix is invertible (nonsingular). if and only if

① $\ker(A) \cap \ker(Q) = \{0\}$. ($Q$, A have no nontrivial kernel). or nullspace

② if $Ax = 0$ and $x \neq 0$. then $x^TQx > 0$ ($Q > 0$ on $\ker(A)$). or

③ $F^TQF > 0$ for $\forall F \in \mathbb{R}^{n\times(n-m)}$ s.t. $\text{im}(F) = \{Fv : v \in \mathbb{R}^{n-m}\} = \ker(A)$.

Proof. "$\Rightarrow$". if $0 \neq x \in \ker(Q) \cap \ker(A)$. $\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = 0$.

"$1 \Rightarrow 2$". $Ax = 0$. $x \neq 0 \Rightarrow Qx \neq 0$. but $x^T Q x = 0 \Rightarrow Qx = 0$. (?)

"$\Leftarrow$". Assume $\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = 0$. need to show $v = w = 0$.

$v^T Q v = v^T(-A^T w) = -(Av)^T w = 0$. contradicts if $Av = 0$. $v \neq 0$

$Qv + A^T w = 0 \Rightarrow A^T w = 0$. contradicts $\mathrm{rank}(A) = m$ if $w \neq 0$.

"$2 \Leftrightarrow 3$". $\mathrm{im}(F) = \ker(A)$. $0 \neq y \in \ker(A) \Leftrightarrow \exists x \neq 0.\ y = Fx$. $\square$

Proposition. Suppose $Q \succeq 0$. Then $Qx = 0$ iff $x^T Q x = 0$.

Proof: Consider the eigendecomposition (orthonormal diagonlization) of Q: $Q = U \Lambda U^T = \sum_{i=1}^{n} \xi_i u_i u_i^T$ $\xi_i \geq 0$ be Q's eigenvalues. Then

$x^T Q x = \sum_{i=1}^{n} \xi_i \| u_i^T x \|^2 = 0 \Rightarrow u_i^T x = 0 \Rightarrow Qx = \sum u_i (\xi_i u_i^T x) = 0$. $\square$

Remark. $\mathrm{rank}(A) = m$ is guaranteed. otherwise $Ax - b$ is redundant or infeasible. Quadratic problem also requires $Q \succeq 0$. but KKT system may not be solvable if $Q \not\succeq 0$. If so. $f(x)$ is unbounded below.

For general optimization problem (not quadratic). recall Newton's:

given x_k. $f(x) \approx \tilde{f}(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$.

$x_{k+1} = \mathrm{argmin}\ \tilde{f}(x)$. if no constraints. Now let $x_{k+1} = \underset{Ax-b=0}{\mathrm{argmin}}\ \tilde{f}(x)$.

Let $d = x_{k+1} - x_k$. Then $\tilde{f}(x) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d$

constraints : $Ad = 0$, since $Ax_{k+1} = Ax_k = b$.

KKT system : $\begin{pmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ 0 \end{pmatrix}$. $x_{k+1} = x_k - d$.

Question : stopping criterion ? $\|\nabla f(x_k)\| \leq \delta$? $d^T \nabla^2 f(x_k) d \leq \delta$ ✓

Example. min $f(x_1, x_2) = x_1^2 + x_2^2$. s.t. $x_1 + x_2 = 1$. start : $(1,0)^T$.

KKT system : $\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} d_1 \\ d_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} -1/2 \\ 1/2 \\ -1 \end{pmatrix}$

Example. min $f(x_1, x_2) = x_1^2$. s.t. $x_1 + 2x_2 = b$. start : $(b, 0)^T$.

KKT system : $\begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} -2b \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} d_1 \\ d_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} -b \\ b/2 \\ 0 \end{pmatrix}$

$\ker(\nabla^2 f) = s \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. $s \in \mathbb{R}$. $\ker(A) = t \cdot \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ $t \in \mathbb{R}$. $F = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$

Example. min $f(x_1, x_2) = e^{x_1^2 + x_2^2}$ s.t. $x_1 + x_2 = 1$. start : $(1, 0)^T$

KKT system : $\begin{pmatrix} 6e & 0 & 1 \\ 0 & 2e & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} -2e \\ 0 \\ 0 \end{pmatrix} \Rightarrow \begin{pmatrix} d_1 \\ d_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} -1/4 \\ 1/4 \\ -e/2 \end{pmatrix}$

$\nabla f(x_1, x_2) = \left(2x_1 e^{x_1^2 + x_2^2}, 2x_2 e^{x_1^2 + x_2^2}\right)^T$. $\nabla^2 f(x_1, x_2) = e^{x_1^2 + x_2^2} \begin{pmatrix} 4x_1^2 + 2, & 4x_1 x_2 \\ 4x_1 x_2, & 4x_2^2 + 2 \end{pmatrix}$

Correctness. 1. $\nabla f(x_k)^T d \leq 0$ 2. x_k optimal if $d^T \nabla^2 f(x_k)^T d = 0$

1. KKT system $\begin{cases} \nabla^2 f(x_k) d + A^T \lambda = -\nabla f(x_k) & ① \\ Ad = 0 & ② \end{cases}$

$d^T ① \Rightarrow d^T \nabla^2 f(x_k) d + d^T A^T \lambda = -d^T \nabla f(x_k) \Rightarrow \nabla f(x_k) \leq 0$

2. If $d^T \nabla^2 f(x_k) d$. then $\nabla^2 f(x_k) d = 0$ by the previous proposition.

So $\nabla^2 f(x_k) d + A^T\lambda = -\nabla f(x_k) \Rightarrow \nabla f(x_k) + A^T\lambda = 0$. (Lagrange)

Convergence: comparison to unconstrained cases.

Example: $\min f(x_1, x_2) = x_1^2 + x_2^2$. s.t. $x_1 + x_2 = 1$. $\min x_1^2 + (1-x_1)^2$.

In general. feasible set $\Omega = \{x : Ax = b\} = \{\tilde{x} + Fz : z \in \mathbb{R}^{n-m}\}$

$\min f(x)$. s.t. $Ax = b \iff \min g(z) = f(\tilde{x} + Fz)$

$\nabla g(z) = F^T \nabla f(\tilde{x} + Fz)$. $\nabla^2 g(z) = F^T \nabla^2 f(\tilde{x} + Fz) F$.

Apply Newton. $z_{k+1} = z_k - (\nabla^2 g(z))^{-1} \nabla g(z)$. claim: $x_k = \tilde{x} + Fz_k$.

By induction. Let dx_k. dz_k be descending directions.

dz_k exists iff $\nabla^2 g(z_k)$ invertible iff KKT invertible iff dx_k exists.

Note that $Ad_{x_k} = 0$. $\mathrm{im}(F) = \ker(A)$. so $\exists u \in \mathbb{R}^{n-m}$. $dx_k = Fu$.

Moreover. $\nabla^2 f(x_k) dx_k + A^T\lambda = -\nabla f(x_k)$. so $\nabla^2 f(x_k) Fu + A^T\lambda = -\nabla f(x_k)$.

Multiply F^T on both sides: $F^T \nabla^2 f(x_k) Fu + F^T A^T \lambda = -F^T \nabla f(x_k)$.

$\ker(A) = \mathrm{im}(F) \Rightarrow \forall v,\ AFv = 0 \Rightarrow AF = \vec{0}$.

So $F^T \nabla^2 f(x_k) Fu = -F^T \nabla f(x_k) \Rightarrow u = -(\nabla^2 g(z_k))^{-1} \nabla g(z_k) = dz_k$.

$dx_k = Fu = Fdz_k \Rightarrow x_{k+1} = x_k + dx_k = \tilde{x} + Fz_k + Fdz_k = \tilde{x} + Fz_{k+1}$.