Lecture 19. Newton's method. KKT condition for ICP.

Recall Newton's method. $\min_d \tilde{f}(x_k+d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T \nabla^2 f(x_k) d$.

Apply to equality constrained problems: $\min_d \tilde{f}(x_k+d)$ s.t. $Ad = 0$.

KKT system for this quadratic problem is: $\begin{pmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d \\ -\lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ 0 \end{pmatrix}$

Newton's method for equality constrained problem: select x_0 s.t. $Ax_0 = b$.

repeat. compute d by solving KKT. $\begin{pmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d \\ -\lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ 0 \end{pmatrix}$

(note if $A=0$. then $d = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$. exactly the same as before.)

(again use backtracking line search). $t \leftarrow$ initial t_0 (usually $t_0 = 1$).

while $f(x_k + td) > f(x_k) + \alpha t \nabla f(x_k)^T d$ do $t \leftarrow \beta t$.

$x_{k+1} \leftarrow x_k + td$. until $\|\nabla f(x_k)\| \leq \delta$? $d^T \nabla^2 f(x_k) d \leq \delta^2$. ✓.

Note it is a feasible descent method. since all x_k are feasible.

and $f(x_{k+1}) < f(x_k)$ unless x_k is optimal. (why?)

$$\left.\begin{aligned} \nabla^2 f(x_k) d - A^T \lambda &= -\nabla f(x_k) \\ Ad &= 0 \end{aligned}\right\} \Rightarrow \begin{aligned} & d^T \nabla^2 f(x_k) d - \underbrace{d^T A^T \lambda}_{=0} = -\nabla f(x_k)^T d \\ & \nabla^2 f(x_k) \succeq 0 \longrightarrow \nabla f(x_k)^T d \leq 0. \end{aligned}$$

By second-order Tayor. $f(x_k + td) \approx f(x_k) + t \nabla f(x_k)^T d + \frac{t^2}{2} \underbrace{d^T \nabla^2 f(x_k) d}_{-\nabla f(x_k)^T d}$.

Moreover. $d^T \nabla^2 f(x_k) d \geq \lambda_{\min} d^T d$. so $\frac{t^2}{2} d^T \nabla^2 f(x_k) d + o(t^2 d^T d) > (1-\alpha) t \nabla f(x_k)^T d$

$\forall$ sufficiently small t.

So if $d^T\nabla^2 f(x) d \le \delta$. Newton's method halts. o.w. backtracking halts.

If $d^T\nabla^2 f(x_k) d = 0$. $\nabla^2 f(x_k) d - A^T\lambda = -\nabla f(x_k)$. we claim $\nabla^2 f(x_k) d = 0$.

Proof. if $Q \succeq 0$. $Q = U^T \Lambda U$. $\Lambda = \mathrm{diag}\{\xi_1, \dots, \xi_n\}$. $U = \{u_1, \dots, u_n\}$.

$$d^T Q d = d^T\left(\sum \xi_i u_i u_i^T\right) d = \sum \xi_i (u_i^T d)^T (u_i^T d) = \sum \xi_i (u_i^T d)^2.$$

$$d^T Q d = 0 \Rightarrow \forall i,\ \xi_i = 0 \text{ or } u_i^T d = 0 \Rightarrow Qd = \sum \xi_i u_i u_i^T d = 0. \quad \square$$

Note $Ax = \begin{pmatrix} a_1^T x \\ \vdots \\ a_m^T x \end{pmatrix}$. $A\lambda = (a_1, \dots, a_m)\lambda = \sum \lambda_i a_i$. So $\nabla f(x_k) = \sum \lambda_i a_i$

Lagrange condition for $\min f(x)$. s.t. $Ax = b$. so x_k is optimal.

Implementation: Gaussian elimination. assume KKT matrix nonsingular.

nonsingular:

① ∘ $\ker(Q) \cap \ker(A) = \{0\}$. Q and A have no nontrivial common kernel.

② ∘ $Ax = 0$. $x \ne 0 \Rightarrow x^T Q x > 0$. $Q > 0$ on $\ker(A)$.

③ ∘ $F^T Q F > 0$ for $\forall F \in \mathbb{R}^{n\times(n-m)}$. s.t. $\mathrm{Im}(F) = \{Fv : v \in \mathbb{R}^{n-m}\} = \ker(A)$

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

"$1 \Rightarrow 2$". $Ax = 0$. $x \ne 0 \Rightarrow Qx \ne 0$. 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 $Av = 0$. $v \ne 0$.

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

"$2 \Leftrightarrow 3$". $Im(F) = \ker(A)$. $0 \neq y \in \ker(A) \Leftrightarrow \exists x \neq 0. \; y = Fx$. $\square$

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

KKT system $\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ -\lambda \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 0 \end{pmatrix} \Rightarrow (d_1, d_2, \lambda) = (-\frac{1}{2}, \frac{1}{2}, 0)$.

set step size $t = 1/2$. next iteration: $(1,0)^T + t(d_1, d_2)^T = \begin{pmatrix} 3/4 \\ 1/4 \end{pmatrix}$.

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

KKT system $\begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ -\lambda \end{pmatrix} = \begin{pmatrix} -2b \\ 0 \\ 0 \end{pmatrix} \Rightarrow (d_1, d_2, \lambda) = (-b, \frac{b}{2}, 0)$.

$\ker(\nabla^2 f) = r(0,1)^T$. $r \in \mathbb{R}$. $\ker(A) = t(-2,1)^T$. $t \in \mathbb{R}$. $F = (-2,1)^T$

Convergence analysis: comparision to unconstrained cases.

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

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

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

$z_{k+1} = z_k - t(\nabla^2 g(z_k))^{-1} \nabla g(z_k)$. We show by induction $x_k = Fz_k + \tilde{x}$.

let dx_k. dz_k be the descent direction. s.t. $\begin{matrix} x_{k+1} = x_k + t\, d_{x_k} \\ z_{k+1} = z_k + t\, d_{z_k} \end{matrix}$

$\exists dz_k$ iff $\exists (\nabla^2 g(z_k))^{-1}$ iff $\nabla^2 g(z_k) \succ 0$ iff KKT nonsingular $\Rightarrow \exists dx_k$

Note $Ad_{x_k} = 0$ and $Im(F) = \ker(A)$. $\Leftrightarrow \exists u \in \mathbb{R}^{n-m}$. $d_{x_k} = Fu$.

Moreover. $\nabla^2 f(x_k) d_{x_k} - A^T \lambda = -\nabla f(x_k)$. so. $\nabla^2 f(x_k) Fu - A^T \lambda = -\nabla f(x_k)$.

Multiply (left) by F^T on both sides. $F^T\nabla^2 f(x_k) F u + (AF)^T \lambda = -F^T \nabla f(x_k)$.

$\ker(A) = \mathrm{Im}(F) \Rightarrow AF = 0$ (why?). $\forall x \in \mathbb{R}^{n-m}$ $AFx = \vec{0}$.

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

$d_{x_k} = Fu = F d_{z_k} \Rightarrow x_{k+1} = x_k + t d_{x_k} = F z_k + \tilde{x} + t F d_{z_k} = F z_{k+1} + \tilde{x}$ □

Recall the convergence result of unconstrained backtracking Newton:

without backtracking: if f is m-strongly convex, $\nabla^2 f$ is M-Lipschitz.

then $\|x_{k+1} - x^*\| \le \frac{M}{2m} \|x_k - x^*\|^2$. $\quad$ ∇f is M-smooth.

with backtracking: further assume f is L-smooth. then global convergence.

let $\eta = \min\{1, 3(1-2\alpha)\} m^2/M$. $\gamma = 2\alpha(1-\alpha)\beta\eta^2 m/L$.

if $\|\nabla f(x_k)\| \le \eta$. $f(x_k + d_k) \le f(x_k) + \alpha \nabla f(x_k)^T d$. $t_k = 1$

then $\|x_{k+1} - x^*\| \le \frac{M}{2m} \|x_k - x^*\|^2$. $f(x_k) - f(x^*) \le \frac{m}{2} \|x_k - x^*\|^2$

if $\|\nabla f(x_k)\| > \eta$. backtracking guarantees $f(x_k) - f(x_{k+1}) \ge \gamma$.

overall. let k_0 be the smallest k that $\|\nabla f(x_k)\| \le \eta$.

$$f(x_k) - f(x^*) \le \begin{cases} f(x_0) - f(x^*) - k\gamma & k \le k_0 \\ \frac{2m^3}{M^2} (1/2)^{2^{k-k_0+1}} & k > k_0 \end{cases}$$

To get $f(x_k) - f(x^*) \le \varepsilon$. it suffices to run

$$\frac{1}{\gamma}(f(x_0) - f(x^*)) + \log\log \frac{2m^3}{M^2 \varepsilon} \text{ steps.}$$

Inequality constrained optimization: $\min f(x)$. s.t. $g(x) = 0$ m
$h(x) \le 0$. k.

Again, our first task is to determine whether a given x^* is optimal.

If $h_j(x^*) = 0$, the j-th inequality constraint $h_j(x) \le 0$ is called active,

otherwise called inactive. ($h_j(x^*) < 0$). Denote $J(x) = \{j : h_j(x) = 0\}$.

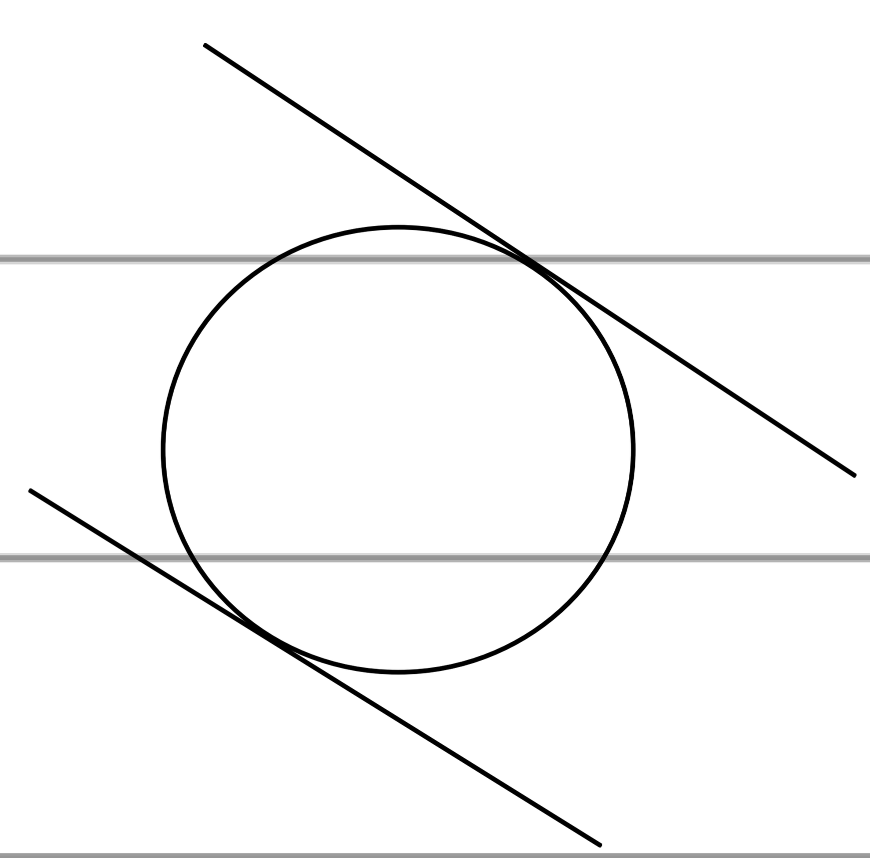
If x^* is a local minimum of original problem, it is also a local

minimum of $\min f(x)$. s.t. $g_i(x) = 0$, $\forall i$; $h_j(x) = 0$, $\forall j \in J(x^*)$.

Apply Lagrange condition. $\exists \lambda^*, \mu^*$. s.t. $\nabla f + \sum_{i=1}^{m} \lambda_i^* \nabla g_i + \sum_{j \in J} \mu_j^* \nabla h_j = 0$

Then we set $\mu_j^* = 0$ for inactive $j \notin J(x^*)$. $\nabla f + \sum_{i=1}^{m} \lambda_i^* \nabla g_i + \sum_{j=1}^{k} \mu_j^* \nabla h_j = 0$.

Is this enough? Consider the example. $\min x_1 + x_2$. s.t. $x_1^2 + x_2^2 \le 2$.



$(1, 1)^T + 2\lambda (x_1, x_2)^T = 0$. $x_1^2 + x_2^2 = 2 \Rightarrow x_1 = x_2 = 1$.
or $\lambda = 0$

If the constraint is equality, then we cannot

distinguish these two cases by first-order condition.

But now …… $-\nabla h_j(x^*)$ is the direction to the interior of the

feasible set, but $-\nabla f(x^*)$ should be the direction to the outside.

$\Rightarrow \mu_j^* > 0$ iff $j \in J(x^*)$. (why this is also true for >1 constraints?).

Note that $j \in J(x^*)$ iff $h_j(x^*) = 0$. o.w $\mu_j^* = 0$. so. $\mu_j^* h_j(x^*) = 0$.

Now we have the Karush-Kuhn-Tucker (KKT) condition.

Theorem. If x^* is a local minimum point of ICP and also a regular point for active constraints. then $\exists$ Lagrange / KKT multipliers. $\lambda_1^*, \dots, \lambda_m^*, \mu_1^*, \dots, \mu_k^*$ such that the following KKT holds.

1. $\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{k} \mu_j^* \nabla h_j(x^*) = 0$.

2. $\mu_j^* \geq 0$. $\forall j = 1, \dots, k$. 3. $\mu_j^* h_j(x^*) = 0$. $\forall j = 1, \dots k$.

Comparison to LP. $\min c^T x$. s.t. $A_1 x = b_1$. $A_2 x \leq b_2$.

KKT condition: $\exists \lambda^*. \mu^*$. s.t. $c + A_1^T \lambda^* + A_2^T \mu^* = 0$.

$\mu_j^* \geq 0$. and $(A_2 x^* - b)^T \mu^* = 0$. $(\mu_j^* \cdot (a_{2j}^T x - b_j) = 0)$.

Duality of LP. $\min (y_1, y_2) \cdot (b_1, b_2)^T$. s.t. $y_1^T A_1 + y_2^T A_2 = -c^T$. $y_2 \geq 0$.

complementary slackness for LP: $(x^*. y^*)$ is optimal for primal and dual res. iff. $y^T(b - Ax) = 0$ and $x^T(A^T y - c) = 0$.

either $y_i = 0$ or $(Ax)_i = b_i$ $y_2^T (A_2 x^* - b) = 0$.

So condition 3 is called complementary slackness for KKT.

KKT is named after Harold W. Kuhn and Albert W. Tucker (1951, published). but later found in master thesis of William Karush (1939) unpublished.