Backtracking line search: get a good enough step. then stop.

$x \leftarrow x_0 \in \mathbb{R}^n$

while $\|\nabla f(x)\| < \delta$ do

  $t \leftarrow t_0$

  while $f(x - t\nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$ do ← Armijo's Rule

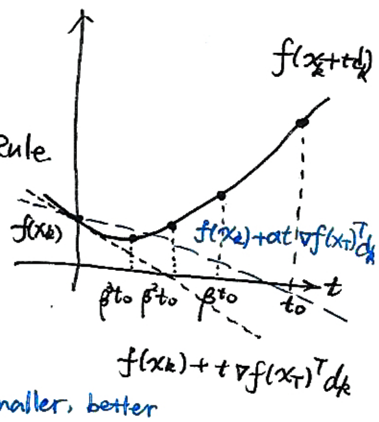
    $t \leftarrow \beta t$

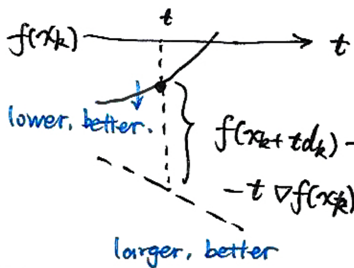
  end while

  $x \leftarrow x - t\nabla f(x)$

end while

return x





$f(x_k+td_k) - f(x_k)$ ← smaller, better

$-t\nabla f(x_k)^T d_k \leq \square \cdot (f(x_k) + t\nabla f(x_k)^T d_k)$

$\to$ $f(x_k) - f(x_k + td_k) \geq \alpha \cdot t[\nabla f(x_k)^T d_k]$

the decrease in function value $\geq$ decrease along tangent line

↓

the decrease is large enough.

Thm. backtracking line search. $f(x_k) - f(x^*) \leq c^k [f(x_0) - f(x^*)]$

where $c = 1 - \min\left\{ 2m\alpha t_0, \frac{4m\beta\alpha(1-\alpha)}{L} \right\} \in (0,1)$

$\underset{\frac{\beta m}{L} < \beta < 1.}{\wedge}$ $\quad \geq 0$

<u>Proof:</u> L-smoothness. $f(x_k - t\nabla f(x_k)) \leq f(x_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x_k)\|^2$

The inner loop terminates for sure if

~~$f(x_k - t\nabla f(x_k)) - f(x_k)$~~ $-t\left(1-\frac{Lt}{2}\right)\|\nabla f(x_k)\|^2 \leq -\alpha t\|\nabla f(x_k)\|^2$.

$\Rightarrow t \leq \frac{2(1-\alpha)}{L}$

$t_k \geq t_0$, $t_k \geq \frac{2(1-\alpha)}{L}$ $\Rightarrow$ $t_k \geq \eta := \max\left\{ t_0, \frac{2\beta(1-\alpha)}{L} \right\}$

$t_{k-1} \leq \frac{\cdot}{\beta}$

Nesterov's Accelerated Gradient Descent (AGD)

Suppose f is L-smooth and m-strongly convex. $(m \geq 0)$

- $x_0 \leftarrow y_0$
- for $k = 0, 1, 2, \cdots$ do

  $x_{k+1} \leftarrow y_k - \frac{1}{L}\nabla f(y_k)$

  $y_{k+1} \leftarrow x_{k+1} + \beta_k(x_{k+1} - x_k)$

  end for

G.S. $(m=0)$. $\quad f(x_k)-f(x^*)=\Omega(\frac{1}{k}). \longrightarrow k=\Omega(\frac{1}{\varepsilon})$.

$(m>0)$ $\quad f(x_k)-f(x^*)=\Omega\left((1-\frac{m}{L})^k\right) \rightarrow k=\Omega(\frac{L}{m}\log\frac{1}{\varepsilon})$.

AGD. $(m=0)$. $\quad f \quad \Omega(\frac{1}{k^2}) \longrightarrow k=\Omega(\frac{1}{\sqrt{\varepsilon}})$

$(m>0)$ $\quad \Omega\left((1-\sqrt{\frac{m}{L}})^k\right) \rightarrow k=\Omega(\sqrt{\frac{L}{m}}\log\frac{1}{\varepsilon})$.

Gradient Descent Applied to Nonconvex Function

Global optimum not guaranteed. even local minimum not guaranteed.

$\hookrightarrow$ find a stationary point.

$$\min\|\nabla f(x_i)\| \le \sqrt{\frac{2(f(x_0)-f^*)}{t(k+1)}}.$$

Better Gradient Descent?

• Newton's Method

Normal Gradient Descent $\rightarrow$ $f(x)\approx \hat{f}_{gd}(x) := f(x_k)+\nabla f(x_k)^T(x-x_k)+\frac{1}{2t_k}\|x-x_k\|^2$

$\downarrow$ Optimization

$\nabla \hat{f}_{gd}(x)=0 \quad \Rightarrow \quad \nabla f(x_k)+\frac{1}{t_k}(x-x_k)=0$.

$\Rightarrow \quad x=x_k-t_k\nabla f(x_k)$.

Newton's method: $f(x)\approx \hat{f}_{nt}(x) := f(x_k)+\nabla f(x_k)^T(x-x_k)+\frac{1}{2}(x-x_k)^T(\nabla^2 f(x_k))(x-x_k)$.

$\downarrow$ Optimization

[2^{nd} order Taylor Expansion]

$\nabla\hat{f}_{nt}(x)=0 \quad \Rightarrow \quad \nabla f(x_k)+\nabla^2 f(x_k)(x-x_k)=0$

$\underline{\nabla^2 f\succ 0}$ $\quad \Rightarrow \quad x=x_k-[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$.

- $x\leftarrow x_0\in\mathbb{R}^n$
- while $\|\nabla f(x)\|>\delta$ do

  $x\leftarrow x-[\nabla^2 f(x)]^{-1}\nabla f(x)$

  end while.

Affine Invariance : invertible A. $\quad g(y)=f(Ay)$. $\quad \nabla g(y)=A^T\nabla f(Ay)$.

$\nabla^2 g(y)=A^T\nabla^2 f(Ay)A$.

$x_0=Ay_0$. Run Newton's Method. $\quad y_1=y_0-[\nabla^2 g(y_0)]^{-1}\nabla g(y_0)=\cdots=y_0-A^{-1}[\nabla^2 f(x_0)]^{-1}\nabla f(x_0)$

$=A^{-1}\left(x_0-(\nabla^2 f(x_0))^{-1}\nabla f(x_0)\right)=A^{-1}x_1$.

$x_1=Ay_1$, $\quad \cdots\cdots \quad \Rightarrow x_n=Ay_n$.

(In GD: not hold) 

Scaling change the "direction"

Newton's method:  ✓.

Connection to Root Finding: 牛顿逼近法 (the original Newton's Method)

$g(x) \approx g(x_k) + g'(x_k)(x - x_k) \Rightarrow x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$. vs Newton's Method $x_{k+1} = x_k - (\nabla^2 f)^{-1}_{x=x_k} \nabla f(x_k)$

Optimization $\Leftrightarrow$ find a root of: $f'(x^*) = 0$

Convergence not guaranteed. If converge. the rate: very fast.

Convergence Analysis

Thm. f: m-strongly convex. f'': L-Lipschitz continuous x^*: a global minimum

$\{x_k\}$ produced by Newton's Method satisfies

$$|x_{k+1} - x^*| \leq \frac{L}{2m} |x_k - x^*|^2$$ ← Local Convergence Result.

(Let $\xi_k = \frac{L}{2m} |x_k - x^*|$. To prove: $\xi_{k+1} \leq \xi_k^2$). $\underline{\xi_0 < 1}$ → converge very fast

Note. $\xi_k \leq \cdots \leq (\xi_0)^{2^k} = \varepsilon$. $\Rightarrow 2^k \log \frac{1}{\xi_0} = \log \frac{1}{\varepsilon}$. $k = \Omega\left(\log_2 \log_{\frac{1}{\xi_0}} \frac{1}{\varepsilon}\right)$ ↖ very fast!

Proof. $|x_{k+1} - x^*| = \left| x_k - \frac{f'(x_k)}{f''(x_k)} - x^* \right| = \left| \frac{1}{f''(x_k)} \left(f''(x_k)(x_k - x^*) - f'(x_k) \right) \right|$

$$= \left| \frac{1}{f''(x_k)} \left(f'(x^*) - f'(x_k) + f''(x_k)(x_k - x^*) \right) \right|$$

$$= \left| \frac{1}{f''(x_k)} \left(f''(\eta_k)(x^* - x_k) - f''(x_k)(x^* - x_k) \right) \right|$$

$$= \frac{1}{|f''(x_k)|} \left| f''(\eta_k) - f''(x_k) \right| |x^* - x_k|$$

$$\leq \frac{1}{|f''(x_k)|} \cdot L \underbrace{|\eta_k - x_k|}_{\leq |x^* - x_k|} |x^* - x_k| \leq \frac{L/2}{|f''(x_k)|} |x^* - x_k|^2 \leq \frac{L}{2m} |x^* - x_k|^2.$$

Def. Matrix Norm

$A \in \mathbb{R}^{m \times n} \longrightarrow \vec{v} = \begin{pmatrix} \vdots \end{pmatrix} \in \mathbb{R}^{mn}$ a vector. $\|A\| := \|v\|$.

Property of vector norm still holds.

Operator Norm. (Induced Norm)

$A: \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto Ax$ (↓ a matrix)

$$\|A\|_{a,b} = \max_{x: x \neq 0} \frac{\|Ax\|_b}{\|x\|_a} = \max_{x: \|x\|_a = 1} \|Ax\|_b = \max_{x: \|x\|_a \leq 1} \|Ax\|_b.$$

$$\|Ax\|_b \leq \|A\|_{a,b} \|x\|_a$$

$\Rightarrow$ Lipschitz: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L \|x - y\|$.

Spectral Norm. : Induced Norm when $a=b=2$.

$$\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}$$

Damped Newton's Method

- $x \leftarrow x_0$
- while $\|\nabla f(x)\| > \delta$ do
  - $d \leftarrow -[\nabla^2 f(x)]^{-1} \nabla f(x)$
  - $t \leftarrow 1$
  - while $f(x+td) > f(x) + \alpha t \nabla f(x)^T d$ do
    - $t \leftarrow \beta t$
  - end while
  - $x \leftarrow x + td$
- end while

Backtracking line $\leftarrow$ Converge globally

to ensure $f(x_k) > f(x_{k+1})$

Thm. f: m-strongly convex. L-smooth.
∇f is M-Lipschitz. x^*: minimum of f.
Damped N.M. satisfies

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

k_0: #(steps until $\|\nabla f(x_{k_0+1})\| \leqslant \eta$. i.e. $k_0 \leqslant \left(\frac{f(x_0)-f(x^*)}{\gamma}\right)$

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

$f(x_k) - f(x^*) \leqslant \varepsilon$.

Need: $\frac{f(x_0)-f(x^*)}{\gamma} + \log_2 \log_2 \frac{\varepsilon_0}{\varepsilon}$. $\quad \varepsilon_0 = \frac{2m^3}{M^2}$

- # Proximal Gradient Descent

G.D: once differentiable. N.M: twice differentiable. what if f is not differentiable?

$$F(x) = \underset{\substack{\text{convex}\\\text{smooth}}}{f(x)} + \underset{\substack{\text{convex}\\\text{not necessarily smooth.}}}{h(x)}$$

e.g. LASSO

$$\min_w \frac{1}{2}\|Xw-y\|_2 \quad \text{s.t. } \|w\|_1 \leqslant t \iff \min_w \frac{1}{2}\|Xw-y\|_2^2 + \lambda\|w\|_1$$

$\lambda = f(t)$.

GD: $x_{k+1} = x_k - t_k \nabla F(x_k) \Rightarrow \nabla F(x_k) + \frac{1}{t_k}(x - x_k) = 0$ $\quad x_{k+1}$

$$x_{k+1} = \underset{x}{\operatorname{argmin}}\Big[\underbrace{\overset{\text{const}}{\cancel{F(x_k)}} + \nabla F(x_k)^T(x-x_k) + \frac{1}{2t_k}\|x-x_k\|_2^2}_{=:\hat{F}(x)}\Big]$$

GD用该方式近似原函数找x_{k+1}！（上上页）

$\|(x-x_k) - t_k\nabla F(x_k)\|^2$

$$= \underset{x}{\operatorname{argmin}} \frac{1}{2t_k}\|x - (x_k - t_k\nabla F(x_k))\|_2^2$$

反正能证. ✓

$\boxed{(\text{In fact, } x_{k+1} = x_k - t_k \nabla F(x_k))}$

Prox:

$$x_{k+1} = \underset{x}{\operatorname{argmin}}\Big[f(x_k) + \nabla f(x_k)^T(x-x_k) + \frac{1}{2t_k}\|x-x_k\|_2^2 + h(x)\Big]$$

$$= \underset{x}{\operatorname{argmin}}\Big(\frac{1}{2}\|x - \underbrace{(x_k - t_k\nabla f(x_k))}_{=z}\|_2^2 + t_k h(x)\|\Big)$$

$\tilde{h}(z) := t_k . h(z)$

"Proximal": x_{k+1} stay close to $x_k - t_k \nabla f(x_k)$

$$\operatorname{prox}_h(x) = \underset{z}{\operatorname{argmin}}\left(\frac{1}{2}\|x-z\|^2 + \tilde{h}(z)\right) \text{ ; i.e. } x_{k+1} \leftarrow \operatorname{prox}_{t_k h}(x_k - t_k\nabla f(x_k))$$

- $x \leftarrow x_0$
- while (stopping criterion == FALSE) do

  $x \leftarrow \mathrm{prox}_{th}(x - t\nabla f(x))$

  end while

$$\mathrm{prox}_h(x) = \underset{z}{\mathrm{argmin}}\left(\frac{1}{2}\|x-z\|_2^2 + h(z)\right)$$

e.g. $h(z) = \frac{\lambda}{2}\|z\|_2^2$ (actually is differentiable) $\Rightarrow \mathrm{prox}_h(x) = \underset{z}{\mathrm{argmin}}\left(\frac{1}{2}\|z-x\|_2^2 + \frac{\lambda}{2}\|z\|_2^2\right) = \frac{x}{1+\lambda}$

$\boxed{l_2 \text{ regularization}}$

$$x_{k+1} = \frac{x_k - t\nabla f(x_k)}{1+\lambda t} = x_k - \frac{t}{1+\lambda t}\left(\nabla f(x_k) + \lambda x_k\right)$$

Note: prox_{th}, $\lambda \to \lambda t$ — exactly GD with step size $t^* = \frac{t}{1+\lambda t}$

l_1 regularization: $h(z) = \lambda|z|$. $(\lambda \geq 0)$.

1D case: $\min_z \frac{1}{2}(z-x)^2 + \lambda|z|$.

$\underline{z^* x \geq 0}$ (minimum). Proof: $\frac{1}{2}(z^*-x)^2 + \cancel{\lambda|z^*|} \leq \frac{1}{2}(-z^*-x)^2 + \cancel{\lambda|-z^*|}$

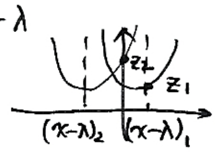
$\Rightarrow -z^*x \leq z^*x \Rightarrow z^*x \geq 0$ □

① $x \geq 0$: $\min_{z\geq 0} \frac{1}{2}(z-x)^2 + \lambda|z| = \min_{z\geq 0} \frac{1}{2}(x-z)^2 + \lambda z$

$z^* x \geq 0$

$\nabla_0 = 0$. $z - x + \lambda = 0 \Rightarrow z = x - \lambda$

$= \max\{0, x-\lambda\}$.



② $x \leq 0$. $\min_{z\leq 0} \frac{1}{2}(z-x)^2 + \lambda|z| = \min\{0, x+\lambda\}$.

Thus, $\mathrm{prox}_{\lambda|\cdot|}(x) = z = S_\lambda(x) = \mathrm{sgn}(x)\left(|x|-\lambda\right)^+ = \begin{cases} x-\lambda & \text{if } x \geq \lambda \\ 0 & \text{if } -\lambda < x < \lambda \\ x+\lambda & \text{if } x \leq -\lambda. \end{cases}$

Soft-thresholding operator

nD case: $\frac{1}{2}\|\vec{z} - \vec{x}\|_2^2 + \lambda\|\vec{z}\|_1 = \sum_{i=1}^n \left(\frac{1}{2}(z_i - x_i)^2 + \lambda|z_i|\right)$

$\therefore \left[\mathrm{prox}_{\lambda|\cdot|}(\vec{x})\right]_i = S_\lambda(x_i)$

$\Rightarrow$ $w_{k+1} = S_{\lambda t}\left(w_k - tX^T(Xw_k - y)\right)$ —— $\underline{\text{ISTA}}$ (Iterative Soft-Thresholding Algorithm)

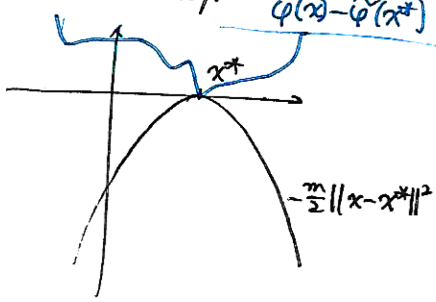
Note: prox_{th}

Lemma. φ: m-strongly convex with a maximum x^*,

$$\varphi(x) \leq \varphi(x^*) + \frac{m}{2}\|x - x^*\|_2^2. \quad \forall x$$

Note. φ is differentiable $\to$ $\varphi(x) \geq \varphi(x^*) + \underbrace{\nabla\varphi(x^*)}_{=0}{}^T(x-x^*) + \frac{m}{2}\|x-x^*\|^2$. easy.

but not guaranteed!



Proof. $\varphi(x) \geq \varphi(x^*)$. $\quad \varphi(x) =: \underbrace{\tilde{\varphi}(x)}_{\text{convex!}} + \frac{m}{2}\|x-x^*\|^2$

$\Downarrow$

$\tilde{\varphi}(x) + \frac{m}{2}\|x-x^*\|^2 \geq \tilde{\varphi}(x^*) \Rightarrow \tilde{\varphi}(x) - \tilde{\varphi}(x^*) \geq -\frac{m}{2}\|x-x^*\|^2$

$\therefore \tilde{\varphi}(x) - \tilde{\varphi}(x^*) \geq 0$ //

(By Separating Hyperplane Thm. $\tilde{\varphi}(x) - \tilde{\varphi}(x^*) \geq 0$) $\longrightarrow$ Proof. TBD.

Thus, $\varphi(x) \geq \tilde{\varphi}(x^*) + \frac{m}{2}\|x - x^*\|_2^2 = \varphi(x^*) + \frac{m}{2}\|x - x^*\|_2^2$

Convergence Analysis.

Let $t \in (0, \frac{1}{L}]$ ($F = f + h$, f: L-Smooth, m-strongly convex)

$F(x_{k+1}) \leq F(x_k)$. $\|x_{k+1} - x^*\|_2 \leq \|x_k - x^*\|_2$

$F(x_k) - F(x^*) \leq \frac{L}{2k}\|x^* - x_0\|_2^2$. if $m > 0$: $\|x^* - x_k\|_2^2 \leq (1 - \frac{m}{L})^k \|x^* - x_0\|_2^2$.

Proof. $\hat{F}(x) = F(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{L}{2}\|x - x_k\|_2^2 + h(x)$ $\qquad$ ($t_k = L^{-1}$)

Smoothness and Strong Convexity: $\frac{m}{2}\|x - x_k\|_2^2 \leq f(x) - f(x_k) - \nabla f(x_k)^T(x - x_k) \leq \frac{L}{2}\|x - x_k\|_2^2$

Plugging into $\hat{F}(x)$: $\qquad F(x) \leq \hat{F}(x) \leq F(x) + \frac{L-m}{2}\|x - x_k\|_2^2$

Note $\hat{F}(x)$ is L-strongly convex. $x_{k+1} = \operatorname{argmin} \hat{F}(x)$

$$F(x_{k+1}) \leq \hat{F}(x_{k+1}) \overset{\text{Lemma}}{\leq} \hat{F}(x) - \frac{L}{2}\|x - x_{k+1}\|_2^2 \leq F(x) + \frac{L-m}{2}\|x - x_k\|_2^2 - \frac{L}{2}\|x - x_{k+1}\|_2^2.$$

$\square$

Convex Problem with Equality Constraints

$$\min_x f(x) \quad \text{s.t.} \quad \underline{a_i^T x = b} \nearrow \underline{x \in X}$$

i.e. $Ax = b$.

i.e. $Ax=b$: linearly independent

Assume $\underline{\operatorname{rank} A = k}$. (full rank)

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

$\mathrm{Null}(A) := \{x : Ax = 0\}$ is the *null space of* A.

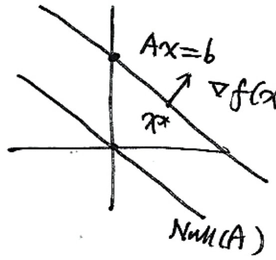
<u>Proof.</u> x^* opt $\Leftrightarrow \nabla f(x^*)^T (x - x^*) \geq 0.\ \forall x \in X$.

$Ax = Ax^* = b \Rightarrow A(x - x^*) = 0$. i.e. $x := x^* + \gamma$. $\gamma \in \mathrm{Null}(A)$.

$A\gamma = 0$. $A(-\gamma) = 0 \quad \therefore -\gamma \in \mathrm{Null}(A)$

$$\left.\begin{array}{l} \nabla f(x^*)^T \gamma \geq 0 \\ \nabla f(x^*)^T (-\gamma) \geq 0 \end{array}\right\} \Rightarrow \nabla f(x^*)^T \gamma = 0. \quad \text{i.e. } \nabla f(x^*)^T (x - x^*) = 0.$$

i.e. $\nabla f(x^*) \perp \mathrm{Null}(A)$



□

<u>Second Proof.</u> $Ax = b$. $A \in \mathbb{R}^{k \times n}$. $\Rightarrow x = x^* + \gamma$. $\gamma = \sum_{i=1}^{n-k} y_i \cdot z_i = F z$.

$\uparrow$ base vec $\quad (F := (y_1, \dots y_{n-k}))$

$g(z) := f(x) = f(x^* + Fz)$.

$\nabla g(0) = \nabla f(x^*) = 0. \Rightarrow \nabla g(0)^T z = 0. \Rightarrow \nabla f(x^*)^T F \cdot z = 0 \Rightarrow \nabla f(x^*)^T (x - x^*) = 0$

$\Rightarrow \nabla f(x^*) \perp \mathrm{Null}(A)$ □

Thm. $x^* \in X$ is optimal iff exists $\lambda^* = (\lambda_1^*, \dots \lambda_k^*)^T \in \mathbb{R}^k$ s.t.

$$\nabla f(x^*) + A^T \lambda^* = 0.$$

$\lambda_1^*, \dots \lambda_k^*$ are called *Lagrange multipliers*.

Lemma. $\mathrm{Null}(A)^\perp = \mathrm{Range}(A^T) := \{A^T v : v \in \mathbb{R}^k\}$.

$\uparrow$ orthogonal complement of $\mathrm{Null}(A)$. i.e. $x \in \mathrm{Null}(A)^\perp \Leftrightarrow x \perp y.\ \forall y \in \mathrm{Null}(A)$

<u>Proof.</u> $\mathrm{Null}(A) = \{x : a_1^T x = 0,\ a_2^T x = 0,\ \dots,\ a_k^T x = 0\}$. $\quad A = \begin{pmatrix} a_1^T \\ \vdots \\ a_k^T \end{pmatrix}$

$A^T = (a_1 \cdots a_k)$

1) $x \in \mathrm{Range}(A^T) \Rightarrow \exists z$ s.t. $A^T z = x$. $\quad z = (\gamma_1 a_1 + \gamma_2 a_2 + \cdots + \gamma_k a_k)$

$\forall y \in \mathrm{Null}(A)$. $\quad x^T y = \gamma_1 \underbrace{a_1^T y}_{0} + \gamma_2 \underbrace{a_2^T y}_{0} + \dots + \gamma_k \underbrace{a_k^T y}_{0} = 0$. i.e. $x \perp y$. $\therefore x \in \mathrm{Null}(A)^\perp$

2) $\dim \mathrm{Range}(A^T) = n - k = n - \dim \mathrm{Null}(A)$.

$\therefore \mathrm{Null}(A)^\perp = \mathrm{Range}(A^T)$.

□

<u>Thm Proof.</u> $\nabla f(x^*) \perp \mathrm{Null}(A) \Rightarrow \nabla f(x^*) \in \mathrm{Null}(A)^\perp = \mathrm{Range}(A^T)$

$= \{x : x = A^T v.\ v \in \mathbb{R}^k\}$

$\therefore \nabla f(x^*) = A^T v = A^T(-\lambda^*)$.

$\Rightarrow \nabla f(x^*) + A^T \lambda^* = 0.$ □

- Lagrange Condition

Def. Lagrangian (Lagrange function) $\mathcal{L}(x,\lambda) = f(x) + \lambda^T (Ax - b)$

$$= f(x) + \sum_{i=1}^{k} \lambda_i (a_i^T x - b_i).$$

~~The optimality of the optimality of original convex problem~~

The optimality condition, becomes KKT equations

$$\begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + A^T \lambda^* = 0 \\ \nabla_\lambda \mathcal{L}(x^*, \lambda^*) = Ax^* - b = 0. \end{cases} \quad (\text{i.e. } \nabla \mathcal{L}(x^*, \lambda^*) = 0.)$$

i.e. (x^*, λ^*) is a stationary point of $\mathcal{L}$

Note. $\nabla_\alpha \mathcal{L} : \left(\frac{\partial \mathcal{L}}{\partial \alpha_1}, \dots \frac{\partial \mathcal{L}}{\partial \alpha_m}\right).$

- General Equality Constrained Problems

e.g. 2D case

$$\min_x f(x) \quad \text{s.t. } h(x) = \|x\|^2 - 1 = 0 \quad \Rightarrow \quad \min_t g(t) := f(x(t)) = f(\cos t, \sin t)$$

opt t^* $\Leftrightarrow \nabla g(t^*) = 0 \Leftrightarrow \nabla f(x(t^*)) = 0 \Leftrightarrow \frac{\partial f}{\partial x_1} \cdot x_1'(t) + \frac{\partial f}{\partial x_2} \cdot x_2'(t) = 0$

$h(x^*) = 0 \Rightarrow h(x(t^*)) = 0 \Rightarrow h' = 0. \quad \frac{\partial h}{\partial x_1} \cdot x_1'(t) + \frac{\partial h}{\partial x_2} \cdot x_2'(t) = 0$

$$\Rightarrow \begin{pmatrix} \nabla f(x^*)^T \\ \nabla h(x^*)^T \end{pmatrix} x'(t^*) = 0 \qquad x'(t^*) \neq 0 \Rightarrow \nabla f(x^*) \text{ and } \nabla h(x^*) : \text{linearly dependent}$$

Meanwhile, $h(x) = \|x\|^2 - 1 = 0. \quad x \neq 0. \quad \nabla h(x^*) = 2x^* \neq 0.$

Thus, $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0$

$\Rightarrow$ **Def.** Lagrangian $\mathcal{L}(x,\lambda) = f(x) + \lambda h(x)$

x^* is a local optimum only if exists λ^* s.t. $\begin{cases} \nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \lambda^* \nabla h(x^*) = 0 \\ \nabla_\lambda \mathcal{L}(x^*, \lambda^*) = h(x^*) = 0 \end{cases}$

At all extrema: $\nabla f \parallel \nabla h, \quad \nabla f \perp X$

Thm. Implicit Function Theorem

$F(x,y)$: continuously differentiable in a neighborhood of (x_0, y_0).

$F(x_0, y_0) = 0. \quad \frac{\partial F}{\partial y}(x_0, y_0) \neq 0 \quad \Rightarrow \exists$ continuously differentiable φ. $y = \varphi(x)$

$$F(x, \varphi(x)) = 0. \quad \varphi'(x) = -\left[\frac{\partial F(x, \varphi(x))}{\partial y}\right]^{-1} \frac{\partial F(x, \varphi(x))}{\partial x}$$

Thm. $\nabla h(x) \neq 0$: x called regular point of h; otherwise: critical point $\leftarrow$ 临界点是critical point.

If x^* is a local extremum of f s.t. $h(x) = 0$. x^* is a regular point of h.
正则点

then exists λ^* s.t. $\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0.$

$\leftarrow$ only this one satisfied: could be neither max…/min…

n-D case:

$$\min_x f(x) \quad \text{s.t.} \quad h_i(x)=0,\ i=1,2,\cdots,k$$

$\nabla h_1(x), \cdots \nabla h_k(x)$ are linearly independent $\to$ regular point; otherwise critical point

Thm. x^*: local extremum of f s.t. $h(x)=0$

$$\nabla f(x^*) + (\lambda^*)^T \nabla h(x^*) = 0$$

Def. Lagrangian function $\mathcal{L}(x,\lambda) = f(x) + \lambda^T h(x)$.

Def. The feasible set $X := \{x : h(x)=0\}$ is a $(n-k)$-dimensional manifold.

Def. A tangent vector of X at $x_0 \in X$: velocity vector $v = x'(t)$ of curve $x(t) \subset X$. $x(0)=x_0$.

By Chain Rule, $Dh(x)v=0$ (导数) i.e. $\nabla h_i(x^*)^T v = 0$. $(i \in \{1,2,\cdots k\})$

The tangent space $T_{x_0}X = \{v \in \mathbb{R}^n : Dh(x)v = 0\}$. (all tangent vectors)

normal space $N_{x_0}X = \text{span}\{\nabla h_1(x_0), \cdots \nabla h_k(x_0)\}$. $\leftarrow$ range space

$$N_{x_0}X = [T_{x_0}X]^{\perp}$$

- KKT Conditions

$$\min_x f(x) \quad \text{s.t.} \quad \begin{array}{ll} g_i(x) \le 0 & i=1,2,\cdots,m \\ h_j(x)=0 & j=1,2,\cdots,m \end{array} \qquad (\text{ICP})$$

convexity uncertain. (not required)

$g_j(x) \le 0$ is active at x_0 if $g_j(x)=0$. otherwise inactive. $J(x_0) = \{j : g_j(x_0)=0\}$

Convention: equality constraints are considered active.

x^*: solution.

$$(\text{ICP}) \Leftrightarrow \min_x f(x) \ \text{s.t.}\ \begin{array}{l} g_i(x)\le 0 \\ h_j(x)=0 \\ x\in B(x^*,\delta) \end{array} \Leftrightarrow \min_x f(x) \ \text{s.t.}\ \begin{array}{l} h_j(x)=0 \\ g_i(x)=0,\ j\in J(x^*) \\ x\in B(x^*,\delta) \end{array}$$

$\leftarrow$ keep all active constraints

If we know who are active at x^*. ICP $\to$ ECP.

At a regular local minimum, we have Lagrangian condition: $\nabla f(x^*) + \sum \lambda_i^* \nabla h_i(x^*) + \sum \mu_j^* \nabla g_j(x^*) = 0$

(We can let $\mu_j^*=0$ if inactive)

KKT Conditions: ① $\nabla f(x^*) + \sum \lambda_i^* \nabla h_i(x^*) + \sum \mu_j^* \nabla g_j(x^*)$

② $\mu_j^* \ge 0$

③ $\mu_j^* g_j(x^*) = 0$. $\leftarrow$ for all inequality constraints (complementary slackness)

another form: ③′ $\sum_j \mu_j g_j(x) = 0$

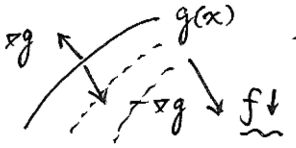
②, ③′. $g_j(x) \le 0 \Rightarrow \mu_j = 0$ or $g_j(x)=0$

Note: $\nabla f + \sum_i \mu_i \nabla g_i = 0 \Rightarrow \nabla f = \sum_{(\text{active})} \mu_j \nabla g_j$ $\leftarrow$ conic combination

(系数非负的线性组合)

Proof that " $\mu_i \geq 0$ " $(\forall i)$

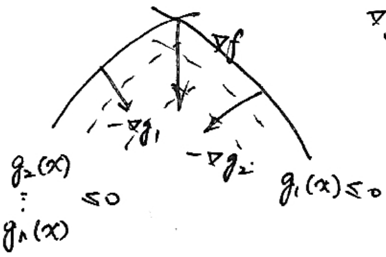
① 1-D case.



$\nabla f = -\mu \nabla g$

Contradict to $f(x^*)_{min}$

② n-D case



$\nabla f = -\mu \nabla g_{other} - \mu \nabla g_1$

To find a direction where f decreases (let it be d)

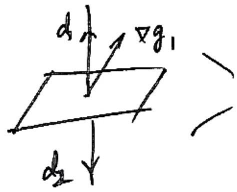
$d^T \nabla f < 0$

$d^T \nabla f = \underbrace{-\mu d^T \nabla g_{other}}_{\text{let it be } 0} - \mu d^T \nabla g_1$. (∇g_1, ∇g_{other} : linearly independent assumption)

contradict!

$d^T \nabla g_1 \neq 0$ otherwise. $\nabla g_1 \in \text{span}\{\nabla g_{other}\}$

d. (perpendicular to span$\{\nabla g_{other}\}$)



at least one is smaller than 0.

d. Found.

Contradict to $f(x^*)_{min}$.

Note: KKT Condition can't be solved easily in most cases.

How to solve KKT Condition?

$$\begin{cases} \mu_i g_i = 0 \\ \mu_i \geq 0 \quad (\forall i \in \{1, \cdots, m\}) \\ \nabla f(x^*) + \sum \lambda_i^* \nabla h_i(x^*) + \sum \mu_j^* \nabla g_j(x^*) = 0 \end{cases}$$

Discuss whether g_i is inactive or not. ($\mu_i = 0$ or not)

$\rightarrow 2^m$ cases.

e.g. Power Allocation : $\max_{P_1, \cdots P_n} \sum_{i=1}^n W_i \log_2 (1 + \frac{P_i}{N_i})$ s.t. $\sum_{i=1}^n P_i \leq P$, $P_i \geq 0$

~~Lagrangian Func.~~

$$\Leftrightarrow \min_P f(P) = -\sum_{i=1}^n W_i \log (1 + \frac{P_i}{N_i}) \quad \text{s.t.} \quad h(P) = \sum_{i=1}^n P_i - P = 0$$

$$g_i(P) = -P_i \leq 0$$

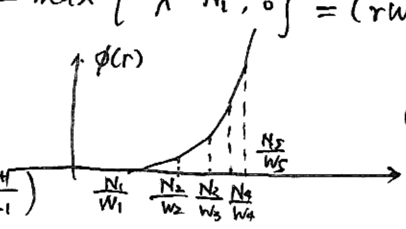
Lagrangian. $\mathcal{L}(P, \lambda, \mu) = -\sum_{i=1}^n W_i \log (1 + \frac{P_i}{N_i}) + \lambda (\sum P_i - P) - \sum_i \mu_i P_i$

$$\partial_{P_i} \mathcal{L} = -\frac{W_i}{P_i + N_i} + \lambda - \mu_i = 0$$

① g_i is inactive. $\mu_i = 0 \Rightarrow -\frac{W_i}{P_i + N_i} + \lambda = 0 \Rightarrow P_i = \frac{W_i}{\lambda} - N_i > 0$

② g_i is active. $P_i = 0$. $\Rightarrow -\frac{W_i}{N_i} + \lambda = \mu_i \geq 0 \Rightarrow \frac{W_i}{\lambda} - N_i \leq 0$.

$$P_i = \begin{cases} \frac{W_i}{\lambda} - N_i , & \frac{W_i}{\lambda} - N_i > 0 \\ 0 , & \frac{W_i}{\lambda} - N_i \leq 0 \end{cases} = \max\{\frac{W_i}{\lambda} - N_i, 0\} = (rW_i - N_i)^+ . \quad (r = \frac{1}{\lambda})$$



$\therefore P = \sum_i (rW_i - N_i)^+$.

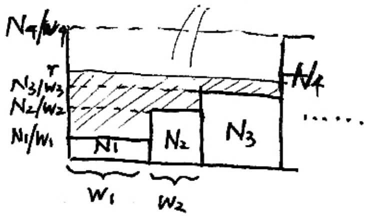
$= \sum_{i=1}^k (rW_i - N_i)$ (if $\frac{N_k}{W_k} \leq r < \frac{N_{k+1}}{W_{k+1}}$)

(Let $\frac{N_1}{W_1} \leq \frac{N_2}{W_2} \leq \cdots \leq \frac{N_n}{W_n}$)

for all k in $\{1, 2, \cdots, n\}$.

$$r = \frac{P + \sum_{i=1}^{k} N_i}{\sum_{i=1}^{k} W_i} .$$ Check whether $\frac{N_k}{W_k} \le r \le \frac{N_{k+1}}{W_{k+1}}$.

→ Water filling problem.



Projected Gradient Descent

- Can we apply gradient descent to constrained problem? $\min_{x \in \mathcal{X}} f(x)$

$\mathcal{X} = \{x : \vec{h}(x) = \vec{0},\ \vec{g}(x) \leq \vec{0}\}$.

Project $x_k - t_k \nabla f(x_k)$ back to the feasible set.

i.e. $$x_{k+1} = \mathcal{P}_{\mathcal{X}}\left(x_k - t_k \nabla f(x_k)\right) = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \left\| x - \left(x_k - t_k \nabla f(x_k)\right) \right\|_2^2$$

Rewrite

Rewrite G.D :

$$x_{k+1} = x_k - t\vec{g}(x_k). \quad \text{where } \vec{g}(x_k) = \frac{1}{t}\left(x - \mathcal{P}_{\mathcal{X}}(x - t\nabla f(x))\right)$$ ← new "gradient"

Stopping criterion:

$$\hat{x} = \mathcal{P}_{\mathcal{X}}(x) \text{ iff. } \langle x - \hat{x}, z - \hat{x} \rangle \leq 0. \quad \forall z \in \mathcal{X}.$$

$$g(x^*) = 0 \iff x^* = \mathcal{P}_{\mathcal{X}}(x^* - t\nabla f(x^*)) \iff \langle x^* - t\nabla f(x^*) - x^*, z - x^* \rangle \leq 0. \quad \forall z \in \mathcal{X}.$$

$$\iff \langle \nabla f(x^*), z - x^* \rangle \geq 0. \quad \forall z \in \mathcal{X}. \xLeftrightarrow[\text{Optimality}]{\text{First-order}} x^* : \text{minimum}.$$

e.g. LASSO. (ℓ_2) $\|y\|_2 \leq t$.

$$\mathcal{P}_{\mathcal{X}}(x) = \begin{cases} x & (\|x\| < t) \\ t\frac{x}{\|x\|} & (\|x\| \geq t) \end{cases}$$

Box: $\{x \in \mathbb{R}^n : a \leq x \leq b\}$

$$\mathcal{P}_{\mathcal{X}}(y) = \min\{b, \max\{a, y\}\}$$

i.e. $$x_i = \begin{cases} a_i & y_i \leq a_i \\ y_i & a_i \leq y_i \leq b_i \\ b_i & y_i \geq b_i \end{cases}$$

$\{x \in \mathbb{R}^n : Ax = b\}$ $$\mathcal{P}_{\mathcal{X}}(y) = y - A^T(AA^T)^{-1}(Ay - b)$$

solve $\min_x \frac{1}{2}\|x - y\|_2^2$ s.t. $Ax = b$.

ℓ_1-ball:

$$\mathcal{P}_{\mathcal{X}}(y) = \begin{cases} y. & \|y\|_1 \leq t \\ \sum(y_i - \mu_0)^+. & \|y\| > t \end{cases}$$

find μ_0

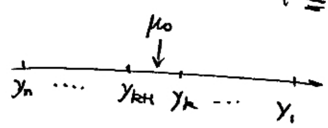
(symmetric → focus on $y > 0$. ⇒ $x > 0$. otherwise. $\hat{x} = -x$ yields better solution)

$$\mathcal{L}(x, \lambda, \mu) = \frac{1}{2}\sum(x_i - y_i)^2 + \mu_0\left(\sum x_i - t\right) + \sum \mu_i x_i$$

1. $x_i = 0$. ⇒ $\mu_i = \mu_0 - y_i$
2. $x_i > 0$. ⇒ $\mu_i = 0$. $x_i = -\mu_0 + y_i$

$x_i = (y_i - \mu_0)^+$.

$\sum(y_i - \mu_0)^+ \leq t$ ⇒ $\begin{cases} < : \text{inside the region. itself} \\ = : \text{find } \mu_0 \end{cases}$



Convergence Analysis

1. Connection to proximal gradient descent

Def. Indicator I_X of a set X:

$$I_X(x) = \begin{cases} 0 & . \ x \in X \\ +\infty & . \ x \notin X. \end{cases}$$

I_X is convex iff. X is convex.

The proximal gradient descent $\operatorname{prox}_{I_X}(y) = \underset{x}{\operatorname{argmin}} \left(\frac{1}{2}\|x-y\|_2^2 + I_X(x) \right)$

$$= \underset{x \in X}{\operatorname{argmin}} \ \frac{1}{2}\|x-y\|_2^2 = P_X(y).$$

By the definition of $I_X(x)$. we know $I_X(x) = t_k I_X(x)$. (for $t_k > 0$).

$$x_{k+1} = P_X(x_k - t_k \nabla f(x_k)) = \operatorname{prox}_{I_X(x)}(x_k - t_k \nabla f(x_k)) = \operatorname{prox}_{t_k I_X(x)}(x_k - t_k \nabla f(x_k))$$

2. **Lemma.** $\varphi: X \to \mathbb{R}$ is μ-strongly convex with a minimum $x^* \in X$. then

$$\varphi(x) \geq \varphi(x^*) + \frac{\mu}{2}\|x - x^*\|_2^2. \quad \forall x \in X.$$

Proof. Fix x. $x_t := tx + \bar{t}x^*$. $t \in [0,1]$ // exist $\underset{\text{convex}}{\tilde{\varphi}(x)}$ s.t. $\varphi(x) = \tilde{\varphi}(x) + \frac{\mu}{2}\|x - x^*\|_2^2$

By convexity of $\tilde{\varphi}$:

$$\underset{\substack{\uparrow \\ \text{minimum}}}{\varphi(x^*)} \leq \varphi(x_t) = \underbrace{\tilde{\varphi}(x_t) + \frac{\mu}{2}\|x_t - x^*\|_2^2 \leq t\tilde{\varphi}(x)}_{\leq} + \bar{t}\underset{\substack{\| \\ \varphi(x^*)}}{\tilde{\varphi}(x^*)} + \frac{\mu t^2}{2}\|x - x^*\|_2^2$$

$$\Rightarrow \quad 0 \leq t\tilde{\varphi}(x) - t\varphi(x^*) + \frac{\mu t^2}{2}\|x - x^*\|_2^2$$

$$\Rightarrow \quad \tilde{\varphi}(x) \geq \varphi(x^*) - \frac{\mu t}{2}\|x - x^*\|_2^2. \qquad (t \in [0,1])$$

$$\Rightarrow \quad \tilde{\varphi}(x) \geq \varphi(x^*)$$

3. Projected Gradient Descent $\Rightarrow$ Proximal Gradient Descent

Equality Constrained Convex QP

$$\min_x f(x) = \frac{1}{2}x^T Q x + g^T x + c \qquad \text{s.t.} \quad Ax = b$$

where $Q \succeq 0$. $A \in \mathbb{R}^{k \times n}$, rank $A = k$.

Lagrangian/KKT Condition: $\begin{cases} \nabla_x \mathcal{L} = Qx^* + g + A^T\lambda^* = 0 \\ \nabla_\lambda \mathcal{L} = Ax^* - b = 0 \end{cases} \Leftrightarrow \underbrace{\begin{pmatrix} Q & A^T \\ A & 0 \end{pmatrix}}_{K := \text{ KKT Matrix}} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} -g \\ b \end{pmatrix}$ KKT System (Might be unsolvable)

$$\text{Null}(K) = \left\{ \begin{pmatrix} x \\ 0 \end{pmatrix} : x \in \text{Null}(A) \cap \text{Null}(Q) \right\}.$$

[Proof] $\begin{cases} Qx + A^T\lambda = 0. \\ \boxed{Ax = 0} \end{cases}$ $\qquad x^TQx = x^T(-A^T\lambda) = -\underset{\overset{\|}{0}}{(Ax)^T}\lambda = 0.$

$Q \succeq 0 \Rightarrow Q = u\Lambda u^T$. $\quad \Lambda = \text{diag}(v_1, \cdots v_n) \quad u = (x_1, \cdots x_n)$

$x = \sum_i \alpha_i x_i$. $\quad x^TQx = x^T\left(\sum_{i=1}^n \alpha_i Q x_i\right) = x^T\left(\sum_{i=1}^n \alpha_i v_i x_i\right) = \sum_i \alpha_i v_i x^T x_i$

Considering $x^T x_i = \sum_j \alpha_j \underbrace{x_j^T x_i}_{} = \alpha_i$, $\quad x^TQx = \sum_{i=1}^n v_i \alpha_i^2 = 0$

~~① $v_i = 0$~~
~~② $v_i > 0 \Rightarrow \alpha_i = 0$.~~

$x = \sum_i \alpha_i x_i \qquad \boxed{Qx = \sum_i \underset{\overset{\|}{0}}{\alpha_i v_i} x_i = 0.}$

$\Rightarrow \quad x \in \text{Null}(A) \cap \text{Null}(Q)$ □

$A^T\lambda = 0. \Rightarrow \lambda = 0.$ □

$\uparrow$ full rank. column

KKT System has no solution iff. $\begin{pmatrix} -g \\ b \end{pmatrix} \notin \text{Range}(K) = \text{Range}(K^T) = \text{Null}(K)^\perp$

$\rightarrow$ exists $\begin{pmatrix} v \\ \lambda \end{pmatrix} \in \text{Null}(K)$ s.t. $\begin{pmatrix} -g \\ b \end{pmatrix}^T \begin{pmatrix} v \\ \lambda \end{pmatrix} \neq 0$. i.e. $g^Tv \neq b^T\lambda$.

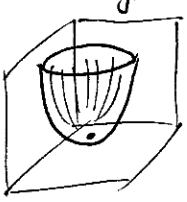
$Av = 0$. $Qv = 0$. $\lambda = 0$. $\Rightarrow g^Tv \neq 0$
$b^T\lambda = 0$.

$$f(x_0 + tv) = f(x_0) + t(Qx_0 + g)^Tv + \frac{1}{2}t^2 v^TQv = f(x_0) + tg^Tv. \longrightarrow -\infty$$

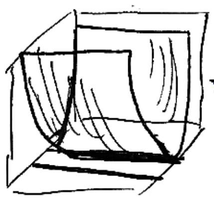
(as $t \rightarrow -\text{sgn}(g^Tv) \cdot \infty$.)

Discuss $f(x) = \frac{1}{2}x^TQx + g^Tx + c$ (unconstrained)

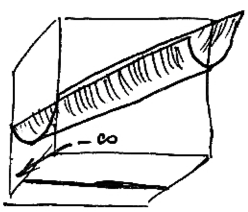
~~K is a nonsingular matrix~~



$Q \succ 0$. unique solution

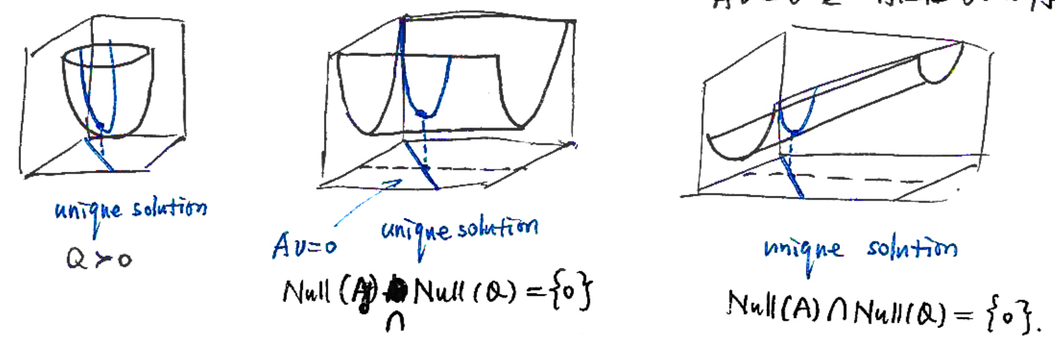


infinite many solutions $g^Tx = 0$ ($g \perp \text{Null}(Q) \neq \{0\}$)



no solution $g \not\perp \text{Null}(Q) \neq \{0\}$

Eq. constrained: $Ax=b \leftarrow A(x_0+v)=b$. $\begin{cases} Ax_0=b \\ Av=0 \end{cases} \leftarrow$ 特征值为0的特征向量 / 也是齐次解



- Nonsingularity of KKT Matrix

KKT Matrix nonsingular $\Longleftrightarrow$ unique solution of KKT Sys. optimal

$Q\succeq 0$. $\operatorname{rank} A = k$ $(A\in\mathbb{R}^{n\times k})$

1) K is nonsingular

2) $Null(Q)\cap Null(A)=\{0\}$ i.e. Q, A have no nontrivial common null space.
i.e. $\begin{cases} Ax=0 \\ Qx=0 \end{cases}$ iff. $x=0$

3) $Ax=0$. $x\neq 0 \Rightarrow x^TQx>0$ i.e. Q is positive definite on $Null(A)$.

4) $F^TQF\succ 0$ for any $F\in\mathbb{R}^{n\times(n-k)}$ s.t. $Range(F)=Null(A)$
i.e. the columns of F are linearly independent solutions of $Ax=0$.

are all equivalent.

[Proof] $K=\begin{pmatrix} Q & A^T \\ A & 0\end{pmatrix}$. $Null(K)=\left\{\begin{pmatrix} x \\ 0\end{pmatrix} : x\in Null(Q)\wedge Null(A)\right\}$.

1) ⇒ 2) $x\in Null(Q)\cap Null(A)$. $K\begin{pmatrix} x \\ 0\end{pmatrix}=0$ $Null(K)=\{0\}$

i.e. $\begin{cases} Qx+A^T\lambda=0 \\ Ax=0\end{cases}$ (Already prove that $\lambda=0$) $\Rightarrow \begin{cases} Qx=0 \\ Ax=0\end{cases}$ iff. $x=0$.

2) ⇒ 1) $\begin{cases} Ax=0 \\ Qx=0\end{cases}$ iff. $x=0$. Let $\lambda=0 \Rightarrow \begin{cases} Qx+A^T\lambda=0 \\ Ax=0\end{cases}$ i.e. $K\begin{pmatrix} x \\ 0\end{pmatrix}=0$
K: nonsingular

2) ⇒ 3). ~~$\left(\begin{cases} Ax=0 \\ Qx=0\end{cases} \Leftrightarrow x=0\right) \Rightarrow Ax=0 \Rightarrow Qx=0$ iff. $x=0$ $\Rightarrow Ax=0 \Rightarrow x^TQx=0$ iff. $x=0$~~
~~$\Rightarrow (Ax=0.\ x\neq 0 \Rightarrow x^TQx>0)$~~

见下页

2) $\Rightarrow$ 3). $\left(\begin{cases} Ax=0 \\ Qx=0 \end{cases} \text{iff. } x=0 \right) \Rightarrow \left(\begin{cases} Ax=0 \\ Qx=0 \end{cases} \Rightarrow x=0 \right) \Rightarrow \left(\begin{cases} Ax=0 \\ x^TQx=0 \end{cases} \Rightarrow x=0 \right)$

$\Rightarrow \left(\begin{cases} Ax=0 \\ x\neq 0 \end{cases} \Rightarrow x^TQx>0 \right)$

3) $\Rightarrow$ 2) . Obvious

3) $\Leftrightarrow$ 4) $x \in Null(A)$ iff $x=Fz$. $Fz\neq 0$ iff $z\neq 0$ (since $y_1, \cdots y_j \cdots$ linearly independent) $F=\{y_1, \cdots y_m\}$

$$Ax=0.\ x\neq 0 \Rightarrow x^TQx>0$$
$$\Updownarrow$$
$$x=Fz.\ z\neq 0 \Rightarrow x^TQx>0$$
$$\Updownarrow$$
$$z\neq 0 \Rightarrow \underbrace{z^TF^TQFz>0}_{\text{definite.}} \Leftrightarrow F^TQF \succ 0.$$

- Newton's Method for Equality Constrained Problem

2nd-order Taylor approximation: at feasible x_k.

$$\min_d\ h(d) := \hat{f}(x_k+d) = f(x_k) + \nabla f(x_k)^Td + \frac{1}{2}d^T\nabla^2 f(x_k)\, d. \quad \text{s.t. } A(x_k+d)=b. \quad (P)$$

x_k: feasible i.e. $Ax_k=b$

(P) $\Rightarrow$ $\min_d h(d)$ s.t. $Ad=0$

(unconstrained case)

$\nabla^2 f(x_k)\cdot d = -\nabla f(x_k)$

without $Ad=0$

KKT System: $\begin{pmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ 0 \end{pmatrix}$ ~ d_k: Newton direction

$x \leftarrow x_0$ $(Ax_0=b)$ ▷ must be feasible! (for $\text{dom} f = \mathbb{R}^n$. use any solution of $Ax=b$ is acceptable)

repeat

  Compute d by solving KKT Sys ▷ The only difference $\rightarrow$ calculate d.

  $t \leftarrow 1$

  while $f(x+td) > f(x) + \alpha t\ \nabla f(x)^Td$ do ▷ backtracking line search

    $t \leftarrow \beta t$

  end while

  $x \leftarrow x+td$

until ① $\|d\| \leq \delta$

② $\sqrt{d^T\nabla^2 f(x)\, d} \leq \delta$ $\leftarrow$ CANNOT USE $\|\nabla f(x)\| \leq \delta$.

(affine invariance) $\nabla f(x^*)=0$ no longer holds for eq-constrained problem

Feasible Descent Method

$$\min_x f(x) \quad \text{s.t.} \quad Ax=b \iff \begin{cases} A\tilde{x}=b \\ \min_z g(z) := f(\underbrace{\tilde{x}+Fz}_{\text{feasible}}) \end{cases}$$

$F=(\underbrace{y_1, \dots y_{n-k}}_{\text{linearly independent}})$ s.t. $Ay_i=0$.

$x_i = Fz = \sum z_i y_i$

<Assume K nonsingular>

[Proof] Just need to prove $x_i = \tilde{x} + Fz_i$. (Induction.)

$$\begin{pmatrix} \nabla^2 f(x_0) & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \Delta x_0 \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} -\nabla f(x_0) \\ 0 \end{pmatrix}$$

Δz_0: $\nabla^2 g(z_0)\, \Delta z_0 = -\nabla g(z_0)$ (solution of ...)

$\Delta x_0 = F\Delta z_0$ $\quad$ ($\Delta x_0 = 0$ iff. $\Delta z = 0$) $\quad$ <proved in Lemma 2. below>

Backtracking line search: stopping criterion

$\|d\| \leq \delta$ i.e. $\|\Delta x\| \leq \delta$. $\iff$ $\|F\Delta z\| \leq \delta$.

$$\lambda_{min}(F^TF)\cdot \|\Delta z\|^2 \leq \underset{\substack{\| \\ (\Delta z)^T F^T F \Delta z}}{\|F\Delta z\|^2} \leq \lambda_{max}(F^TF)\cdot \|\Delta z\|^2$$

$$\therefore \sqrt{\lambda_{min}(F^TF)}\, \|\Delta z\| \leq \underset{\substack{\| \\ \|\Delta x\|}}{\|F\Delta z\|} \leq \sqrt{\lambda_{max}(F^TF)}\, \|\Delta z\| \Rightarrow \underline{\|\Delta z\| \leq \delta'}$$

Lemma 1. K nonsingular $\iff$ $\underset{(Q)}{F^T \nabla^2 f\, F} \succ 0$.

$\nabla^2 g(z_0) = F^T \nabla^2 f\, F$

$\nabla g(z_0) = F^T \nabla f$. $\quad \longleftarrow$ $\Delta x, \Delta z$: well-defined □

Lemma 2. $\nabla^2 f(x_0)\cdot \Delta x_0 + A^T\lambda = -\nabla f(x_0) \implies F^T\nabla^2 f(x_0)\, \Delta x_0 + F^TA^T\lambda = -F^T\nabla f(x_0)$

$A\cdot \Delta x_0 = 0 \quad \Rightarrow \Delta x_0 = Fu.$ $(\sum u_i y_i)$

$$\Rightarrow \underbrace{F^T\nabla^2 f(x_0)\, F}_{\nabla^2 g} u + \cancel{\underset{\substack{\| \\ 0}}{(AF)^T}\lambda} = -\underbrace{F^T\nabla f(x_0)}_{\nabla g}$$

$$\Rightarrow \nabla^2 g \cdot \underset{\substack{\| \\ \Delta z_0}}{u} = -\nabla g \quad (\text{by def.}) \quad \Rightarrow \underline{\Delta x_0 = F\Delta z_0}$$

□

Lemma 3. Backtracking line search

$f(x_0 + t\Delta x) > f(x_0) + \alpha t (\nabla f)^T \Delta x$

$f(x_0) = f(\tilde{x} + Fz_0) = g(z_0)$

$f(x_0 + t\Delta x) = f(\tilde{x} + Fz_0 + tF\Delta z) = f(\tilde{x} + F(z_0 + t\Delta z)) = g(z_0 + t\Delta z)$

$(\nabla f)^T \Delta x = (\nabla f)^T F \Delta z = (F^T\nabla f)^T \Delta z$

Thus, $g(z_0 + t\Delta z) > g(z_0) + \alpha t \nabla g^T \Delta z$.

i.e. Backtracking line search of the two methods give the exactly same t.

cont'd Therefore, $x_i = \tilde{x} + Fz_0 + \Delta x = \tilde{x} + Fz_0 + F\Delta z = \tilde{x} + F(z_0 + \Delta z) = \tilde{x} + Fz_i$.

$\square$

- General Smooth Convex Problems

$$\min_x f(x) \quad \text{s.t.} \quad \begin{array}{l} Ax = b \\ g_j(x) \le 0. \end{array} \qquad \text{(ICP)}$$

① projected G.D.

② No inequality constraints. $\left\{\begin{array}{l}\text{Newton's Method.} \\ \text{projected G.D.}\end{array}\right.$

Interior-point Method (Penalty Method)

reduce the general case to a serial of equality constrained problems by adding penalty term for inequality constraints.

▷ Barrier Method

Feasible set $X = \{x : Ax = b,\ g_j(x) \le 0,\ \forall j\}$

S: interior of the set. $S = \{x : g_j(x) < 0,\ \forall j\}$

Assume $S \ne \emptyset$. Function $B : S \to \mathbb{R}$. (barrier function): $B(x) \to +\infty,\ x (\in S) \to \partial S$.

e.g. Some barrier functions:

$$B(x) = -\sum_{j=1}^m \log(-g_j(x))$$

$$\nabla B(x) = \frac{1}{-g(x)} \nabla g \qquad \underbrace{(\nabla B(x))_k}_{\partial_k B} = \frac{\partial_k g}{-g} \qquad \underbrace{\partial_l(\partial_k B)}_{(\nabla^2 B)_{kl}} = \partial_l\left(\frac{\partial_k g}{-g}\right) = \frac{\partial_l(\partial_k g)}{-g} + \frac{\partial_k g\, \partial_l g}{(-g)^2}$$

$$= \frac{(\nabla^2 g)_{kl}}{-g} + \frac{(\nabla g \nabla g^T)_{kl}}{g^2}$$

$\Rightarrow$ $B(x) = -\sum_{j=1}^m \log(-g_j(x))$ log barrier function

gradient $\nabla B(x) = \sum_{j=1}^m -\frac{1}{g_j(x)} \nabla g_j(x)$

Hessian $\nabla^2 B(x) = \sum_{j=1}^m \frac{1}{g_j^2(x)} \nabla g_j(x) \nabla g_j(x)^T + \sum_{j=1}^m \frac{1}{-g_j(x)} \nabla^2 g_j(x)$

$$I_S(x) = \begin{cases} 0 & x \in S \\ +\infty & x \notin S \end{cases} \rightarrow \text{(ICP)} \Leftrightarrow \min_x f(x) + I_S(x) \quad \text{s.t.} \quad Ax = b$$

indicator of S

$\downarrow$ Approximate

$$\longrightarrow \min_x f(x) + \frac{1}{t} B(x) \quad \text{s.t.} \quad Ax = b$$

feasible requires: $g_j < 0,\ \forall j$. (strictly feasible)

Gradient Descent / Newton's Method: How to make sure each step is within feasible set?

$$\tilde{B}(x) = \begin{cases} -\log(-g(x)) & x<0 \\ +\infty & x \geq 0 \end{cases} \Rightarrow \tilde{f}(x) = \begin{cases} \cdots\cdots & x \in S \\ +\infty & x \notin S \end{cases}$$

<u>Backtracking line search!</u>

while $\left(f(x+td) > f(x) + t\alpha \nabla f(x)^T d \right.$. t too large $\Rightarrow f(x+td) = \underline{+\infty}.\left.\right)$

$t \Leftarrow \beta t$ until finds next step in the feasible set.

Def.

solution of the Problem is called <u>central point</u>. $(x^*(t))$

The curve defined by $x^*(t): (0, +\infty) \longrightarrow X$. is called <u>central path</u>.

<u>$x^*(t) \in S$.</u> $\leftarrow$ The Method called Interior-Point Method.

$t \to \infty$. $x^*(t) \longrightarrow$ optimal solution x^* of the Problem

$\dashrightarrow$ The level curves of $f(x) + \frac{1}{t}B(x) \longrightarrow$ the level curves of $f(x)$.

(The original objective function)

Thm. The Suboptimality of Central Point

Lemma. $x^*(t)$ is $\frac{m}{t}$-suboptimal, i.e. $f(x^*(t)) - f^* \leq \frac{m}{t}$

<u>Proof.</u> Lagrangian? $\leftarrow$ Need to find a form similar to Lagrangian.

$$\nabla f(x^*(t)) + \frac{1}{t}\nabla B(x^*(t)) + A^T\lambda^* = 0$$

$$\nabla f(x^*(t)) + \sum_{j=1}^m \underbrace{-\frac{1}{t g_j(x^*(t))}}_{\mu_j^* > 0} \nabla g_j(x^*(t)) + A^T\lambda^* = 0$$

$$\mu_j^* g_j(x) = -\frac{1}{t}.$$

$$\mathcal{L}(x) := f(x) + \underbrace{\sum_{j=1}^m \underbrace{\mu_j^*}_{>0} \underbrace{g_j(x)}_{\leq 0} + \underbrace{(Ax-b)^T}_{=0} \lambda^*}_{\leq 0}.$$

$\nabla \mathcal{L} = 0 \Rightarrow x^*(t)$

~~obvious. $\mathcal{L}(x^*(t))$~~

~~$\mathcal{L}(x^*(t)) = f(x^*(t)) + \Sigma$~~

To prove $\mathcal{L}(x^*) \leq f^*$. <u>$x^*(t)$ minimize $\mathcal{L}$.</u> $\Rightarrow \mathcal{L}(x^*(t)) \leq \mathcal{L}(x^*)$

$$\mathcal{L}(x^*) = f(x^*) + \underbrace{\sum_{j=1}^m \mu_j^* g_j(x^*)}_{\leq 0} + \underbrace{(Ax^*-b)^T}_{=0} \lambda^* \leq f^*$$

$$\mathcal{L}(x^*(t)) = f(x^*(t)) + \sum_{j=1}^m \underbrace{\mu_j^* g_j(x^*(t))}_{= -\frac{1}{t}} + \underbrace{(Ax^*(t)-b)^T}_{=0} \lambda^*(t)$$

$$= f(x^*(t)) - \frac{m}{t}$$

Thus, $f(x^*(t)) - \frac{m}{t} \leq f(x^*) = f^*$. $\square$

Def. Modified KKT Conditions

$$\begin{cases} Ax^*(t) = b, \quad g_j(x^*(t)) < 0. & \text{— Feasibility} \\ \mu_j^* > 0 & \text{— Nonnegativity} \\ \nabla f(x^*(t)) + A^T \lambda^* + \sum_{j=1}^{m} \mu_j^* \nabla g_j(x^*(t)) = 0 & \text{— Stationarity} \\ \mu_j^* g_j(x^*(t)) = -\frac{1}{t} \approx 0 & \text{— Approximate Complementary Slackness} \end{cases}$$

To achieve ε-suboptimal solution

① $t = \frac{m}{\varepsilon} \longrightarrow$ could be slow

② The barrier method / path-following method : "dynamic" t. t_k increase to $\frac{m}{\varepsilon}$.

To find a better solution based on the solution when t is smaller

▷ The Barrier Method (The path-following Method)

- strictly feasible x_0 s.t. $Ax_0 = b$. $g_j(x_0) < 0 \ \forall j$.
  $t_0 > 0$

- for $k = 0, 1, 2, \cdots$ do

  $x_{k+1} \longleftarrow \underset{x: Ax=b}{\operatorname{argmin}} \left\{ f(x) + \frac{1}{t_k} B(x) \right\}$. starting from x_k

  if $t_k > \frac{m}{\varepsilon}$. then return x_{k+1}. // criterion of stopping

  $t_{k+1} \leftarrow \rho t_k$

$\underline{\rho > 1}$.

In practice. algorithm robust to ρ and t_0. $\rho \in [10, 20]$ recommended.

How do we find a strictly feasible x_0?

- Feasible Problem

$$\min_{x,s} \ s \quad \text{s.t.} \quad Ax = b$$
$$g_j(x) \le s \quad \forall j. \qquad (F) \leftrightarrow \quad \tilde{g}_j(x, s) = g_j(x) - s < 0.$$

Initial Point? $\longrightarrow$ Solve $Ax = b$ $\longrightarrow$ Get a feasible x_0.

$g_j(x_0) \le s \implies s_0 \geq \max\{g_j(x_0)\}$. $\leftarrow$ It's easy to find s_0. $\longrightarrow$ Run barrier method on F.

The original (ICP) is strictly feasible iff. the optimal s value of (F) is negative.

Once we find a $s < 0$ $\longrightarrow$ we get a strictly feasible initial point

↳ Run ICP Barrier Method

Dual Linear Program

Optimization Problem. any feasible x_0 gives an upper bound. $f^* \leq f(x_0)$

What about Lower Bound? $f^* \geq f_{LB} \Leftrightarrow \forall x \in X.\ f(x) \geq f_{LB}$

e.g.

$$\begin{aligned} \min_x \quad & Ax+b \\ \text{s.t.} \quad & g_j(x) \geq 0. \quad \forall j. \end{aligned}$$

Primal LP

$\rightarrow \sum \mu_j g_j(x) \geq 0 \xrightarrow{\text{Separate constants}} \sum \mu_j \hat{g}_j(x) \geq \underbrace{\sum \mu_j g_{cj}}_{\text{constant}}$.

We can choose μ_j to let $\sum \mu_j \hat{g}_j(x) = Ax$

$\Rightarrow$ We get a lower bound!

$$\psi(\mu) := \sum \mu_j g_{cj}. \leq \inf_{x \in X} (Ax+b) \leq f^*$$

$$\Leftrightarrow \quad \boxed{\max_\mu \ \psi(\mu) \quad \text{s.t.} \begin{cases} \sum \mu_j \hat{g}_j(x) = Ax \\ \mu \geq 0 \end{cases} \quad \text{Dual LP}}$$

Def.

$$\begin{aligned} \min_x \quad & f(x) = c^T x \\ \text{s.t.} \quad & Ax = b \\ & Gx \geq h \end{aligned}$$

Primal LP

$\longleftrightarrow$

$$\begin{aligned} \max_{\lambda, \mu} \quad & b^T \lambda + h^T \mu \\ \text{s.t.} \quad & A^T \lambda + G^T \mu = c \\ & \mu \geq 0 \end{aligned}$$

Dual LP

$\sum_{i=1}^{k} (a_i^T x = b_i) \cdot \lambda_k$

$+ \sum_{j=1}^{m} (g_j^T x \geq h_j) \cdot \mu_j. \rightarrow \lambda^T(Ax-b) + \mu^T(Gx-h) \geq 0 \quad (\mu > 0)$

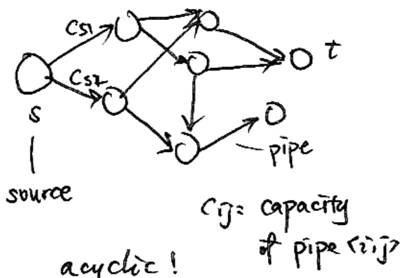
$(\lambda^T A + \mu^T G) x \geq \lambda^T b + \mu^T h. \Rightarrow \underbrace{(A^T\lambda + G^T\mu)^T}_{= c^T} x \geq \underbrace{b^T \lambda + h^T \mu}_{\text{is a number!}}$

Specially, When "$Gx \geq h$" is "$x \geq 0$". $G = I$.

we have

$$\begin{aligned} \max_{\lambda, \mu} \quad & b^T \lambda \\ \text{s.t.} \quad & A^T \lambda + \mu = c \\ & \mu \geq 0 \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \max_{\lambda} \quad & b^T \lambda \\ \text{s.t.} \quad & A^T \lambda \leq c \end{aligned}$$

Def. Maximum Flow Problem (最大流)



acyclic!

$$\max_f \; |f| := \sum_{(s,j)\in E} f_{sj}$$

s.t. $0 \le f_{ij} \le c_{ij} \quad \forall (i,j)\in E$

$$\sum_{(i,k)\in E} f_{ik} = \sum_{(k,j)\in E} f_{kj} \quad \forall k\in V\setminus\{s,t\}.$$

$a_{ij} \times \quad -f_{ij} \le 0$

$b_{ij} \times \quad f_{ij} \le c_{ij}$

$x_k \times \quad \left(\sum_{(i,k)\in E} f_{ik} - \sum_{(k,j)\in E} f_{kj} = 0\right)$

$$\sum_{(i,j)\in E} (-a_{ij} f_{ij} + b_{ij} f_{ij}) + \sum_{k\in V\setminus\{s,t\}} x_k \left(\sum_{(i,k)\in E} f_{ik} - \sum_{(k,j)\in E} f_{kj}\right) = \sum_{(i,j)\in E} c_{ij} b_{ij}$$

$$\| $$

$$\sum_{\substack{k\in V\setminus\{t\}\\ (i,k)\in E}} x_k f_{ik} - \sum_{\substack{k\in V\setminus\{s\}\\ (k,j)\in E}} x_k f_{kj}$$

Consider the coefficient of f_{ij}.

$$\Rightarrow \begin{cases} b_{sj} - a_{sj} + x_j = 1 & \forall (s,j)\in E \\ b_{it} - a_{it} - x_i = 0 & \forall (i,t)\in E \\ b_{ij} - a_{ij} + x_j - x_i = 0 & \forall (i,j)\in E,\ i\ne s,\ j\ne t. \end{cases}$$

to match the form of the obj func: $|f| = \sum_{(s,j)\in E} f_{sj}$

Define $x_s = 1,\ x_t = 0 \longrightarrow \forall (i,j)\in E.\ \ b_{ij} - a_{ij} + x_j - x_i = 0$

The dual LP:

$$\min_{a,b,x} \sum_{(i,j)\in E} c_{ij} b_{ij}$$

s.t. $b_{ij} - a_{ij} + x_j - x_i = 0. \quad \forall (i,j)\in E$

$a_{ij} \ge 0 \quad b_{ij} \ge 0$

i.e.

$$\min_{b,x} \sum_{(i,j)\in E} c_{ij} b_{ij}$$

s.t. $b_{ij} + x_j - x_i \ge 0$

$b_{ij} \ge 0 \quad \forall (i,j)\in E.$

Relaxation

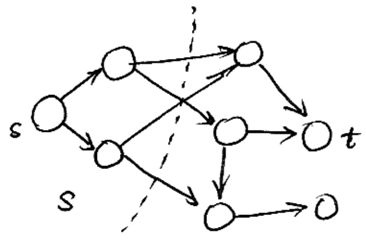
Minimum Cut Problem:

$$\min_{b,x} \sum_{(i,j)\in E} c_{ij} b_{ij}$$

a integer program

s.t. $b_{ij} + x_j - x_i \ge 0$

$b_{ij} \in \{0,1\}.\ \ x_{ij} \in \{0,1\} \quad \forall (i,j)\in E$



a cut
(an s-t cut)

Partition $\{S, \bar{S}\}$.

$s \in S$.

$t \in \bar{S}$.

The capacity of the cut:

$$c(S, \bar{S}) = \sum_{\substack{(i,j)\in E \\ i\in S,\, j\in \bar{S}}} c_{ij}$$

min-cut prob.

$$\min_{(S,\bar{S}) \text{ is a cut}} \quad c(S, \bar{S})$$

$x_i = 0$ for $i \in S$; $x_i = 1$ for $i \in \bar{S}$

$b_{ij} = 1$ if $i \in S$, $j \in \bar{S}$ (cut edge) $\longleftarrow$ $b_{ij} \geq x_i - x_j$

Adjacency
~~Adjacency~~ : represented in C_{ij}

$\geq 1-0$ ~ 1

$\geq 0-0$
$\geq 1-1$ $\sim$ set to 0. (not a cut edge)
$\geq 0-1$

- Weak and Strong Duality

$$\min_x \; c^T x \quad \text{s.t.} \; Ax \geq b, \; x \geq 0 \qquad (P)$$

$$\max_y \; b^T y \quad \text{s.t.} \; A^T y \leq c, \; y \geq 0 \qquad (D)$$

Def. Weak Duality

If x is primal feasible and y is dual feasible, we have

$$c^T x \geq b^T y$$

[Proof.] $c^T x \geq (A^T y)^T x = y^T A x \geq y^T b = b^T y$.

Def. Strong Duality

If one of (P) and (D) has finite optimal value with solution x^*/y^*. So does the other.

Moreover, $c^T x^* = b^T y^*$

Note: finite. optimal value: For min (e.g., WLOG (without loss of generality)) : $< +\infty$ ($\uparrow$ if $= +\infty$ infeasible) $> -\infty$ ($\uparrow$ if $= -\infty$: unbounded below)

Thm. The max-flow value is equal to the min-cut capacity.

[Proof.] Set $f_{ij} = 0 \rightarrow$ feasible (for max-flow)

$0 \leq |f| \leq \sum_{(i,j)\in E} c_{ij}$

$\Big\}$ finite optimal value

Lemma. The min-cut capacity is equal to the optimal value of the relaxation of the LP Problem.

(cont'd) LP: $\displaystyle \min_{x,b} \sum_{ij} c_{ij} b_{ij}$ s.t. $b_{ij} \geq x_i - x_j$, $b_{ij} \geq 0$.

(cont'd) let the optimal value of IP (min-cut) is C_{IP}. of LP: C_{LP}

$C_{IP} \geq C_{LP}$: IP's solution is for sure feasible for LP.

$C_{IP} \leq C_{LP}$: Need to show: exists an s-t cut with capacity $\leq C_{LP}$:

LP optimal solution: b_{ij}^*. x_i^*

Use probabilistic method to prove

$U \sim \text{Uniform}(0,1)$. random variable. —— if $x_i^* > u$. $x_i^* \in S_U$.
$\leq$ $\bar{S}_U$.

$$\mathbb{E}\left(c(S_U, \bar{S}_U)\right) = \int_0^1 c(S_U, \bar{S}_U)\, dU = \sum_{(i,j)\in E} c_{ij}\, \mathbb{E}\left(\mathbb{1}(i \in S_U, j \in \bar{S}_U)\right).$$

0-1 Distribution!

$$= \sum_{(i,j)\in E} c_{ij} \mathbb{P}(i \in S_U, j \in \bar{S}_U) = \sum_{(i,j)\in E} c_{ij}\, \mathbb{P}(u \leq x_i^*, u > x_j^*).$$

$$= \sum_{(i,j)\in E} c_{ij}\, \mathbb{P}(x_j^* \leq u < x_i^*) \leq \sum_{(i,j)\in E} c_{ij}\, (x_i^* - x_j^*)^+$$

$$= \sum_{(i,j)\in E} c_{ij} \cdot b_{ij}^* = C_{LP}.$$

i.e. $\mathbb{E}(C_{IP}) \leq C_{LP}$. exists at least one point s.t. $C_{IP} \leq C_{LP}$.

Thus, $C_{IP} = C_{LP}$.

Qed. □

(Back to ~~proof~~ Thm. proof)

LP is the dual problem of the max-flow prob. i.e. has the same optimal value and optimal solution.

Therefore, max-flow value is equal to the min-cut capacity. □

Qed.

- Dual LP via Lagrangian

$$\mathcal{L}(x, \lambda, \mu) = c^T x - \lambda^T(Ax - b) - \mu^T(Gx - h) \quad \leftrightarrow \quad \begin{array}{ll} \min & c^T x \\ \text{s.t.} & Ax - b = 0, -Gx + h \leq 0. \end{array}$$

$\mu \geq 0$. $x \in \mathcal{X}$.

$$\rightarrow f(x) = c^T x \geq c^T x - \lambda^T \underset{\overset{\|}{0}}{(Ax - b)} - \mu^T \underset{\overset{\vee}{0}}{(Gx - h)} = \mathcal{L}(x, \lambda, \mu)$$

$$f^* = \inf_{x \in \mathcal{X}} f(x) \geq \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda, \mu) \geq \inf_{x} \mathcal{L}(x, \lambda, \mu) =: \phi(\lambda, \mu)$$

Maximize lower bound $\longrightarrow$ $\max\limits_{\lambda, \mu} \phi(\lambda, \mu)$ s.t. $\mu \geq 0$ (P)

$$\mathcal{L}(x, \lambda, \mu) = (c^T - \lambda^T A - \mu^T G)x + \lambda^T b + \mu^T h.$$

$$\underset{\text{value}}{(P)^*} = \begin{cases} b^T\lambda + h^T\mu. & \text{if } c^T - \lambda^T A - \mu^T G = 0 \\ -\infty & \text{otherwise} \end{cases} \iff \begin{array}{l} \text{s.t. } A^T\lambda + G^T\mu = c, \ \mu \geq 0. \\ \max\limits_{\lambda,\mu} \psi(\mu, \lambda) = b^T\lambda + h^T\mu. \end{array}$$

- Lagrangian Dual Function

$$\min_x f(x) \quad \text{s.t.} \quad h_i(x) = 0 , \quad i = 1, 2, \cdots, k \qquad g_j(x) \le 0 , \quad j = 1, 2, \cdots, m$$

$$\longrightarrow \quad \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_i \lambda_i h_i(x) + \sum_j \mu_j g_j(x)$$

The **Lagrangian Dual Function** is

$$\phi(\lambda, \mu) = \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda, \mu) = \inf_{x \in \mathcal{X}} \left(f(x) + \sum_i \lambda_i h_i(x) + \sum_j \mu_j g_j(x) \right)$$

We discuss the case when $\mathcal{X} = \mathbb{R}^n$.

Thm.

$$f^* \ge \phi(\lambda, \mu)$$

i.e. the Lagrangian dual function is always a lower bound.

[Proof.] Infeasible. $f^* = +\infty$. ✓

Feasible.

$$f(x) \ge f(x) + \sum_i \lambda_i \underbrace{h_i(x)}_{=0} + \underbrace{\sum_j \mu_j g_j(x)}_{\le 0} \qquad (\mu_j \ge 0,\ g_j(x) \le 0)$$

$$\big\downarrow \text{min}$$

$$\cdot\ f^* \ge \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda, \mu) =: \phi(\lambda, \mu)$$

Thm. Concavity of Lagrangian Dual Function

[Proof.] Notice when $\alpha_i(x)$ is convex. $\alpha(x) := \sup_i \alpha_i(x)$ is also convex.

$\mathcal{L}(x, \lambda, \mu)$ is affine in (λ, μ). $\longrightarrow$ $\phi(\lambda, \mu)$: inf of a series of affine function.

Thus. $$\phi(\lambda, \mu) = -\sup_{x \in X} \underbrace{\Big(\underbrace{-f(x)}_{\text{constant}} - \underbrace{\sum_i \lambda_i h_i(x) - \sum_j \mu_j g_j(x)}_{\text{convex function}} \Big)}_{\text{convex function}}$$ is concave. Qed. □

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

WEAK DUALITY : $f^* \ge \phi^*$. i.e. $\inf_{x \in X} f(x) \ge \sup_{\lambda, \mu : \mu \ge 0} \phi(\lambda, \mu)$ (always holds)

STRONG DUALITY : $f^* = \phi^*$. (does not hold in general)

Def. duality gap

$$\phi(\lambda, \mu) \le \underbrace{\phi^* \le f^*}_{\text{gap}} \le f(x)$$

When does strong duality hold? — One condition (not necessary) is as follows.

- Slater's Condition

convex problem $\min_{x \in D} f(x)$ $\qquad D = \operatorname{dom} f \cap \bigcap_{j=1}^{m} \operatorname{dom} g_j$

s.t. $g_j(x) \leq 0 \quad (j = 1, 2, \cdots, m)$

$h(x) = Ax - b = 0$ $\qquad$ (P)

Def. / Thm. Slater's Condition

(P) is strictly feasible i.e. $\exists x \in \operatorname{int} D$ s.t. $g_j(x) < 0$ for $j = 1, 2, \cdots, m$ and $Ax = b$.

the domain. most cases $D = \mathbb{R}^n$.

Refined Slater's Condition

If some $g_j(x)$ are affine. "$g_j(x) < 0$" can be relaxed to "$g_j(x) \leq 0$".

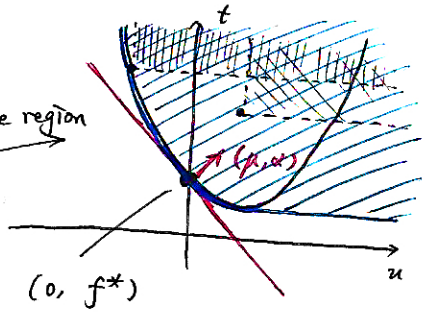
Thm. Slater's Thm.

Strong duality holds for (P) under (refined) Slater's Condition.

Furthermore, if $\phi^* > -\infty$. it is attained by some (λ^*, μ^*).

[Proof.] $g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_j(x) \\ \vdots \\ g_m(x) \end{pmatrix}$ $\qquad \begin{cases} u = g(x) \\ t = f(x) \end{cases}$

$C := \{(u, t) : \exists x \in D \text{ s.t. } g(x) \leq u.\ f(x) \leq t.\}$ → The blue region



$\underline{(0, f(x)) \in C}$ ☆!

C is convex → exists a supporting hyperplane

$$(\mu, \alpha) \cdot (u, t) \geq (\mu, \alpha) \cdot (0, f^*) = \alpha f^*$$

$(+\infty, +\infty) \in C \rightarrow \mu, \alpha \geq 0$

$\downarrow u = g(x),\ t = f(x)$

$$\mu g(x) + \alpha f(x) \geq \alpha f^*$$

if $\alpha > 0$. $\quad \underbrace{\mu^* = \frac{\mu}{\alpha} g(x) + f(x)}_{\mathcal{L}(x, \mu^*)} \geq f^*. \Rightarrow \inf_x \mathcal{L}(x, \mu^*) \geq f^*$

~~$\inf_x \mathcal{L}(x, \mu^*)$~~ $\qquad$ i.e. $\phi(\mu^*) \geq f^*$

$\phi^* \geq f^*$

Meanwhile. $f^* \geq \phi^*$ $\quad \Rightarrow f^* = \phi^*$ $\qquad$ □

if $\alpha = 0$.

Slater's condition: $\exists x_0 \in \operatorname{int} D$. $\subset D$ s.t. $g_j(x) < 0$. $\forall j = 1, 2, \cdots, m$; $Ax = b$

$\rightarrow$. $\exists u_0 < 0$. $(u_0, t_0) \in C$.

$\Rightarrow$ Supporting Hyperplane can't be vertical. $\underline{\alpha \neq 0}$ $\qquad$ □

primal <min>	dual <max>		
feasible	feasible	$\longrightarrow -\infty < f^* = \phi^* < +\infty$	Strong duality
feasible	infeasible	$\longrightarrow f^* = \phi^* = -\infty$	Strong duality
infeasible	feasible	$\longrightarrow f^* = \phi^* = +\infty$	Strong duality
infeasible	infeasible	$f^* = +\infty,\ \phi^* = -\infty$	

$$\min_x f(x) \quad \text{s.t. } h(x)=0,\ g(x)\le 0 \qquad \Big| \qquad \max \phi(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) \quad \text{s.t. } \mu \ge 0$$

Thm. KKT conditions still hold

i.e.

$$h(x^*)=0,\ g(x^*)\le 0$$
$$\mu^* \ge 0$$
$$\nabla_x \mathcal{L}(x^*,\lambda^*,\mu^*)=0$$
$$\mu_j^* g_j(x^*)=0$$

iff.

$$f^* = \phi^* \quad \text{(strong duality)}$$
$$f^* = f(x^*)$$
$$\phi^* = \phi(\lambda^*,\mu^*)$$

[Prof.] $h(x^*)=0.\ g(x^*)\le 0.\ \mu^*\ge 0$: $\Leftrightarrow$ Obvious.

$$\phi(\lambda,\mu) = \inf_x \mathcal{L}(x,\lambda,\mu) \le \mathcal{L}(x,\lambda,\mu) = f(x) + \underbrace{\lambda^T h(x)}_{=0} + \underbrace{\mu^T g(x)}_{\le 0} \le f(x)$$

KKT $\rightarrow \phi(\lambda^*,\mu^*) \le f(x^*)$ $\left(\text{where } f(x^*)+\lambda^{*T}h(x^*)+\mu^{*T}g(x^*) = f(x^*)\right)$

equality holds iff. $\nabla_x \mathcal{L}(x^*,\lambda^*,\mu^*)=0$ (optimality condition of $\mathcal{L}$)

Thus $\phi(\lambda^*,\mu^*) = f(x^*)$. i.e. strong duality holds.

Meanwhile, $\phi(\lambda^*,\mu^*) \le \phi^* \le f^* \le f(x^*) \Rightarrow$ All is equality.

Strong Duality $\rightarrow$ Inverse. Qed.

$\square$

Review

- Convex Set

$x, y \in C.\quad \theta \in [0,1] \Rightarrow \theta x + \bar{\theta} y \in C.$

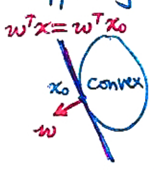
$\sum_{i=1}^{k} \theta_i x_i.\quad \theta_i > 0\ \sum_{i=1}^{k} \theta_i = 0 \longrightarrow$ convex combination

convex null $=:$ $\mathrm{conv}\, S = \left\{ \sum_{i=1}^{m} \theta_i x_i,\ m \in \mathbb{N};\ x_i \in S.\ \theta_i \geq 0;\ \sum_{i=1}^{m} \theta_i = 1 \right\}.$

Intersection / Affine Image $\longrightarrow$ Convex-perserving Operation.

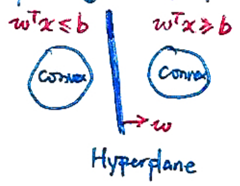
Projection: $P_C(x) = \underset{z \in C}{\mathrm{argmin}}\ \frac{1}{2} \|x - z\|_2^2 .$

Supporting Hyperplane Thm.



C is a nonempty, convex set.
$x_0 \in \partial C.\quad C \subseteq \mathbb{R}^n$
exists $w \in \mathbb{R}^n \setminus \{0\}$ s.t.
$\forall x \in C,\ \langle w, x \rangle \leq \langle w, x_0 \rangle$

Separating Hyperplane Thm.



C_1, C_2 are nonempty, convex sets in $\mathbb{R}^n$
$C_1 \cap C_2 = \emptyset.$
exists $w \in \mathbb{R}^n \setminus \{0\},\ b \in \mathbb{R}$ s.t.
$w^T x \leq b \quad \forall x \in C_1$
$w^T x \geq b \quad \forall x \in C_2$

Proving Methods: 1) Definition; 2) Convex-perserving Operation
3) Sublevel/Superlevel of convex/concave functions
4) Epigraph/Hypograph

- Convex function

$x, y \in \mathrm{dom} f.\quad \theta \in [0,1].\qquad f(\theta x + \bar{\theta} y) \leq \theta f(x) + \bar{\theta} f(y) \longleftarrow$ affine: Both ~~Must~~ convex (and concave)

$\theta \in (0,1)\qquad <\qquad$: strict convexity

$f(x) - \frac{m}{2} \|x\|_2^2$ is convex : m-strong convexity

$\mathrm{epi} f = \{ (x, y) : x \in \mathrm{dom} f,\ y \geq f(x) \}$

$C_\alpha f = \{ x \in \mathrm{dom} f :\ f(x) \leq \alpha \}.\qquad$ e.g. Norm, Affine. $\log \sum x_i$

Commonly Used Method: $g(t) = f(x + td)$
$\uparrow$ direction might be restricted (strongly convex. e.g.)

First-order condition

$f(y) \geq f(x) + \nabla f(x)^T (y - x).\qquad \forall x, y \in \mathrm{dom} f$ —— convex

$>\qquad \forall x \neq y \in \mathrm{dom} f\qquad$ strictly convex

$\geq\qquad + \frac{m}{2} \|x - y\|_2^2 \qquad \forall x, y \in \mathrm{dom} f\qquad$ m-strongly convex

Second-order condition

$\nabla^2 f(x) \succeq 0.$ $\forall x \in \mathrm{dom} f$ — convexity
(eigen value ≥ 0)

$\nabla^2 f(x) \succ 0.$ $\forall x \in \mathrm{dom} f$ — strict convexity
(eigen value > 0)

$\nabla^2 f(x) \succeq mI$ $\forall x \in \mathrm{dom} f$ — m-strong convexity
(eigen value $\geq m$)

Convexity-perserving Operations

$\underset{\forall \text{ convex}}{\sum c_i f_i(x)}$ $\underset{\text{convex}}{f(Ax+b)}$ $h(f_1(x), \dots f_m(x))$ — h: certain composition, f_i: convex $\sup_{i \in I} f_i(x)$ — f_i: convex $\inf_{y \in C} g(x, y)$ — C: convex, g: convex

- Convex Optimization & Optimization Problems

$f^* = \inf_{x \in D} f(x)$. may not be attained.

$$\min_x f(x) \quad \text{s.t. } g(x) \leq 0,\ h(x) = 0$$

$f(x^*) \leq f(x)$ $\forall x \in X$ global minimum

$f(x^*) \leq f(x)$ $\forall x \in X \cap B(x^*, \delta)$ local minimum

C.P.: $\min f(x)$
s.t. $g(x) \leq 0$ — to ensure we optimize on a convex set.
$h(x) = Ax - b = 0$
→ must be convex (Do not transform!) (Unless you are required to.)

$\Downarrow$

a local ~~extrema~~ minimum is a global minimum. (Might not exist / not unique)

Optimality Condition:

f is smooth and convex.

$\nabla f(x^*) = 0.$ — unconstrained

$\nabla f(x^*)^T (x - x^*) \geq 0.$ $\forall x \in X.$ — constrained

$$\begin{cases} \nabla f(x^*)^T + A^T \lambda^* = 0 \\ Ax^* = b. \end{cases}$$ — equality-constrained

A: full rank *

KKT:
(necessary at regular point)

$$\begin{cases} h(x^*) = 0. \quad g(x^*) \leq 0. & \text{primal feasibility} \\ \mu^* \geq 0 & \text{dual feasibility} \\ \nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0 & \text{stationarity} \\ \mu_j^* g_j(x^*) = 0. \quad \forall j & \text{Complementary slackness} \end{cases}$$

$\mathcal{L} = f(x) + \lambda h(x) + \mu g(x)$

inequality-constrained

↳ ① $\mu_j = 0$ ~~[illegible]~~
② $\mu_j \neq 0 \Rightarrow g_j = 0$

or ① g_j inactive ($g_j < 0$) $\Rightarrow \mu_j = 0$
② g_j active ($g_j = 0$)

regular point: $\nabla h(x) \neq 0.$ critical point: $\nabla h(x) = 0$
↖ 边界点

- Lagrange Duality

primal prob. : $\min_x f(x)$ s.t. $g(x) \le 0,\ h(x) = 0$

$\phi(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu)$ where $\mathcal{L}(x, \lambda, \mu) = f(x) + \mu^T g(x) + \lambda^T h(x)$

dual function <always concave> $(\mu \ge 0)$

dual problem : $\max_{\lambda, \mu} \phi(\lambda, \mu)$ s.t. $\mu \ge 0$

→ weak duality : $\phi^* \le f^*$ (always hold) ---> $\phi(\lambda, \mu) \le \phi^* \le f^* \le f(x)$ for $\mu \ge 0$, $x \in X$

→ strong duality: $\phi^* = f^*$.

Slater's condition: $\exists\, x \in \operatorname{int} D$ that is feasible.

KKT conditions hold $\Leftrightarrow$ strong duality + 2× optimality (primal + dual)

$f^* = \phi^*$ $\qquad f^* = f(x^*)$. $\qquad \phi^* = \phi(\lambda^*, \mu^*)$

- Algorithm

  - Smooth ; Unconstrained

    ▷ descent method : $x_{k+1} = x_k + t_k d_k$

    ▷ descent direction

      1) gradient descent $\qquad d_k = -\nabla f(x_k)$

      2) pure Newton's Method $\qquad d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$

    ▷ step size

      1) constant $\qquad t_k = t$

      2) exact line search $\qquad t_k = \operatorname{argmin}_s f(x_k - s \nabla f(x_k))$

      3) backtracking line search (damped Newton's Method for Newton's Method) $\qquad -f(x_k) + f(x_k - t_k \nabla f(x_k)) \ge \alpha t_k \|\nabla f(x_k)\|_2^2$

      * $\underline{t_k = \beta t_{k-1}}$

  * Smooth+nonsmooth $(f+h)$

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

  - Smooth ; Constrained

    ▷ constraint elimination

    ▷ Newton's Method : (for Equality constraints) Solve KKT System $\begin{pmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -\nabla f(x_k) \\ 0 \end{pmatrix}$

    ▷ * Inequality constraints

    Projected GD : $x_{k+1} = P_X(x_k - t_k \nabla f(x_k))$; barrier method.

Review (Part II)

$f(x) = Ax+b \qquad f'(x) = A \qquad \nabla f(x) = A^T$

$f(x) = \frac{1}{2}x^TAx \qquad f'(x) = x^TA \qquad \nabla f(x) = A^Tx$ (A need to be symmetric)

Note: $x^TAx \in \mathbb{R}$. A is not symmetric. $\tilde{A} = \frac{1}{2}(A+A^T) \implies x^TAx = x^T\tilde{A}x$

CHAIN RULE: $h(x) = f(g(x))$. $\nabla h(x) = \nabla g(x)\, \nabla f(g(x))$

$h'(x) = f'(g(x))\, g'(x)$

e.g. $g(x) = f(Ax+b)$, $\nabla g(x) = A^T \nabla f(Ax+b)$

Positive Semidefinite / Definite:

$A \succ 0$ iff. $D_k(A) > 0$. for $k = 1, 2, \cdots, n$ — the leading principal minor of order k. 只考察顺序主子式即可

$A \succeq 0$ iff. $\det A_I \geq 0$ for all $I \subseteq \{1, 2, \cdots, n\}$ — 考察所有主子式; the principal minor of order k

* 主子式定义: $A_I = \begin{pmatrix} a_{i_1i_1} & a_{i_1}a_{i_2} & \cdots & a_{i_1}a_{i_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_ki_1} & a_{i_k}a_{i_2} & \cdots & a_{i_k}a_{i_k} \end{pmatrix}$

$\nabla f(x^*) = 0$. $\nabla^2 f(x^*) \succeq 0 \Leftarrow$ x^*: local minimum

Conversion of LP: 不确定会不会有

LP $\min c^Tx$ s.t. $Bx \leq d$. $Ax = b$ $\longrightarrow$ std LP: $\min c^Tx$ s.t. $\tilde{A}x = \tilde{b}$, $x \geq 0$

① $Bx \leq d \Leftrightarrow \begin{matrix} Bx + s = d \\ s \geq 0 \end{matrix}$

② $x_i = x_i^+ - x_i^-$.

$x_i^+ = \begin{cases} x_i. & x_i > 0 \\ 0. & x_i < 0 \end{cases} \qquad x_i^- = \begin{cases} 0 & x_i > 0 \\ -x_i & x_i < 0 \end{cases}$

inequality form LP: $\min c^Tx$ s.t. $\tilde{A}x \leq \tilde{b}$

① 消去 (eliminate) 一些变量. [Recommended]

② $Ax = b \Leftrightarrow \begin{pmatrix} A \\ -A \end{pmatrix} x \leq \begin{pmatrix} b \\ -b \end{pmatrix}$

L-smoothness: ∇f is L-Lipschitz continuous. $\Leftrightarrow$ $\nabla^2 f \preceq LE$ where $E = \mathrm{diag}(1, 1, \cdots 1)$

$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2$.

Product $\xrightarrow{\ln}$ Sum.