Unconstrained Minimization

(1) unconstrained minimization

minimize $f(x)$ ↳ convex and differentiable

$\Rightarrow x^*$ is optimal $\Leftrightarrow \nabla f(x^*)=0$ optimality condition

iterative algorithm $x^{(0)}, x^{(1)}, \dots, \in \mathrm{dom} f$. $f(x^{(k)}) \to p^*$ as $k\to\infty$

converge if $f(x^{(k)}) - p^* \le \epsilon$

① initial point

要求: $\begin{cases} x^{(0)} \in \mathrm{dom} f \\ \text{sublevel set } S=\{x\in \mathrm{dom} f \mid f(x)\le f(x^{(0)})\} \text{ should be closed.}\end{cases}$

↳ satisfied for all $x^{(0)} \in \mathrm{dom} f$ if f is closed (its sublevel sets are closed)

Definition: a function $f: \mathbb{R}^n \to \mathbb{R}$ is closed if for each α. the sublevel set $\{x\in \mathrm{dom} f \mid f(x)\le \alpha\}$ is closed. then the function is closed.

Properties: ① if f is a continuous function and $\mathrm{dom} f$ is closed, then f is closed ($\mathbb{R}^n$ is closed)

② if f is a continuous function and $\mathrm{dom} f$ is open. then f is closed iff it converges to ∞ along every sequence converging to a boundary point of $\mathrm{dom} f$.

② Strong convexity:

f is strongly convex on S if there exists an $m>0$ such that:

$\nabla^2 f(x) \succeq mI$ → remember f is convex iff $\nabla^2 f(x)\succeq 0$

f is strictly convex iff $\nabla^2 f(x) \succ 0$

↳ 意味着 $\nabla^2 f(x)$ 的最小特征值不小于 m.

Taylor 二次展开: for $y, x\in S$

$$f(y) = f(x)+\nabla f(x)^T(y-x)+\frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x)$$

由于 $\nabla^2 f(x)\succeq mI$ $\Rightarrow$ RHS $\ge f(x)+\nabla f(x)^T(y-x)+\frac{m}{2}(y-x)^T\cdot I\cdot(y-x)$

$= f(x)+\nabla f(x)^T(y-x)+\frac{m}{2}\|y-x\|_2^2$ norm 2: $\|x\|_2 = \sqrt{\langle x,x\rangle} = \sqrt{x^T\cdot x}$

联系 convex function 的 1st-order condition: $f(y)\ge f(x)+\nabla f(x)^T(y-x)$

$\Rightarrow$ 可以说现在 $f(y)\ge f(x)+\nabla f(x)^T(y-x)+\frac{m}{2}\|y-x\|_2^2$ is a better lower bound on $f(y)$ (if $m>0$)

$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2$

↳ convex function for y ⇒ $y = x - \frac{\nabla f(x)}{m}$ 时有最小值

⇒ $RHS = f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|_2^2$

$\geq f(x) + \nabla f(x)^T\left(x - \frac{\nabla f(x)}{m} - x\right) + \frac{m}{2}\left\|x - \frac{\nabla f(x)}{m} - x\right\|_2^2$

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

⇒ $f(y) \geq f(x) + \frac{1}{2m}\|\nabla f(x)\|_2^2$ 成立 for all $y \in S$.

又∵ $f(y) \geq p^*$

⇒ $p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$

⇒ $f(x) - p^* \leq \frac{1}{2m}\|\nabla f(x)\|_2^2$

↳ optimality condition (iteration terminates if $f(x) - p^* \leq \epsilon$)

由上可知 $\|\nabla f(x)\|_2^2$ 可提供 $f(x) - p^*$ 的 upper bound.

⇒ $\|\nabla f(x)\|_2 \leq (2m\epsilon)^{1/2} \Rightarrow f(x) - p^* \leq \epsilon$ conceptual stopping criterion because m&M are rarely known

↳ suboptimality condition

↳ 可证当 $\nabla f(x)$ 很小时, $f(x) \to p^*$. (useful as a stopping criteria)

还可证明得到 $\|x - x^*\|_2 \leq \frac{2}{m}\|\nabla f(x)\|_2$

if f is strongly convex (即 $\nabla^2 f(x) \succeq mI$) ⇒ $\nabla^2 f(x)$ 最大特征值也有界

即 $\nabla^2 f(x) \preceq MI$

同样：有 $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{M}{2}\|y-x\|_2^2$ for $\forall x, y \in S$

⇒ $p^* \leq f(x) + \nabla f(x)^T(y-x) + \frac{M}{2}\|y-x\|_2^2, \forall x, y \in S$

RHS achieves its minimum (for fixed x) if $y = x - \frac{\nabla f(x)}{M}$

⇒ $p^* \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$

⇒ $MI \succeq \nabla^2 f(x) \succeq mI$ ⇒ $\frac{1}{2m}\|\nabla f(x)\|_2^2 \geq f(x) - p^* \geq \frac{1}{2M}\|\nabla f(x)\|_2^2$

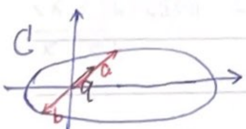
conditional # of $\nabla^2 f(x)$ (最大特征值与最小特征值之比) 上界是 $k = \frac{M}{m}$

理解 conditional #: for a convex set C, the width of C in the direction of q ($\|q\|_2 = 1$)

as $W(C, q) = \sup_{z \in C} q^T z - \inf_{z \in C} q^T z$

minimum width: $W_{min} = \inf_{\|q\|_2 = 1} W(C, q)$

maximum width: $W_{max} = \sup_{\|q\|_2 = 1} W(C, q)$



$W(C, q) = q^T a - q^T b = \|a\|_2 + \|b\|_2$

理解 width

conditional #: $\mathrm{cond}(C) = \frac{W_{max}^2}{W_{min}^2}$

⇒ if conditional # of C is small ⇒ the set has approximately same width in all directions (nearly spherical)

(2) Descent Method → $f(x^{(k+1)}) < f(x^{(k)})$ except when $x^{(k)}$ is optimal

$x^{(k+1)} = x^{(k)} + \underbrace{t^{(k)}}_{\text{step size/step length}} \underbrace{\Delta x^{(k)}}_{\text{step / search direction}}$

又由于 $f(x^{(k+1)}) \geq f(x^{(k)}) + \nabla^T f(x)(x^{(k+1)} - x^{(k)})$

$= f(x^{(k)}) + \nabla^T f(x) \Delta x^{(k)} \cdot t$ (convexity of f)

且 $f(x^{(k+1)}) < f(x^{(k)})$ (descent method)

$\Rightarrow$ 有 $\underline{\nabla^T f(x) \Delta x^{(k)} < 0}$

→ $\Delta x^{(k)}$ is a descent direction

① Line search 找 t

1. Exact line search.

$t = \underset{s \geq 0}{\operatorname{argmin}} f(x + s\Delta x)$

2. Backtracking Line Search (to reduce f "enough")

Start with $t=1$. ($\alpha \in (0, \frac{1}{2})$, $\beta \in (0,1)$)

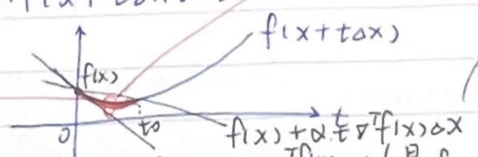
while ~~f(x+Δt)~~ $f(x+t\Delta x) > f(x) + \alpha t \nabla f^T(x) \Delta x$, $t := \beta t$

for fixed x. and fixed Δx: $f(x+t\Delta x)$ is a line in f with single varible t.

$\because f(x)$ is convex

$\therefore f(x+t\Delta x)$ is convex on t

Exact line search result



该区域内 $f(x+t\Delta x) \leq f(x) + \alpha t \nabla^T f(x) \Delta x$

$f(x) + t \cdot \nabla^T f(x) \Delta x$ (是 $f(x+t\Delta x)$ 切线 是 $f(x+t\Delta x)$ 的一阶Taylor展开)

$f(x+t\Delta x) \geq f(x) + t \cdot \nabla^T f(x) \Delta x$

由于 $\nabla^T f(x) \Delta x < 0$ 且 $\alpha \in (0, \frac{1}{2})$

所以 $f(x) + \alpha t \nabla^T f(x) \Delta x \geq f(x) + t \nabla^T f(x) \Delta x$

$t \to 0$: $f(x) \approx$ RHS $\leq$ LHS 达到stopping 要求 ⇒ Backtracking最终必会收敛

$\Rightarrow$ if $t \in (0, t_0)$ 则 satisfy stopping creteria

$\Rightarrow$ finally. $t=1$ or $t \in (\beta t_0, t_0]$

or $t \geq \min\{1, \beta t_0\}$

② 找 Δx.

1. Gradient Descent Method.

$\Delta x = -\nabla f(x)$, $x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x)$

Convergence analysis:

① Exact Line Search

$x^+ = x - t\nabla f(x)$

$f(x^+) = f(x - t\nabla f(x)) \leq f(x) - t\|\nabla f(x)\|_2^2 - \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$

由于 Exact line search 应找到t使得 $f(x^+)$ 最小

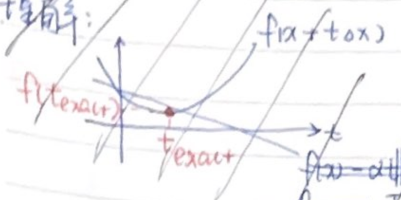
for RHS $= f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$ 最小. when $t = \frac{1}{M}$

$t = \frac{1}{M}$: RHS $= f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$

$\Rightarrow f(t_{exact}) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$

前面也得到了 $f(y) \le f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$ 的结论

理解:



$\cancel{f(x) - \alpha\|\nabla f(x)\|_2^2}$

$f(x) + \nabla f(x)(t\Delta x)$

$= f(x) - t\nabla^T f(x)(\nabla f(x))$

$= f(x) - t\|\nabla f(x)\|_2^2$

$\Rightarrow f(x^+) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$

$f(x^+) - p^* \le f(x) - p^* - \frac{1}{2M}\|\nabla f(x)\|_2^2$

由于 $\|\nabla f(x)\|_2^2 \ge 2m(f(x) - p^*)$

所以 $f(x^+) - p^* \le (1 - \frac{m}{M})(f(x) - p^*)$

$\Rightarrow f(x^{(k)}) - p^* \le (1 - \frac{m}{M})^k (f(x^{(0)}) - p^*)$

$f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$

$c = 1 - \frac{m}{M}$ for exact line search

log(error of kth iteration) $\le k\log(1 - \frac{m}{M})$ + log(error of initial point)

constant

lies below a line on log-linear

log-linear

linear convergence

stopping criteria: $f(x^{(k)}) - p \le \epsilon$

stop if $(1 - \frac{m}{M})^k (f(x^{(0)}) - p^*) \le \epsilon$

$\Rightarrow k \ge \frac{\log((f(x^0) - p^*)/\epsilon)}{\log(1/c)}$ $\quad c = 1 - \frac{m}{M}$

$\Rightarrow$ we have $f(x^{(k)}) - p^* \le \epsilon$ after at most $\frac{\log((f(x^0) - p^*)/\epsilon)}{\log(1/c)}$ iterations

for large $\frac{M}{m}$ $\cancel{\frac{1}{c} = \frac{1}{1-\frac{m}{M}} =}$

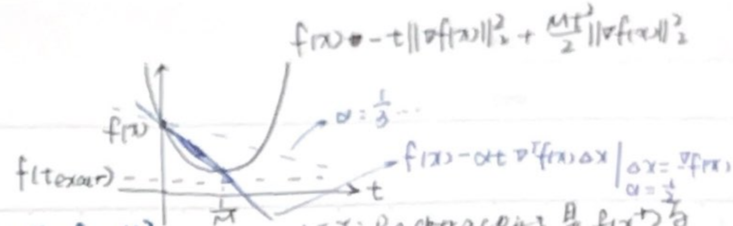
由于 ~~log~~ $-\log x \to -x + 1$ if $x \to 1$

即 $-\log(1 - x) \to x$ if $x \to 0$

$\Rightarrow -\log(1 - \frac{m}{M}) \to \frac{m}{M}$ if $\frac{m}{M} \to 0$

$\Rightarrow$ ~~if $\frac{M}{m}$~~ the bound on the # of iterations increases approximately linearly increase with $\frac{M}{m}$

同理 for backtracking line search.



要保证 $f(t_{back}) \le f(x) - \alpha t\|\nabla f(x)\|_2^2$

$\because f(x^+) \le f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2$

可证 $t \in [0, \frac{1}{M}]$ 时 RHS $\le f(x) - \alpha t\|\nabla f(x)\|_2^2$ (if $\alpha < \frac{1}{2}$)

KEY: Backtracking 是 $f(x^+)$ 与 $f(x) + \alpha t\nabla^T f(x)\Delta x$ (t 的一次函数) 的关系 但由 strongly convex 得到的是 t 的二次函数关系，因此要找二者之间的联系

$\therefore$ Backtracking terminates if $\begin{cases} t=1 \rightarrow f(x^+) \le f(x) - \alpha\|\nabla f(x)\|_2^2 \\ \text{or} \\ t \ge \frac{\beta}{M} \rightarrow f(x^+) \le f(x) - \frac{\alpha\beta}{M}\|\nabla f(x)\|_2^2 \end{cases}$

$\Rightarrow f(x^+) \le f(x) - \min\{\alpha, \frac{\alpha\beta}{M}\}\|\nabla f(x)\|_2^2$

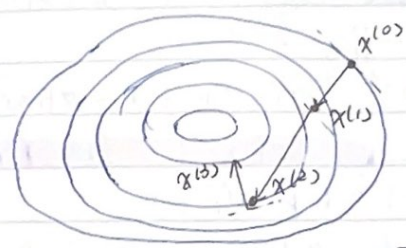
$f(x^+) - p^* \le f(x) - p^* - \min\{\alpha, \frac{\alpha\beta}{M}\}\|\nabla f(x)\|_2^2$

$\le (1 - \min\{2m\alpha, 2\alpha\beta\frac{m}{M}\})(f(x^{(0)}) - p^*)$

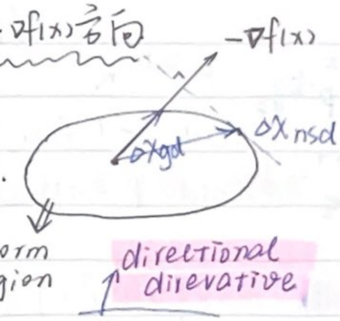
$\Rightarrow f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$

$c = 1 - \min\{2m\alpha, 2\alpha\beta\frac{m}{M}\}$

$\rightarrow$ Gradient Descent works well if conditional # is small



理解 $-\nabla f(x)$ 方向



理解.

directional derivative

使 $f(x+v)$ 的 linear approximation (Taylor 一次展开) 最小化

2. Steepest ~~Grad~~ Descent Method.

linear approximation of f: $f(x+v) \approx \hat{f}(x+v) = f(x) + \nabla f(x)^T \cdot v$

normalized steepest descent direction: $\Delta x_{nsd} = \arg\min\{\nabla f(x)^T v \mid \|v\| \le 1\}$

或 $= \arg\min\{\nabla f(x)^T v \mid \|v\| = 1\}$

Δx_{nsd} is the direction of unit ball of $\|\cdot\|$ that extends farthest in the direction $-\nabla f(x)$

Any Norm

Steepest descent direction: $\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$

$\rightarrow \|\cdot\|_*$ is dual norm (a scalar)

$\|a\|_* = \sup\{x^T a \mid \|x\| = 1\}$ 或 $= \sup\{x^T a \mid \|x\| \le 1\}$

$\nabla f(x)^T \Delta x_{sd} = \nabla f(x)^T \cdot \|\nabla f(x)\|_* \cdot \Delta x_{nsd} = -\|\nabla f(x)\|_*^2$

$\hookrightarrow$ 理解: $\Delta x_{nsd} = \arg\min\{\nabla f(x)^T v \mid \|v\|=1\}$, $\|\nabla f(x)\|_* = \max\{\nabla f(x)^T x \mid \|x\|=1\}$

$\Rightarrow$ ~~$\|\nabla f(x)\|_* \Delta x_{nsd} = \min\{\nabla f(x)$~~ $\nabla f(x)^T \cdot \Delta x_{nsd} = \min\{\nabla f(x)^T v \mid \|v\|=1\}$

$\Rightarrow \nabla f(x)^T \|\nabla f(x)\|_* \Delta x_{nsd} = \|\nabla f(x)\|_* \cdot \min\{\nabla f(x)^T v \mid \|v\|=1\} = -\|\nabla f(x)\|_* \max\{\nabla f(x)^T v \mid \|v\|=1\}$

$= -\|\nabla f(x)\|_*^2$

1) if $\|\cdot\|$ is Euclidean Norm $\rightarrow \Delta x_{sd} = -\nabla f(x)$

$\rightarrow$ Steepest Descent = Gradient Descent

2) Quadratic Norm $\|z\|_P = (z^T P z)^{1/2} = \|P^{1/2} z\|_2$, $P \in S^n_{++}$

$\Delta x_{nsd} = -(\nabla f(x)^T P^{-1} \nabla f(x))^{-1/2} P^{-1} \nabla f(x)$

$\Delta x_{sd} = -P^{-1} \nabla f(x)$ 核心 $\|x\|_P = \|y\|_2$, if 令 $P^{\frac{1}{2}}x$

推导: $\Delta x_{nsd} = \operatorname{argmin}\{\nabla f(x)^T x \mid \|P^{\frac{1}{2}} x\|_2 = 1\}$

let $y = P^{\frac{1}{2}} x$ then $x = P^{-\frac{1}{2}} y$

$\Rightarrow \Delta x_{nsd} = \operatorname{argmin}\{\nabla f(x)^T P^{-\frac{1}{2}} y \mid \|y\|_2 = 1\}$

$\Rightarrow \Delta y_{nsd} = -\dfrac{\nabla f(x)^T P^{-\frac{1}{2}}}{\|\nabla f(x)^T P^{-\frac{1}{2}}\|_2} = -\dfrac{P^{-\frac{1}{2}} \nabla f(x)}{\|\nabla f(x)^T P^{-\frac{1}{2}}\|_2}$

for the Denominator: $\|\nabla f(x)^T P^{-\frac{1}{2}}\|_2 = \cancel{\sqrt{(\nabla f(x)^T P^{-\frac{1}{2}})^T (\nabla f(x)^T P^{-\frac{1}{2}})}}$

$P \in S^n_{++}$. $P^{-\frac{1}{2}} \in S^n_{++}$ $\leftarrow \|P^{-\frac{1}{2}} \nabla f(x)\|_2$

$\Rightarrow (P^{-\frac{1}{2}})^T = P^{-\frac{1}{2}}$ $= \sqrt{\nabla f(x)^T P^{-\frac{1}{2}} P^{-\frac{1}{2}} \nabla f(x)}$

$\|\nabla f(x)\|_* = \|\nabla f(x)^T P^{-\frac{1}{2}}\|_2 = (\nabla f(x)^T P^{-1} \nabla f(x))^{\frac{1}{2}}$

a scalar $= (\nabla f(x)^T P^{-1} \nabla f(x))^{\frac{1}{2}} = \|\nabla f(x)\|_{P^{-1}}$

$\Rightarrow \Delta y_{nsd} = -(\nabla f(x)^T P^{-1} \nabla f(x))^{-\frac{1}{2}} (P^{-\frac{1}{2}} \nabla f(x))$

$\Rightarrow \Delta x_{nsd} = P^{-\frac{1}{2}} \Delta y_{nsd} = -(\nabla f(x)^T P^{-1} \nabla f(x))^{-\frac{1}{2}} (P^{-1}) \nabla f(x)$

$\|\nabla f(x)\|_* = \max\{\nabla f(x)^T x \mid \|x\|_P = 1\} = \max\{\nabla f(x)^T P^{-\frac{1}{2}} y \mid \|y\|_2 = 1\}$

$= \|\nabla f(x)^T P^{-\frac{1}{2}}\|_2$

$\Rightarrow \Delta x_{sd} = \|\nabla f(x)\|_* \, \Delta x_{nsd} = \|\nabla f(x)\|_* \cdot P^{-\frac{1}{2}} \Delta y_{nsd}$

$= \|\nabla f(x)^T P^{-\frac{1}{2}}\|_2 \cdot P^{-\frac{1}{2}} \cdot (-1) \dfrac{P^{-\frac{1}{2}} \nabla f(x)}{\|\nabla f(x)^T P^{-\frac{1}{2}}\|_2}$

$= -P^{-1} \nabla f(x)$

$\rightarrow$ the steepest descent method in $\|\cdot\|_P$ can be thought of as the gradient method applied to the perform after the change of coordinates $y = P^{\frac{1}{2}} x$

同样 Steepest Descent Method 也为 linear convergence

即: 可写成 $f(x^{(k)}) - p^* \le c^k (f(x^{(0)}) - p^*)$

c: constant.

$\rightarrow$ Steepest Descent works well if the transformed problem has moderate conditional #

3. Newton's Method.

$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$

→ 联系 $\Delta x_{sd} = -P^{-1}\nabla f(x)$ ⇒ newton's method 也是 steepest descent method with $\|\cdot\|$ is $\|\cdot\|_{\nabla^2 f(x)}$ 即 $P = \nabla^2 f(x)$

理解①: $\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$ 使 $\hat{f}(x+v)$ 的 Taylor 二次展开最小化

second-order Taylor approximation: $\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$

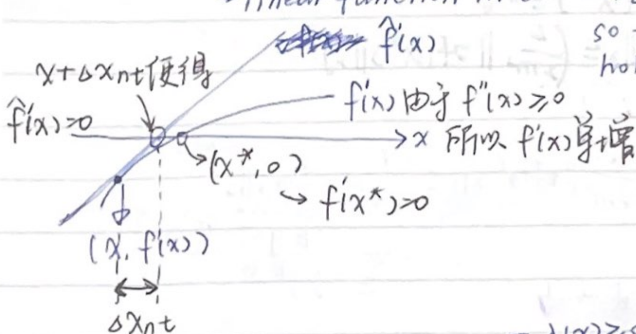
⇒ if $f(x)$ is quadratic (or linear) then $f(x) = \hat{f}(x)$ ⇒ $\Delta x_{nt} + x$ is the exact minimizer of f

理解②: stopping criteria (optimality condition): $\nabla f(x^*) = 0$

原因: we want to search a direction of v so that $\nabla f(x+v) = 0$

$\nabla_v \hat{f}(x) = \nabla f(x) + \nabla^2 f(x) v = 0 \rightarrow v = -\nabla^2 f(x)^{-1} \nabla f(x) = \Delta x_{nt}$

↳ linear function in v. ↳ $\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$ must be added to x so that the linearized optimality condition holds.



Newton Decrement: $\lambda(x) = (\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x))^{\frac{1}{2}}$

$\lambda(x) \geq 0$. ∵ $\nabla^2 f(x) \in S_{+}^{n}$ → $\nabla^2 f(x)^{-1} \in S_{+}^{n}$

⇒ $\forall x$, $x^T(\nabla^2 f(x)^{-1}) x \geq 0$

& $x^T(\nabla^2 f(x)^{-1}) x = 0$ iff $x = 0$

⇒ if ~~$\lambda(x)$~~ $\lambda(x^k) = 0$

it means $\nabla f(x^k) = 0$

⇒ x^k is optimal

理解①: $\lambda(x)$ 描述 the difference between $f(x)$ and the minimum of its quadratic approximation

$f(x) - \inf_{\Delta x} \hat{f}(x+\Delta x) = -\nabla f(x)^T \Delta x_{nt} - \frac{1}{2}\Delta x_{nt}^T \nabla^2 f(x)^{-1} \Delta x_{nt}$

由于 Δx_{nt} 使 $\hat{f}(x+\Delta x)$ 最小化 (quadratic)

$\hat{f}(x+\Delta x) = f(x) + \nabla f(x)^T \Delta x + \frac{1}{2}\Delta x^T \nabla^2 f(x) \Delta x$

将 $\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$ 代入. ⇒ RHS $= \frac{1}{2}\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = \frac{1}{2}\lambda(x)^2$

⇒ $f(x) - \inf_{\Delta x} \hat{f}(x+\Delta x) = \frac{1}{2}\lambda(x)^2$

if $\lambda(x)$ 小 ⇒ $f(x) \to \hat{f}(x+\Delta x)$ ⇒ $f(x) \to p^*$ ⇒ $\lambda(x)$ can act as a stopping ~~critia~~ criteria.

理解②: $\lambda(x) = \|\Delta x_{nt}\|_{\nabla^2 f(x)}$

$\|\Delta x_{nt}\|_{\nabla^2 f(x)} = (-\nabla^2 f(x)^{-1}\nabla f(x))^T \nabla^2 f(x) (-\nabla^2 f(x)^{-1}\nabla f(x))$

$= \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) = \lambda(x)$

$\nabla f(x)^T \Delta x_{nt}$ (directional ~~diff~~ derivative) 将 Δx_{nt} 代入, 可得

理解③: directional direvative ~~of~~ in the newton's direction $\nabla f(x)^T \Delta x_{nt} = -\lambda(x)^2$

convergence analysis

· Assume f is strongly convex $MI \succeq \nabla^2 f(x) \succeq mI$ for all $x \in S$

由 ~~Lipsht~~ Lipschitz continuous on S 可得 $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le L \| x-y\|_2$

if f is a quadratic function, $\nabla^2 f(x) = \nabla^2 f(y) =$ a constant

LHS=0 ⇒ $L=0$ for all quadratic function

⇒ L measures how well f can be approximated by a quadratic model.

⇒ L measures the performance of newton's method.

(newton's method works well if L is small)

Damped Newton Phase ⇒ in this phase # of iteration $\le \frac{f(x^{(0)}) - p^*}{\gamma}$

Exist constants $\eta \in (0, \frac{m^2}{L})$, $\gamma > 0$

① if $\|\nabla f(x)\|_2 \ge \eta$, then $f(x^{k+1}) - f(x^k) \le -\gamma$

② if $\|\nabla f(x)\|_2 < \eta$, then $\frac{L}{2m^2}\|\nabla f(x^{k+1})\|_2 \le \left(\frac{L}{2m^2}\|\nabla f(x^k)\|_2\right)^2$

Pure Newton Phase $(t=1)$

if $\|\nabla f(x)^k\|_2 < \eta$ then $\|\nabla f(x^{k+1})\|_2 < \eta$

$\frac{L}{2m^2}\|\nabla f(x^k)\|_2 \le \left(\frac{L}{2m^2}\|\nabla f(x^l)\|_2\right)^{2^{k-l}}$

$\le \eta \in (0, \frac{m^2}{L})$

$\le (\frac{1}{2})^{2^{k-l}}$

由于 $f(x) - p^* \le \frac{1}{2m}\|\nabla f(x)\|_2^2$

∴ $f(x^k) - p^* \le \frac{1}{2m}\|\nabla f(x^k)\|_2^2 \le \frac{2m^3}{L^2}(\frac{1}{2})^{2^{k-l+1}}$

⇒ $\log$(error of kth iteration) $\le 2^{k-l+1}\log(\frac{1}{2}) + \log(\frac{2m^3}{L^2})$

not a linear bound versus k.

(not linear convergence)

→ Quadratic convergence

→ the convergence is extremely fast once the second condition is satisfied

⇒ in this phase. # of iteration $\le \log_2\log_2(\frac{\epsilon_0}{\epsilon})$

$\epsilon_0 = \frac{2m^3}{L^2}$. $f(x^{(k)}) - p^* \le \epsilon$ (stopping criteria)

⇒ overall. # of iterations until $f(x) - p^* \le \epsilon$ is bounded above by

$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2\log_2(\frac{\epsilon_0}{\epsilon})$ → constant

→ for different ϵ, $\log_2\log_2(\frac{\epsilon_0}{\epsilon})$ 值变化不大

可用6估计其大小

⇒ # of iteration $\approx \frac{f(x^{(0)}) - p^*}{\gamma} + 6$

(3) self-concordance

前面对收敛性的分析依赖于 m. M 与 L. 而它们 are almost unknown in practice

引入 self-concordant function. 其收敛性 bound 与 m. M. L 无关且 affinely invariant

Definition①: a convex function $f: \mathbb{R}\to\mathbb{R}$ is self-concordant if $|f'''(x)| \le 2 f''(x)^{3/2}$ for all $x \in \mathrm{dom} f$ → for function on R

Standard self-concordant inequality

①

但事实上 coefficient 2 并不重要. 所以若 a function $f(x)$ satisfies $|f'''(x)| \le k f''(x)^{\frac{3}{2}}$ for $k>0$. Then 可得到 $|\tilde{f}'''(x)| \le 2 \tilde{f}''(x)^{\frac{3}{2}}$ by constructing $\tilde{f}(x) = \frac{k^2}{4} f(x)$ 即 $\tilde{f}(x)$ is a positive-scaled version of $f(x)$

~~由于 positive scaling 可 perserve self-concordance (explain later~~ ⟨only when a>1⟩

~~所以 对 $f(x)$ is self-concordant <=> $\tilde{f}(x)$ self-concordant~~

~~satisfies $|f'''(x)| \le k f''(x)^{\frac{3}{2}}$ → satisfies standard self-concordant inequality~~

~~→ 系数是否为2并无所谓~~

WRONG 并非所有k都不影响self-concordant

由上面的分析可知 if $k \le 2$ then $f(x)$ is self-concordant <=> $\tilde{f}(x)$ is self-concordant 但 Textbook 上说 any other positive constant k could be used. 但我不知怎么证明

② ~~f(x) is self-concordant. then f(ax+b) is self-concordant~~

let $\hat{f}(x) = f(ax+b)$ then $\hat{f}$ is self-concordant <=> f is self-concordant

变换坐标位不改变 self-concordance

=> self-concordance is affine invariant

Definition②: for function on $\mathbb{R}^n$, a function $f: \mathbb{R}^n \to \mathbb{R}$ is self-concordant if its self-concordance holds along every line in its domain, i.e., if the function $\tilde{f}(t) = f(x + tv)$ is a self-concordant function of t for all $x \in \mathrm{dom} f$ and for all v. ($\in \mathbb{R}^n$)

区别: $x \in \mathbb{R}$. $a, b \in \mathbb{R}$.

Properties: (easy to prove)

① scaling: if f is self-concordant and $a \ge 1$ then af is self-concordant

② Addition: if f_1, f_2 are self-concordant. then $f_1 + f_2$ is self-concordant

③ composition with affine function: if $f: \mathbb{R}^n \to \mathbb{R}$ is self-concordant and $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$, then $f(Ax+b)$ is self-concordant (affine invariant)

④ composition with logarithm: let $g: R \to R$ be a convex function, with $\mathrm{dom}\, g = R_{++}$ and $\left|g'''(x)\right| \le 3\frac{g''(x)}{x}$ for all x. then $f(x) = -\log(-g(x)) - \log x$ is self-concordant on $\{x \mid g(x)<0,\ x>0\}$

注意二次函数一定满足 $|g'''(x)| \le 3\frac{g''(x)}{x}$ ⇒ if $g_1(x)$ satisfies $|g_1'''(x)| \le 3\frac{g_1''(x)}{x}$

则 $g_2(x) = g_1(x) + ax^2 + bx + c$ 也 satisfies $|g_2'''(x)| \le 3\frac{g_2''(x)}{x}$

例: show $f(x,y) = -\log y - \log(\log y - x)$ on $\{(x,y) \mid e^x < y\}$ is self-concordant

学习思路 ① restrict to a line

② use composition with logarithm → 写成 $-\log(\ ) - \log(\ )$ 形式

① restrict to a line

let $x = \hat{x} + tv$. $y = \hat{y} + tw$. where $\hat{x}, \hat{y}, v, w$ are fixed. → 常用设计

⇒ $f(x,y) = -\log y - \log(\log y - x)$ → 已经是 $-\log(\ ) - \log(y)$ 形式, 思路是将

⇒ $f(\hat{x}+tv, \hat{y}+tw) = -\log(\hat{y}+tw) - \log(\log(\hat{y}+tw) - \hat{x} - tv)$ 第一个 log 对象写成 $g(y)$ 且 $g(y)$ 满足 $|g'''(y)| \le 3\frac{g''(y)}{y}$

② use composition with logarithm

(1) case I: if $w=0$

$$RHS = \underbrace{-\log(\hat{y})}_{\text{constant}} - \log(\log\hat{y} - \hat{x} - tv)$$

$= -\log(at+b) + c$ where a, b, c are constants

because $-\log(x)$ is self-concordant. so $-\log(at+b)+c$ is also self-~

⇒ $f(x,y)$ is self-concordant on every line

⇒ $f(x,y)$ is self-concordant

注意推导流程

(2) case II: if $w \ne 0$ ⇒ $t = \frac{y-\hat{y}}{w}$ 常用设计

⇒ $RHS = -\log(y) - \log(\log y - x - \frac{y-\hat{y}}{w} \cdot v)$

$= -\log(y) - \log(\log y - a'y - b')$

let $g(y) = \log y - a'y - b'$

we have $g'(y) = \frac{1}{y} - a'$ $g''(y) = \frac{-1}{y^2}$ $g'''(y) = \frac{2}{y^3}$

~~$|g'''(y)| \le 3\frac{g''(y)}{y}$~~ $|g'''(y)| = \frac{2}{y^3}$ $3\frac{g''(y)}{y} = 3\frac{1}{y^3}$

⇒ $g(y)$ satisfies $|g'''(y)| \le 3\frac{g''(y)}{y}$

⇒ $f(x,y) = -\log(g(y)) - \log y$ is self-concordant convergence analysis

exist constants $\eta \in (0, \frac{1}{4}]$, $\gamma > 0$, suh that

① if $\lambda(x) > \eta$, then $f(x^{k+1}) - f(x^k) \le -\gamma$

② if $\lambda(x) \le \eta$, then $2\lambda(x^{k+1}) \le \left(2\lambda(x^k)\right)^2$

$\Rightarrow$ # of iteration is bounded by $\dfrac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2 (\frac{1}{\epsilon})$