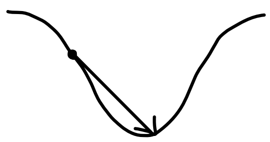
Algorithms 实质上都是在解 KKT 条件. 讨论几种典型的算法.

无约束优化 / 有约束优化

强调: ① 所有优化算法都是迭代算法.

② $x^{k+1} = x^k + \partial^k d^k$ (k 时刻的方向, 与 x 的维数相同) ∂^k: 步长、标量 举例: 下山法

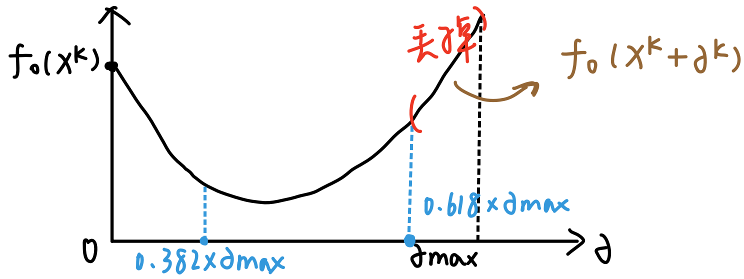


$\partial^k = \underset{\partial \geq 0}{\arg\min}\, f_0(x^k + \partial d^k)$ 一维凸问题.

③ 解步长的方法: line search

一般讨论的是下降的算法.

(1) 迭代算法: 黄金分割法 (优选法)



折半查找的思想

0.618 是收敛速度最快的.

(2) Amijo Rule (还可以的步长)

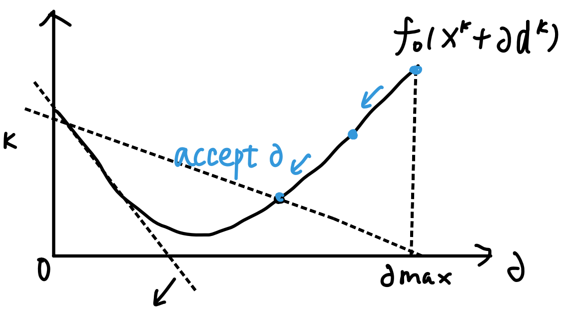
if $f_0(x^k+\partial d^k) > f(x^k) + \gamma \partial \nabla f_0^T(x^k) d^k$ 取 (0, 0.5)

设成 0.5, 收敛性的证明会很漂亮

then, $\partial \leftarrow \partial \cdot \beta$ (0, 1)

otherwise, stop.

Backtracking



Taylor: $f_0(x^k) + \partial \nabla f_0(x^k) d^k$

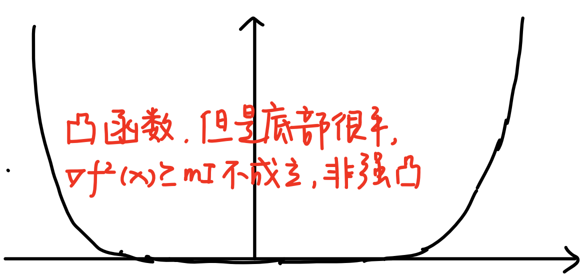
min $f(x)$, $f(x)$ 一阶可微, 最优性条件 $\nabla f(x) = 0$, 假设能找到 x, $\nabla f(x) \approx 0$.

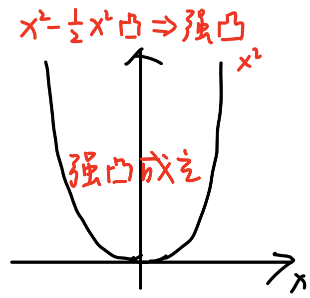
$x \to x^*$? $f(x) \to f(x^*)$?

假设 $f(x)$ 二阶可微, 且有强凸性.

$\exists\, m > 0,\ \forall x \in \mathrm{dom} f,\ \nabla f^2(x) \succeq mI$

(直观解释: 每一点都不能特别平)





强凸的好处: $\forall x,y \in dom f$, $f(y) \ge f(x) + \nabla f^T(x)(y-x) + \frac{1}{2}m\|y-x\|_2^2$

I. $\nabla f(x) \to 0$, $f(x) \to f(x^*)$?

x给定, $f(x) + \nabla f^T(x)(y-x) + \frac{1}{2}m\|y-x\|_2^2$ 是y的凸函数. 仿射项+二次项

一阶偏导 $\nabla f(x) + m(\hat{y}-x) = 0$, $\tilde{y} = x - \frac{\nabla f(x)}{m}$, 代入上式.

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

进而 $f(y) \ge f(x) + \nabla f^T(x)(y-x) + \frac{1}{2}m\|y-x\|_2^2 \geqslant f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$ for $\forall x, \forall y$

取y为最优解, $p^* \geqslant f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$, i.e. $p^* + \frac{1}{2m}\|\nabla f(x)\|_2^2 \ge f(x) \geqslant p^*$

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

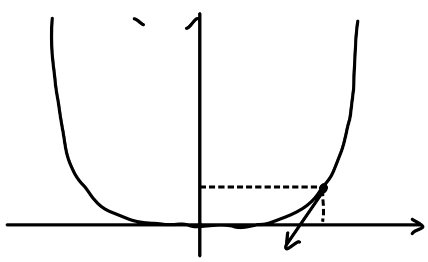
II. $\nabla f(x) \to 0$, $x \to x^*$?

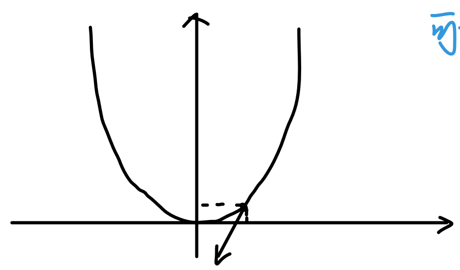
$p^* = f(x^*) \geq f(x) + \nabla f^T(x)(x^*-x) + \frac{m}{2}\|x^*-x\|_2^2$ 没有二范数, 则想办法改写.

回忆 Cachy-Schwarz inequality, $\langle a, b\rangle + \|a\|\|b\| \geq 0$

then $f(x) \geqslant p^* \geqslant f(x) - \|\nabla f(x)\|_2\|x^*-x\|_2 + \frac{m}{2}\|x^*-x\|$

$-\|\nabla f(x)\|_2\|x^*-x\|_2 + \frac{m}{2}\|x^*-x\|_2 \leqslant 0$, i.e. $\|x^*-x\|_2 \leqslant \frac{2}{m}\|\nabla f(x)\|_2$





可量化的损失

(最优值很近, 但是最优解很远)

(最优值和最优解都更近了)

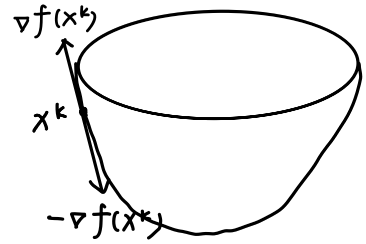
凸性不要太强, 规定上界: $\exists M > 0$, $\forall x \in dom f$, $\nabla^2 f(x) \leqslant MI$ 曲线不能太陡

从正定的角度规定上界和下界

$\forall x, y \in dom f$, $f(y) \leqslant f(x) + \nabla f^T(x)(y-x) + \frac{M}{2}\|y-x\|_2^2$

(+强凸性) 重要结论: $p^* \leqslant f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$

1. 梯度下降法.(负梯度方向, $d^k=-\nabla f(x^k)$)



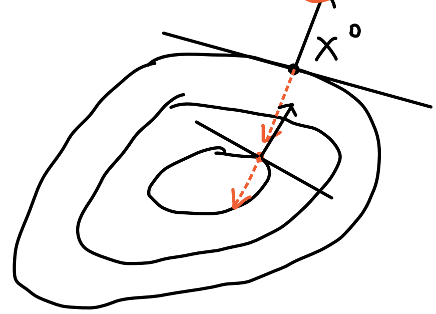
Repeat $\alpha^k=\arg\min f(x^k+\alpha d^k)$ exact or inexact

$\alpha_{max}\geq\alpha\geq 0$

$x^{k+1}=x^k+\alpha^k d^k$

Until Convergence

梯度：函数值增加方向 $\nabla f(x^0)$



• 分析算法收敛性: $\forall x\in dom f$, $MI\succeq\nabla^2 f(x)\succeq mI$

1) exact line search

$$\hat f(\alpha)=f(x^k+\alpha d^k)=f'(\underbrace{x^k-\alpha\nabla f(x^k)}_{x^{k+1}})$$

$$\overset{y}{f(x^{k+1})}\leq \overset{x}{f(x^k)}+\nabla f^T(x^k)(-\alpha\nabla f(x^k))+\frac{M}{2}\|-\alpha\nabla f(x^k)\|_2^2$$

$$\Leftrightarrow \hat f(\alpha)\leq f(x^k)-\alpha\|\nabla f(x^k)\|_2^2+\frac{M\alpha^2}{2}\|\nabla f(x^k)\|_2^2$$

一阶偏导: $-\|\nabla f(x^k)\|_2^2+M\alpha\|\nabla f(x^k)\|_2^2=0$, $\Rightarrow\alpha=\frac{1}{m}$

$$\min_{\alpha}\hat f(\alpha)\leq f(x^k)-\frac{1}{M}\|\nabla f(x^k)\|_2^2+\frac{1}{2M}\|\nabla f(x^k)\|_2^2$$

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

另外一个不等式: $p^*\geq f(x^k)-\frac{1}{2M}\|\nabla f(x^k)\|_2^2$

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

$$-\frac{1}{2m}\|\nabla f(x^k)\|_2^2+f(x^{k+1})-p^*\leq 0$$

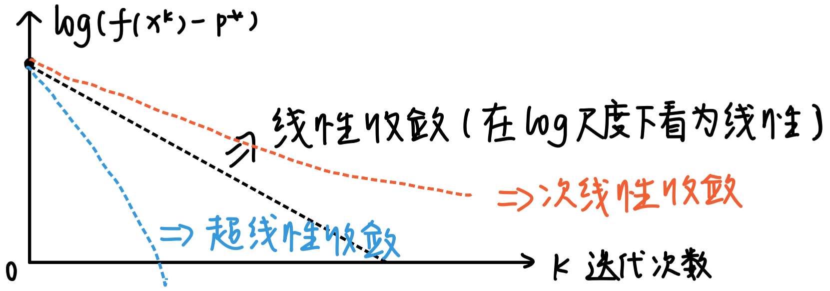
进而, $M(f(x^{k+1})-p^*)+m(f(x^k)-p^*)\leq M(f(x^k)-p^*)$

$\Leftrightarrow (f(x^{k+1})-p^*)\leq(1-\frac{m}{M})(f(x^k)-p^*)$ 可说明每一步离最优解越来越近.

given ε, when $\left\|\frac{f(x^{k+2})-p^*}{f(x^k)-p^*}\right\|\leq\varepsilon$

i.e. $\left\|\frac{f(x^{k+1})-p^*}{f(x^k)-p^*}\right\|\leq\left\|1-\frac{m}{M}\right\|\Rightarrow(1-\frac{m}{M})^\tau=\varepsilon\Rightarrow\tau=\frac{\log\varepsilon}{\log(1-\frac{m}{M})}$ ⇒用 $\frac{m}{M}$ 近似

用收敛曲线解释:



2) Inexact Line Search (Amijo Rule)

当 $0<\alpha<\frac{1}{M}$ 时, 迭代必然停止.

$-\gamma\alpha\|\nabla f^T(x^k)\|_2^2$

$\tilde{f}(\alpha)=f(x^{k+1})\le f(x^k)+\gamma\alpha_0\nabla f^T(x^k)d^k$ 时接受 α.

当 $0\le\alpha\le\frac{1}{M}$, 必有 $(-\alpha+\frac{M\alpha^2}{2})\le-\frac{\alpha}{2}$

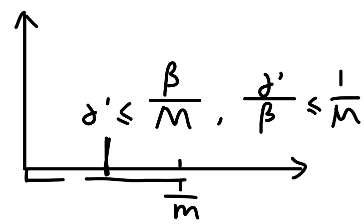
$$\tilde{f}(\alpha)=f(x^{k+1})\le f(x^k)-\alpha\|\nabla f(x^k)\|_2^2+\frac{M\alpha^2}{2}\|\nabla f(x^k)\|_2^2$$

$$\le f(x^k)-\frac{\alpha}{2}\|\nabla f(x^k)\|_2^2$$

$$\le f(x^k)-\gamma\alpha\|\nabla f(x^k)\|_2^2$$

不可能特别小

$\alpha_{inexact}=\alpha_{max}$ 或 $\ge\frac{\beta}{M}$

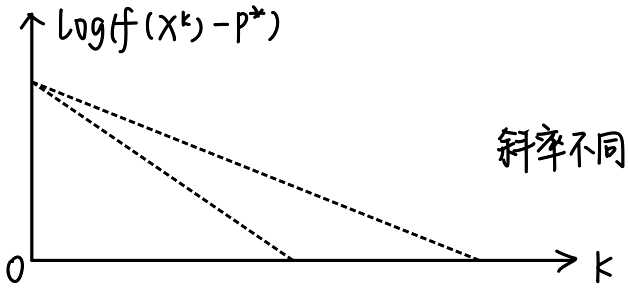


$$f(x^{k+1})=\tilde{f}(\alpha_{exact})\le f(x^k)-\frac{1}{2M}\|\nabla f(x^k)\|_2^2$$

$$f(x^{k+1})\le f(x)-\min\left\{\gamma\alpha_{max},\frac{\gamma\beta}{M}\right\}\|\nabla f(x)\|_2^2$$

$$\frac{f(x^{k+1})-p^*}{f(x^k)-p^*}\le 1-\min\left\{2m\gamma\alpha_{max},\frac{2m\gamma\beta}{M}\right\}\in[0,1]$$

∴仍然是线性收敛.



斜率不同, 收敛速度不同.

例: $f(x)=x^TPx$, $P\in S_+^n$

$\frac{M}{m}$

最大特征值 最小特征值

$\nabla^2 f(x)=P$, $MI\succeq P\succeq mI$

2病态矩阵 $\begin{pmatrix}100 & 0\\0 & 1\end{pmatrix}$ 梯度下降法收敛速度不会特别快

zig-zag: