Convex Optimization Theory and Applications

Topic 13 - Unconstrained Minimization

Li Li

Department of Automation
Tsinghua University

Fall, 2009-2021.

13.0. Outline

13.1. 下降算法与直线搜索

13.2. 梯度下降算法

13.3. 最速下降算法

13.4. Newton 方法

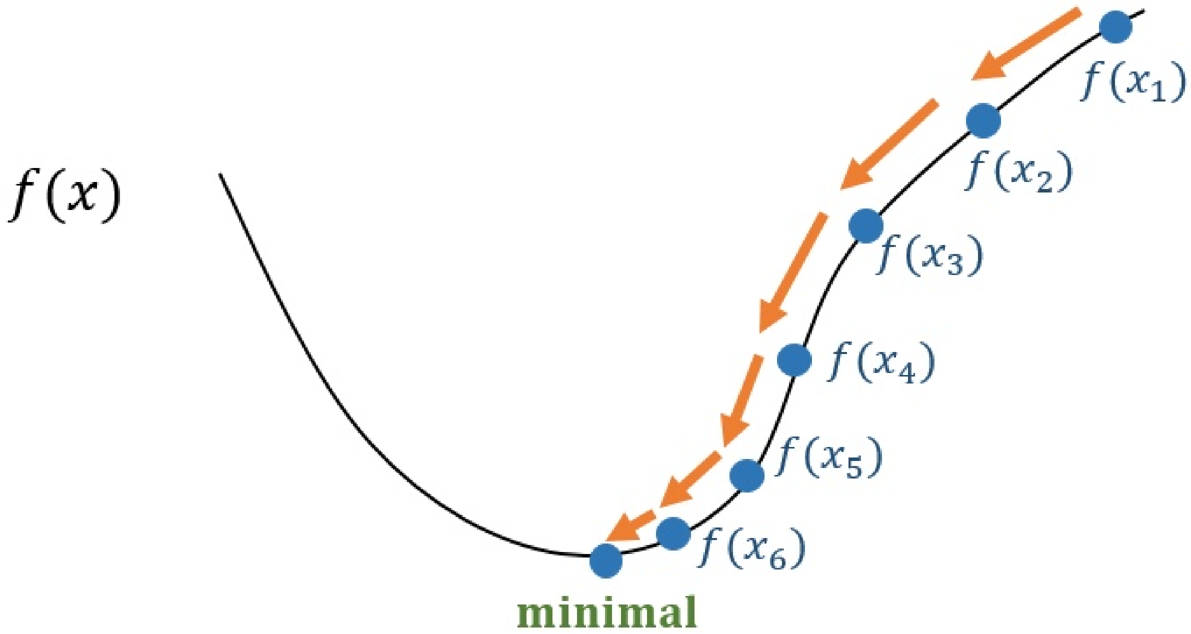
13.5. 收敛性分析

13.6. 自和谐性

13.1. 下降算法与直线搜索

无约束可微优化问题，可以直接利用梯度信息求解

- 如何计算搜索方向
- 如何计算前进步长



13.1. 下降算法与直线搜索

下降算法通用框架

Step 1）确定初始点 $x^0 \in \mathrm{dom}\, f$，令 $k=0$

Step 2）判断是否停止：如果 $\left\|\nabla f\left(x^k\right)\right\| \le \varepsilon$，停止

Step 3）计算 x^k 处的下降方向：确定 d^k 满足

$$\nabla f\left(x^k\right)^T d^k < 0$$

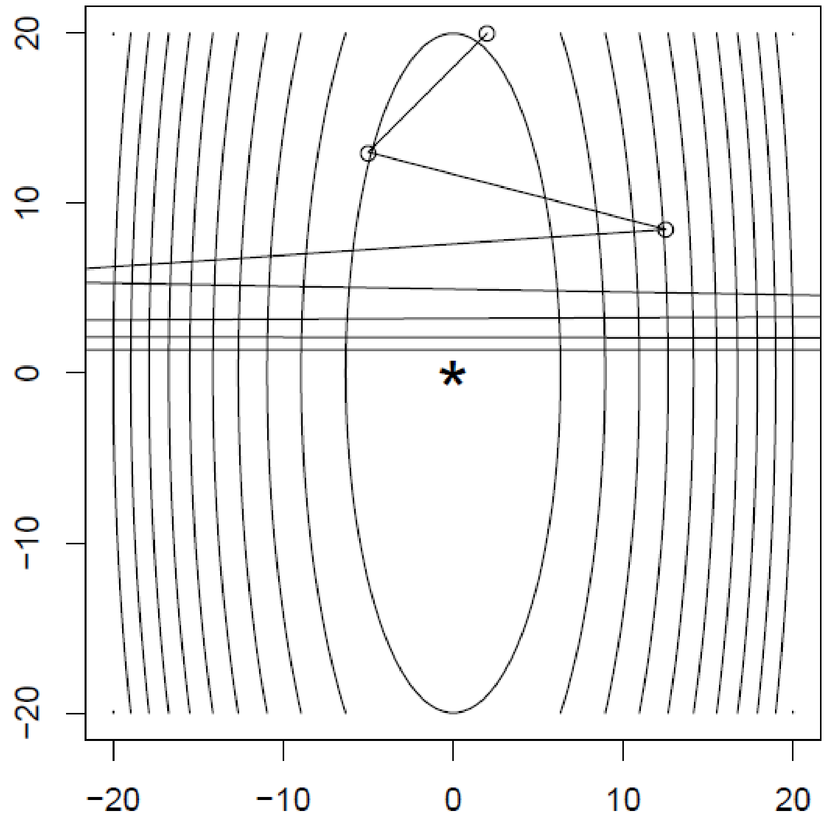
Step 4）确定步长（通常直线搜索）：若 $t^k > 0$ 满足

$$f\left(x^k + t^k d^k\right) < f\left(x^k\right)$$

令 $x^{k+1} = x^k + t^k d^k$， $k \Rightarrow k+1$， 回 Step 2）

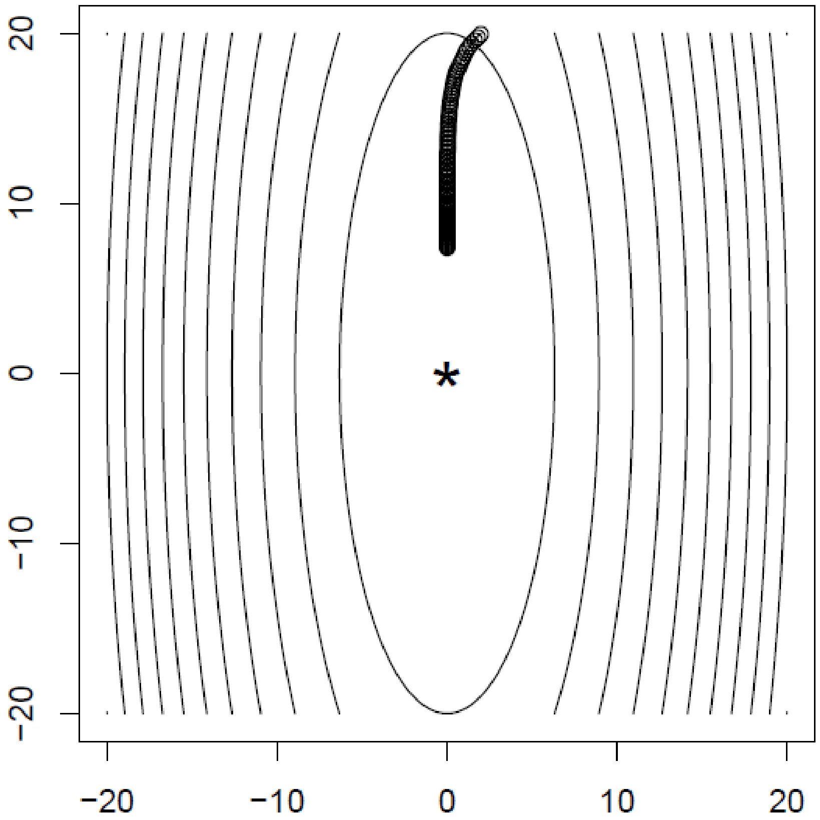
13.1. 下降算法与直线搜索

Simply take $t_k = t$ for all $k = 1, 2, 3, \ldots$, can diverge if t is too big. Consider $f(x) = (10x_1^2 + x_2^2)/2$, gradient descent after 8 steps:



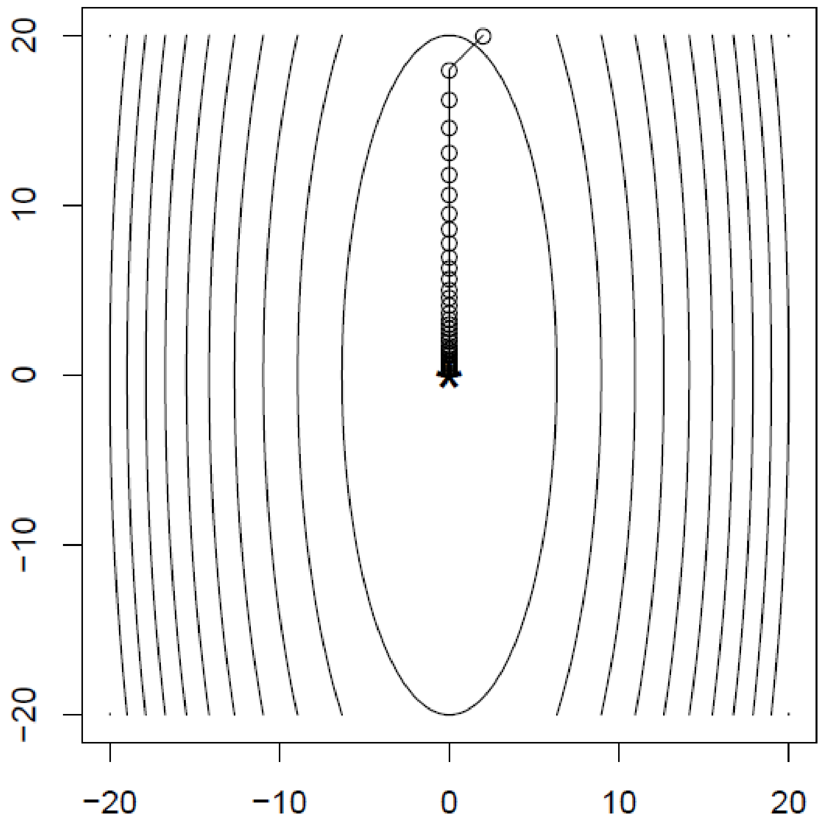
13.1. 下降算法与直线搜索

Can be slow if t is too small. Same example, gradient descent after 100 steps:



13.1. 下降算法与直线搜索

Converges nicely when t is "just right". Same example, 40 steps:



Convergence analysis later will give us a precise idea of "just right"

13.1. 下降算法与直线搜索

一般来说，当我们选定的前进方向，则可以通过精确搜索或者非精确搜索来决定合适的前进步长

精确搜索算法：求延选定方向出发，使得函数最小的步长

$$t^k = \arg\min_{t\geq 0} f\left(x^k + td^k\right)$$

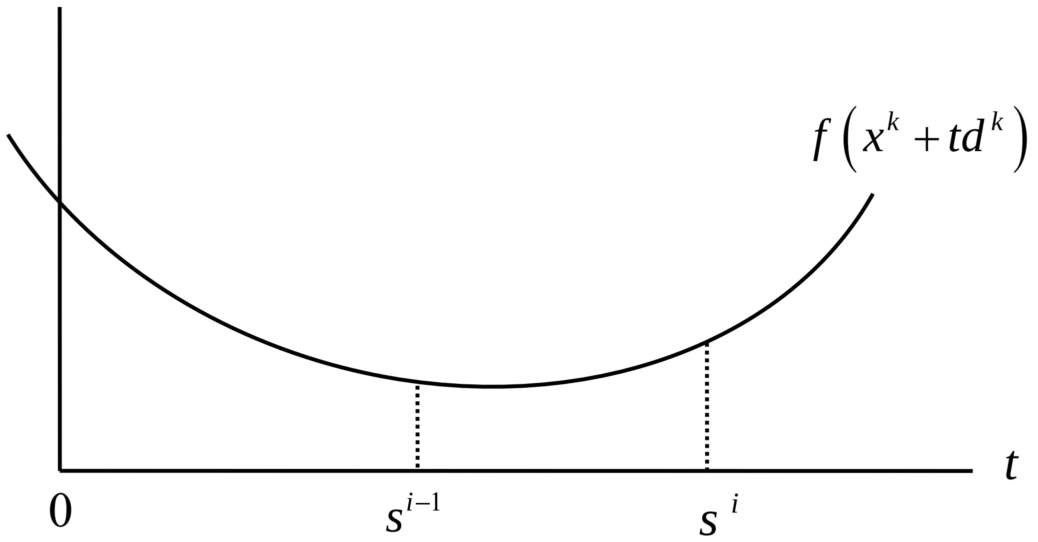
精确搜索算法是单变量优化问题。一般来说，精确搜索所化的时间代价小于求出搜索方向的时间代价

13.1. 下降算法与直线搜索

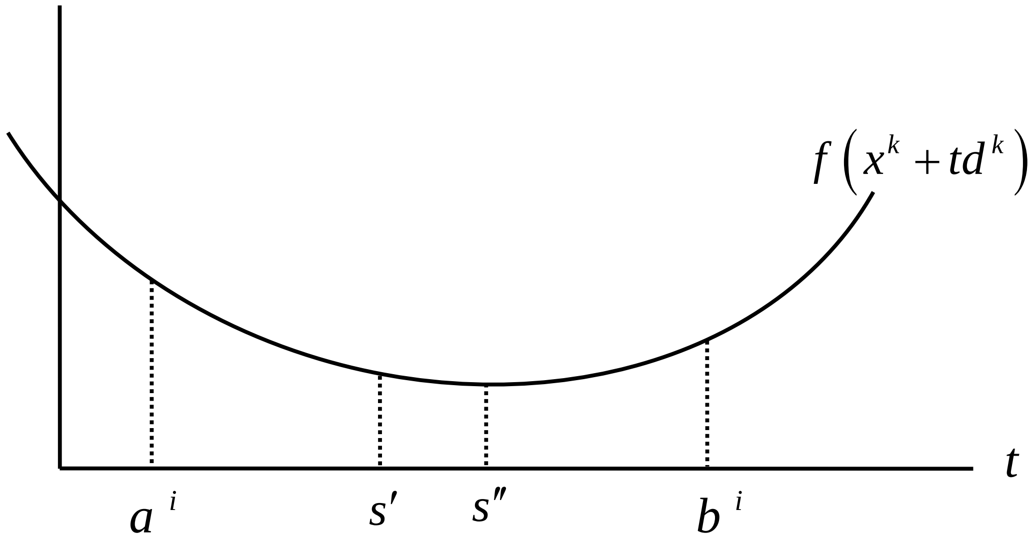
0.618 法（常用精确直线搜索方法）

1）确定单谷区间

选定$\gamma>1$，$\delta>0$，令$s^0=0$，$s^i=\gamma^{i-1}\delta, \forall i\geq 1$, 选取第一个满足$f\left(x^k+s^i d^k\right)>f\left(x^k+s^{i-1}d^k\right)$的$s^i$, 获得单谷区间$\left[0,s^i\right]$



2）确定满足 $f\left(x^k+t^k d^k\right)=\min\limits_{t>0} f\left(x^k+td^k\right)$ 的 t^k



单谷区间为 $\left[a^i,b^i\right]$ 时，取 $s'=a^i+(1-c)\left(b^i-a^i\right)$，$s''=a^i+c\left(b^i-a^i\right)$，其中 $c=\left(\sqrt{5}-1\right)/2\approx 0.618$ 是方程 $(1-c)/c=c/1$ 的正数解。如果 $f\left(x^k+s'd^k\right)>f\left(x^k+s''d^k\right)$，令 $a^{i+1}=s'$，$b^{i+1}=b^i$，用当前 s'' 替换 s'，再令 $s''=a^{i+1}+c\left(b^{i+1}-a^{i+1}\right)$；否则令 $a^{i+1}=a^i$，$b^{i+1}=s''$，用当前 s' 替换 s''，再令 $s'=a^{i+1}+(1-c)\left(b^{i+1}-a^{i+1}\right)$，直至单谷区间长度足够小。

13.1. 下降算法与直线搜索

回溯搜索法（适合凸优化的非精确直线搜索方法 backtracking line search starts with unit step size）

Step 1）选定 $\alpha \in (0, 0.5)$，$\beta \in (0,1)$

Step 2）每次迭代开始时，令 $t^0 = t_{initial}$，通常选 $t_{initial} = 1$

依次迭代令 $t^{k+1} = \beta t^k$，直到找到第一个满足下式（回溯搜索不等式）的 t^k

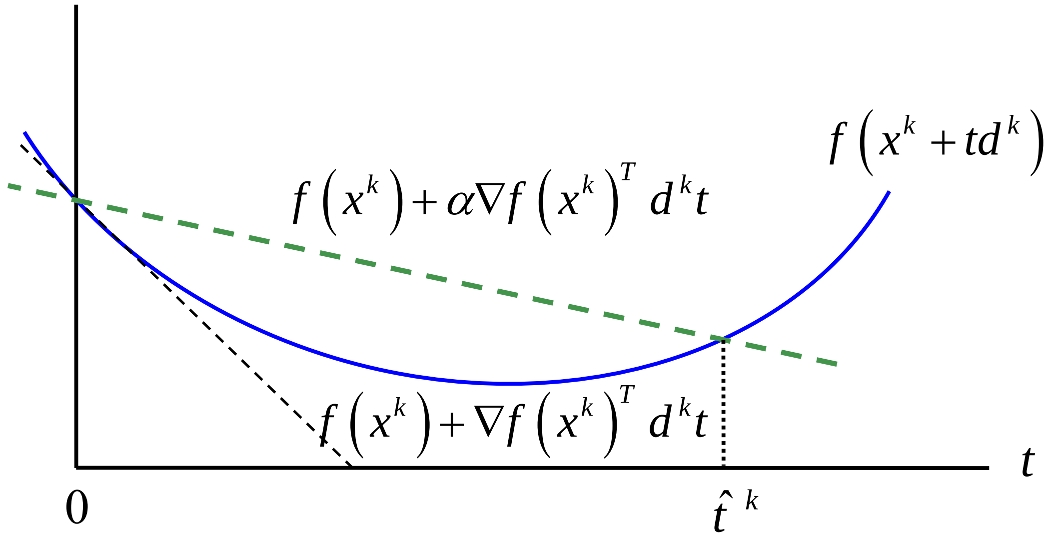
$$f\left(x^k + t^k d^k\right) < f\left(x^k\right) + \alpha t^k \nabla f\left(x^k\right)^T d^k$$

注意到 $\nabla f\left(x^k\right)^T d^k < 0$，我们对于足够小的 t^k，成立

$$f\left(x^k + t^k d^k\right) \approx f\left(x^k\right) + t^k \nabla f\left(x^k\right)^T d^k \le f\left(x^k\right) + \alpha t^k \nabla f\left(x^k\right)^T d^k$$

因此上述算法必然在有限步内停止

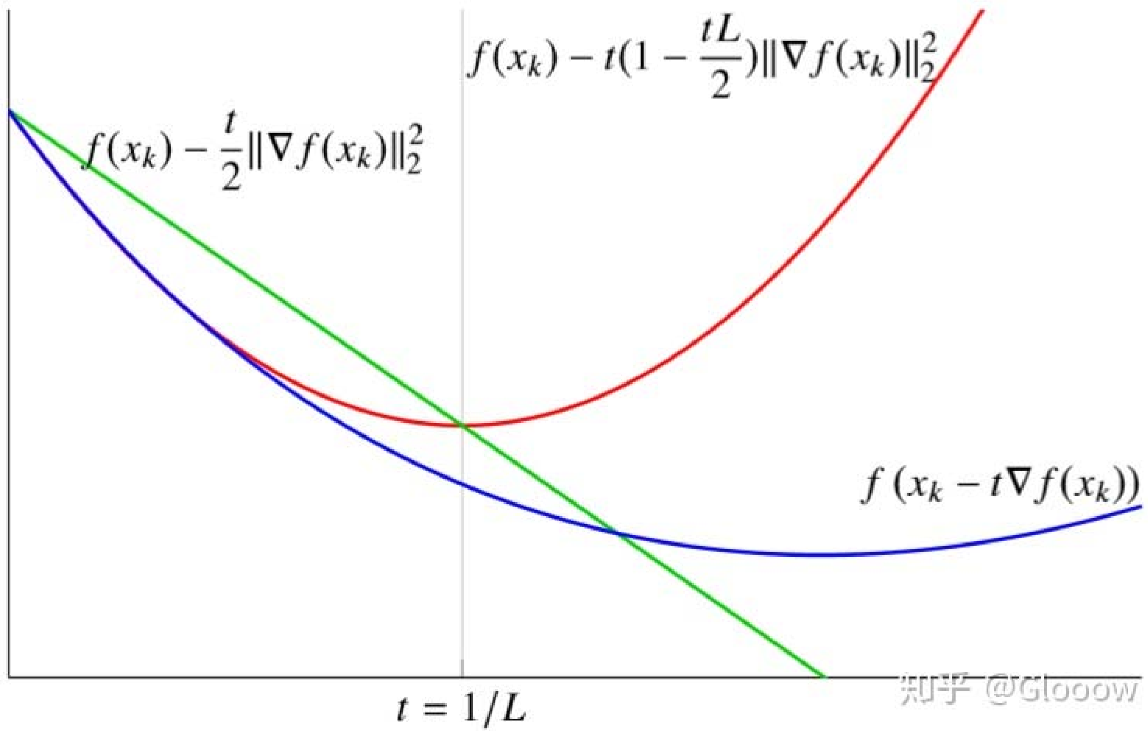
13.1. 下降算法与直线搜索



一开始t^k可能很大，表示梯度下降的步长过大，不能使函数值减小，那我们就减小步长，直到进入绿线与蓝线交点左侧这部分，我们就可以保证一定有$f\left(x^{k+1}\right)<f\left(x^k\right)$，这时就是我们要取的$t^k$

参数α会影响我们的搜索结果。α越大，则上图绿线的斜率越大，那么最终搜索到的满足回溯搜索不等式的t^k应该就越小。也即我们下降算法中采用的每一步的步长都会更小。

实际中往往取$\alpha = 0.5$。这实际是我们用二次上界曲线来近似待优化的函数，而二次上界的最小值点对应的步长就是$t = 1/L$（L为函数f的Lipschitz连续性常数），但由于我们是线搜索，实际得到的t^k一般会比这个值略小一点。



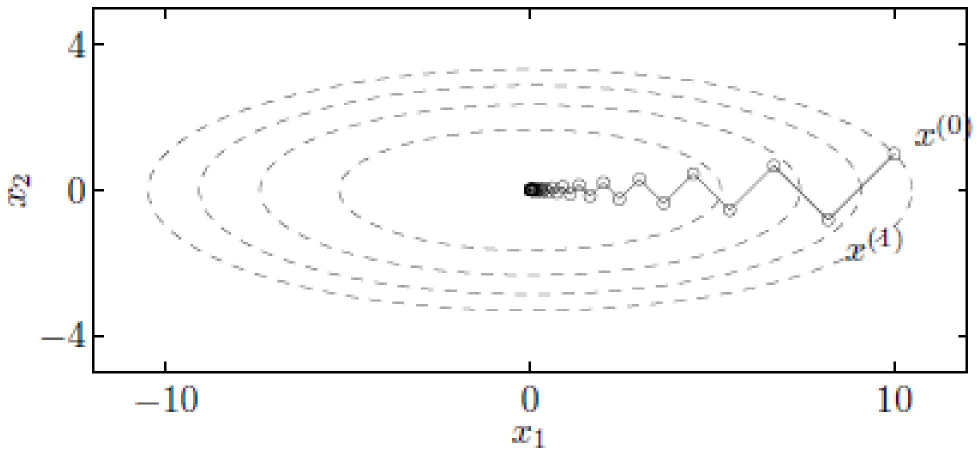
13.2. 梯度下降

例、两个变量的二次规划问题

$$f(x)=\frac{1}{2}\left(x_1^{\ 2}+10x_2^{\ 2}\right)$$

$$\lambda_{\min}\left(\nabla^2 f(z)\right)=1,\ \lambda_{\max}\left(\nabla^2 f(z)\right)=10 \quad \Rightarrow \quad \frac{m}{M}=0.1$$

采用精确直线搜索，迭代过程见下图，注意<u>相邻方向正交</u>



13.2. 梯度下降

精确搜索得到的点和搜索方向之间的关系

设 $\hat{X}'$ 是在 $\hat{X}$ 处沿下降方向 D 进行精确搜索所得到的点，即 $\hat{X}' = \hat{X} + \hat{t}D$，其中 $\hat{t}$ 是优化问题

$$\min_{t>0} f\left(\hat{X} + tD\right)$$

的最优解，应有 $\dfrac{df\left(\hat{X} + \hat{t}D\right)}{dt} = 0$

由 $\dfrac{df\left(\hat{X} + tD\right)}{dt} = \dfrac{\partial f\left(\hat{X} + tD\right)}{\partial X^T} \dfrac{d\left(\hat{X} + tD\right)}{dt} = \nabla^T f\left(\hat{X} + tD\right) D$

可得 $\qquad \nabla^T f\left(\hat{X}'\right) D = 0$

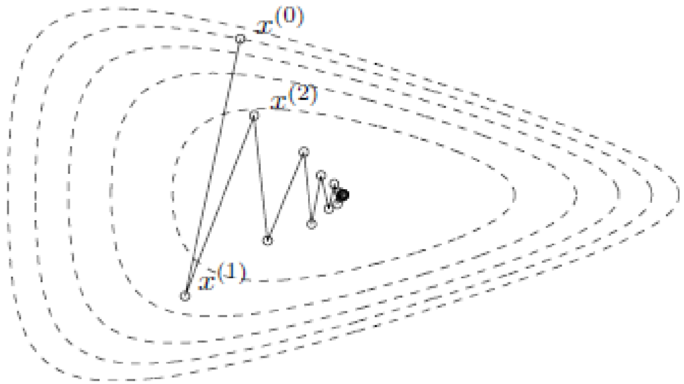
即<u>所得到的点的梯度和所采用的搜索方向垂直</u>

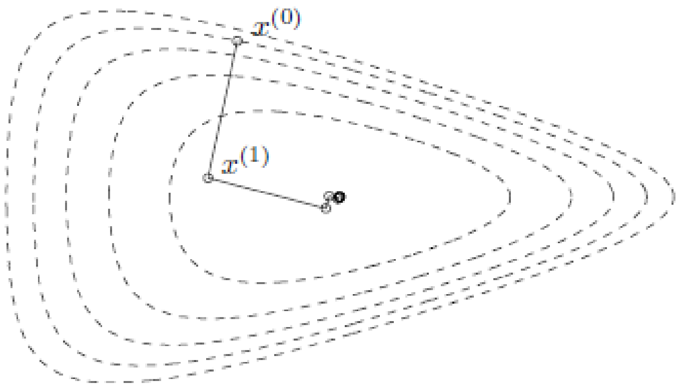
13.2. 梯度下降

例、两个变量的非二次规划问题

$$f\left(x\right)=e^{x_1+3x_2-0.1}+e^{x_1-3x_2-0.1}+e^{-x_1-0.1}$$

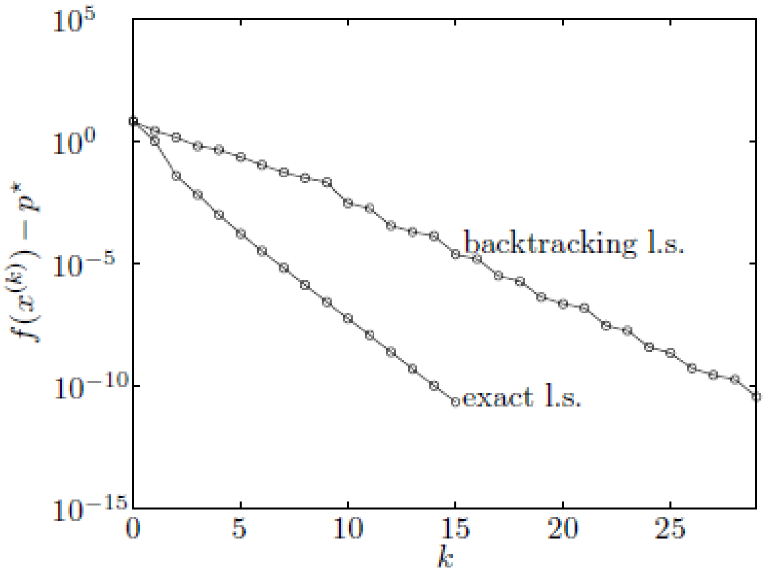
采用 $\alpha=0.1,\beta=0.7$ 的回溯直线搜索的迭代过程（下面左图）和采用精确直线搜索的迭代过程（下面右图）





13.2. 梯度下降

两种直线搜索的迭代误差随迭代次数改变的情况见下图

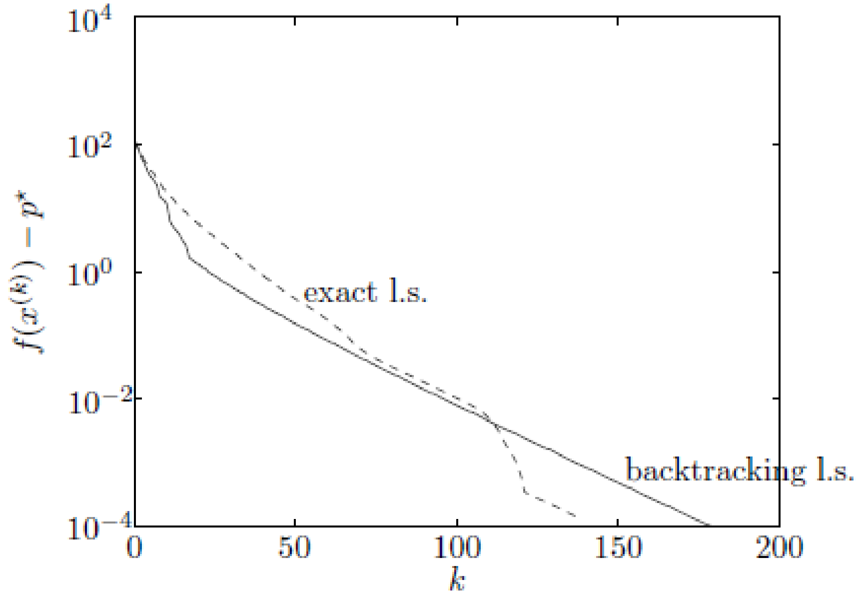


13.2. 梯度下降

例、高维变量的问题

$$f(x)=c^T x-\sum_{i=1}^{500}\log\left(b_i-a_i^T x\right),\quad x\in R^{100}$$

分别采用$\alpha=0.1, \beta=0.5$的回溯直线搜索和精确直线搜索，两种直线搜索的迭代误差随迭代次数改变的情况见下图



13.2. 梯度下降

关于梯度下降法的若干结论

1）误差 $f\left(x^{k}\right)-p^{*}$ 呈现为 k 的几何数列，即 γ^{k}，相似的线性收敛性（Nesterov 证明一阶方法的收敛速度存在极限）

2）回溯参数的选择对收敛速度有影响，精确直线搜索可能会改善收敛性，但都不会出现戏剧性的效果

3）收敛速度强烈依赖于二阶梯度矩阵的条件数，即最小最大特征根的比值

13.3. 最速下降算法

范数$\|\bullet\|$对应的（规范化的）最速下降方向d^k由下式确定

$$\nabla f\left(x^k\right)^T d^k = \min\left\{\nabla f\left(x^k\right)^T d \,\middle|\, \text{s.t. } \|d\| \le 1\right\} = -\max\left\{-\nabla f\left(x^k\right)^T d \,\middle|\, \text{s.t. } \|d\| \le 1\right\}$$

The selected direction is unit-norm step with most negative directional derivative

ℓ_1范数

$$d_i^k = \begin{cases} \operatorname{sgn}\left(-\dfrac{\partial f(x^k)}{\partial x_i}\right) & \text{if } \left|\dfrac{\partial f(x^k)}{\partial x_i}\right| = \left\|\nabla f(x^k)\right\|_\infty \\ 0 & \text{if } \left|\dfrac{\partial f(x^k)}{\partial x_i}\right| \ne \left\|\nabla f(x^k)\right\|_\infty \end{cases}$$（多解选一个）

$$\nabla f\left(x^k\right)^T d^k = -\left\|\nabla f\left(x^k\right)\right\|_\infty$$

13.3. 最速下降算法

ℓ_p 范数（$\|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$，$\forall\ p > 1$）

$$d_i^k = \operatorname{sgn}\left(-\frac{\partial f(x)}{\partial x_i}\right)\left|\frac{\partial f(x)}{\partial x_i}\right|^{q-1}\left(\|\nabla f(x)\|_q\right)^{-\frac{q}{p}}, \quad \forall i$$

$\nabla f\left(x^k\right)^T d^k = -\left\|\nabla f\left(x^k\right)\right\|_q$，　　其中 $q = \dfrac{p}{p-1}$（$p = \dfrac{q}{q-1}$）

ℓ_∞ 范数

$$d_i^k = \operatorname{sgn}\left(-\frac{\partial f(x)}{\partial x_i}\right), \quad \forall i$$

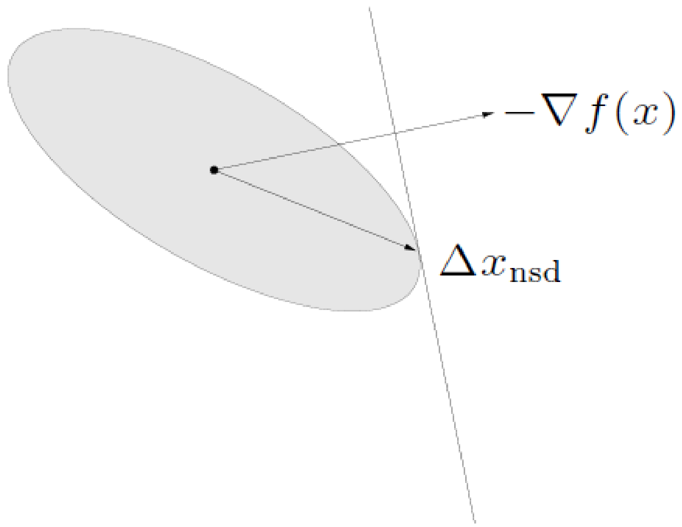
$$\nabla f\left(x^k\right)^T d^k = -\left\|\nabla f\left(x^k\right)\right\|_1$$

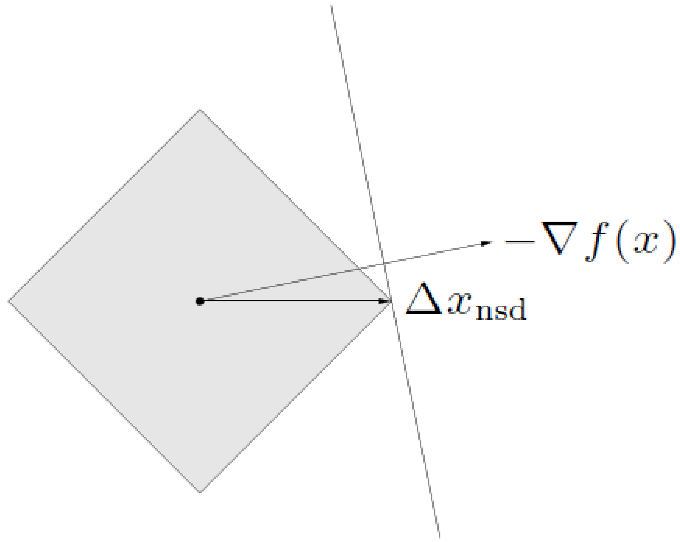
13.3. 最速下降算法

examples

- Euclidean norm: $\Delta x_{\mathrm{sd}} = -\nabla f(x)$
- quadratic norm $\|x\|_P = (x^T P x)^{1/2}$ ($P \in \mathbf{S}^n_{++}$): $\Delta x_{\mathrm{sd}} = -P^{-1}\nabla f(x)$
- ℓ_1-norm: $\Delta x_{\mathrm{sd}} = -(\partial f(x)/\partial x_i)e_i$, where $|\partial f(x)/\partial x_i| = \|\nabla f(x)\|_\infty$

unit balls and normalized steepest descent directions for a quadratic norm and the ℓ_1-norm:





13.3. 最速下降算法

例、两个变量的非二次规划问题

$$f\left(x\right)=e^{x_1+3x_2-0.1}+e^{x_1-3x_2-0.1}+e^{-x_1-0.1}$$

采用二次范数$\|x\|_P=x^TPx$的最速下降算法求解, 其中$P>0$

分别选择

$$P_1=\begin{pmatrix}2 & 0\\0 & 8\end{pmatrix} \quad \text{和} \quad P_1=\begin{pmatrix}8 & 0\\0 & 2\end{pmatrix}$$

同样采用 $\alpha = 0.1, \beta = 0.7$ 的回溯直线搜索，迭代过程见下图

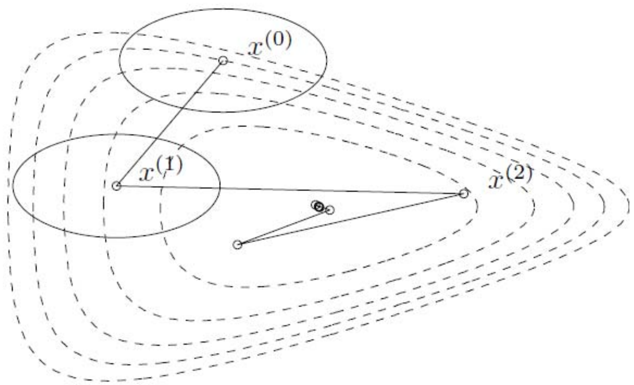


图 9.11　采用二次范数 $\|\cdot\|_{P_1}$ 的最速下降方法。椭圆表示球体 $\{x \mid \|x - x^{(k)}\|_{P_1} \leqslant 1\}$ 在 $x^{(0)}$ 和 $x^{(1)}$ 处的边界。

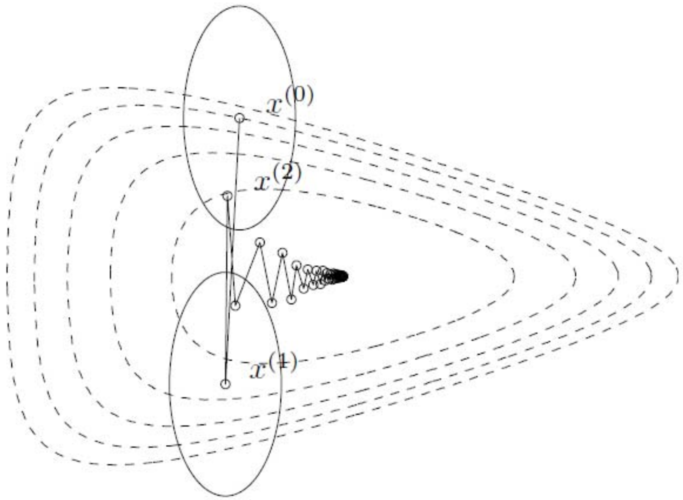


图 9.12　采用二次范数 $\|\cdot\|_{P_2}$ 的最速下降方法。

13.4. Newton 方法

牛顿方向 $d^k = -\nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right)$，显然是可行方向，因为

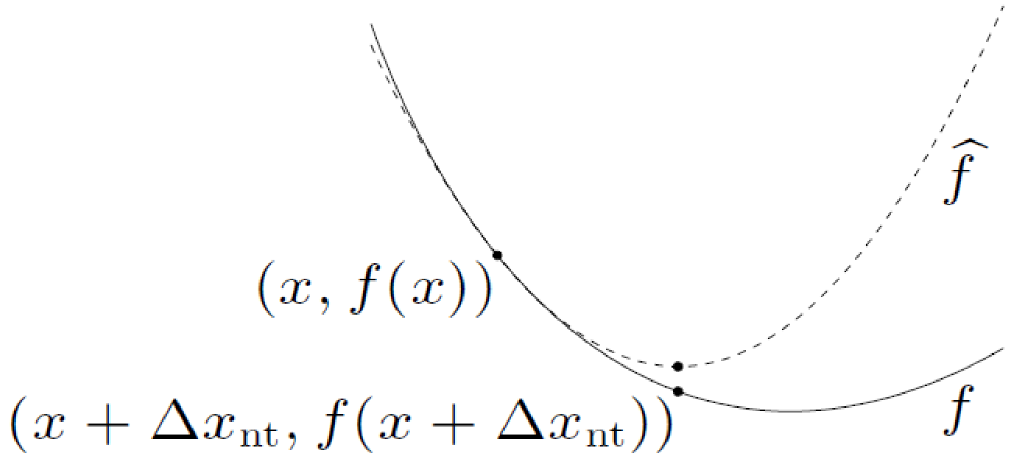
$$\nabla f\left(x^k\right) d^k = -\nabla f\left(x^k\right)^T \nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right) < 0$$

牛顿方法 $x^{k+1} = x^k + t^k d^k = x^k - \nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right)$，因为取 $t^k = 1$

对牛顿方向的第一种理解：目标函数的二阶近似

$$f(x+d) \approx f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d \triangleq \hat{f}(x+d)$$

$$\min_d \hat{f}\left(x^k + d\right) \quad \Rightarrow \quad d^k = -\nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right)$$

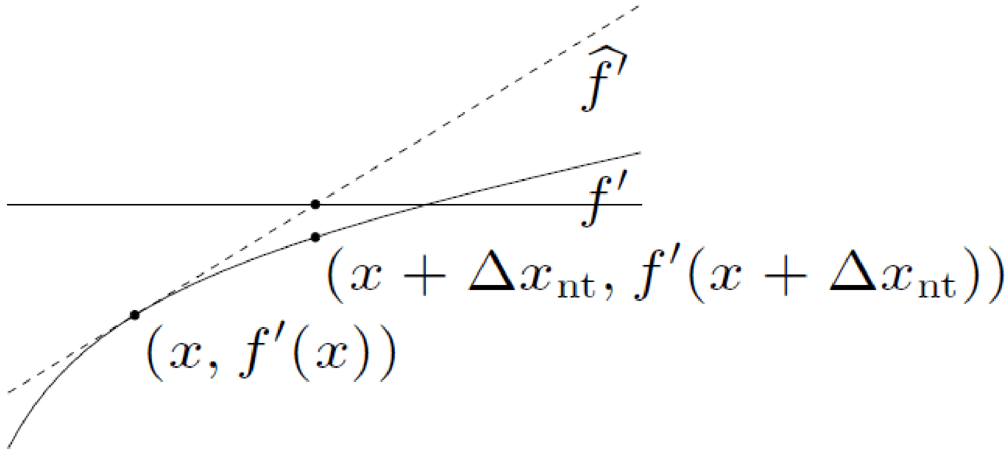


13.4. Newton 方法

对牛顿方向的第二种理解：最优性方程的一阶近似

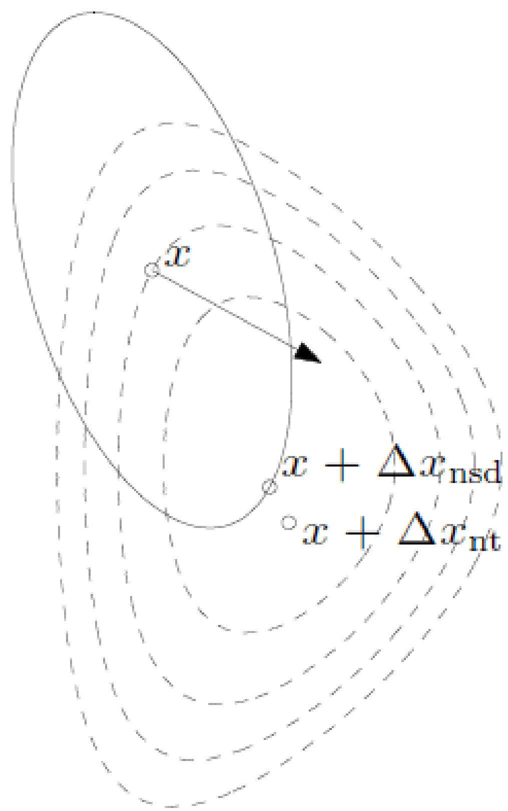
$$\nabla f(x+d) \approx \nabla f(x) + \nabla^2 f(x) d \triangleq \nabla \hat{f}(x+d)$$

$$\nabla \hat{f}(x+d) = 0 \quad \Rightarrow \quad d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$$



- Δx_{nt} is steepest descent direction at x in local Hessian norm

$$\|u\|_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}$$



dashed lines are contour lines of f; ellipse is $\{x + v \mid v^T \nabla^2 f(x) v = 1\}$

arrow shows $-\nabla f(x)$

13.4. Newton 方法

Newton decrement

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

a measure of the proximity of x to $x^\star$

properties

- gives an estimate of $f(x) - p^\star$, using quadratic approximation $\widehat{f}$:

$$f(x) - \inf_y \widehat{f}(y) = \frac{1}{2}\lambda(x)^2$$

- equal to the norm of the Newton step in the quadratic Hessian norm

$$\lambda(x) = \left(\Delta x_{\mathrm{nt}} \nabla^2 f(x) \Delta x_{\mathrm{nt}}\right)^{1/2}$$

- directional derivative in the Newton direction: $\nabla f(x)^T \Delta x_{\mathrm{nt}} = -\lambda(x)^2$
- affine invariant (unlike $\|\nabla f(x)\|_2$)

13.4. Newton 方法

Important property Newton's method: affine invariance. Given f, nonsingular $A \in \mathbb{R}^{n\times n}$. Let $x = Ay$, and $g(y) = f(Ay)$. Newton steps on g are

$$\begin{aligned} y^+ &= y - \left(\nabla^2 g(y)\right)^{-1} \nabla g(y) \\ &= y - \left(A^T \nabla^2 f(Ay) A\right)^{-1} A^T \nabla f(Ay) \\ &= y - A^{-1} \left(\nabla^2 f(Ay)\right)^{-1} \nabla f(Ay) \end{aligned}$$

Hence

$$Ay^+ = Ay - \left(\nabla^2 f(Ay)\right)^{-1} \nabla f(Ay)$$

i.e.,

$$x^+ = x - \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$$

So progress is independent of problem scaling. This is not true of gradient descent!

13.4. Newton 方法

给定起始点 $x \in \mathrm{dom} f$，$\epsilon > 0$

重复：

1.计算牛顿补偿和减量

$$\Delta x_{\mathrm{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x); \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$$

2.停止条件．若 $\lambda^2 / 2 \le \epsilon$ 则停止

3.直线搜索．通过回溯直线搜索选择步长 t

4.更新． $x := x + t\Delta x_{\mathrm{nt}}$

$t^k = 1$ 为 pure Newton method，如果用回溯搜索，则为 damped Newton method 阻尼 Newton 方法

13.4. Newton 方法

步长是通过回溯直线搜索(backtracking search)

参数$0<\alpha\le 1/2, 0<\beta<1$

在每次迭代中，由 t = 1 开始，当

$$f(x+tv) > f(x) + \alpha t \nabla f(x)^T v$$

满足时，缩小$t=\beta t$，否则进行牛顿法更新。

注意，此处：

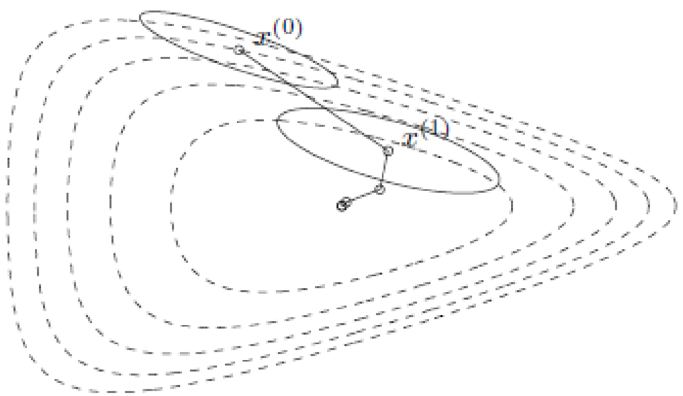
$$v = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x), \text{so } \nabla f(x)^T v = -\lambda^2(x)$$

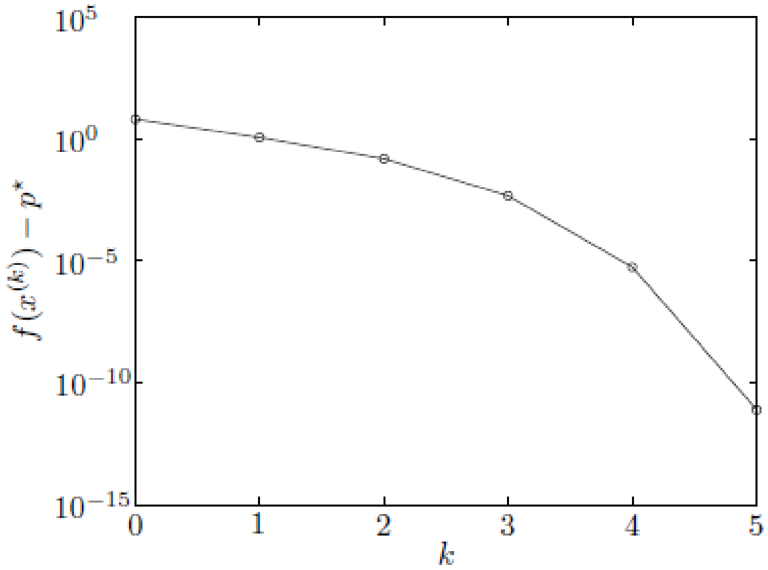
13.4. Newton 方法

例、两个变量的非二次规划问题

$$f(x)=e^{x_1+3x_2-0.1}+e^{x_1-3x_2-0.1}+e^{-x_1-0.1}$$

直线搜索参数 $\alpha=0.1,\ \beta=0.7$， 下图左边显示开始几次迭代点及椭圆 $\left\{x \,\middle|\, \left\|x-x^k\right\|_{\nabla^2 f\left(x^k\right)}=1\right\}$，右边显示优化误差与次数的关系



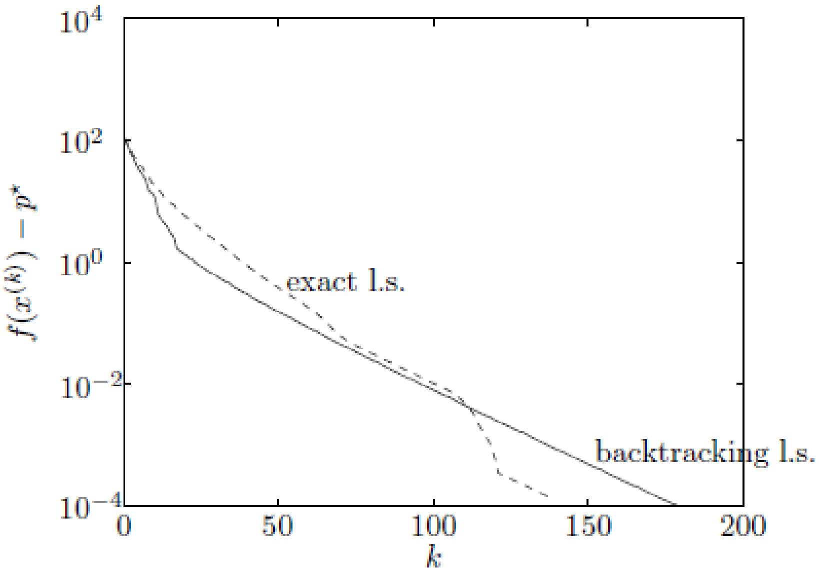


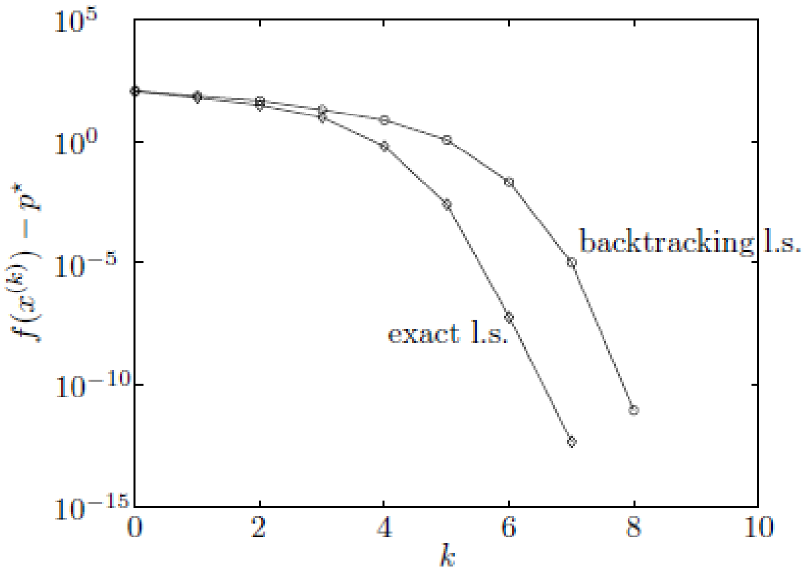
13.4. Newton 方法

例、高维变量的问题

$$f(x) = c^T x - \sum_{i=1}^{500} \log\left(b_i - a_i^T x\right), \quad x \in R^{100}$$

下图左右分别是梯度下降算法和牛顿方法的迭代过程，后者采用$\alpha = 0.01, \beta = 0.5$的回溯直线搜索和精确直线搜索



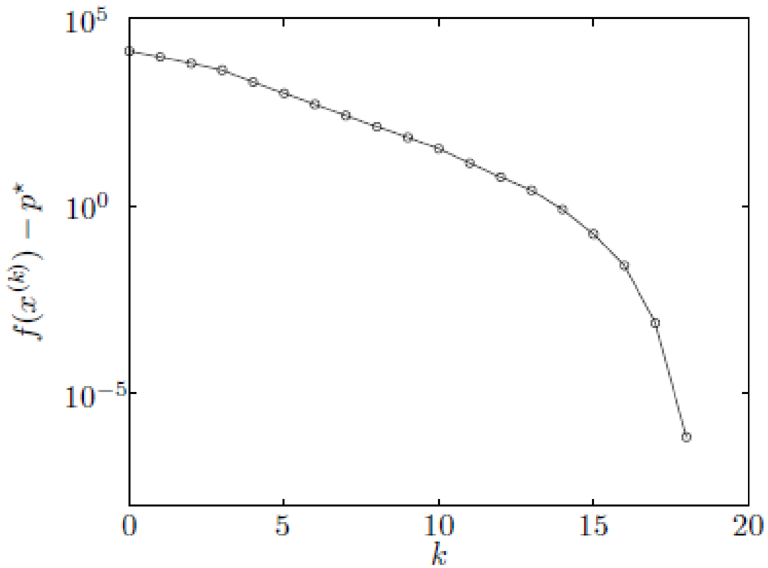


13.4. Newton 方法

例、大规模问题

$$f(x) = -\sum_{i=1}^{10000} \log\left(1 - x_i^2\right) - \sum_{i=1}^{1000} \log\left(b_i - a_i^T x\right), \quad x \in R^{10000}$$

采用 $\alpha = 0.01, \beta = 0.5$ 的回溯直线搜索，下图显示优化误差与次数的关系

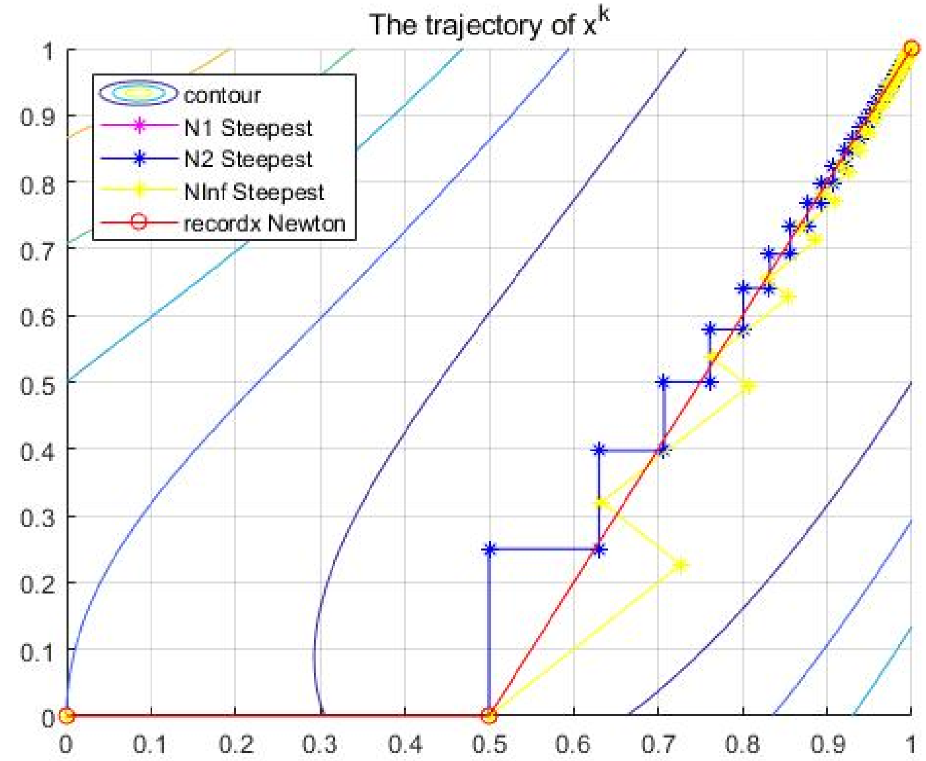


13.4. Newton 方法

例、牛顿法和最速下降法的比较

$$f\left(x\right)=\left(1-x_1\right)^2+2\left(x_2-x_1^{\ 2}\right)^2$$

初始点均为原点，直线搜索均采用精确搜索（0. 618 法）



13.4. Newton 方法

At a high-level:

- Memory: each iteration of Newton's method requires $O(n^2)$ storage (n*n Hessian); each gradient iteration requires O(n) storage (n-dimensional gradient)
- Computation: each Newton iteration requires $O(n^3)$ flops (solving a dense n*n linear system); each gradient iteration requires O(n) flops (scaling/adding n-dimensional vectors)
- Backtracking: backtracking line search has roughly the same cost, both use O(n) flops per inner backtracking step
- Conditioning: Newton's method is not affected by a problem's conditioning, but gradient descent can seriously degrade

主要计算量是确定 Newton 方向 $d^k = -\nabla^2 f\left(x^k\right)^{-1} \nabla f\left(x^k\right)$

13.4. Newton 方法

常见的简化计算方法

1）确定 Cholesky 分解 $\nabla^2 f\left(x^k\right) = LL^T$，其中 L 是下三角矩阵

2）前向代入解方程 $Ly = -\nabla f\left(x^k\right)$ 确定 y

3）后向代入解方程 $L^T d^k = y$ 确定 d^k

之后我们会讨论各种拟牛顿法 Quasi-Newton method

13.5. 收敛性分析

无约束（凸）优化问题： $\min f(x)$，其中 $f: R^n \mapsto R$ 是凸函数

基本假设与作用

1） f 是具有连续二阶导数的凸函数

$\Rightarrow$ $d \in R^n$ 是 $x \in R^n$ 的下降方向的充要条件： $\nabla f(x)^T d < 0$

$x \in R^n$ 是最优解的充要条件： $\nabla f(x) = 0$

2）已知 $x^0 \in \mathrm{dom}\, f$ ， 且满足 $S = \left\{ x \in \mathrm{dom}\, f \,\middle|\, f(x) \le f(x^0) \right\}$ 是闭集

$\Rightarrow$ 下降算法所产生的点列有属于 S 的极限点

3）强凸

4）Lipschitz 连续

13.5. 收敛性分析

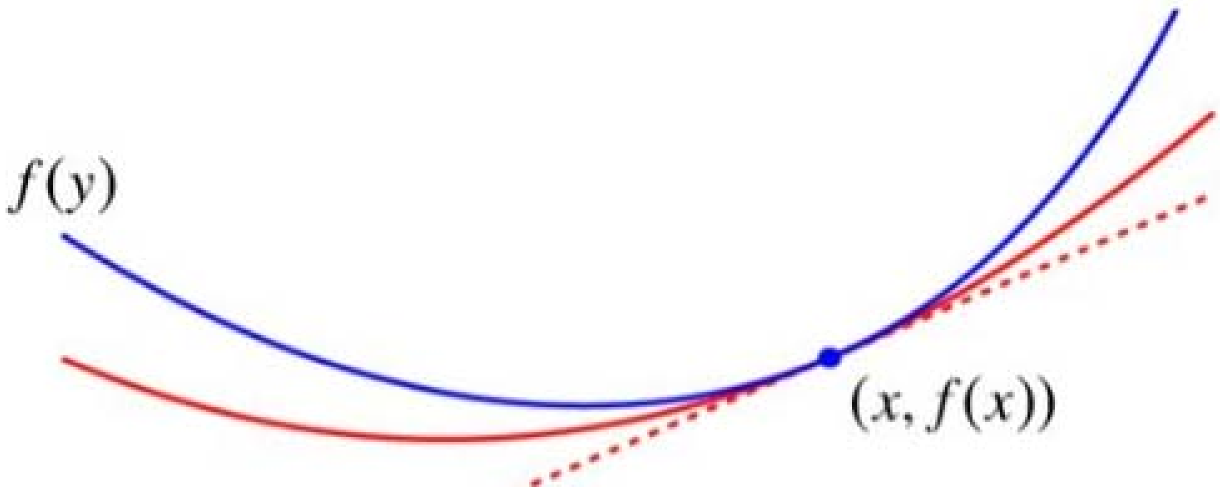
强凸性：对于强凸函数，下面几个式子是等价的

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq m\|x - y\|^2 \quad \text{for all } x, y \in \mathbf{dom}\, f$$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|^2 \quad \text{for all } x, y \in \mathbf{dom}\, f$$

$$g(x) = f(x) - \frac{m}{2}\|x\|^2 \ \text{ is convex}$$

第 3 个式子不等号右边定义了一个二次曲线，这个二次曲线是原函数的下界



13.5. 收敛性分析

强凸函数：

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

由于函数二阶可微：

$$\forall x, y \in domf, f(y) \approx f(x)+\nabla^T f(x)(y-x)+1/2(y-x)^T \nabla^2 f(x)(y-x)$$

结合强凸性：

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

当给定x时，上式不等号右边为y的凸函数，对其关于y求导并令其导数为 0：

$$\nabla f(x)+m(y-x)=0$$

$$\Rightarrow y^*=x-1/m\nabla f(x)$$

13.5. 收敛性分析

再将 y^* 带入，得到

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

结合可知：

$$f(y) \ge f(x) + \nabla^T f(x)(y\text{-}x) + m/2 \|(y\text{-}x)\|_2^2$$

$$\ge f(x) + \nabla^T f(x)\left(y^*\text{-}x\right) + m/2 \left\|\left(y^*\text{-}x\right)\right\|_2^2 = f(x) - \frac{1}{2m} \| \nabla f(x) \|_2^2$$

前面的阐述对任意 $y \in \mathrm{dom} f$ 均成立，故：

$$p^* \ge f(x) - \frac{1}{2m} \| \nabla f(x) \|_2^2 \qquad (7)$$

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

最优解和当前解在函数值上的差值和当前解的梯度有关！

13.5. 收敛性分析

接下来分析梯度很小时，x是否接近最优解：

$$f\left(x^*\right)=p^* \geq f(x)+\nabla^T f(x)\left(x^*-x\right)+m/2\left\|\left(x^*-x\right)\right\|_2^2$$

$$\geq f(x)-\|\nabla f(x)\|\left\|x^*-x\right\|+m/2\left\|\left(x^*-x\right)\right\|_2^2$$

且 $f(x) \geq p^*$

$$\Rightarrow -\|\nabla f(x)\|\left\|x^*-x\right\|+m/2\left\|\left(x^*-x\right)\right\|_2^2<0$$

$$\Rightarrow \left\|x-x^*\right\|_2 \leq 2/m\|\nabla f(x)\|_2$$

把最优解代入进去，我们有下式的左侧

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

$$\Rightarrow \frac{m}{2}\left\|z-x^\star\right\|^2 \leq f(z)-f\left(x^\star\right) \leq \frac{1}{2m}\|\nabla f(z)\|_*^2$$

13.5. 收敛性分析

复习：Lipschitz 连续性

$$\|\nabla f(x)-\nabla f(y)\|_* \le L\|x-y\| \quad \text{for all } x, y \in \mathrm{dom} f$$

下面的式子都是等价的

$$\|\nabla f(x)-\nabla f(y)\|_* \le L\|x-y\| \text{ for all } x, y \in \mathrm{dom} f$$

$$(\nabla f(x)-\nabla f(y))^T(x-y) \le L\|x-y\|^2 \quad \text{for all } x, y \in \mathrm{dom} f$$

$$f(y) \le f(x)+\nabla f(x)^T(y-x)+\frac{L}{2}\|y-x\|^2 \quad \text{for all } x, y \in \mathrm{dom} f$$

$$(\nabla f(x)-\nabla f(y))^T(x-y) \ge \frac{1}{L}\|\nabla f(x)-\nabla f(y)\|_*^2 \quad \text{for all } x, y$$

$$g(x)=\frac{L}{2}\|x\|_2^2-f(x) \text{ is convex}$$

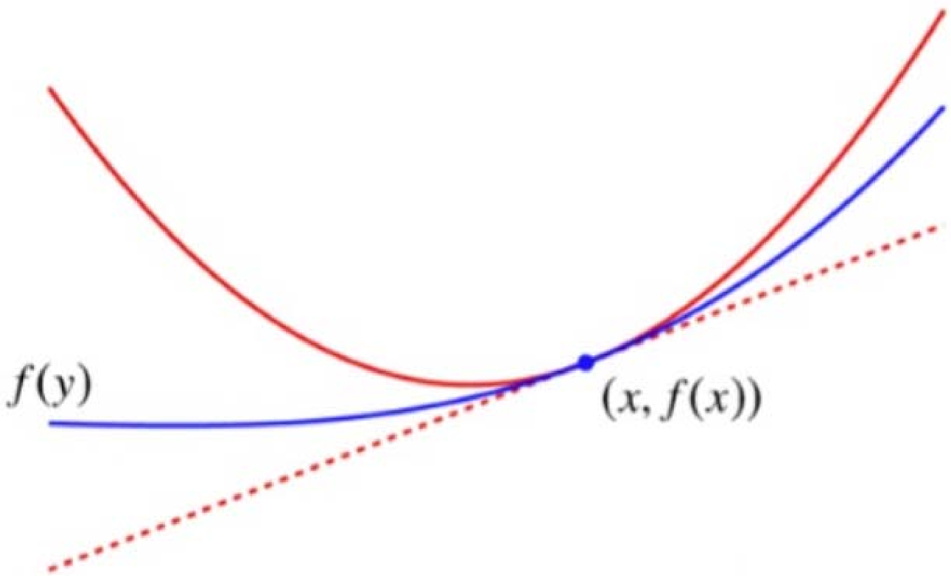
13.5. 收敛性分析

上面第三个式子

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \| y-x \|^2 \quad \text{for all } x, y \in \mathrm{dom} f$$

实际上定义了一个二次曲线，这个曲线是原始函数的

Quadratic upper bound



13.5. 收敛性分析

并且由这个式子可以推导出

$$\frac{1}{2L}\|\nabla f(z)\|_*^2 \le f(z) - f\left(x^\star\right) \le \frac{L}{2}\|z - x^\star\|^2 \quad \text{for all } z$$

这个式子中的上界 $\frac{L}{2}\|z - x^\star\|^2$ 带有 $x^\star$ 是未知的，而下界只与当前值 z 有关,因此在优化过程中我们可以判断当前的 $f(z)$ 与最优值的距离至少为 $\frac{1}{2L}\|\nabla f(z)\|_*^2$ ，如果这个值大于 0，那么我们一定还没得到最优解。

上面的最后一个式子

$$(\nabla f(x) - \nabla f(y))^T (x - y) \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_*^2 \quad \text{for all } x, y$$

被称为 ∇f 的 co-coercivity 性质。（这其实有点像 ∇f 的反函数的 Lipschitz continuous 性质，所以它跟 ∇f 的 Lipschitz continuous 性质是等价的）

13.5. 收敛性分析

精确搜索的收敛性：精确搜索得到的步长 t^k 使得目标函数减少量存在和 $\left|\nabla f\left(x^k\right)^T d^k\right|$ 成比例的下界

$$f\left(x^k\right)-f\left(x^{k+1}\right)\geq\frac{1}{2M}\frac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|^2}$$

理由：$f\left(x^{k+1}\right)=\min_{t>0} f\left(x^k+td^k\right)\leq\min_{t>0} f\left(x^k\right)+t\nabla f\left(x^k\right)d^k+\frac{1}{2}t^2M\left\|d^k\right\|^2$

$$\Rightarrow\qquad f\left(x^{k+1}\right)\leq f\left(x^k\right)-\frac{\left(\nabla f\left(x^k\right)d^k\right)^2}{2M\left\|d^k\right\|^2}$$

请问这里证明利用了强凸性还是 Lipschitz 连续性？

13.5. 收敛性分析

回溯搜索的收敛性：目标函数减少量存在和$\left|\nabla f\left(x^k\right)^T d^k\right|$成比例的下界

$$f\left(x^k\right)-f\left(x^{k+1}\right)\geq\begin{cases}\alpha\left|\nabla f\left(x^k\right)^T d^k\right| & \text{if } t^k=1\\ \dfrac{\alpha\beta}{M}\dfrac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|^2} & \text{if } t^k<1\end{cases}$$

理由：$t^k=1 \quad \Rightarrow \quad f\left(x^{k+1}\right)=f\left(x^k+d^k\right)\leq f\left(x^k\right)+\alpha\nabla f\left(x^k\right)^T d^k$

$$\Rightarrow \quad f\left(x^k\right)-f\left(x^{k+1}\right)\geq-\alpha\nabla f\left(x^k\right)^T d^k=\alpha\left|\nabla f\left(x^k\right)^T d^k\right|$$

13.5. 收敛性分析

$t^k < 1$，首先注意到 $f\left(x^k + \beta^{-1}t^k d^k\right) - f\left(x^k\right) > \alpha \nabla f\left(x^k\right)^T d^k \beta^{-1} t^k$

因此

$$t^k < 1 \quad \Rightarrow \quad f\left(x^k\right) + \alpha \nabla f\left(x^k\right)^T d^k \beta^{-1} t^k < f\left(x^k + \beta^{-1} t^k d^k\right)$$

$$\le f\left(x^k\right) + \nabla f\left(x^k\right)^T d^k \beta^{-1} t^k + \frac{1}{2}\left(\beta^{-1} t^k\right)^2 M \left\|d^k\right\|^2$$

$$\Rightarrow \quad t^k > \frac{2(1-\alpha)\beta}{M} \frac{\left|\nabla f\left(x^k\right)^T d^k\right|}{\left\|d^k\right\|^2} > \frac{\beta}{M} \frac{\left|\nabla f\left(x^k\right)^T d^k\right|}{\left\|d^k\right\|^2}$$

$$\Rightarrow \quad f\left(x^k\right) - f\left(x^{k+1}\right) \ge \frac{\alpha\beta}{M} \frac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|^2}$$

13.5. 收敛性分析

当我们选择负梯度方向为下降方向时， $d^k = -\nabla f\left(x^k\right) \Rightarrow$

$$\nabla f\left(x^k\right)^T d^k = -\left\|\nabla f\left(x^k\right)\right\|^2 = -\left\|d^k\right\|^2$$

满足给定误差阈值的迭代次数上界 K

精确搜索：不小于 $\log\left(\left(f\left(x^0\right)-p^*\right)/\varepsilon\right) \Big/ \left|\log\left(1-m/M\right)\right|$ 的整数

回溯搜索：不小于 $\dfrac{\log\left(\left(f\left(x^0\right)-p^*\right)/\varepsilon\right)}{\left|\log\left(1-2m\alpha\min\left\{1, \dfrac{\beta}{M}\right\}\right)\right|}$ 的整数

$\dfrac{m}{M}$ 越小（$\nabla^2 f(x)$ 的特征根差异越大）， K 越大， 反之亦然

13.5. 收敛性分析

精确搜索上界推导过程

利用不等式 $\|\nabla f(x)\|^2 \geq 2m\left(f(x)-p^*\right)$

$$\Rightarrow \qquad f\left(x^k\right)-f\left(x^{k+1}\right) \geq \frac{1}{2M}\frac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|^2} \geq \frac{m}{M}\left(f\left(x^k\right)-p^*\right)$$

$$\Rightarrow \qquad f\left(x^{k+1}\right)-p^* \leq \left(1-\frac{m}{M}\right)\left(f\left(x^k\right)-p^*\right)$$

$$\Rightarrow \qquad f\left(x^K\right)-p^* \leq \left(1-\frac{m}{M}\right)\left(f\left(x^{K-1}\right)-p^*\right) \leq \cdots \leq \left(1-\frac{m}{M}\right)^K\left(f\left(x^0\right)-p^*\right)$$

$\left(1-\dfrac{m}{M}\right)^K\left(f\left(x^0\right)-p^*\right) \leq \varepsilon$ 线性收敛 $\Rightarrow$ $K \geq \dfrac{\log\left(\left(f\left(x^0\right)-p^*\right)/\varepsilon\right)}{\left|\log\left(1-\dfrac{m}{M}\right)\right|}$

13.5. 收敛性分析

回溯搜索上界推导过程

$t^k = 1$ $$f\left(x^k\right) - f\left(x^{k+1}\right) \geq \alpha \left|\nabla f\left(x^k\right)^T d^k\right| = \alpha \left\|\nabla f\left(x^k\right)\right\|^2$$

$t^k < 1$ $$f\left(x^k\right) - f\left(x^{k+1}\right) \geq \frac{\alpha\beta}{M} \left\|\nabla f\left(x^k\right)\right\|^2$$

$$\Rightarrow \quad f\left(x^k\right) - f\left(x^{k+1}\right) \geq \min\left\{\alpha, \frac{\alpha\beta}{M}\right\} \left\|\nabla f\left(x^k\right)\right\|^2$$

$$\Rightarrow \quad f\left(x^K\right) - p^* \leq \left(1 - \min\left\{2m\alpha, \frac{2m\alpha\beta}{M}\right\}\right)^K \left(f\left(x^0\right) - p^*\right)$$

$$f\left(x^K\right) - p^* \leq \varepsilon \quad \Rightarrow \quad K \geq \frac{\log\left(\left(f\left(x^0\right) - p^*\right)/\varepsilon\right)}{\left|\log\left(1 - 2m\alpha \min\left\{1, \frac{\beta}{M}\right\}\right)\right|}$$

13.5. 收敛性分析

最速下降方法满足给定误差阈值的迭代次数上界

精确搜索：不小于 $\log\left(\left(f\left(x^0\right)-p^*\right)/\varepsilon\right)\Big/\left|\log\left(1-\frac{mC_1{}^2C_2{}^2}{M}\right)\right|$ 的整数

回溯搜索：不小于 $\dfrac{\log\left(\left(f\left(x^0\right)-p^*\right)/\varepsilon\right)}{\left|\log\left(1-2m\alpha C_2{}^2\min\left\{1,\dfrac{\beta C_1{}^2}{M}\right\}\right)\right|}$ 的整数

式中 C_1 和 C_2 是满足 $\|x\|\geq C_1\|x\|_2$, $\|x\|_*\geq C_2\|x\|_2$, $\forall x$ 的常数

$C_1, C_2\in(0,1)$

收敛性质同梯度方法，取决于 $\dfrac{m}{M}$ 的大小

13.5. 收敛性分析

精确搜索上界推导过程

$$f\left(x^k\right)-f\left(x^{k+1}\right)\geq\frac{1}{2M}\frac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|_2^2}=\frac{1}{2M}\frac{\left\|\nabla f\left(x^k\right)\right\|_*^2}{\left\|d^k\right\|_2^2}$$

$$\geq\frac{C_1^2}{2M}\frac{\left\|\nabla f\left(x^k\right)\right\|_*^2}{\left\|d^k\right\|^2}=\frac{C_1^2}{2M}\left\|\nabla f\left(x^k\right)\right\|_*^2\geq\frac{C_1^2C_2^2}{2M}\left\|\nabla f\left(x^k\right)\right\|_2^2$$

再利用$\left\|\nabla f\left(x\right)\right\|_2^2\geq 2m\left(f\left(x\right)-p^*\right),\forall x$可得所需上界

13.5. 收敛性分析

回溯搜索上界推导过程（用$\left\|\nabla f\left(x^k\right)\right\|_* d^k$替换$d^k$）

$t^k = 1$

$$f\left(x^k\right) - f\left(x^{k+1}\right) \geq \alpha \left|\nabla f\left(x^k\right)^T d^k\right| = \alpha \left\|\nabla f\left(x^k\right)\right\|_*^2 \geq \alpha C_2^2 \left\|\nabla f\left(x^k\right)\right\|_2^2$$

$t^k < 1$

$$f\left(x^k\right) - f\left(x^{k+1}\right) \geq \frac{\alpha\beta}{M} \frac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|_2^2}$$

$$\geq \frac{\alpha\beta C_1^2 C_2^2}{M} \left\|\nabla f\left(x^k\right)\right\|_2^2$$

再利用$\left\|\nabla f(x)\right\|_2^2 \geq 2m\left(f(x) - p^*\right), \forall x$可得所需上界

13.5. 收敛性分析

Newton method 收敛性分析基本想法

assumptions

- f strongly convex on S with constant m
- $\nabla^2 f$ is Lipschitz continuous on S, with constant $L > 0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x - y\|_2$$

(L measures how well f can be approximated by a quadratic function)

outline: there exist constants $\eta \in (0, m^2/L)$, $\gamma > 0$ such that

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

$$\frac{L}{2m^2}\|\nabla f(x^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^{(k)})\|_2\right)^2$$

13.5. 收敛性分析

damped Newton phase ($\|\nabla f(x)\|_2 \geq \eta$)

- most iterations require backtracking steps
- function value decreases by at least γ
- if $p^\star > -\infty$, this phase ends after at most $(f(x^{(0)}) - p^\star)/\gamma$ iterations

quadratically convergent phase ($\|\nabla f(x)\|_2 < \eta$)

- all iterations use step size $t = 1$
- $\|\nabla f(x)\|_2$ converges to zero quadratically: if $\|\nabla f(x^{(k)})\|_2 < \eta$, then

$$\frac{L}{2m^2}\|\nabla f(x^l)\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(x^k)\|_2\right)^{2^{l-k}} \leq \left(\frac{1}{2}\right)^{2^{l-k}}, \qquad l \geq k$$

13.5. 收敛性分析

conclusion: number of iterations until $f(x) - p^\star \leq \epsilon$ is bounded above by

$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

- γ, ϵ_0 are constants that depend on m, L, $x^{(0)}$
- second term is small (of the order of 6) and almost constant for practical purposes
- in practice, constants m, L (hence γ, ϵ_0) are usually unknown
- provides qualitative insight in convergence properties ($i.e.$, explains two algorithm phases)

13.5. 收敛性分析

推导迭代次数上界的关键步骤

确定满足如下条件的η

$$\left\|\nabla f\left(x^k\right)\right\| \le \eta \qquad \Rightarrow \qquad t^k = 1,\ \left\|\nabla f\left(x^k + d^k\right)\right\| \le \eta$$

这个条件的作用在于

对任意的$l \ge k$ 成立$t^l = 1$

$$\Rightarrow \quad \text{迭代进入二次收敛阶段} \quad \left\|\nabla f\left(x^{l+1}\right)\right\| \le c\left\|\nabla f\left(x^l\right)\right\|^2$$

确定能保证$t^k=1$的$\left\|\nabla f\left(x^k\right)\right\|$的$\eta$的上界

$$\tilde{f}(t)=f\left(x^k+td^k\right)$$

$$\Rightarrow \quad \tilde{f}''(t)=\left(d^k\right)^T\nabla^2 f\left(x^k+td^k\right)d^k$$

$$\Rightarrow \quad \tilde{f}''(t)\le \tilde{f}''(0)+tL\left\|d^k\right\|^3\le \lambda(x^k)^2+t\frac{L}{m^{3/2}}\lambda(x^k)^3$$ （Lipschitz 条件）

$$\Rightarrow \quad \tilde{f}(t)\le \tilde{f}(0)-t\lambda\left(x^k\right)^2+\frac{t^2}{2}\lambda\left(x^k\right)^2+t^3\frac{L}{6m^{3/2}}\lambda\left(x^k\right)^3$$ （对t两次积分）

$$\Rightarrow \quad f\left(x^k+d^k\right)\le f\left(x^k\right)+\nabla f\left(x^k\right)^T d^k\left(\frac{1}{2}-\frac{L}{6m^{3/2}}\lambda\left(x^k\right)\right)$$ （取$t=1$）

$$\Rightarrow \quad \left(\frac{1}{2}-\frac{L}{6m^{3/2}}\lambda\left(x^k\right)\right)\ge\alpha$$ （满足回溯直线搜索条件）

$$\Rightarrow \quad \left\|\nabla f\left(x^k\right)\right\|\le 3(1-2\alpha)\frac{m^2}{L}$$ （利用$\lambda\left(x^k\right)\le\frac{1}{m^{1/2}}\left\|\nabla f\left(x^k\right)\right\|$）

$$\Rightarrow \quad \eta\le 3(1-2\alpha)\frac{m^2}{L}$$

确定由$\left\|\nabla f\left(x^k\right)\right\| \le \eta$可保证$\left\|\nabla f\left(x^k+d^k\right)\right\| \le \eta$的$\eta$界

$$\frac{d\nabla f\left(x^k+td^k\right)}{dt}=\nabla^2 f\left(x^k+td^k\right)d^k$$

$\Rightarrow \quad \nabla f\left(x^k+d^k\right)=\int_0^1\left(\nabla^2 f\left(x^k+td^k\right)-\nabla^2 f\left(x^k\right)\right)d^k dt$ （必须牛顿方向）

$$\Rightarrow \quad \left\|\nabla f\left(x^k+d^k\right)\right\| \le \frac{L}{2}\left\|d^k\right\|^2 \le \frac{L}{2m^2}\left\|\nabla f\left(x^k\right)\right\|^2$$

$$\left\|\nabla f\left(x^{k+1}\right)\right\| \le \eta \quad \Rightarrow \quad \frac{L}{2m^2}\eta \le 1 \quad \Rightarrow \quad \eta \le \frac{2m^2}{L}$$

综合以上两个不等式，可知以下η能满足要求

$$\eta=\min\left\{2, 3\left(1-2\alpha\right)\right\}\frac{m^2}{L}$$

结论：如果某个$k>0$满足

$$\left\|\nabla f\left(x^k\right)\right\| \le \eta = \min\left\{2, 3(1-2\alpha)\right\}\frac{m^2}{L}$$

那么对任意的$l \ge k$成立

$$f\left(x^l + d^l\right) \le f\left(x^l\right) + \alpha \nabla f\left(x^l\right)^T d^l \quad \Rightarrow \quad t^l = 1$$

$$\frac{L}{2m^2}\left\|\nabla f\left(x^{l+1}\right)\right\| \le \left(\frac{L}{2m^2}\left\|\nabla f\left(x^l\right)\right\|\right)^2$$

$$\Rightarrow \quad \frac{L}{2m^2}\left\|\nabla f\left(x^{k+\tau}\right)\right\| \le \left(\frac{L}{2m^2}\left\|\nabla f\left(x^k\right)\right\|\right)^{2^\tau} \le \left(\frac{L}{2m^2}\eta\right)^{2^\tau}, \quad \forall \tau \ge 0$$

如果再加上$\frac{L}{m^2}\eta \le 1$ $\quad \Rightarrow \quad \frac{L}{2m^2}\left\|\nabla f\left(x^{k+\tau}\right)\right\| \le \left(\frac{1}{2}\right)^{2^\tau}, \quad \forall \tau \ge 0$

因此需要$\eta \le \frac{m^2}{L}$，结合上面的要求可知$\eta = \min\left\{1, 3(1-2\alpha)\right\}\frac{m^2}{L}$

称 $l \geq k$ 前后为二次收敛和阻尼 Newton 阶段

二次收敛阶段迭代次数 K_2 的上界

$$f\left(x^{k+K_2-1}\right)-p^* \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{K_2}}$$

$$\Rightarrow \quad \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{K_2}} \leq \varepsilon \quad \Rightarrow \quad \frac{\varepsilon_0}{\varepsilon} \leq 2^{2^{K_2}}$$ （其中 $\varepsilon_0 \triangleq \dfrac{2m^3}{L^2}$ ）

$$\Rightarrow \quad K_2 \geq \log_2\left(\log_2\left(\frac{\varepsilon_0}{\varepsilon}\right)\right)$$

阻尼 Newton 阶段迭代次数 K_1 的上界

$$\left\|\nabla f\left(x^k\right)\right\| > \eta,\ \ \forall k \le K_1$$

$$\Rightarrow \quad f\left(x^k\right) - f\left(x^{k+1}\right) \ge \frac{\alpha\beta}{M}\frac{\left|\nabla f\left(x^k\right)^T d^k\right|^2}{\left\|d^k\right\|^2} \ge \frac{\alpha\beta m}{M}\left|\nabla f\left(x^k\right)^T d^k\right|$$

$$\ge \frac{\alpha\beta m}{M^2}\left\|\nabla f\left(x^k\right)\right\|^2 > \frac{\alpha\beta m\eta^2}{M^2} \triangleq \gamma,\ \ \forall k \le K_1$$

$$\Rightarrow \quad K_1 \le \frac{f\left(x^0\right) - p^*}{\gamma} = \frac{M^2 L^2}{\alpha\beta m^5 \min\left\{1, 9\left(1-2\alpha\right)^2\right\}}\left(f\left(x^0\right) - p^*\right)$$

总结：Newton 方法总迭代次数 $K = K_1 + K_2$ 为超过下式的整数

$$\frac{M^2L^2}{\alpha\beta m^5 \min\left\{1, 9(1-2\alpha)^2\right\}}\left(f\left(x^0\right)-p^*\right)+\log_2\left(\log_2\left(\frac{\varepsilon_0}{\varepsilon}\right)\right)$$

其中 $\varepsilon_0 = 2m^3/L^2$， L 是满足下述 Lipschitz 条件的 Lipschitz 常数（以下向量范数不标明均指 ℓ_2 范数）

$$\left\|\nabla^2 f(y)-\nabla^2 f(x)\right\| \le L\|y-x\|, \forall y, x \in S$$

式中矩阵范数为 ℓ_2 范数导出的算子范数 $\|A\|^2 \triangleq \lambda_{\max}\left(A^T A\right)$

如果 $\log_2\left(\log_2\left(\varepsilon_0/\varepsilon\right)\right)=6$， $\dfrac{\varepsilon_0}{\varepsilon}=2^{2^6} \approx 2\times10^{19}$

一般情况下可假定 Newton 方法总迭代次数不超过

$$\frac{M^2L^2}{\alpha\beta m^5 \min\left\{1, 9(1-2\alpha)^2\right\}}\left(f\left(x^0\right)-p^*\right)+6$$

13.6. 自和谐性

shortcomings of classical convergence analysis

- depends on unknown constants (m, L, . . .)
- bound is not affinely invariant, although Newton's method is

convergence analysis via self-concordance (Nesterov and Nemirovski)

- does not depend on any unknown constants
- gives affine-invariant bound
- applies to special class of convex functions ('self-concordant' functions)
- developed to analyze polynomial-time interior-point methods for convex optimization

13.6. 自和谐性

definition

- $f : \mathbf{R} \to \mathbf{R}$ is self-concordant if $|f'''(x)| \leq 2f''(x)^{3/2}$ for all $x \in \mathbf{dom}\, f$
- $f : \mathbf{R}^n \to \mathbf{R}$ is self-concordant if $g(t) = f(x + tv)$ is self-concordant for all $x \in \mathbf{dom}\, f$, $v \in \mathbf{R}^n$

examples on R

- linear and quadratic functions
- negative logarithm $f(x) = -\log x$
- negative entropy plus negative logarithm: $f(x) = x \log x - \log x$

affine invariance: if $f : \mathbf{R} \to \mathbf{R}$ is s.c., then $\tilde{f}(y) = f(ay + b)$ is s.c.:

$$\tilde{f}'''(y) = a^3 f'''(ay + b), \qquad \tilde{f}''(y) = a^2 f''(ay + b)$$

13.6. 自和谐性

properties

- preserved under positive scaling $\alpha \geq 1$, and sum
- preserved under composition with affine function
- if g is convex with $\mathbf{dom}\, g = \mathbf{R}_{++}$ and $|g'''(x)| \leq 3g''(x)/x$ then

$$f(x) = \log(-g(x)) - \log x$$

  is self-concordant

examples: properties can be used to show that the following are s.c.

- $f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x)$ on $\{x \mid a_i^T x < b_i,\ i = 1, \ldots, m\}$
- $f(X) = -\log\det X$ on $\mathbf{S}_{++}^n$
- $f(x) = -\log(y^2 - x^T x)$ on $\{(x, y) \mid \|x\|_2 < y\}$

13.6. 自和谐性

假设我们选择v为前进方向，t为步长，$\tilde{f}(t)=f(x+tv)$

自和谐本质$\left|\tilde{f}'''(t)\right|\le 2\tilde{f}''(t)^{3/2} \Leftrightarrow \left|\frac{d}{dt}\left(\tilde{f}''(t)^{-1/2}\right)\right|\le 1$

$$\Rightarrow \qquad \frac{\tilde{f}''(0)}{\left(1+t\tilde{f}''(0)^{1/2}\right)^2}\le \tilde{f}''(t)\le \frac{\tilde{f}''(0)}{\left(1-t\tilde{f}''(0)^{1/2}\right)^2}$$

$$\Rightarrow \qquad \begin{aligned} \tilde{f}(t) &\ge \tilde{f}(0)+t\tilde{f}'(0)+t\tilde{f}''(0)^{\frac{1}{2}}-\log\left(1+t\tilde{f}''(0)^{\frac{1}{2}}\right) \\ \tilde{f}(t) &\le \tilde{f}(0)+t\tilde{f}'(0)-t\tilde{f}''(0)^{\frac{1}{2}}-\log\left(1-t\tilde{f}''(0)^{\frac{1}{2}}\right) \end{aligned}$$

13.6. 自和谐性

1）我们可利用第一个不等式导出类似 $f(x)-p^{*} \le C\|\nabla f(x)\|^{2}$ 的不等式。在不等式右边关于 t 求最小

$$\Rightarrow \qquad \tilde{f}(t) \ge \tilde{f}(0)-\tilde{f}'(0)\tilde{f}''(0)^{-\frac{1}{2}}+\log\left(1+\tilde{f}'(0)\tilde{f}''(0)^{-\frac{1}{2}}\right)$$

因为 $\tilde{f}'(0)=v^{T}\nabla f(x)$， $\tilde{f}''(0)=v^{T}\nabla^{2} f(x)v$， $\lambda(x) \ge \dfrac{-v^{T}\nabla f(x)}{\left(v^{T}\nabla^{2} f(x)v\right)^{\frac{1}{2}}}$，

所以 $\lambda(x) \ge -\tilde{f}'(0)\tilde{f}''(0)^{-\frac{1}{2}}$； 又因为 $u+\log(1-u)$ 是单减函数，所以 $\tilde{f}(t) \ge \tilde{f}(0)+\lambda(x)+\log(1-\lambda(x))$； 再利用对所有的 $0 \le u \le 0.68$ 成立 $u^{2}+u+\log(1-u) \ge 0$，最终可得：当 $\lambda(x) \le 0.68$ 时，成立

$$p^{*}=\inf_{t>0,\mathrm{v}} \tilde{f}(t) \ge \tilde{f}(0)+\lambda(x)+\log(1-\lambda(x)) \ge f(x)-\lambda(x)^{2}$$

2）利用第二个不等式导出回溯直线搜索函数减量下界

取 $v=-\nabla^2 f(x)^{-1}\nabla f(x)$，因为 $\tilde{f}'(0)=-\lambda(x)^2$， $\tilde{f}''(0)=\lambda(x)^2$，所以

$$\tilde{f}(t)\le \tilde{f}(0)-t\lambda(x)^2-t\lambda(x)-\log\big(1-t\lambda(x)\big)$$

取 $\hat{t}=\big(1+\lambda(x)\big)^{-1}$，因为 $\dfrac{u^2}{2(1+u)}-u+\log(1+u)\le 0, \forall u\ge 0$， 所以

$$\begin{aligned}\tilde{f}(\hat{t})&\le \tilde{f}(0)-\hat{t}\lambda(x)\big(1+\lambda(x)\big)-\log\big(1-\hat{t}\lambda(x)\big)\\&=\tilde{f}(0)-\lambda(x)+\log\big(1+\lambda(x)\big)\le \tilde{f}(0)-\alpha\frac{\lambda(x)^2}{1+\lambda(x)}\\&\le \tilde{f}(0)+\alpha v^T\nabla f(x)\hat{t}\end{aligned}$$

由此可知，直线搜索停止时的步长满足 $t>\beta\hat{t}=\beta\big(1+\lambda(x)\big)^{-1}$

$$\Rightarrow\qquad \tilde{f}(0)-\tilde{f}(t)\ge -\alpha v^T\nabla f(x)\beta\hat{t}=\alpha\beta\lambda(x)^2\big(1+\lambda(x)\big)^{-1}$$

也就是说，进入二次收敛的条件是$\lambda\left(x^k\right)\le\frac{(1-2\alpha)}{4}$

利用$u^3+0.5u^2+u+\log(1-u)\ge 0,\ \forall 0<u\le 0.81$

$$
\begin{aligned}
&\lambda\left(x^k\right)\le\frac{1-2\alpha}{4}(\le 0.81)\\
\Rightarrow\quad&\tilde{f}(1)\le\tilde{f}(0)-\lambda\left(x^k\right)^2-\lambda\left(x^k\right)-\log\left(1-\lambda\left(x^k\right)\right)\\
&\quad\le\tilde{f}(0)-0.5\lambda\left(x^k\right)^2+\lambda\left(x^k\right)^3\le\tilde{f}(0)-\alpha\lambda\left(x^k\right)^2\\
\Rightarrow\quad&t=1
\end{aligned}
$$

利用$\lambda(x)\le\frac{1}{4}\Rightarrow\lambda\left(x-\nabla^2 f(x)^{-1}\nabla f(x)\right)\le 2\lambda(x)^2$

$$
\lambda\left(x^k\right)\le\frac{1-2\alpha}{4}\left(\le\frac{1}{4}\right)\quad\Rightarrow\quad 2\lambda\left(x^{k+1}\right)\le\left(2\lambda\left(x^k\right)\right)^2
$$

二次收敛阶段迭代次数上界

取 $\eta = \dfrac{1-2\alpha}{4}$

如果 $\lambda\left(x^k\right) \le \eta$，因为 $t^k = 1,\ 2\lambda\left(x^{k+1}\right) \le \left(2\lambda\left(x^k\right)\right)^2$，对所有 $l \ge k$ 有

$$2\lambda\left(x^l\right) \le \left(2\lambda\left(x^k\right)\right)^{2^{l-k}} \le \left(2\times\frac{1}{4}\right)^{2^{l-k}} = \left(\frac{1}{2}\right)^{2^{l-k}}$$

$$\Rightarrow \qquad f\left(x^l\right) - p^* \le \lambda\left(x^l\right)^2 \le \frac{1}{4}\lambda\left(x^l\right) \le \left(\frac{1}{2}\right)^{2^{l-k+1}}$$

只要 $l - k + 1 \ge \log_2\left(\log_2\left(\dfrac{1}{\varepsilon}\right)\right)$，就有 $f\left(x^l\right) - p^* \le \varepsilon$

阻尼牛顿阶段迭代次数上界

取 $\gamma = \alpha\beta \frac{\eta^2}{1+\eta}$，利用 $f\left(x^k\right) - f\left(x^{k+1}\right) \geq \alpha\beta \frac{\lambda\left(x^k\right)^2}{1+\lambda\left(x^k\right)}$

$$\lambda\left(x^k\right) > \eta \quad \Rightarrow \quad f\left(x^k\right) - f\left(x^{k+1}\right) \geq \alpha\beta \frac{\lambda\left(x^k\right)^2}{1+\lambda\left(x^k\right)} \geq \alpha\beta \frac{\eta^2}{1+\eta} = \gamma$$

$$\Rightarrow \quad K_1 \leq \frac{f\left(x^0\right) - p^*}{\gamma}$$

总的上界

$$\frac{f\left(x^0\right) - p^*}{\gamma} + \log_2\left(\log_2\left(\frac{1}{\varepsilon}\right)\right) = \frac{20-8\alpha}{\alpha\beta(1-2\alpha)^2}\left(f\left(x^0\right) - p^*\right) + \log_2\left(\log_2\left(\frac{1}{\varepsilon}\right)\right)$$

13.7. References

[1] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004. http://www.stanford.edu/~boyd/cvxbook/ http://www.ee.ucla.edu/~vandenbe/cvxbook

[2] http://www.stat.cmu.edu/~ryantibs/convexopt/lectures/grad-descent.pdf

[3] http://www.stat.cmu.edu/~ryantibs/convexopt/lectures/newton.pdf

[4] 为什么在光滑凸优化研究中，Lipschitz gradient比strongly convex更普遍？ https://www.zhihu.com/question/459410340

[5] J. Bolte, S. Sabach, M. Teboulle, Y. Vaisbourd, "First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2131-2151, 2018.

[6] H. H. Bauschke, J. Bolte, J. Chen, M. Teboulle, X. Wang, "On linear convergence of non-Euclidean gradient methods without strong convexity

and lipschitz gradient continuity," *Journal of Optimization Theory and Applications*, vol. 182, no. 3, pp. 1068-1087, 2019.

[7] I. Necoara, Y. Nesterov, F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," *Mathematical Programming*, vol. 175, no. 1, pp. 69-107, 2019.