优化理论与算法

第九章

牛顿与拟牛顿法

郭加熠|助理教授

目录

讲　　员

郭加熠，江波，刘慧康

回顾梯度下降法

重复

1. 搜索方向：$\Delta x := -\nabla f(x)$
2. 线搜索：通过固定或回溯线搜索选择步长 t
3. 更新：$x := x + t\Delta x$

直到满足停止条件为止

- 停止准则通常为形式 $\|\nabla f(x)\|_2 \leq \epsilon$。
- 对于 m-强凸函数 f 有线性收敛结果：

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

  其中 $c \in (0, 1)$ 依赖于 m，线搜索设置
- 非常简单，但通常非常慢，特别是条件数较差情况

牛顿步

牛顿步 Δx_{nt}，既定义方向，又定义长度

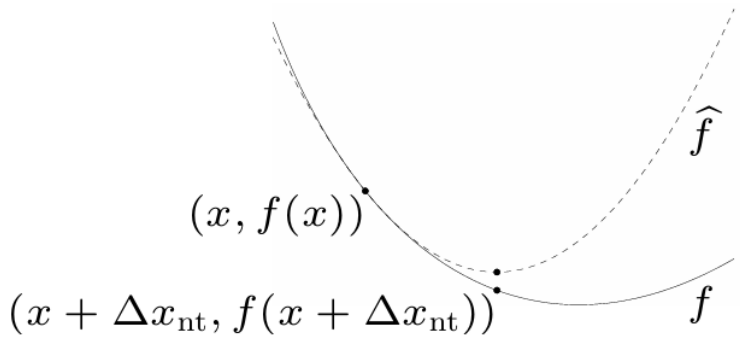
$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

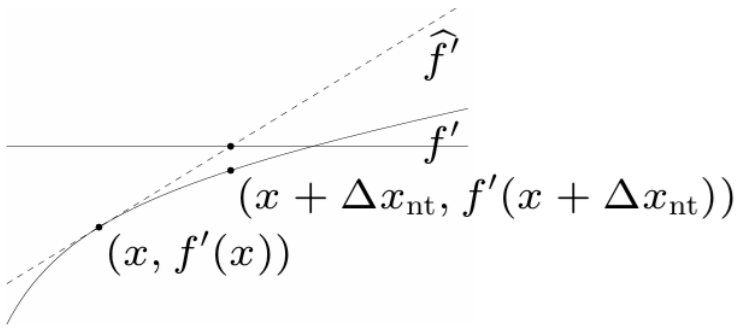
- $x + \Delta x_{\mathrm{nt}}$ 最小化以下二阶近似：

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

- $x + \Delta x_{\mathrm{nt}}$ 为以下最优性条件（线性系统）的解

$$\nabla f(x+v) \approx \nabla \hat{f}(x+v) = \nabla f(x) + \nabla^2 f(x) v = 0$$



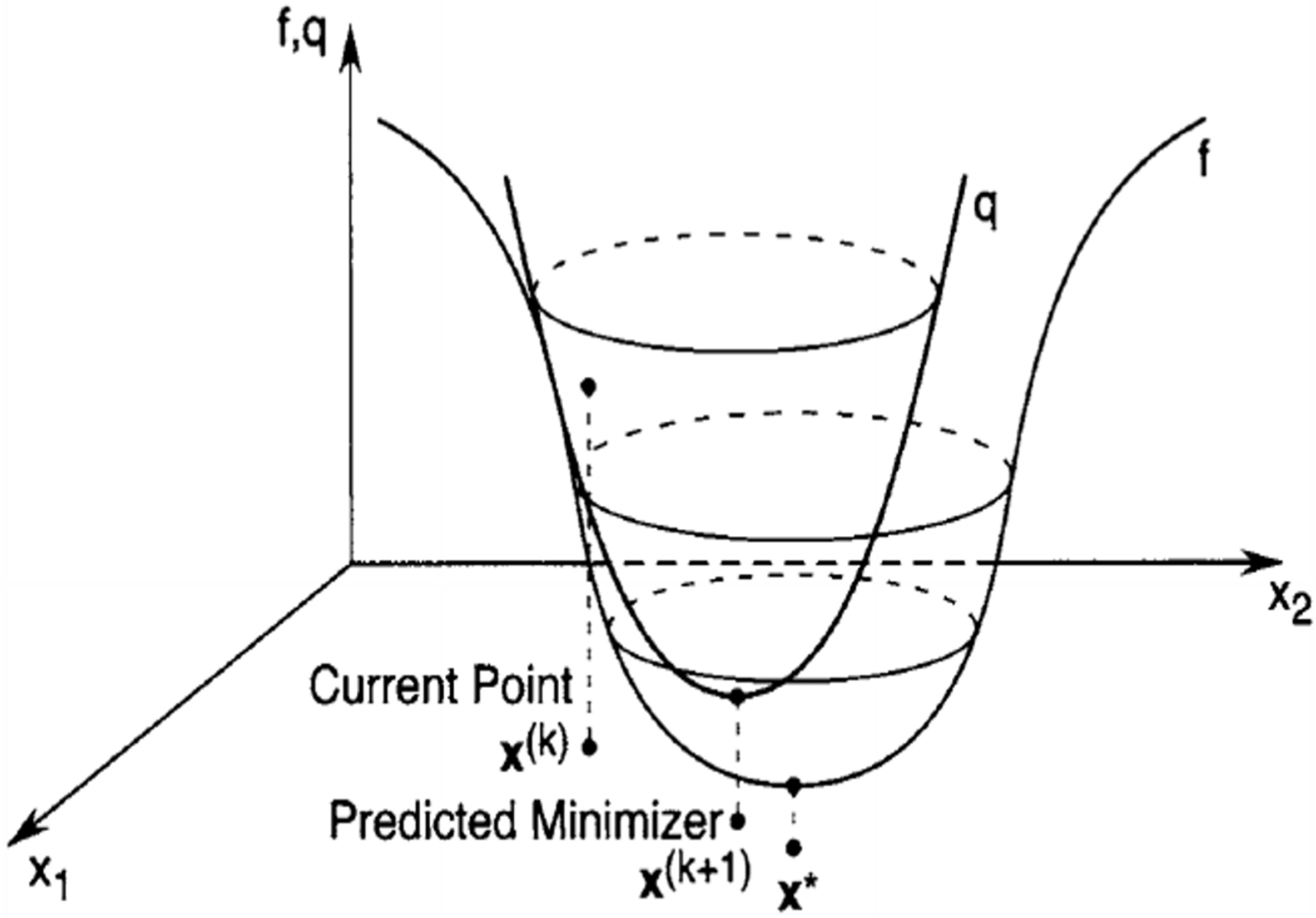


牛顿法

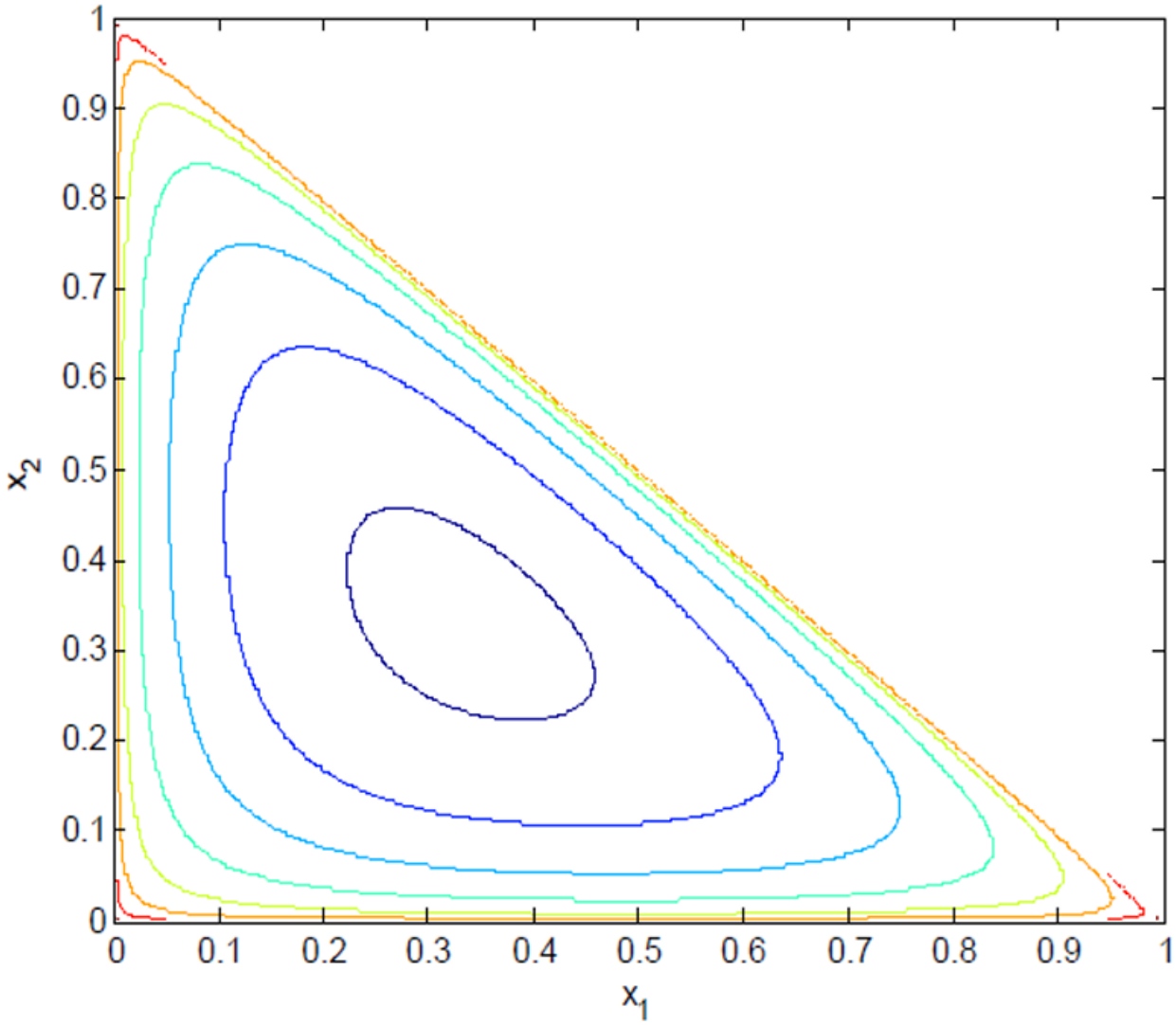
$$\Delta x_{\mathsf{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

$$x_+ = x + \Delta x_{\mathsf{nt}}$$

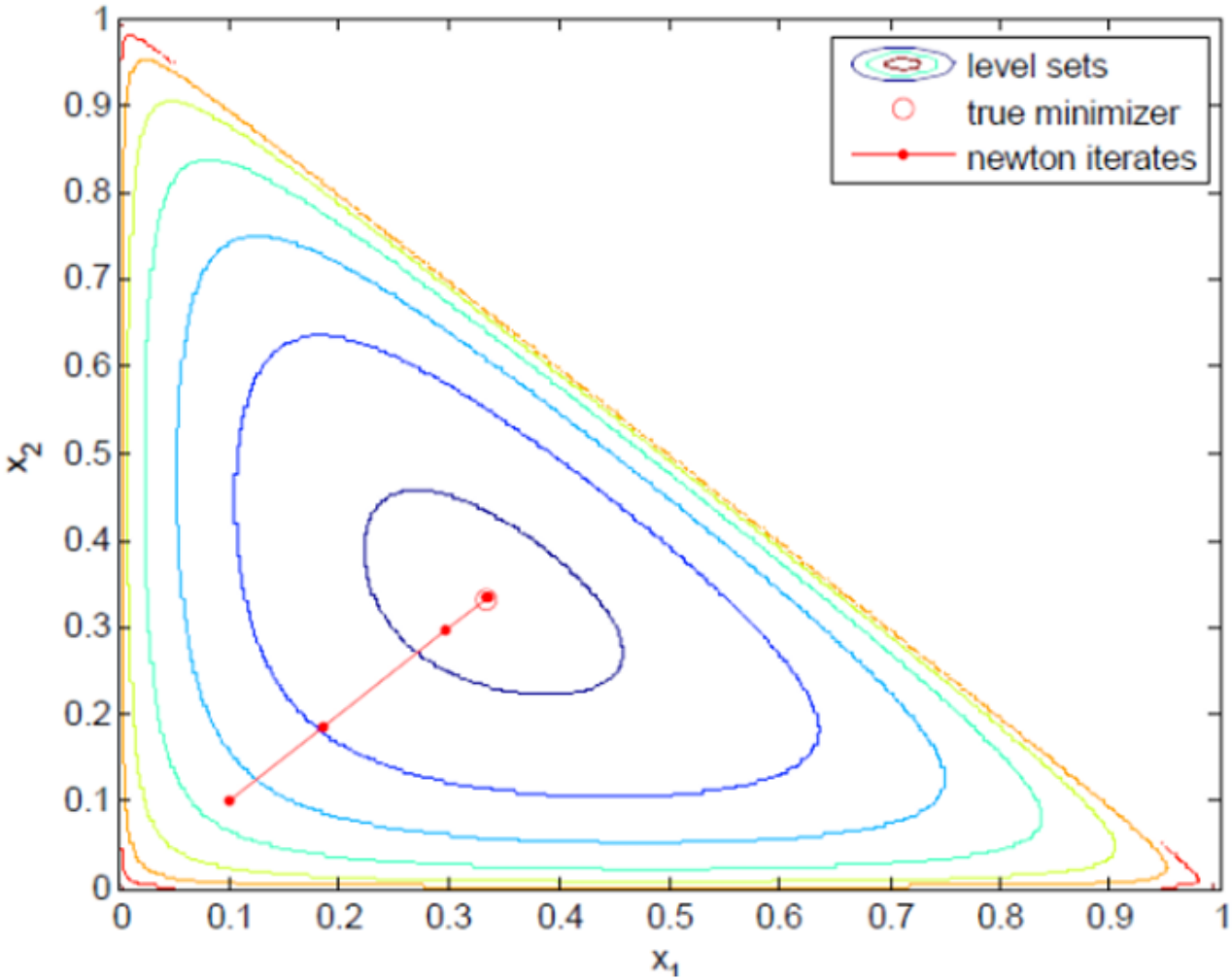
- 即使对于强凸函数，也不保证全局收敛
- 但在距离最优解较近时，局部表现优越
- 对于二次问题 $f(x) = \frac{1}{2} x^T A x + bx$，只需一步（假设 A 可逆）


f,q
q
f
x_2
Current Point
$\mathbf{x}^{(k)}$
Predicted Minimizer
$\mathbf{x}^{(k+1)}$
$\mathbf{x}^*$
x_1

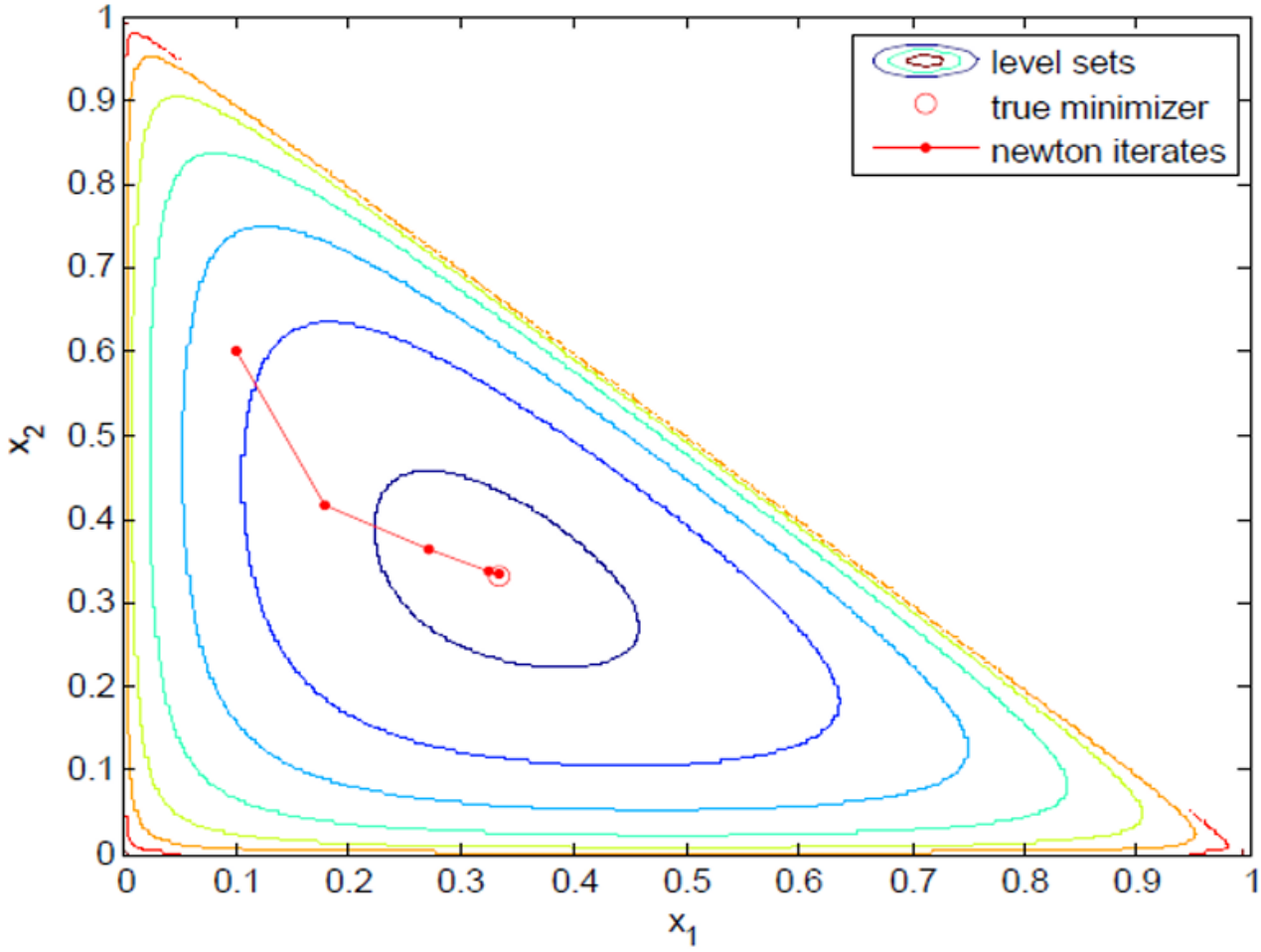
$$f(x_1, x_2) = -\log(1 - x_1 - x_2) - \log(x_1) - \log(x_2)$$



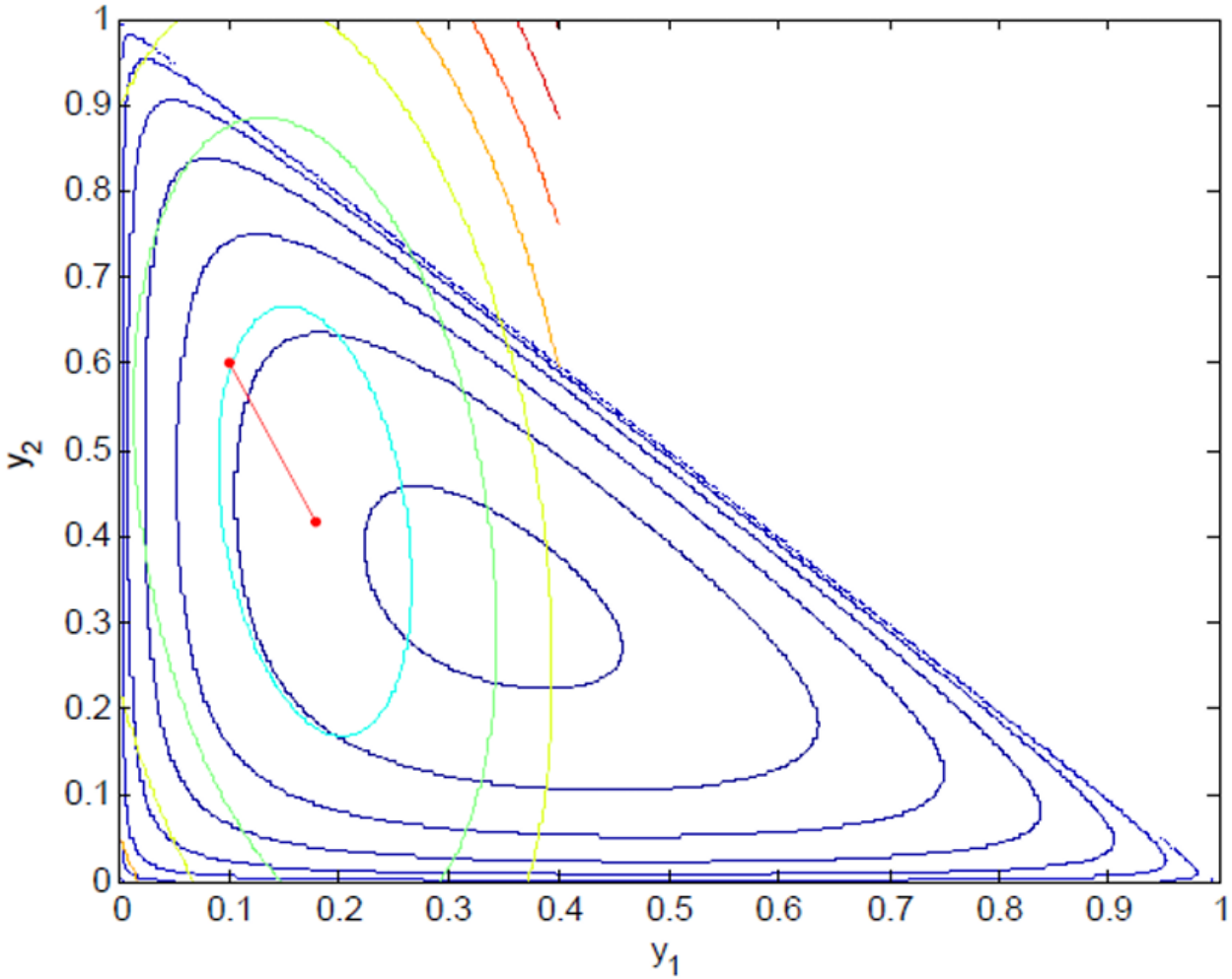
从 $[\frac{1}{10}; \frac{1}{10}]$ 开始牛顿法



从 $[\frac{1}{10}; \frac{6}{10}]$ 开始牛顿法



使用牛顿法执行一步



局部收敛性分析

- 假设 $f(x)$ 是强凸的，二次连续可微，且 Hessian $\nabla^2 f(x)$ 非奇异，可得局部**超线性收敛**
- 在上述条件基础上，继续假设 Hessian $\nabla^2 f(x)$ 还具有 Lipschitz 连续性，可得局部**二次收敛**
- 从 $x^{k+1} - x^* = x^k - x^* - \nabla^2 f(x^k)^{-1} \nabla f(x^k)$ 开始证明
- 证明使用一个关键性质是

$$\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + t(y - x))\,(y - x)\,dt$$

目录

讲　　员

郭加熠，江波，刘慧康

牛顿递减量

牛顿递减量 $\lambda(x)$ 是度量 x 与 x^* 接近程度的一种测度

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

性质

- 利用二阶近似 $\hat{f}$，给出 $f(x) - p^*$ 的估计：

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

- 等于牛顿步的二次 Hessian 范数：

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

- 牛顿方向的方向导数：$\nabla f(x)^T \Delta x_{\mathsf{nt}} = -\lambda(x)^2$
- 仿射不变性

阻尼牛顿法

思想：不采取单位步长，而使用线搜索确定步长

给定一个起始点 $x \in \mathbf{dom}\, f$，容差 $\epsilon > 0$。重复以下步骤：

1. 计算牛顿步和减量

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

2. 停止准则：如果 $\lambda^2/2 \leq \epsilon$，则退出
3. 线搜索：通过回溯线搜索选择步长 t
4. 更新：$x := x + t\Delta x_{\mathsf{nt}}$

线搜索

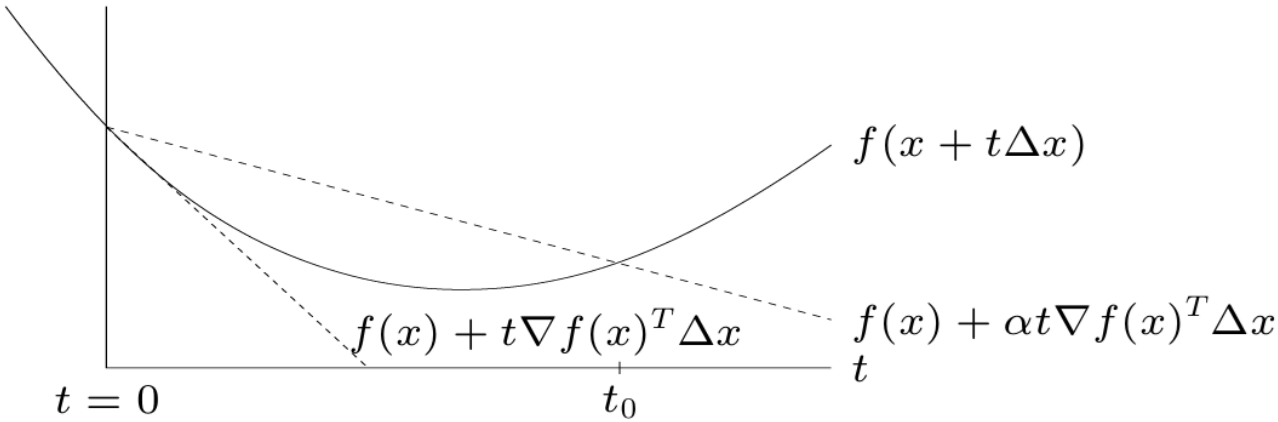
牛顿方向对于凸函数为**下降方向**：t 足够小时，有 $f(x+t\Delta x) < f(x)$

回溯线搜索（参数 $\alpha \in (0, 1/2)$，$\beta \in (0, 1)$）

- 从 $t = 1$ 开始，重复 $t := \beta t$ 直到

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

- 图形解释：回溯直到 $t \leq t_0$



经典收敛性分析

假设

- f 在定义域 S 上为强凸（常数为 m），且 f 为 M-光滑
- $\nabla^2 f$ 在 S 上 Lipschitz 连续，常数为 $L > 0$：

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

（L 衡量 f 可以用二次函数近似的程度）

结论：存在常数 $\eta \in (0, m^2/L)$，$\gamma > 0$，使得

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

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

收敛阶段分析

阻尼牛顿阶段（$\|\nabla f(x)\|_2 \geq \eta$）

- 大多数迭代需要回溯步骤
- 函数值至少减少 γ
- 如果 $p^* > -\infty$，则此阶段最多在 $(f(x^{(0)}) - p^*)/\gamma$ 次迭代后结束

二次收敛阶段（$\|\nabla f(x)\|_2 < \eta$）

- 所有迭代使用单位步长 $t = 1$
- $\|\nabla f(x)\|_2$ 二次收敛到零：如果 $\|\nabla f(x^{(k)})\|_2 < \eta$，则

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

迭代次数

结论：直到 $f(x) - p^* \leq \epsilon$ 的迭代次数

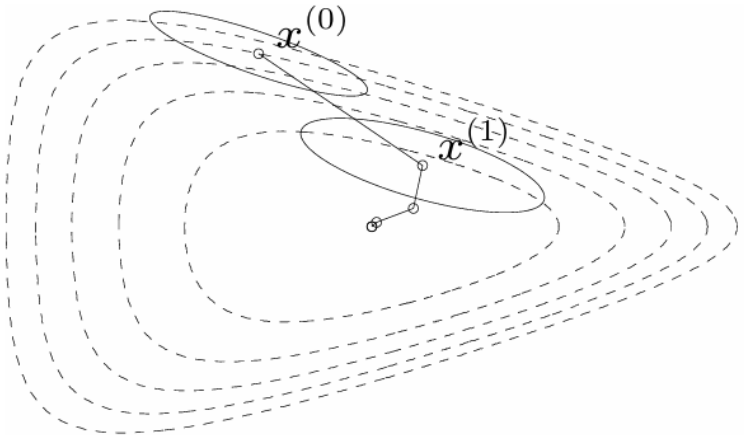
$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2(\epsilon_0/\epsilon)$$

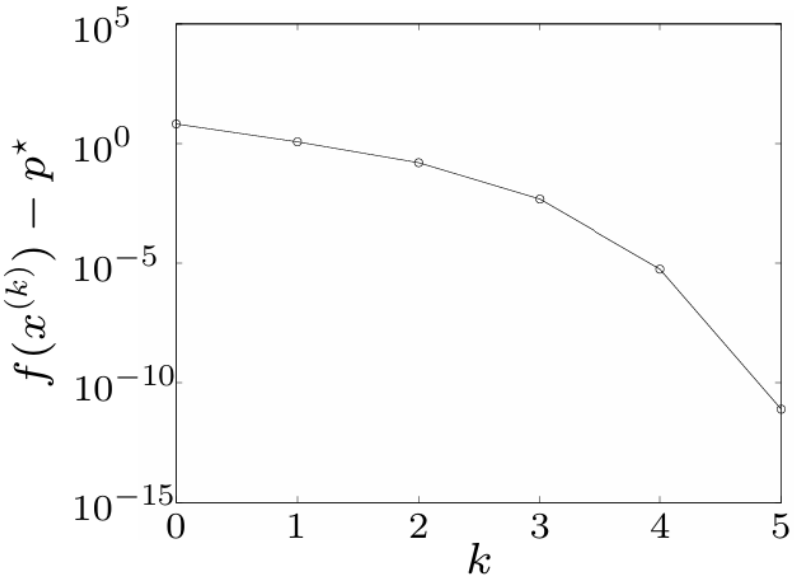
- γ，ϵ_0 是依赖于 m，L，$x^{(0)}$ 的常数
- 第二项很小（数量级为 6），在实际应用中几乎为常数次迭代
- 在实践中，常数 m，L（因此 γ，ϵ_0）通常未知
- 提供了收敛性质的定性分析（即，解释了两个算法收敛阶段）

例子：$x \in \mathbb{R}^2$

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

- 回溯参数 $\alpha = 0.1$, $\beta = 0.7$
- 仅在 5 步内收敛
- 二次局部收敛

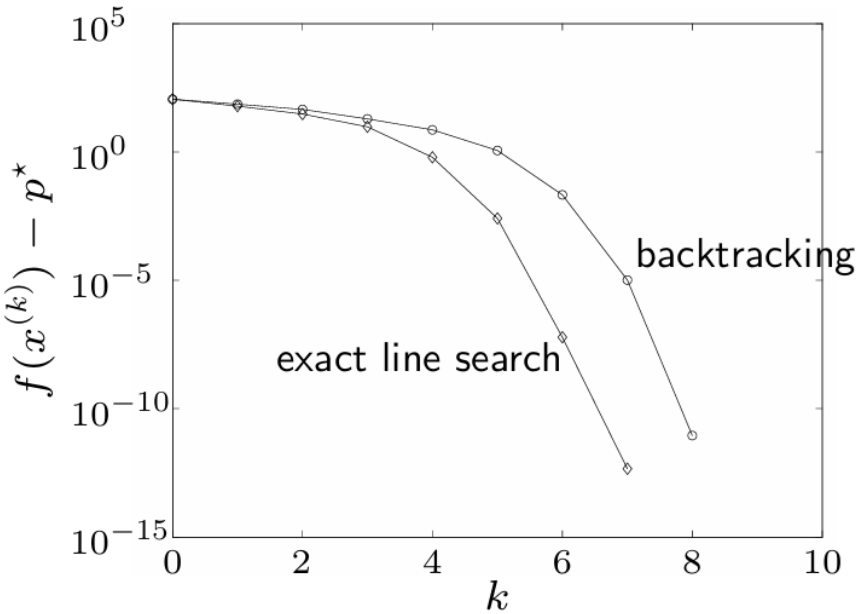


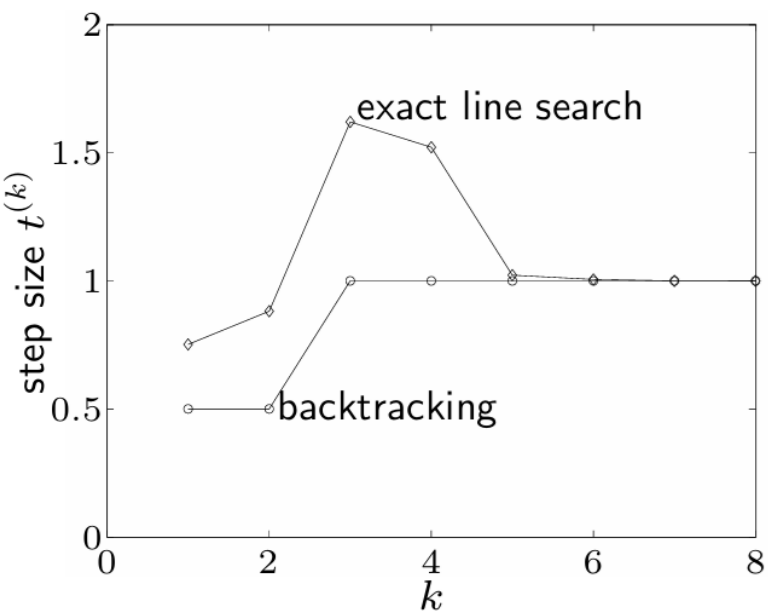


例子：$x \in \mathbb{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$

- 回溯参数 $\alpha = 0.01$, $\beta = 0.5$
- 回溯线搜索几乎和精确线搜索一样快（并且更简单）
- 清楚地显示算法的两个阶段

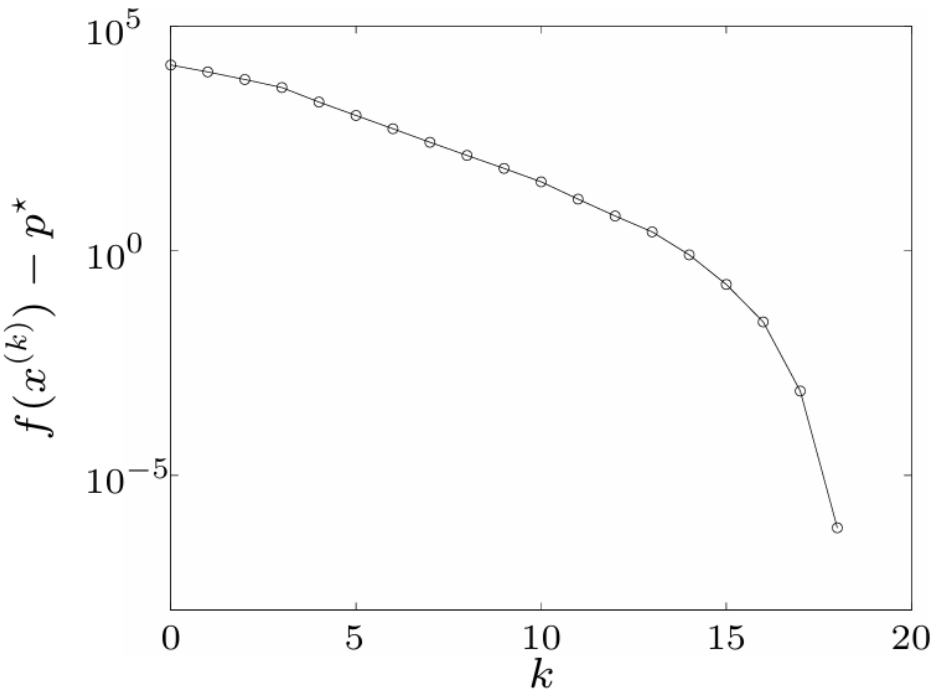




示例在 $\mathbb{R}^{10000}$ 中（具有稀疏 a_i）

$$f(x) = -\sum_{i=1}^{10000} \log(1 - x_i^2) - \sum_{i=1}^{100000} \log(b_i - a_i^T x)$$

- 回溯参数 $\alpha = 0.01$, $\beta = 0.5$.
- 性能与小样本相似



目录

讲　　员

郭加熠，江波，刘慧康

仿射不变性

假设：给定 f，非奇异 $A \in \mathbb{R}^{n\times n}$。设 $x = Ay$，$g(y) = f(Ay)$。

仿射不变性：g 的牛顿迭代结果，恰好为 f 的牛顿迭代结果

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

两边同乘以矩阵 A

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

即，

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

因此，迭代情况与问题放缩（条件数）无关；这在梯度下降中是不成立的

自和谐

经典收敛性分析的缺点

- 依赖于未知常数 $(m, L, \ldots)$
- 收敛上界分析不具有仿射不变性

通过自和谐进行收敛性分析（Nesterov 和 Nemirovski）

- 不依赖于任何未知常数
- 提供仿射不变界限
- 适用于特殊类别的凸函数（“自和谐”函数）
- 在凸优化中，用于分析多项式时间多种算法的收敛性

自和谐函数

定义

- 凸函数 $f:\mathbb{R}\to\mathbb{R}$ 是**自和谐**函数，当 $|f'''(x)|\le 2(f''(x))^{3/2}$ 对所有 $x\in\operatorname{dom} f$ 成立
- 如果 $g(t)=f(x+tv)$ 对所有 $x\in\operatorname{dom} f$，$v\in\mathbb{R}^n$ 是自和谐的，那么 $f:\mathbb{R}^n\to\mathbb{R}$ 是自和谐的

示例

- 线性和二次函数
- 负对数 $f(x)=-\log x$
- $f(X)=-\log\det X$，定义域为 $\mathbb{S}^n_{++}$
- $f(x)=-\sum_{i=1}^m\log(b_i-a_i^Tx)$ 在 $\{x\mid a_i^Tx<b_i,\ i=1,\dots,m\}$ 上

仿射不变性：如果 $f:\mathbb{R}\to\mathbb{R}$ 是自和谐的，那么 $\tilde f(y)=f(ay+b)$ 也是自和谐的：

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

自和谐函数的收敛性分析

总结：存在常数 $\eta \in (0, 1/4]$，$\gamma > 0$，使得

- 如果 $\lambda(x) > \eta$，则 $f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma$
- 如果 $\lambda(x) \leq \eta$，则 $2\lambda(x^{(k+1)}) \leq \left(2\lambda(x^{(k)})\right)^2$

（η 和 γ 仅依赖于回溯参数 α，β）

复杂度界限：牛顿迭代次数由以下界限决定

$$\frac{f(x^{(0)}) - p^*}{\gamma} + \log_2 \log_2(1/\epsilon)$$

对于 $\alpha = 0.1$，$\beta = 0.8$，$\epsilon = 10^{-10}$，界限评估为 $375(f(x^{(0)}) - p^*) + 6$

目录

讲　　员

郭加熠，江波，刘慧康

编程实现

每次迭代的主要计算量：求导数 + 求解牛顿系统

$$H\Delta x = -g$$

其中 $H = \nabla^2 f(x)$，$g = \nabla f(x)$

通过 Cholesky 分解

$$H = LL^T, \quad \Delta x_{\text{nt}} = -L^{-T}L^{-1}g, \quad \lambda(x) = \|L^{-1}g\|_2$$

- 对于非结构化系统，计算量为 $(1/3)n^3$ 次浮点运算
- 如果 H 稀疏，带状，则计算量 $\ll (1/3)n^3$

另一种确定步长的方法：信赖域方法

考虑以下二次优化问题：

$$\begin{array}{ll} \min_d & \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d \\ \text{s.t.} & \|d\|_2 \leq r \end{array}$$

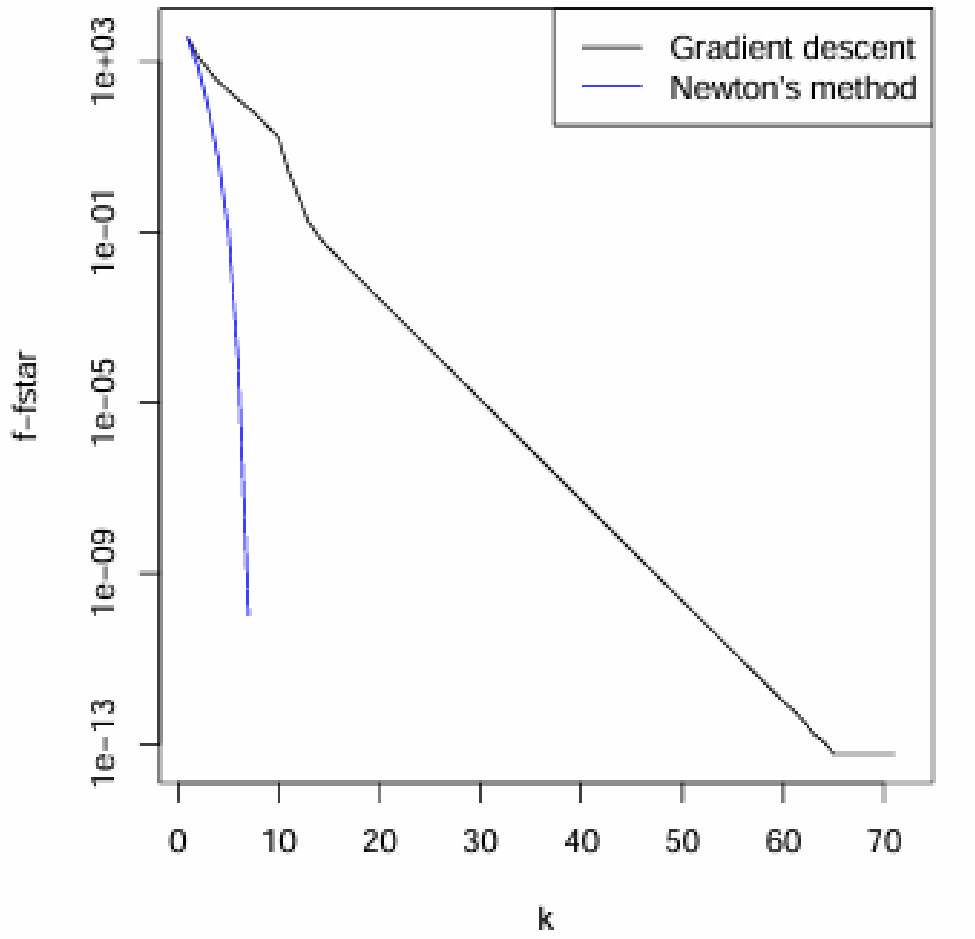
迭代准则

$$x_+ = x + d^*,$$

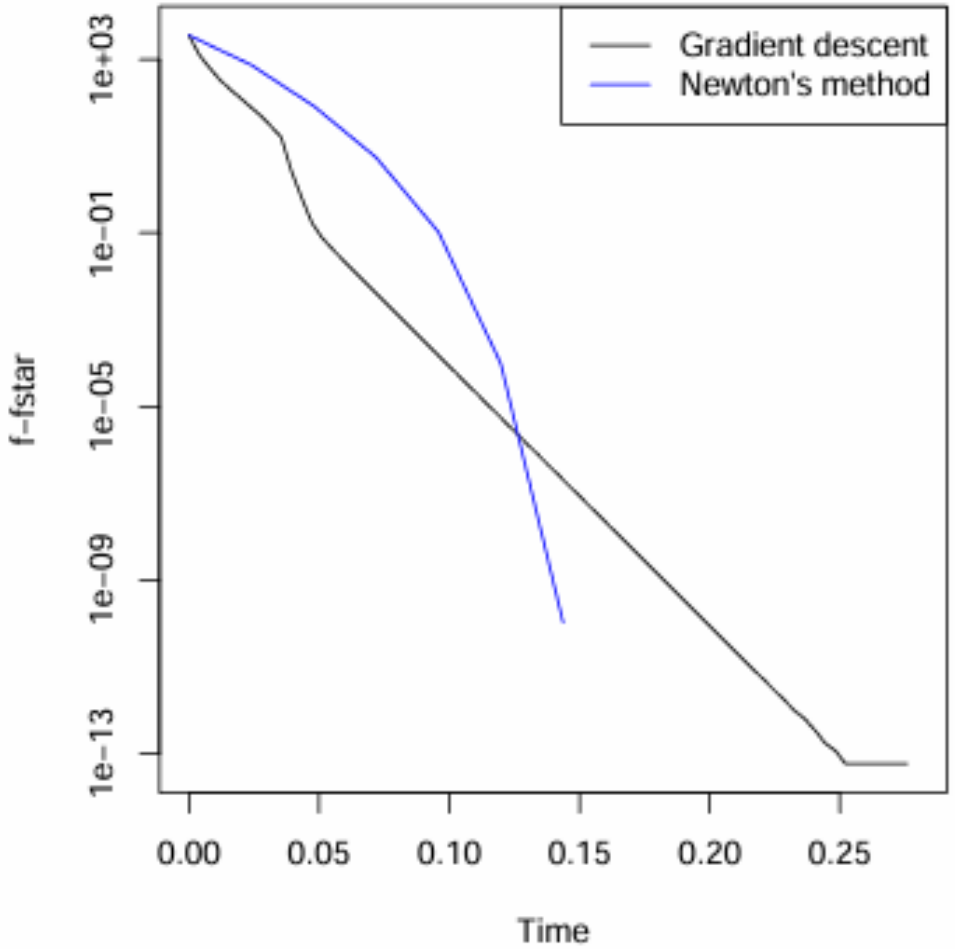
- 以上问题称为信赖域子问题
- 如果 r 很小，则是一种下降方法
- 动态更新 r（例如，效果好就提升 r，效果差就缩小 r）
- 适用于非凸问题

示例：逻辑回归

逻辑回归示例，$n = 500$，$p = 100$：我们比较梯度下降和牛顿法，两者都使用回溯。牛顿法与梯度下降法表现不同。



回到逻辑回归示例：现在 x 轴以每次迭代所需时间为参数



每次梯度下降步骤是 $O(p)$，但每次牛顿步骤是 $O(p^3)$

与一阶方法的比较

特性	牛顿法	梯度下降
内存/迭代	$O(n^2)$	$O(n)$
计算量/迭代	$O(n^3)$	$O(n)$
回溯线搜索	$O(n)$	$O(n)$
局部收敛性	快	慢
条件数	不受影响	影响严重
脆弱性	对 bugs/数值错误更敏感	更稳健