Unconstrained Minimization

Carrson C. Fung
Institute of Electronics
National Chiao Tung University

Closed Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be closed if, for each $\alpha \in \mathbb{R}$, the sublevel set

$$\left\{ \mathbf{x} \in \text{dom } f \middle| f\left(\mathbf{x}\right) \leq \alpha \right\}$$

is closed.

- Equivalent to the condition that the epigraph of f,

$$\text{epi } f = \left\{ \left(\mathbf{x}, t\right) \in \mathbb{R}^{n+1} \middle| \mathbf{x} \in \text{dom } f, f\left(\mathbf{x}\right) \leq t \right\}$$

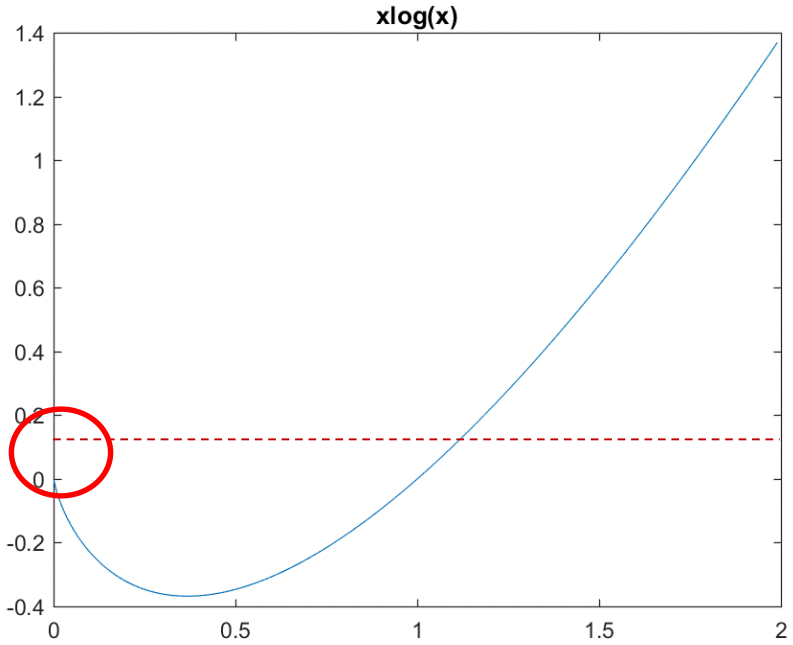
is closed. (This is usually only applied to convex functions.)

- If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, and dom f is closed, then f is closed.
- If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, with dom f open, then f is closed iff f converges to ∞ along every sequence converging to a boundary point of dom f, i.e.

  if $\lim_{i \to \infty} x_i = x \in \text{bd dom } f$, with $x_i \in \text{dom } f$, then $\lim_{i \to \infty} f\left(x_i\right) = \infty$

Examples

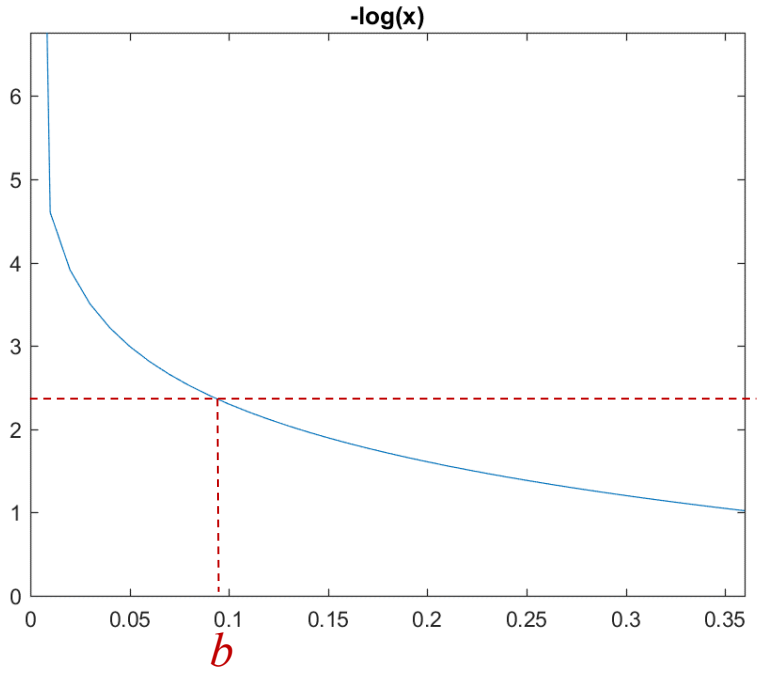
- The function $f : \mathbb{R} \to \mathbb{R},$ with $f\left(x\right) = x \log x$, dom $f = \mathbb{R}_{++}$ is not closed
  Because sublevel set of f is open



Examples

- The function $f : \mathbb{R} \to \mathbb{R}$, with $f(x) = \begin{cases} x \log x, & x > 0 \\ 0 & x = 0 \end{cases}$ $\text{dom } f = \mathbb{R}_+$

  is closed now that we have eliminated the problem in the previous example
- The function $f : \mathbb{R} \to \mathbb{R}$, with $f(x) = -\log x$, $\text{dom } f = \mathbb{R}_{++}$ is closed



The sublevel set contains $x \in [b, \infty)$, and it's closed because its complement $(-\infty, b)$ is open

Unconstrained Minimization Problems

- Will discuss methods for solving the unconstrained optimization problem

$$\min_{\mathbf{x}} f\left(\mathbf{x}\right) \tag{9.1}$$

  where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and twice continuously differentiable (implies that dom f is open)
- Assume problem is solvable, i.e. there exists an optimal point $\mathbf{x}^\star = \inf_{\mathbf{x}} f\left(\mathbf{x}\right)$
- Since f is differentiable and convex, a necessary and sufficient condition for a point $\mathbf{x}^\star$ to be optimal is $\nabla f\left(\mathbf{x}\right) = 0$ (9.2)
- Solving the unconstrained minimization problem (9.1) is the same as finding a solution of (9.2)

Initial Point and Sublevel Set

- The methods described in this chapter require a suitable starting point $\mathbf{x}^{(0)}$. The starting point must lie in dom f and in addition the sublevel set

$$S = \left\{ \mathbf{x} \in \text{dom } f \,\middle|\, f\left(\mathbf{x}\right) \leq f\left(\mathbf{x}^{(0)}\right) \right\} \tag{9.3}$$

  must be closed. This condition is satisfied for all $\mathbf{x}^{(0)} \in$ dom f if the function is closed, i.e. all its sublevel sets are closed.
- The closeness of the sublevel set S is hard to verify, except when all sublevel sets are closed:
  - equivalent to the condition that epi f is closed
  - true if dom $f = \mathbb{R}^n$
  - true if $f\left(\mathbf{x}\right) \to \infty$ as $\mathbf{x} \to$ bd dom f
- Examples of differentiable functions with closed sublevel sets

$$f\left(\mathbf{x}\right) = \log\left(\sum_{i=1}^{m} \exp\left(\mathbf{a}_i^T\mathbf{x} + b_i\right)\right), \quad f\left(\mathbf{x}\right) = -\sum_{i=1}^{m} \log\left(b_i - \mathbf{a}_i^T\mathbf{x}\right)$$

Examples

The general convex quadratic minimization problem has the form

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r, \qquad (9.4)$$

where $\mathbf{P} \in \mathbb{S}_+^n$, $\mathbf{q} \in \mathbb{R}^n$, $r \in \mathbb{R}$

- This problem can be solved via the optimality conditions $\mathbf{P}\mathbf{x}^\star + \mathbf{q} = \mathbf{0}_n$ which is a set of linear equations
  - When $\mathbf{P} \succ 0$, there is a unique solution $\mathbf{x}^\star = -\mathbf{P}^{-1}\mathbf{q}$
  - When $\mathbf{P}$ is not positive definite, any solution $\mathbf{P}\mathbf{x}^\star = -\mathbf{q}$ is optimal
  - If $\mathbf{P}\mathbf{x}^\star = -\mathbf{q}$ does not have a solution, then the problem (9.4) is unbounded below

Example: Least-Squares

Special case of convex quadratic minimization problem is least-squares problem

$$\min_{\mathbf{x}} \left\|\mathbf{A}\mathbf{x}-\mathbf{b}\right\|_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\left(\mathbf{A}^T\mathbf{b}\right)^T\mathbf{x} + \mathbf{b}^T\mathbf{b}$$

where $\mathbf{A}^T\mathbf{A} \in \mathbb{S}_+^n$ because it is a Gram matrix.

- Optimality conditions $\mathbf{A}^T\mathbf{A}\mathbf{x}^\star = \mathbf{A}^T\mathbf{b}$ are called the normal equations of the least-squares problem

Example: Unconstrained Geometric Programming

$$\min_{\mathbf{x}} f\left(\mathbf{x}\right) = \log\left(\sum_{i=1}^{m} \exp\left(\mathbf{a}_i^T \mathbf{x} + b_i\right)\right)$$

- Optimality conditions is

$$\nabla f\left(\mathbf{x}^\star\right) = \frac{1}{\sum_{j=1}^{m} \exp\left(\mathbf{a}_j^T \mathbf{x}^\star + b_j\right)} \sum_{i=1}^{m} \exp\left(\mathbf{a}_i^T \mathbf{x}^\star + b_i\right) \mathbf{a}_i = 0,$$

- Has no analytical solution $\Rightarrow$ resort to iterative algorithm
- For this problem, dom $f = \mathbb{R}^n$, so any point can be chosen as the initial point $\mathbf{x}^{(0)}$

Example: Analytic Center of Linear Inequalities

$$\min_{\mathbf{x}} f(\mathbf{x}) = -\sum_{i=1}^{m} \log\left(b_i - \mathbf{a}_i^T \mathbf{x}\right), \tag{9.5}$$

dom f is the open set $\left\{\mathbf{x} \middle| \mathbf{a}_i^T \mathbf{x} < b_i, i = 1, \ldots, m\right\}$.

- Objective is called logarithmic barrier for $\mathbf{a}_i^T \mathbf{x} \leq b_i$
- Solution, if exists, is called the analytic center of the inequalities
- Initial point $\mathbf{x}^{(0)}$ must satisfy the strict inequalities $\mathbf{a}_i^T \mathbf{x} < b_i, i = 1, \ldots, m$

Example: Analytic Center of Linear Matrix Inequalities

$$\min_{\mathbf{x}} f(\mathbf{x}) = \log\det F(\mathbf{x})^{-1} \qquad (9.6)$$

$F : \mathbb{R}^n \to \mathbb{S}^p$ is affine, i.e. $F(\mathbf{x}) = F_0 + x_1 F_1 + \cdots + x_n F_n$

with $F_i \in \mathbb{S}^P$. dom $f = \{\mathbf{x} \mid F(\mathbf{x}) \succ 0\}$.

- Objective is called logarithmic barrier for the linear matrix inequality $F(\mathbf{x}) \succeq 0$
- Solution, if exists, is called the analytic center of the linear matrix inequality
- Initial point $\mathbf{x}^{(0)}$ must satisfy the strict inequalities $F(\mathbf{x}^{(0)}) \succ 0$.

Strong Convexity and Implications

- The objective function is strongly convex on S, which means that there exists an $m > 0$ such that
$$\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I} \tag{9.7}$$
- Strong convexity has several interesting consequnces. For $\mathbf{x}, \mathbf{y} \in S$, we have
$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x})$$
for some $\mathbf{z}$ on the line segment $[\mathbf{x}, \mathbf{y}]$.
- According to the strong convexity condition in (9.7), we have
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \tag{9.8}$$
for all $\mathbf{x}$ and $\mathbf{y}$ in S

Strong Convexity and Implications

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{m}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \quad (9.8)$$

- Inequality (9.8) can be used to bound $f(\mathbf{x}) - p^\star$
- Righthand side of (9.8) is a convex quadratic function of $\mathbf{y}$ (for fixed $\mathbf{x}$). Set gradient w.r.t. $\mathbf{y}$ to zero $\Rightarrow$ $\tilde{\mathbf{y}} = \mathbf{x} - \frac{1}{m} \nabla f(\mathbf{x})$ minimizes the righthand side

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{m}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \\ &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\tilde{\mathbf{y}} - \mathbf{x}) + \frac{m}{2} \|\tilde{\mathbf{y}} - \mathbf{x}\|_2^2 \\ &= f(\mathbf{x}) - \frac{1}{2m} \|\nabla f(\mathbf{x})\|_2^2 . \end{aligned}$$

- This holds for any $\mathbf{y} \in S$

$$\Rightarrow \quad \boxed{p^\star \geq f(\mathbf{x}) - \frac{1}{2m} \|\nabla f(\mathbf{x})\|_2^2 .} \quad (9.9)$$

$\Rightarrow$ If gradient is small at a point, then the point is nearly optimal.

Strong Convexity and Implications

$$p^\star \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|_2^2. \quad (9.9)$$

(9.9) can also be interpreted as a condition for *suboptimality* (recall from [Ch. 4.1,CLL17], this is known as ϵ-suboptimal) which generalizes the optimality condition (9.2):

$$\text{if } \left\|\nabla f(\mathbf{x})\right\|_2 \leq (2m\epsilon)^{1/2} \quad \Rightarrow \quad p^* \geq f(\mathbf{x}) - \epsilon \quad \Rightarrow \quad f(\mathbf{x}) - p^\star \leq \epsilon \qquad (9.10)$$

Strong Convexity and Implications

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{m}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \quad (9.8)$$

- We can also derive a bound on $\|\mathbf{x} - \mathbf{x}^\star\|_2$, the distance between $\mathbf{x}$ and any optimal point $\mathbf{x}^\star$, in terms of $\|\nabla f(\mathbf{x})\|_2$ given by

$$\|\mathbf{x} - \mathbf{x}^\star\|_2 \leq \frac{2}{m} \|\nabla f(\mathbf{x})\|_2 . \quad (9.11)$$

- To see this, apply (9.8) with $\mathbf{y} = \mathbf{x}^\star$ to obtain

$$\begin{aligned} p^\star = f(\mathbf{x}^\star) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{x}^\star - \mathbf{x}) + \frac{m}{2} \|\mathbf{x}^\star - \mathbf{x}\|_2^2 \\ &\geq f(\mathbf{x}) - \|\nabla f(\mathbf{x})\|_2 \|\mathbf{x}^\star - \mathbf{x}\|_2 + \frac{m}{2} \|\mathbf{x}^\star - \mathbf{x}\|_2^2 , \end{aligned}$$

(CSI is used in 2nd inequality). Since $p^\star \leq f(\mathbf{x})$ then

$$0 \geq p^\star - f(\mathbf{x}) \geq -\|\nabla f(\mathbf{x})\|_2 \|\mathbf{x}^\star - \mathbf{x}\|_2 + \frac{m}{2} \|\mathbf{x}^\star - \mathbf{x}\|_2^2$$

Then we have (9.11)

Strong Convexity and Implications

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{m}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \quad (9.8)$$

Note that the maximum eigenvalue of $\nabla^2 f(\mathbf{x})$, which is a continuous function of $\mathbf{x}$ on S, is bounded above on S, i.e. there exists a constant M such that

$$\nabla^2 f(\mathbf{x}) \preceq M\mathbf{I} \quad (9.12)$$

for all $\mathbf{x} \in S$. This upper bound on the Hessian implies for any $\mathbf{x}, \mathbf{y} \in S$,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) + \frac{M}{2} \|\mathbf{y} - \mathbf{x}\|_2^2, \quad (9.13)$$

which is analogous to (9.8). Minimizing each side over $\mathbf{y}$ yields

$$p^\star \leq f(\mathbf{x}) - \frac{1}{2M} \|\nabla f(\mathbf{x})\|_2^2, \quad (9.14)$$

the counterpart of (9.9).

Condition Number of Sublevel Sets

- From the strong convexity inequality (9.7) and the inequality (9.12), we have

$$m\mathbf{I} \preceq \nabla^2 f\left(\mathbf{x}\right) \preceq M\mathbf{I} \tag{9.15}$$

for all $\mathbf{x} \in S$. The ratio $\kappa = M \,/\, m$ is thus an upper bound on the *condition number* of the matrix of the matrix $\nabla^2 f\left(\mathbf{x}\right)$, i.e. the ratio of its largest eigenvalue to its smallest eigenvalue

- Geometrical interpretation of (9.14): We define the width of a convex set $C \subseteq \mathbb{R}^n$, in the direction $\mathbf{q}$, where $\left\|\mathbf{q}\right\|_2 = 1$, as

$$W\left(C, \mathbf{q}\right) = \sup_{\mathbf{z} \in C} \mathbf{q}^T \mathbf{z} - \inf_{\mathbf{z} \in C} \mathbf{q}^T \mathbf{z}.$$

- The minimum width and maximum width of C are given by

$$W_{\min} = \inf_{\left\|\mathbf{q}\right\|_2 = 1} W\left(C, \mathbf{q}\right) \qquad W_{\max} = \sup_{\left\|\mathbf{q}\right\|_2 = 1} W\left(C, \mathbf{q}\right).$$

Condition Number of Sublevel Sets

- The condition number of the convex set C is defined as

$$\text{cond}\left(C\right) = \frac{W_{\max}^2}{W_{\min}^2},$$

  i.e. the square of the ratio of its maximum width to its minimum width
- The condition number of C gives a measure of its anisotropy or eccentricity
  - If the condition number of a set C is small, say near one, it means that the set has approximately the same width in all directions, i.e., it is nearly spherical
  - If the condition number is large, it means that the set is far wider in some directions than in others

Example 9.1: Condition Number of Ellipsoid

Let $\mathcal{E}$ be the ellipsoid $\mathcal{E} = \left\{ \mathbf{x} \middle| \left(\mathbf{x} - \mathbf{x}_0 \right)^T \mathbf{A}^{-1} \left(\mathbf{x} - \mathbf{x}_0 \right) \le 1 \right\}$, where $\mathbf{A} \in \mathbb{S}^n_{++}$. The width of $\mathcal{E}$ in the direction of $\mathbf{q}$ is

$$\begin{aligned} \sup_{\mathbf{z} \in \mathcal{E}} \mathbf{q}^T \mathbf{z} - \inf_{\mathbf{z} \in \mathcal{E}} \mathbf{q}^T \mathbf{z} &= \left(\left\| \mathbf{A}^{1/2} \mathbf{q} \right\|_2 + \mathbf{q}^T \mathbf{x}_0 \right) - \left(- \left\| \mathbf{A}^{1/2} \mathbf{q} \right\|_2 + \mathbf{q}^T \mathbf{x}_0 \right). \\ &= 2 \left\| \mathbf{A}^{1/2} \mathbf{q} \right\|_2 . \end{aligned}$$

It follows that its minimum and maximum width are

$$W_{\min} = 2\lambda_{\min} \left(\mathbf{A}^{1/2} \right), \qquad W_{\max} = 2\lambda_{\max} \left(\mathbf{A}^{1/2} \right),$$

and its condition number is

$$\operatorname{cond} \left(\mathcal{E} \right) = \frac{\lambda_{\max} \left(\mathbf{A} \right)}{\lambda_{\min} \left(\mathbf{A} \right)} = \kappa \left(\mathbf{A} \right),$$

where $\kappa \left(\mathbf{A} \right)$ denotes the condition number of the matrix $\mathbf{A}$, i.e. the ratio of its maximum singular value to its minimum singular value. Thus the condition number of $\mathcal{E}$ is the same as the condition number of $\mathbf{A}$ that defines it.

Condition number of Sublevels Sets

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \quad (9.8)$$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y}-\mathbf{x}) + \frac{M}{2}\|\mathbf{y}-\mathbf{x}\|_2^2, \quad (9.13)$$

Suppose f satisfies $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}$, $\forall \mathbf{x} \in \mathbb{S}$. We will derive a bound on the condition number of the α-sublevel $C_\alpha = \{\mathbf{x} | f(\mathbf{x}) \leq \alpha\}$, where $p^\star < \alpha \leq f(\mathbf{x}^{(0)})$.

- Applying (9.8) and (9.13) with $\mathbf{x} = \mathbf{x}^\star$, and using (9.2) ($\nabla f(\mathbf{x}^\star) = 0$), we have

$$p^\star + \frac{M}{2}\|\mathbf{y}-\mathbf{x}^\star\|_2^2 \geq f(\mathbf{y}) \geq p^\star + \frac{m}{2}\|\mathbf{y}-\mathbf{x}^\star\|_2^2.$$

Implies that $B_{\text{inner}} \subseteq C_\alpha \subseteq B_{\text{outer}}$, where

$$B_{\text{inner}} = \left\{ \mathbf{y} \,\middle|\, \|\mathbf{y}-\mathbf{x}^\star\|_2 \leq \left(\frac{2(\alpha - p^\star)}{M} \right)^{1/2} \right\},$$

$$B_{\text{outer}} = \left\{ \mathbf{y} \,\middle|\, \|\mathbf{y}-\mathbf{x}^\star\|_2 \leq \left(\frac{2(\alpha - p^\star)}{m} \right)^{1/2} \right\}.$$

The ratio of the radii squared gives an upper bound on the condition number of C_α:

$$\text{cond}(C_\alpha) \leq \frac{M}{m}.$$

Condition Number of Sublevel Sets

Can also give a geometric interpretation of the condition number $\kappa\left(\nabla^2 f\left(\mathbf{x}^\star\right)\right)$ of the Hessian at the optimum. From Taylor series expansion of f around $\mathbf{x}^\star$ (and using (9.2) again)

$$f(\mathbf{y}) \approx p^\star + \frac{1}{2}\left(\mathbf{y}-\mathbf{x}^\star\right)^T \nabla^2 f\left(\mathbf{x}^\star\right)\left(\mathbf{y}-\mathbf{x}^\star\right),$$

we see that, for α close to $p^\star$,

$$C_\alpha \approx \left\{\mathbf{y} \,\middle|\, \left(\mathbf{y}-\mathbf{x}^\star\right)^T \nabla^2 f\left(\mathbf{x}^\star\right)\left(\mathbf{y}-\mathbf{x}^\star\right) \le 2\left(\alpha - p^\star\right)\right\},$$

i.e. the sublevel set is well approximated by an ellipsoid with center $\mathbf{x}^\star$. Hence

$$\lim_{\alpha \to p^\star} \operatorname{cond}\left(C_\alpha\right) = \kappa\left(\nabla^2 f\left(\mathbf{x}^\star\right)\right).$$

Will see that condition number of the sublevel set of f (which is bounded by M/m) has a strong effect on the efficiency of some common methods for unconstrained minimization.

Descent Methods

- The algorithms described in this chapter produce a minimizing sequence $\mathbf{x}^{(k)}$, $k = 1, \ldots,$ where

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)} \Delta \mathbf{x}^{(k)}$$

  and $t^{(k)} > 0$ (except when $x^{(k)}$ is optimal).
- $\Delta \mathbf{x}^{(k)} \in \mathbb{R}^n$ is called the step or search direction (even though it need not have unit norm). The scalar $t^{(k)} \geq 0$ is called the step size or step length at iteration k
- When we focus on one iteration of an algorithm, we sometimes drop the supercripts and use the lighter notation $\mathbf{x}^+ = \mathbf{x} + t\Delta \mathbf{x}$ or $\mathbf{x} := \mathbf{x} + t\Delta \mathbf{x}$ in place of $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)} \Delta \mathbf{x}^{(k)}$.
- All the methods we study are *descent methods*, which means that

$$f\left(\mathbf{x}^{(k+1)}\right) < f\left(\mathbf{x}^{(k)}\right),$$

  except when $\mathbf{x}^{(k)}$ is optimal.

Descent Methods

- From convexity we know that $\nabla f\left(\mathbf{x}^{(k)}\right)^T\left(\mathbf{y}-\mathbf{x}^{(k)}\right)\geq 0$ implies $f\left(\mathbf{y}\right)\geq f\left(\mathbf{x}^{(k)}\right)$ because of 1st-order optimality condition $f\left(\mathbf{y}\right)\geq f\left(\mathbf{x}\right)+\nabla f\left(\mathbf{x}\right)^T\left(\mathbf{y}-\mathbf{x}\right)$, so the search direction in a descent method must satisfy

  $$\nabla f\left(\mathbf{x}\right)^T\Delta\mathbf{x}^{(k)}<0,$$

  plays the role of $f\left(x^{(k+1)}\right)$

  plays the role of $\Delta\mathbf{x}^{(k)}$

  we call such a direction a *descent direction*.

  and $t^{(k)}>0$ (except when $x^{(k)}$ is optimal).

- The outline of a general descent method is as follows. It alternates between two steps: determining a descent direction $\Delta\mathbf{x}$, and the selection of a step size t.

Algorithm 9.1 *General descent method.*

given a starting point $x\in\mathbf{dom}\,f$.

repeat

1. Determine a descent direction Δx.
2. *Line search.* Choose a step size $t>0$.
3. *Update.* $x:=x+t\Delta x$.

until stopping criterion is satisfied.

Descent Methods

- Exact line search

  One line search method sometimes used in practice is exact line search, in which t is chosen to minimize f along the line (ray) $\{\mathbf{x}+t\Delta\mathbf{x} \mid t \geq 0\}$:

$$t = \arg\min_{s \geq 0} f(\mathbf{x}+s\Delta\mathbf{x}). \qquad (9.16)$$

  An exact line search is used when the cost of the minimization problem with one variable is low compared to the cost of computing the search direction itself.

- Backtracking line search
  - Most line searches used in practice are inexact: the step length is chosen to approximately minimize f along the ray $\{\mathbf{x}+t\Delta\mathbf{x} \mid t \geq 0\}$, or even to just reduce f "enough"
  - One inexact line search method that is very simple and quite effective is called backtracking line search. It depends on two constants α, β with $0 < \alpha < 0.5$, $0 < \beta < 1$

Descent Methods

Algorithm 9.2 *Backtracking line search.*

given a descent direction Δx for f at $x \in \mathbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.
$t := 1$.
while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, $t := \beta t$.

- The line search is called backtracking because it starts with unit step size and then reduces it by the factor β until the stopping criterion

Linear approximation of $f(\mathbf{x}+t\Delta\mathbf{x})$ with preassigned α

$$f(\mathbf{x}+t\Delta\mathbf{x}) \le f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x}$$

holds. Since $\Delta\mathbf{x}$ is a descent direction, we have $\nabla f(\mathbf{x})^T \Delta\mathbf{x} < 0$, so for small enought t we have

$$f(\mathbf{x}+t\Delta\mathbf{x}) \approx f(\mathbf{x}) + t\nabla f(\mathbf{x})^T \Delta\mathbf{x} \le f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x},$$

which shows that the backtracking line search eventually terminates.

Descent Methods

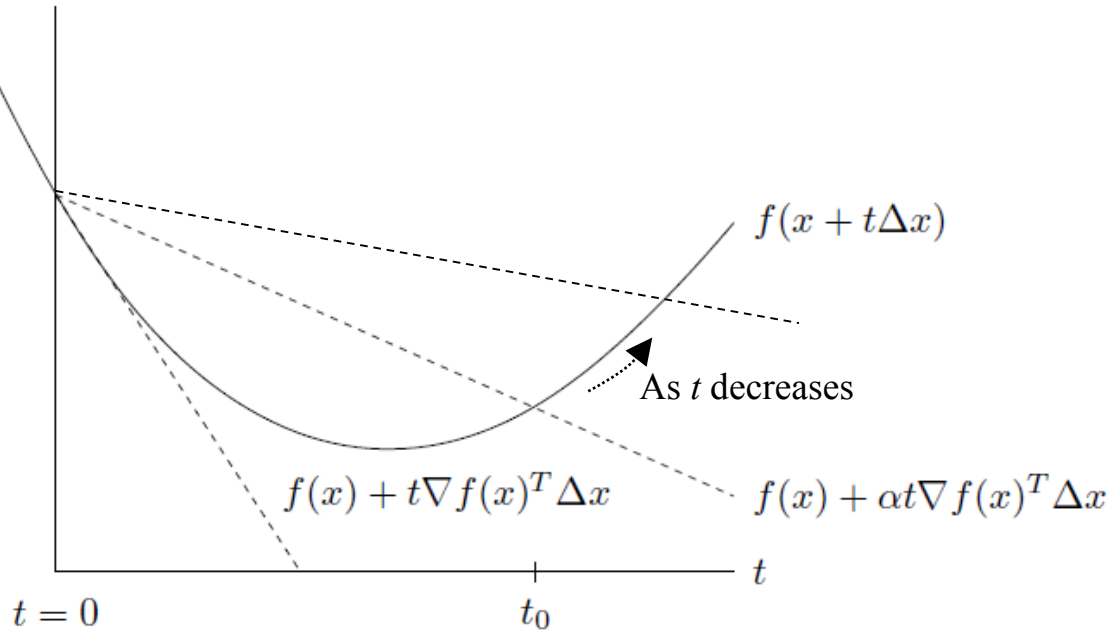
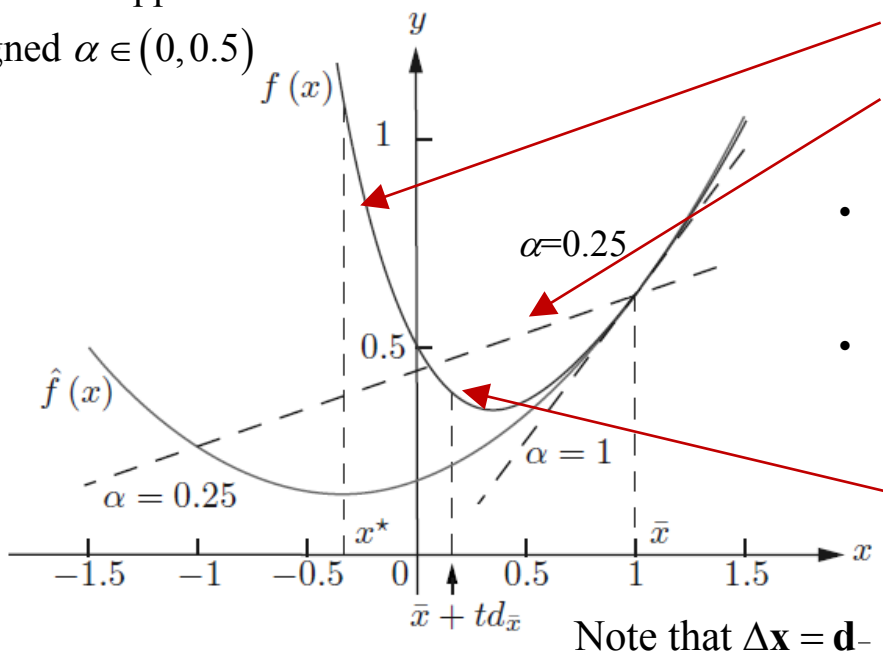


Figure 9.1 *Backtracking line search.* The curve shows f, restricted to the line over which we search. The lower dashed line shows the linear extrapolation of f, and the upper dashed line has a slope a factor of α smaller. The backtracking condition is that f lies below the upper dashed line, *i.e.*, $0 \leq t \leq t_0$.

Backtracking Line Search

keep decreasing t by βt until $f\left(\bar{x}+td_{\bar{x}}\right)$ is below the linear approximation for the preassigned $\alpha \in (0,0.5)$



$f\left(x\right)=e^{-2x}+x-0.5$

Linear approx: $f\left(\bar{x}\right)+\alpha f'\left(\bar{x}\right)\left(x-\bar{x}\right)$

- α describes to what extent the function f is decreased as predicted by linear approximation.
- β determines the accuracy of the backtracking line search

$f\left(\bar{x}+td_{\bar{x}}\right)$

Note that $\Delta \mathbf{x}=\mathbf{d}_{\bar{\mathbf{x}}}$

Figure 10.1 Illustration of backtracking line search by Algorithm 10.1, where $f(x)$ given by (10.14) (blue solid line), the associated quadratic convex function $\hat{f}$ given by (10.7) for $\bar{x}=1$ (red solid line), and the associated linear approximations (dashed lines, one for $\alpha=0.25$ and one for $\alpha=1$) given by (10.15) are depicted. For this case, $x^{\star}=-0.35$ (the optimal solution of $\hat{f}$) and $d_{\bar{x}}=x^{\star}-\bar{x}=-1.35$; an admissible point $\bar{x}+td_{\bar{x}}$ for $t=0.64$ is also indicated by an arrow.

Gradient Descent Method

- A natural choice for the search direction is the negative gradient $\Delta \mathbf{x} = -\nabla f(\mathbf{x})$. The resulting algorithm is called the gradient algorithm or **gradient descent method**.

Algorithm 9.3 *Gradient descent method.*

given a starting point $x \in \mathbf{dom}\, f$.

repeat

1. $\Delta x := -\nabla f(x)$.
2. *Line search.* Choose step size t via exact or backtracking line search.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

- From (9.9), the stopping criterion is usually of the form $\|\nabla f(\mathbf{x})\|_2 \leq \eta$, where η is small and positive.

Gradient Descent Method

- **Convergence analysis**: here we present simple convergence analysis for the gradient method.
  - using the lighter notation $\mathbf{x}^+ = \mathbf{x} + t\Delta\mathbf{x}$, where $\Delta\mathbf{x} = -\nabla f(\mathbf{x})$
  - Assume f is strongly convex on S so that $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}, \ \forall \mathbf{x} \in S$
  - Define the function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by $\tilde{f}(t) = f(\mathbf{x} - t\nabla f(\mathbf{x}))$

With $\mathbf{y} = \mathbf{x} - t\nabla f(\mathbf{x})$, from (9.13) ($f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{M}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$)

then we obtain a quadratic upper bound on $\tilde{f}$:

$$\begin{aligned}\tilde{f}(t) &\le f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{M}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x} - t\nabla f(\mathbf{x}) - \mathbf{x}) + \frac{M}{2}\|\mathbf{x} - t\nabla f(\mathbf{x}) - \mathbf{x}\|_2^2 \\ &= f(\mathbf{x}) - t\|\nabla f(\mathbf{x})\|_2^2 + \frac{Mt^2}{2}\|\nabla f(\mathbf{x})\|_2^2 \end{aligned} \quad (9.17)$$

Gradient Descent Method

$$\tilde{f}(t) \le f(\mathbf{x}) - t\left\|\nabla f(\mathbf{x})\right\|_2^2 + \frac{Mt^2}{2}\left\|\nabla f(\mathbf{x})\right\|_2^2 \quad (9.17)$$

Analysis for exact line search: assume that an exact line search is used, and minmize over t on both sides of the inequality (9.17).

- On the lefthand side we get $\tilde{f}\left(t_{\text{exact}}\right)$, where t_{exact} is the step length that minimizes $\tilde{f}$.
- The righthand side is a simple quadratic, which is minimized by $t = 1/M$ $\left(\frac{d\text{RHS of (9.17)}}{dt} = 0\right)$ and has minimum value

$$f(\mathbf{x}) - \frac{1}{M}\left\|\nabla f(\mathbf{x})\right\|_2^2 + \frac{1}{2M}\left\|\nabla f(\mathbf{x})\right\|_2^2 = f(\mathbf{x}) - \frac{1}{2M}\left\|\nabla f(\mathbf{x})\right\|_2^2.$$

Gradient Descent Method

$$p^\star \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|_2^2 \qquad (9.9)$$

There we have

$$f(\mathbf{x}^+) = \tilde{f}(t_{\text{exact}}) \leq f(\mathbf{x}) - \frac{1}{2M}\left\|\nabla f(\mathbf{x})\right\|_2^2.$$

Substracting $p^\star$ from both sides, we get

$$f(\mathbf{x}^+) - p^\star \leq f(\mathbf{x}) - p^\star - \frac{1}{2M}\left\|\nabla f(\mathbf{x})\right\|_2^2.$$

Combine this with $\left\|\nabla f(\mathbf{x})\right\|_2^2 \geq 2m\left(f(\mathbf{x}) - p^\star\right)$ (which follows from (9.9)) to conclude

$$\begin{aligned} f(\mathbf{x}^+) - p^\star \leq f(\mathbf{x}) - p^\star - \frac{1}{2M}\left\|\nabla f(\mathbf{x})\right\|_2^2 &\leq f(\mathbf{x}) - p^\star - \frac{m}{M}\left(f(\mathbf{x}) - p^\star\right) \\ &= \left(1 - \frac{m}{M}\right)\left(f(\mathbf{x}) - p^\star\right). \end{aligned}$$

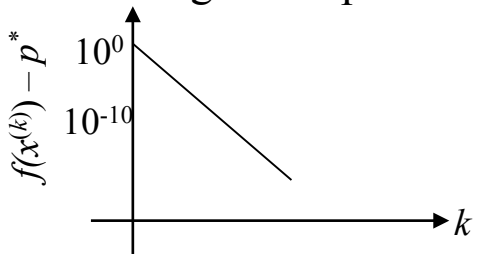
Apply this inequality recursively, we find that

$$f(\mathbf{x}^{(k)}) - p^\star \leq c^k\left(f(\mathbf{x}^{(0)}) - p^\star\right) \qquad (9.18)$$

where $c = 1 - \frac{m}{M} < 1$, which shows that $f(\mathbf{x}^{(k)})$ converge to $p^\star$ as $k \to \infty$

Gradient Descent Method

- Recall from (9.15), $\kappa = M / m$ is an upper bound on the condition number of the Hessian $\nabla^2 f(\mathbf{x})$. Since $c^k = 1 - m/M < 1$, hence the number of iterations required increases approximately linearly with increasing κ. In other words, number of iterations increases when $\nabla^2 f(\mathbf{x})$, near $\mathbf{x}^\star$, has a large κ.
- Conversely, when the sublevel sets of f are relatively isotropic, so that the condition number bound M / m can be chosen to be relatively small, (9.18) shows that convergence is rapid, since c is small, or at least not too close to one. (c.f. $\lim_{\alpha \to p^\star} \text{cond}(C_\alpha) = \kappa\left(\nabla^2 f(\mathbf{x}^\star)\right)$)
- (9.18)(shows that the error $f(\mathbf{x}^{(k)}) - p^\star$ converges to zero at least as fast as geometric series. In the context of iterative numerical methods, this is called *linear convergence*, since the error lies below a line on a log-linear plot of error v.s. iteration number.



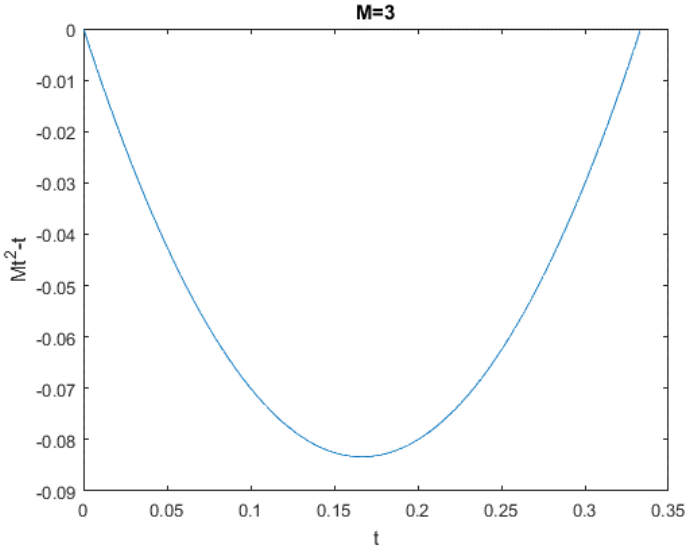
Gradient Descent Method

Analysis for backtracking line search: Consider the case where a backtracking line search is used in the gradient descent method. Will show that the backtracking exit condition

$$\tilde{f}\left(t\right) \le f\left(\mathbf{x}\right) - \alpha t \left\| \nabla f\left(\mathbf{x}\right) \right\|_2^2, \quad \left(\Delta \mathbf{x} = -\nabla f\left(\mathbf{x}\right)\right)$$

is satisfied whenever $0 \le t \le 1/M$. First note that

$$0 \le t \le 1/M \quad \Rightarrow \quad Mt^2 - t \le 0 \quad \Leftrightarrow \quad -t + \frac{Mt^2}{2} \le -\frac{t}{2}$$



Gradient Descent Method

$$\tilde{f}(t) \le f(\mathbf{x}) - t\left\|\nabla f(\mathbf{x})\right\|_2^2 + \frac{Mt^2}{2}\left\|\nabla f(\mathbf{x})\right\|_2^2 \quad (9.17)$$

Using this results and (9.17), for $0 \le t \le 1/M$,

$$\begin{aligned}\tilde{f}(t) &\le f(\mathbf{x}) - t\left\|\nabla f(\mathbf{x})\right\|_2^2 + \frac{Mt^2}{2}\left\|\nabla f(\mathbf{x})\right\|_2^2 \\ &\le f(\mathbf{x}) - \left(\frac{t}{2}\right)\left\|\nabla f(\mathbf{x})\right\|_2^2 \\ &\le f(\mathbf{x}) - \alpha t\left\|\nabla f(\mathbf{x})\right\|_2^2,\end{aligned}$$

since $\alpha < 1/2$.

Gradient Descent Method

$$p^\star \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|_2^2 \qquad (9.9)$$

Therefore the backtracking line search terminates either with $t = 1$ or with a value $t \geq \frac{\beta}{M}$.

This provides a lower bound on the decrease in the objective function.

$$t = 1:\ f(\mathbf{x}^+) \leq f(\mathbf{x}) - \alpha\left\|\nabla f(\mathbf{x})\right\|_2^2, \qquad t = \frac{\beta}{M}:\ f(\mathbf{x}^+) \leq f(\mathbf{x}) - \alpha\frac{\beta}{M}\left\|\nabla f(\mathbf{x})\right\|_2^2$$

Then we have $$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \min\left\{\alpha, \frac{\alpha\beta}{M}\right\}\left\|\nabla f(\mathbf{x})\right\|_2^2.$$

Now can proceed exactly as in the case of exact line search. Subtract $p^\star$ from both sides:

$$f(\mathbf{x}^+) - p^\star \leq f(\mathbf{x}) - p^\star - \min\left\{\alpha, \frac{\alpha\beta}{M}\right\}\left\|\nabla f(\mathbf{x})\right\|_2^2.$$

Using (9.9): $\left\|\nabla f(\mathbf{x})\right\|_2^2 \geq 2m\left(f(\mathbf{x}) - p^\star\right)$, the above becomes

$$f(\mathbf{x}^+) - p^\star \leq f(\mathbf{x}) - p^\star - \min\left\{\alpha, \frac{\alpha\beta}{M}\right\}\left\|\nabla f(\mathbf{x})\right\|_2^2 \leq f(\mathbf{x}) - p^\star - \min\left\{\alpha, \frac{\alpha\beta}{M}\right\}2m\left(f(\mathbf{x}) - p^\star\right)$$

$$= \left(1 - \min\left\{2m\alpha, \frac{2\alpha\beta m}{M}\right\}\right)\left(f(\mathbf{x}) - p^\star\right).$$

Gradient Descent Method

Once again, applying this inequality recursively, we have

$$f\left(\mathbf{x}^{(k)}\right)-p^{\star}\leq c^{k}\left(f\left(\mathbf{x}^{(0)}\right)-p^{\star}\right),$$

with $c=\left(1-\min\left\{2m\alpha,\frac{2\alpha\beta m}{M}\right\}\right)<1.$

- $f\left(\mathbf{x}^{(k)}\right)$ converges to $p^{\star}$ at least as fast as a geometric series with an exponent that depends, at least in part, on the condition bound M / m. Thus, again, the convergence is at least linear.

Example: Quadratic Problem in $\mathbb{R}^2$

Consider $f(\mathbf{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$, where $\gamma > 0$. Clearly, the optimal solution is $\mathbf{x}^\star = \mathbf{0}_2$ and the optimal value is 0.

$$\nabla^2 f(\mathbf{x}) = D\left(\nabla f(\mathbf{x})\right) = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix} \quad \Rightarrow \quad \kappa\left(\nabla^2 f(\mathbf{x})\right) = \frac{M}{m} = \frac{\max\{1,\gamma\}}{\min\{1,\gamma\}} = \max\left\{\gamma, \frac{1}{\gamma}\right\}.$$

Hence the tightest choice for the convexity constants are $m = \min\{1,\gamma\}$ and $M = \max\{1,\gamma\}$

We apply the gradient descent method with exact line search, starting at the point $\mathbf{x}^{(0)} = (\gamma, 1)$.

We can derive the following closed-form expressions for the iterates $\mathbf{x}^{(k)}$ and their function values (see [ex. 9.6, BV04])

$$\mathbf{x}_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \qquad \mathbf{x}_2^{(k)} = \gamma\left(-\frac{\gamma-1}{\gamma+1}\right)^k, \qquad \text{and}$$

$$f\left(\mathbf{x}^{(k)}\right) = \frac{\gamma(\gamma+1)}{2}\left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f\left(\mathbf{x}^{(0)}\right).$$

This is illustrated in Fig. 9.2, for $\gamma = 10$.

Example: Quadratic Problem in $\mathbb{R}^2$

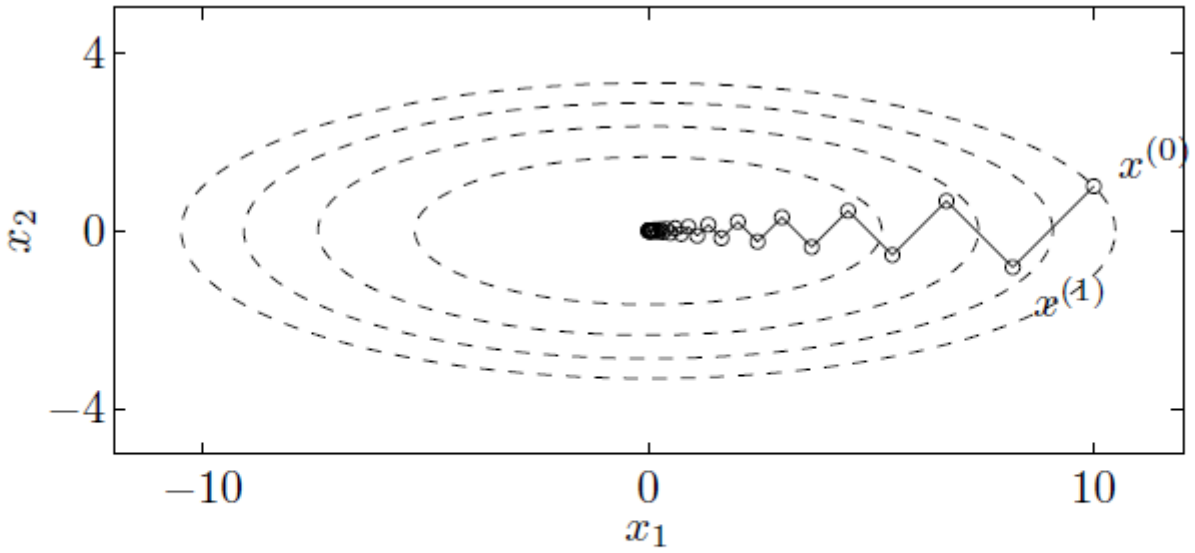


Figure 9.2 Some contour lines of the function $f(x) = (1/2)(x_1^2 + 10x_2^2)$. The condition number of the sublevel sets, which are ellipsoids, is exactly 10. The figure shows the iterates of the gradient method with exact line search, started at $x^{(0)} = (10, 1)$.

Convergence is linear. Error is exactly a geometic series, reduced by a factor of $\left(\frac{\gamma-1}{\gamma+1}\right)^{2k}$ at each iteration. For $\gamma = 1$, exact solution is found in one itreation. Convergence is very slow for $\gamma \gg 1$ or $\ll 1$.

Example: Nonquadratic Problem in $\mathbb{R}^2$

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1} \qquad (9.20)$$

Apply backtracking line search, with $\alpha = 0.1$, $\beta = 0.7$. Fig. 9.3 shows level curves of f, and the iterates $\mathbf{x}^{(k)}$ generated by the gradient method (shown as small circles). The lines connecting successive iterates show the scaled steps $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = -t^{(k)}\nabla f(\mathbf{x})$.

Fig. 9.4 shows the error $f(x^{(k)}) - p^\star$ v.s. k. Graph shows linear rate of convergence.s

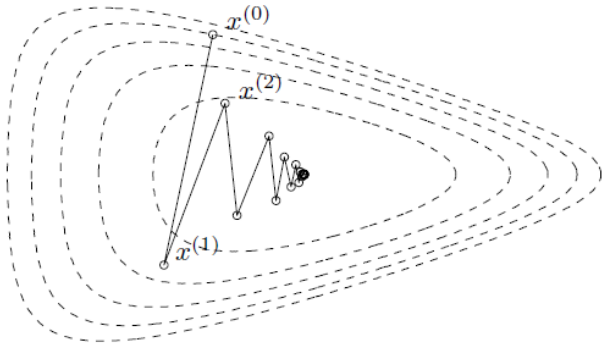


Figure 9.3 Iterates of the gradient method with backtracking line search, for the problem in $\mathbf{R}^2$ with objective f given in (9.20). The dashed curves are level curves of f, and the small circles are the iterates of the gradient method. The solid lines, which connect successive iterates, show the scaled steps $t^{(k)}\Delta x^{(k)}$.

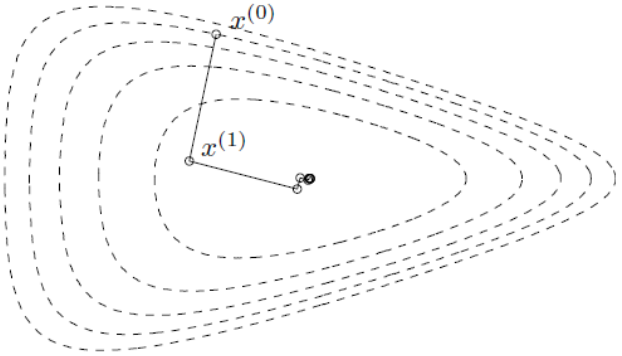


Figure 9.5 Iterates of the gradient method with exact line search for the problem in $\mathbf{R}^2$ with objective f given in (9.20).

Example: Nonquadratic Problem in $\mathbb{R}^2$

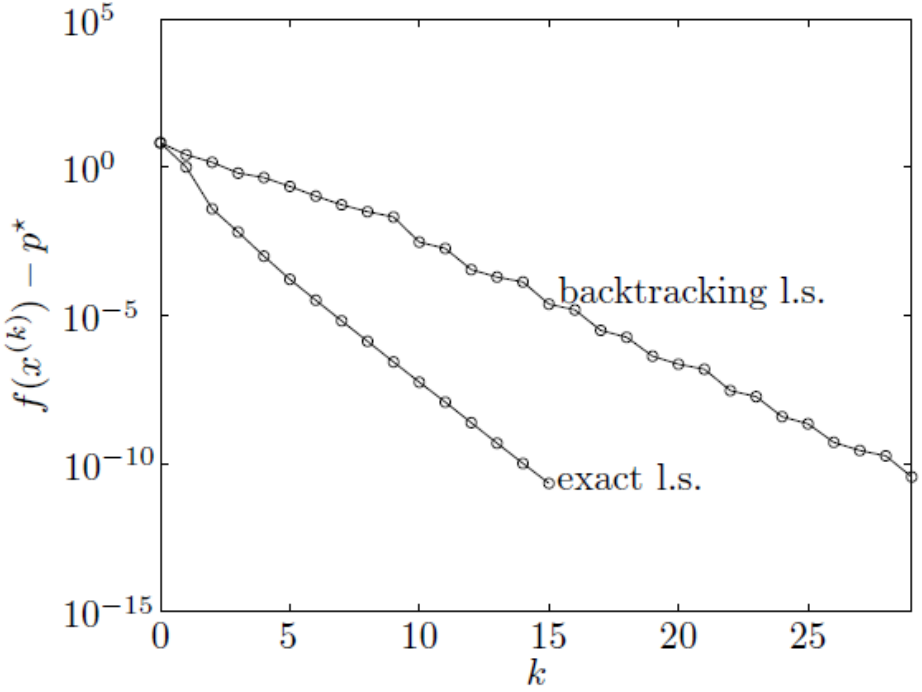


Figure 9.4 Error $f(x^{(k)}) - p^\star$ versus iteration k of the gradient method with backtracking and exact line search, for the problem in $\mathbf{R}^2$ with objective f given in (9.20). The plot shows nearly linear convergence, with the error reduced approximately by the factor 0.4 in each iteration of the gradient method with backtracking line search, and by the factor 0.2 in each iteration of the gradient method with exact line search.

Steepest Descent Method

The main disadvantage of the gradient descent method is the convergence rate depends on $\kappa\left(\nabla^2 f\left(\mathbf{x}\right)\right)$. The steepest descent method improves on this by using a normalized steepest descent direction.

- The first-order Taylor approximation of $f\left(\mathbf{x}+\mathbf{v}\right)$ around $\mathbf{x}$ is (see [(1.54), CLL17])

$$f\left(\mathbf{x}+\mathbf{v}\right) \approx \hat{f}\left(\mathbf{x}+\mathbf{v}\right) = f\left(\mathbf{x}\right) + \nabla f\left(\mathbf{x}\right)^T \mathbf{v}.$$

  The second-term on the righthand side $\nabla f\left(\mathbf{x}\right)^T \mathbf{v}$ is the **directional derivative** of f at $\mathbf{x}$ in the direction of $\mathbf{v}$. It gives the approximate change in f for a small step $\mathbf{v}$. The step $\mathbf{v}$ is a descent direction if the directional derivative is negative.

- Now the question is how to address the question of how to choose $\mathbf{v}$ to *make the directional derivative as negative as possible*. Implying it's nonunique

- We define a *normalized steepest descent direction*

$$\Delta\mathbf{x}_{\text{nsd}} = \arg\min_{\mathbf{v}} \left\{ \nabla f\left(\mathbf{x}\right)^T \mathbf{v} \middle| \left\|\mathbf{v}\right\| = 1 \right\}. \qquad (9.23)$$

  A normalized steepest descent direction $\Delta\mathbf{x}_{\text{nsd}}$ is a step of unit norm that gives the largest decrease in the linear approximation of f.

Steepest Descent Method

- A normalized steepest descent direction can be interpreted geometrically as follows. We can just as well define $\Delta \mathbf{x}_{\text{nsd}}$ as

$$\Delta \mathbf{x}_{\text{nsd}} = \arg\min_{\mathbf{v}} \left\{ \nabla f(\mathbf{x})^T \mathbf{v} \,\middle|\, \|\mathbf{v}\| \le 1 \right\},$$

  - That is, the direction in the unit ball of $\|\cdot\|$ that extends farthest in the direction $-\nabla f(\mathbf{x})$.
- Consider a steepest descent step $\Delta \mathbf{x}_{\text{nsd}}$ that is unnormalized, by scaling the normalized steepest descent direction in a particular way:

$$\Delta \mathbf{x}_{\text{sd}} = \|\nabla f(\mathbf{x})\|_* \, \Delta \mathbf{x}_{\text{nsd}},$$

  where $\|\cdot\|_*$ denotes the dual norm. Note that for the steepest descent step, we have

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

Steepest Descent Method – Dual Norm

To see this, first we derive an expression for the dual norm. By definition

$$\left\|\nabla f(\mathbf{x})\right\|_* = \sup_{\mathbf{u}}\left\{\nabla f(\mathbf{x})^T\mathbf{u}\,\middle|\,\|\mathbf{u}\|\le 1\right\},\ \text{and assume } \|\cdot\| = \|\cdot\|_2 \text{ i.e.}$$

$$\sup_{\|\mathbf{u}\|_2^2\le 1} \nabla f(\mathbf{x})^T\mathbf{u} \quad\Rightarrow\quad \mathcal{L}(\mathbf{u},\lambda) = \nabla f(\mathbf{x})^T\mathbf{u} + \lambda\left(1-\|\mathbf{u}\|_2^2\right)$$

$$\nabla_{\mathbf{u}}\mathcal{L} = 0 \ \Rightarrow\ \mathbf{u}^\star = \frac{1}{2\lambda}\nabla f(\mathbf{x})\ .$$

$$\text{From } \frac{\partial\mathcal{L}}{\partial\lambda} = 0 \ \Rightarrow\ \mathbf{u}^T\mathbf{u} = 1 \ \Rightarrow\ \lambda = \frac{\left\|\nabla f(\mathbf{x})\right\|_2}{2}.\ \text{ So } \mathbf{u}^\star = \frac{1}{2\lambda}\nabla f(\mathbf{x}) = \frac{\nabla f(\mathbf{x})}{\left\|\nabla f(\mathbf{x})\right\|_2}$$

$$\Rightarrow \left\|\nabla f(\mathbf{x})\right\|_* = \nabla f(\mathbf{x})^T\mathbf{u}^\star = \left\|\nabla f(\mathbf{x})\right\|_2$$

Steepest Descent Method For Euclidean Norm

Next recall $\Delta \mathbf{x}_{\text{nsd}} = \arg\min_{\mathbf{v}} \left\{ \nabla f(\mathbf{x})^T \mathbf{v} \,\middle|\, \|\mathbf{v}\| \le 1 \right\}$. Then similar to the derivation from the dual norm except it becomes a minimization problem

$$\min_{\|\mathbf{u}\|_2^2 \le 1} \nabla f(\mathbf{x})^T \mathbf{u} \quad \Rightarrow \quad \mathcal{L}(\mathbf{u}, \lambda) = \nabla f(\mathbf{x})^T \mathbf{u} + \lambda \left(\|\mathbf{u}\|_2^2 - 1 \right)$$

$$\nabla_{\mathbf{u}} \mathcal{L} = 0 \;\Rightarrow\; \mathbf{u}^\star = -\frac{1}{2\lambda} \Delta f(\mathbf{x}). \quad \text{From } \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \;\Rightarrow\; \mathbf{u}^T \mathbf{u} = 1 \;\Rightarrow\; \lambda = \frac{\|\nabla f(\mathbf{x})\|_2}{2}$$

$$\Rightarrow \mathbf{u}^\star = -\frac{1}{2\lambda} \Delta f(\mathbf{x}) = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2} = \Delta x_{\text{nsd}}$$

Since $\|\nabla f(\mathbf{x})\|_* = \|\nabla f(\mathbf{x})\| \quad \Rightarrow \quad \nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\text{nsd}} = -\nabla f(\mathbf{x})^T \frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2} = -\|\nabla f(\mathbf{x})\|_2$

$$= -\|\nabla f(\mathbf{x})\|_*$$

Steepest Descent Method

The steepest descent method uses the steepest descent direction as search direction.

Algorithm 9.4 *Steepest descent method.*

given a starting point $x \in \mathbf{dom}\, f$.

repeat

1. Compute steepest descent direction Δx_{sd}.
2. *Line search.* Choose t via backtracking or exact line search.
3. *Update.* $x := x + t\Delta x_{\mathrm{sd}}$.

until stopping criterion is satisfied.

Steepest Descent Method

- Steepest descent for Euclidean norm:

Recall $\Delta \mathbf{x}_{\text{nsd}} = -\dfrac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2}$. Since $\Delta \mathbf{x}_{\text{sd}} = \|\nabla f(\mathbf{x})\|_* \, \Delta \mathbf{x}_{\text{nsd}}$, and $\|\nabla f(\mathbf{x})\|_* = \|\nabla f(\mathbf{x})\|_2$

$$\Rightarrow \quad \Delta \mathbf{x}_{\text{sd}} = -\nabla f(\mathbf{x})$$

Hence, the steepest descent method for the Euclidean norm coincides with the gradient descent method.

Steepest Descent Method For Quadratic Norm

- Steepest descent for quadratic norm: If we take the norm $\|\cdot\|$ to be quadratic norm

$$\|\mathbf{z}\|_P = \left(\mathbf{z}^T\mathbf{P}\mathbf{z}\right)^{1/2} = \left\|\mathbf{P}^{1/2}\mathbf{z}\right\|_2,$$

where $\mathbf{P} \in \mathbb{S}^n_{++}$. The normalized steepest descent direction is given by

$$\min_{\left\|\mathbf{P}^{1/2}\mathbf{u}\right\|_2^2 \le 1} \nabla f\left(\mathbf{x}\right)^T \mathbf{u} \qquad \Rightarrow \qquad \mathcal{L}\left(\mathbf{u},\lambda\right) = \nabla f\left(\mathbf{x}\right)^T \mathbf{u} + \lambda\left(\left\|\mathbf{P}^{1/2}\mathbf{u}\right\|_2^2 - 1\right)$$

$$\nabla_{\mathbf{u}}\mathcal{L} = 0 \;\Rightarrow\; \mathbf{u}^\star = -\frac{1}{2\lambda}\mathbf{P}^{-1}\Delta f\left(\mathbf{x}\right). \quad \text{From } \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \;\Rightarrow\; \mathbf{u}^T\mathbf{P}\mathbf{u} = 1$$

$$\Rightarrow\; \lambda = \frac{\left(\nabla f\left(\mathbf{x}\right)^T \mathbf{P}^{-1}\nabla f\left(\mathbf{x}\right)\right)^{1/2}}{2}$$

$$\Rightarrow \mathbf{u}^\star = -\left[\nabla f\left(\mathbf{x}\right)^T \mathbf{P}^{-1}\nabla f\left(\mathbf{x}\right)\right]^{-1/2} \mathbf{P}^{-1}\nabla f\left(\mathbf{x}\right) = \Delta\mathbf{x}_{\text{nsd}}$$

Steepest Descent Method For Quadratic Norm (Dual Norm)

The dual norm for the case of the quadratic norm can be derived as

$$\left\|\nabla f(\mathbf{x})\right\|_* = \sup_{\mathbf{u}} \left\{ \nabla f(\mathbf{x})^T \mathbf{u} \,\middle|\, \left\|\mathbf{P}^{1/2}\mathbf{u}\right\|_2 \le 1 \right\},$$

$$\sup_{\left\|\mathbf{P}^{1/2}\mathbf{u}\right\|_2 \le 1} \nabla f(\mathbf{x})^T \mathbf{u} \quad \Rightarrow \quad \mathcal{L}(\mathbf{u},\lambda) = \nabla f(\mathbf{x})^T \mathbf{u} + \lambda\left(1 - \mathbf{u}^T\mathbf{P}\mathbf{u}\right)$$

$$\nabla_{\mathbf{u}}\mathcal{L} = 0 \;\Rightarrow\; \mathbf{u}^\star = \frac{1}{2\lambda}\mathbf{P}^{-1}\nabla f(\mathbf{x}).$$

$$\text{From } \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \;\Rightarrow\; \mathbf{u}^T\mathbf{P}\mathbf{u} = 1 \;\Rightarrow\; \lambda = \frac{\left(\nabla f(\mathbf{x})^T \mathbf{P}^{-1}\nabla f(\mathbf{x})\right)^{1/2}}{2}.$$

$$\Rightarrow \mathbf{u}^\star = \frac{1}{2\lambda}\mathbf{P}^{-1}\nabla f(\mathbf{x}) = \left(\nabla f(\mathbf{x})^T \mathbf{P}^{-1}\nabla f(\mathbf{x})\right)^{-1/2} \mathbf{P}^{-1}\nabla f(\mathbf{x})$$

$$\Rightarrow \left\|\nabla f(\mathbf{x})\right\|_* = \nabla f(\mathbf{x})^T \mathbf{u}^\star = \left(\nabla f(\mathbf{x})^T \mathbf{P}^{-1}\nabla f(\mathbf{x})\right)^{-1/2} \nabla f(\mathbf{x})^T \mathbf{P}^{-1}\nabla f(\mathbf{x})$$

$$= \left(\nabla f(\mathbf{x})^T \mathbf{P}^{-1}\nabla f(\mathbf{x})\right)^{1/2} = \left\|\nabla f(\mathbf{x})\right\|_P$$

Steepest Descent Method For Quadratic Norm

Hence,

$$\Delta \mathbf{x}_{\text{sd}} = \left\|\nabla f\left(\mathbf{x}\right)\right\|_{*} \Delta \mathbf{x}_{\text{nsd}} = -\left(\nabla f\left(\mathbf{x}\right)^{T} \mathbf{P}^{-1} \nabla f\left(\mathbf{x}\right)\right)^{1/2} \left[\nabla f\left(\mathbf{x}\right)^{T} \mathbf{P}^{-1} \nabla f\left(\mathbf{x}\right)\right]^{-1/2} \mathbf{P}^{-1} \nabla f\left(\mathbf{x}\right)$$
$$= -\mathbf{P}^{-1} \nabla f\left(\mathbf{x}\right).$$

The normalized steepest descent direction for a quadratic norm is shown in Fig. 9.9.

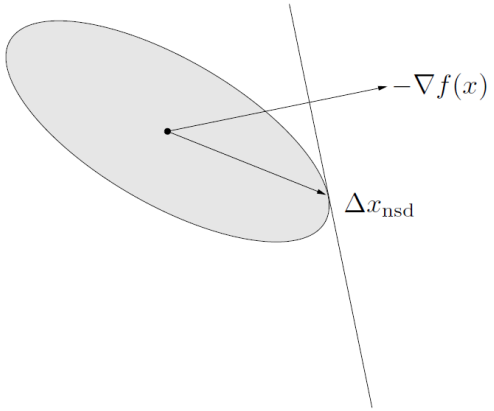


Figure 9.9 Normalized steepest descent direction for a quadratic norm. The ellipsoid shown is the unit ball of the norm, translated to the point x. The normalized steepest descent direction Δx_{nsd} at x extends as far as possible in the direction $-\nabla f(x)$ while staying in the ellipsoid. The gradient and normalized steepest descent directions are shown.

Steepest Descent For ℓ_1-norm

- Steepest descent for ℓ_1-norm:

$$\Delta \mathbf{x}_{\text{nsd}} = \arg\min_{\mathbf{v}} \left\{ \nabla f(\mathbf{x})^T \mathbf{v} \,\middle|\, \|\mathbf{v}\|_1 \leq 1 \right\}$$

is easily characterized. Let i be any index for which $\|\nabla f(\mathbf{x})\|_\infty = \left|[\nabla f(\mathbf{x})]_i\right|$.

- A normalized steepest descent direction $\Delta \mathbf{x}_{\text{nsd}}$ for the ℓ_1-norm is given by

$$\Delta \mathbf{x}_{\text{nsd}} = -\text{sgn}\left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}\right)\mathbf{e}_i,$$

where $\mathbf{e}_i$ is the ith standard basis vector. An unnormalized steepest descent step is then (recall the dual norm is the infinity norm)

$$\Delta \mathbf{x}_{\text{sd}} = \Delta \mathbf{x}_{\text{nsd}} \|\nabla f(\mathbf{x})\|_\infty = -\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}\mathbf{e}_i.$$

Thus, the normalized steepest descent step in ℓ_1-norm can always be chosen to be a standard basis vector (See Fig. 9.10).

Steepest Descent For ℓ_1-norm

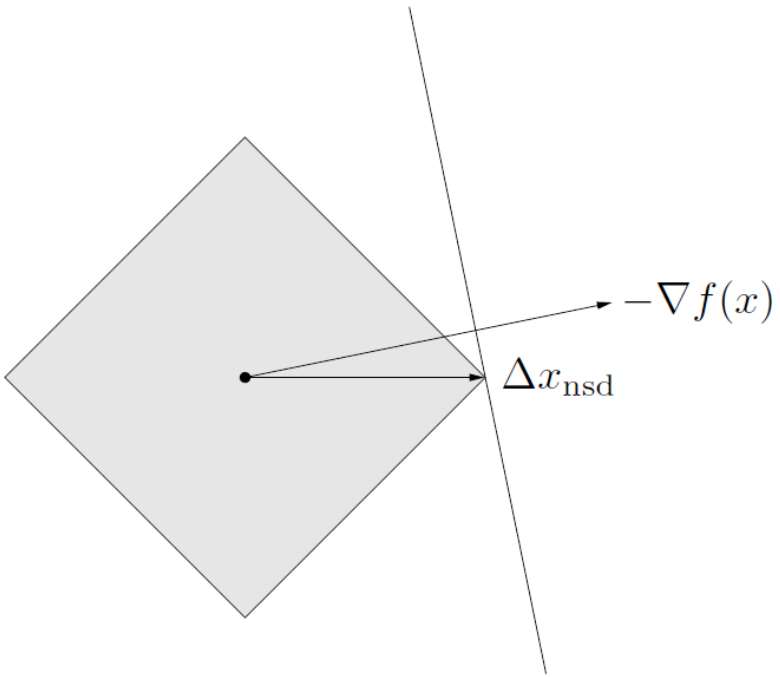


Figure 9.10 Normalized steepest descent direction for the ℓ_1-norm. The diamond is the unit ball of the ℓ_1-norm, translated to the point x. The normalized steepest descent direction can always be chosen in the direction of a standard basis vector; in this example we have $\Delta x_{\text{nsd}} = e_1$.

Newton's Method

The Newton Step: For $\mathbf{x} \in \text{dom } f$, the following vector is called Newton's step

$$\Delta \mathbf{x}_{\text{nt}} = -\nabla^2 f\left(\mathbf{x}\right)^{-1} \nabla f\left(\mathbf{x}\right).$$

Positive definiteness of $\nabla^2 f\left(\mathbf{x}\right)$ implies that

$$\nabla f\left(\mathbf{x}\right)^T \Delta \mathbf{x}_{\text{nt}} = -\nabla f\left(\mathbf{x}\right)^T \nabla^2 f\left(\mathbf{x}\right)^{-1} \nabla f\left(\mathbf{x}\right) \leq 0$$

unless $\nabla f\left(\mathbf{x}\right) = 0,$ so the Newton step is a descent direction (unless $\mathbf{x}$ is optimal).

- Minimizer of second-order approximation:

  The second-order Taylor approximation (or model) $\hat{f}$ of f at $\mathbf{x}$ is

$$\hat{f}\left(\mathbf{x}+\mathbf{v}\right) = f\left(\mathbf{x}\right) + \nabla f\left(\mathbf{x}\right)^T \mathbf{v} + \frac{1}{2}\mathbf{v}^T \nabla^2 f\left(\mathbf{x}\right)\mathbf{v} \qquad \text{[(1.54), CLL17]}$$

$$\text{[(9.28), BV04]}$$

  which is a convex quadratic function of $\mathbf{v}$, and is minimized when $\mathbf{v} = \Delta \mathbf{x}_{\text{nt}}$ $(\nabla_{\mathbf{v}} \hat{f} = \mathbf{0}_n \Rightarrow \mathbf{v}^\star = \Delta \mathbf{x}_{\text{nt}})$

Newton's Method

Thus, the Newton step $\Delta\mathbf{x}_{\text{nt}}$ is what should be added to the point $\mathbf{x}$ to minimize the second-order approximation of f at $\mathbf{x}$. This is shown in Fig. 9.16.

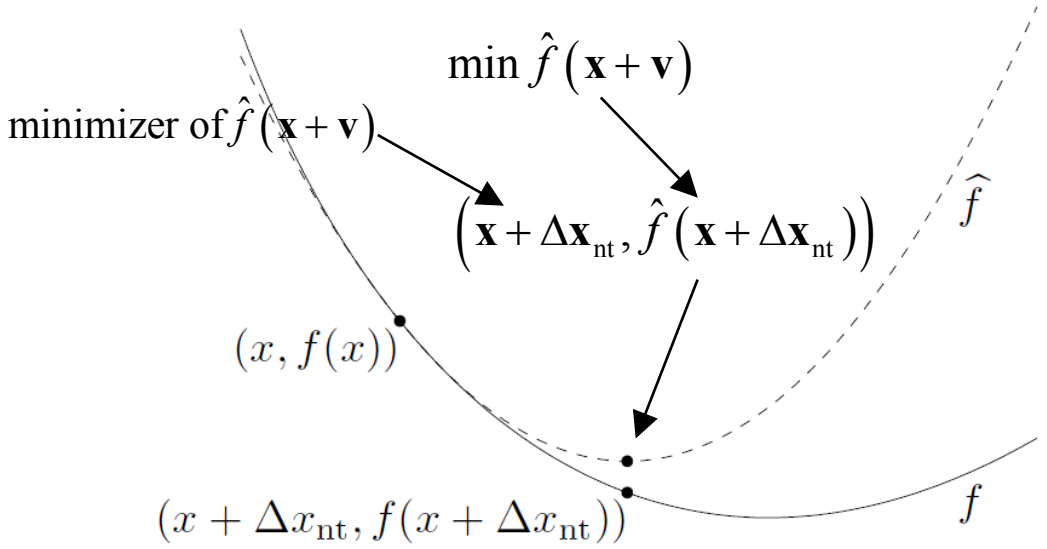


Figure 9.16 The function f (shown solid) and its second-order approximation $\widehat{f}$ at x (dashed). The Newton step Δx_{nt} is what must be added to x to give the minimizer of $\widehat{f}$.

Newton's Step

The interpretation gives us some insight into the Newton step. If the function f is quadratic, then $\mathbf{x} + \Delta\mathbf{x}_{nt}$ is the exact minimizer of f.

If the function f is nearly quadratic, intuition suggests that $\mathbf{x} + \Delta\mathbf{x}_{nt}$ should be a very good estimate of the minimizer of f, i.e. $\mathbf{x}^\star$. Since f is twice differentiable, the quadratic model of f will be very accurate when $\mathbf{x}$ is near $\mathbf{x}^\star$. It follows that when $\mathbf{x}$ is near $\mathbf{x}^\star$, the point $\mathbf{x} + \Delta\mathbf{x}_{nt}$ should be a very good estimate of $\mathbf{x}^\star$.

Newton's Method

From pp. 49, steepest descent for quadratic norm is

$$\Delta \mathbf{x}_{\text{sd}} = \left\|\nabla f(\mathbf{x})\right\|_* \Delta \mathbf{x}_{\text{nsd}} = -\left(\nabla f(\mathbf{x})^T \mathbf{P}^{-1} \nabla f(\mathbf{x})\right)^{1/2} \left[\nabla f(\mathbf{x})^T \mathbf{P}^{-1} \nabla f(\mathbf{x})\right]^{-1/2} \mathbf{P}^{-1} \nabla f(\mathbf{x})$$
$$= -\mathbf{P}^{-1} \nabla f(\mathbf{x}).$$

- Steepest descent direction in Hessian norm: The Newton step is also the steepest descent direction at $\mathbf{x}$ (see pp. 47-49), for the quadratic norm defined by the Hessian $\nabla^2 f(\mathbf{x})$, i.e.

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

- This gives another insight into why the Newton step should be a good search direction and a very good search direction when $\mathbf{x}$ is near $\mathbf{x}^\star$.
- In particular, near $\mathbf{x}^\star$, a very good choice is $\mathbf{P} = \nabla^2 f(\mathbf{x}^\star)$. When $\mathbf{x}$ is near $\mathbf{x}^\star$, we have $\nabla^2 f(\mathbf{x}) \approx \nabla^2 f(\mathbf{x}^\star)$, which explains why the Newton step is a very good choice of search direction. That is illustrated in Fig. 9.17.

Newton's Method

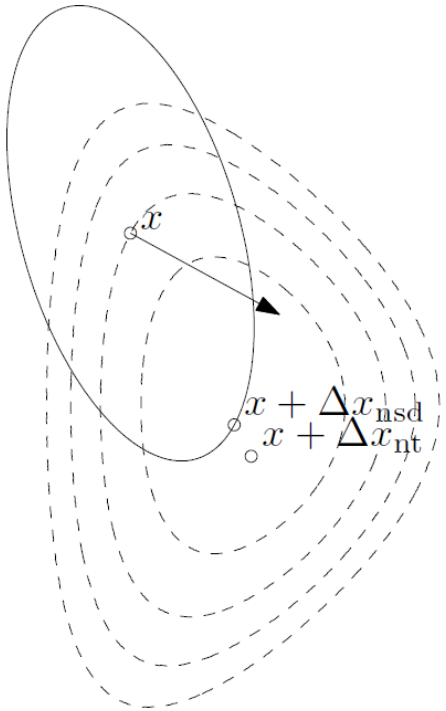


Figure 9.17 The dashed lines are level curves of a convex function. The ellipsoid shown (with solid line) is $\{x + v \mid v^T \nabla^2 f(x) v \leq 1\}$. The arrow shows $-\nabla f(x)$, the gradient descent direction. The Newton step Δx_{nt} is the steepest descent direction in the norm $\|\cdot\|_{\nabla^2 f(x)}$. The figure also shows Δx_{nsd}, the normalized steepest descent direction for the same norm.

Newton's Method

- Solution of linearized optimality condition: If we linearize the optimality condition $\nabla f\left(\mathbf{x}^{\star}\right)=0$ near $\mathbf{x}$ we obtain [(1.53), CLL17]

$$\nabla f\left(\mathbf{x}+\mathbf{v}\right)\approx\nabla f\left(\mathbf{x}\right)+\nabla^{2} f\left(\mathbf{x}\right)\mathbf{v}=\mathbf{0},$$

  which is linear equation in $\mathbf{v}$, with solution $\mathbf{v}=\Delta\mathbf{x}_{\text{nt}}$. So the Newton step $\Delta\mathbf{x}_{\text{nt}}$ is what must be added to $\mathbf{x}$ so that the linearized optimality condition holds.
- When $\mathbf{x}$ is near $\mathbf{x}^{\star}$ (so the optimality conditions almost hold), the update $\mathbf{x}+\Delta\mathbf{x}_{\text{nt}}$ should be a very good approximation of $\mathbf{x}^{\star}$
- When $n=1$, i.e. $f:\mathbb{R}\rightarrow\mathbb{R}$, this interpretation is particularly simple. The solution $\mathbf{x}^{\star}$ of the minimization problem is characterized by $f'\left(\mathbf{x}^{\star}\right)=0$, i.e. it is the zero-crossing of the derivative f', which is monotically increasing since f is convex.
- Given our current approximation $\mathbf{x}$ of the solution, we form a first-order Taylor approximation of f' at $\mathbf{x}$. The zero-crossing of this affine approximation is then $\mathbf{x}+\Delta\mathbf{x}_{\text{nt}}$ (see Fig. 9.18)

Newton's Method

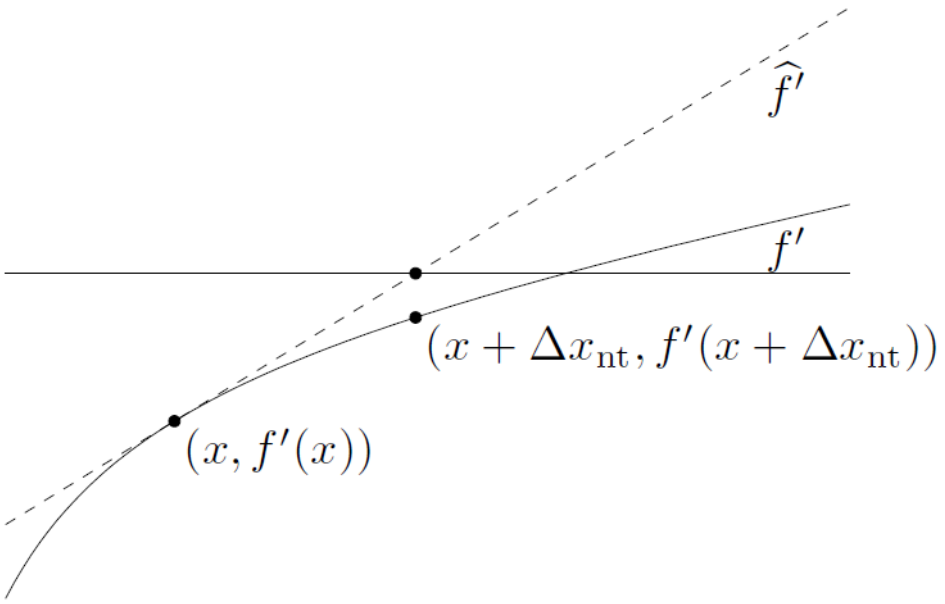


Figure 9.18 The solid curve is the derivative f' of the function f shown in figure 9.16. $\widehat{f'}$ is the linear approximation of f' at x. The Newton step Δx_{nt} is the difference between the root of $\widehat{f'}$ and the point x.

Newton's Method

- The Newton Decrement: The following quantity is called the Newton decrement at $\mathbf{x}$

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

- We can relate the Newton decrement to the quantity $f(\mathbf{x}) - \inf_{\mathbf{y}} \hat{f}(\mathbf{y})$, where $\hat{f}$ is the second-order approximation of f at $\mathbf{x}$:

$$f(\mathbf{x}) - \inf_{\mathbf{y}} \hat{f}(\mathbf{y}) = f(\mathbf{x}) - \hat{f}(\mathbf{x} + \Delta\mathbf{x}_{\text{nt}}) = \frac{1}{2}\lambda(\mathbf{x})^2. \quad \text{Why?}$$

$$\hat{f}(\mathbf{y})\Big|_{\mathbf{y}=\mathbf{x}+\mathbf{v}} = \hat{f}(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \qquad (*)$$

$$\nabla_{\mathbf{y}} \hat{f} = \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\mathbf{y} - \nabla^2 f(\mathbf{x})\mathbf{x} = \mathbf{0}_n$$

$$\Rightarrow \quad \mathbf{y}^\star = \mathbf{x} - \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x}) = \mathbf{x} + \Delta\mathbf{x}_{\text{nt}}$$

Newton's Method

$$\mathbf{y}^\star = \mathbf{x} - \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x}) = \mathbf{x} + \Delta\mathbf{x}_{\text{nt}}$$

$$\inf_{\mathbf{y}} \hat{f}(\mathbf{y}) = \hat{f}(\mathbf{y})\big|_{\mathbf{y}=\mathbf{x}+\mathbf{v}} = \hat{f}(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}) \qquad (*)$$

Plug $\mathbf{y}^\star$ into (*): $\inf_{\mathbf{y}} \hat{f}(\mathbf{y}) = \hat{f}(\mathbf{y}^\star) = f(\mathbf{x}) - \frac{1}{2}\nabla f(\mathbf{x})^T \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})$

$$\Rightarrow f(\mathbf{x}) - \inf_{\mathbf{y}} \hat{f}(\mathbf{y}) = f(\mathbf{x}) - \hat{f}(\mathbf{x} + \Delta\mathbf{x}_{\text{nt}}) = \frac{1}{2}\nabla f(\mathbf{x})^T \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x}) = \frac{1}{2}\lambda(\mathbf{x})^2.$$

- So $\lambda^2 / 2$ is an estimate of $f(\mathbf{x}) - p^\star$, based on the quadratic approximation of f at $\mathbf{x}$.

We can also express the Newton decrement as ($\Delta\mathbf{x}_{\text{nt}} = -\nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})$)

$$\lambda(\mathbf{x}) = \left(\Delta\mathbf{x}_{\text{nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\text{nt}}\right)^{1/2}.$$

This shows that λ is the norm of the Newton step, in the quadratic norm defined by the Hessian, i.e. the norm

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

Newton's Method

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

- The Newton decrement comes up in backtracking line search as well, since we have

$$\nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\text{nt}} = -\lambda(\mathbf{x})^2. \qquad (9.30)$$

This is the constant used in a backtracking line search, and can be interpreted as the directional derivative of f at $\mathbf{x}$ in the direction of the Newton step:

$$-\lambda(\mathbf{x})^2 = \nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\text{nt}} = \frac{d}{dt} f(\mathbf{x} + \Delta \mathbf{x}_{\text{nt}} t)\Big|_{t=0}$$

backtracking LS stopping criterion

$$f(\mathbf{x} + t\Delta \mathbf{x}) \le f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta \mathbf{x}$$

Algorithm 9.5 *Newton's method.*

given a starting point $x \in \mathbf{dom}\, f$, tolerance $\epsilon > 0$.

repeat

1. *Compute the Newton step and decrement.*
   $\Delta x_{\text{nt}} := -\nabla^2 f(x)^{-1} \nabla f(x)$; $\lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)$.
2. *Stopping criterion.* **quit** if $\lambda^2/2 \le \epsilon$.
3. *Line search.* Choose step size t by backtracking line search.
4. *Update.* $x := x + t\Delta x_{\text{nt}}$.

Classical Convergence Analysis

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

$$\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\|_2 \leq L\left\|\mathbf{x} - \mathbf{y}\right\|_2 \tag{9.31}$$

  L can be interpreted as a bound on the third derivative of f, which can be taken to be zero for a quadratic function. Or in general, L measures how well f can be approximated by a quadratic function.

- Outline (before actual proof): there exists constants $\eta \in \left(0, \frac{m^2}{L}\right)$, $\gamma > 0$ such that
  - if $\left\|\nabla f(\mathbf{x})\right\|_2 \geq \eta$, then $f\left(\mathbf{x}^{(k+1)}\right) - f\left(\mathbf{x}^{(k)}\right) \leq -\gamma$ (9.32)
  - if $\left\|\nabla f(\mathbf{x})\right\|_2 < \eta$, then $\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k+1)}\right)\right\|_2 \leq \left(\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2\right)^2$ (9.33)

Classical Convergence Analysis

- Consider the implication of the second condition first.
- Quadratically Convergent Phase ($\left\|\nabla f\left(\mathbf{x}\right)\right\|_2 < \eta$)
  - all iterations use step size $t = 1$
  - $\left\|\nabla f\left(\mathbf{x}\right)\right\|_2$ converges to zero quadratically: if $\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2 < \eta$ and since $\eta \le \frac{m^2}{L}$, then $\left\|\nabla f\left(\mathbf{x}^{(k+1)}\right)\right\|_2 < \eta$, so it will hold for future iterates $\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2 \le \eta,\ \forall \ell \ge k.$

So when algorithm takes a full Newton step $t = 1$, and

$$\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(\ell+1)}\right)\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2\right)^2.$$

Applying this inequality recursively, we find that for $\ell \ge k$,

$$\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2\right)^{2^{\ell-k}} \le \left(\frac{1}{2}\right)^{2^{\ell-k}}, \quad \ell \ge k.$$

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

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

Classical Convergence Analysis

$$p^\star \geq f(\mathbf{x}) - \frac{1}{2m}\left\|\nabla f(\mathbf{x})\right\|_2^2 \qquad (9.9)$$

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

$$\Leftrightarrow \frac{L^2}{2m^4}\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2^2 \leq \left(\frac{1}{2}\right)^{2^{\ell-k+1}}, \quad \ell \geq k$$

$$\Leftrightarrow \frac{1}{2m}\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2^2 \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-k+1}}, \quad \ell \geq k$$

So from (9.9) and above: $f\left(\mathbf{x}^{(\ell)}\right) - p^\star \leq \frac{1}{2m}\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2^2 \leq \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-k+1}}$ (9.35)

Newton's Method

$$f\left(\mathbf{x}^{(\ell)}\right)-p^{\star}\leq\frac{1}{2m}\left\|\nabla f\left(\mathbf{x}^{(\ell)}\right)\right\|_2^2\leq\frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^{\ell-k+1}} \quad (9.35)$$

Last inequality shows that convergence is extremely fast once the second condition is satisfied. This is called quadratic convergence. (9.35) means that after a sufficiently large number of iterations, the number of correct digits doubles at each iteration.

- Newton's method falls into two stages, the quadratically convergent stage, or

- Damped Newton Phase $\left(\left\|\nabla f\left(\mathbf{x}\right)\right\|_2\geq\eta\right)$
  - it is called damped because step size can be $t<1$
  - most iterations require backtracking steps
  - function value decreases by at least γ
  - if $p^{\star}>-\infty$, this phase ends after at most $\left(f\left(\mathbf{x}^{(0)}\right)-p^{\star}\right)/\gamma$ iteratons

Classical Convergence Analysis

- Conclusions:
  - From (9.35) and ϵ-suboptimality,

$$f\left(\mathbf{x}^{(\ell)}\right) - p^\star \leq \epsilon \leq \epsilon_0 \left(\frac{1}{2}\right)^{2^{\ell-k+1}} \quad \Leftrightarrow \log_2 \epsilon \leq \log_2 \epsilon_0 + \log_2 \left(\frac{1}{2}\right)^{2^{\ell-k+1}}$$

$$\Leftrightarrow \log_2 \frac{\epsilon}{\epsilon_0} \leq \left(2^{\ell-k+1}\right) \log_2 \left(\frac{1}{2}\right)$$

$$\Leftrightarrow \log_2 \log_2 \frac{\epsilon}{\epsilon_0} \leq \left(\ell - k + 1\right) + c.$$

which implies that we have $f\left(\mathbf{x}\right) - p^\star \leq \epsilon$ after no more than $\log_2 \log_2 \left(\frac{\epsilon_0}{\epsilon}\right)$ iterations,

where $\epsilon_0 = \frac{2m^2}{L^2}$, with c being a constant, in the quadratically convergent phase.

- Overall, the number of iterations until $f\left(\mathbf{x}\right) - p^\star \leq \epsilon$ is bounded above by

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

Classical Convergence Analysis

- $\log_2 \log_2 \left(\frac{\epsilon_0}{\epsilon} \right)$ grows extremely slowly with required accuracy ϵ and can be considered a constant (say 5 or 6)
  - Six iterations of the quadratically convergent stage gives an accuracy of about $\epsilon \approx 5 \cdot 10^{-20} \epsilon_0$
- in practice, constants m, L (hence γ, ϵ_0) are usually unknown
- Nevertheless, it provides qualitative insight in convergence properties (i.e. explains two algorithm phases)
- Not quite accurately, then, we can say that the number of Newton iterations required to minimize f is bounded above by

$$\frac{f\left(\mathbf{x}^{(0)}\right) - p^\star}{\gamma} + 6. \qquad (9.37)$$

Example in $\mathbb{R}^2$

- We first apply Newton's method with backtracking line search on the test function (9.20): $f\left(x_1, x_2\right) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$, with line search parameters $\alpha = 0.1$, $\beta = 0.7$. Fig. 9.19 shows the Newton iterates, and also the ellipsoids

$$\left\{\mathbf{x} \,\middle|\, \left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_{\nabla^2 f\left(\mathbf{x}^{(k)}\right)} \leq 1\right\}$$

  for the first two iterates $k = 0,1$. The method works well because these ellipsoids give good approximations of the shape of the sublevel sets.
- Fig. 9.20 shows the error vs. iteration number for the same example. This plot shows that convergence to a very high accuracy is achieved in only five iterations. Quadratic convergence is clearly apparent: The last step reduces the error from about 10^{-5} to 10^{-10}.

Example in $\mathbb{R}^2$

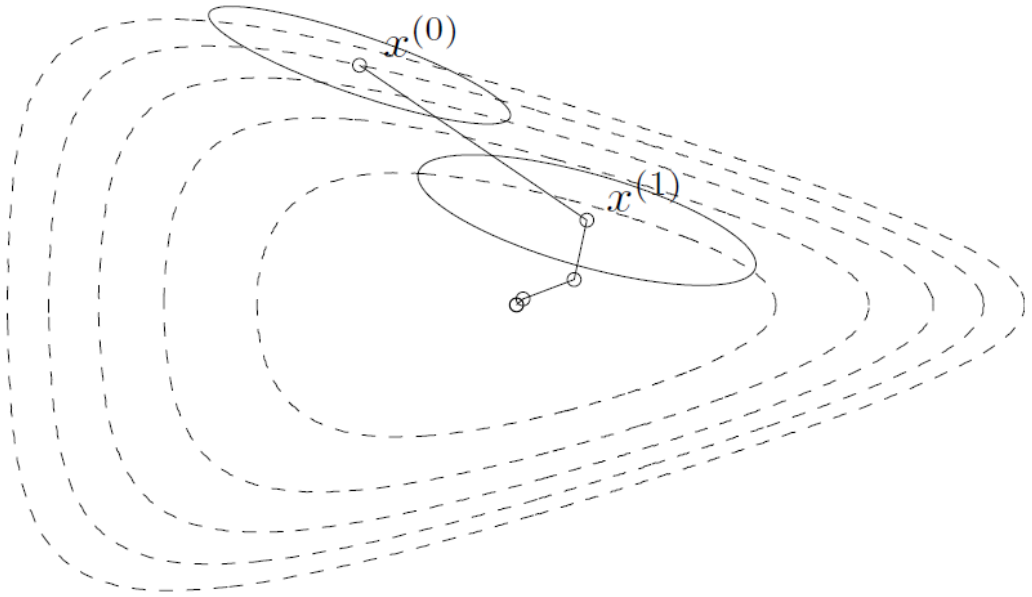


Figure 9.19 Newton's method for the problem in $\mathbf{R}^2$, with objective f given in (9.20), and backtracking line search parameters $\alpha = 0.1$, $\beta = 0.7$. Also shown are the ellipsoids $\{x \mid \|x - x^{(k)}\|_{\nabla^2 f(x^{(k)})} \leq 1\}$ at the first two iterates.

Example in $\mathbb{R}^2$

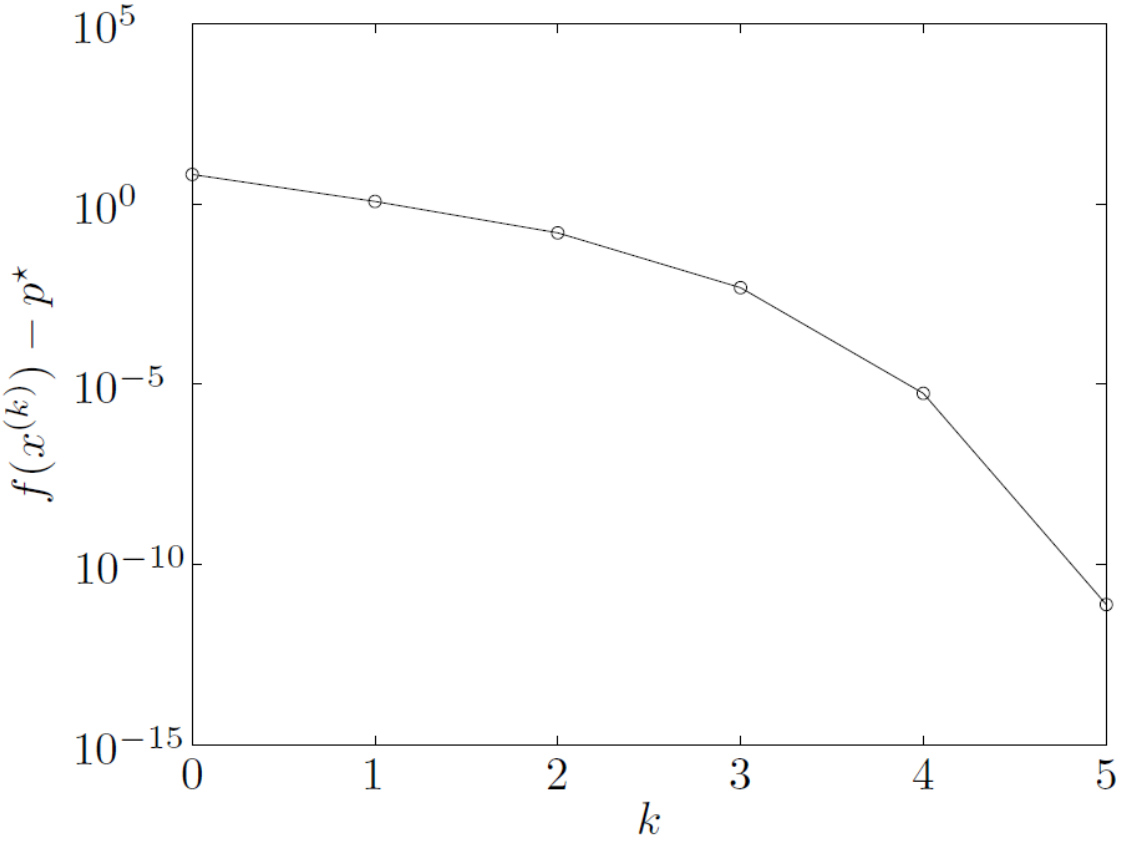


Figure 9.20 Error versus iteration k of Newton's method for the problem in $\mathbf{R}^2$. Convergence to a very high accuracy is achieved in five iterations.

Example in $\mathbb{R}^{100}$

Fig. 9.21 shows the convergence of Newton's method with backtracking and exact line search for a problem in $\mathbb{R}^{100}$ for (9.21): $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} - \sum_{i=1}^{m} \log\left(b_i - \mathbf{a}_i^T\mathbf{x}\right)$, using the same starting point as in Fig. 9.6 (See [BV04] for details)

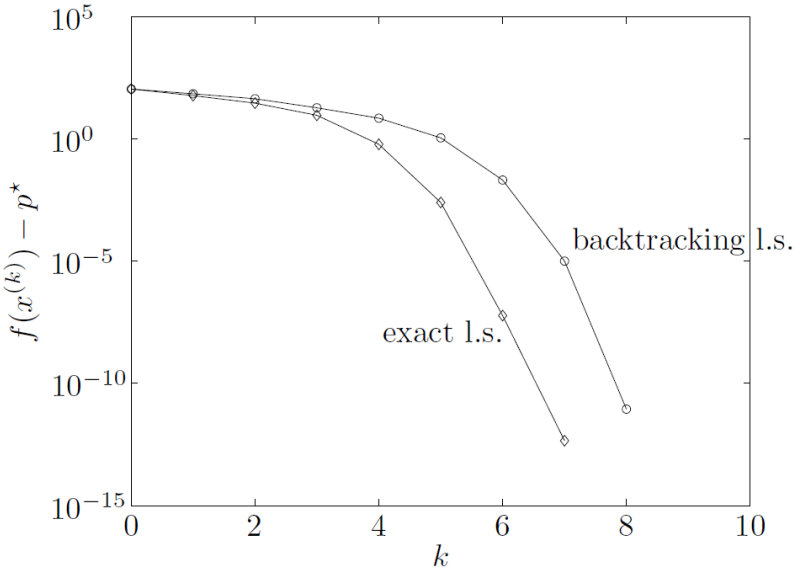


Figure 9.21 Error versus iteration for Newton's method for the problem in $\mathbf{R}^{100}$. The backtracking line search parameters are $\alpha = 0.01$, $\beta = 0.5$. Here too convergence is extremely rapid: a very high accuracy is attained in only seven or eight iterations. The convergence of Newton's method with exact line search is only one iteration faster than with backtracking line search.

Example in $\mathbb{R}^{100}$

- The plot for the backtracking line search shows that a very high accuracy is attained in eight iterations
- Like the example in $\mathbb{R}^2$, quadratic convergence is clearly evident after about the third iteration. The number of iterations in Newton's method with exact line search is only one smaller than with a backtracking line search.
- This is typical. An exact line search usually gives a very small improvement in convergence of Newton's method.
- Fig. 9.22 shows the step sizes for this example. After two damped steps, the steps taken by the backtracking line search are all full, i.e. $t = 1$.

Example in $\mathbb{R}^{100}$

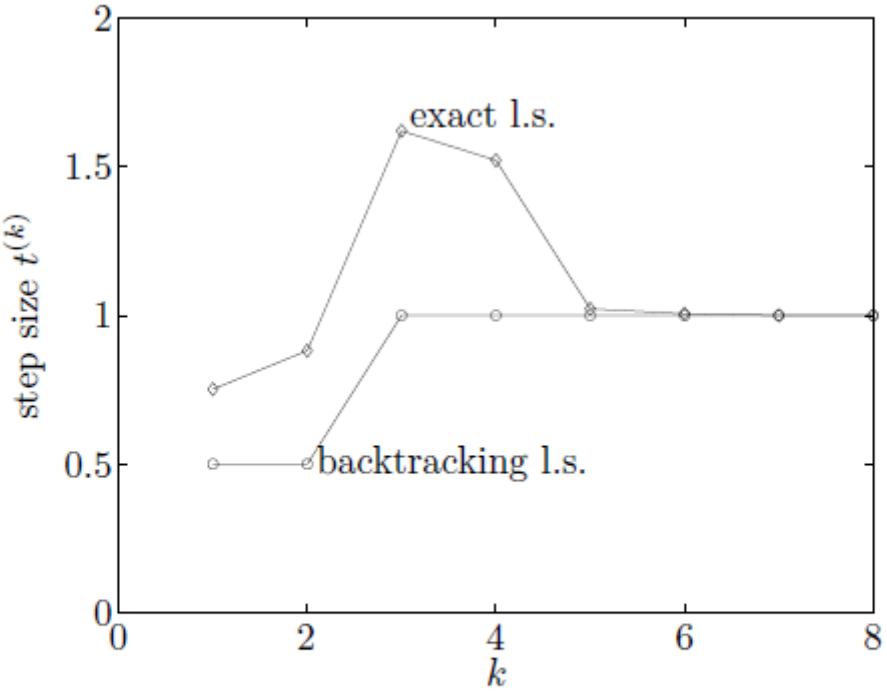


Figure 9.22 The step size t versus iteration for Newton's method with backtracking and exact line search, applied to the problem in $\mathbf{R}^{100}$. The backtracking line search takes one backtracking step in the first two iterations. After the first two iterations it always selects $t = 1$.

Summary for Newton's Method

Newton's method has several very strong advantages over gradient and steepest descent methods:

- Convergence of Newton's method is rapid in general, and quadratic near $\mathbf{x}^\star$. Once the quadratic convergence phase is reached, at most six or so iterations are required to produce a solution of very high accuracy.
- Newton's method is affine invariant. It is insensitive to the choice of coordinates, or the condition number of the sublevel sets of the objective.
- Newton's method scales well with problem size. Its performance on problems in $\mathbb{R}^{10000}$ is similar to its performance on problems in $\mathbb{R}^{10}$, with only a modest increase in the number of steps required.
- The good performance of Newton's method is not dependent on the choice of algorithm parameters. In contrast, the choice of norm for steepest descent plays a critical role in its performance.

Summary for Gradient's Method

- The gradient method often exhibits approximately linear convergence, i.e. the error $f\left(\mathbf{x}^{(k)}\right) - p^{\star}$ converges to zero approximately as a geometric series.
- The choice of backtracking parameters α, β has a noticeable but not dramatic effect on the convergence. An exact line search sometimes improves the convergence of the gradient method, but the effect is not large.
- The convergence rate depends on the condition number of the Hessian, or the sublevel sets. Convergence can be very slow, even for problems that are moderately well conditioned. When the condition number is larger, the gradient method is so slow that is is useless in practice.
- The main advantage of the gradient method is its simplicity. Its main disadvantage is that its convergence rate depends so critically on the condition number of the Hessian or sublevel sets.

Classical Convergence Proof of Newton's Method

$\nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\text{nt}} = -\lambda(\mathbf{x})^2$ (9.30)

if $\|\nabla f(\mathbf{x})\|_2 \geq \eta$, then $f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)}) \leq -\gamma$ (9.32)

Damped Newton's phase:

We now establish the inequality (9.32). Assume $\|\nabla f(\mathbf{x})\|_2 \geq \eta$. First derive a lower bound on the step size selected by the line search. Strong convexity implies that $\nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}$ on S, and therefore

$$f(\mathbf{x} + t\Delta\mathbf{x}_{\text{nt}}) \leq f(\mathbf{x}) + t\nabla f(\mathbf{x})^T \Delta\mathbf{x}_{\text{nt}} + \frac{M\|\Delta\mathbf{x}_{\text{nt}}\|_2^2}{2} t^2$$

$$\leq f(\mathbf{x}) - t\lambda(\mathbf{x})^2 + \frac{M}{2m} t^2 \lambda(\mathbf{x})^2,$$

where (9.30) is used and $\lambda(\mathbf{x})^2 = \Delta\mathbf{x}_{\text{nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\text{nt}} \geq m\|\Delta\mathbf{x}_{\text{nt}}\|_2^2$.

Strong convexity: $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}$

$\lambda(\mathbf{x}) = \Delta\mathbf{x}_{\text{nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\text{nt}}$

$\Rightarrow \quad m\|\Delta\mathbf{x}_{\text{nt}}\|_2^2 \leq \lambda(\mathbf{x})^2 \leq M\|\Delta\mathbf{x}_{\text{nt}}\|_2^2$

The step size $\hat{t} = m/M$ satisfies the exit condition of the line search, since

$$f(\mathbf{x} + \hat{t}\Delta\mathbf{x}_{\text{nt}}) \leq f(\mathbf{x}) - \frac{m}{2M}\lambda(\mathbf{x})^2 \leq f(\mathbf{x}) - \alpha\hat{t}\lambda(\mathbf{x})^2.$$

line search exit criterion

$f(\mathbf{x} + t\Delta\mathbf{x}) \approx f(\mathbf{x}) + t\nabla f(\mathbf{x})^T \Delta\mathbf{x} \leq f(\mathbf{x}) + \alpha t\nabla f(\mathbf{x})^T \Delta\mathbf{x}$

Classical Convergence Proof of Newton's Method

$$\text{if } \left\|\nabla f(\mathbf{x})\right\|_2 \geq \eta, \text{ then } f\left(\mathbf{x}^{(k+1)}\right) - f\left(\mathbf{x}^{(k)}\right) \leq -\gamma \qquad (9.32)$$

So we can use the step size $t \geq \beta m / M$ resulting in a decrease of the objective function

$$\begin{aligned} f\left(\mathbf{x}^{+}\right) - f(\mathbf{x}) &\leq -\alpha t \lambda(\mathbf{x})^2 \\ &\leq -\alpha\beta \frac{m}{M} \lambda(\mathbf{x})^2 \\ &\leq -\alpha\beta \frac{m}{M^2} \left\|\nabla f(\mathbf{x})\right\|_2^2 \\ &\leq -\alpha\beta\eta^2 \frac{m}{M^2}, \end{aligned}$$

where we use $\lambda(\mathbf{x})^2 = \nabla f(\mathbf{x})^T \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})^T \geq (1/M) \left\|\nabla f(\mathbf{x})\right\|_2^2$.

Therefore (9.32) is satisfied with $\gamma = \alpha\beta\eta^2 \frac{m}{M^2}$. (9.38)

Classical Convergence Proof of Newton's Method

$$\text{if } \left\|\nabla f(\mathbf{x})\right\|_2 < \eta, \text{ then } \frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k+1)}\right)\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2\right)^2 \qquad (9.33)$$

Quadratically convergent phase:

We now establish the inequality (9.33). Assume $\left\|\nabla f(\mathbf{x})\right\|_2 < \eta$. We first show that the backtracking line search selects unit steps if

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

By the Lipschitz condition (9.31), we have, for $t \ge 0$,

$$\left\|\nabla^2 f(\mathbf{x}+t\Delta\mathbf{x}_{\text{nt}}) - \nabla^2 f(\mathbf{x})\right\|_2 \le tL\left\|\Delta\mathbf{x}_{\text{nt}}\right\|_2, \text{ and therefore}$$

$$\left|\Delta\mathbf{x}_{\text{nt}}^T\left(\nabla^2 f(\mathbf{x}+t\Delta\mathbf{x}_{\text{nt}}) - \nabla^2 f(\mathbf{x})\right)\Delta\mathbf{x}_{\text{nt}}\right| \le tL\left\|\Delta\mathbf{x}_{\text{nt}}\right\|_2^3.$$

With $\tilde{f}(t) = f(\mathbf{x}+t\Delta\mathbf{x}_{\text{nt}})$, we have $\tilde{f}''(t) = \Delta\mathbf{x}_{\text{nt}}^T\nabla^2 f(\mathbf{x}+t\Delta\mathbf{x}_{\text{nt}})\Delta\mathbf{x}_{\text{nt}}$, so the above inequality is

$$\left|\tilde{f}''(t) - \tilde{f}''(0)\right| \le tL\left\|\Delta\mathbf{x}_{\text{nt}}\right\|_2^3.$$

Classical Convergence Proof of Newton's Method

Strong convexity: $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I} \Leftrightarrow \frac{1}{m}\mathbf{I} \succeq \nabla^2 f(\mathbf{x})^{-1} \succeq \frac{1}{M}\mathbf{I}$

$\lambda(\mathbf{x}) \triangleq \left(\nabla f(\mathbf{x})^T \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})\right)^{1/2} = \Delta\mathbf{x}_{\text{nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\text{nt}}$

$\frac{1}{m}\|\nabla f(\mathbf{x})\|_2^2 \geq \lambda(\mathbf{x})^2 \geq \frac{1}{M}\|\nabla f(\mathbf{x})\|_2^2 \Rightarrow m\|\Delta\mathbf{x}_{\text{nt}}\|_2^2 \leq \lambda(\mathbf{x})^2 \leq M\|\Delta\mathbf{x}_{\text{nt}}\|_2^2$

$\left|\tilde{f}''(t) - \tilde{f}''(0)\right| \leq tL\|\Delta\mathbf{x}_{\text{nt}}\|_2^3$ will be used to determine an upper bound on $\tilde{f}(t)$. We start with

$$\tilde{f}''(t) \leq \tilde{f}''(0) + tL\|\Delta\mathbf{x}_{\text{nt}}\|_2^3 \leq \lambda(\mathbf{x})^2 + t\frac{L}{m^{3/2}}\lambda(\mathbf{x})^3,$$

where we use $\tilde{f}''(0) = \lambda(\mathbf{x})^2$ and $\lambda(\mathbf{x})^2 \geq m\|\Delta\mathbf{x}_{\text{nt}}\|_2^2$. Integrate the inequality to get

$$\tilde{f}'(t) \leq \tilde{f}'(0) + t\lambda(\mathbf{x})^2 + t^2\frac{L}{2m^{3/2}}\lambda(\mathbf{x})^3$$

$$= -\lambda(\mathbf{x})^2 + t\lambda(\mathbf{x})^2 + t^2\frac{L}{2m^{3/2}}\lambda(\mathbf{x})^3,$$

using $\tilde{f}'(0) = f'(\mathbf{x} + t\Delta\mathbf{x}_{\text{nt}})^T \Delta\mathbf{x}_{\text{nt}}\Big|_{t=0} = \nabla f(\mathbf{x})^T \Delta\mathbf{x}_{\text{nt}} = -\lambda(\mathbf{x})^2$ (see (9.30)).

Classical Convergence Proof of Newton's Method

Strong convexity: $m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I} \Leftrightarrow \frac{1}{m}\mathbf{I} \succeq \nabla^2 f(\mathbf{x})^{-1} \succeq \frac{1}{M}\mathbf{I}$

$$\lambda(\mathbf{x}) \triangleq \left(\nabla f(\mathbf{x})^T \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})\right)^{1/2} \Rightarrow \frac{1}{m}\left\|\nabla f(\mathbf{x})\right\|_2^2 \geq \lambda(\mathbf{x})^2 \geq \frac{1}{M}\left\|\nabla f(\mathbf{x})\right\|_2^2$$

$$\Rightarrow \frac{1}{m^{1/2}}\left\|\nabla f(\mathbf{x})\right\|_2 \geq \lambda(\mathbf{x}) \quad \text{(use on this page)}$$

$$= \Delta\mathbf{x}_{\text{nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\text{nt}} \Rightarrow m\left\|\Delta\mathbf{x}_{\text{nt}}\right\|_2^2 \leq \lambda(\mathbf{x})^2 \leq M\left\|\Delta\mathbf{x}_{\text{nt}}\right\|_2^2$$

Integrate one more to get

$$\tilde{f}(t) \leq \tilde{f}(0) - t\lambda(\mathbf{x})^2 + t^2 \frac{1}{2}\lambda(\mathbf{x})^2 + t^3 \frac{L}{6m^{3/2}}\lambda(\mathbf{x})^3.$$

Finally, we take $t = 1$ to obtain

$$f(\mathbf{x} + t\Delta\mathbf{x}_{\text{nt}}) \leq f(\mathbf{x}) - \frac{1}{2}\lambda(\mathbf{x})^2 + \frac{L}{6m^{3/2}}\lambda(\mathbf{x})^3. \tag{9.39}$$

Now suppose $\left\|\nabla f(\mathbf{x})\right\|_2 \leq \eta \leq 3(1-2\alpha)m^2 / L$. By strong convexity, we have

$$\lambda(\mathbf{x}) \leq 3(1-2\alpha)m^{3/2} / L \quad \Leftrightarrow \quad \alpha \leq \frac{1}{2} - \frac{L\lambda(\mathbf{x})}{6m^{3/2}},$$

Classical Convergence Proof of Newton's Method

and by (9.39) we have

$$\begin{aligned} f\left(\mathbf{x}+t\Delta\mathbf{x}_{\text{nt}}\right) &\le f\left(\mathbf{x}\right)-\frac{1}{2}\lambda\left(\mathbf{x}\right)^2+\frac{L}{6m^{3/2}}\lambda\left(\mathbf{x}\right)^3 \\ &\le f\left(\mathbf{x}\right)-\lambda\left(\mathbf{x}\right)^2\left(\frac{1}{2}-\frac{L\lambda\left(\mathbf{x}\right)}{6m^{3/2}}\right) \\ &\le f\left(\mathbf{x}\right)-\alpha\lambda\left(\mathbf{x}\right)^2 \\ &= f\left(\mathbf{x}\right)+\alpha\nabla f\left(\mathbf{x}\right)^T\Delta\mathbf{x}_{\text{nt}}, \end{aligned}$$

which shows that the unit step $t=1$ is accepted by the backtracking line search.

Classical Convergence Proof of Newton's Method

To examine the rate of convergence, apply the Lipschitz condition, we have

$$\begin{aligned}
\left\|\nabla f\left(\mathbf{x}^{+}\right)\right\|_2 &= \left\|\nabla f\left(\mathbf{x}+\Delta\mathbf{x}_{\text{nt}}\right)-\nabla f\left(\mathbf{x}\right)-\nabla^2 f\left(\mathbf{x}\right)\Delta\mathbf{x}_{\text{nt}}\right\|_2 \\
&= \left\|\int_0^1\left(\nabla^2 f\left(\mathbf{x}+t\Delta\mathbf{x}_{\text{nt}}\right)-\nabla^2 f\left(\mathbf{x}\right)\right)\Delta\mathbf{x}_{\text{nt}}dt\right\|_2 \\
&\le \frac{L}{2}\left\|\Delta\mathbf{x}_{\text{nt}}\right\|_2^2 \\
&= \frac{L}{2}\left\|\nabla^2 f\left(\mathbf{x}\right)^{-1}\nabla f\left(\mathbf{x}\right)\right\|_2^2 \\
&\le \frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}\right)\right\|_2^2,
\end{aligned}$$

where the first inequality is from pp. 78 and the second uses CSI and the fact that for strongly convex function, $\frac{1}{m}\mathbf{I}\succeq f\left(\mathbf{x}\right)$. Note that the last equality is the inequality (9.33).

Classical Convergence Proof of Newton's Method

if $\left\|\nabla f\left(\mathbf{x}\right)\right\|_2 < \eta$, then $\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k+1)}\right)\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2\right)^2$ (9.33)

$\frac{f\left(\mathbf{x}^{(0)}\right)-p^\star}{\gamma}+6$ (9.37), $\gamma=\alpha\beta\eta^2\frac{m}{M^2}$ (9.38)

- In conclusion, the algorithm select unit steps and satisfies the condition (9.33) if $\left\|\nabla f\left(\mathbf{x}^{(k)}\right)\right\|_2 < \eta,$ where

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

- Substituting this bound and (9.38) into (9.37), we find that the number of iterations is bounded above by

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