Selecting Descent Direction

Principle of gradient methods

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

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$$

$D^k \in \mathbb{R}^{n\times n}$, $d^k \in \mathbb{R}^n$, and $x^k \in \mathbb{R}^n$.

The descent direction d^k satisfies $\nabla f(x^k)' d^k < 0$. Thus, the matrix D^k should (always) satisfy

$$\nabla f(x^k)' D^k \nabla f(x^k) > 0,$$

meaning _____

Different choice of D^k results in different methods introduced below.

Selecting Descent Direction

- Steepest Descent — Simple, slow convergence
- Newton's Method — Sophisticate, fast convergence
- Diagonally Scaled Steepest Descent — Diagonal approximation to $\nabla^2 f(x^k)^{-1}$
- Modified Newton's Method — $\nabla^2 f(x^0)^{-1}$ replaces $\nabla^2 f(x^k)^{-1}, \forall k$
- Discretized Newton's Method — Finite-difference based approximation to $\nabla^2 f(x^k)$
- Gauss-Newton's Method — for least squares problems

Steepest Descent

$$D^k = I, k = 0, 1, \ldots,$$

where I is the $n \times n$ identity matrix, and

$$d^k = -\nabla f(x^k), \forall k.$$

This idea arises from finding a direction d that solves

$$\min_d d^T \nabla f_k, \ s.t. ||d|| = 1$$

Since $d^T \nabla f_k = ||d|| \cdot ||\nabla f(x^k)|| \cos\theta = ||\nabla f(x^k)|| cos\theta$, where θ is the angle between d and $\nabla f(x^k)$, it is easy to see that the minimizer is attained when $\cos\theta = -1$ and

$$d = \frac{-\nabla f(x^k)}{||\nabla f(x^k)||}$$

Newton's Method

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

The idea in Newton's method is to minimize at each iteration the **quadratic approximation** of f around the current point x^k

$$\min_x f(x^k) + \nabla f(x^k)'(x - x^k) + \frac{1}{2}(x - x^k)'\nabla^2 f(x^k)(x - x^k),$$

Set the first derivative at 0

$$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0$$

Newton's Method

(continued)
The optimal $x = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$ provides a closed form expression of the next iterate x^{k+1}

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

A general form is written as

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

Diagonally Scaled Steepest Descent

D^k is a digonal matrix with ith entry be d_i^k and $d_i^k > 0$

The idea is to approximate the Newton's method by diagonal matrix. That is,

$$d_i^k \sim \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \right)^{-1}$$

Modified Newton's Method

$$D^k = \left(\nabla^2 f(x^0)\right)^{-1}, k = 0, 1, ...,$$

For D^k to be pd, $\nabla^2 f(x^0)$ is also pd.

This method is the same as Newton's method except that the Hessian matrix is not recalculated at each iteration.

Discretized Newton's Method

$$D^k = \left(H(x^k)\right)^{-1}, k = 0, 1, ...$$

where $H(x^k)$ is an approximation of $\nabla^2 f(x^k)$ formed by using finite difference approximation of the second derivatives, based on first derivatives or values of f.

It is required that $H(x^k)$ is pd and symmetric.

Gauss-Newton Method

This method is applicable specifically to the following optimization problem, often encountered in statistical data analysis and neural network training:

$$\begin{aligned} \min \quad & f(x) = \tfrac{1}{2}||g(x)||^2 = \tfrac{1}{2}\sum_{i=1}^{m}(g_i(x))^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \end{aligned}$$

where $g_1, \ldots, g_m$ are real valued functions. We choose

$$D^k = \left(\nabla g(x^k)\nabla g(x^k)'\right)^{-1}$$

The matrix $\nabla g(x^k)\nabla g(x^k)'$ is always psd.
Furthermore, if $\nabla g(x^k)$ has rank n, then $\nabla g(x^k)\nabla g(x^k)'$ is _____ and _____ (Proposition A.20)

Gauss-Newton Method

(continued)
The Gauss-Newton method may be viewed as an approximation to Newton's method. Because $\nabla f(x^k) = \nabla g(x^k) g(x^k)$, the iterate is

$$x^{k+1} = x^k - \alpha^k \left(\nabla g(x^k) \nabla g(x^k)' \right)^{-1} \nabla g(x^k) g(x^k)$$

Others

There are methods with other choices of D^k, such as the class of *Quasi-Newton methods*.

There are methods where the direction d^k is not of the form $-D^k \nabla f(x^k)$, such as *conjugate gradient method* and the *coordinate descent method*.

Selecting Stepsize

The selection of stepsize is usually *after* the determination of a search direction.

- Minimization Rule — α^k such that $f(x^k + \alpha^k d^k)$ is minimized
- Armijo rule — Start with s and continue with βs, $\beta^2 s$,..., until $\beta^m s$ falls within the feasible set of α satisfying
  $f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(x^k)' d^k$.
- Goldstein rule — The first successive stepsize reduction method, more complex than the Armijo rule.
- Constant stepsize — $\alpha^k = constant$
- Diminishing stepsize — $\alpha^k \to 0$ and satisfying
  $\sum_{k=0}^{\infty} \alpha^k = \infty$

Minimization Rule

A stepsize that minimizes the current cost function

$$\min_{\alpha \geq 0} f(x^k + \alpha d^k)$$

Or, the limited minimization rule

$$\min_{\alpha \in [0,s]} f(x^k + \alpha d^k)$$

The above two optimization problems are typically solved by the *one-dimensional line search algorithm*.

Minimization Rule—Line Search

line search? Search along a line with the direction

_____

Three practical methods to search along a line[1]:

- Cubic interpolation
- Quadratic interpolation
- The golden section method

[1] Now go to the Appendix C

Minimization Rule—Line Search

(continued)
In three types of line search methods, we consider minimization of the function

$$g(\alpha) = f(x + \alpha d)$$

The shape of the function $g(\alpha)$, however, is not known.

Three methods are three different "guesses" about the shape of $g(\alpha)$ from computable function value at a point a, $g(a)$, and/or the derivative $g'(a)$ at a.

Successive Stepsize Reduction

An initial stepsize s is chosen and if the corresponding vector $x^k + sd^k$ does not yield an improved value of f, the stepsize should be reduced.

The reduction of stepsize may repeat several times until the value of f is reduced.

While this method often works in practice, it is theoretically unsound because the cost improvement obtained at each iteration may not be substantial enough to guarantee convergence to a minimum:

Successive Stepsize Reduction

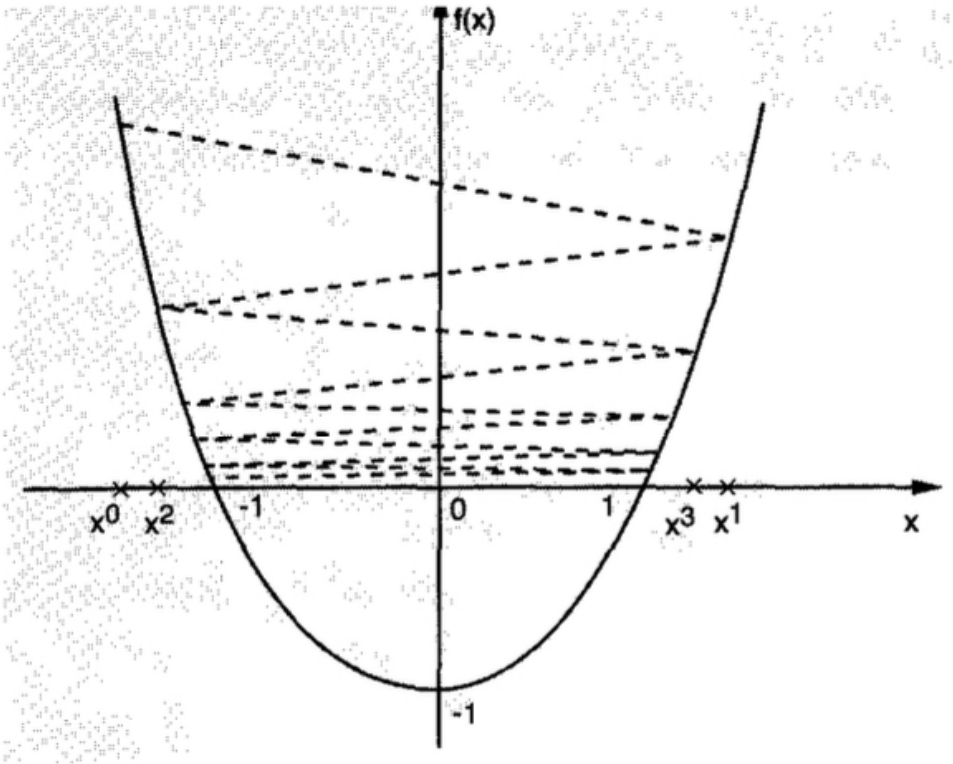


Figure 1.2.6. Example of failure of the successive stepsize reduction rule for the one-dimensional function

Successive Stepsize Reduction—Armijo Rule

The Armijo rule is essentially the successive reduction, suitably modified to eliminate the theoretical convergence difficulty as in the previous figure.

We set the stepsize $\alpha^k = \beta^{m_k} s$, where m_k is the first nonnegative integer m satisfying

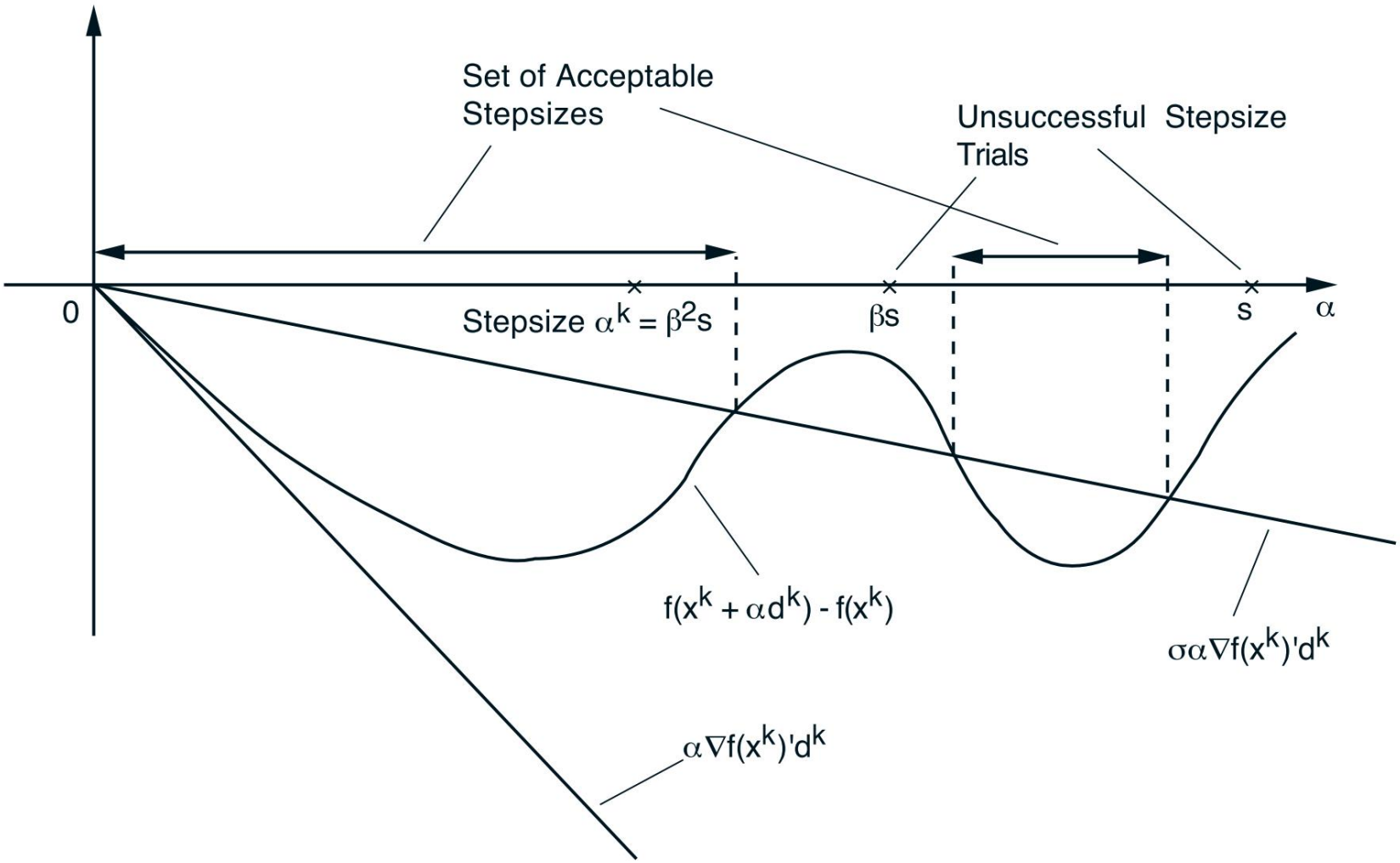
$$f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(x^k)' d^k.$$

And s, $\beta \in (0,1)$, $\sigma \in (0,1)$ are fixed scalars chosen by user[2].
In practice, the stepsizes $\beta^m s$, $m = 0, 1, ...$, are tried successively until the above *Armijo condition* is satisfied.

[2]Common selection: $s = 1$, σ close to zero

Successive Stepsize Reduction—Armijo Rule

The Armijo rule guarantees that the cost improvement must not just positive, but sufficiently large.



Goldstein Rule

The stepsize α^k is chosen to satisfy

$$\sigma \leq \frac{f(x^k + \alpha^k d^k) - f(x^k)}{\alpha^k \nabla f(x^k)' d^k} \leq 1 - \sigma,$$

where $\sigma \in (0, 1/2)$ is a fixed scalar.

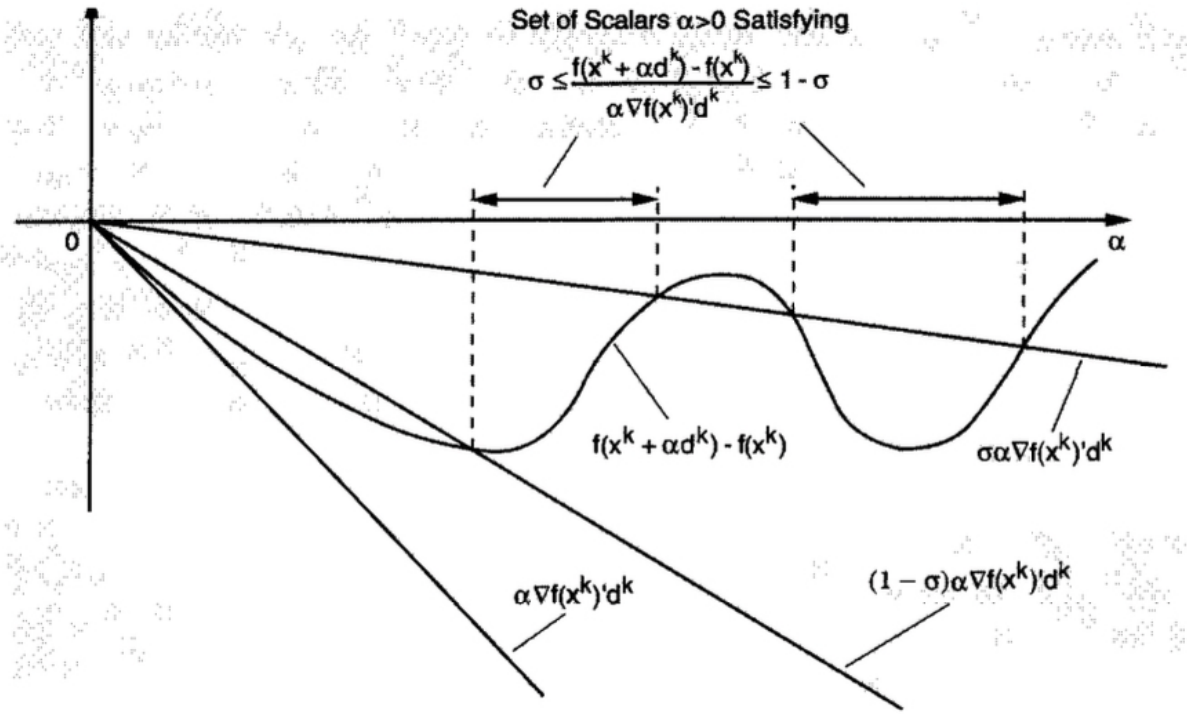


Figure 1.2.8. Illustration of the set of stepsizes that are acceptable in the Goldstein rule.

Goldstein Rule

(continued)
In practice the simpler Armijo rule seems to be universally preferred. The Goldstein rule is included here primarily because of its historical significance: it was the first sound proposal for a general-purpose stepsize rule that did not rely on line minimization.

Constant Stepsize

$$\alpha^k = s,$$

where $s > 0$ is a fixed scalar for all k.

Diminishing Stepsize

In such a design, the stepsize should converge to zero and shouldn't be too small

$$\alpha^k \to 0, \text{ and } \sum_{k=1}^{\infty} \alpha^k = \infty.$$

This stepsize rule is different than the proceeding ones in that it does not guarantee descent at each iteration.

The 2nd condition guarantees that $\{x^k\}$ *does not* converge to a nonstationary point. If $x^k \to \bar{x}$, then for any large indexes m and n $(m > n)$ we have

$$x^m \approx x^n \approx \bar{x}, \quad x^m \approx x^n - \left(\sum_{k=n}^{m-1} \alpha^k \right) \nabla f(\bar{x}),$$

Diminishing Stepsize

(continued)
Since $\sum_{k=n}^{m-1} \alpha^k$ can be made arbitrarily large, the above is a contradiction when $\bar{x}$ is nonstationary.(5 minutes)

The diminishing stepsize converges but the convergence rate tends to be slow. It is used primarily in situations where slow convergence is inevitable; for example, in singular problems or when the gradient is calculated with error.

Convergence to Stationary Point

Is the limit point of a sequence $\{x^k\}$ generated by a gradient method a stationary point?

- Only convergence to stationary points can be guaranteed
- Even convergence to a single limit may be hard to guarantee (capture theorem)
- Danger of nonconvergence if directions d^k tend to be orthogonal to $\nabla f(x^k)$
- Bounded eigenvalues condition
- Gradient related condition

Termination Criteria for infinite convergence

Generally, gradient methods are *not* finitely convergent.

To terminate the iteration:

$$||\nabla f(x^k)|| \leq \epsilon \text{ or } \frac{||\nabla f(x^k)||}{||\nabla f(x^0)||} \leq \epsilon,$$

where ϵ is a small positive scalar.

It is also possible to set the termination criterion as

$$||d^k|| \leq \epsilon.$$

Spacer Steps

In order to achieve the convergence, one inserts an iteration of a convergent algorithm finitely often to one another algorithm, then the theoretical convergence properties of the overall algorithm are quite satisfactory.

Such an inserted iteration is known as a *spacer step*.

Gradient Methods with Random and Nonrandom Errors

Occasionally, the gradient $\nabla f(x^k)$ is not computed exactly, and what is available is a vector

$$g^k = \nabla f(x^k) + e^k,$$

where e^k is an uncontrollable error vector. For example, embedded in a steepest descent method, the iterate is

$$x^{k+1} = x^k - \alpha^k g^k.$$

Convergence results for this setting in many cases are found analogous to those without errors.

Convergence Results

Theorem (Proposition 1.2.1: limit points for Gradient Methods)

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, and assume that $\{d^k\}$ is gradient related and α^k is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of $\{x^k\}$ is a stationary point.

Theorem (Proposition 1.2.2:)

The conclusions of Prop. 1.2.1 hold if $\{d^k\}$ is gradient related and α^k is chosen by the Goldstein rule.

Convergence Results

Theorem (Proposition 1.2.3: constant stepsize)

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that for some constant $L > 0$, we have

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^n,$$

and that for all k we have $d^k \neq 0$ and

$$\epsilon \leq \alpha^k \leq (2 - \epsilon)\bar{\alpha}^k,$$

where

$$\bar{\alpha}^k = \frac{|\nabla f(x^k)' d^k|}{L||d^k||^2},$$

and ϵ is a fixed positive scalar. Then every limit point of $\{x^k\}$ is a stationary point of f.

Convergence Results

Theorem (Proposition 1.2.4: diminishing stepsize)

Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$. Assume that for some constant $L > 0$, we have

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^n,$$

and that there exist positive scalars c_1, c_2 such that for all k we have

$$c_1||\nabla f(x^k)||^2 \leq -\nabla f(x^k)' d^k, \ ||d^k||^2 \leq c_2||\nabla f(x^k)||^2.$$

Suppose also that $\alpha^k \to 0$, $\sum_{k=0}^{\infty} \alpha^k = \infty$. Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f.

Convergence Results

Theorem (**Proposition 1.2.5: [Capture Theorem] tendency of unique limit point**)

Let f be continuously differentiable and let $\{x^k\}$ be a sequence satisfying $f(x^{k+1}) \leq f(x^k)$ for all k and generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, which is convergent in the sense that every limit point of sequence that it generates is a stationary point of f. Assume that there exist scalars $s > 0$ and $c > 0$ such that for all k there holds

$$\alpha^k \leq s, \ ||d^k|| \leq c||\nabla f(x^k)||.$$

Let x^* be a local minimum of f, which is the only stationary point of f within some open set. Then there exists an open set $\mathcal{S}$ containing x^* such that if $x^{\bar{k}} \in \mathcal{S}$ for some $\bar{k} \geq 0$, then $x^k \in \mathcal{S}$ for all $k \geq \bar{k}$ and $\{x^k\} \to x^*$. Furthermore, given any scalar $\bar{\epsilon} > 0$, the set $\mathcal{S}$ can be chosen so that $||x - x^*|| < \bar{\epsilon}$ for all $x \in \mathcal{S}$.

Convergence Results

Theorem (Proposition 1.2.6: spacer steps)

Consider a sequence $\{x^k\}$ *such that*

$$f(x^{k+1}) \leq f(x^k), k = 0, 1, \ldots$$

Assume that there exists an infinite set $\mathcal{K}$ *of integers for which*

$$x^{k+1} = x^k + \alpha^k d^k, \forall k \in \mathcal{K},$$

where $\{d^k\}_{\mathcal{K}}$ *is gradient related and* α^k *is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of the subsequence* $\{d^k\}_{\mathcal{K}}$ *is a stationary point.*

Go to proofs.

Speed of Convergence

Three approaches can be used to quantify the rate of convergence

1. **Computational complexity approach:** an upper bound of the number of required operations
2. **Informational complexity approach:** the number of *function (or gradient) evaluations* needed
3. **Local analysis:** Local analysis describes the behavior of a method near the solution by using Taylor series approximations, but ignores entirely the behavior of the method when far from the solution.

The Local Analysis Approach—Classes of convergence rate

Principle:

Suppose that there is a *unique* limit point x^* to which the sequences $\{x^k\}$ converges. An *error function* can be either defined as the Euclidean distance $e(x) = ||x - x^*||$ or the cost difference $e(x) = |f(x) - f(x^*)|$.

The sequence $\{e(x^k)\}$ is compared with the *geometric progression*

$$\beta^k, k = 0, 1, ...,$$

where $\beta \in (0, 1)$ is some scalar.

The Local Analysis Approach—Classes of convergence rate

Linear convergence

We say that $\{e(x^k)\}$ converges *linearly or geometrically*, if there exist $q > 0$ and $\beta \in (0, 1)$ such that for all k

$$e(x^k) \leq q\beta^k.$$

Alternatively, if for some $\beta \in (0, 1)$ we have

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)} \leq \beta.$$

that is, asymptotically, the error is decreasing by a factor of at least β at each iteration, then a *linear convergence* is obtained.

The Local Analysis Approach—Classes of convergence rate

Superlinear convergence

If for *every* $\beta \in (0,1)$, there exist q such that the condition $e(x^k) \leq q\beta^k$ holds for all k, we say that $\{e(x^k)\}$ converges *superlinearly*. If

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)} = 0.$$

To quantify the notion of superlinear convergence, we compare $\{e(x^k)\}$ with the sequence

$$\beta^{p^k}, k = 0, 1, \ldots,$$

where $\beta \in (0,1)$, and $p > 1$ are some scalars. We say that $\{e(x^k)\}$ converges *at least superlinearly with order* p, if there exist $q > 0$, $\beta \in (0,1)$, and $p > 1$ such that for all k

$$e(x^k) \leq q \cdot \beta^{p^k}.$$

The Local Analysis Approach—Classes of convergence rate

(continued) It is possible to show that superlinear convergence with order p is obtained if

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)^p} < \infty$$

or equivalently, $e(x^{k+1}) = O(e(x^k)^p)$.

Quadratic convergence

The case where $p = 2$ is referred to as *quadratic* convergence.

Convergence Rate of Steepest Descent for Quadratic Program

Suppose that the cost function f is quadratic with positive definite Hessian Q. WLOG, assume f is minimized at $x^* = 0$ and $f(x^*) = 0$ (Otherwise we can use $y = x - x^*$ and subtract the constant $f(x^*)$ from $f(x)$.)

Thus we have

$$f(x) = \frac{1}{2}x'Qx, \nabla f(x) = Qx, \nabla^2 f(x) = Q.$$

Let m: smallest eigenvalue of Q, and M: largest eigenvalue of Q.

Condition number

$$\frac{M}{m} : \text{condition number of } Q$$

Problems with large M/m are referred as *ill-conditioned*.

Convergence Rate of Steepest Descent for Quadratic Program

Three convergence rate results

1. For $x^k \neq 0$, we have

$$\frac{||x^{k+1}||}{||x^k||} \leq \max\{|1-\alpha^k m|, |1-\alpha^k M|\}$$

2. When α^k is chosen by the line minimization rule, we obtain

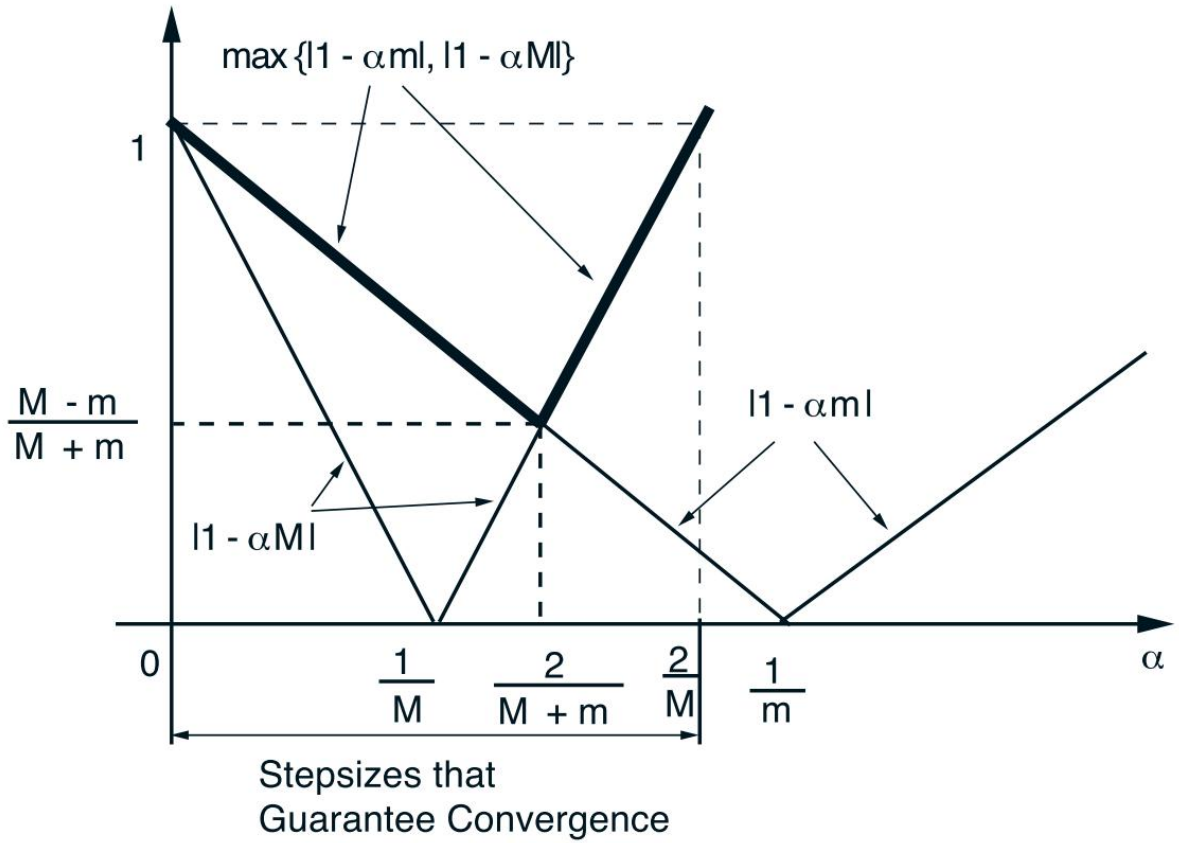
$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2.$$

3. The α^k that minimizes the bound of the condition 1. is $\alpha^* = 2/(M+m)$, in which case

$$\frac{||x^{k+1}||}{||x^k||} \leq \frac{M-m}{M+m}$$

as shown in the following figure.

Convergence Rate of Steepest Descent for Quadratic Program



Convergence Rate of Steepest Descent for Quadratic Program

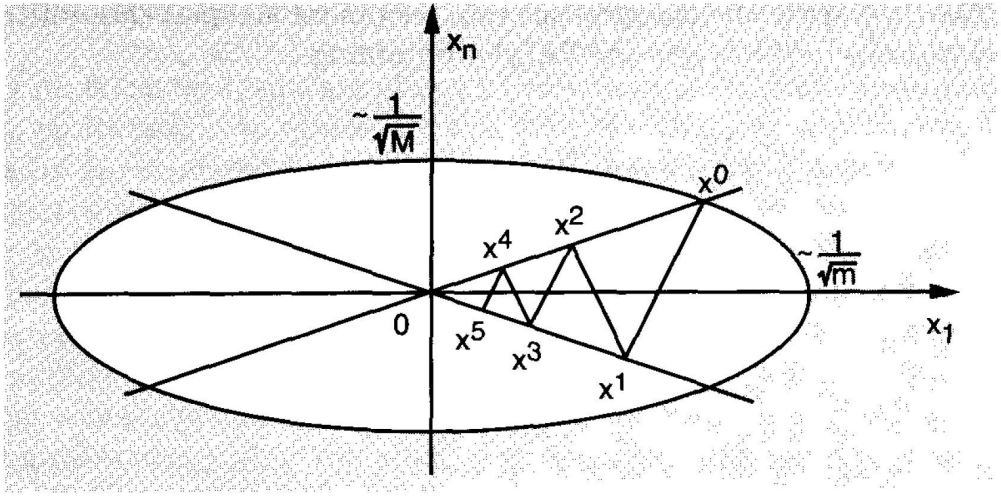


Figure 1.3.2. Example showing that the convergence rate bound

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

is sharp for the steepest descent method with the line minimization rule. Consider the quadratic function

$$f(x) = \tfrac{1}{2}\sum_{i=1}^{n} \lambda_i x_i^2,$$

where $0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M$. Any positive definite quadratic function can be put into this form by transformation of variables. Consider the starting point

$$x^0 = \left(m^{-1}, 0, \ldots, 0, M^{-1}\right)'$$

Convergence Rate of Steepest Descent for Quadratic Program

and apply the steepest descent method $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$ with α^k chosen by the line minimization rule. We have $\nabla f(x^0) = (1, 0, \ldots, 0, 1)'$ and it can be verified that the minimizing stepsize is $\alpha^0 = 2/(M+m)$. Thus we obtain $x_1^1 = 1/m - 2/(M+m)$, $x_n^1 = 1/M - 2/(M+m)$, $x_i^1 = 0$ for $i = 2, \ldots, n-1$. Therefore,

$$x^1 = \left(\frac{M-m}{M+m}\right)\left(m^{-1}, 0, \ldots, 0, -M^{-1}\right)'$$

and, we can verify by induction that for all k,

$$x^{2k} = \left(\frac{M-m}{M+m}\right)^{2k} x^0, \qquad x^{2k+1} = \left(\frac{M-m}{M+m}\right)^{2k} x^1.$$

Thus, there exist starting points on the plane of points x of the form $x = (\xi_1, 0, \ldots, 0, \xi_n)'$, $\xi_1 \in \Re$, $\xi_n \in \Re$, in fact two lines shown in the figure, for which steepest descent converges in a way that the inequality

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

is satisfied as an equation at each iteration.

Convergence Rate of Steepest Descent for Quadratic Program

Scaling

$x = Sy$. Analogous convergence results of $f(y)$ can be obtained.

Nonquadratic cost function and for convergence to nonsingular local minima

Nonquadratic cost function

Let f be twice continuously differentiable.

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k)$$

Assume

$$x^k \to x^*, \nabla f(x^*) = 0, \nabla^2 f(x^*) : \text{pd, and } x^k \neq x^* \forall k.$$

Let m^k : smallest eigenvalue of $(D^k)^{1/2} \nabla^2 f(x^k)(D^k)^{1/2}$,
M^k: largest eigenvalue of $(D^k)^{1/2} \nabla^2 f(x^k)(D^k)^{1/2}$.

Nonquadratic cost function and for convergence to nonsingular local minima

1. There holds

$$\limsup_{k\to\infty} \frac{(x^{k+1}-x^*)'(D^k)^{-1}(x^{k+1}-x^*)}{(x^k-x^*)'(D^k)^{-1}(x^k-x^*)}$$

$$= \limsup_{k\to\infty} \max\{|1-\alpha^k m^k|^2, |1-\alpha^k M^k|^2\}$$

2. If α^k is chosen by the minimization rule, there holds

$$\limsup_{k\to\infty} \frac{f(x^{k+1})-f(x^*)}{f(x^k)-f(x^*)}$$

$$\leq \limsup_{k\to\infty} \left(\frac{M^k-m^k}{M^k+m^k}\right)^2$$

3. When $D^k \to \nabla^2 f(x^*)^{-1}$, we have

$$\lim_{k\to\infty} M^k = \lim_{k\to\infty} m^k = 1$$

Nonquadratic cost function and for convergence to nonsingular local minima

Guideline:

- Choose the matrices D^k as close as possible to $(\nabla^2 f(x^*))^{-1}$, i.e., d^k approaches asymptotically the Newton direction, so the maximum and minimum eigenvalues of $(D^k)^{1/2}\nabla^2 f(x^*)(D^k)^{1/2}$ satisfy $M^k \approx 1$ and $m^k \approx 1$.
- Furthermore, the initial stepsize $s = 1$ is a good choice for the Armijo rule or a starting point of the minimization rule.
- Then a *superlinear* convergence rate is obtained

[one of the most reliable guidelines for designing algorithms for unconstrained NLP]

Difficult cost function and singular local minima

Singular local minimum:

Hessian matrix does not exist or not pd near or at local minima.

Difficult cost function:

(1) flat cost function

Given local minima x^* and direction d

$$\lim_{\alpha \to 0} \frac{\nabla f(x^* + \alpha d)' d - \nabla f(x^*)' d}{\alpha} = 0$$

(2) steep cost function

$$\lim_{\alpha \to 0} \frac{\nabla f(x^* + \alpha d)' d - \nabla f(x^*)' d}{\alpha} = \infty$$

—both with *infinite* condition number, thereby slower than linear convergence for steepest descent.

Difficult cost function and singular local minima

Convergence result:

For flat cost function f, but not for steep cost function, the gradient is *Lipschitz continuous*:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \text{ for some } L,$$

$\forall x, y$ in a neighborhood of x^*.

and there holds

$$f(x^k) - f(x^*) = o(1/k).$$

Difficult cost function and singular local minima

Tips:

- Sophisticated methods, such as Newton-like methods, work well for problems with nonsingular local minima.
- For problems with difficult cost function and singular local minima, simple methods, such as steepest descent with constant or diminishing stepsize, with supplemental features (e.g. heavy ball method) work better.

Convergence Rate Results

Theorem (Proposition 1.3.1)

Consider the quadratic function $f(x) = \frac{1}{2}x'Qx$, *where* Q *is positive definite and symmetric, and the method of steepest descent* $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$, *where the step size* α^k *is chosen according to the minimization rule* $f(x^k - \alpha^k \nabla f(x^k)) = \min_{\alpha \geq 0} f(x^k - \alpha \nabla f(x^k))$. *Then, for all* k,

$$f(x^{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(x^k),$$

where M *and* m *are the largest and smallest eigenvalues of* Q, *respectively.*

Convergence Rate Results

Theorem (Lemma 3.1: Kantorovich Inequality)

Let Q be positive definite and symmetric $n \times n$ matrix. Then for any vector $y \in \mathbb{R}^n$, $y \neq 0$, there holds

$$\frac{(y'y)^2}{(y'Qy)(y'Q^{-1}y)} \geq \frac{4Mm}{(M+m)^2},$$

where M and m are the largest and smallest eigenvalues of Q, respectively.

Convergence Rate Results

Theorem (**Proposition 1.3.2: Superlinear Convergence of Newton-like Methods**)

Let f be twice continuously differentiable. Consider a sequence $\{x^k\}$ generated by the gradient method $x^{k+1} = x^k + \alpha^k d^k$ and suppose that $x^k \to x^*$, $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$: positive definite. Assume further that $\nabla f(x^k) \neq 0$ for all k and

$$\lim_{k\to\infty} \frac{||d^k + (\nabla^2 f(x^*))^{-1}\nabla f(x^k)||}{||\nabla f(x^k)||} = 0$$

Then, if α^k is chosen by means of the Armijo rule with initial stepsize $s = 1$ and $\sigma < 1/2$, we have

$$\lim_{k\to\infty} \frac{||x^{k+1} - x^k||}{||x^k - x^*||} = 0.$$

Furthermore, there exists an integer $\bar{k} \geq 0$ such that $\alpha^k = 1$ for all $k \geq \bar{k}$ (i.e., eventually no reduction will be taking place.)

Convergence Rate Results

- Newton's method, combined with the Armijo rule with initial stepsize $= 1$, converges superlinearly.
- This setting of the Newton's method, however, converges only to a local minimum x^* at which $\nabla^2 f(x^*)$ is positive definite, whenever the starting point is sufficiently close to such a local minima.

Convergence Rate Results

Theorem (Proposition 1.3.3: Convergence rate of gradient methods for singular problems)

Suppose that the cost function f is convex and its gradient satisfies for some L the Lipschitz condition

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^n.$$

Consider a gradient method $x^{k+1} = x^k + \alpha^k d^k$ where α^k is chosen by the minimization rule, and for some $c > 0$ and all k we have

$$\nabla f(x^k)' d^k \leq -c||\nabla f(x^k)||||d^k||.$$

Suppose that the set of global minima of f is nonempty and bounded. Then

$$f(x^k) - f^* = o(1/k),$$

where $f^* = \min_x f(x)$ is the optimal value.

Go to proofs.