I. Function minimization

V. Blobel
Universität Hamburg

April 2001

1. One-dimensional minimization
2. Search methods
3. Unconstrained minimization
4. Constrained minimization

Simple search method

Find initial interval. Start with inital values x_1, x_2 with $f(x_1) > f(x_2)$ and evaluate function at

$$x_i = x_{i-1} + \alpha(x_{i-1} - x_{i-2}) \qquad i = 3, 4, \ldots \qquad \alpha > 1$$

until function value

$$f(x_i) > f(x_{i-1})$$

is found. Then a minimum is included in interval $[x_{i-2}, x_i]$.

Reduction of the interval. The interval $[x_{i-2}, x_i]$ is divided by the inner point x_{i-1} with

$$f(x_{i-1}) < f(x_{i-2}) \qquad f(x_{i-1}) < f(x_i)$$

Based on a function value an another inner point x_t the interval can be reduced,

- either to the interval $[x_i, x_{i-1}]$
- or to the interval $[x_{i-1}, x_t]$

Title

$$\frac{x_b - x_i}{x_b - x_a} = \frac{x_i - x_a}{x_b - x_i} = \tau \qquad \rightarrow \tau = \left(\sqrt{5} - 1\right)/2 = 0.618034$$

Define new point x_t by $(x_t - x_a)/(x_b - x_a) = \tau$:

New interval, depending on function value $f(x_t)$:

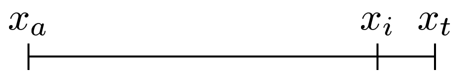
Reduction of length of interval by factor τ, for 10 function evaluations by factor $\tau^{10} \approx 0.01$, not dependent on function behaviour.

Title

Alternative: fit parabola to last three function values and use minimum of parabola as next point.

Many functions to be minimized are parabolic in good approximation $\rightarrow$ minimum of parabola often close to function minimum.

But: method can get stuck with unbalanced section of the interval.



Combined method: use mixture of parabola and golden section method, to avoid unbalanced section of the interval.

2. Search methods in n dimensions

Search methods in n dimensions do not require any derivatives, only functions values

Line search in one variable sequentially in all dimensions - usually rather inefficient

Simplex method by Nelder and Mead : simple, but makes use of earlier function evaluations in an efficient way

Monte Carlo search search for starting values for more efficient methods – meaningful if several local minima may exist

Simplex method in n dimensions

A simplex is formed by $n+1$ points in n-dimensional space

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n, \boldsymbol{x}_{n+1}$$

sorted such that function values $F_j = F(\boldsymbol{x}_j)$ are in the order

$$F_1 \leq F_2 \leq \ldots F_n \leq F_{n+1}$$

In addition: mean of best n points =

$$\text{center of gravity} \qquad \boldsymbol{c} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j$$

Method: sequence of cycles with new point in each cycle, replacing worst point x_{n+1}, with new (upodated) simplex in each cycle.

At start of each cycle new test point $\boldsymbol{x}_r$, obtained by reflexion of worst point at the center of gravity:

$$\boldsymbol{x}_r = \boldsymbol{c} + \alpha \cdot (\boldsymbol{c} - \boldsymbol{x}_{n+1})$$

A cycle in the Nelder and Mead method

Depending on value $F(x_r)$:

[$F_1 \le F_r \le F_n$]: Test point $\boldsymbol{x}_r$ is middle point and is added, the previous worst point is removed.

[$F_r \le F_1$]: Test point $\boldsymbol{x}_r$ is best point, search direction seems to be effective. A new point $\boldsymbol{x}_s = \boldsymbol{c} + \beta \cdot (\boldsymbol{x}_r - \boldsymbol{c})$ (with $\beta > 1$) is determined and the function value is F_s evaluated. For $F_s < F_r$ the extra step is successful, $\boldsymbol{x}_{n+1}$ is replaced by $\boldsymbol{x}_s$ otherwise by $\boldsymbol{x}_r$.

[$F_r > F_n$]: the simplex is too big and has to be reduced in size. For $F_r < F_{n+1}$ the test point $\boldsymbol{x}_r$ replaces the words point $\boldsymbol{x}_{n+1}$. A new test point $\boldsymbol{x}_s$ is defined by $\boldsymbol{x}_s = \boldsymbol{c} - \gamma(\boldsymbol{c} - \boldsymbol{x}_{n+1})$ (contraction) with $0 < \gamma < 1$. If this point $\boldsymbol{x}_s$ with $F_s = F(\boldsymbol{x}_s) < F_{n+1}$ is an improvement, then $\boldsymbol{x}_{n+1}$ is replaced by this point $\boldsymbol{x}_s$. Otherwise a new simplex in defined by replacing all points except $\boldsymbol{x}_1$ by $\boldsymbol{x}_j = \boldsymbol{x}_1 + \delta(\boldsymbol{x}_j - \boldsymbol{x}_1)$ for $j = 2, \ldots n+1$ with $0 < \delta < 1$, which requires n function evaluations.

3. Unconstrained minimization

$$\boxed{\text{minimize } F(\boldsymbol{x}) \qquad \boldsymbol{x} \in R^n}$$

Taylor expansion:

$$\begin{array}{lrcl} \text{function} & F(\boldsymbol{x}_k + \boldsymbol{\Delta x}) & \approx & F_k + \boldsymbol{g}_k^T \boldsymbol{\Delta x} + \frac{1}{2} \boldsymbol{\Delta x}^T \boldsymbol{H}_k \boldsymbol{\Delta x} \\ \text{derivative} & \boldsymbol{g}(\boldsymbol{x}_k + \Delta \boldsymbol{x}) & \approx & \boldsymbol{g}_k + \boldsymbol{H}_k \boldsymbol{\Delta x} \ , \end{array}$$

Function value and derivatives are evaluated at $\boldsymbol{x} = \boldsymbol{x}_k$:

$$\begin{array}{lrcll} \text{function} & F_k & = & F(\boldsymbol{x}_k) & \\ \text{gradient} & (g_k)_i & = & \left.\dfrac{\partial F}{\partial x_i}\right|_{\boldsymbol{x}=\boldsymbol{x}_k} & i = 1,\ 2,\ \ldots n \\ \text{Hesse matrix} & (H_k)_{ij} & = & \left.\dfrac{\partial^2 F}{\partial x_i \partial x_j}\right|_{\boldsymbol{x}=\boldsymbol{x}_k} & i,j = 1,\ 2,\ \ldots n \end{array}$$

Newton step

Step $\boldsymbol{\Delta x}_N$ determined from $\boldsymbol{g}_k + \boldsymbol{H}_k \boldsymbol{\Delta x} = 0$:

$$\boldsymbol{\Delta x}_N = -\boldsymbol{H}_k^{-1} \boldsymbol{g}_k \; .$$

For a quadratic function $F(\boldsymbol{x})$ the Newton step is in length and direction a step to the minimum of the function.

Sometimes large angle ($\approx 90°$) between Newton direction $\boldsymbol{\Delta x}_N$ and $\boldsymbol{g}_k$ (directions of steepest descent).

$$d = -\boldsymbol{g}_k^T \cdot \Delta \boldsymbol{x}_N = \boldsymbol{g}_k^T \boldsymbol{H}_k^{-1} \boldsymbol{g}_k = \boldsymbol{\Delta x}_N^T \boldsymbol{H}_k \boldsymbol{\Delta x}_N > 0$$

if Hessian positive-definite. For a quadratic function the distance to the minimum is $d/2$.

General iteration scheme

Define initial value $\boldsymbol{x}_0$, compute $F(\boldsymbol{x}_0)$ and set $k = 0$

1. **Test for convergence** If the conditions for convergence are satisfied, the algorithm terminates with $\boldsymbol{x}_k$ as the solution. The difference $F_{k-1} - F_k$ and d are used in the test.
2. **Compute a search vector.** A vector $\boldsymbol{\Delta x}$ is computed as the new search vector. The newton search vector is determined from $\boldsymbol{H}_k \boldsymbol{\Delta x}_N = -\boldsymbol{g}_k$.
3. **Line Search** A one-dimensional minimization is done for the function $f(z) = F(\boldsymbol{x}_k + z \cdot \boldsymbol{\Delta x}_N)$. and $z_{\min}$ is determined.
3. **Update.** The point $\boldsymbol{x}_{k+1}$ is defined by $\boldsymbol{x}_k + z_{\min} \cdot \boldsymbol{\Delta x}$, and k is increased by 1.

$$\text{Descent direction} \qquad \boldsymbol{g}_k^T \cdot \Delta \boldsymbol{x} < 0$$

Method of steepest descent

The search vector $\boldsymbol{\Delta x}$ is equal to the negative gradienten

$$\boldsymbol{\Delta x}_{\mathrm{SD}} = -\boldsymbol{g}_k$$

- Step seems to be natural choice;
- only gradient required (no Hesse matrix);
- no step size defined (in contrast to the Newton step);
- rate of convergence only linear:

$$\frac{F(\boldsymbol{x}_{k+1}) - F(\boldsymbol{x}^*)}{F(\boldsymbol{x}_k) - F(\boldsymbol{x}^*)} = c \approx \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 \quad ,$$

($\lambda_{\max}$ and $\lambda_{\min}$ are largest and smallest eigenvalue and κ condition number of Hesse matrix $\boldsymbol{H}$. For a large value of κ c close to one and slow convergence.

Optimal step size, if Hesse matrix known:

$$\boldsymbol{\Delta x}_{\mathrm{SD}} = -\left(\frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \boldsymbol{H}_k \boldsymbol{g}_k}\right) \boldsymbol{g}_k \quad .$$

Problem: Hessian not positive definit

Diagonalization:

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T \ ,$$

(vectors $\boldsymbol{u}_i$are the eigen vectors of $\boldsymbol{H}$, and $\boldsymbol{D}$ is diagonal, with eigenvalues λ_i $\boldsymbol{H}$ in diagonal. At least of the eigenvalues λ_i is zero or less than zero.

1. Strategy: define modified diagonal matrix $\overline{\boldsymbol{D}}$ with elements

$$\overline{\lambda}_i = \max\left(|\lambda_i|, \delta\right) \qquad \delta > 0$$

(which requires diagonalization) and modified positive-definite Hessian

$$\overline{\boldsymbol{H}} = \boldsymbol{U}\overline{\boldsymbol{D}}\boldsymbol{U}^T$$

Problem: Hessian not positive definit

2. Strategy: define modified Hessian by

$$\overline{\boldsymbol{H}} = \boldsymbol{H} + \alpha \boldsymbol{I}_n \qquad (\boldsymbol{I}_n = n \times n \text{ unit matrix}) \ .$$

(simpler)

- large value of α means search direction close to (negative) gradient (safe)
- small value of α means search direction close to Newton direction

factor α allows to interpolate between method of steepeset descent and Newton method.

What is large and what is small: scale defined by

$$\alpha \approx \frac{|\boldsymbol{g}^T \boldsymbol{H} \boldsymbol{g}|}{\boldsymbol{g}^T \boldsymbol{g}}$$

suggested by step of steepest descent method.