Matrix calculus. Line search.

Seminar

Optimization for ML. Faculty of Computer Science. HSE University

Theory recap. Differential

- Differential $df(x)[\cdot] : U \to V$ in point $x \in U$ for $f(\cdot) : U \to V$:

$$f(x+h) - f(x) = \underbrace{df(x)[h]}_{\text{differential}} + \overline{o}(||h||)$$

$U \to V$	$\mathbb{R}$	$\mathbb{R}^n$	$\mathbb{R}^{n\times m}$
$\mathbb{R}$	$f'(x)dx$	$\nabla f(x) dx$	$\nabla f(x) dx$
$\mathbb{R}^n$	$\nabla f(x)^T dx$	$J(x) dx$	—
$\mathbb{R}^{n\times m}$	$tr(\nabla f(X)^T dX)$	—	—

Theory recap. Differential

- Differential $df(x)[\cdot] : U \to V$ in point $x \in U$ for $f(\cdot) : U \to V$:

$$f(x + h) - f(x) = \underbrace{df(x)[h]}_{\text{differential}} + \bar{o}(||h||)$$

- Canonical form of the differential:

$U \to V$	$\mathbb{R}$	$\mathbb{R}^n$	$\mathbb{R}^{n \times m}$
$\mathbb{R}$	$f'(x)dx$	$\nabla f(x)dx$	$\nabla f(x)dx$
$\mathbb{R}^n$	$\nabla f(x)^T dx$	$J(x)dx$	—
$\mathbb{R}^{n \times m}$	$tr(\nabla f(X)^T dX)$	—	—

Theory recap. Differentiation Rules

- Useful differentiation rules and standard derivatives:

Differentiation Rules	Standard Derivatives
$dA = 0$	$d(\langle A,\, X\rangle) = \langle A,\, dX\rangle$
$d(\alpha X) = \alpha(dX)$	$d(\langle Ax,\, x\rangle) = \langle (A + A^T)x,\, dx\rangle$
$d(AXB) = A(dX)B$	$d(Det(X)) = Det(X)\langle X^{-T},\, dX\rangle$
$d(X + Y) = dX + dY$	$d(X^{-1}) = -X^{-1}(dX)X^{-1}$
$d(X^T) = (dX)^T$	
$d(XY) = (dX)Y + X(dY)$	
$d(\langle X,\, Y\rangle) = \langle dX,\, Y\rangle + \langle X,\, dY\rangle$	
$d\left(\frac{X}{\phi}\right) = \frac{\phi dX - (d\phi)X}{\phi^2}$	

Theory recap. Differential and Gradient / Hessian

We can retrieve the gradient using the following formula:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Theory recap. Differential and Gradient / Hessian

We can retrieve the gradient using the following formula:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Then, if we have a differential of the above form and we need to calculate the second derivative of the matrix/vector function, we treat "old" dx as the constant dx_1, then calculate $d(df) = d^2 f(x)$

Theory recap. Differential and Gradient / Hessian

We can retrieve the gradient using the following formula:

$$df(x) = \langle \nabla f(x), dx \rangle$$

Then, if we have a differential of the above form and we need to calculate the second derivative of the matrix/vector function, we treat "old" dx as the constant dx_1, then calculate $d(df) = d^2 f(x)$

$$d^2 f(x) = \langle \nabla^2 f(x) dx_1, dx \rangle = \langle H_f(x) dx_1, dx \rangle$$

Theory recap. Line Search

- Solution localization methods:

Theory recap. Line Search

- Solution localization methods:
    - Dichotomy search method

Theory recap. Line Search

- Solution localization methods:
    - Dichotomy search method
    - Golden selection search method

Theory recap. Line Search

- Solution localization methods:
    - Dichotomy search method
    - Golden selection search method
- Inexact line search:

Theory recap. Line Search

- Solution localization methods:
    - Dichotomy search method
    - Golden selection search method
- Inexact line search:
    - Sufficient decrease

Theory recap. Line Search

- Solution localization methods:
    - Dichotomy search method
    - Golden selection search method
- Inexact line search:
    - Sufficient decrease
    - Goldstein conditions

Theory recap. Line Search

- Solution localization methods:
  - Dichotomy search method
  - Golden selection search method
- Inexact line search:
  - Sufficient decrease
  - Goldstein conditions
  - Curvature conditions

Theory recap. Line Search

- Solution localization methods:
    - Dichotomy search method
    - Golden selection search method
- Inexact line search:
    - Sufficient decrease
    - Goldstein conditions
    - Curvature conditions
    - The idea behind backtracking line search

Matrix Calculus. Problem 1

Example

Find $\nabla f(x)$, if $f(x) = \frac{1}{2}x^T Ax + b^T x + c$.

Matrix Calculus. Problem 2

Example

Find $\nabla f(X)$, if $f(X) = tr(AX^{-1}B)$

Matrix Calculus. Problem 3

Example

Find the gradient $\nabla f(x)$ and hessian $\nabla^2 f(x)$, if $f(x) = \frac{1}{3}\|x\|_2^3$

Line Search. Example 1: Comparison of Methods (Colab ♣)

$$f_1(x) = x(x-2)(x+2)^2$$
$$[a, b] = [-3, 2]$$

Random search: 72 function calls. 36 iterations. $f_1^* = 0.09$
Binary search: 23 function calls. 13 iterations. $f_1^* = 10.00$
Golden search: 19 function calls. 18 iterations. $f_1^* = 10.00$
Parabolic search: 20 function calls. 17 iterations. $f_1^* = 10.00$

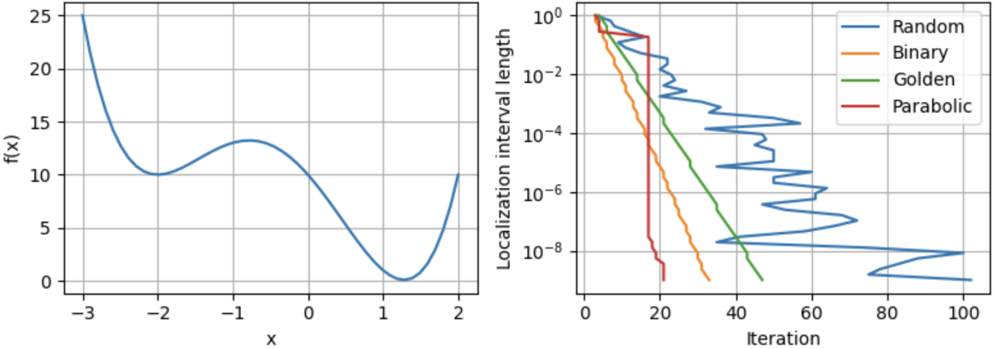

Figure 1: Comparison of different line search algorithms with f_1

Line Search. Example 1: Comparison of Methods (Colab ♣)

$$f_2(x) = -\sqrt{\frac{2}{\pi}}\frac{x^2 e^{-\frac{x^2}{8}}}{8}$$

$$[a, b] = [0, 6]$$

Random search: 68 function calls. 34 iterations. $f_2^* = 0.71$
Binary search: 23 function calls. 13 iterations. $f_2^* = 0.71$
Golden search: 20 function calls. 19 iterations. $f_2^* = 0.71$
Parabolic search: 17 function calls. 14 iterations. $f_2^* = 0.71$

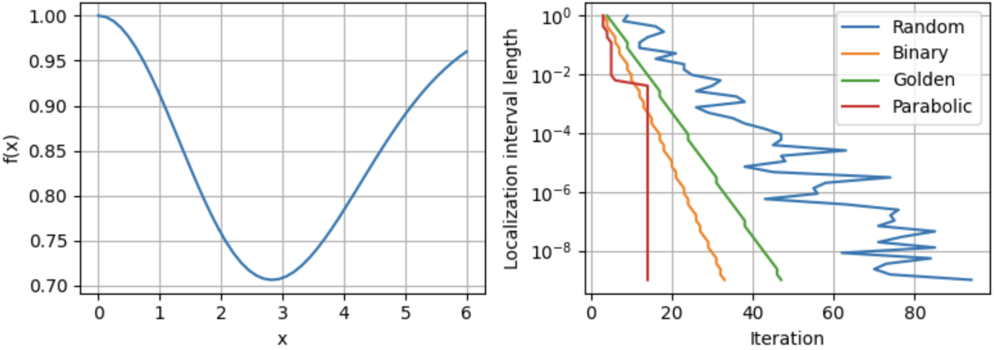

Figure 2: Comparison of different line search algorithms with f_2

Line Search. Example 1: Comparison of Methods (Colab ♣)

$$f_3(x) = \sin\left(\sin\left(\sin\left(\sqrt{\frac{x}{2}}\right)\right)\right)$$

$$[a, b] = [5, 70]$$

Random search: 66 function calls. 33 iterations. $f_3^* = 0.25$
Binary search: 32 function calls. 17 iterations. $f_3^* = 0.25$
Golden search: 25 function calls. 24 iterations. $f_3^* = 0.25$
Parabolic search: 103 function calls. 100 iterations. $f_3^* = 0.25$

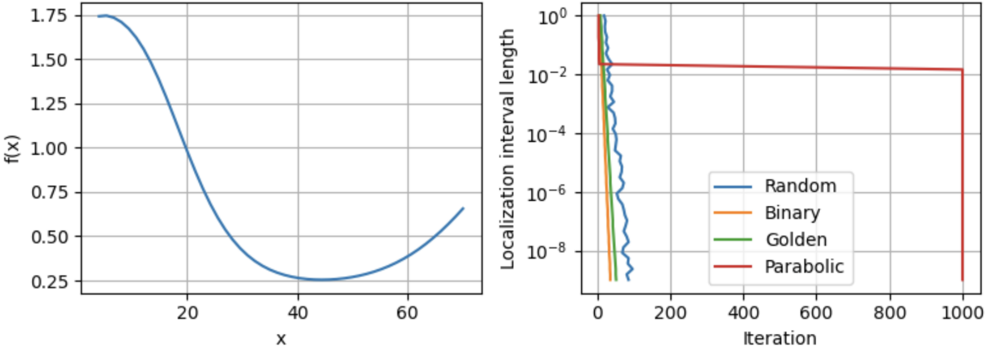

Figure 3: Comparison of different line search algorithms with f_3

Line Search. Example 2: The Brent Method

- Parabolic Interpolation + Golden Search = Brent Method

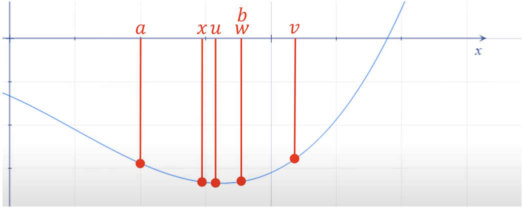

Figure 4: Idea of Brent Method

Line Search. Example 2: The Brent Method

- Parabolic Interpolation + Golden Search = Brent Method
- The key idea of the method is to track the value of the optimized scalar function at six points a, b, x, w, v, u

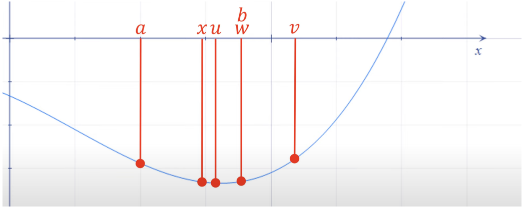

Figure 4: Idea of Brent Method

Line Search. Example 2: The Brent Method

- Parabolic Interpolation + Golden Search = Brent Method
- The key idea of the method is to track the value of the optimized scalar function at six points a, b, x, w, v, u
- $[a, b]$ – localization interval in the current iteration

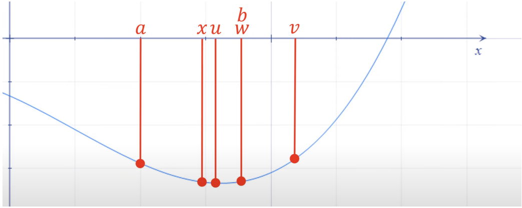

Figure 4: Idea of Brent Method

Line Search. Example 2: The Brent Method

- Parabolic Interpolation + Golden Search = Brent Method
- The key idea of the method is to track the value of the optimized scalar function at six points a, b, x, w, v, u
- $[a, b]$ – localization interval in the current iteration
- The pounts x, w and v such that the inequality $f(x) \leqslant f(w) \leqslant f(v)$ is valid

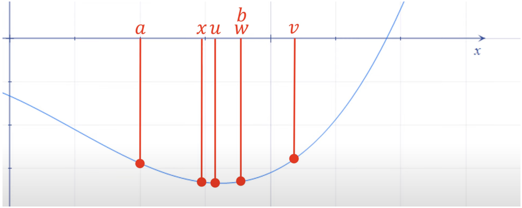

Figure 4: Idea of Brent Method

Line Search. Example 2: The Brent Method

- Parabolic Interpolation + Golden Search = Brent Method
- The key idea of the method is to track the value of the optimized scalar function at six points a, b, x, w, v, u
- $[a, b]$ — localization interval in the current iteration
- The pounts x, w and v such that the inequality $f(x) \leqslant f(w) \leqslant f(v)$ is valid
- u — minimum of a parabola built on points x, w and v or the point of the golden section of the largest of the intervals $[a, x]$ $[x, b]$.

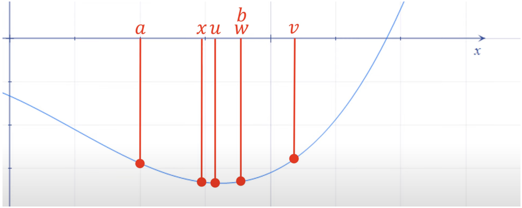

Figure 4: Idea of Brent Method

Line Search. Example 2: The Brent Method

A parabola is constructed only if the points x, w and v are different, and its vertex u^* is taken as the point u only if

- $u^* \in [a, b]$

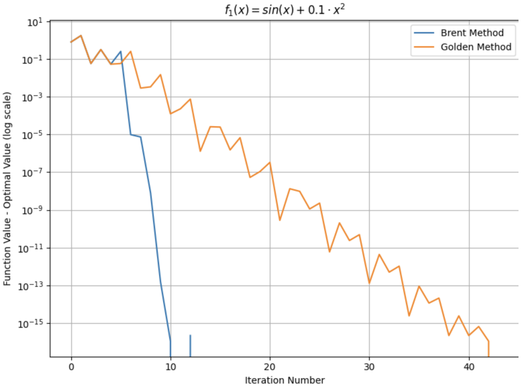

Figure 5: An example of how the Brent Method works

Line Search. Example 2: The Brent Method

A parabola is constructed only if the points x, w and v are different, and its vertex u^* is taken as the point u only if

- $u^* \in [a, b]$
- u^* is no more than half the length of the step that was before the previous one, from the point x

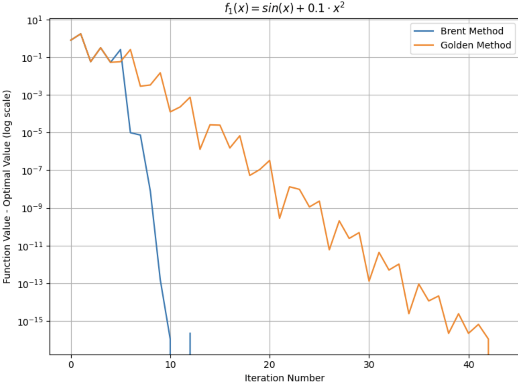

Figure 5: An example of how the Brent Method works

Line Search. Example 2: The Brent Method

A parabola is constructed only if the points x, w and v are different, and its vertex u^* is taken as the point u only if

- $u^* \in [a, b]$
- u^* is no more than half the length of the step that was before the previous one, from the point x
- If the conditions above are not met, then point u is located from the golden search

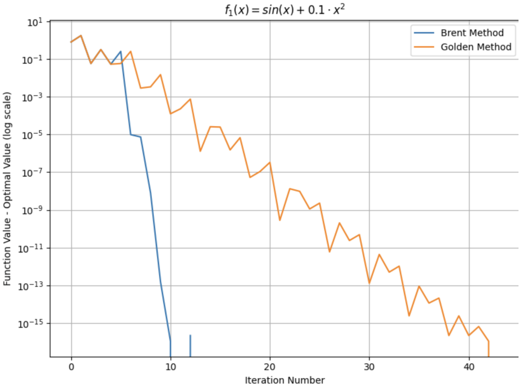

Figure 5: An example of how the Brent Method works

Line Search. Example 2: The Brent Method

A parabola is constructed only if the points x, w and v are different, and its vertex u^* is taken as the point u only if

- $u^* \in [a, b]$
- u^* is no more than half the length of the step that was before the previous one, from the point x
- If the conditions above are not met, then point u is located from the golden search
- Example In Colab ♣

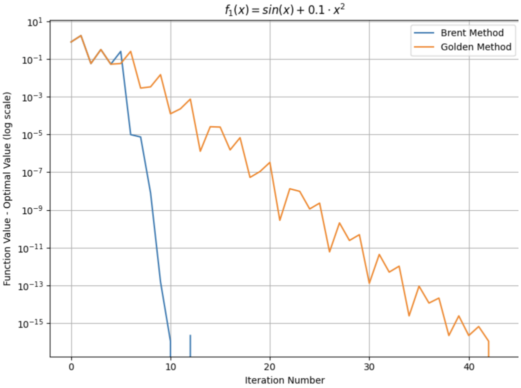

Figure 5: An example of how the Brent Method works