Numerical Optimization

Unconstrained Optimization

Shirish Shevade

Computer Science and Automation
Indian Institute of Science
Bangalore 560 012, India.

NPTEL Course on Numerical Optimization

Unconstrained Minimization Algorithm

(1) Initialize $\boldsymbol{x}^0$ and ϵ, set $k := 0$.

(2) **while** $\|\boldsymbol{g}(\boldsymbol{x}^k)\| > \epsilon$

  (a) Find a descent direction $\boldsymbol{d}^k$ for f at $\boldsymbol{x}^k$

  (b) Find $\alpha^k(> 0)$ along $\boldsymbol{d}^k$ such that

    (i) $f(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k) < f(\boldsymbol{x}^k)$

    (ii) α^k satisfies Armijo-Wolfe conditions

  (c) $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$

  (d) $k := k + 1$

endwhile

Output : $\boldsymbol{x}^* = \boldsymbol{x}^k$, a stationary point of $f(\boldsymbol{x})$.

Does this algorithm converge?

Consider the problem,

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x})$$

- Let $f \in \mathcal{C}^1$ and f be bounded below.
- An optimization algorithm to minimize $f(\boldsymbol{x})$ generates a sequence, $\{\boldsymbol{x}^k\}, k \geq 0$.
- Let the corresponding sequence of function values be $\{f^k\}, k \geq 0$.
- $f^{k+1} < f^k, \; k \geq 0$
- Stopping condition: $\|\boldsymbol{g}^k\| < \epsilon$

What can we say about $\|\boldsymbol{g}^k\|$ as $k \rightarrow \infty$?

Suppose, at every iteration k of the optimization algorithm,

- The direction $\boldsymbol{d}^k$ is chosen such that ${\boldsymbol{g}^k}^T\boldsymbol{d}^k < 0$
- Define $\phi(\alpha) = f(\boldsymbol{x}^k + \alpha\boldsymbol{d}^k)$. $\alpha^k(> 0)$ is chosen such that Armijo-Wolfe conditions are satisfied.

$$\begin{array}{rcl} f^{k+1} & \leq & f^k + c_1\alpha{\boldsymbol{g}^k}^T\boldsymbol{d}^k,\ c_1 \in (0,1) \\ \phi'(\alpha^k) & \geq & c_2\phi'(0),\ c_2 \in (c_1,1) \end{array}$$

- $f^{k+1} < f^k \ \forall\ k \geq 0$
- $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha^k\boldsymbol{d}^k$

Given: $f^{k+1} < f^k \; \forall \; k \geq 0$.
$\{f^k\}$: Monotonically decreasing sequence, which is also bounded below.
$\therefore \{f^k\} \rightarrow f^*$ where $f^* < \infty$.
$\therefore f^0 - f^k < \infty \;\; \forall \; k \geq 0$

$$\therefore \lim_{k \rightarrow \infty} f^0 - f^k < \infty$$

Using Armijo's condition, α^j's are chosen such that

$$\begin{aligned} f^{k+1} &\leq f^k + c_1 \alpha^k \boldsymbol{g}^{k^T} \boldsymbol{d}^k \\ &\leq f^0 + c_1 \sum_{j=0}^{k} \alpha^j \boldsymbol{g}^{j^T} \boldsymbol{d}^j \end{aligned}$$

Therefore,

$$\infty > f^0 - f^{k+1} \geq -c_1 \sum_{j=0}^{k} \alpha^j \boldsymbol{g}^{j^T} \boldsymbol{d}^j$$

$$\therefore\ -c_1 \sum_{j=0}^{\infty} \alpha^j \boldsymbol{g}^{j^T} \boldsymbol{d}^j < \infty$$

$$\therefore \sum_{j=0}^{\infty} \underbrace{-c_1}_{<0} \underbrace{\alpha^j}_{>0} \underbrace{\boldsymbol{g}^{j^T} \boldsymbol{d}^j}_{<0} < \infty$$

Therefore, sum of infinitely many positive terms is *finite*.
This implies, beyond certain iteration k, $\alpha^k \boldsymbol{g}^{k^T} \boldsymbol{d}^k = 0$.
Using Wolfe condition, α^k is chosen such that

$$\begin{aligned}
\phi'(\alpha^k) &\geq c_2\, \phi'(0),\ c_2 \in (c_1, 1) \\
\therefore \boldsymbol{g}^{k+1^T} \boldsymbol{d}^k &\geq c_2\, \boldsymbol{g}^{k^T} \boldsymbol{d}^k \\
\therefore (\boldsymbol{g}^{k+1} - \boldsymbol{g}^k)^T \boldsymbol{d}^k &\geq (c_2 - 1) \boldsymbol{g}^{k^T} \boldsymbol{d}^k
\end{aligned}$$

Let $\boldsymbol{g}$ be Lipschitz continuous. That is, $\exists\, L, 0 < L < \infty$ such that

$$\|\boldsymbol{g}^{k+1} - \boldsymbol{g}^k\| \leq L\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|$$

But, we have, $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$.

$$\begin{aligned} \therefore \|\boldsymbol{g}^{k+1} - \boldsymbol{g}^k\| &\leq L\alpha^k\|\boldsymbol{d}^k\| \\ \therefore (\boldsymbol{g}^{k+1} - \boldsymbol{g}^k)^T\boldsymbol{d}^k &\leq L\alpha^k{\boldsymbol{d}^k}^T\boldsymbol{d}^k \end{aligned}$$

But, using Wolfe conditions, $(\boldsymbol{g}^{k+1} - \boldsymbol{g}^k)^T\boldsymbol{d}^k \geq (c_2 - 1){\boldsymbol{g}^k}^T\boldsymbol{d}^k$. Therefore,

$$\begin{aligned} \alpha^k &\geq \frac{c_2 - 1}{L}\frac{{\boldsymbol{g}^k}^T\boldsymbol{d}^k}{\|\boldsymbol{d}^k\|^2} \\ \therefore \alpha^k{\boldsymbol{g}^k}^T\boldsymbol{d}^k &\leq \frac{c_2 - 1}{L}\frac{({\boldsymbol{g}^k}^T\boldsymbol{d}^k)^2}{\|\boldsymbol{d}^k\|^2} \\ \therefore\ -c_1\alpha^k{\boldsymbol{g}^k}^T\boldsymbol{d}^k &\geq c_1\frac{(1 - c_2)}{L}\frac{({\boldsymbol{g}^k}^T\boldsymbol{d}^k)^2}{\|\boldsymbol{d}^k\|^2} \end{aligned}$$

$$-c_1\alpha^k {\boldsymbol{g}^k}^T \boldsymbol{d}^k \;\; \geq \;\; c_1 \frac{(1-c_2)}{L} \frac{({\boldsymbol{g}^k}^T \boldsymbol{d}^k)^2}{\|\boldsymbol{d}^k\|^2}$$

Let θ_k be the angle between $\boldsymbol{g}^k$ and $\boldsymbol{d}^k$. Therefore,

$$\begin{aligned} -c_1\alpha^k {\boldsymbol{g}^k}^T \boldsymbol{d}^k \;\; &\geq \;\; c_1 \frac{(1-c_2)}{L} \frac{\|\boldsymbol{g}^k\|^2 \|\boldsymbol{d}^k\|^2 \cos^2\theta_k}{\|\boldsymbol{d}^k\|^2} \\ \therefore \; -c_1\alpha^k {\boldsymbol{g}^k}^T \boldsymbol{d}^k \;\; &\geq \;\; c_1 \frac{(1-c_2)}{L} \|\boldsymbol{g}^k\|^2 \cos^2\theta_k \end{aligned}$$

But, using Armijo's conditions, $-c_1 \sum_{k=0}^{\infty} \alpha^k {\boldsymbol{g}^k}^T \boldsymbol{d}^k < \infty$. Therefore,

$$c_1 \frac{(1-c_2)}{L} \sum_{k=0}^{\infty} \|\boldsymbol{g}^k\|^2 \cos^2\theta_k \;\; < \;\; \infty$$

$$c_1 \frac{(1 - c_2)}{L} \sum_{k=0}^{\infty} \|\boldsymbol{g}^k\|^2 \cos^2\theta_k \quad < \quad \infty$$

This implies

$$\|\boldsymbol{g}^k\|^2 \cos^2\theta_k \rightarrow 0.$$

If, at every iteration, $\boldsymbol{d}^k$ is chosen such that,

$$\boldsymbol{g}^{k^T} \boldsymbol{d}^k < 0 \text{ and } \cos^2\theta_k \geq \delta > 0,$$

then, we have,

$$\lim_{k \to \infty} \|\boldsymbol{g}^k\| = 0.$$

Global Convergence Theorem

Global Convergence Theorem [*Zoutendijk*]

Consider the problem to minimize $f(\boldsymbol{x})$ over $\mathbb{R}^n$. Suppose f is bounded below in $\mathbb{R}^n$, $f \in \mathcal{C}^1$ and the gradient, $\nabla f(= \boldsymbol{g})$ is Lipschitz continuous. If at every iteration k of an optimization algorithm, a descent direction $\boldsymbol{d}^k$ is chosen such that $\cos^2 \theta_k > \delta(> 0)$ (where θ_k is the angle between $\boldsymbol{d}^k$ and $\boldsymbol{g}^k$) and α^k satisfies Armijo-Wolfe conditions, then the optimization algorithm either *terminates in a finite number of iterations* or

$$\lim_{k \to \infty} \|\boldsymbol{g}^k\| = 0.$$

Sufficient Decrease and Backtracking

- Armijo-Goldstein Conditions: Choose α^k such that

$$\phi_2(\alpha^k) \leq f(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k) \leq \phi_1(\alpha^k)$$

where $\phi_1(\alpha) = f(\boldsymbol{x}^k) + c_1 \alpha \boldsymbol{g}^{k^T} \boldsymbol{d}^k, \;\; c_1 \in (0,1)$ and $\phi_2(\alpha) = f(\boldsymbol{x}^k) + c_2 \alpha \boldsymbol{g}^{k^T} \boldsymbol{d}^k, \;\; c_2 \in (c_1, 1)$.

- Use of *backtracking* line search with Armijo's condition

Backtracking Line Search

---

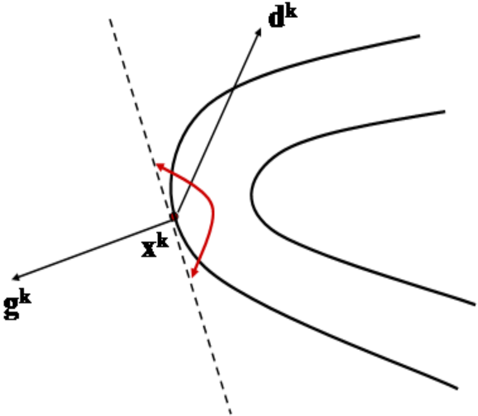
(1) Choose $\hat{\alpha}(> 0), \rho \in (0,1), c_1 \in (0,1)$. Set $\alpha = \hat{\alpha}$.

(2) **while** $f(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k) > f(\boldsymbol{x}^k) + c_1 \alpha \boldsymbol{g}^{k^T} \boldsymbol{d}^k$

$\alpha := \rho\alpha$

endwhile

Output : $\alpha^k = \alpha$



- Descent direction set: $\{\boldsymbol{d} \in \mathbb{R}^n : \boldsymbol{g}^{k^T}\boldsymbol{d} < 0\}$ where $\boldsymbol{g}^k = \boldsymbol{g}(\boldsymbol{x}^k)$

Descent Directions

- Let $\boldsymbol{g}^k \neq \boldsymbol{0}$ and $\boldsymbol{d}^k = -\boldsymbol{A}^k \boldsymbol{g}^k$ where $\boldsymbol{A}^k$ is a symmetric matrix
- If $\boldsymbol{A}^k$ is positive definite,

$$\begin{aligned} \boldsymbol{g}^{k^T} \boldsymbol{d}^k &= -\boldsymbol{g}^{k^T} \boldsymbol{A}^k \boldsymbol{g}^k < 0 \\ &\Rightarrow \boldsymbol{d}^k \text{ is a descent direction} \end{aligned}$$

- $\boldsymbol{d}^k = -\boldsymbol{A}^k \boldsymbol{g}^k$ is a *descent direction* if $\boldsymbol{A}^k$ is positive definite.
- Different optimization algorithms use different $\boldsymbol{A}^k$

How to find $\boldsymbol{d}^k$?

Consider the first order approximation to $f(x)$ about $\boldsymbol{x}^k$:

$$f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}) \stackrel{\Delta}{=} f(\boldsymbol{x}^k) + {\boldsymbol{g}^k}^T(\boldsymbol{x} - \boldsymbol{x}^k) = f(\boldsymbol{x}^k) + {\boldsymbol{g}^k}^T\boldsymbol{d}$$

Maximum decrease in $\hat{f}(\boldsymbol{x})$ is possible by solving (**P1**):

$$\begin{array}{ll} \min_{\boldsymbol{d}} & {\boldsymbol{g}^k}^T\boldsymbol{d} \\ \text{s.t.} & \boldsymbol{d}^T\boldsymbol{d} = 1 \end{array}$$

Let θ_k be the angle between $\boldsymbol{g}^k$ and $\boldsymbol{d}$.

$$\begin{aligned} {\boldsymbol{g}^k}^T\boldsymbol{d} &= \|\boldsymbol{g}^k\|\|\boldsymbol{d}\|\cos\theta_k \\ &= \|\boldsymbol{g}^k\|\cos\theta_k \quad (\because \boldsymbol{d}^T\boldsymbol{d} = 1) \end{aligned}$$

Therefore, the solution to the problem (**P1**) is $\boldsymbol{d}^k = -\boldsymbol{g}^k/\|\boldsymbol{g}^k\|$

Steepest Descent Method

- Uses the steepest descent direction, $\boldsymbol{d}^k = -\boldsymbol{g}^k$

Steepest Descent Algorithm

(1) Initialize $\boldsymbol{x}^0$ and ϵ, set $k := 0$.

(2) **while** $\|\boldsymbol{g}^k\| > \epsilon$

  (a) $\boldsymbol{d}^k = -\boldsymbol{g}^k$

  (b) Find $\alpha^k (> 0)$ along $\boldsymbol{d}^k$ such that

    (i) $f(\boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k) < f(\boldsymbol{x}^k)$

    (ii) α^k satisfies Armijo-Wolfe conditions

  (c) $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha^k \boldsymbol{d}^k$

  (d) $k := k + 1$

endwhile

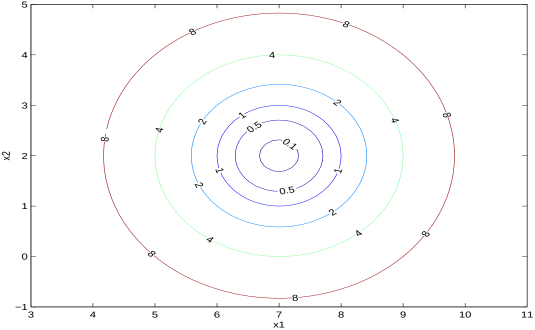
Output : $\boldsymbol{x}^* = \boldsymbol{x}^k$, a stationary point of $f(\boldsymbol{x})$.

- Exact or Backtracking line search can be used in step 2(b)

Example:

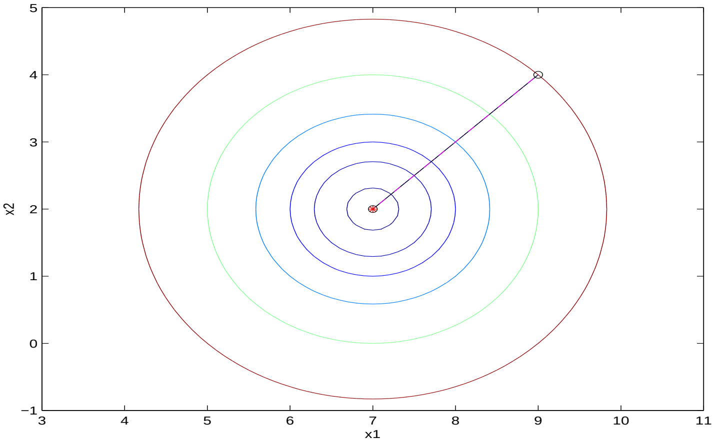
$$\min \ f(\boldsymbol{x}) \triangleq (x_1 - 7)^2 + (x_2 - 2)^2$$

- $\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} 2(x_1 - 7) \\ 2(x_2 - 2) \end{pmatrix}$, $\boldsymbol{H}(\boldsymbol{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$.
- $\boldsymbol{x}^* = \begin{pmatrix} 7 \\ 2 \end{pmatrix}$



Example:

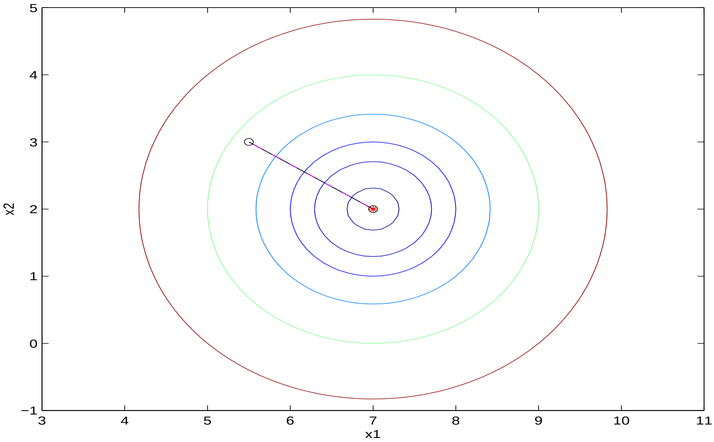
$$\min \quad f(\boldsymbol{x}) \triangleq (x_1 - 7)^2 + (x_2 - 2)^2$$



Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (9, 4)^T$

Example:

$$\min \quad f(\boldsymbol{x}) \triangleq (x_1 - 7)^2 + (x_2 - 2)^2$$

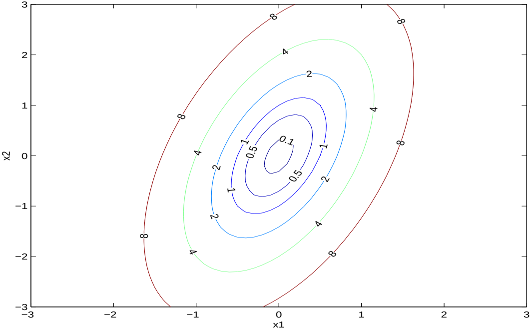


Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (5.5, 3)^T$

Example:

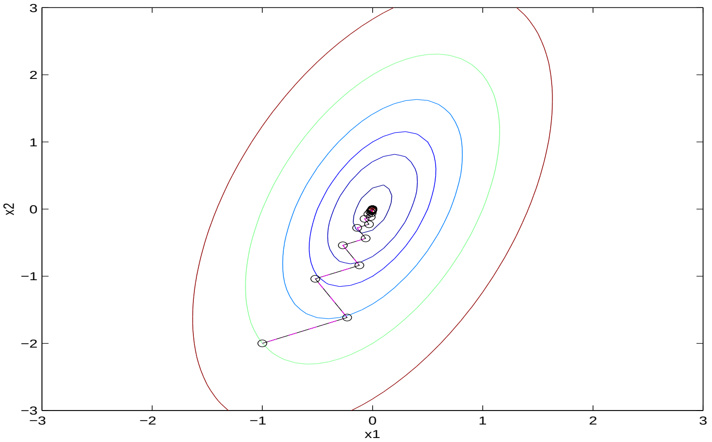
$$\min \; f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$

- $\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} 8x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{pmatrix}$, $\boldsymbol{H}(\boldsymbol{x}) = \begin{pmatrix} 8 & -2 \\ -2 & 2 \end{pmatrix}$.
- $\boldsymbol{x}^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$



Example:

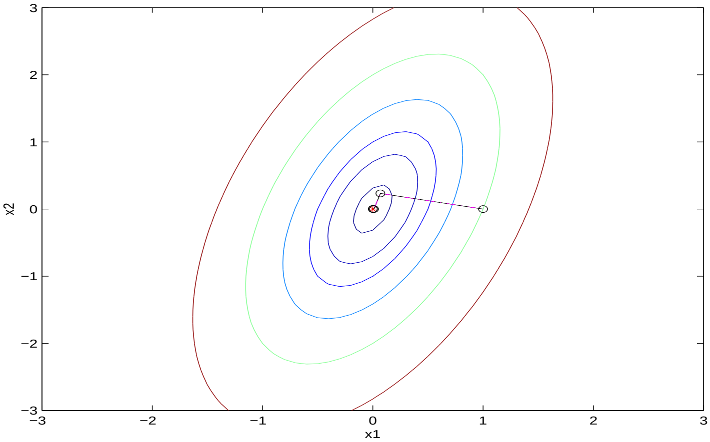
$$\min \quad f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$



Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (-1, -2)^T$

Example:

$$\min \quad f(\boldsymbol{x}) \triangleq 4x_1^2 + x_2^2 - 2x_1x_2$$

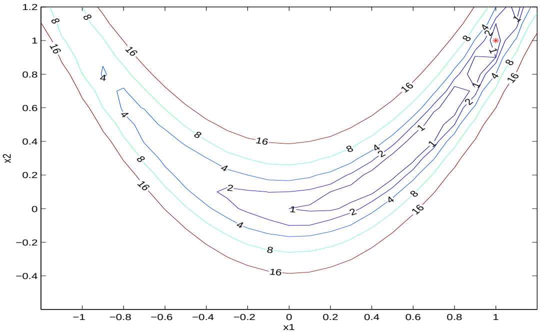


Behaviour of the steepest descent algorithm (with exact line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (1,0)^T$

Example (Rosenbrock function):

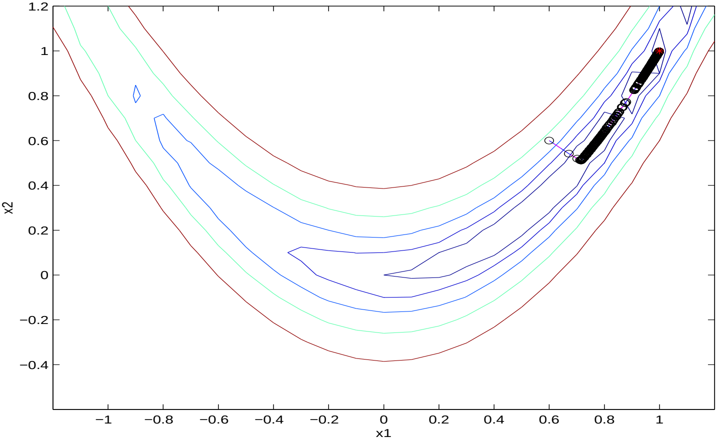
$$\min \quad f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

- $\boldsymbol{x}^* = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$



Example (Rosenbrock function):

$$\min \quad f(\boldsymbol{x}) \stackrel{\Delta}{=} 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$



Behaviour of the steepest descent algorithm (with backtracking line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (0.6, 0.6)^T$

Example (Rosenbrock function):

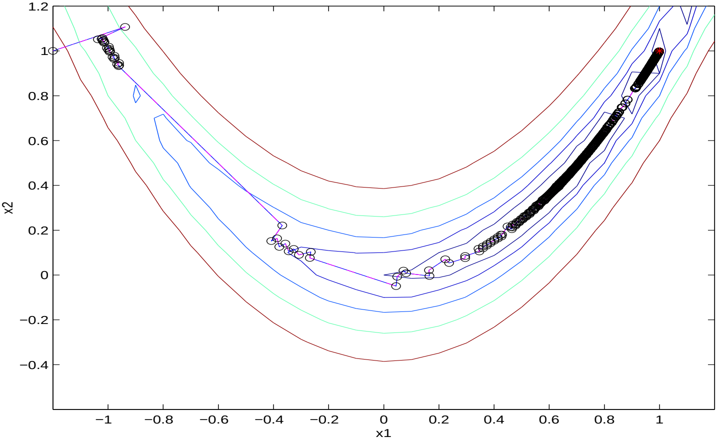
$$\min \quad f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

k	x_1^k	x_2^k	$f(\boldsymbol{x}^k)$	$\|\boldsymbol{x}^k - \boldsymbol{x}^*\|$	$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)$
0	0.6	0.6	5.92	0.5657	5.92
10	0.72	0.52	0.0792	0.5601	0.0782
100	0.78	0.61	0.0465	0.4414	0.0465
1000	0.9914	0.9828	7.45×10^{-5}	0.0192	7.45×10^{-5}
2028	0.9989	0.9978	1.81×10^{-6}	0.0024	1.81×10^{-6}

Table: Steepest descent method (with backtracking line search) applied to Rosenbrock function, using $\boldsymbol{x}^0 = (0.6, 0.6)^T$.

Example (Rosenbrock function):

$$\min \ f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$



Behaviour of the steepest descent algorithm (with backtracking line search) applied to $f(\boldsymbol{x})$ using $\boldsymbol{x}^0 = (-1.2, 1)^T$

Example (Rosenbrock function):

$$\min \; f(\boldsymbol{x}) \triangleq 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

k	x_1^k	x_2^k	$f(\boldsymbol{x}^k)$	$\Vert\boldsymbol{x}^k - \boldsymbol{x}^*\Vert$	$f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)$
0	-1.2	1.0	24.2	2.2	24.2
10	-1.00	1.01	4.02	2.0042	4.02
100	0.57	0.32	0.1867	0.80	0.1867
1000	0.99	0.97	1.99×10^{-4}	0.0314	1.99×10^{-4}
2300	0.9989	0.9979	1.11×10^{-6}	0.0024	1.11×10^{-6}

Table: Steepest descent method (with backtracking lines search) applied to Rosenbrock function, using $\boldsymbol{x}^0 = (-1.2, 1)^T$.