The Steepest Descent Line Search

Qi Gong
Associate Professor
Dept. of Applied Math & Stats.
Baskin School of Engineering
University of California
Santa Cruz, CA

The Steepest Descent Line Search

choose x_0 and tolerance $tol > 0$
$k \leftarrow 0$
evaluate $\nabla f(x_0)$

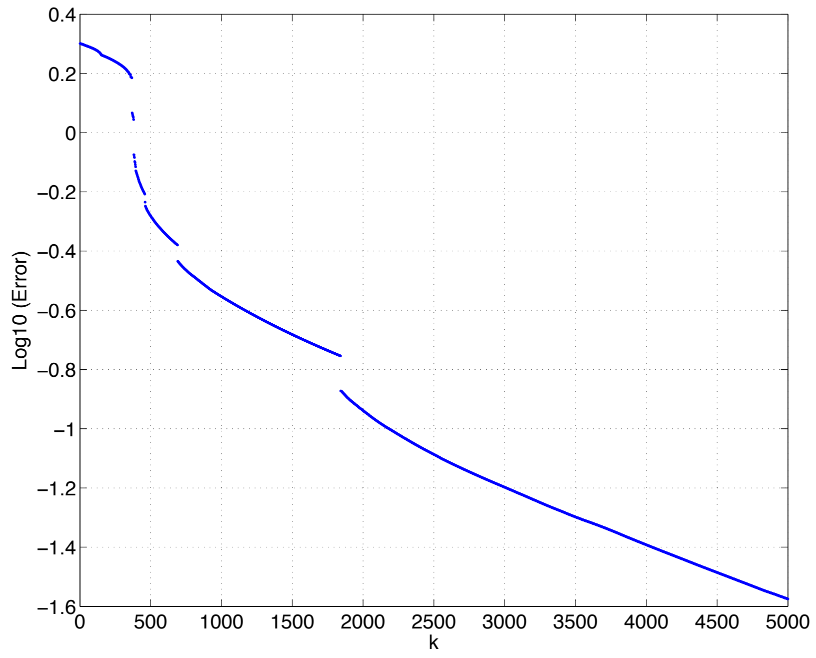
while $\|\nabla f(x_k)\| > tol$
    $p = -\nabla f(x_k)$
    $\alpha \leftarrow$ Wolfe conditions
    $x_{k+1} = x_k + \alpha p$
    $k = k + 1$
    Evaluate $\nabla f(x_k)$
end

- The search direction is set to be the negative gradient.
- The global convergence is guaranteed by Zoutendijk Theorem.
- Along the negative gradient search direction the cost function decreases most rapidly.
- But the convergence rate is slow.

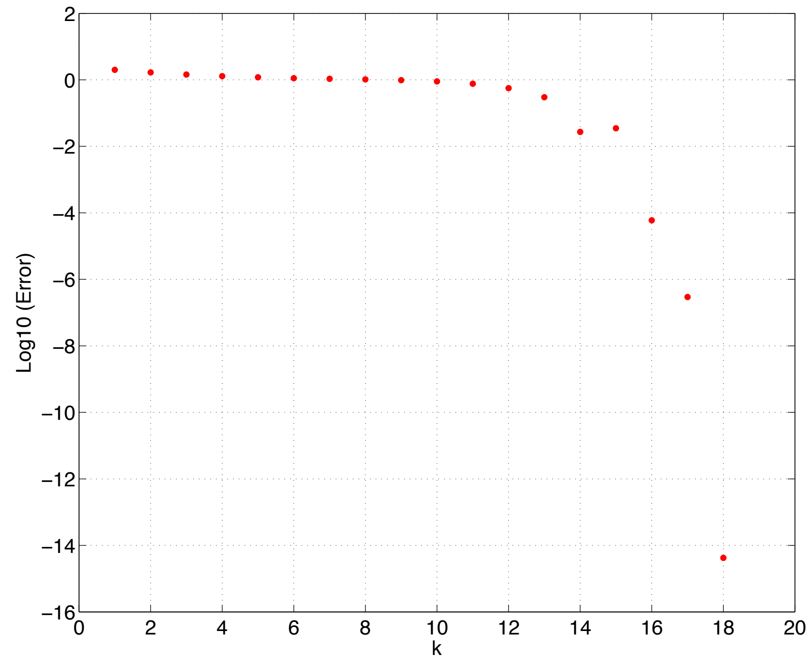
Convergence Comparison

$$\text{Min. } f(x_1, x_2) = (cx_1 - 2)^4 + x_2^2(cx_1 - 2)^2 + (x_2 + 1)^2, \quad c = 80$$

$$\Longrightarrow \quad x_1^* = 2/c, \; x_2^* = -1$$



Steepest Descent Line Search



Newton’s Method

Rate of Convergence – Quadratic Cost Function

Quadratic Example:

$$\boxed{\begin{array}{l}\text{Min. } f(x) = \frac{1}{2}x^T Q x - b^T x \\ \text{where } Q = Q^T \text{ is positive definite.}\end{array}}$$

Exact line search:

$$\begin{aligned} p_k &= -\nabla f(x_k) = -Qx_k + b \\ \alpha_k &= arg \min_{\alpha} f(x_k + \alpha p_k) = \frac{p_k^T p_k}{p_k^T Q p_k} \\ x_{k+1} &= x_k + \alpha_k p_k \end{aligned}$$

Convergence performance?

Convergence Analysis – Quadratic Cost Function

Let $x^* = Q^{-1}b$ (global min), and define $e_k = x_k - x^*$.

The error, e_k, is measured in Q-norm.

$$\|v\|_Q = \sqrt{v^T Q v}, \text{ where } v \in R^n, \text{ and } Q \text{ is positive definite.}$$

$$e_{k+1} = x_{k+1} - x^* = x_k + \alpha_k p_k - x^* = e_k + \alpha_k p_k \quad \Longrightarrow$$

$$\begin{aligned}
\|e_{k+1}\|_Q^2 &= (e_k^T + \alpha_k p_k^T) Q (e_k + \alpha_k p_k) \\
&= \|e_k\|_Q^2 + 2\alpha_k p_k^T Q e_k + \alpha_k^2 p_k^T Q p_k \\
&= \boxed{\left[1 - \frac{\left(p_k^T p_k\right)^2}{\left(p_k^T Q p_k\right)\left(p_k^T Q^{-1} p_k\right)}\right]} \|e_k\|_Q^2
\end{aligned}$$

substitute

$$\alpha_k = \frac{p_k^T p_k}{p_k^T Q p_k}$$

Convergence Analysis – Quadratic Cost Function

Kantorovich Inequality: Let Q be an $n \times n$ symmetric and positive definite matrix. For all $x \in R^n$,

$$\frac{\left(x^T x\right)^2}{\left(x^T Q x\right)\left(x^T Q^{-1} x\right)} \geq \frac{4\lambda_1 \lambda_n}{\left(\lambda_1 + \lambda_n\right)^2},$$

where λ_1 is the smallest eigenvalue of Q and λ_n is the largest eigenvalue of Q.

$$\Longrightarrow \quad 1 - \frac{\left(p_k^T p_k\right)^2}{\left(p_k^T Q p_k\right)\left(p_k^T Q^{-1} p_k\right)} \leq 1 - \frac{4\lambda_1 \lambda_n}{\left(\lambda_1 + \lambda_n\right)^2} = \left(\frac{\lambda_n - \lambda_1}{\lambda_1 + \lambda_n}\right)^2$$

$$\Longrightarrow \quad \|e_{k+1}\|_Q \leq \frac{\lambda_n - \lambda_1}{\lambda_1 + \lambda_n} \|e_k\|_Q$$

$\kappa = \lambda_n / \lambda_1 \geq 1$ is the condition number of Q.

Read Appendix A of the textbook on condition number and matrix norms

Convergence Analysis – Quadratic Cost Function

$$\boxed{\|e_{k+1}\|_Q \leq \frac{\kappa-1}{\kappa+1}\|e_k\|_Q, \ (\kappa = \lambda_n/\lambda_1 \geq 1)} \Longrightarrow$$

- $\lim_{k\to\infty} \|e_k\|_Q = 0$, i.e., $x_k \to x^*$ (global convergence)
- The rate of convergence is linear with $r = \dfrac{\kappa-1}{\kappa+1} \in (0,1)$

> Let $\{x_k\}$ be a sequence in $\mathbb{R}^n$ that converges to x^*. We say that the convergence is *Q-linear* if there is a constant $r \in (0, 1)$ such that
>
> $$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \leq r, \quad \text{for all } k \text{ sufficiently large.} \qquad \text{(A.34)}$$

- $\|e_k\|_Q \leq r\|e_{k-1}\|_Q \leq r^2\|e_{k-2}\|_Q \leq \cdots \leq r^k\|e_0\|_Q$

$$\Longrightarrow \log\|e_k\|_Q \leq \log r^k + \log\|e_0\|_Q = \boxed{(\log r)\,k} + \log\|e_0\|_Q$$

linear in k with a negative slope

- When condition number of Q , i.e., κ is large, $r \approx 1$.
The convergence is very slow.

Convergence Result – General Nonlinear Cost Function

Theorem 3.4.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, and that the iterates generated by the steepest-descent method with exact line searches converge to a point x^* at which the Hessian matrix $\nabla^2 f(x^*)$ is positive definite. Let r be any scalar satisfying

$$r \in \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}, 1 \right),$$

where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\nabla^2 f(x^*)$. Then for all k sufficiently large, we have

$$f(x_{k+1}) - f(x^*) \leq r^2 [f(x_k) - f(x^*)].$$

- The reduction of the cost function is slow if r is close to 1.
- In general we don't expect the rate of convergence will improve if inexact line search (like Wolfe conditions) is used.
- The slow convergence is due to the choice of search direction.