

ISEN 629: Engineering Optimization

Lecture 3

Sergiy Butenko

Industrial and Systems Engineering
Texas A& M University

Fall 2007

1/21

Second order conditions for unconstrained problems

Theorem (SONC for an unconstrained problem)

If $x^* \in \mathbb{R}^n$ is a local minimizer for the problem $\min_{x \in \mathbb{R}^n} f(x)$, where $f(x) \in C^{(2)}(\mathbb{R}^n)$, then $\nabla^2 f(x^*)$ is positive semidefinite. \square

Theorem (SOSC for an unconstrained problem)

If x^* satisfies the FONC and SONC for an unconstrained problem $\min_{x \in \mathbb{R}^n} f(x)$ and $\nabla^2 f(x^*)$ is positive definite, then x^* is a point of strict local minimum for this problem.

2/21

Example: A convex quadratic problem

Consider a quadratic problem

$$\min_{x \in \mathbb{R}^n} q(x),$$

where $q(x) = \frac{1}{2}x^T Qx + c^T x$. If Q is a positive semidefinite matrix, then $q(x)$ is convex and any point satisfying the FONC is a global minimizer. We have

$$\nabla q(x) = 0 \Leftrightarrow Qx = -c.$$

If Q is positive definite, then this system has a unique solution $x^* = -Q^{-1}c$, which is the only global minimizer in this case.

3/21

Level sets

- ▶ For a function $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ and a constant c , the set $S = \{x \in \mathbb{R}^n : f(x) = c\}$ is called the *level set* of f at the level c .
- ▶ A curve γ in a set S is $\gamma = \{x(t) : t \in (a, b)\} \subset S$, where $x(t) : (a, b) \rightarrow S$ is a continuous function.
- ▶ For any curve γ in the level set S of $f(x)$ at the level c , we have

$$f(x(t)) = c, \text{ for } t \in (a, b).$$

Hence, the derivative of $g(t) = f(x(t))$ at any point $t_0 \in (a, b)$ is

$$\frac{dg(t_0)}{dt} = 0.$$

On the other hand, using the chain rule,

$$\frac{dg(t_0)}{dt} = \frac{df(x(t_0))}{dt} = \nabla f(x(t_0))^T x'(t_0).$$

So, if we denote by $x_0 = x(t_0)$, then we have

$$\nabla f(x_0)^T x'(t_0) = 0.$$

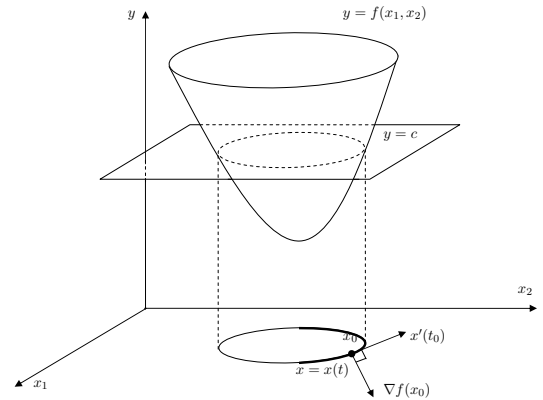
4/21

Level sets

- ▶ Geometrically, this means that vectors $\nabla f(x_0)$ and $x'(t_0)$ are orthogonal.
- ▶ Note that $x'(t_0)$ represents the tangent line to $x(t)$ at x_0 .
- ▶ Thus, for a continuously differentiable function $f(x)$ and a smooth (continuously differentiable) curve $x(t)$ passing through x_0 in the level set of $f(x)$ at the level $c = f(x_0)$, where $x(t_0) = x_0$, $x'(t_0) \neq 0$, the gradient of f at x_0 is orthogonal to the tangent line of $x(t)$ at x_0 .

5/21

Level sets



6/21

Gradient Methods

Given a problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

the numerical methods usually aim to construct a sequence of points $\{x_k : k \geq 0\}$, such that $x_k \rightarrow x^*$, $k \rightarrow \infty$, where x^* is a stationary point of $f(x)$ ($\nabla f(x^*) = 0$). Each next point in this sequence is obtained from the previous point by moving some distance along a feasible direction d :

$$x_{k+1} = x_k + \alpha_k d, \quad k \geq 0.$$

In gradient methods, the direction d depends on the gradient $\nabla f(x_k)$.

7/21

Gradient Methods

- ▶ Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x),$$

where $f(x)$ is a continuously differentiable function.

- ▶ Given $x_0 \in \mathbb{R}^n$ and a direction $d \in \mathbb{R}^n$, the directional derivative of $f(x)$ at x_0 is $\nabla f(x_0)^T d$.
- ▶ Recall that this directional derivative is interpreted as the rate of increase of $f(x)$ at x_0 in the direction d .
- ▶ The Cauchy-Schwartz inequality, stating that for any two vectors $u, v \in \mathbb{R}^n$: $u^T v \leq \|u\| \|v\|$ with equality if and only if $u = \alpha v$ for some scalar $\alpha \geq 0$, allows to find the direction with the largest possible rate of increase.

8/21

Gradient Methods

- ▶ Applying this inequality for d and $\nabla f(x_0)$, we have

$$\nabla f(x_0)^T d \leq \|\nabla f(x_0)\| \|d\|,$$

where equality is possible if and only if $d = \alpha \nabla f(x_0)$ with $\alpha \geq 0$. So, the direction $d = \alpha \nabla f(x_0)$ is the direction of the maximum rate of increase at x_0 .

- ▶ Similarly, for d and $-\nabla f(x_0)$, we have

$$\nabla f(x_0)^T d \geq -\|\nabla f(x_0)\| \|d\|,$$

where equality is possible if and only if $d = -\alpha \nabla f(x_0)$ with $\alpha \geq 0$. So, the direction $d = -\alpha \nabla f(x_0)$ is the direction of the maximum rate of decrease at x_0 . Thus, intuitively, the direction opposite to the gradient is the “best” direction to take in a minimization method.

9/21

Gradient Methods

The general outline of a gradient method is

$$x_{k+1} = x_k - \alpha_k \nabla_k, \quad k \geq 0,$$

where $\alpha_k \geq 0$ and $\nabla_k = \nabla f(x_k)$. Different choices of $\alpha_k, k \geq 0$ result in different variations of gradient methods, but in general α_k is chosen so that the *descent property* is satisfied:

$$f(x_{k+1}) < f(x_k), \quad k \geq 0.$$

10/21

Gradient Methods

Next we show that if $\nabla_k \neq 0$, then α_k can always be chosen such that the sequence $\{f(x_k) : k \geq 0\}$ possesses the descent property. Using Taylor's series, we have

$$f(x_{k+1}) = f(x_k) + \nabla_k^T (x_{k+1} - x_k) + o(\|x_{k+1} - x_k\|).$$

Since $x_{k+1} = x_k - \alpha_k \nabla_k$, we have

$$f(x_{k+1}) - f(x_k) = -\alpha_k \|\nabla_k\|^2 + o(\alpha_k \|\nabla_k\|).$$

This yields that there always exists $\bar{\alpha} > 0$ such that for any positive $\alpha_k \leq \bar{\alpha}$:

$$f(x_{k+1}) - f(x_k) < 0.$$

11/21

Steepest Descent

In the steepest descent method, the step size α_k corresponds to the largest decrease in the objective while moving along the direction $\nabla f(x_k)$ from point x_k :

$$\alpha_k : f(x_k + \alpha_k \nabla_k) = \min_{\alpha \geq 0} f(x_k - \alpha \nabla_k),$$

i.e., $\alpha_k = \arg \min_{\alpha \geq 0} f(x_k - \alpha \nabla_k)$. From the above, it is obvious that such choice of α_k will guarantee the descent property, since

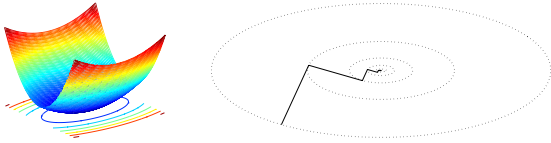
$$f(x_{k+1}) \leq f(x_k - \bar{\alpha} \nabla_k) < f(x_k).$$

Theorem

If $x_k \rightarrow x^*$, where $\{x_k : k \geq 0\}$ is the sequence generated by the steepest descent method, then $\nabla f(x^*) = 0$.

12/21

Steepest Descent



13/21

Convex quadratic case

Consider a convex quadratic function

$$f(x) = \frac{1}{2}x^T Qx + c^T x,$$

where Q is a positive definite matrix. Then, for some x_k ,

$$\nabla_k = \nabla f(x_k) = Qx_k + c$$

and the $k + 1^{\text{st}}$ iteration of the steepest descent method is

$$x_{k+1} = x_k - \frac{\nabla_k^T \nabla_k}{\nabla_k^T Q \nabla_k} \nabla_k, \quad k \geq 0.$$

14/21

Convex quadratic case

Example: For $f(x) = \sum_{i=1}^n x_i^2 = x^T x = \|x\|^2$, for any $x_0 \in \mathbb{R}^n$ we have $Q = I_n$, where I_n is the $n \times n$ identity matrix, and

$$x_1 = x_0 - \frac{4x_0^T x_0}{8x_0^T x_0} 2x_0 = x_0 - x_0 = 0.$$

Thus, we get the global minimizer in one step in this case.

15/21

Speed of convergence

Consider a numerical sequence $\{x_k : k \geq 0\}$ such that $x_k \rightarrow x^*$, $k \rightarrow \infty$. If there exist $C > 0$ and positive integers R, K , such that for any $k \geq K$:

$$\frac{|x_{k+1} - x^*|}{|x_k - x^*|^R} \leq C,$$

then we say that $\{x_k : k \geq 0\}$ has the rate of convergence R . If $R = 1$, we say that the convergence is linear; if $R = 2$, the convergence is quadratic.

If $\lim_{k \rightarrow \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = 0$, the sequence is said to converge superlinearly.

If $\lim_{k \rightarrow \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \mu$, does not hold for any $\mu < 1$ the sequence is said to converge sublinearly.

16/21

Global convergence of the steepest descent method

A numerical method is said to be globally convergent if it converges starting from any point. We discuss the global convergence analysis of the steepest descent method for convex quadratic case only. Consider a convex quadratic function

$$q(x) = \frac{1}{2}(x - x^*)^T Q (x - x^*)$$

The steepest descent iteration for this function is

$$x_{k+1} = x_k - \alpha_k \nabla_k,$$

where

$$\alpha_k = \frac{\nabla_k^T \nabla_k}{\nabla_k^T Q \nabla_k}.$$

It can be shown that

$$q(x_{k+1}) \leq q(x_k) \left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \right),$$

where $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$ are the smallest and the largest eigenvalues of Q , respectively.

17/21

Global convergence of the steepest descent method

$$q(x_{k+1}) \leq q(x_k) \left(1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \right),$$

- ▶ Thus, the steepest descent method is globally convergent for a convex quadratic function. Note that the rate of convergence is linear.
- ▶ From the above inequality we also see that if $\lambda_{\min}(Q) = \lambda_{\max}(Q)$ then we will have convergence in one step (recall the example for $f(x) = x_1^2 + x_2^2$).
- ▶ On the other hand, if $\lambda_{\max}(Q)$ is much larger than $\lambda_{\min}(Q)$, then $1 - \frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)} \approx 1$ and the convergence may be extremely slow in this case.
- ▶ The ratio $k(Q) = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} = \|Q\| \|Q^{-1}\|$ is called the *condition number* of matrix Q . A matrix with a large condition number is called poorly conditioned. This case usually corresponds to "long, narrow" level sets, where the steepest descent moves back and forth ("zigzags") in search for the minimizer.

18/21

Newton's method

As before, we consider the unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x).$$

The Newton's method is based on minimizing the quadratic Taylor's series approximation of $f(x)$ instead of $f(x)$. We have

$$f(x) \approx f(x_k) + \nabla_k^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla_k^2 (x - x_k),$$

where $\nabla_k = \nabla f(x_k)$ and $\nabla_k^2 = \nabla^2 f(x_k)$. If ∇_k^2 is positive definite then the global minimizer of the quadratic approximation

$$q(x) = f(x_k) + \nabla_k^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla_k^2 (x - x_k)$$

is given by

$$x^* = x_k - (\nabla_k^2)^{-1} \nabla_k.$$

Setting $x_{k+1} = x^*$, we obtain an iteration of the Newton's method:

$$x_{k+1} = x_k - (\nabla_k^2)^{-1} \nabla_k, \quad k \geq 0.$$

19/21

Newton's method

Example: Using the Newton's iteration for a convex quadratic function

$$f(x) = \frac{1}{2}x^T Qx + c^T x,$$

we obtain

$$x_{k+1} = x_k - Q^{-1}(Qx_k + c) = -Q^{-1}c.$$

Thus, we get the global minimizer in one step.

20/21

Newton's method: Convergence

Newton's method has the quadratic rate of convergence under certain assumptions. We assume that $f \in C^{(3)}(\mathbb{R}^n)$, ∇_k^2 and $\nabla^2 f(x^*)$ are positive definite, where x^* is a stationary point of $f(x)$ and x_k is a point close to x^* .

Then

$$\begin{aligned}\|x_{k+1} - x^*\| &= \|x_k - (\nabla_k^2)^{-1} \nabla_k - x^*\| \\ &= \|(\nabla_k^2)^{-1} (\nabla f(x^*) - \nabla_k - \nabla_k^2 (x^* - x_k))\| \\ &\leq \|(\nabla_k^2)^{-1}\| \cdot \|\nabla_k - \nabla f(x^*) - \nabla_k^2 (x_k - x^*)\| \\ &\leq c_2 c_1 \|x_k - x^*\|^2.\end{aligned}$$

So, if we start close enough from the stationary point x^* , then the Newton's method converges to x^* with the quadratic rate of convergence.