Numerical Optimization

Lecture 5: Optimality Conditions and Spectral Properties of Matrices

Sangkyun Lee

The content is from Nocedal and Wright (2006) and Bertsekas (2003). Topics marked with ** are optional.

1 Optimality Conditions

In order to check if a point $x^*$ is a local minimizer (recall a point $x^*$ is a local minimizer if $f(x^*) \leq f(x)$ for all $x \in N$ for a neighborhood $N$ of $x^*$), we would have to check the function values at all $x \in N$. However, when $f$ is smooth (e.g. $C^1$ or $C^2$), then there are much more efficient ways characterizing local minimizers using $\nabla f(x^*)$ and $\nabla^2 f(x^*)$.

1.1 A Local Model of an Objective Function

For the characterization of optimality at a point $x$, we use a model of $f(x)$ that behaves similarly to $f(x)$ at the vicinity of $x$. Taylor’s theorem gives such a model built with $\nabla f(x)$ and $\nabla^2 f(x)$.

**Theorem 5.1** (Taylor’s theorem). Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is in $C^1$ and that $p \in \mathbb{R}^n$. Then there exists $t \in (0, 1)$ such that

$$f(x + p) = f(x) + \nabla f(x + tp)^T p.$$  

If $f \in C^2$, then there exists $t \in (0, 1)$ such that

$$\nabla f(x + p) = \nabla f(x) + \int_0^1 \nabla^2 f(x + tp)p \, dt,$$

and that,

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x)p.$$  

**Note.** Alternative forms of Taylor’s theorem:

$$f(x + p) = f(x) + \nabla f(x)^T p + o(\|p\|_2)$$

$$f(x + p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x)p + o(\|p\|_2^2),$$

1
**Order notation** Suppose that $\eta : \mathbb{R} \to \mathbb{R}$.

- **Little-o**: $\eta(\nu) = o(\nu)$ if $\frac{\eta(\nu)}{\nu} \to 0$ as $\nu \to 0$ or $\nu \to \infty$ (this should be clear from context).
- **Big-o**: $\eta(\nu) = O(\nu)$ if there is a constant $C > 0$ such that $|\eta(\nu)| \leq C|\nu|$ for all $\nu \in \mathbb{R}$.

From the alternative forms above, replace $x$ with $x^*$ and $x^* + p$ with $x$. Then $p = x - x^*$, and

$$
f(x) = f(x^*) + \nabla f(x^*)^T (x - x^*) + o(\|x - x^*\|_2)
$$

$$
f(x) = f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|_2^2).
$$

Clearly, we’re interested in the case $x \to x^*$, i.e. $p \to 0$.

### 1.2 First Order Necessary Conditions (FONC)

Consider a first-order approximation of $f \in \mathcal{C}^1$ near $x^*$ given by Taylor’s theorem:

$$
f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^T \Delta x
$$

That is, if $x^*$ is a local minimizer, then the first order cost due to a small variation $\Delta x$ is expected to be nonnegative:

$$
\nabla f(x^*)^T \Delta x \geq 0.
$$

Replacing $\Delta x$ by $-\Delta x$, we also get $\nabla f(x^*) \Delta x \leq 0$. The two inequalities imply that $\nabla f(x^*) \Delta x = 0$ for all $\Delta x$, which in turn implies that $\nabla f(x^*) = 0$.

**Theorem 5.2 (FONC).** If $x^*$ is a local minimizer and $f$ is continuously differentiable in an open neighborhood of $x^*$, then $\nabla f(x^*) = 0$.

**Proof.** For some $d \in \mathbb{R}^n$, $d \neq 0$, consider $g(\alpha) := f(x^* + \alpha d)$ of the scalar $\alpha$. Using the chain rule for differentiation,

$$
\frac{dg(0)}{d\alpha} = d^T \nabla f(x^*).
$$

Also, from the definition of differentiation,

$$
\frac{dg(0)}{d\alpha} = \lim_{\alpha \to 0} \frac{g(\alpha) - g(0)}{\alpha} = \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} \geq 0,
$$

since $x^*$ is local minimizer, e.g. $f(x^* + \alpha d) \geq f(x^*)$. The two results above implies that

$$
d^T \nabla f(x^*) \geq 0.
$$

Since $d$ is arbitrary, the same inequality holds with $d$ replaced by $-d$, e.g. $d^T \nabla f(x^*) \leq 0$. Therefore, $d^T \nabla f(x^*) = 0$ for all $d \in \mathbb{R}^n$, which implies that $\nabla f(x^*) = 0$. 

\[\square\]
Stationary point A point $x^*$ is called a stationary point if $\nabla f(x^*) = 0$. Any local minimizer is a stationary point.

1.3 Second Order Necessary Conditions

Consider a second-order approximation of $f$ near $x^*$ given by Taylor’s theorem:

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x^*) \Delta x.$$ 

We expect that if $x^*$ is a local minimizer, then the second order cost due to a small variation $\Delta x$ is nonnegative. Since $\nabla f(x^*)^T \Delta x = 0$ from FONC, we have

$$\Delta x^T \nabla^2 f(x^*) \Delta x \geq 0.$$ 

This implies that $\nabla^2 f(x^*)$ is positive semidefinite.

**Theorem 5.3 (SONC).** If $x^*$ is a local minimizer of $f$ and $\nabla^2 f$ exists and is continuous in an open neighborhood of $x^*$, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

**Proof.** Since $f$ is twice continuously differentiable near $x^*$, the second order Taylor series expansion yields for a scalar $\alpha$ and a vector $d = d'/\|d'\|_2 \in \mathbb{R}^n$ (for an arbitrary vector $d' \in \mathbb{R}^n$, $d' \neq 0$, so that $\|d\|_2 = 1$),

$$f(x^* + \alpha d) - f(x^*) = \alpha \nabla f(x^*)^T d + \frac{\alpha^2}{2} d^T \nabla^2 f(x^*) d + o(\alpha^2).$$

Using the condition $\nabla f(x^*) = 0$ from the FONC (Theorem 5.2), dividing both sides by $\alpha^2$, and using the fact that $x^*$ is a local minimizer, we obtain

$$0 \leq \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{o(\alpha^2)}{\alpha^2}.$$ 

Taking the limit as $\alpha \to 0$ yields $d^T \nabla^2 f(x^*) d \geq 0$, which implies that $\nabla^2 f(x^*)$ is positive semidefinite since the same inequality holds with replacing $d$ by $d' = \|d'\|_2 d$. 

1.4 Second Order Sufficient Conditions (SOSC)

**Theorem 5.4 (SOSC).** Let $\nabla^2 f$ is continuous in an open neighborhood of $x^*$ and that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then $x^*$ is a strict local minimizer of $f$.

**Proof.** Since the Hessian $\nabla^2 f$ is continuous and positive definite at $x^*$, we can choose a small enough radius $r > 0$ so that $\nabla^2 f(x)$ remains positive.
definite for all $x \in B(x^*, r) := \{z : \|z - x^*\|_2^2 < r\}$. For any nonzero vector $p$ with $\|p\|_2 < r$, we have $x^* + p \in B(x^*, r)$ and therefore,

$$f(x^* + p) - f(x^*) = \nabla f(x^*)^T p + \frac{1}{2} p^T \nabla^2 f(z)p = \frac{1}{2} p^T \nabla^2 f(z)p$$

where $z := x^* + tp$ for some $t \in (0, 1)$. Since $z \in B(x^*, r)$, we have $p^T \nabla^2 f(z)p > 0$. Therefore $f(x^* + p) > f(x^*)$ showing the claim. □

Proof. (alternative) Let $\lambda_1$ and $\lambda_n$ be the smallest and the largest eigenvalues of a positive semidefinite matrix $H \in \mathbb{R}^{n \times n}$, resp. Then $\lambda_1 \|x\|_2^2 \leq x^T H x \leq \lambda_n \|x\|_2^2$ for all $x \in \mathbb{R}^n$.

Let $\lambda$ be the smallest eigenvalue of $\nabla^2 f(x^*)$ ($\lambda > 0$ since $\nabla^2 f(x^*)$ if positive definite). Using the second order Taylor series expansion for a nonzero vector $d$ (with small enough norm $\|d\|_2$ so that $f$ is $C^2$ at $x + d$), we have

$$f(x^* + d) - f(x^*) = \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*)d + o(\|d\|_2^2)$$

$$\geq \frac{\lambda}{2} \|d\|_2^2 + o(\|d\|_2^2)$$

$$= \left(\frac{\lambda}{2} + \frac{o(\|d\|_2^2)}{\|d\|_2^2}\right) \|d\|_2^2 > 0.$$

Added 29.04.14: to be more rigorous about the final statement, consider a scalar sequence $\{v_{\|d\|_2}\}$ indexed by vector norms $\|d\|_2$ such that $v_{\|d\|_2} = o(\|d\|_2^2)/\|d\|_2^2)$. From the definition of the little-o, we have $\lim_{\|d\|_2 \to 0} v_{\|d\|_2} = 0$. That is, there exists $\delta > 0$ for which $\|d\|_2 - 0 < \delta$ and $|v_d - 0| < \epsilon$ for all $\epsilon > 0$. Since the choice of $d$ was arbitrary, we can choose $d \neq 0$ such that $0 < \|d\|_2 < \delta$ and $|v_d| = o(\|d\|_2^2)/\|d\|_2^2) < \lambda/2$. □

2 Spectral Properties of Matrices

For a square matrix $A \in \mathbb{R}^{n \times n}$, $(\lambda, v)$ is a pair of an eigenvalue $\lambda \in \mathbb{C}$ and an eigenvector $v \neq 0$, if

$$Av = \lambda v.$$

- The matrix $A \in \mathbb{R}^{n \times n}$ can have at most $n$ different eigenvalues.
- Different eigenvectors can exist for the same eigenvalue.

From the definition, we have $(A - \lambda I)v = 0$ ($I$ is an $n \times n$ identity matrix), which has a nonzero solution $v$ if and only if $\det(A - \lambda I) = 0$. The LHS of the equation defines a characteristic polynomial of $A$. 

4
Ex. \( A = \begin{bmatrix} 4 & 1 \\ 1 & -2 \end{bmatrix} \),

\[
\det(A - \lambda I) = \det \begin{bmatrix} 4 - \lambda & 1 \\ 1 & -2 - \lambda \end{bmatrix} = (4 - \lambda)(-2 - \lambda) - 1 \cdot 1 = 0
\]

\[
\Rightarrow \lambda^2 - 2\lambda - 9 = (\lambda - 1)^2 - 10 = 0 \quad \Rightarrow \lambda = 1 \pm \sqrt{10}
\]

In \( \mathbb{R} \),

\[
> A = \text{matrix(c(4,1,1,-2), nrow=2, ncol=2, byrow=TRUE)}
\]

\[
> \text{eigen}(A)
\]

$values$

\[
[1] \quad 4.162278 \quad -2.162278
\]

$vectors$

\[
[,1] \quad [,2]
\]

\[
[1,] -0.9870875 \quad 0.1601822
\]

\[
[2,] -0.1601822 \quad -0.9870875
\]

2.1 Properties of Eigenvalues/Eigenvectors

Let \( Av = \lambda v \) for a matrix \( A \in \mathbb{R}^{n\times n} \) (\( A \) doesn’t have to be symmetric), and \( \{\lambda_1, \lambda_2, \ldots, \lambda_n\} \) the set of all eigenvalues of \( A \).

- The eigenvalues of a triangular matrix are equal to its diagonal entries.

- The eigenvalues of \( A + cI \), \( c \in \mathbb{R} \), are equal to \( \lambda_1 + c, \lambda_2 + c, \ldots, \lambda_n + c \), where \( \lambda_1, \lambda_2, \ldots, \lambda_n \) are eigenvalues of \( A \).

- The eigenvalues of a square matrix \( A \) depends continuously on the elements of \( A \).

- \( A^2v = A(Av) = A(\lambda v) = \lambda(Av) = \lambda^2v \). In general, \( A^kv = \lambda^kv \).

- \( A \) is invertible iff all eigenvalues of \( A \) are nonzero. For an invertible \( A \), \( Av = \lambda v \) implies that \( A^{-1}v = (1/\lambda)v \).

- \( \text{tr}(A) := \sum_i A_{ii} = \sum_i \lambda_i \).

- \( \det(A) = \prod_i \lambda_i \).

2.2 Properties of Symmetric Matrices

Let \( A \in n \times n \) be a symmetric matrix. Then the following hold:

- The eigenvalues of \( A \) are real-valued.
• A has a set of \( n \) mutually orthogonal, real, and nonzero eigenvectors \( v_1, v_2, \ldots, v_n \).

• Suppose that the eigenvalues above are normalized so that \( \|v_i\|_2 = 1 \) for each \( i \). Then

\[
A = \sum_{i=1}^{n} \lambda_i v_i v_i^T
\]

where \( \lambda_i \) is the eigenvalue corresponding to \( v_i \). This can be rewritten storing \( v_i \)'s as columns of a matrix \( Q \) and \( \lambda_i \)'s as entires of a diagonal matrix \( \Sigma \),

\[
A = Q \Sigma Q^T \quad \text{(Eigen-decomposition of } A)\]

Note that \( Q^T Q = Q Q^T = I \) from the definitions.

Lemma 5.5. Let \( A \in \mathbb{R}^{n \times n} \) be a symmetric matrix with eigenvalues \( \lambda_1 \leq \cdots \leq \lambda_n \). Let \( v_1, \ldots, v_n \) be the associated orthonormal eigenvectors (\( \|v_i\|_2 = 1 \)). For all \( x \in \mathbb{R}^n \), \( \lambda_1 \|x\|_2^2 \leq x^T A x \leq \lambda_n \|x\|_2^2 \).

Proof. Since \( A \in \mathbb{R}^{n \times n} \) is symmetric, it has a set of \( n \) mutually orthonormal eigenvectors, which forms a basis of \( \mathbb{R}^n \). That is, any vector \( x \in \mathbb{R}^n \) can be expressed as \( x = \sum_{i=1}^{n} c_i v_i \) with each \( c_i \in \mathbb{R} \). Then

\[
x^T A x = x^T (\sum_{i=1}^{n} c_i A v_i) = x^T (\sum_{i=1}^{n} c_i \lambda_i v_i) \\
= (\sum_{i=1}^{n} c_i A v_i)^T (\sum_{i=1}^{n} c_i \lambda_i v_i) = \sum_{i=1}^{n} \lambda_i c_i^2 \|v_i\|_2^2 = \sum_{i=1}^{n} \lambda_i c_i^2
\]

And,

\[
\|x\|_2^2 = x^T x = (\sum_{i=1}^{n} c_i A v_i)^T (\sum_{i=1}^{n} c_i A v_i) = \sum_{i=1}^{n} c_i^2.
\]

These two relations prove the claim. \( \square \)

2.3 Eigenvalues of a positive semidefinite matrix

Definition 5.6. A matrix \( A \in \mathbb{R}^{n \times n} \) is positive semidefinite (p.s.d.) if \( x^T A x \geq 0 \) for all \( x \in \mathbb{R}^n \). A \( A \in \mathbb{R}^{n \times n} \) is positive definite (p.d.) if \( x^T A x > 0 \) for all \( x \neq 0 \in \mathbb{R}^n \).

• We often denote by \( A \succeq 0 \) when \( A \) is psd, and \( A \succ 0 \) when \( A \) is pd.

• When \( A \in \mathbb{R}^{n \times n} \) is symmetric and psd, all eigenvalues of \( A \) are non-negative.
Proof. Let \( \lambda \) and \( v \) be an eigenvalue and its corresponding eigenvector of \( A \). Using \( A \) is psd, we have

\[
0 \leq v^T (Av) = v^T (\lambda v) = \lambda \|v\|^2.
\]

This implies that \( \lambda \geq 0 \). \( \square \)

- When \( A \in \mathbb{R}^{n \times n} \) is symmetric and pd, all eigenvalues of \( A \) are strictly positive.

References
