Numerical Optimization

Lecture 15: Constrained Optimization: Optimality Conditions

Sangkyun Lee

The content is from Nocedal and Wright (2006) and Bertsekas (1999). Topics marked with ** are optional.

1 Examples of Constrained Optimization (Cont’)

Ex. 1

\[ \min f(x_1, x_2) = x_1 + x_2 \quad \text{s.t.} \quad c_1(x_1, x_2) = x_1^2 + x_2^2 - 2 = 0. \]

Consider \( \nabla f(x) \) and \( \nabla c_1(x) \) at \((-1, -1)^T, (1, -1)^T, (1, 1)^T, \) and \((-1, -1)^T\). The solution is obviously \( x^* = (-1, -1)^T \).

We see that with \( \lambda_1^* = -1/2 \),

\[ \nabla f(x^*) = \lambda_1^* \nabla c_1(x^*). \]

This can be derived as follows. Suppose that a feasible \( x \) is not a minimizer and therefore we find a step \( s \) which brings reduction in \( f \). We require \( s \) satisfies the constraint,

\[ 0 = c_1(x + s) \approx c_1(x) + \nabla c_1(x)^T s = \nabla c_1(x)^T s. \]

To produce a decrease in \( f \), we require that

\[ 0 > f(x + s) - f(x) \approx \nabla f(x)^T s. \]

If there is no such step \( s \), then \( x \) is likely to be a local minimizer, which can happen only if \( \nabla f(x) \) and \( \nabla c_1(x) \) are parallel.

We introduce the *Lagrangian* function,

\[ \mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x), \]

and the above condition can be stated that

\[ \nabla_x \mathcal{L}(x^*, \lambda_1^*) = 0. \]

This condition is necessary, but clearly not sufficient for \( x^* \) being a local minimizer since \( x = (1, 1)^T \) with \( \lambda_1 = 1/2 \) satisfies the condition as well, which is a maximizer. Also, the sign of \( \lambda_1^* \) can be changed, by using \( c_1(x) = 2 - x_1^2 - x_2^2 = 0 \) instead (then \( \lambda_1^* = 1/2 \)).

Ex. 2

\[ \min f(x_1, x_2) = x_1 + x_2 \quad \text{s.t.} \quad c_1(x_1, x_2) = 2 - x_1^2 - x_2^2 \geq 0. \]

Suppose that a feasible point \( x \) is not a minimizer. Then we should be able to find a step \( s \) so that

\[ 0 \leq c_1(x + s) \approx c_1(x) + \nabla c_1(x)^T s. \]

1
We can have two cases:

Case 1. \( x \) lies strictly inside of the circle, so that \( c_1(x) > 0 \). When \( \nabla f(x) \neq 0 \), we can find a step \( s \) that gives a decrease in the first order, i.e.

\[
s = -\alpha \nabla f(x)
\]

for some sufficiently small \( \alpha > 0 \).

Case 2. \( x \) lies on the boundary of the circle, so that \( c_1(x) = 0 \). Then a step for decreasing \( f \) must satisfy

\[
\nabla f(x)^T s < 0, \quad \nabla c_1(x)^T s \geq 0.
\]

The intersection of the two regions specified by these (an open half-space and a closed half-space) is empty only when \( \nabla f(x) \) and \( \nabla c_1(x) \) point to the same direction, that is, \( \nabla f(x) = \lambda_1 \nabla c_1(x) \), for some \( \lambda_1 \geq 0 \).

In this case the sign of \( \lambda_1 \) matters.
Again, this can be stated with the Lagrangian function,

\[
\nabla_x L(x^*, \lambda^*_1) = 0, \quad \text{for some} \quad \lambda_1 \geq 0.
\]

where we also require that

\[
\lambda_1^* c_1(x^*) = 0.
\]

This condition is known as a complementarity condition, which implies that the Lagrange multiplier \( \lambda_1 \) can be strictly positive only when the corresponding constraint \( c_1 \) is active. As we see later, this condition plays an important role. In this example, this condition requires that \( \lambda_1^* = 0 \) for case 1, and \( \lambda_1^* \geq 0 \) for case 2.

**Ex 3.**

\[
\min \ f(x) = x_1 + x_2 \quad \text{s.t.} \quad c_1(x) = 2 - x_1^2 - x_2^2 \geq 0, \ c_2(x) = x_2 \geq 0.
\]

The solution is \( x = (-\sqrt{2}, 0)^T \), at which both constraints are active. If \( x \) is not a minimizer, then we expect to find a step \( s \) such that

\[
\nabla c_1(x)^T s \geq 0, \forall i \in I, \quad \nabla f(x)^T s < 0. \tag{15.1}
\]

Clearly, for \( x = (-\sqrt{2}, 0)^T \) there is no such \( s \).

In terms of Lagrangian, for \( \lambda = (\lambda_1, \lambda_2)^T \) we can define

\[
\mathcal{L}(x, \lambda) = f(x) - \lambda^T \begin{bmatrix} c_1(x) \\ c_2(x) \end{bmatrix},
\]

and the optimal variables satisfy

\[
\nabla_x \mathcal{L}(x^*, \lambda^*) = 0, \quad \text{for some} \quad \lambda^* \geq 0.
\]

The complementarity condition becomes

\[
\lambda_1^* c_1(x^*) = 0, \quad \lambda_2^* c_2(x^*) = 0.
\]

When \( x^* = (-\sqrt{2}, 0)^T \), we have

\[
\nabla f(x^*) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \nabla c_1(x^*) = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix}, \quad \nabla c_2(x^*) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},
\]
so it follows that
\[ \lambda^* = \begin{bmatrix} 1/(2\sqrt{2}) \\ 1 \end{bmatrix} > 0. \]

For another point \( x = (\sqrt{2}, 0)^T \), both constraints are active. But in this case we can find \( s = (-1, 0)^T \) that satisfies (15.1). In this case the condition \( \nabla_x L(x, \lambda) = 0 \) is satisfied with \( \lambda = (-1/(2\sqrt{2}), 1)^T \), whose first component is negative.

For the point \( x = (1, 0)^T \), now only the constraint \( c_2 \) is active, and any small enough step \( s \) (satisfying \( c_2 \)) will continue to satisfy \( c_1(x + s) > 0 \). So we need to consider the behavior of only \( c_2 \) to check if \( s \) is a feasible descent step: if \( s \) is descent, then
\[ \nabla c_2(x)^T s \geq 0, \quad \nabla f(x)^T s < 0. \]

It is easy to check that \( s = (-1/2, 1/4)^T \) is such a direction. The optimality conditions in Lagrangian don’t hold: from complementarity, \( \lambda_1 = 0 \) and therefore \( 0 = \nabla_x L(x, \lambda) = \nabla f(x) - \lambda_2 \nabla c_2(x) \), but no such \( \lambda_2 \) exists.

## 2 Optimization Conditions for Constrained Minimization

Consider the constrained minimization of a continuously differentiable function \( f \),
\[ \min_{x \in \Omega} f(x), \]
where the feasible set
\[ \Omega = \{x \in \mathbb{R}^n : c_i(x) = 0, \forall i \in \mathcal{E}, c_i(x) \geq 0, \forall i \in \mathcal{E} \} \]
is nonempty and closed.

Note. A vector \( x^* \) is a global minimizer of \( f \) over \( \Omega \) if \( f(x^*) \leq f(x) \) for all \( x \in \Omega \).

### 2.1 Minimum Principle (Fundamental Optimality Condition)

**Theorem 15.1.** *(Optimality Condition)* Suppose that \( \Omega \) is a convex set.\(^1\) If \( x^* \) is a local minimizer of \( f \) over \( \Omega \), then
\[ \nabla f(x^*)^T (x - x^*) \geq 0, \quad \forall x \in \Omega. \tag{15.2} \]

Moreover, if \( f \) is convex over \( \Omega \), then the above condition is also sufficient for \( x^* \) being a global minimizer of \( f \) over \( \Omega \).

**Proof.** Suppose for contradiction that \( \nabla f(x^*)^T (x - x^*) < 0 \) for some \( x \in \Omega \). By the mean value theorem (Theorem 7.1) applied to a function \( g(\epsilon) = f(x^* + \epsilon(x - x^*)) \), there exists a scalar \( \alpha \in (0, 1) \) such that
\[ f(x^* + \epsilon(x - x^*)) - f(x^*) = \epsilon \nabla f(x^* + \alpha \epsilon(x - x^*))^T (x - x^*). \]

Since \( \nabla f \) is continuous, for all sufficiently small \( \epsilon > 0 \) we have \( \nabla f(x^* + \alpha \epsilon(x - x^*))^T (x - x^*) < 0 \) and therefore \( f(x^* + \epsilon(x - x^*)) < f(x^*) \). The vector \( x^* + \epsilon(x - x^*) \) is feasible for all \( \epsilon \in (0, 1) \) since \( \Omega \) is convex. This contradicts the fact that \( x^* \) is a local minimizer.

If \( f \) is convex, then from Lemma 9.1 we have
\[ f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*), \forall x \in \Omega. \]

If \( \nabla f(x^*)^T (x - x^*) \geq 0 \) holds for all \( x \in \Omega \), then \( x^* \) is a global minimizer of \( f \) over \( \Omega \). \( \square \)

\(^{1}\)This particular theorem requires that \( \Omega \) is convex, but the other optimality conditions will not require the convexity of \( \Omega \). Despite the restriction, the condition is very simple and intuitive.
A vector $x^*$ satisfying the optimality condition (15.2) is called a stationary point. Check that this definition is consistent with the unconstrained cases where $\Omega = \mathbb{R}^n$.

**Ex.** Consider a positive orthant constraint

$$\Omega = \{ x \in \mathbb{R}^n : x_i \geq 0, i = 1, 2, \ldots, n \}.$$  

From the necessary optimality condition, if $x^*$ is a local minimizer then

$$\sum_{i=1}^n \frac{\partial f(x^*)}{\partial x_i} (x_i - x^*) \geq 0, \quad \forall x = (x_1, x_2, \ldots, x_n)^T \geq 0.$$  

(15.3)

Let us fix in index $i$, and let $x_j = x_j^*$ for $j \neq i$ and $x_i = x_i^* + 1$. Then we obtain,

$$\frac{\partial f(x^*)}{\partial x_i} \geq 0, \quad \forall i.$$  

(15.4)

If $x_i^* > 0$, letting $x_j = x_j^*$ for $j \neq i$ and $x_i = \frac{1}{2}x_i^*$ gives

$$\frac{\partial f(x^*)}{\partial x_i} = 0, \quad \text{if } x_i^* > 0.$$  

(15.5)

If $f$ is convex, then (15.4) and (15.5) imply (15.3) and therefore these conditions are also sufficient by Theorem 15.1.

### 2.2 Tangent Cone

Given a feasible point $x \in \Omega$, $\{z_k\}$ is a feasible sequence approaching $x$ if $z_k \in \Omega$ for all sufficiently large $k$ and $z_k \to x$. A tangent is a limiting direction of a feasible sequence.

**Definition 15.2.** The vector $d$ is a tangent to $\Omega$ at a point $x$ if there exists a feasible sequence $\{z_k\}$ approaching $x$ and a sequence of positive scalars $\{t_k\}$ with $t_k \to 0$ such that

$$\lim_{k \to \infty} \frac{z_k - x}{t_k} = d.$$  

The set of all tangents to $\Omega$ at $x^*$ is called the tangent cone, denoted by $T_\Omega(x^*)$.

- $T_\Omega(x^*)$ is indeed a cone.
- $0 \in T_\Omega(x^*)$.
- $T_\Omega(x^*)$ is closed.
- $T_\Omega(x^*)$ is the closure of the set of feasible directions $\{x - x^* : x \in \Omega\}$.

**Definition 15.3.** For a cone $K \subseteq \mathbb{R}^n$, the polar cone of $K$ is defined by

$$K^\circ = \{ d \in \mathbb{R}^n : d^T x \leq 0, \forall x \in K \}.$$  

- $K^\circ$ is indeed a cone.
- (Polar Cone Theorem) For a nonempty closed convex cone $C$, $(C^\circ)^\circ = C$.

**Definition 15.4.** A normal cone to the set $\Omega$ at a point $x \in \Omega$ is defined as

$$N_\Omega(x^*) = T_\Omega^\circ(x^*).$$

---

2A set $C \subseteq \mathbb{R}^n$ is called a cone if $x \in C$, then $\alpha x \in C$ for all $\alpha > 0$.  

---

4
2.3 Tangent Cone Description of Optimality

We refine the minimum principle in Theorem 15.1 using the tangent cone.

**Theorem 15.5.** Suppose that \( x^* \) is a local minimizer. Then
\[
\nabla f(x^*)^T d \geq 0, \quad \forall d \in T_{\Omega}(x^*).
\]
or, equivalently,
\[
-\nabla f(x^*) \in N_{\Omega}(x^*).
\]

**Proof.** Suppose that \( d \in T_{\Omega}(x^*) \). Then there exists a sequences \( \{z_k\} \subset \Omega \) and \( \{t_k\} \) such that
\[
\frac{z_k - x^*}{t_k} = d + o(t_k).
\]
By the mean value theorem (Lemma 7.2), there exists \( \alpha \in (0, 1) \) such that
\[
f(z_k) - f(x^*) = \nabla f(\tilde{z}_k)^T (z_k - x^*), \quad \tilde{z}_k = x^* + \alpha(z_k - x^*).
\]
Therefore,
\[
f(z_k) - f(x^*) = \nabla f(\tilde{x})^T \tilde{d}_k, \quad \tilde{d}_k = t_k d + o(t_k).
\]
If \( d \) is a descent direction of \( f \) at \( x^* \), then \( \nabla f(x^*)^T d < 0 \), and because of the continuity of \( \nabla f \) and the facts that \( \tilde{z}_k \to x^* \) and \( \tilde{d}_k \to d \), we have for sufficiently large \( k \) that \( \nabla f(\tilde{z}_k)^T \tilde{d}_k < 0 \). This implies that \( f(z_k) < f(x^*) \) and contradicts that \( x^* \) is a local minimizer.

Equivalence of the two statement is straightforward, so try for yourself. \( \square \)

2.4 Linearized Feasible Directions (Cone of first order feasible variations)

Given a feasible point \( x \), the set of linearized feasible directions \( \mathcal{F}(x) \) is,
\[
\mathcal{F}(x) := \left\{ d : \begin{array}{l}
d^T \nabla c_i(x) = 0, \forall i \in \mathcal{E} \\
d^T \nabla c_i(x) \geq 0, \forall i \in \mathcal{A}(x) \cap \mathcal{I}
\end{array} \right\}
\]

- \( \mathcal{F}(x) \) is a cone.
- \( T_{\Omega}(x) \subseteq \mathcal{F}(x) \).

**Proof.** Given a feasible point \( x \), we reorder the constraints so that \( c_i, \ i = 1, 2, \ldots, m \) are active at \( x \). Consider sequences \( \{z_k\} \) and \( t_k = \|z_k - x\|_2 \) (in this case \( \|d\|_2 = 1 \)), so that
\[
z_k = x + t_k d + o(t_k).
\]
For \( i \in \mathcal{E} \), using Taylor’s theorem, we have
\[
0 = \frac{1}{t_k} c_i(z_k)
= \frac{1}{t_k} \left[ c_i(x) + t_k \nabla c_i(x)^T d + o(t_k\|d\|_2) \right]
= \nabla c_i(x)^T d + \frac{o(t_k)}{t_k}
\]
Taking \( k \to \infty \), we have \( \nabla c_i(x)^T d = 0 \).
For $i \in I \cup A(x)$, we have similarly that
\[
0 \leq \frac{1}{t_k} c_i(x) = \frac{1}{t_k} [c_i(z_k) + t_k \nabla c_i(x)^T d + o(t_k)]
\]
\[
= \nabla c_i(x)^T d + \frac{o(t_k)}{t_k}.
\]
And therefore $\nabla c_i(x)^T d \geq 0$ by taking $k \to \infty$. Collecting the results show
that $d \in F(x)$.

- $T_\Omega(x)$ does not depend on algebraic specification of $\Omega$ but $F(x)$ does.

**Ex 1 (revisted)**

$$
\min_x f(x) = x_1 + x_2, \quad c_1(x) = x_1^2 + x_2^2 - 2 = 0.
$$

For a non-optimal point $x = (-\sqrt{2}, 0)^T$, we find that
\[
z_k = \begin{bmatrix}
-\sqrt{2} - 1/k^2 \\
-1/k
\end{bmatrix}.
\]
and a tangent $d = (0, -1)$ with defining $t_k = \|z_k - x\|_2$. Another feasible sequence
approaching the same $x$ is
\[
z_k = \begin{bmatrix}
-\sqrt{2} - 1/k^2 \\
1/k
\end{bmatrix}.
\]
and the tangent corresponding to this sequence are $d = (0, \alpha)^T$, $\alpha > 0$. Together
with the previous one, we conclude that the tangent cone at $x = (-\sqrt{2}, 0)^T$ is
\[
T_\Omega(x) = \{(0, d_2) : d_2 \in \mathbb{R}\}.
\]

A vector $d$ in the linearized feasible directions $F(x)$ should satisfy
\[
0 = \nabla c_1(x)^T d = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}^T \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = -2\sqrt{2}d_1.
\]
Therefore, $F(x) = \{(0, d_2)^T : d_2 \in \mathbb{R}\}$ and therefore $T_\Omega(x) = F(x)$ in this case.

Suppose that the feasible set is defined instead of the formula,
\[
\Omega = \{x : c_1(x) = 0\}, \quad c_1(x) = (x_1^2 + x_2^2 - 2)^2 = 0.
\]
Here the set $\Omega$ is the same as before, but its algebraic description has changed. a
vector $d \in F(x)$ if,
\[
0 = \nabla c_1(x)^T d = \begin{bmatrix} 4(x_1^2 + x_2^2 - 2)x_1 \\ 4(x_1^2 + x_2^2 - 2)x_2 \end{bmatrix}^T \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = 0,
\]
for any feasible $x$. Hence, $F(x) = \mathbb{R}^2$, and $T_\Omega(x) \neq F(x)$. 

6
Ex 2 (revisted)

\[
\min_x f(x) = x_1 + x_2, \quad c_1(x) = 2 - x_1^2 - x_2^2 \geq 0.
\]

Check that for \( x = (-\sqrt{2}, 0)^T \), \( T_\Omega(x) = \{(w_1, w_2) : w_1 \geq 0\} \). And a vector \( d \in F(x) \) satisfies

\[
0 \leq \nabla c_1(x)^T d = 2\sqrt{2}d_1,
\]

and therefore \( T_\Omega(x) = F(x) \) in this particular algebraic specification of the feasible set.

References
