

Lecture 33, 34, 35.

Consider the problem of minimizing a function subject to nonnegativity constraints on the variables.

$$\min f(x) \text{ subject to } x \geq 0. \quad (2.2)$$

Here “ $x \geq 0$ ” indicates that all components of the vector x are required to be non-negative. Following Chapter 2, we say that x^* is a local solution of (2.2) if there is a neighborhood \mathcal{N} of x^* such that

$$f(x) \geq f(x^*) \text{ for all } x \in \mathcal{N} \text{ with } x \geq 0.$$

We can derive first-order necessary conditions for x^* to be a local solution of (2.2) by using a slight generalization of the proof technique of Theorem 2.2 in the book. We claim that these conditions are as follows:

$$0 \leq x_i^* \perp \left. \frac{\partial f}{\partial x_i} \right|_{x=x^*} \geq 0, \quad (2.3)$$

where the symbol \perp indicates that the product of these two quantities is zero. We have the following result.

THEOREM 2.1. *If x^* is a local solution of (2.2) and f is continuously differentiable in a neighborhood of x^* , then conditions (2.3) hold.*

Proof. Suppose that conditions (2.3) are violated for a certain index i . If $x_i^* < 0$ then x^* is not even feasible so is clearly not a solution, so we need only consider $x_i^* = 0$ and $x_i^* > 0$.

In the case $x_i^* = 0$, since condition (2.3) is violated, we must have $\partial f(x^*)/\partial x_i < 0$. Consider the direction p in \mathbf{R}^n with $p_j = 0$ for $j \neq i$ and $p_i = -\partial f(x^*)/\partial x_i > 0$. It is easy to see that $x^* + \bar{t}p \geq 0$ for all \bar{t} sufficiently small and positive, and that

$$p^T \nabla f(x^*) = -[\partial f(x^*)/\partial x_i]^2 < 0,$$

and hence that

$$p^T \nabla f(x^* + \bar{t}p) < 0$$

for all \bar{t} sufficiently small. It follows from Taylor’s theorem (see the argument in the proof of Theorem 2.2) that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p^T \nabla f(x^* + \bar{t}p) < f(x^*), \text{ for some } \bar{t} \in (0, \bar{t}).$$

Since any neighborhood \mathcal{N} of x^* will contain points of the form $x^* + \bar{t}p$ for small \bar{t} , and since these points are feasible, x^* does not satisfy the definition of a local minimizer.

Consider now the case of $x_i^* > 0$. Since (2.3) is violated, we must have $\partial f(x^*)/\partial x_i \neq 0$. Define $p \in \mathbf{R}^n$ as above: $p_j = 0$ for $j \neq i$ and $p_i = -\partial f(x^*)/\partial x_i$. Using a similar argument, we can verify that $x^* + \bar{t}p \geq 0$ and that $f(x^* + \bar{t}p) < f(x^*)$ for all \bar{t} sufficiently small and positive, again demonstrating that any neighborhood \mathcal{N} of x^* must contain feasible points that have a lower function value than $f(x^*)$. \square

We now discuss the more general problem of optimization over a closed convex set Ω . It is relatively easy and convenient to develop optimality conditions and the gradient projection algorithm in these general terms, then specialize it to the bound constrained case in which

$$\Omega = \{z \in \mathbf{R}^n \mid z \geq 0\}. \quad (2.4)$$

The general problem is then

$$\min f(x) \text{ subject to } x \in \Omega. \quad (2.5)$$

Given a closed convex set, the projection operator $P : \mathbb{R}^n \rightarrow \Omega$ is defined as follows:

$$P(y) = \arg \min_{z \in \Omega} \|z - y\|_2.$$

That is, $P(y)$ is the point in Ω that is closest to y in the sense of the Euclidean norm. This operator is useful both in defining optimality conditions and in defining algorithms.

We start with a useful result about P .

LEMMA 2.2.

- (i) $(P(y) - z)^T(y - z) \geq 0$ for all $z \in \Omega$, with equality if and only if $z = P(y)$.
- (ii) $(y - P(y))^T(z - P(y)) \leq 0$ for all $z \in \Omega$.

Proof. We prove (i) and leave (ii) as an exercise.

Let z be an arbitrary vector in Ω . We have

$$\begin{aligned} \|P(y) - y\|_2^2 &= \|P(y) - z + z - y\|_2^2 \\ &= \|P(y) - z\|_2^2 + 2(P(y) - z)^T(z - y) + \|z - y\|_2^2 \end{aligned}$$

which implies by rearrangement that

$$2(P(y) - z)^T(y - z) = \|P(y) - z\|_2^2 + [\|z - y\|_2^2 - \|P(y) - y\|_2^2]. \quad (2.6)$$

The term in $[\]$ is nonnegative, from the definition of P . The first term on the right-hand side is trivially nonnegative, so the nonnegativity claim is proved.

If $z = P(y)$, we obviously have $(P(y) - z)^T(y - z) = 0$. If the latter condition holds, then the first term on the right-hand side of (2.6) in particular is zero, so we have $z = P(y)$. \square

We now state the first-order necessary conditions for the problem (2.5).

THEOREM 2.3. *If f is Lipschitz continuously differentiable and x^* is a local solution of (2.5), then*

$$\nabla f(x^*)^T(x - x^*) \geq 0, \text{ for all } x \in \Omega. \quad (2.7)$$

Proof. (Sketch.) Suppose that there is some x such that $\nabla f(x^*)^T(x - x^*) < 0$. Consider a step to a point $x^* + \epsilon(x - x^*)$. We have first by convexity that, since $x \in \Omega$ and $x^* \in \Omega$, that $x^* + \epsilon(x - x^*) \in \Omega$ for all $\epsilon \in [0, 1]$. In addition, we have by Taylor's theorem that

$$f(x^* + \epsilon(x - x^*)) = f(x^*) + \epsilon \nabla f(x^*)^T(x - x^*) + O(\epsilon^2) < f(x^*),$$

for all positive ϵ sufficiently small, since $\nabla f(x^*)^T(x - x^*) < 0$ so the $O(\epsilon^2)$ term is dominated by the first order term for small ϵ . \square

We can combine the two previous results to obtain an equivalent form of the first-order necessary conditions.

THEOREM 2.4. *If*

$$P(x^* - \bar{\alpha}\nabla f(x^*)) = x^*, \text{ for some } \bar{\alpha} > 0, \quad (2.8)$$

then (2.7) holds. Conversely, if (2.7) holds, then

$$P(x^* - \alpha\nabla f(x^*)) = x^*, \text{ for all } \alpha > 0.$$

Proof. Suppose first that (2.8) holds. In Lemma 2.2(ii) we set

$$y = x^* - \bar{\alpha}\nabla f(x^*), \quad P(y) = x^*,$$

and let z be any element of Ω . We then have

$$0 \geq (y - P(y))^T(z - P(y)) = (-\bar{\alpha}\nabla f(x^*))^T(z - x^*),$$

which implies that $\nabla f(x^*)^T(z - x^*) \geq 0$ for all $z \in \Omega$, proving (2.7).

Now supposed that (2.7) holds, and denote

$$x_\alpha = P(x^* - \alpha\nabla f(x^*)).$$

Setting $y = x^* - \alpha\nabla f(x^*)$, $P(y) = x_\alpha$, $z = x^*$ in Lemma 2.2, we have

$$(x^* - \alpha\nabla f(x^*) - x_\alpha)^T(x^* - x_\alpha) \leq 0,$$

which implies that

$$\|x^* - x_\alpha\|_2^2 - \alpha\nabla f(x^*)^T(x^* - x_\alpha) \leq 0. \quad (2.9)$$

By (2.7), we have that

$$-\alpha\nabla f(x^*)^T(x^* - x_\alpha) \geq 0,$$

so both terms on the left-hand side of (2.9) are nonnegative. Hence they are both zero and we have in particular that $x_\alpha = x^*$ as claimed. \square

An immediate consequence of this theorem is that $P(x^* - \bar{\alpha}\nabla f(x^*)) = x^*$ for some $\bar{\alpha} > 0$, then $P(x^* - \alpha\nabla f(x^*)) = x^*$ for all $\alpha > 0$. (Why?)

It is not difficult to show that the general first-order conditions (2.7) reduce to (2.3) in the case of Ω defined by (2.4). We prove this result informally. Suppose first that (2.3) hold. Then

$$\nabla f(x^*)^T(x - x^*) = \sum_{i:(\partial f/\partial x_i) > 0} \frac{\partial f}{\partial x_i}(x_i - x_i^*) + \sum_{i:(\partial f/\partial x_i) = 0} \frac{\partial f}{\partial x_i}(x_i - x_i^*).$$

The second sum is trivially zero. In the first sum, we have $(\partial f/\partial x_i) > 0$ and thus $x_i^* = 0$ by (2.3). Thus $x_i - x_i^* = x_i \geq 0$, so that $\frac{\partial f}{\partial x_i}(x_i - x_i^*) = \frac{\partial f}{\partial x_i}x_i \geq 0$. Hence $\nabla f(x^*)^T(x - x^*) \geq 0$, as required. Now suppose that $\nabla f(x^*)^T(x - x^*) \geq 0$ for all feasible x . Choosing i such that $x_i^* = 0$, we can define x such that $x_j = x_j^*$ for all $j \neq i$, and $x_i = 1$. Thus

$$0 \leq \nabla f(x^*)^T(x - x^*) = \frac{\partial f}{\partial x_i}(x_i - x_i^*) = \frac{\partial f}{\partial x_i},$$

implying that $\frac{\partial f}{\partial x_i} \geq 0$. Hence (2.3) is satisfied for this i . For the other case $x_i^* > 0$, we again choose x with $x_j = x_j^*$ for all $j \neq i$ and obtain

$$0 \leq \nabla f(x^*)^T(x - x^*) = \frac{\partial f}{\partial x_i}(x_i - x_i^*).$$

By choosing $x_i = 0$ and $x_i = 2x_i^*$, we obtain

$$0 \leq \frac{\partial f}{\partial x_i}(-x_i^*), \quad 0 \leq \frac{\partial f}{\partial x_i}x_i^*,$$

from which it follows, using $x_i^* > 0$, that $\frac{\partial f}{\partial x_i} = 0$. Thus (2.3) is satisfied in this case as well.

A basic algorithm is *gradient projection*. In a sense it is the natural extension of steepest descent to the problem (2.5). The search path from a point x is the projection of the steepest descent path onto the feasible set Ω , that is,

$$P(x - \alpha \nabla f(x)), \quad \alpha > 0.$$

We seek an α for which the function decreases, that is,

$$f(x) > f(P(x - \alpha \nabla f(x))).$$

For global convergence we need a stronger condition than this. Ideally, we would like to find an α that minimizes the line search function $\phi(\alpha) \equiv f(P(x - \alpha \nabla f(x)))$. But this is difficult, even for simple functions f . In general it is not a smooth function of α . For example, when Ω is polyhedral and f is quadratic, ϕ is piecewise quadratic. When Ω is polyhedral and f is smooth, ϕ is piecewise smooth. But because of the “kinks” in the search path, we cannot apply line search procedures like the one developed in Chapter 3, which assume smoothness. Backtracking may be more suitable.

To describe this strategy we use the following notation, for convenience:

$$x(\alpha) \stackrel{\text{def}}{=} P(x - \alpha \nabla f(x)).$$

We consider an Armijo backtracking strategy along the projection arc, in which we choose an $\bar{\alpha} > 0$ and take the step α_k to be the first element in the sequence $\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \dots$ for which the following condition is satisfied:

$$f(x_k(\beta^m \bar{\alpha})) \leq f(x_k) + c_1 \nabla f(x_k)^T(x_k(\beta^m \bar{\alpha}) - x_k). \quad (2.10)$$

We can show that the resulting algorithm has stationary limit points. The analysis is quite complicated (see pp. 236–240 of Bertsekas), but we state it in part because the proof techniques are interesting and relevant to convergence results in other settings.

We need a preliminary “geometric” lemma whose proof is omitted—see Lemma 2.3.1 of Bertsekas.

LEMMA 2.5. *For all $x \in \Omega$ and $z \in \mathbb{R}^n$, the function $g : [0, \infty) \rightarrow \mathbb{R}$ defined by*

$$g(s) \stackrel{\text{def}}{=} \frac{\|P(x + sz) - x\|}{s}$$

is monotonically nonincreasing.

Our convergence result is then:

THEOREM 2.6.

(a) For every $x \in \Omega$ there exists a scalar $s_x > 0$ such that

$$f(x) - f(x(s)) \geq c_1 \nabla f(x)^T (x - x(s)), \quad \forall s \in [0, s_x]. \quad (2.11)$$

(b) Let x_k be the sequence generated by the gradient projection method with Armijo backtracking along the projection path, $\bar{\alpha} = 1$, and step acceptance criterion (2.10). Then every limit point of $\{x_k\}$ is stationary

Proof. We follow the proof of Bertsekas Proposition 2.3.3. For Part (a), we have by Lemma 2.2(ii) with $y = x - s\nabla f(x)$ (thus $P(y) = x(s)$) that

$$(x - x(s))^T (x - s\nabla f(x) - x(s)) \leq 0, \quad \forall x \in \Omega, s > 0.$$

By rearrangement this becomes

$$\nabla f(x)^T (x - x(s)) \geq \frac{\|x - x(s)\|^2}{s}, \quad \forall x \in \Omega, s > 0. \quad (2.12)$$

If x is stationary (that is, satisfies conditions and (2.8)) we have $x(s) = x$ for all $s > 0$, so the conclusion (2.11) holds trivially. Otherwise, x is nonstationary, so $x \neq x(s)$ for all $s > 0$. By Taylor's theorem we have

$$f(x) - f(x(s)) = \nabla f(x)^T (x - x(s)) + (\nabla f(\zeta_s) - \nabla f(x))^T (x - x(s)),$$

for some ζ_s on the line segment between x and $x(s)$. Hence, (2.11) can be written as

$$(1 - c_1) \nabla f(x)^T (x - x(s)) \geq (\nabla f(x) - \nabla f(\zeta_s))^T (x - x(s)). \quad (2.13)$$

From (2.12) and Lemma 2.5, we have for all $s \in (0, 1]$ that

$$\nabla f(x)^T (x - x(s)) \geq \frac{\|x - x(s)\|^2}{s} \geq \|x - x(1)\| \|x - x(s)\|.$$

Therefore (2.13) is satisfied for all $s \in (0, 1]$ such that

$$(1 - c_1) \|x - x(1)\| \geq (\nabla f(x) - \nabla f(\zeta_s))^T \frac{x - x(s)}{\|x - x(s)\|}.$$

Obviously the left-hand side of this expression is strictly positive, while the right-hand side goes to zero continuously, as $s \downarrow 0$. Hence there is an s_x with the desired property, so the proof of part (a) is complete.

Note that part (a) ensures that we can find an α_k of the form $\alpha_k = \beta^{m_k}$ that satisfies (2.10) from each point x_k . Suppose there is a subsequence K such that $x_k \rightarrow_{k \in K} \bar{x}$. Since the sequence of function values $\{f(x_k)\}$ is nonincreasing we have $f(x_k) \rightarrow f(\bar{x})$.

We consider two cases. First, suppose that

$$\liminf_{k \in K} \alpha_k \geq \hat{\alpha}$$

for some $\hat{\alpha} > 0$. From (2.12) and Lemma 2.5, we have for all $k \in K$ sufficiently large that

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq c_1 \nabla f(x_k)^T (x_k - x_{k+1}) \\ &\geq c_1 \frac{\|x_k - x_{k+1}\|^2}{\alpha_k} \\ &= \frac{c_1 \alpha_k \|x_k - x_{k+1}\|^2}{\alpha_k^2} \\ &\geq c_1 \hat{\alpha} \|x_k - x_{k+1}\|^2, \end{aligned}$$

since 1 is the initial choice of steplength and $\alpha_k \leq 1$. Taking the limit as $k \rightarrow \infty$, $k \in K$, we have $\bar{x} - \bar{x}(1) = 0$, which implies that \bar{x} is stationary (see (2.8)).

In the second case, suppose that $\liminf_{k \in K, k \rightarrow \infty} \alpha_k = 0$. Then by taking a further subsequence $\bar{K} \subset K$, we have $\lim_{k \in \bar{K}, k \rightarrow \infty} \alpha_k = 0$. For all $k \in \bar{K}$, the Arnijo test (2.10) will fail at least once, so we have

$$f(x_k) - f(x_k(\beta^{-1}\alpha_k)) < c_1 \nabla f(x_k)^T (x_k - x_k(\beta^{-1}\alpha_k)). \quad (2.14)$$

Further, no x_k with $k \in \bar{K}$ can be stationary, since for stationary points we have $\alpha_k = 1$. Hence,

$$\|x_k - x_k(\beta^{-1}\alpha_k)\| > 0. \quad (2.15)$$

By Taylor's theorem we have

$$\begin{aligned} f(x_k) - f(x_k(\beta^{-1}\alpha_k)) &= \nabla f(x_k)^T (x_k - x_k(\beta^{-1}\alpha_k)) \\ &\quad + (\nabla f(\zeta_k) - \nabla f(x_k))^T (x_k - x_k(\beta^{-1}\alpha_k)), \end{aligned}$$

for some ζ_k on the line segment between x_k and $x_k(\beta^{-1}\alpha_k)$. By combining this expression with (2.14), we have

$$(1 - c_1) \nabla f(x_k)^T (x_k - x_k(\beta^{-1}\alpha_k)) < (\nabla f(x_k) - \nabla f(\zeta_k))^T (x_k - x_k(\beta^{-1}\alpha_k)),$$

From (2.12) with Lemma 2.5, we have

$$\begin{aligned} \nabla f(x_k)^T (x_k - x_k(\beta^{-1}\alpha_k)) &\geq \frac{\|x_k - x_k(\beta^{-1}\alpha_k)\|^2}{\beta^{-1}\alpha_k} \\ &\geq \|x_k - x_k(1)\| \|x_k - x_k(\beta^{-1}\alpha_k)\|. \end{aligned}$$

By combining the last two results, and using the Schwartz inequality, we have for large $k \in \bar{K}$ that

$$\begin{aligned} (1 - c_1) \|x_k - x_k(1)\| \|x_k - x_k(\beta^{-1}\alpha_k)\| &< (\nabla f(x_k) - \nabla f(\zeta_k))^T (x_k - x_k(\beta^{-1}\alpha_k)) \\ &\leq \|\nabla f(x_k) - \nabla f(\zeta_k)\| \|x_k - x_k(\beta^{-1}\alpha_k)\|. \end{aligned}$$

Using this expression together with (2.15) we obtain

$$(1 - c_1) \|x_k - x_k(1)\| < \|\nabla f(x_k) - \nabla f(\zeta_k)\|.$$

Since $\alpha_k \rightarrow 0$ and $x_k \rightarrow \bar{x}$ as $k \rightarrow \infty$, $k \in \bar{K}$, it follows that $\zeta_k \rightarrow \bar{x}$. Hence by taking limits in the expression above we obtain that $\bar{x} = \bar{x}(1)$, which implies that \bar{x} is stationary, as claimed. \square