2. $P(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$;
3. $P(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x}$ is feasible, that is, $g_1(\boldsymbol{x}) \leq 0, \ldots, g_p(\boldsymbol{x}) \leq 0$. ∎

Clearly, for the above unconstrained problem to be a good approximation to the original problem, the penalty function P must be appropriately chosen. The role of the penalty function is to "penalize" points that are outside the feasible set. Therefore, it is natural that the penalty function be defined in terms of the constraint functions $g_1, \ldots, g_p$. A possible choice for P is

$$P(\boldsymbol{x}) = \sum_{i=1}^{p} g_i^+(\boldsymbol{x}),$$

where

$$g_i^+(\boldsymbol{x}) = \max(0, g_i(\boldsymbol{x})) = \begin{cases} 0 & \text{if } g_i(\boldsymbol{x}) \leq 0 \\ g_i(\boldsymbol{x}) & \text{if } g_i(\boldsymbol{x}) > 0. \end{cases}$$

We refer to the above penalty function as the *absolute value penalty function*, because it is equal to $\sum |g_i(\boldsymbol{x})|$, where the summation is taken over all constraints that are violated at $\boldsymbol{x}$. We illustrate this penalty function in the following example.

Example 22.1 Let $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ be defined by $g_1(x) = x-2$, $g_2(x) = -(x+1)^3$. The feasible set defined by $\{x \in \mathbb{R} : g_1(x) \leq 0, g_2(x) \leq 0\}$ is simply the interval $[-1, 2]$. In this example, we have

$$\begin{aligned} g_1^+(x) &= \max(0, g_1(x)) = \begin{cases} 0 & \text{if } x \leq 2 \\ x-2 & \text{otherwise} \end{cases} \\ g_2^+(x) &= \max(0, g_2(x)) = \begin{cases} 0 & \text{if } x \geq -1 \\ -(x+1)^3 & \text{otherwise,} \end{cases} \end{aligned}$$

and

$$P(x) = g_1^+(x) + g_2^+(x) = \begin{cases} x-2 & \text{if } x > 2 \\ 0 & \text{if } -1 \leq x \leq 2 \\ -(x+1)^3 & \text{if } x < -1. \end{cases}$$

Figure 22.1 provides a graphical illustration of g^+ for this example. ∎

The absolute value penalty function may not be differentiable at points $\boldsymbol{x}$ where $g_i(\boldsymbol{x}) = 0$, as is the case at the point $x = 2$ in Example 22.1 (notice, though, that in Example 22.1, P is differentiable at $x = -1$). Therefore, in such cases we cannot use techniques for optimization that involve derivatives. A form of the penalty function that is guaranteed to be differentiable is the so-called *Courant-Beltrami penalty function*, given by

$$P(\boldsymbol{x}) = \sum_{i=1}^{p} \left(g_i^+(\boldsymbol{x})\right)^2.$$

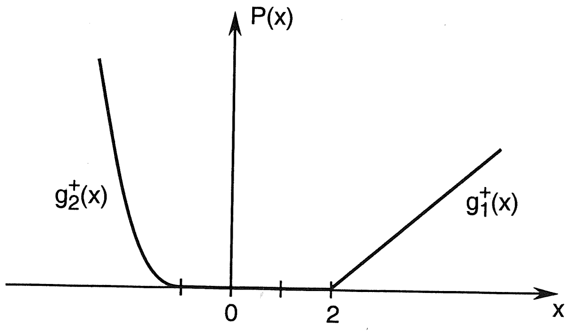

Figure 22.1 g^+ for Example 22.1

In the following discussion, we do not assume any particular form of the penalty function P. We only assume that P satisfies conditions 1–3 given in Definition 22.1.

The penalty function method for solving constrained optimization problems involves constructing and solving an associated unconstrained optimization problem, and using the solution to the unconstrained problem as the solution to the original constrained problem. Of course, the solution to the unconstrained problem (the approximated solution) may not be exactly equal to the solution to the constrained problem (the true solution). Whether or not the solution to the unconstrained problem is a good approximation to the true solution depends on the penalty parameter γ and the penalty function P. We would expect that the larger the value of the penalty parameter γ, the closer the approximated solution will be to the true solution, because points that violate the constraints are penalized more heavily. Ideally, in the limit as $\gamma \to \infty$, the penalty method should yield the true solution to the constrained problem. In the remainder of this section, we analyze this property of the penalty function method.

In our analysis of the penalty method, we adopt the following setting. Recall that the original constrained optimization problem is:

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & g_1(\boldsymbol{x}) \leq 0 \\ & g_2(\boldsymbol{x}) \leq 0 \\ & \vdots \\ & g_p(\boldsymbol{x}) \leq 0. \end{array}$$

Denote by $\boldsymbol{x}^*$ a solution (global minimizer) to the above problem. Let P be a penalty function for the problem. For each $k = 1, 2, \ldots$, let $\gamma_k \in \mathbb{R}$ be a given positive constant. Define an associated function $q(\gamma_k, \cdot) : \mathbb{R}^n \to \mathbb{R}$ by

$$q(\gamma_k, \boldsymbol{x}) = f(\boldsymbol{x}) + \gamma_k P(\boldsymbol{x}).$$

For each k, we can write the following associated unconstrained optimization problem:

$$\text{minimize} \quad q(\gamma_k, \boldsymbol{x}).$$

Denote by $\boldsymbol{x}^{(k)}$ a minimizer of $q(\gamma_k, \boldsymbol{x})$. The following technical lemma describes certain useful relationships between the constrained problem and the associated unconstrained problems.

Lemma 22.2 *Suppose $\{\gamma_k\}$ is a nondecreasing sequence, that is, for each k, we have $\gamma_k \le \gamma_{k+1}$. Then, for each k we have*

1. $q(\gamma_{k+1}, \boldsymbol{x}^{(k+1)}) \ge q(\gamma_k, \boldsymbol{x}^{(k)})$
2. $P(\boldsymbol{x}^{(k+1)}) \le P(\boldsymbol{x}^{(k)})$
3. $f(\boldsymbol{x}^{(k+1)}) \ge f(\boldsymbol{x}^{(k)})$
4. $f(\boldsymbol{x}^*) \ge q(\gamma_k, \boldsymbol{x}^{(k)}) \ge f(\boldsymbol{x}^{(k)})$.

□

Proof. We first prove part 1. From the definition of q and the fact that $\{\gamma_k\}$ is an increasing sequence, we have

$$q(\gamma_{k+1}, \boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k+1)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k+1)}) \ge f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}).$$

Now, because $\boldsymbol{x}^{(k)}$ is a minimizer of $q(\gamma_k, \boldsymbol{x})$,

$$q(\gamma_k, \boldsymbol{x}^{(k)}) = f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}) \le f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}).$$

Combining the above, we get part 1.

We next prove part 2. Because $\boldsymbol{x}^{(k)}$ and $\boldsymbol{x}^{(k+1)}$ minimize $q(\gamma_k, \boldsymbol{x})$ and $q(\gamma_{k+1}, \boldsymbol{x})$, respectively, we can write

$$\begin{aligned} q(\gamma_k, \boldsymbol{x}^{(k)}) &= f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}) \le f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}), \\ q(\gamma_{k+1}, \boldsymbol{x}^{(k+1)}) &= f(\boldsymbol{x}^{(k+1)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k+1)}) \le f(\boldsymbol{x}^{(k)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k)}). \end{aligned}$$

Adding the above inequalities yields

$$\gamma_k P(\boldsymbol{x}^{(k)}) + \gamma_{k+1} P(\boldsymbol{x}^{(k+1)}) \le \gamma_{k+1} P(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}).$$

Rearranging, we get

$$(\gamma_{k+1} - \gamma_k) P(\boldsymbol{x}^{(k+1)}) \le (\gamma_{k+1} - \gamma_k) P(\boldsymbol{x}^{(k)}).$$

We know by assumption that $\gamma_{k+1} \ge \gamma_k$. If $\gamma_{k+1} > \gamma_k$, then we get $P(\boldsymbol{x}^{(k+1)}) \le P(\boldsymbol{x}^{(k)})$. If, on the other hand, $\gamma_{k+1} = \gamma_k$, then clearly $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$ and so $P(\boldsymbol{x}^{(k+1)}) = P(\boldsymbol{x}^{(k)})$. Therefore, in either case, we arrive at part 2.

We now prove part 3. Because $\boldsymbol{x}^{(k)}$ is a minimizer of $q(\gamma_k, \boldsymbol{x})$, we obtain

$$q(\gamma_k, \boldsymbol{x}^{(k)}) = f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}) \le f(\boldsymbol{x}^{(k+1)}) + \gamma_k P(\boldsymbol{x}^{(k+1)}).$$

Therefore,

$$f(\boldsymbol{x}^{(k+1)}) \ge f(\boldsymbol{x}^{(k)}) + \gamma_k (P(\boldsymbol{x}^{(k)}) - P(\boldsymbol{x}^{(k+1)})).$$

From part 2, we have $P(\boldsymbol{x}^{(k)}) - P(\boldsymbol{x}^{(k+1)}) \ge 0$, and $\gamma_k > 0$ by assumption; therefore, we get

$$f(\boldsymbol{x}^{(k+1)}) \ge f(\boldsymbol{x}^{(k)}).$$

Finally, we now prove part 4. Because $\boldsymbol{x}^{(k)}$ is a minimizer of $q(\gamma_k, \boldsymbol{x})$, we get

$$f(\boldsymbol{x}^*) + \gamma_k P(\boldsymbol{x}^*) \ge q(\gamma_k, \boldsymbol{x}^{(k)}) = f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}).$$

Because $\boldsymbol{x}^*$ is a minimizer for the constrained optimization problem, we have $P(\boldsymbol{x}^*) = 0$. Therefore,

$$f(\boldsymbol{x}^*) \ge f(\boldsymbol{x}^{(k)}) + \gamma_k P(\boldsymbol{x}^{(k)}).$$

Because $P(\boldsymbol{x}^{(k)}) \ge 0$ and $\gamma_k \ge 0$,

$$f(\boldsymbol{x}^*) \ge q(\gamma_k, \boldsymbol{x}^{(k)}) \ge f(\boldsymbol{x}^{(k)}),$$

which completes the proof. ∎

With the above lemma, we are now ready to prove the following theorem.

Theorem 22.2 *Suppose the objective function f is continuous, and $\gamma_k \to \infty$ as $k \to \infty$. Then, the limit of any convergent subsequence of the sequence $\{\boldsymbol{x}^{(k)}\}$ is a solution to the constrained optimization problem.* □

Proof. Suppose $\{\boldsymbol{x}^{(m_k)}\}$ is a convergent subsequence of the sequence $\{\boldsymbol{x}^{(k)}\}$. (See Section 5.1 for a discussion of sequences and subsequences.) Let $\hat{\boldsymbol{x}}$ be the limit of $\{\boldsymbol{x}^{(m_k)}\}$. By Lemma 22.2, the sequence $\{q(\gamma_k, \boldsymbol{x}^{(k)})\}$ is nondecreasing and bounded above by $f(\boldsymbol{x}^*)$. Therefore, the sequence $\{q(\gamma_k, \boldsymbol{x}^{(k)})\}$ has a limit $q^* = \lim_{k\to\infty} q(\gamma_k, \boldsymbol{x}^{(k)})$ such that $q^* \le f(\boldsymbol{x}^*)$ (see Theorem 5.3). Because the function f is continuous, and $f(\boldsymbol{x}^{(m_k)}) \le f(\boldsymbol{x}^*)$ by Lemma 22.2, we have

$$\lim_{k\to\infty} f\left(\boldsymbol{x}^{(m_k)}\right) = f\left(\lim_{k\to\infty} \boldsymbol{x}^{(m_k)}\right) = f(\hat{\boldsymbol{x}}) \le f(\boldsymbol{x}^*).$$

Because the sequences $\{f(\boldsymbol{x}^{(m_k)})\}$ and $\{q(\gamma_{m_k}, \boldsymbol{x}^{(m_k)})\}$ both converge, the sequence $\{\gamma_{m_k} P(\boldsymbol{x}^{(m_k)})\} = \{q(\gamma_{m_k}, \boldsymbol{x}^{(m_k)}) - f(\boldsymbol{x}^{(m_k)})\}$ also converges, with

$$\lim_{k\to\infty} \gamma_{m_k} P(\boldsymbol{x}^{(m_k)}) = q^* - f(\hat{\boldsymbol{x}}).$$

By Lemma 22.2, the sequence $\{P(\boldsymbol{x}^{(k)})\}$ is nonincreasing and bounded from below by 0. Therefore, $\{P(\boldsymbol{x}^{(k)})\}$ converges (again see Theorem 5.3), and hence so does $\{P(\boldsymbol{x}^{(m_k)})\}$. Because $\gamma_{m_k} \to \infty$, we conclude that

$$\lim_{k\to\infty} P(\boldsymbol{x}^{(m_k)}) = 0.$$

By continuity of P, we have

$$0 = \lim_{k\to\infty} P(\boldsymbol{x}^{(m_k)}) = P\left(\lim_{k\to\infty} \boldsymbol{x}^{(m_k)}\right) = P(\hat{\boldsymbol{x}}),$$

and hence $\hat{\boldsymbol{x}}$ is a feasible point. Because $f(\boldsymbol{x}^*) \geq f(\hat{\boldsymbol{x}})$ from above, we conclude that $\hat{\boldsymbol{x}}$ must be a solution to the constrained optimization problem. ■

If we perform an infinite number of minimization runs, with the penalty parameter $\gamma_k \to \infty$, then the above theorem ensures that the limit of any convergent subsequence is a minimizer $\boldsymbol{x}^*$ to the original constrained optimization problem. There is clearly a practical limitation in applying this theorem. It is certainly desirable to find a minimizer to the original constrained optimization problem using a *single* minimization run for the unconstrained problem that approximates the original problem using a penalty function. In other words, we desire an exact solution to the original constrained problem by solving the associated unconstrained problem (minimize $f(\boldsymbol{x}) + \gamma P(\boldsymbol{x})$) with a finite $\gamma > 0$. It turns out that indeed this can be accomplished, in which case we say that the penalty function is *exact*. However, it is necessary that exact penalty functions be nondifferentiable, as shown in [7], and illustrated in the following example.

Example 22.2 Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in [0,1], \end{array}$$

where $f(x) = 5 - 3x$. Clearly, the solution is $x^* = 1$.

Suppose we use the penalty method to solve the problem, with a penalty function P that is differentiable at $x^* = 1$. Then, $P'(x^*) = 0$, because $P(x) = 0$ for all $x \in [0,1]$. Hence, if we let $g = f + \gamma P$, then $g'(x^*) = f'(x^*) + \gamma P'(x^*) \neq 0$ for all finite $\gamma > 0$. Hence, $x^* = 1$ does not satisfy the first-order necessary condition to be a local minimizer of g. Thus, P is not an exact penalty function. ■

Here, we prove a result on the necessity of nondifferentiability of exact penalty functions for a special class of problems.

Proposition 22.4 *Consider the problem*

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \Omega, \end{array}$$

with $\Omega \subset \mathbb{R}^n$ convex. Suppose the minimizer $\boldsymbol{x}^*$ lies on the boundary of Ω, and there exists a feasible direction $\boldsymbol{d}$ at $\boldsymbol{x}^*$ such that $\boldsymbol{d}^T\nabla f(\boldsymbol{x}^*) > 0$. If P is an exact penalty function, then P is not differentiable at $\boldsymbol{x}^*$. □

Proof. We use contraposition. Suppose P is differentiable at $\boldsymbol{x}^*$. Then, $\boldsymbol{d}^T\nabla P(\boldsymbol{x}^*) = 0$, because $P(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega$. Hence, if we let $g = f + \gamma P$, then $\boldsymbol{d}^T\nabla g(\boldsymbol{x}^*) > 0$ for all finite $\gamma > 0$, which implies that $\nabla g(\boldsymbol{x}^*) \neq \boldsymbol{0}$. Hence, $\boldsymbol{x}^*$ is not a local minimizer of g, and thus P is not an exact penalty function. ■

Note that the result of the above proposition does not hold if we remove the assumption that $\boldsymbol{d}^T\nabla f(\boldsymbol{x}^*) > 0$. Indeed, consider a convex problem where $\nabla f(\boldsymbol{x}^*) = 0$. Choose P to be differentiable. Clearly, in this case we have $\nabla g(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*) + \gamma\nabla P(\boldsymbol{x}^*) = 0$. The function P is therefore an exact penalty function, although differentiable.

For further reading on the subject of optimization of nondifferentiable functions, see, for example, [25]. The references [8] and [70] provide further discussions on the penalty method, including nondifferentiable exact penalty functions. These references also discuss exact penalty methods involving differentiable functions; these methods go beyond the elementary type of penalty method introduced in this chapter.

EXERCISES

22.1 Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m < n$, rank $\boldsymbol{A} = m$, and $\boldsymbol{b} \in \mathbb{R}^m$. Define $\Omega = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$ and let $\boldsymbol{x}_0 \in \Omega$. Show that for any $\boldsymbol{y} \in \mathbb{R}^n$,

$$\boldsymbol{\Pi}[\boldsymbol{x}_0 + \boldsymbol{y}] = \boldsymbol{x}_0 + \boldsymbol{P}\boldsymbol{y},$$

where $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A}$.

Hint: Use Exercise 6.4 and Example 12.4.

22.2 Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{c}$, where $\boldsymbol{Q} = \boldsymbol{Q}^T > 0$. We wish to minimize f over $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m < n$, and rank $\boldsymbol{A} = m$. Show that the projected steepest descent algorithm for this case takes the form

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \left(\frac{\boldsymbol{g}^{(k)T}\boldsymbol{P}\boldsymbol{g}^{(k)}}{\boldsymbol{g}^{(k)T}\boldsymbol{P}\boldsymbol{Q}\boldsymbol{P}\boldsymbol{g}^{(k)}}\right)\boldsymbol{P}\boldsymbol{g}^{(k)},$$

where

$$\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{Q}\boldsymbol{x}^{(k)} - \boldsymbol{c},$$

and $\boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}\boldsymbol{A}$.

22.3 Consider the problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\boldsymbol{x}\|^2 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \end{array}$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m < n$, and rank $\boldsymbol{A} = m$. Show that if $\boldsymbol{x}^{(0)} \in \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}$, then the projected steepest descent algorithm converges to the solution in one step.