An optimal solution is an admissible solution of minimum cost. That is easy to say, but it contains a lot. It means, first, that there is a ***cost function*** $C(x)$ to be minimized. It also implies some ***admissibility constraints*** on x: the flow out must equal the flow in, or capacities cannot be exceeded, or the structure must withstand the load. If x represents a set of prices, or a set of shipments, or a set of probabilities, they cannot be negative. The constraints probably prevent the cost C from reaching its absolute minimum, and the problem is one of ***constrained optimization***.

Without constraints, the minimization of $C(x)$ is an ordinary problem in calculus. When x is a vector we need multivariable calculus, and when it is a function we need the calculus of variations, but these just give extensions of the same basic idea: The derivative C' should be zero at the minimum. That is the grandfather of all optimality conditions! The constraints seem to destroy this simple rule, and the chief goal of the theory is to save as much of it as possible.

Thanks to Lagrange, it is almost entirely saved. We look first at the case of linear constraints, then at nonlinear constraints, and finally at inequality constraints. In each case there is a test on the derivative C' at the optimal point x^*, but it applies *only in the directions permitted by the constraints*. In the other directions we cannot move from x^* without leaving the admissible set. This is clear from the geometry, and we will rely more than usual on insight and examples. At the end, however, it is mathematics that produces the critical quantities—the Lagrange multipliers y. They convert the constrained minimization of $C(x)$ into an unconstrained minimization of $L(x)=C(x)+y^T(A(x)-b)$. A chief object of this section is to find and understand these magic numbers y.

Before we start, it is worth thinking for a moment about inequality constraints. Suppose there are m of them, linear or nonlinear, and they are written as $A_i(x)\le b_i$. At the optimal x^* they fall into two groups—those for which there is equality $A_i(x^*)=b_i$ and a "tight" constraint, and those for which there is strict inequality $A_i(x^*)<b_i$. The first constraints are ***active*** at x^*, the others are ***inactive***. As far as inactive constraints are concerned, the test for a minimum hardly notices them. They will be satisfied by any x near x^*, and their Lagrange multipliers y_i will be zero. It is the active constraints that need Lagrange multipliers. When $A_i(x^*)-b_i$ is zero, y_i is almost certainly nonzero; it measures the force of the constraint, preventing $A_i(x)$ from exceeding b_i as it would like to.

One or the other of these quantities, either y_i or $A_i(x^*)-b_i$, is always zero. In the case of equality constraints, $A_i(x^*)-b_i$ is zero because that is the constraint. In every case the product $y_i(A_i(x^*)-b_i)$ is zero, and the sum of all m products—which is the inner product of y with $A(x^*)-b$—is also zero:

$$\boxed{y(A(x^*)-b)=0.} \tag{1}$$

This is the ***complementarity condition***, a nonlinear copy of the previous section's

$y(Ax-b)=0$. The zeros in y (we should really write y^*) are in complementary positions to the zeros in $A(x^*)-b$. When b_i is large, and the ith material is oversupplied, the constraint $A_i(x) \le b_i$ is not really restrictive. The material becomes a free good, and its price y_i drops to zero. At an equilibrium point x^*, each Lagrange multiplier y_i tells the real price of its constraint—by revealing how much a small change in b_i would affect the minimum cost.†

You might say that (1) must automatically hold, if the constrained minimum of C and the unconstrained minimum of $L=C+y(Ax-b)$ are the same. That is true! The y_i have the remarkable property that the minima of C and L are attained *at the same point* x^*. Of course the y_i are not known—they might be regarded as the fundamental unknowns, if the final unconstrained problem is easy—and it cannot be predicted in advance which of them are nonzero. We do not know which constraints are active, until the whole problem is solved. Just as in the simplex method, the correct y_i emerge at the same time as the correct x_j.

Numerically, the simplex method is a very limited model for solving nonlinear programs. For ***quadratic programming***—when C is a quadratic and the constraints are linear equations and inequalities—you can see what will happen. Moving along an edge of the feasible set, the cost looks like a parabola. It is decreasing at the start of the edge or we would not move. If it is still decreasing at the end, we stop there. But unlike the linear case, a parabola may start down and later go up; in that case we stop at its minimum, and take the next step from there. This requires only small changes in the simplex method, and quadratic programming is not excessively hard—but the minimum may occur *inside* the feasible set. The solution is not always at a corner.

For a general nonlinear program there are many possibilities, and no algorithm is the clear winner. We will choose a direction d_k in which to move from the current guess x_k. We may conduct a line search in that direction—to find the new $x_{k+1}=x_k+sd_k$ that minimizes the cost $C(x)$ while remaining admissible (or close to it). This search is one-dimensional, with a scalar unknown—the step size s. The direction d_k is the critical choice. Frequently the gradient of C is projected onto the subspace of directions permitted by the active constraints (as Karmarkar did). It is like Newton's method, linearizing near the current x_k and venturing a step on the basis of C'. In fact Newton's method becomes a quadratic program at each step; the cost and the active constraints are decided at x_k. That may be the best. At the end of the step new constraints will be active, just as one component became nonzero and another became zero in each simplex step. But nonlinear constraints bring extra difficulties, and a full discussion is hardly possible.

We might remark on the choice between minimizing C and solving an equation like $C'+yA'=0$. With inequality constraints most algorithms choose the minimization; with equality constraints $L'=C'+yA'=0$ becomes reasonable. In structural optimization there was a war between these two camps, recently settled by a

† When b_i goes up by δ, more vectors x become admissible and the minimum cost goes

compromise—in which the multipliers y are improved at the same time as the primal unknowns x. Duality won again.

Conditions for a Constrained Minimum

Imagine that the cost function $C(x)$ has a bowl-shaped graph. If it comes from a positive definite matrix, $C(x)=\frac{1}{2}x^TMx$, then the bowl is perfect. Its cross-sections $C=\text{constant}$, called "level curves" or "level surfaces," are ellipsoids. In general C will not be exactly a quadratic and the graph will be more uneven; its level curves are sketched in Fig. 8.9. The inner curves come from the low values of C, near the bottom of the bowl. The cost increases as we move out and up the bowl. The problem is to find the lowest point that satisfies the constraints, and we proceed in four steps.

1. To begin, let there be ***one equality constraint*** and let it be linear: $a_1x_1+\cdots+a_nx_n=b$. This is $A(x)=b$. It gives a vertical plane that slices through the graph of C. The constrained problem looks for the lowest point of their intersection, the minimum of $C(x)$ subject to $A(x)=b$. The idea is to look down from above on the graph of C, and watch what happens at that lowest point x^*—marked P in Fig. 8.9. The level curve through P is *tangent* to the cutting plane $Ax=b$.†

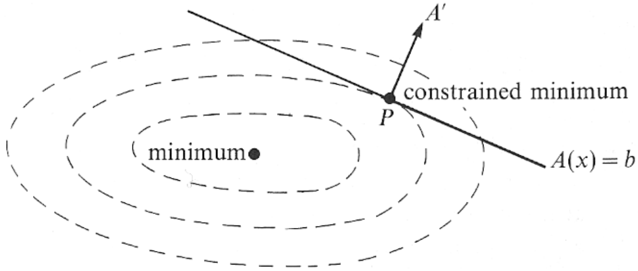


Fig. 8.9. Constraint $A(x)=b$ tangent to level curve at the solution x^*.

When two surfaces $A(x)=b$ and $C(x)=\text{constant}$ are tangent, their perpendicular directions are the same. One perpendicular comes from the vector $a=(a_1,\ldots,a_n)$, and the other comes from the gradient $C'=(\partial C/\partial x_1,\ldots,\partial C/\partial x_n)$. Since these vectors are in the same direction, the gradient must be a multiple of a:

$$\boxed{C'+ya=0 \text{ for some multiplying factor } y.} \qquad (2)$$

That is the key to constrained optimization. The partial derivatives of C may not be zero at the point P, but ***the derivatives of*** $C+yA$ ***are zero***. There are $n+1$

† The level curve cannot pass through the plane, or there would be an even lower level curve inside it still touching the plane. Since the curve just touches at P, it must be tangent.

unknowns $x_1, \ldots, x_n, y$, and $n+1$ equations

$$\frac{\partial C}{\partial x_1} + ya_1 = 0, \ldots, \frac{\partial C}{\partial x_n} + ya_n = 0 \text{ and } A(x) = b. \tag{3}$$

EXAMPLE 1 *Minimize* $C = \frac{1}{4}(x_1^4 + x_2^4)$ *subject to* $x_1 + 8x_2 = 34$.
The $n+1 = 3$ equations are

$$\begin{aligned} \frac{\partial C}{\partial x_1} + ya_1 &= 0, \quad \text{or} \quad x_1^3 + y = 0 \\ \frac{\partial C}{\partial x_2} + ya_2 &= 0, \quad \text{or} \quad x_2^3 + 8y = 0 \\ A(x) &= b, \quad \text{or} \quad x_1 + 8x_2 = 34. \end{aligned}$$

To solve them, $x_1 = -y^{1/3}$ and $x_2 = -2y^{1/3}$ give

$$x_1 + 8x_2 = -17y^{1/3} = 34, \quad \text{or} \quad y^{1/3} = -2, \quad \text{or} \quad y = -8.$$

The point P has $x_1 = 2$, $x_2 = 4$, and the minimum is $C = \frac{1}{4}(16 + 256) = 68$.

2. The step to ***one nonlinear constraint*** is easy. The surface $A(x) = b$ becomes curved instead of flat, but it is still tangent to the level surface of C at the point $x^* = P$. Therefore the two perpendicular vectors still go in the same direction. One is C', as before, and the other is A'. Previously, $A(x) = a_1x_1 + \cdots + a_nx_n$ was linear and its gradient was always $A' = (a_1, \ldots, a_n)$. Now A' varies from point to point, as C' does, but what matters is the situation at P—where they are in the same direction: $C' + yA' = 0$ for some multiplier y. In the next example the constraint surface is a circle $x_1^2 + x_2^2 = 1$.

EXAMPLE 2 *Minimize* $C = ax_1^2 + 2bx_1x_2 + cx_2^2$ *subject to* $x_1^2 + x_2^2 = 1$.
The $n+1$ equations are

$$\begin{aligned} \frac{\partial C}{\partial x_1} + y\frac{\partial A}{\partial x_1} &= 0, \quad \text{or} \quad 2(ax_1 + bx_2 + yx_1) = 0 \\ \frac{\partial C}{\partial x_2} + y\frac{\partial A}{\partial x_2} &= 0, \quad \text{or} \quad 2(bx_1 + cx_2 + yx_2) = 0 \\ A(x) &= b, \quad \text{or} \quad x_1^2 + x_2^2 = 1. \end{aligned}$$

Cancelling the factor 2, the first equations are

$$\begin{aligned} ax_1 + bx_2 &= -yx_1 \\ bx_1 + cx_2 &= -yx_2 \end{aligned} \quad \text{or} \quad \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -y \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Thus the optimal $x=(x_1, x_2)$ is an *eigenvector* of this matrix M; we have $Mx = -yx$. The minimum value of C is exactly the *smallest eigenvalue*, since

$$C=[x_1 \quad x_2]\begin{bmatrix} a & b \\ b & c \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -y[x_1 \quad x_2]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -y.$$

Geometrically, the level curve of C is an ellipse that touches the circle $x_1^2+x_2^2=1$ at the ends of its longest axis—and that axis points along the eigenvector. It is like minimizing the Rayleigh quotient

$$\frac{ax_1^2+2bx_1x_2+cx_2^2}{x_1^2+x_2^2}=\frac{x^TMx}{x^Tx}.$$

For any symmetric matrix, of any size, the minimum of $C(x)=x^TMx$ subject to $x^Tx=1$ is the smallest eigenvalue of M.

3. The next step is to admit ***two or more constraints***. Separately they are $A_i(x)=b_i$, and collectively (in a vector equation) they are $A(x)=b$. If they are linear, the gradients A_i' are the rows of a fixed matrix. If they are nonlinear, the A_i' depend on x. In either case, the minimizing point P must satisfy $A(x)=b$. If we move away from P, staying on the surface $A(x)=b$, the cost $C(x)$ must not decrease; otherwise P would not be minimal. Therefore the vector C', which is orthogonal at P to the surface $C=$ constant, must also be orthogonal to the surface $A(x)=b$.

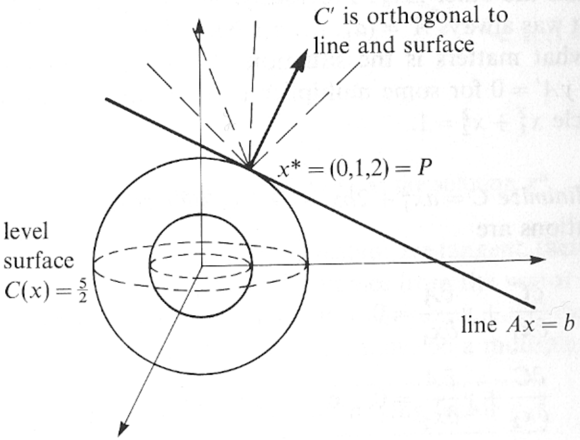


Fig. 8.10. The derivative C' at x^* is orthogonal to $A(x)=b$.

When there was one constraint, this orthogonality determined the direction of C'. With more constraints the surface $A(x)=b$ will be lower dimensional, like a line in three dimensions, and many directions are orthogonal to it. The gradient C' goes in one of those directions (Fig. 8.10), so

$$C' + y_1 A_1' + \cdots + y_m A_m' = 0 \text{ for some multiplying factors } y_i. \tag{4}$$

That is subtle but important. Think of the linear case, with constraint $Ax = 0$. The admissible x are in the nullspace of A. Each row of A is orthogonal to the nullspace, and C' can be any combination of those rows. That is equation (4), which has $n + m$ unknowns $x_1, \ldots, x_n, y_1, \ldots, y_m$. There are also $n + m$ equations, the n given by (4) and the m constraints. They are combined by saying that ***all partial derivatives of*** $L = C + y(Ax - b)$ ***are zero***:

$$\boxed{\frac{\partial L}{\partial x_j} = 0, j = 1, \ldots, n, \quad \text{and} \quad \frac{\partial L}{\partial y_i} = 0, i = 1, \ldots, m.} \tag{5}$$

The last m equations are the constraints $A(x) = b$.

EXAMPLE 3 *Minimize* $C = \frac{1}{2}x^T Hx - x^T f$ *subject to* $Ax = b$.
The gradient C' is $Hx - f$ and the $n + m$ equations are

$$\begin{aligned} \partial L/\partial x_j = 0: &\quad Hx + A^T y^T = f \\ \partial L/\partial y_i = 0: &\quad Ax \qquad\quad\;\; = b \end{aligned} \quad \text{or} \quad \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ y^T \end{bmatrix} = \begin{bmatrix} f \\ b \end{bmatrix}. \tag{6}$$

These are the all-important equations from Chapter 2, with the notations reversed: $x \leftrightarrow y$, $b \leftrightarrow f$, $A \leftrightarrow A^T$. We denoted the matrix by H, since C is now cost, and y has become a row vector: $y_1 A_1' + \cdots + y_m A_m'$ is $A^T y^T$. But the underlying problem is identical with the one in Chapter 2, to minimize a quadratic with linear constraints.

Figure 8.10 has $C = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$, with spheres as level surfaces. The rows from the constraint are $A_1' = [1 \quad 0 \quad 0]$ and $A_2' = [0 \quad 1 \quad 2]$. At the solution C' is A_2' and the optimality condition (4) is satisfied.

4. The final step is to allow ***inequality constraints***. They may be active or inactive—either they alter the minimum or they don't. Both possibilities appear in the simplest problem, *to minimize the number* x^2 *subject to* $x \leq b$. If b is positive, the minimum of x^2 is at $x = 0$. It is the absolute minimum and whether $b = 10$ or $b = 1000$ makes no difference. But if b is negative, and $x = 0$ is inadmissible because it violates $x \leq b$, the minimum is changed. It occurs where x *equals* b. The constraint becomes active, and it changes the minimum of x^2 from 0 to b^2.

How is this reflected in the Lagrange multiplier y? When the constraint is active, the derivatives of $x^2 + y(x - b)$ are

$$\frac{\partial L}{\partial x} = 2x + y = 0$$

$$\frac{\partial L}{\partial y} = \quad x - b = 0.$$

Thus $y = -2b$.† When the constraint is inactive, the multiplier y is zero and the minimum is zero (at $x = 0$).

The pattern is the same when there are n unknowns and m constraints, and it is the fundamental condition for optimality:

8H (Kuhn-Tucker optimality conditions) The minimum of $C(x_1, \ldots, x_n)$ subject to $A_i(x) \le b_i$ occurs where

$$\frac{\partial C}{\partial x_j} + y_1 \frac{\partial A_1}{\partial x_j} + \cdots + y_m \frac{\partial A_m}{\partial x_j} = 0, j = 1, \ldots, n \tag{7}$$

with y and x also subject to

$$y_i \ge 0,\ A_i(x) \le b_i,\ y_i(A_i(x) - b_i) = 0, i = 1, \ldots, m. \tag{8}$$

There are $n + m$ equations, but we cannot predict in (8) whether $y_i = 0$ or $A_i(x) = b_i$. Either the constraint is active or the multiplier is zero; the right equation is not known in advance.

In a full-scale treatment of optimization we would have to discuss the extra hypotheses that make this literally true. First, the functions C and A_i have been assumed smooth. If the graph of C has a corner, there is a whole family of "derivatives" and any one is acceptable in (7). Second, the vectors A_i' should be independent at the minimizing point or the y_i are not well determined. Third, and most important, the functions C and A_i should be ***convex*** (see below). Convexity is the prime requirement in proving that there is a constrained minimum. Without it the solution to (7–8) can be a saddle point, or a maximum, or fail to exist. With a strengthened form of convexity, the minimization succeeds.

EXAMPLE 4 (Linear programming) *Minimize* $C = c_1x_1 + \cdots + c_nx_n$ *subject to* $Ax \le b$.

The equations (7) in vector notation are $c + yA = 0$, and (8) is complementary slackness:

$$y \ge 0,\ Ax \le b,\ y(Ax - b) = 0. \tag{9}$$

They are the optimality conditions connecting the primal to the dual, when there is no sign constraint on x.

This is an example in which C is convex but not *strictly* convex—its second derivatives are zero. The zero matrix is positive semidefinite but certainly not positive definite. Therefore the minimum may fail to exist. It is $-\infty$, if we minimize $2x$ subject to $x \le 4$.

In the final example C is strictly convex; the matrix H is to be positive definite.

† y is the derivative of the minimum value $x^2 = b^2$, but with opposite sign.

EXAMPLE 5 (Quadratic programming) *Minimize* $C = \frac{1}{2}x^T Hx$ *subject to* $x_1 \le b_1, \ldots, x_n \le b_n$.

There are n constraints, n multipliers y_i, and $n + n$ equations:

$$\text{(7) becomes } Hx + y^T = 0 \tag{10}$$

$$\text{(8) becomes } y \ge 0,\ x \le b,\ y(x-b) = 0. \tag{11}$$

For $n = 1$ we are back to the two possibilities $x = b$ (active constraint) or $y = 0$ (inactive constraint). For $n = 2$ it is reasonable to test all four possibilities. As n increases, the number of combinations climbs to 2^n; each constraint can be active or inactive. For large n a good algorithm finds the right combination without trying them all.

Convex Functions

We need to know which functions $C(x)$ fit naturally into these minimum problems. They will not be the only functions that can be minimized, but they are the best ones. They were described earlier in terms of a "bowl-shaped graph"—which was intuitively correct but not overwhelmingly precise. The right description is in the definition of a ***convex function***, which extends one of the basic ideas of calculus—that the second derivative satisfies $f'' \ge 0$ at a minimum. The first requirement is $f' = 0$, and in our constrained problems that became $L' = C' + yA' = 0$. Without this *first-order condition*, a point is not even a candidate for a minimum. But for points which survive that test, there has to be a *second-order condition* (involving the second derivatives of L) to distinguish between minima and maxima and saddle points. A convex function will pass that second-order test, and a strictly convex function will pass with something to spare.

A convex function is defined in the same way as a convex set:

> A set E is convex if the line segment between any two of its points stays within the set.
>
> A function F is convex if the line segment between any two points of its graph lies on or above the graph.

If all line segments go strictly inside the set E, or strictly above the graph of F, then the set or the function is *strictly convex*. There are no flat segments on the boundary of the set or on the graph. A function $F(x) =$ constant, or a linear function $F(x) = a^T x$, or a feasible set in linear programming, is convex but not strictly convex.

There are three ways to test for convexity. The first two come directly from the definition and the third, which extends $f'' \ge 0$, is provided by calculus.

(1) The set of points on or above the graph of f should be a convex set. This set is called the "epigraph."

(2) At every point $x = cx_1 + (1-c)x_2$ between x_1 and x_2, the value of $F(x)$ must not be above the straight line value $cF(x_1) + (1-c)F(x_2)$:

$$F(cx_1+(1-c)x_2) \le cF(x_1)+(1-c)F(x_2) \quad \text{for } 0 \le c \le 1. \tag{12}$$

(3) The matrix H of second derivatives of F, $H_{ij} = \partial^2 F/\partial x_i \partial x_j$, must be positive semidefinite at every point.

The link between test 2 and test 3 comes from Taylor series expansions:

$$F(x) = F(x_1) + (x-x_1)^T \nabla F(x_1) + \tfrac{1}{2}(x-x_1)^T H(x-x_1) + \text{cubic terms.}$$

Expanding both sides of (12), the quadratic terms produce $(x_2-x_1)^T H(x_2 - x_1) \ge 0$. This is the positive semidefiniteness of H. It means that all eigenvalues of H are ≥ 0, all pivots are ≥ 0, and all symmetrically placed subdeterminants are ≥ 0. If H is positive definite, these numbers are strictly positive, (12) is true with strict inequalities, and F is strictly convex.

EXAMPLE The third test confirms the convexity of $F_1 = x^4 + y^6$ and it refutes the convexity of $F_2 = x^2y^2$. The function F_2 is not convex even though x^2 and y^2 are separately convex. The matrices H, called second gradients or Hessians, are easy to find from the second derivatives of F_1 and F_2:

$$H_1 = \begin{bmatrix} 12x^2 & 0 \\ 0 & 30y^4 \end{bmatrix} \quad \text{and} \quad H_2 = \begin{bmatrix} 2y^2 & 4xy \\ 4xy & 2x^2 \end{bmatrix}.$$

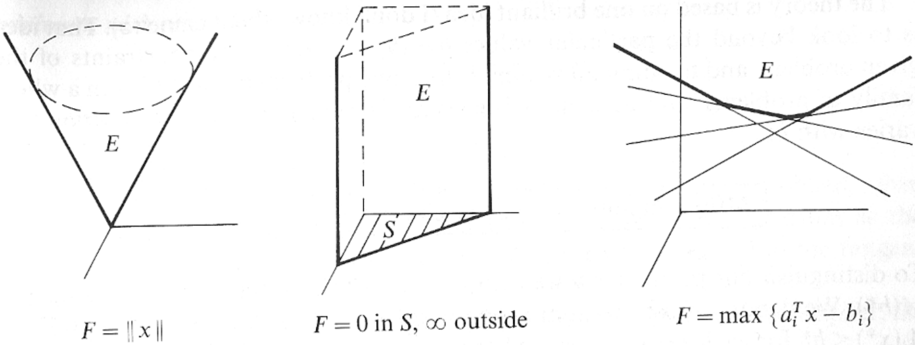
The determinant of H_2 is negative so one of its eigenvalues must be negative. The graph of F_2 is a parabola in the x and y directions, but overall it cannot be convex. $F_2 = x^2y^2$ is zero at $x=0$, $y=2$ and also at $x=2$, $y=0$, but between them it is positive. It goes *above* the line segment connecting those points, and all tests for convexity must fail.

This definiteness condition on the second derivatives is completely successful when F is smooth, but it is crucial to recognize that there are nonsmooth possibilities:

(a) A convex function like the absolute value $|x|$ has a corner, where the second derivative becomes a delta-function. The graph resembles the letter V. In n dimensions it turns into the length function $F(x) = \|x\|$, whose graph is the cone in Fig. 8.11. The set E above the graph is pointed but still convex, and condition (12) becomes the triangle inequality

$$\|cx_1+(1-c)x_2\| \le \|cx_1\| + \|(1-c)x_2\|.$$

(b) A convex function may be ***infinite*** at some points. Condition (12) keeps it finite on the line between x_1 and x_2, if it is finite at those two points. Therefore the set on which F is finite must be convex. One particular function is important: $F(x)=0$ when x is in the convex set S, and $F(x) = +\infty$ when x is not in S. It is known as the "indicator function" $I(x)$ of the set S, and it is extremely useful for constraints:



$F = \|x\|$ $F = 0$ in S, ∞ outside $F = \max \{a_i^T x - b_i\}$

Fig. 8.11. Convex sets E above the graphs of convex functions.

To minimize $C(x)$ subject to x in S, we minimize $C(x) + I(x)$ for all x. The points not in S lead immediately to $+\infty$, so effectively it is a minimum of $C(x)$ over the set S.

The sum of convex functions is automatically convex. So is the maximum of two or more convex functions, and the third graph in Fig. 8.11 is an example. With an infinite number of planes, the graph of the maximum can be curved (but still convex!). We could even reach the other two graphs in the figure, by choosing the right planes. In fact every convex function is the maximum of a family of linear functions—and to understand that we go back to convex sets.

The great property of a convex set E is that through every boundary point there is at least one ***supporting hyperplane***. This is a plane that touches the boundary of E, keeping the rest of the set on one side. We assume the boundary is included in E; E is a "closed" convex set. Where the boundary is smooth, the only supporting plane is the one tangent to E. Where the boundary is pointed, at a corner, there are a lot of supporting planes through the boundary point. If we know all these planes, we can reconstruct E.† It will be the intersection of all the halfspaces cut out by the supporting planes.

Moving from convex sets to convex functions, these supporting planes become *tangent planes* to the graph. Their slopes are the *derivatives* of the function. It is the existence of these derivatives—the fact that there is at least one tangent plane at every point of the graph (and more than one, at the vertex of the cone)—that will produce a saddle point. The planes are exactly what is needed for a general duality theorem, bringing together all the specific cases proved earlier in this book.

Convexity and Duality

Suppose the cost $C(x)$ and the constraints $A_i(x)$ are all convex functions. If they are linear, we have linear programming. If C is quadratic, we have quadratic programming. In general we have nonlinear programming, and we are ready to show how it is transformed by Lagrange multipliers.

† This is basic to duality: A convex set can be described by saying which points the set contains, or by saying which half-spaces contain the set.

The theory is based on one brilliant idea (I don't know who it came to). That idea is to look beyond the particular values $b=(b_1, \ldots, b_m)$ in the constraints of the given problem, and to admit all vectors b. Our one problem is embedded in a whole family of problems, and their minimum values produce a ***minimum function*** that varies with b:

$$\boxed{M(b) = \text{minimum value of } C(x) \text{ subject to } A(x) \leq b.} \tag{13}$$

To distinguish our particular b we denote it by b^*; the specific problem is to find $M(b^*)$. We denote by x^* the minimizing point in that problem (if it exists). Thus $A_i(x^*) \leq b_i^*$ for each $i=1, \ldots, m$, and the minimum cost is $M(b^*)=C(x^*)$.

Now enters the hypothesis that C and the A_i are convex functions. It follows that the minimum value $M(b)$ is not only decreasing as b increases (because more candidates x are admitted). *It is also a convex function of b.* The example with cost $C=x^2$ and constraint $x \leq b$ is sketched in Fig. 8.12—its minimum $M(b)$ equals b^2 for negative b and zero for positive b. The fact that M is convex when C and A are convex is verified in the exercises. This convexity allows the general theory to make its contribution: At any point like b^*, the graph of M has a supporting tangent plane. The plane has height $M(b^*)=C(x^*)$ at the point b^*, and at no point does the plane go above the graph of M.

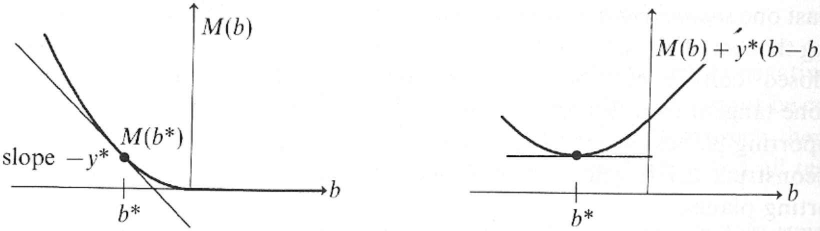


Fig. 8.12. $M(b) =$ minimum of x^2 with $x \leq b$.

If you tilt your head to make the plane horizontal, the whole curve has its minimum at b^*. That is the point of Fig. 8.12b; we produce an *unconstrained minimum* by adding a linear term that comes from the plane. If the slope of the plane is $-y^*$, the linear term to add is $y^*(b-b^*)$. It equals zero at b^*, but for larger b it raises the curve and for smaller b it lowers it. Then the value $C(x^*)$ at $b=b^*$ becomes an absolute minimum:

$$C(x^*) = \min_{b} \min_{A(x) \leq b} [C(x) + y^*(b-b^*)]. \tag{14}$$

This y^* is the right Lagrange multiplier for the original problem. It is nonnegative, as expected with inequality constraints. Its components satisfy $y_i^* \geq 0$, because $-y^*$ is the slope of a decreasing function $M(b)$. That slope is the ***sensitivity of the***

minimum with respect to b:

$$\boxed{-y_i^* = \frac{\partial M}{\partial b_i} \text{ at } b = b^*.} \tag{15}$$

The Lagrange multiplier—which will be the solution to the dual problem, when that appears—is the ***marginal cost***: y^* gives the change in the minimum as the constraints are changed. The whole theory of sensitivity comes from the tangent plane to $M(b)$.

To go from sensitivity to duality, and to show that y^* is the correct multiplier, the technical step is to verify from (14) that also

$$C(x^*) = \min_x \,[C(x) + y^*(A(x) - b^*)]. \tag{16}$$

We do it quickly. For any x in (16), choose the same x in (14) and choose $b = A(x)$. Then the two expressions agree, and since (14) allows other choices we have $(16) \geq (14)$. On the other hand, for any b and x the requirement $A(x) \leq b$ makes $(16) \leq (14)$, remembering $y^* \geq 0$. Therefore the two are equal.

At the point x^* something special must happen. The term $y^*(A(x^*) - b^*)$ could not be negative, or equation (16) would be ridiculous. Since $y^* \geq 0$ and $A(x^*) \leq b^*$, the only alternative is the ***complementarity condition***:

$$\boxed{\text{for each } i, \text{ either } y_i^* = 0 \text{ or } A_i(x^*) = b^*.} \tag{17}$$

Then the inner product $y^*(A(x^*) - b^*)$ is zero, which is the Kuhn-Tucker condition. We have reached the main result of Lagrange duality:

8I If the cost $C(x)$ and the constraint functions $A_i(x)$ are strictly convex, then

$$\min_{A(x) \leq b^*} C(x) = \max_{y \geq 0} \min_x \,[C(x) + y(A(x) - b^*)]. \tag{18}$$

The constrained minimization splits into an unconstrained minimization of L with a parameter y, followed by a maximization (the dual problem) over y. The optimal x^* in the primal problem on the left and the optimal y^* in the dual problem on the right are related by the Kuhn-Tucker conditions (7) and (8).

Proof At $y = y^*$ the two minima in (18) agree; that is $(14) = (16)$. For other $y \geq 0$ the right side could not be larger than the left. This is weak duality, which is always easy:

$$\text{if } y \geq 0 \text{ and } A(x) \leq b^*, \quad \text{then } C(x) \geq C(x) + y(A(x) - b^*).$$

Therefore (18) is correct, the minimum equals the maximum, and duality holds. Our simplest example will illustrate it best.

EXAMPLE *Minimize $C = x^2$ subject to $x \le b$*
The unconstrained minimum of $L = x^2 + y(x - b)$ comes first:

$$L' = 2x + y = 0 \text{ at } x = -\tfrac{1}{2}y, \text{ so the minimum is } L = -\tfrac{1}{4}y^2 - by.$$

Then the maximum over $y \ge 0$ (the dual problem) is

$$\max\left(-\tfrac{1}{4}y^2 - by\right) = \begin{cases} b^2 & \text{at} \quad y = -2b, \quad \text{if} \quad b \le 0 \\ 0 & \text{at} \quad y = 0, \quad \text{if} \quad b \ge 0 \end{cases}$$

This is the right side of (18); the minimum of C is b^2 or zero.†

Conjugate Convex Functions

Behind this analysis of duality lies a beautiful piece of geometry. We have hinted at it, twice at least, and the book will be incomplete until we say what it is. It brings together the applications, and then we are finished.

The geometry starts with a convex function, for example $F(x) = x^2$. The question is: Which straight lines lie below this parabola? The graph of $yx - d$ is a line with slope y, and the line is below the parabola if $yx - d \le x^2$ for all x. That is true if the depth d is large enough:

$$d \ge yx - x^2 \text{ for all } x. \tag{19}$$

The right side is largest where its derivative is zero. At that point $y - 2x = 0$, or $x = \frac{1}{2}y$, and the requirement becomes $d \ge \frac{1}{4}y^2$. The line will just touch the parabola, as in Fig. 8.13, if $d = \frac{1}{4}y^2$.

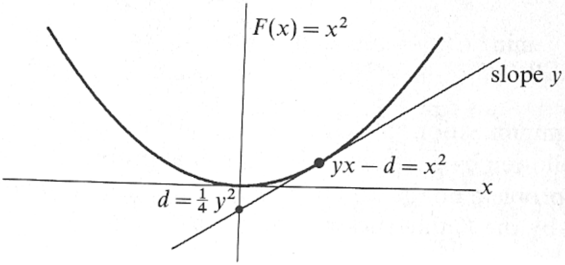


Fig. 8.13. Tangent line under $F(x)$, at depth $d = F^*(y)$.

† I always doubted that duality could make a problem easier (here it doesn't). But for a continuous maximum flow problem with capacity $|v| \le 1$ in the unit square, duality shows that div $v = 2 + \sqrt{\pi}$ is possible. There is a prize of 10,000 yen for v.

Now consider all these touching lines, with different slopes y. Their envelope is the parabola! We get back to x^2 by looking always for the highest line:

$$\max_y \left[yx - \frac{y^2}{4} \right] = x^2. \tag{20}$$

The maximum is at $y = 2x$, and at that point (20) is $2x^2 - x^2 = x^2$.

This duality between convex functions and tangent lines extends far beyond parabolas. The functions depend on x and the tangents depend on y. It is usual to write $F^*(y)$ rather than $d(y)$, to emphasize the parallel between F and F^*. This ***conjugate function*** F^* looks so ordinary and innocent; it is the constant term d that raises or lowers the line until it touches the parabola. However its construction is at the center of convex analysis, and the steps that succeeded for a parabola in (19) and (20) will succeed for every F:

8J Suppose $F(x)$ is a convex function. For each slope y let

$$d = F^*(y) = \max_x \left[yx - F(x) \right]. \tag{21}$$

This conjugate function F^* is also convex, and for every x and y it satisfies $F^* \geq yx - F$, or $F \geq yx - F^*$. Then the maximum over the tangent lines $yx - d$ brings back F:

$$F(x) = \max_y \left[yx - F^*(y) \right]. \tag{22}$$

Since (22) repeats the operation in (21), the conjugate of F^* is $F^{**} = F$. In other words, the dual of the dual is the primal.

The step from F to F^* is the *Legendre-Fenchel transform*—named after the mathematician who used it in physics and the one who saw its possibilities as mathematics. Legendre concentrated on smooth functions; Fenchel allowed corners, and jumps to infinity. For smooth F the maximum occurs where the derivative of $yx - F(x)$ is zero: $y = F'(x)$. For the parabola this was $y = 2x$ and it gave $x = \frac{1}{2}y$. Notice that we are looking for x! The equation $y = F'(x)$ has to be solved—the function F' has to be "inverted" to find $x = (F')^{-1}(y)$—and this is the subtle point in the calculation. Fortunately F' is an increasing function (since F was convex). Then transforming from F^* back to $F = F^{**}$ reverses this process. The derivative in (22) is zero at $x = (F^*)'(y)$; for the parabola this was $x = \frac{1}{2}y$. We are looking for y and we rediscover $y = 2x$. Let me try to put that relationship into words:

> There is a pairing between slopes y and points x. The tangent with slope y touches the graph of F at the corresponding x. For each pair that means
>
> $$F(x) + F^*(y) = xy, \quad \text{or} \quad F^* = xF'(x) - F(x). \tag{23}$$

Furthermore the derivatives $G=F'$ and $H=(F^*)'$ are inverse to each other: $H(G(x))=x$ and $G(H(y))=y$.

In the example F was x^2, F^* was $\frac{1}{4}y^2$, and the slope paired with x was $y=2x$. You can check that (23) is correct. The last statement is easy, since G was multiplication by 2 and H was multiplication by $\frac{1}{2}$.

In Section 3.6 the transformation from F to F^* took the Lagrangian to the Hamiltonian. The pairing was between velocity v and momentum $p=mv$; the kinetic energy was given equally by $F=\frac{1}{2}mv^2$ and $F^*=\frac{1}{2}p^2/m$. Other pairs are fundamental in science and engineering. In statics the pairing is between strain and stress; F and F^* are strain energy and complementary energy. In thermodynamics one pair is pressure and volume, and the Gibbs free energy and the Helmholtz free energy. For electrical networks there is potential difference and current. In n dimensions the tangent lines become tangent planes, but the geometry holds on.

Applications: Minimum Norms

As we end this part of the book, devoted to optimization and duality, it is amazing to see how many applications come from a single source. That source was present at the beginning, in the first example of Chapter 2—the distance to a line. In n-dimensional space, it would become the distance to a flat surface $Ax=b$. And if different norms (measures of distance) are allowed, including the familiar $\|x\|^2=x_1^2+\cdots+x_n^2$ along with others, then the applications begin to appear. They all minimize that distance, subject to some form $Ax=b$ of Kirchhoff's current law, which permits a brief review of the whole subject:

(1) ***Transportation problem***: Minimize the shipment cost $C_1x_1+\cdots+C_nx_n$
(2) ***Resistive network***: Minimize the heat dissipation $x_1^2R_1+\cdots+x_n^2R_n$
(3) ***Maximal flow*** (stated differently): Minimize the maximum of $|x_1|/c_1, \ldots, |x_n|/c_n$.

If the costs and resistances and capacities are all 1, the distances are

$$\|x\|_1=|x_1|+\cdots+|x_n|;\ \|x\|_2=(x_1^2+\cdots+x_n^2)^{1/2};\ \|x\|_\infty=\max|x_i|.$$

The first and third are the "l^1 norm" and "l^∞ norm" of x. They are associated with linear programming. If one of them appears in the primal, the other will appear in the dual. In between these two, and dual to itself, is the ordinary Euclidean length $\|x\|_2$. This is the "l^2 norm" of x. It is squared in the electrical problem and in all of Chapter 2, and it leads to quadratic programming.

I realize now that all these problems lead back to the first example of duality in this book: ***The minimum distance to a line equals the maximum distance to planes through that line.*** It is true that the line has become a higher-dimensional surface; it is the graph of $Ax=b$. Originally I was thinking of an ordinary line, with 2 equations in 3-dimensional space, but one virtue of algebra is its freedom from that limitation. The dual problems are associated with Federal Express, and Sprint, and a minimum cut. Perhaps even the marriage problem fits into this framework; it must. So the whole of duality theory in these applications has come to depend on one final calculation, the distance from a point to a plane.