3 Optimization Methods: One-to-One Distribution

Readings for Chapter 3

Newell (1971) shows how to find an optimal sequence of headways for a transportation route serving a changing demand over time with a continuum approximation method that avoids "details." This problem is mathematically analogous to the problems with time dependent demand addressed in this chapter, which are traditionally solved with dynamic programming. Section 3.3, is based on this reference. Daganzo (1987) shows that a continuous approximation of a function and its variables can be *more* accurate than the exact, detailed and discontinuous world representation they replace. This result is discussed in Section 3.2.

3.1 Initial Remarks

This chapter describes logistics problems linking one origin and one destination (one-to-one problems) and the methods used to solve them. The following points, mentioned in Chapter 1, will be revisited:

(i) Accurate cost estimates can be obtained without precise, detailed input data,

(ii) Departures from an optimal decision by a moderate percentage do not increase cost significantly. Since there is no need to seek the most accurate estimate of the optimum, there may be little use for highly detailed data,

(iii) Detailed data may get in the way of the optimization, actually *hindering* the search for an optimum,

(iv) Thus, we advocate a two-step solution approach to logistics problems: the first (analytical) step involves little detail and yields broad solution concepts; the second (or fine tuning) step leads to specific solutions, consistent with the ideals revealed by the first – it uses all the relevant detailed information.

These points will be illustrated with simple extensions of the EOQ model introduced in Chapter 2. Section 3.2 analyzes one-to-one systems with constant production and consumption rates; the discussion focuses on the robustness and accuracy of the results. Section 3.3 examines the same problem when the demand varies over time; it describes numerical methods and a continuous approximation (CA) analytical approach that is based on summarized data. Section 3.4 illustrates how the CA approach can be used for a location problem that has an analogous structure, and Section 3.5 demonstrates the accuracy of the CA solutions.

As a prelude to the more complex problems explored in forthcoming chapters, Section 3.6 explains how the CA approach can be extended to multidimensional problems with constraints, and Section 3.7 discusses network design issues.

3.2 The Lot Size Problem with Constant Demand

Let us now explore the optimization problem for the optimum shipment size, v^*, described in the previous chapter:

$$z = \min\left\{ Av + \frac{B}{v} : \; v \le v_{\max} \right\}. \tag{3.1}$$

Consider first the case $v_{max} = \infty$. Then v^* is the value of v which minimizes the convex expression $Av + Bv^{-1}$:

$$v^* = \sqrt{\frac{B}{A}}. \tag{3.2}$$

Remember that B represented the fixed motion costs, c_f, and A the holding cost per item, c_h/D'. Note that v^* is the value which makes both terms of the objective function equal. That is, for an optimal shipment size, holding cost = motion cost.

The optimum cost per item is:

$$z^* = (cost/item)^* = 2\sqrt{AB}, \tag{3.3}$$

which is easy to remember as "twice the square root of the product" of the two terms of (3.1). It will be convenient to memorize Eqs. (3.2) and (3.3), since EOQ minimization expressions will arise frequently.

As a function of c_f, c_h and D', the optimum cost per item increases at a decreasing rate with c_f and c_h and decreases with the item flow D'. There are *economies of scale*, since higher item flows lead to lesser average cost.

In the remainder of this section we examine the sensitivity of the resulting cost to errors in: (i) the decision variable, v , (ii) the inputs (A or B), and (iii) the functional form of the equation.

3.2.1 Robustness in the Decision Variable

Suppose that instead of v^*, the chosen shipment size is $v^0 = \gamma v^*$, where γ[1] is a number close to 1, capturing the relative error in v^0. Then, the ratio of the actual to optimum cost z^0/z^* will be a number, γ', greater than 1, satisfying:

$$\gamma' = \left[A\sqrt{\frac{B}{A}}\gamma + B\sqrt{\frac{A}{B}}\frac{1}{\gamma} \right] \Big/ \left[2\sqrt{AB} \right] = \frac{1}{2}\left[\gamma + \frac{1}{\gamma} \right]. \qquad (3.4)$$

Independent of A and B , this relationship between input and output relative errors holds for all EOQ models. It indicates that if γ is between 0.5 and 2, so that the optimal shipment size is approximated to within a factor of 2, then $\gamma' \le 1.25$. If γ is between 0.8 and 1.25, then $\gamma' \le 1.025$. Thus, a cost within 2.5 percent of the optimum can be reached if the decision variable is within 25 percent of optimal. On the other hand, if γ is several times larger (or smaller) than 1, then the cost penalty is severe, i.e., $\gamma' \approx \gamma$ (or $\gamma' \approx 1/\gamma$). Obviously, thus, while it is important to get reasonably close to the optimal value of the decision variable (say to within 20 to 40 percent), from a practical standpoint it may not be imperative to refine the decision beyond this level.

3.2.2 Robustness in Data Errors

Let us now assume that one of the cost coefficients A (or B) is not known precisely. If it is believed to be $A' = A\delta$ (or $B' = B\delta$), for some $\delta \approx 1$, then the optimal decision with this erroneous cost structure is:

---

[1] This symbol is unrelated to the coefficient of variation of the prior chapter, also denoted by γ.

$$v^{*'} = \left(\sqrt{\frac{B}{A}}\right)\delta^{-1/2} = v^* \,\delta^{-1/2} \qquad \text{if} \quad A' = A\delta \,,$$

or

$$= v^* \,\delta^{1/2} \qquad \text{if} \quad B' = B\delta \,.$$

Because the actual to optimal shipment size ratio, $v^{*'}/v^*$, is either $\delta^{-1/2}$ or $\delta^{1/2}$ (see Eq. (3.2)), the cost penalty paid is as if $\gamma = \delta^{1/2}$. Thus, the resulting cost is even less sensitive to the data than it is to the decision variables. For example, if the input is known to within a factor of 2 $(0.5 \le \delta \le 2)$, then $0.7 \le \gamma \le 1.4$ and $\gamma' \le 1.1$. The cost penalty would be about 10 percent, whereas before it was 25 percent. The penalty declines quickly as δ approaches 1. This robustness to data errors is fortunate because, as we pointed out in Chapter 2, the cost coefficients (for waiting cost especially) are rarely known accurately.

3.2.3 Robustness in Model Errors

A cost penalty is also paid if the EOQ formula itself is inaccurate. To illustrate the impact of such functional errors, we assume that the actual cost, a complicated (perhaps unknown) expression, can be bounded by two EOQ expressions; the cost penalty can then be related to the width of the bounds.

Suppose, for example, that the actual holding cost $z_h(v)$ is not exactly equal to the EOQ term (Av), but it satisfies:

$$Av - \Delta/2 \le z_h(v) \le Av + \Delta/2$$

for some small Δ. (Such a situation could happen, for example, if storage space could only be obtained in discrete amounts.) Because Δ is small, the EOQ lot size v^* is adopted. Clearly, then the absolute difference between the actual cost $[z_h(v^*) + B/v^*]$ and the predicted EOQ cost z^* cannot exceed $\Delta/2$. It is also easy to see that the difference between the optimal cost with perfect information, $\min\{z_h(v) + B/v\}$, and z^* cannot exceed $\Delta/2$ either. As a result, the difference between the actual and theoretical minimum costs – the cost penalty – is bounded by Δ.

Usually, though, this penalty will be significantly smaller than the maximum possible; Figure 3.1 illustrates the unusual conditions generating the largest penalty. Thus, if Δ is small compared to z^* (e.g., within 10 percent) the functional form error should be inconsequential. The same conclusion is reached if the motion cost is also inaccurate. In general, the EOQ solution will be reasonable if it is accurate to within a small fraction of its predicted optimal cost.

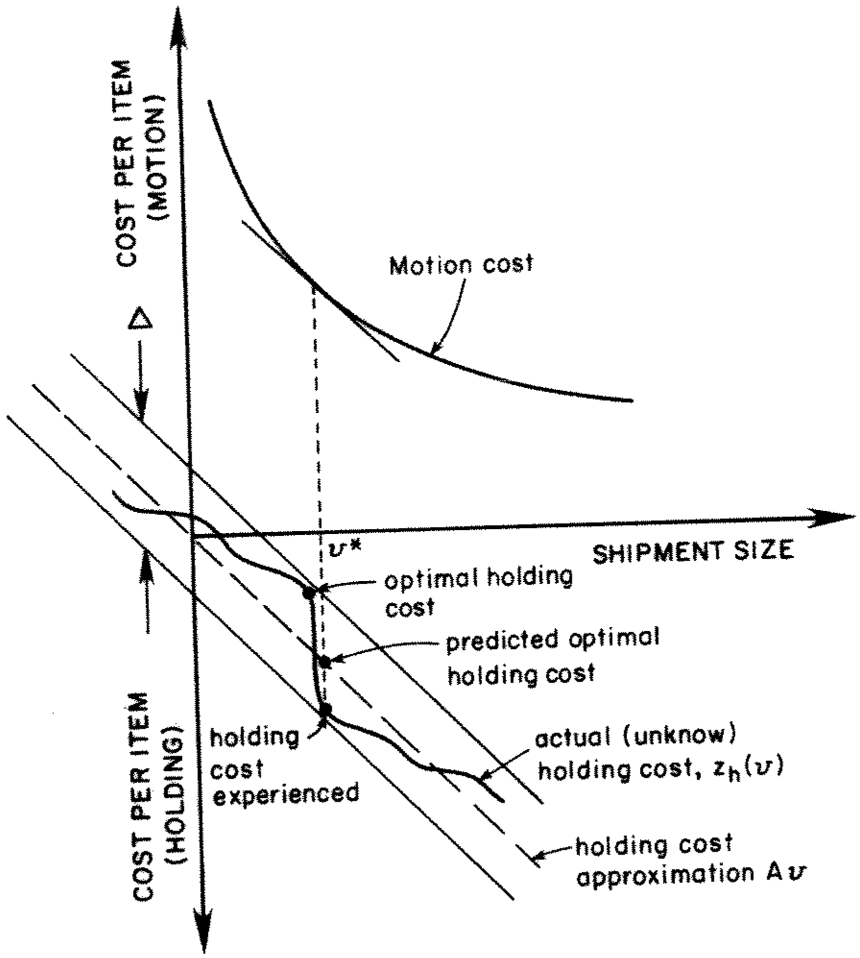

Fig. 3.1 Cost penalty resulting from errors in the holding cost function

3.2.4 Error Combinations

If errors of the three types exist, one would expect the cost penalty to be greater. Fortunately though, when dealing with errors the whole (the combined penalty) is not as great as the sum of its parts.

Suppose for example that the lot size recipe is not followed very precisely (because, e.g., lots are chosen to be multiples of a box, only certain dispatching times are feasible, etc.) and that as a result 40 percent discrepancies are expected between the calculated and actual lot sizes. We have already seen that such discrepancies can be expected to increase cost by about 10 percent. Let us assume, in addition, that one of the inputs (A or B) is suspected to be in error by a factor of 2, which taken alone would also increase cost by about 10 percent. Would it then be reasonable to expect a 20 percent cost increase? The answer is no; it should be intuitive that the penalty paid by introducing an input error when the lot size decision does not follow the recipe accurately should be smaller than the penalty paid if the decision follows the recipe. In our example, the combined likely increase is 14 percent [the square root of the sum of the squared errors: $.14 = (.1^2 + .1^2)^{1/2}$]. Statistical analysis of error propagation through models reveals similar composition laws in more general contexts (see e.g., Daganzo, 1985). This subject, however, is beyond the scope of this monograph. Further information can be found in Taylor (1997).

The above example illustrated how input and decision errors propagate. Although model errors follow similar laws – the whole is still less than the sum of the parts – for some approximate models the results are surprising. The composed (data and model) error can be actually *smaller* than the data error alone with the exact model! (Daganzo, 1987). This fortuitous phenomenon, illustrated by problem 3.1, has a special significance because it arises when, as recommended in this monograph, certain discontinuous models with discrete inputs are approximated by continuous functions and data. A more detailed discussion of this issue can be found in Daganzo (1987).

For ease of exposition, our discussion of robustness and errors ignored the $v \leq v_{max}$ constraint of Eq. (3.1), although similar remarks could have been made for the constrained solution and other non-EOQ models (see exercise 3.10). The constrained EOQ solution is now presented rather briefly, before turning our attention to the lot size problem with variable demand.

If, in solving the unconstrained EOQ problem, we find that $v^* > v_{max}$, then the solution is not feasible. In that case, choosing $v = v_{max}$ is optimal. Hence, the optimal EOQ solution can be expressed as:

$$v^* = \min\left\{\sqrt{\frac{B}{A}}, v_{max}\right\}, \qquad (3.5a)$$

and the optimal cost per item z^* is:

$$\begin{aligned} z^* &= 2\sqrt{AB} && \text{if } \sqrt{B/A} \le v_{max} \\ &= Av_{max} + \frac{B}{v_{max}} && \text{if } \sqrt{B/A} > v_{max} \end{aligned} \tag{3.5b}$$

Note that z^* is an increasing and concave function of A , and also of B (see Fig. 3.2a and b). As a function of $1/A = D'/c_h$, and thus of D' , z^* is decreasing and convex; the economies of scale continue to exist for all ranges of D'. Finally, note that the total cost per unit time, $D'z^*$, is proportional to $D'^{1/2}$ until the capacity constraint is reached, and from then on increases linearly with D'. The critical point is $D'_{crit} = (v_{max})^2/c_f$. The general form of the relationship is depicted in Fig. 3.2c.

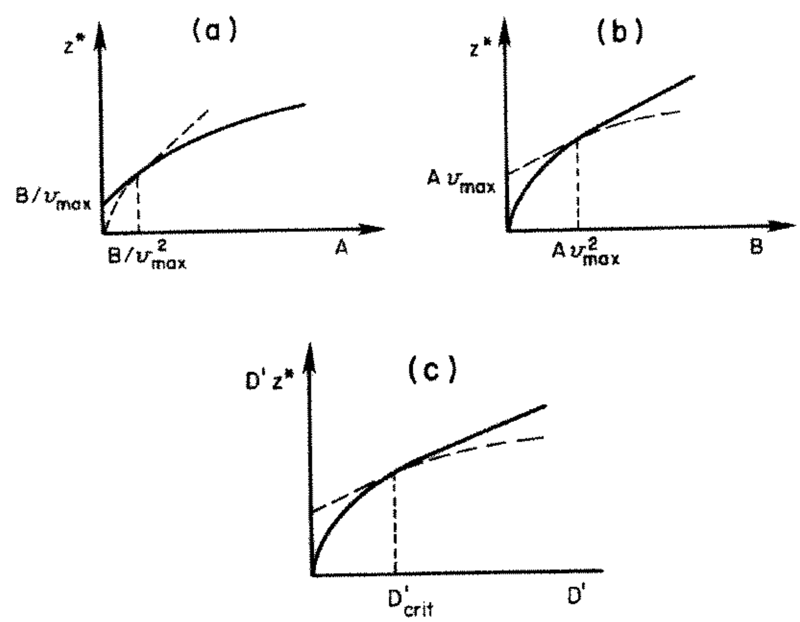

Fig. 3.2 Optimal EOQ cost as a function various parameters: (a) holding cost per item, A; (b) fixed motion costs, B; and (c) demand rate, D'. Dashed lines are the unused branches of Eq. (3.5b)

3.3 The Lot Size Problem with Variable Demand

Let us now consider the EOQ problem over a finite time horizon when the consumption rate D' changes with time in a predictable manner. The demand pattern, an input to our problem, is characterized by a function D(t) that gives the cumulative number of items demanded between times 0 (the beginning of the study period) and t . The time derivative of this function D'(t) represents the variable demand rate. We then seek the set of times when shipments are to be received ($t_0 = 0$, t_1 , ... , t_{n-1}), and the shipment sizes (v_0, v_1 , ... , v_{n-1}), that will minimize the sum of the motion plus holding costs over our horizon, $t \in [0, t_{max}]$.

As in Chapter 2, we also define as inputs to our problem a fixed (motion) cost per vehicle dispatch c_f , a holding cost per item-time $c_h = c_r + c_i$, and a maximum lot size v_{max} . With an infinite horizon and a constant demand, D(t) = D't , this formulation reduces to the EOQ problem examined in Section 3.2, where $A = c_h/D'$ and $B = c_f$.

For most of this section, we assume that the v_{max} constraint can be ignored. We will relax this restriction in Section 3.6. Subsection 3.3.1, below, examines the variable demand problem when rent costs are the dominant part of holding cost; a simple solution can then be obtained. Subsection 3.3.2 shows that if inventory (waiting) costs are dominant, then the solution is not quite as apparent; two solution methods are then described: a numerical method in subsection 3.3.3 and an analytical method in subsection 3.3.4.

3.3.1 *Solution When Holding Cost Is Close to the Rent Cost*

If inventory cost is negligible, $c_i << c_r$, then holding cost approximately equals rent cost $c_h \approx c_r$. We have already mentioned that rent cost increases with the *maximum* inventory accumulation (regardless of when it is held), and that otherwise the cost is rather insensitive to the accumulations at other times. This property of holding cost simplifies the solution to our problem.

Recall from Sec. 2.3 that given a set of n shipments, the motion cost during the period of analysis, $c_f n$, is independent of the shipment times and sizes. The problem, then, is to find the sets of shipment times and sizes that will minimize holding cost. A lower bound to the maximum accumulation at the destination is the size of the largest shipment received, which is minimized when all the shipments are equal. Hence, the largest shipment – and, thus, the maximum accumulation – must exceed or at least equal $D(t_{max})/n$. If a set of times and shipment sizes is found for which the maximum accumulation equals $D(t_{max})/n$, the set is an optimal way of sending n shipments with rent cost per unit time: $c_r D(t_{max})/n$. Figure 3.3 depicts such a solution for a hypothetical cumulative consumption curve D(t) . Each shipment is just large enough to meet the demand until the next shipment; the consumption between consecutive receiving times, the same in all cases, is $D(t_{max})/n$. Clearly then, the following strategy is optimal:

(i) Divide the ordinate axis between 0 and $D(t_{max})$ into n equal segments and find the times t_i for which D(t) equals $(i/n)D(t_{max})$ for $i = 0, \dots, n - 1$. These are the shipment times,

(ii) Dispatch barely enough to cover the demand until the following shipment.

One must now find the optimal n by minimizing the resulting cost. Interestingly, it does not depend on the t_i , only on n:

$$cost/time = c_r\left(D(t_{max})/n\right) + c_f\left(n/t_{max}\right), \quad and$$
$$cost/item = \left(\frac{c_r}{\overline{D'}}\right)\left(\frac{D(t_{max})}{n}\right) + c_f\left(\frac{n}{D(t_{max})}\right) , \tag{3.6}$$

where $\overline{D'}$ is the average consumption rate:

$$\overline{D'} = D(t_{max})/t_{max} .$$

Note that (3.6) is the EOQ expression with $v = D(t_{max})/n$. The solution now requires that n be an integer (there are constraints on v), but we have already seen that any v close to the unconstrained v* is near optimal. As a result, unless the time horizon is so short that n* = 1 or 2, the optimal cost per item should be close to the cost with constant demand.

It should be intuitive that if $v_{max} < \infty$, the solution procedure does not change. It is still optimal to have equal shipment sizes, but the number of shipments should be large enough to satisfy: $D(t_{max})/n \le v_{max}$. The solution is still of the form (3.5), with v^{-1} restricted to being an integer multiple of $D(t_{max})^{-1}$.

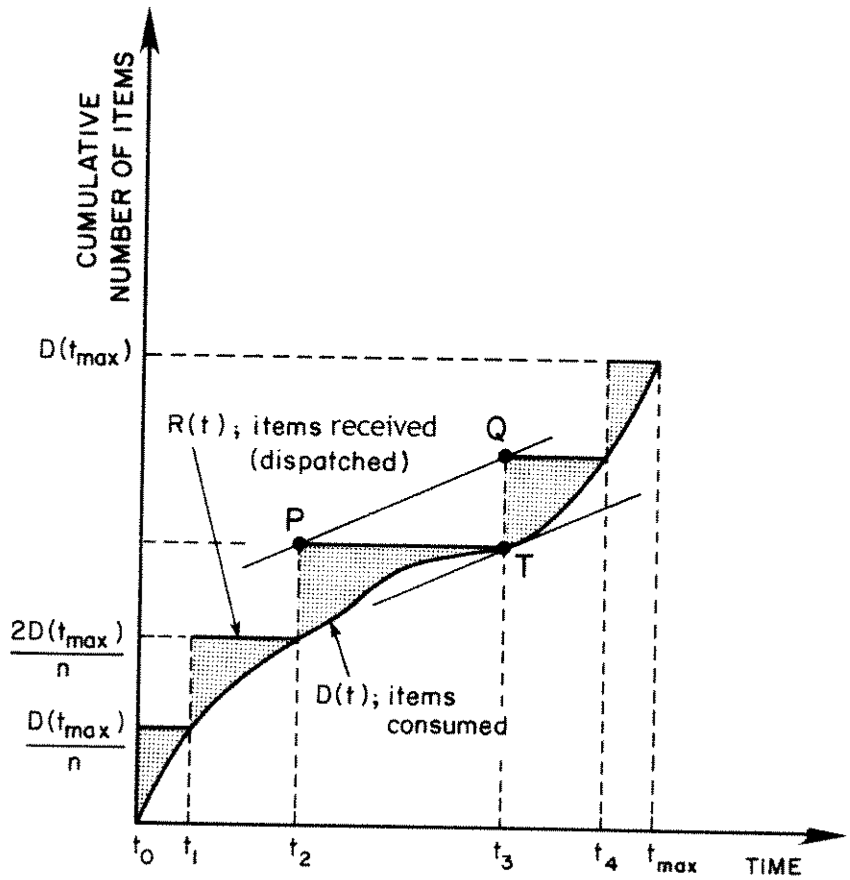

Fig. 3.3 Selection of shipment times for least holding cost

3.3.2 *Solution when Rent Cost Is Negligible*

Let us now examine another extreme but common situation, where items are so small and expensive, that most of the holding cost arises from the item-hours spent in inventory, and not from the rent for the space to hold them. In this case the destination's holding cost should be proportional to the shaded area of Fig. 3.3.

The combined origin-destination holding cost will also be proportional to this area if (i) the origin holding cost can be ignored, or (ii) if it is proportional to the area. Situation (i) arises if the origin produces generic items for so many destinations that the part of its costs that would be prorated to each destination is negligible. The second situation arises if the production strategy at the origin is as described in Fig. 2.10. Then, we see from that figure that the total wait at the origin that can be attributed to the shipping strategy must be similar to that of the destination; i.e., it would also be proportional to the shaded area of Fig. 3.3. A third scenario arises with typical passenger transportation systems.

When holding costs are proportional to the area of Fig. 3.3 they are no longer a function of n alone. Newell (1971) points out that for a set of points (t_1 ... t_{n-1}) to be optimal, each line $\overline{PQ}$ (of Fig. 3.3) must be parallel to the tangent line to D(t) at the receiving time (point T in the figure). The reader can verify that if this condition is not satisfied, then it is possible to reduce the total shaded area by either advancing or delaying the receiving time by a small amount.

Unfortunately, the smallest shaded area – and thus the waiting cost – no longer can be expressed as a function of n alone, independently of D(t). Thus, it seems that a simple expression for the optimal cost cannot be obtained for any D(t) . (Subsection 3.3.4 develops an approximation when D(t) varies slowly with t).

3.3.3 Numerical Solution

There are different ways in which this problem can be solved numerically. For example, it can be formulated as a dynamic program in which a shipment time, t_i , is chosen at each stage (i = 1 , ..., n - 1) , and where the state of the system is the prior shipment time, t_{i-1} . The dynamic programming procedure yields an optimum holding cost for a given n , $z^*_i(n)$, which can be substituted for the first term of Eq. (3.6) to yield n^*.

The following procedure, based on Newell's property, is less laborious and works particularly well if D(t) is smooth, without bends or jumps (refer to Figure 3.4 for the explanation):

(i) Choose a point P_1 on the ordinates axis and move across to T_1,
(ii) Draw from P_1 a line parallel to the tangent to D(t) at T_1, and draw from T_1 a vertical line. Label the point of intersection P_2 .

Steps (i) and (ii) identify a point P_2 from a point P_1 . They should be repeated to identify P_3 from P_2, P_4 from P_3, etc. ..., defining in this manner a receiving step curve, R(t). If R(t) does not pass through the end point, (t_{max} , D(t_{max})), the position of P_1 should be perturbed until it does.

If a different point P_1 is chosen, a different number of steps may result, and the motion cost will change.[2] The holding cost for the given P_1 is proportional to the area between R(t) and D(t) ; it will also change if P_1 is moved. The overall optimum can be found by shifting the position of P_1 and comparing the sum of the holding and motion costs.

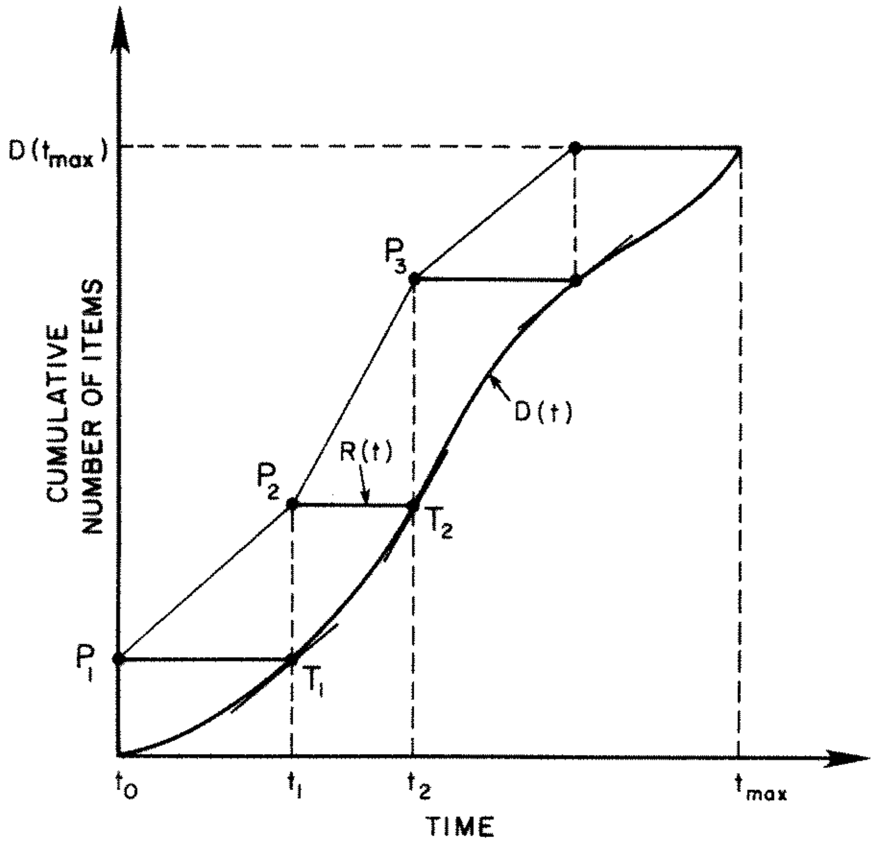

Fig. 3.4 Construction method for the cumulative number of items shipped versus time

3.3.4 The Continuous Approximation Method

The method about to be described, proposed by Newell (1971), replaces the search for $\{t_i\}$ by a search for a continuous function, whose knowledge

[2] As illustrated with problem 3.6, there may be more than one solution with the same number of steps.

yields a set of t_i with near minimal cost. It works well when D'(t) does not change rapidly; i.e., if $D'(t_i) \approx D'(t_{i+1})$ for all i . A by-product is a simple expression and decomposition principle for the total cost.

Let us assume that an optimal solution has been found, and denote by I_i the ith interval between consecutive receiving times: $[t_{i-1}, t_i)$, i = 1, 2, ... Then, divide the total cost during the study period into portions "$cost_i$" corresponding to each interval. That is, "$cost_i$" includes the cost, c_f , of dispatching one shipment plus the product of c_i and the shaded area for interval I_i:

$$cost_i = c_f + c_i(area_i).$$

Clearly, the sum of the prorated costs will equal the total cost. Since D'(t) is continuous, it should be intuitive that there is a point t'_i in each interval I_i for which the area above D(t) satisfies: $area_i = \frac{1}{2}(t_i - t_{i-1})^2 D'(t'_i)$. To see this informally, consider the triangle defined by the horizontal and vertical lines passing through a point P_i in the figure and a straight line passing through T_i with a slope that yields "$area_i$" for the triangle; i.e. slope $D'(t'_i)$. Since such a slanted line must intersect D(t) (otherwise the areas above D(t) and above the slanted line could not be equal) there must be a point between T_i and the point of intersection where the two lines have the same slope. The abscissa of this point is t'_i. Therefore we can write:

$$area_i = \frac{1}{2}(t_i - t_{i-1})^2 D'(t'_i) = \int_{t_{i-1}}^{t_i} \frac{1}{2}(t_i - t_{i-1}) D'(t'_i)\, dt. \qquad (3.7)$$

If we now define $H_s(t)$ as a step function such that $H_s(t) = t_i - t_{i-1}$ if $t \in I_i$ (see Figure 3.5 for an example), then the cost per interval can be expressed as:

$$cost_i = \int_{t_{i-1}}^{t_i} \left[\frac{c_f}{H_s(t)} + \frac{c_i H_s(t)}{2} D'(t'_i) \right] dt. \qquad (3.8)$$

Note that this is an exact expression.

If we now approximate $D'(t'_i)$ by D'(t) – which is reasonable if D'(t) varies slowly – the total cost over the whole study period can be expressed as the following integral:

$$cost \approx \int_0^{t_{max}} \left[\frac{c_f}{H_s(t)} + \frac{c_i H_s(t)}{2} D'(t) \right] dt. \qquad (3.9)$$

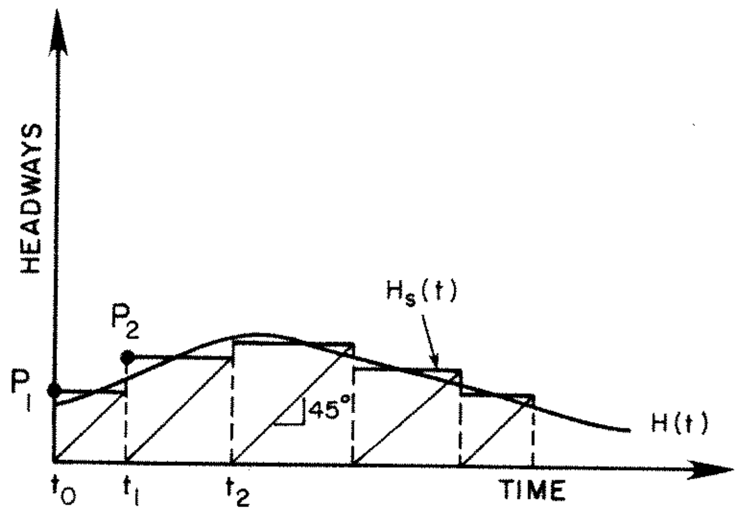

Fig. 3.5 Obtaining a set of dispatching times from H(t)

We seek the function $H_s(t)$, which minimizes (3.9). Unfortunately, this is akin to determining the $\{t_i\}$ themselves. A closed form solution can be obtained if in (3.9) $H_s(t)$ is replaced by a smooth function, $H(t)$, as shown in Fig. 3.5. That is:

$$cost \approx \int_0^{t_{max}} \left[\frac{c_f}{H(t)} + \frac{c_i H(t)}{2} D'(t) \right] dt . \tag{3.10}$$

Now, instead of finding $H_s(t)$, we can find the $H(t)$ which minimizes (3.10) – a much easier task – and then choose a set of shipment times (i.e., $H_s(t)$) consistent with $H(t)$.

Clearly, the $H(t)$ which minimizes (3.10) minimizes the integrand at every t ; thus:

$$H(t) = \left[2c_f / \left(c_i D'(t) \right) \right]^{1/2} . \tag{3.11a}$$

This is the time between dispatches (*headway*) for the EOQ problem with constant demand $D' = D'(t)$.

A set of shipment times consistent with H(t) can be found easily since H(t) varies slowly with t ; see (3.11a). Figure 3.5 suggests how this can be done systematically: Starting at the origin (point t_0) draw a 45° line and find a horizontal segment from a point on the vertical axis, such as P_1 in the figure, to the intersection with the 45° line. The elevation of P_1 should be such that the area below the segment equals the area below H(t). The abscissa of the point of intersection is the next shipment time, t_1 . This locates t_1 , given t_0 . The construction is then repeated from t_1 to locate t_2 , from t_2 , to locate t_3 , etc. In practice one does not need to be quite so precise, since we have already seen that small deviations from optimality have a minor effect.

Replacing the right side of (3.11a) for H(t) in integral (3.10) yields a simple expression for the optimal cost:

$$\textit{Total cost} \approx \int_0^{t_{\max}} \left[2c_i c_f D'(t)\right]^{1/2} dt . \tag{3.11b}$$

The integrand of this expression is the optimal EOQ cost per unit time if D' = D'(t).

Note that the integrand of Eq. (3.11b) can be written as:

$$\left[2c_i c_f / D'(t)\right]^{1/2} \left[D'(t)\, dt\right],$$

where the first factor represents the optimal cost *per item* for an EOQ problem with constant demand, D'(t) ; see Eq. (3.3). The average cost per item (across all the items) is obtained by dividing (3.11b) by the total number of items,

$$D(t_{\max}) = \int_0^{t_{\max}} D'(t)\, dt$$

The result is:

$$\left(\frac{cost^*}{item}\right) \approx \frac{\int_0^{t_{\max}} \left[2c_i c_f / D'(t)\right]^{1/2} D'(t)\, dt}{\int_0^{t_{\max}} D'(t)\, dt} . \tag{3.11c}$$

In practical terms this equation indicates that the average optimal cost per item can be obtained by *averaging the cost of all the items, as if each one of these was given by the EOQ formula with a (constant) demand rate equal to the demand rate at the time when the item is consumed.*

Equation (3.11b) has a similar (decomposition) interpretation: the expression indicates that, given a partition of $[0, t_{max}]$ into a collection of short time intervals, the optimum cost can be approximated by the sum of the EOQ costs for each one of the intervals considered isolated from the others.

Equations (3.11) are so simple that they can be used as building blocks for the study of more complex problems as we shall see in later chapters. This is one of the attractive features of the CA approach; it yields cost estimates without having to develop, or even define, a detailed solution to the problem.

The CA approach can also be used to locate points on any line (time or otherwise) *provided that the total cost can be prorated approximately to (short) intervals on the line, while ensuring that the prorated cost to any interval only depends on the characteristics of said interval.* In the previous discussion, the integrand of (3.10) is the prorated cost in $[t, t+dt)$, which does not depend on the demand rate outside the interval.

The CA approach can also be used to locate points in multidimensional space, when the total cost can be expressed as a sum of neighborhood costs dependent only on their local characteristics. Newell (1973) argues that the CA approach is comparatively more useful then, because in the multidimensional case it is much more difficult for exact numerical methods to deal with the complex boundary conditions that arise. Because the CA approach will be used in forthcoming chapters repeatedly, the next section discusses two additional (one-dimensional) examples.

3.4 Other One-Dimensional Location Problems

The CA technique was originally proposed to find a near-optimal bus departure schedule from a depot (Newell, 1971). Given the cumulative number of people $D(t)$ demanding service by time t, the fixed cost of a bus dispatch c_f, and the cost of each person-hour waited c_i, the objective was to minimize the sum of the bus dispatch (motion) and waiting (holding) costs. With an unlimited bus capacity, this problem is almost identical to the one we have just solved; except for $D(t)$, which now represents the cumulative number of people (items) *entering* the system and not the number leaving. Equations (3.11), however, still hold (see problem 3.2) This should be intuitive. Although the graphical construction of Figure 3.4 is now slightly different (i.e., the sought passenger departure curve R(t) now touches D(t) from below) consideration shows that the new and old figures become qualitatively identical if one of them is rotated 180 degrees. Since such a rotation cannot change the mathematical relationships between the elements of the figure, it shouldn't be surprising that Eqs. (3.11) remain valid.

The second example locates freight terminals on a distance line between 0 and d_{max}. This interval contains origins, which send items to a depot.

The distance line extends from the origin, O , to a depot, located at $d = \tilde{d} \geq d_{max}$. The flow of freight (number of items per day) that originates between O and d is a function of d , D(d), which increases from 0 to v_{tot} (see Figure 3.6). Items are individually carried to the terminals at a cost c_d' per unit distance per item. Each day a vehicle travels the route collecting the items accumulated at each terminal and takes them to the depot.

The motion cost for this operation has three components: the handling cost at the terminals, assumed to be constant and therefore ignored, the access cost to the terminals, and the line-haul cost of operating the vehicle from the terminals to the depot. The access cost is given by the product of c'_d and the total item-miles of access traveled per day; it increases with the separation between stops as will be explained in a moment. The line-haul cost has the form of Eq. (2.5d):

$$\begin{pmatrix} \textit{line-haul} \\ \textit{cost/day} \end{pmatrix} = c_s(1 + n_s) + c_d(\tilde{d}) + c'_s(v_{tot}),$$

where n_s is the number of stops (excluding the depot) and v_{tot} is the total size of the shipment arriving at the depot. Note that the line-haul cost does not depend on the specific stop locations and that in contrast to the access cost, it increases with n_s. As a function of n_s we express it as:

$$\begin{pmatrix} \textit{line-haul} \\ \textit{cost/day} \end{pmatrix} = c^o + c_s n_s , \tag{3.12}$$

where c^o is a constant that will be ignored for design purposes.

As the problem has been formulated, with one trip per day, the sum of the holding costs at all stops can be ignored – consideration reveals that the sum is constant. Pipeline inventory costs do depend on the decision variables (they should increase with n_s) but for cheap freight the effect is negligible relative to (3.12). Thus, all inventory and holding costs are neglected. The stops will be located as the result of a trade-off between line-haul and access costs. Without this simplification, which is inappropriate for passenger transportation, the problem is equivalent to the transit stop location problem solved by Vuchic and Newell (1968) with dynamic programming, and later by Hurdle (1973), and Wirasinghe and Ghoneim (1981) with the CA method. (See problem 3.3).

Figure 3.6 depicts the location of three terminals (at points d_1, d_2, and d_3) and a curve, R(d), depicting the number of items in the vehicle as a function of its position. This curve increases in steps at each terminal location. The size of each step equals the number of items collected. To minimize access (and total) cost each item is routed to the nearest terminal, and as a result the step curve passes through the midpoints, M_i , shown in the figure. (The coordinates of M_i are $m_i = (d_i + d_{i+1})/2$ and $D(m_i)$; with $m_0 = 0$ and $m_{ns} = d_{max}$).

Let us see how the total cost can be prorated to short intervals, by considering the partition of $(0, d_{max}]$ into the following intervals surrounding each terminal: $I_1 = (0, m_1]$, $I_2 = (m_1, m_2]$, ... , $I_{ns} = (m_{ns-1} ;d_{max}]$. Each interval, I_i , adds an access cost proportional to the daily item-miles traveled for access to terminal i . This is given by the shaded area on the two quasi-triangular segments next to the location of the terminal, $(area)_i$, thus:

$$access\ cost_i = (area)_i\, c'_d \,.$$

For slowly varying D(d), the access cost can be rewritten as:

$$access\ cost_i \approx \frac{1}{4}(m_i - m_{i-1})^2\, D'(d_i)\, c'_d \,.$$

Since each terminal adds c_s to the daily line-haul cost (see Eq. (3.12)), the share of the total cost prorated to I_i is:

$$\begin{pmatrix} Total\ cost \\ per\ day \end{pmatrix}_i \approx c_s + \frac{c'_d}{4}(m_i - m_{i-1})^2\, D'(d_i).$$

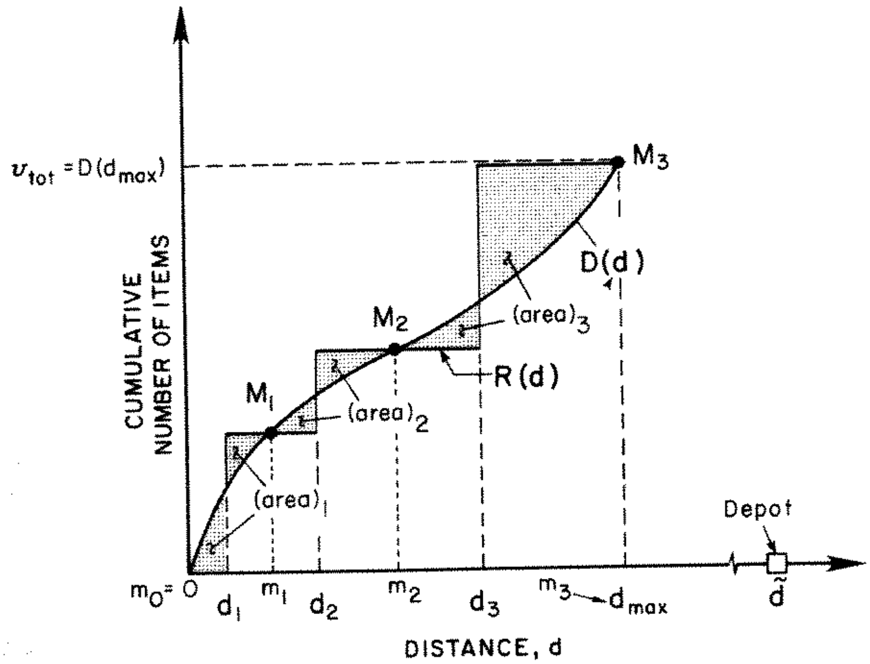

Fig. 3.6 Geometrical construction for a terminal location problem

Since D'(d) ≈ D'(d_i) for d∈I_i (we stated that D'(d) varied slowly), the above expression can be approximated by:

$$\begin{pmatrix} Total\ cost \\ per\ day \end{pmatrix}_i \approx \int_{m_{i-1}}^{m_i} \left\{ \frac{c_s}{(m_i - m_{i-1})} + \frac{c'_d}{4}(m_i - m_{i-1})D'(d) \right\} dd\,.$$

If we now let s(d) denote a slowly varying function such that $s(d_i) = m_i - m_{i-1}$ (the function, used later to locate the terminals, indicates the size of a terminal's influence area depending on location), then we can re-write the last expression once again, using s(d) instead of $m_i - m_{i-1}$:

$$\begin{pmatrix} Total\ cost \\ per\ day \end{pmatrix}_i \approx \int_{m_{i-1}}^{m_i} \left\{ \frac{c_s}{s(d)} + \frac{c'_d}{4} s(d)\, D'(d) \right\} dd\,.$$

The total cost for the system is then:

$$\begin{pmatrix}\text{Total cost}\\ \text{per day}\end{pmatrix} \approx \int_0^{d_{max}} \left\{ \frac{c_s}{s(d)} + \frac{c'_d}{4}\, s(d)\, D'(d) \right\} dd \,. \qquad (3.13)$$

As with Eqs. (3.11), the least cost s(d) minimizes the integrand at every point; given its EOQ analytical form, we find:

$$s(d) \approx 2\left[\, c_s \;/\; \left(c'_d\, D'(d)\right)\right]^{1/2}. \qquad (3.14a)$$

(Note that if D' varies slowly, s(d) will vary slowly as we had assumed.)

The expressions for the minimum total and average (per item) cost are similar to (3.11b) and (3.11c); the partition/decomposition principle still holds.

$$\begin{pmatrix}\text{Total cost}\\ \text{per unit}\\ \text{of time}\end{pmatrix}^{*} \approx \int_0^{d_{max}} \left[c_s c'_d \;\; D'(d)\,\right]^{1/2} dd \;, \qquad (3.14b)$$

$$Cost^{*}/item \approx \int_0^{d_{max}} \left[c_s c'_d \,/\, D'(d)\right]^{1/2} D'(d)\, dd \;\Big/ \int_0^{d_{max}} D'(d)\, dd \,. \qquad (3.14c)$$

To locate the terminals, one first divides $(0, d_{max}]$ into non-overlapping intervals of approximately correct, length I_1, I_2, etc. ..., by starting at one end and using (3.14a) repeatedly. If the last interval is not of correct length, then the difference can be absorbed by small changes to the other intervals. If d_{max} is large (so that there are at least several intervals), then the final partition should satisfy $s(d) \approx m_i - m_{i-1}$ if $d \in I_i$, and the approximations leading to (3.14) should be valid. With the influence areas defined in this manner, the terminals are located next. They should be positioned within each interval so that the boundary between neighboring intervals is equidistant from the terminals. For a general sequence of intervals (e.g., of rapidly fluctuating lengths) this may be difficult (even impossible) to do, but for our problem with $|I_i| \approx |I_{i+1}|$ the best locations should be near the center of each interval; in fact little is lost by locating the terminals at the centers.

3.5 Accuracy of the CA Expression

Although a systematic analysis of its errors has not been reported, experience indicates that the CA approach is very accurate when the descriptive characteristics of the problem (D'(t) in the text's examples) vary slowly as assumed. Also quite robust, the approach is effective even if the variation in conditions is fairly rapid – in our case, accurate results are obtained even if D'(t) varies by a factor of two within the influence areas. Perhaps this should not be surprising, in light of the EOQ robustness discussed in Section 3.2.

When conditions are unfavorable, the CA method can both over- and underpredict the optimal cost. The following two examples identify said conditions, with the first example illustrating over-estimation and the second underestimation. The basis for comparison will be the exact solution, which for our problem can be obtained readily, as described below.

3.5.1 *An Exact Procedure and Two Examples*

A construction similar to that in Fig. 3.4 can also be used for the terminal location problem.

Note first that, given n_s , for a set of locations to be optimal the line D(d) of Fig. 3.6 must bisect in two equal halves every vertical segment of R(d) . Otherwise, the terminal (e.g., terminal 3 of Fig. 3.6) could be moved slightly to decrease access cost. The optimal solution can then be found by comparing all the possible R(t) with the above property.

For a given d_1 , draw a vertical step that is bisected by D(d), and move across horizontally so that the horizontal segment is also bisected by D(t) . This identifies d_2 . Repeat the construction to find d_3, d_4, etc. (Only those values of d_1 for which the last vertical segment is bisected by D(t) need to be considered seriously.) The optimal solution corresponds to a d_1 which minimizes the sum of the stop cost and access cost. The procedure is so simple that it can be implemented in spreadsheet form. (The user selects d_1 and the spreadsheet returns the graphs, and the cost; it is then easy to find the solution either interactively or automatically with the computer.) The examples can now be discussed.

Example 1:

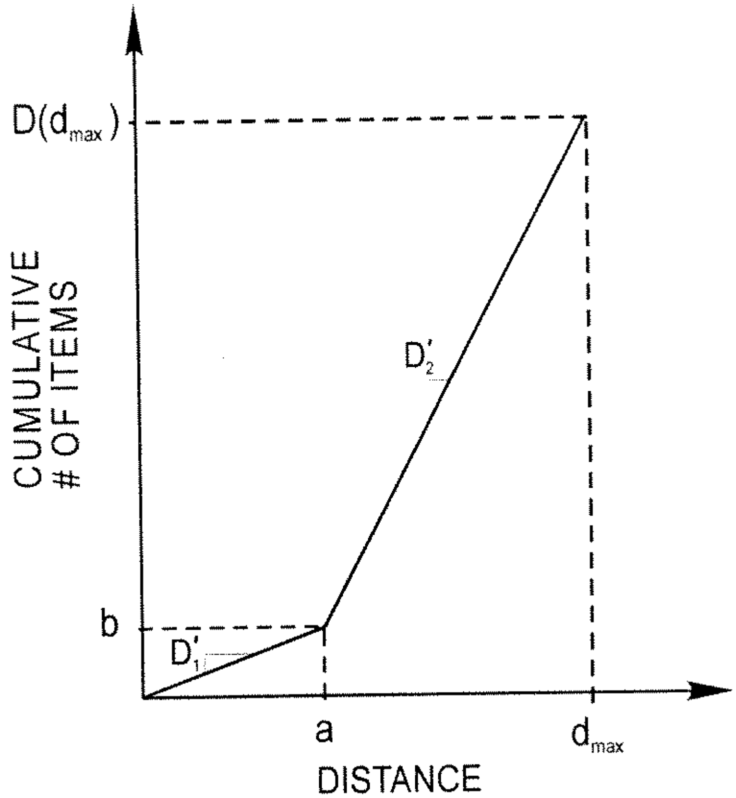

Fig. 3.7 Cumulative demand versus distance for example 1

Terminals are to be located on two adjoining regions with high and low demand. Figure 3.7 depicts a generic piece-wise linear cumulative demand curve of this type. The coordinates of the break-point (distance, item number) are given by parameters "a" and "b". They, of course must be consistent with the specified values for d_{max} , $D(d_{max})$, D'_1 and D'_2 . For this problem the continuum approximation approach yields – see Eq. (3.14b):

$$Total\ cost^* \approx (c_s c'_d)^{1/2} \left\{ a\sqrt{D'_1} + (d_{max} - a)\sqrt{D'_2} \right\}$$

A possible set of parameters is $d_{max} = 500$, $D(d_{max}) = 1700$, $D'_1 = 1$, $D'_2 = 5$, $a = b = 200$, $c'_d = 1$ and $c_s = 160{,}000$. This choice has been made because a systematic analysis shows that it produces the largest overprediction error in percentage terms. The predicted cost is:

$$\textit{Total cost}^* \approx 348{,}328\,.$$

In actuality the least possible cost is 8% smaller. It arises when a single terminal is located at d = 330. The reader can verify that the exact access cost for this location is 160,500 units. Since the terminal cost is 160,000 units (for one terminal), the grand total is 320,500 < 348,328.

This rather extreme example illustrates that the CA approach can overestimate the optimum cost. To understand why this happens let us decompose the CA costs into its components. Note first that the ideal spacing between terminals predicted by the CA method with (3.14a) is:

$s(d) = 800$ in the low demand section, and
$s(d) = 357$ in the high demand section.

Thus, the CA access cost is calculated as if the average access distance was $s(d)/4 = 200$ in the low demand section and 89.25 units in the high demand section. Since there are 200 items in the low density region and 1500 in the high density region, the total CA access cost is approximately: $200{\times}200 + 89.25{\times}1500 \approx 173{,}875$. The CA stop cost is calculated by integrating the density of terminals over the service region, $(200/800 + 300/357) \approx 1.09$, and multiplying this result by the cost of a terminal: $1.09{\times}160{,}000 = 174{,}400$. The grand total is therefore: $173{,}875 + 174{,}400 = 348{,}275 \approx 348{,}328$.

It turns out, however, that just a single terminal in the high density region can serve both, the low density points with an average distance barely greater than the CA access distance, and the high demand section with an average access distance considerably inferior to the corresponding CA distance. For our chosen location (d = 330) the actual average access distances are: 230 units for the low density section (200 with the CA method) and 76 for the high density section (89 with CA method). Since we are using only one terminal, the final cost is lower.

The overprediction effect arises because the demand curve varies significantly and very favorably between the terminal and the edge of the service region, and the CA approach does not exploit this variation. The variation is so favorable that it allows a terminal provided for the high density points to double up efficiently as a terminal for the low density points.

Favorable conditions are unusual, however. When the demand does not vary rapidly the CA approach consistently underestimates demand.

Example 2: An example where the CA approach underestimates cost is easy to construct. By its nature, the CA approach ignores that the number of terminals must be an integer; any situation with a finite region size (or time horizon) will exhibit this error type. To exclude the overprediction error type illustrated by example 1, the demand per unit length of region is set constant: $D'(d) = D'$. This also allows closed form comparisons to be made.

The CA solution (3.14b) is:

$$Total\ cost^* = \sqrt{c_s c'_d}\ \sqrt{D'} d_{\max}$$

Without losing generality, we choose the units of distance, item quantity and money so that $d_{max} = 1$, $D(d_{max}) = 1$ and $c_s = 1$. Thus, $D' = 1$ and only the parameter c'_d remains. The above expression becomes:

$$Total\ cost^* = \sqrt{c'_d}\ . \tag{3.15a}$$

If the exact optimal solution has n_s terminals, the distance line will be partitioned into n_s intervals of equal length: $I_i = ((i - 1)/n_s, i/n_s]$. The total cost is then:

$$Total\ cost\ (n_s) = n_s + 2n_s \left\{ \left(\frac{1}{2\ n_s} \right)^2 \frac{c'_d}{2} \right\} = n_s + \frac{c'_d}{4n_s} \tag{3.15b}$$

which is an EOQ expression in n_s . Its minimum over $n_s = 1, 2, 3, \ldots$ is the optimal cost.

This least cost will always be greater or equal to the right side of Eq. (3.15a) because (3.15a) is the minimum of (3.15b) with unrestricted n_s , obtained for $n^*_s = (c'_d/4)^{1/2}$. Clearly, the underprediction will be most significant when n^*_s is close to an odd multiple of 0.5, or close to zero. Equation (3.4), which described the sensitivity of the EOQ cost expression to errors in the decision variables, also quantifies this underprediction; as n^*_s increases the underprediction quickly vanishes. Once $c'_d > 16$ (n^*_s is greater than 2) the difference is below one percent. If $c'_d > 4$ (the value at which $n^*_s = 1$) then the maximum difference stays below six percent. Although for smaller c'_d the difference can grow arbitrarily large as $c'_d \to 0$, that is not the case that is likely to be of interest; the large spacing between terminals recommended by the CA method (much larger than d_{max}) indicates that operating line-haul vehicles is probably an overkill. If it were of interest, and a terminal had to be provided, one could force the solution to the CA approach to satisfy the constraint $n_s \geq 1$. The next section will discuss how more involved constraints can be accommodated within a general CA framework.

Although exhibiting different errors types, both examples shared a common trait when their errors were largest: the ideal terminal spacing in an interval with constant demand exceeded the length of the interval; i.e., demand varied significantly within the spacing. Errors arose because this property violates the stated requirement for the CA approach: D'(d) should vary slowly over distances comparable with s(d). Conversely, the numerical results prove that an error below one percent results if D(d) is piecewise linear with segments at least three times as long as each s(d). Thus, any demand function that can be approximated in this manner should also yield accurate results.

3.6 Generalization of the CA Approach

The CA method can be applied to more complex problems – even problems that defy exact numerical solution. In forthcoming chapters it will be used to locate points in multidimensional (time-space) domains while satisfying decision variable constraints.

All that is needed is that the input data vary slowly with position, either in one or multiple dimensions, that the total cost can be expressed as a sum of costs over non-overlapping (small) regions of the location domain, and that these component costs (and constraints) depend only on the decisions made in their regions. If this is true, the decomposition principle holds and the CA results approximate the optimal cost accurately.

As a one-dimensional illustration, let us return to the inventory control problem of Eqs. (3.7) to (3.11), and let us assume that there is a capacity constraint on shipment size:

$$D(t_i) - D(t_{i-1}) \leq v_{max} .$$

This constraint has a local nature because it only involves quantities determined by events close to the time of shipment; i.e., by two neighboring dispatching times and by the amount of consumption between them. For any time t, thus, it should be possible to write the constraint approximately as an inequality including only variables and data specific to time t.

Recalling the definition of $H_s(t)$ (see Fig. 3.6), and using the slow-varying property of $D'(t)$, we can write:

$$D(t_i) - D(t_{i-1}) \approx H_s(t)D'(t) \approx H(t)D'(t)$$

and the constraint can be replaced by the approximation *based only on conditions at t:*

$$H(t)D'(t) \leq v_{max}, \quad or \quad H(t) \leq v_{max}/D'(t),$$

which must be satisfied for all t.

An approximate solution to our problem, thus, is an $H(t)$ that minimizes (3.10) subject to this constraint. The solution is of the form indicated by Eqs. (3.5); i.e., the optimal $H(t)$ is the least of: (i) the right side of (3.11a), $(2c_f/c_i\ D'(t))^{1/2}$, and (ii) $v_{max}/D'(t)$. Letting $\Psi\{x\}$ denote the increasing concave function $\{x^{1/2}$ if $x \leq 1$; or $\frac{1}{2}[1 + x]$ if $x > 1\}$, we can express the minimum cost per unit time concisely in terms of the dimensionless quantity, $2c_fD'(t)/(c_i\ v_{max}^2)$:

$$c_i v_{max} \psi\left\{ 2c_f D'(t)/c_i v_{max}^2 \right\}.$$

Integrated from 0 to t_{max} , this expression approximates the optimal total cost, as in Eq. (3.11b). Note that when the argument of Ψ is less than one, as would happen if v_{max} is very large, then the expression coincides with the integrand of (3.11b), $[2c_ic_fD'(t)]^{1/2}$. An average cost per item can also be obtained as in Eq. (3.11c); its interpretation as a cost average across items (calculated as if each item was part of a problem with constant conditions, equal to the local conditions for the item) is still valid.

In practical cases, a per-item cost estimate can be obtained easily with the following two-step procedure:

(i) Solve the problem with constant conditions for a representative sample of items and input data,

(ii) Average the solution across all the sampled items to obtain the result.

Note that the cost estimate can be obtained even without defining the decision variables in step (i). Problem 3.5 illustrates the accuracy of the CA method under capacity constraints.

3.6.1 *Practical Considerations*

While for simple problems, such as the one solved above, the solution can be easily automated, more complex situations may benefit from decision support tools with substantial human intervention. The following two-step human/machine procedure is recommended: (i) first, recognizing that its recommendations may need fine-tuning adjustments, the CA (or other simplified) method is applied to a basic version of the problem without secondary details; (ii) then, trained humans develop implementable solutions that account for the details, perhaps aided by numerical methods that can benefit from the output of the first step.

In some cases, when time is of the essence humans alone may have to carry out this second step because efficient numerical methods capturing peculiar details may not be readily available, and developing them may be prohibitively time consuming. Furthermore, even without time pressures, if the details are so complex (or so vaguely understood) that they cannot be quantified properly, pursuing automation for the fine-tuning step would seem ill-advised. Fortunately, this is not a serious drawback; as argued earlier, significant departures from ideal situations should not increase cost significantly, leaving humans considerable latitude for accommodating details.

As an illustration of these concepts, problem 3.6 re-examines the terminal location problem of Section 3.4. when only 50 specific locations are feasible. The cost of the two-step procedure (fine-tuned by hand) is compared to the ideal cost without restrictions, and (optionally) to the exact optimal cost obtained with dynamic programming. The reader will find that the fine-tuning step often identifies the exact optimum, and when it does not, the difference between the two-step and the exact optimal costs is measured by a fraction of a percentage point. Furthermore, the two-step and one-step (or ideal) costs are very close; of course, provided that n^*_s is not greater than 50.

3.7 Network Design Issues

In all the scenarios discussed so far, the items followed a predetermined path. Real logistics problems, however, often involve the choice of alternative routes (e.g., alternative ways of shipping) between origins and destinations, in addition to the choice of when and how much to dispatch. In some instances one may even be interested in whether certain routes should be provided at all; or even in the design of an entirely new physical distribution network.

We also found in Section 3.1 that there were economies of scale in flow; i.e., the optimal cost per item decreased with D'. Later in this monograph we will have to consider logistics problems with multiple destinations, where an item's route is not predetermined and cost decreases with flow. We discuss here some key features of these problems, and conclude the chapter with a comparison of detailed and non-detailed approaches for logistic system design.

3.7.1 The Effect of Flow Scale Economies on Route Choice

A simple example with one origin and two destinations (see Figure 3.8) effectively illustrates the properties of optimal system designs with and without flow economies of scale. The origin, O , produces items of type i (i = 1, 2) for destination P_i at a constant rate, given by the parenthetical numbers in the figure: $D'_1 = D'_2 = 4$ items per unit time. The combined production rate at the origin is $D'_1 + D'_2 = 8$ items/unit time. The arrows in the figure depict possible shipment trips; these transportation links are numbered 1, 2, 3. While all the items traveling to P_1 , must travel directly between O and P_1 , the items traveling to P_2 may go either directly or via P_1 .

Let us assume that a fraction (to be decided) x , of the items for P_2 are sent via P_1 and the rest are shipped directly. This establishes a flow $x_1 = 4(1+x)$ on link 1 (OP_1), a flow $x_3 = 4x$ on link 3 (P_1P_2) and a flow $x_2 = 4(1 - x)$ on link 2 (OP_2).

We also assume that the total cost on the network can be expressed as a sum of link costs, and that these depend only on their own flows. This is a reasonable assumption if no attempt is made to coordinate the shipping schedules on the three links, as then the prorated cost to each link should be close to the EOQ expression with demand rate equal to the link flow. Thus, if we let $z_i(x_i)$ denote the cost per item on link i when the flow is x_i, the total system cost per unit time is:

With economies of scale, the functions $x_iz_i(x_i)$ increase at a decreasing rate (are concave) as in

$$Total\ cost = \sum_{i=1}^{3} x_i z_i(x_i). \tag{3.16}$$

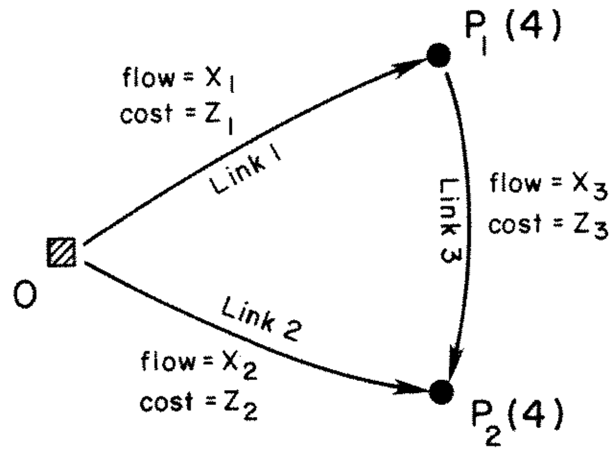

Fig. 3.8 Flows and costs for a simple 3-node network

With economies of scale, the functions $x_i z_i(x_i)$ increase at a decreasing rate (are concave) as in Figure 3.2c. Because the x_i's are linear in the split x , the total cost is a concave function of the split – this (concave) dependence of cost on splits (decision variables) also holds for general networks.

Suppose, for example, that

$$z_1 = x_1^{-1/2}, \quad z_2 = 3x_2^{-1/2} \ and \ \ z_3 = 1\,;$$
$$x_1 z_1 = x_1^{1/2}\ , x_2 z_2 \ = 3x_2^{1/2} \ \ and\, x_3 z_3 = x_3 .$$

Then, as a function of x , (3.16) becomes:

$$Total\ cost = 2(1+x)^{1/2} + 6\ (1-x)^{1/2} + 4x. \qquad (3.17)$$

This relationship is plotted on Fig. 3.9; as stated, the total cost is a concave function of the split, x . Like any concave function, it reaches a minimum at one of the ends of the feasibility interval. For our data the optimal solution is $x^* = 1$, indicating that everything should be shipped through P_1. The total cost is 6.8. Although shipping everything direct may be better for different data, clearly one would never want to split the flow to P_2 among the two routes (OP_2 and OP_1P_2) .

A similar "all-or-nothing" principle holds for networks with multiple origins and destinations if the total cost is a concave function of all the link flows (Zangwill, 1968). In that case all the flow from any origin to any destination should be allocated to only one route. This is not difficult to see: one can define a split between any two routes joining an origin and a destination, and since the link flows are linear in that split, the total cost is concave in the split; thus, only one of the routes can carry flow. Networks with diseconomies of scale behave in an opposite manner. In that case the total cost function is convex in the splits and there is an incentive to spread out the flow among routes. In fact, if for a one origin and one destination network, there exist several routes with identical cost functions (with diseconomies); it is not difficult to prove that the total flow should be evenly divided among *all* the routes.

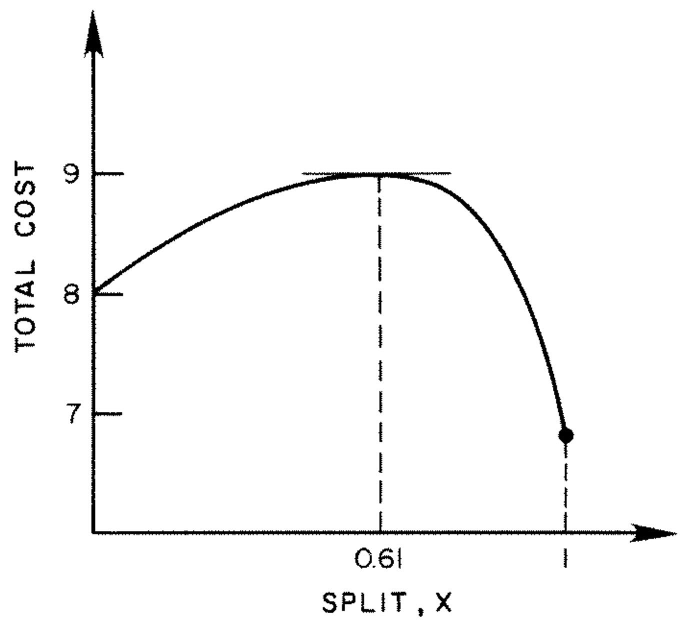

Fig. 3.9 Concave cost function in the split

Networks with flow economies of scale also respond in a different manner to changes in conditions. While, with diseconomies, a small improvement to one of the routes would lead to a small change in the optimal flow distribution (see exercise 3.8), with economies, the optimal flows either stay the same or change by a discrete amount. This can be seen with the example of (3.17). As long as $z_3 < [2 - 2^{-1/2}] \approx 1.3$, x^* equals 1, but if z_3 is increased beyond this value ever so slightly, the solution jumps to $x^* = 0$. This is typical of concave cost problems: minor changes to the input data can induce large changes in the optimal solution. Fortunately, the cost does not behave in such manner; despite the jump in our example the cost is a continuous function of z_3:

$$\text{Total cost}^* = 8 \qquad \text{if } z_3 \leq 2 - 2^{-1/2}$$
$$= 8^{1/2} + 4z_3 \qquad \text{if } z_3 > 2 - 2^{-1/2}$$

Let us now turn our attention to solution methods.

3.7.2 Solution Methods

The nature of the solution is not the only difference between networks with economies and diseconomies; the way to find it is also different. While networks with diseconomies are well behaved optimization problems without local minima that are not global, networks with economies are not. The books by Steenbrink (1974), Newell (1980), and Sheffi (1985) discuss networks with scale diseconomies in detail; Popken (1988) reviews the sparser (traditional/detailed) network design literature for networks with scale economies. Further information on this subject can be found in Ball et al (1995 and 1995a).

Although local search algorithms can be used to find near optimal solutions for large detailed networks with convex costs, the same procedures fail with concave cost networks. The task is then much more complicated, and the network sizes that can be handled by numerical methods much smaller.

Except for technical details, all local search algorithms work in the same manner. First, the total cost is evaluated for an initial feasible solution, described by a set of variables that uniquely identify the decisions; e.g., the set of splits for all origin destination pairs. A small cost-reducing perturbation to the feasible solution (e.g., a differential change to the splits) is then sought. If not found, the search stops because the initial solution is a local minimum; i.e., a solution that cannot be improved without substantial changes. Otherwise, an improved larger perturbation obtained from the original small perturbations is identified, and then used to construct a new improved feasible solution. The process is then iterated (seeking small cost-reducing perturbations to the new solution, etc.) until no significant improvements result.

Local search techniques work acceptably for networks with scale diseconomies, because in those instances any local minimum is a global minimum. Unfortunately, this is not the case with economies of scale. Figure 3.9 reveals that our simple problem has two local minima: $x = 0$ and $x = 1$. If a local search algorithm is applied to our example, any starting solution with $x < 0.61$ (the maximum in the figure) will converge suboptimally to $x = 0$. While for our simple example this can be corrected simply by starting with different x's , the task is daunting for large, highly detailed networks. In that case, the number of potential traps for a local search – all local minima regardless of cost – increases exponentially with the amount of detail.

This is illustrated with an example, where items from a large number N of origins are shipped to one destination using two transportation modes (1 and 2). We use x_i to denote the split of production from origin i sent on mode 1, and assume that (to satisfy an agreement with the providers of type-1 transportation) each x_i must satisfy $x_i \geq h_i$ for some constant $h_i \geq 0$. Transportation by mode 2 is assumed to be more attractive, but limited in capacity; that is, the sum of the x_i's must exceed a value h.

For a set of splits to be feasible, thus, the following must be true:

$$\sum_{i=1}^{N} x_i \geq h, \text{and } h_i \leq x_i \leq 1, \forall i \tag{3.18a}$$

We seek the set of feasible splits that minimize the total cost, or equivalently the penalty paid because not all the items can be shipped by mode 2. The penalty paid for each origin is assumed to increase with x_i , except at certain values where a fixed amount ε_i is saved – perhaps because shipments can then be multiples of a box, requiring less handling. To simplify the exposition, let us assume that there is only one such value δ_i for every origin, and that away from this value the penalty equals x_i; otherwise the penalty is x_i - ε_i. If we define $\varepsilon_i(x_i)$ to be: ε_i if $x_i = \delta_i$ and 0 otherwise, then the combined penalty for all the origins can be expressed as:

$$\sum_{i=1}^{N} \left[x_i - \varepsilon_i(x_i)\right]. \tag{3.18b}$$

Note that each one of the terms in this summation for which $\delta_i > h_i$ exhibits two local minima in the range of feasibility $[h_i,1]$: $x_i = h_i$ and $x_i = \delta_i$.

Any combination of x's, each equaling either h_i or δ_i , and satisfying (3.18a) is a local minimum, which could stop a search. If the δ_i and the h_i are uniformly distributed between 0 and 1, and h is small, there will be $O(2^{N/2})$ local optima. With so many traps, local search algorithms are doomed to failure for this problem – not because (3.18b) is discontinuous, but because it is not convex. A different method must be used.

Certainly, one could search exhaustively over all the possible solutions with a combinatorial tool such as branch and bound, but these methods can only handle problems of small size – typically with $O(10^2)$ decision variables or less.

Alternatively, one could try to exploit the peculiar *mathematical structure* of Eqs. (3.18) – or whichever problem is at hand – to develop a suitable algorithm. If successful, the approach would find a solution with all its detail. In our case, the optimization of (3.18) can be reduced to a knapsack problem that can be solved easily (see exercise 3.8); in other instances it may be possible to decompose the problem into a collection of small easy problems. Very often, however, a simple solution method cannot be found. In our case, this would happen if there were more than one (ε_i, δ_i) for each origin. Traditionally one then turns to ad hoc intuitive solution methods (known as heuristics) which one *hopes* will yield reasonable solutions.

There is also another approach. If while inspecting the formulation, or even better in the process of formulating the problem, one realizes that certain details are of little importance one should leave them out. Our example illustrates how removing minor details can turn a nightmare into an easy problem. If the ε_i's are so small that the $\varepsilon_i(x_i)$ in (3.18b) can be neglected, then the objective function reduces to $\Sigma_i x_i$. Former sources of difficulty, the δ_i and ε_i no longer enter the formulation. With less detail, the problem becomes well behaved (convex), and even admits a closed form solution; e.g., if $\Sigma_i h_i \geq h$ then the optimal splits are $x_i = h_i$ and the total cost is $\Sigma_i h_i = N\bar{h}$.

Note that the optimal cost is given by an average (there is no need to know precisely each individual h_i in order to estimate the optimal cost), and that the optimal solution can be described with the simple rule "make every split as small as possible", which can be stated without making reference to the h_i's.

In the rest of this monograph we will seek solutions to logistic problems using as little detail as possible, describing (as in the example) the solution in terms of guidelines which are developed based on broad averages instead of detailed data. We recognize that the solutions obtained from such guidelines may benefit from fine-tuning once detailed data become available; but also note that incorporating all the details into the model early will increase the effort for gathering data, and, as illustrated, may even get in the way of obtaining a good solution.

Observation of mother nature's logistics networks suggests that many logistics systems can be designed in this manner. Trees can be viewed as a logistic system for carrying nutrients from the soil to an above-ground region (the leaves) to meet the sun's rays. While every individual tree of a species is distinct from other individuals, we also see that the members of a species share many common characteristics *on average*. There is order at the macroscopic level. This is not surprising, since members of the species have adapted to similar environmental conditions, also filling the same niche in the eco-system. The detailed characteristics of an individual tree are (like our logistic systems) developed from two levels of data in two different ways:

(i) Members of the same species share a genetic code, which has evolved in response to the typical or average conditions that can be expected. This code is analogous to the guidelines of a simple model; e.g., "make each split as small as possible."

(ii) In response to the detailed conditions of its location, a tree develops an individuality within the guidelines of the genetic code, better to exploit the local conditions. This would be analogous to the fine tuning that could have taken place if the ε_i, h_i, and δ_i had been given in our example.

The same could be said for other logistic systems encountered in nature, such as the circulating and nervous systems of the human body.

On further inspection we notice that, not only average characteristics, but some *specific* traits are also the same for all individuals, (e.g., some tree species have always one trunk, all humans have one aorta artery, etc.). It is as if nature had decided that these items of commonality are optimal for almost any conditions that can be encountered; therefore, that part of the design is not open to fine tuning. Perhaps the same can be said of logistics systems.

The logistics systems of nature also have economies of scale. It takes less energy to move a certain flow through one single pipe than through two pipes with one-half the cross section. As in our networks with concave costs, there is an incentive to *consolidate* flow into single routes that can handle great volumes efficiently. Nature has responded to this challenge by evolving hierarchical systems of conveyance, such as the three hierarchy network of Fig. 3.10.

Scientists have begun to realize that apparently very complex ("fractal") structures, such as a fern leaf, can be replicated and/or described with just a few rules and parameters (Gleick, 1988, provides an entertaining description of these ideas). For the example of Fig. 3.10 the separations between "nodes" (e.g., A_1 and A_2) for each hierarchy might be found to be relatively constant, perhaps varying with the distance from the root, as might be the number of branches at every node and the relative size of the main and secondary branches at nodes of the same hierarchy. The latter may also vary with the distance from the "root."

A physical distribution network should probably be organized in a similar way with the root becoming the depot, the leaves the customers, and the nodes intermediate transshipment centers or terminals. Physical distribution networks that serve similar purposes, just as in nature, should likely share the same hierarchical organization and overall traits even if the specific details differ. As in nature, it should be possible to describe their near optimal configuration with just a few simple rules and parameters (see problem 3.11).

In this spirit, the chapters that follow will try to get at the "genetic code" of logistics systems; i.e., describe how general classes of logistics networks should be organized, with guidelines for obtaining an optimal structure developed without using detailed data. Building on the simple EOQ model, we gradually consider more complex systems.

Chapter 4 describes problems with a single hierarchy consisting of one origin and many destinations (or the reverse); i.e., "one-to-many" problems. Chapter 5 describes "one-to-many" problems with transshipments (multiple hierarchies), and Chapter 6 concludes with "many-to-many" problems.

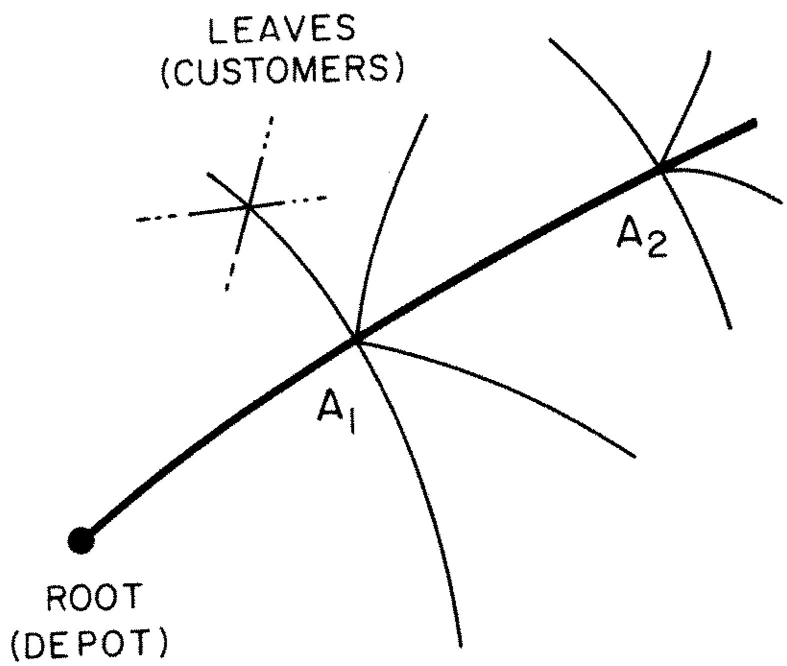

Fig. 3.10 Schematic representation of naturally occurring logistics systems