1

Introduction

1.1 Optimization and evolution

Evolution is directed by natural selection. Those sets of genes which enable animals to survive and reproduce best are most likely to be transmitted to subsequent generations. The ability to survive and reproduce can be measured by the quantity known as '*fitness*'. If a particular set of genes is possessed by n_1 members of the current generation and n_2 members of the next generation (the counts being made at the same stage of the life history in each generation), the fitness of that set of genes is n_2/n_1.

Evolution favours genotypes of high fitness but it does not generally increase fitness in the species as a whole. The reason is that fitness depends on the competitors which have to be faced, as well as on other features of the environment. A genotype which has high fitness now may have much lower fitness at some time in the future when a new, improved genotype has become common. In due course it will probably be eliminated, as evolution proceeds.

Though fitness itself may not increase, other qualities which affect fitness tend to improve in the course of evolution. For instance, natural selection generally favours characters that make animals use food energy more efficiently, enabling them to survive better when food is scarce and to divert more energy to reproduction when food is more plentiful. Natural selection generally favours characters which enable animals to collect food faster, so that they can either collect more food or devote more time to other activities such as reproduction. Natural selection generally favours characters that enable animals to hide or escape from predators more effectively. It favours characters that in these and other ways fit the animal best for life in its present environment.

A shopper looking for the best buy chooses the cheapest article among several of equal quality, or the best among several of equal price. Similarly, natural selection favours sets of genes which minimize costs or maximize benefits. The costs can often be identified as mortality or energy losses, the benefits as fecundity or energy gains.

Optimization is the process of minimizing costs or maximizing benefits, or obtaining the best possible compromise between the two. Evolution by natural selection is a process of optimization. Learning can also be an optimizing process: the animal discovers the most effective technique for some purpose by trial and error. Subsequent chapters show some of the ways in which optima are approached, in the structure and lives of animals. They are therefore concerned largely with maxima and minima. The rest of this chapter is about maxima and minima and ways in which they can be found. It introduces, as simply as possible, some of the mathematical concepts and techniques that are applied to zoological problems in later chapters.

1.2 Maxima and minima

In Fig. 1-1(a), y has a maximum value when $x = x_{\max}$. In Fig. 1-1(b), y has a minimum value when $x = x_{\min}$. It is easy enough to see where the maximum and minimum are when the graphs are plotted but it will be convenient to have a method for finding the maxima and minima of algebraic expressions, without drawing graphs. The most generally useful method is supplied by the branch of mathematics called differential calculus. The rest of this section can be skipped by readers who already know a little calculus.

The method of finding maxima and minima depends on the study of gradients. Figure 1-1(c) shows a straight line passing through the points (x_1, y_1) and (x_2, y_2). The gradient (slope) of this line is $(y_2 - y_1)/(x_2 - x_1)$. In this case $y_2 > y_1$ and $x_2 > x_1$ so the gradient is positive. In the case illustrated in Fig. 1-1(d), however, $y_2 < y_1$ and the gradient is negative.

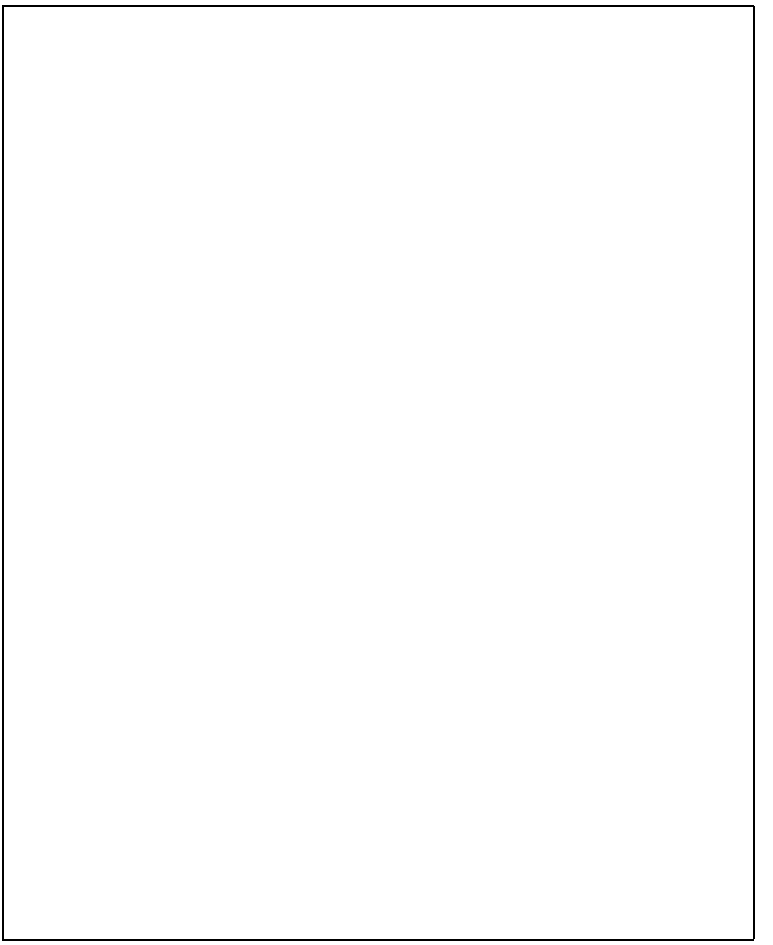

Figure 1-1. Graphs illustrating (**a**) a maximum; (**b**) a minimum; (**c**) a positive gradient and (**d**) a negative gradient.

A straight line has the same gradient all along its length. A curve can be thought of as a chain of very short straight lines of different gradients, joined end to end, so different parts of a curve have different gradients. In Fig. 1-1(a) the gradient is positive before the maximum (i.e. at lower values of x), zero at the maximum and negative after the maximum. Similarly in Fig. 1-1(b) the gradient is

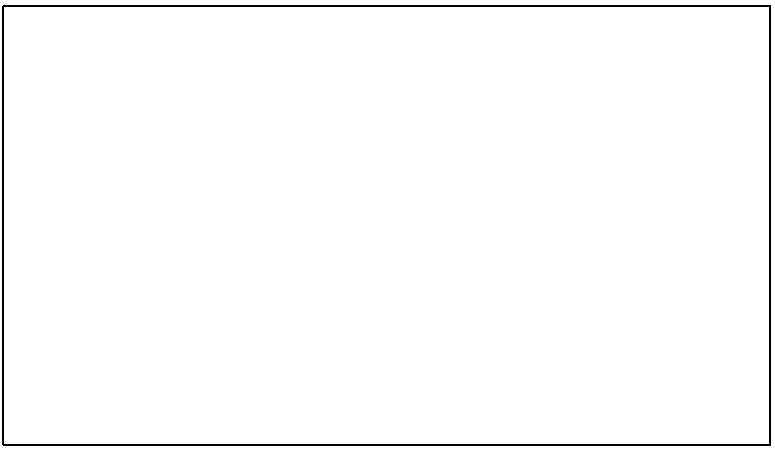

Figure 1-2. (**a**) A graph of y against x. (**b**) A graph of $\mathrm{d}y/\mathrm{d}x$ against x. In both cases, $y = x^2$.

negative before the minimum, zero at it and positive after it. At a maximum the gradient is zero and decreasing, but at a minimum it is zero and increasing.

An example will show how gradients can be calculated. Figure 1-2(a) is a graph of $y = x^2$. It shows that y has just one minimum and no maximum, and that the minimum occurs when $x = 0$. Consider two points very close together on the graph, (x, y) and $(x + \delta x, y + \delta y)$. The symbol δx means a small increase in x and δy means the corresponding increase in y. The gradient of a straight line joining the two points is $\delta y/\delta x$. The smaller δx is, the more nearly is $\delta y/\delta x$ equal to the gradient of the curve at (x, y), which is represented by the symbol $\mathrm{d}y/\mathrm{d}x$. If δx is infinitesimally small, $\delta y/\delta x = \mathrm{d}y/\mathrm{d}x$. Note that $\mathrm{d}y/\mathrm{d}x$ should be read as a single symbol: it is not a quantity $\mathrm{d}y$ divided by a quantity $\mathrm{d}x$, still less can it be interpreted as $(\mathrm{d} \times y) \div (\mathrm{d} \times x)$.

In this particular case $y = x^2$

and also
$$\begin{aligned} y + \delta y &= (x + \delta x)^2 \\ &= x^2 + 2x.\delta x + (\delta x)^2 \end{aligned}$$

Subtracting the first equation from the second

$$\begin{aligned} \delta y &= 2x.\delta x + (\delta x)^2 \\ \delta y/\delta x &= 2x + \delta x \end{aligned}$$

If δx is infinitesimally small, the term δx on the right hand side of the equation can be neglected, and $\delta y/\delta x = \mathrm{d}y/\mathrm{d}x$, so

$$\mathrm{d}y/\mathrm{d}x = 2x$$

It can be shown by similar arguments that if k is a constant

$$\text{when } y = kx \qquad \mathrm{d}y/\mathrm{d}x = k \tag{1.1}$$

$$\text{when } y = kx^2 \qquad \mathrm{d}y/\mathrm{d}x = 2kx \tag{1.2}$$

$$\text{when } y = kx^3 \qquad \mathrm{d}y/\mathrm{d}x = 3kx^2 \tag{1.3}$$

$$\text{when } y = k/x \qquad \mathrm{d}y/\mathrm{d}x = -k/x^2 \tag{1.4}$$

$$\text{and when } y = k \qquad \mathrm{d}y/\mathrm{d}x = 0 \tag{1.5}$$

All these cases are summarized by the general statement

$$\text{when } y = kx^n \qquad \mathrm{d}y/\mathrm{d}x = nkx^{n-1} \tag{1.6}$$

Similar arguments can also be used to obtain expressions for the gradients of other functions of x, for instance

$$\text{when } y = k.\log_e x \qquad \mathrm{d}y/\mathrm{d}x = k/x \tag{1.7}$$

$$\text{and when } y = \exp(kx)^* \qquad \mathrm{d}y/\mathrm{d}x = k.\exp(kx) = ky \tag{1.8}$$

Notice particularly that in this last case, $\mathrm{d}y/\mathrm{d}x$ is proportional to y. If x represented time this case would represent exponential growth, in which rate of growth is proportional to present size.

Other examples can be found in books on calculus. The process of obtaining $\mathrm{d}y/\mathrm{d}x$ from y is called differentiation.

*$\exp(kx)$ means e^{kx}; $e = 2.718$

More complicated expressions are often easy to differentiate. Let y and z be two different functions of x. Then

$$\text{if } u = y + z \qquad \mathrm{d}u/\mathrm{d}x = \mathrm{d}y/\mathrm{d}x + \mathrm{d}z/\mathrm{d}x \tag{1.9}$$

$$\text{if } u = yz \qquad \frac{\mathrm{d}u}{\mathrm{d}x} = z \cdot \frac{\mathrm{d}y}{\mathrm{d}x} + y \cdot \frac{\mathrm{d}z}{\mathrm{d}x} \tag{1.10}$$

$$\text{and if } u = y/z \qquad \frac{\mathrm{d}u}{\mathrm{d}x} = \left(z \cdot \frac{\mathrm{d}y}{\mathrm{d}x} - y \cdot \frac{\mathrm{d}z}{\mathrm{d}x}\right)/z^2 \tag{1.11}$$

Also, if u is a function of y which in turn is a function of x

$$\frac{\mathrm{d}u}{\mathrm{d}x} = \frac{\mathrm{d}u}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x} \tag{1.12}$$

For instance to differentiate $u = (1+x^2)^{\frac{1}{2}}$, write $y = 1+x^2$. Then $\mathrm{d}u/\mathrm{d}y = \mathrm{d}(y^{\frac{1}{2}})/\mathrm{d}y = \frac{1}{2}y^{-\frac{1}{2}}$ and $\mathrm{d}y/\mathrm{d}x = 2x$, so

$$\mathrm{d}u/\mathrm{d}x = xy^{-\frac{1}{2}} = x/(1+x^2)^{\frac{1}{2}}$$

Equations *(1.1)* to *(1.12)* enable us to discover values of x for which $\mathrm{d}y/\mathrm{d}x$ is zero, in particular cases. Thus they help us to find maxima and minima. To distinguish between maxima and minima we also need to know whether $\mathrm{d}y/\mathrm{d}x$ is increasing or decreasing. In other words, we need to know whether the gradient of a graph of $\mathrm{d}y/\mathrm{d}x$ against x is positive (as in Fig. 1-2b) or negative, at the appropriate value of x. This can be discovered by differentiating again.

Since the symbol $\mathrm{d}y/\mathrm{d}x$ was used for the gradient of a graph of y against x, it would be logical to use $\mathrm{d}(\mathrm{d}y/\mathrm{d}x)/\mathrm{d}x$ for the gradient of a graph of $\mathrm{d}y/\mathrm{d}x$ against x. It is customary to write instead, for brevity, $\mathrm{d}^2y/\mathrm{d}x^2$.

At a minimum $\mathrm{d}y/\mathrm{d}x = 0$ and $\mathrm{d}^2y/\mathrm{d}x^2$ is positive.* (*1.13*)

At a maximum $\mathrm{d}y/\mathrm{d}x = 0$ and $\mathrm{d}^2y/\mathrm{d}x^2$ is negative.* (*1.14*)

These rules will be applied to the case $y = x^2$. Differentiation gives $\mathrm{d}y/\mathrm{d}x = 2x$, which is zero when $x = 0$. A second differentiation (using equation 1.1) gives $\mathrm{d}^2y/\mathrm{d}x^2 = 2$, which is positive

*There are some exceptional cases of maxima and minima at which $\mathrm{d}^2y/\mathrm{d}x^2$ is zero. See section 7.1.

when $x = 0$ (and for all other values of x). Hence y has a minimum when $x = 0$.

Most of the mathematical functions discussed in this book have just one maximum and no minima, or just one minimum and no maxima. Many other functions have a maximum and a minimum, or several of each. In such cases the rules *(1.13)* and *(1.14)* are generally adequate to find all the maxima and minima.

1.3 Optima for aircraft

The method which has just been explained will be illustrated by a simple example. It is about aeroplanes, not animals, but the same equation will be applied to birds in chapter 3. The power P required to propel an aeroplane in level flight at constant velocity u is given by the equation

$$P = Au^3 + BL^2/u \tag{1.15}$$

where A and B are constants for the particular aircraft and L is the lift, the upward aerodynamic force which supports the weight of the aircraft. The lift is produced by the wings deflecting air downwards, and the term BL^2/u represents the power required for this purpose. It is called induced power. The term Au^3 represents power which would be needed to drive the aircraft through the air, even if no lift were required. It is called profile power. As the speed u increases, profile power increases but induced power decreases, and there is a particular speed at which the total power P is a minimum. It can be found by differential calculus.

Differentiate equation *(1.15)*, using equations *(1.3)* and *(1.4.)*

$$\mathrm{d}P/\mathrm{d}u = 3Au^2 - BL^2/u^2 \tag{1.16}$$

which is zero when u has the value $u_{\min P}$ given by

$$\begin{aligned} 3Au_{\min P}^2 &= BL^2/u_{\min P}^2 \\ u_{\min P} &= (BL^2/3A)^{\frac{1}{4}} \end{aligned} \tag{1.17}$$

Differentiate equation *(1.16)*

$$\mathrm{d}^2P/\mathrm{d}u^2 = 6Au + 2BL^2/u^3$$

which is positive for all positive values of u. This confirms that the power is a minimum, at the speed given by equation *(1.17)*.

This is the speed at which least power is needed to propel the aircraft, but it may not be the optimum speed. If the aircraft flew faster it would need more power but it would reach its destination sooner and might use less fuel for the journey. The energy E required to travel unit distance is given by

$$\begin{aligned} E &= P/u \\ &= Au^2 + BL^2/u^2 \end{aligned}$$

Differentiation gives

$$\mathrm{d}E/\mathrm{d}u = 2Au - 2BL^2/u^3$$

and E has its minimum value at the speed $u_{\min E}$ given by

$$u_{\min E} = (BL^2/A)^{\frac{1}{4}} \qquad (1.18)$$

This is 32% faster than the speed $u_{\min P}$. A pilot wishing to remain airborne for as long as possible without re-fuelling should fly at $u_{\min P}$ but a pilot wishing to fly as far as possible without re-fuelling should fly at $u_{\min E}$.

1.4 Fitting lines

Figure 1-3(a) is a graph of a quantity y that depends on two variables, a and b. Since a, b and y all have to be represented, this has to be a three-dimensional graph. The third dimension is represented by the contours which give values of y. These contours show that y has a minimum value when $a = 2$ and $b = 3$.

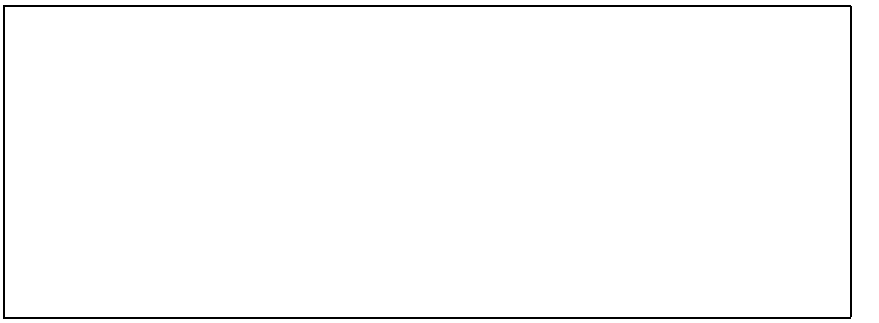

Figure 1-3. These graphs are explained in the text.

Section 1.2 explained how maxima and minima can be found for functions of one variable. The rule for functions of two variables is very similar. At a maximum or minimum

$$\partial y/\partial a = 0 \text{ and } \partial y/\partial b = 0 \qquad (1.19)$$

Remember that dy/dx means the gradient of a graph of y against x. The symbol $\partial y/\partial a$ means the gradient of a graph of y against a, with b (and any other variables) held constant. Similarly $\partial y/\partial b$ means the gradient of a graph of y against b with all other variables held constant. A point at which equations *(1.19)* both hold may be a maximum, or a minimum, or neither. The rules for deciding which are more complicated than conditions *(1.13)* and *(1.14)* and are explained in books about calculus.

The usefulness of conditions *(1.19)* will be illustrated by explaining a standard statistical procedure. Experiments often lead to graphs like Fig. 1-3(b). The points are scattered (due perhaps to experimental errors) but suggest a sloping line. What straight line fits them best? This is a problem in optimization.

Any straight line can be represented by an equation

$$y = ax + b \qquad (1.20)$$

where a and b are constants (a is the gradient and b is the intercept). The problem is to find the best values of a and b, to fit the

points (x_1, y_1), (x_2, y_2) etc. Suppose that particular values have been chosen. The first point is a height h_1 above or below the line, the second is h_2 above or below the line and so on. The standard method of choosing the best line is to choose the values of a and b which minimize the function

$$\Phi = h_1^2 + h_2^2 + h_3^2 + \ldots\ldots h_n^2$$

This can also be written

$$\Phi = \sum_{i=1}^{n} h_i^2$$

which means exactly the same thing. In this equation h_i means any of the heights h_1, h_2 etc. The symbol Σ means sum (add together). The letters above and below it show that all the h_i^2 have to be added together, from the first $(i = 1)$ to the last $(i = n)$.

The y coordinate of the first point is y_1 but the equation *(1.20)* suggests it should be $(ax_1 + b)$. Thus the height h_1 is $(y_1 - ax_1 - b)$ and the function to be minimized is

$$\Phi = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \qquad (1.21)$$

A graph of Φ against a and b would look rather like Fig. 1-3(a). Differentiating Φ with respect to a and b gives

$$\begin{aligned} \partial\Phi/\partial a &= \sum_{i=1}^{n} [-2x_i(y_i - ax_i - b)] \\ \partial\Phi/\partial b &= \sum_{i=1}^{n} [-2(y_i - ax_i - b)] \end{aligned}$$

(The rule for differentiating a function of a function, equation *1.12*, was used). At the minimum, these must both be zero

$$\left.\begin{aligned} \sum_{i=1}^{n} [-2x_i(y_i - ax_i - b)] = 0 \\ \sum_{i=1}^{n} [-2(y_i - ax_i - b)] = 0 \end{aligned}\right\} \qquad (1.22)$$

The optimum values of a and b can be found, in any particular case, by solving *(1.22)* as a pair of simultaneous equations. This is the standard statistical technique of least-squares regression. Further details are explained in books on statistics.

If you have reached this point in the book, you should now understand what maxima and minima are, for functions of one or two variables. You should know the basic rules of differential calculus and understand how they can be used to find maxima and minima. This is sufficient mathematical preparation for the first half of the book, and the next two sections are intended to help you with the second half (particularly sections 4.6, 5.2, 5.3 and 5.4). I suggest that you read them now, but in case you prefer to go immediately to chapter 2 I have inserted references back to these sections, later in the book.

1.5 The best shape for cans

A cylindrical can is to be made to contain a volume V of food. What is the best shape (the best ratio of height h to diameter D)? It will be shown that two different approaches lead to the same answer.

It will be assumed that the best shape is the one that requires the least area of tinplate. The top and bottom of the can are circular discs, each of area $\pi D^2/4$. The circumference of the can is πD so the area of the rectangle of metal needed for its sides is πDh. The total area of metal needed is $(\pi D^2/2) + \pi Dh$. The volume of the can is $\pi D^2 h/4$. The problem can thus be stated

$$\begin{aligned} &\text{minimize } \Phi = (\pi D^2/2) + \pi Dh \\ &\text{subject to } \Psi = (\pi D^2 h/4) - V = 0 \end{aligned} \tag{1.23}$$

Notice that this has the form, 'minimize the function Φ subject to the constraint that the function Ψ is zero'. Problems like this often arise and can be solved in several different ways.

Here is an obvious method that can be used in this case, but is

difficult or impossible for some problems. The constraint gives

$$h = 4V/\pi D^2$$

Substitution of this in the expression for Φ gives

$$\Phi = (\pi D^2/2) + (4V/D)$$

and differentiation gives

$$\begin{aligned} \mathrm{d}\Phi/\mathrm{d}D &= \pi D - 4V/D^2 \\ &= \pi D - \pi h \end{aligned}$$

(using the constraint again). When Φ has its minimum value, $\mathrm{d}\Phi/\mathrm{d}D$ must be zero and so h must equal D. To make a can of given volume from least metal, the height and diameter should be made equal.

There is another method for getting the same result, which is sometimes more convenient. It will be presented here simply as a recipe, but explanations of why it works can be found in books on optimization (for instance Koo, 1977). Define a new function

$$L = \Phi + \lambda\Psi \tag{1.24}$$

The symbol λ represents a constant called a Lagrange multiplier. It can be shown that when D and h have the values which represent the solution to the problem, $\partial L/\partial D$ and $\partial L/\partial h$ must both be zero. In this particular problem

$$L = (\pi D^2/2) + \pi Dh + \lambda[(\pi D^2 h/4) - V] \tag{1.25}$$

so that at the minimum

$$\partial L/\partial D = \pi D + \pi h + \lambda[(\pi Dh/2) - 0] = 0 \tag{1.26}$$

and

$$\partial L/\partial h = 0 + \pi D + \lambda[(\pi D^2/4) - 0] = 0 \tag{1.27}$$

Equation *(1.27)* requires either $D = 0$ (which is impossible, if the can is to hold anything) or $\lambda = -4/D$. By putting this value of

λ in equation *(1.26)* we get

$$\begin{aligned} \pi D + \pi h - 2\pi h &= 0 \\ h &= D \end{aligned}$$

so this method gives the same answer as the other one. The height should be made equal to the diameter.

Most of the food cans in my store cupboard have heights greater than their diameters. Their manufacturers were plainly not minimizing the area of metal used, but they may have been applying some other optimization criterion.

1.6 The shortest path

The problems of the aeroplane and of the line of best fit were answered by finding the optimum values for variables. Some problems require instead, the discovery of an optimum function. Consider, for instance, the problem of finding the shortest path between two points. The required answer is not a number but an equation, the equation of the shortest line that joins them. If we did not happen to know already that the answer is a straight line we would not know what mathematical form the equation should take, whether it should be $y = ax + b$ or $y = ax^b$ or $y = a^{bx}$ or something more complicated. Problems like this can sometimes be solved by a method called the calculus of variations.

Figure 1-4(a) shows a line (any line) joining the points P and Q. Consider a short segment of the line, from the point (x, y) to $(x + \delta x, y + \delta y)$. The length of this segment is (by Pythagoras) $[(\delta x)^2 + (\delta y)^2]^{\frac{1}{2}}$. Since δx and δy are short, $\delta y/\delta x$ is almost exactly equal to the gradient $\mathrm{d}y/\mathrm{d}x$ so δy is approximately $(\mathrm{d}y/\mathrm{d}x).\delta x$. The length δl of the segment is given by

$$\delta l = [1 + (\mathrm{d}y/\mathrm{d}x)^2]^{\frac{1}{2}}.\delta x$$

To make subsequent equations less clumsy the symbol y' will be used to represent $\mathrm{d}y/\mathrm{d}x$.

$$\delta l = (1 + y'^2)^{\frac{1}{2}}.\delta x \qquad (1.28)$$

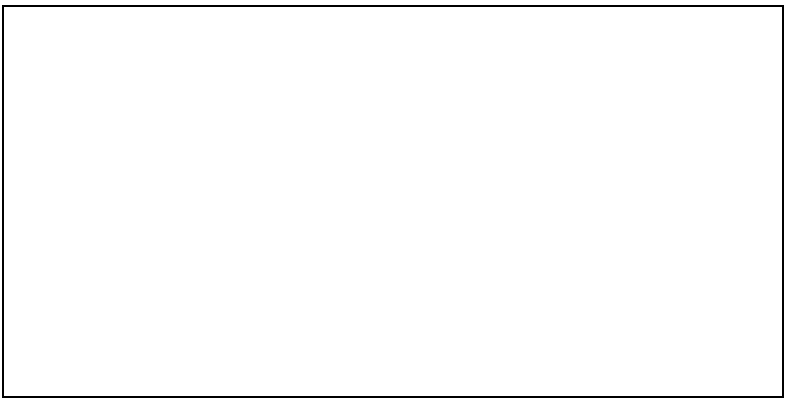

Figure 1-4. These graphs are explained in the text.

Figure 1-4(b) is a graph of $(1 + y'^2)^{\frac{1}{2}}$ against x. The narrow hatched strip under it is $(1 + y'^2)^{\frac{1}{2}}$ high and δx wide, so its area gives the value of δl. Other strips could be drawn representing the lengths of all the other segments of the line. Their areas could be added together to obtain the total length l of the line from P to Q. Thus the area of the whole stippled region under the graph gives the length l of the line. This is expressed by writing

$$l = \int_{x_P}^{x_Q} (1 + y'^2)^{\frac{1}{2}} . \mathrm{d}x \qquad (1.29)$$

The expression on the right hand side of the equation, representing the area under a graph, is called an integral. The symbols x_P and x_Q at the bottom and top of the integral sign say where the area starts and ends. Books on calculus show how integrals can be evaluated, but the technique of integration is not needed here. It is sufficient for readers to understand what integrals are.

The problem of finding the shortest line joining two points is the problem of finding the equation relating x and y that minimizes the length l. This is a problem of the kind that the calculus of variations is designed to solve.

The standard problem of the calculus of variations is to find the equation relating x and y that minimizes (or maximizes) some

function Φ that has the form

$$\Phi = \int_{x_a}^{x_b} f(x, y, y').dx \qquad (1.30)$$

In this equation, $f(x, y, y')$ means some function of any or all of x, y and y'. The most important theorem of the calculus of variations is that any solution to the problem must satisfy the Euler equation

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0 \qquad (1.31)$$

for all values of x between x_a and x_b. A proof of the theorem is given in books by Koo (1977) and others.

In the problem of finding the shortest line, the function f is $(1 + y'^2)^{\frac{1}{2}}$. Since y does not appear in this expression (but only y'), $\partial f/\partial y = 0$. The example following equation *(1.12)* shows that $\partial f/\partial y' = y'/(1 + y'^2)^{\frac{1}{2}}$. Thus the Euler equation says

$$\frac{d}{dx}\left(\frac{y'}{(1 + y'^2)^{\frac{1}{2}}}\right) = 0$$

for all values of x between x_P and x_Q. This can only be true if y' $(= dy/dx)$ is a constant, independent of x. Thus the gradient of the shortest path from P to Q (Fig. 1-4a) is constant and that path is a straight line.

This may seem a long-winded way of proving the obvious, but the calculus of variations is useful for solving other, less easy, problems of optimization.

1.7 Conclusion

This chapter has introduced some of the mathematics that will be applied to zoological problems in the rest of the book. A few more mathematical ideas will be introduced and explained when the need for them arises. This chapter has also illustrated an important general point. It is not sensible to claim in general terms that (for instance) a particular speed is optimal for aircraft, or that

a particular shape is optimal for cylindrical cans. It is essential to state the criteria for optimization, that the speed is to be chosen (for instance) to minimize the power requirement and the shape to minimize the area of tinplate.