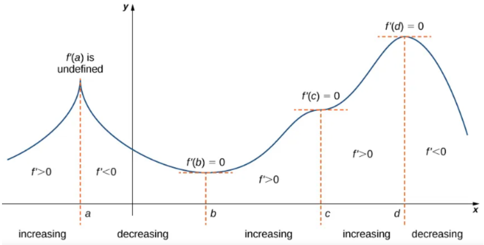
Optimization of $f(x)$

DEFINITION

A function f has a **local maximum** at c if there exists an open interval I containing c such that I is contained in the domain of f and $f(c) \geq f(x)$ for all $x \in I$. A function f has a **local minimum** at c if there exists an open interval I containing c such that I is contained in the domain of f and $f(c) \leq f(x)$ for all $x \in I$. A function f has a **local extremum** at c if f has a local maximum at c or f has a local minimum at c.

Fermat's Theorem

If f has a local extremum at c and f is differentiable at c, then $f'(c) = 0$.



Optimization of $f(x)$

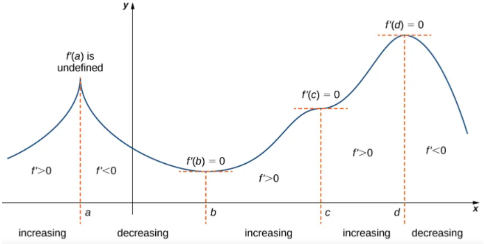
First Derivative Test

Suppose that f is a continuous function over an interval I containing a critical point c. If f is differentiable over I, except possibly at point c, then $f(c)$ satisfies one of the following descriptions:

i. If f' changes sign from positive when $x < c$ to negative when $x > c$, then $f(c)$ is a local maximum of f.

ii. If f' changes sign from negative when $x < c$ to positive when $x > c$, then $f(c)$ is a local minimum of f.

iii. If f' has the same sign for $x < c$ and $x > c$, then $f(c)$ is neither a local maximum nor a local minimum of f.



Optimization of $f(x_1, x_2, \ldots, x_n)$

DEFINITION

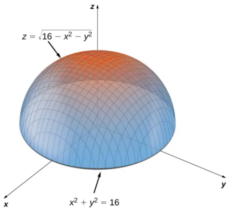
Let $z = f(x, y)$ be a function of two variables that is defined on an open set containing the point (x_0, y_0). The point (x_0, y_0) is called a **critical point of a function of two variables** f if one of the two following conditions holds:

1. $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$
2. Either $f_x(x_0, y_0)$ or $f_y(x_0, y_0)$ does not exist.

THEOREM 4.16

Fermat's Theorem for Functions of Two Variables

Let $z = f(x, y)$ be a function of two variables that is defined and continuous on an open set containing the point (x_0, y_0). Suppose f_x and f_y each exists at (x_0, y_0). If f has a local extremum at (x_0, y_0), then (x_0, y_0) is a critical point of f.



Optimization of $f(x_1, x_2, \ldots, x_n)$

Let $z = f(x, y)$ be a function of two variables for which the first- and second-order partial derivatives are continuous on some disk containing the point (x_0, y_0). Suppose $f_x(x_0, y_0) = 0$ and $f_y(x_0, y_0) = 0$. Define the quantity

$$D = f_{xx}(x_0, y_0) f_{yy}(x_0, y_0) - (f_{xy}(x_0, y_0))^2. \quad (4.43)$$

i. If $D > 0$ and $f_{xx}(x_0, y_0) > 0$, then f has a local minimum at (x_0, y_0).
ii. If $D > 0$ and $f_{xx}(x_0, y_0) < 0$, then f has a local maximum at (x_0, y_0).
iii. If $D < 0$, then f has a saddle point at (x_0, y_0).
iv. If $D = 0$, then the test is inconclusive.

See Figure 4.49.

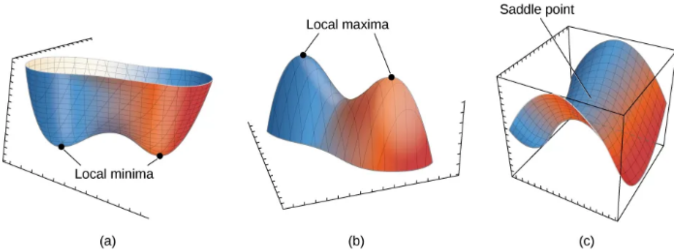


Figure 4.49 The second derivative test can often determine whether a function of two variables has a local minima (a), a local

Optimization of $f(x_1, x_2, \ldots, x_n)$: Gradient Descent

DEFINITION

Let $z = f(x, y)$ be a function of x and y such that f_x and f_y exist. The vector $\nabla f(x, y)$ is called the **gradient** of f and is defined as

$$\nabla f(x, y) = f_x(x, y)\,\mathbf{i} + f_y(x, y)\,\mathbf{j}. \tag{4.39}$$

The vector $\nabla f(x, y)$ is also written as "grad f."

Properties of the Gradient

Suppose the function $z = f(x, y)$ is differentiable at (x_0, y_0) (Figure 4.41).

i. If $\nabla f(x_0, y_0) = \mathbf{0}$, then $D_{\mathbf{u}} f(x_0, y_0) = 0$ for any unit vector $\mathbf{u}$.
ii. If $\nabla f(x_0, y_0) \neq \mathbf{0}$, then $D_{\mathbf{u}} f(x_0, y_0)$ is maximized when $\mathbf{u}$ points in the same direction as $\nabla f(x_0, y_0)$. The maximum value of $D_{\mathbf{u}} f(x_0, y_0)$ is $\|\nabla f(x_0, y_0)\|$.
iii. If $\nabla f(x_0, y_0) \neq \mathbf{0}$, then $D_{\mathbf{u}} f(x_0, y_0)$ is minimized when $\mathbf{u}$ points in the opposite direction from $\nabla f(x_0, y_0)$. The minimum value of $D_{\mathbf{u}} f(x_0, y_0)$ is $-\|\nabla f(x_0, y_0)\|$.

Optimization of $f(x_1, x_2, \ldots, x_n)$: Gradient Descent

Gradient Descent Algorithm:

Let $J(x_1, x_2, \ldots, x_n)$ be a differentiable function:

$w \leftarrow$ random values
for i < epoch **do**
    grad $\leftarrow \nabla_w J$
    $w \leftarrow w - \eta \times$grad
    $i \leftarrow i + 1$
end for

Supervised Learning (I)

More generally, suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon. \tag{2.1}$$

Prediction

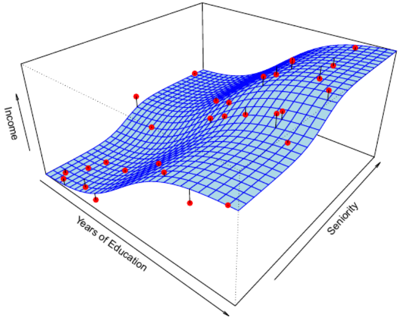
In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2.2}$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. In this setting, $\hat{f}$ is often treated as a *black box*, in the sense that one is not typically concerned with the exact form of $\hat{f}$, provided that it yields accurate predictions for Y.

How to estimate $f(x)$: Parametric Methods

- Use training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.



- Make an assumption about the functional form of $f(x)$.
- The most common assumption about $f(x)$ is that it is linear:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Simple Linear Regression

Assume you have training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We must find $\hat{\beta}_0 + \hat{\beta}_1 x$.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*—this is the difference between the ith observed response value and the ith response value that is predicted by our linear model. We define the *residual sum of squares* (RSS) as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad (3.3)$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \quad (3.4)$$

Question: Why?