Chapter 1 Optimization and Solving Nonlinear Equations

1.1 Definitions and Notations

- **Optimization Problem:** We want to find a point $\theta^* \in \Theta$ (for example, $\Theta = \mathbb{R}^p$) to maximize (or minimize) an *objective function* $g(\theta)$, denoted by

$$\theta^* = \arg\max_{\theta \in \Theta} g(\theta).$$

- **Maximum Likelihood Estimate (MLE):** Let $X_1, X_2, \cdots, X_n$ be a random sample following a distribution with probability density function (PDF) $f(x_1, \cdots, x_n; \theta)$, where θ is a $p \times 1$ vector.
  - For each given $x_1, \cdots, x_n$, $f(x_1, \cdots, x_n; \theta)$ considered as a function of the parameter θ is called the *likelihood function* and denoted by $l(\theta)$.
  - The MLE of θ is

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} l(\theta) = \arg\max_{\theta \in \Theta} \log l(\theta).$$

1.1 Definitions and Notations

- – Suppose $X_1, \cdots, X_n$ are i.i.d. following a distribution with PDF $f(x;\theta_0)$, where θ_0 is the true parameter. Let $\hat{\theta}_{MLE}$ be the MLE of θ using $X_1, \cdots, X_n$. Then under certain regulation conditions,

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_{\theta} f(X_1, \cdots, X_n; \theta) \\
&= \arg\max_{\theta} \log f(X_1, \cdots, X_n; \theta) \\
&= \arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log f(X_i; \theta) \\
&\approx \arg\max_{\theta} \int [\log f(x;\theta)]\, f(x;\theta_0)\, dx \\
&= \theta_0.
\end{aligned}
$$

1.1 Definitions and Notations

- **Jensen's Inequality:** Suppose that X is a random variable with $E|X| < \infty$ and $\phi(\cdot)$ is a concave function, then

$$E[\phi(X)] \leq \phi[E(X)].$$

- By Jensen's inequality,

$$\begin{aligned}\int \left[\log \frac{f(x;\theta)}{f(x;\theta_0)}\right] f(x;\theta_0)\, dx &= E_{\theta_0}\left[\log \frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right] \\ &\leq \log\left\{E_{\theta_0}\left[\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right]\right\} = 0.\end{aligned}$$

1.1 Definitions and Notations

- **Method of Moments Estimate (MME):** Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (i.i.d.) random variables following a distribution with probability density function $f(x;\theta_0)$, where θ_0 is a $p \times 1$ vector.

  – The MME of θ_0 is the solution of equations

  $$\frac{1}{n}\sum_{i=1}^{n} m(X_i) = E_\theta\big[m(X_i)\big] = \int m(x) f(x;\theta)\, dx,$$

  where $m(X_i) = \big(m_1(X_i), \cdots, m_p(X_i)\big)'$, *e.g.*, $m(X_i) = \big(X_i, \cdots, X_i^p\big)'$

  – We have

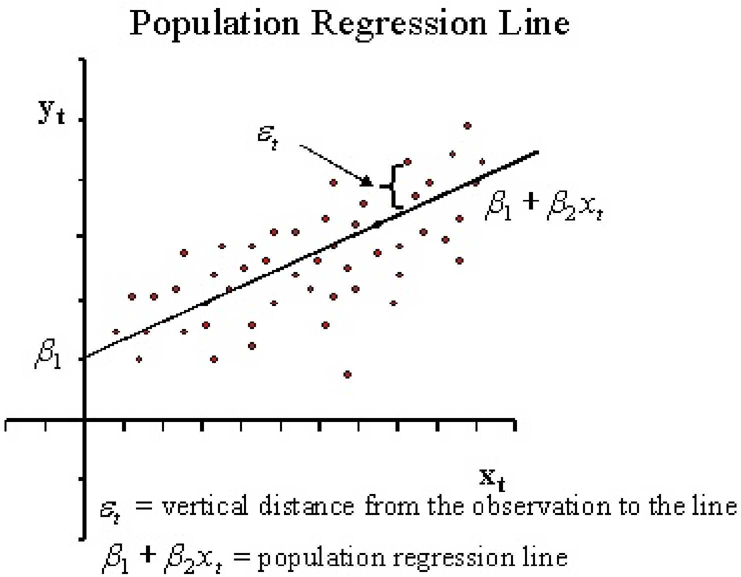
  $$\hat{\theta}_{MME} = \arg\max_{\theta\in\Theta} \Big\{ - \parallel \frac{1}{n}\sum_{i=1}^{n} m(X_i) - E_\theta\big[m(X_i)\big] \parallel_2^2 \Big\},$$

  where $\parallel u \parallel_2 := \Big(\sum_{k=1}^{p} u_k^2\Big)^{1/2}$ if $u = (u_1, \cdots, u_p)'$.

1.1 Definitions and Notations

- **Ordinary Least Square (OLS):** Let X_i and Y_i be the height and weight of person i, respectively, for $i = 1, \cdots, n$. We want to find a linear relationship between X_i and Y_i, that is, find $\theta = (\beta_1, \beta_2)'$ so that $Y_i \approx \beta_1 + \beta_2 X_i$. The parameter θ can be estimated by

$$\hat{\theta} = \arg \min_{(\beta_1,\beta_2)} \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_i)^2 = \arg \max_{(\beta_1,\beta_2)} \left\{ -\sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 X_i)^2 \right\}.$$



1.1 Definitions and Notations

- **Example: A Transportation Problem.** A chemical company has 2 factories F_1 and F_2 and a dozen retail outlets $R_1, R_2, \cdots, R_{12}$.
  - Each factory F_i can produce a_i tons of a certain chemical product each week, and each retail outlet R_j has a known weekly demand of b_j tons of the product. Suppose that $a_1 + a_2 \geq b_1 + \cdots + b_{12}$.
  - The cost of shipping one ton of the product from factory F_i to retail outlet R_j is c_{ij}.
  - Let θ_{ij} is the number of tons of the product shipped from factory F_i to retail outlet R_j. We find to find the "optimal" $\{\theta_{ij}, i = 1, 2; j = 1, \cdots, 12\}$ to minimize the transportation cost, under the constraint that demands of all outlets are satisfied.

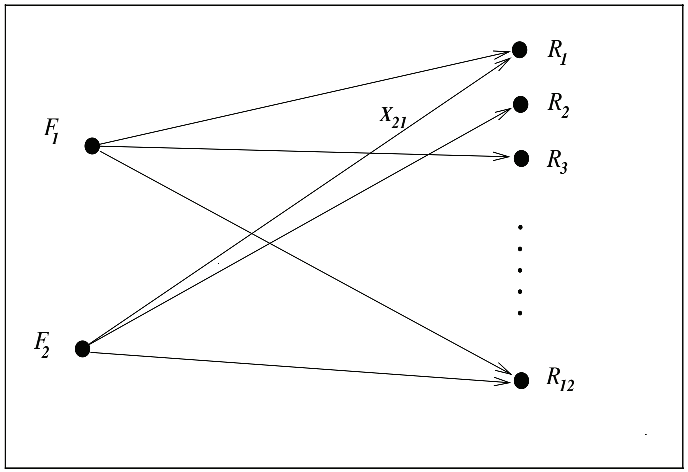
1.1 Definitions and Notations

- – Consider the optimization problem

$$\min_{\theta_{ij}} \sum_{i=1}^{2} \sum_{j=1}^{12} \theta_{ij} c_{ij} \quad \text{or} \quad \max_{\theta_{ij}} \left\{ - \sum_{i=1}^{2} \sum_{j=1}^{12} \theta_{ij} c_{ij} \right\}$$

  subject to

$$\theta_{ij} \geq 0, \quad \sum_{j=1}^{12} \theta_{ij} \leq a_i, \quad \sum_{i=1}^{2} \theta_{ij} \geq b_j, \quad i = 1, 2;\ j = 1, \cdots, 12.$$



1.1 Definitions and Notations

- **Unconstrained Optimization:** Find a point $\theta^* \in \Theta$ to minimize the *objective function* $g(\theta)$, that is,

$$\theta^* = \arg\max_{\theta \in \Theta} g(\theta).$$

- **Constrained Optimization:** Find the solution of

$$\max_{\theta \in \Theta} g(\theta)$$

subject to

$$\begin{aligned} c_i(\theta) &= 0, \quad i = 1, \cdots, m_1; \\ c_i(\theta) &\geq 0, \quad i = m_1 + 1, \cdots, m, \end{aligned}$$

where $c_i(\theta) = 0$ is called the *equality constraint*, and $c_i(\theta) \geq 0$ is called the *inequality constraint*.

- In this chapter, we focus on **unconstrained optimization** problems.

1.1 Definitions and Notations

- **Global Maximizer:** A point θ^* is called a *global maximizer* of $g(\theta)$ if

$$g(\theta^*) \geq g(\theta) \quad \text{for all } \theta \in \Theta.$$

  – The global maximizer **may not exist**, for example, $g(\theta) = -\frac{1}{1+\theta^2}$.

  – The global maximizer **may not be unique**, for example, $g(\theta) = \min\{-|\theta|, -1\}$.

- **Local Maximizer:** A point θ^* is called a *local maximizer* if there is a neighborhood $\mathcal{N} \subset \Theta$ of θ^* such that

$$g(\theta^*) \geq g(\theta) \quad \text{for all } \theta \in \mathcal{N}.$$

  It is called a *strict local maximizer* if

$$g(\theta^*) > g(\theta) \quad \text{for all } \theta \in \mathcal{N} \text{ and } \theta \neq \theta^*.$$

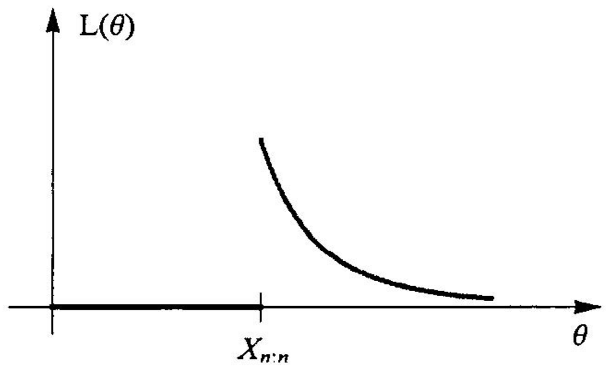
1.1 Definitions and Notations

- **Gradient and Hessian Matrix:** The *gradient* of g at θ is denoted by $\nabla g(\theta) = (\frac{\partial g(\theta)}{\partial \theta_1}, \cdots, \frac{\partial g(\theta)}{\partial \theta_p})'$, and the *Hessian matrix* of g at θ is denoted by

$$\nabla^2 g(\theta) = \left\{ \frac{\partial^2 g(\theta)}{\partial \theta_i \partial \theta_j} \right\}_{p \times p}.$$

  - The Hessian Matrix is a symmetric matrix.
  - A point satisfying $\nabla g(\theta) = \mathbf{0}$ is called a *stationary point* of $g(\theta)$, where $\mathbf{0} = (0, \cdots, 0)'$.
  - Let θ^* be a local maximizer of g. If θ^* is an interior point of Θ and $\nabla g(\theta^*)$ exists, then $\nabla g(\theta^*) = \mathbf{0}$. (**Why?**)
  - When $\nabla g(\theta^*) = \mathbf{0}$, θ^* may not be a local maximizer (or minimizer) of g, for example, $g(\theta) = \theta^3$, consider the point $\theta^* = 0$.

- **Example:** When $g(\theta)$ is not differentiable, the local maximizer (or local minimizer) may not be a stationary point. Consider random variables $X_1, \cdots, X_n$ i.i.d. from the $U[0, \theta]$ distribution. The likelihood function is
$$l(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \le X_i \le \theta) = \frac{1}{\theta^n} I\big(\max\{X_1, \cdots, X_n\} \le \theta\big) \prod_{i=1}^{n} I(X_i > 0).$$
The MLE is $\hat{\theta}_{MLE} = \max\{X_1, \cdots, X_n\}$, which is not a stationary point.



1.1 Definitions and Notations

- **Theorem:** Suppose that $\nabla^2 g(\theta)$ is continuous at θ^* and that $\nabla g(\theta^*) = 0$ and $\nabla^2 g(\theta^*)$ is negative definite. Then θ^* is a strict local maximizer of g.

- **Proof.**

  – Because $\nabla^2 g(\theta^*)$ is negative definite and $\nabla^2 g(\theta)$ is continuous at θ^*, so $\nabla^2 g(\theta)$ is negative definite for all $\theta \in \mathcal{D} := \{u : \| u - \theta^* \|_2 < r\}$ for some $r > 0$.

  – For any $\theta \in \mathcal{D}$, we have

$$\begin{aligned} g(\theta) &= g(\theta^*) + \nabla g(\theta^*)^T (\theta - \theta^*) + \frac{1}{2}(\theta - \theta^*)^T \nabla^2 g(\bar{\theta})(\theta - \theta^*) \\ &= g(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T \nabla^2 g(\bar{\theta})(\theta - \theta^*), \end{aligned}$$

  where $\bar{\theta} \in \mathcal{D}$ is a point between θ^* and θ.

  – Since $\nabla^2 g(\bar{\theta})$ is negative definite, we have $g(\theta) < g(\theta^*)$ if $\theta \in \mathcal{D}$ and $\theta \neq \theta^*$. Therefore, θ^* is a strict local maximizer of g.

1.1 Definitions and Notations

- A local maximizer may not be a global maximizer. Many algorithms for nonlinear optimization problems seek only a local maximizer. We often need to try different initial points.

- **Convex Set:** A set $\mathcal{S} \subset \mathbb{R}^p$ is a *convex set* if for any two points $\theta_1 \in \mathcal{S}$ and $\theta_2 \in \mathcal{S}$, we have $\alpha\theta_1 + (1-\alpha)\theta_2 \in \mathcal{S}$ for all $0 \leq \alpha \leq 1$.

- **Convex Function and Concave Function:** A function $g(\theta)$ defined on a convex set Θ is called a *convex function* if

$$g\big(\alpha\theta_1 + (1-\alpha)\theta_2\big) \leq \alpha g(\theta_1) + (1-\alpha)g(\theta_2) \quad \text{for all } 0 \leq \alpha \leq 1.$$

It is called a *concave function* if

$$g\big(\alpha\theta_1 + (1-\alpha)\theta_2\big) \geq \alpha g(\theta_1) + (1-\alpha)g(\theta_2) \quad \text{for all } 0 \leq \alpha \leq 1.$$

- **Remark:** If $g(\theta)$ is convex, then $-g(\theta)$ is concave.

1.1 Definitions and Notations

- **Theorem:** When g is concave, any local maximizer θ^* is a global maximizer of g. If in addition g is differentiable, then any stationary point θ^* is a global maximizer of g.

- **Proof.**

  – Suppose that θ^* is a local but not a global maximizer. Then there exist a point θ_0 with $g(\theta_0) > g(\theta^*)$. For any point $\theta_1 = \alpha\theta^* + (1-\alpha)\theta_0$, $0 < \alpha < 1$, between θ^* and θ_0, we have

$$\begin{aligned} g(\theta_1) &= g\big(\alpha\theta^* + (1-\alpha)\theta_0\big) \\ &\geq \alpha g(\theta^*) + (1-\alpha)g(\theta_0) > g(\theta^*), \end{aligned}$$

  which implies θ^* is not a local maximizer. So we arrive at a contradiction. Therefore, θ^* must be a global maximizer.

- – If g is differentiable, suppose that $\nabla g(\theta^*) = \mathbf{0}$ but θ^* is not a global maximizer. Let θ_0 be the point satisfying $g(\theta_0) > g(\theta^*)$. Consider

$$\begin{aligned}
\frac{d}{d\lambda} g\big(\theta^* + \lambda(\theta_0 - \theta^*)\big)\Big|_{\lambda=0} &= \lim_{\lambda \to 0+} \frac{g\big(\theta^* + \lambda(\theta_0 - \theta^*)\big) - g(\theta^*)}{\lambda} \\
&\geq \lim_{\lambda \to 0+} \frac{\lambda g(\theta_0) + (1-\lambda) g(\theta^*) - g(\theta^*)}{\lambda} \\
&= g(\theta_0) - g(\theta^*) > 0.
\end{aligned}$$

But we also have

$$\frac{d}{d\lambda} g\big(\theta^* + \lambda(\theta_0 - \theta^*)\big)\Big|_{\lambda=0} = \nabla g(\theta^*)^T (\theta_0 - \theta^*) = 0,$$

and arrive at a contradiction. Hence, θ^* is a global maximizer.

1.2 Univariate Problems

- **Smooth Cases:** Suppose $g''(\theta)$ exists and is continuous, and the maximizer is in the interior of Θ. We want to find the stationary point of g. Let θ^* be a solution of the equation $g'(\theta) = 0$.
  - If $g''(\theta^*) < 0$, then θ^* is a local maximizer of $g(\theta)$.
  - If g is concave, θ^* is a global maximizer of $g(\theta)$.

- In the following, we focus on finding the solution of the equation

$$h(\theta) = 0$$

with $h(\theta) = g'(\theta)$.

- **Example:** Consider the optimization problem

$$\max_{0<\theta<\infty} \left\{ \frac{\log \theta}{1+\theta} \right\}.$$

We want to find

$$\theta^* = \arg\max_{\theta} g(\theta) \quad \text{with } g(\theta) = \frac{\log \theta}{1+\theta},$$

Equivalently, we want to find

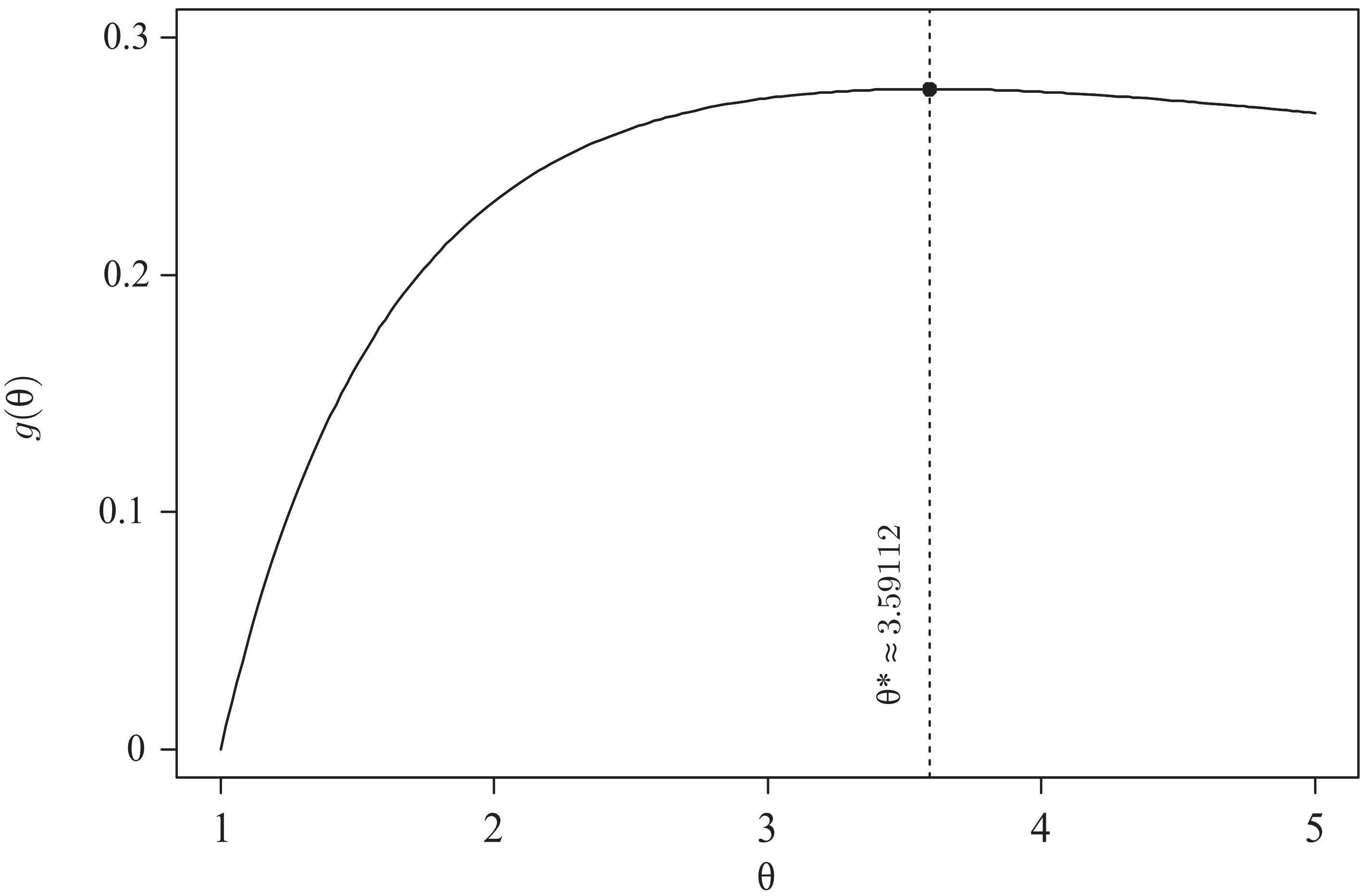
$$\theta^* = \arg\operatorname{zero}_\theta h(\theta),$$

where

$$h(\theta) = g'(\theta) = \frac{1 + 1/\theta - \log \theta}{(1+\theta)^2}.$$

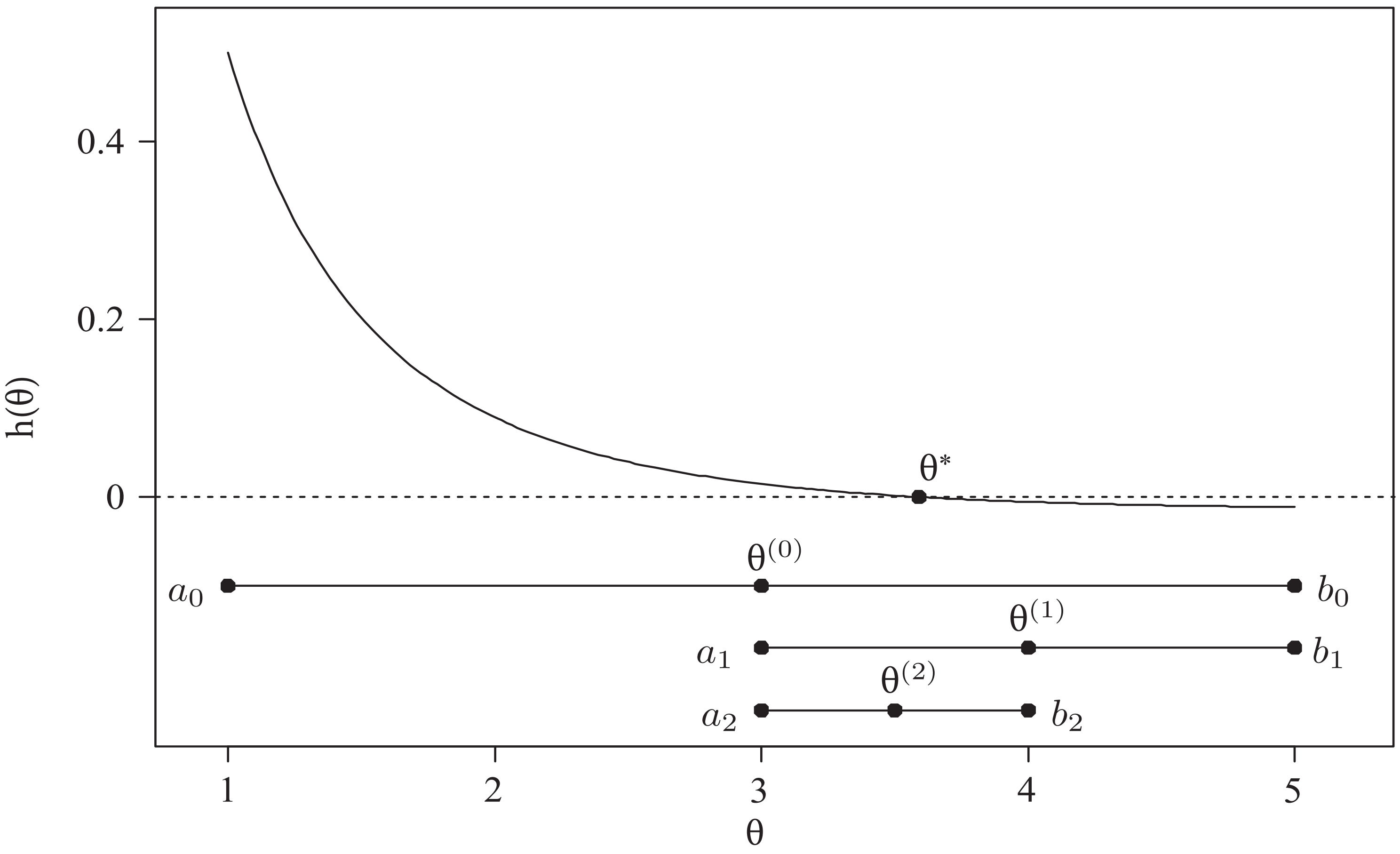
This problem does not have an analytic solution, we need to use **numerical method** to find θ^*.

1.2 Univariate Problems



1.2 Univariate Problems

- **Bisection Method:** Suppose $h(\theta)$ is **continuous**. If $a < b$ and $h(a)h(b) < 0$, there must be a zero point within the interval (a, b).
  - Let $t = 0$, find $a_0 < b_0$ such that $h(a_0)h(b_0) < 0$. Let $h_a = h(a_0)$ and $h_b = h(b_0)$.
  - Until $|b_t - a_t| < \epsilon$:
    * Let $c \leftarrow (a_t + b_t)/2$ and $h_c = h(c)$ (if $h_c = 0$, we find the solution).
    * If $h_a h_c < 0$, let $a_{t+1} = a_t$, $b_{t+1} = c$ and $h_b = h_c$; if $h_b h_c < 0$, let $a_{t+1} = c$, $h_a = h_c$ and $b_{t+1} = b_t$.
    * $t \leftarrow t + 1$.
  - Let $\hat{\theta}^* = (a_t + b_t)/2$.



Bisection Method

1.2 Univariate Problems

- **Remark:** If we want to find the maximizer of $g(\theta)$, we would expect $h(a_0) = g'(a_0) > 0$ and $h(b_0) = g'(b_0) < 0$.

- How to determine the initial interval (a_0, b_0) in bisection method? Suppose we want to find the maximizer of $g(\theta)$ on $(-\infty, \infty)$.

- We first take a point θ_0. If $h(\theta_0) > 0$, let $a_0 = \theta_0$ and $h_a = h(\theta_0)$; if $h(\theta_0) < 0$, let $b_0 = \theta_0$ and $h_b = h(\theta_0)$. Assume that $h(\theta_0) > 0$, we need to determine b_0.

  - Set $h > 0$ and $\gamma > 1$. Let $b_0 = a_0 + h$ and $h_b = h(b_0)$.
  - Until $h_a h_b < 0$:
    - Let $h \leftarrow \gamma\, h$.
    - Compute $b_0 = a_0 + h$ and $h_b = h(b_0)$.

1.2 Univariate Problems

- **Newton's Method:** Suppose at iteration t, we have $\theta^{(t)}$. Consider the Taylor series expansion

$$h(\theta) \approx h(\theta^{(t)}) + h'(\theta^{(t)})(\theta - \theta^{(t)}).$$

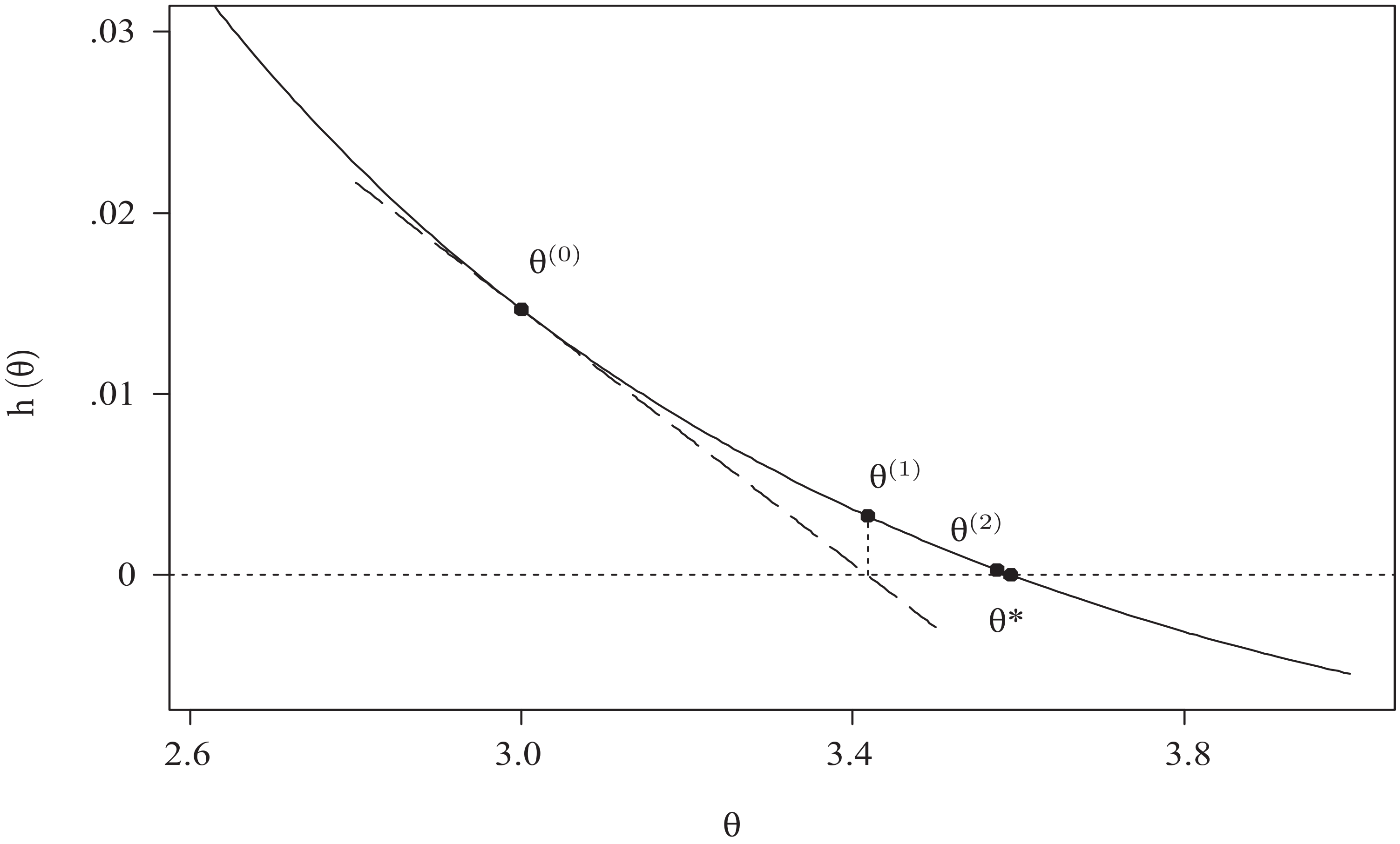
So we can let

$$\theta^{(t+1)} = \theta^{(t)} - h(\theta^{(t)})/h'(\theta^{(t)}).$$

- **Remarks:**
  - The method is also called the *Newton-Raphson method.*
  - We can stop the algorithm when $h(\theta^{(t)})$ is close to 0.
  - When $h(\theta) = g'(\theta)$, we have

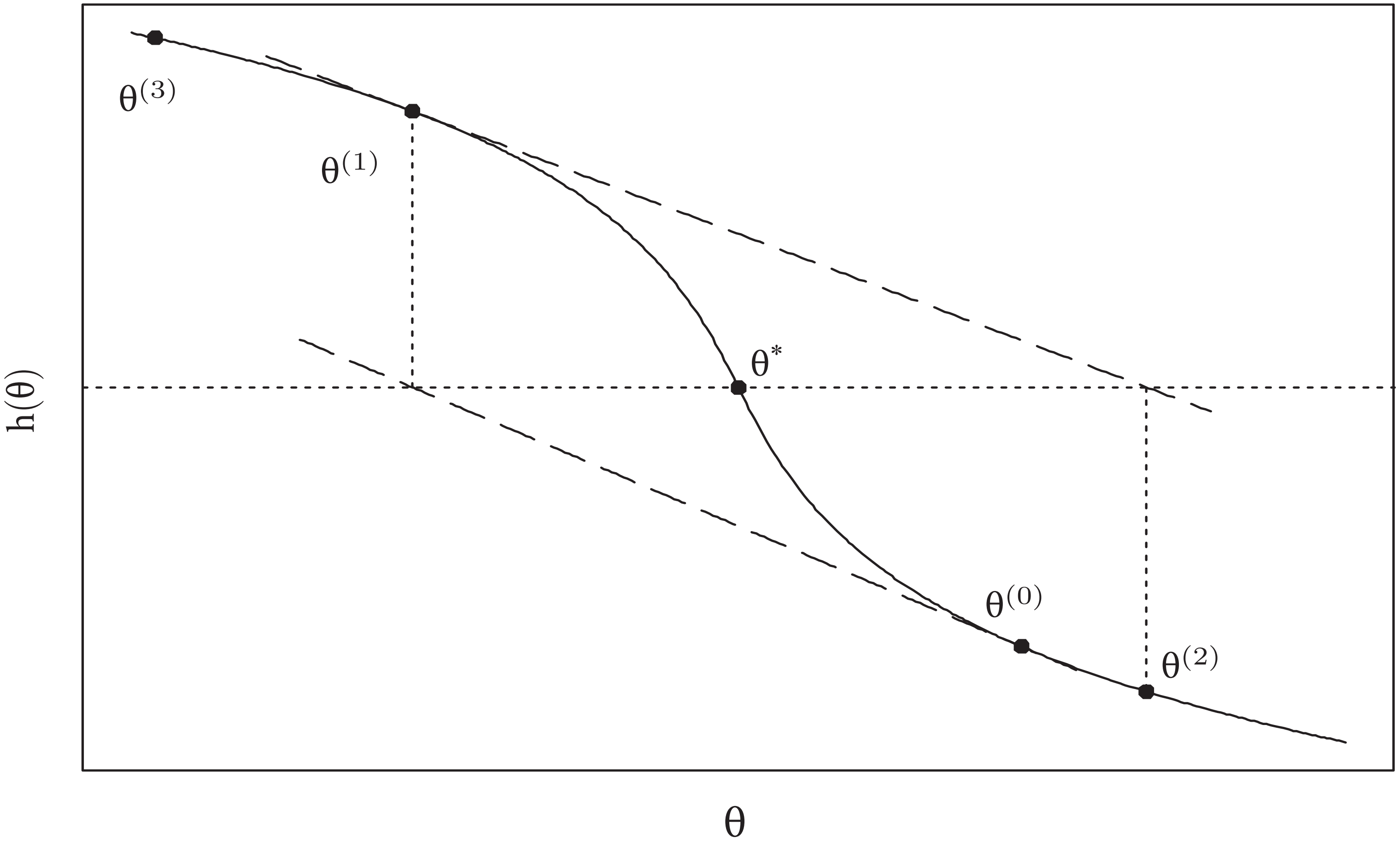
$$\theta^{(t+1)} = \theta^{(t)} - g'(\theta^{(t)})/g''(\theta^{(t)}).$$

  - If we want to find the maximizer of $g(\theta)$, we would expect $g''(\theta^{(t)}) < 0$.



Newton's Method

- – Different from the bisection method, the Newton's method could diverge.



- **Secant Method:** When it is difficult to calculate $h'(\theta^{(t)})$, we can use $\frac{h(\theta^{(t)})-h(\theta^{(t-1)})}{\theta^{(t)}-\theta^{(t-1)}}$ to replace it, then we have

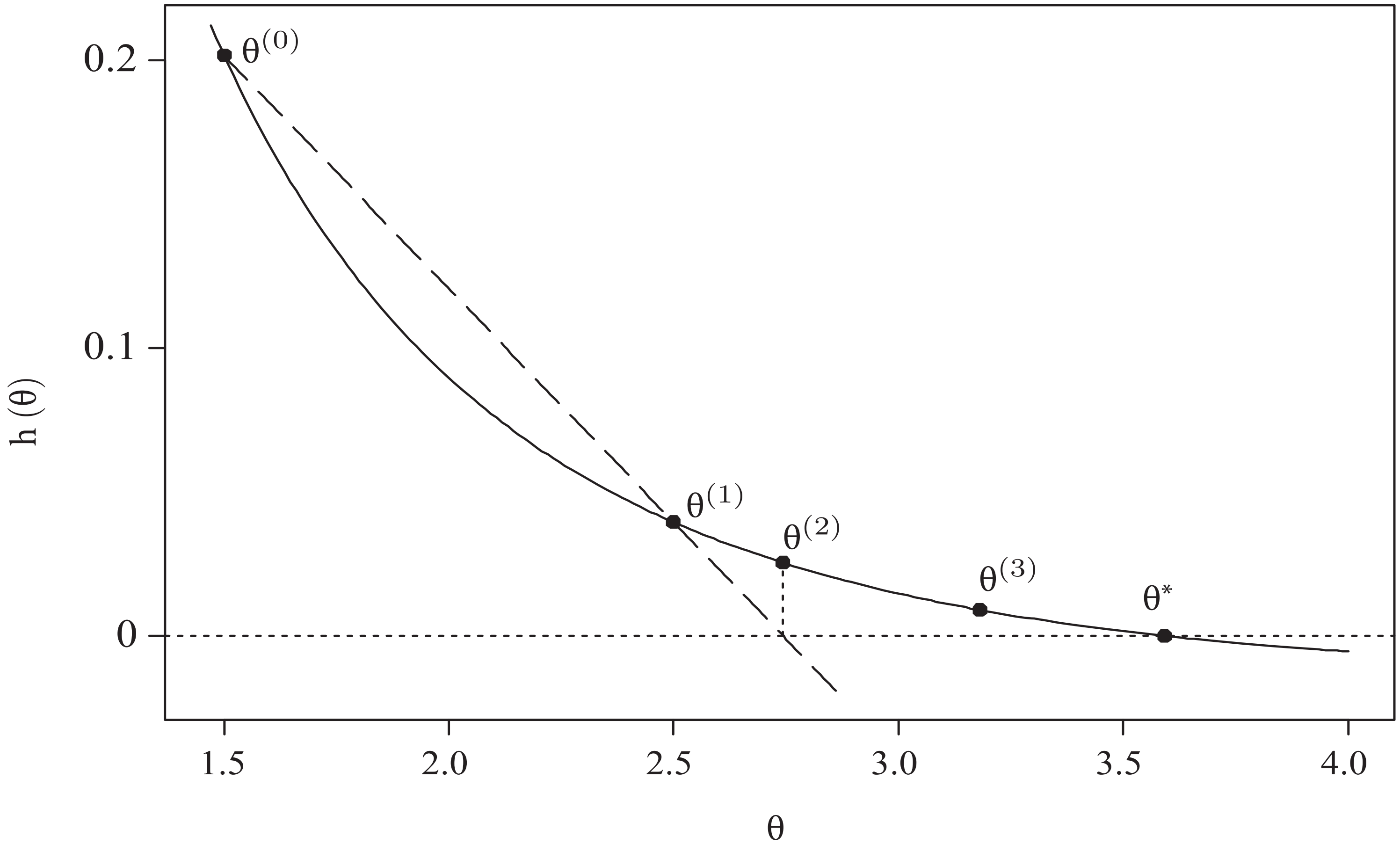
$$\theta^{(t+1)} = \theta^{(t)} - h(\theta^{(t)}) \cdot \frac{\theta^{(t)} - \theta^{(t-1)}}{h(\theta^{(t)}) - h(\theta^{(t-1)})}.$$

- **Remarks:**

  – When $h(\theta) = g'(\theta)$, we have

$$\theta^{(t+1)} = \theta^{(t)} - g'(\theta^{(t)}) \cdot \frac{\theta^{(t)} - \theta^{(t-1)}}{g'(\theta^{(t)}) - g'(\theta^{(t-1)})}.$$

  – The secant method need two starting points $\theta^{(0)}$ and $\theta^{(1)}$.



Secant Method

- **Convergence Order:** The performance of a root-finding approach is typically measured by its order of convergence to the true root θ^*. Let $\epsilon^{(t)} = \theta^{(t)} - \theta^*$. A method has *convergence order* β if $\lim_{t\to\infty} \epsilon^{(t)} = 0$ and

$$\lim_{t\to\infty} \frac{|\epsilon^{(t)}|}{|\epsilon^{(t-1)}|^\beta} = c$$

for some constant $c > 0$ and $\beta > 0$.

- **Remarks:**
  - If a method has the convergence order β with constant c, then

$$|\epsilon^{(t)}| \approx c|\epsilon^{(t-1)}|^\beta \approx \cdots \approx c^{1+\cdots+\beta^{t-1}} \cdot |\epsilon^{(0)}|^{\beta^t}.$$

  - When $\beta = 1$ and $c < 1$, we call the method has a *linear convergence order* .

- – For the **bisection method**, $\lim_{t\to\infty} \frac{|\epsilon^{(t)}|}{|\epsilon^{(t-1)}|^\beta}$ may not exist for any β, but we have $\frac{|\epsilon^{(t)}|}{|\epsilon^{(t-1)}|} \approx \frac{|b_t - a_t|}{|b_{t-1} - a_{t-1}|} = 0.5$. It approximately has a linear convergence order.

  – For the **Newton's method**,

  * We have

$$\epsilon^{(t+1)} = \theta^{(t+1)} - \theta^* = \theta^{(t)} - \theta^* - h(\theta^{(t)})/h'(\theta^{(t)}).$$

  * Note that

$$0 = h(\theta^*) = h(\theta^{(t)}) + h'(\theta^{(t)})(\theta^* - \theta^{(t)}) + \frac{1}{2}h''(u_t)(\theta^* - \theta^{(t)})^2,$$

where u_t is a point between $\theta^{(t)}$ and θ^*. Then

$$\theta^{(t)} - \theta^* = \left[h(\theta^{(t)}) + \frac{1}{2}h''(u_t)(\theta^* - \theta^{(t)})^2\right]/h'(\theta^{(t)}).$$

- – * Thus,

$$\epsilon^{(t+1)} = \theta^{(t)} - \theta^* - h(\theta^{(t)})/h'(\theta^{(t)}) = \frac{1}{2}h''(u_t)(\epsilon^{(t)})^2/h'(\theta^{(t)}).$$

  * Suppose we already know that $\theta^{(t)} \to \theta^*$ as $t \to \infty$, then under some regulation conditions,

$$\lim_{t\to\infty} \frac{|\epsilon^{(t+1)}|}{|\epsilon^{(t)}|^2} \to \frac{1}{2}|h''(\theta^*)|/|h'(\theta^*)|,$$

    and the Newton's method has a convergence order $\beta = 2$ when $h'(\theta^*) \neq 0$.

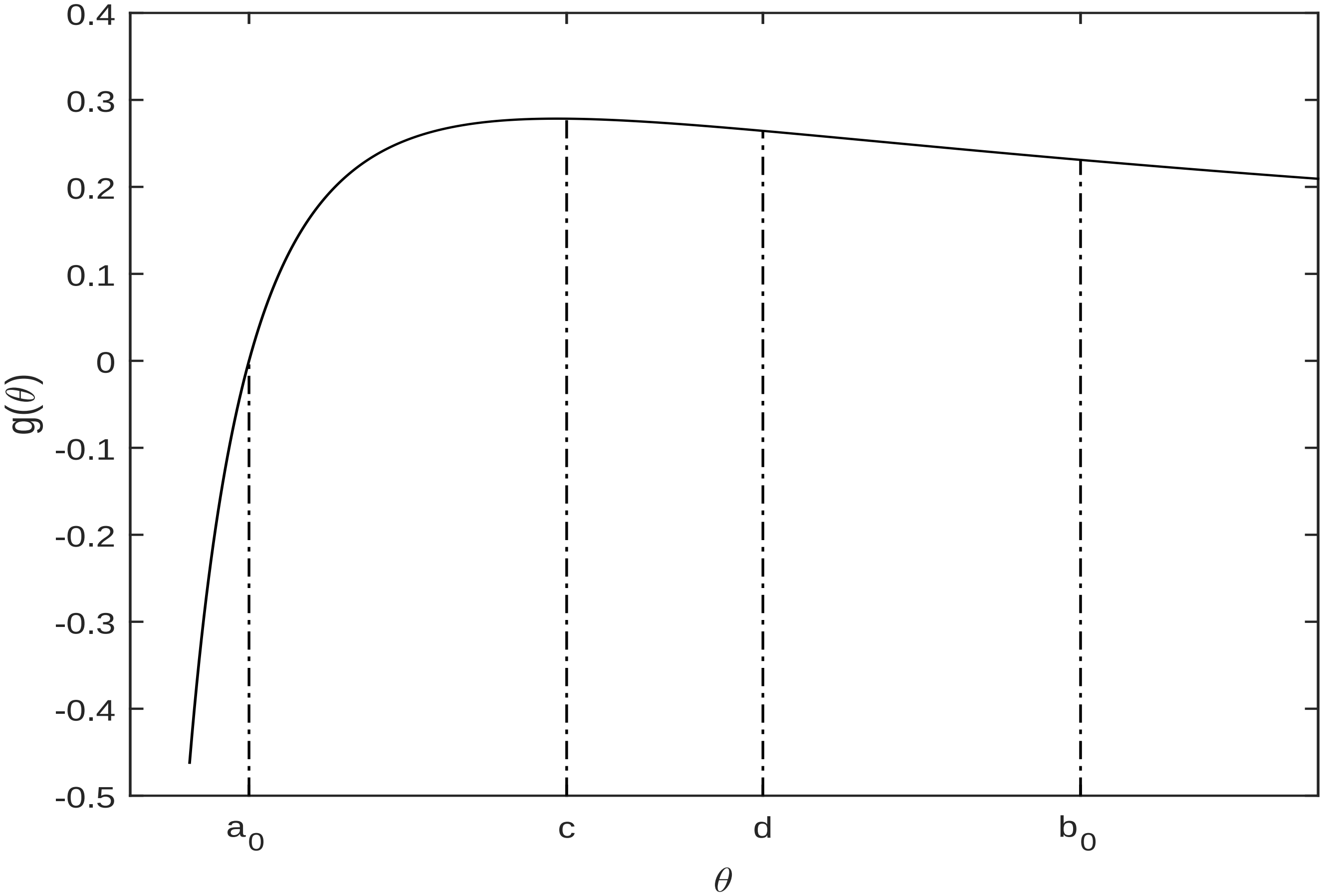
  – Under certain conditions, we can show the **secant method** has a convergence order $\beta = (1 + \sqrt{5})/2 \approx 1.62$.

  – Under certain conditions, the Newton's method converges faster than the secant method and the bisection method.

1.2 Univariate Problems

- Now we consider finding $\theta^* = \arg\max_\theta g(\theta)$ when $g(\theta)$ is not differentiable or when $h(\theta) = g'(\theta)$ is difficult to calculate.

- **Golden Section Search Method:** Suppose we know that $g(\theta)$ is concave in $[a_0, b_0]$ and its maximizer $\theta^* \in (a_0, b_0)$. Let $c < d$ be two points within the interval (a_0, b_0).
  - If $g(c) \leq g(d)$, then the maximizer $\theta^* \in (c, b_0)$. **(Why?)**
  - If $g(c) \geq g(d)$, then the maximizer $\theta^* \in (a_0, d)$.
  - We take $c = a_0 + (1-\rho)(b_0 - a_0) = \rho a_0 + (1-\rho) b_0$ and $d = a_0 + \rho(b_0 - a_0) = (1-\rho)a_0 + \rho b_0$, where $\rho = \frac{\sqrt{5}-1}{2} \approx 0.618$. Here c and d are golden section points of the interval (a_0, b_0).
  - Note that c is a golden section point of (a_0, d) and d is a golden section point of (c, b).



Golden Section Search Method

- **Golden Section Search Method:**
  - When $t = 0$, set $c = \rho a_0 + (1-\rho)b_0$ and $d = (1-\rho)a_0 + \rho b_0$. Compute $g_c = g(c)$ and $g_d = g(d)$.
  - Until $|b_t - a_t| < \epsilon$:
    * If $g_c < g_d$, let $a_{t+1} \leftarrow c$, $b_{t+1} \leftarrow b_t$, $c \leftarrow d$ and $g_c \leftarrow g_d$. Then set $d = (1-\rho)a_{t+1} + \rho b_{t+1}$ and compute $g_d = g(d)$.
    * Else, let $a_{t+1} \leftarrow a_t$, $b_{t+1} \leftarrow d$, $d \leftarrow c$ and $g_d \leftarrow g_c$. Then set $c = \rho a_{t+1} + (1-\rho)b_{t+1}$ and compute $g_c = g(c)$.
    * $t \leftarrow t+1$.
  - Let $\hat{\theta}^* = (a_t + b_t)/2$.
- **Remark:** We have $|b_{t+1} - a_{t+1}|/|b_t - a_t| = \rho \approx 0.618$. The golden section search method approximately has a linear convergence order.

1.3 Multivariate Problems

- **Multivariate Maximization:** We want to find a point $\theta^* \in \Theta \subset \mathbb{R}^p$ to maximize the objective function $g(\theta)$, where the dimension of θ is $p > 1$.

- **Line Search Strategy:** Suppose we have $\theta^{(t)}$ at iteration t. Then we choose a direction $u^{(t)}$, and search along this direction from $\theta^{(t)}$ for a new point with a larger function value. Particularly, we can let

$$\alpha_t = \arg\max_{\alpha > 0} g(\theta^{(t)} + \alpha u^{(t)})$$

and set

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t u^{(t)}.$$

- **Remarks:**
  - Consider the Taylor series expansion $g\big(\theta^{(t)} + \alpha u^{(t)}\big) = g(\theta^{(t)}) + \alpha \nabla g(\theta^{(t)})^T u^{(t)} + o(\alpha)$. We should choose $u^{(t)}$ so that $\nabla g(\theta^{(t)})^T u^{(t)} > 0$.

1.3 Multivariate Problems

- – We can let $u^{(t)} = \frac{g(\theta^{(t)})}{\partial \theta_k} \cdot e_k$ if $t = lp + k$ for integers l and k, where $e_k = (0, \cdots, 0, 1, 0, \cdots, 0)$ is a $p \times 1$ vector with the kth entry being 1 and other entries being 0. It is called the *coordinate ascent method.*

  – Suppose $u^{(t)}$ is a unit direction, that is, $\| u^{(t)} \|_2 = 1$. Then

$$\begin{aligned} g\big(\theta^{(t)} + \alpha u^{(t)}\big) &\approx g(\theta^{(t)}) + \alpha \nabla g(\theta^{(t)})^T u^{(t)} \\ &\leq g(\theta^{(t)}) + \alpha \| \nabla g(\theta^{(t)}) \|_2 . \end{aligned}$$

  The value of $g\big(\theta^{(t)} + \alpha u^{(t)}\big)$ has the most rapid decrease when $u^{(t)} = \frac{\nabla g(\theta^{(t)})}{\|\nabla g(\theta^{(t)})\|_2}$. If we set $u^{(t)} = \nabla g(\theta^{(t)})$, the method is called the *steepest ascent method.*

  – If we want to solve a **minimization problem**, we should let $u^{(t)} = -\nabla g(\theta^{(t)})$ and the method is called the *steepest descent method.*

1.3 Multivariate Problems

- **Newton's Method:** Suppose we have $\theta^{(t)}$ at iteration t. Consider

$$\begin{aligned}
& g(\theta) \\
\approx\ & g(\theta^{(t)}) + \nabla g(\theta^{(t)})^T(\theta - \theta^{(t)}) + \frac{1}{2}(\theta - \theta^{(t)})^T \nabla^2 g(\theta^{(t)})(\theta - \theta^{(t)}) \\
=\ & \frac{1}{2}\Big\{\theta - \theta^{(t)} + \big[\nabla^2 g(\theta^{(t)})\big]^{-1}\nabla g(\theta^{(t)})\Big\}^T \nabla^2 g(\theta^{(t)})\Big\{\theta - \theta^{(t)} + \big[\nabla^2 g(\theta^{(t)})\big]^{-1}\nabla g(\theta^{(t)})\Big\} \\
& \qquad -\frac{1}{2}\nabla g(\theta^{(t)})^T\big[\nabla^2 g(\theta^{(t)})\big]^{-1}\nabla g(\theta^{(t)}) + g(\theta^{(t)}).
\end{aligned}$$

The Newton's method let

$$\theta^{(t+1)} = \theta^{(t)} - \big[\nabla^2 g(\theta^{(t)})\big]^{-1}\nabla g(\theta^{(t)}).$$

- **Remarks:**
  - If we want to solve a maximization (minimization) problem, we would expect $\nabla^2 g(\theta^{(t)})$ to be a negative (positive) definite matrix.

1.3 Multivariate Problems

- – The Newton's method does not require a line search step.

  – When $\nabla^2 g(\theta^{(t)})$ is negative definite, we have

$$\nabla g(\theta^{(t)})^T \Big[- \big[\nabla^2 g(\theta^{(t)})\big]^{-1} \nabla g(\theta^{(t)}) \Big] > 0.$$

  So we can also consider a line search strategy with $u^{(t)} = -\big[\nabla^2 g(\theta^{(t)})\big]^{-1} \nabla g(\theta^{(t)})$, that is,

$$\theta^{(t+1)} = \theta^{(t)} - \alpha_t \big[\nabla^2 g(\theta^{(t)})\big]^{-1} \nabla g(\theta^{(t)}).$$

  where

$$\alpha_t = \arg\max_{\alpha > 0} g\Big(\theta^{(t)} - \alpha \big[\nabla^2 g(\theta^{(t)})\big]^{-1} \nabla g(\theta^{(t)})\Big).$$

  Here $u^{(t)} = -\big[\nabla^2 g(\theta^{(t)})\big]^{-1} \nabla g(\theta^{(t)})$ is also called the *Newton direction*.

- **Example: Logistic Regression.** Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be independent and identically distributed random vectors, where $Y_i \in \{0, 1\}$ and X_i, $i = 1, \cdots, n$, are d-dimensional random vectors. Consider a logistic regression model

$$P(Y_i = 1 \,|\, X_i = x_i, \beta) = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}.$$

Given observations $(x_1, y_1), \cdots, (x_n, y_n)$, we want to estimate $\beta = (\beta_1, \cdots, \beta_d)^T$.

  – A *logit function* is defined as $\text{logit}(u) = \log\left(\frac{u}{1-u}\right)$ for $0 < u < 1$. The logistic regression model assumes

$$\text{logit}\big(P(Y_i = 1 \,|\, X_i = x_i; \beta)\big) = x_i^T \beta.$$

  – Note that the conditional probability mass function

$$f(y_i \,|\, x_i; \beta) := P(Y_i = y_i \,|\, X_i = x_i; \beta) = \frac{\big(\exp\{x_i^T \beta\}\big)^{y_i}}{1 + \exp\{x_i^T \beta\}}.$$

1.3 Multivariate Problems

- – The likelihood function can be written as

$$
\begin{aligned}
l(\beta) &= f(y_1, \cdots, y_n, x_1, \cdots, x_n; \beta) \\
&= \prod_{i=1}^{n} f(y_i \,|\, x_i; \beta)\, f(x_i) = \prod_{i=1}^{n} \frac{\left(\exp\{x_i^T \beta\}\right)^{y_i}}{1 + \exp\{x_i^T \beta\}} \cdot f(x_i),
\end{aligned}
$$

  and the log-likelihood function is

$$
\log l(\beta) = c + \sum_{i=1}^{n} y_i \cdot x_i^T \beta - \sum_{i=1}^{n} \log\left[1 + \exp\{x_i^T \beta\}\right].
$$

  – The gradient and Hessian of $\log l(\beta)$ are

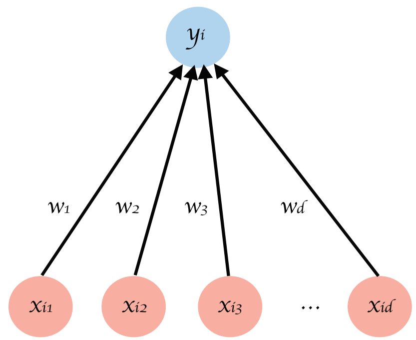
$$
\nabla \log l(\beta) = \sum_{i=1}^{n} y_i \cdot x_i - \sum_{i=1}^{n} \frac{1}{1 + \exp\{-x_i^T \beta\}} \cdot x_i,
$$

  and

$$
\nabla^2 \log l(\beta) = -\sum_{i=1}^{n} \frac{\exp\{-x_i^T \beta\}}{\left[1 + \exp\{-x_i^T \beta\}\right]^2} \cdot x_i x_i^T.
$$

  It is easy to see that $\log l(\beta)$ is concave. We can use the Newton's method to find $\hat{\beta}_{MLE}$.

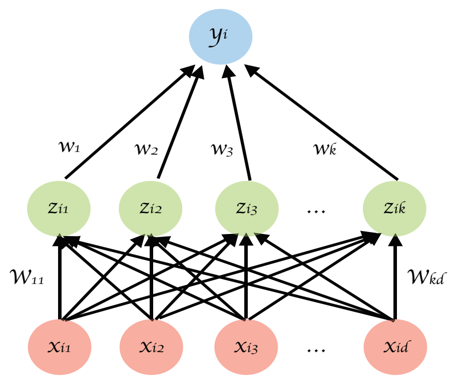
- **Example: Neural Network.** Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be independent and identically distributed random vectors, where $Y_i \in \{0, 1\}$ and $X_i = (X_{i1}, \cdots, X_{id})^T$, $i = 1, \cdots, n$, are d-dimensional random vectors. We want to find a model to fit $P(Y_i = 1 \,|\, X_i = x_i)$.
  - The logistic model assumes $P(Y_i = 1 \,|\, X_i = x_i) := y_i^* = h\big(\sum_{j=1}^d w_j x_{ij}\big)$, where $h(u) = 1/(1 + e^{-u}) = e^u/(1 + e^u)$ is called the *logistic function* or *sigmoid function* (in machine learning).



Logistic model (Julie Nutini, University of British Columbia)

1.3 Multivariate Problems

- – Neural network assumes a structure with multiple layers.
    - $*$ Input $x_i = (x_{i1}, \cdots, x_{id})^T$.
    - $*$ For $l = 1, \cdots, k$, compute $z_{il} = h\big(\sum_{j=1}^{d} W_{lj} x_{ij}\big)$.
    - $*$ Output $P(Y_i = 1 \mid X_i = x_i) = y_i^* = h\big(\sum_{l=1}^{k} w_l z_{il}\big)$.



Neural network (Julie Nutini, University of British Columbia)

1.3 Multivariate Problems

- – Training of a neural network:
  - ∗ The model parameters are $\theta = (w_1, \cdots, w_l, W_{11}, \cdots, W_{kd})^T$.
  - ∗ Given observations $(x_1, y_1), \cdots, (x_n, y_n)$, we want to find θ to maximize the log-likelihood function

$$
\begin{aligned}
\log l(\theta) &= \sum_{i=1}^{n} \log \left[f(y_i \,|\, x_i; \theta) f(x_i) \right] \\
&= c + \sum_{i=1}^{n} \log \left\{ \left[P(Y_i = 1 \,|\, X_i = x_i; \theta) \right]^{y_i} \left[P(Y_i = 0 \,|\, X_i = x_i; \theta) \right]^{1-y_i} \right\} \\
&= c + \sum_{i=1}^{n} \left[y_i \log P(Y_i = 1 \,|\, X_i = x_i; \theta) + (1 - y_i) \log P(Y_i = 0 \,|\, X_i = x_i; \theta) \right] \\
&= c + \sum_{i=1}^{n} \left[y_i \log y_i^* + (1 - y_i) \log(1 - y_i^*) \right].
\end{aligned}
$$

- – Define

  $$L_i(\theta) = y_i \log y_i^* + (1 - y_i)\log(1 - y_i^*),$$

  where $y_i^* := P(Y_i = 1 \mid X_i = x_i; \theta)$. Then $\log l(\theta) = c + \sum_{i=1}^n L_i(\theta)$.

  – **Training steps:**

  ∗ Repeat until some stopping criterion is satisfied:

  · For $i = 1, \cdots, n$, let

  $$\theta^{(t+1)} \leftarrow \theta^{(t)} + \alpha \nabla_\theta L_i(\theta^{(t)}),$$

  where α is a small positive number.

  – For a given θ, how to compute $\nabla_\theta L_i(\theta)$?

- – We can compute $\nabla_\theta L_i(\theta)$ using *backpropagation* (chain rule).
    - ∗ Compute $dL_i(\theta)/dy_i^* = y_i/y_i^* - (1-y_i)/(1-y_i^*)$.
    - ∗ For $l = 1, \cdots, k$, compute

$$\partial L_i(\theta)/\partial w_l = dL_i(\theta)/dy_i^* \cdot \partial y_i^*/\partial w_l,$$

where $\partial y_i^*/\partial w_l = h'\big(\sum_{l=1}^k w_l z_{il}\big) \cdot z_{il}$.

- ∗ For $l = 1, \cdots, k$,
    - · Compute

$$\partial L_i(\theta)/\partial z_{il} = dL_i(\theta)/dy_i^* \cdot \partial y_i^*/\partial z_{il},$$

where $\partial y_i^*/\partial z_{il} = h'\big(\sum_{l=1}^k w_l z_{il}\big) \cdot w_l$.

- · For $j = 1, \cdots, d$, compute

$$\partial L_i(\theta)/\partial W_{lj} = \partial L_i(\theta)/\partial z_{il} \cdot \partial z_{il}/\partial W_{lj},$$

where $\partial z_{il}/\partial W_{lj} = h'\big(\sum_{j=1}^d W_{lj} x_{ij}\big) \cdot x_{ij}$.

1.3 Multivariate Problems

- **Profile MLE:** Suppose the parameter $\theta = (\theta_1, \theta_2)$, where θ_1 is a $p_1 \times 1$ vector and θ_2 is a $p_2 \times 1$ vector. The dimension of θ is $p = p_1 + p_2$. We want to find the MLE

$$(\hat{\theta}_{1,MLE}, \hat{\theta}_{2,MLE}) = \arg\max_{\theta_1,\theta_2} \log l(\theta_1, \theta_2).$$

  - Assume that for any given θ_1, it is easy to find

$$\hat{\theta}_2(\theta_1) = \arg\max_{\theta_2} \log l(\theta_1, \theta_2).$$

  - Then we can reduce the p-dimensional optimization problem to a p_1-dimensional optimization problem, that is,

$$\hat{\theta}_{1,MLE} = \arg\max_{\theta_1} \log l\big[\theta_1, \hat{\theta}_2(\theta_1)\big]$$

and

$$\hat{\theta}_{2,MLE} = \hat{\theta}_2\big(\hat{\theta}_{1,MLE}\big).$$

- **Example:** Assume that $X_1, \cdots, X_n$ are i.i.d. following a Gamma(α, λ) distribution with the probability density function

$$f(x; \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} & x > 0, \\ 0 & \text{otherwise}, \end{cases}$$

where $\alpha > 0$, $\lambda > 0$ and $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.

  - The log-likelihood function is

$$\log l(\alpha, \lambda) = n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^n \log(X_i) - \lambda \sum_{i=1}^n X_i.$$

  - For each given $\alpha > 0$, it is easy to find

$$\hat{\lambda}(\alpha) = \arg\max_\lambda \log l(\alpha, \lambda) = \alpha / \overline{X}_n,$$

where $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$

1.3 Multivariate Problems

- – To find the MLE of α and λ, we need to solve the following maximization problem

$$\hat{\alpha}_{MLE} = \max_{\alpha} \Big\{ n\alpha \log \alpha - n\alpha \log \overline{X}_n \\ - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(X_i) - n\alpha \Big\}$$

  and let $\hat{\lambda}_{MLE} = \hat{\alpha}_{MLE} / \overline{X}_n$.

Homework

1. If $\mathcal{X} \subset \mathbb{R}^p$ is a convex set, prove that $\mathcal{Y} = \{y = Ax + b : x \in \mathcal{X}\}$ is also a convex set, where A is a $q \times p$ matrix and b is a $q \times 1$ vector.

2. Suppose $g_1(\theta)$ and $g_2(\theta)$ are convex functions. Prove that (1) for any $\alpha > 0$ and $\beta > 0$, $\alpha g_1(\theta) + \beta g_2(\theta)$ is convex; (2) $\max \big\{ g_1(\theta), g_2(\theta) \big\}$ is convex.

3. Let $g(\theta) = \frac{1}{2}\,\theta^T A\,\theta + b^T\theta + c$, where A is a $p \times p$ negative definite matrix, b is a $p \times 1$ vector, and c is a scalar. Find the maximum value and maximizer of $g(\theta)$.

4. The following data are i.i.d. samples from a Cauchy(θ,1), $-\infty < \theta < \infty$, distribution with the probability density function $f(x;\theta) = \frac{1}{\pi\left(1+(x-\theta)^2\right)}$, $-\infty < x < \infty$:

   1.77, -0.23, 2.76, 3.80, 3.47, 56.75, -1.34, 4.24, -2.44, 3.29, 3.71, -2.40, 4.53, -0.07, -1.05, -13.87, -2.53, -1.75, 0.27, 43.21.

   (a) Graph the log-likelihood function. Find the MLE for θ using the Newton's method. Try all of the following starting points: -11, -1, 0, 1.5, 8 and 38. Discuss your results.

   (b) Apply the bisection method with starting points -1 and 1. Use additional runs to illustrate manners in which the bisection method may fail to find the global maximum.

   (c) From starting values of $(\theta^{(0)}, \theta^{(1)}) = (-2, -1)$, apply the secant method to estimate θ. What happens when $(\theta^{(0)}, \theta^{(1)}) = (-3, 3)$?

5. Consider the probability density function $f(x;\theta) = \big(1 - \cos(x - \theta)\big)/2\pi$ for $0 \le x \le 2\pi$, where θ is a parameter between $-\pi$ and π. The following i.i.d. data arise from this density: 3.91, 4.85, 2.28, 4.06, 3.70, 4.04, 5.46, 3.53, 2.28, 1.96, 2.53, 3.88, 2.22, 3.47, 4.82, 2.46, 2.99, 2.54, 0.52, 2.50. We want to estimate θ.

   (a) Graph the log likelihood function between $-\pi$ and π.

   (b) Find the method of moments estimator of θ.

   (c) Find the MLE for θ using the Newton's method, using the result from (b) as the starting value. What solutions do you find when you start at -2.7 and 2.7?

Homework

6. Suppose (X_{1i}, X_{2i}, Y_i), $i = 1, \cdots, n$, are i.i.d. following the logistic model

$$P(Y_i = 1 \mid X_{1i}, X_{2i}; \beta) = \frac{\exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}\}}{1 + \exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}\}}.$$

Find the MLE of $\beta = (\beta_0, \beta_1, \beta_2)$.