OR4030 OPTIMIZATION Chapter 1
Some Optimization Models

1.1 Least Squares Data Fitting

Suppose that in a model, the relationship between two variables b and t are governed by a quadratic function:

$$b(t) = x_1 + x_2 t + x_3 t^2.$$

We need to use some experimental data $(t_1, b_1), (t_2, b_2), \ldots,$ (t_m, b_m) to determine x_1, x_2 and x_3.

If $m = 3$, and

$$\begin{cases} (t_1, b_1) = (2, 1) \\ (t_2, b_2) = (3, 6) \\ (t_3, b_3) = (5, 4) \end{cases}$$

then

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 5 & 25 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 4 \end{bmatrix}.$$

Solving the above linear equations gives

$$(x_1, x_2, x_3) = (-21, 15, -2),$$

so that

$$b(t) = -21 + 15t - 2t^2.$$

It is common that we collect more data points than the number of variables in the model. And it is expected that each of the measurements will be in error.

For example, if we also have

$$(t_4, b_4) = (7, -15),$$

and if we use the first three data points to determine (x_1, x_2, x_3) and $b(t)$, then

$$b(7) = -21 + 15 \cdot 7 - 2 \cdot 7^2 = -14 \neq -15.$$

Since each data point may have some error, we must use them "collectively"

Define the residual vector

$$r = b - Ax = \begin{bmatrix} b_1 - (x_1 + x_2 t_1 + x_3 t_1^2) \\ b_2 - (x_1 + x_2 t_2 + x_3 t_2^2) \\ \cdot \\ \cdot \\ \cdot \\ b_m - (x_1 + x_2 t_m + x_3 t_m^2) \end{bmatrix},$$

where

$$r = \begin{bmatrix} r_1 \\ \cdot \\ \cdot \\ \cdot \\ r_m \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 1 & t_1 & t_1^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & t_m & t_m^2 \end{bmatrix},$$

and make r as small as possible. As r is a vector, we specify the request of making r small as letting some norm of r small.

Least Squares Data Fitting

Least squares data fitting is the most commonly used approach to make r small, i.e., to make $\|r\|_2$ the smallest:

$$\min_x \sum_{i=1}^{m} r_i^2 = \sum_{i=1}^{m} [b_i - (x_1 + x_2 t_i + x_3 t_i^2)]^2.$$

The right hand side of the above equality is a non-linear function of x. Hence, the problem is a nonlinear optimization problem.

In this model, in the function

$$b(t) = x_1 + x_2 t + x_3 t^2,$$

all x_j occur linearly, and hence the model is called a linear least squares model.

Other Examples of Linear Model

$$b(t) = x_1 + x_2 \sin t + x_3 \sin 2t + \ldots + x_{k+1} \sin kt.$$

or

$$b(t) = x_1 + \frac{x_2}{1+t^2}.$$

Note that even though the functions of t are nonlinear, variables x_1, x_2 and x_3 appear linearly.

Nonlinear Least Squares Model: Some $x_1, x_2, \ldots x_k$ in function $b(t)$ occur nonlinearly.

Examples Consider the following models where

$$b(t) = x_1 + x_2 e^{x_3 t} + x_4 e^{x_5 t}, \tag{1}$$

or

$$b(t) = x_1 + \frac{x_2}{1 + x_3 t^2}.$$

For the model (1), the least squares problem is

$$\min_x \sum_{i=1}^{m} r_i(x)^2,$$

where

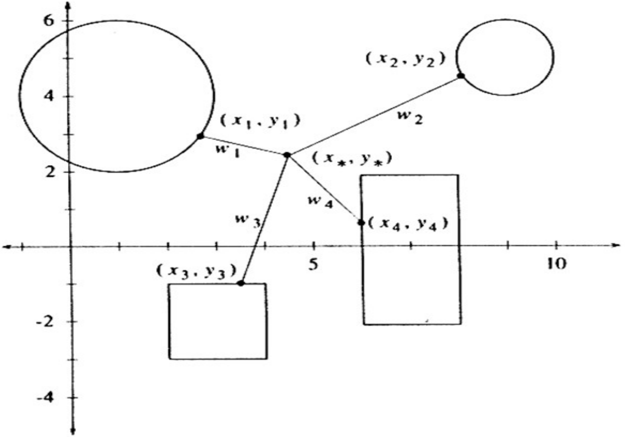
$$r_i(x) = b_i - (x_1 + x_2 e^{x_3 t_i} + x_4 e^{x_5 t_i}).$$

The minimization problem is a nonlinear optimization problem. Note that no matter the least squares (LS) problem is a linear LS model, or a nonlinear LS model, it is always a nonlinear optimization problem. But we shall see later that linear LS problem is easier to solve.

1.2 Nonlinear Programming

In this course, nonlinear programming means optimization of a function $f(x)$ subject to some constraints of the type $g_i(x) = 0$ or $h_j(x) \leq 0$, where at least one function (either the objective function f, or a constraint function g_i or h_j) is nonlinear.

Example 1 The figure below gives locations of two round buildings and two rectangular shaped buildings. Find a point (x^*, y^*) such that the total distance from (x^*, y^*) to all buildings is minimum. (here the distance from a point to a building means the distance from the point to the closest point in the building)


6
4
2
-2
-4
5
10
(x_1, y_1)
(x_2, y_2)
(x_3, y_3)
(x_4, y_4)
(x_*, y_*)
w_1
w_2
w_3
w_4

The model can be formulated as:

$$\begin{aligned}
\min \quad & w_1 + w_2 + w_3 + w_4 \\
s.t. \quad & w_i = \sqrt{(x_i - x^*)^2 + (y_i - y^*)^2}, \quad i = 1, 2, 3, 4 \\
& (x_1 - 1)^2 + (y_1 - 4)^2 \leq 4, \\
& (x_2 - 9)^2 + (y_2 - 5)^2 \leq 1, \\
& 2 \leq x_3 \leq 4, \quad -3 \leq y_3 \leq -1, \\
& 6 \leq x_4 \leq 8, \quad -2 \leq y_4 \leq 2.
\end{aligned}$$

In the model (x_i, y_i) represents the closest point in building i $(i = 1, 2, 3, 4)$ to the point (x^*, y^*). They are also unknown variables, and should be determined by this minimization model.

Example 2 Consider a portion (or say, a segment) of a sphere shown in the graph below. As we know, the volume of the segment is $\pi h^2(r - \frac{h}{3})$, and its surface area is $2\pi rh$. (without including the bottom part)

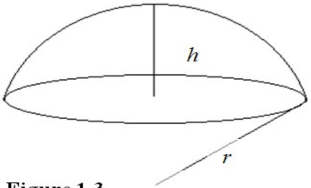


Figure 1.3
Archimedes' problem

The problem is how to choose r and h so as to maximize the volume of the segment, but where the surface area A of the segment is fixed., i.e., A is a given constant.

This problem can be formulated as the following model:

$$\begin{array}{ll} \max & v(r,h) = \pi h^2 (r - \frac{h}{3}) \\ s.t. & 2\pi rh = A. \end{array}$$

Here $v(r,h) = \pi h^2 (r - \frac{h}{3})$ is a nonlinear objective function of r and h, and $2\pi rh = A$ is a nonlinear equality constraint of r and h. So, in this problem both objective function and the constraint are nonlinear.

Example 3 (Portfolio selection). We know that
security: stock, bond, ... ;
portfolio: a collection of securities.

For each security j, its return r_j is in fact a random variable, and let

- μ_j: the expected annual return, i.e., the mean of the random variable r_j;
- v_{jj}: the variance of the random variable r_j (large variance means the return of security j is unstable);
- v_{ij}: the covariance of the returns between securities i and j (large v_{ij} means the performances of the two securities are closely related).

It is well known that diversifying investment can reduce risk. So, in stead of buying a single security, we now establish a portfolio by investing on several securities. Suppose all together there are N available securities, and denote

- x_j: the number of shares of security j to be invested;
- a_j: current price per share of security j.

Suppose a portfolio P consists of

$$x_j \text{ units of security } j, \quad j = 1, 2, \ldots, N,$$

and let

$\bar{r}_P$: the expected return of the portfolio P, which equals

$$\bar{r}_P = \sum_{j=1}^{N} x_j \mu_j = \mu^T x;$$

σ_P: the standard deviation of the portfolio P, i.e.,

$$\sigma_P = (\sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j v_{ij})^{\frac{1}{2}} = (x^T V x)^{\frac{1}{2}},$$

where

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ v_{21} & v_{22} & \cdots & v_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ v_{N1} & v_{N2} & \cdots & v_{NN} \end{bmatrix}.$$

In the above model, σ_P or $\sigma_P^2 = x^T V x$ can measure the risk level.

Generally, investors want to have

- large $\bar{r}_P$ (high expected return)
- small σ_P^2 (low risk)

for their portfolios. But these are two conflict targets.

Let the total available budget be B, the portfolio selection model can be formulated in several ways:

(1) Maximize a combination of $\mu^T x$ and $-x^T Vx$:

$$\begin{aligned} \max \quad & \mu^T x - \alpha x^T Vx \\ s.t. \quad & a^T x \leq B \\ & x \geq 0, \end{aligned}$$

where α is a weight coefficient to measure the relative importance of the two parts $\mu^T x$ and $-x^T Vx$. Note that

- $\alpha = 0$: risk is totally ignored;
- large value of α: risk is heavily concerned.

(2) Minimize $x^T Vx$ while setting a lower bound for the return:

$$\begin{aligned} \min \quad & x^T Vx \\ s.t. \quad & \mu^T x \geq p \\ & a^T x \leq B \\ & x \geq 0, \end{aligned}$$

where p stands for the minimum acceptable annual return.

(3) Maximize $\mu^T x$ while setting an upper bound for the risk level:

$$\begin{aligned} \max \quad & \mu^T x \\ s.t. \quad & x^T Vx \leq q \\ & a^T x \leq B \\ & x \geq 0, \end{aligned}$$

where q stands for the maximum acceptable risk level.

The model may have more restrictions. For example, if no more than 5% of total investment can be put into each individual security, then we should add constraints:

$$\frac{a_j x_j}{B} \leq 0.05, \quad j = 1, 2, \ldots, N$$

i.e.,

$$a_j x_j \leq 0.05B, \quad j = 1, 2, \ldots, N.$$

The three models are all nonlinear programming problems.

The above method for portfolio selection was first suggested by H.M. Markowitz in 1959, who won 1990 Nobel Prize for economics due to this contribution to the development of finance and economics.

Example 4 (Minimize the total travel time).
Let

- t_{ij}: the travel time between points i and j when the traffic is light;
- x_{ij}: the number of cars on the road between points i and j;
- c_{ij}: the capacity of the road between i and j;
- α_{ij}: a coefficient to measure how rapidly the travel time increases as the traffic gets heavier.

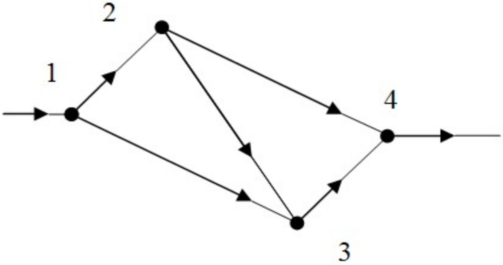
According to transportation science, the travel time between points i and j when the number of cars on the road is x_{ij} can be estimated by the formula:

$$T_{ij}(x_{ij}) = t_{ij} + \alpha_{ij}\frac{x_{ij}}{1 - \frac{x_{ij}}{c_{ij}}}.$$

Two extreme cases:
when $x_{ij} = 0$, the travel time is t_{ij};
when $x_{ij} = c_{ij}$, the travel time tends to $+\infty$.

Now suppose in the network below (see Figure 1.4), the total number of cars through the network (i.e., from point 1 to point 4) is X. How should we distribute the cars traveling on each road, so that the total travel time of all cars is minimized?

Figure 1.4
Traffic network



The model can be expressed as

$$\begin{aligned}
\min \quad & f(x) = \sum_{(i,j)} x_{ij} T_{ij}(x_{ij}) \\
s.t. \quad & x_{1,2} + x_{1,3} = X \\
& x_{2,3} + x_{2,4} - x_{1,2} = 0 \\
& x_{3,4} - x_{1,3} - x_{2,3} = 0 \\
& x_{2,4} + x_{3,4} = X \\
& T_{ij}(x_{ij}) = t_{ij} + \alpha_{ij} \frac{x_{ij}}{1 - \frac{x_{ij}}{c_{ij}}}, \text{for } ij = 12, 13, 23, 24, 34 \\
& x_{ij} \geq 0, \text{for } ij = 12, 13, 23, 24, 34.
\end{aligned}$$

This problem is again a nonlinear programming problem as the functions $T_{ij}(x_{ij})$ and $f(x)$ are nonlinear functions.

Example 5 (Maximum likelihood estimation).
The method of maximum likelihood has been regarded as one of the most effective methods of estimating distribution parameters.

Let $x^1, \dots x^N$ be a set of samples from a population with probability density function $g(x, \theta)$, where $\theta = (\theta_1, \dots, \theta_s)^T$ are distribution parameters. $\theta \in D$, where D is the parameter space in R^s.

We need to estimate the values of $\theta_1, \ldots, \theta_s$ by the samples. For the purpose, we introduce the likelihood function

$$L(\theta) = \prod_{i=1}^{N} g(x^i, \theta)$$

and find the maximum likelihood estimate $\hat{\theta}$ which maximizes the function $L(\theta)$:

$$L(\hat{\theta}) = \max_{\theta \in D} L(\theta).$$

or equivalently, we can maximize the logarithmic likelihood function

$$LL(\theta) = \sum_{i=1}^{N} \log g(x^i, \theta).$$

A special probability density function is a Weibull distribution with three parameters: location parameter δ, scale parameter β, and shape parameter α:

$$g(x; \delta, \beta, \alpha) = \frac{\alpha}{\beta^\alpha}(x - \delta)^{\alpha-1} exp\{-[(x - \delta)/\beta]^\alpha\}.$$

We see that maximizing its likelihood and logarithmic likelihood functions

$$L(\delta, \beta, \alpha) = \prod_{i=1}^{N} g(x^i; \delta, \beta, \alpha),$$

and

$$LL(\delta, \beta, \alpha) = \sum_{i=1}^{N} \log g(x^i; \delta, \beta, \alpha)$$

are both nonlinear optimization problems.

Example 6 (A two-class classification problem).

Suppose there is a set of historical data $\{(x^1, y_1), (x^2, y_2), \ldots, (x^N, y_N)\}$, where each $x^i \in R^m$ and $y_i \in R^1$, $i = 1, \ldots, N$. The m-dimensional vectors x represent some measurable factors, whereas each y has only two possible values: 1 or -1 which refers to two possible outcomes: either the result is in class one ($y = 1$) or in class two ($y = -1$).

We call these $\{(x^i, y_i)\}$ training data, and we want to find a line (if m=2), or a plane (if m=3), or generally a hyperplane in space R^m

$$(w, x) + b = 0$$

that separates the two classes of data points (here the symbol (w, x) represents the scalar product of the two vectors w and x).

That is,
$(w, x^i) + b > 0$, if the corresponding $y_i = 1$;
$(w, x^i) + b < 0$, if the corresponding $y_i = -1$.

The plane is called a separating plane. Once such a plane is found, if later we obtain another input x, what should the associate y be, 1 or -1?

We can answer that

$$y = sign\{(w, x) + b\},$$

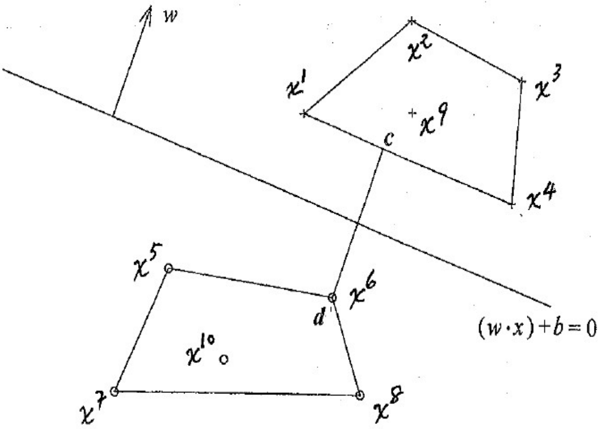
that is, $y = 1$ if $(w, x) + b > 0$ and -1 otherwise.

There are several methods to obtain such a separating plane. Here we introduce a method, called halving the shortest distance method. Its main idea is

1. find the closest points between convex hulls of the two classes of points, say points c and d;
2. link the two points to obtain line segment (c, d) and obtain its middle point M;
3. then take the plane vertical to line segment (c, d) and passing through point M as the wanted separating plane.

(we may use the graph on next page to understand the idea)

The method can be realized by the algorithm on page 31.


w
x^1
x^2
x^3
x^4
x^9
c
x^5
x^6
d
x^{10}
x^7
x^8
$(w \cdot x) + b = 0$

1. Formulate and solve the following optimization problem:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}||\sum_{y_i=1} \alpha_i x^i - \sum_{y_i=-1} \alpha_i x^i||^2, \\ s.t. \quad & \sum_{y_i=1} \alpha_i = 1, \quad \sum_{y_i=-1} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq 1, \quad i = 1, \ldots, N \end{aligned}$$

   and obtain an optimal solution $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots \hat{\alpha}_N)^T$.

2. Obtain the closest points between the two convex hulls:

$$c = \sum_{y_i=1} \hat{\alpha}_i x^i, \quad d = \sum_{y_i=-1} \hat{\alpha}_i x^i.$$

3. Construct the separating hyperplane

$$(\hat{w}, x) + \hat{b} = 0,$$

   where

$$\hat{w} = c - d = \sum_{i=1}^{N} y_i \hat{\alpha}_i x^i, \quad \hat{b} = -\frac{1}{2}((c-d), (c+d)).$$

We see that we have used the training set of data to establish the separating plane.

To check if a new input x belongs to class $y = 1$ or class $y = -1$, we just let

$$y = sign\{(\hat{w}, x) + \hat{b}\}.$$

This type of methods to find separating hyperplane or surface is called **support vector machine** which forms an important part of the subject **data mining**. Actually the key step is step 1, i.e., to solve a nonlinear optimization problem (more particularly, a quadratic optimization problem).

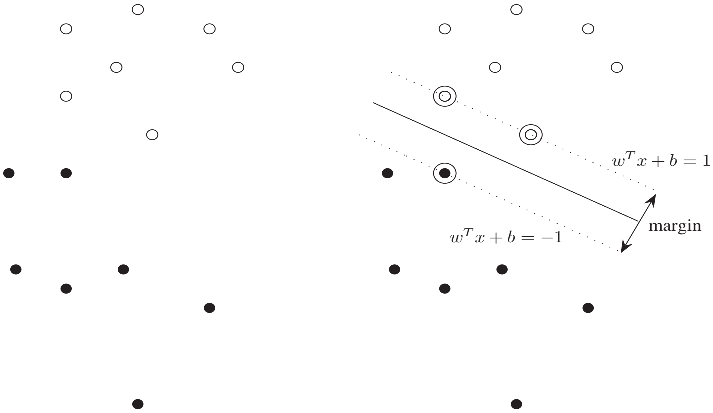
Another Model for the Classification Problem

To better separate the two groups of data points, we wish to have a band area centered at the separating plane $(\hat{w}, x_i) + \hat{b} = 0$, such that no data points would be in the area， see the graph on next page. That is, we wish that

$$\begin{aligned} (\hat{w}, x_i) + \hat{b} &\geq +k, \qquad \text{for } y_i = +1, \\ (\hat{w}, x_i) + \hat{b} &\leq -k, \qquad \text{for } y_i = -1. \end{aligned}$$

Without loss of generality, we may assume the number k to be 1, because if there is any number k satisfying the above two inequalities, then by changing $\hat{w}$ to a new one $w = \hat{w}/k$ and $\hat{b}$ to a new constant $b = \hat{b}/k$, the data points would satisfy

$$\begin{aligned} (w, x_i) + b &\geq +1, \qquad \text{for } y_i = +1, \\ (w, x_i) + b &\leq -1, \qquad \text{for } y_i = -1. \end{aligned}$$


$w^Tx + b = 1$
margin
$w^Tx + b = -1$

The width of the band is called its margin (see the graph). According to the fact that the normal directions of these planes are the same, which is w, we can obtain that the margin is $2/\|w\|$. If the margin is large, the two sets of data points would be more clearly separated. So, we wish to maximize $1/\|w\|$, or equivalently minimize $\|w\|^2$. Now the second model to determine the separating plane is:

$$\begin{array}{ll} \min & \|w\|^2 \\ s.t. & (w, x_i) + b \geq +1, \quad \text{for } y_i = +1, \\ & (w, x_i) + b \leq -1, \quad \text{for } y_i = -1. \end{array}$$

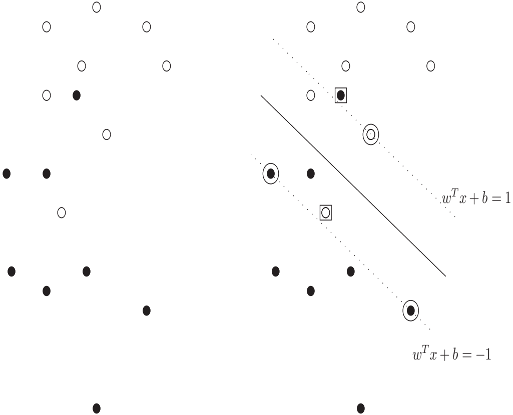
It can be proved that the two methods actually obtain the same plane.

Approximately Linearly Separable Case

So far we have assumed that the two data sets are linearly separable, i.e., there exists a plane to separate them exactly. But it is possible that such plane does not exist, see the next graph. Under such case, we need to relax the request, and let the plane separate as more data points correctly as possible, but with some misclassified data points.

We allow some points to violate the separation constraints, but impose a penalty for the violation. Let the nonnegative amount ξ_i be the amount by which the point x_i violates the constraint:

$$\begin{aligned} (w, x_i) + b &\geq +1 - \xi_i, \qquad \text{for } y_i = +1, \\ (w, x_i) + b &\leq -1 + \xi_i, \qquad \text{for } y_i = -1. \end{aligned}$$


$w^T x + b = 1$
$w^T x + b = -1$

The total violation can be measured by $\sum_i \xi_i$. As we want to minimize both $\|w\|^2$ and $\sum_i \xi_i$, the model can be formulated as:

$$\begin{aligned}
\min \quad & \|w\|^2 + C \sum_i \xi_i \\
s.t. \quad & (w, x_i) + b \geq +1 - \xi_i, \quad \text{for } y_i = +1, \\
& (w, x_i) + b \leq -1 + \xi_i, \quad \text{for } y_i = -1, \\
& \text{all } \xi_i \geq 0.
\end{aligned}$$

C is a control parameter. The larger C is, the larger the penalty for violating the separation is imposed. In practice, we may try several values of C to get a suitable one.

Note that if in the solution, $\sum_i \xi_i > 0$, it means that there is no true separating plane, that is, the two data sets are in fact linearly inseparable, and the obtained plane $(w, x) + b = 0$ is only an approximate separating plane with some errors.