Introduction

George Michailidis

gmichail@ucla.edu

STAT 102B

Optimization in Statistics and Machine Learning

Optimization problems are at the core of Statistics and Machine Learning

In courses focusing on statistical models, the first objective is to formulate the model $\mathcal{M}$; e.g., linear/logistic regression, principal components, support vector machines, feed-forward neural network, etc.

The second objective is to translate model $\mathcal{M}$ into an optimization problem of the form

$$\min_{x \in D} f(x) \tag{1}$$

where $f(\cdot)$ is the objective function we aim to optimize x is the (multivariate) argument of the function f and D is the domain of interest of the argument x

The objective of STAT 102B is to learn/study algorithms of how to solve (1)

Remark on the notation

The lecture notes follow the standard notation in optimization as outlined in (1)

The algorithms studied in STAT 102B will be presented following this standard notation

However, the function $f(\cdot)$, the argument x and the domain D will be specified explicitly in illustrative examples of statistical and machine learning models

Why bother?

For all the standard and widely used in practice statistical/machine learning models, "efficient" algorithms to solve the corresponding optimization problem have been studied in depth and moreover they are available in `R/Python` code

Why bother?

For all the standard and widely used in practice statistical/machine learning models, "efficient" algorithms to solve the corresponding optimization problem have been studied in depth and moreover they are available in `R/Python` code

Selected reasons for studying optimization:

- For the same model $\mathcal{M}$, different algorithms may be available in the software package;
  each algorithm may be preferable (or work "better") in different instances
- The study of model $\mathcal{M}$ through an optimization lens can provide deeper insights about the properties of the model; typical example, the problem of multicollinearity in linear regression and the introduction of ridge regression to mitigate its impact
- Understanding details of the optimization procedure for model $\mathcal{M}$ can help postulate another model $\mathcal{M}'$, which may prove more useful in practice

Illustration: The linear regression model - I

A linear regression model with p predictor variables $x_1, \cdots, x_p$ and a response variable y is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i, \quad i = 1, \cdots, n, \tag{2}$$

where β_0 denotes the intercept, $\beta_j, j = 1, \cdots, p$ the regression coefficients for the p predictors and ϵ_i is a random noise term.

The random noise satisfies $\mathbb{E}(\epsilon_i) = 0,\ \mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$

For ease of presentation, assume that $\beta_0 = 0$

The regression model can be written in compact matrix form as

$$y = X\beta + \epsilon \tag{3}$$

wherein y is a n-dimensional column vector containing the n values of the response variable, X an $n \times p$ matrix containing the n values of the p predictors and ϵ an n-dimensional column vector

A Concrete Example - I

We have data on sales of $n = 200$ products (response variable) and 3 predictors: the amount of advertising about each of the n products on facebook, youtube and in newspapers

```
data("marketing", package = "datarium")
head(marketing, 4)
  youtube facebook newspaper sales
1  276.12    45.36     83.04 26.52
2   53.40    47.16     54.12 12.48
3   20.64    55.08     83.16 11.16
4  181.80    49.56     70.20 22.20
```

A concrete example - II

For this example, the formulation of the regression model $y = X\beta + \epsilon$ is as follows:

$$\begin{bmatrix} 26.52 \\ 12.48 \\ 11.16 \\ 22.20 \\ y_5 \\ \cdots \\ y_{200} \end{bmatrix} = \begin{bmatrix} 276.12 & 45.36 & 83.04 & \\ 53.40 & 47.16 & 54.12 & \\ 20.64 & 55.08 & 83.16 & \\ 181.80 & 49.56 & 70.20 & \\ x_{1,5} & x_{2,5} & x_{3,5} & \\ \cdots & \cdots & \cdots & \cdots \\ x_{1,200} & x_{2,200} & x_{3,200} & \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \cdots \\ \epsilon_{200} \end{bmatrix}$$

Illustration: The linear regression model - II

Objective:
Given a data set $\{y_i, X_i\}_{i=1}^n$ obtain an estimate $\hat{\beta}$

We have a statistical model $\mathcal{M}$ (given in (2)) and an objective to obtain an estimate $\hat{\beta}$ from data

Illustration: The linear regression model - II

Objective:
Given a data set $\{y_i, X_i\}_{i=1}^n$ obtain an estimate $\hat{\beta}$

We have a statistical model $\mathcal{M}$ (given in (2)) and an objective to obtain an estimate $\hat{\beta}$ from data

It is customarily to obtain β by solving the following optimization problem

$$\min_{\beta \in \mathbb{R}^p} f(\beta) \equiv \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ji})^2 \equiv \min_{\beta \in \mathbb{R}^p} \frac{1}{2n} ||y - X\beta||_2^2 \qquad (4)$$

where $\|\cdot\|$ denotes the vector ℓ_2 norm

Illustration: The linear regression model - III

In the standard notation of optimization problems given in (1), the linear regression problem in (4) takes the form

- $x \leftarrow \beta$
- $D \leftarrow \mathbb{R}^p$ (no constraints)
- $f(\beta) \leftarrow \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ji})^2 \equiv \frac{1}{2n} \|y - X\beta\|_2^2$

The objective function $f(\cdot)$ is known as the least squares criterion or squared error loss function or sum of squared errors

Digression: Squared error loss function

Note that (3) can be rewritten as

$$\epsilon = y - X\beta \tag{5}$$

The objective function then becomes

$$f(\beta) = \frac{1}{2n}\|\epsilon\|_2^2 \equiv \frac{1}{2n}\left(\epsilon^\top \epsilon\right) \equiv \frac{1}{2n}\sum_{i=1}^{n}\epsilon_i^2 \tag{6}$$

The last expression explains the name of squared error loss function or sum of squared errors function

We will discuss more about loss functions during the rest of the term

Illustration: The linear regression model - IV

The use of the least squares function in (4) yields the well known least squares estimator for $\hat{\beta}$ given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top y \tag{7}$$

One can obtain $\hat{\beta}$ for a particular data set using the `lm()` command in `R`

Of course, one can also obtain $\hat{\beta}$ in `R` by using the following command

```
beta=ginv(t(X)%*%X)%*%(t(X)%*%t(y))
```

i.e., translating the formula in (7) directly into code

or in `Python`

```
beta = np.linalg.inv(X.T @ X) @ X.T @ y
```

Illustration: The linear regression model - V

How do we obtain the least squares solution?

Next, its derivation is quickly reviewed that reveals several interesting facts of more general interest

But first, a brief overview of some useful facts about vectors and matrices that come in handy in the derivation of the least squares solution

Review of Useful Facts - Vectors and Matrices - I

An array z of n real numbers $z_1, \cdots, z_n$ ($z \in \mathbb{R}^n$) is called a column vector and is written as

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \cdots \\ z_n \end{bmatrix}$$

Further, $z^\top = [z_1 \cdots z_n]$ denotes its transpose (a row vector)

- A vector z can be represented geometrically as a directed line in n dimensions with components z_k along the k-th axis
- Vector multiplication by a scalar $a \in \mathbb{R}$:
  Consider a row vector z; then $az = [az_1 \cdots az_n]$
- Vector addition: consider two row vectors z, w both in $\mathbb{R}^n$. Then,

$$z + w = [z_1 + w_1 \cdots z_n + w_n]$$

Review of Useful Facts - Vectors and Matrices - II

- Inner product of two vectors z, w both in $\mathbb{R}^n$. Then,

$$\langle z, w \rangle = z^\top w = \sum_{i=1}^{n} z_i w_i$$

- (Euclidean) Length of a vector $z \in \mathbb{R}^n$:

$$||z||_2 = \sqrt{z_i^2 + \cdots + z_n^2}$$

  Note that $||z||_2 = \sqrt{z^\top z}$. Further, $||z||_2^2 = z^\top z$

Remark:
the function $f(\beta)$ (squared error loss or sum of squared errors) in the regression model formulation (4) simply corresponds to the squared length of the error vector ϵ; i.e., $f(\beta) = ||y - X\beta||_2^2 = ||\epsilon||_2^2 = \epsilon^\top \epsilon$

Review of Useful Facts - Vectors and Matrices - III

A real matrix of dimensions $n \times m$ corresponds to a collection of m column vectors of size n;
e.g., let $z, v, w \in \mathbb{R}^n$ be three column vectors, then

$$A = [z \ v \ w] \in \mathbb{R}^{n \times m}$$

or a collection of n row vectors of size m;
e.g., let $z, v, w \in \mathbb{R}^m$ be three row vectors, then

$$B = \begin{bmatrix} z \\ v \\ w \end{bmatrix}$$

Review of Useful Facts - Vectors and Matrices - III

- Transpose:
  the transpose of a matrix A of size $n \times m$ is denoted by $A^\top$ and it is a matrix of size $m \times n$ with the original columns of A being now rows in $A^\top$ (and similarly for the rows of A being columns of $A^\top$)
- Addition of same size matrices:
  Let $A, B \in \mathbb{R}^{n \times m}$. Then, $C = A + B$ is a matrix of size $n \times m$, with elements $C(i,j) = A(i,j) + B(i,j),\ i = 1, \cdots, n, j = 1, \cdots, m$
- Matrix multiplication of commuting matrices
  Let $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times q}$. Then,
  $C = AB$ is a matrix of size $n \times q$ with entries

$$C(i,j) = A(i,:)^\top B(:,j),$$

  i.e., the (i,j)-th entry of C is the inner product of the i-th row of A and the j-th column of B

Review of Useful Facts - Vectors and Matrices - IV

Some special matrices:

- Square:
  A is square if it has the same number of rows and columns; i.e., $A \in \mathbb{R}^{n \times n}$
- Identity:
  denoted by I and is a square matrix with ones as its diagonal entries $(A(i, i\_1, i = 1, \cdots, n)$ and zeros in all other entries
- Diagonal:
  A square matrix is diagonal if $A(i, j) = 0$, for all $i \neq j, i, j = 1, \cdots, n$
- Symmetric:
  A is symmetric, if
  $$A^\top = A$$
  It can be seen that a symmetric matrix must also be square
- Orthogonal:
  A is orthogonal, if
  $$A^\top A = AA^\top = I,$$
  where I denotes the the identity matrix. It can be seen that an orthogonal matrix must also be square.

Review of Useful Facts - Vectors and Matrices - V

The inverse of a matrix (is defined only for square matrices)
Let A be a square matrix of size $n \times n$. Its inverse, denoted by A^{-1} is a square matrix of size $n \times n$ that satisfies

$$AA^{-1} = A^{-1}A = I$$

Back to the SSE function of the Regression Model

Recall that

$$\begin{aligned} f(\beta) = \text{SSE}(\beta) &= \frac{1}{2n}\|\epsilon\|_2^2 = \frac{1}{2n}\left(\epsilon^\top \epsilon\right) = \frac{1}{2n}(y - X\beta)^\top (y - X\beta) \\ &= \frac{1}{2n}\left[y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta\right] \\ &= \frac{1}{2n}\left[y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta\right] \end{aligned}$$

Objective function:

$$\min_{\beta \in \mathbb{R}^p} \text{SSE}(\beta) = \min_{\beta} \ \frac{1}{2n}\left[y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta\right]$$

Review - Derivatives of multivariate functions

Let $f(w) : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function with domain $\mathbb{R}^n$ and range $\mathbb{R}$

Then, the gradient of g with respect to the multivariate argument w is a column vector denoted by $\nabla g(w)$ and defined as

$$\nabla g(w) = \begin{bmatrix} \frac{\partial g(w)}{\partial w_1} \\ \frac{\partial g(w)}{\partial w_2} \\ \cdots \\ \frac{\partial g(w)}{\partial w_n} \end{bmatrix}$$

The column vector $\nabla g(w)$ contains all the partial derivatives of $g(w)$

Example 1: Quadratic function in $\mathbb{R}^2$

Consider the quadratic function:

$$f(x) = f(x_1, x_2) = ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f$$

Its gradient is given by:

$$\nabla f(x) \equiv \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2ax_1 + cx_2 + d \\ 2bx_2 + cx_1 + e \end{bmatrix}$$

Example 2: Logistic Function in $\mathbb{R}^2$

The logistic function is defined as:

$$f(x) = f(x_1, x_2) = \frac{1}{1 + e^{-(ax_1 + bx_2 + c)}}$$

To compute the gradient, we apply the chain rule

Let

$$g(x) = g(x_1, x_2) = ax_1 + bx_2 + c$$

Then, the logistic function can be rewritten as:

$$f(g) = \frac{1}{1 + e^{-g}}$$

Differentiating $f(g)$ with respect to $g \in \mathbb{R}$ (a scalar argument) we obtain after some algebra:

$$\frac{df}{dg} = \frac{e^{-g}}{(1 + e^{-g})^2} = f(g)(1 - f(g))$$

Example 2: Logistic Function in $\mathbb{R}^2$ (ctd)

Using the chain rule:

$$\nabla f(x_1, x_2) = \frac{df}{dg} \times \nabla_x g$$

where:

$$\nabla_x g(x_1, x_2) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$$

Thus, the gradient of the logistic function is given by:

$$\nabla f(x) = \nabla f(x_1, x_2) = f(x_1, x_2)(1 - f(x_1, x_2)) \begin{bmatrix} a \\ b \end{bmatrix}$$

Note that $\nabla f(x)$ is a 2×1 column vector (as required)

Review of Two Useful Facts - Derivatives of linear and quadratic multivariate functions - II

- Let $z \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$ and consider the linear function $g(w) = w^\top z$. Then,

$$\nabla g(w) = z$$

- Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ be a square matrix and consider the function $g(w) = Aw$. Then,

$$\nabla g(w) = A^\top$$

- Let $A \in \mathbb{R}^{n \times n}$ be a square matrix and consider the quadratic function $g(w) = w^\top Aw$. Then,

$$\nabla g(w) = (A + A^\top)w$$

If A is also symmetric, then we get $\nabla g(w) = 2Aw$

Back to Deriving the Least Squares Solution $\hat{\beta}$

Recall that we aim to $\min_\beta \text{SSE}(\beta)$

To do so, analogously to the case in univariate calculus (namely setting the derivative equal to 0) we will solve for $\nabla\text{SSE}(\beta) = 0$ (namely setting the gradient equal to the zero vector)

We computed

$$\text{SSE}(\beta) = \frac{1}{2n}\left[y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta\right]$$

Using the rules to obtain the gradient of the linear function $-2\beta^\top(X^\top y)$ and the quadratic function $\beta^\top(X^\top X)\beta$, we get

$$\nabla\text{SSE}(\beta) = 0 \Longrightarrow \frac{1}{2n}\left[-2X^\top y + 2X^\top X\beta\right] = 0$$

Assuming $(X^\top X)^{-1}$ exists, some straightforward algebra gives

$$\hat{\beta} = (X^\top X)^{-1}X^\top y$$

Back to Deriving the Least Squares Solution $\hat{\beta}$

Recall that we aim to $\min_\beta \text{SSE}(\beta)$

To do so, analogously to the case in univariate calculus (namely setting the derivative equal to 0) we will solve for $\nabla\text{SSE}(\beta) = 0$ (namely setting the gradient equal to the zero vector)

We computed

$$\text{SSE}(\beta) = \frac{1}{2n}\left[y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta\right]$$

Using the rules to obtain the gradient of the linear function $-2\beta^\top(X^\top y)$ and the quadratic function $\beta^\top(X^\top X)\beta$, we get

$$\nabla\text{SSE}(\beta) = 0 \Longrightarrow \frac{1}{2n}\left[-2X^\top y + 2X^\top X\beta\right] = 0$$

Assuming $(X^\top X)^{-1}$ exists, some straightforward algebra gives

$$\hat{\beta} = (X^\top X)^{-1}X^\top y$$

The assumption that the inverse of $X^\top X$ exists also guarantees that the solution of the equation $\nabla\text{SSE}(\beta) = 0$ is unique; hence, one global minimizer

Digression: when does $(X^\top X)^{-1}$ exist?

Recall that one of the standard assumptions for the linear regression model is that the $n \times p$ matrix X is of full column rank

It can be established that

$$\text{rank}(X^\top X) = \text{rank}(X) = p \tag{8}$$

Since $X^\top X$ is a $p \times p$ matrix, it is of full rank and hence invertible

Digression: when does $(X^\top X)^{-1}$ exist?

Recall that one of the standard assumptions for the linear regression model is that the $n \times p$ matrix X is of full column rank

It can be established that

$$\text{rank}(X^\top X) = \text{rank}(X) = p \tag{8}$$

Since $X^\top X$ is a $p \times p$ matrix, it is of full rank and hence invertible

To show (8), it suffices to show that if $(X^\top X)v = 0$ for some vector $v \in \mathbb{R}^p$, then $v = 0$

If $(X^\top X)v = 0$ then multiplying both sides by v we get
$0 = v^\top (X^\top X)v = (Xv)^\top (Xv) = ||Xv||_2^2$
Hence, $Xv = 0$ and since X is assumed to be full column rank, it implies that $v = 0$

Ensuring that $\hat{\beta}$ is indeed a Minimum

Recall that in univariate calculus, to obtain the minimum of a twice differentiable function $h(w) : \mathbb{R} \to \mathbb{R}$ we do/check the following:

1. $h'(w) = 0$ (set the first derivative equal to zero and solve for $\hat{w}$)
2. $h''(\hat{w}) > 0$ (calculate the second derivative and check that for $\hat{w}$ it is positive)

Note that there may be multiple $\hat{w}$ that are solutions of $h'(w) = 0$
Then, for all of them, $h''(\hat{w}) > 0$ must hold

Review of Useful Facts - Second derivatives of vector and matrix functions

Let $g(w) : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function with domain $\mathbb{R}^n$ and range $\mathbb{R}$

We have reviewed how to calculate its gradient $\nabla g(w)$

The Hessian matrix contains all second partial derivatives of $g(w)$ and is defined as

$$\nabla^2 g(w) \equiv H(w) = \begin{bmatrix} \frac{\partial^2 g(w)}{\partial w_1 \partial w_1} & \frac{\partial^2 g(w)}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 g(w)}{\partial w_1 \partial w_n} \\ \frac{\partial^2 g(w)}{\partial w_2 \partial w_1} & \frac{\partial^2 g(w)}{\partial w_2 \partial w_2} & \cdots & \frac{\partial^2 g(w)}{\partial w_2 \partial w_n} \\ & \cdots & \cdots & \\ \frac{\partial^2 g(w)}{\partial w_n \partial w_1} & \frac{\partial^2 g(w)}{\partial w_n \partial w_2} & \cdots & \frac{\partial^2 g(w)}{\partial w_n \partial w_n} \end{bmatrix}$$

Remarks:

1. It can be seen that $H(w)$ is a $n \times n$ matrix
2. It can also be seen that $H(w) = \nabla \tilde{g}(w)$, where $\tilde{g}(w) : \mathbb{R}^n \to \mathbb{R}^n$, with $\tilde{g}(w) = \nabla g(w)$

Review of Useful Facts - Positive Definite Matrices and Characterization of Function Minimum

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix

A is positive definite, if for any vector $v \in \mathbb{R}^n$ with $v \neq 0$, the following holds:

$$v^\top A v > 0$$

Consider a twice differentiable function $g(w) : \mathbb{R}^n \to \mathbb{R}$. A point $\hat{w}$ that is a solution to $\nabla g(w) = 0$ is a minimum, if

$$H(\hat{w}) \quad \text{is positive definite at} \quad \hat{w},$$

i.e., the Hessian matrix of second partial derivatives of the function $g(w)$ is positive definite when evaluated at $\hat{w}$

Remark: Note that this is the matrix generalization of the condition $h''(\hat{w}) > 0$ for univariate twice differentiable functions

Ensuring that $\hat{\beta}$ is indeed a Minimum

We previously calculated that $\nabla \text{SSE}(\beta) = \frac{1}{2n}\left[-2X^\top y + 2X^\top X\beta\right]$

Note that the first term does not depend on β, while the gradient of the second term with respect to β is $\frac{1}{n}(X^\top X)$

Hence, $H(\beta) = \frac{1}{n}(X^\top X)$ and does not depend on β

Since we assumed that $X^\top X$ has an inverse, this implies that $X^\top X$ is positive definite (why?) and hence $\hat{\beta}$ is indeed a minimum!

Since $\hat{\beta}$ is the unique solution of $\nabla \text{SSE}(\beta) = 0$, it is then the global minimum

Why is $X^\top X$ positive definite?

We have shown that if $X^\top X$ has an inverse, then X is full column rank

Then, for any vector $\in \mathbb{R}^p$ with $v \neq 0, Xv \neq 0$

Let $v \neq 0$ and compute

$$v^\top (X^\top X) v = (Xv)^\top (Xv) = \|Xv\|_2^2 > 0$$

which is the definition of positive definiteness

Why do we need another algorithm to minimize SSE(β)?

Note that the calculation of $\hat{\beta}$ involves taking the inverse of $X^\top X$, which can be an expensive operation on the computer for large p (many predictors)

Need for general purpose optimization algorithms - I

The previous discussion established that to

$$\min_{\beta \in \mathbb{R}^p} \text{SSE}(\beta) \iff \hat{\beta} \text{ can be explicitly be computed from the data} \qquad (9)$$

Need for general purpose optimization algorithms - I

The previous discussion established that to

$$\min_{\beta \in \mathbb{R}^p} \text{SSE}(\beta) \Longleftrightarrow \hat{\beta} \text{ can be explicitly be computed from the data} \qquad (9)$$

However, small changes in the regression problem formulation render the previous strategy inoperable since finding a root of the equation $\nabla \text{SSE}(\beta) = 0$ may become challenging

Examples include:

1. $\min_{\{\beta:\ \beta_j > 0,\ j=1,\cdots,p\}} \|y - X\beta\|_2^2$
   constrain the domain of the function $f(\cdot)$ from $f : \mathbb{R}^p \to \mathbb{R}$ to $f(\cdot) : \mathbb{R}_+^p \to \mathbb{R}$
2. $\min_{\{\beta:\ ||\beta||_2^2 < t\}} \|y - X\beta\|_2^2$
   this is the formulation for ridge regression
3. $\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_1$
   change the criterion function and instead of the squared ℓ_2 norm, use the ℓ_1 norm

Need for general purpose optimization algorithms - II

More generally, for most statistical/machine learning models (e.g., logistic regression, generalized linear models, kernel regression, multi-layer perceptron), no close form solution for the model parameter exists;
i.e., the solution of the equation $\nabla f(x) = 0$ is complicated

In the sequel, we will discuss a general purpose optimization algorithm (i.e., gradient descent and its variants) that allows us to estimate parameters of interest for any statistical/machine learning model