Johns Hopkins Engineering

Applied Machine Learning for Mechanical Engineers

Optimization, Part 1, C

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Introduction to Optimization Algorithms

- By the end of this lecture you will be able to:
  - Describe optimization algorithms in general
  - Describe the application of derivatives and gradients in optimization problems

Introduction to Optimization Algorithms

- An optimization **model**, **algorithm**, or **technique** describes steps to be taken to discover the best (optimum) solution (if exists) to achieve a purpose considering every existing limitations.
  - Trial and error (costly)

What if we had thousands of variables?

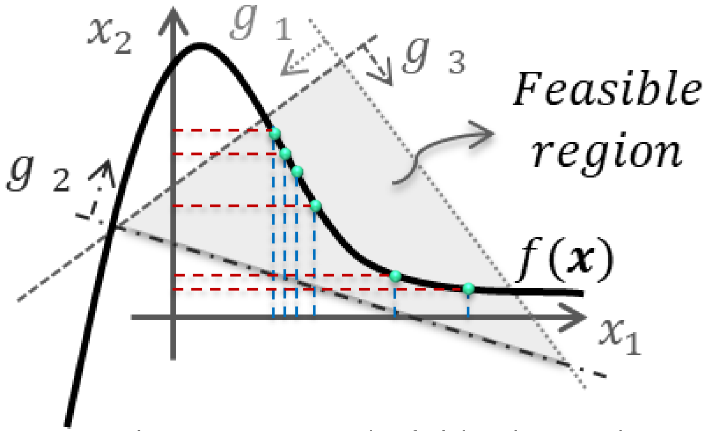

Figure 1-2 An example of trial and error using random points in the feasible region

  - Mathematical, statistical, and nature-inspired optimization algorithms (cost-effective)

Introduction to Optimization Algorithms

- Mathematical, statistical, and nature-inspired optimization algorithms are cost-effective techniques.
    - Derivatives and gradients (mathematical)
    - Simplex algorithm and gradient descent algorithm (mathematical)
    - Neural dynamic algorithm (mathematical)
    - Stochastic gradient descent algorithm (statistical)
    - Genetic algorithm and particle swarm optimization algorithm (nature-inspired)

Introduction to Optimization Algorithms

- Derivatives and gradients (mathematical)
  - Extreme points (i.e. minimum and maximum points) of a function can be discovered by investigating its derivatives or gradients with respect to its variables.

$$
\begin{gathered}
f(x) = x^2 + x + 1 \\
\frac{df}{dx} = 2x + 1 = 0 \\
x = -0.5 \\
\frac{d^2 f}{d^2 x} = 2 > 0 \\
x_{minimum} = -0.5 \\
f_{min} = f(-0.5) = 0.75
\end{gathered}
\qquad (1-4)
$$

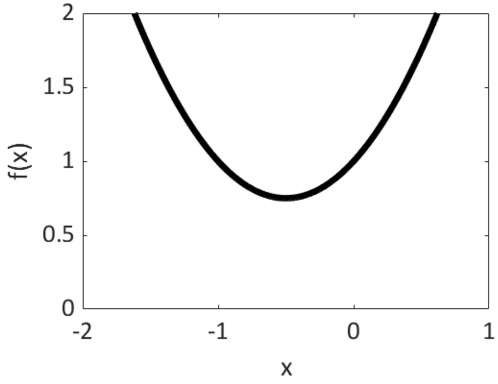

Figure 1-3 Curve corresponding to Eq. 1-4

Introduction to Optimization Algorithms

- Derivatives and gradients (mathematical)
  - Extreme points (i.e. minimum and maximum points) of a function can be discovered by investigating its derivatives or gradients with respect to its variables.

$$f(x_1, x_2) = x_1^2 + x_2^2 + 1$$

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}\right) = (2x_1, 2x_2) = 0$$

$$x_1 = 0 \ and \ x_2 = 0 \qquad (1-5)$$

$$\nabla^2 f = \left(\frac{\partial^2 f}{\partial^2 x_1}, \frac{\partial^2 f}{\partial^2 x_2}\right) = (2 > 0, 2 > 0)$$

$$f_{min} = f(x_1 = 0, x_2 = 0) = 1$$

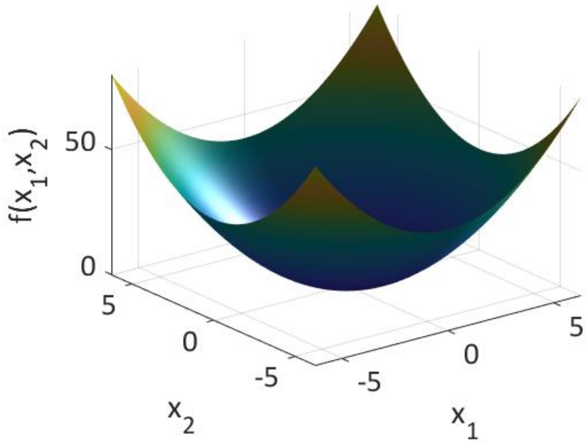

Figure 1-4 Curve corresponding to Eq. 1-5

Introduction to Optimization Algorithms

- One application of gradients can be illustrated in linear regression problems.
- Linear regression is about fitting a hyperplane in data such the residuals become minimized. Hence, linear regression is an optimization problem
- In 2D, this hyperplane is a line.
- This line shall estimate y for any x

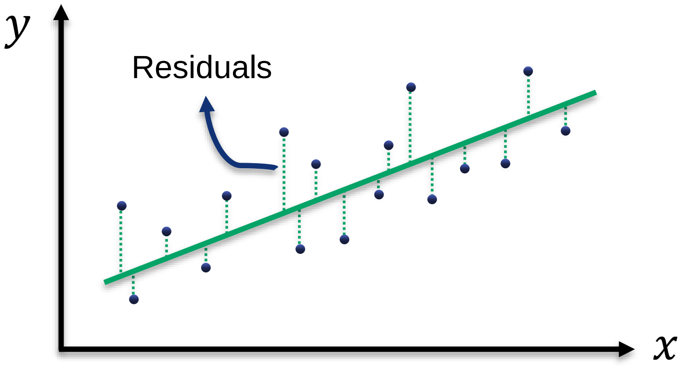

Figure 1-5 Example of residuals and the fitted line in 2D

Introduction to Optimization Algorithms

- In linear regression, the hyperplane is presented as follows:

$$y = \boldsymbol{x}\boldsymbol{\alpha}^T \qquad (1-6)$$

where $\boldsymbol{x} = [1, x_1, x_2, \dots, x_I]$ and $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_I]$ are row vectors, and T represents the transpose function. The 1 in $\boldsymbol{x}$ is to turn α_0 in $\boldsymbol{\alpha}$ as the constant of the hyperplane. The dependent variable, y, and independent variables, $\boldsymbol{x}$, have a linear relation.

Introduction to Optimization Algorithms

- If there are N points to have a hyperplane fitted for, where the n^{th} point is represented by row vector $\boldsymbol{X}_n = \left[1, X_{n,1}, X_{n,2}, \dots, X_{n,I}\right]$ and Y_n, the linear regression optimization problem is defined as minimizing the following objective function:

$$f(\boldsymbol{\alpha}) = \left\|\widehat{\boldsymbol{X}}\boldsymbol{\alpha}^T - \widehat{\boldsymbol{Y}}\right\|^2 = \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{X}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{Y}} + \boldsymbol{\alpha}^T\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}}\boldsymbol{\alpha} \qquad (1-7)$$

where $\widehat{\boldsymbol{X}}$ is an $N \times I$ matrix where the n^{th} row is $\boldsymbol{X}_n$, $\widehat{\boldsymbol{Y}} = [Y_1, Y_2, \dots. Y_N]$ is a column vector, and $\|.\|$ is the sum of squared function. There is no constraints.

Introduction to Optimization Algorithms

- The second gradient of Eq. 1-7 is always positive; hence, the minimum of Eq. 1-7 is where $\nabla f(\boldsymbol{\alpha}) = 0$:

$$\begin{gathered}\nabla f(\boldsymbol{\alpha}) = -2\widehat{\boldsymbol{Y}}^T\widehat{\boldsymbol{X}} + 2\boldsymbol{\alpha}^T\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}} = 0 \\ \boldsymbol{\alpha} = \left(\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}^T\widehat{\boldsymbol{Y}}\end{gathered} \qquad (1-8)$$

- By plugging $\boldsymbol{\alpha}$ into Eq. 1-7, the best hyperplane is retrieved.

Introduction to Optimization Algorithms

- In this lecture, you learned about:
  - Optimization algorithms in general
  - Application of derivatives and gradients in optimization problems
- In the next lecture, we will talk about mathematical and nature-inspired optimization algorithms.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING