Module 4: Supplementary Slides

(Optional, for Interested Students)

Obtaining Taylor Formula for Multivariable Functions

$$f(\boldsymbol{x}+h\boldsymbol{s}) = f(\boldsymbol{x}) + h\nabla f(\boldsymbol{x})^T\boldsymbol{s} + o(h)$$

$$f(\boldsymbol{x}+h\boldsymbol{s}) = f(\boldsymbol{x}) + h\nabla f(\boldsymbol{x})^T\boldsymbol{s} + \frac{1}{2}h^2\boldsymbol{s}^T H(\boldsymbol{x})\boldsymbol{s} + o\left(h^{\mathbf{2}}\right)$$

As before let us define:

$$w(h) = f(\boldsymbol{x}+h\boldsymbol{s}).$$

From the lecture, we know that

$$w'(0) = \nabla f(\boldsymbol{x})^T\boldsymbol{s}$$

and

$$w''(0) = \boldsymbol{s}^{\boldsymbol{T}} H(\boldsymbol{x})\boldsymbol{s}.$$

Now, expanding the Taylor formula for the single variable function $w(h)$ around $h = 0$, we have:

$$f(\boldsymbol{x}+h\boldsymbol{s}) = w(h) = w(0) + hw'(0) + o(h) = f(\boldsymbol{x}) + h\nabla f(\boldsymbol{x})^T\boldsymbol{s} + o(h)$$

$$f(\boldsymbol{x}+h\boldsymbol{s}) = w(h) = w(0) + hw'(0) + \frac{1}{2}h^2w''(0) + o\left(h^2\right) = f(\boldsymbol{x}) + h\nabla f(\boldsymbol{x})^T\boldsymbol{s} + \frac{1}{2}h^2\boldsymbol{s}^{\boldsymbol{T}} H(\boldsymbol{x})\boldsymbol{s} + o\left(h^{\mathbf{2}}\right)$$

Approximating a Multivariable Function

Approximating a Multivariable Function

$f(\boldsymbol{x} + \boldsymbol{\Delta x})$ can be approximated by using the Taylor expansion as:

$$f(\boldsymbol{x} + \boldsymbol{\Delta x}) \approx f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x}) \cdot \boldsymbol{\Delta x} = f(\boldsymbol{x}) + \sum_{k=1}^{n} \frac{\partial f}{\partial x_k} \Delta x_k$$

Example: If $f(x, y) = ln(x) + ln(y)$, then approximate $f(1.06, 1.02)$ using differentials.

Solution:

$$f(x + \Delta x, y + \Delta y) \approx f(x, y) + \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y$$

$$ln(x + \Delta x) + ln(y + \Delta y) \approx [ln(x) + ln(y)] + \frac{1}{x} \Delta x + \frac{1}{y} \Delta y$$

$$ln(1.06) + ln(1.02) \approx [ln(1) + ln(1)] + \frac{1}{1}(0.06) + \frac{1}{1}(0.02) = 0.08$$

Matrix Calculus

Matrix Calculus

Two Most Useful Results

Proposition 1: *Let* $\boldsymbol{x}, \boldsymbol{c} \in \mathbb{R}^{n \times 1}$. *Then*

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{c}^T \boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^T \boldsymbol{c}) = \boldsymbol{c}$$

Proof: To obtain $\nabla_x \boldsymbol{c}^T \boldsymbol{x}$, we first expand $\boldsymbol{c}^T \boldsymbol{x}$:

$$\boldsymbol{c}^T \boldsymbol{x} = [c_1 x_1 + c_2 x_2 + \cdots + c_n x_n]$$

The partial derivative with respect to a single coordinate is:

$$\frac{\partial}{\partial x_i} \boldsymbol{c}^T \boldsymbol{x} = c_i$$

Thus, the gradient is:

$$\nabla_x \boldsymbol{c}^T \boldsymbol{x} = \nabla_x \boldsymbol{x}^T \boldsymbol{c} = \boldsymbol{c}$$

Matrix Calculus

Two Most Useful Results

Proposition 2: *Let* $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$ *and* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. *Then*

$$\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$$

Proof: To obtain $\nabla_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}$ for a square matrix $\boldsymbol{A}$, we first expand $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}$:

$$\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T \begin{bmatrix} x_1 a_{11} + x_2 a_{12} + \cdots + x_n a_{1n} \\ x_1 a_{21} + x_2 a_{22} + \cdots + x_n a_{2n} \\ \vdots \\ x_1 a_{n1} + x_2 a_{n2} + \cdots + x_n a_{nn} \end{bmatrix} = \begin{matrix} x_1^2 a_{11} + x_2^2 a_{12} + \cdots + x_n^2 a_{1n} \\ x_1^2 a_{21} + x_2^2 a_{22} + \cdots + x_n^2 a_{2n} \\ \vdots \\ x_1^2 a_{n1} + x_2^2 a_{n2} + \cdots + x_n^2 a_{nn} \end{matrix}$$

The partial derivative with respect to the *i*th component is:

$$\frac{\partial}{\partial x_i} \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} = \sum_{j=1}^{n} x_j \left(a_{ij} + a_{ji}\right)$$

Matrix Calculus

Proof (continued): Thus, the gradient is:

$$\nabla_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \begin{bmatrix} \sum_{j=1}^{n} x_j \left(a_{1j} + a_{j1}\right) \\ \sum_{j=1}^{n} x_j \left(a_{2j} + a_{j2}\right) \\ \vdots \\ \sum_{j=1}^{n} x_j \left(a_{nj} + a_{jn}\right) \end{bmatrix} = \begin{bmatrix} a_{11} + a_{11} & a_{21} + a_{12} & \cdots & a_{1n} + a_{n1} \\ a_{12} + a_{21} & a_{22} + a_{22} & \cdots & a_{n2} + a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} + a_{n1} & a_{2n} + a_{n2} & \cdots & a_{nn} + a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \left(\boldsymbol{A} + \boldsymbol{A}^T\right)\boldsymbol{x}$$

Matrix Calculus

Other Helpful and General Results

Let $a \in \mathbb{R}$, $\boldsymbol{x}, \boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{U}(\boldsymbol{x}), \boldsymbol{V}(\boldsymbol{x}), \boldsymbol{A} \in \mathbb{R}^{n\times n}$. *Then*

1. $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{c}^T\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{c}) = \boldsymbol{c}$ (Proposition 1)
2. $\frac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$ (Proposition 2)
3. $\frac{\partial \boldsymbol{x}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$ (Special Case of Proposition 2)
4. $\frac{\partial^2 \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T} = \boldsymbol{A} + \boldsymbol{A}^T$ (Resulting from using Prop 1 on Prop 2)
5. $\frac{\partial \boldsymbol{c}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^T\boldsymbol{c}$ (Another representation of Proposition 1)
6. $\frac{\partial a}{\partial \boldsymbol{x}} = 0$ (Trivial result)
7. $\frac{\partial}{\partial \boldsymbol{x}}\left(a\boldsymbol{U}(\boldsymbol{x})\right) = a\frac{\partial \boldsymbol{U}(\boldsymbol{x})}{\partial \boldsymbol{x}}$ (Trivial result)
8. $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{U}(\boldsymbol{x}) + \boldsymbol{V}(\boldsymbol{x})) = \frac{\partial \boldsymbol{U}(\boldsymbol{x})}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{V}(\boldsymbol{x})}{\partial \boldsymbol{x}}$
9. $\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{U}(\boldsymbol{x})^T\boldsymbol{V}(\boldsymbol{x})\right) = \frac{\partial \boldsymbol{U}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{V}(\boldsymbol{x}) + \frac{\partial \boldsymbol{V}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{U}(\boldsymbol{x})$ (Product rule)
10. $\frac{\partial}{\partial \boldsymbol{x}}[\boldsymbol{U}(\boldsymbol{x})^T\boldsymbol{A}\boldsymbol{V}(\boldsymbol{x})] = \frac{\partial \boldsymbol{U}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{A}\boldsymbol{V}(\boldsymbol{x}) + \frac{\partial \boldsymbol{V}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{A}^T\boldsymbol{V}(\boldsymbol{x})$ (Product rule)
11. $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{c}^T\boldsymbol{U}(\boldsymbol{x})) = \frac{\partial \boldsymbol{U}(\boldsymbol{x})}{\partial \boldsymbol{x}}\boldsymbol{c}$ (Generalization to Proposition 1)
12. $\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{c}^T\boldsymbol{x}^T\boldsymbol{c}\boldsymbol{x}\right) = \frac{\partial}{\partial \boldsymbol{x}}\left((\boldsymbol{x}\boldsymbol{c})^T\boldsymbol{c}\boldsymbol{x}\right) = 2\boldsymbol{c}\boldsymbol{c}^T\boldsymbol{x}$

Application of Matrix Calculus: Least Square Linear Regression

Linear Regression

In a multiple regression, we examine the linear relationship between one dependent variable and several independent variables. Let m observations of one dependent variable be denoted with $\boldsymbol{y} = [y_i]_{m\times 1}$ and the corresponding m observations of independent variables be denoted by $\boldsymbol{X} = [\mathbf{1}\ \boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_k}] = \left[x_{ij}\right]_{m\times(k+1)}$, $\boldsymbol{x_j} \in \mathbb{R}^m$, $\mathbf{1} = [1]_{m\times 1}$ where x_{ij} is the i^{th} observation of the independent variable $\boldsymbol{x_j}$. We wish to estimate the value of $\boldsymbol{Y}$ by considering a linear function in the form of $\hat{y}_i = \beta_0(1) + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ where $\hat{y}_i$ is the predicted value for observation i. In a concise form, we can write $\hat{\boldsymbol{y}} = \boldsymbol{\beta X} \equiv [\hat{y}_i]_{m\times 1}$, and

$$\hat{\boldsymbol{y}}_{\boldsymbol{m\times 1}} = \boldsymbol{X\beta}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mk} \end{bmatrix}_{m\times(k+1)}$$

The value $\mathrm{h}_i = y_i - \hat{y}_i$ is prediction error corresponding to the i^{th} observations $[1\ x_{i1}, x_{i2}, \dots, x_{ik}]$ and y_i.

Linear Regression

Ordinary Least Squares Regression (OLS): The OLS regression line is the linear function $\hat{\boldsymbol{y}} = \boldsymbol{X\beta}$ *that minimizes the sum of squared error values, i.e.,* $\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \left(\boldsymbol{e}^T\boldsymbol{e}\right), \boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\beta}$. *The OLS regression line coefficient vector is given by:*

$$\boldsymbol{\beta}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Proof:

Expanding $\boldsymbol{e}^T\boldsymbol{e}$ gives:

$$\boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) = \left(\boldsymbol{y}^T - \boldsymbol{\beta}^T\boldsymbol{X}^T\right)(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}.$$

From Proposition 1, we note that

$$\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{y}^T\boldsymbol{X\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{y}.$$

Also, by Proposition 2,

$$\frac{\partial}{\partial\boldsymbol{\beta}}\left(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta}\right) = \boldsymbol{X}^T\boldsymbol{X} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^T.$$

Linear Regression

Using the first derivative test to find the minimum values, using matrix calculus properties, we have:

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left((\boldsymbol{y}-\boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{X\beta})\right) = 0 - \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{y} + \left(\boldsymbol{X}^T\boldsymbol{X} + \left(\boldsymbol{X}^T\boldsymbol{X}\right)^T\right)\boldsymbol{\beta} = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X\beta} = -2\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{X\beta}).$$

Solving for the optimal coefficients vector gives:

$$-\mathbf{2}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{X\beta}^*) = \mathbf{0} \Rightarrow \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X\beta}^* = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \Rightarrow \boldsymbol{\beta}^* = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}.$$

Now, checking the second derivative test to verify that $\boldsymbol{\beta}^*$ is a minimizer, we have

$$\frac{\partial}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\left((\boldsymbol{y}-\boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{X\beta})\right) = \frac{\partial}{\partial \boldsymbol{\beta}}\left(-2\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X\beta}\right) = 2\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X},$$

in which we used Proposition 1, to calculate $\frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X\beta}\right) = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^T = \boldsymbol{X}^T\boldsymbol{X}.$ Also, any constant comes out of derivative.

Note that $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is positive semidefinite because $\forall\, \boldsymbol{w} \in \mathbb{R}^m$:

$$\boldsymbol{w}^T\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} = (\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{X}\boldsymbol{w}) = \|\boldsymbol{X}\boldsymbol{w}\|^2 \geq 0.$$

Hence, $\boldsymbol{\beta}^*$ is a minimizer.

Linear Regression

Example

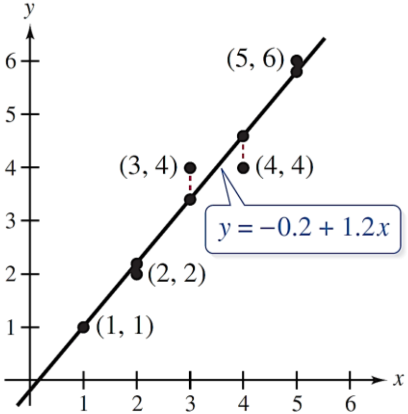
Find the linear regression line for the points $(1,1), (2,2), (3,4), (4,4), (5,6)$.

Solution.

The matrices $\boldsymbol{X}$ and $\boldsymbol{y}$ are $\boldsymbol{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 4 \\ 6 \end{bmatrix}$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}, \qquad \boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 4 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 17 \\ 63 \end{bmatrix}$$

$$\boldsymbol{\beta}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{1}{50}\begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix}\begin{bmatrix} 17 \\ 63 \end{bmatrix} = \begin{bmatrix} -0.2 \\ 1.2 \end{bmatrix}$$



Least Squares Regression Line

Hence, the least squares regression line is $y = -0.2 + 1.2x$.

Constrained Optimization

KKT Conditions for Multiple Constraints

Equality Constraint

Proof for One Constraint Using Gradients

Consider an optimization problem with a single equality constraint

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{subject to } \; g(\boldsymbol{x}) = 0$$

where f and g have continuous partial derivatives. Then $f(\boldsymbol{x})$ has an interior local minimum at $\boldsymbol{x}^*$ if, for any feasible direction $\boldsymbol{s}$,

$$\nabla f(\boldsymbol{x}^*) \cdot \boldsymbol{s} = 0 \Rightarrow \nabla f(\boldsymbol{x}^*) \perp \boldsymbol{s}$$

Feasible directions are those that satisfy:

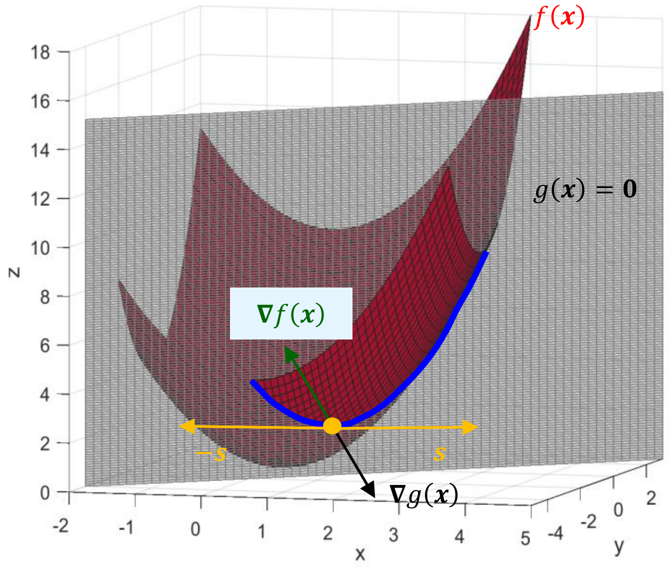
$$g(\boldsymbol{x} + h\boldsymbol{s}) = g(\boldsymbol{x}) = 0 \Rightarrow Dg_{\boldsymbol{s}}(\boldsymbol{x}) = 0 \Rightarrow \nabla g(\boldsymbol{x}) \cdot \boldsymbol{s} = 0 \Rightarrow \nabla g(\boldsymbol{x}) \perp \boldsymbol{s}$$

So, at the optimal point $\boldsymbol{x}^*$:

$$\nabla f(\boldsymbol{x}^*) || \nabla g(\boldsymbol{x}^*) \Rightarrow \nabla f(\boldsymbol{x}^*) = \lambda \nabla g(\boldsymbol{x}^*)$$

λ is called a Lagrange multiplier and $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$ is called Lagrangian function. At $\boldsymbol{x}^*$:

$$\nabla L(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*) - \lambda \nabla g(\boldsymbol{x}^*) = \boldsymbol{0}$$



Inequality Constraints

Consider an optimization problem with a single inequality constraint

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{subject to } \; g(\boldsymbol{x}) \leq 0$$

We know that if the solution lies at the constraint boundary, the Lagrange multiplier holds: $\nabla f(\boldsymbol{x}^*) - \lambda \nabla g(\boldsymbol{x}^*) = \mathbf{0}$, otherwise, we have $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$. We could optimize the problem by introducing an infinite step penalty for infeasible points:

$$f_{\infty-\text{step}}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \text{if} \quad g(\boldsymbol{x}) \leq 0 \\ \infty & \text{if} \quad g(\boldsymbol{x}) > 0 \end{cases} \Rightarrow f_{\infty-\text{step}}(\boldsymbol{x}) = f(\boldsymbol{x}) + \infty(g(\boldsymbol{x}) > 0)$$

Unfortunately, $f_{\infty-\text{step}}(\boldsymbol{x})$ is inconvenient to optimize because it is discontinuous and nondifferentiable. We can instead use a linear penalty $\lambda g(\boldsymbol{x})$, which forms a lower bound on $\infty(g(\boldsymbol{x}) > 0)$ and penalizes the objective as long as $\lambda > 0$. We can use this linear penalty to construct a Lagrangian function:

$$L(\boldsymbol{x}, \mu) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$$

We can recover $f_{\infty-\text{step}}(\boldsymbol{x})$ by maximizing with respect to μ.

$$f_{\infty-\text{step}}(\boldsymbol{x}) = \max_{\mu \geq 0} L(\boldsymbol{x}, \mu)$$

For any infeasible $\boldsymbol{x}$, we get infinity and for any feasible $\boldsymbol{x}$, we get $f(\boldsymbol{x})$.

Inequality Constraints

KKT Conditions

The new optimization problem is thus

$$\min_{x} \max_{\lambda \geq 0} L(\boldsymbol{x}, \lambda)$$

This formulation is known as the primal problem. Optimizing the primal problem will require finding $\boldsymbol{x}^*$ such that:

1. (Prime) Feasibility: $g(\boldsymbol{x}^*) \leq 0$
   - the point is feasible
2. Dual Feasibility: $\lambda \geq 0$
   - The penalty must point in the right direction. μ is also called a dual variable.
3. Complementary Slackness: $\lambda g(\boldsymbol{x}^*) = 0$
   - A feasible point on the boundary will have $g(\boldsymbol{x}^*) = 0$ whereas a feasible point with $g(\boldsymbol{x}^*) < 0$ must have $\lambda = 0$ to recover $f(\boldsymbol{x}^*)$ from Lagrangian
4. Stationarity: $\boldsymbol{\nabla} f(\boldsymbol{x}^*) - \lambda \boldsymbol{\nabla} g(\boldsymbol{x}^*) = \boldsymbol{0}$
   - When the constraint is active, we require the Lagrange multiplier. When it is inactive, we require $\boldsymbol{\nabla} f(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\lambda = \boldsymbol{0}$.

These conditions are known as Karush-Kuhn-Tucker (KKT) conditions. (Mostly used for verification of optimizality of a solution. It is hard to use it directly to find an optimal value)

Constrained Optimization

KKT Conditions For Multiple Constraints

KKT conditions: Consider the following problem:

$$\max_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

subject to :

$$c_k(\boldsymbol{x}) \leq 0, \quad k = 1, \dots, K$$

The first derivative necessary conditions for $\boldsymbol{x}^* \in \mathbb{R}^m$ to be a local maximizer is that $\exists\ \boldsymbol{\lambda}^* \in \mathbb{R}^K$ such that :

1. (Primal) feasibility: $c_k(\boldsymbol{x}) \leq 0, \quad k = 1, \dots, K$
2. Stationarity: $\frac{\partial f(\boldsymbol{x})}{\partial x_i} - \sum_{k=1}^{K} \lambda_k \frac{\partial c_k(\boldsymbol{x})}{\partial x_i} = 0,\ i = 1, \dots, n$
3. Complementary Slackness: $\lambda_k^* c_k(\boldsymbol{x}) = 0,\ k = 1, \dots, \mathrm{K},$
4. Positivity (/Dual feasibility): $\lambda_k^* \geq 0,\ k = 1, \dots, K$

KKT Conditions For Multiple Constraints

Example

Solve the following problem:

$$\max_{x_1,x_2} f(x_1,x_2) = (x_1+1)^2 + (x_2+1)^2$$

subject to:

$$\lambda_1 \qquad x_1^2 + x_2^2 \le 2,$$

$$\lambda_2 \qquad x_2 \le 1$$

Solution: The KKT conditions are

1) $2(x_1^*+1) - \lambda_1(2x_1^*) = 0 \Rightarrow x_1^*+1 = \lambda_1^* x_1^*$

2) $2(x_2^*+1) - \lambda_1(2x_2^*) - \lambda_2(1) = 0 \Rightarrow x_2^*+1 = \lambda_1^* x_2^* + \frac{1}{2}\lambda_2^*$

3) $\lambda_1^*\left({x^*}_1^2 + {x^*}_2^2 - 2\right) = 0$

4) $\lambda_2^*(x_2^* - 1) = 0$

5) $\lambda_1^* \ge 0,\ \lambda_2^* \ge 0$

KKT Conditions For Multiple Constraints

Solution

We can examine the possibilities for the complementarity condition:

- $\lambda_1^* = 0,\ \lambda_2^* = 0$

$$x_1^* + 1 = 0,\ x_2^* + 1 = 0\ \Rightarrow (x_1^*, x_2^*) = (-1, -1) \Rightarrow f(x_1^*, x_2^*) = 0$$ max

- $\lambda_1^* = 0,\ \lambda_2^* > 0$

$$x_2 + 1 = \frac{1}{2}\lambda_2^*,\ \lambda_2^*(x_2^* - 1) = 0 \Rightarrow (x_1^*, x_2^*) = (-1, 1),\ \lambda_2^* = 4 \Rightarrow f(x_1^*, x_2^*) = -4$$ Saddle point

- $\lambda_1^* > 0,\ \lambda_2^* = 0$

$$x_1^* + 1 = \lambda_1^* x_1^*,\ x_2^* + 1 = \lambda_1^* x_2^*,\ \lambda_1^*\left(x_1^2 + x_2^2 - 2\right) = 0 \Rightarrow (x_1^*, x_2^*) = (1,1),\ \lambda_1^* = 2 \Rightarrow f(x_1^*, x_2^*) = -8$$ min

- $\lambda_1^* > 0,\ \lambda_2^* > 0$

$${x^*}_1^2 + {x^*}_2^2 = 2, x_2^* = 1 \Rightarrow (x_1^*, x_2^*) = (\pm 1, -1)$$ But the answers coincide the previous items.

KKT Conditions For Multiple Equality and Inequality Constraints

Consider the following problem:

$$\max_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

Subject to :

$$c_k(\boldsymbol{x}) \leq 0, \quad k = 1, \dots, K$$

$$d_l(\boldsymbol{x}) = 0, \quad l = 1, \dots, L$$

The first derivative necessary conditions for $\boldsymbol{x}^* \in \mathbb{R}^m$ to be a local maximizer is that $\exists\ \boldsymbol{\lambda}^* \in \mathbb{R}^K$ and $\exists\ \boldsymbol{\gamma}^* \in \mathbb{R}^L$ such that :

1. (Primal) feasibility: $c_k(\boldsymbol{x}) \leq 0, \quad k = 1, \dots, K,$ and $d_l(\boldsymbol{x}) = 0, \quad l = 1, \dots, L$
2. Stationarity: $\frac{\partial f(\boldsymbol{x})}{\partial x_i} - \sum_{k=1}^{K} \lambda_k \frac{\partial c_k(\boldsymbol{x})}{\partial x_i} - \sum_{l=1}^{L} \gamma_k \frac{\partial d_k(\boldsymbol{x})}{\partial x_i} = 0, \ i = 1, \dots, n$
3. Complementary Slackness: $\lambda_k^* c_k(\boldsymbol{x}) = 0, \ k = 1, \dots, K,$ and $\gamma_l^* c_l(\boldsymbol{x}) = 0, \ l = 1, \dots, L$
4. Positivity (/Dual feasibility): $\lambda_k^* \geq 0, \ k = 1, \dots, K$ (only for inequality constraints)

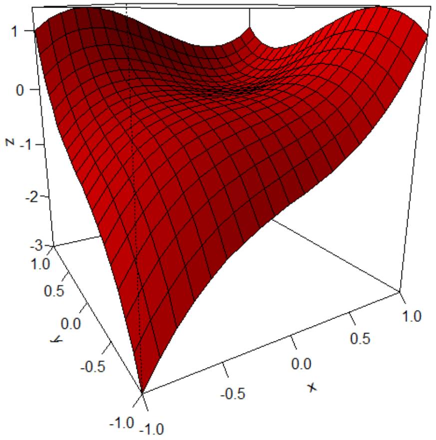
Additional Examples in R

Plotting 3D in R

```
persp (x,y,z)
Creates a 3D surface graphic for the vectors (x,y,z) where z is a function of x and
y defined previously by the "outer" code.
```

Example: Plot $f(x,y) = x^3 + y^3 - xy$

```
x=y=seq(-1,1,length=20)
f=function(x,y)x^3+y^3-x*y
z=outer(x,y,f)
persp(x,y,z,
    main= "Perspective Plot", # (optional)
    zlab = "Height", # (optional)
    theta  = -30,       # Rotation (vertical) (optional)
    phi    = 0,       # Rotation (horizontal) (optional)
    expand =  1,       # Shrinking/growing factor (optional)
    col    =  "red", # Shrinking/growing factor (optional)
    shade  =  0.6, # color shade (optional)
    ticktype = "detailed" # puts numbers on x,y, and z axes
    )
```



Constrained Optimization

Example: Equality Constraint

Minimize $c(q_1, q_2) = 2q_1^2 + q_1q_2 + q_2^2 + 200$ subject to $q_1 + q_2 = 200$.

Solution:

```
f=function(x) (2*x[1]^2+x[1]*x[2]+x[2]^2+200)
equalities=function(x){
  h=0
  h[1]=200-x[1]-x[2]
  return(h)}
p0=c(0,0)
y=constrOptim.nl(p0,f,heq = equalities);
print(y$par)
[1] 49.9983 150.0017
print(y$value)
[1] 35200
```

Constrained Optimization

Example: Equality Constraint

Maximize $f(x_1, x_2) = 4x_1^2 + 10x_2^2$ subject to $x_1^2 + x_2^2 = 4$.

Solution: Write the function as minimizing $-f(x_1, x_2)$.

```
f=function(x) -(4*x[1]^2+10*x[2]^2)
Equalities=function(x){
  h=0
  h[1]=(x[1]^2+x[2]^2-4)
  return(h)}
p0=c(2,0)
y=constrOptim.nl(p0,f, heq=Equalities);
print(y$par)
[1] -4.99786e-05 -2.00000e+00
```

$(x_1^*, x_2^*) = (0, -2)$

```
print(y$value)
[1] -40
```

$f(x_1^*, x_2^*) = 40$

Constrained Optimization

Example: Inequality Constraint

Maximize $f(x_1, x_2) = 4x_1^2 + 10x_2^2$ subject to $x_1^2 + x_2^2 \leq 4$.

Solution: Write the function as minimizing $-f$.

```
f=function(x) -(4*x[1]^2+10*x[2]^2)
inequalities=function(x){
  h=0
  h[1]= 4-x[1]^2-x[2]^2 #note that h[1] must be positive, i.e., h[1]>=0
  return(h)}
p0=c(0,0)
y=constrOptim.nl(p0,f,hin=inequalities);
par: 0.02919113 1.999787
fval: -39.99489
```