EMSE 6035: Marketing of Technology

Intro to Maximum Likelihood Estimation & Optimization

John Paul Helveston, Ph.D.
Assistant Professor
Engineering Management & Systems Engineering
The George Washington University

Background: Random Utility Model

Utility can be broken into two parts:

$$\tilde{u}_j = v_j + \tilde{\varepsilon}_j$$

Things we can observe / measure ("Observed Utility")

Things we can't observe / measure ("Error")

We define v_j as a function of observable product attributes, x_j:

$$v_j = f(x_j) = \boxed{\beta_1} x_{j1} + \boxed{\beta_2} x_{j2} + \ldots$$

Weights that denote the *relative* value of attributes x_{j1} and x_{j2}

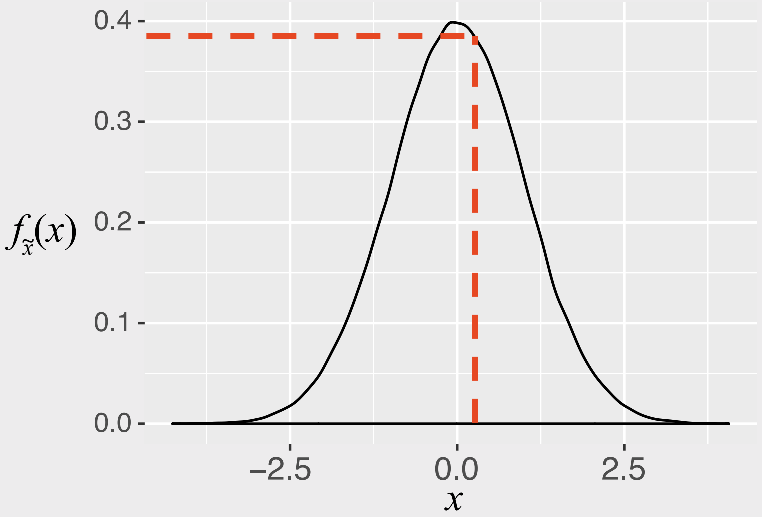
Estimate model coefficients, β_1 , β_2 , … , by maximizing the likelihood function

The likelihood function is a function of the parameters of a statistical model, given observed data

Probability

$$\Pr(\tilde{x} = x \mid \boldsymbol{\theta})$$

Example:
$\tilde{x}$ follows a normal distribution
with two parameters ($\boldsymbol{\theta}$) :
- Mean ($\mu = 0$)
- Standard deviation ($\sigma = 1$)



$$\Pr(\tilde{x} = 0 \mid \boldsymbol{\theta}) \\ = f_{\tilde{x}}(0) \\ \cong 0.4$$

Likelihood

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x})$$

Example:
We assume $\tilde{x}$ follows a normal distribution
We have the following observations

0.2	-0.5	-1	0.2	0.1	1.6	0.6	0.5	-1.9	-0.4

What is the likelihood that the parameters are:
- Mean ($\mu = 0$)
- Standard deviation ($\sigma = 1$)

$f_{\tilde{x}}(\mathbf{x}) =$

0.39	0.35	0.24	0.39	0.40	0.11	0.33	0.35	0.07	0.37

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = f_{\tilde{x}}(x_1)\, f_{\tilde{x}}(x_2) \ldots f_{\tilde{x}}(x_n) = 1.63\text{e-}6$$

Take the log of the likelihood to convert multiplication to addition

0.39	0.35	0.24	0.39	0.40	0.11	0.33	0.35	0.07	0.37

$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = f_{\tilde{x}}(x_1)\, f_{\tilde{x}}(x_2) \ldots f_{\tilde{x}}(x_n)$ = 1.63e-6

log $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = f_{\tilde{x}}(x_1) + f_{\tilde{x}}(x_2) + \ldots + f_{\tilde{x}}(x_n) = 3$

Maximum likelihood estimation is about finding the parameters that produce the highest log-likelihood

<u>Observations</u>

0.2	-0.5	-1	0.2	0.1	1.6	0.6	0.5	-1.9	-0.4

μ	$\boldsymbol{\sigma}$	Probability of $\tilde{x} = x$										$\log \mathcal{L}(\boldsymbol{\theta}\|\mathbf{x})$
-1	1	0.19	0.35	0.40	0.19	0.22	0.01	0.11	0.13	0.27	0.33	2.2
0	1	0.39	0.35	0.24	0.39	0.40	0.11	0.33	0.35	0.07	0.37	3
1	2	0.18	0.15	0.12	0.18	0.18	0.19	0.20	0.19	0.07	0.16	1.62

Practice Question 1

Observations: Height of students (inches)

65	69	66	67	68	72	68	69	63	70

a. Let's say we know that the height of students, $\tilde{x}$, in a classroom follows a normal distribution. A professor obtains the above height measurements students in her classroom. What is the log-likelihood that $\tilde{x} \sim N(68, 4)$? In other words, compute log$\mathcal{L}(\mu = 68, \sigma = 4|\mathbf{x})$.

Hints:

1. The log-likelihood is computed by: $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = f_{\tilde{x}}(x_1)+f_{\tilde{x}}(x_2)+\ldots+f_{\tilde{x}}(x_n)$
2. The *dnorm(x, mean, sd)* function in *R* returns the value of $f_{\tilde{x}}(x)$ for a normal distribution with a given mean (*mean*) and standard deviation (*sd*).

b. Compute the log-likelihood function using the same standard deviation ($\sigma = 4$) but with the following different values for the mean, μ: 66, 67, 68, 69, 70. How do the results compare? Which value for μ produces the highest log-likelihood?

Use the data we observe, $\mathbf{x}$, to estimate the parameters, $\boldsymbol{\theta}$, of an assumed model

$$\text{maximize}\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = f_{\tilde{x}}(x_1)+f_{\tilde{x}}(x_2)+\ldots+f_{\tilde{x}}(x_n) = \sum_{i=1}^{n} f_{\tilde{x}}(x_i|\boldsymbol{\theta})$$

with respect to $\boldsymbol{\theta}$

Solving this is known as
"Maximum Likelihood Estimation"

This is an optimization problem!

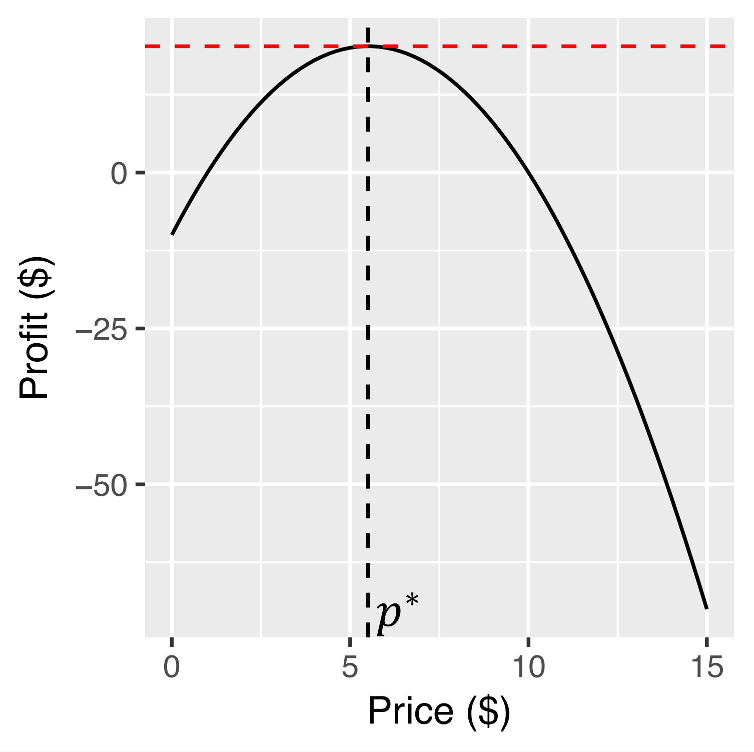
Optimization:
Find the value, x, that maximizes the function $f(x)$

Example: Find what price, p, will maximize profit , π, for the following model:

Profit: $\pi(p) = q(p - c)$
Demand: $q = 10 - p$
Cost: c

maximize $\quad \pi(p)$
with respect to p
subject to $p \geq 0$

Profit if $c = 1$:



$$\pi(p) = q(p - c)$$
$$= (10 - p)(p - c)$$
$$= -p^2 + (10 + c)p \; - 10c$$

$$\frac{\partial \pi}{\partial p} = -2p + 10 + c = 0$$

Solve for p:

$$p^* = \frac{10 + c}{2}$$

If $c = 1$, $p^* = \frac{11}{2}$ =5.5

Optimality Conditions

Optimality conditions

First order necessary condition

x^* is a "stationary point" when

$$\frac{df(x^*)}{dx} = 0$$

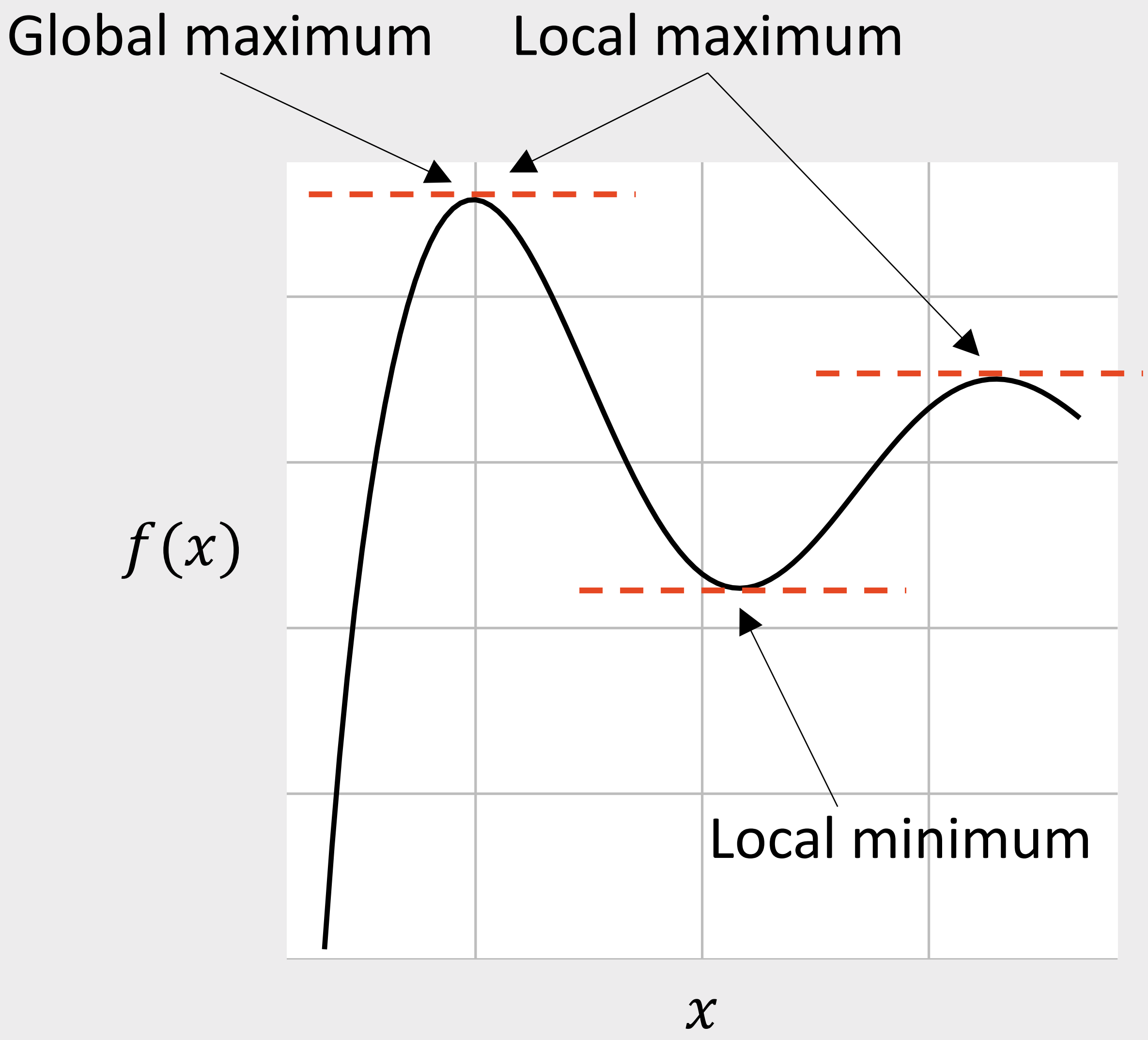
Second order sufficiency condition

x^* is a local *maximum* when

$$\frac{d^2f(x^*)}{dx^2} < 0$$

x^* is a local *minimum* when

$$\frac{d^2f(x^*)}{dx^2} > 0$$



Optimality conditions

First order necessary condition

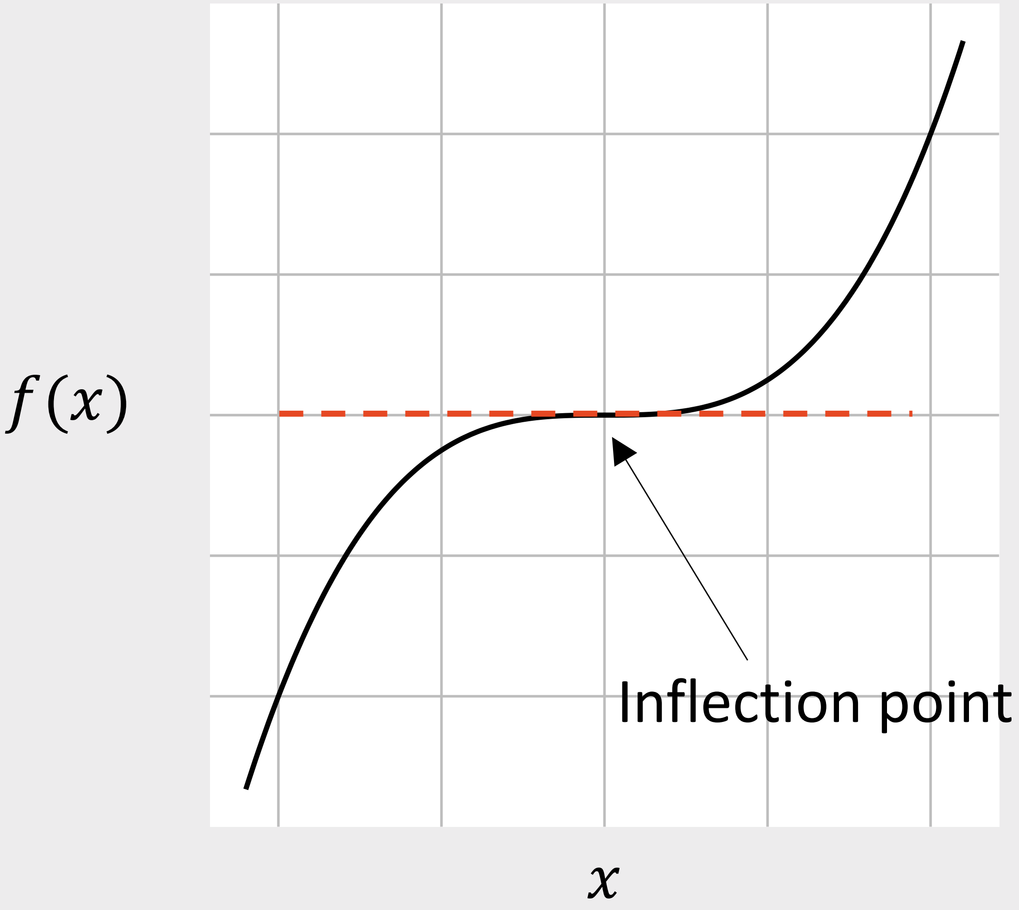
x^* is a "stationary point" when

$$\frac{df(x^*)}{dx} = 0$$

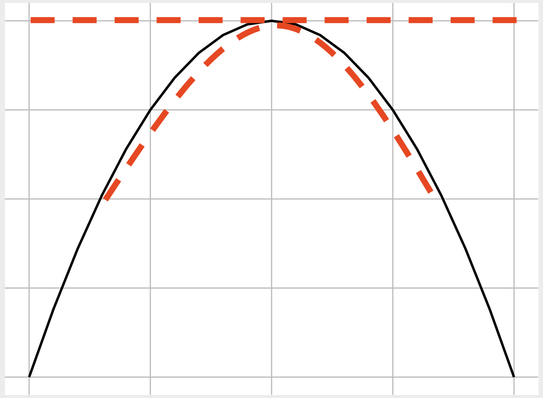
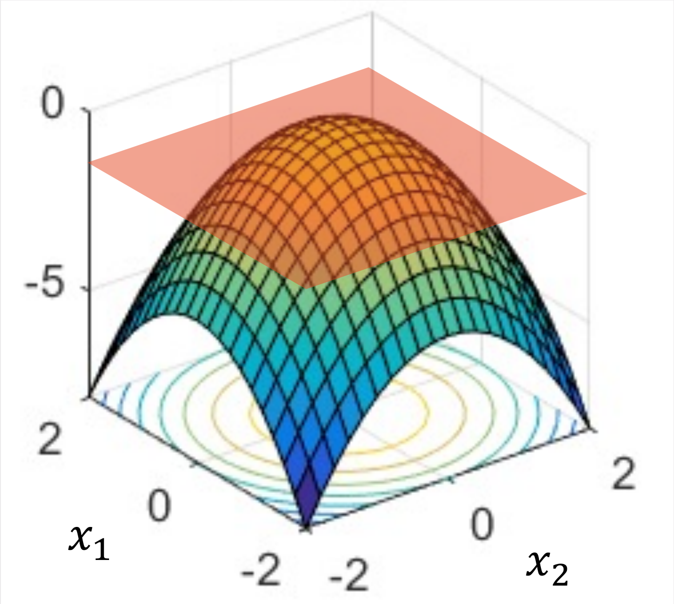
Second order sufficiency condition

x^* is an *inflection point* when

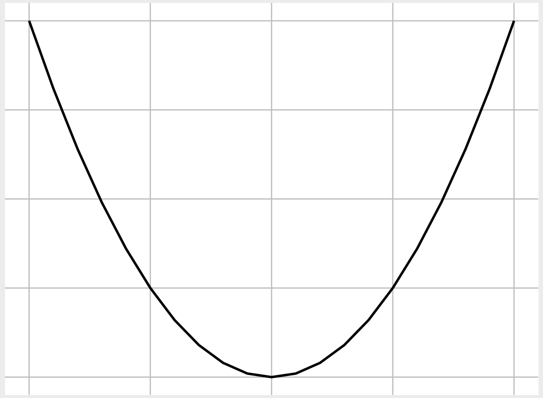
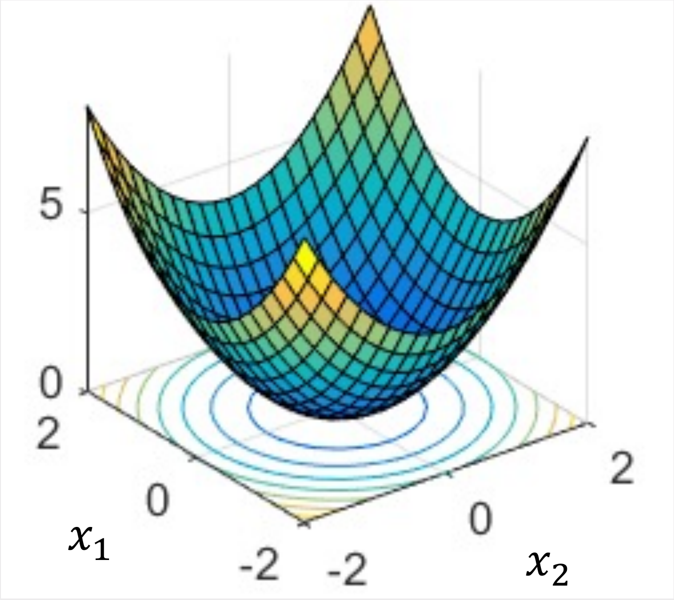
$$\frac{d^2 f(x^*)}{dx^2} = 0$$



Optimality conditions for local **maximum**

Number of dimensions	First order condition	Second order condition	Example
One	$\frac{df(x^*)}{dx} = 0$	$\frac{d^2 f(x^*)}{dx^2} < 0$	
Multiple	"Gradient" $\nabla f(x_1, x_2, \ldots x_n)$ $= \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]$ $= \color{red}{[0,0,\ldots,0]}$	"Hessian" $\nabla^2 f(x_1, x_2, \ldots x_n)$ $= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$ Must be "negative definite"	

Optimality conditions for local **minimum**

Number of dimensions	First order condition	Second order condition	Example
One	$\frac{df(x^*)}{dx} = 0$	$\frac{d^2 f(x^*)}{dx^2} > 0$	
Multiple	"Gradient" $\nabla f(x_1, x_2, \dots x_n)$ $= \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right]$ $=$ $[0,0, \dots, 0]$	"Hessian" $\nabla^2 f(x_1, x_2, \dots x_n)$ $= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$ Must be "positive definite"	

Optimization Convention: "Negative Null Form"

maximize $f(x)$ with respect to x subject to…	$\longrightarrow$	minimize $-f(x)$ with respect to x subject to…

Optimization Approaches:

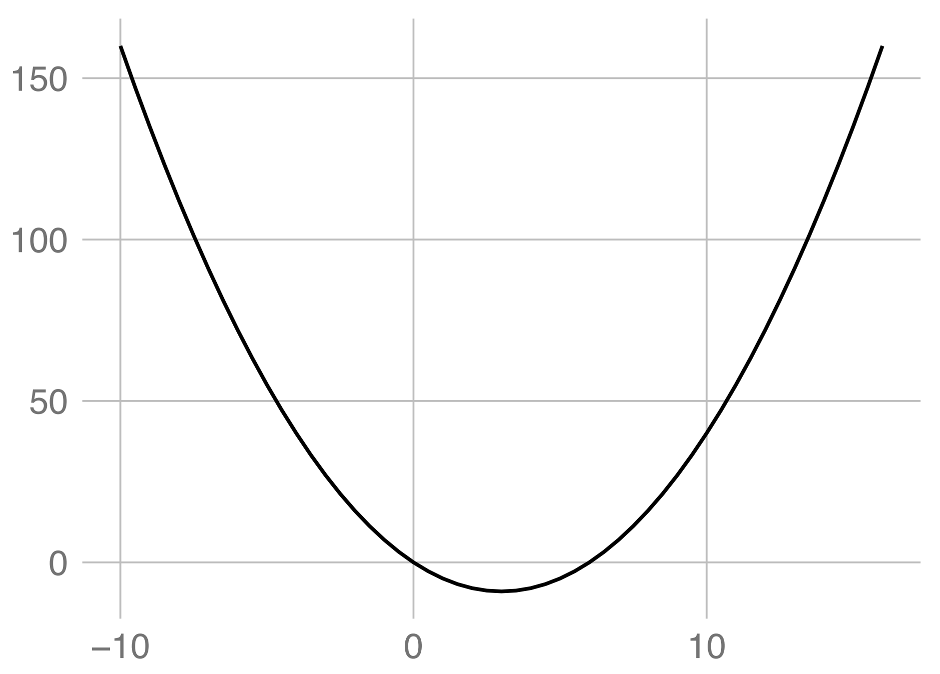
1. Analytic
2. Algorithmic

Analytical Optimization

Ex: Find what value for x will maximize the function $f(x) = -x^2 + 6x$

minimize $\quad f(x) = x^2 - 6x$

with respect to x



First order necessary condition

x^* is a "stationary point" when

$$\frac{df(x^*)}{dx} = 0$$

$$\frac{df}{dx} = 2x - 6 = 0 \longrightarrow x^* = 3$$

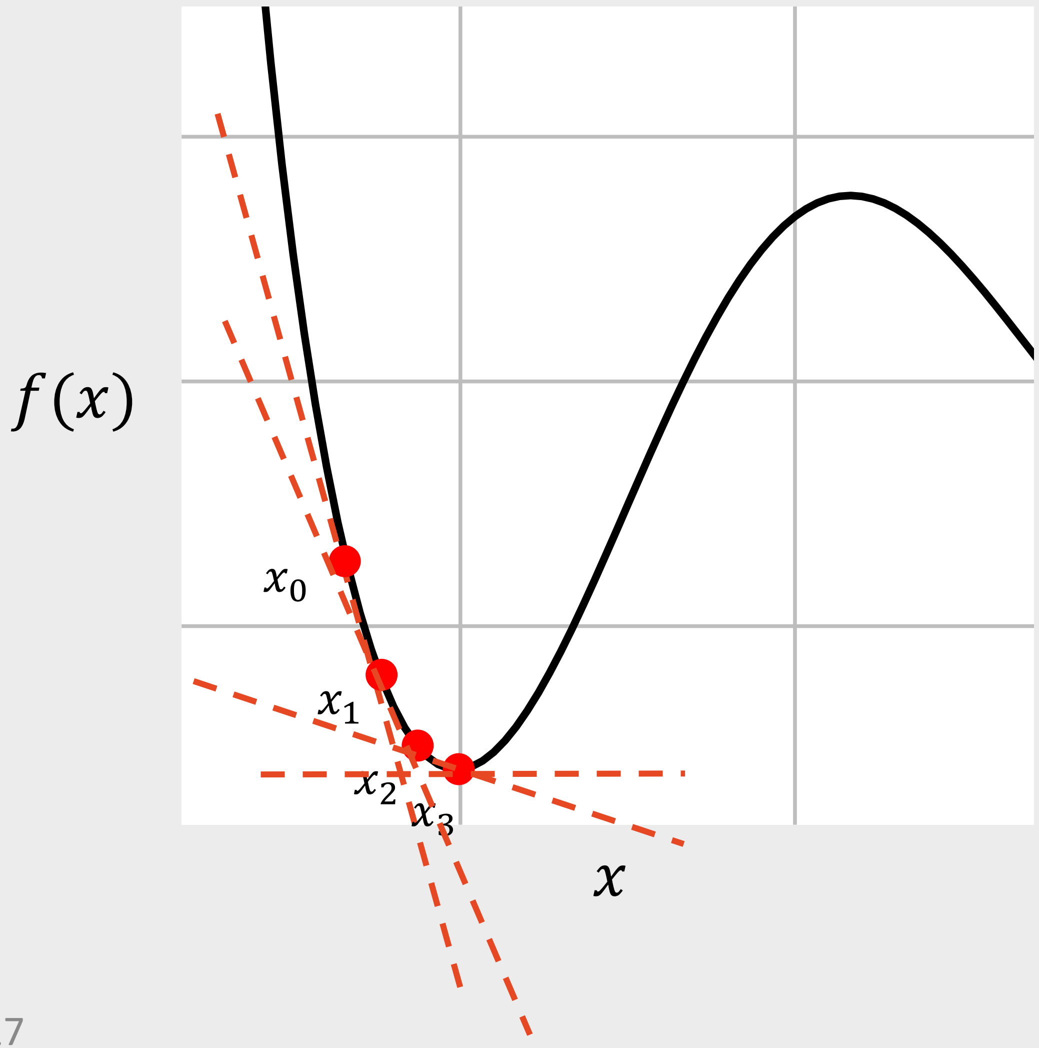
Second order sufficiency condition

x^* is a local maximum / minimum when

$$\frac{d^2f(x^*)}{dx^2} < 0 \qquad \frac{d^2f(x^*)}{dx^2} > 0$$

$$\frac{d^2f}{dx^2} = 2 \longrightarrow$$ $x^*$is a local minimum

Optimization Algorithms



<u>Gradient Descent Method</u>:

1. Choose a starting point, x_0
2. At that point, compute the gradient, $\nabla f(x_0)$
3. Compute the next point, with a step size γ :

$$x_{n+1} = x_n - \gamma \nabla f(x_n)$$

Very small number

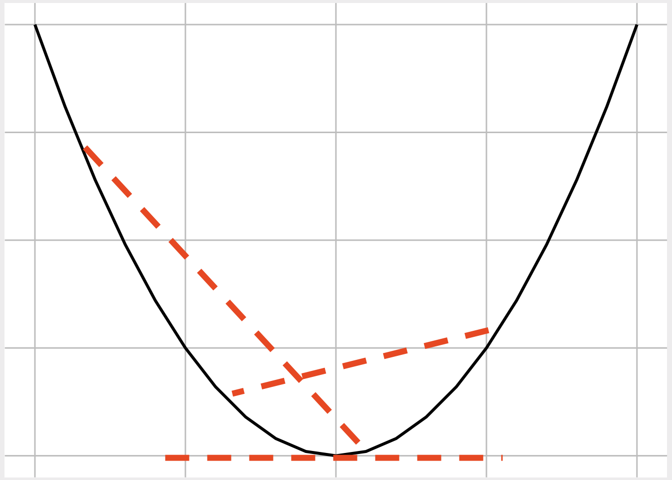
*Stop when $\nabla f(x_n) < \delta$

or

*Stop when $(x_{n+1} - x_n) < \delta$

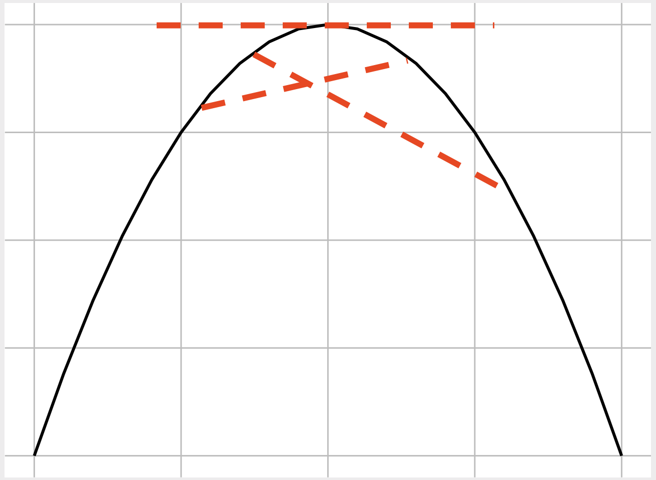
Convex & Concave Functions

Convex



When minimizing a convex function, any *local* minimum is a *global* minimum

Concave



When maximizing a concave function, any *local* maximum is a *global* maximum

Practice Question 2

Consider the following function:
$f(x) = x^2 - 6x$

The gradient is:
$\nabla f(x) = 2x - 6$

Using the starting point x =1 and the step size $\gamma = 0.3$, apply the gradient descent method to compute the next **three** points in the search algorithm.

Hints:

1. Remember the gradient descent method:

$$x_{n+1} = x_n - \gamma \nabla f(x_n)$$

Practice Question 3

Consider the following function:

$f(\underline{x}) = x_1^2 + 4x_2^2$

The gradient is:

$\nabla f(\underline{x}) = \begin{bmatrix} 2x_1 \\ 8x_2 \end{bmatrix}$

Using the starting point $\underline{x}_0$ =[1, 1] and the step size $\gamma = 0.15$, apply the gradient descent method to compute the next **three** points in the search algorithm.

Hints:

1. Remember the gradient descent method:
$$x_{n+1} = x_n - \gamma \nabla f(x_n)$$
2. In *R*, use the c() function to create a vector.

Estimating Utility Model Coefficients Using Maximum Likelihood Estimation

$$\begin{aligned}\tilde{u}_j &= v_j + \tilde{\varepsilon}_j \\ &= \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \tilde{\varepsilon}_j \\ &= \boldsymbol{\beta}' \mathbf{x}_j + \tilde{\varepsilon}_j\end{aligned}$$

Estimate $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_n]$
by maximizing the likelihood function

$$\boxed{\text{minimize } -log\mathcal{L} = -\sum_{j=1}^{J} P_j(\boldsymbol{\beta}|\mathbf{x})^{y_j} \quad \text{with respect to } \boldsymbol{\beta}}$$

$y_j = 1$ if alternative j was chosen
$y_j = 0$ if alternative j was not chosen

For logit model:

$$P_j = \frac{e^{v_j}}{\sum_{k=1}^{J} e^{v_k}} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_j}}{\sum_{k=1}^{J} e^{\boldsymbol{\beta}'\mathbf{x}_k}}$$