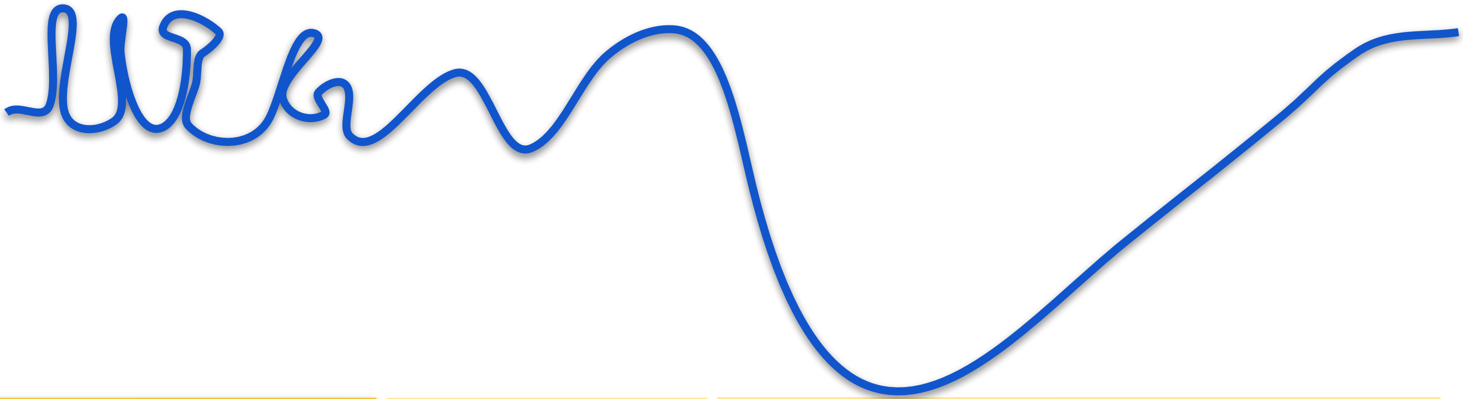
BTech, MTech, PhD Open Elective Course

DA623: Computing with Signals

Neeraj Kumar Sharma (Neeks)
neerajs@iitg.ac.in

Computing with Signals



DA 623

Jan - May 2024

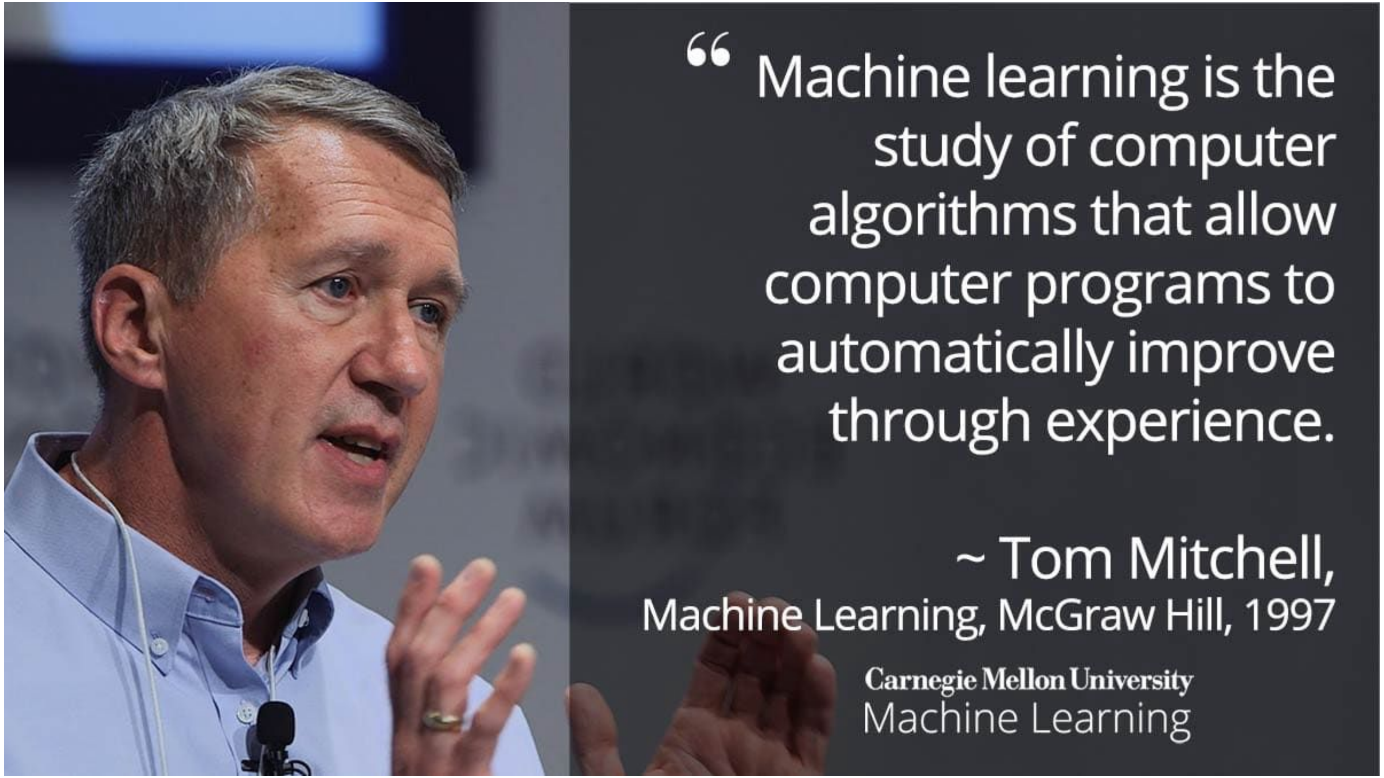
IIT Guwahati

Instructors: Neeraj Sharma

Lecture-25

Week 13
Machine Learning

What is Machine Learning?

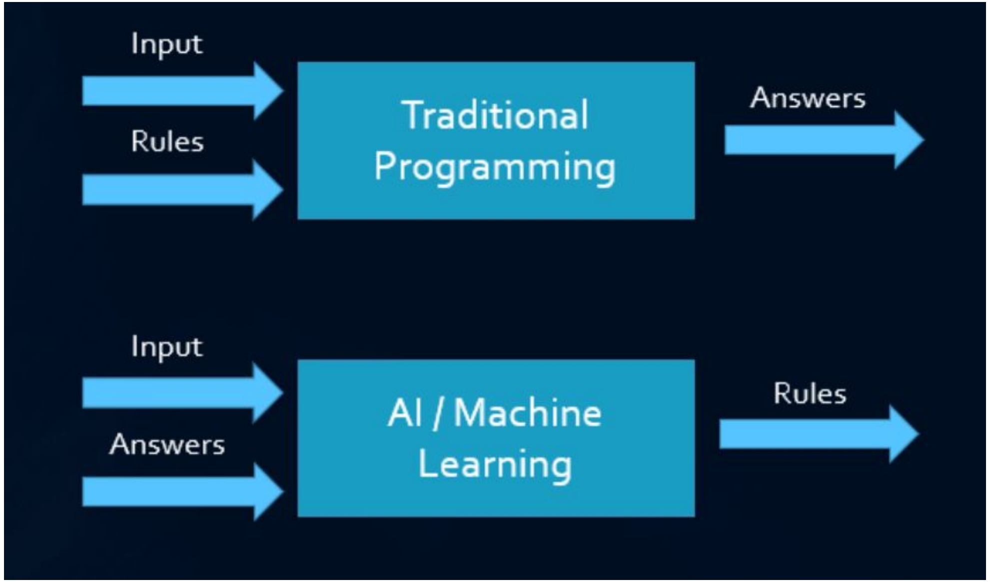


What is Machine Learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Tom Mitchell, CMU)

What is Machine Learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." (Tom Mitchell, CMU)



Rule-based AI Systems

These systems operate by applying a set of predefined rules to incoming data or situations to make decisions or take actions.

Also known as Expert systems or Knowledge-based systems.

Rule-based AI Systems

These systems operate by applying a set of predefined rules to incoming data or situations to make decisions or take actions.

Also known as Expert systems or Knowledge-based systems.

These systems were designed to emulate the decision-making process of human experts by encoding their knowledge and expertise into a set of rules.

```
Rule 1: If patient has fever and cough, then diagnose as flu.
Rule 2: If patient has rash and itching, then diagnose as allergic reaction.
Rule 3: If patient has chest pain and shortness of breath, then diagnose as
heart attack.
```

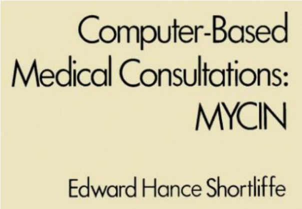
Rule-based AI Systems

MYCIN, a rule-based AI systems, was developed in the 1970s at Stanford University to assist physicians in diagnosing infectious blood diseases.

It operated using a fairly simple inference engine and a knowledge base of certain number of rules. It would query the physician running the program via a long series of simple yes/no or textual questions.

The output will include:

- possible culprit bacteria, and its confidence in each diagnosis' probability
- reasoning behind each diagnosis
- recommended a course of drug treatment



What is Machine Learning?

“The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.”
(Kevin Patrick Murphy, Machine Learning: A Probabilistic Perspective)

What is Machine Learning?

"The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest."
(Kevin Patrick Murphy, Machine Learning: A Probabilistic Perspective)

"Machine learning is thus closely related to the fields of statistics and data mining, but differs slightly in terms of emphasis ..."

Machine Learning versus Statistics

Both statistics and machine learning aim to “detect patterns in data” by building predictive models.

Statistics: use this as a tool to learn something about the world.

- Focus on simple, interpretable models
- Develop theoretical analysis
- Work out statistical guarantees under some assumptions

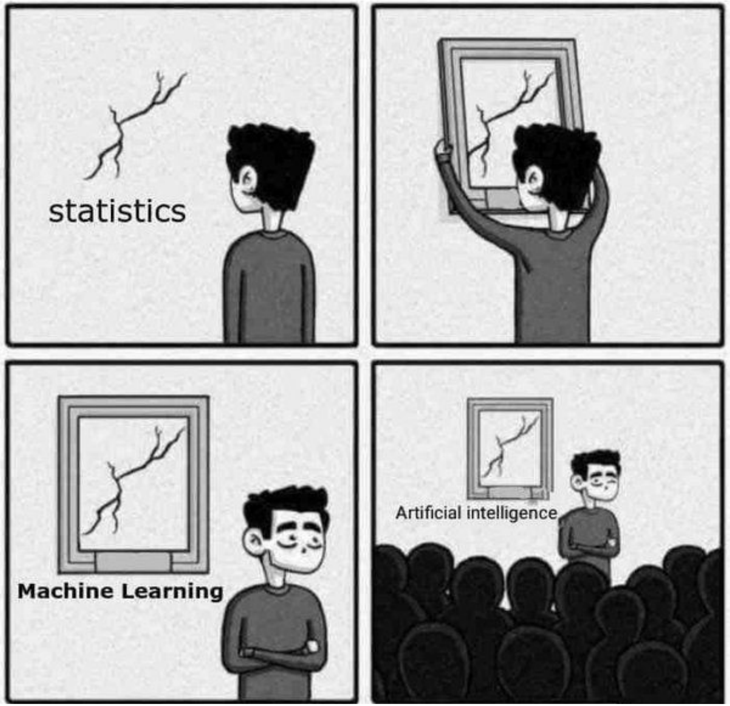
Machine learning: use this as a tool to actually make useful predictions.

- Focus on complicated, competitive models
- Use large datasets
- Be pragmatic
- Give up on inference (significance)

Machine Learning versus Statistics

In practice, it is difficult to categorize.

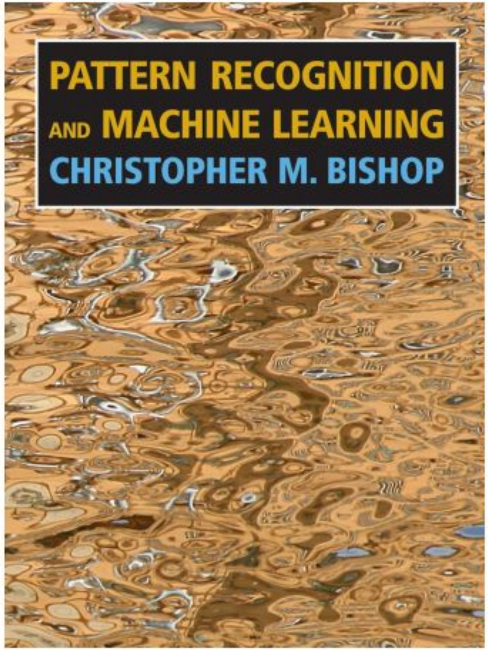
Both can be considered to overlap over a continuous spectrum.

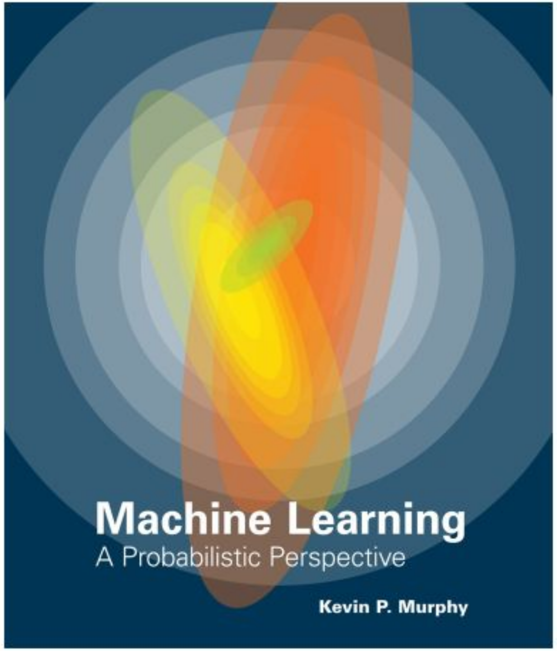


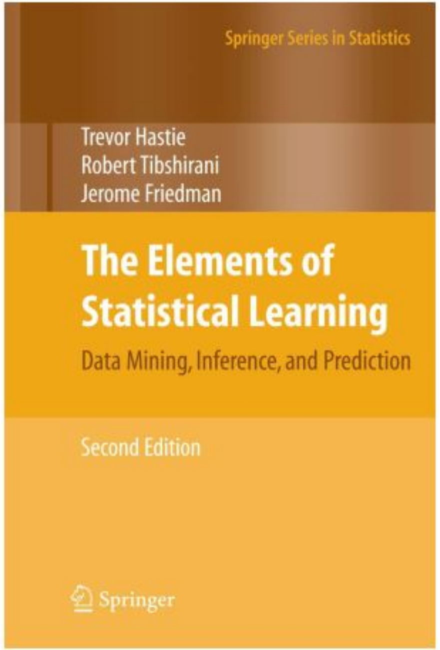
Ref: Internet Image Search. Could not find the source.

Statistical Learning

Statistics + Machine Learning







Types of Machine Learning Problems

Supervised Learning
Example: distinguish photos of goose from photos of dogs

Unsupervised Learning
Example: figure out that goose and dogs are different animals

Reinforcement Learning
Example: learning to ride a cycle, playing a game

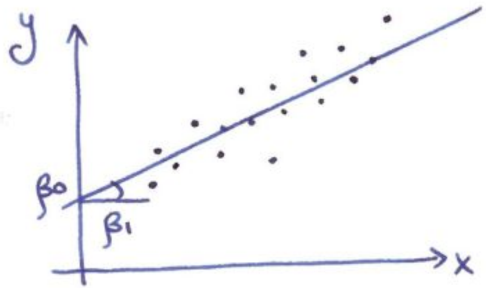
Simple Linear Regression

Acknowledgement

Dmitry Kobak

Videos: Tübingen Machine Learning / Introduction to Machine Learning

Simple Linear Regression



Supervised learning problem. Regression (not classification) problem.

Training data: $\{(x_i, y_i)\}_{i=1}^{n}$.

Model: $\hat{y} = f(x) = \beta_0 + \beta_1 x$.

Two coefficients: *intercept* and *slope*. We want to *fit* the model to the data.

Loss Function

To fit the model means to find β_0 and β_1 so that $f(x_i) \approx y_i$.

Loss function (aka cost function):

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

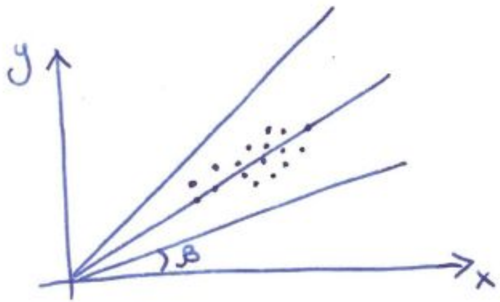
Mean squared error (MSE). Why MSE?

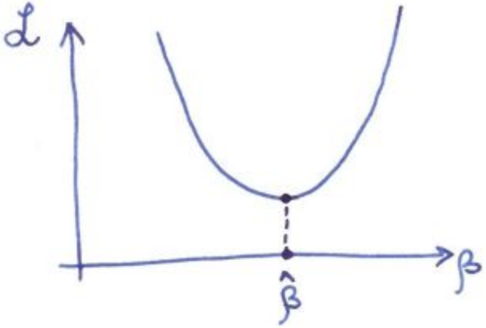
Ordinary least squares (OLS).

Loss Function

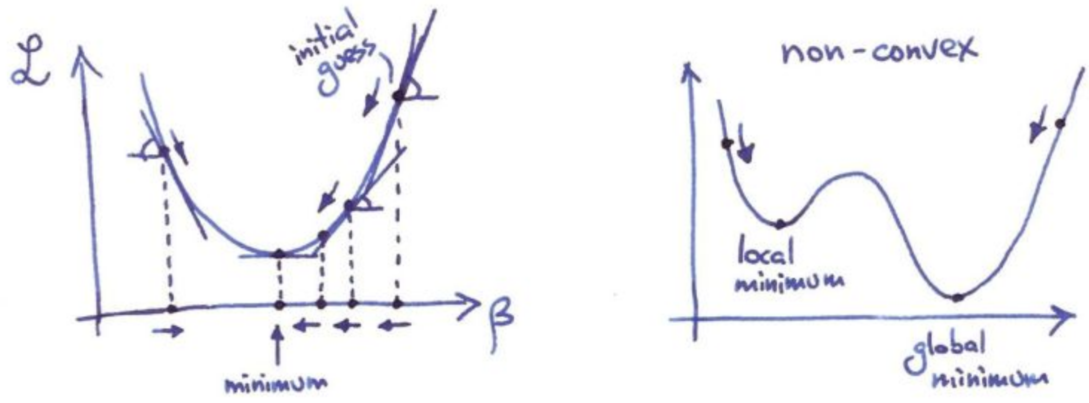
Consider slope-only model: $f(x) = \beta x$.

The loss: $\mathcal{L}(\beta) = \frac{1}{n}\sum_i (y_i - \beta x_i)^2$.





Solution using Gradient Descent



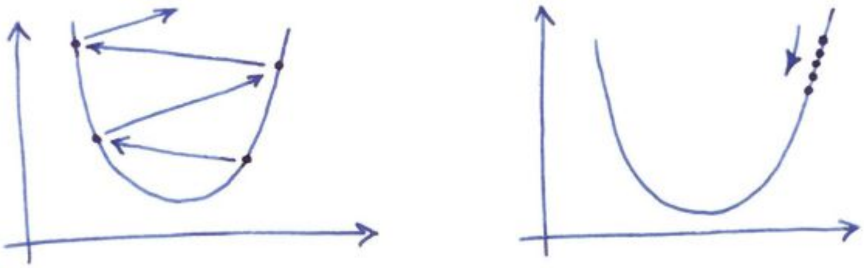
Update rule:

$$\beta \leftarrow \beta - \eta \frac{d\mathcal{L}(\beta)}{d\beta}.$$

Here η is the learning rate.

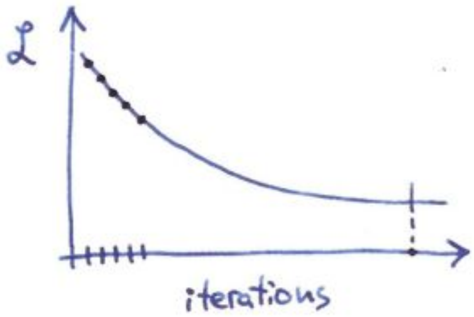
Learning Rate

$$\beta \leftarrow \beta - \eta \frac{d\mathcal{L}(\beta)}{d\beta}$$



Too large η — divergence. Too small η — slow convergence.

Learning Rate



$$\beta \leftarrow \beta - \eta \frac{d\mathcal{L}(\beta)}{d\beta}$$

Gradient Descent

We need to compute the derivative of the loss:

$$\mathcal{L}(\beta) = \frac{1}{n}\sum_i (y_i - \beta x_i)^2.$$

We get:

$$\mathcal{L}'(\beta) = \frac{1}{n}\sum_i 2(y_i - \beta x_i)(-x_i) = -\frac{2}{n}\sum_i x_i(y_i - \beta x_i).$$

Analytical Solution

$$\mathcal{L}'(\beta) = -\frac{2}{n}\sum_i x_i(y_i - \beta x_i).$$

At the minimum:

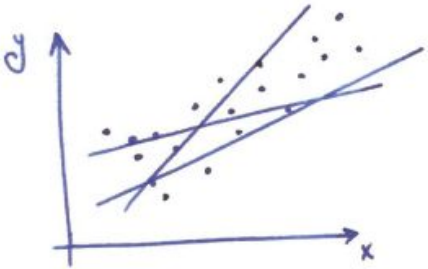
$$\sum_i x_i y_i - \hat{\beta}\sum_i x_i^2 = 0.$$

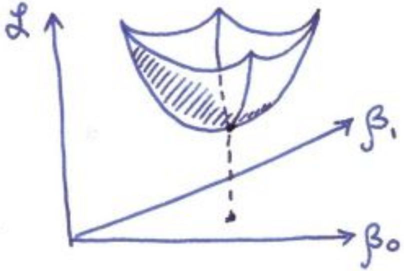
We obtain:

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

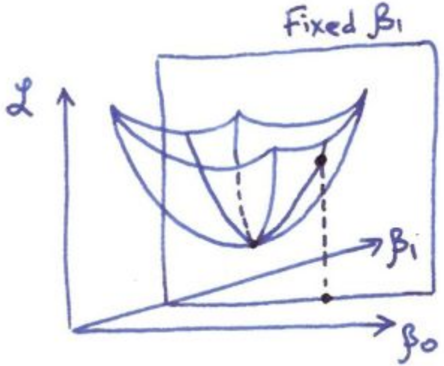
Simple Linear Regression

$$\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$





Using Partial Derivatives



$$\beta_0 \leftarrow \beta_0 - \eta \frac{\partial \mathcal{L}}{\partial \beta_0}$$

$$\beta_1 \leftarrow \beta_1 - \eta \frac{\partial \mathcal{L}}{\partial \beta_1}$$

Gradients

Update rules for each parameter:

$$\beta_0 \leftarrow \beta_0 - \eta \frac{\partial \mathcal{L}}{\partial \beta_0}$$
$$\beta_1 \leftarrow \beta_1 - \eta \frac{\partial \mathcal{L}}{\partial \beta_1}$$

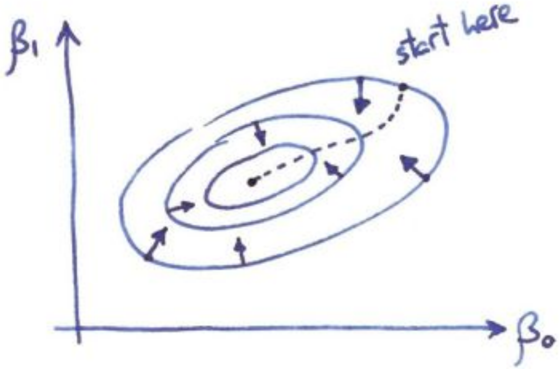
In vector form:

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \nabla \mathcal{L}.$$

Gradients

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\vec{\beta} \leftarrow \vec{\beta} - \eta\nabla\mathcal{L}$$



Computing the Gradient

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

We need partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = -\frac{2}{n}\sum(y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial \mathcal{L}}{\partial \beta_1} = -\frac{2}{n}\sum(y_i - \beta_0 - \beta_1 x_i)x_i$$

Exercise: derive the analytical solution for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Name	Operator	Function	Maps	Example Value
Derivative	$\frac{d}{dx}$	$\frac{df}{dx}(x)$	$\mathbb{R} \to \mathbb{R}$	2.5
Partial Derivative	$\frac{\partial}{\partial x}$	$\frac{\partial f}{\partial x}(x, y, z)$	$\mathbb{R}^3 \to \mathbb{R}$	2.5
Gradient	∇	$\nabla f(x, y, z)$	$\mathbb{R}^3 \to \mathbb{R}^3$	$\begin{bmatrix} 2.5 \\ 0 \\ -1 \end{bmatrix}$

Name	Operator	Function	Maps	Example Value
Derivative	$\frac{d}{dx}$	$\frac{df}{dx}(x)$	$\mathbb{R} \to \mathbb{R}$	2.5
Partial Derivative	$\frac{\partial}{\partial x}$	$\frac{\partial f}{\partial x}(x, y, z)$	$\mathbb{R}^3 \to \mathbb{R}$	2.5
Gradient	∇	$\nabla f(x, y, z)$	$\mathbb{R}^3 \to \mathbb{R}^3$	$\begin{bmatrix} 2.5 \\ 0 \\ -1 \end{bmatrix}$

Since a gradient is a vector, we can talk about its magnitude and direction.

- The magnitude is $\|\nabla f\|$ and tells us **how fast things are changing**.
- The direction is $\frac{\nabla f}{\|\nabla f\|}$ and tells us **the direction of fastest change** or the steepest direction.

Taylor Series in 1-D

***Theorem* (Taylor)**: *Suppose n is a positive integer and $f : \mathbb{R} \to \mathbb{R}$ is n times differentiable at a point x_0. Then*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + R_n(x, x_0),$$

where the remainder R_n satisfies

$$R_n(x, x_0) = o(|x - x_0|^n) \text{ as } x \to x_0.$$

Taylor Series in d-D

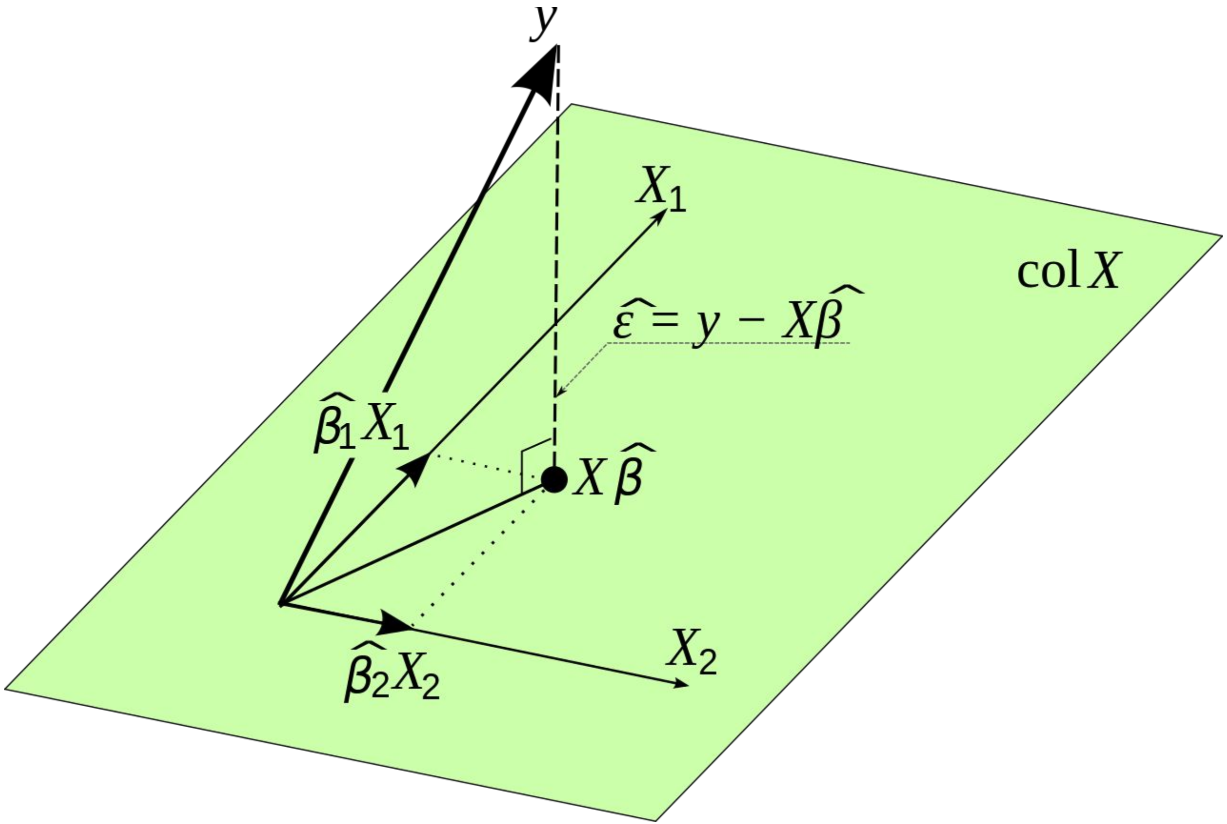
Theorem: *Suppose* $f : \mathbb{R}^d \to \mathbb{R}$ *is differentiable on* $N_r(\mathbf{x}_0)$. *Then for any* $\mathbf{x} \in N_r(\mathbf{x}_0)$, *there exists* $\bar{\mathbf{x}}$ *on the line segment connecting* $\mathbf{x}$ *and* $\mathbf{x}_0$ *such that*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\bar{\mathbf{x}})^\top(\mathbf{x} - \mathbf{x}_0)$$

Taylor Series in d-D

Theorem: *Suppose* $f : \mathbb{R}^d \to \mathbb{R}$ *is differentiable on* $N_r(\mathbf{x}_0)$. *Then for any* $\mathbf{x} \in N_r(\mathbf{x}_0)$, *there exists* $\bar{\mathbf{x}}$ *on the line segment connecting* $\mathbf{x}$ *and* $\mathbf{x}_0$ *such that*

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\bar{\mathbf{x}})^\top (\mathbf{x} - \mathbf{x}_0)$$


y
X_1
col X
$\widehat{\varepsilon} = y - X\widehat{\beta}$
$\widehat{\beta}_1 X_1$
$X\widehat{\beta}$
$\widehat{\beta}_2 X_2$
X_2