IOT Online Course

Fundamentals of Artificial Intelligence

F-AI-4: Regression

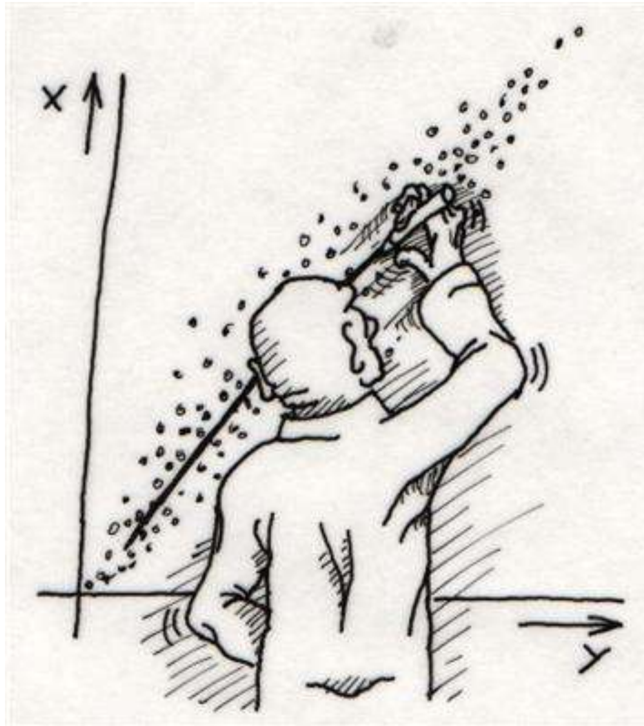
Prof. Congduc Pham
http://www.univ-pau.fr/~cpham
Université de Pau, France

Slides realized by J. Mantilla

IoT – from idea to reality

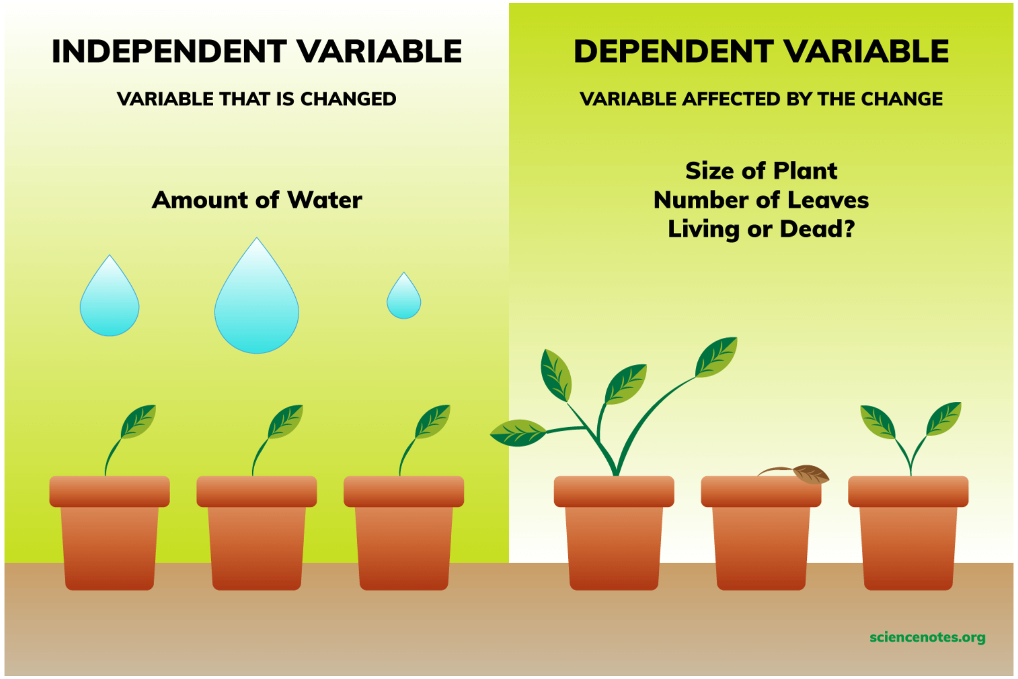
What is Regression?

- Regression takes a group of random events, thought to be predicting another event, and tries to find a mathematical relationship between them.
- Events are represented by variables.
- Used in statistics to find trends in data.
- Implemented in Machine Learning (**ML**) as supervised algorithm where the predicted output is expressed in Real numbers.
- Used to predict values within a continuous range.



Types of Variables

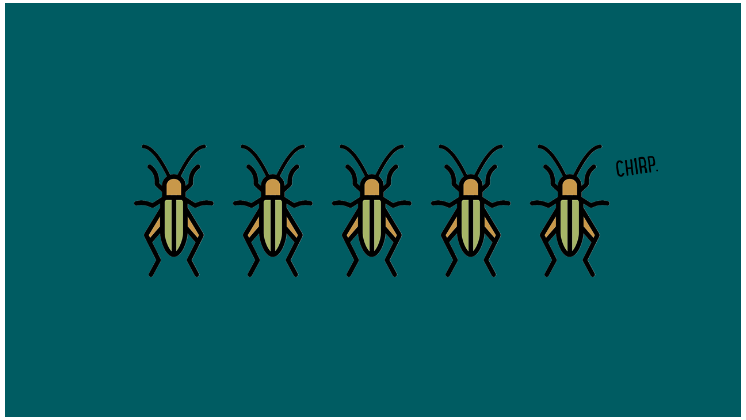
- **Dependent variables:** the main event or factor to understand or predict. Also known as *explanatory variable.*
- **Independent variables:** the events or factors suspected to have an impact on the dependent variable. Also known as *response variable.*

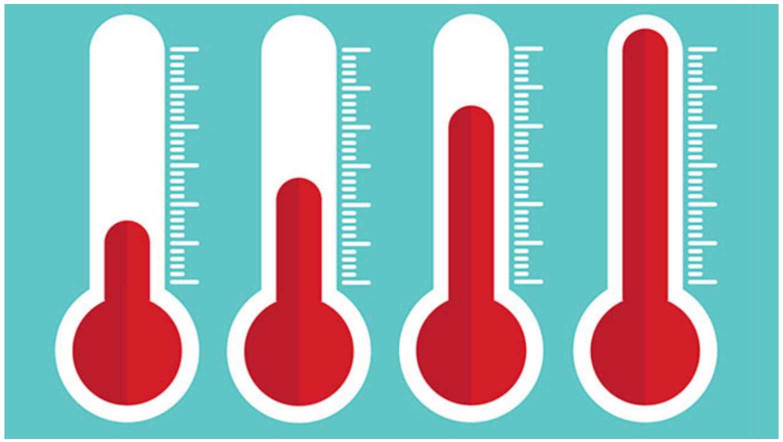


Example

Number of times a population of crickets chirp to predict the temperature.

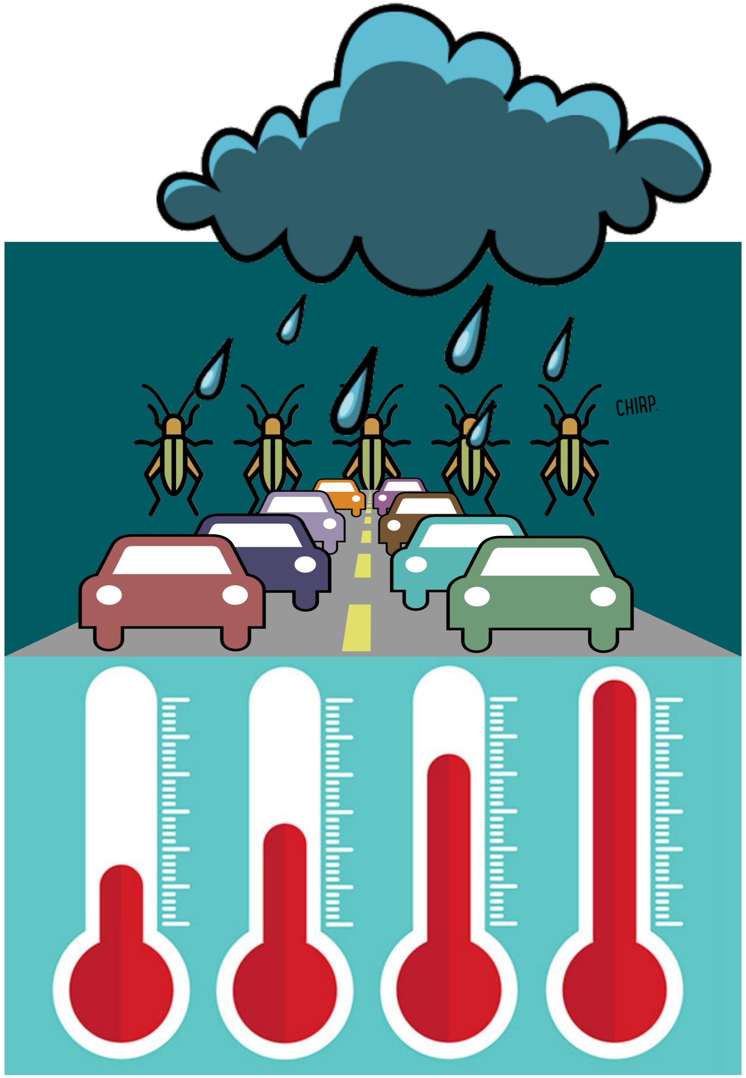
- The number of chirps is the **independent** variable
- The temperature is the **dependent** variable





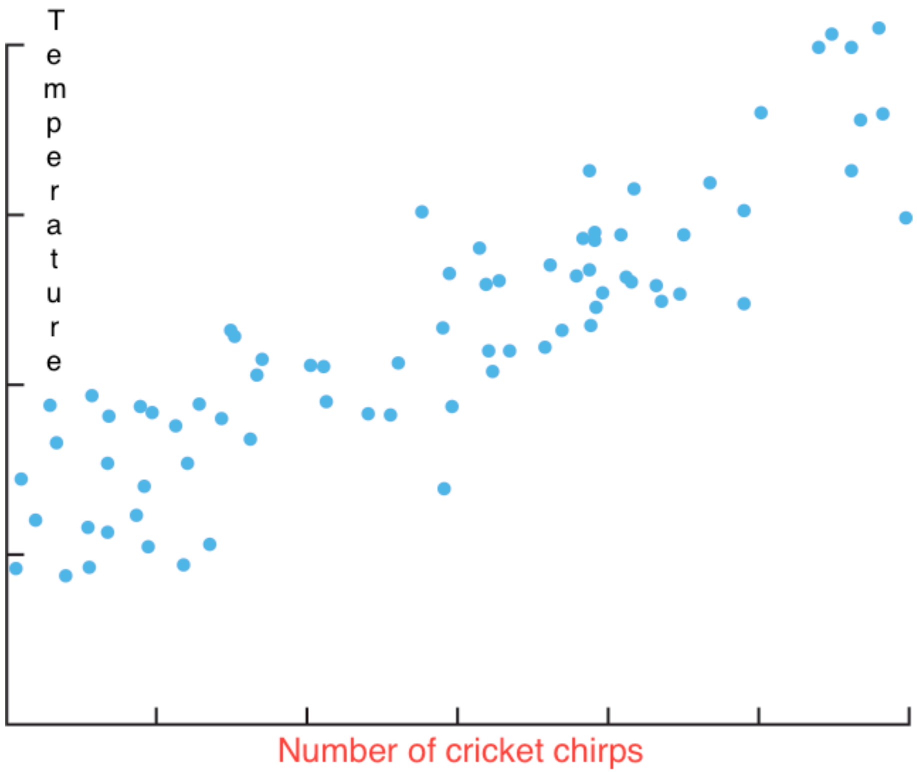
Types of Regression

- **Simple regression:** single independent variable for a single dependent variable. It is very common to name the independent variable as x as and **Y** as the dependent variable.
  x: number of cricket chirps
  Y: temperature
- **Multivariable regression:** multiple independent variables, x_1, x_2, x_3, for a dependent variable **Y**.
  x_1: number of cricket chirps
  x_2: rainfall
  x_3: automobile traffic
  Y: temperature



Scatter Plot

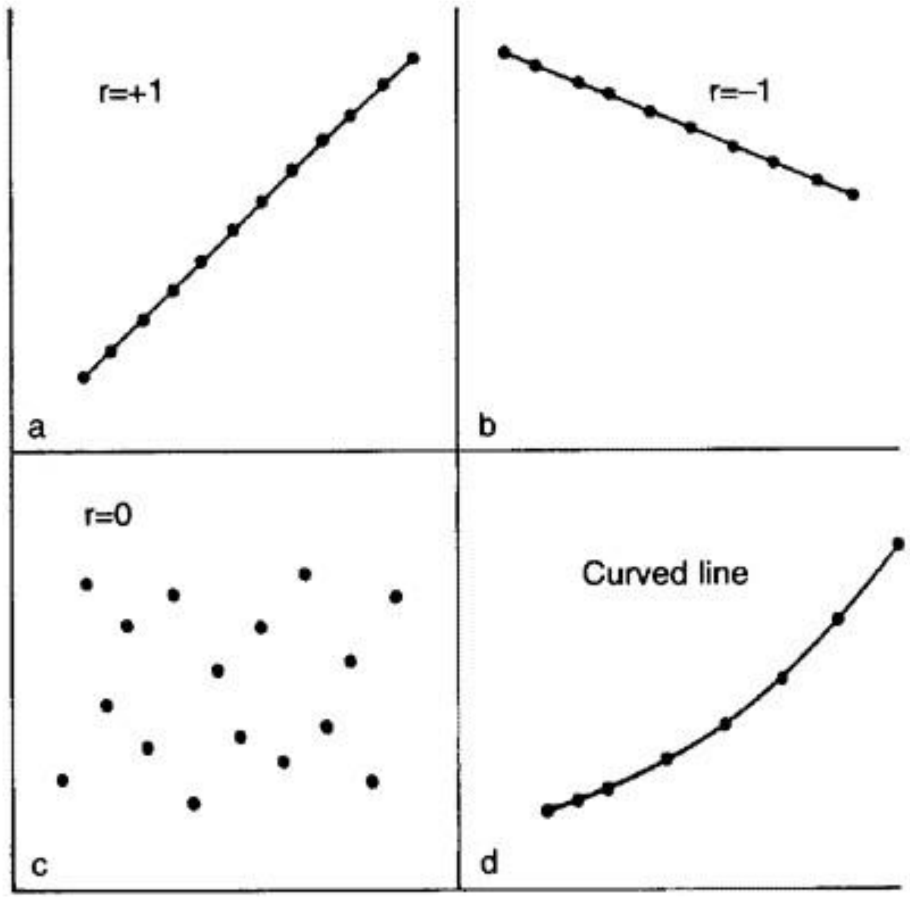
- Data gathering on the variables in question
- The vertical scale represents one set of measurements and the horizontal scale the other



Correlation Coefficient

- Measures the degree of association, denoted by r or R, sometimes called Pearson's correlation.
- Measured on a scale that varies from + 1 through 0 to – 1.
- When one variable increases as the other increases the correlation is *positive*
- When one variable decreases as the other increases the correlation is *negative*.
- Complete correlation between two variables is expressed by either + 1 or -1.
- Complete absence of correlation is represented by 0.
- If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

Correlation Chart



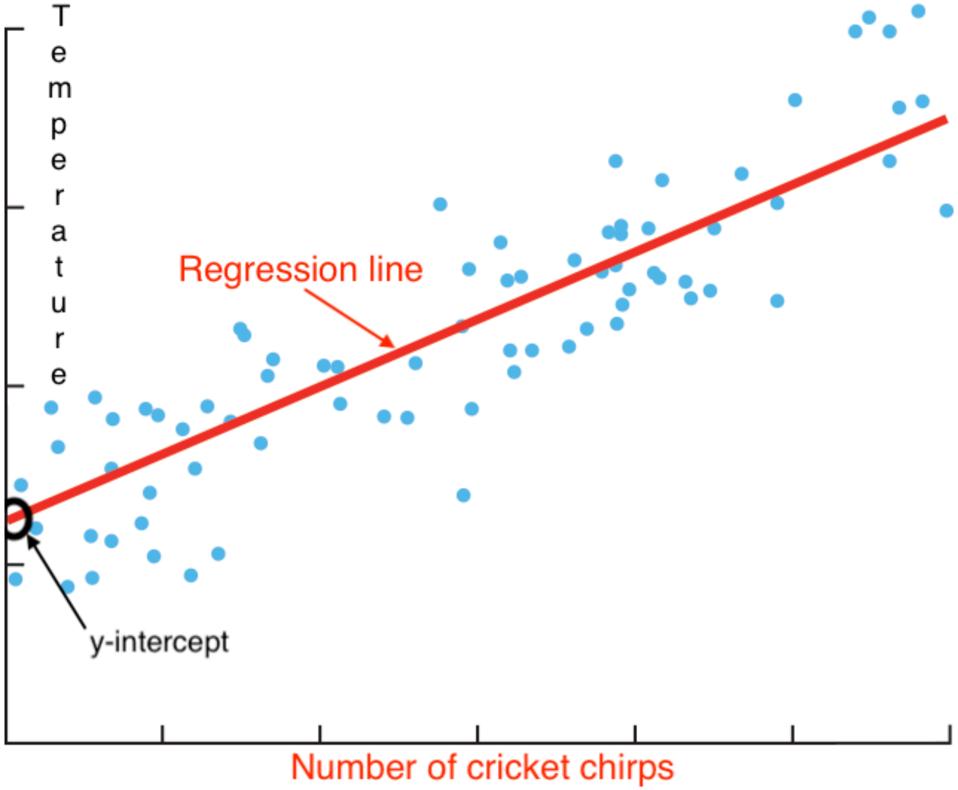
Guidelines

- It's widely stated in the literature, as rule of thumb, to have more than 100 items in the sample (100 dots in the scatter plot). Safer if at least 200 observations or better yet—more than 400.
- Don't do a regression analysis unless you have already found at least a moderately strong correlation between the variables.
- A good rule of thumb is a correlation coefficient at 0.5 or beyond, either positive or negative
- If data don't resemble a line to begin with, don't try to use a line to fit the data and make predictions.

Prof. Congduc Pham
http://www.univ-pau.fr/~cpham

Linear Regression

- A linear relationship to predict the (average) numerical value of **Y** for a given value of x using a straight line, called the ***regression line***.
- Knowing the *slope* and the **y-**intercept of that regression line, it is possible to plug in a value for x and *predict* the average value for **Y**. In other words, predict the average **Y** from x.



Equations

- Simple linear regression: $\mathbf{Y} = a\boldsymbol{x} + b + u$
- Multiple linear regression: $\mathbf{Y} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + a_3\boldsymbol{x}_3 + \ldots + a_i\boldsymbol{x}_i + b + u$

Y: the variable to predict

$\boldsymbol{x}$: the variable used to predict **Y**

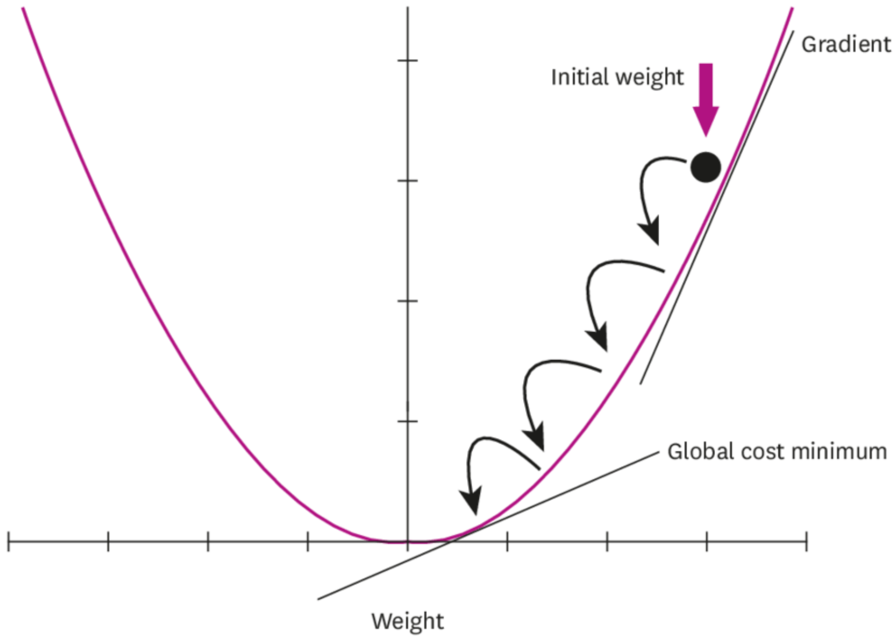
a: the slope

b: the **y**-intercept

u: the regression residual.

Cost Function

- Helps to figure out the best possible values for a (the slope) and b (the **y**-intercept) which would provide the best *regression line* for data points (dots in the scatter plot).
- To find best values for a and b, this search problem is converted into a minimization problem whereby to *minimize* the **error** between the **predicted** value and the **actual** value.



Mean Squared Error (MSE)

MSE is used as cost function to *minimize* the **error**

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

- $\boldsymbol{n}$ is the total number of observations (data points)
- $\mathbf{y}_i$ is the actual value
- $\tilde{y}_i$ is the predicted value

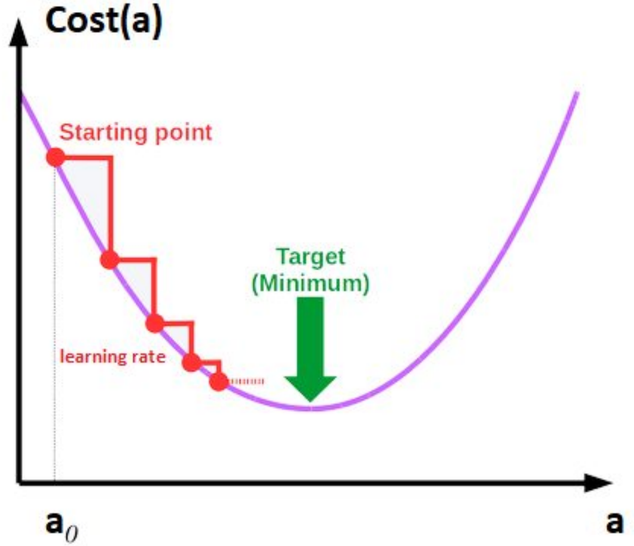
Square the error difference and sum over all data points and divide that value by the total number of data points.

Provides the average squared error over all the data points.

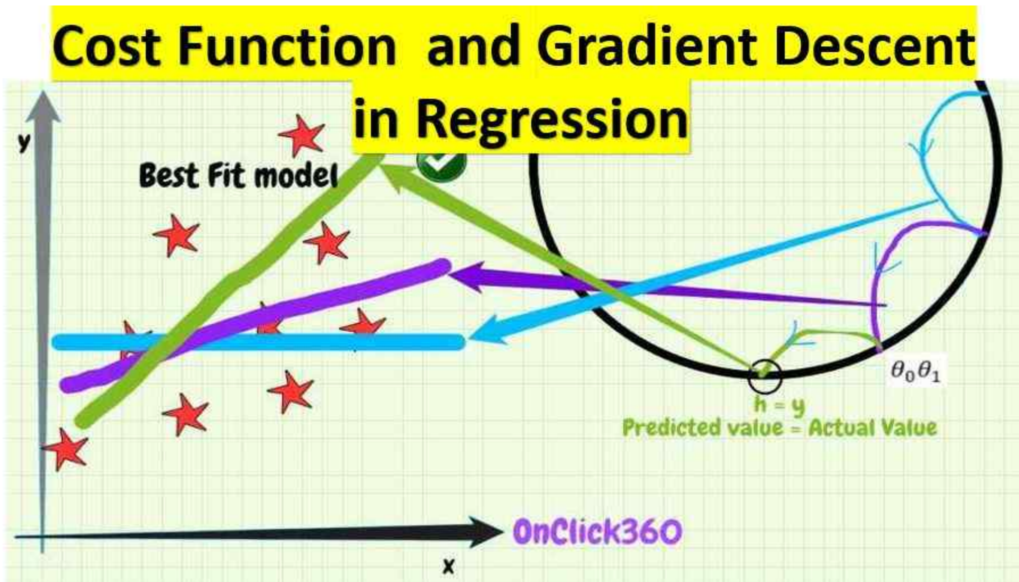
Find a *minimum* **error** value using the Cost Function (MSE).

Gradient Descent

- Gradient Descent is a method of updating a and b to reduce the Cost Function (MSE).
- The idea is to start with some values for a and b then change these values iteratively to reduce the cost.
- Gradient Descent helps on how to change the values.



http://aishelf.org/gradient-descent/



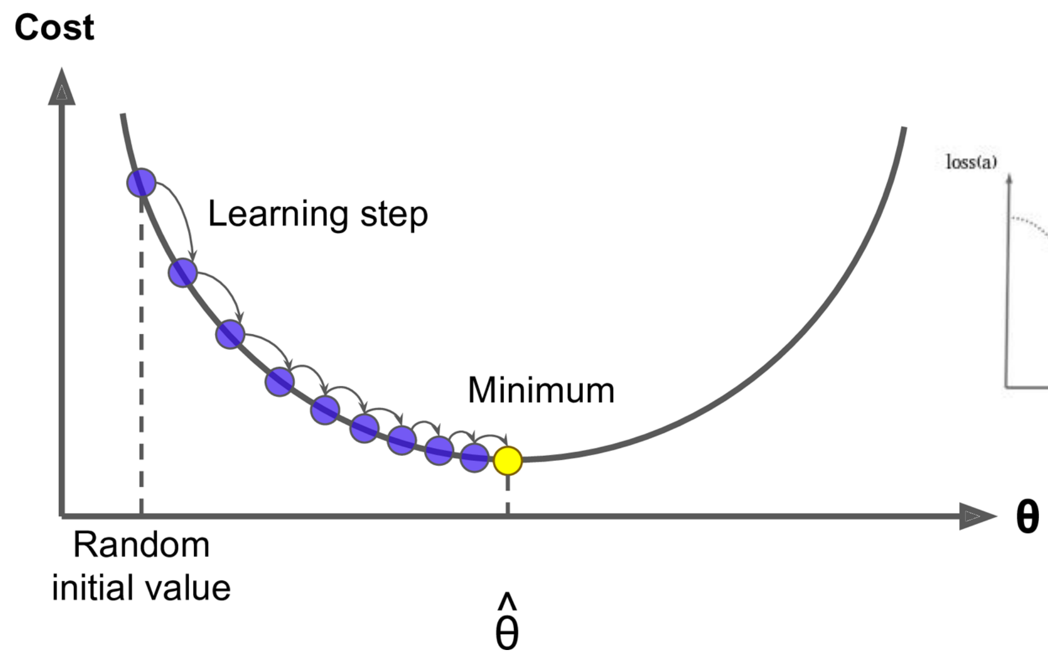
Partial Derivatives of Cost Function

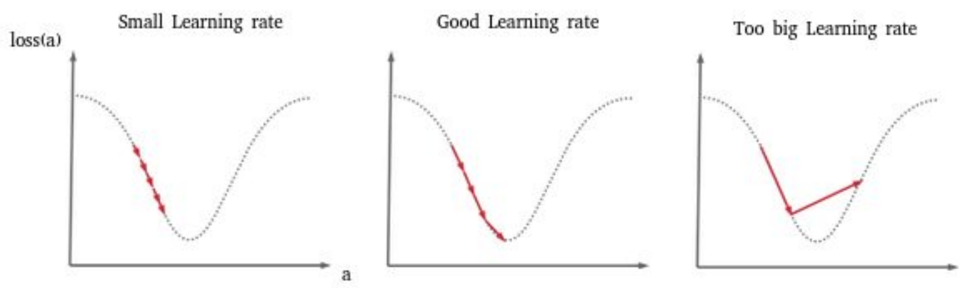
- The gradients of the Cost Function (MSE) are used to update a and b values.
- To find these gradients, take **partial derivatives** (because we have several variables) with respect to a and b
- Alpha (α) is the *learning rate* which is a hyper-parameter specified by the human observer.

$$
\begin{aligned}
&\text{repeat until convergence } \{ \\
&\quad \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \\
&\quad \theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)} \\
&\}
\end{aligned}
$$

Learning Rate

- A smaller *learning rate* could get closer to the *minima* but takes more time to reach the *minima*
- A larger *learning rate* converges sooner but there is a chance that you could overshoot the *minima*.





Regression Line

- Finally, got the **best fit line** or **best regression line.**
- Can predict new **Y** values using the best *regression line*.