# LOGISTIC REGRESSION

# WHAT IS LOGISTIC REGRESSION ?

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
- The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.
- In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

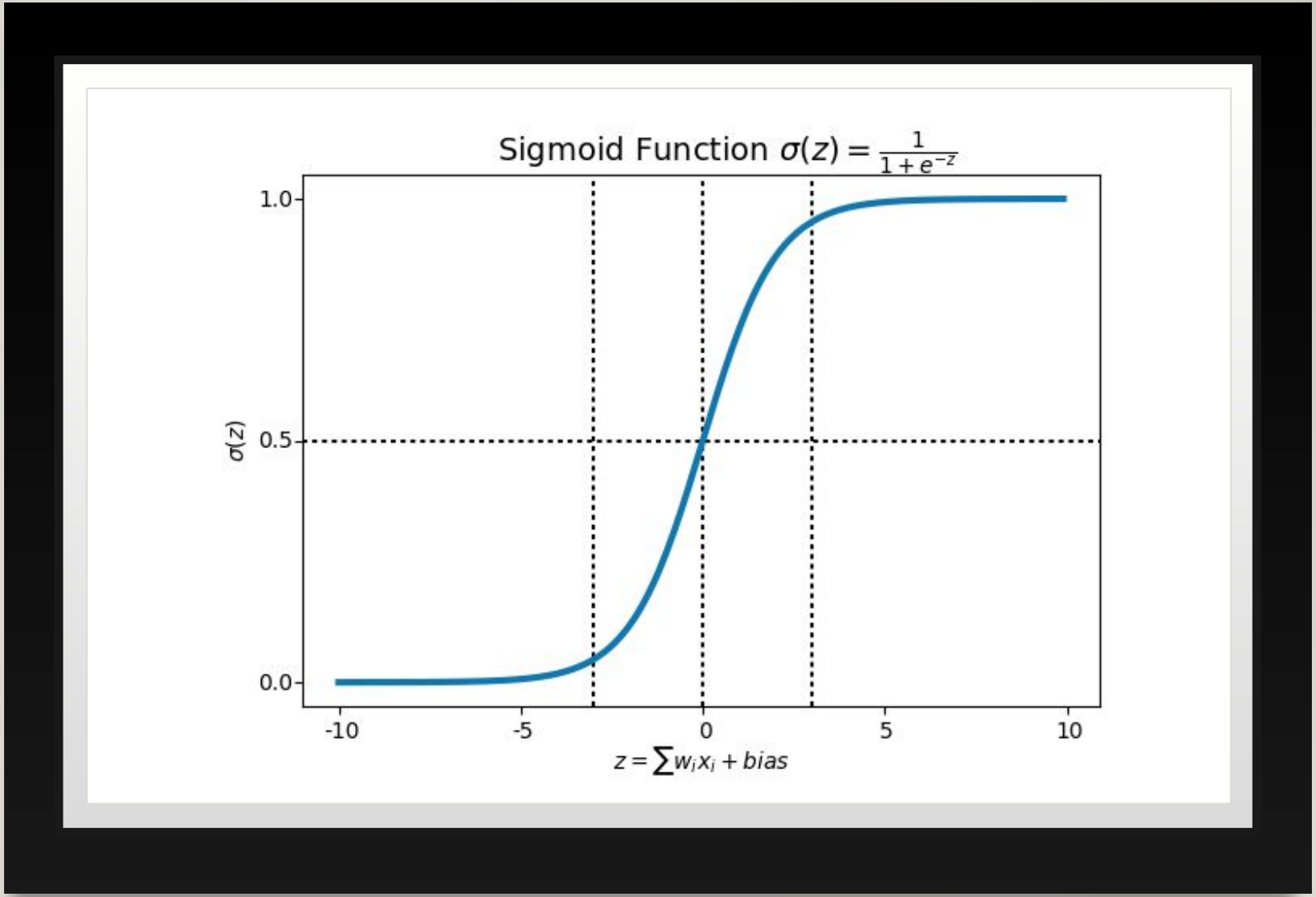
# TYPES OF LOGISTIC REGRESSION

- Based on number of categories, Logistic regression can be divided into following types −
  - Binary/Binomial: In such a kind of classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.
  - Multinomial: In such a kind of classification, dependent variable can have 3 or more possible ***unordered*** types or the types having no quantitative significance. For example, these variables may represent "Type A" or "Type B" or "Type C".
  - Ordinal: In such a kind of classification, dependent variable can have 3 or more possible ***ordered*** types or the types having a quantitative significance. For example, these variables may represent "poor" or "good", "very good", "Excellent" and each category can have the scores like 0,1,2,3.

# LOGISTIC FUNCTION

- Logistic regression is named for the function used at the core of the method, the logistic function.
- The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment.
- It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
- Formula: 1 / (1 + e^(-value)), where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform.

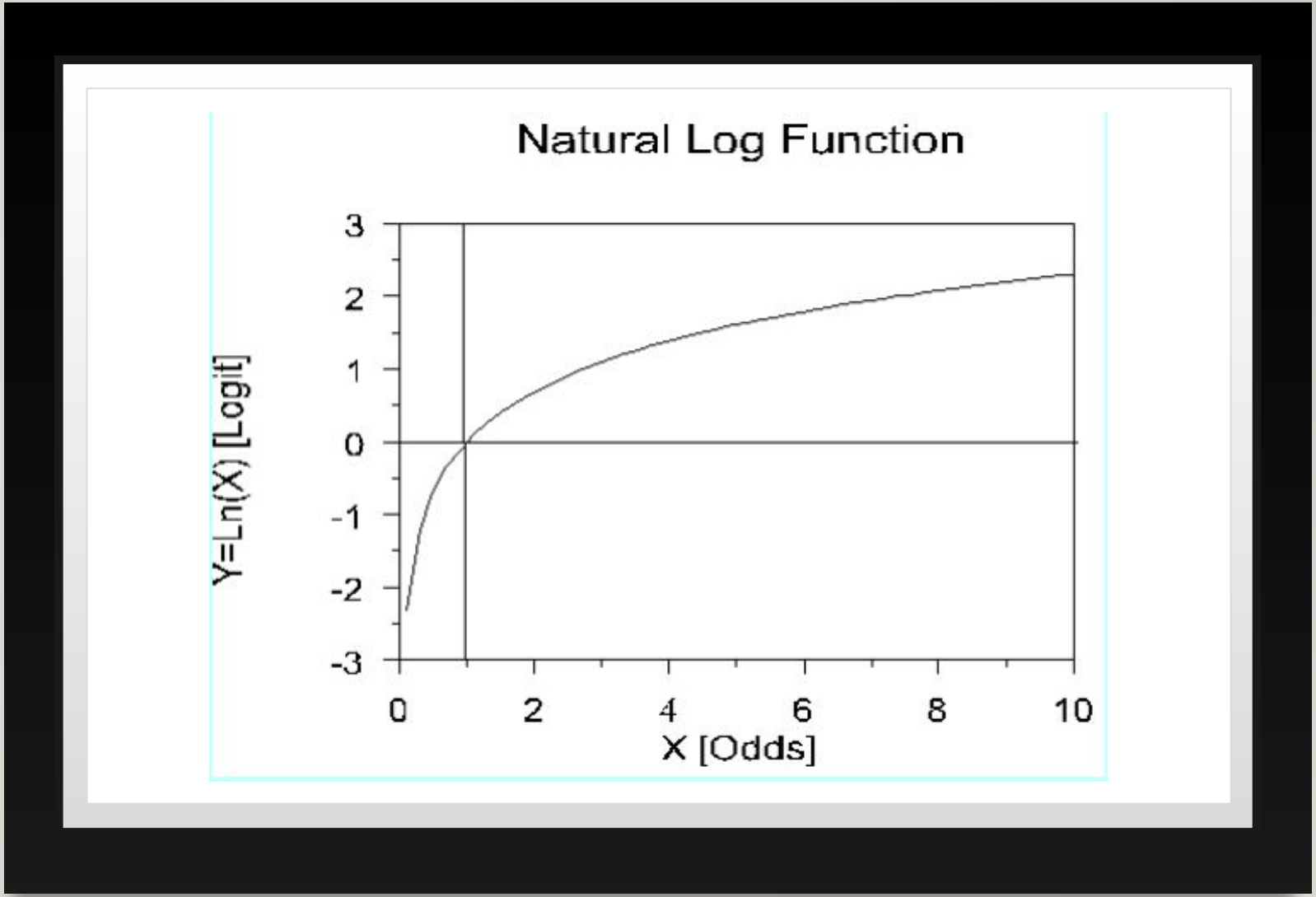
# LOGISTIC FUNCTION EXAMPLE



# REPRESENTATION USED FOR LOGISTIC REGRESSION FUNCTION

- Logistic regression uses an equation as the representation, very much like linear regression.
- Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.
- Example of Logistic Regression Equation: y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x)); where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x).
- Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

# NATURAL LOG FUNCTION



# LOGIT FUNCTION

$$\log(\text{odds}) = \text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

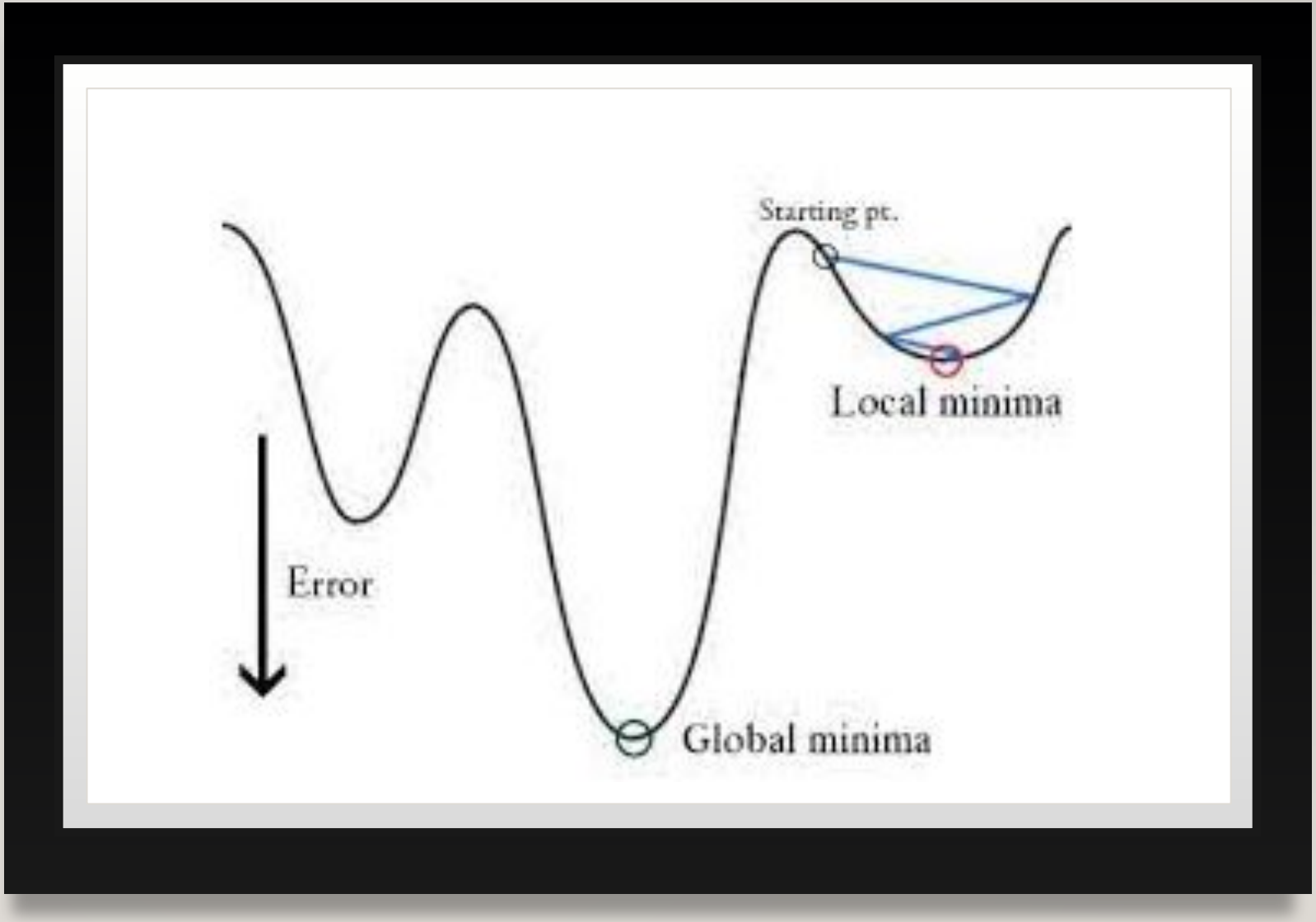
# EQUATIONS IN LOGISTIC REGRESSION

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\frac{P}{1-P} = e^{a+bX}$$

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

# COST FUNCTION GRAPH



# COST FUNCTION IN LOGISTIC REGRESSION: LOG LOSS

- Log Loss is the most important classification metric based on probabilities. It's hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models.
- For any given problem, a lower log loss value means better predictions.
- *Mathematical interpretation:* Log Loss is the negative average of the log of corrected predicted probabilities for each instance.

# COST FUNCTION FOR LOGISTIC REGRESSION

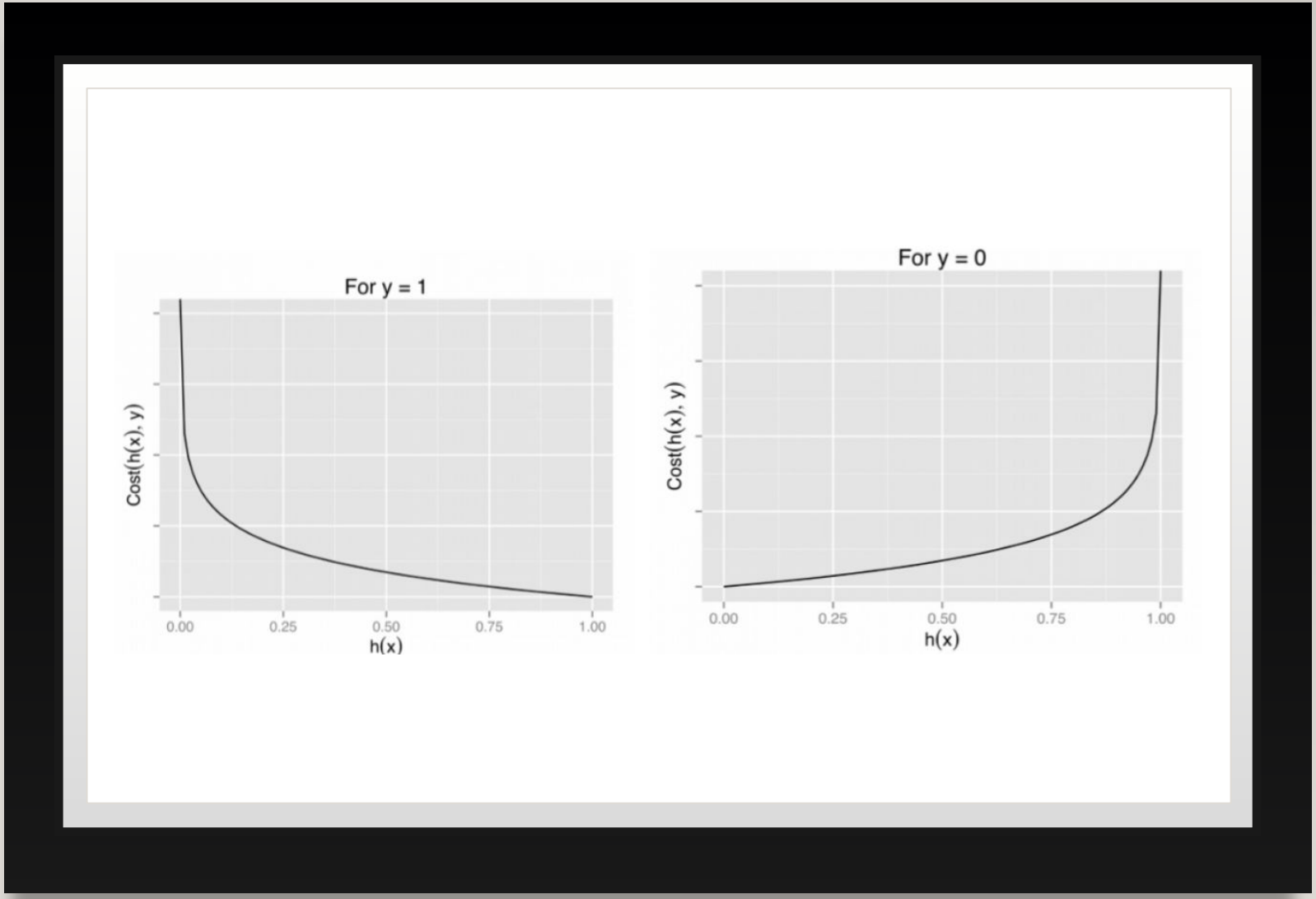
For logistic regression, the Cost function is defined as:

$-log(\mathrm{h}\theta(\mathrm{x}))$ *if* $y = 1$

$-log(1-\mathrm{h}\theta(\mathrm{x}))$ *if* $y = 0$

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

# GRAPHS

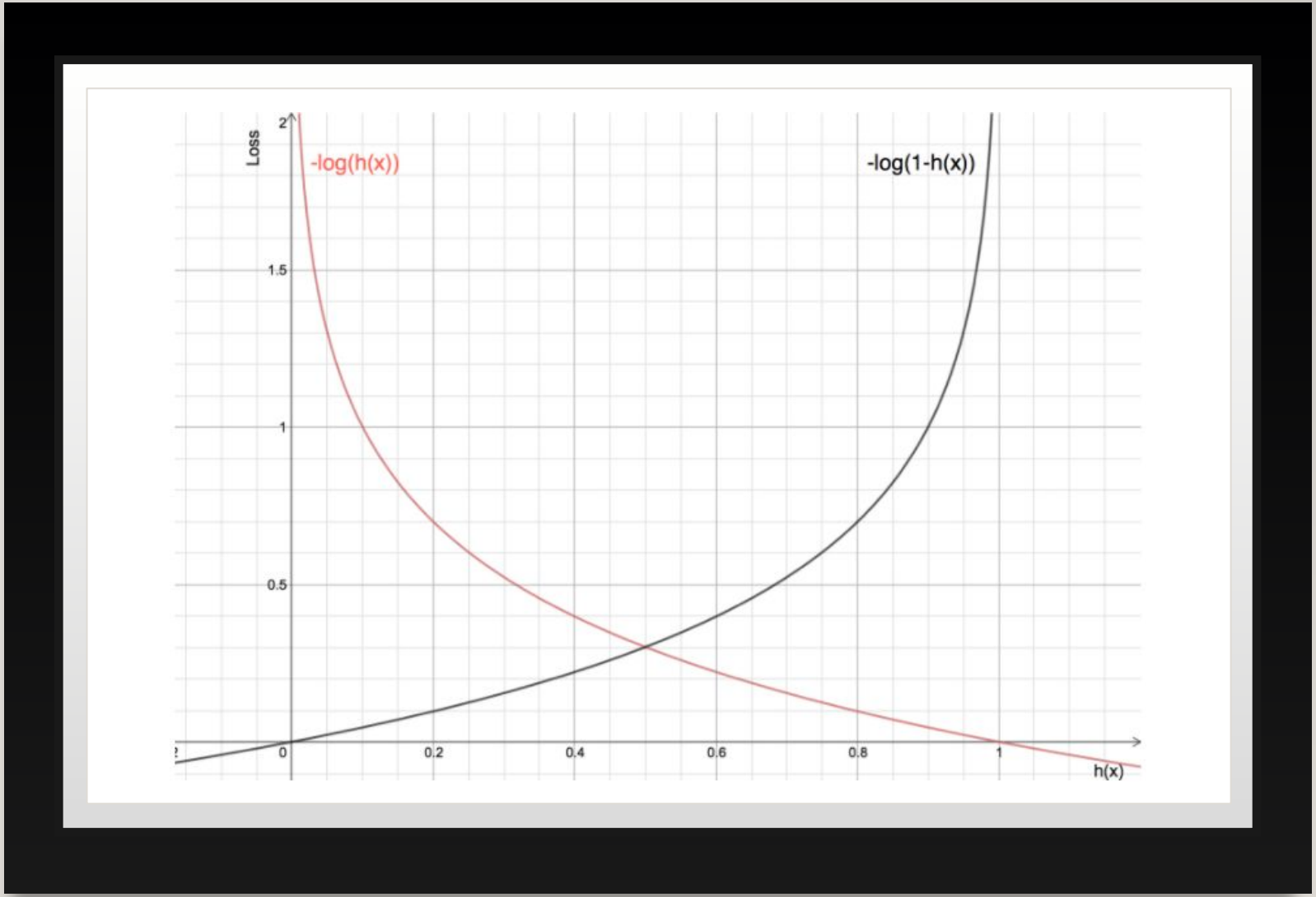

For y = 1
Cost(h(x), y)
0.00
0.25
0.50
0.75
1.00
h(x)
For y = 0
Cost(h(x), y)
0.00
0.25
0.50
0.75
1.00
h(x)

# BINARY CROSS ENTROPHY

The above two functions can be compressed into a single function i.e.

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1 - y^{(i)}\right)\log(1 - h\theta(x(i)))\right]$$

# LOG LOSS GRAPH



# USE OF GRADIENT DESCENT TO MINIMIZE COST FUNCTION

$$\theta j := \theta j - \alpha \frac{\partial}{\partial \theta j} J(\theta)$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all $\theta_j$)

$\}$

# PREPARING DATA FOR LOGISTIC REGRESSION

- Make output variable binary.
- Remove Noise
- Remove Outliers
- Gaussian Distribution: Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output.
- Remove correlated inputs
- Fail to Converge: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in your data or the data is very sparse.

# ASSUMPTIONS (1/2)

- First, binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
- Second, logistic regression requires the observations to be independent of each other.
- Third, logistic regression requires there to be little or no multicollinearity among the independent variables.
- Fourth, logistic regression assumes linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

# ASSUMPTIONS (2/2)

- Fifth, logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10*5 / .10).
- Lastly, Logistic regression expects that the data should not have extreme outliers.

# ADVANTAGES

- Logistic regression is easier to implement, interpret, and very efficient to train.
- It makes no assumptions about distributions of classes in feature space.
- It can easily extend to multiple classes(multinomial regression) and a natural probabilistic view of class predictions.
- It not only provides a measure of how appropriate a predictor(coefficient size) is, but also its direction of association (positive or negative).
- It is very fast at classifying unknown records.
- Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
- It can interpret model coefficients as indicators of feature importance.
- Logistic regression is less inclined to over-fitting.

# DISADVANTAGES

- If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
- It constructs linear boundaries. Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
- The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
- It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
- In Linear Regression independent and dependent variables are related linearly. But Logistic Regression needs that independent variables are linearly related to the log odds (log(p/(1-p)).

# APPLICATIONS

- Finance Industry: Credit card fraud, Credit scoring, Loan Approval.
- Medicine Industry: Whether the patient would have a disease or not.
- Hotel/Travel Booking
- Stock Market prediction
- Email (Spam/Not Spam)
- Handwriting Recognition

# THANK YOU