Machine Learning
ICT-4261

By-

Dr. Jesmin Akhter

Professor

Institute of Information Technology

Jahangirnagar University

Contents

The course will mainly cover the following topics:

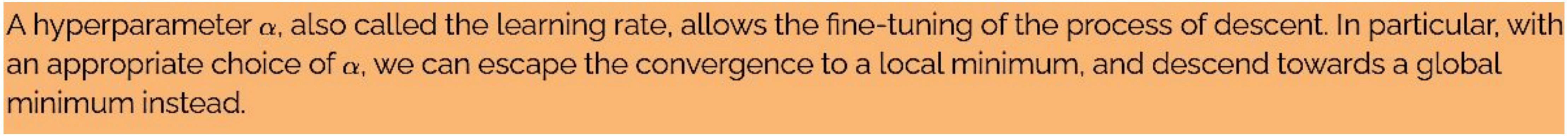
- ✔ A Gentle Introduction to Machine Learning
- ✔ Linear Regression
- ✔ Logistic Regression
- ✔ Naive Bayes
- ✔ Support Vector Machines
- ✔ Decision Trees and Ensemble Learning
- ✔ Clustering Fundamentals
- ✔ Hierarchical Clustering
- ✔ Neural Networks and Deep Learning
- ✔ Unsupervised Learning

Outline

- ✔ Logistic Regression
  - Gradient Descent
  - Linear regression vs logistic regression
  - Types of logistic regression
  - Key properties of the logistic regression equation
  - Stochastic gradient descent algorithms
  - Linear classification

Gradient Descent

- ✔ Gradient descent is an iterative process and in each step, it tries to move down the slope and get closer to the local minimum through which we optimize the parameters of a machine learning model.

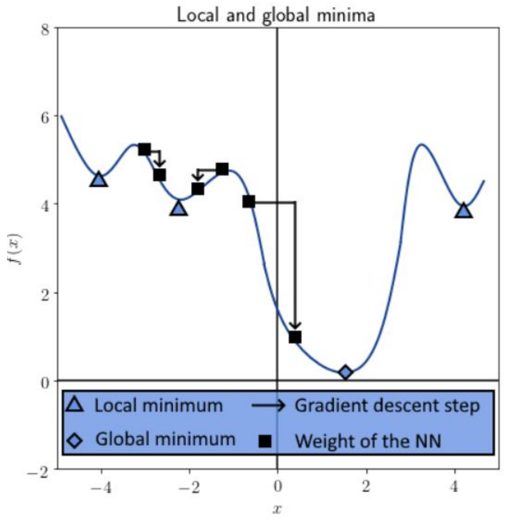


A hyperparameter α, also called the learning rate, allows the fine-tuning of the process of descent. In particular, with an appropriate choice of α, we can escape the convergence to a local minimum, and descend towards a global minimum instead.

- ✔ The gradient is calculated with respect to a vector of parameters for the model, typically the weights theta
- ✔ The sign of the gradient allows us to decide the direction of the closest minimum to the cost function. For a given parameter α, we iteratively optimize the vector by computing:

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

- ✔ And this is the graphical representation



Gradient Descent

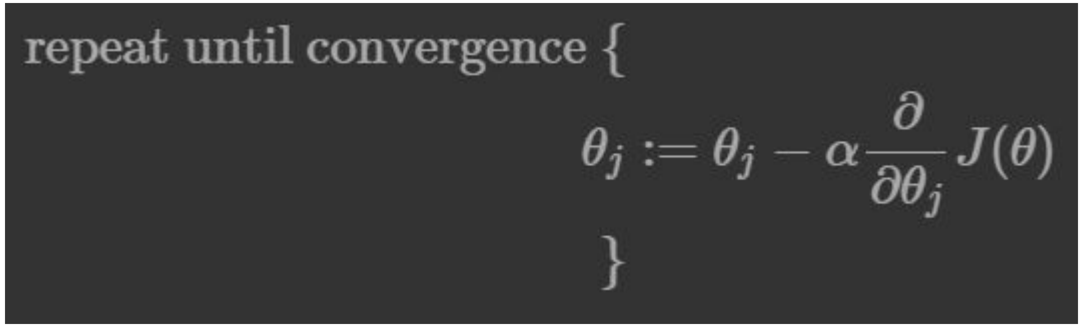
✔ At step *j* , the weights are all modified by the product of the hyperparameter alpha times the gradient of the cost function, computed with those weights. **If the gradient is positive, then we decrease the weights;** and conversely, if the gradient is negative, then we increase them.

✔ we can summarize the Gradient Descent Algorithm as:

✔

Start with random

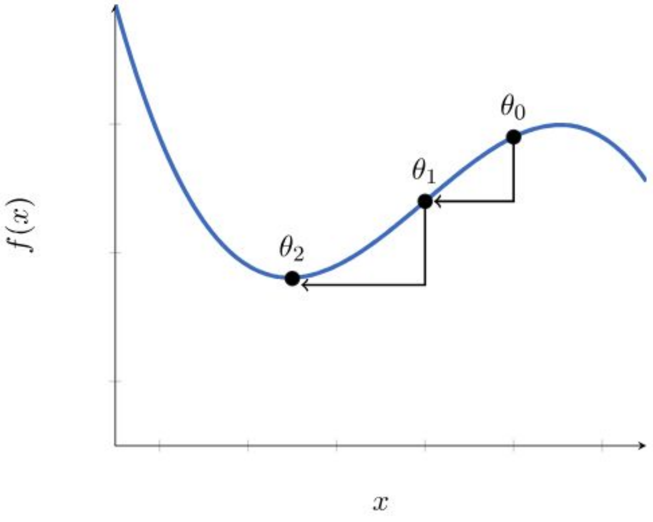
- **Loop until convergence:**
  - **Compute Gradient**
  - **Update**
- **Return**

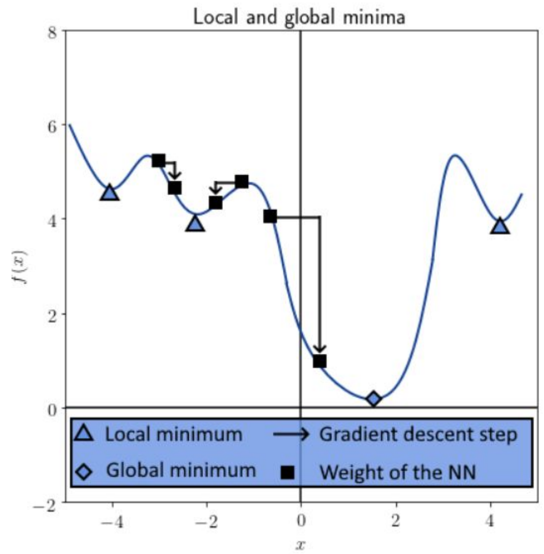


```
repeat until convergence {
    $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$
}
```

Minimizing the Cost with Gradient Descent

- **How gradient descent, can iteratively approximate the local minimum of a function with an arbitrary degree of precision?**
- We start with identifying a starting point that is sufficient proximity to the function’s local minimum, in this case, θ_0:





- Then, iteratively, we move towards the closest local minimum by exploiting the gradient of the function around θ_2.

Minimizing the Cost with Gradient Descent

- ✔ **Gradient descent is an iterative optimization algorithm, which finds the minimum of a differentiable function.** In this process, we try different values and update them to reach the optimal ones, minimizing the output. $\min_{\theta} J(\theta)$

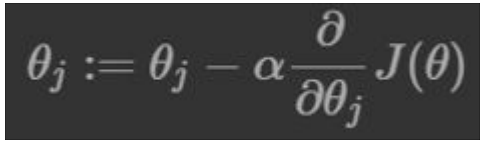
- ✔ we can apply this method to the cost function of logistic regression. This way, we can find an optimal solution minimizing the cost over model parameters: θ_j

- ✔ We’re using the sigmoid function as the hypothesis function in logistic regression.
- ✔ Assume we have a total of n features. In this case, we have n parameters for the theta vector. To minimize our cost function, we need to run the gradient descent on each parameter :

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

Minimizing the Cost with Gradient Descent

- ✔ Furthermore, we need to update each parameter simultaneously for each iteration **in the direction that decreases the cost function**. In other words, we need to loop through the parameters



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

In the case of logistic regression, analogously, we use a cost function that contains a logarithmic expression and we apply gradient descent algorithm on it.

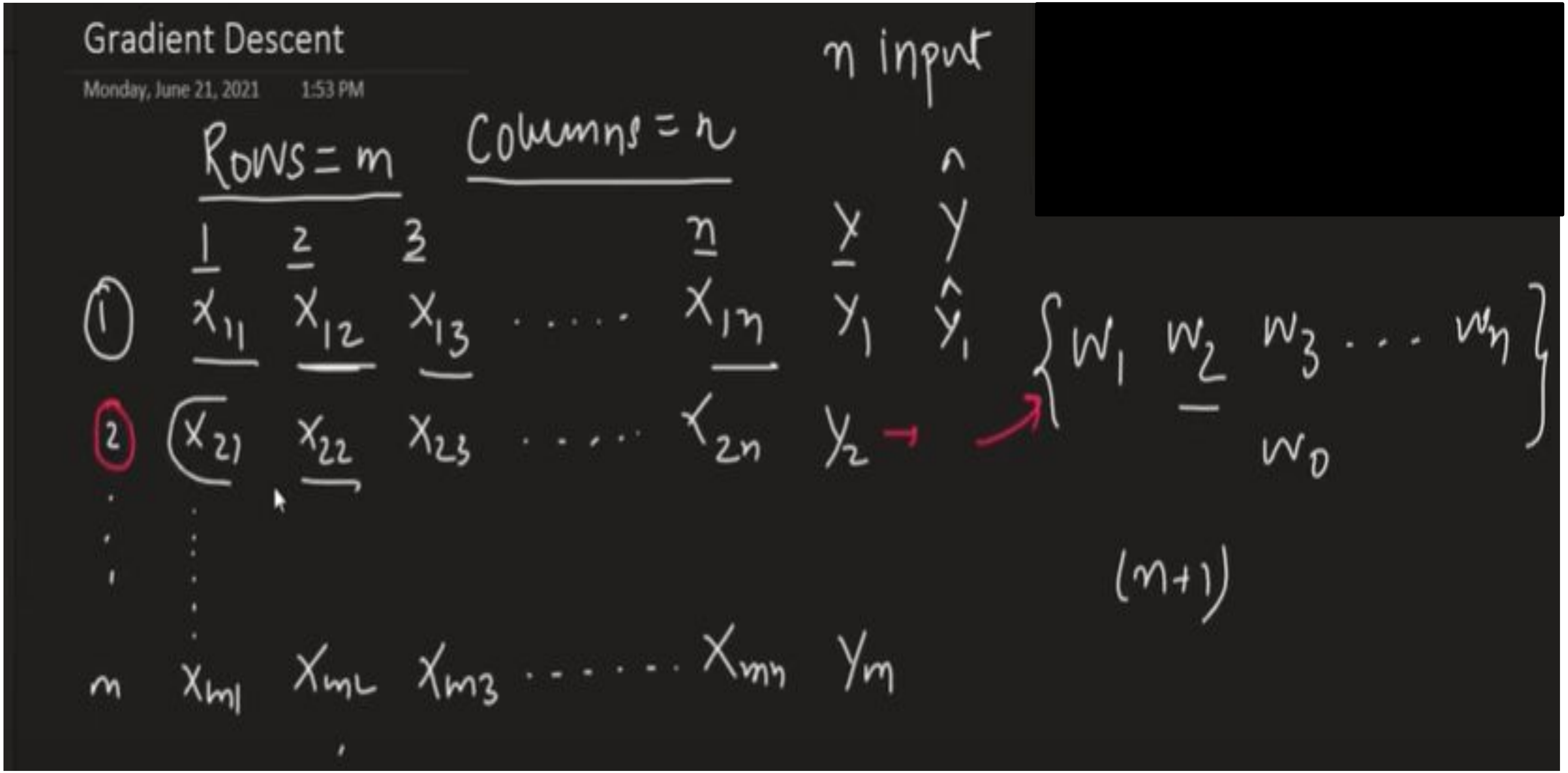
$\theta_0, \theta_1, \ldots, \theta_n$ in vector $\theta = [\theta_0, \theta_1, \ldots, \theta_n]$.

To complete the algorithm, we need the value of $\frac{\partial}{\partial \theta_j} J(\theta)$, which is:

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

- ✔ Plugging this into the gradient descent function leads to the update rule.
- ✔ Surprisingly, the update rule is the same as the one derived by using the sum of the squared errors in linear regression. As a result, we can use the same gradient descent formula for logistic regression as well.
- ✔ By iterating over the training samples until convergence, we reach the optimal parameters theta leading to minimum cost.

Derive the loss function with matrix form



Gradient Descent

Monday, June 21, 2021 1:53 PM

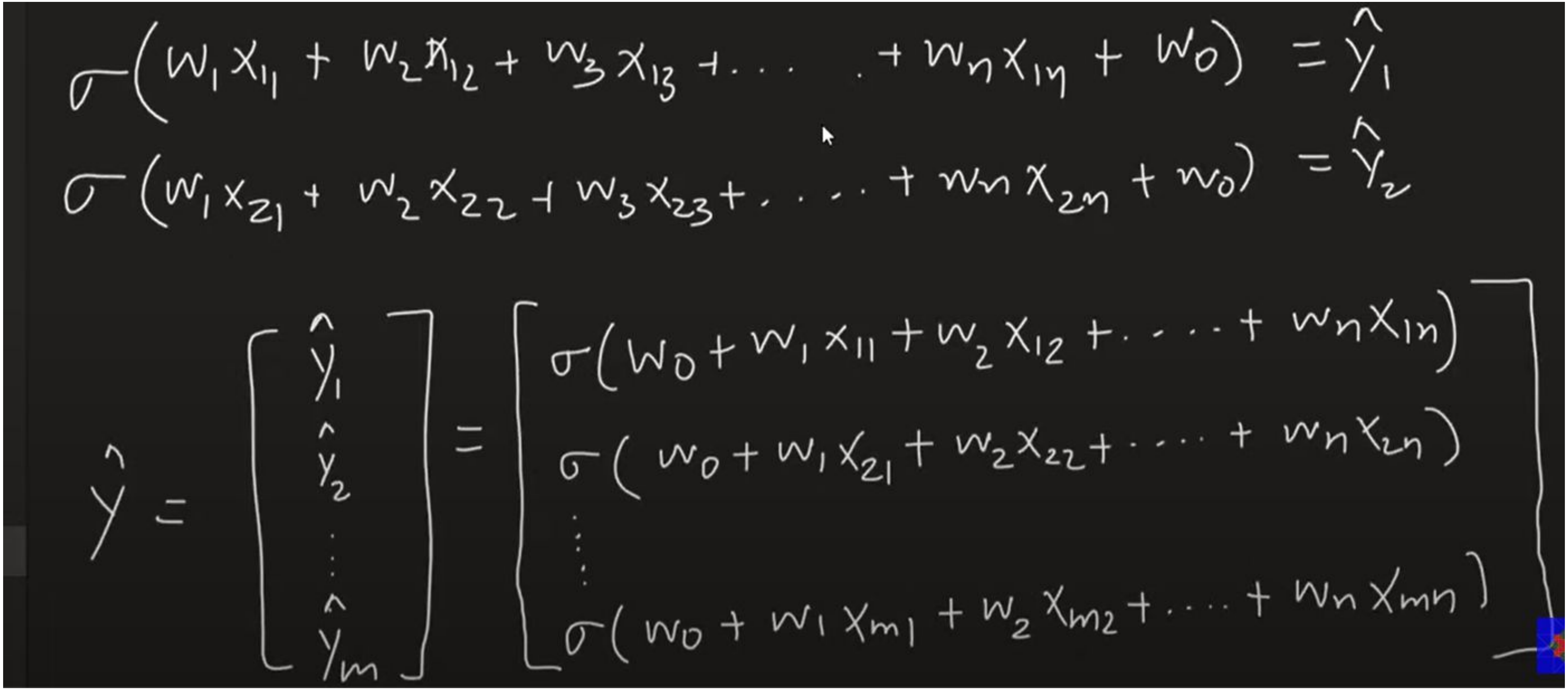
[redacted]

Rows = m   Columns = n   n input

$$\begin{array}{c|ccccccc}
 & \underline{1} & \underline{2} & \underline{3} & & \underline{n} & \underline{y} & \hat{y} \\
① & \underline{x_{11}} & \underline{x_{12}} & \underline{x_{13}} & \cdots\cdots & \underline{x_{1n}} & y_1 & \hat{y}_1 \\
② & (x_{21} & \underline{x_{22}} & x_{23} & \cdots\cdots & x_{2n} & y_2 \rightarrow & \\
\vdots & \vdots & & & & & & \\
m & x_{m1} & x_{m2} & x_{m3} & \cdots\cdots & x_{mn} & y_m &
\end{array}$$

$$\nearrow \left\{ w_1 \ \underline{w_2} \ w_3 \cdots w_n \atop w_0 \right\}$$

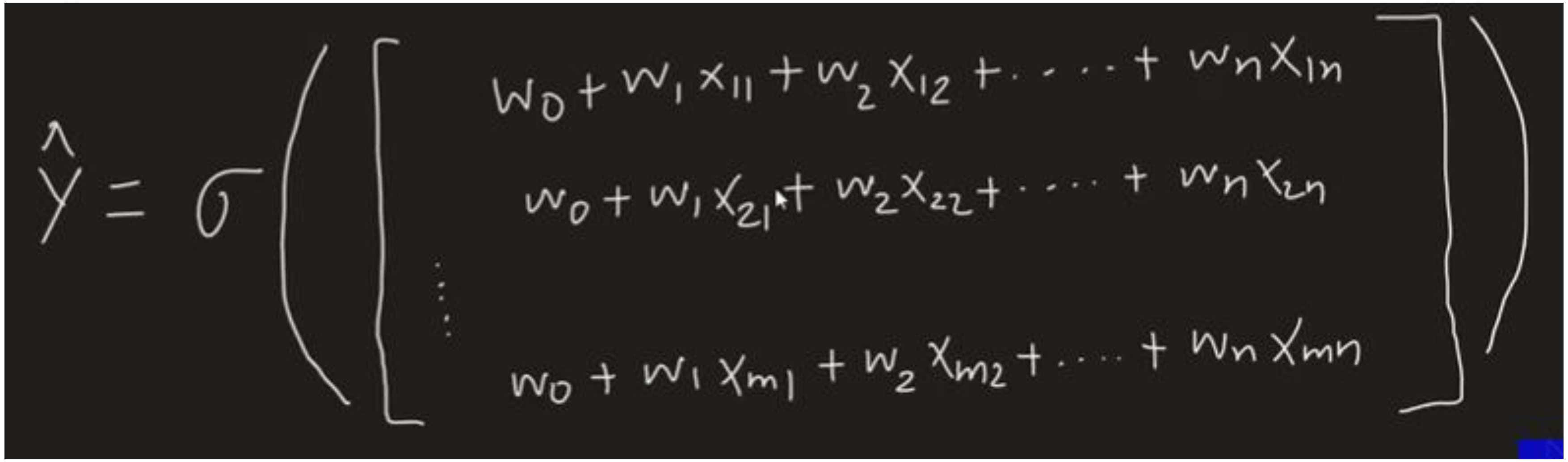
$(n+1)$



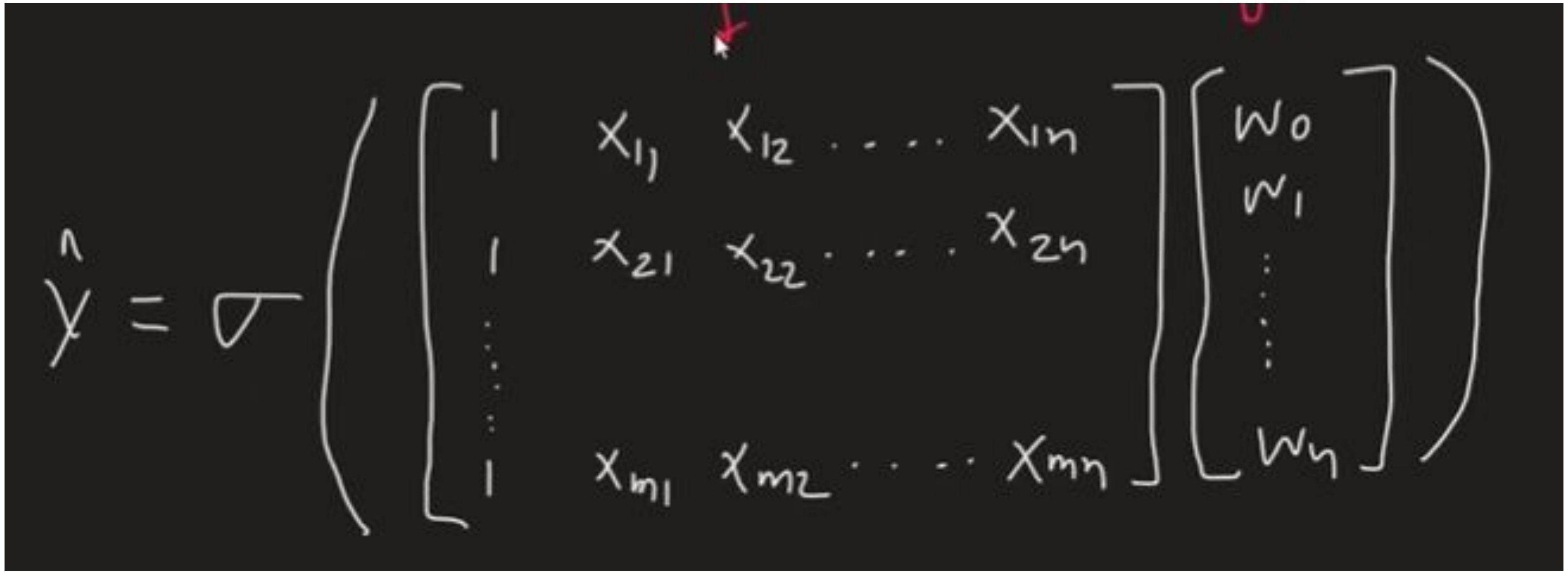
$$\sigma(w_1 x_{11} + w_2 x_{12} + w_3 x_{13} + \dots + w_n x_{1n} + w_0) = \hat{y}_1$$

$$\sigma(w_1 x_{21} + w_2 x_{22} + w_3 x_{23} + \dots + w_n x_{2n} + w_0) = \hat{y}_2$$

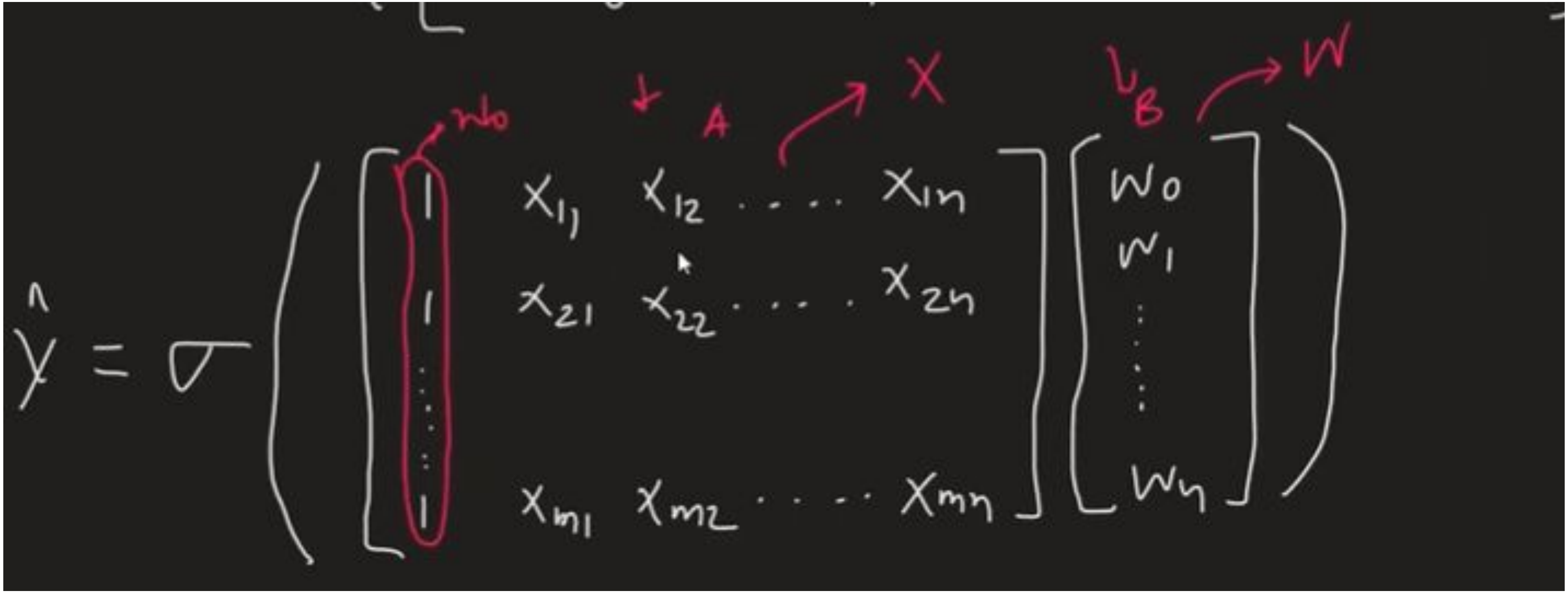
$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \sigma(w_0 + w_1 x_{11} + w_2 x_{12} + \dots + w_n x_{1n}) \\ \sigma(w_0 + w_1 x_{21} + w_2 x_{22} + \dots + w_n x_{2n}) \\ \vdots \\ \sigma(w_0 + w_1 x_{m1} + w_2 x_{m2} + \dots + w_n x_{mn}) \end{bmatrix}$$



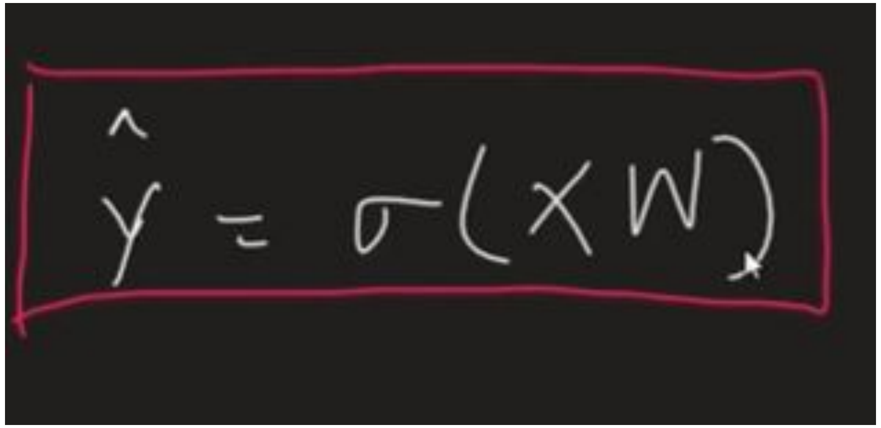
$$\hat{y} = \sigma\left(\begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} + \cdots + w_n x_{1n} \\ w_0 + w_1 x_{21} + w_2 x_{22} + \cdots + w_n x_{2n} \\ \vdots \\ w_0 + w_1 x_{m1} + w_2 x_{m2} + \cdots + w_n x_{mn} \end{bmatrix}\right)$$



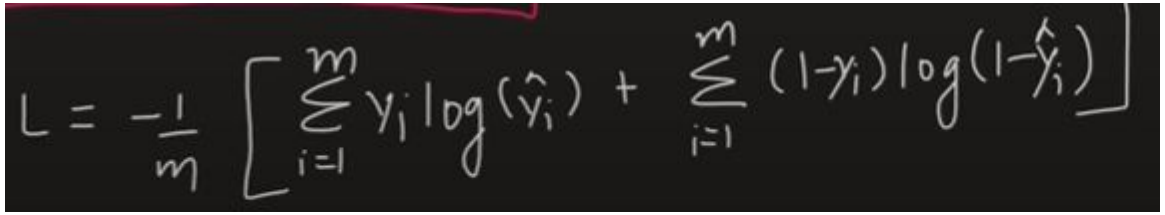
$$\hat{y} = \sigma\left(\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}\right)$$



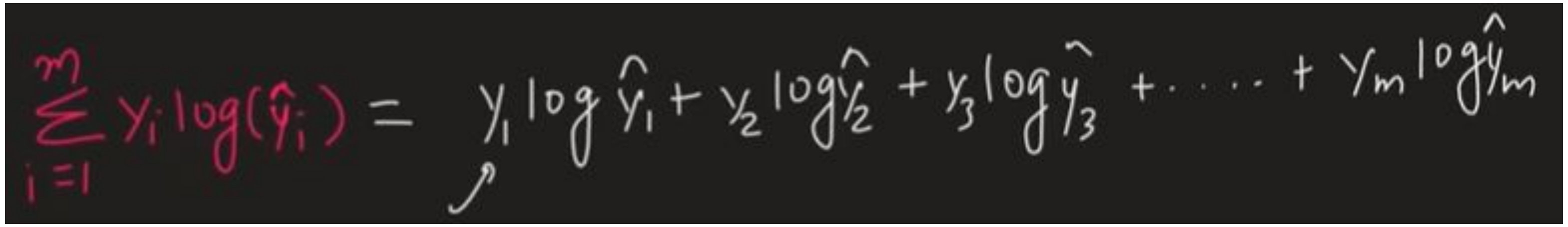
$$\hat{y} = \sigma\left(\overset{\downarrow\ A \nearrow X}{\begin{bmatrix} \overset{x_0}{1} & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}} \overset{\downarrow B \nearrow W}{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}} \right)$$



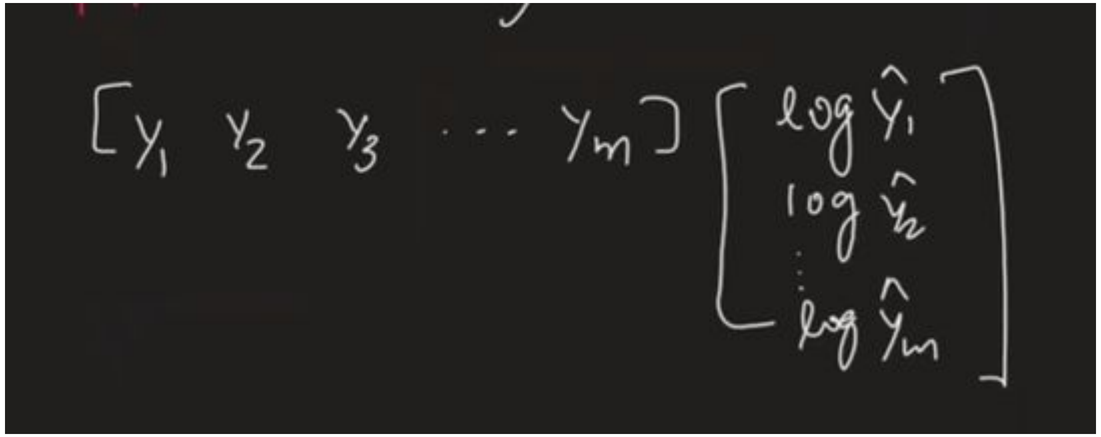
$$\boxed{\hat{y} = \sigma(XW)}$$



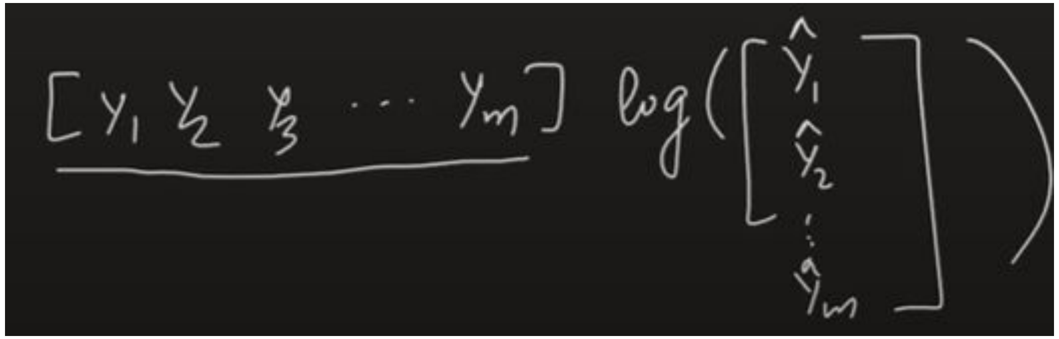
$$L = -\frac{1}{m}\left[\sum_{i=1}^{m} y_i \log(\hat{y}_i) + \sum_{i=1}^{m} (1-y_i)\log(1-\hat{y}_i)\right]$$



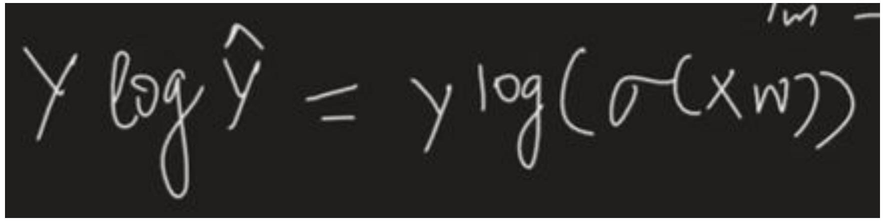
$$\sum_{i=1}^{m} y_i \log(\hat{y}_i) = y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + y_3 \log \hat{y}_3 + \dots + y_m \log \hat{y}_m$$



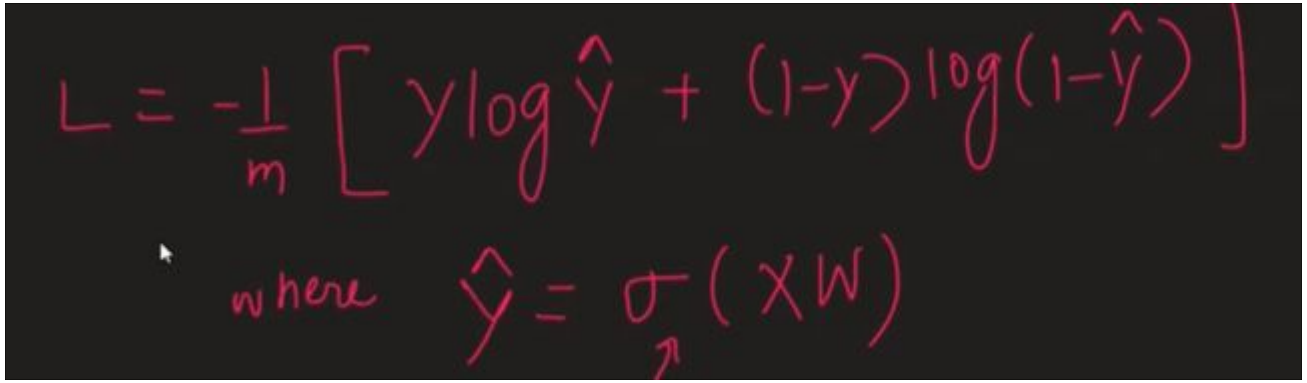
$$\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix} \begin{bmatrix} \log \hat{y}_1 \\ \log \hat{y}_2 \\ \vdots \\ \log \hat{y}_m \end{bmatrix}$$



$$\underline{\begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix}} \log\left(\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} \right)$$

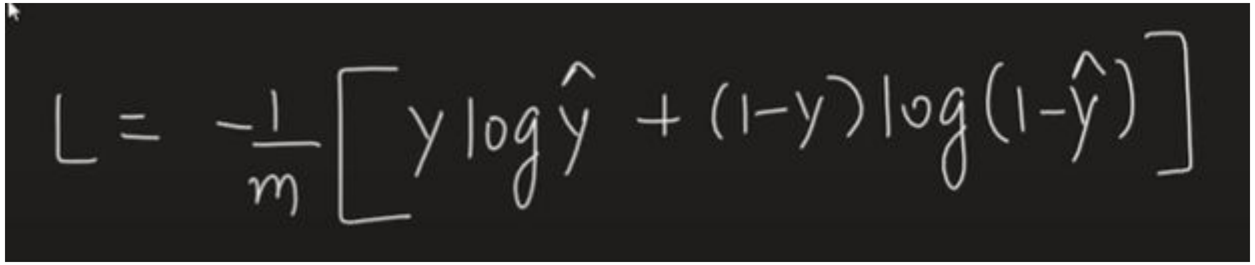


$$Y \log \hat{Y} = y \log(\sigma(XW))$$

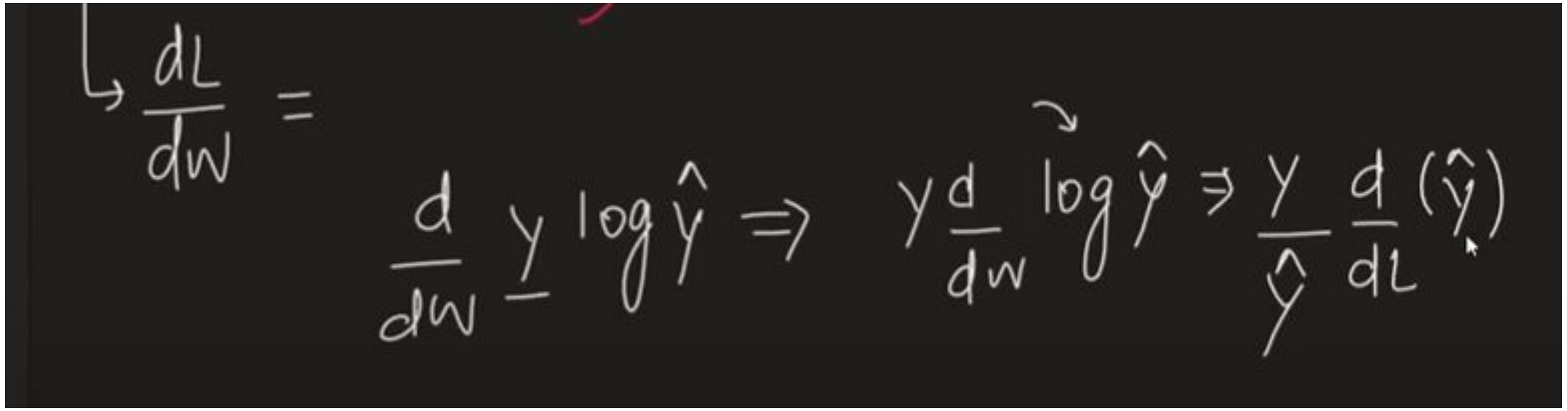


$$L = -\frac{1}{m}\left[y \log \hat{y} + (1-y)\log(1-\hat{y}) \right]$$

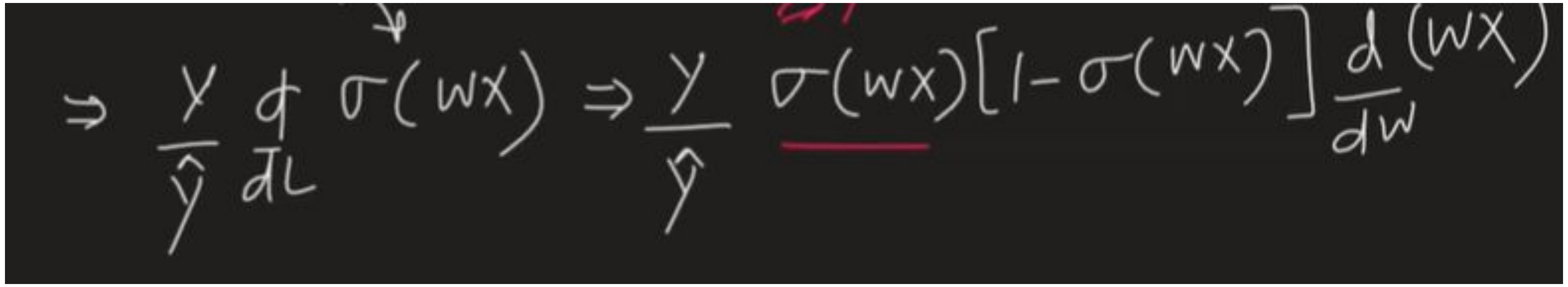
where $\hat{y} = \sigma(XW)$



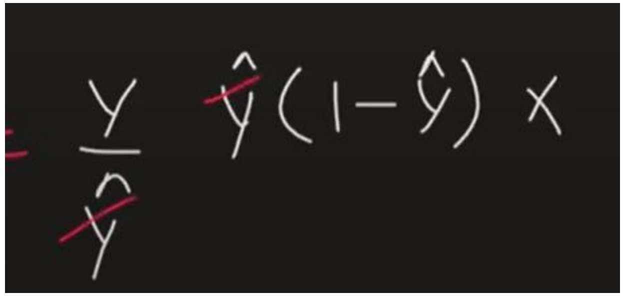
$$L = -\frac{1}{m}\left[y\log\hat{y} + (1-y)\log(1-\hat{y})\right]$$



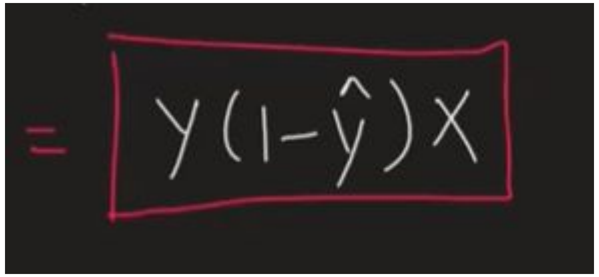
$$\hookrightarrow \frac{dL}{dw} = \frac{d}{dw}\underline{y}\log\hat{y} \Rightarrow y\frac{d}{dw}\log\hat{y} \Rightarrow \frac{y}{\hat{y}}\frac{d}{dL}(\hat{y})$$



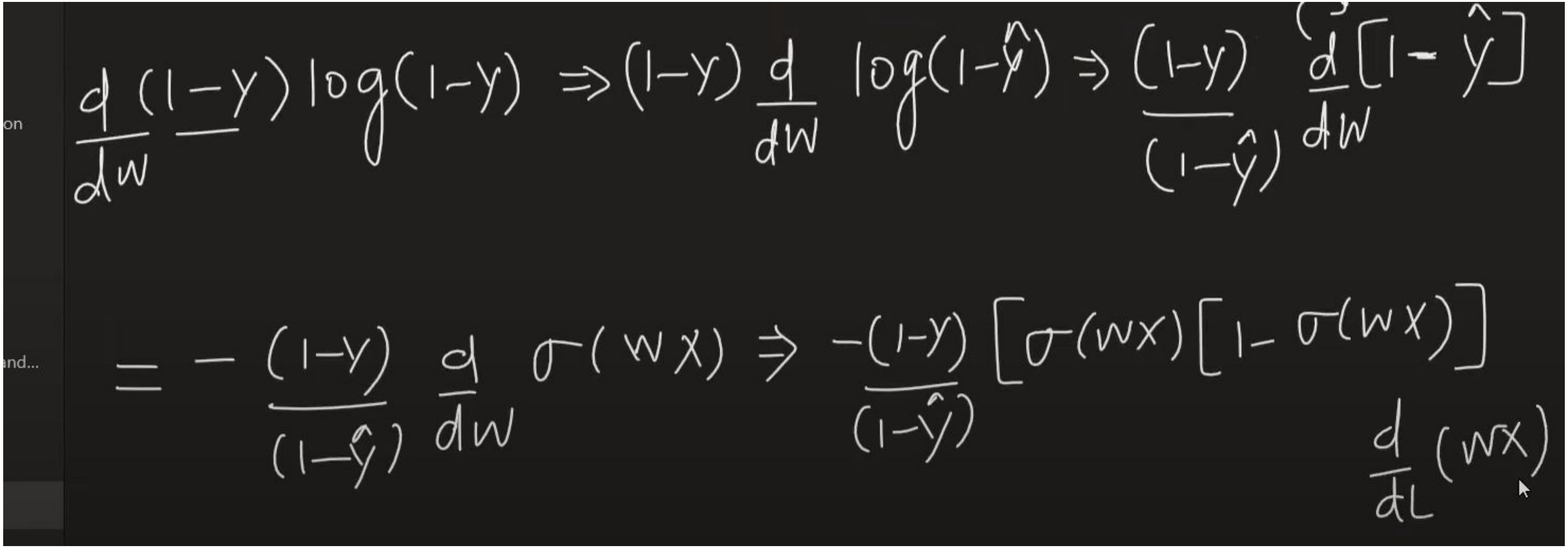
$$\Rightarrow \frac{y}{\hat{y}}\frac{d}{dL}\sigma(wx) \Rightarrow \frac{y}{\hat{y}}\underline{\sigma(wx)}[1-\sigma(wx)]\frac{d}{dw}(wx)$$



$$= \frac{y}{\cancel{\hat{y}}}\cancel{\hat{y}}(1-\hat{y})\,x$$



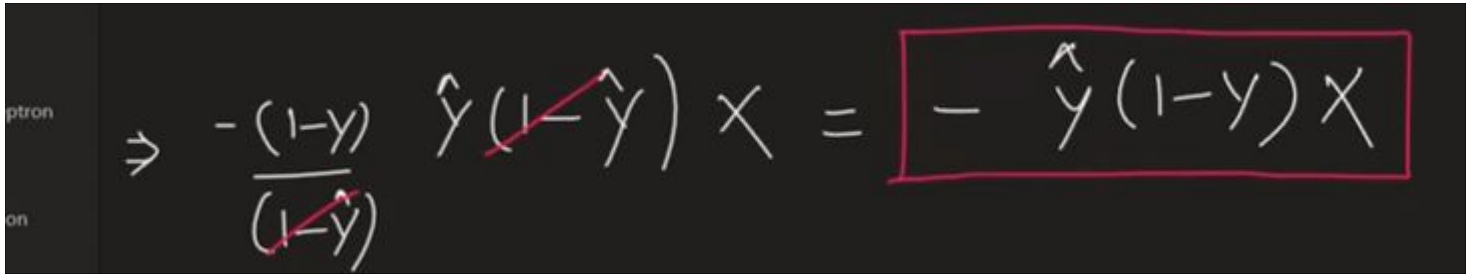
$$= \boxed{y(1-\hat{y})X}$$



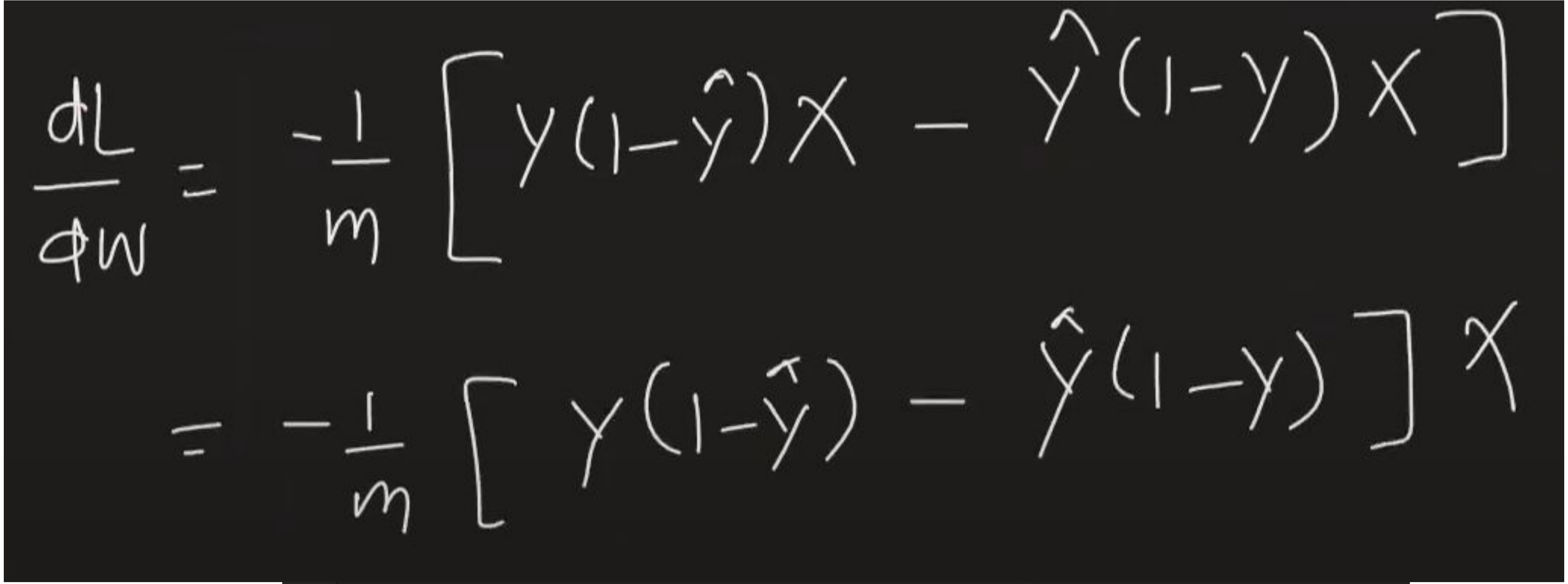
$$\frac{d}{dW}\underline{(1-y)}\log(1-y) \Rightarrow (1-y)\frac{d}{dW}\log(1-\hat{y}) \Rightarrow \frac{(1-y)}{(1-\hat{y})}\frac{d}{dW}[1-\hat{y}]$$

$$= -\frac{(1-y)}{(1-\hat{y})}\frac{d}{dW}\sigma(WX) \Rightarrow -\frac{(1-y)}{(1-\hat{y})}\left[\sigma(wx)[1-\sigma(wx)]\right]$$

$$\frac{d}{dL}(wx)$$

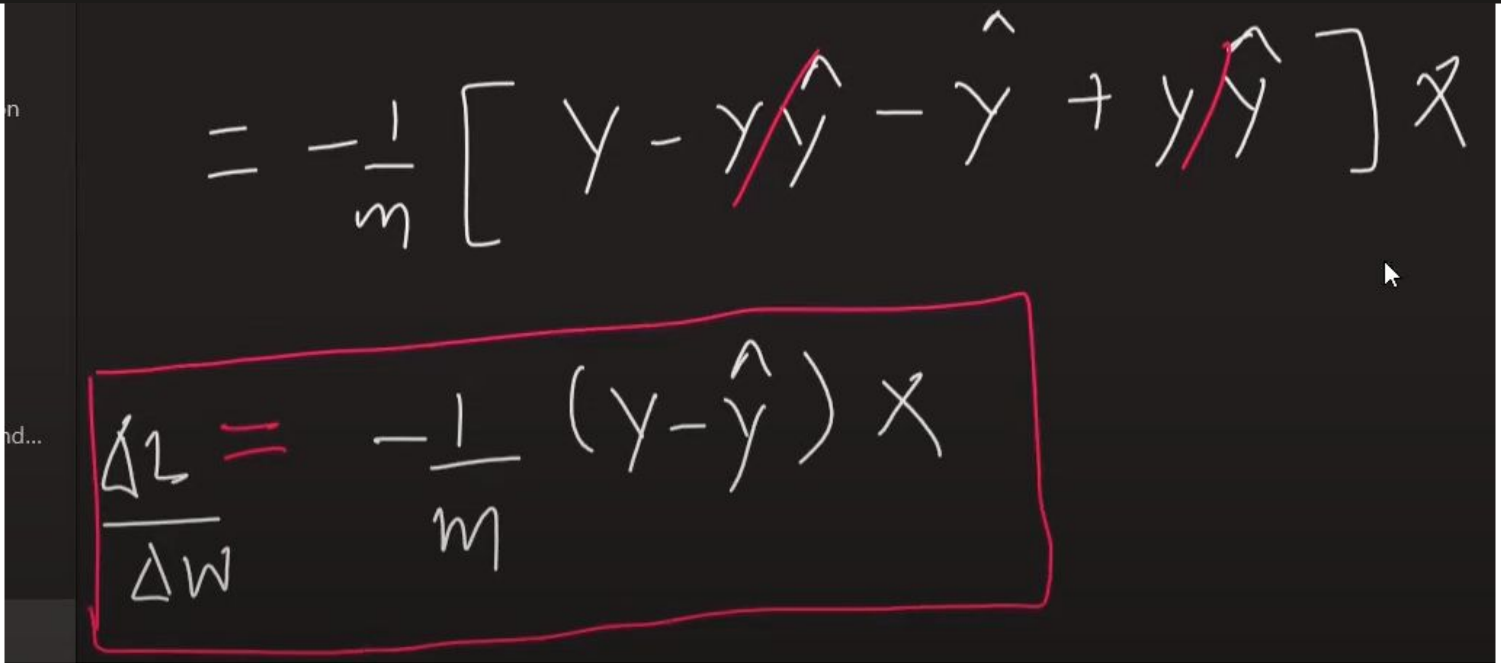


$$\Rightarrow -\frac{(1-y)}{\cancel{(1-\hat{y})}}\hat{y}\cancel{(1-\hat{y})}X = \boxed{-\hat{y}(1-y)X}$$



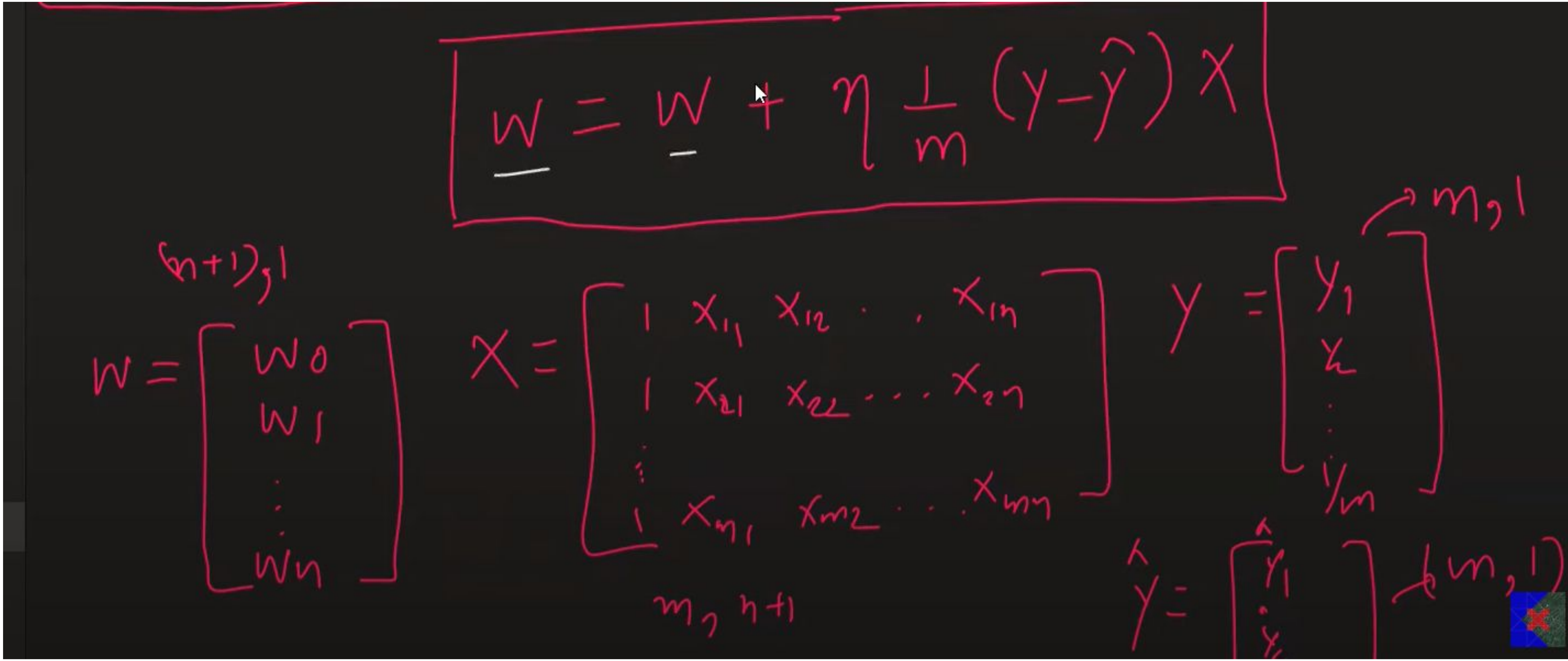
$$\frac{dL}{dw} = -\frac{1}{m}\left[y(1-\hat{y})X - \hat{y}(1-y)X \right]$$

$$= -\frac{1}{m}\left[y(1-\hat{y}) - \hat{y}(1-y) \right] X$$



$$= -\frac{1}{m}\left[y - \cancel{y\hat{y}} - \hat{y} + \cancel{y\hat{y}} \right] X$$

$$\boxed{\frac{\Delta L}{\Delta w} = -\frac{1}{m}(y-\hat{y})X}$$



$$\boxed{\underline{W} = \underline{W} + \eta \frac{1}{m}(y - \hat{y}) X}$$

$(n+1), 1$

$$W = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1n} \\ 1 & X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & & & & \\ 1 & X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}$$

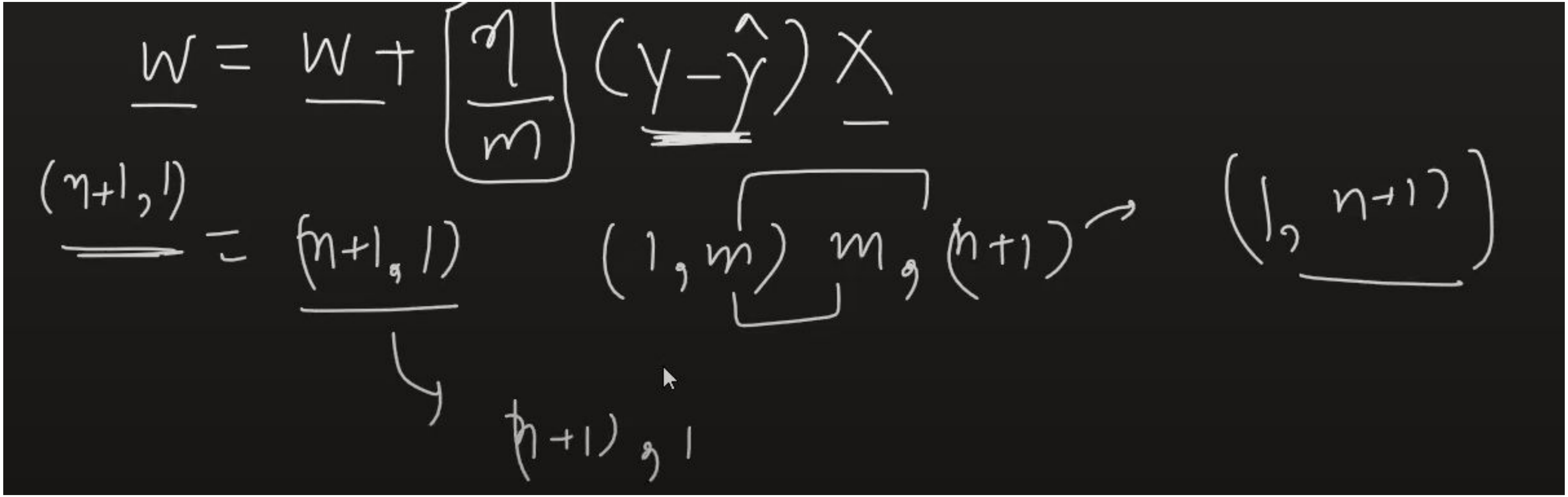
$m, n+1$

$m, 1$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ y \end{bmatrix}$$

$(m, 1)$



$$\underline{W} = \underline{W} + \boxed{\frac{\eta}{m}} (\underline{y - \hat{y}}) \underline{X}$$

$$\underline{(n+1, 1)} = \underline{(n+1, 1)} \qquad (1, m) \quad m, (n+1) \rightarrow \underline{(1, n+1)}$$

$$(n+1), 1$$

Linear regression vs logistic regression

Sr. No	Linear Regresssion	Logistic Regression
1	Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic regression is used to predict the categorical dependent variable using a given set of independent variables.
2	Linear regression is used for solving Regression problem.	It is used for solving classification problems.
3	The relationship between the dependent variable and independent variable must be linear.	The relationship DOES NOT need to be linear between the dependent and independent variables.
4	We are finding and using the line of best fit to help us easily predict outputs.	We are using the S-curve (Sigmoid) to help us classify predicted outputs.
5	Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for Estimation of accuracy.
6	The output must be continuous value,such as price,age,etc.	Output is must be categorical value such as 0 or 1, Yes or no, etc.
7	There is a possibility of collinearity between the independent variables.	There should not be any collinearity between the independent variable.

Some Examples

Some examples of such classifications and instances where the binary response is expected or implied are:

- ✔ **Fraud detection:** Logistic regression models can help teams identify data anomalies, which are predictive of fraud. Certain behaviors or characteristics may have a higher association with fraudulent activities, which is particularly helpful to banking and other financial institutions in protecting their clients..
- ✔ **Disease prediction:** In medicine, this analytics approach can be used to predict the likelihood of disease or illness for a given population. Healthcare organizations can set up preventative care for individuals that show higher tendency for specific illnesses.
  - **Determine the probability of heart attacks**: With the help of a logistic model, medical practitioners can determine the relationship between variables such as the weight, exercise, etc., of an individual and use it to predict whether the person will suffer from a heart attack or any other medical complication.

Some Examples

- ✔ **Possibility of enrolling into a university**: Application aggregators can determine the probability of a student getting accepted to a particular university or a degree course in a college by studying the relationship between the estimator variables, such as GRE, GMAT, or TOEFL scores.
- ✔ **Identifying spam emails**: Email inboxes are filtered to determine if the email communication is promotional/spam by understanding the predictor variables and applying a logistic regression algorithm to check its authenticity.

Types of logistic regression

There are three types of logistic regression models, which are defined based on categorical response.

- ✔ Binary logistic regression
- ✔ Multinomial logistic regression
- ✔ Ordinal logistic regression

- ✔ **Binary logistic regression**
  - In this approach, the response or **dependent variable is dichotomous in nature—**i.e. it has only two possible outcomes (e.g. success/failure, 0/1, or true/false). Within logistic regression, it is one of the most common classifiers for binary classification.
  - Some popular examples of its use include
    - predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant.
    - Deciding on whether or not to offer a loan to a bank customer: Outcome = yes or no.
    - Evaluating the risk of cancer: Outcome = high or low.
    - Predicting a team's win in a football match: Outcome = yes or no.

Types of logistic regression

- ✔ **Multinomial logistic regression**
  - – A categorical dependent variable has three or more discrete outcomes in a multinomial regression type. This implies that this regression type has more than two possible outcomes.
  - – ; however, these values have no specified order such as "cat", "dogs", or "sheep“

Examples:

- – Let’s say you want to predict the most popular transportation type for 2040. Here, transport type equates to the dependent variable, and the possible outcomes can be electric cars, electric trains, electric buses, and electric bikes.
- – Predicting whether a student will join a college, vocational/trade school, or corporate industry.
- – Estimating the type of food consumed by pets, the outcome may be wet food, dry food, or junk food.

Types of logistic regression

- ✔ **Ordinal logistic regression**
  - – This type of logistic regression model is applied when the response variable has three or more possible outcome, but in this case, these values do have a defined order.
    - **Examples**: Ordered types of dependent variables represent,
    - "low", "Medium", or "High“, grading scales from A to F or rating scales from 1 to 5.
    - Formal shirt size: Outcomes = XS/S/M/L/XL
    - Survey answers: Outcomes = Agree/Disagree/Unsure
    - Scores on a math test: Outcomes = Poor/Average/Good

Key Advantages of Logistic Regression

1. **Easier to implement machine learning methods**: A machine learning model can be effectively set up with the help of training and testing. The training identifies patterns in the input data (image) and associates them with some form of output (label). Training a logistic model with a regression algorithm does not demand higher computational power. As such, logistic regression is easier to implement, interpret, and train than other ML methods.

2. **Suitable for linearly separable datasets**: A linearly separable dataset refers to a graph where a straight line separates the two data classes. In logistic regression, the y variable takes only two values. Hence, one can effectively classify data into two separate classes if linearly separable data is used.

3. **Provides valuable insights**: Logistic regression measures how relevant or appropriate an independent/predictor variable is (coefficient size) and also reveals the direction of their relationship or association (positive or negative).

Thank You

Derivation of Cost Function:

$$J = -\sum_{i=1}^{m} y_i \log\left(h_\theta\left(x_i\right)\right) + \left(1 - y_i\right) \log\left(1 - h_\theta\left(x_i\right)\right)$$

- ✔ Now, we will derive the cost function with the help of the chain rule as it allows us to calculate complex partial derivatives by breaking them down.
- ✔ **Step-1: Use chain rule and break the partial derivative of log-likelihood.**

$$-\frac{\partial LL(\theta)}{\partial \theta_j} = -\frac{\partial LL(\theta)}{\partial p} \cdot \frac{\partial p}{\partial \theta} \qquad \textit{where } p = \sigma\left[\theta^T x\right]$$

$$= -\frac{\partial LL(\theta)}{\partial p} \cdot \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial \theta_j} \qquad \textit{where } z = \theta^T x$$

- ✔ **Step-2: Find derivative of log-likelihood w.r.t p**

We know,

$$LL(\theta) = y \log(p) + (1 - y)\log(1 - p) \qquad \text{where } p = \sigma\left[\theta^T x\right]$$

$$\frac{\partial LL(\theta)}{\partial p} = \frac{y}{p} + \frac{(1 - y)}{(1 - p)}$$

✔ **Step-3: Find derivative of *'p'* w.r.t *'z'***

$$p = \sigma(z)$$

$$\frac{\partial p}{\partial z} = \frac{\partial[\sigma(z)]}{\partial z}$$

We know the derivative of sigmoid function is $\sigma\left[\theta^{T}x\right]\left[1-\sigma\left(\theta^{T}x\right)\right]$

$$\Rightarrow \frac{\partial p}{\partial z} = \sigma[z][1-\sigma(z)]$$

✔ ***Step-4: Find derivate of z w.r.t* θ**

$$z = \theta^{T}x$$

$$\frac{\partial z}{\partial \theta_{j}} = x_{j}$$

✔ **Step-5: Put all the derivatives in equation 1**

$$-\frac{\partial LL(\theta)}{\partial \theta_{j}} = -\frac{\partial LL(\theta)}{\partial p} \cdot \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial \theta_{j}}$$

$$-\frac{\partial LL(\theta)}{\partial \theta_{j}} = -\left[\frac{y}{p} + \frac{(1-y)}{(1-p)}\right].\sigma(z)\sigma(1-(z)).x_{j}$$

$$= -\left[\frac{y}{p} + \frac{(1-y)}{(1-p)}\right].p[1-p].x_{j} \qquad \text{since } p = \sigma[z]$$

$$= -[y(1-p) - p(1-y)].x_{j}$$

$$= -[y-p].x_{j}$$

$$\Rightarrow [p-y].x_{j} = [\sigma(\theta^{T}x) - y].x_{j}$$

- Hence the derivative of our cost function is:

$$\theta_{new} = \theta_{old} - \alpha \left[\sigma\left(\theta^T x \right) - y \right] . x_j$$

- Now since we have our derivative of the cost function, we can write our gradient descent algorithm as:
  - If the slope is negative (downward slope) then our gradient descent will add some value to our new value of the parameter directing it towards the minimum point of the convex curve. Whereas if the slope is positive (upward slope) the gradient descent will minus some value to direct it towards the minimum point.

$$\theta x^i := \theta_0 + \theta_1 x_1^i + \cdots + \theta_p x_p^i.$$

Then

$$\log h_\theta(x^i) = \log \frac{1}{1+e^{-\theta x^i}} = -\log(1+e^{-\theta x^i}),$$

$$\log(1-h_\theta(x^i)) = \log(1-\frac{1}{1+e^{-\theta x^i}}) = \log(e^{-\theta x^i}) - \log(1+e^{-\theta x^i}) = -\theta x^i - \log (1+e^{-\theta x^i}),$$

[this used: $1 = \frac{(1+e^{-\theta x^i})}{(1+e^{-\theta x^i})}$, the 1's in numerator cancel, then we used: $\log(x/y) = \log(x) - \log(y)$]

Since our original cost function is the form of:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i))$$

Plugging in the two simplified expressions above, we obtain

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

, which can be simplified to:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i\theta x^i - \theta x^i - \log(1+e^{-\theta x^i})\right] = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i\theta x^i - \log(1+e^{\theta x^i})\right], \quad (*)$$

where the second equality follows from

$$-\theta x^i - \log(1+e^{-\theta x^i}) = -\left[\log e^{\theta x^i} + \log(1+e^{-\theta x^i})\right] = -\log(1+e^{\theta x^i}).$$

[we used $\log(x) + \log(y) = log(xy)$]

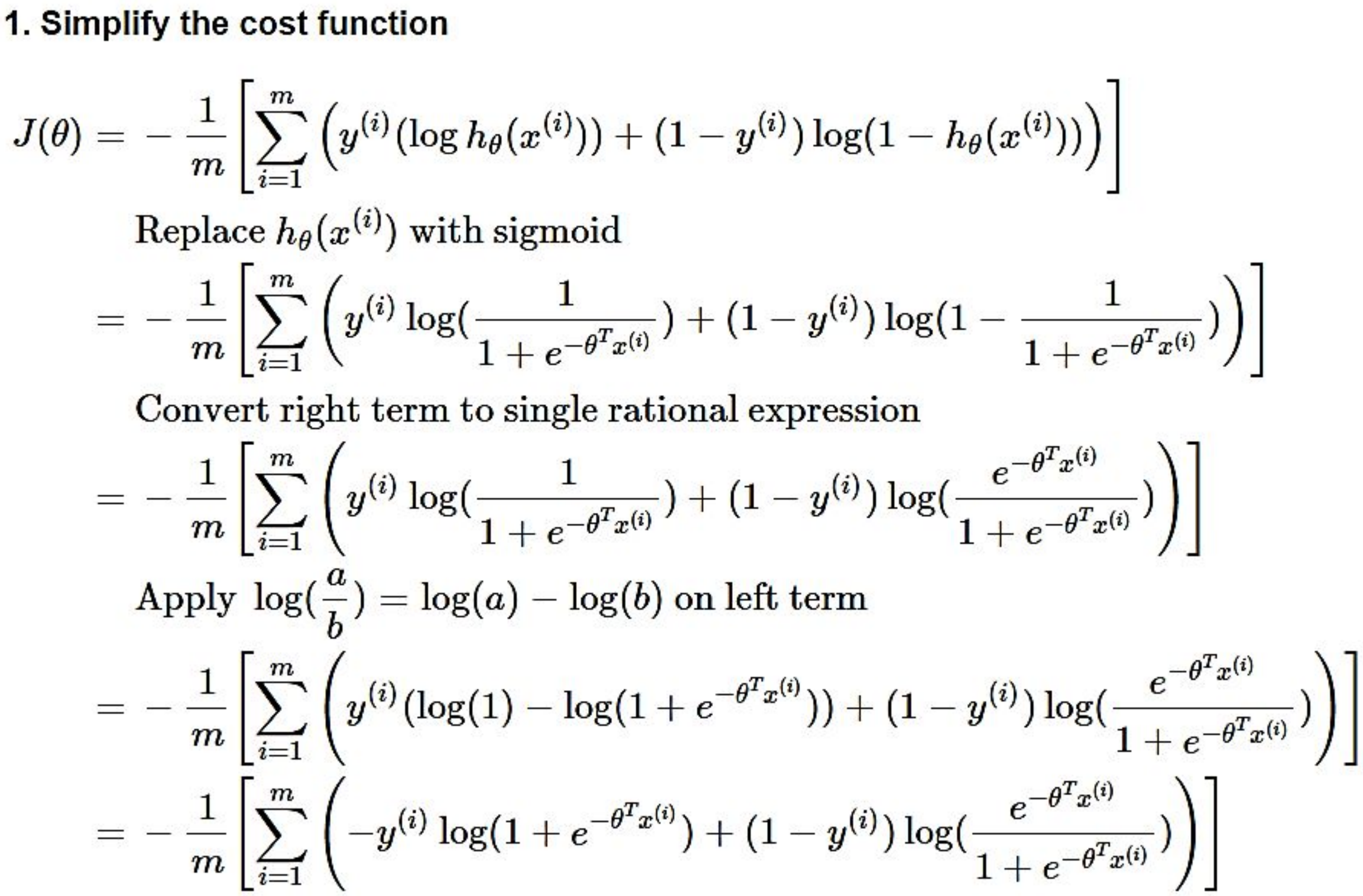
All you need now is to compute the partial derivatives of $(*)$ w.r.t. θ_j. As

$$\frac{\partial}{\partial\theta_j} y_i\theta x^i = y_i x^i_j,$$

$$\frac{\partial}{\partial\theta_j}\log(1+e^{\theta x^i}) = \frac{x^i_j e^{\theta x^i}}{1+e^{\theta x^i}} = x^i_j h_\theta(x^i),$$

the thesis follows.

Deriving the Gradient Descent formula for Logistic Regression



1. Simplify the cost function

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

Replace $h_\theta(x^{(i)})$ with sigmoid

$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\left(y^{(i)}\log(\frac{1}{1+e^{-\theta^T x^{(i)}}}) + (1-y^{(i)})\log(1-\frac{1}{1+e^{-\theta^T x^{(i)}}})\right)\right]$$

Convert right term to single rational expression

$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\left(y^{(i)}\log(\frac{1}{1+e^{-\theta^T x^{(i)}}}) + (1-y^{(i)})\log(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}})\right)\right]$$

Apply $\log(\frac{a}{b}) = \log(a) - \log(b)$ on left term

$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\left(y^{(i)}(\log(1) - \log(1+e^{-\theta^T x^{(i)}})) + (1-y^{(i)})\log(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}})\right)\right]$$

$$= -\frac{1}{m}\left[\sum_{i=1}^{m}\left(-y^{(i)}\log(1+e^{-\theta^T x^{(i)}}) + (1-y^{(i)})\log(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}})\right)\right]$$

Deriving the Gradient Descent formula for Logistic Regression

Apply $\log(\frac{a}{b}) = \log(a) - \log(b)$ to right term

$$-\frac{1}{m}\left[\sum_{i=1}^{m}\left(-y^{(i)}\log(1+e^{-\theta^T x^{(i)}}) + (1-y^{(i)})\log(e^{-\theta^T x^{(i)}}) - (1-y^{(i)})(\log(1+e^{-\theta^T x^{(i)}}))\right)\right]$$

Apply $\log(e^{a}) = a$ to right term

$$-\frac{1}{m}\left[\sum_{i=1}^{m}\left(-y^{(i)}\log(1+e^{-\theta^T x^{(i)}}) + (1-y^{(i)})(-\theta^T x^{(i)}) - (1-y^{(i)})(\log(1+e^{-\theta^T x^{(i)}}))\right)\right]$$

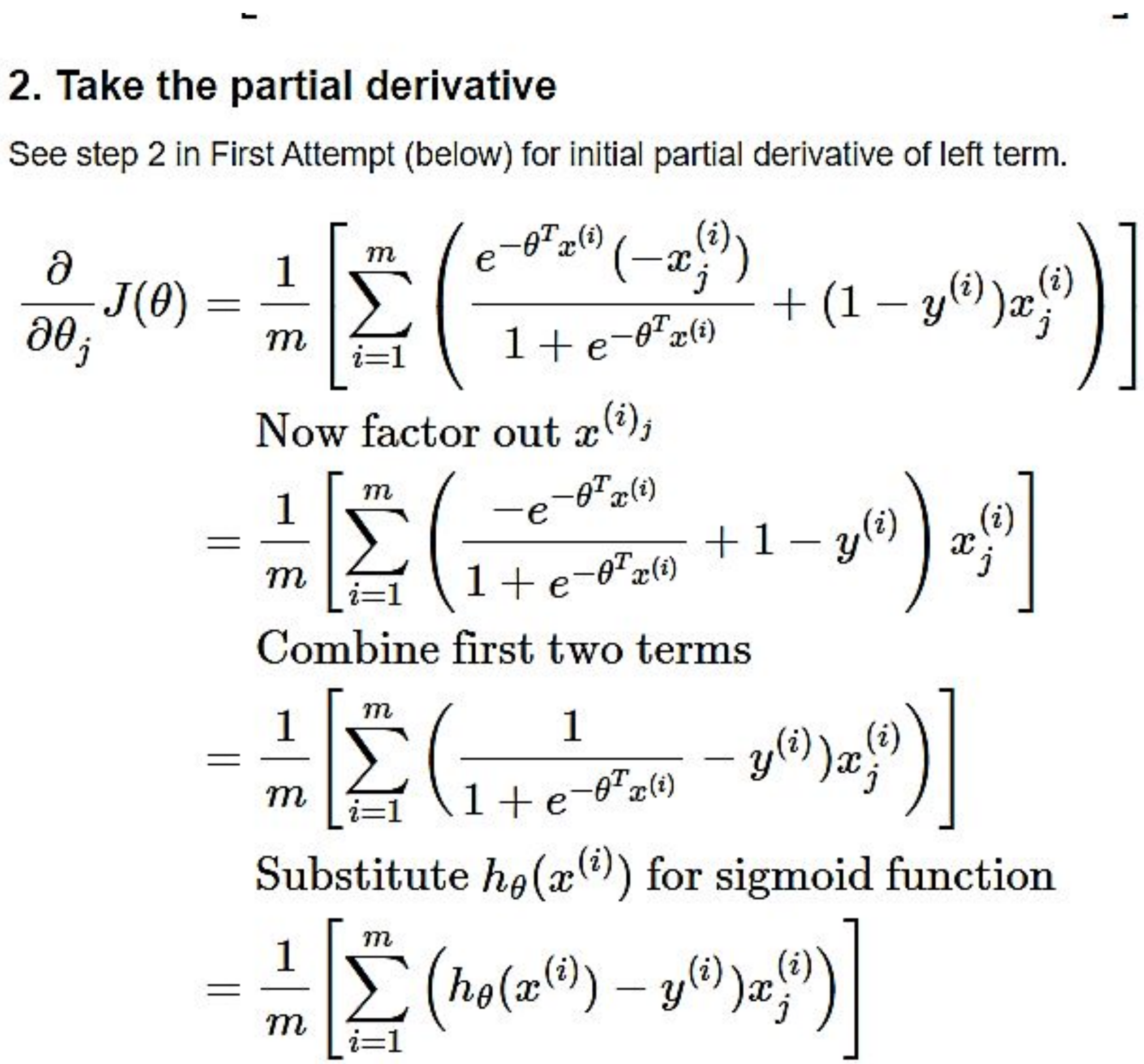
Move minus sign inside $\sum$

$$\frac{1}{m}\left[\sum_{i=1}^{m}\left(y^{(i)}\log(1+e^{-\theta^T x^{(i)}}) + (1-y^{(i)})(\theta^T x^{(i)}) + (1-y^{(i)})(\log(1+e^{-\theta^T x^{(i)}}))\right)\right]$$

Combine first and third terms

$$\frac{1}{m}\left[\sum_{i=1}^{m}\left(\log(1+e^{-\theta^T x^{(i)}}) + (1-y^{(i)})(\theta^T x^{(i)})\right)\right]$$

Deriving the Gradient Descent formula for Logistic Regression



2. Take the partial derivative

See step 2 in First Attempt (below) for initial partial derivative of left term.

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m}\left[\sum_{i=1}^{m}\left(\frac{e^{-\theta^T x^{(i)}}(-x_j^{(i)})}{1+e^{-\theta^T x^{(i)}}} + (1-y^{(i)})x_j^{(i)}\right)\right]$$

Now factor out $x^{(i)}{}_j$

$$= \frac{1}{m}\left[\sum_{i=1}^{m}\left(\frac{-e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}} + 1 - y^{(i)}\right) x_j^{(i)}\right]$$

Combine first two terms

$$= \frac{1}{m}\left[\sum_{i=1}^{m}\left(\frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)})x_j^{(i)}\right)\right]$$

Substitute $h_\theta(x^{(i)})$ for sigmoid function

$$= \frac{1}{m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right)\right]$$