ChatGPT 5.2

Give me a complete mathematical derivation of the algorithm from perceptron to single classification logistic regression
Follow this exact structure strictly and provide only mathematical and technical content (no intuition, no examples, no story narration, no analogies).

Structure:
Problem Definition

State the inputs, outputs, assumptions, and objective function clearly using symbols only.
Mathematical Setup

Define all variables, parameters, and functions used.
Convert the algorithm into equations and formal expressions.
Derivation of the Core Equations

Derive the hypothesis/model equation.
Derive the cost/loss function (if applicable).
Show all partial derivatives w.r.t parameters step-by-step.
Solving for Optimal Parameters

Solve the system of equations created by setting derivatives to zero.
Derive closed-form solutions (if they exist).
Otherwise, derive the update rules (e.g., gradient descent, SGD, EM, dynamic programming).
Final Formula Summary

Present the final compact mathematical expressions.
Show vector/matrix form if appropriate.
Computational Form

Provide final implementable formulas (gradient update or closed form).
Use mathematician-level notation and keep every step explicit.
No intuition, no diagrams, no storytelling — only pure derivation.
explain only what i asked , now cover what happen in

stochastic gradient descent, and mini batch gradient descent

2/3

Below is a **pure mathematical and technical derivation**, strictly following your requested structure, covering the progression **from Perceptron → Single-variable Logistic Regression**, and concluding with **Batch GD, SGD, and Mini-Batch GD**.
No intuition, no examples, no narration.

---

1. Problem Definition

Inputs

$$\{(x_i, y_i)\}_{i=1}^{N}, \quad x_i \in \mathbb{R}^d,\ y_i \in \{0,1\}$$

Outputs

$$\hat{y}_i \in \{0,1\}, \quad \hat{y}_i = \mathbb{I}(f(x_i;\theta) \geq 0)$$

Parameters

$$\theta = (w, b), \quad w \in \mathbb{R}^d,\ b \in \mathbb{R}$$

Assumptions

$$P(y=1|x;\theta) = \sigma(w^\top x + b)$$

Objective Function

$$\min_{\theta}\ \mathcal{L}(\theta)$$

---

2. Mathematical Setup

Variables

$$z_i = w^\top x_i + b$$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$\hat{y}_i = \sigma(z_i)$$

Indicator Function (Perceptron)

$$\phi(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

---

3. Derivation of the Core Equations

3.1 Perceptron Model

Hypothesis

$$\hat{y}_i = \phi(w^\top x_i + b)$$

Loss Function

$$\ell_i(w, b) = \begin{cases} 0 & y_i(w^\top x_i + b) \geq 0 \\ -y_i(w^\top x_i + b) & \text{otherwise} \end{cases}$$

Update Rule (Derived from Subgradient)

$$w^{(t+1)} = w^{(t)} + \eta y_i x_i$$

$$b^{(t+1)} = b^{(t)} + \eta y_i$$

3.2 Logistic Regression Model

Hypothesis Function

$$\hat{y}_i = \sigma(w^\top x_i + b)$$

Likelihood Function

$$P(y_i | x_i; \theta) = \hat{y}_i^{y_i} (1 - \hat{y}_i)^{1 - y_i}$$

Log-Likelihood

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

Cost Function (Negative Log-Likelihood)

$$J(w,b) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)]$$

3.3 Partial Derivatives

Derivative of Sigmoid

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1-\sigma(z))$$

Gradient w.r.t. w

$$\frac{\partial J}{\partial w} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i}{\hat{y}_i} - \frac{1-y_i}{1-\hat{y}_i}\right]\cdot \hat{y}_i(1-\hat{y}_i)x_i$$

Simplification:

$$\frac{\partial J}{\partial w} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i$$

Gradient w.r.t. b

$$\frac{\partial J}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$

---

4. Solving for Optimal Parameters

4.1 Stationary Condition

$$\nabla J(w,b) = 0$$

$$\sum_{i=1}^{N}(\hat{y}_i - y_i)x_i = 0$$

This system is **nonlinear** in w.

4.2 Closed-Form Solution

$$\text{No closed-form solution exists}$$

4.3 Batch Gradient Descent

$$w^{(t+1)} = w^{(t)} - \eta \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i) x_i$$

$$b^{(t+1)} = b^{(t)} - \eta \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

4.4 Stochastic Gradient Descent (SGD)

Single sample i:

$$w^{(t+1)} = w^{(t)} - \eta (\hat{y}_i - y_i) x_i$$

$$b^{(t+1)} = b^{(t)} - \eta (\hat{y}_i - y_i)$$

4.5 Mini-Batch Gradient Descent

Mini-batch $\mathcal{B} \subset \{1, \dots, N\}$, $|\mathcal{B}| = m$:

$$w^{(t+1)} = w^{(t)} - \eta \frac{1}{m} \sum_{i \in \mathcal{B}} (\hat{y}_i - y_i) x_i$$

$$b^{(t+1)} = b^{(t)} - \eta \frac{1}{m} \sum_{i \in \mathcal{B}} (\hat{y}_i - y_i)$$

---

5. Final Formula Summary

Vector Form

$$X \in \mathbb{R}^{N \times d}, \quad y \in \mathbb{R}^N$$

$$\hat{y} = \sigma(Xw + b\mathbf{1})$$

$$\nabla_w J = \frac{1}{N} X^\top (\hat{y} - y)$$

$$\nabla_b J = \frac{1}{N} \mathbf{1}^\top (\hat{y} - y)$$

---

6. Computational Form

Batch Gradient Descent

$$w \leftarrow w - \eta \frac{1}{N} X^\top (\hat{y} - y)$$

$$b \leftarrow b - \eta \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)$$

Stochastic Gradient Descent

$$w \leftarrow w - \eta (\sigma(w^\top x_i + b) - y_i) x_i$$

$$b \leftarrow b - \eta (\sigma(w^\top x_i + b) - y_i)$$

Mini-Batch Gradient Descent

$$w \leftarrow w - \eta \frac{1}{m} X_{\mathcal{B}}^\top (\hat{y}_{\mathcal{B}} - y_{\mathcal{B}})$$

$$b \leftarrow b - \eta \frac{1}{m} \sum_{i \in \mathcal{B}} (\hat{y}_i - y_i)$$

If you want next:

- **Extension to multi-class (softmax)**
- **Newton / IRLS derivation**
- **Connection to maximum entropy**
- **Convergence proofs**

say which one.