Date
Page

Weighted least Square, Logistics Regression Newtons method

~~Ton Regression, yon~~

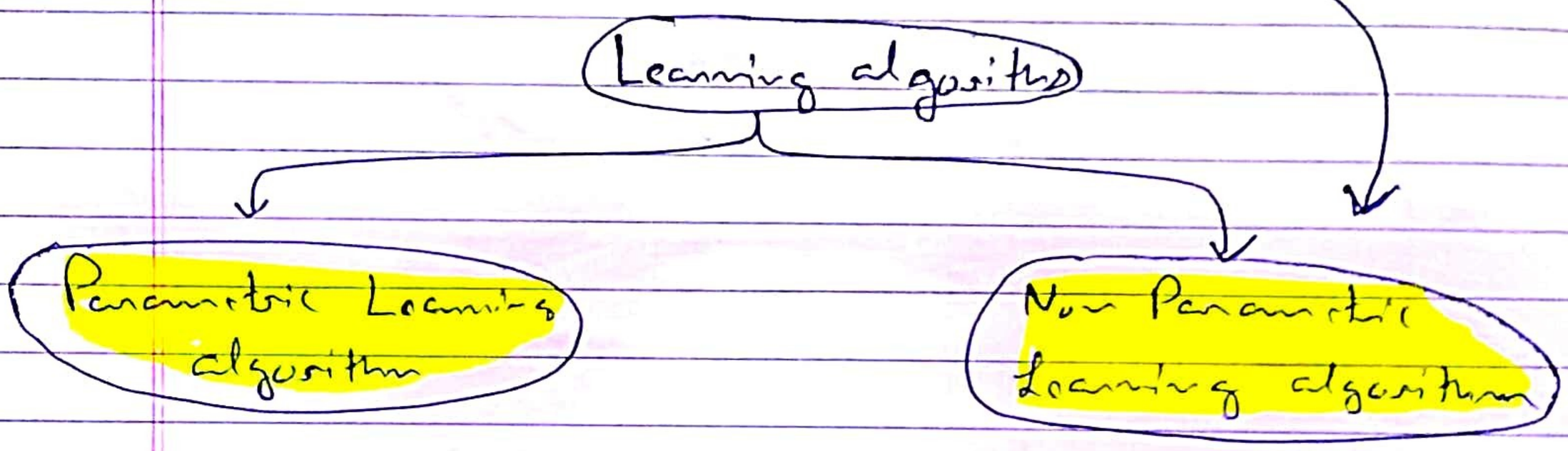
⇒ Using Linear Regression you can also train a non linear model by using a simple trick.

$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_1^2$ {Non linear model in X_1}

Let new feature $X_2 = X_1^2$

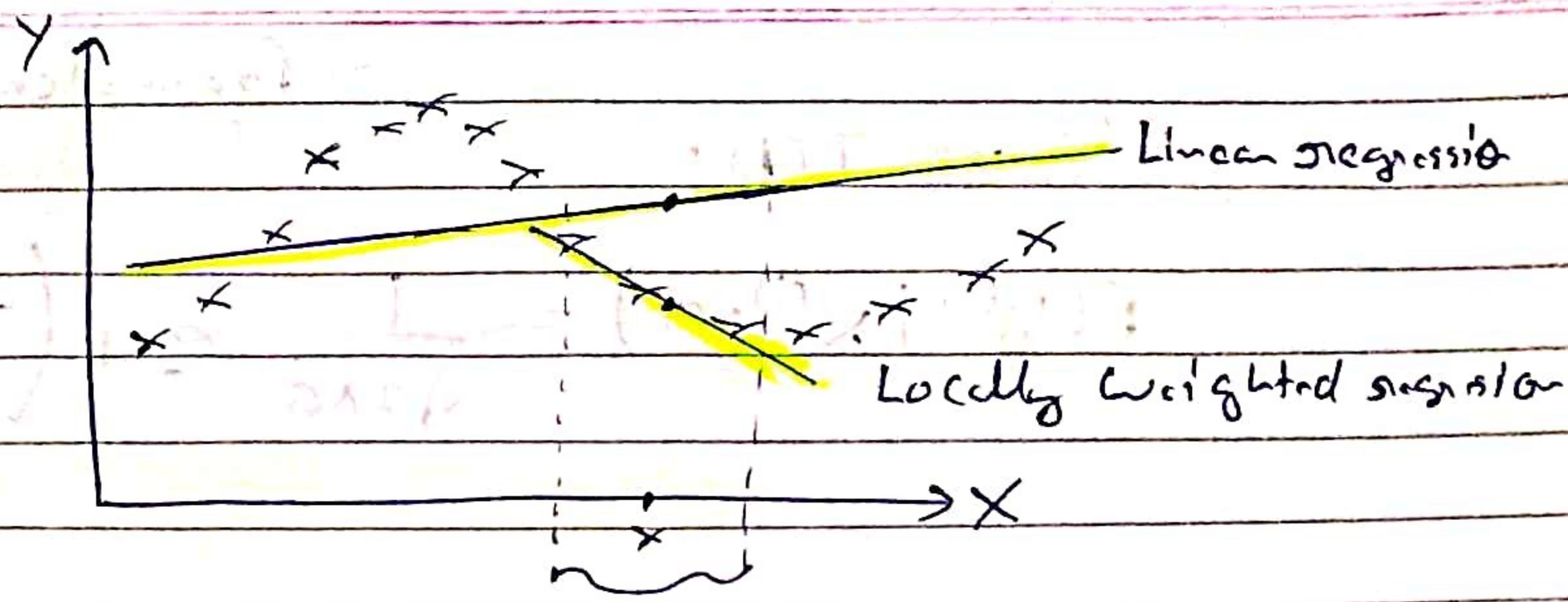
So $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2$ {Now the model is linear in two feature X_1 & X_2}

★ Locally weighted regression



{fit a fixed set of parameters (θ_i) to data}

{Amount of data (parameters) you need to keep grows with size of data}



Fit Θ to minimize

$$J = \sum_{i=1}^{m} w^{(i)} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

→ Weight function

$$w^{(i)} = \exp\left(\frac{-(x^{(i)} - x)^2}{2\tau^2}\right)$$ { Common choice of weighting function }

⇒ Generally used when you have a relatively low dimensional dataset and we have a lot of data.

★ Probabilistic interpretation

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$

→ { Unmodeled effect, random noise etc }

Let $\varepsilon^{(i)} \sim N(0, \sigma^2)$

$$P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon^{(i)2}}{2\sigma^2}\right)$$

$\varepsilon^{(i)}$ are IID.

$$P(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

→ Parameterized by θ

$$y^{(i)} \mid x^{(i)}, \theta \sim N(\theta^T x^{(i)}, \sigma^2)$$

$$P(y \mid X; \theta) = \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta)$$

$$\log(P(y \mid X; \theta)) = -\sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} + m \log \frac{1}{\sqrt{2\pi}\sigma}$$

⇒ Using MAP estimation:

$$\theta = \min_{\theta} \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

$J(\theta)$

13:50

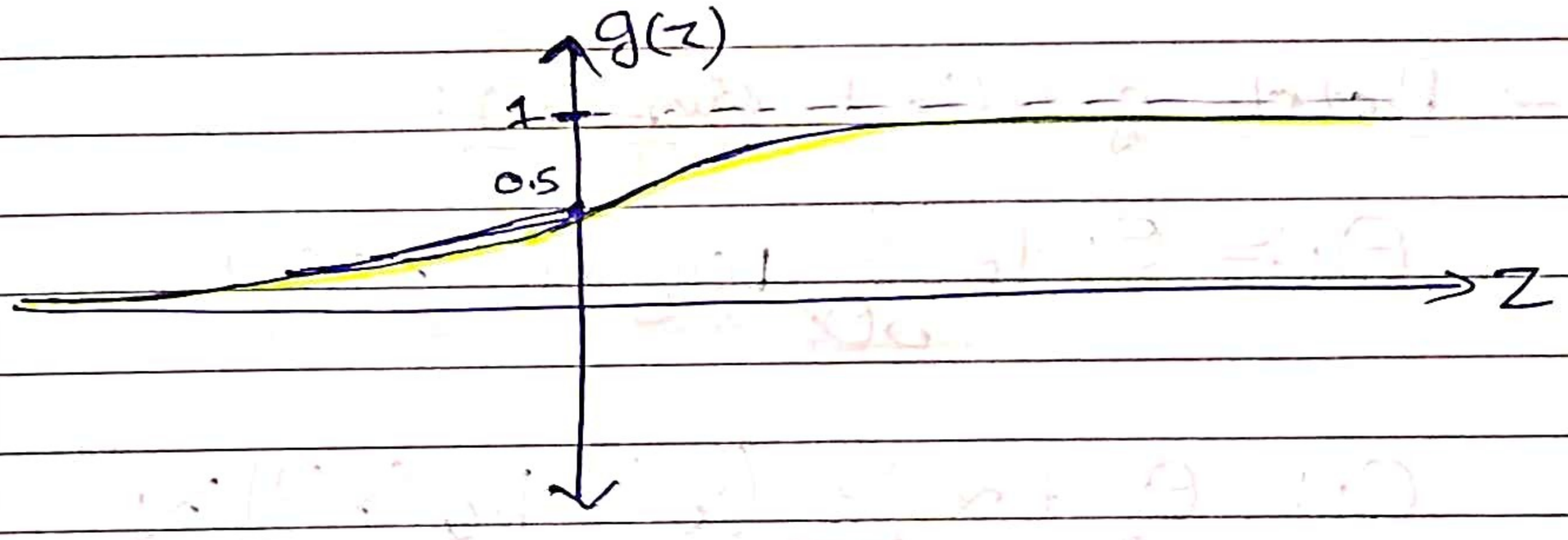
Date
Page

★ Classification

① Binary Classification $\{y \in [0, 1]\}$

{Logistic Regression}

$$\boxed{g(z) = \frac{1}{1+e^{-z}}} \quad \{\text{Sigmoid or Logistic function}\}$$



Let $\boxed{h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}}$ {Hypothesis function}

$$\begin{cases} P(Y=1 \mid X;\theta) = h_\theta(x) \\ P(Y=0 \mid X;\theta) = 1 - h_\theta(x) \end{cases}$$

$$\rightarrow P(Y \mid X;\theta) = h_\theta(x)^{y} (1 - h_\theta(x))^{1-y}$$

$$P(Y|X;\theta) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)};\theta)$$

$$= \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}} (1-h_\theta(x))^{1-y^{(i)}}$$

$$\log P(Y|X;\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))$$

⇒ Choose θ to maximize $\log P(Y|X;\theta)$.

⇒ Batch gradient ascent:

$$\theta_j := \theta_j + \alpha \frac{\delta}{\delta\theta_j} \log P(Y|X;\theta)$$

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

⇒ The above function has only one optimum, no local optimum.

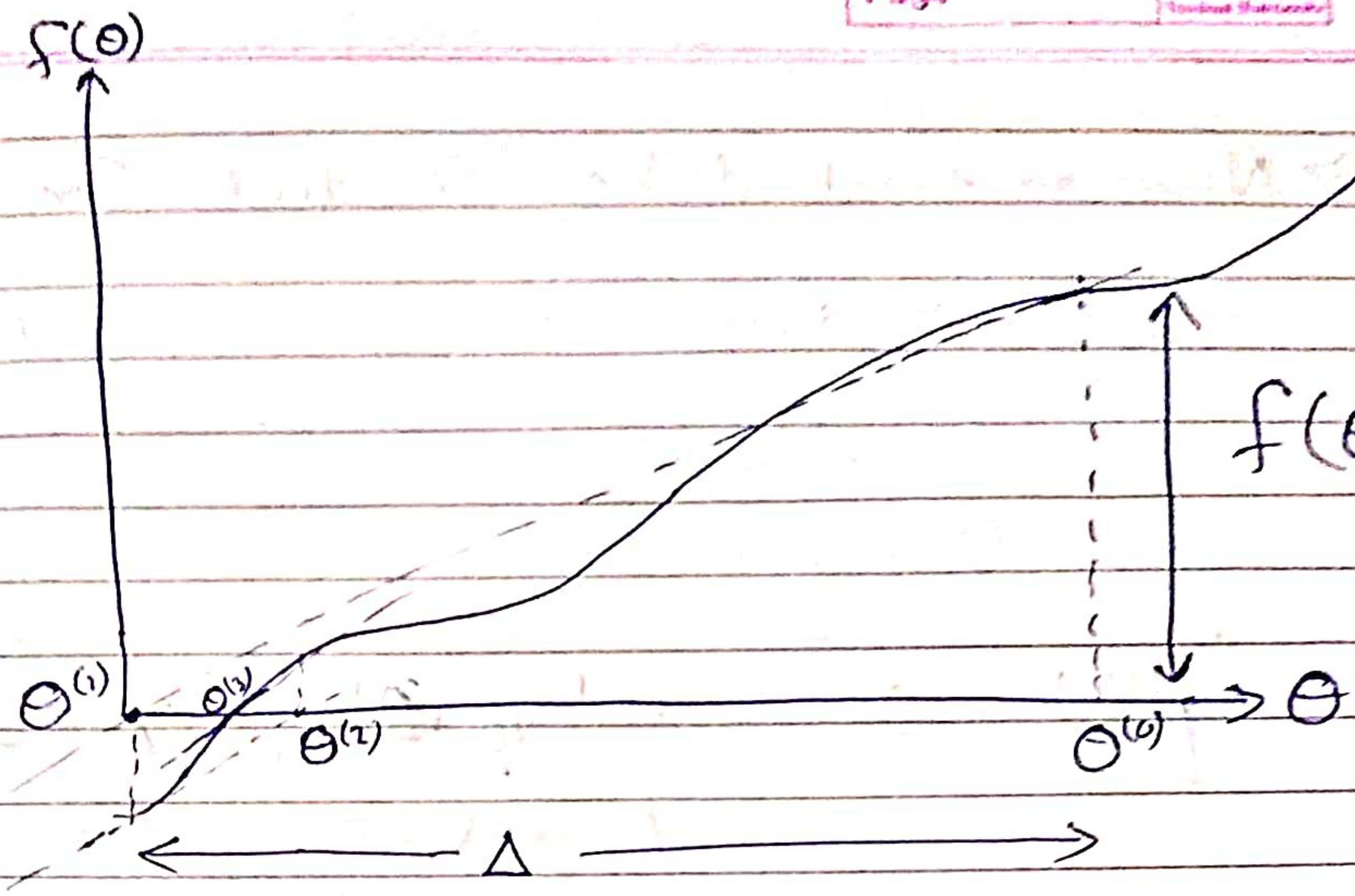
★ Newtons Method

"Given f, find θ, st $f(\theta)=0$"

Approach: (1) Start with a guess $\theta^{(0)}$.

(2) $\theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$

(3) Iterate



$$\theta^{(1)} = \theta^{(0)} - \Delta \qquad f'(\theta^{(0)}) = \frac{f(\theta^{(0)})}{\Delta}$$

$$\theta^{(1)} = \theta^{(0)} - \frac{f(\theta^{(0)})}{f'(\theta^{(0)})}$$

On general,

$$\theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$$

$\Rightarrow$ Let $f(\theta) = \frac{\delta}{\delta\theta} \log(P(y|x;\theta))$

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\frac{\delta}{\delta\theta}\log(P(y|x;\theta))}{\frac{\delta^2}{\delta\theta^2}\log(P(y|x;\theta))}$$

Date
Page

OM
Student Notebooks

⇒ Newtons method has "Quadratic Convergence".

$0.01 \rightarrow 0.0001 \rightarrow 0.00000001$

Error

⇒ When θ is a vector:

$$\theta^{(t+1)} = \theta^{(t)} + H^{-1} \nabla_\theta \log(P(Y|X;\theta))$$

Gradient

Hessian matrix

$$H_{ij} = \frac{\delta^2 f}{\delta\theta_i \delta\theta_j}$$