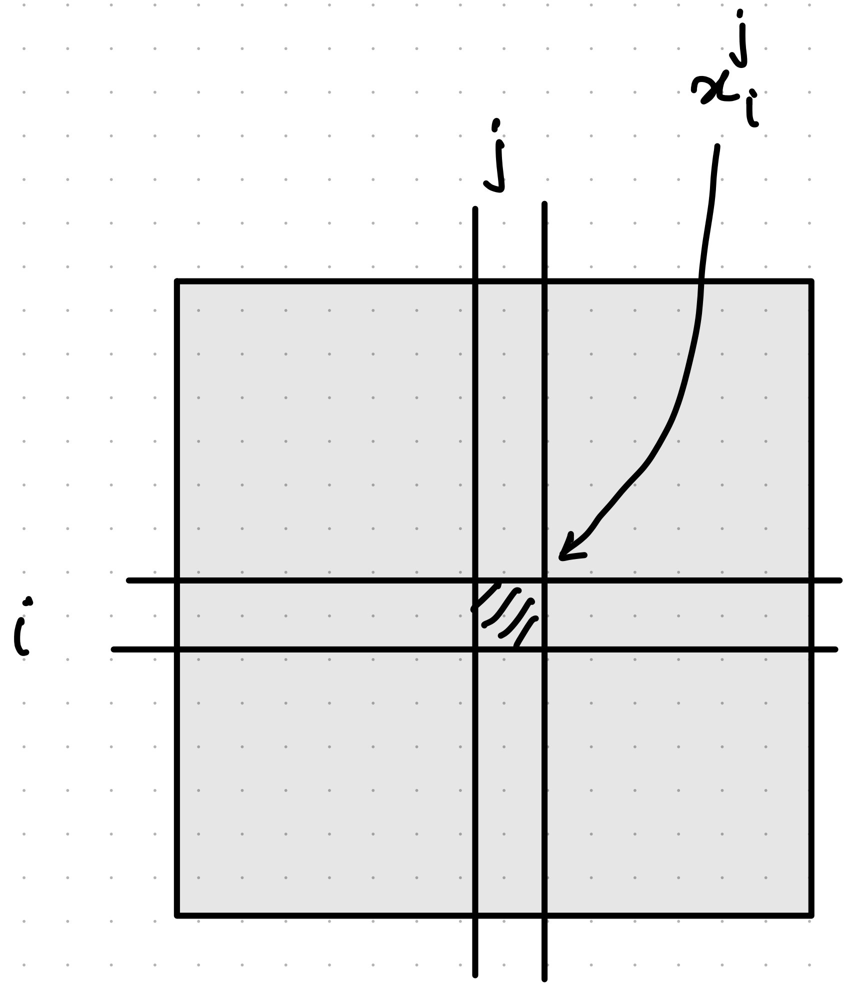
CONVEXITY OF CROSS ENTROPY LOSS

$$J(\theta) = -\sum_{i=1}^{N} \left(y_i \log \hat{y}_i + (1-y_i) \log (1-\hat{y}_i) \right)$$

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

Where

x_i^j — $j \leftarrow j^{th}$ element, $i \leftarrow i^{th}$ data point



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

$$H = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \frac{\partial J(\theta)}{\partial \theta_1} & \ddots & \\ \frac{\partial}{\partial \theta_2} \frac{\partial J(\theta)}{\partial \theta_1} & & \\ \vdots & & \\ & & \frac{\partial}{\partial \theta_d} \frac{\partial J(\theta)}{\partial \theta_d} \end{bmatrix}$$

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

$$H = \begin{bmatrix} \frac{\partial}{\partial \theta_1}\frac{\partial J(\theta)}{\partial \theta_1} & \ddots & \\ \frac{\partial}{\partial \theta_2}\frac{\partial J(\theta)}{\partial \theta_1} & & \\ \vdots & & \\ & & \frac{\partial}{\partial \theta_d}\frac{\partial J(\theta)}{\partial \theta_d} \end{bmatrix}$$

Let us compute a $\frac{\partial}{\partial \theta_j}\frac{\partial J(\theta)}{\partial \theta_k}$ as H_{jk} entry

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

Let us compute a $\frac{\partial}{\partial \theta_j} \frac{\partial J(\theta)}{\partial \theta_k}$ as H_{jk} entry

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

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

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

Let us compute a $\frac{\partial}{\partial \theta_j} \frac{\partial J(\theta)}{\partial \theta_k}$ as H_{jk} entry

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

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

$$H_{jk} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{N} \hat{y}_i \, x_i^k$$

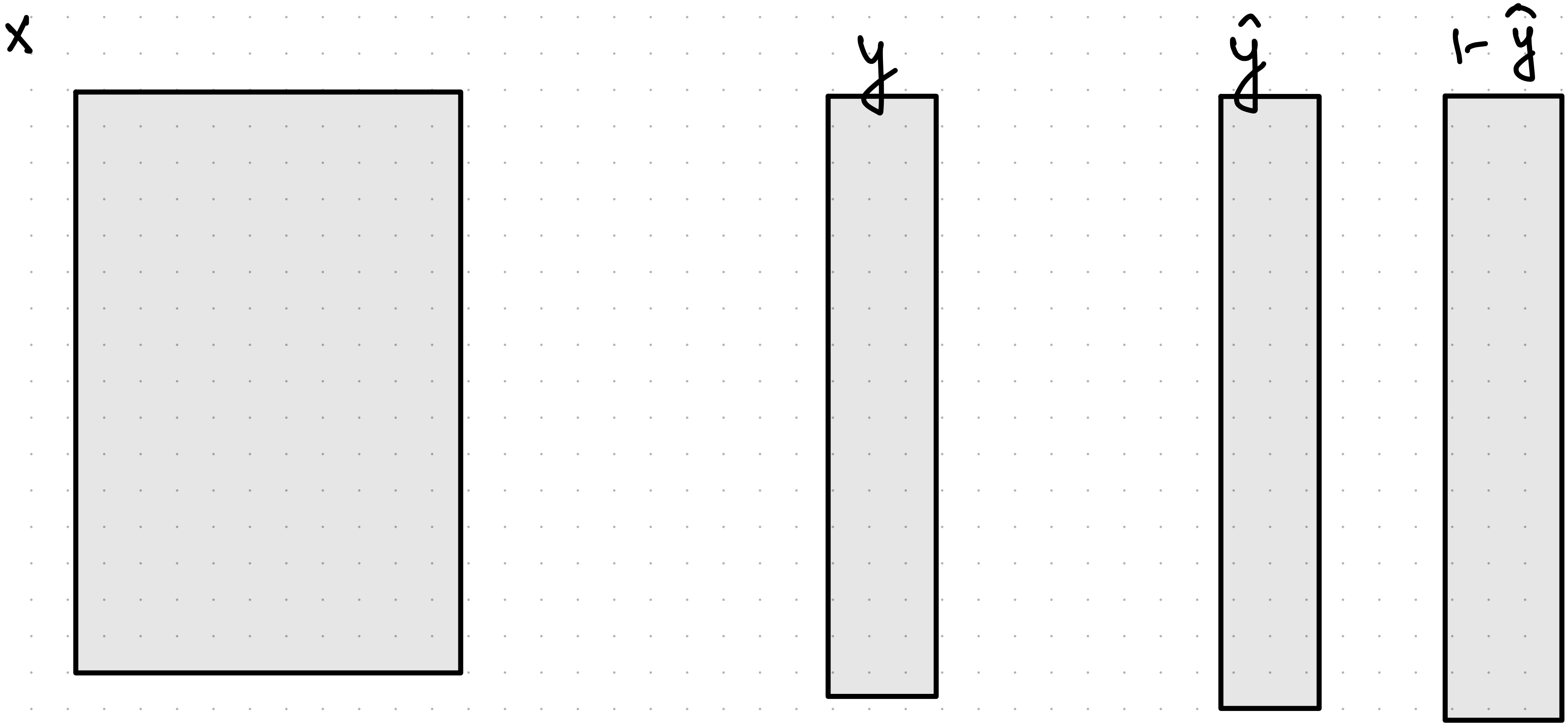
$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_{i=1}^{N} (\hat{y}_i - y_i) x_i^j$$

$$H_{jk} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{N} \hat{y}_i x_i^k$$

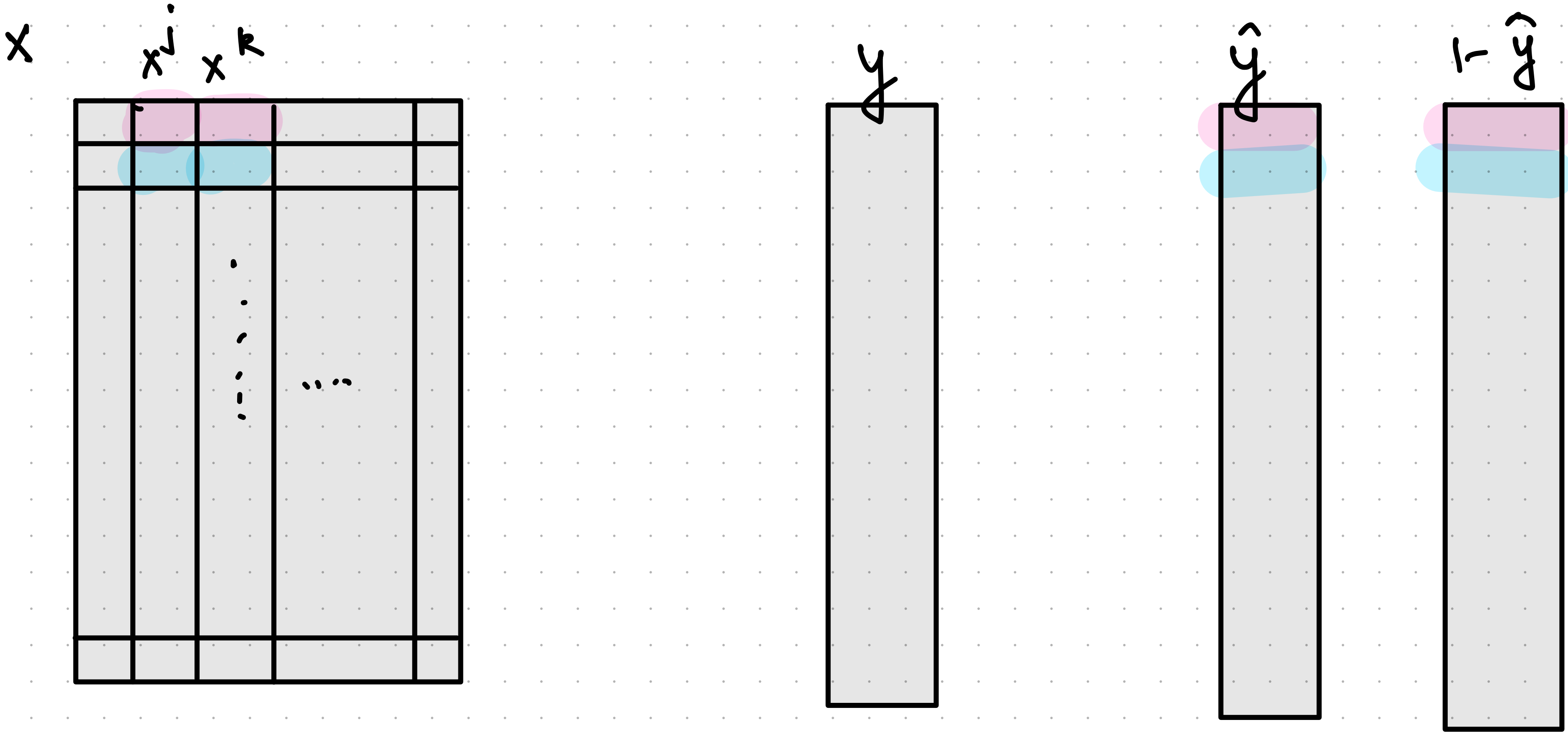
$$= \sum_{i=1}^{N} \hat{y}_i (1-\hat{y}_i) x_i^k * \frac{\partial}{\partial \theta_j} (x_i^1 \theta_1 + x_i^2 \theta_2 + \dots)$$

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

$$H_{jk} = \sum_{i=1}^{N} \hat{y}_i (1-\hat{y}_i)\, x_i^k\, x_i^j$$

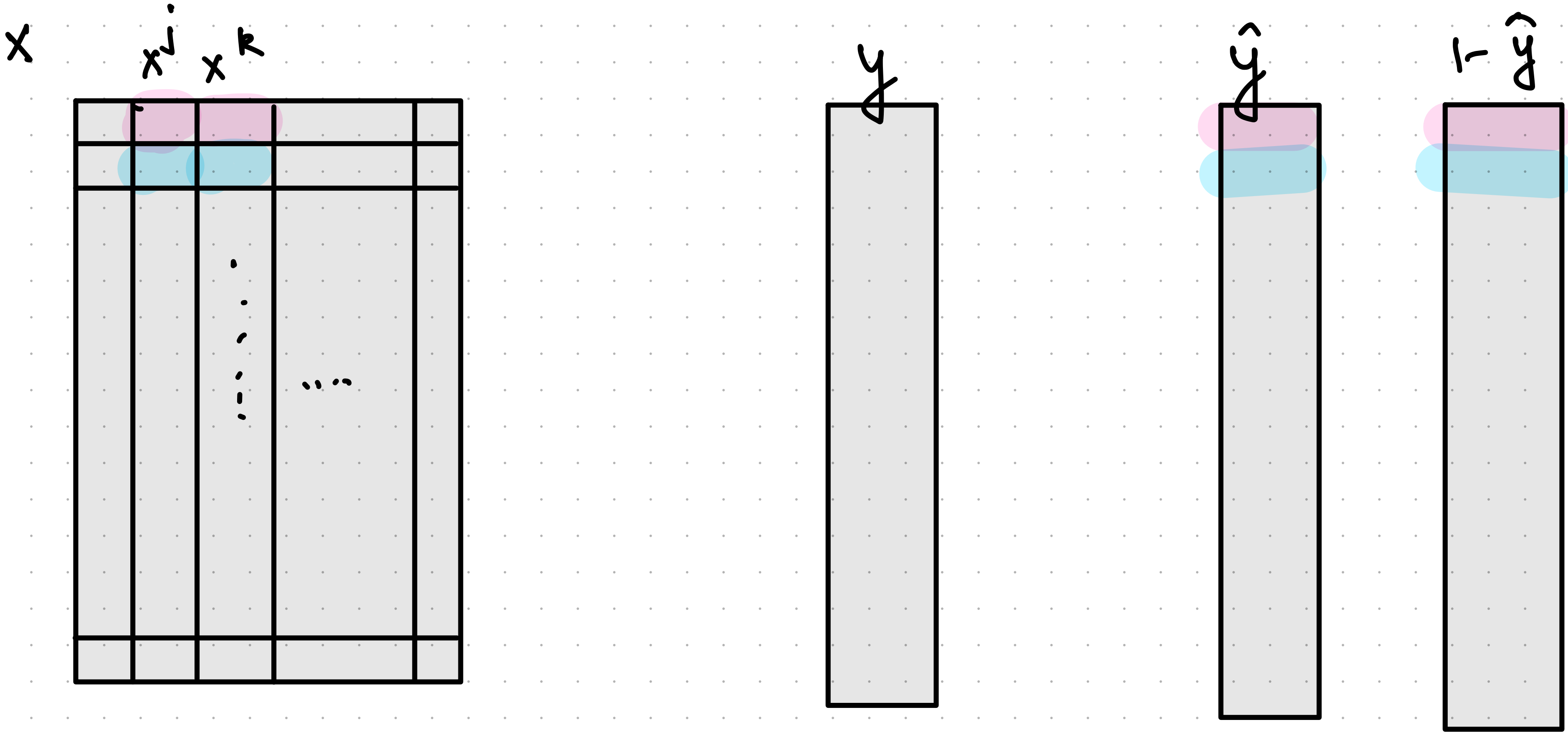


$$H_{jk} = \sum_{i=1}^{N} \hat{y}_i (1 - \hat{y}_i) x_i^k x_i^j$$



Say $j = 2; k = 3$

$$H_{jk} = \sum_{i=1}^{N} \hat{y}_i (1-\hat{y}_i)\, x_i^k\, x_i^j$$



$$H_{jk} = \underbrace{{x^j}^T}_{1\times N} \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & \cdots & 0 & 0 \\ & \ddots & & \\ & & & \hat{y}_N(1-\hat{y}_N) \end{bmatrix}_{N\times N} x^k_{N\times 1} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y}))\, x^k$$

Element-wise mult

$$H_{jk} = \sum_{i=1}^{N} \hat{y_i}(1-\hat{y}_i)\, x_i^k \, x_i^j$$

$$H_{jk} = \underset{1 \times N}{X^{j^T}} \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & \cdots & 0 & 0 \\ & \ddots & & \\ & & & \hat{y}_N(1-\hat{y}_N) \end{bmatrix}_{N \times N} \underset{N \times 1}{X^k}$$

$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y}))\, x^k$$

$$H_{jk} = \sum_{i=1}^{N} \hat{y}_i (1-\hat{y}_i)\, x_i^k\, x_i^j$$

$$H_{jk} = \underset{1\times N}{x^{j\,T}} \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & \cdots & 0 & 0 \\ & \ddots & & \\ & & & \hat{y}_N(1-\hat{y}_N) \end{bmatrix}_{N\times N} x^k_{N\times 1}$$

$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y}))\, x^k = x_j^T D\, x^k$$

What can we say about D?

$$H_{jk} = \sum_{i=1}^{N} \hat{y}_i (1-\hat{y}_i)\, x_i^k\, x_i^j$$

$$H_{jk} = \underset{1\times N}{x^{j^T}} \begin{bmatrix} \hat{y}_1(1-\hat{y}_1) & \cdots & 0 & 0 \\ & \ddots & & \\ & & & \hat{y}_N(1-\hat{y}_N) \end{bmatrix}_{N\times N} \underset{N\times 1}{x^k}$$

$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y}))\, x^k = x_j^T D\, x^k$$

What can we say about D?

Each $\hat{y} \in [0,1]$; $1-\hat{y} \in [0,1]$

Max $\hat{y}[1-\hat{y}] = 0.5 * 0.5 = 0.25$

$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y})) x^k = x_j^T D x^k$$

What can we say about D?

Each $\hat{y} \in [0,1]$; $1-\hat{y} \in [0,1]$

Max $\hat{y}[1-\hat{y}] = 0.5 * 0.5 = 0.25$

Thus D is of type

$$D = \begin{bmatrix} 0 \leq D_{11} \leq 0.25 & 0 & 0 & \cdots \\ 0 & \ddots & & \\ 0 & & \ddots & \\ \vdots & & & 0 \leq D_{NN} \leq 0.25 \end{bmatrix}$$

$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y})) x^k = x_j^T D x^k$$

What can we say about D?

Each $\hat{y} \in [0,1]$; $1-\hat{y} \in [0,1]$

Max $\hat{y}[1-\hat{y}] = 0.5 * 0.5 = 0.25$

Thus D is diagonal matrix where diagonal entries are b/w 0 and 0.25

$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y}))x^k = x_j^T D x^k$$

What does H look like?

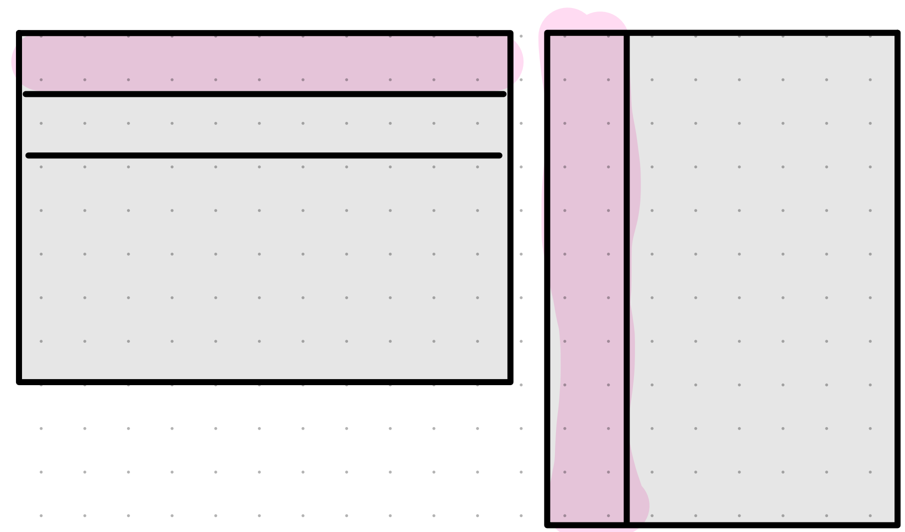
$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y}))\, x^k = x^{j^T} D\, x^k$$

What does H look like?

$$H_{11} = x^{1^T} D\, x^1$$

$$H_{21} = x^{2^T} D\, x^1$$

$$\vdots$$



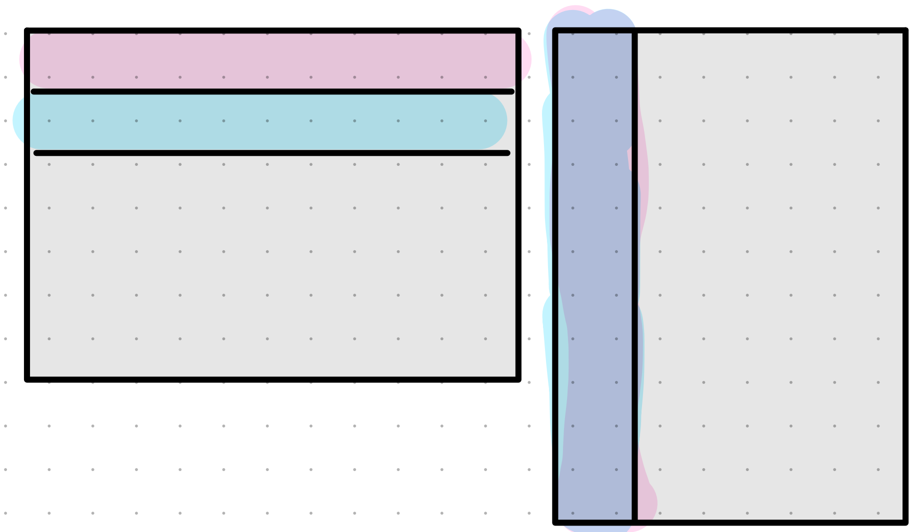
$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y})) x^k = x^{j^T} D \, x^k$$

What does H look like?

$$H_{11} = x^{1^T} D \, x^1$$

$$H_{21} = x^{2^T} D \, x^1$$

$$\vdots$$



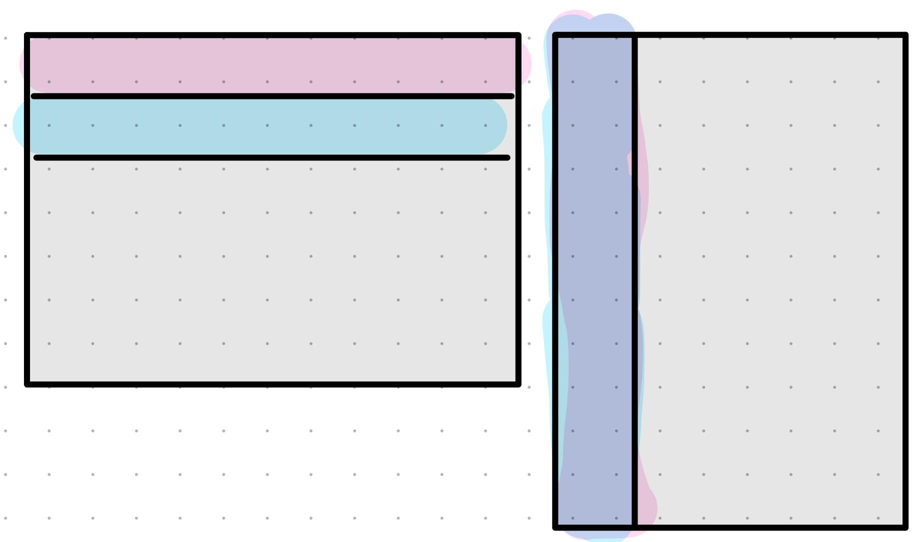
$$H_{jk} = x_j^T \operatorname{diag}(\hat{y} \odot (1-\hat{y})) x^k = x^{j^T} D x^k$$

What does H look like?

$$H_{11} = x^{1^T} D x^1$$

$$H_{21} = x^{2^T} D x^1$$

$$\vdots$$



$$\boxed{H = X^T D X}$$

$$H = X^T D X$$

Q: Prove H is P.S.D matrix

$$H = X^T D X$$

Q: Prove H is P.S.D matrix

$$v^T H v = v^T X^T D X v$$

let $Xv = Z$

$$v^T H v = Z^T D Z$$

$$= \sum_{i=1}^{N} d_{ii} Z_i^2 \quad \text{where} \quad 0 \leq d_{ii} \leq 0.25$$

≥ 0 $\therefore$ H is P.S.D. $\Rightarrow J(\theta)$ is convex

Iteratively Reweighted Least Squares

I) First Order Update Rule

$$\theta_{t+1} = \theta_t - \alpha \nabla J(\theta)$$

Typically g: gradient $\nabla J(\theta)$

H: Hessian

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = \theta_t - H^{-1} g$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = \theta_t - H^{-1} g$$

For logistic regression:

$$\theta_{t+1} = \theta_t - (X^T D X)^{-1} X^T (\hat{y} - y)$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = \theta_t - H^{-1} g$$

For logistic regression:

$$\theta_{t+1} = \theta_t - (X^T D X)^{-1} X^T (\hat{y} - y)$$

$$= (X^T D X)^{-1} \left[X^T D X \theta_t - X^T (\hat{y} - y) \right]$$

$$\theta_{t+1} = (X^T D X)^{-1} X^T D \left[X \theta_t - D^{-1} (\hat{y} - y) \right]$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = (X^T D X)^{-1} X^T D \left[X\theta_t - D^{-1}(\hat{y} - y) \right]$$

$$= (X^T D X)^{-1} X^T D z_t \qquad ; z_t = X\theta_t - D^{-1}(\hat{y} - y)$$

Contrast w/ weighted linear regression

$$\hat{\theta} = (X^T D X)^{-1} X^T D y$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = (X^T D X)^{-1} X^T D \left[X\theta_t - D^{-1}(\hat{y} - y) \right]$$

$$= (X^T D X)^{-1} X^T D z_t \qquad ; z_t = X\theta_t - D^{-1}(\hat{y} - y)$$

Contrast w/ weighted linear regression

$$\hat{\theta} = (X^T D X)^{-1} X^T D y$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = (X^T D X)^{-1} X^T D \left[X\theta_t - D^{-1}(\hat{y} - y) \right]$$

$$= (X^T D X)^{-1} X^T D z_t \quad ; \; z_t = X\theta_t - D^{-1}(\hat{y} - y)$$

Contrast w/ weighted linear regression

$$\hat{\theta} = (X^T D X)^{-1} X^T D y$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = (X^T D X)^{-1} X^T D \left[X\theta_t - D^{-1}(\hat{y} - y) \right]$$

$$= (X^T D X)^{-1} X^T D z_t \quad ; \; z_t = X\theta_t - D^{-1}(\hat{y} - y)$$

$$D = \begin{bmatrix} \hat{y}_1(1-\hat{y}) & & \\ & \ddots & \\ & & \ddots \end{bmatrix}$$

Contrast w/ weighted linear regression

$$\hat{\theta} = (X^T D X)^{-1} X^T D y$$

Iteratively Reweighted Least Squares

II) Second Order Update Rule

$$\theta_{t+1} = (X^T D X)^{-1} X^T D \left[X\theta_t - D^{-1}(\hat{y} - y) \right]$$

$$= (X^T D X)^{-1} X^T D z_t \quad ; z_t = X\theta_t - D^{-1}(\hat{y} - y)$$

$$D = \begin{bmatrix} \hat{y}_1(1-\hat{y}) & & \\ & \ddots & \\ & & \ddots \end{bmatrix}$$

$f(\theta_t)$

Contrast w/ weighted linear regression

$$\hat{\theta} = (X^T D X)^{-1} X^T D y$$