$$\mathcal{L}(\mathbf{w}) = \frac{1}{N} NLL(\mathbf{w}) + \frac{1}{2}\lambda||\mathbf{w}||_2^2$$

$$NLL(w) = -\log p(y|X,w) = \sum_{i=1}^{N} -\log p(y_i|x_i,w) = \sum_{i=1}^{N} L_i(w)$$

$$x_i = (1, \text{—} x_i \text{—}) \qquad w = (b, \text{—} w \text{—})$$

$$\nabla_w NLL(w) = \sum_{i=1}^{N} \nabla_w L_i(w) = \sum_{i=1}^{N} (\sigma(w^T x_i) - y_i) \cdot x_i$$

$$(1\times1)\,(D\times1) \quad (1\times1) \quad (D\times1)$$

$$\boxed{\nabla_w L_i(w)} = \nabla_w\left(-y_i w^T x_i + \ln(1+e^{w^T x_i})\right) = -y_i \cdot x_i + \overset{\alpha}{\sigma}(w^T x_i)\, x_i = \boxed{(\sigma(w^T x_i) - y_i)\cdot x_i}$$

$$\boxed{\nabla_w(-y_i w^T x_i)} = -y_i \nabla_w(w^T x_i) = \boxed{-y_i \cdot x_i}$$

$$\boxed{\nabla_w \ln(1+e^{w^T x_i})} = \left(\frac{\partial \ln(1+e^{w^T x_i})}{\partial(1+e^{w^T x_i})} \cdot \frac{\partial(1+e^{w^T x_i})}{\partial(w^T x_i)} \cdot \underbrace{\frac{\partial(w^T x_i)}{\partial w}}_{x_i^T}\right)^T$$

$$\frac{\partial \ln(z)}{\partial z} = \frac{1}{z}$$

$$\frac{\partial(1+e^a)}{\partial a} = e^a$$

$$= \underbrace{\frac{1}{1+e^{w^T x_i}}}_{1\times1} \cdot \underbrace{e^{w^T x_i}}_{1\times1} \cdot \underbrace{x_i}_{D\times1}$$

$$= \frac{e^{w^T x_i}}{1+e^{w^T x_i}} \cdot x_i = \frac{1}{1+e^{-w^T x_i}} \cdot x_i$$

$$= \boxed{\sigma(w^T x_i)\cdot x_i}$$

$$\overbrace{[1, \dots, 1]}^{D}$$

$$\sum_{i=1}^{N} \underbrace{(\sigma(w^T x_i) - y_i)}_{\alpha_i} \cdot x_i = \sum_{i=1}^{N} \alpha_i x_i = \mathrm{sum}(\alpha \cdot \mathbb{1}_D^T \odot X, \text{axis}=0)$$

$$\cancel{\alpha \odot X}$$

X $(N \times D)$

w (D) $(D\times1)$

y (N)

$\alpha = (\text{sigmoid}(X @ w) - y)$ # shape $(N \times 1)$

np dot(X, w)

$$\mathrm{np.sum}(\alpha[:, \mathrm{np.newaxis}] * X, \mathrm{axis}=0)$$

$$X^T\alpha$$

$$\alpha^T X \quad \mapsto \quad \checkmark$$

$$(1\times N)\ (N\times D) \qquad (1\times D)$$

$$[\alpha^T X]_j = \sum_{i=1}^{N} \alpha_i \cdot x_{ij} = \frac{\partial}{\partial w_j} NLL(w)$$

where $\alpha_i = (\sigma(w^T x_i) - y_i)$

$$\nabla_w NLL(w) = \sum_{i=1}^{N} (\sigma(w^T x_i) - y_i) \cdot x_i = X^T \alpha$$