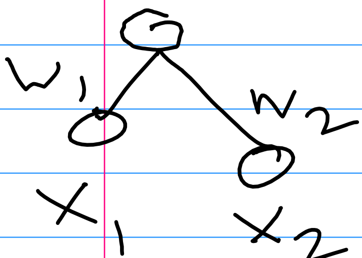
$\bar{y} = w_1 x_1 + w_2 x_2$ $\quad \mathcal{L}(w) = \sum_i \frac{1}{2}(y^i - \bar{y}^i)^2$

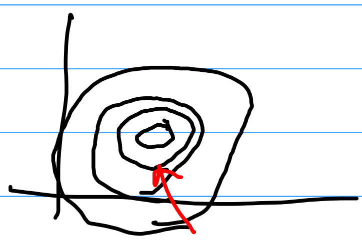
(x^i, y^i)



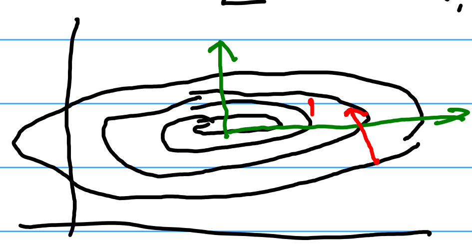
$$\mathcal{L}(w) = \frac{1}{2} w^T \underbrace{\left[\sum_i x^i x^{i^T}\right]}_{H} w + \cdots$$

$$H_{pq} = \frac{\partial^2 \mathcal{L}}{\partial w_p \partial w_q} = \frac{\partial\left(\frac{\partial \mathcal{L}}{\partial w_p}\right)}{\partial w_q}$$

$H = cI$



$$H = \begin{bmatrix} h_{11} & & 0 \\ & h_{22} & \\ 0 & & \ddots \end{bmatrix}$$



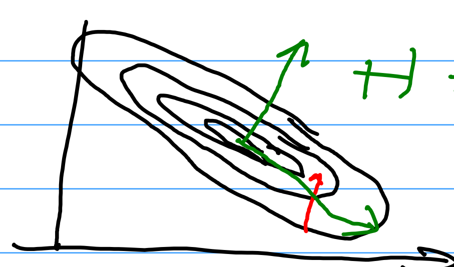
$$\eta_i = \frac{c}{h_{ii}}$$

$\boxed{\text{diagonal Levenberg-Marquardt}}$

$$\eta_i = \frac{c}{h_{ii} + \gamma}$$

$$\eta_i = \frac{c}{\max(h_{ii}, \gamma)}$$

H = full matrix



$H = Q^T \Lambda Q$

$$H \Delta W = \frac{\partial \mathcal{L}}{\partial W}$$ Newton's algo

$$W \leftarrow W - H^{-1} \frac{\partial \mathcal{L}}{\partial W}$$

Gauss-Newton Approx

$$\mathcal{L}(w) = \sum_i \frac{1}{2} \| Y^i - G(w, x^i) \|^2$$

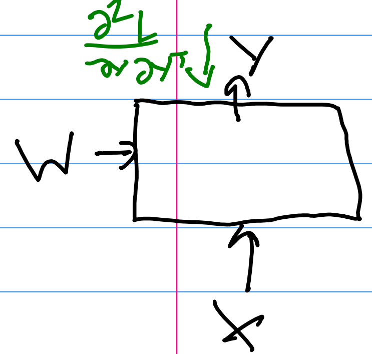
$$\frac{\partial \mathcal{L}}{\partial w} = \sum_i -(Y^i - G(w, x^i))^T \frac{\partial G(w, x^i)}{\partial w}$$ ← Jacobian

$$\frac{\partial^2 \mathcal{L}}{\partial w \partial w^T} = \frac{\partial \left[\frac{\partial \mathcal{L}}{\partial w}\right]}{\partial w^T} = \underbrace{\sum_i \frac{\partial G(w, x^i)}{\partial w}^T \frac{\partial G(w, x_i)}{\partial w}}_{\text{Square Jacobian} = \tilde{H}} + \ldots$$

$$\ldots \; (Y^i - G(w, x^i))^T \frac{\partial^2 G(w, x^i)}{\partial w \, \partial w^T}$$

Levenberg-Marquardt: $(\tilde{H} + \lambda I)\Delta W = \frac{\partial \mathcal{L}(w)}{\partial w}$

2nd derivatives in a deep net



$$\frac{\partial^2 L}{\partial Y \partial Y^T} \qquad \frac{\partial^2 L}{\partial x \partial x^T} \, ? \qquad \frac{\partial^2 L}{\partial W \partial W^T} \, ?$$

$$\frac{\partial^2 L}{\partial W \partial W^T} = \frac{\partial Y^T}{\partial W} \frac{\partial^2 L}{\partial Y \partial Y^T} \frac{\partial Y}{\partial W} + \frac{\partial L}{\partial Y} \cdot \frac{\partial^2 Y}{\partial W \partial W^T}$$

$N \times N \quad N \times M \quad M \times M \quad M \times N$

$$\frac{\partial \left[\frac{\partial L}{\partial W}\right]}{\partial W^T} = \frac{\partial \left[\frac{\partial L}{\partial Y} \frac{\partial Y}{\partial W}\right]}{\partial W^T} = \frac{\partial \left[\frac{\partial L}{\partial Y}\right]}{\partial W} \circ \frac{\partial Y}{\partial W} + \frac{\partial L}{\partial Y} \frac{\partial^2 Y}{\partial W \partial W^T}$$

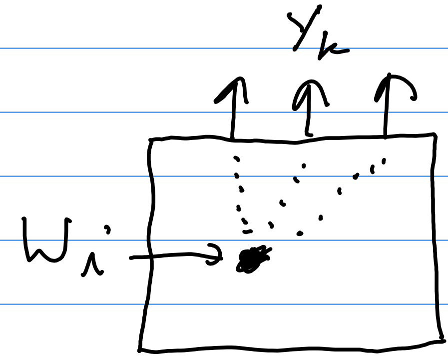
$$\frac{\partial Y^T}{\partial W} \frac{\partial \left[\frac{\partial L}{\partial Y}\right]}{\partial Y} \frac{\partial Y}{\partial W} + \underbrace{\frac{\partial L}{\partial Y} \cdot \frac{\partial^2 Y}{\partial W \partial W^T}}_{\text{ignore}} = \frac{\partial Y^T}{\partial W} \frac{\partial^2 L}{\partial Y \partial Y^T} \frac{\partial Y}{\partial W} + \cdots$$

$$\boxed{\frac{\partial^2 L}{\partial W \partial W^T} \approx \frac{\partial Y^T}{\partial W} \frac{\partial^2 L}{\partial Y \partial Y^T} \frac{\partial Y}{\partial W}}$$

$$\boxed{\frac{\partial^2 L}{\partial x \partial x^T} = \frac{\partial Y^T}{\partial x} \frac{\partial^2 L}{\partial Y \partial Y^T} \frac{\partial Y}{\partial x}}$$

Diagonal approximation

$$\frac{\partial^2 L}{\partial W_i^2} = \sum_k \frac{\partial^2 L}{\partial Y_k^2} \left(\frac{\partial Y_k}{\partial W_i}\right)^2$$



$$Y = \sum_j W_j X_j \qquad \frac{\partial Y}{\partial W_i} = X_i$$

$$\frac{\partial^2 L}{\partial W_i^2} = \frac{\partial^2 L}{\partial Y} (X_i)^2 \qquad Y_k = \sum_j W_{kj} X_j$$

$$\frac{\partial^2 L}{\partial X_i^2} = \frac{\partial^2 L}{\partial Y} (W_i)^2 \qquad \frac{\partial^2 L}{\partial X_i^2} = \sum_k \frac{\partial^2 L}{\partial Y_k} (W_{ki})^2$$

(1) take a sample

(2) fprop, bprop, bbprop

$$\tilde{Y} \qquad \frac{\partial L}{\partial W_k}. \qquad \frac{\partial^2 L}{\partial W_k^2}.$$

accumulate the $\frac{\partial^2 L}{\partial W_k^2}$ on many samples

$$h_{kk} \approx \sum_i \frac{\partial^2 L}{\partial W_k^2}(x^i, y^i)$$

$$\eta_k = \frac{c}{\max(h_{kk}, \gamma)}$$