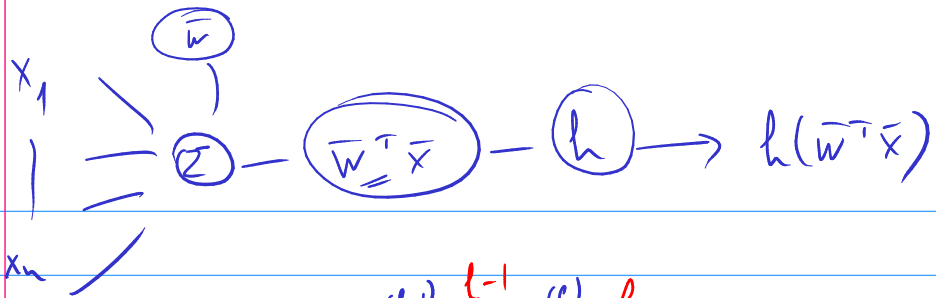
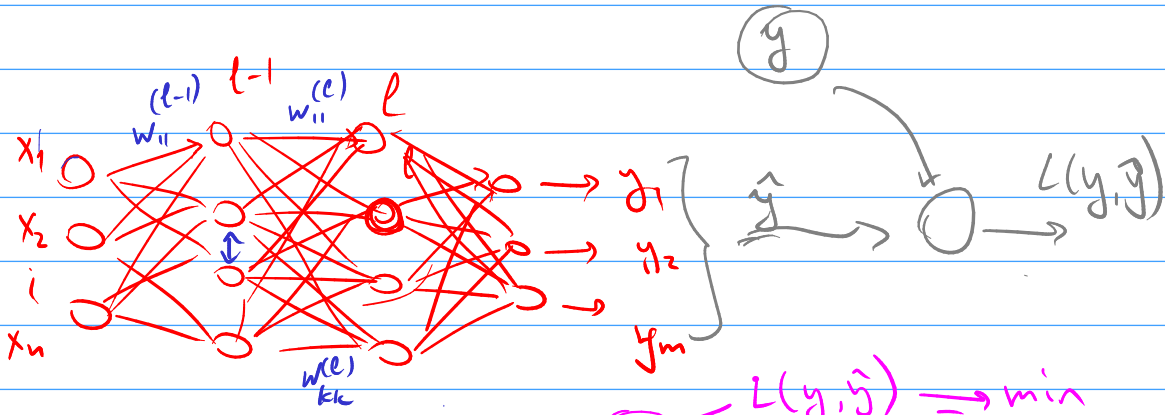
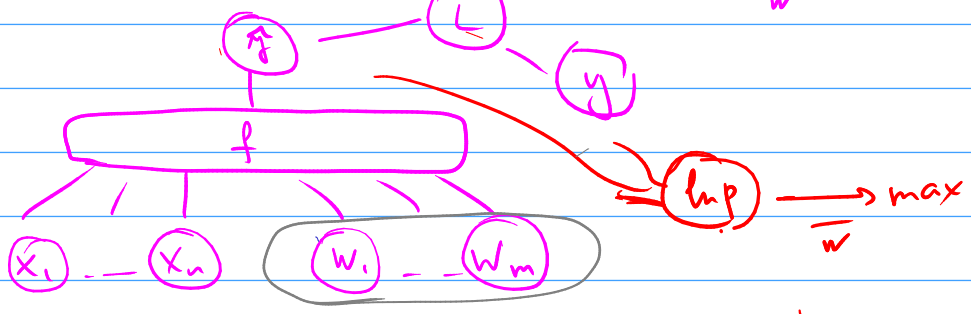




Computational graph

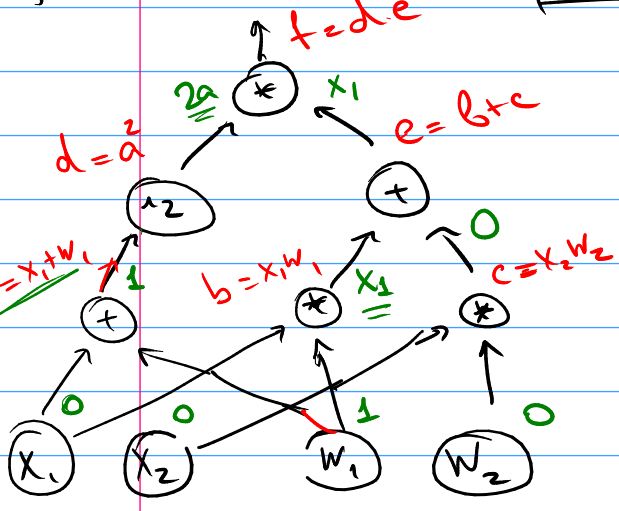


$\nabla_{\bar{w}} L$

$\bar{w} := \bar{w} - \eta \nabla_{\bar{w}} L$ → $H = \nabla_{\bar{w}}(\nabla_{\bar{w}} L)$

quasi-Newton $H = uv^T$ L-BFGS

$f = (x_1 + w_1)^2 (x_1 w_1 + x_2 w_2)$



$\nabla_{\bar{w}} f = \begin{pmatrix} \partial f/\partial w_1 \\ \partial f/\partial w_2 \end{pmatrix}$

$\frac{\partial f(g_1(x), \dots, g_k(x))}{\partial x} = \frac{\partial f}{\partial g_1}\frac{\partial g_1}{\partial x} + \dots + \frac{\partial f}{\partial g_k}\frac{\partial g_k}{\partial x}$

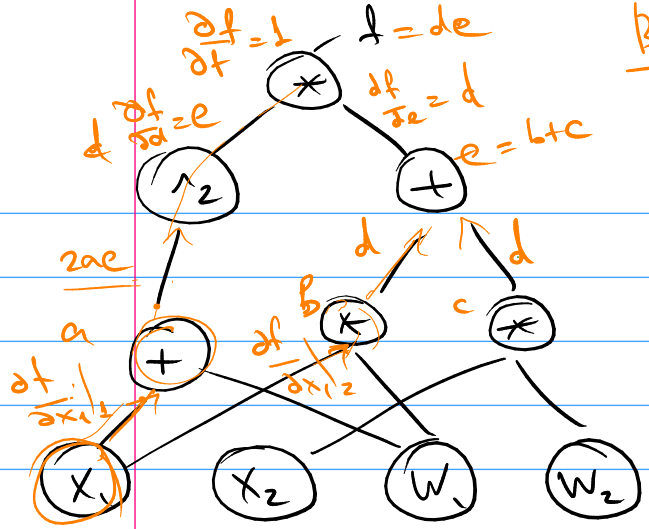
forward propagation

$\frac{\partial a}{\partial w_1} = \frac{\partial a}{\partial x_1}\frac{\partial x_1}{\partial w_1} + \frac{\partial a}{\partial w_1}\frac{\partial w_1}{\partial w_1} = 1$

$\frac{\partial b}{\partial w_1} = x_1$ $\quad \frac{\partial c}{\partial w_1} = 0$ $\quad \frac{\partial d}{\partial w_1} = \frac{\partial d}{\partial a}\frac{\partial a}{\partial w_1} = 2a$

$\frac{\partial e}{\partial w_1} = x_1$ $\quad \frac{\partial f}{\partial w_1} = d \cdot \frac{\partial e}{\partial w_1} + e \cdot \frac{\partial d}{\partial w_1} = 2ae + dx_1 =$

$= 2(x_1 + w_1)(x_1 w_1 + x_2 w_2) + (x_1 + w_1)^2 x_1$



Backpropagation

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial e}\frac{\partial e}{\partial b} = d \cdot 1 = d$$

$$\frac{\partial f}{\partial c} = d$$

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial d}\frac{\partial d}{\partial a} = 2a \cdot e$$

$$\frac{\partial f}{\partial x_1} = \frac{\partial f}{\partial x_1}\Big|_a + \frac{\partial f}{\partial x_1}\Big|_b = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x_1} + \frac{\partial f}{\partial b}\frac{\partial b}{\partial x_1} = 2ae + dw_1$$

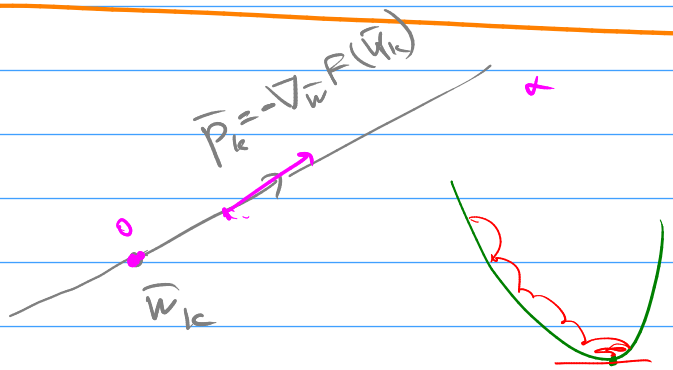
$$\frac{\partial f}{\partial w_1} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial w_1} + \frac{\partial f}{\partial b}\frac{\partial b}{\partial w_1} = 2ae + dx_1$$

$\nabla_{\bar{w}} f$

$$\frac{\partial f}{\partial w_2} = \frac{\partial f}{\partial c}\frac{\partial c}{\partial w_2} = d \cdot x_2 = (x_1+w_1)^2 \cdot x_2$$

$$F(\bar{w}) \longrightarrow \min$$

$$\bar{w}_{k+1} = \bar{w}_k - \alpha \nabla_{\bar{w}} F(\bar{w}_k)$$



$$\min_{\alpha} \underbrace{F(\bar{w}_k + \alpha \bar{p}_k)}_{\psi_k(\alpha)}$$

Armijo rule

$$\psi_k(\alpha) \le \psi_k(0) + c_1 \alpha \cdot \psi_k'(0)$$

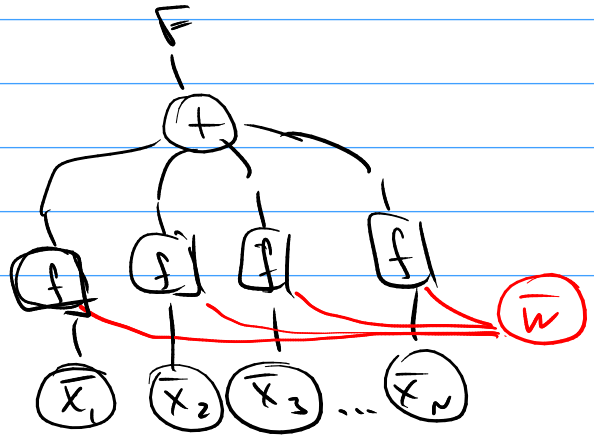
$$0 < c_1 < 1/2$$

$$|\psi_k'(\alpha)| \le c_2 |\psi_k'(0)|$$

$$F(\bar{w}; D) = \sum_{d \in D} f(\bar{w}, d)$$

$$\ln p(D|\bar{w}) = \sum_{d \in D} \ln p(d|\bar{w})$$

$$\nabla_{\bar{w}} F = \sum_d \nabla_{\bar{w}} f(\bar{w}, d)$$

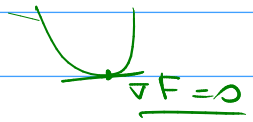


Stochastic gradient descent

mini-batch $\quad \bar{w}_{k+1} = \bar{w}_k - \alpha \nabla_{\bar{w}} F^{(k)}(\bar{w})$

Stochastic optimization $\quad F^{(k)}(\bar{w}) = \sum_{d \in Batch_k} f(\bar{w}, d)$

$$F(\bar{w}) = \mathbb{E}_{q(y)}[f(\bar{w}, \bar{y})] \xrightarrow[\bar{w}]{} \min$$



$$F(\bar{w}) = \mathbb{E}_{d \sim Unif(D)}[f(\bar{w}, d)] \xrightarrow[\bar{w}]{} \min$$

$$\Rightarrow \frac{1}{N} \cdot \sum f(\bar{w}, d)$$

$$\hat{F}(\bar{w}) = \frac{1}{m} \sum_{i=1}^{m} f(\bar{w}, d_i), \quad \text{где} \quad d_i \sim Unif(D)$$

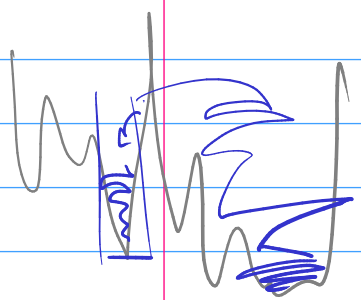
SGD $\quad F(\bar{w}) = \mathbb{E}_q[f(\bar{w}, \bar{y})]$

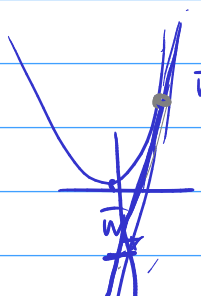
$$\bar{w}_{k+1} = \bar{w}_k - \alpha_k \hat{\bar{g}}_k, \qquad \mathbb{E}[\hat{\bar{g}}_k] = \bar{g}_k = \nabla F(\bar{w}_k)$$

$$\mathbb{E}_q\left[\|\bar{w}_{k+1} - \bar{w}_*\|^2\right] = \|\bar{w}_k - \alpha_k \hat{\bar{g}}_k - \bar{w}_*\|^2 =$$

$$= \|\bar{w}_k - \bar{w}_*\|^2 - 2\alpha_k \hat{\bar{g}}_k^T (\bar{w}_k - \bar{w}_*) + \alpha_k^2 \|\hat{\bar{g}}_k\|^2$$

$$\|\bar{w}_k - \bar{w}_*\|^2 - 2\alpha_k \bar{g}_k^T (\bar{w}_k - \bar{w}_*) + \alpha_k^2 \mathbb{E}[\|\hat{\bar{g}}_k\|^2]$$





$$f(\bar{w}_*) \geq f(\bar{w}_k) + \bar{g}_k^T (\bar{w}_* - \bar{w}_k)$$

$$f(\bar{w}_k) - f(\bar{w}_*) \leq \bar{g}_k^T (\bar{w}_k - \bar{w}_*)$$

$$\mathbb{E}\left[\alpha_k (f(\bar{w}_k) - f(\bar{w}_*))\right] \leq \alpha_k \bar{g}_k^T (\bar{w}_k - \bar{w}_*) =$$

$$= \frac{1}{2}\|\bar{w}_k - \bar{w}_*\|^2 + \frac{1}{2}\alpha_k^2 \mathbb{E}[\|\hat{\bar{g}}_k\|^2] - \frac{1}{2}\mathbb{E}[\|\bar{w}_{k+1} - \bar{w}_*\|^2]$$

$\xi_0 \quad \xi_1 + \xi_1 - \xi_2 + \ldots + \xi_k - \xi_{k+1}$

$$\sum_{i=0}^{k} \alpha_i \left(\mathbb{E}[F(\bar{w}_i)] - F(\bar{w}_*) \right) \leq \frac{1}{2}\|\bar{w}_0 - \bar{w}_*\|^2 + \frac{1}{2}\sum_{i=0}^{k} \alpha_i^2 \, \mathbb{E}\left[\|\hat{g}_i\|^2\right] - \cancel{\frac{1}{2}\mathbb{E}\left[\|\bar{w}_{k+1} - \bar{w}_*\|^2\right]}$$

$$\mathbb{E}\left[\frac{\sum_{i=0}^{k} \alpha_i F(\bar{w}_i)}{\sum \alpha_i}\right] \geq \mathbb{E}\left[F\left(\frac{\sum \alpha_i F(\bar{w}_i)}{\sum \alpha_i}\right)\right]$$

$$\mathbb{E}\left[F\left(\frac{\sum \alpha_i F(\bar{w}_i)}{\sum \alpha_i}\right) - F(\bar{w}_*)\right] \leq \frac{\frac{1}{2}\|\bar{w}_0 - \bar{w}_*\|^2 + \frac{1}{2}\sum_{i=0}^{k} \alpha_i^2 \, \mathbb{E}\left[\|\hat{g}_i\|^2\right]}{\sum_{i=0}^{k} \alpha_i}$$

$$\|\bar{w}_0 - \bar{w}_*\| \leq R$$

$$\mathbb{E}\left[\|\hat{g}_i\|^2\right] \leq G^2$$

$$\mathbb{E}\left[F(\bar{v}_k) - F(\bar{w}_*)\right] \leq \frac{R^2 + G^2 \sum_{0}^{k} \alpha_i^2}{2\sum_{0}^{k} \alpha_i} \longrightarrow 0 \quad (k \to \infty)$$

$\sum_{0}^{k} \alpha_i^2 = h^2 \cdot (k+1)$

$\sum_{0}^{k} \alpha_i = h(k+1)$

$\alpha_i = h$

$$\sum_{0}^{\infty} \alpha_i = \infty$$

$$\sum_{0}^{\infty} \alpha_i^2 < \infty$$

$\alpha_i = \frac{1}{i}$

$$\mathbb{E}\left[F(\bar{v}_k) - F(\bar{w}_*)\right] \leq \frac{R^2}{2h(k+1)} + \frac{1}{2}G^2 h \longrightarrow \frac{1}{2}G^2 h \quad (k \to \infty)$$