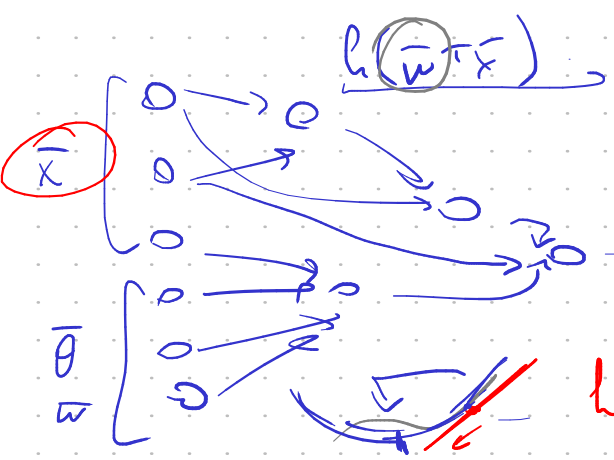


$$-\log p(\bar{x}|\bar{w}) - \log p(\bar{w}) \to f(\bar{x}, \bar{w})$$

$$L(\bar{w}) = \sum_{\bar{x} \in D} \left[\log p(\bar{x}|\bar{w}) + \log p(\bar{w}) \right]$$

$$= N \cdot \mathbb{E}_{\bar{x} \sim \text{Unif}(D)} \left[\log p(\bar{x}|\bar{w}) + \log p(\bar{w}) \right]$$

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

$$\bar{w}^{(0)}, \quad \bar{w}^{(k+1)} := \bar{w}^{(k)} - \eta \cdot \nabla_{\bar{w}} F(\bar{w}^{(k)}) \quad \Big\} \text{GD} - \text{gradient descent}$$

$\bar{g}_k$

$$F(\bar{w}) \approx \bar{w}^{(k)} + (\bar{w} - \bar{w}^{(k)}) \cdot \nabla_{\bar{w}} F(\bar{w}^{(k)}) + \frac{1}{2} (\bar{w} - \bar{w}^{(k)})^T H_k (\bar{w} - \bar{w}^{(k)})$$

$$\left(\partial^2 F / \partial w_i \partial w_j \right)$$

$$-\bar{g}_k + H_k \bar{w} - H_k \bar{w}^{(k)} = 0$$

$$\bar{w}^{(k+1)} = \bar{w}^{(k)} + H_k^{-1} \bar{g}_k$$

quasi-Newton algorithms — L-BFGS — $\bar{g}_k \to$ low rank $\approx H^{-1}$

SGD – stochastic gradient descent

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

$$\begin{cases} q(\bar{y}) = \text{Unif}(D) \\ q(\bar{y}) \neq p(\bar{y}|D) \end{cases}$$

$$G(\bar{x}) = \nabla_{\bar{x}} F(\bar{x}) = \mathbb{E}_{q(\bar{y})} \nabla_{\bar{x}} f(\bar{x}, \bar{y})$$

$$\hat{F}(\bar{x}) = \frac{1}{m} \sum_{i=1}^{m} f(\bar{x}, \bar{y}_i)$$

$$\hat{g}(\bar{x}) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\bar{x}} f(\bar{x}, \bar{y}_i) \quad \text{где } \bar{y}_i \sim q(\bar{y})$$

$$\mathbb{E}\, \hat{g}_k = \bar{g}_k$$

$$\bar{x}_{k+1} = \bar{x}_k - \alpha_k \cdot \hat{\bar{g}}_k \quad - \text{SGD}$$

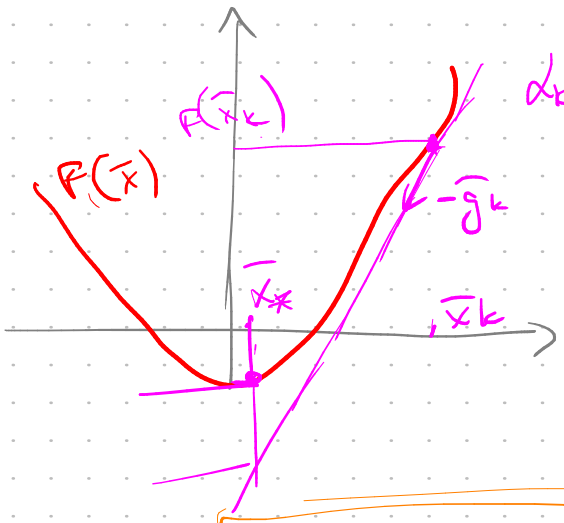
GD $\left[F(\bar{x}) \to \min \right.$ $\bar{g}_k$ $F(\bar{x}_k - \alpha \bar{g}_k) \xrightarrow{\alpha} \min$ (?) $\left. \right]$

$\bar{x}_*$: сходится

$$\|\bar{x}_{k+1} - \bar{x}_*\|^2 = \|\bar{x}_k - \bar{x}_* - \alpha_k \hat{\bar{g}}_k\|^2 = \|\bar{x}_k - \bar{x}_*\|^2 - 2\alpha_k \hat{\bar{g}}_k^T (\bar{x}_k - \bar{x}_*) + \alpha_k^2 \|\hat{\bar{g}}_k\|^2$$

$$\geq \alpha_k (F(\bar{x}_k) - F(\bar{x}_*))$$

$$\mathbb{E} \|\bar{x}_{k+1} - \bar{x}_*\|^2 = \mathbb{E} \|\bar{x}_k - \bar{x}_*\|^2 + 2\alpha_k \bar{g}_k^T (\bar{x}_* - \bar{x}_k) + \alpha_k^2 \mathbb{E} \|\hat{\bar{g}}_k\|^2$$



$$\alpha_k \cdot F(\bar{x}_*) \geq \alpha_k \left(F(\bar{x}_k) - (\bar{x}_* - \bar{x}_k)^T \bar{g}_k \right)$$

$$\sum_k \left(\mathbb{E}\|\bar{x}_{k+1} - \bar{x}_*\|^2 - \mathbb{E}\|\bar{x}_k - \bar{x}_*\|^2 \leq \right.$$

$$\left. \leq \alpha_k^2 \mathbb{E}\|\hat{\bar{g}}_k\|^2 - 2\alpha_k (F(\bar{x}_k) - F(\bar{x}_*)) \right)$$

$$\sum_{i=0}^{k} 2\alpha_i (F(\bar{x}_i) - F(\bar{x}_*)) \leq \sum_{i=0}^{k} \alpha_i^2 \mathbb{E}\|\hat{g}_i\|^2 + \mathbb{E}\|\bar{x}_0 - \bar{x}_*\|^2 - \mathbb{E}\|\bar{x}_{k+1} - \bar{x}_*\|^2$$

$$F\left(\frac{\sum \alpha_i \bar{x}_i}{\sum \alpha_i} \right) \leq \frac{\sum \alpha_i F(\bar{x}_i)}{\sum \alpha_i}$$

$$F(\bar{x}_k') - F(\bar{x}_*) \leq \frac{\sum \alpha_i (F(\bar{x}_i) - F(\bar{x}_*))}{\sum \alpha_i}$$

$$\mathbb{E} F(\bar{x}_k') - F(\bar{x}_*) \leq \frac{\mathbb{E}\|\bar{x}_0 - \bar{x}_*\|^2 + \sum_{i=0}^{k} \alpha_i^2 \mathbb{E}\|\hat{\bar{g}}_i\|^2}{2 \cdot \sum_{i=0}^{k} \alpha_i}$$

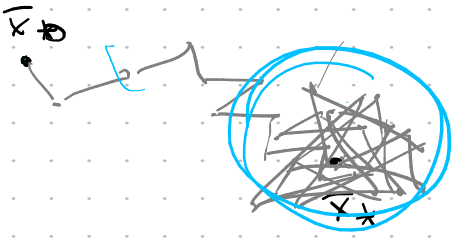
$$\|\bar{x}_0 - \bar{x}_*\| \leq R, \quad \mathbb{E}\|\bar{g}_k\|^2 \leq G^2$$

$$\mathbb{E} F(\bar{x}_k') - F(\bar{x}_*) \leq \frac{R^2 + G^2 \cdot \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i} \to 0, \quad k \to \infty$$

$\sum \alpha_i^2 < \infty$, $\sum \alpha_i = \infty$, $\alpha_i = \frac{1}{i}$

$$\alpha_i = \alpha \quad \leq \frac{R^2 + G^2 \cdot \alpha^2 \cdot k}{2 \cdot \alpha k} = \frac{R^2}{2\alpha k} + \frac{G^2 \cdot \alpha}{2}$$

$k \to \infty$, 0



Вариации SGD

$\bar{x}_t$ $\bar{u}_t = -\alpha \bar{g}_t$ $\bar{x}_{t+1} = \bar{x}_t - \alpha \cdot \bar{g}_t$ $\bar{u}_t = \bar{x}_{t+1} - \bar{x}_t$

1) SGD w/ momentum

$\gamma = 0.99, 0.995$

$\gamma^2 \cdot \bar{u}_{t-2} - \gamma \cdot \alpha \cdot \bar{g}_{t-1}$

$\bar{x}_{t-1}$ $\bar{u}_{t-1}$ $\bar{x}_t$ $-\alpha \cdot \bar{g}_t$ $\gamma \cdot \bar{u}_{t-1}$

$\bar{u}_t = \gamma \cdot \bar{u}_{t-1} - \alpha \cdot \bar{g}_t$

$\bar{x}_{t+1} = \bar{x}_t + \bar{u}_t$

2) Nesterov accelerated gradient (NAG)

$\bar{x}_{t-1}$ $\bar{u}_{t-1}$ $\bar{x}_t$ $\gamma \cdot \bar{u}_{t-1}$

$\bar{x}_{t+1} = \bar{x}_t + \gamma \bar{u}_{t-1} - \alpha \cdot \bar{g}'_t$

$-\alpha \bar{g}'_t = -\alpha \cdot \nabla_{\bar{x}} f(\bar{x}_t + \gamma \cdot \bar{u}_{t-1})$

3) Adaptive SGD

- **Adagrad** $G_{0,i} = 0$

$$G_{t,i} = G_{t-1,i} + g_{t,i}^2$$

$$\underbrace{x_{t+1,i}}_{c} = \underbrace{x_{t,i}}_{c} - \frac{\alpha}{\sqrt{G_{t,i} + \varepsilon}} \cdot \underbrace{g_{t,i}}_{м/с}$$

$\frac{1}{м/с}$

- **RMSProp**

$$G_{t,i} = G_{t-1,i} \cdot \gamma + (1-\gamma) \cdot g_{t,i}^2$$

- **Adadelta** $\bar{x}$ – секунды, $f(\bar{x})$ – метры, $\bar{g}$ – $м/с$

$$\bar{x} := \bar{x} - \alpha \bar{g}_{м/с}$$

$$\bar{x}_{t+1} = \bar{x}_t - \frac{1}{2} H^{-1} \bar{g}_t$$

$$H^{-1} \sim (м/с^2)^{-1}$$

$$x_{t+1,i} = x_{t,i} - \alpha \cdot \frac{\sqrt{R_{t,i} + \varepsilon}}{\sqrt{G_{t,i} + \varepsilon}} \cdot g_{t,i}, \quad R_{t,i} = \gamma \cdot R_{t-1,i} + (1-\gamma) u_{t,i}^2$$

- Adam

$$G_{ti} = \beta_2 G_{t-1,i} + (1-\beta_2) g_{ti}^2$$

$$x_{t+1,i} = x_{ti} - \frac{\alpha}{\sqrt{G_{ti} + \varepsilon}} \cdot m_{ti}, \text{ где } m_{ti} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{ti}$$

$\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-7,8}$

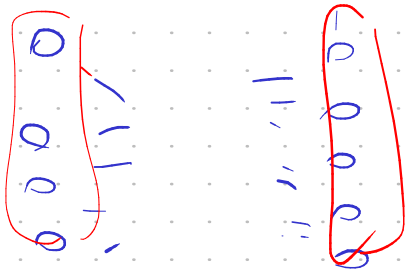
Weight decay

L_2: $L(\bar{w}) + \alpha \|\bar{w}\|^2$

weight decay: $\bar{w} := (1-\alpha)\cdot \bar{w} - \alpha \cdot \nabla_{\bar{w}} L(\bar{w}) =$

$= \bar{w} - \alpha \cdot \nabla_{\bar{w}} \left(L(\bar{w}) + \frac{\gamma}{2\alpha} \cdot \|\bar{w}\|^2 \right)$

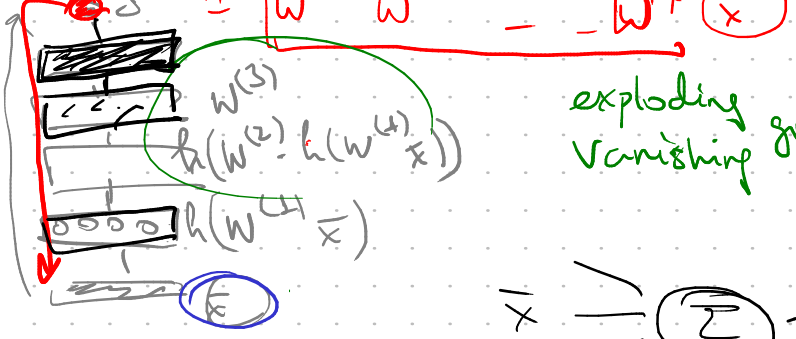
$W^{(k)} \cdot A$ $\quad$ $A^{-1} \cdot W^{(k+1)}$



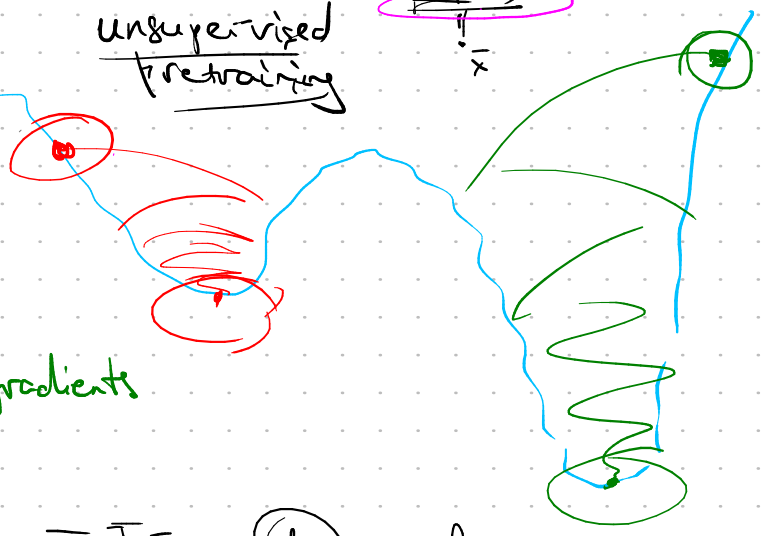
unsupervised pretraining

Weight initialization

$L(\bar{y})$ $\quad \bar{y} = W^{(k)} W^{(k-1)} \dots W^{(1)} \bar{x}$



exploding / vanishing gradients



$x = \sum \Sigma - \bar{w}^T \bar{x} - h - h(\bar{w}^T \bar{x})$

$y = \sum_i^n w_i x_i = \sum_i^n y_i$

$\mathrm{Var}[y_i] = \mathrm{Var}[w_i x_i] =$

$= E[w_i^2 x_i^2] - E[w_i x_i]^2 = E[x_i]^2 \cdot \mathrm{Var}[w_i] + E[w_i]^2 \mathrm{Var}[x_i] +$

$+ \mathrm{Var}[w_i] \cdot \mathrm{Var}[x_i]$

$\mathrm{Var}[y_i] = \mathrm{Var}[w_i] \cdot \mathrm{Var}[x_i]$

$\mathrm{Var}[y] = n \cdot \mathrm{Var}[w_i] \cdot \mathrm{Var}[x_i]$

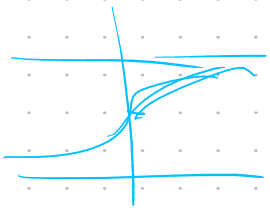
$\approx 1 = O(1/n)$

Xavier init

$w_i \sim \mathrm{Unif}\left(\left[-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}} \right] \right)$

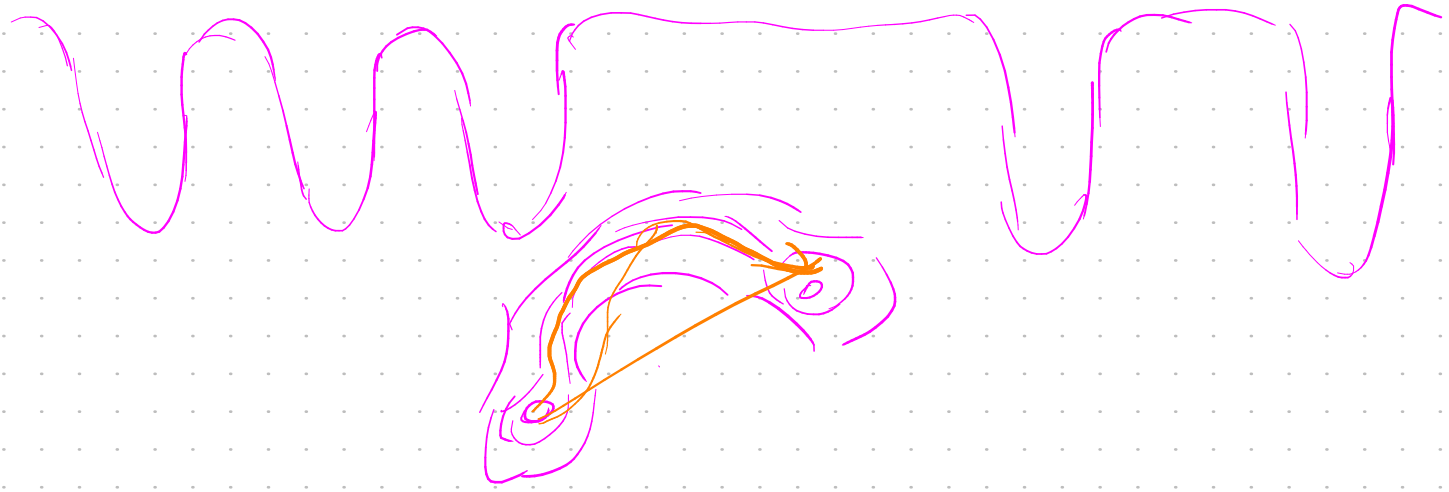
$\mathrm{Var}\{\mathrm{unif}([a,b])\} = (b-a)^2/12$

$$Var\left[Unif\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]\right] = \frac{(2/\sqrt{n})^2}{12} = \frac{1}{3n}$$



$Var[w_i] = 2/n$ - ReLU

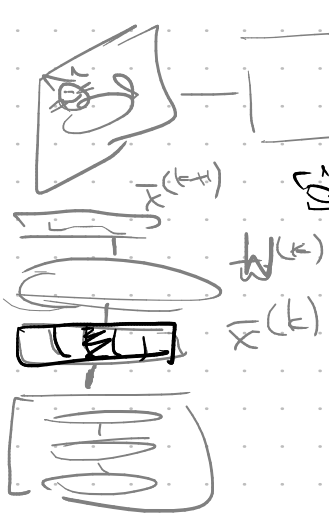
He init



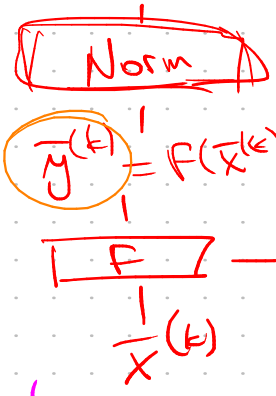
Batch normalization

Covariate shift

Internal covariate shift

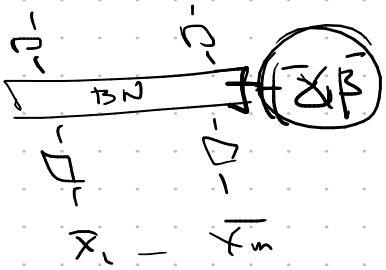


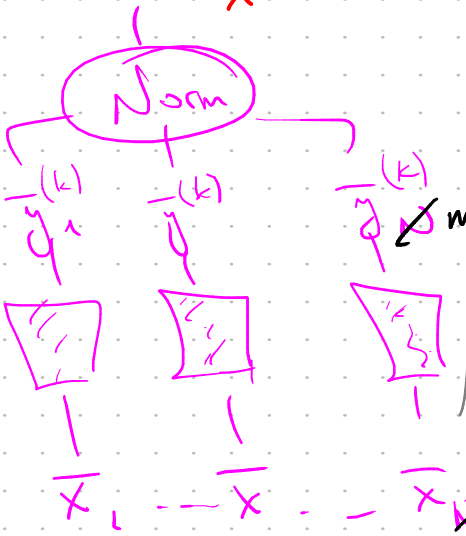
$$\bar{x}^{(k+1)} = Norm(\bar{y}^{(k)})$$



Layer normalization

$$\bar{x}^{(k+1)} = \frac{\bar{y}^{(k)} - E_D[\bar{y}^{(k)}]}{\sqrt{Var_D[\bar{y}^{(k)}]}}$$





$$BN(\bar{y}^{(k)}) = \frac{\bar{y}^{(k)} - E_m[\bar{y}^{(k)}]}{\sqrt{Var_m[\bar{y}^{(k)}]}} \odot \bar{\gamma} + \bar{\beta}$$