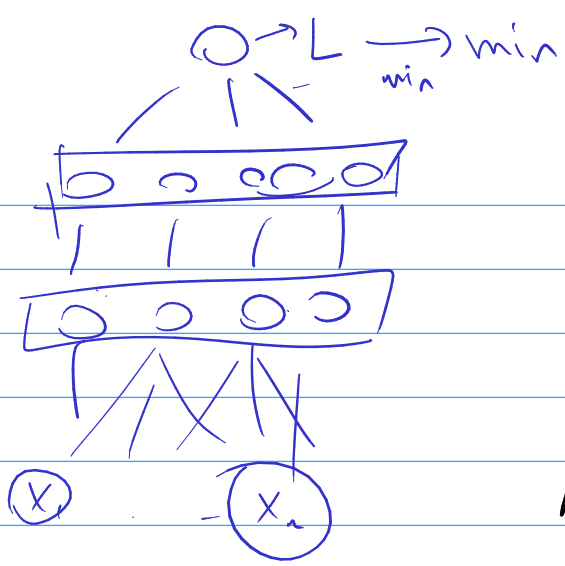
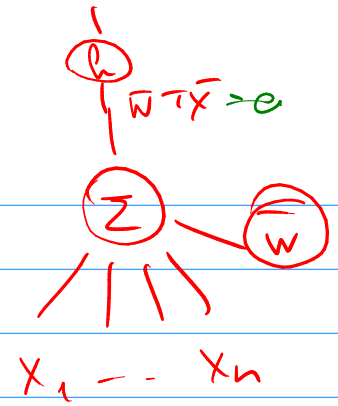




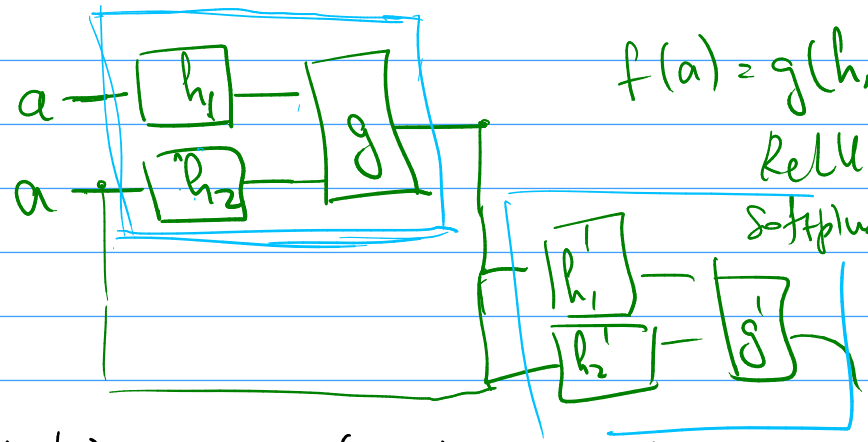
$h(a)$

$h: \mathbb{R} \to \mathbb{R}$

Neural architecture search

Quoc Le

2017



$f(a) = g(h_1(a), h_2(a))$

$ReLU = \max(a, 0)$

$Softplus = \log(1 + e^x)$

$Swish(a) = a \cdot \sigma(\beta \cdot a) = \dfrac{a}{1 + e^{-\beta a}}$

Hard

$\max(\cdot_1, \cdots, \cdot)$

Soft

$S_\beta(x_1 \dots x_n) = \dfrac{\sum x_i \cdot e^{\beta x_i}}{\sum e^{\beta x_i}}$

$S_\beta(g(x), h(x)) = g \cdot \dfrac{e^{\beta g}}{e^{\beta g} + e^{\beta h}} + h \cdot \dfrac{e^{\beta h}}{e^{\beta g} + e^{\beta h}} = \sigma(-x) = 1 - \sigma(x)$

$\sigma(x) = \dfrac{1}{1 + e^{-x}}$

$= g \cdot \sigma(\beta(g - h)) + h \cdot \sigma(\beta(h - g))$

$= (g(x) - h(x))\,\sigma(\beta \cdot (g(x) - h(x)) + h(x)$

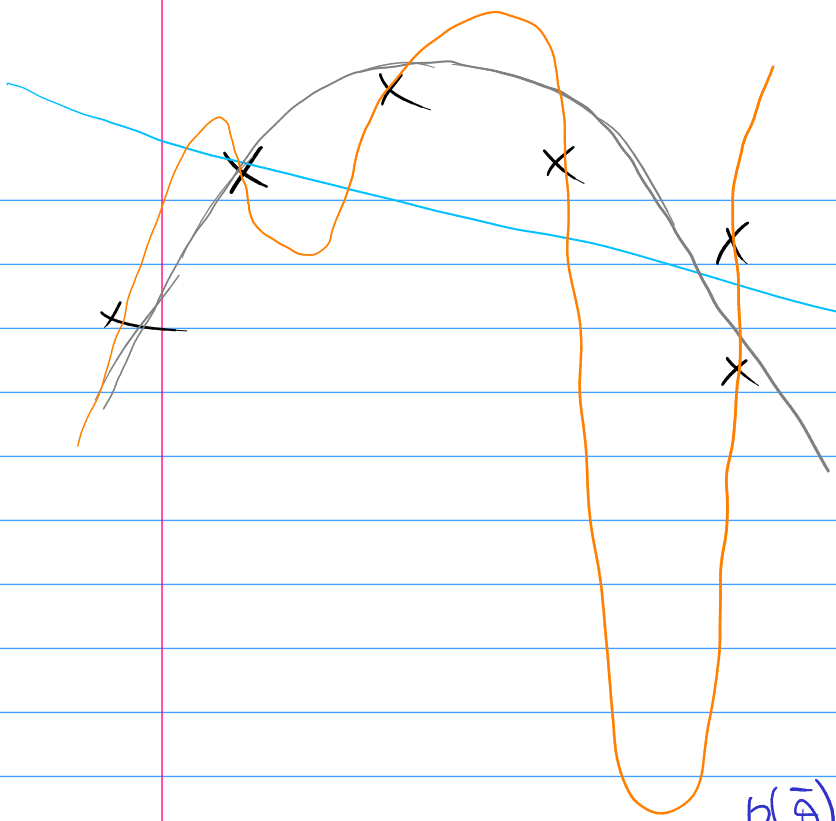
$S_\beta(x, 0) = x \cdot \sigma(\beta x)$

$S_\beta(x, ax) = (1 - a)x \cdot \sigma(\beta(1 - a)x) + ax$

$ACON = S_\beta(a_1 x, a_2 x)$

$ReLU(x) = \max(0, x)$

$LReLU(x) = \max(x, ax)$



параметры данные

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}$$

posterior — prior — likelihood

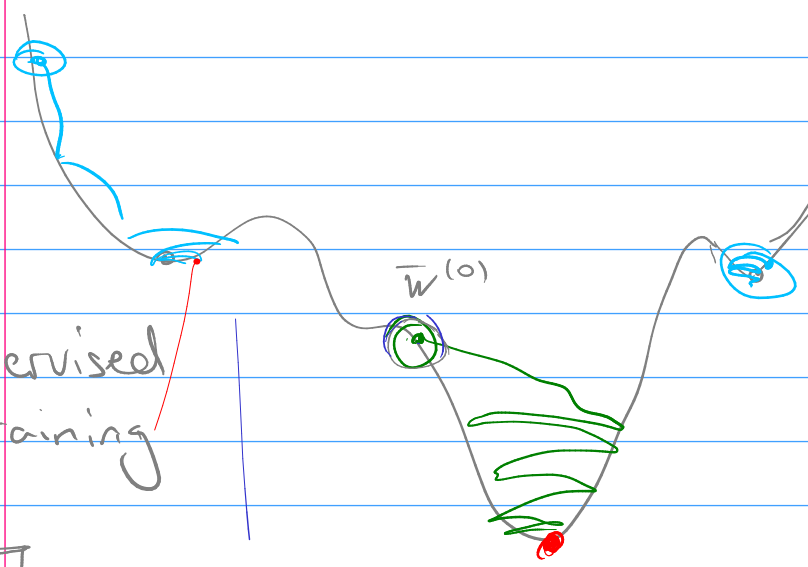
$$\log p(\theta|D) = \text{const} + \log p(D|\theta) + \log p(\theta)$$

regularizer

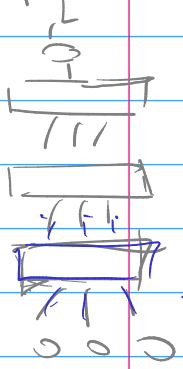
$$p(\bar{\theta}) = \mathcal{N}(\bar{\theta}|\bar{0}, \lambda \cdot I) \qquad -\frac{\lambda}{2}\|\bar{\theta}\|_2^2$$

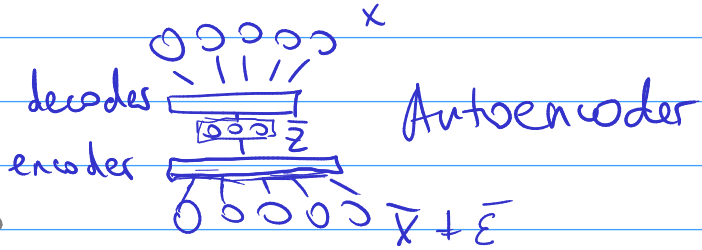
L_2-регул.

$$L(\bar{w}) \to \min_{\bar{w}}$$

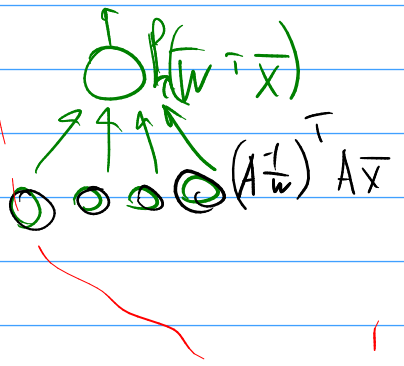


Unsupervised pretraining

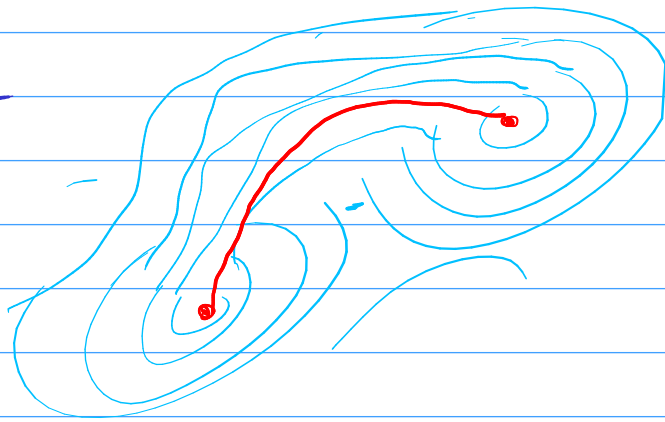


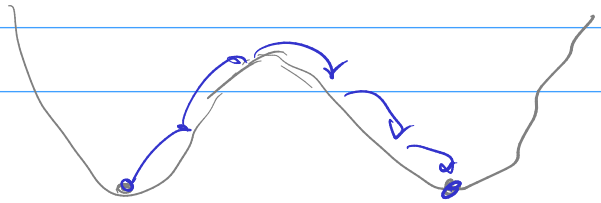


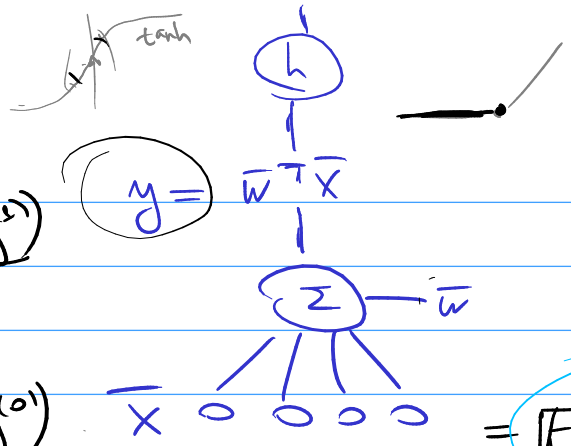
Autoencoder



А. Вепров







$$y = \bar{w}^T \bar{x} = \sum \underbrace{w_i x_i}_{y_i} = \sum y_i$$

$x^{(2)} = h(y^{(2)})$

$x^{(1)} = h(y^{(0)})$

$\bar{x} = \bar{x}^{(0)}$

$$Var[y_i] = Var[w_i x_i] =$$

$$= \mathbb{E}[w_i^2 x_i^2] - (\mathbb{E}[w_i x_i])^2$$

$$= \mathbb{E}[x_i]^2 Var[w_i] + \mathbb{E}[w_i]^2 Var[x_i] + Var[x_i] Var[w_i]$$

$$Var[y_i] = Var[x_i] Var[w_i]$$

$$Var[y] = \sum Var[y_i] = n \cdot Var[w_0] \cdot Var[x_i] \quad (\approx \tfrac{1}{3})$$

$w_i \sim \mathcal{N}(w_i | 0, \frac{1}{n})$

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

2010 Xavier Glorot

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

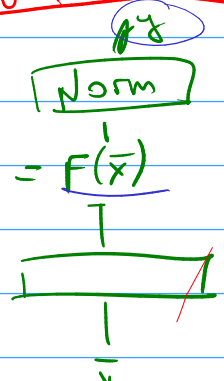
Xavier init

$$Var w_i = \frac{1}{3n}$$

$$Var[y_i] = \mathbb{E}[x_i^2] Var[w_i] + Var[x_i] Var[w_i]$$

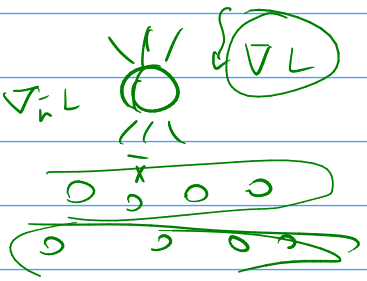
Kaiming He 2014

He init

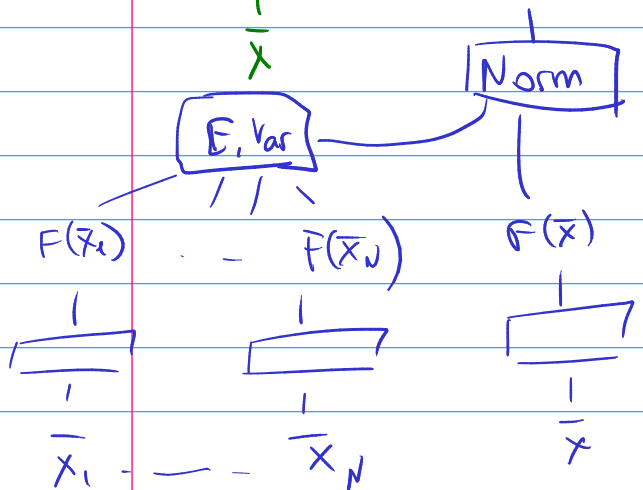


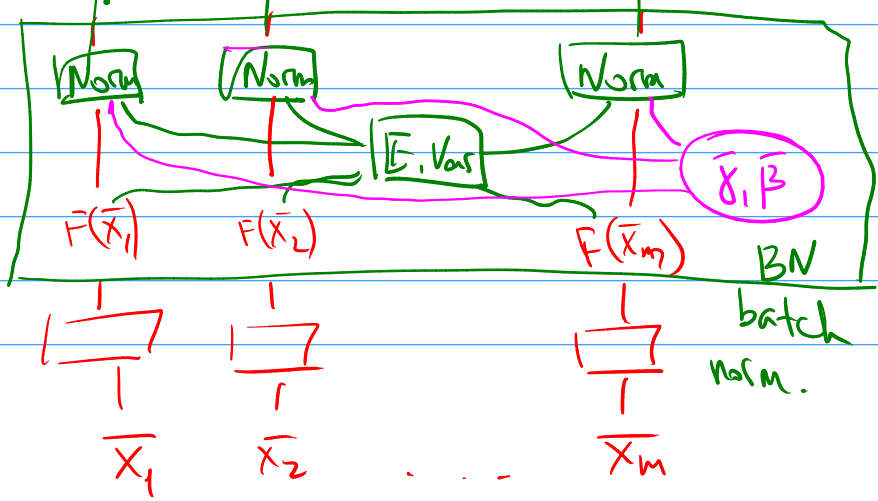
$$y_k = \left(\frac{z_k - \mathbb{E} z_k}{Var[z]_k}\right) \gamma_k + \beta_k$$

$F(\bar{x}) \to F(\bar{x}) + \varepsilon$

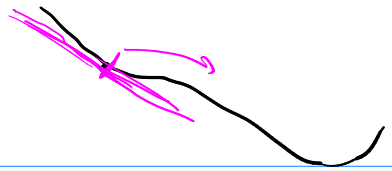


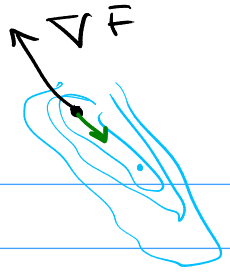
Mini-batches $\bar{x}_1, \ldots, \bar{x}_m$





$$F(\bar{x}) \to \min$$

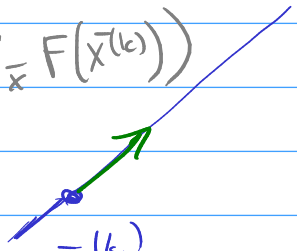




$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \alpha_k \cdot \nabla_{\bar{x}} f$$

$$\alpha_k = \arg\min F\left(\bar{x}^{(k)} - \alpha_k \nabla_{\bar{x}} F(\bar{x}^{(k)})\right)$$

Armijo, Wolfe



$$F(x,y) = x^2 + py^2$$

quasi-Newton

L-BFGS

holy grail

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \alpha_k \left(H_k^{-1} \nabla_{\bar{x}} F\right)$$

GD:

$$\nabla_{\bar{w}} L = \nabla_{\bar{w}} \left[\frac{1}{N} \sum_{n=1}^{N} \ell(\bar{x}_n, y_n, \bar{w}) \right] = \frac{1}{N} \sum_n \nabla_{\bar{w}} \ell(\bar{x}_n, y_n, \bar{w})$$

SGD:

- $\bar{x}_1, \dots, \bar{x}_m \sim D$
- $\bar{w}^{(k+1)} = \bar{w}^{(k)} - \alpha_k \cdot \nabla L_m(\bar{w})$

$$F(\bar{w}) = \frac{1}{N} \sum_1^N f(\bar{x}_n, \bar{w}) = \mathbb{E}\left[f(\bar{x}, \bar{w}) \right] \quad \bar{x} \sim \text{Unif}[D]$$

$$\frac{1}{m} \sum_{i=1}^{m} \nabla_{\bar{w}} \ell(\bar{x}_i, y_i, \bar{w})$$

$$\hat{F}_k(\bar{w}) = \frac{1}{m} \sum_1^m f(\bar{x}_i, \bar{w}) \approx \mathbb{E}\left[f(\bar{x}, \bar{w}) \right]$$

$$\nabla_{\bar{w}} \hat{F}_k(\bar{w}) \approx \nabla_{\bar{w}} F(\bar{w})$$