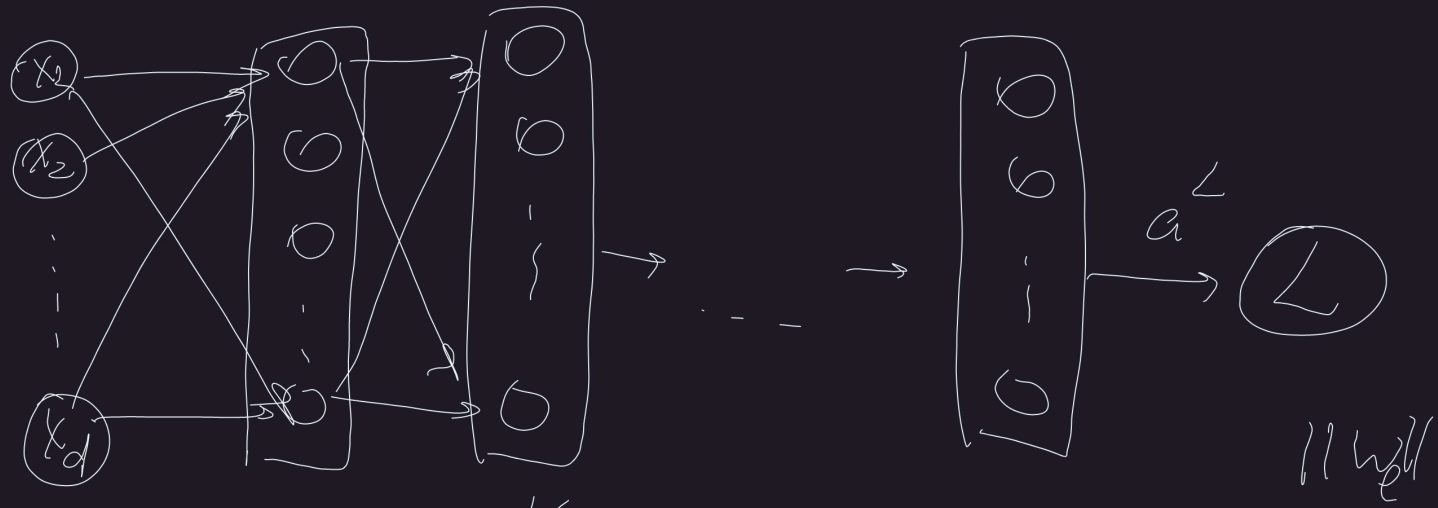


$$F(W) = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, a^L(x_i, W)\right) \longrightarrow \min_W$$

GD: Init W_0

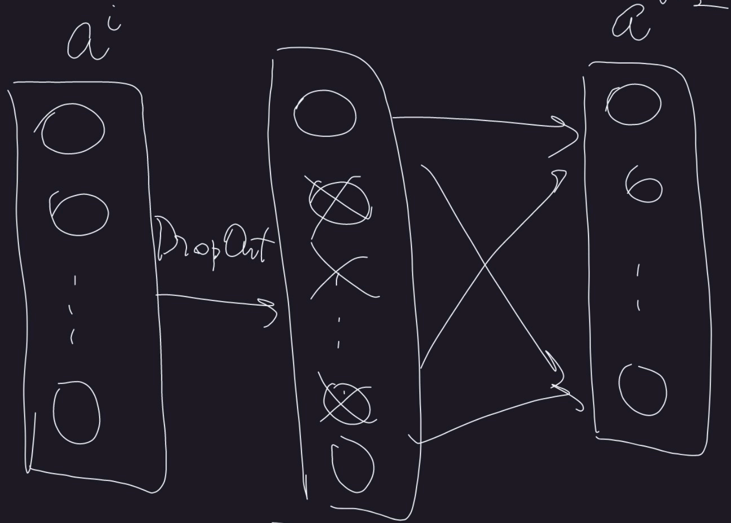
until convergence: $W_{k+1} = W_k - \alpha_t \nabla F(W_k)$

Regularization

$$F(W) = \underbrace{\frac{1}{N}\sum_{i=1}^{N} L(y_i, a^L(x_i, W))}_{L_{data}(W)} + \lambda \sum_{\ell=1}^{L} \|W_\ell\|_F^2 \to \min_W$$

$$f(x) + \sum_i \lambda_i g_i(x) \to \min_x \iff \begin{cases} f(x) \to \min_x \\ g_i(x) \le ?_i \quad \forall i \end{cases}$$

Drop Out



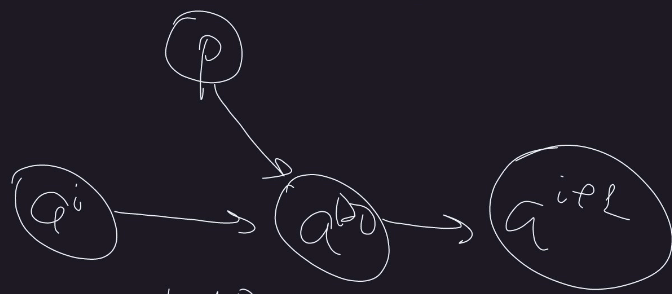
y_j is diff. for diff. mini-batches

$$\mathbb{E}_{y_j \sim Bern}\, a_j^{DO} = \mathbb{E}\, y_j\, a_j^i = a_j^i (1-p)$$

$$a_j^{DO} = y_j \cdot a_j^i$$

$y_j \sim Bern(y \mid 1-p)$:

0	1
p	1-p



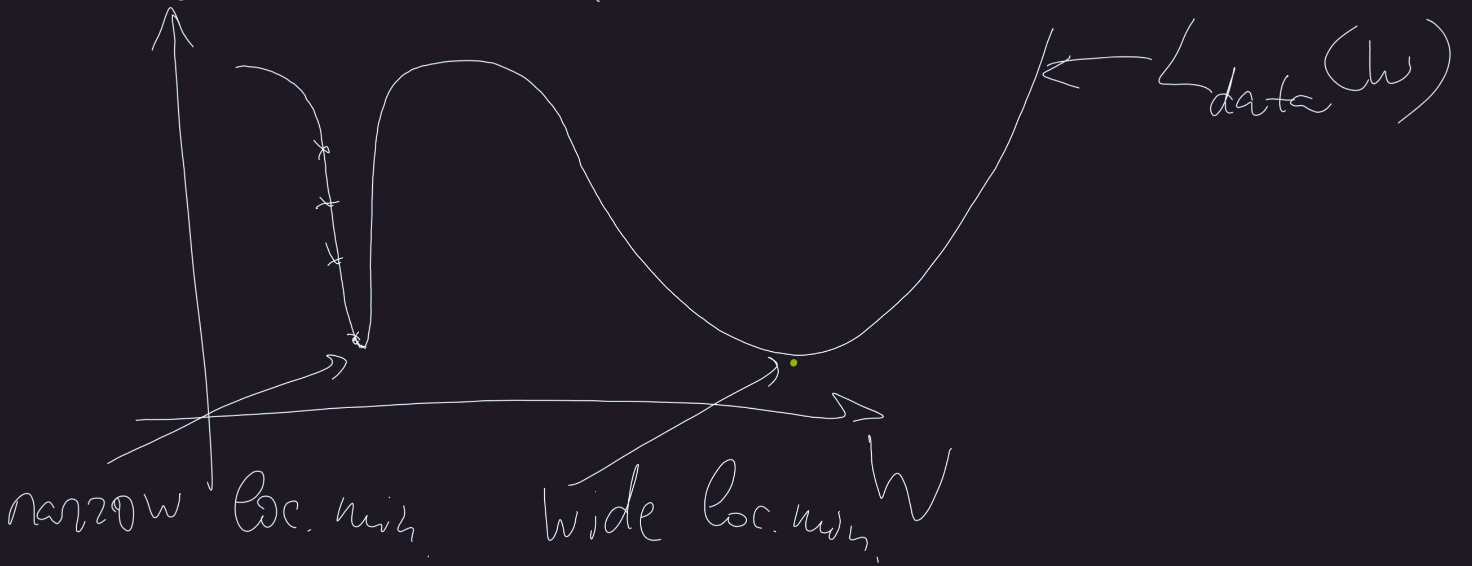
inverted DO

$$a_j = \frac{1}{1-p}\, y_j \cdot a_j^i$$

$$\mathbb{E}\, a_j^{invDO} = a_j^i$$

Gradient Penalty reg. or R_1 reg.

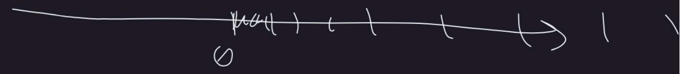
$$F(w) = L_{data}(w) + \lambda \|\nabla_w L_{data}(w)\|_2^2 \to \min_w$$



Batch Normalization

$$fl(x) = s \cdot M \cdot 2^{E}$$

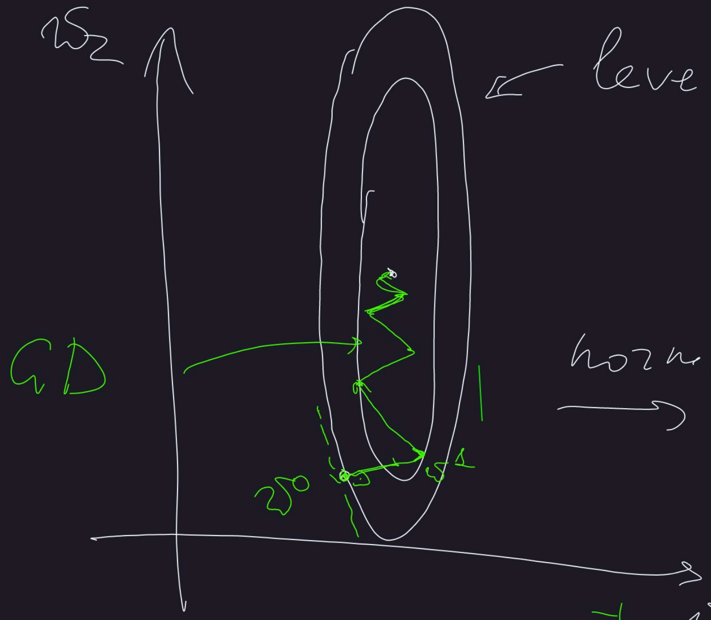
$$\{x_1, \dots, x_N\} \xrightarrow{\text{norm.}} \{\hat{x}_1, \dots, \hat{x}_N\}$$



$$\hat{x}_{ij} = \frac{x_{ij} - m_j}{\sqrt{\sigma_j^2 + \varepsilon}}, \qquad m_j = \frac{1}{N} \sum_{i=1}^{N} x_{ij}$$

$$\sigma_j^2 = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} - m_j)^2$$

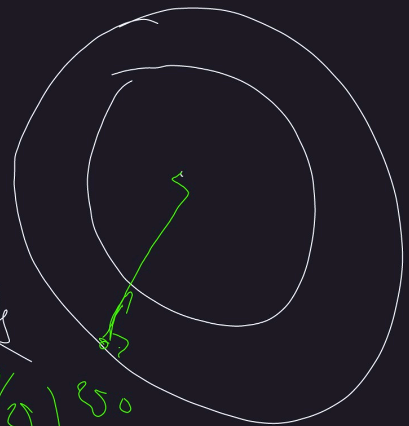
$$w^T x = \sum_{j=1}^{d} w_j x_j = \sum_{j} \hat{w}_j \hat{x}_j$$



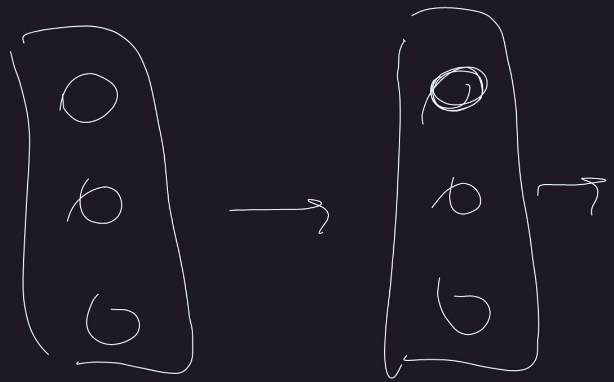
level set for loss

$\lambda \|w\|^2 = \lambda (w_1^2 + w_2^2)$

norm.



$$w_{k+1} = w_k - \alpha_k \left(\nabla^2 F(w_k)\right)^{-1} \nabla F(w_k)$$



$$m_j = \frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} x_{ij}$$

$$\sigma_j^2 = \frac{1}{N_{batch}} \sum_i \left(x_{ij} - m_j\right)^2$$

$$y_{ij} = \frac{x_{ij} - m_j}{\sqrt{\sigma_j^2 + \varepsilon}}; \quad \hat{x}_{ij} = \gamma_j y_{ij} + \delta_j$$

$$\{\gamma_j, \delta_j\}_{j=1}^{d}$$

$$\{x_{ij}\}_{i,j=1}^{N_{batch}, d} \longrightarrow \boxed{BN} \longrightarrow \{\hat{x}_{ij}\}_{i,j=1}^{N_{batch}, d}$$

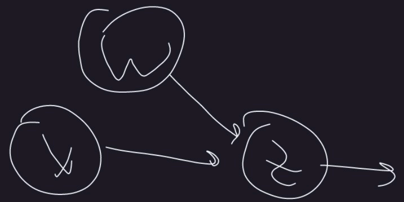
$$\gamma_j = \sqrt{\sigma_j^2 + \varepsilon}$$

$$\delta_j = m_j$$

$$m_j^{new} = \alpha m_j + (1-\alpha) m_j^{prev}$$

Weight init.

$x, \quad z = Wx$



$\nabla_z L$

$$(\nabla_x L)_j = \sum_{i=1}^{n_{outputs}} W_{ij} (\nabla_z L)_i$$

$$dL = \nabla_z L^T dz =$$

$$= \underbrace{\nabla_z L^T W}_{\nabla_x L^T} dx$$

$$\nabla_x L = W^T \nabla_z L$$

$$Var(W_{ij}) = \frac{1}{n_{outputs}}$$

$$Var((\nabla_x L)_j) = n_{outputs} Var(W_{ij}) Var((\nabla_z L)_i)$$

Kaiming He

Xavier Glorot

$$z = Wx$$

$$z_i = \sum_j W_{ij} x_j$$

$$Var(W_{ij}) = \frac{1}{n_{inputs}}$$

$$x_j \sim N(x_j \mid 0, 1)$$

$$Var(W_{ij}) = \frac{2}{n_{inputs} + n_{outputs}}$$

$$W_{ij} \sim N(W_{ij} \mid 0, Var(W_{ij}))$$

$$\mathbb{E} z_i = \mathbb{E} \sum_j W_{ij} x_j = \sum_j \mathbb{E} W_{ij} \cdot \underbrace{\mathbb{E} x_j}_{=0} = 0$$

$$Var(z_i) = Var\left(\sum_j W_{ij} x_j\right) = \sum_{j=1}^{n_{input}} Var(W_{ij}) \cdot \underbrace{Var(x_j)}_{=1} = n_{inputs} \cdot Var(W_{ij})$$

$$W_{ij} \sim R\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]$$

$$Var(W_{ij}) = \frac{1}{3n} = \frac{2}{n_{inputs} + n_{outputs}}$$

$$\Rightarrow n = \frac{n_{inputs} + n_{outputs}}{6} \; ; \; W_{ij} \sim R\left[-\sqrt{\frac{6}{n_{inputs} + n_{outputs}}}, \sqrt{\frac{6}{\cdots}}\right]$$

Orth. init.

$W \in \mathbb{R}^{n \times d}$

$W^T W = I$

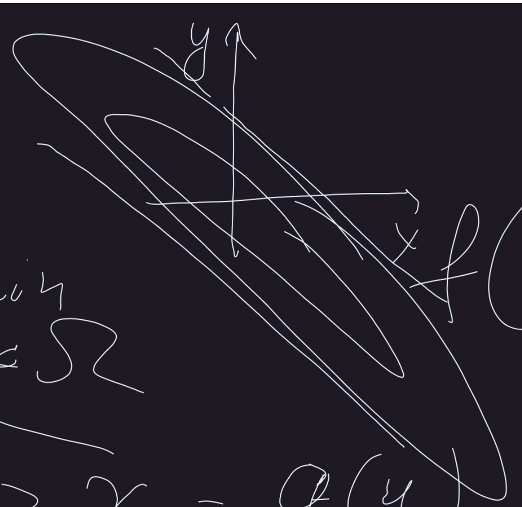
$F(W) \to \min_{W \in \text{Orth.}}$

$z = Wx$

$\|z\|_2 = \|Wx\|_2 = \|x\|_2$

$\nabla_x L = W^T \nabla_z L$

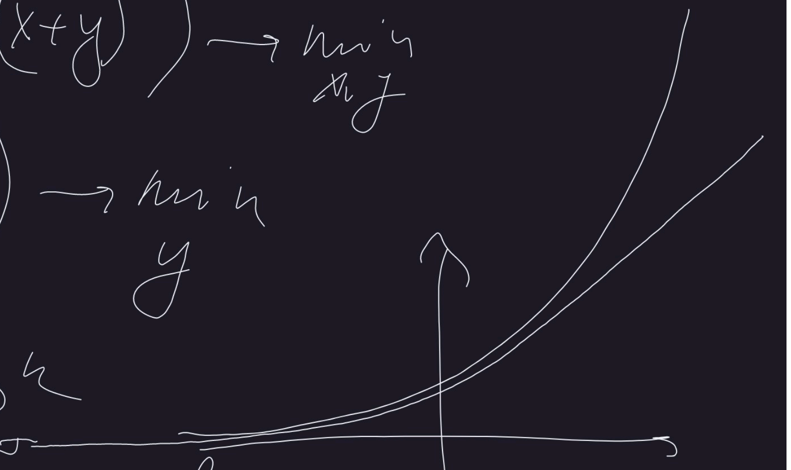
$\|\nabla_x L\|_2 = \|W^T \nabla_z L\|_2 \overset{n=d}{=} \|\nabla_z L\|_2$



$$f(x) \to \min_{x \in \Omega}$$

$$f((x+y)^2) \to \min_{x,y}$$

$$f(g(y)) \to \min_{y}$$



$$x \in \Omega \Leftrightarrow x = g(y),\ y \in R^n$$

$$[a,b] \ni g(y) = a + (b-a)\,\sigma(y)$$

$$g(y) = \log(1+\exp(y))$$

$$g(y) = \exp(y) > 0$$

$$g(y) = y^2$$

$$W = \operatorname{expm}(V - V^T)$$

$$p(y|x) = N(y \mid NN_1(x,\theta),\ NN_2(x,\theta))$$