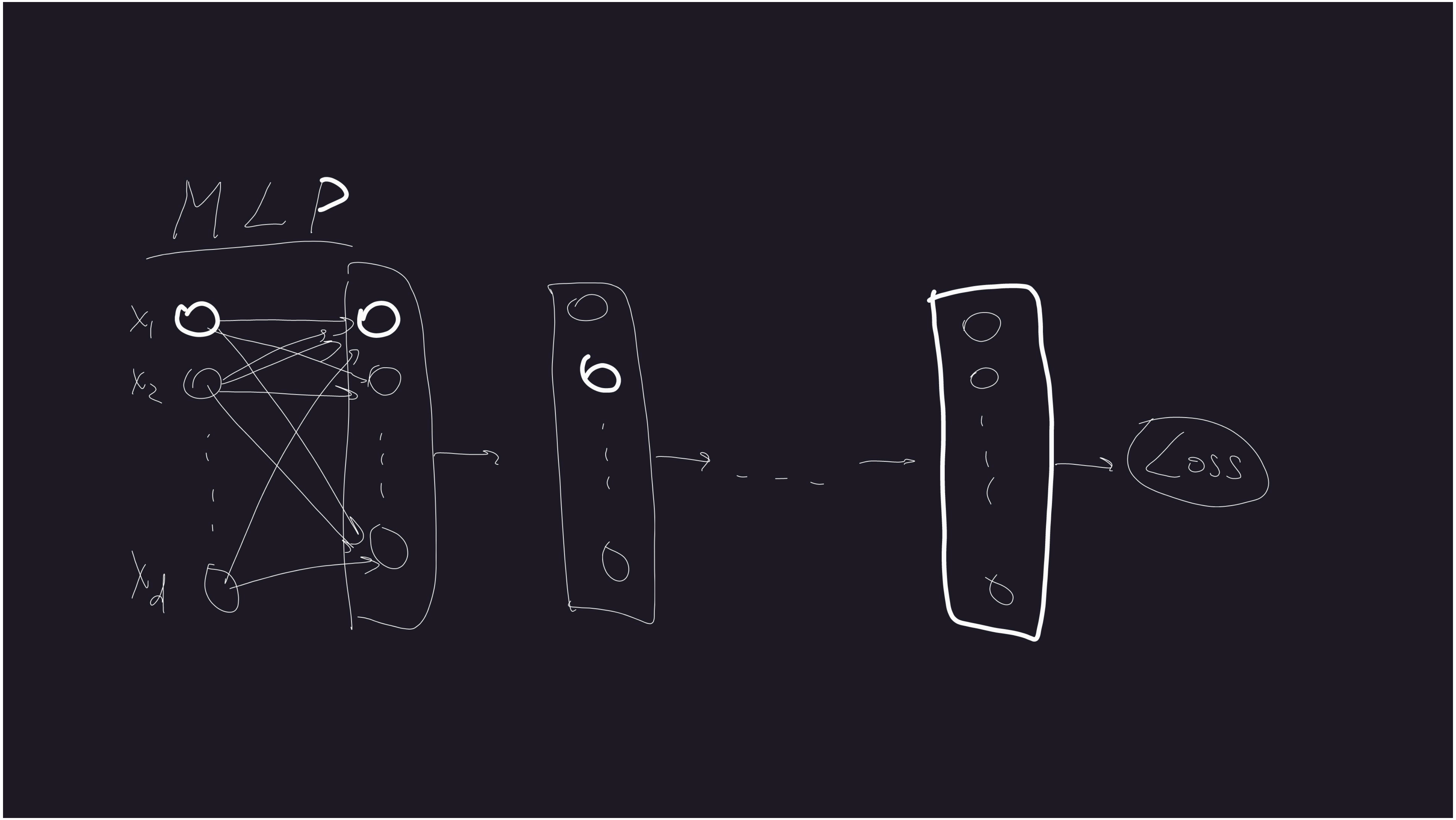

MLP
x_1
x_2
x_d
Loss

$$\frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f(x_i, \Theta)\big) + \frac{\lambda}{2}\|\Theta\|_2^2 \to \min_{\Theta}$$

$$\Updownarrow$$

$$\begin{cases} \frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f(x_i, \Theta)\big) \to \min_{\Theta} \\ \|\Theta\|_2^2 \leq \eta \end{cases}$$

Drop Out

$z = Wx + b$

$a = g(z)$

$y_i \sim \text{Bern}:$ $\begin{array}{cc} 0 & 1 \\ p & 1-p \end{array}$

$a_i^{DO} = a_i \cdot y_i$

$\mathbb{E}\, a_i^{DO} = \mathbb{E}_{y_i \sim \text{Bern}}\, a_i \cdot y_i = p(0 \cdot a_i) + (1-p)\, 1 \cdot a_i = a_i(1-p)$

Classic DO

Train: $y_i \sim \text{Bern},\ a_i^{DO} = a_i \cdot y_i$

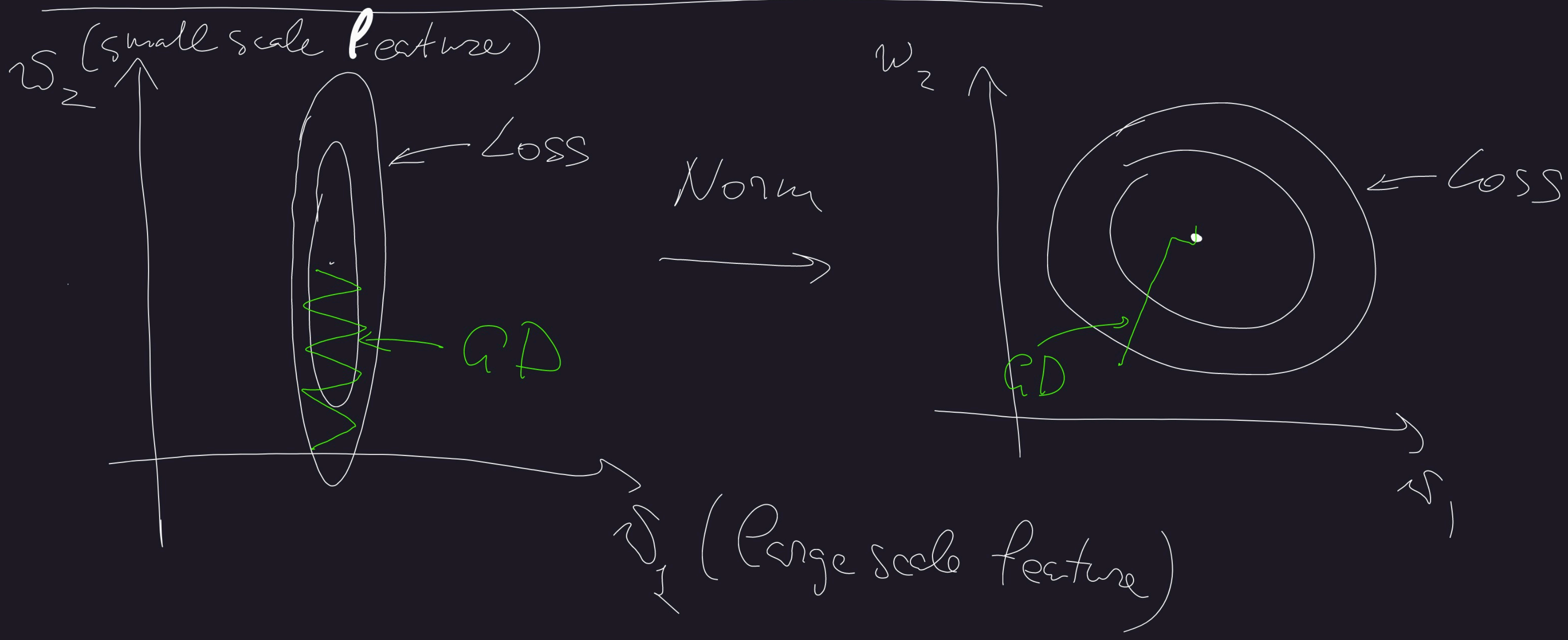
Test: $a_i^{DO} = \dfrac{a_i}{1-p}$

Inverted DO:

Train: $y_i \sim \text{Bern},\ a_i^{DO} = \dfrac{1}{1-p}\, a_i \cdot y_i$

Test: $a_i^{DO} = a_i$

Batch Normalization



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

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

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

$$\sigma_j^2 = \frac{1}{N_{batch}} \sum_i (x_{ij} - m_j)^2 \quad \forall j$$

$$z_{ij} = \frac{x_{ij} - m_j}{\sqrt{\sigma_j^2 + \varepsilon}}$$

$$\hat{x}_{ij} = \gamma_j z_{ij} + \delta_j$$

Train: compute m_j, σ_j on mini-batch and learn γ_j, δ_j

Test: use saved m_j, σ_j for trainset

Image $x \in \mathbb{R}^{H \times W \times C}$

Batch $\in \mathbb{R}^{B \times H \times W \times C}$

BatchNorm: for each C average $B \times H \times W$

LayerNorm: for each B average $H \times W \times C$

InstanceNorm: for each B, C average $H \times W$

Weight Initialization

$$z = Wx$$

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

$$X_j \sim \mathcal{N}(0, 1), \quad W_{ij} \sim \mathcal{N}(0, Var(W))$$

$$z_i = \sum_{j=1}^{n_{input}} W_{ij} x_j, \quad \mathbb{E} z_i = \mathbb{E} \sum_j W_{ij} x_j = \sum_j \mathbb{E} W_{ij} x_j =$$

$$= \sum_j \mathbb{E} W_{ij} \mathbb{E} x_j = 0$$

$i = 1, \ldots, n_{output}$

$j = 1, \ldots, n_{input}$

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

$$\frac{\partial f}{\partial x_j} = \sum_{i=1}^{n_{outputs}} (W^T)_{ji} \frac{\partial f}{\partial z_i}$$

$$\boxed{Var\left(\frac{\partial f}{\partial x_j}\right) = n_{outputs} \cdot Var(W)}$$

Xavier (Glorot) init.

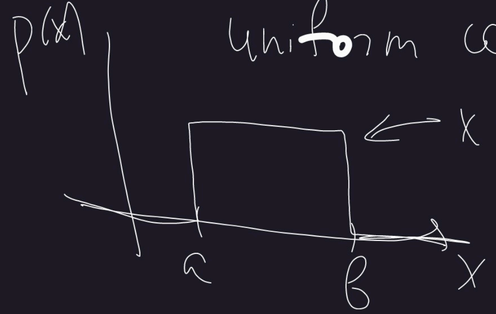
Activations: sigmoid, tanh

$$Var(W) = \frac{2}{n_{input} + n_{output}}$$

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

$$W_{ij} \sim R\left(-\sqrt{\frac{6}{n_{input} + n_{output}}}, \sqrt{\frac{6}{n_{input} + n_{output}}}\right)$$

uniform cont. distr.



$x \sim R[a, b]$

Kaiming (He) init.

$$Var(W) = \frac{2}{n_{input}}$$

Activations: ReLu,
Leaky ReLu

$$W_j \sim \mathcal{N}(0, Var(W))$$

$$W_j \sim R\left(-\sqrt{\frac{6}{n_{input}}}, \sqrt{\frac{6}{n_{input}}}\right)$$

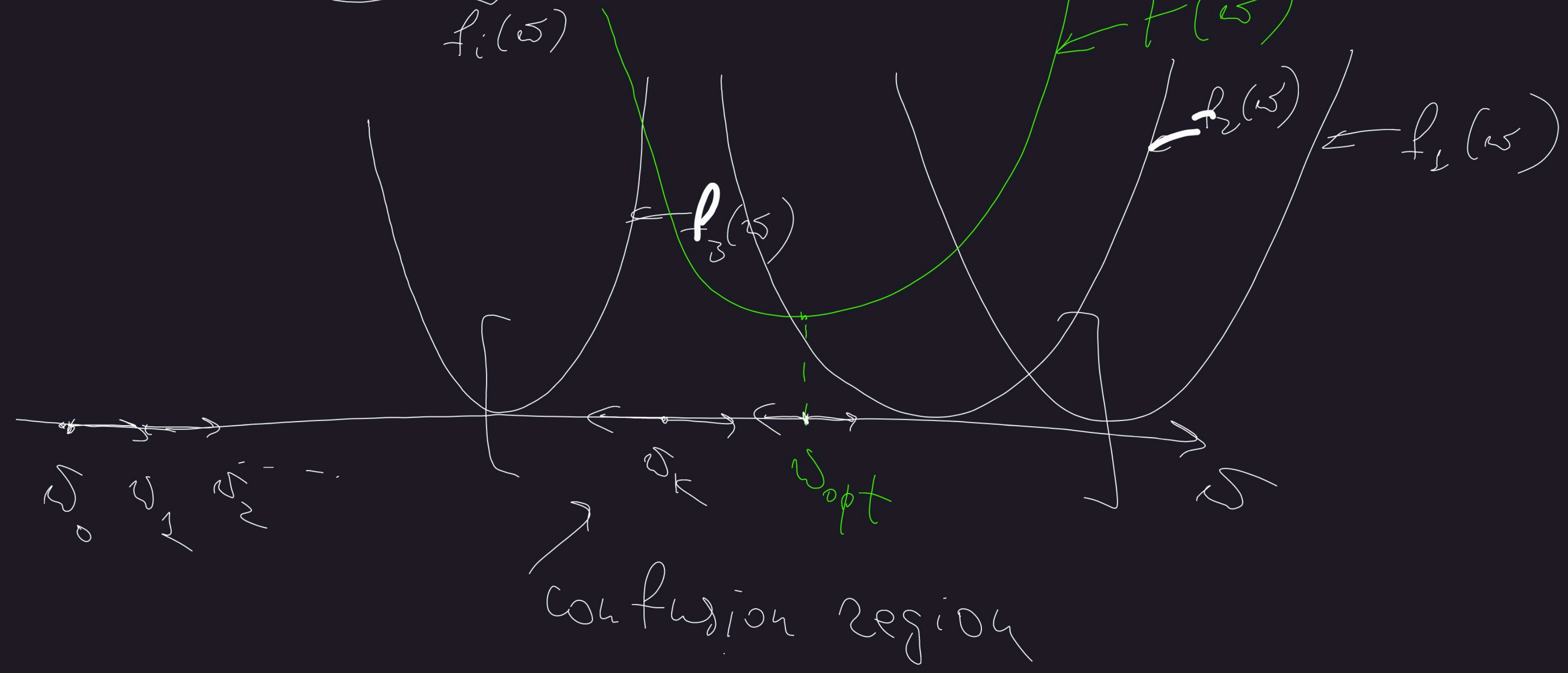
$$F(w) = \frac{1}{N}\sum_{i=1}^{N} f_i(w) \to \min_{w} \quad ; \quad N >> 1$$

$$\nabla F(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(w)$$

SGD

$$\begin{cases} I_k \subset \mathrm{Unif}(1, \ldots, N) \\ g_k = \frac{1}{|I_k|}\sum_{i \in I_k} \nabla f_i(w_k) \\ w_{k+1} = w_k - \alpha_k g_k \end{cases}$$

$$F(w) = \frac{1}{N} \sum_i \underbrace{(w x_i - y_i)^2}_{f_i(w)} \to \min_w \qquad x_i, y_i \in \mathbb{R}$$



$$\hat{g}_k = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(w_k)$$

$$x \sim \mathcal{N}(x \mid \mu, \sigma^2)$$

$$x = \mu + \sigma \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1)$$

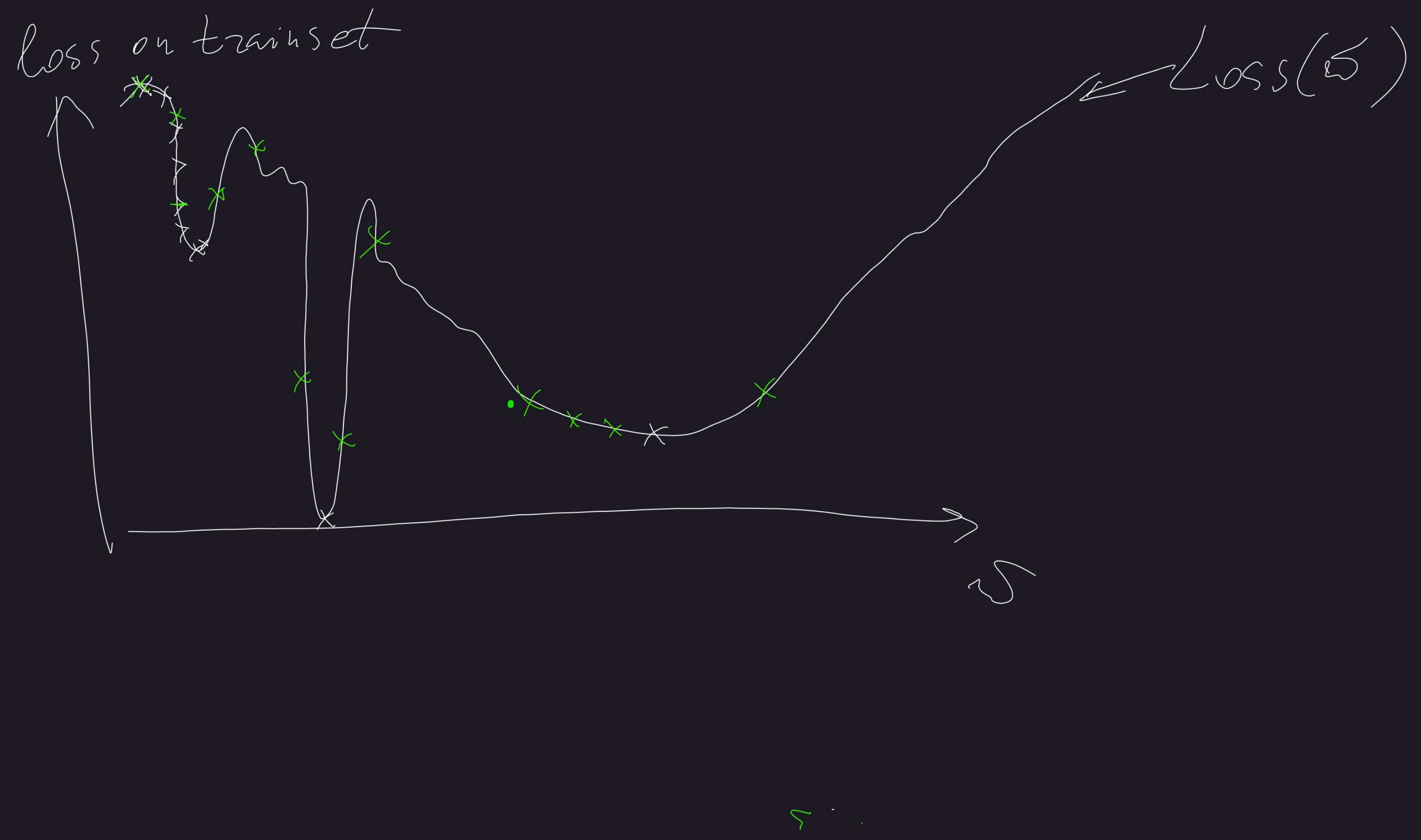
$$g_k = \frac{1}{|I_k|} \sum_{i \in I_k} \nabla f_i(w_k)$$

$$g_k \sim \mathcal{N}(\hat{g}_k, \hat{\Sigma}_k), \quad \hat{\Sigma}_k \to 0 \text{ if } |I_k| \to N$$

$$\text{SGD: } w_{k+1} = w_k - \alpha_k g_k = w_k - \alpha_k (\hat{g}_k + \varepsilon_k)$$

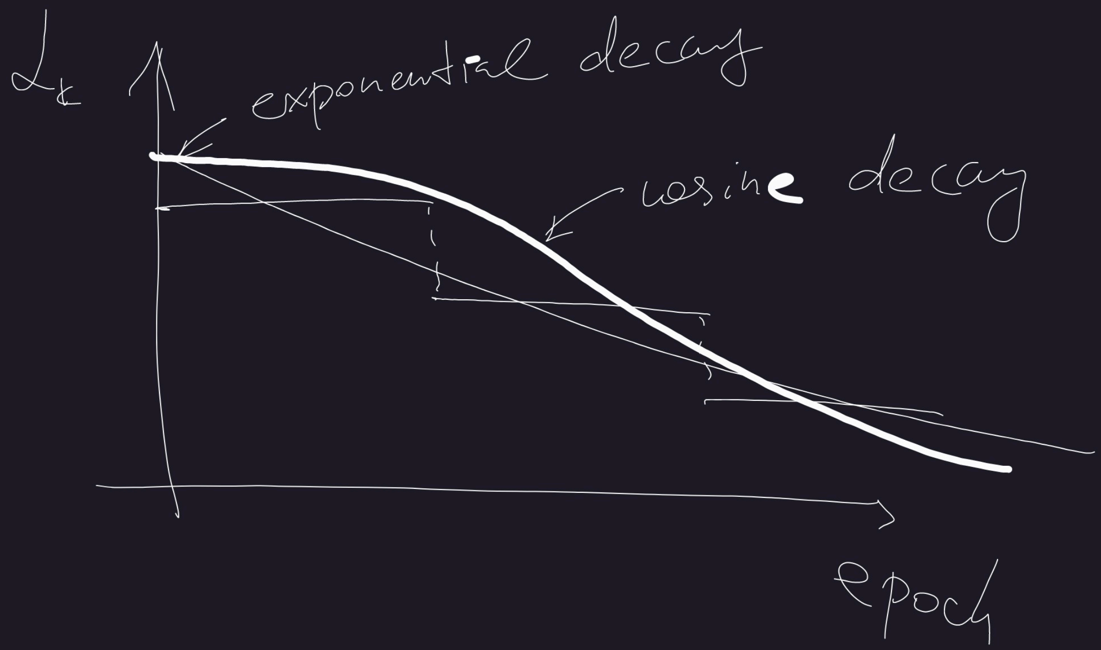
Noisy full GD

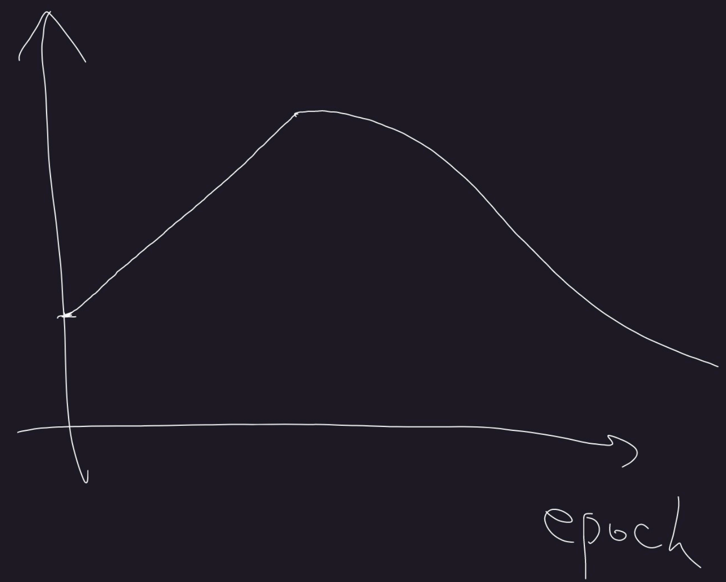
$$\varepsilon_k \sim \mathcal{N}(0, \Sigma_k)$$

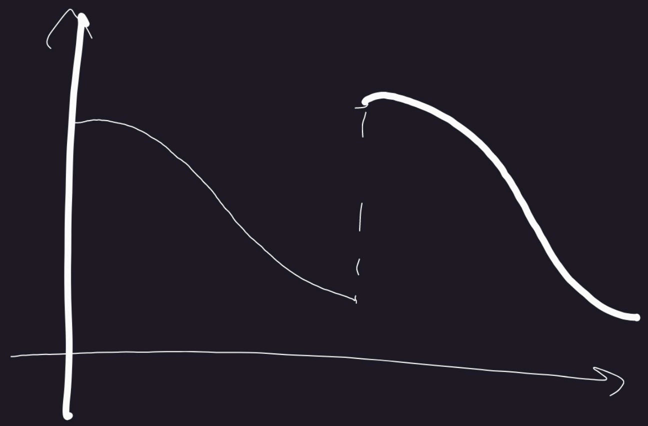


loss on trainset
Loss(ω)
ω

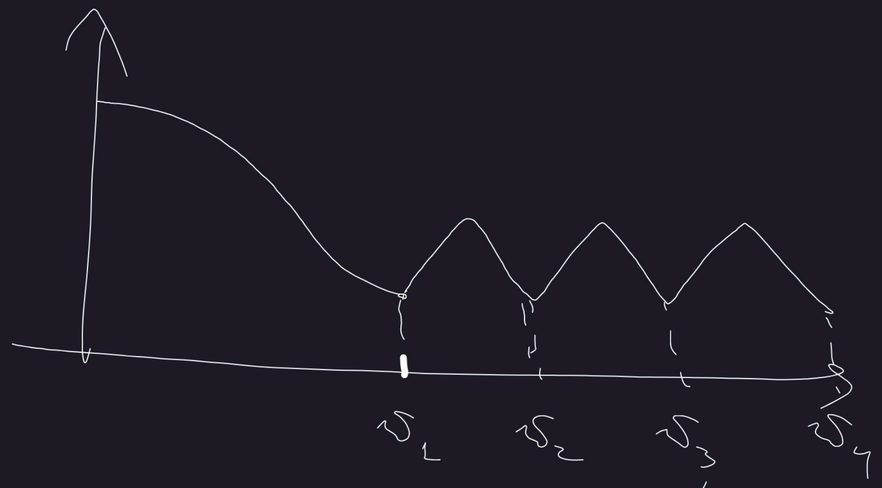
Learning Rates Schedule







SWA



$$w = \frac{1}{4} \cdot \sum_{i=1}^{4} w_i$$