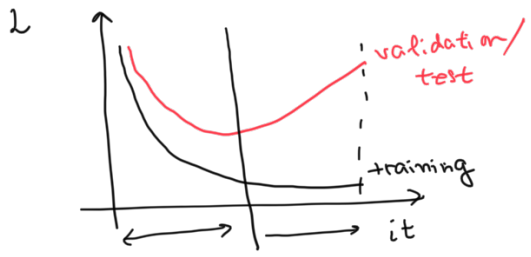
Overfitting and regularization :

$$\mathcal{L}(\theta) := \underbrace{\frac{1}{2}\sum_{i=1}^{n}\left(f_\theta(x_i) - y_i\right)^2}$$

$$\theta := \{W^{(1)}, b^{(1)}, \dots, W^{(L)}, b^{(L)}, W^{(0)}, b^{(0)}\}$$

1.) Early stopping :



Idea : Monitor the validation loss and stop the training early (as soon as the validation loss starts to diverge).

2.) L_1 / L_2 - parameter regularization (MAP estimation)

$$\mathcal{L}(\theta) = \underbrace{\frac{1}{2}\sum_{i=1}^{n}\left[f_\theta(x_i) - y_i\right]^2}_{\text{data fit}} + \underbrace{\frac{\lambda_1}{2}\|W\|_1}_{\text{Laplace prior}} + \underbrace{\frac{\lambda_2}{2}\|W\|_2}_{\text{Gaussian prior}} \quad \text{(model complexity)}$$

λ_1, λ_2 $\Theta(10^{-1} - 10^{-2})$

$$p(\theta|x,y) \propto \underbrace{p(y|x,\theta)}_{\text{Gaussian likelihood}} \; \underbrace{p(\theta)}$$

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

3.) Dropout :

Recall : $H^{(\ell)} = f\left(H^{(\ell-1)} W^{(\ell)} + b^{(\ell)}\right)$

$r_j^{(\ell)} \sim \text{Bernalli}(p)$, $j = 1, \dots, Q^{(\ell)}$: # of neurons in the ℓ-th layer

$$\tilde{z}^{(\ell)} = r^{(\ell)} \odot H^{(\ell-1)}$$

$$H^{(\ell)} = f\left(z^{(\ell)} \cdot W^{(\ell)} + b^{(\ell)}\right)$$

4.) Data augmentation (classification tasks)

$$\hat{X} = \frac{X - \mathbb{E}[X]}{std[X]}$$

Network initialization:

$D := \{X, Y\}$ is a "standardized" data-set : $\mathbb{E}[X] = 0$, $Var = 1$

$\mathbb{E}[Y] = 0$, $Var = 1$

$X \in \mathbb{R}^{d_{in}}$, $Y \in \mathbb{R}^{d_{out}}$, X, Y i.i.d.

Linear regression : $Y = W_1 x_1 + \cdots + W_d x_{d_{in}}$, W_i are zero-mean

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

$$\overset{i.i.d.}{\Rightarrow} \quad Var[Y] = d_{in} \cdot Var[x_i]\, Var[w_i]$$

$$\longrightarrow \quad Var[w_i] = \frac{1}{d_{in}} \qquad (1)$$

Repeat this analysis for the back-propagated gradient signal

$$\longrightarrow \quad Var[w_i] = \frac{1}{d_{out}} \qquad (2)$$

Empirical rule for initializing W :

$$Var[w_i] = \frac{2}{d_{in} + d_{out}} \quad \checkmark$$

Glorot initialization:

- $W \sim U\left[-\frac{\sqrt{6}}{\sqrt{d_{in}+d_{out}}}, \frac{\sqrt{6}}{\sqrt{d_{in}+d_{out}}}\right]$ Glorot uniform initialization

- $W \sim \mathcal{N}\left(0, \frac{2}{d_{in}+d_{out}}\right)$ Glorot normal initialization