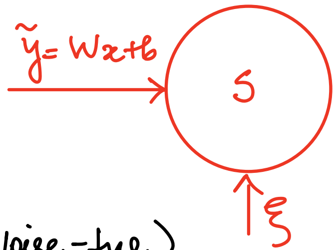


(additive noise)

Consider input: (Noise-free)

$$\tilde{y} = Wx + b$$

Subject to (gaussian) noise

e.g. $\xi(t) = N(0;\sigma)$

$$y = \tilde{y} + \xi$$

Neuron is in state S given by:

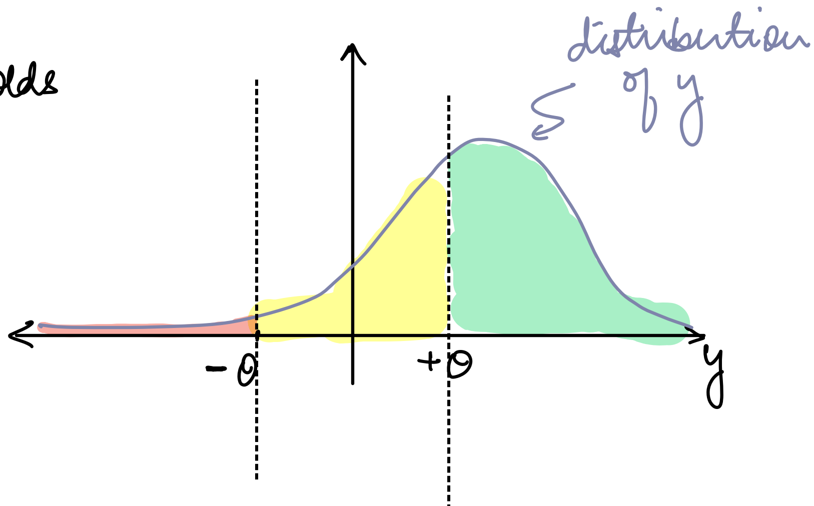
$$S = h(y)$$

The goal is to find "Expected activation" of the neuron.

Let the neuron take states $S \in \{S_1, S_2, \dots, S_k\}$

The neuron assumes a state $S = S_i$ as a function of the noisy input arriving at it y. Let that function be $h(y)$

θ's: thresholds



for a $k=3$:

$$S = \{-1, 0, +1\}$$

we can define thresholds as $\pm\theta$.

as an example: say h is the activation:

$$h(y) = \begin{cases} -1 & ; -\infty < y < -\theta \\ 0 & ; -\theta < y < \theta \\ 1 & ; +\theta < y < +\infty \end{cases}$$

$\mathbb{E}[h(y)] \doteq$ Expected value of the state of the neuron.

In general, if the controllable inputs are $\tilde{y} = Wx + b$;

$P(S=s_i \mid \tilde{y}) \rightarrow$ probability of finding a neuron in state s_i given NOISE FREE input $\tilde{y}$.

Let $\theta \in \{\theta_0, \theta_1, \dots, \theta_{k-1}, \theta_k\}$ be the thresholds. $\theta_0 = -\infty$; $\theta_k = +\infty$

$$P(S=s_i \mid \tilde{y}) = P(y \in \{\theta_{i-1}, \theta_i\})$$

$$= \int_{\theta_{i-1}}^{\theta} P(y \mid \tilde{y})\, dy$$

(due to noise, y can overshoot or undershoot a threshold)

Define $\int_{-\infty}^{\theta} P(y \mid \tilde{y})\, dy = F(\theta)$ (CDF of $P(y \mid \tilde{y})$)

$$\therefore P(S=s_i \mid \tilde{y}) = F(\theta_i) - F(\theta_{i-1})$$

$$\therefore \mathbb{E}[(S \mid \tilde{y})] = \sum_{i=1}^{k} s_i\, P(S=s_i \mid \tilde{y})$$

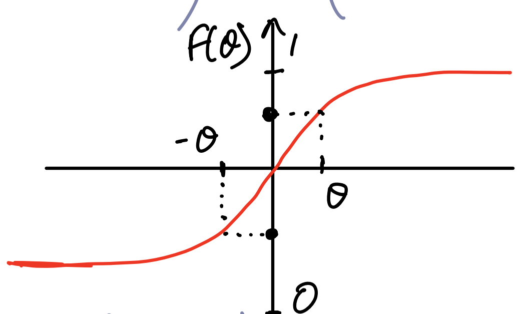
$$\boxed{\mathbb{E}[(S \mid \tilde{y})] = \sum_{i=1}^{k} s_i \left(F(\theta_i) - F(\theta_{i-1}) \right)}$$

if $S \in \{-1, 0, 1\}$; $\theta_1 = -\theta$, $\theta_2 = +\theta$

$$\mathbb{E}[(S \mid \tilde{y})] = -1 \cdot \left(F(-\theta) - F(-\infty)\right) + 0 \cdot \left(F(\theta) - F(-\theta)\right) + 1 \cdot \left(F(\infty) - F(\theta)\right)$$

$$= -F(\theta) + \underbrace{F(\infty)}_{1} - F(\theta)$$

$$= 1 - \left(F(\theta) + F(-\theta)\right)$$



if $f(\cdot)$ is sigmoidal function (erf, tanh, ...)

$$f(-\theta) = 1 - f(\theta)$$

$$\therefore \mathbb{E}[(s|\tilde{y})] = 1 - (f(\theta) + 1 - f(\theta))$$

$$\boxed{\mathbb{E}[(s|\tilde{y})] = 0}$$

The expected state is always zero (?)

- Bringing in the effect of input:

$$\mathbb{E}[(s|\tilde{y})] = \mathbb{E}[(h(y)|\tilde{y})]$$

and if $h(\cdot)$ is simply thresholding function; under additive noise

$$h(y) \rightarrow \int_{\theta_{i-1}}^{\theta_i} P(y|\tilde{y})\, dy \; ; \; \forall i = 1, \ldots, k$$

$$\therefore \frac{d}{dy} h(y) = \frac{d}{dy} \int_{\theta_{-i}}^{\theta_i} P(y|\tilde{y})\, dy$$

Fundamental theorem of calculus:

$$\boxed{\frac{d}{dy} h(y) = P(y = \theta_i | \tilde{y}) - P(y = \theta_{i-1} | \tilde{y})}$$

Now consider change in the expected state w.r.t. input

$$\frac{d}{dy} \mathbb{E}[(s|\tilde{y})] = \frac{d}{dy} \mathbb{E}[(h(y)|\tilde{y})]$$

$$= h(y) \cdot P(y \in \{\theta_i, \theta_{i-1}\}) \cdot \frac{d}{dy} h(y)$$

$$= h(y) \cdot \underbrace{\left[P(y=\theta_i|\tilde{y}) - P(y=\theta_{i-1}|\tilde{y}) \right]}_{\frac{d}{dy}h(y)} \cdot \int_{\theta_{i-1}}^{\theta_i} P(y|\tilde{y})\, dy$$

$$\boxed{\frac{d}{dy}\mathbb{E}\left[(s|\tilde{y})\right] = h(y) \cdot \left[P(y=\theta_i|\tilde{y}) - P(y=\theta_{i-1}|\tilde{y}) \right] \cdot \left[F(\theta_i) - F(\theta_{i-1}) \right]}$$

- Now credit assignment can be done with chain rule:

$$\frac{d}{dW}\mathbb{E}\left[(s|\tilde{y})\right] = \frac{d}{dy}\mathbb{E}\left[(s|\tilde{y})\right] \cdot \frac{dy}{dW}$$

$$\because \quad y = \underbrace{Wx + b}_{\tilde{y}} + \xi$$

$$\therefore \quad \frac{dy}{dW} = x \quad \left(\cancel{+ \frac{d\xi}{dW}} \to 0 \right)$$

$\because$ ξ does not depend on W

$$\therefore \quad \boxed{\frac{d}{dW}\mathbb{E}\left[(s|\tilde{y})\right] = \frac{d}{dy}\mathbb{E}\left[(s|\tilde{y})\right] \cdot x} \quad \text{(Single weight case)}$$

→ This would turn into outer product update if W is matrix & x & y are vectors

$$\Delta W = \eta \frac{d}{dW}\mathbb{E}\left[(s|\tilde{y})\right] \otimes \vec{x}$$

learning rate