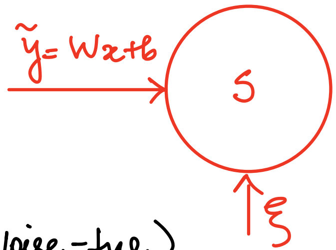
"TriDeNT"



(additive noise)

Consider input: (Noise-free)

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

Subject to (gaussian) noise

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

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

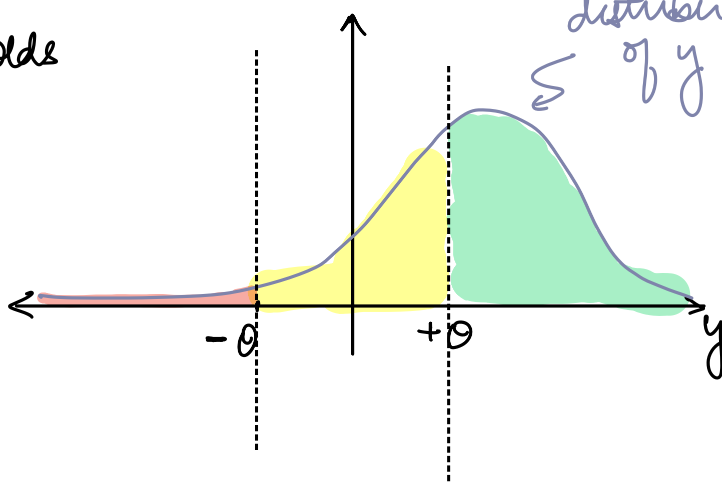
Neuron is in state s given by:

$$s = h(y)$$

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

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

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

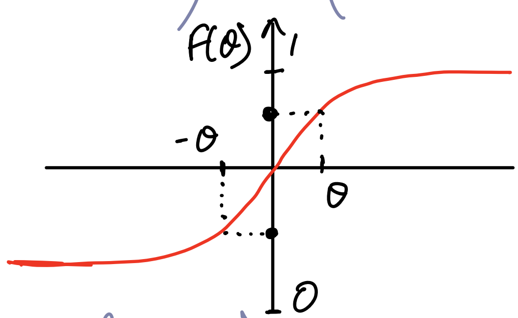
$$\boxed{E[(S \mid \tilde{y})] = \sum_{i=1}^{k} s_i \left(F(\theta_i) - F(\theta_{i-1})\right)}$$

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

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

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

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



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

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

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

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

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

$$= \frac{d}{dy}\sum_{i=1}^{k} s_i\, P(y \in \{\theta_{i-1}, \theta_i\}|\tilde{y})$$

$$= \sum_{i=1}^{k} s_i \frac{d}{dy} \int_{\theta_{i-1}}^{\theta_i} p(y|\tilde{y})\, dy$$

$$\boxed{\frac{d}{dy} \mathbb{E}\left[(s|\tilde{y})\right] = \sum_{i=1}^{k} s_i \left[p(y=\theta_i|\tilde{y}) - p(y=\theta_{i-1}|\tilde{y}) \right]}$$

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

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

- Now credit assignment can be done with chain rule:

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

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

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

$\because \xi$ does not depend on W

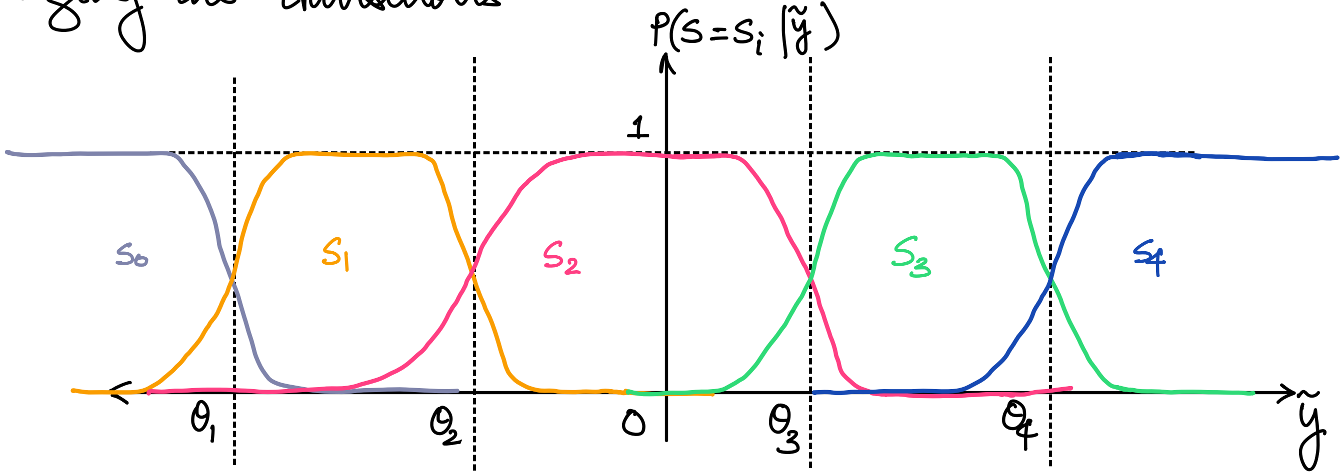
$$\therefore \boxed{\frac{d}{dW}\mathbb{E}\left[(s|\tilde{y})\right] = \frac{d}{dy}\mathbb{E}\left[(s|\tilde{y})\right]\cdot x} \quad \text{(Single weight case)}$$

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

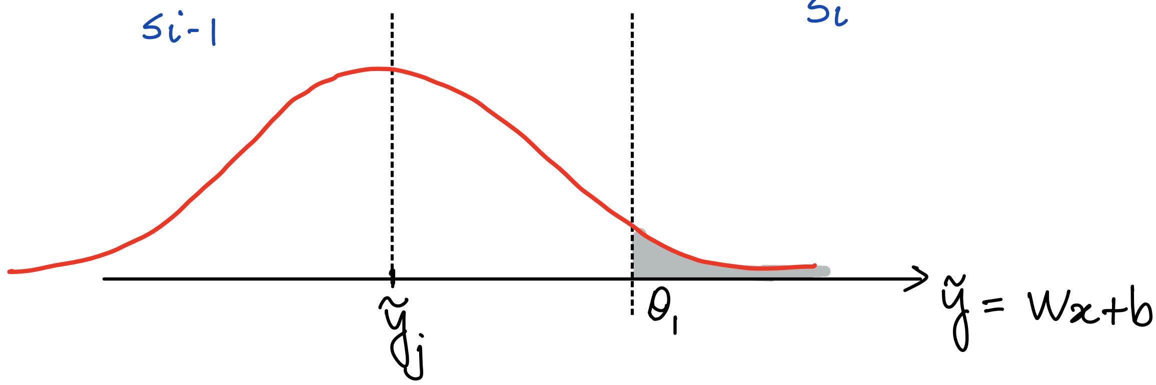
$$\Delta W = -\eta\,\frac{d}{dW}\mathbb{E}\left[(s|\tilde{y})\right]\otimes\vec{x}$$

↑ learning rate

• Visualizing the transitions



Consider a threshold θ_i around which the model goes from state S_{i-1} to S_i

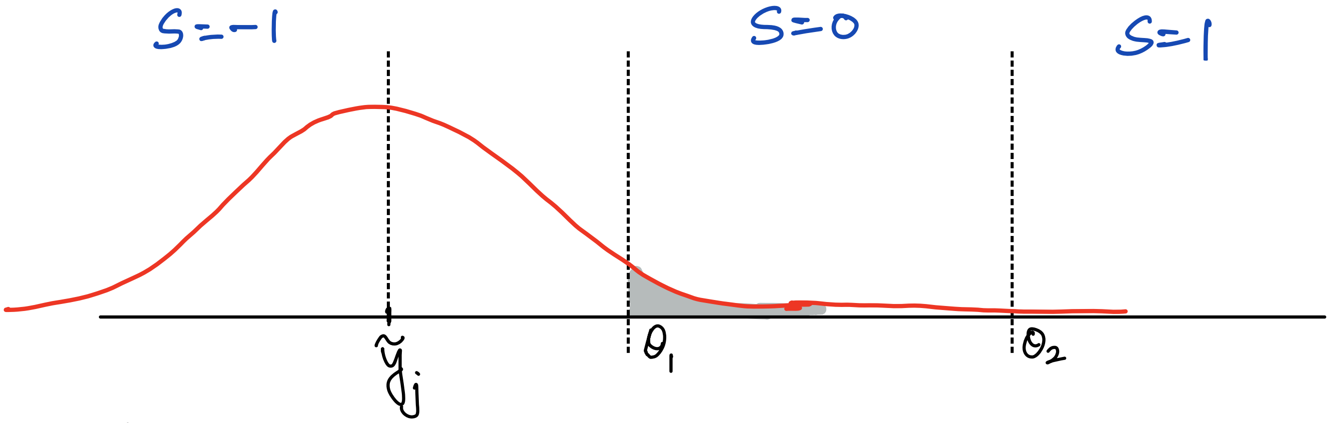


Let $\tilde{y}_j$ be the input for j^{th} sample datum. As the decision variable $y_j = \tilde{y}_j + \xi_j$, the probability of the model switching to state S_i is:

$$P(y>\theta_1|\tilde{y}) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\theta_1}^{\infty} e^{-\frac{(u-y_j)^2}{2\sigma^2}}\,du$$

& $$P(y<\theta_1|\tilde{y}) = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\theta_1} e^{-\frac{(u-y_j)^2}{2\sigma^2}}\,du.$$

for $s \in \{-1, 0, +1\}$



$$\mathbb{E}\left[(s|\tilde{y})\right] = (-1)\cdot P(y<\theta_1|\tilde{y}) + 0\cdot\left(P(\theta_1<y<\theta_2|\tilde{y})\right) + (+1)\cdot P(y>\theta_1|\tilde{y})$$

$$= (-1)\cdot \frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\theta_1} e^{-\frac{(u-\tilde{y}_j)^2}{2\sigma^2}}\,du \;+\; \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\theta_2}^{\infty} e^{-\frac{(v-\tilde{y}_j)^2}{2\sigma^2}}\,dv$$

$$\boxed{= \underbrace{\frac{1}{\sqrt{2\pi}\,\sigma}}_{C}\left[\int_{\theta_2}^{\infty} e^{-\frac{(v-\tilde{y}_j)^2}{2\sigma^2}}\,dv - \int_{-\infty}^{\theta_1} e^{-\frac{(u-\tilde{y}_j)^2}{2\sigma^2}}\,du\right]}$$

$$\frac{d}{d\tilde{y}_j}\mathbb{E}\left[(s|\tilde{y}_j)\right] = c\cdot\left[-\left\{e^{-\infty} - e^{-\frac{(\theta_2-\tilde{y}_j)^2}{2\sigma^2}}\right\} - (-1)\left\{e^{-\frac{(\theta_1-\tilde{y}_j)^2}{2\sigma^2}} - e^{\infty}\right\}\right]$$

$$\boxed{\frac{d}{d\tilde{y}_j}\mathbb{E}\left[(s|\tilde{y}_j)\right] = c\cdot\left[N(\theta_2;\tilde{y}_j,\sigma) + N(\theta_1;\tilde{y}_j,\sigma)\right]}$$

- Discretizing the gradients

→ Sigmoid everywhere, computing der. → E

→ Experiments: 4 figures

1. Net. architecture, overview
   ↓
   - 3panel idea
   - Net topology
   - Backprop.

→ 3 result figures.

→ How does this scale?

↳ Not sure how to show that.

$$\begin{array}{r|c|c|c|} & & p & \\ \text{label} & 0 & 1 & 0 \\ \end{array}$$

$$\begin{array}{r|c|c|c|} z & .01 & .9 & .3 \\ \end{array}$$

$z[p] = .9$

$$\underbrace{\sum \max(0, .01 - .9)}_{0} + \max(\quad)$$

$$\frac{d}{dy} \underbrace{E[(h(y)|\tilde{y})]}_{} = \underbrace{\frac{dE}{dh}} \cdot \underbrace{\frac{dh}{dy}}$$

$$= \underbrace{\sum_{i=1}^{k} h_i(y) p(h_i(y)|\tilde{y})}$$

$$\frac{dE}{dh(y)} \cdot \frac{dh}{dy} = \frac{d}{dh_j} \sum_{i=1}^{k} h_i(y)\, P(h(y)=S_i|\tilde{y}) \cdot \frac{dh_i}{dy}$$

$$= P(\underbrace{h_j(y)}_{s_j}|\tilde{y}) \cdot \frac{dh}{dy}$$

$$y \in \{\theta_{j-1}, \theta_j\}$$

C.R.V

$$E[(h(y)|\tilde{y})] = \int_{-\infty}^{\infty} h(y)\, p(h(y)|\tilde{y})\, dh$$

$$\frac{\partial E}{\partial h} = \int \frac{\partial}{\partial h}\left(h(y)\, p(h(y) \mid \tilde{y})\right) dh$$

$$= \int \left[p(h(y) \mid \tilde{y}) + h(y) \frac{\partial}{\partial h} p(h(y) \mid \tilde{y})\right] dh$$

$$= 1 + h(y) \frac{\partial}{\partial h} p(h(y) \mid \tilde{y})$$

D.R.U

$$E\left[p(h(y) \mid \tilde{y})\right] = \sum_{i=1}^{K} h_i(y)\, p(h = h_i(y) \mid \tilde{y})$$

$$\frac{\partial E}{\partial h_j} = \frac{\partial}{\partial h_j} \sum_i h_i(y)\, p(h = h_i \mid \tilde{y}) = p(h = h_j \mid \tilde{y}) + \underbrace{h_j(y) \frac{\partial}{\partial h_j} p(h = h_j \mid \tilde{y})}_{?}$$

what's happening to this term

$$E\left[(s \mid \tilde{y})\right] = \sum_i s_i\, p(y \mid \tilde{y})$$

$$\frac{d}{d\tilde{y}} \int e^{-\frac{(u-\tilde{y})}{2\sigma^2}}\, du$$

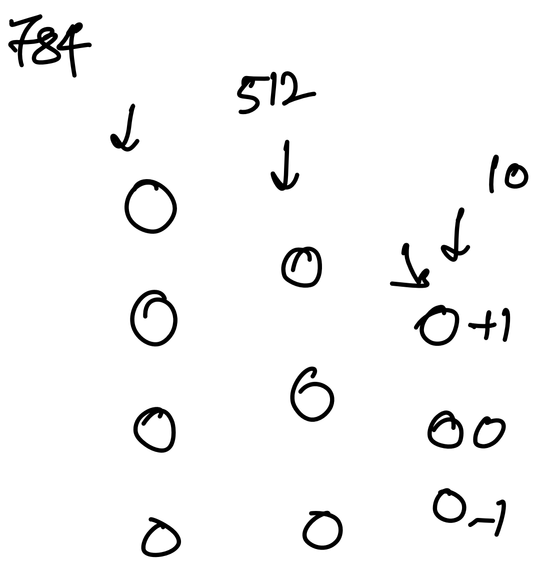
$$= \int \frac{d}{d\tilde{y}}\, e^{-\frac{(u-\tilde{y})^2}{2\sigma^2}}\, du$$

$$= \int e^{-\frac{(u-\tilde{y})^2}{2\sigma^2}} \cdot \frac{1}{2\sigma^2} \cdot (-2)(u-\tilde{y}_j)\, du$$

$$= (-1) \int \overbrace{\frac{\cancel{2}}{2\sigma^2}(u-\tilde{y}_j)}^{z'}\, e^{\overbrace{-\frac{(\ \ )}{2\sigma^2}}^{z}}\, du$$

$$= -1 \cdot e^{-\frac{(u-\tilde{y})^2}{2\sigma^2}}$$

$$\underbrace{\int x \cdot e^{-x^2}\, dx}_{-\cancel{e^{-x^2}}}$$



1. Expected state
2. N/w topology
3. Loss : Cross entropy

   $l \rightarrow$ label

   $p \rightarrow$ model $\rightarrow$ softmax$(w_f x_f + b_f)$

$$\mathcal{L} = -\sum_{c=1}^{10} l_c \log p_c$$

$= 0 \cdot \log 0.01 +$

$0 \log 0.01 +$

$1 \underbrace{\log 0.9}_{\simeq 1} \simeq 0$

0	0	1

.01	0.01	0.9

.9	.01	.01

$0 \cdot \log_{10} .9 + 0 \cdot \log_{10} .01$

$+ 1 \cdot \log_{10} .01$ ← -ve number

$(-1)(-2) = 2$

$$\sum_{c=1}^{10} p_c \log \frac{p_c}{l_c + \epsilon}$$

Potential Paper title(s)

1. Deep Learning with ternary Stochastic Neurons
2. Deep Stochastic Ternary NNs