Logistic $\sigma(x) = \frac{1}{1+e^{-x}} \in (0,1)$

$\sigma'(x) = \sigma(x)(1-\sigma(x))$

Hypobolic tangent

$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \in (-1,1)$

$\tanh'(x) = 1 - \tanh^2(x)$

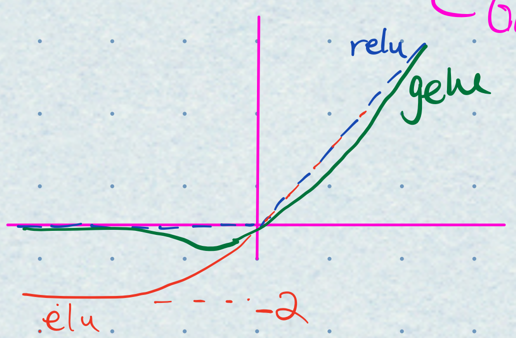
ReLU $\text{relu}(x) = \max(0, x)$

Leaky ReLU $f(x) = \begin{cases} 0.01x, & \text{if } x<0 \\ x, & \text{if } x \geq 0 \end{cases}$

GELU $\text{gelu}(x) = x\Phi(x) \approx x\,\sigma(1.702x)$

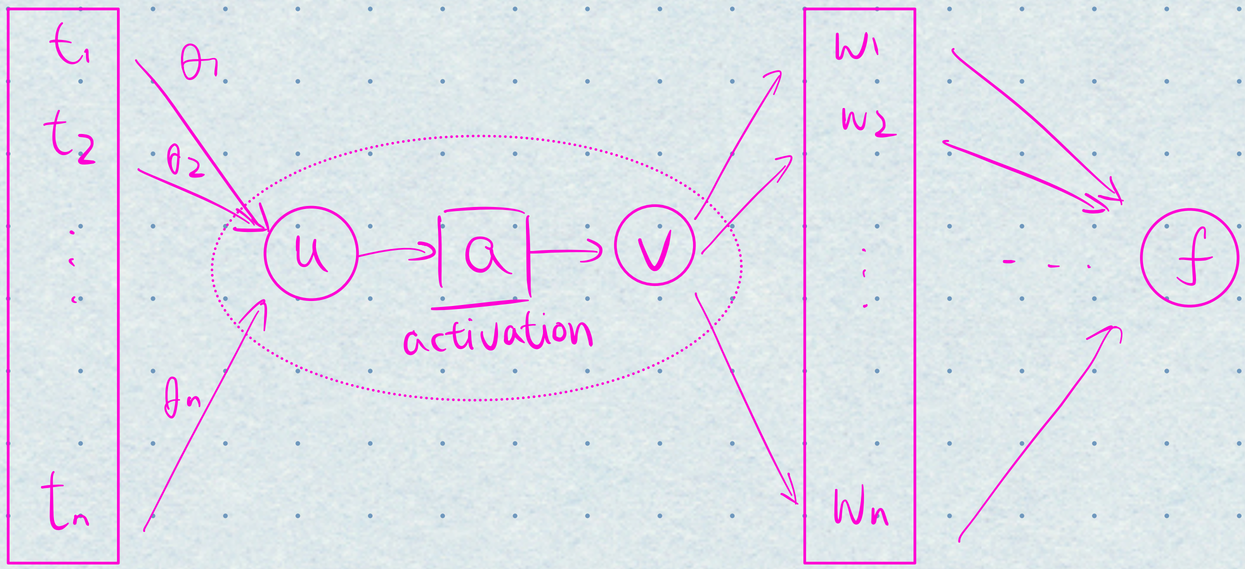
Gaussian CDF

SiLU $= x\sigma(x)$




ELU $\text{elu}(x) = \begin{cases} \alpha(e^{x}-1), & \text{if } x<0 \\ x, & \text{if } x \geq 0 \end{cases}$

# Back propagation




$$\frac{\partial f}{\partial \theta_1} = \underbrace{\frac{\partial f}{\partial v}}_{?} \underbrace{\frac{\partial v}{\partial u}}_{a'(\cdot)} \underbrace{\frac{\partial u}{\partial \theta_1}}_{t_1}$$

$$\frac{\partial f}{\partial v} = \sum_j \frac{\partial f}{\partial w_j} \underbrace{\frac{\partial w_j}{\partial v}}_{\text{compute on } w_j}$$

$$\frac{\partial f}{\partial \theta_i} = \left(\sum_j \frac{\partial f}{\partial w_j} \frac{\partial w_j}{\partial v}\right) \frac{\partial v}{\partial \theta_i} = \overbrace{\left(\sum_j \left(\sum_k \frac{\partial f}{\partial z_k} \frac{\partial z_k}{\partial w_j}\right) \frac{\partial w_j}{\partial v}\right)}^{\text{repeated terms}} \frac{\partial v}{\partial \theta_i}$$

$$\frac{\partial f}{\partial \theta_k} = \left(\sum_j \frac{\partial f}{\partial w_j} \frac{\partial w_j}{\partial v}\right) \frac{\partial v}{\partial \theta_k} = \left(\sum_j \left(\sum_k \frac{\partial f}{\partial z_k} \frac{\partial z_k}{\partial w_j}\right) \frac{\partial w_j}{\partial v}\right) \frac{\partial v}{\partial \theta_k}$$

(Shannon) Entropy:

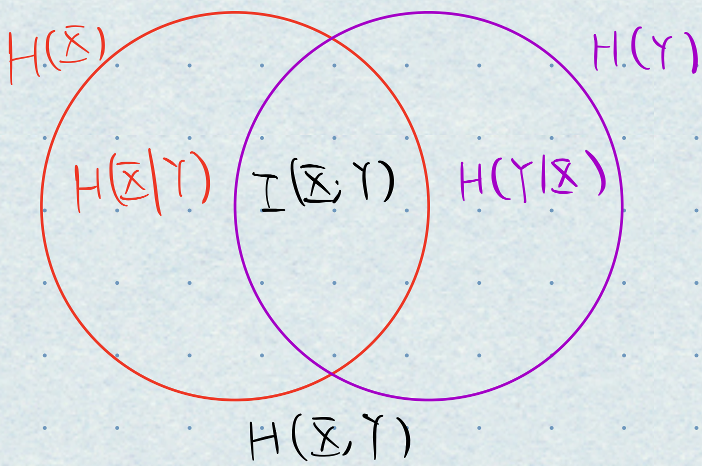
$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

Conditional entropy:

$$H(Y|X) = \sum_{x} P_X(x)\, H(Y|X=x)$$

$$H(Y|X=x) = -\sum_{y} P_{Y|X}(y|x) \log P_{Y|X}(y|x)$$

$$\Rightarrow H(Y|X) = \sum_{x,y} P(x,y) \log \frac{P(x)}{P(x,y)}$$




Chain rule: $H(X_1, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \dots, X_{i-1})$

Mutual information

$$I(X;Y) = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)\,P_Y(y)}$$

$$I(X;Y) = D_{KL}(P_{X,Y} \,||\, P_X P_Y)$$

Relative entropy (KL divergence)

$$D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Cross entropy

$$CE(P,Q) = -\sum_x P(x) \log Q(x)$$

$$= H(P) + D_{KL}(P||Q)$$

Diversity: Hill number

$${}^{q}D = \left(\sum_{i=1}^{R} p_i^{q}\right)^{\frac{1}{q-1}}$$

$${}^{1}D = \exp\left\{-\sum_{i=1}^{R} p_i \ln(p_i)\right\}$$ Shannon index

# Gradient descent methods in NN

Cost function : $J(\theta)$ , $\theta \in \mathbb{R}^d$ NN params

Gradient : $\nabla_\theta J(\theta) = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \vdots \\ \frac{\partial J}{\partial \theta_d} \end{bmatrix}$

Generally speaking, $\theta^{(t+1)} = \theta^{(t)} - \Delta^{(t)}$

① Gradient descent

$$\Delta^{(t)} = \eta \nabla_\theta J(\theta^{(t)})$$

$\eta$ : learning rate

② Momentum

$$\Delta^{(t)} = \gamma \Delta^{(t-1)} + \eta \nabla_\theta J(\theta^{(t)})$$

$\gamma$ : decay factor, 0.9

③ Adagrad

$$\Delta^{(t)} = \frac{\eta}{\sqrt{G^{(t)} + \varepsilon I}} \nabla_\theta J(\theta^{(t)})$$

$$G^{(t)} = \sum^{t} \begin{bmatrix} \frac{\partial J}{\partial \theta_1} & & \\ & \ddots & \\ & & \end{bmatrix}^2$$

$$i=1 \qquad \frac{\partial J}{\partial \theta_d}\Bigg]_{\theta=\theta^{(i)}}$$

④ RMSprop

$$\Delta^{(t)} = \frac{\eta}{\sqrt{G^{(t)} + \varepsilon I}} \nabla_\theta J(\theta^{(t)})$$

$$G^{(t)} = \underbrace{\gamma}_{0.9} G^{(t-1)} + (1-\gamma) \begin{bmatrix} \frac{\partial J}{\partial \theta_1} & & \\ & \ddots & \\ & & \frac{\partial J}{\partial \theta_d} \end{bmatrix}^2_{\theta=\theta^{(t)}}$$

⑤ Adam

$$m^{(t)} = \beta_1 m^{(t-1)} + (1-\beta_1) \nabla_\theta J(\theta^{(t)})$$

$$v^{(t)} = \beta_2 v^{(t-1)} + (1-\beta_2) (\nabla_\theta J(\theta^{(t)}))^2$$

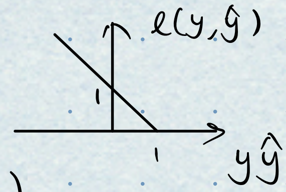
element wise

$$\hat{m}^{(t)} = \frac{m^{(t)}}{1-\beta_1^t}, \quad \hat{v}^{(t)} = \frac{v^{(t)}}{1-\beta_2^t}$$

$$\Delta^{(t)} = \frac{\eta}{\sqrt{\hat{v}^{(t)}} + \varepsilon} \hat{m}^{(t)}$$

Typical values: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$

## Hinge loss




$(-\infty, +\infty)$

Binary : $\ell(y, \hat{y}) = \max(0, 1 - y\hat{y})$

$\{\pm 1\}$

N-class : $(-\infty, +\infty)$

$$\hat{y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N)$$

$$\ell(y, \hat{y}) = \sum_{i \neq y} \max(0, 1 + \hat{y}_i - \hat{y}_y)$$

## Hinge embedding loss

$$\ell(x, y) = \begin{cases} x & \text{if } y = 1 \\ \max(0, \Delta - x) & \text{if } y = -1 \end{cases}$$

$x$ measures the distance between a pair instances

$y = 1/-1$ means the pair is similar/dissimilar

## Cross entropy loss

$\{0,1\}$ $(0,1)$

Binary : $\ell(y, \hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$

N-class : $\hat{y} = (\hat{y}_1, \cdots, \hat{y}_N)$ $\leftarrow (0,1)$

$$\ell(y, \hat{y}) = -\sum_{i=1}^{N} \mathbb{1}_{y=i} \log \hat{y}_i$$

## Logistic loss

$\{\pm 1\}$

Binary : $\ln(1 + e^{-y\hat{y}})$ $\leftarrow$ equal to CE loss applied on NN output with $\sigma(\cdot)$