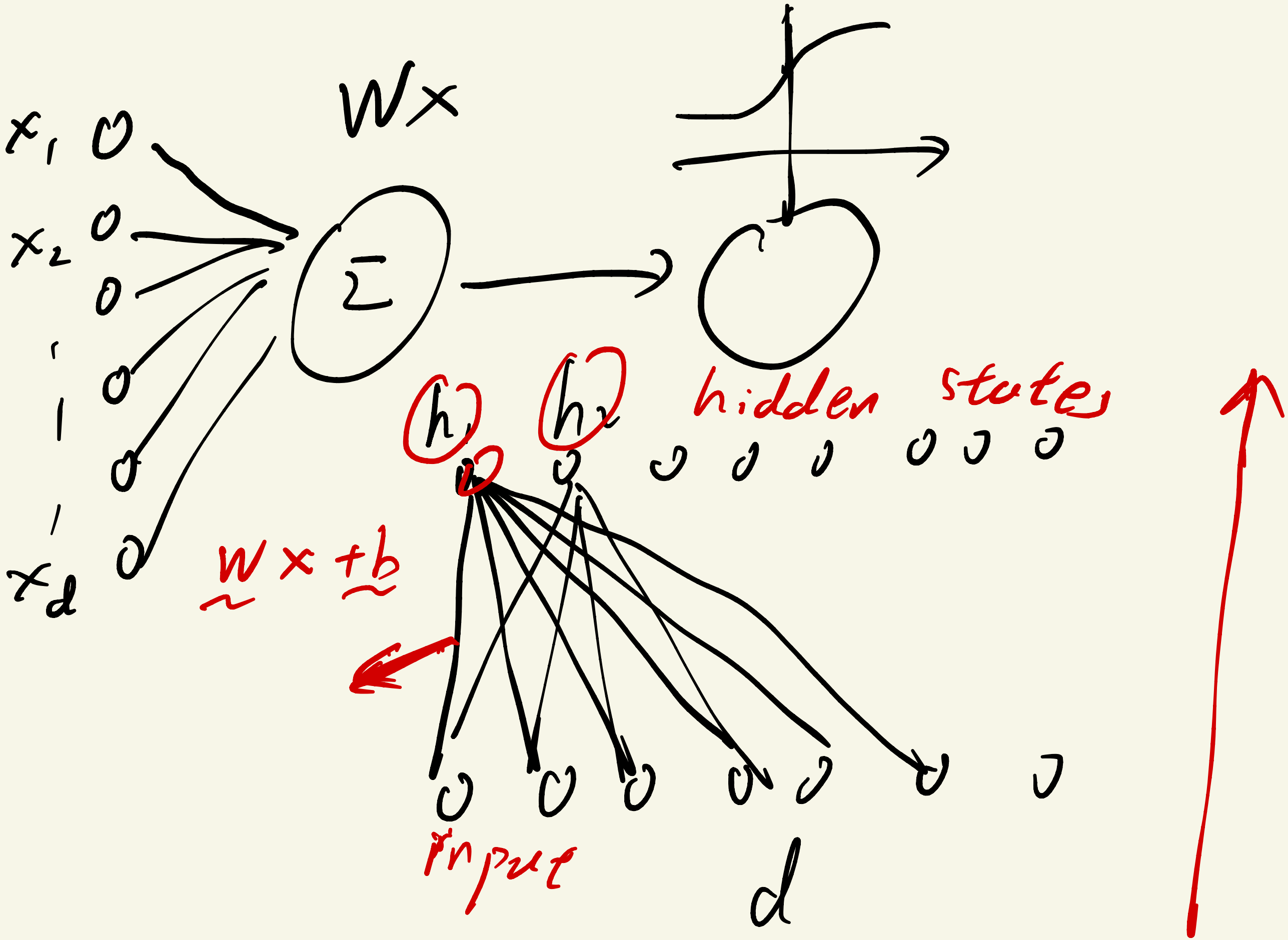
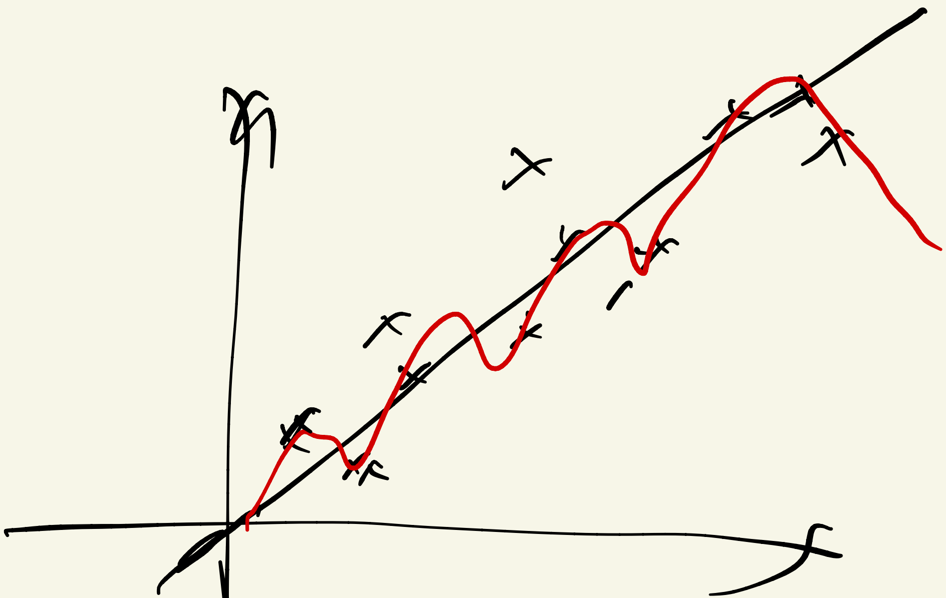
Lecture 18 NN, back propagation



x_1
x_2
x_d
Wx
Σ
h_1
h_2
hidden states
$Wx+b$
input
d

non-linear → non-convex

over parameterized



15

100 parameters

HMM → non-convex

$P(S_t | S_{t-1})$ $P(O_t | S_t)$

MSE

$$\frac{1}{2}(y - f(x))^2$$

$$0 \to f(x) \to$$

$$\frac{1}{1+e^{-f(x)}}$$

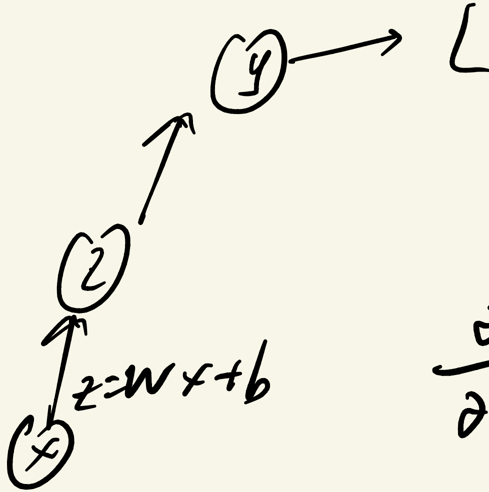
$L \rightarrow$ loss

$w \rightarrow$ weight

$\alpha \rightarrow$ learning rate

$$net = \sum_{i=0}^{n} w_i x_i$$

$$\frac{\partial o}{\partial w_i} = \frac{\partial o}{\partial net} \cdot \frac{\partial net}{\partial w_i}$$



$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w}$$

fan-out

$$
\begin{tikzcd}
 & x \arrow[dr] & \\
t \arrow[ur] \arrow[dr] & & f \\
 & y \arrow[ur] &
\end{tikzcd}
\qquad \frac{\partial f}{\partial t}
$$

$$
\frac{\partial f}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial f}{\partial y} \cdot \frac{dy}{dt}
$$

$$\frac{\partial L}{\partial w_{11}}$$

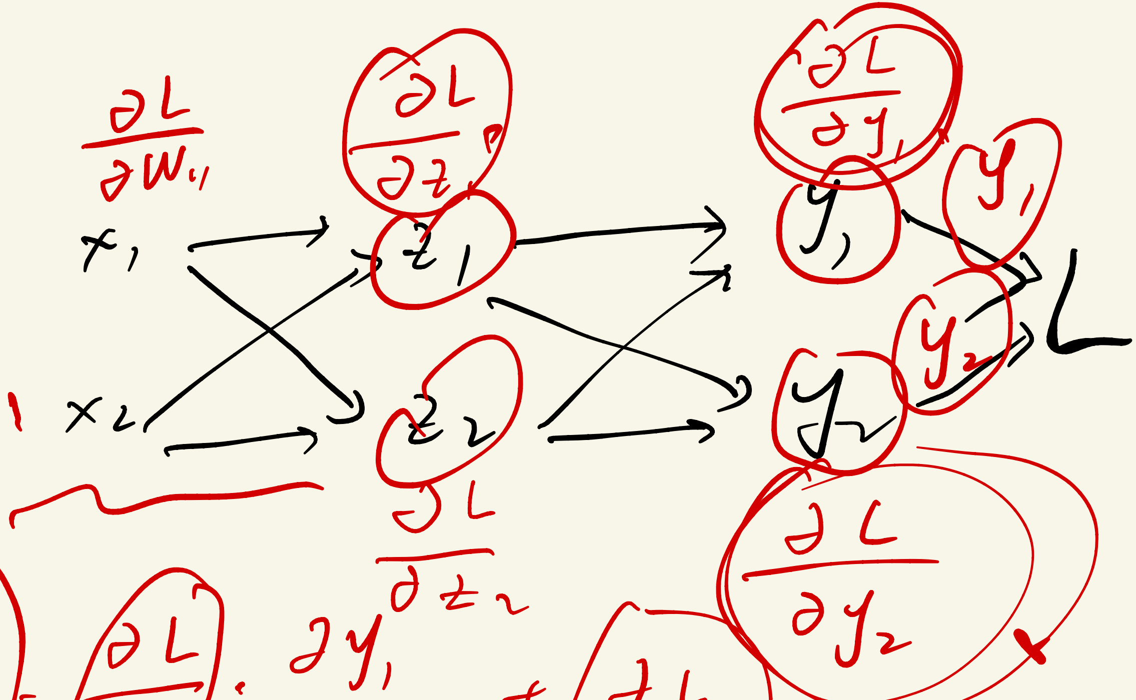
$$\frac{\partial L}{\partial y_1} = -\frac{t_1}{y_1} \qquad \frac{\partial L}{\partial y_2} = \frac{-t_2}{y_2}$$

$$\frac{\partial L}{\partial y_1} \qquad \frac{\partial L}{\partial y_2} \qquad \frac{\partial L}{\partial z_1} \qquad \frac{\partial y_1}{\partial z_1} = y_1 \cdot (1 - y_1)$$

$$\frac{\partial L}{\partial w_{12}}$$

$$\frac{\partial L}{\partial w_{11}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{11}} \rightarrow x_1$$

$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial y_1} \cdot \frac{\partial y_1}{\partial z_1} + \frac{\partial L}{\partial y_2} \cdot \frac{\partial y_2}{\partial z_2}$$



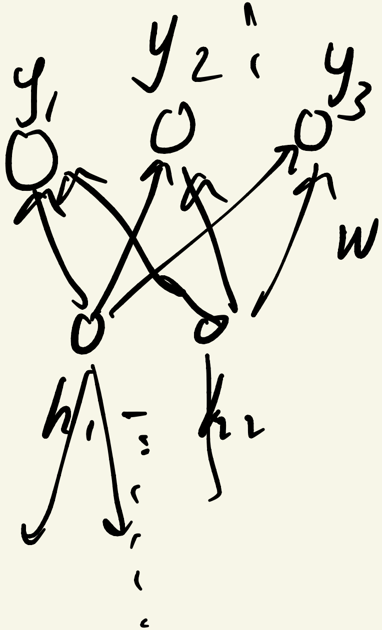
$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial y_1} \cdot \frac{\partial y_1}{\partial z_1} + \frac{\partial L}{\partial y_2} \cdot \frac{\partial y_2}{\partial z_2}$$

$$\frac{\partial L}{\partial W_{11}} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z}{\partial W_{11}}$$

$$\sum_j w_{ij} x_j$$

2017 torch dynet caffee

sigmoid



$$\vec{y} = W\vec{h}$$

$$\frac{\partial L}{\partial y_1} \quad \frac{\partial L}{\partial y_2} \quad \frac{\partial L}{\partial y_K}$$

$$\frac{\partial L}{\partial h_i} = \underbrace{\sum_k \bar{y}_k W_{ki}}$$

$$\underline{O(K^2)}$$

$$\frac{\partial L}{\partial w_1 \partial w_2}$$

$$\frac{\partial^2 L}{\partial w_1^2}$$

$\boxed{\text{Hessian}}$

$$\underbrace{\hat{E}_{x\sim P_{data}}}$$

i.i.d

$$x_1 \quad \cdots \quad \overset{\text{training}}{x_N}$$

unbiased samples from P_{data}

MCMC

$$E_{x \sim p} f(x) \leftarrow \frac{1}{N}$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

$\downarrow$

$n = 32$

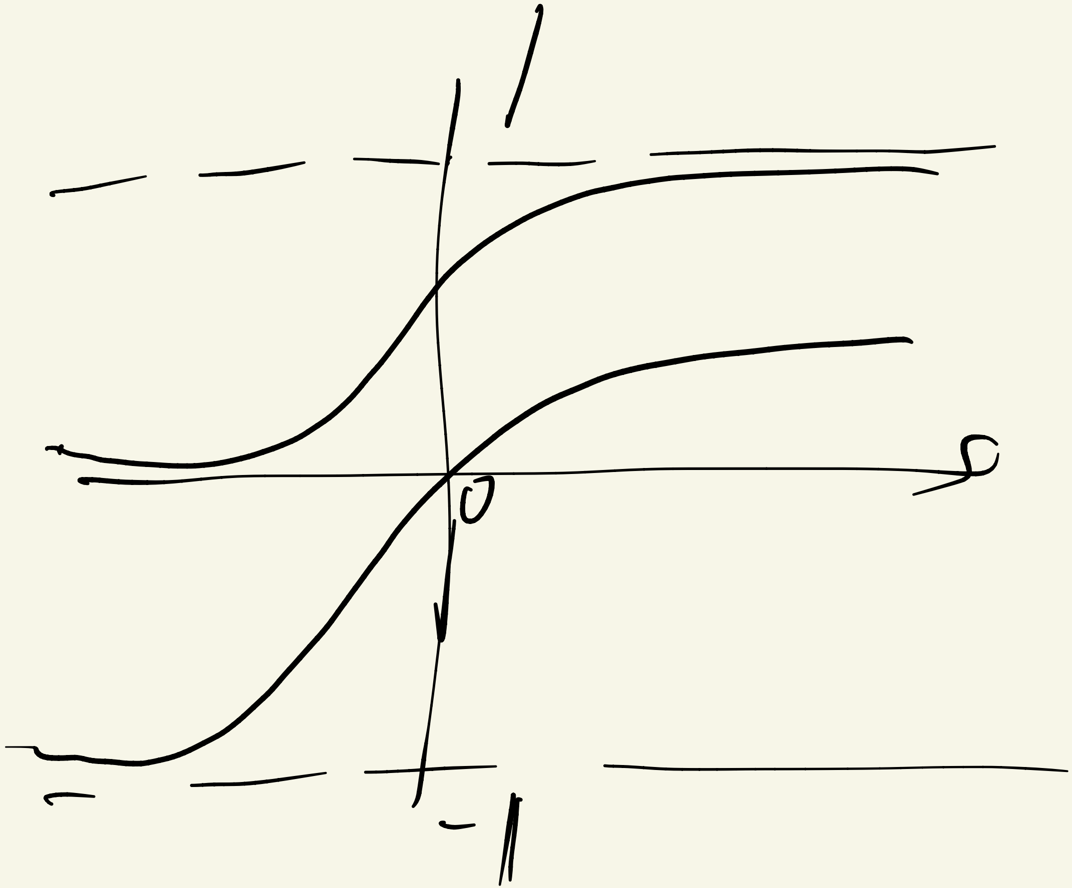
$n = 1$

dropout

MCMC

(un)biased estimator

Variance



1
0
-1

$y = \max(0, x)$

fully connected