

Intro
Statistical
LEARNING

R^2

~~Density Estimation~~

机器学习与量化交易实战

第六课

$$\frac{\partial y}{\partial w}$$

$K_{\text{msprop}} \approx \text{SGD}$

X_{10}

Outline



训练集

优化

Feature
selection

• 特征选择

• 遗传算法

• 深入理解BP算法

• RNN

RMS1

上次作业cont.

周六 最后一节课 PPT

最后一个数据集

benchmark R^2 on

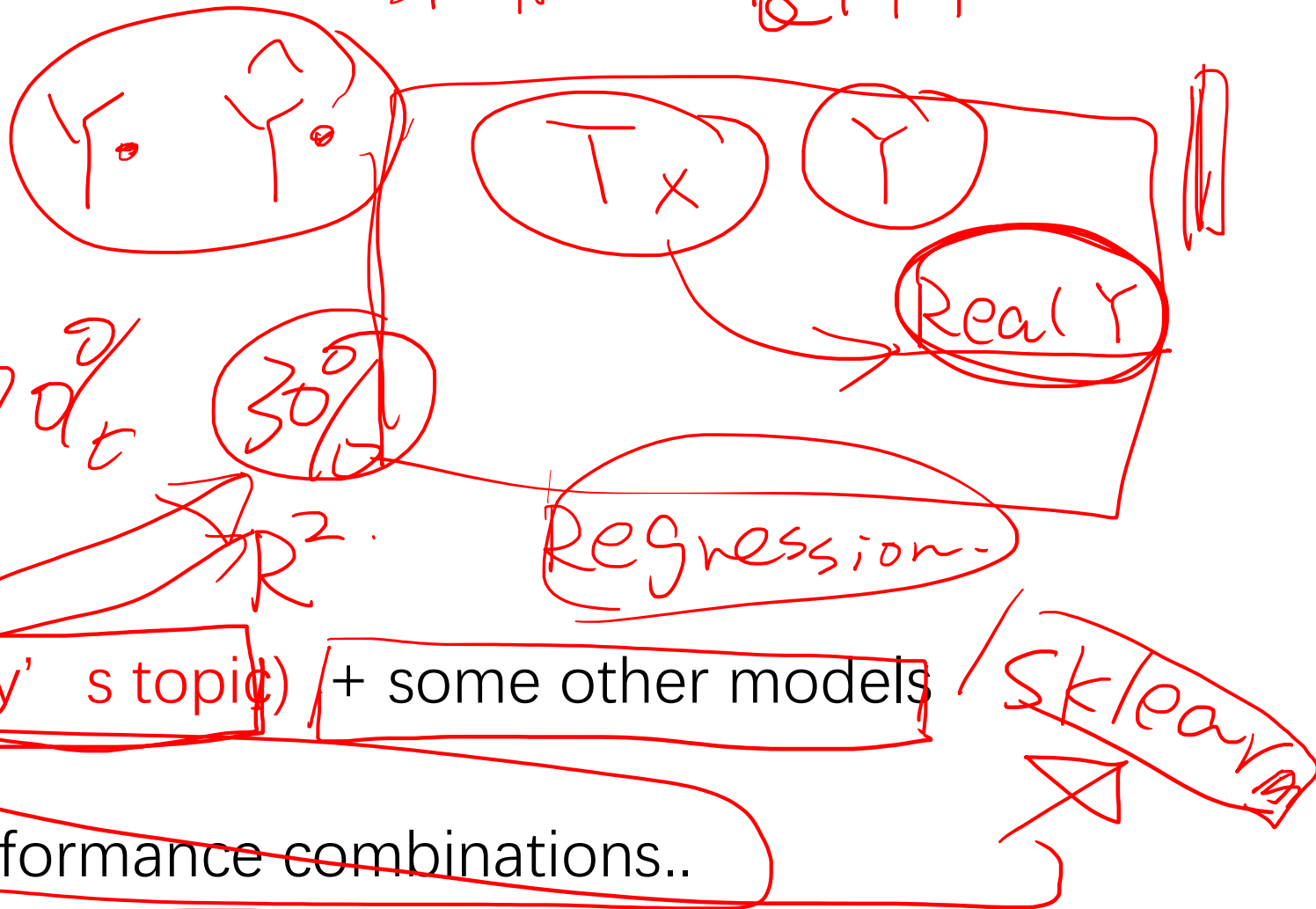
KNN

And OLS

Try:

Feature Selection (today's topic) + some other models

..and report the best performance combinations..



$\vec{X} = [X_1, \dots, X_i, \dots, X_p]$

• **Subset Selection.** We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.

• **Shrinkage.** We fit a model involving all p predictors, but the estimated coefficients are shrunk towards zero relative to the least squares estimates. This shrinkage (also known as **regularization**) has the effect of reducing variance and can also perform variable selection.

• **Dimension Reduction.** We project the p predictors into a M -dimensional subspace, where $M < p$. This is achieved by computing M different **linear combinations**, or **projections**, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

$$f(X_{300 \times 1}) \rightarrow Z_{20}$$

Subset Selection

Best subset and stepwise model selection procedures

Handwritten notes in red ink:

	1	2	3
		300	
		2	3
	50	(300)	(300)
		2	3

Below the table, there is a red checkmark and a red arrow pointing to the right.

Best Subset Selection

1. Let \mathcal{M}_0 denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \dots, p$:
 - (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
 - (b) Pick the best among these $\binom{p}{k}$ models, and call it \mathcal{M}_k . Here *best* is defined as having the smallest RSS, or equivalently largest R^2 .
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

Forward Stepwise Selection

- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest *additional* improvement to the fit is added to the model.

In Detail

Backward

Forward Stepwise Selection

$\{X_1, \dots, X_p\}$

$\{X_8\}$

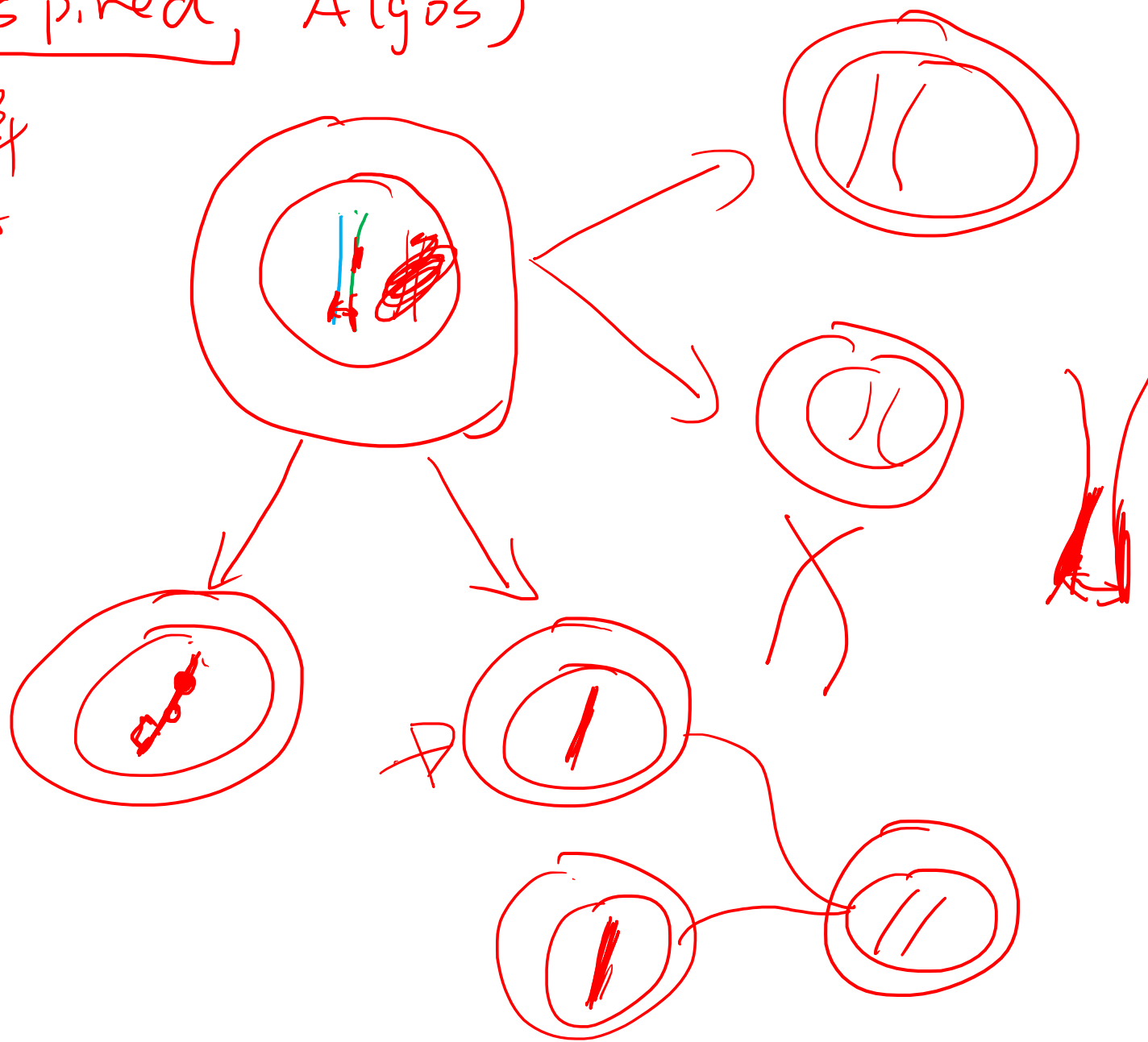
$(\begin{smallmatrix} 0 \\ 23 \end{smallmatrix})$

1. Let \mathcal{M}_0 denote the *null* model, which contains no predictors.
2. For $k = 0, \dots, p-1$:
 - 2.1 Consider all $p - k$ models that augment the predictors in \mathcal{M}_k with one additional predictor.
 - 2.2 Choose the *best* among these $p - k$ models, and call it \mathcal{M}_{k+1} . Here *best* is defined as having smallest RSS or highest R^2 .
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2 .

① GA (inspired, Algos)

1. 种群

2. DNA



ATCG
V (Base4)

0101.

Y

X₁ X₃₀₀

→ 表现型

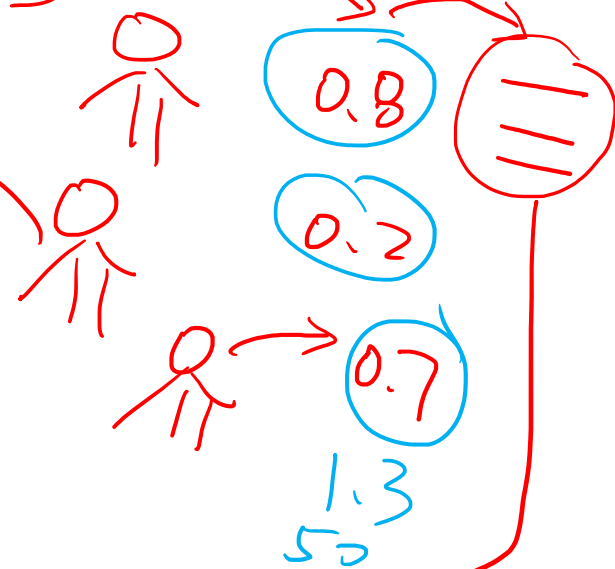
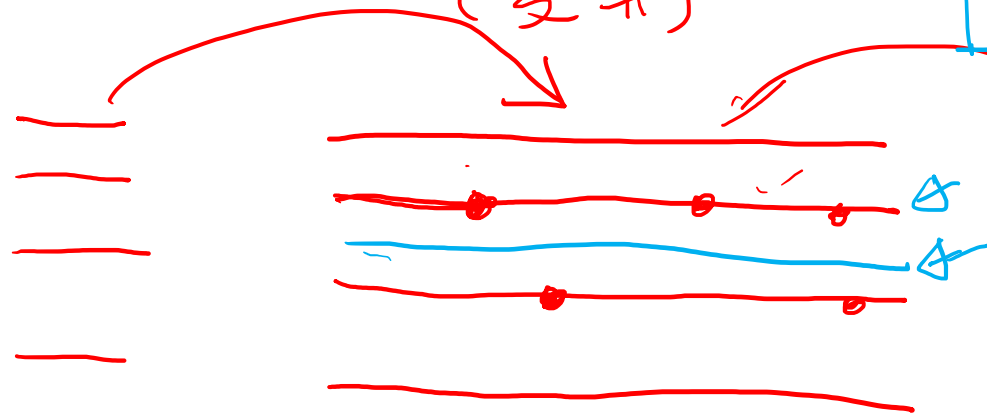
MUTATION
(变异)

→ [1 0011001]
1.2 53 3.7

→ 基因型

CROSSOVER

FITNESS



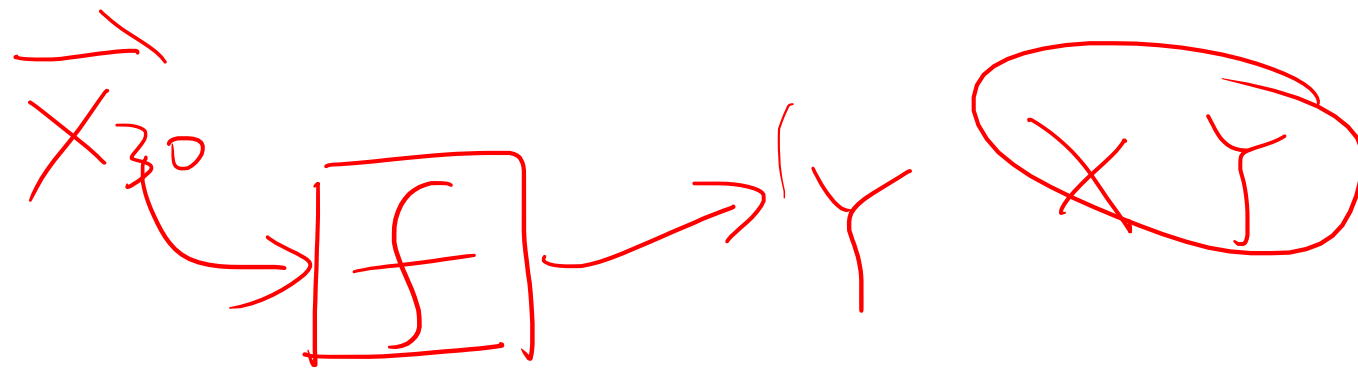
Population

P = 0.05%

Step Size

f()

R²



目的 $\vec{x}_{30} = ?$ Y 最大

1000 { $[0.1 \ 1.2 \ \dots \ 9.9]_{30}$ $f(\vec{x})$

\sum $\overset{N}{001001}$
 \uparrow \uparrow

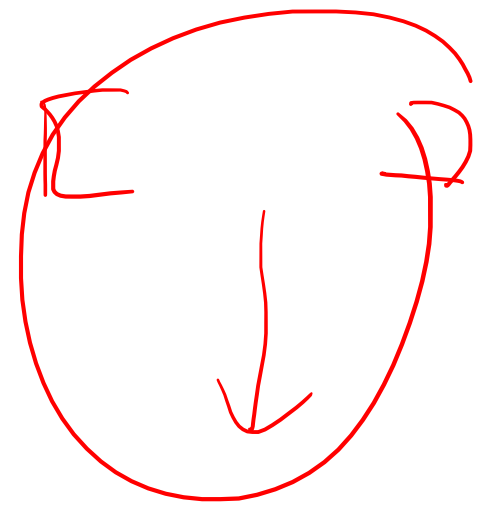
\perp

$-$

$=$

$\frac{P_{t+1}}{R_{t+1}}$

\downarrow



GOAL

X

~~f(x)~~ x^2

10
25
0 x^*

max
min
 x

FITNESS (x)

100 - x [10000]

1
100
625
0

1. 根元部.



1 —

2 —

3 —

4 —

5 —

6 —

FITNESS

(1.2)

0.3

0.1

1.5

⋮

0 1
Random Num
(r.n - 0.6) ~~0~~ 0

np.random.rand

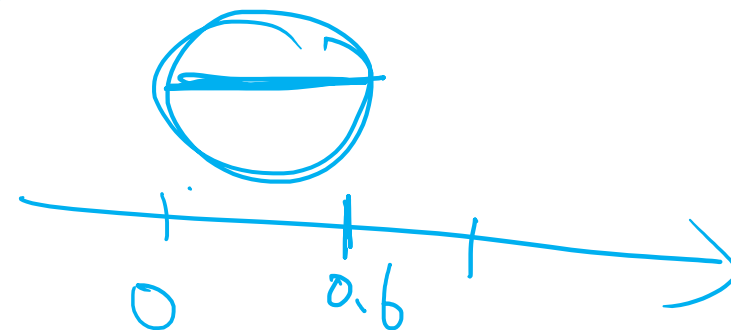
[0.1]

P₁

P₂

$$P_1 = \frac{f_1}{\sum f_i} = 0.6$$

$$P_2 = \frac{f_2}{\sum f_i}$$

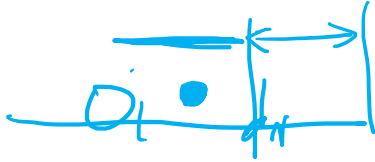


Implementation ~~Q.D~~
Trick

60%

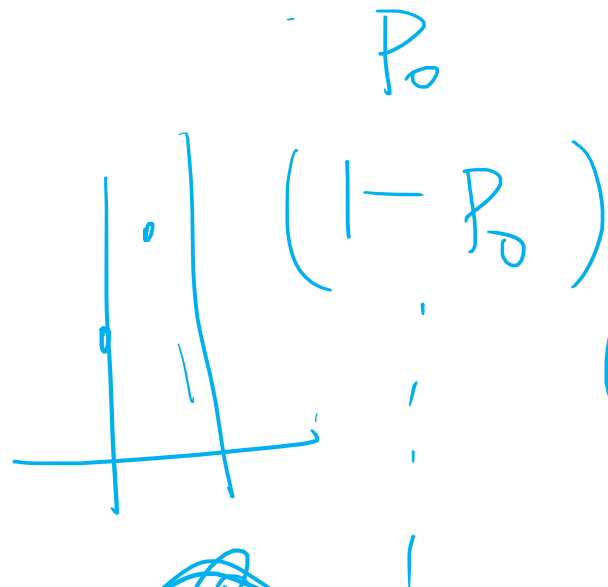
X_1

[]



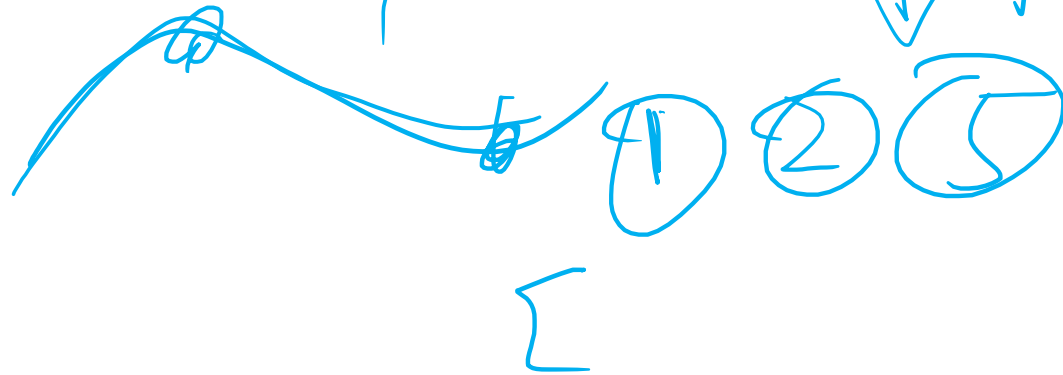
RANK.

DIVERSITY



var

Entropy



$T = \text{"HELLO WORLD."}$

$f(S) = \# \text{ Char which is correct}$

[0] [25] [6] -

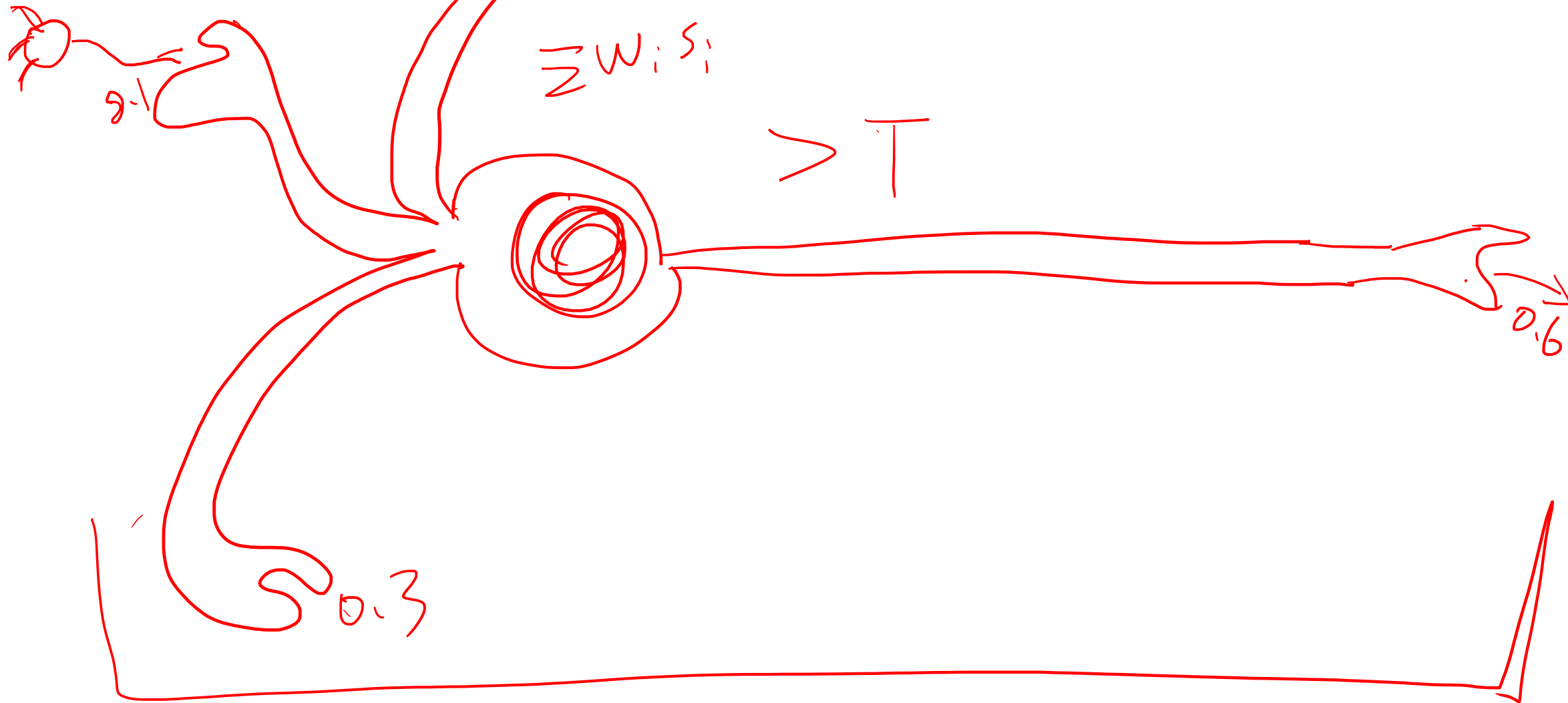
0~24

$Y = X^2$

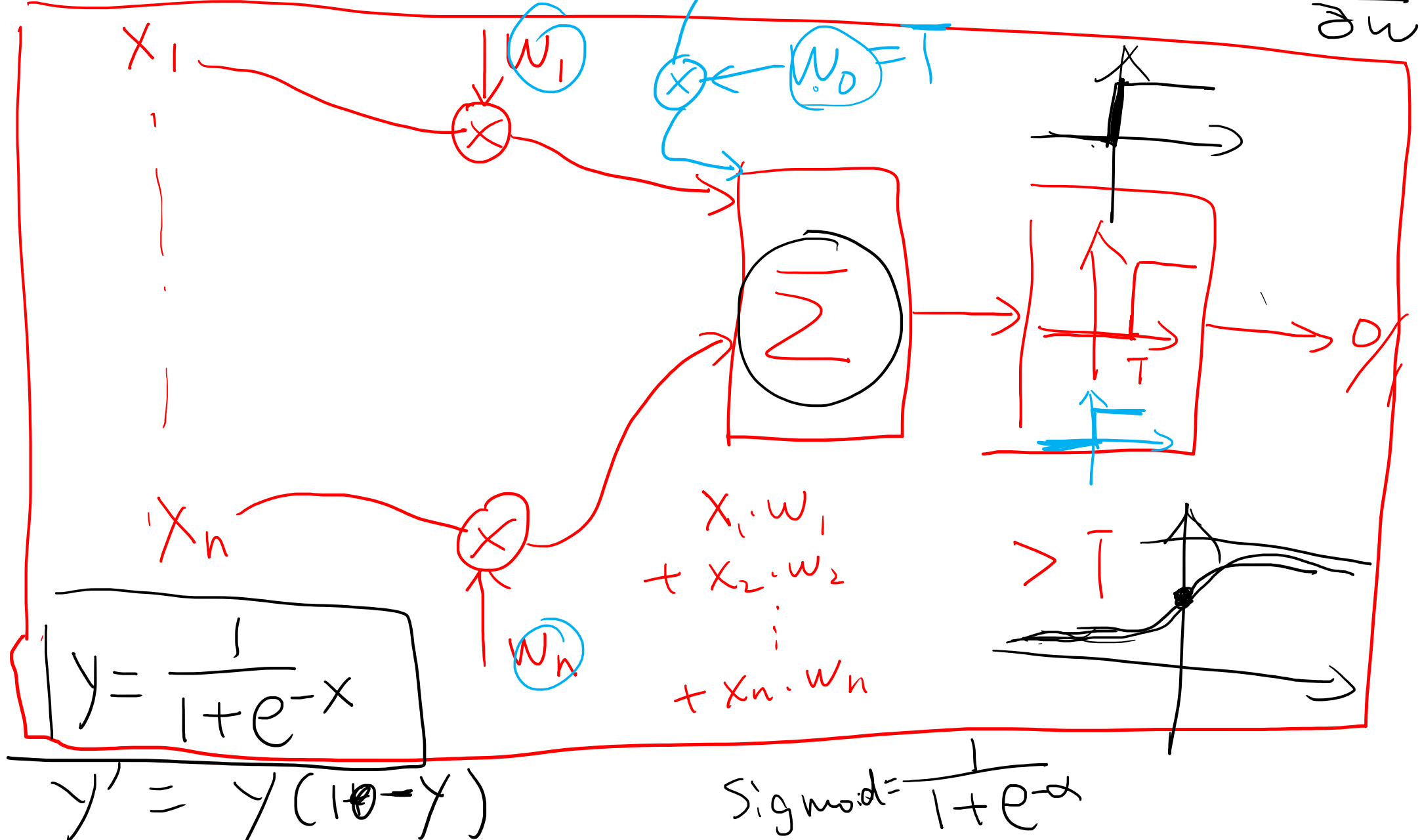
PPPOS

||NIPS||

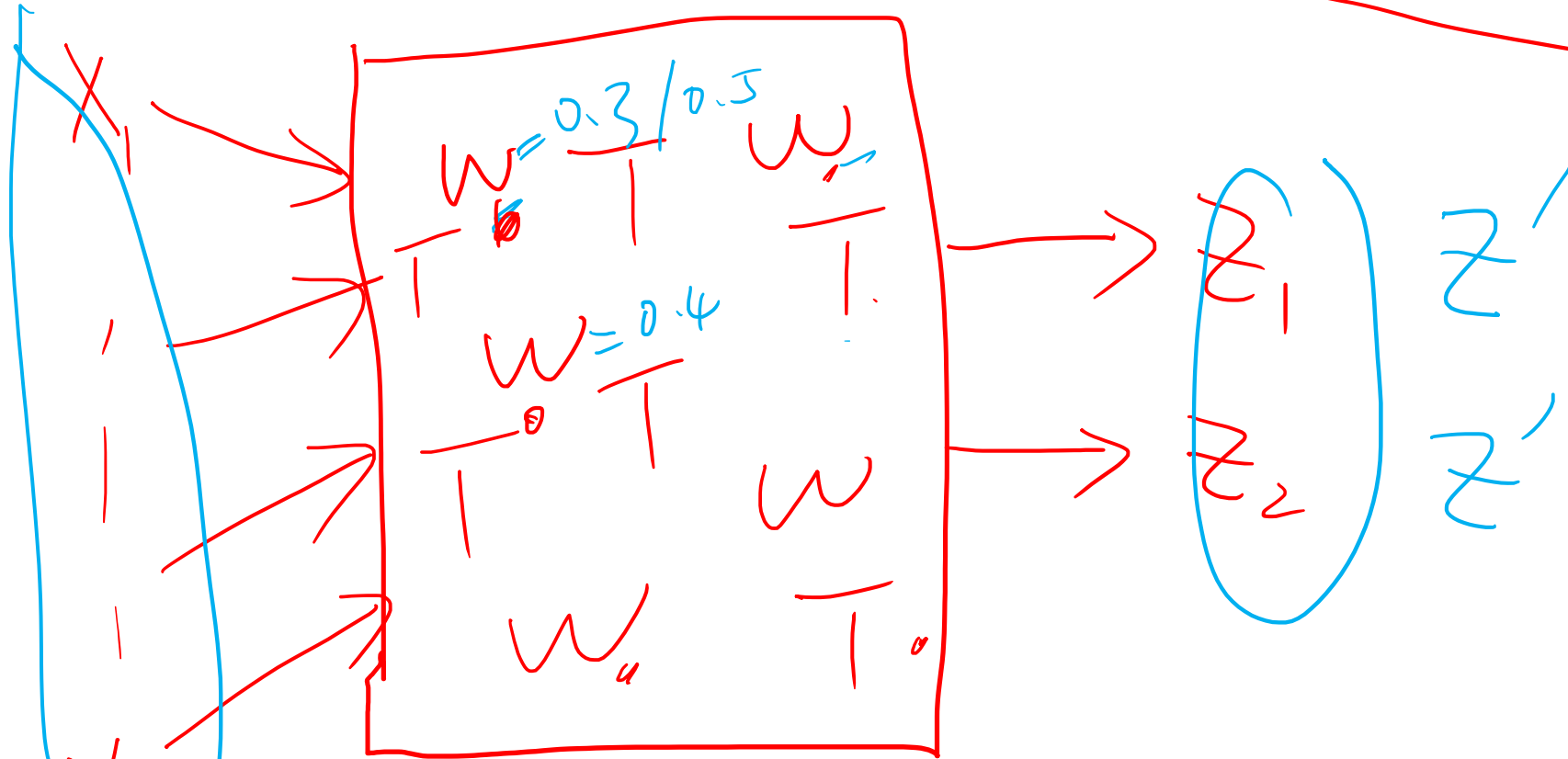
ANN



x_{300} y_1 bias term $\frac{\partial P}{\partial u}$



$$\vec{z} = f(\vec{x}, \vec{w}, T)$$



$$\frac{\partial P}{\partial w}$$

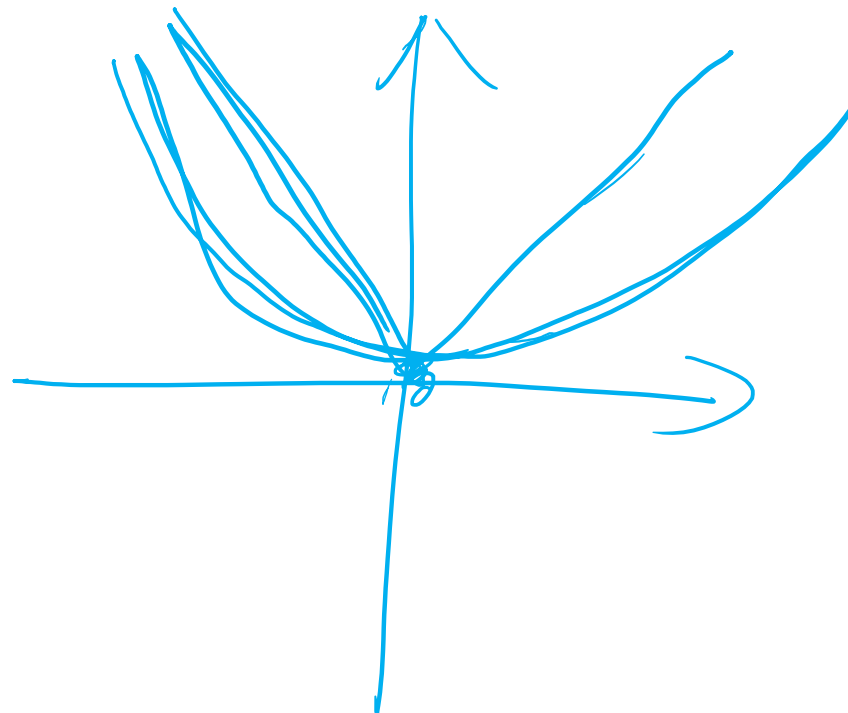
$$\vec{z} = f(\vec{w}, \vec{x})$$

$$P = ||z - d||$$

$$\begin{cases} w_1 = 0.2 \\ w_2 = 0.8 \end{cases}$$

$$\vec{d} = g(\vec{x})$$

$$P = \frac{1}{2} || \underbrace{\vec{z}}_{\text{真实}} - \underbrace{\vec{d}}_{\text{预测}} ||^2$$

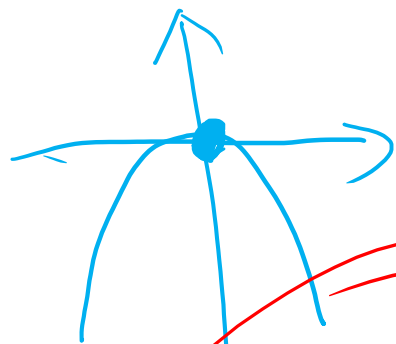


$$P = -\frac{1}{2} \|z - d\|^2$$

$$\frac{\partial P}{\partial w_1} = 1.2$$

w

$$\frac{\partial P}{\partial w}$$



Andrew
NG


SGD

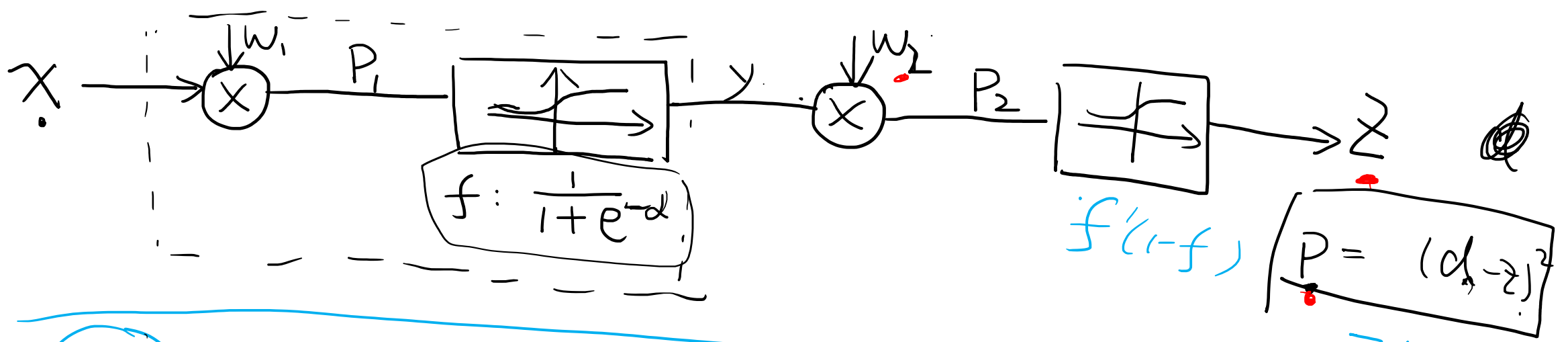
w_2



$$\Rightarrow \left| \begin{array}{l} 1.2 \\ w_1 \end{array} \right| \leftarrow \begin{array}{l} 1.2 \\ w_1 \end{array} - \begin{array}{l} 0.03 \\ \propto \\ 0.03 \end{array} \left(\frac{\partial P}{\partial w_1} \right)$$

$$\boxed{X_{160 \times 1} \quad Y_{1 \times 1}}$$

$$f: X_1 \rightarrow Y_1$$




$$\frac{\partial P}{\partial w_2} = \frac{\partial P}{\partial z} \cdot \frac{\partial z}{\partial w_2} \quad \frac{\partial z}{\partial P_2} \cdot \frac{\partial P_2}{\partial w_2}$$

$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z} \cdot \frac{\partial z}{\partial P_2} \cdot \frac{\partial P_2}{\partial y} \cdot \frac{\partial y}{\partial P_1} \cdot \frac{\partial P_1}{\partial w_1}$$

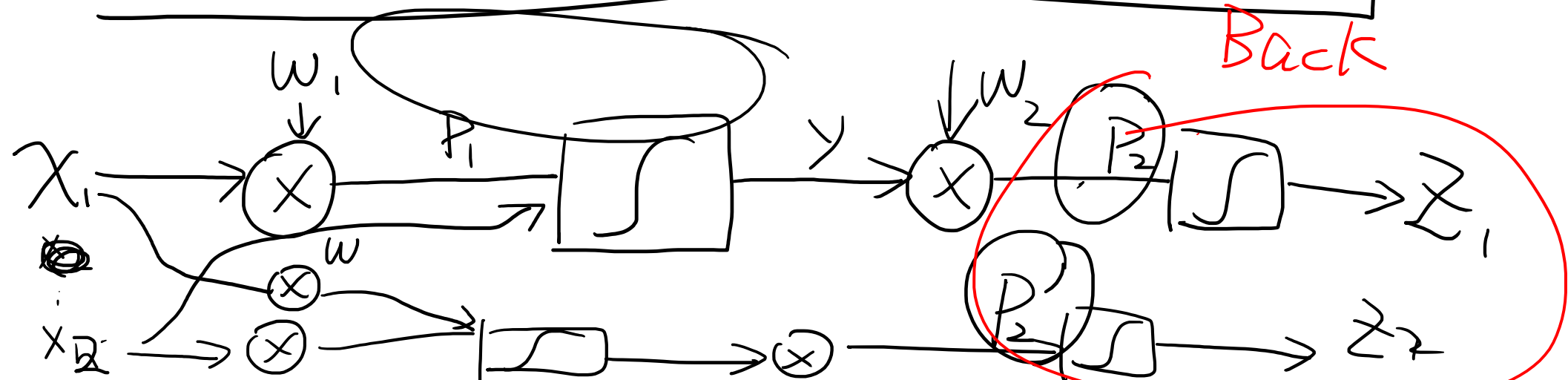
$$P = (d - z)^2$$

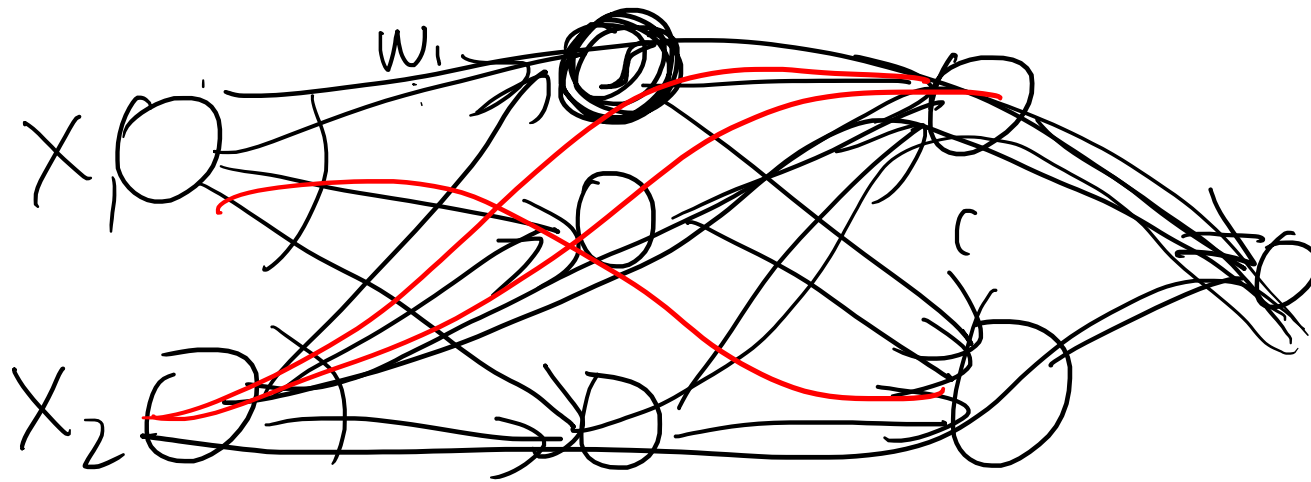
$$\frac{\partial P}{\partial z} = ?$$

$$\frac{dP}{dz} = 2(d - z)$$

$$\frac{\partial P}{\partial w_2} = \frac{\partial P_2}{\partial w_2} \cdot \left(\frac{\partial z}{\partial P_2} \cdot \frac{\partial P}{\partial z} \right)$$

$$\frac{\partial P}{\partial w_1} = \frac{\partial P_1}{\partial w_1} \cdot \frac{\partial y}{\partial P_1} \cdot \frac{\partial P_2}{\partial y} \cdot \left(\frac{\partial z}{\partial P_2} \cdot \frac{\partial P}{\partial z} \right)$$

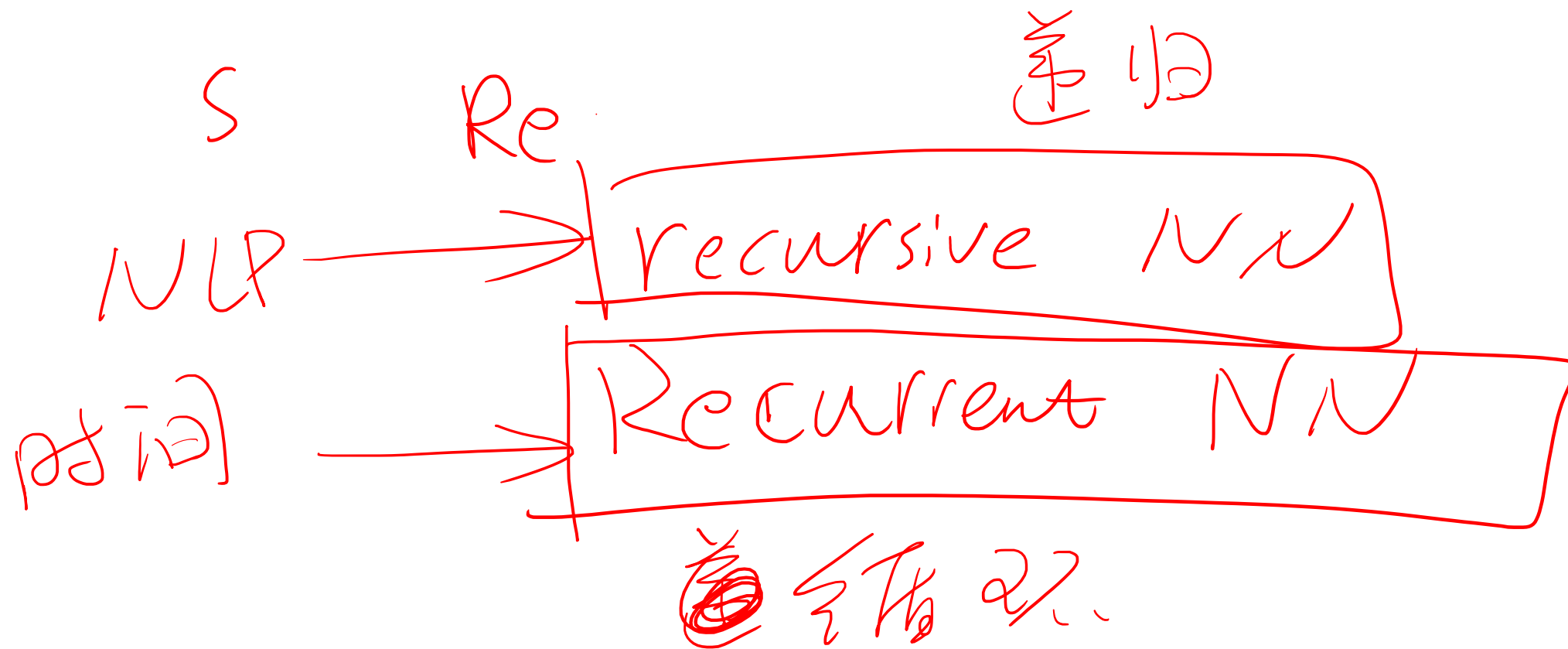


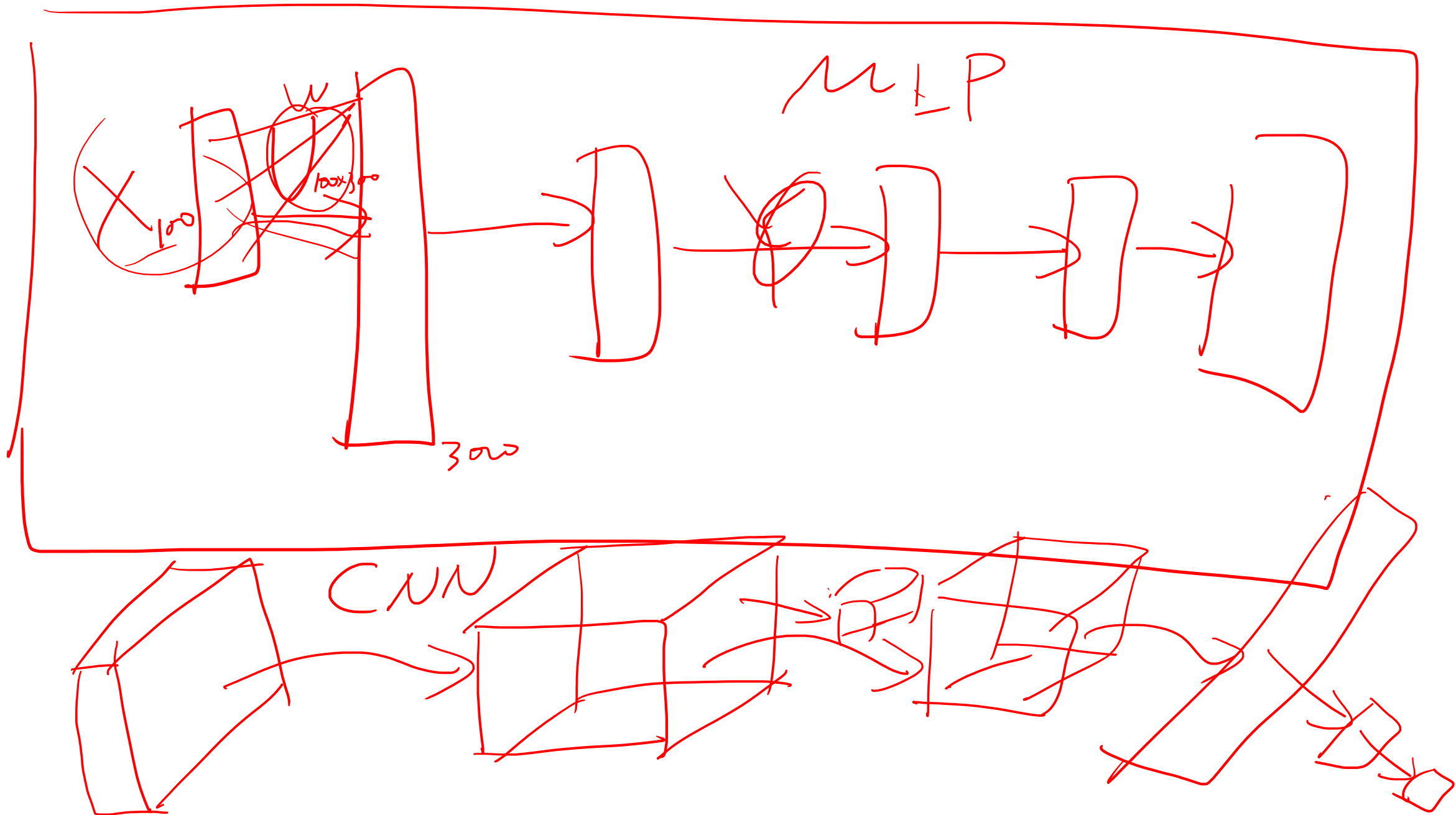


$N \uparrow$ # possible PATH

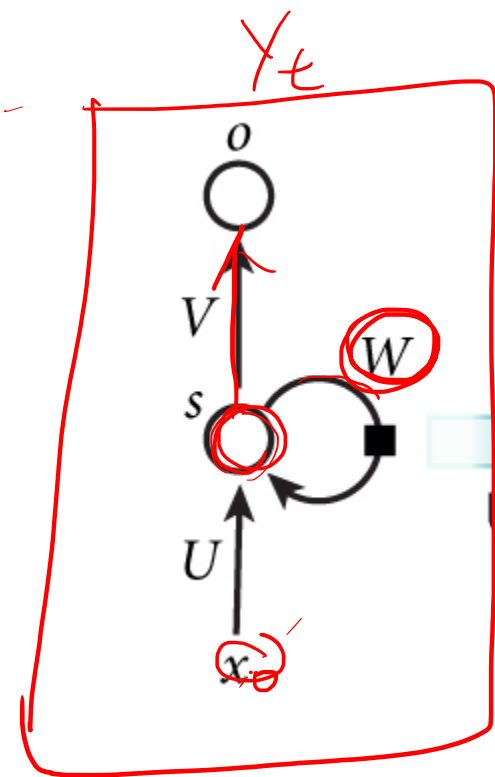
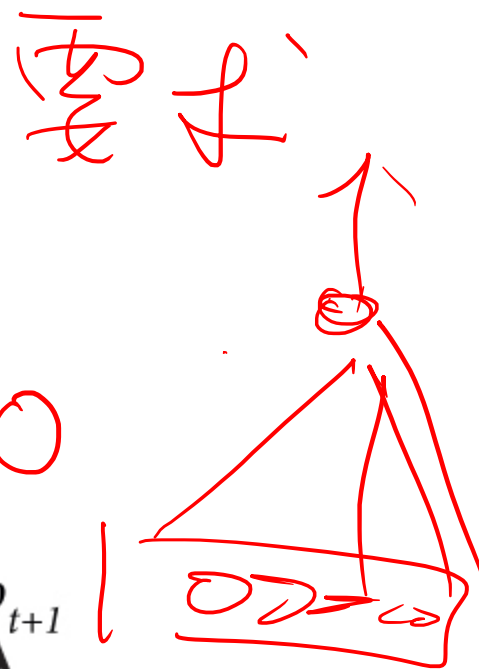
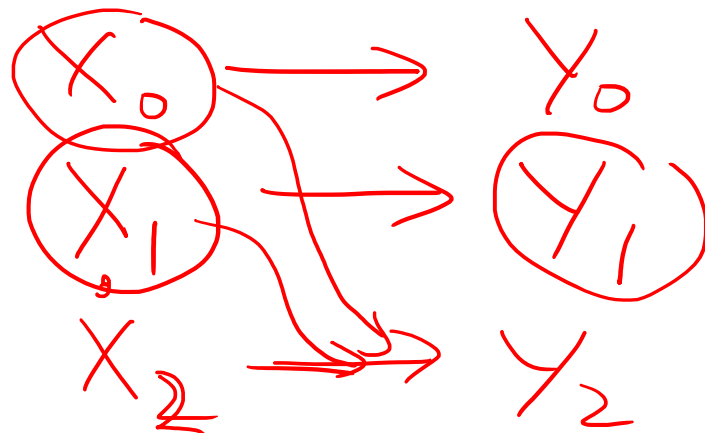
$\uparrow \uparrow \uparrow$

$O(\# \text{ Hidden Layer})$

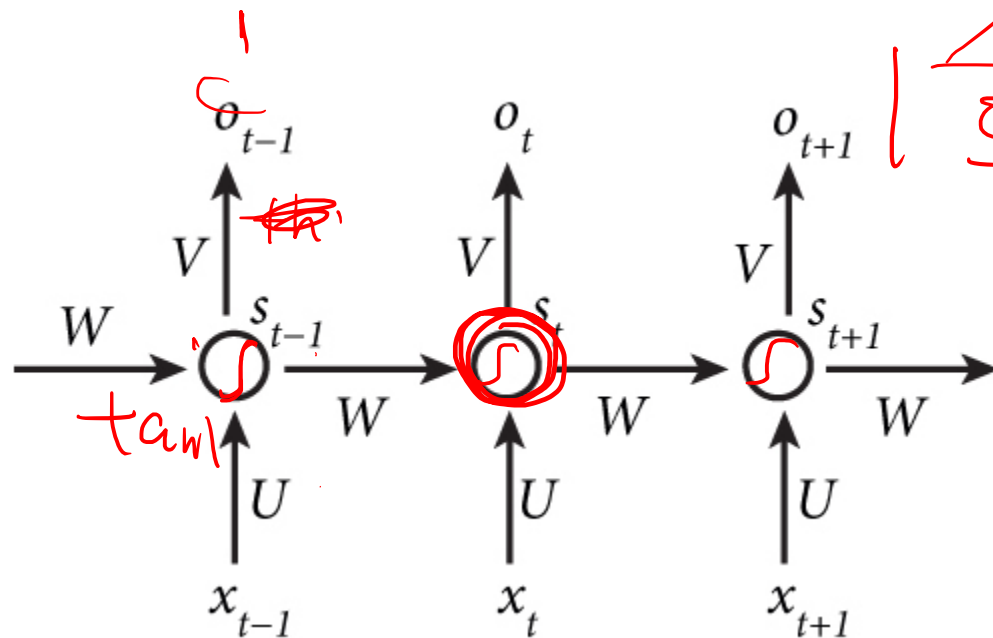




Recurrent
 $N \rightarrow Y_{t+1}$



Unfold



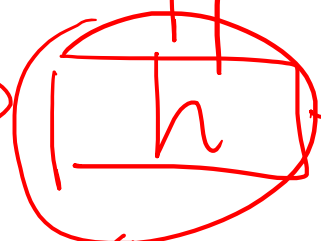
$$\frac{\partial y}{\partial w_1}$$

w_1

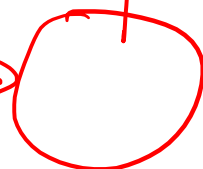
$$\frac{\partial y_{t+3}}{\partial w_1}$$



w



w



w



$$f(w \cdot h_{t-1} + V \cdot x_t)$$

$U_{300 \times h}$

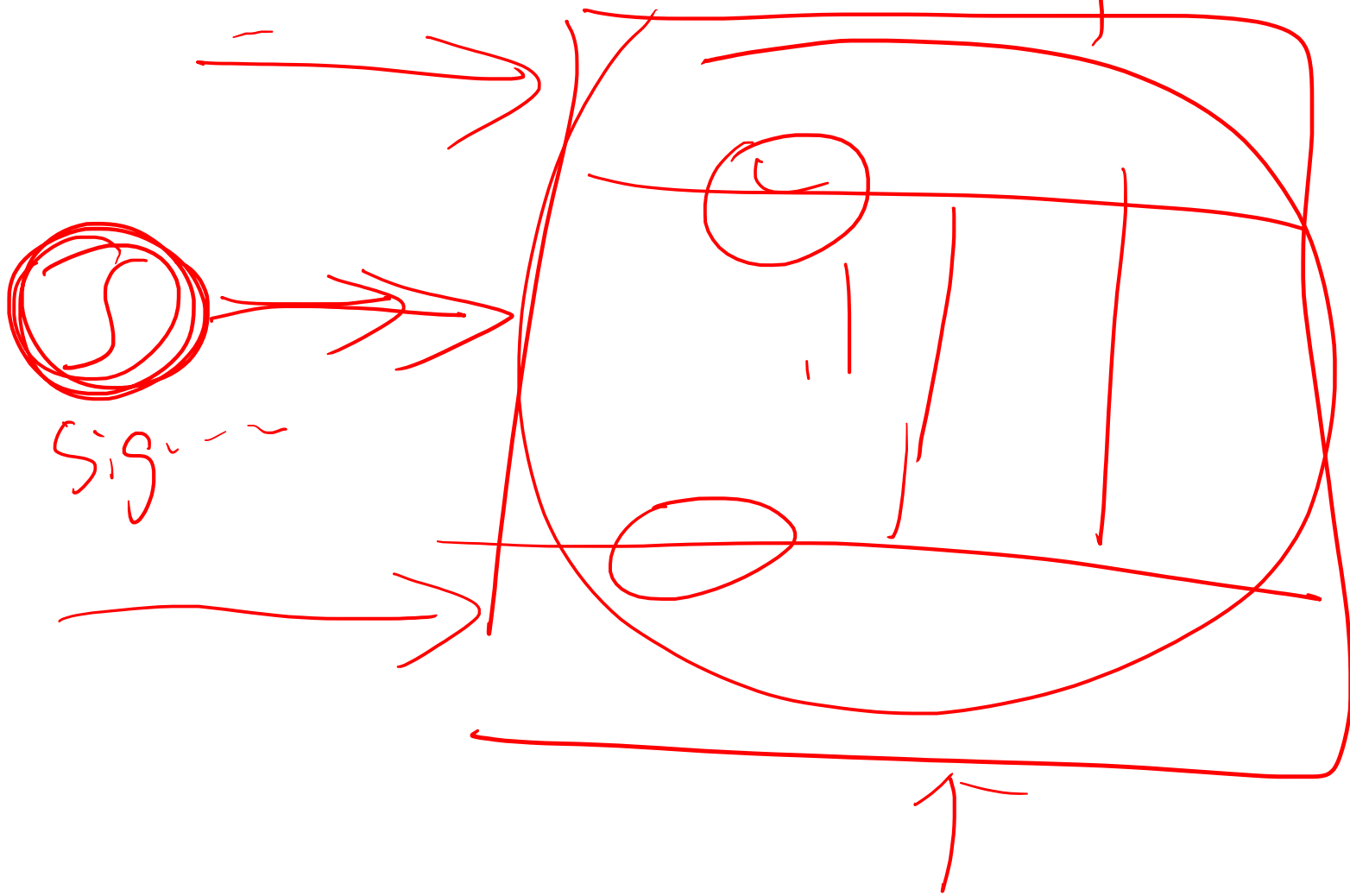
X_{300}

X

Vanishing Gradient

LSTM

4



Keras
LSTM

