# CS 6180 09/22

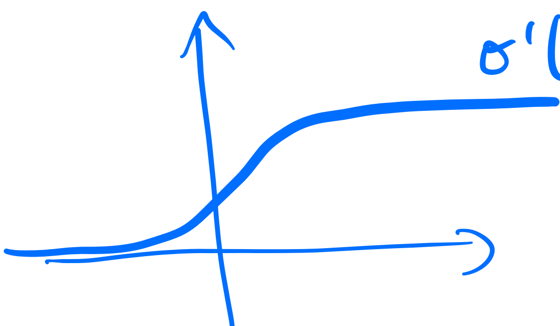
$$\frac{\partial \mathcal{L}^{(i)}}{\partial W_h} = \sum_{j=1}^{i} \frac{\partial \mathcal{L}^{(i)}}{\partial \vec{h}^{(j)}} \frac{\partial \vec{h}^{(j)}}{\partial W_h}\Big|_{(j)}$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial \vec{h}^{(1)}} = \frac{\partial \mathcal{L}^{(i)}}{\partial \vec{h}^{(i)}} \cdot \frac{\partial \vec{h}^{(i)}}{\partial \vec{h}^{(i-1)}} \cdots \frac{\partial \vec{h}^{(2)}}{\partial \vec{h}^{(1)}}$$

$$h^{(i)} = \sigma\big(\underbrace{W_h \vec{h}^{(i-1)} + W_e \vec{e}^{(i)} + \vec{b}_1}_{\vec{a}^{(i)}}\big)$$

$$\frac{\partial \vec{h}^{(i)}}{\partial \vec{h}^{(i-1)}} = \operatorname{diag}\big(\sigma'(\vec{a}^{(i)})\big) \cdot W_h$$

values of derivative of sigmoid



$$\sigma'(z) = \underbrace{\sigma(z)}_{0<} \cdot \underbrace{[1-\sigma(z)]}_{<1}$$

$$0 < \sigma(z) < 1$$

multiplying many derivatives of the sigmoid $(<1)$ leads to a very small gradient

$$\underbrace{\frac{\partial \mathcal{L}^{(i)}}{\partial \vec{h}^{(1)}}}_{\approx 0} = \frac{\partial \mathcal{L}^{(i)}}{\partial \vec{h}^{(i)}} \cdot \underbrace{\frac{\partial \vec{h}^{(i)}}{\partial \vec{h}^{(i-1)}}}_{<1} \cdots \underbrace{\frac{\partial \vec{h}^{(2)}}{\partial \vec{h}^{(1)}}}_{<1}$$

$\Rightarrow$ model is not properly

learning.

but derivative can also rely on $W_h$ which can cause some changes in the learning.

$\Rightarrow$ eigenvalues

same

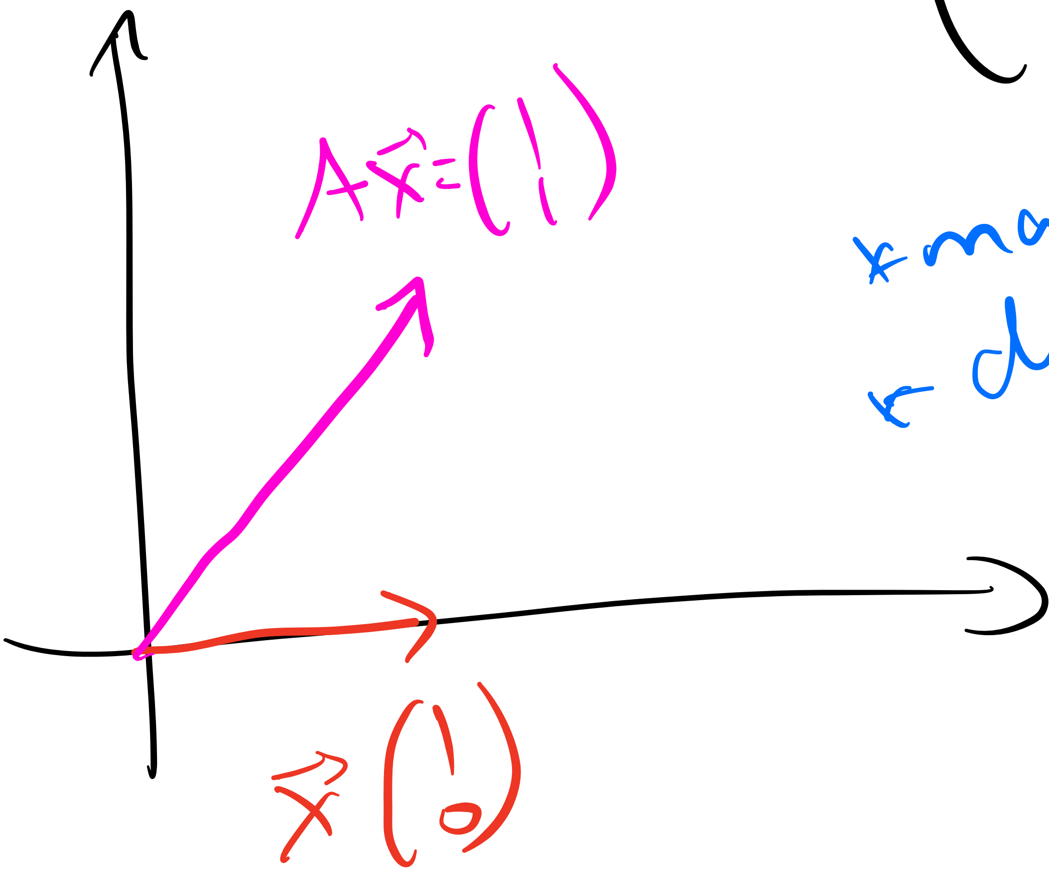
# Background on Eigenvalues

$$A\vec{x} = \lambda \vec{x}$$

lambda: eigenvalue

$$\vec{x} \neq \vec{0}$$

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$



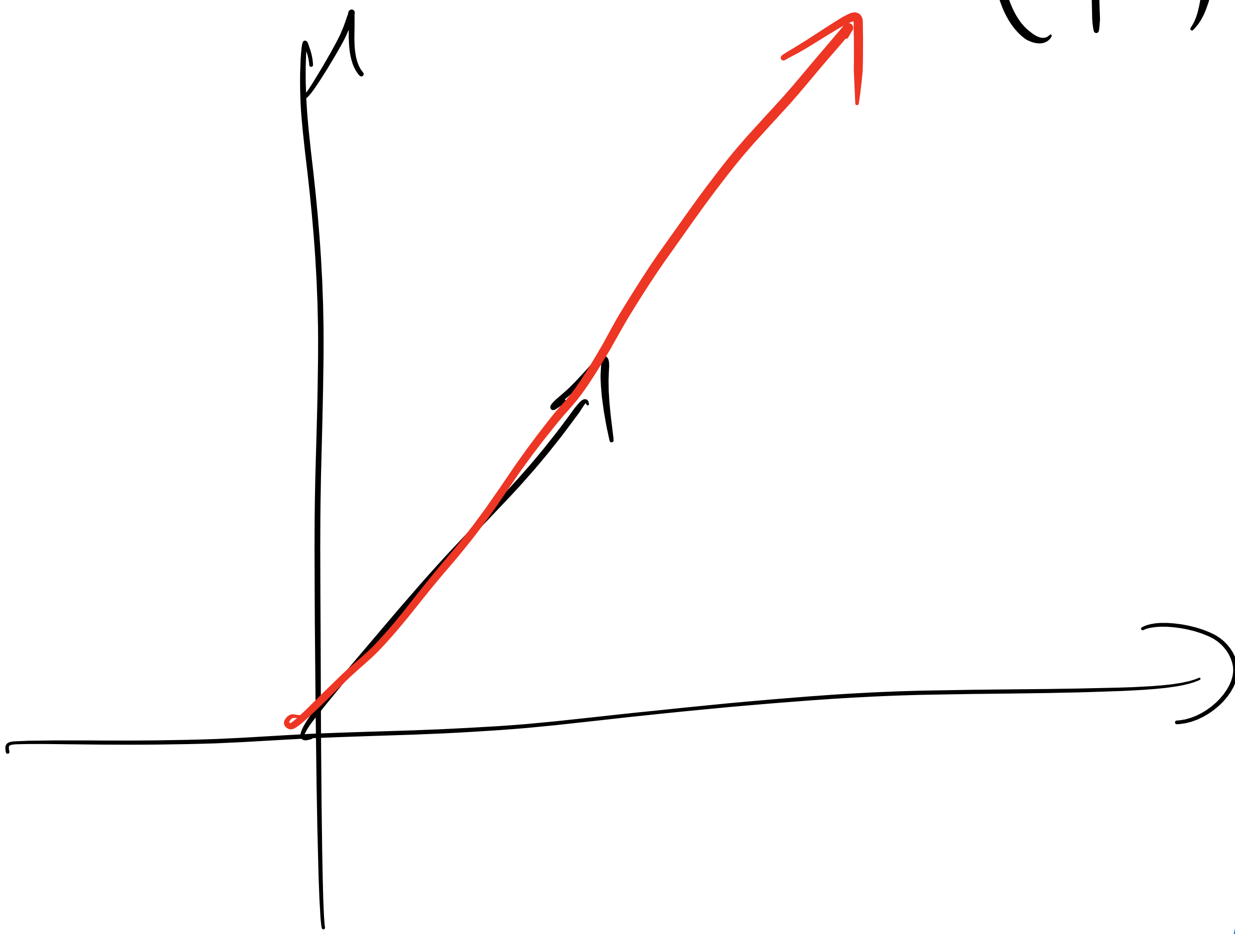

× magnitude

✓ direction

$$A\vec{x} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$A\begin{pmatrix}1\\1\end{pmatrix}=\begin{pmatrix}1&1\\1&1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}=\begin{pmatrix}2\\2\end{pmatrix}$$

$$=2\begin{pmatrix}1\\1\end{pmatrix}$$




value of the eigenvalue

$\lambda > 1 \Rightarrow$ growth in the vector

$|\lambda| = 1 \Rightarrow$ no growth

$-1 < \lambda < 1 \Rightarrow$ decay

$\lambda < -1 \Rightarrow$ oscillating growth

$A\vec{x} \approx \vec{x}$

bridge

wind

eigenvalue close to 1

$$W_h^{(new)} = W_h^{(old)} - \alpha \frac{\partial \mathcal{L}}{\partial W_h}$$

So eigenvalues of $W_h$ play an important role here

How is it that $W_h$ may have some large/or small eigenvalues?

$\Rightarrow$ may happen because of initialization of $W_h$

normal distribution

can have large

eigenvalues → exploding gradients

what to do?

- how can we resolve exploding gradients?

grad = ...

if grad > $10^2$:
    grad = $10^2$

gradient clipping

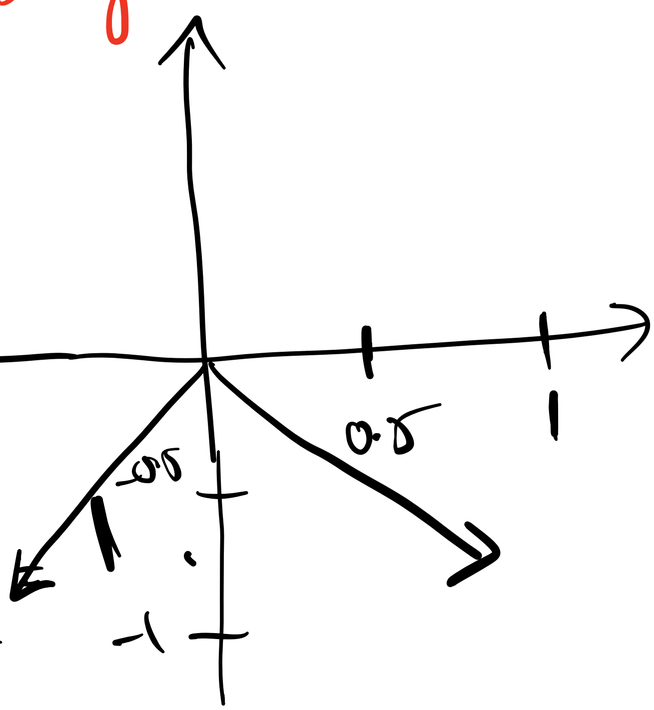
(sure helps with exploding numbers but the model is also losing some of the learning) how to transform $_L W_h$ to the initial have eigenvalues that around 1?

$$\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} = 1\cdot\begin{pmatrix}x_1\\ x_2\end{pmatrix}$$

eigenvalue of 1

orthogonal matrix

$$Q = \begin{pmatrix}\boxed{\begin{matrix}1/\sqrt{2}\\ -1/\sqrt{2}\end{matrix}} & \boxed{\begin{matrix}-1/\sqrt{2}\\ -1/\sqrt{2}\end{matrix}}\end{pmatrix}$$




$$\sqrt{\left(1/\sqrt{2}\right)^2 + \left(-1/\sqrt{2}\right)^2} = \sqrt{\frac{1}{2}+\frac{1}{2}}$$

columns are

orthonormal
(orthogonal and unit length)

$$Q\vec{x} = \lambda \vec{x}$$

$$\Rightarrow Q\vec{x} - \lambda\vec{x} = \vec{0}$$

$$\Rightarrow \underbrace{(Q - \lambda I)}\vec{x} = \vec{0}$$

$$\vec{x} \neq \vec{0}$$

determinant $\rightarrow$ matrix invertible or not

$$\det(A) \rightarrow \text{scalar} \begin{cases} = 0 & \text{not invertible} \\ \neq 0 & \text{invertible} \end{cases}$$

$$\boxed{\det(Q - \lambda I) = 0}$$

$$\det\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = ad - bc$$

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\boxed{\begin{pmatrix} 2 & 0 \\ 1 & -1 \end{pmatrix}} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{bmatrix} 2x_1 \\ x_1 - x_2 \end{bmatrix}$$

$$A\vec{x} = \vec{0} \Rightarrow \vec{x} = \vec{0}$$

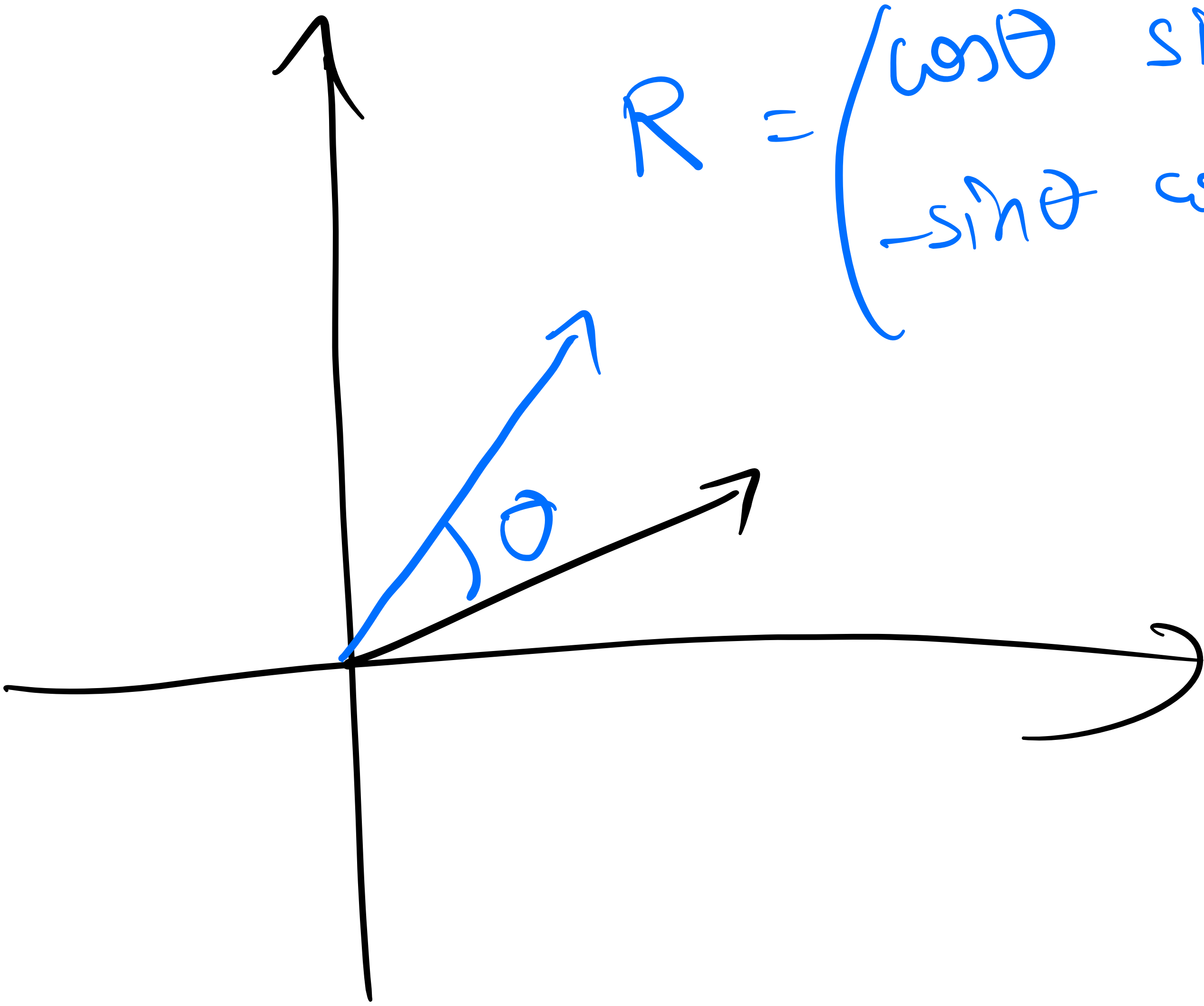
---

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

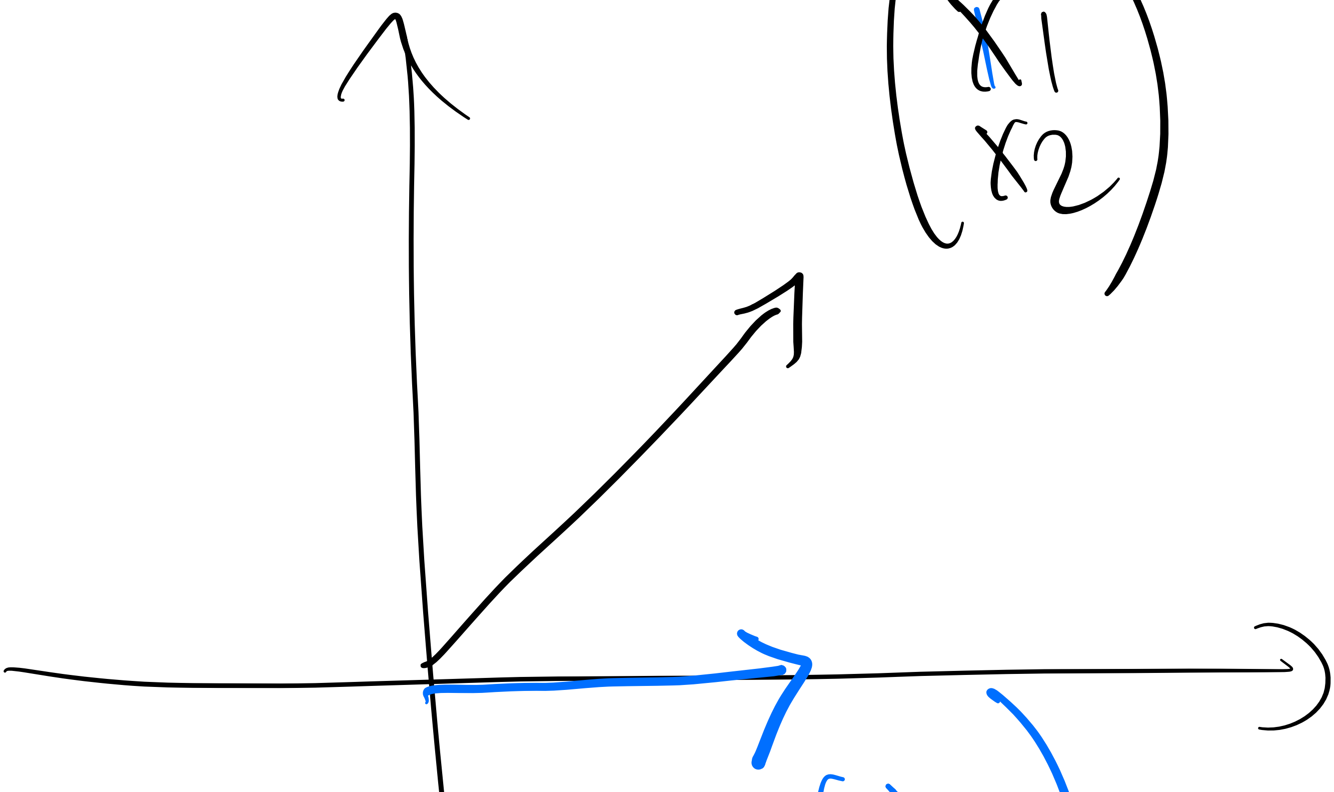
$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$B\vec{x} = \vec{0} \not\Rightarrow \vec{x} = \vec{0}$$

$$R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$




$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



$$\begin{pmatrix} x_1 \end{pmatrix}$$

$$\begin{pmatrix} 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \longrightarrow \text{not invertible}$$

$$Q - \lambda I$$

$$= \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$$= \begin{pmatrix} 1/\sqrt{2} - \lambda & -1/\sqrt{2} \\ -1/\sqrt{2} & -1/\sqrt{2} - \lambda \end{pmatrix}$$

$$\det(Q-\lambda I) = \left(\tfrac{1}{\sqrt{2}}-\lambda\right)\left(-\tfrac{1}{\sqrt{2}}-\lambda\right) - \frac{1}{2} = 0$$

$$\Rightarrow -\frac{1}{2} - \cancel{\frac{\lambda}{\sqrt{2}}} + \cancel{\frac{\lambda}{\sqrt{2}}} + \lambda^2 - \frac{1}{2} = 0$$

$$\Rightarrow \lambda^2 = 1 \qquad \Rightarrow \lambda = \pm 1$$

$$\Rightarrow |\lambda| = 1$$

---

Orthogonal Initialization

import numpy as np

$W_h \longrightarrow$ random normal distr

$W_{h\text{-}orth} =$ np.linalg.qr(A)

matrix $\longrightarrow$ orthogonal matrix

gram schmidt process.

orthogonal initialization $\xrightarrow{\text{yay}}$ exploding gradients

$\longrightarrow$ vanishing gradients

alit (but not fully)

Later words in a sentence are more affected by vanishing gradients issue because we would be including more derivatives of sigmoids.

$\Rightarrow$ truncate the number of previous words used in the prediction of the next word.

# of previous words small $\longrightarrow$

The boy who is in France, a country ---- ..., ... - - . croissant . ------- ) is a student.

small window of previous words could lead to the model not properly learning the relationship between "student" and "boy".

We can use a larger # of words
→ vanishing gradients again
→ computations

Truncated Backpropagation through time

typical numbers for the window
→ 20-50 words
→ 100 words. (meh, risk vanish grad, and more computations)

⇒ Still could have vanishing gradients issues or bad learning.

⇒ Long Short term Memory (LSTM) models

still an RNN, a more improved RNN