Shortcomings of RNN

Backward Propagation

- $\partial L/\partial W = (\partial L/\partial \hat{y}) \,.\, (\partial \hat{y}/\partial H_3) \,.$
  $(\partial g(z_3)/\partial z_3)[\ H_2 +$
  $W((\partial g(z_2)/\partial z_2)[\ H_1 +$
  $W((\partial g(z_1)/\partial z_1)[\ H_0 +$
  $W(\partial H_0/\partial W)])]\ )]$

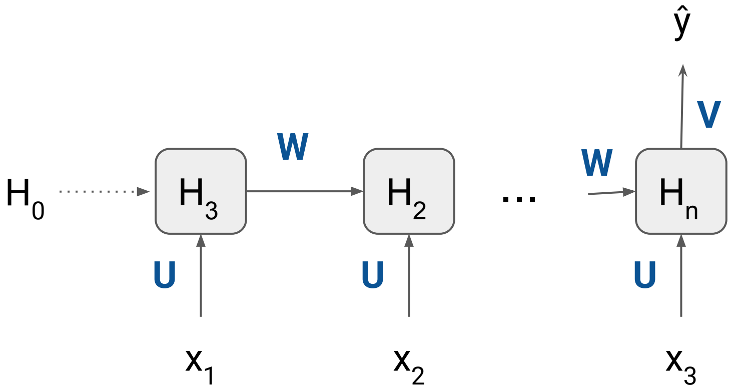
- $\partial L/\partial U = (\partial L/\partial \hat{y}) \,.\, (\partial \hat{y}/\partial H_3) \,.$
  $(\partial g(z_3)/\partial z_3)[\ x_3 +$
  $(W\ (\partial g(z_2)/\partial z_2) \,.\, [\ x_2 +$
  $(W\ (\partial g(z_1)/\partial z_1) \,.\, [\ x_1 +$
  $(\partial W H_0/\partial U)\ ])\ ])\ ]$

Shortcomings of RNN

- **∂L/∂W**

- **∂L/∂U**

L = Loss

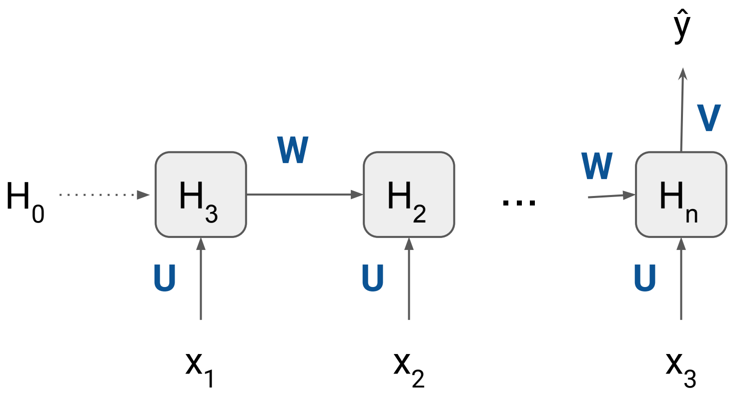


Shortcomings of RNN

- **$\partial L/\partial W$** → $\partial H_n/\partial W$

- **$\partial L/\partial U$** → $\partial H_n/\partial U$

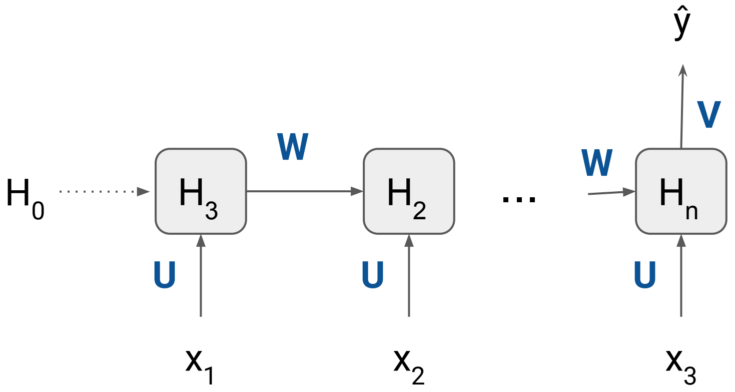
L = Loss



Shortcomings of RNN

- **$\partial L/\partial W$** → $\partial H_n/\partial W$ → $\partial H_{n-1}/\partial W$
- **$\partial L/\partial U$** → $\partial H_n/\partial U$ → $\partial H_{n-1}/\partial U$

L = Loss

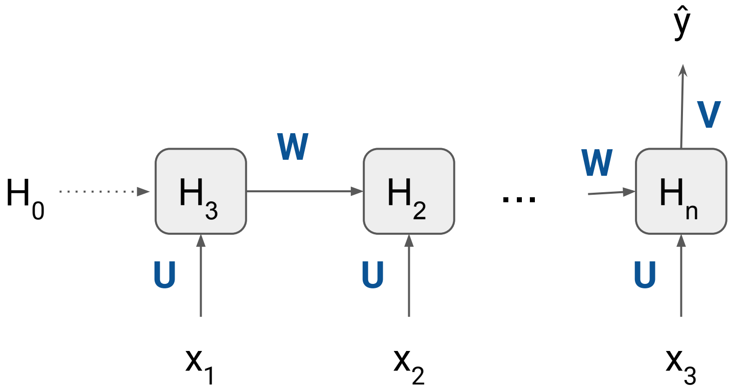


Shortcomings of RNN

- $\boldsymbol{\partial L/\partial W} \rightarrow \partial H_n/\partial W \rightarrow \partial H_{n-1}/\partial W \ldots \partial H_0/\partial W$

- $\boldsymbol{\partial L/\partial U} \rightarrow \partial H_n/\partial U \rightarrow \partial H_{n-1}/\partial U \ldots \partial H_0/\partial U$

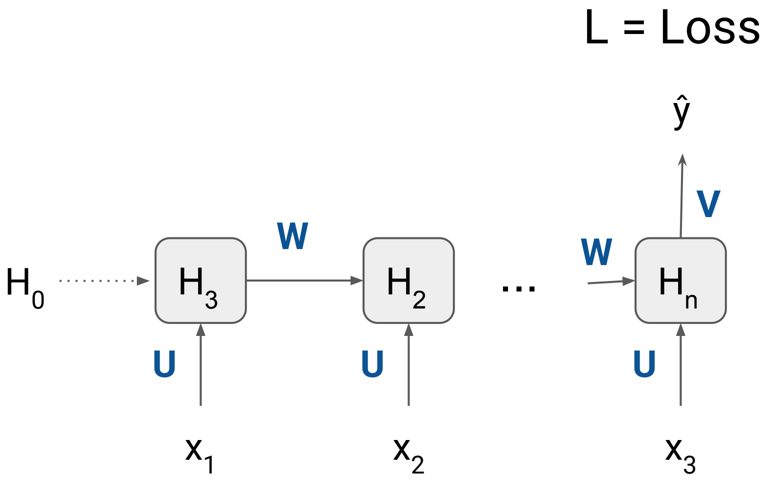
L = Loss



Vanishing Gradient

- **$\partial L/\partial W$** → $\partial H_n/\partial W$ → $\partial H_{n-1}/\partial W$ … $\partial H_0/\partial W$

- **$\partial L/\partial U$** → $\partial H_n/\partial U$ → $\partial H_{n-1}/\partial U$ … $\partial H_0/\partial U$



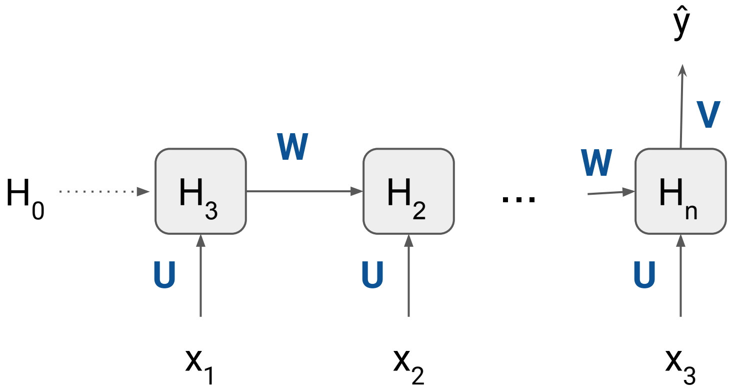
Vanishing Gradient

- $\mathbf{\partial L/\partial W} \rightarrow \partial H_n/\partial W \rightarrow \partial H_{n-1}/\partial W \ldots \partial H_0/\partial W$

- $\mathbf{\partial L/\partial U} \rightarrow \partial H_n/\partial U \rightarrow \partial H_{n-1}/\partial U \ldots \partial H_0/\partial U$

- If gradients < 1

L = Loss



Vanishing Gradient

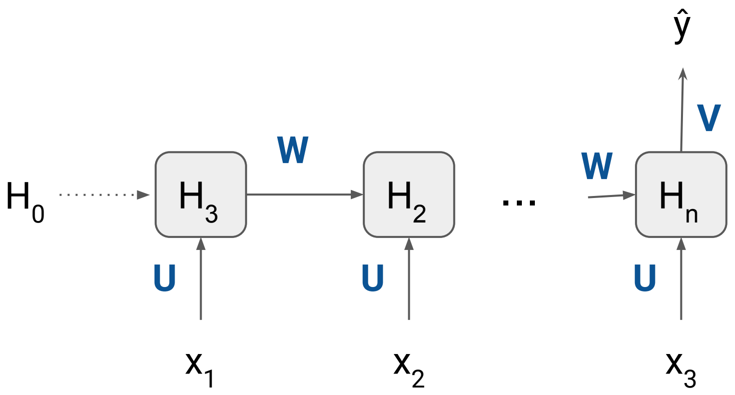
- $\boldsymbol{\partial L/\partial W} \rightarrow \partial H_n/\partial W \rightarrow \partial H_{n-1}/\partial W \ldots \partial H_0/\partial W$

- $\boldsymbol{\partial L/\partial U} \rightarrow \partial H_n/\partial U \rightarrow \partial H_{n-1}/\partial U \ldots \partial H_0/\partial U$

- If gradients < 1

  then **∂L/∂W** and **∂L/∂U** will be infinitesimally small

L = Loss



Vanishing Gradient

- **∂L/∂W** → $\partial H_n/\partial W$ → $\partial H_{n-1}/\partial W$ … $\partial H_0/\partial W$

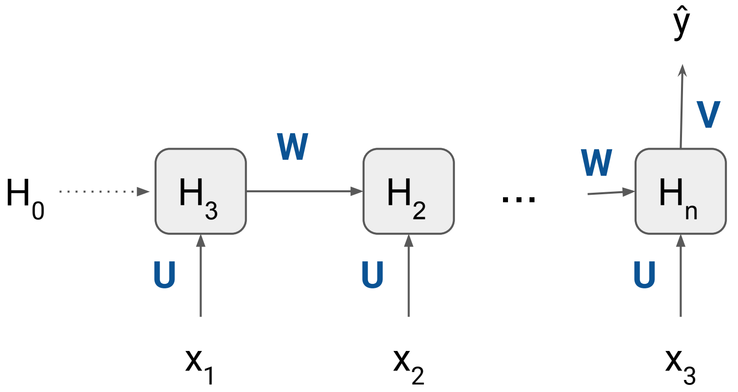
- **∂L/∂U** → $\partial H_n/\partial U$ → $\partial H_{n-1}/\partial U$ … $\partial H_0/\partial U$

- If gradients < 1

  then **∂L/∂W** and **∂L/∂U** will be infinitesimally small

  (0.5)^10 = 0.00097

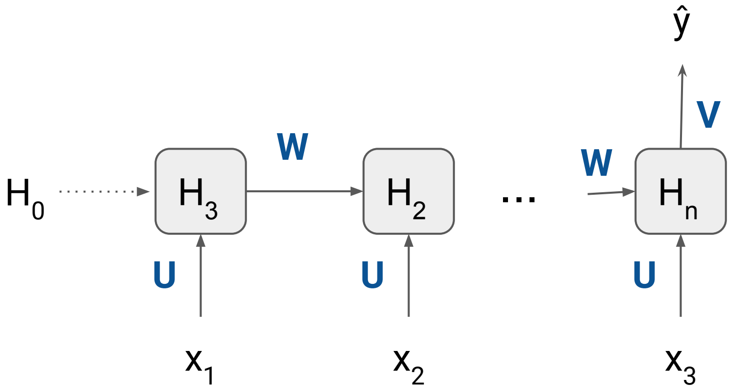
L = Loss



Vanishing Gradient

- $\partial L/\partial W \approx \partial L/\partial U \approx 0$

L = Loss

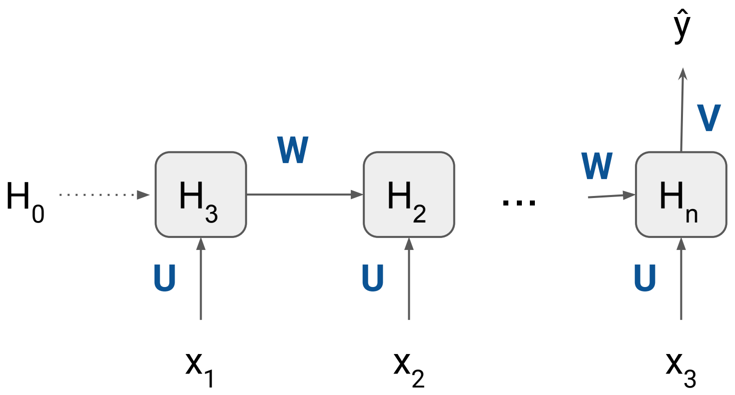


Vanishing Gradient

- $\partial L/\partial W \approx \partial L/\partial U \approx 0$
- $W = W - \alpha\,(\partial L/\partial W)$
  $U = U - \alpha\,(\partial L/\partial U)$

updating weights

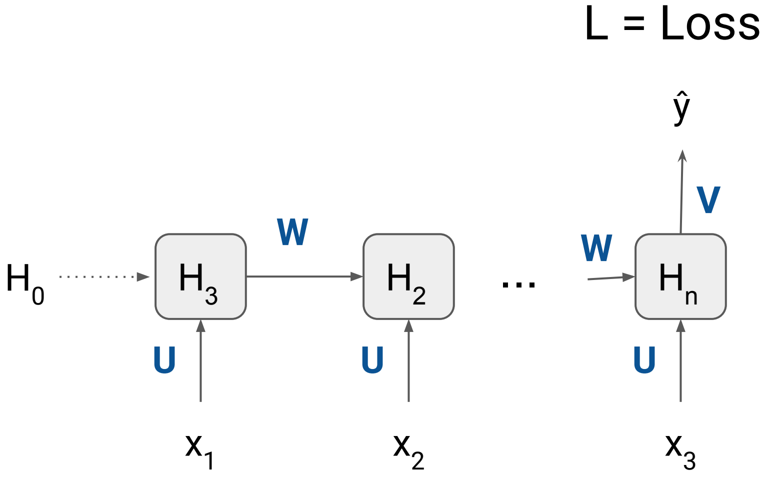
L = Loss



Vanishing Gradient

- $\partial L/\partial W \approx \partial L/\partial U \approx 0$
- $W = W - \alpha\,(\partial L/\partial W)$
  $U = U - \alpha\,(\partial L/\partial U)$
- $W_{before} \approx W_{after}$
  $U_{before} \approx U_{after}$

updating weights

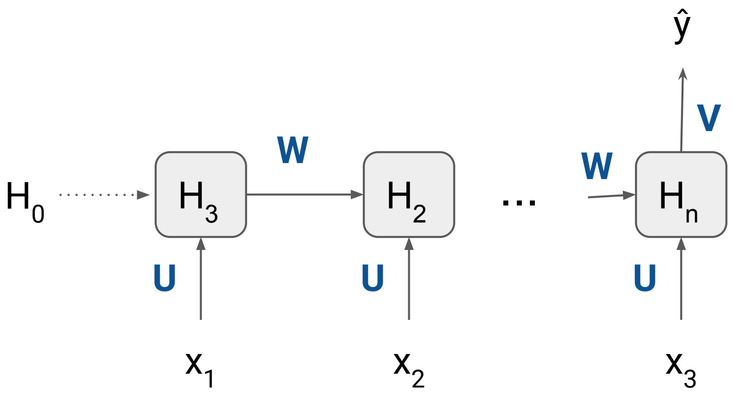


Exploding Gradient

- **∂L/∂W** → $\partial H_n/\partial W \rightarrow \partial H_{n-1}/\partial W \ldots \partial H_0/\partial W$

- **∂L/∂U** → $\partial H_n/\partial U \rightarrow \partial H_{n-1}/\partial U \ldots \partial H_0/\partial U$

L = Loss



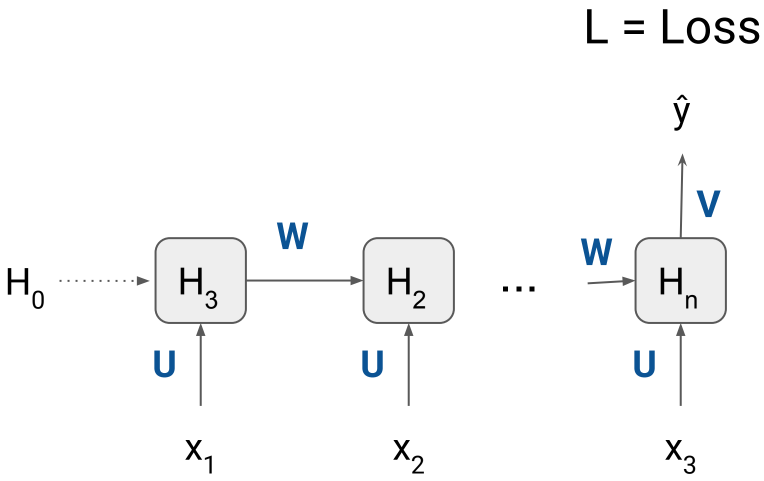
Exploding Gradient

- **$\partial L/\partial W$** → $\partial H_n/\partial W$ → $\partial H_{n-1}/\partial W$ … $\partial H_0/\partial W$

- **$\partial L/\partial U$** → $\partial H_n/\partial U$ → $\partial H_{n-1}/\partial U$ … $\partial H_0/\partial U$

- If gradients > 1

  then $\partial L/\partial W$ and $\partial L/\partial U$ are very large



Exploding Gradient

- **$\partial L/\partial W$** → $\partial H_n/\partial W$ → $\partial H_{n-1}/\partial W$ … $\partial H_0/\partial W$

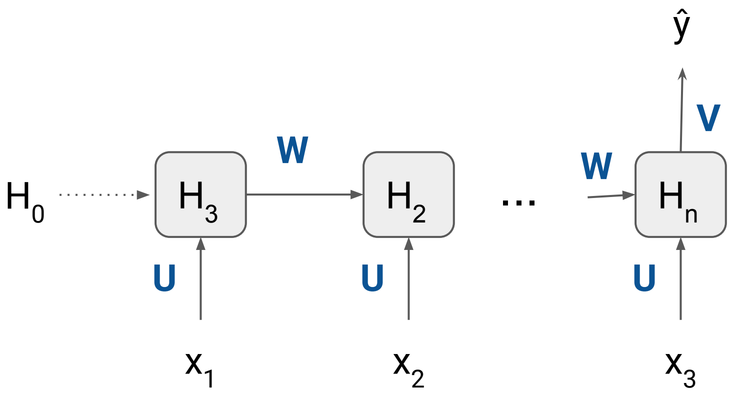
- **$\partial L/\partial U$** → $\partial H_n/\partial U$ → $\partial H_{n-1}/\partial U$ … $\partial H_0/\partial U$

- If gradients > 1

  then $\partial L/\partial W$ and $\partial L/\partial U$ are very large

  (1.5)^10 = 57.665

L = Loss



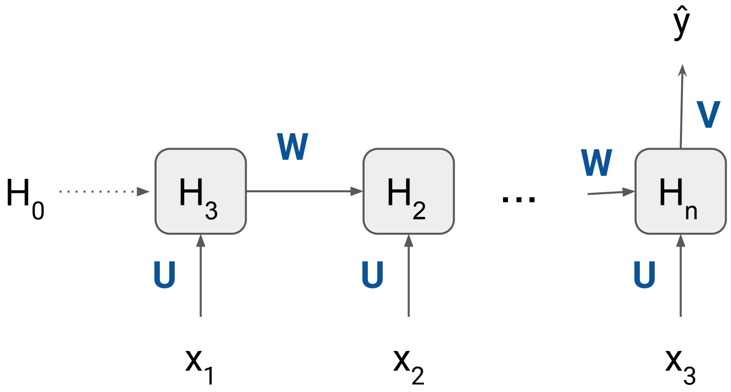
Analytics Vidhya
Learn everything about analytics

Exploding Gradient

- $W = W - \alpha (\partial L / \partial W)$
  $U = U - \alpha (\partial L / \partial U)$

updating weights

L = Loss



Issues with RNN

- **Vanishing Gradient**

  RNN does not perform well for long sentences

Issues with RNN

- **Vanishing Gradient**

RNN does not perform well for long sentences

Eg: The writer of the books

Issues with RNN

- **Vanishing Gradient**

  RNN does not perform well for long sentences

  Eg: The writer of the books

Issues with RNN

- **Vanishing Gradient**

  RNN does not perform well for long sentences

  Eg: The writer of the books is

Issues with RNN

- **Vanishing Gradient**

  RNN does not perform well for long sentences

  Eg: The writer of the books is
  The writer of the books are

Issues with RNN

- **Vanishing Gradient**

  RNN does not perform well for long sentences

  Eg: The writer of the books is
  The writer of the books are

- **Exploding Gradient**

  - Gradients are large

Issues with RNN

- **Vanishing Gradient**

  RNN does not perform well for long sentences

  Eg: The writer of the books is
  The writer of the books are

- **Exploding Gradient**

  - Gradients are large
  - Poor predictions

Issues with RNN

How to mitigate exploding and vanishing gradients?

Issues with RNN

How to mitigate exploding and vanishing gradients?

Exploding Gradients
Gradient Clipping

Issues with RNN

How to mitigate exploding and vanishing gradients?

Exploding Gradients
Gradient Clipping

Threshold = 0.1

- if gradient > threshold, then gradient = threshold
- if gradient <= threshold, then gradient = gradient

Issues with RNN

How to mitigate exploding and vanishing gradients?

Exploding Gradients	Vanishing Gradients
Gradient Clipping	LSTM or GRU

Thank You