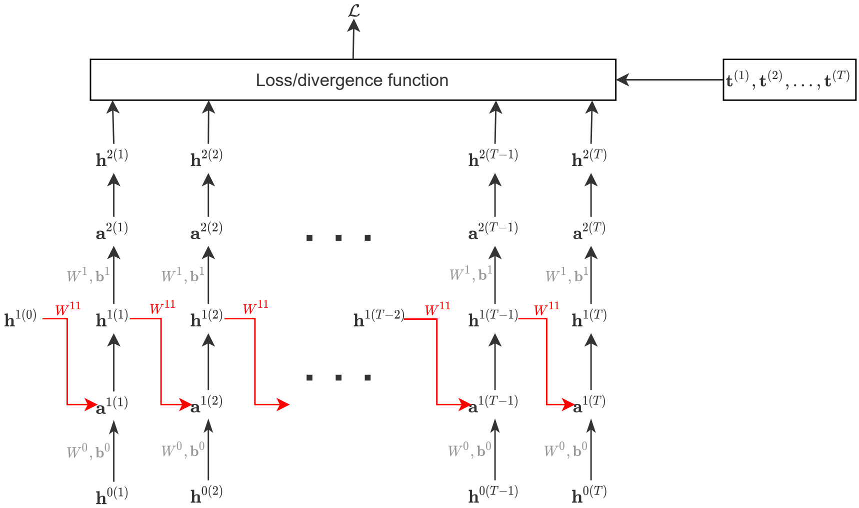
Deep Learning

Syed Irtaza Muzaffar

Backpropagation Through Time

RNN Unfolded in Time



Backpropagation Through Time (BPTT)

- In order to train a single hidden layer RNN, we need 5 derivatives:
  1. $\nabla_{W^1} \mathcal{L} \in \mathbb{R}^{M \times K}$
  2. $\nabla_{\mathbf{b}^1} \mathcal{L} \in \mathbb{R}^{1 \times K}$
  3. $\nabla_{W^{11}} \mathcal{L} \in \mathbb{R}^{M \times M}$
  4. $\nabla_{W^0} \mathcal{L} \in \mathbb{R}^{D \times M}$
  5. $\nabla_{\mathbf{b}^0} \mathcal{L} \in \mathbb{R}^{1 \times M}$
- They correspond to backpropagation through space as well as time.

Background
Multivariate Chain Rule

- Recall the *multivariate* chain rule of differentiation

$$\frac{df(u(x),v(x))}{dx} = \frac{\partial f}{\partial u}\frac{du}{dx} + \frac{\partial f}{\partial v}\frac{dv}{dx}$$

$$\begin{tikzcd} & u(x) \arrow[dr] & \\ x \arrow[ur] \arrow[dr] & & f(u(x),v(x)) \\ & v(x) \arrow[ur] & \end{tikzcd}$$

Background
Matrix and Vector Calculus

For scalars $x, y \in \mathbb{R}$, vectors $\mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^k$ and matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$, we will use the following conventions for writing matrix and vector derivatives.

$$\text{Scalar w.r.t vector: } \nabla_{\mathbf{x}} y = \frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_d} \end{bmatrix}$$

$$\text{Vector w.r.t scalar: } \nabla_{x} \mathbf{y} = \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_k}{\partial x} \end{bmatrix}$$

$$\text{Vector w.r.t vector: } \nabla_{\mathbf{x}} \mathbf{y} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \nabla_{\mathbf{x}} y_1 \\ \nabla_{\mathbf{x}} y_2 \\ \vdots \\ \nabla_{\mathbf{x}} y_k \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_d} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_k}{\partial x_1} & \frac{\partial y_k}{\partial x_2} & \cdots & \frac{\partial y_k}{\partial x_d} \end{bmatrix}}_{k\times d}$$

Background
Matrix and Vector Calculus

Scalar w.r.t matrix: $\nabla_{\mathbf{X}} y = \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{21}} & \cdots & \frac{\partial y}{\partial x_{m1}} \\ \frac{\partial y}{\partial x_{12}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1n}} & \frac{\partial y}{\partial x_{2n}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix}$

Matrix w.r.t scalar: $\nabla_{x} \mathbf{Y} = \frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{12}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \frac{\partial y_{21}}{\partial x} & \frac{\partial y_{22}}{\partial x} & \cdots & \frac{\partial y_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \frac{\partial y_{m2}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$

Background
Matrix and Vector Calculus

For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and matrices $\mathbf{M} \in \mathbb{R}^{k \times d}$ and $\mathbf{A} \in \mathbb{R}^{d \times d}$

- $\nabla_{\mathbf{x}}(\mathbf{y}^T\mathbf{x}) = \nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \mathbf{y}^T$
- $\nabla_{\mathbf{x}}(\mathbf{M}\mathbf{x}) = \mathbf{M}$
- $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{x}^T(\mathbf{A}^T + \mathbf{A})$
- For symmetric $\mathbf{A}$, $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = 2(\mathbf{A}\mathbf{x})^T$

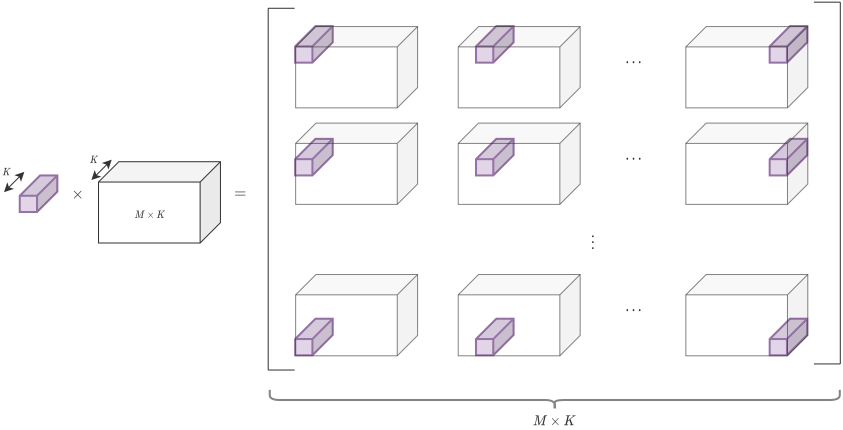
Detour
$\nabla_W \mathcal{L}(W\mathbf{x})$

- Derivative of *scalar* loss function $\mathcal{L}(\mathbf{y})$ of *vector* output $\mathbf{y} = W\mathbf{x}$ w.r.t matrix $W \in \mathbb{R}^{K \times M}$.

$$\underbrace{\nabla_W \mathcal{L}}_{M \times K} = \underbrace{\nabla_{\mathbf{y}} \mathcal{L}}_{1 \times K} \underbrace{\nabla_W \mathbf{y}}_{K \times (M \times K)}$$

Detour
$\nabla_W \mathcal{L}(W\mathbf{x})$

Multiplication of 1D array with 3D tensor



Detour
$\nabla_W \mathcal{L}(W\mathbf{x})$

$$
\begin{bmatrix} y_1 \\ \\ \\ \end{bmatrix}
=
\underbrace{\begin{bmatrix} W_{11} & W_{12} & W_{13} & W_{14} \\ & & & \\ & & & \end{bmatrix}}_{3 \times 4}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}
$$

$$
\nabla_W y_1 =
\begin{bmatrix}
\frac{\partial y_1}{\partial W_{11}} & \frac{\partial y_1}{\partial W_{21}} & \frac{\partial y_1}{\partial W_{31}} \\
\frac{\partial y_1}{\partial W_{12}} & \frac{\partial y_1}{\partial W_{22}} & \frac{\partial y_1}{\partial W_{32}} \\
\frac{\partial y_1}{\partial W_{13}} & \frac{\partial y_1}{\partial W_{23}} & \frac{\partial y_1}{\partial W_{33}} \\
\frac{\partial y_1}{\partial W_{14}} & \frac{\partial y_1}{\partial W_{24}} & \frac{\partial y_1}{\partial W_{34}}
\end{bmatrix}
=
\begin{bmatrix}
x_1 & 0 & 0 \\
x_2 & 0 & 0 \\
x_3 & 0 & 0 \\
x_4 & 0 & 0
\end{bmatrix}
$$

Detour
$\nabla_W \mathcal{L}(W\mathbf{x})$

Derivative of vector with respect to matrix

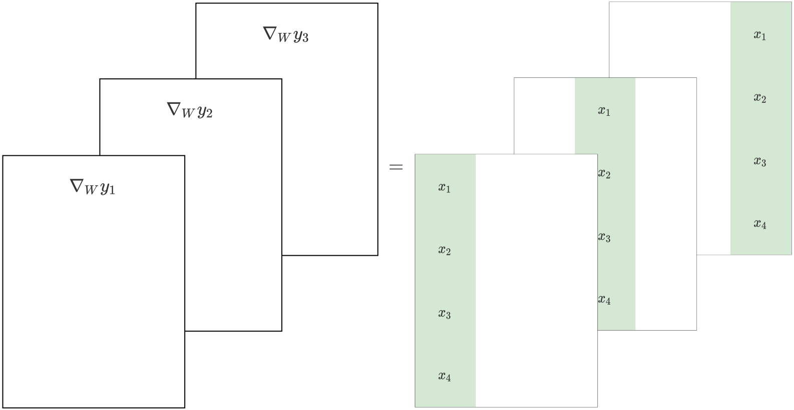
$$\nabla_W y_2 = \begin{bmatrix} \frac{\partial y_2}{\partial W_{11}} & \frac{\partial y_2}{\partial W_{21}} & \frac{\partial y_2}{\partial W_{31}} \\ \frac{\partial y_2}{\partial W_{12}} & \frac{\partial y_2}{\partial W_{22}} & \frac{\partial y_2}{\partial W_{32}} \\ \frac{\partial y_2}{\partial W_{13}} & \frac{\partial y_2}{\partial W_{23}} & \frac{\partial y_2}{\partial W_{33}} \\ \frac{\partial y_2}{\partial W_{14}} & \frac{\partial y_2}{\partial W_{24}} & \frac{\partial y_2}{\partial W_{34}} \end{bmatrix} = \begin{bmatrix} 0 & x_1 & 0 \\ 0 & x_2 & 0 \\ 0 & x_3 & 0 \\ 0 & x_4 & 0 \end{bmatrix}$$

$$\nabla_W y_3 = \begin{bmatrix} \frac{\partial y_3}{\partial W_{11}} & \frac{\partial y_3}{\partial W_{21}} & \frac{\partial y_3}{\partial W_{31}} \\ \frac{\partial y_3}{\partial W_{12}} & \frac{\partial y_3}{\partial W_{22}} & \frac{\partial y_3}{\partial W_{32}} \\ \frac{\partial y_3}{\partial W_{13}} & \frac{\partial y_3}{\partial W_{23}} & \frac{\partial y_3}{\partial W_{33}} \\ \frac{\partial y_3}{\partial W_{14}} & \frac{\partial y_3}{\partial W_{24}} & \frac{\partial y_3}{\partial W_{34}} \end{bmatrix} = \begin{bmatrix} 0 & 0 & x_1 \\ 0 & 0 & x_2 \\ 0 & 0 & x_3 \\ 0 & 0 & x_4 \end{bmatrix}$$

Detour
$\nabla_W \mathcal{L}(W\mathbf{x})$

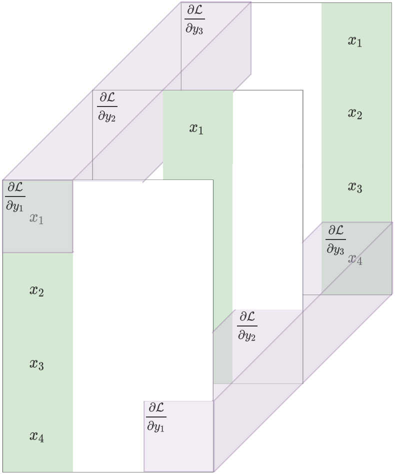
Derivative of vector with respect to matrix

$$\nabla_W \mathbf{y} = \nabla_W W\mathbf{x} =$$



Detour

$\nabla_W \mathcal{L}(W\mathbf{x})$



$$= \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial y_1}x_1 & \frac{\partial \mathcal{L}}{\partial y_2}x_1 & \frac{\partial \mathcal{L}}{\partial y_3}x_1 \\ \frac{\partial \mathcal{L}}{\partial y_1}x_2 & \frac{\partial \mathcal{L}}{\partial y_2}x_2 & \frac{\partial \mathcal{L}}{\partial y_3}x_2 \\ \frac{\partial \mathcal{L}}{\partial y_1}x_3 & \frac{\partial \mathcal{L}}{\partial y_2}x_3 & \frac{\partial \mathcal{L}}{\partial y_3}x_3 \\ \frac{\partial \mathcal{L}}{\partial y_1}x_4 & \frac{\partial \mathcal{L}}{\partial y_2}x_4 & \frac{\partial \mathcal{L}}{\partial y_3}x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial y_1} & \frac{\partial \mathcal{L}}{\partial y_2} & \frac{\partial \mathcal{L}}{\partial y_3} \end{bmatrix}$$

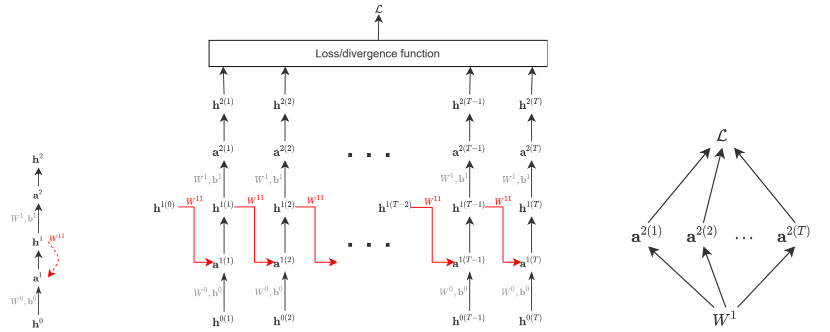
$$\nabla_{\mathbf{y}} \mathcal{L} \times \nabla_W \mathbf{y} \quad = \quad \mathbf{x} \nabla_{\mathbf{y}} \mathcal{L}$$

BPTT

Derivative number 1: $\nabla_{W^1}\mathcal{L}$

- Notice that W^1 affects loss $\mathcal{L}$ through $\mathbf{a}^{2(t)}$ at each time t.

$$\mathcal{L}(\underbrace{\mathbf{a}^{2(1)}(W^1)}_{t=1}, \underbrace{\mathbf{a}^{2(2)}(W^1)}_{t=2}, \ldots, \underbrace{\mathbf{a}^{2(T)}(W^1)}_{t=T})$$



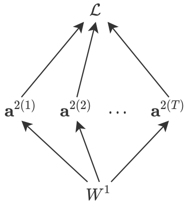
RNN

Unfolded in time

Influence diagram

BPTT

Derivative number 1: $\nabla_{W^1}\mathcal{L}$



$$\mathbf{h}^{2(t)} = f(\mathbf{a}^{2(t)})$$

$$\boxed{\mathbf{a}^{2(t)} = W^1\mathbf{h}^{1(t)} + \mathbf{b}^1}$$

$$\mathbf{h}^{1(t)} = \tanh(\mathbf{a}^{1(t)})$$

$$\mathbf{a}^{1(t)} = W^0\mathbf{h}^{0(t)} + W^{11}\mathbf{h}^{1(t-1)} + \mathbf{b}^0$$

- Using the multivariate chain rule over time

$$\underbrace{\nabla_{W^1}\mathcal{L}}_{M\times K} = \sum_{t=1}^{T} \underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}}_{1\times K}\, \underbrace{\nabla_{W^1}\mathbf{a}^{2(t)}}_{K\times(M\times K)}$$

$$= \sum_{t=T}^{1} \underbrace{\mathbf{h}^{1(t)}}_{M\times 1}\, \underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}}_{1\times K}$$

- Computation of $\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}$ is described next.

BPTT
$\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}$

- The derivatives of loss $\mathcal{L}$ w.r.t pre-activations $\mathbf{a}^{2(t)}$ can be computed as

$$\underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}}_{1\times K} = \underbrace{\nabla_{\mathbf{h}^{2(t)}}\mathcal{L}}_{1\times K}\,\underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathbf{h}^{2(t)}}_{K\times K} = \nabla_{\mathbf{h}^{2(t)}}\mathcal{L}\underbrace{\begin{bmatrix} \partial_{a_1}h_1 & \partial_{a_2}h_1 & \dots & \partial_{a_K}h_1 \\ \partial_{a_1}h_2 & \partial_{a_2}h_2 & \dots & \partial_{a_K}h_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{a_1}h_K & \partial_{a_2}h_K & \dots & \partial_{a_K}h_K \end{bmatrix}^{2(t)}}_{\text{Jacobian matrix}}$$

- The Jacobian matrix is the derivative of outputs with respect to inputs.
- In 1D, the term $\frac{dy}{dx}$ is the 1×1 Jacobian matrix of $y = f(x)$.
- Jacobian matrix is
  - diagonal for scalar activation functions (logistic sigmoid, tanh, ReLU), and
  - dense for vector activation functions (softmax).

BPTT

Derivative number 2: $\nabla_{\mathbf{b}^1}\mathcal{L}$

- Following the same reasoning as used for $\nabla_{W^1}\mathcal{L}$ above, we can compute

$$\underbrace{\nabla_{\mathbf{b}^1}\mathcal{L}}_{1\times K} = \sum_{t=T}^{1} \underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}}_{1\times K}$$

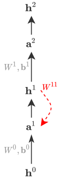
where we have used the fact that $\nabla_{\mathbf{b}^1}\mathbf{a}^{2(t)} = I_K$.

BPTT

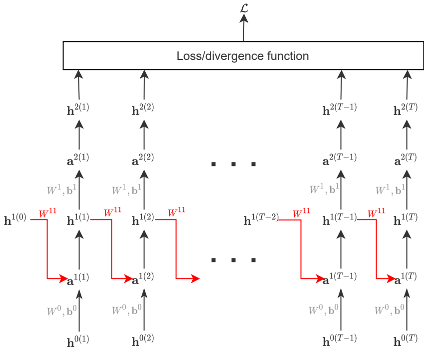
Derivative number 3: $\nabla_{W^{11}}\mathcal{L}$

- Notice that W^{11} affects loss $\mathcal{L}$ through $\mathbf{a}^{1(t)}$ at each time t.

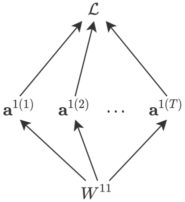
$$\mathcal{L}(\underbrace{\mathbf{a}^{1(1)}(W^{11})}_{t=1}, \underbrace{\mathbf{a}^{1(2)}(W^{11})}_{t=2}, \ldots, \underbrace{\mathbf{a}^{1(T)}(W^{11})}_{t=T})$$



RNN



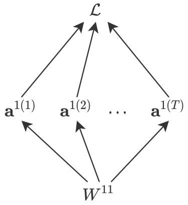
Unfolded in time



Influence diagram

BPTT

Derivative number 3: $\nabla_{W^{11}}\mathcal{L}$



$$\mathbf{h}^{2(t)} = f(\mathbf{a}^{2(t)})$$

$$\mathbf{a}^{2(t)} = W^1\mathbf{h}^{1(t)} + \mathbf{b}^1$$

$$\mathbf{h}^{1(t)} = \tanh(\mathbf{a}^{1(t)})$$

$$\boxed{\mathbf{a}^{1(t)} = W^0\mathbf{h}^{0(t)} + W^{11}\mathbf{h}^{1(t-1)} + \mathbf{b}^0}$$

- Using the multivariate chain rule over time

$$\underbrace{\nabla_{W^{11}}\mathcal{L}}_{M\times M} = \sum_{t=1}^{T} \underbrace{\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}}_{1\times M}\,\underbrace{\nabla_{W^{11}}\mathbf{a}^{1(t)}}_{M\times(M\times M)}$$

$$= \sum_{t=T}^{1} \underbrace{\mathbf{h}^{1(t-1)}}_{M\times 1}\,\underbrace{\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}}_{1\times M}$$

- Computation of $\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}$ is described next.

BPTT
$\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}$

- The derivatives of loss $\mathcal{L}$ w.r.t pre-activations $\mathbf{a}^{1(t)}$ can be computed as

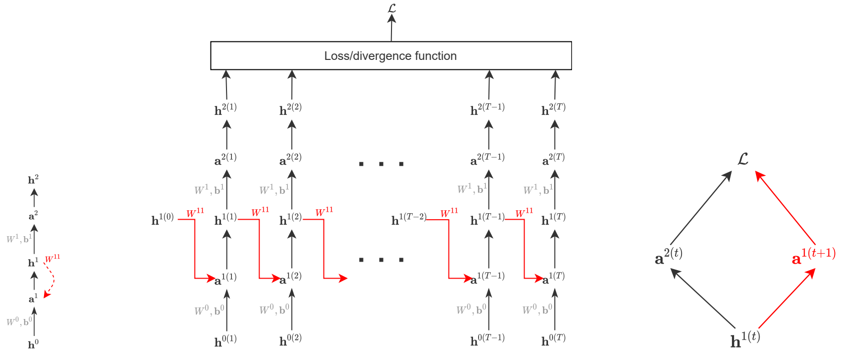
$$\underbrace{\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}}_{1\times M} = \underbrace{\nabla_{\mathbf{h}^{1(t)}}\mathcal{L}}_{1\times M}\underbrace{\nabla_{\mathbf{a}^{1(t)}}\mathbf{h}^{1(t)}}_{M\times M} = \nabla_{\mathbf{h}^{1(t)}}\mathcal{L}\underbrace{\begin{bmatrix} \partial_{a_1}h_1 & \partial_{a_2}h_1 & \dots & \partial_{a_M}h_1 \\ \partial_{a_1}h_2 & \partial_{a_2}h_2 & \dots & \partial_{a_M}h_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{a_1}h_M & \partial_{a_2}h_M & \dots & \partial_{a_M}h_M \end{bmatrix}^{1(t)}}_{\text{Jacobian matrix}}$$

- Computation of $\nabla_{\mathbf{h}^{1(t)}}\mathcal{L}$ is described next.

BPTT

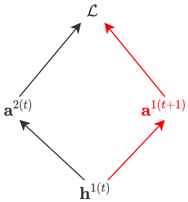
$\nabla_{\mathbf{h}^{1(t)}}\mathcal{L}$

- Notice that $\mathbf{h}^{1(t)}$ affects loss $\mathcal{L}$
  1. through $\mathbf{a}^{2(t)}$ at each time t, and
  2. through $\mathbf{a}^{1(t+1)}$ at each time $t+1$.



RNN | Unfolded in time | Influence diagram

BPTT
$\nabla_{\mathbf{h}^{1(t)}}\mathcal{L}$



$$\mathbf{h}^{2(t)} = f(\mathbf{a}^{2(t)})$$

$$\boxed{\mathbf{a}^{2(t)} = W^1\mathbf{h}^{1(t)} + \mathbf{b}^1}$$

$$\mathbf{h}^{1(t)} = \tanh(\mathbf{a}^{1(t)})$$

$$\boxed{\mathbf{a}^{1(t)} = W^0\mathbf{h}^{0(t)} + W^{11}\mathbf{h}^{1(t-1)} + \mathbf{b}^0}$$

- Using the multivariate chain rule over these 2 time steps

$$\underbrace{\nabla_{\mathbf{h}^{1(t)}}\mathcal{L}}_{1\times M} = \nabla_{\mathbf{a}^{2(t)}}\mathcal{L}\nabla_{\mathbf{h}^{1(t)}}\mathbf{a}^{2(t)} + \underbrace{\nabla_{\mathbf{a}^{1(t+1)}}\mathcal{L}\nabla_{\mathbf{h}^{1(t)}}\mathbf{a}^{1(t+1)}}_{\text{Not required when } t=T}$$

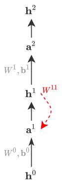
$$= \underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}}_{1\times K}\underbrace{W^1}_{K\times M} + \underbrace{\nabla_{\mathbf{a}^{1(t+1)}}\mathcal{L}}_{1\times M}\underbrace{W^{11}}_{M\times M}$$

BPTT

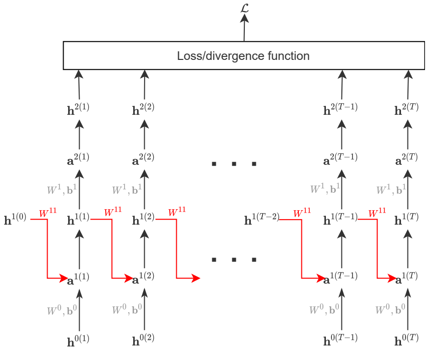
Derivative number 4: $\nabla_{W^0}\mathcal{L}$

- Notice that W^0 affects loss $\mathcal{L}$ through $\mathbf{a}^{1(t)}$ at each time t.

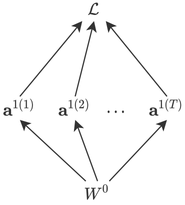
$$\mathcal{L}(\underbrace{\mathbf{a}^{1(1)}(W^0)}_{t=1}, \underbrace{\mathbf{a}^{1(2)}(W^0)}_{t=2}, \ldots, \underbrace{\mathbf{a}^{1(T)}(W^0)}_{t=T})$$



RNN



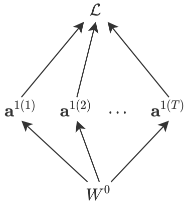
Unfolded in time



Influence diagram

BPTT

Derivative number 4: $\nabla_{W^0}\mathcal{L}$



$$\mathbf{h}^{2(t)} = f(\mathbf{a}^{2(t)})$$

$$\mathbf{a}^{2(t)} = W^1\mathbf{h}^{1(t)} + \mathbf{b}^1$$

$$\mathbf{h}^{1(t)} = \tanh(\mathbf{a}^{1(t)})$$

$$\boxed{\mathbf{a}^{1(t)} = W^0\mathbf{h}^{0(t)} + {\color{red}W^{11}\mathbf{h}^{1(t-1)}} + \mathbf{b}^0}$$

- Using the multivariate chain rule over time

$$\underbrace{\nabla_{W^0}\mathcal{L}}_{D\times M} = \sum_{t=1}^{T} \underbrace{{\color{green}\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}}}_{1\times M}\,\underbrace{{\color{blue}\nabla_{W^0}\mathbf{a}^{1(t)}}}_{M\times(D\times M)}$$

$$= \sum_{t=T}^{1} \underbrace{{\color{blue}\mathbf{h}^{0(t)}}}_{D\times 1}\,\underbrace{{\color{green}\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}}}_{1\times M}$$

BPTT

Derivative number 5: $\nabla_{\mathbf{b}^0}\mathcal{L}$

- Following the same reasoning as used for $\nabla_{W^0}\mathcal{L}$ above, we can compute

$$\underbrace{\nabla_{\mathbf{b}^0}\mathcal{L}}_{1\times M} = \sum_{t=T}^{1} \underbrace{\nabla_{\mathbf{a}^{1(t)}}\mathcal{L}}_{1\times M}$$

where we have used the fact that $\nabla_{\mathbf{b}^0}\mathbf{a}^{1(t)} = I_M$.

Now we have all 5 derivatives required to train an RNN with 1 hidden layer.

Please note that all 5 derivatives will be transposed to obtain the gradients used in gradient descent.

Note about biases

- Notice that, throughout the course, derivative with respect to bias has been the sum of δ-values.
- This was the case for
  - Neural Networks
  - Convolutional Neural Networks, and now
  - Recurrent Neural Networks

Summary

Output layer

$$\nabla_{\mathbf{a}^{2(t)}}\mathcal{L} = \nabla_{\mathbf{h}^{2(t)}}\mathcal{L}\,\underbrace{\nabla_{\mathbf{a}^{2(t)}}\mathbf{h}^{2(t)}}_{\text{Jacobian}}$$

$$\nabla_{W^1}\mathcal{L} = \sum_{t=T}^{1} \mathbf{h}^{1(t)}\nabla_{\mathbf{a}^{2(t)}}\mathcal{L}$$

$$\nabla_{\mathbf{b}^1}\mathcal{L} = \sum_{t=T}^{1} \nabla_{\mathbf{a}^{2(t)}}\mathcal{L}$$

Summary

Hidden layer

$$\nabla_{\mathbf{h}^{1(t)}}\mathcal{L} = \nabla_{\mathbf{a}^{2(t)}}\mathcal{L} W^1 + \underbrace{\nabla_{\mathbf{a}^{1(t+1)}}\mathcal{L} W^{11}}_{\text{Not required when } t = T}$$

$$\nabla_{\mathbf{a}^{1(t)}}\mathcal{L} = \nabla_{\mathbf{h}^{1(t)}}\mathcal{L} \underbrace{\nabla_{\mathbf{a}^{1(t)}}\mathbf{h}^{1(t)}}_{\text{Jacobian}}$$

$$\nabla_{W^{11}}\mathcal{L} = \sum_{t=T}^{1} \mathbf{h}^{1(t-1)} \nabla_{\mathbf{a}^{1(t)}}\mathcal{L}$$

$$\nabla_{W^{0}}\mathcal{L} = \sum_{t=T}^{1} \mathbf{h}^{0(t)} \nabla_{\mathbf{a}^{1(t)}}\mathcal{L}$$

$$\nabla_{\mathbf{b}^{0}}\mathcal{L} = \sum_{t=T}^{1} \nabla_{\mathbf{a}^{1(t)}}\mathcal{L}$$