

How to compute a derivative

Computing derivatives of complicated functions

- How do you compute the derivatives in an LSTM or GRU cell?
- How do you compute derivatives of complicated functions *in general*
- In these slides we will give you some hints
- In the slides we will assume vector functions and vector activations
- But we will also give you scalar versions of the equations to provide intuition
- The two sets will be almost identical, except that when we deal with vector functions
 - The notation becomes uglier and less intuitive
 - We must ensure that the dimensions come out right
- Please compare vector versions of equations to their scalar counterparts for better intuition, if needed

First: Some notation and conventions

- We will refer to the derivative of scalar L with respect to x as $\nabla_x L$
 - Regardless of whether the derivative is a scalar, vector, matrix or tensor
- The derivative of a scalar L w.r.t an $N \times 1$ column vector x is a $1 \times N$ row vector $\nabla_x L$
- The derivative of a L w.r.t an $N \times M$ matrix X is a $M \times N$ matrix $\nabla_X L$
 - Remember our gradient update rule : $W = W - \eta \nabla_W L^T$
- The derivative of an $N \times 1$ vector Y w.r.t an $M \times 1$ vector X is an $N \times M$ matrix $J_X(Y)$
 - The Jacobian

Rules: 1 (scalar)

$$z = Wx$$

- All terms are scalars
- $\frac{\partial L}{\partial z}$ is known

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} W$$

$$\frac{\partial L}{\partial W} = x \frac{\partial L}{\partial z}$$

Rules: 1 (vector)

$$z = Wx$$

- z is an $N \times 1$ vector
- x is an $M \times 1$ vector
- W is an $N \times M$ matrix
- L is a function of z
- $\nabla_z L$ is known (and is a $1 \times N$ vector)

$$\nabla_x L = (\nabla_z L)W$$

$$\nabla_W L = x(\nabla_z L)$$

Please verify that the dimensions match!

Rules: 2 (vector, *schur* multiply)

$$z = x \circ y$$

- x, y and z are all $N \times 1$ vectors
- “ \circ ” represents component-wise multiplication
- $\nabla_z L$ is known (and is a $1 \times N$ vector)

$$\nabla_x L = (\nabla_z L) \circ y^T$$

$$\nabla_y L = (\nabla_z L) \circ x^T$$

Please verify that the dimensions match!

Rules: 3 (scalar)

$$z = x + y$$

- All terms are scalars
- $\frac{\partial L}{\partial z}$ is known

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}$$

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

Rules: 3 (vector)

$$z = x + y$$

- x, y and z are all $N \times 1$ vectors
- $\nabla_z L$ is known (and is a $1 \times N$ vector)

$$\nabla_x L = \nabla_z L$$

$$\nabla_y L = \nabla_z L$$

Please verify that the dimensions match!

Rules: 4 (scalar)

$$z = g(x)$$

- x and z are scalars
- $\frac{\partial L}{\partial z}$ is known

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} g'(x)$$

Rules: 4 (vector)

$$z = g(x)$$

- x and z are $N \times 1$ vectors
- $\nabla_z L$ is known (and is a $1 \times N$ vector)
- $J_x g$ is the *Jacobian* of $g(x)$ with respect to x
 - May be a diagonal matrix

$$\nabla_x L = \nabla_z L J_x g$$

Please verify that the dimensions match!

Rules: 4b (vector) component-wise multiply notation

$$z = g(x)$$

- x and z are $N \times 1$ vectors
- $\nabla_z L$ is known (and is a $1 \times N$ vector)
- $g'(x)$ is a column vector consisting of the derivatives of the individual components of $g(x)$ w.r.t individual components of x

$$\nabla_x L = \nabla_z L \circ g'(x)^T$$

Please verify that the dimensions match!

Rule 5: Addition of derivatives

- Given two variables

$$\begin{aligned}z &= g(x) \\ y &= h(x)\end{aligned}$$

- And given $\frac{\partial L}{\partial y}$ and $\frac{\partial L}{\partial z}$

- we get

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} g'(x) + \frac{\partial L}{\partial y} h'(x)$$

- The rule also extends to vector derivatives

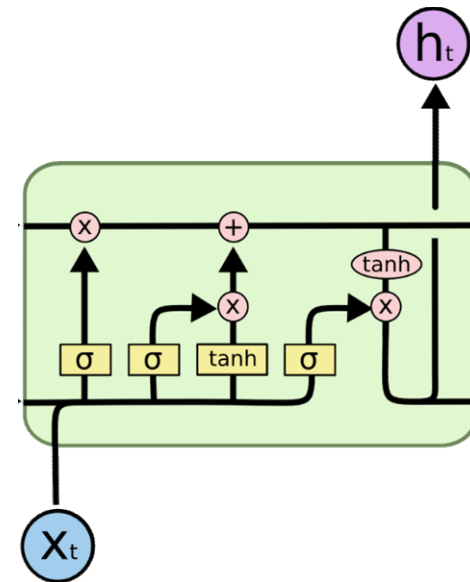
Computing derivatives of complex functions

- We now are prepared to compute very complex derivatives
- Procedure:
 - Express the computation as a series of computations of intermediate values
 - Each computation must comprise either a unary or binary relation
 - Unary relation: RHS has one argument, e.g. $y = g(x)$
 - Binary relation: RHS has two arguments
e.g. $z = x + y$ or $z = xy$
 - Work your way backward through the derivatives of the simple relations

Example: LSTM

- Full set of LSTM equations (in the order in which they must be computed)

- 1 $f_t = \sigma(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f)$
- 2 $i_t = \sigma(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i)$
- 3 $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- 4 $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
- 5 $o_t = \sigma(W_o \cdot [C_t, h_{t-1}, x_t] + b_o)$
- 6 $h_t = o_t * \tanh(C_t)$



- Its actually much cleaner to separate the individual components, so lets do that first

LSTM

$$1. \quad f_t = \sigma(W_{fC}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$$

$$2. \quad i_t = \sigma(W_{iC}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$$

$$3. \quad \tilde{C}_t = \sigma(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C)$$

$$4. \quad C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$

$$5. \quad o_t = \sigma(W_{oC}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$$

$$6. \quad h_t = o_t \circ \tanh(C_t)$$

- This is the full set of equations *in the order in which they must be computed*
- Lets rewrite these in terms of unary and binary operations

LSTM

1. $f_t = \sigma(W_{fC}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$
2. $i_t = \sigma(W_{iC}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$
3. $\tilde{C}_t = \sigma(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C)$
4. $C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$
5. $o_t = \sigma(W_{oC}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$
6. $h_t = o_t \circ \tanh(C_t)$

$$\begin{aligned}z_1 &= W_{fC}C_{t-1} \\z_2 &= W_{fh}h_{t-1} \\z_3 &= z_1 + z_2 \\z_4 &= W_{fx}x_t \\z_5 &= z_3 + z_4 \\z_6 &= z_5 + b_f \\f_t &= \sigma(z_6)\end{aligned}$$

- Lets rewrite these in terms of unary and binary operations

LSTM

1. $z_1 = W_{fc} C_{t-1}$

2. $z_2 = W_{fh} h_{t-1}$

3. $z_3 = z_1 + z_2$

4. $z_4 = W_{fx} x_t$

5. $z_5 = z_3 + z_4$

6. $z_6 = z_5 + b_f$

7. $f_t = \sigma(z_6)$

LSTM

1. $f_t = \sigma(W_{fc}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$
2. $i_t = \sigma(W_{ic}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$
3. $\tilde{C}_t = \sigma(W_{ch}h_{t-1} + W_{cx}x_t + b_c)$
4. $C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$
5. $o_t = \sigma(W_{oc}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$
6. $h_t = o_t \circ \tanh(C_t)$

$$\begin{aligned}z_7 &= W_{ic}C_{t-1} \\z_8 &= W_{ih}h_{t-1} \\z_9 &= z_7 + z_8 \\z_{10} &= W_{ix}x_t \\z_{11} &= z_9 + z_{10} \\z_{12} &= z_{11} + b_i \\i_t &= \sigma(z_{12})\end{aligned}$$

- Lets rewrite these in terms of unary and binary operations

LSTM

1. $z_1 = W_{fC} C_{t-1}$

2. $z_2 = W_{fh} h_{t-1}$

3. $z_3 = z_1 + z_2$

4. $z_4 = W_{fx} x_t$

5. $z_5 = z_3 + z_4$

6. $z_6 = z_5 + b_f$

7. $f_t = \sigma(z_6)$

8. $z_7 = W_{iC} C_{t-1}$

9. $z_8 = W_{ih} h_{t-1}$

10. $z_9 = z_7 + z_8$

11. $z_{10} = W_{ix} x_t$

12. $z_{11} = z_9 + z_{10}$

13. $z_{12} = z_{11} + b_i$

14. $i_t = \sigma(z_{12})$

LSTM

1. $f_t = \sigma(W_{fC}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$
2. $i_t = \sigma(W_{iC}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$
3. $\tilde{C}_t = \sigma(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C)$
4. $C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$
5. $o_t = \sigma(W_{oC}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$
6. $h_t = o_t \circ \tanh(C_t)$

$$\begin{aligned}z_{13} &= W_{Ch}h_{t-1} \\z_{14} &= W_{Cx}x_t \\z_{15} &= z_{13} + z_{14} \\z_{16} &= z_{15} + b_C \\ \tilde{C}_t &= \sigma(z_{16})\end{aligned}$$

- Lets rewrite these in terms of unary and binary operations

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

LSTM

1. $f_t = \sigma(W_{fC}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$
2. $i_t = \sigma(W_{iC}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$
3. $\tilde{C}_t = \sigma(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C)$
4. $C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$
5. $o_t = \sigma(W_{oC}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$
6. $h_t = o_t \circ \tanh(C_t)$

$$\begin{aligned}z_{17} &= f_t \circ C_{t-1} \\z_{18} &= i_t \circ \tilde{C}_t \\C_t &= z_{17} + z_{18}\end{aligned}$$

- Lets rewrite these in terms of unary and binary operations

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

LSTM

1. $f_t = \sigma(W_{fC}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$
2. $i_t = \sigma(W_{iC}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$
3. $\tilde{C}_t = \sigma(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C)$
4. $C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$
5. $o_t = \sigma(W_{oC}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$
6. $h_t = o_t \circ \tanh(C_t)$

$$\begin{aligned}z_{19} &= W_{oC}C_{t-1} \\z_{20} &= W_{oh}h_{t-1} \\z_{21} &= z_{19} + z_{20} \\z_{22} &= W_{ox}x_t \\z_{23} &= z_{21} + z_{22} \\z_{24} &= z_{23} + b_o \\o_t &= \sigma(z_{24})\end{aligned}$$

- Lets rewrite these in terms of unary and binary operations

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$23. z_{19} = W_{oC} C_{t-1}$$

$$24. z_{20} = W_{oh} h_{t-1}$$

$$25. z_{21} = z_{19} + z_{20}$$

$$26. z_{22} = W_{ox} x_t$$

$$27. z_{23} = z_{21} + z_{22}$$

$$28. z_{24} = z_{23} + b_o$$

$$29. o_t = \sigma(z_{24})$$

LSTM

1. $f_t = \sigma(W_{fC}C_{t-1} + W_{fh}h_{t-1} + W_{fx}x_t + b_f)$
2. $i_t = \sigma(W_{iC}C_{t-1} + W_{ih}h_{t-1} + W_{ix}x_t + b_i)$
3. $\tilde{C}_t = \sigma(W_{Ch}h_{t-1} + W_{Cx}x_t + b_C)$
4. $C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$
5. $o_t = \sigma(W_{oC}C_{t-1} + W_{oh}h_{t-1} + W_{ox}x_t + b_o)$
6. $h_t = o_t \circ \tanh(C_t)$

$$z_{25} = \tanh(C_t)$$
$$h_t = o_t \circ z_{25}$$

- Lets rewrite these in terms of unary and binary operations

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$23. z_{19} = W_{oC} C_{t-1}$$

$$24. z_{20} = W_{oh} h_{t-1}$$

$$25. z_{21} = z_{19} + z_{20}$$

$$26. z_{22} = W_{ox} x_t$$

$$27. z_{23} = z_{21} + z_{22}$$

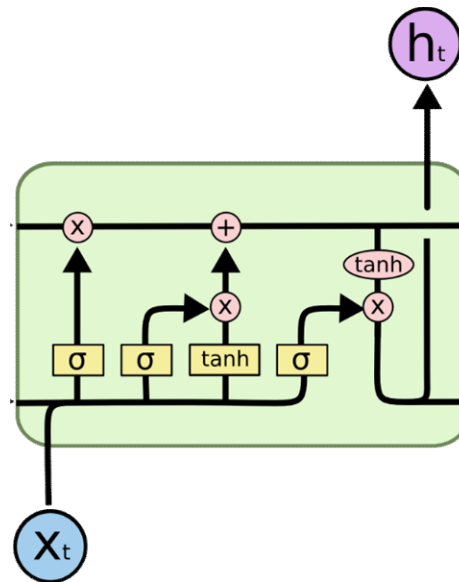
$$28. z_{24} = z_{23} + b_o$$

$$29. o_t = \sigma(z_{24})$$

$$30. z_{25} = \tanh(C_t)$$

$$31. h_t = o_t \circ z_{25}$$

LSTM forward



- The full forward computation of the LSTM can be performed by computing Equations 1-31 in sequence
- Every one of these equations is unary or binary

LSTM

1. $z_1 = W_{fC} C_{t-1}$

2. $z_2 = W_{fh} h_{t-1}$

3. $z_3 = z_1 + z_2$

4. $z_4 = W_{fx} x_t$

5. $z_5 = z_3 + z_4$

6. $z_6 = z_5 + b_f$

7. $f_t = \sigma(z_6)$

8. $z_7 = W_{iC} C_{t-1}$

9. $z_8 = W_{ih} h_{t-1}$

10. $z_9 = z_7 + z_8$

11. $z_{10} = W_{ix} x_t$

12. $z_{11} = z_9 + z_{10}$

13. $z_{12} = z_{11} + b_i$

14. $i_t = \sigma(z_{12})$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$23. z_{19} = W_{oC} C_{t-1}$$

$$24. z_{20} = W_{oh} h_{t-1}$$

$$25. z_{21} = z_{19} + z_{20}$$

$$26. z_{22} = W_{ox} x_t$$

$$27. z_{23} = z_{21} + z_{22}$$

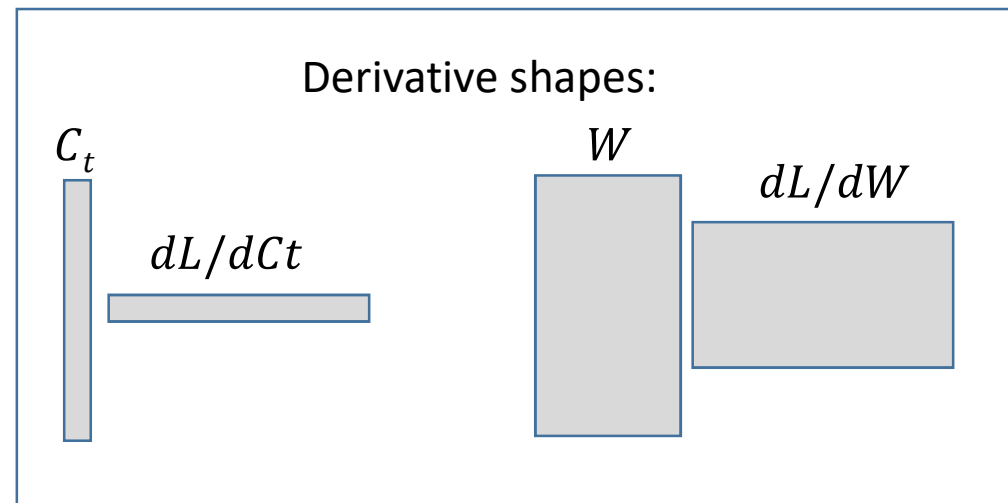
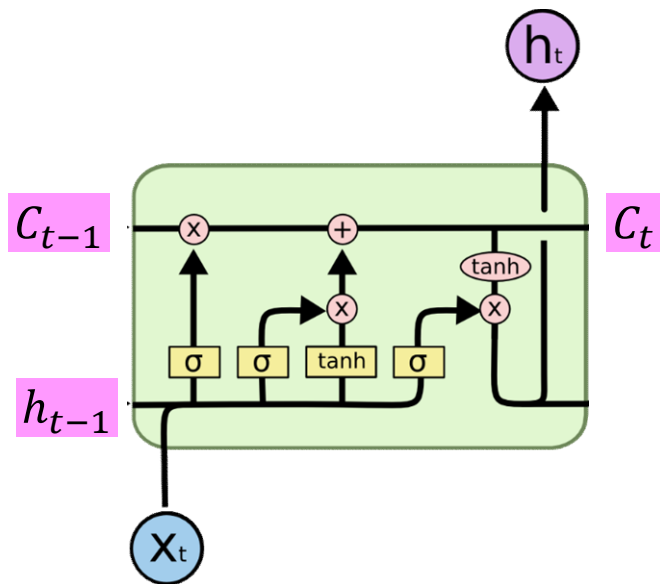
$$28. z_{24} = z_{23} + b_o$$

$$29. o_t = \sigma(z_{24})$$

$$30. z_{25} = \tanh(C_t)$$

$$31. h_t = o_t \circ z_{25}$$

Computing derivatives



- We will now work our way backward
- We assume derivatives $\frac{dL}{dh_t}$ and $\frac{dL}{dC_t}$ of the loss w.r.t h_t and C_t are given
- We must compute $\frac{dL}{dC_{t-1}}$, $\frac{dL}{dh_{t-1}}$ and $\frac{dL}{dx_t}$
 - And also derivatives w.r.t the parameters within the cell
- Recall: the shape of the derivative for any variable will be transposed with respect to that variable

LSTM

$$1. \quad \nabla_{o_t} L = \nabla_{h_t} L \circ z_{25}^T$$

$$2. \quad \nabla_{z_{25}} L = \nabla_{h_t} L \circ o_t^T$$

$$23. \quad z_{19} = W_{oc} C_{t-1}$$

$$24. \quad z_{20} = W_{oh} h_{t-1}$$

$$25. \quad z_{21} = z_{19} + z_{20}$$

$$26. \quad z_{22} = W_{ox} x_t$$

$$27. \quad z_{23} = z_{21} + z_{22}$$

$$28. \quad z_{24} = z_{23} + b_o$$

$$29. \quad o_t = \sigma(z_{24})$$

$$30. \quad z_{25} = \tanh(C_t)$$

$$31. \quad h_t = o_t \circ z_{25}$$

LSTM

1. $\nabla_{o_t} L = \nabla_{h_t} L \circ z_{25}^T$
2. $\nabla_{z_{25}} L = \nabla_{h_t} L \circ o_t^T$
3. $\nabla_{C_t} L = \nabla_{z_{25}} L \circ (1 - \tanh^2(C_t))^T$

23. $z_{19} = W_{oc} C_{t-1}$

24. $z_{20} = W_{oh} h_{t-1}$

25. $z_{21} = z_{19} + z_{20}$

26. $z_{22} = W_{ox} x_t$

27. $z_{23} = z_{21} + z_{22}$

28. $z_{24} = z_{23} + b_o$

29. $o_t = \sigma(z_{24})$

30. $z_{25} = \tanh(C_t)$

31. $h_t = o_t \circ z_{25}$

LSTM

1. $\nabla_{o_t} L = \nabla_{h_t} L \circ z_{25}^T$
2. $\nabla_{z_{25}} L = \nabla_{h_t} L \circ o_t^T$
3. $\nabla_{C_t} L = \nabla_{z_{25}} L \circ (1 - \tanh^2(C_t))^T$
4. $\nabla_{z_{24}} L = \nabla_{o_t} L \circ \sigma(z_{24})^T \circ (1 - \sigma(z_{24}))^T$

23. $z_{19} = W_{oc} C_{t-1}$

24. $z_{20} = W_{oh} h_{t-1}$

25. $z_{21} = z_{19} + z_{20}$

26. $z_{22} = W_{ox} x_t$

27. $z_{23} = z_{21} + z_{22}$

28. $z_{24} = z_{23} + b_o$

29. $o_t = \sigma(z_{24})$

30. $z_{25} = \tanh(C_t)$

31. $h_t = o_t \circ z_{25}$

LSTM

1. $\nabla_{o_t} L = \nabla_{h_t} L \circ z_{25}^T$
2. $\nabla_{z_{25}} L = \nabla_{h_t} L \circ o_t^T$
3. $\nabla_{C_t} L = \nabla_{z_{25}} L \circ (1 - \tanh^2(C_t))^T$
4. $\nabla_{z_{24}} L = \nabla_{o_t} L \circ \sigma(z_{24})^T \circ (1 - \sigma(z_{24}))^T$
5. $\nabla_{z_{23}} L = \nabla_{z_{24}} L$
6. $\nabla_{b_o} L = \nabla_{z_{24}} L$

23. $z_{19} = W_{oc} C_{t-1}$
24. $z_{20} = W_{oh} h_{t-1}$
25. $z_{21} = z_{19} + z_{20}$
26. $z_{22} = W_{ox} x_t$
27. $z_{23} = z_{21} + z_{22}$
28. $z_{24} = z_{23} + b_o$
29. $o_t = \sigma(z_{24})$
30. $z_{25} = \tanh(C_t)$
31. $h_t = o_t \circ z_{25}$

Equations highlighted in yellow show derivatives w.r.t. parameters

LSTM

$$7. \quad \nabla_{z_{22}} L = \nabla_{z_{23}} L$$

$$8. \quad \nabla_{z_{21}} L = \nabla_{z_{23}} L$$

$$23. \quad z_{19} = W_{oc} C_{t-1}$$

$$24. \quad z_{20} = W_{oh} h_{t-1}$$

$$25. \quad z_{21} = z_{19} + z_{20}$$

$$26. \quad z_{22} = W_{ox} x_t$$

$$27. \quad z_{23} = z_{21} + z_{22}$$

$$28. \quad z_{24} = z_{23} + b_o$$

$$29. \quad o_t = \sigma(z_{24})$$

$$30. \quad z_{25} = \tanh(C_t)$$

$$31. \quad h_t = o_t \circ z_{25}$$

LSTM

$$7. \quad \nabla_{z_{22}} L = \nabla_{z_{23}} L$$

$$8. \quad \nabla_{z_{21}} L = \nabla_{z_{23}} L$$

$$9. \quad \nabla_{W_{ox}} L = x_t \nabla_{z_{22}} L$$

$$10. \quad \nabla_{x_t} L = \nabla_{z_{22}} L W_{ox}$$

$$23. \quad z_{19} = W_{oc} C_{t-1}$$

$$24. \quad z_{20} = W_{oh} h_{t-1}$$

$$25. \quad z_{21} = z_{19} + z_{20}$$

$$26. \quad z_{22} = W_{ox} x_t$$

$$27. \quad z_{23} = z_{21} + z_{22}$$

$$28. \quad z_{24} = z_{23} + b_o$$

$$29. \quad o_t = \sigma(z_{24})$$

$$30. \quad z_{25} = \tanh(C_t)$$

$$31. \quad h_t = o_t \circ z_{25}$$

LSTM

$$7. \quad \nabla_{z_{22}} L = \nabla_{z_{23}} L$$

$$8. \quad \nabla_{z_{21}} L = \nabla_{z_{23}} L$$

$$9. \quad \nabla_{W_{ox}} L = x_t \nabla_{z_{22}} L$$

$$10. \quad \nabla_{x_t} L = \nabla_{z_{22}} L W_{ox}$$

$$11. \quad \nabla_{z_{20}} L = \nabla_{z_{21}} L$$

$$12. \quad \nabla_{z_{19}} L = \nabla_{z_{21}} L$$

$$23. \quad z_{19} = W_{oc} C_{t-1}$$

$$24. \quad z_{20} = W_{oh} h_{t-1}$$

$$25. \quad z_{21} = z_{19} + z_{20}$$

$$26. \quad z_{22} = W_{ox} x_t$$

$$27. \quad z_{23} = z_{21} + z_{22}$$

$$28. \quad z_{24} = z_{23} + b_o$$

$$29. \quad o_t = \sigma(z_{24})$$

$$30. \quad z_{25} = \tanh(C_t)$$

$$31. \quad h_t = o_t \circ z_{25}$$

LSTM

$$7. \quad \nabla_{z_{22}} L = \nabla_{z_{23}} L$$

$$8. \quad \nabla_{z_{21}} L = \nabla_{z_{23}} L$$

$$9. \quad \nabla_{W_{ox}} L = x_t \nabla_{z_{22}} L$$

$$10. \quad \nabla_{x_t} L = \nabla_{z_{22}} L W_{ox}$$

$$11. \quad \nabla_{z_{20}} L = \nabla_{z_{21}} L$$

$$12. \quad \nabla_{z_{19}} L = \nabla_{z_{21}} L$$

$$13. \quad \nabla_{W_{oh}} L = h_{t-1} \nabla_{z_{20}} L$$

$$14. \quad \nabla_{h_{t-1}} L = \nabla_{z_{20}} L W_{oh}$$

$$23. \quad z_{19} = W_{oc} C_{t-1}$$

$$24. \quad z_{20} = W_{oh} h_{t-1}$$

$$25. \quad z_{21} = z_{19} + z_{20}$$

$$26. \quad z_{22} = W_{ox} x_t$$

$$27. \quad z_{23} = z_{21} + z_{22}$$

$$28. \quad z_{24} = z_{23} + b_o$$

$$29. \quad o_t = \sigma(z_{24})$$

$$30. \quad z_{25} = \tanh(C_t)$$

$$31. \quad h_t = o_t \circ z_{25}$$

LSTM

$$7. \quad \nabla_{z_{22}} L = \nabla_{z_{23}} L$$

$$8. \quad \nabla_{z_{21}} L = \nabla_{z_{23}} L$$

$$9. \quad \nabla_{W_{ox}} L = x_t \nabla_{z_{22}} L$$

$$10. \quad \nabla_{x_t} L = \nabla_{z_{22}} L W_{ox}$$

$$11. \quad \nabla_{z_{20}} L = \nabla_{z_{21}} L$$

$$12. \quad \nabla_{z_{19}} L = \nabla_{z_{21}} L$$

$$13. \quad \nabla_{W_{oh}} L = h_{t-1} \nabla_{z_{20}} L$$

$$14. \quad \nabla_{h_{t-1}} L = \nabla_{z_{20}} L W_{oh}$$

$$15. \quad \nabla_{W_{oc}} L = C_{t-1} \nabla_{z_{19}} L$$

$$16. \quad \nabla_{C_{t-1}} L = \nabla_{z_{19}} L W_{oc}$$

$$23. \quad z_{19} = W_{oc} C_{t-1}$$

$$24. \quad z_{20} = W_{oh} h_{t-1}$$

$$25. \quad z_{21} = z_{19} + z_{20}$$

$$26. \quad z_{22} = W_{ox} x_t$$

$$27. \quad z_{23} = z_{21} + z_{22}$$

$$28. \quad z_{24} = z_{23} + b_o$$

$$29. \quad o_t = \sigma(z_{24})$$

$$30. \quad z_{25} = \tanh(C_t)$$

$$31. \quad h_t = o_t \circ z_{25}$$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$7. \nabla_{z_{18}} L = \nabla_{C_t} L$$

$$8. \nabla_{z_{17}} L = \nabla_{C_t} L$$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$7. \nabla_{z_{18}} L = \nabla_{C_t} L$$

$$8. \nabla_{z_{17}} L = \nabla_{C_t} L$$

$$9. \nabla_{i_t} L = \nabla_{z_{18}} L \circ \tilde{C}_t^T$$

$$10. \nabla_{\tilde{C}_t} L = \nabla_{z_{18}} L \circ i_t^T$$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$7. \nabla_{z_{18}} L = \nabla_{C_t} L$$

$$8. \nabla_{z_{17}} L = \nabla_{C_t} L$$

$$9. \nabla_{i_t} L = \nabla_{z_{18}} L \circ \tilde{C}_t^T$$

$$10. \nabla_{\tilde{C}_t} L = \nabla_{z_{18}} L \circ i_t^T$$

$$11. \nabla_{C_{t-1}} L += \nabla_{z_{17}} L \circ f_t^T$$

$$12. \nabla_{f_t} L = \nabla_{z_{17}} L \circ C_{t-1}^T$$

Second time we're computing a derivative for C_{t-1} , so we increment the derivative ("+=")

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$7. \nabla_{z_{18}} L = \nabla_{C_t} L$$

$$8. \nabla_{z_{17}} L = \nabla_{C_t} L$$

$$9. \nabla_{i_t} L = \nabla_{z_{18}} L \circ \tilde{C}_t^T$$

$$10. \nabla_{\tilde{C}_t} L = \nabla_{z_{18}} L \circ i_t^T$$

$$11. \nabla_{C_{t-1}} L += \nabla_{z_{17}} L \circ f_t^T$$

$$12. \nabla_{f_t} L = \nabla_{z_{17}} L \circ C_{t-1}^T$$

$$13. \nabla_{z_{16}} L = \nabla_{\tilde{C}_t} L \circ \sigma(z_{16})^T \circ (1 - \sigma(z_{16}))^T$$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$14. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$15. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$14. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$15. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

$$16. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$17. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$14. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$15. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

$$16. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$17. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

$$18. \nabla_{W_{Cx}} L = x_t \nabla_{z_{14}} L$$

$$19. \nabla_{x_t} L += \nabla_{z_{14}} L W_{Cx}$$

Note the "+="

LSTM

$$15. z_{13} = W_{Ch} h_{t-1}$$

$$16. z_{14} = W_{Cx} x_t$$

$$17. z_{15} = z_{13} + z_{14}$$

$$18. z_{16} = z_{15} + b_C$$

$$19. \tilde{C}_t = \sigma(z_{16})$$

$$20. z_{17} = f_t \circ C_{t-1}$$

$$21. z_{18} = i_t \circ \tilde{C}_t$$

$$22. C_t = z_{17} + z_{18}$$

$$14. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$15. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

$$16. \nabla_{b_C} L = \nabla_{z_{16}} L$$

$$17. \nabla_{z_{15}} L = \nabla_{z_{16}} L$$

$$18. \nabla_{W_{Cx}} L = x_t \nabla_{z_{14}} L$$

$$19. \nabla_{x_t} L += \nabla_{z_{14}} L W_{Cx}$$

$$20. \nabla_{W_{Ch}} L = h_{t-1} \nabla_{z_{14}} L$$

$$21. \nabla_{h_{t-1}} L += \nabla_{z_{13}} L W_{Ch}$$

Note the "+="

Continuing the computation

- Continue the backward progression until the derivatives from forward Equation 1 have been computed
- At this point all derivatives will be computed.

Overall procedure

- Express the overall computation as a sequence of unary or binary operations
 - Can be automated
- Computes derivatives incrementally, going backward over the sequence of equations!
- Since each atomic computation is simple and belongs to one of a small set of possibilities, the conversion to derivatives is trivial once the computation is serialized as above

May be easier to think of it in terms of a “derivative” routine

- Define a routine that returns derivatives for unary and binary operations
- **SCALAR version (all variables are scalars)**

```
function deriv(dz, x, y, operator)
    case operator:
        `none' : return dx
        `*' : return x*dz, dz*x
        `+' : return dz, dz
        `-' : return dz, -dz
        # Single argument operations
        `tanh' : return dz(1-tanh2(x))
        `sigmoid' : return dz sigmoid(x) (1-sigmoid(x))
```

Derivative routine, vector version

- Note distinction between component-wise and matrix multiplies
- Observe also that matrix and vector dimensions are correctly handled (locally)
- “ \circ ” is component-wise multiply
- “ $*$ ” is matrix multiply

```
function deriv(dz, x, y, operator)
    case operator:
        'none' : return dx
        # component-wise "schur" multiply
        'o' : return dz o xT, dz o yT
        # Matrix multiply. X must be a matrix
        '*' : return x*dz, dz*x
        '+' : return dz, dz
        '-' : return dz, -dz
        # The following will expect a single argument
        'tanh' : return dz o (1-tanh2(x))T
        'sigmoid' : return dz o sigmoid(x)T o (1-sigmoid(x))T
        # The jacobian is the full derivative matrix of the sigmoid
        'softmax' : return dz*Jacobian(sigmoid,x)
```

When to use “=” vs “+=”

- In the forward computation a variable may be used multiple times to compute other intermediate variables
- During backward computations, the first time the derivative is computed for the variable, the we will use “=”
- In subsequent computations we use “+=”
- It may be difficult to keep track of when we first compute the derivative for a variable
 - When to use “=” vs when to use “+=”
- Cheap trick:
 - Initialize all derivatives to 0 during computation
 - *Always* use “+=”
 - You will get the correct answer (why?)

```

[dCt-1, dxt, dht-1, d[W, b]] = LSTM_derivative(dCt dht)
initialize d(variable)=0 (all variables)
# Derivative of eq. 31  $h_t = o_t \circ z_{25}$ 
[dot, dz25] += deriv(dht, ot, z25, 'o')
# Derivative of eq. 30  $z_{25} = \tanh(C_t)$ 
[dCt] += deriv(dz25, Ct, 'tanh')
# Derivative of eq. 29  $o_t = \sigma(z_{24})$ 
[dz25] += deriv(dot, z25, 'sigmoid')
# Derivative of eq. 28  $z_{24} = z_{23} + b_o$ 
[dz23, dbo] += deriv(dz24, z23, bo, '+')
# Derivative of eq. 27  $z_{23} = z_{21} + z_{22}$ 
[dz21, dz22] += deriv(dz23, z21, z22, '+')
# Derivative of eq. 26  $z_{22} = W_{ox}x_t$ 
[dWox, dxt] += deriv(dz22, Wox, xt, '*')
# Derivative of eq. 25  $z_{21} = z_{19} + z_{20}$ 
[dz19, dz20] += deriv(dz21, z19, z20, '+')
# Derivative of eq. 24  $z_{20} = W_{oh}h_{t-1}$ 
[dWoh, dht-1] += deriv(dz20, Woh, ht-1, '*')
# Derivative of eq. 23  $z_{19} = W_{oc}C_{t-1}$ 
[dWoc, dCt-1] += deriv(dz19, Woc, Ct-1, '*')

```

```

23.  $z_{19} = W_{oc}C_{t-1}$ 
24.  $z_{20} = W_{oh}h_{t-1}$ 
25.  $z_{21} = z_{19} + z_{20}$ 
26.  $z_{22} = W_{ox}x_t$ 
27.  $z_{23} = z_{21} + z_{22}$ 
28.  $z_{24} = z_{23} + b_o$ 
29.  $o_t = \sigma(z_{24})$ 
30.  $z_{25} = \tanh(C_t)$ 
31.  $h_t = o_t \circ z_{25}$ 

```

... continued from previous slide

Derivative of eq. 22 $C_t = z_{17} + z_{18}$

[dz₁₇, dz₁₈] += deriv(dC_t, z₁₈, z₁₈, '+')

Derivative of eq. 21 $z_{18} = i_t \circ \tilde{C}_t$

[di_t, dtildeC_t] += deriv(dz₁₈, i_t, dtildeC_t, 'o')

Derivative of eq. 20 $z_{17} = f_t \circ C_{t-1}$

[df_t, dC_{t-1}] += deriv(dz₁₇, f_t, C_{t-1}, 'o')

Derivative of eq. 19 $\tilde{C}_t = \sigma(z_{16})$

[dz₁₆] += deriv(dtildedeC_t, 'sigmoid')

Derivative of eq. 18 $z_{16} = z_{15} + b_c$

[dz₁₅, db_c] += deriv(dz₁₆, z₁₅, b_c, '+')

Derivative of eq. 17 $z_{15} = z_{13} + z_{14}$

[dz₁₃, dz₁₄] += deriv(dz₁₅, z₁₃, z₁₄, '+')

Derivative of eq. 16 $z_{14} = W_{cx} x_t$

[dW_{cx}, dx_t] += deriv(dz₁₄, W_{cx}, x_t, '*')

Derivative of eq. 15 $z_{13} = W_{ch} h_{t-1}$

[dW_{ch}, dh_{t-1}] += deriv(dz₁₃, W_{ch}, h_{t-1}, '*')

15. $z_{13} = W_{ch} h_{t-1}$

16. $z_{14} = W_{cx} x_t$

17. $z_{15} = z_{13} + z_{14}$

18. $z_{16} = z_{15} + b_c$

19. $\tilde{C}_t = \sigma(z_{16})$

20. $z_{17} = f_t \circ C_{t-1}$

21. $z_{18} = i_t \circ \tilde{C}_t$

22. $C_t = z_{17} + z_{18}$

... continued from previous slide

Derivative of eq. 14 $i_t = \sigma(z_{12})$

[dz₁₂] += deriv(di_t, 'sigmoid')

Derivative of eq. 13 $z_{12} = z_{11} + b_f$

[dz₁₁, db_i] += deriv(dz₁₂, z₁₁, b_i, '+')

Derivative of eq. 12 $z_{11} = z_9 + z_{10}$

[dz₉, dz₁₀] += deriv(dz₁₁, z₉, z₁₀, '+')

Derivative of eq. 11 $z_{10} = W_{ix}x_t$

[dW_{ix}, dx_t] += deriv(dz₁₀, W_{ix}, x_t, '+')

Derivative of eq. 10 $z_9 = z_7 + z_8$

[dz₇, dz₈] += deriv(dz₉, z₇, z₈, '+')

Derivative of eq. 9 $z_8 = W_{ih}h_{t-1}$

[dW_{ih}, dh_{t-1}] += deriv(dz₈, W_{ih}, h_{t-1}, '*')

Derivative of eq. 8 $z_7 = W_{ic}C_{t-1}$

[dW_{ic}, dC_{t-1}] += deriv(dz₇, W_{ic}, C_{t-1}, '*')

8. $z_7 = W_{ic}C_{t-1}$

9. $z_8 = W_{ih}h_{t-1}$

10. $z_9 = z_7 + z_8$

11. $z_{10} = W_{ix}x_t$

12. $z_{11} = z_9 + z_{10}$

13. $z_{12} = z_{11} + b_i$

14. $i_t = \sigma(z_{12})$

... continued from previous slide

```
# Derivative of eq. 7  $f_t = \sigma(z_6)$ 
[dz_6] += deriv(df_t, 'sigmoid')
# Derivative of eq. 6  $z_6 = z_5 + b_f$ 
[dz_5, db_f] += deriv(dz_6, z_5, b_f, '+')
# Derivative of eq. 5  $z_5 = z_3 + z_4$ 
[dz_3, dz_4] += deriv(dz_5, z_3, z_4, '+')
# Derivative of eq. 4  $z_4 = W_{fx} x_t$ 
[dW_fx, dx_t] += deriv(dz_4, W_fx, x_t, '*')
# Derivative of eq. 3  $z_3 = z_1 + z_2$ 
[dz_1, dz_2] += deriv(dz_3, z_1, z_2, '+')
# Derivative of eq. 2  $z_2 = W_{fh} h_{t-1}$ 
[dW_fh, dh_{t-1}] += deriv(dz_2, W_fh, h_{t-1}, '*')
# Derivative of eq. 1  $z_1 = W_{fc} C_{t-1}$ 
[dW_fc, dC_{t-1}] += deriv(dz_1, W_fc, C_{t-1}, '*')

return dC_{t-1}, dh_{t-1}, dx_t, d[W, b]
```

1. $z_1 = W_{fc} C_{t-1}$
2. $z_2 = W_{fh} h_{t-1}$
3. $z_3 = z_1 + z_2$
4. $z_4 = W_{fx} x_t$
5. $z_5 = z_3 + z_4$
6. $z_6 = z_5 + b_f$
7. $f_t = \sigma(z_6)$

Caveats

- The `deriv()` routine given is missing several operators
 - Operations involving constants ($z = 2y$, $z = 1-y$, $z = 3+y$)
 - Division and inversion (e.g. $z = x/y$, $z = 1/y$, $z = A^{-1}$)
 - You may have to extend it to deal with these, or rewrite your equations to eliminate such operations if possible
- In practice many of the operations will be grouped together for computational efficiency
 - And to take advantage of parallel processing capabilities
- But the basic principle applies to *any* computation that can be expressed as a serial operation of unary and binary relations
 - If you can do it on a computer, you can express it as a serial operation
- In fact the preceding logic is *exactly* what we use to compute derivatives in backprop
 - We saw this explicitly in the vector version of BP for MLPs.