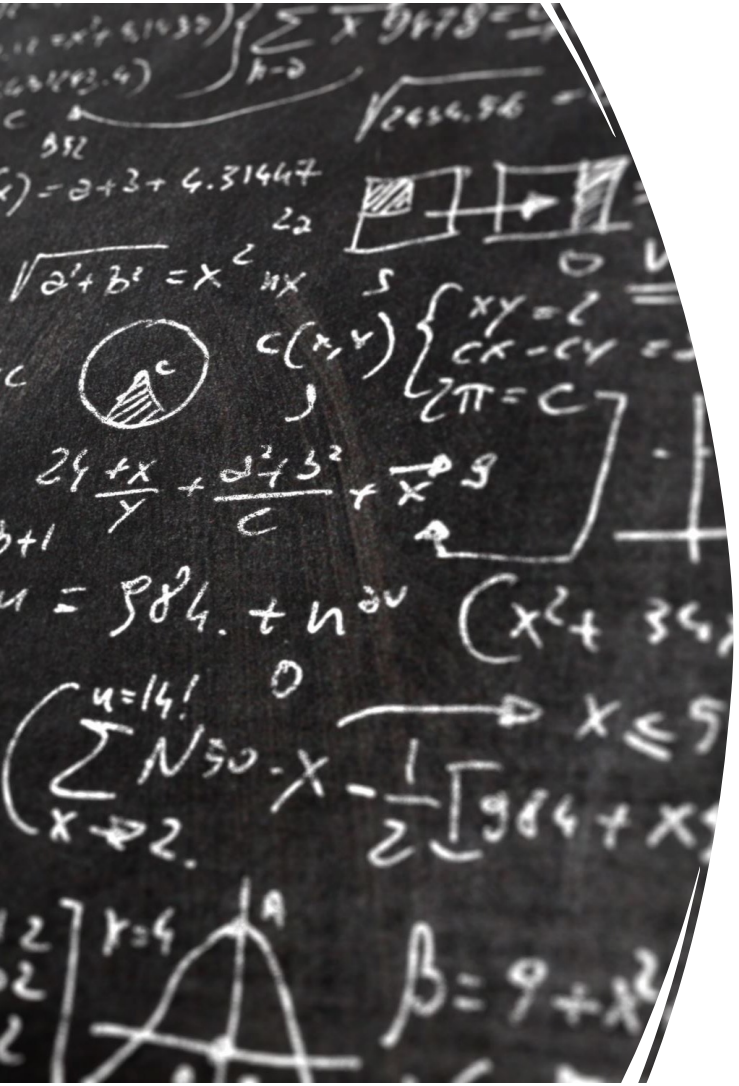


CS 383 – Computational Text Analysis

Lecture 10
Feed Forward Networks

Adam Poliak

02/20/2023

Slides adapted Dan Jurafsky, Jordan Boyd-Graber, Daniel Khashabi

Announcements

- HW04
  - Due next Wednesday 03/01
- Reading 05
  - Due Monday 02/27 - CTA/TADA/CSS papers using Word Embeddings,
  - lets move it to next Wednesday if I release today
- Office hours:
  - 3-4 pm today
  - 4-5 Thursday (tomorrow)

Final Project

Deliverables:

- Ideation
  250 write up – what idea do you have, who are you working with
  Due before Spring break

- Proposal
  Due after spring break

- Presentation
  Maybe last day of class

- Writeup, code, data
  End of finals?

Outline

Updating weights in LR

Computation Graph

Backpropagation

Issues when training NNs

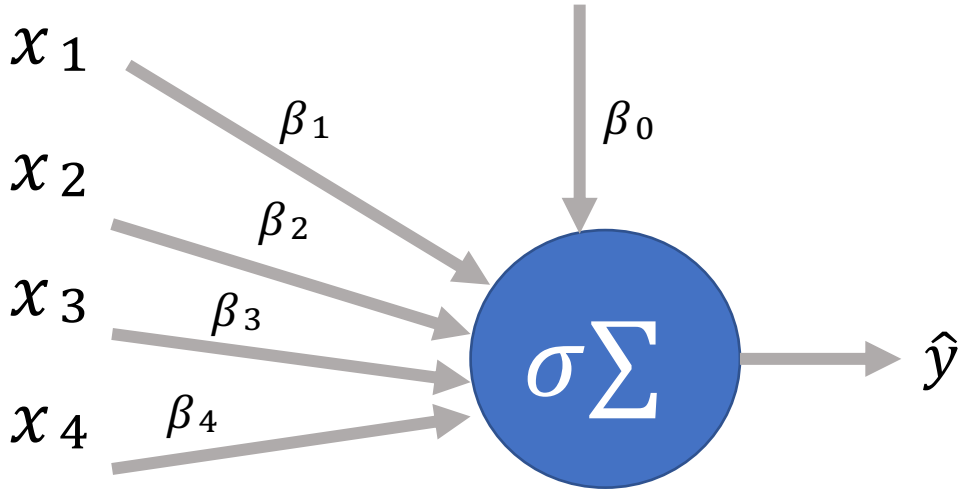
Pytorch

Deep Averaging Neural Network

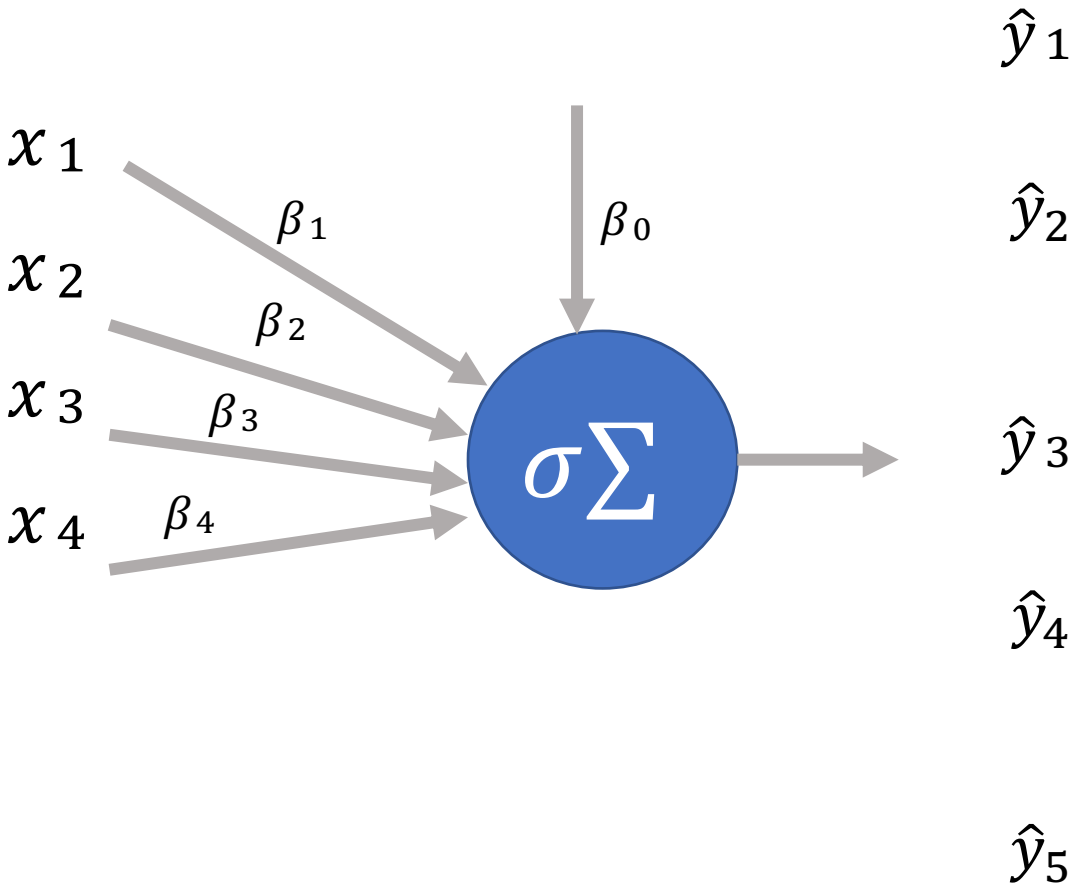
Logistic Regression

We make a prediction by applying sigmoid to the doc product of the features (covariates) and weights (coefficients)

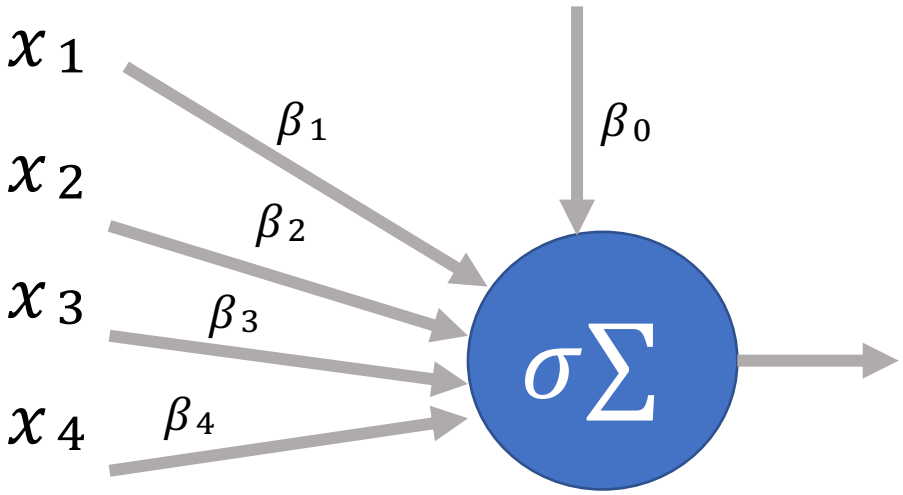
$$\sigma(\boldsymbol{\beta} \cdot \boldsymbol{x}) = \sum_i^j \sigma(\beta_i \cdot x_i)$$



Softmax


x_1
x_2
x_3
x_4
β_1
β_2
β_3
β_4
β_0
$\sigma\sum$
$\hat{y}_1$
$\hat{y}_2$
$\hat{y}_3$
$\hat{y}_4$
$\hat{y}_5$

Softmax



$\hat{y}_1$ 0.1

$\hat{y}_2$ 0.32

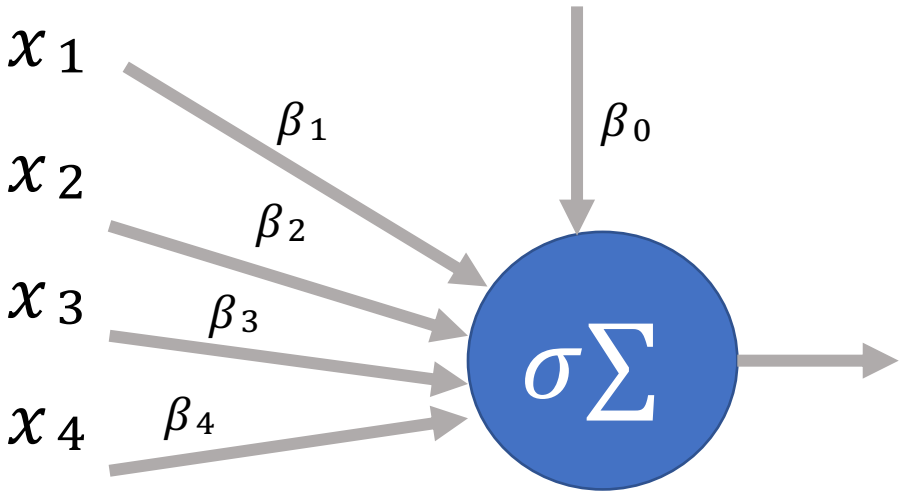
$\hat{y}_3$ 0.21

$\hat{y}_4$ 0.16

$\hat{y}_5$ 0.21

$$softmax(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)} \qquad 1 \leq i \leq k$$

Softmax



$\hat{y}_1$ 0.1

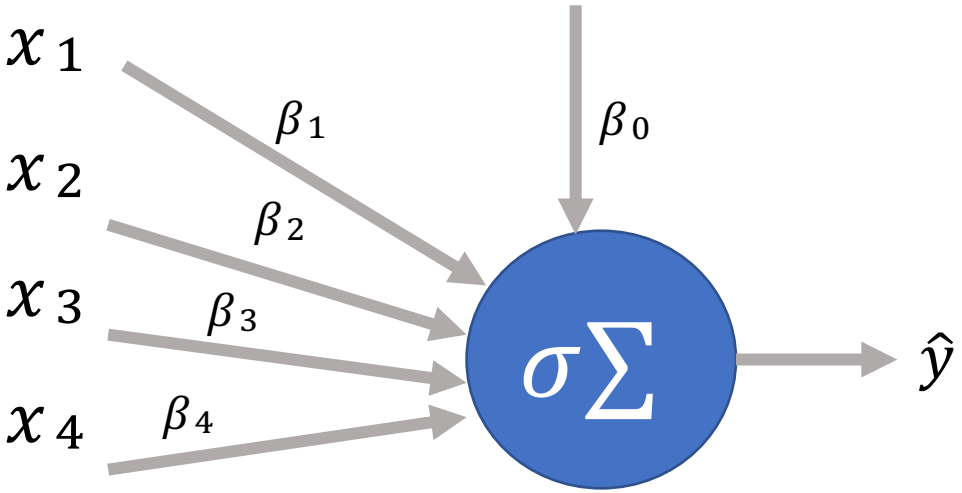
$\hat{y}_2$ 0.32

$\hat{y}_3$ 0.21

$\hat{y}_4$ 0.16

$\hat{y}_5$ 0.21

A neuron



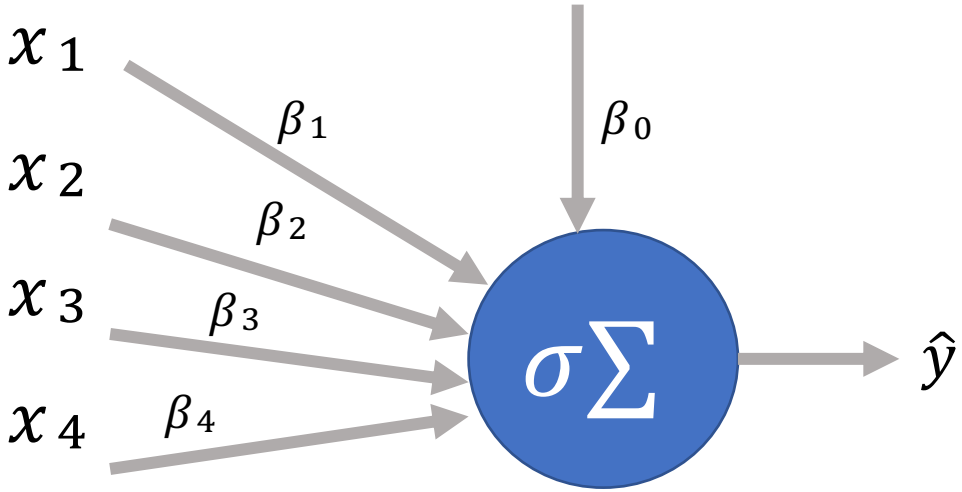
Logistic Regression as NN

A single layer neural network

Input layer: features

Output layer: prediction

We can pass the output of the neuron to another neuron

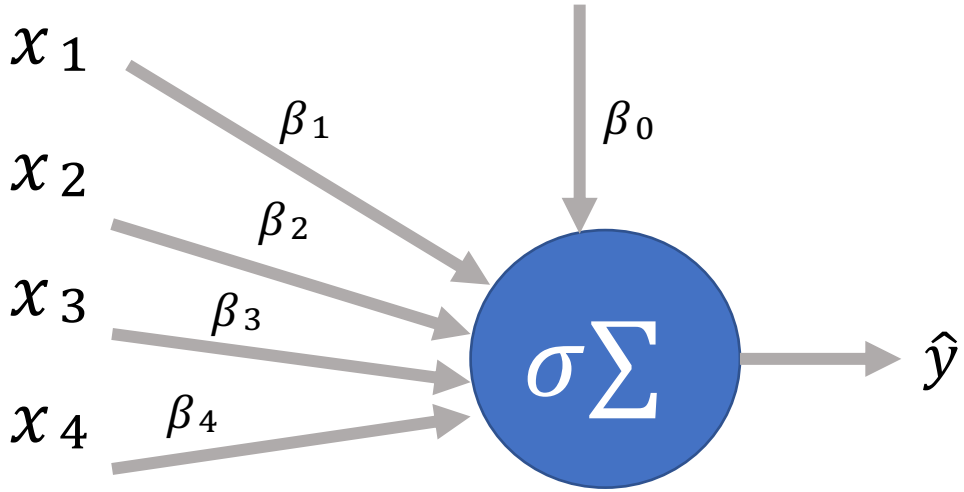


Updating weights

We want to minimize $\mathcal{L}(\hat{y}, y)$

For every weight we ask, how does changing the weight affect $\mathcal{L}(\hat{y}, y)$

Use partial derivatives to answer the question:



Updating weights

We want to minimize $\mathcal{L}(\hat{y}, y)$

For every weight we ask, how does changing the weight affect $\mathcal{L}(\hat{y}, y)$

Use partial derivatives to answer the question:

$$\nabla\mathcal{L}(\hat{y}, y) = \begin{bmatrix} \frac{\partial\mathcal{L}(\hat{y},y)}{\partial\beta_1} \\ \frac{\partial\mathcal{L}(\hat{y},y)}{\partial\beta_2} \\ \frac{\partial\mathcal{L}(\hat{y},y)}{\partial\beta_3} \\ \frac{\partial\mathcal{L}(\hat{y},y)}{\partial\beta_4} \\ \frac{\partial\mathcal{L}(\hat{y},y)}{\partial\beta_0} \end{bmatrix}$$

New loss function:

$$L(a, b, c) = b(a + 2b + ac^2)$$

We want to minimize this L

For every parameter we ask, how does changing the parameter affect this loss

Use partial derivatives to answer the question:

$$\nabla \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a} \\ \frac{\partial \mathcal{L}}{\partial b} \\ \frac{\partial \mathcal{L}}{\partial c} \end{bmatrix}$$

New loss function:

$$L(a, b, c) = b(a + 2b + ac^2)$$

What is $\frac{\partial \mathcal{L}}{\partial a}$?

$(a + 2b + ac^2)$ (1 + 2c)

What is $\frac{\partial \mathcal{L}}{\partial b}$?

$(a + 2b + ac^2)(2)$

What is $\frac{\partial \mathcal{L}}{\partial c}$?

$(a + 2b + ac^2)(2c)$

$$\nabla \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a} \\ \frac{\partial \mathcal{L}}{\partial b} \\ \frac{\partial \mathcal{L}}{\partial c} \end{bmatrix}$$

New loss function:

$$L(a, b, c) = b(a + 2b + ac^2)$$

What is $\frac{\partial \mathcal{L}}{\partial a}$?

$(a + 2b + ac^2)$ (1 + 2c)

What is $\frac{\partial \mathcal{L}}{\partial b}$?

$(a + 2b + ac^2)(2)$

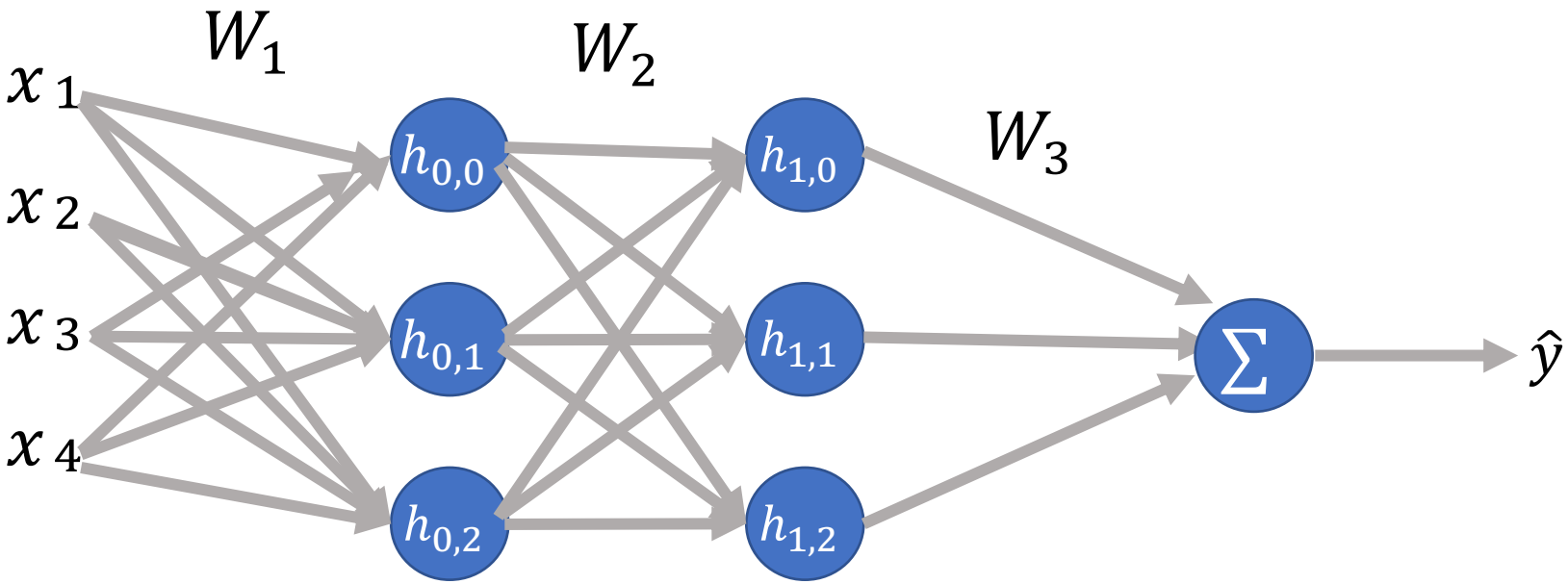
What is $\frac{\partial \mathcal{L}}{\partial c}$?

$(a + 2b + ac^2)(2c)$

$$\nabla \mathcal{L} = \begin{bmatrix} \frac{\partial \mathcal{L}}{\partial a} \\ \frac{\partial \mathcal{L}}{\partial b} \\ \frac{\partial \mathcal{L}}{\partial c} \end{bmatrix}$$

$$= \begin{bmatrix} (a + 2b + ac^2)(1 + 2\mathrm{c}) \\ (a + 2b + ac^2)(2) \\ (a + 2b + ac^2)(2c) \end{bmatrix}$$

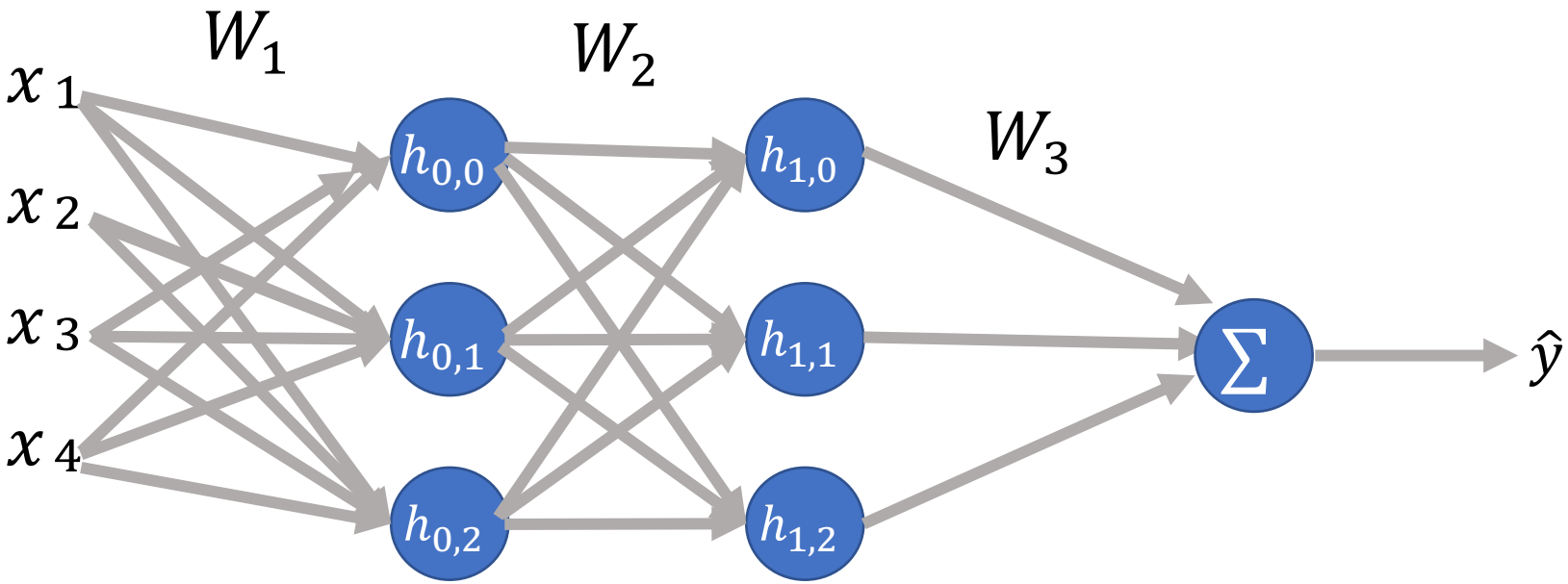
Feed forward NN



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

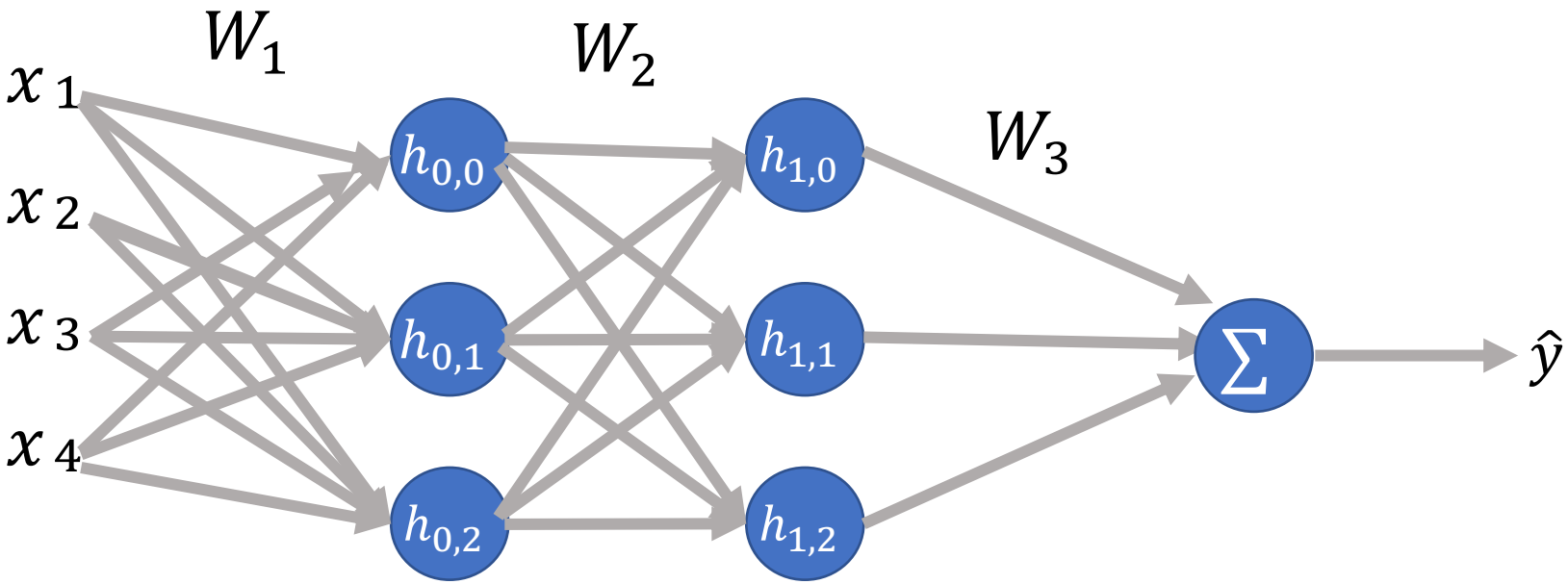
$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

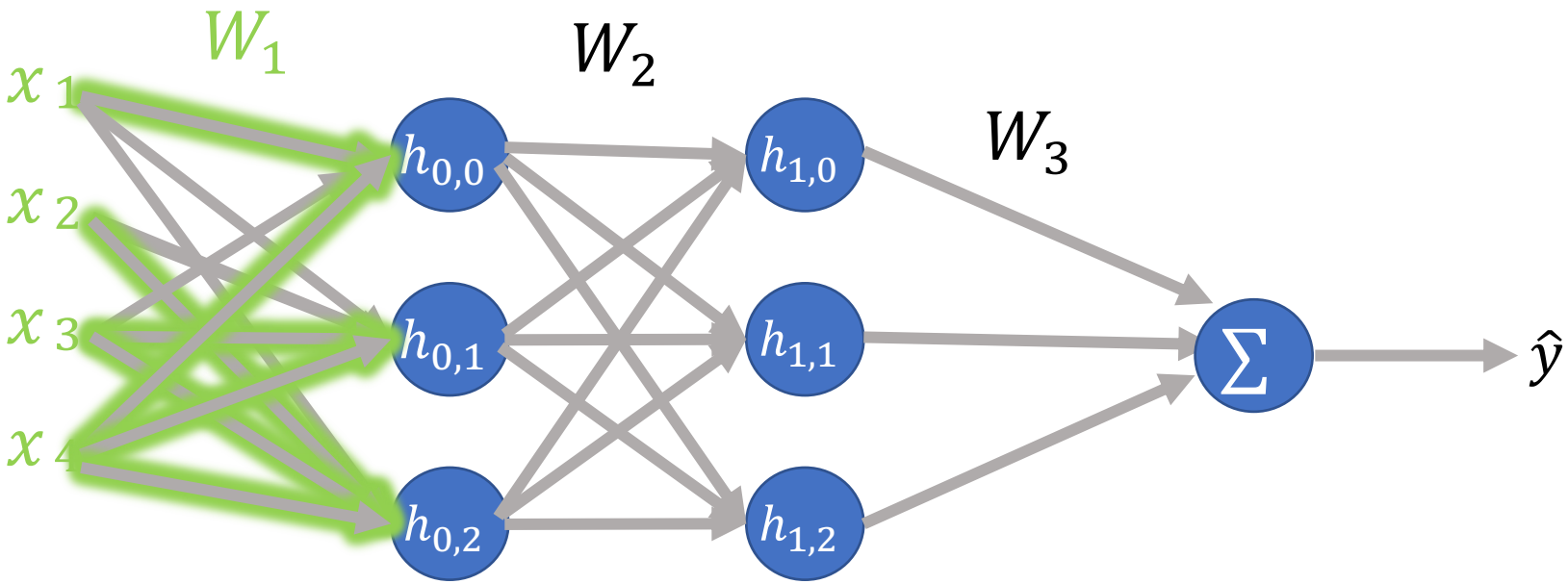
$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{xW}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

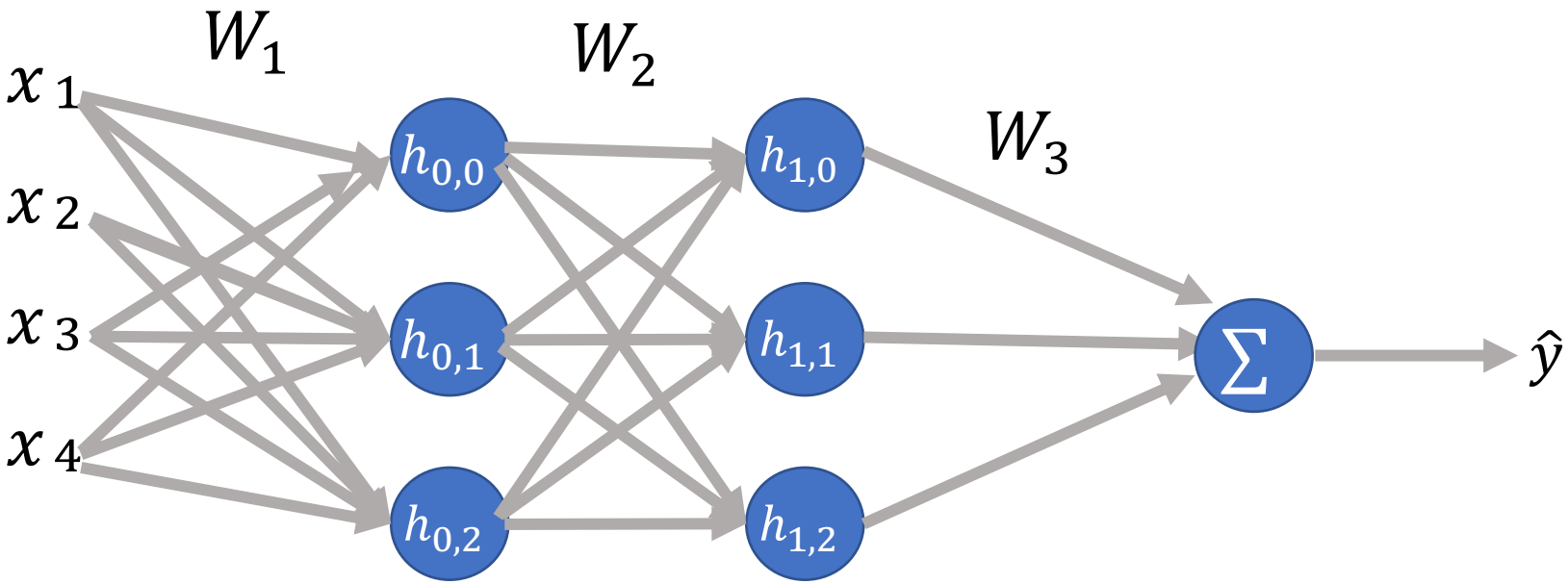
$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{xW}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

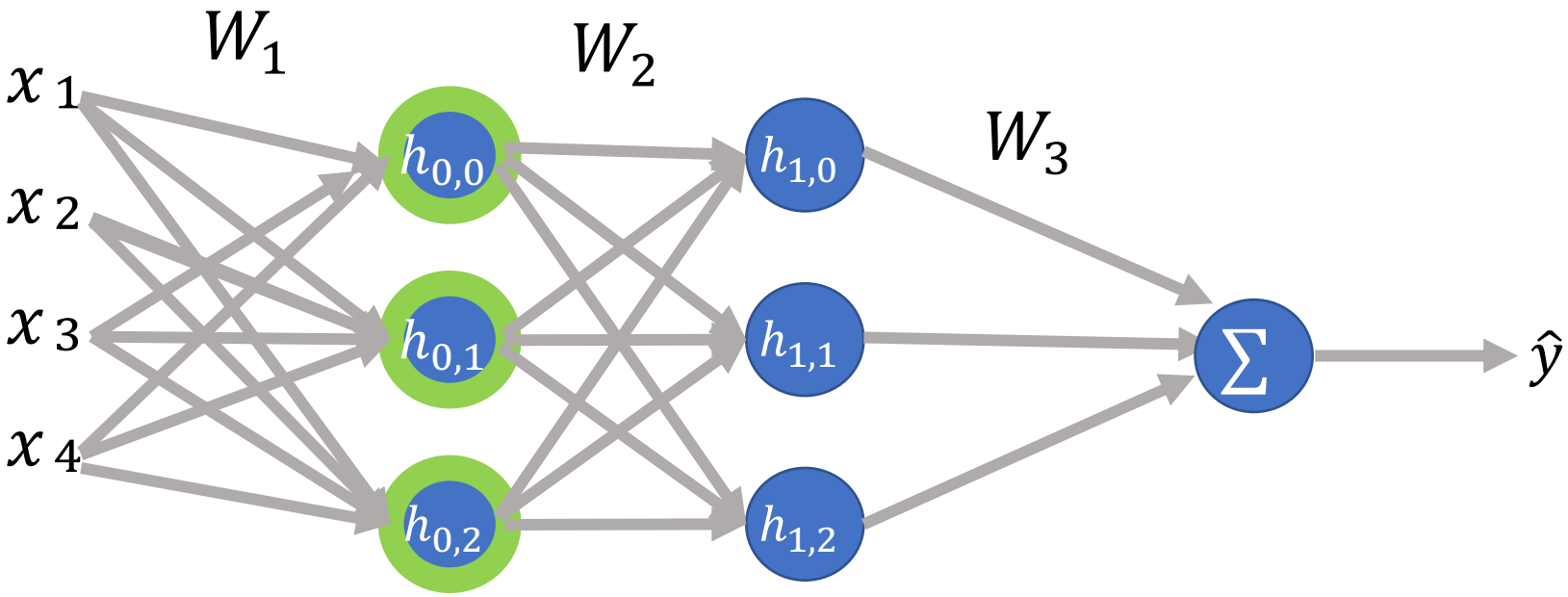
$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

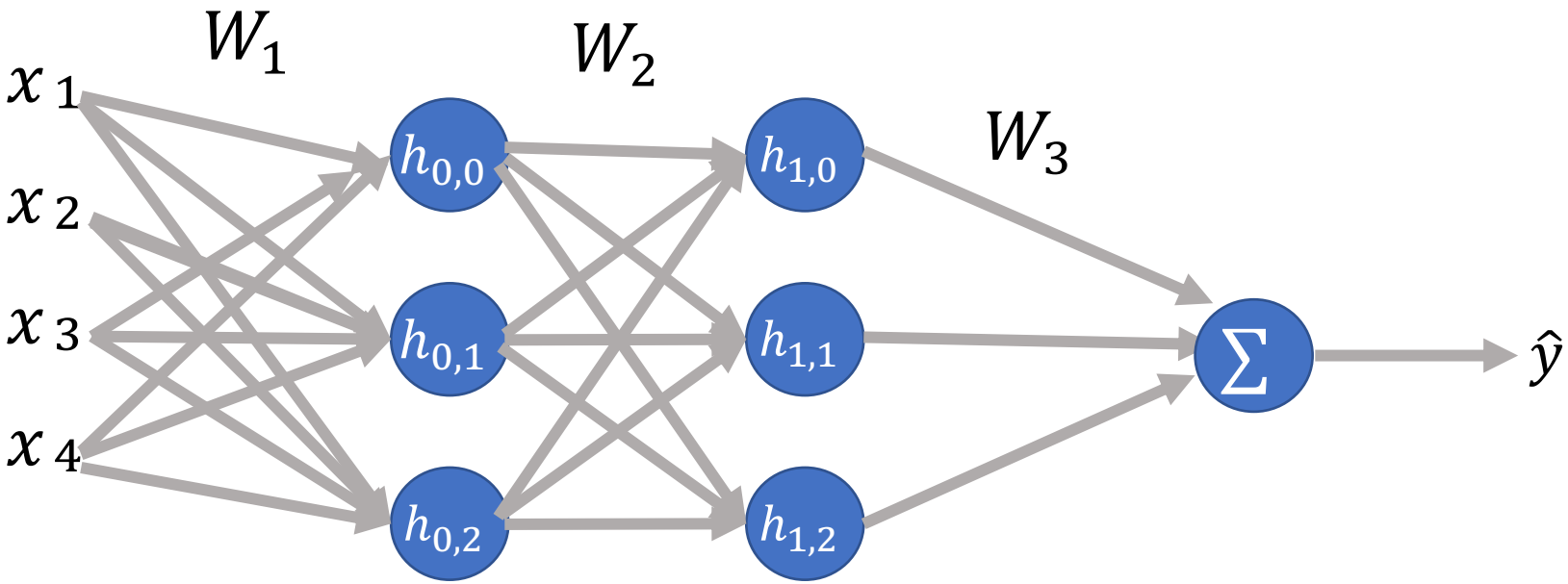
$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

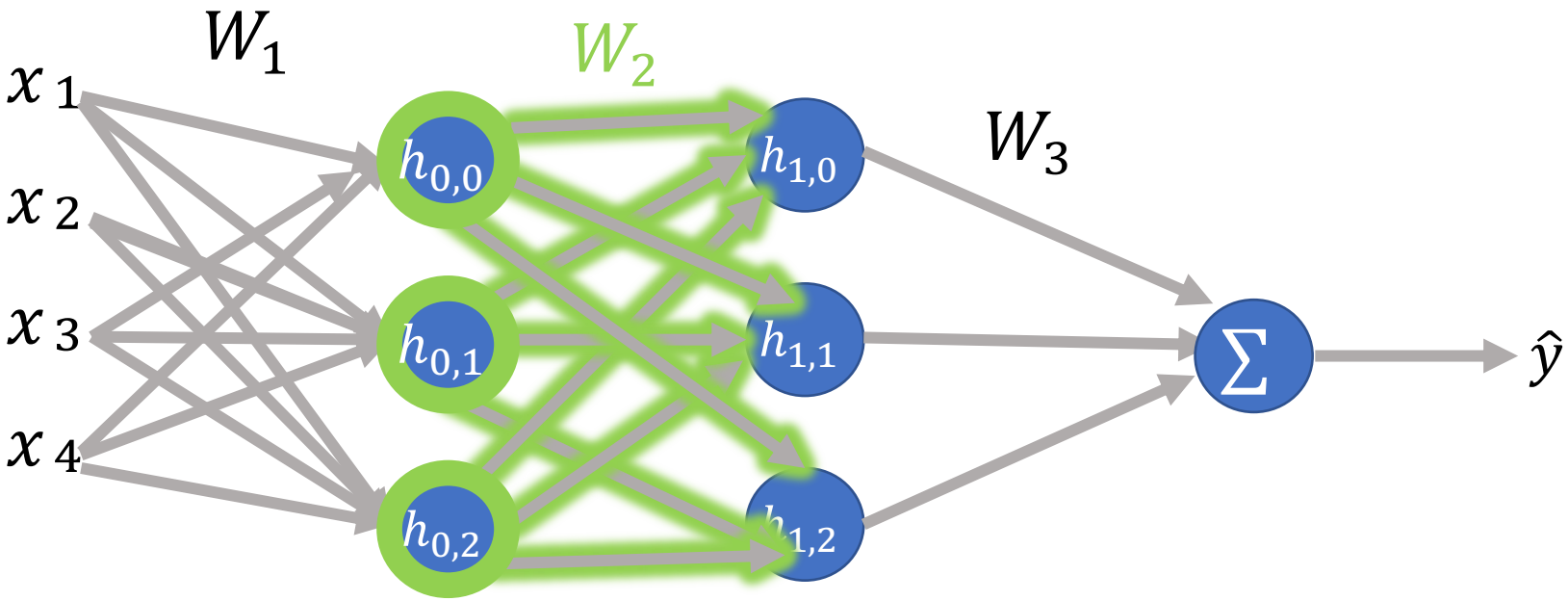
$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

$$L(\hat{y}, y) =$$

$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b})) + \boldsymbol{b})), \boldsymbol{y})$$



$$\boldsymbol{h}_0 = \sigma(xW_1)$$

Loss function:

$$L(\sigma(\boldsymbol{W}_3\sigma(\boldsymbol{W}_2\,\sigma(\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}))+\boldsymbol{b})),\boldsymbol{y})$$

How many parameters do we have?

W_1 is 4 x 3

W_2 is 3 x 3

W_3 is 3 x 1

3 bias terms

12 + 9 + 3 + 3 = 27 parameters

$$\nabla\mathcal{L}=\begin{bmatrix}\frac{\partial\mathcal{L}}{\partial\theta_1}\\ \frac{\partial\mathcal{L}}{\partial\theta_2}\\ \cdots\\ \cdots\\ \frac{\partial\mathcal{L}}{\partial\theta_{27}}\\ \frac{\partial\mathcal{L}}{\partial b_1}\end{bmatrix}$$

Outline

Updating weights in LR

Computation Graph

Backpropagation

2 layered Feed-Forward Neural Network

Issues when training NNs

Pytorch

Deep Averaging Neural Network.

Computation graph

$$L(a, b, c) = b(a + 2b + ac^2)$$

Computations:

$$d = c^2$$
$$\mathrm{e} = \mathrm{a} * \mathrm{d}$$
$$\mathrm{f} = 2 * \mathrm{b}$$
$$g = \mathrm{a} + \mathrm{f}$$
$$\mathrm{h} = \mathrm{g} + \mathrm{e}$$
$$\mathrm{i} = \mathrm{b} * \mathrm{h}$$

Computation graph

$L(a, b, c) = b(a + 2b + ac^2)$

Computations:

$d = c^2$

$e = a * d$

$f = 2 * b$

$g = a + f$

$h = g + e$

$i = b * h$

Computation graph

$L(a, b, c) = b(a + 2b + ac^2)$

Computations:

$d = c^2$

$\mathrm{e} = \mathrm{a} * \mathrm{d}$

$\mathrm{f} = 2 * \mathrm{b}$

$g = \mathrm{a} + \mathrm{f}$

$\mathrm{h} = \mathrm{g} + \mathrm{e}$

$\mathrm{i} = \mathrm{b} * \mathrm{h}$

$\mathrm{i} = h * b$

$d = c^2$

Computation graph

$L(a, b, c) = b(a + 2b + ac^2)$

Computations:

$d = c^2$

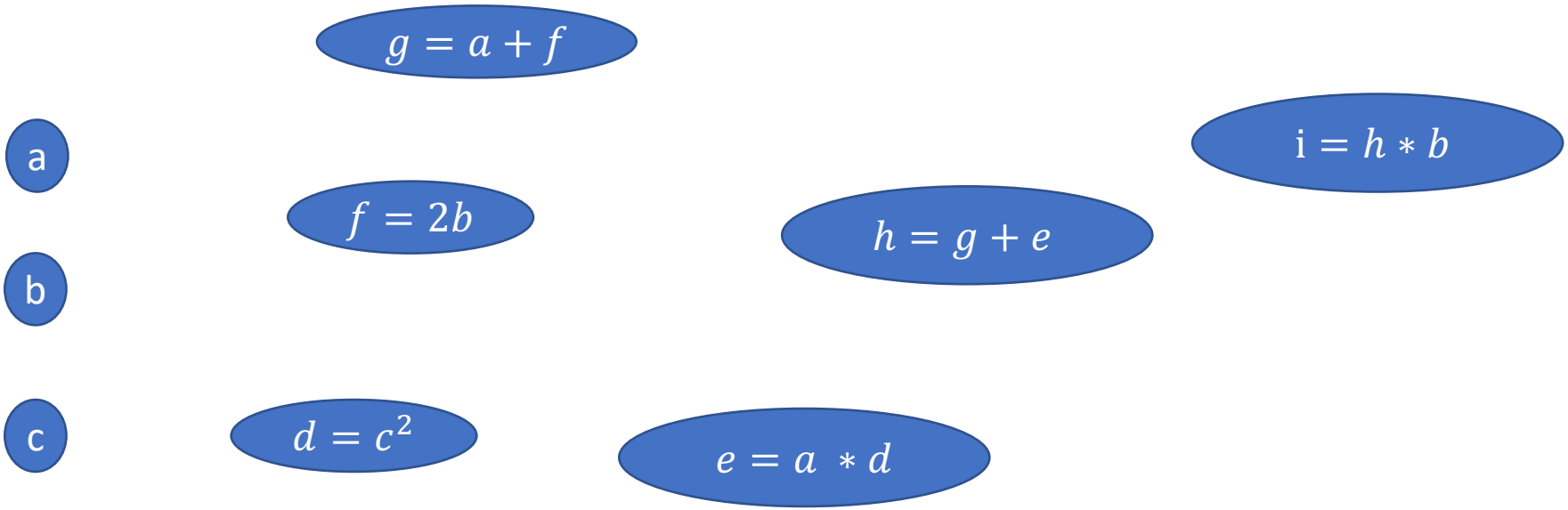
$e = a * d$

$f = 2 * b$

$g = a + f$

$h = g + e$

$i = b * h$



Computation graph

$L(a, b, c) = b(a + 2b + ac^2)$

Computations:

$d = c^2$

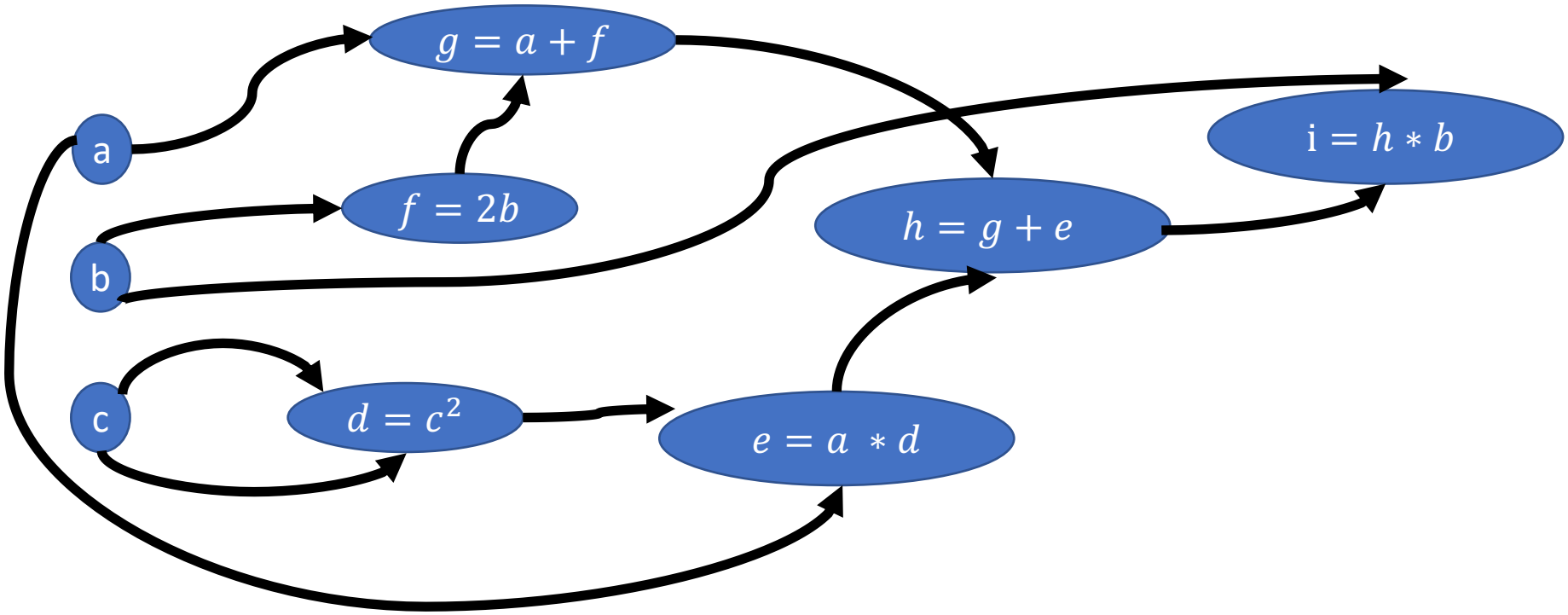
e = a ∗ d

f = 2 ∗ b

g = a + f

h = g + e

i = b ∗ h



Computation graph

$L(a, b, c) = b(a + 2b + ac^2)$

$L(3, 1, -2) = b(a + 2b + ac^2)$

Computations:

$d = c^2$

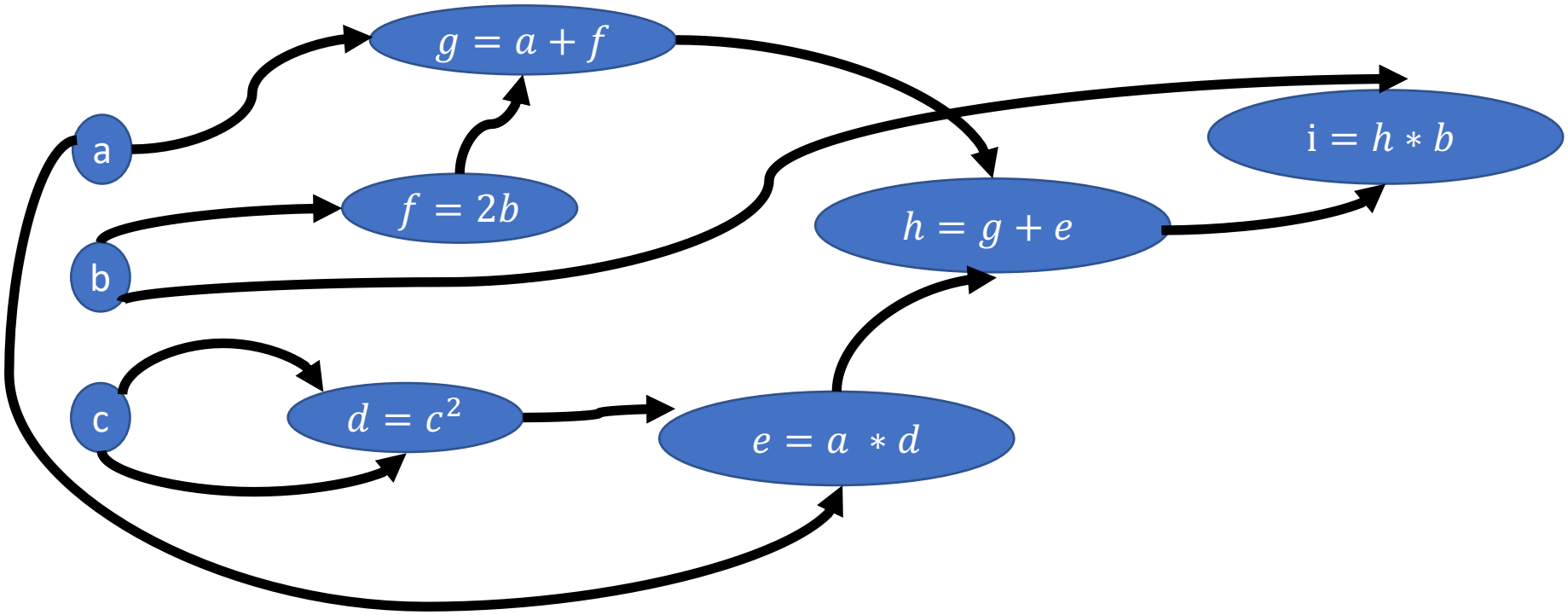
$e = a * d$

$f = 2 * b$

$g = a + f$

$h = g + e$

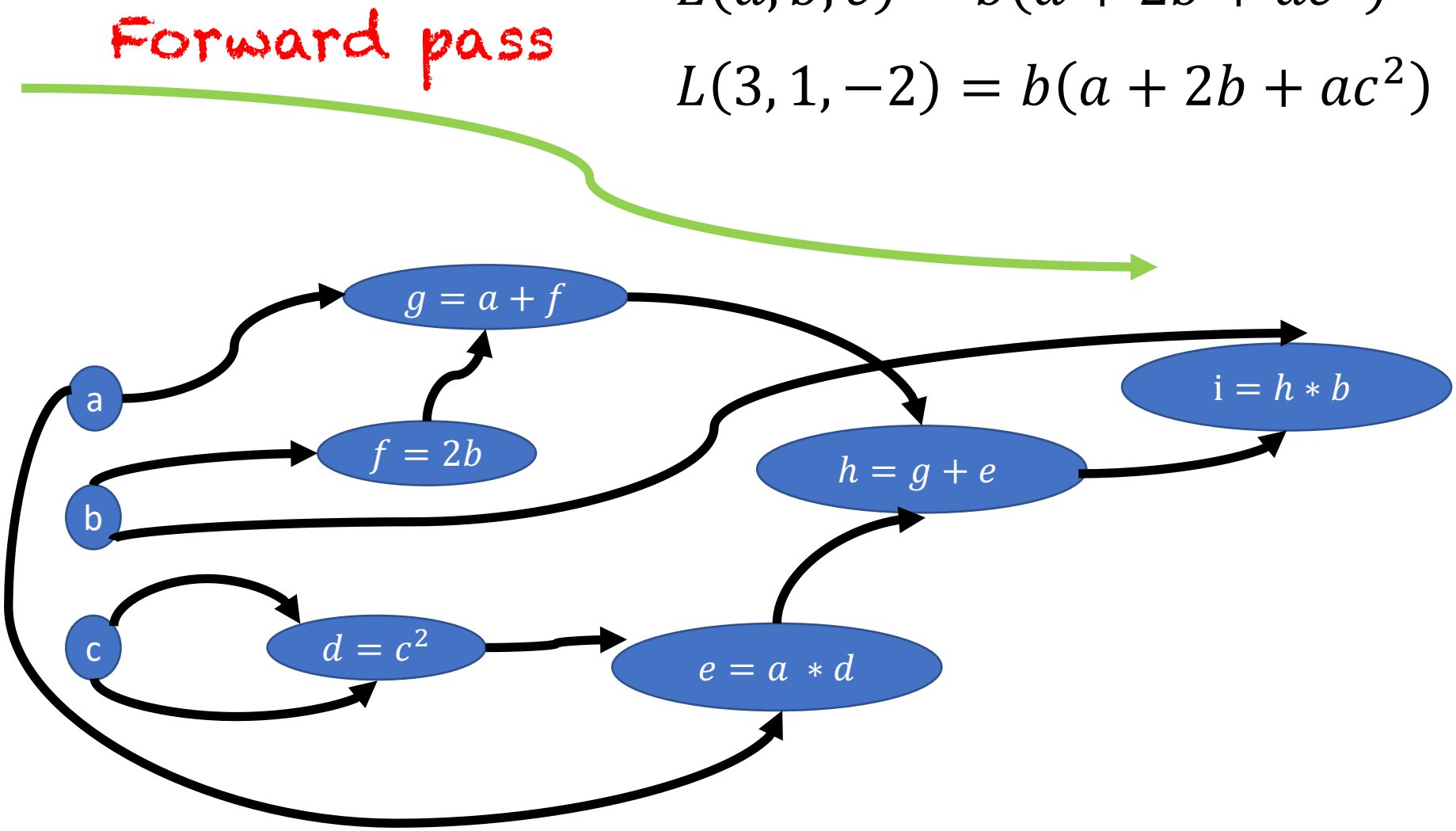
$i = h * b$



Computation graph

$$L(a,b,c) = b(a + 2b + ac^2)$$

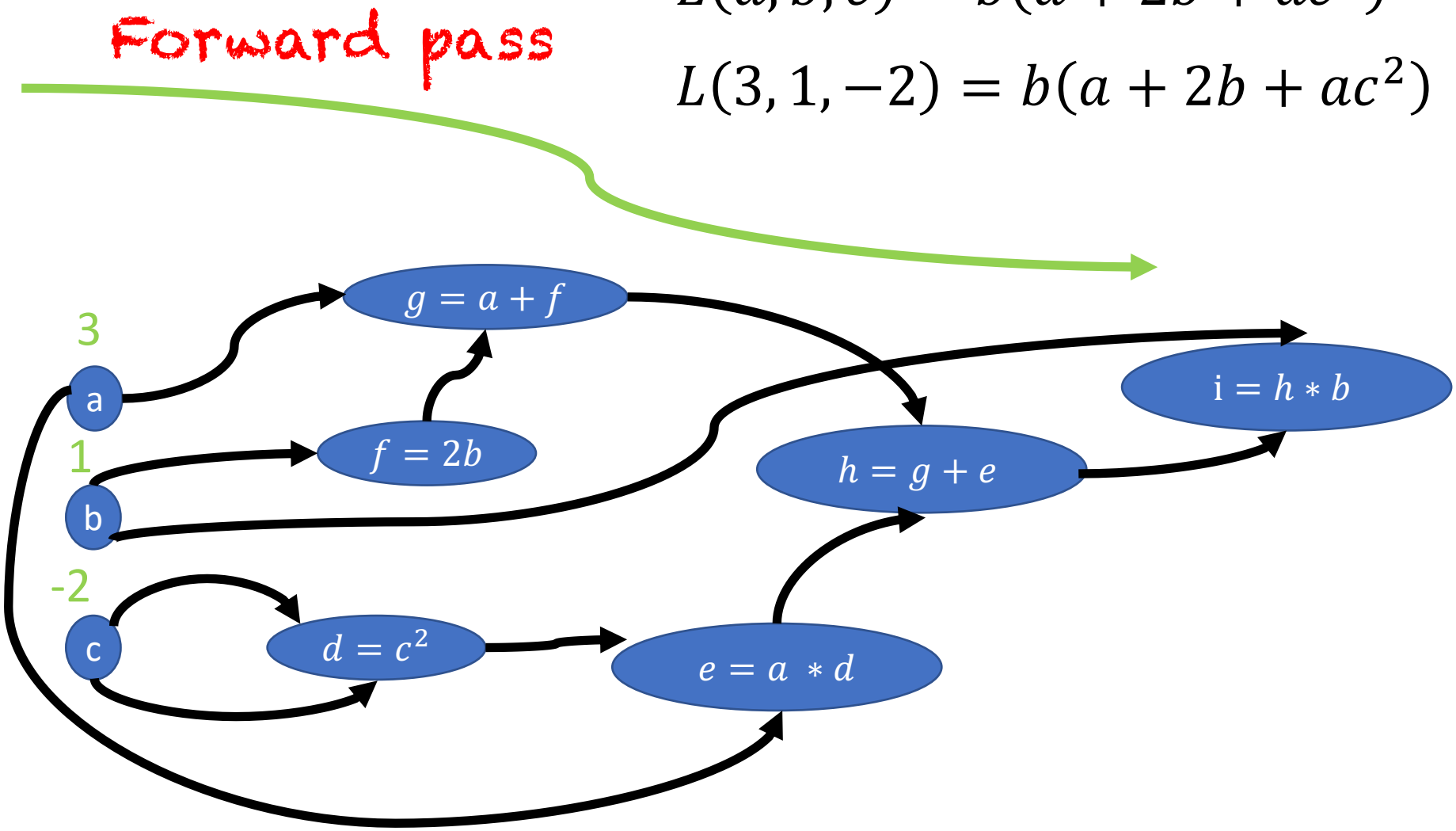
$$L(3,1,-2) = b(a + 2b + ac^2)$$



Computation graph

$$L(a,b,c) = b(a + 2b + ac^2)$$

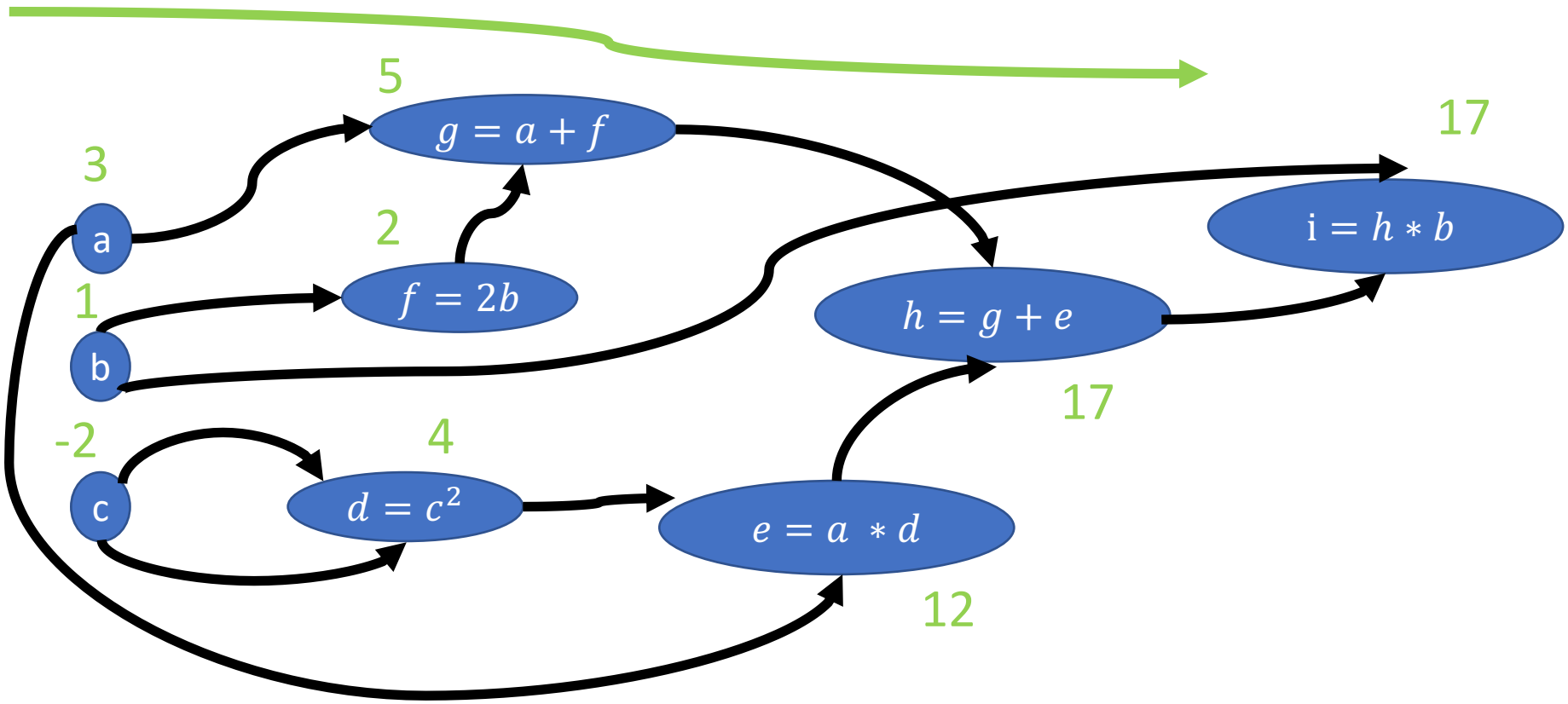
$$L(3,1,-2) = b(a + 2b + ac^2)$$



Computation graph $L(a,b,c) = b(a + 2b + ac^2)$

$$L(3,1,-2) = b(a + 2b + ac^2)$$

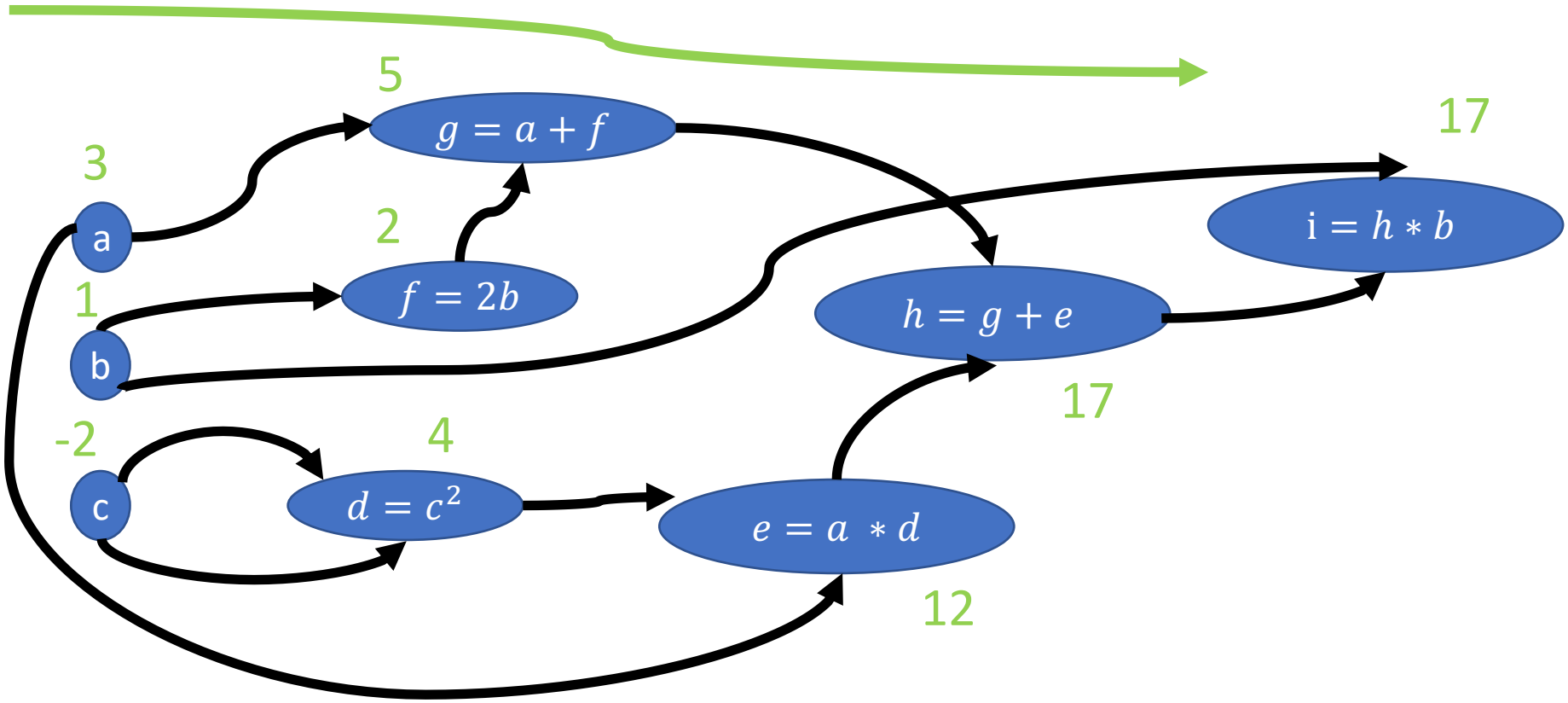
Forward pass



Computation graph $L(a,b,c) = b(a + 2b + ac^2)$

Forward pass

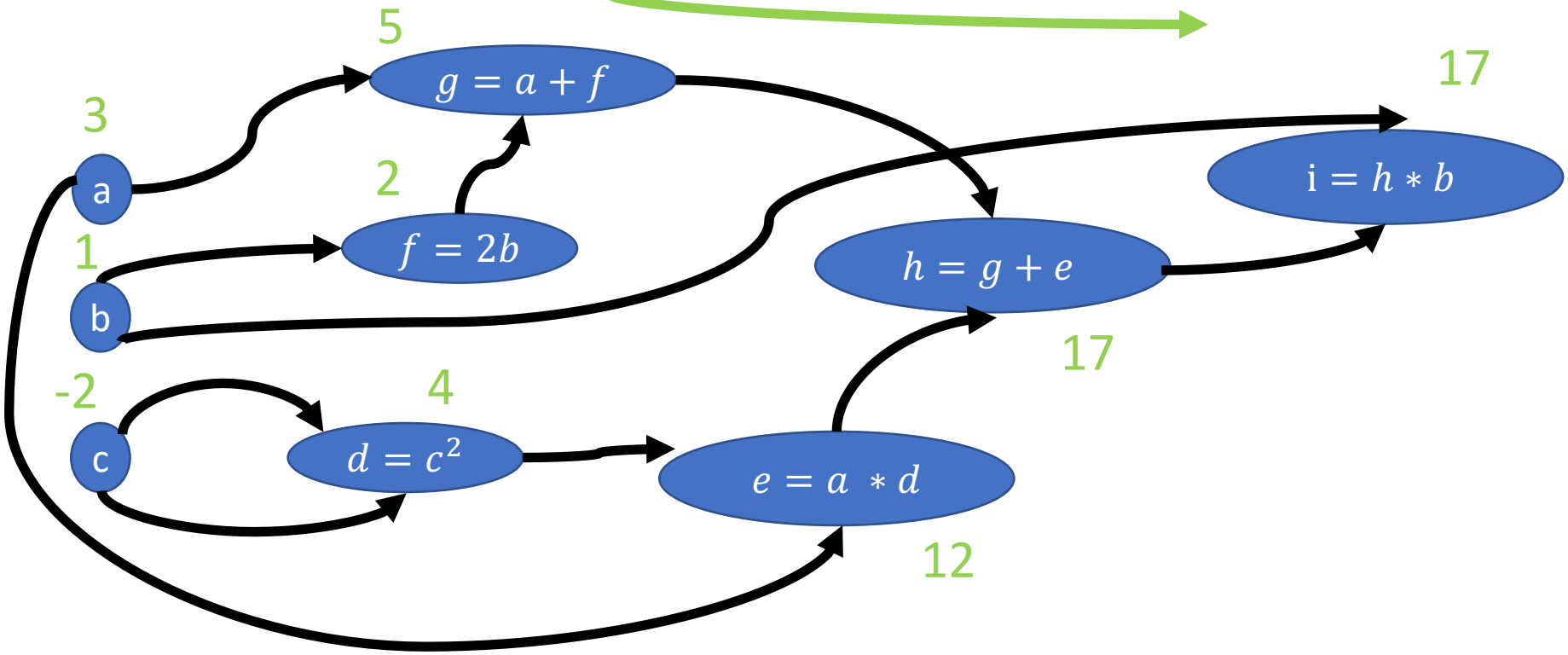
$$L(3,1,-2) = b(a + 2b + ac^2)$$
$$L(3,1,-2) = 1(3 + 2(1) + 3(-2)^2)$$



Computation graph $L(a,b,c) = b(a + 2b + ac^2)$

Forward pass

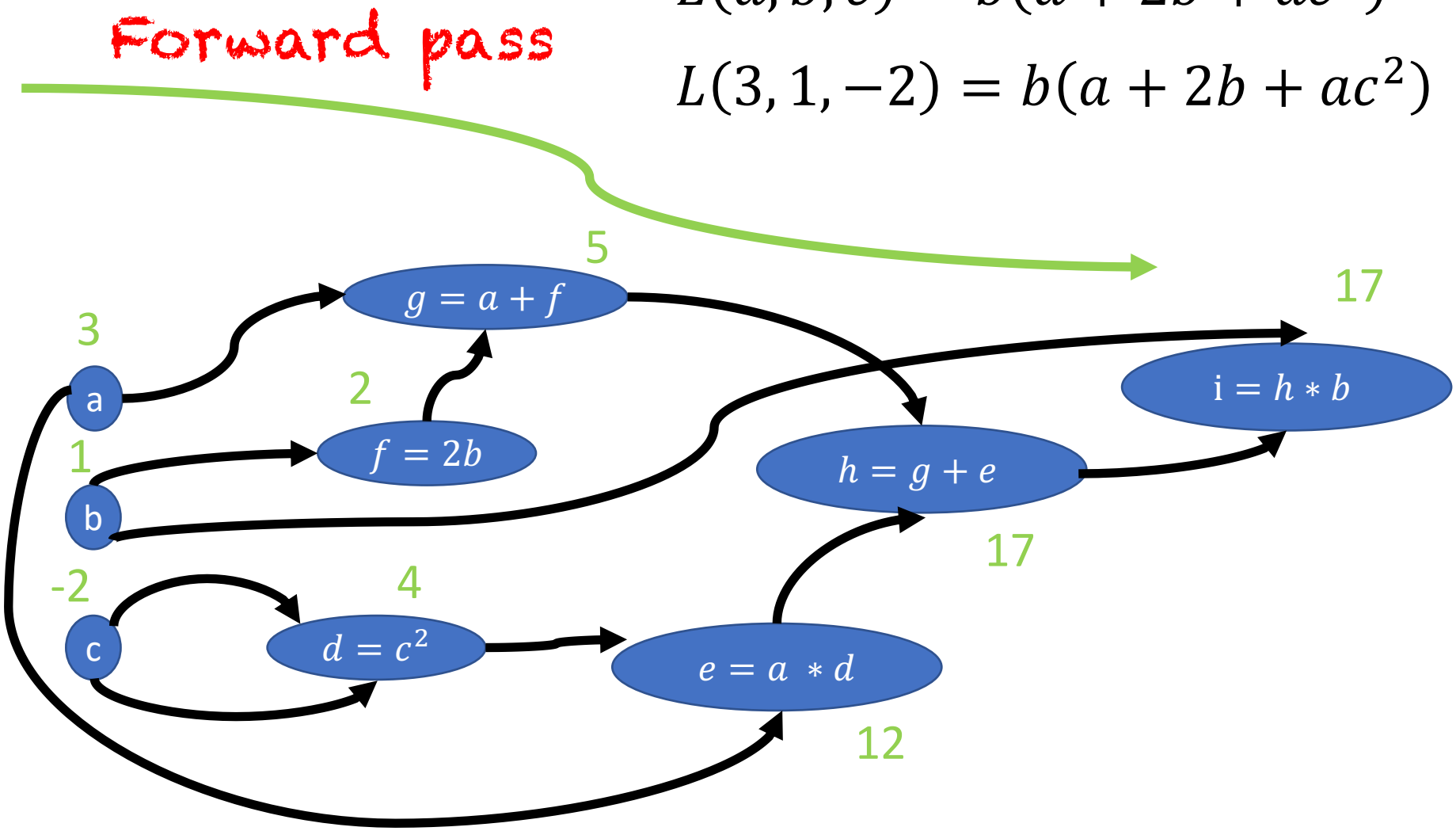
$$L(3,1,-2) = b(a + 2b + ac^2)$$
$$L(3,1,-2) = 1(3 + 2(1) + 3(-2)^2)$$
$$L(3,1,-2) = 1(3 + 2 + 12) = 17$$



Computation graph

$$L(a,b,c) = b(a + 2b + ac^2)$$

$$L(3,1,-2) = b(a + 2b + ac^2)$$



One node view

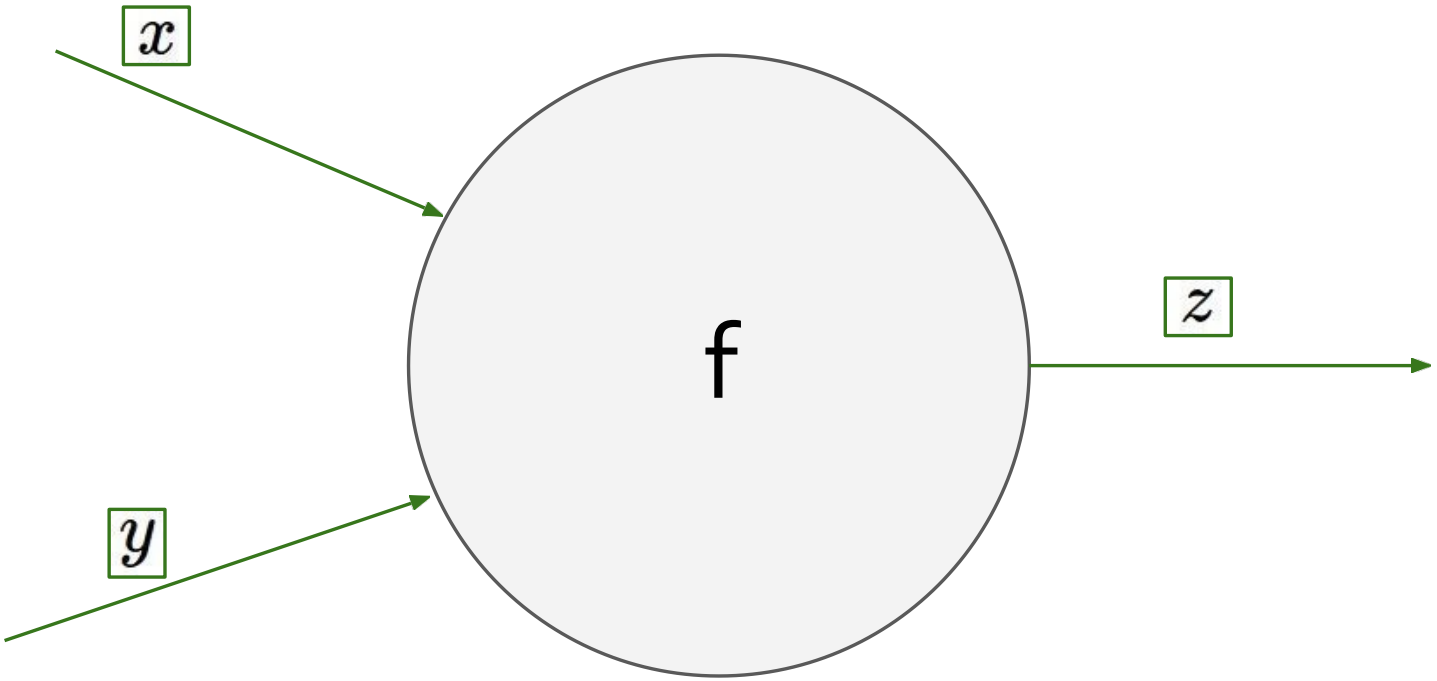


Figure from Andrej Karpathy

One node view

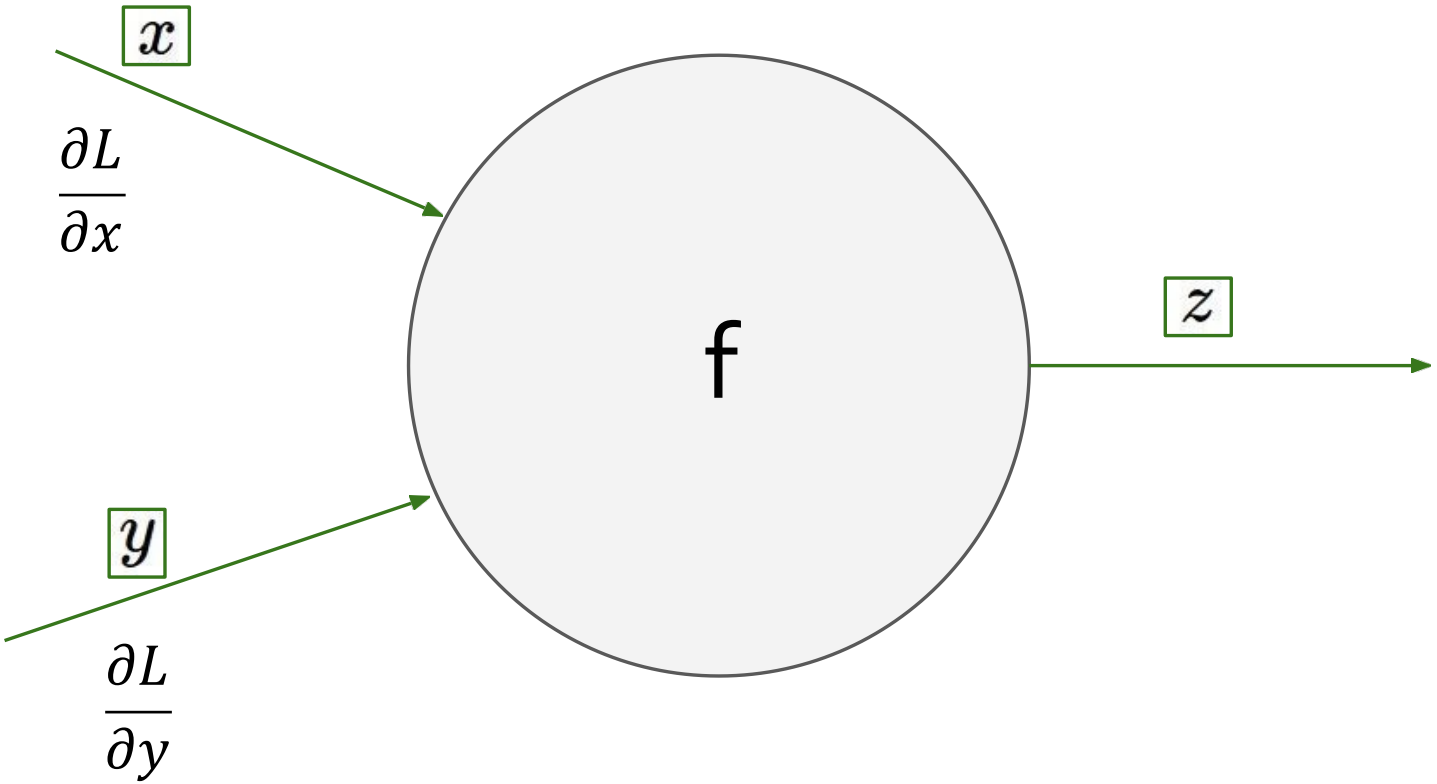


Figure from Andrej Karpathy

One node view

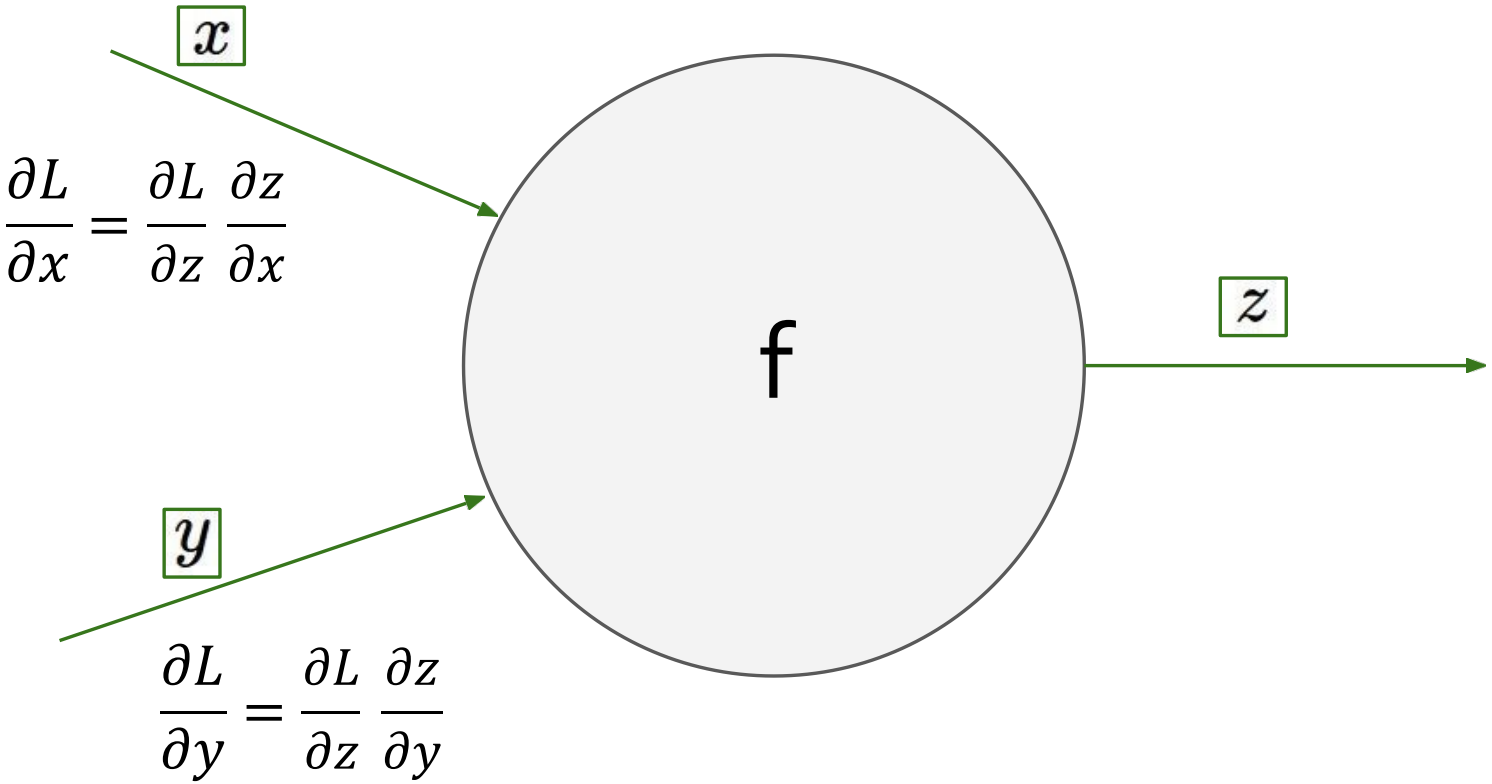


Figure from Andrej Karpathy

One node view

Local gradient

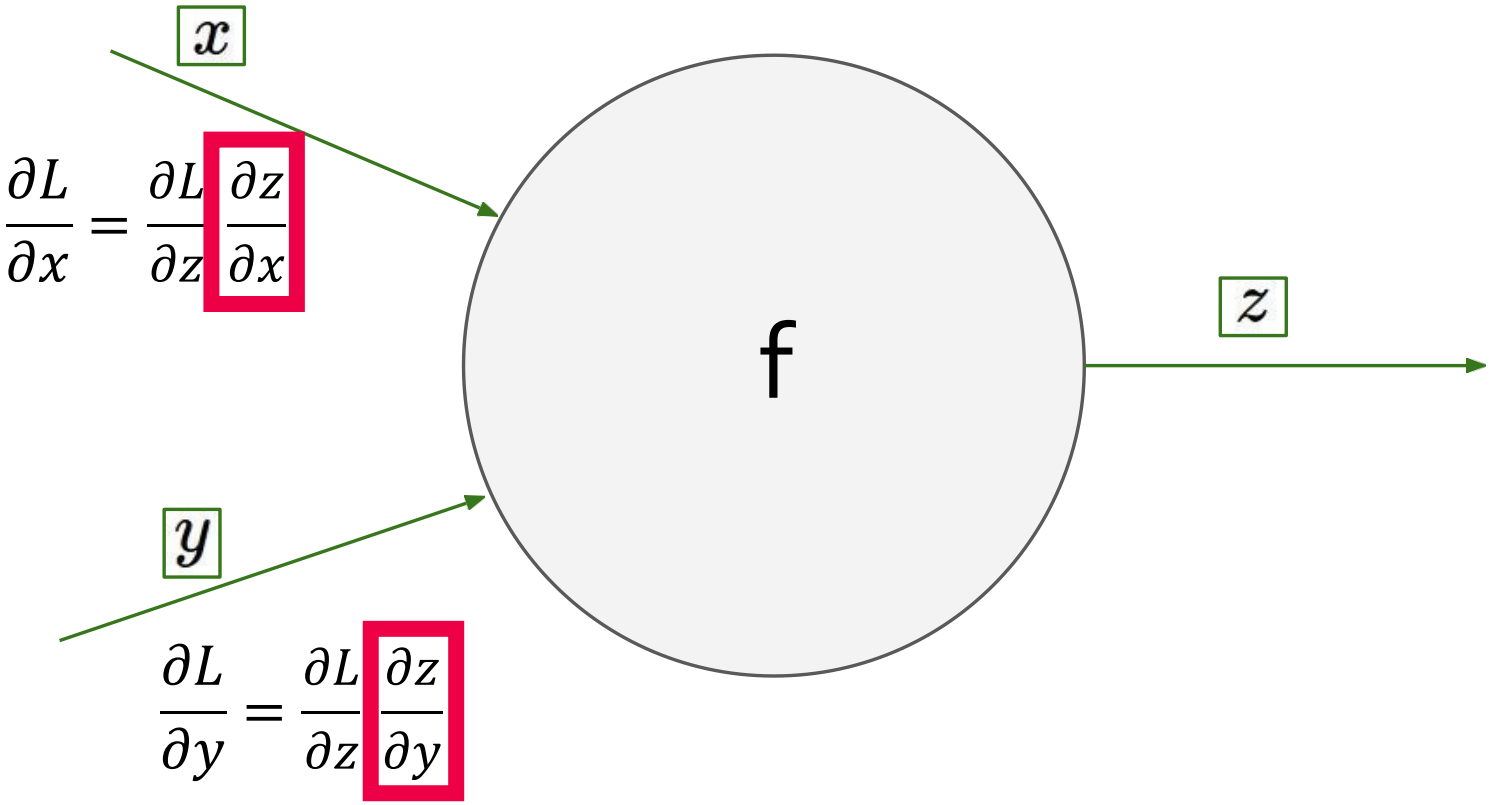


Figure from Andrej Karpathy

One node view

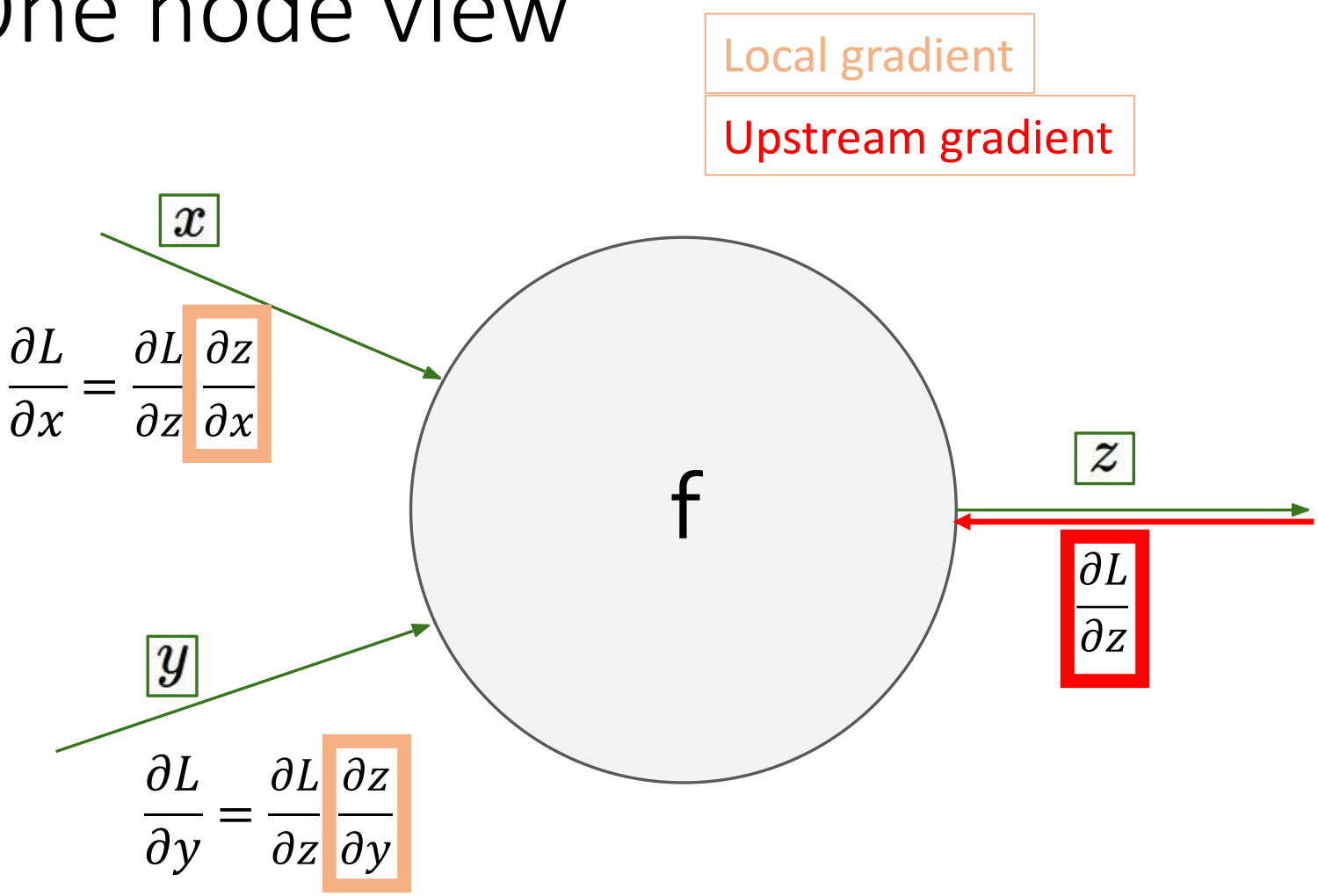


Figure from Andrej Karpathy

One node view

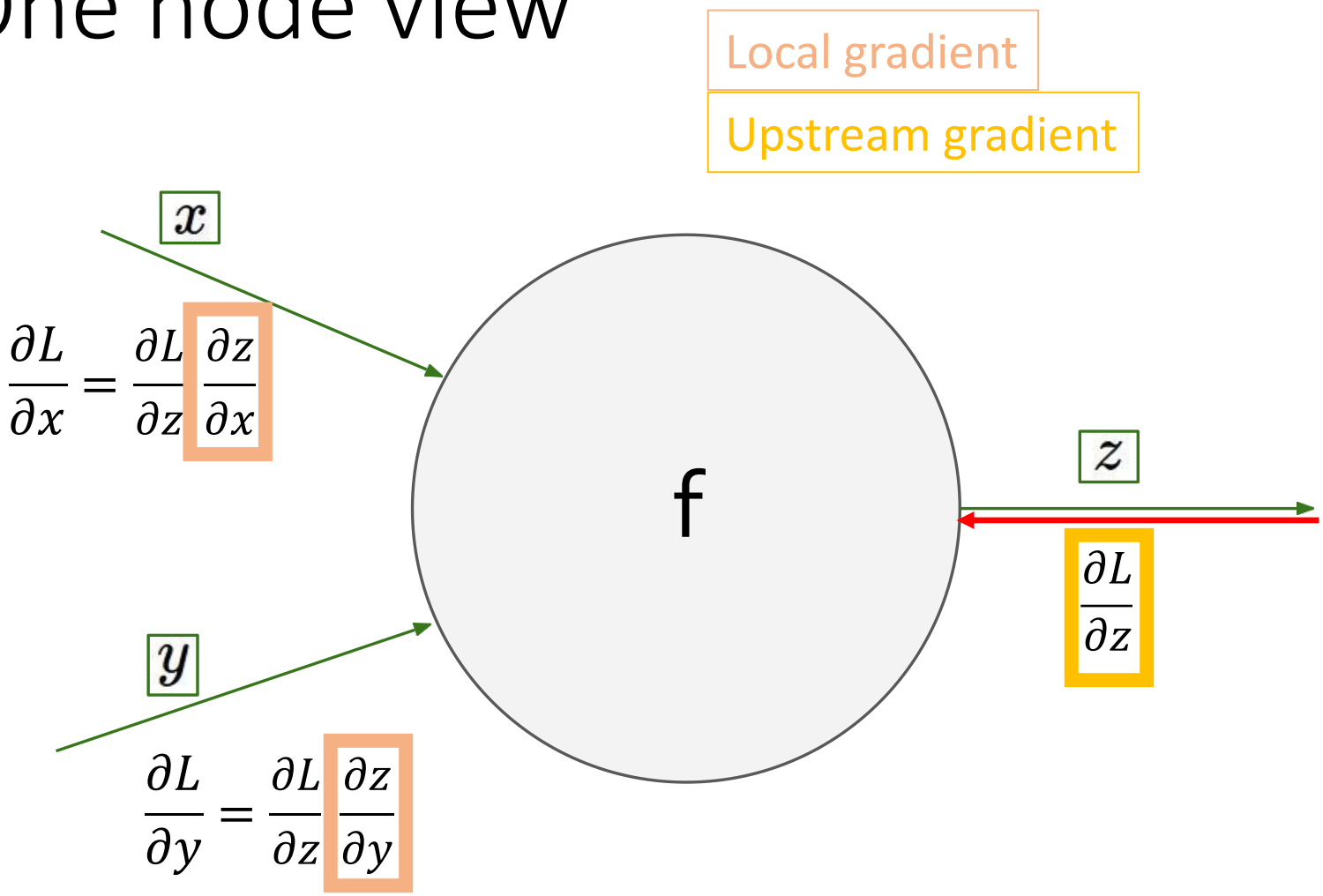


Figure from Andrej Karpathy

One node view

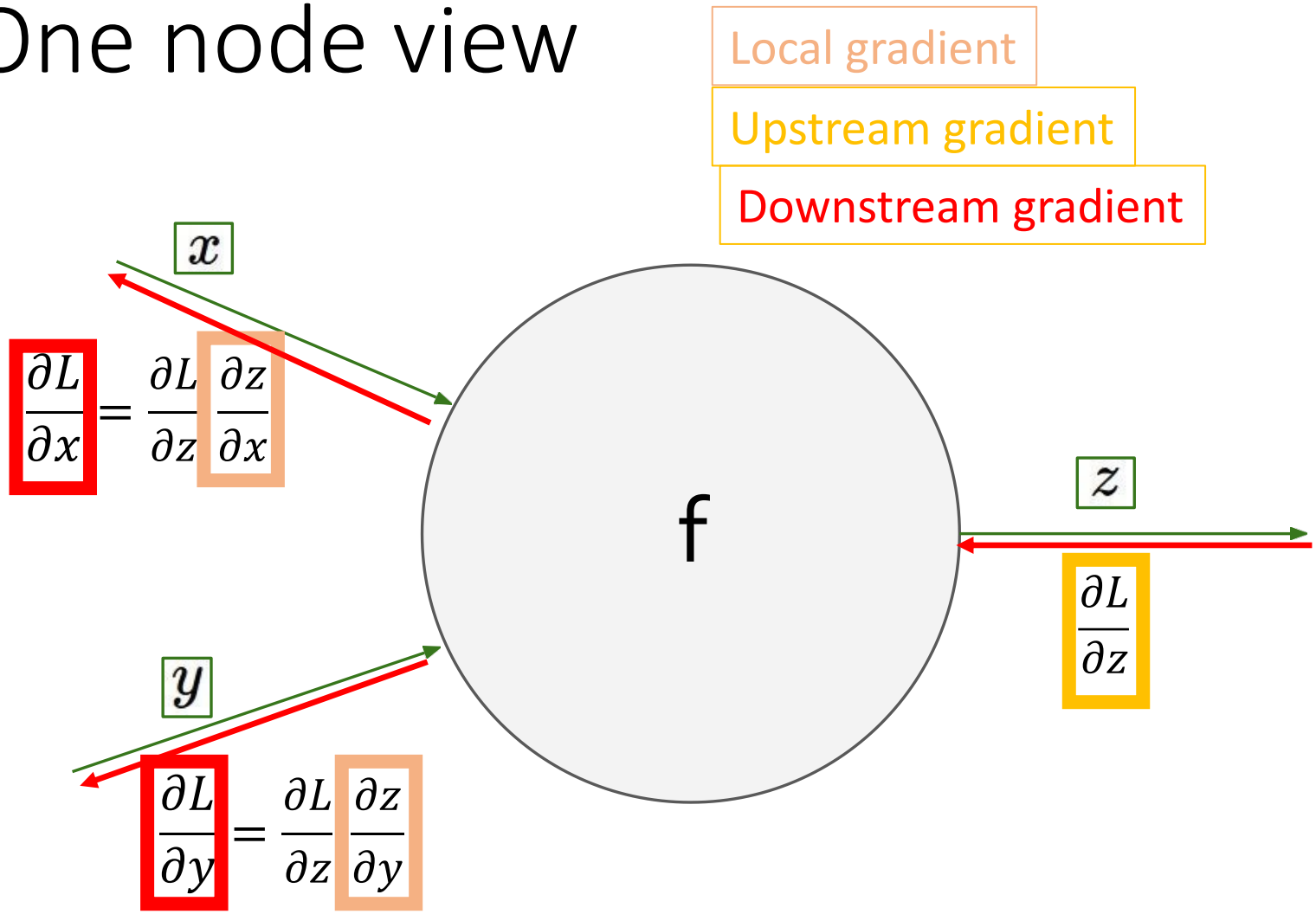


Figure from Andrej Karpathy

One node view

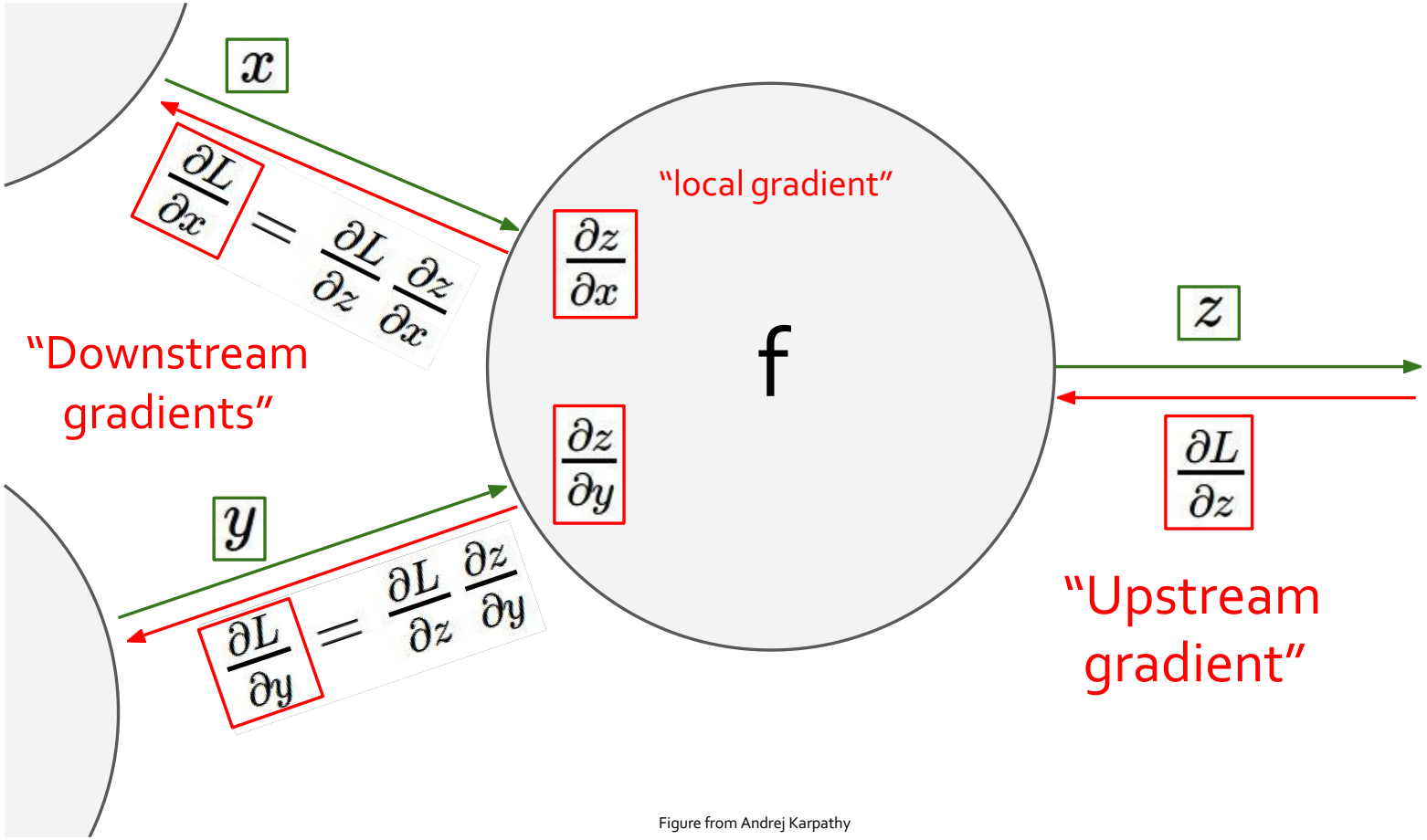
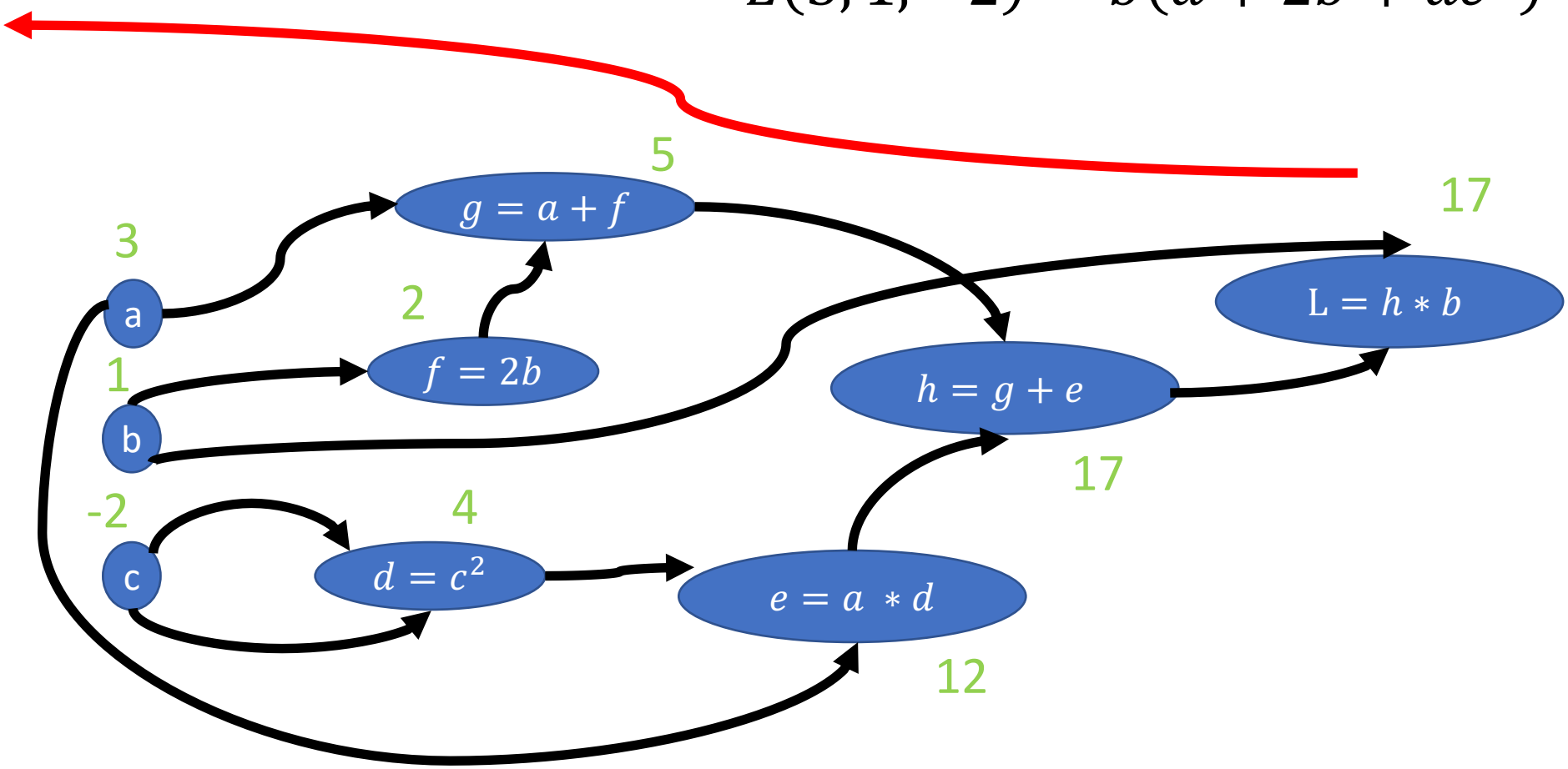


Figure from Andrej Karpathy

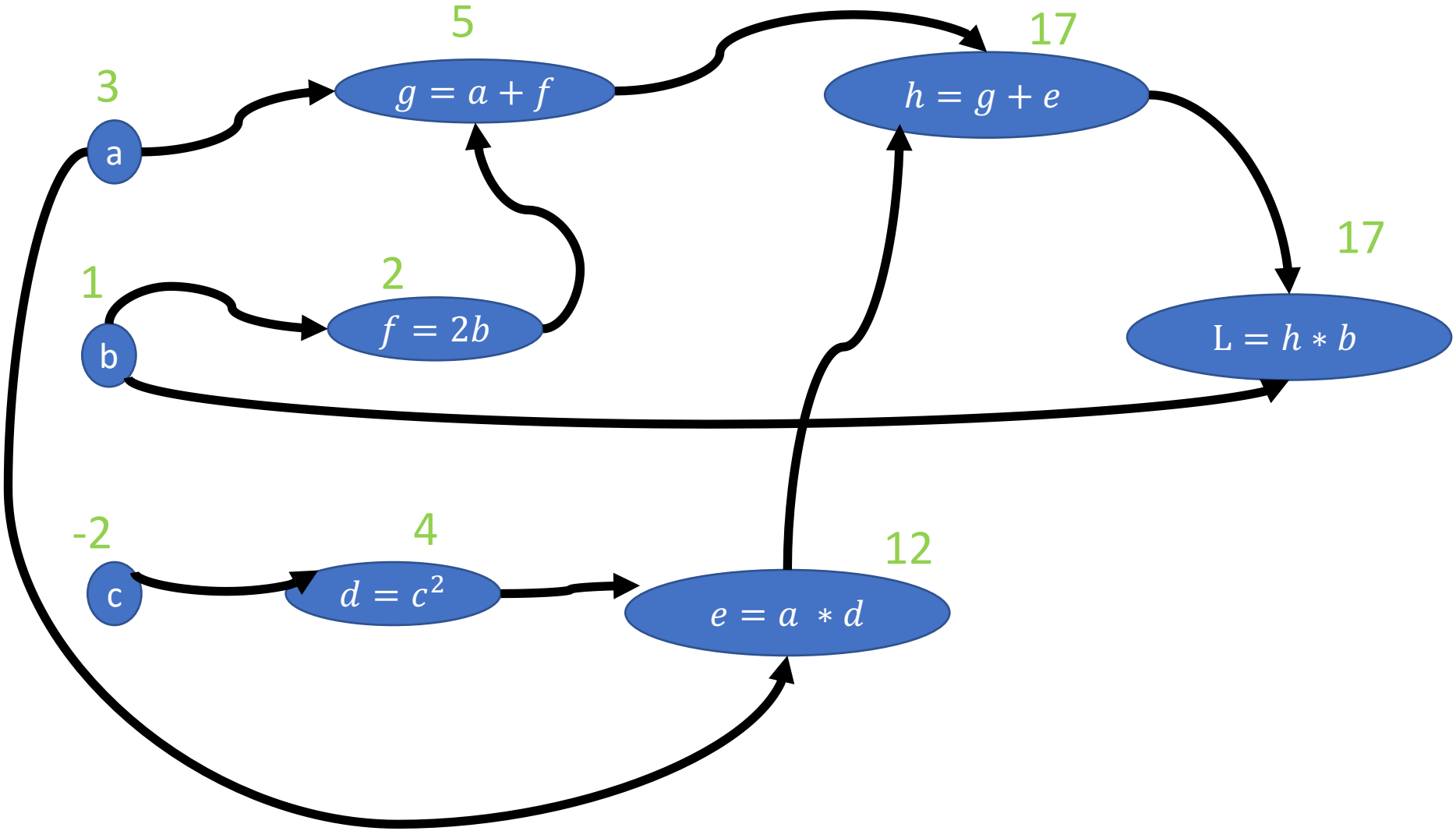
Computation graph

$$L(a,b,c) = b(a + 2b + ac^2)$$

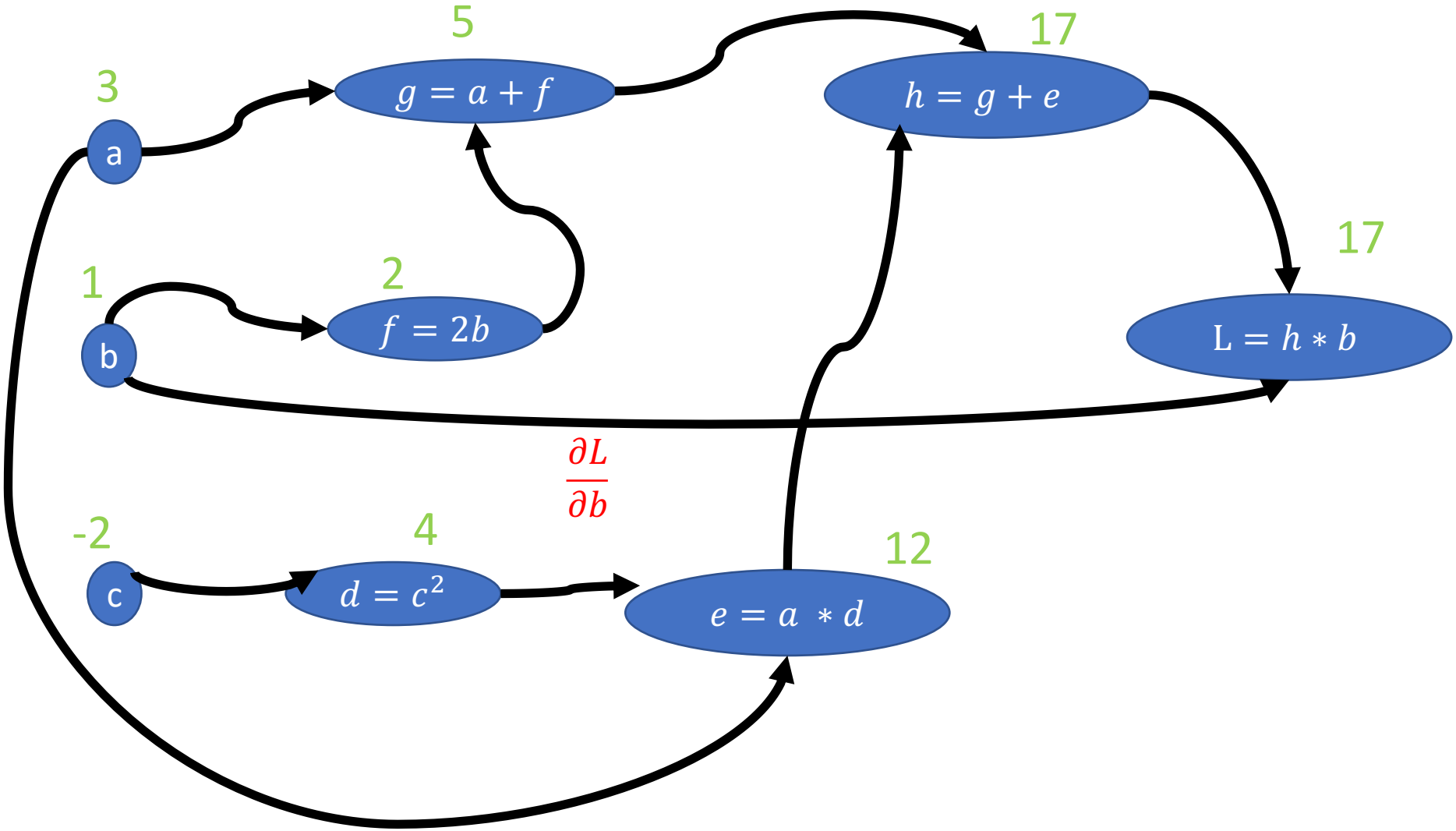
$$L(3,1,-2) = b(a + 2b + ac^2)$$



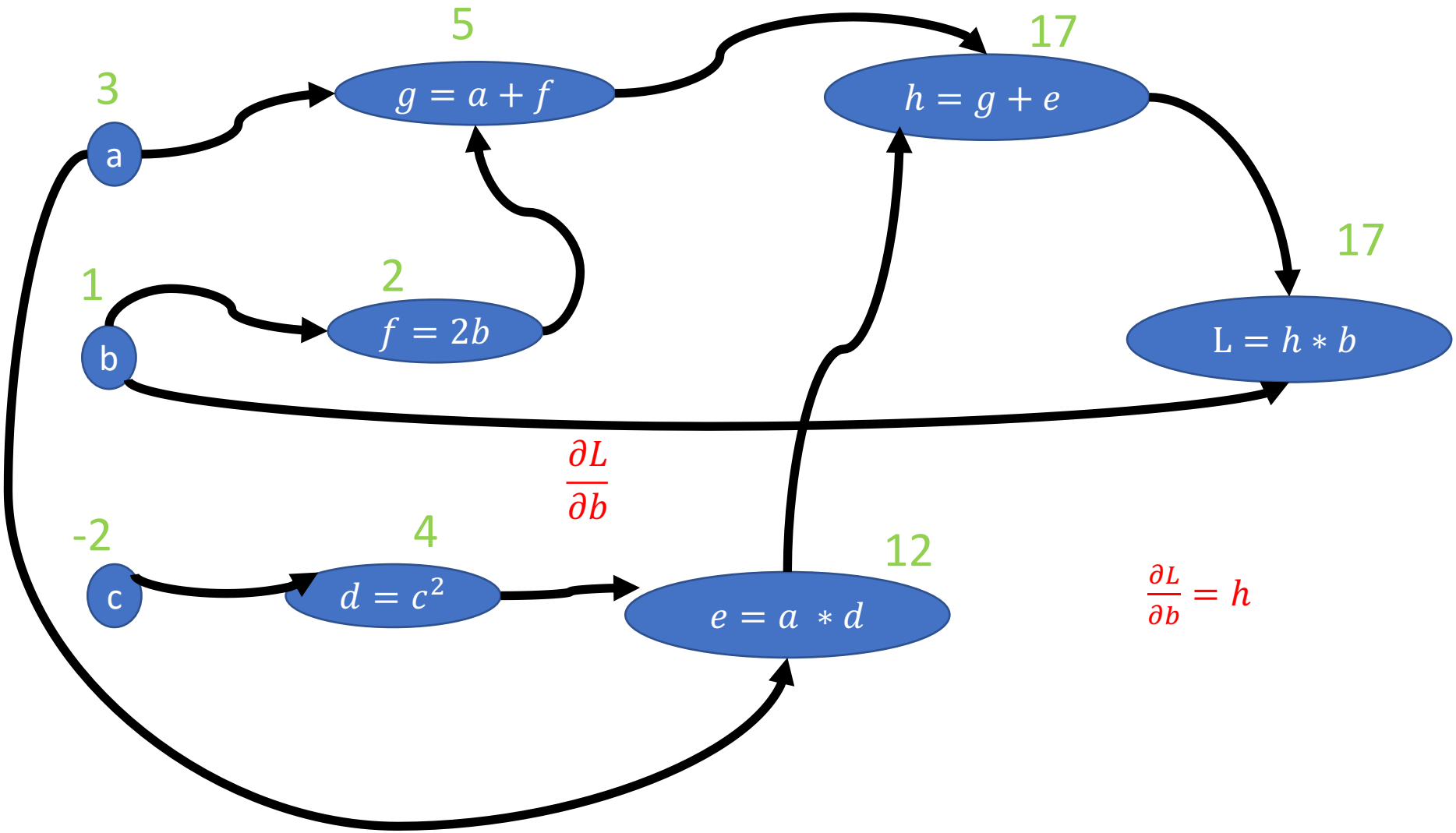
backward pass



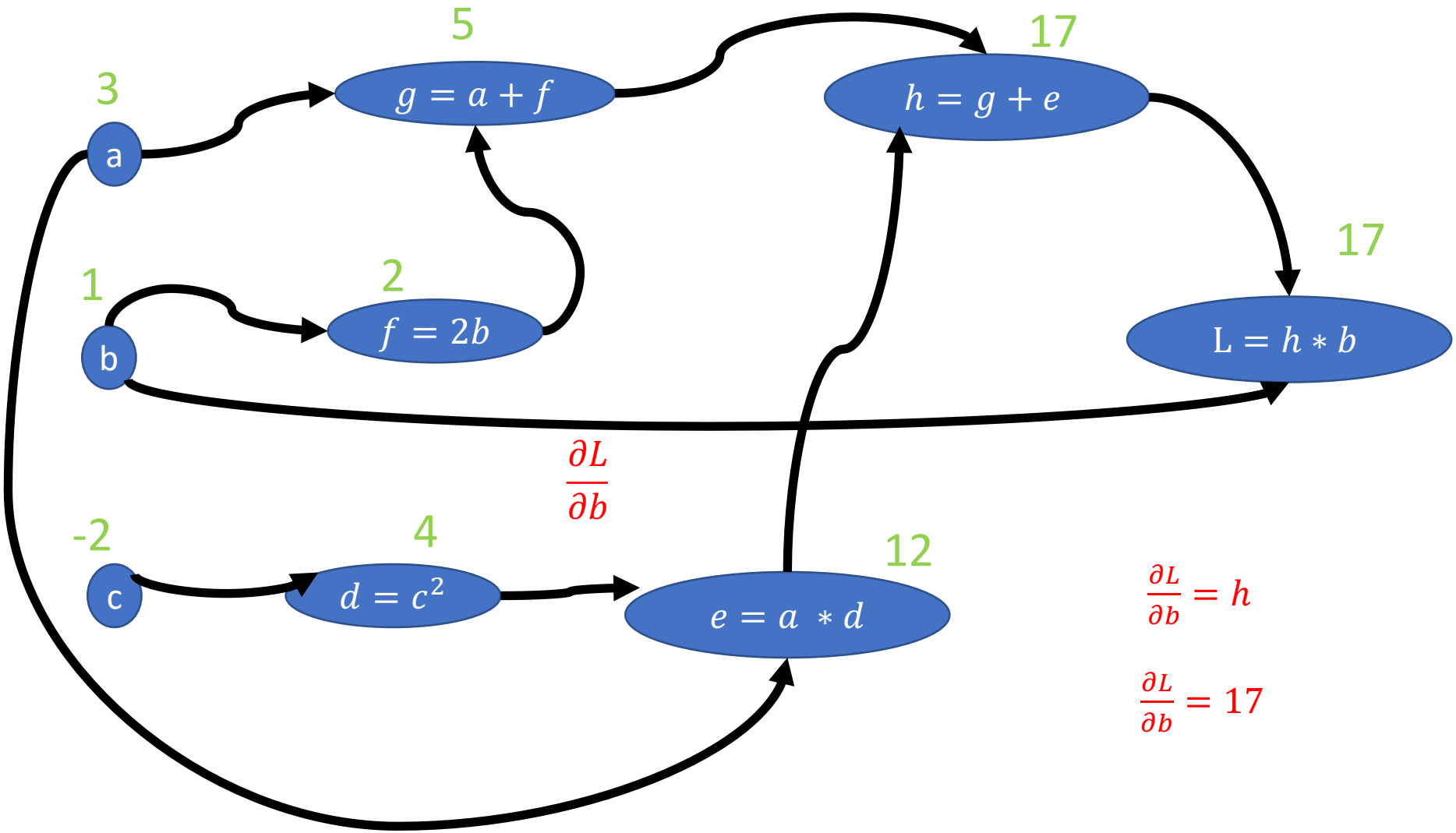
backward pass



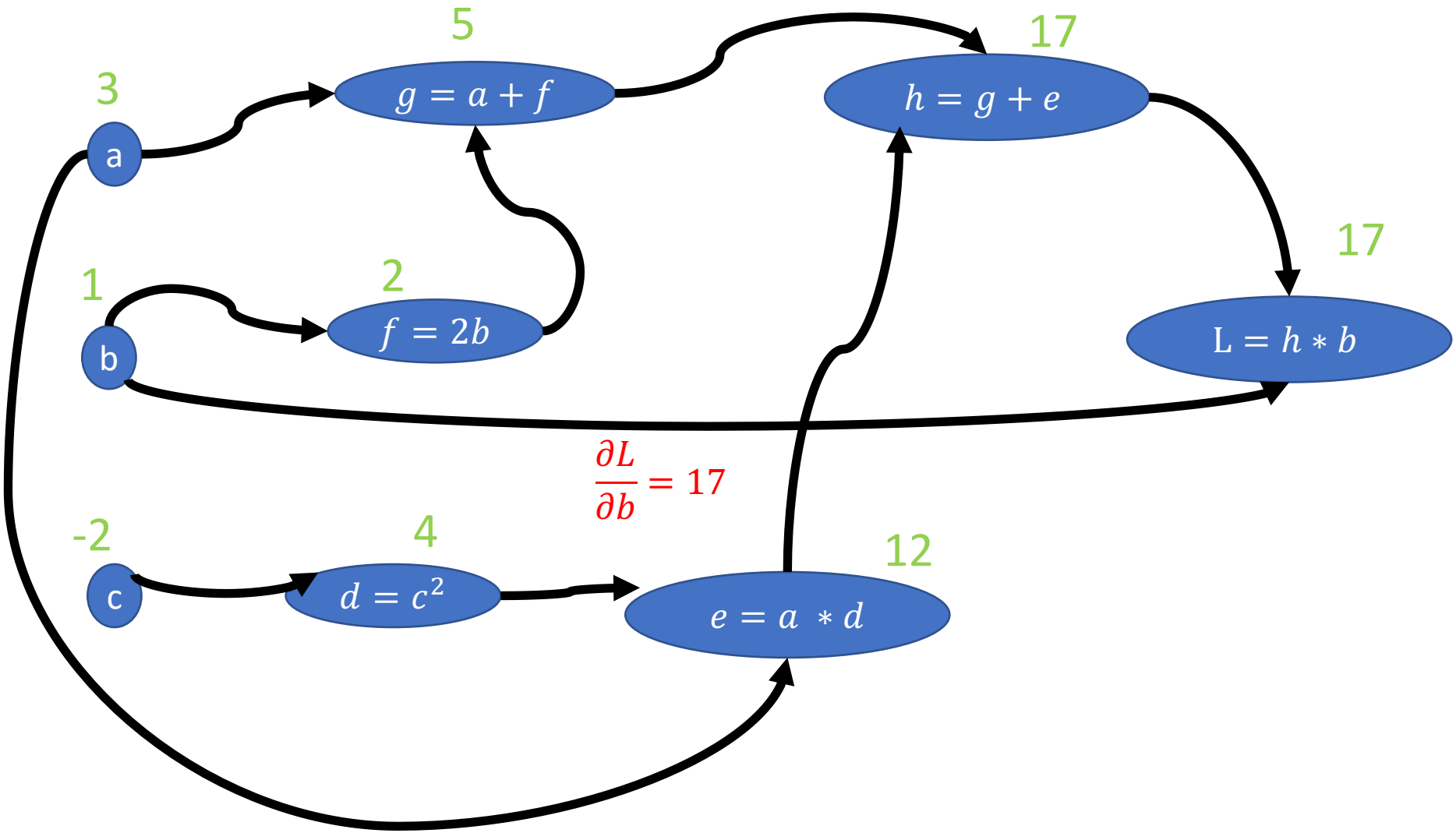
backward pass



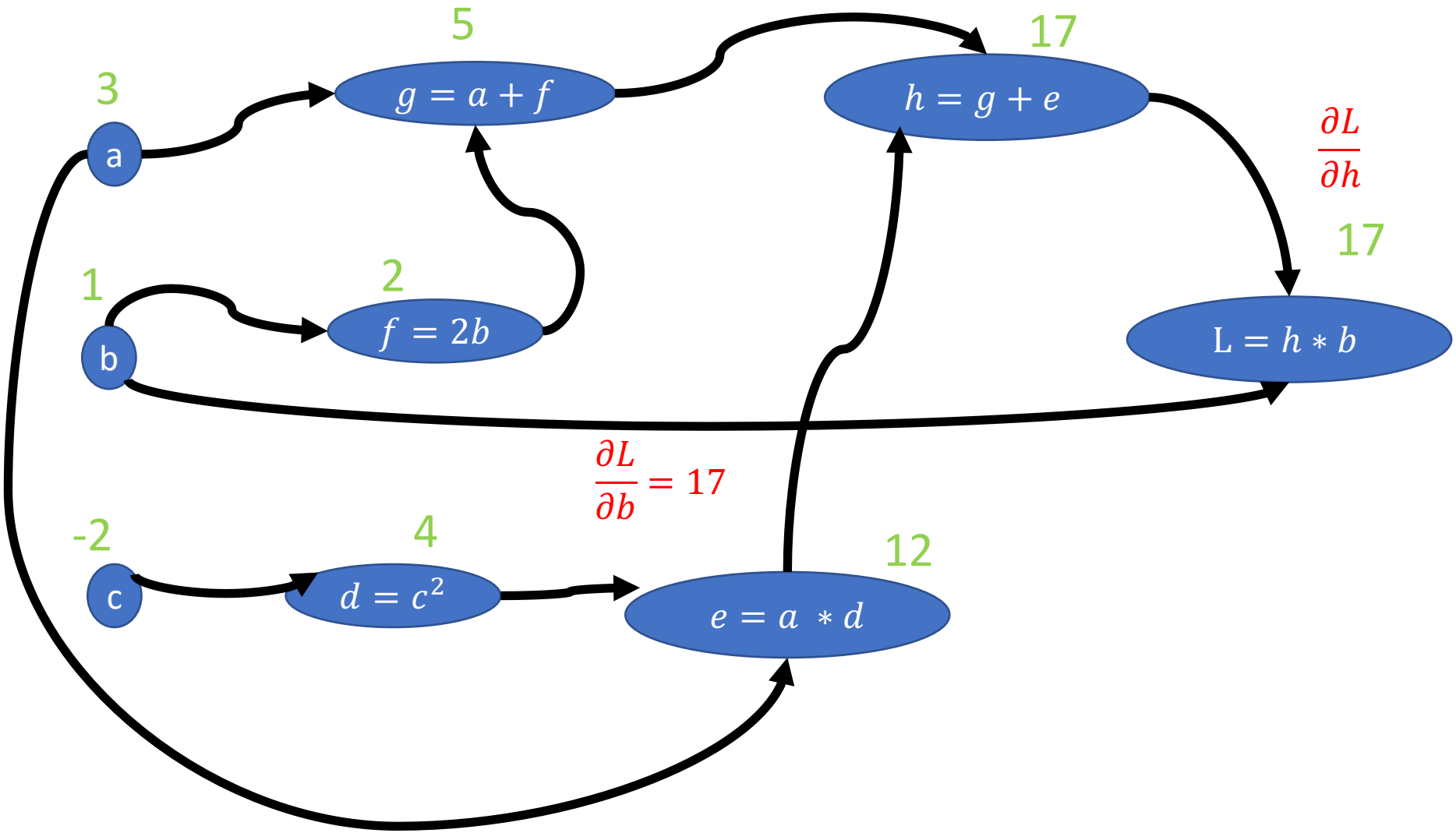
backward pass



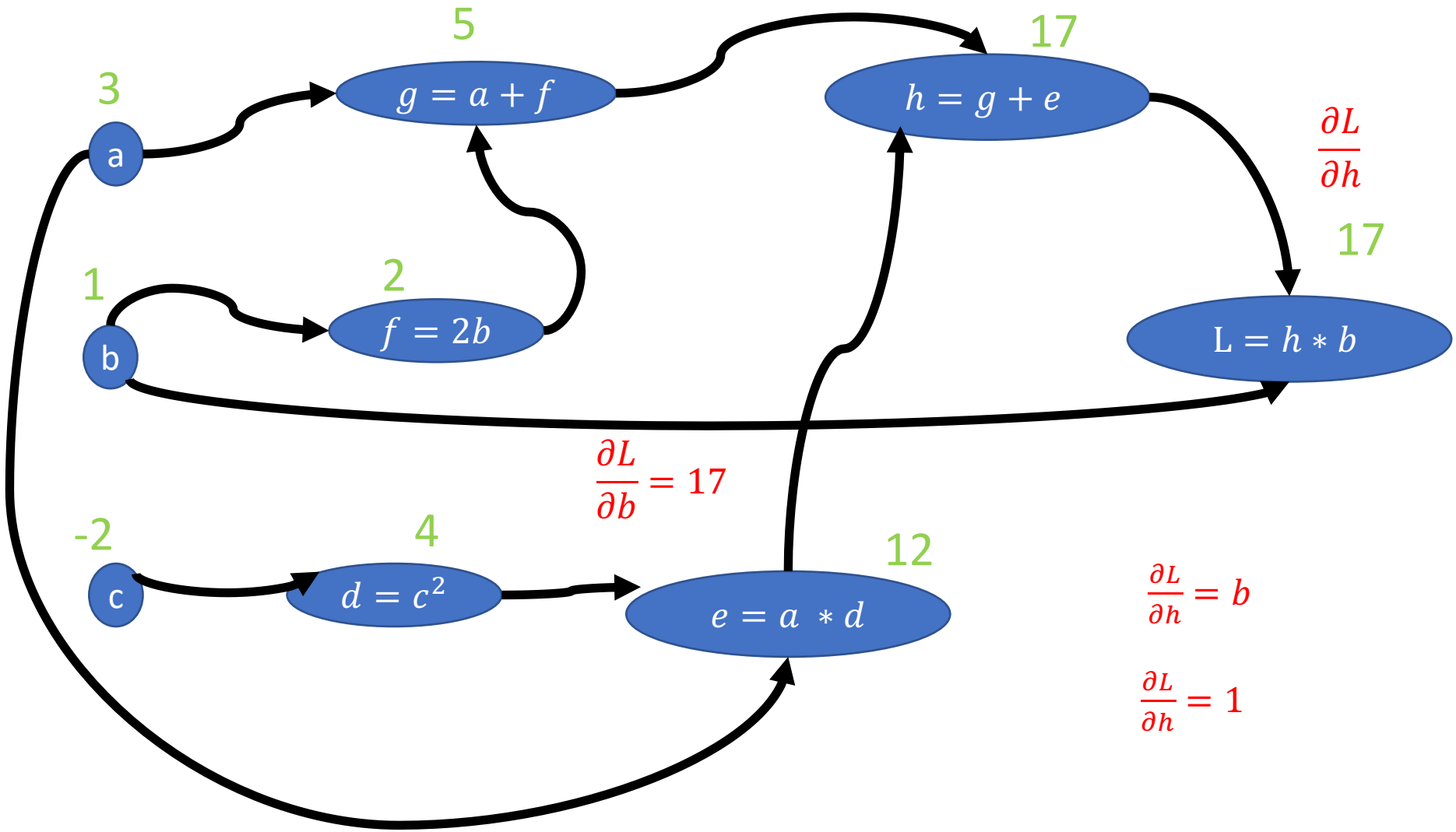
backward pass



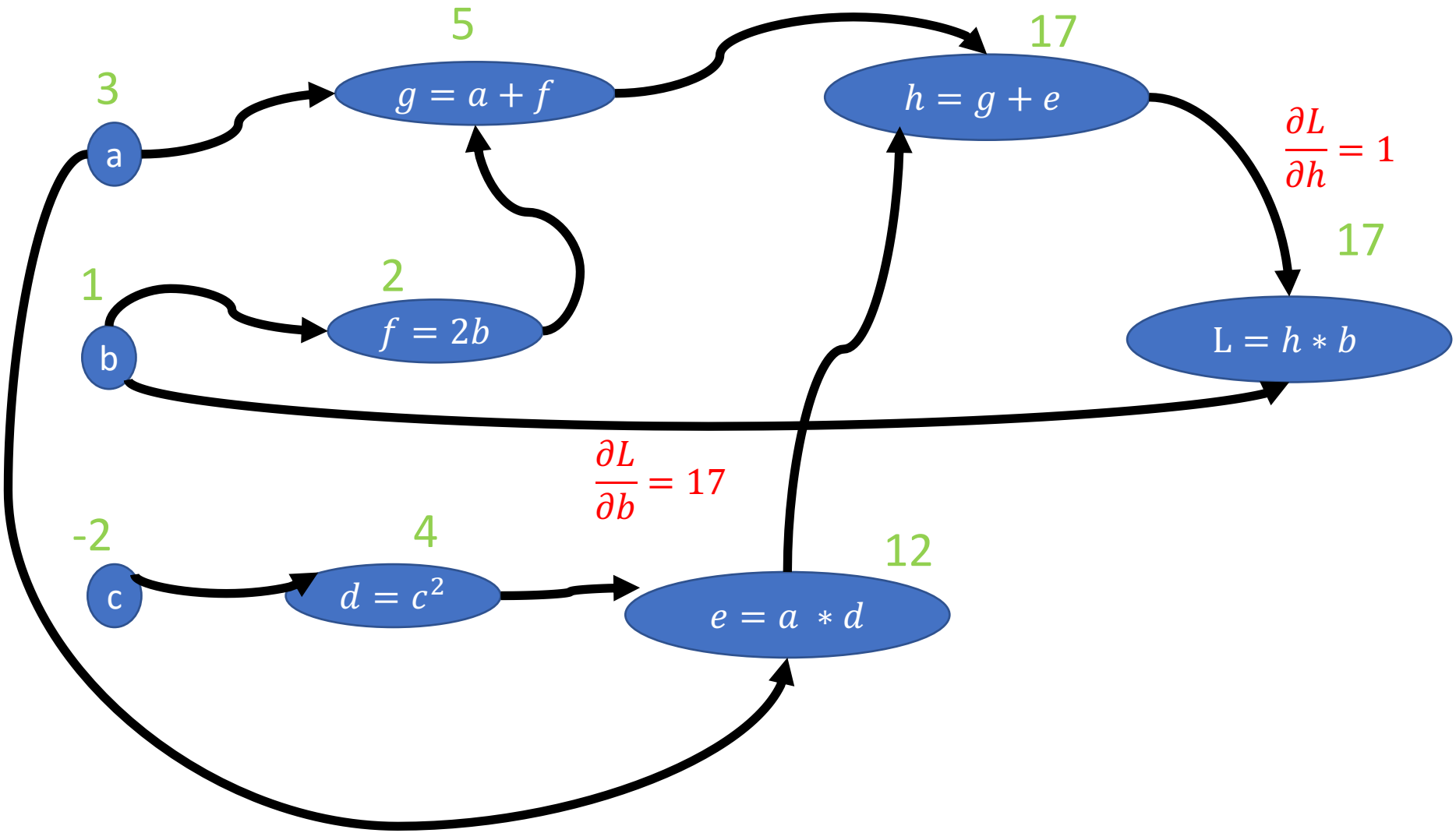
backward pass

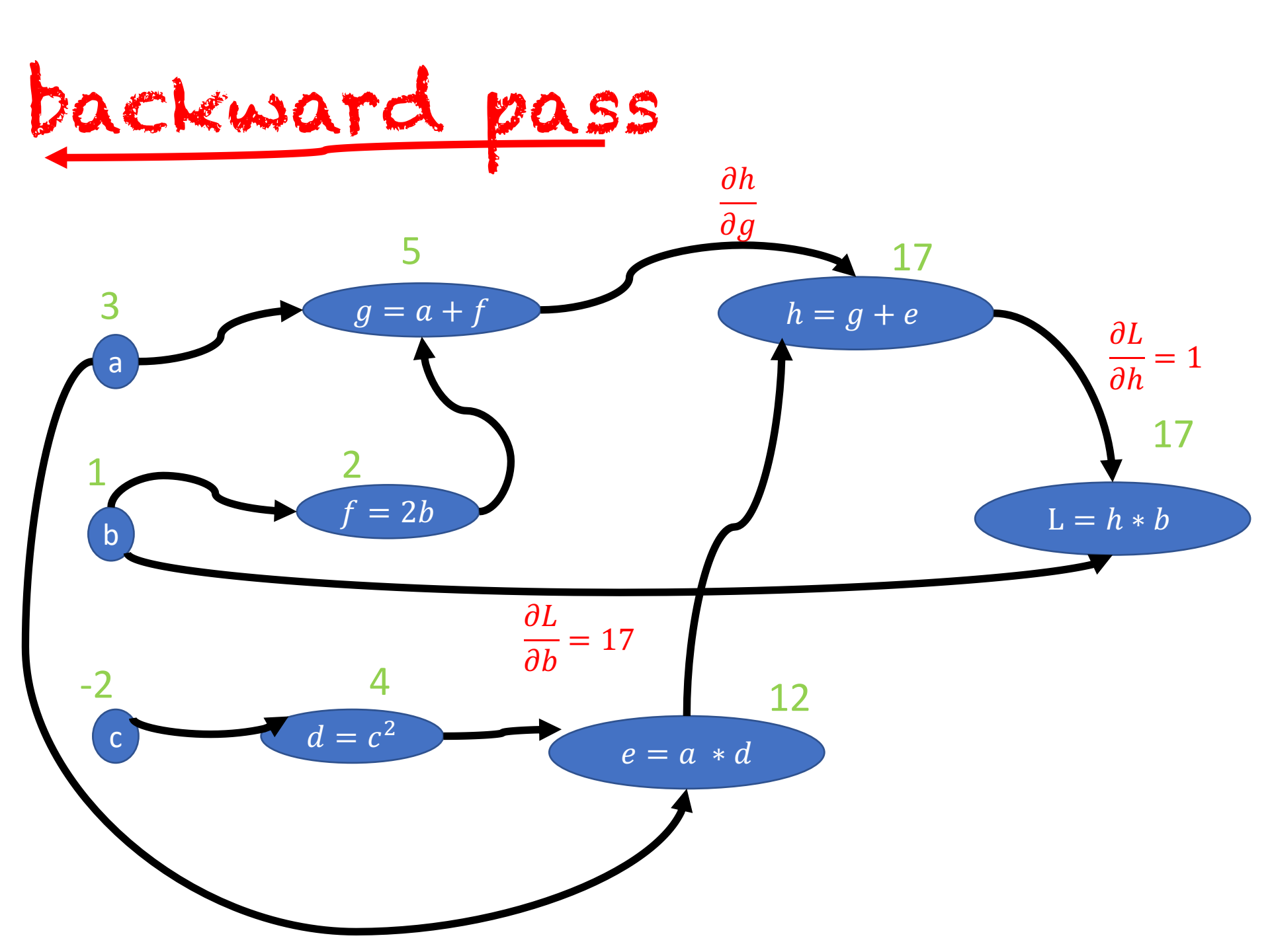


backward pass

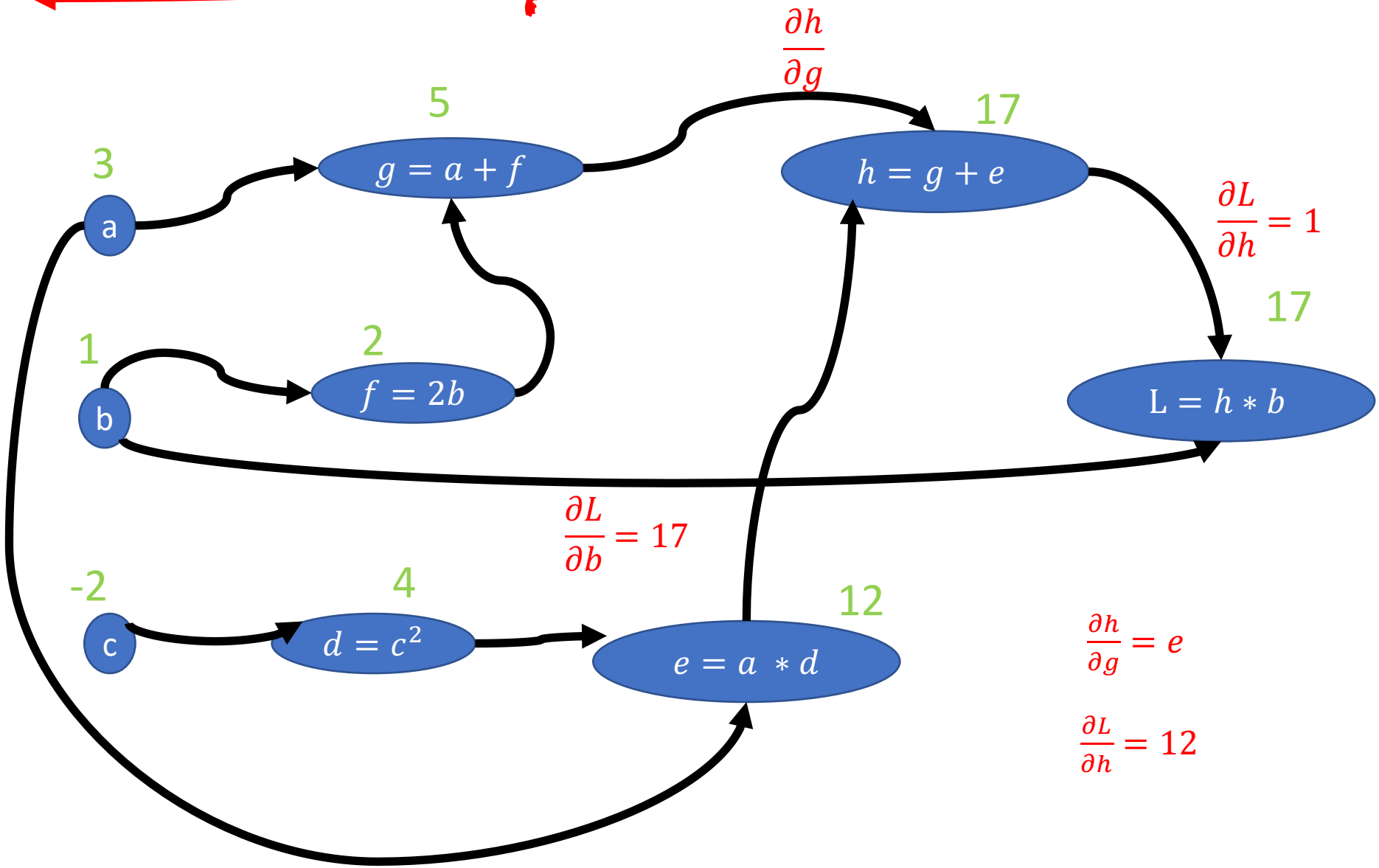


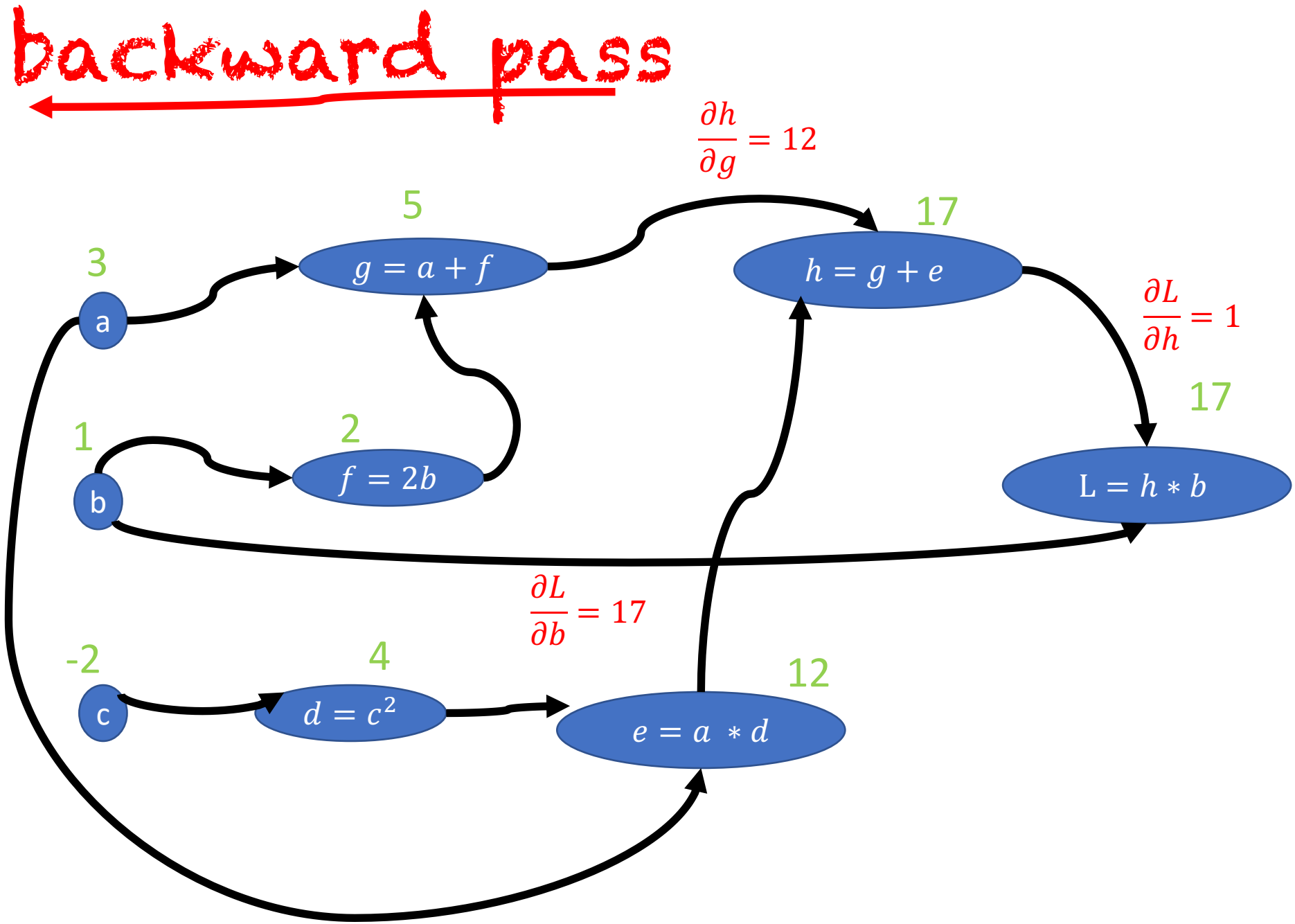
backward pass

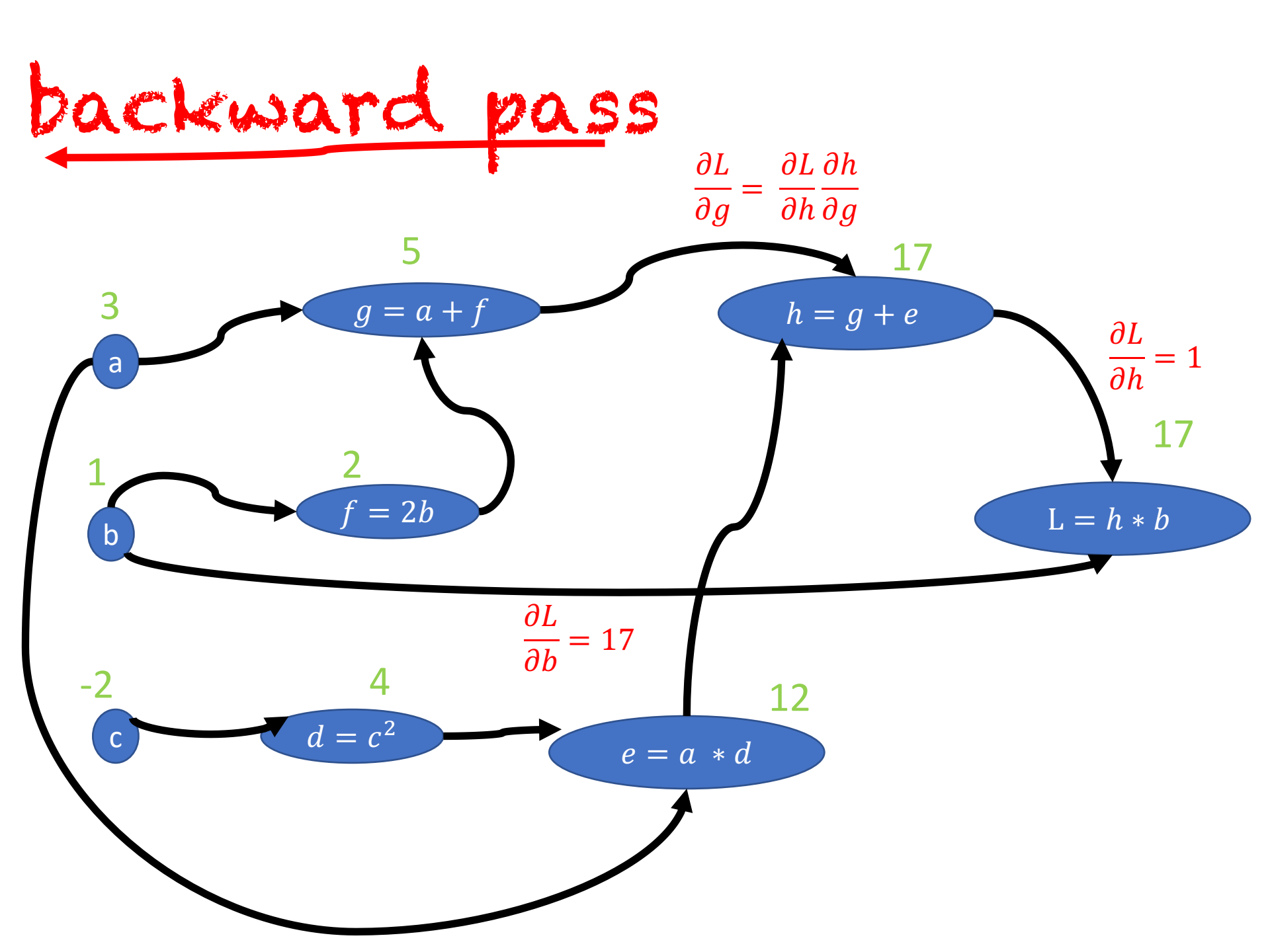


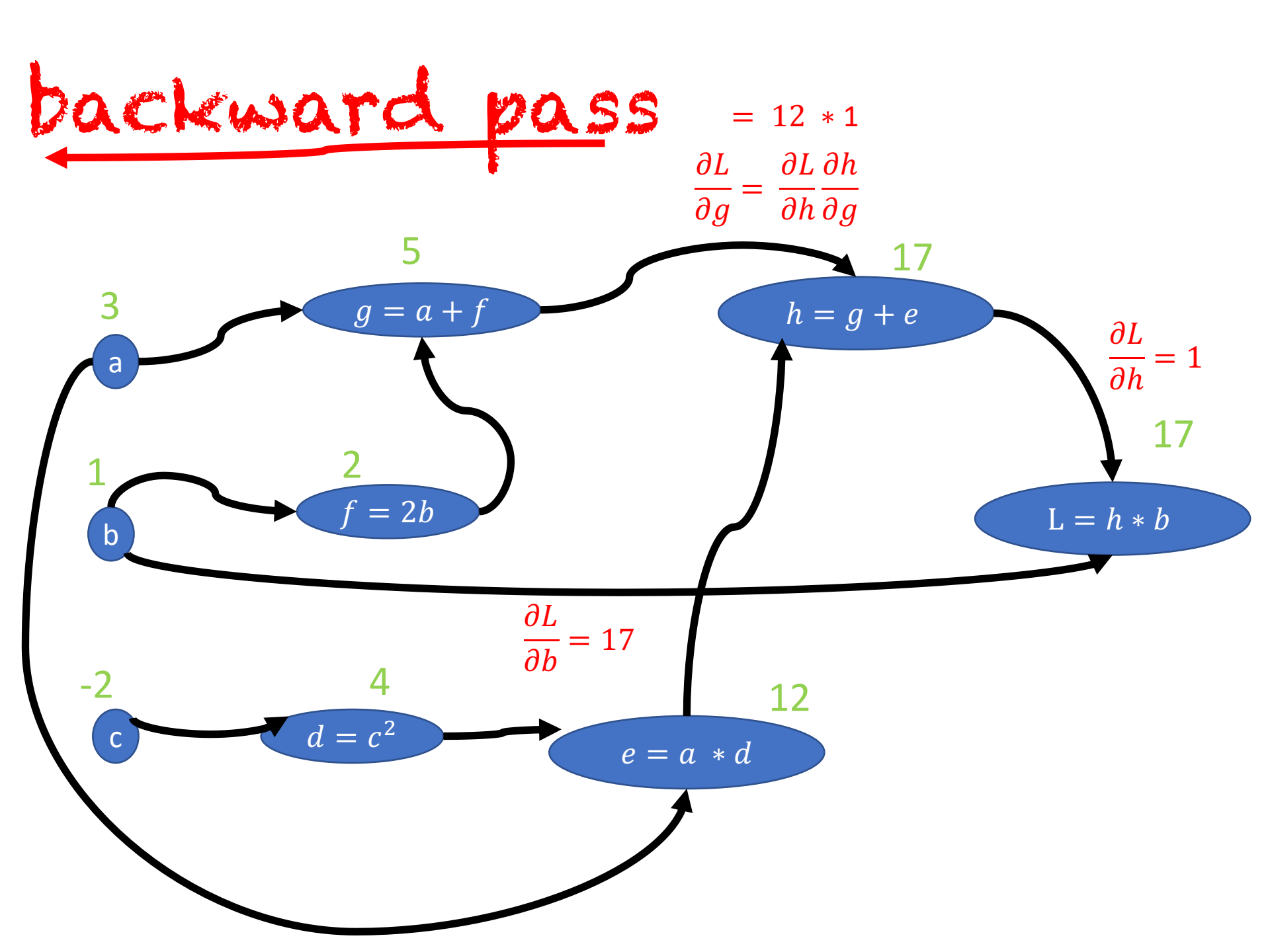

backward pass
$\frac{\partial h}{\partial g}$
5
17
3
$g = a + f$
$h = g + e$
$\frac{\partial L}{\partial h} = 1$
a
17
1
2
$f = 2b$
$L = h * b$
b
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$

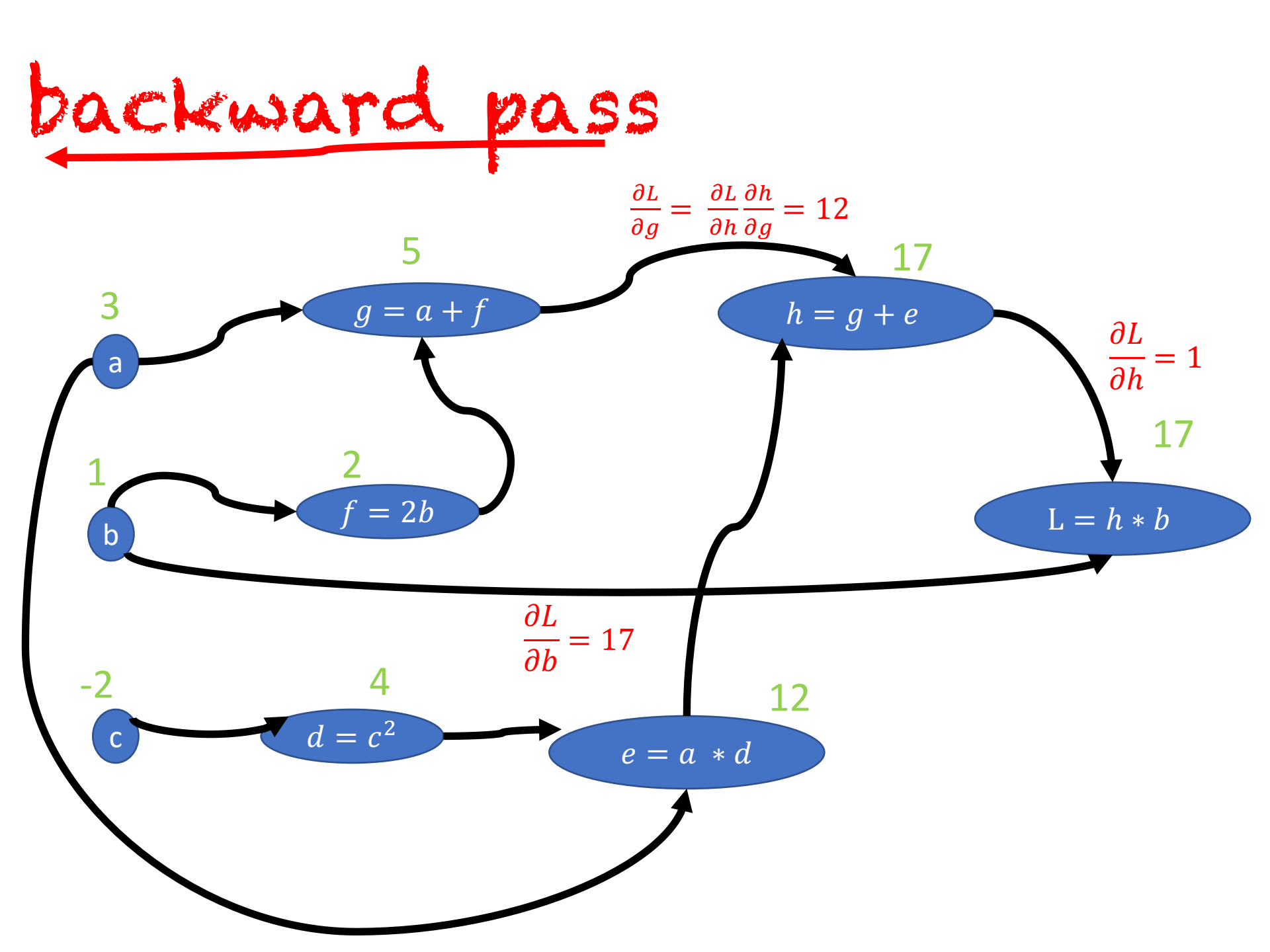
backward pass

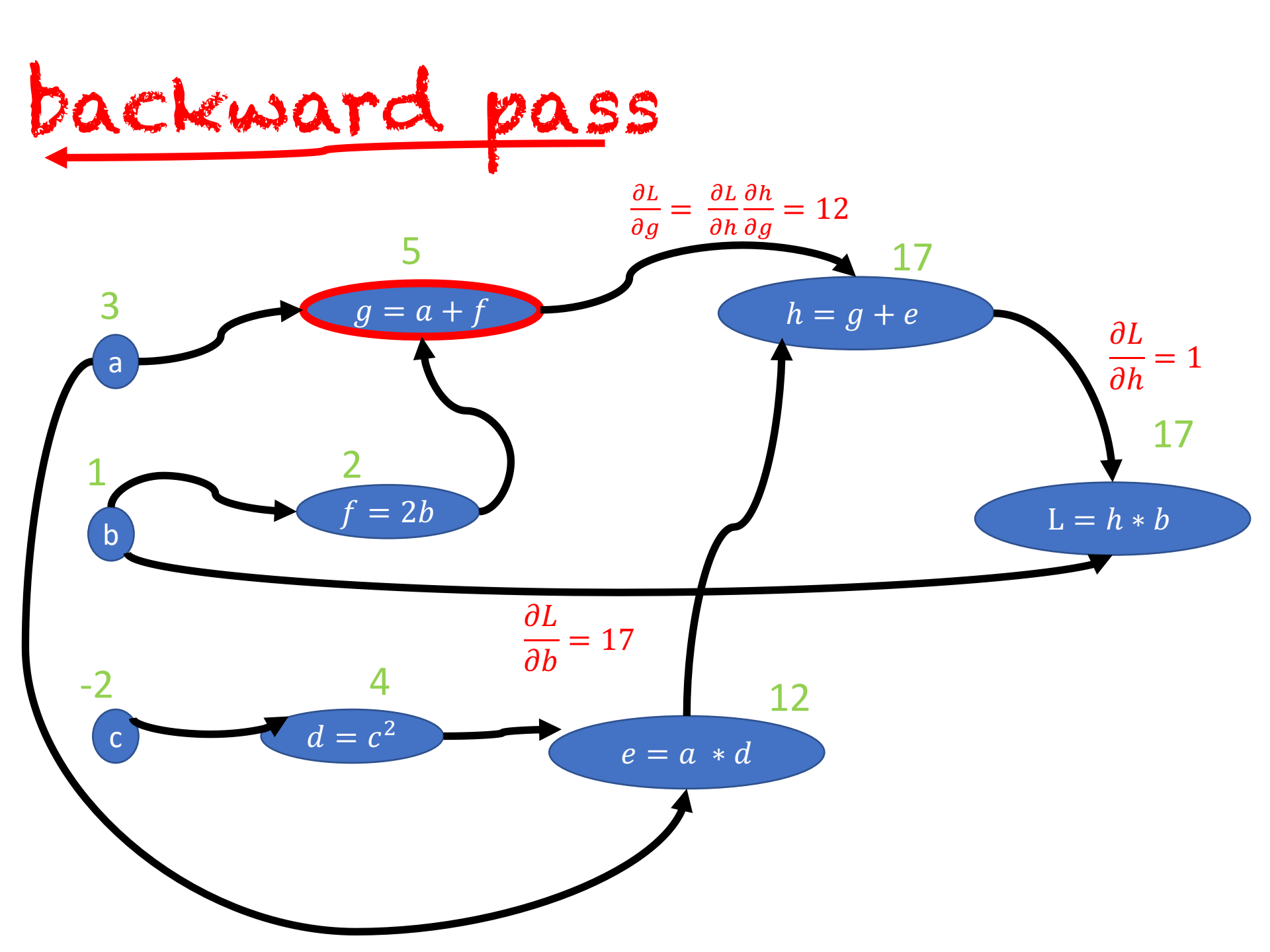


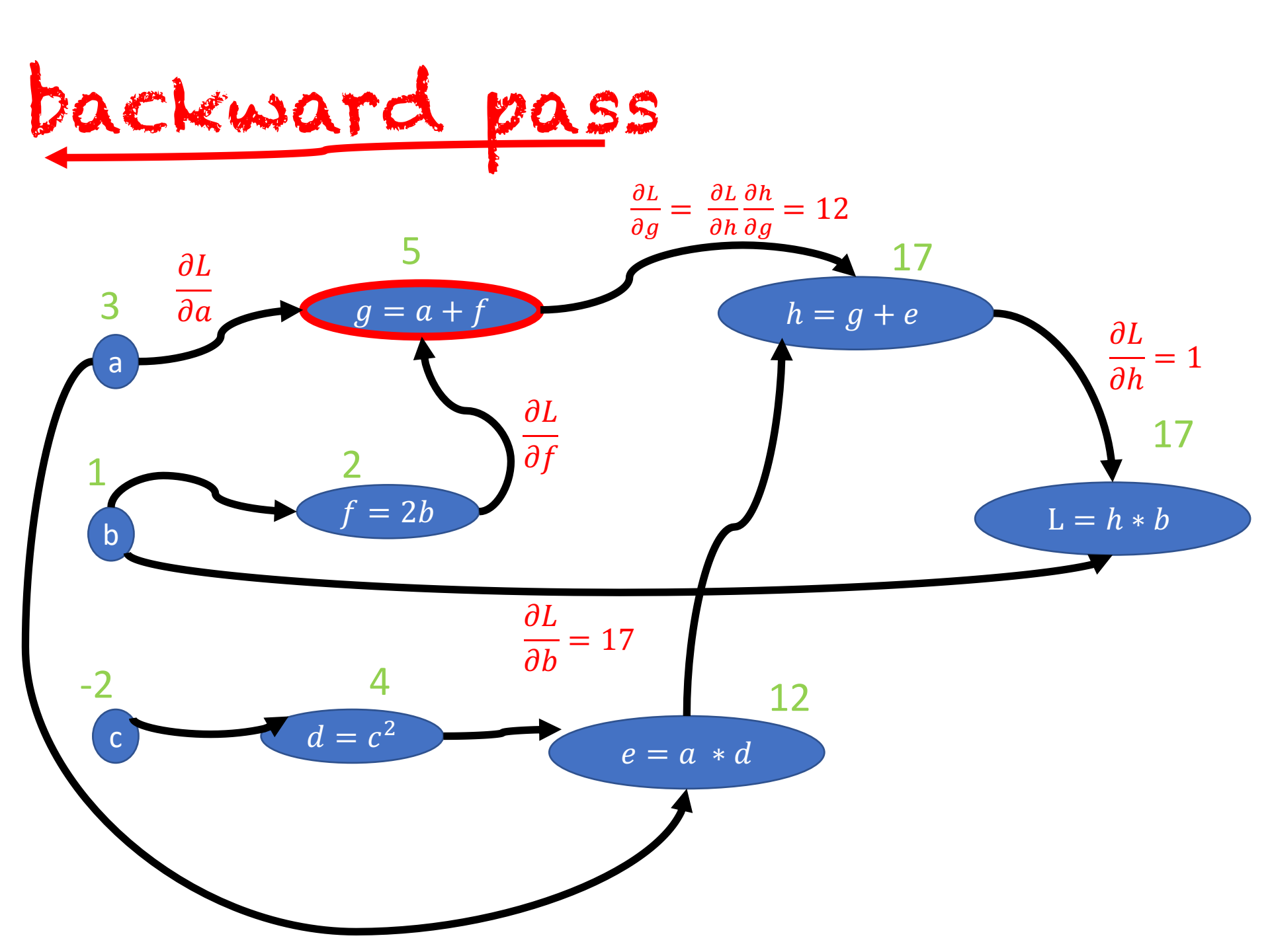

backward pass
$\frac{\partial h}{\partial g} = 12$
5
17
3
$g = a + f$
$h = g + e$
a
$\frac{\partial L}{\partial h} = 1$
17
1
2
$f = 2b$
$L = h * b$
b
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$

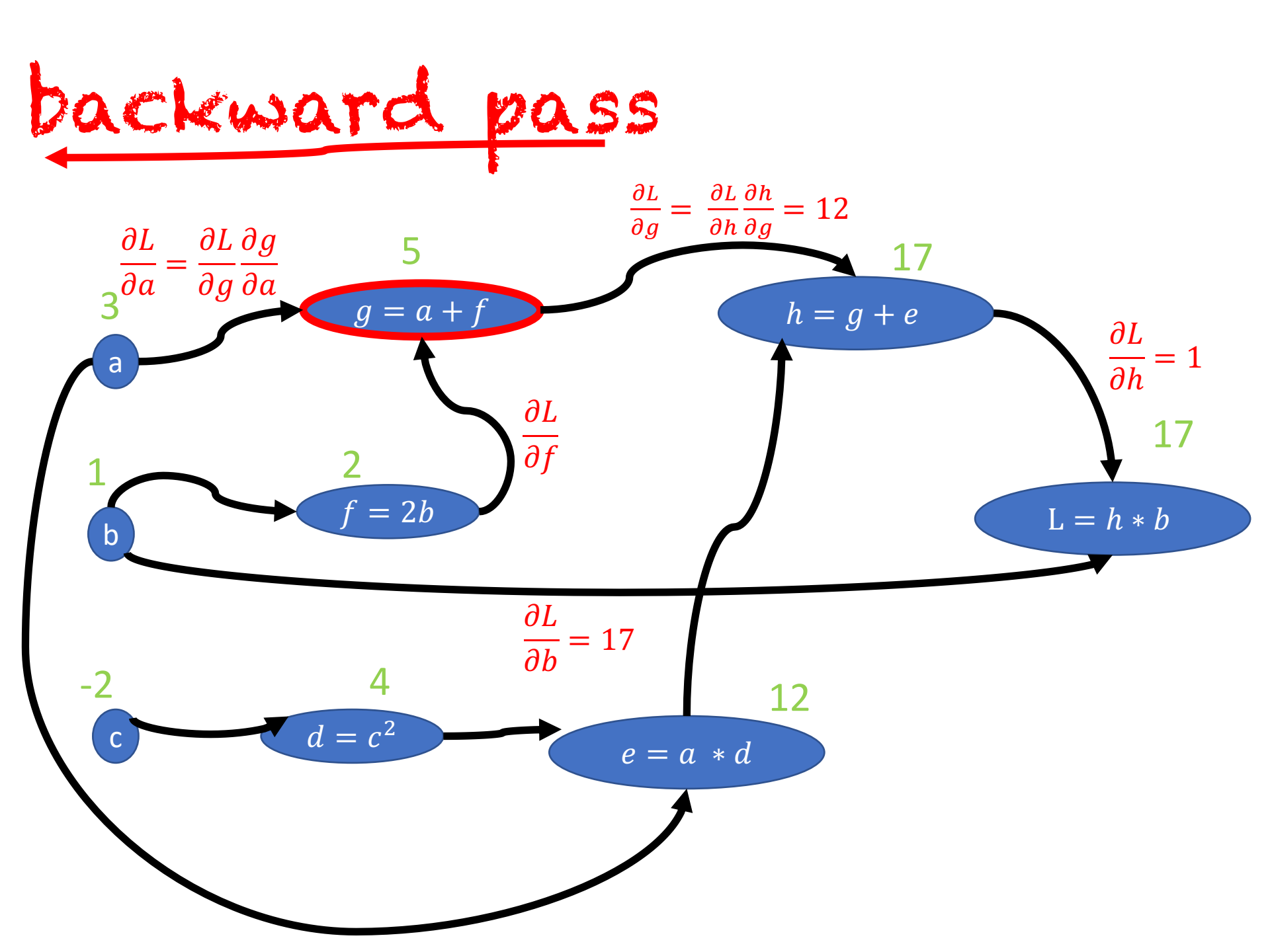

backward pass
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g}$
5
17
3
$g = a + f$
$h = g + e$
a
$\frac{\partial L}{\partial h} = 1$
17
1
2
$f = 2b$
$L = h * b$
b
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$

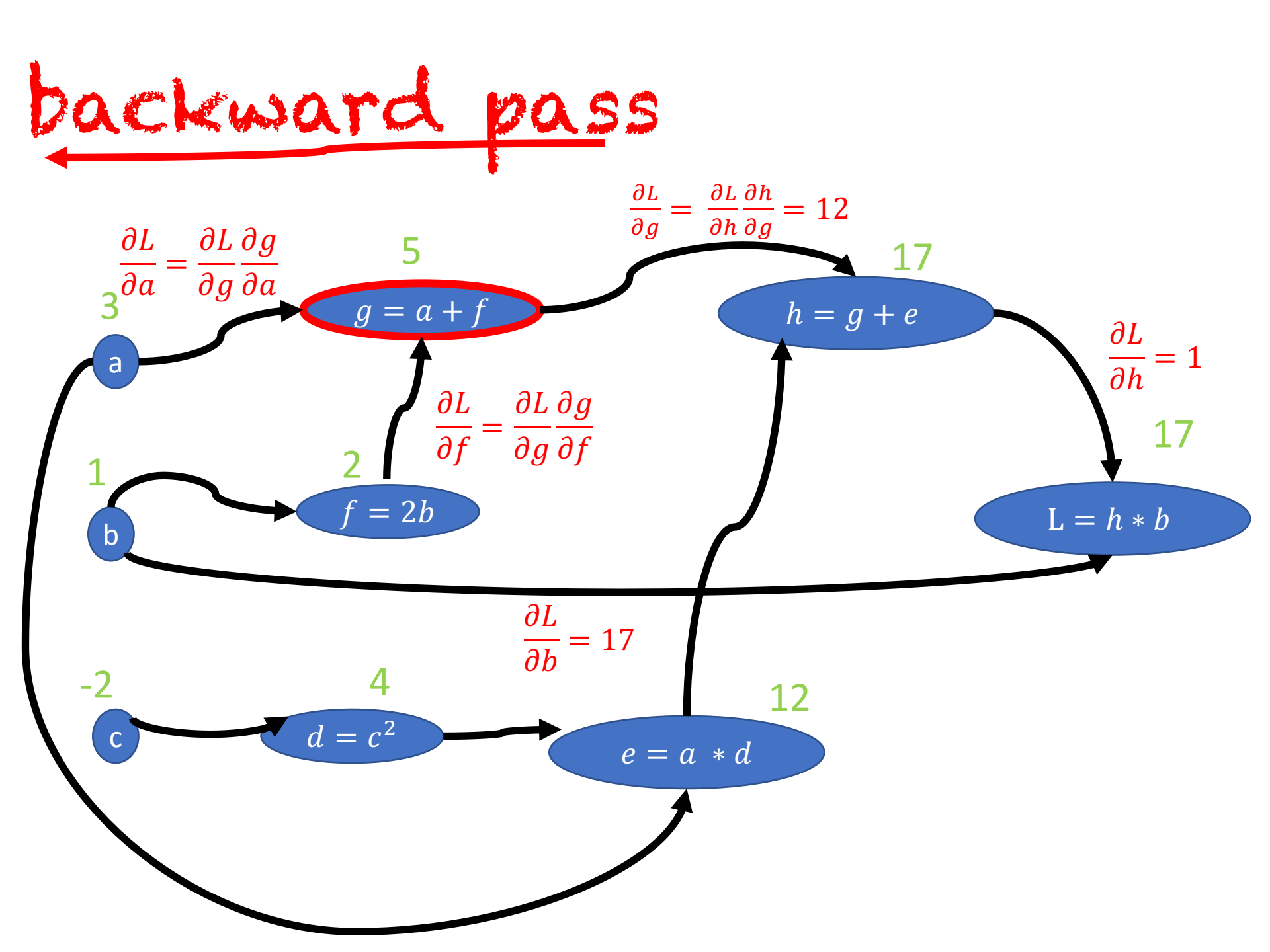

backward pass
$= 12 * 1$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g}$
5
17
3
$g = a + f$
$h = g + e$
$\frac{\partial L}{\partial h} = 1$
a
17
1
2
$f = 2b$
$L = h * b$
b
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$


backward pass
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
5
17
3
$g = a + f$
$h = g + e$
a
$\frac{\partial L}{\partial h} = 1$
17
1
2
$f = 2b$
b
$L = h * b$
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$

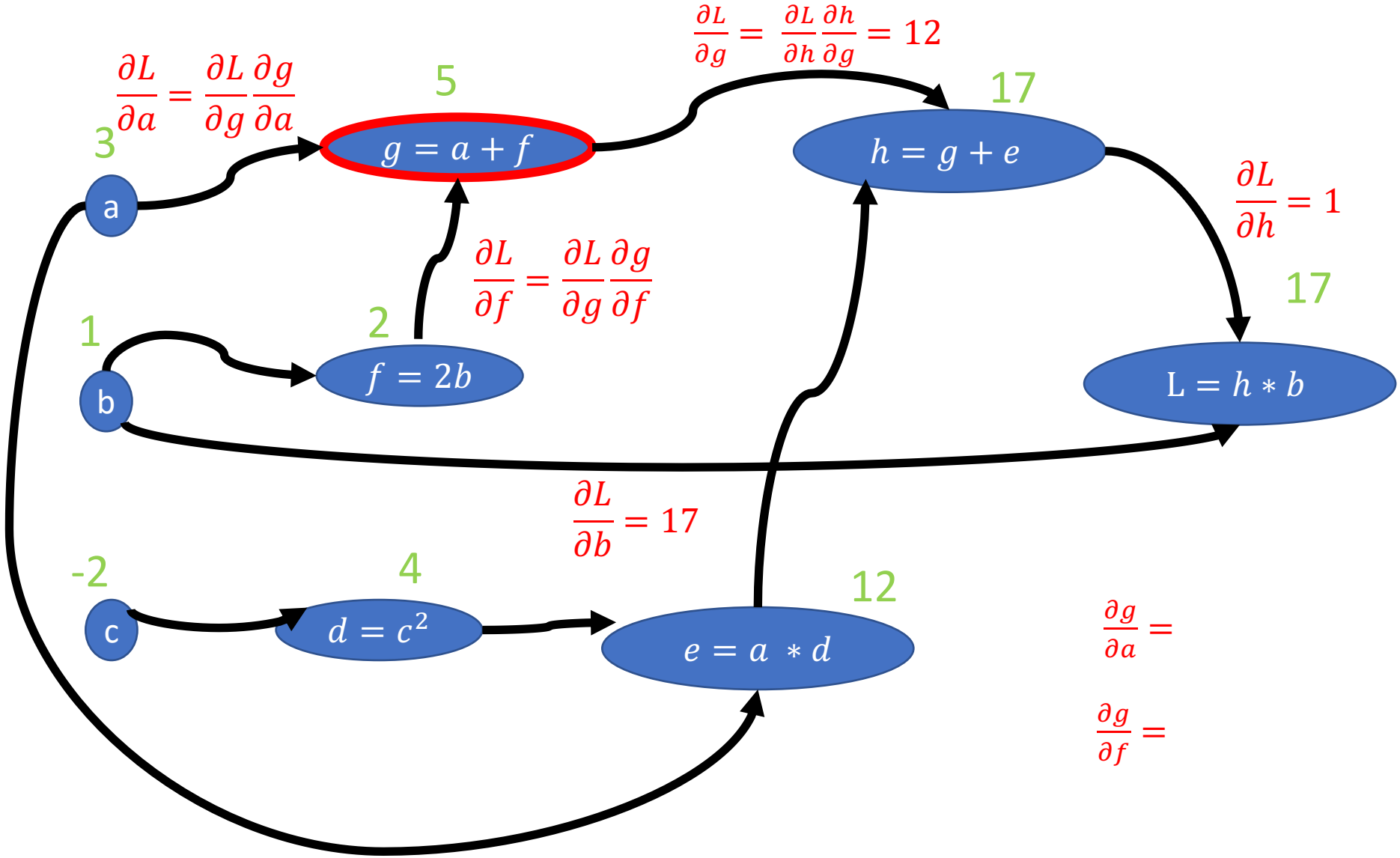

backward pass
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
5
17
3
$g = a + f$
$h = g + e$
$\frac{\partial L}{\partial h} = 1$
17
1
2
$f = 2b$
$L = h * b$
b
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$

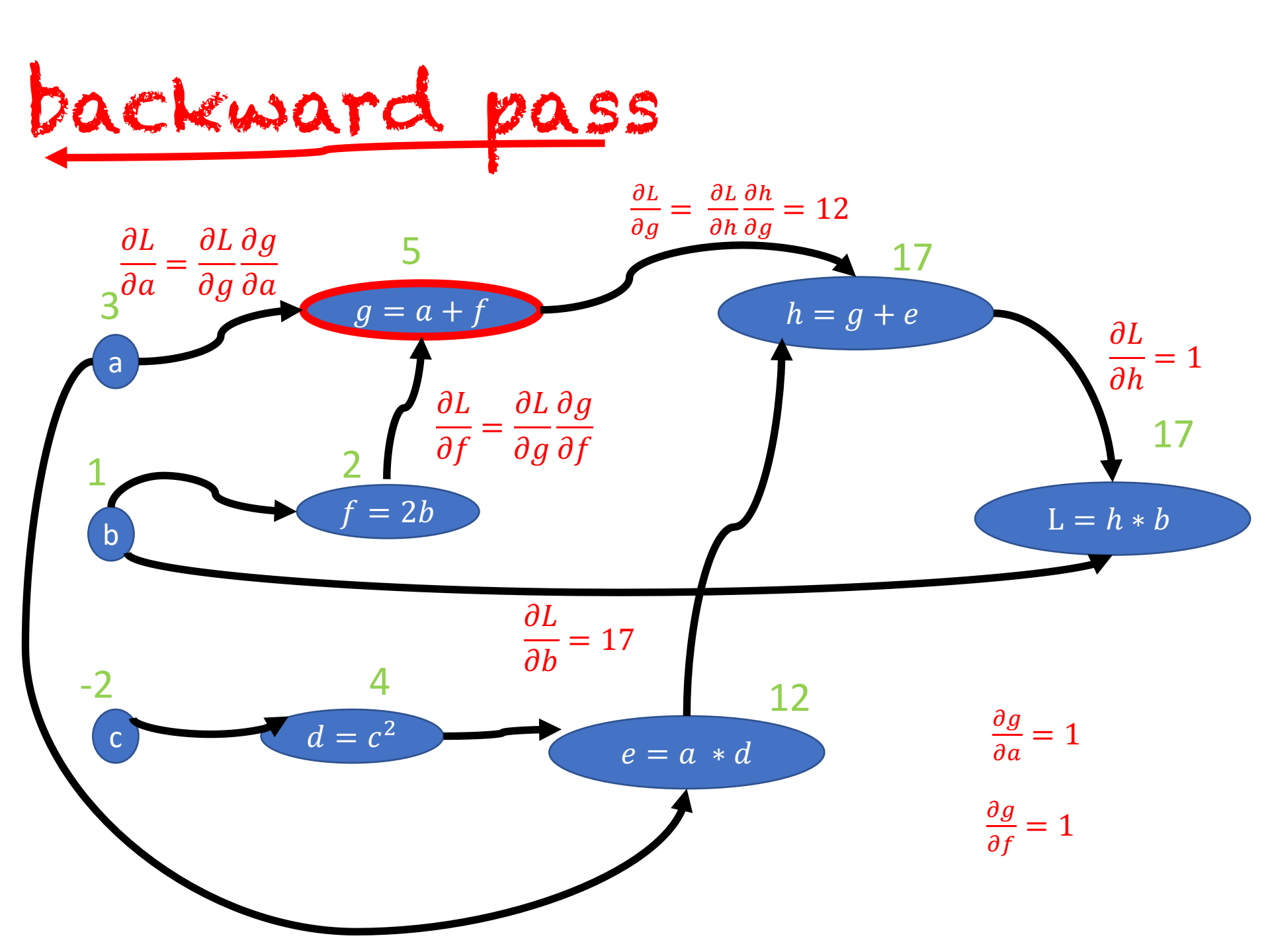

backward pass
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
5
17
$\frac{\partial L}{\partial a}$
3
a
$g = a + f$
$h = g + e$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial f}$
17
1
2
$f = 2b$
b
$L = h * b$
$\frac{\partial L}{\partial b} = 17$
-2
4
12
c
$d = c^2$
$e = a * d$

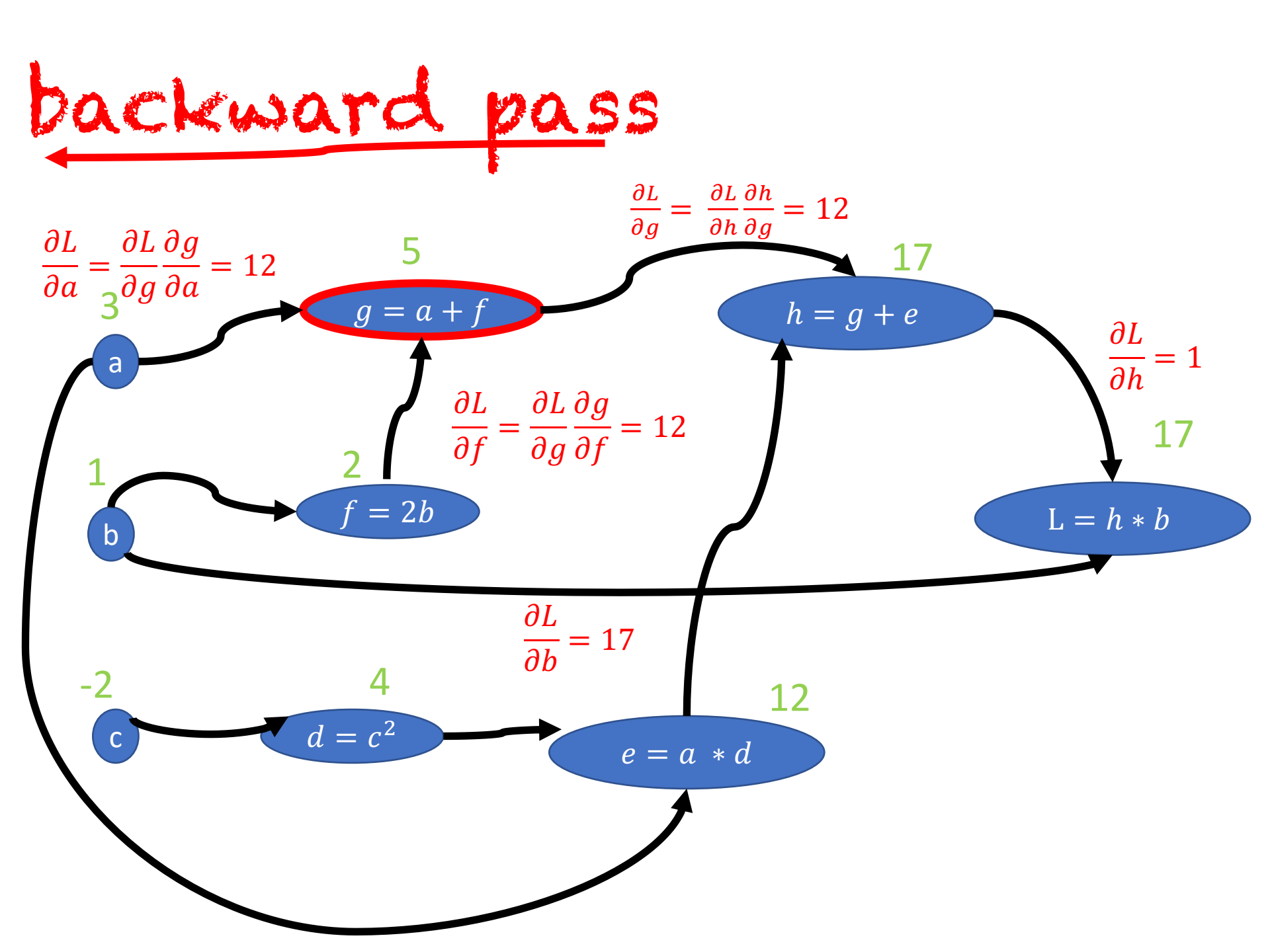

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a}$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial f}$
$\frac{\partial L}{\partial b} = 17$
3
5
17
17
1
2
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$d = c^2$
$e = a * d$
$L = h * b$

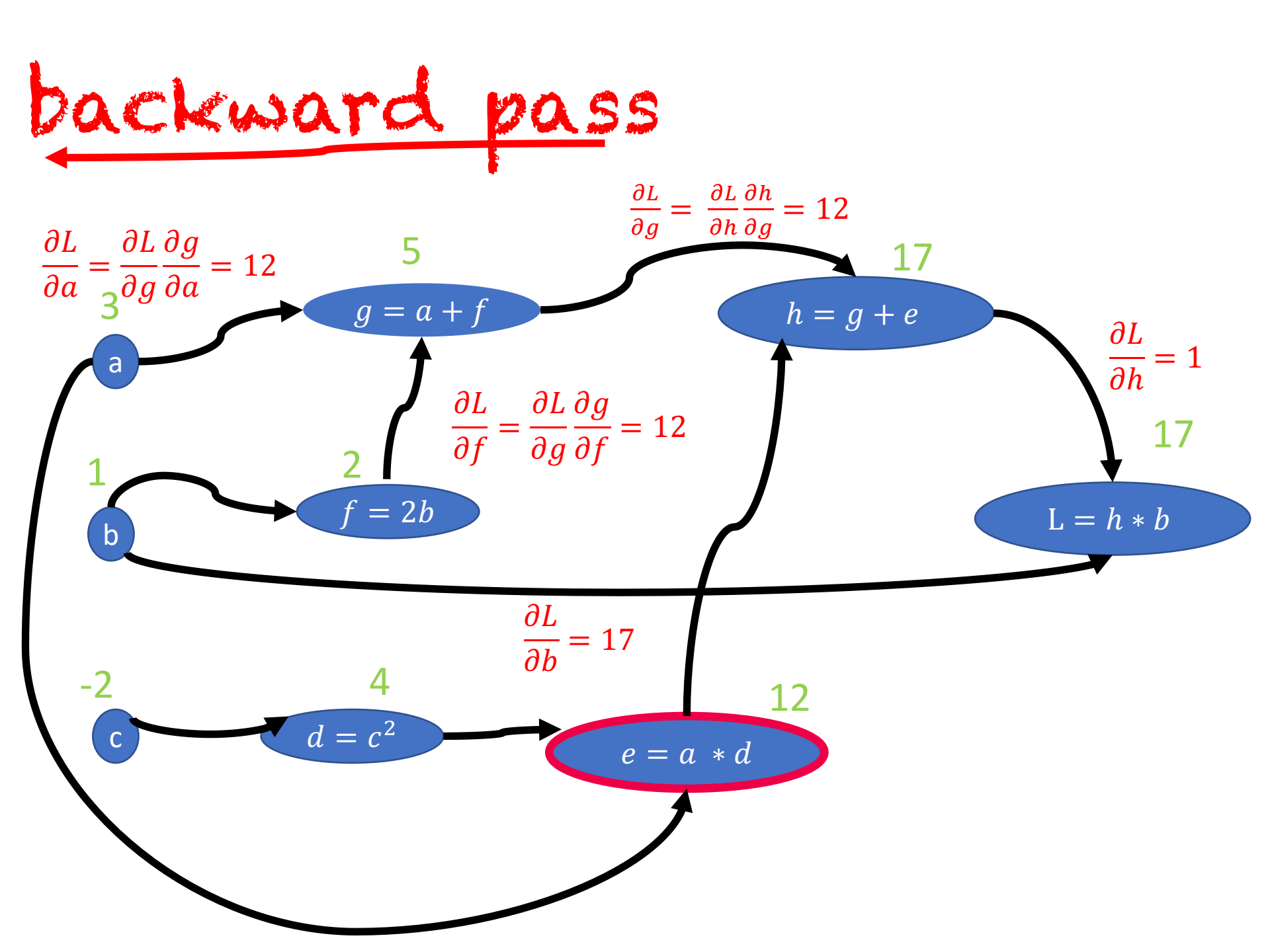

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a}$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f}$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
3
5
17
17
1
2
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$\mathrm{L} = h * b$
$d = c^2$
$e = a * d$

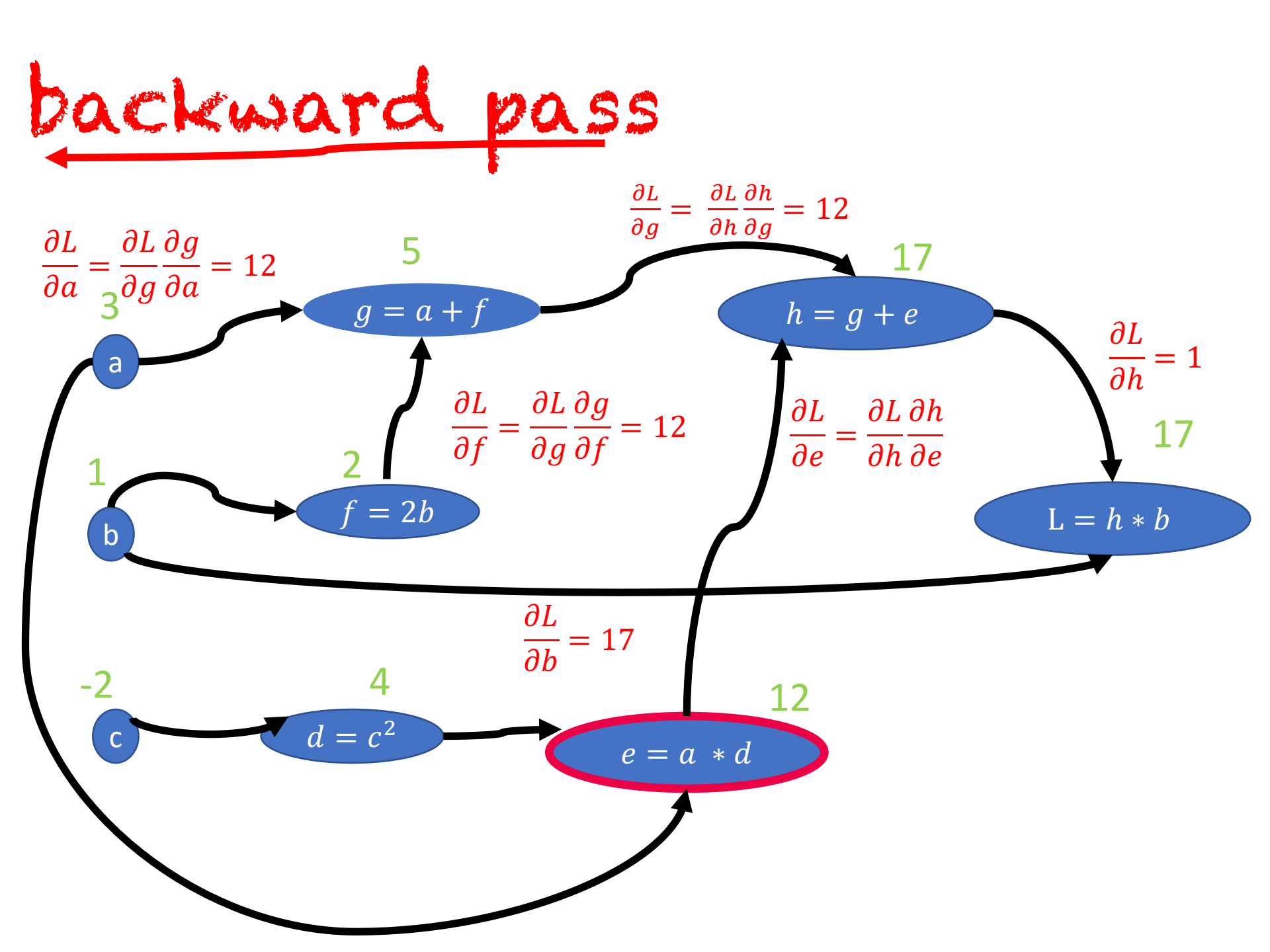
backward pass

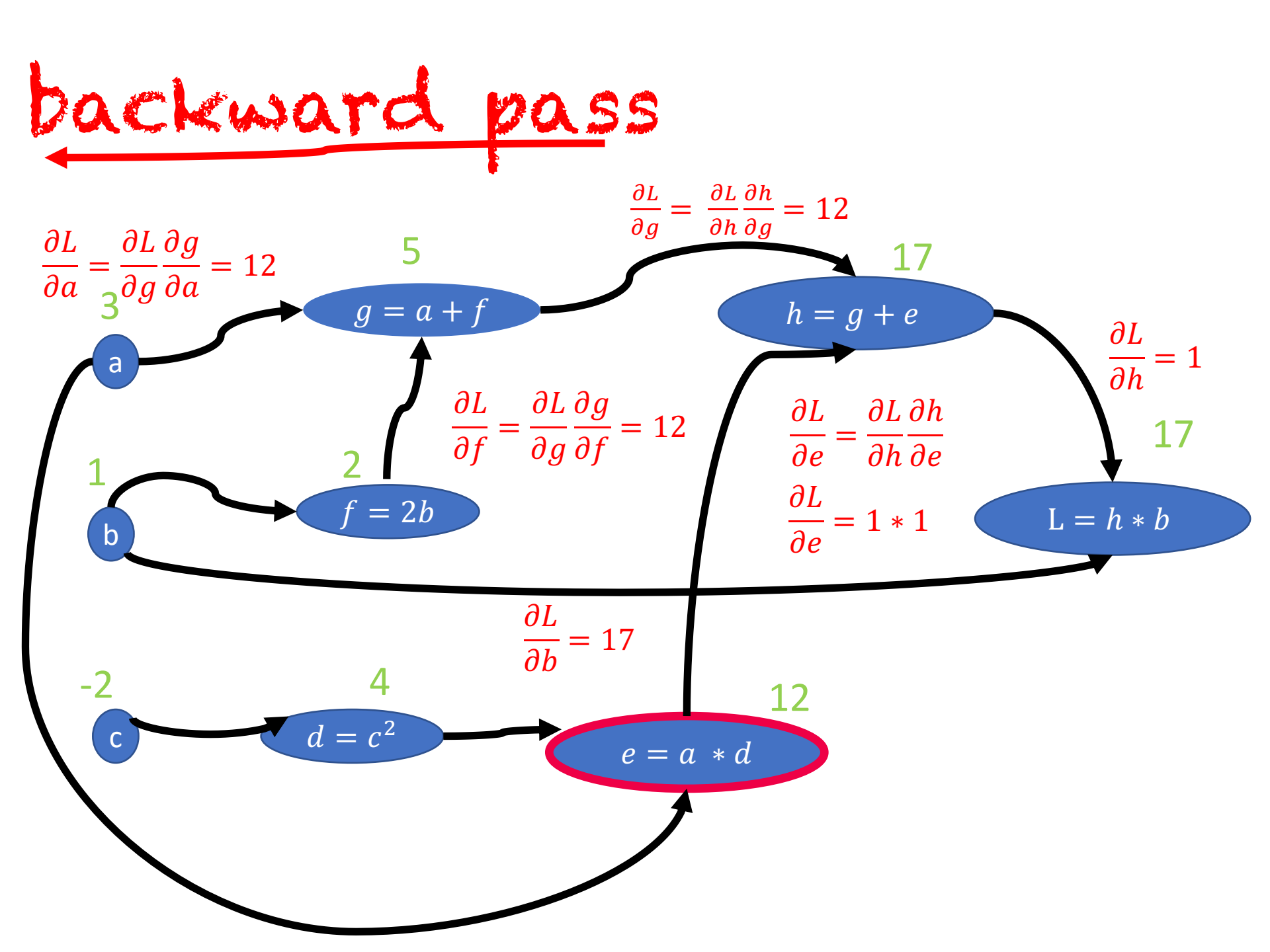


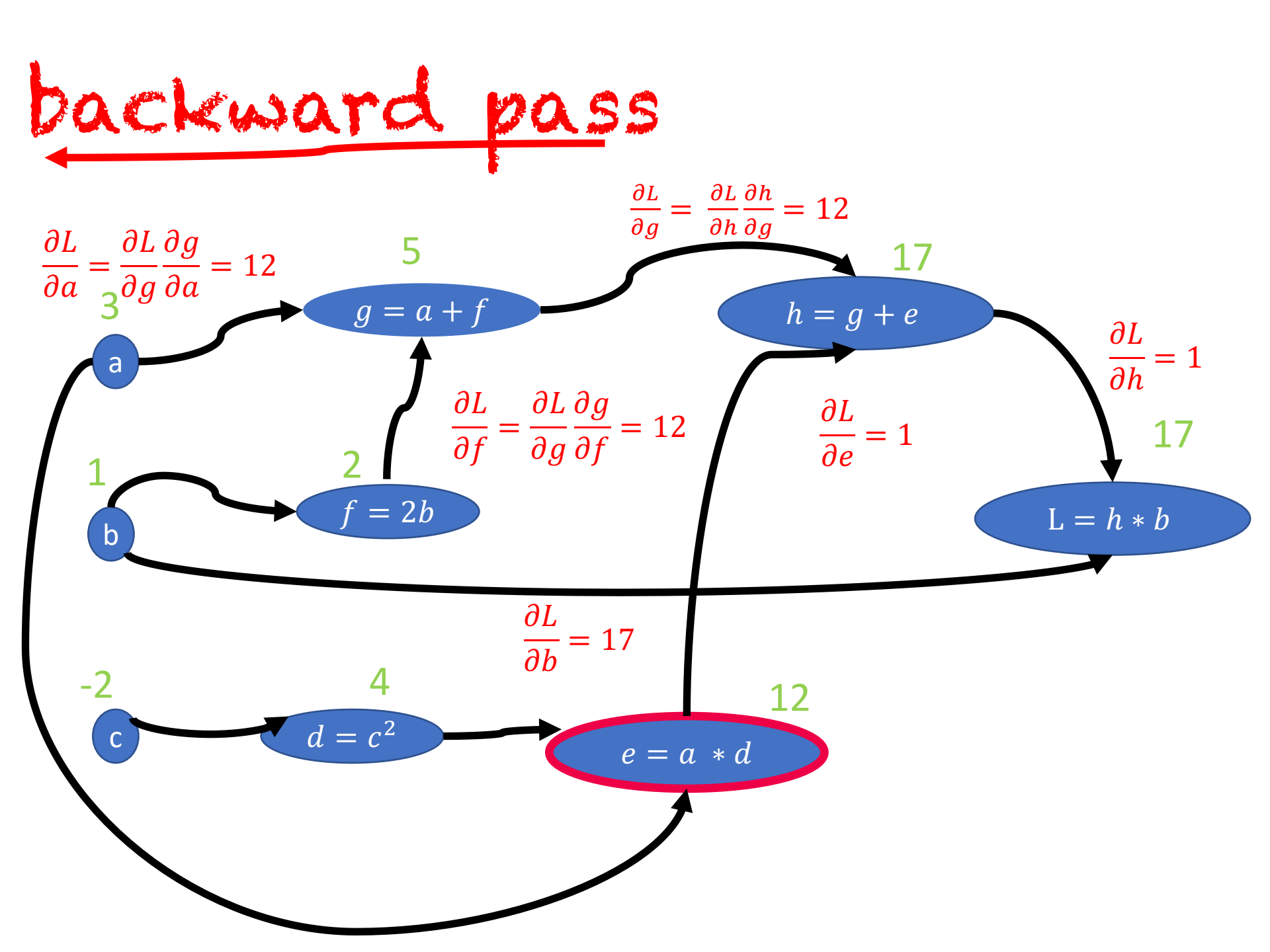

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a}$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f}$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial g}{\partial a} = 1$
$\frac{\partial g}{\partial f} = 1$
3
5
17
17
1
2
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

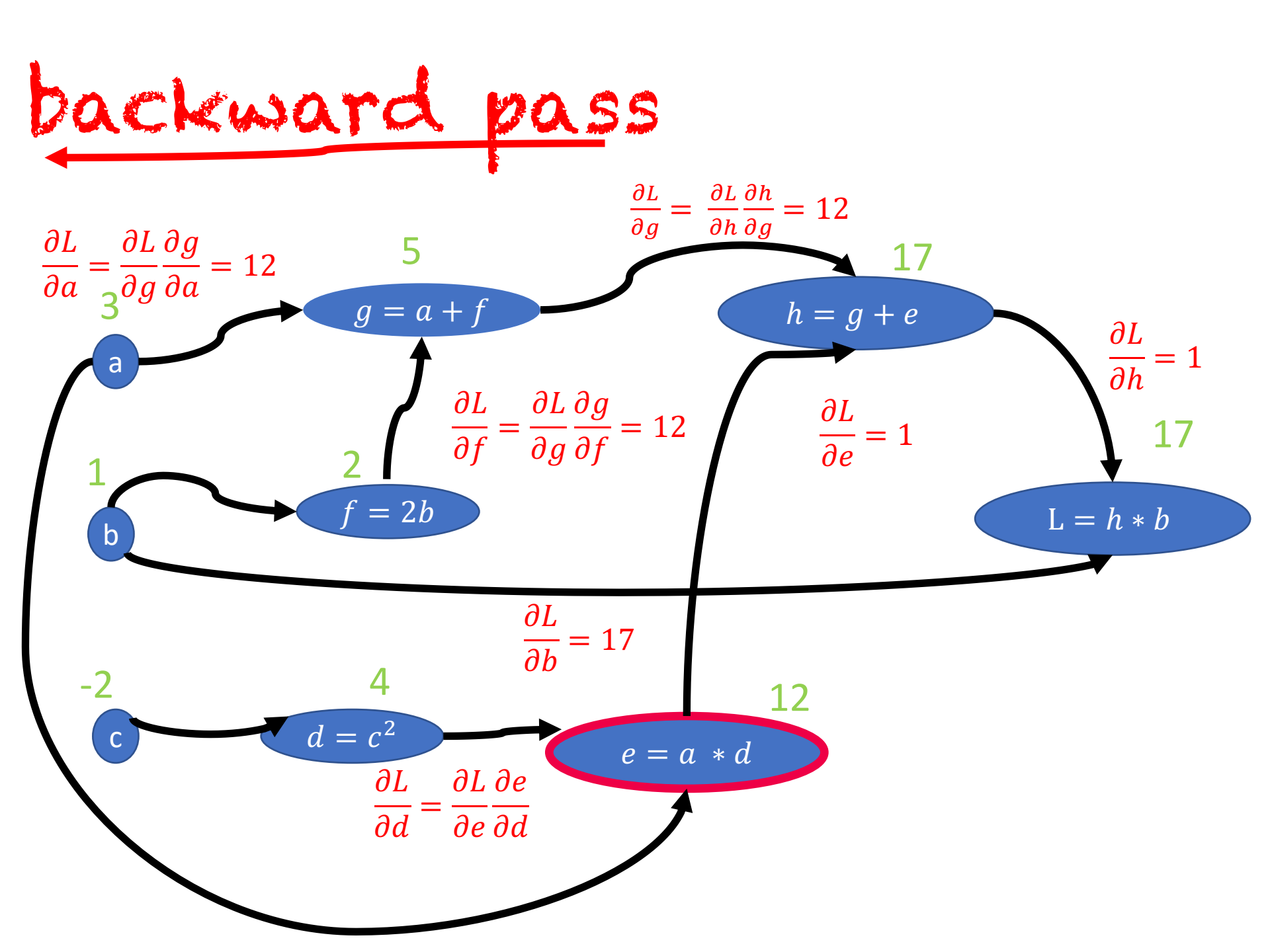

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
3
5
17
17
1
2
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

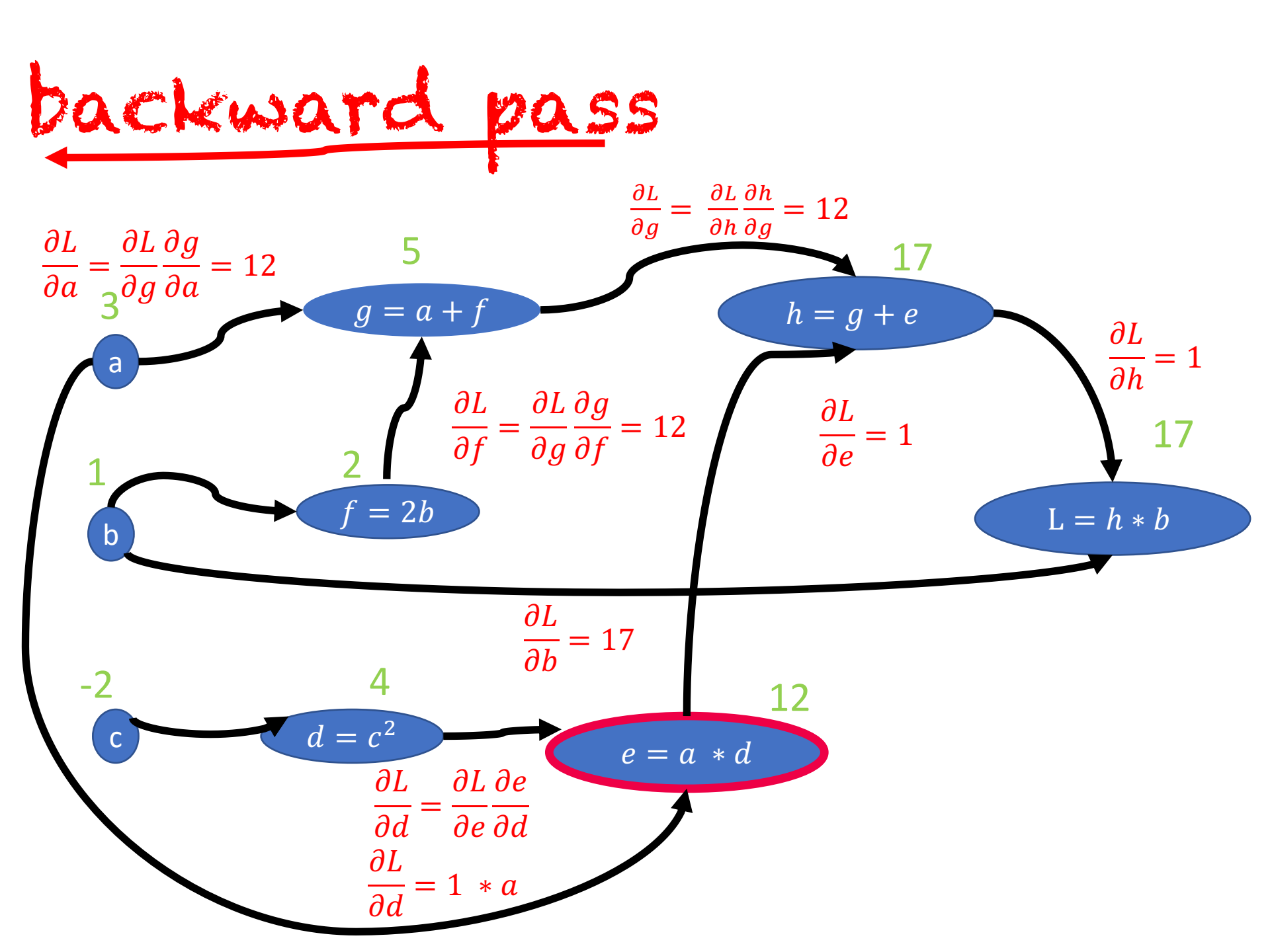

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
3
5
17
17
1
2
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

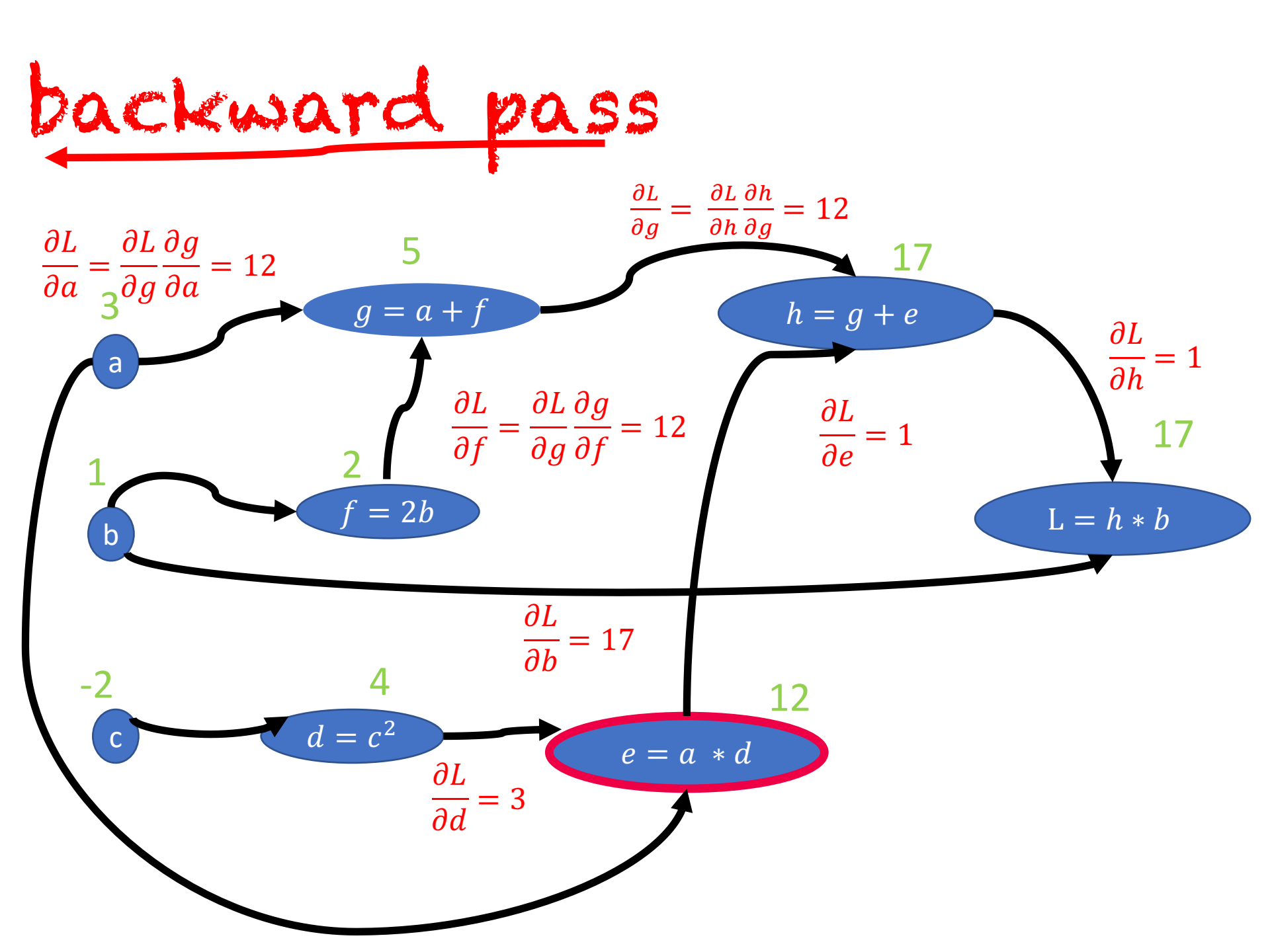

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial e}$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
3
5
17
17
1
2
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

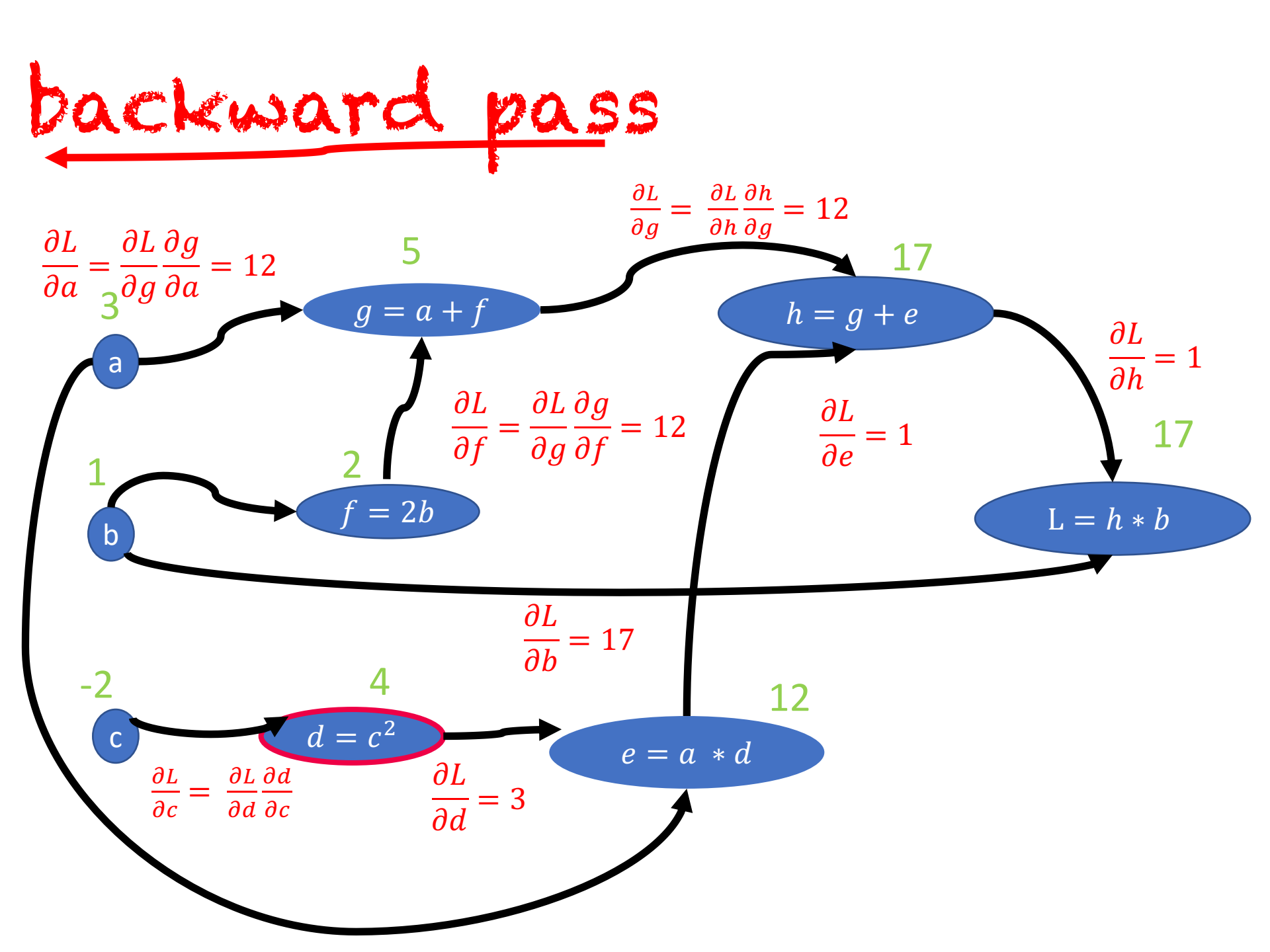

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial e} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial e}$
$\frac{\partial L}{\partial e} = 1 * 1$
$\frac{\partial L}{\partial b} = 17$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$\mathrm{L} = h * b$
$d = c^2$
$e = a * d$

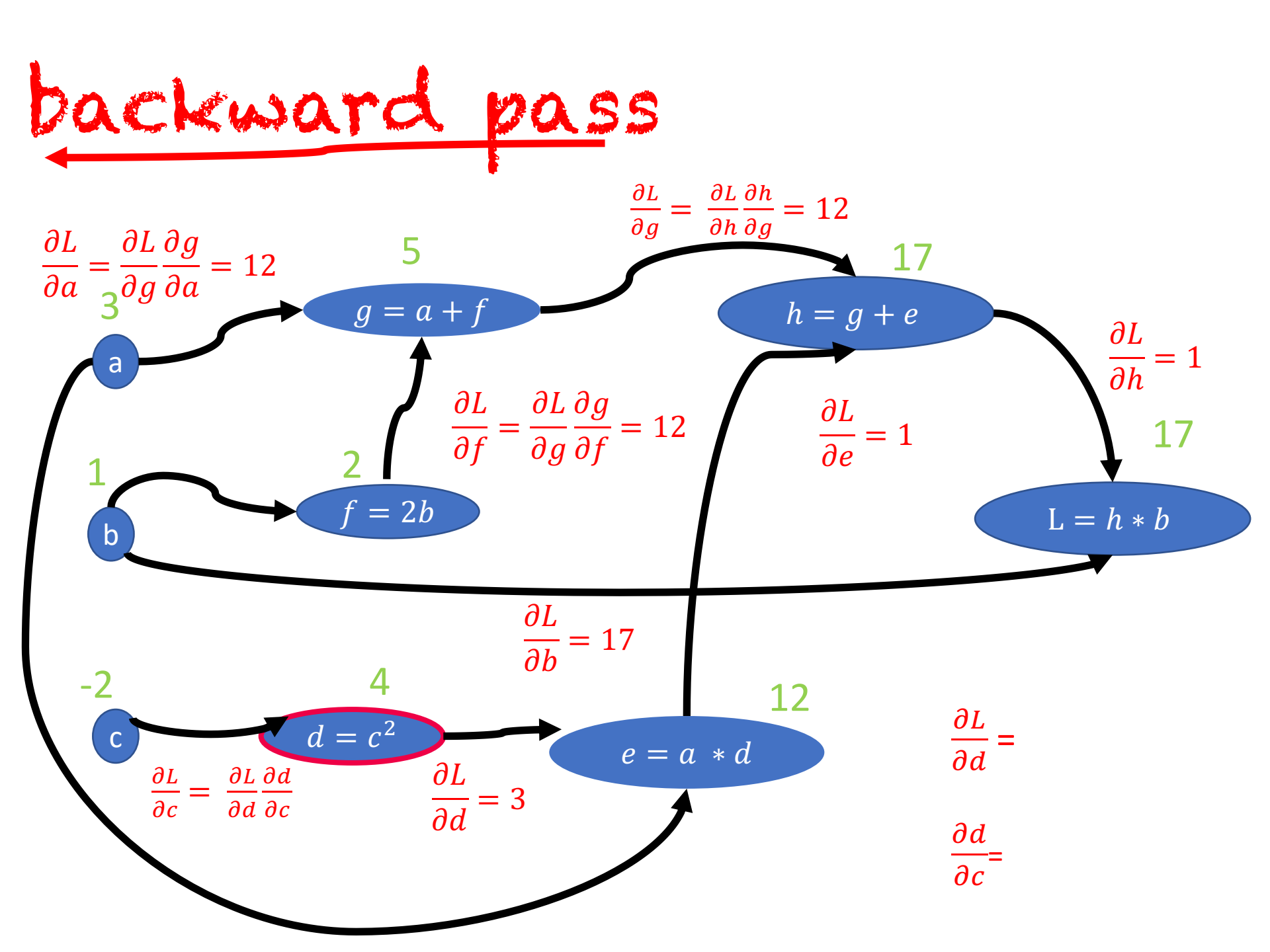

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

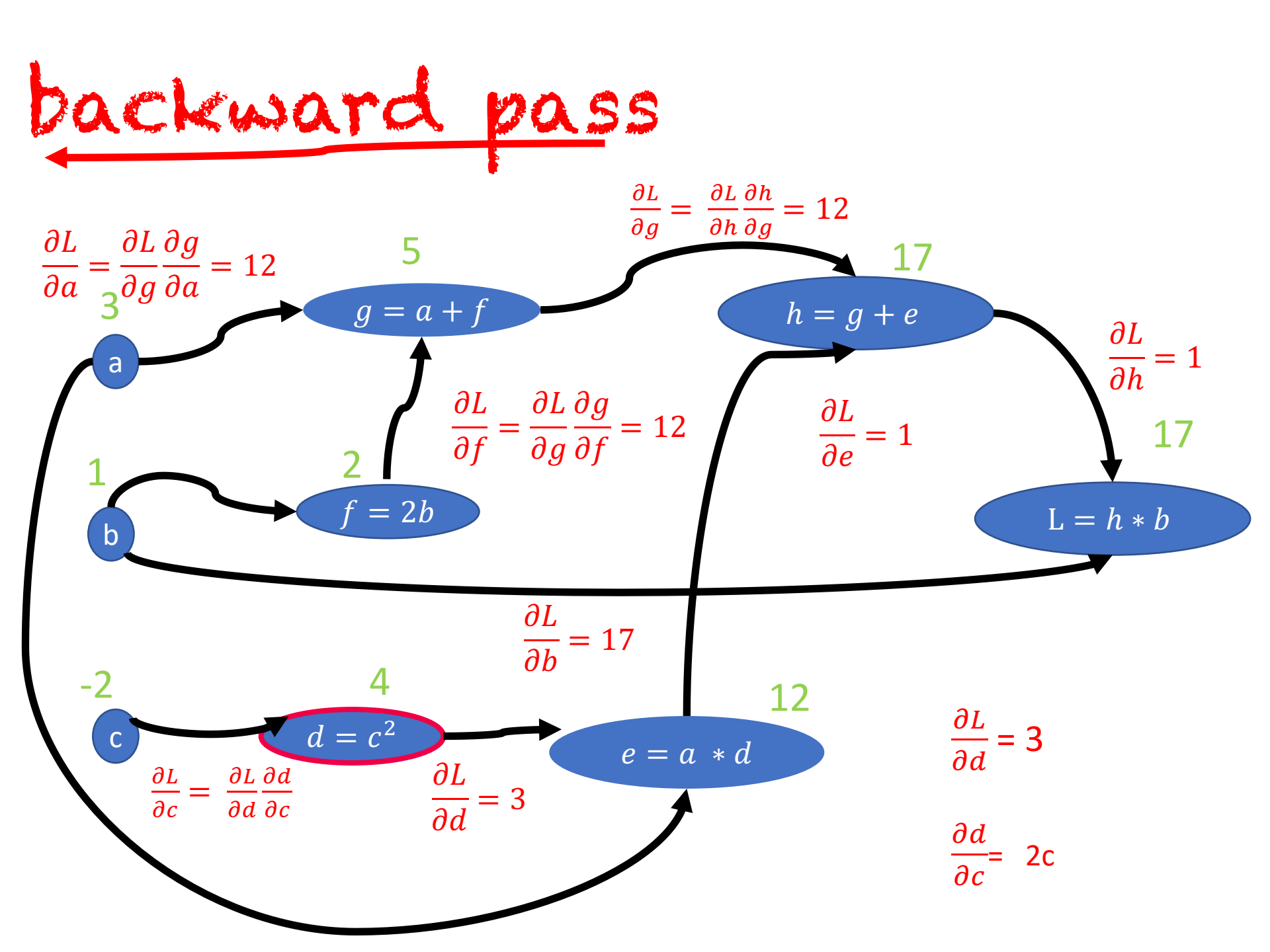

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial d} = \frac{\partial L}{\partial e}\frac{\partial e}{\partial d}$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

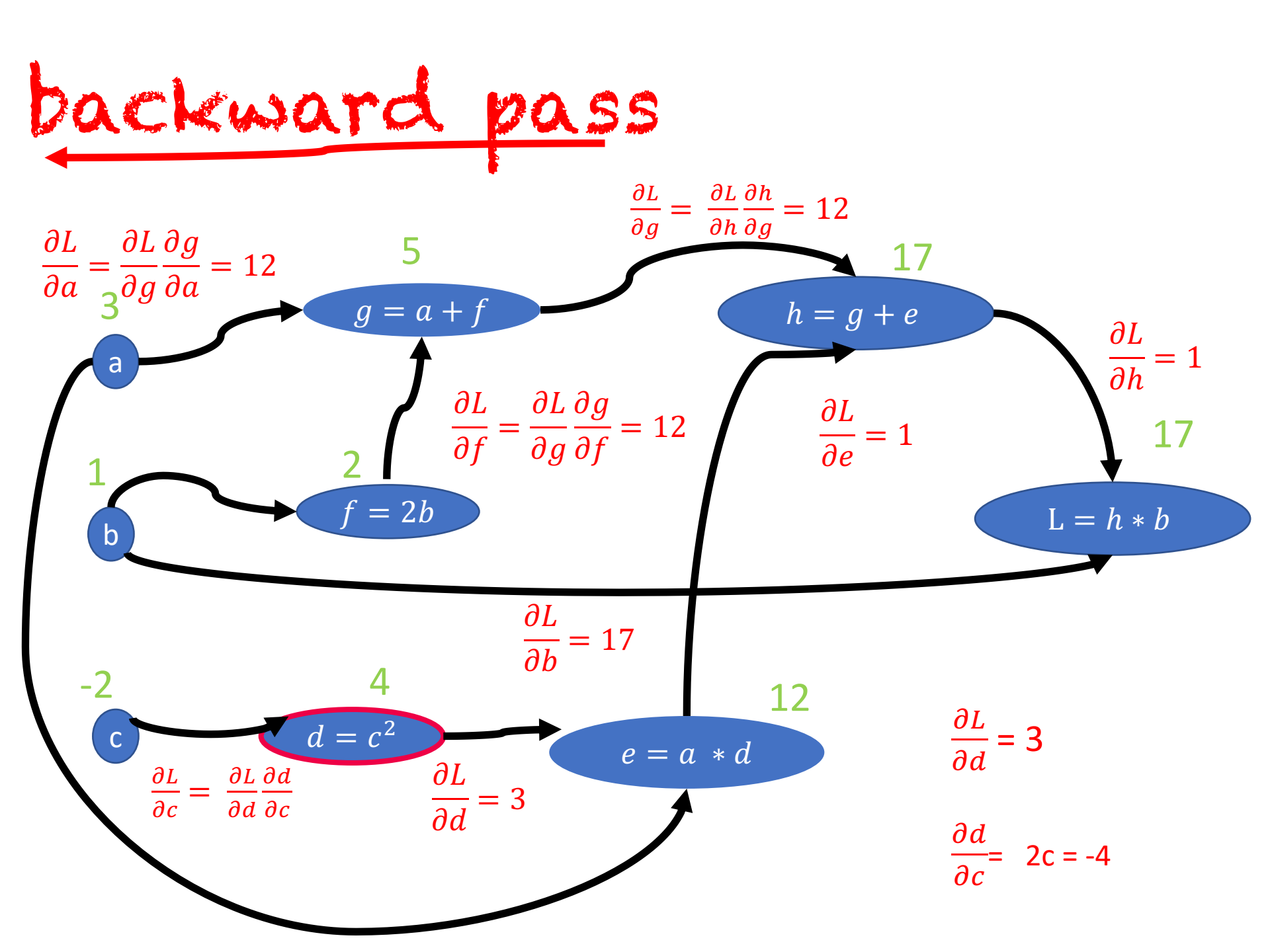

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial d} = \frac{\partial L}{\partial e}\frac{\partial e}{\partial d}$
$\frac{\partial L}{\partial d} = 1 * a$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

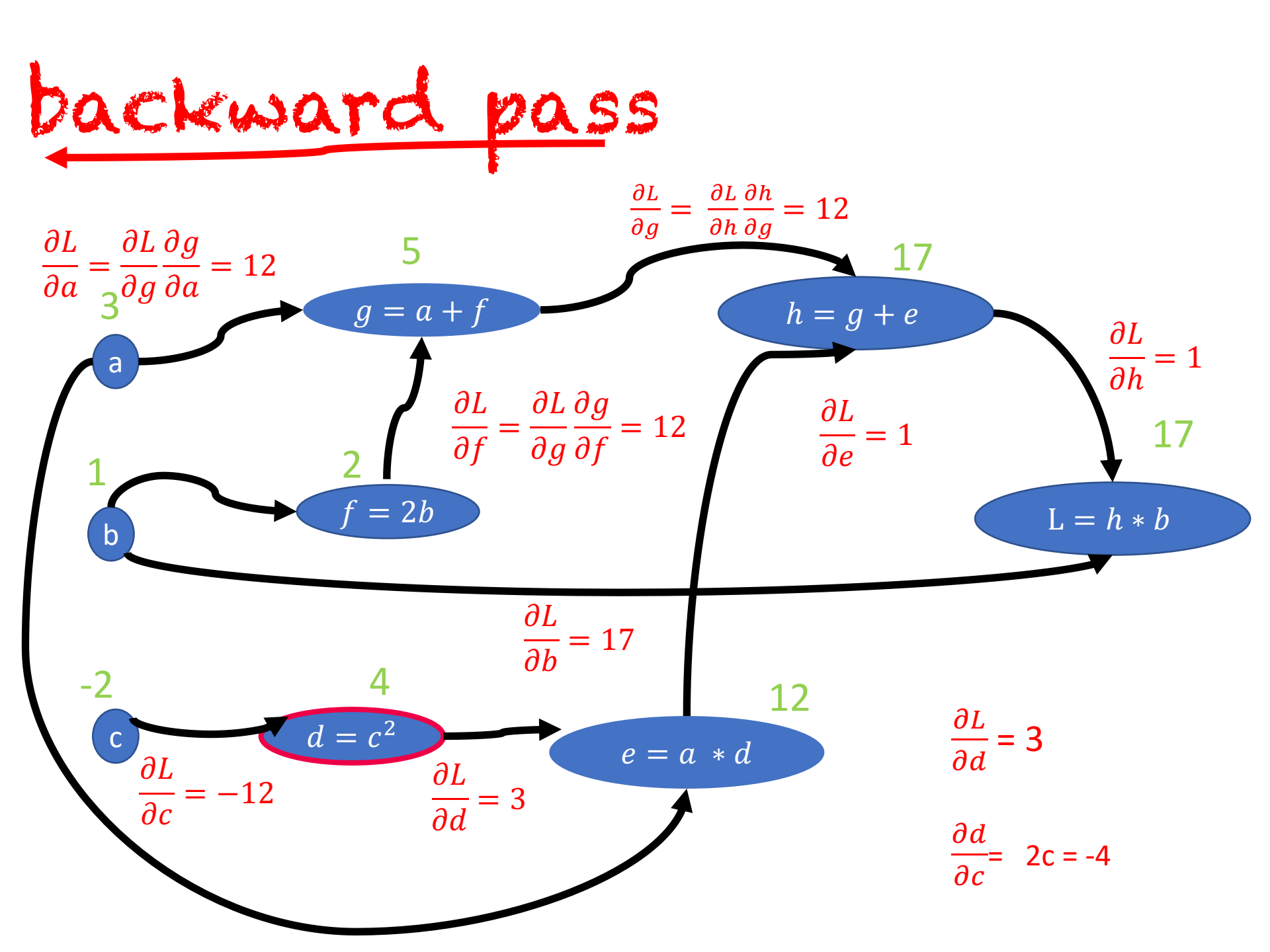

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial d} = 3$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$


backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = \frac{\partial L}{\partial d}\frac{\partial d}{\partial c}$
$\frac{\partial L}{\partial d} = 3$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$


backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = \frac{\partial L}{\partial d}\frac{\partial d}{\partial c}$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial L}{\partial d} =$
$\frac{\partial d}{\partial c} =$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

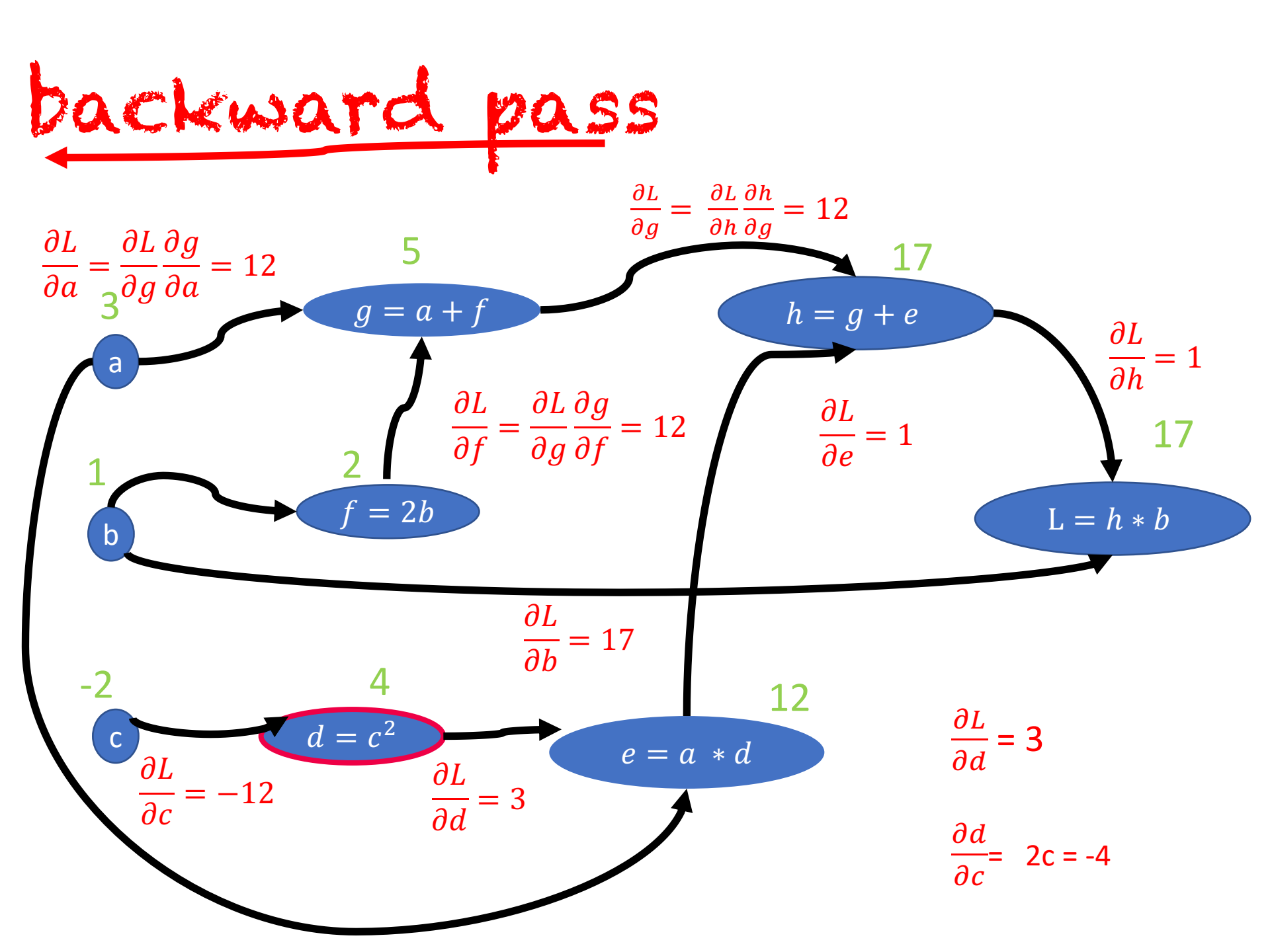

backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = \frac{\partial L}{\partial d}\frac{\partial d}{\partial c}$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial d}{\partial c} = 2c$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$


backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = \frac{\partial L}{\partial d}\frac{\partial d}{\partial c}$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial d}{\partial c} = 2c = -4$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$


backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = -12$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial d}{\partial c} = 2c = -4$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

Backwards differentiation in computation graphs

- The importance of the computation graph comes from the backward pass
- This is used to compute the derivatives that we'll need for the weight update.
- For training, we need the derivative of the loss with respect to weights in early layers of the network
- But loss is computed only at the very end of the network!
- Solution: **backward differentiation**
- Given a computation graph and the derivatives of all the functions in it we can automatically compute the derivative of the loss with respect to these early weights.


backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = -12$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial d}{\partial c} = 2c = -4$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

Exploding gradient

The gradient can accumulate, becoming very big

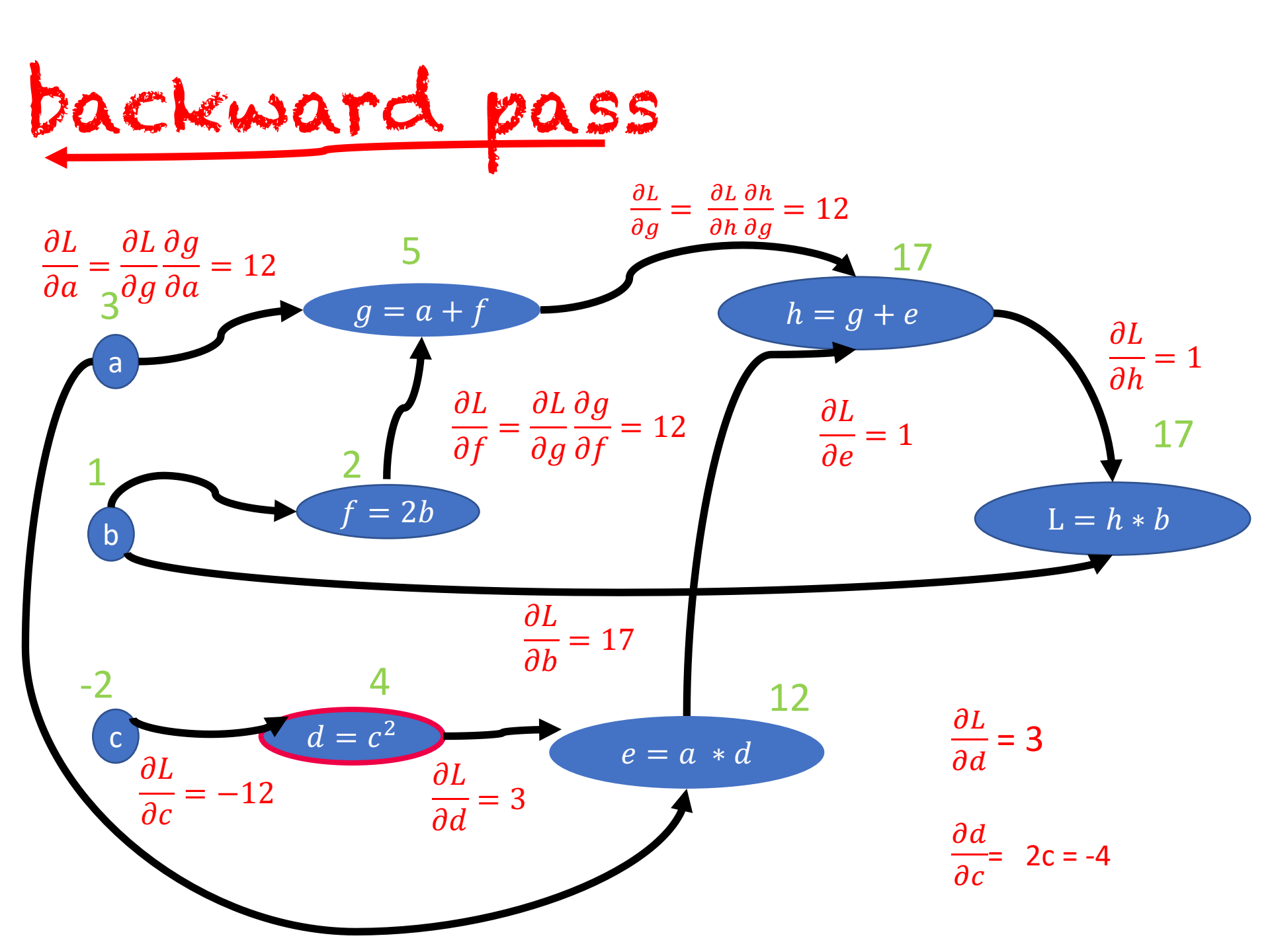
Issues:

- might move our weights too much
- result in Nan

Solution:

Clipping

- Maximum value for gradients
- Can be dynamic


backward pass
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial a} = 12$
$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial h}\frac{\partial h}{\partial g} = 12$
$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial f} = 12$
$\frac{\partial L}{\partial e} = 1$
$\frac{\partial L}{\partial h} = 1$
$\frac{\partial L}{\partial b} = 17$
$\frac{\partial L}{\partial c} = -12$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial L}{\partial d} = 3$
$\frac{\partial d}{\partial c} = 2c = -4$
3
5
17
1
2
17
-2
4
12
a
b
c
$g = a + f$
$h = g + e$
$f = 2b$
$L = h * b$
$d = c^2$
$e = a * d$

Vanishing gradient

The gradient become 0

Issues:

wont be able to update weights (because 0 gets passed all the way back)

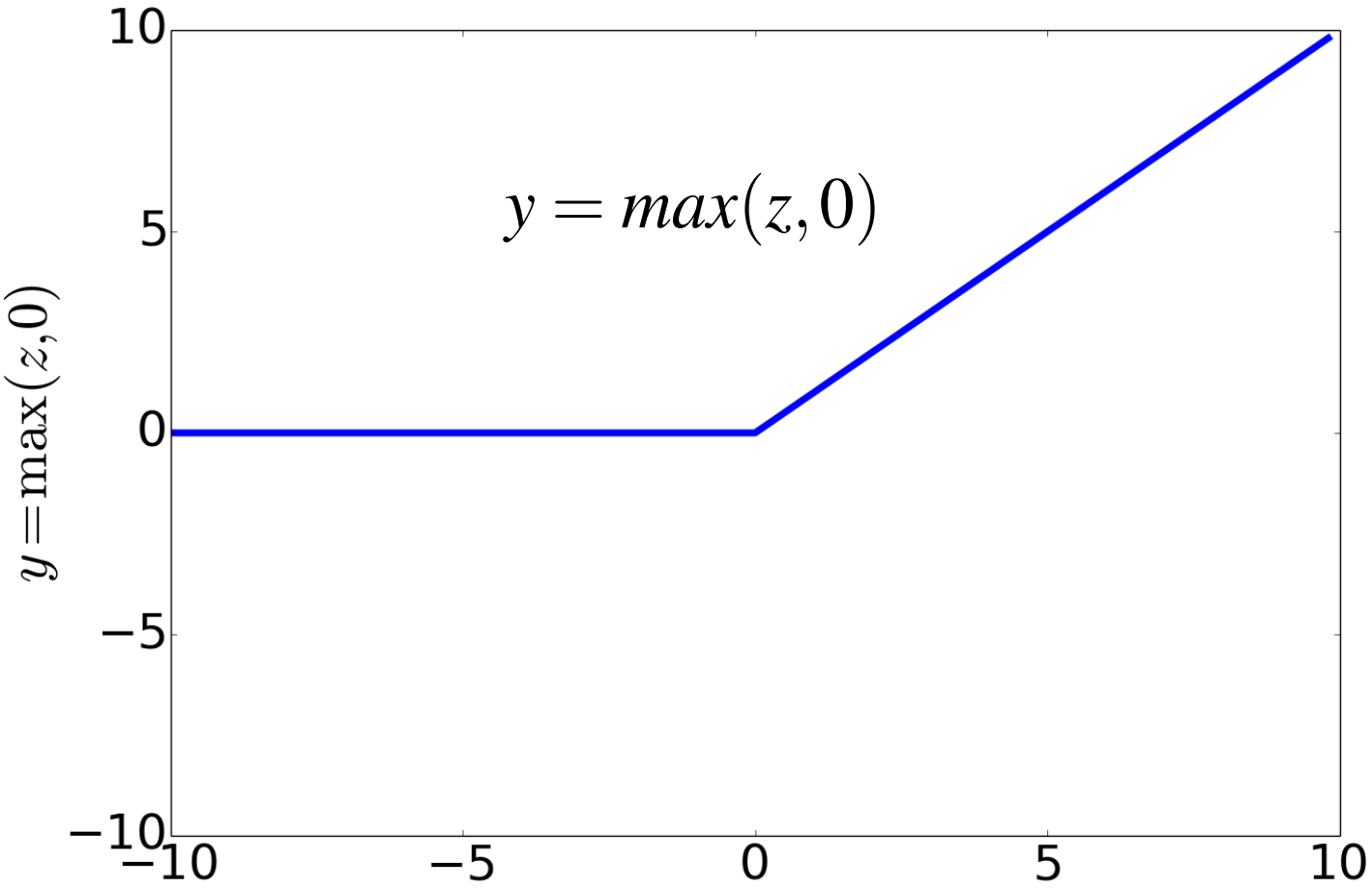
stuck in a local optima

Solution:

ReLU activation function

$z = \max(0, z)$

ReLU



One node view

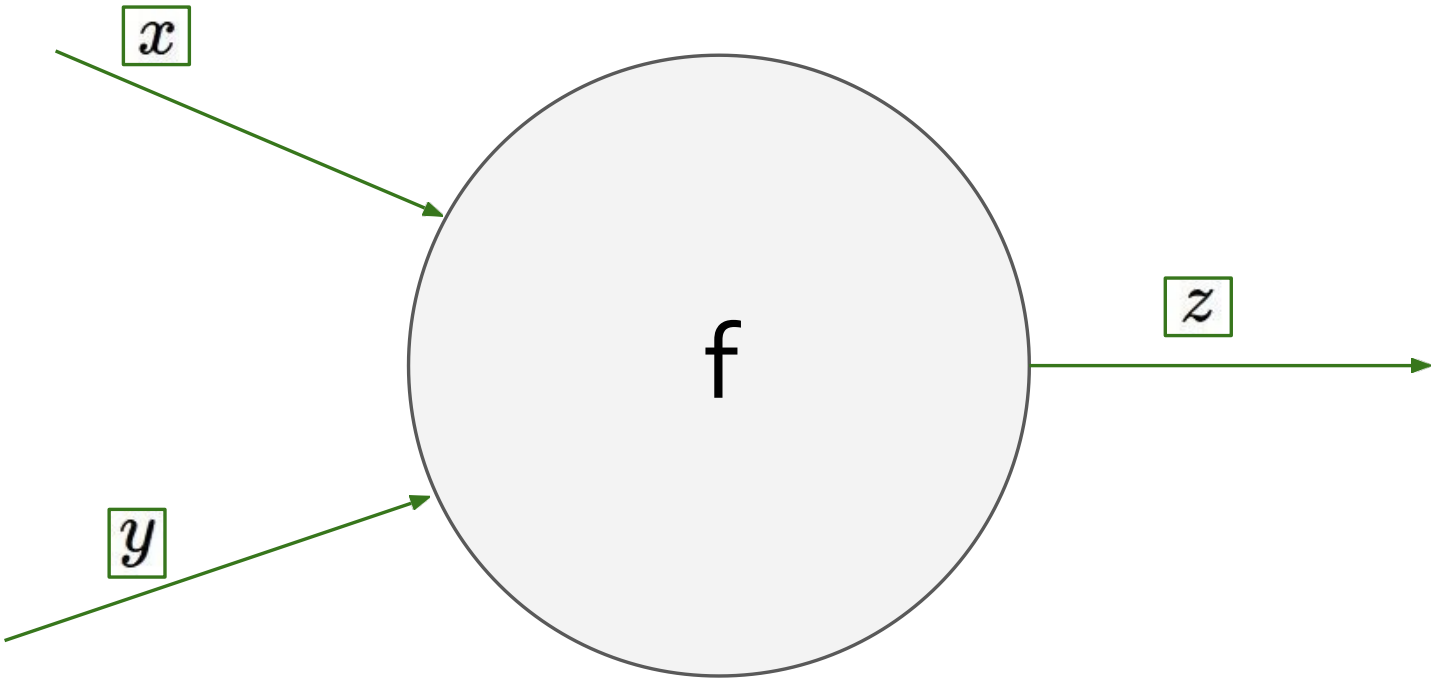


Figure from Andrej Karpathy

Dead neuron

In forward pass, output of a node w/ ReLU activation often will be 0

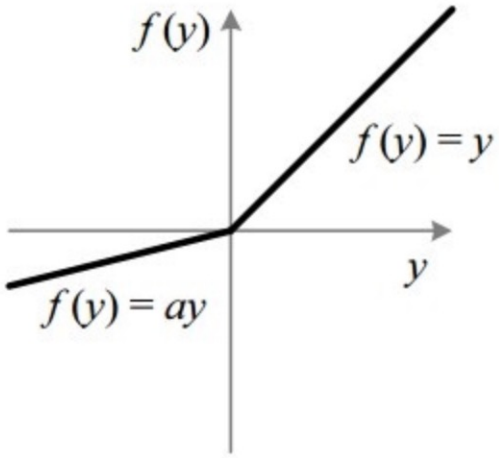
Issues:

wont pass information from one node to the next

lots of useless nodes

Solution:

Leaky ReLU activation function



Outline

Updating weights in LR

Computation Graph

Backpropagation

Issues when training NNs

Pytorch

Deep Averaging Neural Network

Pytorch

Torch: Facebook's deep learning framework

Originally written in Lua (C backend)

Optimized to run computations on GPU

Mature, industry-supported framework

Defining a model

```
import torch
from torch import nn


class LogisticRegression(nn.Module):
    def __init__(self, input_size, num_classes):
        super(LogisticRegression, self).__init__()
        self.linear = nn.Linear(input_size, num_classes)

    def forward(self, x):
        out = self.linear(x)
        return out
```

nn.Module

Base class for all neural network modules.

Creates a computation graph

Define the model in `__init__`

Specify how to make predictions in `forward`

- If only use built-in modules, no need to implement backprop

Defining a model

```
import torch
from torch import nn

class FNN(nn.Module):
    def __init__(self, input_size, hidden_size, num_classes):
        super(FNN, self).__init__()
        self.input_size = input_size
        self.l1 = nn.Linear(input_size, hidden_size)
        self.relu = nn.ReLU()
        self.l2 = nn.Linear(hidden_size, num_classes)

    def forward(self, x):
        out = self.l1(x)
        out = self.relu(out)
        out = self.l2(out)
        # no activation and no softmax at the end
        return out
```

Train a model

Define:

- Loss function
- Learning algorithm (e.g. SGD)
- Learning rate
- Number of epochs

```
num_epochs = 100
learning_rate = 0.003
optimizer = torch.optim.SGD(model.parameters(), lr=learning_rate)
loss_fn = nn.CrossEntropyLoss()
```

Train a model

In each iteration:

- Make a prediction
- Compute the loss
- Autograd (Automatic differentiation), backprop
- Update the weights

```
optimizer.zero_grad()
prediction = model(X[i])
loss_val = loss_fn(prediction, labels[0][i])
loss_val.backward()
optimizer.step()
```

Train a model

```
# Training the Model
for epoch in range(num_epochs):
  num_correct = 0
  for i in range(100):
    optimizer.zero_grad()
    prediction = model(X[i])
    loss_val = loss_fn(prediction, labels[0][i])
    loss_val.backward()
    optimizer.step()

  print(f"loss at epoch {epoch}: {loss_val}")
  print(f"accuracy at epoch {epoch}: {num_correct / 100}")
```

FFN's issues

Fixed input size

Solutions:

1. Create a fixed length representation
2. Recurrent Neural Networks