Module 4.4: Backpropagation (Intuition)

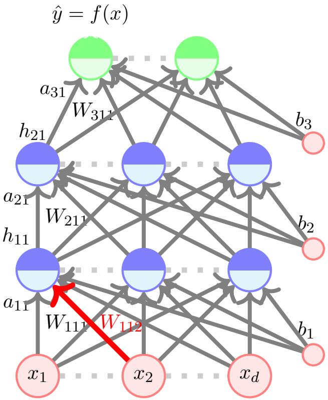
We need to answer two questions

- How to choose the loss function $\mathscr{L}(\theta)$?
- How to compute $\nabla\theta$ which is composed of $\nabla W_1, \nabla W_2, ..., \nabla W_{L-1} \in \mathbb{R}^{n\times n}, \nabla W_L \in \mathbb{R}^{n\times k}$,
  $\nabla b_1, \nabla b_2, ..., \nabla b_{L-1} \in \mathbb{R}^n$ and $\nabla b_L \in \mathbb{R}^k$?

We need to answer two questions

- How to choose the loss function $\mathscr{L}(\theta)$?
- How to compute $\nabla\theta$ which is composed of $\nabla W_1, \nabla W_2, ..., \nabla W_{L-1} \in \mathbb{R}^{n \times n}, \nabla W_L \in \mathbb{R}^{n \times k}$,
  $\nabla b_1, \nabla b_2, ..., \nabla b_{L-1} \in \mathbb{R}^{n}$ and $\nabla b_L \in \mathbb{R}^{k}$?

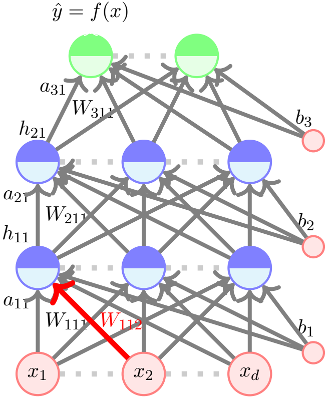
- Let us focus on this one weight (W_{112}).



Algorithm: gradient descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
$Initialize \quad \theta_0$;
while $t{+}{+} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

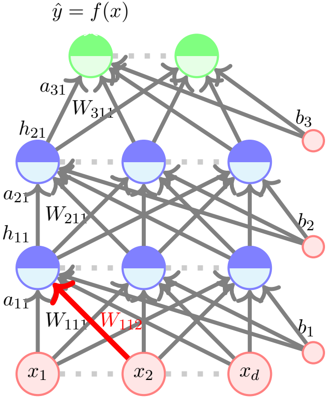
- Let us focus on this one weight (W_{112}).
- To learn this weight using SGD we need a formula for $\frac{\partial \mathscr{L}(\theta)}{\partial W_{112}}$.



Algorithm: gradient descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
$Initialize \quad \theta_0$;
while $t{++} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

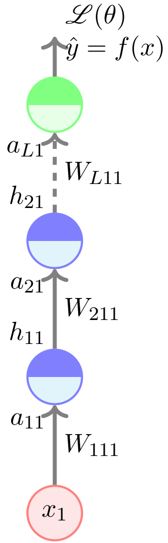
- Let us focus on this one weight (W_{112}).
- To learn this weight using SGD we need a formula for $\frac{\partial \mathscr{L}(\theta)}{\partial W_{112}}$.
- We will see how to calculate this.



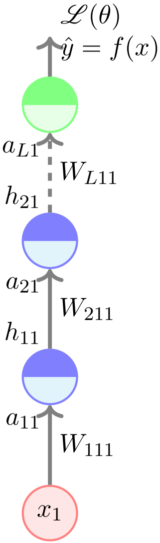
Algorithm: gradient descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
Initialize θ_0;
while $t{+}{+} < max\_iterations$ **do**
  $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \theta_t$;
end

- First let us take the simple case when we have a deep but thin network.

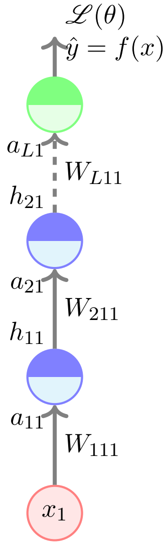


- First let us take the simple case when we have a deep but thin network.
- In this case it is easy to find the derivative by chain rule.



- First let us take the simple case when we have a deep but thin network.
- In this case it is easy to find the derivative by chain rule.

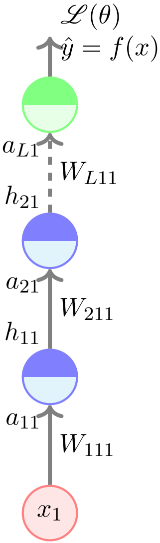
$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$



- First let us take the simple case when we have a deep but thin network.
- In this case it is easy to find the derivative by chain rule.

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}} \qquad \text{(just compressing the chain rule)}$$

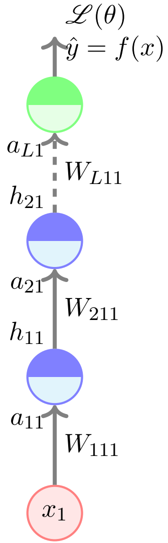


- First let us take the simple case when we have a deep but thin network.
- In this case it is easy to find the derivative by chain rule.

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}} \qquad \text{(just compressing the chain rule)}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{211}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{21}} \frac{\partial h_{21}}{\partial W_{211}}$$



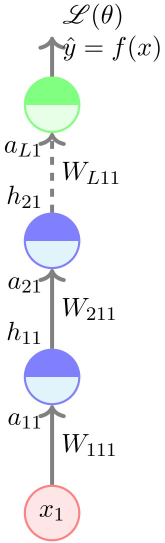
- First let us take the simple case when we have a deep but thin network.
- In this case it is easy to find the derivative by chain rule.

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_{L11}} \frac{\partial a_{L11}}{\partial h_{21}} \frac{\partial h_{21}}{\partial a_{21}} \frac{\partial a_{21}}{\partial h_{11}} \frac{\partial h_{11}}{\partial a_{11}} \frac{\partial a_{11}}{\partial W_{111}}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{11}} \frac{\partial h_{11}}{\partial W_{111}} \quad \text{(just compressing the chain rule)}$$

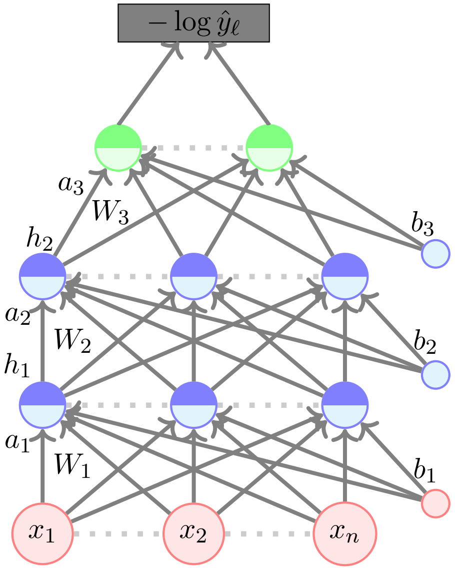
$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{211}} = \frac{\partial \mathscr{L}(\theta)}{\partial h_{21}} \frac{\partial h_{21}}{\partial W_{211}}$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial W_{L11}} = \frac{\partial \mathscr{L}(\theta)}{\partial a_{L1}} \frac{\partial a_{L1}}{\partial W_{L11}}$$

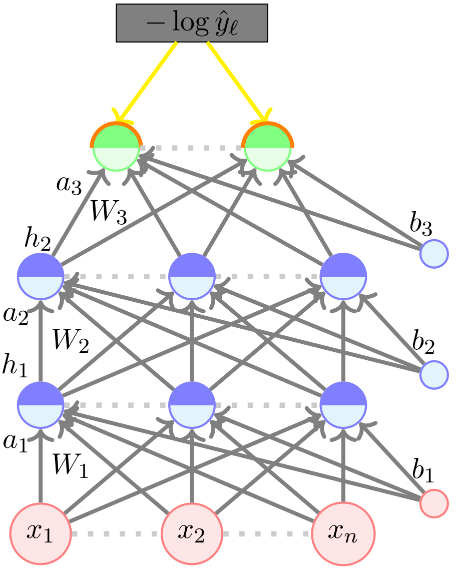


Let us see an intuitive explanation of backpropagation before we get into the mathematical details

- We get a certain loss at the output and we try to figure out who is responsible for this loss

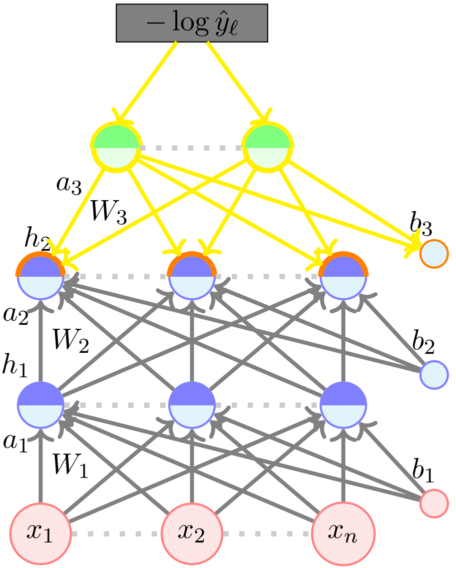


- We get a certain loss at the output and we try to figure out who is responsible for this loss
- So, we talk to the output layer and say “Hey! You are not producing the desired output, better take responsibility”.

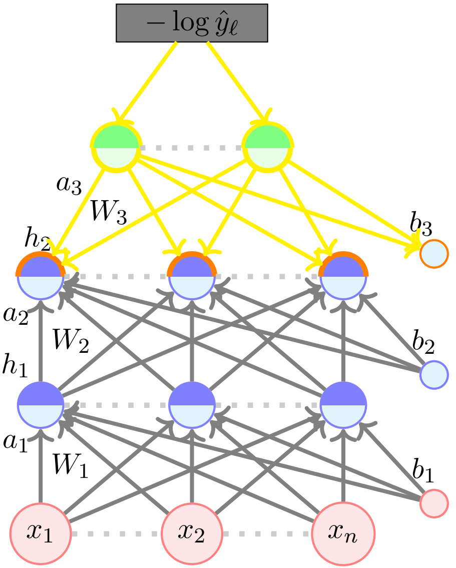


- We get a certain loss at the output and we try to figure out who is responsible for this loss
- So, we talk to the output layer and say "Hey! You are not producing the desired output, better take responsibility".
- The output layer says "Well, I take responsibility for my part but please understand that I am only as the good as the hidden layer and weights below me". After all ...

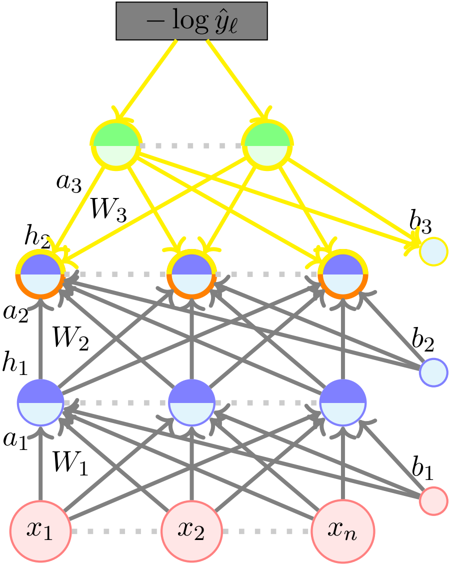
$$f(x) = \hat{y} = O(W_L h_{L-1} + b_L)$$



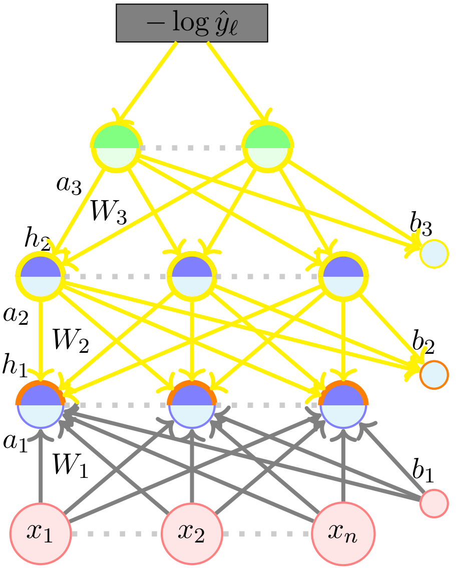
- So, we talk to W_L, b_L and h_L and ask them "What is wrong with you?"



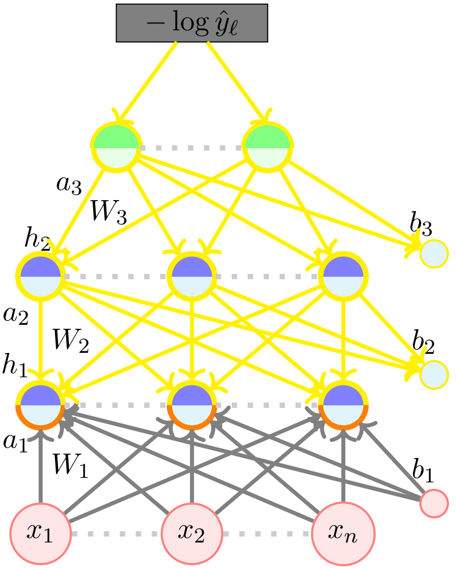
- So, we talk to W_L, b_L and h_L and ask them "What is wrong with you?"
- W_L and b_L take full responsibility but h_L says "Well, please understand that I am only as good as the pre-activation layer"



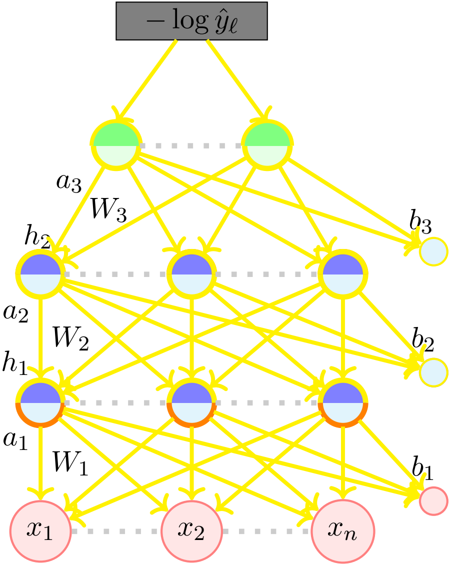
- So, we talk to W_L, b_L and h_L and ask them "What is wrong with you?"
- W_L and b_L take full responsibility but h_L says "Well, please understand that I am only as good as the pre-activation layer"
- The pre-activation layer in turn says that I am only as good as the hidden layer and weights below me.



- So, we talk to W_L, b_L and h_L and ask them "What is wrong with you?"
- W_L and b_L take full responsibility but h_L says "Well, please understand that I am only as good as the pre-activation layer"
- The pre-activation layer in turn says that I am only as good as the hidden layer and weights below me.
- We continue in this manner and realize that the responsibility lies with all the weights and biases (i.e. all the parameters of the model)

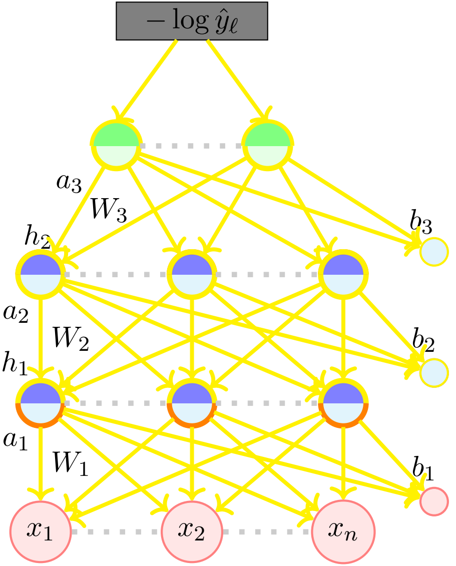


- So, we talk to W_L, b_L and h_L and ask them "What is wrong with you?"
- W_L and b_L take full responsibility but h_L says "Well, please understand that I am only as good as the pre-activation layer"
- The pre-activation layer in turn says that I am only as good as the hidden layer and weights below me.
- We continue in this manner and realize that the responsibility lies with all the weights and biases (i.e. all the parameters of the model)
- But instead of talking to them directly, it is easier to talk to them through the hidden layers and output layers (and this is exactly what the chain rule allows us to do)



- So, we talk to W_L, b_L and h_L and ask them "What is wrong with you?"
- W_L and b_L take full responsibility but h_L says "Well, please understand that I am only as good as the pre-activation layer"
- The pre-activation layer in turn says that I am only as good as the hidden layer and weights below me.
- We continue in this manner and realize that the responsibility lies with all the weights and biases (i.e. all the parameters of the model)
- But instead of talking to them directly, it is easier to talk to them through the hidden layers and output layers (and this is exactly what the chain rule allows us to do)

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\substack{\text{Talk to the} \\ \text{weight directly}}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\substack{\text{Talk to the} \\ \text{output layer}}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\substack{\text{Talk to the} \\ \text{previous hidden} \\ \text{layer}}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\substack{\text{Talk to the} \\ \text{previous} \\ \text{hidden layer}}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\substack{\text{and now} \\ \text{talk to} \\ \text{the} \\ \text{weights}}}$$



$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\substack{\text{Talk to the}\\ \text{weight directly}}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\substack{\text{Talk to the}\\ \text{output layer}}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\substack{\text{Talk to the}\\ \text{previous hidden}\\ \text{layer}}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\substack{\text{Talk to the}\\ \text{previous}\\ \text{hidden layer}}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\substack{\text{and now}\\ \text{talk to}\\ \text{the}\\ \text{weights}}}$$

Quantities of interest (roadmap for the remaining part):

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\text{Talk to the weight directly}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\text{Talk to the output layer}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\text{Talk to the previous hidden layer}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\text{and now talk to the weights}}$$

Quantities of interest (roadmap for the remaining part):

- Gradient w.r.t. output units

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\substack{\text{Talk to the}\\ \text{weight directly}}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial a_3}}_{\substack{\text{Talk to the}\\ \text{output layer}}} \underbrace{\frac{\partial a_3}{\partial h_2}\frac{\partial h_2}{\partial a_2}}_{\substack{\text{Talk to the}\\ \text{previous hidden}\\ \text{layer}}} \underbrace{\frac{\partial a_2}{\partial h_1}\frac{\partial h_1}{\partial a_1}}_{\substack{\text{Talk to the}\\ \text{previous}\\ \text{hidden layer}}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\substack{\text{and now}\\ \text{talk to}\\ \text{the}\\ \text{weights}}}$$

Quantities of interest (roadmap for the remaining part):

- Gradient w.r.t. output units
- Gradient w.r.t. hidden units

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\substack{\text{Talk to the}\\ \text{weight directly}}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\substack{\text{Talk to the}\\ \text{output layer}}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\substack{\text{Talk to the}\\ \text{previous hidden}\\ \text{layer}}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\substack{\text{Talk to the}\\ \text{previous}\\ \text{hidden layer}}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\substack{\text{and now}\\ \text{talk to}\\ \text{the}\\ \text{weights}}}$$

Quantities of interest (roadmap for the remaining part):

- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\substack{\text{Talk to the} \\ \text{weight directly}}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\substack{\text{Talk to the} \\ \text{output layer}}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\substack{\text{Talk to the} \\ \text{previous hidden} \\ \text{layer}}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\substack{\text{Talk to the} \\ \text{previous} \\ \text{hidden layer}}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\substack{\text{and now} \\ \text{talk to} \\ \text{the} \\ \text{weights}}}$$

Quantities of interest (roadmap for the remaining part):

- Gradient w.r.t. output units
- Gradient w.r.t. hidden units
- Gradient w.r.t. weights and biases

$$\underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial W_{111}}}_{\substack{\text{Talk to the}\\\text{weight directly}}} = \underbrace{\frac{\partial \mathscr{L}(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_3}}_{\substack{\text{Talk to the}\\\text{output layer}}} \underbrace{\frac{\partial a_3}{\partial h_2} \frac{\partial h_2}{\partial a_2}}_{\substack{\text{Talk to the}\\\text{previous hidden}\\\text{layer}}} \underbrace{\frac{\partial a_2}{\partial h_1} \frac{\partial h_1}{\partial a_1}}_{\substack{\text{Talk to the}\\\text{previous}\\\text{hidden layer}}} \underbrace{\frac{\partial a_1}{\partial W_{111}}}_{\substack{\text{and now}\\\text{talk to}\\\text{the}\\\text{weights}}}$$

- Our focus is on *Cross entropy loss* and *Softmax* output.