Exploding and vanishing gradients
Problem statement

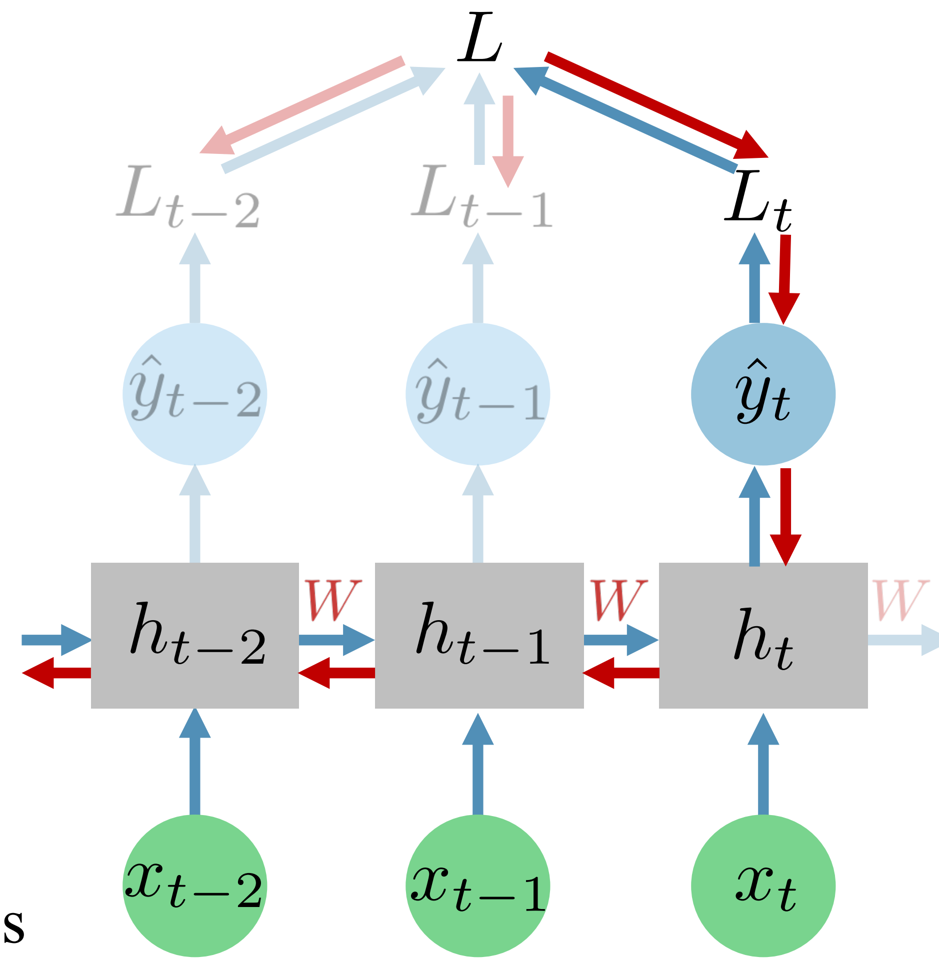
Previously on this week: BPTT

To train an RNN we need to backpropagate through layers and time

$$\frac{\partial L}{\partial W} = \sum_{i=0}^{T} \frac{\partial L_i}{\partial W}$$

$$\frac{\partial L_t}{\partial W} \propto \sum_{k=0}^{t} \boxed{\left(\prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}}$$

Contribution of a state at time step ***k*** to the gradient of the loss at time step ***t***



Let’s look at the gradient

$$\frac{\partial L_t}{\partial W} \propto \underbrace{\sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}}_{\text{These are jacobian matrices.}}$$

The more steps between the time moments *k* and *t*, the more elements are in this product

⬇

Values of these Jacobian matrices have particularly severe impact on the contributions from faraway steps

Let's look at the gradient

$$\frac{\partial L_t}{\partial W} \propto \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}$$

Let's suppose for a moment that h_i is a scalar and consequently $\frac{\partial h_i}{\partial h_{i-1}}$ is also a scalar

$\left| \frac{\partial h_i}{\partial h_{i-1}} \right| < 1$ ⇒ The product goes to 0 exponentially fast

$\left| \frac{\partial h_i}{\partial h_{i-1}} \right| > 1$ ⇒ The product goes to infinity exponentially fast

Let's look at the gradient

$$\frac{\partial L_t}{\partial W} \propto \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}$$

Let's suppose for a moment that h_i is a scalar and consequently $\frac{\partial h_i}{\partial h_{i-1}}$ is also a scalar

$$\left| \frac{\partial h_i}{\partial h_{i-1}} \right| < 1 \quad \Rightarrow$$

Vanishing gradients

- contributions from faraway steps vanish and don't affect the training
- difficult to learn long-range dependencies

Let's look at the gradient

$$\frac{\partial L_t}{\partial W} \propto \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}$$

Let's suppose for a moment that h_i is a scalar and consequently $\frac{\partial h_i}{\partial h_{i-1}}$ is also a scalar

Exploding gradients

$$\left| \frac{\partial h_i}{\partial h_{i-1}} \right| > 1 \Rightarrow$$

- make the learning process unstable
- gradient could even become a NaN

Let's look at the gradient

$$\frac{\partial L_t}{\partial W} \propto \sum_{k=0}^{t} \left(\prod_{i=k+1}^{t} \frac{\partial h_i}{\partial h_{i-1}} \right) \frac{\partial h_k}{\partial W}$$

The same is true for matrices but with the spectral matrix norm instead of the absolute value:

The spectral norm is the maximum singular value of a matrix (presumably, in its singular value decomposition). Intuitively, you can think of it as the maximum 'scale', by which the matrix can 'STRETCH' a vector.
See http://ee263.stanford.edu/lectures/svd-v2.pdf for more details.

$$\left\| \frac{\partial h_i}{\partial h_{i-1}} \right\|_2 < 1 \quad \Rightarrow$$

The product goes to zero-norm matrix exponentially fast

$$\left\| \frac{\partial h_i}{\partial h_{i-1}} \right\|_2 > 1 \quad \Rightarrow$$

The product goes to a matrix of infinite norm exponentially fast

Is it really a problem in practice?

$$h_t = f_h(Vx_t + Wh_{t-1} + b_h) = f_h(pr_t)$$

$$\boxed{\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial h_t}{\partial pr_t}\frac{\partial pr_t}{\partial h_{t-1}} = diag(f_h'(pr_t)) \cdot \mathbf{?}}$$

diag(f'_h(pr_t)) basically means:

We have a diagonal matrix where each diagonal entry
is a partial derivative w.r.t pr_t

Is it really a problem in practice?

$$h_t = f_h(Vx_t + Wh_{t-1} + b_h) = f_h(pr_t)$$

d pr_t / d h_{t-1} = W. See equation above:

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial h_t}{\partial pr_t}\frac{\partial pr_t}{\partial h_{t-1}} = diag(f_h'(pr_t)) \cdot W$$

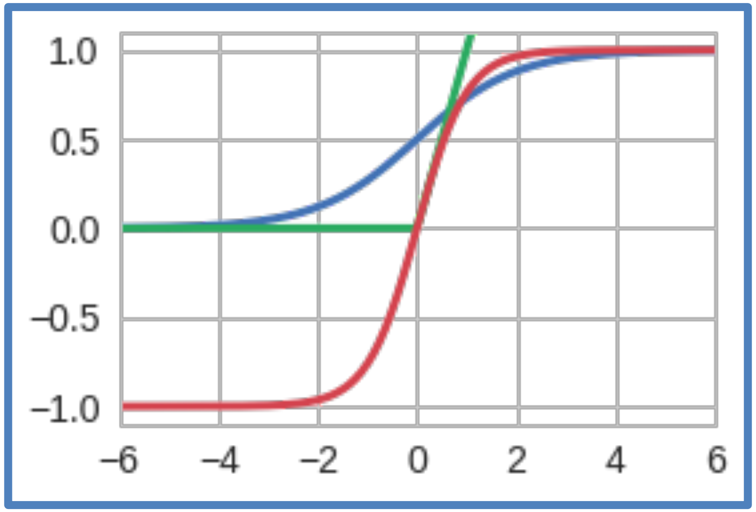
Is it really a problem in practice?

$$h_t = f_h(Vx_t + Wh_{t-1} + b_h) = f_h(pr_t)$$

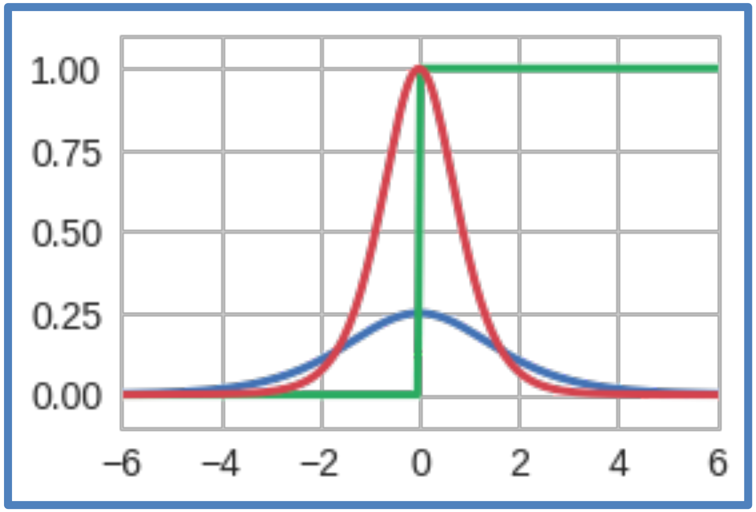
$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial h_t}{\partial pr_t}\frac{\partial pr_t}{\partial h_{t-1}} = \boxed{diag(f_h'(pr_t))} \cdot W$$

ReLU may still vanish if always negative.

sigmoid, tanh, ReLU



Derivatives



Vanishing gradients are very likely especially with sigmoid and tanh

Yes. This is because their gradients have limit to 0 on both -\infty and \infty.

Is it really a problem in practice?

$$h_t = f_h(Vx_t + Wh_{t-1} + b_h) = f_h(pr_t)$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial h_t}{\partial pr_t}\frac{\partial pr_t}{\partial h_{t-1}} = diag(f_h'(pr_t)) \cdot \boxed{W}$$

$||W||$ may be either **small** or **large**

Question: What is the use of the spectral norm?

Small $||W||$ could aggravate the vanishing gradient problem

Large $||W||$ could cause exploding gradients (especially with ReLU)

Summary

- In practice vanishing and exploding gradients are common for RNNs. These problems also occur in deep Feedforward NNs.
- Vanishing gradients make the learning of long-range dependencies very difficult.
- Exploding gradients make the learning process very unstable and may even crash it.

In the next video:

How to deal with these issues?