Vanishing Gradient Problem

Recurrent Neural Networks (RNNs)

Recurrent neural networks have hidden layers which repeat themselves in time. More simply, RNNs take the original inputs together with the output of the previous step together as the input of each computation step. See Figure 1. The equations (1) and (2) represent the model. Vanishing and exploding gradients are two very significant problems in Recurrent Neural Networks which lead to lack of learning.

$$\mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{u}_t, \theta) \tag{1}$$

$$\mathbf{x}_t = \mathbf{W}_{rec}\sigma(\mathbf{x}_{t-1}) + \mathbf{W}_{in}\mathbf{u}_t + \mathbf{b} \tag{2}$$

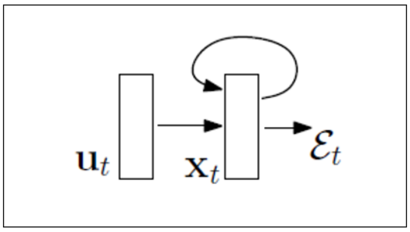


Figure 1

Introducing Vanishing and Exploding Gradients

Vanishing Gradient Problem refers to the behaviour, when long term components go exponentially fast to norm 0, making it impossible for the model to learn correlations between events.

Exploding Gradient Problem refers to the large increase in the norm of the gradient during training.

Back Propagation Through Time

See Figure 2 as a visualization of back propagation on the unrolled structure.

As this is a new type of neural network architecture, the computation of errors also differs. We calculate the errors at each time step and take their sum for the total error at that epoch (training step).

For back propagation, we need to take the derivative of the error with respect to θ, which represents the weights and biases. The equations (5), (6) and (7) show these derivatives. The last term inside the sum of equation (6) stands for immediate partial derivative which means we take x_{k-1} as constant with respect to θ. Lastly on equation (7), the middle term inside the sum of equation (6) is written as the product of *t-k* Jacobien matrices.

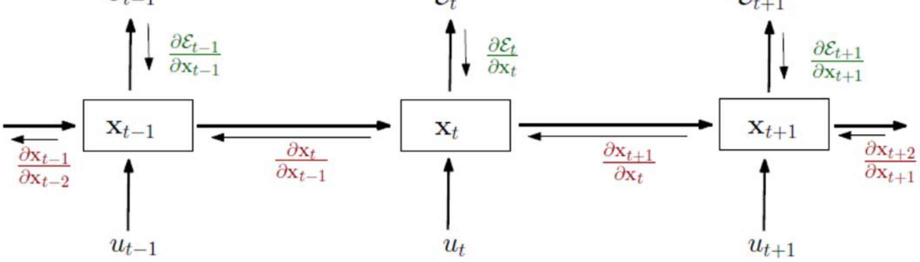

$\mathcal{E}_{t-1}$
$\frac{\partial \mathcal{E}_{t-1}}{\partial \mathbf{x}_{t-1}}$
$\mathcal{E}_t$
$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t}$
$\mathcal{E}_{t+1}$
$\frac{\partial \mathcal{E}_{t+1}}{\partial \mathbf{x}_{t+1}}$
$\mathbf{x}_{t-1}$
$\mathbf{x}_t$
$\mathbf{x}_{t+1}$
$\frac{\partial \mathbf{x}_{t-1}}{\partial \mathbf{x}_{t-2}}$
$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_{t-1}}$
$\frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t}$
$\frac{\partial \mathbf{x}_{t+2}}{\partial \mathbf{x}_{t+1}}$
u_{t-1}
u_t
u_{t+1}

Figure 2

$$\mathcal{E} = \textstyle\sum_{1\leq t\leq T} \mathcal{E}_t \tag{3}$$

$$\mathcal{E}_t = \mathcal{L}(\mathbf{x}_t) \tag{4}$$

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1\leq t\leq T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{5}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1\leq k\leq t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{6}$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t\geq i>k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t\geq i>k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{7}$$

Sufficient Conditions

It is sufficient for the largest eigenvalue of the recurrent weight matrix to be smaller than 1 for long term components to vanish (as $t\rightarrow\infty$) and necessary for it to be larger than 1 for gradients to explode. The information that is necessary for the proof and the condition we claim can be seen below. λ_1 stands for the largest eigenvalue of the recurrent weight matrix.

$|\sigma'(x)|$ is bounded by a value $\gamma \in \mathbb{R}$ and therefore $\|diag(\sigma'(\mathbf{x}_k))\| \leq \gamma$.

$$\prod_{t\geq i>k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t\geq i>k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1}))$$

$\frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k}$ is given by $\mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_k))$.

$$\lambda_1 < \frac{1}{\gamma}$$

With remembering equation (7), we can further prove the sufficient condition as follows.

$$\forall k, \left\| \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\| \leq \left\| \mathbf{W}_{rec}^T \right\| \left\| diag(\sigma'(\mathbf{x}_k)) \right\| < \frac{1}{\gamma}\gamma < 1 \tag{8}$$

$$\text{Let } \eta \in \mathbb{R} \text{ be such that } \forall k, \left\| \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\| \leq \eta < 1 \tag{9}$$

$$\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \left(\prod_{i=k}^{t-1} \frac{\partial \mathbf{x}_{i+1}}{\partial \mathbf{x}_i} \right) \leq \eta^{t-k} \frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \tag{10}$$

Geometrical Interpretation

To make it easier to understand, a system without inputs is assumed as in equation (11), with identity function as sigma and a bias of 0. It can be generalized from the geometrical interpretation that when gradients explode they explode along some direction **v,** equation (14).

$$x_t = w\sigma(x_{t-1}) + b \tag{11}$$

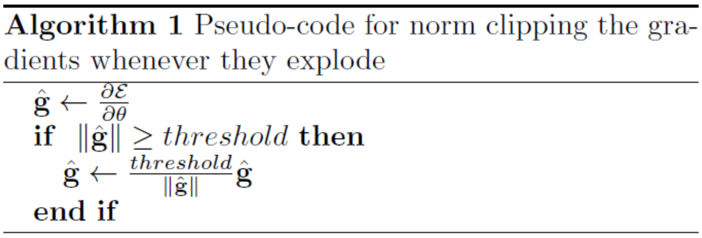
$$x_t = x_0 w^t \tag{12}$$

$$\frac{\partial x_t}{\partial w} = t x_0 w^{t-1} \text{ and } \frac{\partial^2 x_t}{\partial w^2} = t(t-1) x_0 w^{t-2} \tag{13}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta}\mathbf{v} \geq C\alpha^t \qquad C, \alpha \in \mathbb{R} \text{ and } \alpha > 1 \tag{14}$$

Solutions

The suggested approach for the exploding gradients is *Gradient Clipping*. In this method we have a threshold value and when the norm of the gradient goes over this threshold the algorithm rescales it. See Figure 3 for the pseudo-code.



Algorithm 1 Pseudo-code for norm clipping the gradients whenever they explode

```
$\hat{\mathbf{g}} \leftarrow \frac{\partial \mathcal{E}}{\partial \theta}$
if $\|\hat{\mathbf{g}}\| \geq threshold$ then
    $\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$
end if
```

Figure 3

The suggested approach for the vanishing gradients is *Regularization*. The regularizer omega (See equation (15)) will act on the error signals which preserve their norms as we travel back through time.

$$\Omega = \sum_k \Omega_k = \sum_k \left(\frac{\left\| \frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}} \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\|}{\left\| \frac{\partial \mathcal{E}}{\partial \mathbf{x}_{k+1}} \right\|} - 1 \right)^2 \tag{15}$$

References

[1] Pascanu, R., Mikolov, T., Bengio, Y., (2013). On the difficulty of training Recurrent Neural Networks