Gradients for an RNN

Carter N Brown

January 4, 2017

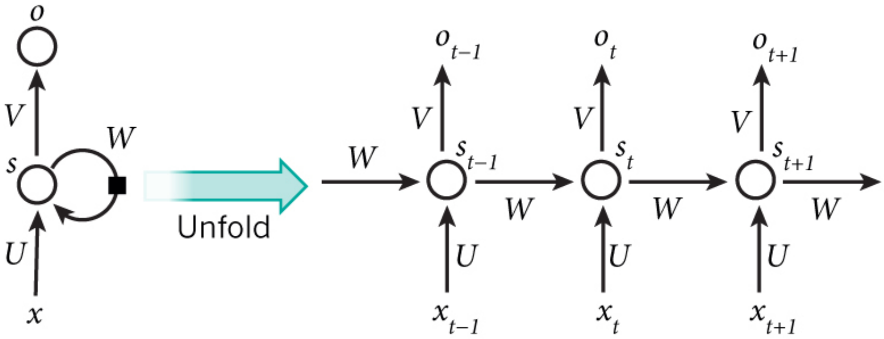
1 Overview

In this document we will go through the derivation of the gradient for an Recurrent Neural Network (RNN). The formalism and names for everything are consistent with WildML's RNN tutorial. The purpose is to walk through the math in the tutorial in greater detail.

1.1 RNN Recap

The RNN structure can be seen below (image from WildML):

Figure 1: RNN structure and its unfolding



The equations for s_t and o_t are:

$$s_t = \tanh(Ux_t + Ws_{t-1}), \tag{1a}$$

$$\hat{y}_t = \text{softmax}(Vs_t). \tag{1b}$$

Our loss function is:

$$L(y, \hat{y}) = -\frac{1}{N}\sum_t y_t \log \hat{y}_t. \tag{2}$$

For the sake of computational ease later, we define:

$$E_t = -y_t \log \hat{y}_t. \tag{3}$$

N.B. that the loss functions are dot products between the vectors y_t and element-wise logarithm of $\hat{y}_t$

1.2 Math Recap

The important math concepts here are Einstein Summation, chain rule, and matrix derivatives. For the summation notation, we won't be concerned with the dual basis, i.e. all indices will be on the bottom of the variable for ease. N.B. we will not denote vectors or matrices with either arrows or boldface.

Einstein summation notation is useful here to help manage the chain rule and matrix derivatives. For example, suppose we have a function $f(x, y)$ where $x, y \in \mathbb{R}^N$. Furthermore, suppose that x and y are functions of $r \in \mathbb{R}$, i.e. $x = x(r)$ and $y = y(r)$. Then,

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x_i}\frac{\partial x_i}{\partial r} + \frac{\partial f}{\partial y_j}\frac{\partial y_j}{\partial r} \tag{4}$$

where we sum over the dummy indices i and j. Here "dummy" means that they're only being summed over, i.e. they aren't a key part of the definition. An example of an index that isn't a dummy is m in the equation $v_m = T_{mn}u_n$ (while n is a dummy index).

A useful sanity checks is whether the left and right sides of the equation have the same free indices, i.e. indices that are not summed over. In our chain rule example (4) the left side has no indices, and the right side has no free indices. In our second example, with v_m, both the left and right have m as a free index and no others.

A nice rule of thumb for chain rule here comes with summed indices: For each pair, one index will appear in the numerator of a derivative and the other will appear in the denominator of a derivative.

Now, suppose we have a matrix $V \in \mathbb{R}^{N\times M}$ and a function $g(M)$, and we want $\frac{\partial g}{\partial V}$. Then,

$$\left(\frac{\partial g}{\partial V}\right)_{ij} = \frac{\partial g}{\partial V_{ij}}. \tag{5}$$

Now we can use our new summation notation to clarify definition (3) for our loss function as

$$E_t = -y_{t_i} \log \hat{y}_{t_i} \tag{6}$$

2 Gradient Calculations

The parameters of our RNN are U, V, and W, so we must compute the gradient of our loss function with respect to these matrices. This will be done in order of increasing difficulty.

2.1 V

The parameter V is present only in the function $\hat{y}$. Let $q_t = Vs_t$. Then,

$$\frac{\partial E_t}{\partial V_{ij}} = \frac{\partial E_t}{\partial \hat{y}_{t_k}}\frac{\partial \hat{y}_{t_k}}{\partial q_{t_l}}\frac{\partial q_{t_l}}{\partial V_{ij}}. \tag{7}$$

From our definition of E_t (6), we have that

$$\frac{\partial E_t}{\partial \hat{y}_{t_k}} = -\frac{y_{t_k}}{\hat{y}_{t_k}}. \tag{8}$$

Our function $\hat{y}$ is just the sigmoid function, so

$$\frac{\partial \hat{y}_{t_k}}{\partial q_{t_l}} = \begin{cases} -\sigma(q_t)_k\,\sigma(q_t)_l, & k \neq l \\ \sigma(q_t)_k\,(1-\sigma(q_t)_k)\,, & k = l \end{cases} \tag{9a}$$

$$= \begin{cases} -\hat{y}_{t_k}\hat{y}_{t_l}, & k \neq l \\ \hat{y}_{t_k}\,(1-\hat{y}_{t_k})\,, & k = l \end{cases}. \tag{9b}$$

Putting together (8) and (9b) gives us a sum over all values of k to obtain $\frac{\partial E_t}{\partial q_{t_l}}$:

$$-\frac{y_{t_l}}{\hat{y}_{t_l}}\left(\hat{y}_{t_l}\left(1-\hat{y}_{t_l}\right)+\sum_{k\neq l}\left(-\frac{y_{t_k}}{\hat{y}_{t_k}}\right)\left(-\hat{y}_{t_k}\hat{y}_{t_l}\right)\right)=-y_{t_l}+y_{t_l}\hat{y}_{t_l}+\sum_{k\neq l}y_{t_k}\hat{y}_{t_l} \tag{10a}$$

$$=-y_{t_l}+\hat{y}_{t_l}\sum_k y_{t_k}. \tag{10b}$$

And, if you'll recall that y_t are all one-hot vectors, then that sum is just equal to 1, so

$$\frac{\partial E_t}{\partial q_{t_l}}=\hat{y}_{t_l}-y_{t_l} \tag{11}$$

Lastly, $q_t = Vs_t$, so $q_{t_l} = V_{lm}s_{t_m}$. Then,

$$\frac{\partial q_{t_l}}{\partial V_{ij}}=\frac{\partial}{\partial V_{ij}}\left(V_{lm}s_{t_m}\right) \tag{12a}$$

$$=\delta_{il}\delta_{jm}s_{t_m} \tag{12b}$$

$$=\delta_{il}s_{t_j}. \tag{12c}$$

Now we combine (11) and (12c) to obtain:

$$\frac{\partial E_t}{\partial V_{ij}}=\left(\hat{y}_{t_i}-y_{t_i}\right)s_{t_j}, \tag{13}$$

which is recognizable as the outerproduct. Hence,

$$\frac{\partial E_t}{\partial V}=\left(\hat{y}_t-y_t\right)\otimes s_t, \tag{14}$$

where $\otimes$ is the outer product.

2.2 W

The parameter W appears in the argument for s_t, so we will have to check the gradient in both s_t and $\hat{y}_t$. We must also make note that $\hat{y}_t$ depends on W both directly and indirectly (through s_{t-1}). Let $z_t = Ux_t + Ws_{t-1}$. Then $s_t = \tanh(z_t)$.
At first it seems that by the chain rule we have:

$$\frac{\partial E_t}{\partial W_{ij}}=\frac{\partial E_t}{\partial \hat{y}_{t_k}}\frac{\partial \hat{y}_{t_k}}{\partial q_{t_l}}\frac{\partial q_{t_l}}{\partial s_{t_m}}\frac{\partial s_{t_m}}{\partial W_{ij}} \tag{15}$$

Note that of these four terms, we have already calculated the first two, and the third is simple:

$$\frac{\partial q_{t_l}}{\partial s_{t_m}}=\frac{\partial}{\partial s_{t_m}}\left(V_{lb}s_{t_b}\right) \tag{16a}$$

$$=V_{lb}\delta_{bm} \tag{16b}$$

$$=V_{lm}. \tag{16c}$$

The final term, however, requires us to notice that there is an implicit dependence of s_t on W_{ij} through s_{t-1} as well as a direct dependence. Hence, we have

$$\frac{\partial s_{t_m}}{\partial W_{ij}}\to\frac{\partial s_{t_m}}{\partial W_{ij}}+\frac{\partial s_{t_m}}{\partial s_{t-1_n}}\frac{\partial s_{t-1_n}}{\partial W_{ij}}. \tag{17}$$

But we can just apply this again to yield:

$$\frac{\partial s_{t_m}}{\partial W_{ij}} \rightarrow \frac{\partial s_{t_m}}{\partial W_{ij}} + \frac{\partial s_{t_m}}{\partial s_{t-1_n}} \frac{\partial s_{t-1_n}}{\partial W_{ij}} + \frac{\partial s_{t_m}}{\partial s_{t-1_n}} \frac{\partial s_{t-1_n}}{\partial s_{t-2_p}} \frac{\partial s_{t-2_p}}{\partial W_{ij}}. \tag{18}$$

This process continues until we reach s_{-1}, which was initialized to a vector of zeros. Notice that the last term in (18) collapses to $\frac{\partial s_{t_m}}{\partial s_{t-2_n}} \frac{\partial s_{t-2n}}{\partial W_{ij}}$ and we can turn the first term into $\frac{\partial s_{t_m}}{\partial s_{t_n}} \frac{\partial s_{t_n}}{\partial W_{ij}}$. Then, we arrive at the compact form

$$\frac{\partial s_{t_m}}{\partial W_{ij}} = \frac{\partial s_{t_m}}{\partial s_{r_n}} \frac{\partial s_{r_n}}{\partial W_{ij}}, \tag{19}$$

where we sum over all values of r less than t in addition to the standard dummy index n. More clearly, this is written as:

$$\frac{\partial s_{t_m}}{\partial W_{ij}} = \sum_{r=0}^{t} \frac{\partial s_{t_m}}{\partial s_{r_n}} \frac{\partial s_{r_n}}{\partial W_{ij}}, \tag{20}$$

which the referenced WildML tutorial indicates as the term responsible for the vanishing gradient problem.
Combining all of these yields:

$$\frac{\partial E_t}{\partial W_{ij}} = (\hat{y}_{t_l} - y_{t_l})\, V_{lm} \sum_{r=0}^{t} \frac{\partial s_{t_m}}{\partial s_{r_n}} \frac{\partial s_{r_n}}{\partial W_{ij}}. \tag{21}$$

2.3 U

Taking the gradient of U is similar to doing it for W since they both require taking sequential derivativs of an s_t vector. We have

$$\frac{\partial E_t}{\partial U_{ij}} = \frac{\partial E_t}{\partial \hat{y}_{t_k}} \frac{\partial \hat{y}_{t_k}}{\partial q_{t_l}} \frac{\partial q_{t_l}}{\partial s_{t_m}} \frac{\partial s_{t_m}}{\partial U_{ij}}. \tag{22}$$

Note that we only need to calculate the last term now. Following the same procedure as for W, we find that

$$\frac{\partial s_{t_m}}{\partial U_{ij}} = \sum_{r=0}^{t} \frac{\partial s_{t_m}}{\partial s_{r_n}} \frac{\partial s_{r_n}}{\partial U_{ij}}, \tag{23}$$

and thus we have:

$$\frac{\partial E_t}{\partial U_{ij}} = (\hat{y}_{t_l} - y_{t_l})\, V_{lm} \sum_{r=0}^{t} \frac{\partial s_{t_m}}{\partial s_{r_n}} \frac{\partial s_{r_n}}{\partial U_{ij}}. \tag{24}$$

The difference between U and W appears in the actual implementation since the values of $\frac{\partial s_{r_n}}{\partial U_{ij}}$ and $\frac{\partial s_{r_n}}{\partial W_{ij}}$ differ.

2.4 Total Gradient

Since our loss function (2) is just a summation of the E_t's, we can just sum up these values we've calculated over all relevant time-steps for a given backprop to calculate our total gradient.