Mathematical derivation of BPTT for RNN, and gradient vanishing and exploding problems

Susovan PAL

July 2019

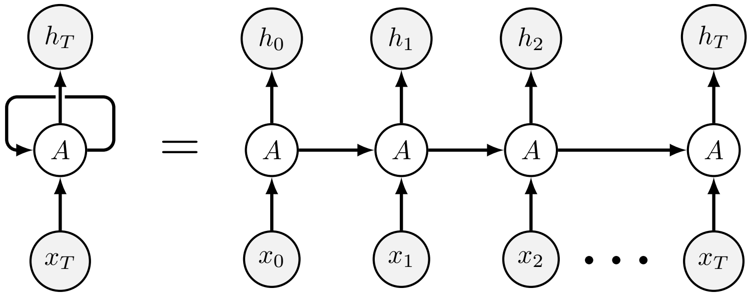
Contents

1 Introduction

In this short article, we attempt to derive the BPTT for RNN's in the most complete way as possible, and subsequently derive the gradient vanishing and exploding problems. We'll make many references to the following the ideas in the well known paper by Pascanu et. al. (link: `https://arxiv.org/pdf/1211.5063.pdf`), that derived the gradient vanishing and exploding problems in full detail, but the mathematical derivation of BPTT there (equation (4)) seems to be incomplete. I'll also make reference to some other blogs or videos where the derivation of BPTT, or equivalently, a similar equation as above have been mentioned, but I think in each of them, they lack rigor. This motivated me to write this article, where I attempt to give a complete derivation.

2 Notation

For the rest of the article, we'll use the following notations for RNN, confirming to the diagram below:



In above, the RNN is unrolled until time $t, 1 \leq t \leq T$, and let:
$x_t :=$ input at time t
$h_t :=$ hidden state activation at time t
$o_t =$ predicted output from the t-th layer.
$y_t :=$ target corresponding to the t-th layer.
$W_R :=$ weight matrix connecting two successive hidden layers, called recurrent weight matrix.
$W_I :=$ weight matrix connecting the input with the hidden layer
$W_O :=$ weight matrix connecting the putput with the hidden layer
$E_t :=$ error caused by the layer t, i.e. $E_t := E(o_t, y_t)$, where E is a pre-defined error or loss function, e.g. MSE or cross-entropy.

Then the generic equation of an RNN is:

$$h_t = \sigma(W_R h_{t-1} + W_I x_t + b_R)....(G1)$$
$$o_t = softmax(W_O h_t + b_O)....(G2)$$

Note above that if we set: $\tilde{h_t} := \sigma^{-1}(h_t), \forall t$, then the above two equations become:

$$\tilde{h_t} = W_R\sigma(\tilde{h_{t-1}}) + W_I x_t + b_R ...(R1)$$
$$o_t = softmax(W_O\sigma(\tilde{h_t}) + b_O)...(R2)$$

Given the equivalence between two seemingly different set of equations above, it makes sense to write the simpler form of RNN as follows, which is the one we're going to use for the rest of this paper:

$$h_t = W_R\sigma(h_{t-1}) + W_I x_t + b_R....(1)$$
$$o_t = softmax(W_O h_t + b_O)....(2)$$

3 Back propagation through time (BPTT)

3.1 Reducing BPTT with respect to just one parameter and for the loss from just one layer

While training the above RNN, we can use stochastic gradient descent (SGD) to optimally estimate the parameter set Θ for the RNN, namely: $\Theta = (W_R, W_I, W_O, b_R, b_I, b_O)$. Since SGD is a first order differential method, we'll eventually estimate $\frac{\partial E}{\partial \Theta}$. To keep the notations from being too clumsy, we'll instead just estimate only one component of the above derivative, namely: $\frac{\partial E}{\partial W_R}$, whereas the derivation of the formulae for the other components are similar. Next, we'll derive the formula just for $\frac{\partial E_t}{\partial W_R}$, where t is a fixed time step between 1 and T. This is enough, as we note that: $\frac{\partial E}{\partial W_R} = \Sigma_{t=1}^{T} \frac{\partial E_t}{\partial W_R}$

3.2 Derivation of $\frac{\partial E_t}{\partial W_R}$:

Assume $dim(W_R) = n \times n$. Recall that, following section $2, E_t$ is a function of the predicted output from the t-th layer o_t, which in turn is a function of the hidden state h_t, which in turn is a function of the matrix W_R. Also note that $\frac{\partial E_t}{\partial W_R}$ is actually a linear map from $\frac{\partial E_t}{\partial W_R} : \mathbb{R}^{n\times n} \to \mathbb{R}^n$, whereas $\frac{\partial E_t}{\partial W_R}(W_R)$ is the same linear map evaluated at W_R, hence it's an element of $\mathbb{R}^n$. However, for ease of notations, we'll identify $\frac{\partial E_t}{\partial W_R}(W_R)$ with $\frac{\partial E_t}{\partial W_R}$, a practice we'll continue for the rest of this article.

Hence we have:

$$\frac{\partial E_t}{\partial W_R} = \Sigma_{k=1}^{t} \frac{\partial E_t}{\partial h_t}\frac{\partial h_t}{\partial W_R} = \Sigma_{k=1}^{t} \frac{\partial E_t}{\partial o_t}\frac{\partial o_t}{\partial h_t}\frac{\partial h_t}{\partial W_R}......(3)$$

Dimensionality check: Following the discussion in the beginning of section 3.2 above, we note that both the left and right hand sides of the above equation are in $\mathbb{R}^n$. This is clear as $E_t : h_t \mapsto E_t(h_t) : \mathbb{R}^n \to \mathbb{R}, h_t : W_R \mapsto h_t(W_R) : \mathbb{R}^{n\times n} = \mathbb{R}^{n^2} \to \mathbb{R}^n$. Hence $\frac{\partial E_t}{\partial h_t}(h_t) \in \mathbb{R}^{1\times n} = \mathbb{R}^n$, and $\frac{\partial h_t}{\partial W_R}(W_R) \in \mathbb{R}^{n\times n^2}$.

Note above that, except the last term in each summand, namely $\frac{\partial h_t}{\partial W_R}$, the other two terms $\frac{\partial E_t}{\partial o_t}, \frac{\partial o_t}{\partial h_t}$ are straightforward to derive, thanks to the equations (1) and (2). The derivation of the last term $\frac{\partial h_t}{\partial W_R}$ is not so straightforward, as h_t is recursively defined in terms of h_{t-1}. So determining $\frac{\partial h_t}{\partial W_R}$ alone will determine $\frac{\partial E_t}{\partial W_R}$.

3.3 Expression of $\frac{\partial h_t}{\partial W_R}$, and its derivation

Theorem 1 *Let $v \in \mathbb{R}^n = \mathbb{R}^{n\times 1}$ be a column vector. Denote by $HV(v) \in \mathbb{R}^{n\times n^2}$ the matrix of dimension n by n^2 that is obtained by concatenating horizontally and vertically $n \times n = n^2$ copies of $v^T \in \mathbb{R}^{1\times n}$, so that the resulting matrix becomes one of dimension $n \times n^2$. More precisely, $HV(v)$ is defined by the $n \times n$ dimensional block matrix, where each block is v^T:*

$$\begin{bmatrix} v^T & v^T & ... & v^T \\ v^T & v^T & ... & v^T \\ v^T & v^T & ... & v^T \end{bmatrix}$$

Then for the RNN described by equation (1), we have:

$$\frac{\partial h_t}{\partial W_R} \equiv \frac{\partial h_t}{\partial W_R}(W_R) = \Sigma_{k=1}^{t-1}(\Pi_{r=k+1}^{t} W_R.diag(\sigma'(h_{r-1}))).HV(\sigma(h_{k-1})) + HV(\sigma(h_{t-1}))$$

Remark: The expression on the right hand side is not a polynomial in W_R unless w_R is a scalar, as it doesn't contain powers of the matrix W_R, as it does not in general commute with the matrix $diag(\sigma'(h_r))$, because $diag(\sigma'(h_r))$, although a diagonal matrix, doesn't have the same elements across its diagonal, as h_r is a vector, not a scalar, following equation (1). Also, note that, the expression contains highest $t-1$ factors of W_R, among all its summands (the products), which happens precisely in the product $k = 1$.

The proof is divided into several parts or steps.

Treating the other parameters than w_R as constants, we instead use the symbol $\frac{dE_t}{dW_R}$ instead of $\frac{\partial E_t}{\partial W_R}$, as this'll helps us simplify the notations further, and avoid confusion.

Lemma 1 *Let $\{h_t\}$ denote a recursively defined dynamical system which is a function of both and only of W_R and h_{t-1}. Then:* $\frac{dh_t}{dW_R} = \frac{\partial h_t}{\partial W_R} + \frac{\partial h_t}{\partial h_{t-1}} \frac{dh_{t-1}}{dW_R}$.

Remark: Before proving the lemma, note that here $\frac{\partial h_t}{\partial W_R}$ takes the derivative of h_t w.r.t. only W_R and not h_{t-1}. So it doesn't give us the total derivative of h_t w.r.t. W_R, as it doesn't take into account any contribution from h_{t-1}, which is also a function of W_R. The $\frac{\partial h_t}{\partial W_R}$ we defined earlier in this article, e.g. in equation (3) doesn't do the same: it takes the derivative of h_t w.r.t. W_R as a whole, using the recursive equation (1), and hence taking into account full contribution from h_{t-1}, not treating it as a constant.

To give a concrete example, consider the dynamical system where $\sigma = Id$, and h_0 is a non-zero constant (because if random initialization). So from (1), $h_t = W_R h_{t-1}$. Then $h_2 = W_R h_1 = W_R(W_R h_0) = W_R^2 h_0$. Hence $\frac{dh_2}{dW_R} = 2W_R h_0$ (this is the total derivative of h_t w.r.t. W_R), where $\frac{\partial h_2}{\partial W_R} = h_1 = W_R h_0$. Hence the old and new notations of $\frac{\partial h_t}{\partial W_R}$ are not the same.

Below is a proof of the above lemma.

Proof 1 *Using chain rule, and the fact $h_t = h_t(W_R, h_{t-1})$ (function of both), we've:*

$$\frac{dh_t}{dW_R} = \frac{\partial h_t}{\partial W_R}.\frac{dW_R}{dW_R} + \frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial W_R} = \frac{\partial h_t}{\partial W_R} + \frac{\partial h_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial W_R}$$

Lemma 2 *For an RNN following equation (1) and (2), we've:*

$$\frac{dh_t}{dW_R} = HV(\sigma(h_{t-1})) + W_R\sigma'(h_{t-1})\frac{dh_{t-1}}{dW_R}$$

Proof 2 *In lemma 1, just set the equation for RNN defined by equation (1), i.e.: $h_t = W_R\sigma(h_{t-1}) + W_I x_t + b_R$. Then we've: $\frac{\partial h_t}{\partial W_R} = HV(\sigma(h_{t-1}))$, and $\frac{\partial h_t}{\partial h_{t-1}} = W_R diag(\sigma'(h_{t-1}))$. This finishes the proof of the lemma.*

Finally, we're ready to obtain the expression in Theorem 1, using recursion. We'll derive the expression for $t = 1, 2$, and the general expression will be clear.

Proof for $t = 1$:

Note that for $t = 1$, $\frac{dh_1}{dW_R} = \frac{d(W_R\sigma(h_0)+W_I x_0+B_R}{dW_R} = \sigma(h_0)$, which matches with the expression of Theorem (1), as the left part doesn't simply exist, because here r starts from $k + 1 = 2$ when $k = 1$, but $t = 1$. Therefore, the product(s) do(es)n't exist, and so doesn't their sum.

Proof for $t = 2$:

By lemma (2), we've:

$\frac{dh_2}{dW_R}$
$= \sigma(h_1) + W_R diag(\sigma'(h_1))) \frac{dh_1}{dW_R}$
$= \sigma(h_1) + W_R diag(\sigma'(h_1))(\sigma(h_0) + W_R \sigma'(h_0) \frac{dh_0}{dW_R})$
$= \sigma(h_1) + W_R diag(\sigma'(h_1))\sigma(h_0)$ [as $\frac{dh_0}{dW_R} = 0$]
$= W_R diag(\sigma'(h_1))\sigma(h_0) + \sigma(h_1)$
$= \Sigma_{k=1}^{t-1=1}(\Pi_{r=k+1}^{t=2} W_R.diag(\sigma'(h_{r-1}))).\sigma(h_{k-1}) + \sigma(h_1)$

In the same way, we can prove Theorem 1 for general t.

4 Gradient vanishing problem for long RNN's

4.1 Mathematical derivation

In this section, we'll first do a mathematical derivation, and then interpret in the next section what gradient vanishing problem is, what what it is not. To ease notations, we write in Theorem 1:

$$\frac{\partial h_t}{\partial W_R} = \Sigma_{k=1}^{t-1} F(k,t) + HV(\sigma(h_{t-1}))$$

We state the following theorem on vanishing gradient:

Theorem 2 *Denote by λ_1 the eigenvalue of W_R that has the largest absolute value. Assume $||diag(\sigma'(.))|| \leq \gamma$, and $\lambda_1 < 1/\gamma$. Then for any fixed $k = k_0$, we have: $lim_{t\to\infty} F(k_0, t) = 0$.*

Proof 3 *Note that:*
$||F(k_0, t)||$
$= ||\Pi_{r=k_0+1}^{t} W_R.diag(\sigma'(h_{r-1}))).\sigma(h_{k-1})||$
$\leq ||W_R||^{t-k_0}.||diag(\sigma'(h_{r-1}))||^{t-k_0}.||\sigma(h_{t-1})||^{t-k_0}$
$\leq (||W_R||.||\sigma(h_{t-1})||)^{t-k_0}$
$\leq (\lambda_1 ||\sigma(h_{t-1})||)^{t-k_0}$
$\leq (\lambda_1 \gamma)^{t-k_0}$
$\leq \eta^{t-k_0}$*, where* $\eta = \lambda_1 \gamma < 1$
$\to 0$ *as* $t \to \infty$

4.2 Interpretation of Theorem 2: learning difficulty for long RNN's

As shown in Theorem 2, the contributions of the long terms ($t >> k_0$) in the expression $\frac{\partial h_t}{\partial W_R}$, and hence of $\frac{\partial E_t}{\partial W_R}$ of will be very small when t is large. This means that, when we use (stochantic/minibatch) gradient descent for large t, the values of $\frac{\partial E_t}{\partial W_R}$ for large t will not be affected much by the new inputs corresponding to those large t's. But this is not what we expected when we wanted to design RNN's, as we should expect the longer terms to affect the "learning process", i.e. we expect the values of the gradient $\frac{\partial E_t}{\partial W_R}$ to be affected by the long term inputs. This is not the case here, as this gradient above become stabilized for long terms. This is the "gradient vanishing problem".

Remark: We should be cautioned that the "gradient vanishing" doesn't mean that the gradient term $\frac{\partial E_t}{\partial W_R}$ vanish: clearly it does not, as we note in the expression above in Theorem 1, that the term $\sigma(h_{t-1})$ will always be present in the expression of $\frac{\partial h_t}{\partial W_R}$, and so will the first few summands in $\frac{\partial h_t}{\partial W_R}$, i.e. when $k = 1, 2, 3$ etc. So learning will still happen, but it won't be affected by the long term inputs.

4.3 What happens when "gradient vanishing" doesn't happen?

COMING UP.

5 Gradient exploding problem for long RNN's

First we state and prove a special case of the above phenomenon.

Theorem 3 *Consider the the dynamical system where the hidden state has the equation: $h_t = W_R\sigma(h_{t-1}) + W_I x_t + b_R$ with $\sigma = Id$. Let the recurrent weight matrix W_R has dimension $n \times n$, and has n distinct eigenvalues $\{\lambda_1, \lambda_2, ...\lambda_n\}$ in the decreasing order of absolute values, i.e. $\lambda_1 > \lambda_2 > ... > \lambda_n$, and let $\{q_1, q_2...q_n\}$ be the corresponding eigenvectors. Let $\frac{\partial E_t}{\partial h_t} = \Sigma_{i=1}^N c_i q_i$, and j be a positive integer such that for each $j' < j, c_{j'} = 0$. Let $\lambda_j > 1$.*

Then $\frac{\partial E_t}{\partial W_R}$ grows exponentially along the direction of q_j, as $t \to \infty$.

For a proof of this special case, see Pascanu's paper: `https://arxiv.org/pdf/1211.5063.pdf`. We'll, however, prove a more general version of the theorem that'll work with any activation function σ, not just identity. The corresponding theorem is:

ATTENTION: The theorem and proof of theorem 4 may be wrong!!! WILL CHECK!!!

Theorem 4 *Choose any $\epsilon > 0$. Assume that $||W_R|| \geq (1+\epsilon)/(\lambda^*(diag(\sigma'(*))))$. Then: $||\frac{\partial E_t}{\partial W_R}|| \to \infty$ as $t \to \infty$.*

To prove the above theorem, first, we state a lemma without proof: for a proof, see `https://math.stackexchange.com/questions/268316/lower-bound-on-norm-of-product-of-two-ma`

Lemma 3 *Let A, B be two square matrices. Let $\lambda^*(A)$ denote the minimum absolute value of all the eigenvalues of A. Then, for the 2-norms or operator norms, we have: $||AB|| \geq \lambda^*(A)||B||$.*

Remark: Following the above lemma, we'll use the symbol $\lambda^*(M)$ to denote the absolute value of the minimum eigenvalues of the matrix M.
Note as before that, to prove the above theorem, it's enough to prove it for $\frac{\partial h_t}{\partial W_R}$.

Proof 4 *Recall the expression of Theorem 1: $\frac{\partial h_t}{\partial W_R} = \Sigma_{k=1}^{t-1}(\Pi_{r=k+1}^{t} W_R.diag(\sigma'(h_{r-1}))).HV(\sigma(h_{k-1})^T)+ HV(\sigma(h_{t-1})^T)$*

Consider for a fixed k, part (=a factor) of the k-th summand above, i.e. $\Pi_{r=k+1}^{t} W_R.diag(\sigma'(h_{r-1}))).HV(\sigma(h_{k-1}$ when $r = k+1$, i.e. consider just the factor $\Pi_{r=k+1}^{t} W_R.diag(\sigma'(h_{r-1})))$. Let's consider it's (operator or equivalently 2-) norm, which is the largest absolute value of all its eigenvalues.

Upon using Lemma 3 above repeatedly with $A = diag(\sigma'(h_{r-1}))$, we get:

$||\Pi_{r=k+1}^{t} W_R.diag(\sigma'(h_{r-1})))||$
$\geq (\Pi_{r=k+1}^{t}(\lambda^*(diag(\sigma'(h_{r-1}))))).||W_R||^{t-k}$
$\geq (1+\epsilon)^{t-k}$*, because of the assumption on the norm* $||W_R||$
$\rightarrow \infty$ *as* $t \rightarrow \infty$*.*