On the Difficulty of Training RNNs

R. Pascanu, T. Mikolov, Y. Bengio

ICML 2013

Presented by OUR NAMES: Hanna M. Dettki

1. Introduction & Background

TODO:

1.1 Context & Motivation

- **Importance of sequence modeling**
  - e.g., language, time-series in finance
- Identifying gradient problems (Bengio et al., 1994)
  - **Vanishing gradient problem**: impossible to learn long-term dependencies
  - **Exploding gradient problem**: numerical instabilities → unstable training
- → Why stable gradient flow is critical for learning temporal dependencies (paper's contribution)

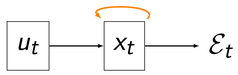


Fig. 1

1.2 Schematic & formal def. of RNN

$$x_t = F(x_{t-1}, u_t, \theta) \qquad (1)\ \text{General}$$
$$x_t = W_{\text{rec}}\,\sigma(x_{t-1}) + W_{in}u_t + \mathbf{b} \qquad (2)$$

where u_t: input, x_t: state, t: time step, $\mathbf{b}$: bias, $\mathcal{E}_t = \mathcal{L}(x_t)$ (error)

The recurrent connections in the hidden layer allow information to persist from one input to another.

1.3 Training RNNs: Backprop Through Time (BPTT) on Unrolled RNN

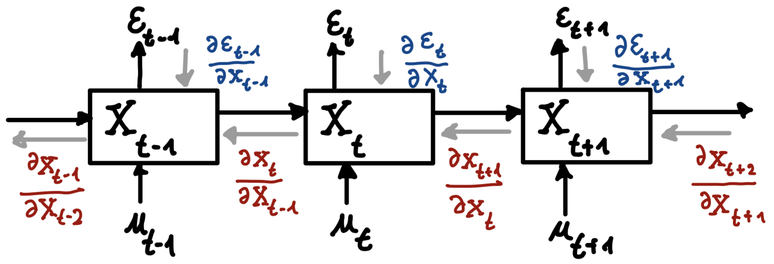


Fig. 2: Unrolled RNN

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{E}_t}{\partial \theta} \qquad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{k=1}^{t} \left(\frac{\partial \mathcal{E}_t}{\partial x_t} \frac{\partial x_t}{\partial x_k} \frac{\partial^+ x_k}{\partial \theta} \right) \qquad (4)$$

$$\frac{\partial x_t}{\partial x_k} = \prod_{i=k+1}^{t} W_{\text{rec}}{}^{\top} \cdot \text{diag}\left(\sigma'(x_{i-1})\right) \qquad (5)$$

where $\frac{\partial^+ x_k}{\partial \theta}$ denotes the "immediate" partial derivative (treating x_{k-1} as constant).

Blue: total gradient over time. Red: temporal error contribution.

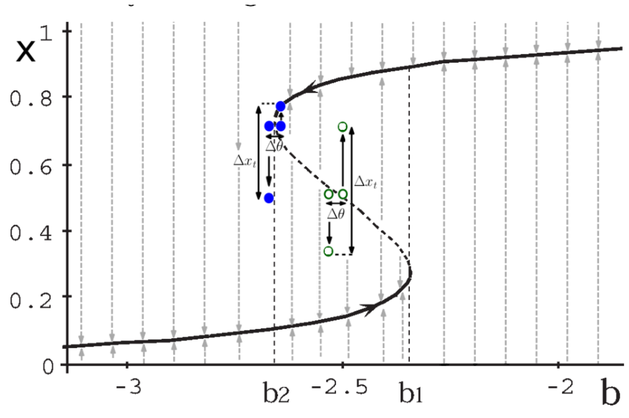
2. The Problem

Mechanics of Exploding and Vanishing Gradients:

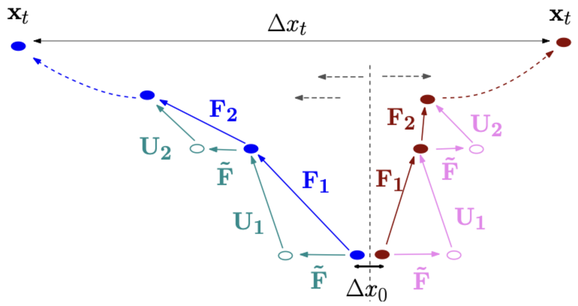
These issues occur in RNNs due to repeated multiplication of Jacobian matrices during backpropagation. If the spectral radius ρ of the recurrent weight matrix W_{rec} is less than 1, gradients vanish; if greater than 1, they explode. For non-linear activations with bounded derivatives (e.g., $\gamma = 1$ for tanh), gradients vanish when the largest singular value $\lambda_1 < \gamma^{-1}$.

Dynamical Systems View:

An RNN's hidden state evolves like a dynamical system converging to attractors. As parameters change, the system may cross bifurcation points, causing drastic changes in state evolution. Crossing basin boundaries can result in gradient explosions. Inputs can shift the system into different attractor basins, intensifying this instability.



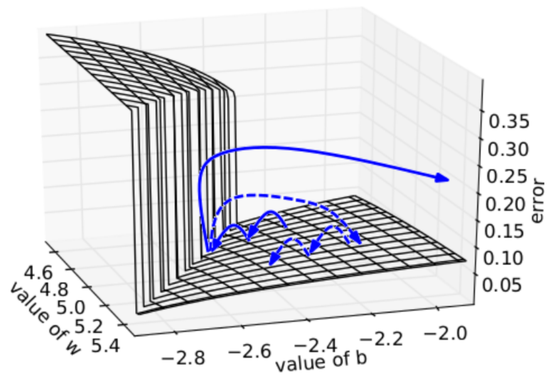
Bifurcation diagram showing attractor transitions



Gradient explosion due to basin crossing

Geometric Interpretation:

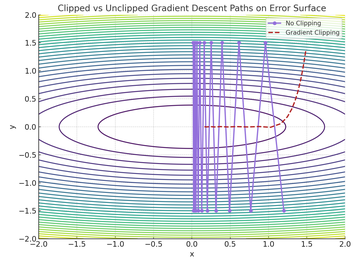
Consider $x_t = w\sigma(x_{t-1}) + b$ with $x_0 = 0.5$. In the linear case ($b = 0$), gradients are $\frac{\partial x_t}{\partial w} = tw^{t-1}x_0$, showing exponential growth. Exploding gradients align with steep directions in the error surface, forming sharp walls that SGD struggles to traverse, disrupting convergence.



Steep error surface caused by exploding gradients

3. Solution & Experiments

Gradient Clipping



- Pseudo-code: $\hat{g} \leftarrow \nabla E$; if $\|\hat{g}\|_2 \geq \tau$ then $\hat{g} \leftarrow \tau \cdot \frac{\hat{g}}{\|\hat{g}\|_2}$
- Gradient clipping introduces a hyperparameter: the threshold. A common heuristic sets this value based on the average gradient norm over early training steps.
- Compared to clipping individual gradient components by value, norm-based clipping preserves the direction of the gradient vector and is generally more robust in high-dimensional settings.

Vanishing Gradient Regularization

- Regularizer:

$$\Omega = \sum_k \Omega_k = \sum_k \left(\frac{\left\| \frac{\partial E}{\partial x_{k+1}} \cdot \frac{\partial x_{k+1}}{\partial x_k} \right\|}{\left\| \frac{\partial E}{\partial x_{k+1}} \right\|} - 1 \right)^2$$

$$\frac{\partial^+ \Omega}{\partial W_{\text{rec}}} = \sum_k \frac{\partial^+}{\partial W_{\text{rec}}} \left(\frac{\left\| \frac{\partial E}{\partial x_{k+1}} \cdot W_{\text{rec}}^{\top} \cdot \text{diag}\left(\sigma'(x_k)\right) \right\|^2}{\left\| \frac{\partial E}{\partial x_{k+1}} \right\|^2} - 1 \right)^2$$

- The regularization term only enforces norm preservation of the Jacobian matrix $\frac{\partial x_{k+1}}{\partial x_k}$ in the direction of the error signal $\frac{\partial E}{\partial x_{k+1}}$, not in all directions.
- The soft constraint does not guarantee perfect norm preservation, so exploding gradients may still occur, particularly during early training or unstable updates. To mitigate this, we combine the regularizer with gradient clipping for more stable and effective learning.

Experiments and Results

Experiments and Results To edit

This section presents the experimental setup and results.
evaluated the performance of different initialization methods.
The temporal order problem was tested across various sequence lengths.
Three initialization techniques were compared: sigmoid, basic tanh, and smart tanh.
The results highlight the impact of initialization on success rates.
Detailed analysis shows smart tanh outperforming the others.
See the figure below for a visual comparison of the results.

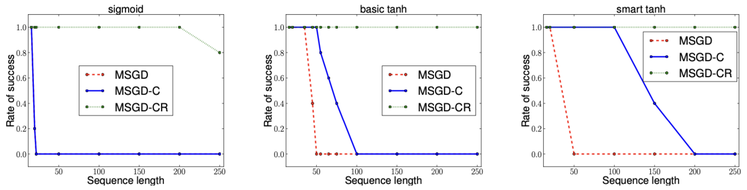


Figure 7: Rate of success for solving the temporal order problem versus sequence length for different initializations (from left to right: sigmoid, basic tanh and smart tanh)

Results - what worked

Line 1
Line 2.
Line 3.
Line 4
Line 5
Line 6
Line 7
Line 8
Line 9
Line 10

4. Relevance today & SOTA techniques

Exploding gradients: Clipping is still relevant!

Vanishing gradients: Alternatives to regularization:

- Residual connections
- Gating mechanisms
- Attention mechanism
- Gradient checkpointing
- Layer normalization
- Positional encoding