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)

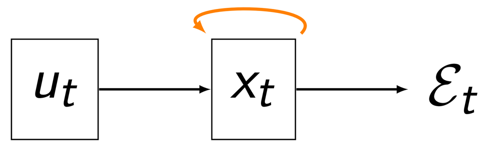


Fig. 1

1.2 Schematic & formal def. of RNN

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

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

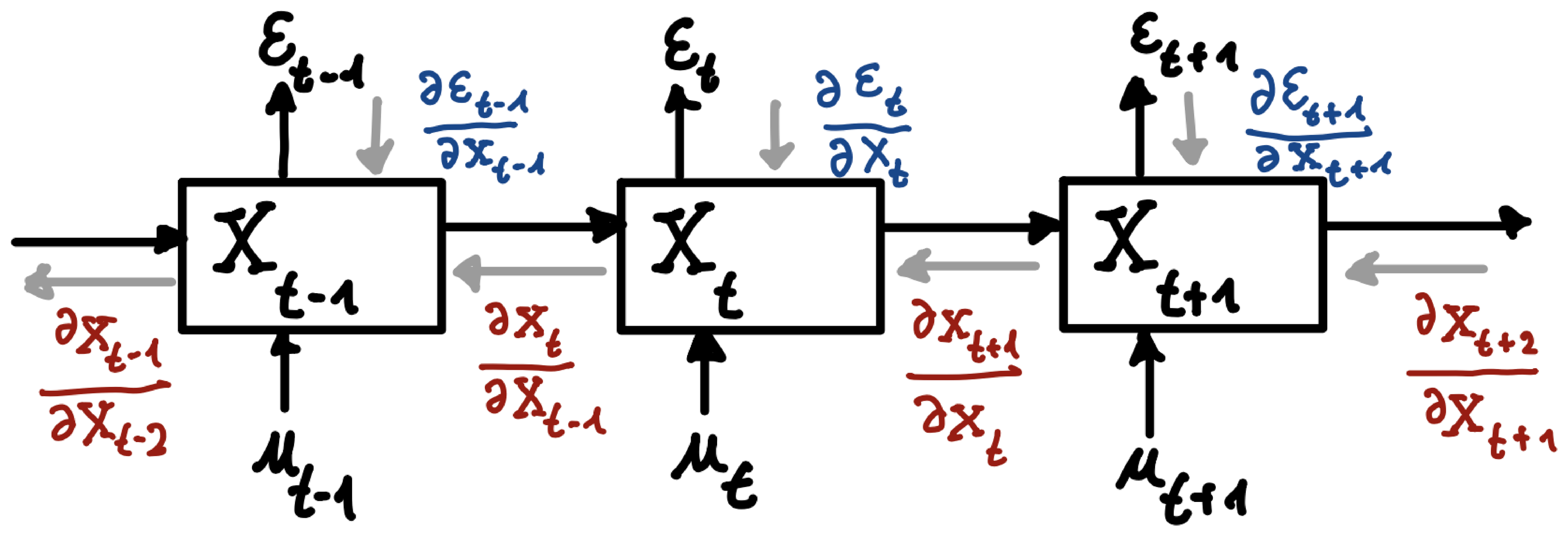


Fig. 2: Unrolled RNN

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

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

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

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.

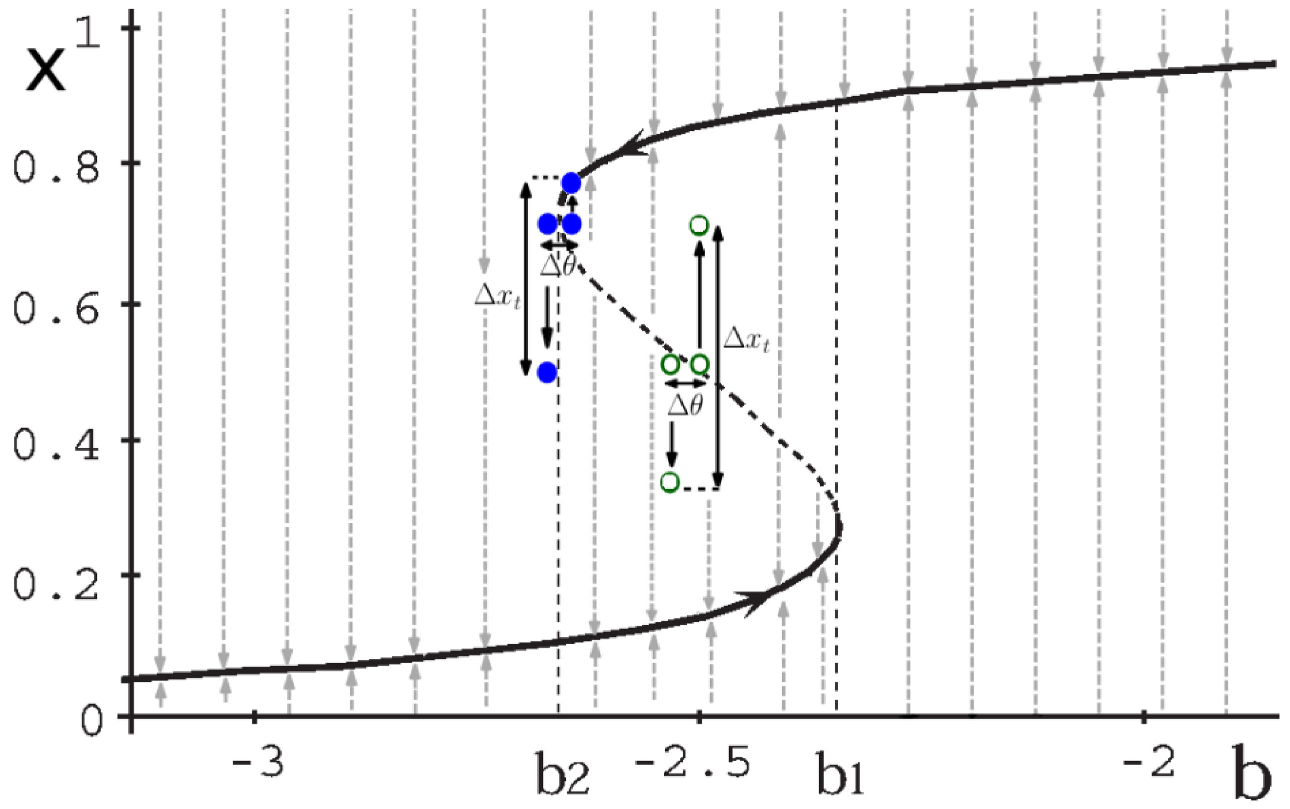
2. The Problem

Mechanics of Exploding and Vanishing Gradients:

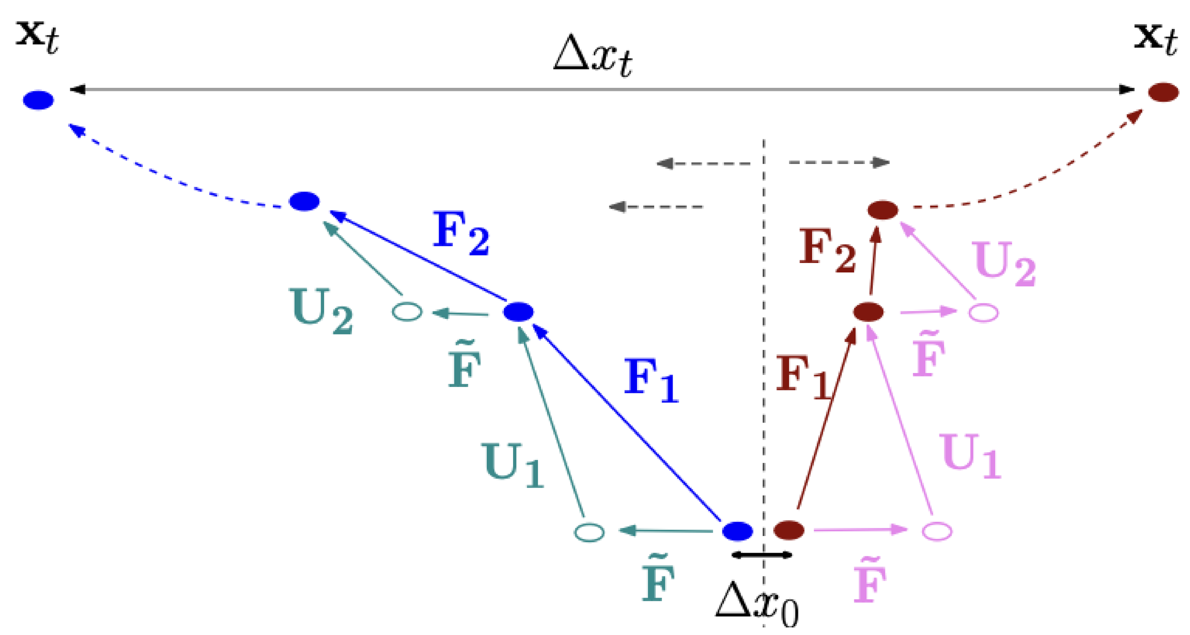
These issues occur in RNNs due to repeated multiplication of Jacobian matrices during backpropagation. If the spectral radius ρ of the recurrent weight matrix W_{rec} is less than 1, gradients vanish; if greater than 1, they explode. For non-linear activations with bounded derivatives (e.g., $\gamma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



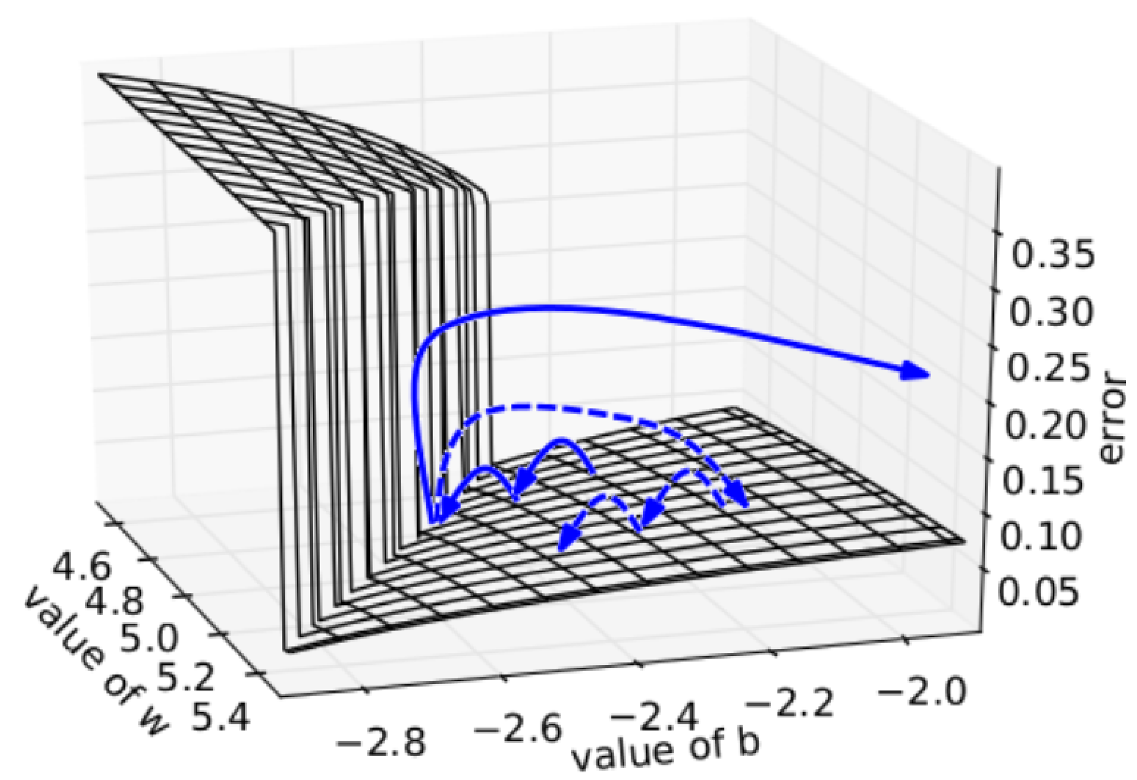
Bifurcation diagram showing attractor transitions



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

3.1 Gradient Clipping

- see Fig. 3 for *motivation*
- Pseudo-code for norm-clipping
- **TODO:**

3.2 VG-Regularization

- see paper eq. 9 & 10
- **TODO:**

3.3 MSGD-CR

- **TODO:**
- (combines 3.1 & 3.2)

3.4 Initialization Strategies

- see experiments in Sec. 4

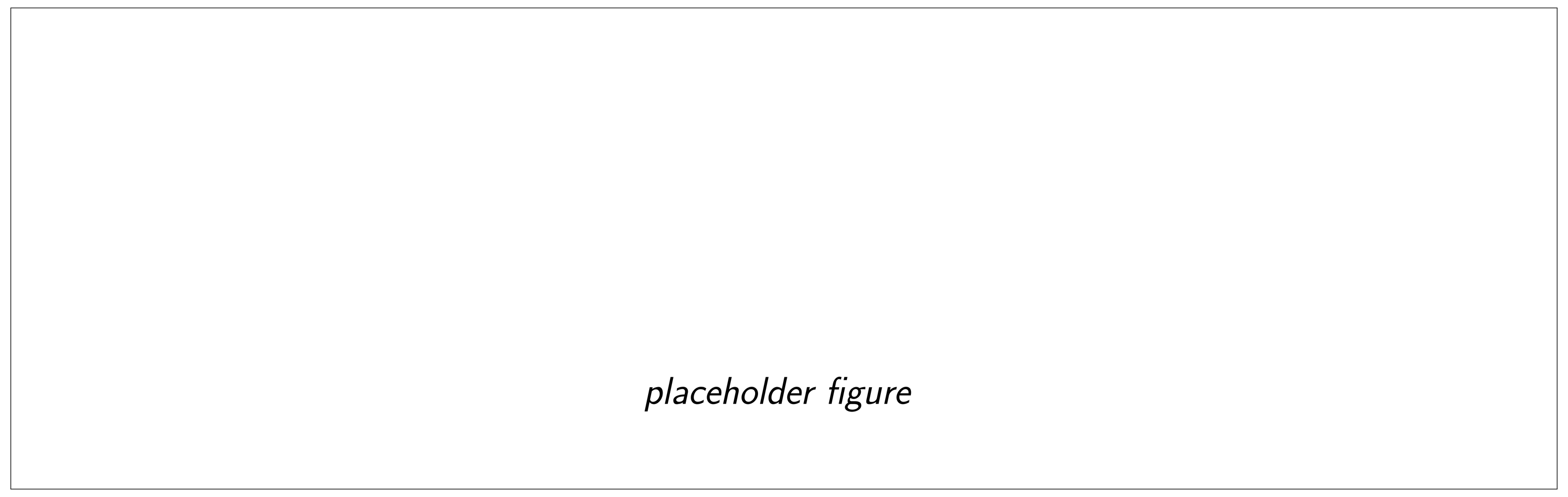


Fig. 4

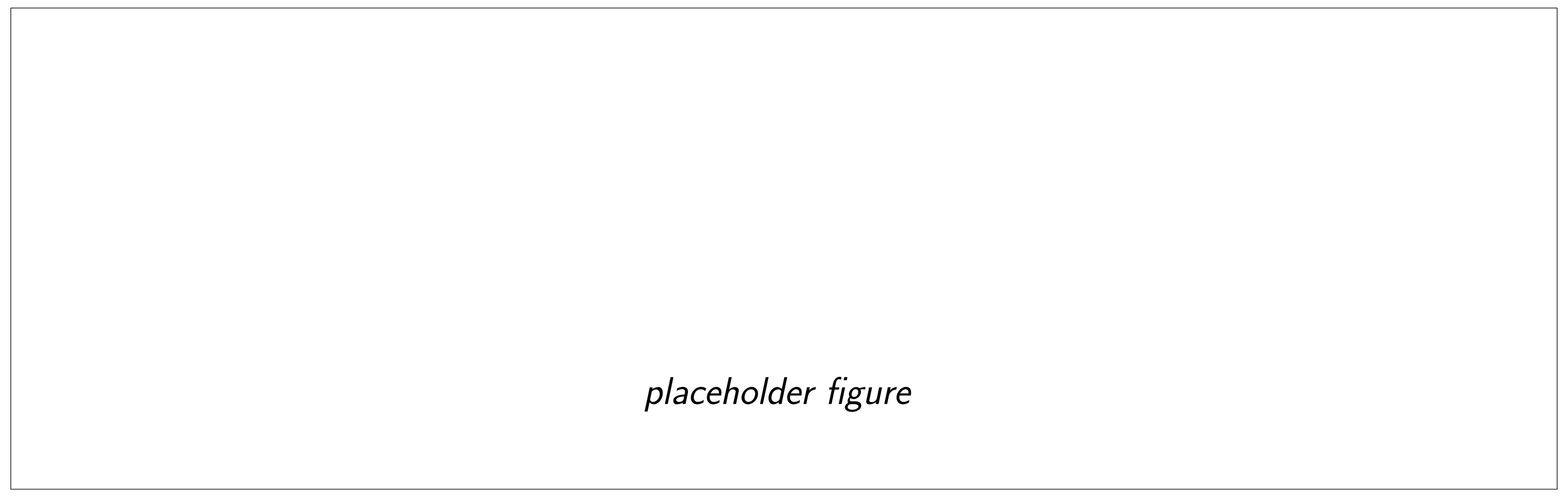


Fig. 5

4. Relevance today & SOTA techniques

TODO:

- Clipping still relevant!
- Instead of regularization:
  - residual connections
  - gradient checkpointing
  - gating mechanisms
  - layer normalization
  - attention mechanism
  - positional encoding