LSTM Network - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

A Long Short-Term Memory (LSTM) network augments the vanilla RNN with a *cell state* that supports nearly constant-gradient flow across long time spans via multiplicative *gates* (input, forget, output). Each gate is a learned, data-dependent control signal that regulates writing to, retaining in, and reading from the cell. We derive the forward equations, a full Backpropagation Through Time (BPTT) for LSTM, and analyze the gradient dynamics ("constant error carousel") that ameliorate vanishing/exploding gradients.

1.2 Data and Notation

Let a sequence be $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{T}$ with

$$\mathbf{x}_t \in \mathbb{R}^{d_x}, \qquad \mathbf{y}_t \in \{0,1\}^K \text{ (one-hot target) or } \mathbb{R}^K.$$

Hidden/output state $\mathbf{h}_t \in \mathbb{R}^{d_h}$; *cell state* $\mathbf{c}_t \in \mathbb{R}^{d_h}$. Gates and candidate:

$$\mathbf{i}_t,\ \mathbf{f}_t,\ \mathbf{o}_t,\ \mathbf{g}_t \in \mathbb{R}^{d_h}.$$

Parameters (no peepholes in the base derivation):

$$\mathbf{W}_{x\bullet} \in \mathbb{R}^{d_h \times d_x}, \quad \mathbf{W}_{h\bullet} \in \mathbb{R}^{d_h \times d_h}, \quad \mathbf{b}_\bullet \in \mathbb{R}^{d_h}, \qquad \bullet \in \{i, f, o, g\}.$$

Readout: $\mathbf{W}_{hy} \in \mathbb{R}^{K \times d_h}$, $\mathbf{b}_y \in \mathbb{R}^K$. Elementwise non/linearities: logistic $\sigma(u) = 1/(1+e^{-u})$, $\tanh(u)$; Hadamard product $\odot$; $\texttt{Diag}(\cdot)$ forms a diagonal matrix from a vector.

1.3 Model Formulation (Forward Dynamics)

Given $(\mathbf{h}_0, \mathbf{c}_0)$ (zeros or learned), define gate pre-activations

$$\mathbf{a}_t^{(i)} = \mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i, \qquad \mathbf{i}_t = \sigma\big(\mathbf{a}_t^{(i)}\big), \tag{1}$$

$$\mathbf{a}_t^{(f)} = \mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f, \qquad \mathbf{f}_t = \sigma\big(\mathbf{a}_t^{(f)}\big), \tag{2}$$

$$\mathbf{a}_t^{(g)} = \mathbf{W}_{xg}\mathbf{x}_t + \mathbf{W}_{hg}\mathbf{h}_{t-1} + \mathbf{b}_g, \qquad \mathbf{g}_t = \tanh\big(\mathbf{a}_t^{(g)}\big), \tag{3}$$

$$\mathbf{a}_t^{(o)} = \mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o, \qquad \mathbf{o}_t = \sigma\big(\mathbf{a}_t^{(o)}\big). \tag{4}$$

Cell update and hidden/output:

$$\text{Cell:} \quad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \tag{5}$$

$$\text{Hidden:} \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \tag{6}$$

Readout logits and (for classification) softmax:

$$\mathbf{z}_t = \mathbf{W}_{hy}\mathbf{h}_t + \mathbf{b}_y \in \mathbb{R}^K, \qquad \hat{\mathbf{p}}_t = \mathrm{softmax}(\mathbf{z}_t). \tag{7}$$

Many-to-many vs. many-to-one. Use all $\{\mathbf{z}_t\}$ for per-time predictions (many-to-many) or only $t{=}T$ for sequence classification.

1.4 Training Objective (Empirical Risk)

Sum of per-time losses (cross-entropy shown; squared loss is analogous):

$$\mathcal{L} \;=\; \sum_{t=1}^{T} \ell_t, \qquad \ell_t \;=\; -\,\mathbf{y}_t^\top \log \hat{\mathbf{p}}_t. \tag{8}$$

With weight decay $\frac{\lambda}{2}\sum_\bullet\big(\|\mathbf{W}_{x\bullet}\|_F^2 + \|\mathbf{W}_{h\bullet}\|_F^2\big) + \frac{\lambda}{2}\|\mathbf{W}_{hy}\|_F^2$ if desired.

1.5 Backpropagation Through Time (BPTT) for LSTM

We unroll Eqns. (1)–(7) over $t = 1{:}T$ and apply reverse-mode AD. Define the logits error

$$\boldsymbol{\delta}_t^{(z)} \;\triangleq\; \frac{\partial \ell_t}{\partial \mathbf{z}_t} \;=\; \hat{\mathbf{p}}_t - \mathbf{y}_t \in \mathbb{R}^K. \tag{9}$$

We track adjoints (total derivatives) for $\mathbf{h}_t$ and $\mathbf{c}_t$:

$$\mathbf{g}_t \;\triangleq\; \frac{\partial \mathcal{L}}{\partial \mathbf{h}_t} \in \mathbb{R}^{d_h}, \qquad \mathbf{q}_t \;\triangleq\; \frac{\partial \mathcal{L}}{\partial \mathbf{c}_t} \in \mathbb{R}^{d_h}.$$

Initialize $\mathbf{g}_{T+1} = \mathbf{0}$, $\mathbf{q}_{T+1} = \mathbf{0}$.

Backward recurrences (per time step $t = T{:}1$).

1. Accumulate loss-to-hidden via readout:

$$\mathbf{g}_t \leftarrow \mathbf{g}_t + \mathbf{W}_{hy}^\top \boldsymbol{\delta}_t^{(z)}.$$

2. From $\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$:

$$\text{to output gate: } \bar{\mathbf{o}}_t = \mathbf{g}_t \odot \tanh(\mathbf{c}_t), \qquad \text{to cell:} \qquad \tilde{\mathbf{q}}_t = \mathbf{g}_t \odot \mathbf{o}_t \odot \big(1 - \tanh^2(\mathbf{c}_t)\big).$$

3. Accumulate cell adjoint (including future through $\mathbf{c}_{t+1}$):

$$\mathbf{q}_t \leftarrow \mathbf{q}_t + \tilde{\mathbf{q}}_t.$$

4. From $\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$:

$$\bar{\mathbf{f}}_t = \mathbf{q}_t \odot \mathbf{c}_{t-1}, \quad \bar{\mathbf{i}}_t = \mathbf{q}_t \odot \mathbf{g}_t, \quad \bar{\mathbf{g}}_t = \mathbf{q}_t \odot \mathbf{i}_t, \quad \text{and} \quad \mathbf{q}_{t-1} \leftarrow \mathbf{q}_{t-1} + \mathbf{q}_t \odot \mathbf{f}_t.$$

5. Pass through gate nonlinearities to *pre-activations* (elementwise):

$$\boldsymbol{\delta}_t^{(o)} = \bar{\mathbf{o}}_t \odot \mathbf{o}_t \odot (1 - \mathbf{o}_t), \tag{10}$$

$$\boldsymbol{\delta}_t^{(f)} = \bar{\mathbf{f}}_t \odot \mathbf{f}_t \odot (1 - \mathbf{f}_t), \tag{11}$$

$$\boldsymbol{\delta}_t^{(i)} = \bar{\mathbf{i}}_t \odot \mathbf{i}_t \odot (1 - \mathbf{i}_t), \tag{12}$$

$$\boldsymbol{\delta}_t^{(g)} = \bar{\mathbf{g}}_t \odot \left(1 - \mathbf{g}_t^{\odot 2}\right). \tag{13}$$

6. Parameter gradients (outer-product accumulations):

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{xo}} \mathrel{+}= \boldsymbol{\delta}_t^{(o)} \mathbf{x}_t^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ho}} \mathrel{+}= \boldsymbol{\delta}_t^{(o)} \mathbf{h}_{t-1}^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}_o} \mathrel{+}= \boldsymbol{\delta}_t^{(o)}, \tag{14}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{xf}} \mathrel{+}= \boldsymbol{\delta}_t^{(f)} \mathbf{x}_t^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hf}} \mathrel{+}= \boldsymbol{\delta}_t^{(f)} \mathbf{h}_{t-1}^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}_f} \mathrel{+}= \boldsymbol{\delta}_t^{(f)}, \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{xi}} \mathrel{+}= \boldsymbol{\delta}_t^{(i)} \mathbf{x}_t^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hi}} \mathrel{+}= \boldsymbol{\delta}_t^{(i)} \mathbf{h}_{t-1}^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}_i} \mathrel{+}= \boldsymbol{\delta}_t^{(i)}, \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{xg}} \mathrel{+}= \boldsymbol{\delta}_t^{(g)} \mathbf{x}_t^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hg}} \mathrel{+}= \boldsymbol{\delta}_t^{(g)} \mathbf{h}_{t-1}^\top, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{b}_g} \mathrel{+}= \boldsymbol{\delta}_t^{(g)}. \tag{17}$$

7. Propagate to the previous hidden:

$$\mathbf{g}_{t-1} \leftarrow \mathbf{g}_{t-1} + \mathbf{W}_{ho}^\top \boldsymbol{\delta}_t^{(o)} + \mathbf{W}_{hf}^\top \boldsymbol{\delta}_t^{(f)} + \mathbf{W}_{hi}^\top \boldsymbol{\delta}_t^{(i)} + \mathbf{W}_{hg}^\top \boldsymbol{\delta}_t^{(g)}.$$

Finally, the readout gradients are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hy}} = \sum_{t=1}^{T} \boldsymbol{\delta}_t^{(z)} \mathbf{h}_t^\top, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{b}_y} = \sum_{t=1}^{T} \boldsymbol{\delta}_t^{(z)}. \tag{18}$$

Correctness (chain-rule sketch). From Eqn. (6), $\partial \ell / \partial \mathbf{c}_t = \mathbf{g}_t \odot \mathbf{o}_t \odot (1 - \tanh^2 \mathbf{c}_t)$ and $\partial \ell / \partial \mathbf{o}_t = \mathbf{g}_t \odot \tanh(\mathbf{c}_t)$. From Eqn. (5), $\partial \ell / \partial \mathbf{c}_{t-1} = \mathbf{q}_t \odot \mathbf{f}_t$ and $\partial \ell / \partial \mathbf{f}_t = \mathbf{q}_t \odot \mathbf{c}_{t-1}$, $\partial \ell / \partial \mathbf{i}_t = \mathbf{q}_t \odot \mathbf{g}_t$, $\partial \ell / \partial \mathbf{g}_t = \mathbf{q}_t \odot \mathbf{i}_t$. Applying elementwise derivatives of σ and tanh yields the preactivation deltas above; linear maps then produce parameter/hidden gradients. ■

1.6 Gradient Dynamics: Constant Error Carousel (CEC)

From Eqn. (5), the *exact* Jacobian of the cell across time is

$$\frac{\partial \mathbf{c}_t}{\partial \mathbf{c}_{t-1}} = \mathtt{Diag}(\mathbf{f}_t). \tag{19}$$

Thus the backpropagated cell gradient satisfies

$$\mathbf{q}_{t-\tau} = \left(\prod_{s=t-\tau+1}^{t} \mathtt{Diag}(\mathbf{f}_s) \right) \mathbf{q}_t + \cdots, \tag{20}$$

where "$\cdots$" are local contributions at intermediate steps. If $\mathbf{f}_s \approx \mathbf{1}$ (forget gates near 1), the product stays near the identity and gradients remain *nearly constant* over long spans; if $\|\mathbf{f}_s\| < 1$ uniformly, gradients decay. This multiplicative control explains LSTM's robustness to vanishing/exploding gradients relative to vanilla RNNs.

1.7 Variants and Options

- **Peephole LSTM.** Add terms $\mathbf{W}_{co}\mathbf{c}_t$, $\mathbf{W}_{cf}\mathbf{c}_{t-1}$, $\mathbf{W}_{ci}\mathbf{c}_{t-1}$ in Eqns. (1)–(4). Backprop adds paths from $\mathbf{c}$ to gate preactivations.
- **Coupled inputforget (CIFG).** Set $\mathbf{f}_t = \mathbf{1} - \mathbf{i}_t$ to reduce parameters; adjust backprop accordingly.
- **Bias init.** Initialize $\mathbf{b}_f$ to a positive value (e.g., 1) to encourage long memory early in training.

1.8 Optimization and Practicalities

- **Truncated BPTT.** Backprop over windows of length $L \ll T$ to limit memory/compute.
- **Regularization.** Weight decay, dropout on inputs/outputs (and variational dropout on recurrent connections), gradient clipping.
- **Initialization.** Orthogonal $\mathbf{W}_{h\bullet}$, Xavier/He for $\mathbf{W}_{x\bullet}$; $\mathbf{h}_0, \mathbf{c}_0$ zeros or learned.

1.9 Algorithm (LSTM + BPTT)

1. **Input:** sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{T}$; learning rate η.
2. **Forward for $t = 1{:}T$:** compute gates Eqns. (1)–(4), cell/hidden Eqns. (5)–(6), logits/probs Eqn. (7); accumulate $\mathcal{L}$ via Eqn. (8).
3. **Backward for $t = T{:}1$:** compute $\boldsymbol{\delta}_t^{(z)}$ Eqn. (9), run steps 17 above; accumulate parameter gradients.
4. **Update:** $\Theta \leftarrow \Theta - \eta \nabla \mathcal{L}$ with SGD/Adam (clip if $\|\nabla\|_2 > c$).

1.10 Computational Aspects

Per step, forward/backward are $\mathcal{O}(d_h d_x + d_h^2 + K d_h)$, with a constant ≈ 4 over vanilla RNN due to four gate/candidate pathways. Memory $\mathcal{O}(T d_h)$ to store $\{\mathbf{a}_t^{(\bullet)}, \mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t, \mathbf{g}_t, \mathbf{c}_t, \mathbf{h}_t\}$.

1.11 Summary of Variables and Their Dimensions

- $\mathbf{x}_t \in \mathbb{R}^{d_x}$: input at time t; $\mathbf{y}_t \in \{0,1\}^K$ or $\mathbb{R}^K$: target.
- $\mathbf{h}_t \in \mathbb{R}^{d_h}$: hidden/output state; $\mathbf{c}_t \in \mathbb{R}^{d_h}$: cell state.
- Gates/candidate: $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t, \mathbf{g}_t \in \mathbb{R}^{d_h}$.
- Preactivations: $\mathbf{a}_t^{(i)}, \mathbf{a}_t^{(f)}, \mathbf{a}_t^{(o)}, \mathbf{a}_t^{(g)} \in \mathbb{R}^{d_h}$.
- Parameters: $\mathbf{W}_{x\bullet} \in \mathbb{R}^{d_h \times d_x}$, $\mathbf{W}_{h\bullet} \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{b}_\bullet \in \mathbb{R}^{d_h}$ for $\bullet \in \{i, f, o, g\}$; readout $\mathbf{W}_{hy} \in \mathbb{R}^{K \times d_h}$, $\mathbf{b}_y \in \mathbb{R}^K$.
- Backprop adjoints: $\boldsymbol{\delta}_t^{(z)} \in \mathbb{R}^K$; $\mathbf{g}_t = \partial\mathcal{L}/\partial\mathbf{h}_t \in \mathbb{R}^{d_h}$; $\mathbf{q}_t = \partial\mathcal{L}/\partial\mathbf{c}_t \in \mathbb{R}^{d_h}$; gate deltas $\boldsymbol{\delta}_t^{(\bullet)} \in \mathbb{R}^{d_h}$.

Summary

From first principles, an LSTM computes gates $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$ and a candidate $\mathbf{g}_t$ to update the cell $\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$ and emit $\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$. Unrolling and applying the chain rule yields exact BPTT with cell-gradient recursion $\mathbf{q}_{t-1} = \mathbf{q}_t \odot \mathbf{f}_t$ (the CEC), gate/candidate deltas via elementwise derivatives, and parameter gradients as outer-product sums. The multiplicative forget gate controls the spectrum of the time Jacobian Eqn. (19), enabling sustained, stable gradient flow and effective long-range credit assignment.