Notes on Edge of Chaos in LSTM

Tommaso Fioratti

1 Introduction

The aim of this work is to better understand how neural networks should be initialized. In particular, it is well known that deep learning models while offering greater expressivity are more difficult to train, as the input signal may fail to propagate through the entire network, or similar inputs may be collapsed into the same representation.
This line of research originates primarily from the seminal work of Sompolinsky et al. [3], who analyzed chaotic dynamics in large random recurrent neural networks via dynamical mean-field theory. These ideas were later extended to deep feedforward networks by Poole et al. [1], and Schoenholz et al. [2].
The common framework underlying all these analyses is the use of the *mean-field approximation*, a concept rooted in statistical physics. The idea is simple and widely used: we want to reduce a complicated system with many degrees of freedom (the neurons of the network) into a simpler one with a single effective degree of freedom. This is done by taking the thermodynamic limit and applying concentration results, such as the law of large numbers and the central limit theorem.
We will later see that the mean-field approximation reveals that neural dynamics can fall into different regimes depending on the initialization of weights and biases. In particular, we will see that neural networks can exhibit either ordered or chaotic dynamics, and that optimal initialization corresponds to the critical point at the phase transition between these two regimes: the so called edge of chaos. In the following we will mainly consider LSTM for their practical use in sequence modeling tasks.

2 LSTM Dynamics

Let us consider the equations defining a standard Long Short-Term Memory (LSTM) recurrent neural network with hidden state dimension N and input dimension d. At each time step $t \in \mathbb{N}$, given an input vector $x_t \in \mathbb{R}^d$ and the previous hidden state $h_{t-1} \in \mathbb{R}^N$, the LSTM updates its hidden and cell states according to the following equations:

$$
\begin{aligned}
f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\
i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\
o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\
\tilde{c}_t &= \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Here, the symbol $\odot$ denotes the Hadamard (element-wise) product, and $\sigma(\cdot)$ is the logistic sigmoid function defined by $\sigma(z) = (1 + e^{-z})^{-1}$. The matrices $W_g \in \mathbb{R}^{N \times d}$, $U_g \in \mathbb{R}^{N \times N}$, and biases

$b_g \in \mathbb{R}^N$ for $g \in \{f, i, o, c\}$ are the learnable parameters associated to the forget, input, output gates, and cell input, respectively.

In what follows, we will analyze the dynamics of the above equations in the absence of external field, i.e., assuming $x_t \equiv 0$, and $b_g = 0$ for $g \in \{f, i, o, c\}$, focusing on the autonomous evolution of the system. Let $i \in \{1, \dots, N\}$ index the coordinates of each hidden unit. For each i, we denote the i-th entry of a vector $v \in \mathbb{R}^N$ by v_i. In particular, in the case specified above we can write:

$$f_{t,i} = \sigma\left([U_f h_{t-1}]_i\right), \quad i_{t,i} = \sigma\left([U_i h_{t-1}]_i\right),$$
$$o_{t,i} = \sigma\left([U_o h_{t-1}]_i\right), \quad \tilde{c}_{t,i} = \tanh\left([U_c h_{t-1}]_i\right),$$

where we have assumed, as before, that all biases and the input x_t are identically zero. The cell and hidden state updates then reduce to the following recursive equations, applied coordinate-wise:

$$c_{t+1,i} = f_{t+1,i} \cdot c_{t,i} + i_{t+1,i} \cdot \tilde{c}_{t+1,i},$$
$$h_{t+1,i} = o_{t+1,i} \cdot \tanh(c_{t+1,i}).$$

We now rewrite the coordinate-wise cell update equation in a factored form that exposes its interpretation as an exponential smoothing filter. We define the total gain:

$$A_{t+1,i} := f_{t+1,i} + i_{t+1,i},$$

and the normalized input gate coefficient

$$\alpha_{t+1,i} := \frac{i_{t+1,i}}{A_{t+1,i}} \in [0, 1].$$

Then, using the identities:

$$\frac{f_{t+1,i}}{A_{t+1,i}} = 1 - \alpha_{t+1,i}, \qquad \frac{i_{t+1,i}}{A_{t+1,i}} = \alpha_{t+1,i},$$

we factor the update as:

$$\begin{aligned} c_{t+1,i} &= A_{t+1,i}\left(\frac{f_{t+1,i}}{A_{t+1,i}} \cdot c_{t,i} + \frac{i_{t+1,i}}{A_{t+1,i}} \cdot \tilde{c}_{t+1,i}\right) \\ &= A_{t+1,i}\left((1 - \alpha_{t+1,i}) \cdot c_{t,i} + \alpha_{t+1,i} \cdot \tilde{c}_{t+1,i}\right). \end{aligned} \tag{1}$$

This shows that the cell update is a convex combination between the previous memory content $c_{t,i}$ and the candidate input $\tilde{c}_{t+1,i}$, scaled by a gain factor $A_{t+1,i}$.

2.1 First–order stationary approximation of the gates

Now we shall assume that for every gate $g \in \{f, i, o, c\}$ the matrix entries $\{U_{g,ij}\}_{i,j=1}^N$ are i.i.d. with

$$\mathbb{E}[U_{g,ij}] = 0, \qquad \mathrm{Var}(U_{g,ij}) = \frac{\sigma_g^2}{N}.$$

Independence is taken both across g and across coordinates. We assume also that the initial hidden state $h_0 \in \mathbb{R}^N$ has i.i.d. centred coordinates, independent of the weights, and $\mathbb{E}[h_{0,i}^2] = q_0 < \infty$. We want to formally justify the following heuristic approximation in the mean-field approximation under the phase transition, i.e. as we take the width of the network to go to infinity and the network is in a stable regime ($\|h_{t-1}\|^2 \le B \quad \forall t$):

$$A_{t,i} \simeq 1, \qquad \alpha_{t,i} \simeq o_{t,i} \simeq \frac{1}{2}$$

Theorem 1 (Asymptotic decorrelation of weights and activations)**.** *Under the standing assumptions, for every gate $g \in \{f, i, o, c\}$, every pair of indices $i \neq j$, and every time step $t \geq 0$, we have*

$$\lim_{N\to\infty} \mathbb{E}\big[U_{g,ij}\, h_{t,j}\big] \;=\; \lim_{N\to\infty} \mathbb{E}[U_{g,ij}] \cdot \mathbb{E}[h_{t,j}] \;=\; 0.$$

Proof. We prove it by induction, we have that h_0 is independent from every matrix weight, so let us suppose that h_{t-1} is independent from U_g.
Let us consider the case for $j \neq i$:

$$U_{g,ij}\, h_{t,j} = U_{g,ij}\, g\Big(\sum_k U_{g,jk}\, h_{t-1,k}\Big)$$

from this we can see that the weight entry $U_{g,ij}$ does not appear in the computation of $h_{t,j}$ and by the independence from the previous time step we can conclude that

$$\mathbb{E}[U_{g,ij} h_{t,j}] = \mathbb{E}[U_{g,ij}]\, \mathbb{E}[h_{t,j}] = 0.$$

For the case $i = j$ we have

$$\mathbb{E}[U_{g,ii}\, h_{t,i}] = \mathbb{E}\Big[U_{g,ii}\, g\Big(\sum_{k\neq i} U_{g,ik}\, h_{t-1,k} + h_{t-1,i}\, U_{g,ii}\Big)\Big]$$

and we exploit the case showed before to add a zero term under expectation, since the expected value factorizes and $U_{g,ii}$ is zero-mean

$$\mathbb{E}[U_{g,ii}\, h_{t,i}] = \mathbb{E}\Big[U_{g,ii}\, \Big(g\Big(\sum_{k\neq i} U_{g,ik}\, h_{t-1,k} + h_{t-1,i}\, U_{g,ii}\Big) - g\Big(\sum_{k\neq i} U_{g,ik}\, h_{t-1,k}\Big)\Big)\Big].$$

At this point by Cauchy-Schwarz:

$$\mathbb{E}[U_{g,ii}\, h_{t,i}] \leq \big(\mathbb{E}[U_{g,ii}^2]\big)^{1/2} \Big(\mathbb{E}\Big[\Big(g\Big(\sum_{k\neq i} U_{g,ik}\, h_{t-1,k} + h_{t-1,i}\, U_{g,ii}\Big) - g\Big(\sum_{k\neq i} U_{g,ik}\, h_{t-1,k}\Big)\Big)^2\Big]\Big)^{1/2},$$

and lipschitz-continuity of the gates

$$\mathbb{E}[U_{g,ii}\, h_{t,i}] \leq \big(\mathbb{E}[U_{g,ii}^2]\big)^{1/2} \big(L^2\, \mathbb{E}[h_{t-1,i}^2 U_{g,ii}^2]\big)^{1/2}.$$

Now rearranging and by independence of the hidden state at the previous temporal step we get:

$$\mathbb{E}[U_{g,ii}\, h_{t,i}] \leq L\, \big(\mathbb{E}[U_{g,ii}^2]\big)\, \big(\mathbb{E}[h_{t-1,i}^2]\big)^{1/2} = L\, \left(\frac{\sigma^2}{N}\right) \big(\mathbb{E}[h_{t-1,i}^2]\big)^{1/2}$$

and this goes to zero as $N \to \infty$ as long as the norm of the hidden state remains bounded.

□

Now we shall note that the sigmoid function satisfies the identity

$$\sigma(-x) = 1 - \sigma(x),$$

and letting s be a random variable with a distribution symmetric around zero, i.e., $s \overset{d}{=} -s$ we get the following

$$\mathbb{E}[\sigma(s)] = \mathbb{E}[\sigma(-s)] = \mathbb{E}[1 - \sigma(s)] = 1 - \mathbb{E}[\sigma(s)].$$

Solving this equation gives

$$\mathbb{E}[\sigma(s)] = \frac{1}{2}.$$

So that if the preactivation is symmetric and it is zero-mean in the mean-field limit (as we have shown above), input, output and forget gate have mean 1/2.
Now the last thing we shall prove is that as N goes to infinity the gate concentrates around its expectation.

Theorem 2 (Concentration of gates around their expected value)**.** *Under the standing assumptions, input, output and forget gate concentrate in probability around* $\frac{1}{2}$*:*

$$\lim_{N\to\infty} \mathbb{P}\left(\left|\sigma(s_{g,i}) - \frac{1}{2}\right| > t\right) = 0$$

Proof. Let the pre-activation of coordinate i be $s_{g,i} = \sum_{j=1}^{N} U_{g,ij} h_{t-1,j}$ where $U_{g,ij} \sim \mathcal{N}(0, \sigma^2/N)$ are i.i.d., and $h_{t-1,j}$ are fixed values. Define $T_j = U_{g,ij} h_{t-1,j}$ for all $j = 1, \dots, N$.
Then, conditioned on h_{t-1}, we have $T_j \mid h_{t-1} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N} h_{t-1,j}^2\right)$, which implies that each T_j is subgaussian with parameter $\lambda_j^2 = \frac{\sigma^2}{N} h_{t-1,j}^2$.
Since the T_j are independent given h_{t-1}, the sum $s_{g,i} = \sum_{j=1}^{N} T_j$ is also subgaussian, with parameter $\lambda_{s_{g,i}}^2 = \sum_{j=1}^{N} \lambda_j^2 = \frac{\sigma^2}{N} \sum_{j=1}^{N} h_{t-1,j}^2 = \frac{\sigma^2}{N} \|h_{t-1}\|_2^2$.
Now we write:

$$\mathbb{E}\left[e^{\theta s_{g,i}}\right] = \mathbb{E}_{h_{t-1}}\left[\mathbb{E}\left[e^{\theta s_{g,i}} \mid h_{t-1}\right]\right] \le \mathbb{E}_{h_{t-1}}\left[\exp\left(\frac{\theta^2\sigma^2}{2N}\|h_{t-1}\|_2^2\right)\right].$$

Now since we are below the phase transition $\|h_{t-1}\|_2^2 \le B$, we get:

$$\mathbb{E}\left[e^{\theta s_{g,i}}\right] \le \exp\left(\frac{\theta^2\sigma^2 B}{2N}\right).$$

Thus, we have shown that the preactivation is subgaussian with parameter $\lambda^2 = \frac{\sigma^2 B}{N}$.
Now by standard tail bound for Lipschitz functions of subgaussian random variables we get:

$$\mathbb{P}\left(|f(s_{g,i}) - \mathbb{E}[f(s_{g,i})]| > t\right) \le 2\exp\left(-\frac{t^2}{2L^2\lambda^2}\right) = 2\exp\left(-\frac{Nt^2}{2L^2\sigma^2 B}\right).$$

In particular, for the sigmoid function $f(x) = \sigma(x)$, which is 1/4-Lipschitz, we obtain

$$\mathbb{P}\left(|\sigma(s_{g,i}) - \mathbb{E}[\sigma(s_{g,i})]| > t\right) \le 2\exp\left(-\frac{8Nt^2}{\sigma^2 C}\right).$$

□

In conclusion, by proving that the preactivation in the thermodynamic limit are mean-zero, symmetric, subagussian random variable, we managed to show that the gates concentrate at 1/2, thus the approximation are justified.

2.2 Langevin-Like Dynamics

We have shown that in the subcritical regime $i_t = f_t = o_t \approx \frac{1}{2}$, now assuming again that the hidden state and cell state are small with high probability (should I ask more specific conditions?), the use of the linear approximations:

$$\tanh(c_t) \approx c_t, \qquad \tanh(h_t) \approx h_t,$$

is justified. Exploiting the following:

$$h_t = o_t \tanh(c_t) \approx \frac{1}{2} c_t,$$

and remembering the definition

$$\tilde{c}_t = \tanh\left(U_c h_{t-1}\right) \approx U_c h_{t-1},$$

we can write from (1)

$$\begin{aligned}
c_{t+1,i} &= A_{t+1,i}\left((1-\alpha_{t+1,i}) \cdot c_{t,i} + \alpha_{t+1,i} \cdot \tilde{c}_{t+1,i}\right), \\
2h_{t+1,i} &= A_{t+1,i}\left((1-\alpha_{t+1,i}) \cdot 2h_{t,i} + \alpha_{t+1,i} \cdot \sum_{j=1}^{N} U_{g,ij} h_{t,j}\right), \\
h_{t+1,i} &= A_{t+1,i}\left((1-\alpha_{t+1,i}) \cdot h_{t,i} + \frac{\alpha_{t+1,i}}{2} \cdot \sum_{j=1}^{N} U_{g,ij} h_{t,j}\right),
\end{aligned}$$

Now by using the concentration of gates

$$h_{t+1,i} - h_{t,i} = \left(-\frac{\cdot h_{t,i}}{2} + \cdot \sum_{j=1}^{N} \frac{U_{g,ij}}{4} h_{t,j}\right),$$

So that we have a discretization of the following Langevin dynamics:

$$\frac{dh_i}{dt} = \left(-\frac{\cdot h_{t,i}}{2} + \cdot \sum_{j=1}^{N} \frac{U_{g,ij}}{4} h_{t,j}\right). \tag{2}$$

which is a linearization of the stochastic differential equation which has been studied in the seminal work of Sompolinsky, Crisanti, and Sommers [3].

2.3 Naif stability observations

From (2) an immediate analysis show us that $h^* = 0$ is a fixed point of the SDE. If the fixed point is (locally) stable, then small perturbations around it decay over time. In this case, for a broad class of initial conditions, the network dynamics will converge to a regime where all hidden states vanish and the network effectively shuts down. Understanding the local stability of this fixed point is thus essential. To that end, we want to understand the dynamics of perturbations around the fixed point $h(t) = h^* + \delta h(t) = 0 + \delta h(t)$.

By writing the equation in vector form we can employ RMT to understand the distribution, and thus the sign, of the eigenvalues of the random matrix U, and consequently of the Jacobian.

$$\delta \dot{h}_t = \frac{1}{2}\left(-I + \frac{1}{2}U_g\right)\delta h_t.$$

Now it is easy to see that the fixed point is locally stable if and only if every eigenvalue of the Jacobian $J := \left(-I + \frac{1}{2}U_g\right)$ has negative real part.
As $N \to \infty$ by circular law we know that the spectrum of U_g fills a disc of radius g in the complex plane; the extreme real part therefore tends to

$$\max_i \operatorname{Re}\mu_i \xrightarrow{N\to\infty} g.$$

So that

$$\max_i \operatorname{Re}\lambda_i = \frac{1}{2}\left(-1 + \frac{g}{2}\right).$$

Stability requires it to be smaller than 0, i.e.

$$-1 + \frac{g}{2} < 0 \quad \Longrightarrow \quad g < 2.$$

Hence the critical gain at which the real part crosses zero is

$$\boxed{g_c = 2}$$

We have proved that in the mean-field limit approximation for $g < g_c$ all eigenvalues satisfy $\operatorname{Re}\lambda_i < 0$ and the fixed point $h_t \equiv 0$ is linearly stable; for $g > g_c$ at least the maximum eigenvalue of J acquires positive real part, i.e. a small perturbation around zero diverges away from the fixed point. Note that a linear system cannot exhibit chaos per se, but we will see later on how this same condition appears in the non-linear dynamics.

3 Dynamical Mean Field Theory

Now from the analysis made in the previous section we have seen that arbitrarily close to the phase transition the LSTM dynamics reduces to the linearized dynamics, but this is a too simple model (sketchy) so we reintroduce back a non-linearity from (2) and obtain the famous model from [3] with the additional scaling factors coming from the gates to gain more insight of what is happening. This corresponds to

$$\dot{h}_i(t) = -\frac{h_i(t)}{2} + \sum_{j=1}^{N} \frac{J_{ij}}{4}\tanh(h_j(t)), \quad \text{with} \quad J_{ij} \sim \mathcal{N}\left(0, \frac{g^2}{N}\right). \tag{3}$$

3.1 No bias

Now, considering a fully connected random neural network, we have N coupled equations of the form (3). To reduce this complexity, we aim to take the thermodynamic limit $N \to \infty$ and invoke a Central Limit Theorem (CLT) argument to prove that the recurrent input converges to an effective Gaussian process. This would enable the derivation of a dynamical mean-field (DMF) equation governing the behavior of a typical unit.

However, this step is more delicate than it might appear. At fixed time t, the variables $h_j(t)$ are random and depend on the disorder realization of the coupling matrix J. Consequently, the terms in the sum $\sum_j J_{ij} \tanh(h_j(t))$ are statistically dependent, and a straightforward application of the classical CLT is not valid.

To address this, we employ a *cavity method*: by analyzing the effect of removing a single neuron from the network, one can argue that the dependence between J_{ij} and $h_j(t)$ induces corrections of order $O(N^{-1/2})$, which become negligible in the thermodynamic limit. This justifies the approximation of the input as a Gaussian process with self-consistent statistics, forming the basis of the DMFT formulation.

Let us consider the evolution equation for a generic neuron $h_\alpha^{(i)}$ in the cavity network, i.e., in the network where neuron i has been removed:

$$\dot{h}_\alpha^{(i)}(t) = -\frac{h_\alpha^{(i)}(t)}{2} + \sum_{j\neq i} \frac{J_{\alpha j}}{4} \tanh\left(h_j^{(i)}(t)\right).$$

Reintroducing neuron i perturbs the input current of neuron α by an amount of order $O(N^{-1/2})$, which we can treat perturbatively:

$$\delta I_\alpha(t) = \frac{J_{\alpha i}}{4} \tanh\left(h_i(t)\right).$$

The resulting change in $h_\alpha(t)$ is given by linear response theory:

$$\delta h_\alpha(t) \approx \int_0^t ds \sum_{k\neq i} G_{\alpha k}^{(i)}(t,s)\, \delta I_k(s) = \int_0^t ds \sum_{k\neq i} G_{\alpha k}^{(i)}(t,s)\, \frac{J_{ki}}{4} \tanh\left(h_i(s)\right),$$

where the retarded Green's function of the cavity dynamics is defined as

$$G_{\alpha k}^{(i)}(t,s) := \frac{\delta h_\alpha^{(i)}(t)}{\delta I_k(s)}.$$

We can now expand the full dynamics of neuron i around the cavity fields:

$$\dot{h}_i(t) = -\frac{h_i(t)}{2} + \sum_{j=1}^N \frac{J_{ij}}{4} \tanh\left(h_j(t)\right) = -\frac{h_i(t)}{2} + \sum_{j=1}^N \frac{J_{ij}}{4} \tanh\left(h_j^{(i)}(t) + \delta h_j(t)\right).$$

Expanding the nonlinearity to first order:

$$\dot{h}_i(t) = -\frac{h_i(t)}{2} + \underbrace{\sum_{j=1}^N \frac{J_{ij}}{4} \tanh\left(h_j^{(i)}(t)\right)}_{\text{bulk term}} + \underbrace{\sum_{j=1}^N \frac{J_{ij}}{4} \tanh'\left(h_j^{(i)}(t)\right) \delta h_j(t)}_{\text{memory term}} + \text{higher order terms}.$$

Now in our case, where the random matrix has no structure and all entries are independent, the memory term is zero in the thermodynamic limit:

$$\text{Memory term} = \frac{1}{16} \int_0^t ds\ \tanh\left(h_i(s)\right) \sum_{j=1}^N \sum_{k\neq i} J_{ij} J_{ki}\ \tanh'\left(h_j^{(i)}(t)\right)\ G_{jk}^{(i)}(t,s)$$

since its expectation is zero and its variance goes to zero as N goes to infinity. From a physical point of view, this reflects the absence of structural correlations between the recurrent input received by a neuron and the feedback it induces on the rest of the network. In the fully asymmetric case, the path by which a perturbation propagates from neuron i to neuron j, and then returns via k to i, carries no coherent information back to neuron i. As a result, all such feedback contributions interfere incoherently and self-average to zero, leaving no effective memory in the dynamics.
The only term is the bulk one:

$$\text{Bulk term} = \sum_{j=1}^{N} \frac{J_{ij}}{4} \tanh\left(h_j^{(i)}(t)\right)$$

Since now the terms in the sum are independent, the bulk term is a gaussian process with mean zero and covariance $\Sigma(t,s)$:

$$\Sigma^{(i)}(t,s) = \frac{g^2}{16}\frac{1}{N}\sum_{j=1}^{N} \tanh\Big(h_j^{(i)}(t)\Big)\tanh\Big(h_j^{(i)}(s)\Big)$$

and we assume that in the limit this quantity is self averaging and converges to the mean (according to the law of the limiting process h_i) of the product of the terms:

$$\lim_{N\to\infty} \Sigma^{(i)}(t,s) = \frac{g^2}{16}\mathbb{E}[\tanh(h_i(t))\tanh(h_i(s))]$$

where $\mathbb{E}[\cdot]$ stands for the expected value wrt to the law of h_i.
Thus we have the following result for the no bias dynamics [3]:

$$\dot{h}(t) = -\frac{1}{2}h(t) + \eta(t), \quad \eta(t) \sim \mathcal{GP}\left(0, \frac{g^2}{16}\Delta(t,s)\right), \tag{4}$$

$$\text{with} \quad \Delta(t,s) = \mathbb{E}\left[\tanh(h(t))\tanh(h(s))\right]$$

where we have dropped the index of the neuron since it holds for every neuron.

3.2 Memory of the Network

Solving (4) formally yields:

$$h(t) = \int_{-\infty}^{t} du\, e^{-(t-u)/2}\,\eta(u)$$

$$\mathbb{E}[h(t)h(s)] = \int_{-\infty}^{t} du \int_{-\infty}^{s} dv\, e^{-(t-u)/2}\, e^{-(s-v)/2}\,\mathbb{E}[\eta(u)\,\eta(v)]$$

Now noting that the exponential kernel is the green function of the operator $\partial_t + \lambda$ this can be equivalently written as:

$$\left(\partial_t + \frac{1}{2}\right)\left(\partial_s + \frac{1}{2}\right) C(t,s) = \frac{g^2}{16}\,\Delta(t,s)$$

where $C(t,s) := \mathbb{E}[h(t)h(s)$. We assume now stationarity, i.e. $C(t,s) = C(|t-s|) = C(\tau)$ so that by changing variables we obtain:

$$(-\partial_\tau^2 + \frac{1}{4})C(\tau) = \frac{g^2}{16}\Delta(\tau)$$

Now we rearrange $\Delta(\tau)$ in such a way to explicit is dependence on $C(\tau)$. By definition we have:

$$\Delta(\tau) = \frac{g^2}{16} \cdot \frac{1}{2\pi\sqrt{\det \Sigma(\tau)}} \iint_{\mathbb{R}^2} \tanh(h(t))\, \tanh(h(t+\tau)) \times \exp\left(-\frac{1}{2}\begin{pmatrix} h(t) \\ h(t+\tau) \end{pmatrix}^{\top} \Sigma(\tau)^{-1} \begin{pmatrix} h(t) \\ h(t+\tau) \end{pmatrix}\right) dh(t)\, dh(t+\tau)$$

By changing variables through the cholesky decomposition of Σ

$$\Delta(C(\tau), C_0) = \frac{g^2}{16} \iint Dz_1\, Dz_2\ \tanh\left(\frac{C(\tau)}{\sqrt{C_0}}\, z_2 + \sqrt{C_0 - \frac{C(\tau)^2}{C_0}}\, z_1\right) \cdot \tanh\left(\sqrt{C_0}\, z_2\right) \tag{5}$$

So that now we have the following non-linear ODE for the correlation function:

$$(-\partial_\tau^2 + \frac{1}{4})C(\tau) = \frac{g^2}{16}\Delta(C(\tau), C_0) \tag{6}$$

Where C_0 can be deduced self-consistently from the same ODE:

$$C_0 = \frac{g^2}{4} \int_{-\infty}^{\infty} Dz\ \tanh^2(\sqrt{C_0}z)$$

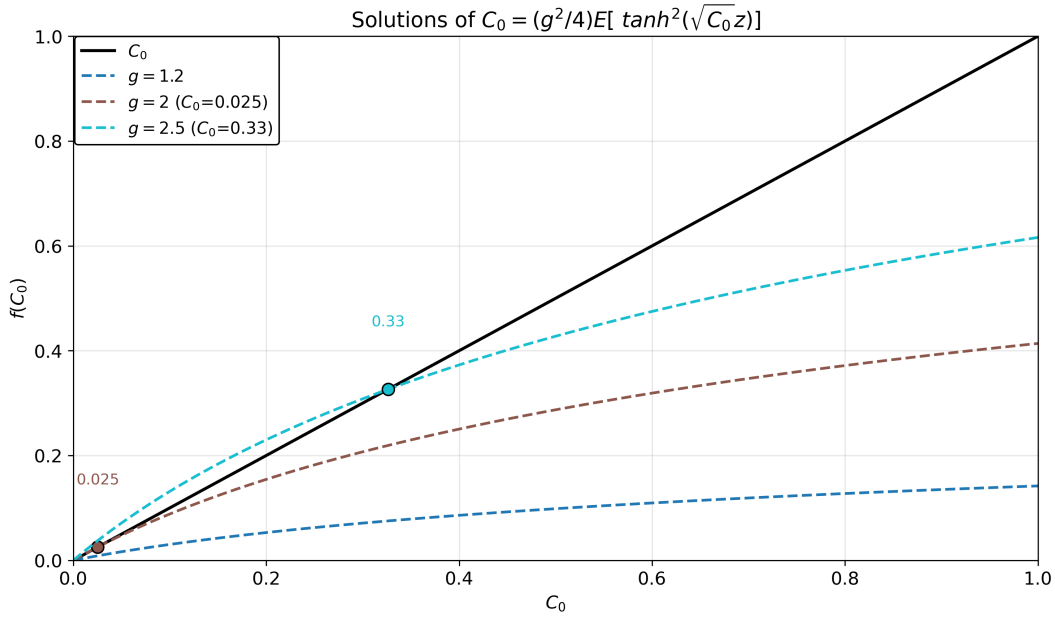


Figure 1: Variance solutions C_0 e $f_g(C_0)$ for different g

Reminding the meaning of C_0 this means that if we look at the hidden states of the LSTM after a temporary transient and we have $g < 2$ all the h will have collapsed on the zero fixed point, thus the network turns off and there is no dynamics left, i.e. the autocorrelation is identically zero. For a non-zero fixed point, i.e. above the threshold $g = 2$ instead we have a non-zero stationary variance of the hidden states of the network.

Now we can analyze equation (6) expanding the gaussian integral in a neighbourhood of C_0 to understand for example the linear stability of the fixed point.

$$\Delta(C(\tau), C_0) = \Delta(C_0, C_0) + \left(\mathbb{E}_z\left[\text{sech}^4(\sqrt{C_0}z)\right]\right)^{\cdot}(C(\tau) - C_0) + o(C(\tau) - C_0)$$

and by substituting $C(\tau) = C_0 + \delta C(\tau)$:

$$\left(-\partial_\tau^2 + \frac{1}{4}\right)(C_0 + \delta C(\tau)) = \frac{g^2}{16}\big(\Delta(C_0, C_0) + \left(\mathbb{E}_z\left[\text{sech}^4(\sqrt{C_0}z)\right]\right)^{\cdot}\delta C(\tau)) + o(\delta C(\tau))\big)$$

and since the fixed point satisfies the equation we get the following linearized dynamics:

$$\left(-\partial_\tau^2 + \frac{1}{4}\right)\delta C(\tau) = \frac{g^2}{16}\left(\mathbb{E}_z\left[\text{sech}^4(\sqrt{C_0}z)\right]\right)^{\cdot}\delta C(\tau)$$

with eigenvalues

$$\lambda^2 = \frac{1}{4} - \frac{g^2}{16}\mathbb{E}_z\left[\text{sech}^4(\sqrt{C_0}z)\right]$$

and relaxation time:

$$\tau_\infty = \frac{2}{\sqrt{(1 - \frac{g^2}{4}\,\mathbb{E}_z\left[\text{sech}^4(\sqrt{C_0}z)\right])}}$$

which diverges iff

$$g_c = \frac{2}{\sqrt{\mathbb{E}_z\left[\text{sech}^4(\sqrt{C_0}z)\right]}} \tag{7}$$

which is consistent with what we obtained by approaching from below the phase transition in the linearized model (2). Now this condition can be actually shown to predict also chaotic dynamics (one should see at the Schrodinger equation for the lyapunov exponent [3].

3.3 bias

Let us now add a constant (in time) bias $b \sim \mathcal{N}(0, \sigma_b^2)$ for the single neuron Langevin-dynamics we deduced before. This guarantees that we preserve the zero-mean and the gaussianity of the process.

$$\dot{h}(t) = -\frac{1}{2}h(t) + b + \eta(t),$$

Formally solving the SDE yields

$$h(t) = \int_{-\infty}^{t} du\, e^{-\frac{(t-u)}{2}}\left[b + \eta(u)\right] = 2b + \underbrace{\int_{-\infty}^{t} du\, e^{-\frac{(t-u)}{2}}\eta(u)}_{\delta h(t)}, \tag{8}$$

Now note that the perturbation follows the dynamics without bias:

$$\dot{\delta h}(t) = -\frac{1}{2}\delta h(t) + \eta(t)$$

so that it is clear that the dynamics of the process is the same, as the covariance of the bias dynamics is just a translation of the covariance of the free-bias dynamics.

$$C^{tot}(t, s) = \mathbb{E}[h(t)h(s)] = 4\sigma_b^2 + C(t, s), \qquad C(t, s) \equiv \mathbb{E}[\delta h(t)\,\delta h(s)].$$

From before we know that the covariance of the free-bias satisfies the following self-consistent equation.

$$C(t,s) = \int_{-\infty}^{t} du \int_{-\infty}^{s} dv\; e^{-\frac{(t-u)}{2}} e^{-\frac{(s-v)}{2}}\, \mathbb{E}[\eta(u)\eta(v)].$$

And by the same trick as before:

$$(\partial_t + \tfrac{1}{2})(\partial_s + \tfrac{1}{2})\, C^{tot}(t,s) = \sigma_b^2 + (\partial_t + \tfrac{1}{2})(\partial_s + \tfrac{1}{2})\, C(t,s) = \sigma_b^2 + \frac{g^2}{16}\Delta(t,s)$$

Changing variables we get:

$$(-\partial_\tau^2 + \frac{1}{4})C(\tau) = \frac{g^2}{16}\Delta(C(\tau), C_0) + \sigma_b^2 \tag{9}$$

From which we deduce the fixed point of the variance:

$$C_0 = \frac{g^2}{4}\int_{-\infty}^{\infty} Dz\, \tanh^2(\sqrt{C_0}z) + 4\sigma_b^2 \tag{10}$$

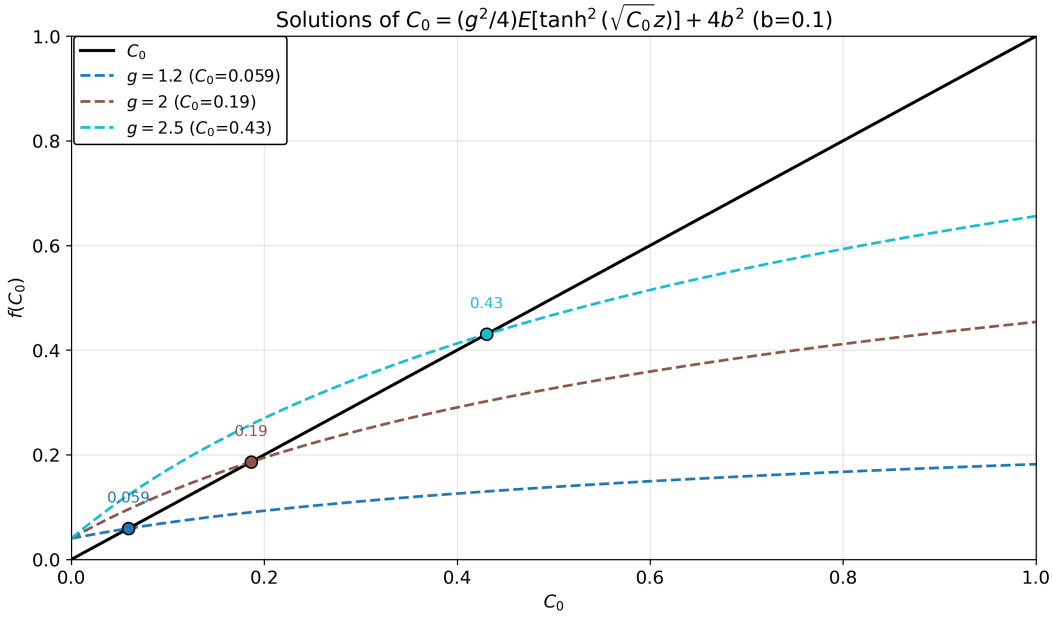


Figure 2: Variance fixed points for different g

Now there are two important things to note: the first one is that the presence of the bias creates always an intersection for every value of g, this means that the dynamics cannot turn off as in the case before under the threshold which means that the bias can be seen as an external magnetic field (analogy with spin glass). Secondly, the structure of the problem for quenched bias remains basically the same apart from the change in the stationary variance but all the considerations made above still hold. In particular (7) still holds with a new C_0 given by (10).

$$g_c \;=\; \frac{2}{\sqrt{\mathbb{E}_Z\left[\operatorname{sech}^4\left(\sqrt{C_0}\,Z\right)\right]}}, \qquad Z \sim \mathcal{N}(0,1).$$

Now we can taylor expand for small biases (thus small C_0) to understand the functional dependency of g_c with the bias.

$$\begin{cases} C_0 = \dfrac{g_c^2}{4}\,\mathbb{E}_z\left[\tanh^2(\sqrt{C_0}z)\right] + 4\sigma_b^2, \\ 1 = \dfrac{g_c^2}{4}\,\mathbb{E}_z\left[\operatorname{sech}^4(\sqrt{C_0}z)\right], \end{cases}$$

By taylor expansion of tanh and sech and the first sixth moments of a standard gaussian ($\mathbb{E}[z^k] = (2k-1)!!$ for even powers) we get:

$$\tanh^2 u = u^2 - \frac{2}{3}u^4 + \frac{17}{45}u^6 + O(u^8),$$

$$\mathbb{E}_z[\tanh^2(\sqrt{C_0}z)] = C_0 - 2C_0^2 + \frac{17}{3}C_0^3 + O(C_0^4),$$

$$\operatorname{sech}^4 u = 1 - 2u^2 + \frac{7}{3}u^4 - \frac{94}{45}u^6 + O(u^8),$$

$$\mathbb{E}_z[\operatorname{sech}^4(\sqrt{C_0}z)] = 1 - 2C_0 + 7C_0^2 - \frac{94}{3}C_0^3 + O(C_0^4),$$

Thus, we have the following equation:

$$\frac{g_c^2}{4} = \frac{1}{1 - 2C_0 + 7C_0^2 - \frac{94}{3}C_0^3 + O(C_0^4)} = 1 + 2C_0 - 3C_0^2 + \frac{50}{3}C_0^3 + O(C_0^4),$$

and using the autoconsistent equation for C_0

$$C_0 = \left(1 + 2C_0 - 3C_0^2 + \dots\right)\left(C_0 - 2C_0^2 + \frac{17}{3}C_0^3 + \dots\right) + 4\sigma_b^2, = C_0 - \frac{4}{3}C_0^3 + O(C_0^4) + 4\sigma_b^2,$$

From which we deduce

$$C_0 = 3^{1/3}\sigma_b^{2/3} + O(\sigma_b^{4/3}),$$

now back to g_c

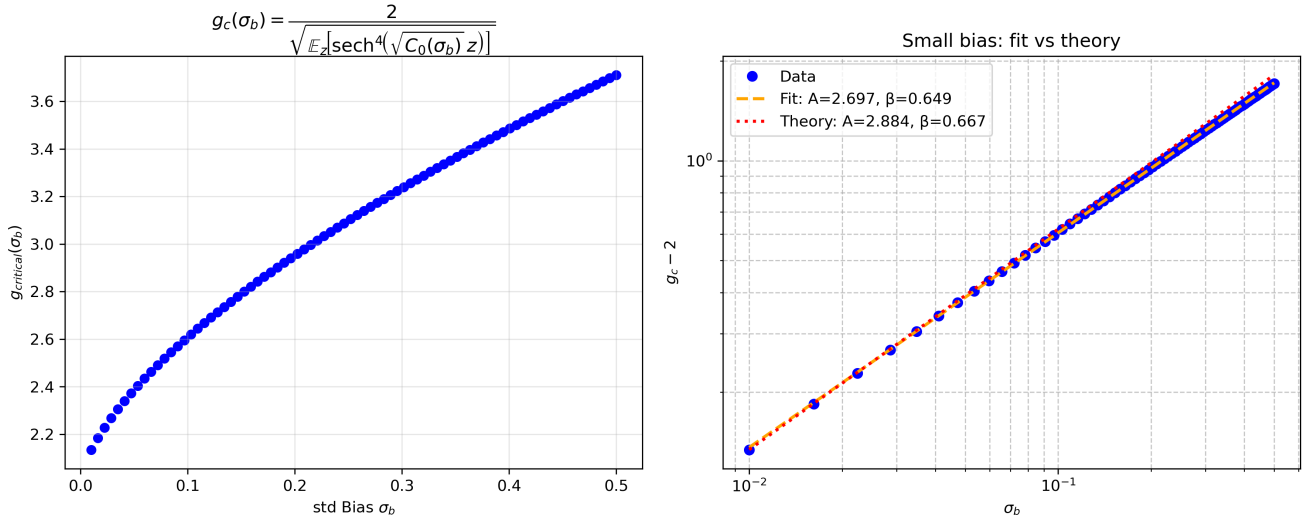
$$\frac{g_c^2}{4} = 1 + 2C_0 - 3C_0^2 + O(C_0^3),$$

$$g_c = 2\sqrt{1 + 2C_0 - 3C_0^2 + O(C_0^3)},$$

$$g_c = 2 + 2C_0 + O(C_0^2),$$

$$g_c - 2 = 2 \cdot 3^{1/3}\sigma_b^{2/3} + O(\sigma_b^{4/3}), \approx 2.8845 \cdot \sigma_b^{2/3}$$

forse dovrei mediare su piu risoluzioni autoconsistenti per maggior precisione



4 Chaos and Lyapunov exponents

1) è vero che LSTM è sompolinsky? 2) dimostrare che condizione chaos = condizione lineare instabilità

References

[1] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[2] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *International Conference on Learning Representations (ICLR)*, 2017.

[3] Haim Sompolinsky, Andrea Crisanti, and Hermann J. Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262, 1988.