Edge of Chaos in LSTM

Tommaso Fioratti

1 Introduction

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 $i \in \{1, \ldots, N\}$ 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:

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

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:

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

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 ($\|h_{t-1}\|^2 \leq B \quad \forall t$):

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

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:

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

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

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

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:

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

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

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

□

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:

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

which has been studied in the seminal work of Sompolinsky, Crisanti, and Sommers [1]

3 Stability and Lyapunov exponents

In this section, we analyze the stability properties of the Langevin dynamics derived in the mean-field limit, focusing on the behavior of small perturbations around the fixed points. Our goal is to characterize how the spectral properties of the random weight matrix U_g influence the evolution of the system and determine the onset of instability.

3.1 No Bias

We can rewrite the scalar equation in vectorial form by introducing the vectors $h_t \in \mathbb{R}^N$ and the weight matrix $U_g \in \mathbb{R}^{N\times N}$. Then the equation becomes:

$$
\frac{dh_t}{dt} = \frac{1}{2}\left(-I + \frac{1}{2}U_g\right) h_t.
$$

Now it is easy to see that $h_t \equiv 0$ is a fixed point and it is stable iff every eigenvalue of the Jacobian J 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

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

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 (maximum lyapunov exponent) acquires positive real part, i.e. a small perturbation around zero diverges away from the fixed point, thus the system enters the chaotic regime.

3.2 Bias

Let us now consider naively (since the derivation before assumed no bias) the following model:

$$\frac{dh_i}{dt} = -\frac{1}{2}h_i(t) + \sum_{j=1}^{N} \frac{U_{g,ij}}{4} \tanh(h_j(t)) + b_i$$

In vector form, this becomes:

$$\dot{\boldsymbol{h}}(t) = -\frac{1}{2}\boldsymbol{h}(t) + \frac{1}{4}U_g \tanh(\boldsymbol{h}(t)) + \boldsymbol{b}$$

The fixed point $\boldsymbol{h}^*$ satisfies:

$$\boldsymbol{h}^* = \frac{1}{2}U_g \tanh(\boldsymbol{h}^*) + 2\boldsymbol{b} \tag{2}$$

Linearizing around the fixed point, we obtain the dynamics:

$$\dot{\boldsymbol{\delta h}}(t) = \left(-\frac{1}{2}I + \frac{1}{4}U_g \cdot \operatorname{diag}\left[1 - \tanh^2(\boldsymbol{h}^*)\right]\right)\boldsymbol{\delta h}(t)$$

We define the Jacobian matrix as:

$$J = -\frac{1}{2}I + \frac{1}{4}U_g - \frac{1}{4}U_g \cdot \operatorname{diag}[\tanh^2(\boldsymbol{h}^*)] \tag{3}$$

From the fixed point equation (2), we isolate:

$$\tanh(\boldsymbol{h}^*) = 2U_g^{-1}(\boldsymbol{h}^* - 2\boldsymbol{b}) \Rightarrow \tanh^2(\boldsymbol{h}^*) = 4U_g^{-2}(\boldsymbol{h}^* - 2\boldsymbol{b})^2$$

Substituting into (3), we get:

$$J = -\frac{1}{2}I + \frac{1}{4}U_g - U_g \cdot \operatorname{diag}\left[U_g^{-2}(\boldsymbol{h}^* - 2\boldsymbol{b})^2\right] \tag{4}$$

In the mean-field limit, assume that the components of $(\boldsymbol{h}^* - 2\boldsymbol{b})^2$ concentrate around their expectation. Then $h_i^* \sim \mathcal{N}(2b, \sigma_h^2)$, where σ_h satisfies the following self consistent equation (take variance of (2):

$$\sigma_h^2 = \frac{g^2}{4}\,\mathbb{E}_z[\tanh^2(2b + \sigma_h z)] \tag{5}$$

Using a first-order approximation:

$$\mathbb{E}(\tanh^2(2b + \sigma_h z)) \approx 4b^2 + \sigma_h^2$$

Substituting into (5), we obtain:

$$\sigma_h^2 = \frac{g^2}{4}(4b^2 + \sigma_h^2)$$

Solving for σ_h^2:

$$\sigma_h^2 = \frac{g^2 b^2}{1 - \frac{g^2}{4}} \tag{6}$$

Now, heuristically approximating the largest eigenvalue of J with a mean-field scalar form:

$$\lambda_{\max}(J) \approx -\frac{1}{2} + \frac{g}{4} - \frac{g}{g^2}\sigma_h^2 \tag{7}$$

Substituting (6) into (7):

$$\lambda_{\max}(J) = -\frac{1}{2} + \frac{g}{4} - \frac{gb^2}{1 - \frac{g^2}{4}} \tag{8}$$

Ora per piccoli b dove vale la nostra analisi dovremmo sempre avere $\lambda_{max} < 0$. L'unica cosa che riesco a dire che però non è abbastanza è che se assumiamo che nel limite termodinamico h^* sia 2b (è troppo semplificata sta cosa) allora abbiamo che

$$J := -\frac{1}{2}I + \frac{1}{4}U_g D, \quad \text{with } D := \operatorname{diag}\left(\operatorname{sech}^2(2b)\right)$$

quindi abbiamo che nel limite

$$\lambda_{max} = -\frac{1}{2} + \frac{g}{4} sech^2(2b) = 0 \implies g_c = \frac{2}{sech^2(2b)} \approx \frac{2}{1 - 4b^2}$$

4 Dynamical Mean Field Theory

4.1 No bias

By Sompolinsky et al. we have the following result for the no bias dynamics:

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

$$\text{with} \quad \Delta(t,s) = \frac{g^2}{16}\mathbb{E}\left[\tanh(h(t))\tanh(h(s))\right]$$

Solving 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 for small h $\Delta(t,s) = \frac{g^2}{16}\,\mathbb{E}\left[\tanh(h(t))\tanh(h(s))\right] \approx \frac{g^2}{16}\,\mathbb{E}\left[h(t)h(s)\right]$ so that we can write

$$\Delta(t,s) = \frac{g^2}{16}\int_0^t du \int_0^s dv\, e^{-(t-u)/2}\, e^{-(s-v)/2}\Delta(u,v)$$

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

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

Now from this last equation we can analyze the phase transition, we are interested in the evolution of the quantity $\Delta(t,t)$ that describes the temporal evolution of the power of the signal.

$$\left(\partial_t^2\Delta + \partial_t\Delta + (\frac{1}{4} - \frac{g^2}{16})\right)(t) = 0$$

So that plugging an exponential ansatz we get

$$\lambda_{max} = -\frac{1}{2} + \frac{g}{4}$$

and setting it to zero we get the known condition $g = 2$

4.2 bias

Let us now add a costant bias for the single neuron Langevin-dynamics we deduced before.

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

Formally solving the SDE yields

$$h(t) = \int_{-\infty}^{t} du\, e^{-\frac{(t-u)}{2}}\left[b + \eta(u)\right] = 2b + \underbrace{\int_{-\infty}^{t} du\, e^{-\frac{(t-u)}{2}}\eta(u)}_{\delta h(t)}, \tag{9}$$

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.

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

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

$$C(t,s) = \int_{-\infty}^{t} du \int_{-\infty}^{s} dv\; e^{-\frac{(t-u)}{2}} e^{-\frac{(s-v)}{2}}\,\mathbb{E}[\eta(u)\eta(v)]. \tag{10}$$

Now we cannot expand anymore *tanh* around zero, since the bias shifts the fixed point of the SDE. We expand now near the new fixed point yielding: $\mathbb{E}[\eta(u)\eta(v)] \approx \frac{g^2}{16}[(\mathbb{E}[h^{*2}] + \mathbb{E}[sech^2(h^*)]C(u,v)]$ And by the same trick as before:

$$(\partial_t + \tfrac{1}{2})(\partial_s + \tfrac{1}{2})\, C(t,s) = \frac{g^2}{16}[(\mathbb{E}[h^{*2}] + \mathbb{E}[sech^2(h^*)]C(u,v)]. \tag{11}$$

Now, the forcing costant term does not change the dynamics of the process, so that we can focus on analyzing the homogeneous ODE ($t = s$) and obtain:

$$\partial_t^2 \tilde{C}(t,s) + \partial_t \tilde{C}(t,s) + (\frac{1}{4} - \frac{g_{\text{eff}}^2}{16})\tilde{C}(t,s) = 0 \quad \text{con } g_{\text{eff}} \equiv g\,\sqrt{\mathbb{E}[sech^2(h^*).]} \tag{10}$$

From which the maximum lyapunov exponent is

$$\lambda_{max} = -\frac{1}{2} + \frac{g_{eff}}{4}.$$

If b and σ_h are small and of the same order of magnitude we can expand as follows:

$$\mathbb{E}(sech^2(2b + \sigma_h z) = 1 - \int tanh^2(2b + \sigma_h z) Dz \approx 1 - (2b^2 + \sigma_h^2)$$

where σ_h satisfies

$$\sigma_h^2 = \frac{g^2}{4}\,\mathbb{E}_z[\tanh^2(2b + \sigma_h z)]$$

which in the same approximation of above yields

$$\sigma_h^2 = \frac{g^2 b^2}{1 - \frac{g^2}{4}}$$

so that we have

$$\begin{aligned}\lambda_{\max} &= -\frac{1}{2} + \frac{g}{4}\sqrt{1 - \left(2b^2 + \frac{g^2 b^2}{1 - \frac{g^2}{4}}\right)} = -\frac{1}{2} + \frac{g}{4}\sqrt{1 - b^2\left[2 + \frac{g^2}{1 - \frac{g^2}{4}}\right]} \\ &= -\frac{1}{2} + \frac{g}{4}\left[1 - \frac{1}{2}b^2\left(2 + \frac{g^2}{1 - \frac{g^2}{4}}\right) + \mathcal{O}(b^4)\right] = -\frac{1}{2} + \frac{g}{4} - \frac{g(4 + g^2)}{4(4 - g^2)}b^2 + \mathcal{O}(b^4),\end{aligned}$$

This analysis tells us that for small biases the role of the bias is the one of shifting the eigenvalues of the Jacobian by a negative term which actually becomes dominant as g approaches the old critical value. This means that at first order approximation the fixed point of the dynamics is always stable and the phase transition is smoothed out.

References

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