Local dynamics in trained recurrent Neural networks - Supplemental Material

Alexander Rivkind and Omri Barak

Local Fading Memory Property

The main text assumes stability of the open loop system, both to guarantee the existence of the fixed point $\bar{x}$, and to justify the application of the Nyquist criterion. Here we argue that this condition holds for networks trainable by Echo State protocol [1] or by FORCE [2].

The terms "echo state property" and "fading memory property" coined in [1] refer to the requirement for globally convergent dynamics of the reservoir network. This property was originally claimed to be a necessary condition for reservoir network training to succeed. However, Sussillo and Abbott [2] showed that intrinsically chaotic networks ($g > 1$) can be successfully trained as well. In principle, there are two hypothetical mechanisms by which a rank one perturbation $w_{FB}w_{out}^T$, added to a stochastic matrix W with spectral radius greater than unity, might enable the formation of a stable trajectory:

1. Linear: State feedback stabilization [3] of unstable eigenvalues of $W - I$ by $w_{FB}w_{out}^T$. The spectrum of W can be arbitrarily modified (and, in particular, stabilized, similarly to [4]) by a suitable w_{out}, subject to the controllability matrix $\mathscr{C} \triangleq [w_{FB}, Ww_{FB}, W^2w_{FB}, ...W^{N-1}w_{FB}]$ being full rank. For W independent of w_{FB} and finite N, one can conjecture that the rank of $\mathscr{C}$ is full and controllability condition holds. However, $\mathscr{C}$ is expected to be very non-regular for large N, resulting in nonregularity of the resulting stabilizing readout w_{out}.

2. Non-linear**:** The trajectory obtained by training is such that chaos is suppressed in the open loop system [5, 6]. While the spectral radius of W exceeds unity, characteristic amplitudes of the state vector elements are large enough to keep the derivatives of $r(x)$ low and linearization of $Wr(x(t))$ w.r.t. $x(t)$ stable.

According to our numerical investigation, only mechanism #2 contributes to stabilization, namely, chaos *must* be suppressed along the open loop trajectory for the training to succeed completely. This is true for both online update of w_{out} (FORCE algorithm [2]) and a batch update [1]. If chaos is not suppressed in the reservoir, the output will either be noisy or totally lose its relation to the target. Thus the assumption of Open-loop stability is valid for the training techniques of both [1] and [2].

Figure 1 shows that in our case of fixed point attractors, successful training of the full network (main text eq. 1) by FORCE, is possible only if the *open loop* system (main text eq. 2) is stable at the target point A (here the training is for a *single* target fixed point). We also verified this behavior for more complex patterns - the combinations of sine waves used in [2]. Here, we demonstrate that training by FORCE does not reduce the fraction of chaotic (non-periodic) activity in the internal state $x(t)$, compared to open loop network (main text eq. 2) driven by the target pattern f_t (Figure 2). The output error $\left\langle (z(t) - f_t(t))^2 \right\rangle$ is also non-zero as long as chaotic activity persists in the reservoir. Here, following [7], we quantify the amount chaos in state dynamics $x(t)$, by the non-periodic fraction of the autocorrelation function $C(\tau) = \langle x^T(t)x(t+\tau) \rangle$ ([7], Fig. 3a).

Remarkably, for simple output patterns, e.g. a fixed point or a pure sine wave, we observe that training does succeed to some extent even in the chaotic regime, and output, while noisy, visually resembles the target. Formal characterization of such a partial success can be of a great interest, but falls out of the scope of the current study.

Derivation of open loop gain for multiple fixed points

In analysis of large random RNN [8], products of a form $\sum_{j=1}^{N} W_{ij}\phi(x_j(t))$ were approximated by a random field $\eta_i(t)$ independent of the site potential $x_i(t)$. In general, given arbitrary vectors $a, b \in \mathbb{R}^N$ and the ensemble of random Gaussian matrices W, second order statistics of $a' = Wa$ and $b' = Wb$ are given by: $a'_i \sim \mathscr{N}(0, g^2N^{-1}a^Ta)$ and $\mathbb{E}\left(b^TW^TWa\right) = g^2b^Ta$. Furthermore, the elements a'_i, b'_i are *jointly* Gaussian. Assuming self averaging properties of W, the element-wise distribution are the same as the ensemble ones, and covariance is given by $< a', b' >= N^{-1}a'^Tb' = g^2 < a, b >$.

In our derivations joint second order statistics of two complex vectors are not required, however joint statistics of a real and a complex vector are extensively used. In the case of one of the vectors being complex, i.e. $c \in \mathbb{C}^N$ covariance with a real vector $a \in \mathbb{R}^N$ can be represented by a complex scalar, rather than by a matrix. Namely, one defines $q \in \mathbb{C}$ such that $Re\{q\} =< Re\{c\}, a >$ and $Im\{q\} =< Im\{c\}, a >$. Furthermore for zero-mean a, c, it is convenient

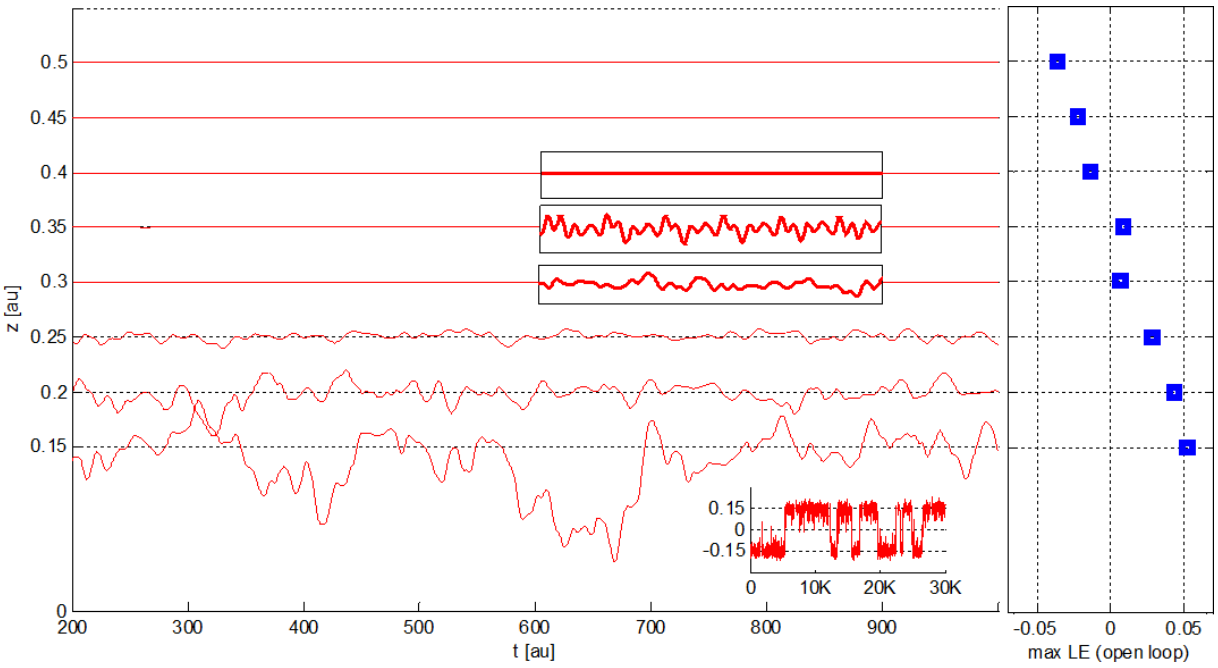

Figure 1: Results of FORCE training targeting fixed a single point $A = 0.15, 0.20, ...0.5$. Maximal Real Lyapunov exponent in the corresponding open loop system is shown on right. Insets provide zoom for graphs with fluctuations too small to be distinguishable at the main panel. Bottom inset shows that a longer simulation time for target $A = 0.15$.

to define $\alpha = q < a, a >^{-1}$ and then use notation of $c_{\|} = \alpha a$ and $c_{\perp} = c - c_{\|}$. The latter notation of orthogonality makes sense since one can easily verify that $< c_{\perp}, a >= 0$.

The readout in multiple fixed point case is given by $w_{out} = N^{-1} \sum_{n=1}^{M} k_n \bar{r}_n$. The coefficient vector k is derived from $k = (A_1...A_M)C^{-1}$ with $(A_1...A_M)$ denoting a row vector composed from the desired output values and the correlation matrix C is given by $C_{nm} = N^{-1} \bar{r}_n^T \bar{r}_m \approx g^{-2} c_{nm} \sigma_n \sigma_m$, with c_{nm}, σ_n and σ_m representing the second order statistics of elements of $\bar{x}$. Given variances σ_n^2, σ_m^2 of states, associated with fixed points $z(t) \equiv A_n$, $z(t) \equiv A_m$ and obtained according to (main text eq. 8), one can use self consistency to find the correlation coefficient c_{nm}, namely, we have:

$$c_{nm}\sigma_n\sigma_m =< \bar{x}_m, \bar{x}_n >= g^2 < \bar{r}_n, \bar{r}_m >= \\ = g^2 \int \mathscr{D}w Dy_1 Dy_2 \phi\left(wA_n + \sigma_n y_1\right) \times \\ \times \phi\left(wA_m + \sigma_m\left(c_{nm}y_1 + \sqrt{1-c_{nm}^2}y_2\right)\right) \quad (1)$$

and coefficients k_m for $1 \leq m \leq M$ follow.

Note that for $A_m = A_n$, the complete correlation, $c_{mn} = 1$ is always a solution of (1). If this solution is unique then uniqueness of a steady state $\bar{x}_n$ in an open loop system, driven by $f = A_n$ follows. For $\phi(x) = \tanh(x)$ we are not aware of a case where an additional solution for (1) emerges.

Next, we compute the projections $G_{nm}(\omega) =< \bar{r}_m^T R_n' X_n(\omega) >$. Similarly to $G_{00}(\omega)$ of (main text eq. 12) we have $G_{nn} = (1 + i\omega)\alpha_n(\omega)\sigma_n^2$. To proceed with G_{nm} $(m \neq n)$, we decompose $\bar{x}_m$ as in (1) and decompose X_n accordingly: $X_n = X^0 + \alpha_n \sigma_n y_1 + \alpha_{mn}\sigma_m y_2 + X_{n\perp}^1$(here y_1, y_2 are i.i.d unity variance Gaussian vectors). Using

$$c_{nm}\sigma_m y_1 + \sqrt{1-c_{nm}^2}\sigma_m y_2 = \bar{x}_m^1 = W\bar{r}_m$$

and

$$(1 + i\omega)X_n^1 = WR_n' X_n$$

we have:

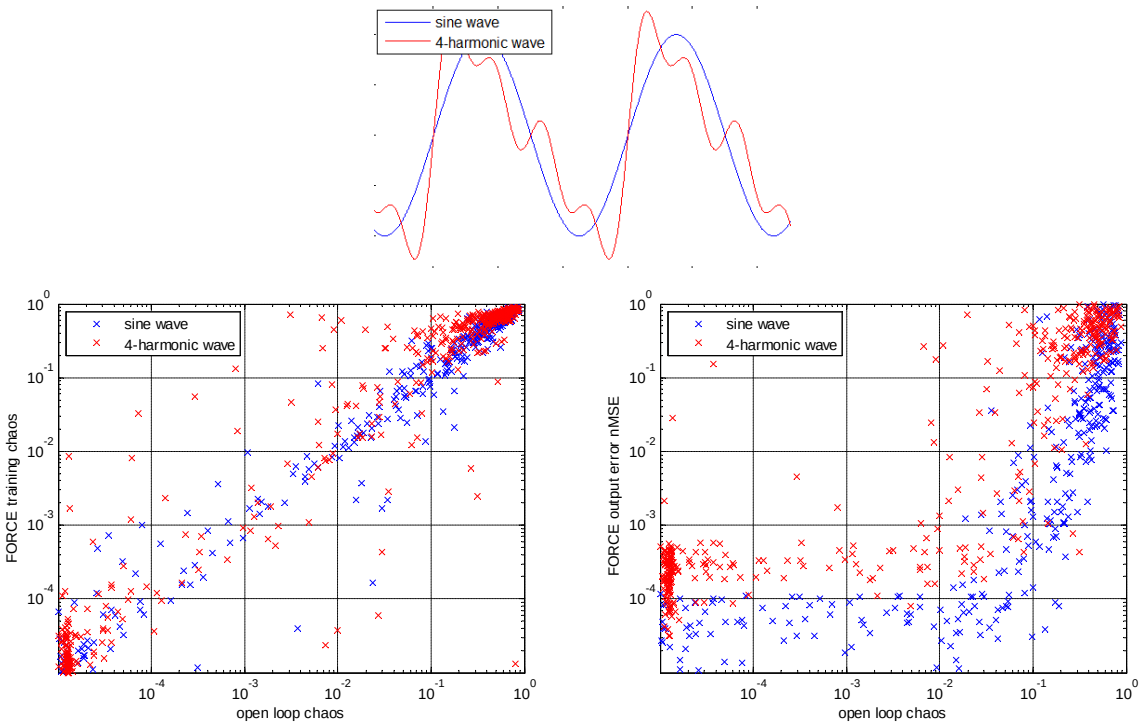

Figure 2: Chaotic (non-periodic) fraction of energy in the open loop system (main text eq. 2) plotted vs. chaotic energy in the trained system (left) and versus normalized mean squared error (right).

$$< \left(c_{nm}\sigma_m y_1 + \sqrt{1-c_{nm}^2}\sigma_m y_2\right), (1+i\omega)\left(\alpha_n\sigma_n y_1 + \alpha_{mn}\sigma_m y_2 + X^1_{n\perp}\right) >=$$
$$=< W\bar{r}_m, WR'_n X_n(\omega) >= g^2 < \bar{r}_m^T R'_n X_n(\omega) >= g^2 G_{nm}(\omega)$$

The resulting self consistency equation for α_{mn} is:

$$g^2 G_{nm}(\omega) = (1+i\omega)\left(c_{nm}\sigma_n\sigma_m\alpha_n + \sqrt{1-c_{nm}^2}\sigma_m^2\alpha_{nm}\right) = (1+i\omega)^{-1}\gamma_0 + \gamma_1\alpha_n + \gamma_2\alpha_{nm} \tag{2}$$

with coefficients $\gamma_{0,1,2}$ defined as

$$\gamma_{0,1,2} \triangleq g^2 \int \mathscr{D}w \int Dy_1 Dy_2 \phi\left(wA_m + \sigma_m\left(c_{nm}y_1 + \sqrt{1-c_{nm}^2}y_2\right)\right)\phi'\left(wA_n + \sigma_n y_1\right)\xi_{0,1,2}$$

with $\xi_0 = w$, $\xi_1 = \sigma_n y_1$ and $\xi_2 = \sigma_m y_2$. And general term $G_{nm}(\omega)$ follows:

$$G_{nm}(\omega) = \frac{K_{nm}(i\omega - z_{nm})}{(i\omega - p_{nn})(i\omega - p_{nm})} \tag{3}$$

with poles at

$$p_{nn} = \beta_1 - 1$$

and

$$p_{nm} = \gamma_2(1-c_{nm}^2)^{-\frac{1}{2}}\sigma_m^{-1} - 1,$$

a zero at

$$z_{nm} = p_{nn} - \beta_0\gamma_0^{-1}\left(\gamma_1 - \gamma_2 c_{nm}(1-c_{nm}^2)^{-\frac{1}{2}}\sigma_n\sigma_m^{-1}\right)$$

and a gain coefficient

$$K_{nm} = \gamma_0$$

While the pole and zero values are not fully interpretable, some insights are available, revealing typical frequency range for non-trivial phenomena. First one can show that:

$$p_{nm} = \gamma_2(1 - c_{nm}^2)^{-\frac{1}{2}}\sigma_m^{-1} - 1 = g^2 < r'_m r'_n > -1$$

Remarkably, this expression remains always negative for a network compliant with the fading memory property which implies $1 > max(\rho_m^2, \rho_n^2) \geq g^2 < r'_m r'_n >$. Calculation is done by integrating γ_2 by parts with respect to y_2 leading to:

$$\begin{aligned}
\gamma_2 = g^2 \int \mathscr{D}w \int Dy_1 Dy_2 \phi\left(wA_m + \sigma_m\left(c_{nm}y_1 + \sqrt{1-c_{nm}^2}y_2\right)\right)\phi'\left(wA_n + \sigma_n y_1\right)\sigma_m y_2 = \\
(1-c_{nm}^2)^{\frac{1}{2}} g^2 \sigma_m \int \mathscr{D}w \int Dy_1 Dy_2 \phi'\left(wA_m + \sigma_m\left(c_{nm}y_1 + \sqrt{1-c_{nm}^2}y_2\right)\right)\phi'\left(wA_n + \sigma_n y_1\right) + \\
- g^2\sigma_m \int \mathscr{D}w \int Dy_1\, \phi\left(wA_m + \sigma_m\left(c_{nm}y_1 + \sqrt{1-c_{nm}^2}y_2\right)\right)\phi'\left(wA_n + \sigma_n y_1\right)\exp(y_2^2/2)\Big|_{y_2=-\infty}^{y_2=\infty} = \\
= (1-c_{nm}^2)^{\frac{1}{2}}\sigma_m g^2 < r'_m r'_n >
\end{aligned}$$

Similarly, β_1 can be expressed as follows:.

$$\begin{aligned}
\beta_1 = g^2\sigma^{-2} \int Dy \mathscr{D}w \phi(wA + \sigma y)\phi'(wA + \sigma y)\sigma y \\
= g^2\sigma^{-2}(\sigma^2 \int Dy \mathscr{D}w \phi'(wA + \sigma y)\phi'(wA + \sigma y) + \\
\sigma^2 \int Dy \mathscr{D}w \phi(wA + \sigma y)\phi''(wA + \sigma y)) - \sigma \mathscr{D}w \phi(wA + \sigma y)\phi'(wA + \sigma y)\exp(y^2/2)\big|_{y=-\infty}^{y=\infty})
\end{aligned}$$

and hence for sub-exponential activation functions ϕ we have:

$$\beta_1 = \rho^2 + g^2 \int Dy \mathscr{D}w \phi(wA + \sigma y)\phi''(wA + \sigma y)$$

For sigmoids centered at origin, the correction term $\eta = g^2 \int Dy \mathscr{D}w \phi(wA + \sigma y)\phi''(wA + \sigma y)$ is always negative. We thus have:

$$\beta_1 = \rho^2 + \eta < \rho^2$$

For g of order of *one* and $\phi(x) = \tanh(x)$, η remains of the order of *one* as well. For $g \to +\infty$ sigmoidal ϕ implies $\sigma \propto g$ and here further work is needed to elucidate the balance between growing g and vanishing importance of the second derivative $\phi''(wA + \sigma y)$ w.r.t the Gaussian measure.

For the zero z_{nm}, we are not aware of a meaningful simplification for a general case. However, in a limit of close training targets $A_n - A_m \to 0$, we have $p_{mn} = \rho^2 - 1$, where ρ is the spectral radius of an open loop system (main text eq. 2). This term, potentially diverging at the edge of the chaos [9, 10] , is approximately canceled at the above limit by the zero z_{mn} (exact cancellation occurs for $A_m \equiv A_n$ accounting for the particularly simple solution for the single fixed point case (main text eq. 13)). For training targets that are *not* infinitesimally close, no cancellation occurs, and a complex frequency characteristic of both individual terms G_{mn} and of their weighted sum (main text eq. 17) can emerge.

In conclusion, we may expect frequencies where resonance occurs to be of the order magnitude of the poles and zeros elaborated above.

---

[1] H. Jaeger, Bonn, Germany: German National Research Center for Information Technology GMD Technical Report **148**, 34 (2001).

[2] D. Sussillo and L. F. Abbott, Neuron **63**(4), 544 (2009).
[3] K. J. Aström and R. M. Murray, *Feedback systems: an introduction for scientists and engineers* (Princeton university press, 2010), chap. 6.
[4] G. Hennequin, T. P. Vogels, and W. Gerstner, Neuron **82**(6), 1394 (2014).
[5] I. B. Yildiz, H. Jaeger, and S. J. Kiebel, Neural networks **35**, 1 (2012).
[6] G. Manjunath and H. Jaeger, Neural computation **25**(3), 671 (2013).
[7] K. Rajan, L. F. Abbott, and H. Sompolinsky, Phys. Rev. E **82**, 011903 (Jul 2010), http://link.aps.org/doi/10.1103/PhysRevE.82.011903.
[8] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Phys. Rev. Lett. **61**, 259 (Jul 1988), http://link.aps.org/doi/10.1103/PhysRevLett.61.259.
[9] M. Massar and S. Massar, Phys. Rev. E **87**, 042809 (Apr 2013), http://link.aps.org/doi/10.1103/PhysRevE.87.042809.
[10] Y. Ahmadian, F. Fumarola, and K. D. Miller, Phys. Rev. E **91**, 012820 (Jan 2015), http://link.aps.org/doi/10.1103/PhysRevE.91.012820.