Transient chaotic SNR amplification

Strongly chaotic non-linear networks strongly separate inputs, but are believed to be useless for classification tasks because also irrelevant (noise) differences within any class are exacerbated, leading to bad generalization. We show this is actually not the case during the initial time period following input presentation: During this time, the representation is dominated by expansion, but not by mixing, and larger differences (between classes) expand faster than smaller differences (within classes). Therefore, the representation is disentangled by the dynamics, and when classifying the network state by linear readouts, the signal-to-noise ratio (SNR) actually increases, before it eventually deteriorates when mixing begins to dominate. We show that this is a general effect in high-dimensional non-linear chaotic systems, and demonstrate it in spiking, continuous rate, and LSTM networks. The transient SNR amplification is always fast (within 50 ms) for spiking networks, while its timescale in continuous valued networks depends on the distance to the edge of chaos. Moreover, this fast, noise-resilient transient disentanglement of representations is in line with empirical evidence: the olfactory bulb, for example, rapidly enhances the separability of sensory representations in a single recurrent layer, being the initial processing stage of a relatively flat hierarchy.

Additional detail

We exemplify and analyze the effect of transient chaotic SNR amplification in the following setting: Consider a classification task consisting of P classes of L-dimensional vectors $x_0^p \in [-1,1]^L$. Each class p is a Gaussian distribution with a random mean vector given by $\bar{x}_0^p$ and standard deviation σ controlling the variability within the class. If $P \gg 2L$, the linear separability of the classes is very low, relating to Cover's theorem [1]. We employ a reservoir computing paradigm to solve the task: A random recurrent network of $N \gg L$ neurons is presented at t_0 with an L - dimensional pattern drawn from the task distribution, and for any fixed time t, linear readouts are trained encode by a one-hot vector the class that the pattern was drawn from. As the first network model, we consider classical continuous-valued units

$$\tau \partial_t \boldsymbol{h} = -\boldsymbol{h} + \boldsymbol{J}\,\mathrm{T}(\boldsymbol{h})\,, \tag{1}$$

where T is a sigmoid non-linearity, the entries of $\boldsymbol{J} \in \mathbb{R}^{N\times N}$ are drawn from $\mathcal{N}(0, g^2/N)$, and g controls the network dynamics to be in the regular ($g<1$) or chaotic ($g>1$) regime [2]. It has been extensively argued that the parameter regime close to the edge of chaos, that is $g \approx 1$, optimizes the computational power of such networks when used in reservoir computing [e.g., 3, 4]. Here we instead investigate the strongly chaotic regime, $g \gg 1$. What is the characteristic feature of such networks? Any two close points in state space are drawn apart by the dynamics. While this could benefit a separation of entangled representations, it has been argued that also any noise is strongly amplified, and therefore computational utility should be poor in the strongly chaotic regime [3]. However, if the task does not require long memory, the network can perform powerful transient computation. This is related to how quickly points in state space diverge depending on their distance, which can be calculated analytically by a replica calculation using dynamical mean-field theory, leading to [5, 6]

$$(\partial_t + 1)(\partial_s + 1) Q^{(12)}(t,s) = g^2\, f_{\mathrm{T}}(Q_0, Q^{(12)}), \tag{2}$$

where $Q^{(12)}(t,s)$ is the average correlation between two trajectories at times t, s and $f_{\mathrm{T}}(Q_0, Q^{(12)}) = \langle \mathrm{T}(h_1)\mathrm{T}(h_2)\rangle$ with the average taken with respect to $(h_1, h_2) \sim \mathcal{N}\Big(0, \begin{pmatrix} Q_0 & Q^{(12)} \\ Q^{(12)} & Q_0 \end{pmatrix}\Big)$. Integrating this equation, we find that short after stimulus presentation, two points that are close diverge slower than two points that are further apart (Fig 1**a**). Ultimately, distances saturate at the maximal distance determined by the bounded state space volume. What does this mean for the classification task? Initially after presentation, the representation has dimensionality $L \ll N$ and is not linearly separable by the readouts because also $2L \ll P$. But when the chaotic dynamics begins to expand the representation, the larger differences between classes expand faster than the smaller within class variability. Crucially, the larger differences are also more strongly affected by the non-linearity, which embeds them into the higher dimensional state space and improves the linear separability of the classes. This trend finally reverses when the expansion saturates and mixing begins to dominate. Therefore the classification accuracy transiently peaks, Fig 1**a** (inset). We argue that this is a general effect in high-dimensional, non-linear chaotic systems. Here we demonstrate it also for a recurrent LSTM network [7], Fig 1**b**, and an inhibitory LIF (leaky-integrate-and-fire) network, Fig 1**c**.

While for the continuous network (1) the link between distances and dimensionality was explained by a qualitative argument, this link can be made explicitly for the spiking LIF network: Binning the spike trains by a moving window

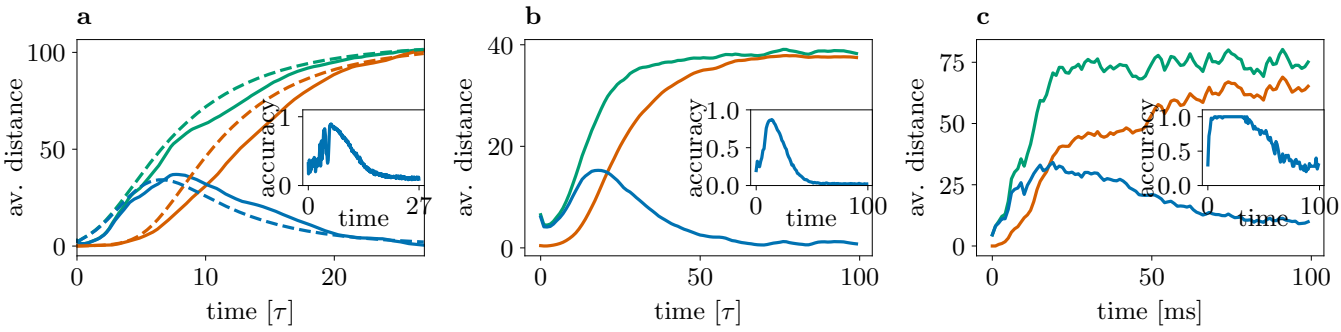

Figure 1. **Transient chaotic SNR amplification in continuous valued (a), LSTM (b) and LIF (c) networks. a** Average inter-class distances (green), intra-class distances (orange) and difference between the two (blue). Numerical solutions of equation (2) (dashed). Inset shows classification accuracy on unseen examples. Network parameters: $N = 250$, $g = 5.8$. Task parameters: $P = 50$, $L = 8$ and $\sigma = 0.3$, readouts trained on 100 examples per pattern. **b** Same for LSTM network with $N = 200$. **c** Same for inhibitory LIF network in the asychronous-irregular regime with fixed in-degree $K = 125$, average firing rate $\nu = 19\,\mathrm{Hz}$, and $N = 500$.

whose width represents the membrane time constant of the readout neurons, we obtain at each time a binary vector $\in \{0,1\}^N$ if the activity is sparse enough to make the occurrence of two spikes in one bin unlikely. Due to the geometry of the $\{0,1\}^N$ hyper-cube, a set of points with average mutual distances $d \ll N$ must also span d dimensions. Thus the average distance between trajectories on the hypercube can be interpreted directly as the dimensionality of the representation. In Fig 1**c** we again plot the average distance d_s between two patterns of different classes, the average distance d_n between two patterns of the same class, and their difference $\Delta d = d_s - d_n$, which can thus be interpreted as the effective dimensionality of the representation that is not corrupted by noise, and is therefore predictive of the classification accuracy, Fig 1**c** (inset). The divergence speed of trajectories in the LIF network can be related to the divergence rate between flux tubes, which are locally stable environments of trajectories coexisting with global instability [8]. Also, their existence introduces an additional computational effect: Because any variability within a flux tube is quenched by the network, a fraction of the noise is 'swallowed' and not amplified. This causes a slowed down rise of the noise dimensionality, and a non-zero plateau of the classification accuracy at late times (Fig 1**c**).

Finally, the olfactory system is a good candidate to rely on the transient computational mechanism we have described, because it is specialized on pattern classification. Indeed, recordings in the antennal lobe of locusts [9], and the olfactory bulb of zebrafish [10] and rats [11] show responses consistent with the here found mechanism, in that the spatio-temporal activity patterns following odor presentation quickly decorrelate even for similar odors, and the linear decoding accuracy of odor identity peaks during the initial transient, not after a stable state has been reached. We formulate the following testable predictions for neural systems that implement classification by transient chaotic SNR amplification:

1. Variability is small or quenched at stimulus onset, then transiently increases and reaches a stable value.
2. Not only the inter-class distances, but also the intra-class (noise) distances increase, although initially slower.
3. Decoding accuracy based on linear readouts trained at each time point shows a peak, and this peak occurs before the distances saturate.

[1] T. M. Cover, IEEE Transactions on Electronic Computers **EC-14**, 326 (1965).
[2] H. Sompolinsky, A. Crisanti, and H. J. Sommers, Phys. Rev. Lett. **61**, 259 (1988).
[3] R. Legenstein and W. Maass, Neural Networks **20**, 323 (2007).
[4] T. Toyoizumi and L. F. Abbott, Phys. Rev. E **84**, 051908 (2011).
[5] J. Schuecker, S. Goedeke, and M. Helias, Phys Rev X **8**, 041029 (2018).
[6] J. Kadmon and H. Sompolinsky, Phys. Rev. X **5**, 041030 (2015).
[7] S. Hochreiter and J. Schmidhuber, Neural computation **9**, 1735 (1997).
[8] M. Monteforte and F. Wolf, Physical Review X **2**, 041007 (2012).
[9] O. Mazor and G. Laurent, Neuron **48**, 661 (2005).
[10] R. Friedrich and G. Laurent, Science **291**, 889 (2001).
[11] K. M. Cury and N. Uchida, Neuron **68**, 570 (2010).