Supplementary material: Information processing capacity of dynamical systems

Joni Dambre, David Verstraeten, Benjamin Schrauwen and Serge Massar

June 28, 2012

1 Proofs

1.1 Idea of the proofs and geometrical intuition

To derive the theorems stated in the main text we use two key mathematical concepts. The first is the ergodic theorem which states that for i.i.d. input signals $u(t)$ with distribution $p(u)$, and under appropriate technical conditions, one can replace time averages by ensemble averages, i.e., that $\langle\cdot\rangle_T = \langle\cdot\rangle_{U^\infty}$ [1, 2, 3]. The second is that by introducing the Hilbert space of fading memory functions one can reason in a space that has much more structure than the set of time series, and thereby get stronger results and better intuition. We first sketch how the theorems should follow from these two concepts. The detailed proofs are given in full below.

We place ourselves in the conditions of Theorem 7, namely we have N linearly independent fading memory dynamical variables $x_i(u^{-\infty}) \in \mathcal{H}_{U^\infty}$. We denote by $\mathcal{H}_X \subset \mathcal{H}_{U^\infty}$ the N dimensional subspace consisting of the span of memory functions $x_i(u^{-\infty})$. An orthonormal basis for $\mathcal{H}_X$ can be constructed by taking suitable linear combinations of the original output functions: $\bar{x}_j = \sum_i \Lambda_{ji} x_i$, $\langle \bar{x}_i, \bar{x}_j\rangle_{U^\infty} = \delta_{ij}$, $i,j = 1 \dots N$. (Note that the capacity $C_T[X,z]$ does not change under an invertible linear transformation carried out on the x_i). Let $\Pi_X(\cdot)$ be the projector onto $\mathcal{H}_X$, which can be written as $\Pi_X(\cdot) = \sum_i \langle \cdot, \bar{x}_i\rangle_{U^\infty} \bar{x}_i$. Using the definition eq. (1.6) and ergodicity, the capacity for reconstructing a function z using a linear combination of the observation functions can be written as:

$$\lim_{T\to\infty} C_T[X,z] = \frac{1}{\|z\|^2_{U^\infty}} \sum_{i=1}^{N} \langle z, \bar{x}_i\rangle_{U^\infty} \langle \bar{x}_i, z\rangle_{U^\infty} = \frac{\|\Pi(z)\|^2_{U^\infty}}{\|z\|^2_{U^\infty}}, \tag{1.1}$$

which is ratio of the the squared norm of the projection of z onto $\mathcal{H}_X$ to the squared norm of z. This implies the normalization condition $0 \le \lim_{T\to\infty} C_T[X,z] \le 1$. [1]

Let us now consider a finite orthonormal set of functions in $\mathcal{H}_{U^\infty}$: $Y_L = \{y_1, \dots, y_L\}$. Denote by Π_{Y_L} the projector onto the space spanned by Y_L. Then (once again using ergodicity) the sum of the capacities for the y_l

$$\sum_{l=1}^{L} \lim_{T\to\infty} C_T[X,y_l] = \sum_{l=1}^{L}\sum_{i=1}^{N} \langle y_l, \bar{x}_i\rangle_{U^\infty} \langle \bar{x}_i, y_l\rangle_{U^\infty} = \sum_{i=1}^{N} \|\Pi_{Y_L}(\bar{x}_i)\|^2_{U^\infty} \tag{1.2}$$

$$\le \sum_{i=1}^{N} \|\bar{x}_i\|^2_{U^\infty} = N \tag{1.3}$$

is the sum of the norm square of the projection of the $\bar{x}_i$ onto the space spanned by the y_l. Each of these terms is bounded by the norm square of the $\bar{x}_i$ (this is known as Bessel's identity) which is 1, giving an overall bound of N. In the limit where Y_L constitutes an orthonormal basis for $\mathcal{H}_{U^\infty}$ the projection Π_{Y_L} is the identity operator, $\|\Pi_{Y_L}(\bar{x}_i)\|^2_{U^\infty} = \|\bar{x}_i\|^2_{U^\infty}$, and one has equality in eq. (1.2).

1.2 Normalization of Capacity

Here we prove Proposition 3.

[1] Note that Prop. 3 is stronger since it states that the normalization holds for all finite times T, and also for dynamical systems that are not fading memory.

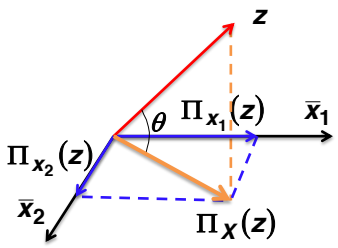

Figure 1.1: Geometrical picture of dynamical system and capacities. The fading memory dynamical system has two variables, x_1 and x_2 (not pictured). The vectors $\bar{x}_1$ and $\bar{x}_2$ (pictured) are an orthonormal basis of the space spanned by x_1 and x_2. For the sake of the representation, we consider a three dimensional subspace of the Hilbert space containing the two vectors x_1 and x_2. We consider a function z. Its projection on the space spanned by $\bar{x}_1$ and $\bar{x}_2$ is denoted $\Pi_X(z)$, and the corresponding components along $\bar{x}_1$ and $\bar{x}_2$ by $\Pi_{x_1}(z)$ and $\Pi_{x_2}(z)$. The capacity to reconstruct z is given by $\cos^2(\theta)$, where the θ is the angle between z and $\Pi_X(z)$. Note that $\Pi_X(z) = \|z\| \cos(\theta)$.

Proof. Any time series $f(t)$, $t = 1, ..., T$ can be viewed as a vector in $\mathbb{R}^T$. Hence the time average of a product

$$\frac{1}{T}\sum_{t=1}^{T} f(t)g(t) = \langle f, g\rangle_T \tag{1.4}$$

can be viewed as the standard scalar product on $\mathbb{R}^T$.

Consider the output time series $x_i(t)$, $t = 1, ..., T$. We denote the correlation matrix by

$$R_{ii'} = \langle x_i, x_{i'}\rangle_T \tag{1.5}$$

and by $R^{-1}_{ii'}$ the inverse of the matrix $R_{ii'}$. (For simplicity we suppose that R has full rank, otherwise we restrict the analysis to the subspace on which R is nonzero). Since $R_{ii'}$ is a symmetric matrix it can be diagonalized, and in particular one can define a matrix $\Lambda_{ii'}$ such that $\Lambda^T R \Lambda = I$ (i.e., $\sum_{kl} \Lambda_{ki} R_{kl} \Lambda_{lj} = \delta_{ij}$). This allows us to define new internal variables through $\bar{x}_i = \sum_j \Lambda_{ki} x_j$. In terms of these new variables we have

$$\bar{R}_{ii'} = \langle \bar{x}_i, \bar{x}_{i'}\rangle_T = \delta_{ii'} \ . \tag{1.6}$$

That is the time series $\bar{x}_i(t)$, $t = 1, ..., T$ are orthogonal normalized vectors of $\mathbb{R}^T$ for the scalar product eq. (1.4).

Consider a target function $z(t)$ and an estimator $\hat{z}(t) = \sum_i W_i x_i(t)$. When varying the MSE with respect to W_i, one finds that the optimal linear estimator is given by $W_i = \sum_{i'} R^{-1}_{ii'} P_{i'}$ with $P_i = \langle x_i, z\rangle_T$. This implies the following identities for the optimal linear estimator:

$$\hat{z}(t) = \sum_i \bar{x}_i(t)\langle \bar{x}_i, z\rangle_T \tag{1.7}$$

$$\langle \hat{z}^2\rangle_T = \langle \hat{z}\hat{z}\rangle_T = \sum_{i,j=1}^{N} P_i R^{-1}_{ij} P_j = \sum_{i=1}^{N} \langle \bar{x}_i, z\rangle_T^2 \tag{1.8}$$

$$MMSE_T = \langle z^2\rangle_T - \langle \hat{z}^2\rangle_T \tag{1.9}$$

$$C_T[X, z] = \frac{\langle \hat{z}^2\rangle_T}{\langle z^2\rangle_T} = \frac{\sum_{i,j=1}^{N} P_i R^{-1}_{ij} P_j}{\langle z^2\rangle_T} = \frac{\sum_{i=1}^{N} \langle \bar{x}_i, z\rangle_T^2}{\langle z^2\rangle_T} \tag{1.10}$$

Equation (1.7) shows that the best linear estimator $\hat{z}(t) = \sum_i W_i x_i(t)$ of a time series $z(t)$ is the orthogonal projection of $z(t)$ onto the subspace of $\mathbb{R}^T$spanned by the $x_i(t)$. Equation (1.10) then implies that the capacity for reconstructing z is the ratio of the norm squares of the projection $\hat{z}(t)$ and of the original time series $z(t)$: $C_T[X, z] = \langle \hat{z}^2\rangle_T / \langle z^2\rangle_T$. Hence the capacity to reconstruct z is normalized according to $0 \leq C_T[X, z] \leq 1$ for any time series $z(t)$ and any output time series $x_i(t)$. □

1.3 Ergodicity

Central to the proofs below is the theory of stochastic processes. For an introduction see, e.g., [1, 2, 3]. The main tool we will use is the ergodic theorem, that is the possibility, for most functions $y : U^\infty \to \mathbb{R}$, to replace the time

average by the ensemble average:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}y(t)=E_{U^\infty}[y]=\langle y\rangle_{U^\infty}. \tag{1.11}$$

To formalize this notion it is useful to introduce the time translation operator: $\mathcal{U}_\tau$ which acts as follows: $\mathcal{U}_\tau y[...,u_{-2},u_{-1,}u_0]=y[...,u_{-2-\tau},u_{-1-\tau},u_{-\tau}]$. We also introduce the covariance $Cov_y(\tau)=E_{U^\infty}\left[(y-\langle y\rangle_{U^\infty})\left(U_\tau y-\langle y\rangle_{U^\infty}\right)\right]$. We will use the following result which is sufficiently simple that for completeness we give the proof:

Proposition 8. (Ergodic Theorem). Consider a function $y:U^\infty\to\mathbb{R}$ with finite variance $var[y]=E_{U^\infty}\left[(y-\langle y\rangle_{U^\infty})^2\right]<\infty$ and such that its covariance tends to zero for long times, $\lim_{\tau\to\infty}Cov_y(\tau)=0$. Then the time average of y converges to the ensemble average in mean square:

$$\lim_{T\to\infty}E_{U^\infty}\left[\left(\frac{1}{T}\sum_{t=1}^{T}y(t)-\langle y\rangle_{U^\infty}\right)^2\right]=0. \tag{1.12}$$

Note that convergence in mean square implies convergence in probability: for all $\epsilon,\delta>0$, there exists $T>0$ such that

$$Prob\left[\frac{1}{T}\sum_{t=1}^{T}y(t)-\langle y\rangle_{U^\infty}>\epsilon\right]<\delta. \tag{1.13}$$

Proof. First note that the Cauchy-Shwarz inequality implies $Cov_y(\tau)<Var[y]$ for all τ.

Second, note that $\lim_{\tau\to\infty}Cov_y(\tau)=0$ is equivalent to: for all $\epsilon>0$, there exists $\tau>0$ such that for all $t>\tau$, $Cov_y(\tau)<\epsilon$.

Third, fix $\epsilon>0$ and $\tau>0$ as above, and rewrite

$$E_{U^\infty}\left[\left(\frac{1}{T}\sum_{t=1}^{T}y(t)-\langle y\rangle_{U^\infty}\right)^2\right] = \frac{1}{T}Var[y]+\frac{2}{T^2}\sum_{t=1}^{\tau}(T-t)Cov_y(t) \tag{1.14}$$

$$+\frac{2}{T^2}\sum_{t=\tau+1}^{T}(T-t)Cov_y(t) \tag{1.15}$$

from which we obtain $E_{U^\infty}\left[\left(\frac{1}{T}\sum_{t=1}^{T}y(t)-\langle y\rangle_{U^\infty}\right)^2\right]\le\frac{(1+2\tau)}{T}Var[y]+2\epsilon$. Hence for sufficiently large T, we have $E_{U^\infty}\left[\left(\frac{1}{T}\sum_{t=1}^{T}y(t)-\langle y\rangle_{U^\infty}\right)^2\right]<3\epsilon$ which proves the result. □

1.4 Proof of bound on capacity

Here we prove Theorem 4.

Proof. We need to show that

$$\sum_{l=1}^{L}C_T[X,y_l]=\sum_{i=1}^{N}\sum_{l=1}^{L}\frac{\langle\bar{x}_i,y_l\rangle_T^2}{\langle y_l^2\rangle_T} \tag{1.16}$$

is bounded by N in the limit $T\to\infty$ if $\langle y_l y_{l'}\rangle_{U^h}=\delta_{ll'}$ and $\langle y_l^4\rangle_{U^h}<\infty$, where we have used the variables $\bar{x}_i(t)$ defined such that their correlation matrix is the identity, see eq. (1.6).

The idea of the proof is that, since $\langle y_l y_{l'}\rangle_{U^h}=\delta_{ll'}$, the ergodic theorem implies that $\sum_l y_l(t)y_l(t')/\langle y_l^2\rangle_T$, viewed as $T\times T$ matrix, is "almost" an orthogonal projector onto a L dimensional subspace of $\mathbb{R}^T$. Then one should have $\sum_l\frac{\langle\bar{x}_i,y_l\rangle_T^2}{\langle y_l^2\rangle_T}\lesssim\langle\bar{x}_i^2\rangle_T$ from which follows the result. The technicalities consist of taking care of the "almost".

Ergodicity (which is where the condition $\langle y_l^4\rangle_{U^h}<\infty$ comes in) implies that, with probability at least $1-\delta$, for sufficiently large T, $\langle y_l y_{l'}\rangle_T=\delta_{ll'}+\epsilon_{ll'}$ with $|\epsilon_{ll'}|<\epsilon$ (we use the fact that $L<\infty$ is finite to ensure that δ,ϵ are independent of l,l').

Let us define $z_k(t)=\sum_{k'}O_{kk'}y_{k'}(t)$ with $O_{kk'}$ an orthogonal matrix, such that $\langle z_k z_{k'}\rangle_T=\lambda_k\delta_{kk'}$ is diagonal. We have

$$\lambda_k=\langle z_k^2\rangle_T=1+\sum_{k'k''}O_{kk'}O_{kk''}\epsilon_{k'k''}\ .$$

Hence with probability at least $1-\delta$, $|\lambda_k - 1| < L^2\epsilon$ where we use that $|O_{kk'}| < 1$.

We can express y_k in terms of z_k as $y_k(t) = \sum_{k'} O_{k'k} z_{k'}(t)$. Using this relation and the fact that $O_{kk'}$ is an orthogonal matrix we obtain that with probability at least $1-\delta$:

$$\begin{aligned}
\sum_{k=1}^{L} C_T[X, y_k] &= \sum_{i=1}^{N}\sum_{k=1}^{L} \frac{\sum_{k'k''\in K}\langle \bar{x}_i, z_{k'}\rangle_T O_{k'k} O_{k''k}\langle z_{k''}, \bar{x}_i\rangle_T}{(1+\epsilon_{kk})} \ . \\
&\leq \frac{1}{1-\epsilon}\sum_{i=1}^{N}\sum_{k'k''=1}^{L}\langle \bar{x}_i, z_{k'}\rangle_T \left(\sum_{k=1}^{L} O_{k'k}O_{k''k}\right)\langle z_{k''}, \bar{x}_i\rangle_T \\
&\leq \frac{1}{1-\epsilon}\sum_{i=1}^{N}\sum_{k=1}^{L}\langle \bar{x}_i, z_k\rangle_T\langle z_k, \bar{x}_i\rangle_T \\
&\leq \frac{1+L^2\epsilon}{1-\epsilon}\sum_{i=1}^{N}\sum_{k=1}^{L}\frac{1}{\lambda_k}\langle \bar{x}_i, z_k\rangle_T\langle z_k, \bar{x}_i\rangle_T
\end{aligned}$$

We define the functions $\bar{x}_k(t) = z_k(t)/\sqrt{\lambda_k}$ which form an orthonormal basis of $\mathbb{R}^T$. Therefore, with probability at least $1-\delta$, we have

$$\sum_{k=1}^{L} C_T[X, y_k] \leq \frac{1+L^2\epsilon}{1-\epsilon}\sum_{i=1}^{N}\sum_{k=1}^{L}\langle \bar{x}_i, \bar{x}_k\rangle_T^2.$$

Bessel's identity implies that $\sum_{k=1}^{L}\langle \bar{x}_i, \bar{x}_k\rangle_T^2 \leq \langle \bar{x}_i^2\rangle_T = 1$. Hence with probability at least $1-\delta$,

$$\sum_{k=1}^{L} C_T[X, P_k] \leq \frac{1+L^2\epsilon}{1-\epsilon} N \ .$$

Taking $\epsilon \to 0$ (and hence $T \to \infty$) concludes the proof. □

1.5 Hilbert space of fading memory functions

The Hilbert space of fading memory functions plays a central role in our analysis. We introduce here the basic notions we need. For introductions to the theory of Hilbert spaces, see, e.g., [4].

We recall that $\mathcal{H}_U$ is the Hilbert space of functions $f : U \to \mathbb{R}$ that depend on a single input $u \in U$, with scalar product given by the probability measure on U: $\langle x, y\rangle_U = \mathbb{E}_U[xy]$. It is well known that $\mathcal{H}_U$ constitutes a Hilbert space, known as the weighted L^2 space. We make the hypothesis that the Hilbert space $\mathcal{H}_U$ is separable, i.e., that any basis of $\mathcal{H}_U$ is either finite number or denumerable. We choose a basis of $\mathcal{H}_U$ composed of the function $y_j(u)$, $j \in \mathbb{N}$, with the first basis vector equal to the constant function $y_0(u) = 1$.

The Hilbert space $\mathcal{H}_{U^h}$ of functions that depend on a finite number h of inputs is the tensor product of h copies of $\mathcal{H}_U$: $H_{U^h} = \otimes_{j=1}^{h}\mathcal{H}_{U_j}$. The scalar product on $\mathcal{H}_{U^h}$ is given by the probability measure on U^h: $\langle x, y\rangle_{U^h} = \mathbb{E}_{U^h}[xy]$. The set of all products $y_{j_1}y_{j_2}...y_{j_h} : U^h \to \mathbb{R}$ constitutes a basis of $\mathcal{H}_{U^h}$.

The Hilbert space $\mathcal{H}_{U^\infty}$ of fading memory functions is a subspace of the space of functions from $U^\infty \to \mathbb{R}$. We give two definitions of $\mathcal{H}_{U^\infty}$ and then show that they are equivalent:

Definition 9. *Hilbert space of fading memory functions: Cauchy sequences.* The Hilbert space $\mathcal{H}_{U^\infty}$ of fading memory functions is defined as the limit of functions x_h on H_{U^h}, as h increases as follows:

1. If $x \in \mathcal{H}_{U^h}$, then $x \in \mathcal{H}_{U^\infty}$, and $\langle x, x\rangle_{U^\infty} = \langle x, x\rangle_{U^h}$.
2. Consider a sequence $x_h \in \mathcal{H}_{U^h}$, $h \in \mathbb{N}$ belonging to larger and larger tensor products $\mathcal{H}_{U^h}$. Then the limit $\lim_{h\to\infty} x_h$ exists and is in $\mathcal{H}_{U^\infty}$ if for all $\epsilon > 0$, there exits $h_0 \in \mathbb{N}$, such that for all $h, h' > h_0$, $\| x_h - x_{h'} \|^2_{U^{\max(h,h')}} < \epsilon$.
3. Conversely, all $x \in \mathcal{H}_{U^\infty}$ are the limit of a sequence of the type given in 2).
4. If $x, x' \in \mathcal{H}_{U^\infty}$, then their scalar product is defined as $\langle x', x\rangle_{U^\infty} = \lim_{h,h'\to\infty}\langle x'_{h'}, x_h\rangle$, where $x_h \to x$ and $x'_h \to x'$ are any two Cauchy sequences that converge to x and x' according to 2) above.

Definition 10. *Hilbert space of fading memory functions: basis construction.* A basis of the Hilbert space $\mathcal{H}_{U^\infty}$ of fading memory functions is given by all products $y_{\mathbf{j}} = y_{j_1}(u_1)y_{j_2}(u_2)...y_{j_h}(u_k)... : U^\infty \to \mathbb{R}$ where only a finite number of terms are different from the constant function $y_0 = 1$.

Proposition 11. *Definitions 9 and 10 are equivalent.*

Proof. First note that the basis vectors according to definition 10 are labeled by $j = j_0 j_1 j_2...$ where the $j_k \in \mathbb{N}$ ($k \in \mathbb{N}$) are positive integers, all but a finite number of which are different from zero. Hence the basis given in definition 10 is denumerable, hence the space $\mathcal{H}_{U^\infty}$ according to definition 10 is isomorphic to l^2.

For any basis function y_j with label $j = j_0 j_1 j_2...$, we denote by $h(j) \in \mathbb{N}$ the largest index k such that $j_k \neq 0$. Note that for any $h \in \mathbb{N}$, any j such that $h(j) \leq h$, the corresponding basis function y_j belongs to $\mathcal{H}_{U^h}$. Therefore according to both definitions 9 and 10 the space $\mathcal{H}_{U^\infty}$ is defined as the limit of functions in $\mathcal{H}_{U^h}$ for h increasing. It remains to show that there is a one to one correspondence between the elements of the spaces built according to definitions 9 and 10.

Consider any vector $x = \sum_{\mathbf{j}} c_{\mathbf{j}} y_{\mathbf{j}} \in \mathcal{H}_{U^\infty}$ according to definition 10. We can write $x = \sum_{\mathbf{j};\mathrm{m}(\mathbf{j})\leq \mathrm{h}} c_{\mathbf{j}} y_{\mathbf{j}} + \sum_{\mathbf{j};\mathrm{m}(\mathbf{j})>\mathrm{h}} c_{\mathbf{j}} y_{\mathbf{j}}$ and define the vectors $x_h = \sum_{\mathbf{j};\mathrm{m}(\mathbf{j})\leq \mathrm{h}} c_{\mathbf{j}} y_{\mathbf{j}} \in \mathcal{H}_{U^h}$. The set of vectors x_h are a Cauchy sequence according to both definitions 10 and 9. Thus any vector according to definition 10corresponds to a unique vector according to definition 9.

Consider any vector $x \in \mathcal{H}_{U^\infty}$ according to definition 9. If x_h is a Cauchy sequence converging to x according to 9, then we can write $x_h = \sum_{\mathbf{j};\mathrm{m}(\mathbf{j})\leq \mathrm{h}} c_{\mathbf{j}}^h y_{\mathbf{j}} \in \mathcal{H}_{U^h}$ and $(c_{\mathbf{j}}^h)$ is a Cauchy sequence in l^2. We can therefore identify the limit $\lim_{h\to\infty}(c_{\mathbf{j}}^h)$ with the corresponding limit vector in $\mathcal{H}_{U^\infty}$ according to definition 10: $\lim_{h\to\infty} x_h = \sum_{\mathbf{j}} \left(\lim_{h\to\infty} c_{\mathbf{j}}^h\right) y_{\mathbf{j}}$. Thus any vector according to definition 9 corresponds to a unique vector according to definition 10.

Third, it is obvious that the above one to one correspondence between the two spaces preserves the scalar product, hence it is an isomorphism. □

Proposition 12. If $x \in \mathcal{H}_{U^\infty}$, then the covariance of x tends to zero at large times: $\lim_{\tau\to\infty} Cov_x(\tau)$=0

Proof. Suppose without loss of generality that $\langle x\rangle_{U^\infty} = 0$. Since $x \in \mathcal{H}_{U^\infty}$, there exists a sequence $x_h \in \mathcal{H}_{U^h}$, such that for any $\epsilon > 0$, there exists h_0 such that $h \geq h_0$ implies $d(x - x_h) < \epsilon$. For any $\tau > h_0$, we have

$$\begin{aligned} Cov_x(\tau) &= d(x, \mathcal{U}_\tau x) \\ &\leq d(x, x_{h_0}) + d(x_{h_0}, \mathcal{U}_\tau x_{h_0}) + d(\mathcal{U}_\tau x_{h_0}, \mathcal{U}_\tau x) \\ &\leq 2\epsilon \end{aligned}$$

which proves the result. □

1.6 Saturation of total capacity for fading memory systems

Because the dynamical system has fading memory, its output functions $x_i(t)$ are uniquely determined by the previous inputs, and can be identified with fading memory functions $x_i(t) = x_i\left[u^{-\infty}(t)\right]$ with $x_i[u^{-\infty}]$ in $\mathcal{H}_{U^\infty}$. Indeed the convergence condition $E_{U^{t+T'}}\left(x_i(t) - x_i^h\left[u^{-h}(t)\right]\right)^2 < \epsilon$ for all $h > h_0$ in the definition of fading memory dynamical systems is precisely the condition that the functions x_i^h have a limit in $\mathcal{H}_{U^\infty}$. This limit will coincide with the output functions $x_i(t)$ in the limit where the initialization time T' tends to infinity.

We now prove Theorem 7.

Proof. The idea of the proof is to use ergodicity to rewrite the memory capacity as an ensemble average, and then to use Perceval's identity.

We define the correlation function $\tilde{R}_{ij} = \langle x_i x_j\rangle_{U^\infty}$. Because in the limit $T \to \infty$, the correlation matrix $R_{ii'} = \langle x_i, x_{i'}\rangle_T$ has rank N, it follows from the Ergodic theorem that $\tilde{R}$ also has rank N. The matrix $\tilde{R}$ can therefore be diagonalized. We introduce an invertible matrix Λ_{ij} and new function $\overline{x}_i\left[u^{-\infty}(n)\right] = \sum_j \Lambda_{ij} x_i\left[u^{-\infty}(n)\right]$ such that $\langle \overline{x}_i \overline{x}_j\rangle_{U^{-\infty}} = \delta_{ij}$.

For any finite time T, the memory capacity for target function $y_k[u^{-h}]$ can be re-expressed in terms of the $\overline{x}$ as:

$$C_T\left[X, y_k\right] = \sum_{i,j=1}^{N} \frac{\langle y_k, \overline{x}_i\rangle_T \langle \overline{x}_i \overline{x}_j\rangle_T^{-1} \langle \overline{x}_j, y_k\rangle_T}{\langle y_k^2\rangle_T}$$

Taking the limit $T \to \infty$, all quantities $\langle y_k, \overline{x}_i \rangle_T$, $\langle \overline{x}_i \overline{x}_j \rangle_T$, $\langle y_k^2 \rangle_T$ tend towards their ensemble average (because of the Ergodic Theorem, which can be applied in view of the conditions $\langle x_i^4 \rangle_{U^\infty} < \infty$, $\langle y_k^4 \rangle_{U^\infty} < \infty$). Using the same kind of reasoning as in the proof of Theorem 4, we can show that

$$\lim_{T\to\infty} C_T\left[X, y_k\right] = \sum_{i,j=1}^{N} \frac{\langle y_k, \overline{x}_i \rangle_{U^\infty} \langle \overline{x}_i \overline{x}_j \rangle_{U^\infty}^{-1} \langle \overline{x}_j, y_k \rangle_{U^\infty}}{\langle y_k^2 \rangle_{U^\infty}} = \sum_{i=1}^{N} \langle y_k, \overline{x}_i \rangle_{U^\infty}^2 \ .$$

In the limit $K \to K_{CS}$, where K_{CS} denotes a complete set of functions, the functions $y_k[u^{-\infty}]$ are an orthonormal basis of functions on $\mathcal{H}_{U^\infty}$. Hence any function $f \in \mathcal{H}_{U^\infty}$ can be expanded as $f = \sum_{k\in K_{CS}} c_k P_k$ with $c_k = \langle y_k f \rangle_{U^\infty}$. Perceval's identity states that $\sum_{k\in K_{CS}} |c_k|^2 = \langle f^2 \rangle_{U^\infty}$. In particular we can expand $\overline{x}_i = \sum_{k\in K_{CS}} \bar{x}_{ik} P_k$ and we have $\sum_{k\in K_{CS}} |\bar{x}_{ik}|^2 = 1$. Hence

$$\lim_{T\to\infty} \sum_{k\in K} C_T\left[X, y_k\right] = \sum_{i=1}^{N} \sum_{k\in K} |\bar{x}_{ik}|^2 \ .$$

Taking the limit $K \to K_{CS}$, we have

$$\lim_{K\to K_{CS}} \lim_{T\to\infty} \sum_{k\in K} C_T\left[X, y_k\right] = \sum_{i=1}^{N} \lim_{K\to K_{CS}} \sum_{k\in K} |x_{ik}|^2 = \sum_{i=1}^{N} = N \ .$$

which is the result we wanted to prove. □

2 Defining Capacities in terms of Correlation Coefficients

Rather than definition eq. (1.6), one can also define the capacities as

$$C'_T[X, z] = \frac{\sum_{ij} \mathrm{cov}(z x_i) \mathrm{cov}(x_i x_j)^{-1} \mathrm{cov}(x_j z)}{\mathrm{var}(z)} \tag{2.1}$$

where $cov(x, y) = \langle xy \rangle_T - \langle x \rangle_T \langle y \rangle_T$ and $var(x) = cov(x, x)$.

The two definitions are inequivalent but all the results for C_T also hold for C'_T. The original definition eq. (1.6) is more natural from a geometric and Hilbert space point of view, because definition eq. (2.1) implies that the constant function $y\left[u^{-h}\right] = 1$ is treated specially, while from the point of view of Hilbert space $\mathcal{H}_{U^h}$ it is just one function among many others. However, from the point of view of signal processing and statistical analysis, definition eq. (2.1) may seem more natural. It also makes contact with the work of [5] which used this definition.

For completeness we state the properties of C'_T. First we have the normalization:

$$0 \leq C'_T[X, z] \leq 1 \tag{2.2}$$

SecondTheorem 4 holds provided its statement is replaced by:

Corollary 13. *Consider a dynamical system as described above with N output functions $x_i(t)$ and choose a positive integer $h \in \mathbb{N}$. Consider any finite set $Y'_L = \{y'_1, ..., y'_L\}$ of size $|Y'_L| = L$ of functions $y_l \in \mathcal{H}_{U^h}$ obeying the orthogonality condition $\langle y'_l, y'_{l'} \rangle_{U^h} - \langle y'_l \rangle_{U^h} \langle y'_{l'} \rangle_{U^h} = {c_l}^2 \delta_{ll'}$, $l, l' = 1, ..., L$. We further require that the fourth moment of the y'_l is finite $\langle {y'_l}^4 \rangle_{U^h} < \infty$. Then, in the limit of an infinite data set $T \to \infty$, the sum of the capacities for these functions is bounded by the number N of output functions (independently of h, of the set Y'_L, or of its size $|Y'_L| = L$):*

$$\lim_{T\to\infty} \sum_{l=1}^{L} C'_T\left[x, y'_l\right] \leq N. \tag{2.3}$$

Finally Theorem 7 holds provided its statement is replaced by

Corollary 14. *Consider a dynamical system with fading memory as described above with N accessible variables $x_i(t)$. Because the dynamical system has fading memory, we can identify the output functions with functions $x_i(u^{-\infty})$ in $\mathcal{H}_{U^\infty}$. Consider an increasing family of functions in $\mathcal{H}_{U^\infty}$: $Y'_L = \{y'_1, ..., y'_L\}$ with $Y_L \subseteq Y_{L'}$ if $L' \geq L$ and $y'_l \in \mathcal{H}_{U^\infty}$, $\langle y'_l, y'_{l'} \rangle_{U^h} - \langle y'_l \rangle_{U^h} \langle y'_{l'} \rangle_{U^h} = {c_l}^2 \delta_{ll'}$, $l, l' = 1, ..., L.$, such that in the limit $L \to \infty$, the sets Y_L tend towards a complete set of functions in $\mathcal{H}_{U^\infty}$. Suppose that the readout functions $x_i(u^{-\infty})$ and the basis functions*

$y'_l(u^{-\infty})$ have finite fourth order moment: $\langle x_i^4\rangle_{U^\infty} < \infty$, $\langle y_l'^4\rangle_{U^\infty} < \infty$. Consider the limit of an infinite data set $T \to \infty$ and infinite initialization time $T' \to \infty$. Suppose the covariance matrix $R_{ii'} = \lim_{T,T'\to\infty} \mathrm{cov}(x_i x_{i'})$ has rank N. Then the sum of the capacities $C'_T[X, y_l]$ for the sets Y'_L tends towards the number N of output functions:

$$\lim_{Y'_L \to complete\ set} \left[\lim_{T,T'\to\infty} \sum_{l=1}^{L} C'_T[X, y'_l] \right] = N \ . \tag{2.4}$$

We sketch the proofs.

Eq. (2.2) is proven by inserting $x'_i = x_i - \langle x_i\rangle_T$ and $z' = z - \langle z\rangle_T$ into eq. (1.7) .

To prove Corollaries 13 and 14, we replace the original dynamical system with variables $X \equiv (x_1, ..., x_N)$, by a new dynamical system with variables $X' \equiv (x'_1, ..., x'_N)$ where $x'_i = x_i - \lim_{T\to\infty}\langle x_i\rangle_T$, and we replace all sets of functions $Y'_L = \{y'_1, ..., y'_L\}$ by $Y_L = \{y_1, ..., y_L\}$ with $y_i = y'_i - \langle y'_i\rangle_{U^\infty}$. Theorems 4 and 7 hold for the new dynamical system X' and the sets Y_L. Using the fact that $\lim_{T\to\infty} C'_T(X, y'_l) = \lim_{T\to\infty} C_T(X, y'_l)$ we recover Corollaries 13 and 14.

3 Estimating capacities in practice

3.1 Why the sum of capacities do not saturate the bound Theorem 7

Theorem 7 states that under very general conditions, when the state of a fading memory dynamical system depends only on the history of its inputs $u(t)$, the sum of the capacities should tend towards the number of linearly independent readout functions. In practice this limit is generally not attained precisely. There are several possible reasons for this underestimation:

1. The fading memory condition of theorem 7 is not satisfied. This can in principle be checked by measuring the Lyapunov exponents of the input driven system. That is, starting the system in slightly different initial conditions, and checking whether the system states diverge. As the result can depend on the specific input stream and operating point of the system, a very large number of samples would need to be taken. Another indication of the fading memory properties of the system can be given by measuring the so-called maximal *local Lyapunov exponent* of the system [9]. This is an estimate of the mean maximal singular value of the system's Jacobian, where the mean is taken across the operating points of the driven system.

2. The system is affected by noise. Particularly for any experimental system, one expects that the system is affected by many noise sources. These can be thought of as additional inputs. However the Theorem 7 only applies if one takes into account all inputs to the system, including the noise. If some of the inputs are unknown, then one is not summing over a complete set of functions so the bound will not saturate.

3. The convergence in theorem 7 is too slow. More precisely if we have at our disposal a finite data set $x(t)$ collected in time T. When computing the capacities $C_T[X, y_l]$ one is essentially computing the coefficients of the decomposition of the function x_i onto the chosen basis y_l. Because of the finite statistics, we must restrict the estimation to a finite set $l = 1, ..., L$. In general the capacities for $l > L$ outside this set will be nonzero, leading to an underestimate. In addition this decomposition will have many very small but nonzero coefficients. A finite data set will not be sufficient to estimate them reliably. As discussed below, contributions that are too small to be estimated accurately should be set to zero, resulting in an underestimate of the total capacity. A related issue is that it may be that one of the capacities $C_T[X, y_l]$ is large, but that the index l is very large, and because of finite statistics and finite analysis time, we did not estimated this specific capacity.

4. The outputs are linearly independent but very similar. In this case, the covariance matrix $\langle x_i x_j\rangle_T$ of the outputs may be ill conditioned and its inversion, which is needed to compute the capacities, may become inaccurate. This is also expressed by a very large value of the condition number (the ratio between the largest and smallest eigenvalues) of $\langle x_i x_j\rangle_T$. As was discussed in [6], when the condition number is very large, the information carried by the corresponding principal components can be hard to extract accurately.

3.2 Overestimating small capacities

Proposition 3 states that the capacities are normalized $0 \leq C_T[X, y_l] \leq 1$. If y_l constitute a complete set, then most of the capacities $\lim_{T\to\infty} C_T[X, y_l]$ will be very small. However, as we discuss below and have extensively checked

on numerical examples, for finite times T the estimates of the capacities are affected by a systematic positive error of order $O(N/T)$.

To understand the origin of this systematic error, consider the case where $\lim_{T\to\infty}\langle\overline{x}_i\overline{x}_j\rangle_T = \delta_{ij}$, $\lim_{T\to\infty}\langle y^2\rangle_T = 1$, and $\lim_{T\to\infty}\langle\bar{x}_i y\rangle_T = 0$, so that $\lim_{T\to\infty} C_T[X,y] = 0$. Write the estimated capacity as

$$C_T[X,y] = \sum_{i,j=1}^{N} \frac{\langle y,\overline{x}_i\rangle_T \langle\overline{x}_i\overline{x}_j\rangle_T^{-1}\langle\overline{x}_j,y\rangle_T}{\langle y^2\rangle_T} \simeq \sum_{i=1}^{N}\langle y,\overline{x}_i\rangle_T^2 = \sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} y(t)\overline{x}_i(t)\right)^2 .$$

We first consider the simple case where $\overline{x}_i(t)$ and $y(t)$ are independent identically distributed (i.i.d.) random variables for all $t = 1,..,T$. From the normalization these random variables have variance 1, and independence implies that $\lim_{T\to\infty}\langle\overline{x}_i^2 y^2\rangle_T = 1$. Furthermore, the central limit theorem then implies that $\frac{1}{T}\sum_{t=1}^{T} y(t)\overline{x}_i(t) \sim N(0,\frac{1}{\sqrt{T}})$. Therefore the estimated capacity $C_T[X,y]$ is the sum of squares of normally distributed random variables, hence following a $\chi^2(N)$ distribution: $C_T[X,y] \sim \frac{1}{T}\chi^2(N)$. It is positive, has mean N/T and variance $2N/T^2$.

In most cases $\overline{x}_i(t)$ and $y(t)$ will not be independent. A good illustrative example consists of taking $y(t) = u(t)$ and $\bar{x}_i(t) = u(t)\,u(t-1)$. We still expect that $\frac{1}{T}\sum_{t=1}^{T} y(t)\overline{x}_i(t)$ has a gaussian distribution. If $\lim_{T\to\infty} C_T[X,y] = 0$, then the mean of this gaussian is zero. Its variance is in general difficult to compute. For illustrative purposes, we consider the case where $\overline{x}_i(t)$ and $y(t)$ at different times are independent. In this case the variance is given by $\frac{1}{T}\lim_{T\to\infty}\langle\overline{x}_i^2 y^2\rangle_T \geq \frac{1}{T}\lim_{T\to\infty}\langle\overline{x}_i^2\rangle_T\langle y^2\rangle_T = \frac{1}{T}$. Therefore $C_T[X,y]$ will be distributed as a sum of square of gaussians, each with mean zero and variance c_i/T with $c_i \geq 1$, i.e., $C_T[X,y]$ is given by a generalised $\chi^2(N)$ distribution with mean Nm/T and variance $2Nv/T^2$ for some constants $m, v \geq 1$.

This indicates that for finite times T the estimates of very small capacities are affected by a systematic positive error of order $O(N/T)$. Unfortunately the proportionality constant in this systematic error cannot be estimated a priori, except for independent variables.

In practice we proceed as follows: we choose a small probability p of mistakenly assigning a positive value to capacities which are in fact zero (for concreteness we take $p = 10^{-4}$ – this value should be adjusted according to how many capacities $C_T[X,y_l]$ one wants to estimate). We define the threshold t such that $P\left(\chi^2(N) \geq t\right) = p$. We then double this threshold to take into account that the estimate we have made only applies to independent variables. All capacities for which the estimate $C_T[X,y] < \frac{2t}{T}$ is found to be below this threshold (the factor 2 is the doubling just mentioned) are replaced by the value zero.

3.3 Searching for Non-zero capacities

To estimate the capacities we simulate numerically the dynamical system, recording a long set of data (typically $T = 10^5$ to 10^6 time steps). We then use the recordings of $x_i(t)$ to estimate the capacities. We proceed as follows:

We recall that the basis functions we use consist of all products of Legendre polynomials $y_l = \prod_i \mathcal{P}_{d_i}(u(t-i))$. Using finite data, we cannot estimate the capacities for all basis functions, since the set of indices $\{d_i\}$ is infinite. Also, as we do not know in advance the type of computation a system performs, we prefer not to fix in advance the set of basis functions we will consider. Rather, we have assumed an exploration strategy of the capacity space that, to a large extent, succeeds in determining the relevant set of basis functions autonomously. We assume that the capacities will generally decrease when the degree of the polynomial and the delay increase. The former assumption was found to be true in all systems we studied experimentally, whereas the latter is generally true in average for physical systems with fading memory. In practice we go over the indices $\{d_i\}$ using 5 nested loops. These loops are chosen such that the basis functions are enumerated by "increasing complexity".

From outer loop to inner loop, these are: degree (the total degree of the basis function, i.e., the sum of all individual polynomial degrees); variables (i.e., the number of time steps the basis function depends on, bounded from above by the degree; power list (i.e., the degrees of the individual polynomials); window (i.e., the largest delay minus the smallest delay plus 1, bounded from below by the number of variables; positions (i.e., delay values of the variables within the window); delay (i.e., the smallest delay used in any of the basis functions). Using this loop nest ensures that each basis function is counted only once.

In order to decide when to stop each loop, we assume that the capacities become smaller and smaller as the complexity of the polynomial increases, i.e., as we go further and further in each loop. The threshold of section 3.2 is exploited exiting each loop whenever no scores above threshold are found.

This procedure is further modified as follows: in the outermost loop we treat the even and odd degrees separately. We note also that for systems with delayed response or damped oscillations, one should not stop the delay loop too soon. Depending on the system, the following approaches still allow autonomous exploration: ignoring the loop

exit conditions up to a certain delay and/or to using a low-passed version of the measured capacity in the loop exit condition.

3.4 Reaction Diffusion system

The Gray-Scott Reaction Diffusion (RD) system, supplemented by a time dependent input, is described by the equations

$$\begin{cases} \frac{\partial a_{xy}}{\partial \tau} = d_a \nabla^2 a_{xy} - a_{xy} b_{xy}^2 + f\left(1.0 + \iota w_{xy} u\left(\tau\right) - a_{xy}\right) \\ \frac{\partial b_{xy}}{\partial \tau} = d_b \nabla^2 b_{xy} + a_{xy} b_{xy}^2 - \left(f + k\right)\, b_{xy} \end{cases}$$

In these equations, the first term models the diffusion term, the second term the conversion of reagent A into B (with normalised reaction rate). Reagent B is removed from the RD system through a semi-permeable membrane with permeability $f + k$. The concentration of B on the other side of the membrane is 0.

Reagent A is supplied to the RD system through a semipermeable membrane with permeability f. In the usual Gray-Scott system the abundance of A on the other side of the membrane is 1.0, and the rate at which A is supplied is $f(1.0 - a_{xy})$.

In our work we drive the RD system with a time dependent input $u(t)$ where t is discrete time. We first convert this into a continuous time input through $u(\tau) = u(t)$ for $tT_S \leq \tau < (t+1)T_S$ for some sampling period T_S. We suppose that the concentration of A on the other side of the membrane is given by $1.0 + \iota w_{xy} u\left(\tau\right)$ where $w_{xy} \in [-1, +1]$ is a function of x and y, and $\iota \in [-1, +1]$ is a parameter used to determine how strong is the external input. The spatial dependence of w_{xy} is important, as it breaks the translational symmetry of the system.

We chose as parameters $d_a = 0.01$ and $d_b = 0.02$, $f = 0.022$, $k = 0.02$. Linear stability analysis indicates that in this parameter region the system is bistable. We took the spatial extent of the system to be 40×40 (in arbitrary spatial units), with periodic boundary conditions. Readout probes measuring the concentration of A were placed in a regular square grid with spacing=8.0, yielding $N = 25$ readouts. One sample is taken per sampling period T_S, yielding the readout functions $x_i(t)$.

For numerical integration, we used a spatial discretization step $\Delta x = 0.4$ (yielding a 100x100 effective grid size). The integration scheme used (Cellular Array model) is described in [7] and further discussed in [8]. In these works it is recommended to use an internal time step of $\Delta t = \Delta x^2/(6.max(d_a, d_b))$ which corresponds in our case to $\Delta t = 4/3$ (in arbitrary time units).

At each of the 100x100 grid points, an input weight w_{xy} was drawn uniformly at random from the interval $[-1, +1]$.

In our work we investigated 10 different sampling periods comprised between $4 \leq T_S \leq 40$. Using this parameter to change how the system processes the external information is interesting as it does not change the undriven dynamics. We observe that as T_S increases, the system looses linear memory, and becomes increasingly nonlinear. If T_S got too large, the system would evolve to a steady state for each input, loosing all memory.

References

[1] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, third edition, 2001, Oxford University Press, Oxford, U.K.

[2] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, second edition, 1975, Academic Press, New York

[3] Bruce Hajek, *An Exploration of Random Processes for Engineers*, 2011

[4] Christopher G. Small, Don L. McLeish, *Hilbert Space Methods in Probability and Statistical Inference*, Wiley Series in Probability and Statistics

[5] H. Jaeger, *Short Term Memory in Echo State Networks*, Fraunhofer Institute for Autonomous Intelligent Systems, Technical report:GMD report 152, 2002

[6] Michiel Hermans and Benjamin Schrauwen, *Memory in linear recurrent neural networks in continuous time,* Neural Networks, Vol. 23(3), pp. 341-355, 2010

[7] Rui Dilao and Joaquim Sainhas (1997) Validation and Calibration of Models for Reaction-Diffusion Systems, axXiv:patt-sol/9712007v1

[8] A. Adamatzki, Computing in nonlinear media and automata collectives, Institute of Physics (IoP) Publishing, 2001, ISBN 0 7503 0751

[9] H. Wolff, Local Lyapunov exponents: looking closely at chaos, Journal of the Royal Statistical Society, Series B (methodological), pp. 353-371, 1992