Dynamical systems as temporal feature spaces

Tino, Peter

Journal of Machine Learning Research 21 (2020) 1-42 Submitted 7/19; Revised 12/19; Published 2/20

Dynamical Systems as Temporal Feature Spaces

Peter Tino P.Tino@cs.bham.ac.uk
School of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK

Editor: Sayan Mukherjee

Abstract

Parametrised state space models in the form of recurrent networks are often used in machine learning to learn from data streams exhibiting temporal dependencies. To break the black box nature of such models it is important to understand the dynamical features of the input-driving time series that are formed in the state space. We propose a framework for rigorous analysis of such state representations in vanishing memory state space models such as echo state networks (ESN). In particular, we consider the state space a temporal feature space and the readout mapping from the state space a kernel machine operating in that feature space. We show that: **(1)** The usual ESN strategy of randomly generating input-to-state, as well as state coupling leads to shallow memory time series representations, corresponding to cross-correlation operator with fast exponentially decaying coefficients; **(2)** Imposing symmetry on dynamic coupling yields a constrained dynamic kernel matching the input time series with straightforward exponentially decaying motifs or exponentially decaying motifs of the highest frequency; **(3)** Simple ring (cycle) high-dimensional reservoir topology specified only through two free parameters can implement deep memory dynamic kernels with a rich variety of matching motifs. We quantify richness of feature representations imposed by dynamic kernels and demonstrate that for dynamic kernel associated with cycle reservoir topology, the kernel richness undergoes a phase transition close to the edge of stability.

Keywords: Recurrent Neural Network, Echo State Network, Dynamical Systems, Time Series, Kernel Machines

1. Introduction

When dealing with time series data, techniques of machine learning and signal processing must account in some way for temporal dependencies in the data stream. One popular option is to impose a parametric state-space model structure in which the state vector is supposed to dynamically code for the input time series processed so far and the output is determined through a static readout from the state. Recurrent neural networks (e.g. (Downey et al., 2017)), Kalman filters (Kalman, 1960) or hidden Markov models (Baum and Petrie, 1966) represent just a few examples of this approach. In some cases the state space and transition structure is (at least partially) imposed based on the relevant prior knowledge (Yoon, 2009), but usually it is learnt from the data along with the readout map. In the case of uncountable state space and non-linear state dynamics, the use of gradient methods in learning the state transition dynamics is hampered by the well known

"information latching problem" (Bengio et al., 1993). As temporal sequences increase in length, the influence of early components of the sequence have less impact on the network output. This causes the partial gradients, used to update the weights, to (exponentially) shrink to zero as the sequence length increases. Several approaches have been suggested to overcome this challenge, e.g. (Bengio et al., 1994; Hochreiter and Schmidhuber, 1997; Jing et al., 2019).

One possibility to avoid having to train the state transition part in a state space model is to simply initialise it randomly to a 'sensible' fading memory dynamic filter and only train the static readout part of the model. Models following this philosophy (Jaeger, 2001; Maass et al., 2002; Tino and Dorffner, 2001) have been termed "reservoir computation (RC) models" (Lukosevicius and Jaeger, 2009). Perhaps the simplest form of a RC model is the Echo State Network (ESN) (Jaeger, 2001, 2002a,b; Jaeger and Hass, 2004). Briefly, ESN is a recurrent neural network with a non-trainable state transition part (reservoir) and a simple trainable linear readout. Connection weights in the ESN reservoir, as well as the input weights are randomly generated. The reservoir weights are scaled so as to ensure the *"Echo State Property"* (ESP): the reservoir state is an *"echo"* of the entire input history and does not depend on the initial state. Scaling reservoir weights so that the largest singular value is smaller than 1 makes the reservoir dynamics contractive and guarantees the ESP. In practice, sometimes it is the spectral radius that guides the scaling. In this case, however, spectral radius < 1 does not guarantee the ESP.

ESNs has been successfully applied in a variety of tasks (Jaeger and Hass, 2004; Skowronski and Harris, 2006; Bush and Anderson, July 2005; Tong et al., 2007). Many extensions of the classical ESN have been suggested in the literature, e.g. deep ESN (Gallicchio et al., 2017), intrinsic plasticity (Schrauwen et al., 2008; Steil, 2007), decoupled reservoirs (Xue et al., 2007), leaky-integrator reservoir units (Jaeger et al., 2007), filter neurons with delay-and-sum readout (Holzmann and Hauser, 2009) etc. However, there are properties of the reservoir that are poorly understood (Xue et al., 2007) and specification of the reservoir and input connections require numerous trails and even luck (Xue et al., 2007). Furthermore, imposing a constraint on spectral radius or largest singular value of the reservoir matrix is a weak tool to properly set the reservoir parameters (Ozturk et al., 2007). Finally, random connectivity and weight structure of the reservoir is unlikely to be optimal and such a setting prevents us from providing a clear and systematic insight into the reservoir dynamics organisation (Ozturk et al., 2007; Rodan and Tino, 2010). Rodan and Tino (2010) demonstrated that even an extremely simple setting of a high-dimensional state space structure governed by only two free parameters set deterministically can yield modelling capabilities on par with other ESN architectures. However, a deeper understanding of why this is so has been missing.

In order to theoretically understand the workings of parametrised state space models as machine learning tools to process and learn from temporal data, there has been a lively research activity to formulate and assess different aspects of computational power and information processing capacity in such systems (e.g. (Dambre et al., 2012; Obst and Boedecker, 2014; Hammer, 2001; Hammer and Tino, 2003; Siegelmann and Sontag, 1994; Tino and Hammer, 2004)). For example, tools of information theory have been used to assess information storage or transfer within systems of this kind (Lizier et al., 2007, 2012; Obst et al., 2010; Bossomaier et al., 2016). To specifically characterise capability of input-driven dy-

namical systems to keep in their state-space information about past inputs, several memory quantifiers were proposed, for example "short term memory capacity" (Jaeger, 2002a) and "Fisher memory curve" (Ganguli et al., 2008; Tino, 2018). Even though those two measures have been developed from completely different perspectives, deep connections exist between them (Tino and Rodan, 2013). The concept of memory capacity, originally developed for univariate input streams, was generalised to multivariate inputs in (Grigoryeva et al., 2016). Couillet et al. Couillet et al. (2016) rigorously studied mean-square error of linear dynamical systems used as dynamical filters in regression tasks and suggested memory quantities that generalise the short term memory capacity and Fisher memory curve measures. Finally, Ganguli and Sompolinsky (2010) showed an interesting connection between memory in dynamical systems and their capacity to perform dynamical compressed sensing of past inputs.

In this contribution, we suggest a novel framework for characterising richness of dynamic representations of input time series in the form of states of a dynamical system, which is the core part of any state space model used as a learning machine. Our framework is based on the observation that the idea of fixed dynamic reservoir with simple static linear mapping build on top of it strikingly resembles the philosophy of kernel machines (Legenstein and Maass, 2007). There, the inputs are transformed using a fixed mapping (usually only implicitly defined) into a feature space that is "rich enough" so that in that space it is sufficient to train linear models only. The key tool for building linear models in the feature space is the inner product. One can grasp workings of a kernel machine by understanding of how the data is mapped to the feature space and what "data similarity" in the original space means when expressed as the inner product in the feature space. We will view the reservoir state space as a "temporal feature space" in which the linear readout is operating. In this view, the input time series seen by the reservoir model results in a state that codes all history of the presented input items so far and thus forms a feature representation of the time series. Different forms of coupling in the reservoir dynamical system will result in different temporal feature spaces with different feature representations of input time series, implying different notions of similarity between time series, expressed as inner products of their feature space representations. We will ask if and how the feature spaces differ in cases of traditional randomly generated reservoir models, as well as more constrained reservoir constructions studied in the literature.

Since RC models are input-driven non-autonomous dynamical systems, theoretical studies linking their information processing capabilities to the reservoir coupling structures have been performed mostly in the context of linear dynamics, e.g. (Ganguli et al., 2008; Couillet et al., 2016; Couillet et al., 2016; Tino, 2018). While such studies are of interest by themselves, in the context of the present work, studying linear dynamics can shed light on a wide class of RC models whose approximation capabilities equal those of non-linear systems. In particular, Grigoryeva, Gonon and Ortega recently proved a series of important results concerning universality of RC models (Grigoryeva and Ortega, 2018b,a; Gonon and Ortega, 2019). The universality can be obtained even if the state transition dynamics is linear, provided the readout map is polynomial (or a neural network)[1]. However, univer-

1. Universal approximation capability was first established in the L^∞ sense for deterministic, as well as almost surely uniformly bounded stochastic inputs (Grigoryeva and Ortega, 2018b). This was later

sality is a property of a whole *family* of RC models. For appropriate classes of filters[2] and input sources, it guarantees that for any filter and approximation precision, there exists a RC model approximating the filter to that precision. This is an existential statement that does relate individual filters to their approximating RC models. Our new framework will enable us to reason about what kind of RC model setup is necessary if filters with deeper memory were to be approximated. In particular, we will first investigate properties of linear dynamical readout kernels obtained on top of linear dynamical systems. Crucially, memory properties of such kernels can not be enhanced by moving from linear to polynomial static readout kernels. Loosely speaking, if feature representation $\mathbf{x}$ of a time series $\mathbf{u}$ captures properties of $\mathbf{u}$ only up to some look-back time $t-\tau_0$ from the last observation time t, then no nonlinear transformation γ of $\mathbf{x}$ can prolong memory τ_0 in the feature representation $\gamma(\mathbf{x})$ of $\mathbf{u}$. Hence, we will be able to make statements regarding appropriate settings of the linear dynamics that are necessary for universal approximation of deeper memory filters.

The paper has the following organisation: In section 2 we set the scene and outline the main intuitions driving this work. Section 3 formally introduces the notion of temporal kernel and provides some useful properties of the kernel to be used further in our study. In section 4 we will setup basic tools for characterising dynamic kernels - motifs and their corresponding motif weights. Starting from section 5, we will analyse dynamic kernels corresponding to different settings of the dynamical system. In particular, dynamical kernels associated with fully random, symmetric and highly constrained coupling of the dynamical system are analysed in sections 5, 6 and 7, respectively. We provide examples illustrating the developed theory and compare the motif richness of different parameter settings of the dynamical system in section 8. The paper finishes with discussion and conclusions in section 9.

2. Preliminary concepts and intuitions

We consider fading memory state space models with linear input driven dynamics in an N-dimensional state space and univariate inputs and outputs. Note that in the ESN metaphor, the state dimensions correspond to reservoir units coupled to the input $u(t)$ via an N-dimensional weight vector $\mathbf{w} \in \mathbb{R}^N$. Denoting the state vector at time t by $\mathbf{x}(t) \in \mathbb{R}^N$, the dynamical system evolves as

$$\mathbf{x}(t) = \mathbf{W}\ \mathbf{x}(t-1) + \mathbf{w}\ u(t), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is an $N \times N$ weight matrix providing the dynamical coupling. In state space models, the output $y(t)$ is often determined solely based on the current state $\mathbf{x}(t)$ through a readout function h:

$$y(t) = h(\mathbf{x}(t)). \tag{2}$$

The readout map h is typically trained (offline or online) by minimising the (normalised) mean square error between the targets and reservoir readouts $y(t)$.

extended in (Gonon and Ortega, 2019) to L^p, $1 \leq p < \infty$ and not necessarily almost surely uniformly bounded stochastic inputs.

2. transforming semi-infinite input sequences into outputs

Denote the set of natural numbers (including zero) by $\mathbb{N}_0$. In this contribution, we study how the dynamical system (1) extracts potentially useful information about the left infinite input time series $..., u(t-2), u(t-1), u(t)$, $u(-j) \in \mathbb{R}, j \in \mathbb{N}_0$, in its state $\mathbf{x}(t) \in \mathbb{R}^N$, since it is only the state $\mathbf{x}(t)$ that will be used to produce the predictive output $y(t)$ upon seeing the input time series up to time t. In particular, we will consider readout maps constructed in the framework of kernel machines. For example, in the case of linear Support Vector Machine (SVM) regression, the readout from the state space at time t has the form

$$y(t) = h(\mathbf{x}(t)) = \sum_i \beta_i \ \langle \mathbf{x}(t_i), \mathbf{x}(t) \rangle + b, \tag{3}$$

where $\beta_i \in \mathbb{R}$ and $b \in \mathbb{R}$ are weight coefficients and bias term, respectively and $\mathbf{x}(t_i)$ are support state vectors observed at important "support time instances" t_i. In the spirit of state space modelling discussed above, we consider the state $\mathbf{x}(t') \in \mathbb{R}^N$ reached after observing the time series $..., u(t'-2), u(t'-1), u(t')$, the feature state space representation of that time series. Hence (3) can also be written as

$$y(t) = \sum_i \beta_i \ K([...u(t_i-1), u(t_i)], [...u(t-1), u(t)]) + b, \tag{4}$$

where $K(\cdot, \cdot)$ is a time series kernel associated with the dynamical system (1),

$$K([...u(t_i-1), u(t_i)], [...u(t_j-1), u(t_j)]) = \langle \mathbf{x}(t_i), \mathbf{x}(t_j) \rangle . \tag{5}$$

In this context, the support time instances t_i can be viewed as end times of the "support time series" $..., u(t_i-2), u(t_i-1), u(t_i)$ observed in the past and deemed "important" for producing the outputs by the training algorithm trained on the history of the time series before the time step t.

The suggested viewpoint is illustrated in figure 1. There are three support time series $(..., u(t_1-2), u(t_1-1), u(t_1))$, $(..., u(t_2-2), u(t_2-1), u(t_2))$ and $(..., u(t_3-2), u(t_3-1), u(t_3))$ represented through the states $\mathbf{x}(t_1)$, $\mathbf{x}(t_2)$ and $\mathbf{x}(t_3)$, respectively. To evaluate the output at time t, the current feature space representation $\mathbf{x}(t)$ of $..., u(t-2), u(t-1), u(t)$ is compared with feature space representations $\mathbf{x}(t_i)$, $i = 1, 2, 3$, of the support time series through dot products.

We will next formalise these intuitions and then investigate the properties of state space feature representations of time series by dynamical systems. In particular, we will be interested in how different forms of dynamic coupling $\mathbf{W}$ influence richness of such feature representations and how they map to properties of the corresponding temporal kernel.

3. Temporal kernel defined by dynamical system

Without loss of generality, we will study feature state space representations under the dynamical system (1) of left-infinite time series $..., u(-2), u(-1), u(0)$, $\quad u(-j) \in \mathbb{R}$, $j \in \mathbb{N}_0$. We will assume that the largest singular value ν of the dynamic coupling $\mathbf{W}$ is strictly less than 1, making the dynamics (1) contractive. This means that the echo state property is fulfilled and for sufficiently long past horizons $\tau \gg 1$, the influence of initial state $\mathbf{x}(-\tau)$ on the feature representation of

$$u(-\tau+1), u(-\tau+2), ..., u(-1), u(0)$$

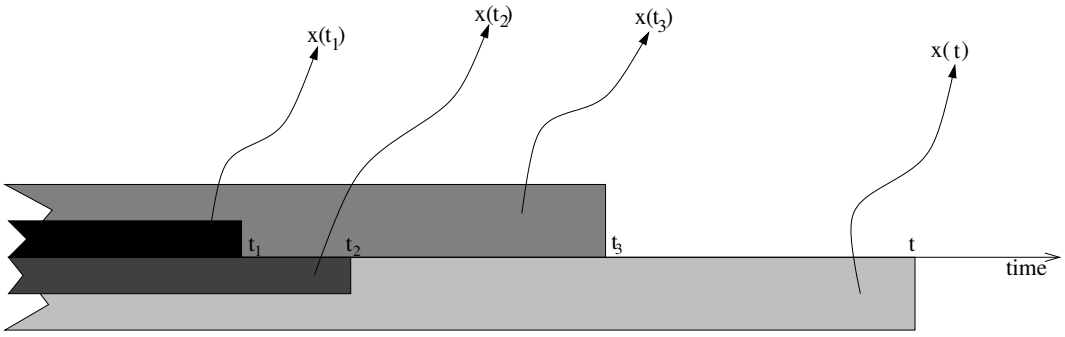

Figure 1: Illustration of the workings of kernel machine producing an output at time t after observing $..., u(t-1), u(t)$. The time series $..., u(t-1), u(t)$ is compared with the three support time series $(..., u(t_i - 1), u(t_i))$, $i = 1, 2, 3$, by evaluating dot products between their feature space representations $\mathbf{x}(t)$ and $\mathbf{x}(t_i)$.

is negligible. Note that $\nu < 1$ is a sufficient condition for the echo state property, but the property may actually be achieved under milder conditions, especially when particular input streams are considered (for formal treatment and further details see e.g. (Yildiz et al., 2012; Manjunath and Jaeger, 2013)). In this contribution we use $\nu < 1$, since it allows us **(1)** to consider arbitrary input streams over a bounded domain (the ESP is always guaranteed) and **(2)** to explicitly bound, in terms of properties of $\mathbf{W}$, the norm of dynamical states, as well as the extent to which the initial state influences the temporal kernel.

More formally, given a past horizon $\tau \gg 1$, we will represent the time series $u(-\tau + 1), u(-\tau + 2), ..., u(-1), u(0)$ as a vector $\mathbf{u}(\tau) = (u_1, u_2, ..., u_\tau)^\top \in \mathbb{R}^\tau$, where $u_i = u(-i + 1), i = 1, 2, ...\tau$. In other words $\mathbf{u}(\tau) = (u(0), u(-1), ..., u(-\tau + 1))^\top$.

Consider a state $\mathbf{x}(-\tau) \in \mathbb{R}^N$ at time $-\tau$. After seeing the input series $u(-\tau+1), u(-\tau+2), ..., u(-1), u(0)$, the new state of the dynamics (1) will be[3]

$$\mathbf{x}(0) = \mathbf{W}^\tau \mathbf{x}(-\tau) + \sum_{j=1}^{\tau} u(j-\tau)\mathbf{W}^{\tau-j}\mathbf{w}.$$

As discussed in the previous section, the state $\mathbf{x}(0)$ reached from the initial condition $\mathbf{x}(-\tau)$ after seeing $\mathbf{u}(\tau)$ codes for information content in $\mathbf{u}(\tau)$ and will be considered the "feature space representation" of $\mathbf{u}(\tau)$ through the dynamical system (1):

$$\begin{aligned} \phi(\mathbf{u}(\tau);\ \mathbf{x}(-\tau)) &= \mathbf{x}(0) \\ &= \mathbf{W}^\tau \mathbf{x}(-\tau) + \sum_{i=1}^{\tau} u(1-i)\mathbf{W}^{i-1}\mathbf{w} \\ &= \mathbf{W}^\tau \mathbf{x}(-\tau) + \sum_{i=1}^{\tau} u_i \mathbf{W}^{i-1}\mathbf{w}. \end{aligned} \tag{6}$$

Given two time series at past horizon τ represented through $\mathbf{u}(\tau) = (u_1, u_2, ..., u_\tau)^\top$ and $\mathbf{v}(\tau) = (v_1, v_2, ..., v_\tau)^\top$, the temporal kernel defined by dynamical system (1) evaluated on

3. $\mathbf{W}^0 = \mathbf{I}_{N\times N}$, the $N \times N$ identity matrix.

$\mathbf{u}(\tau)$ and $\mathbf{v}(\tau)$ reads:

$$K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{x}(-\tau)) = \langle \phi(\mathbf{u}(\tau);\ \mathbf{x}(-\tau)), \phi(\mathbf{v}(\tau);\ \mathbf{x}(-\tau)) \rangle . \tag{7}$$

We will now show that, as expected given the contractive nature of (1), for sufficiently long past time horizons $\tau \gg 1$ on input streams over bounded domain[4], the kernel evaluation is insensitive to the initial condition $\mathbf{x}(-\tau)$. This will allow us to simplify the presentation by setting $\mathbf{x}(-\tau)$ to the origin in the rest of the paper.

Theorem 1 *Consider the dynamical system (1) driven by time series over a bounded domain $[-U, U]$, $0 < U < \infty$, with a past time horizon $\tau > 1$. Assume that the largest singular value ν of the dynamic coupling $\mathbf{W}$ is strictly smaller than 1 and that the norm of the input coupling $\mathbf{w}$ satisfies $\|\mathbf{w}\| \leq B$. Assume further that the norm of the initial condition is upper bounded by $\|\mathbf{x}(-\tau)\| \leq A(\tau) = c \cdot \zeta^{-\tau}$, where $\nu < \zeta < 1$ and $c > 0$ is a large enough positive constant satisfying*

$$c \geq \frac{B \cdot U}{(1-\nu) \cdot \left(1 - \frac{\nu}{\zeta}\right)}. \tag{8}$$

Then, for any $\mathbf{u}(\tau), \mathbf{v}(\tau) \in [-U, U]^\tau$, it holds

$$K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{x}(-\tau)) = K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{0}) + \epsilon,$$

where

$$-\eta^\tau \left[\frac{2c}{1-\nu} \cdot B \cdot U\right] \leq\ \epsilon\ \leq \eta^\tau \left[c^2\ \eta^\tau + \frac{2c}{1-\nu} \cdot B \cdot U\right], \tag{9}$$

with $\eta = \nu/\zeta < 1$.

Proof Note that

$$\phi(\mathbf{u}(\tau);\ \mathbf{x}(-\tau)) = \mathbf{W}^\tau \mathbf{x}(-\tau) + \phi(\mathbf{u}(\tau);\ \mathbf{0})$$

and therefore, denoting

$$\phi(\mathbf{u}(\tau);\ \mathbf{0}) = \sum_{i=1}^{\tau} u_i \mathbf{W}^{i-1} \mathbf{w} \tag{10}$$

by $\phi_0(\mathbf{u}(\tau))$, we have

$$\begin{aligned} K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{x}(-\tau)) &= \langle \mathbf{W}^\tau \mathbf{x}(-\tau) + \phi_0(\mathbf{u}(\tau)), \mathbf{W}^\tau \mathbf{x}(-\tau) + \phi_0(\mathbf{v}(\tau)) \rangle \\ &= \|\mathbf{W}^\tau \mathbf{x}(-\tau)\|_2^2 + \langle \mathbf{W}^\tau \mathbf{x}(-\tau), \phi_0(\mathbf{u}(\tau)) + \phi_0(\mathbf{v}(\tau)) \rangle \\ &+ K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{0}), \end{aligned}$$

4. It is common in the ESN literature to consider input streams over a bounded domain (e.g. Jaeger (2001)). In the recent work on universality of ESNs Grigoryeva and Ortega (2018b) consider almost surely uniformly bounded stochastic inputs. This is further relaxed in (Gonon and Ortega, 2019).

where

$$K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{0}) = \langle \phi_0(\mathbf{u}(\tau)), \phi_0(\mathbf{v}(\tau)) \rangle$$

is the dynamic kernel evaluated using initial condition $\mathbf{x}(-\tau)$ set to the origin $\mathbf{0}$. We have,

$$\|\mathbf{W}^\tau \mathbf{x}(-\tau)\|_2^2 \leq \nu^{2\tau} \cdot (A(\tau))^2. \tag{11}$$

Note that

$$\langle \mathbf{W}^\tau \mathbf{x}(-\tau), \phi_0(\mathbf{u}(\tau)) \rangle \leq \|\mathbf{W}^\tau \mathbf{x}(-\tau)\| \cdot \|\phi_0(\mathbf{u}(\tau))\|$$

and (see (10))

$$\begin{aligned} \|\phi_0(\mathbf{u}(\tau))\| &\leq \sum_{i=1}^{\tau} \nu^{i-1} \cdot U \cdot \|\mathbf{w}\| \\ &\leq B \cdot U \cdot \frac{1}{1-\nu}, \end{aligned} \tag{12}$$

yielding

$$\langle \mathbf{W}^\tau \mathbf{x}(-\tau), \phi_0(\mathbf{u}(\tau)) \rangle \leq \frac{\nu^\tau}{1-\nu} \cdot A(\tau) \cdot B \cdot U.$$

We thus have

$$K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{x}(-\tau)) = K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{0}) + \epsilon,$$

with

$$\epsilon \leq \nu^\tau \left[\nu^\tau (A(\tau))^2 + \frac{2}{1-\nu} \cdot A(\tau) \cdot B \cdot U \right].$$

To evaluate the lower bound on ϵ, note that

$$\begin{aligned} \langle \mathbf{W}^\tau \mathbf{x}(-\tau), \phi_0(\mathbf{u}(\tau)) \rangle &\geq -\|\mathbf{W}^\tau \mathbf{x}(-\tau)\| \cdot \|\phi_0(\mathbf{u}(\tau))\| \\ &\geq \frac{-\nu^\tau}{1-\nu} \cdot A(\tau) \cdot B \cdot U. \end{aligned}$$

Since, trivially, $\|\mathbf{W}^\tau \mathbf{x}(-\tau)\|_2^2 \geq 0$, we have

$$\epsilon \geq -\frac{2\nu^\tau}{1-\nu} \cdot A \cdot B \cdot U.$$

We have thus obtained,

$$-\nu^\tau \left[\frac{2}{1-\nu} \cdot A(\tau) \cdot B \cdot U \right] \leq \epsilon \leq \nu^\tau \left[\nu^\tau (A(\tau))^2 + \frac{2}{1-\nu} \cdot A(\tau) \cdot B \cdot U \right],$$

which is equivalent to (9).

In order to reconcile this setting with the dynamics (1), consider a past horizon $\tau + \tau_0$ for some additional look-back time $\tau_0 \geq 1$. We require,

$$\|\mathbf{x}(-\tau)\| \leq A(\tau) = c \cdot \zeta^{-\tau} \quad \text{and} \quad \|\mathbf{x}(-\tau - \tau_0)\| \leq A(\tau + \tau_0) = c \cdot \zeta^{-\tau-\tau_0}.$$

But from the dynamics (1), we also have (see eqs.(6), (11) and (12)),

$$\begin{aligned} \|\mathbf{x}(-\tau)\| &\leq \|\mathbf{W}^{\tau_0}\ \mathbf{x}(-\tau - \tau_0)\| + \frac{B \cdot U}{1-\nu} \\ &\leq \nu^{\tau_0} \cdot A(\tau + \tau_0) + \frac{B \cdot U}{1-\nu}. \end{aligned} \tag{13}$$

We would like the norm of the state $\mathbf{x}(-\tau)$ reached from the initial state $\mathbf{x}(-\tau-\tau_0)$ (bounded in norm by $A(\tau+\tau_0)$) to be within the required bound $A(\tau)$. In other words, we would like

$$A(\tau + \tau_0) \cdot \nu^{\tau_0} + \frac{B \cdot U}{1-\nu} < c \cdot \zeta^{-\tau} \tag{14}$$

to hold. Using $A(\tau + \tau_0) = c \cdot \zeta^{-\tau-\tau_0}$, we conclude that the inequality (14) holds when

$$c > \zeta^{\tau} \cdot \frac{B \cdot U}{(1-\nu)} \cdot \frac{1}{1 - \left(\frac{\nu}{\zeta}\right)^{\tau_0}}. \tag{15}$$

Since $0 < \eta = \nu/\zeta < 1$, for $\tau, \tau_0 \geq 1$,

$$\zeta^{\tau} \cdot \frac{B \cdot U}{(1-\nu)} \cdot \frac{1}{1-\eta^{\tau_0}} \ < \ \frac{B \cdot U}{(1-\nu) \cdot (1-\eta)}$$

we have that the inequality (15) is definitely satisfied when

$$c > \frac{B \cdot U}{(1-\nu) \cdot (1 - \frac{\nu}{\zeta})}.$$

■

Theorem 1 formally states that because the dynamical system (1) is contractive, the influence of the initial condition $\mathbf{x}(-\tau)$ on the kernel value $K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{x}(-\tau))$ decays exponentially with the past time horizon τ. For sufficiently long past time horizons $\tau \gg 1$ we can thus set $\mathbf{x}(-\tau) = \mathbf{0}$. Hence, in the rest of this study we will assume $\tau \geq N$ and (unless necessary) we will drop specific reference to τ by writing $\mathbf{u}$ instead of $\mathbf{u}(\tau) \in \mathbb{R}^{\tau}$. In fact, it will be easier to think of time horizons in units of N, so that $\tau = \ell \cdot N$, for some sufficiently large integer $\ell > 1$. Furthermore, we will refer to $\phi(\mathbf{u}(\tau);\ \mathbf{0})$ and $K(\mathbf{u}(\tau), \mathbf{v}(\tau);\ \mathbf{0})$ simply as $\phi(\mathbf{u})$ and $K(\mathbf{u}, \mathbf{v})$, respectively.

4. Temporal kernel and its motifs

In the previous section we established that the temporal kernel associated with dynamical system (1) and acting on time series with past time horizon $\tau \gg 1$ is defined as

$$K(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle \,. \tag{16}$$

In order to analyse the action of $K(\mathbf{u}, \mathbf{v})$ on time series $\mathbf{u}, \mathbf{v}$, we need to find its expression directly in terms of $\mathbf{u}$ and $\mathbf{v}$. The next theorem shows that there exists a matrix $\mathbf{Q}$ of rank at most N that acts as a metric tensor on a subspace of $\mathbb{R}^\tau$ (of dimensionality at most N), so that $K(\mathbf{u}, \mathbf{v})$ can be expressed as a quadratic form $\mathbf{u}^\top \mathbf{Q}\ \mathbf{v}$. This will allow us to study properties of $K(\mathbf{u}, \mathbf{v})$ by analysing the associated metric tensor $\mathbf{Q}$.

Theorem 2 *Consider the dynamical system (1) of state dimensionality N and a dynamic coupling $\mathbf{W}$ with largest singular value $0 < \nu < 1$. Let $K(\mathbf{u}, \mathbf{v})$ (16) be the temporal kernel associated with system (1). Then for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\tau$,*

$$K(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{Q}\ \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{Q}} \,,$$

where $\mathbf{Q}$ is a symmetric, positive semi-definite $\tau \times \tau$ matrix of rank $N_m = \mathrm{rank}(\mathbf{Q}) \leq N$ and elements

$$Q_{i,j} = \mathbf{w}^\top \left(\mathbf{W}^\top \right)^{i-1}\ \mathbf{W}^{j-1}\ \mathbf{w}, \quad i, j = 1, 2, ..., \tau. \tag{17}$$

The upper bound on absolute values of $Q_{i,j}$ decays exponentially with increasing time indices $i, j = 1, 2, ..., \tau$, as

$$|Q_{i,j}| \ \leq\ \nu^{i+j-2}\ \|\mathbf{w}\|_2^2. \tag{18}$$

Proof First, we write

$$\begin{aligned}
K(\mathbf{u}, \mathbf{v}) &= \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle \\
&= \left\langle \sum_{i=1}^{\tau} u_i \mathbf{W}^{i-1} \mathbf{w}, \sum_{j=1}^{\tau} v_j \mathbf{W}^{j-1} \mathbf{w} \right\rangle \quad \text{(eq. (10))} \\
&= \sum_{i,j=1}^{\tau} u_i\ v_j\ \left\langle \mathbf{W}^{i-1} \mathbf{w}, \mathbf{W}^{j-1} \mathbf{w} \right\rangle \\
&= \sum_{i,j=1}^{\tau} u_i\ v_j\ Q_{i,j} \\
&= \mathbf{u}^\top \mathbf{Q}\ \mathbf{v}.
\end{aligned}$$

Second, $\phi(\mathbf{u})$ can be written as $\phi(\mathbf{u}) = \Phi \mathbf{u}$, where Φ is an $N \times \tau$ matrix whose i-th column is equal to $\mathbf{W}^{i-1}\mathbf{w}$. Hence, $K(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \Phi^\top \Phi\ \mathbf{v}$ and $\mathbf{Q} = \Phi^\top \Phi$ is symmetric positive semi-definite with rank at most $N \leq \tau$.

Finally, since $\|\mathbf{W}^i\mathbf{w}\| \leq \|\mathbf{W}^i\|\|\mathbf{w}\| \leq \nu^i\|\mathbf{w}\|$, we have

$$\begin{aligned} |Q_{i,j}| &= |\left\langle \mathbf{W}^{i-1}\mathbf{w}, \mathbf{W}^{j-1}\mathbf{w} \right\rangle| \\ &\leq \|\mathbf{W}^{i-1}\mathbf{w}\|_2 \cdot \|\mathbf{W}^{j-1}\mathbf{w}\|_2 \\ &\leq \nu^{i+j-2} \ \|\mathbf{w}\|_2^2. \end{aligned}$$

■

Note that $K(\cdot,\cdot)$ is a semi-inner product on $\mathbb{R}^\tau$. In other words, time series $\mathbf{u} \in \ker(\mathbf{Q})$ from the kernel of the linear operator $\mathbf{Q}$ have zero length. It acts as an inner product in the quotient of $\mathbb{R}^\tau$ by $\ker(\mathbf{Q})$, $\mathbb{R}^\tau / \ker(\mathbf{Q})$ (image of $\mathbf{Q}$). Since this distinction is not crucial for our argumentation, in order not to unnecessarily complicate the presentation, (slightly abusing mathematical terminology) we will refer to $K(\cdot,\cdot)$ as temporal kernel and to $\mathbf{Q}$ as the associated metric tensor.

Theorem 2 tells us that $K(\cdot,\cdot)$ is a fading memory temporal kernel and we can unveil its inner workings through eigen-analysis of $\mathbf{Q}$:

$$\mathbf{Q} = \mathbf{M}\Lambda\mathbf{M}^\top, \tag{19}$$

where the columns of $\mathbf{M}$ are the eigenvectors $\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_\tau \in \mathbb{R}^\tau$ of $\mathbf{Q}$ with the corresponding real non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_\tau$ arranged on the diagonal of the diagonal matrix Λ. Based on theorem 2, there are $N_m \leq N \leq \tau$ eigenvectors $\mathbf{m}_i$ with positive eigenvalue $\lambda_i > 0$.

Given two time series $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\tau$ of past time horizon τ, the temporal kernel value is

$$\begin{aligned} K(\mathbf{u},\mathbf{v}) &= \mathbf{u}^\top \ \mathbf{Q} \ \mathbf{v} \\ &= \left(\Lambda^{\frac{1}{2}}\mathbf{M}^\top\mathbf{u}\right)^\top \ \Lambda^{\frac{1}{2}}\mathbf{M}^\top\mathbf{v}. \end{aligned} \tag{20}$$

This has the following interpretation. In order to determine the kernel based "similarity" $K(\mathbf{u},\mathbf{v})$ of two time time series $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\tau$, both time series are first represented through a series of matching scores with respect to a potentially small number of filters $\mathbf{m}_i$ ($N_m \leq N \leq \tau$) , weighted by $\lambda_i^{1/2}$:

$$\widetilde{\mathbf{u}} = \left(\lambda_1^{1/2} \ \langle \mathbf{m}_1, \mathbf{u} \rangle, \lambda_2^{1/2} \ \langle \mathbf{m}_2, \mathbf{u} \rangle, ..., \lambda_{N_m}^{1/2} \ \langle \mathbf{m}_{N_m}, \mathbf{u} \rangle\right)^\top \in \mathbb{R}^{N_m} \tag{21}$$

and

$$\widetilde{\mathbf{v}} = \left(\lambda_1^{1/2} \ \langle \mathbf{m}_1, \mathbf{v} \rangle, \lambda_2^{1/2} \ \langle \mathbf{m}_2, \mathbf{v} \rangle, ..., \lambda_{N_m}^{1/2} \ \langle \mathbf{m}_{N_m}, \mathbf{v} \rangle\right)^\top \in \mathbb{R}^{N_m}.$$

Similarity between $\mathbf{u} \in \mathbb{R}^\tau$ and $\mathbf{v} \in \mathbb{R}^\tau$ is then evaluated as the degree to which both $\mathbf{u}$ and $\mathbf{v}$ match in the same way the highly weighted filters $\mathbf{m}_i$. Hence, instead of direct matching of $\mathbf{u}$ and $\mathbf{v}$, as would be the case for $\langle \mathbf{u}, \mathbf{v} \rangle$, we consider $\mathbf{u}, \mathbf{v}$ "similar" if $\langle \widetilde{\mathbf{u}}, \widetilde{\mathbf{v}} \rangle$ is high, in other words, if *both* $\mathbf{u}$ and $\mathbf{v}$ match well a number of significant filters $\mathbf{m}_i$ of high weight $\lambda_i^{1/2}$.

The matching scores $\langle \mathbf{m}_i, \mathbf{u} \rangle$ can be viewed as projections of the input time series $\mathbf{u}$ unto "prototypical" time series motifs $\mathbf{m}_i$ that characterise what features of the input time series are used by the kernel to assess their similarity. Loosely speaking, a temporal kernel employing a rich set of slowly decaying high-weight ("significant") motifs with deep memory will be able to perform more nuanced time series similarity evaluation than a kernel with a small set of highly constrained and fast decaying short memory motifs. In what follows we refer to $\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_{N_m} \in \mathbb{R}^\tau$ as *motifs of the temporal kernel* $K(\cdot, \cdot)$ with the associated *motif weights* given by $\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq ... \geq \lambda_{N_m}^{1/2} > 0$. In the light of the comments above, $K(\cdot, \cdot)$ acts as semi-inner product on $\mathbb{R}^\tau$ and as inner product on the span of the motifs, $\text{span}\{\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_{N_m}\}$.

In the case of SVM regression, the readout output for a time series $\mathbf{v} \in \mathbb{R}^\tau$, based on the state space representation of $\mathbf{v}$ through (1) would have the form (see eq.(4))

$$\sum_i \beta_i \; K(\mathbf{u}_i, \mathbf{v}) + b,$$

where $\mathbf{u}_i \in \mathbb{R}^\tau$ are the support vectors ("support time series"). This can be rewritten as a linear model $\mathbf{a}^\top \mathbf{v} + b$ with weight vector $\mathbf{a} \in \mathbb{R}^\tau$:

$$\begin{aligned} \mathbf{a}^\top &= \sum_i \beta_i \; \mathbf{u}_i^\top \; \mathbf{Q} \\ &= \sum_i \beta_i \; \sum_{j=1}^{M} \lambda_j \; \langle \mathbf{m}_j, \mathbf{u}_i \rangle \; \mathbf{m}_j^\top. \end{aligned} \tag{22}$$

Free parameters of the output-producing function are the coefficients β_i corresponding to the support time series $\mathbf{u}_i$. In contrast, motifs $\mathbf{m}_j$ and motif weights $\lambda_j^{1/2}$ are fixed by the dynamical system (1). Hence, whatever setting of the free parameters β_i one can come up with, the inherent memory and time series structures one can access in past data in order to produce the output for a newly observed time series are determined by the richness and memory depth characteristics of the motif set $\{\mathbf{m}_j\}_{j=1}^{N_m}$. In what follows we will take this viewpoint when analysing temporal kernels corresponding to the dynamical system (1) for different types of state space coupling $\mathbf{W} \in \mathbb{R}^{N \times N}$.

5. Random dynamic coupling W with zero-mean i.i.d. entries

It has been common practice in the reservoir computation community to generate dynamic coupling $\mathbf{W}$ of ESNs randomly (Lukosevicius and Jaeger, 2009), typically with elements of $\mathbf{W}$ generated independently from a zero-mean symmetric distribution and then normalised so that $\mathbf{W}$ has a desirable scaling property (e.g. certain spectral radius or largest singular value). In this section we investigate temporal kernels associated with such an ESN setting. We will see that the nature of motifs is remarkably stable (small set of shallow memory motifs), even though the couplings $\mathbf{W}$ are generated from a wide variety of distributions.

Consider a random matrix $\widetilde{\mathbf{W}}$ with elements $\widetilde{W}_{i,j}$, $i, j = 1, 2, ..., N$, generated i.i.d. from a zero-mean distribution with variance $\sigma_0^2 > 0$ and finite fourth moment. Such a realisation $\widetilde{\mathbf{W}} \in \mathbb{R}^{N \times N}$ will be rescaled to the desired largest singular value $\nu \in (0, 1)$:

$$\mathbf{W} = \frac{\nu}{\sigma_{max}(\widetilde{\mathbf{W}})} \widetilde{\mathbf{W}},$$

where $\sigma_{max}(\widetilde{\mathbf{W}})$ is the maximum singular value of $\widetilde{\mathbf{W}}$.

For large N, the largest eigenvalue of $N^{-1}\widetilde{\mathbf{W}}^\top\widetilde{\mathbf{W}}$ converges to $4\sigma_0^2$ almost surely (Rudelson and Vershynin, 2010; Tino, 2018). Hence, the largest singular value of $N^{-1/2}\ \widetilde{\mathbf{W}}$ approaches $2\sigma_0$. It follows that for large N, $\sigma_{max}(\widetilde{\mathbf{W}})$ approaches $2\sqrt{N}\sigma_0$. Rescaling

$$\mathbf{W} = \frac{\nu}{2\sqrt{N}\sigma_0}\widetilde{\mathbf{W}}$$

can be equivalently thought of as generating $W_{i,j}$ i.i.d. from a zero-mean distribution with standard deviation

$$\sigma = \sigma_0\frac{\nu}{2\sqrt{N}\sigma_0} = \frac{\nu}{2\sqrt{N}}. \tag{23}$$

We would like to reason, under the assumption of high state space dimensionality N of the dynamical system (1), about the properties of the metric tensor $\mathbf{Q}$ with elements $Q_{i,j}$ given by eq. (17).

To ease the mathematical notation, we denote the matrix $(\mathbf{W}^\top)^i\ \mathbf{W}^j$ by $\mathbf{A}^{(i,j)}$. Hence,

$$Q_{i,j} = \mathbf{w}^\top\ \mathbf{A}^{(i-1,j-1)}\ \mathbf{w}. \tag{24}$$

5.1. Diagonal elements of Q

The first diagonal element of $\mathbf{Q}$, $Q_{1,1}$, is trivially equal to $\|\mathbf{w}\|_2^2$, so let us first concentrate on $\mathbf{A}^{(1,1)}$ corresponding to $Q_{2,2}$.

$$\begin{aligned}\mathbf{A}^{(1,1)}_{j,j} &= N\left[\frac{1}{N}\sum_{i=1}^{N} W_{i,j}^2\right] \\ &\approx N\sigma^2 \\ &= \frac{\nu^2}{4}. \qquad \text{(see eq. (23))}\end{aligned} \tag{25}$$

The off-diagonal terms of $\mathbf{A}^{(1,1)}$ get asymptotically small as

$$\mathbf{A}^{(1,1)}_{i,j} = N\left[\frac{1}{N}\sum_{k=1}^{N} W_{k,i}\ W_{k,j}\right] \approx 0$$

since for $i \neq j$, $W_{k,i}$ and $W_{k,j}$ are uncorrelated and generated from zero-mean distribution with standard deviation $\sigma = \mathcal{O}(1/\sqrt{N})$ (see (23)). For large N we can thus approximate $\mathbf{A}^{(1,1)}$ as

$$\mathbf{A}^{(1,1)} \approx \frac{\nu^2}{4}\ \mathbf{I}_{N\times N}, \tag{26}$$

where $\mathbf{I}_{N\times N}$ is the identity matrix of rank N.

To approximate $\mathbf{A}^{(2,2)}$, we write

$$\begin{aligned}
\mathbf{A}^{(2,2)} &= (\mathbf{W}^\top)^2\ \mathbf{W}^2 \\
&= \mathbf{W}^\top\ \mathbf{A}^{(1,1)}\ \mathbf{W} \\
&\approx \frac{\nu^2}{4}\ \mathbf{W}^\top\ \mathbf{W} \\
&= \frac{\nu^2}{4}\mathbf{A}^{(1,1)} \\
&\approx \left(\frac{\nu^2}{4}\right)^2\ \mathbf{I}_{N\times N}.
\end{aligned} \tag{27}$$

Proceeding inductively, we obtain

$$\begin{aligned}
\mathbf{A}^{(k,k)} &= \mathbf{W}^\top\ \mathbf{A}^{(k-1,k-1)}\ \mathbf{W} \\
&\approx \frac{\nu^2}{4}\ \left(\frac{\nu^2}{4}\right)^{k-1}\ \mathbf{I}_{N\times N} \\
&= \left(\frac{\nu^2}{4}\right)^{k}\ \mathbf{I}_{N\times N}, \quad k = 2, 3, ..., \tau - 1.
\end{aligned} \tag{28}$$

We can thus approximate $Q_{j,j}$ as

$$\begin{aligned}
Q_{j,j} &= \mathbf{w}^\top\ \mathbf{A}^{(j-1,j-1)}\ \mathbf{w} \\
&\approx \left(\frac{\nu^2}{4}\right)^{j-1}\ \mathbf{w}^\top\ \mathbf{w} \\
&= \left(\frac{\nu}{2}\right)^{2(j-1)} \|\mathbf{w}\|_2^2.
\end{aligned} \tag{29}$$

Hence, the diagonal elements of $\mathbf{Q}$, necessarily non-negative since $\mathbf{A}^{(j,j)}$ are positive semi-definite, decay much faster (exponentially so, by the factor of $4^{-(j-1)}$) than the upper bound (18) of theorem 2,

$$Q_{j,j} \approx 4^{-(j-1)}\ \nu^{2(j-1)}\ \|\mathbf{w}\|_2^2. \tag{30}$$

In particular, if all elements of the input coupling $\mathbf{w}$ have the same absolute value w (with possibly different signs), we have

$$Q_{j,j} \approx N w^2 \left(\frac{\nu}{2}\right)^{2(j-1)}. \tag{31}$$

5.2. Off-diagonal elements of Q

We now investigate terms $Q_{i,j}$ for $i \neq j$. Since Q is symmetric, without loss of generality we can assume $j > i$. Then,

$$\begin{aligned}
\mathbf{A}^{(i-1,j-1)} &= (\mathbf{W}^\top)^{i-1}\ \mathbf{W}^{i-1}\ \mathbf{W}^{j-i} \\
&= \mathbf{A}^{(i-1,i-1)}\ \mathbf{W}^{j-i} \\
&\approx \left(\frac{\nu}{2}\right)^{2(i-1)}\ \mathbf{W}^{j-i}
\end{aligned} \tag{32}$$

The elements of $\mathbf{A}^{(i-1,j-1)}$ decay exponentially with increasing i (deeper past in the time series). We will now approximate $\mathbf{W}^{j-i}$.

Concentrate first on the sub- and super-diagonal elements of $\mathbf{Q}$. We have

$$\mathbf{A}^{(j,j+1)} \approx \left(\frac{\nu}{2}\right)^{2(j-1)} \mathbf{W}$$

and so besides the main diagonal elements $Q_{j,j} \approx (\nu/2)^{2(j-1)} \|\mathbf{w}\|_2^2$ we have sub- and super-diagonal elements

$$Q_{j+1,j} = Q_{j,j+1} \approx \left(\frac{\nu}{2}\right)^{2(j-1)} \mathbf{w}^\top \mathbf{W}\mathbf{w},$$

which, depending on $\mathbf{W}$ and $\mathbf{w}$, can be substantially smaller than $Q_{j,j}$. For example, if both $\mathbf{W}$ and $\mathbf{w}$ are generated element-wise independently from zero mean distributions, then for large N, $\mathbf{W}\mathbf{w} \approx 0$. This is because each row of $\mathbf{W}$ contains i.i.d. realisations of a random variable uncorrelated with random variable whose realisations are stored as elements of $\mathbf{w}$. Then $\mathbf{w}^\top \mathbf{W}\mathbf{w}$ is negligible.

For elements $Q_{i,j}$ further away from the diagonal, we first analyse properties of the matrix $\mathbf{B} = \mathbf{W}^2$.

$$\begin{aligned} B_{i,i} &= \sum_{k=1}^{N} W_{i,k}\, W_{k,i} \\ &= W_{i,i}^2 + \sum_{k\neq i} W_{i,k}\, W_{k,i} \\ &\approx W_{i,i}^2, \end{aligned}$$

because of uncorrelated $W_{i,k}$ and $W_{k,i}$ for $k \neq i$. Similarly, for $i \neq j$,

$$B_{i,j} = \sum_{k=1}^{N} W_{i,k}\, W_{k,j} \approx 0.$$

We have

$$\begin{aligned} Q_{j+2,j} = Q_{j,j+2} &\approx \left(\frac{\nu}{2}\right)^{2(j-1)} \mathbf{w}^\top\, \mathbf{B}\, \mathbf{w} \\ &\approx \left(\frac{\nu}{2}\right)^{2(j-1)} \sum_{i=1}^{N} W_{i,i}^2\, w_i^2. \end{aligned} \tag{33}$$

Note that in order to scale a large matrix $\widetilde{\mathbf{W}}$ generated i.i.d. from a zero-mean distribution to spectral radius less than one, the individual elements $W_{i,j}$ of the scaled matrix $\mathbf{W}$ need to be necessarily small, increasingly so for increasing dimensionality N. In particular, based on (23), the mean of $W_{i,i}^2$ is approximately $\nu^2/(4N)$. Comparing (see eq. (29))

$$Q_{j,j} \approx \left(\frac{\nu}{2}\right)^{2(j-1)} \sum_{i=1}^{N} w_i^2$$

with eq. (33), we see that there will be an increasing gap (with increasing state space dimensionality N) between the diagonal elements $Q_{j,j}$ of $\mathbf{Q}$ and the corresponding elements $Q_{j+2,j} = Q_{j,j+2}$ two places off the diagonal.

Continuing the preceding argumentation inductively, we can conclude that compared to the diagonal terms $Q_{j,j}$ of $\mathbf{Q}$, for the approximation purposes, the off-diagonal terms can be neglected and the metric tensor can be approximated by a diagonal matrix

$$\mathbf{Q} \approx \widehat{\mathbf{Q}} = \|\mathbf{w}\|_2^2 \ \ \mathrm{diag}\left(1, \left(\frac{\nu}{2}\right)^2, \left(\frac{\nu}{2}\right)^4, ..., \left(\frac{\nu}{2}\right)^{2(\tau-1)}\right). \tag{34}$$

5.3. Temporal kernel motifs generated by random W

The eigen-decomposition of $\widehat{\mathbf{Q}}$ is straightforward: The eigenvectors form the standard basis $\{\mathbf{e}_i\}$, each vector $\mathbf{e}_i$ containing zeros, except for the i-th element, which is equal to 1. The corresponding eigenvalues are equal to the diagonal elements of $\widehat{\mathbf{Q}}$,

$$\widehat{\lambda}_i = \|\mathbf{w}\|^2 \ \left(\frac{\nu}{2}\right)^{2(i-1)}. \tag{35}$$

This means that the motif $\mathbf{m}_i = \mathbf{e}_i$ extracts the i-th element from the history of the time series and weights it with the weight $\|\mathbf{w}\| \ (\nu/2)^{i-1}$.

Perhaps surprisingly, the temporal kernel defined by the dynamical system (1) with random coupling $\mathbf{W}$ generated i.i.d. from a zero mean distribution has a rigid Markovian flavor with shallow memory. In particular, the kernel

$$\begin{aligned} K(\mathbf{u}, \mathbf{v}) &= \sum_{i=1}^{N} \lambda_i \ \langle \mathbf{m}_i, \mathbf{u} \rangle \ \langle \mathbf{m}_i, \mathbf{v} \rangle \\ &\approx \sum_{i=1}^{N} \widehat{\lambda}_i \ \langle \mathbf{e}_i, \mathbf{u} \rangle \ \langle \mathbf{e}_i, \mathbf{v} \rangle \\ &\approx \|\mathbf{w}\|^2 \ \sum_{i=1}^{N} \left(\frac{\nu}{2}\right)^{2(i-1)} u_i \ v_i, \end{aligned}$$

compares the corresponding recent entries of the time series and weights down comparisons of past elements with rapidly decaying weights.

To illustrate this approximation, as well as the rapidly decaying memory of such temporal kernels, we considered 100-dimensional state space ($N = 100$) and generated 100 realisations of $\widetilde{\mathbf{W}}$ with elements $W_{i,j}$ randomly distributed according to the standard normal distribution $N(0, 1)$. Each $\widetilde{\mathbf{W}}$ was renormalised to $\mathbf{W}$ of largest singular value $\nu = 0.995$ and an input coupling vector $\mathbf{w}$ was generated as a random vector with elements generated i.i.d. according to $N(0, 1)$ and then renormalised to unit vector (length 1). We then imposed a past horizon $\tau = 200$ and calculated the metric tensor $\mathbf{Q}$, as well as its approximation $\widehat{\mathbf{Q}}$ (eq. (34)). In figure 2 we show the true motifs $\mathbf{m}_i$ (eigenvectors of $\mathbf{Q}$ for the first four dominant motifs (motifs with the largest 4 motif weights) as the mean and standard deviations across the 100 realisations. For clarity, we only show the first 10 dimensions. It is clear that the motifs approximately correspond to the first four standard basis vectors $\mathbf{e}_i$, $i = 1, 2, 3, 4$,

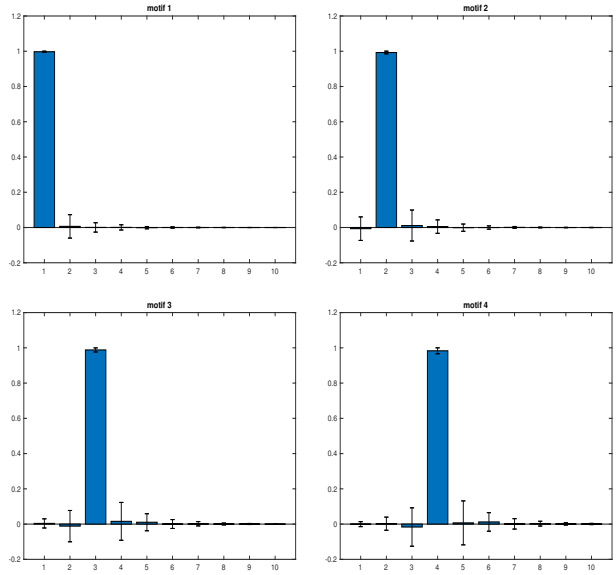

Figure 2: The first 10 elements of the four most dominant kernel motifs corresponding to $\mathbf{W} \in \mathbb{R}^{100\times100}$ generated element-wise i.i.d. from $N(0,1)$ and renormalised to largest singular value $\nu = 0.995$. The input coupling $\mathbf{w}$ was generated element-wise i.i.d. from $N(0,1)$ and renormalised to unit length. Shown are the means and standard deviations across 100 joint realisations of $\mathbf{W}$ and $\mathbf{w}$.

as predicted by our theory. Figure 3 presents the corresponding eigenvalues - solid bars correspond to the means of the actual eigenvalues λ_i across the 100 realisations (also shown are standard deviations). The theoretically predicted values (eq. (35)) are shown as the red line. Again, there is a strong agreement between the theoretical approximations $\widehat{\lambda}_i$ and the real eigenvalues λ_i. We illustrate generality of this result in appendix A, where motifs and their weights were obtained under the same conditions, but with the input coupling vector $\mathbf{w}$ generated as a vector of all 1s with randomly flipped signs (with equal probability 0.5 in each dimension). We also tried setting where both $\widetilde{\mathbf{W}}$ and $\mathbf{w}$ consist of all 1s with signs flipped independently element-wise with probability 0.5. In both cases the Markovian motifs and their weights are almost indistinguishable from those shown in figures 2 and 3.

It is notable that even though the state space dimensionality is quite high ($N = 100$), the rapidly decaying motif weights basically prevent the kernel to be able to dig deeper into the history of the time series $\mathbf{u}$, $\mathbf{v}$ when creating a quantitative evaluation of their similarity, $K(\mathbf{u},\mathbf{v})$. If the time series are zero-mean, the kernel is estimating a weighted covariance of $\mathbf{u}$ and $\mathbf{v}$ with weights $\left(\frac{\nu}{2}\right)^{2(i-1)}$ exponentially decreasing at the rate much faster than the upper bound $(\nu)^{2(i-1)}$, given the contractive dynamics of (1) with spectral radius ν.

We conclude this section by noting that for large random $\mathbf{W}$, the spectral radius $\rho \approx \nu/2$. Hence, the resulting temporal kernel can be readily interpreted from the standpoint of spectral radius: The Markovian motifs $\mathbf{e}_i$ have weights $\|\mathbf{w}\|\ \rho^{i-1}$, leading to temporal

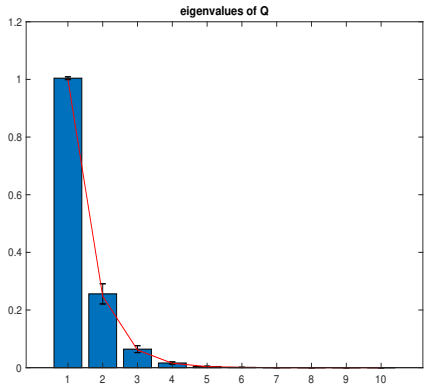

Figure 3: Eigenvalues (squared motif weights) of the metric tensor $\mathbf{Q}$ for random setting of the dynamical system (1) as described in fig. 2. Solid bars correspond to the means of the actual eigenvalues λ_i across the 100 realisations of $\mathbf{W}$ and $\mathbf{w}$ (also shown are standard deviations). The theoretically predicted values (eq. (35)) are shown as the red line.

kernel

$$K(\mathbf{u},\mathbf{v}) \approx \|\mathbf{w}\|^2 \sum_{i=1}^{N} \rho^{2(i-1)} u_i \; v_i.$$

Zhang et al. (2012) studied echo state networks with i.i.d. random weights in $\mathbf{W}$. They showed that the dynamic mapping (1) can be contractive with high probability even when only the spectral radius ρ (as opposed to maximum singular value ν) is less than one.

6. Symmetric dynamic coupling W

In this section we investigate how the nature of the temporal kernel changes if we impose symmetry on the dynamic coupling $\mathbf{W}$ of system (1): $W_{i,j} = W_{j,i}$, $i,j = 1,2,...,N$. In this case, the largest singular value ν of $\mathbf{W}$ is equal to its spectral radius. Memory capacity of such systems was rigorously analysed in (Tino and Rodan, 2013; Tino, 2018). In terms of memory capacity, the role of self-couplings in large systems was shown to be negligible. In (Couillet et al., 2016) systems with symmetric coupling were shown to lead to inferior performance on memory tasks, when compared with unconstrained dynamic coupling. Similar observation was made in the context of forecasting realised variances of stock market indices (Ficura, 2017).

Recall that given N_k kernels $K^{(a)}(\cdot,\cdot)$ operating on a space $\mathcal{X}$ and positive real numbers $\alpha_a > 0$, $a = 1,2,...,N_k$, the linear combination $K(\cdot,\cdot) = \sum_{a=1}^{N_k} \alpha_a K^{(a)}(\cdot,\cdot)$ is a valid kernel on $\mathcal{X}$. We will show that in case of symmetric $\mathbf{W}$, the corresponding temporal kernel can be understood as a linear combination of simple kernels, each with a unique exponentially decaying motif.

Theorem 3 *Consider the dynamical system (1) of state dimensionality N with symmetric coupling $\mathbf{W}$ of rank $N_k \leq N$. Let $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{N_k}$, be the eigenvectors of $\mathbf{W}$ corresponding to non-zero eigenvalues $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{N_k}$. Denote by $\widetilde{w}_a = \mathbf{s}_a^\top \mathbf{w}$ the projection of the input coupling $\mathbf{w}$ onto the eigenvector $\mathbf{s}_a$. Then the temporal kernel $K(\cdot,\cdot)$ associated with system (1) is a linear combination of N_k kernels $K^{(a)}(\cdot,\cdot)$,*

$$K(\cdot,\cdot) = \sum_{a=1}^{N_k} \widetilde{w}_a^2 \; K^{(a)}(\cdot,\cdot), \tag{36}$$

each kernel $K^{(a)}$ with a single motif

$$\mathbf{m}^{(a)} = (1, \sigma_a, \sigma_a^2, ..., \sigma_a^{\tau-1})^\top \in \mathbb{R}^\tau. \tag{37}$$

Proof Since $\mathbf{W}$ is symmetric, it can be decomposed as

$$\mathbf{W} = \mathbf{S} \; \Sigma \; \mathbf{S}^\top,$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{N_k}]$ is an $N \times N_k$ matrix storing the eigenvectors of $\mathbf{W}$ as columns, with the associated eigenvalues organised along the diagonal of $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_{N_k})$. The powers of $\mathbf{W}$ can be then expressed simply through powers of Σ: $\mathbf{W}^i = \mathbf{S} \; \Sigma^i \; \mathbf{S}^\top$. We thus have

$$\begin{aligned} Q_{i,j} &= \mathbf{w}^\top \; (\mathbf{W}^\top)^{i-1} \; \mathbf{W}^{j-1} \; \mathbf{w} \\ &= \mathbf{w}^\top \; \mathbf{W}^{i+j-2} \; \mathbf{w} \qquad \text{(by symmetry of } \mathbf{W}) \\ &= \mathbf{w}^\top \mathbf{S} \; \Sigma^{i+j-2} \; \mathbf{S}^\top \mathbf{w} \\ &= \widetilde{\mathbf{w}}^\top \; \Sigma^{i+j-2} \; \widetilde{\mathbf{w}}, \end{aligned} \tag{38}$$

where $\widetilde{\mathbf{w}} = \mathbf{S}^\top \mathbf{w}$ is the projection of input coupling $\mathbf{w}$ onto the orthonormal eigen-basis of $\mathbf{W}$.

Writing (38) as a quadratic form, we obtain

$$\begin{aligned} Q_{i,j} &= \sum_{a,l=1}^{N_k} \widetilde{w}_a \; \widetilde{w}_l \; \Sigma_{a,l}^{i+j-2} \\ &= \sum_{a=1}^{N_k} \widetilde{w}_a^2 \; \sigma_a^{i+j-2}, \end{aligned} \tag{39}$$

because Σ is a diagonal matrix.

Let us define N_k matrices $\mathbf{Q}^{(a)} \in \mathbb{R}^{\tau \times \tau}$, $a = 1, 2, ..., N_k$, as

$$Q_{i,j}^{(a)} = \sigma_a^{i+j-2}.$$

Then,

$$\mathbf{Q} = \sum_{a=1}^{N_k} \widetilde{w}_a^2 \; \mathbf{Q}^{(a)}. \tag{40}$$

Note that $\mathbf{Q}^{(a)}$ are rank-1 positive semi-definite matrices $\mathbf{Q}^{(a)} = \mathbf{m}^{(a)} \ (\mathbf{m}^{(a)})^\top$. Since

$$\mathbf{Q}^{(a)} \ \mathbf{m}^{(a)} = \mathbf{m}^{(a)} \ (\mathbf{m}^{(a)})^\top \ \mathbf{m}^{(a)} = \|\mathbf{m}^{(a)}\|_2^2 \ \mathbf{m}^{(a)},$$

we have that $\mathbf{m}^{(a)}$ is the only eigenvector of $\mathbf{Q}^{(a)}$ with a non-zero eigenvalue, i.e. $\mathbf{m}^{(a)}$ is the only motif of the kernel

$$K^{(a)}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \ \mathbf{Q}^{(a)} \ \mathbf{v}$$

with non-zero motif weight. From (40) it follows that $K(\mathbf{u}, \mathbf{v}) = \sum_{a=1}^{N_k} \widetilde{w}_a^2 \ K^{(a)}(\mathbf{u}, \mathbf{v})$. ■

Theorem 3 states that the temporal kernel of a system (1) with symmetric coupling is a linear combination of several kernels, each of which corresponds to a single non-zero eigenvalue σ_a of $\mathbf{W}$. Each such kernel has a unique motif $\mathbf{m}^{(a)} \in \mathbb{R}^\tau$ associated with it. The motifs $\mathbf{m}^{(a)}$ can only be of two kinds: Either an exponentially decaying profile $(1, \sigma_a, \sigma_a^2, \sigma_a^3, \sigma_a^4, ...)$, if σ_a is positive, or an exponentially decaying profile with high oscillation frequency $(1, -|\sigma_a|, \sigma_a^2, -|\sigma_a^3|, \sigma_a^4, ...)$, if σ_a is negative. This is obviously quite limiting, precluding the component kernels $K^{(a)}(\cdot, \cdot)$ to capture more diverse range of possible dynamic behaviours.

A word of caution is in order. The individual motifs $\mathbf{m}^{(a)}$ are indeed motifs of the component kernels $K^{(a)}(\cdot, \cdot)$, but they are not motifs of the kernel $K(\cdot, \cdot)$. Even though one can write

$$\mathbf{Q} = \mathbf{V} \ \Sigma_W \ \mathbf{V}^\top,$$

where the matrix $\mathbf{V} = [\mathbf{m}^{(1)}, \mathbf{m}^{(2)}, ..., \mathbf{m}^{(N_k)}]$ stores component motifs $\mathbf{m}^{(a)}$ as columns and $\Sigma_W = \text{diag}(\widetilde{w}_1^2, \widetilde{w}_a^2, ..., \widetilde{w}_{N_k}^2)$, the component motifs $\mathbf{m}^{(a)}$ are not orthogonal. Hence, in general there is no non-zero number κ, such that

$$\mathbf{Q} \ \mathbf{m}^{(a)} = \mathbf{V} \ \Sigma_W \ \mathbf{V}^\top \ \mathbf{m}^{(a)} = \kappa \ \mathbf{m}^{(a)}.$$

Unlike in the previous section, because of the imposed symmetry on $\mathbf{W}$, it is much more difficult to approximate the structure of $\mathbf{Q}$. We can recover the upper bound (18) of theorem 2 on absolute values of $Q_{i,j}$. From Theorem 2 and eq. (39) we have

$$Q_{i,j} = \sum_{a=1}^{N} \widetilde{w}_a^2 \ \sigma_a^{i+j-2}$$

and

$$\begin{aligned} |Q_{i,j}| &\leq \sum_{a=1}^{N} \widetilde{w}_a^2 \ |\sigma_a^{i+j-2}| \\ &\leq \nu_a^{i+j-2} \sum_{a=1}^{N} \widetilde{w}_a^2 \qquad (41) \\ &\leq \nu^{i+j-2} \ \|\mathbf{w}\|_2^2. \qquad (42) \end{aligned}$$

Here (42) follows from (41) since the norm of the input coupling $\mathbf{w}$ is invariant with respect to orthonormal change of basis. The inequality in (42) becomes equality if $\mathbf{W}$ is full rank.

7. W as a scaled permutation matrix

We will now consider a strongly constrained dynamical coupling $\mathbf{W}$ in the form of cyclic $N \times N$ permutation matrix $\mathbf{P}$, scaled by ν, so that the largest singular value, as well as the spectral radius of $\mathbf{W} = \nu \cdot \mathbf{P}$ is equal to ν. This follows from a theorem by Frobenius that states that for a non-negative matrix $\mathbf{W}$, its spectral radius is lower and upper bounded by the minimum and maximum row sum, respectively (e.g. (Minc, 1988)). Since in our case all rows of $\mathbf{W}$ sum to ν, the spectral radius must be[5] ν.

Without loss of generality[6] we will consider cyclic permutation $\{1 \to 2, 2 \to 3, ..., N-1 \to N, N \to 1\}$, represented by $P_{i+1,i} = 1$, $i = 1, 2, ..., N-1$ and $P_{1,N} = 1$, all the other elements of $\mathbf{P}$ are zero. Dynamic couplings in the form of scaled cyclic permutation matrix correspond to the setting of simple cycle reservoir (Rodan and Tino, 2011), where the reservoir units are connected in a uni-directional ring structure, with the same weight value on all connections in the ring. Analogously, setting of the input coupling $\mathbf{w}$ can be very simple, controlled again by a single amplitude value $w > 0$ for all input weights. Intuitively, all the input weights should not have the same value w, as this would greatly symmetrise the ESN architecture. To break the symmetry, Rodan and Tino (2011) suggest to apply an a-periodic sign pattern to the input weights (e.g. according to binary expansion of an irrational number). While such a reservoir structure has the advantage of being extremely simple and completely deterministic, the predictive performance of the associated ESNs in a variety of tasks on time series of different origins and memory structure was shown to be on par (and sometimes even better) with the usual random reservoir constructions (Rodan and Tino, 2011). Similar observations were made in (Strauss et al., 2012). This is of great practical importance, since many optics-based physical constructions of reservoir models follow the ring topology structure, which can be implemented using a single unit with multiple delays (Röhm and Lüdge, 2018; Tanaka et al., 2019; Appeltant et al., 2011). Yet, it has been unclear, why such a simple setting can be competitive in real word tasks, or why indeed the breaking of symmetry through a-periodic sign pattern in the input weights is so crucial. In this section, we will study the nature of motifs associated with ring reservoir topologies and the consequences of adopting periodic, rather than a-periodic input weight sign patterns.

Given a time horizon $\tau = \ell N$, for some positive integer $\ell > 1$, we will now show that the temporal kernel motifs corresponding to the dynamical system (1) with scaled permutation coupling $\mathbf{W} = \nu \cdot \mathbf{P}$ have an intricate block structure.

Theorem 4 *Consider the dynamical system (1) of state space dimensionality N, with coupling $\mathbf{W} = \nu \cdot \mathbf{P}$, where $\nu \in (0, 1)$ and $\mathbf{P}$ is the $N \times N$ cyclic permutation matrix. Let $\widetilde{\mathbf{m}}_i \in \mathbb{R}^N$, $i = 1, 2, ..., N$, be motifs of the temporal kernel associated with (1) under past time horizon equal to N. Denote the corresponding motif weights by $\widetilde{\omega}_i$. Then, given a different past time horizon $\tau = \ell \cdot N$, for some positive integer $\ell > 1$, the temporal kernel motifs $\mathbf{m}_i \in \mathbb{R}^\tau$ associated with (1) have the following block form:*

$$\mathbf{m}_i = \left(\widetilde{\mathbf{m}}_i^\top, \nu^N \widetilde{\mathbf{m}}_i^\top, \nu^{2N} \widetilde{\mathbf{m}}_i^\top, ..., \nu^{(\ell-1)N} \widetilde{\mathbf{m}}_i^\top\right)^\top, \quad i = 1, 2, ...N.$$

5. Alternatively, this can be shown by arguing that $\mathbf{W}$ is a normal matrix.
6. We can always renumber the state space dimensions.

The corresponding motif weights are equal to

$$\omega_i = \widetilde{\omega}_i \ \left(\frac{1-\nu^{2\tau}}{1-\nu^{2N}}\right)^{\frac{1}{2}}.$$

Proof Note that because $\mathbf{P}$ is a permutation matrix, for any non-negative integer $i \in \mathbb{N}_0$, we have

$$\mathbf{P}^i = \mathbf{P}^{N\cdot(i\backslash N)} \ \mathbf{P}^{i \bmod N},$$

where $\bmod$ and $\backslash$ denote the modulo and integer division operations. Since $\mathbf{P}^{N\cdot(i\backslash N)} = \mathbf{I}_{N\times N}$, we have $\mathbf{P}^i = \mathbf{P}^{i \bmod N}$. Consequently,

$$\mathbf{W}^i = \nu^i \cdot \mathbf{P}^{i \bmod N}.$$

Furthermore, since $\mathbf{P}$ is orthogonal, $\mathbf{P}^{-1} = \mathbf{P}^\top$. We can now write the elements of $\mathbf{Q}$ as (see eq. (17)),

$$\begin{aligned} Q_{i,j} &= \mathbf{w}^\top \ (\mathbf{W}^\top)^{i-1} \ \mathbf{W}^{j-1} \ \mathbf{w} \\ &= \nu^{i+j-2} \ \mathbf{w}^\top \ (\mathbf{P}^\top)^{i-1} \ \mathbf{P}^{j-1} \ \mathbf{w} \\ &= \nu^{i+j-2} \ \mathbf{w}^\top \ \mathbf{P}^{j-i} \ \mathbf{w}. \\ &= \nu^{i+j-2} \ \mathbf{w}^\top \ \mathbf{P}^{(j-i) \bmod N} \ \mathbf{w}. \end{aligned} \tag{43}$$

For $k \in \{-N+1, -N+2, ..., -1, 0, 1, ...N-1\}$, if k is positive, $\mathbf{P}^k\mathbf{w}$ is the vector with elements of $\mathbf{w}$ rotated k places to the right. In case k is negative, the rotation is to the left. From (43) it is clear the $\mathbf{Q} \in \mathbb{R}^{\tau\times\tau}$ has the following block structure:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}^{(1,1)} & \mathbf{Q}^{(1,2)} & \dots & \mathbf{Q}^{(1,\ell)} \\ \mathbf{Q}^{(2,1)} & \mathbf{Q}^{(2,2)} & \dots & \mathbf{Q}^{(2,\ell)} \\ \dots & \dots & \dots & \dots \\ \mathbf{Q}^{(\ell,1)} & \mathbf{Q}^{(\ell,2)} & \dots & \mathbf{Q}^{(\ell,\ell)} \end{bmatrix},$$

where each matrix $\mathbf{Q}^{(a,b)} \in \mathbb{R}^{N\times N}$, $a, b = 1, 2, .., , \ell$, has elements

$$Q^{(a,b)}_{i,j} = \nu^{(a+b-2)N} \ \nu^{i+j-2} \ \mathbf{w}^\top \ \mathbf{P}^{j-i} \ \mathbf{w}, \quad i, j = 1, 2, ..., N.$$

Define an $N \times N$ matrix $\mathbf{R}$ with elements

$$R_{i,j} = \nu^{i+j-2} \ \mathbf{w}^\top \ \mathbf{P}^{j-i} \ \mathbf{w}, \quad i, j = 1, 2, ..., N, \tag{44}$$

yielding $\mathbf{Q}^{(a,b)} = \nu^{(a+b-2)N} \ \mathbf{R}$. Note that $\mathbf{R}$ is the metric tensor of the temporal kernel associated with (1) under the past time horizon N. Let $\widetilde{\mathbf{m}}_i \in \mathbb{R}^N$ be the i-th eigenvector of $\mathbf{R}$ with eigenvalue $\widetilde{\lambda}_i$. Then,

$$\mathbf{Q}^{(a,b)} \ \widetilde{\mathbf{m}}_i = \nu^{(a+b-2)N} \ \widetilde{\lambda}_i \ \widetilde{\mathbf{m}}_i,$$

and so

$$\left[\mathbf{Q}^{(a,1)}, \mathbf{Q}^{(a,2)}, \cdots, \mathbf{Q}^{(a,\ell)}\right] \begin{bmatrix} \widetilde{\mathbf{m}_i} \\ \widetilde{\mathbf{m}}_i \\ \cdots \\ \widetilde{\mathbf{m}}_i \end{bmatrix} = \widetilde{\lambda}_i \left(\sum_{j=1}^{\ell} \nu^{(j-1)N}\right) \nu^{(a-1)N} \widetilde{\mathbf{m}_i}. \tag{45}$$

It follows that for each $a \in \{1, 2, ..., \ell\}$,

$$\left[\mathbf{Q}^{(a,1)}, \mathbf{Q}^{(a,2)}, \cdots, \mathbf{Q}^{(a,\ell)}\right] \begin{bmatrix} \widetilde{\mathbf{m}_i} \\ \nu^N \widetilde{\mathbf{m}}_i \\ \cdots \\ \nu^{(\ell-1)N} \widetilde{\mathbf{m}}_i \end{bmatrix} = \widetilde{\lambda}_i \left(\sum_{j=1}^{\ell} \nu^{2(j-1)N}\right) \nu^{(a-1)N} \widetilde{\mathbf{m}_i}. \tag{46}$$

We can thus conclude that the vector

$$\mathbf{m}_i = \left(\widetilde{\mathbf{m}}_i^\top, \nu^N \widetilde{\mathbf{m}}_i^\top, \nu^{2N} \widetilde{\mathbf{m}}_i^\top, ..., \nu^{(\ell-1)N} \widetilde{\mathbf{m}}_i^\top\right)^\top$$

is an eigenvector of $\mathbf{Q}$ with eigenvalue

$$\begin{aligned} \lambda_i &= \widetilde{\lambda}_i \sum_{j=0}^{\ell-1} \left(\nu^{2N}\right)^j \\ &= \widetilde{\lambda}_i \frac{1-\nu^{2N\ell}}{1-\nu^{2N}}. \\ &= \widetilde{\lambda}_i \frac{1-\nu^{2\tau}}{1-\nu^{2N}}. \end{aligned}$$

■

7.1. Periodic input coupling w

It has been empirically shown in (Rodan and Tino, 2011) that when the dynamic coupling $\mathbf{W}$ is formed by a scaled permutation matrix, a very simple setting of input coupling $\mathbf{w}$ is sufficient: all elements of $\mathbf{w}$ can have the same absolute value, but the sign pattern should be aperiodic. Intuitively, it is clear that for such $\mathbf{W}$ a periodic input coupling $\mathbf{w}$ will induce symmetry in the dynamic processing of (1) and such a symmetry should be broken. However, in this section we would like to ask exactly what representational capabilities are lost by imposing a periodicity in $\mathbf{w}$.

We will start by considering a general case of periodic $\mathbf{w} \in \mathbb{R}^N$ formed by $k > 1$ copies of a periodic block $\mathbf{s} \in \mathbb{R}^p$, $\mathbf{w} = (\mathbf{s}^\top, \mathbf{s}^\top, ..., \mathbf{s}^\top)^\top$. Obviously, $N = k \cdot p$.

Denote by $\overline{\mathbf{P}} \in \mathbb{R}^{p \times p}$ the top left $p \times p$ block of the right shift permutation matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$. In other words, $\overline{\mathbf{P}}$ is the right shift permutation matrix operating on vectors from $\mathbb{R}^p$. Furthermore, we introduce matrix $\mathbf{T} \in \mathbb{R}^{p \times p}$ with elements

$$T_{i,j} = \nu^{i+j-2} \left\langle \mathbf{s}, \overline{\mathbf{P}}^{|j-i|} \mathbf{s} \right\rangle, \quad i, j = 1, 2, ..., p. \tag{47}$$

Theorem 5 *Consider the dynamical system (1) of state space dimensionality N, with coupling $\mathbf{W} = \nu \cdot \mathbf{P}$, where $\nu \in (0,1)$ and $\mathbf{P}$ is the $N \times N$ cyclic permutation matrix. Let the input coupling $\mathbf{w} \in \mathbb{R}^N$ consist of $k > 1$ copies of a periodic block $\mathbf{s} \in \mathbb{R}^p$. Denote by $\overline{\mathbf{m}}_i \in \mathbb{R}^p$, $i = 1, 2, ..., p$, eigenvectors of the matrix $\mathbf{T}$ (47) with the corresponding eigenvalues $\overline{\lambda}_i$. Then, given a past time horizon $\tau = \ell \cdot N$, for some positive integer $\ell > 1$, there are at most p temporal kernel motifs $\mathbf{m}_i \in \mathbb{R}^\tau$ associated with (1) of non-zero motif weight. Furthermore, the kernel motifs have the following block form,*

$$\mathbf{m}_i = (\overline{\mathbf{m}}_i^\top, \nu^p\ \overline{\mathbf{m}}_i^\top, \nu^{2p}\ \overline{\mathbf{m}}_i^\top, ..., \nu^{\tau-p}\ \overline{\mathbf{m}}_i^\top)^\top, \quad i = 1, 2, ...p,$$

with the corresponding motif weights

$$\omega_i = \left(\overline{\lambda}_i\ \frac{1-\nu^{2\tau}}{1-\nu^{2p}} \right)^{\frac{1}{2}}.$$

Proof By Theorem 4, to determine motifs of the temporal kernel associated with (1), it is sufficient to perform eigen-analysis of the block matrix $\mathbf{Q}^{(1,1)} = \mathbf{R}$ (eq. (44)).

For $a = 0, 1, 2, ..., N-1$,

$$\langle \mathbf{w}, \mathbf{P}^a \mathbf{w} \rangle = k \cdot \langle \mathbf{s}, \overline{\mathbf{P}}^{\,a} \mathbf{s} \rangle = k \cdot \left\langle \mathbf{s}, \overline{\mathbf{P}}^{\,a \bmod p}\ \mathbf{s} \right\rangle$$

and since $\mathbf{R}$ is symmetric, from (eq. (44)) we have

$$Q_{i,j}^{(1,1)} = k \cdot \nu^{i+j-2} \cdot \left\langle \mathbf{s}, \overline{\mathbf{P}}^{\,|j-i| \bmod p}\ \mathbf{s} \right\rangle, \quad i, j = 1, 2, ..., N. \tag{48}$$

Therefore, $\mathbf{Q}^{(1,1)}$ can be decomposed into blocks of $p \times p$ matrices

$$\mathbf{Q}^{(1,1)} = \begin{bmatrix} \mathbf{C}^{(1,1)} & \mathbf{C}^{(1,2)} & \dots & \mathbf{C}^{(1,k)} \\ \mathbf{C}^{(2,1)} & \mathbf{C}^{(2,2)} & \dots & \mathbf{C}^{(2,k)} \\ \dots & \dots & \dots & \dots \\ \mathbf{C}^{k,1)} & \mathbf{C}^{(k,2)} & \dots & \mathbf{C}^{(k,k)} \end{bmatrix},$$

where

$$\mathbf{C}^{(c,d)} = \nu^{(c+d-2)p}\ \mathbf{C}^{(1,1)}, \quad c, d = 1, 2, ..., k$$

and

$$C_{i,j}^{(1,1)} = \nu^{i+j-2} \cdot \left\langle \mathbf{s}, \overline{\mathbf{P}}^{\,|j-i|}\ \mathbf{s} \right\rangle, \quad i, j = 1, 2, ..., p.$$

Now, let $\overline{\mathbf{m}}_i \in \mathbb{R}^p$ be the i-th eigenvector of $\mathbf{C}^{(1,1)} = \mathbf{T}$ with eigenvalue $\overline{\lambda}_i$. Then,

$$\mathbf{C}^{(c,d)}\ \overline{\mathbf{m}}_i = \nu^{(c+d-2)p}\ \mathbf{C}^{(1,1)}\ \overline{\mathbf{m}}_i = \nu^{(c+d-2)p}\ \overline{\lambda}_i\ \overline{\mathbf{m}}_i.$$

We have

$$\left[\mathbf{C}^{(c,1)}, \mathbf{C}^{(c,2)}, \cdots, \mathbf{C}^{(c,k)} \right] \begin{bmatrix} \overline{\mathbf{m}}_i \\ \nu^p\ \overline{\mathbf{m}}_i \\ \cdots \\ \nu^{(k-1)p}\ \overline{\mathbf{m}}_i \end{bmatrix} = \overline{\lambda}_i\ \left(\sum_{j=1}^{k} \nu^{2(j-1)p} \right) \nu^{(c-1)p}\ \overline{\mathbf{m}}_i \tag{49}$$

for $c = 1, 2, ..., k$. Hence,

$$\widetilde{\mathbf{m}}_i = (\overline{\mathbf{m}}_i^\top, \nu^p\, \overline{\mathbf{m}}_i^\top, \nu^{2p}\, \overline{\mathbf{m}}_i^\top, ..., \nu^{(k-1)p}\, \overline{\mathbf{m}}_i^\top)^\top$$

is an eigenvector of $\mathbf{Q}^{(1,1)}$ with eigenvalue

$$\begin{aligned}
\widetilde{\lambda}_i &= \overline{\lambda}_i \sum_{j=0}^{k-1} \left(\nu^{2p}\right)^j \\
&= \overline{\lambda}_i \frac{1-\nu^{2pk}}{1-\nu^{2p}}. \\
&= \overline{\lambda}_i \frac{1-\nu^{2N}}{1-\nu^{2p}}.
\end{aligned}$$

By Theorem 4, the corresponding eigenvector $\mathbf{m}_i$ of $\mathbf{Q}$ reads:

$$\begin{aligned}
\mathbf{m}_i &= (\widetilde{\mathbf{m}}_i^\top, \nu^N \widetilde{\mathbf{m}}_i^\top, ..., \nu^{(\ell-1)N} \widetilde{\mathbf{m}}_i^\top)^\top \\
&= (\overline{\mathbf{m}}_i^\top, \nu^p\, \overline{\mathbf{m}}_i^\top, ..., \nu^{(k-1)p}\, \overline{\mathbf{m}}_i^\top, \nu^N\, \overline{\mathbf{m}}_i^\top, \nu^{N+p}\, \overline{\mathbf{m}}_i^\top, ..., \nu^{(\ell-1)N+(k-1)p}\, \overline{\mathbf{m}}_i^\top)^\top \\
&= (\overline{\mathbf{m}}_i^\top, \nu^p\, \overline{\mathbf{m}}_i^\top, \nu^{2p}\, ..., \nu^{\tau-p}\, \overline{\mathbf{m}}_i^\top)^\top.
\end{aligned}$$

The last equality holds since from $\tau = \ell N$ and $N = kp$, we have $(\ell-1)N + (k-1)p = \tau - p$. We can calculate the corresponding eigenvalue as

$$\begin{aligned}
\lambda_i &= \widetilde{\lambda}_i \frac{1-\nu^{2\tau}}{1-\nu^{2N}} \\
&= \overline{\lambda}_i \frac{1-\nu^{2N}}{1-\nu^{2p}} \frac{1-\nu^{2\tau}}{1-\nu^{2N}} = \overline{\lambda}_i \frac{1-\nu^{2\tau}}{1-\nu^{2p}}.
\end{aligned}$$

■

Theorem 5 formally specifies consequences for the dynamical kernel of having a periodic input coupling $\mathbf{w}$ of period p in the system (1). First, the number of potentially useful kernel motifs of non-zero weight is reduced from N (the state space dimensionality) to p. Second, the motif structure is even more restricted than in the case of general $\mathbf{w}$. If the past horizon is $\tau = \ell N$, then in general, by theorem 4, each motif $\mathbf{m}_i \in \mathbb{R}^\tau$ consists of a series of ℓ copies of the same "core motif" $\widetilde{\mathbf{m}}_i \in \mathbb{R}^N$, down-weighted by exponential decay. In the case of periodic $\mathbf{w}$, motifs $\mathbf{m}_i \in \mathbb{R}^\tau$ are formed by a series of ℓk copies of the same small block $\overline{\mathbf{m}}_i \in \mathbb{R}^p$, down-weighted by exponential decay.

We will now investigate special settings of the periodic input coupling $\mathbf{w} \in \mathbb{R}^N$. Consider first the binary setting, i.e. the core periodic block is $\mathbf{s} = (1, 0, 0, ..., 0)^\top \in \{0, 1\}^p$. Assume $\mathbf{w}$ contains k such blocks ($N = k \cdot p$). Then, since for $a = 0, 1, 2, ..., p-1$,

$$\left\langle \mathbf{s}, \overline{\mathbf{P}}^{\,a}\, \mathbf{s} \right\rangle = \begin{cases} 1, & \text{if } \ a = 0 \\ 0, & \text{otherwise,} \end{cases}$$

the matrix $\mathbf{T} \in \mathbb{R}^{p \times p}$ (eq. (47)) will have a diagonal form, $\mathbf{T} = \mathrm{diag}(1, \nu^2, ..., \nu^{2(p-1)})$. The eigenvectors $\overline{\mathbf{m}}_i \in \mathbb{R}^p$ of $\mathbf{T}$, $i = 1, 2, ..., p$, correspond to the standard basis $\overline{\mathbf{e}}_i$ of $\mathbb{R}^p$, i.e.

all elements of $\overline{\mathbf{e}}_i$ are zeros, except for the i-th element, which is 1. The corresponding eigenvalues are $\overline{\lambda}_i = \nu^{2(i-1)}$. By theorem 5, each motif

$$\mathbf{m}_i = (\overline{\mathbf{e}}_i^\top, \nu^p\ \overline{\mathbf{e}}_i^\top, \nu^{2p}\ \overline{\mathbf{e}}_i^\top, ..., \nu^{\tau-p}\ \overline{\mathbf{e}}_i^\top)^\top, \tag{50}$$

with motif weight

$$\omega_i = \nu^{i-1} \left(\frac{1-\nu^{2\tau}}{1-\nu^{2p}} \right)^{\frac{1}{2}}. \tag{51}$$

is a periodic exponentially decaying motif that picks up elements of time series driving (1) with periodicity p and initial lag i. Given a time series $\mathbf{u} \in \mathbb{R}^\tau$,

$$\langle \mathbf{m}_i, \mathbf{u} \rangle = \sum_{j=1}^{\ell \cdot k} \nu^{(j-1)p}\ u_{i+(j-1)p}.$$

In the representation of (eq. (21)) we then have

$$\widetilde{\mathbf{u}} = \left(\frac{1-\nu^{2\tau}}{1-\nu^{2p}} \right)^{\frac{1}{2}} \cdot (\langle \mathbf{m}_1, \mathbf{u} \rangle, \nu\ \langle \mathbf{m}_2, \mathbf{u} \rangle, ..., \nu^{p-1}\ \langle \mathbf{m}_p, \mathbf{u} \rangle)^\top \in \mathbb{R}^p.$$

Given another time series $\mathbf{v} \in \mathbb{R}^\tau$, the temporal kernel gives

$$\begin{aligned} K(\mathbf{u}, \mathbf{v}) &= \langle \widetilde{\mathbf{u}}, \widetilde{\mathbf{v}} \rangle \\ &= \frac{1-\nu^{2\tau}}{1-\nu^{2p}} \sum_{i=1}^{p} \nu^{2(i-1)}\ \langle \mathbf{m}_i, \mathbf{u} \rangle\ \langle \mathbf{m}_i, \mathbf{v} \rangle. \end{aligned} \tag{52}$$

In the case of all-ones $\mathbf{w}$ with a periodic sign pattern, the core periodic block is $\mathbf{s} = (+1, -1, -1, ..., -1)^\top \in \{-1, +1\}^p$. For $a = 0, 1, 2, ..., p-1$, we have

$$\langle \mathbf{s}, \overline{\mathbf{P}}^{\,a}\ \mathbf{s} \rangle = \begin{cases} p, & \text{if } a = 0 \\ p-4, & \text{otherwise,} \end{cases}$$

From (eq. (47)), the matrix $\mathbf{T} \in \mathbb{R}^{p \times p}$ with elements

$$T_{i,j} = \begin{cases} \nu^{2(i-1)}\ p, & \text{if } i = j \\ \nu^{i+j-2}\ (p-4), & \text{otherwise,} \end{cases}$$

can yield a richer set of eigenvectors $\overline{\mathbf{m}}_i$ than the standard basis $\overline{\mathbf{e}}_i$ in $\mathbb{R}^p$. An exception is the case of period-4 sign pattern, $p = 4$. In that case, $\mathbf{T}$ is a diagonal matrix $\mathbf{T} = p \cdot \mathrm{diag}(1, \nu^2, ..., \nu^{2(p-1)})$, exactly the scaled version of $\mathbf{T}$ analysed above, when $\mathbf{w}$ was the binary vector composed of a series of k blocks of $\overline{\mathbf{e}}_1 \in \mathbb{R}^p$. Hence the four motifs $\mathbf{m}_i$ will have the form suggested by eq. (50) and the motif weights (51) will be scaled by $\sqrt{p} = 2$. We have thus established:

Corollary 6 *Under the assumptions of Theorem 5, assume that the input coupling* $\mathbf{w} \in \{0,1\}^N$ *consists of* $k > 1$ *copies of the binary standard basis block* $\mathbf{s} = \overline{\mathbf{e}}_1 \in \{0,1\}^p$. *Then there are* p *non-zero wight motifs of the dynamic kernel associated with (1),*

$$\mathbf{m}_i = (\overline{\mathbf{e}}_i^\top, \nu^p\ \overline{\mathbf{e}}_i^\top, \nu^{2p}\ \overline{\mathbf{e}}_i^\top, ..., \nu^{\tau-p}\ \overline{\mathbf{e}}_i^\top)^\top,$$

with motif weights

$$\omega_i = \nu^{i-1} \left(\frac{1-\nu^{2\tau}}{1-\nu^{2p}} \right)^{\frac{1}{2}}.$$

Each $\mathbf{m}_i$ *is thus a periodic exponentially decaying motif that picks up elements of input time series with periodicity* p *and initial lag* i.

Furthermore, if the bipolar input coupling $\mathbf{w} \in \{-1,+1\}^N$ *consists of* $k > 1$ *copies of the block* $\mathbf{s} = 2\,\overline{\mathbf{e}}_1 - \mathbf{1} \in \{-1,+1\}^4$ *of period* $p = 4$, *then there are four non-zero wight motifs* $\mathbf{m}_i$ *(50) with motif weights* $2\omega_i$.

8. Illustrative examples

In this section we will illustrate the results obtained so far showing the influence of the dynamic and input coupling, $\mathbf{W}$ and $\mathbf{w}$, respectively, on the strength and richness of motifs of the temporal kernel associated with the dynamical system (1). In all illustrations we will use state space dimensionality $N = 100$ and re-normalise the dynamic coupling $\mathbf{W} \in \mathbb{R}^{100\times100}$ to largest singular value $\nu = 0.995$. The input coupling $\mathbf{w}$ is renormalized to unit length. The past horizon will be set to $\tau = 200$. We will show motifs with motif weights up to 10^{-2} of the highest motif weight.

Figure 4 (left) shows motifs of the temporal kernel given by random coupling $\mathbf{W}$, where all elements $W_{i,j}$ were generated i.i.d. from normal distribution $N(0,1)$. The motifs are shown in a column-wise fashion, i.e. the x-axis indexes the individual motifs, while the motif values are shown along the y-axis. The associated motif weights are presented in the right plot.

As explained in section 5, each of the Markovian motifs picks an element from the recent time series history, yielding a shallow memory involved in the kernel evaluation, with rapidly decaying motif weights. Almost identical results were obtained for $W_{i,j}$ and w_i generated i.i.d. from other distributions (e.g. uniform over $[-1,+1]$, Bernoulli over $\{-1,+1\}$ or $\{0,1\}$), as well as for many other settings of $\mathbf{w}$, including the all-ones vector $\mathbf{w} = \mathbf{1}$.

Introduction of a structure into random $\mathbf{W}$ by imposing symmetry (Wigner $\mathbf{W}$) leads to a slightly richer motif set, albeit still with shallow memory (see figure 5). Note the high frequency nature of some motifs, as discussed in section 6. Again, the number and nature of the motifs stayed unchanged across a variety of generative mechanisms for $\mathbf{W}$ and $\mathbf{w}$ described above.

The situation changes dramatically when $\mathbf{W}$ is set to the scaled permutation matrix of section 7. Figure 6 shows an example of motif and motif weight structure for $\mathbf{w}$ generated randomly i.i.d. from $N(0,1)$. To demonstrate that what really matters, as argued in section 7, is the aperiodicity of $\mathbf{w}$, we show in figures 7 and 8 motifs and motif weights when $\mathbf{w}$ is simply a vector of ones with signs prescribed by the first $N = 100$ digits of binary

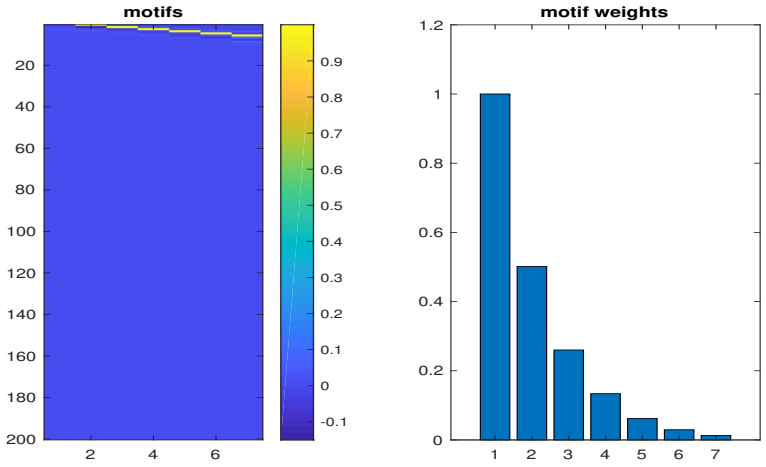

Figure 4: Temporal kernel motifs and the corresponding motif weights for randomly generated $\mathbf{W}$ and $\mathbf{w}$.

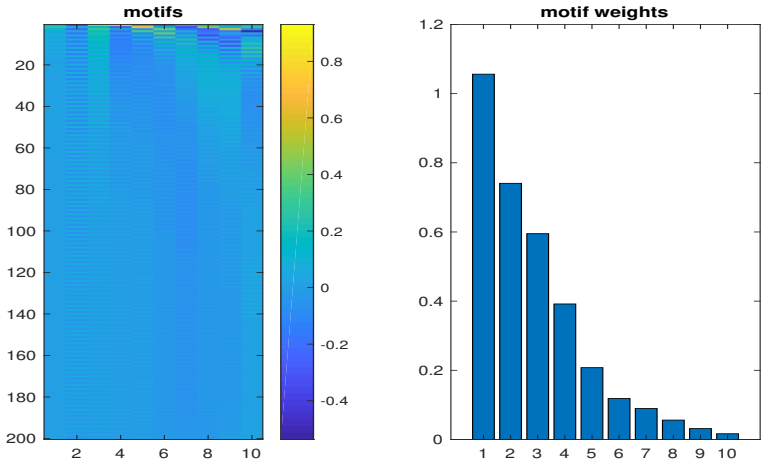

Figure 5: Temporal kernel motifs and the corresponding motif weights for random symmetric Wigner $\mathbf{W}$ and random $\mathbf{w}$.

expansion of π and e, respectively. This was suggested in (Rodan and Tino, 2011) as a simple controlled way of generating aperiodic input couplings. Such settings admit a full set of $N = 100$ highly variable motifs. The scaled block structure of motifs proved in section 7 is clearly visible. In striking contrast, as suggested in section 7, we present in figure 9 motifs and motif weights for the case of periodic $\mathbf{w}$ with period $p = 10$. As predicted by the theory, the shrunk motif set contains $p = 10$ simple periodic motifs given by repeated blocks of permuted standard basis $\overline{\mathbf{e}}_i$ (with possibly flipped sign).

As an example, in figure 10 we show in greater detail six temporal kernel motifs from figure 7, all with high motif weights. Compared with the setting of random or symmetric $\mathbf{W}$, besides the sheer number of motifs with higher weight, there is a much richer motif

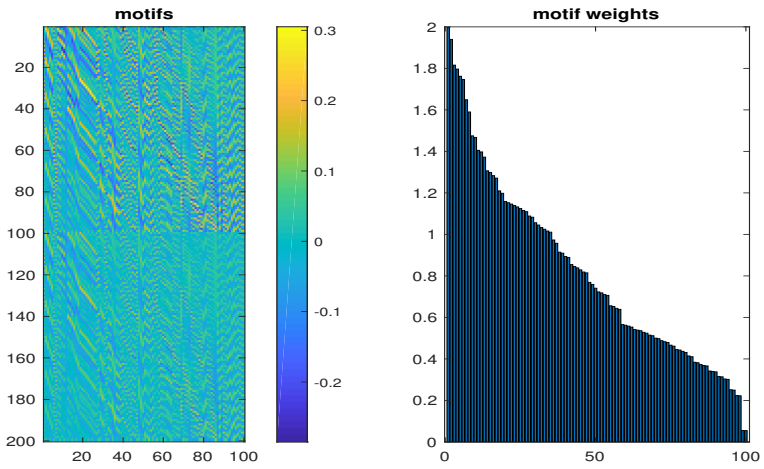

Figure 6: Temporal kernel motifs and the corresponding motif weights for scaled permutation matrix $\mathbf{W}$ and random $\mathbf{w}$.

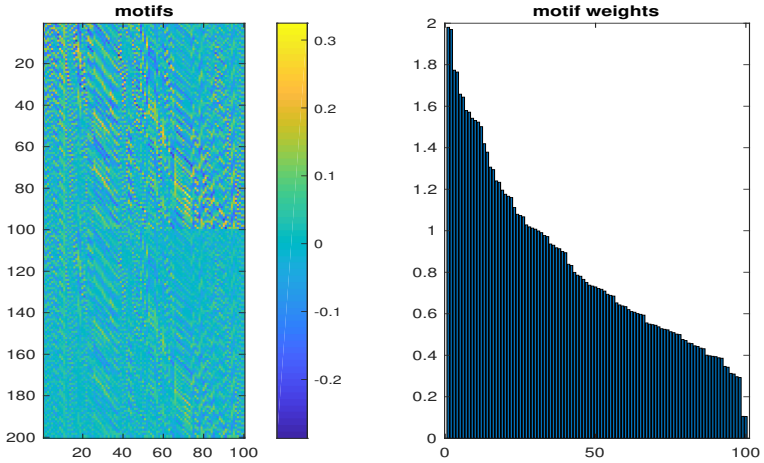

Figure 7: Temporal kernel motifs and the corresponding motif weights for scaled permutation matrix $\mathbf{W}$ and aperiodic all-ones vector $\mathbf{w}$ with signs following binary expansion of π.

variety and longer memory. Note that in accordance with theorem 4, since the state space dimensionality and past horizon are set to $N = 100$ and $\tau = 200$, respectively, the second half of each motif $\mathbf{m}_i$ is the scaled version of the first half, $\mathbf{m}_{i,101:200} = \nu^{100} \cdot \widetilde{\mathbf{m}}_i$, $\widetilde{\mathbf{m}}_i = \mathbf{m}_{i,1:100}$. In order to quantify "motif richness", we perform Fast Fourier Transform (FFT) of motifs $\mathbf{m}_i$ with motif weights ω_i up to 10^{-2} of the highest motif weight. We collect the Fourier coefficients $z_{i,k} \in \mathbb{C}$ of each motif $\mathbf{m}_i$ along with the corresponding motif weight $q_{i,k} = \omega_i$ in a set $F_i = \{(z_{i,k}, q_{i,k})\}_k$. The total coefficient set is then $F = \bigcup_i F_i$. Figure 11 presents distribution of motif Fourier coefficients from F for different settings of spectral radius $\nu \in \{0.96, 0.99, 0.996, 1.0\}$.

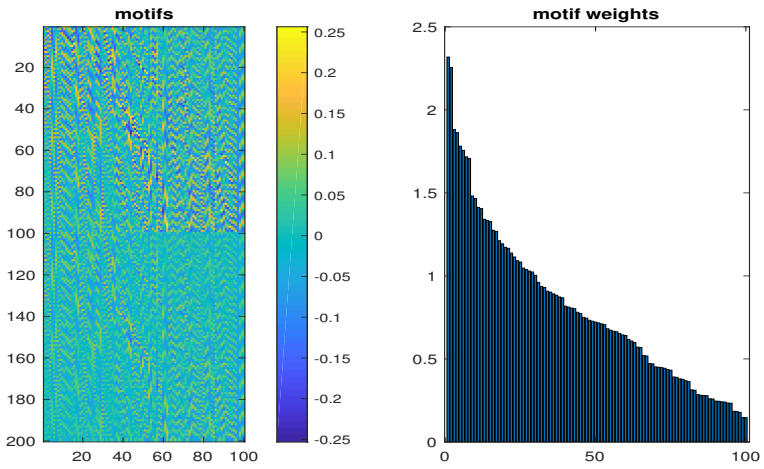

Figure 8: Temporal kernel motifs and the corresponding motif weights for scaled permutation matrix $\mathbf{W}$ and aperiodic all-ones vector $\mathbf{w}$ with signs following binary expansion of e.

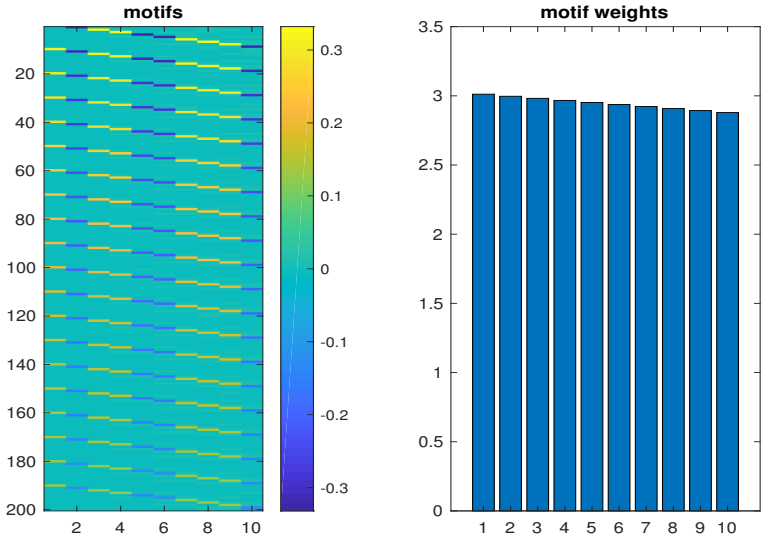

Figure 9: Temporal kernel motifs and the corresponding motif weights for scaled permutation matrix $\mathbf{W}$ and periodic binary vector $\mathbf{w}$ with period $p = 10$.

We designed two measures to characterise distribution of Fourier coefficients from F. The first is simply the area occupied by the coefficients $z_{i,k}$. In particular, the coefficient space $[-7, 7]^2$ in the complex plane was covered by regular grid of cells with side length 0.05. The *relative area* covered by F is then the ratio of the number of cells visited by the coefficients $z_{i,k}$ to the total number of cells. Figure 12 shows the relative area occupied by motif Fourier coefficients, as a function of the scaling largest singular value ν. It is

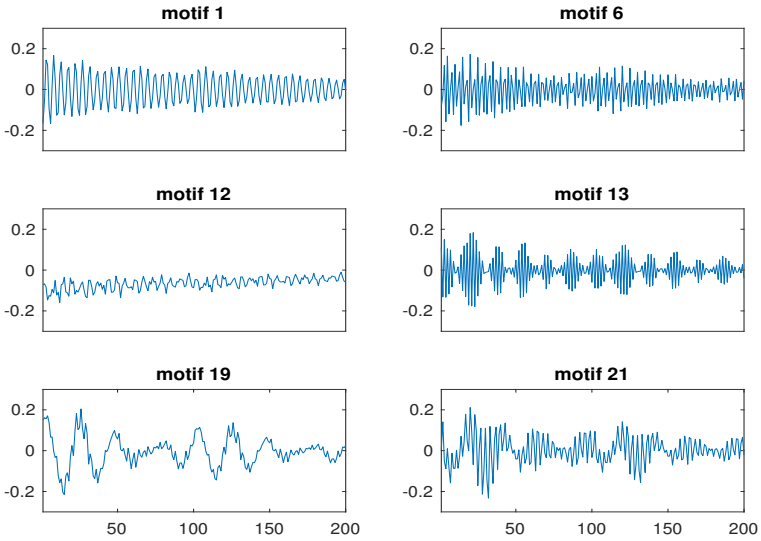

Figure 10: Selection of motifs ($\tau = 200$) of temporal kernel associated with dynamical system (1) of state space dimensionality $N = 100$. Dynamic coupling $\mathbf{W}$ was the scaled permutation matrix with spectral radius $\nu = 0.995$ and the input coupling $\mathbf{w}$ was formed by vector of all 1s with signs distributed according to the first $N = 100$ digits of the binary expansion of π.

remarkable how stable the behaviour of the relative area is for the scaled permutation matrix $\mathbf{W}$ (black lines), as long as the input coupling $\mathbf{w}$ is aperiodic: we tried vector of all 1s with signs distributed randomly (stars), according to the first N digits of the binary expansion of π (circles) and e (crosses), i.i.d. elements w_i of $\mathbf{w}$ from $N(0, 1)$ (squares) and uniform distribution over $[-1, +1]$ (diamonds). In all cases, the motif richness (measured by relative area covered by Fourier coefficients) monotonically increases with ν up to $\nu = 0.99$ (dashed red vertical line), where there is a phase transition marking the onset of a rapid decline in motif richness. No such behaviour can be observed for random $\mathbf{W}$ ($W_{i,j}$ generated i.i.d. from $N(0, 1)$ (dashed green lines), where the motif richness is consistently low.

Our second measure of motif richness takes into account motif weights, instead of simply noting whether a particular small cell in the complex plane of Fourier coefficients was visited or not. To that end the motif weights were first normalised to the total sum 1. In each cell c we calculate the mean $\bar{q}_c$ of the weights $q_{i,k}$ of coefficients $z_{i,k}$ that landed in that cell. The *relative weighted area* covered by F is the ratio of the accumulated mean weight in cells visited by the coefficients $z_{i,k}$, $\sum_c \bar{q}_c$, to the total number of cells. Figure 13 shows that the relative weighted area exhibits the same universal behaviour as a function of spectral radius ν of $\mathbf{W}$ as that followed by the relative area (figure 12).

9. Discussion and Conclusion

Parametrised state space models have been used extensively in the machine learning community, e.g. in the form of recurrent neural networks. Since learning of the dynamic part

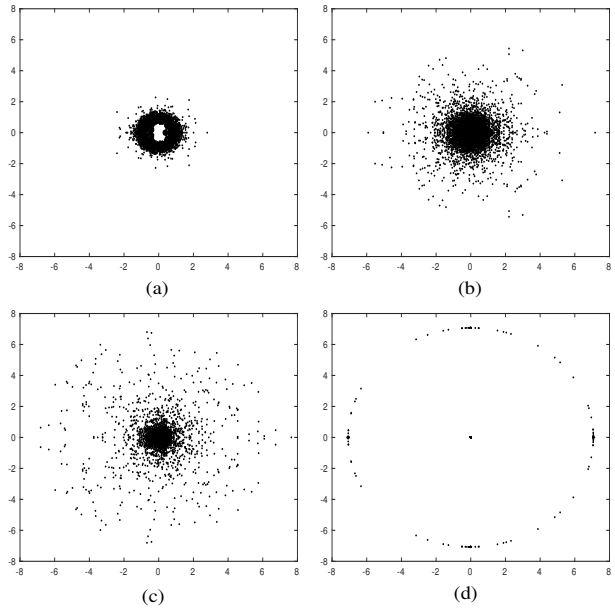

Figure 11: Fourier coefficient distribution of motifs of the temporal kernel associated with dynamical system (1) of state space dimensionality $N = 100$. Dynamic coupling $\mathbf{W}$ was the scaled permutation matrix with spectral radius $\nu = 0.96$ (a), 0.99 (b), 0.996 (c) and 1.0 (d). The input coupling $\mathbf{w}$ was formed by vector of all 1s with signs distributed according to the first $N = 100$ digits of the binary expansion of π. Motifs used have motif weights from the maximal motif weight ω_{max} to $10^{-2}\omega_{max}$.

is known to be difficult, the key idea of reservoir computation is to restrict learning only to the static readout part from the state space, while keeping the underlying dynamical system fixed. Furthermore, the readout is usually a simple linear mapping. This is very similar in spirit to the idea of kernel machines: Transform the inputs using a fixed mapping (usually only implicitly defined) into another "rich" feature space and only train a linear model in that space, with the dot product as the canonical tool. The key to understanding the workings of kernel machines is to understand the feature space: How are the data mapped from the original space into the "richer" feature space? What similarity notions does the inner product in the feature space express in terms of the original data space? This paper is the first study suggesting to formalise and rigorously analyse the connection between fixed dynamics in reservoir computation models and fixed kernel-based transformations to feature spaces in kernel machines. So far, theoretical tools at our disposal that would allow us to make statements regarding appropriateness of different settings of dynamic coupling in reservoir computation models have been rather limited. The new framework introduced in this paper allows us to investigate richness of internal representations of input time series in terms of dynamic states and how they operate to produce desired outputs in terms of matching with a set of temporal motifs defined by the structure of the dynamic coupling. Our investigations lead to several rather surprising results:

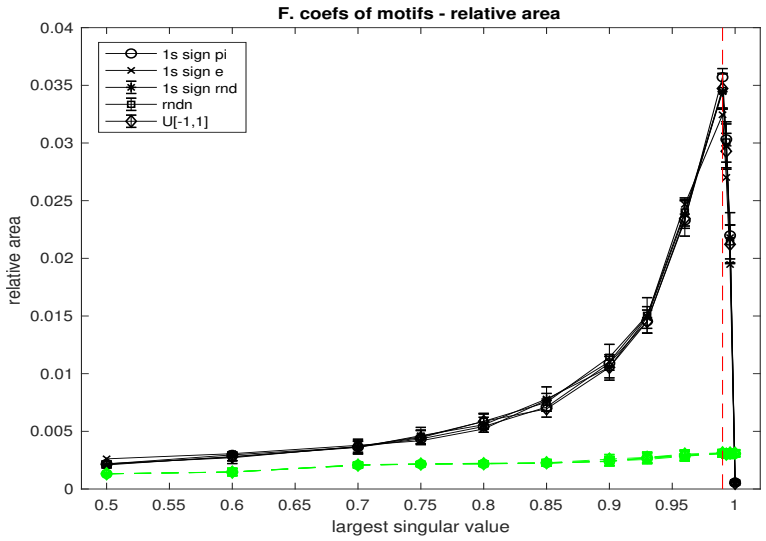

Figure 12: Relative area covered by Fourier coefficients of the temporal kernel motifs as a function of largest singular value of $\mathbf{W}$. Dimensionality of the dynamical system (1) was set to $N = 100$ with dynamic coupling $\mathbf{W}$ formed by the scaled permutation matrix (black lines) or random matrix with individual elements generated i.i.d. from $N(0, 1)$ (dashed green lines). The input coupling $\mathbf{w}$ was formed by vector of all 1s with signs distributed randomly (stars), according to the first $N = 100$ digits of the binary expansion of π (circles) and e (crosses). We also show the case where the elements of $\mathbf{w}$ were generated i.i.d. from $N(0, 1)$ (squares) and uniform distribution over $[-1, +1]$ (diamonds). The vectors $\mathbf{w}$ were renormalised to unit length. In case of deterministic $\mathbf{w}$ but stochastic generation of $\mathbf{W}$, the experiment was repeated 30 times. When both $\mathbf{W}$ and $\mathbf{w}$ were generated randomly, the experiment was repeated 60 times. In those cases, we show the means and standard deviations of the relative area covered by the Fourier coefficients. In each experimental run, the motifs used have their weights ranging from the maximal motif weight ω_{max} to $10^{-2}\omega_{max}$.

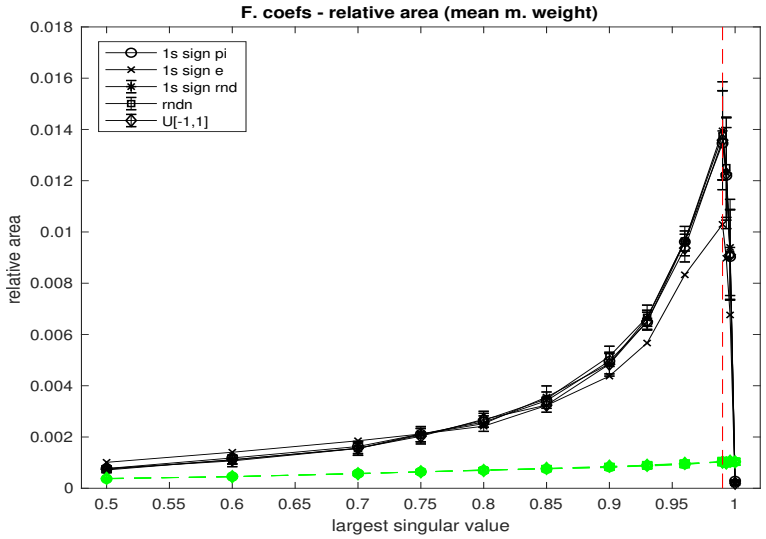

Figure 13: Relative area measured using local mean motif weight of Fourier coefficients of temporal kernel motifs. All other settings are equal to those of figure 12.

1. The usual strategy of random generation (i.i.d.) of input and dynamic coupling weights in the reservoir of ESN leads to shallow memory time series representations, corresponding to cross-correlation operator with fast exponentially decaying coefficients. This finding is quite robust with respect to distributions with which the ESN parameters are generated.

2. Imposing symmetry on coupling weights of the dynamical system yields a constrained dynamic kernel that is a linear combination of trivial kernels matching the input time series with either a straightforward exponentially decaying motif or an exponentially decaying motif of the highest frequency.

3. The simple cycle reservoir topology has been empirically demonstrated to have the potential to equal the performance of more complex reservoir settings (Rodan and Tino, 2010). The dynamical system can have high state space dimensionality, but it is specified only through two free parameters, namely a constant coupling weight $r > 0$ between consecutive reservoir units in the cycle topology[7] and a constant weight $v > 0$ of the input-to-state coupling. The crucial constraint is that the input coupling vector, while all its elements have the same absolute value v, has a-periodically distributed signs. In this paper we have provided rigorous arguments for the need of aperiodic sign distribution in the input coupling, showing that compared with periodic sign patterns, the feature representations of time series in case of a-periodically distributed signs are much richer. In addition, even though such settings of the dynamical system are extremely simple (two free parameters) and completely deterministic[8], the number

7. Note that this also automatically makes the spectral radius ν of $\mathbf{W}$ equal to r, so no separate tuning of ν is needed.
8. The sign distribution can follow binary expansion of an irrational number, such as π or e.

and variety of dynamic motifs of the associated dynamic kernel are far superior to the standard configurations of ESN that rely on stochastic generation of coupling weights.

4. By quantifying motif richness of the dynamic kernel associated with cycle reservoir topology, we showed that there is a phase transition in motif richness at spectral radius values close to, but strictly less than 1. This confirms previous findings in the ESN literature on the importance of tuning the dynamical system at the edge of stability (Bertschinger and Natschlger, 2004).

The arguments in this paper were developed under the assumption of linear dynamical system with linear readout map. However, it has been proved that even linear dynamical systems can be universal, provided they are equipped with polynomial readout maps (Grigoryeva and Ortega, 2018b,a; Gonon and Ortega, 2019). In our setting, this corresponds to considering instead of the linear kernel (eq. (16)) a polynomial kernel (of some degree d),

$$K(\mathbf{u}, \mathbf{v}) = (\langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle + a)^d. \tag{53}$$

Clearly, memory characteristics of such a kernel will not change with offset $a \in \mathbb{R}$ or increasing polynomial degree d. By eqs. (20-21), the polynomial kernel can be written as

$$K(\mathbf{u}, \mathbf{v}) = (\langle \widetilde{\mathbf{u}}, \widetilde{\mathbf{v}} \rangle + a)^d, \tag{54}$$

where the elements $\widetilde{u}_i, \widetilde{v}_i,\ i = 1, 2, ..., N_m$, of $\widetilde{\mathbf{u}}, \widetilde{\mathbf{v}} \in \mathbb{R}^{N_m}$ are projections of $\mathbf{u}, \mathbf{v} \in \mathbb{R}^\tau$ onto motifs $\mathbf{m}_i \in \mathbb{R}^\tau$, scaled by the motif weight. Non-linear manipulation of $\widetilde{u}_i, \widetilde{v}_i$ can increase the capacity of the readout mapping *but only at the level of memory and feature set defined by the motifs*. Randomly generated or symmetric reservoir couplings will still lead to constrained shallow memory kernels. We have shown that simple cycle reservoirs tuned at the edge of stability, with aperiodic sign patterns in input coupling are among the ESN architectures capable of approximating deep memory processes when linear dynamical system and polynomial readout are used. Of course, when non-linearity is allowed in the dynamical system (for example, by employing a logistic sigmoid transfer function), even randomly generated reservoirs may be able to capture deeper memory.

Our study contributes to the debate about what characteristics of the dynamical system are desirable to make it a 'universal' temporal filter capable of producing rich representations of input time series in its state space. Such representations can then be further utilised by readouts, purpose-build for a variety tasks. Ozturk et al. (2007) hypothesised that the distribution of reservoir activations should have high entropy and suggested that it was desirable for the reservoir weight matrix to have eigenvalues uniformly distributed inside the unit circle. In this way the system dynamics would include uniform coverage of time constants (related to the uniform distribution of the poles) (Ozturk et al., 2007). Our work suggests a counterargument when linear reservoirs and non-linear readouts are used: Uniform distribution of eigenvalues inside the unit circle can be achieved by random generation of the reservoir matrix. However, this leads to a highly constrained set of shallow memory motifs of the associated dynamic kernel that describes how features of time series seen in the past contribute to the production of the model output. On the other hand, a very simple setting of high dimensional dynamical system governed by just two free parameters can achieve a much richer and deeper memory motifs of the dynamic kernel. Note that in

this case, the eigenvalues of the reservoir coupling matrix are distributed uniformly along a circle with radius equal to the spectral radius of the reservoir matrix.

Acknowledgement:
This work was supported by the European Commission Horizon 2020 Innovative Training Network SUNDIAL (SUrvey Network for Deep Imaging Analysis and Learning), Project ID: 721463.

Appendix A. Markovian motifs resulting from random dynamical coupling W

In this appendix we demonstrate that the approximations in section 5 of motifs and their weights in the case of random dynamic coupling $\mathbf{W}$ hold for diverse distributions used to generate elements of $\mathbf{W}$. In particular, in figure 14 we show kernel motifs obtained from $\mathbf{W} \in \mathbb{R}^{100\times 100}$ generated element-wise i.i.d. from $N(0,1)$ and renormalised to largest singular value $\nu = 0.995$ (the setting used in section 5) and input coupling vector $\mathbf{w}$ generated as a vector of all 1s with randomly flipped signs (in each dimension with equal probability 0.5), renormalised to unit vector. The associated squared motif weights are presented in figure 15. We also show in figures 16 and 17 the kernel motifs and eigenvalues of $\mathbf{Q}$ when both $\widetilde{\mathbf{W}}$ and $\mathbf{w}$ consist of all 1s with signs flipped independently element-wise with probability 0.5 (dynamical coupling renormalised to largest singular value $\nu = 0.995$ and $\mathbf{w}$ to unit vector). In both cases, the Markovian motifs and their weights are almost indistinguishable from those shown in section 5 (figures 2 and 3).

References

L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer. Information processing using a single dynamical node as complex system. *Nature Comm.*, 2, 2011.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 12 1966.

Y. Bengio, P. Frasconi, and P Simard. The problem of learning long-term dependencies in recurrent networks. In *Proceedings of the 1993 IEEE International Conference on Neural Networks*, volume 3, pages 1183–1188, 1993.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Nils Bertschinger and Thomas Natschlger. Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7):1413–1436, 2004.

Terry Bossomaier, Lionel Barnett, Michael Harr, and Joseph T. Lizier. *An Introduction to Transfer Entropy: Information Flow in Complex Systems*. Springer Publishing Company, Incorporated, 1st edition, 2016.

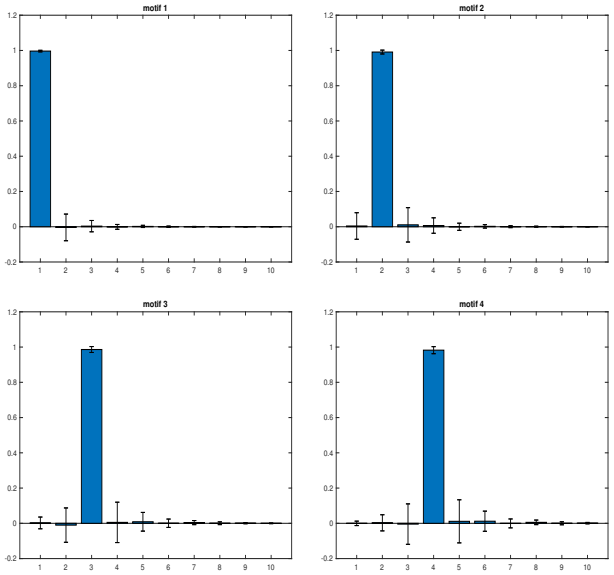

Figure 14: The first 10 elements of the four most dominant kernel motifs corresponding to $\mathbf{W} \in \mathbb{R}^{100\times100}$ generated element-wise i.i.d. from $N(0,1)$ and renormalised to largest singular value $\nu = 0.995$. The input coupling $\mathbf{w}$ was generated as a vector of all 1s with randomly flipped signs (in each dimension with equal probability 0.5). Shown are the means and standard deviations across 100 joint realisations of $\mathbf{W}$ and $\mathbf{w}$.

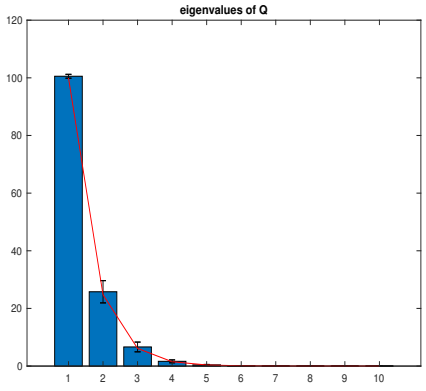

Figure 15: Eigenvalues (squared motif weights) of the metric tensor $\mathbf{Q}$ for random setting of the dynamical system (1) as described in fig. 14. Solid bars correspond to the means of the actual eigenvalues λ_i across the 100 realisations of $\mathbf{W}$ and $\mathbf{w}$ (also shown are standard deviations). The theoretically predicted values (eq. (35)) are shown as the red line.

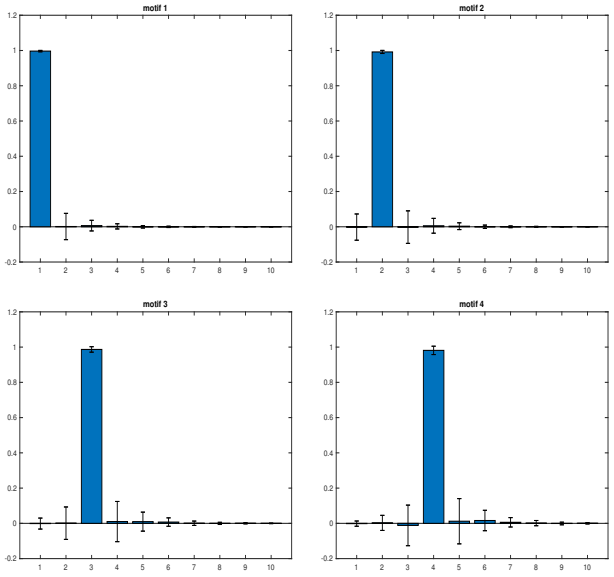

Figure 16: The first 10 elements of the four most dominant kernel motifs corresponding to $\mathbf{W} \in \mathbb{R}^{100\times 100}$ and $\mathbf{w}$, both consisting of all 1s with signs flipped independently element-wise with probability 0.5. $\mathbf{W}$ was renormalised to largest singular value $\nu = 0.995$. Shown are the means and standard deviations across 100 joint realisations of $\mathbf{W}$ and $\mathbf{w}$.

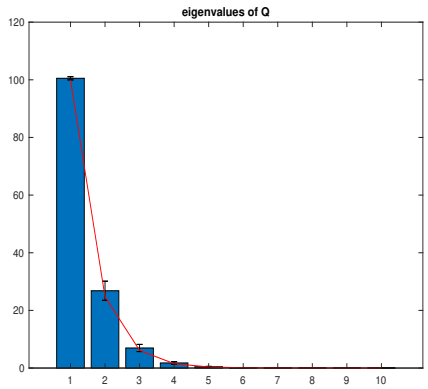

Figure 17: Eigenvalues (squared motif weights) of the metric tensor $\mathbf{Q}$ for random setting of the dynamical system (1) as described in fig. 16. Solid bars correspond to the means of the actual eigenvalues λ_i across the 100 realisations of $\mathbf{W}$ and $\mathbf{w}$ (also shown are standard deviations). The theoretically predicted values (eq. (35)) are shown as the red line.

K. Bush and C. Anderson. Modeling reward functions for incomplete state representations via echo state networks. In *Proceedings of the International Joint Conference on Neural Networks, Montreal, Quebec*, July 2005.

R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali. Training performance of echo state neural networks. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–4, 2016.

Romain Couillet, Gilles Wainrib, Harry Sevi, and Hafiz Tiomoko Ali. The asymptotic performance of linear echo state neural networks. *Journal of Machine Learning Research*, 17(178):1–35, 2016.

J Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information processing capacity of dynamical systems. *Scientific reports*, 2:514, 07 2012.

Carlton Downey, Ahmed Hefny, Boyue Li, Byron Boots, and Geoff Gordon. Predictive state recurrent neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6055–6066, USA, 2017. Curran Associates Inc.

M. Ficura. Forecasting stock market realized variance with echo state neural networks. *European Financial and Accounting Journal*, 3, 2017. doi: 10.18267/j.efaj.193.

C. Gallicchio, A. Micheli, and L. Pedrelli. Deep reservoir computing: A critical experimental analysis. *Neurocomputing*, 268:87 – 99, 2017.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105:18970–18975, 2008.

Surya Ganguli and Haim Sompolinsky. Short-term memory in neuronal networks through dynamical compressed sensing. In *Advances in neural information processing systems*, pages 667–675, 2010.

Lukas Gonon and Juan-Pablo Ortega. Reservoir computing universality with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, To appear, 02 2019.

L. Grigoryeva and J.-P. Ortega. Echo state networks are universal. *Neural Networks*, 108: 495 – 508, 2018a.

L. Grigoryeva and J.-P. Ortega. Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *J. Mach. Learn. Res.*, 19(1):892–931, January 2018b.

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. *Neural Computation*, 28(7):1411–1451, 2016.

B. Hammer. Generalization ability of folding networks. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):196–206, 2001.

B. Hammer and P. Tino. Recurrent neural networks with small weights implement definite memory machines. *Neural Computation*, 15(8):1897–1929, 2003.

J. Hochreiter and J. Schmidhuber. Long short term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

G. Holzmann and H. Hauser. Echo state networks with filter neurons and a delay and sum readout. *Neural Networks*, 32(2):244–256, 2009.

H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical report gmd report 148, German National Research Center for Information Technology, 2001.

H. Jaeger. Short term memory in echo state networks. Technical report gmd report 152, German National Research Center for Information Technology, 2002a.

H. Jaeger. A tutorial on training recurrent neural networks, covering bppt, rtrl, ekf and the "echo state network" approach. Technical report gmd report 159, German National Research Center for Information Technology, 2002b.

H. Jaeger and H. Hass. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless telecommunication. *Science*, 304:78–80, 2004.

H. Jaeger, M. Lukosevicius, D. Popovici, and U. Siewert. Optimisation and applications of echo state networks with leaky-integrator neurons. *Neural Networks*, 20(3):335–352, 2007.

Li Jing, Caglar Gulcehre, John Peurifoy, Yichen Shen, Max Tegmark, Marin Soljacic, and Yoshua Bengio. Gated orthogonal recurrent units: On learning to forget. *Neural Computation*, 31(4):765–783, 2019.

R. E. Kalman. Approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35, 1960.

R. Legenstein and W. Maass. What makes a dynamical system computationally powerful? In S. Haykin, J. C. Principe, T. Sejnowski, and J. McWhirter, editors, *New Directions in Statistical Signal Processing: From Systems to Brains*, pages 127–154. MIT Press, 2007.

Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. Detecting non-trivial computation in complex dynamics. In *Proceedings of the 9th European Conference on Advances in Artificial Life*, ECAL'07, pages 895–904, Berlin, Heidelberg, 2007. Springer-Verlag.

Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. Local measures of information storage in complex distributed computation. *Inf. Sci.*, 208:39–54, November 2012.

M. Lukosevicius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

W. Maass, T. Natschlager, and H. Markram. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

G. Manjunath and H. Jaeger. Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. *Neural Computation*, 25(3):671–696, 2013. ISSN 0899-7667.

Henryk Minc. *Nonnegative Matrices*. John Wiley and Sons Ltd-Interscience Series in Discrete Mathematics and Optimization, 1988.

O. Obst and J. Boedecker. Guided self-organization of input-driven recurrent neural networks. In *In Guided Self-Organization: Inception. Emergence, Complexity and Computation*, pages 319–340. Springer, Berlin, Heidelberg, 2014.

Oliver Obst, Joschka Boedecker, and Minoru Asada. Improving recurrent neural network performance using transfer entropy. In *Proceedings of the 17th International Conference on Neural Information Processing: Models and Applications - Volume Part II*, ICONIP'10, pages 193–200, Berlin, Heidelberg, 2010. Springer-Verlag.

M. C. Ozturk, D. Xu, and J. Principe. Analysis and design of echo state network. *Neural Computation*, 19(1):111–138, 2007.

A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–144, 2011.

Ali Rodan and Peter Tino. Minimum complexity echo state network. ieee trans neural netw. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 22:131–44, 11 2010. doi: 10.1109/TNN.2010.2089641.

André Röhm and Kathy Lüdge. Multiplexed networks: reservoir computing with virtual and real nodes. *Journal of Physics Communications*, 2(8):085007, aug 2018.

M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *In Proceedings of the International Congress of Mathematicians. Volume III, Hindustan Book Agency, New Delhi*, pages 1576–1602, 2010.

B. Schrauwen, M. Wardermann, D. Verstraeten, J.J. Steil, and D Stroobandt. Improving reservoirs using intrinsic plasticity. *Neurocomputing*, 71(7-9):1159–1171, 2008.

Hava T. Siegelmann and Eduardo D. Sontag. Analog computation via neural networks. *Theoretical Computer Science*, 131(2):331–360, 1994.

M.D. Skowronski and J.G. Harris. Minimum mean squared error time series classification using an echo state network prediction model. In *IEEE International Symposium on Circuits and Systems, Island of Kos, Greece, pp. 3153-3156*, 2006.

J. Steil. Online reservoir adaptation by intrinsic plasticity for backpropagation-decorrelation and echo state learning. *Neural Networks*, 20:353–364, 2007.

Tobias Strauss, Welf Wustlich, and Roger Labahn. Design strategies for weight matrices of echo state networks. *Neural Computation*, 24(12):3246–3276, 2012. ISSN 0899-7667.

Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Hroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing: A review. *Neural Networks*, 115:100 – 123, 2019.

P. Tino. Asymptotic fisher memory of randomized linear symmetric echo state networks. *Neurocomputing*, 298:4–8, 2018.

P. Tino and G. Dorffner. Predicting the future of discrete sequences from fractal representations of the past. *Machine Learning*, 45(2):187–218, 2001.

P. Tino and B. Hammer. Architectural bias in recurrent neural networks: Fractal analysis. *Neural Computation*, 15(8):1931–1957, 2004.

P. Tino and A. Rodan. Short term memory in input-driven linear dynamical systems. *Neurocomputing*, 112:58–63, 2013.

M. H. Tong, A.D. Bicket, E.M. Christiansen, and G.W. Cottrell. Learning grammatical structure with echo state network. *Neural Networks*, 20:424–432, 2007.

Y. Xue, L. Yang, and S. Haykin. Decoupled echo state networks with lateral inhibition. *Neural Networks*, 20:365–376, 2007.

Izzet B. Yildiz, Herbert Jaeger, and Stefan J. Kiebel. Re-visiting the echo state property. *Neural Networks*, 35:1 – 9, 2012. ISSN 0893-6080.

Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10:402–15, 09 2009.

B. Zhang, D. J. Miller, and Y. Wang. Nonlinear system modeling with random matrices: Echo state networks revisited. *IEEE Transactions on Neural Networks and Learning Systems*, 23(1):175–182, Jan 2012. ISSN 2162-2388. doi: 10.1109/TNNLS.2011.2178562.