On the Linearizing Effect of Spatial Averaging in Large-Scale Populations of Homogeneous Nonlinear Systems

Sabbir Ahmed Erfan Nozari

***Abstract*— Understanding the dynamics resulting from large-scale populations of systems poses one of the greatest challenges ahead of modern science. While it is often expected that the emerging dynamics from such populations compound in complexity, we here show that, on the contrary, the aggregation of complex individual dynamics can in fact lead to simpler behavior overall. In particular, mounting empirical evidence from neuroscience and beyond has pointed out the linearity of macroscopic dynamics that result from the interaction of large populations of microscopic subsystems, despite the highly nonlinear dynamics possessed by the individual subsystems. Rigorous analyses and theoretical grounds for such observations, however, have remained lacking. In this paper, we develop a general theoretical framework showing that the average dynamics of a broad family of populations of nonlinear stochastic subsystems converge to linear time-varying (LTV) dynamics transiently and to linear time-invariant (LTI) dynamics in steady state. Simulations are provided to illustrate this effect in populations of static (feedforward) nonlinear maps as well as a wide range of nonlinear systems exhibiting bistable, limit cycle, and chaotic dynamics.**

I. Introduction

A critical aspect in the study of large-scale systems, from engineered systems such as the traffic and power networks to natural phenomena exhibited by social, neural, or epidemic dynamics, is the spatial scale of analysis [1]–[6]. While complex and potentially nonlinear dynamics are often exhibited by individual (microscopic) subsystems in each case, greater interest often exists in the characterization of the overall (macroscopic) dynamics of large-scale populations [7]–[12]. Nonlinear analyses, however, often scale poorly to large-scale systems, thus motivating the use of linear models as local linearizations around nominal equilibria or merely as approximations in favor of tractability. In this work, we contribute to this broad body of research by rigorously showing that under certain assumptions, linear models provide an accurate description of the average dynamics of a population of homogeneous stochastic subsystems, even when the subsystem dynamics are nonlinear and/or non-Gaussian.

Literature review. The present work is particularly motivated by our prior work [13] on the dynamic modeling of macroscopic brain networks, even though our theoretical analyses herein are broad and applicable to a wide range of nonlinear systems. Therein, we provided empirical evidence based on real and simulated data that the aggregate dynamics of large-scale populations of neurons appear to be most accurately described by a stochastic linear model even though the nonlinearity of the dynamics of each individual neuron is both theoretically and empirically established [14], [15]. Similar observations of linearity have also emerged in recent works on large-scale biological [16]–[19] and artificial [20]–[22] neural networks. For example, in modeling large-scale brain dynamics resulting from electrical stimulation [16] as well as those that are relevant for a subject's motor behavior [17], nonlinear models have failed to perform better than linear ones, thus questioning the presence and/or extent of nonlinearity in the dynamics of large-scale brain networks. Similarly, nonlinear models are found to be no more accurate than linear ones for age or sex prediction from structural or functional brain scans [18] while [19] found direct linear relationships between visual stimuli and macroscopic brain recordings, further pointing to a potential lack of nonlinearity in data acquired from large-scale brain networks. On the other hand, in artificial neural networks (ANNs) with nonlinear activation functions, the linearity of gradient descent dynamics with respect to network parameters has been shown in [20], [21] when the network width is increased to infinity. Similar results on the asymptotic linearity of ANNs with respect to their parameters have also been shown previously in [22], [23]. Motivated by this body of mounting evidence, we here provide rigorous proofs for the linearity of aggregate activity of large-scale populations of both static (feedforward) and dynamic (recurrent) nonlinear subsystems.

The emergence of linear dynamics from a population of nonlinear subsystems is also similar to stochastic linearization, a.k.a. quasilinearization [24]. In stochastic linearization, however, a nonlinear mapping $y = f(x)$ is *approximated* by $\hat{y} = \mathbb{E}[\frac{\partial y}{\partial x}]x$, whereas the asymptotically linear dynamics described in this work provide an exact (stochastic) characterization of the aggregate dynamics of large-scale populations. Finally, our proposed framework relies on tools from probability theory [25], [26] including, in particular, the multivariate Central Limit Theorem (CLT), as well as tools from stochastic differential equations and the theory of continuous-space Markov Chains [27], [28].

Statement of contributions. In the present work, we propose a theoretical framework to explain the linearizing effects of spatial averaging observed in many populations of nonlinear dynamical subsystems. Our contributions in this regard are threefold. Starting from static (feedforward) nonlinear maps, our first contribution involves the characterization of

Sabbir Ahmed is with the Department of Electrical and Computer Engineering, University of California, Riverside, sahme094@ucr.edu. Erfan Nozari is with the Departments of Mechanical Engineering, Electrical and Computer Engineering, and Bioengineering, University of California, Riverside, erfan.nozari@ucr.edu.

the convergence of the average input-output relationship of large homogeneous populations of nonlinear subsystems to a linear relationship as the size of the population grows. Interestingly, we show that this asymptotic linearity holds regardless of the functional form of the subsystems' nonlinearity and their input distributions so long as their inputs and outputs have finite first and second moments. Our second contribution then involves the generalization of this result to populations of dynamical systems. Here, we also show that under mild assumptions on the dynamics of each subsystem, the average population dynamics converge to a linear time-varying (LTV) system as the size of the population grows. In our extensive simulations, however, we often observe the emergence of linear time-invariant (LTI) dynamics from such large-scale populations. This motivated our third and final contribution where we prove, under additional assumptions on the subsystem dynamics, the convergence of the aggregate population dynamics to an LTI system as both the size of the population and time grow to infinity. Overall, our results provide an important first step in the rigorous characterization of a robust and pervasive phenomenon in large-scale systems that can be of broad interest to the systems and control community and beyond.

II. Notation

We use $\mathbb{R}$ and $\mathbb{Z}$ to denote the set of reals and integers, respectively. $\mathbf{0}_n$ and $\mathbf{0}_{m\times n}$ stand for the the n-vector of all zeros and the m-by-n zero matrix. The subscripts are omitted when clear from the context. When a vector $\mathbf{y}$ or matrix $\mathbf{A}$ are block-partitioned, $\mathbf{y}_i$ and $\mathbf{A}_{ij}$ refer to the ith block of $\mathbf{y}$ and (i,j)th block of $\mathbf{A}$, respectively. $\|\cdot\|$ denotes the vector l_2-norm. For any $t\in\mathbb{R}$, $\lfloor t\rfloor = \max\{k\in\mathbb{Z} | k\le t\}$ and $\lceil t\rceil = \min\{k\in\mathbb{Z} | k\ge t\}$.

Throughout this work, all probabilities are defined on measurable spaces consisting of a Euclidean space (or a subset thereof) and the associated Borel σ-algebra. Hence, when clear from the context, the space over which each probability is defined is omitted. $\mathbb{E}[\cdot]$ and $\mathbb{P}\{\cdot\}$ denote expectation and probability, respectively. For two random vectors $\mathbf{x}\in\mathbb{R}^n$ and $\mathbf{y}\in\mathbb{R}^m$, $\mathrm{Cov}(\mathbf{x},\mathbf{y})\in\mathbb{R}^{n\times m}$ denotes the covariance between them. If π is a probability distribution on $\mathbb{R}^n$ and $A\subseteq\mathbb{R}^n$, $\pi(A)=\int_A d\pi$, i.e., the probability assigned to A by π.

III. Problem Statement

As detailed in Section I, a mounting body of empirical works have observed the linearity of macroscopic biological and artificial neural dynamics despite the nonlinearity of the constituting subsystems. Interestingly, this phenomenon can be observed in the aggregate dynamics of a large range of nonlinear systems exhibiting various distinctively nonlinear phenomena such as bistability, limit cycles, and chaos (Figure 1). In this work, we seek the development of a general theoretical framework that can explain and rigorously characterize this observed emergent linearity at the macroscale.

Consider a homogeneous population of N dynamical subsystems each represented by the general nonlinear discrete-time form

$$\begin{aligned} \mathbf{x}_i(t+1) &= f(\mathbf{x}_i(t),\mathbf{w}_i(t)),\\ \mathbf{w}_i(t) &\sim p_w,\\ \mathbf{x}_i(0) &\sim p_0, \qquad i=1,\dots,N, \end{aligned} \tag{1}$$

where $\mathbf{x}_i(t)\in\mathbb{R}^n$ represents the state of i-th dynamical subsystem at time t, $\mathbf{w}_i(t)\in\mathbb{R}^m$ represents the noise at time t with the time-invariant distribution p_w, and p_0 represents the distribution of the initial state $\mathbf{x}_i(0)$. Given how universal the linearizing property of spatial averaging appears to be based on our simulations (cf., e.g., Figure 1), we allow the functional form of the nonlinearity $f(\cdot)$ and the distributions of the initial conditions and noise inputs to be general, except for a limited set of assumptions described later in Section IV-B. In this context, we can formulate our problem of interest as follows.

Problem 1. ***(Effects of spatial averaging on aggregate dynamics of population of nonlinear dynamical systems).*** *Consider a homogenous population of discrete-time nonlinear dynamical systems described by* (1). *Describe the resulting average dynamics of the population, namely, that of the average state vector*

$$\bar{\mathbf{x}}(t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\mathbf{x}_i(t)-\mathbb{E}[\mathbf{x}_i(t)]), \tag{2}$$

and show its asymptotic convergence to a linear system as $N\to\infty$. □

In defining the average state vector in (2), it is essential to mean-center each $\mathbf{x}_i(t)$ before averaging in order to prevent the drift of the expected value of $\bar{\mathbf{x}}(t)$ to infinity while, at the same time, scaling the average by $\frac{1}{\sqrt{N}}$ (instead of $\frac{1}{N}$) to prevent the decay of the variance of $\bar{\mathbf{x}}(t)$ to zero. We should point out, however, that mean-centering and normalization only matter from a technical perspective as we study the asymptotic limit of $N\to\infty$, while the effect of spatial averaging is often significant and robust enough to fully linearize aggregate population dynamics even for small values of N, as shown in Figure 1.

IV. Main Results

In this section we present the main findings of our paper with regards to Problem 1. To begin, we investigate in Section IV-A a simplified version of Problem 1 including only static and scalar nonlinear mappings. This simplified setup, in addition to being of significant independent interest, allows us to focus on the key ingredients that underlie the linearizing effect of spatial averaging on dynamical systems addressed in Section IV-B, without involving the additional complexities that arise therein.

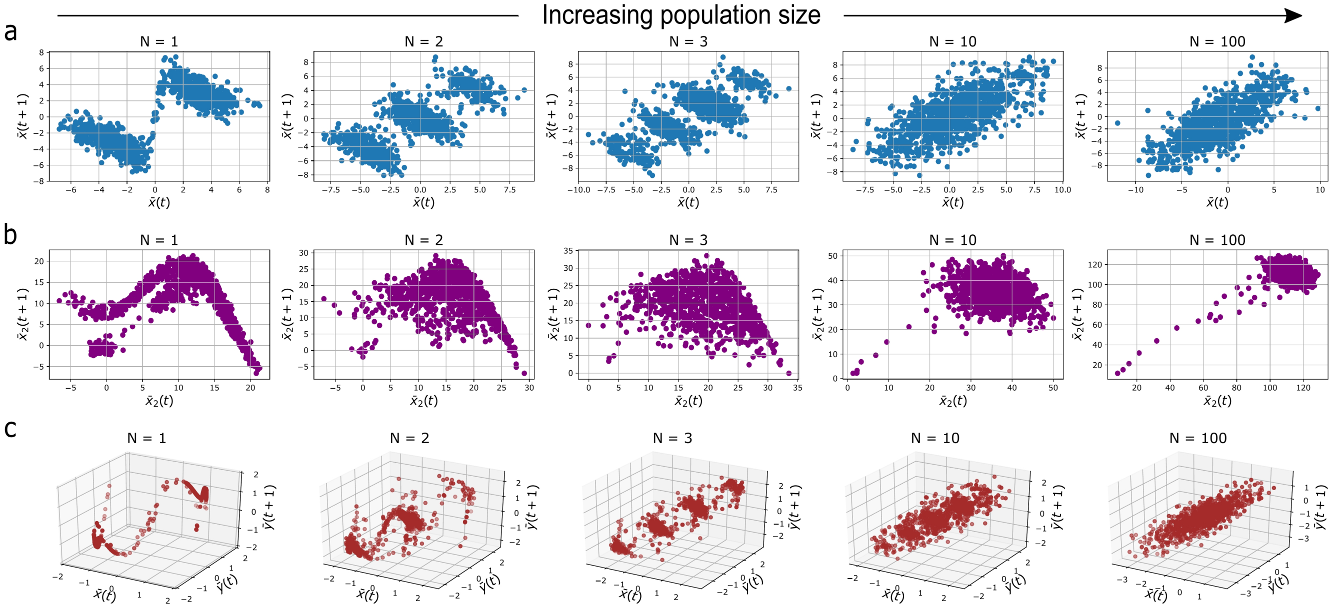

Fig. 1: **Spatial averaging can linearize diverse families of nonlinear dynamical systems. (a)** The linearizing effect of spatial averaging on a population of dynamical subsystems with bistable dynamics. Each subsystem is given by $x_i(t+1) = 10x_i(t)/(1+x_i^2(t)) + w_i(t)$ and the average state $\bar{x}(t) = 1/\sqrt{N}\sum_{i=1}^N x_i(t)$ is plotted at times t and $t+1$ on the x and y axes, respectively. **(b)** Similar to (a) but for subsystem dynamics exhibiting limit cycles. Each subsystem has planar dynamics described by $x_{i1}(t+1) = x_{i2}(t), x_{i2}(t+1) = 0.4x_{i1}(t) + 0.4x_{i2}(t) + \sin(0.2x_{i2}(t))x_{i2}(t) + w_i(t)$. The average state $\bar{x}_2(t) = 1/\sqrt{N}\sum_{i=1}^N x_{i2}(t)$ is plotted at times t and $t+1$ on the x and y axes, respectively. **(c)** Similar to (a,b) but for a population of subsystems exhibiting chaotic behavior (Duffling map). Each subsystem has the form $x_i(t+1) = y_i(t)$, $y_i(t+1) = -0.2x_i(t) + 2.75y_i(t) - y_i^3(t) + 10^{-3/2}w_i(t)$. The average states $\bar{x}(t) = 1/\sqrt{N}\sum_{i=1}^N x_i(t)$, $\bar{y}(t) = 1/\sqrt{N}\sum_{i=1}^N y_i(t)$ and $\bar{y}(t+1)$ are plotted on the x, y, and z axes, respectively. The noise $w_i(t)$ in all cases has a white standard normal distribution. Despite the high degree of nonlinearity in the dynamics of individual subsystems in each case, the average dynamics become linear for even small values of N.

A. Linearizing Effect of Averaging on Static Nonlinearities

At the core of Problem 1 is the effect of spatial averaging on the vector field $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$. In fact, by taking the average of both sides of (1), after appropriate mean-centering and scaling, we have

$$\bar{\mathbf{x}}(t+1) = \frac{1}{\sqrt{N}}\sum_{i=1}^N f(\mathbf{x}_i(t), \mathbf{w}_i(t)) - \mathbb{E}[f(\mathbf{x}_i(t), \mathbf{w}_i(t))]$$

and thus Problem 1 primarily concerns the relationship between the average of $f(\mathbf{x}_i, \mathbf{w}_i)$ and the averages of its arguments $\mathbf{x}_i$ and $\mathbf{w}_i$. As such, in this section we start tackling Problem 1 by first showing how spatial averaging can linearize static scalar nonlinear maps, and then tackle the additional complexities that arise from the dynamic nature of Problem 1 in Section IV-B.

Consider a sequence of N, independent and identically distributed (i.i.d.) random variables $x_1, x_2, \ldots, x_N$ and let

$$y_i = f(x_i), \qquad i = 1, 2, \ldots, N \tag{3}$$

where $f : \mathbb{R} \to \mathbb{R}$. Define the average variables

$$\bar{x} = \frac{1}{\sqrt{N}}\sum_{i=1}^N (x_i - \mu_x), \qquad \bar{y} = \frac{1}{\sqrt{N}}\sum_{i=1}^N (y_i - \mu_y) \tag{4}$$

where μ_x, μ_y are the means of x_i and y_i, respectively. Similar to the linearizing effects depicted in Figure 1, extensive simulations with various distributions of x_i and nonlinear forms for f show that the relationship between $\bar{x}$ and $\bar{y}$ becomes asymptotically linear as $N \to \infty$ (see, e.g., Figure 2).

As seen in Figure 2, the (linear) relationship between $\bar{x}$ and $\bar{y}$ is stochastic, even when the relationship between each x_i and y_i is deterministic. Therefore, in order to prove that this relationship becomes linear, one has to show that the conditional expectation

$$\mathbb{E}[\bar{y}|\bar{x} = \xi]$$

converges to a linear function of ξ as $N \to \infty$. A numerical approximation of this value is also shown in Figure 2. Note that the linearity of this conditional expectation is precisely the linearity assumption of simple linear regression [29], and can be shown for a broad class of nonlinear functions f and distributions of $x_1, \ldots, x_N$, as follows.

Theorem IV.1. ***(Linearizing effect of averaging on i.i.d. data).*** *Consider N i.i.d. random variables $x_1, x_2, \ldots, x_N$ and their nonlinear maps $y_1, y_2, \ldots, y_N$ as defined in* (3). *Assume that the distribution of x_i and the nonlinear map $f(\cdot)$ are such that x_i and y_i have finite mean and variance. Then the relationship between the average variables $\bar{y}$ and*

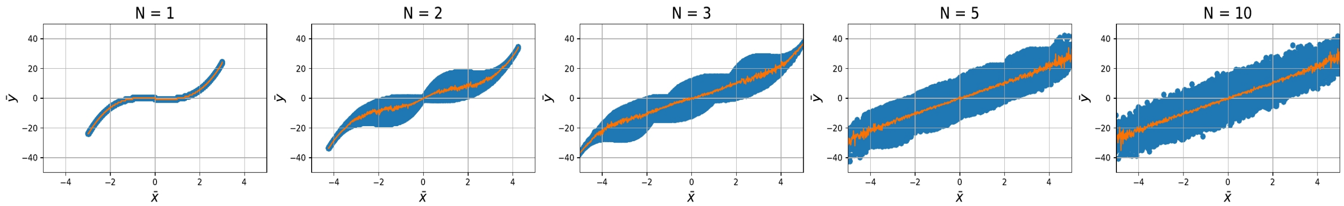

Fig. 2: **Spatial averaging of a static nonlinearity.** The panels show the linearizing effect of spatial averaging for an example static scalar map of the form $y_i = \lfloor x_i^3 - x_i \rfloor$ with $x_i \sim \mathcal{U}[-3, 3]$. Each panel illustrates 10^5 samples of the average variables $\bar{x}$ and $\bar{y}$, each averaged over a varying number N of subsystems. Each panel also demonstrates a sample estimate of $\mathbb{E}[\bar{y}|\bar{x} = x_0]$ (orange line) which becomes increasingly linear even for small values of N. The conditional expectations of the form $\mathbb{E}[\bar{y}|\bar{x} = x_0]$ are approximated by a Gaussian-weighted sample average of y_i where each y_i is weighted based on the distance between the corresponding x_i and x_0.

$\bar{x}$ as defined in (4) *is asymptotically linear, i.e.,*

$$\mathbb{E}[\bar{y}|\bar{x} = \xi] \to \frac{\sigma_{xy}}{\sigma_x^2}\xi$$

as $N \to \infty$, where σ_{xy} is the covariance between x_i and y_i and σ_x^2 is the variance of x_i. □

The proof of Theorem IV.1 is a special case of the proof of Theorem IV.2 and therefore omitted. This result is, however, of far-reaching independent interest as it proves a fundamental and general relationship between averaging and linearity, akin to (and based on) the fundamental relationship between averaging and normality established by the CLT. In the next section, we build on the core idea of Theorem IV.1 and address the additional challenges that arise from the dynamic nature of Problem 1.

B. Linearizing Effect of Averaging on Nonlinear Dynamics

In this section, we discuss the main problem outlined in Problem 1. Throughout this section, we make the following assumptions on the population dynamics in (1).

Assumption 1. ***(Standing assumptions).***

(A1) *The function $f(.)$ is bounded, i.e.,*

$$||f(\mathbf{x}, \mathbf{w})|| \leq M < \infty, \qquad \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{w} \in \mathbb{R}^m.$$

(A2) *The noise input is zero mean, white, and finite-variance, so for all i,*

$$\begin{aligned} &\mathbb{E}[\mathbf{w}_i(t)] = \mathbf{0}, && \forall t, \\ &\mathbb{E}[\mathbf{w}_i(t)\mathbf{w}_i(\tau)^T] = \mathbf{0}, && \forall t \neq \tau, \\ &||\mathbb{E}[\mathbf{w}_i(t)\mathbf{w}_i(t)^T]|| \leq C < \infty, && \forall t. \end{aligned}$$

(A3) *The state $\mathbf{x}_i(t)$ at time t is independent of the noise $\mathbf{w}_i(t)$ at the same time t, i.e., for all $\boldsymbol{\xi}, \boldsymbol{\zeta}, t$, and i,*

$$\mathbb{P}\{\mathbf{x}_i(t) \leq \boldsymbol{\xi}, \mathbf{w}_i(t) \leq \boldsymbol{\zeta}\} = \mathbb{P}\{\mathbf{x}_i(t) \leq \boldsymbol{\xi}\}\mathbb{P}\{\mathbf{w}_i(t) \leq \boldsymbol{\zeta}\}.$$

(A4) *The initial conditions $\mathbf{x}_i(0)$ and noise inputs $\mathbf{w}_i(t)$ are independent among the subsystems, i.e.,*

$$\begin{aligned} &\mathbb{P}\{\mathbf{x}_i(0) \leq \boldsymbol{\xi}_i, \mathbf{w}_i(t) \leq \boldsymbol{\zeta}_i, \mathbf{x}_j(0) \leq \boldsymbol{\xi}_j, \mathbf{w}_j(t) \leq \boldsymbol{\zeta}_j\} \\ &= \mathbb{P}\{\mathbf{x}_i(0) \leq \boldsymbol{\xi}_i\}\mathbb{P}\{\mathbf{w}_i(t) \leq \boldsymbol{\zeta}_i\}\mathbb{P}\{\mathbf{x}_j(0) \leq \boldsymbol{\xi}_j\} \\ &\quad \mathbb{P}\{\mathbf{w}_j(t) \leq \boldsymbol{\zeta}_j\}, \end{aligned}$$

for all $i \neq j$, t, and $\boldsymbol{\xi}_i, \boldsymbol{\xi}_j, \boldsymbol{\zeta}_i, \boldsymbol{\zeta}_j$. □

A number of remarks about the foregoing assumptions are in order. First, note that (A1) is satisfied by models of most, if not all, real-world systems due to the inherent boundedness of any states that represent a physical quantity. For models with unbounded vector fields but uniformly bounded trajectories (often used to describe the dynamics of real-world systems in certain "working" regions of their state space), f can be augmented with sufficiently large saturations $f_i \leftarrow \max\{\min\{f_i, M\}, -M\}$ to satisfy (A1) without affecting the effective subsystem dynamics. Since f is otherwise arbitrary, (A2) is also not restrictive in practice since the color (internal dynamics) or non-zero mean of any noise input can be absorbed into f. Given the whiteness of $\mathbf{w}_i(t)$, (A3) is mild as well since it takes at least one time step before the information in $\mathbf{w}_i(t)$ reflects in the state $\mathbf{x}_i(t+1)$. Our strongest assumption is (A4) which, in turn, does not allow for correlations between the states of the subsystems, as would be the case, e.g., for networked systems. This assumption is nevertheless needed only from a technical perspective and its relaxation will be pursued in future extensions of this work to subsystems with correlated activity.

With the stated assumptions above, we are ready to characterize the asymptotic aggregate dynamics of the population of dynamical systems in (1), as follows.

Theorem IV.2. ***(Linearizing effect of spatial averaging on population of dynamical systems).*** *Consider the population dynamics in* (1) *and assume that assumptions* (A1)-(A4) *hold. Let*

$$\begin{aligned} \boldsymbol{\mu}_x(t) &= \mathbb{E}[\mathbf{x}_i(t)] \\ \boldsymbol{\Sigma}_{xx}(t) &= \mathrm{Cov}(\mathbf{x}_i(t), \mathbf{x}_i(t)) \\ \boldsymbol{\Sigma}_{ww}(t) &= \mathrm{Cov}(\mathbf{w}_i(t), \mathbf{w}_i(t)) \\ \boldsymbol{\Sigma}_{x^+x}(t) &= \mathrm{Cov}(\mathbf{x}_i(t+1), \mathbf{x}_i(t)) \\ \boldsymbol{\Sigma}_{x^+w}(t) &= \mathrm{Cov}(\mathbf{x}_i(t+1), \mathbf{w}_i(t)) \end{aligned}$$

and define the average population variables

$$\bar{\mathbf{x}}(t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\mathbf{x}_i(t) - \boldsymbol{\mu}_x(t)),$$

$$\bar{\mathbf{w}}(t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{w}_i(t).$$

Then the relationship between $\bar{\mathbf{x}}(t+1)$, $\bar{\mathbf{x}}(t)$ and $\bar{\mathbf{w}}(t)$ becomes asymptotically linear as $N \to \infty$, i.e., for any $\boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{\omega} \in \mathbb{R}^m$,

$$\mathbb{E}\left[\bar{\mathbf{x}}(t+1) \middle| \begin{bmatrix}\bar{\mathbf{x}}(t)\\ \bar{\mathbf{w}}(t)\end{bmatrix} = \begin{bmatrix}\boldsymbol{\xi}\\ \boldsymbol{\omega}\end{bmatrix}\right] \to \mathbf{A}(t)\boldsymbol{\xi} + \mathbf{B}(t)\boldsymbol{\omega} \tag{5}$$

as $N \to \infty$, where

$$\begin{aligned}\mathbf{A}(t) &= \boldsymbol{\Sigma}_{x^+x}(t)\boldsymbol{\Sigma}_{xx}(t)^{-1}\\ \mathbf{B}(t) &= \boldsymbol{\Sigma}_{x^+w}(t)\boldsymbol{\Sigma}_{ww}(t)^{-1}.\end{aligned} \tag{6}$$

Proof. First, we show that the average population states and noises jointly approach normality. Then, we show that the mean of the population average state, when conditioned on the past state and noise, approaches a linear function thereof.

To prove joint normality, let

$$\mathbf{z}_i(t) = \begin{bmatrix}\mathbf{x}_i(t+1)^T & \mathbf{x}_i(t)^T & \mathbf{w}_i(t)^T\end{bmatrix}^T$$

which has the mean and covariance

$$\begin{aligned}\boldsymbol{\mu}_z(t) &= \begin{bmatrix}\boldsymbol{\mu}_x(t+1)^T & \boldsymbol{\mu}_x(t)^T & \mathbf{0}_m^T\end{bmatrix}^T,\\ \boldsymbol{\Sigma}_{zz}(t) &= \begin{bmatrix}\boldsymbol{\Sigma}_{xx}(t+1) & \boldsymbol{\Sigma}_{x^+x}(t) & \boldsymbol{\Sigma}_{x^+w}(t)\\ \boldsymbol{\Sigma}_{x^+x}(t)^T & \boldsymbol{\Sigma}_{xx}(t) & \mathbf{0}_{n\times m}\\ \boldsymbol{\Sigma}_{x^+w}(t)^T & \mathbf{0}_{m\times n} & \boldsymbol{\Sigma}_{ww}(t)\end{bmatrix}.\end{aligned}$$

Both of $\|\boldsymbol{\mu}_z(t)\|$ and $\|\boldsymbol{\Sigma}_{zz}(t)\|$ are uniformly bounded by Assumptions (A1) and (A2). Therefore, by the multivariate CLT [25, Thm 3.9.6], the population average of $\mathbf{z}_i(t)$, i.e.,

$$\begin{aligned}\bar{\mathbf{z}}(t) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\mathbf{z}_i(t) - \boldsymbol{\mu}_z(t))\\ &= \begin{bmatrix}\bar{\mathbf{x}}(t+1)^T & \bar{\mathbf{x}}(t)^T & \bar{\mathbf{w}}(t)^T\end{bmatrix}^T\end{aligned}$$

converges in distribution to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{zz}(t))$ as $N \to \infty$ for each t. Therefore, any integral with respect to this sequence of distributions, such as the conditional expectation in (5) [25, Ex. 5.1.4], will also converge to the same integral with respect to the limit distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{zz}(t))$ [30].

Therefore, for any t, let $\mathbf{z}^*(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{zz}(t))$ and partition it as

$$\mathbf{z}^*(t) = \begin{bmatrix}\mathbf{z}_1^*(t)\\ \mathbf{z}_2^*(t)\end{bmatrix}$$

where $\mathbf{z}_1^*(t) \in \mathbb{R}^n$ and $\mathbf{z}_2^*(t) \in \mathbb{R}^{n+m}$. The covariance matrix $\boldsymbol{\Sigma}_{zz}(t)$ can also be partitioned accordingly, i.e.,

$$\begin{aligned}\boldsymbol{\Sigma}_{zz}(t) &= \left[\begin{array}{c|cc}\boldsymbol{\Sigma}_{xx}(t+1) & \boldsymbol{\Sigma}_{x^+x}(t) & \boldsymbol{\Sigma}_{x^+w}(t)\\ \hline \boldsymbol{\Sigma}_{x^+x}(t)^T & \boldsymbol{\Sigma}_{xx}(t) & \mathbf{0}\\ \boldsymbol{\Sigma}_{x^+w}(t)^T & \mathbf{0} & \boldsymbol{\Sigma}_{ww}(t)\end{array}\right]\\ &\triangleq \begin{bmatrix}\boldsymbol{\Sigma}_{11}(t) & \boldsymbol{\Sigma}_{12}(t)\\ \boldsymbol{\Sigma}_{21}(t) & \boldsymbol{\Sigma}_{22}(t)\end{bmatrix}\end{aligned}$$

Then, the conditional distribution of $\mathbf{z}_1^*$ given $\mathbf{z}_2^*$ also has a multivariate normal distribution [26] with mean

$$\begin{aligned}\mathbb{E}\left[\mathbf{z}_1^*(t) \middle| \mathbf{z}_2^*(t) = \begin{bmatrix}\boldsymbol{\xi}\\ \boldsymbol{\omega}\end{bmatrix}\right] &= \boldsymbol{\Sigma}_{12}(t)\boldsymbol{\Sigma}_{22}(t)^{-1}\begin{bmatrix}\boldsymbol{\xi}\\ \boldsymbol{\omega}\end{bmatrix}\\ &= \mathbf{A}(t)\boldsymbol{\xi} + \mathbf{B}(t)\boldsymbol{\omega},\end{aligned}$$

completing the proof. □

As with the subtle distinction between nonlinear and linear time-varying (LTV) systems in general, it is important to note that the matrices $\mathbf{A}(t)$ and $\mathbf{B}(t)$ in Theorem IV.2 depend on the *distribution* of $\mathbf{x}_i(t)$ but not on $\mathbf{x}_i(t)$ themselves. In other words, for any choice of model in (1) (i.e., for any choice of $f(\cdot)$, p_w, and p_0), the matrices $\mathbf{A}(t)$ and $\mathbf{B}(t)$ can be pre-computed for all $t \geq 0$ and would be the same for all realizations of (the noise and therefore) $\mathbf{x}_i(t)$. As such, the asymptotic dynamics in (5) are indeed linear, hence the significance of the theorem.

Example IV.3. ***(LTV vs. Nonlinear system).*** *Consider dynamics of a system*

$$y(t+1) = a(t)y(t), \quad y(0) \sim \mathcal{N}(e, 1)$$

and three cases, as follows.

(i) $a(t) = e^{2^t}$. Here the system is clearly LTV.

(ii) $a(t) = e^{y(t)}$. Here the system is clearly nonlinear.

(iii) $a(t) = \mathbb{E}[y(t)]$. Although the system here may look nonlinear, the coefficient $\mathbb{E}[y(t)]$ can pre-computed for all $t \geq 0$ given the distribution of $y(0)$. It is straightforward to see that $a(t) = e^{2^t}$, making the system in (iii) equivalent to that of (i) and similarly LTV. □

Interestingly, if the dynamics in (1) satisfy additional assumptions, the average population dynamics tend not only to a linear system but further to an LTI one.

Assumption 2. ***(Lipschitzness).*** *To prove the convergence of the solutions of* (1) *to stationarity, we need to make additional assumptions, including the following.*

(A5) *The function $f(\cdot)$ is globally Lipschitz in $\mathbf{x}$, i.e., for any $\mathbf{w} \in \mathbb{R}^m$ there exists $L(\mathbf{w}) \geq 0$ such that*

$$\|f(\mathbf{x}_1, \mathbf{w}) - f(\mathbf{x}_2, \mathbf{w})\| \leq L(\mathbf{w})\|\mathbf{x}_1 - \mathbf{x}_2\|, \tag{7}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$. □

Assumption (A5), while being restrictive on the space of all functions, becomes mild when considering only bounded functions as required by assumption (A1). Any nonlinear polynomial, e.g., is not globally Lipschitz but becomes so when augmented with a saturation as in $\max\{\min\{f_i, M\}, -M\}$ suggested earlier. Indeed, bounded functions exist that are not globally Lipschitz, such as $\cos(x^2)$, but they are often contrived examples with little use in the modeling of real-world systems.

The next result builds on Assumptions (A1)-(A5), as well as Assumptions (A6)-(A7) stated therein, to prove that all solutions of (1) converge to a stationary process.

Theorem IV.4. ***(Steady-state convergence to stationarity).*** *Consider the population dynamics in* (1) *and assume that assumptions* (A1)-(A5) *hold. Then, all subsystems have a stationary solution*

$$\mathbf{x}^*(t) \sim p^*, \qquad t \geq 0.$$

Further, all $\mathbf{x}_i(t)$ converge to $\mathbf{x}^*(t)$ in distribution as $t \to \infty$ if

(A6) *the initial distribution p_0 is such that $p_0(A) = 0$ for any set A where $p^*(A) = 0$;*

(A7) *for any set A with $p^*(A) > 0$, the noise distribution p_w is such that for all $\mathbf{x} \in \mathbb{R}^n$, there exists a non-zero probability that $f(\mathbf{x}, \mathbf{w}) \in A$.*

Proof. Throughout the proof, we drop the index i since it is arbitrary and all subsystems have the same dynamics (so $\mathbf{x}(t)$ in the following is still n-dimensional, *not* a concatenation of all $\mathbf{x}_i(t)$).

To prove the existence of a stationary solution, we will follow the exposition in [27, Thm 2.4] for continuous-time systems. Since the solution to (1) is fully determined by its initial condition and noise (input) process, it is sufficient to prove that

$$\begin{aligned}&\mathbb{P}\{\mathbf{x}(t) \in A_0, \mathbf{w}(t+s_1) \in A_1, \ldots, \mathbf{w}(t+s_m) \in A_m\}\\ &\quad= \mathbb{P}\{\mathbf{x}_0 \in A_0, \mathbf{w}(s_1) \in A_1, \ldots, \mathbf{w}(s_m) \in A_m\}\end{aligned}$$

for all t, m, $s_1, \ldots, s_m$, and Borel sets $A_0, A_1, \ldots, A_m$, where $\mathbf{x}(t)$ is the solution starting from $\mathbf{x}(0) = \mathbf{x}_0$. Consider an arbitrary initial distribution $\bar{p}_0$ and let $\bar{\mathbf{x}}(t)$ be the solution of (1) starting from $\bar{\mathbf{x}}_0 \sim \bar{p}_0$. Let τ_k be a discrete random variable uniformly distributed in $[0, k]$ independent of $\mathbf{w}(t)$ and $\bar{\mathbf{x}}_0$ and define

$$\mathbf{x}^{(k)}(t) = \bar{\mathbf{x}}(t + \tau_k), \qquad \mathbf{x}_0^{(k)} = \bar{\mathbf{x}}(\tau_k), \tag{8a}$$

$$\mathbf{w}^{(k)}(t) = \mathbf{w}(t + \tau_k). \tag{8b}$$

By the total probability law [31, Ch. 2],

$$\begin{aligned}&\mathbb{P}\{\mathbf{x}^{(k)}(t) \in A_0, \mathbf{w}^{(k)}(s_1) \in A_1, \ldots, \mathbf{w}^{(k)}(s_m) \in A_m\} = \\ &\frac{1}{k}\sum_{s=0}^{k} \mathbb{P}\{\bar{\mathbf{x}}(t+s) \in A_0, \mathbf{w}(s_1+s) \in A_1, \ldots, \mathbf{w}(s_m+s) \in A_m\}.\end{aligned} \tag{9}$$

In particular, for all k and t,

$$\mathbb{P}\{\|\mathbf{x}_0^{(k)}\| > R\} = \frac{1}{k}\sum_{s=0}^{k} \mathbb{P}\{\|\bar{\mathbf{x}}(s)\| > R\} = 0, \quad \forall R > M,$$

$$\mathbb{P}\{\|\mathbf{w}^{(k)}(t)\| > R\} = \frac{1}{k}\sum_{s=0}^{k} \mathbb{P}\{\|\mathbf{w}(s)\| > R\} \to 0 \text{ as } R \to \infty,$$

where we used Assumption (A1) in the former. These, together with the stationarity (due to whiteness) of $\mathbf{w}(t)$ ensure that the sequence of random processes $(\mathbf{x}_0^{(k)}, \mathbf{w}^{(k)}(t))$ satisfy the conditions of Lemma A.1. Let $(\tilde{\mathbf{x}}_0^{(n_k)}, \tilde{\mathbf{w}}^{(n_k)}(t))$ and $(\tilde{\mathbf{x}}_0, \tilde{\mathbf{w}}(t))$ be the subsequence and limit process provided by Lemma A.1. By the stationarity of $\mathbf{w}(t)$, (8b), and Lemma A.1, $\mathbf{w}(t)$ and $\tilde{\mathbf{w}}(t)$ have the same distribution. Let $\hat{\mathbf{x}}_0^{(n_k)}$ and $\mathbf{x}_0^*$ be random variables in the original probability space that have the same joint distribution with $\mathbf{w}(t)$ as the joint distribution of $\tilde{\mathbf{x}}_0^{(n_k)}$, $\tilde{\mathbf{x}}_0$, and $\tilde{\mathbf{w}}(t)$. They can always be constructed, e.g., by inverse transform sampling [32]. Define $\hat{\mathbf{x}}^{(n_k)}(t)$ and $\mathbf{x}^*(t)$ to be the solutions of (1) with the initial conditions $\mathbf{x}_0^{(n_k)}$ and $\mathbf{x}_0^*$, respectively. From (7),

$$\|\hat{\mathbf{x}}^{(n_k)}(t) - \mathbf{x}^*(t)\| < \Pi_{\tau=0}^{t-1} L(\mathbf{w}(\tau)) \|\hat{\mathbf{x}}_0^{(n_k)} - \mathbf{x}_0^*\|.$$

Therefore, by Lemma A.1, $\hat{\mathbf{x}}^{(n_k)}(t) \to \mathbf{x}^*(t)$ in probability for every t. Now, let $h(.)$ be any bounded continuous function. Then from [27, Thm 2.2], we can write,

$$\begin{aligned}
&\mathbb{E}[h(\mathbf{x}^*(t), \mathbf{w}(t+s_1), \ldots, \mathbf{w}(t+s_m))]\\
&\overset{(a)}{=} \lim_{k\to\infty} \mathbb{E}[h(\hat{\mathbf{x}}^{(n_k)}(t), \mathbf{w}(t+s_1), \ldots, \mathbf{w}(t+s_m))]\\
&\overset{(b)}{=} \lim_{k\to\infty} \mathbb{E}[h(\mathbf{x}^{(n_k)}(t), \mathbf{w}^{(n_k)}(t+s_1), \ldots, \mathbf{w}^{(n_k)}(t+s_m))]\\
&\overset{(c)}{=} \lim_{k\to\infty} \frac{1}{n_k}\sum_{u=0}^{n_k} \mathbb{E}[h(\bar{\mathbf{x}}(t+u), \mathbf{w}(t+s_1+u), \ldots, \mathbf{w}(t+s_m+u))]\\
&\overset{(d)}{=} \lim_{k\to\infty} \frac{1}{n_k}\sum_{s=0}^{n_k} \mathbb{E}[h(\bar{\mathbf{x}}(s), \mathbf{w}(s+s_1), \ldots, \mathbf{w}(s+s_m))]\\
&\overset{(e)}{=} \lim_{k\to\infty} \mathbb{E}[h(\mathbf{x}_0^{(n_k)}, \mathbf{w}^{(n_k)}(s_1), \ldots, \mathbf{w}^{(n_k)}(s_m))]\\
&\overset{(f)}{=} \lim_{k\to\infty} \mathbb{E}[h(\tilde{\mathbf{x}}_0^{(n_k)}, \mathbf{w}^{(n_k)}(s_1), \ldots, \mathbf{w}^{(n_k)}(s_m))]\\
&\overset{(g)}{=} \mathbb{E}[h(\tilde{\mathbf{x}}_0, \mathbf{w}(s_1), \ldots, \mathbf{w}(s_m))]\\
&\overset{(h)}{=} \mathbb{E}[h(\mathbf{x}_0^*, \mathbf{w}(s_1), \ldots, \mathbf{w}(s_m))],
\end{aligned} \tag{10}$$

where (a) follows from the fact that $\hat{\mathbf{x}}^{(n_k)}(t) \to \mathbf{x}^*(t)$ in probability, (b) follows from the stationarity of $\mathbf{w}(t)$, (c, e) follow from the law of total expectation [25], (d) follows from a change of variables $s = t+u$ and the facts that $n_k \to \infty$ as $k \to \infty$ while $h(.)$ and t are bounded, (f, g) follow from Lemma A.1, and (h) follows from the construction of $\mathbf{x}_0^*$. The claim of the theorem then follows from (10) and the fact that h is an arbitrary bounded continuous function [33].

To prove the convergence of all $\mathbf{x}(t)$ to $\mathbf{x}^*(t)$ under (A6) and (A7), note that (1) defines a continuous-state Markov chain. Let P denote its transition kernel [28]. By (A7), there is a non-zero probability that the state moves from an arbitrary state to any A with $p^*(A) > 0$ in an arbitrary number of steps, ensuring that P is strongly p^*-irreducible [28]. Therefore, according to [28, Theorem 1], there exists $U \subseteq \mathbb{R}^n$ such that $p^*(\mathbb{R}^n \setminus U) = 0$ and for any initial distribution on U, $\mathbf{x}(t) \to \mathbf{x}^*(t)$ in distribution. By (A6), p_0 defines a distribution on U, hence the claim of the theorem. □

A note is warranted on the assumptions (A6) and (A7). Let $R \subset \mathbb{R}^n$ be the (bounded) range of f. Given the continuous nature of $\mathbf{x}_i(t)$ and the often continuous distribution of noise in real-world systems, it is expected for p^* to also have a continuous distribution and therefore a density function on R where $p^*(A) = 0$ if and only if A has Lebesgue measure zero. If so, then any continuous initial distribution p_0 with a density function on R satisfies (A6). Also, assumption (A7) would then only require the possibility of transitioning from any $\mathbf{x}$ to a set A of positive Lebesgue measure with positive probability. This can be satisfied by various forms of f and distributions p_w, including systems with additive noise of the form $f(\mathbf{x}, \mathbf{w}) = f_1(f_2(\mathbf{x}) + \mathbf{w})$ and noise distributions that have a sufficiently large support over $\mathbb{R}^m$.

Combining the linearity of Theorem IV.2 and the stationarity of Theorem IV.4 ensures the convergence of the average population dynamics to an LTI system, as formalized next.

Theorem IV.5. ***(LTI average population dynamics).*** *Consider the population dynamics* (1) *and assume that assumptions* (A1)-(A7) *hold. Then,*

$$\mathbb{E}\left[\bar{\mathbf{x}}(t+1) \middle| \begin{bmatrix} \bar{\mathbf{x}}(t) \\ \bar{\mathbf{w}}(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\upsilon}_0 \\ \boldsymbol{\zeta}_0 \end{bmatrix}\right] \to \mathbf{A}\boldsymbol{\upsilon}_0 + \mathbf{B}\boldsymbol{\zeta}_0$$

as $N, t \to \infty$, *where*

$$\mathbf{A} = \lim_{t\to\infty} \mathbf{A}(t), \quad \mathbf{B} = \lim_{t\to\infty} \mathbf{B}(t), \tag{11}$$

and $\mathbf{A}(t)$ *and* $\mathbf{B}(t)$ *are defined in* (6).

Proof. The result follows primarily from Theorems IV.2 and IV.4. All that remains to be proven is that the limits in (11) exist. For this, it is sufficient to show that $\mathbf{\Sigma}_{x^+x}(t)$, $\mathbf{\Sigma}_{x^+w}(t)$, $\mathbf{\Sigma}_{xx}(t)$, and $\mathbf{\Sigma}_{ww}(t)$ all converge as $t \to \infty$. The latter holds trivially as $\mathbf{w}(t)$ is i.i.d. across time. From Theorem IV.4, we have that the sequence of distributions of $\mathbf{x}(t)$ converges to p^* as $t \to \infty$. Hence, any integral with respect to this sequence of distributions, including $\mathbf{\Sigma}_{xx}(t)$, also converges to the same integral with respect to p^* [30].

To prove the convergence of $\mathbf{\Sigma}_{x^+x}(t)$ and $\mathbf{\Sigma}_{x^+w}(t)$, the same argument we used for $\mathbf{\Sigma}_{xx}(t)$ applies provided that we can show the convergence of the joint distributions of $(\mathbf{x}(t+1), \mathbf{x}(t))$ and $(\mathbf{x}(t+1), \mathbf{w}(t))$ to respective limits. Note that (dropping the subindex i as in the proof of Theorem IV.4),

$$\begin{aligned} F_{x^+xw}(\boldsymbol{\xi}^+, \boldsymbol{\xi}, \boldsymbol{\omega}, t) &= \mathbb{P}\{\mathbf{x}(t+1) \le \boldsymbol{\xi}^+, \mathbf{x}(t) \le \boldsymbol{\xi}, \mathbf{w}(t) \le \boldsymbol{\omega}\} \\ &= \int_{S_{\boldsymbol{\xi}^+,\boldsymbol{\xi},\boldsymbol{\omega}}} dp_{xw}(t) \end{aligned}$$

where

$$S_{\boldsymbol{\xi}^+,\boldsymbol{\xi},\boldsymbol{\omega}} = \{(\mathbf{x}, \mathbf{w}) | f(\mathbf{x}, \mathbf{w}) \le \boldsymbol{\xi}^+, \mathbf{x} \le \boldsymbol{\xi}, \mathbf{w} \le \boldsymbol{\omega}\}$$

is independent of t and $p_{xw}(t)$ is the joint distribution of $\mathbf{x}(t)$ and $\mathbf{w}(t)$. By Assumption (A3), the latter is equal to the product of the marginals $p_x(t)p_w(t)$, which converges to a stationary distribution by Assumption (A2) and Theorem IV.4. Therefore, the sequence of joint distributions $p_{x^+xw}(t)$ and any integral with respect to it, including $\mathbf{\Sigma}_{x^+x}(t)$ and $\mathbf{\Sigma}_{x^+w}(t)$, converge to a stationary limit as $t \to \infty$. □

V. Conclusions and Future Work

In this work we provided a theoretical framework for understanding the linearity of spatially averaged dynamics of populations of dynamical subsystems, inspired by the observed linearity of macroscopic biological and artificial neural dynamics. Our results apply to a broad range of nonlinear dynamics and were presented separately for populations of static (feedforward) and dynamic (recurrent) nonlinear subsystems. In the latter, we further distinguished between the transient and steady-state dynamics of the average state variables and proved that they converge to an LTV and LTI system, respectively. Despite their generality, our results are still limited in their need for independence between the individual subsystems, thus limiting the application of our results to networked systems. Future work will therefore focus on extending our framework to networked systems with correlated and potentially heterogeneous subsystems as well as data-driven validations of our results and their extensions to the dual linearizing effect of temporal averaging (low pass filtering) on nonlinear dynamics.

References

[1] M. Breakspear, "Dynamic models of large-scale brain activity," *Nature neuroscience*, vol. 20, no. 3, pp. 340–352, 2017.

[2] W. Burghout, H. N. Koutsopoulos, and I. Andreasson, "Hybrid mesoscopic–microscopic traffic simulation," *Transportation Research Record*, vol. 1934, no. 1, pp. 218–225, 2005.

[3] M. Marchiori and L. Possamai, "Micro-macro analysis of complex networks," *PloS one*, vol. 10, no. 1, p. e0116670, 2015.

[4] I. Lampropoulos, G. M. Vanalme, and W. L. Kling, "A methodology for modeling the behavior of electricity prosumers within the smart grid," in *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*. IEEE, 2010, pp. 1–8.

[5] M. O. Jackson, B. W. Rogers, and Y. Zenou, "The economic consequences of social-network structure," *Journal of Economic Literature*, vol. 55, no. 1, pp. 49–95, 2017.

[6] A. Banos, N. Corson, B. Gaudou, V. Laperrière, and S. R. Coyrehourcq, "Coupling micro and macro dynamics models on networks: Application to disease spread," in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. Springer, 2015, pp. 19–33.

[7] C. Zang, P. Cui, and C. Faloutsos, "Beyond sigmoids: The nettide model for social network growth, and its applications," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 2015–2024.

[8] F. Milano and R. Zárate-Miñano, "A systematic method to model power systems as stochastic differential algebraic equations," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4537–4544, 2013.

[9] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy, "Neural population dynamics during reaching," *Nature*, vol. 487, no. 7405, pp. 51–56, 2012.

[10] A. A. Ivanova, M. Schrimpf, S. Anzellotti, N. Zaslavsky, E. Fedorenko, and L. Isik, “Is it that simple? linear mapping models in cognitive neuroscience,” *bioRxiv*, 2021.

[11] U. A. Ernst and C. W. Eurich, “Cortical population dynamics and psychophysics,” *Handbook of brain theory and neural networks*, pp. 294–300, 2002.

[12] M. J. Keeling and J. V. Ross, “On methods for studying stochastic disease dynamics,” *Journal of the Royal Society Interface*, vol. 5, no. 19, pp. 171–181, 2008.

[13] E. Nozari, M. A. Bertolero, J. Stiso, L. Caciagli, E. J. Cornblath, X. He, A. S. Mahadevan, G. J. Pappas, and D. S. Bassett, “Is the brain macroscopically linear? a system identification of resting state dynamics,” *Nature Biomedical Engineering*, 2022, under Review.

[14] V. Booth and J. Rinzel, “A minimal, compartmental model for a dendritic origin of bistability of motoneuron firing patterns,” *Journal of computational neuroscience*, vol. 2, no. 4, pp. 299–312, 1995.

[15] H. R. Wilson and J. D. Cowan, “Excitatory and inhibitory interactions in localized populations of model neurons,” *Biophysical journal*, vol. 12, no. 1, pp. 1–24, 1972.

[16] Y. Yang, S. Qiao, O. G. Sani, J. I. Sedillo, B. Ferrentino, B. Pesaran, and M. M. Shanechi, “Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation,” *Nature biomedical engineering*, vol. 5, no. 4, pp. 324–345, 2021.

[17] O. G. Sani, H. Abbaspourazad, Y. T. Wong, B. Pesaran, and M. M. Shanechi, “Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification,” *Nature Neuroscience*, vol. 24, no. 1, pp. 140–149, 2021.

[18] M.-A. Schulz, T. Yeo, J. Vogelstein, J. Mourao-Miranada, J. Kather, K. Kording, B. A. Richards, and D. Bzdok, “Deep learning for brains?: Different linear and nonlinear scaling in uk biobank brain images vs. machine-learning datasets,” *BioRxiv*, p. 757054, 2019.

[19] Z. Liu, C. Rios, N. Zhang, L. Yang, W. Chen, and B. He, “Linear and nonlinear relationships between visual stimuli, eeg and bold fmri signals,” *Neuroimage*, vol. 50, no. 3, pp. 1054–1066, 2010.

[20] T. Luo, Z.-Q. J. Xu, Z. Ma, and Y. Zhang, “Phase diagram for two-layer relu neural networks at infinite-width limit,” *Journal of Machine Learning Research*, vol. 22, no. 71, pp. 1–47, 2021.

[21] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, “Wide neural networks of any depth evolve as linear models under gradient descent,” *Advances in neural information processing systems*, vol. 32, 2019.

[22] C. Liu, L. Zhu, and M. Belkin, “On the linearity of large non-linear models: when and why the tangent kernel is constant,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 954–15 964, 2020.

[23] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” *Advances in neural information processing systems*, vol. 31, 2018.

[24] S. Ching, Y. Eun, P. T. Kabamba, and S. M. Meerkov, “Quasilinear control theory: An overview,” *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 10 958–10 963, 2011.

[25] R. Durrett, *Probability: theory and examples*, 4.1 ed. Cambridge university press, April 2013, vol. 49.

[26] J. Jensen, L. Lake, P. Corbett, and D. Goggin, *Statistics for Petroleum Engineers and Geoscientists*. Prentice Hall, 1997.

[27] R. Khasminskii, *Stochastic stability of differential equations*, ser. Stochastic Modelling and Applied Probability. Berlin, Germany: Springer, Nov. 2013.

[28] P. Diaconis and D. Freedman, “On markov chains with continuous state space,” *Technical Report no. 501*, 1997.

[29] M. DeGroot and M. Schervish, *Probability and Statistics*. Addison-Wesley, 2012.

[30] H. Royden and P. Fitzpatrick, *Real Analysis*, 4th ed. Pearson, 2017.

[31] A. Papoulis and S. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. McGraw-Hill, 2002.

[32] L. Devroye, “Sample-based non-uniform random variate generation,” in *Proceedings of the 18th conference on Winter simulation*, 1986, pp. 260–265.

[33] Y. V. Prokhorov, “Convergence of random processes and limit theorems in probability theory,” *Theory of Probability & Its Applications*, vol. 1, no. 2, pp. 157–214, 1956. [Online]. Available: https://doi.org/10.1137/1101016

Appendix A. Auxiliary Results

Lemma A.1. ***(Existence of sequence with weakly convergent finite-dimensional distributions in different probability space).*** *Consider a sequence of stochastic processes $\mathbf{v}^{(k)}(t), k \geq 1, t \geq 0$ and assume that it is uniformly bounded in probability, i.e.,*

$$\lim_{R\to\infty} \sup_{t,k} \mathbb{P}\{\|\mathbf{v}^{(k)}(t)\| > R\} = 0.$$

Then, there exists a subsequence $\mathbf{v}^{(n_k)}(t)$ of this sequence as well as a (sub)sequence of stochastic processes $\tilde{\mathbf{v}}^{(n_k)}(t)$ and a stochastic process $\tilde{\mathbf{v}}(t)$ in a (potentially different) probability space such that for all t, $\tilde{\mathbf{v}}^{(n_k)}(t) \xrightarrow{k\to\infty} \tilde{\mathbf{v}}(t)$ in probability and, for all k, $\mathbf{v}^{(n_k)}(t)$ and $\tilde{\mathbf{v}}^{(n_k)}(t)$ have the same finite-dimensional joint distributions across time.

Proof. For all k, we extend the discrete-time stochastic process $\mathbf{v}^{(k)}(t)$ to a continuous-time process $\mathbf{v}_c^{(k)}(t)$ by linear interpolation, i.e., for any $t \in \mathbb{R}$, $t \geq 0$,

$$\mathbf{v}_c^{(k)}(t) = \mathbf{v}^{(k)}(\lfloor t\rfloor) + \frac{\mathbf{v}^{(k)}(\lceil t\rceil) - \mathbf{v}^{(k)}(\lfloor t\rfloor)}{\lceil t\rceil - \lfloor t\rfloor}(t - \lfloor t\rfloor).$$

By construction, $\mathbf{v}_c^{(k)}(t)$ is uniformly stochastically continuous, i.e.,

$$\sup_{k,|s_1-s_2|<h} \mathbb{P}\{|\mathbf{v}_c^{(k)}(s_1) - \mathbf{v}_c^{(k)}(s_2)| > \epsilon\} \to 0,$$

as $h \to 0$. Therefore, $\mathbf{v}_c^{(k)}(t)$ satisfies the conditions of [27, Thm 2.1, 2.2] and, as such, there exists a subsequence $\mathbf{v}_c^{(n_k)}(t)$ of it as well as a (sub)sequence of stochastic processes $\tilde{\mathbf{v}}_c^{(n_k)}(t)$ and a stochastic process $\tilde{\mathbf{v}}_c(t)$ in a potentially different probability space such that for all t, $\tilde{\mathbf{v}}_c^{(n_k)}(t) \xrightarrow{k\to\infty} \tilde{\mathbf{v}}_c(t)$ in probability and, for all k, $\mathbf{v}_c^{(n_k)}(t)$ and $\tilde{\mathbf{v}}_c^{(n_k)}(t)$ have the same finite-dimensional joint distributions. The claim of the theorem then follows from sampling $\tilde{\mathbf{v}}_c^{(n_k)}(t)$ and $\tilde{\mathbf{v}}_c(t)$ to obtain $\tilde{\mathbf{v}}^{(n_k)}(t)$ and $\tilde{\mathbf{v}}(t)$, respectively. $\square$