Synchronization and Redundancy: Implications for Robustness of Neural Learning and Decision Making

Jake Bouvrie
jvb@math.duke.edu
Department of Mathematics, Duke University, Durham, NC 27708, U.S.A.

Jean-Jacques Slotine
jjs@mit.edu
Nonlinear Systems Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Learning and decision making in the brain are key processes critical to survival, and yet are processes implemented by nonideal biological building blocks that can impose significant error. We explore quantitatively how the brain might cope with this inherent source of error by taking advantage of two ubiquitous mechanisms, redundancy and synchronization. In particular we consider a neural process whose goal is to learn a decision function by implementing a nonlinear gradient dynamics. The dynamics, however, are assumed to be corrupted by perturbations modeling the error, which might be incurred due to limitations of the biology, intrinsic neuronal noise, and imperfect measurements. We show that error, and the associated uncertainty surrounding a learned solution, can be controlled in large part by trading off synchronization strength among multiple redundant neural systems against the noise amplitude. The impact of the coupling between such redundant systems is quantified by the spectrum of the network Laplacian, and we discuss the role of network topology in synchronization and in reducing the effect of noise. We discuss range of situations in which the mechanisms we model arise in brain science and draw attention to experimental evidence suggesting that cortical circuits capable of implementing the computations of interest here can be found on several scales. Finally, simulations comparing theoretical bounds to the relevant empirical quantities show that the theoretical estimates we derive can be tight.

1 Introduction

Learning and decision making in the brain are key processes critical to survival, and yet are processes implemented by imperfect biological building blocks that can impose significant error. We suggest that the brain can cope with this inherent source of error by taking advantage of two ubiquitous

mechanisms: redundancy and sharing of information. These concepts will be made precise in the context of a specific model and learning scenario that together can serve as a conceptual tool for illustrating the effect of redundancy and sharing.

Motivated by the problem of learning to discriminate, we consider a neural process whose goal is to learn a decision function by implementing a nonlinear gradient dynamics. The dynamics, however, are assumed to be corrupted by perturbations modeling the error that might be incurred. This general perspective is intended to capture a range of possible learning instances occurring at different anatomical scales. The neural process can involve whole brain areas communicating by behavioral, motor, or sensory pathways (Schnitzler & Gross, 2005), as in the case of the multiple amygdala-thalamus loops assumed to underpin fear conditioning, for instance (LeDoux, 2000; Maren, 2001). Interacting local field potentials (LFPs) may also be modeled as both direct—by long-range phase locking, for example, in olfactory systems (Friedrich, Habermann, & Laurent, 2004)—or indirect measurements of coordination and interaction among large assemblies of neurons. The learning dynamics may alternatively model smaller ensembles of individual neurons, as in primary motor cortex, though we do not emphasize biological realism in our models at this scale. Nevertheless, one may still draw useful conclusions as to the role of redundancy and information sharing. The error too may be treated at different scales and may take the form of noise intrinsic to the neural environment (Faisal, Selen, & Wolpert, 2008) on a large, aggregate scale (e.g., in the case of LFPs) or on a small scale involving localized populations of neurons.

If there is noise corrupting the learning process, an immediate question is whether it is possible to gauge the accuracy of the predictions of the learned function and to what extent the organism can reduce uncertainty in its decisions by taking advantage of a simple, common information-sharing mechanism. If there is redundancy in the form of multiple independent copies of the dynamical circuit (Adams, 1998; Fernando, Goldstein, & Szathmary, 2010), it is reasonable to expect that averaging over the different solutions might reduce noise via cancellation effects. In the case of learning in the brain, however, this approach is problematic because neurons are susceptible to saturation of their firing rates, and on large scales aggregate signal amplitudes will also saturate; the macroscopic dynamics that neuron populations and assemblies obey can be strongly nonlinear. When the dynamics followed by different dynamical systems are nonlinear, one cannot expect to gain a meaningful signal by linear averaging (see Tabareau, Slotine, & Pham, 2010). As a simple illustration of this phenomenon, consider a collection of noisy sinusoidal oscillators allowed to run starting from different initial conditions, with identical frequencies and independent noise terms. The oscillators will be out of phase from each other, so an average over the trajectories will not yield anything close to a clean version of a sinusoid at the desired frequency. On the other hand, it is

reasonable to suppose that synchronization across neuron populations or between macroscale cortical loops may provide sufficient phase alignment to make linear averaging, and thus "consensus," a powerful proposal for reducing the effects of noise (Tabareau et al., 2010; Masuda, Kawamura, & Kori, 2010; Cao, Stewart, & Leonard, 2010; Young, Scardovi, & Leonard, 2010; Poulakakis, Scardovi, & Leonard, 2010; Gigante, Mattia, Braun, & Del Giudice, 2009). Indeed, synchrony within a system of coupled dynamical elements provides (quantifiable) robustness to perturbations occurring in any one element's dynamics (Needleman, Tiesinga, & Sejnowski, 2001; Wang & Slotine, 2005; Pham & Slotine, 2007).

We will place much emphasis on exploring quantitatively the role of synchrony in controlling uncertainty arising from noise modeling neural error. In particular, we base our work on the argument that noisy, nonlinear trajectories can be linearly averaged if fluctuations due to noise can be made small and that fluctuations can be made small by coupling the dynamical elements appropriately. In the stochastic setting adopted here, synchronization refers to state synchrony: the tendency for individual elements' trajectories to move toward a common trajectory in a quantifiable sense. The estimates we present directly characterize the trade-off between the network's tendency toward synchrony and the noise, and ultimately address the specific role this trade-off plays in determining uncertainty surrounding a function learned by an imperfect learning system.

We further show how and where the topology of the network of neural ensembles affect the extent to which the noise, and therefore uncertainty, can be controlled. The estimates we provide take into account in a fundamental way both the nonlinearity in the dynamics and the noise. More generally, the work discussed here also has implications in other related domains, such as networks of coupled learners or adaptive sensor networks, and it can be extended to multitask online or dynamic learning settings. The difficulty inherent in analyzing dynamic learning systems, such as hierarchies with feedback (Mumford, 1992; Lee & Mumford, 2003), poses a challenge. But considering dynamic systems can yield substantial benefits: transients can be important, as suggested by the literature on regularization paths and early stopping (Yao, Rosasco, & Caponnetto, 2007). Furthermore, the role of feedback and backprojections and attention-like mechanisms in learning and recognition systems, both biological and artificial, is known to be important but is not well understood (Hahnloser, Douglas, Mahovald, & Hepp, 1999; Itti & Koch, 2001; Hung, Kreiman, Poggio, & DiCarlo, 2005).

The letter is organized as follows. In section 2 we consider a specific learning problem and define a system of stochastic differential equations (SDEs) modeling a simple dynamic learning process. We then discuss stability and network topology in the context of synchronization. In section 3 we present the main theoretical results, a set of uncertainty estimates, postponing proofs until later. Then in section 5, we provide simulations and compare empirical estimates to the theoretical quantities predicted by the

theorems in section 3. Section 4 provides a discussion addressing the significance and applicability of our theoretical contributions to neuroscience and behavior. Finally, in section 6, we give proofs of the results stated in section 3.

2 Biological Learning as a Stochastic Network Model

The learning process we will model is that of a one-dimensional linear fitting problem described by gradient-based minimization of a square loss objective, in the spirit of Rao and Ballard (1999). This is perhaps the simplest and most fundamental abstract learning problem that an organism might be confronted with—that of using experiential evidence to infer correlations and ultimately discover causal relationships that govern the environment and can be used to make predictions about the future. The model realizing this learning process is also simple, in that we capture neural communication as an abstract process "in which a neural element (a single neuron or a population of neurons) conveys certain aspects of its functional state to another neural element" (Schnitzler & Gross, 2005). In doing so, we focus on the underlying computations taking place in the nervous system rather than dwell on neural representations. Even this simple setting becomes involved technically and is rich enough to explore all of the key themes. Our model also supports nonlinear decision functions in the sense that we might consider taking a linear function of nonlinear variables whose values might be computed upstream. In this case, the development would be similar but extended to the multidimensional case. The model may also be extended to richer function classes and more exotic loss functions directly; however, for our purposes, the additional generality does not yield significant further insight and might raise biological plausibility concerns.[1]

2.1 Problem Setup. To make the setting more concrete, we begin by assuming that we have observed a set of input-output examples $\{x_i \in \mathbb{R}, y_i \in \mathbb{R}\}_{i=1}^m$, each representing a generic unit of sensory experience, and want to estimate the linear regression function $f_w(x) = wx$. When the square loss is adopted, the total error incurred on the observations by f_w is given by the familiar expression

$$E(w) = \sum_{i=1}^{m}(y_i - f_w(x_i))^2 = \sum_{i=1}^{m}(y_i - wx_i)^2.$$

We will model adaptation (training) by a noisy gradient descent process, with biologically plausible dynamics, on this squared prediction error loss

---

[1]In the sense that one would have to carefully justify biologically the particular nonlinearities going into a nonlinear decision function on a case-by-case basis.

function. The trajectory of the slope parameter over time, $w(t)$, and its governing dynamics may be represented in the biology in various forms. Stochastic rate codes, average activities in populations of neurons and population codes, localized direct electrical signals, and chemical concentration gradients are some possibilities occurring across a range of scales. The dynamical system may also be interpreted as modeling the noisy, time-varying strength of a local field potential or other macro electrophysiological signal when there are multiple, interacting brain regions. We discuss these possibilities further in section 4.

The gradient of E with respect to the weight parameter is given by $\nabla_w E = -\sum_{i=1}^{m}(y_i - wx_i)x_i$ and serves as the starting point. The gradient dynamics $\dot{w} = -\nabla_w E(w)$ are both linear and noise free. Following the discussion above, we modify these dynamics to capture nonlinear saturation effects as well as (often substantial) noise modeling error. Saturation effects lead to a saturated gradient, which we model in the form of the hyperbolic tangent nonlinearity,

$$\dot{w} = -\tanh(a\,\nabla_w E(w)),$$

where a is a slope parameter. Note that the saturated dynamics need not be interpretable as itself the gradient of an appropriate loss function. The fundamental learning problem is defined by the square loss, but it is implemented using an imperfect mechanism that imposes the nonlinearity.[2] The error is modeled with an additional diffusion (noise) term giving the SDE,

$$dw_t = -\tanh(a\,\nabla_w E(w_t))\,dt + \sigma\,dB_t, \tag{2.1}$$

where dB_t denotes the standard one-dimensional Wiener increment process with standard deviation $\sigma > 0$. This noise term $\sigma\,dB_t$ and corresponding error are due to intrinsic neuronal noise (Faisal et al., 2008), aggregated or localized, and possible interference between large assemblies of neurons or circuits and parallels the more general concept of measurement error in networks of coupled dynamical systems.

2.2 Synchronization and Noise. We now consider the effect of having n independent copies of the neural system or pathway implementing the dynamics 2.1, with associated parameters $\{w_1(t), \ldots, w_n(t)\}$. Since these dynamics are nonlinear, the effect of the noise cannot be reduced by simply averaging over the independent trajectories. However, if the circuit copies

---

[2]Put differently, in our setting, the nonlinearity is not part of the learning problem, and so the saturated gradient dynamics should not be viewed as the gradient of another loss function.

are coupled strongly enough, they will attempt to synchronize, and averaging over the copies becomes a potentially powerful way to reduce the effect of the noise (Sherman & Rinzel, 1991; Needleman et al., 2001). The noise can be potentially large (we do not make any small-noise assumptions) and will of course act to break the synchrony. We will explore how well the noise can be reduced by synchronization and redundancy in the sections that follow.

Given n diffusively coupled copies of the noisy neural system and setting $a = 1$ in equation 2.1, we have the following system of nonlinear SDEs,

$$d w_i(t) = - \tanh \left[\sum_{\ell=1}^{m} (w_i(t) x_\ell - y_\ell) x_\ell \right] dt$$

$$+ \sum_{j=1}^{n} W_{ij}(w_j - w_i) dt + \sigma d B_t^{(i)}, \tag{2.2}$$

for $i = 1, \ldots, n$, where $B_t^{(i)}$ are independent standard Wiener processes. The diffusive couplings here should be interpreted as modeling abstract inter-communication between and among different neural circuits, populations, or pathways. In such a general setting, diffusive coupling is a natural and mathematically tractable choice that can capture the key, aggregate aspects of communication among neural systems. Electrical connections such as those implemented by gap junctions in the mammalian cortex (Fukuda, Kosaka, Singer, & Galuske, 2006; Bennett & Zukin, 2004) are also modeled well by diffusive coupling terms when individual neurons are being discussed; however, we emphasize that the system in equation 2.2 is a conceptual model involving possibly large brain regions and do not make assumptions at a level of biological detail that would invoke or require gap-junction-type connectivity.

Each copy of the basic neural circuit is corrupted by independent noise processes but follows the same noise-free dynamics as the others, modulo initial conditions. In fact these coupled systems may start from very different initial conditions. We will assume for simplicity uniform symmetric weights $W_{ji} = W_{ij} = \kappa > 0$ when element i is connected to element j. Defining $(\mathbf{w})_i$ to be the (scalar) output of the ith circuit, we can rewrite system 2.2 in vector form as

$$d\mathbf{w}(t) = - \left(\tanh \left[\sum_{i=1}^{m} (\mathbf{w}(t) x_i - y_i \mathbb{1}) x_i \right] + L\mathbf{w}(t) \right) dt + \sigma d\mathbf{B}_t, \tag{2.3}$$

where $L = \text{diag}(W\mathbb{1}) - W$ is the network Laplacian and $\mathbf{B}_t$ is the standard n-dimensional Wiener process. The spectrum of the network Laplacian captures important properties of the network's topology and will

play a key role. Finally, the change of variable $X_t := \mathbf{w}\|\mathbf{x}\|^2 - \langle \mathbf{x}, \mathbf{y} \rangle \mathbb{1}$, with $(\mathbf{x})_i = x_i$, $(\mathbf{y})_i = y_i$, yields a system that will be easier to analyze:

$$d X_t = -\left(\tanh(X_t)\|\mathbf{x}\|^2 + L X_t \right) dt + \tilde{\sigma} d\mathbf{B}_t, \tag{2.4}$$

where we have defined $\tilde{\sigma} := \sigma \|\mathbf{x}\|^2$. The unique globally stable equilibrium point for the deterministic part of equation 2.4 is seen to be $X^* = 0$, which checks with the fact that the solution to the linear regression problem is $w^* = \langle \mathbf{x}, \mathbf{y} \rangle / \langle \mathbf{x}, \mathbf{x} \rangle$ in this simple case.

2.3 Role of Network Topology. The topology of a network of dynamical systems strongly influences synchronization, to include the rate at which elements synchronize and whether sync (or the tendency to sync) can occur at all in the first place. Thus, the pattern of interconnections among neural systems plays an important role in controlling uncertainty by way of synchronization properties. In a network of stochastic systems of the general (diffusive) type described in section 2, topology can be seen to influence the robustness of synchrony to noise through the spectrum of the network Laplacian. Laplacians arising in various interesting networks and applications have received much attention, in both biological decision making and the context of synchronization of dynamical systems more generally (Kopell & Ermentrout, 1986; Kopell, 2000; Jadbabaie, Lin, & Morse, 2003; Wang & Slotine, 2005; Taylor, Tinsley, Wang, Huang, & Showalter, 2009; Poulakakis et al., 2010).

We consider four important network graphs here, and these arrangements will be helpful examples to keep in mind when interpreting the results given in section 3. The simplest graph of coupled elements is perhaps the full all-to-all graph. This network is also the easiest to synchronize since each element can speak directly to the others. The spectrum of the network Laplacian $\lambda(L)$ for this graph shows why it might be especially effective for reducing uncertainty in the context of equation 2.3. With uniform coupling strength $\kappa > 0$ and n denoting the number of elements in the network, one can check that $\lambda(L) = \{0, n\kappa, \ldots, n\kappa\}$. Denote by λ_- the smallest nonzero (Fiedler) eigenvalue and by λ_+ the largest eigenvalue. Here $\lambda_- = \lambda_+ = n\kappa$, and it is these eigenvalues that control synchronization for any given network. As we will show in theorem 1, the effect of the noise can be reduced particularly quickly precisely because the nonzero eigenvalues depend on both parameters, κ and n.

If fewer connections are made in the network, it becomes harder to synchronize, and we move away from the all-to-all ideal. Figure 1 shows some other common network graphs. The undirected ring graph, in the middle, has spectrum $\lambda_i(L) = 2\kappa[1 - \cos(\frac{2\pi}{n}(i - 1))]$, $i = 1, \ldots, n$. If the single edge connecting the first and last elements is removed to make a chain, as shown on the left in the figure, the network becomes considerably harder
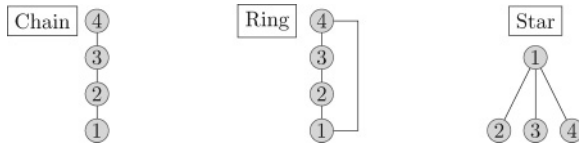
Figure 1: Examples of (undirected) network graphs.

to synchronize (Kopell & Ermentrout, 1986), although the spectrum of the chain looks similar: $\lambda_i(L) = 2\kappa[1 - \cos(\frac{\pi}{n}(i-1))]$. This makes intuitive sense because information is constrained to flow through only one path, and with possibly significant delays. Finally, the star graph shown on the right in the figure has spectrum $\lambda(L) = \{0, \kappa, \dots, \kappa, n\kappa\}$, and we can see that the key Fiedler eigenvalue $\lambda_- = \kappa$ does not grow with the size of the network n. The theorems in section 3 then predict that it will be impossible to increase the synchronization rate simply by incorporating more copies of the neural circuit. The coupling strength must also increase to make fluctuations from the common trajectory (synchronization subspace) small. We will discuss this case in more detail. As might be particularly relevant to brain anatomy, random graphs and directed graphs may also be considered and have been studied extensively (Bollobas, 2001).

In neuroscience-related models, each connection in a network has an associated biophysical cost in terms of energy and space requirements. All-to-all networks, with n^2 connections among n circuits or neurons, is often criticized as being biologically unrealistic because of this cost. However, it has been noted that all-to-all connectivity can be implemented with $2n$ connections using quorum sensing ideas (Taylor et al., 2009), wherein a global average is computed and shared. The global average is computed given inputs from all n elements, and this average is sent back to each circuit via other n connections. The shared variable may be communicated by synapses or sensed chemically or electrically. Although quorum sensing cannot realize any set of n^2 connections, the global average may be a weighted average, or there may be several common variables organized hierarchically. This allows for a rich set of networks with $\mathcal{O}(n)$ connectivity that behave more like networks with all-to-all connectivity for synchronization and stability purposes. Furthermore, dynamics in the computation of the quorum variable itself, when appropriate for modeling purposes, does not necessarily pose any special difficulty for establishing synchronization properties if virtual systems are used (Russo & Slotine, 2010).

The difficulty with which synchrony may be imposed can be normalized by the number of connections in many cases to obtain a comparison between synchronization properties of various graphs that take biological cost into account. Using quorum variables where appropriate, graphs whose spectrums depend on n are thus roughly comparable on equal biological terms.

Cost-normalized comparisons of synchronization properties are not always possible or meaningful, however. Consider the ring and chain networks introduced above. There is a difference of one edge between the two, but in the noise-free setting, for example, the chain requires asymptotically four times more effort to synchronize than the ring architecture (see Wang & Slotine, 2005).

2.4 A Comment on Stability and Contraction. We turn to analyzing the stability of the nonlinear system given by equation 2.4. We will argue that this is difficult for two reasons: the presence of noise and the fact that the (noise-free) dynamics saturate in magnitude. Indeed, without additional assumptions, one cannot in general show that the system is globally exponentially stable. A common method for studying the stability properties of a noiseless nonlinear dynamical system is via Lyapunov theory (Slotine & Li, 1991); however, in the presence of noise, system trajectories along the Lyapunov surface may not be strictly decreasing. Contraction analysis (Lohmiller and Slotine, 1998; Wang & Slotine, 2005) is a differential formalism related to Lyapunov exponents and captures the notion that a system is stable in some region if initial conditions or temporary disturbances are forgotten. If all neighboring trajectories converge to each other, global exponential convergence to a single trajectory can be concluded:

Definition 1 (Contraction). *Given the system equations $\dot{x} = f(x, t)$, a region of the state space is called a contraction region if the Jacobian $J_f = \frac{\partial f}{\partial x}$ is uniformly negative definite in that region. Furthermore, the contraction rate is given by β, where $\frac{1}{2}(J_f + J_f^\top) \le \beta I < 0$.*

An analogous definition in the case stochastic dynamics has also been developed (Pham, Tabareau, & Slotine, 2009) and requires contraction of the noise-free dynamics as well as a uniform upper bound on the variance of the noise. However, for system 2.4, the Jacobian is found to be

$$J(\mathbf{w}) = \|\mathbf{x}\|^2 \mathrm{diag}(\tanh^2(\mathbf{w}) - 1) - L,$$

so that $\lambda(J(\mathbf{w})) < -\lambda(L) \le -\lambda_{\min}(L) = 0$. The subspace of constant vectors is a flow-invariant subspace, and L does not contribute to the dynamics in this flow invariant space since L has a 0 eigenvalue corresponding to its constant eigenvector. This difficulty can arise whenever one considers diffusively coupled elements, and in such cases, the usual way around this difficulty is to work with an auxiliary or virtual system (as in Pham & Slotine, 2007) and study contraction to the flow-invariant subspace starting from initial conditions outside. However, since $\tanh'(x) = 1 - \tanh^2(x)$, we are still left with the difficulty that the noise-free dynamics can have a convergence rate to equilibrium arbitrarily close to 0 as one travels far out

to the tails of the tanh function; the system is not necessarily contracting. Indeed, for any saturated dynamics, $\tanh\big(\mathbf{f}(\mathbf{x}, t)\big)$, the rate can be arbitrarily small. Thus, one cannot easily determine the rate of convergence to equilibrium using standard techniques. The analysis we provide in the succeeding sections attempts to get around these difficulties by separately exploring the system's behavior in and out of the flow-invariant (synchronization) subspace of constant vectors.

3 Controlling Uncertainty in Learning

In this section, we present and interpret the main results of the letter. The argument we put forward is that noisy, nonlinear trajectories can be linearly averaged to reduce the noise if fluctuations due to noise can be made small. We show that the fluctuations can be made small by coupling the dynamical systems and that one can precisely control the size of the fluctuations. In particular, we give estimates that show that the trade-off between noise and coupling strength among neural circuits determines the amount of uncertainty surrounding the decisions made by the neural system. Proofs of the theorems are postponed until section 6.

3.1 Preliminaries. We begin by decomposing the stochastic process $\{X_t \in \mathbb{R}^n\}_{t \geq 0}$ into a sum describing fluctuations about the center of mass. Let $P = I - (1/n)\mathbf{1}\mathbf{1}^\top$, the canonical projection onto the zero-mean subspace of $\mathbb{R}^n$, and define $Q = I - P$. Then for all $t \geq 0$, $X_t = P X_t + Q X_t$. Clearly, $\ker P = \mathrm{im}\, Q$ is the subspace of constant vectors. We will adopt the notation $\widetilde{X}_t$ for $P X_t$ and $\overline{X}_t \mathbf{1}$ for $Q X_t$ (along with the analogous notation $\widetilde{\mathbf{w}}_t$ and $\bar{w}_t$) and derive expressions for these quantities based on equation 2.4. The macroscopic variable $\overline{X}_t$ satisfies

$$d\overline{X}_t = \frac{1}{n}\mathbf{1}^\top d X_t = -\frac{\|\mathbf{x}\|^2}{n}\mathbf{1}^\top \tanh(X_t)\, dt + \frac{\tilde{\sigma}}{\sqrt{n}} d B_t \tag{3.1}$$

and thus

$$d\widetilde{X}_t = d X_t - d\overline{X}_t \mathbf{1}$$

$$= -\Big(\tanh(X_t)\|\mathbf{x}\|^2 + L X_t - \frac{\|\mathbf{x}\|^2}{n}\mathbf{1}^\top \tanh(X_t)\mathbf{1}\Big) dt$$

$$+ \tilde{\sigma} d\mathbf{B}_t - \frac{\tilde{\sigma}}{\sqrt{n}} d B_t \mathbf{1}. \tag{3.2}$$

In terms of the original variable $\mathbf{w}$, the fluctuations $\widetilde{\mathbf{w}}_t$ are purely due to the noise, while $\bar{w}_t$ parameterizes the average decision function. Because the decision function we consider is linear, the uncertainty in the decisions

is directly equivalent to uncertainty in the parameter w. We will study the evolution of both the mean and the fluctuation processes over time; however, to assess uncertainty, the central quantity of interest will be the size of the ball containing the fluctuations (the "microscopic" variables). We characterize the magnitude of the fluctuations via the squared norm process satisfying

$$\frac{d\|\widetilde{X}_t\|^2}{2} = -\left(\|\mathbf{x}\|^2\langle\widetilde{X}_t, \tanh(X_t)\rangle + \langle\widetilde{X}_t, L\,X_t\rangle\right)dt$$

$$+\frac{1}{2}\tilde{\sigma}^2(n-1)\,dt + \tilde{\sigma}\|\widetilde{X}_t\|d\,B_t \tag{3.3}$$

which follows from equation 3.2 applying Ito's lemma to the function $h(\widetilde{X}_t) = \frac{1}{2}\langle\widetilde{X}_t, \widetilde{X}_t\rangle$ and the fact that $\langle\widetilde{X}_t, d\mathbf{B}_t\rangle = \|\widetilde{X}_t\|d\,B_t$ in law.

3.2 Uncertainty Estimates. The first—and central—result says that the ball centered at $\bar{w}$ (the center of mass) containing the fluctuations can be controlled in expectation from above and below by the coupling strength and in most cases the number of circuit copies, via the spectrum of the network Laplacian L. We note that lower bounds are typically ignored in the dynamical systems literature, possible because they are less important for stability analyses. We have found, however, that such bounds can be derived in the case of saturated gradient dynamics and that control from below can yield further insight into the problem of neural learning.

Let λ_+ be the largest eigenvalue of L, and let λ_- be the smallest nonzero eigenvalue of L.

Theorem 1 (*fluctuations can be made small*). *After transients of rate $2\lambda_-$,*

$$\frac{(n-1)\sigma^2}{2\lambda_+}\left(1 - \frac{\|\mathbf{x}\|^2}{\lambda_-}\right) \le \mathbb{E}\|\tilde{w}(t)\|^2 \le \frac{(n-1)\sigma^2}{2\lambda_-}$$

where $\tilde{w} = P\mathbf{w}(t)$.

Clearly the lower bound is informative only when $\lambda_- > \|\mathbf{x}\|^2$. Although we do not explicitly assume any particular bound on the size of the examples $\|\mathbf{x}\|$, it is reasonable that $\lambda_- \gg \|\mathbf{x}\|^2$ since λ_- can depend on the number of circuits n and will always depend on the coupling strength κ, which can be large. Large coupling strengths can be found in a variety of circumstances, particularly in the case of motor control circuits (Grandhe, Abbas, & Jung, 1999; Kiemel, Gormley, Guan, Williams, & Cohen, 2003) for example.

In the next theorem, we give the variance of the fluctuations via a higher moment of $\|\widetilde{w}\|$. This result makes use of the lower bound in theorem 1, and

leads to a result that gives control of the fluctuations in probability rather than in expectation.

Theorem 2 *(variance of the trajectory distances to the center of mass). After transients of rate $2\lambda_-$,*

$$\mathsf{var}\left(\|\tilde{\boldsymbol{w}}(t)\|^2\right) \le \left(\frac{(n-1)\sigma^2}{2\lambda_-}\right)^2 \left(2 + \frac{4}{n-1}\right)$$
$$- \left(\frac{(n-1)\sigma^2}{2\lambda_+}\right)^2 \left(1 - \frac{\|\boldsymbol{x}\|^2}{\lambda_-}\right)^2.$$

Chebyshev's inequality combined with theorem 2 immediately gives the following corollary:

Corollary 1. *After transients of rate $2\lambda_-$,*

$$\mathbb{P}\left[\left\|\tilde{\boldsymbol{w}}(t)\right\|^2 - \mathbb{E}\left\|\tilde{\boldsymbol{w}}(t)\right\|^2 \ge \varepsilon\right] \le \frac{\mathsf{var}\left(\|\tilde{\mathbf{w}}(t)\|^2\right)}{\varepsilon^2}. \tag{3.4}$$

Since any connected network graph has nontrivial eigenvalues that depend on the uniform coupling strength κ, we see that for fixed n as $\kappa \to \infty$, $\mathsf{var}(\|\tilde{\mathbf{w}}(t)\|^2) \to 0$. In the case of the all-to-all network topology, for example, the eigenvalues of L depend on both n and κ so that $\mathsf{var}(\|\tilde{\mathbf{w}}(t)\|^2) = \mathcal{O}(\kappa^{-2})$, giving a power law decay of order $\mathcal{O}(\kappa^{-2}\varepsilon^{-2})$ on the right-hand side of equation 3.4.

Finally, we turn to estimating in expectation the steady-state average distance between the trajectories of the circuit copies and the noise-free solution. As we argued in section 2.4, the rate of convergence to equilibrium of the trajectories $w_i(t)$ can be arbitrarily small. Although from theorems 3.1 and 3.21 the fluctuations can be made small, one cannot in general make a similar statement about the center of mass $\bar{w}_t$ process unless assumptions about the initial conditions are made (and, by extension, the same holds true for the trajectories $w_i(t)$). Such an assumption would lead to control over the contribution of the tanh terms and establishes a lower bound on the contraction rate. Rather than make a specific assumption, however, we state a general result: we again provide a lower bound, this time following from the law of large numbers governing sums of i.i.d. gaussian random variables and the lower bound on the fluctuations provided by theorem 1:

Theorem 3 (*average distance to the noise-free trajectory*). *Denote by w^* the minimizer of the squared-error objective, equation 2.1. After transients of rate $2\lambda_-$,*

$$\frac{\sigma^2}{n} + \left[\frac{(n-1)\sigma^2}{2n\lambda_+} \left(1 - \frac{\|\boldsymbol{x}\|^2}{\lambda_-} \right) \right]^+$$

$$\leq \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (w_i(t) - w^*)^2 \right] \leq \frac{\sigma^2}{2\lambda_-} + \mathbb{E}[(\bar{w}_t - w^*)^2],$$

where $[\cdot]^+ \equiv \max(0, \cdot)$.

Theorem 3 says that the average closeness of the noisy system to that of the noise-free optimum is controlled by the trade-off between the noise and the coupling strength and the number of circuit copies n. The former controls in large part the magnitude of the fluctuations, as discussed above. The latter quantity is the unavoidable linear averaging component and can be brought to 0 only as fast as the law of large numbers allows— $\mathcal{O}(n^{-1/2})$ at best. For fixed n as $\lambda_-^{(n)} \to \infty$, the upper and lower bounds coincide since $\mathbb{E}[(\bar{w}(t) - w^*)^2] \to \sigma^2/n$. As both $n \to \infty$ and $\kappa \to \infty$, theorem 3 confirms that $w_i(t) \to w^*$ in expectation. If the fluctuations are not made small, however, linear averaging will be wrong, and the error will be greater. Just how bad linear averaging is when the fluctuations are allowed to be large is described in large part by the maximum curvature of the noise-free dynamics.[3]

Finally, we note that the estimates above depend on the number of samples m only through the norm of the examples $\mathbf{x}$, and it is reasonable to assume that this quantity may be appropriately normalized based on the maximum values conveyed by subsystems or rates of neurons comprising the circuit in the case of population or rate codes, or maximum field strengths in the case of LFPs. However, the requirement that the organism must collect m observations before learning can proceed is not essential. We may also consider the online learning setting, where data are observed sequentially and updates to the parameters $(w)_i$ are made separately on the basis of each observation in temporal order. The analysis above studies convergence to and distance from the solution in the steady state, whatever that solution may be, given m pieces of evidence. Thus, the online setting can also be considered as long as the time between observations is longer than the transient periods. Indeed, in many scenarios, learning and decision-making processes in the brain can take place on short timescales relative to the timescale on which experience is accumulated. In this case, when

---

[3]One way to see this is to take the first-order Taylor expansion of the dynamics with integral remainder. The remainder term can be upper-bounded by the spectral radius of the Hessian matrix, which is related to curvature (see Tabareau et al., 2010).

another piece of information arrives, the system moves to a region defined (stochastically) around a new steady state. A complication can arise when the new point arrives during the transient period of the previous learning process—before the system has had a chance to settle, on average, into the new equilibrium—however, we do not attempt to model this situation here.

4 Discussion

The estimates given in section 3 quantify the trade-off between the degree of synchronization and the noise (error) and the role this trade-off plays in determining the uncertainty of a decision function learned by way of a stochastic, nonlinear dynamics. Estimates in both expectation and probability were derived. We showed how and where both the coupling strength and the topology of the network of neural ensembles affect the extent to which the noise, and therefore uncertainty due to error, can be controlled. In particular, for most networks (see section 2.3), the effect of the noise can be reduced by increasing either the coupling strength or the number of redundant systems (or both), leading to a steady-state solution that is going to be closer to the ideal, error-free solution with high probability. From a technical standpoint, this is because fluctuations about the common trajectory are exactly the way in which the noise enters the picture. When the fluctuations are made small, the error is made small. In this way, an organism may mitigate error imposed by a noisy, imperfect learning apparatus and solve a learning task to greater accuracy. Furthermore, synchronization and redundancy can both improve the speed of learning, in the sense that the rate of convergence to the steady-state solution also depends on these mechanisms. Each of the bounds presented in section 3 holds after transient terms of order $e^{-t\lambda_-}$ vanish, where λ_- is the smallest nonzero eigenvalue of the network Laplacian. For any stable connected network, strong coupling strengths directly improve convergence rates to the steady state, as seen by the dependence of λ_- on κ. In the case of all-to-all (including approximately all-to-all and many random graphs), $\lambda_- = \mathcal{O}(n\kappa)$, so that both increased redundancy and sync will improve the speed of learning.

Our overarching goal has been to explore quantitatively the role of redundancy and synchronization in reducing error incurred by a stochastic, nonideal learning and decision-making neural process. We have gone about this by considering a model that emphasizes the underlying computations taking place rather than particular neural representations. Looking at the appropriate scale, we seek to address the precise meaning of ensemble measurements and population codes, as well as the information these codes convey about the underlying dynamics and signals. The results derived support the notion that synchronization and redundancy play a more functional role in the context of learning processes in the brain rather than being a mere epiphenomenon.

4.1 Synchronization and Redundancy in the Brain. Synchronization has been suggested, over a diverse history of experimental work, as a fundamental mechanism for improvement in precision and reduction of uncertainty in the nervous system (see Needleman et al., 2001; Enright, 1980). Redundancy too is an important and commonly occurring mechanism. In retinal ganglion cells (Croner, Purpura, & Kaplan, 1993; Puchalla, Schneidman, Harris, & Berry, 2005) and heart cells (Clay & DeHaan, 1979), the spatial mean across coupled cells cancels out noise. Populations of hair cells in otoliths perform redundant, collaborative computations to achieve robustness (Kandel, Schwartz, & Jessell, 2000; Eliasmith & Anderson, 2004), and it has been suggested that multiple cortical (amygdala-thalamus) loops contribute to fear response and conditioning, and emotion more generally (LeDoux, 2000). With motor tasks such as reaching or standing, it has been argued that planning and representation occur at least partially in redundant coordinate systems and involve redundant degrees of freedom (Scholz & Schoner, 1999). (Todorov, 2008) maintains that redundancy and noise combine to give rise to optimal muscle control policies, raising the interesting possibility that in some cases, the impact of the noise may need to be adjusted but not necessarily eliminated altogether. On a more localized scale, reach direction has also been found to be conveyed by populations of neurons with overlapping tuning curves (Georgopoulos, Kalaska, Caminiti, & Massey, 1982), where synchrony within such populations plays an important role (Grammont & Riehle, 1999). Multiple sensorimotor transformations involving disparate brain regions may be at play in the parietal cortex, where redundant sensory inputs from multiple modalities must be mapped into motor responses (Ting, 2007; Pouget & Sejnowski, 1997). In the ascending auditory pathway, varying degrees of redundancy have been noted and contribute to the robust representation of frequency and more complex auditory objects (Chechik et al., 2006). Ensemble measurements have also been connected to behavior and have been suggested as inputs to brain-machine interfaces, while in stochastic neural decision making, it has been suggested that it is the collective behavior across multiple populations of neurons that is responsible for perception and decision making, rather than activity of a single neuron or population of neurons (Gigante et al., 2009).

In these examples and more generally, we suggest that redundancy plus feedback synchronization is a mechanism that may be used to improve the accuracy, robustness, and speed of a learning process involving the relevant respective brain areas. This is separate from, and in contrast to, redundancies that are harnessed to specifically increase storage capacity, as in the case of associate memory models (Hertz, Krogh, & Palmer, 1991). There, robustness to corruption is also achieved (via pattern completion dynamics), but the degree of robustness must be traded off against capacity. The primary function of such populations of neurons is to ostensibly store and retrieve memory patterns rather than to implement adaptive, learning dynamics while eliminating noise.

Another theme emerging from these instances of sync and redundancy is that key computations may be seen as implemented by distant brain regions coupled together by way of long-distance projections and network hubs. Recent experimental observations in *C. elegans* casts this interpretation in a developmental light (Varier & Kaiser, 2011) and suggests that such interactions occur from an early stage in life and are important for normal development in even simple organisms. Learning processes realized by such computations and interactions are certainly susceptible to noise and must cope with this noise one way or another. We suggest that synchronization and redundancy are not only present and possible but provide a ready, natural solution.

The ability to learn and make decisions reliably in the presence of uncertainty is of fundamental importance for the survival of any organism. This uncertainty can be seen to arise from three distinct sources, and the approach discussed here treats only the first two: intrinsic neuronal noise, both local and in the aggregate, and noise in the form of measurement error, under which we include error due to limitations in precision and nonlinearity in biological systems. A third, and equally important, source of error is that of uncertainty in the inference process itself (Yang & Shadlen, 2007; Kiani & Shadlen, 2009). This uncertainty is specific to and inherent in the decision problem and is characterized by the posterior distribution over decisions given the experiential evidence. Our work considers uncertainty only beyond that of the inference process and, as such, is one part of a larger puzzle. We argue that intrinsic noise is both experimentally and theoretically important—and involved enough technically—to be addressed in isolation, while holding all other variables constant. Indeed, intrinsic noise intensities can be large. The role of the network's topology and coupling mechanism also strongly influences the overall picture, often in surprising or subtle ways. But it is also possible that the methods recruited here can be applied toward understanding some aspect of the inference error if different inferences from the same observations can be made by different "expert" (circuits), each with its own biases. Then averaging, nonlinearity and the uncertainty could potentially be treated in a similar framework.

4.2 Extensions and Generalizations. Asymptotic stability of the stochastic system considered here is guaranteed as long as there is coupling. In general, if the dynamics of a stochastic system are contracting or can be made contracting with feedback, then combinations (e.g., parallel, serial, hierarchical) of such systems will be contracting (Pham et al., 2009; Lohmiller & Slotine, 1998). In the present setting, the system governing the fluctuations about the mean trajectory is contracting with a rate dependent on the coupling strength and the noise variance. Thus, combinations of learning systems of the general type considered here can enjoy strong stability guarantees automatically since the individual systems are contracting.

Finally, we have assumed throughout that the errors affecting the collection of redundant neural circuits or systems are mutually independent. This is not an unreasonable modeling assumption. For large-scale learning processes involving different brain areas, noise imposed by local spike irregularities is largely unrelated to noise present in distant circuits. Within small populations of neurons, it is likely that dependence among intrinsic neuronal noise sources decays rapidly in space so that nearest neighbors may experience somewhat correlated noise, but beyond this are not significantly affected by other members of the population. Because noise in a biological environment can never be fully dependent (whether due to thermal or chemical-kinetic factors, or otherwise), partial dependence among noise inputs may be explicitly modeled as, for example, mixing processes if desired (Doukhan, 1994). Estimates of the form discussed here would then be augmented with mixing terms, leading to results that make identical qualitative statements about the role of redundancy and sync. Fluctuations, and the effect of the noise, would still be reducible but would require larger coupling strengths or more redundancy compared to what would be necessary if the noise sources were independent.

5 Simulations

To empirically test the estimates given in section 3 we simulated several systems of SDEs given by equation 2.4 using Euler-Maruyama integration (over time $t \in [0, 10s]$, 10^5 regularly spaced sample points), for different settings of the parameters n (number of circuits or elements), κ (coupling strength) and σ (noise standard deviation). Initial conditions were randomly drawn from the uniform distribution on $[-5, 5]$, and we fixed $\|\mathbf{x}\|^2 = 1$ and the coupling arrangement to all-to-all coupling with fixed strength determined by κ. For simplicity the simulated systems had an equilibrium point at 0, corresponding to $\mathbf{y} = 0$, so that $\langle x, y \rangle = 0$ and $X^* = w^*$ (the change of variables is the identity map and we can identify X_t with w_t).

For comparison purposes, we show at the left in Figure 2 typical simulated trajectories of uncoupled (top) and coupled (bottom) populations when $n = 20, \kappa = 5, \sigma = 10$. Both populations are driven by the same noise and the same set of initial conditions; however, each element is driven by noise independent from the others as assumed above. From the units on the vertical axes, one can see that coupling clearly reduces intertrajectory fluctuations as expected. On the right in Figure 2, we show the coupled and uncoupled populations' respective center of mass trajectories for this particular simulation instance. One can see from this figure that the average of the coupled system tends closer to 0 (X^*) and is less affected by large noise excursions.

To empirically test the tightness of the estimates given in section 3, we repeated simulations of each respective system 5000 times and averaged the relevant outcomes to approximate the expectations appearing in
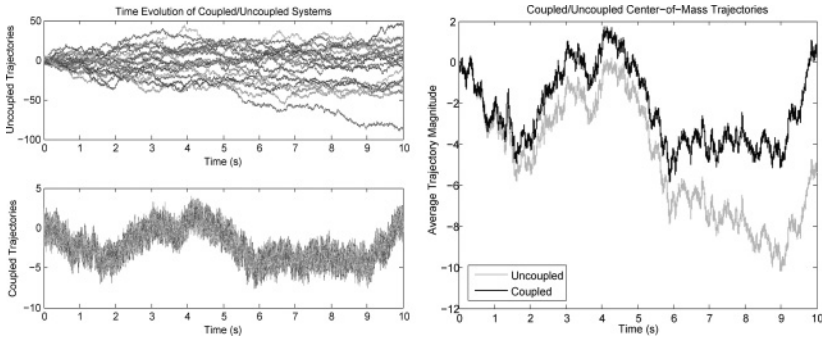
Figure 2: (Left) Typical simulated trajectories for coupled and uncoupled networks driven by the same noise. (Right) Population average trajectories for the coupled and uncoupled systems.

the bounds. Transient periods were excluded in all cases. In Table 1, we show the values predicted by the bounds and the corresponding simulated quantities, for each respective triple of system parameter settings. Sample standard deviations of the simulated averages (expectations) are given in parentheses. In Figure 3 we show theoretical versus simulated expected magnitudes of the fluctuations $\mathbb{E}\|\widetilde{X}_t\|^2$ when $n = 200$ and $\sigma = 10$ over a range of coupling strengths. The solid dark trace is the upper bound of theorem 1, while the open circles are the average simulated quantities (again 5000 separate simulations were run for each κ). Error bars are also given for the simulated expectations. Note that the magnitude scale (y-axis) is logarithmic, so the error bars are also plotted on a log scale. We omitted the lower theoretical bound from the plot because it is too close to the upper bound to visualize well relative to the scale of the bounds.

Generally the estimates relating to the magnitude of the fluctuations are seen to be tight, and the variance estimate is within an order of magnitude. For the experiments with large noise amplitudes, the empirical estimates can appear to slightly violate the bounds where the bounds are tight since the variance across simulations is large. The lower bound estimating the distance of the center of mass to the noise-free solution is also seen to be reasonably good. For comparison, we give the upper estimate where the empirical distance is substituted in place of the expectation in order to show closeness to the lower bound. Theorem 3 predicts that the upper and lower estimates will eventually coincide if κ and/or n are chosen large enough.

6 Proofs

In this section we provide proofs of the results discussed in section 3.

Table 1: Estimated Versus Simulated Quantities.

Quantity	Lower Bound	Simulated	Upper Bound
$n = 20, \kappa = 5, \sigma = 10$			
$\mathbb{E}\|\widetilde{X}_t\|^2$	9.405	9.497 (std = 3.1)	9.500
$\mathrm{var}(\|\widetilde{X}_t\|^2)$	-	9.450 (std = 14.7)	111.046
$\frac{1}{n}\mathbb{E}\|X_t - X^*\mathbb{1}\|^2$	5.470	12.249 (std = 22.2)	12.249 (std = 22.2)
$n = 20, \kappa = 1, \sigma = 5$			
$\mathbb{E}\|\widetilde{X}_t\|^2$	11.281	11.719 (std = 3.8)	11.875
$\mathrm{var}(\|\widetilde{X}_t\|^2)$	-	14.261 (std = 23.0)	184.45
$\frac{1}{n}\mathbb{E}\|X_t - X^*\mathbb{1}\|^2$	1.814	1.933 (std = 2.5)	1.946 (std = 2.4)
$n = 20, \kappa = 1, \sigma = 10$			
$\mathbb{E}\|\widetilde{X}_t\|^2$	45.125	47.053 (std = 15.2)	47.500
$\mathrm{var}(\|\widetilde{X}_t\|^2)$	-	230.275 (std = 373.4)	2951.234
$\frac{1}{n}\mathbb{E}\|X_t - X^*\mathbb{1}\|^2$	7.256	14.761 (std = 24.1)	14.784 (std = 24.1)
$n = 100, \kappa = 1, \sigma = 10$			
$\mathbb{E}\|\widetilde{X}_t\|^2$	49.005	49.556 (std = 7.0)	49.500
$\mathrm{var}(\|\widetilde{X}_t\|^2)$	-	49.332 (std = 70.6)	2598
$\frac{1}{n}\mathbb{E}\|X_t - X^*\mathbb{1}\|^2$	1.490	1.449 (std = 1.6)	1.449 (std = 1.6)
$n = 100, \kappa = 5, \sigma = 10$			
$\mathbb{E}\|\widetilde{X}_t\|^2$	9.880	10.137 (std = 1.5)	9.900
$\mathrm{var}(\|\widetilde{X}_t\|^2)$	-	2.151 (std = 3.2)	102.362
$\frac{1}{n}\mathbb{E}\|X_t - X^*\mathbb{1}\|^2$	1.099	1.496 (std = 1.5)	1.496 (std = 1.5)

We first introduce a key lemma to be used in the development immediately below.

Lemma 1. *Let* $P = I - (1/n)\mathbb{1}\mathbb{1}^\top$, *the canonical projection onto the zero mean subspace of* $\mathbb{R}^n$. *Then for all* $x \in \mathbb{R}^n$,

$$0 \le \langle Px, \tanh(x)\rangle \le \|Px\|^2,$$

where the hyperbolic tangent applies elementwise.

Proof. Given $x \in \mathbb{R}^n$, define the index sets $I = \{1, \ldots, n\}$, $I_+ = \{i \in I \mid (Px)_i \ge 0\}$, and $I_- = I \setminus I_+$. Since Px is zero mean, $\sum_{i \in I_+}(Px)_i = \sum_{i \in I_-}|(Px)_i|$. We will express the hyperbolic tangent as $\tanh(z) = 2s(2z) - 1$, where $s(z) = (1 + e^{-z})^{-1}$ is the logistic sigmoid function. If we let $\mu = \frac{1}{n}\mathbb{1}^\top x$ be the center of mass of x, $(Px)_i = x_i - \mu \ge 0$ implies $s(x_i) \ge s(\mu)$ by monotonicity of s. Likewise, $(Px)_i < 0$ implies $s(x_i) < s(\mu)$. Finally, note that since
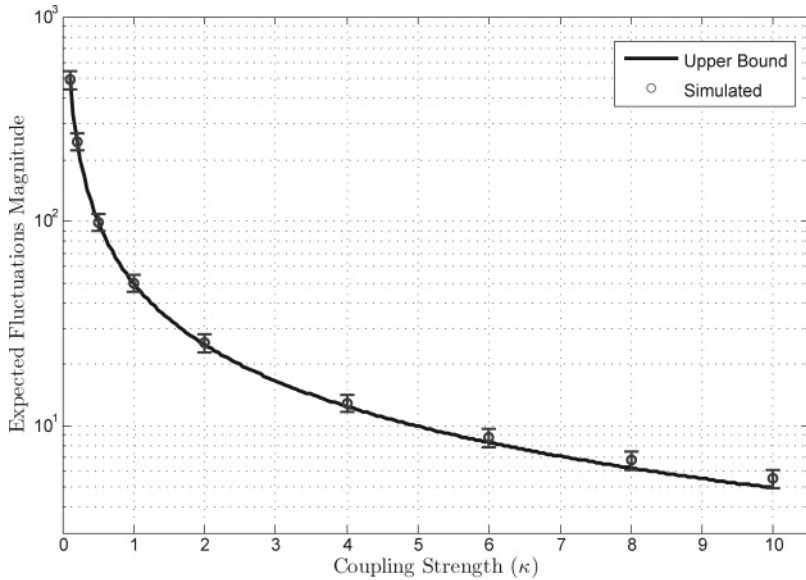
Figure 3: Simulated versus theoretical upper-bound estimates of the fluctuations' expected magnitude over a range of coupling strengths κ. Here $n = 200$ circuits and $\sigma = 10$.

$P^2 = P$ and $1 \in \ker P$, $\langle Px, \tanh(x) \rangle = \langle Px, P(2s(2x) - 1) \rangle = 2\langle Px, s(2x) \rangle$. Using these facts, we prove the lower bound first:

$$\langle Px, \tanh(x) \rangle = 2 \sum_{i \in I_+} (Px)_i s(2x_i) - 2 \sum_{i \in I_-} |(Px)_i| s(2x_i)$$

$$\geq 2s(2\mu) \sum_{i \in I_+} (Px)_i - 2s(2\mu) \sum_{i \in I_-} |(Px)_i|$$

$$= 2s(2\mu) \cdot 0 = 0.$$

Turning to the upper bound, we prove the equivalent statement $\langle Px, s(2x) - x \rangle \leq 0$. First, if $\mu = 0$, then $Px = x$ so $\langle Px, \tanh(x) \rangle = \langle x, \tanh(x) \rangle \leq \|x\| \|\tanh(x)\| \leq \|x\|^2 = \|Px\|^2$, since $\|\tanh(x)\| \leq \|x\|$ by virtue of the fact that $|\tanh(z)| = \tanh(|z|) \leq |z|$ for any $z \in \mathbb{R}$. Now suppose that $\mu > 0$. If $z \geq \mu > 0$, we can upper-bound $s(2z)$ by the line tangent to the point $(\mu, s(2\mu))$: $s(2z) \leq mz + b$ with $m < \frac{1}{2}$ and $b > \frac{1}{2}$. If $z < \mu$, we can take the lower bound $s(2z) > \frac{1}{2}z + \frac{1}{2}\mu - s(2\mu)$. Using these estimates,

we have that

$$\langle Px, s(2x) - x\rangle = \sum_{i\in I_+}(Px)_i(s(2x_i) - x_i) + \sum_{i\in I_-}|(Px)_i|(x_i - s(2x_i))$$

$$\leq \sum_{i\in I_+}(Px)_i(b - (1 - m)x_i)$$

$$+ \sum_{i\in I_-}|(Px)_i|\left(\frac{1}{2}x_i + \frac{1}{2}\mu - s(2\mu)\right)$$

$$\leq \sum_{i\in I_+}(Px)_i(b - (1 - m)\mu)$$

$$+ \sum_{i\in I_-}|(Px)_i|\left(\frac{1}{2}\mu + \frac{1}{2}\mu - s(2\mu)\right)$$

$$= \left(\sum_{i\in I_+}(Px)_i\right)(b + m\mu - s(2\mu)) = 0.$$

The second inequality follows from the fact that $(1 - m) > 0$, $x_i \geq \mu$ for $i \in I_+$ and $x_i < \mu$ for $i \in I_-$. Since $\sum_{i\in I_+}(Px)_i = \sum_{i\in I_-}|(Px)_i|$, and recalling that by definition b satisfies $m\mu + b = s(2\mu)$, the final equalities follow. If $\mu < 0$, then the proof is similar, taking the line tangent to the point $(\mu, s(2\mu))$ as a lower bound for $s(2z)$ and the line $\frac{1}{2}(z - \mu) + s(2\mu)$ as an upper bound.

6.1 Fluctuations Estimates: Proof of Theorem 1. We begin by adding $\lambda\|X_t\|^2 dt$, with $\lambda \in (0, \infty)$, to both sides of equation 3.3 to obtain

$$\frac{1}{2}d\|\widetilde{X}_t\|^2 + \lambda\|\widetilde{X}_t\|^2 dt = -\|\mathbf{x}\|^2\langle\tanh X_t, \widetilde{X}_t\rangle dt + (\lambda\|\widetilde{X}_t\|^2 - \langle LX_t, \widetilde{X}_t\rangle)dt$$

$$+ \frac{1}{2}(n - 1)\tilde{\sigma}^2 dt + \tilde{\sigma}\|\widetilde{X}_t\|dB_t = e^{-2\lambda t}d\left(\frac{1}{2}\|\widetilde{X}_t\|^2 e^{2\lambda t}\right),$$

where the second equality follows, noticing that the right-hand side is the total Ito derivative of the left-hand side of the first equality. Now multiply both sides by $e^{2\lambda t}$, switch to integral form, and multiply both sides by $e^{-2\lambda t}$ to arrive at

$$\frac{1}{2}\|\widetilde{X}_t\|^2 = e^{-2\lambda t}\|\widetilde{X}_0\|^2 + \int_0^t e^{2\lambda(s-t)}\left(\frac{1}{2}(n - 1)\tilde{\sigma}^2 - \|\mathbf{x}\|^2\langle\tanh X_s, \widetilde{X}_s\rangle\right)ds$$

$$+ \int_0^t e^{2\lambda(s-t)}(\lambda\|\widetilde{X}_s\|^2 - \langle LX_s, \widetilde{X}_s\rangle)ds + \tilde{\sigma}\int_0^t e^{2\lambda(s-t)}\|\widetilde{X}_s\|dB_s.$$

$$(6.1)$$

Upper bound: Next, note that $\langle L X_t, \widetilde{X}_t \rangle = \langle L \widetilde{X}_t, \widetilde{X}_t \rangle$ since $L(\overline{X}_t 1) = 0$, and that $\widetilde{X}_t$ is by definition orthogonal to any constant vector. For all t, we also have that

$$
\begin{aligned}
\lambda_- \|\widetilde{X}_t\|^2 - \langle L\widetilde{X}_t, \widetilde{X}_t \rangle &\leq 0 \\
-\langle \tanh X_t, \widetilde{X}_t \rangle &\leq 0
\end{aligned}
\tag{6.2}
$$

almost surely. The first inequality follows from the fact that for all $x \in \operatorname{im} P$,

$$
\lambda_- \|\widetilde{X}_t\|^2 \leq \langle L\widetilde{X}_t, \widetilde{X}_t \rangle \leq \lambda_+ \|\widetilde{X}_t\|^2,
$$

if λ_- is the Fiedler eigenvalue of L and λ_+ is the largest eigenvalue of L. The second inequality is given by lemma 1. Setting $\lambda \equiv \lambda_-$ and applying the inequalities 6.2 to equation 6.1 gives the estimate

$$
\begin{aligned}
\frac{1}{2}\|\widetilde{X}_t\|^2 &\leq e^{-2\lambda_- t}\|\widetilde{X}_0\|^2 + \frac{(n-1)\tilde{\sigma}^2}{2}\int_0^t e^{2\lambda_-(s-t)}ds + \tilde{\sigma}\int_0^t e^{2\lambda_-(s-t)}\|\widetilde{X}_s\|\,dB_s \\
&= e^{-2\lambda_- t}\|\widetilde{X}_0\|^2 + \frac{(n-1)\tilde{\sigma}^2}{4\lambda_-}(1 - e^{-2\lambda_- t}) + \tilde{\sigma}\int_0^t e^{2\lambda_-(s-t)}\|\widetilde{X}_s\|\,dB_s
\end{aligned}
\tag{6.3}
$$

almost surely. Taking expectations and noting that $\mathbb{E}[\int_0^t e^{2\lambda_-(s-t)}\|\widetilde{X}_s\|\,dB_s] = 0$, we have that

$$
\mathbb{E}\|\widetilde{X}_t\|^2 \leq \frac{(n-1)\tilde{\sigma}^2}{2\lambda_-}
\tag{6.4}
$$

after transients of rate $2\lambda_-$.

Lower bound: We show that $\mathbb{E}\|\widetilde{X}_t\|^2$ has a lower bound that can also be expressed in terms of the coupling strength and the noise level. The derivation is similar to that of the upper bound, and we begin with equation 6.1. We set $\lambda \equiv \lambda_+$ and apply the estimates $\lambda_+\|\widetilde{X}_s\|^2 - \langle L\widetilde{X}_s, \widetilde{X}_s \rangle \geq 0$ and $\langle \tanh X_s, \widetilde{X}_s \rangle \leq \|\widetilde{X}_s\|^2$ for all s a.s., yielding

$$
\begin{aligned}
\frac{1}{2}\|\widetilde{X}_t\|^2 &\geq e^{-2\lambda_+ t}\|\widetilde{X}_0\|^2 + \int_0^t e^{2\lambda_+(s-t)}\left(\frac{1}{2}(n-1)\tilde{\sigma}^2 - \|\mathbf{x}\|^2\|\widetilde{X}_s\|^2\right)ds \\
&\quad + \tilde{\sigma}\int_0^t e^{2\lambda_+(s-t)}\|\widetilde{X}_s\|\,dB_s.
\end{aligned}
$$

Taking expectations and integrating the Ito term, we have

$$
\begin{aligned}
\frac{1}{2}\mathbb{E}\|\widetilde{X}_t\|^2 &\geq e^{-2\lambda_+ t}\mathbb{E}\|\widetilde{X}_0\|^2 + \frac{(n-1)\tilde{\sigma}^2}{4\lambda_+}(1 - e^{-2\lambda_+ t}) \\
&\quad - \|\mathbf{x}\|^2\int_0^t e^{2\lambda_+(s-t)}\mathbb{E}\|\widetilde{X}_s\|^2 ds.
\end{aligned}
$$

After transients of rate $2\lambda_-$, we can apply equation 6.4 to estimate the remaining integral and lower-bound the above equation by

$$e^{-2\lambda_+ t}\mathbb{E}\|\widetilde{X}_0\|^2 + \frac{(n-1)\tilde\sigma^2}{4\lambda_+}(1 - e^{-2\lambda_+ t}) - \|\mathbf{x}\|^2 \frac{(n-1)\tilde\sigma^2}{4\lambda_-\lambda_+}(1 - e^{-2\lambda_+ t}).$$

Since $\lambda_- \le \lambda_+$, transients of rate $2\lambda_+$ have already transpired if we suppose that we have waited for transients of rate $2\lambda_-$. Therefore, we can say that after transients of rate $2\lambda_-$,

$$\mathbb{E}\|\widetilde{X}_t\|^2 \ge \frac{(n-1)\tilde\sigma^2}{2\lambda_+}\left(1 - \frac{\|\mathbf{x}\|^2}{\lambda_-}\right). \tag{6.5}$$

6.1.1 Inverting the Change of Variables. Finally, we can obtain corresponding upper and lower bounds for the original system 2.3, noting that since $\widetilde{X}_t = P\big(\mathbf{w}(t)\|\mathbf{x}\|^2 - \langle\mathbf{x}, \mathbf{y}\rangle\mathbb{1}\big) = \|\mathbf{x}\|^2 P\mathbf{w}(t)$, we have $\mathbb{E}\|\tilde{\mathbf{w}}\|^2 = \mathbb{E}\|\widetilde{X}_t\|^2/\|\mathbf{x}\|^4$, where we have used the notation $\tilde{\mathbf{w}}$ for $P\mathbf{w}$. The $\|\mathbf{x}\|^4$ in the denominator then cancels with the same quantity occurring in $\tilde\sigma^2$ in equations 6.4 and 6.5, giving the final form shown in theorem 1.

6.2 Fluctuations Estimates: Proof of Theorem 2. We first derive the fourth moment of the norm of the fluctuations. Starting from equation 6.3, allow transients of rate $2\lambda_-$ to pass so that we are left with the integral inequality

$$\frac{1}{2}\|\widetilde{X}_t\|^2 \le \frac{(n-1)\tilde\sigma^2}{4\lambda_-} + \tilde\sigma \int_0^t e^{2\lambda_-(s-t)}\|\widetilde{X}_s\| dB_s.$$

Squaring both sides, we can apply the identity $(a+b)^2 \le 2a^2 + 2b^2$ to obtain

$$\|\widetilde{X}_t\|^4 \le \left(\frac{(n-1)\tilde\sigma^2}{\sqrt{2}\lambda_-}\right)^2 + 8\tilde\sigma^2\left(\int_0^t e^{2\lambda_-(s-t)}\|\widetilde{X}_s\| dB_s\right)^2.$$

Taking expectations and invoking Ito's isometry for the second term leads to

$$\mathbb{E}\|\widetilde{X}_t\|^4 \le \left(\frac{(n-1)\tilde\sigma^2}{\sqrt{2}\lambda_-}\right)^2 + 8\tilde\sigma^2\int_0^t e^{4\lambda_-(s-t)}\mathbb{E}\|\widetilde{X}_s\|^2 ds$$

$$\le \left(\frac{(n-1)\tilde\sigma^2}{\sqrt{2}\lambda_-}\right)^2 + \frac{8\tilde\sigma^2}{4\lambda_-}\left(\frac{(n-1)\tilde\sigma^2}{2\lambda_-}\right)$$

$$= \left(\frac{(n-1)\tilde\sigma^2}{2\lambda_-}\right)^2\left(2 + \frac{4}{n-1}\right),$$

where the estimate 6.4 has been substituted in for $\mathbb{E}\|\widetilde{X}_s\|^2$. An upper bound on the variance is then obtained from the identity $\mathrm{var}(Z^2) = \mathbb{E}[Z^4] - (\mathbb{E}Z^2)^2$ and the lower estimate given in equation 6.5. Reversing the change of variables as in section 6.1.1 yields the final result.

6.3 Distance to the Noise-Free Trajectory: Proof of Theorem 3. Theorem 1 can be applied providing a lower bound for the average distance between the noisy trajectories of the neural circuit and the noise-free solution to the learning problem. First, observe that from the orthogonal decomposition $X_t = P X_t + Q X_t$ and the change of variables mapping equation 2.3 to 2.4,

$$\|X_t\|^2 = \|\overline{X}_t 1\|^2 + \|\widetilde{X}_t\|^2 = \|\mathbf{x}\|^4 \|\mathbf{w} - w^* 1\|^2. \tag{6.6}$$

Furthermore, we have that

$$\overline{X}_t = n^{-1} \sum_i X_i(t) = n^{-1} \sum_i (w_i \|\mathbf{x}\|^2 - \langle \mathbf{x}, \mathbf{y} \rangle),$$

so evidently $\|\mathbf{x}\|^{-4} \mathbb{E}\overline{X}_t^2 = \mathbb{E}[(\bar{w}_t - w^*)^2]$. Next, note that if the fluctuations are small, the trajectories $(w_i(t))_{i=1}^n$ are close to one another, and the average trajectory $\bar{w}_t = n^{-1}\mathbf{w}(t)^\top 1$ evolves essentially as $\bar{w}_t \sim w^* + \frac{\sigma}{\sqrt{n}} W_t$, where W_t is interpreted as a white noise process. In this case, we then have that $\mathbb{E}[(\bar{w}_t - w^*)^2] = \frac{\sigma^2}{n}$, and we see that $\mathbb{E}[(\bar{w}_t - w^*)^2] \geq \frac{\sigma^2}{n}$ when the fluctuations are not necessarily small. So we have that $\|\mathbf{x}\|^{-4}\mathbb{E}\overline{X}_t^2 \geq \frac{\sigma^2}{n}$. Combining the above with theorem 1,

$$\frac{\sigma^2}{n} + \left[\frac{(n-1)\sigma^2}{2n\lambda_+} \left(1 - \frac{\|\mathbf{x}\|^2}{\lambda_-} \right) \right]^+ \leq \frac{\mathbb{E}\overline{X}_t^2}{\|\mathbf{x}\|^4} + \frac{\mathbb{E}\|\widetilde{X}_t\|^2}{n\|\mathbf{x}\|^4}$$

$$\leq \frac{\sigma^2}{2\lambda_-} + \mathbb{E}[(\bar{w}_t - w^*)^2]$$

with the notation $[\cdot]^+ \equiv \max(0, \cdot)$. Equation 6.6 then shows that the middle quantity above is equal to $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (w_i(t) - w^*)^2 \right]$.

Acknowledgments ⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯⎯

References

Adams, P. (1998). Hebb and Darwin. *Journal of Theoretical Biology, 195*(4), 419–438.

Bennett, M., & Zukin, R. (2004). Electrical coupling and neuronal synchronization in the mammalian brain. *Neuron, 41*(4), 495–511.

Bollobas, B. (2001). *Random graphs* (2nd ed.). Cambridge: Cambridge University Press.

Cao, M., Stewart, A., & Leonard, N. (2010). Convergence in human decision-making dynamics. *Systems and Control Letters, 59*, 87–97.

Chechik, G., Anderson, M., Bar-Yosef, O., Young, E., Tishby, N., & Nelken, I. (2006). Reduction of information redundancy in the ascending auditory pathway. *Neuron, 51*(3), 278–280.

Clay, J. R., & DeHaan, R. L. (1979). Fluctuations in interbeat interval in rhythmic heart-cell clusters: Role of membrane voltage noise. *Biophysical Journal, 28*(3), 377–389.

Croner, L. J., Purpura, K., & Kaplan, E. (1993). Response variability in retinal ganglion cells of primates. *Proceedings of the National Academy of Sciences of the United States of America, 90*(17), 8128–8130.

Doukhan, P. (1994). *Mixing: Properties and examples*. Berlin: Springer-Verlag.

Eliasmith, C., & Anderson, C. (2004). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Enright, J. (1980). Temporal precision in circadian systems: A reliable neuronal clock from unreliable components? *Science, 209*(4464), 1542–1545.

Faisal, A., Selen, L., & Wolpert, D. (2008). Noise in the nervous system. *Nat. Rev. Neurosci., 9*, 292–303.

Fernando, C., Goldstein, R., & Szathmary, E. (2010). The neuronal replicator hypothesis. *Neural Computation, 22*(11), 2809–2857.

Friedrich, R. W., Habermann, C. J., & Laurent, G. (2004). Multiplexing using synchrony in the zebrafish olfactory bulb. *Nature Neurosci., 7*, 862–871.

Fukuda, T., Kosaka, T., Singer, W., & Galuske, R. (2006). Gap junctions among dendrites of cortical Gabaergic neurons establish a dense and widespread intercolumnar network. *J. Neurosci., 26*(13), 3434–3443.

Georgopoulos, A., Kalaska, J., Caminiti, R., & Massey, J. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci., 2*(11), 1527–1537.

Gigante, G., Mattia, M., Braun, J., & Del Giudice, P. (2009). Bistable perception modeled as competing stochastic integrations at two levels. *PLoS Comput Biol., 5*(7), e1000430.

Grammont, F., & Riehle, A. (1999). Precise spike synchronization in monkey motor cortex involved in preparation for movement. *Exp. Brain. Res., 128*, 118–122.

Grandhe, S., Abbas, J., & Jung, R. (1999). Brain-spinal cord interactions stabilize the locomotor rhythm to an external perturbation. *Biomed. Sci. Instrum., 35*, 175–180.

Hahnloser, R., Douglas, R. J., Mahovald, M., & Hepp, K. (1999). Feedback interactions between neuronal pointers and maps for attentional processing. *Nat. Neurosci., 2*, 746–752.

Hertz, J., Krogh, A. S., & Palmer, R. (1991). *Introduction to the theory of neural computation.* Reading, MA: Addison-Wesley.

Hung, C., Kreiman, G., Poggio, T., & DiCarlo, J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310*(5749), 863–866.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci., 2*(3), 194–203.

Jadbabaie, A., Lin, J., & Morse, A. S. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control, 48*(6), 988–1001.

Kandel, E., Schwartz, J., & Jessell, T. (2000). *Principles of neural science* (4th ed.). New York: McGraw-Hill.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science, 324*(5928), 759–764.

Kiemel, T., Gormley, K. M., Guan, L., Williams, T. L., & Cohen, A. H. (2003). Estimating the strength and direction of functional coupling in the lamprey spinal cord. *J. Comput. Neurosci., 15*(2), 233–245.

Kopell, N. (2000). We got rhythm: Dynamical systems of the nervous system. *Notices of the AMS, 47* 6–16.

Kopell, N., & Ermentrout, G. (1986). Symmetry and phase-locking in chains of weakly coupled oscillators. *Communications on Pure and Applied Mathematics, 39*, 623–660.

LeDoux, J. (2000). Emotion circuits in the brain. *Annu. Rev. Neurosci., 23*, 155–184.

Lee, T., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A: Opt. Image. Sci. Vis., 20*(7), 1434–1448.

Lohmiller, W., & Slotine, J. (1998). On contraction analysis for non-linear systems. *Automatica, 34*, 683–696.

Maren, S. (2001). Neurobiology of Pavlovian fear conditioning. *Annu. Rev. Neurosci., 24*, 897–931.

Masuda, N., Kawamura, Y., & Kori, H. (2010). Collective fluctuations in networks of noisy components. *New Journal of Physics, 12*(9), 093007.

Mumford, D. (1992). On the computational architecture of the neocortex—II: The role of cortico-cortical loops. *Biol. Cyb., 66*, 241–251.

Needleman, D. J., Tiesinga, P. H., & Sejnowski, T. J. (2001). Collective enhancement of precision in networks of coupled oscillators. *Physica D: Nonlinear Phenomena, 155*(3–4), 324–336.

Pham, Q.-C., & Slotine, J.-J. (2007). Stable concurrent synchronization in dynamic system networks. *Neural Networks, 20*(1), 62–77.

Pham, Q.-C., Tabareau, N., & Slotine, J.-J. (2009). A contraction theory approach to stochastic incremental stability. *IEEE Transactions on Automatic Control, 54*(4), 816–820.

Pouget, A., & Sejnowski, T. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience, 9*(2), 222–237.

Poulakakis, I., Scardovi, L., & Leonard, N. (2010). Coupled stochastic differential equations and collective decision making in the two-alternative forced-choice task. In *Proc. American Control Conference*. Piscataway, NJ: IEEE.

Puchalla, J. L., Schneidman, E., Harris, R. A., & Berry, M. J. (2005). Redundancy in the population code of the retina. *Neuron, 46*(3), 493–504.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci., 2*, 79–87.

Russo, G., & Slotine, J.-J. (2010). Global convergence of quorum sensing networks. *Physical Review E*, 28.

Schnitzler, A., & Gross, J. (2005). Normal and pathological oscillatory communication in the brain. *Nature Reviews Neuroscience, 6*, 285–296.

Scholz, J., & Schoner, G. (1999). The uncontrolled manifold concept: Identifying control variables for a functional task. *Exp. Brain Res., 126*(3), 289–306.

Sherman, A., & Rinzel, J. (1991). Model for synchronization of pancreatic beta-cells by gap junction coupling. *Biophysical Journal, 59*(3), 547–559.

Slotine, J.-J. E., & Li, W. (1991). *Applied nonlinear control*. Upper Saddle River, NJ: Prentice Hall.

Tabareau, N., Slotine, J.-J., & Pham, Q.-C. (2010). How synchronization protects from noise. *PLoS Comput Biol, 6*(1), e1000637.

Taylor, A., Tinsley, M. R., Wang, F., Huang, Z., & Showalter, K. (2009). Dynamical quorum sensing and synchronization in large populations of chemical oscillators. *Science, 323*(5914), 614–617.

Ting, L. (2007). Dimensional reduction in sensorimotor systems: A framework for understanding muscle coordination of posture. In P. Cisek, T. Drew, & J. Kalaska (Eds.), *Computational neuroscience: Theoretical insights into brain function* (pp. 301–325). New York: Elsevier.

Todorov, E. (2008). *Recurrent neural networks trained in the presence of noise give rise to mixed muscle-movement representations*. Available at http://www.cs.washington.edu/homes/todorov/papers/mixed.pdf.

Varier, S., & Kaiser, M. (2011). Neural development features: Spatio-temporal development of the *Caenorhabditis elegans* neuronal network. *PLoS Computational Biology, 7*(1), 1–9.

Wang, W., & Slotine, J. (2005). On partial contraction analysis for coupled nonlinear oscillators. *Biological Cybernetics, 92*(1), 38–53.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature, 447*(7148), 1075–1080.

Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation, 26*(2), 289–315.

Young, G., Scardovi, L., & Leonard, N. (2010). Robustness of noisy consensus dynamics with directed communication. In *Proc. American Control Conference*. Piscataway, NJ: IEEE.