Synchronization can Control Regularization in Neural Systems via Correlated Noise Processes

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

Jean-Jacques Slotine
Nonlinear Systems Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02138
jjs@mit.edu

Abstract

To learn reliable rules that can generalize to novel situations, the brain must be capable of imposing some form of regularization. Here we suggest, through theoretical and computational arguments, that the combination of noise with synchronization provides a plausible mechanism for regularization in the nervous system. The functional role of regularization is considered in a general context in which coupled computational systems receive inputs corrupted by correlated noise. Noise on the inputs is shown to impose regularization, and when synchronization upstream induces time-varying correlations across noise variables, the degree of regularization can be calibrated over time. The resulting qualitative behavior matches experimental data from visual cortex.

1 Introduction

The problem of learning from examples is in most circumstances ill-posed. This is particularly true for biological organisms, where the "examples" are often complex and few in number, and the ability to adapt is a matter of survival. Theoretical work in inverse problems has long established that regularization restores well-posedness [5, 20] and furthermore, implies stability and generalization of a learned rule [2]. How the nervous system imposes regularization is not entirely clear, however. Bayesian theories of learning and decision making [14, 12, 29] hold that that brain is able to represent prior distributions and assign (time-varying) uncertainty to sensory measurements. By way of a Bayesian integration, the brain may effectively work with hypothesis spaces of limited complexity when appropriate, trading off prior knowledge against new evidence [9]. But while these mechanisms can effect regularization, it is still not clear how to calibrate it: when to cease adaptation or how to fix a hypothesis space suited to a given task. A second possible explanation is that regularization – and a representation of uncertainty – may emerge naturally due to noise. Intuitively, if noise is allowed to "smear" observations presented to a learning apparatus, overfitting may be mitigated – a well known phenomenon in artificial neural networks [1].

In this paper we argue that noise provides an appealing, plausible mechanism for regularization in the nervous system. We consider a general context in which coupled computational circuits subject to independent noise receive common inputs corrupted by spatially correlated noise. Information processing pathways in the mammalian visual cortex, for instance, fall under such an organizational pattern [10, 24, 7]. The computational systems in this setting represent high-level processing stages, downstream from localized populations of neurons which encode sensory input. Noise correlations in the latter arise from, for instance, within-population recurrent connections, shared feed-forward inputs, and common stimulus preferences [24]. Independent noise impacting higher-level computational elements may arise from more intrinsic, ambient neuronal noise sources, and may be roughly independent due to broader spatial distribution [6].

To help understand the functional role of noise in inducing regularization, we propose a high-level model that can explain quantitatively how noise translates into regularization, and how regularization may be calibrated over time. The ability to adjust regularization is key: as an organism accumulates

experience, its models of the world should be able to adjust to the complexity of the relationships and phenomena it encounters, as well as reconcile new information with prior probabilities. Our point of view is complementary to Bayesian theories of learning; the representation and integration of sensory uncertainty is closely related to a regularization interpretation of learning in ill-posed settings. We postulate that regularization may be plausibly controlled by one of the most ubiquitous mechanisms in the brain: synchronization. A simple, one-dimensional regression (association) problem in the presence of both independent ambient noise and correlated measurement noise suffices to illustrate the core ideas.

When a learner is presented with a collection of noisy observations, we show that synchronization may be used to adjust the dependence between observational noise variables, and that this in turn leads to a quantifiable change in the degree of regularization imposed upon the learning task. Regularization is further shown to both improve the convergence rate towards the solution to the regression problem, *and* reduce the negative impact of ambient noise. The model's qualitative behavior coincides with experimental data from visual tracking tasks [10] (area MT) and from anesthetized animals [24] (area V1), in which correlated noise impacts sensory measurements and correlations increase over short time scales. Other experiments involving perceptual learning tasks have shown that noise correlations decrease with long-term training [8]. The mechanism we propose suggests that changes in noise correlations arising from feedback synchronization can calibrate regularization, possibly leading to improved convergence properties or better solutions. Collectively, the experimental evidence lends credence to the hypothesis that, at a high level, the brain may be optimizing its learning processes by adapting dependence among noise variables, with regularization an underlying computational theme.

Lastly, we consider how continuous dynamics solving a given learning problem might be efficiently computed in cortex. In addition to supporting regularization, noise can be harnessed to facilitate distributed computation of the gradients needed to implement a dynamic optimization process. Following from this observation, we analyze a stochastic finite difference scheme approximating derivatives of quadratic objectives. Difference signals and approximately independent perturbations are the only required computational components. This distributed approach to the implementation of dynamic learning processes further highlights a connection between parallel stochastic gradient descent algorithms [25, 15, 28], and neural computation.

2 Learning as noisy gradient descent on a network

The learning process we will consider is that of a one-dimensional linear fitting problem described by a dynamic gradient based minimization of a square loss objective, in the spirit of Rao & Ballard [21]. 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 which govern the environment and which 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" [22]. In doing so, we focus on the underlying computations taking place in the nervous system rather than particular neural representations. The analysis that follows, however, may be extended more generally to multi-layer feedback hierarchies.

To make the setting more concrete, assume that we have observed a set of input-output examples $\{x_i \in \mathbb{R}, y_i \in \mathbb{R}\}_{i=1}^m$, with each x_i representing a generic unit of sensory experience, and want to estimate the linear regression function $f_w(x) = wx$ (we assume the intercept is 0 for simplicity). Adopting the square loss, the total prediction error incurred on the observations by the rule f_w is given by

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

Note that there is no explicit regularization penalty here. We will model adaptation (training) by a noisy gradient descent process on this squared prediction error loss function. The gradient of E with respect to the slope parameter is given by $\nabla_w E = -\sum_{i=1}^m (y_i - wx_i)x_i$, and generates the continuous-time, noise-free gradient dynamics

$$\dot{w} = -\nabla_w E(w). \tag{2}$$

The learning dynamics we will consider, however, are assumed to be corrupted by two distinct kinds of noise:

(N1) Sensory observations $(x_i)_i$ are corrupted by time-varying, correlated noise processes.

(N2) The dynamics are themselves corrupted by additive "ambient" noise.

To accommodate (N1) we will borrow an averaging or, *homogenization*, technique for multi-scale systems of stochastic differential equations (SDEs) that will drastically simplify analysis. We have discussed the origins of (N1) above. The noise (N2) may be significant (we do not take small noise limits) and can be attributed to some or all of: error in computing and sensing a gradient, intrinsic neuronal noise [6] (aggregated or localized), or interference between large assemblies of neurons or circuits.

Synchronization among circuits and/or populations will be modeled by considering multiple coupled dynamical systems, each receiving the same noisy observations. Such networks of systems capture common pooling or averaging computations, and provides a means for studying variance reduction. The *collective enhancement of precision* hypothesis suggests that the nervous system copes with noise by averaging over collections of signals in order to reduce variation in behavior and improve computational accuracy [23, 13, 26, 3]. Coupling synchronizes the collection of dynamical systems so that each tends to a common "consensus" trajectory having reduced variance. If the coupling is strong enough, then the variance of the consensus trajectory decreases as $\mathcal{O}(1/n)$ after transients, if there are n signals or circuits [23, 17, 19, 3]. We will consider regularization in the context of networks of coupled SDEs, and investigate the impact of coupling, redundancy (n) and regularization upon the convergence behavior of the system. Considering networks will allow a more general analysis of the interplay between different mechanisms for coping with noise, however n can be small or 1 in some situations.

Formally, the noise-free flow (2) can be modified to include noise sources (N1) and (N2) as follows. Noise (N1) may be modeled as a white-noise limit of Ornstein-Uhlenbeck (OU) processes $(\mathbf{Z}_t)_i$, and (N2) as an additive diffusive noise term. In differential form, we have

$$dw_t = -\big(w_t\|\mathbf{x}+\mathbf{Z}_t\|^2 - \langle \mathbf{x}+\mathbf{Z}_t, \mathbf{y}\rangle\big)dt + \sigma dB_t \tag{3a}$$

$$dZ_t^i = -\frac{Z_t^i}{\varepsilon}dt + \frac{\sqrt{2}\gamma}{\sqrt{\varepsilon}}dB_t^i, \qquad i=1,\dots,m. \tag{3b}$$

Here, B_t denotes the standard 1-dimensional Brownian motion and captures noise source (N2). The observations $(\mathbf{x})_i = x_i$ are corrupted by the noise processes $(\mathbf{Z}_t)_i = Z_t^i$, following (N1). For the moment, the Z_t^i are independent, but we will relax this assumption later. The parameter $0 < \varepsilon \ll 1$ controls the correlation time of a given noise process. In the limit as $\varepsilon \to 0$, Z_t^i may be viewed as a family of independent zero-mean Gaussian random variables indexed by t. Characterizing the noise $\mathbf{Z}_t$ as (3b) with $\varepsilon \to 0$ serves as both a modeling approximation/idealization and an analytical tool.

2.1 Homogenization

The system (3a)-(3b) above is a classic "fast-slow" system: the gradient descent trajectory w_t evolves on a timescale much longer than the $\mathcal{O}(\varepsilon)$ stochastic perturbations $\mathbf{Z}_t$. Homogenization considers the dynamics of w_t after averaging out the effect of the fast variable $\mathbf{Z}_t$. In the limit as $\varepsilon \to 0$ in (3b), the solution to the averaged SDE converges (in a sense to be discussed below) to the solution of the original SDE (3a).

The following Theorem is an instance of [18, Thm. 3], adapted to the present setting.

Theorem 2.1. *Let $0 < \varepsilon \ll 1$, $\sigma, \gamma > 0$ and let $\mathcal{X}, \mathcal{Y}$ denote finite-dimensional Euclidean spaces. Consider the system*

$$dx = f(x,y)dt + \gamma dW_t, \qquad x(0) = x_0 \tag{4a}$$

$$dy = \varepsilon^{-1}g(y)dt + \varepsilon^{-1/2}\sigma dB_t, \qquad y(0) = y_0, \tag{4b}$$

where $x \in \mathcal{X}, y \in \mathcal{Y}$, and $W_t \in \mathcal{X}, B_t \in \mathcal{Y}$ are independent multivariate Brownian motions. Assume that for all $x \in \mathcal{X}, y \in \mathcal{Y}$ the following conditions on (4) *hold:*

$$\begin{aligned} \langle g(y), y/\|y\|\rangle &\le -r\|y\|^\alpha, \\ \|f(x,y) - f(x',y)\| &\le C(y)\|x-x'\| \\ \|f(x,y)\| &\le K(1+\|x\|)(1+\|y\|^q), \end{aligned}$$

with $r > 0, \alpha \ge 0, q < \infty$, and where $C(y)$ is a constant depending on y. If the SDE (4b) *is ergodic, then there exists a unique invariant measure μ_∞ characterizing the probability distribution of y_t in*

the steady state, and we may define the vector field $F(x) \triangleq \mathbb{E}_{\mu_\infty}[f(x,y)] = \int_{\mathcal{Y}} f(x,y)\mu_\infty(dy)$. *Furthermore,* $x(t)$ *solving* (4a) *is closely approximated by* $X(t)$ *solving*

$$dX = F(X)dt + \gamma dW_t, \quad X(0) = x_0$$

in the sense that, for any $t \in [0,T]$, $x(t) \Rightarrow X(t)$ *in* $C([0,T],\mathcal{X})$ *as* $\varepsilon \to 0$.

It may be readily shown that the system (3) satisfies the conditions of Theorem 2.1. Moreover, the OU process (3b) on $\mathbb{R}^m$ is known to be ergodic with stationary distribution $\mathbf{Z}_\infty \sim \mathcal{N}(\mathbf{0}, \gamma^2 I)$ (see e.g. [11]), where $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian distribution with mean μ and covariance Σ. Averaging over the fast variable $\mathbf{Z}_t$ appearing in (3a) with respect to this distribution gives

$$dw_t = -\big[w_t(\|\mathbf{x}\|^2 + m\gamma^2) - \langle \mathbf{x}, \mathbf{y}\rangle\big]dt + \sigma dB_t, \tag{5}$$

and by Theorem 2.1, we can conclude that Equation (5) well-approximates (3a) when $\varepsilon \to 0$ in (3b) in the sense of weak convergence of probability measures.

2.2 Network structure

Now consider $n \geq 1$ diffusively coupled neural systems implementing the dynamics (5), with associated parameters $\mathbf{w}(t) = \big(w_1(t), \ldots, w_n(t)\big)$. If $W_{ij} \geq 0$ is the coupling strength between systems i and j, $L = \text{diag}(W\mathbf{1}) - W$ is the *network Laplacian* [16]. We assume here that L is symmetric and defines a connected network graph. Letting $\alpha := \|\mathbf{x}\|^2 + m\gamma^2$, $\beta := \langle \mathbf{x}, \mathbf{y}\rangle$ and $\boldsymbol{\mu} := (\beta/\alpha)\mathbf{1}$, the coupled system can be written concisely as

$$\begin{aligned} d\mathbf{w}_t &= -(L + \alpha I)\mathbf{w}_t dt + \beta \mathbf{1} dt + \sigma d\mathbf{B}_t \\ &= (L + \alpha I)(\boldsymbol{\mu} - \mathbf{w}_t)dt + \sigma d\mathbf{B}_t\,, \end{aligned} \tag{6}$$

with $\mathbf{B}_t$ an n-dimensional Brownian motion. The diffusive couplings here should be interpreted as modeling abstract intercommunication 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. Note that one can equivalently consider n systems (3a) and then homogenize assuming n copies of the same noise process $\mathbf{Z}_t$, *or* n independent noise processes $\{\mathbf{Z}_t^{(i)}\}_i$; either choice also leads to (6).

3 Learning with noisy data imposes regularization

Equation (6) is seen by inspection to be an OU process, and has solution (see e.g. [11])

$$\mathbf{w}(t) = e^{-(L+\alpha I)t}\mathbf{w}(0) + \big(I - e^{-(L+\alpha I)t}\big)\boldsymbol{\mu} + \sigma \int_0^t e^{-(L+\alpha I)(t-s)} d\mathbf{B}_s. \tag{7}$$

Integrals of Brownian motion are normally distributed, so $\mathbf{w}(t)$ is a Gaussian process and can be characterized entirely by its time-dependent mean and covariance, $\mathbf{w}(t) \sim \mathcal{N}\big(\boldsymbol{\mu}_w(t), \Sigma_w(t)\big)$. A straightforward manipulation (details omitted due to lack of space) gives

$$\begin{aligned} \boldsymbol{\mu}_w(t) :&= \mathbb{E}[\mathbf{w}(t)] = e^{-(L+\alpha I)t}\,\mathbb{E}[\mathbf{w}(0)] + \big(I - e^{-(L+\alpha I)t}\big)\boldsymbol{\mu} \\ \Sigma_w(t) :&= \mathbb{E}\left[\big(\mathbf{w}(t) - \mathbb{E}\,\mathbf{w}(t)\big)\big(\mathbf{w}(t) - \mathbb{E}\,\mathbf{w}(t)\big)^\top\right] \\ &= e^{-(L+\alpha I)t}\,\mathbb{E}[\mathbf{w}(0)\mathbf{w}(0)^\top]e^{-(L+\alpha I)t} + \frac{\sigma^2}{2}(L + \alpha I)^{-1}\big(I - e^{-2(L+\alpha I)t}\big). \end{aligned} \tag{8}$$

The solution to the noise-free regression problem (minimizing (1)) is given by $w^* = \langle \mathbf{x}, \mathbf{y}\rangle / \|\mathbf{x}\|^2$, however (7) together with (8) reveals that, for any $i \in \{1, \ldots, n\}$,

$$\mathbb{E}[w_i(t)] \xrightarrow{t\to\infty} (\boldsymbol{\mu})_i = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|^2 + m\gamma^2} \tag{9}$$

which is exactly the solution to the *regularized* regression problem

$$\min_{w\in\mathbb{R}} \|\mathbf{y} - w\mathbf{x}\|^2 + \lambda w^2$$

with regularization parameter $\lambda := m\gamma^2$. To summarize, we have considered a network of coupled, noisy gradient flows implementing unregularized linear regression. When the observations $\mathbf{x}$ are noisy, all elements of the network converge in expectation to a common equilibrium point representing a regularized solution to the original regression problem.

3.1 Convergence behavior

In the previous section we showed that the network converges to the solution of a regularized regression problem, but left open a few important questions: What determines the convergence rate? How does the noise (N1),(N2) impact convergence? How does coupling and redundancy (number of circuits n) impact convergence? How do these quantities affect the variance of the error? We can address these questions by decomposing $\mathbf{w}(t)$ into orthogonal components, $\mathbf{w}(t) = \bar{w}(t)\mathbf{1} + \widetilde{\mathbf{w}}(t)$, representing the mean-field trajectory $\bar{w} = \frac{1}{n}\mathbf{1}^\top\mathbf{w}$, and fluctuations about the mean $\widetilde{\mathbf{w}} = \mathbf{w} - \bar{w}\mathbf{1}$. We may then study the error

$$\mathbb{E}\big[\tfrac{1}{n}\|\mathbf{w}(t) - \boldsymbol{\mu}\|^2\big] = \mathbb{E}\big[\tfrac{1}{n}\|\widetilde{\mathbf{w}}(t)\|^2\big] + \mathbb{E}\big[\tfrac{1}{n}\|\bar{w}(t)\mathbf{1} - \boldsymbol{\mu}\|^2\big] \tag{10}$$

by studying each term separately. Decomposing the error into fluctuations about the average and the distance between the average and the noise-free equilibrium allows one to see that there are actually two different convergence rates governing the system: one determines convergence towards the synchronization subspace (where $\widetilde{\mathbf{w}} = 0$), and the another determines convergence to the equilibrium point $\boldsymbol{\mu}$. The following result provides quantitative answers to the questions posed above:

Theorem 3.1. *Let $\widetilde{C}, \overline{C}$ be constants which do not depend on time, and let $\underline{\lambda}$ denote the smallest non-zero eigenvalue of L. Set $\alpha := \|\mathbf{x}\|^2 + m\gamma^2$ and $\boldsymbol{\mu} := (\langle \mathbf{x}, \mathbf{y}\rangle/\alpha)\mathbf{1}$, as before. Then for all $t > 0$,*

$$\mathbb{E}\big[\tfrac{1}{n}\|\mathbf{w}(t) - \boldsymbol{\mu}\|^2\big] \leq \widetilde{C}e^{-2(\underline{\lambda}+\alpha)t} + \overline{C}e^{-2\alpha t} + \frac{\sigma^2}{2}\left(\frac{1}{\underline{\lambda}+\alpha} + \frac{1}{\alpha n}\right). \tag{11}$$

A proof is given in the supplementary material. The first term of (11) estimates the transient part of the fluctuations term in (10), and we find that the rate of convergence to the synchronization subspace is $2(\underline{\lambda} + \alpha)$. The second term term estimates the transient part of the centroid's trajectory, and we see that the rate of convergence of the mean trajectory to equilibrium is 2α. In the presence of noise, however, the system will neither synchronize nor reach equilibrium exactly. After transients, we see that the residual error is given by the last term in (11). This term quantifies the steady-state interaction between: gradient noise (σ); regularization (α, via the observation noise γ); network topology (via $\underline{\lambda}$), coupling strength (via $\underline{\lambda}$), and redundancy (n; possibly $\underline{\lambda}$).

3.2 Discussion

From the results above we can draw a few conclusions about networks of noisy learning systems:

1. Regularization improves both the synchronization rate and the rate of convergence to equilibrium.
2. Regularization contributes towards *reducing* the effect of the gradient noise σ: (N1) counteracts (N2).
3. Regularization *changes the solution*, so we cannot view regularization as a "free-parameter" that can be used solely to improve convergence or reduce noise. Faster convergence rates and noise reduction should be viewed as beneficial side-effects, while the appropriate degree of regularization primarily depends on the learning problem at hand.
4. The number of circuits n and the coupling strength contribute towards reducing the effect of the gradient noise (N2) (that is, the variance of the error) and improve the synchronization rate, but do not affect the rate of convergence toward equilibrium.
5. Coupling strength and redundancy *cannot* be used to control the degree of regularization, since the equilibrium solution $\boldsymbol{\mu}$ does not depend on n or the spectrum of L. This is true no matter how the coupling weights W_{ij} are chosen, since constants will always be in the null space of L and $\boldsymbol{\mu}$ is a constant vector.

In the next section we will show that if the noise processes $\{Z_t^i\}_i$ are *themselves* trajectories of a coupled network, then synchronization *can* be a mechanism for controlling the regularization imposed on a learning process.

4 Calibrating regularization with synchronization

If instead of assuming independent noise processes corrupting the data as in (3b), we consider correlated noise variables $(Z_t^i)_{i=1}^m$, it is possible for synchronization to control the regularization which the noise imposes on a learning system of the form (3a). A collection of dependent observational noise processes is perhaps most conveniently modeled by coupling the OU dynamics (3b) introduced

before through another (symmetric) network Laplacian L_z:

$$d\mathbf{Z}_t = -\frac{1}{\varepsilon}(L_z + \eta I)\mathbf{Z}_t dt + \frac{\sqrt{2}\gamma}{\sqrt{\varepsilon}} d\mathbf{B}_t, \tag{12}$$

for some $\eta > 0$. We now have two networks: the first network of gradient systems is the same as before, but the observational noise process $\mathbf{Z}_t$ is now generated by another network. For purposes of analysis, this model suffices to capture generalized correlated noise sources. In the actual biology, however, correlations may arise in a number of possible ways, which may or may not include diffusively coupled dynamic noise processes.

To analyze what happens when a network of learning systems (3a) is driven by observation noise of the form (12), we take an approach similar to that of the previous Section. The first step is again homogenization. The system (12) may be viewed as a zero-mean variation of (6), and its solution $\mathbf{Z}_t \sim \mathcal{N}\big(\boldsymbol{\mu}_z(t), \Sigma_z(t)\big)$ is a Gaussian process characterized by

$$\boldsymbol{\mu}_z(t) = e^{-(L_z+\eta I)t/\varepsilon}\,\mathbb{E}[\mathbf{Z}(0)] \tag{13a}$$

$$\Sigma_z(t) = e^{-(L_z+\eta I)t/\varepsilon}\,\mathbb{E}[\mathbf{Z}(0)\mathbf{Z}(0)^\top]e^{-(L_z+\eta I)t/\varepsilon} + \gamma^2(L_z+\eta I)^{-1}\big(I - e^{-2(L_z+\eta I)t/\varepsilon}\big). \tag{13b}$$

Taking $t \to \infty$ in (13) yields the stationary distribution $\mu_\infty = \mathcal{N}\big(\mathbf{0}, \gamma^2(L_z+\eta I)^{-1}\big)$. We can now consider (3a) defined with $\mathbf{Z}_t$ governed by (12), and average with respect to μ_∞:

$$\begin{aligned} dw_t &= -\,\mathbb{E}_{\mu_\infty}\Big\{\big(w_t\|\mathbf{x}+\mathbf{Z}_t\|^2 - \langle \mathbf{x}+\mathbf{Z}_t, \mathbf{y}\rangle\big)\Big\}dt + \sigma dB_t \\ &= -\Big[w_t\big(\|\mathbf{x}\|^2 + \gamma^2\,\mathrm{tr}(L_z+\eta I)^{-1}\big) - \langle \mathbf{x}, \mathbf{y}\rangle\Big]dt + \sigma dB_t \end{aligned}$$

where we have used that $\mathbb{E}[\|\mathbf{Z}_t\|^2] = \gamma^2\,\mathrm{tr}(L_z+\eta I)^{-1}$. As before, the averaged approximation is good when $\varepsilon \to 0$. An expression identical to (6),

$$d\mathbf{w}_t = (L + \alpha I)(\boldsymbol{\mu} - \mathbf{w}_t)dt + \sigma d\mathbf{B}_t \tag{14}$$

is obtained by redefining $\alpha := \|\mathbf{x}\|^2 + \gamma^2\,\mathrm{tr}(L_z+\eta I)^{-1}$ and $\boldsymbol{\mu} := (\langle \mathbf{x}, \mathbf{y}\rangle/\alpha)\mathbf{1}$. In this case,

$$\lambda = \alpha - \|\mathbf{x}\|^2 = \gamma^2\,\mathrm{tr}(L_z+\eta I)^{-1}.$$

Theorem 3.1 may be immediately applied to understand (14). As before, the covariance of $\mathbf{Z}_t$ figures into the regularization parameter. However now the covariance of $\mathbf{Z}_t$ is a function of the network Laplacian $L_z = L_z(t)$, which is defined by the topology and potentially *time-varying* coupling strengths of the noise network. By adjusting the coupling in (12), we adjust the regularization λ imposed upon (14). When coupling increases, the dependence among the Z_t^i increases and $\mathrm{tr}(L_z+\eta I)^{-1}$ (and therefore α) decreases. Thus, *increased correlation among observational noise variables implies decreased regularization*.

In the case of all-to-all coupling with uniform strength $\kappa \geq 0$, for example, L_z has eigenvalues $0 = \lambda_0 < \lambda_1 = \cdots = \lambda_m = m\kappa$. The regularization may in this case range over the interval

$$\inf_\kappa \mathrm{tr}(L_z+\eta I)^{-1} = \frac{1}{\eta} < \frac{\lambda}{\gamma^2} \leq \frac{m}{\eta} = \sup_\kappa \mathrm{tr}(L_z+\eta I)^{-1}$$

by adjusting the coupling strength $\kappa \in [0, \infty)$. Note that all-to-all coupling may be plausibly implemented with $\mathcal{O}(n)$ connections using mechanisms such as *quorum sensing* (see [3, §2.3], [27]).

5 Distributed computation with noise

We have argued that noise can serve as a mechanism for regularization. Noise may also be harnessed, in a different sense, to compute dynamics of the type discussed above. The distributed nature of the mechanism we will explore adheres to the general theme of parallel computation in the brain, and provides one possible explanation for how the gradients introduced previously might be estimated. The development is closely related to stochastic gradient descent (SGD) ideas appearing in stochastic approximation [25, 15] and adaptive optics [28].

5.1 Parallel stochastic gradient descent

Let $J(\mathbf{u}) : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz Lyapunov cost functional we wish to minimize with respect to some set of control signals $\mathbf{u}(t) \in \mathbb{R}^d$. Gradient descent on J can be described by the collection of flows

$$\frac{du_i(t)}{dt} = -\gamma\frac{\partial J}{\partial u_i}(u_1, \ldots, u_d), \qquad i = 1, \ldots, d.$$

We consider the case where the gradients above are estimated via finite difference approximations of the form

$$\frac{\partial J(\mathbf{u})}{\partial u_i} \approx \frac{J(u_1, \ldots, u_i + \delta u_i, \ldots, u_d) - J(u_1, \ldots, u_i, \ldots, u_d)}{\delta u_i},$$

where δu_i is a small perturbation applied to the i-th input. *Parallel stochastic gradient descent* (PSGD, see e.g. [28]) involves applying i.i.d. stochastic perturbations δu_i simultaneously to all inputs in parallel, so that the gradients $\partial_i J(\mathbf{u})$ are estimated as

$$\frac{\partial J(\mathbf{u})}{\partial u_i} \approx \delta J \delta u_i, \qquad i = 1, \ldots, d \tag{15}$$

where $\delta J = J(u_1 + \delta u_1, \ldots, u_i + \delta u_i, \ldots, u_d + \delta u_d) - J(u_1, \ldots, u_i, \ldots, u_d)$. If δu_i are symmetric random variables with mean zero and variance σ^2, then $\sigma^{-2}\,\mathbb{E}[\delta J \delta u_i]$ is accurate to $\mathcal{O}(\sigma^2)$ [28].

5.2 Stochastic gradient model

The parallel finite difference approximation (15) suggests a more biologically plausible mechanism for implementing gradient dynamics. If the perturbations δu_i are taken to be Gaussian i.i.d. random variables, we can model parallel stochastic gradient descent as an Ito process:

$$d\mathbf{u}_t = -\gamma\big[J(\mathbf{u}_t + \mathbf{Z}_t) - J(\mathbf{u}_t)\big]\mathbf{Z}_t dt, \qquad \mathbf{u}(0) = u_0 \tag{16a}$$

$$d\mathbf{Z}_t = -\frac{1}{\varepsilon}\mathbf{Z}_t dt + \frac{\sigma}{\sqrt{\varepsilon}} d\mathbf{B}_t, \qquad \mathbf{Z}(0) = z_0 \tag{16b}$$

where $\mathbf{B}_t$ is a standard d-dimensional Brownian motion. Additive noise affecting the gradient has been omitted from (16a) for simplicity, and does not change the fundamental results discussed in this section. The perturbation noise $\mathbf{Z}_t$ has again been modeled as a white-noise limit of Ornstein-Uhlenbeck processes (16b). When $\varepsilon \to 0$, Equation (16a) implements PSGD using the approximation given by Equation (15) with δu_i zero-mean i.i.d. Gaussian random variables.

We will proceed with an analysis of (16) in the particular case where J is chosen from the quadratic family of cost functionals of the form $J(\mathbf{u}) = \mathbf{u}^\top A \mathbf{u}$ where A is a symmetric, bounded and strictly positive definite matrix[1]. In this setting the analysis is simpler and suffices to illustrate the main points. This cost function satisfies $\min_{\mathbf{u}\in\mathbb{R}^d} J(\mathbf{u}) = 0$ with minimizer $\mathbf{u}^* = 0$, and J is a Lyapunov function. Equation (16a) now takes the form

$$d\mathbf{u}_t = -\gamma\big(2\mathbf{u}_t^\top A \mathbf{Z}_t + \mathbf{Z}_t^\top A \mathbf{Z}_t\big)\mathbf{Z}_t dt, \qquad \mathbf{u}(0) = u_0. \tag{17}$$

5.3 Convergence of continuous-time PSGD with quadratic cost

We turn to studying the convergence behavior of (17) and the precise role of the stochastic perturbations $\mathbf{Z}_t$ used to estimate the gradients. These perturbations must be small in order to obtain accurate approximations of the gradients. However, one may also expect that the noise will play an important role in determining convergence properties since it is the noise that ultimately kicks the system "downhill" towards equilibrium. Homogenizing (17) with respect to $\mathbf{Z}_t$ leads to the following Theorem, the proof of which is given in the supplementary material.

Theorem 5.1. *For any $0 \le t \le T < \infty$, the solution $\mathbf{u}(t)$ to (17) satisfies*

$$\lim_{\varepsilon \to 0} \mathbb{E}[\mathbf{u}(t)] = e^{-\gamma\sigma^2 A t}\mathbf{u}(0). \tag{18}$$

It is clear from this result that the PSGD system (16), for $\varepsilon \to 0$, converges in expectation globally and exponentially to the minimum of J when J is a positive definite quadratic form. Our earlier intuition that the perturbation noise σ should play a role in the rate of convergence is also confirmed: greater noise amplitudes lead to faster convergence. However this comes at a price. The covariance of $\mathbf{u}(t)$ after transients is exactly the covariance of $\mathbf{Z}_t$. Thus an inherent tradeoff between speed and accuracy must be resolved by any organism implementing PSGD-like mechanisms.

[1] Without loss of generality we may assume A is symmetric since the antisymmetric part does not contribute to the quadratic form. In addition, objectives of the form $u^\top A u + b^\top u + c$ may be expressed in the homogeneous form $u^\top A u$ by a suitable change of variables.

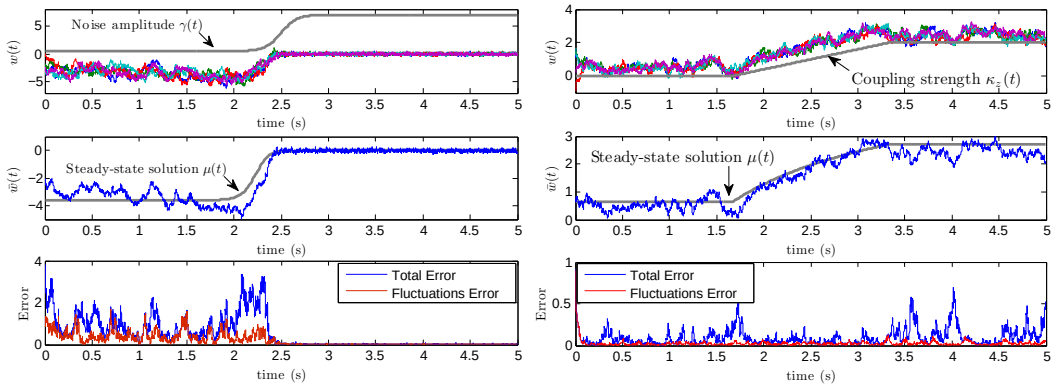

Figure 1: *(Left stack) Increased observation noise imposes greater regularization, and leads to a reduction in ambient noise. (Right stack) Stronger coupling/correlation between observation noise processes decreases regularization. See text for details.*

6 Simulations

We first simulated a network of gradient dynamics with uncoupled observation noise processes obeying (3). To illustrate the effect of increasing observation noise variance, the parameter γ in (3b) was increased from 0.5 to 7 along a monotonic, sigmoidal path over the duration of the simulation. We used $n = 5$ systems (3a) with $\sigma = 4$, coupled all-to-all with uniform strength $\kappa = 2$. Observations were sampled according to $(\mathbf{x})_i \sim \mathcal{N}(0, 0.04)$, $(\mathbf{y})_i \sim \text{Uniform}[0, 20]$ with $m = 20$ entries, once and for all, at the beginning of the experiment. Initial conditions were drawn according to $\mathbf{w}(0) \sim \text{Uniform}[-3, 3]$, and $\mathbf{Z}(0)$ was set to 0. Figure 1 (left three plots) verifies some of main conclusions of Section 3.2. The top plot shows the sample paths $\mathbf{w}(t)$ and time course of the observational noise deviation $\gamma(t)$ (grey labeled trace). When the noise increases near $t = 2.5s$, a dramatic drop in the variance of $\mathbf{w}(t)$ is visible. The middle plot shows the center of mass (mean-field) trajectory $\bar{w}(t)$ superimposed upon the time-varying noise-free solution $\mu(t)$ (gray labeled trace). Because the observation noise is increasing, the regularization $\lambda = m\gamma^2$ increases and the solution $\mu(t)$ to the regularized problem decreases in magnitude following (9). The bottom plot shows the mean-squared distance to the time-dependent noise-free solution $\boldsymbol{\mu}(t)$, and the mean-squared size of the fluctuations about the centroid $\bar{w}^2$. It is clear that the error rapidly drops off when $\gamma(t)$ increases, confirming the apparent reduction in the variance of $\mathbf{w}(t)$ in the top plot.

A second experiment, described by the right-hand stack of plots in Figure 1, shows how synchronization can function to adjust regularization over time. This simulation is inspired by the experimental study of noise correlations in cortical area MT due to [10], where it was suggested that time-varying correlations between pairs of neurons play a significant role in explaining behavioral variation in smooth-pursuit eye movements. In particular, the findings in [10] and [4] suggest that short-term increases in noise correlations are likely to occur after feedback arrives and neurons within and upstream from MT synchronize. We simulated a collection of correlated observation noise processes obeying (12) ($\varepsilon = 10^{-3}$, $\eta = 3$) with all-to-all topology and uniform coupling strength $\kappa_z(t)$ increasing from 0 to 2 along the profile shown in Figure 1 (top-right plot, labeled gray trace). This noise process $\mathbf{Z}_t$ was then fed to a population of $n = 5$ units obeying (3a), with ambient noise $\sigma = 1$ and all-to-all coupling at fixed strength $W_{ij} = \kappa = 2$. New data $\mathbf{x}, \mathbf{y}$ and initial conditions were chosen as in the previous experiment. The middle plot on the right-hand side shows the effect of increasing synchronization among the observation noise processes. As the coupling increases, the noise becomes more correlated and regularization decreases. This in turn causes the desired solution $\mu(t)$ to the regression problem to increase in magnitude (labeled gray trace). With decreased regularization, the ambient noise is more pronounced. The bottom-right plot shows the mean fluctuation size and distance to the noise-free solution (total error). An increase in the noise variance is apparent following the increase in observational noise correlation.

[2]These quantities are similar to those defined in (10), but represent only this single simulation – not in expectation. Here, ergodic theory allows one to (very roughly) infer ensemble averages by visually estimating time averages.

Acknowledgments
The authors are grateful to Rodolfo Llinas for pointing out the plausible analogy between gradient search in adaptive optics and learning mechanisms in the brain. JB was supported under DARPA FA8650-11-1-7150 SUB#7-3130298, NSF IIS-08-03293 and WA State U. SUB#113054 G002745.

References

[1] C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[2] O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2(3):499–526, 2002.

[3] J. Bouvrie and J.-J. Slotine. Synchronization and redundancy: Implications for robustness of neural learning and decision making. *Neural Computation*, 23(11):2915–2941, 2011.

[4] S. C. de Oliveira, A. Thiele, and K. P. Hoffmann. Synchronization of neuronal activity during stimulus expectation in a direction discrimination task. *J Neurosci.*, 17(23):9248–60, 1997.

[5] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, 1996.

[6] A. Faisal, L. Selen, and D. Wolpert. Noise in the nervous system. *Nat. Rev. Neurosci.*, 9:292–303, April 2008.

[7] T. J. Gawne and B. J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J Neurosci.*, 13(7):2758–71, 1993.

[8] Y. Gu, S. Liu, C. R. Fetsch, Y. Yang, S. Fok, A. Sunkara, G. C. DeAngelis, and D.E. Angelaki. Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron*, 71(4):750 – 761, 2011.

[9] T. D. Hanks, M. E. Mazurek, R. Kiani, E. Hopp, and M. N. Shadlen. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *J. Neurosci.*, 31(17):6339–52, 2011.

[10] X. Huang and S. G. Lisberger. Noise correlations in cortical area MT and their potential impact on trial-by-trial variation in the direction and speed of smooth-pursuit eye movements. *J. Neurophysiol*, 101:3012–3030, 2009.

[11] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

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

[13] T. Kinard, G. De Vries, A. Sherman, and L. Satin. Modulation of the bursting properties of single mouse pancreatic β-cells by artificial conductances. *Biophysical Journal*, 76(3):1423–1435, 1999.

[14] K. P. Körding and D. M. Wolpert. Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7):319–326, 2006.

[15] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

[16] M. Mesbahi and M. Egerstedt. *Graph Theoretic Methods in Multiagent Networks*. Princeton U. Press, 2010.

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

[18] E. Pardoux and A. Yu. Veretennikov. On the Poisson equation and diffusion approximation. I. *Annals of Probability*, 29(3):1061–1085, 2001.

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

[20] T. Poggio and S. Smale. The mathematics of learning: dealing with data. *Notices Amer. Math. Soc.*, 50(5):537–544, 2003.

[21] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2:79–87, 1999.

[22] A. Schnitzler and J. Gross. Normal and pathological oscillatory communication in the brain. *Nature Reviews Neuroscience*, 6:285–296, 2005.

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

[24] M. A. Smith and A. Kohn. Spatial and temporal scales of neuronal correlation in primary visual cortex. *J Neurosci.*, 28(48):12591–12603, 2008.

[25] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:332–341, 1992.

[26] N. Tabareau, J.-J. Slotine, and Q.-C. Pham. How synchronization protects from noise. *PLoS Comput Biol*, 6(1):e1000637, Jan 2010.

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

[28] M. A. Vorontsov, G. W. Carhart, and J. C. Ricklin. Adaptive phase-distortion correction based on parallel gradient-descent optimization. *Opt. Lett.*, 22(12):907–909, Jun 1997.

[29] T. Yang and M. N. Shadlen. Probabilistic reasoning by neurons. *Nature*, 447(7148):1075–1080, 2007.