*LETTER* Communicated by Anke Meyer-Base

# Timescale Separation in Recurrent Neural Networks

**Thomas Flynn**
*tflynn@gradcenter.cuny.edu*
*Graduate Center, City University of New York, New York, NY 10016, U.S.A.*

**Supervised learning in recurrent neural networks involves two processes: the neuron activity from which gradients are estimated and the process on connection parameters induced by these measurements. A problem such algorithms must address is how to balance the relative rates of these activities so that accurate sensitivity estimates are obtained while still allowing synaptic modification to take place at a rate sufficient for learning. We show how to calculate a sufficient timescale separation between these two processes for a class of contracting neural networks.**

## 1 Introduction

Given a fixed input, the neural networks studied here compute by convergence to a steady state. The output of the network is taken to be this fixed point or some simple readout thereof. It is possible to implement backpropagation-like algorithms, which take into account the recurrent nature of the network. Early works that explored this include Atiya (1988), Pineda (1989), and Almeida (1987). A desirable property of such algorithms is that they operate locally in time and space. *Local in time* means the algorithm uses only information along the trajectory of the network, as opposed to requiring asymptotic data (such as the fixed point of the system) at each step. Spatial locality requires that during training, information must be transferred using the existing network topology. In a neural network, this means the process taking place on a parameter along a particular edge requires information only from the two nodes incident at that edge. Such a process, which we review here, can be derived naturally from the equations describing the gradient in a recurrent network and comprises two activities: accumulating sensitivities along the trajectory of the network and a flow on the parameters. In the absence of dynamics on the weights, this process simply computes the sensitivities, and one would expect that accurate estimates can still be obtained if the weights change slowly enough (alternatively, that the sensitivity accumulation happens fast enough). If the weights change too quickly, the sensitivity estimates may not be able to keep up and the optimization will fail. Thus, controlling the relative rates of these two activities is crucial for ensuring these optimization schemes work.

*Neural Computation* **27, 1321–1344** (2015) 
doi:10.1162/NECO_a_00740

We show under reasonable assumptions on the network architecture that finite and fixed rates can be chosen so that this localized learning scheme has the same long-time behavior as the gradient system associated with the network, which from the distributed perspective requires infinite power and nonlocal communication. The assumptions are that the network obeys a stability condition known as contraction uniformly for all parameter values and that the various first and second derivatives of the activity at each node are bounded. The main technical tools we use are stability results applicable to contracting systems and hierarchical combinations of such systems. The optimization problem is approached as a gradient system with error, and the timescale is chosen so as to control this error.

The assumption of contraction plays a key role in our results. This stability criterion, which says roughly that any two trajectories of a system converge toward each other, has been fruitfully applied to problems in systems biology and the analysis of networks of dynamical elements generally. This has been used to study, for instance, phenomena such as synchronization (Wang & Slotine, 2005) and entrainment to periodic signals (Russo, Di Bernardo, & Sontag, 2010). Here we show how it can be applied in the context of the distributed gradient-based optimization of such networks.

The main theoretical tool used in previous work to analyze this problem was singular perturbation theory. The basic idea is to analyze the process on the weights and the forward motion of the network separately, according to their natural timescales. If certain stability criteria are met in each of these cases, then the stability of the original system can be concluded when the weights are adapted slowly enough (or the forward motion happens fast enough), that is, using an appropriate timescale separation. Using general results for systems with multiple timescales, Riaza and Zufiria (2003) analyzed a gradient-based supervised scheme from this perspective. In this situation, the stable points of the parameter space are local minima of an objective function. They showed local stability of the system around such points, again for a sufficient choice of the timescale separation. Those authors also looked at the problem in Riaza & Zufiria (2002), where they also noted that the problem of learning a periodic trajectory could be reduced to a set of fixed-point learning problems. A supervised training scheme with a different learning rule was also analyzed by Tseng and Šiljak (1995), who reached a similar conclusion. That work also investigated conditions that guarantee the neuron dynamics remain stable as the connection weights are modified during training, which is important because in general, dynamics on the weights may destroy stability. The issue of timescales in supervised learning was also noted in Atiya (1988), Pineda (1989), and Baldi (1995), where heuristics were proposed for choosing it in practice.

In the above cited theoretical analyses, the essential soundness of the schemes is established only when the procedure begins near a local minimum of the objective function. In addition, previous work did not attempt to quantify this requirement or the bound on the timescale in terms of

information that might be available about the model. Here, the specific set of assumptions and the tools we develop allows one to conclude that the optimization will work under much milder conditions on the initial state of the system. It is necessary only that the neuron activity be close enough to its stable point before the optimization begins. This initial requirement and a sufficient timescale parameter are quantified in terms of information that may be known, such as bounds on derivatives and the rate of convergence of neuron dynamics. Given enough information about the model, we show how to bootstrap the system so that the procedure can be carried out from any starting point.

In this letter, we focus on supervised learning, where one has a specific objective on the networks' behavior. A related line of research investigates the stability of Hebbian learning in recurrent networks. Dong and Hopfield (1992) show stability by constructing a Lyapunov function on the joint system consisting of the forward activity of the network and a Hebbian process on the weights and use this to show that the total system converges to a local fixed point. Singular perturbation methods were applied to an unsupervised learning rule in Meyer-Baese, Ohl, & Scheich (1996).

Similar considerations apply to probabilistic neural networks. For instance, in order to calculate gradients in the Boltzmann machine, it is necessary to compute expectations over the stationary distribution of a Markov chain. Hence, the applicability of a multiple timescale approach, consisting of a Markov chain from which statistics are gathered and a comparatively slower weight adaptation process. Some theoretical aspects of such algorithms were analyzed in Younes (1999). Yuille (2005) noted a connection between this class of stochastic optimization procedures and the contrastive divergence algorithm for training restricted Boltzmann machines. This algorithm was shown to give favorable results on some machine learning tasks in Tieleman (2008). A variant of this procedure also plays an important role in training algorithms for deep Boltzmann machines (Salakhutdinov & Hinton, 2009). Neal (1992) considered the problem for sigmoid belief networks, which are directed acyclic models, and suggested a two timescale approach using persistent Markov chains for these models as well.

## 2 Optimization System

In this section we introduce the optimization system, which consists of the forward and adjoint system and the dynamics on the weights. The derivation here is somewhat standard (see also Baldi, 1995). We start with the system whose asymptotic behavior we wish to optimize:

$$\dot{x} = f(x, w, \rho). \tag{2.1}$$

Here $x \in \mathbb{R}^n$ is the state of the neural network, $w$ are the parameters, and $\rho$ is some input vector that remains fixed while the network is running. For

instance, in the case of neural networks typically used in machine learning, the $f$ might be $f(x, w, \rho) = -x + \sigma(wx + \rho)$, where $\sigma$ is a sigmoid function applied component-wise. For now we drop the notation indicating the input $\rho$. We assume that $f$ has a unique fixed point for each choice of the parameter vector $w$, and that the system tends toward this point regardless of the starting configuration, in a specific sense that we identify in the next section. We want to design a flow on the parameters $w$ of $f$ that finds a local minimum of some objective function on the equilibrium point. Let $x^*(w)$ give the equilibrium point of equation 2.1, that is, the solution of the equation

$$f(x^*(w), w) = 0, \tag{2.2}$$

and let $g$ be a function that calculates fitness of points in the phase space. To minimize or at least find a stationary point of $g(x^*(w))$, a natural process would be the gradient system,

$$\dot{w} = -D_w(g \circ x^*). \tag{2.3}$$

According to the implicit function theorem, $x^*$ will be differentiable at $w$ so long as $D_x f$ is invertible at the fixed point. Differentiating equation 2.2 and using the chain rule, one can then express the $w$ dynamics as

$$\dot{w} = \left[ \left(D_x g\right) \left(D_x f\right)^{-1} \left(D_w f\right) \right]_{(x^*(w), w)}, \tag{2.4}$$

where the derivatives are understood to be evaluated at the point $(x^*(w), w)$. As is, this system poses problems from both the distributed and computational points of view. This is due to the presence of terms involving the equilibrium point $x^*(w)$ and the inverse of $D_x f$. That is, it is local in neither time nor space. We introduce a new variable $y$ with dynamics

$$\dot{y} = y \left(D_x f\right)_{(x^*(w), w)} - \left(D_x g\right)_{(x^*(w))}, \tag{2.5}$$

which essentially solves the equation $y(D_x f) = (D_x g)$ in a distributed fashion. Taking $y^* = y^*(w)$ to be the fixed point of this system we can rewrite equation 2.4 as

$$\dot{w} = y^* \left(D_w f\right)_{(x^*(w), w)}. \tag{2.6}$$

So far this equation defines the same flow on $w$ as equation 2.3 does. Now we consider the localized system consisting of $x$, $y$, and $w$, where we use only

information along the trajectories of $x$ and $y$, as opposed to their equilibrium points:

$$\tau \dot{x} = f(x, w), \tag{2.7a}$$

$$\tau \dot{y} = y \left(D_x f\right)_{(x,w)} - (D_x g)_{(x)}, \tag{2.7b}$$

$$\dot{w} = y \left(D_w f\right)_{(x,w)}. \tag{2.7c}$$

Here we have introduced two separate timescales via the parameter $\tau$. Note that by setting $\tau = 0$, equation 2.7c becomes equation 2.6. These three equations define the distributed optimization system. As Riaza and Zufiria (2003) noted, this can also be viewed as a three-timescale algorithm by introducing separate rates $\tau_x$ and $\tau_y$ on the corresponding subsystems. For simplicity, we calculate a single rate that suffices for both systems. We refer to the individual systems 2.7a and 2.7b as the forward and adjoint systems, respectively. At times we refer to $z = (x, y)$ as the joint system.

Our goal in the analysis of the system is to show that under proper conditions, this flow does decrease the error and that it does so at a sufficient pace that its long-time behavior matches that of equation 2.3. This can be ensured by keeping $z$ close enough to $z^*(w)$, in a sense we make precise in section 4. In the next section, we specify the stability properties we require of the forward system and give conditions under which this extends to the joint system. We then formalize the condition that the weight process 2.7c should satisfy in section 4, taking the view that it is a gradient system with error. In section 5 we present a general result that provides timescales for this type of condition, and in section 6 we verify that this is applicable to the optimization system. We demonstrate several neural network architectures to which these results apply in section 7.

## 3 Stability Criteria

Our focus is on systems that converge to a fixed point. To be precise, we focus on systems having the *contraction* property. To define contraction, we first recall the notion of a matrix measure. Fix a vector norm $\|\cdot\|$ on $\mathbb{R}^n$. The matrix measure induced by this norm is the real-valued function $\mu$ defined on $n \times n$ matrices such that

$$\mu(A) = \lim_{h \to 0^+} \frac{\|I + hA\| - 1}{h},$$

where the norm in the numerator is the induced matrix norm, and this limit is taken as $h \to 0$ from the right. The matrix measure, also known as the logarithmic norm, is well defined for all matrices and any vector norm. A

proof of this fact and other properties of matrix measures is in Vidyasagar (2002) and Desoer and Haneda (1972). For $\beta > 0$, the system $\dot{z} = u(z, w(t))$ is said to be contracting with rate $\beta$ when

$$\mu\left(D_z u\right) \leq -\beta \tag{3.1}$$

for all $z$ and $t$. In nonautonomous systems, this condition guarantees that trajectories started from different locations converge toward each other (see Lohmiller & Slotine, 1998, and Sontag, 2010). Precisely, letting $z_1(t), z_2(t)$ be any two solutions of the system $\dot{z} = u(z, w(t))$ corresponding to different initial conditions, one has

$$\|z_1(t) - z_2(t)\| \leq \|z_1(0) - z_2(0)\| e^{-\beta t} \tag{3.2}$$

in the norm in which the system is contracting. In autonomous systems, the contraction property guarantees the existence of and convergence to a unique equilibrium point. In the context of neural networks, if the system defined by equation 2.1 obeys a contraction condition for all inputs, then the network will converge to a unique fixed point for each input.

An important feature of contraction is that the property is preserved under various types of system combinations. The following result, from Sontag (2010, theorem 3), regards hierarchies of contracting systems:

**Theorem 1.** *Consider a hierarchy of contracting systems of the form*

$$\dot{x} = f(x),$$
$$\dot{y} = g(y, x),$$

*where $x$ is contracting with rate $\beta_x$ in the norm $\|\cdot\|_X$, $y$ is contracting with rate $\beta_y$ in the norm $\|\cdot\|_Y$, and $\|D_x g\|_{X,Y} \leq k$ for some $k$. Then for any positive numbers $p_1$, $p_2$ such that $\beta_x - \frac{p_2}{p_1} k > 0$, the joint system $(x, y)$ is contracting with rate $\beta$ in the norm $\|(x, y)\| = p_1\|x\|_X + p_2\|y\|_Y$ where $\beta = min\,\{\beta_x - \frac{p_2}{p_1} k, \beta_y\}$.*

We will show that the $y$ system, equation 2.7b, is contracting whenever the $x$ system, equation 2.7a, is, and that theorem 1 may be applied to conclude that the joint system $(x, y)$ is contracting. This will allow us to carry $(x, y)$ as a single system $z$ in the analysis of the learning procedure. Specifically, the next result shows that the $y$ system is contracting in the dual norm $\|x\|_* = \sup_{\|y\|=1} |x^T y|$.

**Proposition 1.** *If the $x$ system, equation 2.7a, is contracting with rate $\beta$ in a given norm $\|\cdot\|$ then the $y$ system, equation 2.7b, is contracting with rate $\beta$ in the dual norm $\|\cdot\|_*$*

**Proof.** It is easily seen that the Jacobian of the adjoint system, equation 2.7b is $(D_x f)^T$, that is, the transpose of the Jacobian of the forward system, equation 2.7a. It is well known that for any vector norm, the induced matrix norm satisfies $\|A^T\|_* = \|A\|$. Then for any number $h$, $\|I + hA^T\|_* = \|I + hA\|$ and therefore $\mu_*(A^T) = \mu(A)$, where $\mu_*$ is the matrix measure with respect to the dual norm. In particular $\mu_*((D_x f)^T) = \mu(D_x f)$.

As an example, in section 7 we consider neural networks contracting in the $\infty$-norm $\|x\|_\infty = \max_i \|x_i\|$. In this case, proposition 1 is easy to verify: the matrix norm is $\|A\|_\infty = \max_i\{\sum_j |A_{i,j}|\}$, and the measure is given by

$$\mu(A)_\infty = \max_i \left\{ A_{i,i} + \sum_{j \neq i} |A_{i,j}| \right\}.$$

The dual norm is $\|x\|_{\infty^*} = \|x\|_1 = \sum_i |x_i|$. The matrix norm and measure are given by $\|A\|_1 = \max_j\{\sum_i |A_{i,j}|\}$ and

$$\mu(A)_1 = \max_j \left\{ A_{j,j} + \sum_{i \neq j} |A_{i,j}| \right\},$$

from which it is immediate that $\mu(A)_\infty = \mu(A^T)_1$. We note that when speaking of a dual norm, as in the norm $\|v\|_{X^*}$ of a vector in $\mathbb{R}^n$, this is meant in the sense of a particular norm on the vector space $\mathbb{R}^n$. And for a matrix $M \in \mathbb{R}^{n \times m}$ and an expression like $\|M\|_{X^*,Y}$, we are referring to the norm of $M$ as a linear map between the vector spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ with the norms $\|\cdot\|_{X^*}$ and $\|\cdot\|_Y$, respectively.

The assumptions on the network dynamics $f$ and the loss function $g$ we use are that the first and second derivatives are uniformly bounded for all values of the parameter vector $w$. These properties may be verified in any norm that is convenient or by just showing that all (mixed) partials are bounded. However, the specific timescale $\tau$ that we obtain is phrased in terms of bounds on these quantities with respect to specific norms that we identify throughout.

The $y$ system is a linear time-varying system, and the following bound applies:

**Proposition 2.** *Assume $x$ is contracting and $\|D_x g\|_* \leq (L_x g)$ for some $L_x g$. Then $\|y\|_* \leq \frac{L_x g}{\beta}$ is forward invariant for $y$.*

**Proof.** Say $\|y\|_* > \frac{L_x g}{\beta}$. Then $\|y(t) + h(D_t y(t))\|_* \leq \|I + h(D_x f)^T\|_* \|y\|_* + h\|D_x g\|_*$ and therefore $D_t^+ \|y\|_* \leq \mu(D_x f)\|y\|_* + \|D_x g\|_*$. Under the assumption on $\|y\|_*$, $D_t^+ \|y\|_* < 0$.

We are now in a position to show that the joint system is contracting. Technically the system is contracting on the set $\mathbb{R}^n \times B_*(\frac{L_x g}{\beta})$ where $B_*(r) = \{y \in \mathbb{R}^n | \|y\|_* \leq r\}$ is the ball of a given radius in the norm $\|\cdot\|_*$. The proof is provided in the appendix.

**Proposition 3.** *Assume $\|D_x^2 g\|_{X,X^*} \leq L_{x^2} g$, $\|D_x^2 f\| \leq L_{x^2} f$, and that the conditions of proposition 2 hold. Let $\beta_x$ be the contraction rate of the forward system, and let $p_1$, $p_2$ be two positive numbers such that $\beta_x - \frac{p_2}{p_1} k > 0$ where $k = \frac{(L_x g)}{\beta_x}(L_{x^2} f) + (L_{x^2} g)$. Then the joint system, equation 2.7a and 2.7b, is contracting with rate $\beta$ in the norm $\|\cdot\|_Z$ on the set $U$ where $\beta = \beta_x - \frac{p_2}{p_1} k$, $\|(x, y)\|_Z = p_1\|x\| + p_2\|y\|_*$, and $U = \mathbb{R}^n \times B_*(\frac{L_x g}{\beta_x})$.*

This next result summarizes two important facts about contracting systems that we use later. The first is how the fixed point $z^*$ changes with the parameter $w$. The second is how $\|u\|$, which may be thought of as a measure of energy for a contracting system, changes when there is some dynamics on the parameter $w$. A proof is in the appendix.

**Proposition 4.** *Assume $\dot{z} = u(z, w(t))$ is contracting with rate $\beta$ in $\|\cdot\|_Z$ and the function $u$ satisfies $\|D_w u\|_{W,Z} \leq (L_w u)$ for a norm $\|\cdot\|_W$. Then $\|D_w z^*\|_{W,Z} \leq \frac{L_w u}{\beta}$ and $D_t^+ \|u\|_Z \leq -\beta\|u\|_Z + (L_w u)\|\dot{w}\|_W$.*

## 4 Optimization Criteria

Here we formalize the property that trajectories of equation 2.7, should satisfy, taking the view that it is a gradient system with a certain type of error. We seek to apply the following result, a proof of which is in the appendix:

**Theorem 2.** *Let $v(z) : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function such that $D_z v$ is Lipschitz continuous and $v$ is bounded from below. Consider the perturbed gradient system*

$$\dot{z} = -(D_z v)_{(z)} + u(t)$$

*where*

$$\|u(t)\|_2 \leq \alpha \|(D_z v)_{(z(t))}\|_2$$

*for $\alpha < 1$. Then $v(z(t))$ converges, and $(D_z v)_{(z(t))} \to 0$.*

**Corollary 1.** *Let $v$ be as in theorem 2 and consider the system*

$$\dot{z} = -u(t)$$

*where*

$$\|(D_z v)_{(z(t))} - u(t)\|_2 \leq \alpha \|u(t)\|_2$$

*for $\alpha < \frac{1}{2}$. Then $v(z(t))$ converges, and $(D_z v)_{(z(t))} \to 0$.*

**Proof.** We can write $-u(t) = -(D_z v) + e(t)$ where $e(t) = (D_z v) - u(t)$, and it is easily seen that $\|e(t)\|_2 \leq \frac{\alpha}{1-\alpha}\|D_z v\|_2$. Since $\alpha < \frac{1}{2}$, $e(t)$ satisfies the conditions of theorem 2.

Let $h = h(x, y, w) = y(D_w f)_{(x,w)}$ and $h^* = h(x^*(w), y^*(w), w)$. With corollary 1 in mind and noting that $h^*$ is the negative of the gradient of the error function $g(x^*(w))$, we aim to satisfy the condition

$$\|h - h^*\|_2 \leq \alpha \|h\|_2 \tag{4.1}$$

for some $0 \leq \alpha < \frac{1}{2}$. Such conditions or closely related criteria are commonly used in discrete gradient descent schemes (see Bertsekas & Tsitsiklis, 2000). By itself, condition 4.1 guarantees descent of the objective function, and when the other conditions of corollary 1 are satisfied, convergence can be concluded as well. We now reformulate this condition so that we may use information about the online behavior of the joint system $z$ that is available in the norm $\|\cdot\|_Z$. Norms in Euclidean space being equivalent, a sufficient condition for equation 4.1 is that

$$\|h - h^*\|_W \leq \frac{\alpha}{k}\|h\|_W$$

for some $k$ such that

$$\frac{1}{\sqrt{k}}\|\cdot\|_W \leq \|\cdot\|_2 \leq \sqrt{k}\|\cdot\|_W. \tag{4.2}$$

When $D_z h$ is bounded, a sufficient condition for equation 4.1 is then

$$\|z - z^*\|_Z \leq \frac{\alpha}{k(L_z h)}\|h\|_W, \tag{4.3}$$

where $L_z h$ is such that $\|D_z h\|_{Z,W} \leq L_z h$. In any contracting system, we have

$$\beta\|z_1 - z_2\|_Z \leq \|u(z_1) - u(z_2)\|_Z \tag{4.4}$$

for two states $z_1, z_2$, where $\beta$ is the contraction rate (Söderlind, 1984). In particular, $\beta\|z - z^*\|_Z \leq \|u(z)\|_W$, so it suffices that

$$\|u\|_Z \leq \frac{\alpha\beta}{k(L_z h)}\|h\|_W. \tag{4.5}$$

In the next section, we show that this condition can be satisfied for all $\alpha$ below a certain threshold by choosing a corresponding rate $\tau$ small enough.

## 5 Timescale Selection

One approach to the optimization problem is to view the system 2.7 as a singularly perturbed version of equation 2.3 and then apply standard results for such systems. This has been used in the context of Hebbian learning in Meyer-Baese et al. (1996) and in the supervised case in Riaza and Zufiria (2003). Riaza and Zufiria (2003) investigate the behavior of a "semirelaxed" form of the learning problem, equation 2.4, in which only a temporal relaxation is introduced, and show that near a joint equilibrium of the total system $(x, w)$, one can conclude convergence of the relaxed learning algorithm to this local minimum, for an adequate choice of the timescale parameter. Using the strong assumption of contraction, we are able to have weaker conditions on the optimization problem and the starting conditions. We do not assume that the original gradient system, equation 2.3, obeys any particular stability criteria, only that it is Lipschitz continuous. Our condition on the starting point of the algorithm, stated in theorem 3, is that the joint system $(x, y)$ system starts sufficiently close to its equilibrium point. This is always possible to do by running the system for a warm-up period, since when there is no optimization happening, the joint system is convergent. Del Vecchio and Slotine (2013) revisited some topics of singular perturbation theory from the perspective of contraction theory. For instance, their results may be applied to conclude that $\tau$ and the starting point may be chosen so that $z$ remains within an arbitrarily small radius around $z^*$ for all time.

As noted above, the specific property we require of the trajectory allows $z$ to be far from $z^*$ when the gradient is large, but requires that the distance go to zero as the gradient does. The theorem we now prove is a general result for this type of condition, and in the next section, we apply it to the distributed optimization system.

**Theorem 3.** *Consider a system of the form*

$$\tau\dot{z} = u(z, w)$$
$$\dot{w} = h(z, w)$$

*where $z$ is contracting with rate $\beta$ in the norm $\|\cdot\|_Z$ and $\|\cdot\|_W$ is a norm on $w$ such that $\|D_w u\|_{W,Z} \le L_w u$, $\|D_z h\|_{Z,W} \le L_z h$ and $\|D_w h\|_W \le L_w h$. For any $0 < \alpha < 1$, if*

$$c = \frac{\alpha\beta}{(L_z h)}$$

*and*

$$\tau \le \frac{(1-\alpha)\beta c}{c(L_w h) + (L_w u)},$$

*then $c\|h(z,w)\|_W \ge \|u(z,w)\|_Z$ is forward invariant for the system $(z,w)$.*

**Proof.** Let $H = \|h\|_W$, $U = \|u\|_Z$, and $B = cH - U$. We show that $D_t^+ B > 0$ when $B < 0$. Note that the conditions of proposition 4 are satisfied so that

$$D_t^+ U \le -\frac{\beta}{\tau} U + (L_w u) H.$$

On the other hand, since in general $D_t^+ \|h(t)\|_W \ge -\|D_t h(t)\|_W$, we have

$$D_t^+ H \ge -\frac{(L_z h)}{\tau} U - (L_w h) H.$$

Combining these, it follows that for $B$,

$$D_t^+ B \ge c\left(-\frac{(L_z h)}{\tau} U - (L_w h) H\right) - (L_w u) H + \frac{\beta}{\tau} U.$$

When $B < 0$, we have $cH < U$. Using this fact and rearranging terms, we obtain

$$D_t^+ B > U\left(-(L_w h) - \frac{(L_w u)}{c} + \frac{1}{\tau}(\beta - c(L_z h))\right) \ge 0,$$

where the last inequality follows by assumption on $c$ and $\tau$.

## 6 Application to the Optimization System

To show that the above result may be applied to the distributed optimization problem, it suffices to exhibit the bounds $L_w u$, $L_z h$, $L_w h$, and also show that it is possible to choose $c$ small enough so that condition 4.5 may be satisfied. The latter problem is trivial once we show that $D_z h$ is bounded, since

the theorem provides a timescale for arbitrarily *small* values of the parameter $c$. Demonstrating the bounds on these derivatives is a straightforward application of the assumptions, and we defer the proofs to the appendix.

**Proposition 5.** *Assume the conditions of proposition 3. Let* $\|(x, y)\|_Z = p_1\|x\| + p_2\|y\|_*$ *be the norm provided by that result, let* $\beta_x$ *be the contraction rate of the forward system, and further assume there are numbers* $L_w f$, $L_{x,w} f$, *and* $L_{w^2} f$ *such that* $\|D_w f\|_{W^*,X} \leq L_w f$, $\|D^2_{x,w} f\|_{X,W^*,X} \leq L_{x,w} f$, *and* $\|D^2_w f\|_{W,W^*,X} \leq L_{w^2} f$. *Then* $\|D_z h\|_{Z,W} \leq max\{\frac{(L_x g)}{\beta_x}\frac{(L_{x,w} f)}{p_1}, \frac{(L_w f)}{p_2}\}$ *and* $\|D_w h\|_W \leq \frac{(L_x g)}{\beta_x}(L_{w^2} f)$.

Using this, constraint 4.5 with $L_z h$ equal to the bound on $\|D_z h\|_{Z,W}$ given by proposition 5 is sufficient to ensure descent during the learning process.

The next result says the other conditions of the proposition are also true: $D_w u$ and $D_z u$ are uniformly bounded.

**Proposition 6.** *Assume the conditions of proposition 3. Let* $\|(x, y)\|_Z = p_1\|x\| + p_2\|y\|_*$ *be the norm provided by that result, let* $\beta_x$ *be the contraction rate of the forward system, and further assume that* $\|D_w f\|_{W,X} \leq L_w f$, $\|D^2_{x,w}\|_{X,W,X} \leq L_{x,w} f$, $\|D_x f\|_{X,X} \leq L_x f$, *and* $\|D^2_x f\|_{X,X,X} \leq L_{x^2} f$ *for some numbers* $L_w f$, $L_{x,w} f$, $L_x f$, *and* $L_{x^2} f$. *Then* $\|D_z u\|_Z \leq (L_x f) + \frac{(L_x g)}{\beta_x}\frac{p_2}{p_1}(L_{x^2} f)$ *and* $\|D_w u\|_{W,Z} \leq p_1(L_w f) + p_2(L_{x,w} f)\frac{L_x g}{\beta_x}$.

The following theorem summarizes the result for the optimization system 2.7. We write $A_u v$ for the assumption that the derivative $D_u v$ is uniformly bounded.

**Theorem 4.** *Assume that the forward system 2.7a is contracting, and properties* $A_x f$, $A_{x^2} f$, $A_w f$, $A_{w^2} f$, $A_{x,w} f$, $A_x g$, *and* $A_{x^2} g$. *Then there are* $0 < \alpha < \frac{1}{2}$ *and* $\tau > 0$ *such that the optimization system 2.7 verifies the descent condition 4.1 along the whole trajectory, given suitable initial conditions on the joint system* $(x, y)$. *If, in addition,* $g$ *is bounded from below, then* $(g \circ x^*)(w(t))$ *converges and* $D_w(g \circ x^*)_{(w(t))} \to 0$.

**Proof.** The conditions of proposition 3 are satisfied so one may construct a norm $\|\cdot\|_Z$ in which the joint system $z = (x, y)$ is contracting, with contraction rate $\beta$. The conditions of propositions 5 and 6 are also satisfied, so that the bounds on $D_w u$, $D_z h$, and $D_w h$ exist. Then by theorem 3, for all $c$ sufficiently small, there is a rate $\tau$ so that $\|u\|_Z \leq c\|h\|_W$ is forward invariant for the optimization system, 2.7. In particular we can choose $c$ small enough so that

$$c < \frac{\beta}{2k(L_z h)},$$

where $k$ is defined as in equation 4.2 and the descent condition, equation 4.1, will be satisfied. Note that this is equivalent to selecting $\alpha < \frac{1}{2k}$ in theorem 3.

Next we show that $D_w(g \circ x^*)(w)$ is Lipschitz continuous. Let $L_w h$, $L_z h$ be the bounds on $\|D_w h\|_W$ and $\|D_z h\|_{Z,W}$, respectively, provided by proposition 5. Note that $-D_w(g \circ x^*)(w) = h(x^*(w), y^*(w), w) = h(z^*(w), w)$. For any two $w_1$, $w_2$, we have $\|h(z^*(w_1), w_1) - h(z^*(w_1), w_2)\|_W \leq (L_w h)\|w_1 - w_2\|_W$ and

$$\begin{aligned}\|h(z^*(w_1), w_2) - h(z^*(w_2), w_2)\|_W &\leq (L_z h)\|z^*(w_1) - z^*(w_2)\|_Z \\ &\leq (L_z h)K\|w_1 - w_2\|_W,\end{aligned}$$

where $K$ is the bound on $D_w z^*$ given by proposition 4. Combining these,

$$\|h(z^*(w_1), w_1) - h(z^*(w_2), w_2)\| \leq ((L_w h) + (L_z h)K)\|w_1 - w_2\|_W.$$

Therefore, corollary 1 can be applied.

Finally, we address how to bootstrap the optimization system. This is the problem of finding some initial $(x, y)$ that verify equation 4.5 for a given $\alpha$. This can be done by running the joint system with $w$ fixed until condition 4.5 is satisfied. The left- and right-side terms can be measured at run time if the norms $\|\cdot\|_W$ and $\|\cdot\|_Z$ can be computed and the other constants are available. So long as $h^* \neq 0$, this inequality must eventually be verified, since $h \to h^*$ and $u \to 0$ exponentially fast. After this, the optimization can be started.

## 7 Neural Networks

We now verify whether the above assumptions hold for various neural network architectures. These models use a sigmoid function that is bounded and has bounded derivatives. These can be, for instance, the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$ or the hyperbolic tangent $\sigma_H(x) = \tanh(x) = \frac{1-e^{-2x}}{1+e^{-2x}}$. We assume the existence of three numbers $L_\sigma$, $L_{\sigma'}$, and $L_{\sigma''}$ such that $|\sigma(x)| \leq L_\sigma$, $|\sigma'(x)| \leq L_{\sigma'}$, and $|\sigma''(x)| \leq L_{\sigma''}$. The error function $g$ should have $\|D_x g\|$ and $\|D^2_{xg}\|$ bounded.

Various conditions have been identified that ensure convergence to a unique equilibrium point in neural networks. Atiya (1988) shows global stability when the sum of squares of the weights is small. Matrix measure conditions like the ones referred to here were also explored in Fang and Kincaid (1996). Many of the conditions can be enforced by selecting a suitable norm for the weight matrix and requiring that the weight matrix is not too large when measured this way. For instance, consider the model where

Table 1: Stability Conditions for the Network $\dot{x} = f(x, w) = -x + \sigma(wx + \rho)$ for Different Choices of the Vector Norm.

| Vector Norm | $\mu(D_x f)$ | Stability Criteria |
|---|---|---|
| $\|\cdot\|_2$ | $\max_i\{\lambda_i(-I + \frac{1}{2}(\sigma'(u)w + (\sigma'(u)w)^T))\}$ | $\|w\|_2 < \frac{1}{L_{\sigma'}}$ |
| $\|\cdot\|_1$ | $-1 + \max_j\{\sum_i \sigma'(u_i)\|w_{i,j}\|\}$ | $\|w\|_1 < \frac{1}{L_{\sigma'}}$ |
| $\|\cdot\|_\infty$ | $-1 + \max_i\{\sigma'(u_i)\sum_j \|w_{i,j}\|\}$ | $\|w\|_\infty < \frac{1}{L_{\sigma'}}$ |

Note: In the first row, $\sigma'(u)$ refers to the diagonal matrix whose $(j, j)$ entry is $\sigma'(u_j)$. Note that the third column uses a matrix norm.

unit $i$ obeys

$$\dot{x}_i = -x_i + \sigma\left(\sum_j x_j w_{i,j} + \rho_i\right), \tag{7.1}$$

where $\rho$ is an input vector and we assume $w_{i,i} = 0$. When we write $u_i = \sum_j x_j w_{i,j} + \rho_i$, the Jacobian of this system is

$$(D_x f)_{i,j} = -\delta_{i,j} + \sigma'(u_i) w_{i,j}.$$

From here, one may obtain different conditions on the weight matrix that ensure convergence to a unique equilibrium point by choosing different norms on the state $x$. Several examples of this are summarized in Table 1. Note that any feedforward networks satisfying equation 7.1 may be seen to be contracting by recursive application of theorem 1.

To determine whether the results in this letter are applicable to a given model, it is necessary to verify the uniform stability in a particular norm, while the uniform boundedness of the maps $D^i_x$, $D^i_w f$ for $i = 1, 2$ and $D^2_{w,x} f$ may be verified in any norm. As mentioned above, to actually set the rates and verify the initial conditions, it is necessary to have bounds in particular norms.

**Example 1.** The simplest case to analyze is a "fixed feedback" network. The parameters have been partitioned into a matrix $v$ holding the weights between internal nodes $x$ and a matrix $w$ holding the weights between the internal nodes and the input $\rho$. We assume $v_{i,i} = 0$. The forward dynamics are

$$\dot{x} = f(x, w) = -x + \sigma(vx + w\rho). \tag{7.2}$$

Table 2: Derivatives and Bounds for the Recurrent Network, Equation 7.2.

| Derivative | Values | Norm Required | Upper Bound |
|---|---|---|---|
| $(D_w f)_{i,(i,k)}$ | $\sigma'(u_i)\rho_k$ | $\|D_w f\|_{F,2}$ | $L_{\sigma'}\|\rho\|_2$ |
| $(D^2_{x,w} f)_{i,j,(i,r)}$ | $\sigma''(u_i)v_{i,j}\rho_r$ | $\|D^2_{x,w} f\|_{2,F,2}$ | $L_{\sigma''}\|\rho\|_2\|v\|_F$ |
| $(D^2_w f)_{i,(i,k),(i,r)}$ | $\sigma''(u_i)\rho_k\rho_r$ | $\|D^2_w f\|_{F,F,2}$ | $L_{\sigma''}\|\rho\|_2^2$ |
| $(D^2_x f)_{i,j,k}$ | $\sigma''(u_i)v_{i,j}v_{i,k}$ | $\|D^2_x f\|_{2,2,2}$ | $L_{\sigma''}\|v\|_F^2$ |

The system will be contracting for $\|v\|$ small enough in some norm, since $\sigma'$ is bounded and $\rho$ is fixed. Optimization takes place on the parameters $w$. It can be verified that all first and second derivatives with respect to $x$ and $w$, along with the mixed partials $D^2_{x,w} f$, are bounded uniformly. These are shown in Table 2. This shows the distributed optimization procedure is feasible for this model.

For this example, we describe in detail how one may choose a suitable pair of the timescale $\tau$ and the parameter $\alpha$, which controls how small the error in the gradient is. We start with the following assumptions:

- The 2-norm is used on the network state $x$, and the Frobenius norm is used for weight matrix $w$, that is, $\|w\|_F = (\sum_i \sum_j (w_{i,j})^2)^{\frac{1}{2}}$.
- The feedback matrix $v$ has $\|v\|_F < \frac{1}{L_{\sigma'}}$, so that the network is contracting in the 2-norm with rate $\beta_x = 1 - L_{\sigma'}\|v\|_F$.
- The error function is $g(x) = \frac{1}{2}\|x - t\|_2^2$ for some target vector $t \in [0,1]^n$ where $n$ is the number of nodes in the network. Then $\|D_x g\| = \|x - t\| \leq \sqrt{n}$ and $\|D^2_x g\| = \|I\| = 1$.

This choice of norm on the weight matrix has the advantage that it is self-dual, $\|w\|_{F^*} = \|w\|_F$, so there are fewer bounds to compute. For instance, in general, one needs bounds on both $\|D_w f\|_{W,X}$ and $\|D_w f\|_{W^*,X}$ (to compute the results of propositions 5 and 6, respectively), but these coincide when using the Frobenius norm. It is also consistent with the 2-norm in the sense that $\|wx\|_2 \leq \|w\|_F \|x\|_2$, for a vector $x$, which simplifies computations of the bounds. Table 2 lists bounds on the derivatives of $f$ that are needed for computing the timescale parameter $\tau$. To show how these bounds are computed, a proof is given in the appendix for the bounds on the first and second derivatives with respect to $w$. The two other bounds (on $D^2_{w,x} f$ and $D^2_x f$) may be computed similarly.

**Proposition 7.** *Consider the neural network defined by equation 7.2. The derivatives of f in this case satisfy* $\|D_w f\|_{F,2} \leq L_{\sigma'}\|\rho\|_2$ *and* $\|D^2_w f\|_{F,F,2} \leq L_{\sigma''}\|\rho\|_2^2$.

Given this, the rest of the constants may be computed as follows:

- Define a norm and compute the contraction rate on the joint system 2.7a and 2.8b:
  Define $K = \frac{L_x g}{\beta_x} L_{x^2} + L_{x^2} g$.
  Choose $p_1, p_2 > 0$ such that $\beta_x - \frac{p_2}{p_1} K > 0$.
  Define $\beta = \beta_x - \frac{p_2}{p_1} K$.
- Compute bounds provided by propositions 5 and 6:
  Define $L_z h = \max \left\{ \frac{L_x g}{\beta_x} \frac{L_{x,w} f}{p_1}, \frac{L_w f}{p_2} \right\}$.
  Define $L_w h = \frac{L_x g}{\beta_x} L_{w^2} f$.
  Define $L_w u = p_1 (L_w f) + p_2 (L_{x,w} f) \frac{L_x g}{\beta_x}$.
- Compute a timescale parameter:
  Choose $0 \leq \alpha < \frac{1}{2}$.
  Define $c = \frac{\alpha \beta}{L_z h}$.
  Define $\tau = \frac{(1-\alpha)\beta c}{c L_w h + L_w u}$.

There are a number of choices to be made when calculating the timescale parameter. Beyond the choice of norms $\| \cdot \|_X$ and $\| \cdot \|_W$, one must also decide on the coefficients $p_1, p_2$ for the norm $\| \cdot \|_Z$ and the parameter $\alpha$, which controls in the error in the gradient. Intuitively, a small $\alpha$ means less error in the gradient, and therefore a better guarantee in the rate of descent of the objective $g(x^*)$, and one can see in the above list that there is a trade-off between $\alpha$ and $\tau$: A small $\alpha$ requires a small $\tau$, which means more energy must be expended running the adjoint system. The nature of the trade-off is modulated by the choice of norms, and presumably one would want to select norms that exhibit an efficient trade-off. This ends example 1.

A problem that can arise in recurrent network optimization is that the dynamics on the weights may destroy the global stability of the network. That is, bifurcations may occur that result in the creation of spurious fixed points or other undesired phenomena. The results in this letter require that the network tends to a unique fixed point for each parameter value at a uniform rate. If this rate is not uniform and the convergence rate is allowed to approach zero during training, a finite timescale separation may not be sufficient to ensure the descent condition. The network in example 1 did not present any of these issues since we were able to bound the contraction rate independent of the parameters being optimized. Jin and Gupta (1999) explored several approaches to stability in the discrete time case. In one suggested approach, a penalty term is added to the objective to keep the parameters within a stable regime. In Atiya (1988), the training procedure scales the weights down if they grow so large that global stability cannot be guaranteed. Tseng and Šiljak (1995) applied the concept of connective

stability to derive conditions under which the network will remain stable during weight adaptation.

To get around the possibility of bifurcations, these next examples have uniform stability built in by some mechanism that allows one to a priori bound the weight on each connection. Instead of having the dynamics act directly on the connection strengths, in several of the models, we have the weight on the connection from $j$ to $i$ pass through another sigmoid function. For instance, applying this to equation 7.1, the forward dynamics would then consist of

$$\dot{x} = -x + \sigma(\omega x + \rho),$$
$$\omega = \sigma(w).$$

The point of introducing this model is that although it still may be possible for the weights $w_{i,j}$ to grow to infinity, the relevant quantities for optimization (the first and second derivatives of $f$ with respect to $x$ and $w$) remain bounded. It also enables one to separately control the magnitude and sign of each connection. Note that once the network is trained, it is only necessary to retain $\omega_{i,j}$ as the weight from $i$ to $j$ and $w_{i,j}$ can be discarded.

**Example 2.** We extend example 1 by adding a readout of the feedback module, which is also subject to optimization:

$$\dot{x_1} = -x_1 + \sigma(\omega x_2),$$
$$\omega = \sigma_H(w),$$
$$\dot{x_2} = -x_2 + \sigma(v x_2 + u\rho),$$

which has three sets of parameters $v$, $w$, and $u$, of which $w$ and $u$ are subject to optimization. The presence of a fixed feedback module and dynamics on the read-out weights is a hallmark of reservoir computing (Lukoševičius & Jaeger, 2009). This system is a hierarchy of the form $x_2 \to x_1$, and theorem 1 may be applied to conclude that the overall system $(x_1, x_2)$ is contracting. The boundedness of the various derivatives holds as well.

**Example 3.** We now turn to a network where optimization takes place on all connections. Let $in(i)$ and $out(i)$ denote the in-degree and out-degree, respectively, of node $i$, and let $E$ be the set of edges present in the network:

$$\dot{x}_i = -x_i + \sigma\left(\sum_{j:(j,i)\in E} \omega_{i,j} x_j + \rho_i\right), \tag{7.3a}$$
$$\omega_{i,j} = \alpha_i \sigma_H(w_{i,j}). \tag{7.3b}$$

Table 3: Derivatives for the Model Defined by Equations 7.3 and 7.4.

| | |
|---|---|
| $(D_x f)_{i,j}$ | $\sigma'(u_i)\omega_{i,j}$ |
| $(D_w f)_{i,(i,k)}$ | $\sigma'(u_i)x_k\alpha_k\sigma'(w_{i,k})$ |
| $(D_x^2 f)_{i,j,k}$ | $\sigma''(u_i)\omega_{i,j}\omega_{i,k}$ |
| $(D_w^2 f)_{i,(i,k),(i,r)}$ | $\sigma''(u_i)x_k x_r\alpha_k\alpha_r\sigma'(w_{i,k})\sigma'(w_{i,r})$ |
| $(D_{x,w}^2 f)_{i,j,(i,r)}$ | $\sigma''(u_i)\omega_{i,j}x_r\alpha_r\sigma'(w_{i,r})$ |

This system is contracting in the $\infty$-norm, for instance, when

$$0 \leq \alpha_i < \frac{1}{L_{\sigma_H} L_{\sigma'} in(i)}.$$

This can be easily seen from the definition of $\mu_\infty$. Due to the $\omega$ terms being bounded, the other conditions required by the optimization procedure may also be verified. If we replace equation 7.3a in this example with the definition

$$\omega_{i,j} = \alpha_j \sigma(w_{i,j}) \tag{7.4}$$

then the system, equations 7.3a and 7.4 is contracting in the 1-norm when

$$|\alpha_j| < \frac{1}{L_\sigma L_{\sigma'} out(j)}.$$

Here the sign of $\alpha_j$ determines whether connections emanating from that node are facilitory or inhibitory. Additionally, as can be seen from Table 3, all the relevant derivatives are bounded. Therefore, the optimization procedure can be applied in this case as well.

## 8 Conclusion

We believe that in addition to the differentiability of the system components, the existence of appropriate timescales as demonstrated here is key to concluding that optimization is feasible in recurrent networks. This allows one to implement a flow on the parameters that is naturally adapted to the distributed nature of the network in the sense that it can be expressed in terms of local information. This is relevant not only to distributed or parallel implementations. When the optimization is phrased in terms of the system 2.7, one avoids having to solve a set of nonlinear equations at each step and can rely on simple computations involving the derivatives of the activation functions. We also believe that the conditions required here are not too restrictive since, as shown in section 7, a wide variety of criteria

may be obtained by different choices of norm. Several of these conditions allow one to conclude stability based on local information at each node. The corresponding contraction rate also depends only on local data. This is important if the procedure is to be regulated in a decentralized manner. As is, the rate $\tau$ provided by these results does not have this property, since it involves the equivalence between the 2-norm and the norm in which the system is contracting, which may depend on global properties such as the number of nodes in the network. Recently the memristor has been proposed as building block for synapses in neuromorphic computing architectures. A possible extension to our work would be to investigate models containing memristors. The stability problem for such systems has been studied in, for instance, Wu and Zeng (2012). Other interesting extensions of this analysis might consider changes to the network topology during learning, discrete time implementations of this scheme, and the application to probabilistic models.

## Appendix

**Proof of Proposition 3.** We verify the conditions of theorem 1. By proposition 1, we conclude that the adjoint system $y$ is contracting in the norm $\|\cdot\|_*$. Define $e(x, y, w) = y\left(D_x f\right)_{(x,w)} - (D_x g)_{(x)}$, so that $\dot{y} = e(x, y, w)$. We will show that $\|D_x e\|_{X,X^*} \leq k$. Let $x_1, x_2$ be two possible states of the $x$ system. Holding $y, w$ fixed, we have

$$\begin{aligned}\|e(x_1) - e(x_2)\|_* \leq \|((D_x f)_{(x_1)} - (D_x f)_{(x_2)})^T y\|_* \\ + \|(D_x g)_{(x_1)} - (D_x g)_{(x_2)}\|_*.\end{aligned}$$

Using the fact that $\|M^T\|_* = \|M\|$ for any matrix $M$, one has

$$\begin{aligned}\|\left((D_x f)_{(x_1)} - (D_x f)_{(x_2)}\right)^T y\|_* &\leq \|((D_x f)_{(x_1)} - (D_x f)_{(x_2)})^T\|_{*,*}\|y\|_* \\ &\leq \|(D_x f)_{(x_1)} - (D_x f)_{(x_2)}\|\|y\|_* \\ &\leq (L_{x^2} f)\|x_1 - x_2\|\|y\|_*.\end{aligned}$$

By assumption on $D_x g$, we have $\|(D_x g)_{(x_1)} - (D_x g)_{(x_2)}\|_* \leq (L_{x^2} g)\|x_1 - x_2\|$. By definition of $U$, we may assume $\|y\|_*$ is bounded. Therefore, theorem 1 can be applied.

**Proof of Proposition 4.** In general for a matrix $A$, if $\mu(A) \leq -\alpha$ for $\alpha > 0$, then $\|A^{-1}\| \leq \frac{1}{\alpha}$ (Desoer & Haneda, 1972). Turning to the system $z$, by the chain rule, we have $(D_w z^*)_{(w)} = -(D_z u)^{-1}(D_w u)_{(z^*(w),w)}$. Since $\mu(D_z u) \leq -\beta$ where $\beta > 0$, $D_z u$ is invertible and $\|(D_z u)^{-1}\|_Z \leq \frac{1}{\beta}$. Then $\|D_w z^*\|_{W,Z} \leq \frac{1}{\beta}\|D_w u\|_{W,Z} \leq \frac{L_w u}{\beta}$. For the second statement, we proceed as in

the proof of Vidyasagar (2002, theorem 2.5.3). It is sufficient to show that $\lim_{h\to 0^+} \frac{1}{h}(\|u(t)+hD_t u(t)\|_Z - \|u(t)\|_Z) \le -\beta\|u\|_Z + (L_w u)\|\dot{w}\|_W$.
Note that

$$\begin{aligned}\|u(t)+h(D_t u(t))\|_Z &= \|u(t)+h((D_z u)u(t)+(D_w u)\dot{w})\|_Z \\ &\le \|I+h(D_z u)\|_Z\|u\|_Z + h\|D_w u\|_{W,Z}\|\dot{w}\|_W.\end{aligned}$$

From here, one can subtract $\|u\|_Z$ from both sides, divide by $h$, and take limits as $h \to 0^+$, while using the definition of matrix measure.

**Proof of Theorem 4.1.** The norm used here is the 2-norm. Let $L$ be the Lipschitz constant of $D_z v$, so that $\|(D_z v)_{(z_1)} - (D_z v)_{(z_2)}\| \le L\|z_1 - z_2\|$ for all $z_1, z_2$. It can be easily seen that if $\alpha < 1$, then $v(z(t))$ is decreasing. Under the additional assumption that $v$ is bounded from below, we conclude that $v(z(t))$ converges to some value $v^*$. We show that $D_z v \to 0$. Let $\phi(z,t)$ be the map that takes initial conditions $v(0) = z$ to the state of the gradient system at time $t$. Using the estimate

$$\|\phi(z,t) - \phi(z,t+h)\| \le \frac{\|(D_z v)_{(z_t)}\|}{L}\left(e^{(1+\alpha)Lh} - 1\right)$$

and fixing an initial condition, one may obtain from Taylor's formula that

$$\begin{aligned}v(z(t+h)) \le v(z(t)) + \|(D_z v)_{(z_t)}\|^2 \Bigg(&-2h + \frac{1}{L}(e^{(1+\alpha)Lh} - 1) \\ &+ \frac{h^2}{2L}(e^{(1+\alpha)Lh} - 1)^2\Bigg).\end{aligned}$$

Note that this can be written as

$$v(z(t+h)) \le v(z(t)) + \|(D_z v)_{(z_t)}\|^2 r(h),$$

where $r(h) = -2h + \frac{1}{L}(e^{(1+\alpha)Lh} - 1) + h^2 k(L,\alpha,h)$ for some function $k$. Since $r'(0) = -1 + \alpha$, it follows that if $\alpha < 1$, then there exists some $h, \beta$ both positive and depending only on $L, \alpha, h$, so that for all $t$,

$$v(z(t+h)) \le v(z(t)) - \beta\|(D_z v)_{(z(t))}\|^2.$$

We use this estimate to derive a contradiction in the case that $D_z v \not\to 0$. If this is so, then there is an $\epsilon$ so that for every $t$, there is a $t' > t$ where $\|(D_z v)_{(z_{t'})}\| \ge \epsilon$. In particular, there is an increasing sequence $\{t_i\}$ of times where $t_{n+1} - t_n > h$ and $\|(D_z v)_{(z(t_i))}\| \ge \epsilon$ for all $i$. Setting $y_n = v(z(t_n)) - v(z(t_{n-1}))$, it follows

that $v(z(t_0)) + \sum_{n=1}^{\infty} y_n \to f^*$ where the series converges. However, since $t_{n+1} - t_n \geq h$, and $f$ is strictly decreasing, we must have

$$\begin{aligned} v(z(t_{n+1})) - v(z(t_n)) &\leq v(z(t_n + h)) - v(z(t_n)) \\ &\leq -\beta \|(D_z v)(z(t_n))\|^2 \\ &\leq -\beta\epsilon^2, \end{aligned}$$

which implies that $\sum_{n=1}^{\infty} y_n \to -\infty$.

**Proof of Proposition 5.** By proposition 2, we have $\|y\|_* \leq \frac{L_x g}{\beta_x}$. Fix $w$ and let $z_1 = (x_1, y_1), z_2 = (x_2, y_y)$ be two states of the joint system. Note that for any $x$, $\|(D_w f)^T_{(x)}\|_{X^*,W} = \|(D_w f)_{(x)}\|_{W^*,X} \leq L_w f$ and

$$\|((D_w f)_{(x_1)} - (D_w f)_{(x_2)})^T\|_{X^*,W} = \|(D_w f)_{(x_1)} - (D_w f)_{(x_2)}\|_{W^*,X}.$$

Then we have

$$\begin{aligned} \|h(z_1) - h(z_2)\|_W &= \|(D_w f)^T_{(x_1)} y_1 - (D_w f)^T_{(x_2)} y_2\|_W \\ &\leq \| \left( (D_w f)_{(x_1)} - (D_w f)_{(x_2)} \right)^T y_1 \|_W \\ &\quad + \|(D_w f)^T_{(x_2)} (y_1 - y_2)\|_W \\ &\leq (L_{x,w} f)\|x_1 - x_2\|_X \|y_1\|_* + (L_w f)\|y_1 - y_2\|_*. \end{aligned}$$

Setting $k = \max\{\frac{(L_x g)}{\beta_x} \frac{(L_{x,w} f)}{p_1}, \frac{L_w f}{p_2}\}$, it follows that $\|h(z_1) - h(z_2)\|_W \leq k\|z_1 - z_2\|_Z$. For $D_w h$, fix some $z$, and let $w_1$, $w_2$ be two parameter vectors. Then

$$\begin{aligned} \|h(w_1) - h(w_2)\|_W &= \| \left( (D_w f)_{(w_1)} - (D_w f)_{(w_2)} \right)^T y \|_W \\ &\leq \|(D_w f)_{(w_1)} - (D_w f)_{(w_2)}\|_{W^*,X} \|y\|_* \\ &\leq (L_{w^2} f)\|y\|_* \|w_1 - w_2\|_W. \end{aligned}$$

**Proof of Proposition 6.** Fixing $x$, we have $\|f(w_1) - f(w_2)\|_X \leq (L_w f)\|w_1 - w_2\|_W$. For the $y$ system, we have

$$\begin{aligned} \|((D_x f)_{(w_1)} - (D_x f)_{(w_2)})^T y\|_* &\leq \|((D_x f)_{(w_1)} - (D_x f)_{(w_2)})^T\|_{X^*,X^*} \|y\|_* \\ &\leq \|(D_x f)_{(w_1)} - (D_x f)_{(w_2)}\|_{X,X} \|y\|_* \\ &\leq (L_{x,w} f)\|w_1 - w_2\|_W \|y\|_*. \end{aligned}$$

Then

$$\begin{aligned}\|u(w_1) - u(w_2)\|_Z &\le p_1(L_w f)\|w_1 - w_2\|_W + p_2(L_{w,x} f)\|y\|_*\|w_1 - w_2\|_W \\ &= (p_1(L_w f) + p_2(L_{x,w} f)\|y\|_*)\|w_1 - w_2\|_W.\end{aligned}$$

We now bound $D_z u$. Fixing $w$, let $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$ be two states of the enlarged system. Then $\|f(x_1) - f(x_2)\|_X \le (L_x f)\|x_1 - x_2\|_X$, while for the $y$ system, we have

$$\begin{aligned}\|((D_x f)_{(x_1)} - (D_x f)_{(x_2)})^T y_1\|_* &\le \|((D_x f)_{(x_1)} - (D_x f)_{(x_2)})^T\|_{X^*,X^*}\|y\|_* \\ &\le \|(D_x f)_{(x_1)} - (D_x f)_{(x_2)}\|_{X,X}\|y\|_* \\ &\le (L_{x^2} f)\|x_1 - x_2\|_X\|y\|_*\end{aligned}$$

and $\|(D_x f)^T_{(x_2)}(y_1 - y_2)\|_* \le \|(D_x f)^T_{(x_2)}\|_*\|y_1 - y_2\|_* \le (L_x f)\|y_1 - y_2\|_*$. It follows that

$$\begin{aligned}\|(D_x f)^T_{(x_1)} y_1 - (D_x f)^T_{(x_2)} y_2\|_* \le{}& (L_{x^2} f)\|x_1 - x_2\|_X\|y\|_* \\ &+ (L_x f)\|y_1 - y_2\|_*.\end{aligned}$$

Then

$$\begin{aligned}\|u(z_1) - u(z_2)\|_Z \le{}& (p_1(L_x f) + p_2(L_{x^2} f)\|y\|_*)\|x_1 - x_2\|_X \\ &+ p_2(L_x f)\|y_1 - y_2\|_*.\end{aligned}$$

Setting $k = (L_x f) + \frac{p_2}{p_1}(L_{x^2} f)\|y\|_*$, it follows that $\|u(z_1) - u(z_2)\|_Z \le k\|z_1 - z_2\|_Z$.

**Proof of Proposition 7.** The entries of the matrix $D_w f$ are $(D_w f)_{i,(j,k)} = \delta_{i,j}\sigma'(u_j)\rho_k$. We show that for a general matrix $n \times (n \times n)$ matrix $M$ with $M_{i,(j,k)} = \delta_{i,j} a_j b_k$, the norm $\|M\|_{F,2}$ satisfies $\|M\|_{F,2} \le \|a\|_\infty \|b\|_2$. The first result follows from this by setting $a_j = \sigma'(u_j)$ and $b = \rho$. By definition, $\|M\|_{F,2} = \sup_{\|N\|_F = 1} \|MN\|_2$. Straightforward computation shows that for any matrix $N$, the entries of the vector $MN$ are $(MN)_i = a_i(N\rho)_i$. Therefore, $MN = D(a)(N\rho)$ where $D(a)$ is the diagonal matrix with the vector $a$ on the diagonal. Then $\|MN\|_2 \le \|D(a)\|_2\|N\rho\|_2$. In the 2-norm, a diagonal matrix $D(a)$ has $\|D(a)\|_2 = \|a\|_\infty$. Using this and the fact that the Frobenius norm is consistent, we get $\|MN\|_2 \le \|a\|_\infty\|\rho\|_2\|N\|_F$. Since the $N$ is arbitrary, this shows $\|M\|_{F,2} \le \|a\|_\infty\|\rho\|_2$.

To show the second part, fix $x$ and let $w_1, w_2$ be two different weight matrices. Then $(D_w f)_{(w_1)} - (D_w f)_{(w_2)} = M$, where $M_{i,(j,k)} = \delta_{i,j} a_j b_k$,

where $a_j = \sigma'((vx + w_1\rho)_j) - \sigma'((vx + w_2\rho)_j)$ and $b = \rho$. By the above argument, $\|M\|_{F,2} \leq \|a\|_\infty \|b\|_2$. In general $\|a\|_\infty \leq \|a\|_2$, and in this case, $\|a\|_2 \leq L_{\sigma''}\|(w_1 - w_2)\rho\|_2 \leq L_{\sigma''}\|\rho\|_2\|w_1 - w_2\|_F$. This shows $\|(D_w f)_{(w_1)} - (D_w f)_{(w_2)}\|_{F,2} \leq L_{\sigma''}\|\rho\|_2^2\|w_1 - w_2\|_F$, from which the second result follows.

## References


Almeida, L. B. (1987). A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *Proceedings of the IEEE First International Conference on Neural Networks*. Piscataway, NJ: IEEE.

Atiya, A. F. (1988). Learning on a general network. In D. Anderson (Ed.), *Neural information processing systems* (pp. 22–30). Melville, NY: American Institute of Physics.

Baldi, P. (1995). Gradient descent learning algorithm overview: A general dynamical systems perspective. *IEEE Transactions on Neural Networks*, *6*(1), 182–195.

Bertsekas, D. P., & Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, *10*(3), 627–642.

Del Vecchio, D., & Slotine, J.-J.E. (2013). A contraction theory approach to singularly perturbed systems. *IEEE Transactions on Automatic Control*, *58*(3), 752–757.

Desoer, C., & Haneda, H. (1972). The measure of a matrix as a tool to analyze computer algorithms for circuit analysis. *IEEE Transactions on Circuit Theory*, *19*(5), 480–486.

Dong, D. W., & Hopfield, J. J. (1992). Dynamic properties of neural networks with adapting synapses. *Network: Computation in Neural Systems*, *3*(3), 267–283.

Fang, Y., & Kincaid, T. G. (1996). Stability analysis of dynamical neural networks. *IEEE Transactions on Neural Networks*, *7*(4), 996–1006.

Jin, L., & Gupta, M. M. (1999). Stable dynamic backpropagation learning in recurrent neural networks. *IEEE Transactions on Neural Networks*, *10*(6), 1321–1334.

Lohmiller, W., & Slotine, J.-J.E. (1998). On contraction analysis for non-linear systems. *Automatica*, *34*(6), 683–696.

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

Meyer-Baese, A., Ohl, F., & Scheich, H. (1996). Singular perturbation analysis of competitive neural networks with different time scales. *Neural Computation*, *8*(8), 1731–1742.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, *56*(1), 71–113.

Pineda, F. J. (1989). Recurrent backpropagation and the dynamical approach to adaptive neural computation. *Neural Computation*, *1*(2), 161–172.

Riaza, R., & Zufiria, P. (2002). Time-scaling in recurrent neural learning. In J. Dorronsoro (Ed.), *Notes in Computer Science: Artificial Neural Networks ICANN 2002, Vol. 2415* (pp. 1371–1376). Berlin: Springer.

Riaza, R., & Zufiria, P. J. (2003). Differential-algebraic equations and singular perturbation methods in recurrent neural learning. *Dynamical Systems: An International Journal*, *18*(1), 89–105.

Russo, G., Di Bernardo, M., & Sontag, E. D. (2010). Global entrainment of transcriptional systems to periodic inputs. *PLoS Computational Biology*, *6*(4), e1000739.

Salakhutdinov, R., & Hinton, G. E. (2009). Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (pp. 448–455). JMLR.

Söderlind, G. (1984). On nonlinear difference and differential equations. *BIT Numerical Mathematics*, *24*(4), 667–680.

Sontag, E. D. (2010). Contractive systems with inputs. In J. Willems, S. Hara, Y. Ohta, & H. Fujioka (Eds.), *Perspectives in mathematical System Theory, Control, and Signal Processing* (pp. 217–228). New York: Springer.

Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 1064–1071). New York: ACM.

Tseng, H. C., & Šiljak, D. (1995). A learning scheme for dynamic neural networks: Equilibrium manifold and connective stability. *Neural Networks*, *8*(6), 853–864.

Vidyasagar, M. (2002). *Nonlinear systems analysis* (2nd ed.). Philadelphia: Society for Industrial and Applied Mathematics.

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

Wu, A., & Zeng, Z. (2012). Dynamic behaviors of memristor-based recurrent neural networks with time-varying delays. *Neural Networks*, *36*, 1–10.

Younes, L. (1999). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics: An International Journal of Probability and Stochastic Processes*, *65*(3–4), 177–228.

Yuille, A. L. (2005). The convergence of contrastive divergences. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 1593–1600). Cambridge, MA: MIT Press.

Received November 19, 2014; accepted February 8, 2015.