Adaptive Capability of Recurrent Neural Networks with Fixed Weights for Series-Parallel System Identification

James Ting-Ho Lo
jameslo@umbc.edu
Department of Mathematics and Statistics, University of Maryland
Baltimore County, Baltimore, MD 21250, U.S.A.

By a fundamental neural filtering theorem, a recurrent neural network with fixed weights is known to be capable of adapting to an uncertain environment. This letter reports some mathematical results on the performance of such adaptation for series-parallel identification of a dynamical system as compared with the performance of the best series-parallel identifier possible under the assumption that the precise value of the uncertain environmental process is given. In short, if an uncertain environmental process is observable (not necessarily constant) from the output of a dynamical system or constant (not necessarily observable), then a recurrent neural network exists as a series-parallel identifier of the dynamical system whose output approaches the output of an optimal series-parallel identifier using the environmental process as an additional input.

1 Introduction

In 1992, a fundamental theorem on recursive neural filtering was reported (Lo, 1992a, 1992b, 1994b) in a general formulation of filtering (Liptser & Shiryaev, 1977; Bucy & Joseph, 1987; Kallianpur & Karandikar, 1988). The theorem states that a recursive neural network exists that inputs the measurement process and outputs a virtually optimal estimate of a part or the whole of the signal process, where the estimate can be made as close as desired to the conditional expectation of the signal process given the history of the measurement process that has been processed. The theorem and its proof show that this is true whether or not an uncertain environmental process is involved.

Based on the theorem, it was observed (Lo, 1994a; Lo & Yu, 1995) that if the environmental process has a small variability and if its probability distribution is fully represented in the training data, a neural filter with fixed weights that is trained on the data is capable of adapting to the environmental process, and online adjustment of the weights of the adaptive filter is unnecessary. In fact, such adaptation can be viewed as a manifestation of an estimation of the observable environmental process performed implicitly inside the recurrent neural network.

The adaptive capability of a fixed-weight recurrent neural network was also confirmed by other researchers (e.g., Feldkamp, Puskorius, & Moore, 1996; Feldkamp & Puskorius, 1997; Younger, Conwell, & Cotter, 1999; Younger, Hochreiter, & Conwell, 2001; Prokhorov, Feldkamp, & Tyukin, 2002; Prokhorov, 2006). In engineering, a processor such as a filter or a controller whose weights or parameters are adjusted online to adapt to an uncertain environment is called an adaptive processor (Widrow & Stearns, 1985; Haykin, 1996, 1999; Principe, Euliano, & Lefebvre, 2000). To distinguish an adaptive processor that adapts without online weight or parameter adjustment from an adaptive processor that adapts with such adjustment, the former is called an accommodative processor (e.g., accommodative filter and accommodative controller).

Lo (1994a) and Lo and Yu (1995) remarked that an accommodative neural filter is not expected to work as well as an adaptive neural filter if the environmental process has large variability or is not properly represented in the offline training data. This remark was confirmed in Lo and Bassu (2001) by numerical comparison of the performance of accommodative neural networks and that of adaptive neural networks with long- and short-term memories for system identification. Nevertheless, in many applications of practical importance, not requiring online adjustment for adaptation is a highly desirable advantage of the accommodative neural network, especially when online training data are not available.

A natural question is under what condition and how well an uncertain environmental process can be adapted to by an accommodative processor. This letter makes a mathematical attempt toward answering the question in the context of series-parallel system identification (Ljung & Soderstrom, 1983; Widrow & Stearns, 1985; Ljung, 1987; Haykin, 1996). This letter is based on the fundamental neural filtering theorem in Lo (1992a, 1992b, 1994b), which applies to system identification as a special case.

Obviously the best performance that an accommodative processor can achieve is the performance achievable given the precise value of the environmental process. If an accommodative processor exists whose performance converges to this best performance achievable, the environmental process is said to be adaptation-fit for the processing involved. A precise statement of this definition of adaptation fitness for series-parallel system identification is given in section 3, after a general series-parallel system identification problem is stated in section 2. It is proved in section 3 that there exists an accommodative neural network with an identification performance allowed by the degree of adaptation fitness of the environmental process. The proof is based on the fundamental neural filtering theorem in Lo (1992a, 1992b, 1994b).
Observability of a stochastic environmental process from the dynamical state of a stochastic system is defined in section 4. As expected, the degree of adaptation fitness of the environmental process depends on the degree of its observability, as well as the smoothness of the conditional expectation of the signal process with respect to the environmental process. The main theorem in section 4 gives an intuitively appealing quantification of the dependence.

A surprising result in this letter is that a constant environmental process does not have to be observable to be adaptation-fit to any degree. This is stated and proved as the main theorem in section 5. The proof of this theorem is based on a theorem that justifies the use of adaptive multilayer perceptrons with long- and short-term memories in Lo and Bassu (1999, 2002).

Physical justification of the accommodative capability of recurrent neural networks with fixed weights can best be given with an engineering application. In many applications of active acoustic or structural vibration control, the vibration propagation paths vary due to changes in temperature, humidity, loading, and speed, and allowing the vibration sources to collect data for modeling the paths and adjusting the controller is not acceptable. In Lo and Bassu (2003), accommodative identifiers and controllers are described that do not need to be adjusted online but can adapt to the changing propagation paths.

2 Formulations of System Identification

If a mathematical description of a dynamical system is unavailable and is constructed using only measurements of inputs and outputs of the dynamical system, the construction is called the identification of the dynamical system. In this letter, dynamical systems to be identified satisfy the following general difference equation:

\[ y(t, \theta_i) = f[y_{-p}^{t-1}(\theta), u_{-q}^{t-q}(\theta), \xi_t] \tag{2.1} \]

\[ y_{-p}^{t-1}(\theta) := (y(t - 1, \theta_{t-1}), y(t - 2, \theta_{t-2}), \ldots, y(t - p, \theta_{t-p})) \tag{2.2} \]

\[ u_{-q}^{t-q}(\theta) := (u(t), \ldots, u(t - q)) \tag{2.3} \]

where \( \theta_i \) denotes the value of an uncertain environmental process at time \( t \) with values from the compact set \( \Theta \), the pair \((u(t), y(t, \theta_i))\) is the input-output pair at time \( t \), and \( \xi_t \) is a random sequence. The dynamical system to be identified is called a plant. The function \( f \) and the integers, \( p \) and \( q \), are unknown, but \( p < T_0 \) and \( q < T_0 \) for some known \( T_0 < \infty \). It is assumed that there is a positive number \( B \) such that for all \( t, \|u(t)\| < B \), \( \|y(t, \theta_i)\| < B \), and \( \|f(y_{-p}^{t-1}(\theta), u_{-q}^{t-q}(\theta), \xi_t)\| < B \). The plant is a deterministic or stochastic system depending on whether the random disturbance \( \xi_t = 0 \) for all \( t \). For simplicity, the output \( y(t, \theta_i) \) is often denoted by \( y(t) \).

There are two formulations of the problem of identifying the plant: the series-parallel formulation and parallel formulation, depending on whether the plant output \( y(t - 1) \) is used as an input to the system identifier at time \( t \).

In the series-parallel formulation, at time \( t \), both the plant input \( u(t) \) and output \( y(t - 1) \) are used as inputs to the system identifier, and the plant output \( y(t) \) is the target output of the system identifier. In the parallel formulation, at time \( t \), only the plant input \( u(t) \) is used as an input to the system identifier. The series-parallel system identifier is also looked on as a one-step predictor and the parallel system identifier as an input-output model of the plant.

For a system identifier to show accommodative ability, the system identifier has to input information about the environmental process. Such information is contained in the outputs, equation 2.2, of the plant, but usually not in the inputs, equation 2.3. Since outputs of the plant are not used as inputs to a parallel identifier, accommodative identification of a plant is usually performed in the series-parallel formulation, which is what we concentrate on in this letter.

In the rest of this letter, the inputs \( u(t) \) are assumed known wherever required and therefore not explicitly mentioned in the sequel for simplicity.

3 Accommodating an Adaptation-Fit Environmental Process

It is intuitively clear that successful adaptation to an uncertain environmental process requires sufficient information about it. If sufficient information about the environmental process is contained in the outputs of a plant for series-parallel identification of the plant to a certain accuracy, the environmental process is called adaptation-fit to that accuracy. This definition is precisely stated as definition 1. In the main theorem of this section, it is proven that if the environmental process is adaptation-fit to a certain accuracy, the series-parallel identification of the plant to the same accuracy can be realized with a recurrent neural network with fixed weights. Such a recurrent neural network does not require online weight adjustment and is called an accommodative neural network.

In the rest of the letter, we denote the Euclidean norm by \( \| \cdot \| \), the plant output \( y(t, \theta_i) \) by \( y_t \), the conditional expectation of a random variable \( x \) given another \( y \) by \( E[x | y] \), the transpose of a matrix \( A \) by \( A' \), and the sequences \( \{\theta_t, \tau = 1, \ldots, l\} \), \( \{y_t, \tau = 1, \ldots, l\} \), and \( \{y_t, \tau = s, \ldots, t\} \) by \( \theta^t \), \( y^t \) and \( y^t_s \), respectively.

Definition 1. Assume that the environmental process \( \theta \) is a stochastic process. The environmental process \( \theta \) is said to be adaptation-fit to within an error of \( \epsilon \) for
identifying the plant, equation 2.1, in the series-parallel formulation, if there is a positive integer \( N(\epsilon) \) such that for all \( t > N(\epsilon) \),

\[
E[\|\hat{y}(t) - \hat{y}(t)\|_2^2] < \epsilon
\]

where \( \| \cdot \| \) denotes the Euclidean norm, and \( \hat{y}(t) \) and \( \hat{y}(t \mid \theta) \) denote the conditional expectations \( E[y(t) \mid y^{t-1}] \) and \( E[y(t) \mid y^{t-1}, \theta^{t-1}] \), respectively.

We need the fundamental neural filtering theorem (Lo, 1992a, 1992b, 1994b) to prove the main theorem of this section.

**Theorem 1.** Consider an \( n \)-dimensional stochastic process \( x(t) \) and an \( m \)-dimensional stochastic process \( y(t) \), \( t = 1, \ldots, T \) defined on a probability space \((\Omega, \mathcal{A}, P)\). Assume that the range \( \{ y(t, \omega) \mid t = 1, \ldots, T, \omega \in \Omega \} \subset \mathbb{R}^m \) is compact and \( \psi \) is an arbitrary \( k \)-dimensional Borel function of \( x(t) \) with finite second moments \( E[\| \psi(x(t)) \|^2] \), \( t = 1, \ldots, T \). Let \( \alpha(t) \) denote the \( k \)-dimensional output at time \( t \) of a recurrent neural network (e.g., multilayer perceptron with interconnected neurons within each hidden layer) that has taken the inputs, \( y(1), \ldots, y(t) \), in the given order.

1. Given \( \epsilon > 0 \), there exists a recurrent neural network with one hidden layer of fully interconnected neurons such that

\[
\frac{1}{T} \sum_{t=1}^{T} E[\| \alpha(t) - E[\psi(x(t)) \mid y^{t-1}] \|^2] < \epsilon.
\]

2. If the recurrent neural network has one hidden layer of \( N \) neurons, which are fully interconnected and the output \( \alpha(t) \) is written as \( \alpha(t \mid N) \) here to indicate its dependency on \( N \), then

\[
r(N) := \min_{\alpha} \frac{1}{T} \sum_{t=1}^{T} E[\| \alpha(t, N) - E[\psi(x(t)) \mid y^{t-1}] \|^2]
\]

is monotone decreasing and converges to 0 as \( N \) approaches infinity.

The following corollary follows immediately the above fundamental neural filtering theorem:

**Corollary 1.** Consider an \( n \)-dimensional stochastic process \( x(t) \) and an \( m \)-dimensional stochastic process \( y(t) \), \( t = 1, \ldots, T \) defined on a probability space \((\Omega, \mathcal{A}, P)\). Assume that the range \( \{ y(t, \omega) \mid t = 1, \ldots, T, \omega \in \Omega \} \subset \mathbb{R}^m \) is compact and \( \psi \) is an arbitrary \( k \)-dimensional Borel function of \( x(t) \) with finite second moments \( E[\| \psi(x(t)) \|^2] \), \( t = 1, \ldots, T \). Let \( \alpha(t) \) denote the \( k \)-dimensional output at time \( t \) of a recurrent neural network that has taken the inputs, \( y(1), \ldots, y(t) \), in the given order.

Given \( \epsilon > 0 \), there exists a recurrent neural network with one hidden layer of fully interconnected neurons such that for all \( t = 1, \ldots, T \),

\[
E[\| \alpha(t) - E[\psi(x(t)) \mid y^{t-1}] \|_2^2] < \epsilon.
\]

We are now ready to state the first main theorem of this letter:

**Theorem 2.** Consider a plant (see equation 2.1) where the environmental process \( \theta \) is adaptation-fit to within an error of \( \epsilon \) for identifying the plant in the series-parallel formulation. For identifying the plant over a time interval, \( 1 \leq t \leq T \), in the series-parallel formulation, is a fixed-weight recurrent neural network, which has only one hidden layer of neurons, inputs \( y(t-1, \theta, \ldots) \) and outputs an estimate \( \alpha(t) \) of the plant output \( y(t, \theta) \) at time \( t \) exist as a series-parallel identifier such that

\[
E[\| \hat{y}(t \mid \theta) - \alpha(t) \|_2^2] < \epsilon
\]

for all \( t \) less than \( T \) and greater than some positive integer \( N(\epsilon) \), where \( \hat{y}(t \mid \theta) \) denotes the conditional expectation \( E[y(t) \mid y^{t-1}, \theta^{t-1}] \).

**Proof.** By definition 1, for the given \( \epsilon \), there is a positive integer \( N(\epsilon) \) such that for all \( t > N(\epsilon) \),

\[
E[\| \hat{y}(t \mid \theta) - \hat{y}(t) \|_2^2] < \epsilon. \tag{3.2}
\]

Let \( \epsilon_1 \) be a positive number less than \( \epsilon - E[\| \hat{y}(t \mid \theta) - \hat{y}(t) \|_2^2] > 0 \). Notice that both \( \hat{y}(t) \) and \( \alpha(t) \) are measurable with respect to \( y^{t-1} \). It follows that

\[
E[(\hat{y}(t \mid \theta) - \hat{y}(t))(\hat{y}(t \mid \theta) - \hat{y}(t) \mid y^{t-1})] = E[\hat{y}(t \mid \theta) - \hat{y}(t)](\hat{y}(t \mid \theta) - \hat{y}(t) \mid \theta) = 0.
\]

By the smoothing property of the conditional expectation, we have, for \( t > N(\epsilon) \),

\[
E[\| \hat{y}(t \mid \theta) - \alpha(t) \|_2^2]
\]
Adaptive Capability of Recurrent Neural Networks

This shows that the cross-term resulting from expanding the second expression below is zero and establishes the second equality below:

\[ E \left\| \hat{y}(t | \theta') - g \left( y_{t-p}^{t-1}, \hat{\theta} \right) \right\|^2 = E \left\| (\hat{y}(t | \theta') - \hat{y}(t)) + (\hat{y}(t) - g \left( y_{t-p}^{t-1}, \hat{\theta} \right)) \right\|^2 = E \left\| \hat{y}(t | \theta') - \hat{y}(t) \right\|^2 + E \left\| \hat{y}(t) - g \left( y_{t-p}^{t-1}, \hat{\theta} \right) \right\|^2. \]

Hence,

\[ E \left\| \hat{y}(t | \theta) - \hat{y}(t) \right\|^2 \leq E \left\| \hat{y}(t | \theta) - g \left( y_{t-p}^{t-1}, \hat{\theta} \right) \right\|^2 = E \left\| g \left( y_{t-p}^{t-1}, \hat{\theta} \right) - g \left( y_{t-p}^{t-1}, \hat{\theta} \right) \right\|^2. \]

By the mean value theorem,

\[ \left\| g \left( y_{t-p}^{t-1}, \hat{\theta} \right) - g \left( y_{t-p}^{t-1}, \hat{\theta} \right) \right\| \leq \left\| \frac{\partial g \left( y_{t-p}^{t-1}, \hat{\theta} \right)}{\partial \theta} \right\| \left\| \hat{\theta} - \hat{\theta} \right\| \leq M \left\| \hat{\theta} - \hat{\theta} \right\|. \]

Therefore, \( E \left\| \hat{y}(t | \theta) - \hat{y}(t) \right\|^2 \leq M^2 E \left\| \hat{\theta} - \hat{\theta} \right\|^2 < M^2 \varepsilon, \) for some \( N(\varepsilon). \)

Example 1. Assume that the plant satisfies a difference equation,

\[ y(t, \theta) = f \left( y(t-1), u(t), \theta \right) + \xi, \]

where all the variables are real-valued, \( |df(y(t-1), u(t), \theta)|/|d\theta| < M, \) \( \theta \) is observable to within an error of \( \varepsilon > 0, \) and \( \xi \) is a white sequence with mean 0, which is statistically independent of \( \theta. \) Then

\[ g \left( y_{t-p}^{t-1}, \theta \right) = E \left[ y(t) | y_{t-p}^{t-1}, \theta \right] = E \left[ f \left( y(t-1), u(t), \theta \right) + \xi | y_{t-p}^{t-1}, \theta \right] = f \left( y(t-1), u(t), \theta \right). \]

Obviously \( |d(g(y_{t-p}^{t-1}, \theta))/d\theta| = |df(y(t-1), u(t), \theta)|/|d\theta| < M. \) By theorem 3, \( \theta \) is adaptation-fit to within an error of \( M^2 \varepsilon \) for series-parallel identification of the plant. By theorem 2, for identifying the plant over a time interval, \( 1 \leq t \leq T, \) in the series-parallel formulation, a fixed-weight recurrent neural network, which has only one hidden layer of neurons and inputs
y(t - 1, \theta_{t-1}) and outputs an estimate \alpha(t) of the plant output y(t, \theta_t) at time t, exists as a series-parallel identifier such that

\[ E[\|y(t | \theta) - \alpha(t)\|^2] < M^2 \epsilon \]

for all t less than T and greater than some positive integer N(\epsilon).

**Remark.** \( \epsilon \) in theorem 3 reflects the predictability of \( \theta_t \) and \( M \) reflects the smoothness of the conditional expectation \( E[y(t | y_{t-1}^{-1}, \theta_t)] \) with respect to \( \theta_t \). Theorem 3 provides only a quantification of our intuition that adaptation fitness increases as the predictability of \( \theta \) and the smoothness of the conditional expectation \( E[y(t | y_{t-1}^{-1}, \theta_t)] \) with respect to \( \theta_t \) increase.

### 5 Constancy Implies Adaptation Fitness

In this section, we show that if \( \theta \) is a random variable, which is constant over time, then it is adaptation-fit to within an error of any \( \epsilon \) for series-parallel identification of the plant, whether \( \theta \) is observable from the outputs of the plant or not.

Theorem 4 in the following is taken from Lo (1998) and Lo and Bassu (2002) and will be used in proving theorem 5, the main theorem of this section. Theorem 4 justifies the use of an adaptive MLP (an MLP with long- and short-term memories) for adaptively approximating \( f(x, \theta) \), where \( \theta \) is a piecewise-constant environmental process.

**Theorem 4.** Let \( a \) be a continuous or nondecreasing sigmoidal function with \((-1, 1)\) as its range, \( X \) and \( \Theta \) be compact sets in \( R^n \) and \( R^m \), respectively; \( p \) be a positive number; \( \mu \) be a finite measure on \( X \times \Theta \); and \( f(x, \theta) \) be a Borel measurable function from \( X \times \Theta \) into \( R^1 \) with \( \int_{X \times \Theta} \| f(x, \theta) \|^2 \, d\mu(x, \theta) < \infty. \) For any \( \epsilon > 0 \), there is a sum

\[ f(x, \theta) = \sum_{j=1}^{N} v_j(\theta) a(u_j x + u_{j0}), \quad (5.1) \]

where \( u_j \) and \( u_{j0}, \) \( j = 1, \ldots, N, \) are constant vectors and real numbers, and \( v_j(\theta), \) \( j = 1, \ldots, N, \) are continuous functions of \( \theta, \) such that

\[ \int_{X \times \Theta} \| f(x, \theta) - f(x, \theta) \|^2 \, d\mu(x, \theta) < \epsilon. \quad (5.2) \]

**Theorem 5.** Let the random sequence \( \xi \) in the plant, equation 2.1, be white, and the environmental process \( \theta \) in the plant be a random variable on a compact set \( \Theta \subset R^\theta. \) If \( E[y(t) | y(t)] \) is a finite matrix and if \( \eta(t) := y(t) - E[y(t) | y_{t-1}^{-1}, \theta] \) has mean zero and covariance \( \sigma^2 I, \) then the environmental process is adaptation-fit to within an error of any \( \epsilon \) for series-parallel identification of the plant.

**Proof.** We will prove the theorem for a plant with a scalar-valued output \( y(t). \) The proof can be easily extended for the vector case.

Since the random sequence \( \xi \) in the plant is white, it follows from equation 2.1 that \( \hat{y}(t | \theta) = E[y(t) | y_{t-1}^{-1}, \theta]. \) Note that the prediction error \( \eta_t := y(t, \theta) - E[y(t) | y_{t-1}^{-1}, \theta] \) is a white sequence. It is easy to see that

\[
E[\|y(t) - E[y(t) | y_{t-1}^{-1}, \theta]\|^2 | y_{t-1}^{-1}, \theta] = E[y^2(t) | y_{t-1}^{-1}, \theta] - (E[y(t) | y_{t-1}^{-1}, \theta])^2.
\]

Since the left side is greater than or equal to 0, so are the right side and its expectation. Hence, \( E[\|E[y(t) | y_{t-1}^{-1}, \theta]\|^2] \leq E[y^2(t)] < \infty. \) By theorem 4, \( E[y(t) | y_{t-1}^{-1}, \theta] \), which is a Borel function of \( y_{t-1}^{-1} \) and \( \theta, \) to be denoted by \( g(y_{t-1}^{-1}, \theta), \) can be approximated with an adaptive two-layer MLP \( \hat{g}(y_{t-1}^{-1}, \theta) \) (with a long-term memory independent of \( \theta \)) such that for any \( \epsilon > 0 \)

\[
E[\|\hat{g}(y_{t-1}^{-1}, \theta) - g(y_{t-1}^{-1}, \theta)\|^2] < \epsilon.
\]

Note that \( y(t, \theta) = \hat{g}(y_{t-1}^{-1}, \theta) + \zeta_t + \eta_t, \) where \( \zeta_t := g(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}, \theta) \) with \( E[\zeta^2_t] < \epsilon. \) It is appropriate to assume that \( \zeta_t \) form a zero-mean white random sequence independent of \( \eta_t. \)

Denote the activation level of neuron \( i \) of layer 1 of the adaptive two-layer MLP \( \hat{g}(y_{t-1}^{-1}, \theta) \) by \( \beta_i(y_{t-1}^{-1}) \) and its corresponding linear weight by \( c_i(\theta). \) Then \( \hat{g}(y_{t-1}^{-1}, \theta) = \sum_{i=0}^{N} c_i(\theta) \beta_i(y_{t-1}^{-1}), \) where \( N \) is the number of neurons in layer 1 and \( \beta_0(y_{t-1}^{-1}) = 1. \) Let the linear weights \( c_i(\theta) \) of the adaptive MLP be determined online by minimizing the following online error criterion,

\[
Q_0(c) = \sum_{k=t-1}^{t-1} \left( y(k) - \sum_{i=0}^{N} c_i(\theta) \beta_i(y_{t-1}^{-1}) \right)^2.
\]

by the variation of \( c = [c_0, \ldots, c_N]^T \) for a given integer \( T. \) Denoting \( \beta_i(y_{t-1}^{-1}) \) by \( \beta_i^{k-1} \) for notational simplicity, the minimization is a least-squares regression for

\[
y(k) = \sum_{i=0}^{N} c_i(\theta) \beta_i^{k-1} + \zeta_k + \eta_k, \quad k = t - T, \ldots, t - 1.
\]
By the well-known linear regression formulas, we obtain the following expressions of the estimate \( \hat{c} := [c_0 \cdots c_N]^\top \), its error \( c - \hat{c} \), and its covariance \( V = E(c - \hat{c})(c - \hat{c})^\top \):

\[
\hat{c} = \left( \sum_{k=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) \right)^{-1} \left( \sum_{k=1}^{T-1} \beta_k^{-1}y(k) \right)
\]

\[
c - \hat{c} = \left( \sum_{k=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) \right)^{-1} \left( \sum_{k=1}^{T-1} \beta_k^{-1}(c_k + \eta_k) \right)
\]

\[
V = \left( \sum_{k=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) \right)^{-1} (\delta^2 + \sigma^2),
\]

where \( \beta_k^{-1} := [\beta_0^{-1} \cdots \beta_N^{-1}]^\top \) and \( \delta^2 := E[c_k^2] \). Notice that \( \beta_k^{-1} \) and \( \delta_k \) are Borel functions of \( y_{t-1}^{-1} \) and \( y_{t-1}^{-1} \), respectively.

Denoting \( \hat{g}(y_{t-1}^{-1}, \theta) \) by \( \hat{g} \) below for notational simplicity, the variance of \( g(t) := g(y_{t-1}^{-1}, \theta) - \sum_{i=0}^{N} \hat{c}_i \beta_i^{-1} \) is

\[
E(g(t))^2 = E(\left( \hat{g}(y_{t-1}^{-1}, \theta) - \hat{g} - \sum_{i=0}^{N} \hat{c}_i \beta_i^{-1} \right)^2) 
\]

\[
\leq 2E(\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g})^2 + \left( \hat{g} - \sum_{i=0}^{N} \hat{c}_i \beta_i^{-1} \right)^2 
\]

\[
\leq 2\varepsilon + 2(\delta^2 + \sigma^2)(\beta_1^{-1})^2
\cdot \left( \sum_{k=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) \right)^{-1}
\cdot \left( \sum_{k=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) \right)^{-1}
\cdot \left( \sum_{k=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) \right)^{-1}.
\]

Denoting the vector of the activation levels of the neurons in layer 1 (including \( \beta_0 = 1 \)) by \( \beta \), we note that \( \lim_{T \to \infty} \sum_{t=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) / T = E[\beta \beta^\top] = E[(\beta - E\beta)(\beta - E\beta)] + (E\beta)(E\beta)^\top \). It is known that if \( E[(\beta - E\beta)(\beta - E\beta)] \) is singular, \( \beta \) lies in some hyperplane \( b^\top \beta = c \) with \( b \neq 0 \) with probability one. Obviously if this is the case, at least one neuron in layer 1 is redundant for approximating \( g(y_{t-1}^{-1}, \theta) \) and can be removed. Therefore, it can be assumed that \( E[\beta \beta^\top] \) is positive definite and invertible.

Hence, \( \lim_{T \to \infty} \left( \sum_{t=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) / T \right) = (E[\beta \beta^\top])^{-1} \), and for any positive-definite matrix \( M \) with positive entries, \( \left( \sum_{t=1}^{T-1} \beta_k^{-1}(\beta_k^{-1}y) / T \right) < (E[\beta \beta^\top])^{-1} + M \) for \( T \) sufficiently large. It follows that

\[
E(g(t))^2 \leq 2\varepsilon + 2(\delta^2 + \sigma^2)(\beta_1^{-1})^2
\cdot (E[\beta \beta^\top])^{-1} + M \beta_1^{-1} / T.
\]

Recalling that \( \beta_1^{-1} \) is the vector of the activation levels of the neurons in layer 1 whose components lie in the range \((-1, 1)\), and that \( \varepsilon \) can be made arbitrarily small, it follows that so can be \( E(g(t))^2 \) for \( T \) sufficiently large.

Since \( \hat{c}_i \) and \( \beta_1^{-1} \) are Borel functions of \( y_{t-1}^{-1} \), so is \( \sum_{i=0}^{N} \hat{c}_i \beta_1^{-1} \), which we denote by \( h(y_{t-1}^{-1}) \). By the smoothing property of the conditional expectation, we obtain

\[
E\left( (\hat{g}(y_{t-1}^{-1}, \theta) - h(y_{t-1}^{-1}))^2 \right)
\]

\[
= E\left( (\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}) + \hat{g}(y_{t-1}^{-1}) - h(y_{t-1}^{-1}))^2 \right)
\]

\[
= E[(\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}))^2] + E[(\hat{g}(y_{t-1}^{-1}) - h(y_{t-1}^{-1}))^2]
\]

\[
\leq 2E\left( (\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}))^2 + E[(\hat{g}(y_{t-1}^{-1}) - h(y_{t-1}^{-1}))^2]
\]

\[
= E[(\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}))^2] + E[(\hat{g}(y_{t-1}^{-1}) - h(y_{t-1}^{-1}))^2]
\]

\[
= E[(\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}))^2] + E[(\hat{g}(y_{t-1}^{-1}) - h(y_{t-1}^{-1}))^2]
\]

Hence, \( E[(\hat{g}(y_{t-1}^{-1}, \theta) - \hat{g}(y_{t-1}^{-1}))^2] \leq E[(\hat{g}(y_{t-1}^{-1}, \theta) - h(y_{t-1}^{-1}))^2] = E[\gamma^2(t)] \), which can be made arbitrarily small for sufficiently large \( T \).

Intuitively, theorem 4 shows that whether the environmental parameter \( \theta \) is observable or not, the dependency of the function \( f(x, \theta) \) to be pushed to the linear weights of the last layer of the MLP in an offline batch training of the same. The linear weights can then be adjusted to adapt to \( \theta \) in an online training. These linear weights are functions of \( \theta \) and are observable within an error of \( \varepsilon \) in accordance with the definition of observability in section 4. This explains why \( \theta \) can be adaptation-fit whether it is observable, as stated in theorem 5.

6 Conclusion

This letter examines the performance of an accommodative neural network used as a series-parallel identifier of a dynamical system as compared with the optimal performance that is achievable under the assumption that the environmental process is available and used. More specifically, it defines adaptation-fit-identifiable processes and proves that if the environmental process is adaptation-fit, there exists an accommodative neural network as a series-parallel identifier whose output approaches the output of an optimal series-parallel identifier that uses the environmental process as an additional input. This letter also defines the observability of an environmental process and proves that if the environmental process is observable,
then it is adaptation-fit. The most surprising result is that a constant and nonobservable environmental process is actually adaptation-fit.

References


Received June 4, 2007; accepted March 12, 2009.