Information Fusion Method For System Identification Based On Sensitivity Analysis

Jacob Barhen, Nageswara S. V. Rao
Center for Engineering Science Advanced Research
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6355
{barhenj,raons}@ornl.gov

Abstract - *We consider the identification of a parametrized time-invariant non-linear plant using a smooth model such as a sigmoid non-linear network. There is measurement noise associated with the plant parameters as well as it's input and output. An initial plant model is obtained by utilizing the domain-specific knowledge in terms of the fundamental plant equations, which in general only partially capture the plant dynamics. Once the initial model is fixed, measurements are collected on the plant parameters and input/output. We show that the iid measurements can be fused with the initial plant model by recomputing the parameters. The updated parameters yield a more accurate identifier of the original plant both in parameter and input/output space. The method is based on empirical versions of the closed-form solutions derived in the nuclear engineering literature for an ideal version of the problem based the sensitivity analysis. We show the asymptotic convergence of our computational procedure as well as derive its finite sample results. We illustrate the method using an identifier based on a sigmoid feedforward neural network.*

Keywords: Sensitivity and uncertainty analysis, system identification, neural network, information fusion.

1 Introduction

In a number of engineering systems, one needs to accurately identify a plant for the purposes such as predicting abnormal operation, fault diagnosis, and designing controllers [14, 9]. A variety of identification methods are currently employed [8], and recently, sigmoidal neural networks were shown to be very useful for designing such identifiers [7]. Identification of complicated dynamical systems, such as the transient upflow of high pressure water in light water reactor heat transfer, was originally developed in the area of nuclear engineering using sensitivity and uncertainty analysis [1]. In practical systems, an initial model of the system is developed based on the overall system equations of the plant, which is then refined by utilizing plant measurements. Such methods were shown to be particularly effective in practice, since the basic functionality is captured by the well-established relationships, whereas the residual unmodeled parts are handled by combining plant measurements with the initial model. These information fusion methods have been shown to be very effective even under limited number of measurements, although no convergence results are known. One of our objectives is to extend these information fusion methods to new classes of systems, and bring the corresponding techniques to the information fusion community. Furthermore, we relate this technique to the empirical risk minimization method of Vapnik [17, 18] by showing it to be a particular implementation, namely using a empirical estimate of the closed-form solution derived for the ideal case. By utilizing the smoothness properties of the identifier, we show the convergence of the proposed computational method in terms of asymptotic consistency as well as distribution-free finite sample guarantee. For the latter, we show that given large enough iid (independently and identically distributed) measurements, the error of the computed solution can be made arbitrarily close to the optimal error with a specified probability in the sense of Probably Approximately Correct (PAC) learning [15, 5].

We consider a time-invariant non-linear plant which is characterized by certain key plant *parameters*. There are random errors associated with measuring the parameters as well as the plant input and output. We consider that an initial plant model is designed based on the equations modeling the basic plant phenomenology, and then some measurements of plant parameters and inputs/outputs have been collected. In this paper, we present an information-fusion method for combining the plant measurements with the initial model so that the fused model more accurately identifies the plant both in terms of plant parameters and output. Without loss of generality and to simplify the technical treatment, the input parameters that cannot be measured without errors are combined with the original parameters to be treated as parameters. The inputs that can be measured without errors are left as explicit inputs. The method is based on the uncertainty analysis performed on a locally linearized plant model [2, 1, 19]. The computational complexity of our method is $O(n^3)$, where n is the sample size, and the computation involves matrix operations such as inverse and product. This computational aspect must be contrasted with the general empirical risk minimization solutions for fusion problems [12] that typically are

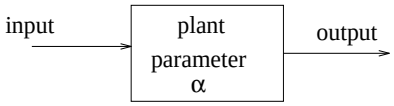

Figure 1: *Plant Model.*

not polynomial-time solvable. Sensitivity-based uncertainty analysis methods have been used extensively in uncertainty reduction in the area of nuclear engineering for the analysis of plants such as light water reactors [20, 1]. In this paper, we present a general information fusion method and apply it to the identifiers based on sigmoid feedforward networks to compute the connection weights.

In section 2, we present a formulation of the problem and describe the information fusion operation. In Section 3, we describe a computational method to implement the fusion method. We show in Section 3.1 the asymptotic convergence of the proposed method and also derive finite sample results for the simple case where the correlation between parameters and outputs are ignored. The performance guarantees are shown in Section 3.2 for the case when these correlations are considered. In Section 4, we provide the details about an implementation using sigmoid feedforward neural networks.

2 Problem Formulation

We consider a general non-linear plant with input $x \in \Re^{d_x}$, output $y \in \Re^{d_y}$, and parameter $\alpha \in \Re^{d_\alpha}$ as shown in Figure 1. Note that inputs that cannot be measured without error are included as components of the parameter α. The input x represents only those inputs that can be measured without error. There is a probabilistic relation between α and y denoted by the joint distribution $P_{\alpha,y}$, which is *unknown*. We assume that the plant operates at a fixed and known value for x. On the other hand, the parameter α is fixed but is unknown, and y varies (due to randomness) even for a fixed value of α.

There are errors in measuring α and y as shown in Figure 2. The corresponding measurements, $\hat{\alpha}$ and $\hat{y}$, are distributed according to the *unknown* conditional distributions $P_{\hat{\alpha}|\alpha}$ and $P_{\hat{y}|y}$, respectively. We assume that both measurements are zero mean in that $E[\hat{\alpha}|\alpha] = \alpha$ and $E[\hat{y}|y] = y$.

The objective is to build a plant model or identifier that accurately represents both the parameters and the output. Let $\langle \tilde{\alpha}, M(x, \tilde{\alpha}) \rangle$ represent such identifier, where $\tilde{\alpha}$ is an estimate of the parameter α, and $M : \Re^{d_x} \times \Re^{d_\alpha} \mapsto \Re^{d_y}$ produces an estimated value $M(x, \tilde{\alpha})$ of the plant output as shown in Figure 3. Dependence on x is not critical to certain parts of the paper, and hence $M(x, \tilde{\alpha})$ is sometimes denoted by $M(\tilde{\alpha})$. We consider that $M(.,.)$ has been derived from the basic plant equations. In general, $M(.)$ does not adequately model the plant noise (due to $P_{y|\alpha}$), and does not account for the measurement

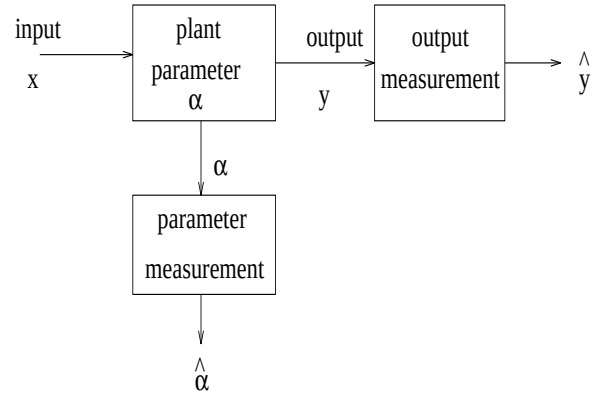

Figure 2: *Measurement Model.*

noise, since measurements are not a part of the plant. To account for the unmodeled parts of plant equations and the measurement noise, a set of measurements $(\hat{\alpha}_1, \hat{y}_1), (\hat{\alpha}_2, \hat{y}_2), \ldots, (\hat{\alpha}_n, \hat{y}_n)$ are collected from the plant ($\hat{\alpha}_i \in \Re^{d_\alpha}$ and $\hat{y}_i \in \Re^{d_y}$ for $i = 1, 2, \ldots n$). Our goal is to *combine* or *fuse* the information contained in the predetermined model $M(.)$ with the measurements to estimate $\tilde{\alpha}$ such that

(i) $\tilde{\alpha}$ is close to the unknown parameter α, and

(ii) $M(x, \tilde{\alpha})$ is close to the unknown output y.

Note that the requirement of the fusion operation is two-fold in that both the estimates of parameter and output must be close to the actual values. Note that even though α is fixed, the observable $\hat{\alpha}$ is random. On the other hand, randomness in $\hat{y}$ is due to the plant characteristics given by $P_{y|\alpha}$ *as well as* the measurement noise given by $P_{\hat{y}|y}$.

Let $\alpha = (a_1, a_2, \ldots, a_{d_\alpha})$ and $\tilde{\alpha} = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{d_\alpha})$ A cost measure that reflects solely the criterion (i) above is given by

$$[\tilde{\alpha} - \alpha] C_\alpha^{-1} [\tilde{\alpha} - \alpha]^T$$

where $C_\sigma = [c_{ij}]$ is the covariance matrix such that $c_{ij} = E[(\tilde{a}_i - a_i)(\tilde{a}_j - a_j)]$. Note here that α itself is not a random variable. Similarly, a cost measure based solely on the criterion (ii) above is given by

$$[M(\tilde{\alpha}) - y] C_y^{-1} [M(\tilde{\alpha}) - y]^T$$

where C_y is the covariance matrix such that $C_y = E[(y - E[y])(y - E[y])^T]$. By simultaneously utilizing the criteria (i) and (ii), the fusion operation is achieved at the minima of the random cost function

$$Q(\tilde{\alpha}) = [(\tilde{\alpha} - \alpha)\ (M(\tilde{\alpha}) - y)] \begin{bmatrix} C_\alpha C_{\alpha,y} \\ C_{y,\alpha} C_y \end{bmatrix}^{-1} [(\tilde{\alpha} - \alpha)(M(\tilde{\alpha}) - y)]^T, \qquad (2.1)$$

where $C_{y,\alpha} = [d_{ij}]$ for $d_{ij} = E[a_i y_i]$ and $C_{\alpha,y} = [e_{ij}]$ for $e_{ij} = E[a_i y_i]$. The correlations between parameters and outputs are captured by $C_{y,\alpha}$ and $C_{\alpha,y}$. Since this is a random cost, it it not possible to minimize it. We

now define a cost function by taking expectation to obtain

$$R(\tilde{\alpha}) = [(\tilde{\alpha} - E[\alpha])\,(M(\tilde{\alpha}) - E[y])] \begin{bmatrix} C_\alpha C_{\alpha,y} \\ C_{y,\alpha} C_y \end{bmatrix}^{-1} [(\tilde{\alpha} - E[\alpha])(M(\tilde{\alpha}) - E[y])]^T, \qquad (2.2)$$

where $E[\alpha]$ refers to expectation with respect to the measurement noise. We seek to minimize $R(.)$ to obtain the estimator $\tilde{\alpha}^*$. In the present case, however, we cannot explicitly minimize $R(.)$ since it depends on the expectations and covariance matrices that depend on the unknown distributions. When a set of measurements are given, such problems are studied extensively under the method of empirical risk minimization due to Vapnik [17]. In this method, a minima is computed for an empirical version of $R(.)$, denoted by $\hat{R}(.)$, as will be shown in the next section. In general, the minimizer of $\hat{R}(.)$ is shown to be close to that of $R(.)$ with a high probability. The minima of $\hat{R}(.)$ is computed in general by solving a non-linear optimization problem. In the next section, we show that the existing solutions to Eq (2.2) derived in [3, 19] can be used to obtain a simple expression for the solution. This solution is very easy to compute as opposed to the minimization problem for $\hat{R}(.)$, which is computationally intractable in general. The validity of our results are based on the smoothness of the model and small deviations of $\hat{\alpha}$ from α as described in the next section.

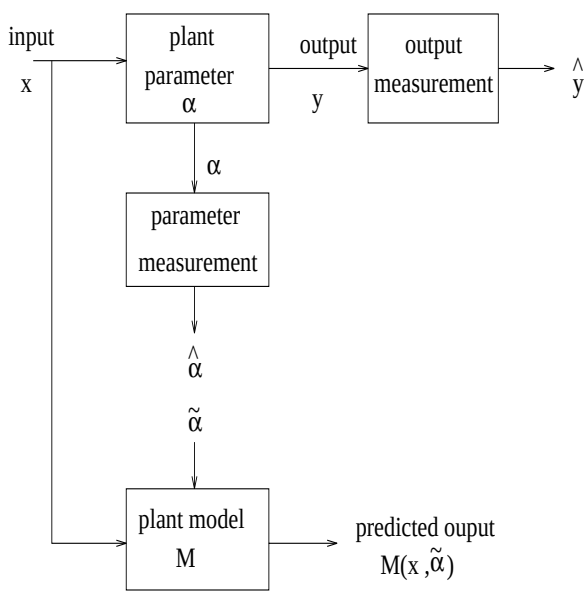

Figure 3: *Identification Model.*

3 Sensitivity-Based Uncertainty Analysis

We now describe sample-based computational methods to minimize the cost $\hat{R}(.)$. One of the simplest way to is to use an observation $\hat{\alpha}$ as an estimate such that $\tilde{\alpha} = \hat{\alpha}$, and the corresponding predicted output is $M(\hat{\alpha})$. Assume that $M(x, \tilde{\alpha})$ is *differentiable* in $\tilde{\alpha}$, and $M(x, \tilde{\alpha}) = (m_1, m_2, \ldots, m_{d_y})$, $\tilde{\alpha} = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{d_\alpha})$. Then, for smooth $M(.)$ and small $\delta\alpha$ such that $\tilde{\alpha} = \alpha + \delta\alpha$, we have

$$M(\hat{\alpha}) = M(\alpha + \delta\alpha) \approx M(\alpha) + S\delta\alpha, \qquad (3.1)$$

where $S = \left[\frac{\partial m_i}{\partial \tilde{a}_j}\right]$ is the sensitivity matrix of the identifier $M(.)$. If $M(.)$ is a good representation of the regression $E[y|.]$, then $M(\alpha)$ will be a good estimate of y. Recall that $M(.)$ is a priori chosen by utilizing the basic plant equations and the domain-specific knowledge; thus, the basic functionality is captured in $M(.)$. We can compute $\tilde{\alpha}$, which will close to α, and and $M(\tilde{\alpha})$ will be close to y.

3.1 Uncorrelated Parameters and Outputs

We first consider a simpler case studied in [19] that ignores the correlations between parameters and output, which results in a simpler version of the cost function in Eq (2.1)

$$Q(\tilde{\alpha}) = [\tilde{\alpha} - \alpha] C_\alpha^{-1} [\tilde{\alpha} - \alpha]^T + [M(\tilde{\alpha}) - y] C_y^{-1} [M(\tilde{\alpha}) - y]^T.$$

For this case, under the condition Eq (3.1), the solution is given by the closed-form expression (see [19] for the derivation)

$$\tilde{\alpha} = \alpha - C_\alpha S^T (C_y + C_M)^{-1} (M(\sigma) - y),$$

where $C_M = S C_\alpha S^T$ is the covariance matrix of M. This solution is not computable because α and y are not known, and the covariances depend on the unknown distributions. Such computation is not possible even in theory since it will not be possible to predict a random number. Instead, we replace α and y by their expectations, which yields the solution

$$\tilde{\alpha}^* = E[\alpha] - C_\alpha S^T (C_y + C_M)^{-1} (M(E[\sigma]) - E[y]),$$

which minimizes the cost function

$$\begin{aligned} R(\tilde{\alpha}) &= [\tilde{\alpha} - E[\alpha]] C_\alpha^{-1} [\tilde{\alpha} - E[\alpha]]^T \\ &\quad + [M(\tilde{\alpha}) - E[y]] C_y^{-1} [M(\tilde{\alpha}) - E[y]]^T. \end{aligned}$$

Now, $\tilde{\alpha}^*$ is computable in principle since the right hand side is a deterministic quantity. In practice, however, it is not possible to compute it since it depends on the covariances C_α and C_y which in turn depend on the unknown distributions.

Now consider that the measurements $(\hat{\alpha}_1, \hat{y}_1), (\hat{\alpha}_2, \hat{y}_2), \ldots, (\hat{\alpha}_n, \hat{y}_n)$ are independently and identically distributed (iid). Then we consider a *sample-based cost function* defined as follows:

$$\begin{aligned} \hat{R}(\tilde{\alpha}) &= [\tilde{\alpha} - \bar{\alpha}] \hat{C}_\alpha^{-1} [\tilde{\alpha} - \bar{\alpha}]^T \\ &\quad + [M(\tilde{\alpha}) - \bar{y}] \hat{C}_y^{-1} [M(\tilde{\alpha}) - \bar{y}]^T, \qquad (3.2) \end{aligned}$$

where (i) $\bar{\alpha}$ and $\bar{y}$ are the sample means of α and y based on measurements, and (ii) $\hat{C}_\alpha$ and $\hat{C}_y$ are the empirical covariances computed based on the sample, corresponding to C_α and C_y, respectively. The cost in

Eq (3.2) is minimized by $\tilde{\alpha}_{\text{sample}}$, which is the solution of [19] where distributions are taken as the empirical distribution of the sample.

$$\tilde{\alpha}_{\text{sample}} = \bar{\alpha} - \hat{C}_\alpha S^T (\hat{C}_y + C_M)^{-1} (\hat{M}(\bar{\alpha}) - \bar{y}).$$

Note that $\tilde{\alpha}_{\text{sample}}$ can be computed using the sample. We employ $\tilde{\alpha}_{\text{sample}}$ as an estimator of $\tilde{\alpha}^*$ (which can only be computed if the distributions are known). We now show that $\tilde{\alpha}_{\text{sample}}$ is a good estimator of $\tilde{\alpha}^*$ in the following theorem.

Theorem 3.1 *Consider the bounded parameters and outputs such that $\alpha \in [-A, A]^{d_\alpha}$ and $y \in [-B, B]^{d_y}$ and the condition Eq (3.1) be satisfied. Let Λ_α and Λ_y denote the upperbounds on the entries of C_α^{-1} and C_y^{-1}, respectively. Let D be an upperbound on $\frac{\partial m_i}{\partial \bar{a}_j}$ for all i, j, and $R(\tilde{\alpha}) \leq \tau$ for all $\tilde{\alpha}$. Let*

$$L = 4(d_\alpha \Lambda_\alpha A + d_y \Lambda_y BD).$$

Given a sample of size at least

$$n = \frac{512\tau}{\epsilon^2} \left[d_\alpha \ln(32/L\epsilon) + \ln(8/\delta) \right],$$

we have, for any δ and ϵ,

$$P\left[R(\tilde{\alpha}_{\text{sample}}) - R(\tilde{\alpha}^*) > \epsilon \right] < \delta.$$

Furthermore, $R(\tilde{\alpha}_{\text{sample}}) \to R(\tilde{\alpha}^*)$, as $n \to \infty$.

Proof: Consider the function class

$$\mathcal{R} = \{ R(\tilde{\alpha}) : \tilde{\alpha} \in [-A, A]^{d_\alpha} \}.$$

By the result of Vapnik [16] (page 41), we have

$$\begin{aligned} & P\left[R(\tilde{\alpha_{\text{sample}}}) - R(\tilde{\alpha}^*) > \epsilon \right] \\ & \quad \leq P\left[\sup_{r \in [-A,A]^{d_\alpha}} |\hat{R}(r) - R(r)| > \epsilon/2 \right]. \end{aligned}$$

To see this result, consider the condition

$$P\left[\sup_{r \in [-A,A]^{d_\alpha}} |\hat{R}(r) - R(r)| > \epsilon/2 \right] < \delta$$

or equivalently

$$P\left[\sup_{r \in [-A,A]^{d_\alpha}} |\hat{R}(r) - R(r)| < \epsilon/2 \right] > 1 - \delta.$$

Then, with probability $1 - \delta$, we have

$$\begin{aligned} R(\tilde{\alpha}^*) &\leq \hat{R}(\tilde{\alpha}^*) + \epsilon/2 \\ &\leq \hat{R}(\tilde{\alpha}_{\text{sample}}) + \epsilon/2 \\ &\leq R(\tilde{\alpha}_{\text{sample}}) + \epsilon \end{aligned}$$

where the first and third inequalities are due to the application of supremum bound for $\tilde{\alpha}^*$ and $\tilde{\alpha}_{\text{sample}}$, respectively, and the second inequality is due to the condition $\hat{R}(\tilde{\alpha}^*) \geq \hat{R}(\tilde{\alpha}_{\text{sample}})$. As a result, we have

$$P\left[R(\tilde{\alpha}_{\text{sample}}) - R(\tilde{\alpha}^*) > \epsilon \right] \leq \delta,$$

which shows the above Vapnik's result.

For any function $g : [-A, A]^d \mapsto \Re$, let

$$\| g(r) \|_\infty = \sup_{r \in [-A,A]^{d_\alpha}} |g(r)|.$$

The *covering number* $N_\infty(\epsilon, \mathcal{G})$ of a function class $\mathcal{G}$ is the smallest cardinality for a subclass $\mathcal{G}^*$ of $\mathcal{G}$ such that

$$\min_{g^* \in \mathcal{G}^*} \| g - g^* \|_\infty \leq \epsilon$$

for each $g \in \mathcal{G}$.

Based on Lemma 3 of Appendix of Krzyzak *et al.* [6] (which itself is based on Pollard [10]), we have

$$\begin{aligned} & P\left[\sup_{r \in [-A,A]^{d_\alpha}} |\hat{R}(r) - R(r)| > \epsilon \right] \\ & \quad \leq 8 N_\infty(\epsilon/8, \mathcal{R}) e^{-\epsilon^2 n / 128\tau} \end{aligned}$$

where $R(\tilde{\alpha}) \leq \tau$, for all $\tilde{\alpha} \in [-A, A]^{d_\alpha}$.

By combining the above two results, we have

$$\begin{aligned} & P\left[R(\tilde{\alpha}_{\text{sample}}) - R(\tilde{\alpha}^*) > \epsilon \right] \\ & \quad \leq 8 N_\infty(\epsilon/16, \mathcal{R}) e^{-\epsilon^2 n / 512\tau}. \end{aligned}$$

Thus, the sample size required to ensure the result of the theorem is given by

$$\frac{512\tau}{\epsilon^2} \left[\ln N_\infty(\epsilon/16, \mathcal{R}) + \ln(8/\delta) \right].$$

We subsequently compute an upper bound for $N_\infty(\epsilon/16, \mathcal{R})$, which proves the theorem. Consider that $R(\tilde{\alpha})$ is *Lipschitz* with the constant L that is for all $\tilde{\alpha}_1, \tilde{\alpha}_2 \in [-A, A]^{d_\alpha}$ we have

$$\| R(\tilde{\alpha}_1) - R(\tilde{\alpha}_2) \|_\infty \leq L \| \tilde{\alpha}_1 - \tilde{\alpha}_2 \|_\infty .$$

By using a mesh with points ϵ apart, we can cover $[-A, A]^{d_\alpha}$ with $\left[\frac{2A}{\epsilon}\right]^{d_\alpha}$ points. Then, because of the Lipschitz property of $R(.)$, we have

$$N_\infty(\epsilon, \mathcal{R}) \leq N_\infty(\epsilon/L, [-A, A]^{d_\alpha}) \leq \left[\frac{2A}{L\epsilon} \right]^{d_\alpha}.$$

Thus, the last step of the theorem is to show that $R(\tilde{\alpha})$ is *Lipschitz* with the constant

$$L = 4(d_\alpha \Lambda_\alpha A + d_y \Lambda_y BD).$$

Note that $\hat{R}(.)$ is the sum of two quadratic forms in $\tilde{\alpha} - \bar{\alpha}$. We note that Lipschitz constant is no higher than the magnitude of the largest value for the derivative $\frac{\partial R}{\partial \tilde{a}_i}$, where $\tilde{\alpha} = (\tilde{a}_1, \ldots, \tilde{a}_{d_\alpha})$. For $\bar{\alpha} = (\bar{a}_1, \ldots, \bar{a}_{d_\alpha})$, we have the following upper bound

$$\begin{aligned} \frac{\partial R}{\partial \tilde{a}_i} &\leq 2 d_\alpha \Lambda_\alpha \max_{\tilde{a}_i, \bar{a}_i} \{ \tilde{a}_i - \bar{a}_i \} \\ &\quad + 2 d_y \Lambda_y \frac{\partial M}{\partial \tilde{a}_i} \max_{\tilde{a}_i, \bar{a}_i} \{ \tilde{a}_i - \bar{a}_i \} \\ &\leq 4 d_\alpha \Lambda_\alpha A + 4 d_y \Lambda_y DB. \end{aligned}$$

The sample size follows from by using this value for L in the bound for $N_\infty(\epsilon/16, \mathcal{R})$.

To show the asymptotic convergence, let

$$\delta(n,\epsilon)8N_\infty(\epsilon/16,\mathcal{R})e^{-\epsilon^2 n/512\tau}$$

show the explicit dependence of δ on n and ϵ. Under the finiteness of $N_\infty(\epsilon,\mathcal{R})$, the consistency result follows from the Borel-Cantelli Lemma [4] if

$$\sum_{l=1}^{\infty} 8N_\infty(\epsilon/16,\mathcal{R})e^{-n\epsilon^2/512\tau} < \infty$$

for every $\epsilon > 0$. This condition is true since

$$\begin{aligned}\sum_{n=1}^{\infty} e^{-n\epsilon^2/512\tau} &\leq \int_{x=0}^{\infty} e^{-x\epsilon^2/512\tau}dx \\ &\leq \frac{512\tau}{\epsilon^2}e^{-\epsilon^2/512\tau}\end{aligned}$$

which is finite for all $\epsilon > 0$. □

The first part of the theorem provides a distribution-free finite sample result. It says that given a sufficiently large sample size, with a probability $1-\delta$, we have

$$R(\tilde{\alpha}_{\text{sample}}) - R(\tilde{\alpha}^*) \leq \epsilon,$$

i. e. the cost of the sample-based solution is within ϵ of the lowest achievable cost (which can only be computed if all error distributions are known). This result is distribution-free, and only depends on the sample size. It is best possible result in that stronger results such as showing $\delta = 0$ is not possible, since $\tilde{\alpha}^*$ depends on a distribution (which is not finite-dimensional) and $\tilde{\alpha}_{\text{sample}}$ depends on a *finite* sample. The asymptotic results such as $R(\tilde{\alpha}_{\text{sample}}) \to R(\tilde{\alpha}^*)$ are more common in statistics literature [11]. The finite sample result is stronger in that it implies the asymptotic result, and also establishes that the method is justified even for small sample sizes. To our knowledge, these are the first convergence results shown for this method of identification.

3.2 Correlated Parameters and Outputs

We now consider the case studied in [3] based on the cost function in Eq (2.1). The sample-based cost in this case is given by

$$\hat{R}(\tilde{\alpha}) = [(\tilde{\alpha}-\bar{\alpha})\,(\hat{M}(\bar{\alpha}) - E[y])] \begin{bmatrix} \hat{C}_\alpha \hat{C}_{\alpha,y} \\ \hat{C}_{y,\alpha}\hat{C}_y \end{bmatrix}^{-1} [(\tilde{\alpha}-\bar{\alpha}(M(\bar{\alpha})-\bar{y})]^T, \qquad (3.3)$$

where (i) $\bar{\alpha}$ and $\bar{y}$ are the sample means of the measurements of α and y, respectively, and (ii) $\hat{C}_\alpha$, $\hat{C}_{\alpha,y}$, $\hat{C}_{y,\alpha}$ and $\hat{C}_y$ are the empirical covariances computed based on the sample corresponding to C_α, $C_{\alpha,y}$, $C_{y,\alpha}$, and C_y, respectively. The solution to this equation is is given by (see [3] for the details of the derivation)

$$\tilde{\alpha}_{\text{sample}} = \bar{\alpha} + (\hat{C}_{\alpha y} - \hat{C}_{\alpha\alpha}S^T)\hat{C}_{dd}^{-1}(\hat{M}(\alpha) - \bar{y})$$

where $\hat{C}_{dd} = \hat{C}_y + C_M - S\hat{C}_{\alpha y} - \hat{C}_{y\alpha}S^T$. Let $\tilde{\alpha}^*$ minimize $R(.)$ in Eq (2.2).

The discussion of this case follows along the lines of previous section, except the technical details are more involved.

Theorem 3.2 *Consider the bounded parameters and outputs such that $\alpha \in [-A,A]^{d_\alpha}$ and $y \in [-B,B]^{d_y}$, and the condition Eq (3.1) be satisfied. Let Λ_α, $\Lambda_{\alpha,y}$, $\Lambda_{y,\alpha}$, and Λ_y denotes the upperbounds on the entries of C_α^{-1}, $C_{\alpha,y}^{-1}$, $C_{y,\alpha}^{-1}$, and C_y^{-1}, respectively. Let D and E be upperbounds on $\frac{\partial m_i}{\partial \tilde{a}_j}$ for all i,j and $M(.)$, respectively, and let $R(\tilde{\alpha}) \leq \tau$ for all $\tilde{\alpha}$. Let*

$$\begin{aligned} L &= 4(d_\alpha\Lambda_\alpha A + d_y\Lambda_y BD) \\ &\quad +(d_\alpha+d_y)(\Lambda_{\alpha,y}+\Lambda_{y,\alpha})[E+2B+2DA]. \end{aligned}$$

Given a sample of size at least

$$n = \frac{512\tau}{\epsilon^2}\left[d_\alpha \ln(32A/L\epsilon) + \ln(8/\delta)\right],$$

we have, for any δ and ϵ,

$$P\left[R(\tilde{\alpha}_{\text{sample}}) - R(\tilde{\alpha}^*) > \epsilon\right] < \delta.$$

Furthermore, $R(\tilde{\alpha}_{\text{sample}}) \to R(\tilde{\alpha}^*)$, as $n \to \infty$.

Proof: The proof is almost identical to that in Theorem 3.1 except for the estimation of the Lipschitz constant L. For $\bar{\alpha} = (\bar{a}_1,\ldots,\bar{a}_{d_\alpha})$, we have the following upper bound

$$\begin{aligned} \frac{\partial R}{\partial \tilde{a}_i} &\leq 2d_\alpha\Lambda_\alpha \max_{\tilde{a}_i,\bar{a}_i}\{\tilde{a}_i - \bar{a}_i\} \\ &\quad +2d_y\Lambda_y\frac{\partial M}{\partial \tilde{a}_i}\max_{\tilde{a}_i,\bar{a}_i}\{\tilde{a}_i-\bar{a}_i\} \\ &\quad +(d_\alpha+d_y)\Lambda_{\alpha,y} \\ &\qquad \left[\max_{\alpha,y}\{M(\alpha)-y\} + \frac{\partial M}{\partial \tilde{a}_i}\max_{\tilde{a}_i,\bar{a}_i}\{\tilde{a}_i-\bar{a}_i\}\right] \\ &\quad +(d_\alpha+d_y)\Lambda_{y,\alpha} \\ &\qquad \left[\max_{\alpha,y}\{M(\alpha)-y\} + \frac{\partial M}{\partial \tilde{a}_i}\max_{\tilde{a}_i,\bar{a}_i}\{\tilde{a}_i-\bar{a}_i\}\right] \\ &\leq 4d_\alpha\Lambda_\alpha A \\ &\quad +(d_\alpha+d_y)(\Lambda_{\alpha,y}+\Lambda_{y,\alpha})[E+2B+2DA] \\ &\quad +4d_y\Lambda_y DB, \end{aligned}$$

which yields the required bound on L. □

4 Feedfoward Sigmoid Networks

We consider the class of feedforward neural networks with a single hidden layer of l nodes and a single output node. The output of the network corresponding to input $x \in [-C,C]^{d_x}$ is given by

$$f_{w,\gamma}(x) = \sum_{j=1}^{l} a_j\sigma(b_j^T x + t_j)$$

where: (i) $\sigma(z) = \frac{1}{1+e^{-\gamma z}}$, for $0 < \gamma < \infty$, called the *gain*, and (ii) $w = (w_1, w_2, \ldots, w_{l(d+2)})$ is the *weight* or *parameter vector* consisting of a_1, a_2, …, a_l,

$b_{11}, b_{12}, \ldots, b_{1d}, \ldots, b_{l1}, \ldots, b_{ld}$ and $t_1, t_2, \ldots, t_l$. We employ the function $f_{w,\gamma}(.)$ for $w \in [-W, W]^{l(d+2)}\}$ to implement $M(\alpha, x) = f_{\alpha,\gamma}(x)$ and $A = W$. The gain γ is a priori fixed, and we set $w = \tilde{\alpha}$, which has $d_\alpha = l(d+2)$ components. For simplicity we consider y to be 1-dimensional.

For computing the solution based on neural networks, we need its sensitivity matrix S, which is obtained by the following partial derivatives of f_w with respect to w_i [13]. We have, for $x = (x_1, x_2, \ldots, x_d)$, $b_j = (b_{j1}, b_{j2}, \ldots, b_{jd})$, and $b_j^T x = \sum_{i=1}^{d} b_{ji} x_i$,

$$\frac{\partial f_w}{\partial a_j} = \sigma(\sum_{i=1}^{d} b_{ji} x_i + t_j)$$

$$\frac{\partial f_w}{\partial b_{ji}} = a_j \sigma'(\sum_{i=1}^{d} b_{ji} x_i + t_j)\gamma x_i$$

$$\frac{\partial f_w}{\partial t_j} = a_j \sigma'(\sum_{i=1}^{d} b_{ji} x_i + t_j)\gamma.$$

To obtain the bound D, first note that

$$\frac{d\sigma(z)}{dz} = \gamma\sigma(z)[1 - \sigma(z)] \leq \gamma/4$$

since the right hand side is maximized at $\sigma(z) = 1/2$. Then we have

$$\frac{\partial f_w}{\partial a_j} \leq 1$$

$$\frac{\partial f_w}{\partial b_{ji}} \leq WC\gamma^2/4$$

$$\frac{\partial f_w}{\partial t_j} \leq W\gamma^2/4.$$

Thus, the bound on the derivative D required in Theorems 3.1 and 3.2 is given by

$$D = \max(1, CW\gamma^2/4, W\gamma^2/4).$$

In addition, the bound E in Theorem 3.1 is given by lA, since $\sigma(z) \leq 1$ for all z.

5 Conclusions

We presented a method for the identification of a parametrized time-invariant non-linear plant using a smooth model, where plant parameters and inputs/outputs can be measured with noise. An initial plant model is obtained by utilizing the domain-specific knowledge in terms of fundamental plant equations, which in general only partially capture the plant dynamics. Once the initial model is fixed, measurements are collected on the plant parameters and output. We showed that the iid measurements can be fused with the initial plant model by recomputing the parameters. We established the asymptotic convergence of our computational procedure, and derived its finite sample results of the PAC learning kind. Our method is based on the empirical versions of the closed-form solutions derived in nuclear engineering literature for an ideal version of the problem based on the sensitivity analysis of the identifier. The computation is of low-order polynomial complexity, and involves matrix multiplication and inversion. We also provided the details about an identifier based on a sigmoid feedforward neural network.

We assume that various distributions are stationary. If they are time-varying [1], then observations must be time-windowed to implement the proposed method. The time-window must be carefully chosen to provide the suitable convergence properties. These aspects are of future interest.

Acknowledgements

This research is sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp., and the Office of Naval Research under order N00014-96-F-0415.

References

[1] J. Barhen, D. G. Cacuci, J. J. Wagschal, M . A. Bjerke, and C. B. Mullins. Uncertainity analysis of time-dependent nonlinear systems: Theory and application to transient thermal hydraulics. *Nuclear Sciene and Engineering*, 81:23–44, 1982.

[2] J. Barhen, J. J. Wagschal, and Y. Yeivin. Lagrange multipliers in data adjustments. *Transactions of Americal Nuclear Society*, 34:31–32, 1980.

[3] J. Barhen, J. J. Wagschal, and Y. Yeivin. Response-parameter correlations in uncertainty analysis. *Transactions of Americal Nuclear Society*, 35:246–247, 1980.

[4] P. Billingsley. *Probability and Measure*. John Wiley and Sons, New York, second edition, 1986.

[5] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachisetts, 1994.

[6] A. Krzyzak, T. Linder, and G. Lugosi. Nonparametric estimation and classification using radial basis function nets and empirical risk minimization. *IEEE Transactions on Neural Networks*, 7(2):475–487, 1996.

[7] A. U. Levin and K. S. Narendra. Recursive identification using feedforward neural networks. *International Journal of Control*, 61(3):533–547, 1995.

[8] L. Ljung. *System Identification*. Prentice Hall, Engelwood Cliffs, NJ, 1987.

[9] K. S. Narendra and A. M. Annaswamy. *Stable Adaptive Systems*. Prentice- Hall, 1989.

[10] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[11] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Academic Press, New York, 1983.

[12] N. S. V. Rao. Multiple sensor fusion under unknown distributions. *Journal of Franklin Institute*, 336(2):285–299, 1999.

[13] N. S. V. Rao. Simple sample bound for feedforward sigmoid networks with bounded weights. *Neurocomputing*, 29:115–122, 1999.

[14] E. D. Sontag. *Mathematical Control Theory*. Springer-Verlag, 1990.

[15] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[16] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[17] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[18] V. N. Vapnik. *Statistical Learning Theory*. John-Wiley and Sons, New York, 1998.

[19] J. J. Wagschal and Y. Yeivin. The significance of Lagrange multipliers in cross-section adjustment. *Transactions of Americal Nuclear Society*, 34:776–777, 1980.

[20] C. R. Weisbin, R. W. Peelle, J. H. Marable, P. Collins, E. Kujawski, E. Greenspan, and G. de Saussure. Sensitivity and uncertainity analysis of reactor performance parameters. In *Advances in Nuclear Science and Technology*, volume 14. Plenum Press, New York, 1981.