

Chapter 1

MULTISENSOR FUSION UNDER UNKNOWN DISTRIBUTIONS

Finite-Sample Performance Guarantees

Nageswara S. V. Rao

Center for Engineering Science Advanced Research

Computer Science and Mathematics Division

Oak Ridge National Laboratory

Oak Ridge, TN 37831-6355

raons@ornl.gov

Abstract We consider a multiple sensor system such that for each sensor the outputs are related to the actual feature values according to a certain probability distribution. We present an overview of informational and computational aspects of a fuser that is required to combine the sensor outputs to more accurately predict the feature, when the sensor distributions are unknown but iid measurements are given. Our focus is on methods to compute a fuser with probabilistic guarantees in terms of distribution-free performance bounds based on a finite sample. We first discuss a number of methods based on the empirical risk minimization approach. These methods yield a fuser which is guaranteed, with a high probability, to be close to an optimal fuser (computable only under a complete knowledge of sensor distributions). Then we describe the isolation fusers that are guaranteed to perform at least as good as the best sensor, and the projective fusers that are guaranteed to perform at least as good as the best subset of sensors. Then we consider physical systems wherein the training data consisting of actual physical values is not available. We discuss methods that utilize the physical laws to obtain a suitable fuser under these conditions.

Keywords: Sensor fusion, information fusion, empirical risk minimization, vector spaces, neural networks.

1. INTRODUCTION

Fusion or combination of information from multiple sources to achieve performances exceeding those of individual sources has been recognized for

centuries in diverse areas such as political economy models (Grofman and Owen, 1986), and composite methods (de Laplace, 1818); a brief overview of these works can be found in (Madan and Rao, 1999). In the twentieth century, such fusion methods continued to be applied in a wide spectrum of areas such as reliability (von Neumann, 1956), forecasting (Granger, 1989), pattern recognition (Chow, 1965), neural networks (Hashem et al., 1994), decision fusion (Dasarathy, 1994; Varshney, 1996), and statistical estimation (Brieman, 1996; Juditsky and Nemirovski, 1996).

In engineering systems, the fusion methods have been proven to be particularly important since they can provide multiple sensor capabilities, which are significantly beyond those of single sensor systems. In particular, many researchers realized the fundamental limitations of single sensor systems in a number of application areas such as robotics (Brady, 1988; Abidi and Gonzalez, 1992), and tracking (Bar-Shalom and Li, 1995), thereby motivating the deployment of multiple sensors. Information fusion is very important in such multiple sensor systems, and is often referred to as *multisensor fusion* or simply *sensor fusion*. The overall objective is the same across the disciplines, namely to “fuse” the information from many different sensors to overcome the limitations of a single sensor. Although one would expect a fuser to perform better than any of the sensors, it is not clear as to how to design such fuser. Thus, systematic approaches to fuser design are very critical to the overall performance, for an inappropriate fuser can render the system worse than the worst individual sensor.

In engineering systems, the sensor fusion problems present technical challenges in ways unprecedented in other disciplines. Early information fusion methods require either independence of sensor errors or closed-form analytical expressions for sensor distributions. In the first case, a general majority rule suffices, while in the second a fusion rule can be computed using the Bayesian methods. Several popular distributed decision fusion methods belong to the latter class (Dasarathy, 1994; Varshney, 1996). In engineering systems, however, independence can seldom be assured and, in fact, may not be satisfied. Bayesian methods are not particularly conducive either from an information or from a cost perspective. Typically, the fusion rules are “selected” from a specific function class chosen by the designer to ensure the convergence of the fuser computation. On the other hand, the sensor distributions are not within the hands of the user, and can be arbitrarily complicated in complex engineering systems. As a result, the problem of obtaining the sensor distributions required by the Bayesian methods is more difficult, in an information-theoretic sense, than the fusion problem itself (Vapnik, 1982). In addition, deriving closed form expressions for sensor distributions is a very difficult and expensive task in these systems since it requires the knowledge of a variety of areas such as device physics, electrical engineering, and statistical modeling.

The fusion problems arising in operational engineering and robotic systems, however, have a positive side: it is easy to collect “data” by sensing objects and environments with known parameters. Thus, practical solutions to these sensor fusion problems must exploit the empirical data available from the observation and/or experimentation, which is the main topic of this paper. We present a summary of methods for fusion rule estimation from empirical data, which has become possible largely due to the developments in empirical process theory and computational learning theory. Our main focus is on methods that provide *performance guarantees based on finite samples* from a statistical perspective. In particular, we do not cover adhoc fusion rules with no performance bounds or results based on asymptotic guarantees that are valid *only* as the sample size approaches infinity. Our approach is based on a statistical formulation of the problem for the most part with the exception of physical systems in Section 8. In this respect, our solutions do not fully capture the non-statistical aspects such as calibration and registration, which can be incorporated into a suitable cost function. But, our results provide an analytical justification of sample-based approaches to the sensor fusion problem, and establish the basic tractability of the solutions. We believe that our performance bounds can be improved in specific cases by suitably incorporating the application specific details.

The organization of this paper is as follows. We present the problem formulation in Section 2. In Section 3, we present several solutions based on the empirical risk minimization methods. In Section 4, we present solutions based on non-linear statistical estimators. We describe applications of these methods in Section 5. In Section 6, we address the issues of relative performance of the composite system and the individual sensors. We describe the metausers in Section 7. Then, in Section 8, we discuss a special class of the fusion problem which can be efficiently solved by utilizing the physical laws. In Section 9, we present an approach to incorporating known conditional distributions into the fuser computation. Our presentation is tutorial in nature in that we describe only the main results, and the specific details can be found in the references.

2. PROBLEM FORMULATION

In a *generic sensor system* of N sensors, the sensor S_i , $i = 1, 2, \dots, N$, outputs $Y^{(i)} \in \mathfrak{R}^d$ corresponding to input $X \in \mathfrak{R}^d$ according to the distribution $P_{Y^{(i)}|X}$. The input X is the quantity that needs to be “estimated” or “measured” by the sensors, such as the presence of a target or a value of the feature vector. The *expected error* of the sensor S_i is given by

$$I(S_i) = \int C(X, Y^{(i)}) dP_{Y^{(i)}, X},$$

where $C : \mathfrak{R}^d \times \mathfrak{R}^d \mapsto \mathfrak{R}$ is a cost function. $I(S_i)$ is a measure of how good the sensor S_i is in “sensing” the input feature X . For example, if S_i is a target

detector or classifier, a choice could be $X \in \{0, 1\}$ and $Y^{(i)} \in \{0, 1\}$, where $X = 1$ (0) corresponds to the presence (absence) of a target. Then

$$I(S_i) = \int [X \oplus Y^{(i)}] dP_{Y^{(i)}, X}$$

corresponds to the probability of misclassification (false alarm or missed detection) of S_i , where \oplus is the exclusive-OR operation.

There are two types of errors that a sensor can make. The *measurement error* corresponds to the randomness involved in measuring a particular value of the feature X , which is distributed according to $P_{Y^{(i)}|X}$. The *systematic error* at X corresponds to $E[C(X, Y^{(i)})|X]$ which must be 0 in the case of a perfect sensor.

Example 2.1: Consider a sensor system consisting of two sensors. For the first sensor, we have $Y^{(1)} = a_1 X + Z$, where Z is normally distributed with zero mean, and is independent of X . That is, first sensor has a scaling error and a random additive error. For the second sensor, we have $Y^{(2)} = a_2 X + b_2$, which has a scaling and bias error. Let X be uniformly distributed over $[0, 1]$, and $C[X, Y] = (X - Y)^2$. Then, we have $I(S_1) = (1 - a_1)^2$ and $I(S_2) = (1 - a_2 - b_2)^2$, which are non zero in general. \square

We consider a fuser $f : \mathfrak{R}^{Nd} \mapsto \mathfrak{R}^d$ that combines the outputs of sensors $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(N)})$ to produce the fused output $f(Y)$. The *expected error* of the fuser f is given by

$$I_F(f) = \int C(X, f(Y)) dP_{Y, X}$$

where $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(N)})$. The objective of fusion is to ensure that $I_F(f)$ is as small as possible. Note that a fuser must account for both the systematic and measurement errors of the sensors in order to achieve a low expected error. The fuser can be chosen from a family of fusion rules $\mathcal{F} = \{f : \mathfrak{R}^N \mapsto \mathfrak{R}\}$ and, the *expected best* fusion rule f^* minimizes $I_F(\cdot)$ over \mathcal{F} , i.e.

$$I_F(f^*) = \min_{f \in \mathcal{F}} I_F(f).$$

For example, \mathcal{F} could be the set of sigmoidal neural networks obtained by varying the weight vector for a fixed architecture. In this case $f^* = f_{w^*}$ corresponding to the weight vector w^* that minimizes $I_F(\cdot)$ over all the weight vectors.

Example 2.1: (Continued) Consider the fuser

$$f(Y^{(1)}, Y^{(2)}) = \frac{Y^{(1)}}{2a_1} + \frac{1}{2a_2}(Y^{(2)} - b).$$

For this fuser, we have $I_F(f) = 0$, since the bias b is subtracted from $Y^{(2)}$ and the multipliers cancel the scaling error. In practice, however, such fuser can be designed only when the sensor distributions are known. \square

In our formulation, since $I_F(\cdot)$ depends on the *unknown* error distribution $P_{Y,X}$, f^* cannot be computed even in principle. We consider that only an independently and identically distributed (iid) l -sample

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_l, Y_l)$$

is given, where $Y_i = (Y_i^{(1)}, Y_i^{(2)}, \dots, Y_i^{(N)})$ and $Y_i^{(j)}$ is the output of S_j in response to input X_i . Now the question is what type of performance guarantees are reasonable to expect under this formulation? The answer can be found in the area of Probably Approximately Correct (PAC) learning (Vapnik, 1982; Valiant, 1984). We consider methods to compute an estimator \hat{f} , based *only* on a sufficiently large sample, such that

$$P_{Y,X}^l \left[I_F(\hat{f}) - I_F(f^*) > \epsilon \right] < \delta \quad (1.1)$$

where $\epsilon > 0$ and $0 < \delta < 1$, and $P_{Y,X}^l$ is the distribution of iid l -samples. For simplicity, we denote $P_{Y,X}^l$ by P . Informally, this condition states that the “error” of \hat{f} is within ϵ of the optimal error (of f^*) with an arbitrary high probability $1 - \delta$, *irrespective* of the underlying sensor distributions. Intuitively, it is a reasonable criterion since \hat{f} is to be “chosen” from an infinite set, namely \mathcal{F} , based only on a finite sample. More concretely, conditions that are strictly stronger than Eq 1.1 are generally not possible. For example, consider the condition $P_{Y,X}^l [I_F(\hat{f}) > \epsilon] < \delta$ for the case $\mathcal{F} = \{f : [0, 1]^N \mapsto \{0, 1\}\}$. This condition cannot be satisfied, since for any $f \in \mathcal{F}$, there exists a distribution for which $I_F(f) > 1/2 - \rho$ for any $\rho \in [0, 1]$; see Theorem 7.1 of (Devroye et al., 1996) for details.

Example 2.1: (Continued) To illustrate the effects of finite samples, consider that we generate three values for X given by $\{0.1, 0.5, 0.9\}$ with corresponding Z values given by $\{0.1, -0.1, -0.3\}$. The corresponding values for $Y^{(1)}$ and $Y^{(2)}$ are given by $\{0.1a_1 + 0.1, 0.5a_1 - 0.1, 0.9a_1 - 0.3\}$ and $\{0.1a_2 + b_2, 0.5a_2 + b_2, 0.9a_1 + b_2\}$ respectively. Consider the class of linear fusers such $f(Y^{(1)}, Y^{(2)}) = w_1Y^{(1)} + w_2Y^{(2)} + w_3$. Based on the measurements, the following weights enable the fuser outputs to exactly match X values for each of the measurements:

$$w_1 = \frac{1}{0.2 - 0.4a_1}, w_2 = \frac{1}{0.4a_2} \quad \text{and} \quad w_3 = \frac{0.1a_1 + 0.1}{0.4a_1 + 0.1} - \frac{0.1a_2 + b_2}{0.4a_2}.$$

While the fuser with these weights achieves zero error on the measurements it does not achieve zero value for I_F . Note that a fuser with zero expected

error exists, and can be computed if the sensor distributions are given. The idea behind the criterion in Eq 1.1 is to achieve performances close to optimal fuser using only a sample. To meet this criterion one needs to select a suitable \mathcal{F} , and then achieve small error on a sufficiently large sample, as will be illustrated subsequently. \square

The generic sensor fusion problem formulated here is fairly general and requires very little information about the sensors: a sensor could be a hardware device, a software module or a combination. We now briefly describe some concrete examples, which are described in detail in Section 5. In the door detection example (Rao, 1997b), S_i is an ultrasonic or infrared detector, and the objective is to detect doors that are wide enough for a robot to pass through. In the example of function estimation (Rao, 1997a), S_i is a software module that predicts a function value with certain errors, and the objective is to fuse a number of such modules to improve the prediction accuracy. In both examples, the sensors are available so that the training sample can be collected by experimentation.

2.1 RELATED WORKS

Due to the generic nature of the sensor fusion problem described here, it is related to a number of similar problems applied in a wide variety of areas. Here we briefly show its relationship to some of the well-known methods in engineering areas. If the sensor error distributions are known, several fusion rule estimation problems have been solved by methods not requiring the samples. Some of the earlier work in this direction is due to (Chow, 1965). This problem is also related to the group decision models studied extensively in political economy; for example see (Grofman and Owen, 1986). Indeed, early majority methods of combining the outputs of probabilistic Boolean elements date back to 1786 under the name of Condorcet jury models. The distributed detection problem based on probabilistic formulations has been extensively studied (Varshney, 1996). These problems are special cases of the generic fusion problem such that $X \in \{0, 1\}$ and $Y \in \{0, 1\}^N$, but the difference is that they assume that various probabilities are available. In the systems where only measurements are available the existing methods are not useful, but the solutions to the generic sensor fusion problem are applicable. In general, however, much tighter performance bounds are possible since distribution detection is a special (namely Boolean) case of the generic sensor fusion problem (Rao and Oblow, 1994; Rao and Oblow, 1997). Also, in many cases the solutions based on known distributions case can be converted to sample-based ones (Rao, 1996).

Many of the existing information integration techniques are based on maximizing a posteriori probabilities of hypotheses under a suitable probabilistic model. However, in situations where the probability densities are unknown (or

difficult to estimate) such methods are ineffective. One alternative is to estimate the density based on a sample. But, as illustrated in general by (Vapnik, 1982), the density estimation is more difficult than the subsequent problem of estimating a function chosen from a family with bounded capacity. This property holds for several pattern recognition and regression estimation problems (Vapnik, 1982). In the context of feedforward neural networks that are employed to “learn” a function based on sample, the problem is to identify the weights of a network of chosen architecture. The choice of weights picks a particular network \hat{f} from a family \mathcal{F} of neural networks of a particular architecture. This family \mathcal{F} satisfies bounded capacity (Anthony, 1994) and Lipschitz property (Tang and Koehler, 1994). Both these properties are conducive for the statistical estimation of \hat{f} as explained in the next section. On the other hand, no such information is available about the class from which the unknown density is chosen, which makes it difficult to estimate the density.

3. EMPIRICAL RISK MINIMIZATION

In this section we present fusion solutions based on the empirical risk minimization methods (Vapnik, 1982). Consider that the empirical estimate

$$I_{emp}(f) = \frac{1}{l} \sum_{i=1}^l \left[X_i - f \left(Y_i^{(1)}, Y_i^{(2)}, \dots, Y_i^{(N)} \right) \right]^2$$

is minimized by $\hat{f} \in \mathcal{F}$. Using Vapnik’s empirical risk minimization method (Vapnik, 1982), for example, we can show (Rao, 1995) that if \mathcal{F} has finite capacity, then under bounded error, or bounded relative error for sufficiently large sample

$$P_{Y,X}^l \left[I_F(\hat{f}) - I_F(f^*) > \epsilon \right] < \delta$$

for arbitrarily specified $\epsilon > 0$ and $\delta, 0 < \delta < 1$. Typically, the required sample size is expressed in terms of ϵ and δ and the parameters of \mathcal{F} .

In this section, we describe several different classes of \mathcal{F} and their sample size estimators. We first describe the most general condition available on \mathcal{F} under which the performance guarantee in Eq 1.1 can be assured. The sample sizes in this case are based on the scale-sensitive dimension (Alon et al., 1993). Then we present specific cases of feedforward networks and vector spaces that will also provide the performance guarantees. We apply the specific properties of these function classes to derive the sample size estimates, which are in general tighter than the general bounds.

3.1 SCALE-SENSITIVE DIMENSION

We first present the definition of scale-sensitive dimension (Alon et al., 1993); also see (Anthony and Bartlett, 1999) for a more detailed discussion. Let \mathcal{F} be

a class of $[0, 1]$ -valued functions on some domain set D and let ρ be a positive real number. We say that \mathcal{F} P_ρ -shatters a set $A \subseteq D$ if there exists a function $s : A \mapsto [0, 1]$ such that for every $E \subseteq A$ there exists some $f_E \in \mathcal{F}$ satisfying: for every $x \in A - E$, $f_E(x) \leq s(x) - \rho$, and for every $x \in E$, $f_E(x) \geq s(x) + \rho$. Let the P_ρ -dimension of \mathcal{F} , denoted by $P_\rho\text{-dim}(\mathcal{F})$, be the maximal cardinality of a set $A \subseteq D$ that is P_ρ -shattered by \mathcal{F} .

The following theorem presents an estimate of the sample size to ensure the condition in Eq 1.1 when \mathcal{F} has finite a scale-sensitive dimension (Alon et al., 1993) and $X \in [0, 1]$.

Theorem 1 (Rao, 1999b) *Let f^* and \hat{f} denote the expected best and empirical best fusion rules chosen from a function class \mathcal{F} with range $[0, 1]$. Given a sample of size*

$$\frac{5040}{\epsilon^2} \max \left\{ d_S \ln^2 \frac{50d_S}{\epsilon}, \ln \frac{48}{\delta} \right\}$$

where $d_S = P_{\epsilon/4}\text{-dim}(\mathcal{F})$, we have $P \left[I_F(\hat{f}) - I_F(f^*) > \epsilon \right] < \delta$.

Theorem 1 provides us the sufficient sample size as a function of ϵ , δ and d_S . One needs to simply compute \hat{f} that minimizes the empirical error on a sample of sufficient size to ensure the performance condition. The scale-sensitive dimension is known for several classes such as neural networks and linear combinations (Anthony and Bartlett, 1999). While the above bound is not tight (due to its general applicability), it does establish that the fuser rule estimation is a tractable problem from a statistical standpoint. The result of Theorem 1 is more general than that in (Rao, 1994; Rao, 1995) which is based on capacity of \mathcal{F} (Vapnik, 1982) in that finiteness of capacity implies that of scale-sensitive dimension but not vice versa.

This theorem can be generalized in a straight forward manner to handle the cases: (a) $Y^{(i)}$ is a multi-dimensional vector from \mathfrak{R}^d , and/or (b) $X \in [0, \tau]$, $\tau > 0$. The cost function can also be generalized to Lipschitz cost functions with an appropriate change in the sample size (Rao and Protopopescu, 1998).

The sample bound is based on uniform convergence of empirical means to their expectations for function classes, which are available from the empirical process theory (Pollard, 1990; van der Vaart and Wellner, 1996; Talagrand, 1994) and its applications to machine learning (Vapnik, 1995; Haussler, 1992). Results of this kind are available based on a number of characterizations of \mathcal{F} such as pseudo-dimension (Pollard, 1990), fat VC-dimension (Kearns and Schapire, 1994), scale-sensitive dimension (Alon et al., 1993), graph dimension (Dudley, 1987), and Euclidean parameters (Talagrand, 1994; Nolan and Pollard, 1987), which can be used to obtain sample size estimates along the lines of Theorem 1. Finiteness of these parameters is only sufficient for the ‘‘learnability’’ of bounded functions, while that of the scale sensitive dimension is both

necessary and sufficient (Alon et al., 1993). Moreover, the latter is only such deterministic quantity known to us, while other similar quantities are based on expected capacity or entropy (Vapnik, 1995; van der Vaart and Wellner, 1996).

The empirical risk minimization simply requires that \hat{f} minimize the empirical error, and does not specify methods to *compute* it. In general computation of \hat{f} in this general framework is intractable. For example, in the special case that \mathcal{F} is set of feedforward neural networks with threshold hidden units, this problem is NP-complete even for simple architectures (Blum and Rivest, 1992). Also, since no restrictions are placed on the information of the sensors, even if there is no probabilistic error in the sensors, several multisensor fusion problems are NP-complete (Tsitsiklis and Athans, 1985; Rao, 1991; Rao et al., 1993). For the more restrictive cases where \mathcal{F} is chosen to be a special class, the computational problems is not always easier. Indeed very few subclasses of the fusion estimation problems are known to be inherently polynomial-time solvable. In linearly separable systems (Rao, 1994), the associated computational problem can be solved (exactly) as a quadratic programming problem, which be solved in polynomial time.

In the next two sections, we describe two practical solutions to the sensor fusion problem. The first method is based on feedforward neural networks. It is applied in a number of practical cases with good results although the underlying computational problem is not known to be polynomial-time computable. The second method is based on the vector space method which includes linear fusers as a special case. This method is widely used in practice and has the advantage of being polynomial-time solvable. Thus, the choice of \mathcal{F} has a significant effect on both the sample complexity as well as computational complexity.

3.2 FEEDFORWARD SIGMOIDAL NETWORKS

We consider a feedforward network with a single hidden layer of k nodes and a single output node. The output of the j th hidden node is $\sigma(b_j^T y + t_j)$, where $y \in [-B, B]^N$, $b_j \in \mathfrak{R}^N$, $t_j \in \mathfrak{R}$, and the nondecreasing $\sigma : \mathfrak{R} \mapsto [-1, +1]$ is called the *activation function*. The output of the network corresponding to input y is given by

$$f_w(y) = \sum_{j=1}^k a_j \sigma(b_j^T y + t_j)$$

where $w = (w_1, w_2, \dots, w_{k(d+2)})$ is the *weight vector* of the network consisting of $a_1, a_2, \dots, a_k, b_{11}, b_{12}, \dots, b_{1d}, \dots, b_{k1}, \dots, b_{kd}$, and t_1, t_2, \dots, t_k . Let the set of *sigmoidal feedforward networks with bounded weights* be denoted by

$$\mathcal{F}_W^\gamma = \{f_w : w \in [-W, W]^{k(d+2)}\}$$

where $0 < \gamma < \infty$, and $\sigma(z) = \tanh^{-1}(\gamma z)$, $0 < W < \infty$.

The following theorem provides several sample size estimates for the fusion rule estimation problem based on the different properties of neural networks.

Theorem 2 (Rao, 1999a) *Consider the class of feedforward neural networks $\mathcal{F}_{\mathcal{W}}^\gamma$. Let $X \in [-A, A]$ and $R = 8(A + kW)^2$. Given a sample of size at least*

$$\frac{16R}{\epsilon^2} \left(\ln(18/\delta) + 2 \ln(8R/\epsilon^2) + \ln(2\gamma^2 W^2 kR/\epsilon) + \frac{\gamma W^2 kR}{\epsilon} \left[\left(\frac{\gamma W^2 kR}{\epsilon} - 1 \right)^{N-1} + 1 \right] \right),$$

the empirical best neural network \hat{f}_w in $\mathcal{F}_{\mathcal{W}}^\gamma$ approximates the expected best f_w^ in $\mathcal{F}_{\mathcal{W}}^\gamma$ such that*

$$P \left[I_F(\hat{f}_w) - I_F(f_w^*) > \epsilon \right] < \delta.$$

The same condition can also be ensured under the sample size

$$\frac{16R}{\epsilon^2} \left(\ln(18/\delta) + 2 \ln(8R/\epsilon^2) + k(d+2) \ln(L_w R/\epsilon) \right)$$

where $L_w = \max(1, WB\gamma^2/4, W\gamma^2/4)$, or, for $\gamma = 1$,

$$\frac{128R}{\epsilon^2} \max \left\{ \ln \left(\frac{8}{\delta} \right), \ln \left(\frac{16e(k+1)R}{\epsilon} \right) \right\}.$$

These sample sizes are based on three qualitatively different parameters of \mathcal{F} , namely, (a) Lipschitz property of $f(y) \in \mathcal{F}$ with respect to input y (Rao, 1994), (b) compactness of weight set and smoothness of $f \in \mathcal{F}$ with respect to weights, and (c) VC-dimension of translates of sigmoid units (Lugosi and Zeger, 1995). The three sample estimates provide three different means for controlling the sample size depending on the available information and intrinsic characteristics of the neural network class $\mathcal{F}_{\mathcal{W}}^\gamma$. The sample sizes in the first and second bounds can be modified by changing the parameter γ . For example, by choosing $\gamma = \frac{\epsilon}{W^2 k R}$ the first sample size can be reduced to a simpler form

$$\frac{16R}{\epsilon^2} \left[\ln(18/\delta) + \ln \left(\frac{128R}{W^2 k \epsilon^2} \right) \right].$$

Also, by choosing $\gamma^2 = \frac{4}{W \max(1, B)}$, we have a simpler form of the second sample size estimate

$$\frac{16R}{\epsilon^2} \left[\ln \left(\frac{1152}{\delta} \right) + k(d+2) \ln(R/\epsilon) \right],$$

for $R \geq 1$. In practice, it could be useful to compute all three bounds and choose the smallest one.

These three sample bounds are based on utilizing the boundedness of weights of neural network. If boundedness is not satisfied, one utilizes either VC-bounds (Koiran and Sontag, 1997) or scale-sensitive dimension (Anthony and Bartlett, 1999) to estimate the sample size, which in general results in larger sample bounds; see (Rao, 1999e) for details. We believe, however, in practical implementations the weights are always bounded.

The problem of computing the empirical best neural network \hat{f}_w is still difficult, and is NP-complete for very general subclass of $\mathcal{F}_{\mathcal{V}}^\gamma$ (Sima, 1996). However, in several practical cases very good results have been obtained using the backpropagation algorithm which provides an approximation to \hat{f}_w . This algorithm is very easy to implement, and is also available in a number of commercial neural network packages. For the vector space method in the next section, the computation problem is polynomial-time solvable.

3.3 VECTOR SPACE METHODS

We now present a method that is attractive from an analytical point of view. Consider that \mathcal{F} forms a finite dimensional vector space. In this case: (a) sample size is a simple function of the dimensionality of \mathcal{F} , (b) \hat{f} can be easily computed by well-known least square methods in polynomial time, and (c) no smoothness conditions are required on the functions or distributions.

Theorem 3 (Rao, 1998c) *Let f^* and \hat{f} denote the expected best and empirical best fusion functions chosen from a vector space \mathcal{F} of dimension d_V and range $[0, 1]$. Given an iid sample of size*

$$\frac{512}{\epsilon^2} \left[d_V \ln \left(\frac{64e}{\epsilon} + \ln \frac{64e}{\epsilon} \right) + \ln(8/\delta) \right],$$

we have $P \left[I_F(\hat{f}) - I_F(f^) > \epsilon \right] < \delta$.*

If $\{f_1, f_2, \dots, f_{d_V}\}$ is a basis of \mathcal{F} , $f \in \mathcal{F}$ can be written as $f(y) = \sum_{i=1}^{d_V} a_i f_i(y)$ for $a_i \in \Re$. Then consider

$$\hat{f} = \sum_{i=1}^{d_V} \hat{a}_i f_i(y)$$

such that $\hat{a} = (\hat{a}_1, \hat{a}_2, \dots, \hat{a}_{d_V})$ minimizes the cost expressed as

$$I_{emp}(a) = \frac{1}{l} \sum_{k=1}^l \left(X_k - \sum_{i=1}^{d_V} a_i f_i(Y_k) \right)^2,$$

where $a = (a_1, a_2, \dots, a_{d_V})$. Now $I_{emp}(a)$ can be written in the quadratic form $a^T C a + a^T D$, where $C = [c_{ij}]$ is a positive definite symmetric matrix, and D is a vector. This problem can be solved in polynomial-time using quadratic programming methods (Vavasis, 1991).

This method subsumes two very important cases:

- (a) *Potential Functions*: The potential functions of Aizerman *et al.* (Aizerman *et al.*, 1970), where $f_i(y)$ is of the form $\exp((y - \alpha)^2/\beta)$ for suitably chosen constants α and β , constitute an example of the vector space methods. An incremental algorithm was originally proposed for the computation of the coefficient vector a , for which finite sample results have been derived recently (Rao *et al.*, 1996) under certain conditions. The sample size estimate of Theorem 3 is simpler compared to the existing finite sample results. Note that the above sample size is valid only for the method that minimizes $I_{emp}(\cdot)$ and is not valid for the original incremental algorithm of the potential functions.
- (b) *Special Neural Networks*: In two-layer sigmoidal networks of (Kurkova, 1992), the unknown weights are only in the output layer. The specific form of these networks enables us to express each network in the form $\sum_{k=1}^{d_V} a_k \eta_k(y)$ where $\eta_k(\cdot)$'s are universal. These networks have been shown to approximate classes of the continuous functions with arbitrarily specified precision in a manner similar to the general single layer sigmoidal networks as shown in (Cybenko, 1989).

3.4 ASYMPTOTIC CONSISTENCY

In statistics literature, the asymptotic consistency results are more common. The finite sample guarantees in Eq 1.1 often lead to asymptotic results by a direct application of the Borel-Cantelli Lemma (Billingsley, 1986) as will be shown here. The estimate \hat{f} is *consistent* if $I_F(\hat{f}) \rightarrow I_F(f^*)$ with probability one as $l \rightarrow \infty$. In several cases, δ can be rewritten as a function of the sample size, ϵ , and parameters of \mathcal{F} . We consider the general case in Section 3.1. In this case we have (Rao, 1999b)

$$\delta = 48l \left(\frac{4608l}{\epsilon^2} \right)^{d_S \log_2(96en/(d\epsilon))} e^{-\epsilon^2 l/144}$$

where $d_S = P_{\epsilon/4}$ -dim (\mathcal{F}) . Under the finiteness of d_S , the consistency result follows from the Borel-Cantelli if

$$\sum_{l=1}^{\infty} 48l \left(\frac{4608l}{\epsilon^2} \right)^{d \log_2(96en/(d_S \epsilon))} e^{-\epsilon^2 l/144} \leq \infty$$

for every $\epsilon > 0$, which can be easily shown. Similar results can be shown for various conditions used for sigmoid neural networks, and vector space methods.

4. STATISTICAL ESTIMATORS

The fusion rule estimation problem is very similar to the regression estimation problem. In this section we present a polynomial-time (in sample size l) computable estimator which guarantees the criterion in Eq 1.1 under additional smoothness conditions. Our presentation is based on Nadaraya-Watson estimator applied to the sensor fusion problem. As is the case with most available statistical estimators, their original performance guarantees are asymptotic. In general finite sample results had to be derived for these estimators to be applied the sensor fusion problem under criterion in Eq 1.1.

We first present some preliminaries needed for the main result. Let Q denote the unit cube $[0, 1]^N$ and $\mathcal{C}(Q)$ denote the set of all continuous functions defined on Q . The modulus of smoothness of $f \in \mathcal{C}(Q)$ is defined as

$$\omega_\infty(f; r) = \sup_{\|y-z\|_\infty < r, y, z \in Q} |f(y) - f(z)|$$

where $\|y - z\|_\infty = \max_{i=1}^M |y_i - z_i|$. For $m = 0, 1, \dots$, let Q_m denote a family of dyadic cubes (Haar system) such that $Q = \bigcup_{J \in Q_m} J$, $J \cap J' = \emptyset$ for $J \neq J'$, and

the N -dimensional volume of J , denoted by $|J|$, is 2^{-Nm} . Let $1_J(y)$ denote the indicator function of $J \in Q_m$: $1_J(y) = 1$ if $y \in J$, and $1_J(y) = 0$ otherwise. For given m , we define the map P_m on $\mathcal{C}(Q)$ as follows: for $f \in \mathcal{C}(Q)$, we have $P_m(f) = P_m f$ defined by

$$P_m f(y) = \frac{1}{|J|} \int_J f(z) dz$$

for $y \in J$ and $J \in Q_m$ (Ciesielski, 1988). Note that $P_m f : Q \mapsto [0, 1]$ is a discontinuous (in general) function which takes constant values on each $J \in Q_m$. The *Haar kernel* is given by

$$P_m(y, z) = \frac{1}{|J|} \sum_{J \in Q_m} 1_J(y) 1_J(z)$$

for $y, z \in Q$.

Given l -sample, the Nadaraya-Watson estimator based on Haar kernels is defined by

$$\hat{f}_{m,l}(y) = \frac{\sum_{j=1}^l X_j P_m(y, Y_j)}{\sum_{j=1}^l P_m(y, Y_j)} = \frac{\sum_{Y_j \in J} X_j}{\sum_{Y_j \in J} 1_J(Y_j)}$$

for $y \in J$ (Rao, 1983; Engel, 1994). The second expression indicates that $\hat{f}_{m,l}(y)$ is the mean of the function values corresponding to Y_j 's in J that contains y . This property is the key to efficient computation of the estimate (Rao and Protopopescu, 1996).

The Nadaraya-Watson estimator based on more general kernels is well-known in statistics literature (Nadaraya, 1989). This estimator was found to be very effective in a number of applications involving nonlinear regression estimation. The typical results about the performance of this estimator are in terms of asymptotic results, and are not particularly targeted towards fast computation. The above computationally efficient version based on Haar kernels is due to (Engel, 1994), which was subsequently shown to yield finite sample guarantees in (Rao and Protopopescu, 1996). The result of (Rao and Protopopescu, 1996) requires finiteness of capacity of \mathcal{F} in addition to smoothness, and the following theorem specifies the sample size based only on smoothness.

Theorem 4 (Rao, 1997a) *Consider a family of functions $\mathcal{F} \subseteq \mathcal{C}(Q)$ with range $[0, 1]$ such that $\omega_\infty(f; r) \leq kr$ for some $0 < k < \infty$. We assume that: (i) there exists a family of densities $\mathcal{P} \subseteq \mathcal{C}(Q)$; (ii) for each $p \in \mathcal{P}$, $\omega_\infty(p; r) \leq kr$; and (iii) there exists $\mu > 0$ such that for each $p \in \mathcal{P}$, $p(y) > \mu$ for all $y \in [0, 1]^N$. Suppose that the sample size, l , is larger than*

$$\frac{2^{2m+4}}{\epsilon_1^2} \left[\left(\frac{k2^m}{\epsilon_1} \left[\left(\frac{k2^m}{\epsilon_1} - 1 \right)^{N-1} + 1 \right] + m \right) \ln \left(2^{m+1} k / \epsilon_1 \right) \right. \\ \left. + \ln \left(\frac{2^{2m+6}}{(\delta - \lambda) \epsilon_1^4} \right) \right]$$

where $\epsilon_1 = \epsilon(\mu - \epsilon)/4$, $0 < \beta < \frac{N}{2(N+1)}$, $m = \lceil \frac{\log l \beta}{N} \rceil$ and

$$\lambda = b \left(\frac{2}{\epsilon} \right)^{1/N+1-1/2\beta} + b \left(\frac{2}{\epsilon_1} \right)^{1/N+1-1/2\beta}.$$

Then for any $f \in \mathcal{F}$, we have $P \left[|I_F(\hat{f}_{m,l}) - I_F(f^*)| > \epsilon \right] < \delta$.

The value of $\hat{f}_{m,l}(y)$ at a given y is the ratio of local sum of X_i 's to the number of Y_i 's in J that contains y . The range-tree (Preparata and Shamos, 1985) can be constructed to store the cells J that contain at least one Y_i ; with each such cell, we store the number of the Y_i 's that are contained in J and the sum of the corresponding X_i 's. The time complexity of this construction is $O(l(\log l)^{N-1})$ (Preparata and Shamos, 1985). Using the range tree, the values of J containing y can be retrieved in $O((\log l)^N)$ time (Rao and Protopopescu, 1996).

Training Set	Testing Set	Nadaraya-Watson	Nearest Neighbor	Neural Network
100	10	0.000902	0.002430	0.048654
1000	100	0.001955	0.003538	0.049281
10000	1000	0.001948	0.003743	0.050942

(a) $d = 3$

Training Set	Testing Set	Nadaraya-Watson	Nearest Neighbor	Neural Network
100	10	0.004421	0.014400	0.018042
1000	100	0.002944	0.003737	0.021447
10000	1000	0.001949	0.003490	0.023953

(b) $d = 5$

Table 1.1 Fusion of function estimators: mean square error over test set.

The smoothness conditions required in Theorem 4 are not very easy to verify in practice. However, this estimator is found to perform well in a number of applications including those that do not have smoothness properties (see next section). Several other statistical estimators can also be used for fusion rule estimation, but finite sample results must be derived to ensure the condition in Eq 1.1. Such finite sample results are available for adapted nearest-neighbor rules and regressograms (Rao and Protopopescu, 1996) which can also be applied for the fuser estimation problem.

5. APPLICATIONS

We present three concrete applications to illustrate the performance of methods described in the previous sections – the first two are simulation examples and the third one is an experimental system. In addition, the first two examples also provide results obtained with the nearest neighbor rule, which is analyzed elsewhere (Rao, 1994). In the second example, we also consider another estimate, namely, the empirical decision rule described in (Rao and Iyengar, 1996). Pseudo random number generators are used in both the simulation examples.

Example 5.1: *Fusion of Noisy Function Estimators:* (Rao, 1997a) Consider five estimators of a function $g : [0, 1] \mapsto [0, 1]$ such that i th estimator out-

puts a corrupted value $Y^{(i)} = g_i(X)$ of $g(X)$ when presented with input $X \in [0, 1]^d$. The fused estimate $f(g_1(X), \dots, g_5(X))$ must closely approximate $g(X)$. Here g is realized by a feedforward neural network, and, for $i = 1, 2, \dots, 5$, $g_i(X) = g(X)(1/2 + iZ/10)$ where Z is uniformly distributed over $[-1, 1]$. Thus we have $1/2 - i/10 \leq g_i(X)/g(X) \leq 1/2 + i/10$. Table 1.1 corresponds to the mean square error in the estimation of f for $d = 3$ and $d = 5$, respectively, using the Nadaraya-Watson estimator, nearest neighbor rule, and a feedforward neural network with backpropagation learning algorithm. Note the superior performance of the Nadaraya-Watson estimator. \square

Example 5.2: Decision Fusion: (Rao and Iyengar, 1996; Rao, 1999a) We consider 5 sensors such that $Y \in \{H_0, H_1\}^5$ such that $X \in \{H_0, H_1\}$ corresponds to “correct” decision, which is generated with equal probabilities, i. e., $P(X = H_0) = P(X = H_1) = 1/2$. The error of sensor S_i , $i = 1, 2, \dots, 5$, is described as follows: the output $Y^{(i)}$ is correct decision with probability of $1 - i/10$, and is the opposite with probability $i/10$. The task is to combine the outputs of the sensors to predict the correct decision. The percentage error of the individual detectors and the fused system based on the Nadaraya-Watson estimator is presented in Table 1.2. Note that the fuser is consistently better than the best sensor S_1 beyond the sample sizes of the order of 1000. The performance results of the Nadaraya-Watson estimator, empirical decision rule, nearest neighbor rule, and the Bayesian rule based on the analytical formulae are presented in Table 1.3. The Bayesian rule is computed based on the formulae used in the data generation and is provided for comparison only. \square

Example 5.4: Door Detection Using Ultrasonic and Infrared Sensors: Consider the problem of recognizing a door (an opening) wide enough for a mobile robot to move through. The mobile robot (TRC Labmate) is equipped with an

Sample Size	Test set	S_1	S_2	S_3	S_4	S_5	Nadaraya-Watson
100	100	7.0	20.0	33.0	35.0	55.0	12.0
1000	1000	11.3	18.5	29.8	38.7	51.6	10.6
10000	10000	9.5	20.1	30.3	39.8	49.6	8.58
50000	50000	10.0	20.1	29.8	39.9	50.1	8.860

Table 1.2 Performance of Nadaraya-Watson estimator for decision fusion.

Sample Size	Test Size	Bayesian Fuser	Empirical Decision	Nearest Neighbor	Nadaraya-Watson
100	100	91.91	23.00	82.83	88.00
1000	1000	91.99	82.58	90.39	89.40
10000	10000	91.11	90.15	90.81	91.42
50000	50000	91.19	90.99	91.13	91.14

Table 1.3 Comparative Performance.

array of four ultrasonic and four infrared Boolean sensors on each of four sides as shown in Figure 1.1. We address only the problem of detecting a wide enough door when the sensor array of any side is facing it. The ultrasonic sensors return a measurement corresponding to distance to an object within a certain cone as illustrated in Figure 1.1. The infrared sensors return Boolean values based on the light reflected by an object in the line-of-sight of the sensor; white smooth objects are detected due to high reflectivity, while objects with black or rough surface are generally not detected. In practice, both ultrasonic and infrared sensors are unreliable, and it is very difficult to obtain accurate error distributions of these sensors. The ultrasonic sensors are susceptible to multiple reflections and the profiles of the edges of the door. The infrared sensors are susceptible to surface texture and color of the wall and edges of the door. Accurate derivation of probabilistic models for these sensors requires a detailed knowledge of the physics and engineering of the device as well as a priori statistical information. Consequently, a Bayesian solution to this problem is very hard to implement. On the other hand, it is relatively easy to collect experimental data by presenting to the robot doors that are wide enough as well as those that are narrower than the robot. We employ the Nadaraya-Watson estimator to derive a non-linear relationship between the width of the door and the sensor readings. Here the training sample is generated by actually recording the measurements while the sensor system is facing the door. Positive examples are generated if the door is wide enough for the robot, and the sensory system is facing the door. Negative examples are generated when the door is not wide enough or the sensory system is not correctly facing a door (wide enough or not). The robot is manually located in various positions to generate the data. Consider the sensor array of a particular side of the mobile robot. Here $Y^{(1)}, Y^{(2)}, Y^{(3)}, Y^{(4)}$ correspond to the normalized distance measurements from the four ultrasonic sensors, and $Y^{(5)}, Y^{(6)}, Y^{(7)}, Y^{(8)}$ correspond to the Boolean measurements of the infrared sensors. X is 1 if the sensor system is correctly facing a wide enough door, and

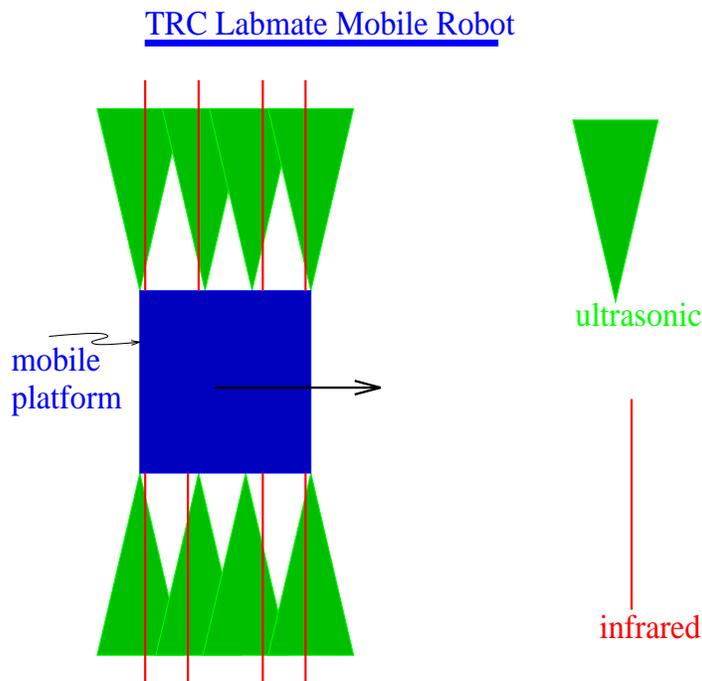


Figure 1.1 Schematic of sensory system (only the side sensor arrays are shown for simplicity).

is 0 otherwise. The training data included 6 positive examples and 12 negative examples. The test data included 3 positive examples and 7 negative examples. The Nadaraya-Watson estimator predicted the correct output in all examples of test data. \square

6. PERFORMANCE OF FUSED SYSTEM

In the empirical risk minimization methods $I_F(\hat{f})$ is shown to be close to $I_F(f^*)$, which depends on \mathcal{F} . In general $I_F(f^*)$ could be very large for particular fuser classes. Note that one cannot simply choose an arbitrary large \mathcal{F} : if so, the performance guarantees of the type in Eq1.1 cannot be guaranteed. If $I_F(f^*) > I(S_i)$, then fusion is not useful, since one is better off just using S_i . In practice, however, such condition cannot be verified if the distributions are not known. In this section, we address the issue of the relative performance of the composite system, composed of the fuser and S_1, S_2, \dots, S_N , and the individual sensors or sensor subsets. We obtain sufficiency conditions under which the composite system can be shown to be at least as good as the best sensor or best subset of sensors.

For simplicity, we consider a system of N sensors such that $X \in [0, 1]$, $Y^{(i)} \in [0, 1]$ and the *expected square error* is given by

$$I_S(S_i) = \int [X - Y^{(i)}]^2 dP_{Y^{(i)}, X}.$$

The *expected square error* of the fuser f is given by

$$I_F(f) = \int [X - f(Y)]^2 dP_{Y, X}$$

respectively, where $Y = (Y^{(1)}, Y^{(2)}, \dots, Y^{(N)})$.

6.1 ISOLATION FUSERS

If the distributions are known, one can derive the best sensor S_{i^*} such that

$$I_S(S_{i^*}) = \min_{i=1}^N I_S(S_i).$$

In the present formulation, the availability of *only* a sample makes the selection (with probability 1) of the best sensor infeasible, even in the special case of the target detection problem (Devroye et al., 1996). In this section, we present a method that circumvents this difficulty by fusing the sensors such that the performance of best sensor is achieved as a minimum. The method is fully sample-based in that no comparative performance of the sensors is needed – in particular, the best sensor may be unknown.

A function class $\mathcal{F} = \{f : [0, 1]^k \mapsto [0, 1]\}$ has the *isolation property* if it contains the functions $f^i(y_1, y_2, \dots, y_k) = y_i$ for all $i = 1, 2, \dots, k$. If \mathcal{F} has the isolation property, we have

$$\begin{aligned} I_F(f^*) &= \min_{f \in \mathcal{F}} \int (X - f(Y))^2 dP_{Y, X} \leq \int (X - f^i(Y))^2 dP_{Y, X} \\ &= \int (X - Y^{(i)})^2 dP_{Y, X} = I_S(S_i), \end{aligned}$$

which implies $I_F(f^*) = \min_{i=1}^N I_S(S_i) - \Delta$ for some $\Delta \in [0, \infty)$. Due to the isolation property, we have $\Delta \geq 0$, which implies that the error of f^* is no higher than $I(S_{i^*})$, but can be significantly smaller. The precise value of Δ depends on \mathcal{F} , but the isolation property guarantees that $I_F(f^*) \leq \min_{i=1}^N I_S(S_i)$ as a minimum.

Let the set S be equipped with a pseudometric ρ . The *covering number* $N_C(\epsilon, \rho, S)$ under metric ρ is defined as the smallest number of closed balls of radius ϵ , and centers in S , whose union covers S . For a set of functions

$\mathcal{G} = \{g : \mathfrak{R}^M \mapsto [0, 1]\}$, we consider two metrics defined as follows: for $g_1, g_2 \in \mathcal{G}$ we have

$$d_P(g_1, g_2) = \int_{z \in \mathfrak{R}^M} |g_1(z) - g_2(z)| dP,$$

for the probability distribution P defined on \mathfrak{R}^M , and

$$d_\infty(g_1, g_2) = \sup_{z \in \mathfrak{R}^M} |g_1(z) - g_2(z)|.$$

This definition is applied to functions defined on $A \subseteq \mathfrak{R}^M$ by extending them to take value 0 on $\mathfrak{R}^M \setminus A$.

Theorem 5 (Rao, 1998a) Consider a fuser class $\mathcal{F} = \{f : [0, 1]^N \mapsto [0, 1]\}$, such that $I_F(f^*) = \min_{f \in \mathcal{F}} I_F(f)$ and $\hat{I}_F(\hat{f}) = \min_{f \in \mathcal{F}} \hat{I}_F(f)$. If \mathcal{F} has the isolation property, we have

$$I_F(f^*) = \min_{i=1}^N I_S(S_i) - \Delta,$$

for $\Delta \in [0, \infty)$, and

$$P_{Y,X}^l \left[I_F(\hat{f}) - \min_{i=1}^N I_S(S_i) + \Delta > \epsilon \right] < \delta$$

given the sample size l of at least

$$\frac{2048}{\epsilon^2} [\ln N_C(\epsilon/64, \mathcal{F}) + \ln(4/\delta)]$$

for cases: (i) $N_C(\epsilon, \mathcal{F}) = N_C(\epsilon, d_\infty, \mathcal{F})$, and (ii) $N_C(\epsilon, \mathcal{F}) = N_C(\epsilon, d_P, \mathcal{F})$ for all distributions P .

If \mathcal{F} has the isolation property, the fuser is guaranteed to perform at least as good as the best sensor in PAC sense. No information other than the iid sample is needed to ensure this result. Since $\Delta \geq 0$, under the sample size of Theorem 5, we trivially have

$$P \left[I_F(\hat{f}) - \min_{i=1}^N I_S(S_i) > \epsilon \right] < \delta.$$

The sample size needed is expressed in terms of d_∞ or distribution-free covers for \mathcal{F} . Note that for smooth fusers such as sigmoid neural networks, we have simple d_∞ cover bounds. The pseudo-dimension and scale-sensitive dimension of \mathcal{F} provide the distribution-free cover bounds needed in Theorem 5 when smoothness conditions may not be satisfied.

The isolation property was first proposed in (Rao et al., 1994b; Rao, 1994) for concept and sensor fusion problems. For linear combinations, i. e.

$$f(y_1, y_2, \dots, y_k) = w_1 y_1 + w_2 y_2 + \dots + w_k y_k,$$

for $w_i \in \mathfrak{R}$, this property is trivially satisfied. For several well-known function classes such as potential functions (Aizerman et al., 1970) and feedforward sigmoid networks (Roychowdhury et al., 1994), this property is not satisfied in general; see (Rao, 2000) for a more detailed discussion on the isolation property and various function classes that have this property.

Consider the special case where S_i 's are classifiers obtained using different methods as in (Mojirsheibani, 1997). For Boolean functions, the isolation property is satisfied if \mathcal{F} contains all Boolean functions on k variables. S_i computed based an iid l -sample is *consistent* if $I_S(S_i) \rightarrow I_S(S^*)$, where S^* is the Bayes classifier. By the isolation property, if one of the classifiers is consistent, the fused classifier system (trained by l -sample independent from n -sample used by the classifiers) can be seen to be consistent. Such result was obtained in (Mojirsheibani, 1997) for linear combinations (for which $N_C(\epsilon, \mathcal{F})$ is finite). Our result does not require the linearity, but pinpoints the essential property, namely the isolation. More generally, linear combinations have been extensively used as fusers in various applications such as combining neural network estimators (Hashem, 1997), regression estimators (Brieman, 1996; Taniguchi and Tresp, 1997), and classifiers (Mojirsheibani, 1997). Since the linear combinations possess the isolation property, Theorem 5 can be viewed as providing some analytical justification for these methods. A more detailed treatment of Boolean problems, namely for classifier problems, can be found in (Rao, 1998b).

6.2 PROJECTIVE FUSERS

A *projective fuser* (Rao, 1999c), f_P , corresponding to a *partition* $P = \{\pi_1, \pi_2, \dots, \pi_k\}$, $k \leq N$, of input space \mathfrak{R}^d ($\pi_i \subseteq [0, 1]^d$, $\bigcup_{i=1}^k \pi_i = \mathfrak{R}^d$, and $\pi_i \cap \pi_j = \emptyset$ for $i \neq j$), assigns each block π_i to a sensor S_j such that

$$f_P(Y) = Y^{(j)}$$

for all $X \in \pi_i$, i.e. the fuser simply transfers the output of the sensor S_j for every point in π_i . An *optimal projective fuser*, denoted by f_{P^*} , minimizes $I(\cdot)$ over all projective fusers corresponding to all partitions of \mathfrak{R}^d and assignments of blocks to sensors S_1, S_2, \dots, S_N .

We define the *error regression* of the sensor S_i and fuser f_F as

$$\mathcal{E}(X, S_i) = \int C(X, Y^{(i)}) dP_{Y|X}$$

and

$$\mathcal{E}(X, f_P) = \int C(X, f_P(Y)) dP_{Y|X},$$

respectively. The *projective fuser* based on the lower envelope of error regressions of sensors is defined by

$$f_{LE}(Y) = Y^{(i_{LE}(X))}$$

where

$$i_{LE}(X) = \arg \min_{i=1,2,\dots,N} \mathcal{E}(X, S_i).$$

We have $\mathcal{E}(X, f_{LE}) = \min_{i=1,\dots,N} \mathcal{E}(X, S_i)$, or equivalently the error regression of f_{LE} is the lower envelope with respect to X of the set of error regressions of sensors given by $\{\mathcal{E}(X, S_1), \dots, \mathcal{E}(X, S_N)\}$.

Example 6.1: (Rao, 1999c) Consider that X is uniformly distributed over $[0, 1]$, which is measured by two sensors S_1 and S_2 . Let $C(X, Y^{(i)}) = (X - Y^{(i)})^2$. Consider $Y^{(1)} = X + |X - 1/2| + U$ and $Y^{(2)} = X + 1/[4(1 + |X - 1/2|)] + U$, where U is an independent random variable with zero mean. Thus, for both sensors the measurement error at any X is represented by U . Note that

$$E[Y^{(1)} - X] = |X - 1/2|$$

$$E[Y^{(2)} - X] = 1/[4(1 + |X - 1/2|)].$$

Thus, S_1 achieves a low error in the middle of the range $[0, 1]$, and S_2 achieves a low error towards the end point of the range $[0, 1]$. The error regressions of the sensors are given by

$$\mathcal{E}(X, S_1) = (X - 1/2)^2 + E[U^2]$$

$$\mathcal{E}(X, S_2) = 1/[16(1 + |X - 1/2|)^2] + E[U^2].$$

We have

$$I(S_1) = 0.0833 + E[U^2] \quad \text{and} \quad I(S_2) = 0.125 + E[U^2],$$

which indicates that S_1 is the better of the two sensors. Now consider the projective fuser f_{LE} specified as follows, which corresponds to the lower envelope of $\mathcal{E}(X, S_1)$ and $\mathcal{E}(X, S_2)$.

range for X	sensor to be projected
$[0, 0.134]$	S_2
$[0.134, 0.866]$	S_1
$[0.866, 1]$	S_2

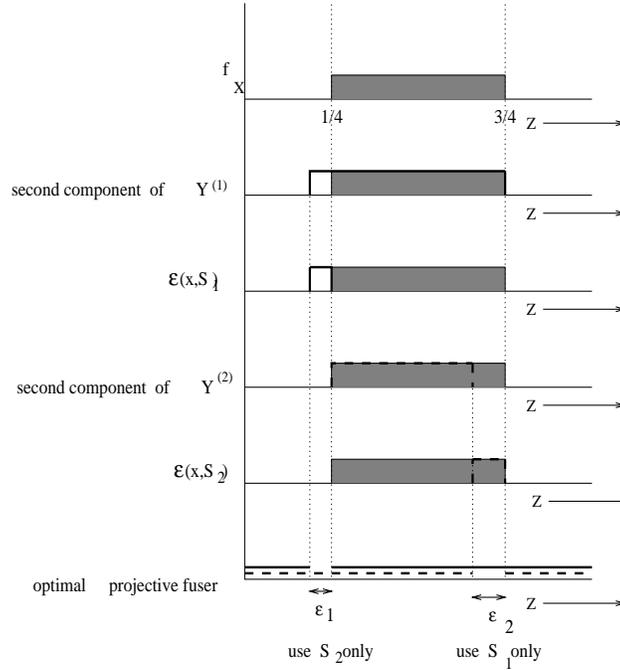


Figure 1.2 Illustration for example

Then, we have $I(f_{LE}) = 0.0828 + E[U^2]$, which is lower than that of the best sensor. \square

Example 6.2: (Rao, 1999c) We consider a classification example such that $X \in [0, 1] \times \{0, 1\}$ is specified by a function $f_X = 1_{[1/4, 3/4]}$, where $1_A(z)$ is the *indicator function* (which has a value 1 if and only if $z \in A$ and has value 0 otherwise). The value of X is generated as follows: a random variable Z is generated uniformly in the interval $[0, 1]$ as the first component, and then $f_X(Z)$ forms the second component, i.e. $X = (Z, f_X(Z))$. In the context of the detection problem, the second component of X corresponds to the presence ($f_X(Z) = 1$) or absence ($f_X(Z) = 0$) of a target, which is represented by a feature Z taking a value in the interval $[1/4, 3/4]$. Each sensor consists of a device to measure the first component of X and an algorithm to compute the second component. We consider that S_1 and S_2 have ideal devices that measure Z without an error, but make errors in utilizing the measured features. Consider that $Y^{(1)} = (Z, 1_{[1/4-\epsilon_1, 3/4]}(Z))$ and $Y^{(2)} = (Z, 1_{[1/4, 3/4-\epsilon_2]}(Z))$ for some $0 < \epsilon_1, \epsilon_2 < 1/4$ (see Figure 1.2). In other words, there is no measurement noise in the sensors but just a systematic error due to how the feature value is

utilized; addition of independent measurement noise as in Example 6.1 does not change the basic conclusions of the example. Now consider the quadratic cost function $C(X, Y^{(i)}) = (X - Y^{(i)})^T (X - Y^{(i)})$. The error regressions are given by $\mathcal{E}(X, S_1) = 1_{[1/4 - \epsilon_1, 1/4]}(Z)$ and $\mathcal{E}(X, S_2) = 1_{[3/4 - \epsilon_2, 3/4]}(Z)$, which corresponds to disjoint intervals of Z as shown in Figure 1.3. The lower envelope of the two regressions is the zero function hence $I(f_{LE}) = 0$, where as both $I(S_1)$ and $I(S_2)$ are positive. The profile of f_{LE} is shown at the bottom of Figure 1.2, wherein S_1 and S_2 are projected based on the first component of X in the intervals $[3/4 - \epsilon_2, 3/4]$ and $[1/4 - \epsilon_1, 1/4]$, respectively, and in other regions either sensor can be projected. \square

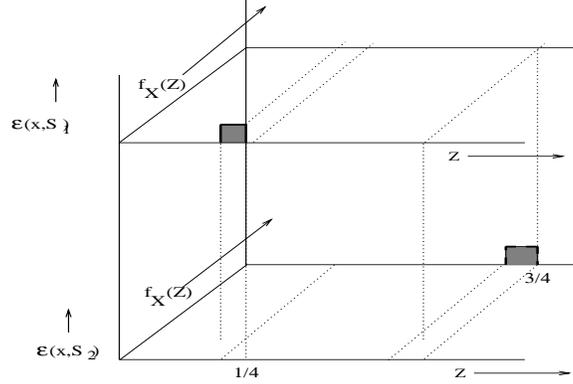


Figure 1.3 Illustration of error regressions.

The projective fuser based on error regressions is optimal as in the following theorem.

Theorem 6 (Rao, 1999c) *The projective fuser based on the lower envelope of error regressions is optimal among all projective fusers.*

A special case of this theorem for function estimation can be found in (Rao, 1999d), and for classifiers can be found in (Rao, 1998b). A sample-based version of projective fuser is discussed in (Rao, 1999c).

We close this section by emphasizing that f_{LE} may not be optimal in a larger class of fusers where some function of the sensor output (as opposed to just the output) can be projected.

Example 6.3: (Rao, 1999c) In Example 2.2, consider $f_X = 1_{[1/4, 3/4]}$,

$$Y^{(1)}(X) = (Z, 1_{[1/4 - \epsilon_1, 3/4 - \epsilon_1]}(Z))$$

$$Y^{(2)}(X) = (Z, 1_{[1/4, 3/4 - \epsilon_2]}(Z))$$

for some $0 < \epsilon_1, \epsilon_2 < 1/8$, and $\epsilon_1 < \epsilon_2$. Thus, we have $\mathcal{E}(X, S_1) = 1_{[1/4-\epsilon_1, 1/4]}(Z)$ and $\mathcal{E}(X, S_2) = 1_{[3/4-\epsilon_2, 3/4]}(Z)$, whose lower envelope is not the zero function. Thus, we have $\mathcal{E}(X, f_{LE}) = 1_{[3/4-\epsilon_2, 3/4-\epsilon_1]}(Z)$ and $I(f_{LE}) = \int_{[3/4-\epsilon_2, 3/4-\epsilon_1]} dP_Z$. By changing the assignment $Y^{(1)}$ of f_{LE} to $1 - Y^{(1)}$ for $Z \in [3/4 - \epsilon_2, 3/4 - \epsilon_1]$, one can easily achieve zero error. \square

7. METAFUSERS

In this section, we first compare projective fusers to linear fusers and show that they have complementary performances. Thus it is natural to combine the fusers – much the same ways as sensors – to exploit the relative merits of individual fusers. Such approach leads to the idea of *metafusers*. There are many possible ways of designing metafusers, and here we describe how isolation property can be utilized to develop systems that exploit the complementary performances sensors as well as fusers.

The output of linear fuser corresponding to input X and sensor output $Y = (Y^{(1)}, \dots, Y^{(N)})$ is defined as

$$f_L(Y) = \sum_{i=1}^N \alpha_i Y^{(i)}$$

where α_i is a $d \times d$ matrix. For simplicity, we consider the case $d = 1$ such that $(\alpha_1, \dots, \alpha_N) \in \mathfrak{R}^N$. An *optimal linear combination fuser*, denoted by f_{L^*} , minimizes $I(\cdot)$ over all linear combinations. In terms of relative performance, f_{LE} is better than f_{L^*} if the individual sensors perform better in certain localized regions of \mathfrak{R}^d . On the other hand, if the sensors are equally distributed around certain values in global sense, f_L performs better as illustrated follows.

Example 6.4: (Rao, 1999c) In the Example 6.2, for $f_L = \alpha_1 Y^{(1)} + \alpha_2 Y^{(N)}$, we have

$$\begin{aligned} I(f_L) &= \alpha_1^2 \int_{[1/4-\epsilon_1, 1/4]} dP_Z \\ &+ (1 - \alpha_1 - \alpha_2)^2 \int_{[1/4, 3/4-\epsilon_2]} dP_Z \\ &+ (1 - \alpha_1)^2 \int_{[3/4-\epsilon_2, 3/4]} dP_Z \end{aligned}$$

which is non-zero no matter what the coefficient are. The error regressions of S_1 and S_2 take non-zero values in the intervals $[1/4 - \epsilon_1, 1/4]$ and $[3/4 - \epsilon_2, 3/4]$

of Z , respectively. Since, these intervals are disjoint, there is no possibility of the error of one sensor being canceled by a scalar multiplier of the other. \square

In general, the argument of Example 6.4 is true: if the error regressions of the sensors take non-zero values on disjoint intervals, then any linear fuser will have non-zero error. On the other hand, the disjointness yields $\mathcal{E}(X, f_{LE}) = 0$, for all X , and hence $I(f_{LE}) = 0$. We now present an example where a linear fuser outperforms f_{LE} .

Example 6.5: Consider that in Example 6.2, $f_X = 1$ for $Z \in [0, 1]$,

$$Y^{(1)}(X) = (Z, \epsilon Z + 1 - \epsilon)$$

$$Y^{(2)}(X) = (Z, -\epsilon Z + 1 + \epsilon),$$

for $0 < \epsilon < 1$. The optimal linear fuser is given by $f_{L^*}(Y) = 1/2(Y^{(1)} + Y^{(2)}) = 1$, and $I(f_{L^*}) = 0$. At every $X \in [0, 1]$, we have

$$\mathcal{E}(X, S_1) = \mathcal{E}(X, S_2) = \epsilon^2(1 - Z)^2 = \mathcal{E}(X, f_{LE}).$$

Thus, $I(f_{LE}) = \epsilon^2 \int_{[0,1]} (1 - Z)^2 dP_Z > 0$, whereas $I(f_{L^*}) = 0$. \square

In summary, the performance of the optimal linear and projective fusers are complementary in general. We now combine linear and projective fusers to realize various meta-fusers that are guaranteed to be at least as good as the best sensor as well as best sensor. By including the optimal linear combination as S_{N+1} , we can guarantee that $I(f_{LE}) \leq I(f_{L^*})$ by the isolation property of projective fusers (Rao, 1999c). Since linear combinations also satisfy the isolation property, we in turn have $I(f_{L^*}) \leq \min_{i=1, \dots, N} I(S_i)$.

The roles of f_{L^*} and f_{LE} can be switched – by including f_{LE} as one of the components of f_{L^*} – to show that

$$I(f_{L^*}) \leq I(f_{LE}) \leq \min_{i=1, \dots, N} I(S_i).$$

One can design a meta-fuser by utilizing the available sensors which are combined using a number of fusers including a fuser based on isolation property (for example, a linear combination). Consider that we employ a meta-fuser based on a linear combination of the fusers. Then the fused system is guaranteed to be at least as good as the best of the fusers as well as the best sensor. If at a latter point, a new sensor or a fuser is developed, it can be easily integrated into the system by retraining the fuser and/or meta-fuser as needed. As a result, we have a system guaranteed (in PAC sense) to perform at least as good as the best available sensor and fuser at all times. Also, the computational problem of updating the fuser and/or meta-fuser is a simple least squares estimation that can be solved using a number of available methods.

8. FUSERS FOR PHYSICAL SYSTEMS

In this section, we consider a multisensor fusion problem that does not exactly fit into the classical formulation of the fusion problem. Consider a physical system described by a set of parameters, such that each parameter is either measured by a number of sensors or estimated by a set of computer programs using sensor measurements. As a result, the resultant parameter values could be widely varying. In comparison with the traditional fusion problems, there is no training set that provides the actual parameter values. Furthermore, since every parameter is measured or estimated, there are no parameters whose actual values are known. In this section, we describe a fuser based on the least violation of the physical laws that relate the parameters. Under very general conditions on the physical law, we derive distribution-free performance bounds based on finite samples. We illustrate the effectiveness of this method for a practical problem of fusing well-log data in methane hydrate exploration.

8.1 PHYSICAL LAWS

We consider a physical system specified by the parameters

$$P(z) = (p_1(z), p_2(z), \dots, p_n(z))$$

with $p_i(z) \in \mathfrak{R}$, where z is one-dimensional variable such as time or position. Each parameter p_i is measured by a_i instruments and estimated by b_i estimators ($a_i \geq 0$, $b_i \geq 0$, and $a_i + b_i \geq 1$). The measurements corresponding to $p_i(z)$ are denoted by

$$m_i(z) = \{m_{i,1}(z), m_{i,2}(z), \dots, m_{i,a_i}(z)\}$$

and the corresponding estimators are denoted by

$$e_i(z) = \{e_{i,1}(z), e_{i,2}(z), \dots, e_{i,b_i}(z)\}.$$

The measurements are noisy in that repeated measurements by a sensor of $p_i(z) = x$ for a fixed value are distributed independently according to the distribution $P_{m_{i,j}|x}$, which is denoted by $P_{m_{i,j}|p_i(z)}$. Thus, $m_{i,j}$ is a random variable. The estimator $e_{i,j}$ is a (deterministic) function of the measurements, and hence is also a random variable. The joint distribution of the measurements is denoted by $P_{m_1, m_2, \dots, m_n | p_1, p_2, \dots, p_n}$. Note that there are $a_i + b_i$ competing values for each parameter, and in general we do not know which one is more accurate.

There is a physical law

$$L[p_1(z), p_2(z), \dots, p_n(z)] = 0$$

which relates the actual parameters corresponding to z . We assume that $L[\cdot]$ satisfies the reasonable *monotonicity* condition: for any y_1, y_2 , $|y_1| \leq |y_2|$, we

have

$$|L[p_1(z), \dots, p_i(z) + y_1, \dots, p_n(z)]| \leq |L[p_1(z), \dots, p_i(z) + y_2, \dots, p_n(z)]|.$$

This condition means that accurate parameter estimators yield no lesser “magnitude” of violation of the law compared to less accurate estimators.

Let us choose a single estimator or measurement \hat{p}_i for the parameter. The closeness of $L[\hat{p}_1(z), \hat{p}_2(z), \dots, \hat{p}_n(z)]$ to 0 determines how closely the law is satisfied. Let a *basic set*, denoted by S , be a set of measurements and estimators such that for each parameter we choose precisely one measurement or estimator (but not both). The total error due to S is given by

$$\hat{E}(S) = \sum_z L[\hat{p}_1(z), \hat{p}_2(z), \dots, \hat{p}_n(z)].$$

Note that in all there are $\prod_{i=1}^n (a_i + b_i)$ possible basic sets, and \hat{S} be the one with least error such that $\hat{E}(\hat{S}) = \min_S \hat{E}(S)$. The expected error of S is denoted by

$$E(S) = \sum_z \int L[\hat{p}_1(z), \hat{p}_2(z), \dots, \hat{p}_n(z)] P_{m_1, \dots, m_n | p_1, \dots, p_n},$$

and let S^* be the one with the least expected error such that $E(S^*) = \min_S E(S)$.

Note that S^* minimizes the expected error but \hat{S} in general does not. More detailed discussion of the physical laws can be found in (Rao et al., 2000a).

Example 8.1: (Rao et al., 2000a) We consider a simple illustrative example of known mass m subjected to a constant force f in a friction-free environment. In this case the physical law is $f = ma$, which can be rewritten as $L[f, a] = (f - ma)^2 = 0$, where a is acceleration and f is force. Let $p_1(z) = f$ and $p_2(z) = a$. We are given a sensor that measures force and two sensors that measure acceleration. The force measurements have simple bias error such that $m_{1,1}(z) = f + \epsilon$, for some deterministic ϵ . The acceleration measurements are given by $m_{2,1} = a + \delta$, $m_{2,2} = 0.7a$, where δ is a small normally distributed error. Then, we have

$$L[m_{1,1}, m_{2,1}] = (\epsilon - m\delta)^2$$

$$L[m_{1,1}, m_{2,2}] = (\epsilon + 0.3ma)^2$$

Consider that $a > 0$, and $\epsilon \geq m\delta$. If $|\delta| \leq |0.3a|$, we have

$$L[m_{1,1}, m_{2,1}] \leq L[m_{1,1}, m_{2,2}],$$

i.e. for large values of a , the better choice is $m_{2,1}$, otherwise $m_{2,2}$ is a better choice. \square

8.2 FUSER COMPUTATION

The *fusion function* $f_i \in \mathcal{F}_i$ for parameter p_i combines the measurements and estimators such that $f_i(m_i(z), e_i(z))$ is an estimate of $p_i(z)$. Let $f = (f_1, \dots, f_n)$ denote the *fuser* for all parameters. The expected error due to the fused estimate is

$$E(f) = \sum_z \int L[f_1(m_1(z), e_1(z)), \dots, f_n(m_n(z), e_n(z))] dP_{m_1, \dots, m_n | p_1, \dots, p_n},$$

and let $f^* \in \mathcal{F}_1 \times \dots \times \mathcal{F}_n$ be the one with the least expected error. As before $E(f)$ cannot be computed if the error distributions are not known, and hence f^* is not computable. In stead, we compute \hat{f} that minimizes the empirical cost given by

$$\hat{E}(f) = \sum_z L[f_1(m_1(z), e_1(z)), \dots, f_n(m_n(z), e_n(z))],$$

based on a set of iid measurements (also called the sample)

$$\{(m_1(z), e_1(z)), \dots, (m_n(z), e_n(z))\}_{z=1, \dots, s}.$$

We now describe methods that ensure $E(f^*) \leq E(S^*)$, and more importantly based on a computable \hat{f} that

$$E(\hat{f}) < E(S^*),$$

with a specified probability based entirely on the measurements and without any knowledge of the underlying distributions.

If each \mathcal{F}_i satisfies the isolation property, then the following conditions are directly satisfied.

$$E(f^*) \leq E(S^*) \quad \text{and} \quad \hat{E}(\hat{f}) \leq \hat{E}(\hat{S}).$$

The first condition is useful only if f^* can be computed, which in turn requires the knowledge of the distributions. If the distributions are not known, then \hat{f} can be used as an approximation. We subsequently show finite samples guarantees for such fuser.

8.3 SMOOTH PHYSICAL LAWS

For any function $g : [-A, A]^d \mapsto \mathfrak{R}$, let

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

A function $g(y) : [-A, A]^d \mapsto \mathfrak{R}^a$ is *Lipschitz* with constant k_g if for all $y_1, y_2 \in [-A, A]^d$, we have

$$\|g(y_1) - g(y_2)\|_\infty \leq k_g \|y_1 - y_2\|_\infty.$$

For example, the sigmoid neural networks are Lipschitz with constant specified by the parameters of the network (Rao, 1999e).

Theorem 7 (Rao et al., 2000a) *Consider that the physical law is Lipschitz with constant k_L , and parameters, estimators and measurements are bounded such that $p \in [-C, C]^n$, $m_i \in [-A, A]^{a_i}$, and $e_i \in [-B, B]^{b_i}$. Let each fuser class \mathcal{F}_i be Lipschitz with constant k_{f_i} . Let $d = \sum_{i=1}^n (a_i + b_i)$ and $k = k_L \max(k_{f_1}, k_{f_2}, \dots, k_{f_n})$. Then given a sample size*

$$s = \frac{512k(A+B)}{\epsilon^2} \left[d \ln \left(\frac{32k(A+B)}{\epsilon} \right) + \ln(8/\delta) \right],$$

we have

$$P \left[E(\hat{f}) - E(f^*) > \epsilon \right] \leq \delta,$$

irrespective of the sensor distributions. Furthermore, $E(\hat{f}) \rightarrow E(f^*)$, as $s \rightarrow \infty$.

This theorem provides a distribution-free finite sample result: given a sufficiently large sample size, with a probability $1 - \delta$, 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). Results similar to the asymptotic result shown in the above theorem are more common in the statistics literature (Rao, 1983). The finite sample result, however, is stronger in that it implies the asymptotic result, and also establishes that the method is justified even for small sample sizes.

The smoothness conditions required in this theorem are quite reasonable. The Lipschitz condition is satisfied for a number of physical laws, although not always guaranteed. Similar condition was used in converting the decision fusion rules designed for known distributions to sample-based ones in (Rao, 1996). The isolation and Lipschitz properties required of the fusers are satisfied in a number of cases such as linear combinations with bounded coefficients and piecewise linear feedforward networks (Rao, 1998a; Rao, 2000).

Example 8.2: (Rao et al., 2000a) Consider the scenario in Example 8.1, and

$$f_2(m_{2,1}, m_{2,2}) = w_1 m_{2,1} + w_2 m_{2,2}.$$

For the choice $w_1 = 0.3$ and $w_2 = 1$, we have $f_2(m_{2,1}, m_{2,2}) = a + 0.3\delta$, and

$$L[m_{1,1}, f_2(m_{2,1}, m_{2,2})] = (\epsilon - 0.3m\delta)^2,$$

which is always smaller than $L[m_{1,1}, m_{2,1}]$, and is smaller than $L[m_{1,1}, m_{2,2}]$ for small $\delta \leq a$ which is true with probability one since δ is a zero-mean random variable. \square

8.4 NON-SMOOTH PHYSICAL LAWS

In general physical laws may not be Lipschitz, especially if they involve discrete components or discontinuities. For example, consider the simple case of H_2O heated in a container, where p_1 denotes the temperature and $p_2 \in \{0, 1\}$ is the state, i. e. $p_2 = 0$ denotes liquid and $p_2 = 1$ denotes steam. Let T_0 denote the boiling temperature under this condition. Then, one of the physical laws is: $p_2 = 0$ if $p_1 < T_0$ and $p_2 = 1$ otherwise. This law can be represented as

$$L[p_1, p_2] = p_2 1_{\{p_1 < T_0\}} + (p_2 - 1) 1_{\{p_1 \geq T_0\}} = 0,$$

where the indicator function 1_C is 1 if condition C is true and is 0 otherwise. Here, $L[\cdot]$ is not Lipschitz. The class of functions with bounded variation (Apostol, 1974), allow for discontinuities and discrete values, and include Lipschitz functions as a subclass. We now describe a generalization of the results of last section, which enable the utilization of non-smooth physical laws.

Consider a function one-dimensional function $h : [-A, A] \mapsto \mathfrak{R}$. For $A < \infty$, a set of points $P = \{x_0, x_1, \dots, x_n\}$ such that $-A = x_0 < x_1 < \dots < x_n = A$ is called a *partition* of $[-A, A]$. The collection of all possible partitions of $[-A, A]$ is denoted by $\mathcal{P}[-A, A]$. A function $g : [-A, A] \mapsto \mathfrak{R}$ is of *bounded variation*, if there exists M such that for any partition $P = \{x_0, x_1, \dots, x_n\}$, we have $\sum(P) = \sum_{k=1}^n |f(x_k) - f(x_{k-1})| \leq M$. A multivariate function $g : [-A, A]^d \mapsto \mathfrak{R}$ is of bounded variation if it is so in each of its input variable for every value of the other input variables.

The following are facts about the functions of bounded variation: (i) not all continuous functions are of bounded variation, e.g. $g(x) = x \cos(\pi/(2x))$ for $x \neq 0$ and $g(0) = 0$; (ii) differentiable functions on compact domains are of bounded variation; and (iii) absolutely continuous functions, which include Lipschitz functions, are of bounded variation.

For the case of non-smooth physical laws, we utilize the fuser classes with finite pseudo-dimension (Anthony and Bartlett, 1999), which is described as follows. Let G be a set of functions mapping from a domain X to \mathfrak{R} and suppose that $S = \{x_1, x_2, \dots, x_m\} \subseteq X$. Then S is *pseudo-shattered* by F if there are real numbers r_1, r_2, \dots, r_m such that for each $b \in \{0, 1\}^m$ there is a function g_0 in \mathcal{G} with $\text{sgn}(f_b(x_i) - r_i) = b_i$ for $1 \leq i \leq m$. Then \mathcal{G} has the *pseudo-dimension* d if d is the maximum cardinality of a subset S of X that is pseudo-shattered by \mathcal{G} . If no such maximum exists, we say that \mathcal{G} has infinite pseudo-dimension. The pseudo-dimension of \mathcal{G} is denoted $\text{Pdim}(\mathcal{G})$. Pseudo-dimensions are known for several classes such as sigmoid neural networks, vector spaces, and linear combinations (Anthony and Bartlett, 1999).

Theorem 8 (Rao et al., 2000b) *Consider that the physical law is of bounded variation such that $|L(p)| \leq M_L$ for all p . Let parameters, estimators and mea-*

measurements are bounded. Let each fuser class \mathcal{F}_i have finite pseudo-dimension d_i , and each fuser function g be bounded such that $|g(\cdot)| \leq M$ for all parameters. Let $d = \sum_{i=1}^n d_i$. Then given a sample of size

$$s = \frac{256M_L^2}{\epsilon^2} \left[4d \ln \left(\frac{128eM}{\epsilon} \right) + (n+1) \ln(4/\delta) \right],$$

we have

$$\mathbf{P} \left[E(\hat{f}) - E(f^*) > \epsilon \right] \leq \delta,$$

irrespective of the sensor distributions. Furthermore, $E(\hat{f}) \rightarrow E(f^*)$, as $s \rightarrow \infty$.

The following corollary is a weaker version of Theorem 8 since $E(f^*) \leq E(S^*) \leq E(\hat{S})$.

Corollary 1 *Let \mathcal{F}_i satisfy the isolation property for all $i = 1, 2, \dots, n$. Under the same conditions as Theorem 8, we have following conditions satisfied.*

$$\mathbf{P} \left[E(\hat{f}) - E(S^*) > \epsilon \right] \leq \delta \quad \text{and} \quad \mathbf{P} \left[E(\hat{f}) - E(\hat{S}) > \epsilon \right] \leq \delta.$$

Informally speaking, this corollary shows that the error of the computed fuser \hat{f} is not likely to be much higher than that of the best basic set, and could be much smaller. Theorem 8 offers a stronger result: \hat{f} will be closer to f^* which can have much smaller error than S^* .

8.5 METHANE HYDRATES WELL LOGS

We now briefly describe a practical application of the proposed fusion method based on physical laws. Gas hydrates are crystalline substances composed of water and gas, in which gas molecules are contained in cage-like lattices formed by solid water. A challenge is to predict the presence of hydrates using measurements collected at wells located in certain locations such as off the US coast in mid-Atlantic and Mackenzie Delta in Northwest Canada. A number of measurements are collected at each well using a suite of sensors. These measurements include density, neutron porosity, acoustic transit-time, and electric resistivity, collected at various depths in the well (Dallimore et al., 1999). We consider only the estimation of the *porosity* at various depths. Our data consists of 3045 sets of measurements each collected at different depths in a single well. There are a variety of methods to estimate porosity based on different principles and utilizing different measurements. We employed six known methods for estimating the porosity based on neutron measurements ($\hat{\phi}_1$), density measurements ($\hat{\phi}_2$), fluid velocity equation ($\hat{\phi}_3$), acoustic travel

time based on S-wave ($\hat{\phi}_4$), time-average equation based on P-wave ($\hat{\phi}_5$), and Wood's equation ($\hat{\phi}_6$).

We consider one of the well-established physical laws relates the parameters of porosity (ϕ), density (ρ), and hydrate concentration (ψ), as follows

$$L[\phi, \psi, \rho] = (\phi[\rho_m - (1 - \psi)\rho_w + \psi\rho_h] - \rho + \rho_m)^2 = 0,$$

where ρ_m , ρ_w , and ρ_h are known constants. In this equation, we use the only one measurement for density $\hat{\rho}$ and a single estimator $\hat{\psi}$ for the hydrate concentration using the Archie's equation. We consider a fuser based on the linear combination of the estimators

$$\hat{\phi}_F = w_7 + \sum_{i=1}^6 w_i \hat{\phi}_i,$$

where $(w_1, \dots, w_7) \in \mathfrak{R}^7$ is the weight vector that minimizes the error based on measurements. The error achieved by $\hat{\phi}_F$ is about 20 times better than that of the best estimator $\hat{\phi}_4$ (details can be found in (Rao et al., 2000a)). Note that $L[\cdot]$ and the fusers employed here satisfy the conditions of Corollary 1. Incidentally, they also satisfy the smoothness conditions of (Rao et al., 2000a).

9. KNOWN DISTRIBUTIONS

It is not necessary that all sensor distributions be unknown to apply the main results of this paper. Consider that the joint conditional distribution $P_{Y^{(1)}, \dots, Y^{(M)} | Y^{(M+1)}, \dots, Y^{(N)}}$ of the sensors S_1, \dots, S_M , for $M < N$, is known. Then we can rewrite

$$I_F(f) = \int \Phi(X, f(Y)) dP_{Y^{(M+1)}, \dots, Y^{(N)}, X}$$

where $\Phi(\cdot)$ is suitably derived from the original cost function $C(\cdot)$ and the known part of the conditional distribution. Then various theorems can be applied to this new cost function with only a minor modification. Since the number of variables with unknown distribution is reduced now, the statistical estimation process is easier. It is important note that it is not sufficient to know the individual distributions of the sensors, but the *joint* conditional distributions are required. In the special case of statistical independence of sensors the joint distribution is just the product, which makes the transformation easier. In general for sensor fusion problems, however, the interdependence between the sensors is a main feature to be exploited to overcome the limitations of single sensors.

10. CONCLUSIONS

In a multiple sensor system, we considered that for each sensor the outputs are related to the actual feature values according to a certain probability distribution.

We presented an overview of informational and computational aspects of a fuser that combines the sensor outputs to more accurately predict the feature, when the sensor distributions are unknown but iid measurements are given. Our performance criterion is the probabilistic guarantees in terms of distribution-free sample bounds based entirely on a finite sample. We first discussed a number of methods based on the empirical risk minimization approach, which yield a fuser that is guaranteed, with a high probability, to be close to an optimal fuser. Note that the optimal fuser is computable only under a complete knowledge of sensor distributions. Then we described the isolation fusers that are guaranteed to perform at least as good as the best sensor. We then discussed the projective fusers that are guaranteed to perform at least as good as the best subset of sensors. We briefly discussed the notion of meta-fusers that can combine fusers of different types. We considered physical systems wherein the training data consisting of actual physical values is not available. For this case, we discussed methods that utilize the physical laws to obtain a suitable fuser.

The overall focus of this paper is very limited: we only considered sample-based fuser methods that provide finite sample performance guarantees. Even then, there are a number of important issues for the fuser rule computation that have not been covered in this paper. We did not discuss stochastic algorithms for fuser computation (Rao, 1994; Rao et al., 1994a). Another important aspect is the utilization of fusers that have been designed for known distribution cases for the sample-based case. In many important cases, the fuser formulas expressed in terms of probabilities can be converted into sample-based ones by utilizing suitable estimators (Rao, 1996). For the most part, we only considered stationary systems, and it would be of future interest to study sample-based fusers for time-varying systems.

Acknowledgments

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-AC05-00OR22725 with UT-Battelle, LLC, Office of Naval Research under order No. N00014-96-F-0415, and Ballistic Missile Defense Organization under MIPR No. 0100568954.

References

- Abidi, M. A. and Gonzalez, R. C., editors (1992). *Data Fusion in Robotics and Machine Intelligence*. Academic Press, New York.
- Aizerman, M. A., Braverman, E. M., and Rozonoer, L. I. (1970). *Extrapolative problems in automatic control and method of potential functions*, volume 87 of *American Mathematical Society Translations*, pages 281–303.
- Alon, N., Ben-David, S., Cesa-Bianchi, N., and Hausler, D. (1993). Scale-sensitive dimensions, uniform convergence, and learnability. In *Proc. of 1993 IEEE Symp. on Foundations of Computer Science*.

- Anthony, M. (1994). Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. NeuroCOLT Technical Report Series NC-TR-94-3, Royal Holloway, University of London.
- Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.
- Apostol, T. M. (1974). *Mathematical Analysis*. Addison-Wesley Pub. Co.
- Bar-Shalom, Y. and Li, X. R. (1995). *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS Publishing.
- Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons, New York, second edition.
- Blum, A. L. and Rivest, R. L. (1992). Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127.
- Brady, J. M. (1988). Foreword. *Int. J. Robotics Research*, 7(5):2–4. Special Issue on Sensor Data Fusion.
- Brieman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.
- Chow, C. K. (1965). Statistical independence and threshold functions. *IEEE Trans. Electronic Computers*, EC-16:66–68.
- Ciesielski, Z. (1988). Haar system and nonparametric density estimation in several variables. *Probability and Mathematical Statistics*, 9:1–11.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Controls, Signals, and Systems*, 2:303–314.
- Dallimore, S. R., Uchida, T., and Collett, T. S., editors (1999). *Scientific Results from JAPEX/JNOC/GSC Mallik 2L-38 Gas Hydrate Research Well, Mackenzie Delta: Geological Survey of Canada Bulletin 544*. Geological Survey of Canada, Bulletin 544.
- Dasarathy, B. V. (1994). *Decision Fusion*. IEEE Computer Society Press, Los Alamitos, California.
- de Laplace, P. S. (1818). Deuxième supplément à la théorie analytique des probabilités. Reprinted (1847) in *Oeuvres Complètes de Laplace*, vol. 7 (Paris, Gauthier-Villars) 531–580.
- Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.
- Dudley, R. (1987). Universal Donsker classes and metric entropy. *Annals of Probability*, 15:1306–1326.
- Engel, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 49:242–254.
- Granger, C. W. J. (1989). Combining forecasts — twenty years later. *J. of Forecasting*, 8:167–173.
- Grofman, B. and Owen, G., editors (1986). *Information Pooling and Group Decision Making*. Jai Press Inc., Greenwich, Connecticut.

- Hashem, S. (1997). Optimal linear combinations of neural networks. *Neural Networks*, 10(4):599–614.
- Hashem, S., Schmeiser, B., and Yih, Y. (1994). Optimal linear combinations of neural networks: An overview. In *Proceedings of 1994 IEEE Conf. on Neural Networks*, pages 1507–1512.
- Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150.
- Juditsky, A. and Nemirovski, A. (1996). Functional aggregation for nonparametric estimation. *Publication Interne, IRISA*, (993).
- Kearns, M. J. and Schapire, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *Journal Computer and Systems Sciences*, 48(3):464–.
- Koiran, P. and Sontag, E. D. (1997). Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54:190–198.
- Kurkova, V. (1992). Kolmogorov’s theorem and multilayer neural networks. *Neural Networks*, 5:501–506.
- Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41(3):677–687.
- Madan, R. N. and Rao, N. S. V. (1999). Guest editorial on information/decision fusion with engineering applications. *Journal of Franklin Institute*, 336B(2). 199-204.
- Mojirsheibani, M. (1997). A consistent combined classification rule. *Statistics and Probability Letters*, 36:43–47.
- Nadaraya, E. A. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. Kluwer Academic Publishers, Dordrecht.
- Nolan, D. and Pollard, D. (1987). U-Processes: Rates of convergence. *Annals of Statistics*, 15(2):780–799.
- Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Institute of Mathematical Statistics, Haywood, California.
- Preparata, F. P. and Shamos, I. A. (1985). *Computational Geometry: An Introduction*. Springer-Verlag, New York.
- Rao, B. L. S. P. (1983). *Nonparametric Functional Estimation*. Academic Press, New York.
- Rao, N. S. V. (1991). Computational complexity issues in synthesis of simple distributed detection networks. *IEEE Transactions on Systems, Man and Cybernetics*, 21(5):1071–1081.
- Rao, N. S. V. (1994). Fusion methods for multiple sensor systems with unknown error densities. *Journal of Franklin Institute*, 331B(5):509–530.
- Rao, N. S. V. (1995). Fusion rule estimation in multiple sensor systems using training. In Bunke, H., Kanade, T., and Noltemeier, H., editors, *Modelling and Planning for Sensor Based Intelligent Robot Systems*, pages 179–190. World Scientific Pub.

- Rao, N. S. V. (1996). Distributed decision fusion using empirical estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 33(4):1106–1114.
- Rao, N. S. V. (1997a). Nadaraya-Watson estimator for sensor fusion. *Optical Engineering*, 36(3):642–647.
- Rao, N. S. V. (1997b). Nadaraya-Watson estimator for sensor fusion problems. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 2069–2074.
- Rao, N. S. V. (1998a). A fusion method that performs better than best sensor. In *First International Conference on Multisource-Multisensor Data Fusion*. 19-26.
- Rao, N. S. V. (1998b). To fuse or not to fuse: Fuser versus best classifier. In *SPIE Conference on Sensor Fusion: Architectures, Algorithms, and Applications II*, pages 25–34.
- Rao, N. S. V. (1998c). Vector space methods for sensor fusion problems. *Optical Engineering*, 37(2):499–504.
- Rao, N. S. V. (1999a). Fusion methods in multiple sensor systems using feedforward neural networks. *Intelligent Automation and Soft Computing*, 5(1):21–30.
- Rao, N. S. V. (1999b). Multiple sensor fusion under unknown distributions. *Journal of Franklin Institute*, 336(2):285–299.
- Rao, N. S. V. (1999c). On optimal projective fusers for function estimators. In *IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 1–6.
- Rao, N. S. V. (1999d). On optimal projective fusers for function estimators. In *Second International Conference on Information Fusion*, pages 296–301.
- Rao, N. S. V. (1999e). Simple sample bound for feedforward sigmoid networks with bounded weights. *Neurocomputing*, 29:115–122.
- Rao, N. S. V. (2000). Finite sample performance guarantees of fusers for function estimators. *Information Fusion*, 1(1):35–44.
- Rao, N. S. V. and Iyengar, S. S. (1996). Distributed decision fusion under unknown distributions. *Optical Engineering*, 35(3):617–624.
- Rao, N. S. V., Iyengar, S. S., and Kashyap, R. L. (1993). Computational complexity of distributed detection problems with information constraints. *Computers and Electrical Engineering*, 19(6):445–451.
- Rao, N. S. V. and Oblow, E. M. (1994). Majority and location-based fusers for PAC concept learners. *IEEE Trans. on Syst., Man and Cybernetics*, 24(5):713–727.
- Rao, N. S. V. and Oblow, E. M. (1997). N-learners problem: System of PAC learners. In *Computational Learning Theory and Natural Learning Systems, Vol IV: Making Learning Practical*, pages 189–210. MIT Press.

- Rao, N. S. V., Oblow, E. M., and Glover, C. W. (1994a). Learning separations by Boolean combinations of hyperplanes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(7):765–768.
- Rao, N. S. V., Oblow, E. M., Glover, C. W., and Liepins, G. E. (1994b). N-learners problem: Fusion of concepts. *IEEE Transactions on Systems, Man and Cybernetics*, 24(2):319–327.
- Rao, N. S. V. and Protopopescu, V. (1996). On PAC learning of functions with smoothness properties using feedforward sigmoidal networks. *Proceedings of the IEEE*, 84(10):1562–1569.
- Rao, N. S. V. and Protopopescu, V. (1998). Function estimation by feedforward sigmoidal networks with bounded weights. *Neural Processing Letters*, 7:125–131.
- Rao, N. S. V., Protopopescu, V., Mann, R. C., Oblow, E. M., and Iyengar, S. S. (1996). Learning algorithms for feedforward networks based on finite samples. *IEEE Trans. on Neural Networks*, 7(4):926–940.
- Rao, N. S. V., Reister, D. B., and Barhen, J. (2000a). Fusion method for physical systems based on physical laws. In *Proceedings of 3rd International Conference on Information Fusion*.
- Rao, N. S. V., Reister, D. B., and Barhen, J. (2000b). Information fusion in physical systems using physical laws. In *Proceedings of Eighteenth Symposium on Energy Engineering Sciences*.
- Roychowdhury, V., Siu, K., and Orlitsky, A., editors (1994). *Theoretical Advances in Neural Computation and Learning*. Kluwer Academic Pub., Boston.
- Sima, J. (1996). Back-propagation is not efficient. *Neural Networks*, 9(6):1017–1023.
- Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22(1):28–76.
- Tang, Z. and Koehler, G. J. (1994). Lipschitz properties of feedforward neural networks. Technical report.
- Taniguchi, M. and Tresp, V. (1997). Averaging regularized estimators. *Neural Computation*, 9:1163–1178.
- Tsitsiklis, J. N. and Athans, M. (1985). On the complexity of decentralized decision making and detection problems. *IEEE Transactions on Automatic Control*, AC-30(5):440–446.
- Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergences and Empirical Processes*. Springer-Verlag, New York.
- Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York.
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

- Varshney, P. K. (1996). *Distributed Detection and Data Fusion*. Springer-Verlag.
- Vavasis, S. A. (1991). *Nonlinear Optimization*. Oxford University Press, New York.
- von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. In Shannon, C. E. and McCarthy, J., editors, *Automata Studies*, pages 43–98. Princeton University Press.